TOPOLOGY OF GENEALOGICAL TREES - THEORY AND APPLICATION

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
MARTINA RAUSCHER
aus Bangkok

Köln, 2018

Berichterstatter: Prof. Dr. Thomas Wiehe
Prof. Dr. Michael Nothnagel

Tag der letzten mündlichen Prüfung: 26. Oktober 2018

Acknowledgements

I owe a debt of gratitude to many people who helped and supported me during the last years to achieve my accomplishments.
First of all, I would like to thank my supervisor Prof. Dr. Thomas Wiehe for his patient guidance. I am very grateful not only for his support over the last years, but also for all his advices, knowledge and inspiring ideas for my research project.
My thanks also go to all members of the Wiehe lab, in particular to Jaanus Suurväli, Johannes Wirtz and Yichen Zheng for helpful comments on the manuscript, and to Robert Fürst for always helping me with his technical knowledge when I was facing challenges in programming problems.
I would like to thank my Getti, who had an important influence on my academic path.
I am very grateful to my family. Special thanks go to my parents, who had always been supportive to every decision I have made in my life and have always been there for me.
Finally, I want to express my deepest gratitude to my husband Sebastian, without whom this would not have been possible. His continuous support encouraged me in all of my pursuits.

"Intuition is a human fallacy, the belief that you can predict random events."

Seven of Nine, Star Trek

Zusammenfassung

In der Populationsgenetik ist es vor allem von Interesse, genetische Daten einer Populationsstichprobe zu analysieren und zu verstehen. Hierbei spielt die Koaleszenz Theorie eine wichtige Rolle. Die Koaleszenz Theorie basiert auf der Idee, die genealogischen Eigenschaften einer Population anhand von Datensätzen einer gegenwärtigen Stichprobe von Individuen rückwärts in der Zeit zu analysieren. Wenn bei dieser Rückwärtsbetrachtung zwei Individuen einen gemeinsamen Vorfahren haben, werden diese zusammengefasst, das heißt sie verschmelzen. Grafisch lässt sich das durch einen Baum darstellen. Mit Hilfe dieser Bäume ist es möglich, nicht nur genetische Beziehungen oder Substrukturierung von Populationen zu erkennen, sondern auch Hinweise auf positive Selektion zu erkennen. Der Grundgedanke hierzu beruht darauf, dass sich Loci unter selektiven Einflüssen anders verhalten als Loci unter neutralen Bedingungen. Wenn eine neu aufgetretene Mutation mit Selektionsvorteil in einer Population fixiert wird, steigt nicht nur deren Allelhäufigkeit, sondern auch die Allelhäufigkeit von neutralen Regionen, die mit dem selektierten Locus gekoppelt sind. Als Resultat dieses sogenannten 'Hitchhiking-Effekt' weist die Region in der Umgebung des selektierten Locus eine signifikante Reduktion der genetischen Variabilität auf im Vergleich zu Regionen unter neutralem Einfluss. Dies wirkt sich auf die Topologie des genealogischen Baumes aus. Eine Reduktion der genetischen Variabilität verursacht durch eine positive Selektion wird 'selective sweep' genannt. Den Umstand nutzend, dass 'selective sweeps' extrem unbalancierte genealogische Baum-Topologien in der Umgebung des selektierten Locus erzeugen können, leiten wir daraus einen neuen statistischen Test, basierend auf einer Log-Likelihood-Methode und aufbauend auf dem bereits bekannten T_3-Test, her: den LR_{T_3}-Test. Der Vorteil an statistischen Methoden, die nur die Information der zugrundeliegenden genealogischen Baum-Topologie benötigen, liegt darin, dass diese nicht durch Schwankungen in der Populationsgröße beeinflusst werden. Wir haben alle 26 Populationen des Phase-3-Datensatzes des 1,000 Genome-Projektes mit dem LR_{T_3} Test untersucht, um Kandidatenregionen für positive Selektion zu identifizieren. Darüber hinaus stellen wir ein Maß für die Korrelation von Chromosom-Segmenten an verschiedenen Chromosom-Positionen vor, welches anhand der zu Grunde liegenden genealogischen Baum-Topologie bestimmt werden kann. Auch hierfür werden wir eine praxisorientierte Anwendung anhand der humanen Daten demonstrieren.

Abstract

One of the major interests in population genetics is how genetic variation within and among populations can be explained by evolutionary forces such as natural selection. It is known that recent events of positive selection can leave a specific pattern of polymorphism surrounding the selected site. As a new beneficial mutation arises in a population and eventually becomes fixed, also neutral variants linked to the selected site will increase in frequency. This leads to a reduction of genetic diversity around the selected site, a process known as 'selective sweep'. Still today, identifying loci, which underwent recent selective sweeps is a difficult task, since traces are typically obscured by other evolutionary and demographic factors, such as genetic drift or population bottleneck events. Therefore, several methods have been developed to reliably detecting genomic patterns left by the action of positive selection. The representation of evolutionary history of a sample as a tree is an elementary approach in population genetics. The process in which two lineages merge at a common ancestor, when going back in time, is known as a coalescent event. To detect candidate loci of selective sweeps, we take here an approach which considers the genealogical relationships among individuals and the topological properties of the inferred coalescent tree. Selective sweeps can produce highly unbalanced coalescent tree topologies in region close to a selective sweep site. Building on a previously known test statistic called T_3, which detects bias in the balance of binary genealogical trees, we derive a new test statistic based on a log likelihood approach and we call it the LR_{T_3}-test.

We present the results of genome wide screens of the LR_{T_3}-test applied to the 26 populations of the phase 3 data set of the human 1,000 genomes project. Furthermore, we present a measure of topological linkage disequilibrium (tLD), which is based on clustering individuals with respect to their position in the genealogy rather than clustering alleles and haplotypes. We demonstrate its application to the beforehand processed human data.

Contents

List of Abbreviations

α	population scaled selection coefficient
avg.	Average
BED	Browser Extensible Data format
bp	Base pair(s)
c	Recombination rate per bp per generation
CHR/chr	Chromosome
cM	CentiMorgan
CONT	Continent
dbPSHP	Database of recent positive selection across human populations
Gb/gb	Giga Base Pairs
GO	Gene ontology
GOrilla	Gene ontology enrichment analysis and visualization tool
kb	Kilo Base Pairs
LD	Linkage Disequilibrium
μ	Mutation rate per bp per generation
Mb/mb	Mega Base Pairs
MRCA	Most Recent Common Ancestor
n	Sample size
PAR	Pseudoautosomal region
N	Population size
ρ	$= 4Nc$, population scaled recombination rate (Chapter 5)
s	Selection coefficient, where $(1 + s)$ is the relative fitness of the selected over the ancestral allele (assuming co-dominance h=0.5)
SFS	Site Frequency Spectrum
S/HIC	Soft/Hard Inference through Classification
SNP	Single Nucleotide Polymorphism
ss	Segregating Site(s)
r	$= 4Nc$, population scaled recombination rate

$\boldsymbol{\theta}$	$= 4N\mu$, population scaled mutation rate
tLD	Topological Linkage Disequilibrium
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
UV	Ultra-violet
vcf	Variant call format

Population

ACB	African Caribbean in Barbados
AFR	African superpopulation
AMR	Admixed American superpopulation
ASW	Americans of African Ancestry in Southwest USA
BEB	Bengali from Bangladesh
CDX	Chinese Dai in Xishuangbanna, China
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
CHB	Han Chinese in Beijing, China
CHS	Southern Han Chinese, China
CLM	Colombians from Medellin, Colombia
EAS	East Asian superpopulation
EUR	European superpopulation
ESN	Esan in Nigeria
FIN	Finnish in Finland
GBR	British in England and Scotland
GIH	Gujarati Indian from Houston, Texas
GWD	Gambian in Western Divisions in the Gambia
IBS	Iberian Population in Spain
ITU	Indian Telugu from the UK
JPT	Japanese in Tokyo, Japan
KHV	Kinh in Ho Chi Minh City, Vietnam
LWK	Luhya in Webuye, Kenya
MSL	Mende in Sierra Leone
MXL	Mexican Ancestry from Los Angeles USA
PEL	Peruvians from Lima, Peru
PJL	Punjabi from Lahore, Pakistan
PUR	Puerto Ricans from Puerto Rico
SAS	South Asian superpopulation
STU	Sri Lankan Tamil from the UK
TSI	Toscani in Italia

YRI	Yoruba in Ibadan, Nigeria

Gene Names (appearing in chapter 4)

AF131215.5	AF131215.5
ANXA7	Annexin A7
ARHGEF38	Rho Guanine Nucleotide Exchange Factor 38
ATP6V1D	ATPase H+ Transporting V1 Subunit D
BEND4	BEN Domain Containing 4
BEX5	Brain Expressed X-Linked 5
C1orf185	Chromosome 1 Open Reading Frame 185
CASK	Calcium/Calmodulin Dependent Serine Protein Kinase
CCDC138	Coiled-Coil Domain Containing 138
CHRNA6	Cholinergic Receptor Nicotinic Alpha 6 Subunit
CMSS1	Cms1 Ribosomal Small Subunit Homolog (Yeast)
CNTNAP2	Contactin Associated Protein Like 2
COL8A1	Collagen Type VIII Alpha 1 Chain
DARS	Aspartyl-TRNA Synthetase
DAPK2	Death Associated Protein Kinase 2
DBT	Dihydrolipoamide Branched Chain Transacylase E2
DCAF4L1	DDB1 And CUL4 Associated Factor 4 Like 1
DNAJC9	DnaJ Heat Shock Protein Family (Hsp40) Member C9
ECD	Ecdysoneless Cell Cycle Regulator
EDAR	Ectodysplasin A Receptor
EIF2S1	Eukaryotic Translation Initiation Factor 2 Subunit Alpha
EPAS1	Endothelial PAS Domain Protein 1
EPS15	Epidermal Growth Factor Receptor Pathway Substrate 15
FAM149B1	Family With Sequence Similarity 149 Member B1
FAM71D	Family With Sequence Similarity 71 Member D
FBXL22	F-Box And Leucine Rich Repeat Protein 22
FILIP1L	Filamin A Interacting Protein 1 Like
FNTA	Farnesyltransferase, CAAX Box, Alpha
GCC2	GRIP And Coiled-Coil Domain Containing 2
GPHN	Gephyrin
GPR34	G Protein-Coupled Receptor 34

GPR82	G Protein-Coupled Receptor 82
GSTCD	Glutathione S-Transferase C-Terminal Domain Containing
HERC1	HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase Family Member
HERC2	HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 2
HGSNAT	Heparan-Alpha-Glucosaminide N-Acetyltransferase
HIAT1	Major Facilitator Superfamily Domain Containing 14A
HOOK3	Hook Microtubule Tethering Protein 3
INTS12	Integrator Complex Subunit 12
LCT	Lactase
LIMCH1	LIM And Calponin Homology Domains 1
LIMS1	LIM Zinc Finger Domain Containing 1
LRRC39	Leucine Rich Repeat Containing 39
MCM6	Minichromosome Maintenance Complex Component
MME	Membrane Metalloendopeptidase
MPP5	Membrane Palmitoylated Protein 5
MRPS16	Mitochondrial Ribosomal Protein S16
MSS51	MSS51 Mitochondrial Translational Activator
MYOZ1	Myozenin 1
NELL2	Neural EGFL Like 2
NNT	Nicotinamide Nucleotide Transhydrogenase
NXF5	Nuclear RNA Export Factor 5
OCA2	OCA2 Melanosomal Transmembrane Protein
PCDH15	Protocadherin Related 15
PHOX2B	Paired Like Homeobox 2b
PLEK2	Pleckstrin 2
POMK	Protein-O-Mannose Kinase
PPP3CB	Protein Phosphatase 3 Catalytic Subunit Beta
RALGAPA2	Ral GTPase Activating Protein Catalytic Alpha Subunit 2
RANBP2	RAN Binding Protein 2
RNF11	Ring Finger Protein 11
RNF170	Ring Finger Protein 170
RP11-598P20.5	Gene RP11-598P20.5
RTCA	RNA 3′-Terminal Phosphate Cyclase
SASS6	SAS-6 Centriolar Assembly Protein
SLC30A9	Solute Carrier Family 30 Member 9
SLC35A3	Solute Carrier Family 35 Member A3
SULT1C2	Sulfotransferase Family 1C Member 2
SULT1C4	Sulfotransferase Family 1C Member 4
SYNPO2L	Synaptopodin 2 Like

TCEAL2	Transcription Elongation Factor A Like 2
TCEAL6	Transcription Elongation Factor A Like 6
TCF7L2	Transcription Factor 7 Like 2
THAP1	THAP Domain Containing 1
TMEM117	Transmembrane Protein 117
TMEM229B	Transmembrane Protein 229B
TMEM33	Transmembrane Protein 33
TRMT13	TRNA Methyltransferase 13 Homolog
TTC18	Cilia And Flagella Associated Protein 70
TTC39A	Tetratricopeptide Repeat Domain 39A
USP3	Ubiquitin Specific Peptidase 3
USP54	Ubiquitin Specific Peptidase 54
VTI1A	Vesicle Transport Through Interaction With T-SNAREs 1A
XKR6	XK Related 6
ZMAT1	Zinc Finger Matrin-Type 1
ZRANB3	Zinc Finger RANBP2-Type Containing 3

Chapter 1

Introduction

1.1 Theoretical population genetics

Population geneticists are concerned with how genetic variation within and among populations can be explained by evolutionary factors such as mutation, natural selection, recombination and demography. Using mathematical tools makes the construction of theoretical models possible trying to describe the evolution of genetic patterns under the influence of different components. Although those models rely on simplified representations of the real-world situation in the sense that they are idealised enough to be mathematically tractable, they help us to better understand the rules of inheritance and thus how the genetic composition of a population has evolved. Such a model might even help us to make future predictions about the occurrence of specific alleles or combinations of alleles. These approaches might also be useful in medical research areas, for example, by studying the evolution of drug resistance or developing treatments with regard to the prevention, diagnosis, and treatment of diseases (e.g. Wilson et al., 2016; Polimanti et al., 2014; Carlsten et al., 2014).

The beginnings* of theoretical population genetics started to develop in the late 1920s with the research of Haldane (1927), Fisher (1930), and Wright (1931). Up to then, there had been a big discrepancy between supporters of Mendel's studies of heredity (1865) and supporters of Darwin's theory of evolution, which was first proposed by Darwin and Wallace (1858) stating that beneficial traits which improve an individual's ability to survive and reproduce will become frequent in a population with time. The three pioneers of theoretical population genetics merged the ideas of Darwin's theory and the ideas of Mendel's genetics by reinforcing the consequences of natural selection acting on a population simultaneously fulfilling the

*See also: (Boero, 2015; Okasha, 2016; Charlesworth and Charlesworth, 2017)

Mendelian rules of inheritance with mathematical models. Their work provides a decisive contribution to our understanding of the evolutionary process. It was the start of exploring the consequences of various evolutionary hypotheses by using explicit mathematical arguments. Whereas Fisher and Haldane thought that natural selection was by far the most important factor, Wright was convinced that random factors also played an important role in altering the genetic composition in a population. He proposed the concept of genetic drift which is the random change in allele frequencies in a population.
During the following two decades, in the late 1940s and early 1950s the research on evolution was further extended in several directions. Several attempts to explain mechanisms in evolution were introduced and put into a theoretical framework. Gradually, the idea solidified that allele frequencies in a population may change due to four fundamental forces of evolution: the two previously mentioned forces

- natural selection*
- genetic drift

and in addition to these two

- gene flow, which is the change in allele frequencies due to immigration or migration in populations
- mutation pressure, which is the change of allele frequencies solely due to the same mutations occurring over and over again.

With the introduction of technical tools to sequence DNA in the 1960s, it was then possible to test these theoretical models on real experimental data. However, at that time, data sets were relatively small and hence analyses were limited. Kimura made in (1968) an important discovery: by comparing the average number of nucleotide substitutions from data on amino acid substitutions in hemoglobins and a few other proteins in several mammalian species, he found that the number of mutant substitutions was in disagreement with Haldane's theory of natural selection (1957). The number he found was too large. Building on this discovery, Kimura proposed the *neutral theory* (reviewed in (Kimura, 1983)), which states that most mutations have no or negligible fitness advantage or disadvantage, and consequently most mutations are neutral. Therefore, in Kimura's view randomness takes the leading role in the process of evolution.
To prove his statement, Kimura used a diffusion equation approach to compute the probability and time until mutant alleles become fixated.

*Note, that the general term 'natural selection' refers to different modes of selective pressure. Mostly, these modes are known as: 'Positive selection', where a beneficial allele is selected for in a population, 'negative selection', where deleterious alleles are selected against and thus nature acts to remove them from a population, or 'balancing selection', where the existence of multiple alleles gives a fitness advantage and thus they are maintained in a population. Also note that, since in this thesis we focus on 'positive selection', we will not explain the latter two modes in detail.

The neutral theory was a further pioneering development in population genetics. It laid the foundations for the establishment of statistical methods to test for neutrality. The basic idea is that the neutral theory can be seen as a null hypothesis, and deviations from it may be caused by various kinds of evolutionary forces.
However, determining the evolutionary force and the role of natural selection shaping the observed genomic patterns is to date a difficult task. Most models are established in a setting of idealised assumptions. The two most commonly applied models of a population are the *Wright-Fisher model* (Fisher, 1930; Wright, 1931) and the *Moran model* (Moran, 1958). Whilst, the Wright-Fisher model represents a case of idealised non-overlapping generations, the Moran model represents an idealised case of overlapping generations (see also Box 1.1). In the context of these two models, Kingman introduced a theoretical model to describe the genealogy of populations (1982a; 1982b). In a retrospective view, alleles of a gene of individuals in a population can be traced back to a single ancestral copy in what is then called the most recent common ancestor. Kingman showed that the merging of alleles into a common ancestor can be described by a random process, and he called this process the *coalescent*. Instead of describing how a population will evolve in the future with given parameters, coalescent theory looks backward in time by reconstructing the evolutionary history of a present-day sample. These days, coalescent theory has become of central importance in population genetics. We will look more closely at this in chapter 2.1. A huge advantage of coalescent models is that they enable the efficient simulation of data which can be observed under several evolutionary scenarios. They are mostly easier to implement than diffusion approaches and more time-efficient. As we will see, simulations play a significant role in a population geneticist's daily life. Theoretical genealogies of samples can be generated under various assumptions and scenarios, these simulated samples can then be compared with observed data to test neutral hypotheses or estimating population parameters.

Nowadays, with the theoretical knowledge and background established, several simulation programs exist and are still being developed. Also, the technical improvements in sequencing methods contribute enormously to the continual progress. High-throughput sequencing technology allows the sequencing of entire genomes at low cost in a very short period of time. The availability of a large amount of data sources gives the opportunity to apply theoretical models to experimental data, and also to test the power and reliability of these models.

Box 1.1: Wright-Fisher and Moran Model, a brief overview

Wright-Fisher

- Forward in time
- Population size is constant.
- Random mating (panmictic).
- Discrete and non-overlapping generations.
- Generation $t+1$ is obtained by each offspring individual picking one ancestor at random in the parental generation t. (Hence, all individuals in a population die each generation and are replaced by offspring.)

Moran

- Forward in time
- Population size is constant.
- At discrete time intervals, two individuals are chosen randomly: one to die and one to reproduce. The two individuals can be the same.
- Generations are allowed to overlap.

1.2 Aim and overview of the thesis

One of the main concerns in population genetics is to detect genomic patterns left by the action of natural selection. Several test statistics have been developed in the past. However, many tests suffer from high false positives, mainly due to the confounding impacts of demographic events like population bottleneck events, since they can leave a similar pattern behind as those caused by natural selection.
Some recently introduced test statistics exploit the fact that sweeps produce highly unbalanced coalescent tree topologies. Tree topology based test statistics have the advantage that they are free from the confounding effects caused by varying population sizes (Hudson, 1990; Li, 2011). Therefore, building on a test statistic called T_3 (Li and Wiehe, 2013) which detects bias in the balance of binary genealogical trees, we derived a new test statistic based on a log likelihood approach and we called it the LR_{T_3}-test. Since in general the tree topology is not known, we developed an estimation method using SNP data. We showed, that the estimated tree topology agrees quite well with the true topology. Furthermore, we applied the new test statistic to experimental data. For this end, we screened all 26 populations from the human 1,000 genomes project phase 3 data (Auton et al., 2015) with the LR_{T_3}-test. Results of this screen will be presented.
Moreover, we introduced a measure of topological linkage disequilibrium (tLD)

which is based on clustering individuals with respect to their position in the genealogy rather than clustering alleles and haplotypes (Wirtz, Rauscher, and Wiehe, 2018). Also here, we will demonstrate its practical application.

The thesis is organised as follows:

Chapter 2 gives an overview of the basic concepts of coalescent theory and its classical properties. Furthermore, classical test statistics for detecting traces left by natural selection and their underlying ideas will be presented.

Chapter 3 starts with the concept of the test statistic T_3 (Li and Wiehe, 2013). Further on, we show that the gene tree topology can be well approximated using *single nucleotide polymorphism* (SNP) data. We present a suitable clustering method and show its reliability. Building on the test statistic T_3, we establish the LR_{T_3}-test, based on a log likelihood approach. We will show that the power to detect candidate regions for selective sweeps can be improved by far in that way.

In **Chapter 4**, we apply the LR_{T_3}-test to all 26 populations of the phase 3 release of the human 1,000 genomes project. We found new potential candidate regions which might have undergone selective sweeps, and also many of previously known candidates were confirmed. We present our top candidate genes and discuss their potential beneficial trait they may bring along for their carriers.

Chapter 5 introduces the concept of the *topological linkage disequilibrium* (Wirtz, Rauscher, and Wiehe, 2018). We start with a short introduction recapitulating the concepts of classical *linkage disequilibrium*. Advantages of the *topological linkage disequilibrium* compared to the classical *linkage disequilibrium* are pointed out. We conclude with practical applications.

Finally, in **Chapter 6** we present an overview of the results and conclusions of the thesis. Suggestions of possible future research questions will be given.

Chapter 2

Inferring population history

At one time or another surely the thought of getting to know one's ancestors has crossed most people's minds to discover his or her origins. Besides, questions like how closely humans are related to apes or other animals occasionally decorate the headlines of diverse articles.
Exploring the evolutionary relationship, for instance among various species or between individuals of population samples, has always been of keen interest in human history. A basic approach to this concern is the graphical representation of evolutionary history in form of a 'tree'.
In theoretical population genetics, the introduction of the coalescent theory marked a milestone. It provides mathematical tools to study the evolutionary history of a population and enables the establishment of several test statistics for natural selection. In this chapter, we will start with a brief overview of the basic concepts of coalescent theory and mention some classical properties*.

2.1 Coalescent theory

The first who came up with the idea of describing the common ancestry of two alleles mathematically by looking backwards in time was the French Mathematician Gustave Malécot in the 1940s, see e.g. (Epperson, 1999). He asked, given a Wright-Fisher population (see Box 1.1), how far, on average, do you have to go back in time to find a common ancestor for two randomly chosen alleles?
Looking backward in time, the process in which the lines of descent of two alleles merge at a common ancestor is known as the *coalescent*. Being independently developed by several population geneticists (Ewens, 1972; Tajima, 1983; Hudson, 1983),

*In this chapter, throughout all sections, information content is mainly obtained from the textbooks (Hartl and Clark, 2007; Wakeley, J., 2009; Nielsen and Slatkin, 2013)

the first to record the theory behind a coalescent process as a mathematical model was Kingman (1982a) and he called it the *n-coalescent*. The idea is as follows:
Assuming a population of size $2N$, the probability that two randomly chosen alleles share the same parental allele in the previous generation is $1/2N$, and the probability that they do not share the same parental allele in the previous generation is $(1 - 1/2N)$. In the latter case, we can continue by asking what the probability is that these two alleles share the same grand-parental allele: It is $1/2N$ that they do share, $(1 - 1/2N)$ that they do not share. We can proceed like this and arrive at the probability that two alleles do not coalesce in generation $(t - 1)$, but do coalesce in the t-th generation

$$P(\text{2 alleles coalesce at time } t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}.$$

Now, let us consider a sample of n alleles in which all lineages coalesce independently and only one coalescent event can occur each generation. In any generation, the probability of a pair of alleles coalescing is $1/2N$ and there are $n(n-1)/2$ such pairs. Hence, the probability of coalescent times can be approximated by the exponential distribution (for sufficiently large N)

$$\begin{aligned} P(\text{2 out of } n \text{ alleles coalesce at time } t) &= \left(1 - \frac{n(n-1)}{4N}\right)^{t-1} \frac{n(n-1)}{4N} \\ &\approx \frac{n(n-1)}{4N} e^{\frac{-n(n-1)t}{4N}} \end{aligned} \tag{2.1}$$

with average waiting time T_n for a coalescent event:

$$E[T_n] = \frac{4N}{n(n-1)}.$$

Eventually, all lineages will merge into one node, which is called the *most recent common ancestor* (MRCA). The expected time to the MRCA is equal to the sum of the expected waiting time $E[T_i]$:

$$E[T_{\text{MRCA}}] = \sum_{i=2}^{n} E[T_i] = \sum_{i=2}^{n} \frac{4N}{i(i-1)} = 4N\left(1 - \frac{1}{n}\right).$$

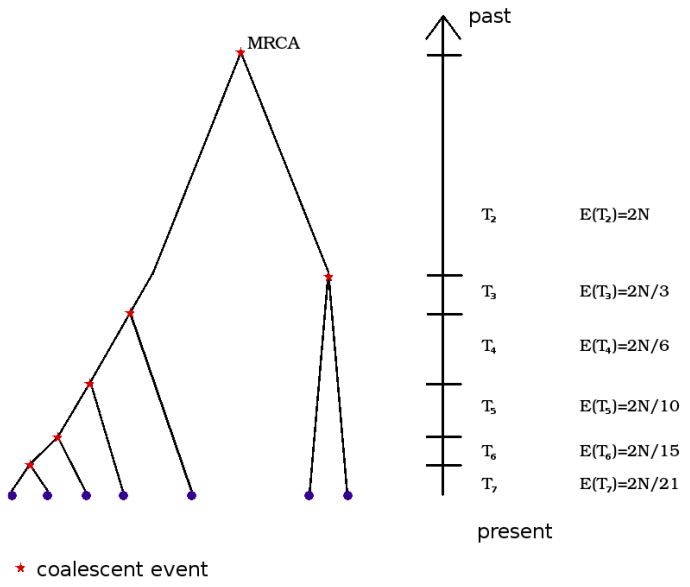


FIGURE 2.1: One possible coalescent tree of a sample of size seven. The lineages are represented by the leaves of the tree. The times between coalescent events are exponentially distributed and are denoted by T_i. On the right side, the respective expected waiting time is given.

The expected complete branch length of the tree $E(T)$ can be computed by summing up the branch lengths $E(T_i)$ for the entire tree:

$$E(T) = E\left(\sum_{i=2}^{n} iT_i\right) = \sum_{i=2}^{n} iE(T_i) = \sum_{i=2}^{n} i\frac{4N}{i(i-1)} = 4N\sum_{i=1}^{n-1}\frac{1}{i}.$$

Note that the coalescent time is increasing as one goes back further in time and the last coalescent time from two alleles to the MRCA is the longest. If n is large, almost half the time is required for the last coalescent event (Felsenstein, 2004).

In this thesis, we consider only binary trees. However, it is worth mentioning that while Kingman's coalescent only produces binary trees, many studies exist dealing with multiple merger coalescent events, e.g. the Λ-coalescent (Pitman, 1999), which allows a coalescent event involving more than two lineages, or the more generalized Ξ-coalescent, which in addition allows simultaneous multiple coalescent events of multiple lineages per generation (Schweinsberg, 2000; Moehle and Sagitov, 2001).

2.1.1 Adding mutation

We now turn to adding mutations to the coalescent model. The *infinite-sites model* is assumed, where each mutation can occur at an infinite number of sites and every

new mutation occurs at a novel site. Mutations are rare events occurring with rate μ during time t per individual. Hence, the number of mutations which occur over coalescent tree branches of a given length is Poisson distributed,

$$P(k \text{ mutations in } t \text{ generations}) = \frac{e^{-t\mu}(t\mu)^k}{k!},$$

and the expected number of mutations is $t\mu$.

Adding mutations to the coalescent tree also means graphically: Mutations affecting only one chromosome can only have occurred on an external branch, mutations affecting many chromosomes have occurred earlier in time, see FIGURE 2.2.

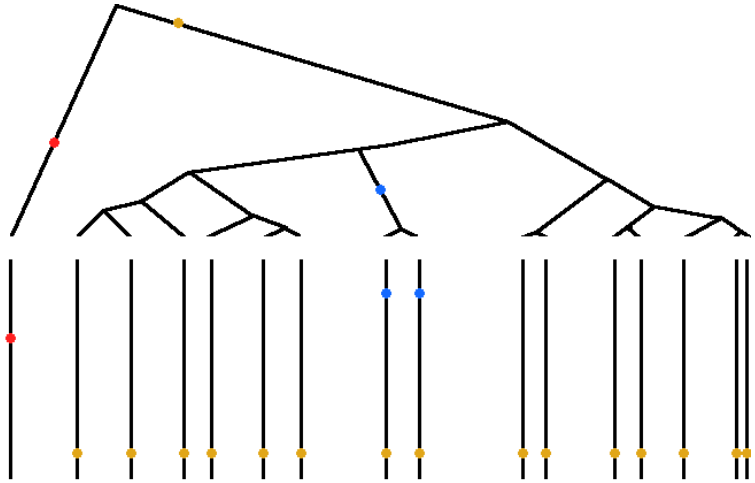


FIGURE 2.2: Coalescent tree for a sample of size $n = 16$, mutations are represented as dots, the respective DNA sequences are drawn below as vertical lines. Colours of mutations indicate the different number of chromosomes which are affected by that mutation: The red one affects only 1 chromosome (= singleton), the blue one two chromosomes (= doubleton), orange affects 15 chromosomes.

One can also compute the expected number of segregating sites $E(S)$. It is

$$E(S) = \mu E(T) = \mu 4N \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}, \tag{2.2}$$

where μ is the per site mutation rate and $\theta := 4N\mu$. θ is also called the population scaled mutation rate.

By rearranging the above equation, it holds that

$$\theta = \frac{E(S)}{\sum_{i=1}^{n-1} \frac{1}{i}}.$$

Actually, Watterson (1975) was the first to derive the expected number of segregating

sites. Nowadays, it is common to use that as means for the estimation of θ. It is also known as 'Watterson's Estimator':

$$\hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}. \tag{2.3}$$

Note that another popular estimator for the population mutation rate is $\hat{\theta}_T$ (or also θ_π), called after Tajima, who first described it (1989a):
The number of nucleotide site differences between a pair of sequences is simply the number of counts of nucleotide positions at which pairwise sequences differ, divided by all possible pairwise comparisons that can be made:

$$\pi = \frac{2}{n(n-1)} \sum_{i<j} d_{ij},$$

where d_{ij} is the number of differences between the ith and jth sequence.
Since the number of nucleotide site differences between a pair of sequences is the same as the number of segregating sites in a sample of size two, from 2.2 we know that an average pair of sequences differs at θ sites. Averaging over all the pairs in a sample doesn't change this, so it follows that

$$E(\pi) = 4N\mu = \theta. \tag{2.4}$$

From this result, one can deduce 'Tajima's Estimator':

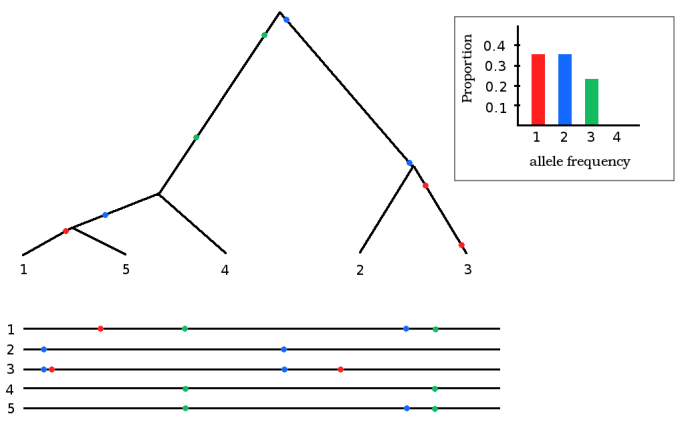
$$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}. \tag{2.5}$$

(The ˆ indicates that these formulas are intended to estimate the parameter θ.)

2.1.2 Site frequency spectrum

Further on, to obtain information about the frequency spectrum of mutations, consider the *site frequency spectrum* (SFS): The (unfolded) SFS is the distribution of the proportion of segregating sites where the derived allele (the mutant) is at the absolute frequency i. For a sample of size n, the SFS can be represented as a vector $f = (f_1, f_2, ..., f_{n-1})$, where f_i denotes the proportion of the derived allele in frequency i. For example, f_1 is the proportion of mutations affecting only one chromosome, also called *singletons*, f_2 is the proportion of mutations affecting two chromosomes, also called *doubletons*, and so forth.

Box 2.1.2: Example: SFS for the DNA sequence



Example of a coalescent for a sample of size 5. The five black horizontally drawn lines on the bottom of the picture represent DNA sequences. The dots indicate mutations, the different colouring represents the (absolute) frequency of the mutation in the sample. There are 8 segregating sites, 3 out of these are singletons, 3 are doubletons and 2 are tripletons. In the upper right corner (the picture in the framed box), the SFS for this DNA data example is given. Note: Tree genealogy can influence the frequency of segregating sites in the sense that the observed patterns are a result of the given genealogical tree.

The expected SFS can be calculated by means of the coalescent and is given by

$$E[f_i] = \frac{\theta/i}{\theta \sum_{k=1}^{n-1} \frac{1}{k}} = \frac{1/i}{\sum_{k=1}^{n-1} \frac{1}{k}}, \quad i = 1, 2, ..., n-1. \tag{2.6}$$

In some cases it is unknown which allele is the derived one and which is the ancestral one. Then one can consider the *folded SFS* which is the distribution of the frequencies or counts of minor alleles in a sample. Obviously, here $i = 1, ..., \lfloor n/2 \rfloor$, and

$$E[f_i] = \frac{\left(\frac{1}{i} + \frac{1}{n-i}\right)}{\sum_{k=1}^{n-1} \frac{1}{k}}, \quad i = 1, 2, ..., \lfloor n/2 \rfloor. \tag{2.7}$$

2.1.3 Adding recombination

In its simplest form, coalescent theory assumes no recombination. Recombination is a process during meiosis by which two DNA sequences exchange genetic material when crossing over occurs. Adding recombination into the coalescent framework is not straight-forward. FIGURE 2.3 illustrates the difficulty.

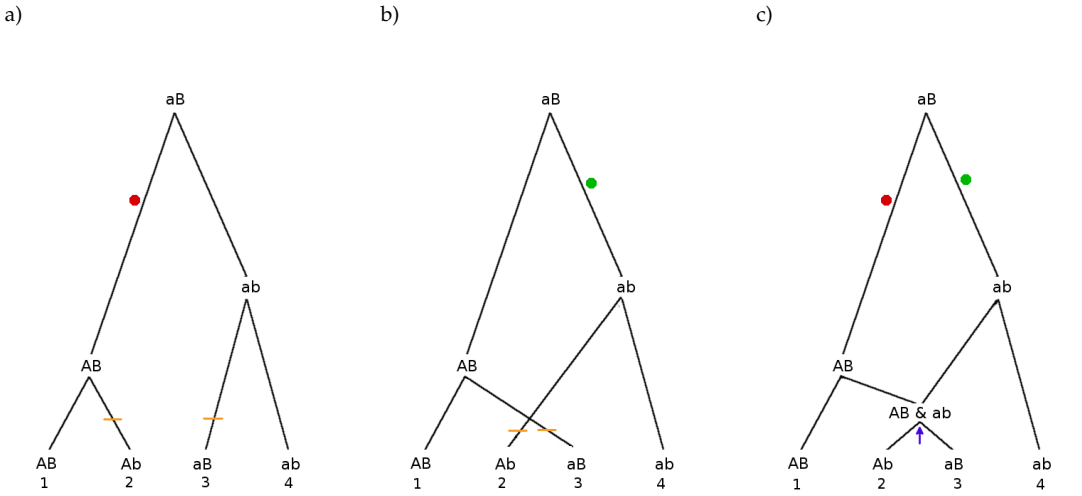


FIGURE 2.3: Picture modified from (Hartl and Clark, 2007, Chapter 3.7, figure 3.17). Shown here is an example of coalescence and recombination in a sample of size $n = 4$. The A/a represents the allele on one site, the B/b allele on the second site. Plot a) shows the coalescent tree with respect to the A and a pair of alleles. The red circle indicates the mutation from a to A. The horizontal lines indicate that one AB-bearing chromosome recombines with an ab-bearing chromosome. Here, suppose the leaves are labelled 1 to 4 from left to right, 1 and 2 are joined together and 3 and 4. Plot b) shows the coalescent tree with respect to the B and b pair of alleles. The green circle indicates the mutation from B to b, and again, the horizontal lines indicate that one AB-bearing chromosome recombines with an ab-bearing chromosome. Here, 1 and 3 are joined together, and 2 and 4. Hence, both trees in a) and b) represent the ancestry of the A, a and B, b pairs of alleles, respectively. But the order of the tree is different. Plot c) A possibility to deal with recombination events: The arrow in the coalescent tree in plot c) points at the coalescence where the recombination took place and the recombinant chromosomes create their own parental node.

Nowadays, recombination and coalescent process is usually studied in the framework of the ancestral recombination graph (ARG), which was introduced by Griffiths and Marjoram (1996). In the ARG, each nucleotide position along the chromosome is associated with a coalescent tree. Due to recombination events, tree topology at different sequence positions may change. Within a chromosome segment with no recombination events, all positions have the same tree topology, the so-called 'marginal tree'. Therefore, by dividing chromosomes into fragments with ideally no recombination events, coalescent trees can be associated to each of a fragment. Recombination is embedded by a random 'prune and re-graft event': A branch of a marginal tree is randomly chosen, pruned and subsequently re-grafted somewhere else above the pruning point or even onto the ancestral lineage of the root. In the latter case, this would lead to a change of root, hence a change of the MRCA.
The ARG can be well approximated by a so-called 'Sequential Markov Coalescent' (McVean and Cardin, 2005; Eriksson, Mahjani, and Mehlig, 2009). The basic idea here is that the ARG is approximated by a process which iteratively determines the genealogy along a chromosome, the local tree at a site depends only on the tree at the previous site.

2.2 A side note on evolutionary trees

In evolutionary biology, the graphically representation of relationships among individuals in the form as a tree has a long history.
So far, we have focused on the coalescent approach. Coalescence theory concentrates on reconstructing possible gene histories to explore what causes might have led to the observation of the underlying genealogy tree. Whilst here the focus lies on the intra-species history, the field of *phylogeny* is interested in inter-species history.
It was the famous zoologist E. Haeckel who coined the word 'phylogeny' in the 1860's, which can be read e.g. in (Dayrat, 2003). A phylogenetic tree represents the evolutionary history of a species observed through time. They are also known as *species trees*. The aim is to reconstruct the 'true' species tree. To build the tree, various data types can be used, however, nowadays it is most common to built phylogenetic trees from molecular data, like DNA or protein data. In molecular phylogenetic analysis, the sequence of a common gene or protein are used to infer the evolutionary relationship of species. The most common methods for estimating the trees are distance-based methods (like UPGMA or neighbour joining algorithms), maximum parsimony methods (i.e. 'choosing' the tree that requires the least amount of mutations to be constructed), and Bayesian methods based on likelihood functions (Yang and Rannala, 2012).
There has been a long-standing debate about which phylogenetic method performs best and how reliable each one is, strongly depending on the type of data used, though. Phylogeneticists are concerned with questions like which the true tree is, if a true tree exists at all.

A species tree might be different from the gene tree. One reason for this phenomenon is called *incomplete lineage sorting*:
If the divergence time was short and the ancestral population sizes were large, it can happen that by the time of the divergence event, not all lineages in a sample from each population have found their MRCA yet. In such a case, one or more lineages from one species will share the MRCA with lineages from the other species (see also FIGURE 2.4).
Other reasons causing the discord between species tree and gene tree can be e.g. horizontal gene transfer (Davidson et al., 2015), gene duplication and loss or hybridization (Szollosi et al., 2015).

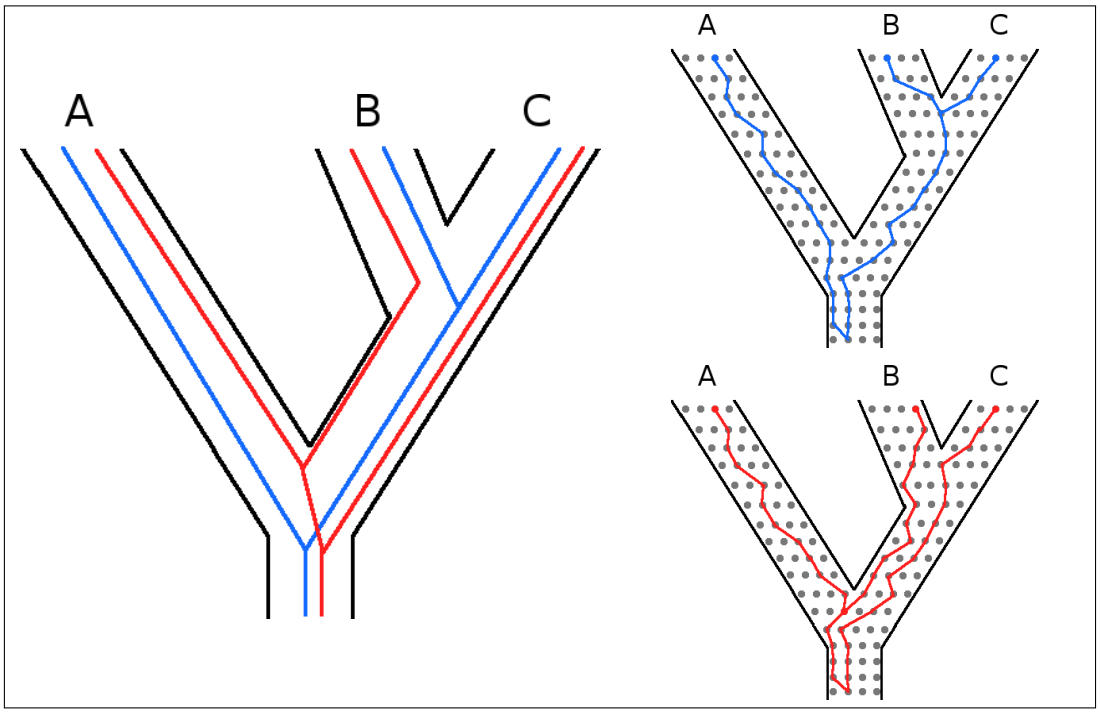


FIGURE 2.4: The gene tree in blue matches the species tree in black, while the gene tree in red does not. Reason for the mismatch might be incomplete lineage sorting: By tracing back three sampled lineages from species A, B and C backward in time, alleles from A and B might succeed (right side: tree on top) or might not succeed (right side: tree on bottom) to coalesce in the common ancestor.

