Integrative approaches to high-throughput data in lymphoid leukemias

(on transcriptomes, the whole-genome mutational landscape, flow cytometry and gene copy-number alterations)

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Giuliano Crispatzu

aus Düren

Köln, Dezember 2016

Berichterstatter/in: **Prof. Dr. Michael Nothnagel**

Prof. Dr. Bernd Wollnik

Tag der mündlichen Prüfung: 20.2.2017

List of Contents

Abbreviations

AARS2: Alanyl-TRNA Synthetase 2, Mitochondrial

ABL: Abelson murine leukemia viral oncogene homolog 1

AGO2: Argonaute 2, RISC Catalytic Componen

AKT1: RAC-alpha serine/threonine-protein kinase 1

ALCL: Anaplastic Large Cell Lymphoma

ANOVA: Analysis of variance

AP-1: Activator protein 1

API: Application Programming Interface

A-T: Ataxia telangiectasia

ATM: Ataxia telangiectasia mutated

ATR: Ataxia Telangiectasia And Rad3-Related Protein

B-cells: B-lymphocytes; mature in bursa of Fabricius (B)

BCL2: B-cell lymphoma 2

BCL: B-cell lymphoma

BCR: B-cell receptor

BFB: Breakage-fusion-bridge

BFS: Breadth-first-search

BRD4: Bromodomain-containing protein 4

CADD: Combined Annotation Dependent Depletion

CASP8: Caspase-8

CBP: CREB binding protein

CD19: Cluster of differentiation 19

CDS: Coding sequence

CHEK2: Checkpoint Kinase 2

CLL: Chronic lymphatic leukemia

CMC4: Cx9C motif-containing protein 4

CM: Central memory

CML: Chronic myeloid leukemia

CN: Copy-number

CNV: Copy-number variation (germline)

COSMIC: Catalogue of Somatic Mutations in Cancer

CTC: Circulating tumor cells

CTCL: Cutaneous T-cell lymphoma

CTLA4: Cytotoxic T-lymphocyte-associated Protein 4

dbSNP: The Single Nucleotide Polymorphism Database

DDR: DNA damage response

DLBCL: Diffuse large B-cell lymphoma

DLEU7: Deleted In Lymphocytic Leukemia 7

DNA-PKcs: DNA-dependent protein kinase catalytic subunit

DNMT3A: DNA (Cytosine-5-)-Methyltransferase 3 Alpha

DSB: Double-strand breaks

EFO: Experimental Factor Ontology

EP300: E1A Binding Protein P300

ERCC6L2: Excision Repair Cross-Complementation Group 6 Like 2

Erα+ breast cancers: Endrogen-receptor alpha positive breast cancers

EZH2: Enhancer of zeste homolog 2

FACS: Fluorescence-activated cell sorting

FAT: Focal adhesion targeting

FCM: Fludarabine, cyclophosphamide, mitoxantrone

FCR: Fludarabine, cyclophosphamide, rituximab

fcs: Flow Cytometry Standard

FDR: False discovery rate

FISH: Fluorescence in situ hybridization

FOS: FBJ Murine Osteosarcoma Viral Oncogene Homolog

FU: Follow-up

GEO: Gene Expression Omnibus

GEP: Gene Expression Profiling

GFI1: Growth factor independent 1 family

GO: Gene Ontology

GPCPD1: Glycerophosphocholine Phosphodiesterase 1

GPD1L: Glycerol-3-Phosphate Dehydrogenase 1-Like

GUI: Graphical user interface

H3K27me3: Histone 3 lysine 27 trimethylation

HapMap: Haplotype map

HCLS IG: Semantic Web Health Care and Life Sciences (HCLS) Interest Group

HDAC: Histone deacetylases

HGNC: HUGO Gene Nomenclature Committee

HMGXB4: HMG (High mobility group)-Box Containing 4

HPC: High performance computing

HR: Homologous repair

HTML: Hypertext Markup Language

HTTP: Hypertext Transfer Protocol

HUGO: Human Genome Organisation

ICGC: International Cancer Genome Consortium

ID: Identifier

IgD: Immunoglobulin D

IGHV: Immunoglobulin heavy chain variable region genes

IHC: Immunohistochemistry

IL-2: Interleukin 2

IGCNU: Intratubular germ cell neoplasia, unclassified type

IP: Internet Protocol

JAK3: Janus kinase 3

JHDM1D: KDM7A; Lysine (K)-Specific Demethylase 7A

JUN: V-Jun Sarcoma Virus 17 Oncogene Homolog

KEGG: Kyoto Encyclopedia of Genes and Genomes

KIAA1211L: KIAA1211-Like

KLHL6: Kelch-like 6

KRAS: V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog

L1TD1: LINE-1 Type Transposase Domain Containing 1

LCK: Lymphocyte Cell-Specific Protein-Tyrosine Kinase

LDT: Lymphocyte doubling time

LOH: Loss of heterozygosity

MAC: Media access control; address

MECOM: MDS1 And EVI1 Complex Locus

mESCs: Mouse embryonic stem cells

MeSH: Medical Subject Headings

MI: Mutual information

MK2: MAPKAP kinase-2

MLH1: mutL homolog 1

MMR: Mismatch repair

MSH3/MSH4: mutS homolog 3 and 4

MSI: MS instability

MS: Microsatellites

MSSQL: Microsoft SQL (Structured Query Language)

MTCL: Mature T-cell lymphoma/leukemia

MTG1: Mitochondrial Ribosome Associated GTPase 1

MuSiC: Mutational Significance In Cancer

MutSigCV: Mutation Significance with Covariates

MYC: Avian myelocytomatosis viral oncogene homolog

MYD88: Myeloid differentiation primary response gene 88

N3: Notation3; non-XML-based, human-readable RDF format

NBC: Normal B-cells

NFKB1: Nuclear Factor Kappa B Subunit 1

NGS: Next-generation sequencing

NHEJ: Non-homologous end-joining

NSCLC: Non-small-cell lung carcinoma

Oct4: Octamer-Binding Protein 4 (now known as POU5F1: POU Class 5 Homeobox 1)

OMIM: Online Mendelian Inheritance in Man

OS: Overall survival

OWL: Web Ontology Language

OxoG: 8-Oxoguanine

p13MTCP1: P13p8mature T-cell proliferation 1

PARP: Poly(ADP-Ribose) Polymerase

PCA: Principal Component Analysis

PFS: Progression-free survival

PHP: PHP - Hypertext Preprocessor

PI3K: Phosphoinositide 3-kinase

PolE: DNA Polymerase E

PolyPhen: Polymorphism Phenotyping

PPI: Protein-protein interaction

PRKDC: Protein Kinase, DNA-Activated, Catalytic Polypeptide

PSMD12: Proteasome 26S Subunit, Non-ATPase 12

PTMA: Prothymosin, Alpha

qRT-PCR: Real-time quantitative PCR (Polymerase chain reaction)

RAB25: Member RAS Oncogene Family

RadialSVM: Radial basis function kernel-based support vector machine

RB1: Retinoblastoma 1

RDF: Resource description framework

REST: Representational State Transfer

RHOH: Ras Homolog Family Member H

RNF11: Ring Finger Protein 11

ROS: Reactive oxygen species

RPRM: Reprimo; TP53-dependent arrest mediator

sCNA: Somatic copy-number alteration

SEPT10: Septin 10

SH: Src-homology domain

SIDER: SIDER Side Effect Resource

SIFT: Sorting Intolerant from Tolerant

SLAMF6: SLAM (Self-ligand receptor of the signaling lymphocytic activation molecule) Family Member 6

SNP: Single-nucleotide polymorphism (germline)

SNV: Single-nucleotide variant (somatic)

SPADE: Spanning-tree Progression Analysis of Density-normalized Events

SPARQL: SPARQL Protocol And RDF Query Language

STAT5B: Signal Transducer And Activator Of Transcription 5B

STK17B: Serine/threonine kinase 17b

STR: (bi- or trinucleotide) Short-tandem repeats

SVM: Support Vector Machine

SV: Structural variation

T-ALL: Precursor T acute lymphoblastic leukemia/lymphoma

TBCD: Tubulin-specific chaperone D

T-cells: T-lymphocytes; mature in thymus from thymocyte

TCGA: The Cancer Genome Atlas

TCL1A: T-cell Leukemia/Lymphoma 1A

TCR: T-cell receptor

TCRαδ: T-cell receptor alpha/delta

TF: Tumor fraction

T-LGL: T-cell large granular lymphocytic leukemia

TNF: Tumor necrosis factor

TNFSF12: Tumor necrosis factor ligand superfamily member 12

TNG1/TNG2: TCL1-Neighboring Gene 1/2 (now known as TCL6)

TNIP: TNFAIP3 (TNF Alpha Induced Protein 3)-interacting protein

TP53: Tumor protein p53

T-PLL: T-cell prolymphocytic leukemia

TRAJ49: T-cell Receptor Alpha Joining 49

TRAV26-2: T-cell Receptor Alpha Variable 26-2

TRIM22: Tripartite motif-containing 22

TUNAR: Tcl1 Upstream Neuron-Associated lincRNA

UBC: Ubiquitin C

UPD: Uniparental disomy

URI: Uniform Resource Locator

URL: Uniform Resource Identifier

VAF: Variant allele fraction

WBC: White blood cell count

WES: Whole-exome sequencing

WGS: Whole-genome sequencing

WHO: World health organization

XBP1: X-box binding protein 1

XML: Extensible Markup Language

XPO1: Exportin 1

xsd: XML Schema Definition

ZAP70: Zeta-Chain (TCR) Associated Protein Kinase 70kDa

β2-M: Beta-2-microglobulin

1. Introduction

The overall goal of this thesis (in form of a cumulative dissertation) is to develop a systems biology framework in which Next-Generation Sequencing (NGS) and other high-throughput data sets are (compatibly) integrated, readable for humans (in form of text and visualizations) and computers (in form of parsable markup flat-files or databases). This approach generates more specific diagnosis criteria and is potentially leading to ultimately earlier and more efficient treatments. In addition, it provides the option to further integrate and accumulate evidence for basic molecular mechanisms.

1.1 Methodological background

1.1.1 Semantic Web

In the Semantic Web paradigm there is, in contrast to other data storage models like the relational one (of *MSSQL* or *MySQL*), no need to create static tables connected through primary keys and to fulfil specific normalization forms. In addition, the Semantic Web paradigm does not require any underlying infrastructure like a database on a dedicated server to be properly read but can be organized as flat-files (on arbitrary hardware volumes), which are easier to share between collaborative scientists. The data is modeled as a directed network where the edges are interpreted as “predicates”, the source nodes as “subjects” and the target nodes as “objects”. Furthermore, each object may itself be a subject in another context. In particular, this allows logical links between different knowledge domains which in many cases are not obvious. The Semantic Web paradigm is being used intensively to develop the “Web 3.0” (Cheung et al. 2008; Hendler 2003), that is, to further exploit and make sense out of information that is available on highly distributed resources worldwide. To highlight the logical construction of the semantic framework, this is usually presented by graphical networks (see **Figure 1.1**). Technically, it is sufficient to create a subject-predicate-object triple in *RDF* (resource description framework) files, a *W3C* recommendation derived from *XML* (Extensible Markup Language) standard, so that the hierarchical model is extended to a network model which makes it also possible to apply algorithms known from graph theory (Deus et al. 2008). *SPARQL* (SPARQL Protocol And RDF Query Language) is usually used to query the data, similar to what SQL does for relational databases. A possible application and the power of this methodology is demonstrated by means of semantic music recommendations as implemented on the commercial website *last.fm.*
After a user has listened to a piece of music, the system offers some recommendations for music by similar artists. These recommendations are based on shared “attributes” (also termed “properties”, shown in **Figure 1.1** as edges) and a similarity measure. If user ':alex:' has listened to music by the 'Beastie Boys', the semantic framework would recommend further listening to music by 'Adam Yauch' because he is one of the 'currentMembers' of this group. On the other hand, it would recommend music from the same 'genre' 'Hardcore_Punk', e.g. by the group 'Black_Flag'. The innovation here is that the inferences are not made by an administrator, but by the computer itself because it can interpret the standardized RDF the schema is written in. An identical procedure has been used on the *“free reference manager and PDF organizer” Mendeley.com* for scientific article recommendations, *Google Knowledge Graph [URL: https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html], or Facebook's* derivative of it.

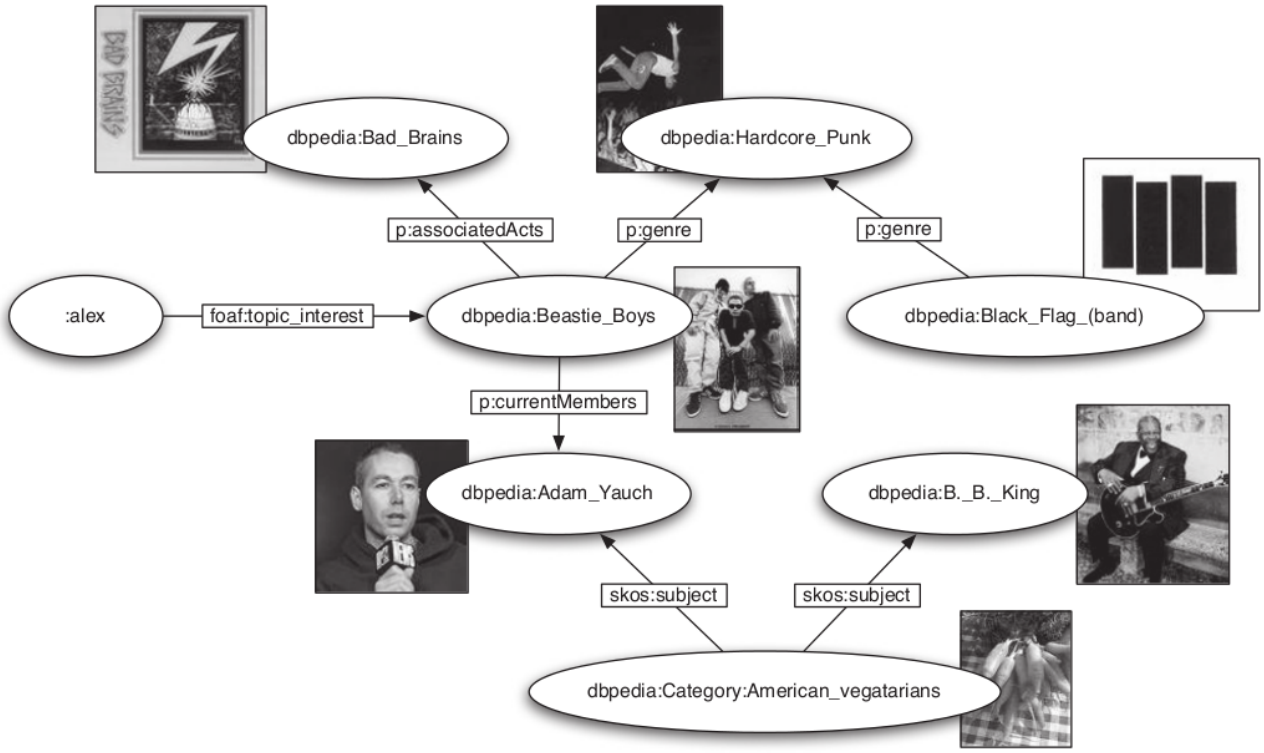


Figure 1.1: Linked dbpedia (Wikipedia entries as RDF) data collections (taken from Dengel et al. 2012)

1.1.2 Current Semantic Web implementations in Life Science

In recent years, there have been tremendous efforts made to set up biological databases into semantic schemas. They are rapidly replacing tedious and unflexible Excel or MySQL table-based data storage and include the EMBL (European Molecular Biology Laboratory) / EBI (European Bioinformatics Institute) RDF platform (Jupp et al. 2014) integrating BioModels, BioSamples, ChEMBL, Ensembl, Expression Atlas, Reactome and UniProt. Each one has its respective SPARQL endpoint and exemplary queries. Bio2RDF (Belleau et al. 2008), which contains as of version 3 in 2014 almost 12M triples, can be used complementary as it has converted mostly data sets not used at the EMBL / EBI, such as dbSNP, BioPortal, DrugBank, KEGG (Kyoto Encyclopedia of Genes and Genomes), MeSH (Medical Subject Headings), OMIM (Online Mendelian Inheritance in Man) or Wormbase. SIDER (SIDER Side Effect Resource), which has been developed at the EBI, but thus far not integrated into their RDF platform, is also available within Bio2RDF. An extension for drug discovery and chemogenomics, called Chem2Bio2RDF, has been previously released (Chen et al. 2010). However, the SPARQL endpoints seem to be taken offline (as of 09/15/2016), but the flat-file can still be obtained and uploaded locally. All these resources use their own controlled vocabulary, or a mash-up obtained through BioPortal (Whetzel et al. 2011) to preset terminology to be used and thus ensure persistent communication between federated SPARQL endpoints and scientists exchanging models written in RDF or a more sophisticated extension capable of inferences, OWL (Web Ontology Language). The most prominent OWL attribute is perhaps 'owl:sameAs' and bidirectionally links two objects / subjects and thus indicating their equality. It can then be inferred in queries that each deregulation coupled to an official gene is the same as the deregulation coupled to its synonym. The EMBL / EBI platforms uses among others, as GO (Gene Ontology) or BioPax (Demir et al. 2010) for pathway or

complex annotation, its own ontology called EFO (Experimental Factor Ontology; Malone et al. 2010) to describe assay results, such as originally gene expression profiling.

Since the creation of ontologies are mostly community-driven, including a long period of feedback-based evaluation (such as surveys), I will limit the description to the level of simple local namespaces and attributes, thus, will not focus on the integration of public data. I will, however, model own data sets within these thesis with terminology already used within public repositories, and comply by recommendations of the *W3C Semantic Web Health Care and Life Sciences Interest Group* (HCLS IG: https://www.w3.org/blog/hcls/). Terminal nodes (subjects) are preferably modelled as URIs (Unified Resource Identifiers) referencing biological entities (e.g. http://bio2rdf.org/hgnc.symbol:TCL1A) linked to persistent URLs (Unified Resource Locations) with human-readable HTML sites when clicked on. These terms can further link two graphs, e.g. overexpression and structural variations affecting the oncogene *TCL1A* (described later), so they have to be consistent thoughout data sets. To achieve this, *identifiers.org* (Juty et al. 2012) offers unambiguous and extensive metadata records. Two commonly used, freely obtainable, „triple stores“ (semantic databases) are OpenRDF Sesame (as used here) and JenaFuseki. The former has the advantage of more precise administration tools and more tolerant data upload, while the latter can be used with the Cytoscape plug-in *RDFscape* (Splendiani et al. 2008) to visualize queries. Both however can be queried using standard SPARQL. Here is an example, where we fetch the official ENSEMBL identifier from the Ensemb SPARQL endpoint (https://www.ebi.ac.uk/rdf/services/ensembl/sparql):

```
PREFIX rdfs: <ht  tp://www.w3.org/2000/01/rdf-schema#>
PREFIX ensembl: <http://rdf.ebi.ac.uk/resource/ensembl/>
SELECT DISTINCT ?ensembl WHERE {
?ensembl rdfs:label "TCL1A" . FILTER(regex(?ensembl, "ENSG[0-9]"))  }
```

Namespaces are in the first two lines, which circumvents writing all the URI bases/prefixes for each corresponding attribute. As the third line is equivalent to SQL, we search for unique („DISTINCT“) Ensembl identifiers („?ensembl“) which („WHERE“) satisfy the graph pattern in the forth line. Variables (objects, predicates or subjects) with a preceding question mark symbolize place holders for specific values matching the graph pattern. Each pattern is terminated with a dot („.“) and further restricts the solution space. In this case only Ensembl identifiers („?ensembl“) with a corresponding label („rdfs:label“) matching „TCL1A“ are returned. Since *TCL1A* has orthologues in many species, we need to further restrict it to the human version. This is facilitated by a regular expression („regex(..)“) with the suffix matching the standard human Ensembl identifier. The query then returns: ensembl:ENSG00000100721. The resulting identifier can further be used to get all triples with it as a subject and thus obtaining meta-information as orthologues, synonyms, UniProt ID or coding information with the following short-cut:

```
PREFIX ensembl: <http://rdf.ebi.ac.uk/resource/ensembl/>
DESCRIBE ensembl:ENSG00000139618
```

Besides directly pasting these queries into the web-interface of the SPARQL endpoint, it can also be adressed on the UNIX command-line taken advantage of the RESTful (Representational state transfer) API (Application programming interface) and result downloads:

```
wget -O test https://www.ebi.ac.uk/rdf/services/ensembl/sparql?query="PREFIX ensembl:
<http://rdf.ebi.ac.uk/resource/ensembl/> DESCRIBE ensembl:ENSG00000139618"
```

Or within the statistical software environment R (R Core Team 2013):

```
library(SPARQL)
query <- "PREFIX ensembl: <http://rdf.ebi.ac.uk/resource/ensembl/> DESCRIBE
ensembl:ENSG00000139618"
SPARQL(url="https://www.ebi.ac.uk/rdf/services/ensembl/sparql", query=query, format="csv")
```

In order to decrease the processing time of SPARQL queries, RDF data sets are stored in different files and uploaded into different graphs. For ease of overview, separate data repositories are further created which can then be accessed using federated queries.

1.2 Biological Background

Throughout evolution, humans developed an immune system to defend against different kinds of pathogens (bacterial, viral, or fungal) encountered over time. As in other vertebrates, but in contrast to e.g. some prokaryotes, this system is subdivided into the innate, as a first barrier and recruiter, and the adaptive immune system with memory capacity. Responsible for the latter system are two kinds of lymphocytes (subtypes of white blood cells).
The first of these present B-cells or B-lymphocytes which mature within the bone marrow and are then released to the blood stream. Once the B-cell receptor (BCR) encounters and binds an antigen (antibody-generating), it secretes antibodies or present them to other immune cells which spawn a response. T-cell progenitors also originate from the bone marrow, but subsequently populate the thymus and differentiate into mature T-cells. Their antigen receptors (T-cell receptors / TCRs) also recognize specific antigens and are further divided into different subtypes. Both, T- and B-cells, go through different differentiation stages, which are characterized by distinct surface marker expressions that are gained or lost, as well as by activation of associated cytokines that are expressed (e.g. interleukin 2, IL-2, by promoting differentiation into e.g. regulatory, memory or effector T-cells) during this transition (**Figure 1.2**). They also generate a large pool (repertoire) of potential responders via somatic hypermutation and rearrangement of their receptor chains, which are positively selected through clonal selection (Hodgkin et al. 2007). Traditionally, and also within this thesis, lymphocytes are characterized using flow cytometry (e.g. FACS, fluorescence-activated cell sorting) in order to identify different surface and cytoplasmic markers. Recently, more deliberate approaches are being developed like measuring marker status by RNA-Seq (transcriptome sequencing) (Jaitin et al. 2014). More concepts of T-cell-mediated immunity are reviewed in Warner, Oberbeck, Schrader et al. (submitted).

Cancer is a proliferative disorder of cells carrying mutations of hierarchical causality, which cumulate over time, mostly leading to loss of function of tumor suppressors and gain of function of proto-oncogenes. These different aberrations are characterized by different hallmarks such as „sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis“ (taken from Hanahan & Weinberg 2011; **Figure 1.3**). Disadvantageous mutations from the perspective of the neoplasm / tumor (and its supporting, surrounding cells; the microenvironment) are selected against (similar to Darwinian selection) and thus give rise to further clones carrying driver mutations

repressing other normal or less fit cells. These descendants may inherit passenger or private (subclonal) mutations. which may only come to effect (rise in variant allele fraction or cancer cell fraction; both measuring tumor allele portion) after additional selective pressure as changing microenvironment or treatment-regimens (e.g. chemo-therapy), thus leading to relapse (Cancer Genome Atlas Research Network 2013).

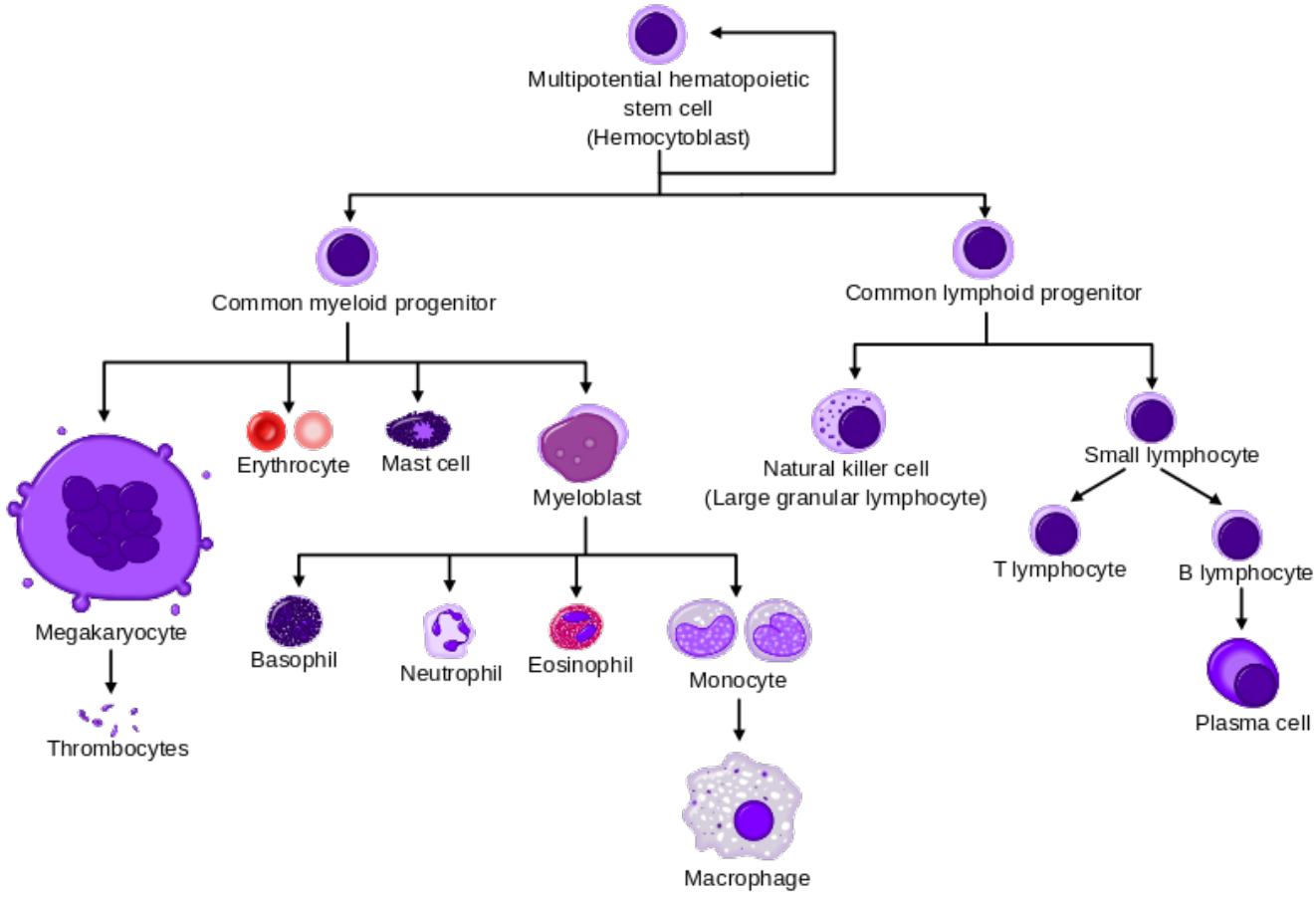


Figure 1.2: Simplified schema of human hematopoiesis. Lymphocyte lineage is depicted on the right. Different activation markers (not shown) are turned on/off during differentiation.
Original by A. Rad, modified by Mikael Häggström. CC BY-SA 3.0

There are five global types of cancers differentiated by tissue of origin, namely carcinomas (of glands and organs), sarcomas (of bone, muscle, fat, or cartilage), melanomas (of the skin), and lymphomas / leukemias (lymphocytes of the blood or lymphoid organs). The latter are often difficult to segregate, especially in the case of (pro)lymphocytic leukemias of mature cells.
Lymphoid neoplasms originate predominantly from cells of the adaptive immune system, namely B- and T-lymphocytes, and are sub-divided into over 50 distinct entities by the WHO (Jaffe 2009). They can occur as primary leukemic forms or as solid lymphomas. They are further divided according to their clinical course in acute and chronic condition, needing either rapid treatment or sequentially evolving and worsening disease status.
Chronic lymphocytic leukemia (CLL), arising from B-cells, and it's T-cell pendant T-cell prolymphocytic leukemia (T-PLL) are two primary leukemic malignancies. Within the scope of my PhD thesis I developed bioinformatical tools to answer fundamental questions concerning the biology of CLL and T-PLL in order to improve treatment and understanding of these, so far, incurable diseases.

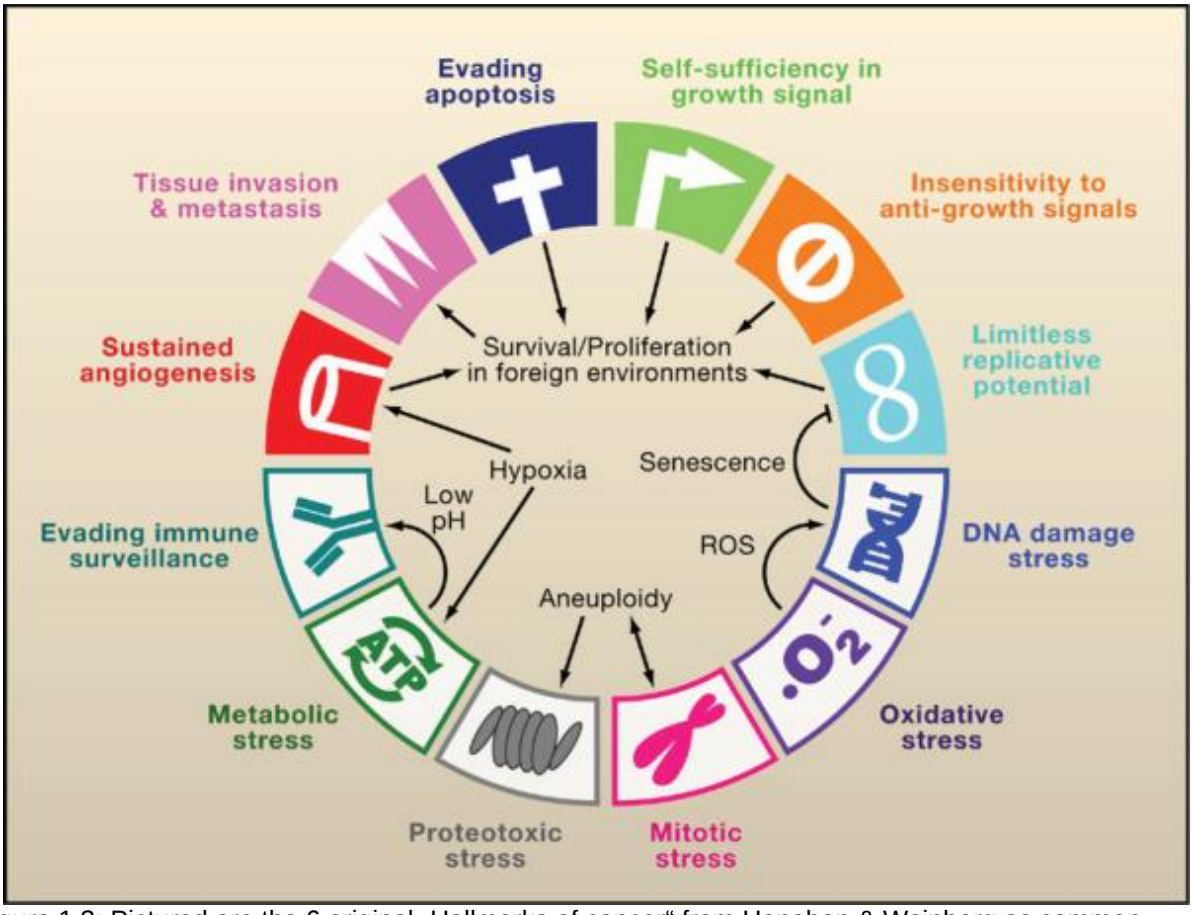


Figure 1.3: Pictured are the 6 original „Hallmarks of cancer“ from Hanahan & Weinberg as common mechanistic ground for cancerogenesis, including 6 additional molecular stress sources from recent investigations. Taken from Luo et al. Cell. 2009 Mar 6; 136(5): 823–837.

1.2.1 Chronic lymphocytic leukemia (CLL)

CLL is the most frequent lymphatic malignancy in Western countries (incidence: 3/100.000) and up to now remains incurable. The median age of diagnosis lies around 72 years and many patients carry relevant comorbidities.
On the molecular level CLL patients are subdivided according to the genetic aberrations found in their malignant lymphocytes, which in many cases are chromosomal deletions affecting tumor suppressors, i.e. in above 50% of cases most often the del(13q14) affecting *mir-15a/mir16-1* and *DLEU7* (not *DLEU2*!) cluster (Klein et al. 2010), as well as del(11q) affecting *ATM* (Ataxia telangiectasia mutated) and del(17p) affecting *TP53* (tumor protein 53), or translocations or duplications like the trisomy 12 (12+) (Doehner et al. 2000; Klein & Dalla-Favera 2010). It has been shown that the expression of specific oncogenic factors like *TCL1A* (T-cell lymphoma/leukemia 1A) is associated with worse prognosis in CLL (Herling et al. 2009). Beta-2-microglobulin (β2-M), *IGHV* (immunoglobulin heavy chain variable region genes) mutational status (pre- or post-germinal center, malignant B-cells of origin) and *ZAP70* (Zeta-Chain (TCR) Associated Protein Kinase 70kDa) serve as further indicators of disease courses, while the latter seems to be the most accurate predictor for genetic risk (Kienle et al. 2010). An interaction of *TCL1A* with *ATM* has previously been described in CLL without 11q- (Garding et al. 2013). High *TCL1A* expression is further correlated with usage of (mutated and unmutated) *IGHV3-21* receptor genes (Mansouri et al. 2010). There is also a morphologically distinct subset to atypical CLL and mantle-cell

lymphoma (another B-cell lymphoma) called B-cell prolymphocytic leukemia and a transformation to a high-grade Non-Hodgkin lymphoma, which ~10% of CLL undergo, called Richter syndrome.
CLL patients are currently diagnosed using a flow cytometry-based analysis of a panel of surface markers. A consensus consists of high CD19 (cluster of differentiation 19), CD23, CD43, CD79a, and intermediate CD20, CD5 expression, as well as weakly expressed surface immunoglobulin M (IgM) and IgD (Gribben 2010). For novel inhibitor studies patient samples are sequenced by targeted capture- or Amplicon-based sequencing. In general CLL is characterized by mutational heterogeneity within *NOTCH1* (notch 1), *XPO1* (exportin 1), *MYD88* (myeloid differentiation primary response gene 88) and *KLHL6* (kelch-like 6) being frequently and clonally mutated (Puente et al. 2011), while subclonal mutations include those in *SF3B1* (splicing factor 3b subunit 1) and *TP53* (Landau et al. 2013).

1.2.2 T-cell prolymphocytic leukemia (T-PLL)

T-PLL is the T-cell pendant to B-CLL/B-PLL and represents with an incidence of 0.6-2.1 per million in Western countries a very rare (approx. 2% of mature lymphocytic leukemias), but also very aggressive mature T-cell leukemia. It is characterized by exponentially rising white blood cell counts (WBC) able to disseminate into spleen and liver, resulting in hepatosplenomegaly (abnormal enlargement of both immunesystemic organs liver and spleen), or skin (in about 20%). The median age at diagnosis is ~65 years, with a median survival of 24 months. Treatment options are currently limited to allogenic stem cell transplantation in younger / physically fit patients and standard chemotherapy (fludarabine, cyclophosphamide, mitoxantrone (FCM)) combined with immunotherapy with Alemtuzumab (anti-CD52). The initiating event in T-PLL is either the inversion or translocation of chromosome 14 (inv(14) / t(14;14)) or the translocation of chromosome X to 14 (t(X;14)) resulting in an juxtaposition of TCL1A / MTCP1 to TCRαδ (T-cell receptor alpha/delta) segments and thus activating *TCL1A* (80%) or p13MTCP1 (P13p8mature T-cell proliferation 1) respectively. Both proto-oncogenes can interact with *AKT1* (RAC-alpha serine/threonine-protein kinase 1) and *AKT2* (PH domain) enhancing their phosphorylation and leading to their nuclear translocation and activation at membrane sites. Common inmmunophenotypes of T-PLL cases include CD4**+**/CD8**-** (60%), CD4**+**/CD8**+** (25%), CD4**-**/CD8**+** (15%), but also CD7**+**, CD5**+** or CD2**+** (Hopfinger et al. 2009). Additional recurrent chromosomal abnormalities (Dürig et al. 2007) involve chromosome 8 (amplifications on 8q (ampl(8q)) believed to affect the oncogene *MYC* (avian myelocytomatosis viral oncogene homolog)), chromosome 11 (deletion affecting *ATM*), but large high-resolution studies have not been performed and the biological mechanisms underlying this fatal disease are still poorly understood. For a more broad overview of profiling data acquired so far in T-PLL, please refer to Schrader, Crispatzu et al. (in review) **Supplementary Table 1.1**.

1.2.3 The proto-oncogene TCL1A

TCL1A's (often abbreviated as TCL1 or Tcl1) usual physiological function is temporally and spatially limited during embryonic development, before it is silenced (Teitell 2005). In adults, the 114-amino-acid protein TCL1A is mainly expressed in (CD3-)CD4-CD8- thymic precursors as immature T-cells, pre B-cells, virgin B-cells (Pekarsky et al. 2001) or plasmacytoid dendritic cells. The post-embryonic activated *TCL1A* is linked to adverse

prognosis in CLL, and T-PLL, as well as diffuse large B-cell lymphoma (DLBCL), where it co-occurs with *MYC* translocation. It is further highly expressed in primary mediastinal B-cell lymphoma (Gualco et al. 2010), blastic natural killer-like T-cell lymphoma (Iqbal et al. 2011), Burkitt's lymphoma, follicular lymphoma, mantle cell lymphoma, nodal marginal zone and splenic marginal zone lymphoma (Aggarwal et al. 2009), as well as in 1-5% cases of Ataxia telangiectasia (Gabellini et al. 2003) where it is linked to telomere dysfunction. In seminoma testicular germ-cell tumors and intratubular germ cell neoplasia, unclassified type (IGCNU) high TCL1A protein expression was recently observed (Lau et al. 2010), suggesting not only a lymphoma/leukemia exclusivity.
Besides the interaction of *TCL1A* with *ATM* described within CLL, other interactions include *NFKB1* (also in CLL) likely by forming a complex wih *EP300* (E1A Binding Protein P300) and *CBP* (CREB binding protein) similar to AKT (Chen et al. 2009) leading to the inhibition of cell death. AP-1-dependent transcription is further inhibited as *TCL1A* interacts with the transcription factors *JUN*, *JUNB* and/or *FOS* (Sivina et al. 2012). Further proof of AKT1 phosphorylation (at site p-Ser.473) comes from Hu et al. 2008, where Oct4 repression reduced Tcl1 expression (and down-regulation of phosphorylated Akt1).
A likely co-activation of other oncogenes may be induced by *TCL1A* and its inhibition of de novo methyltransferases DNMT3A and DNMT3B as proposed in TCL1A-tg (transgenic) mice (Palamarchuk et al. 2012).
Other PPI (protein-protein interaction) partners of TCL1A are visualized in **Figure 1.4a**.
While *TCL1A* is neither frequently somatic mutated, nor is experiencing prominent copy-number losses or gains it is likely activated through enhancer-hijacking (in T-PLL) or mutations and deletions of its regulators experiencing a gain-of-function (e.g. in CLL). Negative regulation is processed by miR targeting of *miR-29b/c*, *miR-181b* (Pekarsky et al. 2006), and *miR-34b/c* (Cardinaud et al. 2009), while *miR-34a* also functions as a (positive) *TP53* inducer (Mraz et al. 2009), as well as the just recently described *miR-484* (Vasyutina et al. 2014) affected by *MECOM* (MDS1 And EVI1 Complex Locus) downregulation in CLL (overview: **Figure 1.4b)**. As there are barely any SNPs or structural variants in germline tissues, one can exclude TCL1-related predispositions for leukemogenesis.
The potential of decreased apoptotic sensitivity induced by *TCL1A* also comes from experiments studying cardiomyopathy, where low protein levels (and *MDR1* SNPs) are associated with higher risk of chemo-induced heart failure in women.
In mice, *TCL1A* is used as a human transgene to induce mouse leukemias resembling either CLL (when transfecting B-cells) or T-PLL (when transfecting T-cells). *MYC* works as a synergistic oncogene, while the overexpression of *TCL1A* may activate the endoplasmic reticulum stress response (Kriss et al. 2012).
TCL1A forms the TCL1 family of oncogenes together with *MTCP1*, *TCL6* (formely split into two isoforms *TNG1* and *TNG2*) and *TCL1B* (formerly known as *TML1*). Although *TCL1B* seems to be not that essential to lymphoma development as TCL1A is, it is overexpressed in lung metastasis-free breast cancers, as well as ERα+ (endrogen-receptor alpha positive) breast cancers when compared with ERα- ones (Badve et al. 2010).
There have also been *TCL1B*-tg mouse models proposed that developed angiosarcoma on the intestinal tract (Hashimoto et al. 2013). *TCL1B* is also highly expressed in mantle cell lymphoma and also interacts with *AKT1*, *AKT2*, *AKT3*, *DNMT3A* and *UBC* (Ubiquitin C). Meanwhile barely is known about *MTCP1* and *TCL6* interactions besides AKT (Auguin et al. 2004).

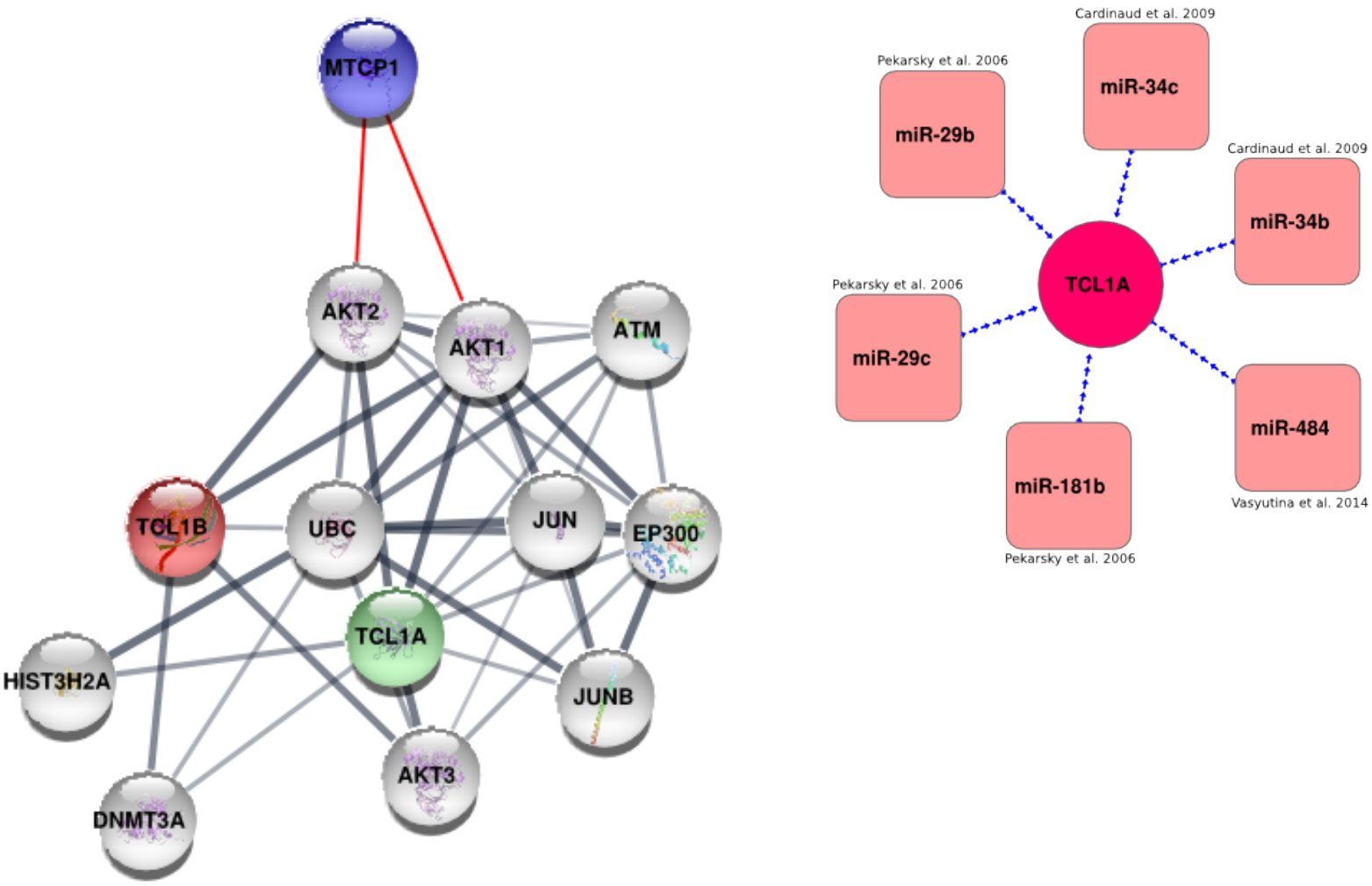


Figure 1.4: a) PPI network of TCL1A and TCL1B obtained from *STRINGdb10* (Szklarczyk et al. 2014). Thickness of edge corresponds to score which in turn corresponds to number of evidence types (co-occurence, data- or literature mining, experimental assays). MTCP1 and its interactions with AKT1 and AKT2 (obtained through *ConsensusPathDB* (Kamburov et al. 2009)) were manually pasted into the graph, because there was no *STRINGdb10* entry. b) Prominent microRNAs targeting *TCL1A* with originating publication.

1.2.4 T-cell receptor (TCR) signaling

The TCRαδ components are located on chromosome 14. In T-PLL, some of its enhancers are translocated/inverted to the *TCL1A* locus likely due to faulty TCR locus recombination events in early thymic immature T-cells (Denny et al., 1986). This genetic alteration does not only lead to post-thymic activation of *TCL1A*, but might also induce TCR hypersensitivity (Herling et al. 2008). In general, genomic enhancers do not have to be in direct proximity to the promoter region, they can rather be several kilobases downstream or upstream (trans-regulatory) in contrast to promoters (cis-regulatory). Direct interactions are created through 3D confirmation changes such as DNA loops (Witte et al. 2015). The expression of a surface TCR (sTCR+) is correlated to and linked to adverse prognosis just like TCL1A+ status and AKT Ser-phosphorylation (Herling et al. 2008).
A similar mechanism of oncogene activation was observed in group 3 and 4 medulloblastoma and coined „enhancer hijacking“ (Northcott et al. 2014), while samples

carrying translocations (or other structural varations) were putting (super-)enhancers next to the proto-oncogenes *GFI1* or *GFI1B* (growth factor independent 1 family).

1.2.5 DNA damage response

DNA damage commonly occurs during replication processes or due to environmental factors like UV light or toxicological substances. The cell has its way of dealing with this by initiating the DNA damage response (DDR) which is comprised of either arresting the cell cycle, repairing the damage or ultimately inducing programmed cell death, called apoptosis. In later stages of genomic instability double-strand breaks (DSB) can appear and are repaired through either homologous repair (HR) by *ATM* among others, or non-homologous end-joining (NHEJ) centrally via DNA-PKcs (DNA-dependent protein kinase catalytic subunit; gene is called *PRKDC*). *ATM* works with *CHEK2* (Checkpoint Kinase 2)/*TP53* to initiate DDR, if this fails the damaged cell goes into apoptosis. Alternatively *ATR* (Ataxia Telangiectasia And Rad3-Related Protein) can induce apoptosis right away through *CHEK1*/*TP53*. Other, earlier lesions such as mismatches are repaired through factors such as *MSH3* or *MSH4* (mutS homolog 3 and 4).
Although, even two decades ago, *ATM* mutations were already characterized in a couple of T-PLL samples (Stilgenbauer et al. 1997), barely any functional validations have been performed on this and other DDR genes within the disease. In CLL however, there is plenty of data available on the consequences and time line occurrences due to the high interest in chemo-resistant 11q- (del(11q)) cases.

Table 1.1: *ATM* deletion and mutation status of both alleles plays a major role in leukemic disease onset. We will later review this in our T-PLL samples by looking at our SNP arrays and whole-exome sequencing data.

1st ATM allele	2nd ATM allele	Consequence
Mutated	Potential wild-type	Normal or impaired DNA damage response. Unknown whether this produces a dysfunctional protein.
Deleted	Potential wild-type	Adverse prognosis in CLL (Döhner et al. 2000)
Deleted	Deleted / mutated	Worse prognosis in CLL (Austen et al. 2007), complete dysfuntional HR. Possible haploinsufficiency.
Germline mutated	Germline mutated	Ataxia telangiectasia (A-T) alias Louis–Bar syndrome and predisposition for lymphoid malignancies

1.2.6 Advanced concepts in tumorigenesis

Due to progress in the field of cancer genomics additional constraints and features have

been added to the „hallmarks of cancer“ to further understand the biology of neoplasms and consequences of invasive treatments.

Gradual changes in chromosomal structure, referred to as genomic instability, give rise to numerous copy-number aberrations for the tumor to pick the most advantageous from. Breakage-fusion-bridge (BFB) cycles represent one of the mechanisms believed to occur on a large scale. When chromatids lack their telomeres (outer ends of human chromosomes), either due to rearrangements or low telomerase activity, they are fused together. During anaphase they are pulled apart and shatter at random places. These aberrations are preserved and get even worse through subsequent replication cycles (modelled in Zakov et al. 2013), e.g. form isochromosomes, where one chromosome arm is lost, while the other is amplified.
An opposing theory is the „catastrophic“ shattering and rearrangement of chromosome parts within a single event in early tumors, referred to as chromothripsis (Zhang et al. 2013; Bassaganyas et al. 2013). Both concepts are recently being investigated using longitudinal sequencing and SNP array data of the same patient.

Telomeres are shortened with each replication cycle, thus symbolizing a type of molecular clock. They are further capped so they do not fuse with other chromosome ends and induce genomic instability. If they reach a critical shortening (Hayflick limit), the carrying cells is either sent to apoptosis through *TP53* (Verdun & Karlseder 2007) or to senescence through *RB1* (Gonzalez-Vasconcellos et al. 2013). *ATM* is also believed to play a role in telomere length maintenance due to its yeast and drosophila homologues Tel1 (Deng et al. 2008).
Evasion of short telomeres from termination is either by inactivation of the senescence or apoptosis pathway genes or by overexpression of telomerase (Röth et al. 2007), which is a polymerase adding TTAGGG repeats at the ends of telomeres.

Many tumors rely on the proliferative and anti-apoptotic signals induced by oncogenes and may be targeted by inhibition of factors encoded by these malfunctioning genes, as done with BCR-ABL (fusion protein) tyrosine-kinase inhibitors for chronic myeloid leukemia (CML; An et al. 2010).
However, other cancer entities or subsets seem not to depend on an oncogene as a central node (hub) to establish an oncogenic phenotype. Disruption of this „non-oncogene addiction“ (Luo et al. 2009) can be achieved by the concept of „synthetic lethality“ (Kaelin. 2005). Genes are considered „synthetic lethal“, where the mutation of one lets the tumor thrive, but the aberration of the other leads to cell death. Candidates are often paralogues or key players in the same or parallel pathway, e.g. as DNA-PKcs within NHEJ and ATM within HR in CLL, or *TP53* and *MK2* (MAPKAP kinase-2) in NSCLC (non-small-cell lung carcinoma) (Morandell et al. 2013).
The ideal cancer treatment consists of selective killing of aberrant cells, while preserving the function of normal, benign cells. In modern times the less precisely form of chemotherapy and radiation exploits DNA damage response and DNA polymerase, while having cytotoxic effect also on benign cells. While more targeted treatments as DNA-PKcs inhibitors (Riabinska et al. 2013) only target cells with a previous deactivation of ATM.

Cancer stem cells carrying properties just like stem cells, i.e. self-renewal capabilities, were identified using experiments in murine model systems, just as the stem memory T-cell population (Zhang et al. 2005; Gattinoni et al. 2009), and it remains to be evaluated to which degree they influence tumor maintenance in humans. It is currently proposed that

they circulate in the blood stream (as CTC; circulating tumor cells) and can re-activate local / non-metastasized tumor cells (Kreso & Dick 2014).

1.2.7 Tumor evolution and clonal hierarchy

Other DNA damage response pathways besides those dealing with DSBs are frequently negatively affected in cancer genomes. „DNA spellchecker“ subsequent to replication are called mismatch repair (MMR) and represent the second most prominent disrupted DNA repair mechanism. The inactivation by mutations of MSH genes (mutS homolog 2 / *MSH2*, *MSH3*, *MSH4*, *MSH5*, *MSH6*) or the hypermethylation of *MLH1* promoter further leads to microsatellite instability (MSI; repeat indels) in colon, rectal, stomach or uterine cancer. Additionally, the inactivation of PolE (proofreading domain of DNA Polymerase E) results in even higher mutation rates. Supek & Lehner 2014 compared single nucleotide variant (SNV) distributions between cancer samples with functional mismatch repair and tumor samples where MMR is rendered dysfunctional. They observed different hotspots suggesting that the MMR mechanism scans damages in essential, euchromatic early replicating genes more efficiently. Once MMR is deficient this bias disappears and the whole genome is equally likely to be affected by somatic mutations. Thus the time of MMR-defiency can be inferred by the flatness of the SNV distribution.

Dating back the malignant cell-of-origin can be achieved by numerous inferences on patient samples. Besides the labor-intensive measurement of 14-C content within a cell (Spalding et al. 2005), computational approaches allow to place samples onto a time line of tumor development. Somatic mutations in microsatellites (MS), bi- or trinucleotide short-tandem repeats (STR), are coupled to mitosis (mostly by replication slippage (mismatches between DNA strands)), and accumulate in normal (and malignant) cells, since there is no (or less) selection pressure in non-coding/non-regulatory regions, and can therefore be used to estimate the depth of a cell by the number of cell divisions since zygote/oocyte (Frumkin et al. 2005). They can also be used, when comparing the diversity of MSI distributions, to estimate tumor specimen age, as was already evaluated in adenocarcinoma and invasive colorectal cancer (Shibata et al. 1996).
Wasserstrom et al. 2008 further observed that animals with mutations in mismatch repair (MMR) genes display very high mutation rates in MS, so 100 alleles were sufficient to estimate with precision above 70% according to simulations. Tandem-repeats can be screened in preferably WGS (whole-genome sequencing) data (including loci with less selective pressure) with tools such as lobSTR (Gymrek et al. 2012) or MSIsensor (Niu et al. 2014).
Another measurement for neoplasia progression in many lymphoid leukemia is the reduction of immune response repertoire, e.g. in TCR (Clemente et al. 2013) especially by the makeup of its Vβ chain (clonal percentage), which is directly responsible for antigen recognition. Tools like miTCR (Bolotin et al. 2013) are able to distinguish common errors (PCR and sequencing) in deep-sequencing assays from somatic hypermutations leading to TCR alpha and beta sequence / peptide variants. Alternatively RNA-seq transcripts can be reconstructed by a combination of de novo assembly and homology search (as done in Warner, Oberbeck, Schrader et al. (review)).
The calculated distances (diversity and similarity) between sample repertoires can be visualized in a phylogenetic tree and thus measuring the degree of monoclonality (time between previous polyclonality and transformation).
Traditionally leukemia samples can be ordered by hematological, kinetical indices like WBC and LDT (lymphocyte doubling time). The former however is reset after cytotoxic

treatment, while the latter is relational i.e. only works with a reference sample of the same patient (Molica et al. 1987).
The most informative approach however still is the sequential sampling of genomics data from patients along their disease course, called clonal evolution or clonality analysis (Landau et al. 2015).
In contrast to the sequential model, as the stepwise adding of subclonal mutations on top of clonal ones due to growth and subsequent selection leading to enhanced survival, recently an alternative model called “Big bang” (Sottoriva et al. 2015) has been proposed, but has only been validated in solid, colorectal cancer. Due to initial intra-tumoral heterogeneity and the absence of selective sweeps (displacement of unfavorable alleles and rise in frequency of favorable alleles due to strong positive selection) in distinct probed sides, an initial carcinogenic “burst” producing clonal and subclonal mutations right away was postulated.

1.3 Aims

1.3.1 Development of a semi-automated pipeline and semantic framework for integrated neoplasia-derived (meta-)data

Demands for bioinformatical analysis tools appropriate for high-throughput data are continuously rising. The Semantic Web paradigm is known to be able to link information previously thought to be unrelated or hard to combine.
In my dissertation, I developed new bioinformatics tools to allow an improved and focussed investigation of high throughput data concerning chronic lymphocytic leukemia (CLL) and T-cell prolymphocytic leukemia (T-PLL). These tools enabled answering fundamental biological questions regarding the biology of these incurable lymphatic malignancies. These questions relate (i) to the most closely related normal counterpart of malignant cells, (ii) to their underlying genetic alterations dividing CLL/T-PLL into new subgroups and (iii) by generating new hypotheses which were then tested in “wet lab” experiments, thus giving rise to the development of more differentiated treatment strategies and disease models.
The core work is focused on the development of new bioinformatics analysis tools for the integration of distinct high-throughput datasets in a semantic manner to automatically generate linkage of knowledge and answer specific questions of lymphoid leukemias. To demonstrate how the Semantic Web paradigm is supposed to be used for this, I will describe in **Chapter 6** the respective semantic models that were applied to the respective data set classes.
In a data-driven approach, key findings were stored in RDF triplets. The information is further retrievable by the user through SPARQL queries and may be combined with other neoplasia data. To achieve this, I used a Java-based software framework that enables specific access and queries to the data models used. This software was then finally used to generate integrated analyses of the data from the laboratories of Dr. M. Herling and Dr. C.D. Herling (formely Schweighofer), in shape of included publications, as well as upcoming manuscripts for the Clinical Research Unit (KFO)-286 ("Exploiting defects in the DNA damage response for the development of novel, targeted CLL therapy"). As a consequence, this process is continuing to be strongly driven by the applications and the nature of the data itself.

1.3.2 Validate the new semantic framework at the informative level in the biological systems of lymphoid leukemias

Applying these newly developed statistical tools, this work aims at an improved biological understanding of the so far incurable diseases T-PLL and CLL. These are based on the integration of distinct data set types. Among the systemic questions that are answered are the following:

- Which mutations are found in genes of a defined expression level?
- Are copy-number variations linked to the transcriptional activity of a gene?
- Is there an allele-specific pattern in the expression of the gained copies?
- Are there mutations that have been acquired in post-transcriptional processes?

Our biological questions for the analysis are focused on three major categories:

A) The genomic landscape of T-PLL

We restrict this sub-aim to T-PLL, as for CLL, recurrent mutations have already shown to discern clinically relevant subsets. Ultimate goal: Derive a refined molecular disease model for T-PLL.

B) Is CLL/T-PLL characterized by a uniform gene expression signature or can heterogeneous subgroups be identified? And how are TCL1 family members affected?

I) Unsupervised**:** if clusters and principal components are formed, what are the genes / gene signatures defining those?

II) Supervised: according to pre-determined strata (categories provided*)*, i.e.

TCL1A status (protein level or at mRNA level *or according to chromosome 14 status*)*: what are these 20% T-PLL that do not express TCL1A and / or show no specific chromosomal aberration?*

Treatment effect: what are the differences between therapy-naive and pre-treated cases?

What are the differences between cases at first diagnosis and samples collected during progressive disease?

Is there a gene signature predicting clinical outcome (long-survivors vs. bad responders)?

Do patient subsets, as defined by their immunophenotype, correlate with subsets on gene expression level?

Do cytogenetic aberrations associate with distinct gene expression profiles?

C) Which normal T-cell subtypes does T-PLL most closely resemble?

To investigate the resemblance of T-PLL tumor cells to physiological counterparts, T-PLL cases were, e.g. compared in their gene expression and immunophenotype profiles to those of normal T-cell controls from different T-cell subsets. To infer on the similarity of phenotypic profiles and clonality, I used (unsupervised) clustering approaches and reconstructed the TCR repertoire.

Human Mutation

Semi-automated cancer genome analysis using high-performance computing

Journal:	*Human Mutation*
Manuscript ID	humu-2016-0221.R1
Wiley - Manuscript type:	Informatics
Date Submitted by the Author:	n/a
Complete List of Authors:	Crispatzu, Giuliano; University of Cologne, Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Bioinformatics Core Facility; University of Cologne, Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Laboratory of Lymphocyte Signalling and Oncoproteome Kulkarni, Pranav; University of Cologne, Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Bioinformatics Core Facility Herling, Marco; University of Cologne, Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Laboratory of Lymphocyte Signalling and Oncoproteome Herling, Carmen; University of Cologne, Department I of Internal Medicine, Center of Integrated Oncology (CIO) Cologne-Bonn, Laboratory for Functional Genomics in Lymphoid Malignancies Frommolt, Peter; University of Cologne, Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Bioinformatics Core Facility
Key Words:	Cancer genomics, Next-Generation Sequencing, Analysis pipeline, Medical bioinformatics

SCHOLARONE™
Manuscripts

Semi-automated cancer genome analysis using high-performance computing

Giuliano Crispatzu[1,2*], Pranav Kulkarni[1*], Marco Herling[2], Carmen D. Herling[3], and Peter Frommolt[1]

[1]Bioinformatics Core Facility, Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Germany

[2]Laboratory of Lymphocyte Signalling and Oncoproteome, Cluster of Excellence on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Germany

[3]Laboratory for Functional Genomics in Lymphoid Malignancies, Department I of Internal Medicine, Center of Integrated Oncology (CIO) Cologne-Bonn, University of Cologne, Germany

* These authors contributed equally to this work.

Corresponding author:

Peter Frommolt

peter.frommolt@uni-koeln.de

This work was supported by the German Research Foundation [grants FR-3313/2-1 to PF, SCHW-1711/1-1 to CDH as part of KFO286, HE-3553/3-1 to MH as part of KFO286] and the German Ministry of Economy and Energy [grant KF2429610MS2 to PF].

For Peer Review

Abstract

Next-Generation Sequencing (NGS) has turned from a new and experimental technology into a standard procedure for cancer genome studies and clinical investigation. While a multitude of software packages for cancer genome data analysis have been made available, these need to be combined into efficient analytical workflows that cover multiple aspects relevant to a clinical environment and that deliver handy results within a reasonable time frame. Here, we introduce *QuickNGS Cancer* as a new suite of bioinformatics pipelines which is focused on cancer genomics and significantly reduces the analytical hurdles that still limit a broader applicability of NGS technology, particularly to clinically driven research. *QuickNGS Cancer* allows a highly efficient analysis of a broad variety of NGS data types, specifically considering cancer-specific issues, such as biases introduced by tumor impurity and aneuploidy or the assessment of genomic variations regarding their biomedical relevance. It delivers highly reproducible analysis results ready for interpretation within only a few days after sequencing, as shown by a re-analysis of 140 tumor/normal pairs from The Cancer Genome Atlas (TCGA). In this re-analysis, the specific calling and filtering strategy of *QuickNGS Cancer* enabled the detection of a significant number of mutations in key cancer genes which were missed by an already well-established mutation calling pipeline.

Introduction

Over the past decade, large-scale cancer genome studies based on Next-Generation Sequencing (NGS) have shed light on tumorigenesis and treatment rationales of a multitude of cancers and novel subtypes (Vogelstein et al., 2013). These efforts were accompanied by the development of many software packages addressing cancer-specific peculiarities in the analysis of the massive amounts of data (reviewed by Ding et al., 2014). Such pecularities are for example the admixture of non-tumor tissue in tumor samples, subclonal heterogeneity through clonal evolution, and chromosomal aneuploidy frequently present in tumor cells. We introduce *QuickNGS Cancer*, an advanced set of computational workflows specifically focused on the analysis of cancer genomics data based on NGS. Our pipeline strongly reduces the time-wise effort for the primary analysis of NGS-based whole-genome (WGS), whole-exome (WXS) as well as whole-transcriptome (RNA-Seq) and targeted sequencing data (amplicon or capture-based) and thus provides significant shortcuts to genetic discoveries of potential clinical and biological importance. The software was

For Peer Review

developed as a computational workflow focused on clinical and experimental cancer research in the context of a large academic hospital.

Materials and Methods

Background: The workflows described in this paper are an important extension to our previously published NGS analysis system *QuickNGS* (Wagle et al., 2015) which is used as a backbone for the basic NGS data workup with *QuickNGS Cancer*. The basic principle of our *QuickNGS* analysis pipelines relies on the organization of available meta data in a MySQL database which is used to control the overall workflow composed of specific software applications for different kinds of analysis. The way in which NGS raw data is processed typically depends on meta information like the NGS library type and sequencing application (WXS, RNA-Seq, etc.), the location of the raw data files on the IT system, the species (human and mouse are supported), details on the submitting laboratory as well as links between samples to be compared from the same individual or patient (i.e. tumor versus non-tumor or follow-up samples). These meta data are fed into the *QuickNGS* database in the background of the pipeline, and the analysis can be started by dropping symbolic links to the raw data files into a dedicated *stack* directory on a multi-node compute cluster. A fully automated and highly standardized analysis procedure is then starting its operations in the background (Figure 1a, Table 1) while extracting all information required for the analysis from the background database. Once the workflow finishes, the results are uploaded into the database and can be accessed by the clinical or experimental scientist on a convenient login-protected website. They are presented in widely used output formats such as Excel tables, PDF files and browsable HTML reports. While the overall workflow is controlled by Bash scripts, the software is based on a careful selection of previously published NGS data analysis software and custom scripts written in Perl and R.

Scope: QuickNGS Cancer specifically extends the *QuickNGS* workflow by the implementation of analytical tools focused on the identification of somatic versus germline gene variation, the visualization of potential tumor-specific genetic alterations, and evaluation of their potential biological and clinical relevance. It encompasses solutions for (1) tumor/normal WXS and (2) tumor/normal WGS, both with automated adoption of tumor purity metrics into downstream analyses, (3) tumor RNA-Seq and (4) targeted sequencing (amplicon or capture-based). As many cancer genome studies currently rely on WXS or WGS data obtained from cancer cells, the WXS

For Peer Review

workflow is the one we focus on to describe in this manuscript.

Analysis approach: In the current version, the workflow for WXS data analysis comprises an initial quality check with FastQC followed by a sequence alignment with BWA (Li and Durbin, 2009) and BAM file post-processing for PCR duplicate removal, local realignment around indels and base quality score recalibration (BQSR) according to the recommendations of the GATK Best Practices (DePristo et al., 2011; van der Auwera et al., 2013). Upon completion of these steps, the pipeline continues the analysis by germline SNP calling with GATK (McKenna et al., 2010) and germline structural variant calling with Delly (Rausch et al., 2012). For somatic mutation calling, the pipeline uses a combination of 4 different mutation callers, namely VarScan2 (Koboldt et al., 2012) for variants with high and MuTect (Cibulkis et al., 2013) for variants with low allele frequency, as well as Strelka (Saunders et al., 2012) and SomaticSniper (Larson et al., 2012), and reports all variants which are detected by at least two of these algorithms. For classification of the germline SNPs as well as somatic mutations regarding their position relative to genes and their effect on protein biosynthesis, the pipeline relies on SnpEff (Cingolani et al., 2012), whereas predictions of their pathogenicity are based on PolyPhen2 (Adzhubei et al., 2010), SIFT (Kumar et al., 2009), MutationTaster (Schwarz et al., 2011), and CADD (Kircher et al., 2014). These predictions are extracted from a database assembled by the developers of ANNOVAR for usage with their software (Wang et al., 2010). The classifications into nonsense, missense and synonymous variants as well as the predictions of their respective pathogenicity are an important vehicle to narrow down the extremely large lists of somatic mutations in order to identify relevant variants of highest clinical or biological interest. In addition, the tumor read fraction as provided in the candidate lists can be used as an indicator to distinguish passenger from driver mutations. For predisposing SNPs, the minor allele frequency as provided in the variant lists can be used to further narrow down the results. The minor allele frequencies are extracted from dbSNP (currently version 147) which includes SNPs from the 1000 Genomes Project as well as the Exome Sequencing Project (ESP). The analysis of somatic copy number gains and losses is based on EXCAVATOR2 (Aurizio et al., 2016). Furthermore, the pipeline uses TitanCNA (Ha et al., 2014) to assess the overall ploidy of the underlying cancer genome and the purity of the tumor samples in a two-step iterative optimization with ploidy set to 2 and purity to 0.5 as initial values. Alternatively, the database can be supplied with an *a priori* known percentage of tumor purity. The estimated purity (or the purity specified in the database) is automatically adopted into the downstream analyses of somatic aberration of the genome. Finally, the Binary Alignment/Map (BAM) files are uploaded into a personalized track hub to be used for visualization on the UCSC Genome Browser (Kent et al., 2002).

For Peer Review

The workflow for WGS analysis is mostly composed of the same steps as the WXS pipeline. However, the WGS copy number analysis is based on a gene-wise segmentation of the genome because ExomeDepth operates on targeted regions such as the exome, not the entire genome. For the analysis of amplicon-based targeted sequencing data, the removal of PCR duplicates is skipped during BAM file preprocessing because the presence of duplicates is actually desired for amplicon sequencing. As targeted panel sequencing frequently does not comprise matched normal samples, our software offers a modification of the WXS workflow where all parts comprising comparisons between tumor and non-tumor samples can be bypassed.

The RNA-Seq pipeline is specifically designed for the discovery of fusion transcripts and the detection of differences in gene expression between tumor and non-tumor samples. In detail, the workflow comprises an initial quality check with FastQC, a basic sequence alignment with Tophat2 (Kim et al., 2013), a search for cancer-specific fusion transcripts with JAFFA (Davidson et al., 2015), gene quantification with Cufflinks2 (Trapnell et al., 2010) as well as the analysis of differentially expressed and differentially spliced genes between tumor and non-tumor samples using DESeq2 (Love et al., 2014) and DEXSeq (Anders et al., 2012). The data is visualized by wiggle files uploaded into a personalized track hub for usage in the UCSC Genome Browser.

Integration into QuickNGS: We have integrated the *QuickNGS Cancer* workflows described in this paper as an add-on into the framework of our previously published *QuickNGS* analysis system. A typical installation is operated by expert staff in a central genomics or bioinformatics lab, whereas clinical or experimental scientists can use the system by getting access to a personalized login area and understand the results without specific knowledge in bioinformatics or NGS analysis. The integration into the *QuickNGS* framework makes the *QuickNGS Cancer* analyses very efficient and, in principle, scalable to cancer cohorts of arbitrary size limited only by the availability of hardware resources.

Software Availability: The source code can be obtained from http://bifacility.uni-koeln.de/quickngs/web under the General Public License (GPL3).

Results

In order to demonstrate the mode of operation and efficiency of *QuickNGS Cancer* and to highlight its potential impact to the cancer genomics field, we have used the system to re-analyze a (random)

John Wiley & Sons, Inc.

For Peer Review

selection of 140 tumor/normal exome pairs from The Cancer Genome Atlas (TCGA). Among the malignancies covered by the analysis are acute myeloid leukemia (AML), urothelial carcinoma, lower grade glioma, invasive breast carcinoma, colon adenocarcinoma, renal clear cell carcinoma, hepatocellular carcinoma, ovarian serous cystadenocarcinoma, pancreatic adenocarcinoma, and prostate adenocarcinoma (10 cases each) as well as lung adenocarcinoma and melanoma (20 cases each). After providing sample information and the raw file locations for the 280 samples (140 tumors and 140 normals) to the *QuickNGS* background database, we linked the 560 FastQ files (forward and reverse reads each) into the *QuickNGS* stack directory. These preparing steps could be finished in less than one hour. The reads for patient TCGA-4T-AA8H, for instance, were 101bp long with an overall count of 135.6M reads (tumor) and 115.4M reads (normal). For this patient, the overall computations took 212.76 CPU hours with a peak memory usage of 32.1 GB. The total time requirement highly depends on the degree of parallelization that can be achieved and thus on the availability of high-performance computing (HPC) resources. Importantly, our approach is scalable to arbitrarily many parallel instances of the pipeline. Upon completion, the software created a browsable analysis report (Figure 2a) providing access to the following files:

- Lists of (1) somatic point mutations (i.e. single nucleotide variants or small insertions and deletions), (2) somatic structural variants, and (3) somatic copy number alterations in three Excel files for each tumor/normal pair. Each table is enriched with comprehensive annotations describing the variants' role in the tumor and potential pathogenic impact
- 21genome in two Excel files for each tumor/normal pair. The tables are enriched with the same annotations as those for the somatic variants
- Per-chromosome plots of somatic copy number aberrations (Figure 2b) and loss of heterozygousity (LOH) for each tumor/normal pair
- Barplots summarizing the total size of somatic copy number aberrations (Figure 2c), the number of somatic mutations (Figure 2d) for all tumor/normal pairs as well as the target enrichment performance in the NGS library preparation
- A table summarizing the characteristics of all tumor/normal pairs (Table 3) and a table with statistics on the NGS libraries for all samples
- Two FastQC reports for the forward and reverse reads of each sample in the analysis
- Link for quick visualization of the BAM files using a local track hub for the UCSC Genome Browser (Kent et al., 2002).

For Peer Review

For instance, *QuickNGS Cancer* discovered 2904 somatic mutations in the 10 renal clear cell carcinoma samples. Among these mutations, we observed 1219 transitions (mutations from a pyrimidine base to a pyrimidine base or from a purine base to a purine base) compared to 1355 transversions (mutations from a pyrimidine base to a purine base or vice versa) and 330 small insertions and deletions. Among the 2904 somatic mutations, 592 were identified to cause a change in the resulting amino acid sequence (non-synonymous mutations).

The results of our meta analysis are summarized in Table 2. We observed the highest rates of non-synonymous mutations for cutaneous melanoma (13.4/Mb), colon adenocarcinoma (11.2/Mb) and lung adenocarcinoma (9/Mb). The mutation rates computed by *QuickNGS Cancer* compare well with previously published mutation rates for the cancer types analyzed (Figure 1 in Kandoth et al., 2013). Next, we compared the mutation counts in the *QuickNGS Cancer* results to those from an official analysis by the TCGA consortium using the *Firehose* pipeline (http://www.broadinstitute.org/cancer/cga/Firehose) which we obtained from the *Firebrowse* portal (http://firebrowse.org). The number of mutations called by *QuickNGS Cancer* deviates by less than 20 mutations from the number called by the *Firehose* pipeline for 6 of the tumor types analyzed (bladder, brain, kidney, liver, ~~ovary,~~ pancreas, and prostate; Table 2a). For 4 tumor types, the average mutation count obtained by *QuickNGS Cancer* exceeded that from the *Firehose* pipeline by 20 or more (AML, breast, ovary and skin), whereas the average count was smaller by 20 or more for the remaining 2 tumor types (colon and lung). In order to also judge the quality of the mutation calls, we checked the mutation status of the 10 most frequently mutated genes in each cancer type according to the International Cancer Genome Consortium (ICGC) Data Portal (e.g. https://dcc.icgc.org/projects/KIRC-US). In total, 182 of the mutations in these key genes could be detected by both analysis approaches (*QuickNGS Cancer* and *Firehose*), 120 were detected only by *QuickNGS Cancer* and 27 could be detected only by the *Firehose* pipeline (Figure 1b, Table 2b; 13 samples from the colon and ovary cohorts excluded). The mutation rates and the actual lists of mutations were extracted from two different tables on the *Firebrowse* portal. Examples of additional calls of *QuickNGS Cancer* are listed in Supplementary Table 1 alongside with the respective read coverages on both alleles in the tumor and normal samples.

The largest average overall size of regions with somatic copy number gain (amplifications) occurred in ovarian serous cystadenocarcinoma (462.7 Mb) and lung adenocarcinoma (281.5 Mb), whereas the largest average overall size of regions with copy number loss (deletions) occurred in ovarian serous cystadenocarcinoma (460.7 Mb) and urothelial carcinoma (294.2 Mb). As no files with processed copy number information based on WXS data are provided by the TCGA data portal, we overlapped somatic copy number aberrations computed by our pipeline with SNP array

data obtained from the TCGA data portal (Table 2a). The given percentage represents the fraction of SNP array-based regions with aberrant copy number which could also be detected from the WXS data after our analysis with *QuickNGS Cancer*. The largest overlap of amplified regions was observed for hepatocellular carcinoma (84.1%) and the smallest overlap for pancreatic adenocarcinoma (49.3%). For regions of copy number loss, the largest overlap was observed for hepatocellular carcinoma (94.5%) and the smallest overlap for prostate adenocarcinoma (34.1%).
The complete list of results for all 140 samples reveals that the mutation rates as well as the overall size of somatic copy number events are highly variable for all cancer types (Supplementary Table 2).

Discussion

We have introduced here *QuickNGS Cancer*, a computational analysis system which allows semi-automated analyses of high-throughput NGS data with a specific focus on the evaluation of genetic data sets obtained from cancer specimens. Our system provides rapid data processing and practical usability with minimal user interaction required. In comparison with other analysis approaches (Table 4), *QuickNGS Cancer* is the only one to automatically estimate the tumor purity from NGS data and use this estimate in the downstream steps of the analysis. Furthermore, *QuickNGS Cancer* can be used for a comparatively large scope of applications and provides a comprehensive and handy report on somatic variation of the cancer genome. In comparison with other analysis approaches for cancer genomics, it is the only software which is also applicable to mouse data. Finally, *QuickNGS Cancer* inherits the efficiency of its overall approach from the actual *QuickNGS* platform. This efficiency is enabled by the overall workflow being controlled by the database at the core of the system. Thus, our new pipelines make the high degree of automation and reproducibility of the actual *QuickNGS* platform accessible also to cancer genome analysis.
We have demonstrated the high efficiency of the overall approach as well as the practical usability of the system to quickly create large-scale analyses with a scope of results that is currently considered state of the art. We have shown this by means of a re-analysis of 140 tumor/normal exome pairs from TCGA. As our pipeline uses an integrative approach by employing a combination of four somatic mutation callers, for instance VarScan2 (Koboldt et al., 2012) for calls of mutations with high and MuTect (Cibulkis et al., 2013) for low allele frequency (Wang et al. 2013), we obtained slight differences in the total number of somatic mutations observed (Table 2a, Supplementary Table 2). *QuickNGS Cancer* was able to detect a significant number of mutations in

key cancer genes which were missed by the *Firehose* pipeline, whereas the *Firehose* pipeline detected only a few mutations missed by *QuickNGS Cancer* (Table 2b). While we have not systematically assessed the false-positive and false-negative rates of the mutation calls from *QuickNGS Cancer*, its proven potential to find mutations in genes which are known to be frequently mutated in the respective cancer type underlines the superiority of our analysis approach over other methods. We believe that this improvement can be attributed to the fact that a consensus calling approach employing several different mutation callers generally outperforms any individual caller (Ewing et al., 2016).

In the comparative analysis of the results on somatic copy number aberrations between *QuickNGS Cancer* and SNP array data obtained from the same samples, the aberrations were highly reproducible for most, but not all samples (Table 2a, Supplementary Table 2). Given that the data were generated by two completely different laboratory assays, a variability between the regions that both methods detect as aberrant is to be expected, in particular in the presence of tumor/normal contamination in both approaches.

In summary, *QuickNGS Cancer* minimizes time and effort for comprehensive cancer genome data analysis based on multiple NGS applications and makes high-quality data analyses accessible also for non-expert researchers within a reasonable time frame. The code is available online and can be adopted by any lab. While scalable to unlimited sample throughput, *QuickNGS Cancer* is capable of boosting cancer genome analyses to population-scale studies enabled by the most recent developments in sequencing technology. Besides its attractive front-end, the *QuickNGS* framework further offers a back-end with an integrated MySQL database making it possible to combine NGS analysis results, e.g. RNA-Seq and WXS, to elucidate allele-specific expression or gain/loss-of-function by expert SQL queries. Future tasks in the development of our pipelines will be an analysis module for the identification of significantly mutated genes, mutation hotspots and co-occurences as described for instance by Cheng et al., 2015, as well as the development of automated searches for viral sequences (transcripts and integration sites) in tumor samples. The visualization can be extended by including graphical representations of structural variations as e.g. with Circos (Krzywinski et al. 2009). Finally, new features could enable *comparative* analyses of cancer genomes between different study cohorts or sequential time points.

For Peer Review

Acknowledgements

The results are based on data generated by The Cancer Genome Atlas (TCGA) research network

(http://cancergenome.nih.gov).

Funding

This work was supported by the German Research Foundation [grants FR-3313/2-1 to PF, SCHW-1711/1-1 to CDH as part of KFO286, HE-3553/3-1 to MH as part of KFO286] and the German Ministry of Economy and Energy [grant KF2429610MS2 to PF].

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-249.

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. Genome Res 22(10):2008–17.

Aurizio RD, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. 2016. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Res; *in press*

van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinformatics 43:11.10.1-11.10.33.

Bao R, Hernandez K, Huang L, Kang W, Bartom E, Onel K, Volchenboum S, Andrade J. 2015. ExScalibur: A High-Performance Cloud-Enabled Suite for Whole Exome Germline and Somatic Mutation Identification. PLoS One 10(8):e0135800.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, et al. 2012. The cBio cancer genomics portal: an open platform for

For Peer Review

exploring multidimensional cancer genomics data. Cancer Discov 2(5):401-4.

Cheng F, Zhao J, Zhao Z. 2015. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. Brief Bioinform 2015:1-15.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31(3):213-9.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. 2012. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet 3:35.

Davidson NM, Majewski IJ, Oshlack A. 2015. JAFFA: High sensitivity transcriptome-focused fusion gene detection. Genome Med 7(1):43.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491-498.

Ding L, Wendl MC, McMichael JF, Raphael BJ. 2014. Expanding the computational toolbox for mining cancer genomes. Nat Rev Genet 15:556-70.

Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY, ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, Kellen MR, Norman TC, Haussler D, Friend SH, Stolovitzky G, Margolin AA, Stuart JM, Boutros PC. 2015. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods 12: 623-30.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451-5.

Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks

For Peer Review

E, Biele J, Ding J, et al. 2014. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res; 24(11):1881-93.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, et al. 2013. Mutational landscape and significance across 12 major cancer types. Nature 502(7471):333-9.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12(6):996–1006.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36.

Kim D, Salzberg SL. 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12(8):R72.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014): A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet; 46(3): 310-5.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22(3):568-76.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an Information aesthetic for comparative genomics. Genome Res 19(9):1639-45.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073-81.

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28(3):311-7.

For Peer Review

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–60.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297-303.

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, Kumararatne D, Doffinger R, Nejentsev S. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28(21):2747-54.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28(18):i333–9.

Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetam RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 28 (14): 1811-7.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7(8):575-6.

Sloggett C, Goonasekera N, Afgan E. 2013. BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 29(13):1685-6.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511-5.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kienzler JW. 2013. Cancer Genome Landscapes. Science 339(6127):1546-58.

For Peer Review

Wagle P, Nikolic M, Frommolt P. 2015. QuickNGS elevates Next-Generation Sequencing data analysis to a new level of automation. BMC Genomics 16:487.

Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, Dahlman KB, Pao W, Zhao Z. 2013. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med 5:91.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164.

For Peer Review

Task	Software (version)	Reference
Basic QC	FastQC (0.10.1)	
Read Alignment	BWA (0.7.7)	Li and Durbin, 2009
BAM file post-processing	GATK (3.3.0): - IndelRealigner - BaseRecalibrator	McKenna et al., 2010
	Picard (1.88): - PCR duplicate removal	http://broadinstitute.github.io/picard
Germline SNP calling	GATK (3.3.0): - UnifiedGenotyper	McKenna et al., 2010
Germline SV calling	Delly (2.0.1)	Rausch et al., 2012
Tumor purity and ploidy	TitanCNA (1.8.0)	Ha et al., 2014
Somatic mutation calling	VarScan2 (2.3.7)	Koboldt et al., 2012
	MuTect (1.1.4)	Cibulkis et al., 2013
	SomaticSniper (1.0.5.0)	Larson et al., 2012
	Strelka (1.0.15)	Saunders et al., 2012
Somatic SV calling	Delly (2.0.1)	Rausch et al., 2012
Evaluation of variants	SNPeff (3.4)	Cingolani et al., 2012
	ANNOVAR (2015-12-14): - PolyPhen2 predictions - SIFT predictions - MutationTaster predictions - CADD predictions	Wang et al., 2010
Copy number analysis	ExomeDepth (1.1.6)	Plagnol et al., 2012
	EXCAVATOR (1.1)	Magi et al., 2013
Raw data visualization	UCSC Genome Browser	Kent et al., 2002

Table 1: List of the software tools used by the WXS pipeline of *QuickNGS Cancer* as of version 1.2.1. The selection of softwares is likely to be modified according to the future evolution of NGS analysis algorithms. The tools used in the most recent *QuickNGS Cancer* release will be available on the *QuickNGS* website which also contains information on the software tools used by other *QuickNGS Cancer* pipelines.

Tissue	n	Purity		Ploidy	Mutations		Mutation Rate	Amplified		Deleted	
		QuickNGS [%]	TCGA [%]	QuickNGS (std. dev.)	QuickNGS [count]	TCGA [count]	QuickNGS [1/Mb]	QuickNGS [Mb]	Overlap w/ TCGA [%]	QuickNGS [Mb]	Overlap w/ TCGA [%]
AML	10	45.5	100.0	0.1	38.5	8.8	1.1	75.6	49.5	59.0	34.1
Bladder	10	62.6	78.2	0.3	192.3	201.4	5.8	188.0	83.9	294.2	78.8
Brain	10	74.6	67.5	0.2	31.0	30.5	0.9	180.2	61.3	251.1	84.4
Breast	10	68.7	75.5	0.2	108.7	36.5	2.0	239.8	81.1	90.2	77.0
Colon	10	59.2	74.5	0.2	503.8	552.7	11.2	267.3	67.3	147.0	57.4
Kidney	10	52.5	78.0	0.4	59.2	47.3	1.5	181.9	79.6	165.7	70.9
Liver	10	79.1	87.5	0.3	77.5	87.5	1.7	247.9	84.1	177.8	94.5
Lung	20	51.8	76.8	0.2	299.5	391.2	9.1	281.5	76.7	227.3	65.9
Ovary	10	82.7	90.8	0.2	70.4	40.5	2.0	462.7	81.7	460.7	83.8
Pancreas	10	43.6	60.5	0.3	56.1	68.6	1.7	137.9	49.3	77.3	44.8
Prostate	10	35.3	71.5	0.2	53.2	43.6	1.6	73.6	49.6	110.6	48.1
Skin	20	60.6	85.8	0.1	442.2	212.7	13.4	190.5	67.9	290.7	62.9

(Table 2a)

Tissue	Key genes	n	Mutations		
			QuickNGS only [count]	TCGA only [count]	Both [count]
AML	PTPN11, TP53, NOTCH1, DNMT3A, KCNJ12, KMT2D, WT1, NRAS, IDH1, KIT	10	4	1	6
Bladder	TP53,LRP1B,LRP2,FGFR3,RYR2,KDM6A,LRP1,SACS, RYR1,COL7A1	10	15	2	11
Brain	IDH1,TP53,ATRX,CIC,NOTCH1,FUBP1,STK19,NF1, PTEN,ARID1A	10	1	0	20
Breast	PIK3CA, TP53, TTN, TTN-AS1, RP11-245C23.3, PCDHGA1, PCDHGA2, PCDHGA3, PCDHA1, CDH1	8	4	1	7
Colon	APC, PCDHA1, PCDHA3, TTN, PCDHA2, CTC-554D6.1, PCDHA4, TTN-AS1, PCDHA5, PCDHA6	0	N/A	N/A	N/A
Kidney	VHL, snoU13, MUC4, PBRM1, TTN, TTN-AS1, MUC16, PCDHGA1, CROCCP2, PCDHGA2	9	10	3	11
Liver	TP53,ARID1A,ALB,LRP1B,RYR2,ARID2,AXIN1,FBN2,ABCA13,APOB	8	6	3	6
Lung	TP53, TTN, TTN-AS1, MUC16, CSMD3, PCDHGA1, RYR2, PCDHGA2, PCDHGA3, ZFHX4	12	24	8	39
Ovary	TP53,BRCA1,RYR2,PKHD1,LRP2,RB1,NF1,TENM2, ABCA3,RYR1	5	4	4	5
Pancreas	TP53,SMAD4,CDKN2A,KRAS,ARID1A,KMT2C,LRP1B,TTN,RYR2, TGFBR2	10	10	3	23
Prostate	TP53,KMT2C,FOXA1,PTEN,RYR2,MYO15A,FBN1, LRP1B,TTN,CACNA1E	10	9	4	2
Skin	BRAF,LRP1B,MGAM,PKHD1L1,SCN11A,SCN10A, NRAS,CACNA1E,SCN5A,MYO18B	19	33	4	56

(Table 2b)

Table 2: Results of a re-analysis of the exomes for 140 tumor/normal pairs from The Cancer Genome Atlas (TCGA). **(a)** The values represent the mean across all samples of the respective cancer entity. For genomic ploidy, the standard deviation is given instead of the mean (which is close to 2 in all cases) in order to highlight how aneuploidy varies across samples. For somatic mutations, the results computed by the *QuickNGS Cancer* pipeline are displayed together with data obtained from the Firebrowse portal (http://firebrowse.org). In addition, the overall sizes of regions with aberrant copy number according to *QuickNGS Cancer* are given together with the percentage of these regions which is also present in SNP array data obtained from *Firebrowse*. **(b)** Number of mutations in key genes of the respective cancer types according to the ICGC Data Portal. *n* represents the number of samples for which detailed mutation data was available not only from *QuickNGS Cancer*, but also from the *Firebrowse* portal, and the number of mutations is reported only for these samples. The table shows how many mutations could be detected only by *QuickNGS Cancer*, only by the *Firehose* pipeline and by both systems.

For Peer Review

Tumor	Normal	Purity	Ploidy	MutRate	Amplified	Deleted
TCGA-A3-3308-01A-01D-0966-08	TCGA-A3-3308-11A-01D-0966-08	80	2.4	1.2	132.4	5.7
TCGA-A3-3317-01A-01D-0966-08	TCGA-A3-3317-11A-01D-0966-08	80	2.1	1.6	32.9	38.4
TCGA-A3-3358-01A-01D-1534-10	TCGA-A3-3358-11A-01D-1534-10	60	1.7	1.4	5.0	147.6
TCGA-A3-A6NL-01A-11D-A33K-10	TCGA-A3-A6NL-11A-11D-A33K-10	85	2.4	2.7	28.9	101.1
TCGA-B0-4818-01A-01D-1501-10	TCGA-B0-4818-11A-01D-1501-10	90	2.1	1.1	155.9	9.1
TCGA-B0-4852-01A-01D-1501-10	TCGA-B0-4852-11A-01D-1501-10	85	2.1	1.6	322.0	66.4
TCGA-B0-5075-01A-01D-1462-08	TCGA-B0-5075-11A-01D-1462-08	75	2.1	2.1	260.1	276.7
TCGA-B0-5077-01A-01D-1462-08	TCGA-B0-5077-11A-01D-1462-08	80	2.3	1.3	229.5	126.5
TCGA-B0-5080-01A-01D-1501-10	TCGA-B0-5080-11A-01D-1501-10	80	2.9	0.6	212.3	0.0
TCGA-B0-5084-01A-01D-1462-08	TCGA-B0-5084-11A-01D-1462-08	85	1.7	1.6	263.0	867.3

Table 3: Tumor statistics provided by the *QuickNGS Cancer* pipeline after an analysis of paired tumor/normal WXS data obtained from 10 renal clear cell carcinoma patients. Key features of the individual tumors are summarized in one table to provide the user with a quick overview of the peculiarities of the cancer exomes analyzed and their variation across patients.

For Peer Review

	QuickNGS Cancer	**Galaxy with BioBlend** (Giardine et al. 2005; Sloggett et al., 2013)	**ExScalibur** (Bao et al., 2015)	**cBio in R (CDGS-R)** (Cerami et al., 2012)
Species	Human and mouse	All	Human	Human
Applications	- germline and somatic SNVs/ InDels and structural variants - copy number analysis - fusion transcripts - differential gene expression and splicing	Universal framework	- germline and somatic SNVs/ InDels	- somatic SNVs/InDels - copy number analysis - differential mRNA expression or protein status
Protocols	WXS, WGS, panel sequencing, RNA-Seq	Universal framework	WXS	WXS, WGS, RNA-Seq, mRNA arrays, SNP arrays, prosphoproteomics
Reproducibility	Results kept in database	Repeat analysis based on workflow file	Results archived	N/A
Purity / ploidy estimates	yes	can be integrated	no	no
Architecture	HPC Database Webserver	HPC Webserver	HPC	Local installation (R API)
User interaction / automation	Low / high	Low / high	Low / high	High / high
Ease of use	- Copy data to HPC cluster - Upload meta data to DB - Link files into stack directory	- Upload data to webserver - Start workflow in a web browser window	- Copy data and configuration files to HPC cluster - Start script in a shell window	Applicable only to processed or public data
Scalibility & extendability	Requires shell programming	Workflow editor	Requires programming	Applicable only to processed or public data

Table 4: Comparison of key features between *QuickNGS Cancer* and other freely available cancer genomics analysis suites. Our software is the only one to employ an integrated purity and ploidy estimation and also uses this for downstream analysis. The scope of applications covered by *QuickNGS Cancer* is the largest among all analysis systems. In contrast to the other softwares, *QuickNGS Cancer* is able to also handle mouse data. Finally, *QuickNGS Cancer* inherits from the actual *QuickNGS* system its database-supported approach by which the sample meta data as well as the analysis results are managed in a very efficient way. This makes all analyses highly reproducible and enables minimum requirements for interactions by the user.

John Wiley & Sons, Inc.

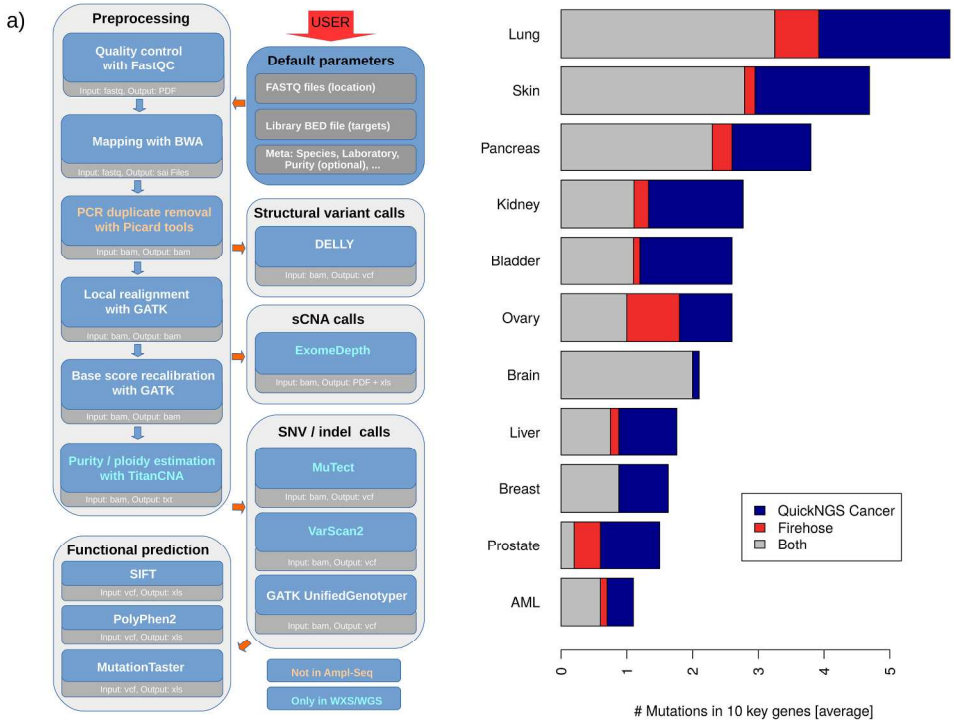


Figure 1: Features of the QuickNGS analysis workflows. (a) Flow chart describing the workflows for targeted gene panels, WXS and WGS. To initiate the analysis, the user uploads a text file with FastQ file names as well as a BED file describing the target library (e.g. TruSeq for targeted panels, NimbleGen SeqCap EZ 2 or 3, Agilent SureSelect V4 for WXS). In addition, the user provides meta information on the samples such as a sample label, the species and the laboratory which has generated the data. Estimates of the tumor purity (e.g. obtained by pathology review or cell sorting) can either be provided by the user or will be estimated with TitanCNA. After this information has been provided, the pipeline is started in a fully automated way. (b) The mutation calling strategy of QuickNGS Cancer enables the detection of mutations in key cancer genes of 140 tumor samples obtained from The Cancer Genome Atlas (TCGA). In particular, the QuickNGS Cancer workflow discovered more key gene mutations in these samples that are missed by the Firehose pipeline than vice versa.

