PhD Thesis

Till Baar

2022

REGRESSION MODELS

for High-Dimensional, Biological Data

REGRESSION MODELS

for High-Dimensional, Biological Data

Till Baar

Institute of Medical Statistics and Computational Biology
Faculty of Medicine, University of Cologne

Thesis Supervisor

Prof. Dr. Achim Tresch

Second Supervisor

Prof. Dr. Andreas Beyer

Regression Models for High-Dimensional, Biological Data

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Till Baar

aus Hamburg

angenommen 2022

Berichterstatter/in: Prof. Dr. Achim Tresch

Prof. Dr. Andreas Beyer

Tag der letzten mündlichen Prüfung:

01.06.2022

Abstract

In this cumulative dissertation, statistical models for regression are discussed in light of high-dimensional, biological data. The dissertation includes three publications:

RNA transcription and degradation of Alu retrotransposons depends on sequence features and evolutionary history examines Alu elements, RNA retrotransposons in the human genome. Their RNA metabolism is poorly understood, and the source of Alu transcripts is still unresolved. We have conducted a transcription shutoff experiment and metabolic RNA labelling to shed further light on the life cycle of Alu transcripts. We furthermore present a novel statistical test for detecting expression quantitative trait loci relying on k-mer sequence representation.

Endoscopic hemostasis makes the difference: Angiographic treatment in patients with lower gastrointestinal bleeding uses retrospective study data from patients receiving either endoscopic or angiographic treatment for lower gastrointestinal bleeding. While a majority of patients can be treated successfully with the usually preferred endoscopic method, in some cases, angiography is required to achieve hemostasis. Using conditional inference trees, we construct a decision tree model predicting if a patient should receive angiographic treatment.

Genetic instability and recurrent *MYC* amplification in *ALK*-translocated NSCLC: a central role of *TP53* mutations investigates a molecular subtype of lung cancer exhibiting rearrangements of the *ALK* gene. This cancer type often resists treatments, and no reliable biomarker to identify patients at risk for relapse is known. Analysing biopsy and cell culture data, we find that mutations in the *TP53* gene can lead to chromosomal instability and thus the amplification of known cancer genes. This, in turn, grants cancer cells a proliferative advantage compared to the wild-type, providing a new approach for diagnosis and treatment.

Contents

1 Introduction

“The temptation to form premature theories upon insufficient data is the bane of our profession.” [Sherlock Holmes, Doyle 2012]. This statement applies not only to fictional consulting detectives but likewise to the field of science. One might even say that it is an inherent tendency of the human mind to jump to conclusions based on incomplete knowledge. Thus, it is the statistician’s task to rigidify all conjectures made in the context of a scientific investigation and base them firmly on observations. This is, of course, easier said than done.

In general, the art of the statistician, in contrast to purely technical aspects, is to specify the model according to the data at hand. Upon contact with data, every mathematical model suffers from the bias-variance tradeoff [von Luxburg 2011], i.e., a model needs to find the right balance between overfitting (high variance) and underfitting (high bias). Bias, in this case, describes the difference between a model’s average predictions and the true values. A model with high bias is oversimplified or even misspecified (i.e., it is inappropriate for describing the data). Variance refers to the variability of a model’s predictions. A model with high variance fails to generalise and cannot make accurate predictions.

Following an example by Hastie 2021, assume we wish to predict an outcome Y given observations $X = \{x_1 \dots x_n\}$ with a relationship of

$$Y = f(X) + \varepsilon$$

where ε describes a normally distributed error term with mean 0 and standard deviation σ_ε. Further assuming a model $\widehat{f}(X)$ of $f(X)$, the expected squared error for a point x becomes

$$\mathrm{E}\,(x) = \mathrm{E}\left[\left(Y - \widehat{f}(x)\right)^2\right]$$

with E denoting the expected value. This further decomposes to

$$\mathrm{E}\,(x) = \underbrace{\left(\mathrm{E}\left[\widehat{f}(x)\right] - f(x)\right)^2}_{\text{Bias}^2} + \underbrace{\mathrm{E}\left[\left(\widehat{f}(x) - \mathrm{E}\left[\widehat{f}(x)\right]\right)^2\right]}_{\text{Variance}} + \sigma_\varepsilon^2$$

where σ_ε^2 is the irreducible error, i.e., the amount of inherent noise in the data that cannot be removed.

In other words, a model that does not capture the pattern from which the observations emerge is underfitted, usually exhibiting high bias and low variance. Vice versa, a model that captures the noisiness of the observations alongside the pattern is overfitted, usually exhibiting low bias and high variance.

The bias-variance tradeoff is connected to the complexity of a model. A model that is too simple and thus undercomplex for the data will underfit. This is because it has too few parameters to model the data adequately. Conversely, a model that is overcomplex will overfit, as it has too many parameters. Of course, one would hope to optimally balance each model's complexity, bias, and variance so to never over- or underfit, or at least come feasible close to this goal.

To approach the optimal model, commonly used methods are regularisation, boosting, and bagging.

Regularisation

Regularisation aims to mitigate the problem of overfitting common to overcomplex models [Deisenroth 2020]. Assume again a model that predicts $f(X)$ and possesses many parameters $\theta_1 \dots \theta_n$. Its associated loss function V governs the training of the model [Rosasco 2004]. Regularisation adds a regularisation term or regulariser R to the loss function that penalises complexity of the model. Thus, the expression to be minimised becomes

$$\min_{\widehat{f}} \sum_{i=1}^{n} V\left(\widehat{f}(x_i)\,, y_i\right) + \lambda\, R\left(\widehat{f}\,\right)$$

with the parameter λ controlling the amount of regularisation that is applied.

Two common applications of regularisation are the linear regression techniques Ridge regression [Hoerl 1970] and Lasso regression [Tibshirani 1996]. While these can be extended to other statistical models, assume simple linear regression for the sake of demonstration. Both techniques add a regularisation term to the loss function that depends directly on the values of the model's parameters θ_i

$$\underbrace{R\left(\theta_i\right)=\lambda \sum_{i=1}^{n} \theta_i^2}_{\text{Ridge regression}} \quad \text{and} \quad \underbrace{R\left(\theta_i\right)=\lambda \sum_{i=1}^{n} |\,\theta_i|}_{\text{Lasso regression}}$$

thus shrinking all but the most influential parameters of the model and thereby reducing model complexity and multicollinearity [Herawati 2018]. The difference between Ridge and Lasso regression lies in the calculation of the applied penalty. While Ridge regression penalises the sum of the squared coefficients (L2 penalty), Lasso regression penalises the sum of their absolute values (L1 penalty). The ultimate consequence is that while Lasso can shrink non-influential parameters to zero, Ridge cannot. On the other hand, this can cause Lasso to eliminate important parameters under multicollinearity, if predictor variables are correlated, as it tends to select one parameter from the correlated group and ignore the rest.

To overcome these limitations, a combination of Ridge and Lasso regression can be applied, elastic net [Zou 2005]. The used regularisation technique combines an L1 and an L2 penalty by using separate λ parameters for each, λ_1 and λ_2. If $\lambda_1 = 0$, the penalty equals Ridge regularisation; if $\lambda_2 = 0$, the penalty equals Lasso regularisation; and if $\lambda_1 > 0$ and $\lambda_2 > 0$, a combination of both is applied.

Boosting

While regularisation is a helpful method to deal with overcomplex models, boosting addresses the problem of poor models in more general terms [Schapire 2009]; a 'poor model', in this case, refers to a weak learner. Valiant [1984] formalised the concept of learnability in the context of computational complexity theory and introduced the probably approximately correct (PAC) model. A problem is PAC- learnable if there exists a model that, with a chance higher than a threshold δ, will arrive at a solution with a generalisation error smaller than a threshold ϵ. Generalisation error or out-of-sample error refers to a model's predictive performance on previously unseen data [Bousquet 2011]. A model satisfying these conditions for any given problem is called a strong learner, while a model that does not is a weak learner.

A problem can benefit from boosting if applying a strong learner is either impossible, as no strong learner exists, or disadvantageous, for example, because the strong learner is prohibitively complex and thus underperformant, or because the available training data is insufficient to apply it. While current machine learning research, especially in the field of deep learning [LeCun 2015], mainly approaches the challenge of more complex problems by fielding stronger algorithms, boosting seeks to improve the results of weak learners.

While Kearns [1994] defined weak learners as models that perform just slightly better than random guessing, Schapire [1990] demonstrated their power if applied correctly, proving that any problem solvable by a strong learner is equally solvable by a collection of weak learners: the hypothesis boosting mechanism. The term 'hypothesis' here describes the solution a model arrives at after training, the model's final parameters. Freund [1995] improved this further, combining many weak learners and using their combined results to arrive at a strong prediction, one weak learner effectively compensating for the shortcomings of another. The next step was AdaBoost, adaptive boosting, for which Freund and Schapire were awarded the Gödel Prize in 2003 [Freund 1997]. This boosting variant scales each weak learner's influence on the final prediction depending on their own error. The current state-of-the-art boosting technique is gradient boosting with its predominant implementation XGBoost [Chen 2016]. In contrast to AdaBoost, which always minimises the exponential loss function, gradient boosting can use any differentiable loss function, which makes it adaptable to many classification and regression tasks.

Following an example by Li [2015], assume again a model $\widehat{f}(X)$. The boosting model iterates over M stages, and each stage m has an associated imperfect model $\widehat{f}(X)_m$, so that at each stage, a new 'hypothesis' is added, a new estimator $\widehat{g}\,(X)_m$.

$$\widehat{f}(X)_{m+1} = \widehat{f}(X)_m + \widehat{g}\,(X)_m$$

As the model iterates over sets of training data $Y_i = \{y_1 \dots y_n\}$, the new estimator is fit to the residual, the difference between the values of the training data Y_i and the estimation of the previous model.

$$\widehat{g}\,(X)_m = Y_i - \widehat{f}(X)_m$$

In this fashion, each new stage $m+1$ attempts to correct for the errors made by the previous stage m.

Bagging

While boosting is an ensemble method that addresses shortcomings of the model, bagging can be regarded as addressing shortcomings of the data [Breiman 1996]. The name 'bagging' derives from bootstrap aggregating. Bootstrapping is a resampling method that estimates statistics by sampling repeatedly from the same data [Efron 1994]. This process makes it possible to assess the accuracy of the estimated statistics, which can be assumed to be an adequate approximation if the empirical distribution of the data represents the true distribution reasonably well.

Bagging applies the principle of bootstrapping to model training. Following an example by Aslam [2007], assume again a set of training data Y. From this data set, new training sets Y_i are created by sampling uniformly from Y with replacement, meaning that each individual entry in Y has the same probability of being drawn and can be drawn again and again. Individual models are then trained on the new training sets Y_i, and their predictions are combined, either by averaging for regression or voting for classification (see Regression and Classification below). Bagging is especially useful for unstable models that can react drastically to small changes in the training data (see Decision Trees and Random Forest on page 10 for an example).

Regression and Classification

While the variety of mathematical models seems almost endless, the models employed in the publications that comprise this dissertation belong predominantly to the field of machine learning. These algorithms encompass models that use sample or training data to learn and make predictions [Mitchell 1997]. Machine learning can be divided into three main approaches, unsupervised and supervised learning, and reinforcement learning. This third category covers the behaviour of intelligent agents that interact with the environment and is of only marginal interest to the topics at hand [Joshi 2021].