2.3 Tests on neutrality

With the advent of new and rapid sequence technologies, a huge amount of DNA data is now available. It is mostly stored in so-called 'gene data banks' and are publicly available; genomic patterns can be actually analysed and extensively studied. These patterns might have been shaped by factors such as demography, natural selection or genetic drift.

However, distinguishing between those can be difficult, for instance, demographic events like population bottlenecks can leave a similar genomic pattern behind as those left by the action of natural selection. The construction of a robust test statistic aiming in identifying the correct underlying dynamic behind, received a high degree of attention for researchers in the past decades.

In this section, we will present the characteristic genomic signatures of positive selection and classical approaches to detect them.

2.3.1 Genomic footprints of positive selection

In a fundamental work, Maynard Smith and Haigh (1974) introduced the following model: When a beneficial mutation arises on a chromosome and subsequently gets fixed in the population, not only the frequency of the advantageous mutation will

increase but so will selectively neutral mutations which are linked to the selected site. This effect occurs due to physical linkage between alleles at different loci, a term called *linkage disequilibrium (LD)* which was first used by Lewontin and Kojima (1960) (more on *LD* in chapter 5).
While the advantageous mutation and the linked neutral variant are swept to high frequency, other neutral variants are swept out of the population, a phenomenon called 'selective sweep' (illustrated in FIGURE 2.5) which results in strongly reduced levels of polymorphism around the selected site.

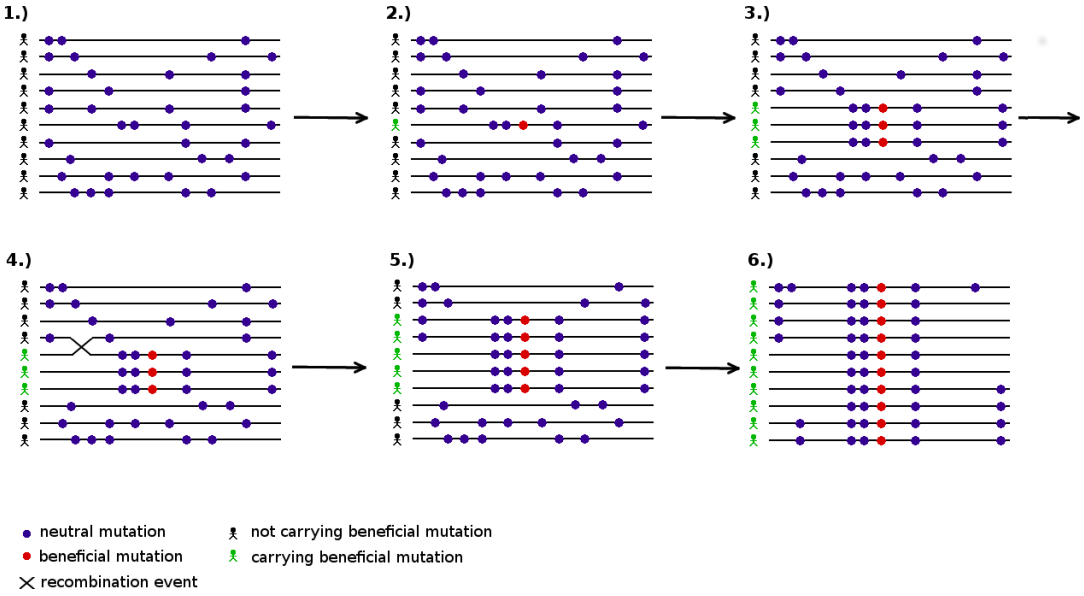


FIGURE 2.5: Consider a sample of size $n = 10$. 1.) Each of the 10 DNA sequence is represented by a horizontal line. Each blue dot represents a neutral mutation, which can be present in more than one sequence. 2.) An advantageous mutation occurs, indicated by a red dot. 3.) The beneficial mutation increases in frequency in the population, and hereby also the frequency of neutral mutations located close to the selected site increase due to their association with the beneficial allele. 4.) A recombination event creates a new combination associated with the selected site. 5.)-6.) The selected site and linked neutral variants increase in frequency and finally are fixed in the population.

Maynard Smith and Haigh called this process 'genetic hitch-hiking'. The work of Maynard Smith and Haigh marked a milestone for population geneticists. Building on this model, a variety of strategies to detect positive selection have been developed. They mostly rely on the idea of detecting specific shifts of the SFS, searching for reduced genomic variation in the genome, or finding specific *LD* patterns. More recently, machine learning approaches gain growing attention, e.g. (Schrider and Kern, 2018).

In the following, we give a short overview of rather 'classical' tests.

2.3.2 Classical neutrality tests

In general, methods detecting selective sweeps can be divided into groups based on their underlying idea. One big group is formed by those based on shifts in the site frequency spectrum (SFS). Selective sweeps affect the SFS in the sense that the SFS creates a shift towards an excess of low- and high-frequency derived alleles (Braverman et al., 1995). In the previous section we have seen that a consequence of a selective sweep is the reduction of genetic diversity around the selected area. Some time after the sweep has been completed, the region will recover from the sweep again, new mutations will occur, however, they can not rise to high frequency due to the short time, creating an excess of rare alleles around the swept region. SFS based neutrality tests exploit this fact. By means of θ-estimators such a shift can be measured. In the section before, we have already seen two estimators for the population scaled mutation rate: $\hat{\theta}_W$ and $\hat{\theta}_\pi$ (equation (2.3) and (2.5) respectively). Under neutrality both estimators are expected to be equal. After a selective sweep, $\hat{\theta}_\pi$ will be smaller than $\hat{\theta}_W$, because mean pairwise differences are less to what is expected from the number of segregating sites. The classical *Tajima's D* test is the comparison between these two quantities (Tajima, 1989a):

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{\mathrm{Var}(\hat{\theta}_\pi - \hat{\theta}_W)}},$$

where $\hat{\theta}_W$ and $\hat{\theta}_\pi$ are given in (2.3) and (2.5), respectively.

There are other estimators for θ than we have seen thus far. E.g. define ξ_1 as the absolute number of singletons, then according to equation (2.6) the $E[\xi_1] = \theta$ and thus

$$\hat{\theta}_e = \xi_1. \tag{2.8}$$

Fu and Li (1993) derived the test statistics *Fu and Li's D* and *Fu and Li's F*, comparing the number of derived singleton mutations and the total number of derived variants:

$$D = \frac{\hat{\theta}_W - \hat{\theta}_e}{\sqrt{\mathrm{Var}(\hat{\theta}_W - \hat{\theta}_e)}}$$

and

$$F = \frac{\hat{\theta}_\pi - \hat{\theta}_e}{\sqrt{\mathrm{Var}(\hat{\theta}_\pi - \hat{\theta}_e)}}.$$

Another noteworthy test from this group is *Fay and Wu'S H* (2000). Their θ-estimator gives in addition much weights to high frequency variants relative to the intermediate-frequency ones. It is defined as

$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i,$$

where ξ_i are the counts of derived allele with absolute frequency i, and hence

$$H = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sqrt{\mathrm{Var}(\hat{\theta}_\pi - \hat{\theta}_H)}}.$$

We now turn to a further big group of neutrality tests: haplotype-based tests. A haplotype is the configuration of segregating sites lying on the same chromosome (see also FIGURE 3.3). In contrast to SFS based tests, these tests also include linkage. In a seminal paper, Sabeti et al. (2002) developed an extended haplotype homozygosity (EHH) which detects long haplotypes at unusually high frequencies in candidate regions. It measures the decay of haplotypes carrying a specified 'core' allele at one end as a function of distance. Building on this, the integrated haplotype score (iHS) was developed by Voight et al. (2006). It measures the amount of EHH at a given site along the ancestral allele relative to the derived allele.

Also a notable consequence of the hitch-hiking effect is that the LD levels are expected to remain high in comparison on each side of the advantageous mutation, and drop drastically for loci across the beneficial mutation, motivating to develop LD-based methods to detect positive selection (Kim and Nielsen, 2004; Wang et al., 2006).

A disadvantage of most statistical tests is that they are affected by the confounding effects of demographic factors (Ramirez-Soriano et al., 2008). Events like population expansions, recoveries from a recent population bottleneck or gene flow lead to shifts in the SFS. For instance, both population expansion or recovery from a recent population bottleneck lead to an excess of low-frequency variants (Fu and Li, 1993; Tajima, 1989a; Tajima, 1989b). Gene flow can result in increasing high-frequency derived variants (De and Durrett, 2007). Also haplotype-based tests suffer from these effects, since they are functions of the recombination rate, the mutation rate and population size (Pritchard and Przeworski, 2001). For instance, LD can be increased by temporary reductions in population size and declines more slowly after the occurrence of such a bottleneck event (Reich et al., 2001).

2.3.3 Tests using coalescent tree topology

We now want to focus on how the tree topology can be used to establish neutrality tests. Suppose an excess of singletons or an excess of rare derived alleles is observed (remember: singletons can only lie on external branches). In terms of tree topology this means that the external branches are likely to be relatively long compared to the short internal branches. Furthermore, after the fixation of a positively selected allele in a population, the tree height is drastically reduced due to its short fixation time at the selected locus. All genealogical branches coalesce at a recent time at the selected site. Not only branch length or tree height in general is affected by a selective sweep, but also the shape. 'Due to the effect of hitch-hiking, one lineage of a neutral locus partially linked to a selected locus may escape from the selective sweep through recombination' (Li, 2011). This lineage will not coalesce with any other lineages before the most recent common ancestor (Kaplan, Hudson, and Langley, 1989; Fay and Wu, 2000) and that leads to a long branch which is linked to the root of the tree. The tree topology is highly asymmetric; the tree is also said to be highly *unbalanced*. Taking the underlying tree topology additionally into account in establishing neutrality tests can provide a more reliable conclusion about what role positive selection might have actually played.

Recently, several test statistics based on coalescent tree topology were established. Li (2011) used the maximum frequency of derived mutations to examine the unbalancedness of the tree of a locus. Furthermore Li showed, that topology-based tests are robust with respect to demographic changes such as bottleneck events. Ferretti et al. (2017) analysed the impact of the structure of genealogical trees upon the SFS by decomposing the SFS in terms of waiting times and tree shape. Yang et al. (2018) took into account the ratio between the lengths of two subtrees in addition to the information of the unbalancedness of the tree.

Li and Wiehe (2013) introduced a simple test for selective sweeps based on microsatellite variation. They called the test statistic T_3 and it only uses tree topology in the sense of tree shape. Basically, the T_3-test is a measurement for the unbalancedness of tree topology. Based on the same model as in (Li and Wiehe, 2013), in the next chapter, we will introduce the T_3 test statistic using SNP data. Furthermore, we will embed the test statistic T_3 in a log likelihood ratio test, and we call it the LR_{T_3}-test. We will show that the power to detect candidate regions for selective sweeps can thus be improved.

Chapter 3

Using genealogical tree topology to detect positive selection

Hudson (1990) proved that in a Wright–Fisher population varying population size does not affect tree topology. Moreover, Li (2011) showed that tree topology is not affected by demographic events like population bottleneck events or size expansion. It therefore stands to reason that tree topology-based statistics are to be considered to search for traces of selective sweeps. As we have already seen in a previous chapter, a selective sweep also leaves visible traces on tree topology: After the fixation of a positively selected allele in a population, the tree height is drastically reduced due to its short fixation time at the selected locus. Genealogical branches will all coalesce in a recent time at the selected site, leading to a tree of low height. Genetic diversity is strongly reduced around that site. But when one moves away, recombination breaks this link, one or a few lineages might escape the selective sweep leading to an unbalance in tree topology. (Kaplan, Hudson, and Langley, 1989; Fay and Wu, 2000)

Most existing coalescent tree topology based tests require more information than just tree topology* (e.g. Li, 2011; Yang et al., 2018). We aim to derive a robust test statistic solely relying on tree topology. Therefore, we will build upon the already known T_3-test (Li and Wiehe, 2013), which is based on the latter idea.

*When we talk about 'tree topology', we mean solely the branching pattern. This means other information like tree height, branch length etc. are of no significance.

3.1 The test statistic T_3

The test statistic T_3 was introduced by Li and Wiehe (2013). A detailed review of the derivation of the T_3-test is provided in the APPENDIX A.1. In the following, only results which will be needed in further sections will be pointed out.

First, we will introduce some terminology:
Consider a binary tree with a fixed number n of leaves. This number is also defined as the size of the tree and represents a sample of size n. The tree has $n-1$ internal nodes, denoted by ν_i, $i = 1, ..., n-1$. The labelling starts at the root of the tree, which also refers to the *most common recent ancestor* (MRCA). As can be seen in FIGURE 3.1, the n leaves of the tree can be divided into two disjoint groups: the left- and right-descendants of root ν_1. The two groups are indicated as L_1 and R_1, respectively.
Further on, let $n = n_1$ and define $\Omega_1 = \min\{|L_1|, |R_1|\}$. Without loss of generality, let $|L_1|$ be smaller than $|R_1|$, thus $\Omega_1 = |L_1|$.
Next, label the root of the subtree consisting of the leaves which belongs to the 'larger' set, in this case the root of subtree with leaf set R_1, with ν_2. This subtree is now of size $n_2 = n_1 - \Omega_1 \geq \frac{n_1}{2}$, since $|R_1| = n_1 - \Omega_1 \geq \frac{n_1}{2}$. Again, divide the n_2 leaves merging at root ν_2 into two disjoint groups: the group containing the right-descendants, $|R_2|$, and the group containing the left-descendants, $|L_2|$. And again, without loss of generality let $|L_2| < |R_2|$, and $\Omega_2 = \min\{|L_2|, |R_2|\} = |L_2|$. In the same manner, we can proceed to determine Ω_3, Ω_4 and so on.

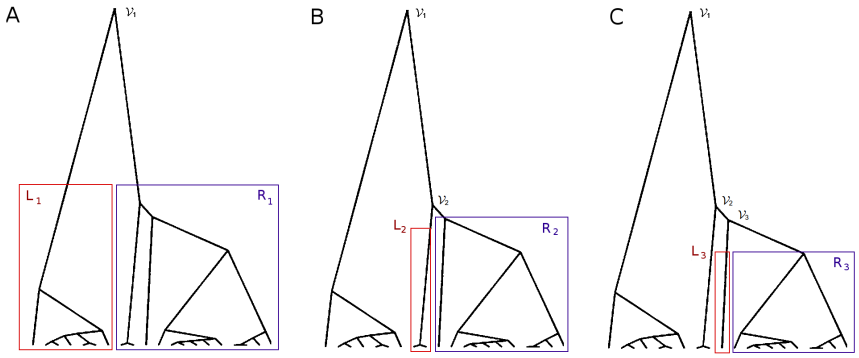


FIGURE 3.1: Example of a binary tree of size $n = 20$. A: Tree with root ν_1, $n = n_1 = 20$, $|L_1| = 7$, $|R_1| = 13$, and thus $\Omega_1 = \min\{|L_1|, |R_1|\} = 7$. B: Label root of set with $\max\{|L_1|, |R_1|\}$ by ν_2, hence $n_2 = 13$, $|L_2| = 2$, $|R_2| = 11$, and $\Omega_2 = 2$. C: Proceed in this way and get $n_3 = 11$, $|L_3| = 1$, $|R_1| = 10$, and thus $\Omega_3 = 1$.

Assuming that trees are generated by the coalescent process, it follows that Ω_1 is a random variable which is 'almost'-uniformly distributed on $\{1, 2, ..., \lfloor n/2 \rfloor\}$ with

$$p(n, \omega_1) := \text{Prob}(\Omega_1 = \omega_1) = \frac{2 - \delta_{\omega_1, n/2}}{n-1},$$

where $\delta_{\cdot,\cdot}$ denotes the Kronecker symbol.
Furthermore, Ω_i given Ω_j, $1 < i < j$, is 'almost'-uniformly distributed on $\{1, 2, ..., \lfloor n_i/2 \rfloor\}$

with

$$p(n_i, \omega_i) := \text{Prob}(\Omega_i = \omega_i),$$

where $n_i = n - \omega_1 - ... - \omega_{i-1}$ and $1 \leq \omega_i \leq \lfloor n_i/2 \rfloor$.
Note that the Ω_i depend on Ω_j, $j = 1, ..., i-1$.

It can be shown that the expectation for Ω_1 is

$$E(\Omega_1) \approx \frac{n}{4}$$

and the variance

$$V(\Omega_1) \approx \frac{n^2}{48}.$$

In general, it holds that

$$E(\Omega_i) \approx \frac{3^{i-1}n}{4^i},$$

and the variance

$$V(\Omega_i) \approx \frac{1}{3}(1 - \frac{3^{i-1}n}{4^i})^2.$$

(see APPENDIX A.1 for more details on calculations.)
By defining the normalised random variables $\Omega_i^* = 2\Omega_i/n_i$, it can be deduced that

$$E(\Omega_1^*) \approx \frac{1}{2} \tag{3.1}$$

and

$$V(\Omega_1^*) \approx \frac{1}{12}.$$

In general, it holds that

$$E(\Omega_i^*) \approx \frac{1}{2}, \tag{3.2}$$

$$V(\Omega_i^*) \approx \frac{1}{12},$$

and hence

$$\sigma(\Omega_i^*) \approx \sqrt{1/12}.$$

We will mostly work with the normalised random variables $\Omega_i^* = 2\Omega_i/n_i$ instead of Ω_i. In this way, they can be well approximated by independent continuous uniforms on the unit interval. With

$$E(\Omega_i^*) \approx 1/2 \qquad \text{and} \qquad \sigma(\Omega_i^*) \approx \sqrt{1/12},$$

it holds that

$$\mathcal{N}(0,1) \sim \sqrt{\frac{1}{k}} \cdot \sum_{i=1}^{k} \frac{(\Omega_i^* - E(\Omega_i^*))}{\sigma(\Omega_i^*)} = \sqrt{\frac{12}{k}} \cdot \sum_{i=1}^{k} \left(\Omega_i^* - \frac{1}{2}\right) =: T_k \tag{3.3}$$

by applying the central limit theorem, which states that the sum of continuous uniforms converges in distribution to a normal random variable.
Already $k = 3$ produces a distribution close enough to a standard normal distribution, as shown in (Li and Wiehe, 2013) and re-checked with simulations (see FIGURE 3.2). Hence, set $k = 3$.
The resulting test statistic T_3 is a measurement for tree balance of binary coalescent trees:

$$T_3 = 2 \cdot \sum_{i=1}^{3} \left(\Omega_i^* - \frac{1}{2}\right) \sim \mathcal{N}(0,1).$$

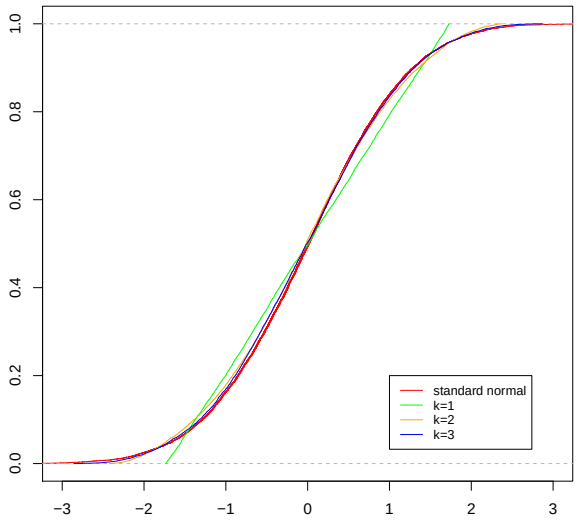


FIGURE 3.2: Agreement of T_k (see equation (3.3)) with the standard normal. As can be seen, already $k = 3$ yields a distribution close to the standard normal distribution.

In the case of neutral evolution T_3 is expected to be standard-normally distributed, i.e. $E(T_3) = 0$, $V(T_3) = 1$. Genealogies after selective sweeps tend to be unbalanced

and produce negative values of T_3 *.

3.2 Estimation of tree topology using SNP data

In practice, tree topology is not known and has to be estimated. Therefore, the reliability of the T_3-test depends on the quality of the reconstruction of the tree topology. Li and Wiehe (2013) showed the application to microsatellite data. They found that the *unweighted pair-group method with arithmetic mean* (UPGMA) yielded a reliable result. The idea was that the microsatellite alleles were grouped into two disjoint sets according to their repeat size and size distance from each other. In the end, the authors could successfully show significance for two microsatellite markers out of the used 16 markers of the *Plasmodium falciparum* surrounding a known drug resistance locus.
In the following, we will demonstrate that the T_3-test can also be well applied to *single nucleotide polymorphism* (SNP) data.

3.2.1 Clustering method

Consider a sample of size $n = n_1$. By using a sliding window approach for a given window length in number of base pairs (bp) and a given step size, we consider the combination of SNPs in each window (see FIGURE 3.3). For clustering the observed haplotypes in two disjoint groups, we apply a 2-means like clustering approach:
We determine the two sequences with maximal Hamming distance. These two most different sequences are now treated as centroids of the two clusters the n_1 sequences have to be grouped into. Next, we assign the remaining $n - 2$ sequences according to their similarity to one of the two 'centroidal' sequences. If the allocation to one of the two groups is not clearly resolvable, for instance when the focal sequence has the same distance to the two 'centroidal' sequences, we randomly assign the alleles to one of the two clusters with equal probability. This gives preference to clusters of balanced size. Once all n_1 sequences have been assigned to one of the two clusters, we are able to determine Ω_1, which is simply the minimum size of the two groups. Now, we can proceed to the next step: Determining Ω_2. For this, we now focus on the remaining $n_2 = n_1 - \Omega_1$ sequences not contributing to Ω_1. The whole clustering procedure is carried out in exactly the same way as before. In this manner, we can estimate Ω_2 and Ω_3.

*An illustration of T_3-profile under different scenarios will be given later, see FIGURE 3.11.

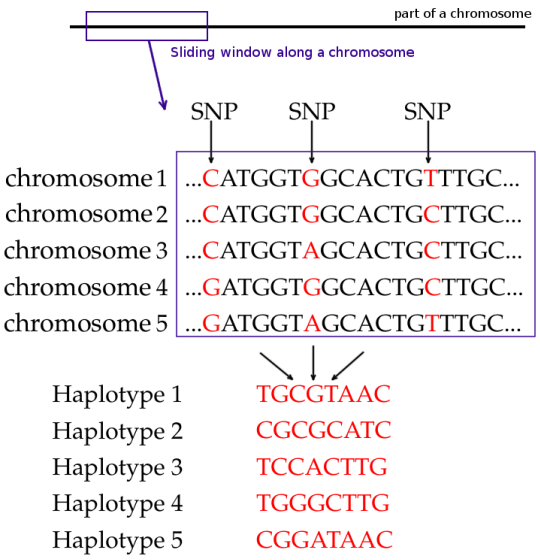


FIGURE 3.3: Part of a chromosome (black line), and a window of a given size, e.g. number of base pairs which slides along this chromosome (blue box). Now let's assume we are analysing this stretch of chromosome for five different sequences in a population: Most of the DNA sequence is identical (black letters), SNPs are indicated in red. A haplotype is made up of a particular combination of alleles nearby SNPs. Here, only SNPs contained in the window are denoted, since this is sufficient to define the haplotypes uniquely.

Number of SNPs and fragment length

Coalescent tree topologies along the chromosome are not independent. Multiple recombination events within a fragment may lead to confounding effects on cluster estimation. This means that fragment length can not be arbitrarily large. But at the same time, it should contain a minimum number of segregating sites to enable a fairly good approximation of the true tree topology.

Minimum number of SNPs

To investigate how many SNPs are at least needed to obtain a good cluster estimation result for the Ω_i's, we generated simulated data for population samples using the simulation program *msms* by Ewing and Hermisson (2010) with varying number of segregating sites. The program *msms* is a coalescent simulation program for genealogies in general structured populations and based on the widely used and well-known simulation program *ms** by Hudson and Kaplan (1988), with the difference of allowing selection at a single locus. Since the output of *msms* provides both SNP data and trees representing the history of the sampled chromosomes in Newick format, in each run we can compare our estimated tree topology from SNP data with the true one (for an example output see FIGURE 3.4).

To choose the appropriate minimum number of segregating sites needed to get a fairly good approximation of the true tree topology, we generated 16 different data sets under neutral assumptions but with various number of segregating sites (ss):

*Note, the difference between *ms* and *msms* is that *msms* contains the option for simulating selection. Both, interface and output format are consistent and therefore, with no selection both can be used equally.

```
1  msms -N 10000 -ms 5 1 -s 20 -T
2  0x48853d412a07f114
3
4  //
5  (((5:0.025,4:0.025):0.249,3:0.274):0.184,(1:0.042,2:0.042):0.416);
6  segsites:  20
7  positions:  0.00805 0.02248 0.03072 0.05581 0.05693 0.09182 0.29899
   0.39859 0.43621 0.48719 0.53773 0.55121 0.61512 0.62242 0.69708
   0.71393 0.91442 0.93375 0.95735 0.96282
8  10001001010100000000
9  10001101010100000000
10 01110010101000111010
11 01000000100011100101
12 01000000100011100100
```

FIGURE 3.4: Example output of a simple *msms* command for the effective population size of N =10000, for 1 sample consisting of 5 sequences, generated assuming that there are 20 segregating sites. The first line of the output is the command line. The second line shows the random number seeds. The history tree in Newick format is represented in line 5, which is triggered by the option *-T* in the command line (see also FIGURE 3.5). Line 6 gives the number of segregating sites in the sample, while in line 7 the positions of the sites are given on a scale of $(0, 1)$. Followed by this line, the haplotypes of each of the 5 sequences are given as a string of '0's, indicating the ancestral allele, and '1's, which stands for the derived allele.

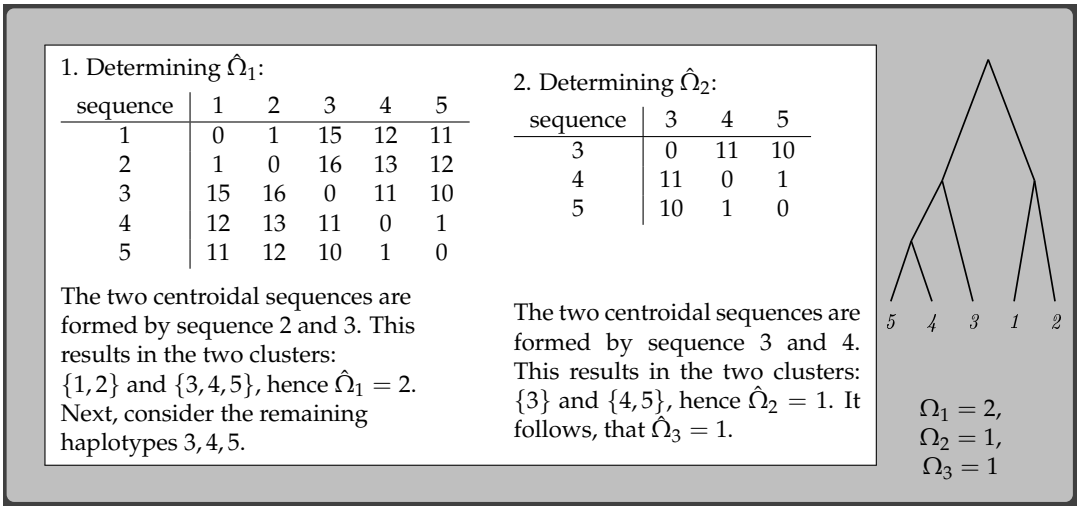


FIGURE 3.5: Example of forming clusters using the SNP data from the output of *msms* in FIGURE 3.4. Here, sequence 1 refers to the haplotype from line 8, sequence 2 to haplotype from line 9 etc. The entries $d_{i,j}$ of the matrix represent the hamming distance between sequences i and j. As it can be seen in the first matrix on the left side, in the first step, the centroidal sequences are formed by sequence two and three, since these two are differing the most from each other (= maximum matrix entry). The remaining sequences will be assigned to one of these two, according to their distance value, which can also be read in the matrix. This leads to two clusters, and thus $\hat{\Omega}_1$ can be determined. In the same way, $\hat{\Omega}_2$ is determined (see distance matrix in the middle of the figure). On the right side, the 'true' tree topology is shown, which refers to the tree presented in Newick format in line 5, FIGURE 3.4. As it can be seen, the 'true' Ω-values are: $\Omega_1 = \min\{|\{3,4,5\}|, |\{1,2\}|\} = 2$, $\Omega_2 = \min\{|\{4,5\}|, |\{3\}|\} = 1, \Omega_3 = 1$.

1, 2, .., 9, 10 ss, 12 ss, 15 ss, 20 ss, 30 ss and 40 ss. In total, 1,000 runs were generated, assuming a sample of size $n = 200$ and effective population size of $N = 10^4$.
Then for each set separately, we determined Ω_1, with the clustering approach explained above using an R-Script written by ourselves* and recorded the average Ω_1 for each set (FIGURE 3.7). In the following, let $\hat{\Omega}_i$ denote the estimated value of Ω_i, $(\cdot)^*$ indicates the normalized value (e.g. $\Omega_i^* = 2\Omega_i/n_i$, $\hat{\Omega}_i^* = 2\hat{\Omega}_i/n_i$).
If we suppose that one segregating site is given, we can obviously form the following two clusters: one consisting of chromosomes carrying the ancestral allele 1, and the other one consisting of those carrying the derived allele 0. The size of the smaller group represents $\hat{\Omega}_1$. We can calculate the theoretically expected estimated $\hat{\Omega}_1^*$ when only one segregating site is used for the cluster estimation. Namely, in this scenario $\hat{\Omega}_1^*$ is equivalent to the minor allele frequency in the sample in each run. By means of the folded SFS (see equation (2.7)), it follows that

$$E[\hat{\Omega}_1^* | (1 \text{ segregating site})] = \frac{\sum_{i=1}^{\lfloor n/2 \rfloor} i \cdot \left(\frac{1}{i} + \frac{1}{n-i}\right)}{a_{n-1}},$$

where n is the sample size and $a_{n-1} = \sum_{i=1}^{n-1} \frac{1}{i}$ is the $(n-1)$-th harmonic number.
For n=200, $E[\hat{\Omega}_1^* | 1\ ss] \approx 0.23$. We obtain a similar value from simulated data ($\approx$ 0.21), (see table in FIGURE 3.7). However, according to equation (3.1) it holds that $E[\Omega_1^*] = 1/2$, hence on average $\hat{\Omega}_1^*$ is underestimated when using 1 segregating site.

Next, we will increase the number of segregating sites by one. Based on the same idea as before, the expected $\hat{\Omega}_1^*$ estimated given two segregating sites can be analytically calculated by means of the folded SFS for two neutral sites.

For the moment, let k be the number of derived alleles at locus one and l the number of derived alleles at locus two, and let the joint two-SFS of two bi-allelic sites be defined as $\xi_{k,l}$ for the sample. One has to be aware of two different cases: the *nested* case, which is when there are chromosomes carrying the two mutations, and the *disjoint* case, when the two mutations are only present in different chromosomes (see

*The original R-script was written by a former Master student S. Bhandari from our lab. Since then, we have performed several modifications and changes to that R-script to meet our requirements. Key differences are:

- If the allocation to one of the two groups is not clearly resolvable, we randomly assign the sequences to one of the two clusters with equal probability.
- Monomorphic sites were excluded (also with regards to Ω_2 and Ω_3 (for determining T_3 later)).
- A detailed assignment of each cluster is given in the output file.
- A window needs to contain at least a given number of SNPs, otherwise it will be extended by 1kb.

FIGURE 3.6.) In the nested case, the haplotypes either carrying the derived version

a) $\xi^N_{1,3}$:	b) $\xi^D_{1,3}$:
–0–0–	–0–1–
–0–1–	–0–1–
–1–1–	–0–1–
–0–0–	–0–0–
–0–1–	–1–0–

FIGURE 3.6: Example of a nested (a) and a disjoint (b) case in a two-locus model, for $n = 5$. Each line represents a chromosome, a 0 indicates that the chromosome has the ancestral allele at that locus, a 1 the derived allele. In both cases, it holds that $k = 1$ and $l = 3$.

at both loci or those carrying the ancestral allele at both loci will form the centroidal sequences for the two clusters, since these two differ the most from each other (they are different at both loci). Haplotypes carrying a derived allele at one locus and an ancestral allele at the other locus are equidistant from both centroidal sequences, meaning that they will be randomly assigned to one of the two clusters. In the disjoint case, haplotypes carrying both mutations are not existent. Here, haplotypes with the derived allele at the first locus and an ancestral allele at the second locus and those haplotypes, which are carrying the opposite combination, will be the centroids of the two clusters. Haplotypes with the ancestral allele at both loci are randomly assigned to one of the two clusters.
The probability of observing k derived alleles at locus one and l derived alleles at locus two, which we define as $\mathrm{P}[\xi_{k,l}]$, is the sum of the nested component $\mathrm{P}[\xi^N_{k,l}]$ and the disjoint $\mathrm{P}[\xi^D_{k,l}]$. In (Ferretti et al., 2018) the respective probabilities were given, where the authors also elaborately provide the derivations of the following equations

$$\mathrm{P}[\xi^{\mathrm{N}}_{k,l}] = \begin{cases} \frac{\beta_n(k)-\beta_n(k+1)}{2} & \text{for } k < l \\ \frac{\beta_n(k)}{2} & \text{for } k = l \\ \frac{\beta_n(l)-\beta_n(l+1)}{2} & \text{for } k > l \end{cases}$$

$$\mathrm{P}[\xi^{\mathrm{D}}_{k,l}] = \begin{cases} \left(\frac{1}{kl} - \frac{\beta_n(k)-\beta_n(k+1)+\beta_n(l)-\beta_n(l+1)}{2}\right)\frac{2-\delta_{k,l}}{2} & \text{for } k+l < n \\ \left(\frac{a_n-a_k}{n-k} + \frac{a_n-a_l}{n-l} - \frac{\beta_n(k)+\beta_n(l)}{2}\right)\frac{2-\delta_{k,l}}{2} & \text{for } k+l = n \\ 0 & \text{for } k+l > n \end{cases}$$

with

$$\beta_n(i) = \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i}.$$

That was the unfolded SFS. Now, we will again turn our focus to the folded SFS. Let k be the number of the minor allele at locus one and l the number of the minor allele

at locus two. For $k < n/2$ and $l > n/2$, or $k > n/2$ and $l < n/2$, the classification of being nested or disjoint will be swapped, when $k < n/2$ and $l < n/2$, or $k > n/2$ and $l > n/2$, the classification remains unchanged. Taking this into account, we can now write down the theoretically expected estimated value $\hat{\Omega}_1^*$ given two segregating sites. This is calculated using following equation(s):

$$\mathrm{E}[\hat{\Omega}_1^* | (2 \text{ segregating sites})] = \mathrm{E}[\xi_{k,l}] = \mathrm{E}[\xi_{k,l}^N] + \mathrm{E}[\xi_{k,l}^D],$$

where

$$\begin{aligned}
\mathrm{E}[\xi_{k,l}^N] = \sum_{k=1}^{n/2}\sum_{l=1}^{n/2} & \left(\min\{k, l, (n - \max\{k, l\})\} + \frac{|k-l|}{2} \right) \cdot \mathrm{P}[\xi_{k,l}^{\mathrm{N}}] \\
& + \Bigg(\min\{(n-k), (n-l), (n - \max\{n-k, n-l\})\} \\
& \qquad + \frac{|(n-k) - (n-l)|}{2} \Bigg) \cdot \mathrm{P}[\xi_{n-k,n-l}^{\mathrm{N}}] \\
& + \left(\min\{n-k, l\} + \frac{k-l}{2} \right) \cdot \mathrm{P}[\xi_{n-k,l}^{\mathrm{D}}] \\
& + \left(\min\{k, n-l\} + \frac{l-k}{2} \right) \cdot \mathrm{P}[\xi_{k,n-l}^{\mathrm{D}}]
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[\xi_{k,l}^D] = \sum_{k=1}^{n/2}\sum_{l=1}^{n/2} & \left(\min\{k, l\} + \frac{(n-k-l)}{2} \right) \cdot \mathrm{P}[\xi_{k,l}^{\mathrm{D}}] \\
& + \left(\min\{(n-k), l, (n - \max\{n-k, l\})\} + \frac{|(n-k) - l|}{2} \right) \cdot \mathrm{P}[\xi_{n-k,l}^{\mathrm{N}}] \\
& + \left(\min\{k, n-l, n - \max\{k, n-l\}\} + \frac{|k - (n-l)|}{2} \right) \cdot \mathrm{P}[\xi_{k,n-l}^{\mathrm{N}}]
\end{aligned}$$

For $n = 200$, $E[\hat{\Omega}_1^* | (2 \text{ segregating sites})] \approx 0.76$. This value is in agreement with the one obtained from simulated data (FIGURE 3.7). Hence, given 2 segregating sites $\hat{\Omega}_1^*$ is on average overestimated.

So far, we have seen that the simulated expected $\hat{\Omega}_1^*$ agreed quite well with the theoretical value. For more than two segregating sites, we determine the expectation $\mathrm{E}[\hat{\Omega}_1^* | (\#\text{ of segregating sites} > 2)]$ by using simulations, because it becomes too complex to be calculated explicitly. The result is illustrated in FIGURE 3.7: With an increase of numbers of segregating sites, the average $\hat{\Omega}_1^*$ eventually approaches 0.5 from above, but never reaches this value. The latter can be explained by the fairly conservative cluster method we are using by always giving preference to clusters of balanced size in not clearly resolvable cases. Moreover, with a minimum number of 10 SNPs the median difference between known Ω_1 and estimated $\hat{\Omega}_1$ is around 0, as illustrated in FIGURE 3.8.

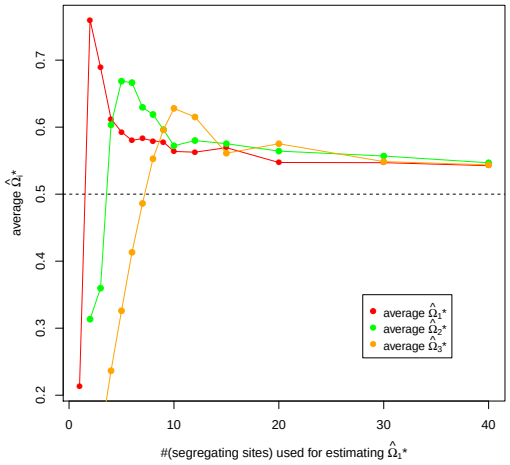


FIGURE 3.7: Average $\hat{\Omega}_1^*$, $\hat{\Omega}_2^*$, $\hat{\Omega}_3^*$ out of 1,000 runs for each scenario, conditioned on the number of segregating sites used for estimating $\hat{\Omega}_1^*$. Dashed horizontal line indicates $E[\hat{\Omega}_i]$ (see (3.2)). For numbers see APPENDIX TABLE A.1

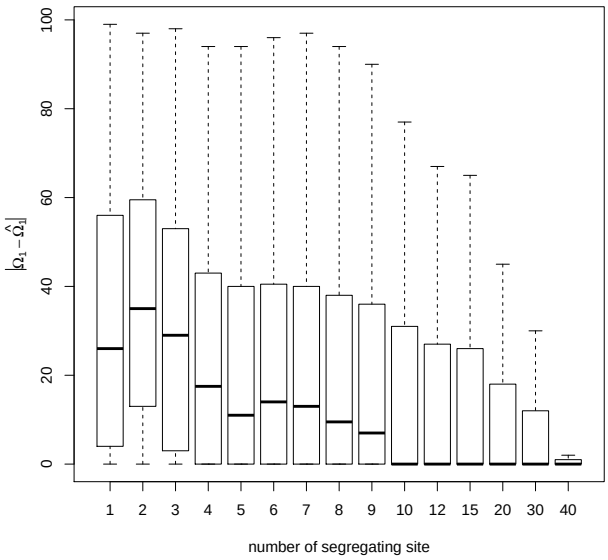


FIGURE 3.8: Absolute difference between Ω_1 and $\hat{\Omega}_1$ (y-axis), where $\hat{\Omega}_1$ was estimated using the number of segregating sites shown on the x-axis. Same simulated data used as before. It can be seen that already with a number of segregating sites of 10, the median is 0.