Figure 1
228x175mm (300 x 300 DPI)

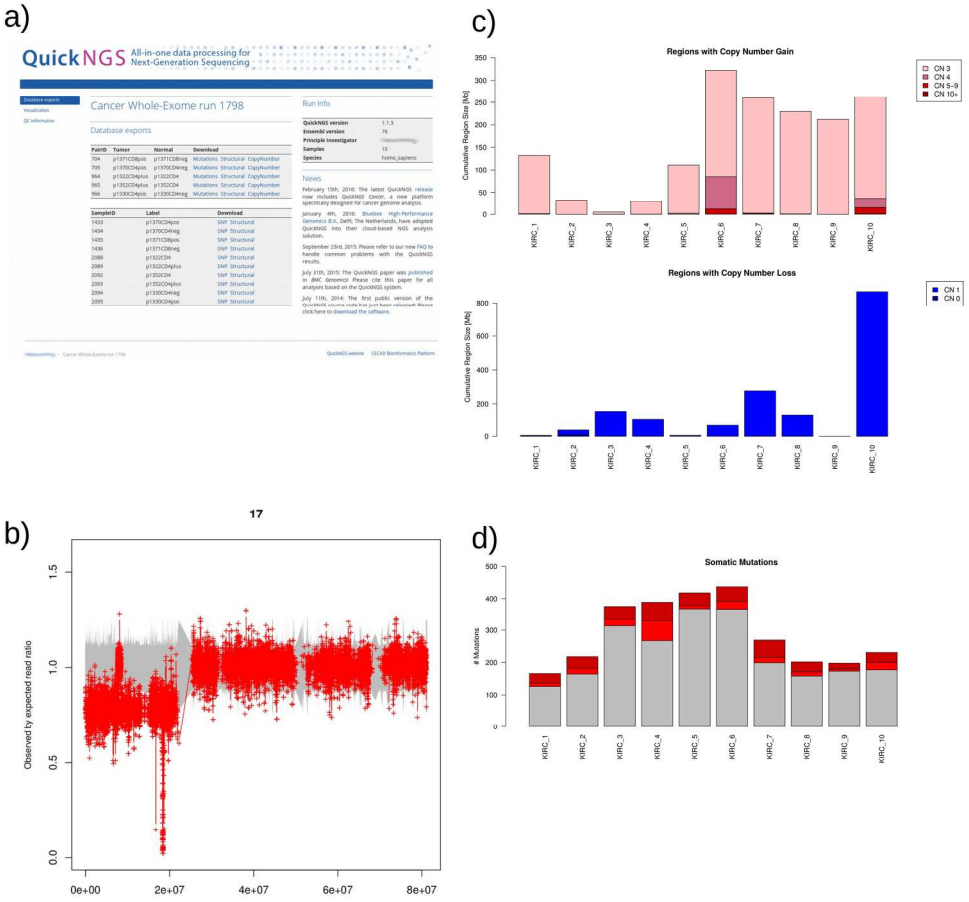


Figure 2: Results for an analysis run of the WXS workflow. (a) Upon completion of the analysis, a password-protected entry point is created for a clinician or experimental scientist. Result tables and graphics are provided as well as general information on the analysis run (PI name blurred). (b) Global graphics on somatic copy number aberrations are generated automatically and made available for download. Depicted here is a deletion of chromosome 17p as observed in chronic lymphocytic leukemia (CLL). (c) Genomic complexity (total size of somatic copy number aberrations) is characterized in two barplots for all samples. Here, the plots are shown for the 10 renal clear cell carcinoma samples from our TCGA meta analysis. The size of regions with a particular number of copies of the genomic locus are displayed in a cumulative way and separately for genomic amplifications (red bars) and deletions (blue bars). (d) Counts of somatic mutations are displayed in cumulative barplots for the entire cohort analyzed. The counts are shown separately for non-synonymous (light red) and potentially damaging mutations (dark red) as well as all other mutations (grey) in a cumulative way. All potentially damaging mutations are also non-synonymous mutations. Thus, the variability of the mutation landscape across the cohort can be captured at a glance.
Figure 2
197x185mm (300 x 300 DPI)

Sheet1

Barcode	Entity	Gene	Chromosome	Position
TCGA-AB-2806	Acute myeloid leukemia	KMT2D	12	49053728
TCGA-AB-2812	Acute myeloid leukemia	WT1	11	32396363
TCGA-AB-2803	Acute myeloid leukemia	WT1	11	32392014
TCGA-AB-2810	Acute myeloid leukemia	KIT	4	54678289
TCGA-BT-A20J	Urothelial carcinoma	LRP2	2	169170537
TCGA-BT-A20J	Urothelial carcinoma	RYR2	1	237445394
TCGA-BT-A2LA	Urothelial carcinoma	TP53	17	7674893
TCGA-BT-A2LA	Urothelial carcinoma	LRP2	2	169185798
TCGA-BT-A2LA	Urothelial carcinoma	KDM6A	X	44961350
TCGA-BT-A2LB	Urothelial carcinoma	COL7A1	3	48586068
TCGA-GC-A3BM	Urothelial carcinoma	LRP1B	2	140950476
TCGA-GC-A3BM	Urothelial carcinoma	RYR1	19	38512145
TCGA-K4-A5RI	Urothelial carcinoma	TP53	17	7674220
TCGA-K4-A5RI	Urothelial carcinoma	LRP1	12	57201707
TCGA-UY-A8OB	Urothelial carcinoma	TP53	17	7674903
TCGA-UY-A8OB	Urothelial carcinoma	RYR1	19	38460341
TCGA-UY-A8OB	Urothelial carcinoma	COL7A1	3	48584550
TCGA-BT-A20Q	Urothelial carcinoma	RYR1	19	38525240
TCGA-2F-A9KQ	Urothelial carcinoma	FGFR3	4	1804372
TCGA-CS-5394	Lower grade glioma	NF1	17	31320344
TCGA-GI-A2C9	Breast invasive carcinoma	PCDHA1	5	140870667
TCGA-BH-A1FC	Breast invasive carcinoma	PCDHGA1	5	141355743
TCGA-3C-AALI	Breast invasive carcinoma	TP53	17	7675064
TCGA-BH-A0B3	Breast invasive carcinoma	PCDHGA3	2	141371420
TCGA-GI-A2C9	Breast invasive carcinoma	TTN-AS1	2	178649908
TCGA-3C-AALI	Breast invasive carcinoma	TTN-AS1	2	178672157
TCGA-A3-3317	Renal clear cell carcinoma	snoU13	4	34966246
TCGA-B0-5084	Renal clear cell carcinoma	MUC4	3	195762138
TCGA-A3-A6NL	Renal clear cell carcinoma	MUC4	3	195780991
TCGA-A3-3358	Renal clear cell carcinoma	MUC4	3	195783887
TCGA-A3-A6NL	Renal clear cell carcinoma	TTN	2	178571586
TCGA-A3-3358	Renal clear cell carcinoma	TTN	2	178582407
TCGA-B0-5077	Renal clear cell carcinoma	TTN-AS1	2	178688975
TCGA-B0-4852	Renal clear cell carcinoma	TTN-AS1	2	178757760
TCGA-A3-3308	Renal clear cell carcinoma	TTN-AS1	2	178757430
TCGA-A3-A6NL	Renal clear cell carcinoma	MUC16	19	8902282
TCGA-B0-4852	Renal clear cell carcinoma	MUC16	19	8894710
TCGA-B0-5084	Renal clear cell carcinoma	CROCCP2	1	16623975
TCGA-A3-A6NL	Renal clear cell carcinoma	CROCCP2	1	16619912
TCGA-2V-A95S	Hepatocellular carcinoma	ARID1A	1	26775642
TCGA-BC-A10Q	Hepatocellular carcinoma	ALB	4	73404298
TCGA-2V-A95S	Hepatocellular carcinoma	ALB	4	73415096
TCGA-2Y-A9GT	Hepatocellular carcinoma	ALB	4	73411964
TCGA-FV-A3I0	Hepatocellular carcinoma	LRP1B	2	140269225
TCGA-DD-A1EI	Hepatocellular carcinoma	LRP1B	2	140269356
TCGA-BC-A10Q	Hepatocellular carcinoma	LRP1B	2	140850350
TCGA-BC-A10U	Hepatocellular carcinoma	RYR2	1	237601816
TCGA-2V-A95S	Hepatocellular carcinoma	RYR2	1	237590862
TCGA-2Y-A9GT	Hepatocellular carcinoma	RYR2	1	237623936
TCGA-BC-A10W	Hepatocellular carcinoma	LRP1B	2	140444546

Sheet1

TCGA-56-7222	Lung squamous cell carcinoma	TP53	17	7675101
TCGA-43-7657	Lung squamous cell carcinoma	TP53	17	7676273
TCGA-56-7580	Lung squamous cell carcinoma	TP53	17	7676185
TCGA-77-8008	Lung squamous cell carcinoma	TP53	17	7673610
TCGA-22-5481	Lung squamous cell carcinoma	TP53	17	7674262
TCGA-77-7338	Lung squamous cell carcinoma	TP53	17	7675136
TCGA-21-5783	Lung squamous cell carcinoma	TP53	17	7676032
TCGA-18-3412	Lung squamous cell carcinoma	TP53	17	7674250
TCGA-09-0367	Ovarian serous cystadenocarcinoma	TP53	17	7675232
TCGA-04-1655	Ovarian serous cystadenocarcinoma	TP53	17	7674252
TCGA-09-1670	Ovarian serous cystadenocarcinoma	TP53	17	7673806
TCGA-09-1673	Ovarian serous cystadenocarcinoma	TP53	17	7675088
TCGA-09-0367	Ovarian serous cystadenocarcinoma	RYR2	1	237792380
TCGA-04-1331	Ovarian serous cystadenocarcinoma	RYR2	1	237792398
TCGA-04-1655	Ovarian serous cystadenocarcinoma	RYR2	1	237237880
TCGA-09-0367	Ovarian serous cystadenocarcinoma	LRP2	2	169212098
TCGA-09-1670	Ovarian serous cystadenocarcinoma	NF1	17	31265251
TCGA-04-1542	Ovarian serous cystadenocarcinoma	RYR1	19	38455542
TCGA-04-1655	Ovarian serous cystadenocarcinoma	RYR1	19	38517570
TCGA-2J-AAB4	Pancreatic adenocarcinoma	SMAD4	18	51058133
TCGA-2J-AAB8	Pancreatic adenocarcinoma	CDKN2A	9	21974777
TCGA-2J-AAB4	Pancreatic adenocarcinoma	KMT2C	7	152235689
TCGA-2J-AAB1	Pancreatic adenocarcinoma	TTN	2	178802347
TCGA-2J-AAB6	Pancreatic adenocarcinoma	RYR2	1	237819279
TCGA-EJ-7782	Prostate adenocarcinoma	TP53	17	7675139
TCGA-EJ-7782	Prostate adenocarcinoma	KMT2C	7	152235676
TCGA-CH-5767	Prostate adenocarcinoma	TTN	2	178539904
TCGA-G9-6499	Prostate adenocarcinoma	KMT2C	7	152273892
TCGA-EJ-7330	Prostate adenocarcinoma	LRP1B	2	140868061
TCGA-BF-A1Q0	Cutaneous melanoma	BRAF	7	140808925
TCGA-BF-A5ER	Cutaneous melanoma	BRAF	7	140753336
TCGA-BF-A1PZ	Cutaneous melanoma	LRP1B	2	140683450
TCGA-BF-A3DL	Cutaneous melanoma	LRP1B	2	140982112
TCGA-BF-A5EO	Cutaneous melanoma	LRP1B	2	140989672
TCGA-BF-A1PX	Cutaneous melanoma	LRP1B	2	140509958
TCGA-D3-A1Q3	Cutaneous melanoma	LRP1B	2	140492584
TCGA-D3-A2J6	Cutaneous melanoma	LRP1B	2	140776176

Sheet1

RefAllele	AltAllele	RefReadsT	AltReadsT	RefReadsN	AltReadsN
G	A	44	7	51	1
C	CGACCG	127	55	211	7
C	T	405	12	420	0
G	A	2	9	5	0
C	G	135	23	121	1
C	T	41	19	88	2
C	T	1	26	32	0
G	A	70	56	143	2
C	T	0	22	19	3
C	T	150	30	170	0
G	A	6	14	18	1
G	A	45	33	82	3
C	T	25	47	83	0
C	A	102	40	116	0
TTC	T	15	18	24	0
C	A	31	7	28	0
G	A	51	30	66	0
G	C	12	11	28	0
A	G	84	67	193	1
A	T	6	5	23	1
T	C	66	10	183	2
C	T	42	15	27	0
G	T	104	193	24	0
C	A	38	20	40	5
A	T	14	11	18	1
C	T	65	18	57	0
G	T	5	6	10	0
C	A	6	9	10	2
G	T	51	24	45	7
C	T	44	7	56	1
C	A	74	6	148	1
C	A	209	10	200	0
A	C	12	9	15	0
C	A	154	55	164	0
T	G	40	9	69	0
T	C	25	8	19	0
C	A	31	13	51	0
C	T	36	11	92	6
T	G	47	15	72	3
GC	G	86	24	68	0
C	T	113	75	377	3
G	T	98	36	128	0
A	G	62	17	85	0
G	T	50	8	43	0
G	A	55	15	71	0
C	A	53	11	92	0
C	A	11	9	13	0
G	A	159	23	143	0
A	T	34	16	64	0
G	A	86	55	266	0

Sheet1

C	A	12	61	93	0
CTGTAGATG	C	162	61	242	1
C	A	162	70	268	0
T	C	56	21	80	0
T	C	11	37	35	0
G	A	53	25	221	0
AGCCCAGAC	A	42	12	73	0
C	G	9	34	66	2
G	A	33	190	130	0
C	A	9	82	108	1
C	T	14	115	94	0
C	T	8	139	83	1
T	C	31	32	26	5
C	T	12	19	31	5
C	T	3	5	9	0
T	A	374	110	564	0
G	T	57	191	311	0
G	A	122	11	162	2
C	T	15	6	30	0
GC	G	166	32	50	0
GGCCA	G	152	22	33	0
A	G	41	5	10	0
G	A	71	20	24	0
G	A	114	28	48	0
C	T	34	6	64	1
CA	C	18	5	10	0
C	T	63	26	157	0
C	A	10	12	14	2
TA	T	15	6	29	1
G	T	234	15	71	0
A	T	51	53	103	0
G	T	6	5	14	0
C	T	35	26	68	0
G	A	60	18	66	0
C	A	126	13	105	0
G	T	98	8	30	0
C	A	66	6	57	0

Sheet1

Callers
strelka,mutect
varscan2,strelka
strelka,mutect
varscan2,somaticsniper.strelka
strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka
strelka,mutect
varscan,somaticsniper
varscan,somaticsniper,strelka
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan2,strelka
varscan2,mutect
Varscan2,somaticsniper,strelka,mutect
Varscan2,somaticsniper,strelka,mutect
Varscan2,somaticsniper,strelka,mutect
Varscan2,somaticsniper,strelka,mutect
Varscan2,somaticsniper,strelka,mutect
Varscan2,strelka
Varscan2,somaticsniper
Varscan2,somaticsniper
Varscan2,strelka
somaticsniper,strelka
strelka,mutect
Varscan2,somaticsniper,strelka
varscan2,somaticsniper,strelka,mutect
varscan2,strelka,mutect
Varscan2,somaticsniper
varscan2,somaticsniper,strelka,mutect
Varscan2,somaticsniper
Varscan2,somaticsniper
varscan2,strelka
varscan2,somaticsniper,strelka,mutect
varscan2,somaticsniper,strelka,mutect
varscan2,somaticsniper,strelka,mutect
strelka,mutect
varscan2,somaticsniper,strelka,mutect
strelka,mutect
varscan2,somaticsniper,strelka,mutect
strelka,mutect
varscan2,somaticsniper,strelka,mutect
varscan2,somaticsniper,strelka,mutect

Peer Review

Sheet1

varscan,somaticsniper,strelka,mutect
varscan,strelka
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,strelka
varscan,somaticsniper,strelka
varscan2,somaticsniper,strelka,mutect
varscan2,somaticsniper,strelka
varscan2,somaticsniper,strelka,mutect
varscan2,somaticsniper,strelka,mutect
varscan2,strelka
varscan2,somaticsniper,strelka
varscan2,somaticsniper,strelka
varscan2,somaticsniper,strelka,mutect
varscan2,somaticsniper,strelka,mutect
strelka,mutect
varscan2,strelka,mutect
varscan2,strelka
varscan2,strelka
varscan2,strelka
varscan2,somaticsniper,strelka,mutect
varscan2,strelka,mutect
varscan2,strelka
varscan2,strelka
varscan2,somaticsniper,strelka,mutect
varscan2,somaticsniper
varscan2,strelka
strelka,mutect
varscan,somaticsniper,strelka,mutect
Varscan,strelka,mutect
varscan,somaticsniper,strelka,mutect
varscan,somaticsniper,strelka,mutect
strelka,mutect
strelka,mutect
strelka,mutect

Peer Review

Sheet1

ListID	TumorBarcode	MatchedNorm
LAML_1	TCGA-AB-2802-03B-01W-0728-08	TCGA-AB-280
LAML_2	TCGA-AB-2803-03B-01W-0728-08	TCGA-AB-280
LAML_3	TCGA-AB-2804-03B-01W-0728-08	TCGA-AB-280
LAML_4	TCGA-AB-2806-03B-01W-0728-08	TCGA-AB-280
LAML_5	TCGA-AB-2813-03B-01W-0728-08	TCGA-AB-281
LAML_6	TCGA-AB-2808-03B-01W-0728-08	TCGA-AB-
LAML_7	TCGA-AB-2809-03D-01W-0755-09	TCGA-AB-280
LAML_8	TCGA-AB-2810-03B-01W-0728-08	TCGA-AB-281
LAML_9	TCGA-AB-2811-03B-01W-0728-08	TCGA-AB-281
LAML_10	TCGA-AB-2812-03B-01W-0728-08	TCGA-AB-
	Mean	
	Standard deviation	
BLCA_1	TCGA-2F-A9KQ-01A-11D-A38G-08	TCGA-2F-A9K
BLCA_2	TCGA-BT-A20J-01A-11D-A14W-08	TCGA-BT-A20
BLCA_3	TCGA-BT-A20Q-01A-11D-A14W-08	TCGA-BT-A20
BLCA_4	TCGA-BT-A20T-01A-11D-A14W-08	TCGA-BT-A20
BLCA_5	TCGA-BT-A20V-01A-11D-A14W-08	TCGA-BT-A20
BLCA_6	TCGA-BT-A2LA-01A-11D-A18F-08	TCGA-BT-A2L
BLCA_7	TCGA-BT-A2LB-01A-11D-A18F-08	TCGA-BT-A2L
BLCA_8	TCGA-GC-A3BM-01A-11D-A22Z-08	TCGA-GC-A3
BLCA_9	TCGA-K4-A5RI-01A-11D-A289-08	TCGA-K4-A5F
BLCA_10	TCGA-UY-A8OB-01A-12D-A42E-08	TCGA-UY-A8
	Mean	
	Standard deviation	
LGG_1	TCGA-CS-4942-01A-01D-1468-08	TCGA-CS-494
LGG_2	TCGA-CS-4943-01A-01D-1468-08	TCGA-CS-494
LGG_3	TCGA-CS-4944-01A-01D-1468-08	TCGA-CS-494
LGG_4	TCGA-CS-5393-01A-01D-1468-08	TCGA-CS-539
LGG_5	TCGA-CS-5394-01A-01D-1468-08	TCGA-CS-539
LGG_6	TCGA-CS-5395-01A-01D-1468-08	TCGA-CS-539
LGG_7	TCGA-CS-6188-01A-11D-1893-08	TCGA-CS-618
LGG_8	TCGA-DB-5277-01A-01D-1468-08	TCGA-DB-527
LGG_9	TCGA-DB-5278-01A-01D-1468-08	TCGA-DB-527
LGG_10	TCGA-DB-5280-01A-01D-1468-08	TCGA-DB-528
	Mean	
	Standard deviation	
BRCA_1	TCGA-E2-A15K-01A-11D-A12Q-09	TCGA-E2-A15
BRCA_2	TCGA-BH-A0DT-01A-21D-A12B-09	TCGA-BH-A0I
BRCA_3	TCGA-BH-A1FC-01A-11D-A13L-09	TCGA-BH-A1I
BRCA_4	TCGA-BH-A0BW-01A-11D-A10Y-09	TCGA-BH-A0I
BRCA_5	TCGA-BH-A18R-01A-11D-A12B-09	TCGA-BH-A18
BRCA_6	TCGA-BH-A0E0-01A-11W-A071-09	TCGA-BH-A0I
BRCA_7	TCGA-GI-A2C9-01A-11D-A21Q-09	TCGA-GI-A2C
BRCA_8	TCGA-BH-A0B3-01A-11W-A071-09	TCGA-BH-A0I
BRCA_9	TCGA-3C-AAAU-01A-11D-A41F-09	TCGA-3C-
BRCA_10	TCGA-3C-AALI-01A-11D-A41F-09	TCGA-3C-
	Mean	
	Standard deviation	
COAD_1	TCGA-AZ-6600-01A-11D-1771-10	TCGA-AZ-660
COAD_2	TCGA-AA-3663-01A-01D-1719-10	TCGA-AA-366

Sheet1

COAD_3	TCGA-AA-3489-01A-21D-1835-10	TCGA-AA-348
COAD_4	TCGA-AA-3655-01A-02D-1719-10	TCGA-AA-365
COAD_5	TCGA-AA-3713-01A-21D-1719-10	TCGA-AA-371
COAD_6	TCGA-AA-3511-01A-21D-1835-10	TCGA-AA-351
COAD_7	TCGA-AZ-6598-01A-11D-1771-10	TCGA-AZ-659
COAD_8	TCGA-4T-AA8H-01A-11D-A40P-10	TCGA-4T-AA8
COAD_9	TCGA-A6-5659-01A-01D-A270-10	TCGA-A6-565
COAD_10	TCGA-F4-6704-01A-11D-1835-10	TCGA-F4-670
	Mean	
	Standard deviation	
KIRC_1	TCGA-A3-3308-01A-01D-0966-08	TCGA-A3-330
KIRC_2	TCGA-A3-3317-01A-01D-0966-08	TCGA-A3-331
KIRC_3	TCGA-A3-3358-01A-01D-1534-10	TCGA-A3-335
KIRC_4	TCGA-A3-A6NL-01A-11D-A33K-10	TCGA-A3-A6N
KIRC_5	TCGA-B0-4818-01A-01D-1501-10	TCGA-B0-481
KIRC_6	TCGA-B0-4852-01A-01D-1501-10	TCGA-B0-485
KIRC_7	TCGA-B0-5075-01A-01D-1462-08	TCGA-B0-507
KIRC_8	TCGA-B0-5077-01A-01D-1462-08	TCGA-B0-507
KIRC_9	TCGA-B0-5080-01A-01D-1501-10	TCGA-B0-
KIRC_10	TCGA-B0-5084-01A-01D-1462-08	TCGA-B0-508
	Mean	
	Standard deviation	
LIHC_1	TCGA-DD-A3A3-01A-11D-A22F-10	TCGA-DD-A3
LIHC_2	TCGA-FV-A3I0-01A-11D-A22F-10	TCGA-FV-A3I
LIHC_3	TCGA-BC-A10Z-01A-11D-A12Z-10	TCGA-BC-A1(
LIHC_4	TCGA-DD-A39X-01A-11D-A20W-10	TCGA-DD-A3
LIHC_5	TCGA-BC-A10W-01A-11D-A12Z-10	TCGA-BC-A1(
LIHC_6	TCGA-DD-A1EI-01A-11D-A12Z-10	TCGA-DD-A1
LIHC_7	TCGA-BC-A10Q-01A-11D-A12Z-10	TCGA-BC-A1(
LIHC_8	TCGA-BC-A10U-01A-11D-A12Z-10	TCGA-BC-A1(
LIHC_9	TCGA-2V-A95S-01A-11D-A36X-10	TCGA-2V-A95
LIHC_10	TCGA-2Y-A9GT-01A-11D-A382-10	TCGA-2Y-A9(
	Mean	
	Standard deviation	
LUSC_1	TCGA-56-7222-01A-11D-2042-08	TCGA-56-722
LUSC_2	TCGA-22-5489-01A-01D-1632-08	TCGA-22-548
LUSC_3	TCGA-43-7657-01A-31D-2122-08	TCGA-43-765
LUSC_4	TCGA-56-7580-01A-11D-2042-08	TCGA-56-758
LUSC_5	TCGA-43-6143-01A-11D-1817-08	TCGA-43-614
LUSC_6	TCGA-77-8008-01A-21D-2184-08	TCGA-77-800
LUSC_7	TCGA-22-5481-01A-31D-1945-08	TCGA-22-548
LUSC_8	TCGA-77-7338-01A-11D-2042-08	TCGA-77-733
LUSC_9	TCGA-56-7731-01A-11D-2122-08	TCGA-56-773
LUSC_10	TCGA-21-5783-01A-41D-2184-08	TCGA-21-
LUSC_11	TCGA-21-5784-01A-01D-1632-08	TCGA-21-578
LUSC_12	TCGA-18-3406-01A-01D-0983-08	TCGA-18-340
LUSC_13	TCGA-18-3407-01A-01D-0983-08	TCGA-18-340
LUSC_14	TCGA-18-3408-01A-01D-0983-08	TCGA-18-340
LUSC_15	TCGA-18-3409-01A-01D-0983-08	TCGA-18-340
LUSC_16	TCGA-18-3410-01A-01D-0983-08	TCGA-18-341
LUSC_17	TCGA-18-3411-01A-01D-0983-08	TCGA-18-341
LUSC_18	TCGA-18-3412-01A-01D-0983-08	TCGA-18-341

Sheet1

LUSC_19	TCGA-18-3414-01A-01D-0983-08	TCGA-18-341
LUSC_20	TCGA-18-3415-01A-01D-0983-08	TCGA-18-341
	Mean	
	Standard deviation	
OV_1	TCGA-04-1331-01A-01W-0486-08	TCGA-04-
OV_2	TCGA-04-1332-01A-01W-0486-08	TCGA-04-133
OV_3	TCGA-04-1341-01A-01W-0486-08	TCGA-04-
OV_4	TCGA-04-1343-01A-01W-0486-08	TCGA-04-
OV_5	TCGA-04-1367-01A-01W-0492-08	TCGA-04-
OV_6	TCGA-04-1542-01A-01W-0553-09	TCGA-04-
OV_7	TCGA-04-1655-01A-01W-0633-09	TCGA-04-
OV_8	TCGA-09-0367-01A-01W-0371-08	TCGA-09-
OV_9	TCGA-09-1670-01A-01W-0633-09	TCGA-09-
OV_10	TCGA-09-1673-01A-01W-0633-09	TCGA-09-
	Mean	
	Standard deviation	
PAAD_1	TCGA-H6-A45N-01A-11D-A26I-08	TCGA-H6-A4
PAAD_2	TCGA-H6-8124-01A-11D-2396-08	TCGA-H6-812
PAAD_3	TCGA-YB-A89D-01A-12D-A36O-08	TCGA-YB-A8
PAAD_4	TCGA-HV-A5A3-01A-11D-A26I-08	TCGA-HV-A5
PAAD_5	TCGA-2J-AAB1-01A-11D-A40W-08	TCGA-2J-AAE
PAAD_6	TCGA-2J-AAB4-01A-12D-A40W-08	TCGA-2J-AAE
PAAD_7	TCGA-2J-AAB6-01A-11D-A40W-08	TCGA-2J-AAE
PAAD_8	TCGA-2J-AAB8-01A-12D-A40W-08	TCGA-2J-AAE
PAAD_9	TCGA-2J-AABE-01A-12D-A40W-08	TCGA-2J-AAE
PAAD_10	TCGA-2J-AABF-01A-31D-A40W-08	TCGA-2J-AAE
	Mean	
	Standard deviation	
PRAD_1	TCGA-2A-A8VL-01A-21D-A377-08	TCGA-2A-A8\
PRAD_2	TCGA-2A-A8VO-01A-11D-A377-08	TCGA-2A-A8\
PRAD_3	TCGA-EJ-7782-01A-11D-2114-08	TCGA-EJ-778
PRAD_4	TCGA-EJ-7785-01A-11D-2114-08	TCGA-EJ-778
PRAD_5	TCGA-EJ-7330-01A-11D-2114-08	TCGA-EJ-733
PRAD_6	TCGA-CH-5767-01A-11D-1786-08	TCGA-CH-57
PRAD_7	TCGA-HC-7740-01A-11D-2114-08	TCGA-HC-77
PRAD_8	TCGA-EJ-7123-01A-11D-1961-08	TCGA-EJ-712
PRAD_9	TCGA-G9-6499-01A-12D-1961-08	TCGA-G9-649
PRAD_10	TCGA-EJ-7331-01A-11D-2114-08	TCGA-EJ-733
	Mean	
	Standard deviation	
SKCM_1	TCGA-BF-A1PU-01A-11D-A19A-08	TCGA-BF-A1
SKCM_2	TCGA-BF-A1PV-01A-11D-A19A-08	TCGA-BF-A1
SKCM_3	TCGA-BF-A1PX-01A-12D-A19A-08	TCGA-BF-A1
SKCM_4	TCGA-BF-A1PZ-01A-11D-A19A-08	TCGA-BF-A1
SKCM_5	TCGA-BF-A1Q0-01A-21D-A19A-08	TCGA-BF-A1
SKCM_6	TCGA-BF-A3DL-01A-11D-A20D-08	TCGA-BF-A3
SKCM_7	TCGA-BF-A3DM-01A-11D-A20D-08	TCGA-BF-A3
SKCM_8	TCGA-BF-A5EO-01A-12D-A27K-08	TCGA-BF-A5
SKCM_9	TCGA-BF-A5EQ-01A-21D-A27K-08	TCGA-BF-A5
SKCM_10	TCGA-BF-A5ER-01A-12D-A27K-08	TCGA-BF-A5
SKCM_11	TCGA-D3-A1Q3-06A-11D-A196-08	TCGA-D3-A1
SKCM_12	TCGA-D3-A1Q5-06A-11D-A196-08	TCGA-D3-A1

For Review

Sheet1

SKCM_13	TCGA-D3-A1Q6-06A-11D-A196-08	TCGA-D3-A1(
SKCM_14	TCGA-D3-A1Q7-06A-11D-A19A-08	TCGA-D3-A1(
SKCM_15	TCGA-D3-A1Q8-06A-11D-A19A-08	TCGA-D3-A1(
SKCM_16	TCGA-D3-A1Q9-06A-11D-A19A-08	TCGA-D3-A1(
SKCM_17	TCGA-D3-A2J6-06A-11D-A19A-08	TCGA-D3-A2J
SKCM_18	TCGA-D3-A2J7-06A-11D-A196-08	TCGA-D3-A2J
SKCM_19	TCGA-D3-A2J8-06A-11D-A196-08	TCGA-D3-A2J
SKCM_20	TCGA-D3-A2J9-06A-11D-A196-08	TCGA-D3-A2J
	Mean	
	Standard deviation	

For Peer Review

Sheet1

Entity	Tumor purity [%]			Tumor ploidy
	QuickNGS	Histology	Difference	QuickNGS
Acute myeloid leukemia	23	100	-77.0	2.2
Acute myeloid leukemia	22	100	-78.0	2.1
Acute myeloid leukemia	34	100	-66.0	1.9
Acute myeloid leukemia	66	100	-34.0	1.9
Acute myeloid leukemia	86	100	-14.0	2.0
Acute myeloid leukemia	73	100	-27.0	2.1
Acute myeloid leukemia	68	100	-32.0	2.0
Acute myeloid leukemia	37	100	-63.0	2.2
Acute myeloid leukemia	23	100	-77.0	2.2
Acute myeloid leukemia	23	100	-77.0	2.2
Acute myeloid leukemia	**45.5**	**100.0**	**-54.5**	**2.1**
Acute myeloid leukemia	**24.9**	**0.0**	**24.9**	**0.1**
Urothelial carcinoma	68	75	-7.0	2.1
Urothelial carcinoma	47	75	-28.0	2.0
Urothelial carcinoma	34	80	-46.0	2.1
Urothelial carcinoma	90	75	15.0	1.8
Urothelial carcinoma	76	60	16.0	2.2
Urothelial carcinoma	65	90	-25.0	2.2
Urothelial carcinoma	63	80	-17.0	2.2
Urothelial carcinoma	51	97	-46.0	2.0
Urothelial carcinoma	73	75	-2.0	2.5
Urothelial carcinoma	59	75	-16.0	2.7
Urothelial carcinoma	**62.6**	**78.2**	**-15.6**	**2.2**
Urothelial carcinoma	**15.9**	**9.9**	**21.8**	**0.3**
Lower grade glioma	74	60	14.0	2.0
Lower grade glioma	94	70	24.0	1.9
Lower grade glioma	77	60	17.0	2.1
Lower grade glioma	75	75	0.0	2.2
Lower grade glioma	76	80	-4.0	2.0
Lower grade glioma	24	75	-51.0	2.3
Lower grade glioma	70	70	0.0	2.2
Lower grade glioma	84	60	24.0	1.9
Lower grade glioma	94	60	34.0	2.1
Lower grade glioma	78	65	13.0	1.8
Lower grade glioma	**74.6**	**67.5**	**7.1**	**2.0**
Lower grade glioma	**19.6**	**7.5**	**23.7**	**0.2**
Breast invasive carcinoma	30	90	-60	1.8
Breast invasive carcinoma	99	70	29	1.8
Breast invasive carcinoma	60	80	-20	2.2
Breast invasive carcinoma	52	70	-18	1.8
Breast invasive carcinoma	73	75	-2	2.1
Breast invasive carcinoma	66	70	-4	2.2
Breast invasive carcinoma	77	90	-13	2.1
Breast invasive carcinoma	58	70	-12	2.0
Breast invasive carcinoma	78	80	-2	2.2
Breast invasive carcinoma	94	60	34	2.4
Breast invasive carcinoma	**68.7**	**75.5**	**-6.8**	**2.1**
Breast invasive carcinoma	**20.3**	**9.6**	**26.2**	**0.2**
Colon adenocarcinoma	51	80	-29	2.4
Colon adenocarcinoma	59	70	-11	2.3

Sheet1

Colon adenocarcinoma	42	70	-28	1.8
Colon adenocarcinoma	76	70	6	2.1
Colon adenocarcinoma	49	70	-21	2.1
Colon adenocarcinoma	64	70	-6	2.1
Colon adenocarcinoma	35	80	-45	2.0
Colon adenocarcinoma	68	70	-2	1.6
Colon adenocarcinoma	84	85	-1	2.0
Colon adenocarcinoma	64	80	-16	2.2
Colon adenocarcinoma	**59.2**	**74.5**	**-15.3**	**2.1**
Colon adenocarcinoma	**15.2**	**6.0**	**15.6**	**0.2**
Renal clear cell carcinoma	28	80	-52	2.4
Renal clear cell carcinoma	45	80	-35	2.1
Renal clear cell carcinoma	42	60	-18	1.7
Renal clear cell carcinoma	52	85	-33	2.4
Renal clear cell carcinoma	88	90	-2	2.1
Renal clear cell carcinoma	42	85	-43	2.1
Renal clear cell carcinoma	46	75	-29	2.1
Renal clear cell carcinoma	45	80	-35	2.3
Renal clear cell carcinoma	58	60	-2	2.9
Renal clear cell carcinoma	79	85	-6	1.7
Renal clear cell carcinoma	**52.5**	**78.0**	**-25.5**	**2.2**
Renal clear cell carcinoma	**18.2**	**10.3**	**17.6**	**0.4**
Hepatocellular carcinoma	85	95	-10	2.3
Hepatocellular carcinoma	89	70	19	2.8
Hepatocellular carcinoma	95	100	-5	1.9
Hepatocellular carcinoma	55	90	-35	1.8
Hepatocellular carcinoma	92	95	-3	1.9
Hepatocellular carcinoma	33	60	-27	2.4
Hepatocellular carcinoma	91	90	1	2.1
Hepatocellular carcinoma	83	95	-12	1.9
Hepatocellular carcinoma	68	80	-12	2.0
Hepatocellular carcinoma	100	100	0	2.0
Hepatocellular carcinoma	**79.1**	**87.5**	**-8.4**	**2.1**
Hepatocellular carcinoma	**21.0**	**13.4**	**15.0**	**0.3**
Lung squamous cell carcinoma	71	60	11	2.3
Lung squamous cell carcinoma	42	80	-38	2.2
Lung squamous cell carcinoma	60	95	-35	2.3
Lung squamous cell carcinoma	55	60	-5	2.1
Lung squamous cell carcinoma	62	85	-23	2.2
Lung squamous cell carcinoma	26	60	-34	2.1
Lung squamous cell carcinoma	66	85	-19	2.0
Lung squamous cell carcinoma	43	60	-17	2.3
Lung squamous cell carcinoma	25	75	-50	2.1
Lung squamous cell carcinoma	50	70	-20	2.2
Lung squamous cell carcinoma	26	70	-44	2.1
Lung squamous cell carcinoma	65	80	-15	2.4
Lung squamous cell carcinoma	42	70	-28	2.2
Lung squamous cell carcinoma	51	80	-29	2.2
Lung squamous cell carcinoma	49	90	-41	1.9
Lung squamous cell carcinoma	73	80	-7	1.9
Lung squamous cell carcinoma	69	85	-16	2.1
Lung squamous cell carcinoma	65	80	-15	1.9

Sheet1

Lung squamous cell carcinoma	49	80	-31	1.8
Lung squamous cell carcinoma	47	90	-43	2.2
Lung squamous cell carcinoma	**51.8**	**76.8**	**-25.0**	**2.1**
Lung squamous cell carcinoma	**14.9**	**10.8**	**15.2**	**0.2**
Ovarian serous cystadenocarcinoma	86	88	-2	2.1
Ovarian serous cystadenocarcinoma	100	89	11	2.0
Ovarian serous cystadenocarcinoma	81	80	1	2.1
Ovarian serous cystadenocarcinoma	71	88	-17	2.0
Ovarian serous cystadenocarcinoma	92	95	-3	2.1
Ovarian serous cystadenocarcinoma	57	80	-23	2.3
Ovarian serous cystadenocarcinoma	95	95	0	2.6
Ovarian serous cystadenocarcinoma	78	95	-17	2.1
Ovarian serous cystadenocarcinoma	76	99	-23	2.1
Ovarian serous cystadenocarcinoma	91	99	-8	2.0
Ovarian serous cystadenocarcinoma	**82.7**	**90.8**	**-8.1**	**2.1**
Ovarian serous cystadenocarcinoma	**12.8**	**7.0**	**11.4**	**0.2**
Pancreatic adenocarcinoma	56	50	6	1.5
Pancreatic adenocarcinoma	79	95	-16	1.9
Pancreatic adenocarcinoma	27	70	-43	2.3
Pancreatic adenocarcinoma	46	60	-14	2.6
Pancreatic adenocarcinoma	27	40	-13	1.8
Pancreatic adenocarcinoma	31	50	-19	2.2
Pancreatic adenocarcinoma	60	80	-20	2.5
Pancreatic adenocarcinoma	26	50	-24	2.2
Pancreatic adenocarcinoma	31	70	-39	2.3
Pancreatic adenocarcinoma	53	40	13	2.0
Pancreatic adenocarcinoma	**43.6**	**60.5**	**-16.9**	**2.1**
Pancreatic adenocarcinoma	**18.1**	**18.0**	**17.2**	**0.3**
Prostate adenocarcinoma	19	70	-51	1.9
Prostate adenocarcinoma	57	85	-28	2.2
Prostate adenocarcinoma	58	70	-12	1.9
Prostate adenocarcinoma	23	70	-47	2.0
Prostate adenocarcinoma	31	80	-49	2.5
Prostate adenocarcinoma	33	60	-27	1.9
Prostate adenocarcinoma	23	65	-42	1.9
Prostate adenocarcinoma	63	65	-2	2.1
Prostate adenocarcinoma	22	70	-48	2.3
Prostate adenocarcinoma	24	80	-56	2.1
Prostate adenocarcinoma	**35.3**	**71.5**	**-36.2**	**2.1**
Prostate adenocarcinoma	**17.2**	**7.8**	**18.2**	**0.2**
Cutaneous melanoma	63	95	-32	2.2
Cutaneous melanoma	48	95	-47	2.0
Cutaneous melanoma	31	85	-54	2.2
Cutaneous melanoma	67	95	-28	2.3
Cutaneous melanoma	58	70	-12	2.0
Cutaneous melanoma	75	95	-20	2.0
Cutaneous melanoma	23	95	-72	2.1
Cutaneous melanoma	86	70	16	1.9
Cutaneous melanoma	100	75	25	1.9
Cutaneous melanoma	80	96	-16	2.1
Cutaneous melanoma (metastatic)	100	95	5	2.0
Cutaneous melanoma (metastatic)	99	85	14	2.1

Sheet1

Cutaneous melanoma (metastatic)	73	85	-12	2.2
Cutaneous melanoma (metastatic)	21	75	-54	2.4
Cutaneous melanoma (metastatic)	46	85	-39	2.1
Cutaneous melanoma (metastatic)	58	85	-27	2.1
Cutaneous melanoma (metastatic)	61	90	-29	1.9
Cutaneous melanoma (metastatic)	51	85	-34	1.9
Cutaneous melanoma (metastatic)	35	75	-40	2.3
Cutaneous melanoma (metastatic)	37	85	-48	2.1
Cutaneous melanoma	**60.6**	**85.8**	**-25.2**	**2.1**
Cutaneous melanoma	**24.7**	**8.8**	**25.7**	**0.1**

For Peer Review

Sheet1

# Mutations			rgetRegion [M	Mutation Rate	# Mutations in key ge	
QuickNGS	TCGA	Difference		QuickNGS	QuickNGS only	TCGA only
15	7	8	32.9	0.5	0	0
20	12	8	32.9	0.6	1	0
33	6	27	32.9	1.0	0	0
152	14	138	32.9	4.6	1	0
35	14	21	32.9	1.1	0	0
62	8	54	38.8	1.6	0	0
13	3	10	38.8	0.3	0	1
27	11	16	32.9	0.8	1	0
14	N/A	N/A	32.9	0.4	0	0
14	4	10	32.9	0.4	1	0
38.5	**8.8**	**32.4**	**34.1**	**1.1**	**0.4**	**0.1**
40.9	**4.1**	**42.2**	**2.5**	**1.3**	**0.5**	**0.3**
63	61	2	32.9	1.9	1	0
362	413	-51	32.9	11.0	2	0
81	89	-8	32.9	2.5	1	0
202	220	-18	32.9	6.1	0	1
81	91	-10	32.9	2.5	0	0
217	149	68	32.9	6.6	3	0
428	481	-53	32.9	13.0	1	0
92	100	-8	32.9	2.8	2	1
118	114	4	32.9	3.6	2	0
279	296	-17	32.9	8.5	3	0
192.3	**201.4**	**-9.1**	**32.9**	**5.8**	**1.5**	**0.2**
128.9	**148.1**	**33.4**	**0.0**	**3.9**	**1.1**	**0.4**
22	36	-14	32.9	0.7	0	0
27	25	2	32.9	0.8	0	0
21	23	-2	32.9	0.6	0	0
26	30	-4	32.9	0.8	0	0
51	16	35	32.9	1.6	1	0
46	54	-8	32.9	1.4	0	0
33	34	-1	32.9	1.0	0	0
46	47	-1	32.9	1.4	0	0
12	11	1	32.9	0.4	0	0
26	29	-3	32.9	0.8	0	0
31.0	**30.5**	**0.5**	**32.9**	**0.9**	**0.1**	**0.0**
12.7	**13.1**	**13.0**	**0.0**	**0.4**	**0.3**	**0.0**
48	28	20	44.1	1.1	0	0
13	7	6	44.1	0.3	0	0
128	55	73	44.1	2.9	1	0
62	52	10	44.1	1.4	0	1
21	14	7	44.1	0.5	0	0
38	20	18	38.8	1.0	0	0
200	92	108	63.6	3.1	2	0
67	24	43	38.8	1.7	1	0
46	N/A	N/A	63.6	0.7	0	N/A
464	N/A	N/A	63.6	7.3	2	N/A
108.7	**36.5**	**35.6**	**48.9**	**2.0**	**0.6**	**0.1**
136.8	**28.1**	**37.0**	**10.4**	**2.1**	**0.8**	**0.4**
133	317	-184	45.1	2.95	3	N/A
1554	1413	141	45.1	34.46	9	N/A

Sheet1

93	87	6	45.1	2.06	7	N/A
82	170	-88	45.1	1.82	1	N/A
796	926	-130	45.1	17.65	9	N/A
122	118	4	45.1	2.71	9	N/A
1845	1678	167	45.1	40.91	9	N/A
110	N/A	N/A	45.1	2.44	8	N/A
223	212	11	45.1	4.94	9	N/A
80	53	27	45.1	1.77	2	N/A
503.8	**552.7**	**-5.1**	**45.1**	**11.2**	**6.6**	**N/A**
669.2	**624.4**	**115.6**	**0.0**	**14.8**	**3.3**	**N/A**
40	N/A	N/A	32.9	1.2	1	2
54	50	4	32.9	1.6	1	0
63	45	18	44.1	1.4	2	0
122	N/A	N/A	45.1	2.7	4	N/A
50	42	8	44.1	1.1	1	0
70	51	19	44.1	1.6	2	0
70	65	5	32.9	2.1	0	1
44	39	5	32.9	1.3	1	0
25	46	-21	44.1	0.6	0	0
54	40	14	32.9	1.6	2	0
59.2	**47.3**	**6.5**	**38.6**	**1.5**	**1.4**	**0.3**
26.0	**8.4**	**12.6**	**6.0**	**0.6**	**1.2**	**0.7**
46	46	0	45.1	1.02	0	0
80	69	11	45.1	1.77	1	0
155	138	17	45.1	3.44	0	1
56	105	-49	45.1	1.24	0	1
102	129	-27	45.1	2.26	1	0
42	65	-23	45.1	0.93	1	0
35	34	1	45.1	0.78	2	0
99	114	-15	45.1	2.20	1	0
85	N/A	N/A	45.1	1.88	3	N/A
75	N/A	N/A	45.1	1.66	2	N/A
77.5	**87.5**	**-10.6**	**45.1**	**1.7**	**1.1**	**0.3**
36.0	**39.1**	**22.0**	**0.0**	**0.8**	**1.0**	**0.5**
303	N/A	N/A	32.9	9.21	9	N/A
199	209	-10	32.9	6.05	2	0
119	N/A	N/A	38.8	3.07	6	N/A
121	N/A	N/A	32.9	3.68	7	N/A
276	268	8	32.9	8.39	4	1
140	N/A	N/A	32.9	4.26	4	N/A
231	N/A	N/A	32.9	7.02	5	N/A
105	N/A	N/A	32.9	3.19	4	N/A
175	N/A	N/A	32.9	5.32	8	N/A
338	N/A	N/A	32.9	10.27	9	N/A
156	185	-29	32.9	4.74	1	1
170	225	-55	32.9	5.17	4	1
116	126	-10	32.9	3.53	1	0
73	78	-5	32.9	2.22	0	1
2358	2432	-74	32.9	71.67	1	1
206	230	-24	32.9	6.26	1	0
293	318	-25	32.9	8.91	2	1
163	159	4	32.9	4.95	3	0

Sheet1

295	306	-11	32.9	8.97	4	1
152	158	-6	32.9	4.62	1	1
299.5	**391.2**	**-19.8**	**33.2**	**9.1**	**3.8**	**0.7**
490.4	**646.6**	**24.0**	**1.3**	**14.9**	**2.8**	**0.5**
123	54	69	32.9	3.7	1	0
64	23	41	32.9	1.9	0	1
41	N/A	N/A	32.9	1.2	0	N/A
79	43	36	32.9	2.4	0	2
100	57	43	32.9	3.0	2	0
89	52	37	38.8	2.3	1	1
47	41	6	38.8	1.2	3	N/A
48	N/A	N/A	32.9	1.5	3	N/A
48	22	26	**38.8**	1.2	2	N/A
65	32	33	**38.8**	1.7	1	N/A
70.4	**40.5**	**36.4**	**35.3**	**2.0**	**1.3**	**0.8**
27.0	**13.7**	**17.6**	**3.0**	**0.9**	**1.2**	**0.8**
52	40	12	32.9	1.58	1	0
35	40	-5	32.9	1.06	1	0
119	192	-73	32.9	3.62	3	0
37	30	7	32.9	1.12	0	0
82	93	-11	32.9	2.49	1	1
30	40	-10	32.9	0.91	2	1
67	76	-9	32.9	2.04	1	0
52	68	-16	32.9	1.58	1	1
35	44	-9	32.9	1.06	0	0
52	63	-11	32.9	1.58	0	0
56.1	**68.6**	**-12.5**	**32.9**	**1.7**	**1.0**	**0.3**
27.4	**47.7**	**23.0**	**0.0**	**0.8**	**0.9**	**0.5**
19	14	5	32.9	0.58	1	0
26	14	12	32.9	0.79	0	0
285	147	138	32.9	8.66	3	1
22	11	11	32.9	0.67	1	0
35	29	6	32.9	1.06	2	1
35	25	10	32.9	1.06	1	0
15	10	5	32.9	0.46	0	0
42	127	-85	32.9	1.28	0	0
36	37	-1	32.9	1.09	1	1
17	22	-5	32.9	0.52	0	1
53.2	**43.6**	**9.6**	**32.9**	**1.6**	**0.9**	**0.4**
82.0	**50.2**	**53.6**	**0.0**	**2.5**	**1.0**	**0.5**
606	N/A	N/A	32.9	18.4	1	0
249	N/A	N/A	32.9	7.6	1	0
374	N/A	N/A	32.9	11.4	2	1
719	N/A	N/A	32.9	21.9	1	0
2058	N/A	N/A	32.9	62.6	7	0
356	N/A	N/A	32.9	10.8	1	1
382	N/A	N/A	32.9	11.6	1	0
480	N/A	N/A	32.9	14.6	3	0
474	N/A	N/A	32.9	14.4	5	0
152	N/A	N/A	32.9	4.6	4	N/A
297	32	265	32.9	9.0	4	0
300	287	13	32.9	9.1	1	0

Sheet1

807	637	170	32.9	24.5	0	0
70	75	-5	32.9	2.1	0	1
183	191	-8	32.9	5.6	2	0
52	32	20	32.9	1.6	0	0
207	15	192	32.9	6.3	2	0
261	294	-33	32.9	7.9	1	0
726	509	217	32.9	22.1	1	0
90	55	35	32.9	2.7	0	0
442.2	**212.7**	**86.6**	**32.9**	**13.4**	**1.9**	**0.2**
440.0	**218.2**	**111.1**	**0.0**	**13.4**	**1.9**	**0.4**

For Peer Review

Sheet1

enes	Amplified [Mb]		Deleted [Mb]	
Both	QuickNGS [Mb]	Overlap w/	QuickNGS [Mb]	Overlap w/ TCGA [%]
3	99.2	4.0	1.6	84.0
0	9.2	59.1	2.9	25.1
0	227.6	42.3	10.3	5.7
0	98.9	94.8	100.1	53.9
1	90.5	82.8	204.2	65.3
1	116.9	20.4	70.8	12.2
0	70.7	46.4	168.8	17.9
0	23.5	45.3	16.6	44.2
1	4.7	68.9	10.0	10.2
0	15.1	31.4	4.5	22.8
0.6	**75.6**	**49.5**	**59.0**	**34.1**
1.0	**68.3**	**27.8**	**75.2**	**26.4**
0	137.5	76.5	452.1	76.8
2	108.4	89.8	58.2	83.9
1	86.0	86.1	175.4	90.4
2	91.3	71.7	414.1	88.5
0	475.2	93.6	181.6	66.3
0	146.5	92.0	499.3	87.7
4	28.2	67.1	104.7	86.6
2	84.9	96.7	695.7	87.0
0	191.1	92.1	234.4	49.4
0	530.6	72.9	126.4	71.4
1.1	**188.0**	**83.9**	**294.2**	**78.8**
1.4	**172.0**	**10.7**	**208.9**	**13.1**
3	286.3	59.5	95.4	86.1
3	161.3	82.9	308.6	96.6
1	147.8	23.8	168.8	94.7
2	233.4	48.9	58.9	81.8
3	48.7	30.2	217.6	63.3
1	319.8	86.1	292.2	68.6
0	227.9	84.5	157.3	97.7
2	343.0	93.9	944.8	83.3
2	30.8	46.2	151.6	99.0
3	2.9	57.1	115.8	72.7
2.0	**180.2**	**61.3**	**251.1**	**84.4**
1.1	**122.4**	**24.6**	**256.5**	**12.9**
1	282.7	79.0	278.0	98.4
1	107.3	89.6	44.8	96.7
1	525.3	72.8	22.4	99.7
1	122.3	98.7	70.9	1.3
0	252.2	64.5	15.3	38.3
1	232.5	80.8	26.5	79.3
1	327.5	87.8	91.3	95.4
1	306.9	82.4	22.2	68.4
N/A	123.7	74.4	142.6	98.5
N/A	117.9	80.7	187.7	93.9
0.9	**239.8**	**81.1**	**90.2**	**77.0**
0.4	**131.4**	**9.6**	**87.5**	**32.9**
N/A	59.0	99.7	133.0	21.7
N/A	249.7	60.5	1.1	95.1

Sheet1

N/A	49.5	12.3	0.1	0.0
N/A	321.6	58.8	250.1	96.4
N/A	148.6	83.1	74.7	0.6
N/A	352.3	72.6	392.1	84.5
N/A	734.2	93.2	261.3	4.0
N/A	246.6	88.1	82.8	92.1
N/A	429.7	41.0	274.4	79.2
N/A	81.4	63.8	0.0	100.0
N/A	**267.3**	**67.3**	**147.0**	**57.4**
N/A	**209.4**	**26.4**	**139.1**	**44.5**
1	132.4	87.3	5.7	17.4
0	32.9	54.4	38.4	52.4
0	5.0	67.0	147.6	56.2
N/A	28.9	90.3	101.1	96.3
3	155.9	98.9	9.1	31.8
3	322.0	74.9	66.4	98.6
1	260.1	87.8	276.7	63.5
2	229.5	90.9	126.5	89.0
0	212.3	67.9	0.0	100.0
0	263.0	63.9	867.3	89.4
1.1	**181.9**	**79.6**	**165.7**	**70.9**
1.3	**102.6**	**14.9**	**277.0**	**31.1**
0	474.6	86.5	86.7	97.9
0	316.4	79.5	509.3	99.6
1	245.6	97.7	252.4	98.2
0	148.1	78.1	55.2	91.6
1	328.2	83.6	316.7	98.2
1	125.6	99.6	17.9	99.3
1	13.3	99.7	128.7	99.4
1	544.6	92.1	341.8	79.3
N/A	160.9	89.9	44.9	97.3
N/A	121.5	34.7	23.9	84.3
0.6	**247.9**	**84.1**	**177.8**	**94.5**
0.5	**168.0**	**19.1**	**168.1**	**7.2**
N/A	258.1	71.8	676.5	97.1
4	308.0	75.7	55.2	74.7
N/A	357.3	77.2	104.8	74.3
N/A	240.9	60.3	27.0	13.2
4	577.7	92.9	622.6	95.3
N/A	66.5	25.7	27.0	22.6
N/A	642.1	87.1	284.6	62.5
N/A	252.6	91.2	111.9	31.0
N/A	66.4	50.5	64.0	17.4
N/A	386.6	87.8	207.8	93.4
3	83.5	88.8	26.4	66.5
3	217.1	74.3	96.7	0.1
2	89.2	88.7	150.5	78.2
1	178.6	91.3	436.8	85
6	0.5	20.9	123.5	79.4
4	106.6	97.3	623.6	95.3
4	515.8	85.9	321.3	55.6
0	449	92.6	220.8	84

Sheet1

4	607.2	78.6	327	93.9
4	225.8	95.2	37.5	99.1
3.3	**281.5**	**76.7**	**227.3**	**65.9**
1.6	**195.1**	**21.8**	**212.3**	**31.8**
3	474.4	80.8	681.6	90.1
0	197.5	66.5	260.1	85.2
N/A	722.5	90.2	85.8	99.9
0	398.3	75.4	179.0	94.8
2	37.7	55.7	243.6	82.3
0	1056.5	83.2	1154.5	87.4
N/A	716.4	83.7	590.1	75.7
N/A	0.0	100.0	432.4	58.1
N/A	541.0	89.8	603.8	73.1
N/A	482.6	91.7	375.8	91.3
1.0	**462.7**	**81.7**	**460.7**	**83.8**
1.4	**326.7**	**13.0**	**313.2**	**12.2**
1	98.1	15.9	12.2	10.1
2	135.5	34.2	109.8	61.9
1	73.8	8.6	2.0	88.5
3	507.0	81.6	440.8	93.6
2	89.2	14.1	4.1	20.7
1	122.1	41.1	14.8	10.4
3	233.1	62.5	133.4	90.3
3	34.9	81.2	29.3	22.0
2	70.3	95.7	17.1	27.6
5	14.5	57.7	9.0	22.5
2.3	**137.9**	**49.3**	**77.3**	**44.8**
1.3	**143.0**	**31.2**	**135.9**	**34.9**
0	63.2	25.6	63.7	69.0
0	93.4	42.9	60.4	81.2
0	101.1	98.4	342.1	72.1
0	11.7	36.8	86.5	1.7
0	5.2	55.8	67.8	11.0
0	170.7	54.4	145.5	97.6
0	16.5	14.9	13.0	39.2
1	105.3	64.6	91.4	39.6
1	131.5	82.1	108.7	59.1
0	37.4	20.4	126.9	10.3
0.2	**73.6**	**49.6**	**110.6**	**48.1**
0.4	**55.9**	**27.0**	**89.5**	**33.0**
3	96.1	57.1	775.2	90.2
2	243.6	72.6	435.5	73.6
4	10.2	43.4	3.7	16.6
1	213.4	68.4	401.4	67.2
2	427.4	86.4	474.2	66.3
4	246.1	78.0	324.0	55.6
3	106.8	89.0	380.7	58.6
2	172.2	91.4	435.5	70.1
5	516.6	74.5	428.8	62.2
N/A	449.6	85.6	271.7	79.1
0	272.9	80.1	16.6	88.3
3	456.4	48.9	620	88.3

Sheet1

7	111.4	47.6	132.7	41.9
1	195.8	48.1	3.7	58.8
2	110.2	89.1	46.7	66.5
2	173.2	80.7	484.4	91.7
0	202.7	85.1	482.6	78.7
6	79.0	67.4	4.2	26.2
6	16.2	21.6	98.7	32.8
0	59.7	49.9	241.6	48.8
2.8	**190.5**	**67.9**	**290.7**	**62.9**
2.1	**139.0**	**20.6**	**236.2**	**22.6**

For Peer Review

Integrated genetic profiles of T-PLL implicate a TCL1/ATM-centered model of aberrant, but actionable damage responses

A. Schrader[1,2*], G. Crispatzu[1,2*], S. Oberbeck[1,2], N. Weit[1,2], K. Warner[1,2,3], S. Pützer[1,2], N. Pflug[1], P. Mayer[1,2], E. Vasyutina[1,2], A. Riabinska[1,2], F. Beier[4], J. Altmüller[5], M. Lanasa[6], T. Haferlach[7], S. Stilgenbauer[8], G. Hopfinger[9], M. Peifer[10], T.H. Brümmendorf[4], P. Nürnberg[5], K.S.J. Elenitoba-Johnson[11], H.C. Reinhardt[1,2], M. Hallek[1,2], M.-H. Stern[12], S. Newrzela[3], P. Frommolt[13], and M. Herling[1,2‡]

[1]Department I of Internal Medicine, Center for Integrated Oncology (CIO) Köln-Bonn, University of Cologne (UoC), Germany, [2]Excellence Cluster for Cellular Stress Response and Aging-Associated Diseases (CECAD), UoC, Germany, [3]Senckenberg Institute of Pathology, Goethe-University, Frankfurt/M., Germany, [4]Department of Hematology, Oncology, and Stem Cell Transplantation, RWTH Aachen University Medical School, Aachen, Germany, [5]Cologne Center for Genomics, UoC, Germany, Institute of Human Genetics, University of Cologne (UoC), Germany, [6]Duke University Medical Center, Durham, NC, USA, [7]MLL Munich Leukemia Laboratory, Munich, Germany, [8]Department III of Internal Medicine, University Hospital Ulm, Germany, [9]Department of Internal Medicine I, Bone Marrow Transplantation Unit, Medical University of Vienna, Vienna, Austria, [10]Department of Translational Genomics, UoC, Germany, [11]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA, [12]INSERM U830, Institut Curie, PSL Research University, Paris, France, [13]Bioinformatics Core Facility, CECAD, UoC, Germany

* These authors contributed equally to this work

Short title: integrated genomic profiles of T-PLL

Key words: T-PLL genomic landscape, *TCL1A*, *ATM*, DNA damage response

Abstract character count: 150 words

Manuscript character count: 4,987 words

Figures: 7

Tables: 0, **Supplementary data** (online only): 3 supplementary files including (1) 23 Tables, (2) 17 Figures, (3) Methods

‡**Corresponding author:** Marco Herling, MD, Laboratory of Lymphocyte Signaling and Oncoproteome, Center for Integrated Oncology (CIO) Köln-Bonn and Cologne Cluster of Excellence in Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Germany
Email: marco.herling@uk-koeln.de
phone +49 221 478-5969
fax +49 221 478-98339

The authors declare no potential conflicts of interest.

ABSTRACT

T-cell prolymphocytic leukemia (T-PLL) is a rare and poor-prognostic mature T-cell malignancy. To address the vastly incomplete molecular concept of T-PLL, we applied large-scale profiling of alterations in gene expression, allelic copy number (CN), and nucleotide variants in 94 well-characterized patients. Key aspects were validated in various experimental models. The dominant trunk of T-PLL's molecular make-up is a unique and functionally synergistic combination of TCL1-overexpression and damaging *ATM* lesions. We identified novel tumor-specific hot-spots for CN variability, fusion molecules, transcript variants, and progression-associated dynamics. Annotated to axes of the DNA damage response, cytokine signaling, and histone modulation, the lesional spectrum of T-PLL determines a specific phenotype of impaired damage sensing and processing, telomere attrition, and chromosomal complexity alongside an abrogated p53-mediated cell death execution. We present a first model of T-PLL evolution resolved for pivotal (epi)genetic alterations integrated with landmarks of cellular dysfunctions and extract from that novel specific drug sensitivities.

STATEMENT OF SIGNIFICANCE

The low incidence of T-PLL impedes systematic studies of this aggressive and highly chemotherapy-resistant mature T-cell leukemia, which continues to be associated with limited therapeutic options and poor patient outcomes. As the first integrative multi-level analysis of genetic lesions on a large set of clinically well-characterized T-PLL, this report provides a comprehensive disease modeling around the central leukemogenic cooperation of overexpressed TCL1 and hypomorphic ATM, that serve as diagnostic hallmarks and that underlie a unique phenotype of selectively impaired DNA damage responses, which in turn can be reinstated by novel epi-/genome targeting compounds.

INTRODUCTION

T-cell prolymphocytic leukemia (T-PLL) is the most frequent mature T-cell leukemia[1], yet with an incidence of ≈0.6/million in Western countries, it is still an orphan disease. It typically presents in the 6-7th decade of life at stages of exponentially rising lymphocyte counts in peripheral blood (PB) accompanied by hepato-splenomegaly, lymphadenopathy, and bone marrow (BM) involvement[1,2]. Its chemo-refractory behavior is reflected in a poor patient survival (usually <3 years)[1,3,4]. Even following responses to the monoclonal antibody alemtuzumab, eventually all patients relapse[3]. A major reason for the limited therapeutic options that aim at the molecular make-up of T-PLL towards sustained clonal eradication is our rudimentary understanding of key mechanisms that underlie progression and resistance.

Karyotypes of T-PLL are often complex[5] and include recurrent rearrangements at chromosome (chr.)14, resulting in juxtaposition of *TCL1A (T-cell leukemia/lymphoma 1A)* at 14q32.1 to T-cell receptor (TCR) gene enhancers[6]. This prevents physiological post-thymic silencing of *TCL1A*. *TCL1A* is the namesake of a 3-paralogue family[7] further including *TCL1B* and *MTCP1*. The X-chromosomal *MTCP1* is involved in rare T-PLL carrying the t(X;14) translocation. Transgenic (tg) mouse models emulating human T-PLL illustrate the T-cell oncogenic potential of *TCL1A*[8] and *MTCP1*[9]. Currently, the best established function of the 14kDa TCL1A protein is an adapter-like engagement in kinase complexes, formed upon antigen-receptor input[10] resulting in enhanced pro-survival signaling.

Deletions of chr.11q leading to losses of the tumor suppressor *ataxia telangiectasia mutated (ATM)* as well as amplifications at chr.8q represent additional highly prevalent abnormalities in T-PLL[5]. While the sporadic form of T-PLL was associated with somatic *ATM* mutations[11], it can also arise in cancer-predisposed adolescents with *ataxia telangiectasia* (A-T) that carry germline *ATM* inactivations[12]. ATM governs the maintenance of genomic integrity by orchestrating a proper DNA damage response (DDR), including double-strand break (DSB) repair, cell cycle control, and apoptosis regulation[13]. There are non-canonical DDRs in the absence of DNA damage, i.e. triggered by telomere, mitotic, replicative, or oxidative stressors[14]. Metabolic or redox-homeostatic roles are also recognized as novel ATM functions[15].

Although small series of genomic and transcriptomic profiling (summary in **TableS1**)

provided important insights, we still face an overall sketchy molecular landscape and disease concept of T-PLL. Here, we report an integrated genetic and functional study on a large T-PLL patient cohort to delineate the spectrum of alterations and their mechanistic impact in T-cell transformation. For relevant clinical associations, we selected treatment-naive samples from patients that were included in prospective multi-centric phase-II trials or that were documented in a nationwide T-PLL registry, providing thorough clinical (e.g. outcomes after uniform front-line therapy), immuno-phenotypic, and cytogenetic data (**TableS2, Fig.S1, Online Methods**).

This study reveals that virtually all cases of T-PLL harbor a dysregulation of a member of the *TCL1* oncogene family predominantly in conjunction with damaging lesions affecting the *ATM* tumor suppressor. Elevated levels of the TCR-signaling enhancer *TCL1A* as the most discerning change in gene expression to normal T-cells were associated with downregulated negative TCR-signaling modulators, e.g. *CTLA4*, implicating the importance of antigen-receptor input in T-PLL. A marked global complexity of gene copy-numbers most frequently includes losses of *ATM* and gains of a chr.8q region commonly involving *AGO2* and *MYC*. The overall mutational profile indicated a genotoxic signature of nucleotide exchanges. Most prevalent were clonally dominant variants in *ATM* with a previously undisclosed domain clustering. Also frequent were subclonal lesions in *JAK1/STAT* genes and in epigenetic regulators. We further describe novel gene fusions, transcript variants, and hierarchic changes upon tumor progression. Essentially, across all platforms, we define T-PLL by a unique combination of *TCL1* overexpression with damaging *ATM* lesions. The effects imposed by TCL1 synergize with compromised ATM towards leukemic outgrowth, associated with a phenotype of impaired damage repair, eroded telomeres, and karyotype complexity. The functionally hypomorphic ATM appears inefficient in alleviating a high redox burden and in evoking a p53-dependent apoptotic response. Novel bi-functional histone-modifying agents reinstated such cell death execution triggered by simultaneously inflicted genotoxic insults.

Overall, we formulate a first comprehensive model of T-PLL pathogenesis. It is centered around the unique combination of constitutive TCL1 and deficient ATM as the common molecular trunk. The leukemogenic cooperation of these initiating core lesions involves perturbations of adequate stress responses, but also represents a specific interventional vulnerability.

RESULTS

The hallmarks of TCL1A overexpression and dysregulated T-cell activation.

PB-isolated tumor cells from 70 T-PLL exhibited a differential expression (|fold-change (fc)| >1.5; q<0.05; p<0.05) of 2569 gene probes as compared to circulating $CD3^+$ pan T-cells from 10 healthy donors (regular $CD4^+/8^+$ ratio of 1.5-2.5). Ingenuity® pathway analysis (IPA) assigned this set of differentially expressed genes to significantly enriched clusters that were functionally annotated to growth regulation, proliferation, cell cycle, chemotaxis, and immune signaling (i.e. cytokine, antigen receptor) (**Fig.S2a, TableS3**). Gene set enrichment analysis (GSEA) highlighted target genes of the transcription factor (TF) and histone-acetylase (HAT) recruiter *MYC* (encoding c-Myc) and signatures of irradiation response or epigenetic remodeling (**Fig.S2b**). We confirmed the deregulated expression of genes associated with T-PLL in meta-comparisons with small published cohorts at the global[16] (GSEA; **Fig.S2b**) and gene-specific level (e.g. *CDKN1B*[17]; **Fig.S2c**). qRT-PCRs validated the differential expression for all of 21 selected transcripts (**Fig.1a**, **S2c**).

Of all genes, *TCL1A* showed the highest degree of dysregulation (fc=33.9; $p=0.3x10^{-13}$; Student t-test; **Fig.1a**). Importantly, as we previously implicated TCL1A as a pro-leukemogenic amplifier of T-cell signaling input[10], the observed *TCL1A* upregulation was accompanied by deregulations of TCR pathway modulators, suggesting a net enhancement of TCR signaling. It included reduced expression of the negative-costimulatory *cytotoxic T-lymphocyte-associated protein 4* (*CTLA4)* (fc=-6.92; $p=0.2x10^{-13}$) and of the repressive T-T homotypic receptor *SLAMF6* (fc=-3.72; $p=0.8x10^{-11}$), or overexpression of the *tumor necrosis factor TNF* (fc=9.98; $p=0.2x10^{-13}$) known to shape TCR signals via *TNFR2* (**Fig.1a**). Upregulation of immunosuppressive *CD83* (**Fig.1a, TableS3**) also indicates immune evasive properties. The other *TCL1* family members were consistently upregulated as well: *TCL1B* (fc=4.53; $p=0.6x10^{-5}$) and $MTCP1^{p13}$ (fc=2.65; $p=0.2x10^{-3}$; **Fig.S2c**). Suggesting an impact of constitutive $MTCP1^{p13}$ comparable to the one by *TCL1A*, there was a considerable overlap of differentially expressed genes (229 of 412 probes; e.g. *CTLA4* and *SLAMF6*) between TCL1A-positive cases and those 4 carrying an *MTCP1*-activating t(X;14). Further implicating a 'uniform' transcriptome of T-PLL, the gene expression profiles (GEPs) of the 2 exclusively TCL1B-positive cases were similar to those of TCL1A-positive or *MTCP1*-rearranged T-PLL. Overall,

proof of TCL1-gene family expression in 90.4% of cases correlated well with cytogenetic detection of locus rearrangements in 94,4% of cases (details in **Fig.S2d**).

Postulating an initiating role of dysregulated *TCL1* genes, we evaluated changes in GEPs in mice with early-onset T-lineage specific overexpression of TCL1A (**Fig.1b**). Already sub-clinical 'chronic' phase expansions (**Fig.S2e**) from spleens of these *Lck*^{pr}*-hTCL1A*^{tg} mice revealed a differential down-regulation of *CTLA4* and *SLAMF6* and other changes, all in common with those observed in human T-PLL ($p<0.05$; $|fc|>2$; **Fig.1b, TableS4**). This signature of T-cell activation in conjunction with TCL1A-drive was preserved at the 'exponential' murine disease stage with additional deregulation of prominent markers of transformation, e.g. *MYC* (**Fig.S2f, TableS4**). The relevance of aberrant TCL1A in overt human leukemia was stressed by the poor prognostic impact of its high-level expression (**Fig.1c**).

Large-scale somatic copy-number alterations (sCNAs) indicate a marked global complexity and involve *ATM* losses and gains of novel genes at 8q.

Based on average abundance of large-fragment genomic lesions, T-PLL (n=83) is positioned near the "complex" end of the sCNA spectrum of hematopoietic and solid cancers (**Fig.2a, Online Methods**). The most frequent sCNAs (compared to pooled germlines from 13 cases and HapMap controls) were found at chr.11 (37%/52%), chr.8 (29%/42%), chr.22 (24%/24%), and chr.13 (14%/14%) (**Fig.2b**). GISTIC2.0 analyses underlined the significance of lesions on chr.11 and chr.8 (**Fig.S3a, TableS5**). The inv(14) and t(14;14) (93% by FISH/karyotyping) were predominantly copy-neutral. We identified recurrent (affected in >20% of cases) gains ($CN>2.5$) in 637 genes and losses ($CN<1.5$) in 1,685 genes (**Fig.S3b, TableS6**). The presence of complex karyotypes (>3 large-scale aberrations), a poor-outcome predictor in other leukemias, was a rather uniform feature (89.5%) and a higher sCNA load tended (low sample size) to associate with an inferior patient survival ($p=0.09$; **Fig.S3c**).

Aberrations on chr.11 and chr.8 are described for T-PLL[5] and have been intuitively linked to alterations of *ATM* and *MYC*. We defined here the minimally deleted and amplified regions (MDR/MAR) of these most prominent hot-spots compared to patient-derived germlines (**Fig.2c, S3d**). The chr.11 MDR was represented by strictly monoallelic losses of *ATM* carried by all MDR affected cases (31/83, 37.4%, average $CN=1.79$; less frequently involved genes in **Fig.S3b**). Identified as often co-deleted

adjacent to the MDR were the P53-suppressor network micro-RNAs miR34b/c. Genes encoding for ATM downstream effectors were affected in a minor subset (*CHEK2* loss 13.3%; *TP53* loss 4.8%).

In contrast to the assumption of *MYC* being the primary target of the chr.8 associated gains, we identified *AGO2* (*argonaute RISC catalytic component 2*), a pro-proliferative/anti-apoptotic mediator of onco-miR/siRNA biogenesis and chromatin remodeling, to define this MAR in 28.9% of cases (51.2% when HapMap controlled; average CN=2.22; **Fig.2c)**). The *AGO2* gain was independently validated using a specifically designed FISH probe (**Fig.2d, S3e**). *MYC* gains were involved in only 70.8% of cases harboring a MAR on chr.8 (average CN=2.17; **Fig.S3b**, **TableS6**). The relevance of genomic alterations of genes encoding for miR/siRNA processing factors, although not mechanistically addressed here, is further underlined by uniparental disomies (UPDs) of *AGO1/-3/-4* (all on chr.1) identified in 68.7% of cases (n=57/83; against HapMap; **TableS6**). Both, *ATM* losses and *AGO2* over-representations (mutually exclusive in 49% of cases), were each associated with a higher degree of CN-lesional complexity (genomic instability) outside their own affected regions (**Fig.2e**) and with specific GEPs (e.g. dysregulated *SLAMF6* with chr.11 MDR or reduced *CTLA4* with chr.8 MAR; **Fig.S4a-c, TableS7,S8**). Among the prominent CN lesions, *ATM* sCNAs were of negative prognostic impact (**Fig.2f**).

Generally, CN losses/gains were not implicitly linked with altered expressions of the affected genes (**Fig.S5a-c**), likely because of not depicted regulatory aspects, including allele-dominance relationships or LOH scenarios. Moreover, chr.11 MDR-independent losses of *ATM* expression and increased MYC levels irrespective of chr.8 gains were commonly observed (**Fig.S4b, S6a-c, TableS3**). This was recapitulated in TCL1A-initiated murine T-PLL: although the proliferations of *Lck*^{pr}-*hTCL1A*^{tg} mice lacked *ATM* and *MYC* sCNAs, they harbored reduced and increased expression of these genes, respectively (**Fig.S6d,e**).

The mutational landscape of T-PLL reveals driver lesions in DDR genes, dominated by clonal variants of *ATM*, but also in those affecting cytokine signaling and epigenetic control.

Samples from 53 patients were subjected to whole-genome (WGS, 3 tumor/germline (t/g)-pairs, 1 tumor 'single'), whole-exome (WES, n=33; 13 t/g-pairs), targeted

amplicon (TAS, n=20), and Sanger resequencing (platform overlap in **Fig.S1a**). Purification and separation of t/g-paired material in a 2-step sorting procedure ensured average tumor purities >98% and contamination rates <2% in germline isolates (**Fig.S1b**). This high purity, together with the general diploid karyotype of T-PLL cells (estimated with TitanCNA[18] based on CNA datasets) facilitated specific somatic calls and reliable variant allele fraction (VAF) analyses for estimations of (sub)clonal sizes or cancer cell fractions. We applied various stringent analytical filters to identify mutations likely to be biologically relevant (**Online Supplements**).

T-PLL displayed a median rate of exonic somatic mutations (~1.45 Mut/Mb) comparable to other hematologic and solid neoplasms (**Fig.3a**; **TableS9**). A global enrichment of G-to-T transversions indicates the presence of high-level genotoxic (i.e. oxidative) stress or an inefficient restorative response[19] (**Fig.S7a**). Genome-wide SNV frequencies (range for individual WES t/g-pairs 38-161) were annotated in exonic regions or at splice sites (predicted to be damaging; **Fig.3b**). GSOA (gene set overrepresentation analysis) identified enrichments of e.g. cell-cell signaling, and histone modification associated gene sets (**Fig.S7b**).

A ranking of the genes affected by those SNVs identified in WES and WGS t/g-pairs (**Fig.3c, S7c**) highlights *ATM* (76.9%, 10/13 cases) and *STAT5B* (53.8%; 7/13) by highest frequencies. Potential biological significance could also be ascribed to less frequently mutated genes based on their clustering in pathways like the DDR, i.e. its branches of nucleotide excision repair (*ERCC1*, *ERCC6L2*) or mismatch repair (*MSH3*, *MSH6*) as well as apoptosis/survival signaling, telomere maintenance, cell cycle regulation, and epigenetic modulation (**TableS9**). Aberrations of mismatch-repair genes like short *MSH3* nucleotide deletions in case *TP002* were not associated with a generally higher number of SNVs (**Fig.3b**), base-exchange preferences, differences in mutation rates by loci, or microsatellite instability (**Fig.S7d**). In contrast to nodal mature T-cell lymphomas[20], no recurrent TCR pathway mutations were enriched for in this set of T-PLL; only single hits targeting e.g. *TEC*, *VAV3*, or *NFATC2*. This suggests 'sufficiency' of the unique consistent overexpression of the TCR-signaling enhancer TCL1A[9] in conjunction with downregulation of negative TCR co-stimulatory receptors (e.g. *CTLA4*, *SLAMF6*, above) to cause net activation of this pathway.

In the 13 WES data sets of paired g/t samples allowing stringent background estimation, 31 genes were identified as significantly mutated (MuSiC with FDR<0.1), including *ATM*, *JAK3*, *STAT5B*, *ILK*, *CDC27*, *CXCR4*, *JAK1***,** and *FBXW10*. When pooled with pseudosomatic singleton WES, genes identified as significantly mutated (n=424 total) further include *CXCR2, TP53, IL7R, EZH2, USP9X, MLH1, MSH4, HIST1H1A, KDM1B, FAT2, DDX11, and FASTKD1*. This confirms the relevance of disturbed DNA repair and cytokine or apoptotic signaling. Importantly, only a small number of SNVs showed high VAFs (80-100%; 7 genes (0.7%), 14 cases; **Fig.S7e** for all SNVs), e.g. *POT1*, *USP9X,* or *FASTK*, but *ATM* was the only recurrently mutated gene with a VAF >80% and thus most likely is an early common-trunk driver (**Fig.3d, TableS10**).

The observed high frequencies of mutations in JAK/STAT signaling components, shown previously also in smaller series[21-23], underline their somatic character. Their low SNVs implicate these lesions as subclonal 'late' events (**Fig.3d**). Combining all sequencing approaches employed here, *JAK1* (10.9%), *JAK3* (21.8%), *IL2RG* (2.8%), or *STAT5B* (36.8%) were mutated in a total of 52.7% of cases. These were predominantly mismatch mutations in the SH2 (*STAT5B*) and pseudo-kinase (*JAK1*/*JAK3*) domains (**Fig.S8a,b**). The presence of these lesions did not translate into elevated JAK/STAT phospho-activation states (**Fig.S8c**), which will likely impede linear deductions of inhibitor sensitivities. Inferences on functional consequences of *JAK/STAT* mutations should also consider altered target binding properties, including dimerization. In fact, these SNVs did reveal associations with specific GEPs (**Fig.S8d, TableS11, 12**) including known JAK/STAT target genes. h*TCL1A*-tg murine T-PLL showed markedly elevated phosphorylation levels of activating JAK3/STAT5B motifs (**Fig.S8e**) corroborating a leukemogenic role of these relays of cytokine responses.

Integration of sCNA and t/g-WES data to speculate on selection for dysfunctional targets revealed that 15 of the 1497 mutated genes (including read-throughs) were affected by gain of function (GOF, CNV>2.2, VAF>0.5) or loss of function (LOF, CNV<1.7, VAF>0.5) aberrations. Somatic mutations combined with focal gains/losses were found in 85% (11/13 t/g-pairs) of cases. They dominantly included the DDR master regulator *ATM* (9/10 mutated WES cases) and the histone-Lysine N-methyltransferases *EZH2* and *KMT2D* (1 case each; **Fig.3e, TableS13**). This

emphasizes the particular relevance of genes associated with DNA repair/damage responses and epigenetic regulation. Further genes simultaneously affected by sCNAs and SNVs included the telomere protective enzymes *POT1*, *JAK1*, and *PCM1,* which are all linked to hematologic malignancies. Associations of SNVs with UPDs were found in 92% of evaluated t/g-pairs, affecting 71 genes including *IL1RAPL1* (2 cases), *STAT5B* (2 cases), *CXCR5* (1 case), and *ATM* (1 case).

SNVs affecting *ATM* were mostly missense mutations (n=35/41 lesions), less frequently nonsense (n=3/41) or frameshift insertion-deletions (InDels; n=3/41) (**Fig.3f,** validations in **Fig.S9a**), unlike the predominantly truncating lesions found in A-T individuals. We catalogued lesions at 23 unreported localizations. In contrast to previous studies suggesting an unbiased distribution of *ATM* SNVs across the entire molecule, our data from somatic *ATM*-SNV carrying T-PLL 35/53 (66%) in conjunction with those from previous series (**Fig.S9b**) revealed for the first time an obvious clustering of mutations in the FRAP/ATM/TRRAP (FAT) and PI3K domains (45/74 total SNVs). It is attractive to speculate whether this mediates selective defects of the various conventional (DNA repair, telomere maintenance) or newly ascribed (e.g. regulations of redox-equilibria, energy metabolism) functions of ATM[15].

The cooperating core lesions of *ATM* functional hypomorphism and *TCL1A* overexpression are accompanied by impaired DNA damage responses.

The vast majority of T-PLL analyzed by GE, sCNA, and SNV profiling was affected by monoallelic CNAs or/and SNVs of *ATM* (42/49, 86%; **Fig.4a**). These cases generally showed a reduced *ATM* transcript abundance (global fc=-2.32, $p=3.6x10^{-14}$ vs normal T-cells). Most frequently, ATM was subject to an LOH event (CN<1.5; *ATM* mutated, VAF>20%) (n=24/42, 57%). *ATM* expression in the 14 T-PLL with *ATM* in SN-wt constellation was unchanged in the CN-biallelic subset (n=7, fc=1.03), but highly deregulated in the CN-monoallelic cases (n=7, fc=-3.26). Very low *ATM* mRNA levels were accompanied by an enriched deregulated expression of other DDR-associated genes (**TableS8**), exemplified by *RAD50* or *FOXO3* and tended to be associated with a poorer patient outcome (**Fig.4b**). The SNV profile of the 7 *ATM* CN-biallelic/SN-wt T-PLL revealed one case with a *TP53* mutation (*TP032* (p.X215Q; VAF 0.23); CN=2), 3 *DDX11* mutated cases (2 COSMIC annotated and one stop-gain SNV), and one case (*TP026*) with multiple damaging mutations in the tumor

suppressors *MSH4*, *FAT3,* and *XRCC2*. Two cases were only subjected to a selected TAS panel, hence, may harbor similar mutations. We conclude that in a minority of T-PLL genome instability is mediated by mutations in regulators of DNA repair or a DDR other than *ATM*.

The complex karyotypes of T-PLL and its chemo-refractory behavior prompted us to causally implicate the consistent multi-level alterations of *ATM*. Therefore, we examined the capacity of leukemic cells to mount an adequate response to DSBs. The induction/resolution kinetics and patterns of induced DSB platforms marked by ATM's target γH2AX were aberrant in 82% of 22 T-PLL (**Fig.4c**). Considering the monoallelic genomic loss at the H2AX locus (*H2FAX* at chr.11q) in 19/83 cases and its T-cell lymphomagenic potential[25], we also recorded activation of KAP1, a rather specific ionizing-irradiation (IR)-induced ATM substrate (**Fig.S10a**). KAP1 mediates relaxation of particularly mutation-prone heterochromatin regions in conjunction with ATM, facilitating repair and regulation of radio-sensitivity. Although 48% (11/23) of T-PLL displayed a markedly diminished biochemical IR response (**Fig.4d, S10b**) that paralleled the altered γH2AX kinetics (**Fig.S10c**), there was a residual pATM/pKAP1 induction in most cases (70% (16/23) of samples with responses of >20% of an ATM-wt control line). Complete abrogation of such IR responses was found in the rare T-PLL with truncating ATM SNVs in analogy to $ATM^{mut/mut}$ lymphoid A-T cells, while *ATM*-biallelic/wt cases showed a more 'regular' pattern (**Fig.4d, S10b,c**). Most importantly, irrespective of any (retained) pATM/pKAP1 activation, T-PLL cells failed to generate a distal pP53 response (all 9 analyzed cases; **Fig.4d**). Given the overall rarity of 17p sCNAs and *TP53* SNVs (above) or their absence in these 9 cases, this generally implicates specific insufficiencies of p53 upstream activators (i.e. ATM). Furthermore, there was strikingly aberrant cytosolic retention of ATM upon DNA damage induction (8/11 cases; **Fig4e, S10d**). This deficient nuclear translocalization was irrespective of genomic *ATM* lesional status. The abnormally high TCR-induced ROS levels (**Fig.S10e**) and the markedly short telomeres of primary T-PLL cells (flow-FISH and WGS, **Fig.4f, S10f-h**) in correlation with the presence of *ATM* lesions further supported the notion of ATM's (partial) functional incompetence. SNVs in actual telomere maintenance genes (3 in *RTEL1*, 1 each in *DKC1*, *POT1*, and *TERT*) implicate other, more direct influences on the telomere attrition phenotype of T-PLL.

Obviously, ATM impairments in T-PLL cells are not associated with elevated chemo-/radiosensitivity. Therefore, we modeled the specific phenotypic impact of anti-apoptotic TCL1A in the mature T-cell leukemia line HH (ATM-biallelic; **Fig.S11a**). In the presence of TCL1A, the extent of DSBs was increased and their processing was markedly protracted, as per kinetics of induced γH2AX, RAD51, and TP53BP1 foci and expression levels (**Fig.4g, S11b-d**). TCL1A also propagated telomere shortening and aneuploidy (**Fig.4h,i**), in line with the data from primary T-PLL cells and pointing towards a functional influence of TCL1A on ATM´s tasks of chromosome end protection and stability maintenance. This impact by TCL1A was likely not attributed to replicative stress, as there was no noticeable TCL1A-induced pro-proliferative effect (**Fig.S11e**). Affirmation of relevance of this TCL1/ATM synergism derived from a generated mouse model. It demonstrated the cooperative pro-T-lymphomagenic outcome of constitutively active TCL1A and ATM-impairment (**Fig.4j, S11f-i**).

Structural variations (SVs) and high-resolution transcript assessment highlight novel fusions and exon usages of pivotal genes.

Somatic intra- and inter-chromosomal SVs detected by WGS revealed a high heterogeneity among cases with COSMIC listed SVs recurrently affecting chr.8, 11, 14, 16, and 21 (**Fig.5a**, **TableS14**). SVs identified in whole-transcriptome sequencing (WTS) data sets reflected fusion transcripts in 13/15 cases (**TableS15**; by TopHat-Fusion). They included the hybrids: *JAK2* (chr.9) - *TCF3* (chr.19) as well as *TRIM22* (chr.11) - *JAK2*, *KANSL1-ARL17A* (both chr.17) in 3 cases, and 3 chr.8-intrinsic fusions involving *PLEC* with varying partners (*CYHR1*, *GRINA*, *SHARPIN)* (**Fig.5b**), the latter most likely generated by the complex rearrangements at chr.8. A SEPT-ABL1 fusion reported in an anecdotal T-PLL[26] or fusions found in nodal mature T-cell lymphomas[27] were not identified. Somatic SVs detected in whole-genomic and exonic regions emphasized the inv(14) or t(14,14) as the most common structural aberrations (n=3/3 by WGS, 10/13 by WES; **Fig.5c, S12a, TableS14**).

In *TP003* the inv(14) links *TCL1A* to *TRAJ49* (TCR-α joining element 49). This newly identified fusion transcript was validated using two additional methods: (1) bioinformatically using STAR 2.5 with STAR-Fusion and (2) by RT-PCR combined with Sanger Sequencing. As the first report of a TCL1A fusion instead of the usually more *in-trans* positioning, it was striking to observe expression of a viable transcript and of neighboring *TCL1B* alongside intermediate TCL1A protein levels (**Fig.5b-d**,

S12b). WES corroborated the SNP-array results of *AGO2* being the most prevalent target of the sCNAs on chr.8 by showing its gains in 61% of cases (n=20/33; *MYC* gain in only one case).

WTS confirmed the prominent overexpression of *TCL1A* and other top-scores from the GEP analysis (**Fig.S13a,b, TableS16**). There were novel variable *TCL1A* transcripts with 4 dominant forms: TCL1A-001 (fc=39.27, $p=3.1x10^{-8}$), a truncated TCL1A-007 (fc=29.2, $p=1.2x10^{-7}$), TCL1A-201 (fc=11.4, $p=3.4x10^{-5}$), and TCL1A-002 (fc=8.3, $p=3.4x10^{-6}$) (**Fig.S13c**). The downregulated *ATM* expression in T-PLL (above) was reflected by the significantly lower expression of 5/7 protein encoding transcript variants (**Fig.S13d**). Allowing inferences on tumor-associated alternative splicing, we identified 2865 genes ($p<0.005$; $q<0.05$; $|log2\text{-}fc|>2$) exhibiting a differential exon usage compared to healthy-donor T-cells (**TableS17**). Among the most significant were those from *ATM*, *ATR*, *BCL2* (a short anti-apoptotic version preferentially expressed), histone modifiers including HDACs -2, -4, -5, -7, and -9, and TCR / cytokine signaling elements (*PIK3R1, RELA*, *NFKB1*, *NFATC1*) (**Fig.5e**).

[redacted]

We interrogated T-PLL cells for exploitable vulnerabilities, especially around their ATM-incompetence. Several strategies to intercept in synthetic lethal relationships including targeting of DNAPKcs, in conjunction with mTOR, even ATM itself, all in the context of etoposide or cyclophosphamide-mediated DNA damage, did not result in marked reductions of cell viability (**Fig.S14a-e**). Instead, several notions prompted us [redacted] (1) treatment resistance is linked to altered epigenetic codes, (2) histone deacetylase inhibitors (HDACi's) show a high activity especially in T-cell tumors and might reprogram resistance in T-PLL[28], (3) DNA-repair depends on histone modifications[29], (4) sufficient ATM activation involves its HAT mediated acetylation[30], and (5) our profiling data identify various recurrent dysregulations in histone modifiers (above; **Fig.S15a, TableS18**). For DSB induction, we opted for the multi-functional nucleoside-like alkylator bendamustine. It recently showed remarkable second-line activity in alemtuzumab-refractory T-PLL[31] and its profile of preferred activation of nucleoside-excision repair (NER) seemed ideal in the face of the multiple NER-gene SNVs discovered in our cohort. Encouraged by a reconstitution of a bendamustine-induced

DDR through the pan-HDACi SAHA (vorinostat), we explored a novel first-in-class

Evolution of genomic events during T-PLL progression.

To reconstruct the chronology of genomic alterations and hierarchic changes in clinically overt T-PLL, follow-up sampling is important. However, given the regularly short survival of T-PLL patients, this imposes a major challenge. Here, sequential samples (diagnosis (t_1) vs follow-up (F/U, t_2)) of up to 5 T-PLL (constant sample purities), were analyzed (**Fig.S16a**). Progression-associated changes were most prominent at the global mRNA level (**Fig.S16b, TableS19**). The few genes with unchanged dysregulated levels were frequently those that most significantly contributed to the difference of T-PLL to normal T-cells, i.e. CTLA4, SLAMF6 (down-) or SERPINA1 (upregulated; **Fig.1, S16b**). We also observed an increase of genomic complexity (in agreement with karyotypic data) with a trend for more sequential gains or losses of genes at t_2, p=0.06 (Mann-Whitney test, **Fig.S17a,b**, **TableS20**). In a time-line resolution of exonic SNVs we observed most mutated genes to overlap between t_1 and t_2 including the prominent mutations in ATM, STAT5B, JAK1, and/or JAK3 (**Fig.7a, TableS21**). Overall, there were t_1-restricted calls, a slightly increased overall number of SNVs at F/U, and affected genes specifically enriched in the progressed / post-therapy relapse sample. These observations point towards ongoing genomic instability affecting large-scale genomic lesions as well as towards dynamic changes of SNV-defined clones, likely also influenced by therapy. This dynamic clonal composition is furthermore highlighted by changes of VAFs of specific SNVs, especially involving ATM and/or JAK1/JAK3/STAT5B (**Fig.7a, S17b**)**.** For ATM in F/U case-1 (TP094), the VAF increase was attributable to the loss of the remaining wt-allele at t_2 (CN<1.5, **Fig.4a**), which was accompanied by an increased

downregulation of *ATM* mRNA: fc^{t1}=-1.63 vs fc^{t2}=-2.35). SVs (WES-based) and their read depths revealed a slight increase in TCL1A/TCR breakpoint frequencies in all cases alongside increased TCL1A mRNA expression (average fc^{t1}=4.24, p=0.09 vs fc^{t2}=11.34, p=0.03) with an additional breakpoint appearing at t_2 in 1 patient.

Although *TCL1A* carried prognostic information, this derived from a rather moderate variability at generally high levels (**Fig.1**). Based on global gene expression changes, we performed regression modeling to more sensitively infer on a yet-indolent phase or a particularly aggressive course after diagnosis through identification of genes with a wider range of expression and outcome-associated changes. A most informative index of 2 differentially expressed genes (*RAB25*, *KIAA1211L*) originated from a learning cohort and provided high discriminatory power toward clinical outcome based on stratified index values in the test cohort (**Fig.S17d, Online Supplements**).

DISCUSSION

In this largest reported cohort of T-PLL, for the first time virtually every case (95.2%) fulfilling the WHO classification criteria[1,34], demonstrated a genomic rearrangement involving a TCL1 gene and/or its overexpression (**Fig.S2d**). TCL1A augments signals from the most central growth receptor of T-cells, the TCR[10]. As in TCL1A-initiated murine T-PLL, this primary step towards perturbation of a protective T-cell homeostasis[35] entails additional downregulation of negative TCR regulators (e.g. *CTLA4, SLAMF6*) upstream of a prominent activation / proliferation profile (e.g. *MYC*, *NFKB2*). This appears as a shared signature by all 3 TCL1 oncogenes.

As a phenotypic hallmark of T-PLL, we identified a pronounced genomic instability, demonstrated by complex losses and gains with newly defined MDRs and MARs and by rearrangements including novel molecular hybrids. *ATM* is the gene most recurrently affected (86%, **Fig.7b,c**) by allelic losses (52%) and/or clonally dominant mutations (66%). Beyond the presence of such LOF lesions, T-PLL cells revealed aberrant DSB-induced recruitment and diminished activation of ATM and its substrates. Similar to *ATM*^{null} A-T cells, p53 activation was severely impaired.

Obviously, major ATM/p53-mediated branches of the DDR to restore genome integrity or to execute a safeguarding apoptotic response, e.g. to oncogenic stressors or to therapy, are disrupted in T-PLL. We show that some functions of ATM (e.g. damage sensing, platform recruitment, selective target engagement) are preserved at sub-sufficient levels, which also might be due to incomplete compensation by stand-in's (i.e. ATR). Importantly, we provide first hints that consequences of functional ATM deficiencies (e.g. in regulated redox homeostasis or maintenance of telomere length[15]) are aggravated by specific effects of TCL1 (**Fig.4**). In support, we previously showed TCL1A to augment mitochondrial ROS biogenesis[37]. As full ATM incompetence per se is pro-apoptotic, the coinciding impact of TCL1 likely perturbs such protective programs making this a powerful pro-leukemogenic liaison. In fact, TCL1A can rescue the apoptotic phenotype of A-T cells while potentiating their chromosome fragility[12,38].

Indeed, our genomic data implicate constitutive activation of TCL1 together with deficient ATM as the central molecular feature (**Fig.7c,** both lesions in 75.9%) that is functionally cooperative to initiate T-PLL (**Fig.4**). This preferred lesional partnership is

not observed in other T-cell lymphomas[34,39]. The spectrum of further pivotal alterations (**Fig.7b-d**) includes amplified programs of *MYC* or miR-based dysregulations (e.g. *AGO* genes, *MIR34* cluster), mutations in the JAK/STAT axis (52.7%) potentially towards late-stage TCR/cytokine independence, or affected cell-cell interaction and immune evasion. Emerging data, e.g. on the impact of JAK/STAT signaling on non-canonical functions of histone modulators like EZH2[40] indicate yet unrecognized cross-talk between the affected functional branches in T-PLL. Our data further indicate no role for viral integration, kataegis/APOBEC events, or chromothripsis.

Based on the overall recurrence of catalogued aberrations, the most commonly affected functional branch was the DDR (**Fig.7c**). However, at the regulatory level the category of epigenetics, predominantly defined by histone modifying molecules (*EZH2*, HDACs, HATs, and HMTs) was most frequently involved. This is intriguing because chromatin modulation is an increasingly recognized determinant of proper DSB processing and dictates treatment resistance[26-28]. In light of the need for non-conventional therapies in T-PLL, we devised from that a successful interventional strategy of a unique customized HDAC-inhibiting/DSB-inducing agent that reconstitutes a sufficient DDR in preclinical T-PLL models and for which a clinical trial has been commenced (NCT02576496).

Overall, the presented molecular profiling and functional interrogations allowed the formulation of a first integrative model of step-wise T-PLL leukemogenesis to be expanded on (**Fig.7e**). It provides a concrete basis for refined diagnostics, prognostication, and therapeutic concepts in this problematic disease.

METHODS

Materials, protocols, associated references, supplementary results, and source data files are available online. Accession codes: GEP, WES, WGS and WTS data sets have been deposited under GEO **XXXXXX** the dbGaP **XXXXXX**.

REFERENCES

1. Herling, M. *et al.* A systematic approach to diagnosis of mature T-cell leukemias reveals heterogeneity among WHO categories. *Blood* **104,** 328–335 (2004).
2. Matutes, E. *et al.* Clinical and laboratory features of 78 cases of T-prolymphocytic leukemia. *Blood* **78,** 3269–74 (1991).
3. Dearden, C. How I treat prolymphocytic leukemia. *Blood* **120,** 538–551 (2012).
4. Hopfinger, G. *et al.* Sequential chemoimmunotherapy of fludarabine, mitoxantrone, and cyclophosphamide induction followed by alemtuzumab consolidation is effective in T-cell prolymphocytic leukemia. *Cancer* **119,** 2258–67 (2013).
5. Delgado, P., Starshak, P., Rao, N. & Tirado, C. A. A Comprehensive Update on Molecular and Cytogenetic Abnormalities in T-cell Prolymphocytic Leukemia (T-pll). *J. Assoc. Genet. Technol.* **38,** 193–8 (2012).
6. Virgilio, L. *et al.* Identification of the TCL1 gene involved in T-cell malignancies. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 12530–4 (1994).
7. Teitell, M. A. The TCL1 family of oncoproteins: co-activators of transformation. *Nat. Rev. Cancer* **5,** 640–8 (2005).
8. Virgilio, L. *et al.* Deregulated expression of TCL1 causes T cell leukemia in mice. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 3885–3889 (1998).
9. Gritti, C. *et al.* Transgenic mice for MTCP1 develop T-cell prolymphocytic leukemia. *Blood* **92,** 368–73 (1998).
10. Herling, M. *et al.* High TCL1 expression and intact T-cell receptor signaling define a hyperproliferative subset of T-cell prolymphocytic leukemia. *Blood* **111,** 328–337 (2008).
11. Stilgenbauer, S. *et al.* Biallelic mutations in the ATM gene in T-prolymphocytic leukemia. *Nat. Med.* **3,** 1155–9 (1997).
12. Petrinelli, P. *et al.* Telomeric associations and chromosome instability in ataxia telangiectasia T cells characterized by TCL1 expression. *Cancer Genet. Cytogenet.* **125,** 46–51 (2001).
13. Ceccaldi, R., Rondinelli, B. & D'Andrea, A. D. Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends Cell Biol.* **26,** 52–64 (2015).
14. Burgess, R. C. & Misteli, T. Not All DDRs Are Created Equal: Non-Canonical DNA Damage Responses. *Cell* **162,** 944–947 (2015).
15. Ambrose, M. & Gatti, R. A. Pathogenesis of ataxia-telangiectasia: the next generation of ATM functions. *Blood* **121,** 4036–45 (2013).
16. Dürig, J. *et al.* Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia* **21,** 2153–63 (2007).
17. Le Toriellec, E. *et al.* Haploinsufficiency of CDKN1B contributes to leukemogenesis in T-cell prolymphocytic leukemia. *Blood* **111,** 2321–2328 (2008).

18. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. **24,** 1881–93 (2014).
19. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis **19,** 169–85 (2004).
20. Vallois, D. et al. Activating mutations in genes related to TCR signaling in angioimmunoblastic and other follicular helper T-cell-derived lymphomas. Blood **128,** 1490–502 (2016).
21. Warner, K. et al. T-Cell Receptor Signaling in Peripheral T-Cell Lymphoma – A Review of Patterns of Alterations in a Central Growth Regulatory Pathway. Curr. Hematol. Malig. Rep. **8,** 163–72 (2013).
22. Bellanger, D. et al. Recurrent JAK1 and JAK3 somatic mutations in T-cell prolymphocytic leukemia. Leukemia **28,** 417–9 (2014).
23. Bergmann, A. K. et al. Recurrent mutation of JAK3 in T-cell prolymphocytic leukemia. Genes. Chromosomes Cancer **53,** 309–16 (2014).
24. Kiel, M. J. et al. Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. Blood **124(9),** 1460–72 (2014).
25. Celeste, A. et al. H2AX haploinsufficiency modifies genomic stability and tumor susceptibility. Cell **114,** 371–83 (2003).
26. Suzuki, R. et al. Identification of a novel SEPT9-ABL1 fusion gene in a patient with T-cell prolymphocytic leukemia. Leuk. Res. Reports **3,** 54–57 (2014).
27. Boddicker, R. L. et al. Integrated mate-pair and RNA sequencing identifies novel, targetable gene fusions in peripheral T-cell lymphoma. Blood **128,** 1234–45 (2016).
28. Hasanali, Z. S. et al. Epigenetic therapy overcomes treatment resistance in T cell prolymphocytic leukemia. Sci. Transl. Med. **7,** 293ra102 (2015).
29. Ziv, Y. et al. Chromatin relaxation in response to DNA double-strand breaks is modulated by a novel ATM- and KAP-1 dependent pathway. Nat. Cell Biol. **8,** 870–6 (2006).
30. Sun, Y., Xu, Y., Roy, K. & Price, B. D. DNA damage-induced acetylation of lysine 3016 of ATM activates ATM kinase activity. Mol. Cell. Biol. **27,** 8502–9 (2007).
31. Herbaux, C. et al. Bendamustine is effective in T-cell prolymphocytic leukaemia. Br. J. Haematol. **168,** 916–9 (2015).
32. [redacted]
33. Heinrich, T. et al. Mature T-cell lymphomagenesis induced by retroviral insertional activation of Janus kinase 1. Mol. Ther. **21,** 1160–8 (2013).
34. Swerdlow, S. H. et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood **127,** 2375–90 (2016).
35. Newrzela, S. et al. Resistance of mature T cells to oncogene transformation. Blood **112,** 2278–86 (2008).
36. Stracker, T. H., Roig, I., Knobel, P. a & Marjanović, M. The ATM signaling network in development and disease. Front. Genet. **4,** 37 (2013).

628 37. Prinz, C. *et al.* Organometallic nucleosides induce non-classical leukemic cell
629 death that is mitochondrial-ROS dependent and facilitated by TCL1-oncogene
630 burden. *Mol. Cancer* **14,** 114 (2015).
631 38. Gabellini, C. *et al.* Telomerase activity, apoptosis and cell cycle progression in
632 ataxia telangiectasia lymphocytes expressing TCL1. *Br. J. Cancer* **89,** 1091–5
633 (2003).
634 39. Palomero, T. *et al.* Recurrent mutations in epigenetic regulators, RHOA and
635 FYN kinase in peripheral T cell lymphomas. *Nat. Genet.* **46,** 166–70 (2014).
636 40. Yan, J. *et al.* EZH2 phosphorylation by JAK3 mediates a switch to
637 noncanonical function in natural killer/T-cell lymphoma. *Blood* **128,** 948–58
638 (2016).
639

FIGURE LEGENDS

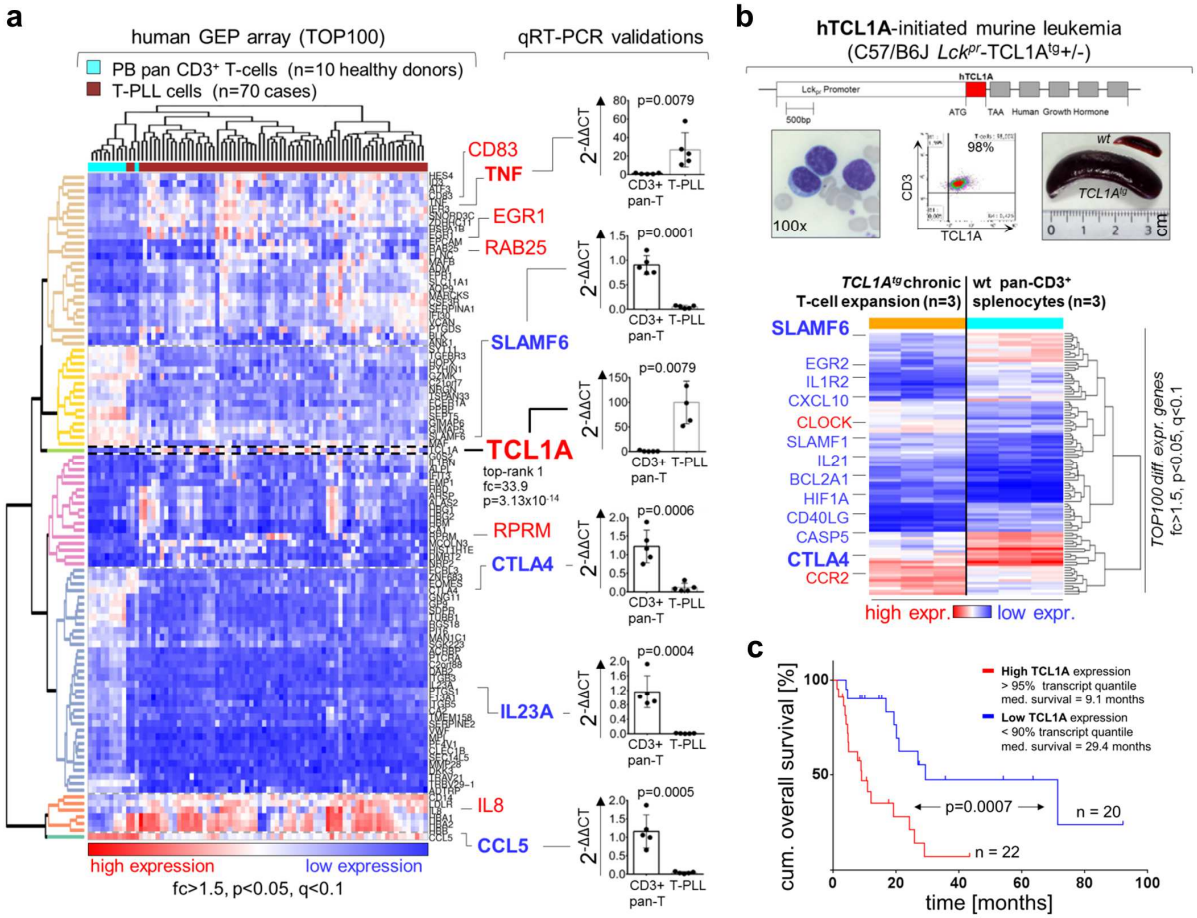


Figure 1: Gene expression profiling highlights the central role of constitutive *TCL1A* in association with dysregulated T-cell receptor signaling modulators.
a) Heatmap: Differentially expressed genes in primary human T-PLL vs normal peripheral blood (PB) T-cells with top-scoring *TCL1A* (even with *MTCP1* rearranged cases not removed). Right: qRT-PCRs for prominent genes in 5 controls/cases each (further genes in **Fig.S2c**). **b)** We re-derived a high-fidelity model resembling human T-PLL[8]. Top: *Lck*^{pr}*-hTCL1A* allele-targeting construct used; below: leukemic PB (left, mid panel) and splenomegaly (right) at overt disease stage. Heatmap: differential GEPs of murine splenic CD8^{+} T-cells at chronic stage (further data **Fig.S2e,f**). Comparison: normal splenic CD3^{+} T-cells from C57BL/6 (background-and age-matched wild-type) animals (3 hybridizations from T-cell pools of 3 mice each (total n=9)). **c)** Kaplan-Meier plot of disease-specific overall survival (log-rank test, time from diagnosis to event) of uniformly treated T-PLL patients stratified by low/high *TCL1A* mRNA expression (n=42, excluding 5% quantile 'buffer').

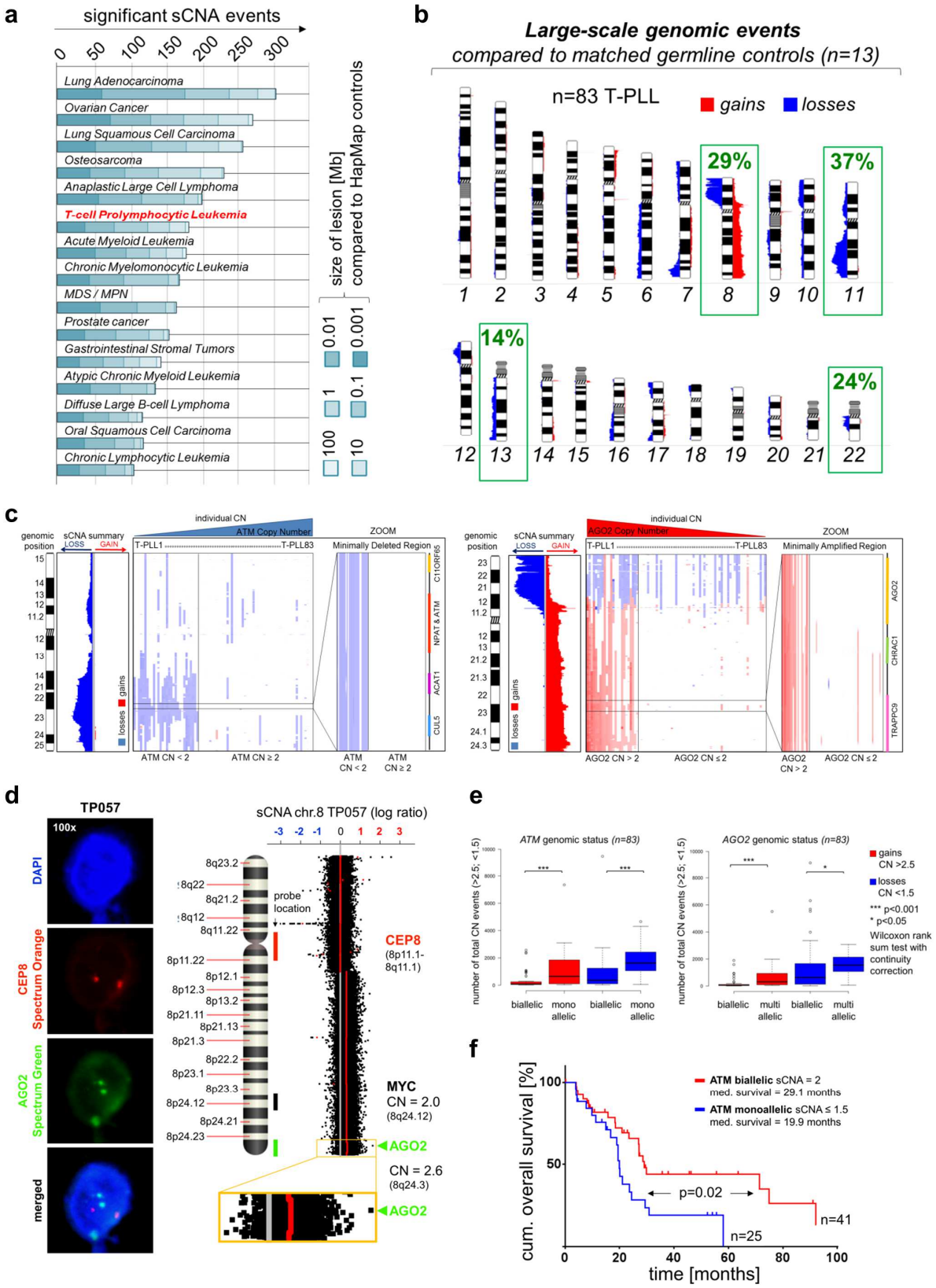


Figure 2: Legend at next page.

Figure 2: Large-scale genomic aberrations dominantly involve losses of *ATM* on chr.11q and gains of *AGO2* and *MYC* on chr.8q.

a) Number of differentially sized sCNA lesions in this T-PLL cohort (n=83) compared to publically available Affymetrix SNP 6.0 primary array data sets (all HapMap controlled, meta-analysis procedure in **Online Supplements**). **b)** Ideograms with average abundance of large-scale genomic lesions (**Fig.S3a**, **TableS5** for GISTIC2.0 analyses). **c)** Minimally deleted region (MDR) on chr.11 centering on *ATM* and minimally amplified region (MAR) on chr.8 defined by *AGO2* (for MDRs on chr.22 and chr.13 see **Fig.S3d**). **d)** Verification of *AGO2* amplification in T-PLL 057 with biallelic *MYC* (CN=2). Circular Binary Segmentation (CBS) with p-value ≤0.01 detects *AGO2,* but not *MYC* as significantly amplified using FISH. **e**) Total number of significant global gains and losses in T-PLL 'monoallelic' (CN≤1.5), "biallelic" (CN=2), and 'multiallelic' (CN≥2.5) for ATM / AGO2 excluding these affected regions. **f)** Different overall survival across 66 T-PLL subjects stratified by *ATM* CN.

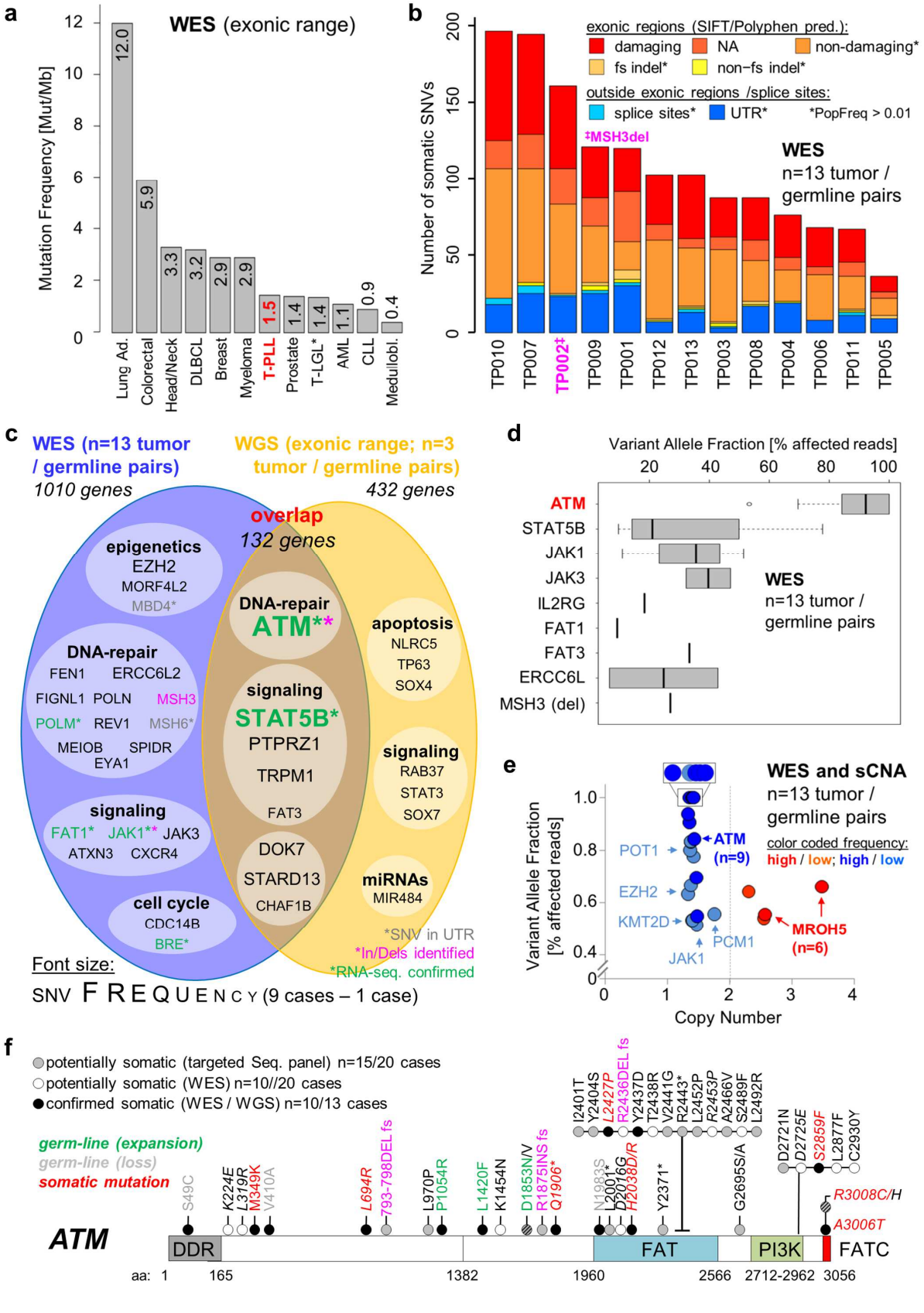


Figure 3: Legend at next page.

Figure 3: The mutational landscape of T-PLL reveals recurrent targeting of specific functional branches and uncovers new monoclonal variations in *ATM*. **a)** WES of 13 T-PLL tumor/germline (t/g)-pairs: meta-analysis (details in **Online Supplements**) comparing mutation frequencies of T-PLL cells to other malignancies (* 2 T-LGL cases sequenced as part of this study). **b)** Number of somatic SNVs per t/g-pair resolved for locations and characteristics (also **TableS9**); overall 1213 distinct SNVs: 5 frameshift insertions, 12 frameshift deletions, 7 non-frameshift deletions, 7 non-frameshift insertions, 38 synonymous, 762 non-synonymous, 19 splice sites, 96 ncRNA CDS, 39 stop-gains, 2 stop-losses, 208 within UTRs, and 17 alterations of unknown function. **c)** Mutated genes (frequencies font-size coded) identified in t/g-pairs by WES and WGS. **d)** Mean VAFs (over all mutated cases) of a selection of mutated genes. **e)** Integrated WES and sCNA data to identify genes with gain of function (GOF, CN>2.2, VAF>0.5) and loss of function (LOF, CN<1.7, VAF>0.5) aberrations. **f)** Mapped *ATM* mutations identified in WES (33 cases) and targeted sequencing (20 cases) data sets and their clustering with FAT domain enrichment. Confirmed somatic: t/g-pairs; potentially somatic: tumor singles; see also **Fig.S9** for validations and integrated meta-data with published *ATM* mutations in T-PLL.

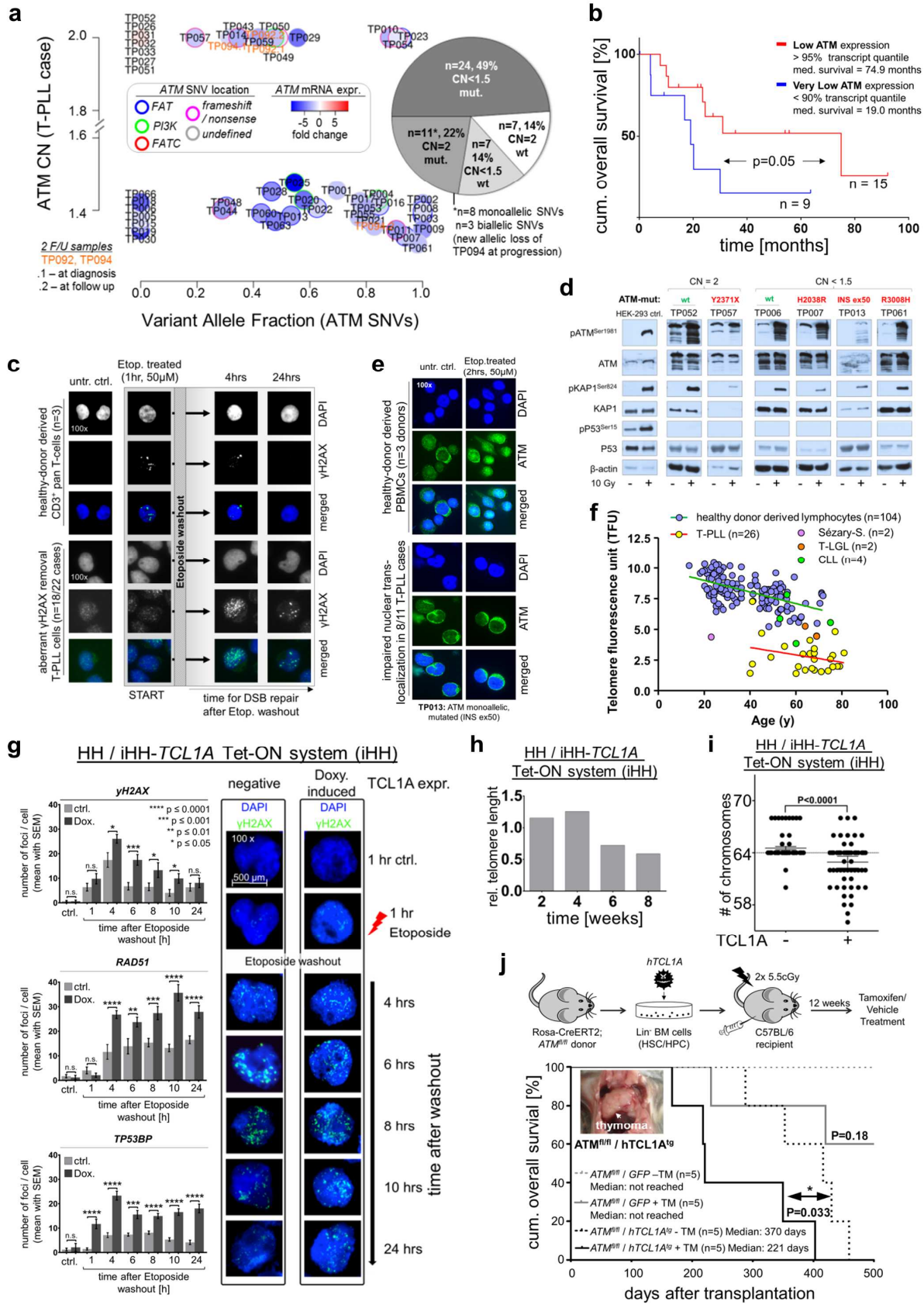
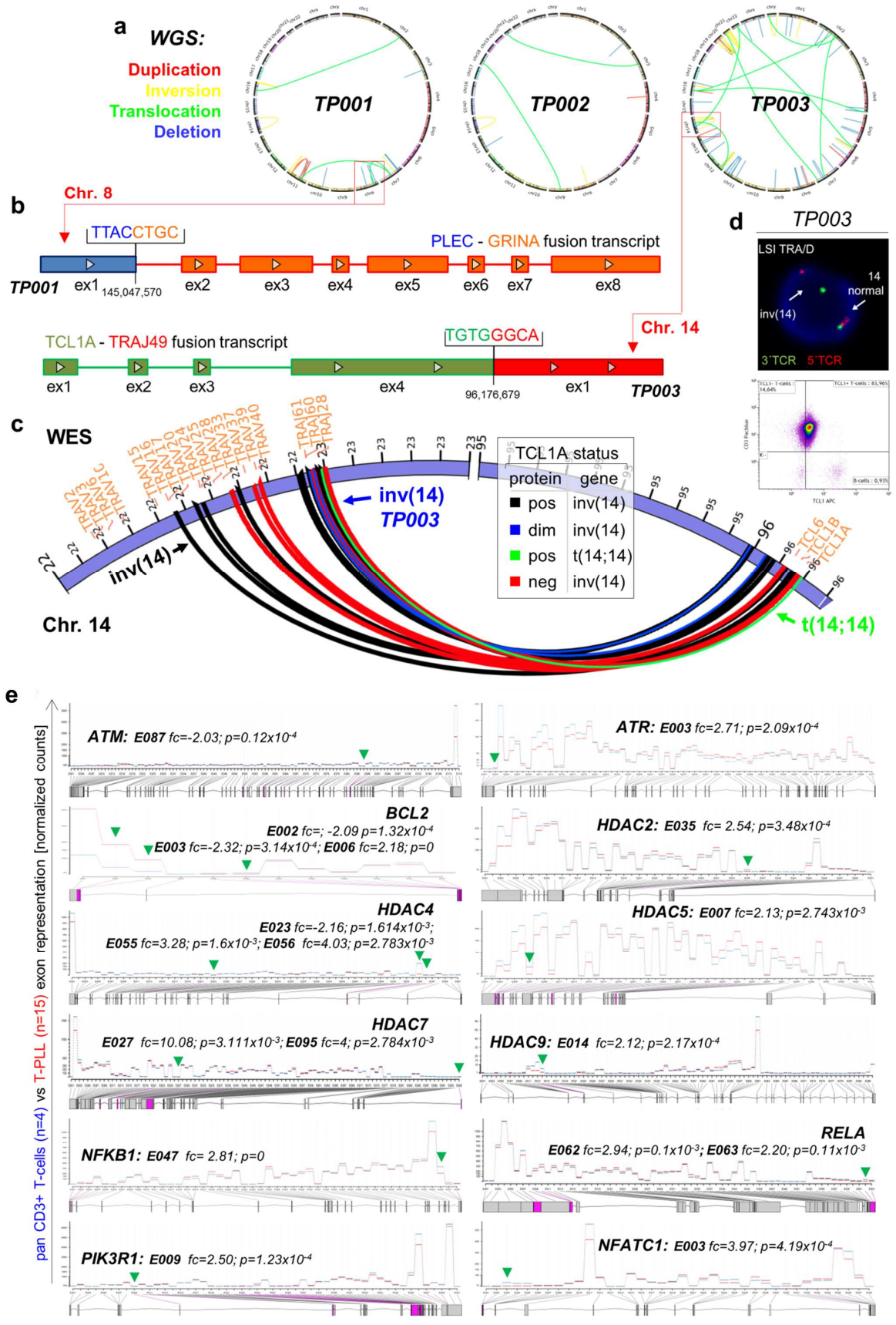


Figure 4: Legend at next page.