Supervised and Unsupervised Learning

Unsupervised learning differs from supervised learning by the type of training data required [Hinton 1999]. Unsupervised models are not reliant on labelled data, meaning data that has been annotated by humans or other models. Instead, unsupervised learning aims to build an internal representation of the space it operates in and capture previously known patterns according to that representation. Some prominent examples of unsupervised learning include clustering [Rokach 2006], dimensionality reduction [Van Der Maaten 2009], and outlier detection [Hawkins 1980]. Supervised learning, on the other hand, does require labelled data. Its goal is to find a function that maps input variables to output variables as best as possible, or in other words, to train a model so that it predicts an outcome as best as possible, given a set of corresponding observations [Mohri 2018]. Some prominent examples for supervised learning include Bayesian inference [Gelman 2014], decision trees [Kamiński 2018], and support vector machines [Cortes 1995]. Most of the methods discussed in detail prior to the included publications are supervised.

Supervised learning is commonly further subdivided into two fields of application: regression and classification. While regression predicts a numerical (i.e., continuous) outcome, classification predicts discrete class labels [Hastie 2021].

Peculiarities of Biological Data

As an empirical and descriptive discipline, the life sciences, and biology, in particular, are founded on data. The nature of this information is thus vital to all scientific enquiries in the field. Following the introductions by Jagadish [2003] and Wooley [2006], biological data can be broadly classified into the following types.

Spatial Data

Biological systems, from strands of DNA in the nucleus of a cell to animal migrations taking place over thousands of kilometres, are three-dimensional in nature and therefore carry spatial information. Measuring and encoding the differences between one region and another in machine-readable form is thus instrumental. Scalar and vector fields can be seen as an extension of this, as they encode phenomena that are continuous in space, such as biochemical properties like concentration gradients or population densities.

Sequence Data

Sequence data is currently one of the most abundant forms of biological information and arguable responsible for the vast majority of progress in the field over the last decades. The amount of available DNA and RNA sequences is also increasing ever more quickly with the development of novel technology, such as single-cell sequencing [Wang 2015] and spatial genomics [Turczyk 2020]. While these two techniques further explore the genetic organisation on the level of individual cells, another approach called metagenomics analyses all sequences present in an ecosystem [Venter 2004, Hugenholtz 2008]. Sequence data can be generalised as strings representing the DNA or RNA bases, including gaps.

Patterns

Within DNA and RNA sequences lie patterns that represent functional units, such as genes in the genome, functional elements like promoters and enhancers [Kim 2015], or restriction sites [Smith 1976]. Patterns can be encoded as context-free grammars [Hopcroft 2001], Hidden Markov Models (HMMs) [Stamp 2018], or regular expressions [Wang 2019].

Models

Another type of information that can be regarded as biological data are the mathematical models created to analyse experimental measurements. With the increasing number of publications, the models contained therein, their structure and parameters, need to be machine- readable to facilitate comparisons.

Images

Images originating from electron and optical microscopy, radiography, and other methods, as well as videos, are another type of data that is especially difficult to convert into a machine-readable form. While storing the raw data digitally is trivial, extracting the features contained in the recordings is not and has spawned the interdisciplinary field of computer vision [Ballard 1982].

Graphs

Relationships such as biochemical and signalling pathways and phylogenetic trees can be represented as graphs, along with gene regulatory networks and laboratory workflows. Even sequence data can be presented in a graph structure to efficiently encode DNA and RNA sequences variability between individuals [Novak 2017]. Geometric arrangements such as the three-dimensional shape of proteins that governs their docking behaviour can also be rendered in graph-form.

Prose

Finally, the literature itself is a form of data, and the annotations, hypotheses, and inferences stated in continuous text are difficult to translate into machine-readable form, as well [Balyan 2017].

As should become clear from this diverse list, biological data can be very heterogeneous, which can further complicate its analysis, as models may need to be found which can integrate the different data types. Epigenetic data, for example, combines spatial, sequence, and pattern information in the form of genome architecture and nucleosome positioning, DNA and RNA sequences, as well as the patterns of promoters and enhancers [Armstrong 2020]. Biological data also originates, in most cases, from laboratory experiments. This has the consequence that equipment- and protocol-dependent biases are almost guaranteed to be present. Even the person performing the experiment can be a confounding factor. It is thus highly unusual that experimental results from different laboratories agree. A promising remedy for this is zero-sum regression by Altenbuchinger [2017], an extension to conventional linear regression that has also been adapted to logistic regression [Kleinbaum 2002] and Cox proportional hazard regression [Fox 2002]. Zero-sum regression, in its simplest from, enforces the zero-sum constraint on a log-linear model, meaning a linear model on log-transformed data where the choice of the reference point can result in sample-wise shifts.

Zero-Sum Regression

Assume again a model with many parameters θ_j that should predict an outcome y_i from predictor x_i with the form

$$\begin{aligned}\widehat{y}_i =&\theta_0 + \sum_{j=1}^{n} \theta_j \log(y_i\, x_{ij})\\ \widehat{y}_i =&\theta_0 + \sum_{j=1}^{n} \theta_j \log(y_i) + \sum_{j=1}^{n} \theta_j \log(x_{ij})\\ \widehat{y}_i =&\theta_0 + \log(y_i) \sum_{j=1}^{n} \theta_j + \sum_{j=1}^{n} \theta_j \log(x_{ij})\end{aligned}$$

By restricting the sum of coefficients (marked in blue) to zero, the linear model is made insensitive to the choice of the reference point. Example reference points include the amount of DNA or RNA included in an experiment or the number of cells. Zero-sum regression assumes that any chosen reference point can be suboptimal. Thus, a model and the resulting biological interpretation should not rely on it, if possible. The systematic differences between experimental conditions are modelled separately and can thus be removed. The intuition behind the method is that, in high-dimensional space, a subspace is found that lies orthogonal to the unwanted shift in the data. Thereby, the subspace becomes invariant to the systematic differences.

Other pitfalls of biological data include its volume, variance, and range. Due to being measurements of inherently noisy phenomena, biological data is usually generated in replicates. This makes it possible to attribute parts of the observed variance to experimental conditions, such as batch effects, while the remainder derives from the phenomenon itself. However, this also means that the data volume is multiplied by the number of replicates. Because sets of raw data generated by modern methods can easily reach several hundreds of gigabytes in size, these measurements can become challenging to handle without appropriate computational resources. Biological phenomena also tend to span several orders of magnitude, for example, in the case of transcript counts associated with individual genomic loci. This complicates matters primarily due to the human factor involved since humans are generally ill-equipped to think analytically in logarithmic terms, even though our senses perceive stimuli on a logarithmic scale [Sun 2012].

Curse of Dimensionality

Finally, biological data also tends to suffer from the curse of dimensionality, a term coined by Richard E. Bellman [Bellman 1957, 1961]. In machine learning, dimensions are synonymous with features, and the curse of dimensionality refers to the pros and cons of a data set with many features; a high-dimensional data set. On the one hand, having many features can be a blessing when it comes to separating data into distinct classes, as points that are difficult, if not impossible, to separate clearly in low dimensions can become easy to separate in higher dimensions. However, on the other hand, when the dimensionality of Euclidean space increases, the distance between points in the space increases, too, as it is proportional to the square root of the number of dimensions [Tabak 2014]. This has the consequence that, with increasing dimensions, Euclidean space becomes vast, and the data becomes sparse. Dimensionality reductions methods like PCA [Pearson 1901], t-SNE [Hinton 2003], UMAP [McInnes 2018], or Autoencoders [Kramer 1991] can serve as a remedy for this problem.

1.1 Method Overview

As many statistical methods are used in more than one instance in the included publications, the following section describes in short the main methods of interest.

Hypothesis Testing

While more complex methods are essential to many findings in the included publications, hypothesis testing is nonetheless a vital foundation for any statistical analysis. In testing theory, a statistical test describes a method used to judge the validity or invalidity of a formal hypothesis [Teunissen 2006]. As sampled data is subject to errors, it is not possible to definitely prove the correctness of such a hypothesis; it is only possible to control the probability of making the wrong decision. In general, a hypothesis test defines two hypotheses, the null hypothesis H_0 which is the standard assumption and holds until it can be rejected with a sufficiently high probability, and the alternative hypothesis H_1 which only applies if H_0 is rejected.

The three most used hypothesis tests in the three included publications are the Brown-Forsythe test, Fisher's exact test, and the Mann-Whitney U test [Fisher 1922, Wilcoxon 1945, Mann 1947, Brown 1974, Winters 2010]. The Brown- Forsythe test assesses if the variances of two groups are equal (homoscedasticity). Fisher's exact test assesses if two or more groups differ with regard to categorical data, while the Mann-Whitney U test, also called the Wilcoxon rank-sum test, is a nonparametric test that assesses if two groups differ with regard to continuous data.

Correlation

Another basic method is correlation analysis. In the broadest sense, correlation describes any relationship between two random variables, or more specifically, the degree to which these variables are linearly related [Mann 1947]. The two most common measures are Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient. Pearson correlation represents a normalised covariance measure and can only report linear relationships. Spearman correlation, on the other hand, uses the rankings of each variable and can thus detect monotonic relationships regardless if they are linear or not.

Decision Trees and Random Forest

The first more involved method used in the included publications are decision trees [Wu 2008]. In general, a decision tree, which can be used for classification or regression, splits the data set first by the predictor variable that best differentiates between the states the outcome variable can take. The resulting subsets are then split again, each by the variable best suited to the subset. This is repeated until an end condition is reached. Both the definition of the best split and the end conditions vary between methods. Decision trees, however, are not very robust and tend to generalise poorly. Random Forest offers a remedy for this by applying the principle of bagging (see page 4) to decision trees [Ho 1995, Breiman 2001]. Not only is the training data subjected to a bagging scheme, but feature bagging is also applied, meaning that a random subset of predictors is considered in each tree of the forest. While random forests thereby achieve greatly improved generalisation compared to decision trees, they also lose the intrinsic interpretability that makes decision trees compelling machine learning models.

Generalised Linear Models

A method focused on regression are generalised linear models (GLMs) [Nelder 1972]. As the name suggests, GLMs generalise linear models (LMs) by introducing a link function. The intuition behind the link function is that it provides a relationship between the linear combination of predictors in the underlying model and another arbitrary distribution function that describes the observations. It converts the expected value of the observation to the scale of the linear predictor. However, the link function is not a data transformation, as it does no operate on individual observations y_i, but on the expectation $\mathrm{E}(Y)$.