Recombination events

Too many recombination events within a fragment should be avoided since this might increase the chance of having confounded tree topologies within one window. This in turn leads to a distortion of the clusters. To drastically reduce correlation of coalescent tree topologies along a recombining chromosome, it takes about 15-20

recombination events (Ferretti, Disanto, and Wiehe, 2013). A sample of size n has experienced on average $4Nca_{n-1}$ recombination events (Hudson and Kaplan, 1985), where a_{n-1} is the $(n-1)$-th harmonic number and c the recombination rate per bp. This corresponds roughly to 6,400-8,520 bp to for a sample of size $n = 200$, $N = 10^4$, recombination rate of $c = 10^{-8}$, since

$$\begin{aligned} 4 \cdot 10^4 \cdot 10^{-8} \cdot \text{length(in bp)} \cdot a_{199} &\stackrel{!}{=} 15 \\ \Rightarrow \text{length(in bp)} &\approx 6388. \end{aligned} \tag{3.4}$$

(For 20 recombination events the calculation is similar.) Above 10kb trees are not strongly correlated anymore. (Correlation based on simulations of the test statistic T_3 with distance is given in APPENDIX FIGURE A.1.)

Window size

Summarising the aforementioned results, we can now conclude the following with regards to the appropriate window size and SNP number for the estimation of tree topology:

We have seen that a minimum number of segregating sites is required to get an acceptable estimation of tree cluster. One segregating site leads on average to underestimation, two segregating sites to overestimation. With an increasing number of segregating sites, the estimated value decreases gradually approximating the theoretical expected value of $E[\Omega_1^*] = 1/2$, though a slight overestimation remains which is a consequence of the rather conservative cluster method, giving preference to clusters of balanced size when the clusters are not clearly resolvable. Furthermore, too many recombination events within a fragment should be avoided. This means that on the one hand, fragment length should not be too large, but on the other hand it should contain a minimum number of segregating sites. Starting from a minimum number of ten SNPs the median difference between the known Ω_1 and the estimated $\hat{\Omega}_1$ is around 0, as illustrated in FIGURE 3.8. Using equation (2.2), we expect to see ten SNPs in a magnitude of about $\sim$ 4,260 bp window length, assuming a sample of size $n = 200$, $N = 10^4$, and a mutation rate of $\mu = 10^{-8}$ per bp. Summarising the results, we suggest to estimate tree topology by using a window size of 5,000 bp with a minimum of ten SNPs. If the latter condition was not fulfilled, we increased the window size by 1,000 bp. The maximum window size was set to 10,000 bp (= 10kb). If still less than ten SNPs were within the maximally extended window, we moved on by a step size of 2,500 bp.

It should be pointed out, that the final choice for fragment length rely on the assumption of a recombination rate of $c = 10^{-8}$ per bp per generation and $\mu = 10^{-8}$ per bp per generation, which are the (average) estimates for human (Roach et al.,

2010; Li and Freudenberg, 2009). Therefore, if applying to species with different mutation and recombination rates as assumed above, the parameters must be changed correspondingly.

3.2.2 Quality of cluster assignment

At the moment, we were only interested in how well the estimated cluster size agreed with the true one. But did we also classify the sequences into the correct cluster? Suppose, the true tree topology T is known. Let $|T| = n$, L_1 and R_1 be the left-descendants and right-descendants, respectively, of root v_1, and let $\hat{L}_1$ and $\hat{R}_1$ be the left-descendants and right-descendants of the estimated version of T. Furthermore, let $\Omega = \Omega_1 = \min\{|L_1|, |R_1|\}$ and $\hat{\Omega} = \hat{\Omega}_1 = \min\{|\hat{L}_1|, |\hat{R}_1|\}$. W.l.o.g. $\Omega = |L_1|$ and $\hat{\Omega} = |\hat{L}_1|$, and let in the following the term *maximum overlap* refer to the maximum total number of sequences classified into the correct clusters (left and right). For instance, suppose $\Omega = \hat{\Omega}$ and all sequences belonging to subset L_1 are correctly assigned to subset $\hat{L}_1$, which implies that all sequences in cluster $\hat{R}_1$ are also assigned correctly. In this case, the *maximum overlap* is equal to the sample size n, since all n sequences are classified correctly to the left and to the right cluster, which represents the optimal case. It is also possible, that some $\Omega - k$ sequences are assigned to the 'the wrong' group, namely to $\hat{R}_1$. In this case, the maximum overlap would be $\max\{(|L_1 \cap \hat{L}_1| + |R_1 \cap \hat{R}_1|), |L_1 \cap \hat{R}_1| + |R_1 \cap \hat{L}_1|)\}$.
Suppose $|L_1 \cap \hat{L}_1| = k$ and $|R_1 \cap \hat{R}_1| = n - \hat{\Omega} - (\Omega - k)$, then the size of the overlap is $n - (\hat{\Omega} + \Omega - 2k)$ (see FIGURE 3.9, A). Otherwise, if left and right are 'swapped', the overlap is $\hat{\Omega} + \Omega - 2k$ (see FIGURE 3.9, B). Hence, the *maximum overlap* is the maximum of these two numbers:
$n - (\hat{\Omega} + \Omega - 2k)$ or $(\hat{\Omega} + \Omega - 2k)$.

As a benchmark for the quality of our clustering method, we want to determine the expected maximum overlap we get by chance, given Ω and $\hat{\Omega}$. We assume that k follows a hypergeometric distribution. Hence

$$P[k|\Omega, \hat{\Omega}] = \frac{\binom{\Omega}{k} \cdot \binom{n-\Omega}{\hat{\Omega}-k}}{\binom{n}{\hat{\Omega}}}.$$

Then, the expected maximum overlap, conditioned on Ω and $\hat{\Omega}$, is

$$E[\text{overlap}_{\text{total}}|\Omega, \hat{\Omega}] = \sum_{k=0}^{\hat{\Omega}} \max\{n - (\hat{\Omega} + \Omega - 2k), (\hat{\Omega} + \Omega - 2k)\} \cdot \frac{\binom{\Omega}{k} \cdot \binom{n-\Omega}{\hat{\Omega}-k}}{\binom{n}{\hat{\Omega}}}. \tag{3.5}$$

Based on equation (3.5), we can calculate the expected *maximum overlap*, conditioned on Ω and $\hat{\Omega}$, if we assign the sequences randomly.

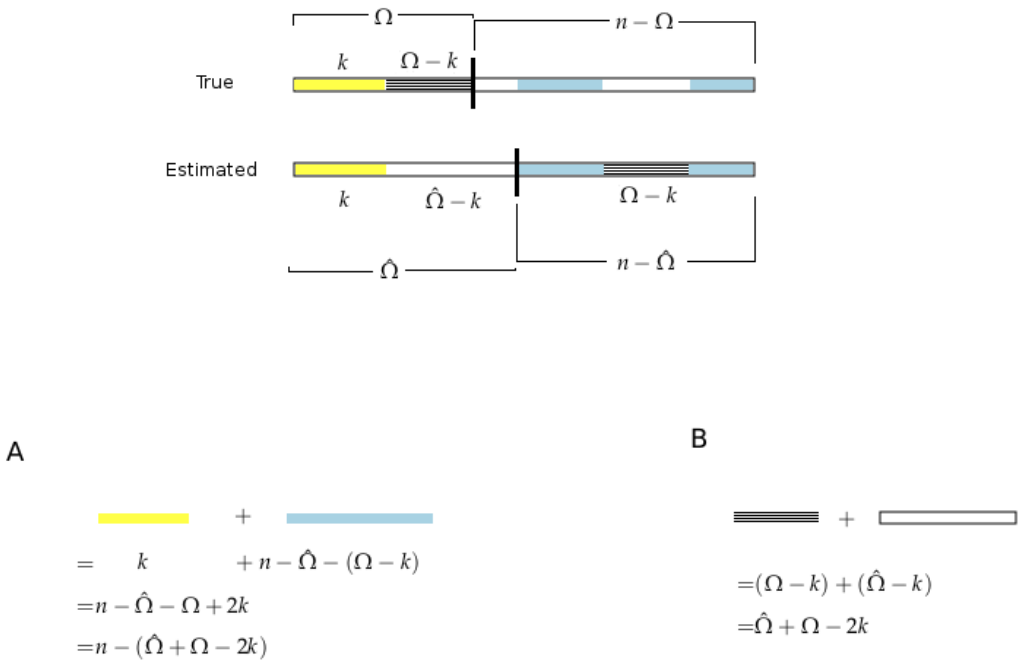


FIGURE 3.9: The two stripes at the top of the picture graphically represents the 'true' cluster and the 'estimated' cluster of a set of size n. The Ω-cluster is further divided into two clusters: the yellow one consisting of k sequences, and the black one consisting of $\Omega - k$ sequences. In this example here, $\hat{\Omega} > \Omega$ (analogous for other cases). To get the overlap of correctly assigned sequences, there are two options (since 'left' and 'right' are interchangeable here): A: $|L_1 \cap \hat{L}_1| + |R_1 \cap \hat{R}_1|$ or B: $|L_1 \cap \hat{R}_1| + |R_1 \cap \hat{L}_1|$. The *maximum overlap* is the maximum of these two.

We estimate tree topologies for 200,000 samples of size $n = 200$ (simulated by *ms*) by using 10 SNPs. We then calculate the average *maximum overlap* conditioned on Ω and $\hat{\Omega}$. We compare this with the expected values calculated using equation (3.5). The result is demonstrated in FIGURE 3.10.

It can be clearly seen, if the estimation of Ω is correct ($\Omega = \hat{\Omega}$) our clustering approach performs very well in assigning all $n = 200$ sequences into the correct cluster, for all $\Omega = \hat{\Omega}$'s. That this is not just a random result, can be seen in particular with increasing Ω. But if $\Omega \neq \hat{\Omega}$, then the quality of the cluster assignment drops quite fast, and is only slightly better than random assignment in extreme cases. Hence, to answer the proposed question from the beginning of this section, it strongly depends on how well we estimate Ω. If $\Omega = \hat{\Omega}$, then the agreement of the assignments is astonishingly good.

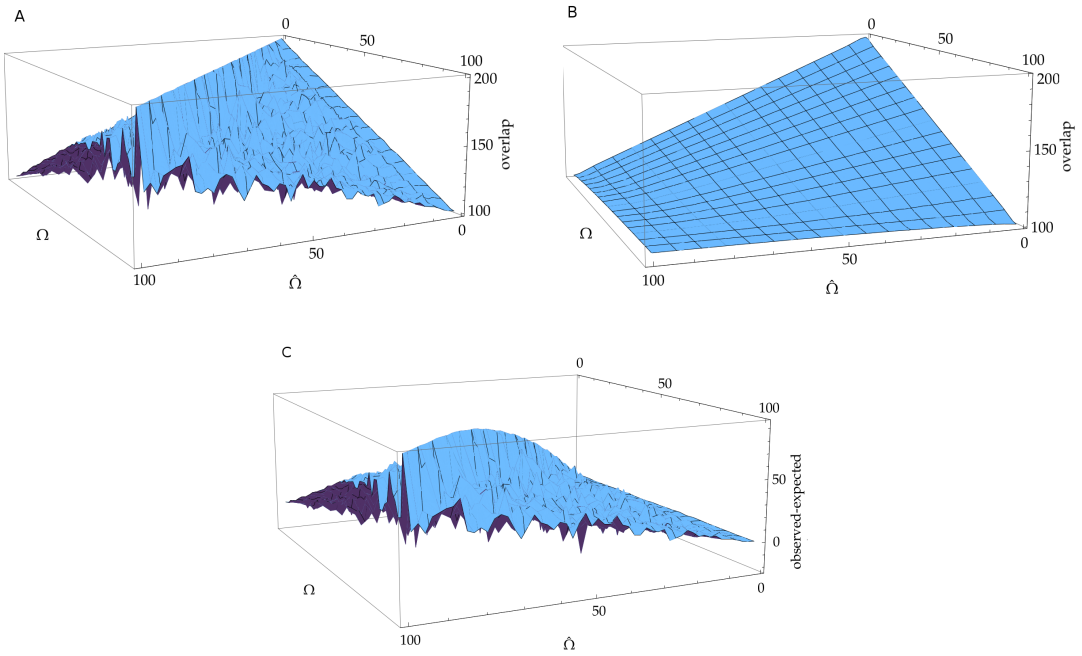


FIGURE 3.10: A: Average maximum overlap of sequences for sample size $n = 200$, conditioned on Ω and $\hat{\Omega}$. If $\Omega = \hat{\Omega}$, sequences are also assigned into the correct cluster. B: Maximum overlap if sequences are randomly assigned into one of the two clusters, given cluster size. C: Difference between observed overlap and expected overlap.

3.3 Robustness to demographic events

3.3.1 Bottleneck events

Distinguishing genomic patterns left by the action of evolutionary forces from those caused by demography has always been challenging, since both events can lead to a reduction in diversity and leave similar footprints behind. Nevertheless, as was already remarked by Li (2011), varying population size does not have an effect on tree topology and hence statistical tests based on tree topology are more robust with respect to this kind of demographic events. This statement is also in accordance with our results tested on simulated data for three different scenarios: neutral, selective sweep and bottleneck. The parameters are $n = 200$, $N = 10^4$, $\theta = 10^3$ and $r = 10^3$, where $r = 4Nc$ is the scaled recombination rate. The choice for $\theta = 10^3$ and $r = 10^3$ refer to a chromosome of size 2.5 Mb with a recombination rate of $c = 10^{-8}$ per bp and mutation rate $\mu = 10^{-8}$ per bp (l =length (in bp)= $2.5 \cdot 10^6$ bp, then $r = 4Nc \cdot l = 10^3$, similarly $l = 2.5 \cdot 10^6$ bp, then $\theta = 4N\mu \cdot l = 10^3$.). For positive selection, we assume that the selected site is located in the very middle of the chromosome, where the strength of selection for the selected allele is given by $\alpha = 2Ns = 1000$, where s is the selection coefficient , and $\tau = 0.0001$, which is the time since the completion of the sweep. For population bottlenecks, we assumed severity 1 and onset 0.01.

Box 3.3.1: Extracting windows from simulated data output.

To cut the sequences from the *msms*-output in appropriate windows, we used the option *mscut* contained in the program package *coatli* provided by A. Klassmann, which can e.g. be downloaded on https://sourceforge.net/p/coatli/wiki/Home/ (or also see (Ferretti et al., 2018)). In general, *mscut* filters ms-output, and retains only those segregating sites whose positions fall into a specified interval. For example, the following command line

```
msms -ms 5 1 -N 10000 -s 1000 | mscut 0 0.01
```

gives the output:

```
1  msms -ms 5 1 -N 10000 -s 1000
2  [null]Window: [0.0000,0.0100[
3
4  //
5  segsites:  10
6  positions:  0.0008 0.0009 0.0015 0.0026 0.0031 0.0068 0.0071
   0.0076 0.0085 0.0090
7  0110010000
8  0100101000
9  1000000101
10 1001000111
11 1000000101
```

First, one run of sample of size $n = 5$, $N = 10^4$ for a chromosome containing 1,000 SNPs is generated. The positions of the SNPs are given on a scale of $(0, 1)$ (compare with 3.4). The command `mscut 0 0.01` retains all SNPs located between 0 and 0.01. The option *msfs* and ntx contained in the same program package *coatli* allows the calculation of Tajima's D. Hereby, *msfs* first calculates the standard frequency spectrum out of the output, then ntx computes Tajima's D value.

The result of the application of the T_3-test for each three scenarios is demonstrated in FIGURE 3.11. When tree topology is estimated based on SNP data, it produces on average slightly larger T_3-values than the true one. This can be explained by the fairly conservative cluster method we are using in always giving preference to clusters of balanced size in not clearly resolvable cases. Furthermore, for reasons of comparison, we also calculated Tajima's D for each set. In FIGURE 3.11 it can be seen, when a population has gone through a bottleneck, the T_3-test is not affected. When tree topology is estimated, it even goes in the opposite direction producing rather positive values. In comparison, Tajima's D is becoming heavily negative, leading to false positives under a bottleneck event.

To cut the simulated sequences into fragments and calculate Tajima's D, we used the program package *coatli* provided by A. Klassmann (Ferretti et al., 2018) (see box 3.3.1).

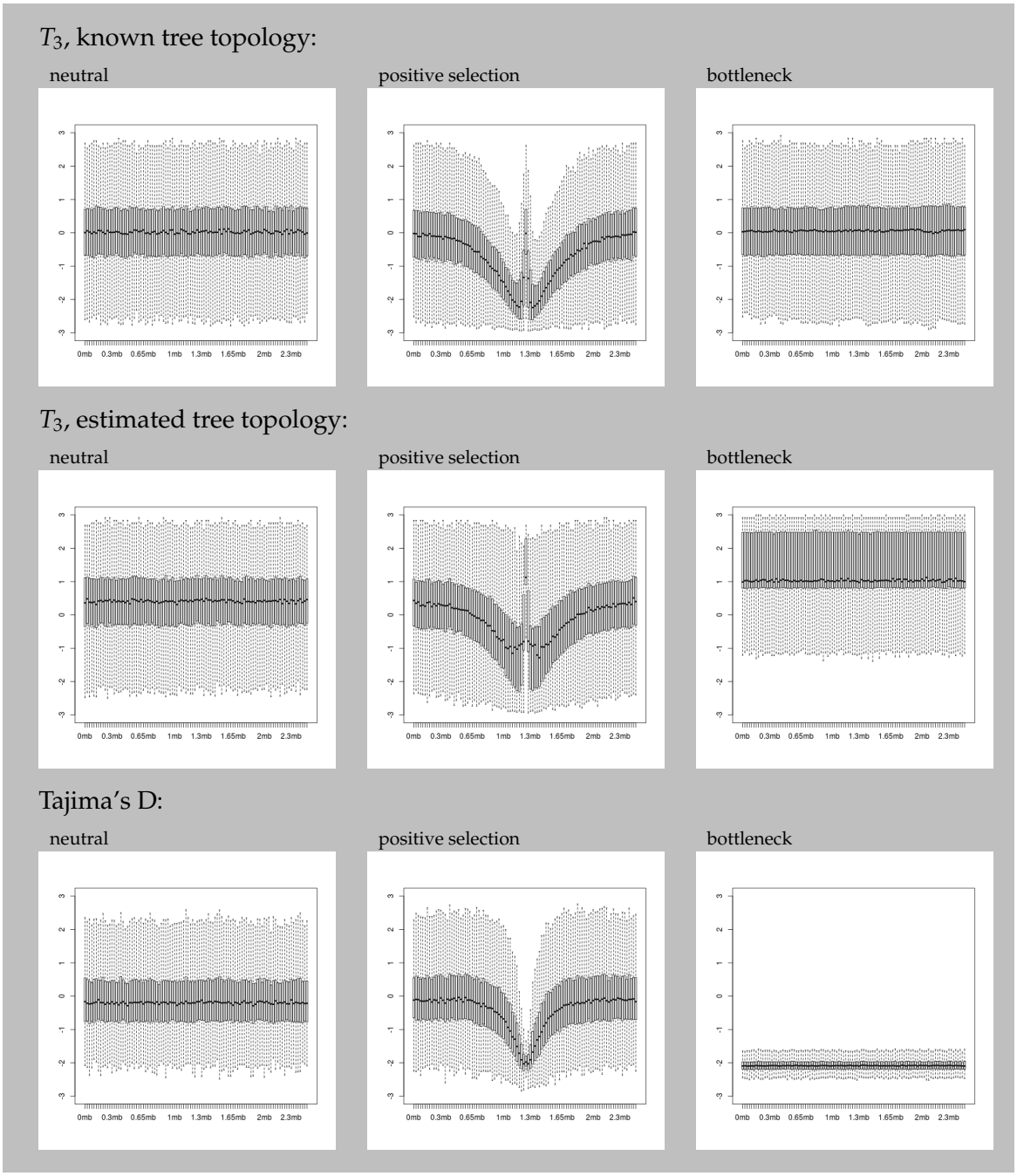


FIGURE 3.11: T_3-profile calculated from simulated data along a recombining chromosome for three different scenarios: neutral, positive selection on a selected site located in the middle of the chromosome, and population bottleneck with severity 1 and onset 0.01. Each scenario is shown for known T_3-values, for estimated T_3-values and for reason of comparison Tajima's D.

3.3.2 Migration events

Another concern for tree-topology based tests are migration events: When a lineage migrates from one subpopulation to another, it may not coalesce with any other

lineages before the most recent common ancestor. Such cases can also cause unbalanced tree topologies. We examined sampling from a population divided into two sub-populations with varying migration rates and varying sampling schemes. Samples were generated using *ms*. As previously, parameters were set such that $N = 10^4$, $n = 200$, $\mu = 10^{-8}$ per nucleotide per generation and recombination rate $c = 10^{-8}$ per nucleotide per generation.

It holds that $n = n_1 + n_2$, where n_1 refers to the number of chromosomes sampled from the first subpopulation and n_2 refers to the number of chromosomes sampled from the second subpopulation. As can be seen in TABLES 3.1 and 3.2, T_3 is affected by the existence of population substructure. When the sampling scheme is heavily biased ($n_1 = 195$ and $n_2 = 5$) and migration rate is low ($4Nm = 0.4$ or $4Nm = 0.04$), T_3 is quite negative (even compared to the selective sweep scenario) leading to a high increase of false negatives. When sampling all chromosomes from only one subpopulation, $n_1 = 200$ and $n_2 = 0$, T_3 is quite robust, however when migration is $4Nm = 0.4$, T_3 seems to be slightly affected (see also APPENDIX FIGURE A.2,A.3 and A.4). In TABLES 3.1 and 3.2 the values for the neutral (panmictic) scenario and the selective sweep scenario from the same data from previous section 3.3.1 are given for reasons of comparison.

4Nm	**subpopulation sample size**	**average T_3-value (known)**	**average T_3-value (estimated)**
4	$n_1 = 180$ and $n_2 = 20$	-0.0968	0.3281
0.4	$n_1 = 180$ and $n_2 = 20$	-0.4899	-0.1302
0.04	$n_1 = 180$ and $n_2 = 20$	-0.5819	-0.2587
4	$n_1 = 195$ and $n_2 = 5$	-0.1188	0.3219
0.4	$n_1 = 195$ and $n_2 = 5$	-0.6254	-0.1968
0.04	$n_1 = 195$ and $n_2 = 5$	-0.8534	-0.5083
4	$n_1 = 200$ and $n_2 = 0$	-0.1031	0.3367
0.4	$n_1 = 200$ and $n_2 = 0$	-0.3234	0.1111
0.04	$n_1 = 200$ and $n_2 = 0$	-0.0826	0.4688
-	neutral scenario (panmictic)	0.0204	0.4376
-	sweep scenario ($\alpha = 1000$)	-0.6283	0.0588

TABLE 3.1: Average T_3-value (known tree topology and estimated tree topology) for different scenarios: substructured populations with varying migration rates and varying sampling schemes, neutral (panmictic) and selective sweep scenario. Average of 1,000 runs.

4Nm	subpopulation sample size	average 1%-threshold (known)	average 1%-threshold (estimated)
4	$n_1 = 180$ and $n_2 = 20$	-2.3007	-2.0139
0.4	$n_1 = 180$ and $n_2 = 20$	-2.5252	-2.4357
0.04	$n_1 = 180$ and $n_2 = 20$	-2.3747	-2.3309
4	$n_1 = 195$ and $n_2 = 5$	-2.3237	-2.0317
0.4	$n_1 = 195$ and $n_2 = 5$	-2.7142	-2.5266
0.04	$n_1 = 195$ and $n_2 = 5$	-2.6516	-2.5298
4	$n_1 = 200$ and $n_2 = 0$	-2.314	-2.0087
0.4	$n_1 = 200$ and $n_2 = 0$	-2.4877	-2.2841
0.04	$n_1 = 200$ and $n_2 = 0$	-2.3088	-2.0626
-	neutral scenario (panmictic)	-2.18	-1.88
-	sweep scenario ($\alpha = 1000$)	-2.71	-2.5259

TABLE 3.2: Average empirically determined 1%-threshold of T_3 (known tree topology and estimated tree topology) for different scenarios: substructured populations with varying migration rates and varying sampling schemes, neutral (panmictic) and selective sweep scenario. Average of 1,000 runs.

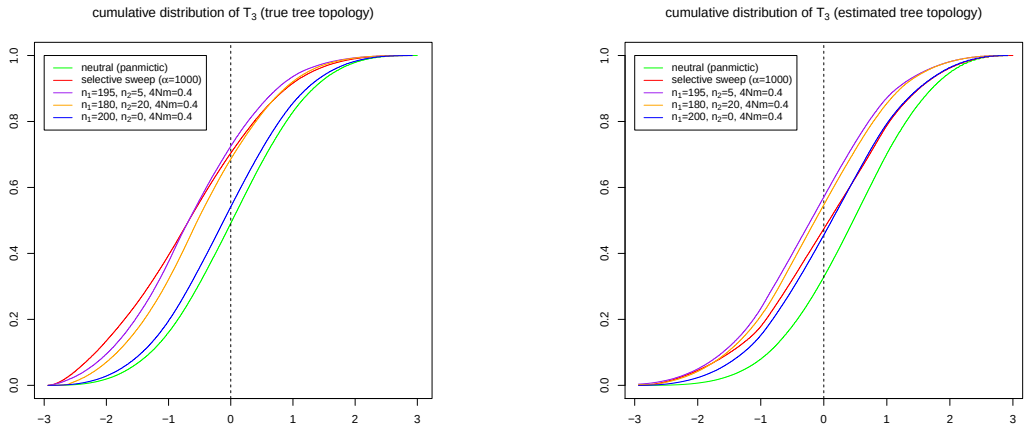


FIGURE 3.12: Cumulative distribution of T_3 for each sampling scheme with migration rate $4Nm = 0.4$.

3.4 Power of the $\mathbf{T_3}$-test

Under neutral assumptions, the probability of observing highly unbalanced tree shapes is quite low (Kirkpatrick and Slatkin, 1993; Blum and Francois, 2006). However, like all neutrality tests, the T_3-test suffers from false positive results.

To check how many of the identified regions are true positives, we simulated a chromosome of size 2.5 Mb experiencing a completed selective sweep with varying strength of selection. As previously, we assumed $n = 200$, $N = 10^4$, mutation rate per bp $\mu = 10^{-8}$ and a recombination rate per bp $c = 10^{-8}$. Simulations were performed with *msms* as in section 3.3.1, 1,000 runs for each setting. The positively

selected site was placed in the middle of the chromosome. For each of the 1,000 runs, we empirically determined the 5% threshold and 1% threshold. Afterwards, when we found a window with a T_3-value below the respective threshold, we recorded the position of this window. The result is illustrated in FIGURE 3.13 for the known tree topology: The y-axis represents the counts of how often a window (located on the x-axis) was significant. As expected, under positive selection we see two peaks located around the selected site. What can be clearly seen, is that the power of the T_3-test depends on the distance to the selected site (see table 3.3). On average, taking a 1% threshold, around 78.86% - 86.12% of the windows identified as being significant were found to be within a distance of 250 kb from the selected site (see table 3.3). However, as just mentioned, it strongly depends on the distance we take into consideration to determine the actual selected site. Still an average of around 20% (by a threshold of 1%) falls outside aforementioned region.

Data set	threshold	average threshold-value	max. 250 kb distant from selected site	max. 500 kb distant from selected site	> 500 kb distant
$\alpha = 500$	5%	-2.17	70.64%	81.56%	18.44%
	1%	-2.58	78.86%	86.66%	13.34%
$\alpha = 1000$	5%	-2.45	79.34%	90.3%	9.7%
	1%	-2.71	85.18%	93.56%	6.44%
$\alpha = 2000$	5%	-2.59	80.23%	92.98%	7.02%
	1%	-2.76	86.12%	94.84%	5.16%
neutral	5%	-1.63	19.60%	39.26%	60.74%
	1%	-2.18	19.60%	39.06%	60.94%

TABLE 3.3: This table shows where on the chromosome, on average, a window with a T_3-value below the respective threshold was found, with regards to the selected site (in the neutral case: middle of the chromosome).

Moreover, in FIGURE 3.14 it can be seen, that if we only consider single windows (regardless of their position from the selected site), the test is not very effective. Suppose, we take a cut-off value of $T_3 = -2.0$, the false positive rate is around 0.019, however the power is only (maximum) 0.23. In the following we want to investigate how the T_3-test can be improved.

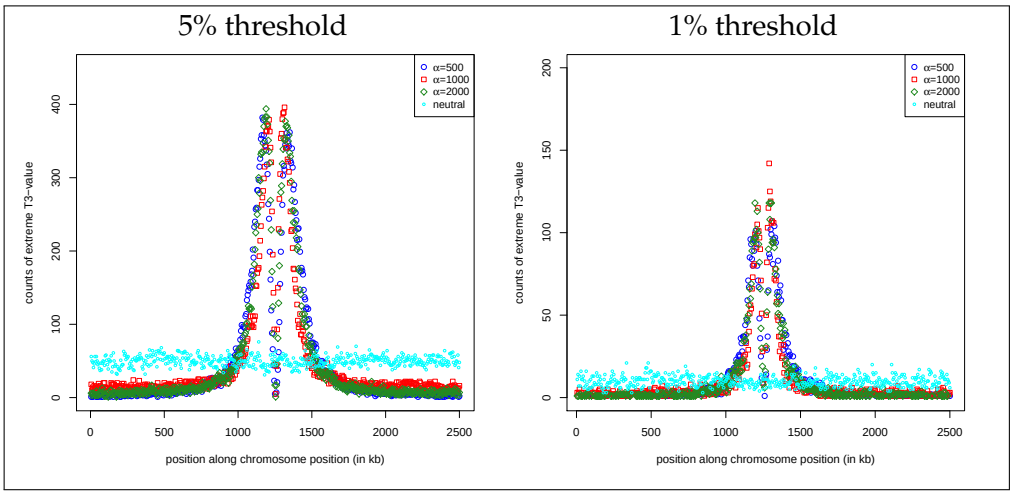


FIGURE 3.13: Absolute counts of how many times out of 1,000 simulations a specific region, shown on the x-axis, was referred to as 'being a significant region'. The selected site is located in the middle of the chromosome.

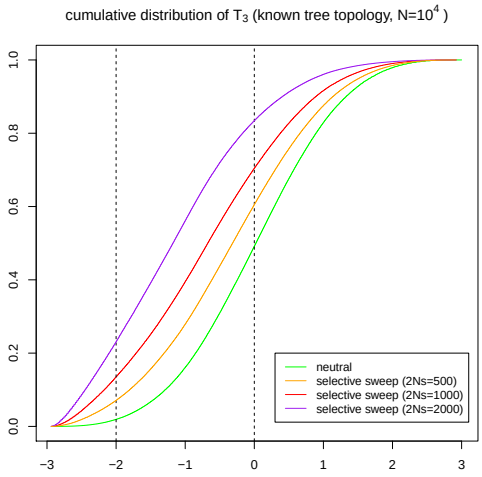


FIGURE 3.14: Shown in this figure is the cumulative frequency distribution of the T_3 values for different simulated data sets. For each scenario, we simulated 1000 runs, with parameter $n = 200$, $N = 10^4$, and recombination rate per bp $c = 10^{-8}$. For positive selection, we assume $\alpha = 500$, $\alpha = 1000$, $\alpha = 2000$, respectively.

3.4.1 Corroborate significance

Re-sampling strategy

For the reconstruction of phylogenetic trees, bootstrapping has long become a common feature to assign confidence to the inferred tree topology (Felsenstein, 1985). Here, we are concerned with the question whether bootstrapping or related re-sampling techniques can contribute to reducing false positives in our case. Of particular interest to us is, if unbalanced tree topologies under neutrality have distinguishable topological features with regards to their subtree structure compared to unbalanced coalescent tree topologies produced by a selective sweep. Hence, the idea

is to re-construct the genealogy of random subsamples of the original sample, so-called *induced subtrees*. The most unbalanced type of tree topology is if Ω_i=1, for all $i = 1, ..., n-1$.

Such a tree is called a *caterpillar tree* (e.g. see FIGURE 3.15). Under the standard neutral model, this tree shape is very unlikely to appear by chance (Blum and Francois, 2006; Kirkpatrick and Slatkin, 1993). A large excess of singleton mutations which is a typical characteristic of a selective sweep, results in the estimation of a star-like tree which takes a caterpillar shape when forced to be binary. Caterpillar trees and their induced subtrees have been analysed before (Disanto and Rosenberg, 2016; Kirkpatrick and Slatkin, 1993), its induced subtrees are also highly unbalanced. This in turn means that a re-sampling strategy surely helps to corroborate candidate regions found. However, as it was already mentioned, the chances to observe such a tree shape in practice is extremely low.

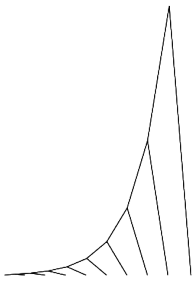


FIGURE 3.15: Example of a caterpillar tree, $n = 10$.

In the following, we tested on simulated data if subtree topologies under neutrality are significantly distinguishable from subtree topologies under selection. To analyse this, we subjected the found regions (with a significance level of 0.01 and 0.05, respectively) in the simulated data sets from subsection 3.4 to a re-sampling strategy. Therefore, independent subsamples of size $n' = 40$ were randomly drawn 100 times, and T_3-value was calculated each time. Then for each region, we determined how many out of the 100 times re-confirmed the candidate region. In the end we reported those, in which at least 30 out of 100 subsamples re-confirmed the candidate.

The following table shows how many of the regions, which were significant in the first step using 'whole' sample (see table 3.3), survived after applying the re-sampling strategy just explained. As in the section before, we demonstrate this for the known tree topology.

As can be seen in TABLE 3.4, on average, unbalanced tree topologies under neutrality seem not to have significant distinguishable topological features with regards to their subtree structure compared to unbalanced coalescent tree topologies produced by a selective sweep. Since, if it were true, windows with a T_3-value below the threshold found close to the selected site should be re-confirmed at a much higher rate than those located far away. However, our results presented in TABLE 3.4 could not confirm this. The reason might be, as mentioned at the beginning of the section, that a re-sampling strategy is only helpful for extreme cases, like caterpillar trees. However to observe a caterpillar tree is extremely unlikely in practice. Deeper analysis is needed concerning 'non-extreme' cases, which are more common to find.

Data set with	threshold	average threshold-value	max. 250 kb distant from selected site	max. 500 kb distant from selected site	> 500 kb distant
$\alpha = 500$	5%	-1.91	63.89%	63.60%	56.98%
	1%	-2.37	22.57%	21.88 %	9.45%
$\alpha = 1000$	5%	-2.14	52.23%	52.92%	38.45%
	1%	-2.48	11.36%	11.14 %	2.79%
$\alpha = 2000$	5%	-2.36	30.74%	33.55%	24.44%
	1%	-2.58	4.02%	4.68%	2.71%
neutral	5%	-1.55	74.37%	74.43%	74.14%
	1%	-2.04	45.17%	46.24%	45.98%

TABLE 3.4: This table shows, how many out of the previously significant regions 3.3 were confirmed after the re-sampling strategy.

Based on our simulation results, and the long running time and large memory needed for this strategy, we then focused on a different approach.

Log likelihood ratio test approach: The $\mathbf{LR_{T_3}}$-test

While a beneficial mutation increases in frequency and is getting fixed in the population, linked neutral variants also increase in frequency, sweeping out the diversity around the selected site. As the distance from the selected site grows, recombination events will allow linked neutral sites to recombine away. However, the level of genomic variation is maintained over a longer chromosomal distance around the selected site than under neutrality; the basis used for haplotype-frequency based neutrality tests, e.g. (Sabeti et al., 2002). Linkage is elevated in regions close to a selected site, recombination events are more rare. That in turn also means that genealogical tree topology should be maintained over a longer chromosomal distance. The probability of observing unbalanced tree topologies in multiple consecutive regions is higher for selected sites than under neutrality. Therefore, we asked: When a candidate region was found on the chromosome, that is for this region its T_3-value is below a previously determined threshold q, how likely is it that also for the following k_l flanking regions to the left and k_r flanking regions to the right, the respective T_3-values of these flanking regions are also below q?
In case of positive selection, the probability that T_3-values are also below q should be higher (compared to the neutral case) for $k_l = 1$ and $k_r = 1$ (the immediate neighbours) and decrease slowly (compared to the neutral case) with growing distance to the 'focal' region which is the region where we start from.
Hence, the idea is to take not only the T_3-value of one window, but also the surrounding ones into account and to construct a test statistic based on the concept of likelihood ratio tests.
Likelihood ratio tests give an idea about how many times less likely the data are

seen under a null model H_0 compared to an alternative model H_1. Here we have

$$H_0 = \text{neutral evolution}$$
$$H_1 = \text{positive selection.}$$

To construct the likelihoods $\hat{P}(\cdot|H_0)$ and $\hat{P}(\cdot|H_1)$, we used our previously simulated data (generated under the neutral scenario and generated under the selective sweep scenario, assuming $\alpha = 1000$) and proceeded as follows:

1. Determine the 1% threshold value from the simulated data under the null hypothesis, namely under the neutral scenario.
2. Start screening the data set from left to right (along the chromosome):

   When a significant region is found (T_3-value is below the 1% threshold determined in 1.) record this region (in the following we will refer to this region as the 'focal region'), and inspect adjacent regions to the left and to the right.

   Record whether the $k_{l/r}$-th neighbour window from the focal region has a T_3-value below the 1% threshold or not. (Note: $k_{l/r} = 1, ..., m_{l/r}$, where $, m_{l/r}$ is the number of consecutive windows investigated to the left and to the right starting from the focal region. The index l and r stand for "left" and "right" side, respectively. See also FIGURE 3.16).
3. Repeat 2. until the end of the chromosome is reached.
4. Calculate the average of how often a T_3-value below the 1% threshold will be found with distance $k_{l/r}$. In the end, obtain a probability distribution of finding another region with a T_3-value below the 1% threshold with respect to the distance of the focal region.

Repeat with simulated data under the alternative hypothesis, namely under the selective sweep scenario.

The construction steps were performed for 1,000 runs of simulated data under neutral assumptions and simulated data under the selective sweep scenario, which we generated in section 3.3*

*As before samples were generated with *msms*. A chromosome of length 2.5 Mb was simulated, mutation and recombination rate as before. The command for the selective sweep scenario was: *ms 200 1000 -N 10000 -t 1000 -r 1000 500 -T -SAA 1000 -SAa 500 -SF 1e-4 -Sp 0.5*. $1 \leq k_{l/r} \leq 250$ and each fragment stands for a window of size 5 kb.

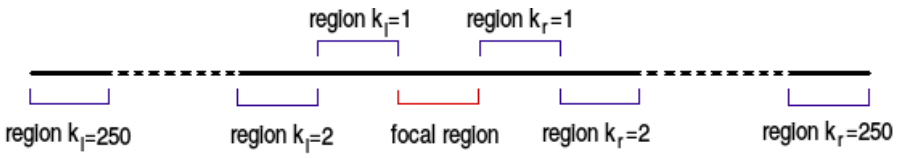


FIGURE 3.16: A simple visualization of step 2. The black line indicates a chromosome, which was divided into 501 fragments, and the red/blue lines indicate the fragments/regions of the chromosome, which a T_3-value is referred to. The red region indicates a region, where a T_3-value under the given threshold was found, suppose it was found in the very middle of the chromosome, thus $k_{l/r} = 1, ..., 250$ for both left and right side from the focal region.

FIGURE 3.17 illustrates the previously computed conditional probabilities. As we can see, it is more likely to observe unbalanced trees in multiple adjacent windows under the selective sweep scenario than under neutrality. (Under neutrality, the probability is almost 0). In the following, we worked with the probabilities calculated for the estimated tree topologies.

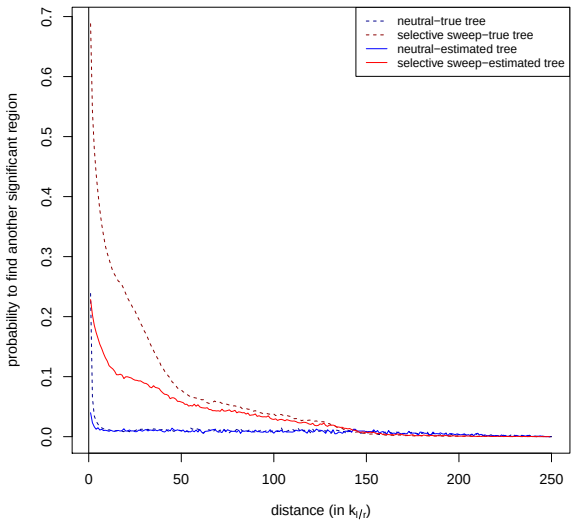


FIGURE 3.17: Probability of finding another highly unbalanced tree at window distance x, given that one was found at $x = 0$.

We defined $(p_{n1}, ..., p_{ni}, ..., p_{n250})$ as the probabilities that, given a window with T_3 below the threshold was found, that neighbour window i from the focal window also has T_3 below the threshold under neutrality, and analogue for p_{si} for the selective sweep scenario.

With this background, we composed a test statistic based on likelihood ratio tests. Each of these two models was separately fitted to the data and the log-likelihood

was recorded, which we defined as LR_{T_3}, and is given by the following equation:

$$\begin{aligned} LR_{T_3} &= -2 \cdot \ln \left(\frac{\hat{P}(\text{data}|H_0)}{\hat{P}(\text{data}|H_1)} \right) \\ &= -2 \cdot \ln \left(\frac{\prod_{i=1}^{k_{l/r}} p^*_{ni}}{\prod_{i=1}^{k_{l/r}} p^*_{si}} \right) \end{aligned} \tag{3.6}$$

Hereby, $p^*_{.i} = p_{.i}$, if the T_3-value in window i is below the threshold in the observed data, otherwise $p^*_{.i} = (1 - p_{.i})$.

Further on, we generated new simulation data, again 1,000 runs under neutral assumptions and 1,000 runs under assumptions of positive selection, using the same parameters as before. These were our test data sets.