Figure 4: *ATM* lesions are accompanied by a phenotype of altered DNA damage response that cooperates with the impact of constitutive TCL1A.

a) *ATM* CNs and VAFs and mRNA expression for 49 T-PLL. Largest subsets among the 42 CN/SNV affected cases: LOH genotype (enriched FAT domain SNVs, p=0.01) followed by ATM-mut./biallelic cases (enriched frameshift or nonsense SNVs, p=0.01 Fisher's exact test). UPDs in 3 cases: *TP010*, *TP023*, and *TP054*. **b)** Shorter overall survival of T-PLL subjects with lower *ATM* mRNA expression (GEP arrays, 5% quantile 'buffer'). **c)** Abnormal formation and kinetics of DSB-induced (etoposide) foci in T-PLL cells (1 case, IF microscopy). Frequently higher basal yH2AX focus counts (and protein levels, not shown) and insufficient or delayed induction with inefficient/protracted removal. **d)** KAP1 (n=23) and p53 (n=9) phosphorylation upon 10Gy ionizing irradiation (IR) in T-PLL cells; ATM/P53-competent HEK293. Representative examples with robust pKAP1 induction (e.g. ATM-wt/CN=2) or with reduced activation (median purity of T-cells 97.5%; also **Fig.S10a-c**). Despite at least weak pATM/pKAP induction for most cases, none showed a pP53 response, irrespective of genomic ATM status (lanes separated for genotype-based ordering, see also **Fig.S10b**). **e)** Aberrant cytoplasmic ATM retention upon DSB induction in T-PLL (also **Fig.S10d**). **f)** Reduced telomere lengths (flow-FISH, age-correlated) in T-PLL (1 telomere fluorescence unit (TFU) corresponds to 1kb pairs); see also **Fig.S10f-h** for WGS-based analyses and associations with *ATM* lesions. **g)** Enforced TCL1A expression in HH T-cell leukemia (doxycycline-inducible iHH) impairs resolution of DSB marks (left, quantified focus counts, **Fig.S11a-e** for controls). TCL1A overexpression mediates telomere shortening (flow-FISH; **h**) and promotes aneuploidy (**i**). **j)** Accelerated T-cell lymphoma onset and shorter animal survival by the $ATM^{fl/fl}/hTCL1A^{tg}$ genotype in a model of inducible ATM-impairment and overexpression of human (h) TCL1A (details in **Fig.S11f,g** and **Online Supplements**).

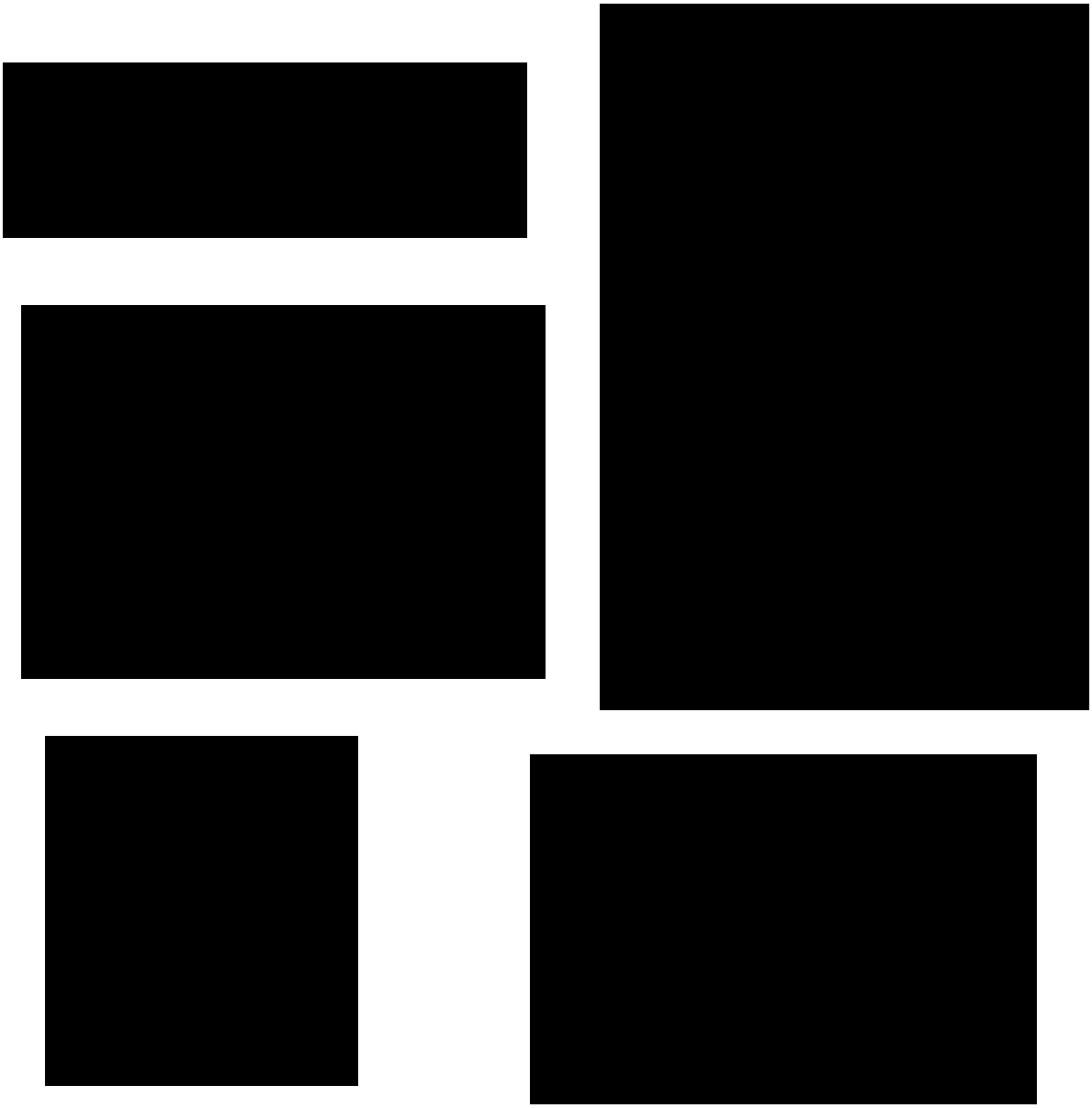


723

724 **Figure 5:** Legend at next page.

Figure 5: Key genes in T-PLL are affected by structural variations that generate fusion transcripts and show differential exon usage.

a) WGS of 3 T-PLL t/g-pairs to map intra- and inter-chromosomal translocations: 6 lesions affecting 4 distinct chromosomes (*TP001*), 10 lesions affecting 5 chromosomes (*TP002*), and 31 lesions affecting 10 chromosomes (*TP003*) (**Fig.S12a** for WES derived data). **b)** Fusion transcripts (n=96, TopHat-Fusion and oncofuse algorithms) identified by WTS of 15 T-PLL compared to healthy donor T-cells (n=4). Two examples: *PLEC-GRINA* from aberrations on chr.8 and *TCL1A-TRAJ49* from inv(14) (**Fig.S12b** for validation). **c)** Mapping of breakpoints involved in inv(14) or t(14;14); WES data on 36 (including 3 sequential) cases. **d)** The FISH-confirmed inv(14) of *TP003* (see **b**; *TCL1A-TRAJ49* fusion) was associated with TCL1A protein expression (flow-cytometry). **e)** Differentially spliced genes (selection from **TableS17**) identified by comparing WTS data of primary T-PLL cells (n=15; red lines) to healthy-donor T-cells (n=4; blue). Green arrow: exons of significantly altered usage.

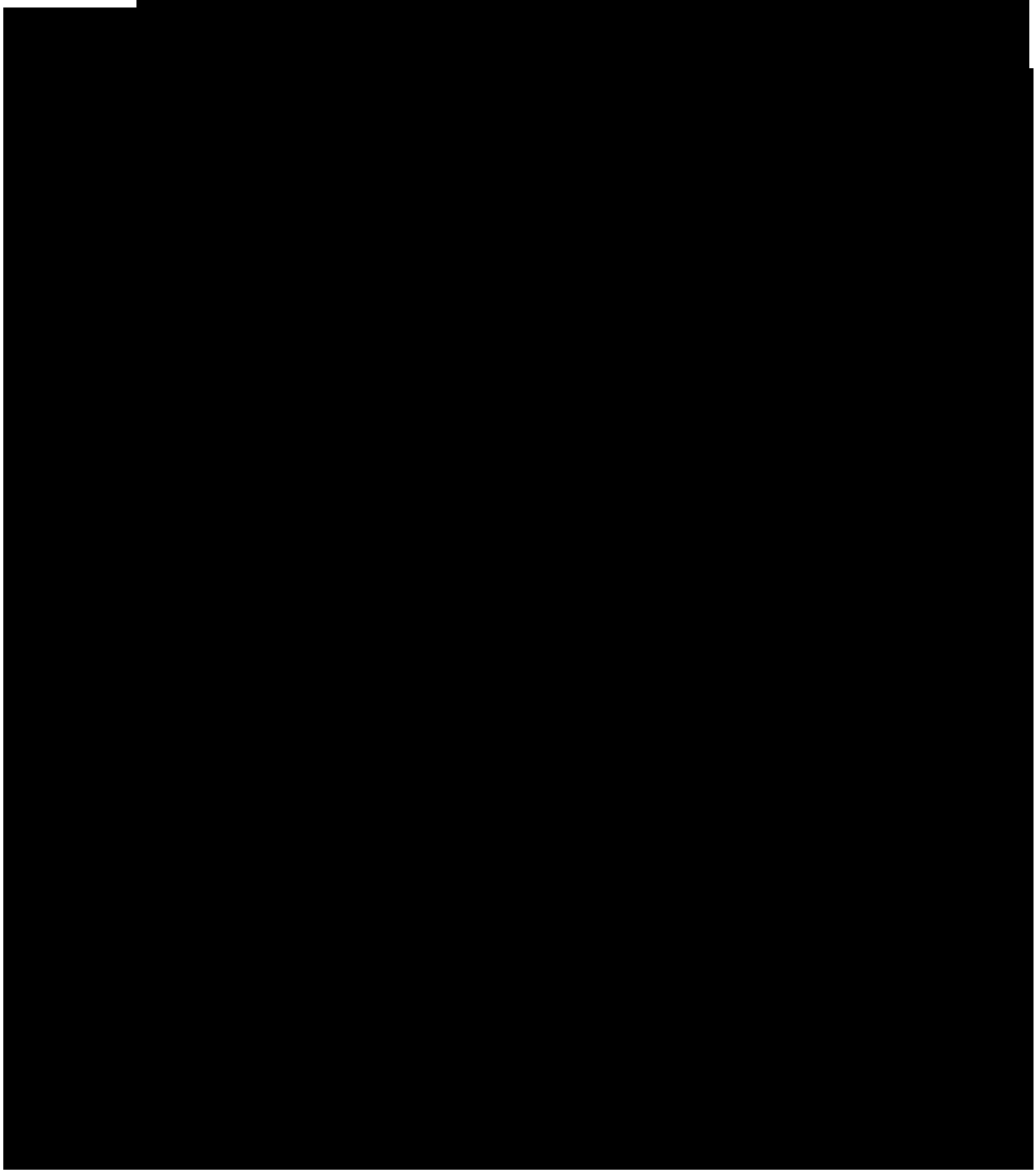


741

742 **Figure 6:** Legend at next page.

743

Figure 6:



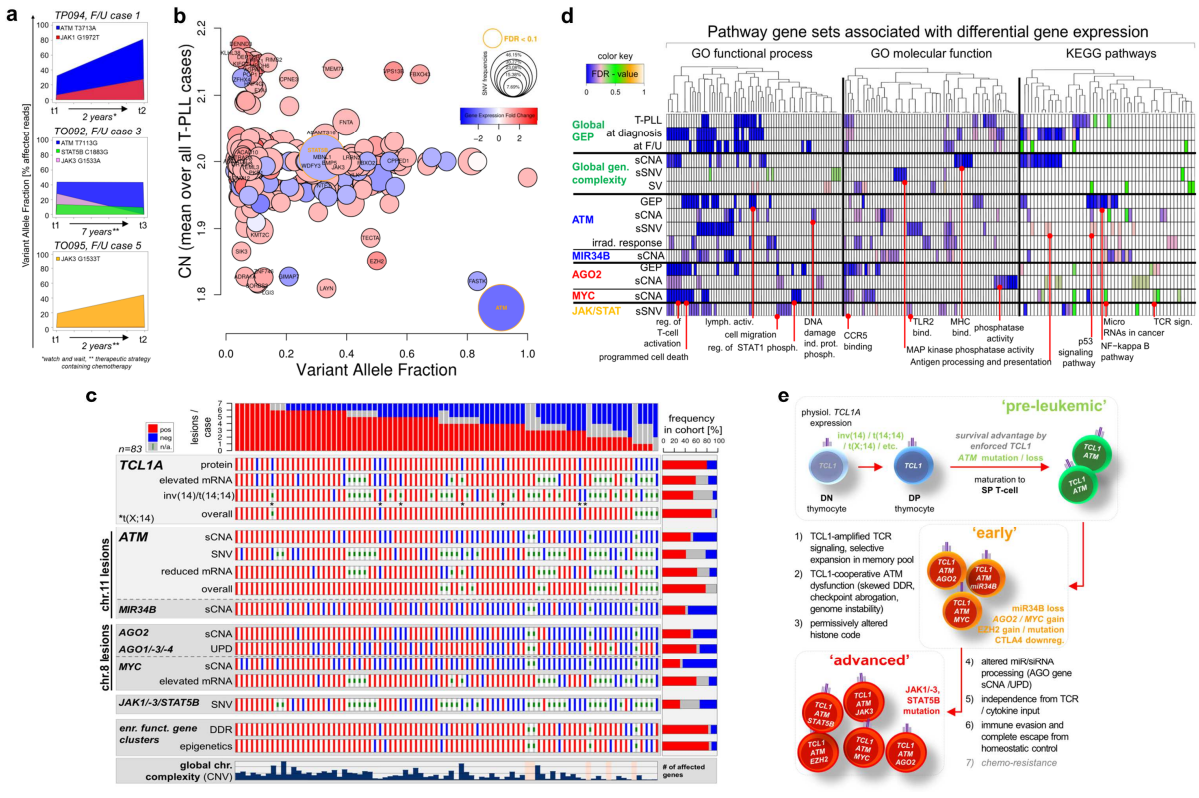


Figure 7: The integrated lesional make-up of T-PLL and a postulated disease model around the core lesions of TCL1/ATM as dominant drivers.

a) VAFs of specific mutations in pivotal T-PLL genes (*ATM*, *JAK1*, *JAK3*, and *STAT5B*) at t_1 and t_2 plotted as continua. No ploidy-correction was necessary as there was no polyploidy (usually diploidy despite single chromosome aneuploidy or CN complexity). Asterisks: therapeutic strategies (details in **Fig.S16a**). **b)** Gene-centric view of global molecular events across all analyzed T-PLL (one circle per gene). Somatic SNVs with at least one damaging prediction considered. Y-axis: sCNA-affected (CN-mean over all T-PLL); x-axis: SNV-affected (mean VAF over all detected mutations); circle size: SNV affected (frequency of SNV detected among all cases); circle border coloring: SNV FDR. **c)** Presence of dominant lesions / oncogenic events detected in GEP (high/low expression), sCNA (gain/loss), and SNV (mutation present/absent) profiling summarized for all T-PLL (red: lesion present, blue: lesion absent, grey: not analyzed). Chromosomal complexity: moderate with <2000 (n=25) and high with >5000 (n=26) sCNA-affected genes. **d)** GEPs summarized in meta-pathway heatmap. For each comparison the 50 most upregulated and 50 most downregulated ($p<0.05$) unique genes were used as an input for gene-set enrichment in GO and KEGG STRINGdb9_05 data bases. For

each GEP the 5 most significant pathways are pooled ($p<0.05$) as a matrix for visual overlap. Intensities correspond to the FDR of each GEP within a gene set and elucidate similar dysfunctions in T-PLL subsets, e.g. ‘regulation of T-cell activation’ in several *AGO2* / *MYC* associated subsets. **e)** Extrapolated model of key aberrations and functional cellular consequences in T-PLL development. Chronology assumptions (i.e. early driver events of TCL1 and ATM in a recent thymic emigrant) are based on identified frequencies in sCNA data and tumor fractions in sequencing data sets. The ‘TCL1’-lesion refers to the deregulation of any TCL1 family member.

ACKNOWLEDGEMENTS

Ma.He. and S.N. are funded by the German Research Foundation (DFG) under HE3553/4-1 and NE1438/4-1 as part of the collaborative research group on mature T-cell lymphomas, “CONTROL-T” (FOR1961), Further support: German Cancer Aid (108029), CECAD, José Carreras Leukemia Foundation (R12/08), Köln Fortune Program, and Fritz Thyssen foundation (10.15.2.034MN) (all to Ma.He.). The Volkswagenstiftung (Lichtenberg Program), the DFG (RE2246/2-1), and the Helmholtz-Gemeinschaft (Preclinical Comprehensive Cancer Center) supported H.C.R.. We thank L. Chessa (Rome, Italy) for A-T cell lines, F. Alt (Harvard Medical School and Howard Hughes Medical Institute) for ATM^{flox} mice, [redacted] [redacted] N. Riet for help with animal experiments. We furthermore thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded High Performance Computing (HPC) system CHEOPS as well as support. We gratefully acknowledge all contributing centers enrolling patients into the trials and registry of the GCLLSG; the GCLLSG staff and the patients with their families for their invaluable contributions.

CONTRIBUTION OF AUTHORS

Design and experimental data analysis: Ma.He., A.S., G.C.. Experiments: A.S., N.W., K.W., P.M., S.O., S.P., A.R., E.V., N.R., F.B., T.B., S.N.. Conduction of GEP, WES, WGS, and WTS analyses: J.A., P.N., K.E-J.. Bioinformatics: G.C., P.F., M.P.. Patient samples, immunophenotypes, and karyotyping: Ma.He., M.L., T.H., S.S.. Clinical data: Ma.He., N.P., G.H., Mi.Ha.. Key reagents: H.C.R., MH.S.. Manuscript preparation: A.S., G.C., Ma.He..

CONFLICTS OF INTEREST DISCLOSURE

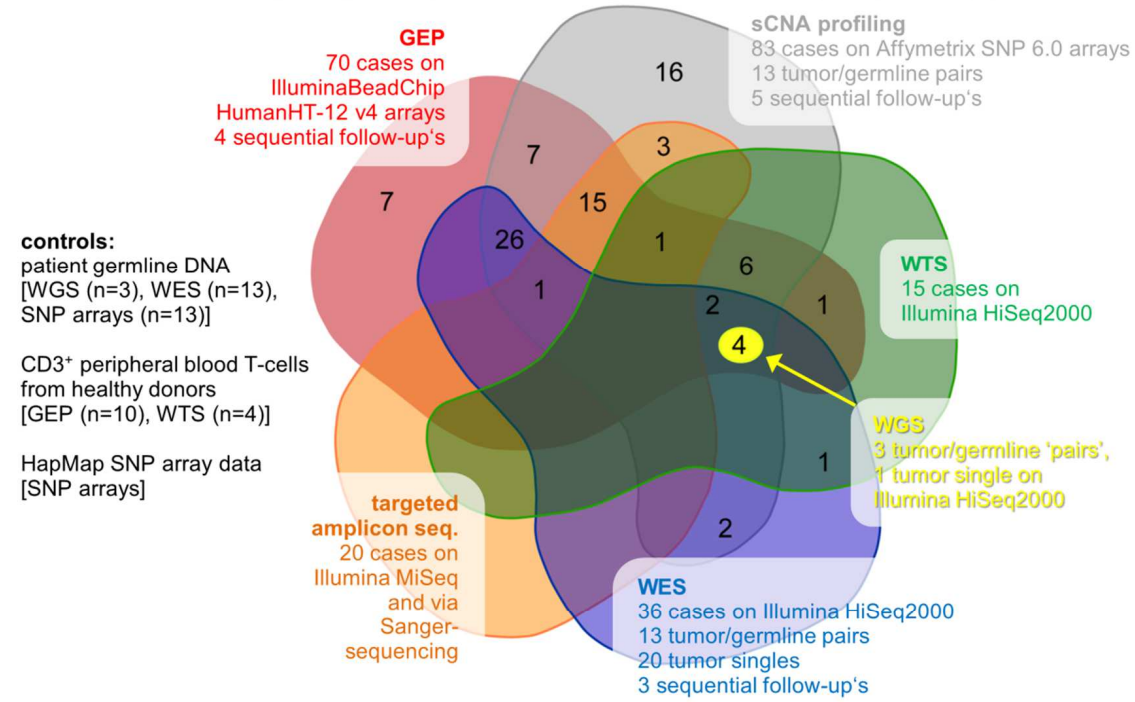
There were no competing interests interfering with the unbiased conduction of this study.

SUPPLEMENTARY FIGURES

CONTENTS

SUPPLEMENTARY FIGURES

a



b

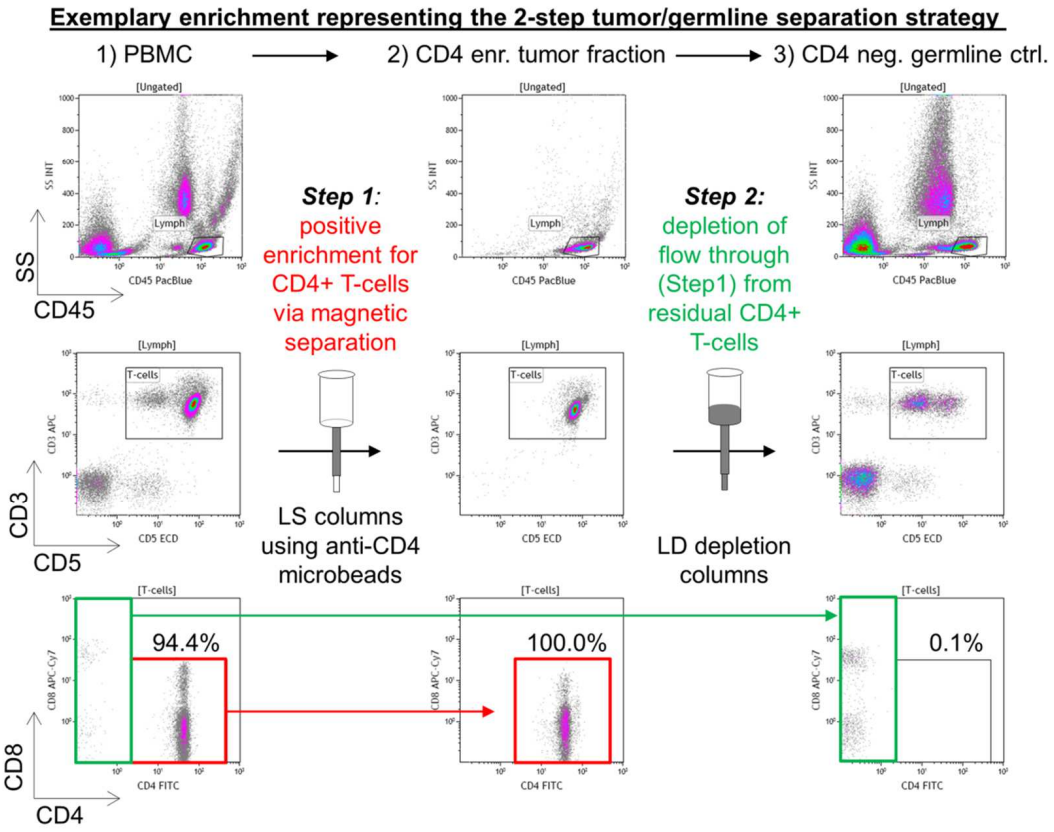


Figure S1: Legend at next page.

Figure S1: Study cohort of 94 T-PLL and controls – platforms and cell isolation.
a) Purified T-cells from 94 T-PLL patients (**TableS2** for additional information) were analyzed using various supplementing high-throughput profiling platforms (overlap indicated): Illumina HumanHT-12 v4 BeadChip arrays (n=70 cases) for gene expression profiling (GEP), Affymetrix SNP 6.0 arrays (n=83 cases) for analysis of somatic copy-number alterations (sCNAs), and the Illumina HiSeq2000 next-generation sequencing (NGS) platform. On the latter, whole-genome sequencing (WGS; n=3 matched pairs of same-patient tumor/germline (t/g) DNA, one tumor single), whole-exome sequencing (WES; n=13 t/g-pairs in addition to n=23 tumor singles including 3 cases with sequential follow-up (F/U) samples), and whole-transcriptome sequencing (WTS; n=15 tumors) were performed. Further cases (n=20 tumor 'singles') were analyzed by a customized targeted amplicon sequencing (TAS) panel including *ATM* (exons 1-63), *JAK1* (exons 9-15), and *JAK3* (exons 10-17) using the Illumina MiSeq platform and *STAT5B* (exon 16) analyzed via Sanger-sequencing based methods. $CD3^+$ pan T-cells isolated from peripheral blood (PB) of healthy donors with a similar age-median were used as "normal" controls for GEP (n=10) and for WTS (n=4). For sCNA profiling patient-derived germline control DNA from 13 t/g pairs of the 83 cases) were used as a pooled reference alone or in combination with publically available HapMap data sets (http://hapmap.ncbi.nlm.nih.gov/).
b) The isolation strategy of PB tumor cells and matched same-sample germline controls from PB mononuclear cells (PBMCs) of T-PLL patients employed a two-step magnetic separation (MACS columns) process (shown is case *TP010)*. (1) Positive enrichment of T-PLL tumor cells: magnetic beads bound to anti-CD4 or anti-CD8 antibodies (Microbeads, Miltenyi Biotec) and LS Columns (Miltenyi Biotec) were used. The specificity of beads was selected according to the individual immunophenotype. (2) Depletion of residual T-PLL cells from the flow-through designated as normal control: Depletion Columns (LD, Miltenyi Biotec) were used to remove residual CD4 or CD8 positive cells from the flow-through obtained from step 1. For further details, see **Online Methods** section.

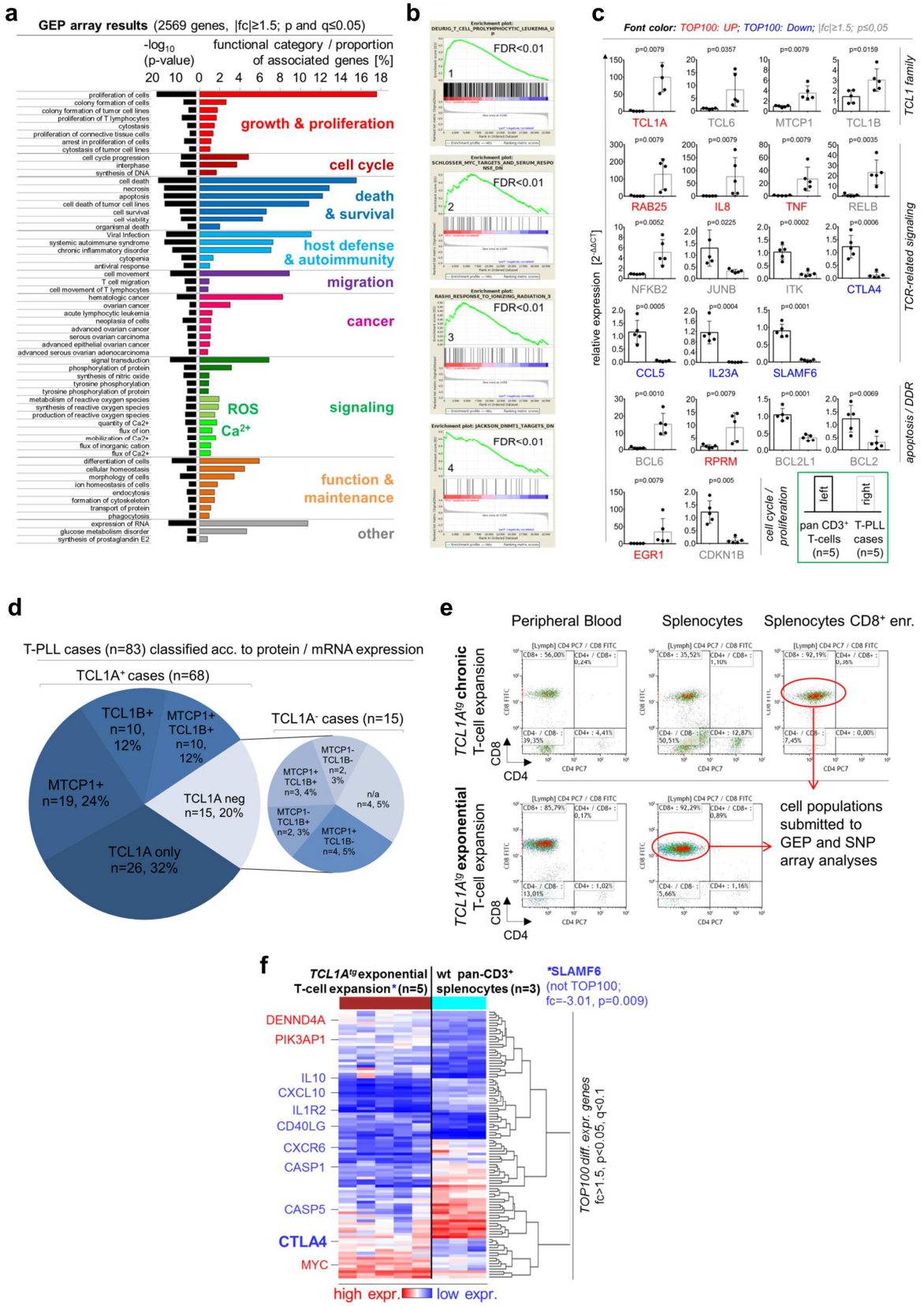


Figure S2: Functional annotations of differentially expressed genes in T-PLL with technical (qRT-PCR) and biological (*Lck*^{pr}-*TCL1A*^{tg} mice) validations.

a) Affiliation of differentially expressed genes (2569 genes; |fc|≥1.5; p ≤0.05; q ≤0.05) to functional groups in Ingenuity® Pathway Analysis (IPA): proportion of genes [%]

associated with the respective process in relation to the total number of differentially expressed genes and specific p-values (black bars). Gene sets belonging to the functional groups of 'growth and proliferation', 'death and survival', and 'host defense and autoimmunity' were most significantly enriched (see **Fig.1a** for a heat map of TOP100; **TableS3** for all differentially expressed genes). **b)** To test whether gene sets previously identified to be deregulated in T-cell malignancies or associated with T-PLL are differentially expressed in our set as well, we analyzed for overlaps using the Broad Institute's GSEA[1,2] platform in addition to general annotations by IPA (FDR<0.01, n=22 gene sets; across all MsigDB gene sets). Four examples of identified functional relevance to T-PLL show significant enrichments of genes that were: (1) previously associated with T-PLL (transcriptomes of 8 $CD3^+$ normal donor-derived PB cell samples vs 5 T- PLL[3]), (2) identified as MYC targets (transcriptional program of lymphocytes in response to MYC expression[4]), (3) activated by ionizing radiation regardless of ATM status in murine lymphoid tissue[5], and (4) identified to be targets of epigenetic modification (microarray analyses of fibroblasts from *DNMT1* knockout mice[6]). **c)** qRT-PCR validations of GEP data, including genes encoding *TCL1* family members (for *TCL6* independent gene status is still controversial[7]), TCR-related signaling molecules, and apoptosis-/DDR-associated factors (5 T-PLL *vs* $CD3^+$ pan T-cells from PB of 5 healthy donors; see **Fig.1a** for examples of TOP100 differentially expressed genes). **d)** TCL1 gene family status by protein / mRNA: TCL1A and/or MTCP1 pos. in 90.4% (n=75/83) *vs* neg. or n/a in 9.6% (8 cases). Of the latter, 2/8 showed elevated TCL1B expression, 2/8 were negative for all 3 TCL1 family members, and for 4/8 no additional data other than lack of TCL1A protein was available (n/a). Genomic data: (not shown): inv(14)/t(14;14) present in 87.0% (n=47/54); t(X;14) in 7.4% (n=4/54). Overall, combining protein/mRNA with genomic information: 95.2% (n=79/83 cases) could be assigned to overexpression or genomic rearrangement of at least one TCL1 family member. GEPs of the 2 exclusively TCL1B-pos. cases or of the 2 cases without detectable expression of any TCL1 family member were similar to those of TCL1A-pos or *MTCP1*-rearranged cases (not shown). **e)** Lck^{pr}-$TCL1A^{+/-}$ T-cells and those of age-matched C57BL/6 (wild-type) mice were enriched from splenic lymphocytes by MACS® protocols. Stages: 'chronic phase' (30-70% tumor cells in PB and spleen, average age 12 months, n=3) and 'exponential phase' (mean PB lymphocyte doubling time (LDT) 12 days; SEM 0.8; >80% tumor cells in PB, >90% in spleen, average age 15 months, n=5). Examples for cell populations submitted to GEP arrays (**Fig.1b, S2f**) and used in immunoblots (**Fig.S6e**). **f)** GEPs of *TCL1A*-induced murine T-cell leukemia at 'exponential phase' (enriched splenic $CD8^+$ T-cells) using Affymetrix GeneChip Mouse Gene 1.0 ST Arrays. Purified splenic $CD3^+$ pan-T-cells isolated from C57BL/6 mice (3 hybridizations from T-cell pools of 3 mice each (total n=9) were used as matched controls. Besides the commonly affected TCR signaling modulators *SLAMF6* and *CTLA4*, we observed an additional deregulation of T-PLL characteristic oncogenes (e.g. *MYC*) in overt murine leukemia at the exponential growth phase. See also **Fig.1b** showing the differential expression of genes in 'chronic-phase' expansions and **TableS4** listing all differentially expressed genes.

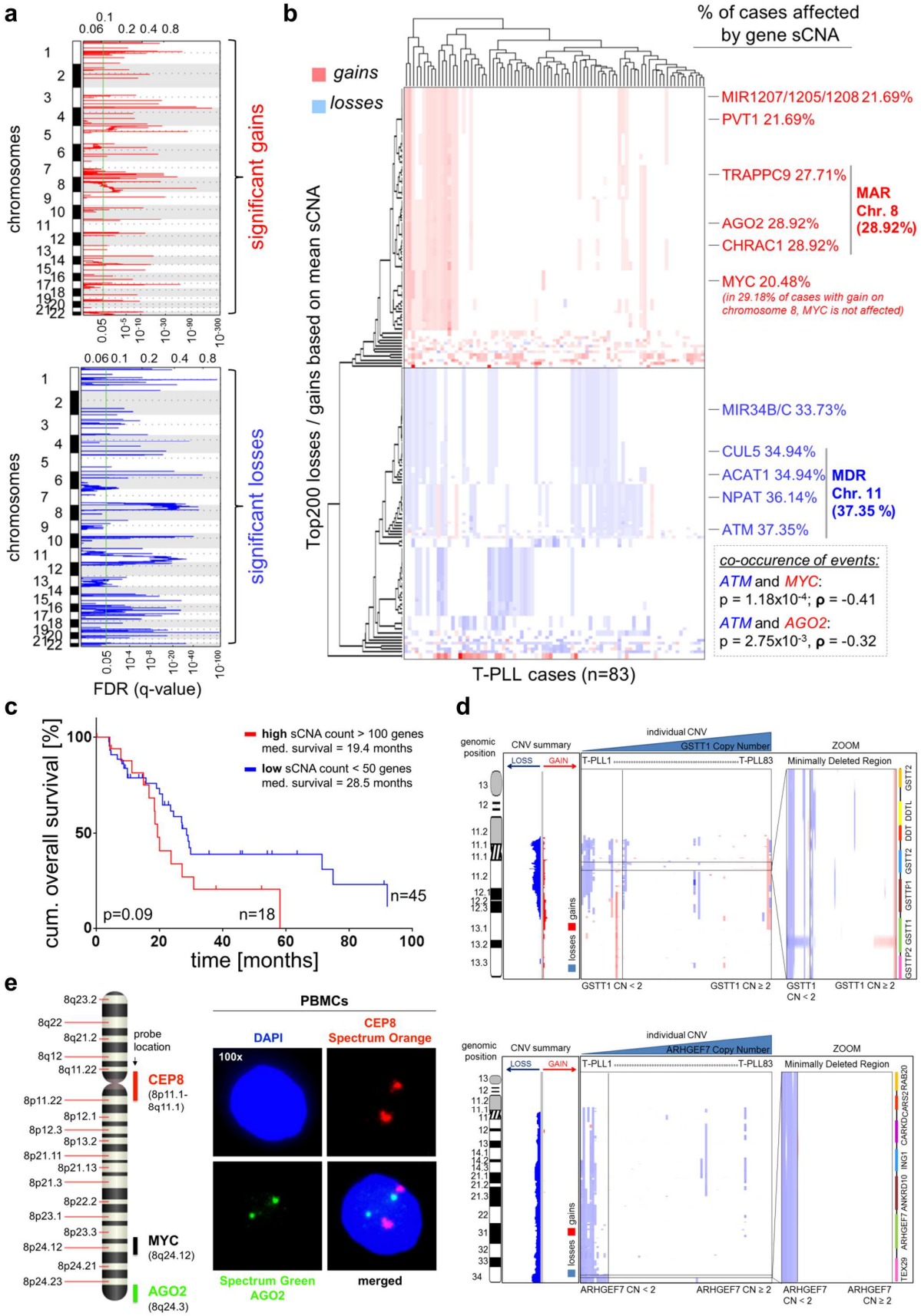
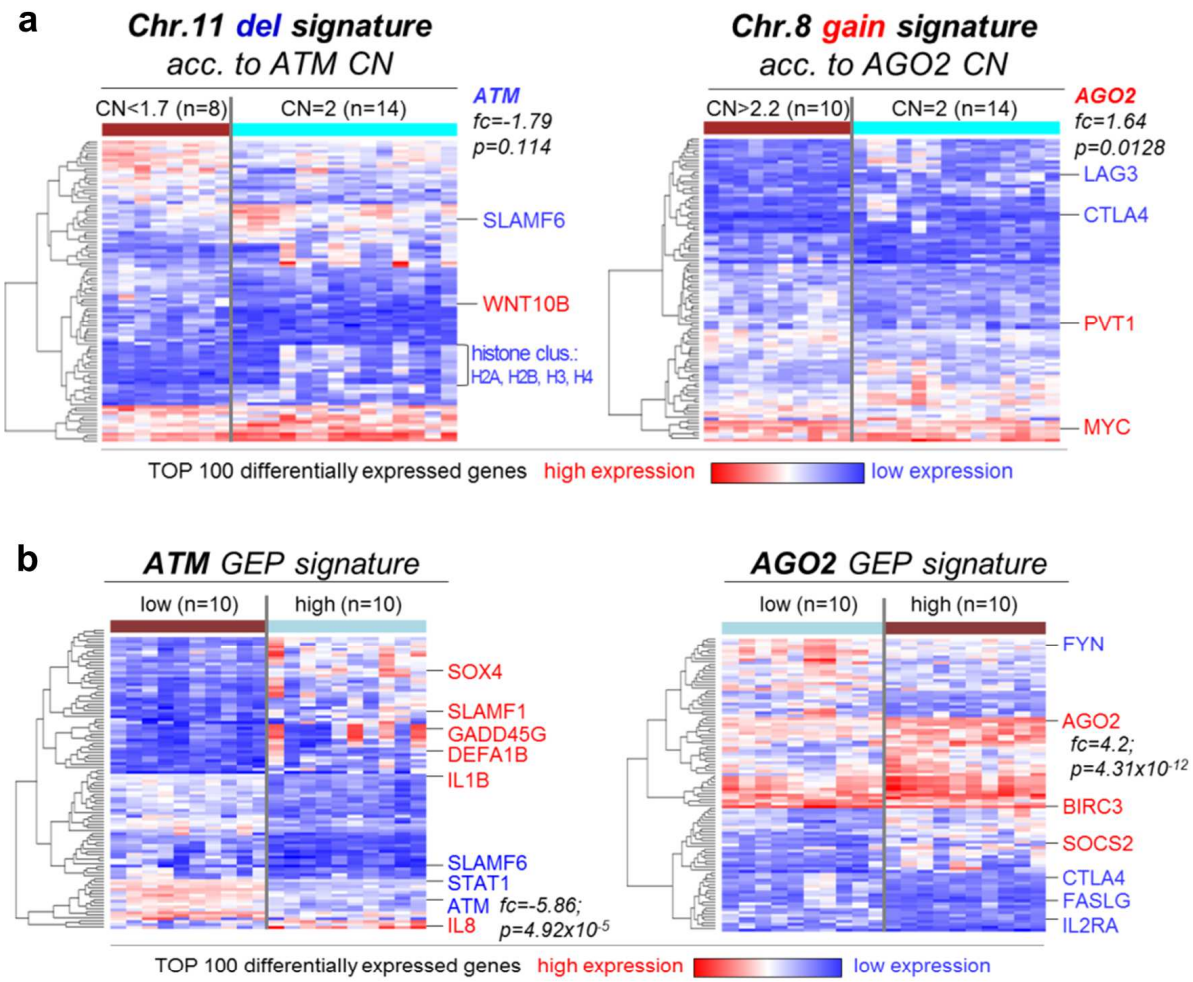


Figure S3: Legend at next page.

Figure S3: Lesions identified in sCNA profiling dominantly include losses at chromosome 11 (*ATM*) and novel gains located on chromosome 8 (*AGO2*).
Globally, we identified gains (CN>2.5) in 19,590 genes and losses (CN<1.5) in 27,193 genes (**TableS6**). The number of sCNA-affected genes (median 3354) varied inter-individually (e.g. 13,862 in *TP038* vs 42 in *TP033*). **a)** GISTIC2.0 [8] analyses showing significant gains and losses in 83 T-PLL compared to 13 patient-derived normal DNAs confirmed the enrichment of lesions on chr.8 and chr.11 (compare **TableS5** and **Fig.2b**). Among the genes that exhibit both focal gains and deletions (centers of wide peaks) with 90%-confidence level are *GSTM1* (Glutathione S-Transferase Mu 1; chr.1; CN=2.57) and *LCE3C* (Late Cornified Envelope 3C; chr.1; CN=1.64), which are also likely due to complex rearrangements. **b)** Heat map showing the color-coded CN of TOP200 gained / lost genes (CN mean across all T-PLL; red: CN>2.5; blue: CN<1.5). Genes characterizing the minimally amplified region (MAR) on chr.8 and the minimally deleted region (MDR) on chr.11 (see **Fig.2c**), were affected at the highest frequencies of CN events (in %; compare **TableS6**). Chr.11 MDR: Slightly less frequently involved than ATM were the cell cycle factor *NPAT*, the mitochondrial acetyltransferase *ACAT1,* and the Ras ubiquitin ligase *CUL5*. Chr.8 MAR: *AGO2* is more frequently overrepresented than *MYC*. **c)** Kaplan-Meier plot of disease-specific overall survival (OS) of T-PLL subjects according to 'CNA complexity'; stratification by total number of sCNAs (high: >100 genes affected; low: <50 genes affected; log-rank test, time from diagnosis to event, n=63). For an association of MDR/MAR lesions with the total number of CN events and the association of *ATM* CN with OS see **Fig.2e,f**. Presence of the MAR on chr.8 did not correlate with OS. **d)** MDRs on chr.22 (top) and chr.13 (bottom) (supplementing data to **Fig.2b,c**) showing restrictions to *GSTT1* (glutathione S-transferase theta 1, lost in 24.1% of cases) and *ANKRD10/ARHGEF7* (ankyrin repeat domain 10 / Rho guanine nucleotide exchange factor, lost in 15.7% of cases), respectively (average CN=1.91 / 1.82). **e)** Verification of biallelic *AGO2* in healthy donor derived PBMCs using FISH (control for the FISH analyses of **Fig.2d)**.



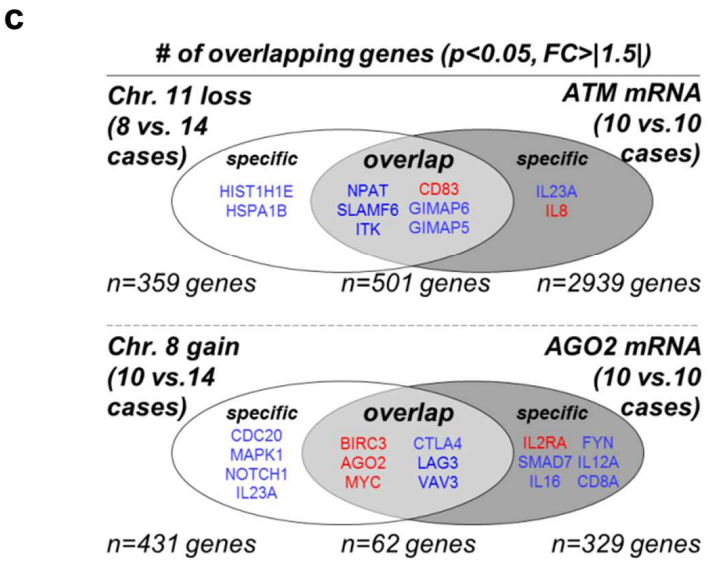


Figure S4: Legend at next page.

Figure S4: Gene expression signatures associated with specific sCNAs or with cases defined by stratified expression of respectively affected genes.

Despite a considerable co-occurrence of the CNAs at chr.11 (MDR) and at chr.8 (MAR) per each case, there was a sizable fraction of T-PLL with discordance between the presence of these CNAs, i.e. 49% of cases with an *ATM* loss did not harbor an *AGO2* gain.

a) Heat maps showing the differential expression (TOP100) of genes specifically associated with chr.11 MDR and with chr.8 MAR. For that, GEPs of cases carrying losses at chr.11 were compared to cases 'biallelic' for chr.11 (*ATM* CN<1.7 vs. CN=2 according to comparison to HapMap controls; chr.8 affected cases excluded) and GEPs of cases with chr.8 gains were compared to cases 'biallelic' for chr.8 (*AGO2* CN>2.2 vs. CN=2 according to comparison to HapMap controls; chr.11 affected cases excluded). Among the genes that 'defined' the global differences of T-PLL cells to normal T-cells regardless of sCNA status (see **Fig.1a**) some were specifically associated with these prominent sCNAs (i.e. *SLAMF6* downregulation with presence of the chr.11 MDR and *CTLA4* downregulation with chr.8 gains (MAR); **TableS7** for additional information). These MDRs/MARs are associated with intuitive fold-changes (fc) of expression of their defining genes, *ATM* and *AGO2*, respectively. There was no association of *ATM* or *AGO2* expression levels with the 'opposite' sCNA lesion.

b) Heat maps showing the differential expression (TOP100) of genes specifically associated with stratified *ATM* and *AGO2* mRNA abundance; comparison: 10 T-PLL with highest vs. 10 cases with lowest expression (fc of *ATM* and *AGO2* expression indicated). *AGO2* mRNA levels are significantly elevated in cases with lowest *ATM* expression (FC= 1.73, p=0.02), while the generally low ATM expression is not different between AGO2 high vs. low cases (see **Table S8**).

c) Gene expression signatures associated with the presence of chr.8 and chr.11 CN lesions (see a) were compared to those derived from stratified *ATM* and *AGO2* mRNA levels (see b). The GEPs of exclusively chr.11- and chr.8-affected cases appeared to be determined to a large degree by the minimal-region defining genes *ATM* and *AGO2,* based on marked overlap of GEPs: 501 of 860 differentially expressed genes associated with the chr.11 MDR are likewise associated with altered ATM mRNA expression; 62 of 493 differentially expressed genes associated with chr.8 aberrations are likewise associated with altered mRNA AGO2 expression.

Together, both frequent sCNAs and the respectively altered expression of their defining genes (*ATM, AGO2*) are associated with unique and joint signatures, but overall with a large number of genes that displayed the most differential expression (vs $CD3^+$ pan T-cells) in the entire cohort of T-PLL (not stratified by any sCNA, **Fig.1a**), i.e. *CD83, SLAMF6, GIMAP5, GIMAP6, CTLA4, or MYC*. Overall, this highlights gene-specific and region-defined contributions to the overall GEP of T-PLL (**TableS7, S8**).

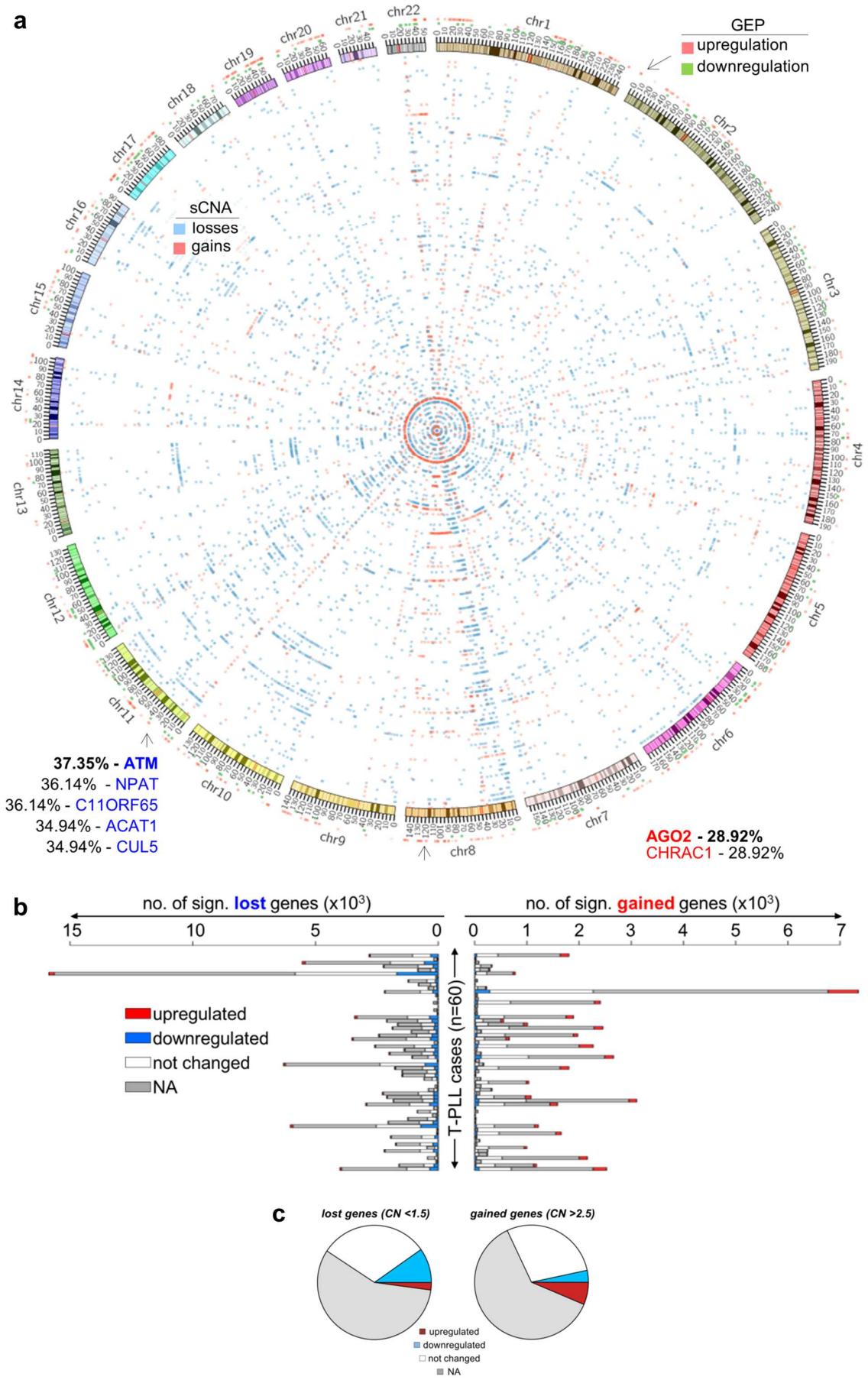


Figure S5: Legend at next page.

Figure S5: Associations of large-scale genomic lesions and deregulations of global gene expression in T-PLL.

a) Circos plot mapping sCNAs and deregulations of gene expression on chromosomal loci (%: frequencies of sCNA events across entire T-PLL cohort). **b)** GEPs superimposed on sCNAs with global data per case. CN lesions (exclusively monoallelic) were correlated with the differential expression of genes located in the respective regions. Although sCNA-associated changes in GEP were of generally intuitive directionality, a larger proportion of genes showed no down- / upregulation in the context of genomic losses / gains. **c)** Summary of b: pie charts illustrating the association of gene-specific sCNA events with differential expression of genes. For the majority of genes, their transcript abundance remained unchanged upon monoallelic losses or gains; a smaller percentage of sCNA-affected genes shows an altered expression intuitively corresponding to the respective genetic change (combination of GEP and sCNA profiling data; n=60 T-PLL cases; blue: downregulated; red: upregulated; white: unchanged; grey: not annotated (N/A). All CNA events are monoallelic.

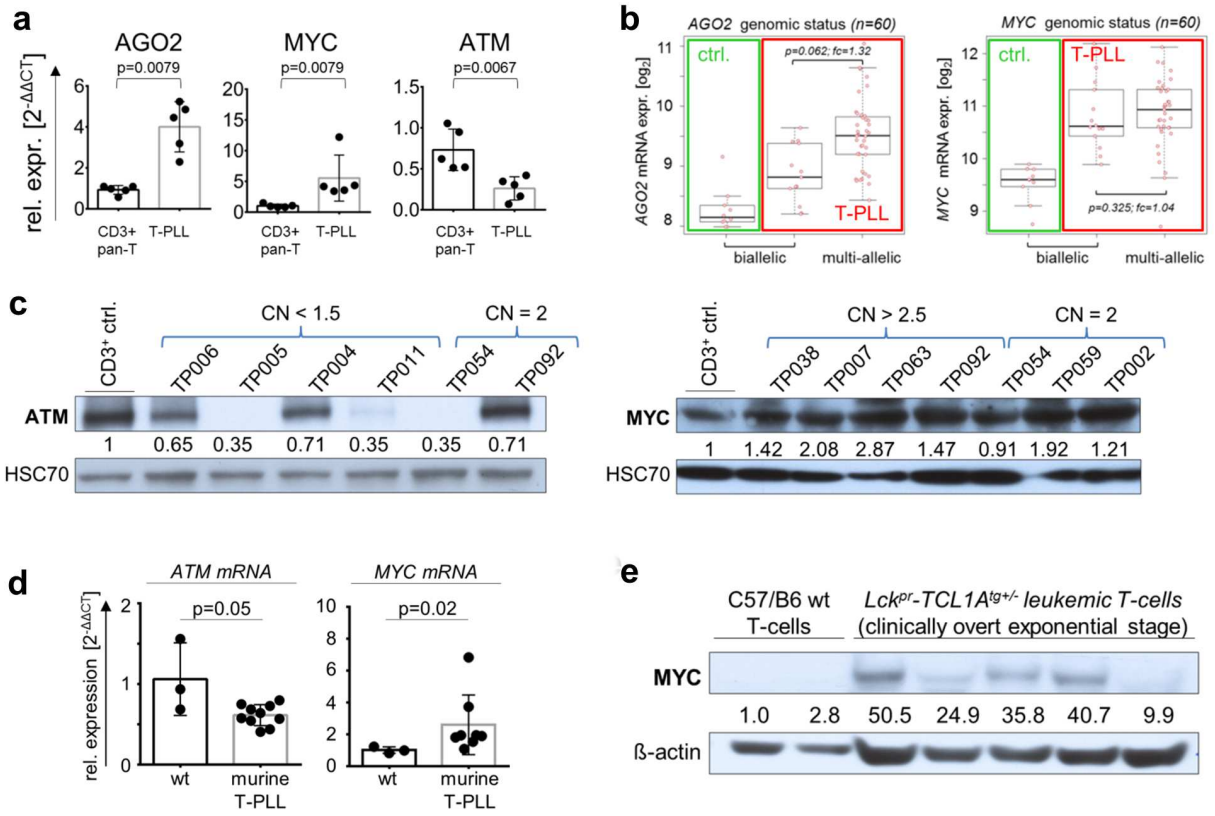


Figure S6: Changes in transcript and protein abundance of *ATM* and *MYC* are not entirely explained by somatic CNA events on chr.11 and chr.8, respectively. **a-c)** Although the genes affected by the chr.11 MDR / chr.8 MAR showed decreased (*ATM*) and increased (*AGO2, MYC*) expression (array-based, qRT-PCR, immunoblots), this was rather generally disease-associated than confined to the presence of the specific genomic CN lesion (see also **Fig.S4**). **a)** qRT-PCR: mRNA expression of *AGO2* and *MYC* was generally upregulated, while *ATM* expression was downregulated in primary T-PLL cells (n=5 cases) vs. $CD3^+$ pan T-cells isolated from PB of healthy donors (n=5); compare GEP data in **TableS3**. **b)** mRNA expression values [log2] of *MYC* and *AGO2* derived from GEP analyses in $CD3^+$ pan T-cells isolated from healthy donors (green box), and T-PLL cases stratified as 'AGO2/MYC biallelic', and 'AGO2/MYC multiallelic' (red box) according to sCNA profiling (compare **Fig.2** and **TableS6**). While *AGO2* mRNA levels showed a trend for a higher expression in 'AGO2 multiallelic' cases, *MYC* mRNA expression seemed to be generally elevated in T-PLL irrespective of the presence of a *MYC* gain, pointing to additional mechanisms upregulating *MYC* expression that are independent of genomic amplification. **c)** Immunoblots on primary human T-PLL cells, n=6 (ATM) and 7 cases (MYC), and $CD3^+$ pan T-cells from PB of healthy donors. Quantifications according to HSC70 loading control via ImageJ®. Protein expression of ATM and MYC was independent of the presence of the respective sCNA lesion, e.g. showing ATM absence (e.g. *TP054* with biallelic *ATM* SNVs) and MYC upregulation in CN-biallelic cases. **d, e)** Murine TCL1A-driven T-PLL-like expansions generally revealed a lower sCNA abundance and recurrence (average 70.7 sCNAs in chronic phase (n=3) and 74.8 sCNAs in exponential phase (n=5; CN<1.8 or >2.2)). **d)** qRT-PCRs of *ATM* and *MYC* mRNA

in splenic T-cells of background-matched wild-type and *Lck*^{pr}*-TCLA1*$^{+/-}$ mice reveals a downregulation of *ATM* and an upregulation of *MYC* although respective genetic CN lesions are not observed in leukemic T-cells of these TCL1A-tg mice, again pointing at CN-independent modes of deregulation (see **Fig.S2e** for cell enrichment, **Fig1b, S2f** and **TableS4** for GEP derived mRNA expression levels). **e)** MYC protein expression in TCL1A-driven murine leukemic T-cell expansions: immunoblot of splenic T-cells from background- and age-matched wild-type control mice (2 T-cell pools of 3 mice each (total n=6)) and from *Lck*^{pr}-TCL1A$^{+/-}$ mice with exponential phase leukemia (for definitions see **Fig.S2**, n=5) corroborated the data on upregulation of *MYC* mRNA in the usually *MYC* 'biallelic' murine leukemias (see **Fig.S2e** for cell enrichments) and paralleled the sCNA-independent MYC upregulation in human T-PLL. Quantification: ß-actin ratio via ImageJ®.

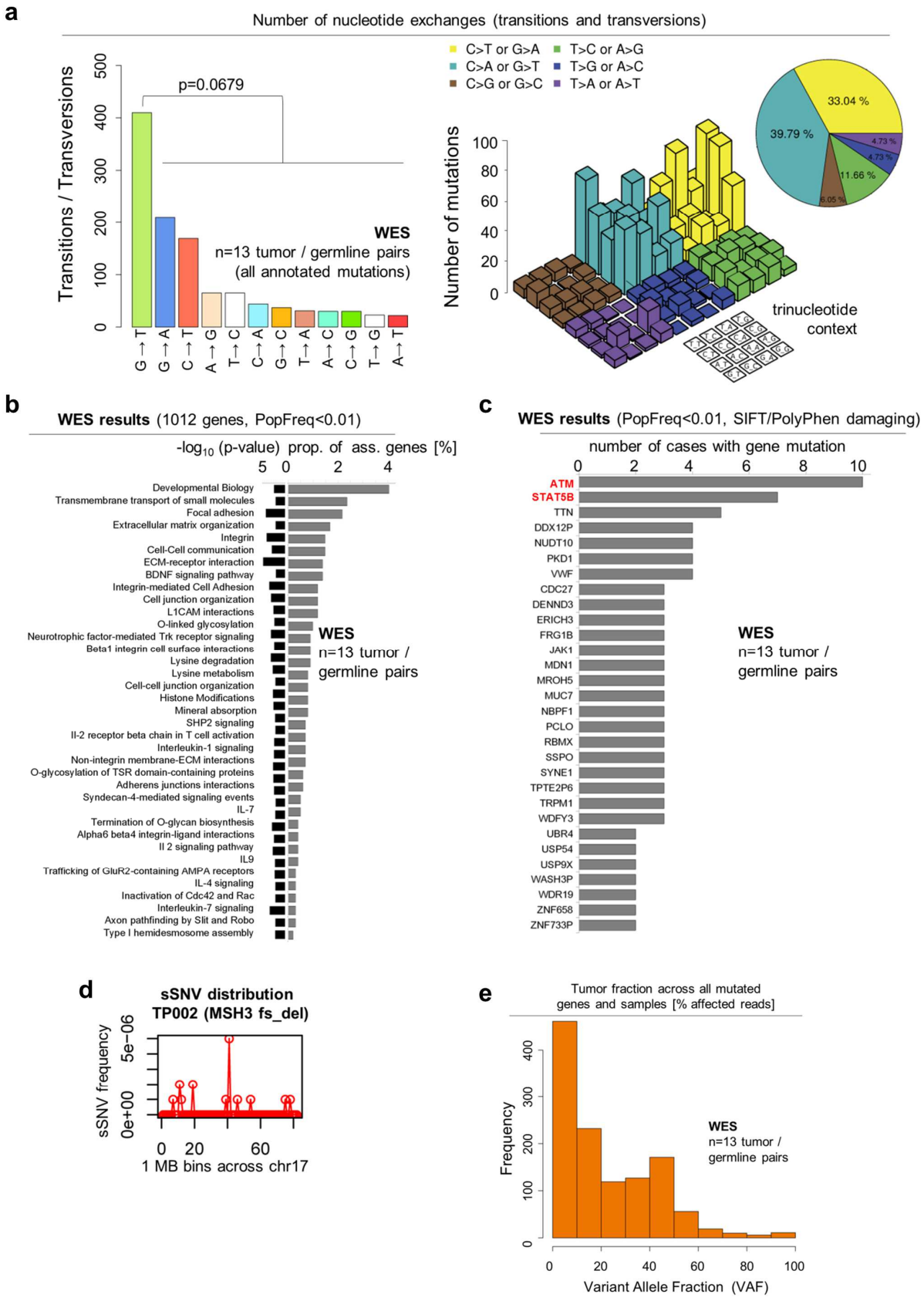


Figure S7: Legend at next page.

Figure S7: Characteristics of WES detected mutations in T-PLL.
Whole exome sequencing (WES) of 13 T-PLL tumor/germline (t/g) control pairs targeted 439,651 exons in 20,000 genes achieving a 44.1Mb target coverage at a minimum of 26-fold for both tumor and matched germlines. Since we observed a high portion of G>T (and C>A) transversions in one batch of WES samples indicative for oxidative DNA damage (8-oxoguanine (8-oxoG) lesions) during sample preparation, we applied additional filters similar to the ones used in *Costello et al. 2013* [9] (see online methods section for details). **a)** Left: Frequencies of somatic base exchanges calculated in the 13 t/g-paired WES data sets revealed a trend toward overrepresentation of G-to-T transversions. Right: Lego plot of SNV (PopFreq<0.01 or COSMIC-annotated, OxoG corrected) frequencies with trinucleotide context and overall percentages in pie chart. C-to-A and G-to-T transversions still represent the largest portion of (39.8%) exchanges observed in a di-thymidine (T_T) context. **b)** GSOA in 1497 genes harboring mutations in exonic regions (PopFreq<0.01) revealed an overrepresentation of generally cancer-associated pathways. Proportion of genes [%] associated with the respective process in relation to the total number of mutated genes (grey bars) and specific p-values (black bars) are given. **c)** The list of genes recurrently mutated with highest frequencies across all analyzed T-PLL cases is headed by *ATM* and *STAT5B* (only SIFT[10]/PolyPhen2 [11] and PopFreq-filtered mutations included; compare **TableS9** and see **Fig.3c** for a selection of functionally annotated genes). *STAT5B* affected cases were enriched for *ATM* SNVs (n=6/7, 85.71%). **d)** Mutation rates by locus mapped on chr.17 (found to carry most mutations) of case *TP002*, carrying a frameshift deletion mutation in the *MSH3* gene encoding for a DNA mismatch repair factor. To assess for a potential regional mutational heterogeneity due to a dysfunctional mismatch-repair (mutations are no longer enriched in late replicating heterochromatin[12]), we binned somatic mutations (paired; PopFreq<0.01 or COSMIC-annotated) into 1Mbp regions and mapped them to chr.17. We observed a regional mutational heterogeneity pointing to no particular defects within the mismatch-repair system. There were no indications for specific microsatellite instability (only 1/11928 sites somatic by MSIsensor; 0.01%). **e)** Tumor fractions (variant allele fractions, VAFs) of all identified mutations detected in WES data (% positive reads) according to their overall frequencies (Y-axis). The incidence of mutations showing a high clonality (80-100% tumor fraction) was rather low (1.48% of all mutations) pointing to a small number of clonal driver mutations compared to a high number of subclonal passenger SNVs (38.04% of all mutated genes with VAFs ≤10%; see also **Fig.3d,e** and **TableS9** for tumor fractions of specific genes).

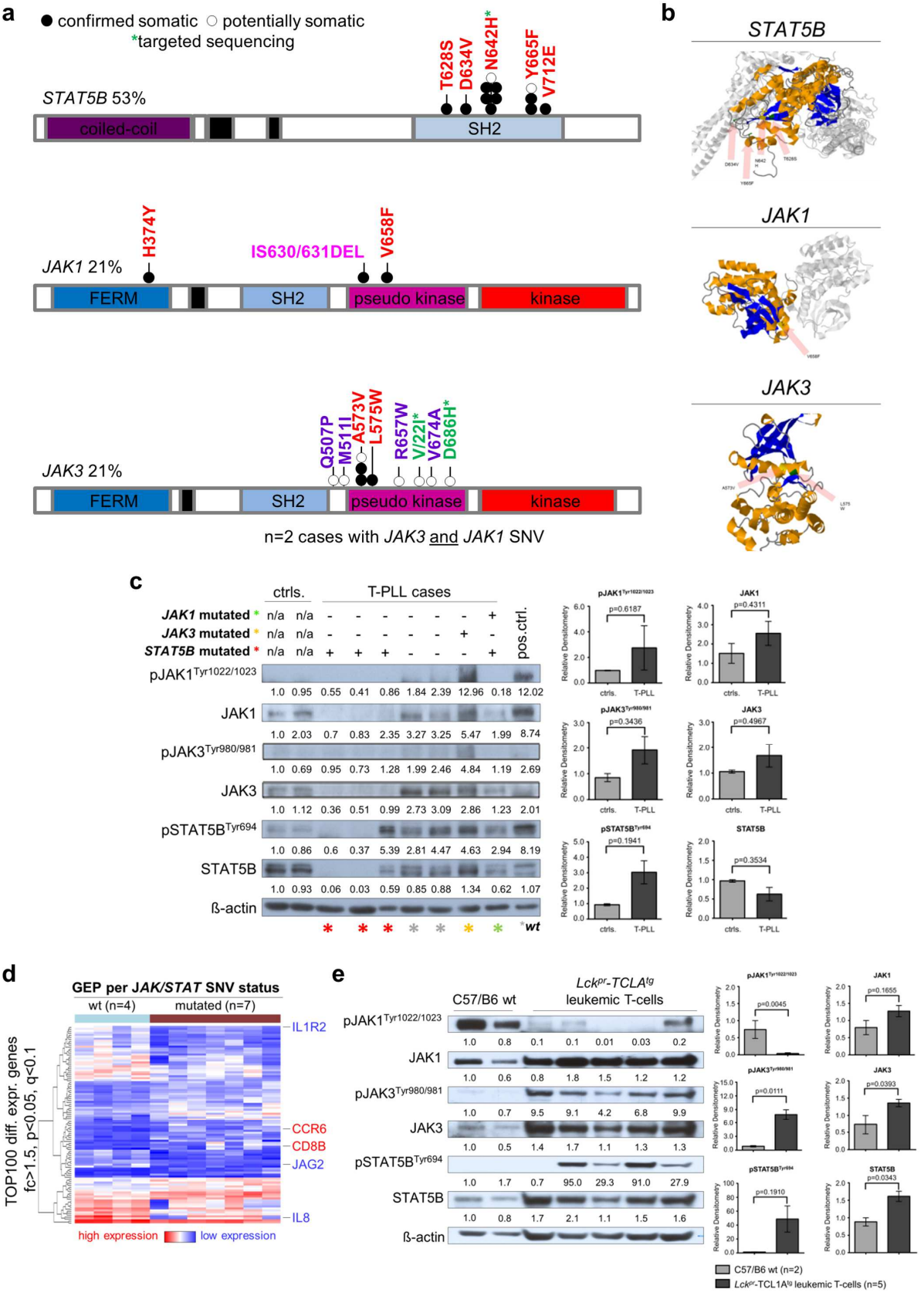


Figure S8: Legend at next page.

Figure S8: The prominent cluster of genomic alterations in *JAK*/*STAT* signaling pathway components confers specific gene expression changes, but respective SNVs do not predict basal JAK/STAT phospho-activation levels.

a) Missense mutations in *STAT5B*, *JAK1*, and *JAK3* genes identified in whole-exome (WES) and targeted amplicon sequencing (TAS) data sets clustered within the SH2 and pseudokinase domains (compare **Fig.3** and **TableS9**). Confirmed somatic: tumor/germline (t/g)-pairs (WES; n=13); potentially somatic: tumor singles (WES: n=20; TAS: n=20). **b)** 3D-molecule structures of *STAT5B*, *JAK1*, and *JAK3* with indicated (red arrow) locations of amino acid (aa) exchanges (via cBioPortal[13]). **c)** Immunoblot analysis showing protein levels with phosphorylation status (activating motifs) of *JAK1*, *JAK3*, and *STAT5B* in primary T-PLL cells (7 cases) with known *STAT5B* / *JAK1* / *JAK3* mutation status. No obvious association of analyzed basal phospho-activation levels with the presence of a respective mutation. Controls: $CD3^+$ pan T-cells isolated from PB of healthy donors (n=2). Lysates from IL2 stimulated HH cells represent positive controls. Quantification: ImageJ®, represented as bar charts, Student's t-test. **d)** Heat map showing the differential expression of genes (TOP100, 178 differentially expressed probes) associated with *STAT5B* / *JAK1* / *JAK3* / *IL2RG* mutations. The comparison included: *STAT5B* / *JAK1* / *JAK3* / *IL2RG* mutated T-PLL (7 cases) vs. 4 T-PLL with wild-type constellation of all of these genes. Differentially expressed genes include e.g. *IL1R2*, *CCR7*, *CD8B*, *JAK2,* and known JAK/STAT target genes like *IL8, MYC, and OAS1*; compare **TableS11** for all differentially expressed genes and **TableS12** for an IPA® analysis showing the functional association of those genes to 'cell death and survival', 'PI3K signaling' and 'interleukin signaling'. **e)** Immunoblots showing protein levels with phospho-activation status of murine JAK1, JAK3, and STAT5B motifs (species-cross reactivity of the antibody) in primary splenic T-cells of Lck^{pr}-$hTCL1A^{+/-}$ mice (overt exponential phase, n=5). Controls: splenic T-cells of genetic-background and age-matched wild-type animals (pools of T-cell isolates from 6 animals). Quantification: ImageJ®, represented as bar charts, Student's t-test.

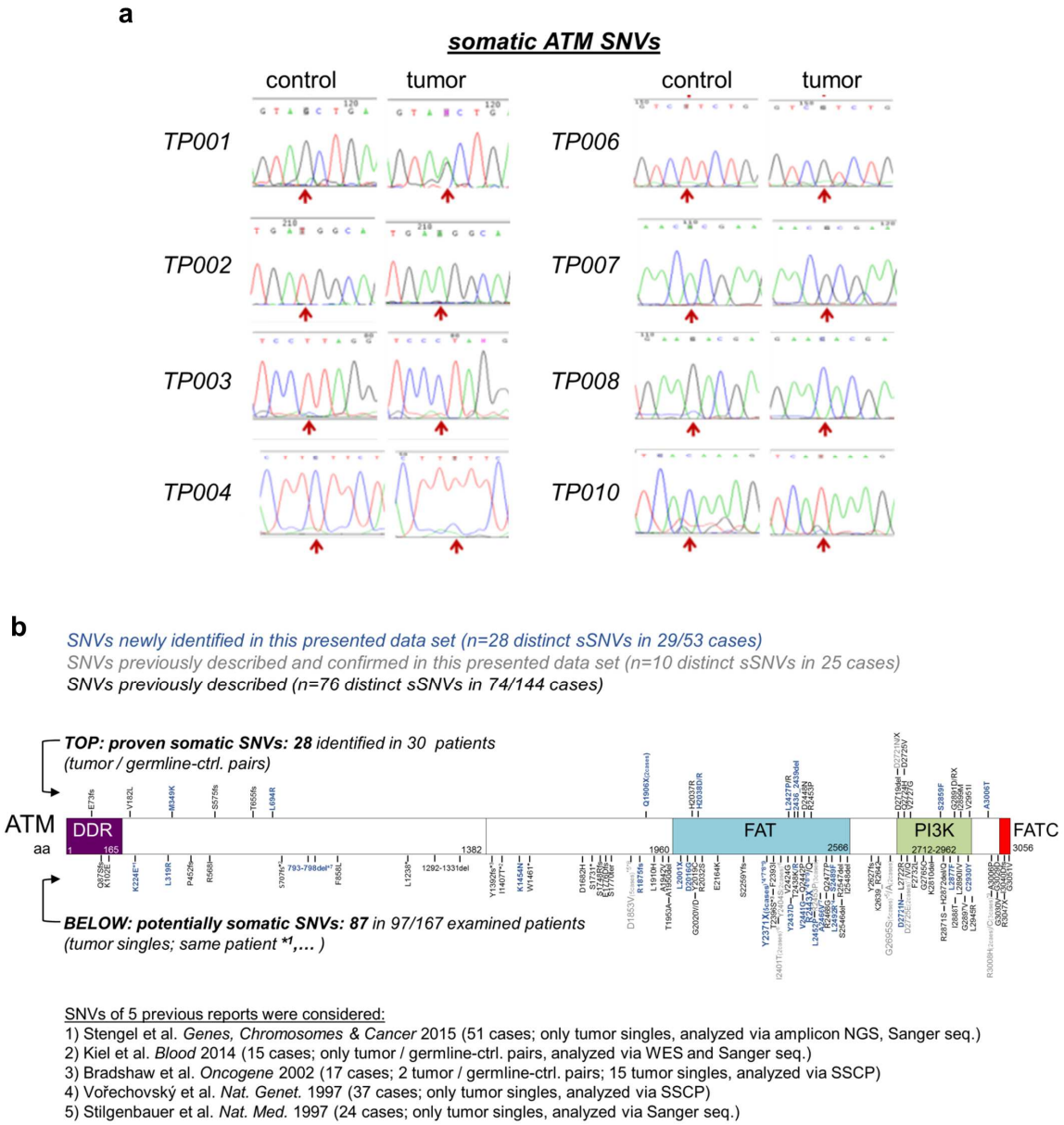


Figure S9: Validations of *ATM* somatic mutations and clustering of *ATM* SNVs in the FAT and PI3K domains.

a) *ATM* mutations detected in tumor/germline (t/g)-pairs by whole-exome sequencing (WES) are validated by Sanger sequencing. Eight distinct SNVs were confirmed as somatic mutations present in 8/9 T-PLL (compare **TableS9**). **b)** Scheme of the ATM molecule with mapping of mutations identified by WES, targeted amplicon sequencing (TAS), and Sanger sequencing (i) according to their description in this series vs previous publications[14–18] (all published data sets carrying sequencing data on *ATM* in T-PLL were selected) and (ii) according to their calling from t/g-pairs (proven somatic, top) vs from tumor singles (potentially somatic, bottom). A clustering in the FAT and PI3K domains is revealed (compare **Fig.3f** for a scheme showing *ATM* SNVs only identified as part of this study) and a dominant missense character of mutations is described (unlike the dominant truncating mutations identified in A.T individuals[19]). Mutations detected in more than one case carry information on the number of respective cases (case numbers in brackets).

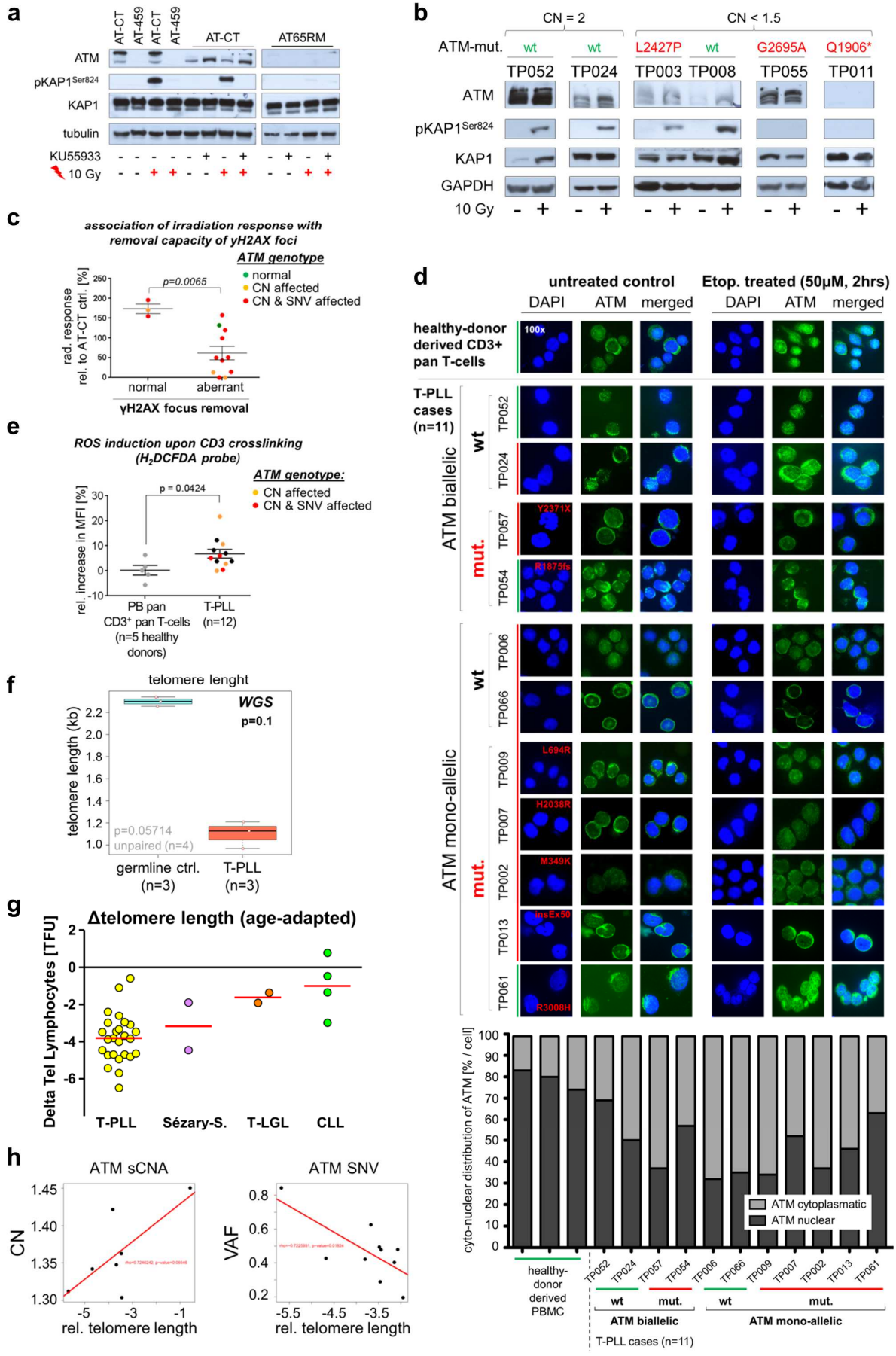


Figure S10: Legend at next page.

Figure S10: ATM in primary T-PLL cells is hypomorphic as per canonical effector functions like yH2AX focus induction, KAP1 and P53 phosphorylation, ATM nuclear translocalization, or telomere maintenance.

a) Control system for activation of the ATM target KAP1 (see **Fig.4d**): lymphoblastoid B-cell lines from A-T patients[20] (AT65RM, $ATM^{\Delta/\Delta}$, c.6573-9G->A/ c.8814_8824del11, ATM protein absent) or unaffected relatives (AT-CT, ATM^{wt}) were pretreated with the ATM kinase inhibitor KU55933 at 50µM for 2hrs. Cells were then exposed to 10Gy ionizing irradiation (IR) and $pKAP1^{Ser824}$ levels were detected 1hr thereafter by Western blot. IR-induced phosphorylation of KAP1 is only detectable in ATM wild-type (wt) cells without KU55933 treatment underlining the specificity of ATM mediated KAP1 phosphorylation. **b)** $KAP1^{Ser824}$ phosphorylation upon 10Gy IR was assessed in primary T-PLL cells of 23 cases. The 6 cases shown serve as informative supplementation to **Fig.4d**. Note that separation of lanes in the presentation of Western blot data was done in order to better assemble cases according to their ATM genotype. Overall, the bulk of cases showed residual pKAP1 induction, despite genomic ATM lesions. T-PLL with ATM in CN-biallelic / SNV-wt constellation usually revealed IR-induced KAP1 phospho-activation, while the rare T-PLL with truncating mutations (Q1906*, no ATM expression, comparable to A-T cells, above) or some few cases with CN monoallelic / ATM mutated status did not (i.e. *TP055*). **c)** There is a correlation of the capacity to phosphorylate KAP1 upon IR with the capacity to induce / remove yH2AX foci following etoposide treatment (see **Fig.4c**). Cases with regular biochemical IR responses show normal yH2AX kinetics. More than half of cases with abnormal yH2AX platform induction / resolution show reduced $pKAP1^{Ser824}$ responses. Quantification of IR response by densitometry of immunoblots: the levels of $pKAP^{Ser824}$ protein relative to pan-KAP1 and housekeeping controls were normalized to induced $pKAP1^{Ser824}$ levels in the AT-CT control cell line (set to 100%). **d)** Subcellular ATM localization in IF microscopy of cytospins of untreated vs etoposide-treated primary T-PLL cells and PBMC controls (supplementary data for **Fig.4e**; here all analyzed cases (top) and quantification (bar chart at bottom). Upper IF panel: Only 3 of 11 cases (green marks) show a predominant nuclear translocalization of ATM upon DSB induction comparable to healthy-donor PBMCs (one representative example of 3 experiments shown). Among cases with regular ATM subcellular kinetics, one harbored an ATM-biallelic / SNV wt constellation, one had an ATM biallelic genotype with a mutation (R1875fs), and one an ATM-monoallelic genotype with a mutation in the FATC domain of ATM (R3008H). The 8 cases without proper ATM translocalization (red marks) harbored heterogeneous, but usually showed affected (7 cases) ATM genotypes. Bottom: The proportion of nuclear ATM in relation to total ATM expression (quantification of fluorescence via ImageJ® software per cell) upon etoposide induced DNA damage is shown as a bar chart (mean of 5 cells). **e)** 2',7'-dichlorodihydrofluorescein diacetate (H_2DCFDA) based measurements of reactive oxygen species (ROS) induction upon T-cell receptor (TCR) activation comparing healthy T-cells (grey dots) to primary T-PLL cases (information on the *ATM* genetic status: orange - CN<1.5, no mutation; red - CN<1.5, mutated; black dots - no genomic ATM status available). Although ROS induction upon CD3/CD28 crosslinking seems to be independent of the

presence of an *ATM* sCNA / SNV, there was a generally higher increase of ROS levels in stimulated T-PLL cells compared to $CD3^+$ pan T-cells isolated from PB of healthy donors. This observation might be linked (1) to a sub-standard performance of the ROS attenuator ATM in T-PLL, (2) to the TCR-sensitizer function of TCL1 [21], (3) to TCL1's effect on mitochondrial ROS generation[22], or (4) to other aberrancies such as inefficient buffer systems. It fits also well with the high abundance of G-to-T transversions observed among all WES-detected SNVs (compare **Fig.S7a**), which can specifically result from ROS induced DNA damage[23]. **f)** Telomere lengths were evaluated according to WGS data using the 'telseq'[24] algorithm. The difference between tumor and germline samples (n=3 paired WGS data sets and the one WGS tumor 'single' included) is of borderline significance (p=0.1, Wilcoxon paired test; p=0.06 unpaired; consider small sample size). **g)** Telomere lengths in 26 primary T-PLL cases (compare **Fig.4f** for an age-adjusted depiction), 4 CLL, 2 T-LGL, and 2 cases of Sézary Syndrome. Measurements were done by flow-FISH and healthy controls were used for age-adaption as described previously[25]; one telomere fluorescence unit (TFU) corresponds to one kilobase pair(s). The data confirm indications of particularly short telomeres in T-PLL in a previous smaller series[26]. **h)** Telomere lengths (flow-FISH) were intuitively associated with ATM lesions (sCNAs and sSNVs) showing shorter telomeres in cases with low *ATM* CNs and high *ATM* VAFs.

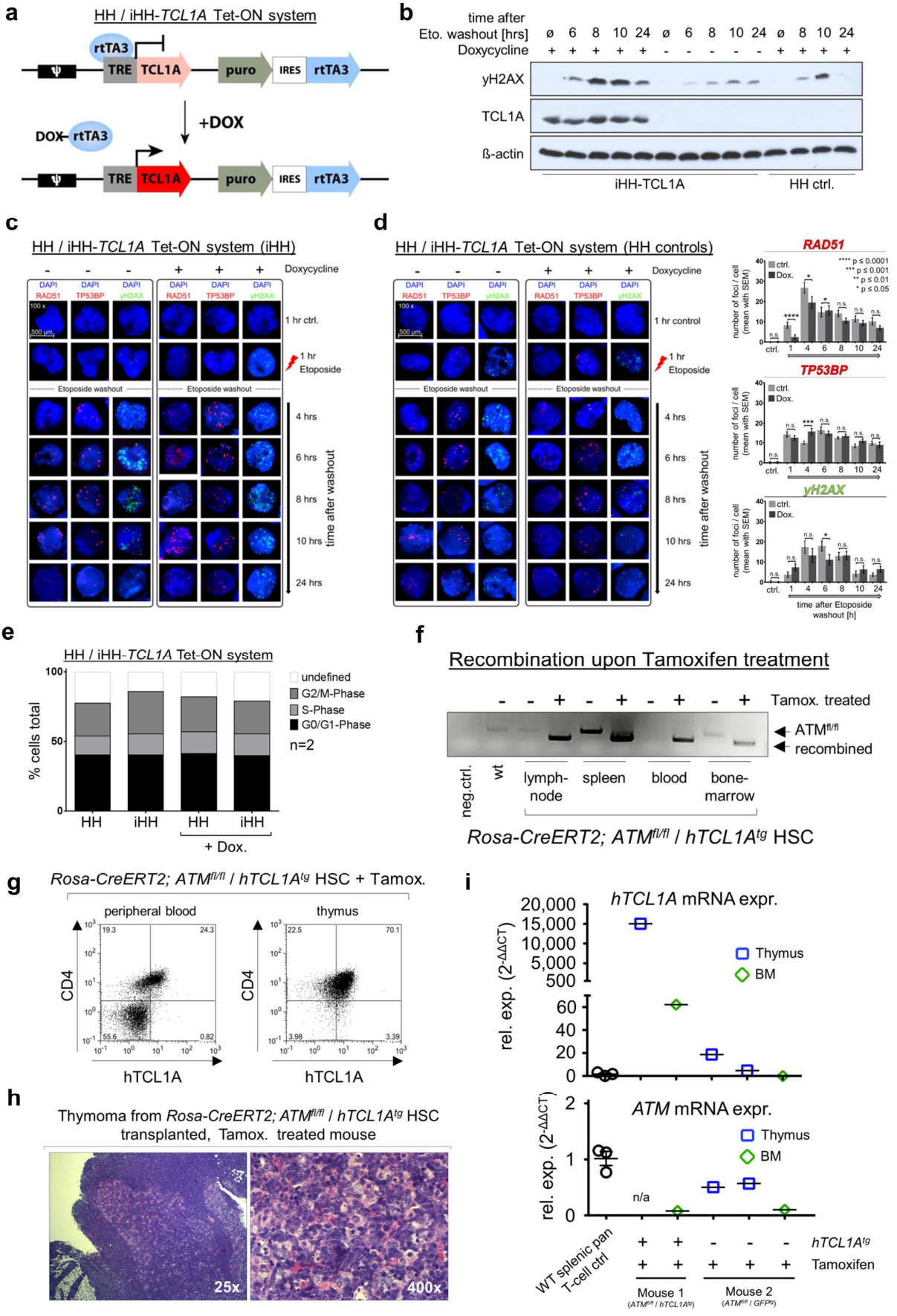


Figure S11: Legend at next page.

Figure S11: Ectopic expression of *TCL1A* affects the DDR and cooperates with *ATM* deficiency towards accelerated T-cell leukemogenesis.

Supplements to **Fig.4g-j**. **a)** Schematic representation of the *TCL1A* expression vector stably transfected in HH mature T-cell leukemia cells (resulting line 'iHH'). TRE: tetracycline responsive element; puromycin: puromycin resistance cassette; IRES: internal ribosomal entry site; rtTA3: reverse tet-transactivator 3. Inducible *TCL1A* expression: upon doxycycline (Dox) treatment, release of the transactivator protein from *TCL1A* promoter binding results in induction of *TCL1A* transcription. **b)** Immunoblots for yH2AX in iHH / HH cells (no *ATM* sCNA, see also DSMZ catalogue #ACC707 for karyotype of HH cells) upon etoposide-induced DSBs (50µM; 1hr) monitored over 24hrs. Doxycycline-induced TCL1A expression enhances yH2AX levels in response to DSBs induction (compare **Fig.4g**, **S11c,d** for parallel time lines of immunofluorescence (IF) microscopy based recordings of yH2AX foci). **c)** IF stainings of cytospins of iHH cells (+/- doxycycline pre-exposure) after DSB induction by etoposide (50µM; 1hr). Ectopic TCL1A expression and its impact on the kinetics of yH2AX, RAD51, and TP53BP1 focus induction / removal: delayed resolution in the presence of TCL1A. Representative images are shown; overall focus quantifications (counts) and representative yH2AX time lines are presented in **Fig.4g**). **d)** As in c) for iHH cells (above), here for the parental HH cells, including doxycycline controls; representative images and focus counts (means, SEM) are shown. In the absence of a transfected TCL1A construct, no difference in focus induction and resolution was detected between the +/- doxycycline conditions. **e)** iHH-TCL1A cells and HH parental controls were treated with doxycycline for 24hrs (1µg/ml). Cell cycle profiles, determined by DNA content assessments using propidium-iodide based flow-cytometry (2 replicates), showed no altered proliferation of TCL1A expressing HH cells, allowing to exclude increased replicative stress as a main cause for the altered DDR (net gain in genomic instability) in the presence of TCL1A.

f) Hematopoetic stem cells (HSCs) of *Rosa-CreERT2;ATM$^{fl/fl}$* mice were retrovirally transduced with *hTCL1A* or a GFP control vector and transplanted into irradiated hosts. Recombination of the *ATM* locus (fl/fl) was induced by tamoxifen (Tamox.) injections 8 weeks after transplantation (1mg/day i.p. injected for 5 consecutive days; see **Fig.4j** for scheme of experimental setup and Kaplan-Meier analysis). Shown are PCR results from animals that were taken out from observation right after the end of tamoxifen injections. Neg. ctrl.: non-template H_20 ctrl.; wt: B6/C57J splenocytes. The shorter PCR product indicates successful recombination at the *Rosa-CreERT2;ATM$^{fl/fl}$* locus. **g)** Evidence of hTCL1A protein expression (flow cytometry) in peripheral blood and thymus of a CD4$^+$/8$^+$ T-cell tumor (221 days post-transplant) from the *ATM$^{fl/fl}$/hTCL1A^{tg}* genotype (**Fig.4j**). **h)** H&E staining of one exemplary thymoma (*ATM$^{fl/fl}$/hTCL1A^{tg}* mouse). **i)** qRT-PCRs of two tumor bearing mice: mouse 1 (*ATM$^{fl/fl}$/hTCL1A^{tg}* Tamox. treated) and mouse 2 (*ATM$^{fl/fl}$/GFP* Tamox. treated). A higher *hTCL1A* mRNA and a lower *ATM* mRNA expression was seen according to the targeted alleles in comparison to WT T-cells. Bone marrow (BM) represents non-tumor bearing hematopoietic tissue and thymus represents tumor tissue of the analyzed diseased *ATM$^{fl/fl}$/hTCL1A^{tg}* and *ATM$^{fl/fl}$/GFP^{tg}* HSC targeted mice. This also speaks to the T-lineage specificity of the leukemogenic TCL1/ATM cooperation

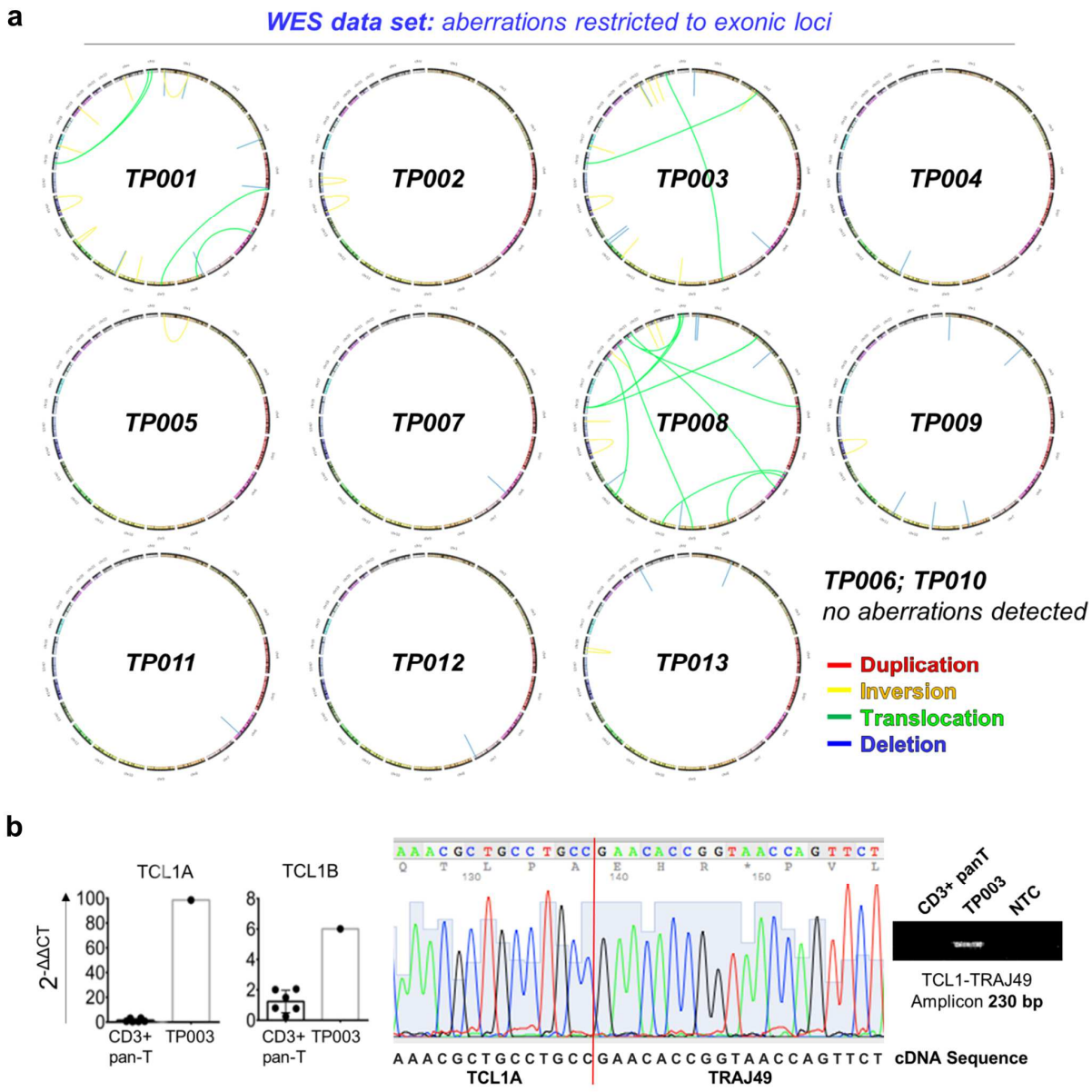


Figure S12: Novel structural variations (SVs) in T-PLL.
a) SVs (color-coded inversions / translocations / deletions) detected in exonic regions are mapped to involved chromosomal loci for all T-PLL tumor/germline-pairs analyzed by WES (data supplementing WGS data of **Fig.5a**, see also **TableS14**). Based on the stringent filters applied, tandem-duplications were not detected and no SVs were detected in *TP006* and *TP010*. **b)** Left: qRT-PCR analysis showing elevated *TCL1A* and *TCL1B* transcript levels in primary T-PLL cells of the *TCL1A-TRAJ49* carrying case *TP003* compared to controls ($CD3^+$ pan T-cells isolated from PB of healthy donors (n=5)). Mid: the fusion transcript was confirmed by Sanger sequencing of cDNA from *TP003* (see **Fig5b** for a schematic representation of the fusion transcript and **Fig.5c,d** for the confirmation of the genomic inv(14) and residual TCL1A protein expression). Right: Validation of the fusion transcript *TCL1A-TRAJ49* expression via RT-PCR in case *TP003* compared to healthy donor derived T-cells (NTC='no template' control).

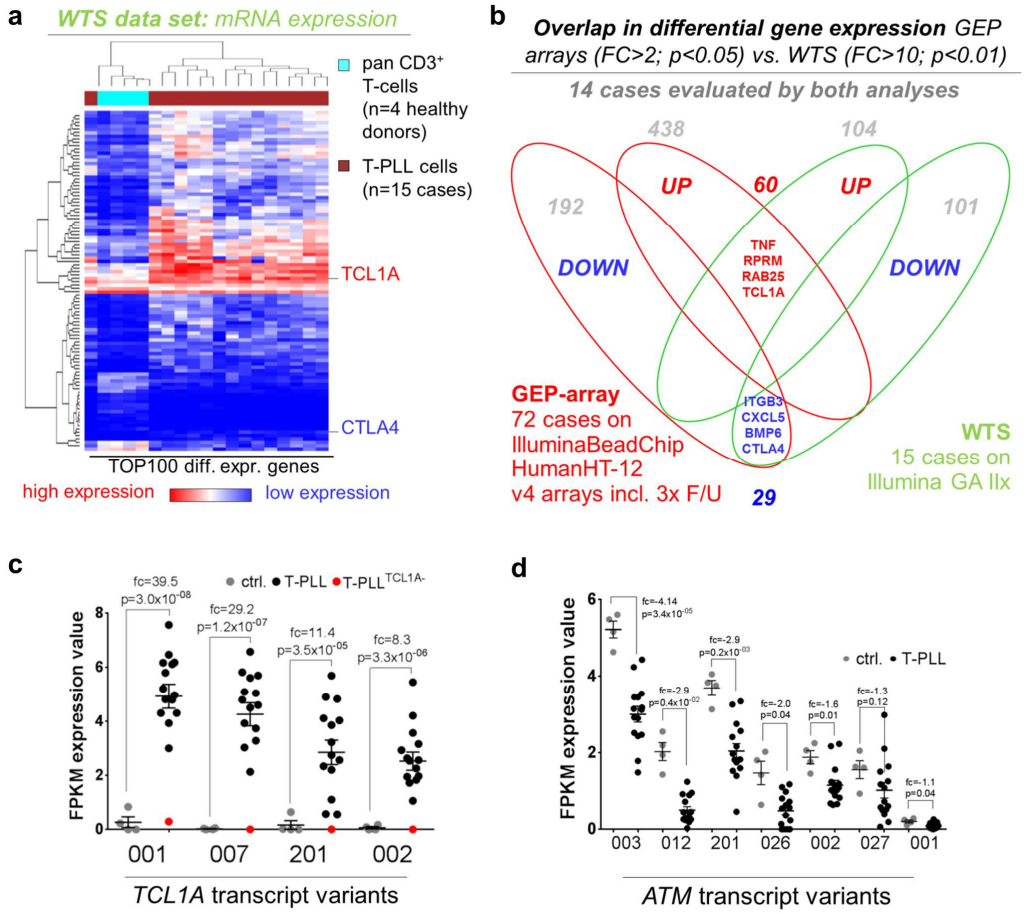


Figure S13: WTS confirms patterns of differential gene expression and identifies transcript variants of *TCL1A* and *ATM*.

a) The TOP100 most variably expressed transcripts, based on the comparison of WTS data from 15 T-PLL to those from CD3$^+$ pan T-cells isolated from PB of healthy donors (n=4) are represented in the heat map (compare **TableS16**). **b)** Overlap of significantly differentially expressed genes in T-PLL cells as detected by WTS data (15 T-PLL) vs GEP arrays (n=70 cases); see **TableS16** for further information. **c)** Differential expression of variant *TCL1A* transcripts in primary T-PLL (n=15) compared to healthy-donor derived CD3$^+$ T-cells ('ctrl.', n=4) revealed a congruent upregulation of all detected *TCL1A* transcripts in '*TCL1A* positive cases' (TCL1A-protein negative case as red dots) and identifies the high expression of a new shorter *TCL1A* variant (*TCL1A-007*). FPKM: fragments per kilobase of exon per million reads mapped. Generally, differential expression of transcripts was assessed using DESeq v1.14.0 by evaluating the expression of respective isoforms through a gapped alignment. In contrast to that, differential exon usage (DEU) as alternative splicing (compare **Fig.5e**), evaluated via DEXSeq v1.8.0, gives a descriptive assessment on whether the particular exon bins (containing merged exons for ORF overlaps of multiple genes) are rather retained or skipped. Here, effects of differential expression were excluded. **d)** Differential expression of variant *ATM* transcripts (PCR) in T-PLL (n=15) compared to healthy-donor CD3$^+$ T-cells ('ctrl.', n=4) confirmed downregulation of 5/7 protein coding *ATM* variants in T-PLL; those not differentially expressed are expressed at generally low levels in both, ctrl. and T-PLL.

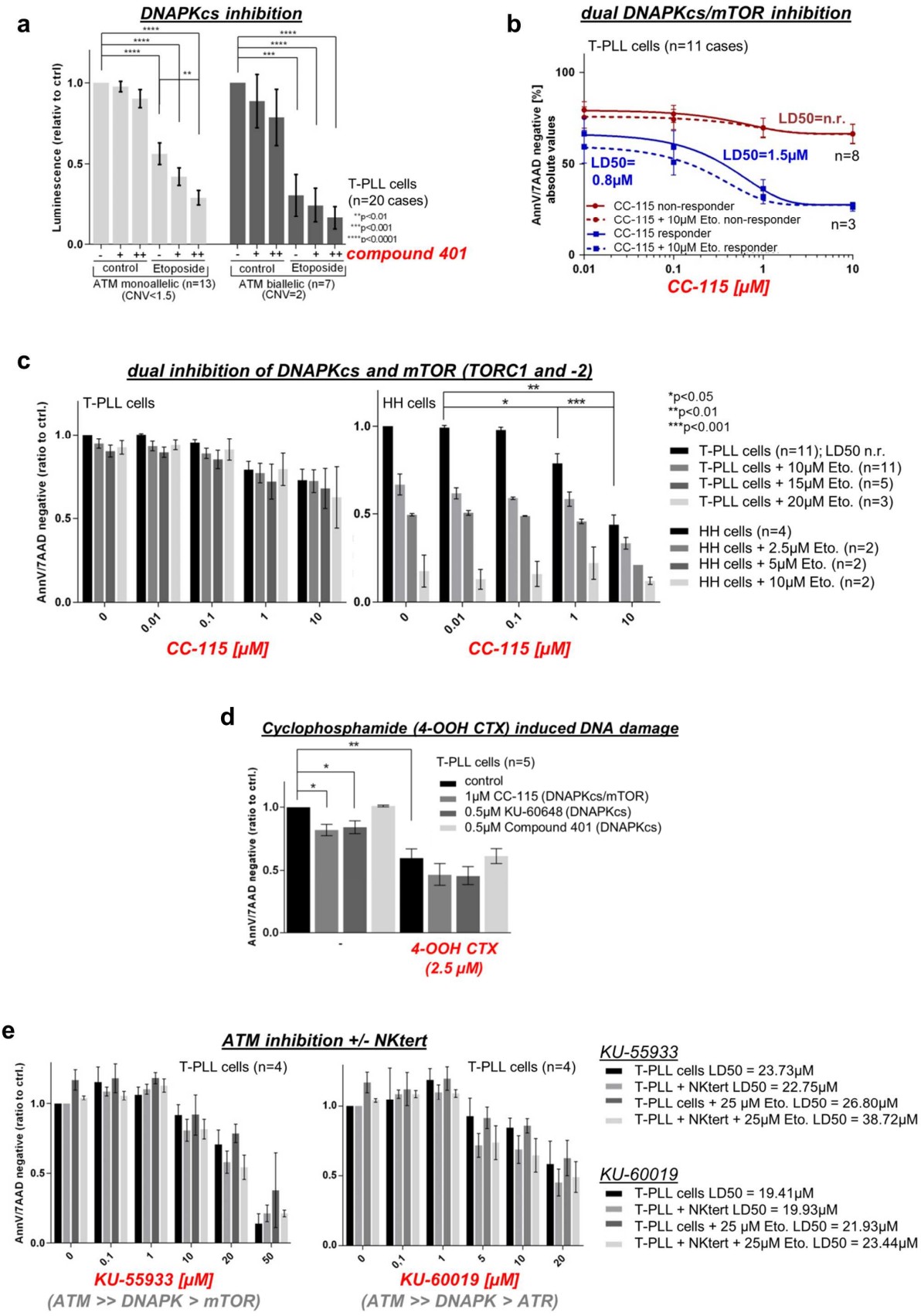
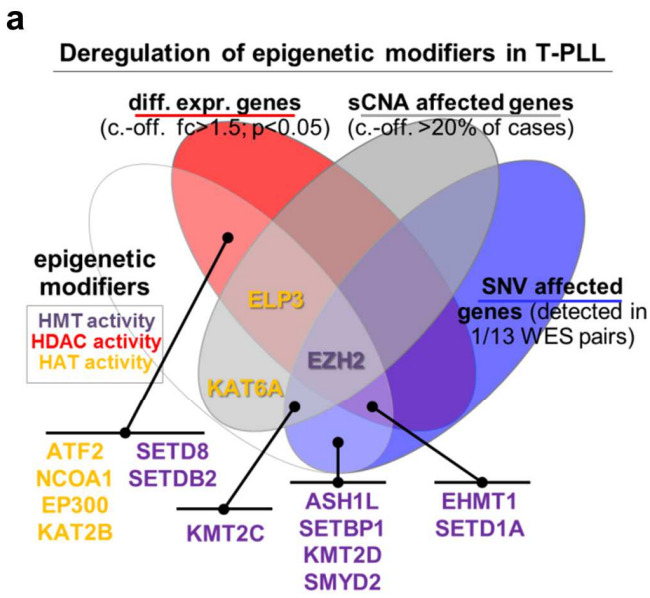


Figure S14: Legend at next page.