Standard LMs can be considered a subclass of GLMs, where the link function is the identity. In nontrivial cases, an exponential-family distribution can be modelled by selecting the appropriate link function. For example, an LM cannot model observations following a binomial distribution in a meaningful way, as it could theoretically predict an outcome above $100\,\%$. Using the appropriate link function rectifies this. Assume a linear predictor

$$\eta = \theta_0 + x_1\theta_1 + x_2\theta_2 + \ldots + x_n\theta_n$$

with parameters θ_j and observations x_j. The link function then takes the form

$$g(\mu) = \eta \quad \text{with} \quad \mu = \mathrm{E}(Y)$$

where the canonical parameter μ is one of the parameters in the standard form of the distribution's density function. For example, in the case of a binomial distribution, the link function becomes

$$g(\mu) = \ln\left(\frac{\mu}{n-\mu}\right)$$

Maximum Likelihood Estimation

Underlying many statistical methods is maximum likelihood estimation (MLE), which aims to fit a distribution to observed data by estimating the distribution's parameters Hastie 2021. Assume a set of observations $x_1, x_2, \ldots, x_n$ that are independent and identically distributed *(iid)* and come from an unknown distribution function f with parameters ϑ. The density function of f can thus be expressed as

$$f(x_1, x_2, \ldots, x_n; \vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta)$$

The density can now be reformulated as a function depending on ϑ to arrive at the likelihood function

$$\mathcal{L}(\vartheta) = \prod_{i=1}^{n} f_\vartheta(x_i)$$

Maximising the likelihood function with respect to the distribution function parameters ϑ thus results in the maximum likelihood estimates for ϑ. Alternatively, the logarithmic likelihood function can be maximised instead, which is often an easier feat and results in the same estimates for ϑ.

$$\ell(\vartheta) = \log\left(\prod_{i=1}^{n} f_\vartheta(x_i)\right)$$

The main drawback of MLE is that the correction underlying distribution the data is sampled from must be used. Should the distribution be assumed wrongly, the results of the MLE will most likely be inconsistent.

2

Evolution Shapes the Alu RNA Metabolism

In this explorative study, we looked into the lifecycle of Alu elements, retrotransposons in the human genome that copy themselves into new genomic positions. We wanted to answer four questions concerning the transcription and degradation of Alu RNAs and the sequence features that influence these processes.

Alu elements, named after a restriction endonuclease of *Athrobacter luteus* which lead to their discovery, classified as short interspersed nuclear elements (SINEs), are around 300 bp long mobile DNA sequences found in the human genome and other species [Schmid 1975, Quentin 1992, Lander 2001, Kriegs 2007, Deininger 2011]. They are RNA retrotransposons, meaning that they are capable of copying themselves into new positions in the genome. Due to this multiplication, Alu elements make up 11 % of

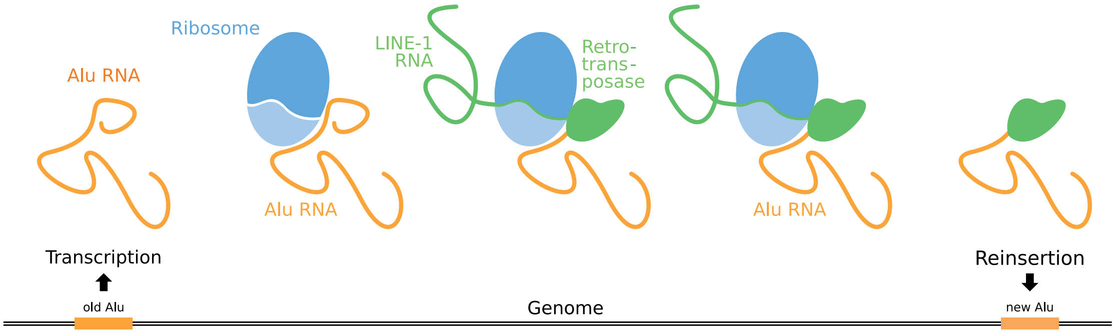

Figure 1: Modification of figure 1a from Baar [2022], showing a schematic representation of Alu retrotransposition (left to right): the Alu element is transcribed – the Alu RNA attaches itself to the exit tunnel of the ribosome through its SRP sequence homolog – a LINE-1 RNA arrives at the ribosome and its retrotransposase is translated – the Alu RNA hijacks the LINE-1 retrotransposase – the LINE-1 retrotransposase reinserts the Alu element into a new genomic position.

the human genome by length, with more than 1 million currently annotated loci [Lander 2001].

The retrotransposition process of Alu elements shown in figure 1, however, is error-prone and thus facilitates an Alu-specific sequence evolution, giving rise to distinct Alu families from the old AluJ family over AluS to the young AluY [Richard Shen 1991, Deininger 1999, Batzer 1996]. In general, Alu elements are composed of a left arm and a right arm separated by a variable A-rich region [Evgen'ev 2007]. The left arm contains an RNA Polymerase III (Pol-III) promoter, and the right arm holds the UGU(NR) motif required for binding to the ribosome (see figure 1) [Paolella 1983, Dagan 2004, Orioli 2012]. If Alu elements serve a function is, so far, unknown. They can be harmful if a new insertion disrupts a gene or other genomic region [Deininger 1999]. They have also been linked to changes in transcriptional activity in general and under heat shock conditions in particular [Mariner 2008, Chen 2017, Zhang 2019].

We wanted to address four questions regarding Alu elements, their transcription, and their sequence features:

1. Are Alu RNAs stable or unstable?

   Previous studies suggested that Alu RNAs should be less stable in the cell than regular mRNAs, meaning that they are degraded quickly [An 2004]. However, these results were obtained using only computational methods extrapolating from Alu sequence features and were not backed up by experimental data.
   We could show that the distribution of Alu RNA half-lives predicted by our experimental approach is very similar to that of mRNAs, suggesting that Alu transcripts are more stable than previously thought.

2. Are Alu elements transcribed primarily by Pol-III or by RNA Polymerase II (Pol-II), as well?

   While Alu elements do contain a Pol-III promoter sequence (see page 13), it is surmised that Alu elements are not only transcribed by Pol-III but also, to a certain extent, by Pol-II [Conti 2015, Zhang 2019]. The experimental evidence for this is mostly indirect [Zhang 2019, Panning 1993, Jagadeeswaran 1981]. We have therefore conducted a Pol-II inhibition experiment and could show that Alu expression does indeed decrease under Pol-II inhibition, suggesting strongly that Alu transcripts arise, at least in part, directly from Pol-II activity.
   This finding carries implications for differential expression

analyses of the past. Often, Alu elements were used as a control group if Pol-II inhibition was performed, assuming wrongly that Alu transcription should be completely dependent on Pol-III [Cordaux 2009].

3. Is Alu transcription a side product of gene transcription?

   If a fraction of Alu transcription depends on Pol-II activity, a possible explanation could be that Alu elements are transcribed alongside regular genes [Conti 2015, Zhang 2019]. The results of our analyses weaken this hypothesis. While we cannot rule out that a fraction of Alu elements might be transcribed alongside genes, it is unlikely that this process contributes substantially to Alu expression.

4. Is Alu expression influenced by sequence features?

   Previous studies unsuccessfully employed regular *de novo* motif search to detect Alu sequence features that influence their expression [Zhang 2019]. We used two less common methods (see Analysis) and could uncover several influential positions and motifs that appear linked to changes in Alu expression. Additionally, some of the motifs match transcription factor binding profiles and may thus present promising targets for future investigations.

Methodology

This study made use of bulk, whole-genome RNA-seq data, partially generated specifically for our investigation and partially repurposed from our earlier publication Schwalb [2016]. The RNA was extracted from K562 cells, an immortalised human suspension cell line of erythroleukemia cells [Andersson 1979]. The sequencing data is noteworthy regarding two of its characteristics:

Firstly, we used dynamic transcriptome analysis (DTA), specifically the 4sUseq and the TT-seq methods [Schwalb 2012, Gressel 2019]. DTA chemically labels newly created transcripts, making them distinguishable from old RNAs still present from before the labelling pulse. This allowed us to calculate the ratio between old and new transcripts and thereby estimate these transcripts' half-life.

Secondly, we performed a Pol-II inhibition experiment using α-amanitin, a toxic substance from the *Amanita phalloides* fungus that blocks Pol-II activity [Lindell 1970, Kedinger 1970, Stirpe 1967, Jacob 1970]. We treated K562 cells with α-amanitin and compared their RNA-seq data with untreated samples. We could thus observe the effect of Pol-II inhibition on different transcript classes, including Alu elements, *bona fide* Pol-II genes, and tRNAs, which are transcribed by the uninhibited Pol-III.

Analysis

Half-Life Estimation

Concerning the question, if Alu RNAs are stable or unstable, we relied on the DTA data (see Methodology), giving us two read counts for each Alu element or gene, one from the labelled fraction and one from the total fraction. With time, new labelled transcripts are created and old unlabelled transcripts decay, meaning that the ratio of labelled RNAs increases until all transcripts are labelled. Assuming exponential decay and steady-state conditions, this ratio r_a gives us the decay rate δ_a for any Alu element or gene a by

$$\begin{aligned} r_a &= {}^{l_a}/_{t_a} = 1 - \exp(-\delta_a \Delta t) \\ \ln r_a &= \ln\left(1 - \exp(-\delta_a \Delta t)\right) \end{aligned}$$

where l_a and t_a are the numbers of labelled or total RNA molecules and Δt is the labelling pulse's duration, meaning the time that passed after the labelling agent was added and before the RNA sequencing was performed. From the decay rate δ_a the half-life $t_{1/2,a}$ is given by

$$t_{1/2,a} = \frac{\ln(2)}{\delta_a}$$

To estimate the ratio between labelled and total RNA molecules from the measured reads, we used maximum likelihood estimation (MLE) [Rossi 2018] (see also Method Overview). We made several assumptions to simplify the estimation, owing to the general paucity of Alu read counts caused by their low expression. We assume steady-state conditions, use a Poisson distribution to model read counts instead of a zero-inflated negative binomial distribution, and neglect non-constant labelling efficiencies for short labelling periods. This means that our estimation can only serve as an assessment to compare the relative half-life distributions of Alu elements and genes. Its predictions do not represent explicit half-life values. Still, the very similar distribution of Alu and gene half-lives suggests that the stability of Alu transcripts is greater than would be expected from sequence features alone.

Differential Expression Analysis

The α-amanitin Pol-II inhibition experiment was the key to investigating the origin of Alu transcripts. To analyse the RNAseq measurements, we used the DESeq2 package for R [Love 2014]. The standard assumption of DESeq2's internal normalisation strategy is that there are no substantial, systematic global expression changes between samples. Due to the inhibition of Pol-II, we have to assume that this does not hold. We, therefore, used mitochondrial transcripts (mtRNAs) for normalisation, which are unaffected by the α-amanitin treatment. This is because the mitochondrial polymerase that transcribes mtRNAs is not inhibited by α-amanitin [Menon 1971, Reid 1971, Saccone 1971]. The resulting differential expression estimates of Alu elements, mRNAs, and tRNAs, which we used as a negative control as they are transcribed by Pol-III, which is also unaffected by α-amanitin, are shown in figure 2 [White 1997]. As expected, tRNAs remain largely unaffected by the α-amanitin treatment, while mRNAs exhibit downregulation. Notably, Alu RNAs also appear downregulated under Pol-II inhibition, suggesting strongly that Alu elements are, at least to a certain extent, transcribed by Pol-II.

Correlation Analysis

To address the follow-up question if the apparent expression of Alu elements by Pol-II may result from Alu RNAs being side products of gene transcription, we argued as follows [Conti 2015, Zhang 2019]: If Alu elements were transcribed alongside genes, those inside or close to genes should show higher expression compared to Alu elements that are far away from genes. The expression of a gene should also correlate with the expression of Alu elements inside or close to it. Finally, Alu elements should show a bias towards lying in sense direction with regard to their associated gene.