We screened the new neutral data set and the new selective sweep data set separately. When a 'focal window' was found, we looked, if procurable, 100 adjacent regions to the left and to the right from the focal region, calculated the likelihood of observing these data under neutrality and under positive selection by means of the previously established probability distributions, afterwards we calculated the log likelihood ratio LR_{T_3} with equation (3.6). For an example work-flow see box 3.4.1.

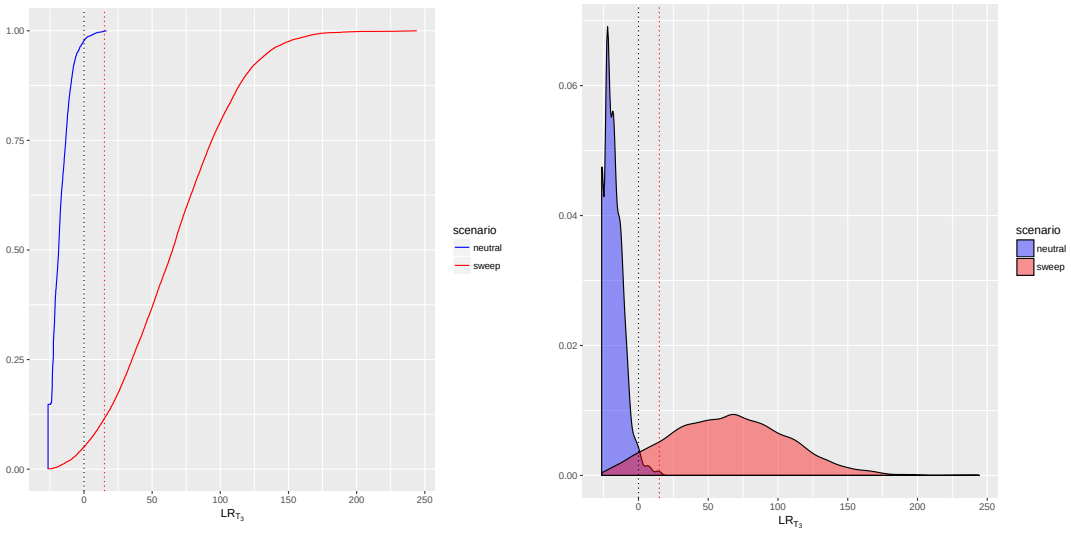


FIGURE 3.18: The picture on the left side illustrates the cumulative distribution of LR_{T_3} and on the right side the density plot of LR_{T_3} for the neutral scenario (blue) and for the selective sweep scenario (red), for the estimated tree topology. For the neutral case, $LR_{T_3} \geq 0$ for 2.26%. For the selective sweep scenario, $LR_{T_3} \geq 0$ for 94.98%. To reduce the false positive rate, we set the threshold of LR_{T_3} at 15 (dashed red line). For the neutral scenario $LR_{T_3} \geq 15$ holds for 0.0007%, for the sweep case that holds for 88.41%.

The result is illustrated in FIGURE 3.18.

We empirically determined the power of this test, and found that by setting the threshold of LR_{T_3} to 0, we get a false positive rate of 2.26%, and a power of 94.98%.

To reduce the false positive rate, we decided to set the threshold-score to 15. In such way, we could reduce the false positive rate to 0.0007%, but at price of reduced power (88.41 %).

Box 3.4.1: Example of calculating $LR_{T_3}-$score.

In the following table the probabilities that the $k_{l/r}$-th neighbouring window (with $k_{l/r}$=1,...,5) from the focal region has a T_3-value below the empirical determined 1% threshold ($:= q_{1\%}$) are given (estimated tree topology.):

$k_{l/r}$ -th neighbour window from focal region	Probability under neutral scenario	Probability under sweep scenario
1	0.0398992 ($=: p_{n1}$)	0.228805 ($=: p_{s1}$)
2	0.0232045 ($=: p_{n2}$)	0.203321 ($=: p_{s2}$)
3	0.0162747 ($=: p_{n3}$)	0.18632 ($=: p_{s3}$)
4	0.0128097 ($=: p_{n4}$)	0.174434 ($=: p_{s4}$)
5	0.0153297 ($=: p_{n5}$)	0.164617 ($=: p_{s5}$)

Suppose, we focus on two windows located at different chromosomal positions, in the following labelled as focal window A and B respectively, where a T_3-value under threshold was found. We now look at 5 adjacent windows to the left and 5 adjacent windows to the right of the focal window and record each time whether the respective T_3-value was below the empirically determined 1% threshold given by $q_{1\%}$ or not:

$k_{l/r}$ -th neighbour window		← to the left					focal window	to the right →				
		5	4	3	2	1		1	2	3	4	5
A:	$T_3 \leq q_{1\%}$?	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
B:	$T_3 \leq q_{1\%}$?	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗

✓= "true" ✗= "false"

By multiplying the probabilities given in the table above we get:

Probability to observe combination around focal window A under neutral scenario
$= (1-p_{n5})p_{n4}p_{n3}p_{n2}p_{n1}p_{n1}p_{n2}p_{n3}(1-p_{n4})p_{n5}$
$= 4.33374e-14$

Probability to observe combination around focal window A under sweep scenario
$= (1-p_{s5})p_{s4}p_{s3}p_{s2}p_{s1}p_{s1}p_{s2}p_{s3}(1-p_{s4})p_{s5}$
$= 1.48785e-06$

Probability to observe combination around focal window B under neutral scenario
$= (1-p_{n5})(1-p_{n4})(1-p_{n3})(1-p_{n2})(1-p_{n1})(1-p_{n1})(1-p_{n2})(1-p_{n3})(1-p_{n4})(1-p_{n5})$
$= 0.804215$

Probability to observe combination around focal window B under sweep scenario
$= (1-p_{s5})(1-p_{s4})(1-p_{s3})(1-p_{s2})(1-p_{s1})(1-p_{s1})(1-p_{s2})(1-p_{s3})(1-p_{s4})(1-p_{s5})$
$= 0.118871$

With equation (3.6), it follows
LR_{T_3}-score of A = 34.7032

LR_{T_3}-score of B = −3.82366 .

3.4.2 LR_{T_3}-test and migration events

In section 3.3.2, we have seen that substructured population and low migration rate affects the T_3-test. Although the LR_{T_3}-test is also affected by migration events, it still performs better than the T_3-test. For instance, when sampling all n chromosomes from only one subpopulation, $n_1 = 200$ and $n_2 = 0$, and by setting a stricter threshold, e.g. $LR_{T_3} = 35$, the false negative rate when migration rate $4Nm = 0.4$ (which was the case influencing the T_3-test most) is only around 0.03, whilst LR_{T_3} has still a high power rate (around 0.75).

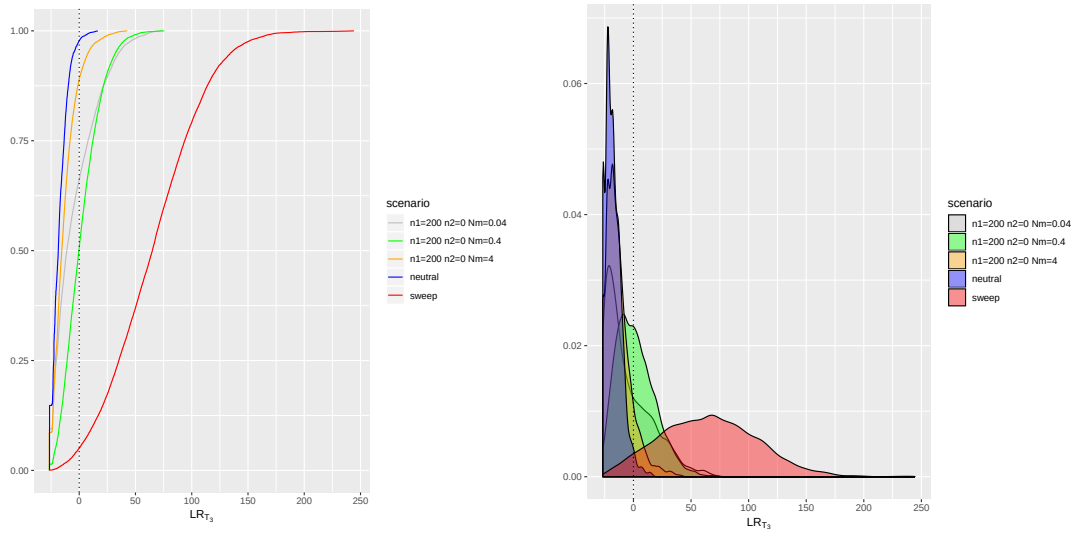


FIGURE 3.19: On the left side: cumulative distribution of LR_{T_3}. On the right side: Density plot of LR_{T_3}. Estimated tree topology.

The distributions for case ($n_1 = 180$ and $n_2 = 20$) and ($n_1 = 195$ and $n_2 = 5$) are given in the APPENDIX FIGURE A.5 and A.6.

3.5 Side note on time point in detection of selective sweep

Thus far, when talking about 'selective sweeps', we referred this term to a 'completed' hard sweep, that is, when the advantageous mutation arises at some time point in the population, quickly increases in frequency and subsequently becomes fixed. However selective sweeps can also be 'incomplete', they have not reached fixation yet and are still ongoing. Whilst methods aiming to detect completed selective sweeps can use the concept of the hitch-hiking process introduced by Maynard Smith and Haigh (1974), see also section 2.3.1, genomic signatures of incomplete sweeps are less clear; several studies exist focusing on identifying incomplete

sweeps (Sabeti et al., 2006; Voight et al., 2006; Ferrer-Admetlla et al., 2014; Vy and Kim, 2015). Paying attention to this mode of selection is essential: Studies of human demography have suggested that the dispersal of humans out of Africa started only 50,000 -100,000 years ago, see e.g. (Nielsen et al., 2017). Within this period of time humans were confronted with new environments and were exposed to constraints like extreme climate conditions, diseases or volatile food supply. Factors like that are supposed to lay the foundation for adaptation and selection. Nevertheless, the amount of time may be too short for new beneficial mutations to occur and to get fixed, giving rise to the conclusion that complete sweeps may be rare in human history (Ferrer-Admetlla et al., 2014).

On the other hand, when some generations have already passed since fixation, the level of diversity around the selected site might have recovered from the sweep through an influx of new mutations, washing out the erstwhile clear signature of the sweep and thus hindering its detection.

In this section, we want to analyse, to what extent time point matters in detecting selective sweep using the LR_{T_3}-test.

Therefore, we used simulated data provided by Yichen Zheng (Y. Zheng, unpublished data, 2018): The data were generated with a customised forward-in-time algorithm. The parameters were set in such way that a DNA sequence of length 600kb was simulated where the mutation rate was $\mu = 10^{-8}$ per bp per generation, the recombination rate was $c = 10^{-8}$ per bp per generation, selection coefficient $s = 0.02$ and population size $N = 10^4$. In total, 100 runs were generated and evolved until 5,000 generations after the fixation time of the selected allele. During one run, twelve so-called 'snapshots' of the genotypes of each sequence were recorded. These 'snapshots' were performed at following time points: when the frequency of the selected allele reached 20%, 40%, 60%, 80%, 99.5% fixation, then 1,000, 2,000, 3,000, 4,000, and 5,000 generations after the selected allele reached 99.5%. On average, out of the 100 runs, it took 269 generations for the selected allele to reach a frequency of 20%, 317 generations to reach a frequency of 40%, 358 generations to reach a frequency of 60%, 407 generations to reach a frequency of 80%, 595 generations to reach a frequency of 99.5% and 1,103 generations to get fixed in the population. From each population snapshot 50 random samples were taken.

First, we determined the T_3-values for the twelve data sets in the same manner as before: With a sliding window approach of window size 5kb and step size 2.5kb, we estimated the respective tree topology for each window. If the window contained less than 10 SNPs, we increased the window by 1kb, however the total window size was not to exceed 10 kb. The result of the T_3-values is illustrated in FIGURE 3.20. Interestingly, the most extreme T_3-value was obtained when the frequency of the selected site reached 80%. If the frequency of the selected site increased, and finally

got fixed, the signal seemed not to be striking. The sparseness of data in the region around the selected site in cases, when the selected allele reached a frequency of 99.5% (which happened on average after 595 generations) until 1,000 generations later when it was fixed, can be attributed to the strong reduction of polymorphism data around the selected site, and therefore no tree can be estimated here.

The boxplots in FIGURE 3.20 indicate the strength of signal depends on time. The strongest signal seems to occur when the selected site has reached a frequency of 80%, thus when the sweep is yet incomplete. But with regards to several recent studies claiming that complete sweeps are rare and incomplete sweeps are dominant, this might be a benefit of the test statistic T_3. Note the rather rapid increase from 60% to 80% and the rapid decrease after fixation.
Further on, we applied the LR_{T_3}-test on these 12 data sets. The result is illustrated in FIGURE 3.21 and FIGURE 3.22. The most significant LR_{T_3}-score can be found when the selected site reached a frequency of 80%, followed by the two scenarios when 99.5% was reached and when it was fixed in the population. The signal increases quite fast within generation 358 (60%) and generation 407 (80%), and starts decreasing after the fixation. Setting the threshold score for LR_{T_3} at 15, the time, when the sweep can be 'reasonably well' detected, starts approximately when the frequency of selected allele is between 60-80% ($\sim$ generation 368) and last approximately to 400 generations after fixation ($\sim$ generation 1502). This gives a time interval of $\sim$ 1134 generations (see FIGURE 3.21), in which the sweep can be well detected.
We conclude that time point matters with regards to detecting selective sweep. When using T_3-based statistics the strongest signal seems to be when the selected site has reached a frequency of 80%. When applying the LR_{T_3}-test, the result was confirmed. Thus, according to this simulation results, our test seems to be applicable not only to recently completed sweeps.
Note that in previous sections, we generated data in such a way that the sweep was already fixed in the population.

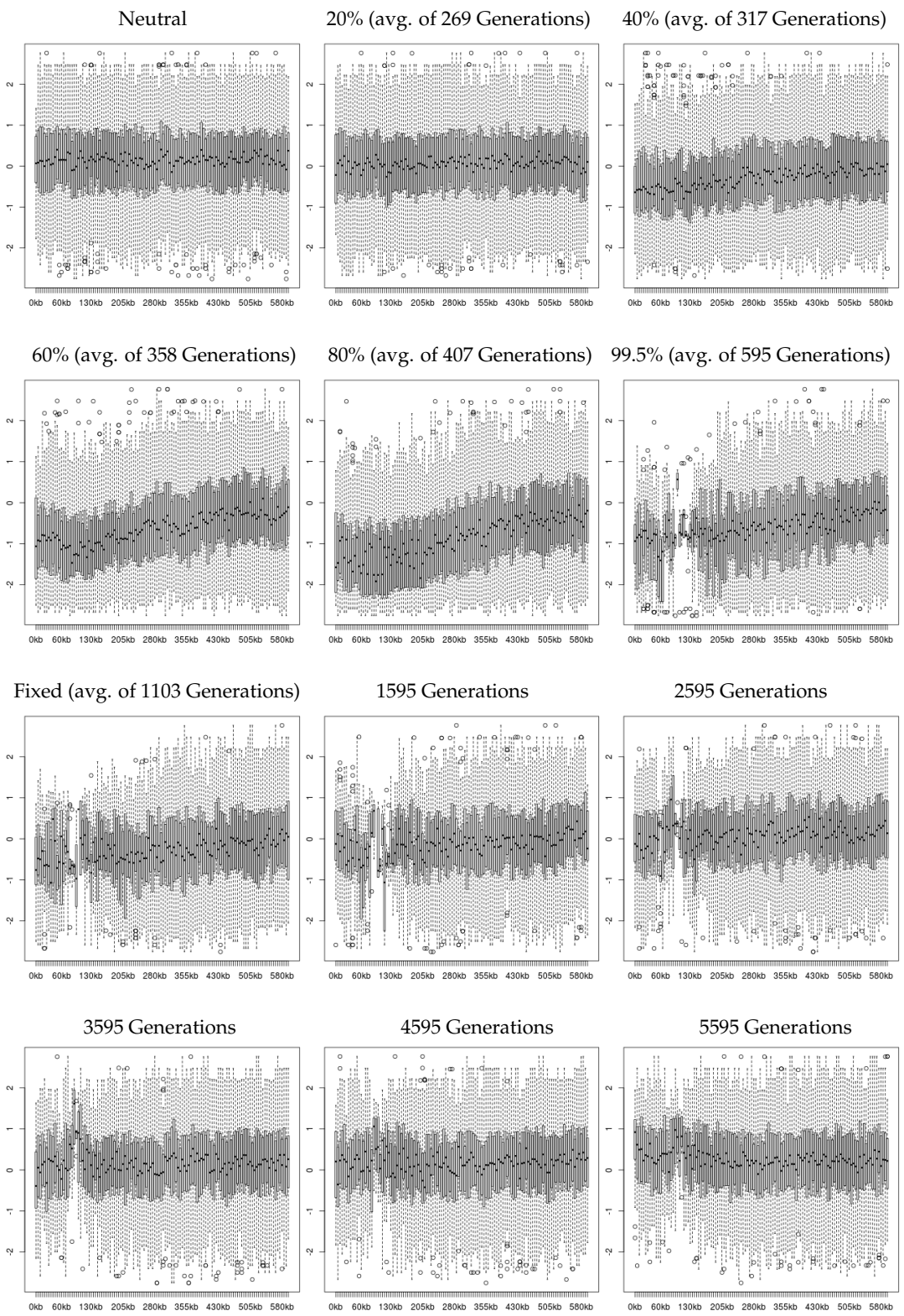


FIGURE 3.20: Distribution of the T_3-values along the 600kb DNA sequence for each twelve different stages explained in the text. The selected site is positioned at chromosomal position 100 kb. The strongest signal seems to be when the selected site has reached a frequency of 80%. The sparseness of data in the region around the selected site in generations 595 (on average) to 1,595 (on average) can be attributed to the strong reduction of polymorphism data around the selected site, and thus we lack of data.

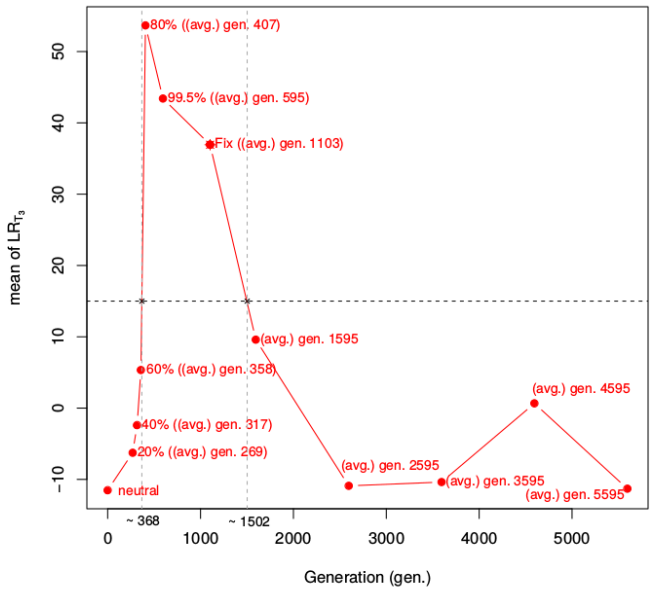


FIGURE 3.21: Mean LR_{T_3}-values of the 12 data sets, mentioned in the text. It can be seen that the most significant LR_{T_3}-score is found when the selected site has reached frequency of 80%, followed by when it has reached 99.5% and then when it was fixed in the population. Dashed black (horizontal) line indicates LR_{T_3}=15, which is the threshold score, see section 3.4.1. Dashed gray (vertical) lines indicates time interval when the sweep is detectable with $LR_{T_3} \geq 15$.

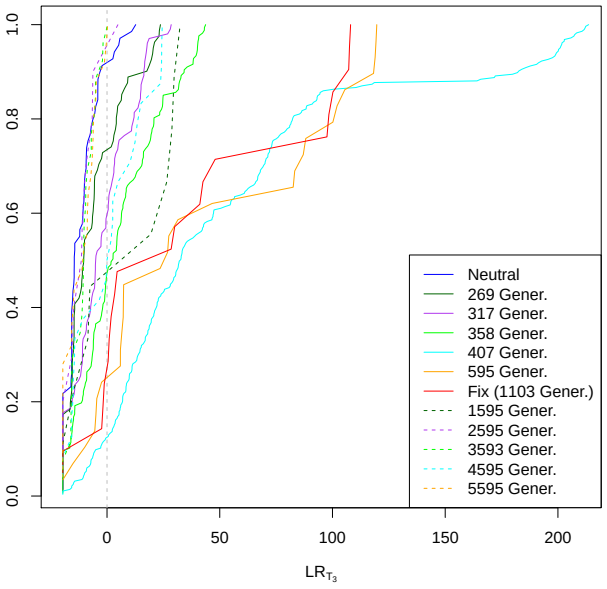


FIGURE 3.22: Cumulative distribution of LR_{T_3} for the 12 datasets, mentioned in the text. In the neutral scenario (dark blue), we can see that approximately 95% have $LR_{T_3} \leq 0$. Notable is the 'jump' towards high LR_{T_3}-values of the distributions of the datasets when the selected site has reached a frequency of 80%, 95.5% and is fixed.

Chapter 4

Application to experimental data

The field of DNA sequencing has been constantly evolving for decades, increasingly becoming both more efficient and more affordable. This has resulted in the generation of massive datasets for a wide spectrum of organisms, including human. The availability of these new data has clearly contributed to recent fundamental advances in population genetics: new models have been designed or existing models have been re-designed, simulation parameters can be chosen more plausibly, genome variation can be reconciled with population histories of admixture, migration or bottlenecks, and genome-wide scans are performed for finding signatures left by natural evolutionary forces leading to a deeper mechanistic understanding of how populations evolve.
In this chapter we show the application of the LR_{T_3}-Test to experimental data. To this end, we performed whole genome screens using human data (phase 3 dataset) from the 1,000 genomes project (Auton et al., 2015). We aimed at identifying new candidate regions which underwent selective sweeps. Furthermore, we expected to confirm many of the previously proposed candidates as well. We took a deeper look at biological functions for potential candidate genes from our 'top' regions to figure out what benefits selection on these genes may have brought along for their carriers.

4.1 The 1,000 Human Genomes Project

The first international effort to map and sequence all genes in the human genome was initiated in 1990 by the Human Genome Project (HGP). However, at that time sequencing the human genome was not only very time consuming, but also very expensive: It took approximately 13 years and $ 2.7 billion to complete the project see: *All About the Human Genome Project*. For instance in comparison to that, in February 2018 a team from the Rady Children's Institute for Genomic Medicine was awarded

with the GUINNESS WORLD RECORD^{TM} for sequencing a child's genome within 19.5 hours (see: *New GUINNESS WORLD RECORDS^{TM} Title Set for Fastest Genetic Diagnosis*). Although this is an extreme example (since the team got assistance from several sequencing companies) it shows what is possible today.
The focus of the 1,000 human genomes project was to create a detailed catalogue of human genetic variation and genotype data from populations all over the world (http://www.internationalgenome.org/). Therefore, more than 1,000 genomes of humans from different ethnic groups were collected. Advances in sequencing technologies allowed the project to be completed much faster than anticipated with less cost. The initial dataset of genomic sequences from 1,092 individuals belonging to 14 populations (also known as the phase 1 dataset) was produced in just four years, from 2008 to 2012 (Abecasis et al., 2012). The final phase of the project (phase 3) was announced in 2015 with a total of 2,504 sequenced human genomes from 26 populations across 5 continents (Auton et al., 2015) (table 4.1). The data include almost 90 million variants in the form of single nucleotide variants, insertions/deletions, and structural variants (source from http://www.internationalgenome.org/, last visited in August 2018).

FIGURE 4.1: Worldwide locations of the 26 population samples from 1,000 genomes project, final phase. Picture from http://www.internationalgenome.org/.
Yellow: African; Red: Admixed American; Green: East Asian; Blue: European; Purple: South Asian.

4.1.1 Examples of known recent human adaptations

The human genome consists of more than 3 billion nucleotide base pairs across 23 pairs of chromosomes (22 pairs of autosomes and one pair of sex chromosomes). There are an estimated 19,000-20,000 protein-coding genes in the human genome

Population	Population Description	Super-Population	Individuals
ACB	African Caribbean in Barbados	AFR	96
ASW	Americans of African Ancestry in Southwest USA	AFR	61
ESN	Esan in Nigeria	AFR	99
GWD	Gambian in Western Divisions in the Gambia	AFR	113
LWK	Luhya in Webuye, Kenya	AFR	99
MSL	Mende in Sierra Leone	AFR	85
YRI	Yoruba in Ibadan, Nigeria	AFR	108
CDX	Chinese Dai in Xishuangbanna, China	EAS	93
CHB	Han Chinese in Beijing, China	EAS	103
CHS	Southern Han Chinese, China	EAS	105
JPT	Japanese in Tokyo, Japan	EAS	104
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS	99
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR	99
FIN	Finnish in Finland	EUR	99
GBR	British in England and Scotland	EUR	91
IBS	Iberian Population in Spain	EUR	107
TSI	Toscani in Italia	EUR	107
BEB	Bengali from Bangladesh	SAS	86
GIH	Gujarati Indian from Houston, Texas	SAS	103
ITU	Indian Telugu from the UK	SAS	102
PJL	Punjabi from Lahore, Pakistan	SAS	96
STU	Sri Lankan Tamil from the UK	SAS	102
MXL	Mexican Ancestry from Los Angeles USA	AMR	64
PUR	Puerto Ricans from Puerto Rico	AMR	104
CLM	Colombians from Medellin, Colombia	AMR	94
PEL	Peruvians from Lima, Peru	AMR	85

TABLE 4.1: Population samples from the final phase (phase 3) of the 1,000 genomes project. There are 26 population samples in the whole dataset, but it can also be divided into five so-called 'superpopulations': African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), South Asian (SAS). Locations illustrated on a world map can be seen in FIGURE 4.1.

(Ezkurdia et al., 2014). The protein-coding sequences account for only a very small fraction of the genome, though. About 98% of the human genome consists of transposons and non-protein-coding sequences, such as non-coding RNA genes, regulatory DNA sequences, introns or sequences for which no function has been determined yet (Lander et al., 2001).

Despite enormous progress since the first human was sequenced, many things are still unknown with regards to the evolution of the human genome. Furthermore, there is much disagreement about the mode, strength and rate of selective sweeps in humans. Identifying loci which underwent recent selective sweeps is difficult because the traces are typically obscured by other evolutionary and demographic forces, e.g. genetic drift or population sub-structuring. It has been proposed that classical selective sweeps are rare in human populations (Hernandez et al., 2011). If

at all, then the majority are incomplete sweeps, soft sweeps (selection on standing variation), or selection on polygenic traits.
However, we do have evidence of differential adaptation of various traits, often associated with the human ancestors successfully establishing (sub)populations throughout the world. Modern humans are assumed to have spread from Africa around 50,000-100,000 years ago (Nielsen et al., 2017; Templeton, 2002), invading a variety of habitats and getting exposed to new environments. Therefore, they had to struggle with different climatic conditions or the availability of new food sources. The combination of selective pressure together with random drift left behind population-specific genetic patterns and phenotypic variations. Below are a few examples of well-documented adaptations in human populations.

Lactose tolerance

One of the standard examples of a gene to have experienced recent positive selection is *LCT*, the gene coding for lactase (lactase-phlorizin hydrolase). Lactase is the enzyme responsible for the ability to tolerate lactose; variants in the *LCT* gene influence whether the ability to digest milk persists into adulthood. Many studies have focused on this gene and the trait of lactase persistence is found in around 35 % of adults living in the world today (Itan et al., 2010). In Europeans, lactase persistence shows quite a strong signal of selection in scans of the entire genome (Bersaglieri et al., 2004). Outside Europe, lactase persistence is found in parts of Africa, the Middle East and India (Schlebusch et al., 2013; Enattah et al., 2008; Segurel and Bon, 2017). A particular allele of the LCT gene is associated with lactase persistence in both European and Indian populations (Gallego Romero et al., 2012). However, in Africa this phenotype appears to be polygenic instead (Gallego Romero et al., 2012; Tishkoff et al., 2007). Thus, lactase persistence evolved several times independently in human evolution in different areas of the world, making it an example of convergent evolution. It is generally thought to be related to the domestication of dairy cattle, as dairy milk is both a valuable source of nutrients during periods of erratic food supply and contains high levels of vitamin D, which is a further advantage in regions with low amount of sunlight, since the production of vitamin D is a UV-dependent process (Parra, 2007; Wacker and Holick, 2013). In any case, despite the numerous studies addressing the issue, much uncertainty remains about the origin of the lactase persistence-associated variants.

High altitude

Another quite well-known example of selection in humans is associated with the adaptation to high altitude, in particular the Tibetans and the Andeans (Beall, 2000).

Compared to the lowlands, mountaintops have less air pressure and lower oxygen content in the air. The physical and genetic changes observed in the Tibeteans and Andeans, in comparison to populations living in the lowlands, thus include mutations affecting the regulatory systems of oxygen respiration and it's transport via blood circulation. Even during pregnancy, blood flow and oxygen delivery to the uterus is increased to reduce the risk of having newborns with low birth weight (Julian, Wilson, and Moore, 2009). Studies suggest that amongst other genes, variants at the *EPAS1* (*Endothelial PAS Domain Protein 1*) locus are involved in the adaptation to high altitude (Peng et al., 2017).

Skin colour

Skin colour variation is one more noteworthy example of adaptation leading to wide-ranging human phenotypic diversity. Whereas dark skin is strongly associated with protection against UV light, lighter skin is subjected to positive selection for reasons such as maintaining vitamin D photosynthesis (Parra, 2007). Unfortunately, it is known that multiple different genes acting in concert are involved for skin (or also hair or eye) pigmentation (Parra, 2007), making it difficult and very complex to pinpoint the exact causative genes. For instance, according to a colour genes database, though focusing primarily on mice and last updated in October 2011, (http://www.espcr.org/micemut/), there are 378 candidate loci for colour genes described in mice and their human and zebrafish homologues, yet apparently only a few of them have been confirmed to have potentially function-altering polymorphisms in humans.
In general, the question is to what extent adaptation has driven evolution and affected patterns of genetic diversity.

4.2 Application of $\mathbf{LR_{T_3}}$-test to human data

As previously mentioned, we have applied the LR_{T_3}-Test to the human 1,000 genomes phase 3 data (Auton et al., 2015), which is publicly available and can be downloaded from the website ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The data is stored in the variant call format (VCF) (Danecek et al., 2011). Each of the 2,504 individuals carries an ID-number. A list of all the samples in the data set and their population, super population and gender can be found at the same public source. (Note: Only variants in form of SNP's were considered for our purposes). For the autosomal chromosomes 1-22, for all individuals the variant calls are diploid and genotypes are phased. Thus, here two haplotypes were constructed for each of the 2,504 individuals, so in total 5,008 haplotypes. However, for the male X chromosome variant calls were shown as haploid, but not in the pseudoautosomal

region (PAR), a part which is common between X and Y chromosome. Here, we modified the X-chromosome data in such a way that only one haplotype (the actual X-chromosome) accounted for a male individual, while for females two haplotype were constructed. This results in a total of 3,775 haplotypes for the X-chromosome. Y-chromosomes are not included in our analysis.

By using VCFtools (Danecek et al., 2011), a program package designed for working with VCF files, we could easily separate the individuals with regard to their population affiliation and thus store it in 26 separate files, each containing the respective individual.

We re-designed the output of the files in such a way that we could apply our T_3-calculations from section 3.2.1 (see example box 4.2). We screened all 26 populations separately by using a sliding window approach across the entire genome.

Box 4.2: Example of formatting a data file

The first table shows an extract of a 1,000 genomes data file. For demonstration purposes, only the following columns are shown (respectively from left to right): the chromosome number, chromosome position/coordinates on which the variant occurs, the reference SNP ID number, the reference (ancestral) allele, the alternative (derived) allele, followed by columns representing the genotype of the sample at this position (here the individuals are represented by X1, X2,... etc. '0' stands for the reference allele, '1' for the derived allele).

For instance, at chromosome 1 position 14464, individual X3 is heterozygous, carrying one copy of each of the reference and derived alleles, while individual X1 is homozygous for the derived allele and individual X2 homozygous for the reference allele.

CHR	POS	ID	REF	ALT	X1	X2	X3	X4	...
1	13110	rs540538026	G	A	0|0	1|0	0|0	0|0	...
1	13116	rs62635286	T	G	0|0	1|0	0|0	0|0	...
1	13118	rs200579949	A	G	0|0	1|0	0|0	0|0	...
1	14464	rs546169444	A	T	1|1	0|0	1|0	0|0	...
1	14599	rs531646671	T	A	0|0	0|1	1|0	0|0	...
1	14604	rs541940975	A	G	0|0	0|1	1|0	0|0	...
1	14930	rs75454623	A	G	1|0	0|1	0|1	1|0	...
1	15211	rs78601809	T	G	0|1	0|1	0|1	0|1	...
1	15820	rs2691315	G	T	1|0	0|1	0|1	0|0	...
1	16949	rs199745162	A	C	0|0	0|0	0|1	0|0	...
1	18643	rs564023708	G	A	0|0	0|0	1|0	0|0	...

We can construct two haplotypes for each individual:

CHR	POS	...	X1.A	X1.B	X2.A	X2.B	X3.A	X3.B	X4.A	X4.B	...
1	13110	...	0	0	1	0	0	0	0	0	...
1	13116	...	0	0	1	0	0	0	0	0	...
1	13118	...	0	0	1	0	0	0	0	0	...
1	14464	...	1	1	0	0	1	0	0	0	...
1	14599	...	0	0	0	1	1	0	0	0	...
1	14604	...	0	0	0	1	1	0	0	0	...
1	14930	...	1	0	0	1	0	1	1	0	...
1	15211	...	0	1	0	1	0	1	0	1	...
1	15820	...	1	0	0	1	0	1	0	0	...
1	16949	...	0	0	0	0	0	1	0	0	...
1	18643	...	0	0	0	0	1	0	0	0	...

Rearrangement leads to an output file similar to the `msms`-output file shown in line 8-12 in FIGURE 3.4.

```
X1.A:   00010010100
X1.B:   00010001000
X2.A:   11100000000
X2.B:   00001111100
X3.A:   00011100001
X3.B:   00000011110
X4.A:   00000010000
X4.B:   00000001000
```

Thus, for position chr1:13110-18643 (= window size of 5533 bp), we have eight haplotypes consisting of 11 SNPs, from which we can now determine Ω_1, Ω_2 and Ω_3 to calculate T_3, in the same manner as in section 3.2.1.

As mentioned before, we took a window of size 5,000 bp and step size 2,500 bp, with the additional condition that the fragment needed to contain at least 10 SNPs. If the latter was not the case, the window size was increased by adding 1,000 bp until the second condition was fulfilled, but with a maximum total window length of 10 kb. Monomorphic sites were excluded, since that would have led to disparities towards balanced trees. For the determination of window size and number of SNPs, re-consider section 3.2.1. The T_3-values were reported for each window as it slides along the chromosome with a step size of 2,500 bp. The result was converted to BED (Browser Extensible Data) format for each 26 population separately. BED files are tab-delimited files with one line for each genomic region. The lines of a BED file have three required fields and additional optional fields with tabs as delimiters. The

first three (required) BED fields are: chromosome, starting position of the region and ending position of the region in the chromosome. In our case the additional optional fourth field represents the T_3-value. Afterwards, we performed the LR_{T_3}-test as described in section 3.4.1.

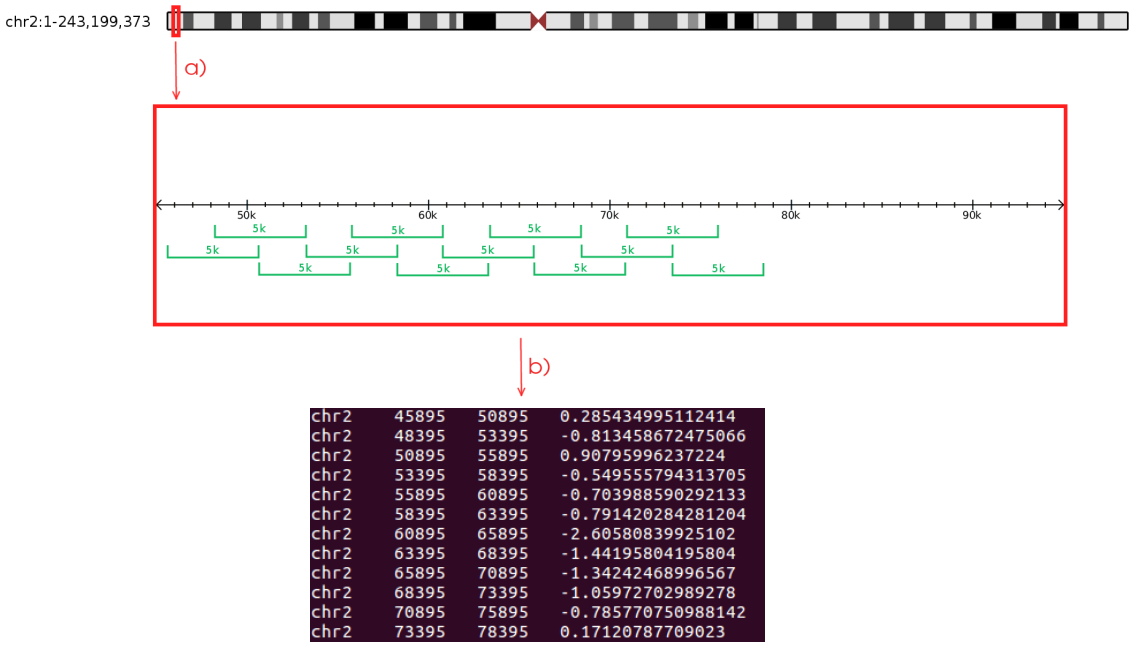


FIGURE 4.2: Visualisation of sliding window approach. Starting from the beginning of the chromosome, a T_3-value is reported for each window (shown as green line) as it slides along the chromosome with a step size of half the window size. a) Zoom-in of a small part of chromosome 2. b) Example of storing the T_3-values in BED format, where the first column contains the chromosome name, the second column the starting position of the window, the third column the ending position and the fourth column the respective T_3-value.

Therefore, we determined the empirical 1% T_3-threshold separately for each population and each chromosome. We identified all regions with a T_3-value under the respective threshold. These identified regions (= 'focal' regions) were then subjected to the LR_{T_3}-test: By looking at 100 adjacent windows to the left and to the right side of the focal region, we recorded for each the respective LR_{T_3}-score. By reason of the previously explained chosen window and step size, the 100 consecutive windows correspond to approximately 250 kb (Note: Since the window length is extended if the minimum SNP number is not fulfilled, this size can vary). The complete result of this screen was also stored in BED format which then can be visualized on the UCSC Browser https://genome.ucsc.edu/, see FIGURE 4.3.
In the end, we considered those regions as candidate regions, if the LR_{T_3}-value was ≥ 15.

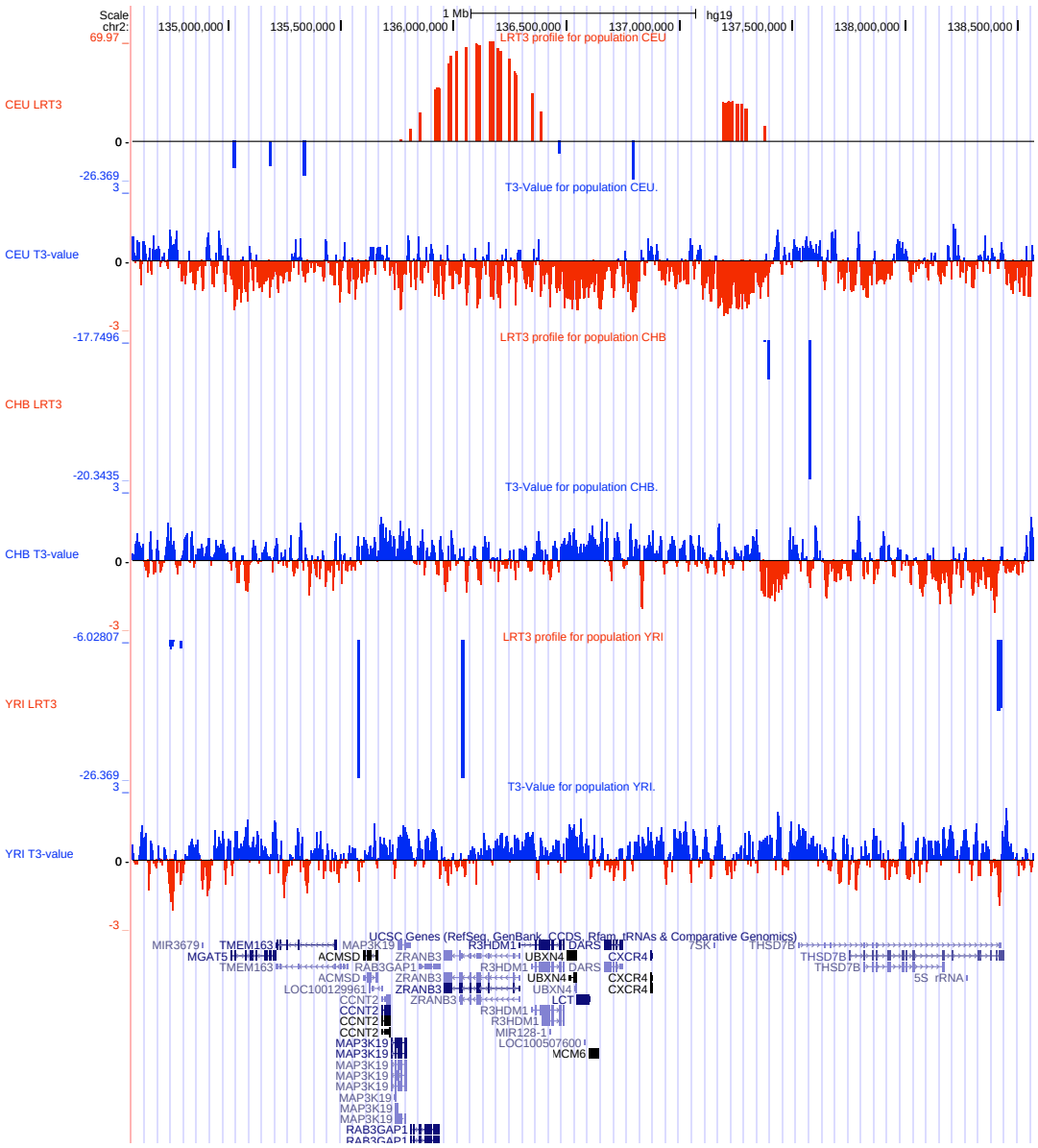
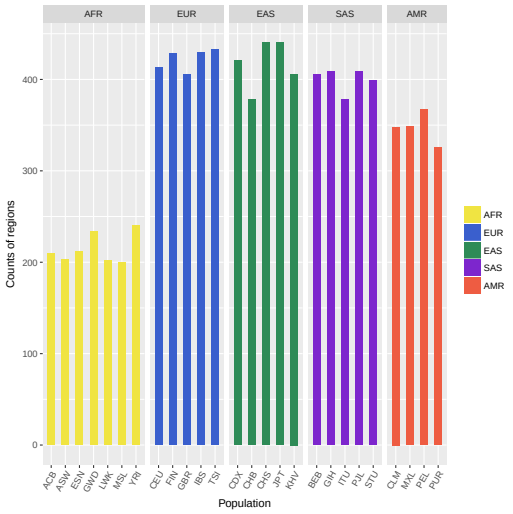


FIGURE 4.3: LR_{T_3}-profile and T_3-profile along the chromosome for region chr2:134,571,975-138,568,190 (Visualisation via UCSC Browser https://genome.ucsc.edu/). Shown are the LR_{T_3}- and T_3-profiles for the three populations: CEU, CHB and YRI, in order from top to bottom. Positive LR_{T_3}-score is shown in red, negative LR_{T_3}-score is shown in blue. Negative T_3-values are shown in red, positive T_3-values in blue. For this area, the populations CHB and YRI hardly contains LR_{T_3}-scores at all, meaning, hardly found significant T_3-windows, and if, then LR_{T_3} is negative. In contrast to CEU, where two location spot seem to be significant as it can be seen. On the bottom of this picture, genes associated to the respective regions are shown.