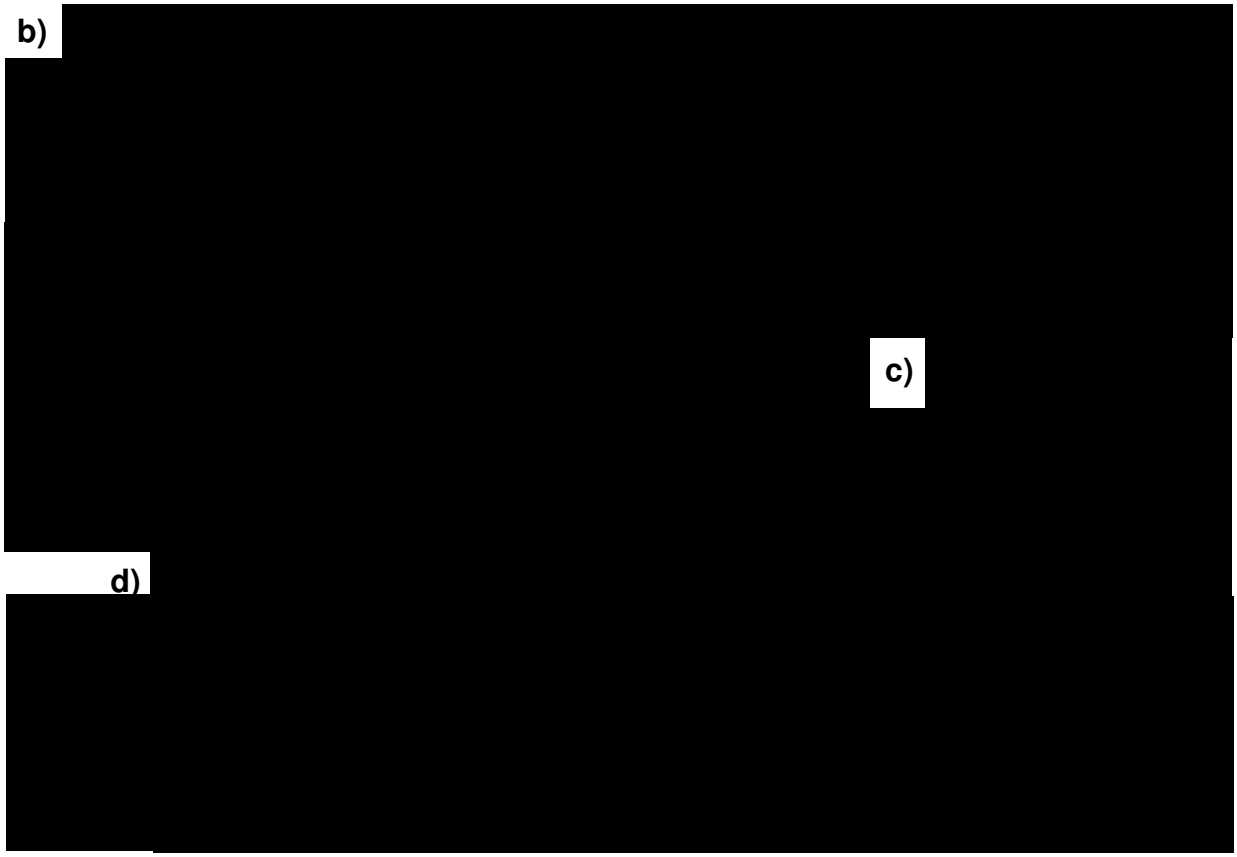
Figure S14: Targeting of factors in potentially synthetic lethal relationships to *ATM* does not affect T-PLL cell viability in the context of DNA damage.
a) Primary T-PLL cells (suspension cultures if not indicated otherwise) from 20 cases were treated *in vitro* with the DNAPKcs inhibitor Compound 401 ('+' 0.25µM and '++' 0.5µM) in the context of etoposide-induced DNA damage (25µM, 96hrs). Cell viability measured as per LumiGlo® assay. Cases were grouped according to their *ATM* genotype in '*ATM* mono-allelic' (CN<1.5, n=13) and '*ATM* bi-allelic' (CN=2.0, n=7). Treatment with the DNAPKcs inhibitor alone showed no reduction of cell viability irrespective of the *ATM* genotype. **b, c)** Primary T-PLL cells (b; 11 cases) as well as HH cells (c; 2 experiments) were treated *in vitro* for 48hrs with the dual DNAPKcs/mTOR inhibitor CC-115 at increasing concentrations in the context of etoposide-induced DNA damage. Apoptotic responses were assessed using AnxV/7AAD staining. **b)** Dose-response curves (LD50; absolute percentages of living cells by AnxV/7AAD flow cytometry). Only a subset of T-PLL (3 responders of 11 cases) showed a dose-related selective cellular sensitivity towards CC-115 in the high nano- / low micro-molar range (LD50 1.5µM), however, which was not much affected by etoposide treatment (LD50 0.8µM; dashed). The distinct response profiles could not be explained by molecular genetic events like *ATM* sCNAs or SNVs. Note the slightly reduced basal 'fitness' of the responders. **c)** Proportions of AnxV/7AAD negative cells (ratio to control) are shown (means, SEM). Left: The minimal activity of CC-115 across all 11 T-PLL reflected the low proportion of cases achieving an LD50 (see b; 36.4%). There was no complete eradication of viable T-PLL cells even at high CC-115 dosages (10µM) combined with high etoposide concentrations (20µM). Right: Although being more sensitive to etoposide treatment in general, HH cells show only minor responses to high doses of (10µM) CC-115 treatment. **d)** DNA damage induction via cyclophosphamide (active metabolite 4-OOH CTX) instead of etoposide: primary T-PLL cultures (n=5 cases) were treated with 2.5µM 4-OOH CTX and apoptotic responses to dual DNAPK/mTOR inhibition (CC-115) and DNAPKcs inhibition (Compound 401, KU-60648) were evaluated by AnxV/7AAD staining. Proportions of AnxV/7AAD negative cells (ratio to control) are shown (means, SEM). The ineffective killing of primary T-PLL cells via DNAPKcs and DNAPKcs/mTOR inhibition in the context of 4-OOH-CTX-induced DNA damage confirmed the low activity of this synthetic lethal approach for T-PLL and excludes a potential etoposide-restricted effect. **e)** Primary T-PLL cells (n=4 cases) in co-cultures with the human bone marrow stromal cell line NKtert were exposed to increasing concentrations of the ATM inhibitors KU-55933 (0.1-50µM), and KU-60019 (0.1-20µM) in the context of etoposide-induced DNA damage (25µM) for 48hrs and cell death was quantified by AnxV/7AAD staining (means, SEM). ATM inhibition did not significantly synergize with etoposide-induced DNA damage. Killing of T-PLL cells was induced only at high inhibitor concentrations at which the (weak) protective effect of NKtert co-cultures is no longer observed.

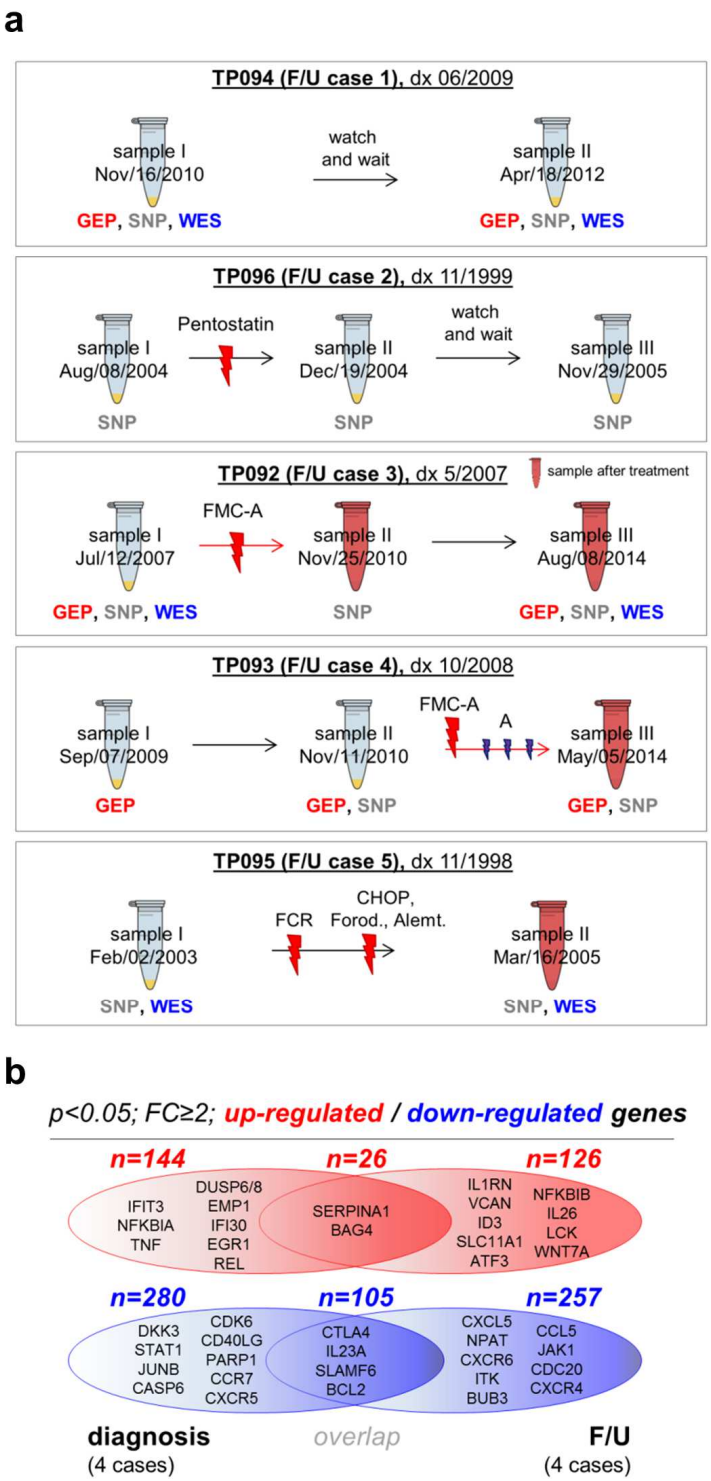


[redacted]

Figure S15: [redacted]

a) Summary of aberrations in epigenetic modifiers (n=77 genes, **TableS18**) called in primary T-PLL cells by profiling of: GEP, sCNA, and SNVs (frequency cut-offs indicated).

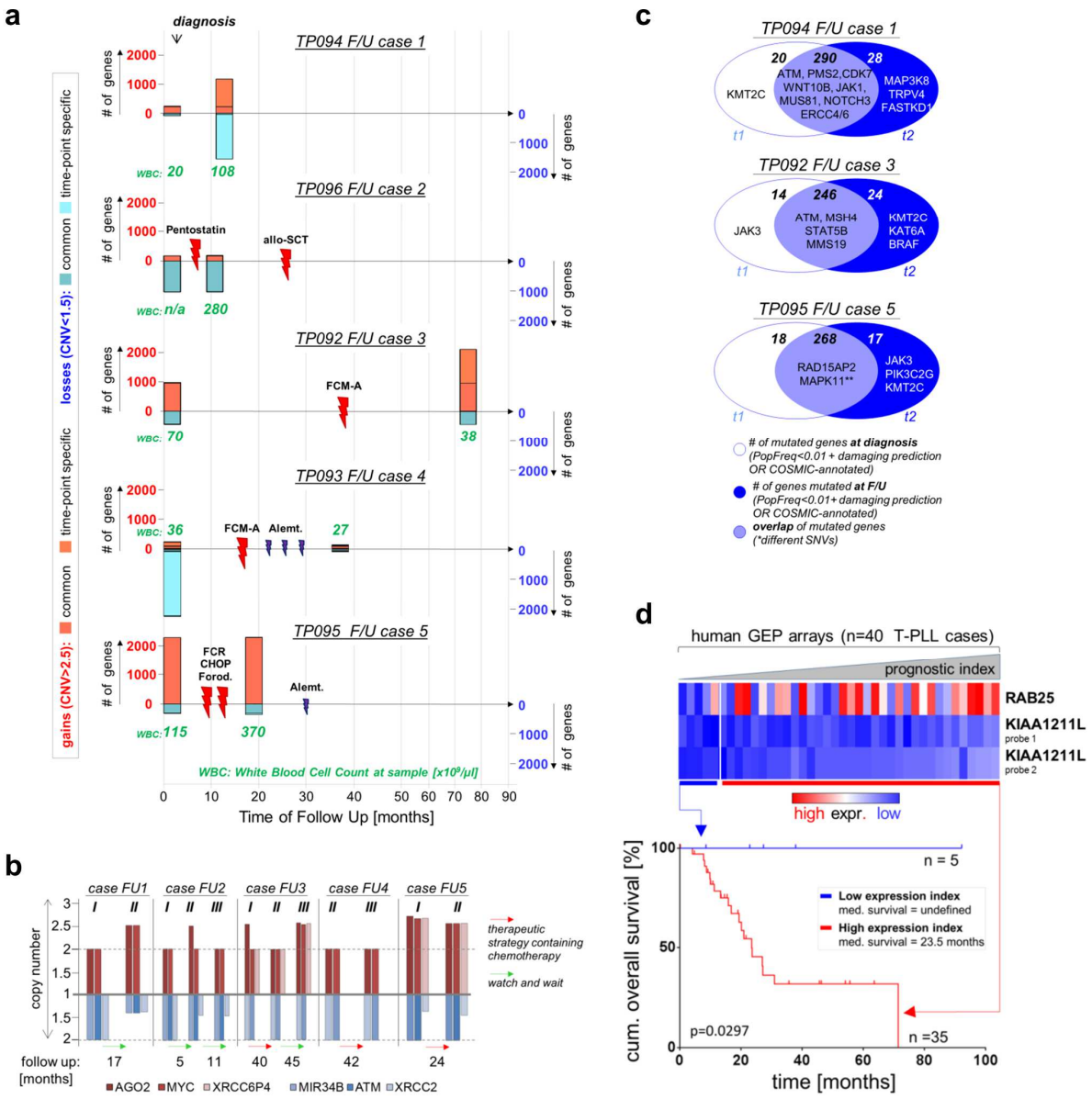

b)
c)
d)



561

562 **Figure S16:** Legend at next page.

Figure S16: **General data of T-PLL cases with available sequential follow-up (F/U) samples and the analysis for evolution of transcriptomic changes.**
a) Among the total n=94 T-PLL cases analyzed, sequential samples were available for n=5 cases with sufficiently long F/U (13 samples, see **Fig.7** for leukocyte counts and further results). The median total F/U time for all cases was 24 months (ranging from 16 to 85) and the median of sample intervals was 20.5 months. The first samples, close to initial diagnosis (treatment naïve) were followed by those after clinically relevant progression or relapse after therapy. These samples were analyzed by at least one of the profiling approaches: GEP, SNP-arrays (for sCNAs), and WES. For F/U case 1, one second sample was collected after 17 months. In F/U case 2, within 16 months 3 samples were collected and analyzed via sCNA profiling. In F/U case 3, 3 sequential samples were collected over a long course of 95 months and subjected to GEP, sCNA profiling, and WES. This patient received an FMC-A chemo-immunotherapy (fludarabine, mitoxantrone, cyclophosphamide; followed by alemtuzumab) between 1^{st} and 2^{nd} sampling. F/U case 4: over 56 months, 3 samples were collected and analyzed via GEP and sCNA profiling. F/U case 5: 2 sequential samples within 24 months. This patient was heavily treated in-between with distinct chemo-immunotherapies: FCR (fludarabine, cyclophosphamide, rituximab), CHOP (cyclophosphamide, doxorubicine, vincristine, and prednisone), forodesine, and single-agent alemtuzumab. Here, sCNA profiling and WES were performed.
b) GEP of 4 cases with available t_1/t_2-pairs. Differential expression calculated separately for each time point (vs healthy-donor T-cells). Selection from lists of differentially up- (red) and down-regulated (blue) genes at t_1, t_2, or with overlap (**TableS19**). The majority of transcripts was specifically restricted to either t_1 or t_2.



588

589 **Figure S17: Changes in sCNAs and sSNVs during evolution of T-PLL through**
590 **progression or relapse and creation of a prognostic gene expression index.**
591 **a**) Total numbers of genes affected by sCNAs (gains=red / losses=blue) plotted for 5
592 T-PLL follow up (F/U) pairs. Treatments and leukocyte counts at sampling are indi-
593 cated (also **TableS2, S20**). **b**) Five cases (*FU1-5*) with available SNP-array based
594 F/U data were analyzed for time-resolved alterations of sCNAs (see also **TableS20**
595 for all identified lesions). Somatic CNAs of selected genes (*AGO2, MYC, XRCC6P4,*
596 *MIR34B, ATM*, and *XRCC2*) are depicted. Data sets (I-III) are ordered according to
597 sampling, and the time intervals (months) are given underneath. Red arrows indicate
598 chemotherapeutic treatment; green arrows correspond to an attentive strategy in-
599 between the samplings. We observe distinct scenarios of sCNA kinetics: respective
600 gains and losses can be present from the outset (*AGO2* in F/U case 1; *AGO2*, *MYC*,
601 *XRCC6P4* and *XRCC2* in F/U case 2) or be acquired at later time points during dis-
602 ease progression (*AGO2*, *MYC*, *MIR34B*, *ATM* and *XRCC2* in F/U case 3; *XRCC2* in
603 F/U case 5).

c) WES (3 pairs) at diagnosis (treatment-naïve, t_1) and relapse/progression (t_2). Numbers of genes mutated specifically at t_1, at t_2, and at both are indicated (exonic; PopFreq.<0.01; predicted to be damaging or COSMIC-annotated); specific examples from **TableS21**. **MAPK11: VAF rose from t_1 (0.28) to t_2 (0.76) in *TP095*.
d) Differential clinical outcome prognosticated by a 2-gene/3-probe gene expression index at the time of diagnosis. Note that T-PLL is a disease with a generally short survival, but with recognition of rare indolent phases. Top: mRNA levels of *RAB25* and both *KIAA1211L* probes (*RAB25* or *KIAA1211L* alone are of insufficient power) as the 2 signature genes filtered through regression from the learning-set of T-PLL subjects (**Online Supplements**). Below: Kaplan-Meier curves as application of the stratified index in the test cohort discriminating the overall survival outcome based on low vs high index values. The oncogenic RAS GTPase *RAB25* was part of the TOP100 T-PLL signature (**Fig.1, TableS3**) providing normal-T vs tumor-cell distinction.

REFERENCES (for Supplementary Figures only)

1. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–50 (2005).
2. Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34,** 267–73 (2003).
3. Dürig, J. *et al.* Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia* **21,** 2153–63 (2007).
4. Schlosser, I. *et al.* Dissection of transcriptional programmes in response to serum and c-Myc in a human B-cell line. *Oncogene* **24,** 520–4 (2005).
5. Rashi-Elkeles, S. *et al.* Parallel induction of ATM-dependent pro- and antiapoptotic signals in response to ionizing radiation in murine lymphoid tissue. *Oncogene* **25,** 1584–92 (2006).
6. Jackson-Grusby, L. *et al.* Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat. Genet.* **27,** 31–9 (2001).
7. Saitou, M., Sugimoto, J., Hatakeyama, T., Russo, G. & Isobe, M. Identification of the TCL6 genes within the breakpoint cluster region on chromosome 14q32 in T-cell leukemia. *Oncogene* **19,** 2796–802 (2000).
8. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
9. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).
10. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).
11. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).
12. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–4 (2015).
13. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** pl1 (2013).
14. Kiel, M. J. *et al.* Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. *Blood* **124(9),** 1460–72 (2014).
15. Bradshaw, P., Condie, A. & Matutes, E. Breakpoints in the ataxia telangiectasia gene arise at the RGYW somatic hypermutation motif. *Oncogene* **58,** 483–487 (2002).
16. Vořechovský, I. *et al.* Clustering of missense mutations in the ataxia-telangiectasia gene in a sporadic T-cell leukaemia. *Nat. Genet.* **17,** 96–99 (1997).
17. Stilgenbauer, S. *et al.* Biallelic mutations in the ATM gene in T-prolymphocytic leukemia. *Nat. Med.* **3,** 1155–9 (1997).
18. Stengel, A. *et al.* Genetic characterization of T-PLL reveals two major biologic subgroups and JAK3 mutations as prognostic marker. *Genes Chromosom. Cancer* **55,** 82–94 (2016).
19. Sandoval, N. *et al.* Characterization of ATM gene mutations in 66 ataxia

telangiectasia families. *Hum. Mol. Genet.* **8,** 69–79 (1999).

20. Delia, D. *et al.* ATM protein and p53-serine 15 phosphorylation in ataxia-telangiectasia (AT) patients and at heterozygotes. *Br. J. Cancer* **82,** 1938–45 (2000).

21. Herling, M. *et al.* High TCL1 expression and intact T-cell receptor signaling define a hyperproliferative subset of T-cell prolymphocytic leukemia. *Blood* **111,** 328–337 (2008).

22. Prinz, C. *et al.* Organometallic nucleosides induce non-classical leukemic cell death that is mitochondrial-ROS dependent and facilitated by TCL1-oncogene burden. *Mol. Cancer* **14,** 114 (2015).

23. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19,** 169–85 (2004).

24. Ding, Z., Mangino, M., Aviv, A., Spector, T. & Durbin, R. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42,** e75 (2014).

25. Weidner, C. I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* **15,** R24 (2014).

26. Röth, A. *et al.* Short telomeres and high telomerase activity in T-cell prolymphocytic leukemia. *Leukemia* **21,** 2456–62 (2007).

SUPPLEMENTARY TABLES

CONTENTS

SUPPLEMENTARY TABLES

Supplementary Table 1: Profiling data in larger cohorts of T-PLL.

We summarize here published studies that presented immunophenotypic, cytogenetic, genomic or transcriptomic data sets on sizable cohorts of T-PLL. Earlier studies, mostly based on clinical and flow-cytometric analyses revealed the non-descript T-cell immunophenotype of T-PLL, its dominant involvement of *TCL1A* affecting cytogenetic lesions, and the loss of *ATM* by Karyotype G-banding, FISH, and microsatellite typing[1–5]. In recent years, smaller series on gene expression profiling (GEP)[6], copy-number (CN) screens[6], targeted amplicon[7–9] and whole exome[10] sequencing (TAS, WES) provided isolated first fragments of genome-wide analyses.

1st Author[ref] year	# of cases	Methods	Main findings / comments
Matutes[5] 1991	78	Flow cytometry, Karyotype G-banding	IP: 65% $CD4^+$ $CD8^-$, 21% $CD4^+$ and $CD8^+$, 13% $CD4^-$ $CD8^+$; genomic abnormalities: inv(14) with breakpoints at 14q11 and 14q32 in 76% of cases, trisomy 8 in 53% of cases
Stilgenbauer[2] 1997	24	Karyotype G-banding, FISH, Sanger seq.	identification of a small commonly deleted segment at 11q22.3-23.1 (*ATM*) in 63% with mutations on the remaining allele in 25% of cases
Stoppa-Lyonnet[1] 1998	15*	LOH by microsatellite typing	inactivation of the *ATM* gene in 67% of cases through LOH
Hetet[3] 2000	21*	LOH by microsatellite typing	loss of heterozygosity of the 12p13 region, including the *ETV6* and *CDKN1B* genes in 43% of cases
Soulier[11] 2001	22	Array CGH	complex pattern of recurrent chromosomal losses and gains at e.g. 8p (86% of cases), 11q (68%), 22q11 (45%), 13q (41%), 8q (82%), 14q32 (50%)
Bradshaw[12] 2002	17	Cloning breakpoints within the *ATM* gene, Southern blot	identification of breakpoints within the *ATM* gene at the RGYW somatic hypermutation motif in 18% of cases
Dürig[6] 2007	5	GEP, SNP-arrays	differentially expressed genes enriched in genomic regions affected by recurrent chromosomal lesions (6p, 8q 6q, 8p, 10p, 11q, and 18p)
Herling[4] 2008	86	Flow cytometry and Karyotype G-banding	IP: 62% $CD4^+$ $CD8^-$, 35% $CD4^+$ and $CD8^+$, 4% $CD4^-$ $CD8^+$; genomic abnormalities: inv(14)(q11;q32.1) or t(14;14) in 40%, trisomy 8 in 35%, -11 or deletion 11q22-23 in 33%, and -17 or isochromosome 17q or deletion 17p in 13% of cases
Le Toriellec[13] 2008	47	Microsatellite typing, Sanger seq.	haploinsufficiency of *CDKN1B* in 43% of cases (partially based on data from *Soulier et al. 2001*)
Bug[14] 2009	12	Karyotype G-banding, GEP, SNP array, FISH	recurrent loss, but lack of mutations, of the *SMARCB1* tumor suppressor gene in 33% of cases
Delgado[15] 2012	-	Review, meta-data	update on molecular and cytogenetic abnormalities
Bellanger[7] 2014	45	Sanger seq.	recurrent *JAK1*/*JAK3* somatic mutations in 49% of cases
Bergmann[8] 2014	32	FISH, Sanger seq.	mutations of *JAK3* in 30% of cases
His[16] 2014	25	Karyotype G-banding, FISH	frequent *TCL1A* rearrangements (75% of cases), losses of *ATM* (64%), and gains of *MYC* (67%)
Kiel[10] 2014	50	WGS, WES, SNP-arrays, Sanger seq.	mutations affecting *EZH2*, *FBXW10,* and *CHEK2*; dominant: JAK/STAT pathway component in 76% of cases
Stengel[9] 2015	51	Karyotype G-banding, FISH, array CGH, amplicon NGS, Sanger seq.	deletions of *ATM* (69% of cases) and TP53 (31%); mutations in *ATM* (73%), *TP53* (14%), *JAK1* (6%), *JAK3* (21%)
López[17] 2016	43	Targeted seq. of *JAK/STAT* genes via Sanger seq.; additional 54-gene panel (recurrently mutated in hematological cancers) by amplicon NGS	activating mutations in *JAK3* (30%) and *STAT5B* (21%) in evaluated hot-spot regions, mutations in genes encoding for epigenetic regulators (*EZH2* 13%; *TET2* 17%; *BCOR* 9%)

Summary on profiling studies in T-PLL. *paired tumor germline samples; IP – Immunophenotype; LOH – loss of heterozygosity, CGH – comparative genomic hybridization, GEP – gene expression profiling, SNP – single-nucleotide polymorphism, FISH – fluorescence in situ hybridization, NGS – next-generation sequencing, WES – whole-exome sequencing, WGS – whole-genome sequencing

List of separate Excel files: Tables S2-S22.

Supplementary Table 2:	Cohort of analyzed T-PLL cases.
Supplementary Table 3:	Gene expression profiling: Differentially expressed genes comparing T-PLL cases to healthy-donor derived T-cells.
Supplementary Table 4:	Gene expression profiling: Differentially expressed genes comparing murine T-PLL like T-cell expansions to healthy wild-type control derived splenic T-cells and overlap of significantly expressed genes to human T-PLL.
Supplementary Table 5:	GISTIC2.0 detected genomic regions significantly affected by sCNAs.
Supplementary Table 6:	sCNA affected genes per patient and sCNA frequencies based on T-PLL patient derived controls and HapMap controls.
Supplementary Table 7:	Differentially expressed genes associated with chr.11 loss and chr.8 gain lesions.
Supplementary Table 8:	Differentially expressed genes associated with low *ATM* and high *AGO2* expression including overlaps of CN lesion associated genes.
Supplementary Table 9:	WES and WGS data sets: Mutated genes.
Supplementary Table 10:	Highly clonal mutations: Genes mutated with a VAF >80%.
Supplementary Table 11:	*JAK/STAT* mutation signature in T-PLL: Differentially expressed genes according to *JAK/STAT* mutation status.
Supplementary Table 12:	IPA of JAK/*STAT* mutation signature.
Supplementary Table 13:	Combined analysis of sCNA and SNV data: Genes affected by gain of function / loss of function alterations.
Supplementary Table 14:	Structural variations detected by WGS and WES.
Supplementary Table 15:	Fusion transcripts detected by WTS analyses.
Supplementary Table 16:	Differentially expressed genes according WTS data set and overlaps with GEP-array based analyses.
Supplementary Table 17:	Differentially used exons according to WTS.
Supplementary Table 18:	Analyzed genes encoding epigenetic modifiers.
Supplementary Table 19:	Differentially expressed genes in sequential T-PLL samples.
Supplementary Table 20:	sCNAs identified in sequential T-PLL samples.
Supplementary Table 21:	SNVs identified in sequential T-PLL samples.
Supplementary Table 22:	Analyzed genes encoding DDR factors.
Supplementary Table 23:	Oligonucleotides.

REFERENCES (for Table S1 only)

1. Stoppa-Lyonnet, D. *et al.* Inactivation of the ATM gene in T-cell prolymphocytic leukemias. *Blood* **91,** 3920–6 (1998).
2. Stilgenbauer, S. *et al.* Biallelic mutations in the ATM gene in T-prolymphocytic leukemia. *Nat. Med.* **3,** 1155–9 (1997).
3. Hetet, G. *et al.* Recurrent molecular deletion of the 12p13 region, centromeric to ETV6/TEL, in T-cell prolymphocytic leukemia. *Hematol. J.* **1,** 42–7 (2000).
4. Herling, M. *et al.* High TCL1 expression and intact T-cell receptor signaling define a hyperproliferative subset of T-cell prolymphocytic leukemia. *Blood* **111,** 328–337 (2008).
5. Matutes, E. *et al.* Clinical and laboratory features of 78 cases of T-prolymphocytic leukemia. *Blood* **78,** 3269–74 (1991).
6. Dürig, J. *et al.* Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia* **21,** 2153–63 (2007).
7. Bellanger, D. *et al.* Recurrent JAK1 and JAK3 somatic mutations in T-cell prolymphocytic leukemia. *Leukemia* **28,** 417–9 (2014).
8. Bergmann, A. K. *et al.* Recurrent mutation of JAK3 in T-cell prolymphocytic leukemia. *Genes. Chromosomes Cancer* **53,** 309–16 (2014).
9. Stengel, A. *et al.* Genetic characterization of T-PLL reveals two major biologic subgroups and JAK3 mutations as prognostic marker. *Genes Chromosom. Cancer* **55,** 82–94 (2016).
10. Kiel, M. J. *et al.* Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. *Blood* **124(9),** 1460–72 (2014).
11. Soulier, J. *et al.* A complex pattern of recurrent chromosomal losses and gains in T-cell prolymphocytic leukemia. *Genes. Chromosomes Cancer* **31,** 248–254 (2001).
12. Bradshaw, P., Condie, A. & Matutes, E. Breakpoints in the ataxia telangiectasia gene arise at the RGYW somatic hypermutation motif. *Oncogene* **58,** 483–487 (2002).
13. Le Toriellec, E. *et al.* Haploinsufficiency of CDKN1B contributes to leukemogenesis in T-cell prolymphocytic leukemia. *Blood* **111,** 2321–2328 (2008).
14. Bug, S. *et al.* Recurrent loss, but lack of mutations, of the SMARCB1 tumor suppressor gene in T-cell prolymphocytic leukemia with TCL1A-TCRAD juxtaposition. *Cancer Genet. Cytogenet.* **192,** 44–7 (2009).
15. Delgado, P., Starshak, P., Rao, N. & Tirado, C. A. A Comprehensive Update on Molecular and Cytogenetic Abnormalities in T-cell Prolymphocytic Leukemia (T-pll). *J. Assoc. Genet. Technol.* **38,** 193–8 (2012).
16. Hsi, A. C. *et al.* T-cell prolymphocytic leukemia frequently shows cutaneous involvement and is associated with gains of MYC, loss of ATM, and TCL1A rearrangement. *Am. J. Surg. Pathol.* **38,** 1468–83 (2014).
17. Lopez, C. *et al.* Genes encoding members of the JAK-STAT pathway or epigenetic regulators are recurrently mutated in T-cell prolymphocytic leukaemia. *Br. J. Haematol.* **173,** 265–273 (2016).

SUPPLEMENTARY METHODS

CONTENTS

MATERIAL AND METHODS

1. Patient samples.

Primary T-PLL cells were isolated from peripheral blood (PB) of 94 T-PLL patients diagnosed according to WHO criteria[1,2]. Differential diagnosis was based on clinical features, immunophenotyping (flow-cytometry and histochemistry; including TCL1A/MTCP1 expression), FISH/karyotypes, and molecular studies (TCR-monoclonality)[3]. Human tumor samples were obtained from patients under IRB-approved protocols following written informed consent according to the Declaration of Helsinki. Collection and use have been approved for research purposes by the ethics committee of the University Hospital of Cologne (#11-319). The cohort was selected based on **uniform front-line treatment (87% of cases)** with either single-agent alemtuzumab or fludarabine-mitoxantrone-cyclophosphamide (FMC) plus alemtuzumab chemo-immunotherapy (similar efficacy, see Refs.[4–6]) as part of the *TPLL1*[4] (NCT00278213) and *TPLL2* (NCT01186640, *unpublished*) prospective clinical trials or as included in the nation-wide T-PLL registry (IRB# 12-146) of the German CLL Study Group (GCLLSG; **TableS2**). Patients had a median age of 62 years at diagnosis and included 1.5-times more men than women. Overall survival (OS) was measured as the time from diagnosis to disease-specific event or censoring. Kaplan-Meier curves were compiled with PRISM6; with log-rank statistics for 2-group comparisons.

A small number of samples from other entities was included as references: T-cell large granular lymphocytic leukemia (T-LGL, n=2) for WES and telomere length assessments, as well as Sézary Syndrome (SS, n=2) and chronic lymphocytic leukemia (CLL, n=4) for telomere length assessments.

2. Flow cytometry, magnetic-bead based cell enrichment, and flow-FISH technique.

Flow cytometry was performed on a Gallios (BeckmanCoulter) cytometer, using antibodies against human CD4-FITC (#317407), CD8-APC-Cy7 (#300926) and TCL1A-Alexa Fluor 647 (#330508; from own developed clone 1-21 [7]), all from BioLegend. Intracellular staining was performed according to the manufacturer's instructions using the IntraPrep kit (BeckmanCoulter). We observed CD4 single positivity in 63%, CD8 single positivity in 24%, and CD4/CD8 double positivity in 14% of cases. Peripheral blood mononuclear cells (PBMCs) of T-PLL patients or healthy volunteers were obtained by density gradient centrifugation (Histopaque, Sigma Aldrich). DNAs of matched tumor/germline (t/g)-pairs were obtained after **magnetic-assisted cell sorting** (MACS), separating $CD4^+$ or $CD8^+$ T-PLL cells from non-tumor hematopoietic cells with a final purity of >98% (**Fig.S1b**). We conceptualized this T-cell enrichment to involve a sequential two-step separation process of which each was carried out according to the manufacturer's (Miltenyi Biotec) instructions: (1) positive enrichment of T-PLL tumor cells followed by (2) depletion of residual T-PLL cells from the flow-through obtained from step 1 to recover a pure non-tumor cell fraction. According to the predominant immunophenotype, samples were first enriched for $CD4^+$ (#130-045-101, Miltenyi Biotec) or $CD8^+$ (#130-045-201, Miltenyi

Biotec) lymphocytes using microbeads of the MACS system (Miltenyi Biotec) and LS Columns (#130-042-401, Miltenyi Biotec). For depletion of the normal control fractions (neutrophils, monocytes, NK-cells, B-cells) by contaminating T-PLL cells, LD Depletion Columns (#130-042-901, Miltenyi Biotec) were used to remove residual $CD4^+$ or $CD8^+$ cells from the flow-through obtained from step 1. Purity of cell populations was assessed by flow cytometry. PBMCs of healthy volunteers were enriched for $CD3^+$ pan T-cells using MACS beads (130-050-101, Miltenyi Biotec). **Flow-FISH** analyses for telomere length assessment was conducted as previously described in detail[8–11]. Healthy control lymphocytes (104 volunteers) were used for age-adaption of telomere length as reported previously[10].

Flow-cytometry based **cell cycle analysis** was performed according to Nicoletti[12]. Briefly, cells were harvested, vortexed intensely in Nicoletti buffer (0,1% w/v Sodium citrate, 0,1% v/v Triton-X100, 50µg/ml propidium iodide freshly added) and incorporation measured on a Gallios (BeckmanCoulter) cytometer.

3. Murine models for T-PLL or T-cell lymphoma.

We re-derived the originally described hemizygous **Lck^{pr}-hTCL1A$^{+/-}$ transgenic** (tg) mice[13] from frozen sperm straws (JAX® mice research, The Jackson Laboratory) by egg fertilization and embryo transfer. They represent an autochthonous model for human T-PLL. Following the early (thymic) onset of constitutive expression of human TCL1A, according to the activity of the proximal Lck promoter, the animals develop a $CD8^+$ disease that resembles human T-PLL[14].

To test drug efficacies ([redacted]) in vivo, transplantable leukemias/lymphomas derived from our **CD2-MTCP1^{p13} tg mice**[15] (predominantly blood, spleen, bone marrow) and from our **ΔJAK1 mice** (more nodal/spleen manifesting mature T-cell lymphoma based on insertional mutagenesis activating JAK1)[16] were i.p. / i.v. injected into background-matched recipients to facilitate the generation of uniform cohorts, which is not possible in the original systems due to long latencies and their wider ranges (despite 100% penetrance) of clinical disease onset. The CD2-MTCP1^{p13} tg system is a T-PLL model analogous to TCL1A-tg, but transplantable lines with slow and fast (latter chosen here) growth kinetics were only established for the **CD2-MTCP1^{p13}** model at the time of study. Transfer model from CD2-MTCP1^{p13} mice: $1x10^7$ cells were i.p. injected into syngeneic recipients (n=26). Starting on day 10 post transplantation (homogeneous distribution of WBC counts), mice were [redacted] (day 10 at 60 mg/kg, days 15, 17, 21 at 20mg/kg), and [redacted] (day 10 at 50 mg/kg, days 15, 17, 21 at 20mg/kg). Animals were randomly assigned to treatment groups (unblinded). Transfer model from ΔJAK1 mice: $2.5x10^6$ cells were transplanted intravenously into Rag-1-deficient mice. Recipients of comparable leukocyte counts were divided randomly into 4 cohorts: 18 mg/kg each for [redacted]

In order to test the **in vivo pro-leukemogenic cooperation of ATM loss with TCL1A overexpression**, hematopoietic stem cells (HSCs) from **Rosa26-CreERT2;ATM$^{fl/fl}$** mice[17] were isolated and retrovirally transduced in vitro with an expression vector for human TCL1A or GFP. Transduced HSCs were re-transplanted

into sub-lethally irradiated background-matched 8-week old recipients and tamoxifen at 1mg/day was i.p. injected for 5 consecutive days to generate *ATM* deficiency from the recombined *ATM*$^{fl/fl}$ alleles. Results (see also **Fig.4j**): At the time of last analysis (500 days), the thymic T-cell lymphomas arising from tamoxifen treated mice transplanted with HSCs harboring the *ATM*$^{fl/fl}$*/hTCL1A*^{tg} genotype showed an accelerated onset and a shorter animal survival (5/5 succumbed; median OS 221 days) compared to reconstitution with single-*hTCL1A* overexpressing (*ATM*$^{fl/fl}$*/hTCL1A*^{tg}, not treated with tamoxifen, 5/5 succumbed, median OS 370 days), single-*ATM*-k.o. (*ATM*$^{fl/fl}$*/GFP,* treated with tamoxifen, 2/5 succumbed, median not reached), or non-targeted control HSCs (0/5 succumbed).

All experiments involving living animals were conducted according to the German Animal Welfare Act (approval numbers: 20.12.A166 (*Lck*^{pr}*-hTCL1A* mice), 2012.A394 (*in vivo* treatment of *CD2-MTCP1*^{p13} and *ΔJAK1* transplants), F21/03_RP_Darmstadt (transplantation of sub-lethally irradiated mice with genetically modified HSCs).

4. Gene expression profiling (GEP).

4.1 GEP of human T-PLL cells.

Sample preparation: PBMCs isolated from T-PLL patients (>95% purity of T-cells) and CD3$^+$ T-cells isolated from PB of healthy donors (see paragraph 2 for detailed descriptions on cell purification) were submitted to RNA isolation using the mirVana kit (Invitrogen). GEP analyses were conducted using Illumina HumanHT-12 v4 BeadChip arrays according to manufacturer's instructions.

Bioinformatics: We used the Illumina proprietary software GenomeStudio v1 to background-correct and to initially annotate the probes of the HumanHT-12 v4 Expression BeadChip. We filtered samples and genes by detection p-values and fluorescence intensities for at least 2/3 hits ($p<0.05$) to reduce false calls. Batch-effects were corrected by the ComBat[18] method which uses an empiric Bayesian model framework[19]. Since the official Illumina HumanHT-12 v4 Expression BeadChip annotation is outdated, we used the data mining tool biomaRt[20], version 75 of the Ensembl database with R, version 3.1.0, and Bioconductor, version 2.10 [21].

T-PLLs (n=70) and normal controls (CD3$^+$ T-cells from 10 healthy donors) were grouped and tested separately for differential expression using the Student's t-test on log-transformed fluorescence values (normally distributed). Fold-changes (fc) were calculated on the fluorescence values without logarithmic transformation. False Discovery Rates (FDRs) were calculated using the R package "qvalue". Hierarchical clustering was carried out using the R package *gplots*, version 2.15.0 (distance function: euclidean; clustering: complete linkage). In **Fig.1a**, the dendrogram was manually cut to obtain clusters with unique expression patterns. Gene expression overlaps between human and mouse were evaluated using Venny®. Functional analyses of (differentially expressed) genes was carried out using Ingenuity® Pathway Analysis (IPA, http://www.ingenuity.com/products/ipa), ConsensusPathDB (GSOA)[22], Broad GSEA 2-2.2.1, and KEGG/GO enrichment from the R package STRINGdb, version 9_05.

For identification of **prognostic GEP signatures**, GEPs of T-PLL cases with longest (>800 days, 5 cases) overall survival (OS; time from diagnosis to death of disease;

no other events included) were compared with GEPs of cases with shortest OS (<300 days, n=5) using "Significance analysis of microarrays" (SAM) analysis in survival mode[23] (1st training set of 10 cases). We only considered cases in which the sampling date was no longer than 6 months from diagnosis and with similar lymphocyte doubling times (LDTs) at presentation. From an initial most informative index-set of 5 differentially expressed probes (*RAB25*, *KIAA1211L-probe1*, *KIAA1211L-probe2, GIMAP6*, *FXYD2;* FDR<0.1), linear regression[24] (and one outlier removal by setting OS<200 days, 2nd training set of 9 cases), followed again by SAM analysis (survival mode), resulted in a 2-gene/3-probe set (*ILMN_1791826* mapping to 4 transcripts including *RAB25-001/ENST00000361084* responsible for standard protein *ENSP00000354376*; *ILMN_1776121* and *ILMN_3243366* both mapping to *KIAA1211L-001/ENST00000397899* responsible for standard protein *ENSP00000380996*; no other probes mapping to both genes) as the most robust predictors (only when combined). Their probe sets were used to calculate an expression index (via additive model fit using Tukey's median polish procedure[25]) on the test set of uniformly treated 40 cases of the GEP-analyzed T-PLL cohort (9 cases of training set (above) excluded) fulfilling the respective criteria (available GEP and OS data). Kaplan-Meier curves (log-rank tests for differences) were created based on stratified values per patient of this "2-gene/3-probe prognostic expression index". Ranking the cases solely based on these expression indices, the 5 T-PLL with the lowest values indeed showed significantly superior OS (only 2 of these 5 cases received an allogenic stem cell transplantation) over those with higher expression index values (index fc=-1.62; **Fig.17d**).

4.2 GEP in murine T-cell leukemia.

Sample preparation: Murine spleens removed *post-mortem* were meshed through a 100µm cell strainer (BD Biosciences) and lymphoid cells were isolated using density gradient centrifugation. Cells were subsequently enriched for $CD8^+$ lymphocytes using MACS beads (130-049-401, Miltenyi Biotec). RNAs were isolated from murine tissues using the mirVana kit (Invitrogen). We hybridized 3 control RNA samples (pooled from $CD3^+$ T-cells enriched from 9 spleens of age- and background-matched wt animals), as well as RNAs isolated from $CD8^+$-enriched splenic T-cells of 3 "chronic phase" and of 5 "exponential phase" *Lck^{pr}-hTCL1A$^{+/-}$* mouse lymphoma samples on Affymetrix Mouse Gene 1.0 ST Arrays. Definition of stages: "chronic" - 30-70% tumor cells in PB and spleen, average age 12 months; "exponential" - mean PB lymphocyte doubling time (LDT) 12 days (SEM 0.8), >80% tumor cells in PB, >90% in spleen, average animal age 15 months.

Bioinformatics: Arrays were pre-processed, background-corrected (RMA), quantile-normalized, and separately analyzed (chronic phase vs ctrl., exponential phase vs ctrl.) with the „affy" R-package. Annotation of mouse probe sets and human orthologues was carried out with biomaRt. We did not only overlap Ensembl IDs, but converted MGI gene names and overlapped them with official gene symbols as well.

5. Somatic copy-number alterations (sCNAs) / loss-of-heterozygosity (LOH).

5.1 sCNAs in human T-PLL cells.

Sample preparation: DNAs were isolated from PBMCs of T-PLL patients (n=83, >95% purity of T-cells) that included 13 CD4/CD8-enriched/depleted tumor/germline (t/g) pairs (see chapter 3.2 for details on cell purification) using the QIAamp DNA Kit (Qiagen). SNP-array analyses were conducted using Affymetrix SNP 6.0 chips according to manufacturer's instructions.

Bioinformatics: To globally infer on **sCNAs across the T-PLL genome,** the T-PLL data sets were compared to the pooled controls (non-tumor hematopoietic cell DNA as 'germline' from T-PLL patients, n=13) obtained by the Affymetrix Power Tools, version 1.14.2 with duplicate SNP/CN markers (by identical position) removed. We segmented the called SNP / copy number (CN) markers by the CBS algorithm (default options, $p<0.01$) within the DNAcopy R-package[26] and converted the output files to .seg files to view them in the "Integrative Genome Viewer"[27]. Since the CBS algorithm only reports significantly altered segments/regions and therefore disregards gene structure (perhaps splits them in two or more segments), we mapped regions on gene CDS (based on version 75 of the Ensembl annotation) within the GenomicRanges R package, version 1.16.4, and clustered CNs by gene names and 100kb regions with the gplots R package. We calculated the frequency by which samples surpassed CN thresholds (CN<1.5 for losses, CN>2.5 for gains) enabling the identification of the **minimal (common) deleted or amplified regions (MDRs/MARs)** and their prevalence across the T-PLL cohort (Parker et al. 2011 [28]). Hot spots of sCNAs were identified by visual inspection, by genes (CDS ranges) assigned to segments called by the CBS algorithm as well as by confirmatory GISTIC2.0[29] analyses (with removal of centromeric and telomeric regions with options: -smallmem 1 -broad 1 -brlen 0.98 -conf 0.99 -armpeel 1 -qvt 0.05).

To evaluate **CNNLOH (copy-number neutral LOH) / UPD (uniparental disomy)**, we focused on those genes that show LOH and are in a biallelic state (CN between 1.9 and 2.1). We obtained genotypes from the SNP array data using Affymetrix Power Tools, version 1.14.2, and the Birdseed[30] algorithm, and mapped specific SNPs to the genes by version 75 of the Ensembl annotation.

A **meta-comparison** of published data on neoplasms hybridized to Affymetrix GenomeWide SNP 6.0 arrays[31–37] available at GEO[38] was performed to compare the spectrum of sCNAs with the one of our T-PLL data set. The HapMap[39] data set "GenomeWideSNP_6.hapmap270.na32.r1.a5.ref" obtained from the Affymetrix support site served as a reference. Each sample was analyzed via CBS and those with significant gains or losses (CN>2.5 or CN<1.5) were selected. We grouped these segments into region size bins for each sample, i.e. one for segments of size from 1bp to 1000bp, one for 1001bp to 10000bp, and so on. This enabled comparisons between the CN spectra across experiments and entities.

5.2 sCNAs in murine T-cell leukemia

Sample preparation: We hybridized DNA samples (QIAamp DNA Kit, Qiagen) onto the Affymetrix MOUSEDIVm520650 chip. We compared 4 controls (DNA isolated from normal liver tissues of age- and background-matched wild-type mice) to 3

'chronic phase' and 5 'exponential phase' (defining features in 4.2) splenic isolates from T-cell leukemia / lymphoma bearing *Lck*^{pr}*-hTCL1A*$^{+/-}$ mice.
Bioinformatics: Arrays were pre-processed and separately analyzed ('chronic phase' vs. ctrl., 'exponential phase' vs. ctrl.) with the 'mouseDivGeno' R-package.

6. Whole-exome sequencing (WES).

Sample preparation: DNAs were isolated from CD4 or CD8 enriched tumor/germline (t/g)-pairs (n=13, see chapter 2 for details on cell purification) using the QIAamp DNA Kit (Qiagen). Exomes were prepared by fragmenting 1µg of DNA using sonication technology (Bioruptor, Diagenode, Liège, Belgium) followed by end repair and adapter ligation including incorporation of Illumina TruSeq index barcodes. After size selection and quantification, pools of 5 libraries were each subjected to enrichment using the SeqCap EZ v2 Library kit from NimbleGen and following the NimbleGen SeqCap EZ Library SR User's Guide version 3.0 protocol[40].
Bioinformatics: We sequenced 13 T-PLL (t/g)-pairs and 26 T-PLL t-single samples (from 23 cases, with F/U samples on 3 of them) using the Illumina HiSeq2000 at the Cologne Center for Genomics (CCG), except for 8 t/g-pairs and 8 tumor singles that were analyzed at another facility (University of Michigan, collaborator/co-author K.E.-J.) for evaluations of data robustness. The mean 30x coverages were: ~422,768 exons for the CCG facility and ~307,245 exons for the outside facility[41]; median insert-sizes: 194bp for CCG facility and 254bp for outside facility. Assembly was performed with BWA 0.6.2 [42] on the UCSC hg19 reference genome. After sorting and indexing of the resulting BAM files with SAMtools, version 0.1.19, PCR duplicates were removed with Picard 1.88. Exonic regions (based on Ensembl 71) were re-aligned and the base quality scores were re-calibrated according to the Genome Analysis Toolkit Best Practices recommendations[43,44]. For 'somatic' comparisons we used the same-patient pair-matched germline if available, otherwise a representative germline sample obtained from the same batch ('predicted somatic') was used.
For **somatic single-nucleotide variants (sSNVs)** MuTect 1.1.4 and MuSic algorithms[45,46] were employed with default parameters, while for somatic **InDels** (insertions and deletions) VarScan 2.3.6 [47] was used. We also used Genome Analysis Toolkit UnifiedGenotyper 2.7-4 [48] for SNVs and InDels. Mutations were annotated using ANNOVAR[49] with the associated packages NCBI dbSNP 138 [50], COSMIC 70 WGS[51], ESP6500-SI (W. NHLBI GO Exome Sequencing Project Seattle), 1000G April 2012 [52], ExAc0.3 (Exome Aggregation Consortium, Cambridge, MA (http://exac.broadinstitute.org [06/08/2015 accessed via ANNOVAR])), NCI60 [53], and clinVar release 20150330 [54]. For proven somatic mutations we used standard MuTect filters, as well as 1000G and/or ESP6500-SI frequency and/or ExAc0.3 with minor allele fraction (MAF) <0.01 ("PopFreq <0.01"). InDel consequences were evaluated by PROVEAN[55].
SNVs were filtered by the (i) exclusion of potential SNPs by eliminating SNVs with a population frequency >0.01 (PopFreq<0.01 considered as SNV, which applies for all reported SNVs), (ii) by determination of genes that are enriched for likely damaging mutations using PolyPhen2 [56](score ≥0.957) and SIFT[57](score ≤0.05) algorithms, followed by a filter for expressed genes (GE arrays), (iii) by a statistical comparison

of observed and expected mutation rates (WUSTL MuSiC). Since we observed a high portion of G>T (and C>A) transversions in one batch of WES samples indicative for oxidative DNA damage (8-oxoguanine (8-oxoG) lesions) during sample preparation, we applied additional filters similar to the ones used in Costello et al. 2013 [58]. First we ran MuTect v2 to obtain FoxoG ratios (fraction of alternate allele supporting reads with G>T on read 1 and C>A on read 2 or vice versa; not to be confused with "strand bias") and tumor loads (estimated log odds that the observed number of alternate allele reads from the tumor sample could have arisen from a reference allele) for each mutation we previously screened with the less stringent MuTect v1. We discarded all G>T and C>A mutations with tumor fractions below 0.5 that were not found by MuTect v2 (and therefore no FoxoG ratios and tumor loads available). We further discarded all G>T and C>A mutations not surpassing the empirical filter of Costello et al. 2013: tumor loads > −10 + (100/3) * FoxoG. A Lego plot of SNV frequencies with **trinucleotide contexts** was prepared using a modified source code by developer Christopher Wardell (https://github.com/cpwardell/3dbarplot).

We calculated the **mutational frequency** without background-correction, by dividing the average number of somatic mutations per sample per target Mb (SeqCap3: 64'000'000 bp). Since we also ran samples on the lower targeting SeqCap2, the mutational frequency is actually underestimated (conservative estimate). Mutation frequencies of other neoplasms were obtained with the same caller[46] ('Published validation rates of calls made by previous versions of MuTect in coding region').

We inferred **structural variations** by mapping distance and order of paired-end reads[59] using DELLY (version 0.5.5 [60]) and filtered for a minimum genotype quality of 100, for no LowQual entries, and for split-read support (more precise breakpoint localization). CN neutral entries in the database of genomic variants (GRCh37_hg19_variants_2013-07-23 [61]) were further used to filter within a 1kb breakpoint window. The resulting list was then annotated with the COSMIC SV data sheet (02/04/2014 last modified; liftOver from hg38 to hg19 with UCSC Utilities web-GUI) and visualized with circos 0.64 [62].

For the **detection of sCNAs in WES data**, we used "ExomeDepth, version 1.0.7" with default settings, which evaluates significant drops of coverage. As the reference set, we pooled all germline samples obtained from the same batch of the respective tumor sample. Potential **microsatellite-instability (MSI)** was assessed using MSIsensor[63] with default settings. **Sequential samples** were compared in a pair-wise fashion: sample at F/U vs sample at diagnosis.

7. Whole-genome sequencing (WGS).

Sample preparation: DNA extraction was performed as described under 5 for WES. Sample processing for WGS was performed as previously reported[40].

Bioinformatics: We sequenced 3 T-PLL t/g-pairs and one T-PLL tumor single on an Illumina HiSeq2000 using the same settings as for WES analysis, except for different target regions for alignment and mutation calling, including non-coding (nc) regions. The Broad Institute hg19 Catalog of long-intergenic non-coding RNAs[64], Gencode lncRNAsv7 summary table (05/02/2012 accessed), mirBase Release 20 [65] (around

2000 validated and over 4000 predicted miRNAs), FANTOM5 hg19 enhancer sites[66] (accession 29/11/2012), and promoter regions derived from version 71 of the Ensembl annotation (-2000 to +200bp of TSS) were used. **Telomere lengths** were analyzed using 'telseq'[67].

8. Whole-transcriptome sequencing (WTS).

Sample preparation: PBMCs of T-PLL patients (>95% purity of T-cells) and $CD3^+$ T-cells isolated from PB of healthy donors (see 2 for details on cell purifications) were subjected to RNA isolation using the mirVana kit (Invitrogen). WTS analyses were conducted using the Illumina HiSeq2000platform as previously described[68].

Bioinformatics: Reads were mapped to the human reference genome, build GRCh37, using Tophat v2.0.10 [69] and the genome annotation based on the Ensembl database, version 75. After duplicate removal, the read counts were further processed using DESeq v1.14.0 [70] and DEXSeq v1.8.0 [71] to analyze **differentially expressed and differentially spliced genes** between all 15 T-PLL samples and 4 healthy-donor derived control T-cell samples. **Fusion events** were analyzed using Tophat-Fusion[72] and the associated downstream filtering pipeline (Tophat-Fusion Post). Alternatively with less stringent quality filters, but with calculation of oncogenic potential, we used oncofuse[73] with two complementary filters: passenger probability <0.001, driver probability >0.999 and minimum support reads >10, as well as passenger probability <0.01, driver probability >0.99, and minimum support reads >100. In a validation approach we aligned reads with STAR_2.5.2a [74] in 2-pass mode to the GRCh37/hg19 reference genome. Sub-routine STAR-Fusion was used to evaluate fusion transcripts. General overlap to results obtained by TopHat-Fusion was quite low (sample-wise: 21/96; global by gene partners: 30/96), however all prominent hits were confirmed: *TCL1A-TRAJ49* as well as *PLEC* with other genes on chr.8 (i.e. *ZC3H3* or *SHARPIN*). WTS samples were screened for SNVs (as anchor points for allele-specific expression) with GATK UnifiedGenotyper 2.7-4 and very low quality thresholds (--filter_mismatching_base_and_quals--filter_reads_with_N_cigar-stand_call_conf5-stand_emit_conf2). Cuffdiff (cufflinks-2.2.1.Linux_x86_64) was used to generate FPKM values (fragments per kilobase of exon per million reads mapped). VirusFinder2.0 [75], did not identify any **viral transcripts** except for J02482/Coliphage phi-X174, a control in the sequencing run. The integration-site file was empty, therefore no whole-genome screens were performed.

9. Targeted amplicon sequencing (TAS) and Sanger sequencing.

T-PLL tumor singles of 20 cases were analyzed by a customized targeted amplicon sequencing (TAS) panel that we designed. It covered *ATM* (ex.1-63), *JAK1* (ex.9-15), *JAK3* (ex.10-17) using the Illumina MiSeq platform, and *STAT5B* (ex.16) using Sanger sequencing (see **Table S23** for oligo-nucleotides).

Sample preparation: Amplicons were generated using standard PCRs. Products were purified using the ZR-96 DNA Clean-up Kit (Zymo Research), and an equimolar amplicon-pool was prepared for each patient. Library preparation was conducted using the TruSeq DNA LT Sample Prep Kit (Illumina) with 1µg amplicon DNA. Amplification was carried out using 8 cycles. The MiSeq Reagent Kit v3 (Illumina)

was used for sequencing and the samples were analyzed on the MiSeq NGS platform. Library preparation and sequencing was performed according to the manufacturer´s instruction at the Cologne Center for Genomics (CCG).

Bioinformatics: For read alignment and further read processing, we followed the same strategy as for WES (above). The Genome Analysis Toolkit Unified Genotyper 2.7-4 [43] was used without down-sampling (dCov=10000) to call mutations (SNVs and InDels). We calculated the VAF with "bam-readcount" (https://github.com/ sjackman/bam-readcount, accessed 19/12/2014) with minimal mapping and a base quality of 20. A Phred-scaled quality of at least 100 and a depth of coverage of at least 10 was presumed to restrict false positives. We further used the same filters as for potential somatic mutations in WES.

Sanger sequencing: Primers spanning all regions of interest were designed and used for PCRs according to standard protocols. PCR products were sequenced using the Big Dye Terminator Sequencing v3.1 kit and ABI PRISM 3730XL DNA Analyzer (Applied Biosystems). Capillary electrophoresis was carried out at the CCG. For electropherogram analysis SnapGene (v2.8.2, SnapGene) and 4Peaks (v1.8, nucleobytes) were used.

10. Integrative approaches of bioinformatic analyses.

Major analysis steps were executed through our 'Cancer Pipeline' (Crispatzu et al. submitted) within the QuickNGS[76] framework and downstream Semantic Web applications. Thus, mutation analysis results are written in the RDF/N3 (resource description framework) format, and stored in a jetty-6.1.26 servlet engine running an OpenRDF Workbench Version 2.6.10 Sesame server. Combinatorial (with patient data) and multiple data set analyses (**Fig.3e**, **Fig.4a**, **Fig.S6b,** and **Fig.7b-d**) as well as sample organization was done by implementing queries that were further processed with the R-package "SPARQL 1.16".

11. Quantitative real-time PCR.

Total RNA was extracted from human $CD3^+$ pan T-cells and murine $CD8^+$ T-cells following manufacturer's instructions (mirVana, Invitrogen and RNeasy Mini Kit, Qiagen). Total sample RNA was reverse-transcribed into polydT cDNAs using SuperScript II reverse transcriptase (Invitrogen). Real-time quantitative PCR on human and murine mRNA was carried out using an ABI 7500 Fast System. Primers were designed using ABI Primer Express software (**Table S23**). Each PCR reaction was performed in duplicates using the Power SYBR Green PCR Master Mix (Applied Biosystems) in 96-well optical reaction plates with the following profile: one cycle at 95°C for 10min, and then 40 cycles, each at 95°C for 30s, at 60°C for 30s, and at 72°C for 30s. Final elongation was carried out at 72°C for 30s. Primers of the genes encoding human and murine β-actin were used as standard references for quantification using the $2^{(-\text{Delta Delta C(T)})}$ method[77].

12. Cell cultures and cell lines.

RPMI-1640 medium (Sigma-Aldrich) supplemented with 1% L-Glutamine (200 mM; Sigma-Aldrich), 10% fetal bovine serum (FBS) (Sigma-Aldrich) and Penicillin /

Streptomycin (100U / 0.1M; PAA) was used for *in-vitro* experimentation on suspension cultures of primary T-PLL cells, the HH / iHH-TCL1A cell lines, the A-T derived B-lymphoblastoid cells, as well as for co-culture experiments with stromal feeder cells (below). Suspension cells were maintained at a density of 1.0-3.0x10^5 cells/ml (HH/iHH-TCL1A system and A-T lines) and of 1.0x10^6 cells/ml (T-PLL cells). Culturing was done in a HERAcell incubator (Thermo Scientific Heraeus) at 37°C and 5% CO_2 with 90% humidity.

The cell lines HH (from the ATCC), NKtert (human bone marrow stromal cells; from RIKEN Cell Bank), and the A-T patient derived lines (gift of L. Chessa, Rome, Italy) were originally acquired in 2011 and before. Only original stock propagated immediately upon arrival for 2 to 3 passages was picked for studies and cultures terminated after the 10^{th} round of passaging (4-6 weeks). Upon thawing for experimentation in 2011-16, all lines were authenticated by characteristic growth behavior and by flow cytometry confirming their characteristic immunophenotype. Each thawed passage was tested for *Mycoplasma* infection by standard PCR protocols (primers: for1: 5'-acaccatgggagytggtaat-3', rev1: 5'-cttcwtcgattycagacccaaggcat-3', for2: 5'-gtgsggmtggatcacctcct-3', rev2: 5'-gcatccaccawawacyctt-3').

HH/iHH-TCL1A system: $CD4^+$ mature T-cell leukemia HH cells were originally isolated from a patient with Sézary Syndrome[78]. iHH-TCL1A cells of inducible TCL1A expression were created by genetic modification of the parental HH (TCL1A negative) line by transfection with lentiviral expression vectors (TRMPVIR system[79]) encoding for human TCL1A under control of the doxycycline-inducible tet-on promotor (**Fig.S11**) and by subsequent puromycin selection. TCL1A was induced in iHH cells by exposure to 1µg/ml doxycycline (in ddH_2O, D9891-1G, Sigma-Aldrich) for 24hrs, or longer if otherwise indicated.

The **A-T patient derived B-lymphoblastoid cell lines**[80] 'AT65RM' ($ATM^{\Delta/\Delta}$: c.6573-9G->A/ c.8814_8824del11; ATM protein absent) and 'AT-CT' (ATM^{WT} control from unaffected relative) were used to assess (ATM related) specificity of $pKAP^{Ser824}$ induction.

For **co-culture** experiments **human bone marrow stromal cells NKtert cells** (RIKEN BRC, Japan) were seeded at concentrations of 1.5x10^4 cells/well and incubated at 37°C in 5% CO_2. After 24hrs NKtert cells at ≈60-80% confluency were inhibited with 0.02mg/ml Mitomycin C for 3hrs and then washed twice with PBS (Life Technologies). After another 24hrs, 4x10^5 T-PLL cells were added per well (with and without feeder cell support) and treated for 24-48hrs with the indicated substances. For detection of levels of **reactive oxygen species (ROS)** 6-well plates (Sarstedt, Germany) were coated with anti-CD3 (OKT3, in house, 10µg/mL) and anti-CD28 (15E8, in house, 20µg/mL) in PBS (Life Technologies) for 1hr at 37°C. The solution was gently aspirated and T-PLL cells (1x10^6/ml) were added. Flow-cytometry based analyses of intracellular ROS levels was conducted as described[81].

13. Chromosome counts.

TCL1A expression in iHH-TCL1A cells was induced by 10µg/ml doxycycline treatment. iHH cells supplemented with doxycycline vs control conditions without doxycy-

cline were cultured in parallel for 8 weeks to allow the accumulation of TCL1-induced changes (i.e. aneuploidy). Cells were maintained at a density of 1.0-3.0x10^5 cells/ml as described above. Sustained TCL1A expression was controlled by flow cytometry and/or immunoblots. Following cell cycle synchronization of cells in mitosis by 14hrs treatment with 10mg/ml nocodazole, metaphase preparations were performed as described[82]. After staining with 1% Giemsa solution, metaphase images were captured at 100x magnification by a Zeiss Axio Scope.A1 fluorescence microscope and chromosome numbers were quantified.

14. FISH analysis and karyotyping.

FISH analysis was conducted according to manufacturer's instructions using probes targeting AGO2 (customized at Empire Genomics, Custom FISH Probe, Clone Library: RPCI-11 (RP11), Clone Name: 628B24), CEP8 (Metasystems, XCE8, D-0808-050-OR), and TCRα / TCRδ sequences (LSI TCR alpha/delta dual color Break Apart rearrangement Probe, 05N41-020, Abbott Molecular). The latter probe set was used to supplement karyotypic data in order to confirm inv(14) or t(14;14) associated rearrangements of TCR gene elements as part of the aberrations that activate TCL1A expression. Karyotyping and metaphase analyses were conducted as previously described[4,83].

15. In vitro drug treatment and cell viability.

The ATM inhibitors KU55933 [84] (118500-2MG) and KU-60019 [85] (4176; TOCRIS Bioscience), the DNAPKcs inhibitor Compound 401 (234501-5MG; Calbiochem / Merck KGaA)[86], the dual DNAPK/mTOR inhibitor CC-115 [87] (Celgene), [redacted] [88] [redacted], the HDAC inhibitor SAHA[89] (vorinostat; SML0061-5mg, Sigma-Aldrich), the nitrogen mustard alkylating agent 4-Hydroperoxy-Cyclophosphamide (active metabolite of cyclophosphamide)[90] (sc-206885, Santa Cruz Biotechnology), and the topoisomerase inhibitor etoposide[91] (E1383, Sigma-Aldrich) were solved in DMSO (Carl Roth). The alkylating agent [redacted] was solved in methanol and the DNA-PK inhibitor KU-60648 [93] (S1570, Selleckchem) was solved in ethanol. Drug exposures were done at the indicated concentrations and times. Dosing was based on published ranges and own IC50/LD50 titrations. 'Non-responders' for a given inhibitor are samples for which the IC50 or LD50 was not reached, while 'responders' could be assigned a concise value. Apoptosis was determined using dual staining for Annexin-V (AnxV) and 7AAD via flow cytometry. The colorimetric MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetra-zolium bromide) assay as well as CellTiter-Glo Luminescent Cell Viability Assay from Promega assessed metabolic activity of cells and by that viability (duplicates per sample).

16. Irradiation response.

Cell lines and primary T-PLL cells were cultured in standard RPMI-1640 medium (plus supplements, see 12) and DNA damage was induced using 10Gy gamma irradiation by a BIOBEAM GM instrument (Gamma-Service Medical) equipped with a

Cs137 radionuclide source. After irradiation, cells were incubated for 1hr at 37°C, 5% CO_2 and subsequently harvested for immunoblotting.

17. Immunoblots.

Western blots on whole-cell protein lysates were performed as previously described[94]. The primary antibodies included: phospho-STAT5B^{Tyr694} (#9359), STAT5B (#9363), phospho-JAK1$^{Tyr1022/1023}$ (#3331), JAK1 (#3344), phospho-JAK3$^{Tyr980/981}$ (#5031), JAK3 (#8863), phospho-p53^{Ser15} (#9286), p53 (#2524), acetyl-p53^{Lys382} (#2525), acetyl-Histone H3^{Lys18} (#D8Z5H), PARP (#9542), γH2AX^{Ser139} (#9418), phospho-TIF1beta^{Ser824} (pKAP1; #4127), TIF1beta (KAP1; #5868), Tubulin (#I1602) and GAPDH (#3683), all from Cell Signaling Technology; ATM (sc-23921), β-actin (sc-1615), and HSC70 (sc-7298), all from Santa Cruz Biotechnology; phospho-ATM^{Ser1981} (LS-C50096) from LifeSpan BioSciences, and c-Myc (DLN07722) from Dianova. Development and use of our anti-TCL1A antibody (clone 1-21) in T-PLL has been described[20]. All primary antibodies were used at 1:1,000 dilutions, except for anti-GAPDH (1:3,000 dilution) and anti-β-actin (1:5,000 dilution). As secondary HRP-coupled antibodies we used: anti-goat (sc-2020), anti-mouse (sc-2314), and anti-rabbit (sc-2313), all from Santa Cruz Biotechnology, according to the manufacturer's instructions. Western blots were developed using Western Bright™ ECL (Advansta). Chemiluminescence was detected using Autoradiography Film Blue, 8x10 (Santa Cruz Biotechnology) and the developer machine CAWOMAT 2000 IR. Signal intensities were recorded by densitometry (ImageJ® software).

18. Immunofluorescence microscopy.

Cytospins were prepared using $1.0x10^5$ primary T-PLL or HH / iHH-TCL1A cells in a Cytospin3 cytocentrifuge (Thermo Shandon) at 800xg for 5min. Cells were fixed for 15min at 4°C in 3% PFA with 2% Sucrose in PBS. Cell permeabilization (10mM PIPES pH6.8, 100mM NaCl, 300mM Sucrose, 3mM $MgCl_2$, 1mM EDTA, 0.5% TritonX 100) and cytoskeleton stripping (10mM Tris-HCl pH7.4, 10mM NaCl, 3mM $MgCl_2$, 2% Tween20, 0.5% Sodium deoxycholate) was performed on ice each for 10min. Blocking was carried out using 5% BSA/PBS for 45min at room temperature. Primary antibodies against yH2AX (#050636, Millipore/Merck Chemicals), RAD51 (#ab63801, Abcam), TP53BP1 (#4937, Cell Signaling Technology/New England Biolabs), and ATM (sc-23921) were used at 1:200 dilution in 5% BSA/PBS over night at 4°C in a wet chamber. The secondary antibodies donkey anti-mouse (AF488 labeled) and donkey anti-rabbit (Cy3 labeled) (#715-545-150 and #711-165-152, Jackson Laboratories/Dianova) were diluted at 1:400 in 5% BSA/PBS. Incubation was carried out for 3hrs at room temperature. Slides were washed 3 times for 10min with 5% BSA/PBS and once shortly with PBS to remove BSA. Slides were coverslipped with Mowiol (Carl Roth) containing Hoechst 33258 (140µM, Sigma-Aldrich). Samples were analyzed using an Axio Scope.A1 fluorescence microscope (Zeiss). Representative images were captured using AxioVision software (Zeiss). Quantification of yH2AX, RAD51, and TP53BP1 foci was performed by manually counting the foci in 30 nuclei per time-point (means with SEM calculated). Cytosolic or nuclear ATM localization was assessed by measuring fluorescence intensity using the ImageJ®

software. Fluorescence signals derived from the whole cell and from the nucleus were determined separately in 5 cells per sample and condition. Whole cell fluorescence was set to 100 % to calculate the percentile distribution of nuclear fluorescence intensity as previously described[95].

REFERENCES (for Supplementary Methods only)

1. Herling, M. *et al.* A systematic approach to diagnosis of mature T-cell leukemias reveals heterogeneity among WHO categories. *Blood* **104,** 328–335 (2004).
2. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127,** 2375–90 (2016).
3. Ravandi, F. *et al.* T-cell prolymphocytic leukemia: a single-institution experience. *Clin. Lymphoma Myeloma* **6,** 234–9 (2005).
4. Hopfinger, G. *et al.* Sequential chemoimmunotherapy of fludarabine, mitoxantrone, and cyclophosphamide induction followed by alemtuzumab consolidation is effective in T-cell prolymphocytic leukemia. *Cancer* **119,** 2258–67 (2013).
5. Dearden, C. How I treat prolymphocytic leukemia. *Blood* **120,** 538–551 (2012).
6. Dearden, C. E. *et al.* Alemtuzumab therapy in T-cell prolymphocytic leukemia: comparing efficacy in a series treated intravenously and a study piloting the subcutaneous route. *Blood* **118,** 5799–802 (2011).
7. Herling, M. *et al.* TCL1 shows a regulated expression pattern in chronic lymphocytic leukemia that correlates with molecular subtypes and proliferative state. *Leukemia* **20,** 280–5 (2006).
8. Werner, B. *et al.* Reconstructing the in vivo dynamics of hematopoietic stem cells from telomere length distributions. *Elife* **4,** e08687 (2015).
9. Baerlocher, G. M., Vulto, I., de Jong, G. & Lansdorp, P. M. Flow cytometry and FISH to measure the average length of telomeres (flow FISH). *Nat. Protoc.* **1,** 2365–76 (2006).
10. Weidner, C. I. *et al.* Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* **15,** R24 (2014).
11. Beier, F. *et al.* Telomere dynamics in patients with del (5q) MDS before and under treatment with lenalidomide. *Leuk. Res.* **39,** 1292–1298 (2015).
12. Nicoletti, I., Migliorati, G., Pagliacci, M. C., Grignani, F. & Riccardi, C. A rapid and simple method for measuring thymocyte apoptosis by propidium iodide staining and flow cytometry. *J. Immunol. Methods* **139,** 271–9 (1991).
13. Virgilio, L. *et al.* Deregulated expression of TCL1 causes T cell leukemia in mice. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 3885–3889 (1998).
14. Warner, K. *et al.* Models for mature T-cell lymphomas-A critical appraisal of experimental systems and their contribution to current T-cell tumorigenic concepts. *Crit. Rev. Oncol. Hematol.* **88,** 680–695 (2013).
15. Gritti, C. *et al.* Transgenic mice for MTCP1 develop T-cell prolymphocytic leukemia. *Blood* **92,** 368–73 (1998).
16. Heinrich, T. *et al.* Mature T-cell lymphomagenesis induced by retroviral insertional activation of Janus kinase 1. *Mol. Ther.* **21,** 1160–8 (2013).
17. Zha, S., Sekiguchi, J., Brush, J. W., Bassing, C. H. & Alt, F. W. Complementary functions of ATM and H2AX in development and suppression of genomic instability. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 9302–6 (2008).
18. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–27 (2007).

19. Crispatzu, G., Schrader, A., Nothnage, M., Herling, M. & Diana Herling, C. A Critical Evaluation of Analytic Aspects of Gene Expression Profiling in Lymphoid Leukemias with Broad Applications to Cancer Genomics. *AIMS Med. Sci.* **3,** 248–271 (2016).
20. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–40 (2005).
21. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
22. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41,** D793–800 (2013).
23. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 5116–21 (2001).
24. Chen, K. Generalized case-cohort sampling. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **63,** 791–809 (2001).
25. Tukey, J. W. *Exploratory Data Analysis*. (Addison-Wesley, 1977).
26. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23,** 657–63 (2007).
27. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–92 (2013).
28. Parker, H. *et al.* 13q deletion anatomy and disease progression in patients with chronic lymphocytic leukemia. *Leukemia* **25,** 489–97 (2011).
29. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
30. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40,** 1253–60 (2008).
31. Wilkerson, M. D. *et al.* Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PLoS One* **7,** e36530 (2012).
32. Weiss, J. *et al.* Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. *Sci. Transl. Med.* **2,** 62ra93 (2010).
33. Boi, M. *et al.* PRDM1/BLIMP1 is commonly inactivated in anaplastic large T-cell lymphoma. *Blood* **122,** 2683–93 (2013).
34. Parkin, B. *et al.* Acquired genomic copy number aberrations and survival in adult acute myelogenous leukemia. *Blood* **116,** 4958–67 (2010).
35. Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat. Genet.* **42,** 722–6 (2010).
36. Monti, S. *et al.* Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* **22,** 359–72 (2012).
37. Edelmann, J. *et al.* High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. *Blood* **120,** 4783–94

(2012).

38. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41,** D991–5 (2013).
39. The International HapMap Project. *Nature* **426,** 789–96 (2003).
40. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524,** 47–53 (2015).
41. Kiel, M. J. *et al.* Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. *Blood* **124(9),** 1460–72 (2014).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).
43. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).
44. Van der Auwera, G. A. *et al. Current Protocols in Bioinformatics*. *Curr. Protoc. Bioinformatics* **11,** (John Wiley & Sons, Inc., 2002).
45. Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22,** 1589–98 (2012).
46. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–9 (2013).
47. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–76 (2012).
48. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
49. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
50. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–11 (2001).
51. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–50 (2011).
52. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
53. Abaan, O. D. *et al.* The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.* **73,** 4372–82 (2013).
54. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–5 (2014).
55. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7,** e46688 (2012).
56. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).
57. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*

4, 1073–81 (2009).

58. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).
59. Xi, R., Kim, T.-M. & Park, P. J. Detecting structural variations in the human genome using next generation sequencing. *Brief. Funct. Genomics* **9,** 405–15 (2010).
60. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).
61. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42,** D986–92 (2014).
62. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–45 (2009).
63. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30,** 1015–6 (2014).
64. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–27 (2011).
65. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68–73 (2014).
66. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16,** 22 (2015).
67. Ding, Z., Mangino, M., Aviv, A., Spector, T. & Durbin, R. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42,** e75 (2014).
68. Fernandez-Cuesta, L. *et al.* Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol.* **16,** 7 (2015).
69. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–11 (2009).
70. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
71. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–17 (2012).
72. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12,** R72 (2011).
73. Shugay, M., Ortiz de Mendíbil, I., Vizmanos, J. L. & Novo, F. J. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics* **29,** 2539–46 (2013).
74. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
75. Wang, Q., Jia, P. & Zhao, Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* **8,** e64465 (2013).
76. Wagle, P., Nikolić, M. & Frommolt, P. QuickNGS elevates Next-Generation

747 Sequencing data analysis to a new level of automation. BMC Genomics **16,**
748 487 (2015).

749 77. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the
750 comparative C(T) method. Nat. Protoc. **3,** 1101–8 (2008).

751 78. Starkebaum, G., Loughran, T. P., Waters, C. A. & Ruscetti, F. W.
752 Establishment of an IL-2 independent, human T-cell line possessing only the
753 p70 IL-2 receptor. Int. J. Cancer **49,** 246–53 (1991).

754 79. Zuber, J. et al. Toolkit for evaluating genes required for proliferation and
755 survival using tetracycline-regulated RNAi. Nat. Biotechnol. **29,** 79–83 (2011).

756 80. Delia, D. et al. ATM protein and p53-serine 15 phosphorylation in ataxia-
757 telangiectasia (AT) patients and at heterozygotes. Br. J. Cancer **82,** 1938–45
758 (2000).

759 81. Prinz, C. et al. Organometallic nucleosides induce non-classical leukemic cell
760 death that is mitochondrial-ROS dependent and facilitated by TCL1-oncogene
761 burden. Mol. Cancer **14,** 114 (2015).

762 82. Guo, W. & Wu, H. Metaphase preparation from murine bone marrow. Nat.
763 Protoc. Exch. (2008). doi:10.1038/nprot.2008.164

764 83. Herling, M. et al. High TCL1 expression and intact T-cell receptor signaling
765 define a hyperproliferative subset of T-cell prolymphocytic leukemia. Blood
766 **111,** 328–337 (2008).

767 84. Li, Y. & Yang, D.-Q. The ATM inhibitor KU-55933 suppresses cell proliferation
768 and induces apoptosis by blocking Akt in cancer cells with overactivated Akt.
769 Mol. Cancer Ther. **9,** 113–25 (2010).

770 85. Golding, S. E. et al. Improved ATM kinase inhibitor KU-60019 radiosensitizes
771 glioma cells, compromises insulin, AKT and ERK prosurvival signaling, and
772 inhibits migration and invasion. Mol. Cancer Ther. **8,** 2894–902 (2009).

773 86. Ballou, L. M., Selinger, E. S., Choi, J. Y., Drueckhammer, D. G. & Lin, R. Z.
774 Inhibition of mammalian target of rapamycin signaling by 2-(morpholin-1-
775 yl)pyrimido[2,1-alpha]isoquinolin-4-one. J. Biol. Chem. **282,** 24463–70 (2007).

776 87. Mortensen, D. S. et al. Optimization of a Series of Triazole Containing
777 Mammalian Target of Rapamycin (mTOR) Kinase Inhibitors and the Discovery
778 of CC-115. J. Med. Chem. **58,** 5599–5608 (2015).

779 88. [redacted]
780
781

782 89. Duvic, M. & Vu, J. Vorinostat: a new oral histone deacetylase inhibitor
783 approved for cutaneous T-cell lymphoma. Expert Opin. Investig. Drugs **16,**
784 1111–1120 (2007).

785 90. Boerrigter, G. H. & Scheper, R. J. Local administration of cytostatic drug 4-
786 hydroperoxy-cyclophosphamide (4-HPCY) facilitates cell-mediated immune
787 reactions. Clin. Exp. Immunol. **58,** 161–6 (1984).

788 91. Kaufmann, S. H. Induction of endonucleolytic DNA cleavage in human acute
789 myelogenous leukemia cells by etoposide, camptothecin, and other cytotoxic
790 anticancer drugs: a cautionary note. Cancer Res. **49,** 5870–8 (1989).

791 92. Leoni, L. M. et al. Bendamustine (Treanda) displays a distinct pattern of
792 cytotoxicity and unique mechanistic features compared with other alkylating
793 agents. Clin. Cancer Res. **14,** 309–17 (2008).

794 93. Munck, J. M. *et al.* Chemosensitization of cancer cells by KU-0060648, a dual
795 inhibitor of DNA-PK and PI-3K. *Mol. Cancer Ther.* **11,** 1789–98 (2012).

796 94. Schrader, A. *et al.* Global gene expression changes of in vitro stimulated
797 human transformed germinal centre B cells as surrogate for oncogenic
798 pathway activation in individual aggressive B cell lymphomas. *Cell Commun.*
799 *Signal. CCS* **10,** 43 (2012).

800 95. Jacquemin, V. *et al.* Underexpression and abnormal localization of ATM
801 products in ataxia telangiectasia patients bearing ATM missense mutations.
802 *Eur. J. Hum. Genet.* **20,** 305–12 (2012).

803

Aberrant effector functions of the memory-type T-PLL cell imply a leukemogenic cooperation of TCL1A with TCR signaling

K. Warner*[1,2], S. Oberbeck*[1,3], A. Schrader*[1,3], G. Crispatzu[1,3], N. Weit[1,3], P. Mayer[1,3], T. Neumann[1,3], S. Pützer[1,3], N. Pflug[1], L. Varghese[1,3], M. Thelen[1], J. Makalowski[1], N. Riet[1], G. Rappl[1], J. Altmüller[4], M. Kotrová[5], S. Stilgenbauer[6], G. Hopfinger[7], J. Dürig[8], T. Haferlach[9], M. Lanasa[10], M. Hallek[1,3], D. Mugiakakos[11], M. von Bergwelt-Baildon[1], M. Brüggemann[5], S. Newrzela[2], H. Abken[1], and M. Herling[1,3,‡]

[1] Department I of Internal Medicine, Center for Integrated Oncology (CIO) Köln-Bonn, University of Cologne (UoC), Germany, [2] Senckenberg Institute of Pathology, Goethe-University, Frankfurt/M., Germany, [3] Excellence Cluster for Cellular Stress Response and Aging-Associated Diseases (CECAD), UoC, Germany, [4] Cologne Center for Genomics, Institute of Human Genetics, UoC, Germany, [5] Medical Department II of Hematology and Oncology, University Hospitals of Schleswig Holstein, Campus Kiel, Germany, [6] Department III of Internal Medicine, University Hospital Ulm, Germany, [7] Department of Internal Medicine I, Bone Marrow Transplantation Unit, Medical University of Vienna, Vienna, Austria, [8] Clinic for Hematology, University Hospital Essen, Essen, Germany, [9] MLL Munich Leukemia Laboratory, Munich, Germany, [10]Duke University Medical Center, Durham, NC, USA, [11] Department of Medicine 5, Haematology and Oncology, University Hospital Erlangen, Germany

* contributed equally

Short title: *TCR signaling and TCL1A in T-PLL*

Key words: T-PLL, T-cell receptor, TCL1A, immune checkpoint, chimeric antigen receptor

Word count manuscript: 4392
Word count abstract: 250
Figures: main - 7; supplements - 7
Tables: main - 0; supplements - 8
References: 39

‡**Corresponding author:** Marco Herling, MD, Laboratory of Lymphocyte Signaling and Oncoproteome, Center for Integrated Oncology (CIO) Köln-Bonn and Cologne Cluster of Excellence in Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Germany

Email: marco.herling@uk-koeln.de

phone +49 221 478-5969; fax +49 221 478-98339

Key Points

- the activated memory-type T-PLL cells differ from normal T-lymphocytes by aberrant TCR-responses including anergy to apoptotic triggers
- the kinase enhancer TCL1A lowers activation thresholds conferring a permissive role of tonic TCR input, implicated in T-PLL pathogenesis

Abstract

T-cell prolymphocytic leukemia (T-PLL) is a rare malignancy, yet represents the most common mature T-cell leukemia. It is a chemotherapy-resistant and poor-prognostic tumor. Its T-cell differentiation stage and effector functions are insufficiently characterized. Constitutive transcriptional activation of the *T-cell leukemia 1A (TCL1A)* oncogene is considered the initiating leukemogenic event, but the concise mechanisms of peripheral T-cell transformation are elusive. We therefore addressed the 'T-cellness' of T-PLL and interrogated the modulatory impact by TCL1A. Immunophenotypic and gene expression profiles revealed a spectrum of memory-type differentiation with predominant central-memory stages and frequent non-canonical patterns. Virtually all T-PLL expressed a T-cell receptor (TCR) and/or CD28-coreceptor, but without overrepresentation of genetic or surface TCR-clonotypes. T-PLL cells revealed an activated phenotype and highest multi-parameter scores correlated with inferior clinical outcomes. Fittingly, they also showed resistance to stimulation-induced cell death. TCR-engagement of T-PLL cells evoked an altered metabolic signature and a prominent Th1-cytokine program. Loss of negative-regulatory TCR-coreceptors and overexpressed TCL1A distinguished the typically TCR-hyperresponsive T-PLL lymphocytes from normal T-cells. In fact, enforced TCL1A enhanced TCR-mediated kinase phosho-activation and second messenger generation and reduced input thresholds for IL-2 release. Such features were resembled in mice of TCL1-initiated protracted T-PLL development. Equipped with monoclonal epitope-defined TCRs, these *Lck^{pr}-TCL1A* T-cells gained a pre-leukemic growth advantage in scenarios of pulsed or continuous low-level receptor stimulation. Overall, we propose a model of TCR-driven T-PLL pathogenesis, in which the presence of constitutively elevated TCL1A enhances TCR-downstream signaling and drives accumulation of memory-type cells that utilize amplified, hence permissive, low-level cognate antigen input.

Introduction

T-cell prolymphocytic leukemia (T-PLL) is the most common mature T-cell leukemia.[1] Characterized by the expansion of peripheral T-cells, T-PLL typically presents with exponentially rising tumor burden in peripheral blood (PB) paralleled by splenomegaly, lymphadenopathy, and bone marrow (BM) infiltration.[2,3] The T-cells of T-PLL show a classical $CD2^+,5^+,7^+$ post-thymic immunophenotype and bear no autoaggressive features.[3–5] T-PLL shows poor responses to conventional cytostatics. The induced remissions after anti-CD52 alemtuzumab are not sustained.[6,7] With a median overall survival (OS) of <3 years, T-PLL patients still face a dismal prognosis.[2,6,7]

The most characteristic molecular hallmark of T-PLL are the rearrangements inv(14)(q11;q32) and t(14;14)(q11;q32), that juxtapose the *T-cell leukemia 1A* (*TCL1A)* oncogene locus under *in-trans* control of *TCRα/δ* gene enhancers.[8] The resulting aberrant expression of TCL1A is found in the majority of T-PLL.[2,9] As peripheral T-cells lack TCL1A expression, this abrogation of TCL1A silencing upon thymic exit is considered causal in the initiation of T-PLL. Transgenic (tg) *TCL1A* is oncogenic in mice by inducing mature T-cell leukemias that resemble human T-PLL.[8]

A mechanistic concept of TCL1A-mediated T-cell transformation is still evolving. We previously showed that in T-PLL cells TCL1A is recruited to TCR-induced protein complexes involving the signaling components ZAP70, LCK, and AKT.[3,9] A physical interaction of TCL1A with the oncogenic Ser/Thr kinase AKT enhances its catalytic activity.[9–11] Given our observation that TCL1A expression itself is inefficient in perturbing the tight homeostatic control in polyclonal settings[12], we postulated a cooperativity of TCL1A with TCR signals to promote clonal escape and leukemic outgrowth.

The maturation stage and effector profile of the T-PLL cell are insufficiently established and cannot be inferred from the non-descript clinical presentations. Furthermore, the mechanisms of T-cell transformation in T-PLL are largely elusive. With focus on the most central receptor in growth and differentiation of T-cells, the TCR, we phenotypically and functionally characterized the T-cells of 105 well-defined T-PLL, and interrogated the modulatory impact of TCL1A. The memory-type T-PLL cells were of variable clonotypic origin and differed from normal T-cells by high activation levels and aberrant TCR-elicited intracellular and effector responses. We provide data of a leukemogenic synergism of ‘tonic’ TCR signaling with elevated TCL1A.

Methods

T-PLL patients, cell lines, and mice.

PB was obtained from 105 T-PLL patients (details in **Table S1**) after informed consent according to GCP guidelines and institutional review-board approved protocols (#11-319). PB mononuclear cells (PBMCs) from T-PLL patients and healthy donors were isolated by density gradient centrifugation. The PB samples of healthy donor-derived (mean age 29 years) normal T-cells, used as controls throughout, were at average composed of 42,1% naïve, 33.7% pan-memory T-cells, 3.5% CD45RO/RA double-positive, and 20,3% double-negative T-cells.

Primary murine mononuclear cells were isolated from spleen and LN. Cell lines and culture methods are described in **online Supplements**. Animal procedures were approved by local officials (2012.A166, 2012.A394, FK/1050, 8.87-50.10.35.08.071, 84-02042012A417, 84-02042012A339).

Detailed descriptions of mouse models, compound preparations, reagents, their sources, protocols for flow cytometry, quantitative real-time PCR (primers), gene expression profiling (GEP), next generation sequencing (NGS), cell-based assays (cell cycle, apoptosis, viability, metabolic activity), transfection and transduction, immunoblots, cytomorphology, immunofluorescence, and statistics are given in the **online Supplements**.

Results

T-PLL cells retain TCR/coreceptor expression and display a spectrum of memory-like phenotypes.

The thymic event of TCL1A activation would implicate a prominent naïve compartment. This contrasts a suggested subtle oncogenic impact of TCL1A without acute maturation blocks. Therefore, the definite differentiation stage of T-PLL had to be resolved. Furthermore, in our TCR-centric oncogenic concept of mature T-cell tumors[13], T-PLL needed to be studied for the presence of viable TCR/corecepor input.

T-PLL cells from 79 patients were subjected to multi-parameter immunophenotyping. Virtually all cases expressed the surface TCRα/β (85%); if TCR-negative, the cells retained the CD3 subunit and/or the CD28 coreceptor (both in 63/74; 85%; **Fig.1A**). No case lacked all 3 receptors. T-PLL cells were predominantly CD4 single-positive (63%), followed by CD8 single-positives (24%), and CD4/CD8 double-positives (14%), independent of CD45RA/RO isoform expression (**Fig.1B**).

The majority of T-PLL (87%) was composed of a dominant T-memory subpopulation, indicated by CD45RO expression (45/79 cases, 57%) or by coexpressing CD45RA and CD45RO (n=19/79, 24%); some cases were composed of 2 distinct populations with at least one of CD45RO phenotype (n=5/79, 6%; **Fig.1B; TableS2**). The most frequent phenotype within $CD45RA^-/RO^+$ or $CD45RA^+/RO^+$ cases was a $CCR7^+/CD62L^+$ pattern of central memory T-cells (TCMs; n=35/64, 55%). Of the few $CD45RA^+/RO^-$ cases (7/79, 9%), 6 resembled classical $CCR7^+/CD62L^+$ naïve T-cells. A small number of cases showed transitional phenotypes of effector memory (EM)-like or of terminally differentiated EM T-cells with CD45RA (T-EMRA). Exemplary cases for each conventional and non-canonical pattern are illustrated in **Fig.S1A,B**. In addition to the manual hierarchical gating, a spanning-tree progression analysis of density-normalized events (SPADE)[14] was applied to the flow cytometry datasets. Using this algorithm, we identified distinct cell populations with similar marker intensities (**Fig.1C**, **S1C**) and confirmed the dominant TCM phenotype in T-PLL.

We next performed array-based gene expression profiling (GEP) of 70 PB T-PLL samples to relate global transcriptomes to those of healthy PB isolated $CD4^+$ naïve T-cells ($CD45RA^+/RO^-$; 10 donors), $CD4^+$ pan-memory T-cells ($CD45RO^+/RA^-$, 10 donors), and CM T-cells ($CD45RO^+/RA^-$, $CCR7^+$; 10 donors; **TableS3**; details on iso-

lation in **online Supplements**). In comparative algorithms, signatures that best discerned these healthy T-cell populations from each other were first identified. Guided by these most informative gene sets, T-PLL expression profiles revealed a higher similarity to memory T-cells, especially TCMs, as compared to naïve T-cells (**Fig.1D,E, S2A,B**). qRT-PCR analyses confirmed the T-cell subset-specific expression of 6/6 best-classifier genes in representative T-PLLs (**Fig.S2C**). GEP-based similarities of T-PLL cells to TCMs were further confirmed by applying published[15] T-memory signatures gene sets (**Fig.S2D**).

Interestingly, the rare tumor-immunophenotype resembling naïve T-cells was associated with a better prognosis compared to cases of CM- and EM-phenotype (**Fig.1F**). Although, this is based on small subsets of short-lived patients, we observed such a relationship already in an independent cohort.[16]

The predominant memory phenotype at the stage of overt leukemia leaves early changes undisclosed. Therefore, we took advantage of *Lck^{pr}-TCL1A* mice with early-onset (thymic) TCL1A overexpression. They develop a CD8$^+$ T-PLL-like disease after a latency of 10-20 months.[8] Splenic T-cells of pre-leukemic *Lck^{pr}-TCL1A* animals (definition of leukemic stages in **online Supplements**) were composed of 25% naïve and 65% TCM, similar to wild-type controls (**Fig.1G**). However, this significantly shifted towards a predominance of CD44$^+$/CD62L$^-$ TEM in spleens of leukemic *Lck^{pr}-TCL1A* mice (means: 9.6% (WT) *vs* 94.1% (*Lck^{pr}-TCL1A*)) with a near-complete exhaustion of the naïve compartment. T-cells in *Lck^{pr}-TCL1A* mice retained expression of CD3 and CD28 throughout leukemic evolution (**Fig.S2E**). The skewing of T-cell subsets in these murine TCL1A-driven expansions hence resembled the dominance of memory T-cells in human T-PLL. Intriguingly, the presence of enforced TCL1A does apparently not impose abnormal post-thymic subset distributions in early development (young mice). We conclude that the memory-pool accumulations at the leukemic stages are activation-enforced over a protracted course.

Overall, T-PLL is predominantly composed of cells at the memory stage. Besides a frequent CM subtype, the spectrum also entails prevalent non-canonical profiles of post-naïve T-cell differentiation. We conclude that there is no evidence for a maturation block in T-PLL cells by constitutive TCL1A expression. We propose that relevant

TCR-mediated T-cell activation had occurred during disease development or is still in place at the overt TCR/corereptor positive leukemia.

The constitutional TCR profile of T-PLL is diverse.

High-throughput sequencing of the rearranged *TCRβ* loci in 105 T-PLL (**Fig.2A**), using consensus primer-sets[17], revealed a random distribution of *TCR-Vβ* chains, with *TRBV20.1* (8%), *TRBV27* (7%), *TRBV12.3* (6%), and *TRBV19* (5%) as the most prevalent. The detected *TCR-Vβ* was usually monoclonal, but a small subset (5% T-PLL) showed a polyclonal *Vβ*-gene composition.

Transcriptome sequencing in 15 T-PLL confirmed the productive mRNA expression by the rearranged *TCR* genes as identified at the genomic level (**Fig.2B**). Compared to the *TCRα* and *TCRβ* diversities of healthy pan-$CD3^+$ T-cells (4 donors), a restricted TCR repertoire of T-PLL samples was evident in the 15 analyzed cases. The marked overall TCR diversity across T-PLL samples was further corroborated by translating the trinucleotide code into amino acid sequences of the *TCRα/β CDR3* regions, which showed no overlap (**TableS4**).

A subset of T-PLL (n=73) was also evaluated for Vβ-chain protein expression via flow cytometry (**Fig.2C**). The panel of antibodies specific to 24 TCR-Vβ families covered ~70% of the whole TCR-Vβ repertoire and proved useful in assessing T-PLL clonality and expressed TCR-specificities, with an inter-method correlation of 67% compared to genomic *TCR* analysis. As most prevalent, clonal TRBV12 (Vβ8) expression was observed in 7% of cases (n=5/73).

This high TCR-repertoire diversity was also observed in leukemias of *Lck^{pr}-TCL1A* mice (**Fig.2D,E**), in which chronologic assessments suggested evolution from a polyconal background. The arising T-cell expansions were evaluated by flow-cytometric Vβ-spectratyping comparing splenic T-cells of young (10 weeks) *vs* old (clinically leukemic; 10-16 months) *Lck^{pr}-TCL1A* mice, and each *vs* age-matched wild-type controls. Young *Lck^{pr}-TCL1A* animals showed the same polyclonal Vβ-spectrum as young and old wild-type controls. In contrast, leukemic *Lck^{pr}-TCL1A* mice showed an oligo/monoclonal Vβ-chain expression, however, as in human T-PLL, without bias towards specific Vβ-chains.

Overall, with the limitations of an under-powered T-PLL sample number against the diversity by TCR*α/β*-recombination, there seems to be no significant overrepresenta-

tion of specific clonal Vα- and Vβ-chains across T-PLL cases. Up to this point, this does not preclude the relevance of a certain antigen or of non-specific general TCR-stimulation in T-PLL. In fact, shared epitopes are detected by TCR-proteins of various genomic constitutions. As typical in memory T-cells, low-level tonic TCR-activation can also occur through self-MHC in the absence of cognate antigen.

T-PLL cells display a markedly activated phenotype.

To assess the basal activation status of T-PLL cells, we profiled up to 75 cases for established T-cell activation and proliferation markers, and for cytokine and chemokine receptors. T-PLL cells showed a heterogeneous, however, an overall elevated expression of CD38, CD69, CD40L, and Ki-67, when compared to healthy PB-derived T-cells. This was also observed for the cytokine receptors CD25 (IL-2Rα), CD122 (IL-2Rβ), CD124, and CD127 (**Fig.3A**). An elevated expression of chemokine-receptors was seen for CCR3 and CCR4, but not for CCR5, CXCR3, and CXCR4 (**Fig.3A**, **S3A**). The pattern of marker expression was not associated with specific T-cell subsets (**Fig.S3B**, **TableS5** for global correlation analysis). Expression of at least 2 activation / proliferation markers (n=31/53 cases; 58.0%) was associated with an inferior OS as opposed to those T-PLL showing a 'low' cell-activation status (0-1 marker; n=22/53 cases; 42.0%, P=0.0012; **Fig.3B**).

An activated T-cell phenotype can be induced by (constant) antigen-driven triggering of the TCR or by downregulation of inhibitory coreceptors. Accordingly, PD-L1, PD1, OX40, 4-1BB, CTLA-4, and LAG3 were found to be downregulated (**Fig.3C,D**) in T-PLL compared to healthy PB-derived T-cells, both at the mRNA and surface protein level. Furthermore and in contrast to normal T-cells, T-PLL lymphocytes did not upregulate these immune checkpoint regulators upon stimulation (**Fig.S3C**). In conjunction with a markedly distinct tumor-to-normal overexpression of the kinase-coactivator TCL1A (**Fig.3C**), this further implicates that the transformed T-cells have escaped from autoregrulatory programs to ensure an elevated net-level of activation.

TCR activation triggers an aberrant T-cell response in T-PLL.

To address whether TCR stimulation produces a functional response in T-PLL cells, prominent sinaling pathways and effector functions were evaluated (**Fig.S4A** for *in-vitro* T-PLL cell viability in response to TCR activation). Upon anti-CD3/CD28 cross-linking, most (67%; 8/12) T-PLL triggered a strong Ca^{2+}-efflux and 33% (4/12)

showed a weak response. Ca^{2+}-releases were enhanced (75%; 9/12) or suppressed (25%; 3/12,) by CD28 costimulation **(Fig.4A**).

T-cell activation entails metabolic changes; we proposed that the malignat T-cell phenotype does as well. Mitochondrial respiration and glycolysis, including their TCR-induced patterns, were assessed in T-PLL (n=4) and healthy T-cells (n=4) by measuring oxygen consumption rates (OCR) and extra-cellular acidification rates (ECAR), respectively. Fitting the more activated leukemic phenotype, an increased basal and stimulation-induced respiration was observed in T-PLL samples (**Fig.S4B**). Suggesting a prominent anaerobic leukemic profile, levels of basal glycolysis were elevated in T-PLL cells (P=0.002; **Fig.4B**) and their ECARs rose to higher levels upon CD3/CD28 engagement (P=0.04). Reactive oxygen species (ROS) as byproducts of the respiratory chain and intracellular signaling intermediates, were induced to higher levels in the leukemic cells (**Fig.4C**). We had previously shown in leukemic B-cells that TCL1A can impose elevated ROS biogenesis.[18]

Furthermore, TCR stimulation induced cell-cycle progression from G1-to-S and G2-M phases more readily in T-PLL cells than in healthy CD3+ pan-T-cells (**Fig.4D)**. T-PLL cells also displayed stimulation-induced changes in memory- and activation-marker expression, similar to normal T-cell controls (**Fig.S4C,D**). The observed re-acquisition of CD45RA upon repeated TCR stimulation (**Fig.S4C)** is a known pattern in re-activated primed T-cells.[19] Inhibitors of ITK (in part also of RLK and JAK3) suppressed the activation-induced stimulation of T-cell viability (**Fig.4E**).

T-PLL cells also revealed an enhanced activation-induced cytokine production. Particularly, there was a more robust secretion of the predominantly T-helper cell type 1 (Th1)-associated cytokines IFNγ, IL-2, IL-10, TNFα/β, GM-CSF, IL-8, IP-10, MIP-1α, and LIF as compared to healthy T-cell controls (**Fig.4F**). The releases of IL-1RA/-2/-6/-10/-13/-17A/-18/-23/-31, TNFα/β, IFNγ, IP-10, GM-CSF, LIF, MCP-1, and MIP-1α were strongly increased; and decreased for RANTES and EGF, in TCR-stimulated T-PLL cells over healthy controls (**TableS6**).

With the limited conclusions from mRNA levels on pathway activities, the profiles of TCR-signaling gene transcripts were altered in human and murine (*Lck^{pr}-TCL1A*) T-PLL (over normal T-cells) and indicated a higher activity state (**Fig.4G**).

T-PLL cells show a reduced propensity to undergo activation-induced or FAS-ligand mediated cell death.

In normal T-cells, activation-induced cell death (AICD) is triggered by interaction of CD95 (FasR) and its ligand CD95L through repeated stimulation of the TCR.[20] A potential incapability of T-PLL cells to undergo AICD could, at least in part, explain the initial uncontrolled expansion of activated T-cells. In fact, there was an almost 3-fold decrease in apoptosis upon repeated stimulation in T-PLL over healthy PB-derived T-cells (**Fig.5A**). Downregulated of CD95 only partly explained this aberrant response, as loss of its surface expression was observed in 50% of cases (n=39/68 cases; P<0.001; **Fig.5B**). The remaining cases even showed upregulation of CD95 (n=29/68; 43%; P=0.002) and CD95L. Since antibody-mediated engagement of CD95 in 12 primary leukemic samples did not induce apoptosis we conclude that the CD95-signaling pathway is not functional in T-PLL (**Fig.5C**). Indeed, both CD95-low and CD95-high cases were resistant to CD95-ligand mediated apoptosis. Fittingly, expression levels of CD95 did not correlate with clinical outcome (data not shown). Other apoptotic achses might also be dysfunctional in T-PLL, as transcript-levels of negative regulators of apoptosis, such as *BCL2* or *FYN* were downregulated and transcript levels of positive regulators, such as *BCL6* and *TNFα* were upregulated (**Fig.S5, TableS7**).

TCL1A enhances the intracellular and effector responses to TCR stimulation.

While peripheral T-cell lack TCL1A, it is overexpressed in the majority of T-PLL cases (94%; n=66/70; **Fig.3C**) and higher levels correlate with a poorer prognosis (**Fig.6A**), as we indicated already in smaller series.[7] In earlier studies, we could associate higher TCL1A protein expression across T-PLL samples with a more robust *in-vitro* growth response to TCR stimuli.[9] To specifically address the impact of high-level TCL1, we used HH human mature T-cell leukemia cells and Jurkat T-cell lines, both modified by constitutive TCL1A transgenes (comparable protein levels to those of human T-PLL, **Fig.S6A**). We used both systems, as it has been controversial and model-related, which TCR-kinase is most affected by TCL1A.[9,21,22] We observed here in stable transfectants of HH and Jurkat T-cells a stronger and earlier phospho(p)-activation of ERK1/2 upon TCR cross-linking in the $TCL1A^{+}$ sublines over their GFP-transfected $TCL1A^{neg.}$ parental lines (**Fig.S6C,D**). In a refined system, TCL1A was modulated by Tet-regulated expression in HH cells (iHH-TCL1A) allow-

ing its titration. Basal TCR-downstream p-kinase levels were slightly increased by TCL1A, but TCR-induced responses were enhanced by earlier and higher increases in pAKT and pERK levels (*vs* doxycycline-untreated controls; **Fig.6B,C**). This signal-enhancing effect was more pronounced and TCL1A-level related for pERK1/2. Ca^{2+} flux assays confirmed the TCR-signal amplifying effect by TCL1A and revealed that peaked and prolonged activation were impacted by TCL1A rather in the context of the CD3 (TCR) signal than under CD28-coreceptor stimulation (**Fig.6D**).

To assess a key distal effector function as well as to address aspects of saturations and signal replacements, we recorded IL-2 release kinetics in experiments of titrated dosages of anti-CD3, anti-CD28, and TCL1A using the iHH-TCL1A system (**Fig.6E, S6E**). The presence of TCL1A potentiated IL-2 secretion at sub-maximal intensities of CD3 engagement, whereas the maximally stimulated levels of IL-2 were independent of TCL1A. There was hardly any effect by TCL1A on isolated or additional CD28-coreceptor stimulation in this system.

As autocrine IL-2 (triggered by TCR signals) is another major growth input of T-cells, we studied TCL1A's influence on IL-2 responses. For that we employed the IL-2 dependent murine T-cell line CTLL-2; with and without transfected human TCL1A (**Fig.6F**). Introduction of TCL1A conferred increased p-levels of AKT and ERK1/2 under conditions of required and supra-maximal IL-2 dosages. This translated into a noticeable growth advantage: cell numbers of TCL1A expressing CTLL-2 cells increased with rising IL-2 concentrations, while CTLL-2 GFP-control cells maintained similar numbers (**Fig.6F**, right). TCL1A did not confer IL-2 independence.

Together, these findings corroborate the proactive impact of TCL1A at the variaous levels of TCR-induced intracellular (p-kinase activation, Ca^{2+} flux) and effector responses (IL-2 secretion, IL-2 dependent growth), those that we also observed to be aberrant in T-PLL cells. This led us to postulate that TCL1A mediates its transforming influence in a synergistic relationship with TCR-singaling. As particularly memory T-cells rely on constant provision of TCR input, TCL1A could be understood as a means to promote signal threshold reduction and permissive growth amplification.

Modelled chronic TCR stimulation facilitates TCL1A-driven transformation.

To assess for a viable TCR-TCL1A cooperation towards T-cell transformation *in vivo*, we utilized ovalbumin (OVA)-specific T-cells from TCR-tg (OT-1) mice for defined

355 TCR stimulation. Isolated OT-1 T-cells were retrovirally transduced to express
356 TCL1A-GFP, transplanted into RAG1$^{-/-}$ mice, and repeatedly stimulated *in vivo* with
357 OVA-peptides (**Fig.7A**). Blood samples from recipient mice were analyzed every 4
358 weeks by flow cytometry (**Fig.7B**). In contrast to unstimulated cohorts, the number of
359 PB circulating CD3^{+} T-cells transiently decreased in the PB of OVA stimulated con-
360 trol (GFP) and TCL1A-expressing OT-1 T-cell recipient mice within 12 weeks after
361 the first OVA injection (**Fig.7B**, left). However, only stimulated mice transplanted with
362 TCL1A^{+} OVA T-cells showed a reemergence of CD3^{+} T-cells in PB subsequent to
363 their initial decline, whereas GFP only OT-1 T-cells remained barely detectable. The
364 number of TCL1A / GFP expressing cells within the CD3^{+} population rose in stimu-
365 lated mice and in those with unstimulated TCL1A-expressing cells, but remained
366 stable in unstimulated mice with GFP-control cells (**Fig.7B**, right). Importantly, the
367 presence of TCL1A combined with TCR stimulation mediated the earliest and
368 strongest T-cell expansion. Interestingly, in this system TCL1A-negative TCR-
369 stimulated T-cells showed the same kinetics as TCL1A-positive TCR-unstimulated T-
370 cells. Knowing that the OT-1 receptor carries intrinsic activity in the absence of
371 OVA[23], this supports our concept of TCL1A promoting low-level TCR input and obvi-
372 ates requirements of strong TCR activation.

373 To address the initial loss of stimulated cells from PB, we performed biolumines-
374 cence imaging 12 weeks after the first stimulation. It revealed that transplanted cells
375 rather relocalized by accumulating in the spleen and other abdominal regions of
376 stimulated recipients of OT-1^{GFP} and OT-1^{TCL1A} cells (**Fig.7C**), with a much stronger
377 signal in the stimulated TCL1A cohort (**Fig.7D**).

378 Immunophenotyping of the PB T-cells revealed a TEM profile for the TCL1A-
379 transduced T-cells in OVA stimulated recipients, based on expression of CD44, but
380 lack of CD69, CCR7, and CD62L (**Fig.S7**). This resembled the phenotype of leuke-
381 mic *Lck^{pr}-TCL1A* mice (**Fig.1G**). The T-cells of the other cohorts showed a slightly
382 different pattern. Their T-cells were of intermediate memory T-cell phenotype, show-
383 ing CD62L expression on almost half of the cells in stimulated GFP only recipients
384 and on all cells in unstimulated cohorts (**Fig.S7**). This implicates the stimulated OT-
385 1^{TCL1A} cells as the most activated cells among these conditions.

Unstimulated recipient mice of OT-1^{TCL1A} T-cells developed lymphoid malignancies between 7-20 months after transplantation, whereas OT-1^{TCL1A} harbouring mice receiving OVA injections showed an earlier onset of the disease between 6.5-13.5 months (**Fig.7E**). Diseased mice had splenomegaly and lymphadenopathy at various extends. Histology and zytomorphology showed medium-sized lymphoid cells with scant basophilic cytoplasm in the spleen and PB (**Fig.7F**). The tumor had the described TEM phenotype (CD3$^+$, CD44$^+$, mostly lacking CCR7 and CD62L; **Fig.7F**)**.**

To corroborate these findings in a refined model system, we we took advantage of the carcinoembryonic antigen (CEA) tg mouse, which delivers a constant low-level CEA recognized by T-cells expressing a chimeric antigen receptor (CAR) with specificity for CEA.[24] In addition to such optimized form of chronic low-input TCR-stimulation, this CAR-mediated type of tissue-associated recognition of a surface self-antigen is MHC-independent. Furthermore, autologous repopulation of the host after lympho-depletion better mimics the homeostatic control enforced by a physiological polyclonal setting and by that places more competition on the experimentally modified cells. Splenocytes from 6-weeks old *CAR^{CEA} vs Lck^{pr}-TCL1A vs CAR^{CEA}xLck^{pr}-TCL1A* mice (for inter-crosses see **Online Supplements**) were transplanted into *CEA-tg* recipients and CD3$^+$ T-cells monitored (**Fig.7G**). To this end, we observed that before the eventual fast incline of only the TCL1A-tg clones and perturbation of cross-control, there was a protracted phase of smoldering expansion. In support of our TCR-TCL1A synergistic concept, at these early stages, there was a growth advantage of *CAR^{CEA}xLck^{pr}-TCL1A* T-cells over *Lck^{pr}-TCL1A* cells (**Fig.7H**).