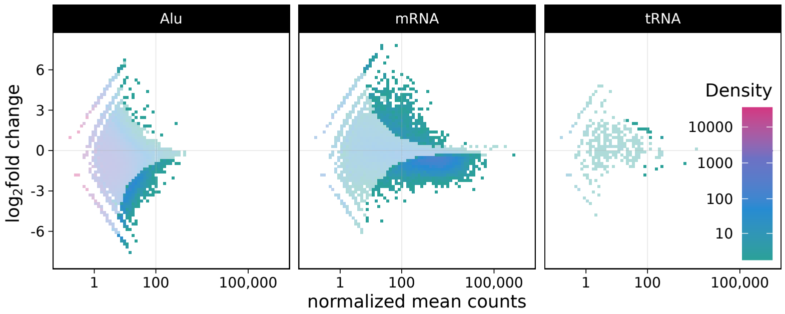

Figure 2: Modification of figure 3e from Baar [2022]. 2D density heatmap showing the DESeq2 differential expression of Alu elements, mRNAs, and tRNAs under α-amanitin Pol-II inhibition. Semi-transparent areas do not pass the significance threshold. Loci with a normalised mean expression below 0.1 are excluded, with affects 51 % of all annotated Alu loci and practically no mRNAs or tRNAs.

While we did detect a significantly higher expression of Alu elements that lie inside or close to genes, this effect can result easily from increased genome accessibility in areas of active gene transcription [Guo 2017]. This also biases the correlation analysis. Therefore, we examined the difference in correlation strength between Alu element and gene transcription if split by sense and antisense Alu direction. As we detected no difference in correlation strength and also found that Alu elements show no preference for inserting themselves in sense direction into genes, we conclude that it is unlikely that Alu transcription is a side product of gene transcription. While our results do not rule out that some Alu transcripts are created alongside genes, this does not appear to be a major source for Alu RNAs.

Generalised Linear Model

To search for sequence features that influence Alu expression, we pursued two different approaches. Firstly, to analyse Alu elements on a per-base level, we used a generalised linear model created with the glmnet package for R [Nelder 1972, Friedman 2010] (see also Method Overview). We assumed a Poisson family distribution response type and used an elastic mixing parameter α of 1 (full Lasso penalty, no ridge regression penalty), no fitted intercept parameter, and $1000\times$ cross-validation. To create the input matrices, we aligned all Alu sequences against the Alu consensus sequence and encoded base exchanges, deletions, and insertions for each position as a binary matrix. For these three types of point mutation, we trained GLMs with the Alu read counts as a response variable. Finally, we used the Euclidean norm of the three obtained effect sizes for each position in the Alu consensus sequence to judge their respective importance, uncovering several influential positions.

De Bruijn Graph

Secondly, to detect larger sequence features, we created a De Bruijn graph of all Alu sequences using bifrost v1.0.5 [Holley 2020]. A De Bruijn genome graph is a way to encode sequence variability in a graph structure [Chikhi 2014]. This method is based on general De Bruijn graphs, which are directed graphs representing the overlap between symbol strings [Sainte-Marie 1894, De Bruijn 1946, Good 1946]. In the context of biological data, the strings are sequences of length k (k-mers), using DNA or RNA bases as symbols. Each node in the graph represents one unique k-mer present in the source data from which the k-mers are generated. Each node also represents its own reverse complement, depending on the direction in which it is traversed. Edges in the graph represent overlaps. For example, a sequence S_1 would be connected by an edge to another sequence S_2 if they overlap except for one base, such that

$$S_1 = (s_1, s_2, \ldots, s_n) \quad \text{and} \quad S_2 = (s_2, s_3, \ldots, s_n, s_{n+1})$$

With 4 biological bases, each node can have up to 16 edges connected to it, 4 in- and 4 out-edged in forward direction and again in reverse complement direction. If a graph is compacted, sequential nodes without branching edges can be combined into a single node representing a sequence with a length greater than k.

As the De Bruijn genome graph we constructed from all Alu sequences was almost complete with an in-degree per node of >7.99, we focused on the constituent k-mers and disregarded the graph structure in downstream analyses. We filtered the k-mers using two criteria. The expression of Alu elements possessing the k-mer needed to be significantly different from those not possessing it. Also, we took each k-mers's suffix and prefix into account.

Assuming a k-mer with the structure X J Y, with X, Y $\in \{A, C, T, G\}$ and J a fixed 2-mer, to assure that X J Y is causal for observed changes in Alu transcription, we test the group of Alu elements containing X J Y against the group containing the k-mer's prefix X J or its suffix J Y, but not the full k-mer. Thus we can be confident that the observed effect is caused by the complete k-mer and not just its partial sequence.

With this method. we uncovered several statistically significant and biologically relevant sequence motifs which previous attempts using regular *de novo* motif search did not [Zhang 2019].

In summary, we found that Alu transcripts appear to be as stable as mRNAs, more stable than previously thought. In addition, we found evidence for Alu elements originating in independent Pol-II transcription, not originating as side products of gene transcription. Finally, we also identified a list of sequence features that influence Alu expression and might therefore be promising targets for future investigations.

My contribution to this publication was the complete bioinformatic and statistical analysis.

> Baar, T., Dümcke, S., Gressel, S., Schwalb, B., Dilthey, A., Cramer, P., Tresch, A. (2022).
>
> RNA transcription and degradation of Alu retrotransposons depends on sequence features and evolutionary history
>
> *G3: Genes | Genomes | Genetics, 2160-1836*

3 Angiography for Gastrointestinal Bleeding

This retrospective study's goal was the identification of variables that increase the chance for a patient suffering from lower gastrointestinal bleeding (LGIB) to benefit from angiography.

LGIB describes any form of gastrointestinal (GI) bleeding occurring in the lower gastrointestinal tract, which includes most of the small intestine and all of the large intestine [Treuting 2018]. GI bleeding can have many causes, including cancer, and the resulting blood loss can lead to shock, syncope, and even death with a chance of around 15 % in general [Rockey 2005, Prasad Kerlin 2013, Wang 2013, Kim 2014]. While the majority of GI bleedings subside on their own or can be arrested through endoscopic treatment, endoscopy cannot detect the cause of LGIB in 40 % of all cases [Yamada 2015, Werner 2018]. Once localised, though, over 90 % of LGIBs can be treated successfully. It is therefore vital that in cases with symptoms severe enough to result in hospitalisation, the source of bleeding is identified quickly and reliably and that hemostasis is achieved, be that through endoscopic treatment or surgery [Strate 2010, Werner 2018].

It is at this point that angiography comes into the picture. Angiography is a medical radiological imaging technique that visualises blood vessels, as well as bleeding. This is achieved by injecting a radio-opaque contrast agent into the bloodstream in conjunction with X-ray imaging [Martin 2015]. Catheter angiography (CA), coupled with transarterial embolisation (TAE), a method to stop the flow of blood to a selected area of tissue, has high technical success rates of 90 % to 100 % and low complication rates of 1 % to 5 % [Tan 2008, Evangelista 2000, Strate 2010, Kim 2017, Lee 2018, Oakland 2019, Pannatier 2019]. However, it also exposes the patient to the contrast agent and X-ray imaging, while endoscopy requires only anaesthesia. Angiography is also a more complex technique and involves the patient's referral to a radiologist. Thus, the decision of when to conclude endoscopic procedures and begin angiographic treatment is challenging. If angiography is initiated too late, the patient is subjected to multiple failed endoscopies, while if angiography is used too early, the patient is needlessly exposed to the side effects of radiological and surgical treatment.

Accordingly, our goal was to construct a decision-making aid for clinicians, assisting them in deciding when to apply endoscopy and when to apply angiography to treat LGIB. While prospective investigations will be required to consolidate our results, the predictors we selected may contribute to the development of future official guidelines.

Methodology

The data for this study was collected over the span of 11 years at a maximum care hospital and included 133 patients. Of these, the treatment group consisted of 41 patients that received CA for LGIB, while the control group of 92 patients was treated for LGIB without angiography. 110 variables were recorded for each patient, of which 20 were designated as being of particular clinical relevance according to expert opinion.

As the data collection period was so extensive and involved many clinicians, no precise statements concerning the methods used for data recording can be made. The data types were also highly diverse, ranging from binary labels, such as clinical success, time intervals and ordinal variables, to numeric laboratory test results.

Analysis

All of our data was ultimately recorded by humans and was thus flawed, which may sound harsh but is true more often than not. For example, Gøtzsche [1989] reports that 76 % of the 196 analysed drug trials to treat rheumatoid arthritis contained "doubtful or invalid statements" [Brown 2018]. Therefore, data cleaning and validation was the first step, an arduous step that is nonetheless crucial.

Descriptive statistics followed, along with naïve pairwise correlation analyses between selected variables and nonparametric tests, using either Fisher's exact test or the Mann-Whitney U test, depending on the data at hand [Winters 2010] (see also Method Overview).

At the start of the principal analysis, two items had to be considered . Firstly, we had to establish that we assume the professional decision of the clinicians to treat a patient either with or without angiography to be founded in their medical expertise. The alternative assumption would be that we cannot equate a patient receiving angiographic treatment with the need of a patient to receive such treatment. Under this alternative assumption, our investigation could not have drawn any meaningful conclusions. It is a natural limitation of a retrospective study, which is also why prospective investigations are necessary before an official guideline can be established. Hence, we assume that a causal link exists between a patient's statistics and treatment.

Secondly, our goal was to create a decision-making aid for clinicians, helping them decide when to switch from endoscopy to angiography for the treatment of LGIB. While we could have constructed a complex regression model to predict the method suitable for a patient as best as possible, such a model would not be applicable in the daily clinic routine. While computer-based predictors to guide treatment decisions may become mundane in the future, this is not yet the case. Consequently, our decision-making aid needed to be easily traceable, allowing the clinician to arrive at a prognosis by following a transparent algorithm. Accepting that, in consequence, our final model may be undercomplex

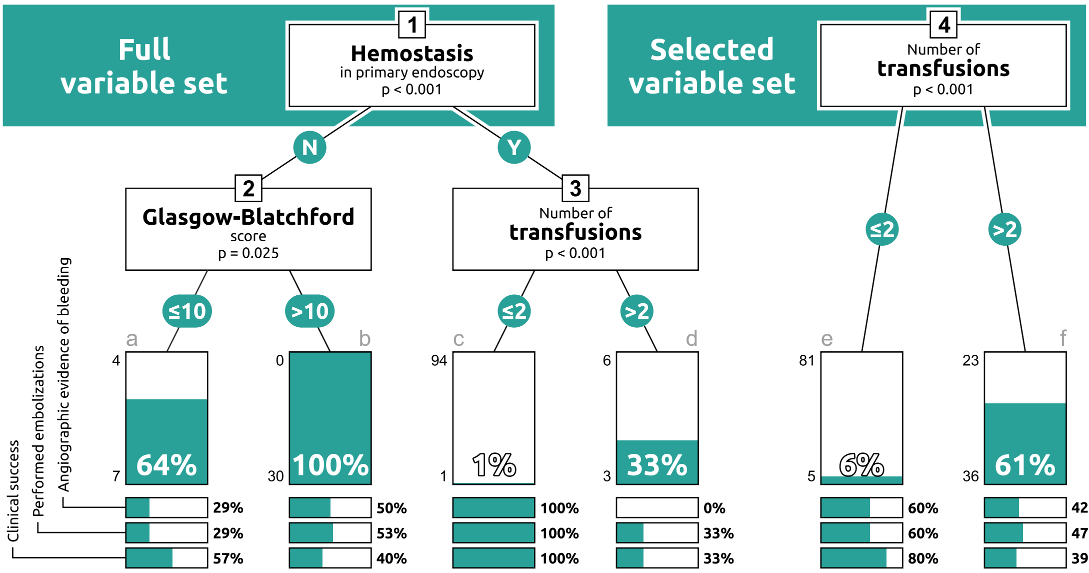

Figure 3: Modification of figures 2 and 3 from Werner [2021]. Conditional inference trees were constructed from either the complete data set (left) or a set of variables selected for clinical relevance (right). Each binary split (numbered boxes 1 to 4) is annotated with its p-value. Each terminal node (vertical bars a to f) shows the percentage of angiography-positive cases.

to some extent and thus prone to underfitting, we elected to use conditional inference trees (see Method Overview).