4.3 Analysis of candidate regions

In the section before, we have screened all 26 populations from the phase 3 release of the human 1,000 genomes data (Auton et al., 2015) with the LR_{T_3}-test. Regions with LR_{T_3}-score ≥ 15 were considered to be a candidate region for having undergone selection. As expected, many of these identified candidate regions were overlapping



POP	ACB	ASW	ESN	GWD	LWK	MSL	YRI
#regions	210	203	212	234	202	200	240

POP	CEU	FIN	GBR	IBS	TSI
#regions	413	428	405	429	433

POP	CDX	CHB	CHS	JPT	KHV
#regions	421	378	440	440	406

POP	BEB	GIH	ITU	PJL	STU
#regions	405	409	378	409	399

POP	CLM	MXL	PEL	PUR
#regions	348	349	367	326

FIGURE 4.4: Number of chromosomal regions, that can be considered candidates for recent selective sweeps, per population. Regions span between 55 kb and 785 kb. More information about regions per chromosome is given in the APPENDIX table B.1.

(a consequence of the sliding window approach). In all such cases, we merged the overlapping regions into a single region. Moreover, motivated by the fact, that the highly unbalanced tree topology is not observed directly but in the vicinity of the selected site, we additionally extended these regions by 25 kb on both sides. Hence, the resulting final candidate regions span lengths between 55 kb and 785 kb. The total numbers of regions per population are illustrated in FIGURE 4.4 (also see APPENDIX B.1).

In general, we found less amount of candidate regions in African populations compared to the rest: We found approximately two times less candidate regions in the African superpopulation compared to the rest (on average 214 in Africans vs $\sim$ 400 on average in the others; see FIGURE 4.4, or APPENDIX table B.1 for details on the numbers). This is consistent with other studies that have found more candidate regions for having undergone selection in non-African populations compared to the African populations (Kayser, Brauer, and Stoneking, 2003; Williamson et al., 2007; Campbell and Tishkoff, 2008). A straightforward explanation might be that while humans dispersed out of Africa 50,000-100,000 years ago (Nielsen et al., 2017; Templeton, 2002), they were forced to adapt to the new environments they encountered (Kayser, Brauer, and Stoneking, 2003; Williamson et al., 2007). However, another possible reason, for swept loci being more identified in non-Africans might be that neutrality test statistics suffer from the confounding effects of demographic events (see chapter 2). During the Out-of-Africa migration, humans were accompanied by bottleneck event(s) (Amos and Hoffman, 2010), a hypothesis mostly studied in the framework to explain why the African population shows a higher level of diversity compared to non-African populations (Campbell and Tishkoff, 2008; Rosenberg and

Kang, 2015).

4.3.1 Identifying candidate genes

To extract genes, we used the `biomaRt` package in `R` (Smedley et al., 2015). `biomaRt` offers an easy way to extract a list of different attributes, which defines the values we are interested in. In our case, we retrieved the gene symbols, chromosomal coordinates, the respective gene biotype, (giving us the information of whether the given transcript is protein-coding or non-coding), and the respective Gene Ontology (GO) term. We use the coordinates for human genome build hg19 for our data, to which phase 3 of the 1,000 genomes project is mapped.

In total we found 9,725 genes that can be considered candidate loci for selection in at least one of the 26 populations. Out of these 9,725, on average 639 are found in African populations, 1,368 in European populations, 1,217 in East Asian populations, 1,205 in South Asian populations and 1,081 in American Admixed populations (see APPENDIX table B.1 for more detail or FIGURE 4.5, A). Furthermore, out of the 9,725 candidate genes 3,956 genes were associated with the biotype „protein-coding“ and the rest with other biotypes. If focusing on protein-coding genes, we found an average of 278 in African populations, 575 in European populations, 497 in East Asian populations, 513 in South Asian populations and 455 in American Admixed populations (see APPENDIX table B.2 for more detail or FIGURE 4.5, B).

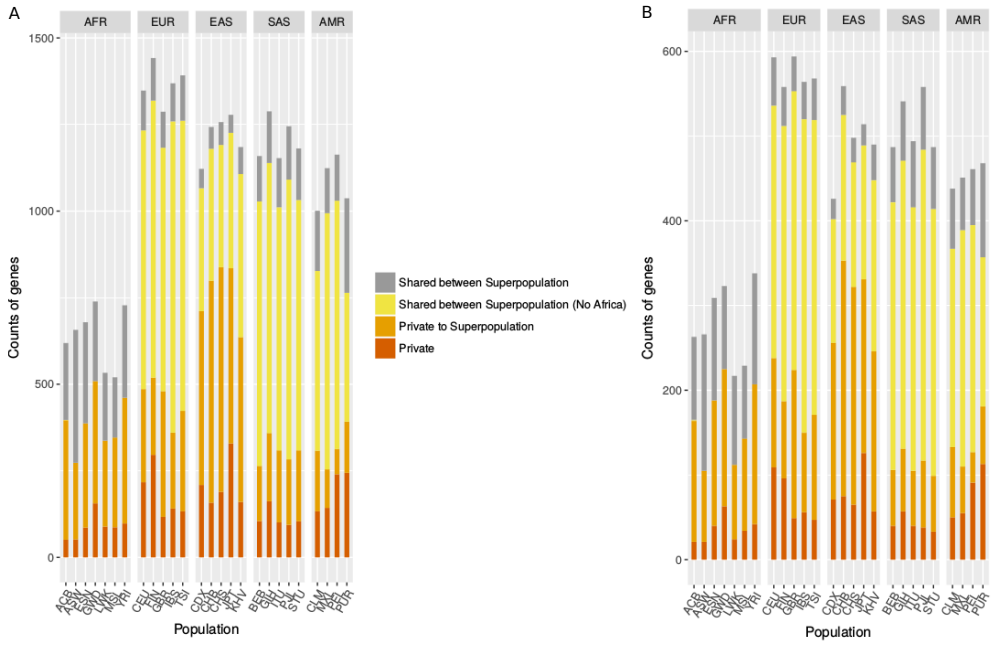


FIGURE 4.5: Shared and private candidate genes. The different colouring indicates the different categories given in the legend. Private-selective sweep candidate in one (super)population. Shared – selective sweep candidate in multiple (super)populations. A: All genes. B: Only protein-coding genes.

Superpopulation	Average number of candidate regions	Average number of all genes	Average number of protein-coding genes
AFR	214	639	278
EUR	422	1,368	575
EAS	417	1,217	497
SAS	400	1,205	513
AMR	348	1,081	455

TABLE 4.2: Overview of average number of candidate regions, average number of all genes and average number of protein-coding genes per superpopulation.

Furthermore, we recorded if the detected genes were found in one population only (private), if they were shared in (at least two) populations belonging to the same superpopulation (private to superpopulation), if they were shared between (at least two) populations not belonging to the same superpopulation (shared between superpopulation), whereby here we additionally made the distinction between superpopulation excluding and including Africa (see FIGURE 4.5). We made the latter distinction since we were interested if the hypothesis that one of the leading forces driving positive selection in non-Africans as the Out-of-Africa migration was reflected in differential patterns and targets concerning the underlying biological function of the selected genes. For instance, one may expect that non-African populations share more positively selected genes involved in metabolic pathways as a response to diverse food source or genetic adaptation as result to diverse climate changes. These genetic adaptations should not be visible in African populations. However, in the African populations one may expect to see local adaptations being prevalent, for instance genetic adaptations providing resistance to the exposure to different pathogens.

Comparison to previous studies

As already mentioned, many previous studies have focused on the detection of genomic regions which might have been targeted by positive selection. For this purpose, several different methods have been established (Vitti, Grossman, and Sabeti, 2013). With the rapid development of genome scale population level DNA genotyping and sequencing in humans, many studies published gene candidates in the human genome that were possibly targeted by selection.
In (Li et al., 2014a), the authors made the effort to collect all candidate sweep regions identified until then, published them and establish a database, called **dbPSHP** (= *database of recent positive selection across human populations*). Intrinsically, the database consists of over 15,000 loci from either publications attempting to study positively selected genomic locus and gene related to specific functions, traits or diseases, or

publications to detect the genome-wide selective signals with different statistical methods. Since the regions recorded in the database vary widely in terms of size, we focused on the candidate genes. Taken together and removing multiple recorded genes, approximately 8,050 unique genes are stored.
Comparing our list of candidate genes with the list in dbPSHP, we confirmed about 1,947 genes, from which 1,853 are protein-coding genes from our list. Since the last update of dbPSHP was, according to the website http://jjwanglab.org/dbpshp (status from July 2018) in May 2014, we took another list of candidate genes into consideration: a list set up by Schrider and Kern, (2017). The authors used a machine learning approach developed by themselves in a previous paper, called *S/HIC* (=*Soft/Hard Inference through Classification*). Their approach should be *'remarkably powerful and robust to non-equilibrium demography'* as quoted from Schrider and Kern, (2017), and allows not only the detection of hard sweeps and soft sweeps, but also the detection of regions closely linked to hard and soft sweeps. It uses 11 population genetic summary statistics (including Tajima's D, Fay and Wu's H and also a number of distinct haplotype based test). If we compared our candidate genes with the genes found in the SHIC paper (where in total 5,939 candidate genes were found), we confirmed in total 1,718 genes, from which 840 were coding genes and 878 were non-coding genes. (From these 1,718 genes, 1,253 are not found in dbPSHP, 383 protein-coding and 870 non-coding.)
However, in the SHIC paper six populations (CEU, JPT, GWD, YRI, LWK, PEL) were analysed, while here we analysed all available 26 populations. If we only took the six populations, 4,551 genes are left in our list. We therefore conclude that with the threshold used here, our test is more stringent than the one used in the SHIC paper. From the aforementioned 4,551 genes we confirmed 912 genes, from which 438 are protein-coding and 474 are non-coding. If the found genes are additional candidates for the same population, then we could confirm 668 genes. Here, 318 are protein-coding and 350 are non-coding.

4.3.2 Analysis of the top candidates

As we have seen in section 4.3, some identified regions ended up to be very large with a region span between 55 kb and 785 kb. Therefore, one region can contain multiple candidate genes. To make a clear decision about which gene is the positively selected one is rather difficult. It has to be noted, that the constituent windows composing the resulting candidate region mostly possess similar high LR_{T_3} scores, making it not easier to determine which the actual 'chosen' region/window is.
In table 4.3, we listed all protein-coding genes associated to regions with very high

LR_{T_3}-score (> 200). We only show the maximum LR_{T_3}-score related to the given region. Although here we focus on protein-coding genes, our method can be also applied to non-coding genes. Generally, the functional role of non-coding genes should not be underestimated. Their functions range from regulation of gene expression at the transcriptional and post-transcriptional level to exhibiting histone modification patterns characteristic of specific functional elements. Recent studies have shown the important role of non-coding RNA in cancer, e.g. (Huang et al., 2013).

max LR_{T_3}	POP	Chr	Position	Size in bp	Coding
316.972	ITU	12	44,342,384-44,904,884	562,500	NELL2, TMEM117
276.577	GBR	14	67,183,154-67,930,654	747,500	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2, **TMEM229B**
260.28	FIN	14	67,220,427-67,905,427	685,000	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2
259.153	TSI	14	67,213,154-67,928,154	715,000	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2 , **TMEM229B**
247.929	CEU	14	67,220,445-67,897,945	677,500	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2
241.559	CHB	X	100,985,920-101,448,420	462,500	**NXF5**, **ZMAT1**, **TCEAL2**, **TCEAL6**, **BEX5**
239.56	CHB	2	108,905,521-109,650,521	745,000	EDAR, RANBP2, LIMS1, CCDC138, GCC2, SULT1C2, **SULT1C4**
238.617	BEB	12	44,307,384-44,927,384	620,000	NELL2, TMEM117
237.469	CHB	15	63,764,703-64,337,203	572,500	HERC1, DAPK2, FBXL22, USP3
233.935	IBS	8	42,643,536-43,378,536	735,000	**HGSNAT**, **POMK**, FNTA, HOOK3, CHRNA6, THAP1, **RNF170**, **RP11-598P20.5**
230.4045	GIH	5	43,588,039-44,073,039	485,000	NNT
226.406	CHB	12	44,354,884-44,699,884	345,000	TMEM117
226.39	CDX	4	41,515,167-42,215,167	700,000	LIMCH1, PHOX2B, TMEM33, DCAF4L1, SLC30A9, BEND4
222.695	CDX	2	108,913,021-109,383,021	470,000	RANBP2, LIMS1, GCC2, SULT1C2, **SULT1C4**
213.416	ACB	20	20,387,585-20,787,585	400,000	**RALGAPA2**
211.634	CHB	3	154,167,942-154,822,942	655,000	MME
205.92	MXL	1	100,410,610-100,790,610	380,000	**SLC35A3**, **HIAT1**, **SASS6**, **TRMT13**, **LRRC39**, DBT, RTCA
205.027	CHB	8	10,725,271-11,112,771	387,500	XKR6, AF131215.5
204.738	JPT	10	55,859,211-56,226,711	367,500	PCDH15
203.813	GIH	4	106,462,667-106,815,167	352,500	ARHGEF38, INTS12, GSTCD
203.628	MXL	10	74,926,660-75,406,660	480,000	SYNPO2L, MYOZ1, USP54, PPP3CB, **MRPS16**, ANXA7, TTC18, **MRPS16**, DNAJC9, FAM149B1, ECD
203.317	FIN	1	51,465,610-52,033,110	567,500	EPS15, TTC39A, RNF11, C1orf185

TABLE 4.3: Protein-coding genes associated to regions with very high LR_{T_3}-score of (> 200). Only the maximum LR_{T_3}-score related to the respective region is shown. The indicated chromosomal position represents the extended coordinates of +/-25kb. Gene names in bold are newly identified candidate loci.

Most of these genes are previously known sweep candidates. Genes we could not re-find either in the 'dbPSHP-list' or the list from Schrider and Kern (2017) are indicated in bold letters. These are potentially new candidate genes. A list with gene names appearing in TABLE 4.3 is provided in the LIST OF ABBREVIATIONS. Although most of these genes have been previously suggested to be under selection (for a reference list where each of these genes have been mentioned before see APPENDIX B.2), the biological function and thus the reason why they should have been selected for is poorly understood. For instance, the region with the highest LR_{T_3}-score is found

in the South Asia population ITU. It contains two protein coding genes: *NELL2* and *TMEM117*. *NELL2* is also a candidate (although smaller LR_{T_3}-score) for: BEB, GIH, PJL and STU (hence all five South Asian populations), and for the European population FIN and for the admixed American populations CLM, MXL and PEL. *TMEM117* is a candidate gene for all five South Asia populations, for all five East Asia populations, for four of five of the European populations (CEU, FIN, GBR and TSI), and for the admixed American populations CLM, MXL and PEL. Therefore, these two genes are candidate genes for almost all non-African populations. As for their function, *NELL2* is a neuronal growth factor; it has been shown to be involved in sexual behaviour and the onset of puberty, at least in rats (Ryu et al., 2011). Interestingly, in (Ramnitz and Lodish, 2013) the authors state that African American girls enter puberty earlier than Caucasian and Hispanic girls. The gene *TMEM117* on the other hand is involved in the maintenance of the mitochondrial membrane (Tamaki et al., 2017).
The second highest LR_{T_3}-score is found in the European population GBR; this region is also a candidate in all other European populations. One possible gene driving selection in this region is *GPHN*, mutations on which affect the nervous system and/or behaviour. Diseases that *GPHN* disruptions might be involved in include hyperekplexia (Rees et al., 2003), Alzheimer's disease, schizophrenia and autism (Lionel et al., 2013; Hales et al., 2013). Another possible candidate from this region is *MPP5*, disruption of which has be associated with cancer and diseases leading to blindness (Li et al., 2014b; Luo et al., 2011), suggesting a possible connection with eyesight. A newly suggested candidate gene for this region might be *TMEM229B*. It is mostly associated in studies with cancer (Stoletov et al., 2018).The strongest candidate region appearing in the East Asian population CHB lies on the X chromosome (see TABLE 4.3) and has not been previously identified as a selection candidate by other works. It contains the protein-coding genes *NXF5*, *ZMAT1*, *TCEAL2*, *TCEAL6* and *BEX5*. This region is also a candidate region for two other two East Asian populations JPT (maximum $LR_{T_3} = 179.14$) and CHS (maximum $LR_{T_3} = 80.48$). Within this region, the gene *NXF5* in particular has been previously associated with mental retardation, kidney failure and female infertility (Jun et al., 2001; Esposito et al., 2013; Fortuno and Labarta, 2014).
A further list of all 'Top Ten per population' candidate region for each 26 population is provided in the APPENDIX B.3.

A few other candidate genes present in our 'Top Ten per population' list, see APPENDIX B.3, piqued our interest due to their functional importance. The *EDAR* gene belongs to a region that is a sweep candidate for all five of the East Asian populations, and in none of the other populations. For four out of the five East Asian

populations it is even a very strong candidate. An LR_{T_3}-profile along this chromosome region is demonstrated in FIGURE 4.6. Similar results for EDAR in East Asian populations have also been reported by other authors (Sabeti et al., 2007; Bryk et al., 2008; Fujimoto et al., 2008; Pickrell et al., 2009).

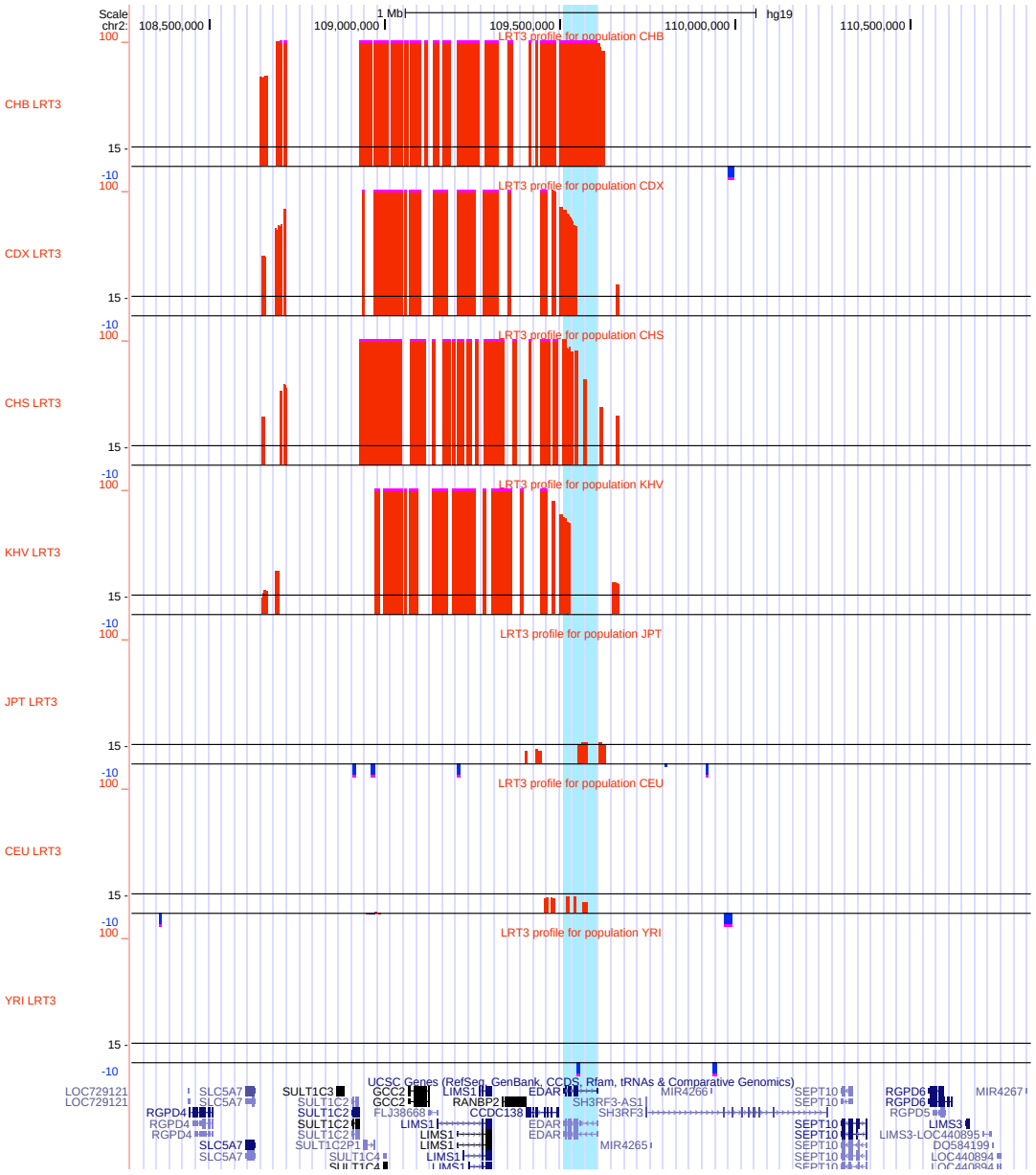


FIGURE 4.6: Strong signal for the *EDAR* gene region for East Asian populations. Only for JPT this is not a strong candidate, conversely a rather weak candidate (maximum $LR_{T_3} = 16.87$). For comparison reason, LR_{T_3}-profile for population CEU and YRI is given at the bottom. Shown is the chromosomal position chr2:108,277,201-110,839,554. *EDAR* is highlighted. Illustration via `https://genome.ucsc.edu/`. Note: Only LR_{T_3}-range from -10 to 100 is shown.

The *EDAR* gene is known to be involved in the development of hair, teeth and sweat glands (Botchkarev and Fessing, 2005; Kamberov et al., 2013). *EDAR* is associated with hair thickness, and the observation that East Asians tend to have thicker hair

than Europeans and Africans, leads to the question of why thicker hair may have been advantageous. Hypotheses range from a simple sexual selection/mating advantage to being a by-product of selection on other functions of the gene (Bryk et al., 2008; Kamberov et al., 2013).
Another noteworthy candidate from our 'Top Ten candidate per population', see APPENDIX B.3, is the gene *CASK*. The region where this belongs to is in the 'Top Ten candidate per population'-list for three African populations: ACB (LR_{T_3}=102.35), YRI (LR_{T_3}=127.87) and LWK (LR_{T_3}=90.2757), and moreover is also significant for further three African populations GWD (LR_{T_3}=22.9285), ASW (LR_{T_3}=38.3957), ESN (LR_{T_3}= 74.7938)¶. An LR_{T_3}-profile along this chromosome region is given in FIGURE 4.8. Although selection on this gene has not received much attention in humans thus far (although it appears in the list in (Frazer et al., 2007)), *CASK* has been suggested to be positively selected in racing pigeons and is implicated in the formation of neuromuscular junctions (Gazda et al., 2018). Hence, the authors suggest this gene to be involved in physical factors contributing to athletic performance.
Another sweep candidate from our list, although from neither of the top lists but rather medium-high LR_{T_3} score, is gene *HERC2*. This gene is suggested having undergone selective sweep for the European population CEU ($LR_{T_3} = 76.53$), GBR ($LR_{T_3} = 50.79$) and FIN ($LR_{T_3} = 58.68$) (for illustration of LR_{T_3}-profile along chromosome region see FIGURE 4.7). It is known that the eye colour is a result of multiple genes interacting together, nevertheless *HERC2* is suggested to belong to one of the key genes being involved for the brown/blue eye colouring. Actually, not the *HERC2* gene itself, but the nearby *OCA2* seems to control the eye pigmentation. Studies have found a region in *HERC2* regulating the activity of the *OCA2* gene which in turn is involved in the production of the pigment melanin. A variant of *HERC2* leads to inhibiting *OCA2* expression, causing a reduction in the production of melanin resulting in blue eyes (Eiberg et al., 2008). However, the advantage of having blue eyes is poorly understood, although there have been speculations that people with blue eyes might be able to deal better with the lack of light (Sturm and Duffy, 2012). Or it might simply be a case of sexual selection.

The last example for this section refers to genes, suggested as 'novel' candidates for African populations in a very recent study (Mughal and DeGiorgio, 2018): *COL8A1*, *CMSS1* and *FILIP1L*. In our analysis, the 'novel' candidates could be confirmed: We recover the candidates in (almost) all seven African populations: *CMSS1*, *FILIP1L* for all seven, *COL8A1* for six without ASW. For an illustration of the LR_{T_3}-profile along chromosome region see APPENDIX, FIGURE B.2. *COL8A1* may be involved in the development of muscle and has been suggested to be positively selected in other

¶Note: The given LR_{T_3}-score refers to the maximum value in the region

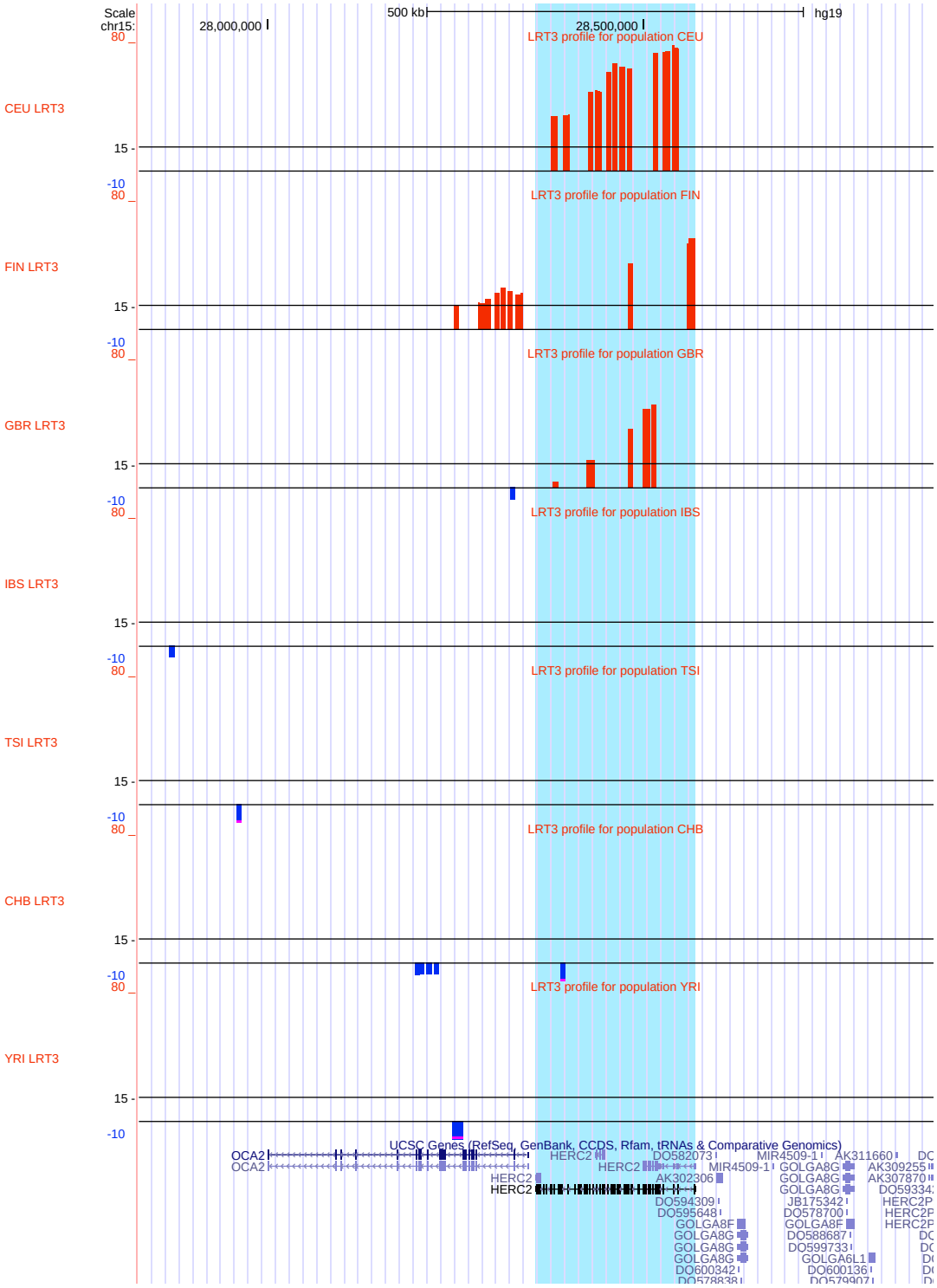


FIGURE 4.7: LR_{T_3}-profile for region around the *HERC2* being significant for CEU, GBR and FIN. For comparison reason, the other two (Mediterranean) European population TSI and IBS is shown (note: no signal can be observed at all), one Asian population CHB and one African population YRI. Shown is the chromosomal position chr15:27,828,393-28,901,088. *HERC2* is highlighted. Illustration via https://genome.ucsc.edu/. Note: Only LR_{T_3}-range from -10 to 80 is shown.

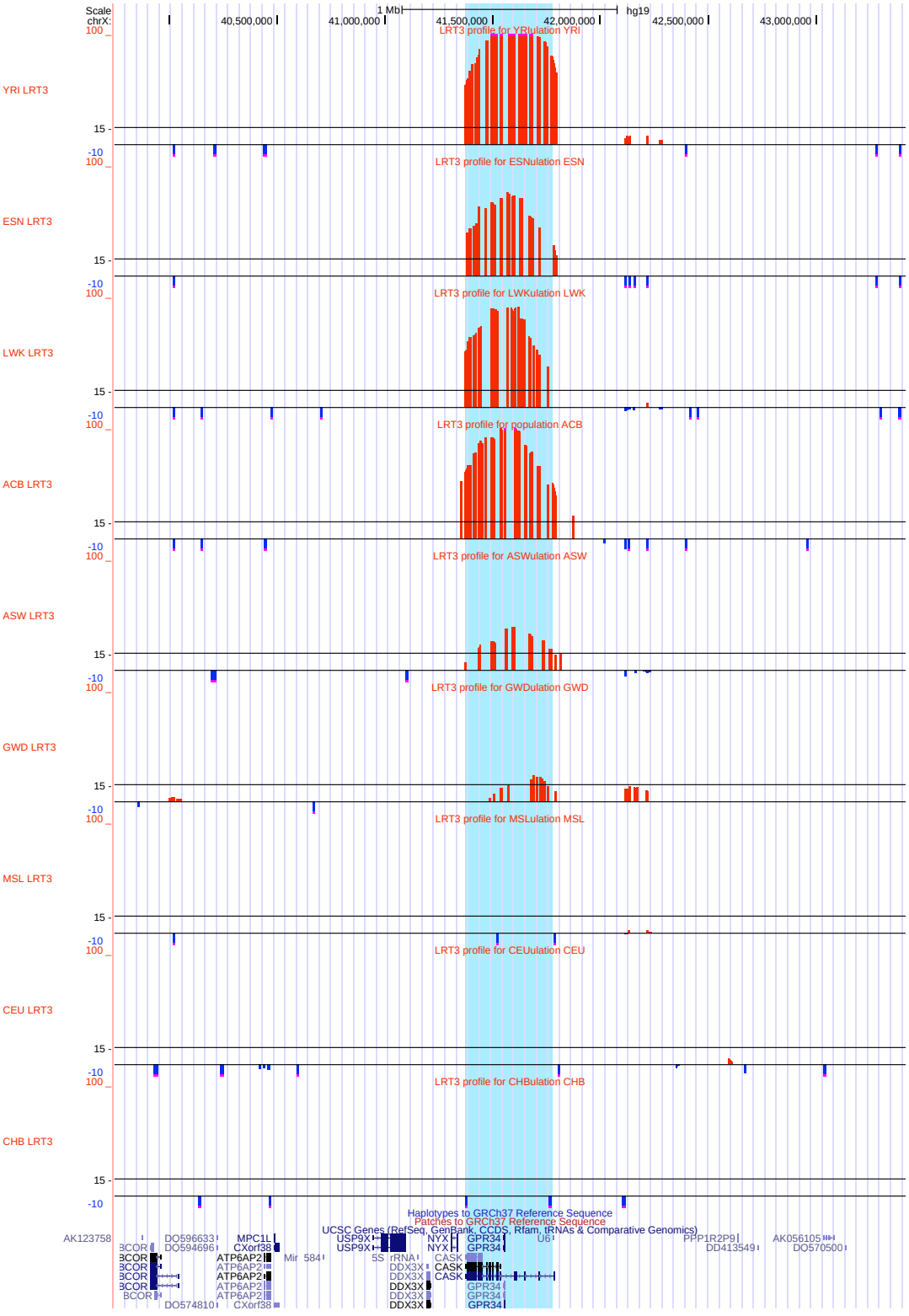


FIGURE 4.8: LR_{T_3}-profile for the region surrounding gene *CASK*, which is a (strong) candidate for almost all African populations. The LR_{T_3}-profile is shown for all seven African populations, for comparison reason, LR_{T_3}-profile for one European population CEU and one East Asia population CHB are given. Shown is the chromosomal position chrX:39,741,793-43,414,683. *CASK* is highlighted. Illustration via https://genome.ucsc.edu/. Note: Only LR_{T_3}-range from -10 to 100 is shown.

species (Utsunomiya et al., 2013; Mughal and DeGiorgio, 2018).

Finally, we want to remark that although our test could confirm many previously known genes, some 'famous' candidates for selective genes do not appear in our candidate list, for instance, *LCT* and the $\sim$ 39kb-distant *MCM6*, which contains regulatory elements for *LCT*, see e.g. (Hubacek et al., 2017). Both are associated with lactose tolerance and enables the carrier of beneficial variants the digestion of milk (see section 4.1.1). However, we do find a rather strong signal from the European populations CEU (see 4.3) and GBR for a zinc finger gene lying 257 kb away from *LCT* and *MCM6*. This gene - *ZRANB3* - was already mentioned in other studies, often in connection to large candidate regions also containing *LCT*. In (Ferrer-Admetlla et al., 2014) it even showed the strongest signal for their haplotype-based statistic nS_L (however for a population from Kenya). We suggest that there are unknown interactions between *ZRANB3* and closely located genes. This hypothesis will be investigated in more detail in chapter 5.

4.3.3 Gene Ontology Enrichment Analysis of top regions

In this section we were investigating whether some gene sets can be associated with functional genetic differences among different continents (or to be more precise: among different superpopulations). Therefore, we performed enrichment analysis on different gene sets by using Gene Ontology (GO) terms (Ashburner et al., 2000; Gene Ontology Consortium, 2017). The GO is a bioinformatics project developed by the Gene Ontology Consortium aiming at providing *a set of structured, controlled vocabularies for community use in annotating genes, gene products and sequences*, as cited from the Gene Ontology Consortium, (2008). GO defines classes which can then be used to describe gene functions, and how these functions are related to each other. Furthermore, GO enrichment analysis allows the assignment of biological meaning to some groups of genes instead of looking at each individually. Generally, GO depicts three functional domains:

- Biological process - represents a biological objective or biological phenomena like limb formation, DNA replication etc.
- Molecular function - describes the activities of a gene product at the molecular level.
- Cellular component - describes the location of the gene relative to cellular compartments and structures.

To find whether there are some functional sets of genes which can be associated with genetic differences among populations located in different continents, we conduct GO enrichment analysis on different lists of our candidate genes.
The principal idea of the analysis is as follows: Given a background gene set and a set of interesting genes, after identifying which GO terms are most commonly associated within the set of interesting genes, ask if this association is significantly different from what would be expected based on the proportions of genes out of the total having each attribute (background gene set) and compute a p-value for the observed association (enrichment).
As a standard approach for identifying enriched GO terms the hypergeometric distribution is used. For the analysis we used the web-based tool *GOrilla* (= ***G**ene **O**ntology en**RI**chment ana**L**ysis and visua**L**iz**A**tion tool*) (Eden et al., 2009).
For the background set, we downloaded a full gene list of human genome on http://grch37.ensembl.org/downloads.html, build hg19/GRCh37. The target sets were produced as follows:
First, we identified the top ten regions for all 26 populations separately and filtered the respective genes belonging to each region (APPENDIX B.3). Then, we built five target sets in grouping together genes according to their superpopulation affiliation.
In the following, we present the top three most significant enriched GO terms for each set, including the description (column 2), the p-value (column 3), the 'FDR q-value'* (column 4) and the relevant annotated genes (column 5).
The most significant results can be found in East Asian populations for a family of histones, which are proteins playing a major role in chromatin packaging (TABLE 4.5). Since DNA is wrapped around histones, they are also important regarding the regulation of gene expression.
However, overall it can be said that the p-values are not remarkably significant (a fortiori the q-values, see TABLE 4.4 and 4.5). The number of genes attributed to the enrichment is quite low and it is thus difficult to make reliable statements or conclusions.
Finally, we could not see any significant differences in biological functions between African and Non-African populations (see also APPENDIX B.6). In this regard, our finding confirms other recent studies (Campbell and Tishkoff, 2008).

Despite what was mentioned above, we did make an intriguing observation concerning the GO Term 'social behaviour', which showed up in the analysis of candidate genes from both Europeans and Admixed Americans (who often have at least some Spanish roots (Montinaro et al., 2015)), see TABLE 4.4 end of this section. When

*'FDR q-value' is the correction of the p-value for multiple testing using the method from (Benjamini and Hochberg, 1995).

performing a GO enrichment analysis for each of the five European populations separately, the GO term associated to 'social behaviour' was also enriched, but only for Spain (IBS) and Italy (TSI) (see APPENDIX B.6). On closer inspection of the genes attributed to the GO Term, we found that most of its genes - *CNTNAP2, ANXA7, PPP3CB, MSS51* - are involved in autism and/or schizophrenia; *CNTNAP2* is even thought to belong to one of the major genes responsible for the autism spectrum disorder (Canali et al., 2018; Liu et al., 2011). Although there are studies showing a lower number of 'Hispanics' diagnosed with autism compared to 'non-Hispanic Whites', it has been suggested to be mainly attributable to socioeconomic factors like the gap in the health care system or the parental understanding of the disease (Palmer et al., 2010). However, other studies have shown that in children of Hispanic origin autism is more likely to be accompanied by other mental disorders (Becerra et al., 2014). In general, comparing global prevalence of autism no conspicuous indication can be found (Elsabbagh et al., 2012), more analysis is needed towards functions these gene might be involved. In any case, our results are in favour of a genetic component being involved in the autism related differences between Hispanic and non-Hispanic people.