Together, both experimental *in vivo* systems of defined chronic TCR-stimulation to TCLA1 overexpressing T-cells expand on our *in-vitro* observations of TCL1A as a TCR-signaling enhancer.

Discussion

The functional features and signal dependencies of the T-PLL cell need to be better understood to develop more effective treatments for this poor-prognostic disease. With this study we describe a cohort of T-PLL cases that is sufficiently large to allow definition of key phenotypic and functional features, including their natural variation.

Based on immunophenotyping and global gene expression, we established a high similarity of T-PLL cells to memory T-cells in the vast majority of cases (>85%), specificially to $CD45RO^+$, $CCR7^+$ CM T-lymphocytes. Previous descriptions also suggested a memory stage of maturation in 40-60% of cases[3,5,9,16,25], but were solely based on CD45RO expression. However, advances in the definition of physiologic T-cell subsets enabled us to refine the spectrum of the memory-type subsets in T-PLL. We revealed a continuum of memory stages with often non-conventional patterns, which in conjunction with the activated phenotype and retained TCR-/coreceptor expression of the tumor cell implicates continued TCR-mediated activation. The high-level CD7 expression observed in 94% (85/90) of our cases (not shown), however, argues agains exhaustion.[26,27] These features can be chronologically recapitulated in models of TCL1A-driven murine T-PLL (**Fig.1G**). An expanding memory pool had also been described for mice with lymphocytic overexpression of the TCL1 gene family member *MTCP1*.[28]

The memory phenotype suggests (auto)antigen experience or at least MHC-driven activation and differentiation of the TCL1A-affected precursor during clonal outgrowth. Chronic antigen stimulation is implicated in other T-cell malignancies as well, e.g. by auto-immune triggers in T-cell large granular lymphocyte leukemia (T-LGL) or by (bacterial) dermatitis in the cutaneous T-cell lymphomas of mycosis fungoides (MF) and Sezary Syndrome (SS).[29] In support, these entities show indications of a biased *TCRβ* gene usage.[30–32] Interestingly, malignant T-cells of MF were characterized as TEM and those of SS as TCM.[33] The diverse *TCRβ* receptoire found in our cohort of T-PLL (**Fig.2**) does not discard an antigen-dependent pathogenesis. In fact, it remains to be determined, if the slight overrepresentations of certain *TCRβ*'s at frequencies of 5-8% constitute receptors that facilitate more permissive signaling. Moreover, even if considering T-PLL of random clonotypic origin, its TCL1A-driven development could involve activation by any TCR-mediated signal (e.g. variety of

443 antigens or sole self-MHCs). Of note, in our TCR-centric concept of T-cell lympho-
444 mas, there are also entities in which the precursor lost TCR expression and survival
445 input is provided by oncogenes acting as TCR-signaling mimics or stand-in's.[13,34]

446 T-PLL cells do not behave like physiologic CM T-cells upon repetitive antigen stimu-
447 lation. Healthy TCM increase CD95 expression to facilitate regulatory apoptotic re-
448 sponses. In T-PLL, CD95 is downregulated or dysfunctional (**Fig.5**). Generally,
449 memory T-cell subsets are characterized by a marked longevity. Especially TCM
450 have been shown to additionally harbor stem-cell like properties representing early
451 differentiated progenitors with self-renewal capacity.[35,36] Acquisition of proliferative
452 stem-cell like capacities and loss of the capability to respond to extrinsic apoptotic
453 signals might be an oncogenic mechanism of persistence of T-PLL cells.

454 Our cell line data demonstrate TCL1A to augment intracellular signaling and effector
455 responses, particularly following low-intensity TCR stimulation. This effect seemed
456 more pronounced in the context of a CD3 (TCR) signal, as also supported by a more
457 obvious pERK1/2 modulation, as compared to the CD28-coreceptor signal, mostly
458 mediated via AKT.[37] This reconciles data from various model systems.[9,21,22] In ex-
459 trapolation, we argue that the inappropriate expression of this proto-oncogene in the
460 affected peripheral T-cells enables sustenance as a quiescent memory fraction by
461 amplifications of low-level TCR input through signal sensitization. Fittingly, human T-
462 PLL cells show such a reduced TCR-activation threshold. The previously unrecog-
463 nized Th1 program elicited by TCR stimulation of T-PLL cells (**Fig.4F**) is in agree-
464 ment with TCL1A's enhancement of IFN-γ production in primed Th1 cells.[21]

465 In our *in-vivo* model systems, TCL1A-transduced T-cells showed an accelerated
466 outgrowth upon repeated TCR stimulation. This provides evidence that the modula-
467 tion by TCL1A in an oncogenic cooperation with TCR signals is indeed relevant. Our
468 data also implicate that TCL1A rather augments TCR responsiveness in the early
469 stages of leukemic development. We speculate that by lowering the TCR signaling
470 threshold, TCL1 propels the transition of naïve T-cells into an expanding T-memory
471 pool as the origin of T-PLL outgrowth. These TCL1A-affected cells would be more
472 self-sustaining due to the ability to more efficiently utilize low affinity TCR signals,
473 potentially through self-antigen. Our system of CARs as powerful TCR surrogates

underlines that physiological levels of an auto-antigen, in this case CEA, can be sufficient to trigger the TCL1A mediated amplification of T-cells *in vivo*.

Our data sustain a concept of T-PLL as an (auto)antigen/(self)MHC-TCR-promoted disease with TCL1A acting as an TCR-signaling enhancer. It entails the accumulation of memory-type cells utilizing low-level TCR activation to acquire competitive advantages towards homeostatic escape and full transformation. Initiated as a TCL1A-affected thymic emigrant rather than being subject of a primary maturation block at the memory-stage, the CM-like phenotype of T-PLL cells likely reflects the terminal line of differentiation at which additional oncogenic forces come to carry to completely perturb the homeostatic control. Future work will have to integrate this T-cell development based model with the defined roles of aberrant pathways instructed by the genomic lesions in ATM or JAK/STAT signaling.[38]

Acknowledgements

Ma.He., H.A., and S.N. are funded by the German Research Foundation (DFG) as part of the collaborative research group on mature T-cell lymphomas, 'CONTROL-T' (FOR1961). Further support: Köln Fortune Program (to A.S.) and Fritz Thyssen foundation (10.15.2.034MN; to Ma.He. and A.S.). We gratefully acknowledge all contributing centers enrolling patients into the trials and registry of the GCLLSG; the GCLLSG staff and the patients with their families for their invaluable contributions. We thank Principia Biopharma, San Francisco, for privision of the ITK Inhibitor PRN-694.

Authorship Contributions

Design and experimental data analysis: Ma.He., S.N., H.A., K.W., S.O., A.S.; Experiments: K.W., S.O., A.S., N.W., P.M., T.N., S.P., L.V., M.T., J.M., N.R., G.R., M.K., J.A., D.M., M.v.B.-B., M.B.; Biostatistical analyses: G.C., M.K.; Patient samples, immunophenotypes, and karyotyping: Ma.He., N.P., S.S., G.H., J.D., T.H., M.L.: Clinical data: Ma.He., N.P., G.H., Mi.Ha.; Manuscript preparation: K.W., S.O., A.S., H.A., Ma.He..

Disclosure of Conflicts of Interest

There were no competing interests interfering with the unbiased conduction of this study.

References

1. Swerdlow S, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;127(20):2375–90.
2. Herling M, Khoury JD, Washington LT, et al. A systematic approach to diagnosis of mature T-cell leukemias reveals heterogeneity among WHO categories. *Blood*. 2004;104(2):328–335.
3. Matutes E, Brito-Babapulle V, Swansbury J, et al. Clinical and laboratory features of 78 cases of T-prolymphocytic leukemia. *Blood*. 1991;78(12):3269–74.
4. Dearden C, Matutes E, Catovsky D. Deoxycoformycin in the treatment of mature T-cell leukaemias. *Br J Cancer*. 1991;64(5):903–906.
5. Garand R, Goasguen J, Brizard A, et al. Indolent course as a relatively frequent presentation in T-prolymphocytic leukaemia. *Br J Haematol*. 1998;103(2):488–494.
6. Dearden C. How I treat prolymphocytic leukemia. *Blood*. 2012;120(3):538–51.
7. Hopfinger G, Busch R, Pflug N, et al. Sequential chemoimmunotherapy of fludarabine, mitoxantrone, and cyclophosphamide induction followed by alemtuzumab consolidation is effective in T-cell prolymphocytic leukemia. *Cancer*. 2013;119(12):2258–67.
8. Virgilio L, Lazzeri C, Bichi R, et al. Deregulated expression of TCL1 causes T cell leukemia in mice. *Proc Natl Acad Sci U S A*. 1998;95(7):3885–3889.
9. Herling M, Patel KA, Teitell MA, et al. High TCL1 expression and intact T-cell receptor signaling define a hyperproliferative subset of T-cell prolymphocytic leukemia. *Blood*. 2008;111(1):328–337.
10. Laine J, Kunstle G, Obata T, Sha M, Noguchi M. The protooncogene TCL1 is an Akt kinase coactivator. *Mol Cell*. 2000;6(2):395–407.
11. Pekarsky Y, Koval A, Hallas C, et al. Tcl1 enhances Akt kinase activity and mediates its nuclear translocation. *Proc Natl Acad Sci U S A*. 2000;97(7):3028–3033.
12. Newrzela S, Cornils K, Li Z, et al. Resistance of mature T cells to oncogene transformation. *Blood*. 2008;112(6):2278–2286.
13. Warner K, Weit N, Crispatzu G, et al. T-Cell Receptor Signaling in Peripheral T-Cell Lymphoma – A Review of Patterns of Alterations in a Central Growth Regulatory Pathway. *Curr Hematol Malig Rep*. 2013;8(3):163–72.
14. Qiu P, Simonds EF, Bendall SC, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*. 2011;29(10):886–91.
15. Haining WN, Ebert BL, Subrmanian A, et al. Identification of an evolutionarily conserved transcriptional signature of CD8 memory differentiation that is shared by T and B cells. *J Immunol*. 2008;181(3):1859–68.
16. Ravandi F, O'Brien S, Jones D, et al. T-cell prolymphocytic leukemia: a single-institution experience. *Clin Lymphoma Myeloma*. 2005;6(3):234–239.
17. van Dongen JJM, Langerak AW, Brüggemann M, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-

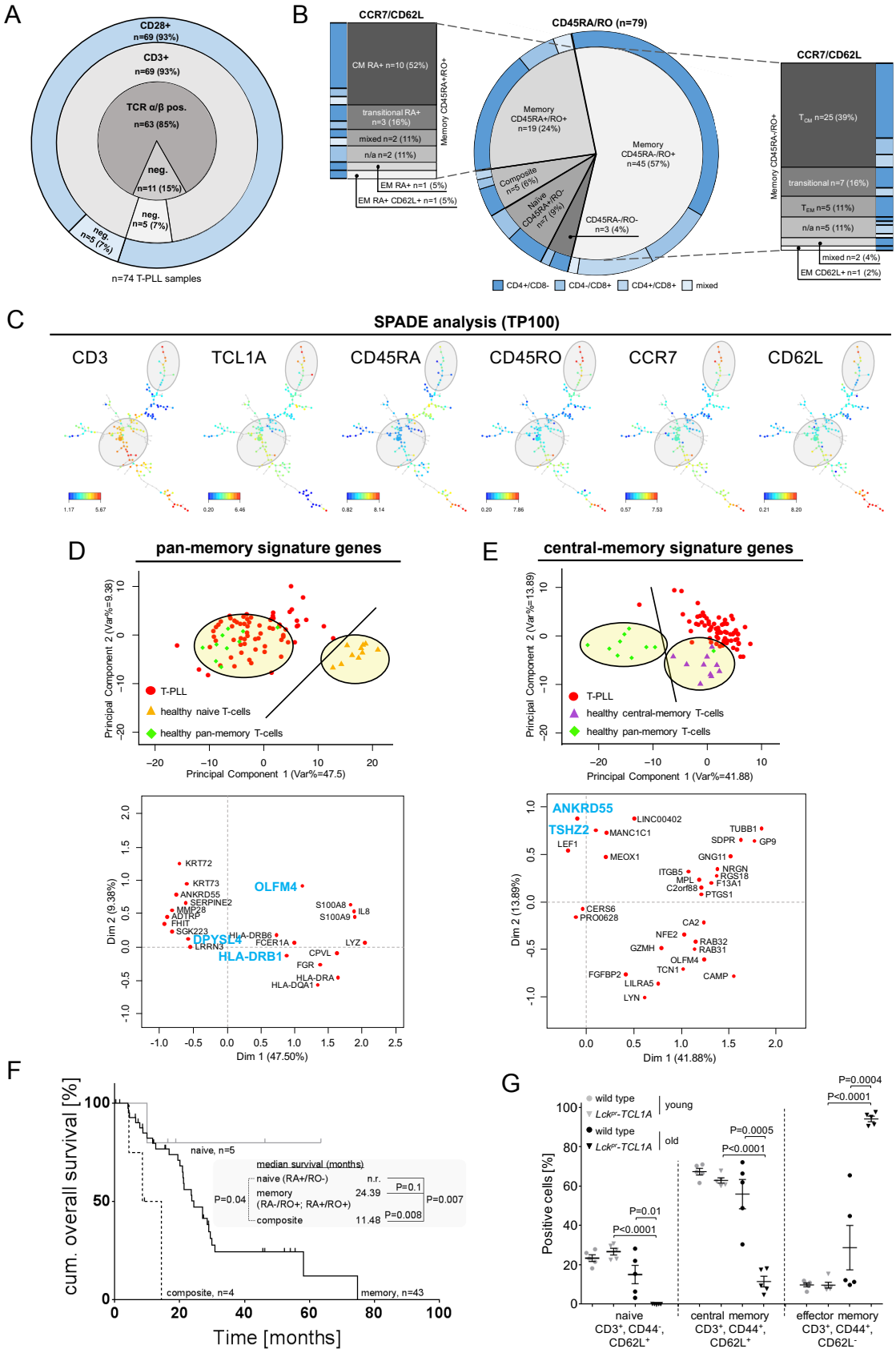
553 3936. *Leukemia*. 2003;17(12):2257–317.
554 18. Prinz C, Vasyutina E, Lohmann G, et al. Organometallic nucleosides induce non-
555 classical leukemic cell death that is mitochondrial-ROS dependent and facilitated
556 by TCL1-oncogene burden. *Mol Cancer*. 2015;14:114.
557 19. Henson SM, Riddell NE, Akbar AN. Properties of end-stage human T cells
558 defined by CD45RA re-expression. *Curr Opin Immunol*. 2012;24(4):476–81.
559 20. Krammer PH. CD95's deadly mission in the immune system. *Nature*.
560 2000;407(6805):789–95.
561 21. Hoyer KK, Herling M, Bagrintseva K, et al. T cell leukemia-1 modulates TCR
562 signal strength and IFN-gamma levels through phosphatidylinositol 3-kinase and
563 protein kinase C pathway activation. *J Immunol*. 2005;175(2):864–873.
564 22. Despouy G, Joiner M, Le Toriellec E, Weil R, Stern MH. The TCL1 oncoprotein
565 inhibits activation-induced cell death by impairing PKC theta and ERK pathways.
566 *Blood*. 2007;110(13):4406–4416.
567 23. Goldrath AW, Bevan MJ. Low-affinity ligands for the TCR drive proliferation of
568 mature CD8+ T cells in lymphopenic hosts. *Immunity*. 1999;11(2):183–90.
569 24. Chmielewski M, Hahn O, Rappl G, et al. T cells that target carcinoembryonic
570 antigen eradicate orthotopic pancreatic carcinomas without inducing
571 autoimmune colitis in mice. *Gastroenterology*. 2012;143(4):1095–107.e2.
572 25. Ascani S, Leoni P, Fraternali Orcioni G, et al. T-cell prolymphocytic leukaemia:
573 does the expression of CD8+ phenotype justify the identification of a new
574 subtype? Description of two cases and review of the literature. *Ann Oncol*.
575 1999;10(6):649–653.
576 26. Aandahl EM, Sandberg JK, Beckerman KP, et al. CD7 is a differentiation marker
577 that identifies multiple CD8 T cell effector subsets. *J Immunol*.
578 2003;170(5):2349–55.
579 27. Rappl G, Schrama D, Hombach A, et al. CD7(-) T cells are late memory cells
580 generated from CD7(+) T cells. *Rejuvenation Res*. 2008;11(3):543–556.
581 28. Joiner M, Le Toriellec E, Despouy G, Stern M-H. The MTCP1 oncogene modifies
582 T-cell homeostasis before leukemogenesis in transgenic mice. *Leukemia*.
583 2007;21(2):362–6.
584 29. Burg G, Kempf W, Haeffner A, et al. From inflammation to neoplasia: new
585 concepts in the pathogenesis of cutaneous lymphomas. *Recent Results Cancer*
586 *Res*. 2002;160:271–280.
587 30. Morgan SM, Hodges E, Mitchell TJ, et al. Molecular analysis of T-cell receptor
588 beta genes in cutaneous T-cell lymphoma reveals Jbeta1 bias. *J Invest*
589 *Dermatol*. 2006;126(8):1893–1899.
590 31. Clemente MJ, Wlodarski MW, Makishima H, et al. Clonal drift demonstrates
591 unexpected dynamics of the T-cell repertoire in T-large granular lymphocyte
592 leukemia. *Blood*. 2011;118(16):4384–93.
593 32. Rodríguez-Caballero A, García-Montero AC, Bárcena P, et al. Expanded cells in
594 monoclonal TCR-alphabeta+/CD4+/NKa+/CD8-/+dim T-LGL lymphocytosis
595 recognize hCMV antigens. *Blood*. 2008;112(12):4609–16.
596 33. Campbell JJ, Clark RA, Watanabe R, Kupper TS. Sezary syndrome and mycosis
597 fungoides arise from distinct T-cell subsets: a biologic rationale for their distinct
598 clinical behaviors. *Blood*. 2010;116(5):767–71.

599 34. Malcolm TIM, Villarese P, Fairbairn CJ, et al. Anaplastic large cell lymphoma
600 arises in thymocytes and requires transient TCR expression for thymic egress.
601 *Nat Commun*. 2016;7:10087.
602 35. Stemberger C, Neuenhahn M, Gebhardt FE, et al. Stem cell-like plasticity of
603 naïve and distinct memory CD8+ T cell subsets. *Semin Immunol*. 2009;21(2):62–
604 8.
605 36. Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T cell subsets,
606 migration patterns, and tissue residence. *Annu Rev Immunol*. 2013;31:137–61.
607 37. Parry R V, Reif K, Smith G, et al. Ligation of the T cell co-stimulatory receptor
608 CD28 activates the serine-threonine protein kinase protein kinase B. *Eur J*
609 *Immunol*. 1997;27(10):2495–501.
610 38. Kiel MJ, Velusamy T, Rolland D, et al. Integrated genomic sequencing reveals
611 mutational landscape of T-cell prolymphocytic leukemia. *Blood*.
612 2014;124(9):1460–1472.
613 39. Dondorf S, Schrader A, Herling M. Interleukin-2-inducible T-cell Kinase (ITK)
614 Targeting by BMS-509744 Does Not Affect Cell Viability in T-cell Prolymphocytic
615 Leukemia (T-PLL). *J Biol Chem*. 2015;290(16):10568–10569.
616

617

Figures and Figure Legends

Figure 1



618

Figure 1: The TCR-positive T-PLL cells cover a spectrum of memory phenotypes with a predominant CM fraction and frequent unconventional patterns.
A-C) Surface (s) marker expression (multi-parameter flow cytometry) in PB-derived primary T-PLL cells. **A**) TCR- and co-stimulatory receptor components: TCRα/β, TCRγ/δ, CD3, CD28 (n=74 cases). Of the 5 sCD3-negative cases, 4 showed cytoplasmic CD3 positivity. **B**) Distribution of T-helper (CD4) / cytotoxic (CD8) markers and of markers of naïve / memory differentiation: CD45RA, CD45RO, CCR7, and CD62L (n=79; **Table S2** for marker-defined subsets). Gating strategies and exemplarily plots per category are given in **Fig.S1A**. **C**) Confirmatory SPADE analyses[14] of markers; one exemplary T-PLL shown, others in **Fig.S1C**). Tree structure of SPADE with cell populations visualized as nodes. Size and colour of these nodes represent the number of cells and intensity of marker expression, respectively. T-PLL nodes (identified by high expression of CD3 and TCL1A), reveal low expression of CD45RA alongside high expression of CD45RO, CCR7, and CD62L, thus, reflecting a central memory (CM) T-cell phenotype. **D, E**) GEP: primary T-PLL cells (n=70 cases), healthy PB-derived naïve, pan-memory, and CM T-cells (n=10 donors each). Principal component analyses (PCA) of signature genes defining healthy naïve and memory T-cell subsets (25 most differentially expressed genes after comparing these healthy donor-derived T-cell substes; +FC sorted; p-value cutoff 0.05). Most informative genes are plotted underneath. For heatmaps showing signature gene expression in T-PLL *vs* control samples (unsupervised clustering) see **Fig.S2A,B**. **D**) PCA for memory T-cell signature genes (pan-memory *vs* naïve T-cells). First 2 dimensions are plotted and account for 47.50% and 9.38% of variance (third dimension: 6.99%). **E**) PCA for CM T-cell signature (CM T-cells *vs* pan-memory T-cells). First 2 dimensions are plotted and account for 41.88% and 13.89% of variance (third dimension: 8.74%). **F**) Kaplan-Meier plot of disease-specific overall survival (log-rank test, time from diagnosis to event) of uniformly treated T-PLL patients stratified by CD45RA/RO surface expression (n=52 cases, 1 EMRA case, 2 $CD45RA^{-}/RO^{-}$ cases excluded). CD4/CD8 expression is not associated with differential disease outcomes (not shown). **G**) TCL1-driven accumulation of EM pool: flow cytometry of murine spleen-derived lymphocytes from young (10 weeks) or old (10-16 months) *Lck^{pr}-TCL1A* mice (n=5) vs age-matched C57/B6J wild-type controls.

Figure 2

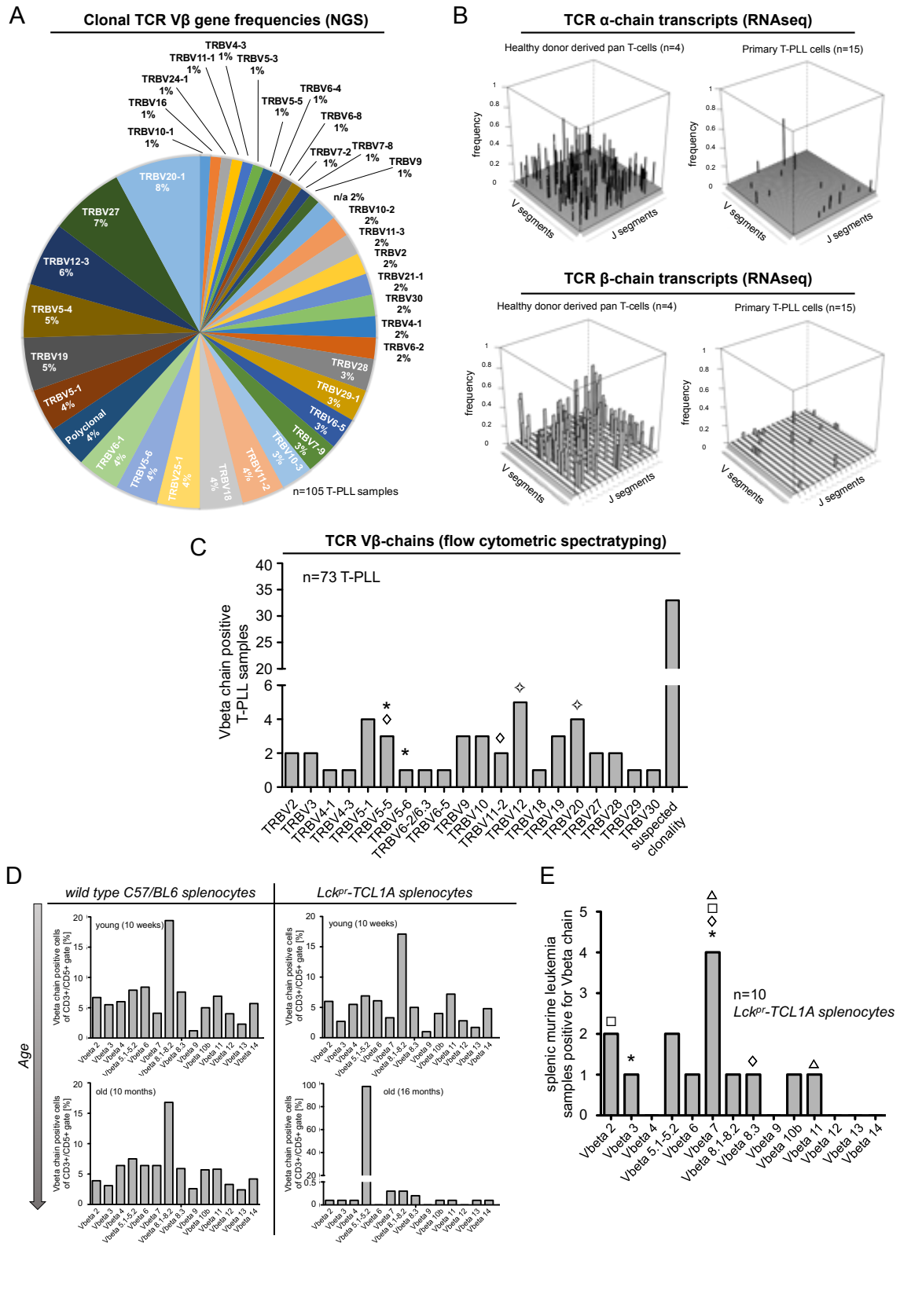

A
Clonal TCR Vβ gene frequencies (NGS)
TRBV20-1 8%
TRBV27 7%
TRBV12-3 6%
TRBV5-4 5%
TRBV19 5%
TRBV5-1 4%
Polyclonal 4%
TRBV6-1 4%
TRBV5-6 4%
TRBV25-1 4%
TRBV18 4%
TRBV11-2 4%
TRBV10-3 3%
TRBV7-9 3%
TRBV6-5 3%
TRBV29-1 3%
TRBV28 3%
TRBV6-2 2%
TRBV4-1 2%
TRBV30 2%
TRBV21-1 2%
TRBV2 2%
TRBV11-3 2%
TRBV10-2 2%
n/a 2%
TRBV9 1%
TRBV7-8 1%
TRBV7-2 1%
TRBV6-8 1%
TRBV6-4 1%
TRBV5-5 1%
TRBV5-3 1%
TRBV4-3 1%
TRBV11-1 1%
TRBV24-1 1%
TRBV16 1%
TRBV10-1 1%
n=105 T-PLL samples
B
TCR α-chain transcripts (RNAseq)
Healthy donor derived pan T-cells (n=4)
Primary T-PLL cells (n=15)
frequency
V segments
J segments
TCR β-chain transcripts (RNAseq)
Healthy donor derived pan T-cells (n=4)
Primary T-PLL cells (n=15)
C
TCR Vβ-chains (flow cytometric spectratyping)
n=73 T-PLL
Vbeta chain positive T-PLL samples
TRBV2 TRBV3 TRBV4-1 TRBV4-3 TRBV5-1 TRBV5-5 TRBV5-6 TRBV6-2/6-3 TRBV6-5 TRBV9 TRBV10 TRBV11-2 TRBV12 TRBV18 TRBV19 TRBV20 TRBV27 TRBV28 TRBV29 TRBV30 suspected clonality
D
wild type C57/BL6 splenocytes
Lckpr-TCL1A splenocytes
Age
young (10 weeks)
old (10 months)
old (16 months)
Vbeta chain positive cells of CD3+/CD5+ gate [%]
E
splenic murine leukemia samples positive for Vbeta chain
n=10
Lckpr-TCL1A splenocytes
Vbeta 2 Vbeta 3 Vbeta 4 Vbeta 5.1-5.2 Vbeta 6 Vbeta 7 Vbeta 8.1-8.2 Vbeta 8.3 Vbeta 9 Vbeta 10b Vbeta 11 Vbeta 12 Vbeta 13 Vbeta 14

Figure 2: The TCRs in T-PLL are not restricted to specific clonotypes.

A) Clonal genomic TRB rearrangements as detected by amplicon based NGS (Illumina MiSeq platform) in primary PB-derived T-PLL cells (n=105 cases). Sequencing libraries were prepared using modified biomed-2 primers for complete *TRB* rearrangements.[17] **B**) Transripts of TCR alpha and beta chains (usage of distinct V- and J-segments) as detected by RNAseq (Illumina HiSeq2000 platform) in primary PB-derived T-PLL cells (n=15 cases; >95% purity of T-cells) and healthy donor PB-derived $CD3^+$ T-cells. **C**) Flow-cytometric Vβ-spectratyping in primary PB-derived T-PLL cells (n=73 cases) using the IO beta mark kit (Beckman Coulter; ~70% coverage of the potential Vβ spectrum). Vβ-negativity (despite CD3 expression, 'suspected clonality') in 33 cases (45%). Distinct expression of at least 1 Vβ-family in 40 cases (55%); of those there were 37 monoclonal cases and 3 cases with 2 clones (indicated by 1 symbol each: *, ◊, ✧). **D, E**) Vβ-chain spectratyping in splenocytes from young (10 weeks) and old leukemic (10-16 months) *Lck^{pr}-TCL1A* mice compared to age- and background-matched wild-type controls (n=5 each) **D**) Representative examples per cohort shown. **E**) Summary of Vβ spectratypes observed in leukemic *Lck^{pr}-TCL1A* mice. Malignant cells show dominant expression of 2 Vβ-chains per sample in 4 animals (indicated by 1 symbol each: ◊, *, Δ, □).

Figure 3

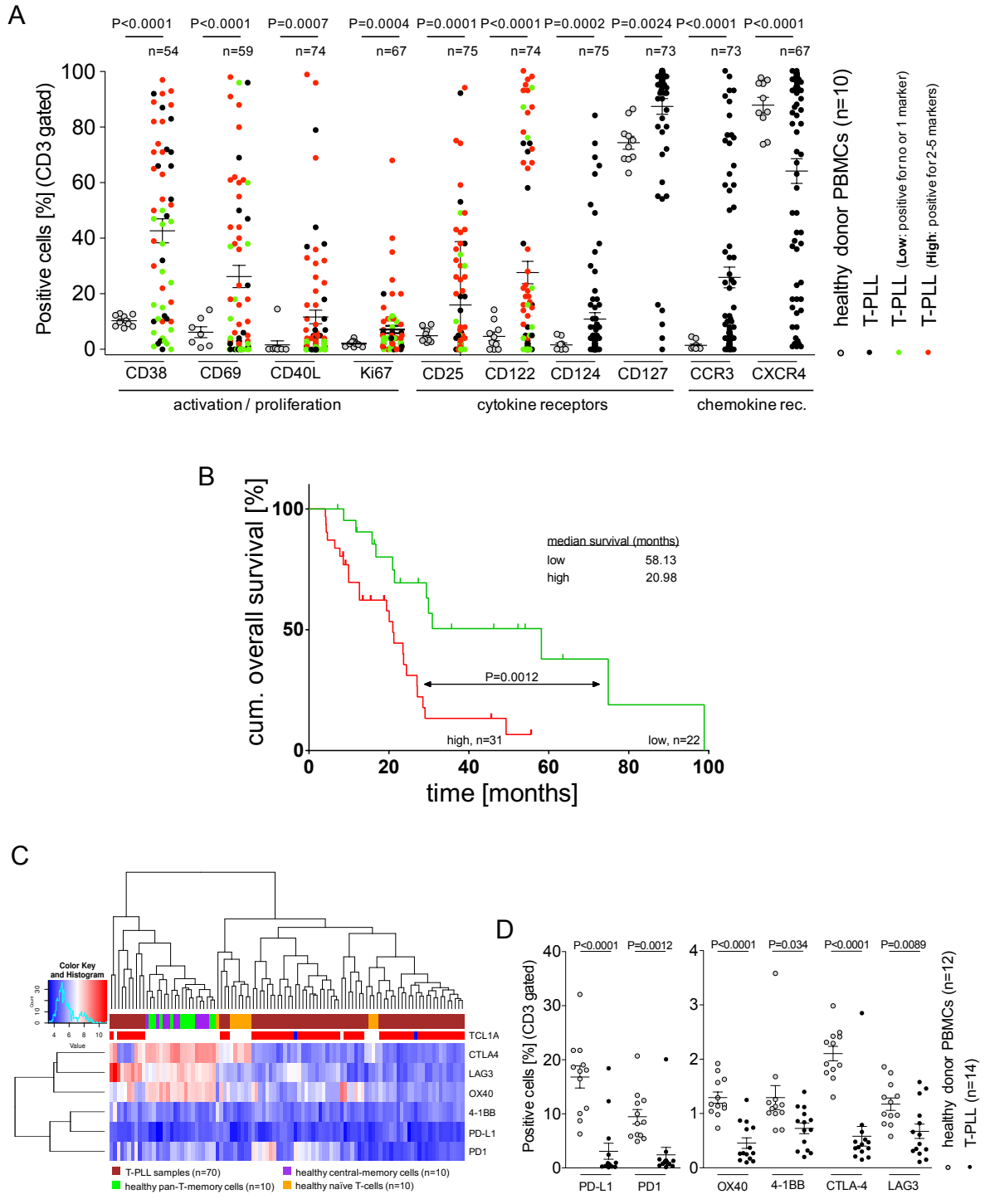

A
P<0.0001 P<0.0001 P=0.0007 P=0.0004 P=0.0001 P<0.0001 P=0.0002 P=0.0024 P<0.0001 P<0.0001
n=54 n=59 n=74 n=67 n=75 n=74 n=75 n=73 n=73 n=67
Positive cells [%] (CD3 gated)
CD38 CD69 CD40L Ki67 CD25 CD122 CD124 CD127 CCR3 CXCR4
activation / proliferation
cytokine receptors
chemokine rec.
healthy donor PBMCs (n=10)
T-PLL
T-PLL (Low: positive for no or 1 marker)
T-PLL (High: positive for 2-5 markers)
B
cum. overall survival [%]
median survival (months)
low 58.13
high 20.98
P=0.0012
high, n=31
low, n=22
time [months]
C
Color Key and Histogram
TCL1A
CTLA4
LAG3
OX40
4-1BB
PD-L1
PD1
T-PLL samples (n=70)
healthy central-memory cells (n=10)
healthy pan-T-memory cells (n=10)
healthy naïve T-cells (n=10)
D
P<0.0001 P=0.0012 P<0.0001 P=0.034 P<0.0001 P=0.0089
Positive cells [%] (CD3 gated)
PD-L1 PD1 OX40 4-1BB CTLA-4 LAG3
healthy donor PBMCs (n=12)
T-PLL (n=14)

Figure 3: T-PLL cells reveal an increased activation status and a signature of immune checkpoint inhibition.

A) Significantly increased percentages of cells (flow cytometry) expressing activation / proliferation markers as well as cytokine (IL-2, -4, -7) and chemokine receptors in T-PLL (up to 75 cases) over normal PB T-lymphocytes (10 healthy donors). T-PLL cases are color coded (green: low activity; red: high activity) according to stratified activation status defined in (B). **B**) Kaplan-Meier plot of disease-specific overall survival (log-rank test, time from diagnosis to event) of uniformly treated T-PLL patients categorized by 'activation phenotype' (flow cytometry): Analyte cut-offs: CD122 (>10% pos. cells), CD25 (>50% pos. cells), CD38 (>50% pos. cells), CD40L (>5% pos. cells), CD69 (>5% pos. cells), and Ki-67 (>20% pos. cells). Strata: number of activation / proliferation markers expressed above threshold (low: 0/1 marker, high: ≥2 markers). **C**, **D**) Elevated *TCL1A* gene expression and significantly reduced expression of negative TCR-regulatory receptors ('immune checkpoint molecules') are distinct features of T-PLL cells over PB normal T-cells. **C)** Transcript abundaces (array-based GEP) in the 3 isolated normal T-cell subsets (each from 10 healthy donors) compared to 70 T-PLL. **D**) Surface receptors (flow cytometry) in CD3 gates of healthy volunteer derived PBMCs (10 donors) vs 14 T-PLL samples. See **Fig.S3B** for impaired TCR-induced increases in immune-checkpoint marker expression in T-PLL cells.

Figure 4

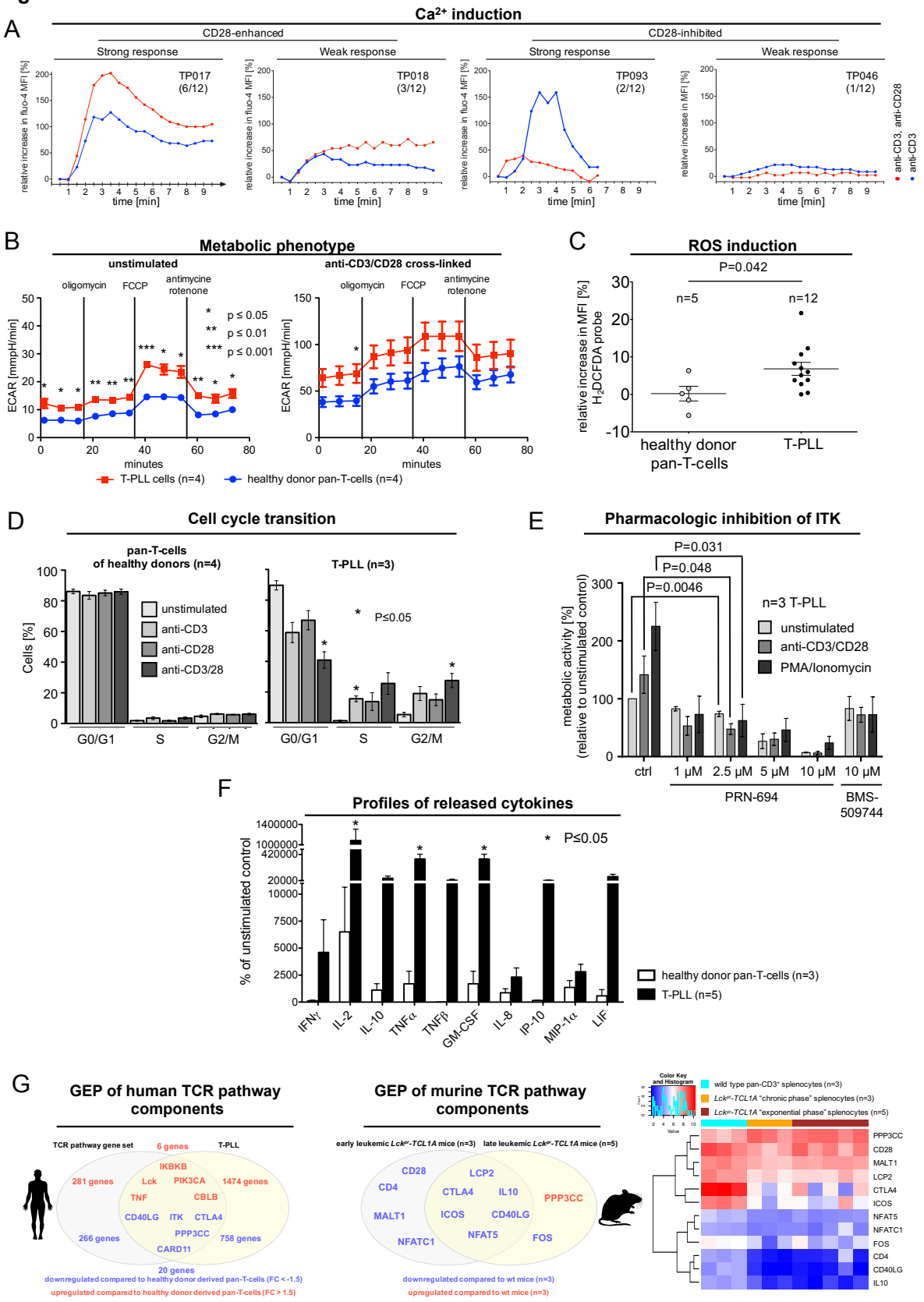
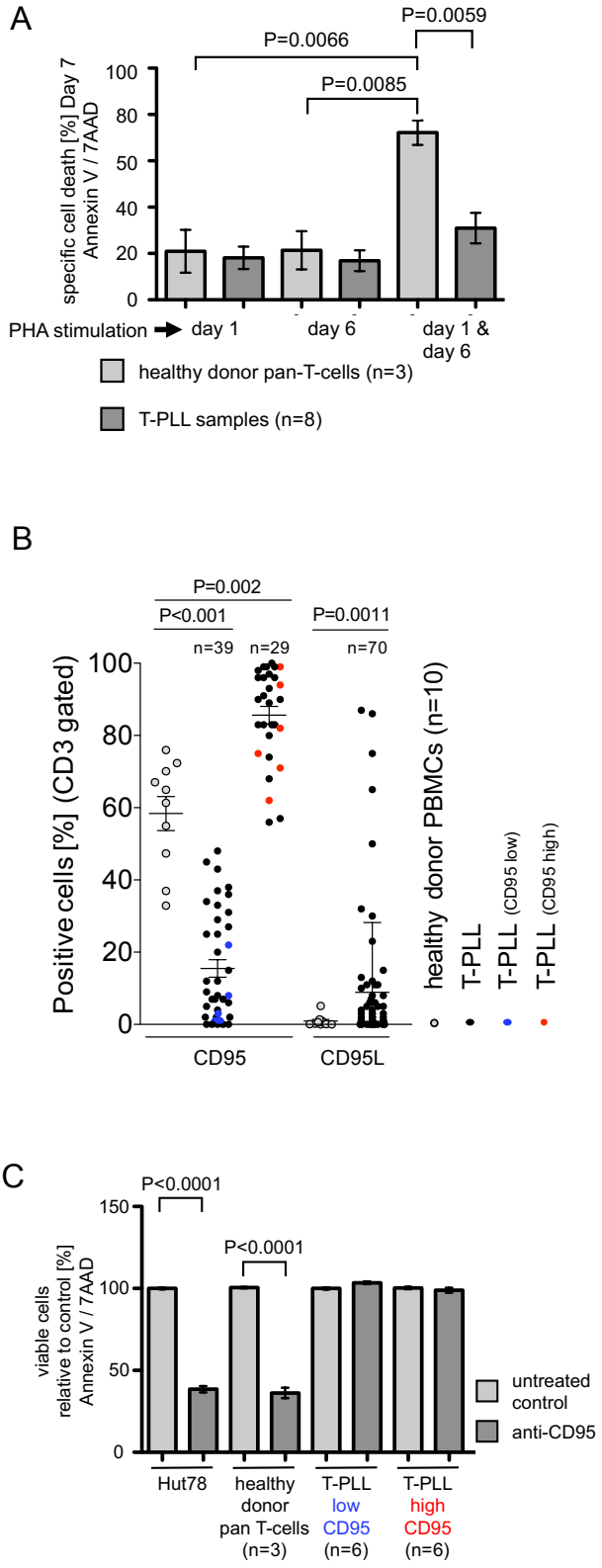

A
Ca2+ induction
CD28-enhanced
CD28-inhibited
Strong response
Weak response
TP017 (6/12)
TP018 (3/12)
TP093 (2/12)
TP046 (1/12)
relative increase in fluo-4 MFI [%]
time [min]
anti-CD3, anti-CD28
anti-CD3
B
Metabolic phenotype
unstimulated
anti-CD3/CD28 cross-linked
oligomycin
FCCP
antimycine rotenone
ECAR [mmpH/min]
minutes
p ≤ 0.05
p ≤ 0.01
p ≤ 0.001
T-PLL cells (n=4)
healthy donor pan-T-cells (n=4)
C
ROS induction
P=0.042
n=5
n=12
relative increase in MFI [%] H2DCFDA probe
healthy donor pan-T-cells
T-PLL
D
Cell cycle transition
pan-T-cells of healthy donors (n=4)
T-PLL (n=3)
Cells [%]
unstimulated
anti-CD3
anti-CD28
anti-CD3/28
P≤0.05
G0/G1
S
G2/M
E
Pharmacologic inhibition of ITK
P=0.031
P=0.048
P=0.0046
n=3 T-PLL
unstimulated
anti-CD3/CD28
PMA/Ionomycin
metabolic activity [%] (relative to unstimulated control)
ctrl
1 µM
2.5 µM
5 µM
10 µM
10 µM
PRN-694
BMS-509744
F
Profiles of released cytokines
P≤0.05
% of unstimulated control
healthy donor pan-T-cells (n=3)
T-PLL (n=5)
IFNγ
IL-2
IL-10
TNFα
TNFβ
GM-CSF
IL-8
IP-10
MIP-1α
LIF
G
GEP of human TCR pathway components
TCR pathway gene set
6 genes
T-PLL
281 genes
1474 genes
266 genes
758 genes
20 genes
IKBKB
Lck
PIK3CA
TNF
CBLB
CD40LG
ITK
CTLA4
PPP3CC
CARD11
downregulated compared to healthy donor derived pan-T-cells (FC < -1.5)
upregulated compared to healthy donor derived pan-T-cells (FC > 1.5)
GEP of murine TCR pathway components
early leukemic Lckpr-TCL1A mice (n=3)
late leukemic Lckpr-TCL1A mice (n=5)
CD28
CD4
MALT1
NFATC1
LCP2
CTLA4
IL10
ICOS
CD40LG
NFAT5
PPP3CC
FOS
downregulated compared to wt mice (n=3)
upregulated compared to wt mice (n=3)
Color Key and Histogram
wild type pan-CD3+ splenocytes (n=3)
Lckpr-TCL1A "chronic phase" splenocytes (n=3)
Lckpr-TCL1A "exponential phase" splenocytes (n=5)
PPP3CC
CD28
MALT1
LCP2
CTLA4
ICOS
NFAT5
NFATC1
FOS
CD4
CD40LG
IL10

Figure 4: A TCR-hyperreactive phenotype is common to T-PLL cells.

A) Ca^{2+} flux upon CD3/28 cross-linking was assessed in primary T-PLL cells (12 cases). Four exemplary cases: TP0173 (strong response, CD28-enhanced), TP093 (strong response, CD28-inhibitory), TP100 (weak response, CD28-enhanced), TP046 (weak response, CD28-inhibitory). **B**) Basal (left) and stimulation-induced (right) extracellular acidification rates (ECARs; by XF96e Extracellular Flux Analyzer, Seahorse Bioscience; see **Fig.S4B** for OCR data) in T-PLL (n=4) and healthy pan-T-cell controls (n=4). T-PLL cells reveal a significantly (P=0.0022) increased basal glycolysis. Upon CD3/28 cross-linking, the increase in ECARs is ≈1.8 times higher in T-PLL than in healthy-donor T-cells (P=0.043). **C**) 2',7'-dichlorodihydrofluorescein diacetate (H_2DCFDA) based measurements of reactive oxygen species (ROS) induction upon TCR activation comparing healthy T-cells (grey dots) to primary T-PLL cases (black dots). A generally higher increase of ROS levels in stimulated T-PLL cells compared to $CD3^+$ pan-T-cells isolated from PB of healthy donors is observed. **D**) The impact of CD3/28 cross-linking on cell-cycle progression was investigated in T-PLL cells (n=3) and healthy T-cells (n=4) using propidium iodide staining and flow cytometry. Upon TCR stimulation, T-PLL cells enter the cell cycle more readily than healthy T-cells (and healthy-donor derived memory (CD45RO+) T-cells, not shown). The combination of CD3 and CD28 engagement has the strongest potential to induce proliferation in T-PLL and controls. **E**) ITK inhibition via PRN-694 (relevant IC50s: ITK - 0.3nM; RLK - 1.4nM; JAK3 - 30nM) in unstimulated and stimulated T-PLL (readout lumiglo). No direct effect on viability at low concentrations, however, induction of an increased level of viability (increased light units compared to control) via CD3/CD28 cross-link (P=0.049) / PMA stimulation (P=0.032) is abolished at 2.5µM PRN-694. Comparable results were achieved for ITK inhibitor BMS-509744 (IC50: ITK - 15nM).[39] **F**) Cytokine secretion of anti-CD3/28-stimulated $CD3^+$ pan-T-cells and T-PLL cells in relation to their unstimulated controls, analyzed on a 45-analyte human cytokine array. Only highly secreted cytokines in TCR-stimulated T-PLL cells are shown. T-PLL cells reveal a stronger response to TCR crosslinking than healthy controls, whereas healthy controls responded better to PMA/Ionomycin stimulation (data not shown). **G**) Genes associated with TCR signaling pathway(s): Differential expression in human T-PLL and in chronic / exponential phase murine leukemic T-cell expansions (*Lck^{pr}-TCL1A* mice).

Figure 5



744
745
746
747

Figure 5: T-PLL cells show a marked defect in the execution of AICD.

A) Apoptosis induction upon repeated TCR activation: Healthy donor PB-derived T-cells (n=3 donors) and PB-derived primary T-PLL cells (n=8 cases) were cultured in the presence of 10U/mL IL-2 and stimulated either once with PHA on day 1 or day 6, or repeatedly on day 1 and day 6 (1 µg/mL). Cells were stained with Annexin V / 7AAD and analyzed by flow cytometry. **B**) CD95L (n=70 T-PLL cases) and CD95 (n=68 T-PLL cases) expression detected by flow cytometry in healthy donor PB-derived T-cell controls (n=10) and PB-derived primary T-PLL cells. The number of CD95L positive cells is heterogeneous but significantly (P=0.0011) increased in T-PLL samples. Expression of CD95 reveals a broader range in the T-PLL samples than in healthy controls allowing an allocation in $CD95^{low}$ (n=39; P<0.001) and $CD95^{high}$ (n=29; P=0.002). **C**) PB-derived primary T-PLL cells (n=12 cases) were investigated for their apoptotic response to agonistic CD95 crosslinking: readout Annexin V / 7AAD staining, flow cytometry analysis. T-PLL samples were classified in groups with low (<50%) and high (≥50%) surface CD95 expression ($CD95^{low}$ (n=6, blue dots marked in 'B' for CD95 expression levels of evaluated cases) and $CD95^{high}$ (n=6, red dots marked in 'B')). T-PLL cells are resistant to extrinsically induced apoptosis via CD95 activation. Positive controls: Hut78 mature T-cell lymphoma line, healthy donor PB-derived pan-T-cells (n=3 donors).

Figure 6

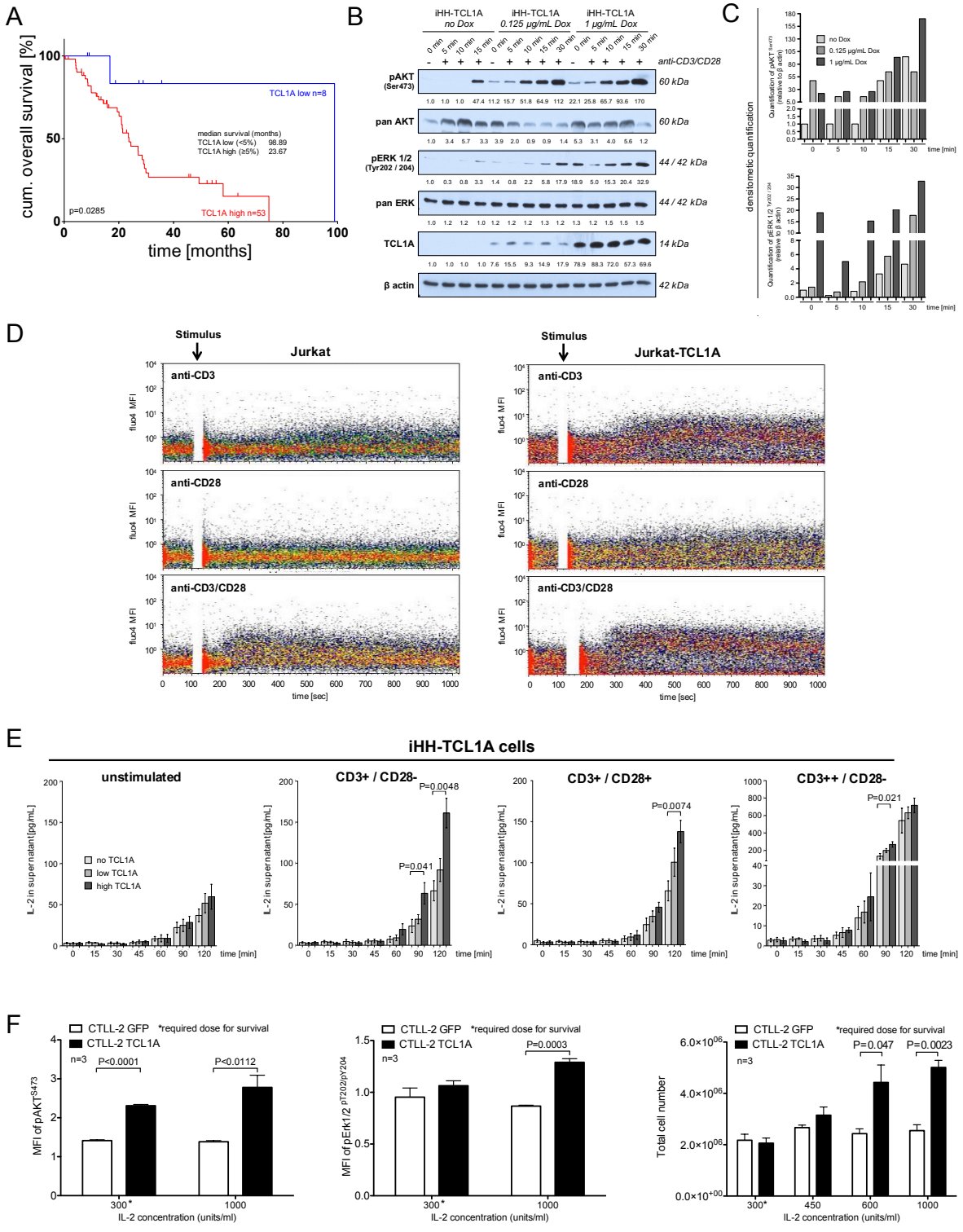
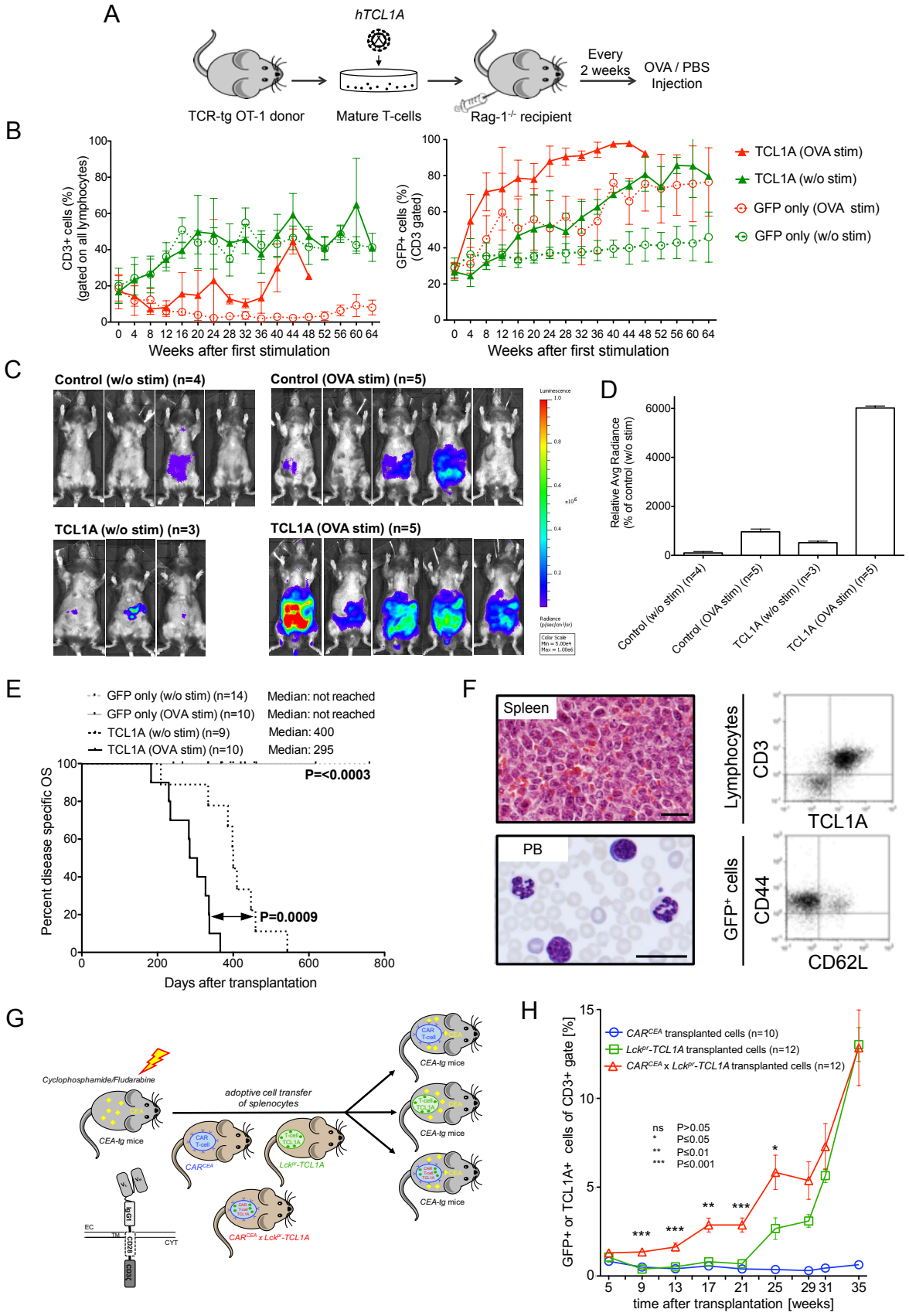

A
cum. overall survival [%]
time [months]
TCL1A low n=8
TCL1A high n=53
median survival (months)
TCL1A low (<5%) 98.89
TCL1A high (≥5%) 23.67
p=0.0285
B
iHH-TCL1A no Dox
iHH-TCL1A 0.125 µg/mL Dox
iHH-TCL1A 1 µg/mL Dox
anti-CD3/CD28
pAKT (Ser473)
pan AKT
pERK 1/2 (Tyr202 / 204)
pan ERK
TCL1A
β actin
60 kDa
44 / 42 kDa
14 kDa
42 kDa
C
densitometric quantification
no Dox
0.125 µg/mL Dox
1 µg/mL Dox
time [min]
D
Stimulus
Jurkat
Jurkat-TCL1A
anti-CD3
anti-CD28
anti-CD3/CD28
fluo4 MFI
time [sec]
E
iHH-TCL1A cells
unstimulated
CD3+ / CD28-
CD3+ / CD28+
CD3++ / CD28-
IL-2 in supernatant [pg/mL]
no TCL1A
low TCL1A
high TCL1A
P=0.0048
P=0.041
P=0.0074
P=0.021
time [min]
F
CTLL-2 GFP
CTLL-2 TCL1A
*required dose for survival
n=3
P<0.0001
P<0.0112
P=0.0003
P=0.047
P=0.0023
MFI of pAKTS473
MFI of pErk1/2 pT202/pY204
Total cell number
IL-2 concentration (units/ml)

Figure 6: TCL1A mediates enhanced TCR-downstream signaling.

A) Kaplan-Meier plot of disease-specific overall survival (log-rank test, time from diagnosis to event) of uniformly treated T-PLL patients stratified by TCL1A protein expression (flow cytometry derived dataset; TCL1A low <5% pos. cells; TCL1A high ≥5% pos. cells). **B**) Enforced low and high level TCL1A expression in HH T-cell leukemia (doxycycline-inducible iHH-TCL1A cell line) expedites and enhances phosphorylation of AKT (pAKT) and pERK1/2 upon CD3/CD28 cross-linking in a concentration dependent manner. The similar expression of surface CD3/CD28 in both HH sublines is shown in **Fig.S6B**. **C**) Desitometric quantification of immunoblot results from 'A'. **D**) Jurkat and Jurkat-TCL1A cells were loaded with the Ca^{2+} indicator dye fluo-4 and were stimulated as indicated with either 10 µg/mL soluble anti-CD3 antibody (OKT3), 20 µg/mL anti-CD28 antibody (15E8) or in combination at t=0. Each antibody was cross-linked using 10 µg/mL anti-IgG antibodies. Changes in intracellular Ca^{2+} levels were measured over time on a single cell level. Isolated anti-CD3 antibody stimulation leads to an increased Ca^{2+} signal in Jurkat-TCL1A cells compared to $TCL1A^{-}$ Jurkat cells. Cross-linking with anti-CD28 antibody alone leads only to a minor Ca^{2+} flux in TCL1A-positive Jurkat cells. Co-stimulation with anti-CD3/CD28 antibodies causes a strong and fast Ca^{2+} signal in Jurkat cells that decreases over time. In comparison, Jurkat-TCL1A cells reveal a stronger extended Ca^{2+} flux. **E**) IL-2 secretion in response to TCL1A modulated TCR activation. Multidimensional titration of TCL1A expression / TCR activation in iHH-TCL1A cell line system via combinations of TCL1A (no, low, high doxycycline), CD3 (low 0.1 µg/mL, high 1.0 µg/mL), and CD28 (low 0.2 µg/mL, high 2.0 µg/mL) crosslinking antibodies. Readout: IL-2 ELISA. TCL1A increases IL-2 secretion upon submaximal levels of TCR stimulation via low dose CD3 antibody (1 µg/mL) disregarding of additional CD28 activation. **F)** IL-2 dependent murine CTLL-2 cell system: CTLL-2 cells transduced with TCL1A or GFP control were treated with IL-2. Left: Phosphorylation of AKT^{S473} and $ERK1/2^{Thr202/Tyr204}$ as detected by flow cytometry. Phospho-kinase responses are generally higher in TCL1A expressing CTLL-2 cells compared to GFP only control cell line. Right: Viablity (total cell number) in response to increasing IL-2 concentrations. Ectopic TCL1A expression does not override IL-2 dependence of CTLL-2 cells, but enables CTLL-2 cells to execute a higher proliferative response upon stimulation with increasing IL-2 concentrations.

Figure 7



818
819

Figure 7: Modelling of chronic TCR-stimulation confers competitive growth benefits to TCL1A expressing T-cells.

A) Schematic outline of the experimental procedure. Cell suspensions from spleens and lymph nodes of OT-1 mice (tg for a monoclonal OVA-peptide responsive TCR) were retrovirally transduced with a TCL1A-GFP or a GFP only cDNA construct and transplanted into lymphodeficient $RAG1^{-/-}$ recipients. Recipient mice received intraperitoneal injections of the OVA peptide (aa 257-264) or PBS every 2 weeks. **B**) Blood samples were taken from unstimulated (green, w/o stim) and stimulated (red, OVA stim) GFP only (circle) and TCL1A (triangle) OT-1 T-cell recipient mice every 4 weeks and analyzed by flow cytometry. Mean percentage of $CD3^+$ cells gated on live cells (left), and GFP^+ cells gated on $CD3^+$ cells (right) was compared between different cohorts throughout the observation time. Experiment was started with 5 mice per group. Mean with SEM. **C**) Unstimulated and stimulated recipient mice of TCL1A-Luc or T-Sapphire-Luc (control) transduced OT-1 T-cells were imaged 12 weeks after the first OVA injection. All pseudocolor images were adjusted to the same threshold. **D**) Quantification of bioluminescence imaging signal intensities in each cohort. Signal intensities (average radiance (photons/s/cm^2/sr)) are shown as relative values setting untreated controls to 100. **E**) Mean overall survival (OS) of unstimulated and stimulated TCL1A-GFP or GFP only transduced OT-1 T-cell recipient mice. **F**) Histopathologic and immunophenotypic characterization of tumors induced by TCL1A-transduced OT-1 T-cells in immunodeficient recipient mice. **G**) Experimental procedure: splenocytes isolated from *Lck^{pr}-TCL1A*, *CAR^{CEA}*, and *CAR^{CEA}xLck^{pr}-TCL1A* mice were transplanted into lympho-depleted (Cyclophosphamide/Fludarabine) *CEA-tg* mice. **H**) Blood samples were taken from *CEA-tg* recipients of *CAR^{CEA}* (blue), *Lck^{pr}-TCL1A* (green), or *CAR^{CEA}xLck^{pr}-TCL1A* (red) tg cells every 4 weeks and analyzed by flow cytometry for repopulation of GFP^+ (CAR) or $TCL1A^+$ cells (gated on $CD3^+$ cells). Statistical significance is shown for *Lck^{pr}-TCL1A* and *CAR^{CEA}xLck^{pr}-TCL1A* recipient mice.

SUPPLEMENTARY METHODS, TABLES, AND FIGURES

CONTENTS

1. Supplementary Figures

Figure S1

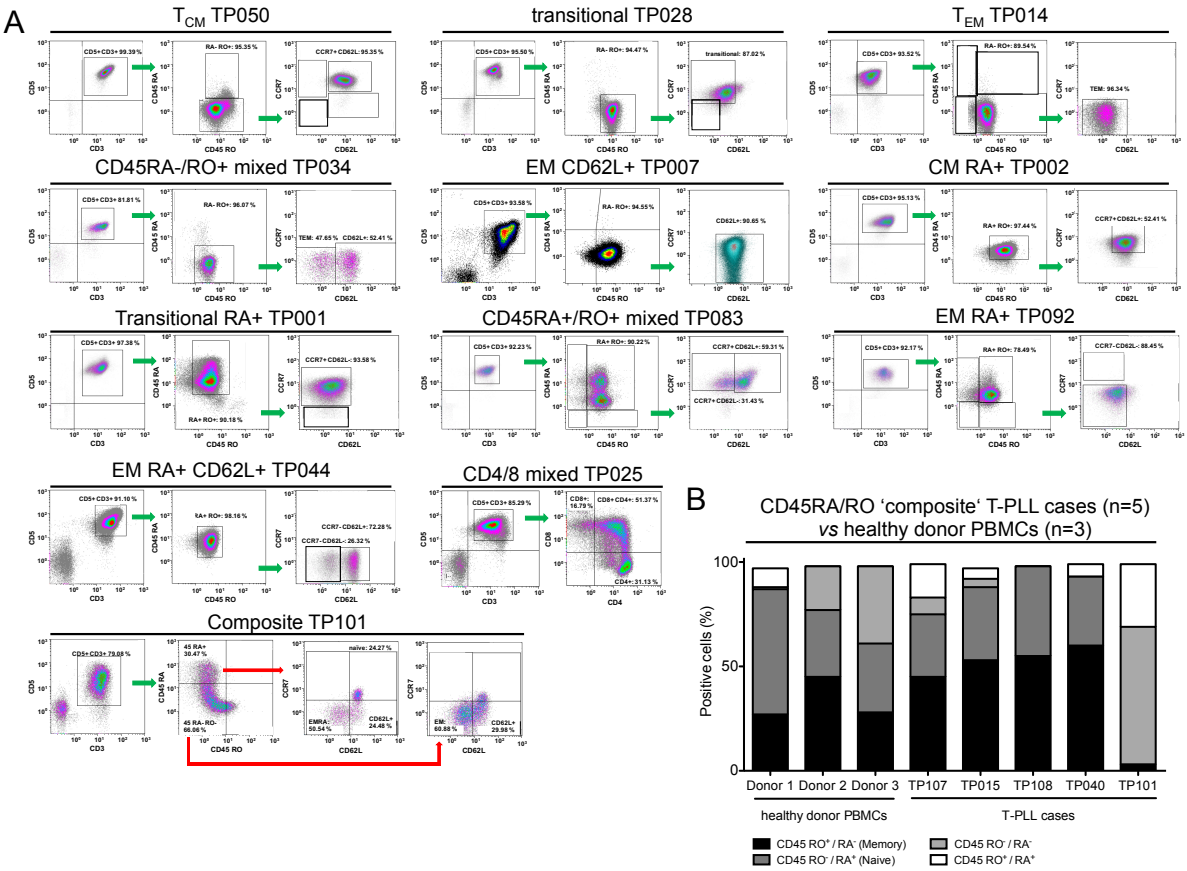


Figure S1: The (central) memory-like T-cell phenotype of T-PLL cells - immunophenotypes.

A) Flow cytometric analysis of CD45RA/CD45RO, CCR7, and CD62L surface expression in exemplary T-PLL cases (data supplementing **Fig.1B**, gating strategy is illustrated via red arrows). For composition / surface marker combinations of single categories please refer to **TableS2. B**) Flow cytometric analysis of 5 T-PLL cases (right) showing a 'composite' expression of CD45RA/RO resembling healthy PB derived distributions of memory / naïve surface makers (left). **C**) SPADE[1] analyses of memory markers in a subsets of T-PLL cases (3 batches analyzed). Batches were defined according to flow cytometry analyses per individuel sample that were performed as one set at the same day using exactly the same flow-cytometer settings.

Figure S2

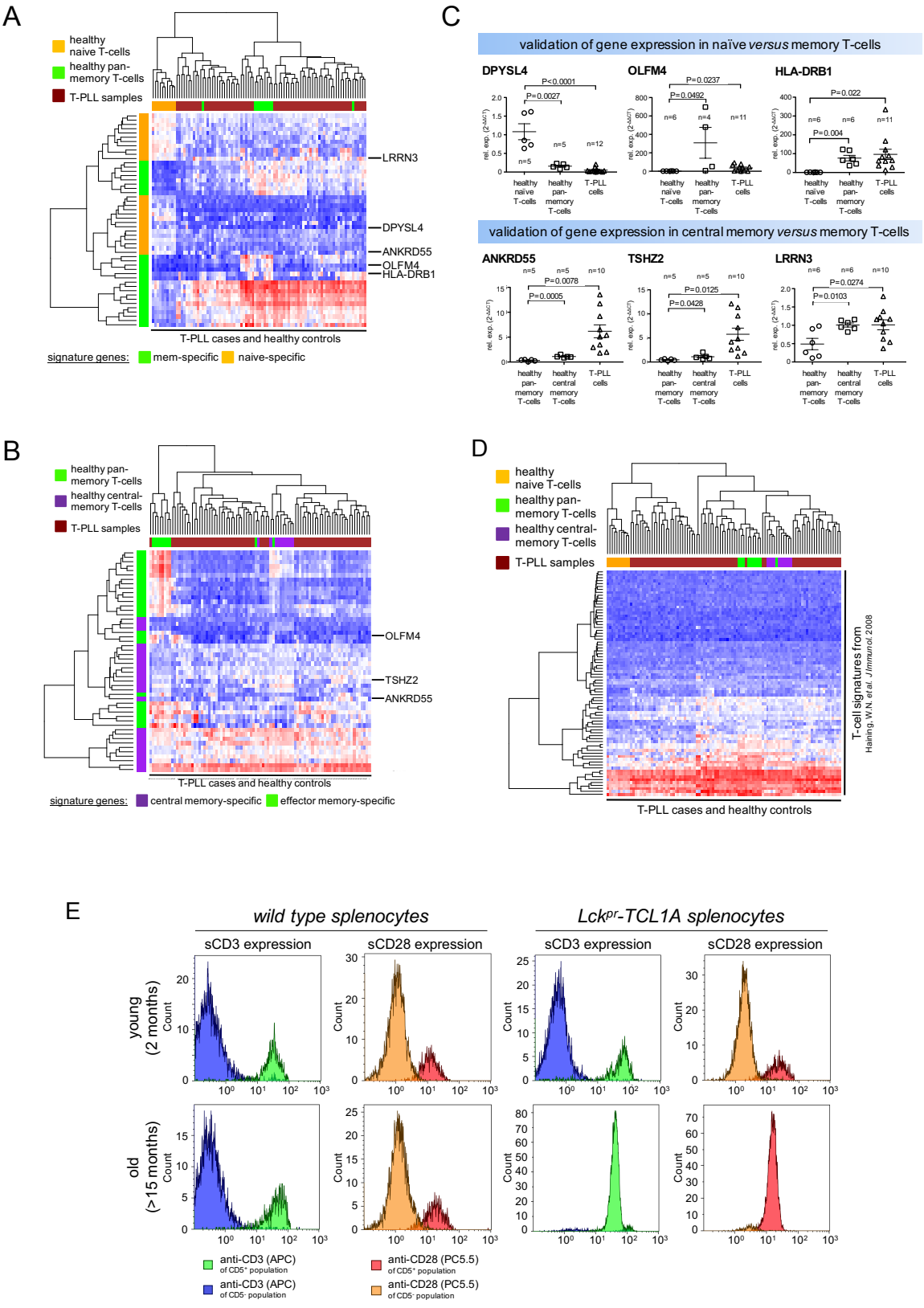

A
healthy naive T-cells
healthy pan-memory T-cells
T-PLL samples
LRRN3
DPYSL4
ANKRD55
OLFM4
HLA-DRB1
T-PLL cases and healthy controls
signature genes: mem-specific naive-specific
C
validation of gene expression in naïve versus memory T-cells
DPYSL4
OLFM4
HLA-DRB1
validation of gene expression in central memory versus memory T-cells
ANKRD55
TSHZ2
LRRN3
B
healthy pan-memory T-cells
healthy central-memory T-cells
T-PLL samples
OLFM4
TSHZ2
ANKRD55
T-PLL cases and healthy controls
signature genes: central memory-specific effector memory-specific
D
healthy naive T-cells
healthy pan-memory T-cells
healthy central-memory T-cells
T-PLL samples
T-cell signatures from Haining, W.N. et al. J Immunol. 2008
T-PLL cases and healthy controls
E
wild type splenocytes
Lckpr-TCL1A splenocytes
sCD3 expression
sCD28 expression
young (2 months)
old (>15 months)
Count
anti-CD3 (APC) of CD5+ population
anti-CD3 (APC) of CD5- population
anti-CD28 (PC5.5) of CD5+ population
anti-CD28 (PC5.5) of CD5- population

Figure S2: The (central) memory-like T-cell phenotype of T-PLL cells – gene expression profiling (GEP).

A, B) Heatmaps showing the expression (red=up-regulation, blue=down-regulation) of (**A**) (pan-) memory *vs* naïve and of (**B**) effector-memory (EM) *vs* central-momory (CM) signature genes in T-PLL and healthy T-cell samples (compare PCA analysis in **Fig.1D,E**). **C**) Confirmations by qRT-PCR: mRNA levels of some of the memory-*vs*-naïve best-distinguishing signature genes (*OLFM4*, *HLA-DRB1*, and *DPYSL4*) in T-PLLs cells or naïve and memory T-cells of healthy donors. Differential gene-expression of the memory-*vs*-central-memory signature genes *ANKRD55*, *TSHZ2*, and *LRRN3* was confirmed in T-PLL cells or central-memory and memory T-cells of healthy donors. **D**) The expression of a T-cell subset specific gene signature as reported by Haining and colleagues[2] was evaluated in the 70 GEPs of primary T-PLL cells and in PB isolated healthy donor-derived naïve *vs* pan-memory *vs* CM T-cells (n=10 donors each (unsupervised hierarchical clustering). Again, the majority of T-PLL cases showed a gene expression most similar to memory T-cells / CM T-cells. **E**) **A-C**) Surface marker expression of CD3 and CD28 (flow cytometry) in murine spleen-derived $CD5^+$ T-cells from young (2 months) and old (>15 months) *Lck*^{pr}-*TCL1A* mice vs age-matched C57/B6J wild-type controls.

Figure S3

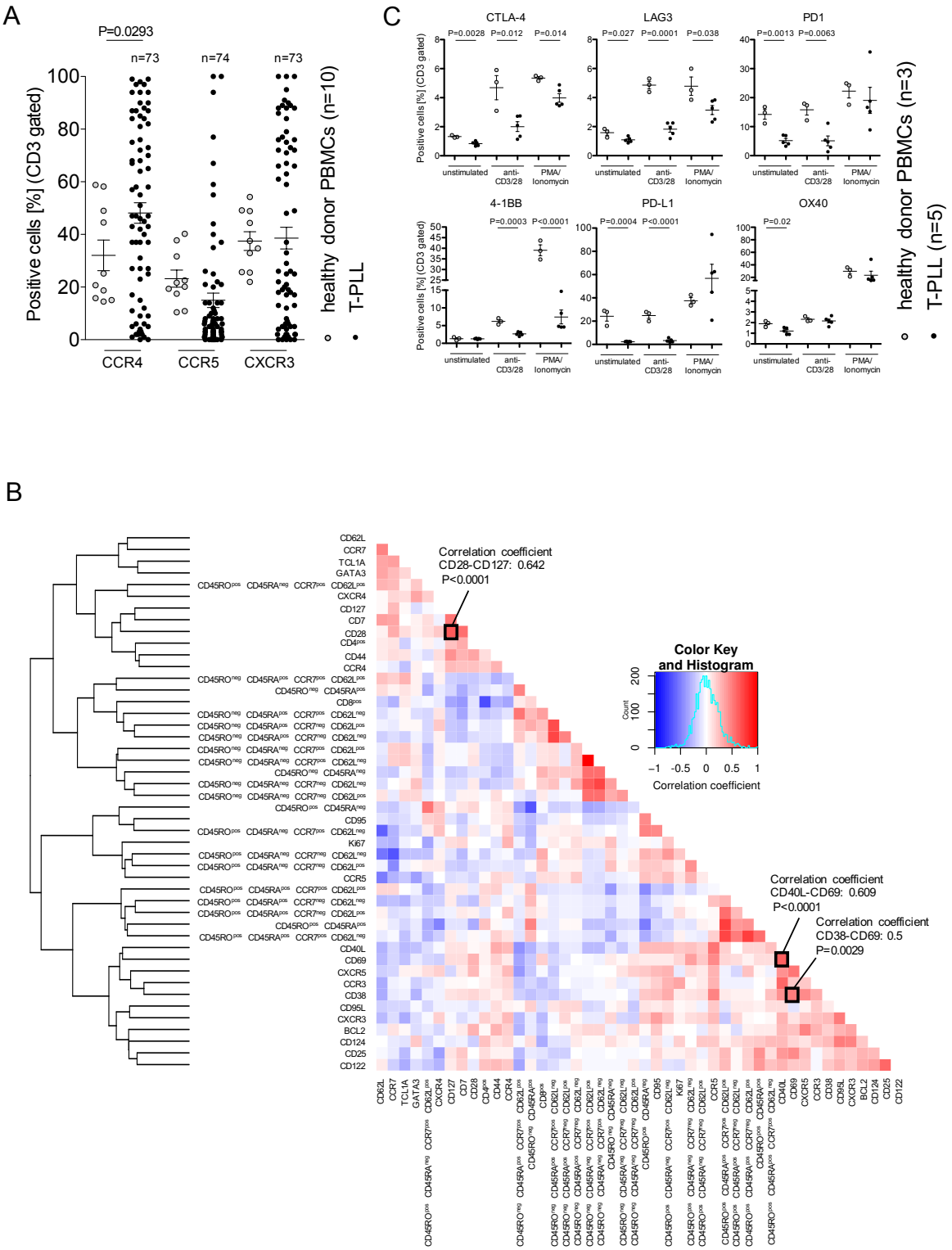

A
P=0.0293
n=73
n=74
n=73
Positive cells [%] (CD3 gated)
CCR4
CCR5
CXCR3
healthy donor PBMCs (n=10)
T-PLL
C
CTLA-4
LAG3
PD1
4-1BB
PD-L1
OX40
unstimulated
anti-CD3/28
PMA/Ionomycin
healthy donor PBMCs (n=3)
T-PLL (n=5)
B
Correlation coefficient
CD28-CD127: 0.642
P<0.0001
Color Key and Histogram
Correlation coefficient
CD40L-CD69: 0.609
P<0.0001
Correlation coefficient
CD38-CD69: 0.5
P=0.0029

Figure S3: Immunophenotypic profiles of T-PLL cells.

A) Healthy PB T-cells (n=10 donors) and PB-derived primary T-PLL cells (n=79 cases) were analyzed by flow cytometry for the expression of chemokine receptors CCR4, CCR5, and CXCR3 (data supplementing **Fig.2A**). **B**) Correlation-matrix (color-coded coefficients): expression of the markers or their informative combinations (subset-defining) quantified via flow cytometry (% positive cells) was correlated across cases of the cohort of 79 T-PLL. Noteworthy positive correlations were: CD28 with CD127 expression (ρ=0.642; P=4.87x10^{-7}), of the activation markers CD40L and CD69 (ρ=0.609; P=0.0283), and of CD38 with CD69 (ρ=0.5; P=8.68Ex10^{-4}). Specific memory marker immunophenotypes (CD45RA/RO, CCR7, CD62L; see **Fig.1B** and **TableS3**) did not correlate with a specific expression pattern of other investigated markers. **C**) Flow-cytometric analysis of immune-checkpoint receptors upon anti-CD3/CD28 and PMA/Ionomycin stimulated primary T-PLL cells *vs* healthy donor-derived CD3 pan-T-cell controls. A certain anergy of T-PLL cells to TCR-induced upregulation of negative auto-regulatory receptors is observed.

Figure S4

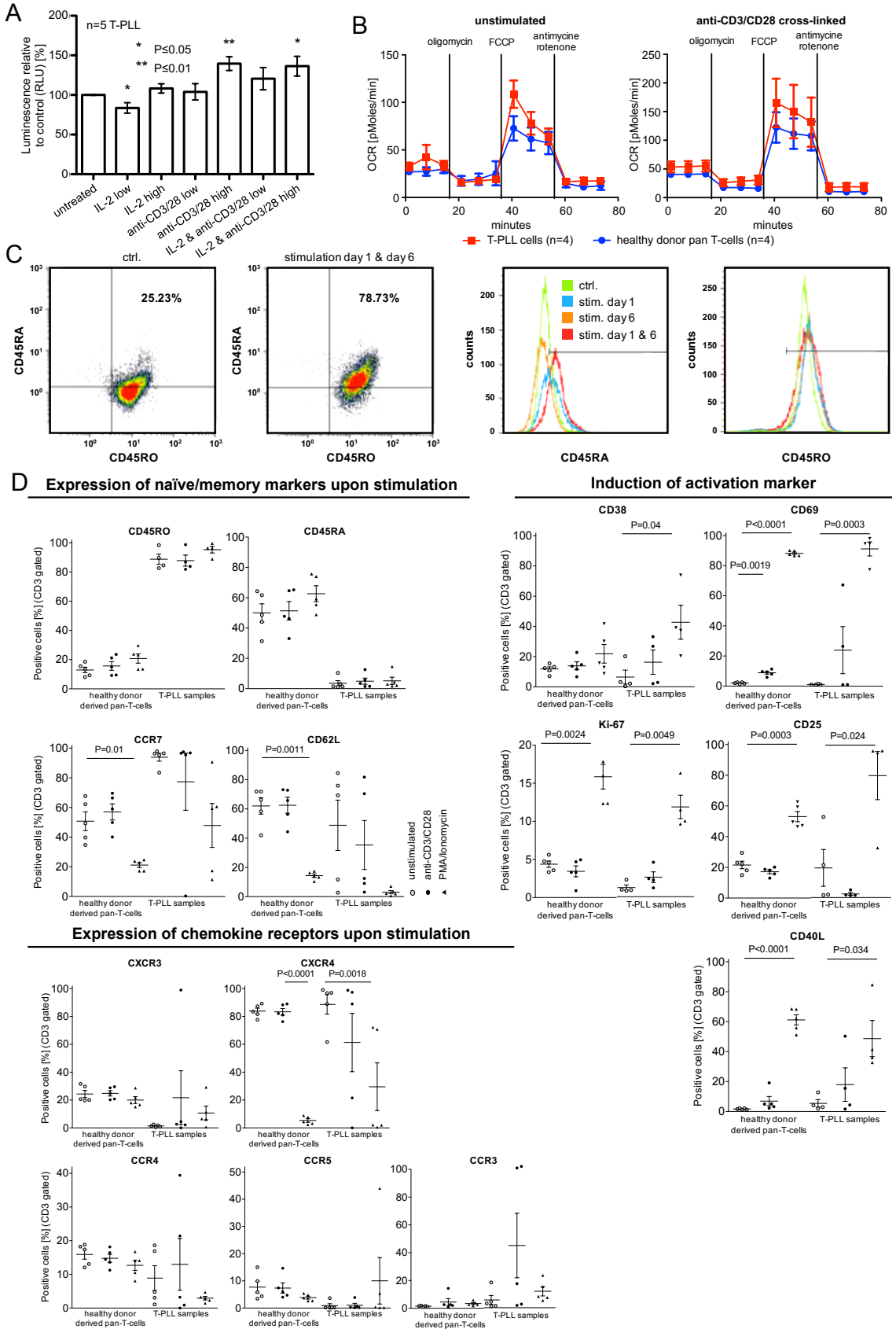

A
n=5 T-PLL
* P≤0.05
** P≤0.01
Luminescence relative to control (RLU) [%]
untreated
IL-2 low
IL-2 high
anti-CD3/28 low
anti-CD3/28 high
IL-2 & anti-CD3/28 low
IL-2 & anti-CD3/28 high
B
unstimulated
anti-CD3/CD28 cross-linked
oligomycin
FCCP
antimycine rotenone
OCR [pMoles/min]
minutes
T-PLL cells (n=4)
healthy donor pan T-cells (n=4)
C
ctrl.
stimulation day 1 & day 6
25.23%
78.73%
CD45RA
CD45RO
counts
ctrl.
stim. day 1
stim. day 6
stim. day 1 & 6
D
Expression of naïve/memory markers upon stimulation
Induction of activation marker
CD45RO
CD45RA
CD38
CD69
CCR7
CD62L
Ki-67
CD25
CD40L
P=0.04
P<0.0001
P=0.0003
P=0.0019
P=0.01
P=0.0011
P=0.0024
P=0.0049
P=0.0003
P=0.024
P<0.0001
P=0.034
Positive cells [%] (CD3 gated)
healthy donor derived pan-T-cells
T-PLL samples
unstimulated
anti-CD3/CD28
PMA/Ionomycin
Expression of chemokine receptors upon stimulation
CXCR3
CXCR4
P<0.0001
P=0.0018
CCR4
CCR5
CCR3

Figure S4: TCR activation triggers proliferation of T-PLL cells.

A) Anti-CD3/CD28 and IL-2 stimulated primary T-PLL cells as suspension cultures. Viability was assessed in T-PLL cells (n=5) using the CellTiter-Glo® luminescent assay by quantifying ATP. All investigated samples surface CD3, CD28, CD25, and CD122 with the exception of one, which revealed no expression of the IL-2 receptor chains CD25 and CD122. Cells were stimulated with different combinations of IL-2 (low 2.5 ng/mL, high 25 ng/mL) and cross-linking antibodies against CD3 (low 0.1 µg/mL, high 1 µg/mL) and CD28 (low 0.2 µg/mL, high 2 µg/mL). Viability was measured after 24 h. **B**) Basal oxygen consumption rate (OCR) as an indicator for mitochondrial respiration was assessed in T-PLL cells (n=4) and healthy donor-derived pan-T-cell controls (n=4). Baseline respiration is slightly increased in T-PLL cells as compared to healthy controls prior to stimulation (not significant). CD3/28 cross-linking leads to an increased OCR with slightly higher levels on T-PLL cells (not significant). **C, D**) The surface expression of CD45RA and activation markers (CD25, CD38, CD69, and Ki-67) increased in T-PLL cells upon stimulation with anti-CD3/28 antibodies and PMA/Ionomycin (24hrs, detection via flow cytometry), comparable to healthy donor-derived pan-T-cells (n=5 donors) and healthy memory (CD45RO+) T-cells (data not shown).

Figure S5

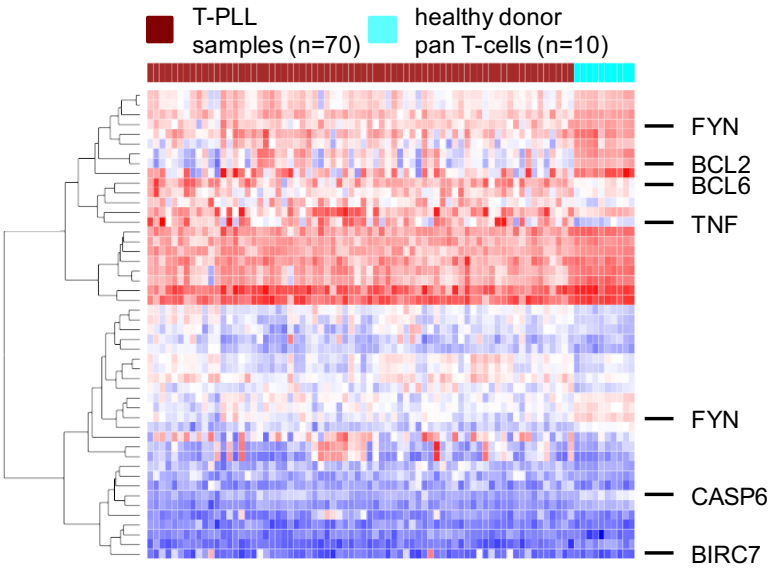


Figure S5: Altered gene expression of apoptosis regulators in T-PLL.

Transcripts of genes regulating apoptotic pathways were found to be differently expressed in T-PLL cells (n=70) compared to healthy donor pan-T-cell controls (n=10). The heatmap shows highly expressed genes in red and downregulated genes in blue. The annotated genes (right) represent the most informative highly differentially expressed genes (P<0.05).

Figure S6

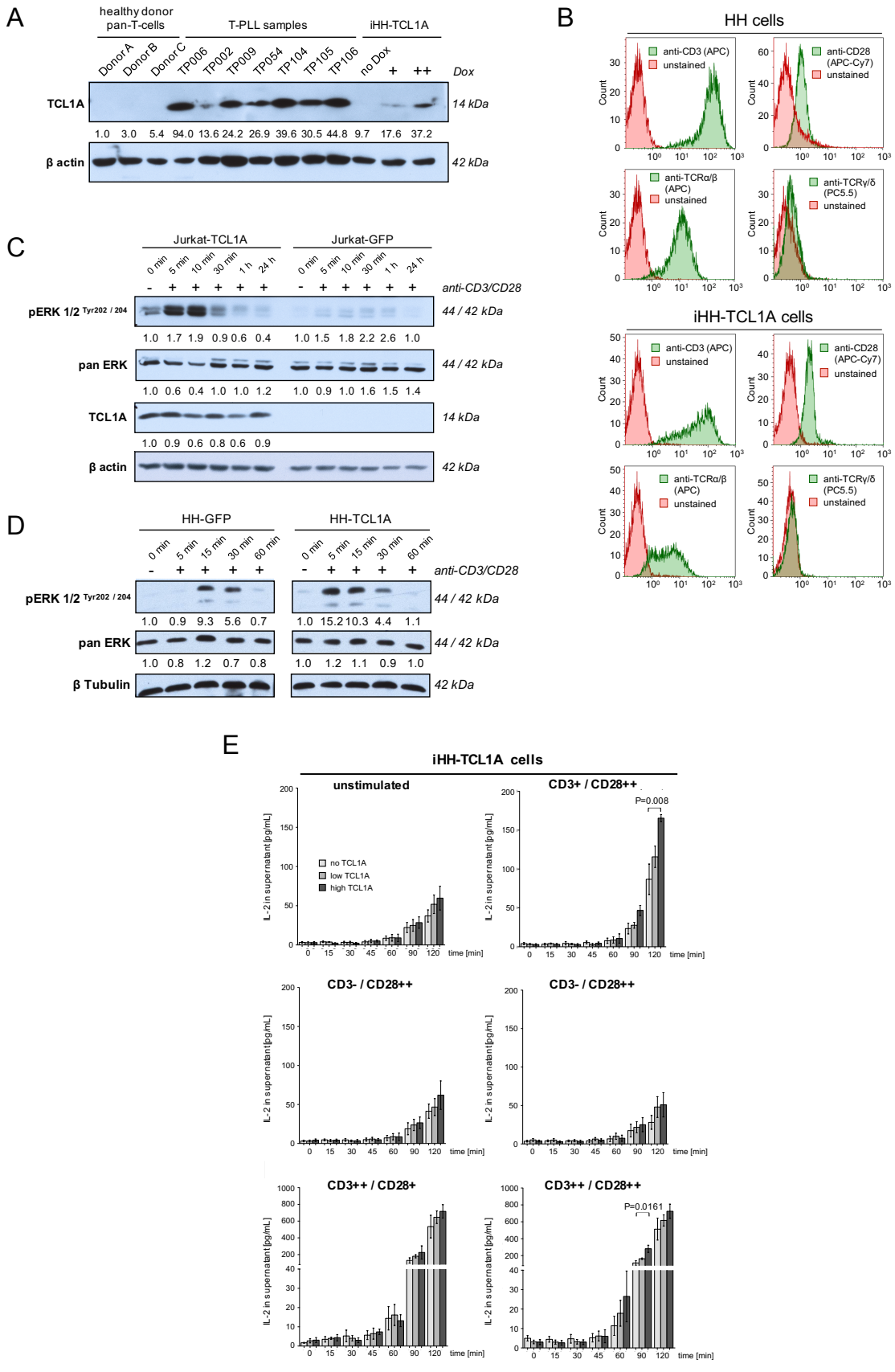

A
healthy donor pan-T-cells
T-PLL samples
iHH-TCL1A
DonorA DonorB DonorC TP006 TP002 TP009 TP054 TP104 TP105 TP106 no Dox + ++ Dox
TCL1A
14 kDa
1.0 3.0 5.4 94.0 13.6 24.2 26.9 39.6 30.5 44.8 9.7 17.6 37.2
β actin
42 kDa
B
HH cells
anti-CD3 (APC)
unstained
anti-CD28 (APC-Cy7)
anti-TCRα/β (APC)
anti-TCRγ/δ (PC5.5)
Count
iHH-TCL1A cells
C
Jurkat-TCL1A
Jurkat-GFP
0 min 5 min 10 min 30 min 1 h 24 h
anti-CD3/CD28
pERK 1/2 Tyr202 / 204
44 / 42 kDa
1.0 1.7 1.9 0.9 0.6 0.4 1.0 1.5 1.8 2.2 2.6 1.0
pan ERK
1.0 0.6 0.4 1.0 1.0 1.2 1.0 0.9 1.0 1.6 1.5 1.4
TCL1A
14 kDa
1.0 0.9 0.6 0.8 0.6 0.9
β actin
42 kDa
D
HH-GFP
HH-TCL1A
0 min 5 min 15 min 30 min 60 min
anti-CD3/CD28
pERK 1/2 Tyr202 / 204
44 / 42 kDa
1.0 0.9 9.3 5.6 0.7 1.0 15.2 10.3 4.4 1.1
pan ERK
44 / 42 kDa
1.0 0.8 1.2 0.7 0.8 1.0 1.2 1.1 0.9 1.0
β Tubulin
42 kDa
E
iHH-TCL1A cells
unstimulated
CD3+ / CD28++
P=0.008
no TCL1A
low TCL1A
high TCL1A
IL-2 in supernatant [pg/mL]
time [min]
CD3- / CD28++
CD3++ / CD28+
CD3++ / CD28++
P=0.0161

Figure S6: TCL1A enhances TCR triggered signaling responses.

A) TCL1A protein expression levels were assessed in the doxycycline-inducible iHH-TCL1A cell line compared to PB-derived primary T-PLL cells. TCL1A expression is even higher in primary cases compared to the iHH-TCL1A system. **B**) CD3 and CD28 surface expression levels in HH and iHH-TCL1A cell line as detected by flowcytometry. **C, D**) TCR responses in stably TCL1A expressing cell lines: The Jurkat T-cell line / the HH T-cell line were stably transfected with either TCL1A or GFP as a control. Transfected cells were stimulated using anti-CD3 (10 µg/mL) and anti-CD28 (20 µg/mL) antibodies and phosphorylation of effector kinase ERK1/2 was investigated by immunoblotting. Phosphorylation of ERK1/2 (pERK1/2) is accelerated and enhanced in the context of overexpressed TCL1A. **E**) Additional conditions from the TCL1A / CD3 / CD28 titration experiment in iHH-TCL1A cells (data supplementing **Fig.6D**). Different concentrations and combinations of TCL1A (no, low, high), and anti-CD3 (low 0.1 µg/mL, high 1.0 µg/mL), and anti-CD28 (low 0.2 µg/mL, high 2.0 µg/mL) crosslinking antibodies were used. Readout: IL-2 release as detected by ELISA.

Figure S7

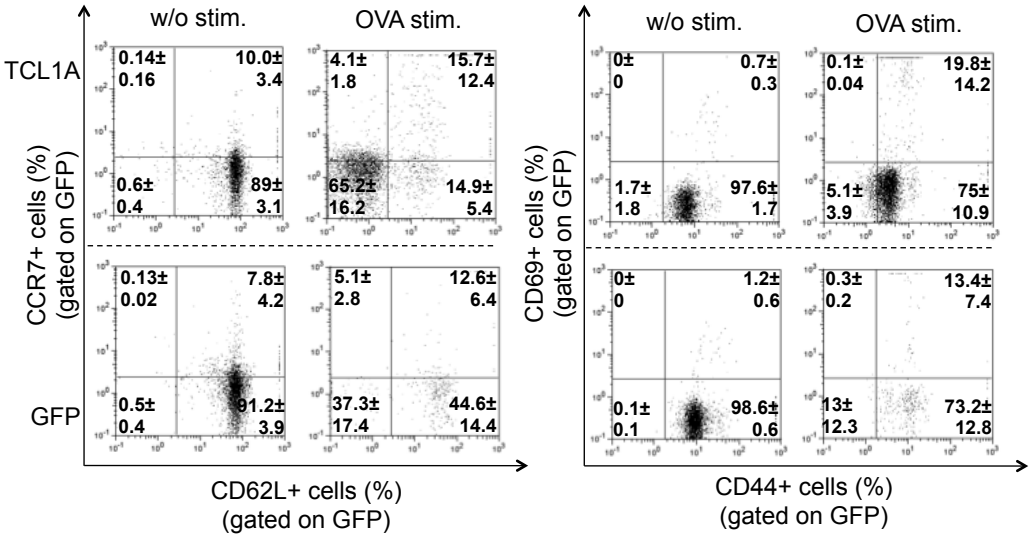


Figure S7: Immunophenotypic characterization of transduced OT-1 T-cells in peripheral blood of recipient mice.

Mature OT-1 T-cells carrying monoclonal TCR's that specifically recognize ovalbumin (OVA) were retrovirally transduced with a TCL1A-GFP or a GFP-only cDNA construct and transplated into immunodeficient $RAG^{-/-}$ mice. Mice were repeatedly stimulated *in vivo* with OVA-peptides (experimental strategy in **Fig.7A**). GFP^{+} cells in the PB of PBS-injected (w/o stim) and OVA-stimulated (OVA-stim) T-cell recipient mice (GFP-only or TCL1A OT-1 T-cells) were characterized by flow cytometry for memory marker expression CCR7, CD62L, CD69, and CD44 at 36 weeks after transplantation. Plots show mean percentages with standard deviations (SD) of sub-gates of 5 analyzed animals.

2. Supplementary Tables

Table S1: Cohort of analyzed T-PLL cases.

(see Excel file)

A total of 105 patients with T-PLL (58 men; 38 women; 9 n/a) were diagnosed at a median age of 66 years. The rearrangements inv(14)(q11;q32) and t(14;14)(q11;q32) were detected in 85% (n=57/67) and TCL1A protein expression by flow cytometry in 79% (n=64/81; >5% TCLA positive cells) of analyzed cases. The immunophenotype $CD4^+/CD8^-$ was present in 59% (n=57/97) of analyzed cases, $CD4^+/CD8^+$ in 25% (n=24/97), $CD8^+/CD4^-$ in 15% (n=15/97), and $CD8^-/CD4^-$ in 1% (n=1/97). The minor differences to the reported frequencies in the Result section orginiated from the variable cohort size of cases that were analyzed for the multi-analyte memory- / activation marker set (n=79) as compared to all 96 cases with available CD45RO/RA data.

Table S2: Surface marker expression defining naïve / memory T-cell like subsets in T-PLL.

Categories shaded in light blue represent T-cell subsets that are observed in healthy individuals[3], categories in dark blue represent aberrant surface marker combinations newly identified in primary T-PLL.

Category	CD45RA	CD45RO	CCR7	CD62L	T-PLL cases n	%
CD45RA+ / CD45RO-					**7**	**9**
Naïve	+	-	+	+	**7**	**9**
CD45RA- / CD45RO+					**46**	**58**
Effector Memory	-	+	-	-	**5**	**6**
Effector Memory CD62L$^+$	-	+	-	+	**1**	**1**
Central Memory	-	+	+	+	**25**	**32**
Transitional	-	+	+	-	**7**	**9**
Memory mixed	-	+	+/-	+/-	**2**	**3**
EMRA	+	-	-	-	**1**	**1**
n/a					**5**	**6**
CD45RA+ / CD45RO+					**19**	**24**
Effector Memory (CD45RA+)	+	+	-	-	**1**	**1**
Effector Memory (CD45RA+) CD62L$^+$	+	+	-	+	**1**	**1**
Central Memory (CD45RA+)	+	+	+	+	**10**	**13**
Transitional (CD45RA+)	+	+	+	-	**3**	**4**
memory mixed (CD45RA+)	+	+	+/-	+/-	**2**	**3**
n/a					**2**	**3**
CD45RA / CD45RO composite	+/-	+/-	+/-	+/-	**5**	**6**
CD45RA- / CD45RO-	-	-	+/-	+/-	**3**	**4**

Table S3: Gene expression signatures defining T-cell differentiation subtypes. 25 genes that represent specific gene expression signatures for naïve (out of a total of n=2078), (pan-) memory (n=2316), CM (n=185) and EM (n=634) T-cells (excluding unannotated probes; P<0.05; q<0.1). They were received by comparative GEP analysis between listed healthy T-cell subsets (see supplementary methods for details).

TOP25 Mem *vs* Naïve (Mem-TC Signature)		**TOP25 Naïve *vs* Mem (Naïve-TC Signature)**		**TOP25 Mem *vs* CM (EM-TC Signature)**		**TOP25 CM *vs* Mem (CM-TC Signature)**	
Gene Symbol (Entrez ID)	**FC**	**Gene Symbol (Entrez ID)**	**FC**	**Gene Symbol (Entrez ID)**	**FC**	**Gene Symbol (Entrez ID)**	**FC**
CAMP (820)	58.15	KRT72 (140807)	14.46	OLFM4 (10562)	10.69	ANKRD55 (79722)	2.63
HLA-DRA (3122)	43.15	MMP28 (79148)	9.57	TMEM158 (25907)	10.19	TSHZ2 (128553)	1.86
LYZ (4069)	42.91	KRT73 (319101)	8.51	CAMP (820)	9.68	SLC22A23 (63027)	1.8
FCER1A (2205)	39.97	EDAR (10931)	6.38	GZMH (2999)	9.59	PRO0628 (29053)	1.79
HLA-DQA1 (3117)	30.26	DACT1 (51339)	5.87	GNG11 (2791)	9.18	TXK (7294)	1.77
FCN1 (2219)	29.83	ADTRP (84830)	5.61	CA2 (760)	8.96	MEOX1 (4222)	1.76
OLFM4 (10562)	27.69	NOG (9241)	5.11	GP9 (2815)	8.91	EPHX2 (2053)	1.75
NAPSB (256236)	27.09	FHIT (2272)	4.93	FGFBP2 (83888)	8.87	CCR7 (1236)	1.73
CPVL (54504)	26.11	LRRN3 (54674)	4.8	F13A1 (2162)	8.83	ABLIM1 (3983)	1.72
S100A9 (6280)	25.51	CELA1 (1990)	4.67	NRGN (4900)	8.64	LINC00402 (100507612)	1.72
HLA-DRB5 (3127)	24.71	DPYSL4 (10570)	4.61	SDPR (8436)	8.54	KLHL3 (26249)	1.71
S100A8 (6279)	23.2	SERPINE2 (5270)	4.39	RGS18 (64407)	8.24	RP11-664D1.1 (105369609)	1.71
HLA-DRB4 (3126)	23.2	SGK223 (157285)	4.39	ITGB5 (3693)	8.19	SNORD104 (692227)	1.67
TCN1 (6947)	19.67	AC084018.1 (338799)	4.04	C2orf88 (84281)	8.19	MAN1C1 (57134)	1.62
HLA-DRB6 (3128)	18.55	BACH2 (60468)	4.02	TUBB1 (81027)	7.98	PRG4 (10216)	1.62
CEBPD (1052)	18.09	PCSK5 (5125)	3.9	TCN1 (6947)	7.94	CERS6 (253782)	1.61
TYROBP (7305)	17.65	ANKRD55 (79722)	3.85	MS4A7 (58475)	7.85	NDUFB1 (4707)	1.6
HLA-DRB1 (3123)	16.77	RNF175 (285533)	3.78	LYN (4067)	7.75	TRABD2A (129293)	1.6
PGLYRP1 (8993)	16.16	DSC1 (1823)	3.75	NFE2 (4778)	7.7	GRAP (10750)	1.6
LILRA3 (11026)	15.83	SCML1 (6322)	3.74	PTGS1 (5742)	7.64	MCF2L-AS1 (100289410)	1.59
CSF1R (1436)	15.54	SATB1 (6304)	3.68	LILRA3 (11026)	7.62	USP10 (9100)	1.59
FGR (2268)	15.23	APBA2 (321)	3.68	RAB32 (10981)	7.55	TCEA3 (6920)	1.58
SIRPA (140885)	15.03	FAM134B (54463)	3.61	SH3BGRL2 (83699)	7.41	TRIB2 (28951)	1.58
IL8 (3576)	14.96	PDE9A (5152)	3.6	MPL (4352)	7.34	LEF1 (51176)	1.58
S100A12 (6283)	14.34	EPHX2 (2053)	3.48	RAB31 (11031)	7.05	C1orf228 (339541)	1.57

Table S4: Amino acid translation of the CDR3 region of T-PLL cells.

Patient ID	Amino Acid Sequence	Length (aa)
TRAV		
TP001	AVRDFSGGYNKLI	13
TP002	ALDENTDKLI	10
TP003	ALDEGNNNDMR	11
TP010	APPAPNQAGTALI	13
TP012	AVNFFGQKLL AMRETLTGNQFY	10 12
TP025	VVIQTGANNLF	11
TP034	AVGDYGGSQGNLI	13
TP035	AVANSNSG AVANSNSGYALN	8 12
TP036	AEYSSASKII	10
TP037	AGSYNTDKLI	10
TP040	ALSDGTNAGKST	12
TP042	AASRVYKLS	9
TRBV		
TP002	ASSLEWGNYEQY	12
TP007	ASRSGRNYGYT ASSLGQGNSPLH	11 12
TP012	AIRENTEAF ASSRSIQETQY ASSLGPPVNEKLF ASSLPRGLDFSYEQY	9 11 13 15
TP025	SVEGGQFYEQY SVEQDSGANVLT ASSPGQGEGYEQY GSSLVGRTGKQETQY	11 12 13 15
TP034	ASSLSYGTGYMNTEAF	16
TP035	AVRGASYEQY	10
TP037	ASSSEGSTDTQY	12
TP038	ASSPGQGAMNTEAF	14
TP040	ASSLVMGREEKLF	13
TP042	EGAGLLQY	8
TP051	ASSLGQGNSPLH	12

Table S5: Correlation Matrix of surface marker expression.

(see Excel file)

Table S6: Cytokine release in TCR-stimulated T-PLL cells.

Analyzed cytokines are listed in 3 categories based on the level of cytokine release that was increased (red), without difference (blue) or decreased (green) in TCR-stimulated T-PLL cells as compared to TCR stimulated healthy donor pan-T-cells.

Cytokine	Healthy donor pan-T-cells (n=3)		T-PLL (n=5)		
pg/ml	Unstimulated (Mean ± SEM)	anti-CD3/CD28 (Mean ± SEM)	Unstimulated (Mean ± SEM)	anti-CD3/CD28 (Mean ± SEM)	
IL-2	0 ± 0	65 ± 41.4	0.4 ± 0.4	4789.0 ± 990.8	
IL-8	25.76 ± 13	223.8 ± 94.3	69.1 ± 33	1605.9 ± 584.4	
GM-CSF	0 ± 0	11.6 ± 11.6	0 ± 0	1205.8 ± 9939.6	
TNF-α	0 ± 0	17 ± 11.6	0.2 ± 0.2	535.6 ± 110	
MIP-1α	1.9 ± 0.2	25.6 ± 12	11.1 ± 4.2	312.2 ± 76.1	
MIP-1β	39.5 ± 5	80.7 ± 14.1	41.3 ± 16.9	404.6 ± 81.4	
TNF-β	0 ± 0	0 ± 0	0 ± 0	258.2 ± 170.8	
SDF-1α	22.8 ± 9.8	44.2 ± 10.6	104 ± 40.9	255.1 ± 92.9	
IL-10	0.36 ± 0.01	4 ± 2.2	0.4 ± 0.1	208.1 ± 119.1	
IP-10	5 ± 0.8	7.6 ± 1.7	0.8 ± 0.8	177 ± 86.3	
IFNγ	3.6 ± 2	4.2 ± 2.6	3.1 ± 3.1	141.7 ± 93.1	
LIF	0.1 ± 0.1	0.6 ± 0.6	0.2 ± 0.1	141.3 ± 62.2	
IL-22	154.75 ± 80.6	12.1 ± 12.1	126.2 ± 57.3	135.9 ± 29.3	
IL-1RA	387.98 ± 201.2	0 ± 0	214.9 ± 189.9	120.6 ± 50.8	
IL-23	30.1 ± 16.1	0 ± 0	65 ± 30.5	113.4 ± 41.6	
MCP-1	4 ± 2	2.3 ± 1.2	37.3 ± 14.6	95.8 ± 41.6	
VEGF-A	21.2 ± 9.5	16.8 ± 1.1	13 ± 9.8	89.7 ± 29.6	
IL-31	148 ± 74.6	0 ± 0	117.4 ± 95.4	56.1 ± 38.9	
IL-9	22.54 ± 15	17.2 ± 4	19.4 ± 12.5	53.6 ± 13.3	
IL-13	7.9 ± 4	0 ± 0	4.5 ± 3.4	13.4 ± 9.7	
IL-15	4.2 ± 4.2	0 ± 0	4.6 ± 4.6	7.6 ± 3.3	
GRO-α	19.4 ± 7.2	9.6 ± 1.1	16.4 ± 8.1	25.6 ± 5.8	
IL-18	35 ± 21.2	0 ± 0	20.6 ± 13	24.5 ± 7.5	
PlGF-1	5.8 ± 1	7.9 ± 0.6	4 ± 2	19.8 ± 5.4	
IL-6	18.4 ± 9.2	0 ± 0	13.3 ± 10.8	18.6 ± 9	
IL-21	10.5 ± 6	0 ± 0	6.4 ± 5.5	8.7 ± 4.8	
IL-4	12.63 ±	0 ± 0	7.0 ± 7.0	7.8 ± 3.4	
VEGF-D	0 ± 0	2 ± 1	1 ± 1	6 ± 1.1	
HGF	0 ± 0	0 ± 0	0.5 ± 0.3	5.6 ± 2.4	
FGF-2	17.8 ± 8.9	0 ± 0	11.8 ± 8.8	5.2 ± 3.6	
IL-1α	0.7 ± 0.1	1.4 ± 0.2	0.8 ± 0.5	5.0 ± 1.7	
IL-27	48.3 ± 25.3	0 ± 0	27.9 ± 26.5	3.3 ± 3.3	
Eotaxin	0.6 ± 0.2	0.91 ± 0.3	0.7 ± 0.49	3.2 ± 0.4	
IL-1β	1.9 ± 1.1	0 ± 0	1.1 ± 1.1	2.0 ± 1.4	
PDGF-BB	6.17 ± 0.8	3.4 ± 1.7	2.3 ± 1	3.6 ± 0	
IL-5	0 ± 0	0 ± 0	0.2 ± 0.2	0 ± 0	
IL-7	0.2 ± 0.1	0.1 ± 0.04	0.2 ± 0.1	0.4 ± 0.13	
IL-12	1.9 ± 0.9	0.57 ± 0.03	1.2 ± 0.7	0.6 ± 0.2	
SCF	2.7 ± 1.4	0 ± 0	2.1 ± 1.9	0.1 ± 0.1	
NGF-β	22.3 ± 17.7	0 ± 0	16.1 ± 16.1	0 ± 0	
BDNF	1.9 ± 0.1	0 ± 0	0.8 ± 0.8	0 ± 0	
IL-17A	0 ± 0	1.1 ± 1.1	0.2 ± 0.2	0 ± 0	
RANTES	15.1 ± 1.4	28.3 ± 1.9	1.9 ± 1.4	16 ± 7	
EGF	1.3 ± 0.2	2.9 ± 0.6	0 ± 0	0.1 ± 0.1	

Table S7: Expression of genes involved in the regulation of apoptotic pathways in T-PLL.