We decided to construct two decision trees shown in figure 3, one based on all recorded patient variables and another based solely on the clinically relevant variables. The trees offer a straightforward way for a clinician to determine if a patient should receive angiography or not. Using the tree based on the full variable set as an example, only two queries are needed for a patient: If hemostasis was achieved in the first endoscopy, the number of blood transfusions a patient has received is needed as input. On the other hand, if the primary endoscopy failed to achieve hemostasis, the Glasgow- Blatchford Bleeding Score (GBS) of the patient becomes the telling factor. The GBS is used to classify the severity of GI bleeding [Laursen 2015].

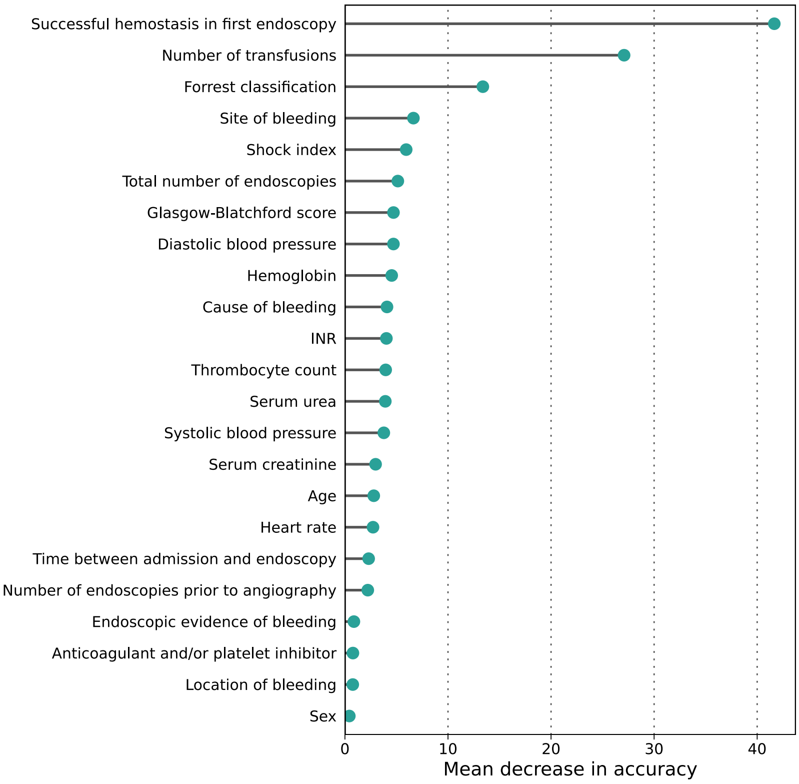

Figure 4: Modification of figure 1 from Werner [2021]. Variable importance in terms of mean decrease in accuracy of the features included in the construction of the decision trees computed using a random forest classifier with 10 000 trees and 25 iterations.

In addition, we also used random forest to compute the mean decrease in accuracy variable importance measure for the variables used in constructing the decision trees [Han 2016] (see also Method Overview). This was done to substantiate the variable choices made by the decision tree models. Mean decrease in accuracy is computed by permuting the out-of-bag (OOB) data, referring to the data not included in an individual bootstrap sample. The error rate on the OOB data is computed once and computed again after permuting each predictor variable. The difference between the two is averaged over all bootstrap samples and normalised by the standard deviation of the differences. The resulting variable importance is a positive value that increases the more influential any given variable is. The results agreed satisfactorily with the decision trees, with the success of achieving hemostasis in the primary endoscopy (binary split 1 in figure 3) and the number of transfusions (binary splits 3 and 4 in figure 3) being the two most important variables.

My contribution to this publication was the complete bioinformatic and statistical analysis.

> Werner, DJ., Baar, T., Kiesslich, R., Wenzel, N., Abusalim, N., Tresch, A., Rey, JW. (2021).
>
> Endoscopic hemostasis makes the difference: Angiographic treatment in patients with lower gastrointestinal bleeding
>
> *World J Gastrointest Endosc, 13(7): 221-232*

4 Theragnosis Biomarkers in Lung Cancer

This study investigates the connection between rearrangement of the anaplastic lymphoma kinase (*ALK*) and *TP53* mutations in human non-small cell lung cancer.

Lung cancer, one of the main causes of death in humans [Siegel 2018], is traditionally divided into two types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), which constitutes over 80 % of all cases [Reck 2014]. However, NSCLC has proven to be too diverse and is now treated as a collection of many different cancer types that each require individual treatment regimes [Boolell 2015]. One of these is *ALK+* lung cancer, in which the *ALK* gene breaks and fuses with other genes [Holla 2017].

The *ALK* gene encodes a receptor tyrosine kinase that is only expressed in early embryonic development and is involved in cell proliferation, survival, and differentiation of the nervous system [Iwahara 1997]. However, in *ALK+* lung cancer, the fusion of *ALK* to other genes causes uncontrolled activation of its downstream signalling paths through the fusion partner's promoter [Holla 2017]. In turn, tyrosine kinase inhibitors (TKI) have been shown to be an effective treatment for *ALK+* lung cancer [Kwak 2010, Reck 2014].

Gainor [2016] found that 33 % of *ALK+* tumours also show mutations in the *TP53* gene, and Aisner [2018] discovered that *TP53* mutations reduced patient survival in *ALK+* lung cancer. *TP53* is classified as a tumour suppressor gene, as it prevents genome mutation [Surget 2013]. It plays a role in cell cycle regulation and apoptosis, activating DNA repair mechanisms when damage has been sustained and halting the cell cycle until the damage is repaired. If the damage is too severe and cannot be repaired, it initiates apoptosis. *TP53* mutations are thus frequent in many cancer types, as inactivation of *TP53* severely compromises tumour suppression [Olivier 2010].

Our assumption, which we could corroborate in the publication, was that mutations in *TP53* lead to genetic instability, which in turn promote the development of resistance mechanisms, reducing patient survival rate in *ALK+* lung cancer [Alidousty 2018].

Methodology

To show that *ALK+* lung cancers with *TP53* mutations do exhibit genetic instability, we examined tumour tissue samples from a total of 423 patients originating in routine molecular diagnostics. However, depending on the used laboratory procedure, not all samples were eligible for analysis.

Bulk, panel-based DNA-seq was used to categorise the tumour samples according to the variants present in a set of genes of interest. In panel-based sequencing, a mix of PCR primers limits the analysis to selected genomic target loci. This has the benefit of increasing the coverage of these loci, as the vast majority of generated reads is focused on the panel regions. As this method could not detect large-scale genomic rearrangements, like the *ALK* translocation, it was paired with fluorescence *in situ* hybridisation (FISH). In this technique, fixed tumour tissue sections are treated with fluorescent molecular probes. The probes target multiple parts of a gene of interest, *ALK*, in this case, hybridising to the sequence's position on the respective chromosome in the nucleus. The tissue sections are then analysed under a microscope. If no rearrangement has taken place, the probes show up as a single fluorescent spot in the nucleus. However, if parts of the *ALK* gene have fused to another gene on another chromosome, multiple spots become visible, as can be seen in figure 2 of the included publication. FISH offers the benefit of being a well- established laboratory technique in cancer diagnostics, reliable at detecting translocation events with clinical relevance. Its shortcomings are that is it a labour-intensive protocol and can only detect specific breakpoints, which is why FISH will most likely be replaced by genome-wide DNA-seq in the future [Skovgaard 2011]. Finally, immunohistochemistry (IHC) antibody staining was used to detect the presence of the TP53 protein in the tissue samples. As *TP53* is only active in early embryonic development, its presence in tumour cells indicates aberrant *TP53* expression.

Nanostring

A fraction of the tissue samples were also analysed using the NanoString nCounter platform [Geiss 2008, Tsang 2017], as described in Kim [2016]. This technology uses fluorescent probes to generate direct counts of DNA molecules in tissue samples. First, a mixture of probes is added to the extracted DNA. Each probe targets a specific sequence unique to a gene of interest so that each molecule containing that sequence is hybridised to a probe. The probes carry unique fluorescent barcodes composed of six spots, each spot being one of four colours. The barcodes are read by the instrument through automated fluorescent microscopy and counted, generating raw counts that report the physical number of DNA molecules containing the sequence of interest on the instrument's slide. This technique has several advantages over sequencing-based technologies, mostly its robustness with regard to fixed samples and the reproducibility of its count data, as no amplification bias is introduced. However, like in panel-based DNA sequencing, the genes that can be analysed are limited by the used probes, of which there are no more than 800 per panel (assuming that each spot needs to be a different colour from its predecessor leaves $4 \cdot 3^5 = 972$ possible barcodes, of which some are needed for quality control and normalisation purposes).

Cell Culture

Finally, the results obtained from the tumour tissue samples were furthermore supplemented by analysing three different *ALK+* human lung cancer cell lines with different *TP53* statuses. Using cell culture samples has the benefit that no fixation procedure is applied and that the amount of available genetic material is unlimited for all practical concerns. This allowed for a chromatin immunoprecipitation DNA sequencing (ChIPseq) analysis to be performed on the cell culture samples. ChIP-seq limits bulk DNA-seq to regions of the genome bound by specific proteins [Park 2009]. First, proteins bound to the genome are chemically

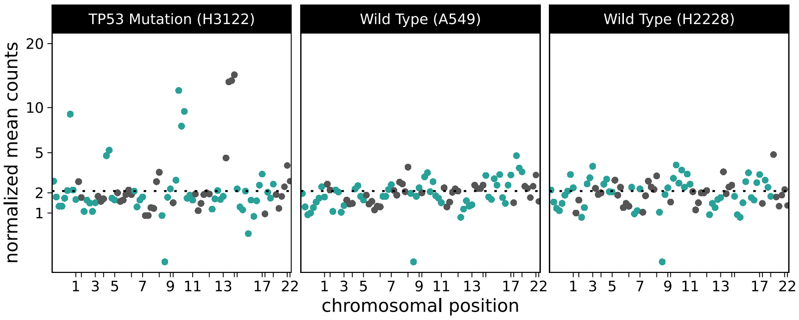

Figure 5: Modification of figure 1c from Alidousty [2018]. Copy number plots of *ALK+* cell lines harbouring wild type *TP53* (middle and right facet) or mutated *TP53* (left facet). Copy numbers of 87 genes were determined by NanoString nCounter technology (see Methodology). Alternating colours denote chromosome boundaries.

immobilised to prevent detachment, and the DNA is fragmented. Then, antibodies are used to pull out only those DNA fragments bound by specific proteins. Finally, these fragments are sequenced and mapped to the genome. In our study, we used ChIP-seq to ascertain the binding of MYC [Dang 2012], a transcription factor that, if overexpressed due to increased *MYC* copy numbers, grants *ALK+ TP53*-mutated tumour cells a proliferative advantage over their wild-type counterpart. We could observe this by monitoring the growth of cell cultures in which we induced transient *MYC* overexpression.