African Population				
AFR - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0006565	L-serine catabolic process	2.03E-4	1E0	SDSL, SDS
GO:0006567	threonine catabolic process	2.03E-4	1E0	SDSL, SDS
GO:0019518	L-threonine catabolic process to glycine	2.03E-4	1E0	SDSL, SDS
AFR - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0004794	L-threonine ammonia-lyase activity	3.4E-5	1.55E-1	SDSL-like, SDS
GO:0003941	L-serine ammonia-lyase activity	1.02E-4	2.32E-1	SDSL-like, SDS
GO:0022834	ligand-gated channel activity	1.37E-4	2.09E-1	GRIK5, TPCN1, SCNN1G, KCNK6, GABRA2, RYR1
AFR - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0031301	integral component of organelle membrane	4.51E-4	8.6E-1	YIF1B, SLC8B1, GABRA2, SYT1, AGK, RYR1
GO:0031300	intrinsic component of organelle membrane	6.89E-4	6.57E-1	YIF1B, SLC8B1, GABRA2, AGK, SYT1, RYR1
GO:0042734	presynaptic membrane	9.84E-4	6.26E-1	GRIK5, CASK, GRM2, SYT1
European Population				
EUR - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0035176	social behaviour	4.04E-5	6.1E-1	ANXA7, PPP3CB, DNAJC9, MSS51, DVL1
GO:0051703	intraspecies interaction between organisms	4.04E-5	3.05E-1	ANXA7, PPP3CB, DNAJC9, MSS51, DVL1
GO:0072593	reactive oxygen species metabolic process	1.1E-4	5.55E-1	NNT, DUOXA2, CYB5R4, DUOXA1, DUOX2, DUOX1
EUR - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0016174	NAD(P)H oxidase activity	7.79E-6	3.55E-2	CYB5R4, DUOX2, DUOX1
GO:0050664	oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor	6.25E-5	1.43E-1	CYB5R4, DUOX2, DUOX1
GO:0005031	tumor necrosis factor-activated receptor activity	5.35E-4	8.14E-1	TNFRSF4, TNFRSF25
EUR - Cellular component				
No GO Enrichment Found.				
Admixed American Population				
AMR - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0072673	lamellipodium morphogenesis	2.93E-6	4.43E-2	PLEKHO1, WASF2, SNX1
GO:0035176	social behaviour	8.38E-6	6.33E-2	ANXA7, CNTNAP2, PPP3CB, DNAJC9, MSS51
GO:0051703	intraspecies interaction between organisms	8.38E-6	4.22E-2	ANXA7, CNTNAP2, PPP3CB, DNAJC9, MSS51
AMR - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0035035	histone acetyltransferase binding	3.95E-4	1E0	BCAS3, TRIP4, ECD
AMR - Cellular component				
No GO Enrichment Found.				

TABLE 4.4: Top three significant GO terms of African, European and Admixed American superpopulations.

East Asian Population				
EAS - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0006334	nucleosome assembly	7.37E-20	1.11E-15	HIST1H1[D/E], HIST1among H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H4[D/E/F/H]
GO:0034728	nucleosome organization	6.71E-18	5.07E-14	HIST1H1[D/E], HIST1H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H4[D/E/F/H]
GO:0065004	protein-DNA complex assembly	3.3E-17	1.66E-13	HIST1H1[D/E], HIST1H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H4[D/E/F/H], GTF2H3
EAS - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0046982	protein heterodimerization activity	6.09E-12	2.78E-8	HIST1H2A[C/D/E], HIST1H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H4[D/E/F/G/H]
GO:0046983	protein dimerization activity	1.95E-6	4.45E-3	HIST1H2A[C/D/E], HIST1H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H4[D/E/F/G/H], RILPL1, TP53I3
GO:0031491	nucleosome binding	2.08E-5	3.16E-2	HIST1H3[D/E/F/G], MLLT10
EAS - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0000786	nucleosome	9.02E-29	1.72E-25	HIST1H1[D/E], HIST1H2A[C/D/E], HIST1H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H2A[C/D/E/F/G], HIST1H4[D/E/F/H]
GO:0044815	DNA packaging complex	6.57E-28	6.27E-25	HIST1H1[D/E], HIST1H2A[C/D/E], HIST1H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H2A[C/D/E/F/G], HIST1H4[D/E/F/H]
GO:0032993	protein-DNA complex	3.36E-24	2.14E-21	HIST1H1[D/E], HIST1H2A[C/D/E], HIST1H2B[C/D/E/F/G/H/I], HIST1H3[D/E/F/G], HIST1H4[D/E/F/H] GTF2H3
South Asian Population				
SAS - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0070059	intrinsic apoptotic signalling pathway in response to endoplasmic reticulum stress	3.63E-4	1E0	TMBIM6, TMEM117, MAP3K5
SAS - Molecular function				
No GO Enrichment Found				
SAS - Cellular component				
No GO Enrichment Found.				

TABLE 4.5: Top three significant GO terms of East Asian and South Asian superpopulations.

Chapter 5

Measuring linkage disequilibrium using genealogical tree topology

In this chapter, we want to demonstrate, that linkage disequilibrium between two chromosomal loci can be measured by means of genealogical tree topology. For this purpose, in (Wirtz, Rauscher, and Wiehe, 2018) a measure of *topological linkage disequilibrium* (*tLD*) was introduced, based on clustering chromosomes with respect to their position in the genealogy rather than defining haplotypes as allele combinations at two loci as in the classical concept of linkage disequilibrium. In (Wirtz, Rauscher, and Wiehe, 2018), the focus lies on the theoretical properties of *tLD* of which the corresponding mathematical proofs were carried out by Johannes Wirtz and thus details on derivations can be read in (Wirtz, Rauscher, and Wiehe, 2018). My contribution was the performance of simulations and the application to experimental data, to analyse the accordance with the theoretical results.
In the following, the concept of *tLD* will be introduced and the application of *tLD* to the 1,000 human phase 3 data will be presented.

5.1 Classical concept of linkage disequilibrium *(LD)*

The classical concept of linkage disequilibrium (*LD*) refers to the non-random associations of alleles at different loci. Consider two markers at different sites. One marker has alleles A and a, and the other marker alleles B and b. Four haplotypes of these markers are possible: AB, Ab, aB and ab. Let p_A be the frequency of allele A in the population, p_a frequency of allele a, p_B of allele B and p_b of allele b. The expected frequency of the haplotypes is the product of the respective allele frequencies, namely $p_{AB} = p_A p_B$, $p_{Ab} = p_A p_b$, $p_{aB} = p_a p_B$ and $p_{ab} = p_a p_b$. Any deviation of the expected haplotype frequencies is linkage disequilibrium, which is typically

indicated by the letter D, and can be calculated by, e.g.

$$D = p_{AB} - p_A p_B.$$

When $D = 0$, the loci are said to be in linkage equilibrium.
In the following, let $x_1 := p_{AB}$, $x_2 := p_{Ab}$, $x_3 := p_{aB}$, $x_4 := p_{ab}$.
Note, that $x_1 = p_A p_B + D$, $x_2 = p_A p_b - D$, $x_3 = p_a p_B - D$ and $x_4 = p_a p_b + D$. Thus, D can be rearranged to

$$D = x_1 x_4 - x_2 x_3.$$

Let c be the recombination rate between the A/a and B/b locus. The frequencies of the haplotypes in the next generation (symbolized in the following by x_1', x_2', x_3' and x_4') can be calculated by, for example,

$$\begin{aligned} x_1' &= x_1^2 + x_1 x_2 + x_1 x_3 + (1-c) x_1 x_4 + c x_2 x_3 \\ &= x_1(x_1 + x_2 + x_3 + x_4) - c(x_1 x_4 - x_2 x_4) \\ &= x_1 - c D_0, \end{aligned}$$

where D_0 is the initial state of LD.
The frequencies of the other haplotypes can be derived likewise, and thus it holds that D in the next generation is

$$\begin{aligned} D_1 &= x_1' x_4' - x_2' x' x_3' \\ &= (1-c) D_0. \end{aligned}$$

It follows by recursion that

$$D_{t+1} = (1-c) D_t = ... = (1-c)^t D_0,$$

where D_t is LD at generation t. Finally, for small c, D in generation t can be approximated by

$$D_t = (1-c)^t D_0 \approx e^{-ct} D_0. \tag{5.1}$$

This shows an important result:
In each generation LD decays at a rate determined by the degree of recombination and particularly, LD depends on recombination rate.

D is easy to calculate, however, its big disadvantage is that its range is dependent on allele frequencies in the population, given by

$$D_{\min} = \max\{-p_A p_B, -p_a p_b\}$$
$$D_{\max} = \min\{p_A p_b, p_a p_B\}.$$

D maximises when allele frequencies are both 0.5, but for example if $p_A = 0.3$ and $p_B = 0.1$, the range is restricted to -0.03 and 0.07.
Lewontin (1964) suggested using a normalisation of D:

$$D' = \begin{cases} \frac{D}{D_{\max}}, & \text{if } D \text{ pos.} \\ \frac{D}{D_{\min}}, & \text{if } D \text{ neg.} \end{cases}$$

D' has the nice property that it is equal to 1 if two sites are in complete *LD* and 0 for no *LD*. Its disadvantage is when alleles are rare or the population size is small, D' tends to be enlarged, making it difficult to be interpreted correctly.
Another way of measuring *LD* is to use a correlation coefficient of the allelic association, first introduced by Hill and Robertson (1968),

$$r = \frac{x_1 x_4 - x_2 x_3}{\sqrt{p_A p_a p_B p_b}}, \tag{5.2}$$

which ranges between -1; strong negative correlation, and 1, strong positive correlation. If r is equal to 0 the two sites are not correlated.
This *LD* measure allows for statistical testing of significance, since r is related to the χ^2-distribution: it holds that $r = \sqrt{\chi^2/n}$. This can be obtained from the 2×2 table of the frequencies x_1, x_2, x_3 and x_4 and n is the total number of haplotypes in the sample.
Mostly, it is common to consider r^2.

In (Wirtz, Rauscher, and Wiehe, 2018), a new approach of defining linkage disequilibrium was introduced in the framework of coalescent theory.

5.2 The topological linkage disequilibrium *(tLD)*

As we have already explained in section 3.1, due to recombination event, tree topology at different sequence positions may change along the chromosome. In the ARG, each nucleotide position along the chromosome is associated with a coalescent tree, and within a chromosome segment with no recombination events all positions have the same tree topology. By dividing chromosomes into recombination-free fragments, coalescent trees can be associated with a fragment.

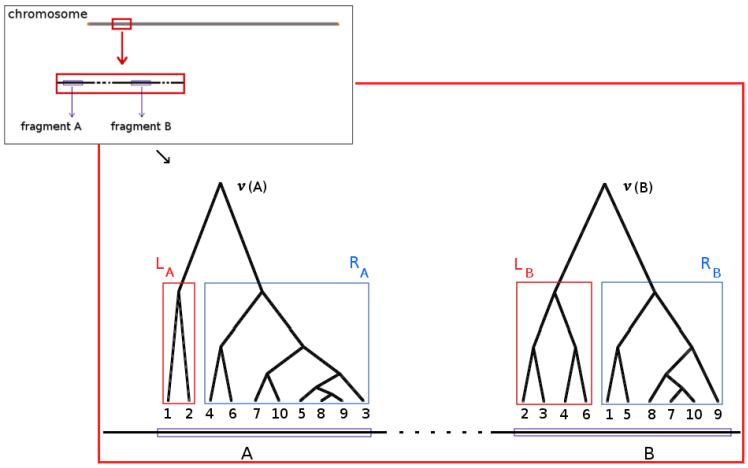


FIGURE 5.1: Coalescent trees along a recombining chromosome of size $n = 10$. Zoom-in of a small part of a chromosome. Consider two fragments of a given window size, labelled as fragment A and fragment B. These two fragments can be associated with a coalescent tree. Recombination events between fragment A and B might have changed not only the tree topology, but also the assignment of chromosomes with regards to the left and right side of the root of the trees.

Likewise in section 3.1, consider a binary tree of size n, the n leaves of the tree can be divided into two disjoint groups: the left and the right-descendants of the root $v_{(\cdot)}$. The two groups are indicated as $L_{(\cdot)}$ and $R_{(\cdot)}$, respectively, and without loss of generality let $L_{(\cdot)}$ be the smaller of the two sets $L_{(\cdot)}$ and $R_{(\cdot)}$. As a consequence of recombination events, when moving along a chromosome, the genealogical tree of fragment A may differ from the tree at fragment B. Moreover, the descendants belonging to the left and right set below the root of the tree associated to fragment A may differ from those of fragment B. In the following, let L_A indicate the left set of the tree associated to fragment A, R_A the right set, and so forth (see FIGURE 5.1).
We can now define a correlation measure as follows:

- Let p_{L_A} be the frequency of chromosomes in L_A:
  $p_{L_A} = |L_A|/n$, and likewise
  $p_{R_A} = |R_A|/n$, $p_{L_B} = |L_B|/n$, $p_{R_B} = |R_B|/n$.
- Let x_1 be the proportion of chromosomes belonging to $L_A \cap L_B$:
  $x_1 = |L_A \cap L_B|/n$, and likewise
  $x_2 = |L_A \cap R_B|/n$, $x_3 = |R_A \cap L_B|/n$, $x_4 = |R_A \cap R_B|/n$.

Then, we define the *topological linkage disequilibrium*, short *tLD*, as

$$r^2_{tLD} = \frac{(x_1x_4 - x_2x_3)^2}{p_{L_A}p_{R_A}p_{L_B}p_{R_B}}. \tag{5.3}$$

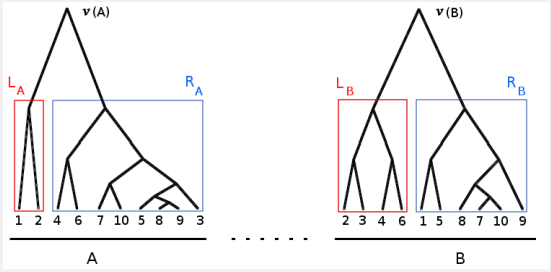
The term is coined *topological* since it is induced by the topology of the coalescent tree.

In the following, we will write r^2 for the conventional *LD*, and r^2_{tLD} for the topological *LD*.

[Remark: In (Wirtz, Rauscher, and Wiehe, 2018), *tLD* is defined as $r^2_{S,U}$, where S and U refers to the two fragments, whilst the conventional *LD* is defined by $r^2_{\alpha,\beta}$, where α and β refers to the two loci.]

Box 5.2: Example of *tLD* using FIGURE 5.1

In this example, we have



- $n = 10$
- $L_A = \{1,2\}$,
  $R_A = \{3,4,5,6,7,8,9,10\}$
- $L_B = \{2,3,4,6\}$,
  $R_B = \{1,5,7,8,9,10\}$

Thus:

- $p_{L_A} = |L_A|/n = 2/10$, $p_{R_A} = |R_A|/n = 8/10$,
- $p_{L_B} = |L_B|/n = 4/10$, $p_{R_B} = |R_B|/n = 6/10$,

and

- $x_1 = |L_A \cap L_B|/n = 1/10$, $x_4 = |R_A \cap R_B|/n = 5/10$,
- $x_3 = |R_A \cap L_B|/n = 3/10$, $x_2 = |L_A \cap R_B|/n = 1/10$.

Substitute in equation (5.3) yields $r^2_{tLD} = 0.0104167$.

Like in the conventional *LD*, the choice of the left and the right set of the root of the tree is not of importance, since it does not have an affect on r^2_{tLD}.
r^2_{tLD} can only be equal 1, if $L_A = L_B$ or $L_A = R_B$.

5.2.1 Properties of *tLD*

As we have seen in equation (5.1), recombination affects *LD*: *LD* decays in each generation at a rate determined by the degree of recombination.
However, if recombination and genetic drift is combined in a finite population N, it is not easy to derive the expected value for r^2. By assuming completely unlinked loci, the configuration of alleles forming a haplotype behaves statistically like a random

2×2-table, and according to Haldane (1940)

$$E[r^2] = \frac{1}{N-1}. \tag{5.4}$$

The question still remains how the expected *LD* decays with respect to the recombination rate. Several efforts to come up with a reasonable formula have been made. Sved (Sved, 1971) approximated the expected equilibrium *LD*

$$E[r^2] \approx \frac{1}{1+4Nc\frac{1-\frac{c}{2}}{(1-c)^2}} \overset{c<<1}{\approx} \frac{1}{4Nc+1}. \tag{5.5}$$

by relating r^2 to the conditional probability of linked identity by descent which is the probability that two chosen haplotypes will be identical copies from some previous generation. This formula illustrates that if $4Nc$ is small, the expected *LD* will approach 1, if $4Nc$ is large, then it will approach 0. If $4Nc$ is large the equation can be approximated by

$$E[r^2] \approx \frac{1}{4Nc}.$$

Note, that we have seen the quantity $4Nc$ before, it is the population recombination rate. To avoid ambiguity, from now on we define the population recombination rate by the Greek letter ρ.

Despite the discrepancy between (5.5) and (5.4), Sved's formula (5.5) has become one of the standard approaches.

Still today, attempts to improve the approximation (5.5) exist and researchers are concerned to find a more suitable formula describing the expected *LD* with respect to the recombination rate , e.g. (Ober et al., 2013). But none of them succeeded to approach Haldane's value.

By using the concept of *tLD*, in (Wirtz, Rauscher, and Wiehe, 2018) a new formula for the decay of expected r^2_{tLD} has been theoretically derived. It has been shown, that

$$E[r^2_{tLD}] \xrightarrow{\rho\to\infty} \frac{1}{1-N},$$

by using arguments derived from coalescent properties.

Thus, *tLD* decays towards the same value as in Haldane's formula (5.4).

Furthermore, by using simulated data, it could have been shown, that *tLD* decays more slowly than the conventional *LD* with chromosomal distance (see FIGURE 5.2). This can be explained by the fact that only a fraction of recombination events affects tree topology at the root. Indeed, in (Wirtz, Rauscher, and Wiehe, 2018, Lemma 2), it could be theoretically deduced, that about $1/3$ of all recombination events lead to changes in such a way that chromosomes from one side of the tree are shifted to

the other. Tree topology is estimated from SNP data in the exact same manner as in previous chapters (re-visit section 3.2.1 for cluster method).

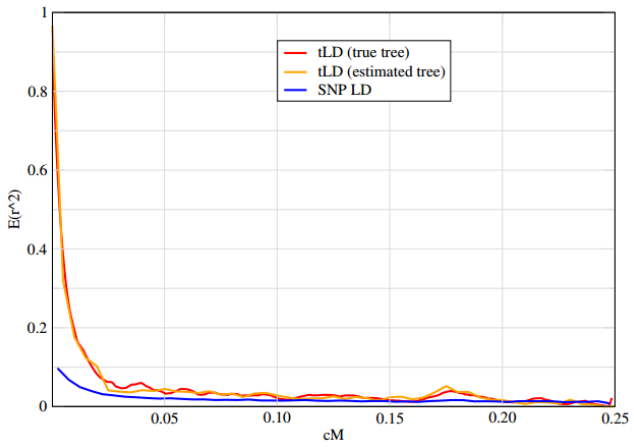


FIGURE 5.2: Figure also shown in (Wirtz, Rauscher, and Wiehe, 2018, Figure 6). Illustration of decay of *tLD* vs. SNP-*LD* with chromosomal distance from simulated data.
Data are from a single simulation run generated with the program *ms*. The parameters were set in such a way that a chromosomal sequence with a recombination rate of 1cM/Mb and length 250kb (0.25cM) was simulated, for $N = 10^4$. The corresponding *ms*-command line was therefore: `ms 200 1 -t 100 -r 100 1000 -T`, where the option `-T` outputs true tree topology in Newick-format (more on *ms* output see FIGURE 3.4).

5.3 Application of *tLD* to 1,000 Humans Data

In this section we will present the application of *tLD* to human data from the human 1,000 genomes project (Auton et al., 2015). The estimation of genealogical tree topology for all 26 populations was already performed previously (see chapter 4). Since the focus lies on the root of the entire tree T for a sample of size n, the MRCA, we only need to consider the first clustering step: the one dividing the n chromosomes into the 'left-descendants' and into the 'right-descendants' of root ν_1, L_1 and R_1 respectively (for terminology, re-visit chapter 3). In contrast to determining T_3, where the size $|L_1|$ or $|R_1|$ is needed, for this concern the 'content' of each cluster is needed. In section 3.2.2, we already analysed how well the assignment of the estimated cluster agrees with the true one. We have shown that if $|L_1| = |\hat{L}_1|$ (or $|R_1| = |\hat{R}_1|$), the clusters agree very well with the true one. Moreover, in chapter 3 we have shown as well that a minimum of 10 SNPs is sufficient to yield a good estimation result also with regards to size: the average difference between known Ω_1 and estimated $\hat{\Omega}_1$ was around 0 (see FIGURE 3.8). That the true and estimated values of *tLD* agree quite well, is once more demonstrated by a heatmap in FIGURE 5.3, where the same simulated data are used as in FIGURE 5.2.

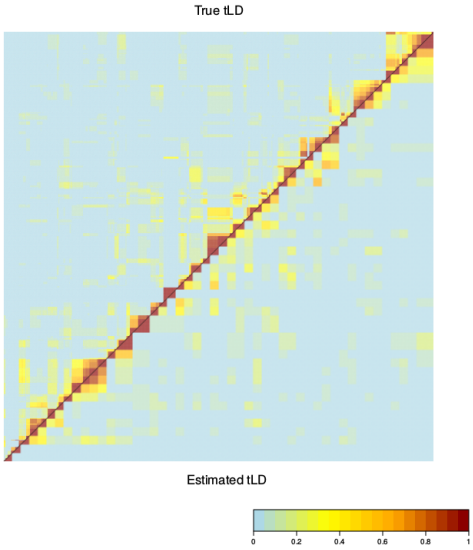


FIGURE 5.3: Figure also shown in (Wirtz, Rauscher, and Wiehe, 2018, Figure 7). Heatmaps of tLD calculated from tree topologies and tLD calculated from estimated tree topologies, performed on the same simulated dataset used in FIGURE 5.2. The diagonal starting from the bottom left corner to the top right corner represents the simulated chromosome sequence, position starts from down left and ends top right. The heatmap on the upper left side of the diagonal represents the *tLD* calculated from true tree topology and the heatmap on the right side below the diagonal *tLD* from estimated data.

We calculated *tLD* for some previously found candidates. First, we determined *tLD* for a 2Mb region on chromosome 2, containing the genes *ZRANB3*, *LCT*, *MCM6*. Remember from previous chapter, that the gene *ZRANB3* was suggested to be under positive selection for some European populations, whilst according to our result the well-known sweep candidate genes *LCT* and *MCM6* were not amongst our list of candidates. We wanted to investigate whether *tLD* provides indications for potential interaction between these genes. For reasons of comparison, we also show the classical *LD* for this example.

First of all, FIGURE 5.4 illustrates a clear signal of elevated level of linkage disequilibrium for the European population CEU in comparison to the African population YRI. Generally, this was expected since African populations are known to show lower levels of linkage disequilibrium in general among loci compared to non-Africans (Campbell and Tishkoff, 2008). However, note that the signals are stronger to be observed for *tLD* than conventional *LD*. This may be not surprising, since *tLD* is calculated over segments and can be therefore seen as 'an average' over blocks of SNPs and as such as a 'coarse-grained' measure for the classical *LD*. However, exactly this can also be seen as the advantage of using *tLD*, since the signal is stronger and thus easier to detect. With regards for visual inspection, this is clearly a benefit. Furthermore, in accordance with results mentioned in the previous section (see FIGURE 5.2), the level of correlation seems to be maintained at a higher level for a longer

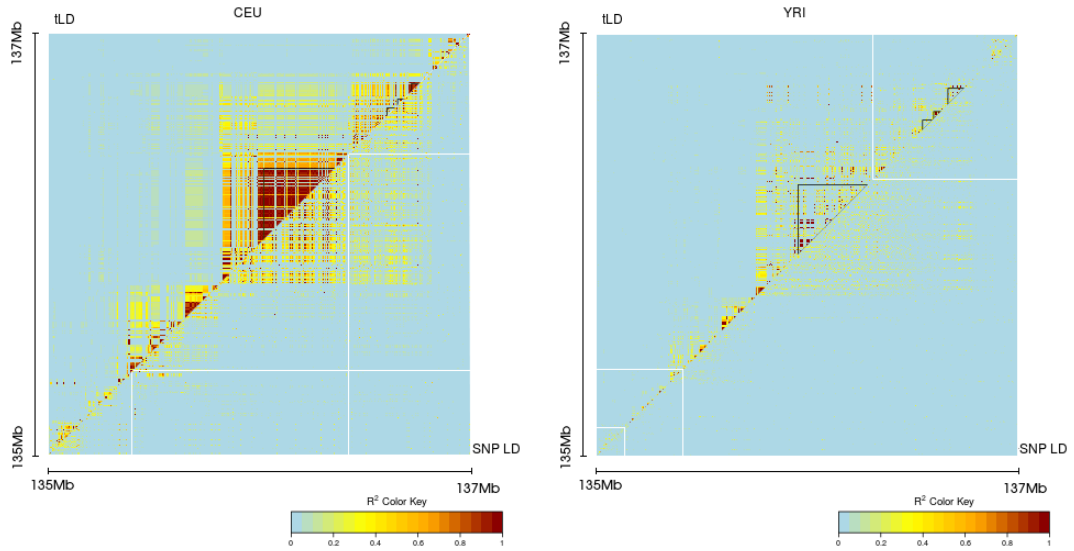


FIGURE 5.4: Heatmaps of *tLD* (upper (left) triangle) for chromosome region chr2:135,000,000-137,000,000 for population CEU and YRI (diagonal from left to right). For reasons of comparison heatmaps of the conventional *LD* (here: *SNP LD* for same chromosome region is shown on lower (right) triangle. Form left to right, the positions of the genes *ZRANB3*, *LCT*, *MCM6* and *DARS* are indicated by the dark triangles within the plot.

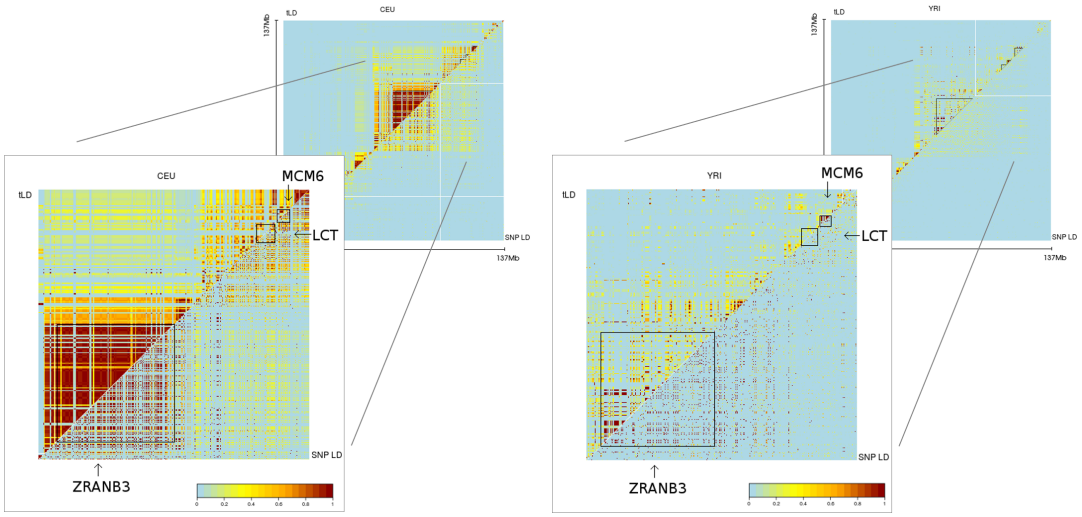


FIGURE 5.5: Zoom-in of region surrounding genes *ZRANB3*, *LCT*, *MCM6*, from FIGURE 5.4.

chromosomal distance compared to classical *LD*. Therefore, *tLD* may be more suitable for detecting long-range linkage disequilibrium.

Our findings show a clearly elevated *tLD* for the region containing the genes *ZRANB3*, *LCT* and *MCM6* (FIGURE 5.5). This might be an indication for interacting functions between ZRANB3 and one of the other genes, responsible for the linkage.

As another example we determined *tLD* for a region on chromosome 15, containing the genes *OCA2* and *HERC2*. In chapter 4, we found *HERC2* to be a sweep candidate

gene for the European population CEU, GBR and FIN, a gene where some of its variants result in blue eyes. Since *HERC2* contains a region regulating the activity of the *OCA2* gene, which in turn controls the eye pigmentation, we were interested if an elevated tLD can be observed in that region. Furthermore, in (Hubacek et al., 2017) a list of alleles was presented, which are suggested to be responsible for the blue eye variant. Therefore, we analysed the region containing these two genes, if elevated linkage can be observed between these regions for the three European population. Our result in FIGURE 5.6 indicate that indeed tLD seemed to be elevated in particular in regions between the three alleles mentioned in (Hubacek et al., 2017) located within the *OCA2* and the (whole) *HERC2*. However, tLD is contiguously high in the regions containing the *HERC2* and *OCA2* gene. Therefore, not directly the three alleles might be responsible for the observed strong signal in this region. Nevertheless, the difference of strength of the signals between the classical and the topological LD in this region is tremendous, in particular for gene *HERC2*, even for the African population YRI. Whilst signals for the classical LD are rather restrained, tLD is quite strong in this region.

SNP ID	Position	Within gene	Gene position
rs4778138	chr15:28,335,820-28,335,820	OCA2	chr15:28,000,023-28,344,458
rs4778241	chr15:28,338,713-28,338,713	OCA2	
rs7495174	chr15:28,344,238-28,344,238	OCA2	
rs1129038	chr15:28,356,859-28,356,859	Herc2	chr15:28,356,183-28,567,298
rs12913832	chr15:28,365,618-28,365,618	Herc2	
rs916977	chr15:28,513,364-28,513,364	Herc2	
rs1667394	chr15:28,530,182-28,530,182	Herc2	

TABLE 5.1: SNPs known to be responsible for the blue eye variant according to (Hubacek et al., 2017).

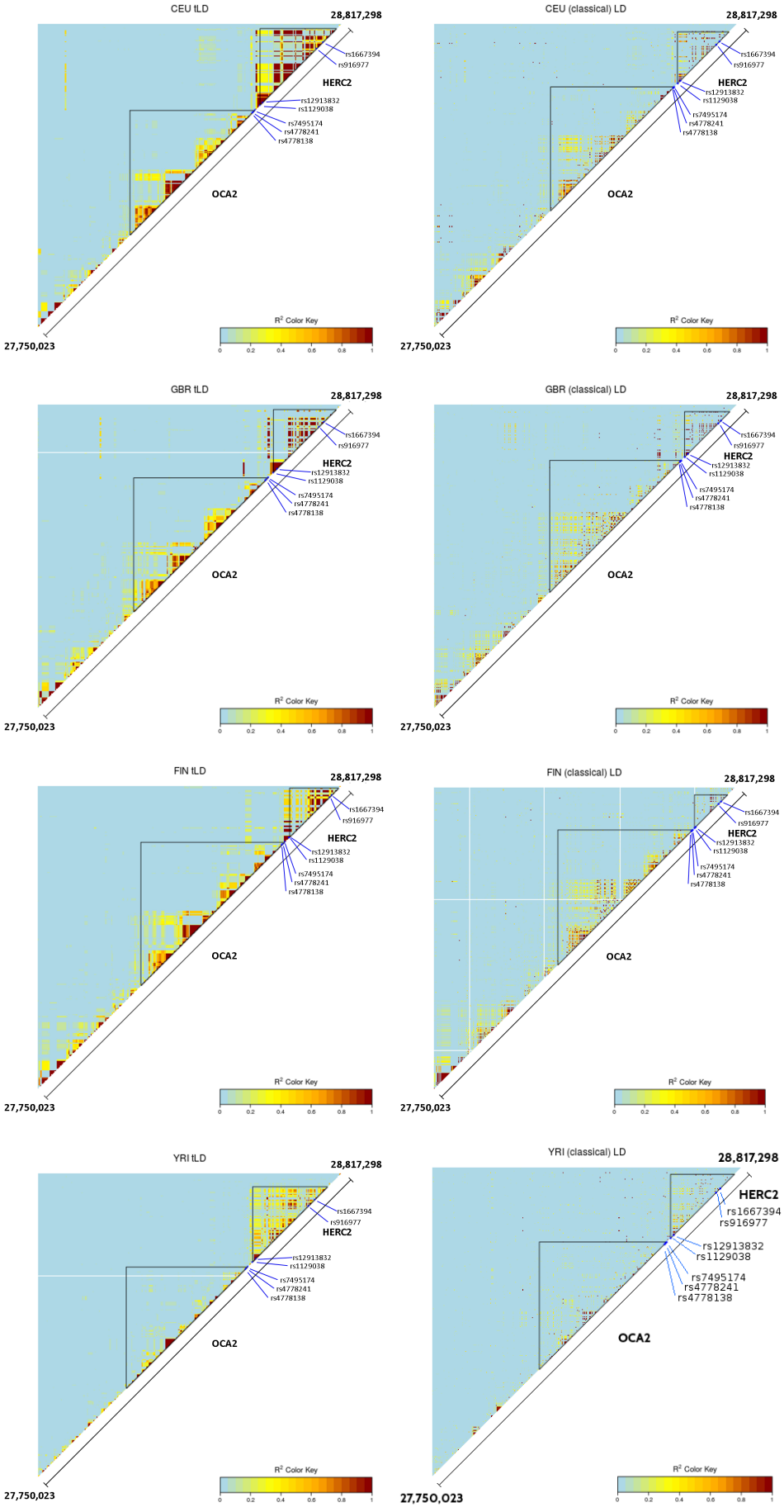


FIGURE 5.6: Heatmaps of *tLD* (left) and classical *LD* (right) for chromosome region chr15:27,750,023-28,817,298 for population CEU, GBR, FIN and YRI. The positions of the genes *OCA2* and *HERC2* are indicated by the dark triangles within the plot (diagonal from left to right).

Chapter 6

Conclusions and outlook

Understanding the role of evolutionary forces leading to the observed genomic patterns in and between different organisms or populations is a challenging task for scientists. These patterns might be shaped by factors such as demographic events, natural selection or simply random drifts. Distinguishing between those can be difficult since demographic events, like population bottlenecks, can leave a similar genomic pattern behind as those left by the action of natural selection. The construction of a robust test statistic aiming in identifying the correct underlying dynamic behind, received a high degree of attention for researchers.

Coalescent tree topology is not affected by varying population size (Hudson, 1990; Li, 2011). This motivated us to investigate the topologies of genealogical trees in more detail, and to establish new methods contributing to the research of evolutionary mechanisms.

In (Li and Wiehe, 2013), the authors proposed a test statistic called T_3, which only uses the information of coalescent tree topology. Selective sweeps can produce highly unbalanced coalescent tree topologies in regions close to a selected site. Under neutral evolution T_3 is expected to be standard-normally distributed. Genealogies after a selective sweep tend to be unbalanced and to produce negative values of T_3 (see section 3.1). Hence, T_3 detects bias in tree balance. However, in practice the tree topology is not known and has to be estimated. Whilst in (Li and Wiehe, 2013) microsatellite data was used for the estimation of tree topology, we show that SNP data provides a good alternative to microsatellite data for estimating the tree topology. In chapter 3 we present in detail, how many SNPs are at least needed to obtain a good cluster estimation result. In the absence of recombination, this number can be arbitrarily large. However many recombination events within a chromosomal segment should be avoided, since this increases the probability of having multiple tree topologies within the segment, leading to confounding tree topologies. In

(Ferretti, Disanto, and Wiehe, 2013) it was shown that it takes about 15-20 recombination events to drastically reduce correlation of coalescent tree topologies along a recombining chromosome. 15-20 recombination events correspond to roughly to 6,400-8,520 bp to for a sample of size $n = 200$, $N = 10^4$ and a recombination rate of $c = 10^{-8}$ per bp (see equation (3.2.1)). We decided to set the maximum window length to 10 kb. In the same section 3.2.1 we demonstrated that a minimum number of SNP is needed to get a fairly good approximation of the true tree topology. A too small number of SNPs led to an under- or overestimation of tree cluster. Besides performing simulations, we underpinned the expected cluster size, conditioned on the number of SNPs used for the estimation, by explicit calculations, as long as these didn't become too complex. We concluded that a minimum number of ten SNPs already yield a cluster size estimation which agrees quite well with the true one. We expect to see ten SNPs in a magnitude of about $\sim$ 4,260 bp window length, see equation (2.2). In such way, we came to the conclusion to estimate tree topology using chromosomal segments of size 5 kb and a step size of 2.5 kb. The chromosomal segment needed to contain at least ten SNPs. If the latter condition was not fulfilled, we extended the window size by 1 kb, up to a maximum window of size 10 kb. If the clusters were not clearly resolvable, we randomly assigned the sequences to one of the two clusters with equal probability. Here, we want to point out, that our choice for the fragment length rely on the assumption of a recombination rate of $c = 10^{-8}$ per bp per generation and $\mu = 10^{-8}$ per bp per generation, which are the (average) estimates for human (Roach et al., 2010; Li and Freudenberg, 2009). Therefore, if applying to species with different mutation and recombination rates as assumed above, the parameters must be changed correspondingly.

To analyse, how the T_3-test, using SNP data for the tree topology estimation, performs under different demographic scenarios, we first generated three data sets: one simulating a population bottleneck scenario, which was compared to the neutral and to the selective sweep scenario (see section 3.3). The results clearly showed, that the T_3-test was quite robust under the population bottleneck scenario, as expected. Furthermore, we examined how the T_3-test performs in presence of population substructure. For this end, we generated various sampling schemes with varying migration rates. We have seen that substructured population and low migration rate affects the T_3-test, in particular when the sampling scheme is heavily biased ($n_1 = 195$ and $n_2 = 5$) and migration rate is low ($4Nm = 0.4$). When sampling all chromosomes from only one subpopulation, $n_1 = 200$ and $n_2 = 0$, T_3 is quite robust when migration rate is moderate ($4Nm = 4$) or very low ($4Nm = 0.04$). When migration rate is low ($4Nm = 0.04$), T_3 seems to be slightly affected, see TABLES 3.1 and 3.2.

Generally, the power of the T_3-test is strongly dependent of the distance to the selected site (see TABLE 3.3). If considering single windows, regardless of their position from the selected site, (although the false positive rate was only around 0.019) the power is only around 0.23 (in case of strong selection, otherwise even below). If we take the distance to the selected site into account, on average, taking a 1% threshold, around 78.86% - 86.12% of the windows identified as being significant were found to be within a distance of 250 kb from the selected site (see table 3.3). However, also an average of around 20% (by a threshold of 1%) falls outside the 250 kb region. Next, we investigated if a re-sampling strategy can help to corroborate significance of previously identified regions. The underlying idea was that induced subtree topologies of unbalanced trees generated under neutrality might be distinguishable from subtree topologies of unbalanced trees generated under selection. It has been shown before, that this is true for the most extreme case of an unbalanced tree, namely a caterpillar tree: its induced subtree is always highly unbalanced. A caterpillar tree can result in a large excess of singleton mutations, which is a typical characteristic of a selective sweep, however a caterpillar tree is also extremely unlikely to be observed in practice (Blum and Francois, 2006; Kirkpatrick and Slatkin, 1993). Our simulation results could not show a considerable improvement in filtering out previously identified false positives (see TABLE 3.4). Therefore, we suggest that the aforementioned hypothesis (that highly unbalanced trees resulting from positive selection inherit this property to their induced subtree whilst highly unbalanced trees generated under neutrality don't) might only hold for 'extreme' cases like caterpillar trees, which are very rare in practice. Besides, on the technical side, this approach requires a long running time and a large memory, making it unsuitable for genome-wide screens.

We then turned our focus to a different strategy. Since unbalanced tree topologies in multiple adjacent regions are more likely to be observed in regions close to the selected sites than by chance, see section 3.4.1, we not only took the T_3-value of one window into account, but also the surrounding ones and thus constructed a test statistic based on the concept of likelihood ratio tests. We called this test the LR_{T_3}-test (see 3.6).

We empirically determined the power of this test, and found that by taking a threshold of $LR_{T_3}-$score to 0, we get a false positive rate of 0.0226, and a power of 0.95. To reduce the false positive rate, we decided to set the threshold-score to 15. In such way, we could reduce the false positive rate to almost 0 (0.0007%), at a price of reduced power: 0.88, however this is still quite good.

In addition, we showed in this chapter, that our test is applicable not only to detect recently completed sweeps, but also incomplete sweeps: the signal was even strongest to be observed when the selected site has reached a frequency of around 80% (section 3.5).

In conclusion, we derived a test statistic solely relying on the knowledge of coalescent tree topology. It is free from the effects of varying population size, from which some test statistics suffer (Ramirez-Soriano et al., 2008), it is slightly affected by migration events, however when sampling scheme is in such way that all chromosomes are sampled from only one subpopulation, it still performs quite well. Furthermore, it is also able to detect incomplete sweeps.
One disadvantage is, that the reliability of the T_3-test depends on the quality of the estimated tree topology. Therefore, one should seek to improve the clustering method. So far, we estimated tree topology according to a sliding window approach; we estimated tree topology for each window independently. But whilst doing so, we are aware of that tree topologies along a chromosome are not independent, but correlated to each other. Instead of estimating tree topology for each window separately, one might also take the topology of the neighbouring windows into account, for example in determining a conditional probability or likelihood of observing the estimated tree topology, given knowledge of the tree topology of the previous window.
In particular in cases, where the clusters were not clearly resolvable (and so far we just randomly assigned the sequences to one of the two clusters with equal probability), or regions, which were 'skipped' due to the lack of data/or monomorphic sites, the additional consideration of the neighbouring regions might help to be more accurate and thus, not to be as conservative. On the contrary, this might lead to an enormous increase of running time, just for the estimation of tree topology. The fact that one need to estimate not only Ω_1, but also Ω_2 and Ω_3 might add to the complexity of the issue.