Gene Symbol	Fold Change	P-value
BAG4	1.85	1.42E-05
	1.96	1.21E-07
BCL11B	-1.7	2.77E-10
BCL2	-2.61	4.58E-15
	-2.55	2.78E-17
BCL2L11	1.98	0.00984
BCL2L13	-1.68	1.40E-12
BCL2L2	1.56	0.000201
BCL3	1.82	0.00984
BCL6	3.16	1.58E-07
BCL7A	1.64	0.0142
BCL7B	1.68	0.000909
BCLAF1	1.93	1.06E-05
BIRC7	3.01	0.00184
BNIP1	1.54	0.0222
	1.55	0.00594
CARD11	-1.54	6.90E-07
CARD8	-1.6	1.63E-07
CASP1	-1.6	1.19E-07
	-1.54	1.46E-10
CASP4	-1.51	1.45E-11
CASP6	-1.88	3.85E-09
	-1.85	0.000118
	-1.57	1.32E-08
CASP8	1.5	0.000885
	1.58	0.00528
CD79A	1.72	8.83E-05
	2.73	5.24E-05
CDKN1B	-2.61	6.74E-12
	-2.15	4.97E-18
FYN	-2.17	2.84E-12
	-1.98	2.27E-13
	-1.91	4.27E-14
GZMA	-2.85	3.68E-07
H1F0	3.96	0.000203
HIST1H1B	1.89	0.00798
HIST1H1C	2.42	0.00431
HIST1H1D	3.9	0.0027
HIST1H1E	5.8	5.96E-05
ITGB1	-2.28	6.12E-05
	-2.03	0.000104
	-1.76	2.62E-05
NLRC3	-1.63	6.60E-06
PIK3R1	-1.53	3.22E-06
PLEC	1.63	0.00143
	1.63	0.00579
PTPN13	-1.68	0.000217
TNF	9.63	6.72E-11

Table S8: Next-generation sequencing of the genomic rearranged TRB locus - PCR conditions.

1st PCR

	Final conc.
PCR Puffer II	1x
$MgCl_2$	3 mM
dNTP-Mix	0.2 mM
Primers	0,05 µM each
AmpliTaq Gold	1U / sample

Reaction volume 50 µl

1 cycle	initial denaturation	94° C	10 min
35 cycles	denaturation	94°C	1 min
	annealing	63°C	1 min
	extension	72° C	30 sec
1 cycle	final extension	72° C	30 min

2nd PCR

	Final conc.
Reaction Buffer with MgCl2	1x
dNTPs	0,2 mM each
Forward primer	0,2 µM
Reverse primer	0,2 µM
Fast Start High Fidelity (Roche)	2.5 U / sample

Reaction volume 50 µl

1 cycle	initial denaturation	95° C	2 min
20 cycles	denaturation	95°C	30 sec
	annealing	63°C	30 sec
	extension	72° C	30 sec
1 cycle	final extension	72° C	5 min

3. Supplementary Methods

3.1 Human T-PLL samples

Patients were diagnosed with T-PLL according to WHO criteria.[4,5] Differential diagnosis was based on clinical features, immunophenotyping (flow-cytometry and histochemistry; including TCL1A/MTCP1 expression), FISH/karyotypes, and molecular studies (TCR-monoclonality).[6] The cohort was selected based on uniform front-line treatment (87% of cases) with either single-agent alemtuzumab or fludarabine-mitoxantrone-cyclophosphamide (FMC) plus alemtuzumab chemo-immunotherapy as part of the *TPLL1* (NCT00278213) and *TPLL2* (NCT01186640, *unpublished*) prospective clinical trials or included in the nation-wide T-PLL registry (IRB# 12-146) of the German CLL Study Group (GCLLSG).

3.2 Mouse models

TCR tg OT-1, RAG1-deficient, and *Lck^{pr}-TCL1A* mice were obtained from the Jackson laboratory (Bar Harbor, ME, USA) and *CAR^{CEA}* mice from the Patterson Insitute, Manchester, UK.[7] *CAR^{CEA}* and *Lck^{pr}-TCL1A* mice were crossbred for ten generations to generate double tg animals (*CAR^{CEA}xLck^{pr}-TCL1A*).

Animals were bred and housed in animal facilities of the Georg-Speyer-Haus (Frankfurt, Germany) and University Hospital Cologne (Cologne, Germany) under specific pathogen-free conditions.

For the OT-1 transplantation model, $5x10^6$ retrovirally transduced OT-1 T-cells were injected intravenously into each RAG1$^{-/-}$ recipient. Recipient mice received intraperitoneal injections of 25µg OVA (257-264) in PBS mixed in a 1:1 ratio with incomplete Freund's adjuvant (IFA) every two weeks for in vivo stimulation of OT-1 T-cells. Control mice received PBS/IFA (1:1) injections.

For the CAR transplantation model, CEA-tg recipient mice (2-7 months old) were treated with cyclophosphamide (200 mg/kg intravenously) on day 1 and fludarabine (150 mg/kg intravenously) on day 4; lympho-depletion was verified by flow cytometry (FSC/SSC) on day 8. Donor splenocytes from *CAR^{CEA}*, *Lck^{pr}-TCL1A* and *CAR^{CEA}xLck^{pr}-TCL1A* mice (12-16 weeks old) were isolated by density gradient centrifugation. Splenocytes ($1x10^7$) were injected intravenously into each *CAR^{CEA}* recipient.

In both models, repopulation of transplanted cells was monitored by flow cytometric analysis of blood samples taken from lateral tail vein. Symptomatic/leukemic were examined for pathological abnormalities, including histology, morphology, white blood cell (WBC) counts, and flow cytometry. Sections of formalin-fixed, paraffin-embedded organs were stained with hematoxylin/eosin (HE) and blood smears with May-Gründwald-Giemsa.

3.3 Cell isolation and flow cytometry

Cell isolation: Healthy T-cell populations were enriched from PBMCs by negative selection using the following kits according to the manufacturer's instructions (Miltenyi Biotec): pan-T-cell isolation kit, naïve $CD4^+$ T-cell isolation kit II, memory $CD4^+$ T-cell isolation kit, and $CD4^+$ central memory T-cell isolation kit. Purity of each population (>98%) was assessed by flow cytometry.

Flow cytometry: The following antibodies from BioLegend (BL), Beckman Coulter (BC), BD Biosciences (BD), eBioscience (eB), and Miltenyi Biotec (MB) were used: TCL1A-PE/APC (eBio1-21, eB), TCL1A-A647 (1-21, BL), CD1a-AF700 (HI149; BL), CD3e-PE (145-2C11; MB), CD3-APC (SK7; BL), CD3-PB (HIT3a; BL), CD4-APC-Cy7/PE (OKT4; BL), CD4-KO (13B8.2; BC), CD5-ECD (BL1a; BC), CD5-PC7 (UCHT2; BL), CD7-FITC (CD7-6B7; BL), CD8-APC-Cy7/AF488 (HIT8a; BL), CD8-PC5.5 (RPA-T8; BL), CD19-APC (HIB19; BL), CD19-ECD (J3-119; BC), CD25-APC (BC96; BL), CD28-AF700 (CD28.2; BL), CD38-PC5.5 (HIT2; BL), CD40L-APC-eF780 (24-31; eB), CD44-VB (IM7.8.1; MB), CD44-PC7 (IM7; BL), CD45-PB (HI30; BL), CD45RA-PE (HI100; BL), CD45RO-AF700 (UCHL1; BL), CD62L-APC (MEL-14; BD), CD62L-APC-Cy7 (DREG-56; BL), CD69-APC-Cy7 (FN50; BL), CD95-PC7 (DX2; BL), CD95L-PE (NOK-1; BL), CD122-APC (TU27; BL), CD124-PE (hIL4R-M57, BD), CD127-AF488 (A019D5; BL), CCR3-AF647 (5E8; BL), CCR4-PC5.5 (TG6/CCR4; BL), CCR5-FITC (HEK/1/85a; BL), CCR7-PC5.5 (G043H7; BL), CXCR3-PB (G025H7; BL), CXCR4-PC7 (12G5; BL), CXCR5-PC5.5 (TG2/CXCR5; BL), Bcl2-AF647 (100; BL), GATA3-PC7 (L50-823; BD), Ki67-FITC (Ki-67; BL). Intracellular staining was performed using the IntraPrep kit (BeckmanCoulter) according to the manufacturer's instructions. TCR clonality was assessed by flow cytometry using the Human IOTest Beta Mark TCR V Kit (BeckmanCoulter) and the Mouse Vβ TCR Screening Panel (BD Pharmingen) according to the manufacturer's

instructions. Analyzes were performed on a Gallios cytometer (BeckmanCoulter) and MACSQuant Analyzer (Miltenyi Biotec) using Kaluza (BeckmanCoulter) and FlowJo software (FlowJo, LLC).

SPADE analysis: FlowCore (R package version 1.38.0) was used to read in/out FCS / LMD data files and compare / match their marker names for each batch. SPADE[1] was used in R for each batch separately with default parameters and excluding fwd / bwd scatter for clustering & tree construction.

3.4 RNA extraction

RNA was extracted from 1x10^7 PBMCs of T-PLL patients (>95% purity of T-cells) and PB T-cell populations (naïve, pan memory and CM) of healthy donors using the mirVana Kit (Invitrogen).

3.5 Gene expression profiling (GEP)

GEP assays were performed on Illumina HumanHT 12 v4 BeadChip arrays according to the manufacturer's instructions. GEP data have been submitted to the GEO database under accession number GSEXXX.

The Illumina proprietary software GenomeStudio v1 was used to background-correct and initially annotate the probes of the HumanHT-12 v4 Expression BeadChip. Batch-effects were corrected by batch-strata and the ComBat method.[8] The data mining tool biomaRt was used via R 3.1.0 Bioconductor 2.10 for probe annotation.[9] Q-values were calculated via the q-value library. Hierarchical (unsupervised) clustering was done with heatmap.2 from the gplots_2.15.0 library (distance function: euclidean; clustering: complete linkage). PCAs were computed with method prcomp() from the stats library.

3.6 Quantitative real-time PCR

Total RNA was reverse-transcribed using SuperScript II reverse transcriptase (Invitrogen). Real-time PCR was performed using an ABI 7500 Fast System (Applied-Biosystems) in the presence of SYBR-green (Applied-Biosystems). Levels of mRNA were quantified using the comparative C_T method and normalized to beta-actin.[10]

The following primers were used: ANKRD55 forward (Fw) 5'-GAAGGCCGAATGTGTCCAGTCACT-3', reverse (Rev) 5'-

GAGGGGGTCGAGTAGGCTCTGTTC-3'; Beta-Actin Fw 5'-TCCCTCACAG CACTAGTATTTCATG-3', Rev 5'-GAATCGGCTGTGTTCTCACAAG-3'; DPYSL4 Fw 5'-AGCGCCTGCCGTGGTCATAAG-3', Rev 5'-CGGGCCCGTCATACAGTCCAC-3'; HLA-DRB1 Fw 5'-GAGCTCCCCACTGGCTTTGTCTG-3', Rev 5'-CTCCCCCACGTC GCTGTCG-3'; LRRN3 Fw 5'-ATGCCACTCCGAATTCATGTGCT-3', REV 5'-CCAAG GCCTGATTTCACACGTACA-3'; OLFM4 Fw 5'-GAT CAAAACACCCCTGTCGTC CAC-3', Rev 5'-TCAATGGCGCCACCCAATACA-3'; TSHZ2 Fw 5'-CTCCTCGTCCG TCCCTGTGTCA-3', Rev 5'-GCCGAGGAGAAAACAGCAGGCAC-3'.

3.7 Cell lines, primary cells, cell culture and *in vitro* stimulation

All cell lines and human primary cells (T-PLL, healthy controls) were cultured in RPMI-1640 Medium (Sigma-Aldrich) supplemented with 1% L-Glutamine (200 mM; Sigma-Aldrich), 10% fetal bovine serum (FBS) (Sigma-Aldrich) and Penicillin / Streptomycin (100U / 0.1M; PAA). Cells were maintained at a density of 1-3x10^5/ml (HH and Jurkat cells) and 1x10^6 cells/ml (T-PLL cells).

For primary human T-cell stimulation, cells (4.5x10^5 cells/mL) were plated into 6- or 96-well plates, which were pre-coated with various concentrations of anti-CD3 epsilon and/or anti-CD28 antibodies at 37°C for 1h or at 4°C overnight. Anti-CD3/CD28 antibodies were either self-produced (OKT3, 15E8) or purchased from Biolegend (OKT3; 28.2). PMA (phorbolmyristylacetate) and ionomycin were used at a final concentration of 100ng/mL and 1mM, respectively.

Primary murine mononuclear cells were isolated from spleen and LNs of TCR tg OT-1 mice and cultivated in RPMI 1640 (Thermo Scientific), supplemented with 10% fetal calf serum (Merck Millipore), 2% L-glutamine (Thermo Scientific), 1% Pen/ Strep (Thermo Scientific), 1% sodium pyruvate (Thermo Scientific), 1% nonessential amino acids (Invitrogen), and 0.1% β-mercaptoethanol (Thermo Scientific) at a density of 2.5x10^6 cells per ml. For OT-1 T-cell stimulation, ovalbumin (OVA) peptide (257-264) (10ng/ml) and IL-2 (10U/ml) were added to the medium.

3.8 Next-generation sequencing of the genomic rearranged TRB locus

The amplicon next generation sequencing (NGS)-based detection of clonal TRB rearrangements was performed on Illumina MiSeq sequencer. Sequencing libraries were prepared as previously described[11], using modified biomed-2 primers for

complete TRB rearrangements[12] for the 1st PCR, and primers harboring Illumina sequencing adaptors and barcodes for the 2nd PCR. The PCR conditions for both PCRs are shown in Supplementary **Table S8**. After the 1st PCR, the PCR products were diluted 0×, 10×, or 100×, depending on the intensity of the band detected by the Agarose gel electrophoresis and 1µl of such PCR product was used for the 2nd PCR reaction. After 2nd PCR, the concentration of the resulting PCR products was measured using the Quant-iT™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific) and the PCR products were pooled into 3 subpools in equimolar ratios. Each subpool was purified via the extraction from the 2% agarose gel, using the MinElute Gel extraction kit (Qiagen). The concentration of each subpool after gel extraction was measured using the Quant-iT™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific) and the final pool with the concentration of 7pM was used for sequencing. Sequencing was performed on the MiSeq sequencer, using the 2x250 bp v2 chemistry, according to the manufacturer's instructions.

The raw sequencing data were demultiplexed using bcl2fastq conversion software (Illumina) with 0 mismatches in barcode sequences allowed. Resulting fastq files were analysed using the bioinformatics tool Vidjil.[13] Only clones with frequency > 15% were reported.

3.9 Reconstruction of TCR chains with RNA-Seq

Whole transcriptome sequencing (RNAseq) analyses were conducted using the Illumina HiSeq2000 platform as previously described.[14] Similarly to the protocol on reconstruction of immunoglobulin chains with RNA-Seq by Bachy and colleagues[15], we de novo assembled T-cell receptor (alpha, beta, gamma and delta) V-D-J transcripts in T-PLL (n=15) and normal CD3+ pan-T-cells (n=4). Reads were aligned with STAR_2.5.2a[16] in 2-pass mode to the GRCh37/hg19 reference genome. Those reads mapping to the TCR alpha, beta, gamma and delta loci were extracted with bedtools.[17] TCR genes and pseudogenes were identified from the Gencode project annotation version 24[18] lifted to GRCh37/hg19. The number of reads mapping to each gene or pseudogene was counted with HTSeq-0.6.1 (www-huber.embl.de/users/anders/HTSeq/doc/overview.html) in default exon-union mode.

Due to TCR segment rearrangements (including fragmentation and fusion of multiple segments) default gapped read aligner fail to align all TCR gene or pseudogenes.

We therefore extracted unmapped reads from the STAR aligment with Picard tools (https://broadinstitute.github.io/picard/) and function SamToFastq (VALIDATION_STRINGENCY=SILENT) in order to reconstruct TCR chains for each sample with the de novo transcription assembler Trinity 2.1.1[19]; ran in non-genome-guided mode, minimum contig length 200, without Jaccard clipping, without digital normalization).

From the individual reconstructed transcriptome, transcripts bearing homology to human TCR genes and pseudogenes were identified with NCBI BLAST[20] using a database downloaded from IMGT.[21] From these homology-bearing Trinity transcripts, we constructed a new reference transcriptome and remapped all reads to it with the bowtie2 short read aligner.[22] Read counts, FPKM (Fragments Per Kilobase Million) and TPM (Transcripts Per Kilobase Million) for each transcript were calculated with eXpress version 1.5.1.[23]

Reconstructed transcripts were further annotated using IgBLAST[24] wrapped in the MIGMAP package (https://github.com/mikessh/migmap: HTS-compatible wrapper for IgBlast V-(D)-J mapping tool). We only accepted following criteria (in decreasing priority): in-frame chains, or chains containing two out of three homologues segments, or containing at least one homologues segment and one ambiguous. The IMGT alignments found in the Supplements can also be used to investigate (the degree of) somatic hypermutations.

Decrease in TCR repertoire was measured by number of unique reference segments with non-zero read count and reconstructed transcripts with non-zero TPM. Both were compared with Wilcoxon rank sum test in T-PLL *vs* CD3+ pan-T-cells and visualized in 3D bar plot (R-3.2.2 library latticeExtra) for segment and chain co-occurences/exclusivity.

3.10 Transfection and transduction

The human cell line HH (TCL1A-negative) was transfected with a doxycycline-inducible TRMPVIR vector containing TCL1A.[25] For induction of TCL1A expression, transfected HH cells (iHH-TCL1A) were treated with 1µg/ml (high expression) or 0.125µg/ml (low expression) doxycycline for 24h. Jurkat-TCL1A and Jurkat-GFP cells were established as previously described.[26] OVA-stimulated OT-1 T-cells were transduced *in vitro* with a retroviral vector co-expressing human TCL1A and a GFP or

luciferase reporter on 2 consecutive days. The retroviral TCL1A plasmid was generated by cloning cDNA of human TCL1A (Michael Teitell, UCLA, USA) into the previously described gamma retroviral vector MP91-GFP.[27] MP91-GFP (GFP only) was used as a control vector. For *in vivo* imaging experiments, GFP was replaced with a firefly luciferase reporter and either T-Sapphire or human TCL1A were cloned in front of the IRES. Retroviral vectors were produced by calcium-phosphate mediated transient transfection of 293T human embryonic kidney cells as previously described [27].

3.11 Cell cycle analysis

For cell cycle analysis the DNA intercalating dye propidium iodide (PI) was used. Analysis was carried out by flow cytometry. Per sample $1*10^6$ cells were washed once with cold PBS and thoroughly resuspended in 500 µL PBS. While vortexing the sample for 30 s 4 mL of ice cold 70 % ethanol were added drop wise. The samples were fixed for at least 2 h or over night at -20°C. The fixed cells were pelleted at 350 x g for 5 min and the ethanol thoroughly decanted. Cells were resuspended in 4 mL PBS and incubated for 1 min at RT before washing. The cells were then resuspended in 500 µL PI staining solution (0.1 % (v/v) Triton X-100, 0.2 mg/mL RNAse A, 0.02 mg/mL propidium iodide in PBS), incubated for 30 min at RT and analyzed immediately.

3.12 Viability assay

Cell viability was assessed using the CellTiter-Glo® (Promega) luminescent cell viability assay according to the manufacturers instructions. The assay was performed in black 96-well plates (BD Biosciences) to reduce scattered light.

3.13 Apoptosis assays

Cell viability was determined by Annexin V and 7AAD (BD Biosciences) staining according to the manufacturer's instructions. Specific cell death was calculated using the formula $((\text{viability}_{\text{baseline}} - \text{viability}_{\text{treated}}) / \text{viability}_{\text{baseline}}*100)$, wherein Annexin V / 7AAD double negative cells are considered as live cells.

Apoptosis was induced using a LEAF (Low Endotoxin, Azide-Free) agonistic CD95 antibody (EOS9.1, BioLegend). 1 x 10^6 T-PLL cells were incubated for 6 h at 37°C

with 1 µg/mL agonistic CD95 antibody in RPMI 1640 medium containing 10 % FCS. H_2O_2 (6%) treated cells were used as a positive control. Apoptosis was assessed by Annexin V and 7AAD staining.

3.14 Assessment of metabolic activity

Bioenergetics of T-PLL samples and MACS enriched pan-T-cells of healthy donorswere determined using the XF96e Extracellular Flux Analyzer (Seahorse Bioscience, North Billerica, MA, USA). Cells were seeded in specialized tissue culture plates (240.000 cells/well) and subsequently immobilized using CELL-TAK (BD Biosciences). One hour prior measurement cells were incubated at 37 °C in a CO2-free atmosphere. First, basal oxygen consumption rate (OCR) (an indicator for mitochondrial respiration) and extracellular acidification rate (ECAR) (an indicator for lactic acid production or glycolysis) were detected. Next, OCR and ECAR responses towards the application of glucose (10 mM), oligomycin (1 µM), and 2-DG (100 mM) were evaluated. Cells were stimulated using activating anti-CD3/CD28 antibodies (T-cell activation/expansion Kit, Miltenyi Biotec) in a cell to bead ratio of 1:2.

3.15 ELISA (enzyme-linked immuno sorbend assay) and cytokine array

Quantification of IL-2 was carried out using the human IL-2 ELISA MAX Deluxe Set (BioLegend) according to the manufacturers instructions. Secreted proteins were detected using the ProcartaPlex Human Cytokine/Chemokine/Growth Factor Panel 1 (45 plex) according to manufacturers instructions.

3.16 Immunoblots

Western Blotting was performed on whole-cell lysates as previously described.[28] The following primary antibodies were used at 1:1,000 dilutions: anti-TCL1A (clone 1-21)[29]; phospho-AKT^{Ser473} (D9E), pan-AKT (40D4), phospho-ERK1/2$^{Thr202/Tyr204}$ (n/a), and ERK1/2 (3A7) from Cell Signaling Technologies; beta-Actin (C-11) and β-Tubulin (H-235) from Santa Cruz Biotechnology. HRP-conjugated species-specific secondary antibodies were purchased from Santa Cruz Biotechnology. Protein bands were visualized by Western Bright ECL (Advansta) and detected using autoradiography films (Blue, 8x10; Santa Cruz Biotechnology) and the X-ray film processor

CAWOMAT 2000 IR. Signal intensities were quantified using ImageJ densitometry software (http://rsb.info.nih.gov/ij/).

3.17 Bioluminescence imaging

In vivo imaging was performed for recipient mice of OT-1 T-cells transduced with luciferase vectors four weeks after transplantation and repeated every four weeks. Biolumuniescence was detected with the IVIS Imaging System Lumina II (PerkinElmer, Waltham, Massachusetts, USA). Anesthetized mice were shaved and injected intraperitoneally with 150µl D-Luciferin (15mg/ml) 10min before imaging. Images were taken in ventro-dorsal and latero-lateral position and acquired after an exposure time of 2 and 5 minutes using binning 4. Signal intensity was quantified as average radiance of photons emitted per second and area (p/s/cm²/sr) within a region of interest (ROI) using the Living Image Software 4.0 (PerkinElmer, Waltham, Massachusetts, USA).

3.18 Statistics

Results are presented as mean ± standard error of the mean (SEM). The student *t* test (GraphPad Prism, version 5.0a) was used to determine statistical significance. P values < .05 were considered significant.

References

1. Qiu P, Simonds EF, Bendall SC, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*. 2011;29(10):886–91.
2. Haining WN, Ebert BL, Subrmanian A, et al. Identification of an evolutionarily conserved transcriptional signature of CD8 memory differentiation that is shared by T and B cells. *J. Immunol.* 2008;181(3):1859–68.
3. Mahnke YD, Brodie TM, Sallusto F, Roederer M, Lugli E. The who's who of T-cell differentiation: human memory T-cell subsets. *Eur J Immunol*. 2013;43(11):2797–809.
4. Swerdlow S, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;127(20):2375–90.
5. Herling M, Khoury JD, Washington LT, et al. A systematic approach to diagnosis of mature T-cell leukemias reveals heterogeneity among WHO categories. *Blood*. 2004;104(2):328–335.
6. Ravandi F, O'Brien S, Jones D, et al. T-cell prolymphocytic leukemia: a single-institution experience. *Clin Lymphoma Myeloma*. 2005;6(3):234–239.
7. Eades-Perner AM, van der Putten H, Hirth A, et al. Mice transgenic for the human carcinoembryonic antigen gene maintain its spatiotemporal expression pattern. *Cancer Res*. 1994;54(15):4169–76.
8. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
9. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.
10. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc*. 2008;3(6):1101–8.
11. Kotrova M, Muzikova K, Mejstrikova E, et al. The predictive strength of next-generation sequencing MRD detection for relapse compared with current methods in childhood ALL. *Blood*. 2015;126(8):1045–7.
12. van Dongen JJM, Langerak AW, Brüggemann M, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*. 2003;17(12):2257–317.
13. Giraud M, Salson M, Duez M, et al. Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics*. 2014;15:409.
14. Fernandez-Cuesta L, Sun R, Menon R, et al. Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol.* 2015;16:7.

15. Blachly JS, Ruppert AS, Zhao W, et al. Immunoglobulin transcript sequence and somatic hypermutation computation from unselected RNA-seq reads in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 2015;112(14):4322–7.

16. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.

17. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.

18. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760–74.

19. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8(8):1494–512.

20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.

21. Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res*. 2005;33(Database issue):D256-61.

22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9.

23. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*. 2013;10(1):71–3.

24. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*. 2013;41(Web Server issue):W34-40.

25. Zuber J, McJunkin K, Fellmann C, et al. Toolkit for evaluating genes required for proliferation and survival using tetracycline-regulated RNAi. *Nat Biotechnol*. 2011;29(1):79–83.

26. Herling M, Patel KA, Teitell MA, et al. High TCL1 expression and intact T-cell receptor signaling define a hyperproliferative subset of T-cell prolymphocytic leukemia. *Blood*. 2008;111(1):328–337.

27. Newrzela S, Cornils K, Li Z, et al. Resistance of mature T cells to oncogene transformation. *Blood*. 2008;112(6):2278–2286.

28. Schrader A, Meyer K, von Bonin F, et al. Global gene expression changes of in vitro stimulated human transformed germinal centre B cells as surrogate for oncogenic pathway activation in individual aggressive B cell lymphomas. *Cell Commun Signal*. 2012;10(1):43.

29. Herling M, Patel KA, Weit N, et al. High TCL1 levels are a marker of B-cell receptor pathway responsiveness and adverse outcome in chronic lymphocytic leukemia. *Blood*. 2009;114(21):4675–4686.

AIMS Medical Science, 3(3): 248–271.
DOI: 10.3934/medsci.2016.3.248
Received 30 July 2016,
Accepted 5 September 2016,
Published 10 September 2016

http://www.aimspress.com/journal/medicalScience

Review article

A Critical Evaluation of Analytic Aspects of Gene Expression Profiling in Lymphoid Leukemias with Broad Applications to Cancer Genomics

Giuliano Crispatzu [1, 2], Alexandra Schrader[1, 2], Michael Nothnagel [3], Marco Herling [1, 2, *], and Carmen Diana Herling [1, *]

[1] Department of Internal Medicine I, Center for Integrated Oncology (CIO) Köln-Bonn, University of Cologne (UoC), Germany;
[2] Excellence Cluster for Cellular Stress Response and Aging-Associated Diseases (CECAD), UoC, Germany;
[3] Cologne Center for Genomics (CCG), Department of Statistical Genetics and Bioinformatics, UoC, Germany

* **Correspondence:** Email: carmen.herling@uk-koeln.de; marco.herling@uk-koeln.de;
Tel: +49-221-478-5969; Fax: +49-221-478-6383

Abstract: In cancer research, transcriptional aberrations are often deduced from mRNA-based gene expression profiling (GEP). Although transcriptome sequencing (RNA-seq) has gained ground in the recent past, mRNA-based microarrays remain a useful asset for high-throughput experiments in many laboratories. Possible reasons are the lower per-sample costs and the opportunity to analyze obtained GEP data in association with published data sets. There are established and widely used methods for the analysis of microarray data, which increase the comparability of different GEP data sets and facilitate data-mining approaches. However, analytic pitfalls, such as batch effects and issues of sample purity, e.g. by complex tissue composition, are often not properly addressed by these standard approaches. Moreover, most of these tools do not capitalize on the full range of public data sources or do not take advantage of the analytic possibilities for functional interpretation or of comprehensive meta-analyses. We present an overview of the most critical steps in the analysis of microarray-based GEP data. We discuss software and database query solutions that may be useful for

each step and for generally overcoming analytic challenges. Aside from machine-learning applications to classify and cluster samples, we describe clinical applications of GEP, including a novel exploratory algorithm to identify potential biomarkers of prognosis in small sample cohorts as demonstrated by exemplary data from lymphatic leukemias. Overall, this review and the attached source code provide guidance to both molecular biologists and bioinformaticians / biostatisticians to properly conduct GEP analyses as well as to evaluate the clinical / biological relevance of obtained results.

Keywords: Cancer genomics; gene expression profiling; microarray; RNA-Seq; survival analysis; CLL; T-PLL; leukemia; lymphoma; TCL1; contamination; SVM; random forest

1. Introduction

Traditionally, gene expression analysis includes reverse transcription of mRNA into cDNA and probing of gene transcripts of interest by specific primers designed for target PCR amplification (gold standard), followed by quantitative, semi-quantitative (e.g. qRT-PCR), or electrophoresis (e.g. Southern blotting) detection methods. Based on efforts provided by the Human Genome Project [1,2] and studies on expressed sequence tags (ESTs) in mammalian genomes, cDNA hybridization array chips have originally been designed to investigate deregulated mRNA expression of distinct and well-characterized gene transcripts in various diseases. Modern mRNA-microarray platforms apply one or two-color fluorescence labeling (i.e. Cyanine3 / Cy3 for green and Cy5 for red dye fluorescence) for one or two samples to be loaded on the chip, respectively, and allow the detection of more than 47 000 transcripts. In contrast to two-color arrays (e.g. HuA1 by Agilent Technologies, Santa Clara, CA, USA), one-color arrays, are most commonly used today (e.g. HG-U133 Plus 2.0 by Affymetrix, Inc., Santa Clara, CA, USA, or BeadArray HT-12v4, Illumina, Inc., San Diego, CA, USA) and represent the focus of this review.

The past few years have seen the advent of transcriptome sequencing (RNA-seq) based on the next-generation sequencing (NGS) technology using high-throughput platforms, such as the GA IIx or HiSeq2000 sequencer from Illumina. RNA-seq does not require the prior design of specific probes, rendering it a highly versatile approach for gene expression profiling (GEP). Accordingly, a number of publications on the genomic landscape of various neoplasms have applied RNA-seq to investigate gene-specific aspects such as differential splicing and exon usage [3], hidden viral transcripts [4], and cancer-specific fusion transcripts [5]. However, published reports using RNA-seq in cancer often lack statistical power for comprehensive gene expression analyses due to a limited sample size. In contrast, mRNA-based microarrays have remained the initial method of choice for high-throughput analyses of gene expression in many laboratories. Reasons for this include the associated lower per-sample costs as well as the availability of already published microarray-derived GEP data in

public databases. Many of these data sets were processed by established and widely used methods, thereby improving their comparability and the suitability for data-mining approaches.

Within this review, we present an overview of critical steps in the analysis of microarray-based GEP data (see overview in Figure 1) and the corresponding library and code information (summarized in Table 1 and 2). We will discuss step-by-step software and database query solutions that may be useful for data analysis, to avoid analytic pitfalls, and to provide an increased capability for clinical and biological interpretation of data. To illustrate the proposed analytic steps, we present analyses on exemplary data of previously published and own GEP data, all obtained in patients with B- and T-cell leukemias or lymphomas.

2. Quality Control can Greatly Differ by Platform

There are various possibilities to apply basic steps of quality control (QC) prior to or during preprocessing of GEP raw data. In order to avoid false estimates of background intensities and false inputs for normalization, removal of potential problematic samples and probes *before* data preprocessing is essential towards a correct interpretation of data. Problematic samples often present as outliers in density distributions or in an unsupervised cluster analysis on global gene expression values (*after* data preprocessing). The latter, e.g. in form of dendrograms (Code 1) or principal component analyses (PCA; Code 2), is created by using the R [7] library *arrayQualityMetrics* from Bioconductor [8] with its informative HTML report per array.

Numerous methods and libraries for R are available for more specific quality assessments for each of the three major microarray platforms. Affymetrix arrays can be analyzed using the *affyQCReport* and *simpleaffy* libraries (see Table 1 for all library references), which normalize expression values using housekeeping genes (e.g. calculating the actin3/actin5 ratio), while the *affyPLM* library allows calculation of important quality measures such as the normalized unscaled standard error (NUSE) and relative log expression (RLE) as well as their plotting across samples (Code 3). The quality of data obtained with Illumina chips can be assessed by statistical standard measurements (mean and standard deviation) or outlier detection using the *lumiQ* function within the *lumi* library (Code 4). Possible slide inhomogeneities (i.e. scratches) or contamination on two-color arrays may be detected with the *imageplot* function of the *limma* library. This package also allows the calculation of the RNA Integrity Number (RIN) as a measure of mRNA degradation with a subsequent option to remove samples below a given threshold.

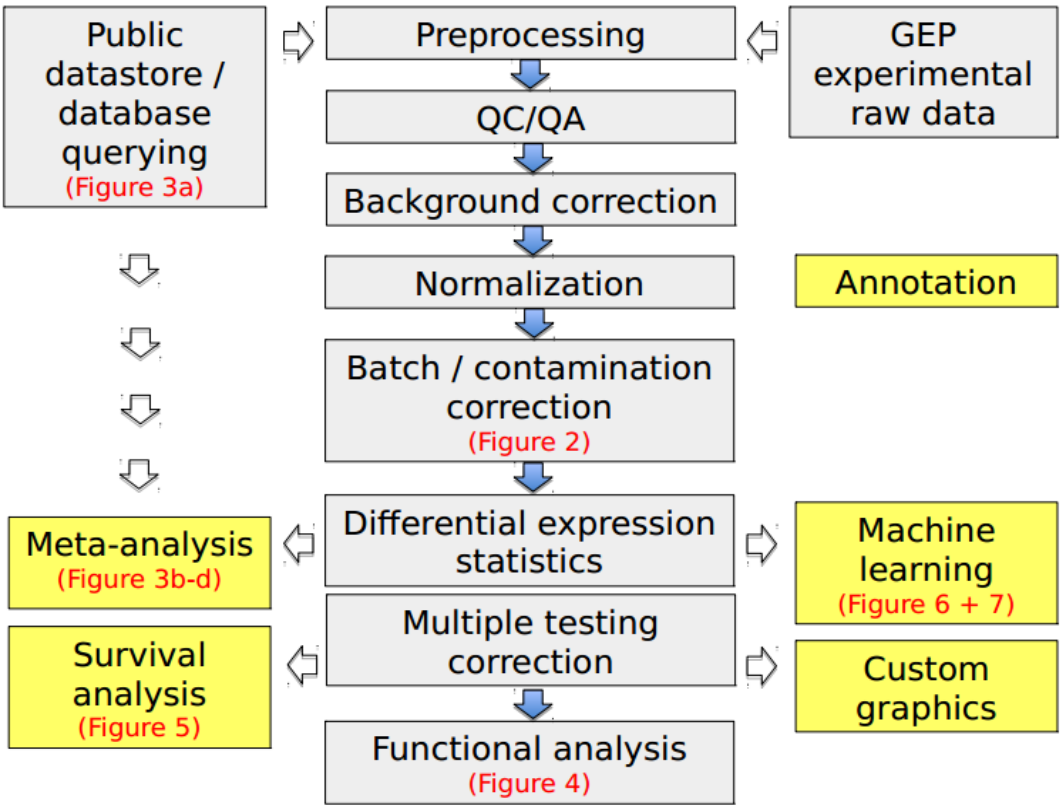


Figure 1. Flow chart describing a suggested GEP protocol. Steps in yellow boxes are modular and may function somewhere independently downstream of the steps in grey boxes. The red text refers to those figures of this review that illustrate the respective step.

3. Proper Preprocessing of Raw Data

A first step in the standard analysis protocol of cDNA microarrays usually is the conversion of hybridization image spots obtained by array scanners into raw gene expression values. For Affymetrix chips this is normally done either by using the freeware *Affymetrix Power Tools* or the R library *affy*. For Illumina's BeadChips the proprietary *GenomeStudio* software or manual decryption via the R library *beadArray* may be used. For two-color arrays, scanner output files, e.g. in TIF format, can easily be read with the *read.maimages* function from the *limma* R library.

In a second step, background correction is conducted by subtracting technical noise from biological variation. This is accomplished by using e.g. *RMA* [9] for Affymetrix arrays or the *bgAdjust* function from the *lumi* R library for Illumina arrays, which employs a similar algorithm as *GenomeStudio* (Code 5). In order to account for outliers and to remove systematic variation, normalization of expression values is required. The most common procedures include quantile-normalization, which preserves the rank, but may eliminate small differences in expression values, and LOESS (locally weighted scatterplot smoothing)-normalization, which does the opposite. Robust splice normalization (RSN) aims to combine the advantages of both methods through a

monotonic splice fit to one reference sample, while simple scaling normalization (SSN) forces samples to have the same scale and background. Both approaches are included in the *lumi* R library for Illumina arrays. For two-color arrays it may be essential to further account for dye biases in the normalization [10] and to normalize within the array itself (between both color-labeled samples) and between all two-color arrays of the cohort, e.g. by use of the *limma* R library. Variance-stabilizing normalization (VSN) constitutes another method for combining background correction and normalization [11], while preserving biological variation. It is implemented in the *vsn* (Code 6) library, applicable to arrays of all major platforms. Within the normalization process raw intensities are usually transformed, either into a log2 scale or glog in case of VSN, in order to smoothen extreme values.

4. Probe Annotation and Deconvolution

Frequent impediments for GEP data analysis are missing array annotations or outdated annotation files provided by the manufacturers (e.g. frequently old GenBank predictions are included). Data-mining tools such as *biomaRt* [12] can be used to acquire up-to-date probe information (Code 7). They may also be helpful in assigning probes to transcripts, thereby enabling filtering for redundancies of probes, which map primarily to transcripts that are prone to nonsense-mediated mRNA decay (NMD) or to unprocessed pseudogenes. Deconvolution of genes with known transcript variants of differential function into probed isoforms may also be important for extrapolations on biological relevance. An example is the apoptosis regulator *myeloid cell leukemia sequence 1 (MCL1),* of which the longer isoform (MCL1-001) has been reported to enhance survival by inhibiting apoptosis, while its shorter isoform (MCL1-002) acts as a pro-apoptotic molecule [13].

5. Exploring Differentially Expressed Genes Considering the "Multiple Comparisons Problem"

Raw data preprocessing and QC is followed by the actual statistical analysis, usually in the form of probe-by-probe hypothesis tests for differential expression including: (1) two-group mean comparisons using a Student's t-test (parametric, i.e. presuming a known statistical distribution), (2) empirical Bayes / moderated t-tests (for low sample size; e.g. $n < 10$; parametric), (3) Mann-Whitney-U tests (for samples with low variability; non-parametric) (Code 8), (4) multiple-group tests by means of an analysis of variance (ANOVA; parametric) (Code 9), or (5), a Jonckheere test (trend test; non-parametric). However, statistical testing of all genes / transcripts detected by an array requires correction for multiple testing, in order to avoid a substantial number of false-positive findings [14,15]. For example, using a significance level of 0.05 for each of 10,000 tests would result in approximately 0.05 * 10,000 = 500 significant rejections by chance, even if all null hypotheses of no differential expression were true. To this end, we can either control the family-wise error rate (FWER) to curtail

the number of statistically significant results, e.g. by use of the (conservative) Bonferroni correction, in which the significance level for each probe-specific test equals the FWER (e.g. 0.05) divided by the total number of tested probes, or by some permutation / resampling approach. Furthermore, we can aim for controlling the false-discovery rate (FDR), i.e. the proportion of falsely rejected null hypotheses, e.g. using the Benjamini-Hochberg's procedure, q-values, or other approaches. It should be noted, however, that control of the FDR, while very helpful in limiting the number of erroneously followed-up probes, does not imply a notion of statistical significance. The procedures by Bonferroni and by Benjamini-Hochberg are implemented in the *multtest* library [16], while the *qvalue* library provides an implementation for the rank-preserving q-value calculation (Code 10).

Nominally differentially expressed probes (e.g. with a single-test level of $p < 0.05$) can also be filtered by multiple-testing correction, for example by applying a q-value / FDR cutoff (common cut-off, e.g. 0.1) to ensure a low proportion of false-positives in the set of probes to be subsequently followed up. To reduce time in the analysis, it may also be useful to exclude genes / probes that are not expected to be differentially expressed either due to biologically low variability in the investigated samples, or due to technically low detectability on the array. This can be achieved either by non-specific filtering of expression values restricted to a given range (e.g. the shortest interval containing half of the data by standard deviation (sd) or interquartile range) or by setting an empirical cut-off to the coefficient of variation (sd/mean), e.g. the top 10 percent or a fixed value of 0.6. Note, however, that this may increase the rate of false-negative findings (Code 11).

6. Pitfalls: Batch-correction and Contamination Estimation

When comparing GEP data obtained in the same laboratory, but with two or more different batches of arrays, the results will deviate from one another beyond the expected biological and array-specific technical variation. Batch correction addresses this issue. Two approaches commonly considered to be performing best [17] are mean-centering and a Bayesian framework named ComBat [18] (Figure 2a–c; Code 12).

A particular problem for cancer transcriptomics / genomics is the contamination of cancer tissues by normal cells (irrespective whether to consider them as actual milieu components) and vice versa. Even in lymphomas and lymphoid leukemias, such problems are encountered in lymph-node samples or in the seemingly 'pure' blood samples, as these are also of mostly multicellular composition. Tools like *ESTIMATE* [19] can weigh specific markers (e.g. indicating an immune or stromal cell origin) within gene expression profiles in the form of gene set enrichment analyses and thus evaluate the degree of purity. Unfortunately, due to intrinsic aberrations of 'immune cell' genes within tumor cells of leukemias / lymphomas, the immune gene set used within *ESTIMATE* is not reliable for the enrichment analysis within these malignancies (Figure 2d; Code 13). An alternative approach especially for leukemias / lymphomas might be *CellMix* [20] which uses gene sets from

specific immune cell subsets, e.g. CD4+ and CD8+ T-lymphocytes, CD14+ monocytes, CD19+ B-lymphocytes, CD56+ natural killer cells, and CD66b+ granulocytes.

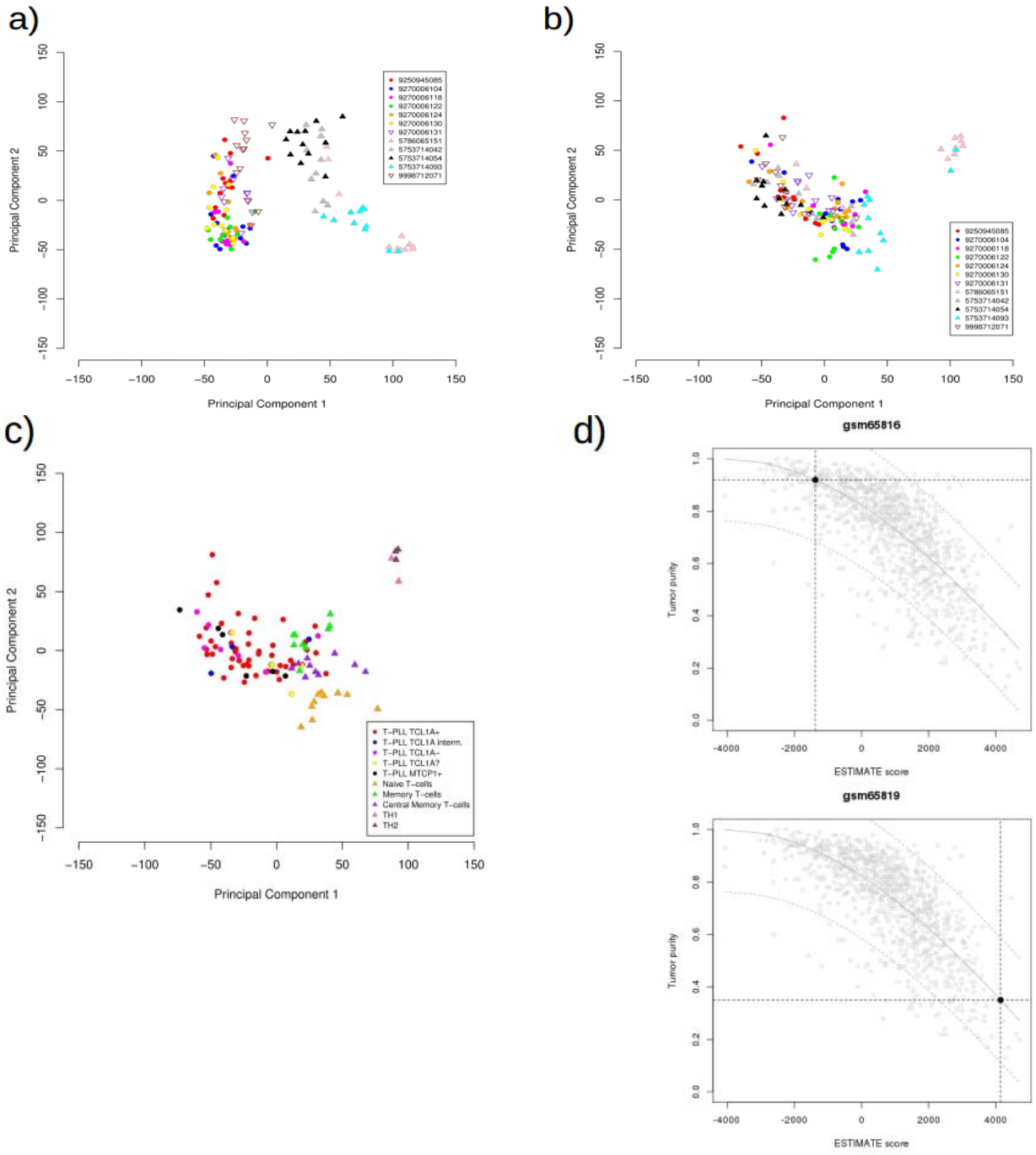


Figure 2. a) PCA (principal component analysis) of the 1000 most variable genes (by variation coefficient) within 12 distinct batches of our T-PLL (T-cell prolymphocytic leukemia) data set reveals batch-specific clustering. **b)** After batch correction samples do not cluster anymore due to technical bias, but rather due to biological information when annotated as in *c)*. **c)** Entity information can be included in ComBat (besides batch information) to fit batches. T-PLL samples (further divided by different oncogene protein status) and normal T-cells form a cloud,

while stimulated T-helper cells (TH1 and TH2) form another cloud. **d)** *ESTIMATE* plots of fitted purities from two samples within the publicly available breast cancer data set GSE2990 [48] (*n* = 189 invasive breast carcinomas; including 64 estrogen receptor (ER)-positive tumors, histologic grade 1 and 3 tumors; Affymetrix HG-U133A). **Upper panel:** When comparing the black dot to gray dots (all other samples), one can observe that the sample is among those with highest purity. **Lower panel:** sample among those with lowest purity.

7. Making Use of Public Databases

Two public databases are commonly used for the comparison of own microarray data with independent data sets, for example in a meta-analysis, namely the GEO (gene expression omnibus) database [21] (http://www.ncbi.nlm.nih.gov/geo) and the ArrayExpress database [22] (https://www.ebi.ac.uk/arrayexpress), with GEO featuring a larger number of integrated samples. Both platforms use distinct annotation / meta-data file systems. In GEO, samples are either described in MIAME Notation in Markup Language (MINiML; pronounced 'minimal') or SOFT formatted family files. In ArrayExpress, sample and data relationships (SDR) are described in the SDRF format, while protocol information is stored in the Investigation Description Format (IDF). Both databases offer processed numerical gene expression values (in the form of matrices) stored in regular text format (txt), or raw data in CEL or idat (for Affymetrix or Illumina chips) files. GEO and ArrayExpress also provide respective R libraries to automate queries and processing of differential expression analyses, namely *GEOquery* and *ArrayExpress*.

Analysis results for data sets within ArrayExpress are further integrated in the 'Gene expression atlas' of the EMBL / EBI (http://www.ebi.ac.uk/gxa). The latter provides information about gene and protein expression in animal and plant samples for different cell types, tissues, developmental stages, diseases, and other conditions from 1572 studies as of August 2015 [23]. The human data sets are currently exported into an RDF version accessible via a SPARQL Endpoint (http://www.ebi.ac.uk/rdf/services/atlas/sparql; accessed 02/21/2016).

Implemented queries include:

- “Query 1: Get experiments where the sample description contains diabetes”
- “Query 2: Get differentially expressed genes where factor is asthma”
- “Query 3: Show expression for ENSG00000129991 (TNNI3)”
- “Query 4: Show expression for ENSG00000129991 (TNNI3) with its GO annotations from Uniprot (Federated query to http://sparql.uniprot.org/sparql)”
- “Query 5: For the genes differentially expressed in asthma, get the gene products associated to a Reactome pathway”
- “Query 6: Get all mappings for a given probe e.g. A-AFFY-1/661_at”

Query 2 and 5 can be further modified in order to compare gene dysregulation in other types of diseases, e.g. in lymphoid leukemias, such as chronic lymphocytic leukemia (CLL; Table 3). User's familiarity with the underlying ontologies (controlled vocabulary; [24]) is, however, necessary to construct queries.

8. Meta Analyses: Exploring Possible Phenotypic Markers across Different Conditions

For conceptualizing a pharmacologic compound (e.g. inhibitor) acting against a specific gene product or for designing specific gene-knockouts within a model organism, it may be particularly important to know in what conditions and disease subtypes expression of a distinct gene is up- or down-regulated and to which degree (basal or extreme). Integrative analyses of expression changes within a multitude of samples of the same entity, or model organism, or any other comparable biological system as well as across initially separately analyzed (and published) series (cohorts) are often called gene expression meta-analyses. In the following we describe multiple ways to conduct a meta-analysis of GEP data with their limitations and advantages.

The first approach includes construction and sending of specific queries to the EMBL / EBI RDF platform. Querying can further be semi-automated using the *SPARQL* R library, which allows the investigation of different data sets in a specific condition, e.g. comparisons of CLL vs. normal B-cells, or between distinct groups of tumor samples stratified by a characteristic of interest, e.g. immunoglobulin heavy chain (*IGHV*) gene mutated vs. unmutated CLL. Results are usually tabularized and fold-changes visualized within a heatmap (Figure 3a; Suppl. Table 1; Code 14).

Since not all 'ArrayExpress' data sets are yet integrated into the EMBL / EBI RDF platform and the GEO database contains additional data sets, the manual download, processing, and integration of such additional data is often necessary.

Therefore, a second, more hands-on approach to meta-analyses is a search by keyword, e.g. 'chronic lymphoid', within GEO and / or ArrayExpress (or any other public database). Once the data set has been picked, it is background-corrected and the annotated replicates can be combined with their original samples by calculating their mean. Afterwards all samples within the data set are normalized (e.g. quantile-normalized).

Probe sets of a gene which map to retained / dysfunctional transcripts (or which map to more retained / dysfunctional transcripts than other probe sets of the same gene) should be removed to obtain meaningful expression values (Suppl. Table 2). For example, *BCL2L11* on Affymetrix HG-U133 Plus 2.0 chips has two probes, one hybridizes two protein-coding and six NMD (nonsense-mediated decay) transcripts, the other one hybridizes two protein-coding and eight NMD transcripts. Thus, ambiguous expression values of this gene have to be evaluated with caution. The residual unambiguous probe sets assigned to a gene are then further summarized by calculation of average expression values per gene.

For further evaluation of the GEP meta-analysis, three different techniques for integration can be used to observe gene expression patterns and entity clustering:

1) The first method quantile-normalizes a matrix of average gene expressions across entities from different experiments and finally gives a visual approximation. If there is also a tumor suppressor gene (very low expression) and an oncogene (very high expression) in the gene set to be evaluated, one can expect an expression range similar to the whole transcriptome. It should be noted that in previous Affymetrix sets, such as HG-U133A, some genes (e.g. *BMF* and *BOK*) are not covered by specific probes on the array and, therefore, need to be imputed by the median of the respective data set. This guarantees that in the heatmap (or PCA) these genes are not visualized as up- or down-regulated; they in fact can be manually labeled (blackened). Expression values from all data sets are merged into one matrix and again quantile-normalized to account for variability in platform specifications and noise. A more suitable approach than normalizing on each gene set separately might be to normalize on the whole combined transcriptome (intersection of all probed genes). However, this would disregard genes not covered by all platforms used. The resulting heatmap (generated by function *heatmap.2*, library *gplots*; Figure 3b) shows the expression of selected genes and transcripts in their respective data set and can be additionally subdivided by the different entities (median across samples of an entity).

2) Batch effects cannot be entirely excluded by using method *1)* as may be observed by a bias in clustering of samples from the same experiment. Therefore, we recommend a novel method called *inSilicoMerge* [25], which combines data sets and removes their batch effect with a choice of various methods, such as the empirical Bayes method ComBat (Figure 3c).

Unfortunately, data sets from different platforms can only be combined gene-wise, meaning that e.g. *MCL1* would not be deconvolutable into its isoforms MCL-001 / MCL1-long and MCL-002 / MCL1-short.

3) For an advanced evaluation, one can further perform differential expression analysis for data sets with different control samples (of varying quality, number, and specificity) available for comparison, such as ’normal‘ non-malignant cells or bulk tissue specimens. Fold-changes with a p-value < 0.1 (trend) or < 0.05 (significant) are extracted to compare normal-matched gene expression between different experiments and probe targets representing different gene transcripts or protein isoforms. The results are again visualized by a heatmap, either in the order obtained by hierarchical clustering (using Euclidean distance) or in order of rows sorted by gene name.

As exemplified by illustration of expression levels of *Death-Associated Protein Kinase* (*DAPK*) gene family members in subsets of CLL and normal B-cells (Figure 3d), this method allows different disease vs. ’normal‘ comparisons and facilitates the evaluation of which genes are exclusively down- or up-regulated and which show no clear pattern or which are specific to small subgroups. In the meta-analysis itself every differential expression analysis is further evaluated by statistical testing. Default setting is the Student’s t-test, except for low variation or non-normal distributions, for which the non-parametric Wilcoxon rank sum test is recommended.

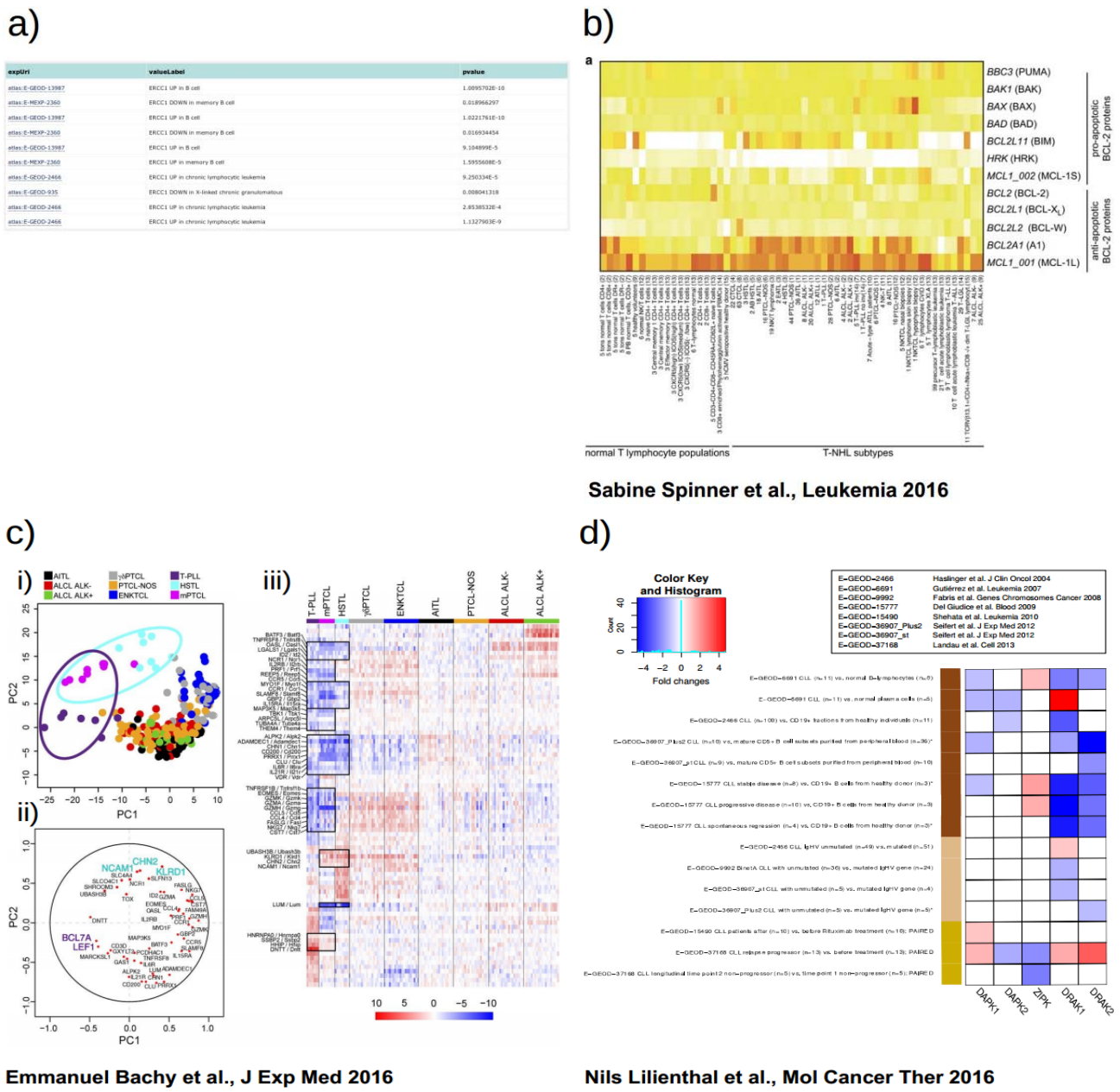


Figure 3. a) Potential ERCC1 deregulations in normals B-cells, B-cell lymphomas / leukemias (mantle-cell lymphoma, chronic lymphocytic leukemia (CLL) and chronic myeloid leukemia (CML)) and chronic conditions are queried within EMBL / EBI Gene Expression Atlas RDF (see Table 3 for exact query). The output, in table format, can be further exported into e.g. csv format. Fold-changes can be further visualized as in *c)*. **b)** Example taken from [49] (Fig. 1a): mature T-cell lymphomas and normal T-cell subsets are grouped by expression of pro- and anti-apoptotic *BCL2* family genes / isoforms. The long *MCL1* isoform seems to be used throughout malignant and benign T-cells, while *BCL2A1* and *BCL2L11* seem to be especially upregulated in malignant T-cells. Samples were quantile-normalized on the basis of 12 markers. **c)** Example taken from [50]. *i+iii)* illustrating different unsupervised clustering results (principal component analysis and heatmap) as *CD1d*-restricted murine

natural killer T-cell lymphoma seems to be most similar to T-cell prolymphocytic leukemia (T-PLL) and hepatosplenic T-cell lymphoma (HSTL). *ii)* Variables factor maps (produced by libraries like *FactoMineR*) show what marker contributes (or correlates) the most to each principal component and thus carries the highest specificity. Platform overlap was reduced to gene level, then batch-corrected using ComBat and quantile-normalized. **d)** Example taken from [51]. Fold-changes were calculated according to labeled comparisons for each *Death-Associated Protein Kinase (DAPK)* gene family member, then the range was cut off and results were visualized. Color bars used for 3 distinct comparisons: (1) CLL vs. normal B cells (various subtypes); (2) CLL with *IGHV* unmutated vs. mutated gene status; (3) CLL with post-to-pretreatment and other clinical comparisons.

9. Functional Analyses: the More the Merrier

In the abundance of genes obtained as significantly dysregulated, the role or function of a specific gene is often unknown and it is therefore encouraged to group them functionally by software tools often coined as ’pathway analysis‘ or ’enrichment‘ tools. One of the most user-friendly, however, costly tools is QIAGEN’s Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity/). Users can upload their differential expression results in the format of Excel tables into the Java GUI (graphical user interface). Annotation in the form of chip design or symbol identifiers (such as Gene Symbol, Ensembl ID or GenBank ID) can be selected for a given column as well as statistical parameters in separate columns, such as p-values, fold-changes, q-values / FDRs or simply expression values (fluorescence in microarrays or FPKM (fragments per kilobase of exon per million reads mapped) for RNA-seq). The list can be further restricted to a given range (e.g. p-value < 0.05). The selected genes are subsequently assembled into manually curated biological or toxicological / pharmacological pathways provided with an E-value (chance of a random hit). One advantage of IPA compared to other tools is the easy visualization of results by intuitive geometric forms, i.e. nodes / genes are drawn as distinct geometric symbols and edges / protein modifications in distinct line types. Similar graphs can be drawn with *igraph* in R, but are restricted to users that are more experienced in bioinformatics.

Other user-friendly and open-source alternatives include DAVID [26], gene set over-representation analysis (GSOA) by ConsensusPathDB [27] (Suppl. Figure 2), and gene set enrichment analysis (GSEA; Figure 4a) by the Broad Institute [28]. All three tools can be operated from web GUIs, while the first two options also offer an R implementation or in the case of GSEA, also a JAVA desktop application.

For more advanced users and those seeking to work with protein identifiers (complementary to above mentioned tools) *STRINGdb10* [29] is a potential alternative. Within the R library PPI (protein-protein interaction) graphs (nodes colored according to fold-change and also reachable via web link) and enrichments (including p-values and number of observed and expected interactions)

are calculated (Figure 4b; Code 15). Therein, inputs are the corresponding proteins of the most significantly dysregulated probes in different gene expression comparisons. Edges between proteins are colored according to evidence level, e.g. co-expression, literature mining, or experimental assays such as yeast2hybrid (y2h). The same R library can also be used for KEGG and GO (gene ontology) enrichment analyses (Code 16). RNA-to-protein inference can however only be approximate due to different half-lifes and decay rates as well as due to variable post-transcriptional and post-translational modifications.

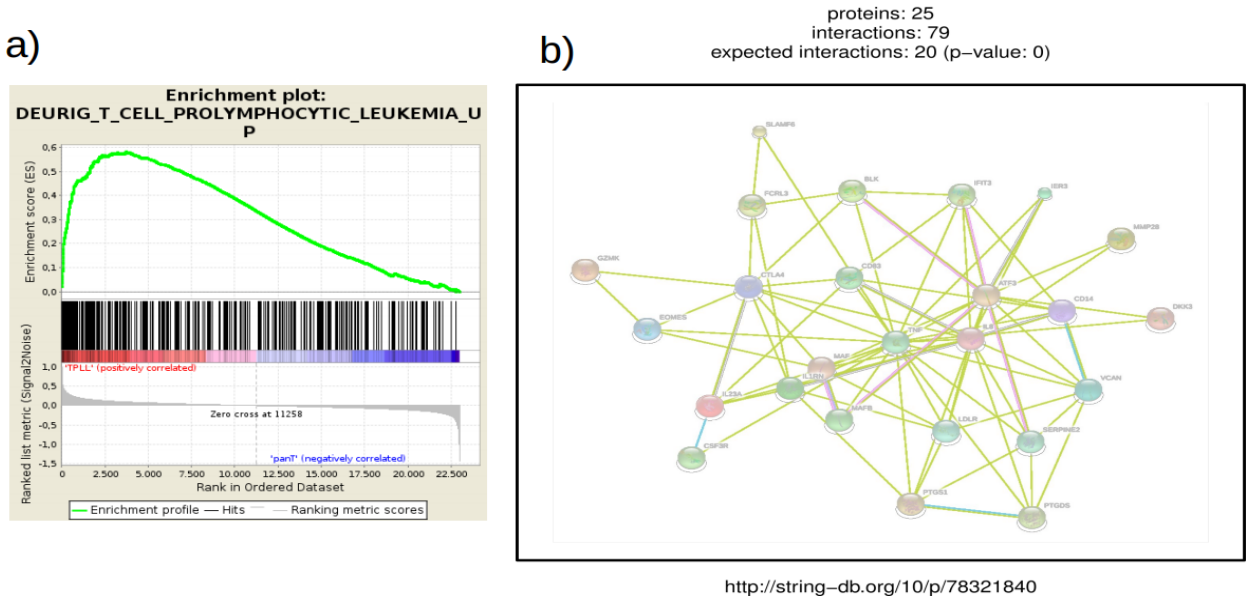


Figure 4. Results of differential expression analysis of 70 samples of T-cell prolymphocytic leukemia (T-PLL) and normal CD3+ T-cells from 10 healthy donors were further functionally annotated. a) Enrichment plot of Broad GSEA (gene set enrichment analysis) of the most deregulated (|fc|>1.5; q < 0.05) genes between T-PLL and normal CD3+ T-cells shows strong correlation (hit accumulation at the front of enrichment profile in dark and peak in green) to the results of a previous T-PLL gene expression data set [52]. **b)** Example of a PPI (protein-protein interaction) graph output from STRINGdb_v10 with a significant enrichment (59 more PPIs than expected). URL at the bottom is automatically generated and serves as an archive for the output.

10. Standard Survival Analysis and An Exploratory / Heuristic Approach

Besides parameters of more established nature (routinely tested), e.g. in CLL those from clinical chemistry, such as β_2 microglobulin [30] or from immunophenotyping, such as ZAP70 [31], the expression of a single gene or a gene set detected by microarray-based GEP can also serve as a marker, or a scored combination of them, that predict clinical outcomes. Such prognostic estimations are predominantly measured in subgroup differences of time-to-event metrics like overall survival (OS; from date of diagnosis or less correctly from first day of treatment or study randomization to last follow-up (FU) or death) or progression-free survival (PFS; from first day of treatment or randomization to disease progression or death). Other measurements include time-to-treatment (TTT; from diagnosis or randomization to first day of treatment), time-to-next-treatment (TTNT; end of first to beginning of next treatment), time-to-treatment-failure (TTF; time from diagnosis or randomization to treatment dismissal), or event-free survival (EFS; time from diagnosis or randomization to disease progression, death or treatment dismissal). These parameters are either right-censored (date of death or progression after study window, thus unknown) or left-censored (study entry is unknown) to deal with missing time points or events (death or progression). Here we focus on right-censored data.

An univariate analysis compares time-to-event parameters for two subgroups divided by a gene expression or other marker status (see [32] for an introduction). For multivariate analysis, multiple genes or markers are considered for a competing subset comparison (see [33] for an introduction). For the former there are standard methods implemented within the R library *survival* with functions *survdiff* to test the differences of survival times with the log-rank test [34] and *survfit* to plot the survival times with the Kaplan-Meier estimator [35] (Code 17). A multivariate analysis allows ranking of the most significant markers contributing to an adverse prognosis. It is usually conducted with the Cox Proportional Hazards [36] (CoxPH) model.

As evidence provided by different data sources and methods strengthens a given hypothesis, it is important to validate identified markers of prognosis in an independent patient cohort. However, this is often difficult due to a limited availability of reasonably-sized data sets for comparison. Possible causes may be a low disease incidence (e.g. notorious for mature T-cell lymphomas) or general difficulties in obtaining primary tumor samples (e.g. due to the need of invasive procedures to be consented by the patient). Another factor imposing limitations on sample size is the uniformity of received treatments, which must apply to a given patient cohort in order to reliably predict related outcomes. For GEP studies in such scenarios, we propose an alternative algorithm for the identification of prognostic gene expression signatures, which we demonstrate by the example of GEP data generated from peripheral blood tumor samples of patients with T-cell prolymphocytic leukemia (T-PLL) and CLL. We obtained gene expression profiles from 49 T-PLL samples with available OS status and from 58 chemoimmunotherapy-treated CLL patients with available PFS data, both from Illumina HumanHT-12 v4.0 Expression BeadChips.

In a first training set of 10 T-PLL, 5 patients with longest OS (time from diagnosis to death of disease, > 800 days) were compared to those with shortest OS (< 300 days, n = 5) using the 'Significance analysis of microarrays' (SAM) analysis in survival mode via the R library *samr* [37]. We only considered expression profiles from patients in whom corresponding samples had been obtained within 6 months from diagnosis (ensuring similarities between specimen and clinical data) and who had presented with similar lymphocyte doubling times as an indicator of disease kinetics at the time of sample. From an initial most informative index-set of 5 differentially expressed probes (*RAB25*, *KIAA1211L*-probe1, *KIAA1211L*-probe2, *GIMAP6*, *FXYD2;* FDR < 0.1), linear regression [38] and removal of one outlier by setting OS<200 days, resulted in a 2nd training set of nine cases. Another subsequent SAM (survival mode) resulted in a 2-gene / 3-probe set as the most robust combined predictor of OS. These probe sets were used to calculate an expression index via an additive model fit using Tukey's median polish procedure [39] (*medpolish* function within the standard *stats* library) on a test set of 40 uniformly treated T-PLL (the nine training cases excluded) fulfilling the criteria of available array data and OS information. Kaplan-Meier curves (log-rank tests for differences) were created based on stratified per patient-values of this "2-gene / 3-probe prognostic expression index" (*RAB25* and the two *KIAA1211LL* transcripts either merged or separated; Figure 5a). Ranking the cases solely based on these expression indices, the five T-PLL cases with the lowest values indeed showed significantly superior OS over those five cases with highest or 35 cases with higher (Figure 5b; Suppl. Figure 3a) expression index values (index fold-change (fc) = −2.37; Figure 5b; index fc = −1.62; Suppl. Figure 3a). A similar approach was used to identify signature genes associated with PFS in chemoimmunotherapy-treated CLL (Figure 5c; Suppl. Figure 3b; Code 18) resulting in a predictive 4-gene / 7-probe index (including *GPD1L, TNFSF12*, *JHDM1D, TBCD, AARS2, MTG1*, and *TNIP*). In both cohorts, the detected differential expression of signature genes and their association with clinical outcome requires further validation, e.g. by qRT-PCR, in independent samples before considering them further as valid markers.

11. Sample Classification by Supervised (Machine Learning) Approaches

When dealing with large data sets (e.g. a gene expression matrix) that incorporate different clinical or molecular information ('features'), and if a group status ('class') of clinical or biological interest (e.g. treatment responder vs. non-responder) is known, the application of discrimination (or supervised learning) methods can be considered. Such methods aim to train classifiers (logistic, linear, or non-linear) that are able to predict the status of future samples based on certain features (e.g. treatment response). In general, it is important to validate classification rules obtained from training data in an independent test set, preferably obtained from another set of patients from a different laboratory / trial group, in order to avoid a biased data interpretation. When there is no independent set available, an internal cross-validation can be performed. Therein, the available patient samples are repeatedly separated into a training set and a test set, while subsequently

observing the average classification performance by the number of false positives and false negatives obtained through the classifier.

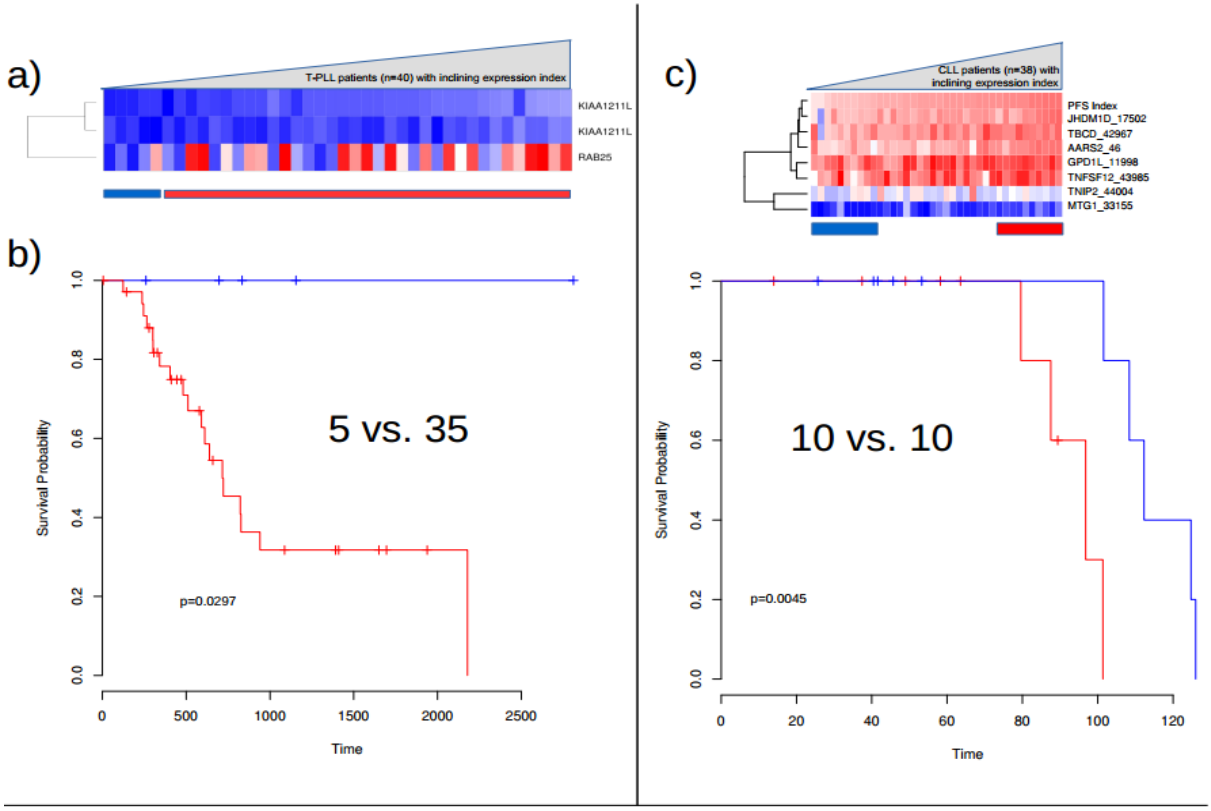


Figure 5. We explored alternative approaches to obtain prognostic values in a 49-case cohort of T-cell prolymphocytic leukemia (T-PLL) (Schrader, Crispatzu et al. submitted) with available overall survival (OS) data as well as in a chemoimmunotherapy-treated cohort of chronic lymphocytic leukemia (CLL) (Herling et al. unpublished; *n* = 58 with available progression-free survival (PFS) status). **a-b)** The five T-PLL patients with each the highest and lowest OS (without censored / alive ones) were considered for a 'Significance analysis of microarrays' (SAM) analysis in survival mode. The resulting probe sets / transcripts were used to calculate an expression index **a)** (via additive model fit using Tukey's median polish procedure) on the test set of residual cases. Kaplan-Meier (log rank; time in days) curves were created based on stratified values per patient of this 'prognostic expression index'. **b)** Five patients with lowest index expression vs. residual 35 patients of test set (see Suppl. Figure 3 for 5 vs. 5). **c)** The same approach was used for ten chemoimmunotherapy-treated CLL with the highest and lowest PFS. The index was calculated on probe set / transcript level and again evaluated in especially indolent and aggressive patient samples (here ten with lowest and highest index expression) within the test set. In both cohorts, of T-PLL and CLL, a high index expression was linked to an adverse prognosis.

A popular supervised learning approach are support vector machines [40] (SVM; R libraries *gmum.r* or *e1071*). They try to separate classes by projecting features and their interactions into high-dimensional space and subsequently by searching for either linear (Figure 6a-b) or non-linear (Figure 6c; Suppl. Figure 4) separating hyperplanes in the original feature space (Code 19).

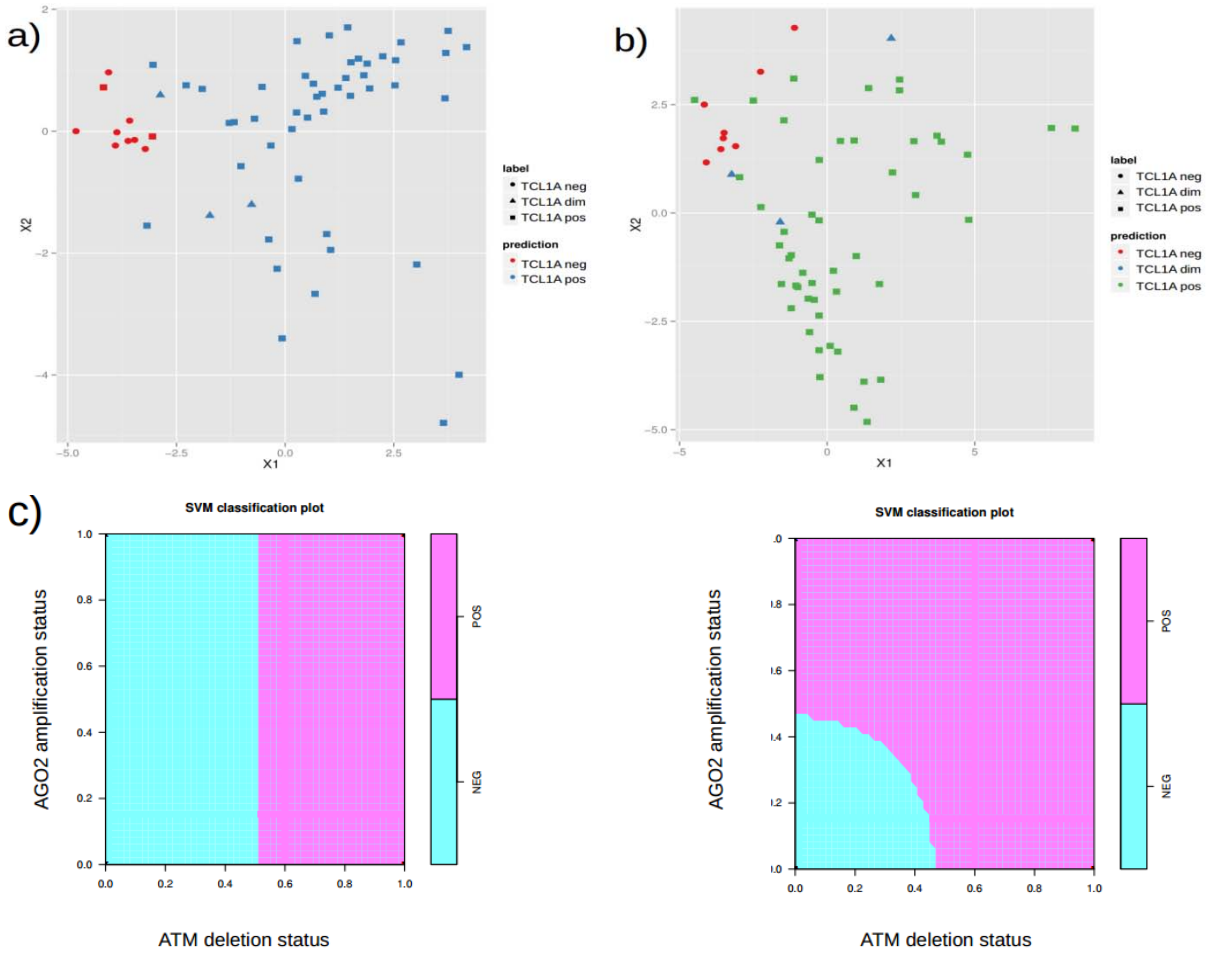
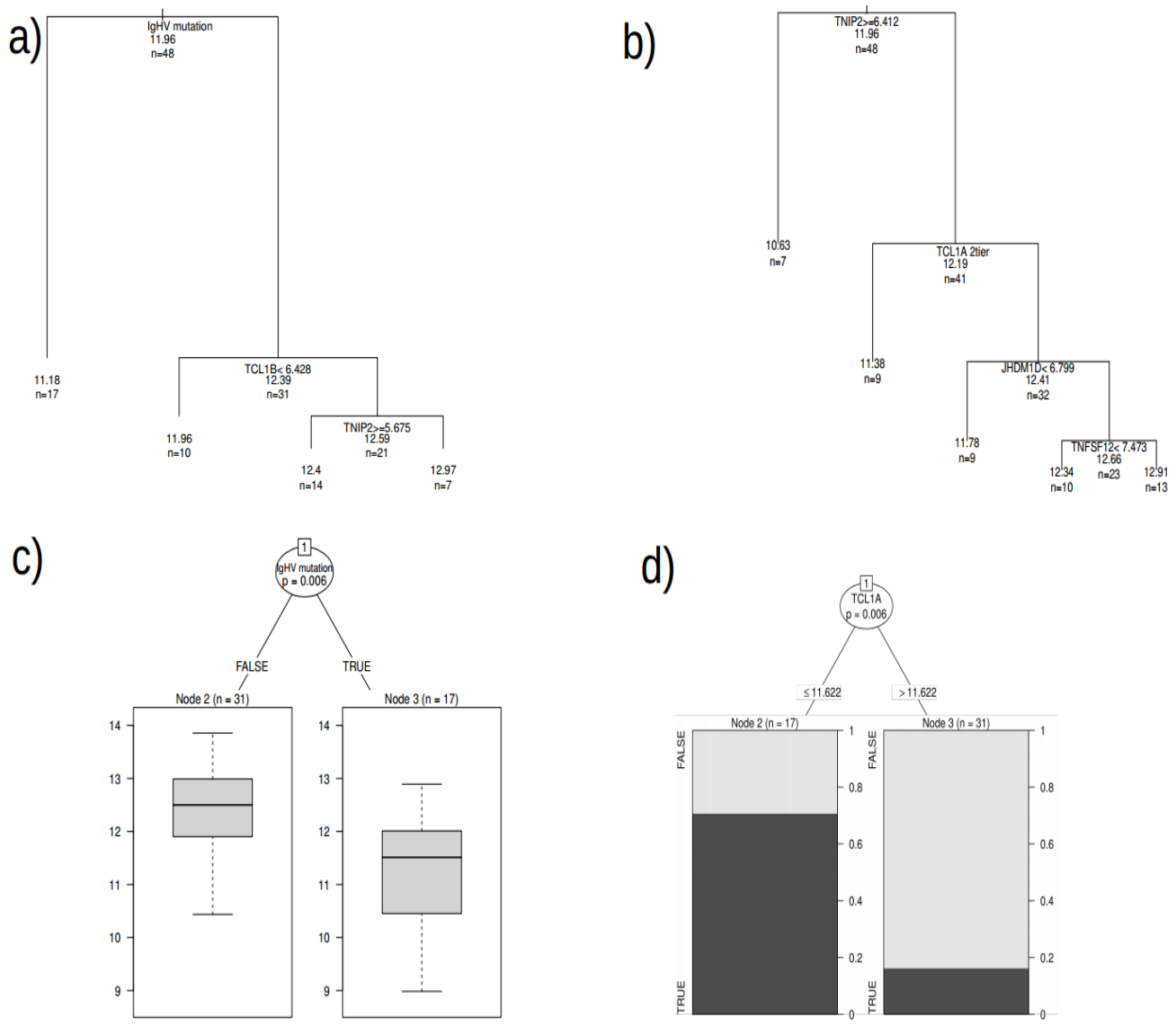


Figure 6. a) Support vector machine (SVM) classifies samples of T-cell prolymphocytic leukemia (T-PLL) based on TCL1A protein status (positive, intermediate, negative; by flow-cytometry) predicted by *TCL1A* and *TCL1B* mRNA expression. As one can see in the top left two samples are misclassified by SVM as TCL1A-negative (red, but squared symbols). **b)** SVM of T-PLL samples of different TCL1A protein status ("dim" being intermediate) by numerous mRNA markers performs more robust classification. **c)** Example of a linear (**upper panel**) and a non-linear, radial / polynomial fit (**lower panel**) of a SVM. T-PLL samples which carry the *ATM* gene in mutated vs. unmutated constitution are classified by their status of *ATM* deletion and *AGO2* amplification. Results, as seen by approximate pattern in linear and more distinct pattern in non-linear classifier, elucidating that *ATM* unmutated samples are more likely to be biallelic for *ATM* and *AGO2*.

Decision trees (R libraries *rpart, tree* or *party;* Code 20) can also divide samples according to a class variable into further most informative binary portions of gene expression signatures (Figure 7a–b) or of other molecular features (i.e. mutational or cytogenetic strata in CLL) (Figure 7c–f; Suppl. Figure 5); measured by ANOVA for numerical or by entropy for categorical values. When looking for a cut-off for adverse prognosis, they can be further used in the form of regression trees [41]. Different parameters can be controlled in this approach, such as the maximum size of a tree or the number of portions / bins. It is recommended to keep these relatively low in the training set to avoid "overfitting" and thus enable re-evaluation in the test set. Random forests [42] (as an assembly of permutated decision trees) can be used to determine the chance of observing random tree branching (library *randomForest*) (Code 21). Both algorithms are also included in the *rattle* library, which offers a user-friendly GUI with interactive plots and a selection menu for class variable and co-variates as well as algorithm and parameter choices. For a more detailed review on current machine learning algorithms in GEP, we refer to [43].



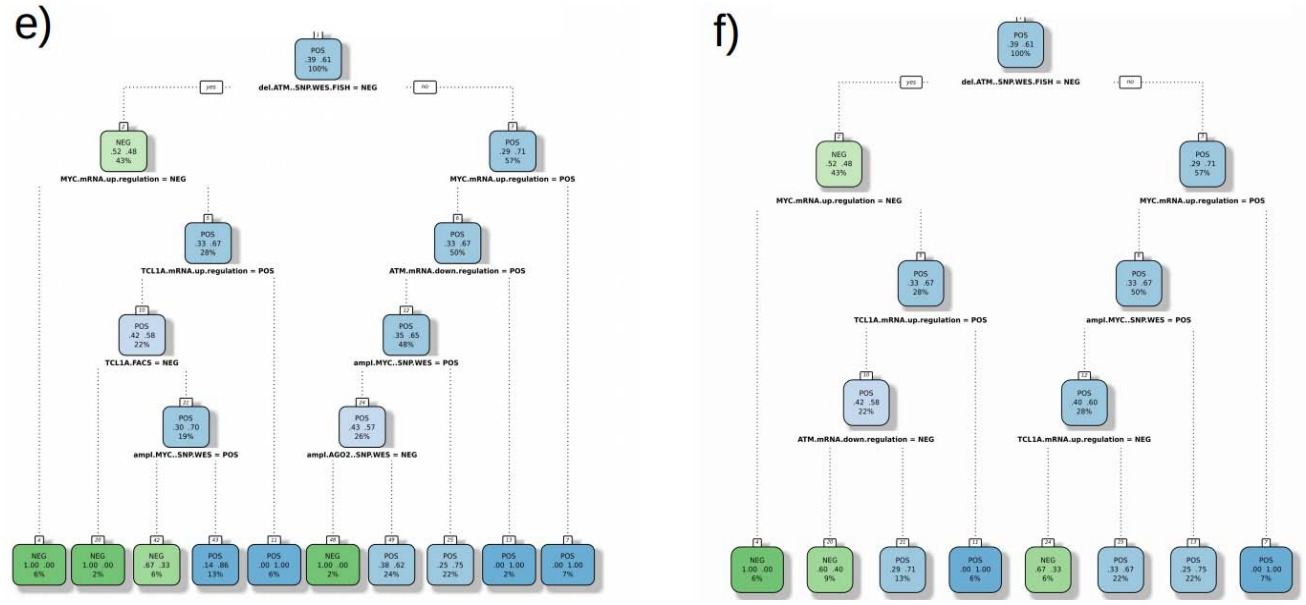


Figure 7. a) Example of a *rpart* decision tree. Chronic lymphocytic leukemia (CLL) samples are stratified according to TCL1A protein status. Model design then includes *IGHV* gene mutation status, mRNA markers linked to adverse prognosis (from algorithm described in **Figure 5c**), and further clinical features. *IGHV* gene mutations status, as seen in top of branch, is the most informative divider. When left out in **b)** an mRNA marker linked to adverse prognosis (with somewhat arbitrary cut-off for illustrative purposes) functions as the most informative divider. **c-e)** *ctree* offers more intuitive visualizations of decision trees. **c)** When stratifying CLL samples by *TCL1A* mRNA expression, *IGHV* mutations status is the most informative divider. **d)** This is confirmed when stratifying CLL samples by *IGHV* mutations status (switching the comparison) hence *TCL1A* mRNA expression is the most informative discriminator. **e)** T-cell prolymphocytic leukemia (T-PLL) samples stratified by *ATM* mutation status. Co-variates include *ATM* deletion, *miR-34B* deletion, *MYC* amplification, *AGO2* amplification, *MYC* mRNA upregulation, *ATM* mRNA downregulation, and *TCL1A* mRNA upregulation. *ATM* deletion status seems to be the most informative co-variate, however due to the excessive size of the tree (controlled by pruning and number of bins) there is a risk of "overfitting". **f)** Shown is a more feasible and smaller decision tree. Again, the most informative co-variate seems to be the status of *ATM* gene deletion. Followed by *AGO2* amplification status. This is further confirmed in random forests (permutated decision trees) in order to circumvent 'overfitting' (not shown).

12. Discussion

In this review we discuss procedures to optimize GEP analyses. We highlight the importance of advanced preprocessing, such as batch correction and admixture modeling, but also appraise the versatility and sophistication of analysis and classification algorithms. Many of the presented

methods, originally established for microarray data analysis, can also be applied to RNA-seq data (on the basis of read counts instead of fluorescence values). In addition to GEP, it is always desirable to aim for additional genetic information, including (somatic) copy-number alterations, structural variation, and genotyping of nucleotide variants for a most comprehensive genetic workup of the investigated cancer specimen. Epigenomic data, e.g. from methylome and ChIP-seq experiments may be added as a second layer. Besides setting up an own data repository in MySQL or RDF for managing internal data, one may also investigate the cBioPortal for Cancer Genomics [44]. TCGA (https://tcga-data.nci.nih.gov/tcga), ICGC (https://dcc.icgc.org), and other large curated data sets provide user-friendly search engines with multiple visualization options. Another helpful tool for combining gene expression data with available genomic knowledge in a network-based analysis is Expander [45]. Overall, this review and the attached source codes may provide guidance to both molecular biologists and bioinformaticians / biostatisticians to properly conduct GEP analyses from microarrays and to go beyond the application of standard analytic tools to optimally interpret the clinical and biological relevance of the obtained results.

Acknowledgements

M.H. (HE3553/3-1) and C.D.H. (SCHW1711/1-1) are funded by the German Research Foundation (DFG) as part of the collaborative research group on “Exploiting the DNA damage response in CLL” (KFO286). M.H. (HE3553/4-1), has been supported by the DFG as part of the collaborative research group on mature T-cell lymphomas “CONTROL-T” (FOR1961). Further support: German Cancer Aid (108029), CECAD, José Carreras Leukemia Foundation (R12/08) (all to M.H.); CLL Global Research Foundation (to M.H. and C.D.H.); Köln Fortune Program and Fritz Thyssen foundation (10.15.2.034MN) (both to M.H. and A.S.).

We gratefully acknowledge all contributing centers and investigators enrolling patients into the trials and registry of the German CLL Study Group (GCLLSG) and at the UT M.D. Anderson Cancer Center (MDACC), Houston/TX, USA; the GCLLSG and UT MDACC staff and the patients with their families for their invaluable contributions.

Contribution of Authors

Data analysis: G.C.; survival analyses: G.C., A.S., M.H.; experiments and conduction of GEP: A.S., C.D.H.; clinical data: M.H., C.D.H.; manuscript preparation: G.C., C.D.H., M.N., M.H.

Conflicts of Interest Disclosure

There were no competing interests interfering with the unbiased conduction of this study.

Patient Samples

Human tumor samples were obtained from patients under IRB-approved protocols following written informed consent according to the Declaration of Helsinki. Collection and use have been approved for research purposes by the ethics committee of the University Hospital of Cologne (#11-319) and UT M.D. Anderson Cancer Research Center. The cohorts were selected based on uniform front-line treatment as part of the TPLL1 [46] (NCT00278213) and TPLL2 (NCT01186640, *unpublished*) prospective clinical trials as well as FCR300 [47] or included in the nation-wide T-PLL and CLL registries of the German CLL Study Group (GCLLSG, IRB# 12-146).

References

1. International Human Genome Mapping Consortium. A physical map of the human genome. (2001) *Nature* 409: 934-941.
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. (2004) *Nature* 431: 931-945.
3. Ferreira PG, Jares P, Rico D, et al. (2014) Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res* 24: 212-226.
4. Ojesina AI, Lichtenstein L, Freeman SS, et al. (2014) Landscape of genomic alterations in cervical carcinomas. *Nature* 506: 371-375.
5. Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507:315-322.
6. Wang C, McKeithan TW, Gong Q, et al. (2015) IDH2R172 mutations define a unique subgroup of patients with angioimmunoblastic T-cell lymphoma. *Blood* 126: 1741-1752.
7. R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.
8. Gentleman RC, Carey VJ, Bates DM, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80. URL: http://www.bioconductor.org/.
9. Irizarry RA, Bolstad BM, Collin F, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15.
10. Yang YH, Dudoit S, Luu P, et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30: e15.
11. Huber W, von Heydebreck A, Sueltmann H, et al. (2002) Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* 18: S96-S104.

12. Durinck S, Moreau Y, Kasprzyk A, et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21: 3439-3440.
13. Bae J, Leo CP, Hsu SY, et al. (2000) MCL-1S, a splicing variant of the antiapoptotic BCL-2 family member MCL-1, encode s a proapoptotic protein possessing only the BH3 domain. J *Biol Chem* 275: 25255-25261.
14. Noble WS (2009) How does multiple testing correction work? *Nat Biotechnol* 27: 1135-1137.
15. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. *Statistical Sci* 18: 71.
16. Pollard KS, Dudoit S, van der Laan MJ (2005) Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer.
17. Kitchen RR, Sabine VS, Sims AH, et al. (2010) Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles. *BMC Genomics* 11: 134.
18. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118-127.
19. Yoshihara K, Shahmoradgoli M, Martínez E, et al. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 4: 2612.
20. Gaujoux R, Seoighe C (2013) CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* 29: 2211-2212.
21. Barrett T, Wilhite SE, Ledoux P, et al. (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41: D991-995.
22. Kolesnikov N, Hastings E, Keays M, et al. (2015) ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 43: D1113-1116.
23. Petryszak R, Keays M, Tang YA, et al. (2016) Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* 44: D746-752.
24. Malone J, Holloway E, Adamusiak T, et al. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26: 1112-1118.
25. Taminau J, Meganck S, Lazar C, et al. (2012) Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. *BMC Bioinformatics* 13: 335.
26. Dennis Jr G, Sherman BT, Hosack DA, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 4: P3.
27. Kamburov A, Stelzl U, Lehrach H, et al. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 41; D793-800.
28. Subramanian A, Tamayo P, Mootha VK, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U. S. A.* 102: 15545-15550.

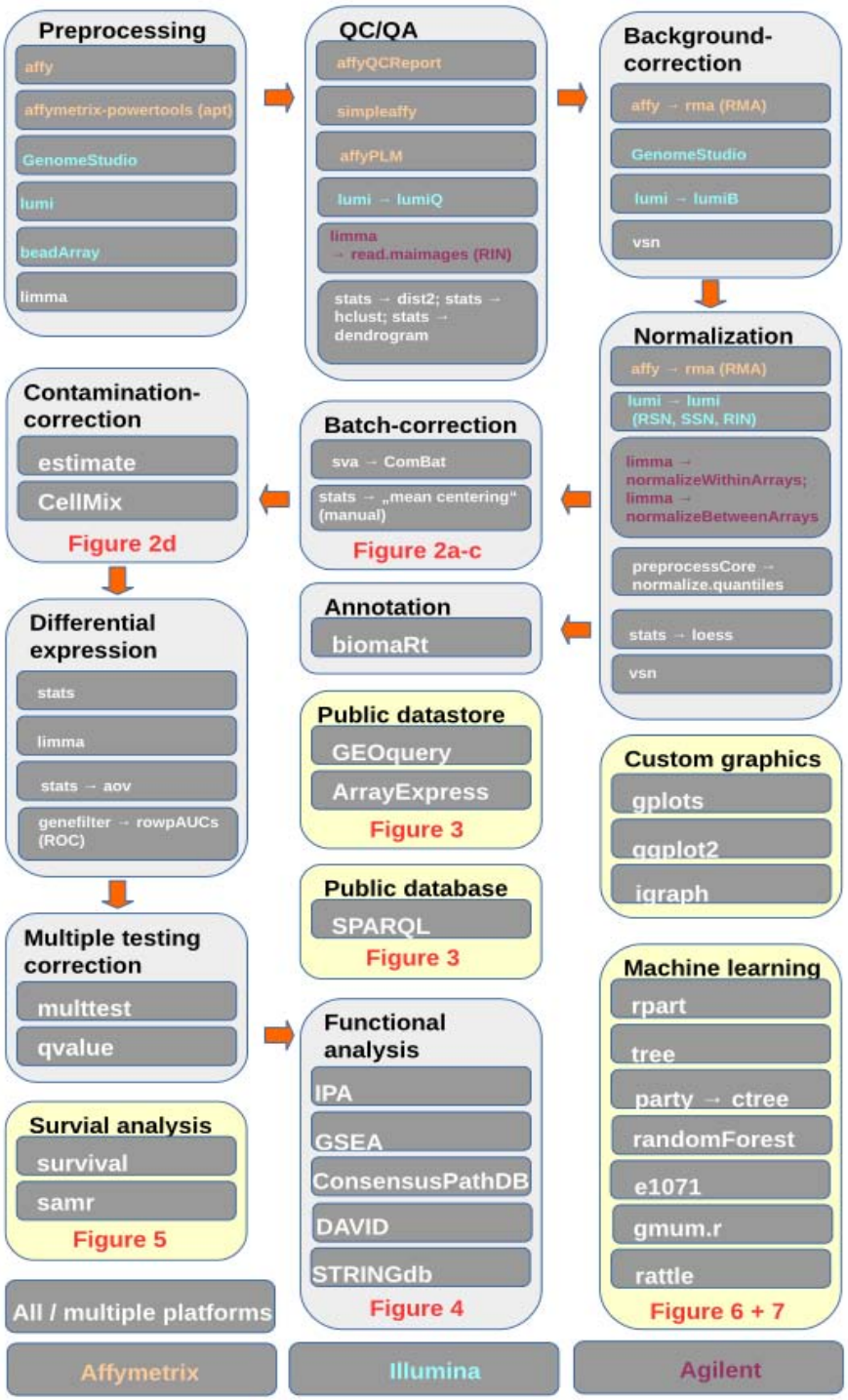
29. Szklarczyk D, Franceschini A, Wyder S, et al. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43: D447-452.
30. Gentile M, Cutrona G, Neri A, et al. (2009) Predictive value of beta2-microglobulin (beta2-m) levels in chronic lymphocytic leukemia since Binet A stages. *Haematologica* 94: 887-888.
31. Wiestner A, Rosenwald A, Barry TS, et al. (2003) ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood* 101: 4944-4951.
32. Clark TG, Bradburn MJ, Love SB, et al. (2003) Survival analysis part I: basic concepts and first analyses. *Br J Cancer* 89:232-238.
33. Bradburn MJ, Clark TG, Love SB, et al. (2003) Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer* 89: 431-436.
34. Peto R, Pike MC, Armitage P, et al. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 35: 1-39
35. Kaplan EL, MeierP (1958) Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53: 457-481
36. Cox DR. (1972) Regression models and life tables (with discussion). *JR Statist Soc B*: 34187-34220.
37. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98: 5116-5121.
38. Chen K (2001) Generalized case-cohort sampling. *Statistical Methodol* 63: 791-809.
39. Tukey JW (1977) Exploratory Data Analysis. Addison-Wesley.
40. Vapnik VN (1995) The Nature of Statistical Learning Theory. Berlin: Springer
41. Breiman L, Friedman J, Stone CJ, et al. (1984) Classification and Regression Trees. Chapman and Hall/CRC.
42. Ho TK (1995) Random Decision Forests Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August pp. 278-282.
43. Hahne F, Huber W, Gentleman R, et al. (2008) Bioconductor Case Studies. Springer Press.
44. Gao J, Aksoy BA, Dogrusoz U, et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6: l1.
45. Ulitsky I, Maron-Katz A, Shavit S, et al. (2010) Expander: from expression microarrays to networks and functions. *Nature Protocols* 5: 303-322.
46. Hopfinger G, Busch R, Pflug N, et al. (2013) Sequential chemoimmunotherapy of fludarabine, mitoxantrone, and cyclophosphamide induction followed by alemtuzumab consolidation is effective in T-cell prolymphocytic leukemia. *Cancer* 119: 2258–2267.
47. Keating MJ, O'Brien S, Albitar M, et al. (2005) Early results of a chemoimmunotherapy regimen of fludarabine, cyclophosphamide, and rituximab as initial therapy for chronic lymphocytic leukemia. *J Clin Oncol* 23: 4079-4088.

48. Sotiriou C, Wirapati P, Loi S, et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 98: 262-272.
49. Spinner S, Crispatzu G, Yi JH, et al. (2016) Re-activation of mitochondrial apoptosis inhibits T-cell lymphoma survival and treatment resistance. Leukemia. Mar 8.
50. Bachy E, Urb M, Chandra S, et al. (2016) CD1d-restricted peripheral T cell lymphoma in mice and humans. *J Exp Med* 213: 841-857.
51. Lilienthal N, Lohmann G, Crispatzu G, et al. (2016) A Novel Recombinant Anti-CD22 Immunokinase Delivers Proapoptotic Activity of Death-Associated Protein Kinase (DAPK) and Mediates Cytotoxicity in Neoplastic B Cells. *Mol Cancer Ther* 29.
52. Dürig J, Bug S, Klein-Hitpass L, et al. (2007) Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia* 21: 2153-2163.

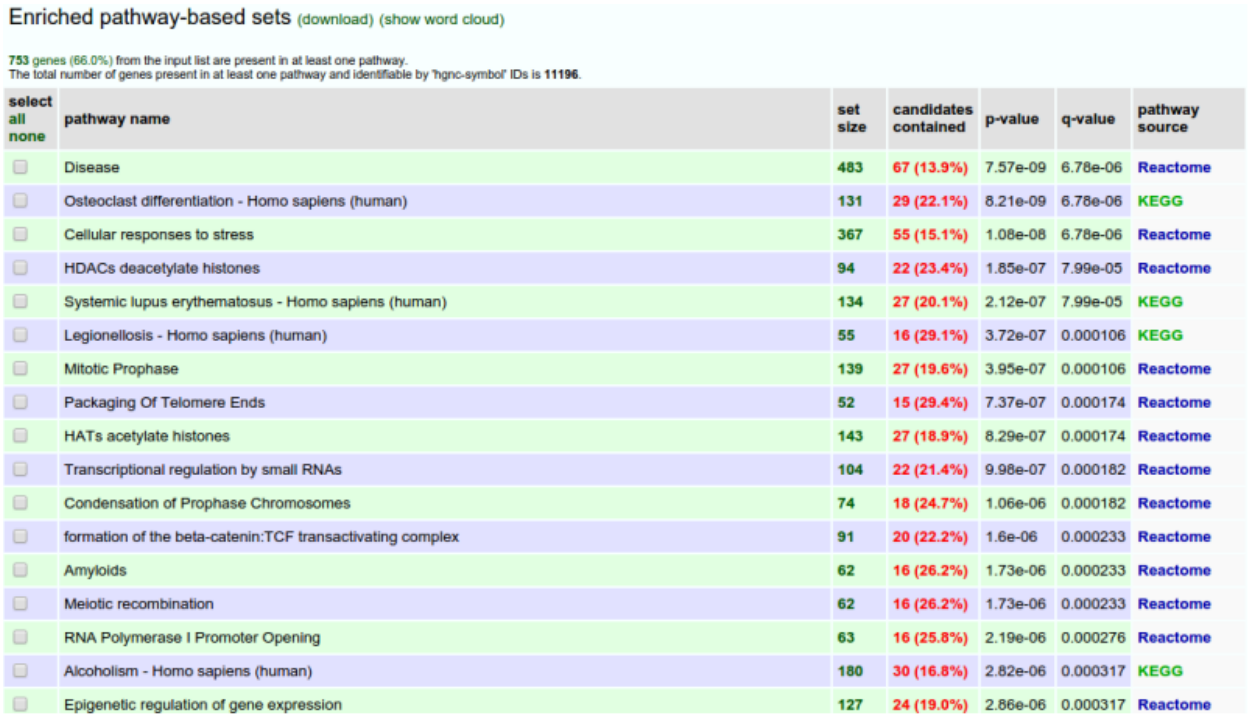
AIMS Press

© 2016 Giuliano Crispatzu et al., licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)

Supplement Materials



Suppl. Figure 1. Flow chart describing a GEP protocol. Steps in yellow boxes are modular and may be applied somewhere downstream.