Analysis

Of the three publications included in this dissertation, this is the one with the most straightforward analysis. The central question was if *ALK+* tumour cells with a *TP53* mutation show higher genetic instability than those without. Genetic instability, in this case, was defined as higher variability in the copy number of genes. In diploid organisms like humans, autosomal genes have two copies each, one for each chromosome [Hartl 2009]. Every deviation from a copy number of 2 can thus be seen as a copy number alteration.

In this case, the choice of laboratory protocol simplified the analysis. As NanoString nCounter technology was employed (see Methodology), the data sets consisted of count matrices giving the number of fluorescent probe detections, automatically normalised between samples through internal controls. Without a prior PCR reaction, no amplification biases could be introduced. We thus chose to use the Brown-Forsythe test for the equality of variances [Brown 1974] (see also Method Overview), which confirmed the apparent increased genetic instability of *ALK+ TP53*-mutated cells, as exemplified in figure 5.

My contribution to this publication's bioinformatic and statistical analysis part was the copy number analysis, which facilitates the study's central finding.

> Alidousty, C., Baar, T., Martelotto, L. G., Heydt, C., Wagener, S., Fassunke, J., Duerbaum, N., Scheel, A. H., Frank, S., Holz, B., Binot, E., Kron, A., Merkelbach-Bruse, S., Ihle, M. A., Wolf, J., Buettner, R., Schultheis, A. M. (2018).
>
> Genetic instability and recurrent MYC amplification in ALK-translocated NSCLC; a central role of TP53 mutations
>
> *J Pathol (July):67–76.*

5 Conclusion

As the three included publications demonstrate, the methods to analyse biological data can be as varied as the data itself. In the following, some more general insights gained after the conclusion of the individual projects are discussed.

Evolution Shapes the Alu RNA Metabolism

Of the three included publications, this project is the methodologically most complex one, as it essentially addresses four separate questions based on the same RNA-seq data. This shows the breadth of possible approaches that can be taken when analysing biological data. The data was used to investigate general expression patterns, differential expression under Pol-II inhibition, and sequence features, too.

We also looked into different ways to answer whether Alu elements are transcribed by Pol-II or Pol-III, but these did not result in conclusive answers. We examined chromatin immunoprecipitation sequencing (ChIPseq), trying to correlate Pol-II and Pol-III peaks with Alu expression, but the resolution of the data we could obtain was not enough to draw any solid conclusions. We also tried to exploit the 5'-cap structure to differentiate transcripts, but this would have required a different experimental setup and did not appear promising in the first place. Finally, we also considered using genomic run-on sequencing (GROseq), which limits the sequencing to nascent RNAs that are currently transcribed by a polymerase, but at the time of data collection, the method was not developed enough to generate both Pol-II and Pol-III data of sufficient quality [Gardini 2017].

To continue the investigation into the life cycle of Alu elements, the main point left unanswered is the individual transcript- or loci-level origin of Alu RNAs. Three strategies present themselves but would require new experiments. Firstly, synthetic inhibitors specific to Pol-II or Pol-III have been developed that work more efficiently than α-amanitin. These could be used in conjunction with deeper RNA-seq to perform a high-resolution differential expression analysis with less biological noise caused by the long incubation time required by α-amanitin. Secondly, GRO-seq could be used to ascertain the origin of individual Alu transcripts. Thirdly, an *in vitro* experiment could be used, combining selected Alu DNA fragments with either Pol-II or Pol-III, observing which of the fragments are targeted by which polymerase.

Angiography for Gastrointestinal Bleeding

This project demonstrates that the goal is not always to choose the method that can model the data most accurately. During the course of the investigation, we discussed several potential candidates, such as GLMs and random forests (see Method Overview). In retrospect, these techniques could have resulted in a model closer to the optimal balance between bias and variance (see Introduction). However, these techniques would also have lacked the interpretability offered by a simple decision tree.

As our goal was to construct a decision-making aid, easy to use in the daily clinic routine, it turned out that, in the end, slight losses in model accuracy and specificity were acceptable trade-offs for improved practicability.

Another method we could have applied that might have resulted in a similarly interpretable model are born-again tree ensembles [Sagi 2020, Vidal 2020]. This method is proven to transform a random forest model back into a single minimal-size decision tree, the born-again (BA) tree, with the optimal number of leaves and a faithful feature space representation. The underlying algorithm was tested on different data sets, including medical data. However, a high number of features can lead to severe decreases in performance, and an upper limit of 20 features is recommended by the authors. Also, the BA tree can still be very complex with over 1000 leaves, which would again make the method impractical for the daily clinic routine. While pruning of BA trees is implemented to simplify the final result, the accuracy of the final BA tree is then no longer guaranteed. Still, testing showed only negligible losses in accuracy. Should this project be continued, possible with a much larger data set, BA trees could be a good model choice.

Finally, this project also showed the importance of data cleaning and validation, and how essential it can be to set up analyses in a reproducible manner. Throughout the investigation, the data set had to be amended and corrected numerous times, as each new descriptive report revealed new inconsistencies and errors in the original records, which is not surprising considering that the data was collected over a ten-year period. Would the analyses have needed to be re-run manually, surely this publication would have spend a few more months in preparation.

Theragnosis Biomarkers in Lung Cancer

This investigation exemplifies how important the choice of the best-suited measurement technology is. Would we not have used the NanoString nCounter method (see page 26) but more conventional bulk RNA-seq, the analysis would most likely have been much more complicated. Copy number variant detection from RNA-seq was, at least at the time of the publication of our study, still very unreliable and largely impossible with panel-based sequencing, which is common in the clinical environment. Since then, advances have been made, and tools like CaSpER could present an alternative route to take in future investigations [Serin Harmanci 2020].

In conclusion, this dissertation shows that the analysis of high-dimensional, biological data, and regression, in particular, is a broad field with many forks in the road. While first and foremost an exact scientific discipline, of course, choosing the proper method to answer a research question is also an art.

Bibliography

Aisner 2018 Aisner, D. L., Sholl, L. M., Berry, L. D., Rossi, M. R., Chen, H., Fujimoto, J., Moreira, A. L., Ramalingam, S. S., Villaruz, L. C., and Otterson, G. A. (2018). The impact of smoking and TP53 mutations in lung adenocarcinoma patients with targetable mutations–The Lung Cancer Mutation Consortium (LCMC2). *Clinical Cancer Research*, 24(5):1038–1047.

Alidousty 2018 Alidousty, C., Baar, T., Martelotto, L. G., Heydt, C., Wagener, S., Fassunke, J., Duerbaum, N., Scheel, A. H., Frank, S., Holz, B., et al. (2018). Genetic instability and recurrent *MYC* amplification in *ALK*-translocated NSCLC; a central role of *TP53* mutations. *The Journal of Pathology*, 246(July):67–76.

Altenbuchinger 2017 Altenbuchinger, M., Rehberg, T., Zacharias, H. U., Stämmler, F., Dettmer, K., Weber, D., Hiergeist, A., Gessner, A., Holler, E., Oefner, P. J., et al. (2017). Reference point insensitive molecular data analysis. *Bioinformatics*, 33(2):219–226.

An 2004 An, H. J., Lee, D., Lee, K. H., and Bhak, J. (2004). The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3′ untranslated regions. *BMC Genomics*, 5(1):1–5.

Andersson 1979 Andersson, L. C., Nilsson, K., and Gahmberg, C. G. (1979). K562–a human erythroleukemic cell line. *International journal of cancer*, 23(2):143–147.

Armstrong 2020 Armstrong, L. (2020). *Epigenetics*. Garland Science.

Aslam 2007 Aslam, J. A., Popa, R. A., and Rivest, R. L. (2007). On estimating the size and confidence of a statistical audit. In *EVT 2007 - 2007 USENIX/ACCURATE Electronic Voting Technology Workshop*.

Baar 2022 Baar, T., Dümcke, S., Gressel, S., Schwalb, B., Dilthey, A., Cramer, P., and Tresch, A. (2022). RNA transcription and degradation of Alu retrotransposons depends on sequence features and evolutionary history. *G3 Genes|Genomes|Genetics*, page jkac054.

Ballard 1982 Ballard, D. H. D. H. and Brown, C. M. (1982). *Computer vision*. Prentice-Hall.

Balyan 2017 Balyan, R., McCarthy, K. S., and McNamara, D. S. (2017). Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension. *Grantee Submission*.

Batzer 1996 Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., Schmid, C. W., Ziętkiewicz, E., and Zuckerkandl, E. (1996). Standardized nomenclature for Alu repeats. In *Journal of Molecular Evolution*, volume 42, pages 3–6.

Bellman 1961 Bellman, R., Bellman, R. E., and Collection, K. M. R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library. Princeton University Press.

Bellman 1957 Bellman, R., Corporation, R., and Collection, K. M. R. (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press.

Boolell 2015 Boolell, V., Alamgeer, M., Watkins, D. N., and Ganju, V. (2015). The Evolution of Therapies in Non-Small Cell Lung Cancer. *Cancers*, 7(3):1815–1846.

Bousquet 2011 Bousquet, O., von Luxburg, U., and Rätsch, G. (2011). *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, volume 3176. Springer.

Breiman 1996 Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman 2001 Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown 2018 Brown, A. W., Kaiser, K. A., and Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, 115(11):2563–2570.

Brown 1974 Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367.

Chen 2017 Chen, L. L. and Yang, L. (2017). ALUternative Regulation for Gene Expression.

Chen 2016 Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 785–794. Association for Computing Machinery.

Chikhi 2014 Chikhi, R., Limasset, A., Jackman, S., Simpson, J. T., and Medvedev, P. (2014). On the representation of de bruijn graphs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8394 LNBI, pages 35–55. Springer, Cham.

Conti 2015 Conti, A., Carnevali, D., Bollati, V., Fustinoni, S., Pellegrini, M., and Dieci, G. (2015). Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Research*, 43(2):817–835.

Cordaux 2009 Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution.

Cortes 1995 Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dagan 2004 Dagan, T., Sorek, R., Sharon, E., Ast, G., and Graur, D. (2004). AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic acids research*, 32(Database issue):D489–92.

Dang 2012 Dang, C. V. (2012). MYC on the path to cancer. *Cell*, 149(1):22–35.

De Bruijn 1946 De Bruijn, N. G. (1946). A combinatorial problem. In *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, volume 49, pages 758–764.

Deininger 2011 Deininger, P., Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2011). Alu elements: know the SINEs. *Genome Biology*, 12(12):236.

Deininger 1999 Deininger, P. L. and Batzer, M. A. (1999). Alu repeats and human disease. *Molecular Genetics and Metabolism*, 67(3):183–193.

Deisenroth 2020 Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2020). *Mathematics for Machine Learning*. Cambridge University Press.

Doyle 2012 Doyle, A. C. and Oakley, J. (2012). *The Complete Sherlock Holmes*. Garden City, N.Y. : Doubleday & Co.

Efron 1994 Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. CRC press.