In chapter 4, we have applied our test statistic LR_{T_3} to the human data from the 1,000 genomes project ((Auton et al., 2015), phase 3). For this end, we performed whole genome screens for all 26 populations; all 22 autosomes and the X chromosome. The 26 populations can be further divided into five so-called 'superpopulations': African, Admixed American, East Asian, European and South Asian (see FIGURE 4.1).
In general, we found approximately two times less candidate regions in the African superpopulation compared to the remaining four superpopulations (see FIGURE 4.4, or APPENDIX table B.1). Our result confirmed previous studies that have found more candidate regions for recent selective sweeps in non-African populations compared to the African populations (Kayser, Brauer, and Stoneking, 2003; Williamson et al., 2007; Campbell and Tishkoff, 2008). We compared our gene candidate list with previous studies. For this purpose we took two lists into consideration: the list from the *database of recent positive selection across human populations* (= dbPSHP) (Li et al.,

2014a), downloaded from http://jjwanglab.org/dbpshp and consisting of about approximately 8,050 candidate genes, and a more recent list taken from (Schrider and Kern, 2017), consisting of about approximately 5,939 candidate genes. The first list is a collection of all candidate sweep regions identified and published until then. For generating the latter list, the authors used a new method developed by themselves in a previous paper, called *S/HIC* (Schrider and Kern, 2016)), which is based on a supervised machine learning approach combining many statistics used to test for selection (including 'classical' tests like Tajima's D, haplotype based tests etc.). In general, the overlap between our candidates and both lists were rather moderate (with the dbPSHP-list: 1,947 genes, with the *S/HIC*-list: 1,718 genes, of which 1,253 genes are not found in dbPSHP). However, other studies have reported a similar result, concerning the small intersection of candidates between different studies (Akey, 2009; Schrider and Kern, 2017). They suggest that it is due to that different methods may produce different false positives and false negatives, resulting in this discord between scans.

Amongst several previously known candidate genes, we also found new potential candidates, for instance the gene *NXF5* on the X chromosome. This gene is involved in the normal functioning of the brain, kidneys and reproductive organs, since its disruptions can lead to disorders of these (Jun et al., 2001; Esposito et al., 2013; Fortuno and Labarta, 2014). The region where this gene is located is the strongest candidate region in the East Asian population CHB (see TABLE 4.3). The region containing this gene was also significant for two other East Asian populations JPT (maximum $LR_{T_3} = 179.14$) and CHS (maximum $LR_{T_3} = 80.48$).

The region where the overall highest LR_{T_3}-score was found (for the South Asia population ITU), is a candidate region for almost all non-African populations. One possible candidate gene driving this selection is *NELL2*. It has been previously recorded to be a sweep candidate, although so far no hypothesis of what the associated beneficial trait of it might be has been suggested. Previous studies have indicated a possible connection for this gene with the onset of puberty in rats (Ryu et al., 2011). As for humans, it is known that girls of the African American population enter puberty earlier than those with Caucasian or Hispanic ancestry (Ramnitz and Lodish, 2013) and we suggest that *NELL2* could be involved in variations of the human onset of puberty in human, although the reason for this trait to be under selection is unclear.

Another candidate gene with clear differences between African and non-African populations was *CASK* (FIGURE 4.8). This gene appears to be a strong candidate for three African populations: ACB (LR_{T_3}=102.35), YRI (LR_{T_3}=127.87) and LWK (LR_{T_3} =90.2757), and it is also significant for further three African populations GWD (LR_{T_3}

=22.9285), ASW (LR_{T_3}=38.3957), ESN (LR_{T_3}= 74.7938). Previously it has been suggested as a candidate for selection in only one of these populations, YRI (Frazer et al., 2007). Gazda et al. (2018) suggested *CASK* to be positively selected in racing pigeons for contributing to athletic performance, since the gene is involved in the formation of neuromuscular junctions. We note that athletes of African origin often perform exceptionally well in competitions and propose that *CASK* gene might be involved in that.

Furthermore, our test could confirm many other previously known genes, from which some of them we mentioned in section 4.3.2. Further on, we investigated whether some gene sets can be associated with functional genetic differences among different superpopulations. Therefore, we performed GO enrichment analysis. In doing so, we were in specific interested, if enrichment can be found for gene sets potentially involved with regards to the adaptation as a result in response to the Out-Of-Africa migration. We came to the conclusion, that in this regard no significant differences in biological functions between African and Non-African populations can be seen (see also APPENDIX B.6). However, our finding is consistent with other studies, for instance in (Campbell and Tishkoff, 2008). The authors pointed out, that despite Africans are more genetically diverse and also possess lower levels of linkage disequilibrium among loci compared to non-Africans, Africans also do have a number of genetic adaptations evolving due to diverse climates and diets. Furthermore, our GO enrichment analysis revealed an intriguing observation between the analysis of candidate genes from the European population Spain (IBS) and Italy (TSI) (see APPENDIX B.6), and the Admixed American superpopulation (see TABLE 4.4): For all of them the GO Term 'social behaviour' showed up to be among the top three most significant enriched GO terms. Most of the genes attributed to the GO Term are involved in autism and/or schizophrenia. According to our findings in section 4.3.3, we suggest that there might be an advantageous genetic component being involved in the autism related differences between Hispanic and non-Hispanic people, but we further suggest that more analysis is needed towards functions where these genes might be involved. In conclusion of chapter 4, we want to point out, that the application of the LR_{T_3}-test on the human 1,000 genomes data performed quite well, not only covering several previously known candidates, but also revealing new candidates. There are still many candidate genes we did not investigate from our list, including all genes not associated with the biotype 'protein-coding'. In particular, out of our candidate list, we found several superpopulation-specific ones. It would be interesting to analyse the biological function of genes driving the selection and the significance of its trait, but this is left for future projects. Another important aspect which has to be mentioned is that the result of the LR_{T_3}-test depends on the

underlying parameters we have set for the selective sweep scenario and the likelihood distributions we have empirically determined in the beginning. One could try to apply the LR_{T_3}-test under changed conditions and assumptions.

In chapter 5 we presented a new measure of *topological linkage disequilibrium* (*tLD*) which is based on the topology of genealogical trees (Wirtz, Rauscher, and Wiehe, 2018). Instead of focusing on haplotypes as allele combinations at two loci as for the classical *LD*, we cluster a sample of chromosomes with respect to their position in the genealogy. Therefore, the focus lies on the first root of the tree (MRCA) which divides the sample into two disjoint groups: the 'left-descendants' and the 'right-descendants' of the root, see section 5.2. The *tLD* is the correlation between the members of each group, see equation (5.3). The advantage of the *tLD* is that it is more sensitive than regular *LD* to detect long range interactions across megabase scales, which can be explained by the fact that only a fraction of recombination events affects tree topology at the root. This could be confirmed by the application of *tLD* to simulated data, see FIGURE 5.2. The tree topology was estimated using the aforementioned method from section 3.2.1, chapter 3. Furthermore, again we could have shown how well the estimated cluster agrees with the true one, 5.3 and also compare section 3.2.2.
We then applied the *tLD* to some previously found candidate genes. In chapter 4, the 'prominent' sweep candidate gene *LCT* did not appear in our list, however we did find a rather strong signal for the *ZRANB3* gene for the two European population CEU and GBR (see 4.3), which lies about 257 kb distant away from the *LCT* gene. Therefore, we were in particular interested if linkage between these two genes can be found. Indeed, our findings show a clearly elevated *tLD* between the genes *ZRANB3* and *LCT*, but also *MCM6*, which contains regulatory elements for *LCT* (Hubacek et al., 2017) (FIGURE 5.5). We suggest that there might be interacting functions between ZRANB3 and one of the other genes, responsible for the linkage. Generally, *tLD* shows stronger signals than the classical *LD*, which is not only a benefit for an easier detection, but also with regards to the visualisation.
This was further demonstrated for the region *HERC2* and *OCA2*, of which *HERC2* was another sweep candidate from our list for the three European population CEU, GBR and FIN. We analysed this region since on the one hand, *HERC2* is suggested to play a key role for the brown/blue eye colouring, but on the other hand the nearby *OCA2* seems to be the one actually controlling the eye pigmentation. According to studies (e.g. (Eiberg et al., 2008)), a region in *HERC2* was found to regulate the activity of the *OCA2* gene. Furthermore, in (Hubacek et al., 2017) a list of alleles was presented, suggested to be responsible for the blue eye variant. Therefore, we analysed the region containing these two genes, if elevated linkage can be observed

in particular between these regions for the three European population. Our result showed, that indeed *tLD* seemed to be elevated in particular in regions between three alleles mentioned in (Hubacek et al., 2017) located within the *OCA2* gene and the (whole) *HERC2* gene, (see FIGURE 5.6). However, *tLD* is contiguously high in the regions containing the *HERC2* and *OCA2* gene. Therefore, not directly the three alleles might be responsible for the observed strong signal in this region, since these are very closely located to the *HERC2* gene. Nevertheless, the difference of strength of signals between the classical and the topological *LD* in this region is tremendous, in particular for gene *HERC2*, even for the African population YRI. Whilst signals for the classical *LD* are rather restrained, *tLD* is quite strong in this region.
Summing up, *tLD* offers a new method for measuring linkage between two loci, which only relies on the genealogical tree topology. Signals from *tLD* are stronger to be observed. Since *tLD* decreases slower than classical *LD* with distance, it may be more suitable to detect linkage disequilibrium in a long-range. To investigate this in detail on experimental data is reserved for future perspectives. One constraint for the *tLD* is, similar to the LR_{T_3}-test, that its reliability is dependent on how well the estimation of tree topology is. As we have seen in section 3.2.1, chapter 3, the assignment of the clusters agrees very well to the true one, given that the correct cluster size was estimated. Whilst for the T_3-test preference is given to the balanced tree in not clearly resolvable cases for the test being conservative, for *tLD* this factor does not need to be taken into account. Furthermore in contrast to the LR_{T_3}-test, one only needs to consider the first 'clustering step'; namely at the root of the tree (MRCA) dividing the sample into the two cluster. As such, in this case it might be less complex (compared to the case of the LR_{T_3}-test) to establish a more suitable clustering method for the use of *tLD*. One might take the cluster assignment of neighbouring windows into account, when estimating the actual tree topology. We propose that as a further future project.

Appendix A

A.1 Derivation of test statistic T_3

In the following, we re-capitulate from (Li and Wiehe, 2013) how the test statistic T_3 was derived. Here, we will provide a somewhat more detailed derivation for the formulas.

Let $p(n, \omega_1) := \text{Prob}(\Omega_1 = \omega_1) = \frac{2-\delta_{\omega_1, n/2}}{n-1}$, where $\delta_{.,.}$ denotes the Kronecker symbol. We will show the calculations for n even. (Same approach, if n uneven).

By applying the formula

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

in the third line, one can derive the expectation

$$\begin{aligned}
E[\Omega_1] &= \sum_{\omega_1=1}^{n/2} \omega_1 p(n, \omega_1) \\
&= 1 \cdot \frac{2}{n-1} + 2 \cdot \frac{2}{n-1} + ... + \left(\frac{n}{2} - 1\right) \cdot \frac{2}{n-1} + \frac{n}{2} \cdot \frac{1}{n-1} \\
&= \frac{2}{n-1} \left(\sum_{k=1}^{\frac{n-2}{2}} k \right) + \frac{n}{2} \cdot \frac{1}{n-1} \\
&= \frac{2}{n-1} \cdot \frac{n^2 - 2n}{8} + \frac{n}{2} \cdot \frac{1}{n-1} \\
&= \frac{1}{n-1} \left(\frac{n^2 - 2n + 2n}{4} \right) \\
&= \frac{n^2}{4(n-1)} \approx \frac{n}{4}.
\end{aligned}$$

The variance is then calculated like following.
By applying the formula

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

in the third line, one gets

$$
\begin{aligned}
V[\Omega_1] &= \sum_{\omega_1=1}^{n/2} \omega^2 p(n,\omega_1) - (E[\Omega_1])^2 \\
&= \left(1^2 \cdot \frac{2}{n-1} + 2^2 \cdot \frac{2}{n-1} + ... + \left(\frac{n}{2}-1\right)^2 \cdot \frac{2}{n-1} + \left(\frac{n}{2}\right)^2 \cdot \frac{1}{n-1}\right) - \left(\frac{n^2}{4(n-1)}\right)^2 \\
&= \left(\frac{2}{n-1} \sum_{k=1}^{\frac{n-2}{2}} k^2 + \frac{n^2}{4} \cdot \frac{1}{n-1}\right) - \frac{n^4}{16(n-1)^2} \\
&= \left(\frac{2}{n-1}\left(\frac{1}{6}\left(\frac{n-2}{2}\left(\frac{n-2}{2}+1\right)\left(\frac{2(n-2)}{2}+1\right)\right)\right) + \frac{n^2}{4} \cdot \frac{1}{n-1}\right) - \frac{n^4}{16(n-1)^2} \\
&= ... = \frac{n^4 - 4n^3 + 8n^2 - 8n}{48(n-1)^2} \\
&= \frac{(n^2-2n)(4+n^2-2n)}{48(n-1)^2} \\
&= \frac{(n-2)n(4+(n-2)n)}{48(n-1)^2} \approx \frac{n^2}{48}
\end{aligned}
$$

And the standard variation is the square root of the variance

$$\sigma(\Omega_1) \approx \frac{n}{2\sqrt{12}}.$$

Note that Ω_i depends on Ω_j, $j = 1, ..., i-1$, $n_i = n - \omega_1 - \omega_2 - ...\omega_{i-1}$.
In a similar calculation like for $E[\Omega_1]$, one gets for $E[\Omega_2]$

$$
\begin{aligned}
E[\Omega_2] &= \sum_{\omega_1=1}^{\frac{n}{2}} p(n_1,\omega_1) \sum_{\omega_2=1}^{\frac{n_2}{2}} \omega_2 p(n_2,\omega_2) \\
&= \sum_{\omega_1=1}^{\frac{n}{2}} p(n_1,\omega_1) \frac{n_2^2}{4(n_2-1)} \\
&\approx \sum_{\omega_1=1}^{\frac{n}{2}} p(n_1,\omega_1) \frac{n_2}{4} \\
&= \sum_{\omega_1=1}^{\frac{n}{2}} p(n_1,\omega_1) \frac{(n-\omega_1)}{4} \\
&= \frac{n(3n-4)}{16(n-1)} \approx \frac{3n}{16} = \frac{3^1 n}{4^2}.
\end{aligned}
$$

In a similar way by evaluating sums iteratively one gets
$E[\Omega_3] \approx \frac{9n}{64} = \frac{3^2 n}{4^3}$, $E[\Omega_4] \approx \frac{3^3 n}{4^4}$, etc, and hence

$$E[\Omega_i] \approx \frac{3^{i-1} n}{4^i}.$$

With similar calculations, it follows

$$V[\Omega_i] \approx \frac{1}{3}\left(1 - \frac{3^{i-1}n}{4^1}\right)^2.$$

Let now $\Omega_1^* := 2\Omega_1/n$ be the nomalised random variables.
Since n is constant, it can be easily deduced that

$$E[\Omega_1^*] = E[2\Omega_1/n] = \frac{2}{n}E[\Omega_1] \approx \frac{1}{2}$$

$$V[\Omega_1^*] = V[2\Omega_1/n] = \left(\frac{2}{n}\right)^2 V[\Omega_1] \approx \frac{1}{12}$$

and hence $\sigma(\Omega_1) \approx \sqrt{\frac{1}{12}}$.
Furthermore, it holds that

$$E[\Omega_i^*] \approx \frac{E[2\Omega_i]}{E[n_i]} = \frac{1}{2},$$

with (by using the geometric series)

$$\begin{aligned} E[n_i] &= E[n - \Omega_1 - ... - \Omega_{i-1}] \\ &\approx n - \frac{3^0 n}{4^1} - ... - \frac{3^{i-2}n}{4^{i-1}} \\ &= n\left(1 - \sum_{k=0}^{i-2}\left(\frac{3^k}{4^{k+1}}\right)\right) \\ &= n\left(1 - \frac{1}{4}\left(\frac{1 - \frac{3}{4}^{i-1}}{1 - \frac{3}{4}}\right)\right) = n\left(\frac{3}{4}\right)^{i-1}. \end{aligned}$$

Similar calculations give

$$V[\Omega_i^*] = \frac{1}{12} + \frac{1}{n^2}\left(\frac{4}{3}\right)^{2i} - \frac{2}{3n}\left(\frac{4}{3}^{i-1}\right) \approx \frac{1}{12}.$$

and hence

$$\sigma[\Omega_i^*] = \sqrt{\frac{1}{12}}.$$

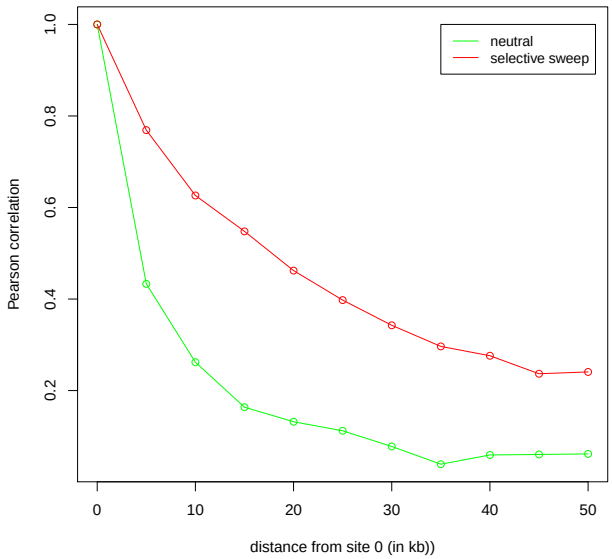
A key result from probability theory is the *central limit theorem*, which states that the sum of continuous uniforms converges in distribution to a normal random variable. Hence, applying this and substitute the expectation and standard variation by

previously results, we arrive at

$$\mathcal{N}(0,1) \sim \sqrt{\frac{1}{k}} \cdot \sum_{i=1}^{k} \frac{(\Omega_i^* - E(\Omega_i^*))}{\sigma(\Omega_i^*)} = \sqrt{\frac{12}{k}} \cdot \sum_{i=1}^{k} \left(\Omega_i^* - \frac{1}{2}\right) =: T_k.$$

# (segregating sites)	average $\hat{\Omega}_1^*$	average $\hat{\Omega}_2^*$	average $\hat{\Omega}_3^*$
1	0.21316	-	-
2	0.75999	0.3134	-
3	0.68913	0.3597	0.1372
4	0.61238	0.6035	0.2365
5	0.59287	0.6686	0.3259
6	0.58012	0.6664	0.4132
7	0.58306	0.6296	0.48613
8	0.57957	0.6189	0.5528
9	0.57788	0.5962	0.5959
10	0.56375	0.5722	0.628
12	0.56263	0.58	0.6151
15	0.56933	0.57529	0.5611
20	0.54727	0.56425	0.5752
30	0.54699	0.5569	0.5485
40	0.54251	0.5468	0.5436

TABLE A.1: Average $\hat{\Omega}_1^*$, $\hat{\Omega}_2^*$, $\hat{\Omega}_3^*$ out of $1,000$ runs for each scenario, conditioned on the number of segregating sites used for estimating $\hat{\Omega}_1^*$. For illustration see FIGURE 3.7



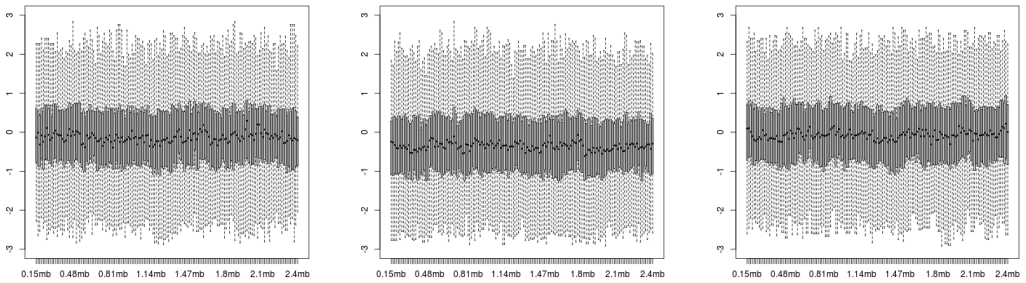
.

FIGURE A.1: Correlation based on simulations of the test statistic T_3 of the true tree. Pearson's correlation coefficient is measured between pairs of T_3-values of trees at position 0 and a position x kb distance away from position 0. In the selected sweep scenario, position 0 refers to the position of the selected site. Average of 1,000 runs.

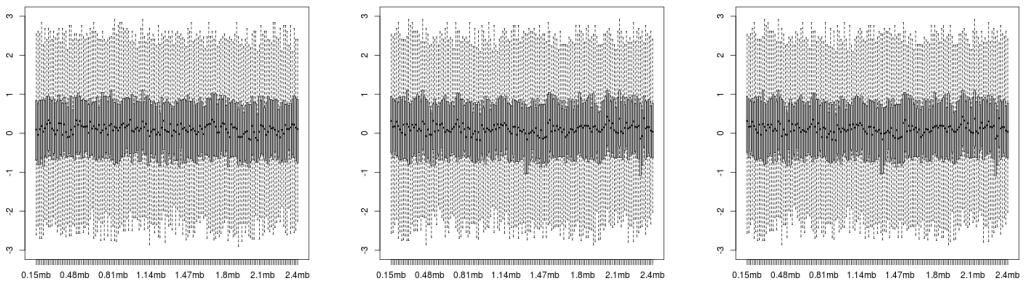
A.2 T_3-distribution along chromosome: Migration events

Sampling scheme: $n_1 = 200$, $n_2 = 0$

T_3, known tree topology:



T_3, estimated tree topology:

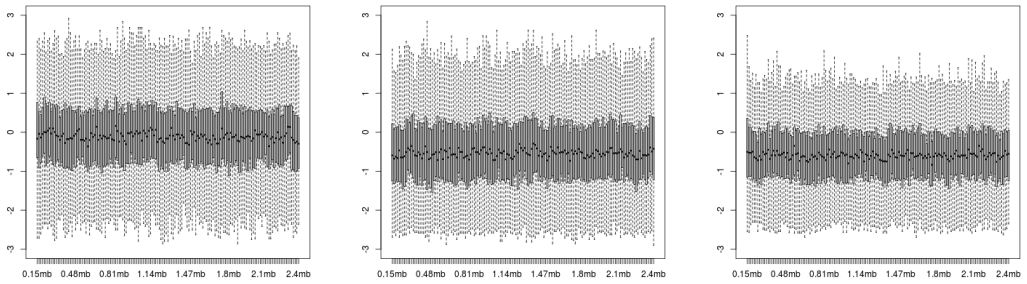


.

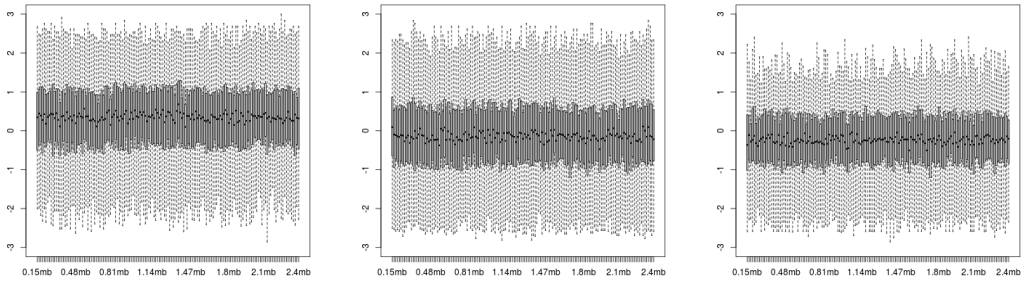
FIGURE A.2: Distribution of T_3 along chromosome. Admixed population. Sample size of sub-population $n_1 = 200$ and $n_2 = 0$. Result from 1000 simulation runs. Populations simulated with *ms*, parameters see section 3.3.
Left: $4Nm = 4$. Middle: $4Nm = 0.4$. Right: $4Nm = 0.04$.

Sampling scheme: $n_1 = 180$**,** $n_2 = 20$

T_3, known tree topology:



T_3, estimated tree topology:

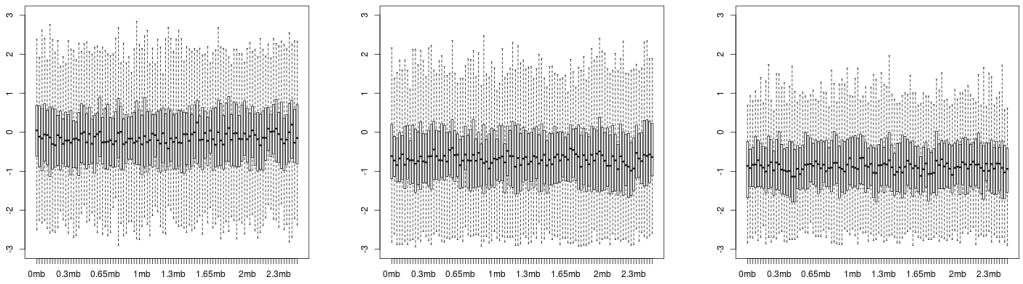


.

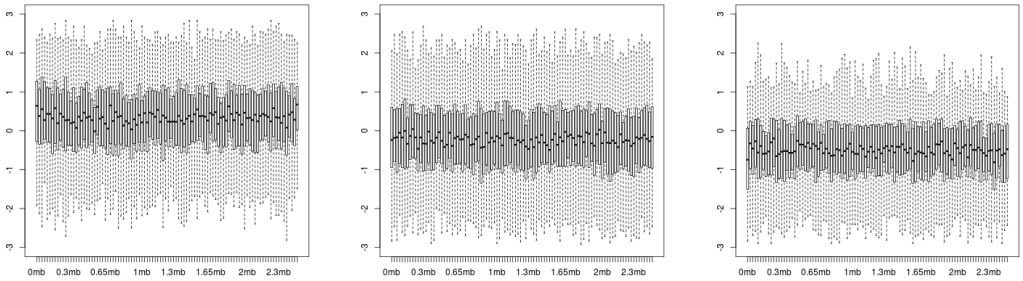
FIGURE A.3: Distribution of T_3 along chromosome. Admixed population. Sample size of sub-population $n_1 = 180$ and $n_2 = 20$. Result from 1000 simulation runs. Populations simulated with *ms*, parameters see section 3.3.
Left: $4Nm = 4$. Middle: $4Nm = 0.4$. Right: $4Nm = 0.04$.

Sampling scheme: $n_1 = 195$, $n_2 = 5$

T_3, known tree topology:



T_3, estimated tree topology:



.

FIGURE A.4: Distribution of T_3 along chromosome. Admixed population. Sample size of sub-population $n_1 = 195$ and $n_2 = 5$. Result from 1000 simulation runs. Populations simulated with *ms*, parameters see section 3.3.
Left: $4Nm = 4$. Middle: $4Nm = 0.4$. Right: $4Nm = 0.04$.

A.3 LR_{T_3}: Migration event $n_1 = 180$ and $n_2 = 20$

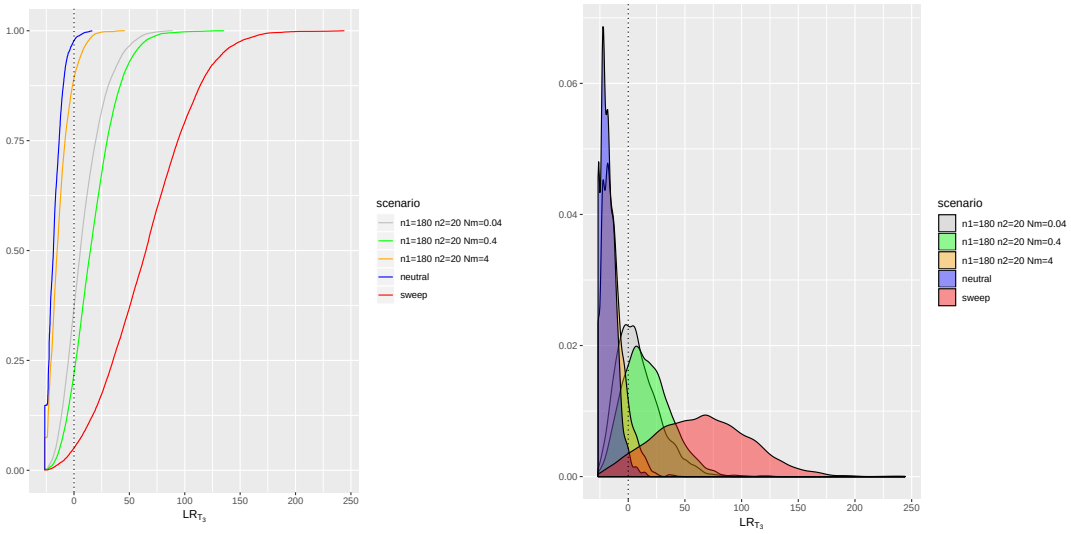


FIGURE A.5: On the left side: cumulative distribution of LR_{T_3}. On the right side: Density plot of LR_{T_3}.

LR_{T_3}: Migration event $n_1 = 195$ and $n_2 = 5$

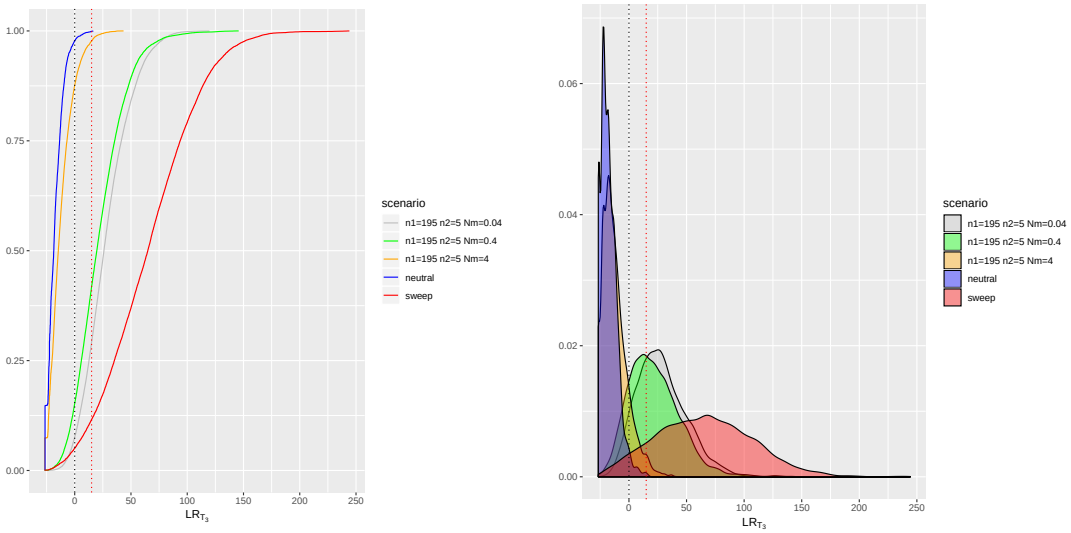


FIGURE A.6: On the left side: cumulative distribution of LR_{T_3}. On the right side: Density plot of LR_{T_3}.

Appendix B

B.1 Analysis of candidate regions

POP	Total	chromosome																						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X
ACB	210	16	18	10	21	19	8	11	12	9	11	6	11	7	4	1	9	2	4	6	2	1	5	17
ASW	203	16	14	12	15	6	15	10	9	12	9	10	17	5	3	1	11	4	5	0	4	2	3	20
ESN	212	24	17	13	17	8	13	9	18	16	3	7	15	4	7	6	5	4	3	5	2	0	3	13
GWD	234	13	20	14	27	17	11	14	14	10	10	10	7	5	4	0	8	4	6	8	7	2	3	20
LWK	202	16	17	5	15	14	18	9	5	11	4	11	19	7	5	3	5	2	3	4	6	0	9	14
MSL	200	19	16	12	17	13	9	9	9	11	9	4	11	4	4	2	5	9	3	5	3	0	7	19
YRI	240	16	15	17	19	16	18	17	16	14	9	4	13	4	4	4	6	10	5	6	7	0	3	17
CEU	413	36	36	22	33	27	28	19	22	21	22	20	25	14	9	14	13	6	11	5	5	5	3	17
FIN	428	26	37	30	34	26	32	22	25	16	17	23	30	13	10	18	11	5	9	5	8	7	4	20
GBR	405	39	40	35	27	26	33	18	18	18	15	18	24	11	8	9	12	9	12	5	6	4	5	13
IBS	429	36	36	24	37	24	27	22	26	18	21	20	21	16	15	13	12	10	10	5	7	4	3	22
TSI	433	31	43	29	40	31	25	23	19	12	10	29	12	20	10	20	10	10	8	6	13	6	3	23
CDX	421	34	52	29	23	26	16	21	28	19	20	22	21	17	9	14	5	10	14	3	4	8	3	23
CHB	378	24	31	30	24	29	18	27	19	13	19	20	22	15	10	11	6	9	12	6	9	4	1	19
CHS	440	33	44	31	21	33	28	25	30	19	21	22	18	14	12	14	7	12	8	5	9	10	7	17
JPT	440	39	45	37	23	26	23	23	19	12	16	25	28	16	10	21	8	13	12	5	7	5	5	22
KHV	406	26	36	31	24	24	27	34	22	17	14	22	15	12	13	8	8	10	15	5	9	7	6	21
BEB	405	25	37	25	31	18	28	25	23	17	17	25	22	16	12	13	8	8	14	8	9	4	2	18
GIH	409	28	45	39	29	23	22	24	28	18	18	16	23	11	11	10	4	7	11	7	10	5	4	16
ITU	378	26	31	31	33	24	23	26	19	15	19	26	14	11	10	8	4	7	9	6	6	5	2	23
PJL	409	30	36	29	31	22	31	25	29	16	13	19	18	12	14	12	8	10	9	9	6	4	4	22
STU	399	33	33	31	34	25	28	20	21	15	16	23	22	13	12	10	5	9	9	6	3	2	2	27
CLM	348	29	28	26	18	20	23	22	17	13	11	17	25	7	10	11	15	6	5	7	7	4	2	25
MXL	349	26	24	29	28	24	20	27	16	12	14	22	19	11	14	7	7	3	11	6	5	5	2	17
PEL	367	34	27	26	30	23	24	17	21	12	16	19	19	16	15	8	4	12	5	3	9	4	4	19
PUR	326	27	24	31	20	13	13	18	18	18	12	17	14	12	8	3	16	7	12	9	7	7	1	19

FIGURE B.1: Numbers of candidate regions found on each chromosome and in each population. Regions span between 55 kb and 785 kb.

Population	Total number of candidate genes	Private to population
ACB	619	51
ASW	657	52
ESN	679	86
GWD	739	156
LWK	533	89
MSL	520	87
YRI	728	98
CEU	1348	217
FIN	1442	296
GBR	1287	117
IBS	1369	141
TSI	1392	133
CDX	1122	209
CHB	1243	158
CHS	1257	189
JPT	1278	329
KHV	1185	160
BEB	1159	105
GIH	1288	162
ITU	1153	101
PJL	1245	94
STU	1181	104
CLM	1001	133
MXL	1124	143
PEL	1163	240
PUR	1037	244

TABLE B.1: Numbers of all genes identified with $LR_{T_3} \geq 15$ per population.

Population	Total number of protein-coding genes	Private to population
ACB	263	21
ASW	266	21
ESN	309	40
GWD	323	63
LWK	217	24
MSL	229	34
YRI	338	42
CEU	593	109
FIN	558	96
GBR	594	49
IBS	564	56
TSI	568	47
CDX	426	71
CHB	559	75
CHS	498	65
JPT	514	126
KHV	490	57
BEB	487	40
GIH	541	57
ITU	494	40
PJL	558	38
STU	487	33
CLM	438	50
MXL	451	55
PEL	461	91
PUR	468	113

TABLE B.2: Numbers of protein-coding genes identified with $LR_{T_3} \geq 15$ per population.

B.2 Top candidates ($LR_{T_3} > 200$), previously known candidates

In the following table, we show an overview of which of the protein-coding genes found in our 'Top Regions' (LR_{T_3}-score > 200 in TABLE 4.3) were previously mentioned in other studies (column 5). The comparison was done using **dbPSHP** (Li et al., 2014a) and a more recent candidate gene list from Schrider and Kern, 2017. Whilst in TABLE 4.3 the population possessing the $LR_{T_3} > 200$ is listed (underlined in column 5), this table also shows when these genes were candidates for other populations.