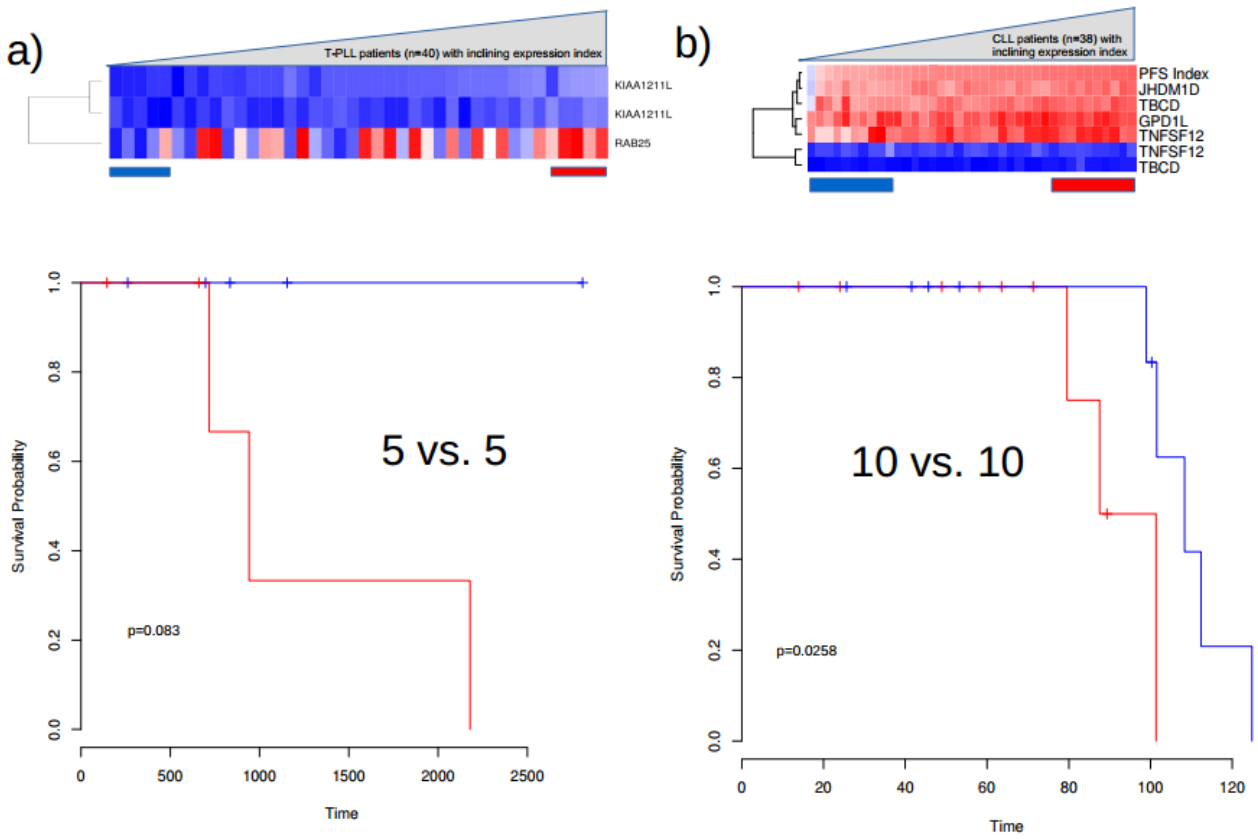


Enriched pathway-based sets (download) (show word cloud)

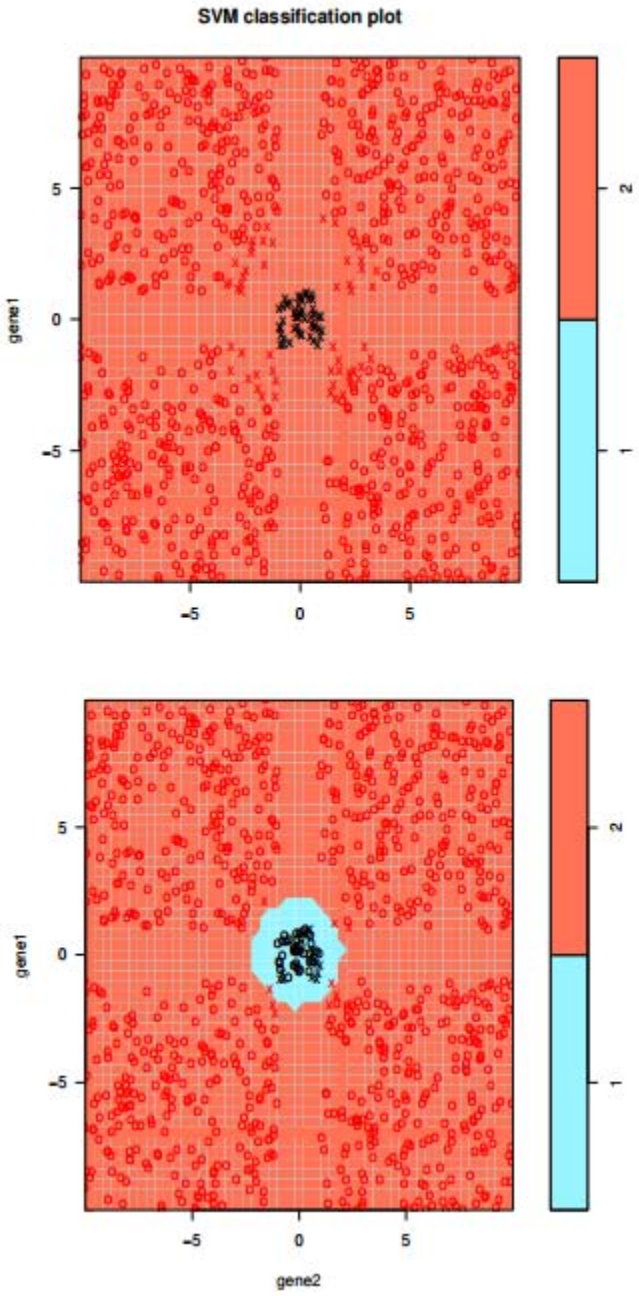
753 genes (66.0%) from the input list are present in at least one pathway.
The total number of genes present in at least one pathway and identifiable by 'hgnc-symbol' IDs is 11196.

select all none	pathway name	set size	candidates contained	p-value	q-value	pathway source
	Disease	483	67 (13.9%)	7.57e-09	6.78e-06	Reactome
	Osteoclast differentiation - Homo sapiens (human)	131	29 (22.1%)	8.21e-09	6.78e-06	KEGG
	Cellular responses to stress	367	55 (15.1%)	1.08e-08	6.78e-06	Reactome
	HDACs deacetylate histones	94	22 (23.4%)	1.85e-07	7.99e-05	Reactome
	Systemic lupus erythematosus - Homo sapiens (human)	134	27 (20.1%)	2.12e-07	7.99e-05	KEGG
	Legionellosis - Homo sapiens (human)	55	16 (29.1%)	3.72e-07	0.000106	KEGG
	Mitotic Prophase	139	27 (19.6%)	3.95e-07	0.000106	Reactome
	Packaging Of Telomere Ends	52	15 (29.4%)	7.37e-07	0.000174	Reactome
	HATs acetylate histones	143	27 (18.9%)	8.29e-07	0.000174	Reactome
	Transcriptional regulation by small RNAs	104	22 (21.4%)	9.98e-07	0.000182	Reactome
	Condensation of Prophase Chromosomes	74	18 (24.7%)	1.06e-06	0.000182	Reactome
	formation of the beta-catenin:TCF transactivating complex	91	20 (22.2%)	1.6e-06	0.000233	Reactome
	Amyloids	62	16 (26.2%)	1.73e-06	0.000233	Reactome
	Meiotic recombination	62	16 (26.2%)	1.73e-06	0.000233	Reactome
	RNA Polymerase I Promoter Opening	63	16 (25.8%)	2.19e-06	0.000276	Reactome
	Alcoholism - Homo sapiens (human)	180	30 (16.8%)	2.82e-06	0.000317	KEGG
	Epigenetic regulation of gene expression	127	24 (19.0%)	2.86e-06	0.000317	Reactome

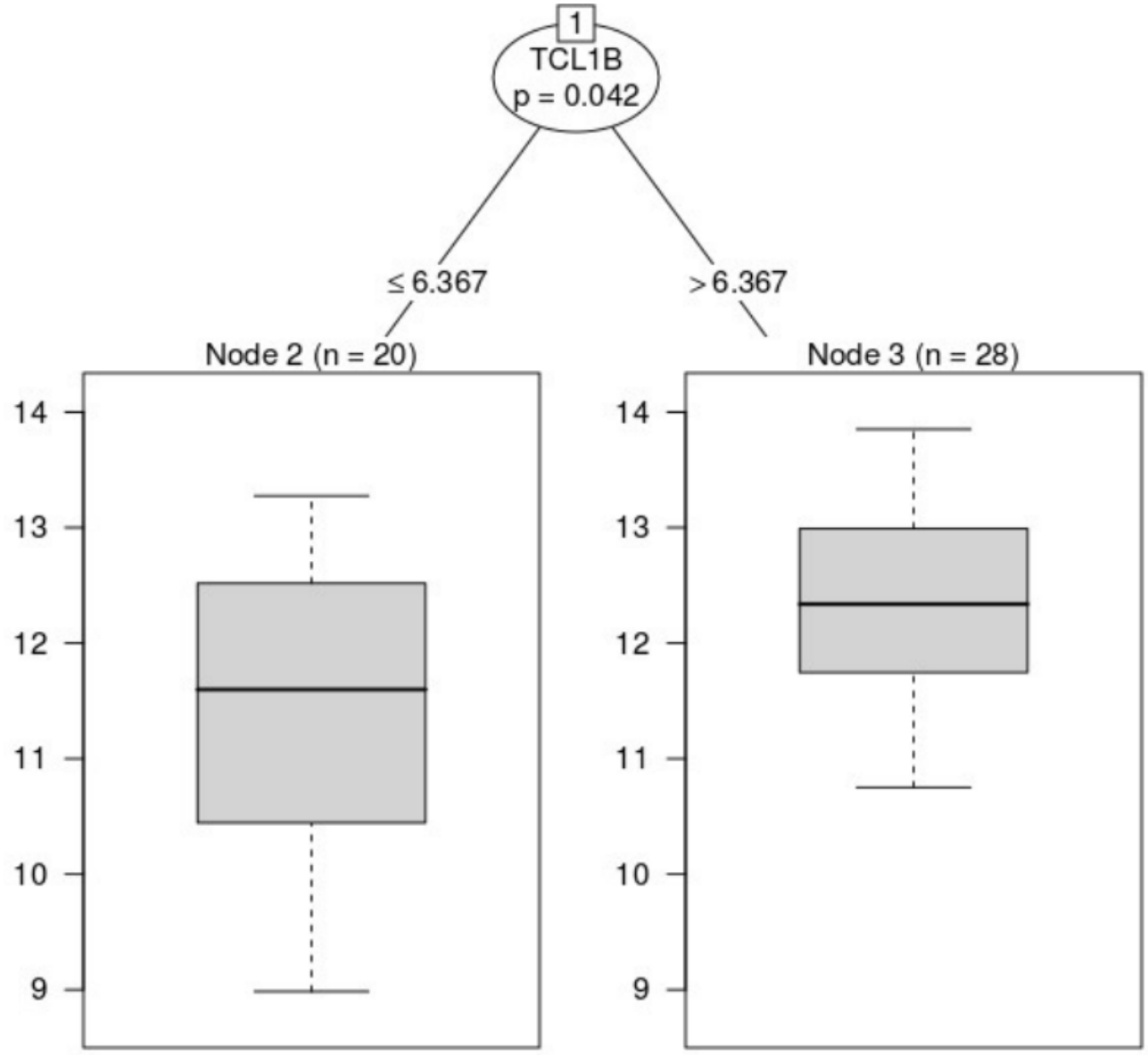
Suppl. Figure 2. Screenshot of the results from the gene set over-representation analysis (GSOA) by ConsensusPathDB with the most upregulated (|fc| > 1.5; *q* < 0.05) genes between T-PLL and normal CD3+ T-cells from healthy donors as input.



Suppl. Figure 3. We explored alternative approaches to obtain prognostic values in a T-PLL cohort (Schrader, Crispatzu et al. submitted; *n* = 49 with available overall survival (OS) status), as well as in a chemoimmunotherapy-treated CLL cohort (Herling et al. unpublished; *n* = 58 with available progression-free surival (PFS) status). a-b) The five T-PLL patients with the highest and lowest OS (without censored / alive ones) were considered for a "Significance analysis of microarrays" (SAM) analysis in survival mode. The resulting probe sets/transcripts were used to calculate an expression index **a)** (via additive model fit using Tukey's median polish procedure) on the test set of residual cases. Kaplan-Meier (log rank; Time in days) curves were created based on stratified values per patient of this "prognostic expression index". **b)** Five patients with lowest index expression vs. five patients with highest index expression within test set. **c)** The same approach was used for ten chemoimmunotherapy-treated CLL with the highest and lowest PFS. The index was calculated on gene level and evaluated in 10 patients with lowest and highest index expression within test set. In both cohorts, of CLL and T-PLL, a high index expression was linked to adverse prognosis.



Suppl. Figure 4. Expression values of two genes within two predetermined classes are simulated to further show importance of appropriate classifier scale. Upper panel: Linear classifier fails to separate classes. **Lower panel:** shows more satisfying example of a non-linear separation through (sets of) hyperplane(s).



Suppl. Figure 5. *ctree* offers more intuitive visualizations of decision trees. When stratifying CLL samples by *TCL1A* mRNA expression, *IGHV* gene mutations status is the most informative divider. This is confirmed when stratifying CLL samples by *IGHV* mutations status (switching the comparison) hence *TCL1A* mRNA expression is the most informative discriminator. When leaving *IGHV* mutation status out in Figure 7 c), then *TCL1B* mRNA expression is the next best divider.

6. Semantic Web approaches: Case studies

To demonstrate how the Semantic Web paradigm is supposed to be used for the development of the semantic framework and how the user can set the framework up, I will describe very detailed below the models that are applied to different data set classes in form of case studies with accompanying source code snippets in R and SPARQL. The information stored in these models can be cross-linked i.e. using a controlled vocabulary (or when further refined: a unified ontology) with unique gene names (e.g. *TCL1A*). This linkage of information brings up new hypotheses, indirect connections and a broader picture of knowledge and is also easily (globally) sharable and self-descriptive.

In a data-driven approach, key findings are modeled in semantic schemas to capture all necessary information. These files, containing millions of RDF triples (*subject predicate object .*), are then stored in a „triple store“ (semantic database; see **Figure 6.1**). The information is retrievable by the user through SPARQL queries and was mainly used here to generate integrated analyses of the data from the M. Herling / C. D. Herling (formely C. D. Schweighofer) group experiments.

Each high-throughput data set and Excel sheet containing clinical data was iteratively refined by adding attributes and URIs (with default namespace *gen:*). If dealing with ambiguous information, a blank node (see **6.2**) was inserted. After conversion by custom-written scripts to the RDF Notation 3 (n3), every model was loaded into an OpenRDF-Sesame data store (URL: http://www.openrdf.org/) wrapped around a Java-based HTTP servlet (URL: http://www.eclipse.org/jetty/ , *jetty-6.1.26*) enabling specific access to the data models used through queries.

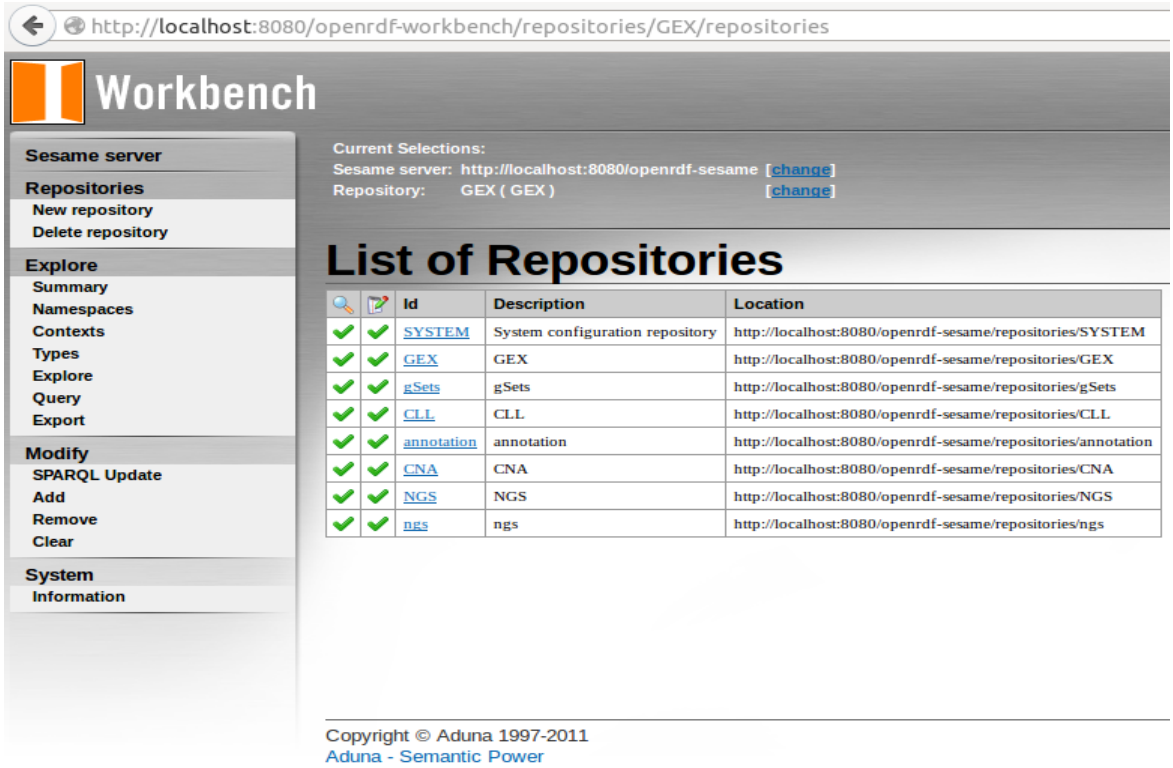
Approximations of these models were visualized with Cytoscape v.2.8.1 (URL: ttp://www.cytoscape.org/). Alternatively Protégé (URL: http://protege.stanford.edu/) can be used to view whole ontologies.

I separated public and private data into two separate triple stores. The latter is only accessible for lab members (by University of Cologne namespace / domain, htaccess password, IP, MAC address or a combination of them) through a web-GUI front-end and direct queries to private data (**Figure 6.2**). The public triple-store can be mirrored with predefined queries through the PHP library *sparqllib.php* (SPARQL RDF library for PHP; ©2010-2012 Christopher Gutteridge, University of Southhampton) and it is thus further possible to navigate through each patient by listing all high-throughput analysis results and the non-dereferencable clinical data **(Figure 6.3a**) and through each gene by listing each alteration in each patient (**Figure 6.3b**).

Upload and automatic conversion into the RDF format (and storage) can be included so that researchers and clinicians can use these models for information exchange, complementary to tables in spreadsheet format. The triple store (making use of the underlying graph structure) then makes it possible to automatically combine knowledge through graph algorithms. Either directly with the Cytoscape plug-in *RDFscape* and Jena or through SPARQL and *RCytoscape* within R in combination with OpenRDF-Sesame (as done here).

Attributes are only linked / compared to ontology terms in **Table S6.1**, and not replaced, to accurately mirror the process of modeling. Often one starts with a set of attributes in order to not overmodel right away, and just then replace iteratively old with existing vocabulary. Sometimes ontology terms are also replaced by more frequently used ones.

a)



b)

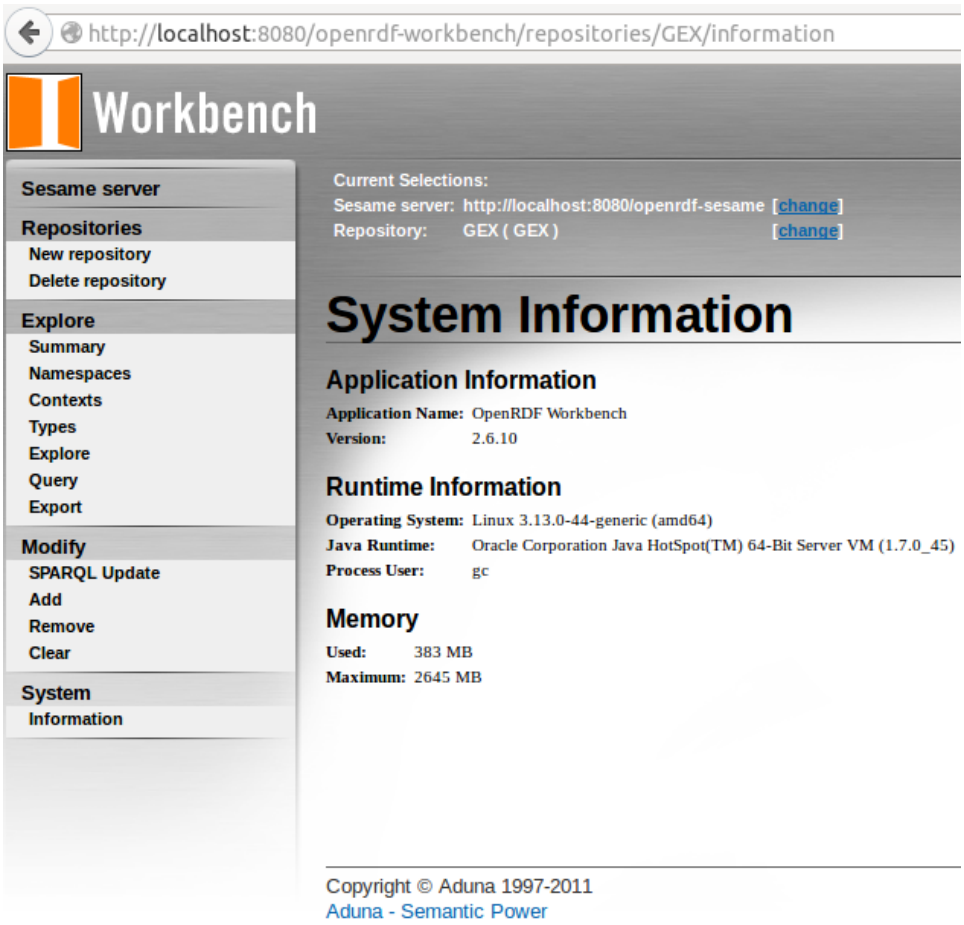
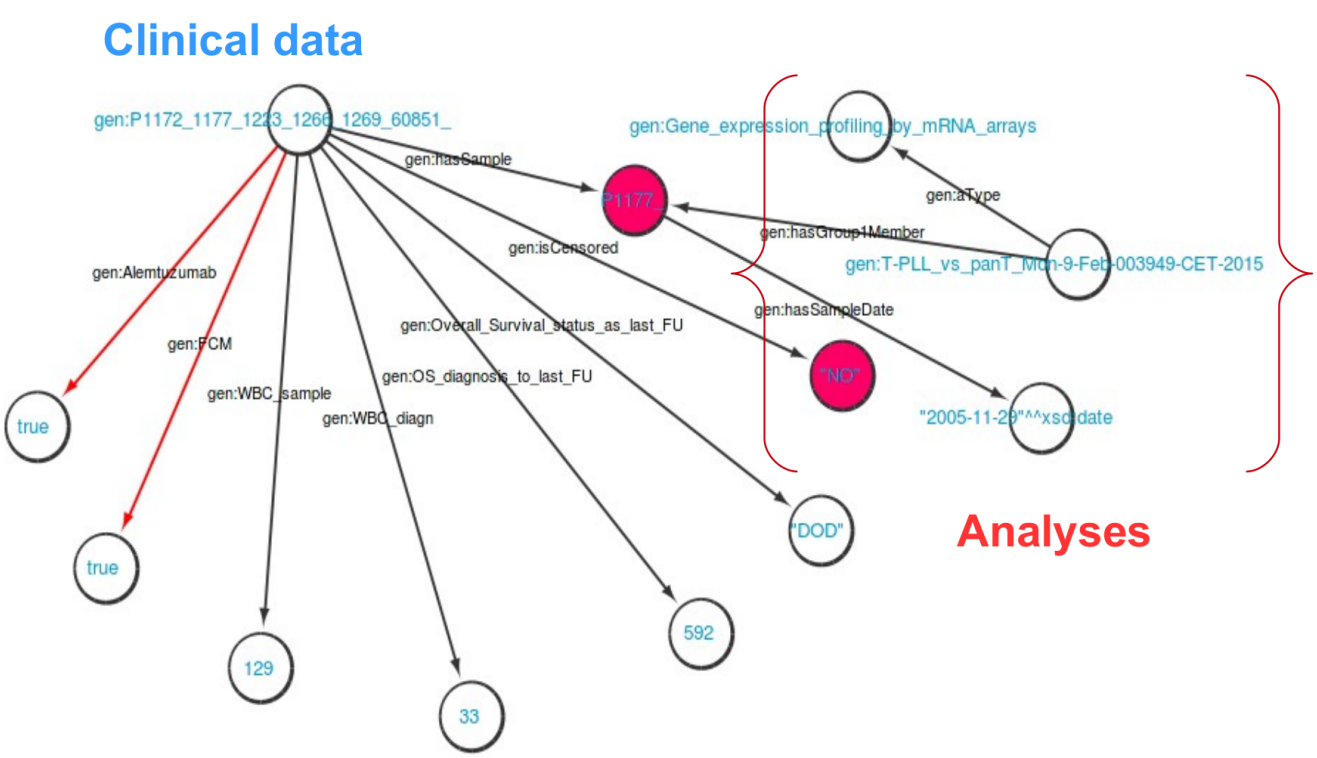


Figure 6.1: **a)** List of data repositories used within our OpenRDF-Sesame triple store. Each repository is created as a „Native Java Store“ to keep the memory usage down and will contain multiple uploaded RDF/n3 files. Multiple repositories are preferred above on single, because it eases export, import and speeds up queries due to lower solution space. **b)** System information of used desktop PC on which OpenRDF-Sesame ran. When not running any queries, only 14.5% memory were used. Default namespace (URI prefix) is: *PREFIX gen: <http://localhost:8080/openrdf-sesame/repositories/general#>*

a)



b)

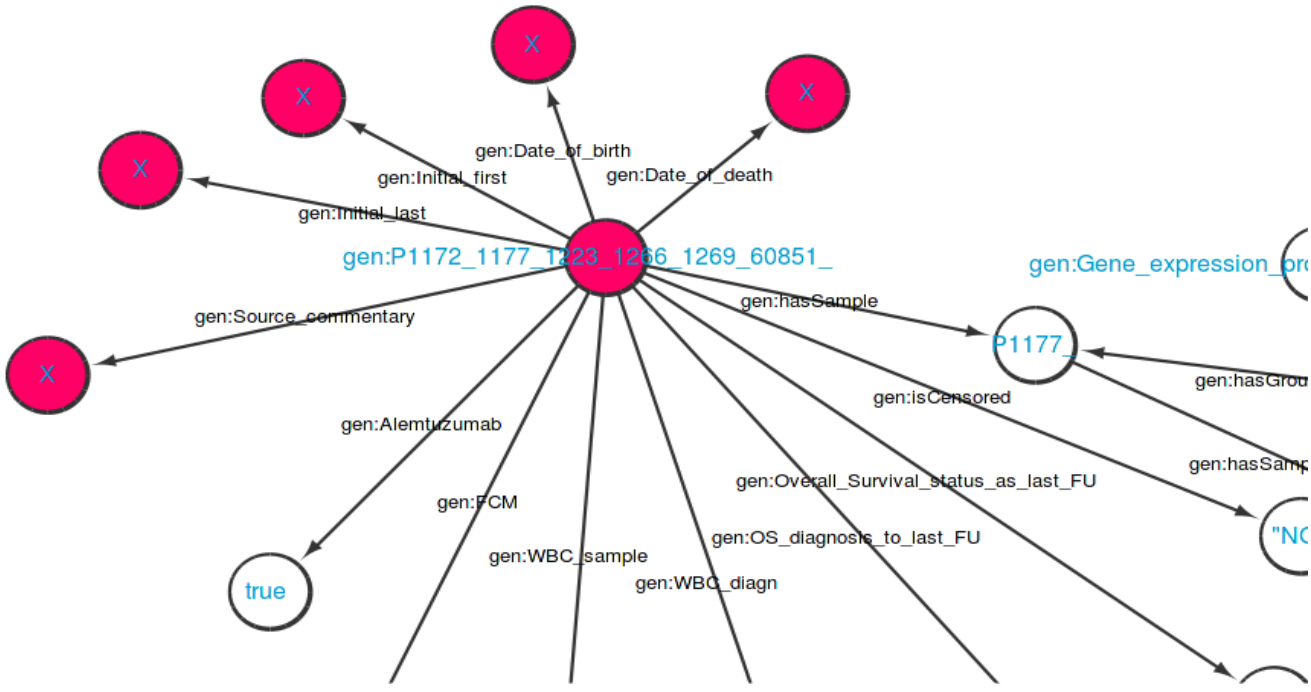


Figure 6.2: **a)** Public store model. Clinical data is enclosed in blue, while high-throughput analyses are enclosed in red. Both can be combined through the *PATIENT_ID* („P1177_“, dark red node) **b)** Private store model with added de-anonymizable data (dark red nodes), which is only accessible within jetty-6.1.26 / OpenRDF-Sesame 2.6.10 with htaccess and IP-restriction or in newer versions (URL: http://rdf4j.org), and better HTTP support, with the jetty security concept *realms*.

a)

CECAD COLOGNE

CECAD RDF Platform

SPARQL Query results

Number of rows: 10 results.

predicate	object
collectedBy	[redacted]
dateSpecimen	2011-04-28
initials	[redacted]
birthDate	[redacted]
AgeAtDiagn	72
sex	male
stageAtDiagnWBC	advanced
DateOfDisease	2011-04-14
DateOfFirstTherapy	2011-04-29
TDT	15

Number of rows: 9 results.

analyses	Type
P1231_Mi-5-Feb-151614-CET-2014	RNA-Sequencing
invy_MH1231_N.list_Fr-4-Apr-180225-CEST-2014	#Inversion_by_Whole-genome_sequencing
invy_MH1231_T.list_Fr-4-Apr-180225-CEST-2014	#Inversion_by_Whole-genome_sequencing
reAl_Exome_mut_3_Do-10-Apr-162915-CEST-2014	somatic_SNV_by_Whole-exome_sequencing
1231_Do-10-Apr-162902-CEST-2014	somatic_SNV_by_Whole-genome_sequencing
1231_Do-10-Apr-162853-CEST-2014	somatic_SNV_by_Whole-genome_sequencing
Sample_20821_Do-10-Apr-173226-CEST-2014	Tandem_duplication_by_Whole-exome_sequencing
Sample_20822_Do-10-Apr-173226-CEST-2014	Tandem_duplication_by_Whole-exome_sequencing
P1231_Di-17-Jun-153856-CEST-2014	Copy-Number_Variation

Number of rows: 10 results.

Whole_exome_mut	tumor_f	chr	pos	ref	alt
http://bio2rdf.org/hugo:SCARNA17.2	0.727273	22	21899247	C	T
http://bio2rdf.org/ensembl:ENSG00000252143	0.727273	22	21899247	C	T
http://bio2rdf.org/hugo:AGGF1P2	0.5	10	135455651	C	T
http://bio2rdf.org/ensembl:ENSG00000233435	0.5	10	135455651	C	T
http://bio2rdf.org/hugo:ADAM20P3	0.444444	4	188668994	G	T
http://bio2rdf.org/ensembl:ENSG00000249162	0.444444	4	188668994	G	T
http://bio2rdf.org/hugo:GOLGA8A	0.363636	15	34691375	A	G
http://bio2rdf.org/ensembl:ENSG00000175265	0.363636	15	34691375	A	G
http://bio2rdf.org/hugo:TEKT4P2	0.333333	21	9907963	G	T
http://bio2rdf.org/ensembl:ENSG00000188681	0.333333	21	9907963	G	T

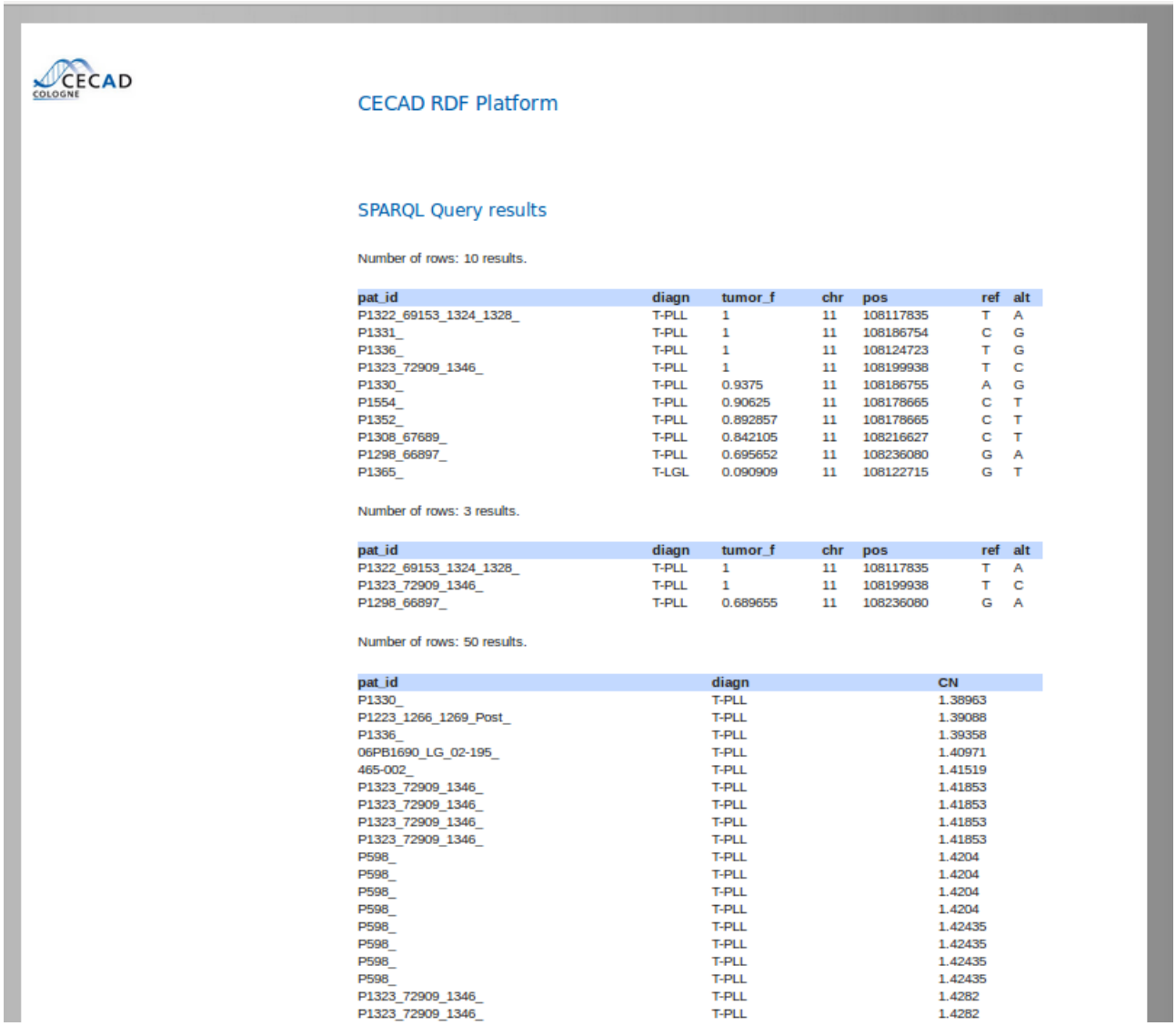
Number of rows: 10 results.

Whole_genome_mut	tumor_f	chr	pos	ref	alt
http://bio2rdf.org/hugo:KCNJ11	0.357143	11	17410432	A	C
http://bio2rdf.org/ensembl:ENSG00000187486	0.357143	11	17410432	A	C
http://bio2rdf.org/hugo:COL4A2-AS2	0.25	13	111109293	G	A
http://bio2rdf.org/ensembl:ENSG00000224821	0.25	13	111109293	G	A
http://bio2rdf.org/hugo:CRIPAK	0.25	4	1388819	T	C
http://bio2rdf.org/ensembl:ENSG00000179979	0.25	4	1388819	T	C
http://bio2rdf.org/hugo:CLIC6	0.235294	21	36089157	C	A
http://bio2rdf.org/ensembl:ENSG00000159212	0.235294	21	36089157	C	A
http://bio2rdf.org/hugo:SLITRK5	0.2	13	88330921	A	C
http://bio2rdf.org/ensembl:ENSG00000165300	0.2	13	88330921	A	C

Number of rows: 10 results.

CopyNumber_Variation	CN
http://bio2rdf.org/hugo:SDHDP6	2.70307
http://bio2rdf.org/hugo:KIR2DS4	2.4844
http://bio2rdf.org/hugo:LRRC37A	2.40677
http://bio2rdf.org/hugo:ARL17A	2.40677
http://bio2rdf.org/hugo:KANSL1-AS1	2.40677
http://bio2rdf.org/hugo:ARL17B	2.40677

b)



CECAD RDF Platform

SPARQL Query results

Number of rows: 10 results.

pat_id	diagn	tumor_f	chr	pos	ref	alt
P1322_69153_1324_1328_	T-PLL	1	11	108117835	T	A
P1331_	T-PLL	1	11	108186754	C	G
P1336_	T-PLL	1	11	108124723	T	G
P1323_72909_1346_	T-PLL	1	11	108199938	T	C
P1330_	T-PLL	0.9375	11	108186755	A	G
P1554_	T-PLL	0.90625	11	108178665	C	T
P1352_	T-PLL	0.892857	11	108178665	C	T
P1308_67689_	T-PLL	0.842105	11	108216627	C	T
P1298_66897_	T-PLL	0.695652	11	108236080	G	A
P1365_	T-LGL	0.090909	11	108122715	G	T

Number of rows: 3 results.

pat_id	diagn	tumor_f	chr	pos	ref	alt
P1322_69153_1324_1328_	T-PLL	1	11	108117835	T	A
P1323_72909_1346_	T-PLL	1	11	108199938	T	C
P1298_66897_	T-PLL	0.689655	11	108236080	G	A

Number of rows: 50 results.

pat_id	diagn	CN
P1330_	T-PLL	1.38963
P1223_1266_1269_Post_	T-PLL	1.39088
P1336_	T-PLL	1.39358
06PB1690_LG_02-195_	T-PLL	1.40971
465-002_	T-PLL	1.41519
P1323_72909_1346_	T-PLL	1.41853
P1323_72909_1346_	T-PLL	1.41853
P1323_72909_1346_	T-PLL	1.41853
P1323_72909_1346_	T-PLL	1.41853
P598_	T-PLL	1.4204
P598_	T-PLL	1.4204
P598_	T-PLL	1.4204
P598_	T-PLL	1.4204
P598_	T-PLL	1.42435
P598_	T-PLL	1.42435
P598_	T-PLL	1.42435
P598_	T-PLL	1.42435
P1323_72909_1346_	T-PLL	1.4282
P1323_72909_1346_	T-PLL	1.4282

Figure 6.3: When the outside user is not supposed to access the SPARQL-point directly, queries can be predefined, inserted into *sparqlllib.php* PHP code and thus mirrored through HTML web pages. **a)** Patient view gives an overview of all aberrations plus clinical data of the individual. Sensitive data are blurred out here and are only accessible through a private store. **b)** Gene view (i.e. here *ATM*) shows that the majority of T-PLL cases carry mono-allelic losses and clonal mutations. The user can further click on an attribute (e.g. a gene within the patient view or *gen:T-PLL* within the gene view) and browse through all results of a generic query where the attribute is either subject or object of any triple in the store.

In the following chapter I further present semantic applications to biological questions (mentioned in **1.3 Aims**) who profit from an integrative design. As well as frequently asked questions in the daily life of a molecular biologists which are hampered by e.g. relational database queries, such as:

- How to add a 'PATIENT_ID' into triple store.
- How to modify a 'PATIENT_ID' in triple store.
- Is gene *X* deregulated in a specific condition or disease?
- Is gene *X* mutated? And which allele or in which clonal fraction?
- How often is gene *X* mutated in a cohort and is a given mutation predicted to be damaging?
- Is gene *X*, which is mutated, also expressed (surpassing a given quantile)?
- Which genes are expressed between strata: gene deleted cases vs. bi-allelic cases, mutated vs. non-mutated cases, treatment-responders vs. non-responders or late vs. early?
- Is gene *X* expressed in other CLLs or B-cell lymphomas?
- Is gene *X* mutated modified in other T-PLL cases of other labs?
- Is gene *X* up- or downregulated in human disease, as well as in murine disease model (i.e. T-PLL and TCL1A-tg mice)?
- Is gene *X* expression correlated with other genes?
- Does gene *X* further interact with other proteins? Or is it further influenced by distal and trans-regulators?
- Which samples are already analysed? Which are planned? What is the platform overlap?
- Which genes exhibit dosage effects?
- Which mutations are generally affected further by copy-number alterations and result in over- or underexpression?
- Describe mutational landscape of gene *X*.
- Describe survival signature of indolent or aggressive phenotype.

6.1 Introduction: Basal functions

Before converting a data table (delimited by a special character; in the data format of *csv*, *tsv* or *XLS*), one has to remove potentially problematic characters or signs (such as umlauts in the German language or other Unicode/UTF-8 non-conform ones, spaces, tabs, question or exclamation marks). These can further disrupt formations of URIs, their representation in RDF formats such as n3/turtle or their downstream processing in HTML via *sparqllib.php*. Commata have to be further replaced by dots when dealing with dates and decimal numbers (01,01,2016 → 01.01.2016; 3,14… → 3.14...).

I integrated an R function to remove common special characters and convert data matrices (with predicate names within header) into n3 format (not XML-based, rather SPARQL-oriented and therefore more intuitive for querying). The function requires a character vector of data types (for the objects/classes) corresponding to the predicate names (therefore same length) and an index for the subject. Standard name spaces or prefixes (e.g. *PREFIX hgnc: <http://bio2rdf.org/ns/hgnc#>*) can be attached. Within the matrix itself lie the object values (either as URI or string which will be converted to *xsd* (XML Schema Definition) data types).

```
# functions needs a matrix to convert, the index of the originating node / subject, the type of the other objects and the standard namespace prefix (e.g. „gen:“).

graphCSV <- function(mat, idx, cTypes, URI_tmp) {

    if((dim(mat)[2]) != 2) {

        for(l in 1:length(mat[,idx])) {

            for(k in 1:(length(cTypes[-idx]))) {

                if(mat[,-idx][l,k] != "" && !is.na(mat[,-idx][l,k])) {

                    if(cTypes[-idx][k] == "URI") {

# print triples consisting of subjects (with index „idx“), predicate and objects (every index except „idx“) with namespace prefixes (URI_tmp) for each matrix entry

                        cat(paste(URI_tmp, mat[l,idx], " ", URI_tmp, colnames(mat)[-idx][k], " ", URI_tmp, mat[,-idx][l,k], " .\n", sep=""))

                    } else {

# print as above, only object is not an URI, but rather a xsd datatype

                        cat(paste(URI_tmp, mat[l,idx], " ", URI_tmp, colnames(mat)[-idx][k], " \"", mat[,-idx][l,k], "\"^^xsd:", cTypes[-idx][k], " .\n", sep=""))

                    }

                }

            }

        }

    }

    if((dim(mat)[2]) == 2) { # if matrix is actually just 2 subject-object vectors

        for(l in 1:length(mat[,idx])) {

            if(mat[,-idx][l] != "" && !is.na(mat[,-idx][l])) {

                if(cTypes[-idx] == "URI") {

                    cat(paste(URI_tmp, mat[,idx][l], " ", URI_tmp, colnames(mat)[-idx], " ", URI_tmp, mat[,-idx][l], " .\n", sep=""))

                } else {
```

```
                    cat(paste(URI_tmp, mat[,idx][l], " ", URI_tmp, colnames(mat)
[-idx], " \"", mat[,-idx][l], "\"^^xsd:", cTypes[-idx], " .\n", sep=""))
                }
            }
        }
    }
}
```

6.2 Sample Organization

The root subject of each analysis is a 'PATIENT_ID', which identifies a human individual (triple: *'PATIENT_ID' http://purl.uniprot.org/core/organism taxon:9606 .*) or a model organism (with *taxon:10090* for mouse, *taxon:10116* for rat, *taxon:6239* for C. elegans, *taxon:7227* for Drosophila or *taxon:7955* for Danio rerio). It then is further defined by global molecular or clinical strata, such as oncogene status (*gen:hasMTCP1status* or *gen:hasTCL1Astatus*), diagnosis or phenotype (*gen:hasDiagn*). Within SPARQL, a 'PATIENT_ID' is inserted (a), deleted (b) and modified (c) as follows:

```
# a) inserting gen:P1_1105_1387_1389_1392_ with clinical information

INSERT DATA { gen:P1_1105_1387_1389_1392_ gen:Skin_infiltration
"false"^^xsd:boolean . }

# b) deleting P1107_

DELETE gen:P1107_ ?a ?c WHERE { gen:P1107_ ?a ?c }

# c) renaming P1107_ to gen:P967_958_1107_53965_

INSERT gen:P967_958_1107_53965_ ?a ?c WHERE { gen:P1107_ ?a ?c }
```

Each 'PATIENT_ID' has to be unique and can have further samples (*gen:hasSample*), either differing by sampling date or labels such as 'replicate', 'early', 'FU' (follow-up) or 'late'. The latter three are corresponding to sequential samples especially important for temporal analyses (**Figure 6.4**). The former is important when combining duplicates into one sample (for variance stabilization). The approach to combine different 'SAMPLE_ID's through the predicate „*owl:sameAs*“ makes it hard to rank or organize them temporally and was therefore not used.

Each sample can have an analysis (*gen:hasAnalysis*) of different classical molecular biology or high-throughput methods, whose RDF attribute names here are inspired by the GEO (Barrett et al. 2013) analyses. Analyses carry results (*gen:hasResult*) in form of blank nodes (e.g. *<analysisName>_<analysisDate>_res<i>*, with integer i=0...n) as an extra step to guarantee unambiguity.

Through the semantic framework it is possible to combine not only different data sets over a 'PATIENT_ID', but also to connect clinical information with the results of high-throughput data sets (also from other laboratories, such as the EMBL/EBI RDF platform). So one can calculate median/mean measurements over all patient or correlate clinical information, as well as sample grouping by using certain thresholds (e.g. TCL1A protein expression > *X*) to diagnose malignancies (e.g. when threshold is surpassed, T-PLL is annotated).

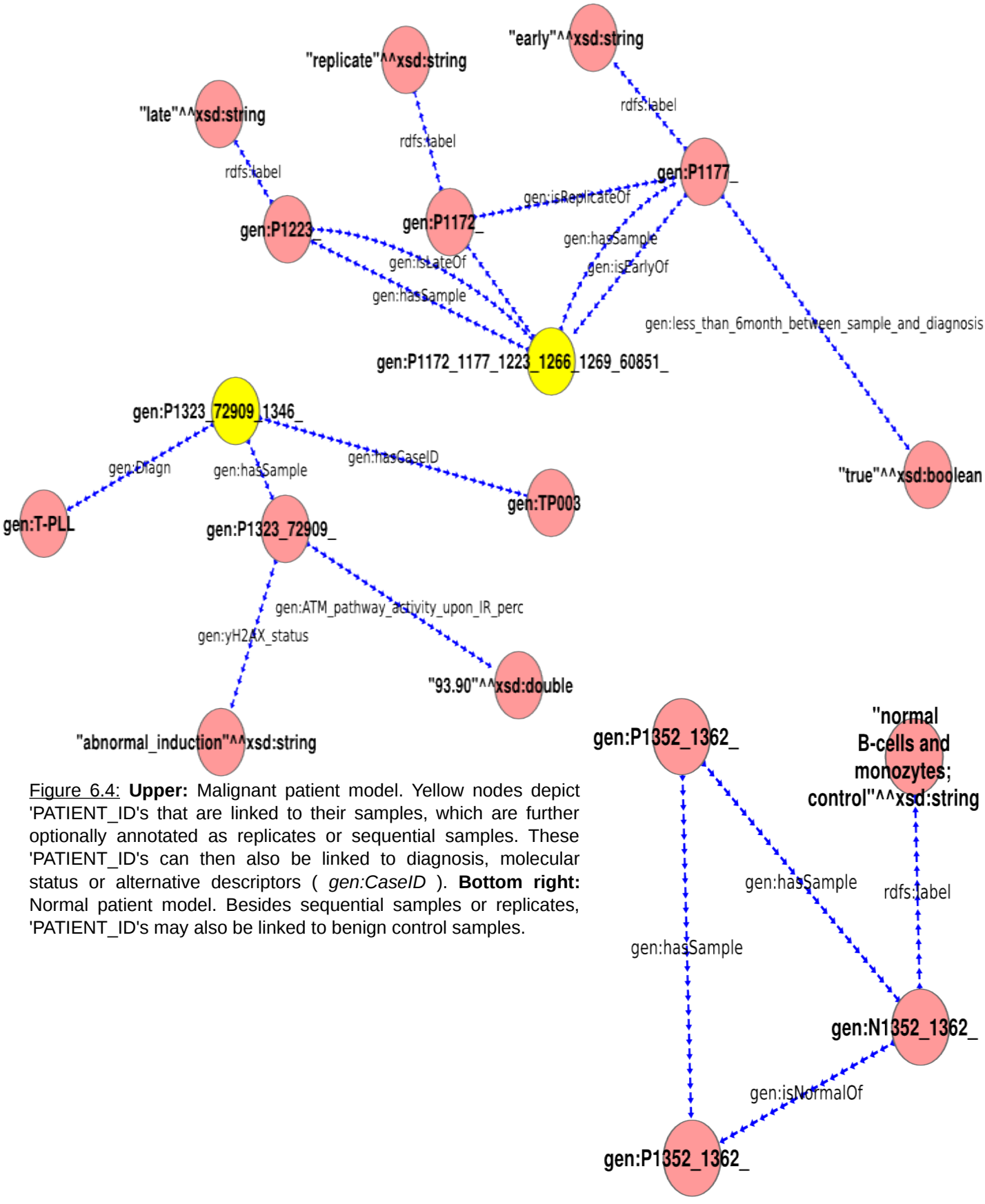


Figure 6.4: **Upper:** Malignant patient model. Yellow nodes depict 'PATIENT_ID's that are linked to their samples, which are further optionally annotated as replicates or sequential samples. These 'PATIENT_ID's can then also be linked to diagnosis, molecular status or alternative descriptors (*gen:CaseID*). **Bottom right:** Normal patient model. Besides sequential samples or replicates, 'PATIENT_ID's may also be linked to benign control samples.

6.3 Semantic model for novel exploratory survival algorithm

The algorithm presented in Crispatzu et al. 2016 relies on clinical and gene expression data, which is in our case modeled and stored in the semantic database. In order to do survival analysis and evaluate potential prognostic marker or marker sets linked to adverse patient outcome, one has to know the dates of certain events or at least their range. Dates can be in the form of 'date of birth' (*gen:Date_of_birth*) and 'date of decease' (*gen:Date_of_death*) or 'age at diagnosis' (*gen:Age_at_Diagn*) to calculate potential age bias / demographic at risk. The latter parameter is favoured in a public database because the other two can easily dereference the patient (Google search of 'date of decease' may come up with an obituary notice and thus clear name) and leading to privacy infringements. However it may be important to double-check, thus the dates should at least be kept in a separate closed database.

Further parameters include 'date of diagnosis' (*gen:Date_of_Diagn*) and status at last follow-up (F/U) to evaluate overall survival (OS; *gen:Overall_Survival_status_as_last_FU*) and progression-free survival (*gen:PFS_date_of_first_therapy_until_date_of_relapse_or_death_or_date_of_last_FU*). In order to censor patients and/or to account only for disease-specific events, one can further use *gen:Disease_related_death*. Dates are stored in *xsd* datatypes (i.e. *xsd:date*) and may have to be converted.

To guarantee the similarity of clinical features and sampled molecular features, one can further restrict the analysis to samples taken at most six month after diagnosis (*gen:less_than_6month_between_sample_and_diagnosis*). This circumvents for example noise introduced by kinetics or disease progression such as LDT or high WBC.

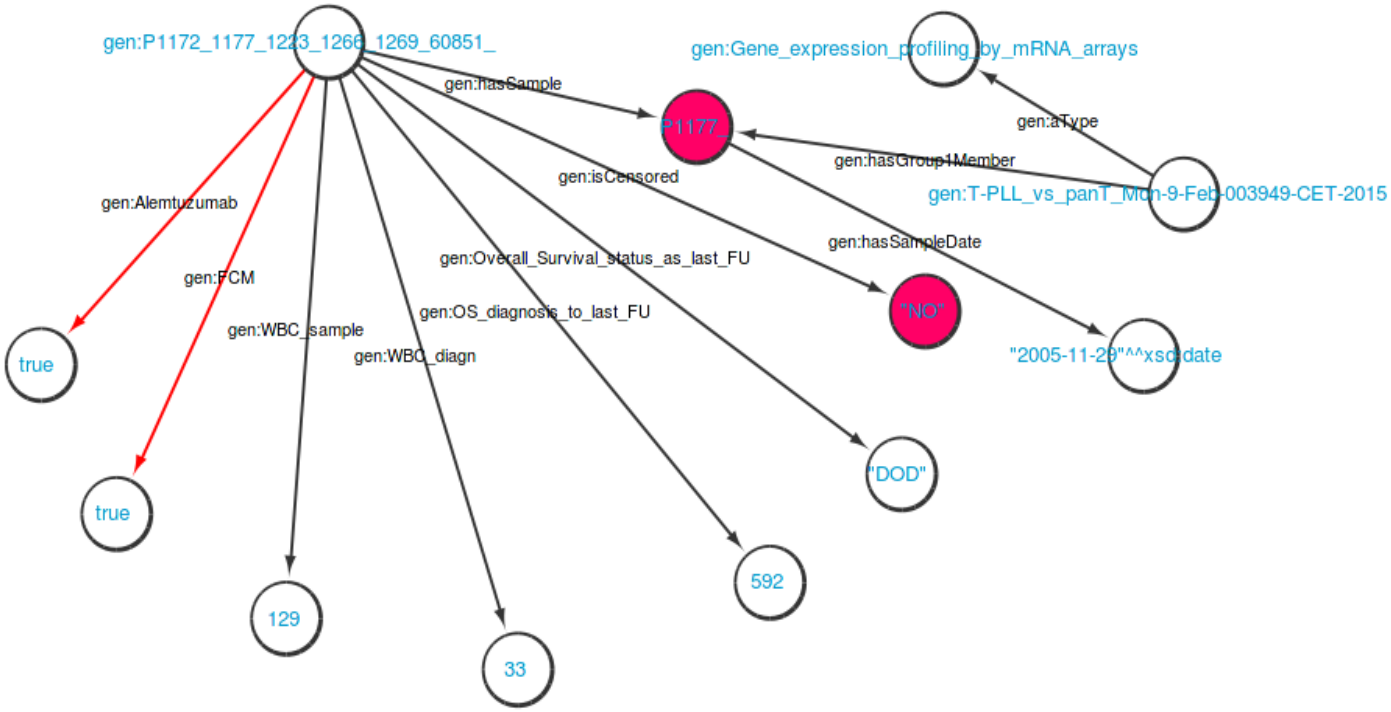


Figure 6.5: Edges in red highlight treatment regimens. While nodes in dark red show the link that the patient is not censored, i.e. died through disease-specific causes and can thus be included into our survival analysis.

The model needed to evaluate survival analysis is shown in **Figure 6.5**. We only considered the 6 month-restricted samples with disease-specific death with the 5 highest and 5 lowest overall survival ('OS hi vs. OS lo') for our T-PLL samples. Sample grouping is done by matching survival annotations with gene expression array names within R via SPARQL library. I further implemented an automated method to facilitate differential expression analysis and getting the 100 (or any other positive integer) most significant deregulated genes sorted by fold change. This gives us a first hint which genes are linked to extreme indolence or aggressiveness instead of calculating Cox statistics and Kaplan-Meier curves for every gene variation and outcome in our T-PLL data sets. The necessary query is as follows:

```
SELECT DISTINCT ?pat ?sample ?os ?mo6 WHERE { ?pat gen:Diagn gen:T-PLL . ?pat
gen:OS_diagnosis_to_last_FU ?os . ?pat gen:hasSample ?sample .

    OPTIONAL { ?sample gen:less_than_6month_between_sample_and_diagnosis ?
mo6 } .

    ?pat gen:Overall_Survival_status_as_last_FU "DOD"^^xsd:string . ?pat
gen:Disease_related_death "true"^^xsd:boolean . }
```

6.4 Gene expression meta-analysis using EMBL / EBI RDF: AtlasRDF

The 'Gene expression atlas' of the EMBL / EBI (http://www.ebi.ac.uk/gxa) “provides information about gene and protein expression in animal and plant samples of different cell types, organism parts, developmental stages, diseases and other conditions” of “1572 studies as of August 2015” (taken from Petryszak et al. 2016). The human data sets are currently exported into an RDF version accessible via SPARQL endpoint (http://www.ebi.ac.uk/rdf/services/atlas/sparql; AtlasRDF accessed on 09/29/2016).

In order to process the SPARQL query results in R, I attached a helpful wrapper function, which parses prefixes, replaces empty entries and deletes namespace prefixes in results:

```
queryProc <- function(queryString, prefixes, endpoint, showPrefixes=T) {
    tmpF <- gsub(" $", "", gsub("^ ","",strsplit(prefixes,"PREFIX")[[1]]))
    tmpF <- gsub(">", "", tmpF[which(tmpF!="")])
    sS <- strsplit(tmpF, " ")
    qStat <- SPARQL(url=endpoint, query=paste(prefixes,queryString))$results
    for(i in 1:length(sS)) {
        for(j in 1:length(qStat[,1])) {
            if(showPrefixes) {
              qStat[j,] <- gsub(">", "", gsub(sS[[i]][2], sS[[i]][1],
qStat[j,]))
            } else {
                qStat[j,] <- gsub(">", "", gsub(sS[[i]][2], "", qStat[j,]))
            }
            qStat[j,] <- gsub("^NA$",NA, qStat[j,])
        }
    }
```

```
    return(qStat)
}
```

To construct another query like Example #2 on AtlasRDF to be semi-automately directed within R SPARQL, one must only be a bit familiar with the underlying ontologies:

```
# read in a txt file with gene names and store it in a vector (here "tel)
# load helper functions
source("/home/gc/workspace/AG_Herling/Semantic_Framework/functions/basic.R")
library("biomaRt")
human <- useMart("ENSEMBL_MART_ENSEMBL",
    dataset="hsapiens_gene_ensembl",
    host="feb2014.archive.ensembl.org",
    path="/biomart/martservice", archive=FALSE)
# convert gene names (e.g. "ATM") to ENSEMBL IDs
bm <- getBM(attributes = c("chromosome_name", "ensembl_gene_id",
"wikigene_name"), filters = "wikigene_name", values = tel, mart = human)
bm <- bm[grep("ENS", bm[,2]),]
# only get IDs from top level assembly
bm <- bm[grep("^[0-9][0-9]*$|^X$|^Y$", bm[,1]),]
qS_all <- data.frame()
# for each ENSEMBL ID, get dysregulations in CLL and other chronic malignancies
for(j in 1:length(bm[,2])) {
    qS <- paste("SELECT distinct ?expUri ?valueLabel ?pvalue \
    WHERE { ?expUri atlasterms:hasAnalysis ?analysis . \
    ?analysis atlasterms:hasExpressionValue ?value . \
    ?value atlasterms:pValue ?pvalue . \
    ?value atlasterms:isMeasurementOf ?probe . \
    ?value rdfs:label ?valueLabel . \
    ?value atlasterms:isMeasurementOf ?probe  . \
    ?probe atlasterms:dbXref identifiers:", bm[j,2] ,"  . \
    FILTER(regex(?valueLabel , \"CLL\") || regex(?valueLabel, \"hronic\")) }",
sep="") # filtered with regular expressions for experiment labels which contain
keywords
    try(qS_res_orig <- queryProc(qS, prefixes,
"http://www.ebi.ac.uk/rdf/services/atlas/sparql", F), silent=T)
    qS_all <- rbind(qS_all, qS_res_orig)
}
write.csv(qS_all, "~/qS_all.csv")
```

6.5 Further mRNA array-based gene expression meta-analyses based on gene sets

Since not all 'Gene expression atlas' data sets are thus far integrated into the EMBL / EBI RDF platform and the GEO database (Barrett et al. 2013) has further data sets, one is obliged to manually integrate these extra ones.

For a more detailed description we further did differential expression analysis for only those data sets with a normal pool (of varying quality, number and specificity) of samples. Fold changes with p-value < 0.1 (trend) or p-value < 0.05 were extracted to compare normal-matched gene regulation per experiment and probe target. Instead of patterns of expression level, one can now observe different disease vs. "normal" comparisons and which genes are exclusively down- or up-regulated and which show no clear pattern or are specific to small subgroups.

The preprocessing of gene sets and sample grouping was further automated by means of the Semantic Web (**Figure 6.6**). Genes are stored in a new RDF data structure similar to a list. The subject identified by the predicate *'gen:is_a_geneset"* has many members with the predicate *'gen:hasGeneSetMember"*, who are then further annotated by HUGO/HGNC gene symbols and Ensembl identifiers (*hgnc:Symbol* and *gen:hasHumanEnsembl* respectively).

Each integrated GEO or ArrayExpress data set is modeled with array platform/manufacturer (*gen:platform*), reference (*gen:Source_of_sample*), a background-corrected, quantile-normalized and annotated *csv* file (*gen:hasFileName*), as well as one or multiple analyses (*atlasterms:hasAnalysis*) differing by sample grouping done by custom-written regular expressions (*gen:hasGroup1RegExpr*, *gen:hasGroup2RegExpr*) (**Figure 6.7**). They can be queried as follows:

```
SELECT DISTINCT ?a ?c ?d ?e ?g1 ?g2 ?lab ?source WHERE { ?a
atlasterms:hasAnalysis ?c . ?a gen:Source_of_sample ?source . ?a gen:hasFileName
?d . ?a gen:platform ?e. ?c rdfs:label ?lab . ?c gen:hasGroup1RegExpr ?g1 . ?c
gen:hasGroup2RegExpr ?g2 . }
```

The analyses names (or labels by *rdfs:label*) further describe to which meta-analysis they belong to (CLL, NBC (normal B-cells), BCL (B-cell lymphoma) or MTCL (mature T-cell lymphoma/leukemia) to name a few). In the meta-analysis itself every differential expression analysis is further annotated by the statistical test used. The default is Student's t-test, except for low variation comparisons whereas Wilcoxon rank sum test is forced.

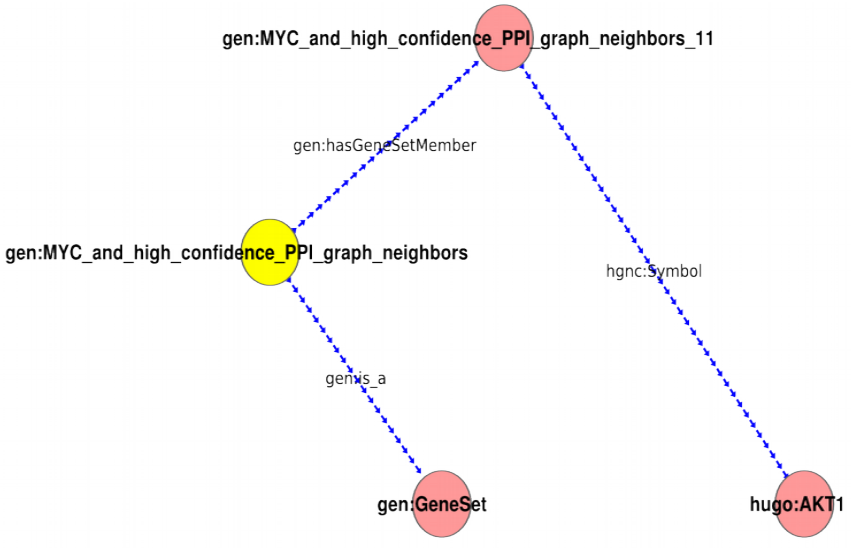


Figure 6.6: GeneSet model linking *AKT1* (et al.) to *MYC's* PPI network neighbors. Name (central node) of the gene set is colored in yellow.

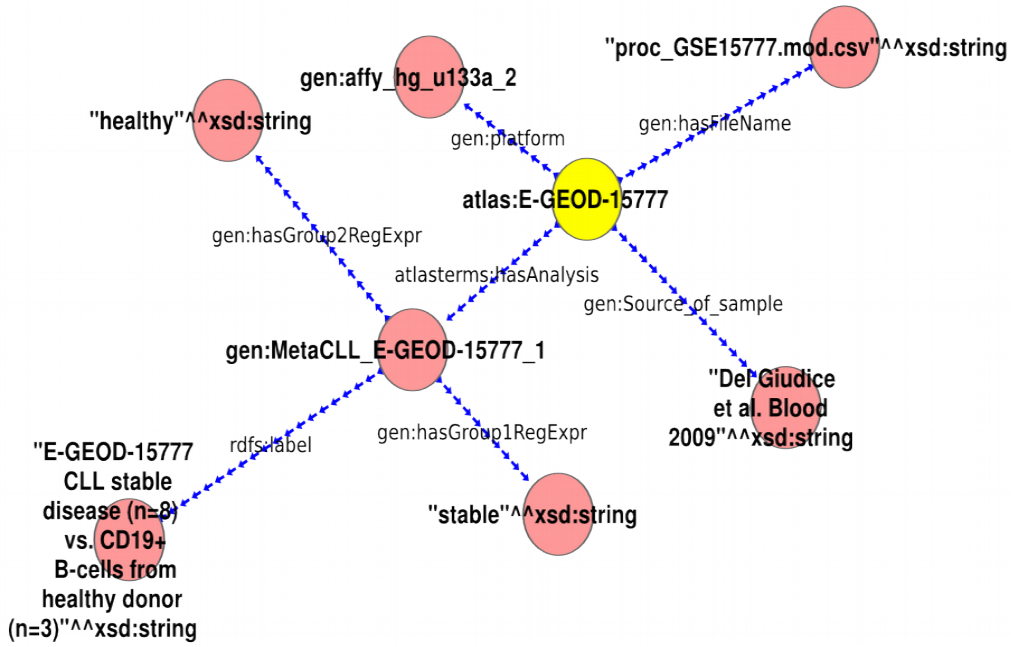


Figure 6.7: Meta model. Central node, naming the data set, is colored in yellow and linked to technical details and analysis. Analysis itself is linked to meta-data of differential expression analysis.

Publicly available data sets investigating normal T-cell subsets are similar modeled. However instead of grouping samples and comparing them pairwise in differential expression analysis and hierarchical clustering, each sample group (e.g. CD4+ or CD8+ T-cells) is visualized in a PCA on the basis of our memory/naive/CM signatures (see Warner, Oberbeck, Schrader et al.). These signatures can be stored as a gene set list or calculated on-the-fly. I further implemented an automated function to run PCAs on a given cohort and given gene set (data not shown).

6.6 Comparative methodology illustrated on copy-number data

Besides traditional platforms like SNP arrays, it is also possible to call copy-number variations by WGS or WES (Nam et al. 2015). The segmentation and (LOH) calling algorithms differ quite a bit due to the unequal variance and noise in read coverage. It is therefore interesting to observe whether and to which extent called copy-number variations overlap. Similar to GEP, every sample can also have a copy-number variation analysis. Either by WES (*Y gen:sample1 X . Y gen:aType gen:Copy-Number_Variation_by_Whole_exome_sequencing .;* **Figure 6.8a**) or SNP arrays (*X gen:hasAnalysis Y . Y gen:aType gen:Copy-Number_Variation_by_SNP_arrays .;* **Figure 6.8b**). Control samples can either be paired (*Y gen:sample2 ?s2*) or pooled (*FILTER(regex(xsd:string(?s2),"pool"))*). Every result of the analysis has (at least) the predicates copy-number (*gen:CopyNumber)* and a HGNC/HUGO gene symbol (*hgnc:Symbol*).

Most segmentation algorithm run on genomic ranges not on coding ranges, therefore a gene can be split in two with different copy-numbers assigned to it. I wrote a function that averages copy-number per gene after SPARQL querying:

```
# functions expects the SPARQL query result table, index for PATIENT_ID to link
average and an index for values to be averaged

mDupl <- function(mat, patIdx, valIdx) {

    dupl <- unique(mat[which(duplicated(mat[, patIdx])), patIdx])

        if(length(dupl) > 0) {

        for(i in 1:length(dupl)) {

            dM <- mean(as.numeric(mat[which( mat[, patIdx]== dupl[i]), valIdx]),
na.rm=T)

            mat[which(mat[, patIdx]== dupl[i])[1], valIdx] <- dM

            lenD <- length(which(mat[, patIdx]== dupl[i]))

            mat <- mat[-which(mat[, patIdx]== dupl[i])[2:lenD],]

        }

    }

    replace_idx <- which(mat[,valIdx] == "NaN")

    if(length(replace_idx) > 0) { mat[replace_idx, valIdx] <- NA }

    return(mat)

}
```

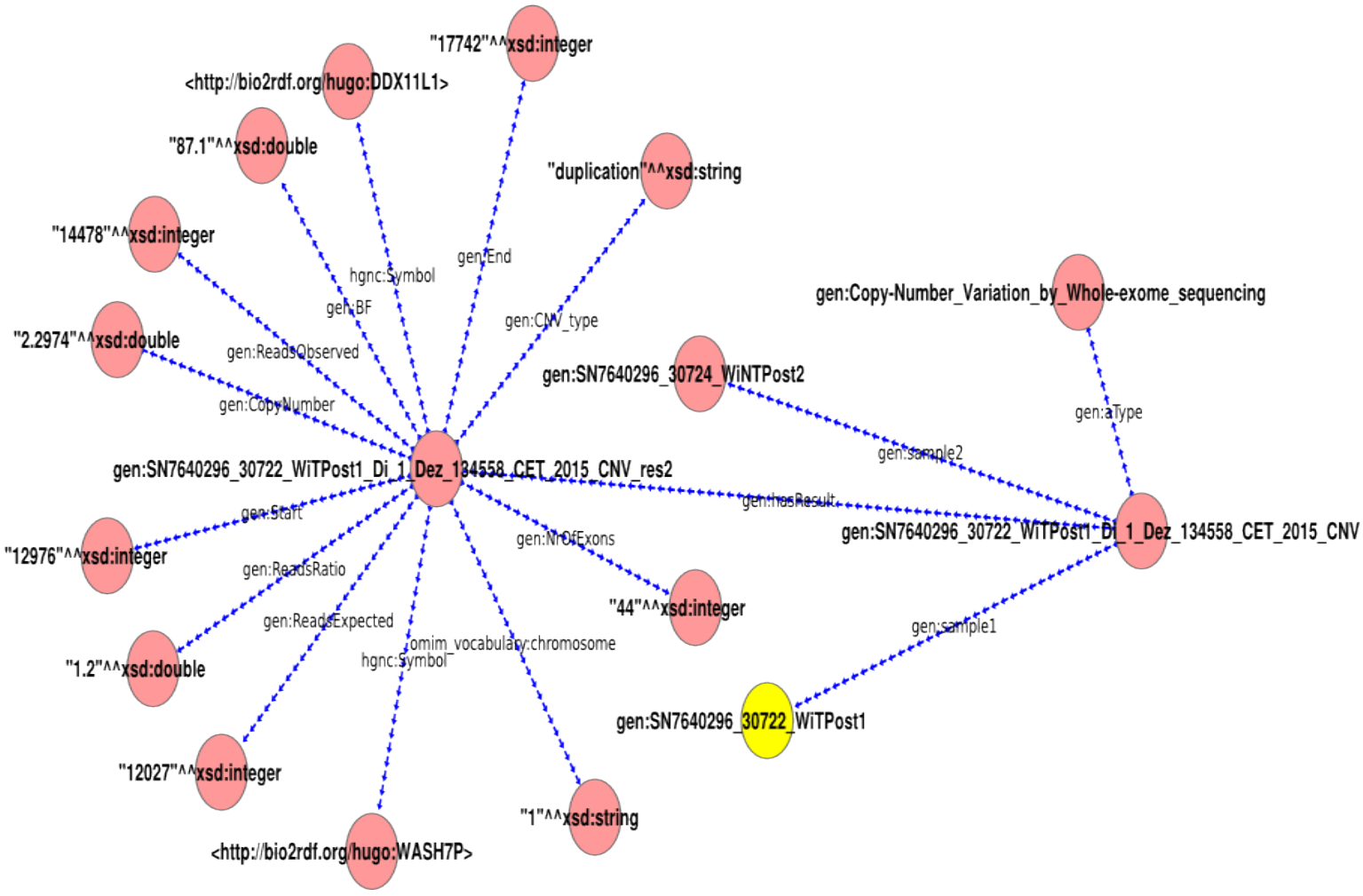
The code to integrate pooled SNP 6.0, SNP 6.0 compared to HapMap and WES copy-numbers is as follows:

```
# loading all basal functions, including „rmDupl()“

source("/home/gc/workspace/AG_Herling/Semantic_Framework/functions/basic.R")

gS <- "SELECT DISTINCT ?gene WHERE {  ?a2 gen:aType gen:Copy-
Number_Variation_by_SNP_arrays . ?a2 gen:hasResult ?r2 . ?r2 gen:CopyNumber ?cn2
. ?r2 hgnc:Symbol ?gene}"

genes <- queryProc(gS, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/CNA/query", F)

ovMat <- matrix(0, ncol=12, nrow=length(genes))

ovMat[,c(1,4,7,10)] <- 2 # default: bi-allelic

for(i in 1:length(genes)) {

    d1 <- paste("SELECT DISTINCT ?pat ?cn ?gene ?orig WHERE {

    SERVICE <http://localhost:8080/openrdf-workbench/repositories/GEX/query> { ?
orig gen:Diagn gen:T-PLL . ?orig gen:hasSample ?pat } .

    ?a1 gen:aType gen:Copy-Number_Variation_by_Whole_exome_sequencing . ?a1
gen:sample1 ?pat . ?a1 gen:sample2 ?s2 . FILTER(regex(xsd:string(?s2),\"pool\"))
. OPTIONAL { ?a1 gen:hasResult ?r1 .?r1 gen:CopyNumber ?cn . ?r1 hgnc:Symbol
<http://bio2rdf.org/hugo:", genes[i],"> } } ORDER by ?orig", sep="")

    r1 <- queryProc(d1, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/ngs/query", F)

    r1 <- rmDupl(r1,4,2)
```

```
    r1[which(is.na(r1[,2])),2] <- 2

    m1 <- mean(as.numeric(r1[,2]))

    g1 <- length(which(as.numeric(r1[,2]) > 2.2)) / length(r1[,2]) * 100

    l1 <- length(which(as.numeric(r1[,2]) < 1.8)) / length(r1[,2]) * 100


    d2 <- paste("SELECT DISTINCT ?pat ?cn2 ?gene ?orig WHERE {

    SERVICE <http://localhost:8080/openrdf-workbench/repositories/GEX/query> { ?
orig gen:Diagn gen:T-PLL . ?orig gen:hasSample ?pat } .

    ?pat gen:hasAnalysis ?a2 . ?a2 rdfs:label ?lab . FILTER(regex(xsd:string(?
lab),\"pool\")) . ?a2 gen:aType gen:Copy-Number_Variation_by_SNP_arrays .
OPTIONAL { ?a2 gen:hasResult ?r2 . ?r2 gen:CopyNumber ?cn2 . ?r2 hgnc:Symbol
<http://bio2rdf.org/hugo:", genes[i],"> } }", sep="")

    r2 <- queryProc(d2, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/CNA/query", F)

    r2 <- rmDupl(r2,4,2)

    r2[which(is.na(r2[,2])),2] <- 2

    m2 <- mean(as.numeric(r2[,2]))

    g2 <- length(which(as.numeric(r2[,2]) > 2.2)) / length(r2[,2]) * 100

    l2 <- length(which(as.numeric(r2[,2]) < 1.8)) / length(r2[,2]) * 100


    d3 <- paste("SELECT DISTINCT ?pat ?cn2 ?gene ?orig WHERE {

    SERVICE <http://localhost:8080/openrdf-workbench/repositories/GEX/query> { ?
orig gen:Diagn gen:T-PLL . ?orig gen:hasSample ?pat } .

    ?pat gen:hasAnalysis ?a2 . ?a2 rdfs:label ?lab . FILTER(regex(xsd:string(?
lab),\"hapmap\")) . ?a2 gen:aType gen:Copy-Number_Variation_by_SNP_arrays .
OPTIONAL { ?a2 gen:hasResult ?r2 . ?r2 gen:CopyNumber ?cn2 . ?r2 hgnc:Symbol
<http://bio2rdf.org/hugo:", genes[i],"> } }", sep="")

    r3 <- queryProc(d3, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/CNA/query", F)

    r3 <- rmDupl(r3,4,2)

    r3[which(is.na(r3[,2])),2] <- 2

    m3 <- mean(as.numeric(r3[,2]))

    g3 <- length(which(as.numeric(r3[,2]) > 2.2)) / length(r3[,2]) * 100

    l3 <- length(which(as.numeric(r3[,2]) < 1.8)) / length(r3[,2]) * 100


    r4 <- t(cbind(cbind(t(r1),t(r2)), t(r3)))

    r4[which(is.na(r4[,2])),2] <- 2

    m4 <- mean(as.numeric(r4[,2]))

    g4 <- length(which(as.numeric(r4[,2]) > 2.2)) / length(r4[,2]) * 100
```

```
    l4 <- length(which(as.numeric(r4[,2]) < 1.8)) / length(r4[,2]) * 100

    ovMat[i,] <- c(m1,l1,g1, m2,l2,g2, m3,l3,g3, m4,l4,g4)
    print(i)
}
colnames(ovMat) <- c("WES Mean", "WES Loss Freq.", "WES Gain Freq.",
            "Ctrl Mean", "Ctrl Loss Freq.", "Ctrl Gain Freq.",
            "Hapmap Mean", "Hapmap Loss Freq.", "Hapmap Gain Freq.",
            "Combi Mean", "Combi Loss Freq.", "Combi Gain Freq.")

write.table(ovMat, "~/ovMat.csv", row.names=F)
### ### ###
```

a)



b)

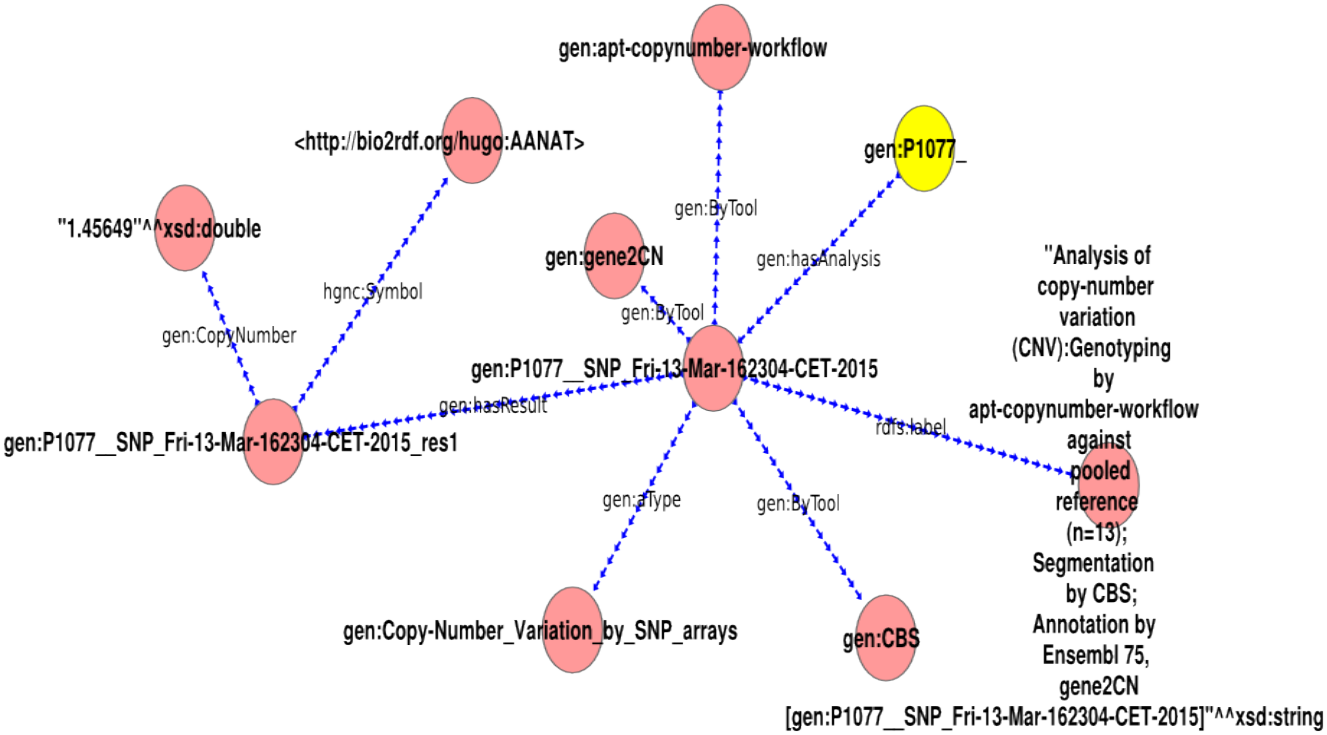


Figure 6.8: **a)** WES CNV model in paired mode. Pooled setting can be described by multiple *gen:sample2* attributes. Except for copy-number and gene name, every other attribute is optional. Result can have multiple affected genes who have to be averaged later on. **b)** SNP 6.0 model with 'PATIENT_ID' depicted in yellow which is further linked to analysis and results with only one affected gene and copy-number. Label on analysis describes protocol and „pooled“ comparison.

ATM deletions are confirmed by FISH, classical cytogenetics and previous reports in Affymetrix SNP 6.0 arrays (e.g. Dürig et al. 2007). We therefore use our FISH and classical cytogenetics annotations of each 'PATIENT_ID' to validate our SNP 6.0 calls, and SNP 6.0 calls to validate CNV calls in WES.

6.7 Dosage effect

For each patient every genes copy-number (*?pat gen:hasSample ?sample . ?sample gen:hasResult ?res . ?res gen:CopyNumber ?cn . ?res hgnc:Symbol ?gene*) was fetched and fold changes of each gene for each patient compared to the CD3+ normal T-cell pool (n=10) were calculated. I only considered those genes who are upregulated and have a gain, those who are downregulated and have a loss (both intuitive) and those who are upregulated and have a loss or downregulated and have a gain (counter-intuitive). Providing us with four categories (2x2 count matrix) or even more when including non-spotted or stable genes. The counter-intuitive cases may be hints for allele-specific expression and/or gene dosages, meaning the allele which is affected by a copy-number event is not favoured by the transcription machinery (see Schrader, Crispatzu et al. **Figure S5b,c**).

In Schrader, Crispatzu et al., we also observed that not *MYC* (Dürig et al. 2007) but actually *AGO2* is the most frequently amplified gene in T-PLL. *MYC* however is still upregulated in many cases not exhibiting a respective amplification or is stable in cases exhibiting a respective amplification. Similar to the investigation of dosage effects, I queried the *AGO2* and *MYC* copy-number for each patient (*?pat gen:hasSample ?sample*

. ?sample gen:hasResult ?res . ?res gen:CopyNumber ?cn . ?res hgnc:Symbol hugo:MYC) and again calculated fold changes in the respective patients. Expression levels of bi-allelic and amplification-carrying patients were then compared and visualized in boxplots. Only *AGO2* (p=0.000503, fc=1.63), not *MYC* (p=0.821, fc=0.0246), seems to respond to sCNAs in T-PLL (Schrader, Crispatzu et al. **Figure S6a,b**). The code to obtain *MYC* boxplots and statistical measurements of its dosage effect is as follows:

```
matMyc <- matrix("", nrow=length(tmpA), ncol=5)

# tmpA are all T-PLL patients

for(i in 1:length(tmpA)) {

    qS <- paste("SELECT DISTINCT ?orig ?snp WHERE { ?orig gen:hasSample gen:",
tmpA[i]," . OPTIONAL { gen:", tmpA[i]," gen:hasAnalysis ?a1 . ?a1 rdfs:label ?
lab . FILTER(regex(xsd:string(?lab),\"pool\")) . ?a1 gen:aType gen:Copy-
Number_Variation_by_SNP_arrays . ?a1 gen:hasResult ?r1 . ?r1 gen:CopyNumber ?snp
. ?r1 hgnc:Symbol <http://bio2rdf.org/hugo:MYC> } . }", sep="")

    matMyc[i,1:2] <- unlist(queryProc(qS, prefixes,
"http://localhost:8080/openrdf-workbench/repositories/NGS/query", F))



    qS <- paste("SELECT DISTINCT ?orig ?snp2 WHERE { ?orig gen:hasSample gen:",
tmpA[i]," . OPTIONAL { gen:", tmpA[i]," gen:hasAnalysis ?a2 . ?a2 rdfs:label ?
lab2 . FILTER(regex(xsd:string(?lab2),\"hapmap\")) . ?a2 gen:aType gen:Copy-
Number_Variation_by_SNP_arrays . ?a2 gen:hasResult ?r2 . ?r2 gen:CopyNumber ?
snp2 . ?r2 hgnc:Symbol <http://bio2rdf.org/hugo:MYC> } . }", sep="")

    matMyc[i,3] <- unlist(queryProc(qS, prefixes,
"http://localhost:8080/openrdf-workbench/repositories/NGS/query", F))[2]



    qS <- paste("SELECT DISTINCT ?orig ?wes WHERE { ?orig gen:hasSample gen:",
tmpA[i]," . OPTIONAL { ?a1 gen:aType gen:Copy-
Number_Variation_by_Whole_exome_sequencing . ?a1 gen:sample1 gen:",
tmpA[i]," . ?a1 gen:sample2 ?s2 . FILTER(regex(xsd:string(?s2),\"pool\")) . ?a1
gen:hasResult ?r1 . ?r1 gen:Exon ?ex . ?r1 gen:CopyNumber ?wes . ?ex hgnc:Symbol
hugo:MYC } . }", sep="")

    matMyc[i,4] <- unlist(queryProc(qS, prefixes,
"http://localhost:8080/openrdf-workbench/repositories/NGS/query", F))[2]



    qS <- paste("SELECT DISTINCT ?orig ?fish WHERE { ?orig gen:hasSample gen:",
tmpA[i]," . OPTIONAL { ?orig gen:MYC_amplification_by_cytogenetics_FISH ?
fish } . }", sep="")

    matMyc[i,5] <- unlist(queryProc(qS, prefixes,
"http://localhost:8080/openrdf-workbench/repositories/NGS/query", F))[2]

}

mycGain <- union(union(union(which(as.numeric(matMyc[,2]) > 2.2),
which(as.numeric(matMyc[,3]) > 2.2)), which(as.numeric(matMyc[,4]) > 2.2)),
which(matMyc[,5] == T) )

mycBi   <- intersect(intersect(intersect(which(is.na(matMyc[,2])),
which(is.na(matMyc[,3]))), which(is.na(matMyc[,3]))), which(matMyc[,5] == F) )
```

```
spray <- as.data.frame(cbind( c( colMeans(combat_edata[myc_idx,
TPLL[mycGain]]) , colMeans(combat_edata[myc_idx, TPLL[mycBi]]),
colMeans(combat_edata[myc_idx, panT]) ), c(rep("ampl(MYC)", length(mycGain)),
rep("MYC~", length(mycBi)), rep("panT", length(panT))) ) )

colnames(spray) <- c("expr","entity")

spray <- as.data.frame(spray)

spray[,1] <- as.numeric(as.character(spray[,1])) #!!!

res <- getDiffExprVal(TPLL[mycGain], TPLL[mycBi], combat_edata)
# [1] "37 vs. 14"

p_myc <- res[[2]]

q_myc <- res[[3]]

fc_myc <- res[[1]]

pdf(file="~/MYC_boxplot_responseToGain_noNA.pdf")

boxplot(expr ~ factor(entity), data=spray, main=paste("Dotplot for MYC
(n=",length(geneIdx),"); p-val=", signif(mean(p_myc[geneIdx]),digits=3),

    ", fc=", signif(mean(fc_myc[geneIdx]),digits=3), sep="") ,
ylab="log2(expr)", las=2)

    stripchart(expr ~ factor(entity), data=spray, vertical = TRUE, method =
"jitter", pch = 21, col = c("maroon"), bg = c("bisque"), add = TRUE)

dev.off()
```

6.8 Dysregulation overlap of human disease to disease model sample

To evaluate how certain models mimic a disease, one can compare the gene expression profilings of the affected animals with patient data (as done in Warner, Oberbeck, Schrader et al. **Figure 4g**). Here, we only overlaped dysregulated genes between late (exponential phase) and early (preleukemic phase) TCL1A-tg mice (**Figure 6.9**) and T-PLL for runtime reasons:

```
# get sign. differentially expressed genes in mice

s1 <- "SELECT DISTINCT ?gene ?p ?fc ?p2 ?fc2 WHERE { \

    ?orig gen:Diagn gen:T-PLL . ?orig gen:hasSample ?sample2 . ?sample2
gen:isEarlyOf ?orig . \

    ?orig gen:hasSample ?sample1 . OPTIONAL { ?sample1 gen:isLateOf ?orig } .
OPTIONAL { ?sample1 gen:isFuOf ?orig } . \

    ?a1 gen:hasGroup1Member ?sample1 . ?a1 gen:hasGroup2Member ?sample2 .  \

    ?a1 gen:aType gen:Gene_expression_profiling_by_mRNA_arrays . ?a1
gen:hasResult ?r1 . \

    ?r1 gen:hasILMN_ID ?ilmn . ?ilmn hgnc:Symbol ?gene . \

    ?r1 gen:fold_change ?fc . ?r1 atlasterms:pValue ?p . ?r1 gen:q-value ?q .
FILTER(abs(?fc) > 2 && ?p < 0.05) }"

t1 <- queryProc(s1, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/GEX/query", F)
```

```
# get sign. differentially expressed genes in humans

s2 <- "SELECT DISTINCT ?gene ?p ?fc ?p2 ?fc2 WHERE { \

    ?g1 gen:Diagn gen:Lckpr_hTCL1A_pn_transgenic_mouse . ?g2 gen:Diagn
gen:Lckpr_hTCL1A_pn_transgenic_mouse . \

    ?a2 gen:hasGroup1Member ?g1 . ?a2 gen:hasGroup2Member ?g2 . ?a2
gen:hasResult ?r2 . \

    ?r2 gen:hasAffymetrixID ?affy . ?affy hgnc:Symbol ?gene . \

    ?g1 rdfs:label \"exponential phase\"^^xsd:string . ?g2 rdfs:label \"chronic
phase\"^^xsd:string . \

    ?r2 gen:fold_change ?fc2 . ?r2 atlasterms:pValue ?p2 . ?r2 gen:q-value ?q2 .
FILTER(abs(?fc2) > 2 && ?p2 < 0.05) }"

t2 <- queryProc(s2, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/GEX/query", F)

intersect( t1[,1], t2[,1]) # overlap of both
```

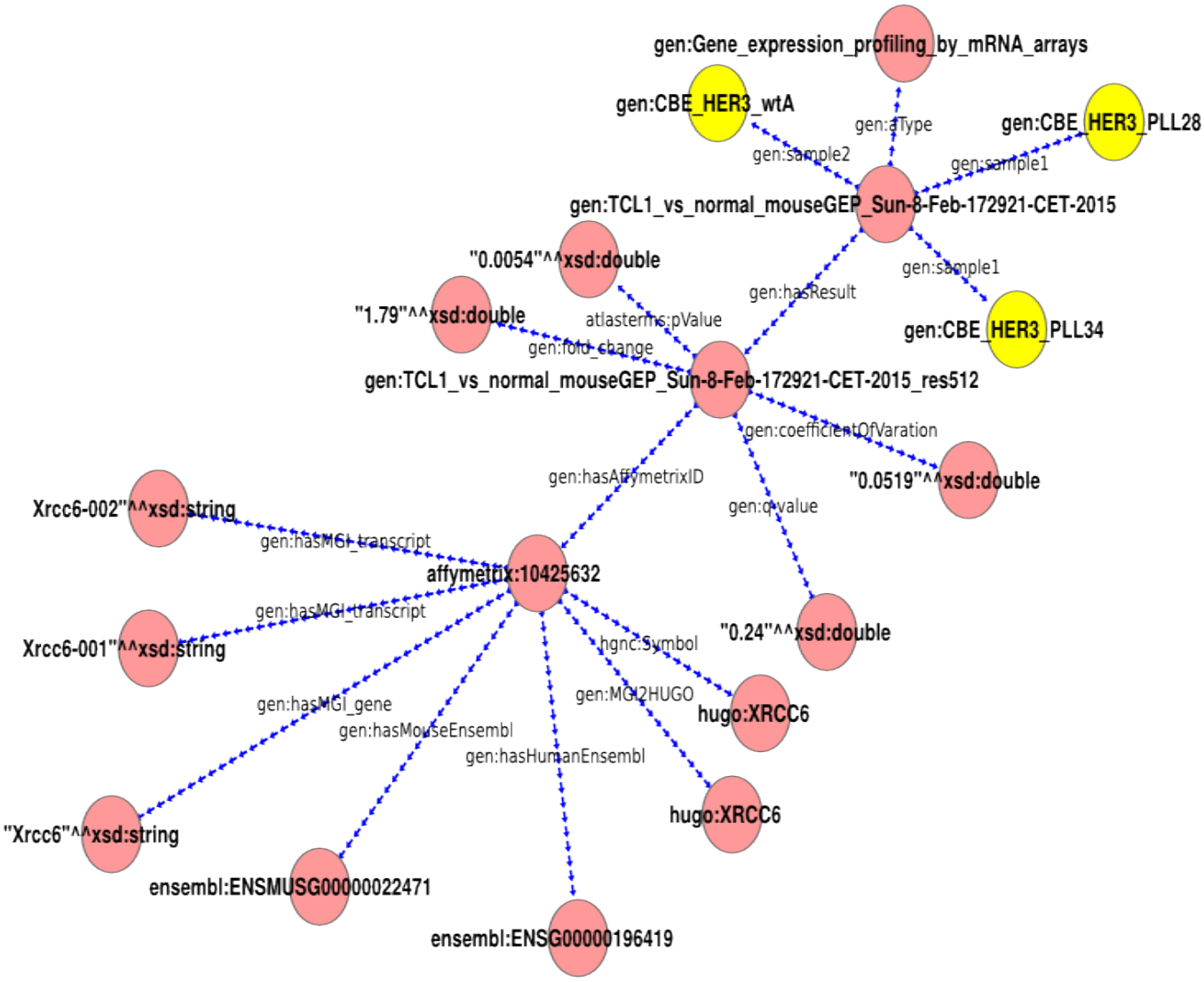


Figure 6.9: Mouse GEP model with malignant and benign samples depicted in yellow. Results consist of different statistical measurements of differential expression and (in lower branches of graphic) annotation of affymetrix probes in order to link results to human orthologues.

6.9 Regulatory gene network analysis

By adding further predicates one can compare co-expression (*?analysis gen:aType gen:Co-expression_analysis_by_mRNA_arrays*) results similar to GEP results with each other. They can further be annotated when assigning pathway links to each gene and test for over-representation. Each co-expression result has two HUGO/HGNC gene symbols *(?res gen:Gene1 ?gene . ?res gen:Gene2 ?gene*), a p-value (*atlasterms:pValue*) and a correlation coefficient (*gen:rho*). One can further filter correlated genes by dysregulation of one or both genes, as we did previously within late vs. early T-PLL and mice models. We therefore looked for co-expressed genes (each gene a query) within our human late vs. early gene set. We then constructed a bidirectional graph (*gene1* correlates with *gene2*, therefore *gene1* also correlates with *gene2*) and passed it on to *Cytoscape 2.8.1* with *RCytoscape:*

```
geneN <- unique(t1[,1])

matR <- matrix("", ncol=4)

for(i in 1:length(geneN)) {

    s3 <- paste("SELECT DISTINCT ?gene1 ?gene2 ?p ?rho WHERE { ?r1 gen:Gene1 ?
gene1 . ?r1 gen:Gene2 ?gene2 . FILTER(regex(xsd:string(?gene1), \"",
geneN[i],"\") || regex(xsd:string(?gene2), \"", geneN[i],"\")) . ?r1 gen:rho ?
rho . ?r1 atlasterms:pValue ?p . FILTER(abs(?rho) > 0.8 && ?p < 0.05) }",
sep="")

    t3 <- tryCatch(as.matrix(queryProc(s3, prefixes,
"http://localhost:8080/openrdf-workbench/repositories/GEX/query", F)),
error=function(e) { t3 <- c("", "", "", "") } )

    matR <- rbind(matR, t3)

}

matR <- matR[-which(matR[,1]==""),]

uMat <- tab <- unique(matR[,c(1,2)])

tt <- unique(unlist(sapply(t[,1], function(a) which(a == uMat[,1]))))

uMat <- uMat[tt,]

tt <- unique(unlist(sapply(t[,2], function(a) which(a == uMat[,1]))))

uMat <- uMat[tt,]

both <- unique(c(t1[,1], t2[,2]))

both <- both[which(!is.na(both))]

rEG <- new("graphNEL", nodes=both, edgemode="directed")


for(k in 1:(dim(tab)[1])) {

    if((length(which(both == tab[k,1])) > 0) && (length(which(both == tab[k,2]))
> 0)) { #!!!

        rEG <- addEdge(tab[k,1], tab[k,2], rEG, 1)

        rEG <- addEdge(tab[k,2], tab[k,1], rEG, 1)

    }
```

```
}

rEGi <- igraph.from.graphNEL(rEG)

write.graph(rEGi, "~/test.gml", format="graphml")

###

library("Rgraphviz")

library("RCytoscape")

rEG <- initEdgeAttribute (rEG, "weight", "numeric", 1.0)

#in Cytoscape 2.8.1: Plugins > CytoscapeRPC > Activate CytoscapeRPC > OK
cw <- new.CytoscapeWindow ('broad', graph=rEG)

displayGraph(cw)

### ### ###
```

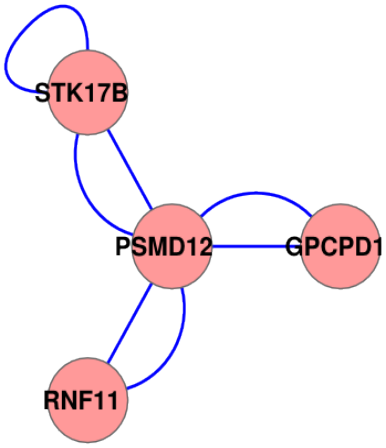
Within *Cytoscape 2.8.1*, we only looked at co-expressed cliques *(***Figure 6.10a***)* and searched for potential mutual cis- and transregulatory elements (e.g. transcription factors activating or miRNAs repressing both co-expressed genes) with *CyTargetLinker* (Kutman et al. 2013; **Figure 6.10c,d**):

```
Press Ctrl+A > Layout > yFiles > Organic

Mark clique manually > Ctrl+I > Delete

Plugins > CyTargetLinker plugin > Load Regulatory Interaction Networks > select
your network attribute (canonicalName) > OK
```

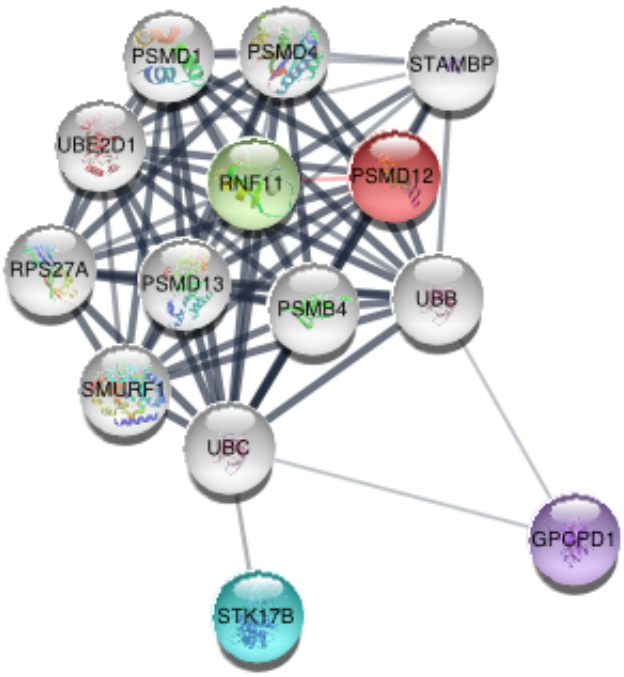
While additionally PPI can be screened for with *stringApp* in *Cytoscape 3.4 (***Figure 6.10b***)*:

```
File > Import > Network > Public Databases... > Data Source: STRING: protein
query > type in genes (i.e. STK17B, PSMD12, RNF11, GPCPD1) > Import

App > STRING > Expand network > OK > Layout > yFiles Layouts > Organic
```

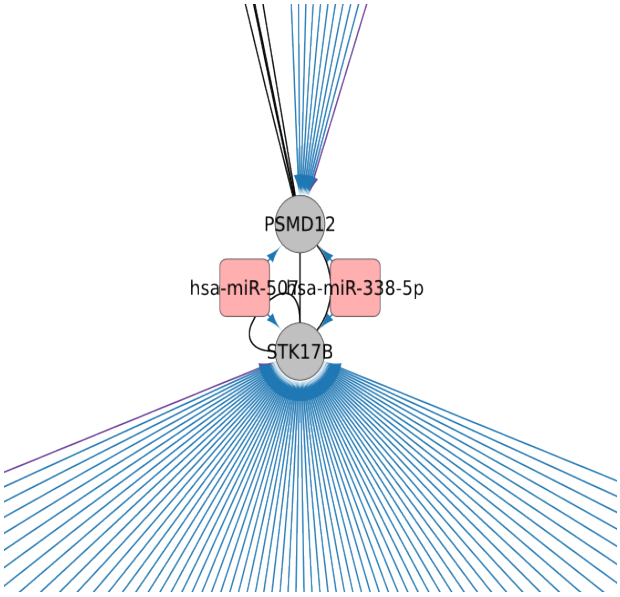
a)



b)



c)



d)

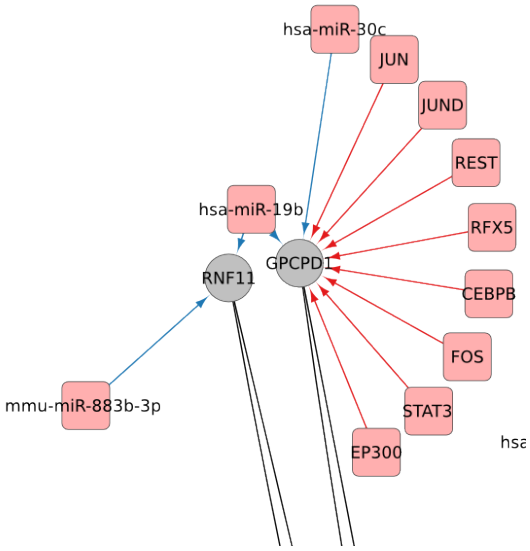


Figure 6.10: **a)** Initial clique graph of four co-expressed genes. Arch of *STK17B* to itself can be interpreted as co-expression between two probes / isoforms and thus be neglected. **b)** Screening for PPI by *stringApp* elucidates that *RNF11and PSMD12* actually interact with each other. Even though the confidence score is rather low, it originates from an experimental assay and not data-mining-based inference. In order to observe potential transitive relationships, the PPI network was expanded (by 10 nodes) and one can see that *STK17B* and *GPCPD1* both interact with UBC. **c)** First part of *CyTargetLinker*-based network integrating data from MicroCosm v5 Homo sapiens, ENCODE network (distal) and ENCODE network (proximal). miR-507 and miR-338-5p both regulate *STK17B* and *PSMD12* expression. **d)** While in the second part, miR-19b regulates both *RNF11* and *GPCPD1*. Further literature mining and re-evaluation in the whole T-PLL cohort (or a second independent one) is needed to evaluate a potential disease mechanism.

6.10 Comparative SNV analysis

Again, SNV or indel screening results in WES or WGS are modeled similar to our GEP or sCNA/CNV analyses (**Figure 6.11**). The only exception is the analysis type (*gen:germline_SNV_by_Sanger_sequencing, gen:germline_SNV_by_Targeted_sequencing, gen:somatic_SNV_by_Whole-exome_sequencing, gen:somatic_SNV_by_Whole-genome_sequencing*) and a couple of added predicates to the results, such as chromosome (*omim_vocabulary:chromosome*), position (*gen:position*), reference allele (*gen:ref_allele*), mutated allele (*gen:alt_allele*), tumor fraction (TF) / variant allele fraction (VAF) (gen:tumor_f), optional predictions by SIFT (*gen:whole-exome_SIFT_score*), PolyPhen2 (*gen:whole-exome_PolyPhen2_HDIV_score*), RadialSVM (*gen:whole-exome_RadialSVM_score*), LR (*gen:whole-exome_LR_score*) and CADD (*gen:whole-exome_CADD_score*), COSMIC 70 annotations (*gen:COSMIC_ID*), read depth (*gen:Depth*), phred-based call quality (*gen:QUAL*) and an optional dbSNP_138 entry (*gen:snp138*).

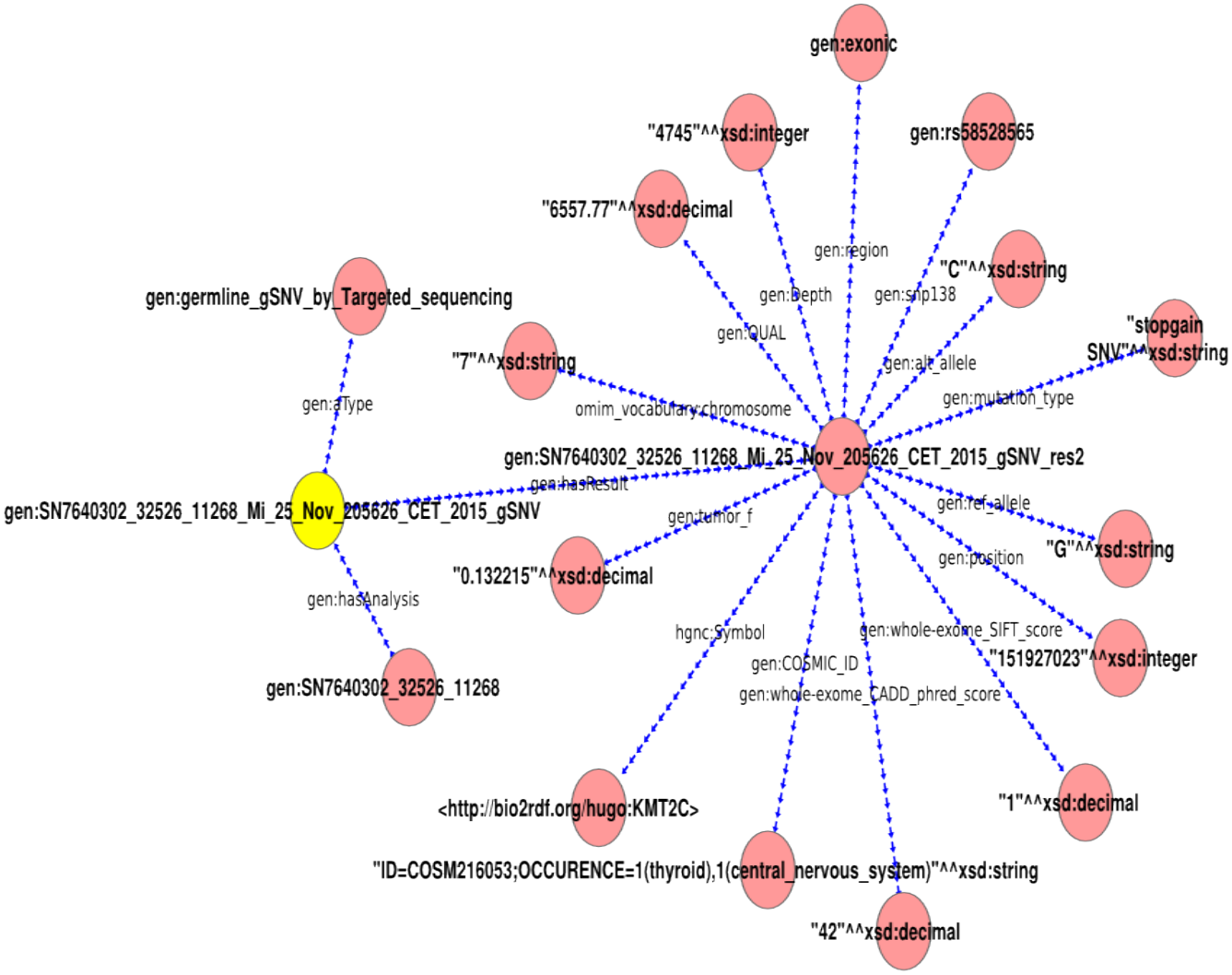


Figure 6.11: WES SNV/indel model. Depicted in yellow is the analysis name (consisting of 'PATIENT_ID', analysis date and analysis type to secure unambiguity) which is linked to a results with a vast amount of mutation annotations.

We thus can convert results from other groups (identifier in PATIENT_ID or analysis description) and compare their findings to ours.

```
SELECT DISTINCT ?pat ?ana ?gene ?lab {

?pat gen:hasAnalysis ?ana .

?pat gen:Diagn gen:T-PLL .

?ana gen:aType gen:somatic_SNV_by_Whole-exome_sequencing .

FILTER(regex(xsd:string(?pat), "Mayo")) . # filter by working group name

?ana gen:hasResult ?res .

?res hgnc:Symbol ?gene .

OPTIONAL { ?res rdfs:label ?lab } .

} ORDER by ?gene
```

6.11 Loss- and gain-of-function analysis

When querying sCNA, UPD and mutations with accompanying VAF, one can further combine them patient-wise and infer loss- and gain-of-function. Meaning that if "the tumor" selects the dysfunctional gene by a second-hit, e.g. when there is a SNV that increases in VAF due to a mono-allelic loss of the other allele (CN < 1.8 & VAF > 0.5) or when there is a SNV that increases in VAF due to a gain (amplification or UPD) of the potential same (!) allele (CN > 2.2 & VAF > 0.5 or VAF > 0.5 & UPD). UPD (*?analysis gen:aType gen:Uniparental_disomy_by_SNP_arrays*) is similarly modeled as sCNA, only it carries no numerical value (nothing like *gen:CopyNumber*), but rather a boolean value (**Figure 6.12**).

```
source("workspace/AG_Herling/Semantic_Framework/functions/basic.R")

# PAIRED SOMATIC + UPD

qS <- "SELECT DISTINCT ?orig ?pat WHERE { SERVICE
<http://localhost:8080/openrdf-workbench/repositories/GEX/query> { \

?orig gen:Diagn gen:T-PLL . ?orig gen:hasSample ?pat } . ?pat gen:hasAnalysis ?
a1 . ?a1 gen:aType ?type . \

FILTER(regex(xsd:string(?type), \"somatic_SNV_by_\")) . FILTER
regex(xsd:string(?a1), \"Mi_7_Okt_123438_CEST_2015_WES_gaIIx\") . }"

iS <- queryProc(qS, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/ngs/query", F)

bAl <- matrix(ncol=3)

for(i in 1:length(iS[,2])) {

    qS <- paste("SELECT DISTINCT ?pat ?gene ?tf WHERE { gen:", iS[i,2], "
gen:hasAnalysis ?a1 . ?a1 gen:aType ?type . FILTER(regex(xsd:string(?type),
\"somatic_SNV_by_\")) . ?a1 gen:hasResult ?r1 . ?r1 gen:alt_allele ?alt . ?r1
gen:position ?pos . ?r1 gen:tumor_f ?tf . ?r1 hgnc:Symbol ?gene . SERVICE
<http://localhost:8080/openrdf-workbench/repositories/CNA/query> { gen:",
iS[i,2], " gen:hasAnalysis ?a2 . ?a2 gen:aType
gen:Uniparental_disomy_by_SNP_arrays . ?a2 gen:hasResult ?r2 . ?r2 hgnc:Symbol ?
gene } } ORDER BY DESC(?orig)", sep="")

    biAl <- try(queryProc(qS, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/ngs/query", F), silent=T)

    if (is(biAl, "try-error")) { biAl <- matrix(c(iS[i,2], "X", "X"), ncol=3); }

    if(dim(biAl)[1] >= 1) { biAl[,1] <- iS[i,2]; }
```

```
    bAl <- t(cbind(t(bAl), t(biAl)))

    print(i)

}

bAl <- bAl[-1,]

library("WriteXLS")

bAl <- as.data.frame(bAl, row.names=F)

colnames(bAl) <- c("Sample_ID", "Gene", "VAF")

WriteXLS(c("bAl"), "~/SNV_UPD_second_hit.xls", SheetNames=c("bi-affected"),
AdjWidth = F, BoldHeaderRow = TRUE, col.names = T, FreezeRow=1)
```

6.12 Combinatorial “bubble” analysis integrating as much data sets as possible to visualize

Building up on **6.11**, one can further include gene expression (by fold changes of each patient compared to the average of a CD3+ normal T-cell pool), UPD status, mutation frequency and FDR of each mutated gene (see Schrader, Crispatzu et al. **Figure 7b**).

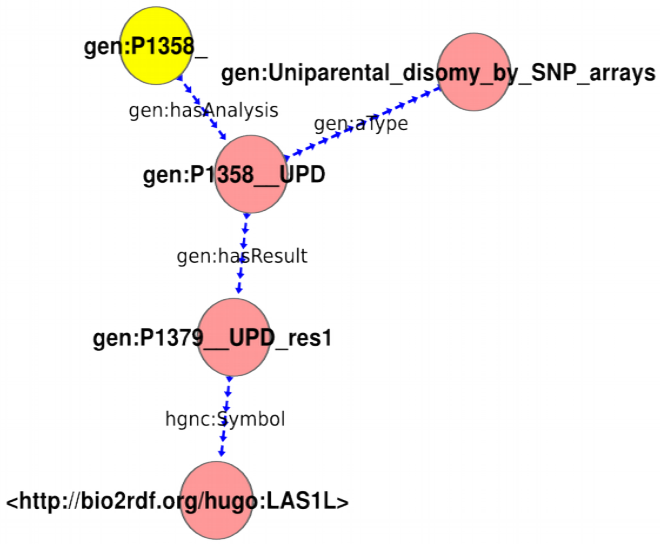


Figure 6.12: UPD model with pseudo-boolean value. 'PATIENT_ID' depicted in yellow either has an UPD result with affected gene („true“) or it does not („false“).