Evangelista 2000 Evangelista, P. T. and Hallisey, M. J. (2000). Transcatheter embolization for acute lower gastrointestinal hemorrhage. *Journal of Vascular and Interventional Radiology*, 11(5):601–606.

Evgen'ev 2007 Evgen'ev, M. B. (2007). Mobile elements and genome evolution. *Molecular Biology*, 41(2):203–213.

Fisher 1922 Fisher, R. A. (1922). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87.

Fox 2002 Fox, J. and Weisberg, S. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, 2002.

Freund 1995 Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285.

Freund 1997 Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman 2010 Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Gainor 2016 Gainor, J. F., Dardaei, L., Yoda, S., Friboulet, L., Leshchiner, I., Katayama, R., Dagogo-Jack, I., Gadgeel, S., Schultz, K., and Singh, M. (2016). Molecular mechanisms of resistance to first-and second-generation ALK inhibitors in ALK-rearranged lung cancer. *Cancer discovery*, 6(10):1118–1133.

Gardini 2017 Gardini, A. (2017). Global run-on sequencing (GRO-Seq). In *Methods in Molecular Biology*, volume 1468, pages 111–120. Humana Press Inc.

Geiss 2008 Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., and Grogan, T. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3):317–325.

Gelman 2014 Gelman, A., Carlin, J. B. B., Stern, H. S. S., and Rubin, D. B. B. (2014). Bayesian Data Analysis, Third Edition. *Book*, page 675.

Good 1946 Good, I. J. (1946). Normal recurring decimals.

Gøtzsche 1989 Gøtzsche, P. C. (1989). Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled clinical trials*, 10(1):31–56.

Gressel 2019 Gressel, S., Lidschreiber, K., and Cramer, P. (2019). Transient transcriptome sequencing : experimental protocol to monitor genome-wide RNA synthesis including enhancer transcription. *protocols.io*, pages 1–20.

Guo 2017 Guo, H., Hu, B., Yan, L., Yong, J., Wu, Y., Gao, Y., Guo, F., Hou, Y., Fan, X., Dong, J., et al. (2017). DNA methylation and chromatin accessibility profiling of mouse and human fetal germ cells. *Cell Research*, 27(2):165–183.

Han 2016 Han, H., Guo, X., and Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th ieee international conference on software engineering and service science (icsess)*, pages 219–224. IEEE.

Hartl 2009 Hartl, D. L. (2009). *Essential Genetics: A Genomics Perspective: A Genomics Perspective*. Jones & Bartlett Publishers.

Hastie 2021 Hastie, T., Tibshirani, R., James, G., and Witten, D. (2021). An Introduction to Statistical Learning (2nd Edition). *Springer Texts*, 102(1998):618.

Hawkins 1980 Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Herawati 2018 Herawati, N., Nisa, K., and Setiawan, E. (2018). Regularized Multiple Regression Methods to Deal with Severe Multicolinearity. *International Journal of Statistics and Applications*, 8(May 2018):167–172.

Hinton 2003 Hinton, G. and Roweis, S. (2003). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*.

Hinton 1999 Hinton, G. and Sejnowski, T. J. (1999). Unsupervised Learning: Foundations of Neural Computation. *MIT Press*.

Ho 1995 Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Hoerl 1970 Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Holla 2017 Holla, V. R., Elamin, Y. Y., Bailey, A. M., Johnson, A. M., Litzenburger, B. C., Khotskaya, Y. B., Sanchez, N. S., Zeng, J., Shufean, M. A., and Shaw, K. R. (2017). ALK: a tyrosine kinase target for cancer therapy. *Molecular Case Studies*, 3(1):a001115.

Holley 2020 Holley, G. and Melsted, P. (2020). Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome biology*, 21(1):249.

Hopcroft 2001 Hopcroft, J. E., Motwani, R., and Ullman, J. D. (2001). Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32(1):60–65.

Hugenholtz 2008 Hugenholtz, P. and Tyson, G. W. (2008). Metagenomics. *Nature*, 455(7212):481–483.

Iwahara 1997 Iwahara, T., Fujimoto, J., Wen, D., Cupples, R., Bucay, N., Arakawa, T., Mori, S., Ratzkin, B., and Yamamoto, T. (1997). Molecular characterization of ALK, a receptor tyrosine kinase expressed specifically in the nervous system. *Oncogene*, 14(4):439–449.

Jacob 1970 Jacob, S. T., Muecke, W., Sajdel, E. M., and Munro, H. N. (1970). Evidence for extranucleolar control of RNA synthesis in the nucleolus. *Biochemical and Biophysical Research Communications*, 40(2):334–342.

Jagadeeswaran 1981 Jagadeeswaran, P., Forget, B. G., and Weissman, S. M. (1981). Short interspersed repetitive DNA elements in eucaryotes: Transposable DNA elements generated by reverse transcription of RNA pol III transcripts?

Jagadish 2003 Jagadish, H. V. and Olken, F. (2003). Data Management for the Biosciences. Report of the NSF. In *NLM Workshop of Data Management for Molecular and Cell Biology*.

Joshi 2021 Joshi, D. J., Kale, I., Gandewar, S., Korate, O., Patwari, D., and Patil, S. (2021). Reinforcement Learning: A Survey. In *Advances in Intelligent Systems and Computing*, volume 1311 AISC, pages 297–308. Springer Science and Business Media Deutschland GmbH.

Kamiński 2018 Kamiński, B., Jakubczyk, M., and Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research*, 26(1):135–159.

Kearns 1994 Kearns, M. and Valiant, L. (1994). Cryptographic Limitations on Learning Boolean Formulae and Finite Automata. *Journal of the ACM (JACM)*, 41(1):67–95.

Kedinger 1970 Kedinger, C., Gniazdowski, M., Mandel, J. L. J., Gissinger, F., and Chambon, P. (1970). Alpha-amanitin: a specific inhibitor of one of two DNA-pendent RNA polymerase activities from calf thymus. *Biochemical and biophysical research communications*, 38(1):165–171.

Kim 2014 Kim, B. S. M. (2014). Diagnosis of gastrointestinal bleeding: A practical guide for clinicians. *World Journal of Gastrointestinal Pathophysiology*, 5(4):467.

Kim 2017 Kim, P. H., Tsauo, J., Shin, J. H., and Yun, S.-C. (2017). Transcatheter arterial embolization of gastrointestinal bleeding with N-butyl cyanoacrylate: a systematic review and meta-analysis of safety and efficacy. *Journal of Vascular and Interventional Radiology*, 28(4):522–531.

Kim 2016 Kim, S. T., Lee, S. J., Park, S. H., Park, J. O., Lim, H. Y., Kang, W. K., Lee, J., and Park, Y. S. (2016). Genomic profiling of metastatic gastroenteropancreatic neuroendocrine tumor (GEP-NET) patients in the personalized-medicine era. *Journal of Cancer*, 7(9):1044.

Kim 2015 Kim, T.-K. and Shiekhattar, R. (2015). Architectural and functional commonalities between enhancers and promoters. *Cell*, 162(5):948–959.

Kleinbaum 2002 Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). *Logistic regression*. Springer.

Kramer 1991 Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

Kriegs 2007 Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. (2007). Evolutionary history of 7SL RNA-derived SINEs in Supraprimates.

Kwak 2010 Kwak, E. L., Bang, Y.-J., Camidge, D. R., Shaw, A. T., Solomon, B., Maki, R. G., Ou, S.-H. I., Dezube, B. J., Jänne, P. A., and Costa, D. B. (2010). Anaplastic lymphoma kinase inhibition in non–small-cell lung cancer. *New England Journal of Medicine*, 363(18):1693–1703.

Lander 2001 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Laursen 2015 Laursen, S. B., Dalton, H. R., Murray, I. A., Michell, N., Johnston, M. R., Schultz, M., Hansen, J. M., de Muckadell, O. B. S., Blatchford, O., and Stanley, A. J. (2015). Performance of new thresholds of the Glasgow Blatchford score in managing patients with upper gastrointestinal bleeding. *Clinical Gastroenterology and Hepatology*, 13(1):115–121.

LeCun 2015 LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee 2018 Lee, H. H., Oh, J. S., Park, J. M., Chun, H. J., Kim, T. H., Cheung, D. Y., Lee, B.-I., Cho, Y.-S., and Choi, M.-G. (2018). Transcatheter embolization effectively controls acute lower gastrointestinal bleeding without localizing bleeding site prior to angiography. *Scandinavian journal of gastroenterology*, 53(9):1089–1096.

Li 2015 Li, C. (2015). A Gentle Introduction to Gradient Boosting. *Novosti Khirurgii*, 23(5):566–569.

Lindell 1970 Lindell, T. J., Weinberg, F., Morris, P. W., Roeder, R. G., and Rutter, W. J. (1970). Specific Inhibition of Nuclear RNA Polymerase II by α-Amanitin. *Science*, 170(3956):447–449.

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

Mann 1947 Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Mariner 2008 Mariner, P. D., Walters, R. D., Espinoza, C. A., Drullinger, L. F., Wagner, S. D., Kugel, J. F., and Goodrich, J. A. (2008). Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. *Molecular Cell*, 29(4):499–509.

Martin 2015 Martin, E. (2015). *Concise Medical Dictionary*. Oxford University Press.

McInnes 2018 McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*.

Menon 1971 Menon, I. A. (1971). Differential Effects of α-Amanitin on RNA Polymerase Activity in Nuclei and Mitochondria. *Canadian Journal of Biochemistry*, 49(12):1395–1398.

Mitchell 1997 Mitchell, T. M. T. M. (1997). Machine Learning. *McGraw-Hill*, page 414.

Mohri 2018 Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Nelder 1972 Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370.

Novak 2017 Novak, A. M., Hickey, G., Garrison, E., Blum, S., Connelly, A., Dilthey, A., Eizenga, J., Elmohamed, M. A. S., Guthrie, S., and Kahles, A. (2017). Genome graphs. *bioRxiv*, page 101378.

Oakland 2019 Oakland, K., Chadwick, G., East, J. E., Guy, R., Humphries, A., Jairath, V., McPherson, S., Metzner, M., Morris, A. J., and Murphy, M. F. (2019). Diagnosis and management of acute lower gastrointestinal bleeding: guidelines from the British Society of Gastroenterology. *Gut*, 68(5):776–789.

Olivier 2010 Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*, 2(1):a001008.

Orioli 2012 Orioli, A., Pascali, C., Pagano, A., Teichmann, M., and Dieci, G. (2012). RNA polymerase III transcription control elements: Themes and variations.

Pannatier 2019 Pannatier, M., Duran, R., Denys, A., Meuli, R., Zingg, T., and Schmidt, S. (2019). Characteristics of patients treated for active lower gastrointestinal bleeding detected by CT angiography: Interventional radiology versus surgery. *European journal of radiology*, 120:108691.

Panning 1993 Panning, B. and Smiley, J. R. (1993). Activation of RNA polymerase III transcription of human Alu repetitive elements by adenovirus type 5: requirement for the E1b 58-kilodalton protein and the products of E4 open reading frames 3 and 6. *Molecular and Cellular Biology*, 13(6):3231–3244.

Paolella 1983 Paolella, G., Lucero, M. A., Murphy, M. H., and Baralle, F. E. (1983). The Alu family repeat promoter has a tRNA-like bipartite structure. *The EMBO Journal*, 2(5):691–696.

Park 2009 Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669–680.

Pearson 1901 Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Prasad Kerlin 2013 Prasad Kerlin, M., Tokar, J. L., Cotton, D., Taichman, D., and Williams, S. (2013). Acute gastrointestinal bleeding.

Quentin 1992 Quentin, Y. (1992). Fusion of a free left alu monomer and a free right alu monometer at the origin of the alu family in the primate genomes. *Nucleic Acids Research*, 20(3):487–493.

Reck 2014 Reck, M., Popat, S., Reinmuth, N., De Ruysscher, D., Kerr, K. M., and Peters, S. (2014). Metastatic non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology : official* *journal of the European Society for Medical Oncology*, 25 Suppl 3:iii27–39.

Reid 1971 Reid, B. D. and Parsons, P. (1971). Partial Purification of Mitochondrial RNA Polymerase from Rat Liver. *Proceedings of the National Academy of Sciences*, 68(11):2830–2834.

Richard Shen 1991 Richard Shen, M., Batzer, M. A., and Deininger, P. L. (1991). Evolution of the master Alu gene(s). *Journal of Molecular Evolution*, 33(4):311–320.

Rockey 2005 Rockey, D. C. (2005). Gastrointestinal bleeding. *Gastroenterology Clinics*, 34(4):581–588.

Rokach 2006 Rokach, L. and Maimon, O. (2006). Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*, pages 321–352. Springer, Boston, MA.

Rosasco 2004 Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., and Verri, A. (2004). Are Loss Functions All the Same? *Neural Computation*, 16(5):1063–1076.

Rossi 2018 Rossi, R. J. (2018). *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons.

Saccone 1971 Saccone, C., Gallerani, R., Gadaleta, M., and Greco, M. (1971). The effect of α-amanitin on RNA synthesis in rat liver mitochondria. *FEBS Letters*, 18(2):339–341.

Sagi 2020 Sagi, O. and Rokach, L. (2020). Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*, 61:124–138.

Sainte-Marie 1894 Sainte-Marie, C. F. (1894). Solution to question nr. 48. *L'intermédiaire des Mathématiciens*, 1:107–110.

Schapire 1990 Schapire, R. E. (1990). The Strength of Weak Learnability. *Machine Learning*, 5(2):197–227.

Schapire 2009 Schapire, R. E. (2009). A Short Introduction to Boosting. *Society*, 14(5):771–780.

Schmid 1975 Schmid, C. W. and Deininger, P. L. (1975). Sequence organization of the human genome. *Cell*, 6(3):345–358.

Schwalb 2016 Schwalb, B., Michel, M., Zacher, B., Hauf, K. F., Demel, C., Tresch, A., Gagneur, J., and Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science*, 352(6290):1225–1228.

Schwalb 2012 Schwalb, B., Schulz, D., Sun, M., Zacher, B., Dümcke, S., Martin, D. E., Cramer, P., and Tresch, A. (2012). Measurement of genome-wide RNA synthesis and decay rates with Dynamic Transcriptome Analysis (DTA). *Bioinformatics*, 28(6):884–885.

Serin Harmanci 2020 Serin Harmanci, A., Harmanci, A. O., and Zhou, X. (2020). CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nature Communications*, 11(1):89.

Siegel 2018 Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30.

Skovgaard 2011 Skovgaard, O., Bak, M., Løbner-Olesen, A., and Tommerup, N. (2011). Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome research*, 21(8):1388–1393.

Smith 1976 Smith, H. O. and Birnstiel, M. L. (1976). A simple method for DNA restriction site mapping. *Nucleic acids research*, 3(9):2387–2398.

Stamp 2018 Stamp, M. (2018). A Revealing Introduction to Hidden Markov Models. *Department of Computer Science San Jose State University*, pages 26–56.

Stirpe 1967 Stirpe, F. and Fiume, L. (1967). Studies on the pathogenesis of liver necrosis by alpha-amanitin. Effect of alpha-amanitin on ribonucleic acid synthesis and on ribonucleic acid polymerase in mouse liver nuclei. *The Biochemical journal*, 105(2):779–782.

Strate 2010 Strate, L. L. and Naumann, C. R. (2010). The role of colonoscopy and radiological procedures in the management of acute lower intestinal bleeding. *Clinical Gastroenterology and Hepatology*, 8(4):333–343.

Sun 2012 Sun, J. Z., Wang, G. I., Goyal, V. K., and Varshney, L. R. (2012). A framework for Bayesian optimality of psychophysical laws. *Journal of Mathematical Psychology*, 56(6):495–501.

Surget 2013 Surget, S., Khoury, M. P., and Bourdon, J.-C. (2013). Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *OncoTargets and therapy*, 7:57–68.

Tabak 2014 Tabak, J. (2014). *Geometry: the language of space and form*. Infobase Publishing.

Tan 2008 Tan, K.-K., Wong, D., and Sim, R. (2008). Superselective embolization for lower gastrointestinal hemorrhage: an institutional review over 7 years. *World journal of surgery*, 32(12):2707–2715.

Teunissen 2006 Teunissen, P. J. G. (2006). *Testing theory*. VSSD Delft.

Tibshirani 1996 Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Treuting 2018 Treuting, P. M., Arends, M. J., and Dintzis, S. M. (2018). Lower gastrointestinal tract. In *Comparative Anatomy and Histology*, pages 213–228. Elsevier.

Tsang 2017 Tsang, H.-F., Xue, V. W., Koh, S.-P., Chiu, Y.-M., Ng, L. P.-W., and Wong, S.-C. C. (2017). NanoString, a novel digital color-coded barcode technology: current and future applications in molecular diagnostics. *Expert Review of Molecular Diagnostics*, 17(1):95–103.

Turczyk 2020 Turczyk, B. M., Busby, M., Martin, A. L., Daugharthy, E. R., Myung, D., Terry, R. C., Inverso, S. A., Kohman, R. E., and Church, G. M. (2020). Spatial sequencing: a perspective. *Journal of Biomolecular Techniques: JBT*, 31(2):44.

Valiant 1984 Valiant, L. G. (1984). A theory of the learnable. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, pages 436–445.

Van Der Maaten 2009 Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13.

Venter 2004 Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., and Nelson, W. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *science*.

Vidal 2020 Vidal, T. and Schiffer, M. (2020). Born-Again tree ensembles. In *37th International Conference on Machine Learning, ICML 2020*, volume PartF16814, pages 9685–9695. International Machine Learning Society (IMLS).

von Luxburg 2011 von Luxburg, U. and Schölkopf, B. (2011). Statistical Learning Theory: Models, Concepts, and Results. *Handbook of the History of Logic*, 10:651–706.

Wang 2013 Wang, J., Bao, Y. X., Bai, M., Zhang, Y. G., Xu, W. D., and Qi, X. S. (2013). Restrictive vs liberal transfusion for upper gastrointestinal bleeding: A meta-analysis of randomized controlled trials. *World Journal of Gastroenterology*, 19(40):6919–6927.

Wang 2019 Wang, P., Bai, G. R., and Stolee, K. T. (2019). Exploring regular expression evolution. In *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 502–513. IEEE.

Wang 2015 Wang, Y. and Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Molecular cell*, 58(4):598–609.

Werner 2021 Werner, D. J., Baar, T., Kiesslich, R., Wenzel, N., Abusalim, N., Tresch, A., and Rey, J. W. (2021). Endoscopic hemostasis makes the difference: Angiographic treatment in patients with lower gastrointestinal bleeding. *http://www.wjgnet.com/*, 13(7):221–232.

Werner 2018 Werner, D. J., Manner, H., Nguyen-Tat, M., Kloeckner, R., Kiesslich, R., Abusalim, N., and Rey, J. W. (2018). Endoscopic and angiographic management of lower gastrointestinal bleeding: Review of the published literature. *United European gastroenterology journal*, 6(3):337–342.

White 1997 White, R. J. (1997). Regulation of RNA polymerases i and III by the retinoblastoma protein: A mechanism for growth control?

Wilcoxon 1945 Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80.

Winters 2010 Winters, R., Winters, A., and Amedee, R. G. (2010). Statistics: a brief overview. *The Ochsner journal*, 10(3):213–216.

Wooley 2006 Wooley, J. C. and Lin, H. S. (2006). *Catalyzing Inquiry at the Interface of Computing and Biology*. National Academies Press.

Wu 2008 Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

Yamada 2015 Yamada, A., Niikura, R., Yoshida, S., Hirata, Y., and Koike, K. (2015). Endoscopic management of colonic diverticular bleeding. *Digestive Endoscopy*, 27(7):721–726.

Zhang 2019 Zhang, X. O., Gingeras, T. R., and Weng, Z. (2019). Genome-wide analysis of polymerase III–transcribed Alu elements suggests cell-type–specific enhancer function. *Genome Research*, 29(9):1402–1414.

Zou 2005 Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

List of Figures

Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten – noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Till Baar

Lebenslauf

Till Baar

28.12.1987, Hamburg (Deutsche Staatsangehörigkeit)

Steinstraße 32, 52372 Kreuzau

Schulische Ausbildung . . . .

1994 – 1998	Grundschule Schortens
1998 – 2000	Franziskusschule Wilhelmshaven (OS)
2000 – 2007	Cäcilienschule Wilhelmshaven freies Gymnasium in kirchl. Trägerschaft (Abschluss: Abitur)

. . . .

2007 – 2008	Zivildienst, FöJ Tierpark Bochum

Universitäre Ausbildung . . . .

2008 – 2009	Bachelor Studiengang Geophysik / Meteorologie, Universität zu Köln
2009 – 2012	Bachelor Studiengang Biologie, Universität zu Köln (Abschluss: B.Sc.)
2013 – 2015	Master Studiengang Biological Sciences, Universität zu Köln (Abschluss: M.Sc.)
2015 – 2017	Wiss. Mitarbeiter Molekularpathologie, Institut für Pathologie, Universitätsklinikum Köln
2017 – 2022	Wiss. Mitarbeiter AG Computational Biology, IMSB, Fakulät für Medizin, Universität zu Köln (Promotion)

Till Baar

Acknowledgements

Throughout my doctoral studies, I have received a great deal of support and assistance.

I would first and foremost like to thank my supervisor, Prof. Dr. Achim Tresch, whose expertise was invaluable in formulating my research questions and methodology. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. He was instrumental in defining the path of my research, and for his guidance through each stage of my doctoral studies, I am immensely grateful.

My heartfelt thanks also go out to the many excellent scientists that collaborated with me on the publications included in my dissertation.

I would also like to thank my colleagues at the Institute of Medical Statistics and Computational Biology, and the Computational Biology workgroup in particular, for their wonderful fellowship. The companionable atmosphere in our office makes every day I spend with them a pleasure.

Furthermore, without the support of Prof. Dr. Michael Melkonian, Prof. Dr. Günter Plickert, and Dr. Lisa Stephan in my early course of studies, I would not have made it this far.

Moreover, I would like to thank my parents, who always believed in my ability to be successful in academia. None of this would have been possible without them and their unending encouragement, for which I am grateful beyond words. I know that they could hardly be any prouder of me, but I hope they take delight in knowing that my small contributions to the sum total of human knowledge bear their subtle brushstrokes, too.

Finally, I could not have completed my dissertation without the support of all of my friends, especially Peter and Dave, who provided invaluable support, stimulating discussions, as well as the essential distraction to rest my mind outside of research.

I thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded (Funding number: INST 216 / 512 / 1FUGG) High Performance Computing (HPC) system CHEOPS as well as support.