6.13 Combinatorial “bubble” analysis restricted to one gene and it's clonal evolution

Similar to **6.12**, one can restrict the bubble plot to only one gene (like *ATM*; Schrader, Crispatzu et al. **Figure 4a**) and further include coloring of sequential cases (*X rdfs:label „early“^^xsd:string* or *rdfs:label „late“^^xsd:string*) and SNV-affected protein domains (like FAT or PI3K modeled with the attribute *gen:AA_Change_refGene*). One can then further divide into mutiple mutation (*gen:ExonicFunc_refGene* or *gen:mutation_type*), single mutation affected, as well as unmutated cases and test for enrichments (by Fisher table count test) of domain disruptions or gene expression dysregulations (comparing fold changes), as well as co-occurence with mutations of other genes (such as *STAT5B* or *TCL1A* mRNA overexpression):

```
# STAT5B SNVs / indels in all patients, incl. pseudo-somatic singletons

qS <- "SELECT DISTINCT ?orig ?pat ?gene ?pos ?type ?tf ?sift ?phen1 ?phen2 WHERE
{ \

SERVICE <http://localhost:8080/openrdf-workbench/repositories/GEX/query> { ?orig
gen:Diagn gen:T-PLL . ?orig gen:hasSample ?pat . } \

?pat gen:hasAnalysis ?a1 . ?a1 gen:aType ?type . FILTER(regex(xsd:string(?type),
\"_SNV_by_\")) . \

OPTIONAL { ?a1 gen:hasResult ?r1 . ?r1 gen:alt_allele ?alt . ?r1 gen:position ?
pos . ?r1 hgnc:Symbol ?gene . \

FILTER(regex(xsd:string(?gene), \"STAT5B$\")) . ?r1 gen:tumor_f ?tf . OPTIONAL {
?r1 gen:whole-exome_SIFT_score ?sift .  \

OPTIONAL { ?r1 gen:PolyPhen2_HDIV_score ?phen1 } . OPTIONAL { ?r1
gen:PolyPhen2_HVAR_score ?phen2 } } } }"

pat_STAT5Bm_paired <- queryProc(qS, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/ngs/query", F)

# ATM SNVs / indels in all patients, incl. pseudo-somatic singletons

qS <- "SELECT DISTINCT ?orig ?pat ?gene ?type ?tf ?sift ?phen1 ?phen2 WHERE { \

SERVICE <http://localhost:8080/openrdf-workbench/repositories/GEX/query> { ?orig
gen:Diagn gen:T-PLL . ?orig gen:hasSample ?pat . } \

?pat gen:hasAnalysis ?a1 . ?a1 gen:aType ?type . FILTER(regex(xsd:string(?type),
\"_SNV_by_\")) . \

OPTIONAL { ?a1 gen:hasResult ?r1 . ?r1 gen:alt_allele ?alt . ?r1 gen:position ?
pos . ?r1 hgnc:Symbol ?gene . \

FILTER(regex(xsd:string(?gene), \"ATM$\")) . ?r1 gen:tumor_f ?tf . OPTIONAL { ?
r1 gen:whole-exome_SIFT_score ?sift .  \

OPTIONAL { ?r1 gen:PolyPhen2_HDIV_score ?phen1 } . OPTIONAL { ?r1
gen:PolyPhen2_HVAR_score ?phen2 } } } }"

pat_ATMm_paired <- queryProc(qS, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/ngs/query", F)


a1 <- unique(pat_ATMm_paired[which(!is.na(pat_ATMm_paired[,3])),1])

a2 <- unique(pat_STAT5Bm_paired[which(pat_STAT5Bm_paired[,4] == "40359729"),1])

pp <- length(intersect(a1,a2))                    # 8

pn <- length(setdiff(a1,a2))                      # 28

np <- length(setdiff(a2,a1))                      # 1

nn <- length(unique(pat_STAT5Bm_paired[,1]))-length(union(a1,a2))   # 18

fisher.test(rbind(c(nn, np), c(pn, pp)))

# p-value = 0.01875 -> 0.1411

fisher.test(rbind(c(nn, np), c(pn, pp)), alternative="greater")

# p-value = 0.009376 -> 0.1052
```

6.14 Gene set to abberations

As mentioned in **6.4** and **6.5** gene sets can be stored in a Semantic Web list structure or just as a regular character vector in R. Each gene can then be iteratively queried for various kinds of mutations, such as sCNA, SNVs or indels (even SVs, GEP or fusion-transcripts). We manually curated three gene sets for DNA damage response (*DDR*), epigenetic modifiers (*EPI*) and telomere maintenance genes (*TELO*), since we previously observed these pathways as overrepresented in GEP and point mutation analysis.

6.15 Binary or gradual summary table

To come up with an initial disease model, observe patient clusters and mutation overview, one can combine above mentioned single steps (**6.6**, **6.11**, **6.14**) and store their results in a binary or multivariate/numerical summary table (see Schrader, Crispatzu et al. **Figure 7c**). We therefore queried first all patients and linked samples, since we want to visualize aberrations throughout the disease course, including classical cytogenetics, such as *TCR* or *TCL1A* locus rearrangements (*gen:Molecular_data_TCR_gene_rearrangement* & *gen:TCL1_rearrangement* respectively). The numerical results from the queries (are converted to binary values and) can be visualized in R with the *tableplot* or *ComplexHeatmap* package. Before passing the summary table to the plotting functions one can order them by frequency of certain genes. One can further use these tables to automatically identify clusters and most informative subsets for a better clinical guidance by machine learning techniques, such as decision trees (within the *rattle* package) and SVM.

In practice decision trees divide a table of different variables (numerical, binary, cardinal) into the most variable (by ANOVA) categories according to a (linear) fit and are able to handle missing information (by using the next best variable). Reasoning can then be obtained by following the tree from leaf to root (see Crispatzu et al. 2016, **Figure 7**).

6.16 Telomere length corralations

As short telomeres are frequently seen in T-PLL (Röth et al. 2007), we measured the telomere length by *flow-FISH* (Baerlocher et al. 2006) of different leukemia/lymphomas in collaboration with the F. Beier group. While the T-PLL has by far the shortest ones (already age-corrected), possible causes still remain unknown (see Schrader, Crispatzu et al. 2016 **Figure 4f**). I therefore modelled telomere length for each analysed patient (*?patient gen:Delta_Lympho ?delta*). It is then possible to correlate telomere lengths against all annotated parameters, such as OS, WBC, *ATM* mRNA expression, *ATM* copy-number, *ATM* VAF and *TCL1A* mRNA expression. We put our focus on *ATM*, hence its homologue in yeast *Tel1* is responsible for telomerere maintenance and *ATM* abberations and high telomerase activity, as well as chromosome instability are frequently seen in A-T (Gabellini et al. 2003; Petrinelli et al. 2001).

```
# querying telomere length and different ATM dysfunctions

qS <- "SELECT DISTINCT ?orig ?pat ?delta ?tf WHERE { ?orig gen:Diagn gen:T-PLL .
?orig gen:hasSample ?pat . ?pat gen:Delta_Lympho ?delta . SERVICE
<http://localhost:8080/openrdf-workbench/repositories/ngs/query> {?pat
gen:hasAnalysis ?a1 . ?a1 gen:aType ?type . FILTER(regex(xsd:string(?type),
\"_SNV_by_\")) . FILTER(!regex(xsd:string(?type), \"Sanger\")) . ?a1
gen:hasResult ?r1 . ?r1 gen:alt_allele ?alt . ?r1 gen:position ?pos . ?r1
hgnc:Symbol ?gene . FILTER(regex(xsd:string(?gene), \"ATM$\")) . ?r1 gen:tumor_f
?tf . } }"
```

```
telo_q <- queryProc(qS, prefixes, "http://localhost:8080/openrdf-
workbench/repositories/GEX/query", F)

# sign. negative correlation between telomere lengths and ATM VAF in T-PLL

cor.test(as.numeric(telo_q[,3]), as.numeric(telo_q[,4]))

# p-value = 0.005047; cor = -0.6638087
```

6.17 Trace back fusion-transcript to structural variation (or copy-number variations)

Fusion-transcripts are modeled similar to co-expression results (**Figure 6.13)**. They also carry predicates for HUGO/HGNC gene symbol pairs for both ends of the fused transcript. In addition they carry different predicates, such as read depth and read support (*gen:NrOfSpanningReads, gen:NrOfSpanningMatePairs, gen:NrOfSpanningMatePAirsOneEndsSpansFusion*).

Both HGNC/HUGO gene symbols can be matched to those of sCNA and thus infer the root causes.

Structural variants have two sets of coordinates (*gen:LeftChr1, gen:RightChr2, gen:LeftPos1, gen:RightPos2*). Their coordinates can further be used to overlap (with a threshold range) the CDS of fusion-transcripts and infer whether the root cause is a SV (**Figure 6.14)**.

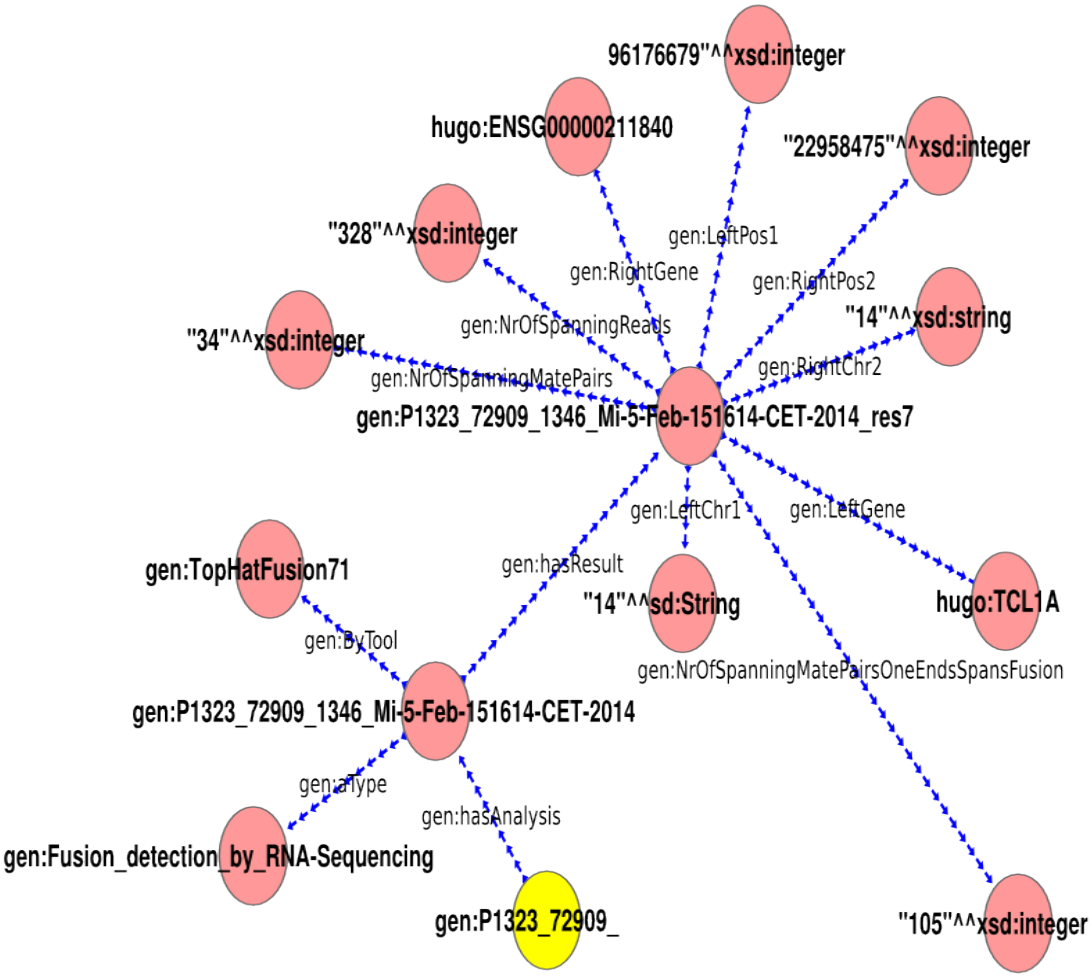


Figure 6.13: Fusion-transcript model with coordinates and coverage information. 'PATIENT_ID' is shown in yellow and can be queried for other analyses.

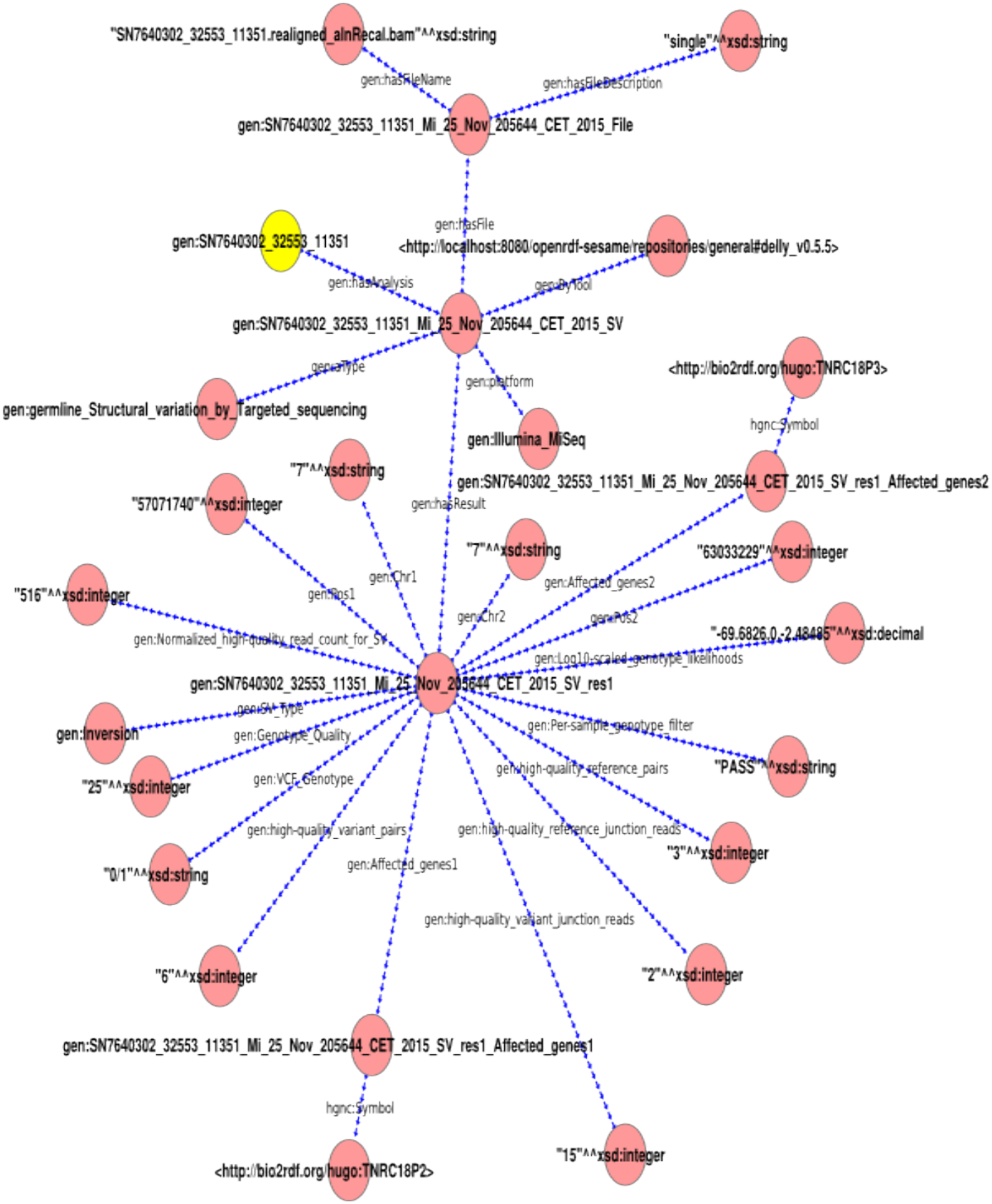


Figure 6.14: SV model with vast amount of coverage and genotype information.

6.18 Correlation of breakpoint distance to affected gene expression

The range information of structural variants can further be combined with mRNA expression levels. So the breakpoint of inversion in chr. 14 can be linked to different TCL1A levels (protein and mRNA) and thus elucidating activation and silencing mechanisms. Pairwise correlation of the mRNA expression of each TCL1 family member probe (*TCL1A, TCL1B*, *TCL6*, *MTCP1*) and average breakpoint distance to the TCL1 locus (end of *TCL1A* CDS) can further be visualized (*pairs()*). No sign. correlation was however observed.

Therefore non-linear correlations were investigated with the calculation of pairwise mutual informations as well (see **Discussion**).

As a FACS cut-off to determine the TCL1A protein status of a T-PLL sample, we used <=5% of cells for negative cases, between 5% and 50% for intermediate/dim cases and >50% for positive cases (*gen:hasTCL1A_FACS*). Since negative cases can have low TCL1A protein, but high mRNA expression levels, it is important not to falsely impute from array data. Only dim cases may be declared as positive cases (*gen:hasTCL1Astatus*), when additionally to their intermediate protein levels, they show high mRNA expression level. When status overlap, *MTCP1* status has priority over *TCL1A* status.

6.19 Correlations of Vbeta chains and surface markers

Surface marker status and vBeta spectratyping for each patient is assigned in continuous expression values (% T-cells gated; *gen:percentage_Vbeta8_CD5p_T-cells_gated*).

Pairwise spearman correlations were calculated to measure co-occurences and thus subpopulations. The correlations of the expression frequencies of immunophenotypic markers in T-PLL cells isolated from peripheral blood, such as surface markers (Warner, Oberbeck, Schrader et al. **Figure S3b**) and Vbeta chains (data not shown) are then visualized in a heatmap.

6.20 FACS sample organization and SPADE analysis

The results of immunophenotyping by manual gating of FACS (fluorescence-activated cell sorting) analyses are by default (e.g. in Beckman & Coulter Gallios Flow Cytometers) stored in *LMD* or *fcs* files. These can be read by proprietary software or in R with the library *flowCore* and accompanying *flowViz.* Besides numerical values used for gating, marker name and descriptions are stored. When dealing with a multitude of these files the semantic database can be used to store all FACS metadata and through its queries full batches or specific tubes with overlapping marker can be selected (**Figure 6.15**). This overlap can then further be used for automatic, agglomerative clustering of cell-sorting values by *SPADE* (Qui et al. 2011). In Warner, Oberbeck, Schrader et al. (**Figure 1c, S1c**) we used this sort of data-mining and unbiased population detection to observe a higher central-memory phenotype compared to a transitional one.

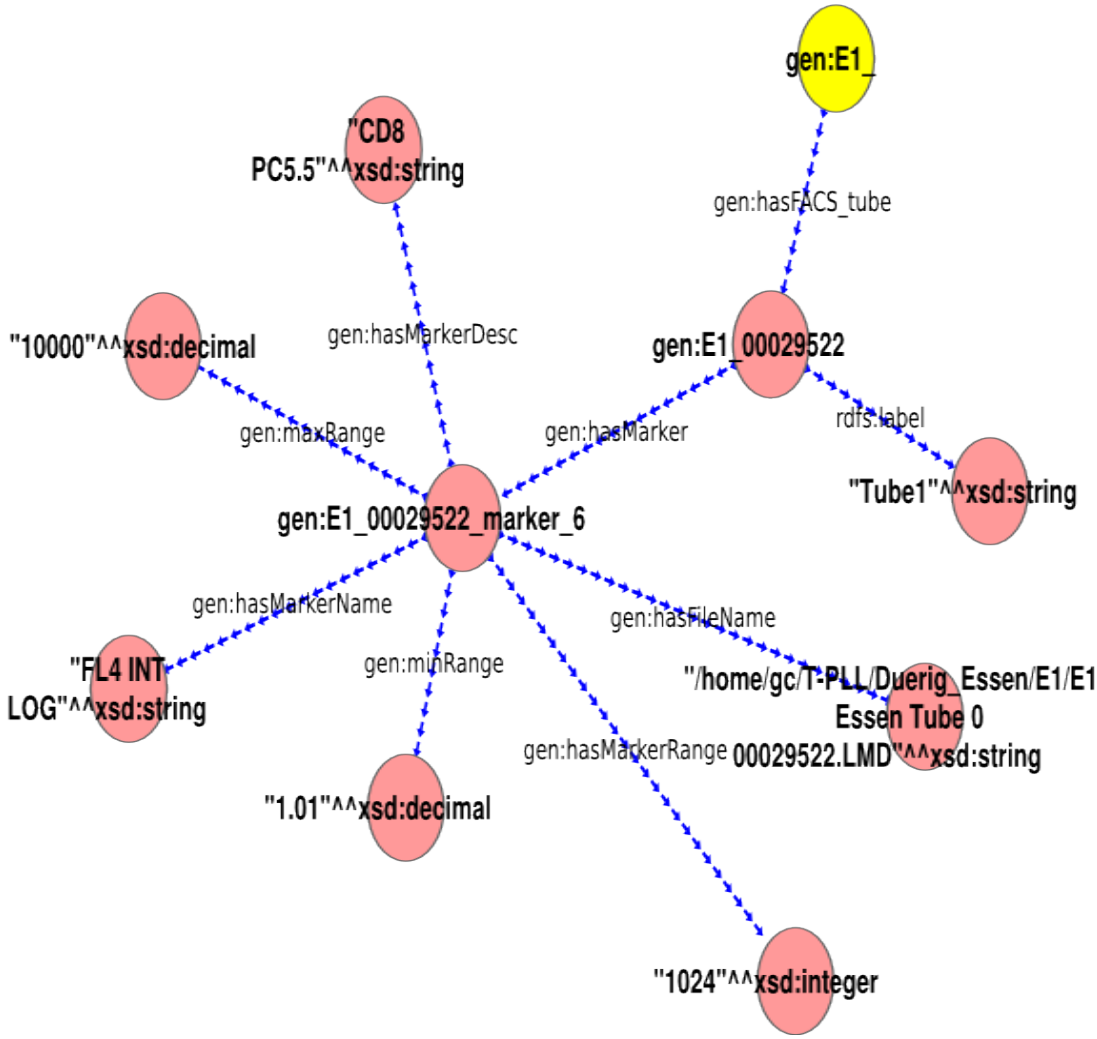


Figure 6.15: FACS model describing a marker within a specific tube. 'PATIENT_ID' depicted in yellow can have several tubes.

6.21 Temporal analysis

Due to the patient and sample organization mentioned in **6.2** one can further investigate the disease course or treatment responses. I queried for each patient, if available, the earliest (*"early"^^xsd:string* or *"FU"^^xsd:string*) and latest sample (*"FU"^^xsd:string* or *"late"^^xsd:string*) and gathered their corresponding gene expression array sample to analyse differential expression (late vs. early). A similar analysis can be done with sCNAs (by *gen:CopyNumber*), SVs (by *gen:NrOfSpanningReads*), or SNVs / indels (by *gen:tumor_f*). If no gene is significantly differential altered (gained or lost, amplified or deleted, re-arranged or point mutated), the load (number of losses & gains, structural variants or point mutations) can be sequentially compared or known (dys)functional genes are case study-wise compared (see Schrader, Crispatzu et al. **Figure S16**).

6.22 Further case studies planned

6.22.1 TCL1A-interactor status and clinical subsets in CLL

Since *TCL1A* is small molecule with no targetable binding pocket and therefore hard to circumvent its *TCR*-modifying and *AKT*-enhancing functions with current inhibitors, it may be possible to inhibit his interactors or the formed complex. We therefore investigated the therapeutic response of *TCL1A* and its interactors in CLL following FCR chemotherapy**.** We used *STRING* (Szklarczyk et al. 2015) and its PPI networks (target network as basis)

for compound screening. When no overlap is present, then we want to know what is the shortest path (BFS; breadth-first-search) to common hubs and co-expressed genes.

6.22.2 Potential compounds

I already mentioned some first practical uses of the EMBL / EBI RDF platforms. Besides a SPARQL endpoint for 'Gene expression Atlas', there are numerous others like the 'ChEMBL' (Gaulton et al. 2012) one. With the information of GEP and SNV enrichment in our T-PLL cohort, one can further link frequently aberrated genes with external information, such as results from compound assays to find possible intervention clues, e.g. screening for possible inhibitors of upregulated genes.

6.22.3 Search for in vivo/in vitro models for selected gene set aberrations

A selected number of genes (gene set) are queried for deregulations in lymphoid leukemias (such as CLL). Each gene should be deregulated in at least three distinct/independent (no re-used samples from same or cooperating investigators) with the same fold change direction (all three up- OR downregulated, not up- AND downregulated) to exclude batch-effects and guarantee consistency. These unambiguous deregulation are then looked for in different model organisms and overlapping inducable therapy, stimuli or other interventions (e.g. „tamoxifen-treated Danio rerio vs. wt Danio rerio“). Ideally they are multiple distinct experiments with the same experimental conditions and same observed deregulations found.

Even though differing from human samples in setting and whole-transcriptome, they may explain or allow to study specific pathway or gene set aberrations.

6.22.4 Boolean networks executable

We calculated pairwise Pearson correlation coefficients between the approx. 25000 annotated genes on our T-PLL GEP Illumina HumanHT-12 v4 Expression BeadChip (n=83) and additionally overlapped the highest absolute values (rho>0.8) and most significant (p-value<0.01) correlations with significant deregulations (|FC| > 2; p-value<0.01) between T-PLL and normal CD3+ T-cells. These co-expression graphs can be overlapped with annotated pathways (as a further restriction) with help of a Semantic framework (Dehmer et al. 2011). These reduced networks can then be used as basis for Boolean networks (Wang 2008) to e.g. investigate *TCL1A*-enhanced *TCR* signaling in T-PLL. In concrete terms Boolean networks can be employed by modeling usual pathway maps as logical gates and play through all possibilities of how one gene activates another by state frequencies. Updates of network nodes can be realized as synchronous or asynchronous (Albert et al. 2008).

6.22.5 Integrative benchmark of high-throughput analyses

The downstream RDF parsers (data not shown) of the *QuickNGS Cancer* pipeline (Crispatzu, Kulkarni et al.) enable us to evaluate the performance of certain NGS tools with differing runtime parameters. Mentioned parsers write a RDF log file with job name (*gen:<SAMPLE_ID>_<analysisDate>*), linked to sample name (*gen:produces*) and software meta-information (**Figure 6.16**). Through the calculation of the average or maximal runtime, one can then avoid possible downtimes or premature aborts of pipeline steps.

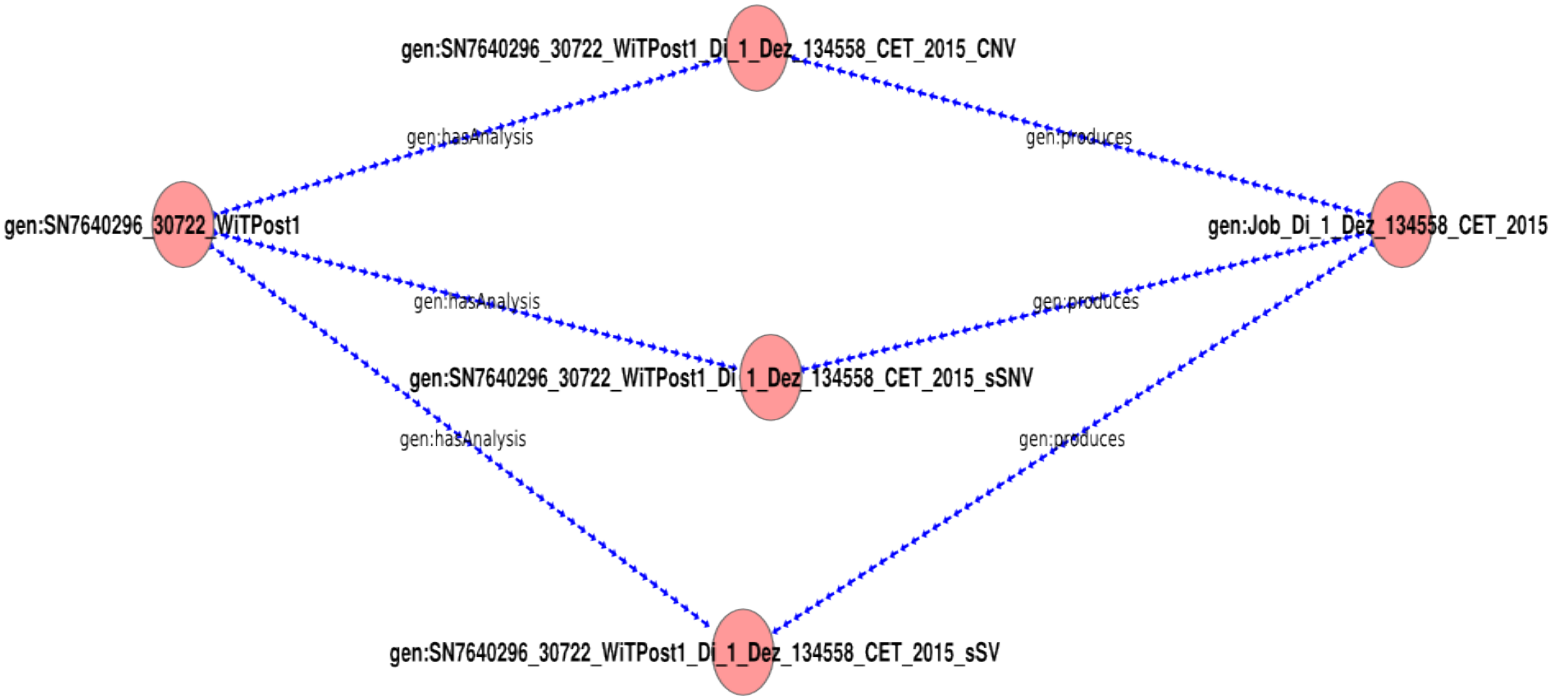


Figure 6.16: Job model which elucidates that one job can produce multiple analyses which in turn can be linked to the 'PATIENT_ID' and its molecular and clinical information. The job node itself can further be annotated with runtime parameters and meta-information of used NGS tools.

Current_terms	Ontology_terms
Splenomegaly	http://sideeffects.embl.de/se/C0038002/
Hepatomegaly	http://sideeffects.embl.de/se/C0019209/
Anemia | Anaemia	http://sideeffects.embl.de/se/C0002871/
isTreatedWith	db:sider/drugs
Fludarabine	http://sideeffects.embl.de/drugs/3367/
cyclophosphamide	http://sideeffects.embl.de/drugs/2907/
Alemtuzumab	https://www.ebi.ac.uk/chembl/compound/inspect/CHEMBL1201587
transcription profiling by array	http://www.ebi.ac.uk/efo/EFO_0002768
genotyping by array	http://www.ebi.ac.uk/efo/EFO_0002767
genotyping by high throughput sequencing	http://www.ebi.ac.uk/efo/EFO_0002771
RNA-seq of coding RNA	http://www.ebi.ac.uk/efo/EFO_0003738
Illumina Genome Analyzer IIx standard manufactur	http://www.ebi.ac.uk/efo/EFO_0005084
Illumina HiSeq 2000 standard manufacturer's proto	http://www.ebi.ac.uk/efo/EFO_0005086
Illumina MiSeq standard manufacturer's protocol	http://www.ebi.ac.uk/efo/EFO_0005087
whole genome shotgun sequencing	http://www.ebi.ac.uk/efo/EFO_0003744
tumor stage	http://www.ebi.ac.uk/efo/EFO_0004925
biological replicate	EFO:0002091
technical replicate	EFO:0002090
investigator	EFO:0001739
control	EFO:0001461
ploidy	EFO:0000659
female	EFO:0001265
male	EFO:0001266
alive	PATO:0001421
dead	PATO:0001422
copy number variation	EFO:0004798
somatic genotype	EFO:0004972
p-value	OBI:0001442
BAM format	EFO:0004157
FASTQ format	EFO:0004155
CEL data file format	EFO:0005630
age at death	EFO:0005056
age at onset	EFO:0004847
age at diagnosis	EFO:0004918
alive at endpoint	EFO:0004951
date of diagnosis	EFO:0004953
event free survival time	EFO:0000482
progression free survival	EFO:0004920

Table S6.1: Currently used terms in our models and proposed replacements for the near future. Ontology matches were investigated using BioPortal (Noy et al. 2009).

7. Discussion

Within this thesis, I presented a novel hands-on approach for integrative analysis of high-throughput data sets. The newly developed tools which enable scientists to answer sophisticated questions across multiple data types in form of queries to a Semantic Web server (without the need to handle dozens of Excel sheets) were applied to lymphoid leukemia data sets of T-PLL and CLL. Upstream of this framework lies a NGS platform for semi-automated analysis of cancer genomics data. An additional plug-in converts the output into RDF. Downstream of the semantic framework, functional data was combined as a contribution to subsequent publications.

7.1 Semi-automated cancer genomics pipeline enables rapid data processing and delivers semantic output for integrative analyses

Within the first publication (Crispatzu, Kulkarni et al.), we introduced a novel semi-automated pipeline for the analysis of cancer genomics data ranging from DNA sequencing data as whole-genome, whole-exome and target capture- or amplicon-based sequencing to RNA-Seq in mice and humans. It allows the identification of basic genetic (not epigenetic thus far) aberrations found in cancer genomes, such as single or multiple nucleotide variants, structural variations, copy-number aberrations, as well as the identification of differential expression and exon usage or fusion transcripts in dependence with important parameters as tumor ploidy and specimen purity (pre-set or automatically inferred). The pipeline is embedded in the MySQL- and HPC (high performance computing)-based framework *QuickNGS* (Wagle et al. 2015) with an easy-to-use graphical front-end. For the integrative, downstream analyses presented in the other three publications (Schrader, Crispatzu et al.; Warner, Oberbeck, Schrader et al.; Crispatzu et al. 2016), I further used a plug-in in form of multiple parser and add-on tools to convert the pipeline results into RDF.
Within Schrader, Crispatzu et al. and a previous version of *QuickNGS Cancer*, we used *ExomeDepth* (Plagnol et al. 2012) to call somatic copy-number aberrations, and *Tophat-Fusion* (Kim et al. 2011) to detect fusion transcripts in T-PLL. Both tools have been recently shown to perform generally adverse (Nam et al. 2016; Liu et al. 2016) compared to e.g. *EXCAVATOR2* (D'Aurizio et al. 2016) and *Jaffa* (Davidson et al. 2015) respectively. They were therefore replaced with the mentioned, superior programs and the data within Schrader, Crispatzu et al. was re-evalulated with SNP array analysis and *STAR-Fusion* (Dobin et al. 2012) / *Jaffa* respectively. This analysis confirmed our results in key findings.
Since the submitted version only marks a ground stone, further modifications according to novel benchmarks and algorithms are necessary. This includes foremost sequential and comparitive sample analysis, as well as detection of significantly mutated genes (*MuSiC* (Dees et al. 2012), *MutSigCV* (Lawrence et al. 2013)), mutation hotspots, contexts and co-occurences in the light of clonal evolution and treatment response. Generally more sophisticated, publication-ready graphical representations, e.g. of structural variations with *Circos* (Krzywinski et al. 2009), are needed as well. More advanced add-ons may then include detection of viral transcripts or integration sites (Li et al. 2015), as well as measurements of chromothripsis and microsatellite instability (Niu et al. 2014) or inference of possible drug targets.

7.2 Integrative framework provides means to describe the ATM/TCL1-centered genomic landscape of T-PLL

We presented the most recent, most diverse (in terms of proto-oncogene status and data sets) and largest reported cohort of T-PLL, including sequential samples. The integrative capabilities of the semantic framework really came to fruition here, as it was applied to possible dosage effects (**6.7**), second hit analysis (**6.11**), clonal evolution (**6.13**) and correlation analysis with molecular and clinical parameters (**6.3**). In terms of gene expression profiling, compared to normal CD3+ T-cells, we found overexpression of *TCL1A*, *MTCP1* or *TCL1B* in the majority of cases possibly leading to further aberrant expression of negative TCR regulatory genes like *SLAMF6* or *CTLA4*, which seem also to be integral in its clonal evolution in patients and murine models mimicking T-PLL. The *TP53*-dependent arrest mediator *RPRM* (reprimo) was further among the most highly and variable expressed genes across T-PLL. In prostate cancer (Ellinger et al. 2008) and gastric cancer (Bernal et al. 2008) *RPRM* is hypermethylated, *in vitro* it is highly expressed in response to DNA damage, and *in vivo* it inhibits tumorigenesis (Ooki et al. 2013). However its exact function in the dysfunctional DDR of T-PLL has to be further evaluated. The lack of much overlap between deregulated and aberrantly lost / amplified genes in T-PLL and *Lck^{pr}-hTCL1A-tg* mice model surprised us. This may be explained by the complex interplay of deregulated factors like *TCL1A, ATM* or *JAK3*, which were not initially perturbed in *Lck^{pr}-hTCL1A-tg* mice. *TCL1A* upregulation itself therefore may not be enough to induce genomic instability, but rather the consequence of the structural rearrangements leading to or descending from the inv(14). This phenotypic hallmark of T-PLL, likely due to failed maintenance of telomeres and aberrant DSB-induced recruitment and diminished activation of *ATM* and its substrates is demonstrated by complex losses and gains. These cumulative copy-number events in T-PLL are ranking above CLL and just below solid tumors and ALCL (Anaplastic large cell lymphoma) when compared by frequencies. Screening for somatic copy-number aberrations in SNP arrays and whole-exome sequencing data confirmed the deletion of chromosome 11q (52%), affecting *ATM* and the *TCL1A* regulator *miR-34b/c,* and the isochromosome 8 or amplification of 8q. However we were able, through FDR correction and FISH confirmation, to dispute *MYC* as being the most frequently gained gene on chromosome 8. Rather the argonaute 2 protein, *AGO2*, was amplified in the majority of cases. Other argonaute family members *AGO1/3/4,* which are located on other chromosomes, were affected by high-frequent UPD (uniparental disomy). Their dysfunction in miR-processing and nuclease activity (only of *AGO2*) within T-PLL, as well as how mutations of *miR-484* (mutated in n=1/3 WGS cases) may affect *TCL1A* and thus T-PLL tumorigenesis, may be investigated in the future by means of comparative microRNA-sequencing between normal T-cells, ampl(*AGO2*) or UPD(*AGO1/2/4*) cases and biallelic, heterozygotic T-PLLs. Within melanoma, *AGO2* is dowregulated only at the protein level, not as mRNA (Völler et al. 2013), while in hepatocellular carcinoma *miR-99a* is overexpressed which in turn downregulates *AGO2* (Zhang et al. 2014). Oncogenic interactions with *KRAS* leading to decreased gene-silencing have also been observed (Shankar et al. 2016).
At the mutational level, we observed clonally dominant *ATM* mutations in the majority of cases (66%), due to loss of the remaining functional allele or UPD of the mutated allele. The residual cases contain mutations of other DDR or MMR genes, such as *ERCC6L2* or *MSH3*, epigenetic regulators (e.g. *EZH2*) or mostly mutually exclusive, subclonal *JAK3* (15.38%) and *STAT5B* (53.84%) SNVs potentially leading towards late-stage TCR/cytokine independence. This is in contrast to T-LGL (T-cell large granular lymphocytic leukemia) patients, the mature T-cell leukemia T-PLL is most often misdiagnosed as, where *STAT5B*

is only mutated in a low fraction (~2%), while the orthologue *STAT3* with its SH domain (exon 21) is being predominantly mutated (28% to 40%; Koskela et al. 2012).
T-PLL samples further exhibit shorter telomeres than any other T-cell lymphoma/leukemia investigated, as well as CLL, where shorter telomeres have been previously only described in T-cells of ZAP70+/CD38+ subtypes (Röth et al. 2008). Whether this is cause or effect of genomic instability is unclear, however reduced telomere length correlated with variant allele fraction (VAF) of *ATM* mutations, as well as *ATM* copy-number decrease.
ATM mutations, while being the most common denominator in T-PLL besides *TCL1A*, are virtually absent in T-LGL, as are other DDR gene defects. Globally, we barely see any clonal mutations in T-PLL besides *ATM*, but an excessive amount of G>T & C>A mutations is observed. After filtering for potential OxoG (8-Oxoguanine) bias during sample preparation (Costello et al. 2013), we propose that this may be due to unrepaired DNA damage induced by functional *ATM* deficiencies in interplay with *TCL1A*-augmented mitochondrial ROS biogenesis (Prinz et al. 2015). The mutational signature most closely resembles the ones of ageing and smoking (**Figure 7.1**), hinting towards a synergy of non-predisposed accumulations and exterior influences or oxidative damage. ROS may further function as an activating molecule in TCR signaling or vice versa (Sena et al. 2013; Williams & Kwon 2004). Since our synthetic lethality approach, i.e. DNA-PKcs inhibitors, failed to induce apoptosis in T-PLL, probably due to residual function of *ATM* and incomplete compensation by stand-in's (i.e. *ATR*), we explored alternative approaches to reconstitute sufficient DDR response and targeting of epigenetic aberrations in T-PLL. We tested an unique customized *HDAC*-inhibiting / DSB-inducing agent that has shown such promising results in primary T-PLL cells, mice transplanted with *JAK1-initiated* and *CD2-hMTCP1^{p13}-tg* mice cells, that a clinical trial has already been commenced (NCT02576496). Possible synergies with telomerase inhibitors (as in Röth et al. 2007) were not investigated. The exact mechanism, which gene signature (or perhaps *TP53* itself) is (de)methylated or (de)acetylated before and after [redacted], can be elucidated by comparative methylome profiling and chromatin immunoprecipitation of treated and non-treated patients in the near future. We can then further investigate the proposed link between *JAK3* mutations in T-PLL (n=3/13) and its possible phosphorylation of *EZH2* (n=2/13 mutated in T-PLL, with second hit deletion), leading to its loss of methyltransferase activity (tumor suppressive) and switching to transcription co-activation (oncogenic) (Yan et al. 2016). *EZH2* can further be phosphorylated by AKT (TCL1 family interaction partner) and thus its H3K27me3 enzyme activity be inhibited (Cha et al. 2005).
Since lesions in *ATM* and *STAT5B* are co-occuring, while *JAK3* is mostly exclusive, inhibitors may be (only) a complementary approach to these parallel aberrations. Within three distinct fusion detection programs, we further observed a multitude of JAK-affected genomic fusions: *TRIM22:JAK2* (n=1/15) in all three (*Tophat-Fusion*, *STAR-Fusion* & *Jaffa*), as well as *JAK1:PTMA* (n=1/15) in only *Jaffa*. We are in the process of validating these by Sanger sequencing and induction experiments.
In summary, we were able to formulate a first integrative model of step-wise T-PLL leukemogenesis (**Figure 7.2**) providing a concrete basis for refined diagnostics, prognostication, and therapeutic concepts in this problematic disease.
When re-considering the sampling of patient data over the last 4 years due to recent drop in costs and availability in standard analysis tools, data like SNP arrays, mRNA arrays and whole-genome sequencing seem now obsolete for our cohort. Somatic copy-number aberrations and structural variation could have been easily called on a larger whole-exome cohort. Non-coding or regulatory mutations showed a heterogeneous pattern that could have been more precisely determined by e.g. miRNA-sequencing or methylome arrays of a medium-sized cohort. While gene expression profiling could have been conducted via

RNA-Seq, further increasing sample size to study differential splicing and fusion detection. It is therefore mandatory to expand our model with novel NGS data, preferably with many sequential samples.

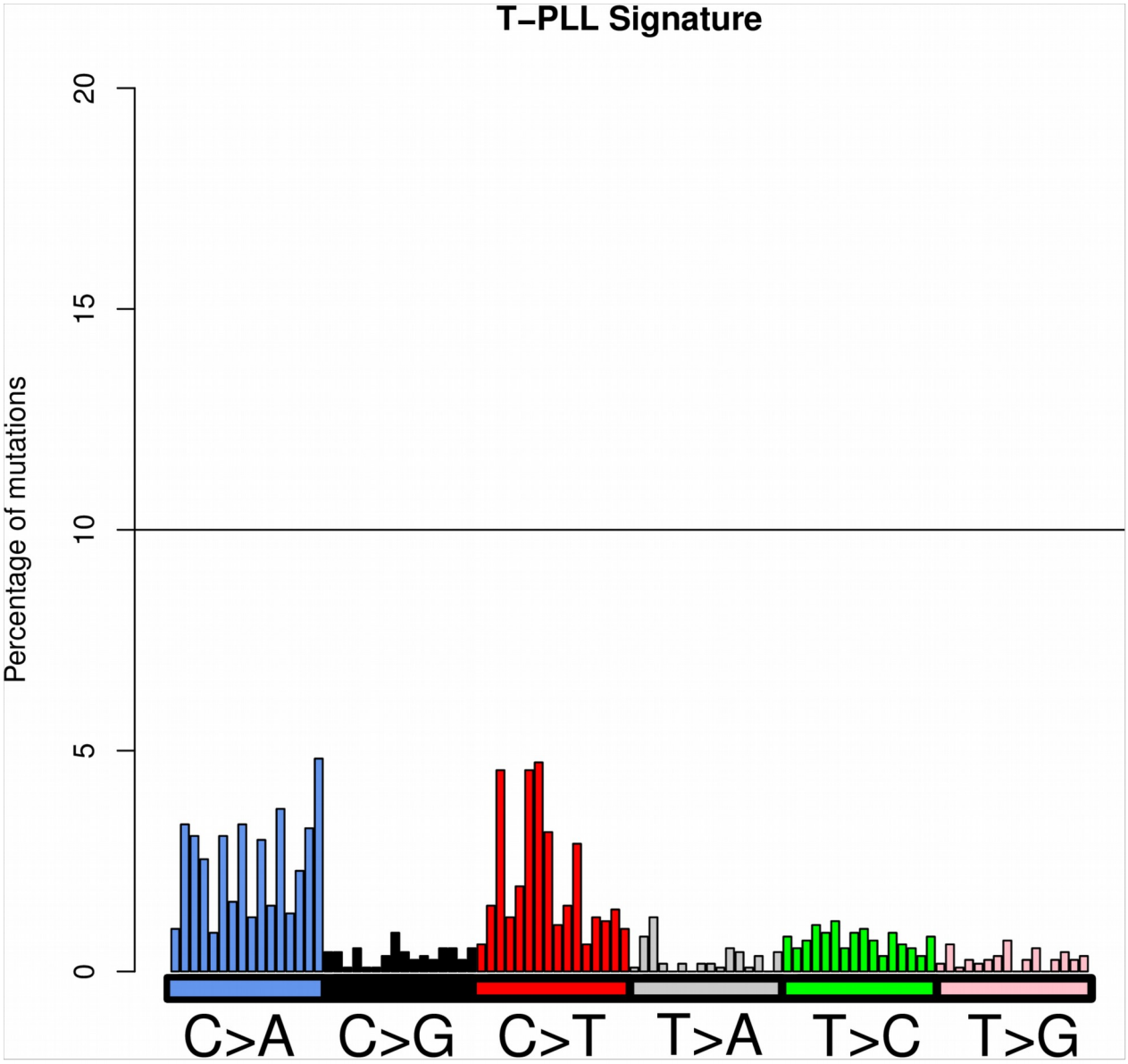
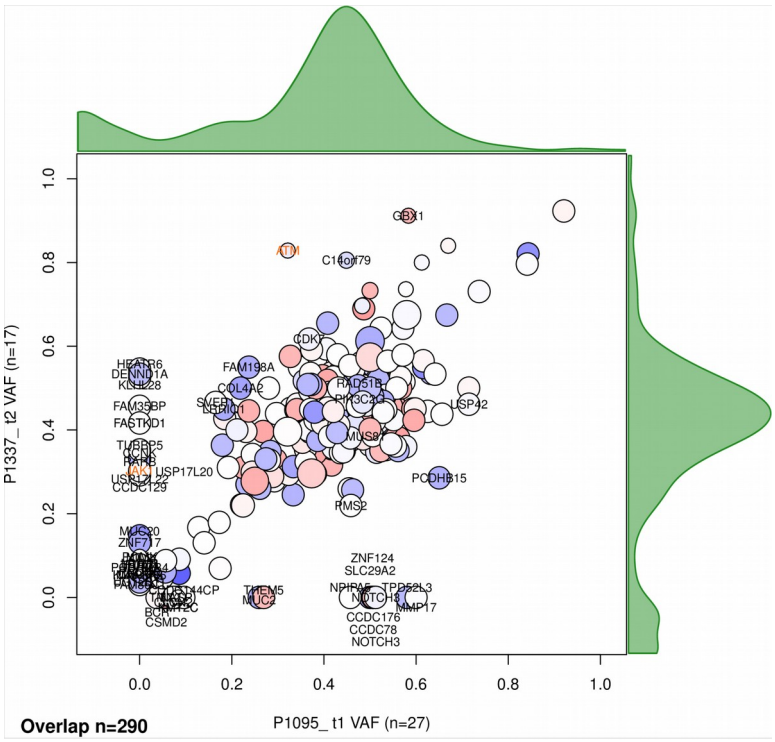


Figure 7.1: Mutational signature of T-PLL (average over 13 paired WES) mostly resembles “Signature 1B” (60.7%; Age; among them CLL) and “Signature 4” (12.1%; Smoking; especially in solid tumors). from Alexandrov et al. 2013.



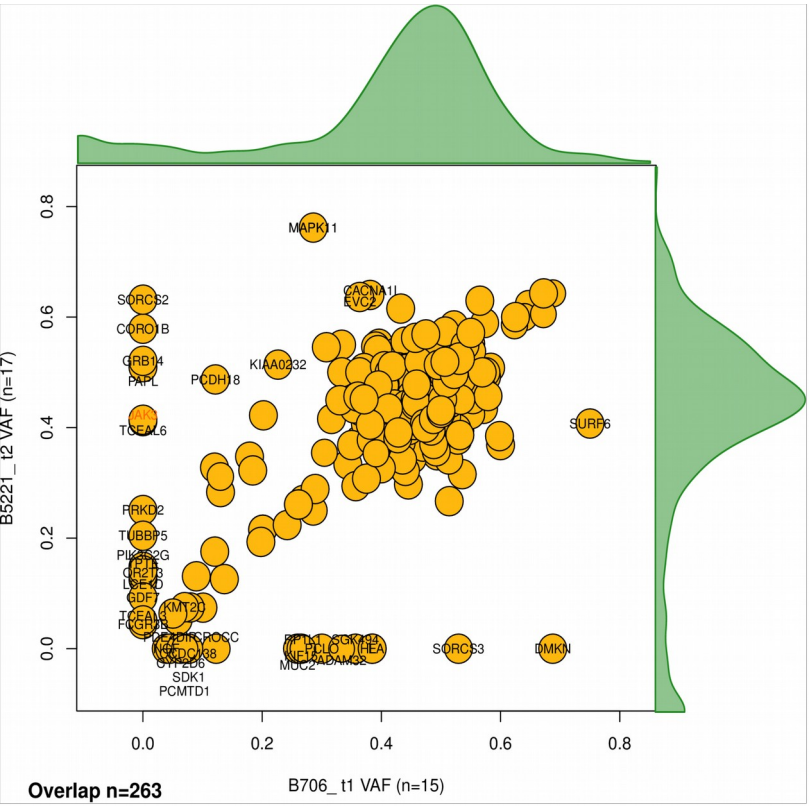


Figure 7.2: Summarized clonality analysis of 3 sequential T-PLL patient pairs. Variant allele fractions (VAFs) are depicted on both axis. Time point 1 (t1) on the x-axis. Time point 2 (t2) on the y-axis. Color of nodes symbolizes degree of up- (red) or down-regulation (blue) of the respective gene between t2 and t1. As sample pair at bottom has no corresponding microarrays, color gradient is absent. Size of the „bubbles“ corresponds to copy-number (CN=1=small; CN=3=large). All values were retrieved from and integrated with the semantic framework. Fitted histograms on tops depict VAFs and show that besides ATM and JAK3 (highlighted in orange font) barely any genes experience clonal shifts.

7.3 T-PLL most closely resembles central memory T-cells as shown through a combination of immunophenotyping, GEP and mouse models

To determine which kind of T-cells T-PLL cells most closely resemble and which functional properties are retained, we used the same gene expression cohort as in Schrader, Crispatzu et al. and combined it with immunophenotyping data through our Semantic Web approach. Both ways of stratification revealed phenotypes of different T-cell differentiation stages. The two most common immunophenotypes, by manual gating and automatic, agglomerative clustering, seem to be a memory phenotype (57% of cases; with predominance of a $CD45RO^{+}$, $CCR7^{+}$ central memory (CM) pattern) and a transitional double-positive (CD45+CD45RA+) effector phenotype (30%).
In T-LGL (chronic lympho-proliferation of CD3+CD57+, activated effector cytotoxic T-lymphocytes) on contrary, a dominant cytotoxic CD8 phenotype is more prevalent than the CD4 one and is subdivided by different TCRs into subtypes CD8+/TCRαβ+, CD4+/TCRαβ+ and TCRγδ+. It is likely that all are chronically stimulated by different auto antigens leading to a survival advantage compared to other T-cells within the patient.
Chronic antigen stimulation is thought to be involved due to dermatitis in the development of CTCL (cutaneous T-cell lymphoma) as well. Malignant T-cells from leukemic CTCL patients were characterized also as CM (Campbell et al. 2010).
We further observed a reduction of TCR repertoire in T-PLL when compared to CD3+ pan T-cells, suggesting that after transformation only especially advantageous chains are selected and the repertoire is rendered monoclonal. It is unknown whether this monoclonality is static (dominant clone persists) or dynamic (dominant clone is overturned, but monoclonality persists) like in T-LGL (Clemente et al. 2013). T-ALL (precursor T acute lymphoblastic leukemia/lymphoma) RNA-Seq data is deposited within the ICGC (International Cancer Genome Consortium) repository, but access is so far further restricted by the submitter. Once obtained, we can compare the reconstructed TCR repertoire of naïve lymphocytes from very young patients with our antigen-experienced T-PLLs.

T-PLL cells may still harbor stem-cell like properties representing early differentiated progenitors with self-renewal capacity (Stemberger et al. 2009; Mueller et al. 2013), even though they do not behave like physiologic CM T-cells upon repetitive antigen stimulation, i.e. CD95 (Fas receptor) is downregulated and cannot react to apoptotic signals. Thus CM-like phenotype of T-PLL cells may represent the differentiation stage where oncogenic forces finally overthrow the homeostatic survival control, rather than T-PLL cells arising from CM cells.

Herling et al. 2008 previously elucidated that TCR-expressing T-PLL with higher *TCL1A* levels show a more robust growth in vitro over those cases with low *TCL1A* levels (associated with reduced TCR response). Here, we postulated *TCL1A* as a sensitizer to TCR signals by reducing the TCR activation threshold for self-antigens to be more efficiently ‘utilized'. Thus driving transition of affected naïve T-cells into an expanding T-memory pool as the origin of T-PLL outgrowth. This may only be required in early stages of leukemia development. Sequential analysis is then not fruitful comparing early and late samples in the clinical course, but rather between pre-clinical samples and those after leukemogenesis onset.

7.4 Semantic database enables exploratory survival analyses and meta-analyses (in lymphoid leukemias) to obtain novel aberration markers

An ideal gene-expression profiling protocol, including batch correction and admixture modeling, as well as classification algorithms, was constructed. We also presented a novel exploratory survival algorithm, with a cut-off still somewhat arbitrary and thus a need for more sophisticated learning algorithm for stratification limit. Still we were able to obtain reproducible gene signatures (CLL: *GPD1L*, *TNFSF12*, *JHDM1D*, *TBCD, AARS2*, *MTG1* & *TNIP* ; T-PLL: *RAB25* & *KIAA1211L*) linked to adverse prognosis in especially indolent and aggressive patient samples. These can be complementary to routinely tested markers (similar to Kienle et al. 2010), e.g. in CLL those from clinical chemistry, such as β2 microglobulin (Gentile et al. 2009) or from immunophenotyping, such as ZAP70 (Wiestner et al. 2003), and in T-PLL *TCL1A* and TCR expression (Herling et al. 2008) and relocation status, as well as *ATM* expression and copy-number status (see Schrader, Crispatzu et. al).
We further exemplified the machine-learning capabilities of R with input from the semantic framework. By means of SVM (support vector machines) we observed that *ATM* unmutated T-PLL samples are more likely to be biallelic for *ATM* and *AGO2*. Whereas in CLL, we found that the most informative variable for positive *TCL1A* status is unmutated *IGHV*, followed just then by *TCL1B* and previously calculated gene expression signature genes using decision trees.

7.5 Refinement of *TCL1A's* role in T-PLL

For the first time virtually every T-PLL case (95.2%) fulfilling the WHO classification criteria (Herling et al. 2004), demonstrated a genomic rearrangement involving a TCL1 gene and/or its overexpression (Schrader, Crispatzu et al. **Fig. S2d**). *TCL1A* augments signals from the most central growth receptor of T-cells, the TCR (Herling et al. 2008) perturbing a protective T-cell homeostasis (Newrzela et al. 2008), as confirmed in TCL1A-tg murine T-PLL. Only one out of 8 cases classified as TCL1A/t(X;14) double-negative (by protein and cytogenetics/WES) showed strictly no inv(14). Four of these cases, carrying GEPs, were associated with a consistent average upregulation of *TCL1B* (fc=1.41; p=0.0045) compared to CD3+ pan T-cells, but revealed gene expression profiles (GEPs) resembling those of 'conventional' *TCL1A*- or *MTCP1*-positive cases. *TCL6* is only slightly upregulated in T-PLL compared to normal CD3+ T-cells (FC=1.61, p=0.0172; q=0.0732), suggesting only a passenger role. Via RNA-Seq, we observed significantly upregulation of *TUNAR* (Tcl1 Upstream Neuron-Associated lincRNA) which is evolutionary conserved in vertebrates, as in mouse embryonic stem cells (mESCs) it was shown to be essential for pluripotency maintenance, while in zebrafish knockdown caused neurological dysfunction (Lin et al. 2014). Its role in T-PLL (and other lymphoid leukemias) remains to be determined, it however is only overexpressed in *TCL1A*+ T-PLLs (**Figure 7.3**).

Since, we observed an increase of *TCL1A* expression comparing early (FC=4.24; p=0.0877) and late follow-up T-PLL (FC=11.3; p=0.0258) samples vs. normal CD3+ T-cells, and a slight increase in the number of inv(14) breakpoints in WES, we postulate that *TCL1A* may be up-modulated with tumor progression as it is still required to uphold its genomic instability program (aneuploidy and telomere attrition) and growth-promoting effects in cooperation with, likely *TCL1B/TCL1A/MTCP1*-declining activity of, TCR (Warner, Oberbeck, Schrader et al.).

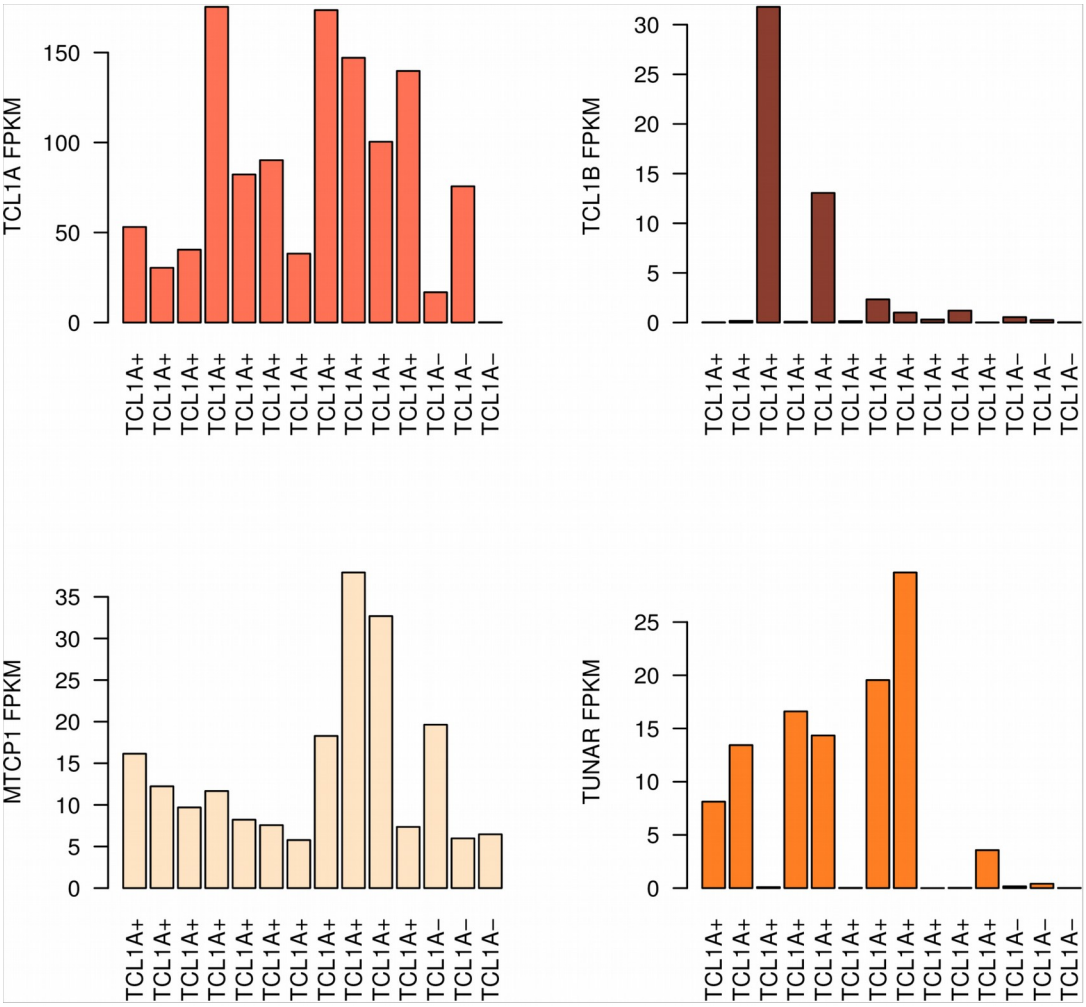


Figure 7.3: FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) values in 15 RNA-Seq T-PLL samples characterizing TCL1 gene family expression. *TCL1B*, as well as novel *TUNAR*, seem to be co-expressed with *TCL1A*. Whereas *MTCP1* is only highly expressed in MTCP1A+ cases, including TCL1A+/MTCP1+ double-positive cases.

To elucidate the molecular mechanisms central to *TCL1A*- T-PLLs, we tested which genes are differentially expressed when comparing T-PLL of different *TCL1A* protein expression levels. Therefore, we divided our samples into three distinct groups: *TCL1A*+ with high expression, *TCL1A*- with very little (should account for ~20% of patients) and *TCL1A* intermediate or “dim” which lie between them and for which FACS analysis is ambiguous. There was a considerable overlap (229/412 probes; e.g. *CTLA4* and *SLAMF6*, but not *MYC*, *ATM*, *RPRM* and *TNF*) of differentially expressed genes between *TCL1A* positive cases and those 6 carrying an *MTCP1*-activating t(X;14), against normal T-cells, respectively. This observation implicates a proliferative impact of constitutive MTCP1^{p13} comparable to *TCL1A*.
Similar to case studies **6.8** and **6.9**, we can query for deregulated TCL1 family members in RDF-converted GEP results and observe potential patterns.

While investigating the expression of *MTCP1*, we found a problem in nomenclatura. *CMC4* (Cx9C motif-containing protein 4; URL: http://www.uniprot.org/uniprot/P56277) is also called *MTCP1B* (MTCP1NB; p8; p8MTCP1) with two transcripts (both coding). While *MTCP1* (URL: http://www.uniprot.org/uniprot/P56278) is equivalent to *MTCP1A* (13;

p13MTCP1) with five transcripts (two of which are coding). Both have the same promoter, hence UniProt and GeneCards elucidates that they "... could be considered the same gene". In contrast to the outdated *TNG1* and *TNG2*, who were merged to *TCL6*, the *MTCP1* gene is split into two isoforms.

Only a few genes (mostly unannotated or pseudogenes) show exclusive gains in the *TCL1A*+ subgroup and exclusive losses in the *TCL1A*- ones or vice versa.*TCL1A* is neither significantly amplified or lost in all entities.

Interestingly, the 8 cases classified as *TCL1A/MTCP1*-negative (by mRNA and protein) showed a detection of an inv(14) in 7/8 cases including classical cytogenetics. Since further inversion breakpoints in the TCR*αδ* and *TCL1A* locus within exonic regions can be observed in 19/36 T-PLL WES cases (across all TCL1A expression statuses, including even *TCL1A*- cases), we propose that the inv(14) seems to be a necessary, but not sufficient condition for *TCL1A* activation in T-PLL. The t(14;14) is only present in one case and is called by *delly* not as a interchromosomal translocation, but rather a deletion of a large segment spanning from 14q11.2 and 14q32 (of one chr. 14) and a similar sized tandem duplication within the other chr. 14.
Since the number of supporting reads in WGS (including non-coding regions) and WES ranges from 10 to 20 and 2 to 34 respectively, as well as the maximum of other inversions/translocations being up to ~100, we conclude that inv(14)/t(14;14) may be only present in multiple subclones (max ~10-35%)**.** Different breakpoints may be due to (early) incorrect DDR of the founder clone within cell replication, while the subclonal status may be sufficient to drive tumorigenesis.
When screening for fusion transcripts within our RNA-Seq samples (n=15), one out of two *TCL1A*- (protein) cases exhibited a fusion-transcript of *TCL1A-TRAJ49* (in middle of the last *TCL1A* exon). The breakpoint can be seen in coverage drop when visualizing the WES and RNA-Seq sample and was confirmed with manually designed primers (Sanger). Interestingly the mRNA level of *TCL1A* is up-regulated compared to normal CD3+ T-cells (in GEP, as well as qRT-PCR), but seems to be degraded later on resulting in negative *TCL1A* protein status as measured by FACS.
The same can probably be observed in the other WES TCL1A- cases (with breakpoint in last *TCL1A* exon, but not RNA-sequenced). In P202_ (TCL1A- patient), the inversion likely results in the fusion-transcripts *TCL1A-TRAJ47* (breakpoint right at start of *TRAJ47*) and *TCL1A-TRAV38-2DV9* (breakpoint right after the end of *TRAV38-2DV9*), as well as the mRNA probe still being highly expressed. P1_/P1387_ also has breakpoints within the last *TCL1A* exon likely resulting in fusion-transcripts *TCL1A-TRAV26-2* (breakpoint after *TRAV26-2*) and *TCL1A-TRAJ10* (breakpoint in middle of *TRAJ10*).
The other *TCL1A*- case in RNA-seq (P1344_1347_) seems to have no breakpoint in *TCL1A*, but its mRNA is down-regulated in contrast to P1323_72909_.
We correlated (with Spearman) the expression of markers and breakpoint distance to *TCL1A* (also between each other) by queries to our semantic database. The “consensus” breakpoint was first averaged by mean and the distance was calculated to the *TCL1A* CDS (resulting in a clonal estimate). No correlation trend between breakpoint distance to *TCL1A* and *MTCP1*, *TCL1B* or *TCL6* expression was observed. However, there was a high mutual information (MI=0.502529; p'<0.002 ; 99.8%-Quantile=0.4667893 (with sampling)) between average breakpoint distance to *TCL1A* and *TCL1A* mRNA e xpression. A pattern can be specifically seen in those cases with breakpoints within and right upstream of the last *TCL1A* exon.
To observe the possible consequences of ‘enhancer hijacking’, as seen also in T-ALL

(inversion in chr. 14 also involving the TCR*αδ* locus), different B-cell lymphomas (translocations involving the *IGHV* locus) or just recently in medulloblastoma (translocation involving *GF1;* Northcott et al. 2014), we tried to correlate the number of juxtapositioned enhancer elements to the *TCL1A* locus with its respective expression value.
We therefore introduced different enhancer coordinates found within the FANTOM5 (FANTOM Consortium and the RIKEN PMI and CLST (DGT)) Phase2.0 project.

We further binned the juxtapositioned regions into 0.1 Mbp and counted the number of enhancers within these bins (taken from the FANTOM5 Phase2.0). Since P1331_ does not differ in counts within the last bin before *TCL1A* from P1323_72909_ (*TCL1A*- case) and has the same *TCL1A* breakpoint as P1323_72909_, we conclude that the breakpoint within the TCR*αδ* locus seems to be of key role here. Likely introducing an in-frame fusion-transcript of *TCL1A*-*TRAJ44*. Unfortunately this case has not been RNA-sequenced, so we have to design new primers for Sanger validation.
So a breakpoint in the last *TCL1A* exon seems to be necessary (P202_, P1323_72909_/P1346_, P1_/P1387_), but not sufficient condition (P1331_) for TCL1A silencing.
Translocation breakpoints in every of our 4 *MTCP1*+ WES cases, suggest that t(X;14) is a necessary and sufficient condition for *MTCP1* activation in T-PLL. Two out of 4 *MTCP1*+ cases further are characterized as *TCL1A*+, while the residual two are *TCL1A*-, suggesting no mutual exclusivity.
Since *TCL1A* has no targetable binding pockets and is thus hard to design an inhibitor for, disruption of its (rearranged TCR) enhancer complexes in T-PLL may be a potential target, as recently done by inhibiting *BRD4* and thus *MYC* in multiple melanomas (Lovén et al. 2013).

7.6 Refinement of *TCL1A's* role in CLL

To investigate the role of the TCL1 proto-oncogene family in between treatment of CLL, we compared 58 FCR (Fludarabine, Cyclophosphamide, Rituximab)-treated patients on Illumina HumanHT-12 v4.0 Expression BeadChips with available PFS data and different *TCL1A* statuses (by mRNA and immunohistochemistry (IHC)) through semantic integration.
Besides a linear gene-by-gene fit for each comparion's p-values, we further used the number of significantly deregulated probes (p-value<0.05 and q-value<0.1) and the *TCL1A* mRNA fold change itself as a metric to judge sample grouping. We found that TCL1A IHC status comparisons perform adverse to the *TCL1A* mRNA expression high (n=12) vs. low (n=12) comparison as it has the most significantly deregulated probes (n=263; |FC|>1.5). *TCL1A* is of course one of them (FC=6.18, p-value=5.560424e-08, q-value=n.s). CLL with *IGHV* mutated (n=17) vs. unmutated (n=31) returned 114 significantly deregulated probes. *TCL1A* is downregulated (FC=-1.98, p-value=0.002, q-value=2.084989e-03) as well.

PPI graphs and enrichments were calculated with STRINGdb10. As input we used the corresponding proteins of the most significantly deregulated probes in different gene expression comparisons. Grouping of gene expression samples and subsequent differential expression by immunohistochemistry status of TCL1A yielded no significantly PPI enrichments. However *TCL1A* mRNA expression high vs. low (http://string-db.org/10/p/55184371) and FCR-treated CLL with *IGHV* mutated vs. unmutated (http://string-db.org/10/p/14054372) did. The former comparison (n=93 proteins) yielded

117 PPI (27 more than expected; p-value: 0.004), while the latter (n=86 proteins) returned 107 PPI (19 more than expected: p-value: 0.029). Nodes are colored according to fold changes (red=upregulated, blue=downregulated). Edges are colored according to evidence level.
Grouping by *IGHV* mutation status and *TCL1A* mRNA expression results in significantly GO enrichments, including biological processes „leukocyte “ and „lymphocyte activation“. While IHC status comparisons yielding no significantly KEGG enrichments.

The meta-analysis of fold changes of known *TCL1A* interactors within different subgroups suggests regulatory resemblance of *IGHV* unmutated and *TCL1A* mRNA high expressing FCR-treated CLL subgroups. Known *TCL1A* interactors were extracted from the STRINGdb10 and their highest fold change was visualized in different comparisons (p-value<0.05; data not shown). Only *B4GALT2* is constantly downregulated in all three IHC comparisons. Interestingly both in *IGHV* unmutated vs. mutated and *TCL1A* mRNA expression high vs. low, the *TCL1A* protein interactors *L1TD1*, *SEPT10*, *TCL1B*, *XBP1* and *ZAP70* are upregulated, while *RHOH*, *HMGXB4* and *AKT3* are downregulated.

7.7 Semantic framework summary

The term "semantic framework" is not meant to describe a single program, but rather a collection of Semantic Web tools to analyse, convert and combine data in the most automatically and consistent manner. The different case studies in **Chapter 6** elucidate this approach and enable the reader to apply mentioned semantic tools to his/her data. Overall, in contrast to other database schemas (**Table 7.1**), the Semantic Web approach made it possible to combine the analyses of such heterogeneous data within all these different projects (Schrader, Crispatzu et al.; Warner, Oberbeck, Schrader et al.; Crispatzu et al. 2016; Crispatzu, Kulkarni et al.). This generated novel hypothesis which were further functionally validated. It can further serve as a starting point in finding context-specific information in multigene regulatory networks, which can be overlapped and the consensus can be visualized.

	RDF-based	SQL-based	Other NoSQL
Data model	Graph- and XML-oriented	Relational	Document-, column- or key-value-oriented
Ontology-based?	Yes	No	Rarely
Scalability & Performance	Medium, but increasing	Low	High
Self-descriptive flat file	Yes	No	Rarely
Querying of databases	Distributed querying	Only one database at a time	Only one database at a time
Possible to concatenate databases?	Yes, may need semi-automated ontology-matching	Yes, but only by manual conversion and linking through primary keys	Yes, but only by manual conversion
Distribution	Medium, but increasing	High, but decreasing	Medium, but increasing
Administration tools	Still under extensive development, e.g. OpenRDF Sesame & Jena	Highly developed, e.g. MySQL & phpMyAdmin	Highly developed, e.g. MongoDb

Table 7.1: Illustrative comparison between common database schemas.

7.8 Semantic framework outlook

One disadvantage of the current models (described in **Chapter 6**) remains the lack of generality to improve distributed querying and overlaps with foreign data sources. However, the more (diverse) data is integrated, the more the underlying vocabulary is tested and expanded through new terms of established ontologies.
As the number of triples increases, it may further be mandatory to set up more sophisticated server solutions. Triples stored in RDF can be converted and further processed into the *Hadoop* framework, thus allowing large-scale computations on cloud-computing architectures like *Amazon Web Servers* (AWS). Our current solution is running a virtual machine hosting a Semantic Web server called from HPC resources (all from our local computing center). Statistical programming can also be upscaled by using Big Data-tailored products like *R Enterprise*.

Acknowledgements

I would like to thank my supervisors, Dr. Marco Herling, Prof. Michael Nothnagel and Dr. Peter Frommolt, as well as my colleagues, especially Dr. Alexandra Schrader and Petra Mayer, and my family, especially my uncle Dr. Dieter Wagner and my mother.

I gratefully acknowledge all contributing centers and investigators enrolling patients into the trials and registry of the German CLL Study Group (GCLLSG) and at the UT M.D. Anderson Cancer Center (MDACC), Houston/TX, USA; the GCLLSG and UT MDACC staff and the patients with their families for their invaluable contributions.

I furthermore thank the Cancer Genome Atlas (TCGA) research network (http://cancergenome.nih.gov) and the International Cancer Genome Consortium (ICGC; http://dcc.icgc.org) for supplying large amounts of NGS data, the BioStars online community (http://www.biostars.org) for their documentation and help regarding current Bioinformatics tools, as well as the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded High Performance Computing (HPC) system CHEOPS as well as support.

References

- Aggarwal M, Villuendas R, Gomez G, Rodriguez-Pinilla SM, Sanchez-Beato M, Alvarez D, Martinez N, Rodriguez A, Castillo ME, Camacho FI, Montes-Moreno S, Garcia-Marco JA, Kimby E, Pisano DG, Piris MA. TCL1A expression delineates biological and clinical variability in B-cell lymphoma. *Mod Pathol.* 2009 Feb;**22**(2):206-15.
- Auguin D, Barthe P, Royer C, Stern MH, Noguchi M, Arold ST, Roumestand C. Structural basis for the co-activation of protein kinase B by T-cell leukemia-1 (TCL1) family proto-oncoproteins. *J Biol Chem.* 2004 Aug 20;**279**(34):35890-902.
- Albert I, Thakar J, Li S, Zhang R, Albert R. Boolean network simulations for life scientists. *Source Code Biol Med.* 2008 Nov 14;**3**:16.
- Alexandrov LB, et al. Signatures of mutational processes in human cancer. *Nature*. 2013 Aug 22;**500**(7463):415-21.
- An X, Tiwari AK, Sun Y, Ding PR, Ashby CR Jr, Chen ZS. BCR-ABL tyrosine kinase inhibitors in the treatment of Philadelphia chromosome positive chronic myeloid leukemia: a review. *Leuk Res.* 2010 Oct;**34**(10):1255-68.
- Austen B, Skowronska A, Baker C, Powell JE, Gardiner A, Oscier D, Majid A, Dyer M, Siebert R, Taylor AM, Moss PA, Stankovic T. Mutation status of the residual ATM allele is an important determinant of the cellular response to chemotherapy and survival in patients with chronic lymphocytic leukemia containing an 11q deletion. *J Clin Oncol.* 2007 Dec 1;**25**(34):5448-57.
- Badve S, Collins NR, Bhat-Nakshatri P, Turbin D, Leung S, Thorat M, Dunn SE, Geistlinger TR, Carroll JS, Brown M, Bose S, Teitell MA, Nakshatri H. Subcellular localization of activated AKT in estrogen receptor- and progesterone receptor-expressing breast cancers: potential clinical implications. *Am J Pathol.* 2010 May;**176**(5):2139-49.
- Baerlocher GM, Vulto I, de Jong G, Lansdorp PM. Flow cytometry and FISH to measure the average length of telomeres (flow FISH). *Nat Protoc.* 2006;**1**(5):2365-76.
- Barrett, T. et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991–5 (2013).
- Bassaganyas L, Beà S, Escaramís G, Tornador C, Salaverria I, Zapata L, Drechsel O, Ferreira PG, Rodriguez-Santiago B, Tubio JM, Navarro A, Martín-García D, López C, Martínez-Trillos A, López-Guillermo A, Gut M, Ossowski S, López-Otín C, Campo E, Estivill X. Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis. *Leukemia.* 2013 Dec;**27**(12):2376-9.
- Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform.* 2008 Oct;**41**(5):706-16.
- Bernal C, Aguayo F, Villarroel C, Vargas M, Díaz I, Ossandon FJ, Santibáñez E, Palma M, Aravena E, Barrientos C, Corvalan AH. Reprimo as a potential biomarker for early detection in gastric cancer. *Clin Cancer Res.* 2008 Oct 1;**14**(19):6264-9.
- Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV, Britanova OV, Chudakov DM. MiTCR: software for T-cell receptor sequencing data analysis. *Nat Methods.* 2013 Sep;**10**(9):813-4.
- Campbell JJ, Clark RA, Watanabe R, Kupper TS. Sezary syndrome and mycosis fungoides arise from distinct T-cell subsets: a biologic rationale for their distinct clinical behaviors. *Blood.* 2010 Aug 5;**116**(5):767-71.
- Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of

adult de novo acute myeloid leukemia. *N Engl J Med.* 2013 May 30;**368**(22):2059-74.
- Cardinaud B, Moreilhon C, Marcet B, Robbe-Sermesant K, LeBrigand K, Mari B, Eclache V, Cymbalista F, Raynaud S, Barbry P. miR-34b/miR-34c: a regulator of TCL1 expression in 11q- chronic lymphocytic leukaemia? *Leukemia*. 2009 Nov;**23**(11):2174-7.
- Cha TL, Zhou BP, Xia W, Wu Y, Yang CC, Chen CT, Ping B, Otte AP, Hung MC. Akt-mediated phosphorylation of EZH2 suppresses methylation of lysine 27 in histone H3. *Science*. 2005 Oct 14;**310**(5746):306-10.
- Chen SS, Raval A, Johnson AJ, Hertlein E, Liu TH, Jin VX, Sherman MH, Liu SJ, Dawson DW, Williams KE, Lanasa M, Liyanarachchi S, Lin TS, Marcucci G, Pekarsky Y, Davuluri R, Croce CM, Guttridge DC, Teitell MA, Byrd JC, Plass C. Epigenetic changes during disease progression in a murine model of human chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A.* 2009 Aug 11;**106**(32):13433-8.
- Cheung KH, Yip KY, Townsend JP, Scotch M. HCLS 2.0/3.0: health care and life sciences data mashup using Web 2.0/3.0. *J Biomed Inform.* 2008 Oct;**41**(5):694-705.
- Clemente MJ, Przychodzen B, Jerez A, Dienes BE, Afable MG, Husseinzadeh H, Rajala HL, Wlodarski MW, Mustjoki S, Maciejewski JP. Deep sequencing of the T-cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes. *Blood*. 2013 Dec 12;**122**(25):4077-85.
- Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, Fostel JL, Friedrich DC, Perrin D, Dionne D, Kim S, Gabriel SB, Lander ES, Fisher S, Getz G. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* 2013 Apr 1;**41**(6):e67.
- Crispatzu G et al. A Critical Evaluation of Analytic Aspects of Gene Expression Profiling in Lymphoid Leukemias with Broad Applications to Cancer Genomics. *AIMS Medical Science,* **3** (3) : 248–271.
- Crispatzu G, Kulkarni P, et al. Semi-automated cancer genome analysis using high-performance computing (accepted with major revisions; *Human Mutation*)
- D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.* 2016 Aug 9.
- Davidson NM, Majewski IJ, Oshlack A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Med.* 2015 May 11;**7**(1):43.
- Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 2012 Aug;**22**(8):1589-98.
- Demir E, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol. 2010 Sep;**28**(9):935-42. Erratum in: *Nat Biotechnol.* 2010 Dec;**28**(12):1308. *Nat Biotechnol.* 2012 Apr;**30**(4):365.
- Dengel A (published by). Semantische Technologien. Grundlagen - Konzepte – Anwendungen. *Spektrum Akademischer Verlag*, 2012. ISBN 978-3-8274-2663-5
- Deng Y, Chan SS, Chang S. Telomere dysfunction and tumour suppression: the senescence connection. *Nat Rev Cancer.* 2008 Jun;**8**(6):450-8.
- Deus HF et al. A Semantic Web management model for integrative biomedical informatics. *PLoS One.* 2008 Aug 13;**3**(8):e2946.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013

Jan 1;**29**(1):15-21.
- Döhner H, Stilgenbauer S, Benner A, Leupolt E, Kröber A, Bullinger L, Döhner K, Bentz M, Lichter P. Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J Med.* 2000 Dec 28;**343**(26):1910-6.
- Dürig J, Bug S, Klein-Hitpass L, Boes T, Jöns T, Martin-Subero JI, Harder L, Baudis M, Dührsen U, Siebert R. Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia*. 2007 Oct;**21**(10):2153-63.
- Ellinger J, Bastian PJ, Jurgan T, Biermann K, Kahl P, Heukamp LC, Wernert N, Müller SC, von Ruecker A. CpG island hypermethylation at multiple gene sites in diagnosis and prognosis of prostate cancer. *Urology*. 2008 Jan;**71**(1):161-7.
- FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature*. 2014 Mar 27;**507**(7493):462-70.
- Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput Biol.* 2005 Oct**;1**(5):e50.
- Gabellini C, Antonelli A, Petrinelli P, Biroccio A, Marcucci L, Nigro G, Russo G, Zupi G, Elli R. Telomerase activity, apoptosis and cell cycle progression in ataxia telangiectasia lymphocytes expressing TCL1. *Br J Cancer.* 2003 Sep 15;**89**(6):1091-5.
- Garding A, Bhattacharya N, Haebe S, Müller F, Weichenhan D, Idler I, Ickstadt K, Stilgenbauer S, Mertens D. TCL1A and ATM are co-expressed in chronic lymphocytic leukemia cells without deletion of 11q. *Haematologica.* 2013 Feb;**98**(2):269-73.
- Gattinoni L, Zhong XS, Palmer DC, Ji Y, Hinrichs CS, Yu Z, Wrzesinski C, Boni A, Cassard L, Garvin LM, Paulos CM, Muranski P, Restifo NP. Wnt signaling arrests effector T cell differentiation and generates CD8+ memory stem cells. *Nat Med.* 2009 Jul;**15**(7):808-13.
- Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012 Jan;**40**(Database issue):D1100-7.
- Gentile M, Cutrona G, Neri A, Molica S, Ferrarini M, Morabito F. Predictive value of beta2-microglobulin (beta2-m) levels in chronic lymphocytic leukemia since Binet A stages. *Haematologica*. 2009 Jun;**94**(6):887-8.
- Gonzalez-Vasconcellos I, Anastasov N, Sanli-Bonazzi B, Klymenko O, Atkinson MJ, Rosemann M. Rb1 haploinsufficiency promotes telomere attrition and radiation-induced genomic instability. *Cancer Res.* 2013 Jul 15;**73**(14):4247-55.
- Gribben JG. How I treat CLL up front. *Blood*. 2010 Jan 14;**115**(2):187-97.
- Gualco G et al. T-cell leukemia 1 expression in nodal Epstein-Barr virus-negative diffuse large B-cell lymphoma and primary mediastinal B-cell lymphoma. *Hum Pathol.* 2010 Sep;**41**(9):1238-44.
- Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* 2012 Jun;**22**(6):1154-62.
- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011 Mar 4;**144**(5):646-74.
- Hashimoto M, Suizu F, Tokuyama W, Noguchi H, Hirata N, Matsuda-Lennikov M, Edamura T, Masuzawa M, Gotoh N, Tanaka S, Noguchi M. Protooncogene TCL1b functions as an Akt kinase co-activator that exhibits oncogenic potency in vivo.

Oncogenesis. 2013 Sep 16;**2**:e70.
- Hendler J. Communication. Science and the semantic web. *Science*. 2003 Jan 24;**299**(5606):520-1.
- Herling M, Khoury JD, Washington LT, Duvic M, Keating MJ, Jones D. A systematic approach to diagnosis of mature T-cell leukemias reveals heterogeneity among WHO categories. *Blood*. 2004 Jul 15;**104**(2):328-35.
- Herling M, Patel KA, Teitell MA, Konopleva M, Ravandi F, Kobayashi R, Jones D. High TCL1 expression and intact T-cell receptor signaling define a hyperproliferative subset of T-cell prolymphocytic leukemia. *Blood*. 2008 Jan 1;**111**(1):328-37.
- Herling M, Patel KA, Weit N, Lilienthal N, Hallek M, Keating MJ, Jones D. High TCL1 levels are a marker of B-cell receptor pathway responsiveness and adverse outcome in chronic lymphocytic leukemia. *Blood*. 2009 Nov 19;**114**(21):4675-86.
- Hodgkin PD, Heath WR, Baxter AG. The clonal selection theory: 50 years since the revolution. *Nat Immunol.* 2007 Oct;**8**(10):1019-26.
- Hopfinger G, Weit N, Herling M. Diagnostische Schritte und Therapieoptionen bei peripheren T-Zell-Neoplasien. *Wiener Klinische Wochenschrift* 2009;**4**:165-76
- Hu T, Liu S, Breiter DR, Wang F, Tang Y, Sun S. Octamer 4 small interfering RNA results in cancer stem cell-like cell apoptosis. *Cancer Res*. 2008 Aug 15;**68**(16):6533-40.
- Iqbal J, Weisenburger DD, Chowdhury A, Tsai MY et al. Natural killer cell lymphoma shares strikingly similar molecular features with a group of non-hepatosplenic γδ T-cell lymphoma and is highly sensitive to a novel aurora kinase A inhibitor in vitro. *Leukemia* 2011 Feb;**25**(2):348-58.
- Jaffe ES. The 2008 WHO classification of lymphomas: implications for clinical practice and translational research. *Hematology Am Soc Hematol Educ Program.* 2009:523-31.
- Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*. 2014 Feb 14;**343**(6172):776-9.
- Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A, Gehant S, Laibe C, Redaschi N, Wimalaratne SM, Martin M, Le Novère N, Parkinson H, Birney E, Jenkinson AM. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*. 2014 May 1;**30**(9):1338-9.
- Kaelin WG Jr. The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer*. 2005 Sep;**5**(9):689-98.
- Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res.* 2009 Jan;**37**(Database issue):D623-8.
- Kienle D, Benner A, Läufle C, Winkler D, Schneider C, Bühler A, Zenz T, Habermann A, Jäger U, Lichter P, Dalla-Favera R, Döhner H, Stilgenbauer S. Gene expression factors as predictors of genetic risk and survival in chronic lymphocytic leukemia. *Haematologica*. 2010 Jan;**95**(1):102-9.
- Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* 2011 Aug 11;**12**(8):R72.
- Klein U, Dalla-Favera R. New insights into the pathogenesis of chronic lymphocytic leukemia. *Semin Cancer Biol.* 2010 Dec;**20**(6):377-83.
- Klein U, Lia M, Crespo M, Siegel R, Shen Q, Mo T, Ambesi-Impiombato A, Califano A, Migliazza A, Bhagat G, Dalla-Favera R. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell.* 2010 Jan 19;**17**(1):28-40.

- Koskela HL, Eldfors S, Ellonen P, van Adrichem AJ, Kuusanmäki H, Andersson EI, Lagström S, Clemente MJ, Olson T, Jalkanen SE, Majumder MM, Almusa H, Edgren H, Lepistö M, Mattila P, Guinta K, Koistinen P, Kuittinen T, Penttinen K, Parsons A, Knowles J, Saarela J, Wennerberg K, Kallioniemi O, Porkka K, Loughran TP Jr, Heckman CA, Maciejewski JP, Mustjoki S. Somatic STAT3 mutations in large granular lymphocytic leukemia. *N Engl J Med.* 2012 May 17;**366**(20):1905-13.
- Kreso A, Dick JE. Evolution of the cancer stem cell model. *Cell Stem Cell*. 2014 Mar 6;**14**(3):275-91.
- Kriss CL, Pinilla-Ibarz JA, Mailloux AW, Powers JJ, Tang CH, Kang CW, Zanesi N, Epling-Burnette PK, Sotomayor EM, Croce CM, Del Valle JR, Hu CC. Overexpression of TCL1 activates the endoplasmic reticulum stress response: a novel mechanism of leukemic progression in mice. *Blood*. 2012 Aug 2;**120**(5):1027-38.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009 Sep;**19(**9):1639-45.
- Kutmon M, Kelder T, Mandaviya P, Evelo CT, Coort SL.CyTargetLinker: a cytoscape app to integrate regulatory interactions in network analysis. *PLoS One.* 2013 Dec 5;**8**(12):e82160.
- Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*. 2013 Feb 14;**152**(4):714-26.
- Landau DA, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature*. 2015 Oct 22;**526**(7574):525-30
- Lau SK, Weiss LM, Chu PG. TCL1 protein expression in testicular germ cell tumors. *Am J Clin Pathol.* 2010 May;**133**(5):762-6.
- Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013 Jul 11;**499**(7457):214-8.
- Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics*. 2013 Mar 1;**29**(5):649-51.
- Lin N, Chang KY, Li Z, Gates K, Rana ZA, Dang J, Zhang D, Han T, Yang CS, Cunningham TJ, Head SR, Duester G, Dong PD, Rana TM. An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Mol Cell.* 2014 Mar 20;**53**(6):1005-19.
- Liu S, Tsai WH, Ding Y, Chen R, Fang Z, Huo Z, Kim S, Ma T, Chang TY, Priedigkeit NM, Lee AV, Luo J, Wang HW, Chung IF, Tseng GC. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res.* 2016 Mar 18;**44**(5):e47.
- Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*. 2013 Apr 11;**153**(2):320-34.
- Luo J, Solimini NL, Elledge SJ. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell*. 2009 Mar 6;**136**(5):823-37.
- Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*. 2010 Apr 15;**26**(8):1112-8.
- Mansouri MR, Sevov M, Aleskog A, Jondal M, Merup M, Sundström C, Osorio L, Rosenquist R. IGHV3-21 gene usage is associated with high TCL1 expression in chronic lymphocytic leukemia. *Eur J Haematol.* 2010 Feb 1;**84**(2):109-16.
- Matthias Dehmer M, Emmert-Streib F, Graber A, Salvador A. Applied Statistics for

Network Biology: Methods in Systems Biology. *Wiley-Blackwell,* April 2011
- Molica S, Alberti A. Prognostic value of the lymphocyte doubling time in chronic lymphocytic leukemia. *Cancer*. 1987 Dec 1;**60**(11):2712-6.
- Morandell S, Reinhardt HC, Cannell IG, Kim JS, Ruf DM, Mitra T, Couvillon AD, Jacks T, Yaffe MB. A reversible gene-targeting strategy identifies synthetic lethal interactions between MK2 and p53 in the DNA damage response in vivo. *Cell Rep.* 2013 Nov 27;**5**(4):868-77.
- Mraz M, Malinova K, Kotaskova J, Pavlova S, Tichy B, Malcikova J, Stano Kozubik K, Smardova J, Brychtova Y, Doubek M, Trbusek M, Mayer J, Pospisilova S. miR-34a, miR-29c and miR-17-5p are downregulated in CLL patients with TP53 abnormalities. *Leukemia*. 2009 Jun;**23**(6):1159-63.
- Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T cell subsets, migration patterns, and tissue residence. *Annu Rev Immunol.* 2013;**31**:137-61.
- Murphy MP, Siegel RM. Mitochondrial ROS fire up T cell activation. *Immunity*. 2013 Feb 21;**38**(2):201-2.
- Nam JY, Kim NK, Kim SC, Joung JG, Xi R, Lee S, Park PJ, Park WY. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Brief Bioinform.* 2016 Mar;**17**(2):185-92
- Newrzela S, Cornils K, Li Z, Baum C, Brugman MH, Hartmann M, Meyer J, Hartmann S, Hansmann ML, Fehse B, von Laer D. Resistance of mature T cells to oncogene transformation. *Blood*. 2008 Sep 15;**112**(6):2278-86.
- Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics*. 2014 Apr 1;**30**(7):1015-6.
- Northcott PA, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*. 2014 Jul 24;**511**(7510):428-34.
- Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 2009 Jul;**37**(Web Server issue):W170-3.
- Ooki A, Yamashita K, Yamaguchi K, Mondal A, Nishimiya H, Watanabe M. DNA damage-inducible gene, reprimo functions as a tumor suppressor and is suppressed by promoter methylation in gastric cancer. *Mol Cancer Res.* 2013 Nov;**11**(11):1362-74.
- Palamarchuk A, Yan PS, Zanesi N, Wang L, Rodrigues B, Murphy M, Balatti V, Bottoni A, Nazaryan N, Alder H, Rassenti L, Kipps TJ, Freitas M, Croce CM, Pekarsky Y. Tcl1 protein functions as an inhibitor of de novo DNA methylation in B-cell chronic lymphocytic leukemia (CLL). *Proc Natl Acad Sci U S A.* 2012 Feb 14;**109**(7):2555-60.
- Pekarsky Y, Hallas C, Croce CM. The role of TCL1 in human T-cell leukemia. *Oncogene*. 2001 Sep 10;**20**(40):5638-43.
- Pekarsky Y, Santanam U, Cimmino A, Palamarchuk A, Efanov A, Maximov V, Volinia S, Alder H, Liu CG, Rassenti L, Calin GA, Hagan JP, Kipps T, Croce CM. Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. *Cancer Res.* 2006 Dec 15;**66**(24):11590-3.
- Petrinelli, P. et al. Telomeric associations and chromosome instability in ataxia telangiectasia T cells characterized by TCL1 expression. *Cancer Genet Cytogenet.* 2001 Feb;**125**(1):46-51.
- Petryszak R et al. Expression Atlas update-an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* 2016 Jan 4;**44**(D1):D746-52.XXXXXXX

- Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, Kumararatne D, Doffinger R, Nejentsev S. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. 2012 Nov 1;**28**(21):2747-54.
- Poulaki V, Mitsiades N, Romero ME, Tsokos M. Fas-mediated apoptosis in neuroblastoma requires mitochondrial activation and is inhibited by FLICE inhibitor protein and Bcl-2. *Cancer Res.* 2001 Jun 15;**61**(12):4864-72.
- Prinz C, Vasyutina E, Lohmann G, Schrader A, Romanski S, Hirschhäuser C, Mayer P, Frias C, Herling CD, Hallek M, Schmalz HG, Prokop A, Mougiakakos D, Herling M. Organometallic nucleosides induce non-classical leukemic cell death that is mitochondrial-ROS dependent and facilitated by TCL1-oncogene burden. *Mol Cancer.* 2015 Jun 4;**14**:114.
- Puente XS, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2011 Jun 5;**475**(7354):101-5.
- Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol.* 2011 Oct 2;**29**(10):886-91.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org/.
- Riabinska A, Daheim M, Herter-Sprie GS, Winkler J, Fritz C, Hallek M, Thomas RK, Kreuzer KA, Frenzel LP, Monfared P, Martins-Boucas J, Chen S, Reinhardt HC. Therapeutic targeting of a robust non-oncogene addiction to PRKDC in ATM-defective tumors. *Sci Transl Med.* 2013 Jun 12;**5**(189):189ra78.
- Röth A, de Beer D, Nückel H, Sellmann L, Dührsen U, Dürig J, Baerlocher GM. Significantly shorter telomeres in T-cells of patients with ZAP-70+/CD38+ chronic lymphocytic leukaemia. *Br J Haematol.* 2008 Nov;**143**(3):383-6.
- Röth A, Dürig J, Himmelreich H, Bug S, Siebert R, Dührsen U, Lansdorp PM, Baerlocher GM. Short telomeres and high telomerase activity in T-cell prolymphocytic leukemia. *Leukemia*. 2007 Dec;**21**(12):2456-62.
- Schrader A, Crispatzu G et al. Integrated genetic profiles of T-PLL implicate a TCL1/ATM-centered model of aberrant, but actionable damage responses (in review; *Cancer Discovery*)
- Shankar S, Pitchiaya S, Malik R, Kothari V, Hosono Y, Yocum AK, Gundlapalli H, White Y, Firestone A, Cao X, Dhanasekaran SM, Stuckey JA, Bollag G, Shannon K, Walter NG, Kumar-Sinha C, Chinnaiyan AM. KRAS Engages AGO2 to Enhance Cellular Transformation. *Cell Rep.* 2016 Feb 16;**14**(6):1448-61.
- Shibata D, Navidi W, Salovaara R, Li ZH, Aaltonen LA. Somatic microsatellite mutations as molecular tumor clocks. *Nat Med.* 1996 Jun;**2**(6):676-81.
- Sivina M, Hartmann E, Vasyutina E, Boucas JM, Breuer A, Keating MJ, Wierda WG, Rosenwald A, Herling M, Burger JA. Stromal cells modulate TCL1 expression, interacting AP-1 components and TCL1-targeting micro-RNAs in chronic lymphocytic leukemia. *Leukemia*. 2012 Aug;**26**(8):1812-20.
- Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, Marjoram P, Siegmund K, Press MF, Shibata D, Curtis C. A Big Bang model of human colorectal tumor growth. *Nat Genet.* 2015 Mar;**47**(3):209-16.
- Spalding KL, Bhardwaj RD, Buchholz BA, Druid H, Frisén J. Retrospective birth dating of cells in humans. *Cell*. 2005 Jul 15;**122**(1):133-43.
- Splendiani A. RDFScape: Semantic Web meets systems biology. *BMC Bioinformatics*. 2008 Apr 25;**9** Suppl 4:S6

- Stemberger C, Neuenhahn M, Gebhardt FE, Schiemann M, Buchholz VR, Busch DH. Stem cell-like plasticity of naïve and distinct memory CD8+ T cell subsets. *Semin Immunol.* 2009 Apr;**21**(2):62-8.
- Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*. 2015 May 7;**521**(7550):81-4.
- Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015 Jan;**43**(Database issue):D447-52.
- Teitell, MA. The TCL1 family of oncoproteins: co-activators of transformation. *Nat Rev Cancer.* 2005 Aug;**5**(8):640-8.
- Vasyutina E, Boucas JM, Bloehdorn J, Aszyk C, Crispatzu G, Stiefelhagen M, Breuer A, Mayer P, Lengerke C, Döhner H, Beutner D, Rosenwald A, Stilgenbauer S, Hallek M, Benner A, Herling M. The regulatory interaction of EVI1 with the TCL1A oncogene impacts cell survival and clinical outcome in CLL. *Leukemia*. 2015 Oct;**29**(10):2003-14.
- Verdun RE, Karlseder J. Replication and protection of telomeres. *Nature*. 2007 Jun 21;**447**(7147):924-31. Review.
- Völler D, Reinders J, Meister G, Bosserhoff AK. Strong reduction of AGO2 expression in melanoma and cellular consequences. *Br J Cancer.* 2013 Dec 10;**109**(12):3116-24.
- Wagle P, Nikolić M, Frommolt P. QuickNGS elevates Next-Generation Sequencing data analysis to a new level of automation. BMC Genomics. 2015 Jul 1;**16**:487.
- Wang E. Cancer Systems Biology. *CRC Press*, May 4, 2010
- Warner K, Oberbeck S, Schrader A, Crispatzu G. et al. Aberrant effector functions of the memory-type T-PLL cell imply a leukemogenic cooperation of TCL1A with TCR signaling (in review; *Blood*)
- Wasserstrom A, Frumkin D, Adar R, Itzkovitz S, Stern T, Kaplan S, Shefer G, Shur I, Zangi L, Reizel Y, Harmelin A, Dor Y, Dekel N, Reisner Y, Benayahu D, Tzahor E, Segal E, Shapiro E. Estimating cell depth from somatic mutations. *PLoS Comput Biol.* 2008 May 9;**4**(4):e1000058.
- Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* 2011 Jul;**39**(Web Server issue):W541-5.
- Wiestner A, Rosenwald A, Barry TS, Wright G, Davis RE, Henrickson SE, Zhao H, Ibbotson RE, Orchard JA, Davis Z, Stetler-Stevenson M, Raffeld M, Arthur DC, Marti GE, Wilson WH, Hamblin TJ, Oscier DG, Staudt LM. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood*. 2003 Jun 15;**101**(12):4944-51.
- Williams MS, Kwon J. T cell receptor stimulation, reactive oxygen species, and cell signaling. *Free Radic Biol Med.* 2004 Oct 15;**37**(8):1144-51. Review.
- Witte S, O'Shea JJ, Vahedi G. Super-enhancers: Asset management in immune cell genomes. *Trends Immunol.* 2015 Sep;**36**(9):519-26.
- Yan J, Li B, Lin B, Lee PT, Chung TH, Tan J, Bi C, Lee XT, Selvarajan V, Ng SB, Yang H, Yu Q, Chng WJ. EZH2 phosphorylation by JAK3 mediates a switch to noncanonical function in natural killer/T-cell lymphoma. *Blood*. 2016 Aug 18;**128**(7):948-58.
- Zakov S, Kinsella M, Bafna V. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc Natl Acad Sci U S A.* 2013 Apr 2;**110**(14):5546-

51.
- Zhang CZ, Leibowitz ML, Pellman D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes Dev.* 2013 Dec 1;**27**(23):2513-30.
- Zhang J, Jin H, Liu H, Lv S, Wang B, Wang R, Liu H, Ding M, Yang Y, Li L, Zhang J, Fu S, Xie D, Wu M, Zhou W, Qian Q. MiRNA-99a directly regulates AGO2 through translational repression in hepatocellular carcinoma. *Oncogenesis*. 2014 Apr 14;**3**:e97.
- Zhang Y, Joe G, Hexner E, Zhu J, Emerson SG. Host-reactive CD8+ memory stem cells in graft-versus-host disease. *Nat Med.* 2005 Dec;**11**(12):1299-305.

Abstract

Within this thesis I developed a new approach for the analysis and integration of heterogeneous leukemic data sets applicable to any high-throughput analysis including basic research. All layers are stored in a semantic graph which facilitates modifications by just adding edges (relationships/attributes) and nodes (values/results) as well as calculating biological consensus and clinical correlation. The front-end is accessible through a GUI (graphical user interface) on a Java-based Semantic Web server. I used this framework to describe the genomic landscape of T-PLL (T-cell prolymphocytic leukemia), which is a rare (~0.6/million) mature T-cell malignancy with aggressive clinical course, notorious treatment resistance, and generally low overall survival.
We have conducted gene expression and copy-number profiling as well as NGS (next-generation sequencing) analyses on a cohort comprising 94 T-PLL cases. *TCL1A* (T-cell leukemia/lymphoma 1A) overexpression and *ATM* (Ataxia Telangiectasia Mutated) impairment represent central hallmarks of T-PLL, predictive for patient survival, T-cell function and proper DNA damage responses. We identified new chromosomal lesions, including a gain of *AGO2* (Argonaute 2, RISC Catalytic Component; 57.14% of cases), which is decisive for the chromosome 8q lesion. While we found significant enrichments of truncating mutations in *ATM* mut/no del (p=0.01365), as well as *FAT* (FAT Atypical Cadherin) domain mutations in *ATM* mut/del (p=0.01156), *JAK3* (Janus Kinase 3) mut/*ATM* del cases may represent another tumor lineage. Using whole-transcriptome sequencing, we identified novel structural variants affecting chromosome 14 that lead to the expression of a TCL1A-TCR (T-cell receptor) fusion transcript and a likely degradated *TCL1A* protein. Two clustering approaches of normal T-cell subsets vs. leukemia gene expression profiles, as well as immunophenotyping-based agglomerative clustering and TCR repertoire reconstruction further revealed a restricted, memory-like T-cell phenotype. This is to date the most comprehensive, multi-level, integrative study on T-PLL and it led to an evolutionary disease model and a *histone deacetylase*-inhibiting / double strand break-inducing treatment that performs better than the current standard of chemoimmunotherapy in preclinical testing.

Zusammenfassung

In dieser Dissertation habe ich eine neue Herangehensweise entwickelt, welche die Analyse und Integration von heterogenen Leukämiedatensätzen erleichtert, sowie anwendbar auf eine Vielzahl hochdurchsatzbasierter Grundlagenexperimente ist. Alle Datenschichten werden in einem semantischem Graphen gespeichert, was wiederum Änderungen in Form des Hinzufügens von Kanten (Beziehungen/Attribute) und Knoten (Werte/Resultate) möglich macht, sowie generell das Errechnen von biologischem Konsens und klinischer Korrelationen. Das System ist erreichbar durch eine graphische Benutzeroberfläche auf einem Java-basiertem Semantic Web-Server. Ich nutzte das Rahmenprogramm weiterhin zum Beschreiben der genomischen Landschaft der T-PLL (T-Zell prolymphozytische Leukämie), einer seltenen (~0.6/Millionen) Erkrankung reifer T-Zellen mit aggressivem klinischem Verlauf, notorischer Behandlungsresistenz und generell niedriger Überlebensrate.
Wir erstellten Genexpressions- und Kopiernummer-Profile, sowie NGS (next-generation sequencing) innerhalb einer Kohorte von 94 T-PLL Patienten. *TCL1A* (T-cell leukemia/lymphoma 1A) Überexpression und *ATM* (Ataxia Telangiectasia Mutated) Beeinträchtigung repräsentieren zentrale Charakteristiken der T-PLL, prädiktiv für das Überleben des Patienten, T-Zell-Funktion und reibungsloses Antworten auf DNA-Schaden. Wir haben neue chromosomale Läsionen identifiziert, einschließlich einer Kopienzahlamplifikation in *AGO2* (Argonaute 2, RISC Catalytic Component; 57.14% der Fälle), welches maßgeblich für die Läsion in Chromosom 8q ist. Weil wir signifikante Anreicherungen von trunkierenden *ATM* Mutationen in *ATM* mutiert/ohne Deletion (p=0.01365), sowie *FAT* (FAT Atypical Cadherin)-Domänen-Mutationen in *ATM* mutiert/deletiert (p=0.01156) fanden, könnte es sich bei *JAK3* (Janus Kinase 3) mutierten/*ATM* deletierten Fällen um Fälle einer separaten Tumor-Enwicklungslaufbahn handeln. Mithilfe von Transkriptom-Sequenzierung identifizierten wir neuartige strukturelle Variationen, die Chromosom 14 beeinflussen und zur Expression eines TCL1A-TCR (T-Zell-Rezeptor) Fusionstranskriptes führen, welches wahrscheinlich in einem degradiertem *TCL1A* Protein resultiert. Zwei Gruppierungsansätze zwischen den Genexpressions-Profilen von normalen T-Zellen und leukämischen Fällen, sowie Immunophenotypisierungs-basiertem agglomerativen Gruppierungen und der Rekonstruktion des TCR-Repertoire veranschaulichten einen restriktiven, memory-like T-Zell Phenotyp. Dies ist damit die bis dato umfangreichste und integrativste Studie der T-PLL, durch welche ein evolutionäres Krankheitsmodell etabliert werden konnte und eine *Histon-Deacetylase-hemmende* / Doppelstrangbrüche-induzierende Behandlung, die besser in prä-klinischen Tests abschneidet als der momentane Standard der Chemoimmuntherapie.

Contributions to publications

- Crispatzu G et al. A Critical Evaluation of Analytic Aspects of Gene Expression Profiling in Lymphoid Leukemias with Broad Applications to Cancer Genomics. *AIMS Medical Science*, **3** (3) : 248–271.
  - All data analysis, most survival analyses and manuscript idea and preparation.

- Crispatzu G*, Kulkarni P* et al. Semi-automated cancer genome analysis using high-performance computing (accepted with major revisions)
  - Establishment of a semi-automated somatic NGS analysis pipeline to help wtih clinical diagnostics and biomedical research. Came up with the idea, while working on targeted sequencing data published in Vollbrecht et al. Plos One 2015. I wrote most of the code and wrapper modules (such as somatic copy-number calling or purity estimation). PF further integrated these into the QuickNGS framework. I further did initial testing. Together with CDH and MH I critically reviewed the manuscript. Re-submission with revisions in first half of October 2016.

- Schrader A*, Crispatzu G* et al. Integrated genetic profiles of T-PLL implicate a TCL1/ATM-centered model of aberrant, but actionable damage responses (in review)
  - Did all the statistics and bioinformatics (except for the initial calling of DEU/DEX by RNA-Seq by PF) with focus on integrative approaches such as meta-analyses in all possible data sets, clonality analysis, mutation enrichments and significance, as well as functional prediction. Help in sample selection and study design. Modeling of T-PLL leukemogenesis. Co-wrote the manuscript with AS and MH.

- Warner K*, Oberbeck S*, Schrader A*, Crispatzu G et al. Aberrant effector functions of the memory-type T-PLL cell imply a leukemogenic cooperation of TCL1A with TCR signaling (in review)
  - Did most of the statistics and bioinformatics (except for the manual gatting of FACS by PM, NW and AS). Again focus on integrative approaches such as meta-analyses in publically available data sets of normal T-cells, vBeta clonality correlation analysis, SPADE analysis, marker clustering, and reconstruction of TCR repertoire by RNA-Seq. Help in sample selection and study design.

* Authors contributed equally to this work.

Declaration

I hereby declare that this PhD thesis submitted by me is the result of my own work. References and methods, as well as tables and figures of others are noted duely.
This dissertation is not submitted to any other faculty or university, nor is it published (except for below mentioned part publications) or am I going to before the rightful end of my doctoral curriculum.
The doctoral regulations are known to me. This dissertation was mentored by Prof. Michael Nothnagel.

Cologne, the 10th of November 2016

Part publications

These already have been published within the scope of this thesis:

- Crispatzu G et al. A Critical Evaluation of Analytic Aspects of Gene Expression Profiling in Lymphoid Leukemias with Broad Applications to Cancer Genomics. *AIMS Medical Science*, **3** (3) : 248–271.
- Crispatzu G*, Kulkarni P* et al. Semi-automated cancer genome analysis using high-performance computing (accepted with major revisions; *Human Mutation*)
- Schrader A*, Crispatzu G* et al. Integrated genetic profiles of T-PLL implicate a TCL1/ATM-centered model of aberrant, but actionable damage responses (in review; *Cancer Discovery*)
- Warner K*, Oberbeck S*, Schrader A*, Crispatzu G et al. Aberrant effector functions of the memory-type T-PLL cell imply a leukemogenic cooperation of TCL1A with TCR signaling (in review; *Blood*)

Errata from 1. of September 2017

Due to legal disputes over pharmaceutical patents, an evaluated substance in Manuscript #2 was redacted. It was further raplaced with novel substances within the most current and submitted manuscript. Please refer to Schrader, Crispatzu et. al 2017 (submitted).

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzen Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen und Abbildungen -, die in anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass eine solche Veröffentlichung vor Abschluss des Promotionsverfahren nicht vornehmen werde.

Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Michael Nothnagel betreut worden.

Köln, den 10.11.2016

Teilpublikationen

Es liegen vier Teilpublikationen vor:

- Crispatzu G et al. A Critical Evaluation of Analytic Aspects of Gene Expression Profiling in Lymphoid Leukemias with Broad Applications to Cancer Genomics. *AIMS Medical Science*, **3** (3) : 248–271.
- Crispatzu G*, Kulkarni P* et al. Semi-automated cancer genome analysis using high-performance computing (accepted with major revisions; *Human Mutation*)
- Schrader A*, Crispatzu G* et al. Integrated genetic profiles of T-PLL implicate a TCL1/ATM-centered model of aberrant, but actionable damage responses (in review; *Cancer Discovery*)
- Warner K*, Oberbeck S*, Schrader A*, Crispatzu G et al. Aberrant effector functions of the memory-type T-PLL cell imply a leukemogenic cooperation of TCL1A with TCR signaling (in review; *Blood*)

Errata vom 1. September 2017

Aufgrund von Patentansprüchen wurde im zweiten Manuskript der Name einer Substanz und Verweise auf diese geschwärzt. Diese wurde im aktuellen Manuskript durch andere Substanzen ersetzt. Es empfiehlt sich deswegen ein Blick in Schrader, Crispatzu et al. 2017 (eingereicht).