CHR	Position	Gene	Found in populations using LR_{T_3}	Found in other studies
chr8	10,983,980-10,987,745	AF131215.5	CDX, CHB, CHS, JPT, KHV	(Schrider and Kern, 2017)
chr10	75,134,859-75,173,834	ANXA7	FIN, GBR, IBS, MXL, TSI	(Carlson et al., 2005), (Kelley et al., 2006), (Cai et al., 2011), (Mendizabal et al., 2012), (Liu et al., 2013)
chr4	106,473,777-106,629,250	ARHGEF38	BEB, CDX, FIN, GBR, GIH, KHV, PJL	(Zhang et al., 2006), (Oleksyk et al., 2008), (Grossman et al., 2013)
chr14	67,761,088-67,826,982	ATP6V1D	CEU, FIN, GBR, IBS, TSI	(Oleksyk et al., 2008), (Han and Abney, 2013), (Wagh et al., 2012), (Liu et al., 2013), (Schrider and Kern, 2017)
chr4	42,112,955-42,154,895	BEND4	CDX, CHB, CHS, GIH, JPT, KHV, STU, TSI	(Barreiro et al., 2008), (Lappalainen et al., 2010), (Grossman et al., 2010), (Grossman et al., 2013), (Liu et al., 2013)
chr1	51,567,906-51,613,752	C1orf185	BEB, CEU, CLM, FIN, GBR, GIH, ITU, PEL, PJL, PUR	(Higasa et al., 2009), (Liu et al., 2013)
chr2	109,403,213-109,501,933	CCDC138	CDX, CHB, CHS, KHV	(Grossman et al., 2010), (Liu et al., 2013)
chr8	42,607,763-42,651,535	CHRNA6	CEU, GIH, IBS, PUR, TSI	(Oleksyk et al., 2008)
chr15	64,199,235-64,364,232	DAPK2	CDX, CHB, CHS, JPT, KHV	(Carlson et al., 2005), (Williamson et al., 2007), (Tang, Thornton, and Stoneking, 2007), (Higasa et al., 2009), (Lopez Herraez et al., 2009), (Cai et al., 2011), (Liu et al., 2013)
chr1	100,652,475-100,715,390	DBT	BEB, FIN, GIH, IBS, ITU, MXL, PEL	(Kelley et al., 2006)
chr4	41,983,713-41,988,476	DCAF4L1	CDX, CHB, CHS, FIN, GIH, ITU, JPT, KHV, MXL, STU, TSI	(Barreiro et al., 2008), (Grossman et al., 2013), (Liu et al., 2013), (Schrider and Kern, 2017)
chr10	75,007,118-75,036,742	DNAJC9	FIN, GBR, IBS, MXL, TSI	(Carlson et al., 2005), (Kelley et al., 2006), (Williamson et al., 2007), (Mendizabal et al., 2012), (Liu et al., 2013)
chr10	74,889,913-74,928,813	ECD	FIN, GBR, IBS, MXL, TSI	(Oleksyk et al., 2008), (Cai et al., 2011), (Mendizabal et al., 2012), (Liu et al., 2013)

chr2	109,510,927-109,605,828	EDAR	CDX, CHB, CHS, JPT, KHV	(Akey et al., 2002), (Carlson et al., 2005), (Kelley et al., 2006), (Williamson et al., 2007), (Tang, Thornton, and Stoneking, 2007), (Frazer et al., 2007), (Sabeti et al., 2007), (Fujimoto et al., 2008), (Barreiro et al., 2008), (Bryk et al., 2008), (Lopez Herraez et al., 2009), (Grossman et al., 2010), (Chun and Fay, 2011), (Peter, Huerta-Sanchez, and Nielsen, 2012), (Kamberov et al., 2013), (Grossman et al., 2013), (Liu et al., 2013), (Tan et al., 2013), (Hider et al., 2013)
chr14	67,826,714-67,853,233	EIF2S1	CEU, FIN, GBR, IBS, TSI	(Oleksyk et al., 2008), (Han and Abney, 2013), (Wagh et al., 2012), (Liu et al., 2013)
chr1	51,819,935-51,985,000	EPS15	ASW, BEB, CEU, CLM, FIN, GBR, GIH, ITU, MXL, PEL, PJL, PUR, STU	(Han and Abney, 2013), (Liu et al., 2013)
chr10	74,927,924-75,004,262	FAM149B1	FIN, GBR, IBS, MXL, TSI	(Cai et al., 2011), (Mendizabal et al., 2012), (Liu et al., 2013)
chr14	67,656,110-67,695,267	FAM71D	CEU, FIN, GBR, IBS, TSI	(Oleksyk et al., 2008), (Wagh et al., 2012), (Liu et al., 2013)
chr15	63,889,552-63,894,627	FBXL22	CDX, CHB, CHS, JPT, KHV	(Carlson et al., 2005), (Barreiro et al., 2008), (Higasa et al., 2009), (Lopez Herraez et al., 2009), (Cai et al., 2011), (Liu et al., 2013), (Karlsson et al., 2013)
chr8	42,889,337-42,940,931	FNTA	ASW, CEU, GBR, GIH, IBS, ITU, PJL, PUR, TSI, YRI	(Oleksyk et al., 2008)
chr2	109,065,017-109,125,871	GCC2	CDX, CHB, CHS, KHV	(Carlson et al., 2005), (Kelley et al., 2006), (Frazer et al., 2007), (Sabeti et al., 2007), (Barreiro et al., 2008), (Kudaravalli et al., 2009), (Grossman et al., 2010), (Grossman et al., 2013), (Liu et al., 2013)
chr14	66,974,125-67,648,520	GPHN	CEU, FIN, GBR, IBS, TSI	(Oleksyk et al., 2008), (Liu et al., 2013)
chr4	106,629,935-106,768,885	GSTCD	BEB, CDX, FIN, GBR, GIH, KHV, PJL	(Barreiro et al., 2008), (Grossman et al., 2010), (Liu et al., 2013), (Karlsson et al., 2013)
chr15	63,900,817-64,126,141	HERC1	CDX, CHB, CHS, JPT, KHV	(Carlson et al., 2005), (Kelley et al., 2006), (Williamson et al., 2007), (Tang, Thornton, and Stoneking, 2007), (Sabeti et al., 2007), (Barreiro et al., 2008), (Higasa et al., 2009), (Grossman et al., 2010), (Cai et al., 2011), (Waldman et al., 2011), (Grossman et al., 2013), (Liu et al., 2013), (Karlsson et al., 2013)

chr8	42,752,075-42,885,682	HOOK3	ASW, CEU, GBR, GIH, IBS, ITU, PJL, PUR, TSI, YRI	(Oleksyk et al., 2008)
chr4	106,603,784-106,817,143	INTS12	BEB, CDX, FIN, GBR, GIH, KHV, PJL	(Lopez Herraez et al., 2009), (Grossman et al., 2010), (Liu et al., 2013), (Karlsson et al., 2013)
chr4	41,361,624-41,702,061	LIMCH1	CDX, CHS	(Barreiro et al., 2008), (Higasa et al., 2009), (Lopez Herraez et al., 2009), (Mizuno et al., 2010)
chr2	109,150,857-109,303,702	LIMS1	CDX, CHB, CHS, KHV	(Carlson et al., 2005), (Frazer et al., 2007), (Sabeti et al., 2007), (Barreiro et al., 2008), (Grossman et al., 2010), (Zhong et al., 2010), (Grossman et al., 2013), (Liu et al., 2013)
chr3	154,741,913-154,901,497	MME	BEB, CHB, CLM, FIN, GBR, GIH, IBS, MXL, PEL, PJL, PUR, TSI	(Schrider and Kern, 2017)
chr14	67,707,826-67,802,536	MPP5	CEU, FIN, GBR, IBS, TSI	(Oleksyk et al., 2008), (Wagh et al., 2012), (Liu et al., 2013), (Schrider and Kern, 2017)
chr10	75,391,412-75,401,515	MYOZ1	FIN, GBR, IBS, MXL, TSI	(Grossman et al., 2013), (Liu et al., 2013)
chr12	44,902,058-45,315,631	NELL2	BEB, FIN, GIH, ITU, MXL, PEL, PJL, STU	(Oleksyk et al., 2008), (Lopez Herraez et al., 2009), (Chen, Patterson, and Reich, 2010), (Wagh et al., 2012), (Liu et al., 2013)
chr5	43,602,794-43,707,507	NNT	ACB, BEB, CEU, CLM, FIN, GBR, GIH, IBS, ITU, MXL, PJL, STU, TSI	(Mendizabal et al., 2012), (Wagh et al., 2012), (Liu et al., 2013)
chr10	55,562,531-57,387,702	PCDH15	CDX, CHB, CHS, JPT	(Williamson et al., 2007), (Frazer et al., 2007), (Sabeti et al., 2007), (Barreiro et al., 2008), (Grossman et al., 2010), (Zhong et al., 2010), (Chun and Fay, 2011), (Grossman et al., 2013), (Liu et al., 2013), (Schrider and Kern, 2017)
chr4	41,746,099-41,750,987	PHOX2B	CDX, CHB, FIN, ITU	(Higasa et al., 2009), (Lopez Herraez et al., 2009)
chr14	67,853,700-67,878,917	PLEK2	CEU, FIN, GBR, IBS, TSI	(Oleksyk et al., 2008), (Lopez Herraez et al., 2009), (Han and Abney, 2013), (Wagh et al., 2012), (Liu et al., 2013)
chr10	75,196,186-75,255,782	PPP3CB	FIN, GBR, GWD, IBS, MXL, TSI	(Carlson et al., 2005), (Kelley et al., 2006), (Cai et al., 2011), (Mendizabal et al., 2012), (Liu et al., 2013)
chr2	109,335,937-109,402,267	RANBP2	CDX, CHB, CHS, KHV	(Carlson et al., 2005), (Kelley et al., 2006), (Tang, Thornton, and Stoneking, 2007), (Frazer et al., 2007), (Sabeti et al., 2007), (Grossman et al., 2010), (Liu et al., 2013)

chr1	51,701,943-51,739,127	RNF11	BEB, CEU, CLM, FIN, GBR, GIH, ITU, MXL, PEL, PJL, PUR, STU	(Storz, Payseur, and Nachman, 2004), (Oleksyk et al., 2008), (Grossman et al., 2013), (Liu et al., 2013), (Schrider and Kern, 2017)
chr1	100,731,763-100,758,325	RTCA	FIN, IBS, MXL	(Higasa et al., 2009), (Liu et al., 2013)
chr4	41,992,489-42,092,474	SLC30A9	CDX, CHB, CHS, FIN, GIH, ITU, JPT, KHV, MXL, STU, TSI	(Carlson et al., 2005), (Kelley et al., 2006), (Williamson et al., 2007), (Frazer et al., 2007), (Sabeti et al., 2007), (Barreiro et al., 2008), (Higasa et al., 2009), (Lappalainen et al., 2010), (Lopez Herraez et al., 2009), (Grossman et al., 2010), (Chen, Patterson, and Reich, 2010), (Grossman et al., 2013), (Liu et al., 2013), (Karlsson et al., 2013), (Schrider and Kern, 2017)
chr2	108,905,095-108,926,371	SULT1C2	CDX, CHB, CHS	(Carlson et al., 2005), (Kelley et al., 2006), (Frazer et al., 2007), (Barreiro et al., 2008), (Lopez Herraez et al., 2009), (Grossman et al., 2013), (Liu et al., 2013)
chr2	108,994,367-109,004,513	SULT1C4	CDX, CHB, CHS, KHV	(Lopez Herraez et al., 2009), (Grossman et al., 2013), (Liu et al., 2013)
chr10	75,404,639-75,423,561	SYNPO2L	FIN, GBR, IBS, MXL, TSI	(Grossman et al., 2010), (Grossman et al., 2013), (Liu et al., 2013)
chr8	42,691,817-42,698,468	THAP1	ASW, CEU, GBR, GIH, IBS, ITU, PUR, TSI, YRI	(Oleksyk et al., 2008)
chr12	44,229,770-44,783,545	TMEM117	BEB, CDX, CEU, CHB, CHS, CLM, FIN, GBR, GIH, ITU, JPT, KHV, MXL, PEL, PJL, STU, TSI	(Barreiro et al., 2008), (Lopez Herraez et al., 2009), (Grossman et al., 2010), (Zhong et al., 2010), (Grossman et al., 2013), (Liu et al., 2013), (Karlsson et al., 2013)
chr4	41,937,137-41,962,589	TMEM33	CDX, CHB, CHS, FIN, GIH, ITU, JPT, KHV, MXL, STU, TSI	(Carlson et al., 2005), (Williamson et al., 2007), (Barreiro et al., 2008), (Higasa et al., 2009), (Lappalainen et al., 2010), (Grossman et al., 2010), (Zhong et al., 2010), (Grossman et al., 2013), (Liu et al., 2013), (Karlsson et al., 2013), (Hider et al., 2013), (Schrider and Kern, 2017)
chr10	75,013,517-75,118,617	TTC18	FIN, GBR, IBS, MXL, TSI	(Carlson et al., 2005), (Kelley et al., 2006), (Williamson et al., 2007), (Cai et al., 2011), (Mendizabal et al., 2012), (Liu et al., 2013)
chr1	51,752,930-51,810,788	TTC39A	BEB, CEU, CLM, FIN, GBR, GIH, ITU, MXL, PEL, PJL, PUR, STU	(Liu et al., 2013)

chr15	63,796,793-63,886,839	USP3	CDX, CHB, CHS, JPT, KHV	(Carlson et al., 2005), (Barreiro et al., 2008), (Higasa et al., 2009), (Lopez Herraez et al., 2009), (Cai et al., 2011), (Liu et al., 2013), (Karlsson et al., 2013)
chr10	75,257,296-75,385,711	USP54	FIN, GBR, GWD, IBS, MXL, TSI	(Carlson et al., 2005), (Kelley et al., 2006), (Mendizabal et al., 2012), (Liu et al., 2013)
chr8	10,753,555-11,058,875	XKR6	CDX, CHB, CHS, JPT, KHV	(Barreiro et al., 2008), (Johansson and Gyllensten, 2008), (Lopez Herraez et al., 2009), (Chen, Patterson, and Reich, 2010), (Wagh et al., 2012), (Liu et al., 2013), (Schrider and Kern, 2017)

B.3 Top ten candidate regions per population

In the following we provide a list containing the 'Top Ten candidate region' for each 26 population and respective genes. To extract the genes, we additionaly expand significant regions with 25kb on each side (shown here). Overlapping regions are put together to one region. To extract the genes, we used the `R` package `biomaRt` (Smedley et al., 2015). We used the coordinates for human genome build hg19 for our data, to which phase 3 of the 1,000 Genomes Project is mapped.

Top ten candidate regions for population ACB

ACB						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr12	87287384	87744884	288.854		RPL23AP68	457.5
chr20	20387585	20787585	213.416	RALGAPA2	EIF4E2P1, RP11-23O13.1, RN7SL607P	400
chr4	107603887	107961387	134.471	DKK2	ACTR6P1	357.5
chr13	52797838	53337838	105.848	THSD1, VPS36, CKAP2, HNRNPA1L2, SUGT1, LECT1	RP11-248G5.8, TPTE2P2, RP13-444H2.1, RNY4P24, LINC00345, RP11-78J21.4, TPTE2P3, MRPS31P4	540
chrX	41326170	41821170	102.353	NYX, CASK, GPR34, GPR82	RP1-169I5.4, CASK-AS1, RNU6-1321P, RN7SL406P, RP11-204C16.4, RN7SL144P, RP5-1174J21.2, RP5-1174J21.1, RNU6-202P	495
chr12	113512384	113729884	99.7771	DTX1, RASAL1, CCDC42B, DDX54, RITA1, IQCD, TPCN1	Y_RNA, AC089999.1, Y_RNA, RP11-545P7.4	217.5
chr5	15345539	15553039	97.9321	FBXL7	MARK2P5, CTD-2313D3.1	207.5
chr17	26268542	26558542	93.5331	NLK	RP11-218F4.1, SCARNA20, RP11-218F4.2, SNORA70, Vault, RPS29P22, AC100852.2, AC100852.1, AC061975.9, AC061975.1, CTD-2008P7.10, AC061975.7, PYY2	290
chr12	113764884	113872384	93.7354	SLC8B1, PLBD2, SDS, SDSL	NONP	107.5
chr12	88009884	88192384	89.801		RP11-248E9.1, CYCSP30, RP11-248E9.4, MKRN9P, RP11-248E9.5	182.5

Top ten candidate regions for population ASW

ASW						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr2	194653645	195356145	200.98		RP11-764E7.1, AC068135.1, GLULP6, HNRNPA1P47, AC018799.1, AC106883.1	702.5
chr12	113512384	113934884	177.352	DTX1, RASAL1, CCDC42B, DDX54, RITA1, IQCD, TPCN1, SLC8B1, PLBD2, SDS, SDSL, LHX5	Y_RNA, AC089999.1, Y_RNA, RP11-545P7.4, RP11-82C23.2	422.5
chr20	20390201	20807701	159.512	RALGAPA2	EIF4E2P1, RP11-23O13.1, RN7SL607P	417.5
chr10	134231660	134629160	144.211	C10orf91, INPP5A, NKX6-2, TTC40	RP11-432J24.2, RP11-432J24.3, RP11-432J24.5, LINC01165, RP11-288G11.3	397.5
chr8	36018715	36383715	121.145		RN7SKP201, RP11-593P24.2, MTND6P19, RP11-593P24.3, RP11-139F9.1, RNU6-533P, RP11-593P24.4	365
chr12	87534884	87674884	119.521		RPL23AP68	140
chr9	102258041	102578041	106.81		RP11-547C13.1, RP11-554F20.1	320
chr8	37363715	37886215	105.415	ZNF703, RP11-863K10.7, ERLIN2, PROSC, GPR124, BRF2, RAB11FIP1, GOT1L1, ADRB3	RP11-150O12.1, RP11-150O12.6, RP11-150O12.5, RP11-150O12.3, RP11-150O12.4, RP11-346L1.2, RNU6-607P, RP11-863K10.2, RP11-863K10.4, RN7SL709P, AC144573.1, KB-1836B5.3	522.5
chr5	15323039	15553039	96.7048	FBXL7	MARK2P5, CTD-2313D3.1	230
chr13	53075338	53337838	92.4284	HNRNPA1L2, SUGT1, LECT1	TPTE2P3, MRPS31P4	262.5

Top ten candidate regions for population ESN

ESN						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr12	87274884	87754884	277.14		RPL23AP68	480
chr13	52772838	53332838	149.988	THSD1, VPS36, CKAP2, HNRNPA1L2, SUGT1, LECT1	RP11-248G5.8, TPTE2P2, RP13-444H2.1, RNY4P24, LINC00345, RP11-78J21.4, TPTE2P3, MRPS31P4	560
chr4	46385167	46765167	144.487	GABRA2, COX7B2	RP11-436F23.1, RNU6-412P, RAC1P2	380
chr4	107602667	107957667	137.085	DKK2	ACTR6P1	355
chr4	87387667	87637667	127.596	MAPK10, PTPN13	MIR4452	250
chr16	22921947	23274447	115.999	HS3ST2, USP31, SCNN1G	RP11-20G6.2, RP11-20G6.3, CTC-391G2.1	352.5
chr2	31862995	32115495	106.936	MEMO1, DPY30	AL133247.3, AL133249.1, AL121652.2, KRT18P52, AL121652.3, AK2P2, RP11-1057B6.1	252.5
chrX	10928670	11163670	105.113	HCCS, ARHGAP6	RP11-120D5.1, Y_RNA	235
chr20	20390085	20787585	102.184	RALGAPA2	EIF4E2P1, RP11-23O13.1, RN7SL607P	397.5
chr12	113509884	113717384	97.5442	DTX1, RASAL1, CCDC42B, DDX54, RITA1, IQCD, TPCN1	Y_RNA, AC089999.1, Y_RNA, RP11-545P7.4	207.5

Top ten candidate regions for population GWD

GWD						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr12	87214884	87727384	199.606	MGAT4C	RP11-202H2.1, RPL23AP68	512.5
chr20	20387701	20775201	189.438	RALGAPA2	EIF4E2P1, RP11-23O13.1, RN7SL607P	387.5
chr7	44363084	44593084	134.381	CAMK2B, NUDCD3, NPC1L1	AC004453.8, RNU6-1097P, AC004938.5	230
chr7	44248084	44360584	110.847	YKT6, CAMK2B	NONP	112.5
chr12	113532384	113729884	97.3028	DTX1, RASAL1, CCDC42B, DDX54, RITA1, IQCD, TPCN1	Y_RNA, AC089999.1, Y_RNA, RP11-545P7.4	197.5
chr6	45095100	45470100	95.9692	SUPT3H, RUNX2	RP11-491H9.3, MIR586, RP1-244F24.1	375
chr3	51257726	51742726	94.0139	DOCK3, MANF, RBM15B, VPRBP, RAD54L2, TEX264, GRM2	RP11-89F17.5, RNU6ATAC29P, RNA5SP132	485
chr4	107602667	107852667	92.9184	DKK2	ACTR6P1	250
chr7	141205584	141460584	91.1882	AGK, KIAA1147, WEE2, SSBP1	RP11-744I24.3, RP11-744I24.2, RP5-894A10.2, RP5-894A10.6, WEE2-AS1, RNU1-82P	255
chr17	26321073	26568573	90.1225	NLK	SCARNA20, RP11-218F4.2, SNORA70, Vault, RPS29P22, AC100852.2, AC100852.1, AC061975.9, AC061975.1, CTD-2008P7.10, AC061975.7, PYY2, CTD-2008P7.9, AC061975.6	247.5

Top ten candidate regions for population LWK

LWK						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr12	87297384	87712384	251.482		RPL23AP68	415
chr4	87317667	87652667	115.258	MAPK10, PTPN13	MIR4452	335
chr4	107565167	107962667	108.496	DKK2	ACTR6P1	397.5
chr10	134319084	134601584	106.023	INPP5A, NKX6-2	RP11-432J24.5, LINC01165, RP11-288G11.3	282.5
chrX	41346170	41786170	90.2757	CASK, GPR34, GPR82	CASK-AS1, RNU6-1321P, RN7SL406P, RP11-204C16.4, RN7SL144P, RP5-1174J21.2, RP5-1174J21.1, RNU6-202P	440
chrX	51643670	51796170	89.2977	MAGED1, RP11-114H20.1	RP11-234P3.2, IPO7P1, RP11-234P3.4, TPMTP3	152.5
chr3	164787719	164972719	87.5841	SI, SLITRK3	Y_RNA, RP11-747D18.1, RP11-85M11.2	185
chr12	87219884	87274884	86.5546	MGAT4C	RP11-202H2.1	55
chrX	51933670	52121170	85.2387	MAGED4, RP11-363G10.2, XAGE2B	SNORA11D	187.5
chr4	148027667	148280167	85.2685		MIR548G	252.5

Top ten candidate regions for population MSL

MSL						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr12	87287384	87694884	258.439		RPL23AP68	407.5
chr4	107610167	107847667	130.319	DKK2	ACTR6P1	237.5
chr4	46367667	46760167	128.475	GABRA2, COX7B2	RP11-436F23.1, RNU6-412P, RAC1P2	392.5
chr4	107850167	107977667	122.342	DKK2	NONP	127.5
chr10	134341660	134629160	119.781	INPP5A, NKX6-2, TTC40	RP11-288G11.3	287.5
chr19	42644984	42832484	101.582	POU2F2, DEDD2, ZNF526, GSK3A, AC006486.9, AC006486.1, ERF, CIC, PAFAH1B3, PRR19, TMEM145, MEGF8	SNORD112, CTC-378H22.2, AC010247.1	187.5
chr19	42432484	42625984	97.6973	ARHGEF1, RABAC1, ATP1A3, GRIK5, ZNF574, POU2F2	CTB-59C6.3	193.5
chr12	87187384	87259884	96.6737	MGAT4C	RP11-202H2.1	72.5
chr16	22934413	23284413	94.2207	USP31, SCNN1G	RP11-20G6.2, RP11-20G6.3, CTC-391G2.1	350
chr3	51082725	51517725	93.8023	DOCK3, MANF, RBM15B, VPRBP	RP11-89F17.5	435

Top ten candidate regions for population YRI

YRI						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr4	107602667	107965167	168.048	DKK2	ACTR6P1	362.5
chr13	52732838	53332838	146.15	NEK3, THSD1, VPS36, CKAP2, HNRNPA1L2, SUGT1, LECT1	MRPS31P5, RP11-248G5.8, TPTE2P2, RP13-444H2.1, RNY4P24, LINC00345, RP11-78J21.4, TPTE2P3, MRPS31P4	600
chr16	22931947	23229447	134.081	USP31, SCNN1G	RP11-20G6.2, RP11-20G6.3, CTC-391G2.1	297.5
chr12	113502384	113929884	132.876	DTX1, RASAL1, CCDC42B, DDX54, RITA1, IQCD, TPCN1, SLC8B1, PLBD2, SDS, SDSL, LHX5	Y_RNA, AC089999.1, Y_RNA, RP11-545P7.4, RP11-82C23.2	427.5
chrX	41346170	41826170	127.872	CASK, GPR34, GPR82	CASK-AS1, RNU6-1321P, RN7SL406P, RP11-204C16.4, RN7SL144P, RP5-1174J21.2, RP5-1174J21.1, RNU6-202P	480
chr4	46452667	46740167	99.1607	GABRA2, COX7B2	RNU6-412P, RAC1P2	287.5
chr20	20410201	20652701	90.3296	RALGAPA2	EIF4E2P1	242.5
chr12	79434884	79594884	88.3779	SYT1	RP11-390N6.1	160
chr19	38754984	38924984	86.3091	SPINT2, CTB-102L5.4, C19orf33, YIF1B, KCNK6, CATSPERG, PSMD8, GGN, SPRED3, FAM98C, RASGRP4, RYR1	Y_RNA, AC026806.2, snoU13, AC005625.1, AC005789.9, AC005789.11	170
chr2	194927995	195200495	85.9595		GLULP6, HNRNPA1P47	272.5

Top ten candidate regions for population CEU

CEU						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr14	67220445	67897945	247.929	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2	CTD-2560C21.1, RP11-862P13.1, RP11-125H8.1, Y_RNA	677.5
chr8	16193536	16471036	199.825	MSR1	MRPL49P2, RP11-13N12.2	277.5
chr19	40427484	40689984	186.635	FCGBP, PSMC4, ZNF546, ZNF780B, ZNF780A	CTC-471F3.4, AC007842.1, CTC-471F3.6, CTC-471F3.5, AC005614.5, AC005614.3, VN1R96P	262.5
chr8	15941036	16166036	178.816	MSR1	RP11-447G11.1	225
chr1	51475610	52005610	177.243	C1orf185, RNF11, TTC39A, EPS15	MIR4421, Y_RNA, CFL1P2, AL162430.2, AL162430.1, RP11-296A18.3, snoU13, RP11-296A18.5, RP11-296A18.6, RP11-275F13.1, RP11-275F13.3, RNU6-877P, RP11-253A20.1, RP11-191G24.1, RNU6-1281P	530
chr1	225048110	225355610	167.216	DNAH14	NONP	307.5
chr11	38073350	38415850	155.019		RP11-436H16.1	342.5
chr11	129788350	130070850	142.815	PRDM10, AP003041.2, APLP2, ST14	LINC00167, RP11-567M21.3, TCEB2P2, RP11-679I18.4, AP003041.1, RPL34P21	282.5
chr4	176176373	176431373	134.246		RP11-287F9.1, RP11-287F9.2, RP11-598D14.1, AC131094.1, TSEN2P1	255
chr5	142055539	142273039	128.967	FGF1, ARHGAP26	AC005592.3, AC005592.1, ARHGAP26-AS1	217.5

Top ten candidate regions for population FIN

FIN						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr14	67220427	67905427	260.28	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2	CTD-2560C21.1, RP11-862P13.1, RP11-125H8.1, Y_RNA	685
chr1	51465610	52033110	203.317	C1orf185, RNF11, TTC39A, EPS15	MIR4421, Y_RNA, CFL1P2, AL162430.2, AL162430.1, RP11-296A18.3, snoU13, RP11-296A18.5, RP11-296A18.6, RP11-275F13.1, RP11-275F13.3, RNU6-877P, RP11-253A20.1, RP11-191G24.1, RNU6-1281P, CALR4P	567.5
chr5	43590539	44078039	189.27	NNT	NNT-AS1, RPL29P12, RP11-8L21.1, RNU6-381P	487.5
chr1	100410610	100785610	186.713	SLC35A3, HIAT1, SASS6, TRMT13, LRRC39, DBT, RTCA	RP5-884G6.2, RNU6-750P, RNU6-1318P, RP4-714D9.5, RP4-714D9.2, RP4-714D9.4, RP11-305E17.7, BRI3P1, RP11-305E17.4, RP11-305E17.6, MIR553	375
chr3	129027748	129302748	161.935	H1FX, EFCAB12, MBD4, IFT122, RHO, H1FOO, PLXND1	H1FX-AS1, NUP210P3, RP13-685P2.8, RP13-685P2.7, RP11-529F4.1, RPL32P3, SNORA7B	275
chr11	129788350	130083350	148.318	PRDM10, AP003041.2, APLP2, ST14	LINC00167, RP11-567M21.3, TCEB2P2, RP11-679I18.4, AP003041.1, RPL34P21	295
chr5	96860539	97303039	144.959		RP11-1E3.1, RP11-72K17.2, RP11-72K17.1, RP11-455B3.1	442.5
chr1	6298110	6473110	141.297	HES3, GPR153, ACOT7, HES2	LINC00337, RP1-202O8.3	175
chr1	6498110	6608110	130.651	ESPN, TNFRSF25, PLEKHG5, NOL9	RP1-202O8.2, RNU6-731P, RP11-58A11.2	110
chr14	66610427	66905427	130.313		Y_RNA, RP11-72M17.1	295

Top ten candidate regions for population GBR

GBR						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr14	67183154	67930654	276.577	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2, TMEM229B	CTD-2560C21.1, RP11-862P13.1, RP11-125H8.1, Y_RNA, MIR5694	747.5
chr11	37818350	38413350	202.999		RP11-159D8.1, RP11-436H16.1	595
chr12	44294884	44742384	199.936	TMEM117	RP11-624G19.1, RP11-46I1.1, RP11-46I1.2	447.5
chr13	64276465	64591465	192.156	AL445989.1	LINC00395, OR7E156P, RP11-473M10.3, RNU6-81P, PPP1R2P10, RP11-394A14.2, OR7E104P, RP11-394A14.4, NFYAP1, LINC00355	315
chr5	43798039	44070539	153.104		RP11-8L21.1, RNU6-381P	272.5
chr7	98853084	99265584	146.374	ARPC1A, ARPC1B, PDAP1, BUD31, PTCD1, ATP5J2-PTCD1, CPSF4, AC073063.1, ATP5J2, ZNF789, ZNF394, ZKSCAN5, FAM200A, ZNF655, GS1-259H13.10, ZSCAN25, CYP3A5	MYH16, snoU13, AC073063.10, AC005020.1, GS1-259H13.2, GS1-259H13.7	412.5
chr11	129753350	130053350	146.406	NFRKB, PRDM10, AP003041.2, APLP2, ST14	LINC00167, RP11-567M21.3, TCEB2P2, RP11-679I18.4, AP003041.1, RPL34P21	300
chr4	81677667	81955167	141.846	C4orf22, BMP3	NONP	277.5
chr8	16213536	16451036	139.258	MSR1	MRPL49P2	237.5
chr6	121367616	121707616	136.962	TBC1D32	RNU6-1286P, Y_RNA, RP1-276J11.2	340

Top ten candidate regions for population IBS

IBS						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**

chr8	42643536	43378536	233.935	CHRNA6, THAP1, RNF170, HOOK3, RP11-598P20.5, FNTA, POMK, HGSNAT	RN7SL806P, MIR4469, Y_RNA, RNU1-124P, RP11-598P20.3, VN1R46P, RP11-726G23.2, RP11-726G23.11, RP11-359P18.2, RP11-726G23.3, AFG3L2P1, RP11-726G23.7, RP11-726G23.10, RP11-726G23.8, POTEA, RNU6-104P, RP11-726G23.12, AC022616.1, RP11-726G23.6, U3, RN7SKP41, RP11-359P18.1, RP11-359P18.7, RP11-359P18.8, SNX18P27	735
chr11	38005850	38423350	233.183		RP11-436H16.1	417.5
chr15	45108305	45438305	198.891	C15orf43, SORD, DUOX2, DUOXA2, DUOXA1, DUOX1	CTD-2008A1.2, CTD-2008A1.1, Y_RNA, RNU1-119P, CTD-2014N11.1, CTD-2014N11.2, RNU6-1108P, RNU6-1332P, CTD-2014N11.3, RNU6-966P, RNU1-78P, RP11-109D20.1, Y_RNA, RP11-109D20.2	330
chr10	74916660	75421660	181.979	ECD, FAM149B1, DNAJC9, MRPS16, TTC18, ANXA7, MSS51, PPP3CB, USP54, MYOZ1, SYNPO2L	Y_RNA, EIF4A2P2, DNAJC9-AS1, RP11-152N13.5, RNU6-833P, snoU13, Y_RNA, RP11-537A6.9, RP11-345K20.2, AL353731.1, RP11-137L10.6, RNU6-883P, RP11-137L10.5, RP11-464F9.20, RP11-464F9.22, RP11-464F9.21	505
chr5	109573039	109955539	155.822	TMEM232	MIR548F3	382.5
chr4	176180167	176522667	149.969		RP11-287F9.1, RP11-287F9.2, RP11-598D14.1, AC131094.1, TSEN2P1, ADAM20P2	342.5
chr1	100428110	100745610	146.109	SLC35A3, HIAT1, SASS6, TRMT13, LRRC39, DBT, RTCA	RP5-884G6.2, RNU6-750P, RNU6-1318P, RP4-714D9.5, RP4-714D9.2, RP4-714D9.4, RP11-305E17.7, BRI3P1, RP11-305E17.4, RP11-305E17.6	317.5
chr6	84525116	84775116	142.254	RIPPLY2, CYB5R4, MRAP2	RP4-676J13.2, RP11-51G5.1	250
chr3	89835226	90125226	137.575		U3	290

chr15	44990805	45098305	136.957	PATL2, B2M, TRIM69	NONP	107.5

Top ten candidate regions for population TSI

TSI						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr14	67213154	67928154	259.153	GPHN, FAM71D, MPP5, ATP6V1D, EIF2S1, PLEK2, TMEM229B	CTD-2560C21.1, RP11-862P13.1, RP11-125H8.1, Y_RNA, MIR5694	715
chr11	38005850	38420850	203.257		RP11-436H16.1	415
chr4	176179190	176424190	161.649		RP11-287F9.1, RP11-287F9.2, RP11-598D14.1, AC131094.1, TSEN2P1	245
chr10	68916691	69286691	139.343	CTNNA3	RP11-93L14.1	370
chr5	43753039	44048039	136.326		RP11-8L21.1	295
chr15	45107600	45360100	136.935	C15orf43, SORD	CTD-2008A1.2, CTD-2008A1.1, Y_RNA, RNU1-119P, CTD-2014N11.1, CTD-2014N11.2, RNU6-1108P, RNU6-1332P, CTD-2014N11.3, RNU6-966P, RNU1-78P, RP11-109D20.1, Y_RNA	252.5
chr18	67553346	67918346	131.467	CD226, RTTN	NONP	365
chr1	1115610	1448110	124.542	TTLL10, TNFRSF18, TNFRSF4, SDF4, B3GALT6, FAM132A, UBE2J2, SCNN1D, ACAP3, PUSL1, CPSF3L, GLTPD1, TAS1R3, DVL1, MXRA8, AURKAIP1, CCNL2, RP4-758J18.2, MRPL20, ANKRD65, TMEM88B, VWA1, ATAD3C, ATAD3B, ATAD3A	RP5-902P8.12, RP5-902P8.10, RP5-890O3.9, RP5-890O3.3, RN7SL657P, RP4-758J18.13, RP4-758J18.7, RP4-758J18.10	332.5
chr10	74686691	75299191	121.593	OIT3, PLA2G12B, P4HA1, NUDT13, ECD, FAM149B1, DNAJC9, MRPS16, TTC18, ANXA7, MSS51, PPP3CB, USP54	RPL17P50, RP11-344N10.4, RP11-344N10.2, Y_RNA, RP11-344N10.5, RP11-152N13.16, SNORA11, Y_RNA, EIF4A2P2, DNAJC9-AS1, RP11-152N13.5, RNU6-833P, snoU13, Y_RNA, RP11-537A6.9, RP11-345K20.2, AL353731.1, RP11-137L10.6, RNU6-883P, RP11-137L10.5	612.5
chr6	110346839	110674339	120.537	WASF1, CDC40, METTL24	NONP	327.5

Top ten candidate regions for population CDX

CDX						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr4	41515167	42215167	226.39	LIMCH1, PHOX2B, TMEM33, DCAF4L1, SLC30A9, BEND4	RP11-227F19.5, OR5M14P, RP11-227F19.1, RP11-227F19.2, RNU1-49P, HMGB1P28, LINC00682, RP11-457P14.5, RP11-457P14.6, RP11-814H16.2, ATP1B1P1	700
chr2	108913021	109383021	222.695	SULT1C2, SULT1C4, GCC2, LIMS1, RANBP2	RP11-443K8.1, SULT1C2P1, RP11-465O11.2, RP11-465O11.1, AC012487.2, AC010095.5, AC010095.6, AC010095.7	470
chr15	63850064	64305064	199.61	USP3, FBXL22, HERC1, DAPK2	USP3-AS1, RP11-317G6.1, MIR422A, RP11-111E14.1	455
chr1	92983116	93438116	167.165	EVI5, RPL5, FAM69A	RP4-593M8.1, HMGB3P9, RNU4-59P, RP11-330C7.3, RP11-330C7.4, CCNJP2, SNORD21, SNORA66, SNORA66, SNORA51, RP11-386I23.1, RNU6-970P	455
chr5	117663039	117963039	161.902		CTD-2281M20.1, RP11-2N5.2, RP11-2N5.1	300
chr13	64277785	64590285	143.195	AL445989.1	LINC00395, OR7E156P, RP11-473M10.3, RNU6-81P, PPP1R2P10, RP11-394A14.2, OR7E104P, RP11-394A14.4, NFYAP1, LINC00355	312.5
chr7	136120584	136395584	142.153		AC009784.3, AC009541.1, hsa-mir-490	275
chr8	10735664	11108164	136.764	XKR6, AF131215.5	MIR598, AF131215.6, AF131215.9, AF131215.2, AF131215.3, AF131215.4, AF131215.1, AF131215.8, LINC00529	372.5
chr3	154165096	154435096	135.404		RP11-656A15.1, CTD-2501O3.2, CTD-2501O3.3, RPL9P15	270
chr3	17570096	17912596	131.583	TBC1D5	U7, AC104451.2	342.5

Top ten candidate regions for population CHB

CHB						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chrX	100985920	101448420	241.559	NXF5, ZMAT1, TCEAL2, TCEAL6, BEX5	RP1-232L22_B.1, RP1-3E10.2, RNU6-345P, RP1-197J16.1, RP1-197J16.2, MTND6P13, TCP11X3P	462.5
chr2	108905521	109650521	239.56	SULT1C2, SULT1C4, GCC2, LIMS1, RANBP2, CCDC138, EDAR	RP11-443K8.1, SULT1C2P1, RP11-465O11.2, RP11-465O11.1, AC012487.2, AC010095.5, AC010095.6, AC010095.7, AC073415.2	745
chr15	63764703	64337203	237.469	USP3, FBXL22, HERC1, DAPK2	USP3-AS1, RP11-317G6.1, MIR422A, RP11-111E14.1	572.5
chr12	44354884	44699884	226.406	TMEM117	RP11-624G19.1, RP11-46I1.1, RP11-46I1.2	345
chr3	154167942	154822942	211.634	MME	RP11-656A15.1, CTD-2501O3.2, CTD-2501O3.3, RPL9P15, RP11-439C8.1, RP11-439C8.2	655
chr8	10725271	11112771	205.027	XKR6, AF131215.5	MIR598, AF131215.6, AF131215.9, AF131215.2, AF131215.3, AF131215.4, AF131215.1, AF131215.8, LINC00529	387.5
chr11	25030850	25368350	179.292	LUZP2	RP11-54J7.2	337.5
chr5	116503039	116743039	162.246		RPL35AP15	240
chr10	21454191	21846691	161.865	NEBL, CASC10, SKIDA1, MLLT10	NEBL-AS1, RP11-565H13.3, LUZP4P1, RNU6-15P, RP11-275N1.1, RNMTL1P1, Y_RNA, U3, MIR1915	392.5
chr3	17560442	17965442	157.582	TBC1D5	U7, AC104451.2, AC104297.1, PDCL3P3	405

Top ten candidate regions for population CHS

CHS						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr2	108905521	109690521	177.833	SULT1C2, SULT1C4, GCC2, LIMS1, RANBP2, CCDC138, EDAR	RP11-443K8.1, SULT1C2P1, RP11-465O11.2, RP11-465O11.1, AC012487.2, AC010095.5, AC010095.6, AC010095.7, AC073415.2	785
chr13	68170269	68472769	166.673		BCRP9, NPM1P22	302.5
chr16	17424477	17694477	148.044	XYLT1	RP11-916L7.1	270
chr12	123977384	124314884	147.365	RILPL1, TMED2, DDX55, EIF2B1, GTF2H3, TCTN2, ATP6V0A2, DNAH10	MIR3908, RP11-486O12.2, SNORA9, RP11-338K17.8, RPL27P12	337.5
chr2	197118021	197820521	146.797	HECW2, CCDC150, GTF3C3, C2orf66, PGAP1	AC020571.3, RN7SL820P, SCARNA16	702.5
chr8	10932815	11105315	142.24	XKR6, AF131215.5	AF131215.9, AF131215.2, AF131215.3, AF131215.4, AF131215.1, AF131215.8, LINC00529	172.5
chr1	92910616	93288116	141.658	GFI1, EVI5	RP4-593M8.1, HMGB3P9, RNU4-59P, RP11-330C7.3, RP11-330C7.4, CCNJP2	377.5
chr3	154172889	154507889	140.519		RP11-656A15.1, CTD-2501O3.2, CTD-2501O3.3, RPL9P15	335
chr2	177600521	177915521	130.548		AC092162.1, FUCA1P1, AC092162.2, AC073636.1, RNU6-187P, AC079305.11	315
chr4	41805167	42142667	129.592	TMEM33, DCAF4L1, SLC30A9, BEND4	RP11-227F19.1, HMGB1P28, LINC00682, RP11-457P14.5, RP11-457P14.6, RP11-814H16.2, ATP1B1P1	337.5

Top ten candidate regions for population JPT

JPT						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr10	55859211	56226711	204.738	PCDH15	AC013737.1, RNU6-687P	367.5
chr3	154170507	154600507	201.772		RP11-656A15.1, CTD-2501O3.2, CTD-2501O3.3, RPL9P15, RP11-439C8.1	430

chr1	92943110	93315610	180.979	GFI1, EVI5, RPL5, FAM69A	RP4-593M8.1, HMGB3P9, RNU4-59P, RP11-330C7.3, RP11-330C7.4, CCNJP2, SNORD21, SNORA66, SNORA66, SNORA51	372.5
chrX	100918625	101443625	179.144	NXF5, ZMAT1, TCEAL2, TCEAL6, BEX5	GHc-602D8.2, RNU6-587P, RP1-232L22_A.1, RP1-232L22_B.1, RP1-3E10.2, RNU6-345P, RP1-197J16.1, RP1-197J16.2, MTND6P13, TCP11X3P	525
chr4	41805167	42215167	170.622	TMEM33, DCAF4L1, SLC30A9, BEND4	RP11-227F19.1, HMGB1P28, LINC00682, RP11-457P14.5, RP11-457P14.6, RP11-814H16.2, ATP1B1P1	410
chr2	24048021	24375521	168.571	ATAD2B, UBXN2A, MFSD2B, C2orf44, FKBP1B, SF3B14, FAM228B, TP53I3, PFN4, RP11-507M3.1	PGAM1P6, AC066692.3, SDHCP3, RN7SL610P, RNU6-370P	327.5
chr2	197155521	197818021	163.049	HECW2, CCDC150, GTF3C3, C2orf66, PGAP1	SCARNA16	662.5
chr6	26120112	26367612	151.624	HIST1H2BC, HIST1H2AC, HIST1H1E, HIST1H2BD, HIST1H2BE, HIST1H4D, HIST1H3D, HIST1H2AD, HIST1H2BF, HIST1H4E, HIST1H2BG, HIST1H2AE, HIST1H3E, HIST1H1D, HIST1H4F, HIST1H4G, HIST1H3F, HIST1H2BH, HIST1H3G, HIST1H2BI, HIST1H4H, BTN3A2	LARP1P1, HIST1H1PS1, RP1-34B20.4, HIST1H2APS3, HIST1H2APS4, HIST1H3PS1, RNU6-1259P, AL021917.1	247.5
chr9	126360904	126725904	124.604	DENND1A	RP11-417B4.2, RP11-417B4.3, PIGFP2	365
chr14	49933016	50410516	118.373	RPS29, AL139099.1, LRR1, RPL36AL, MGAT2, DNAAF2, POLE2, KLHDC1, KLHDC2, NEMF, AL627171.2, AL627171.1, ARF6	RNA5SP384, RPL32P29, RN7SL1, Y_RNA, RHOQP1, RP11-649E7.5, RP11-649E7.7, RP11-831F12.3, RP11-831F12.4, RNU6ATAC30P, RP11-831F12.2, RNU6-539P, RN7SL3, RN7SL2, RNU6-189P, RP11-58E21.4	477.5

Top ten candidate regions for population KHV

KHV						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr5	117648039	117970539	216.287		CTD-2281M20.1, RP11-2N5.2, RP11-2N5.1	322.5
chr5	117323039	117625539	196.657		CTD-3179P9.1, CTD-3179P9.2	302.5
chr8	10710323	11112823	194.053	XKR6, AF131215.5	MIR598, AF131215.6, AF131215.9, AF131215.2, AF131215.3, AF131215.4, AF131215.1, AF131215.8, LINC00529	402.5
chr2	108948021	109553021	193.053	SULT1C4, GCC2, LIMS1, RANBP2, CCDC138, EDAR	SULT1C2P1, RP11-465O11.2, RP11-465O11.1, AC012487.2, AC010095.5, AC010095.6, AC010095.7, AC073415.2	605
chr13	64245285	64590285	179.917	AL445989.1	LINC00395, OR7E156P, RP11-473M10.3, RNU6-81P, PPP1R2P10, RP11-394A14.2, OR7E104P, RP11-394A14.4, NFYAP1, LINC00355	345
chr2	197115521	197815521	164.939	HECW2, CCDC150, GTF3C3, C2orf66, PGAP1	AC020571.3, RN7SL820P, SCARNA16	700
chr15	63860064	64232564	162.322	USP3, FBXL22, HERC1, DAPK2	USP3-AS1, RP11-317G6.1, MIR422A, RP11-111E14.1	372.5
chr1	238933116	239145616	161.77		MIPEPP2	212.5
chr7	136088084	136363084	160.373		AC009784.3, AC009541.1	275
chr12	44397384	44837384	156.351	TMEM117	RP11-624G19.1, RP11-46I1.1, RP11-46I1.2	440

Top ten candidate regions for population BEB

BEB						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr12	44307384	44927384	238.617	TMEM117, NELL2	RP11-624G19.1, RP11-46I1.1, RP11-46I1.2	620
chr1	51565610	52038110	194.666	C1orf185, RNF11, TTC39A, EPS15	Y_RNA, CFL1P2, AL162430.2, AL162430.1, RP11-296A18.3, snoU13, RP11-296A18.5, RP11-296A18.6, RP11-275F13.1, RP11-275F13.3, RNU6-877P, RP11-253A20.1, RP11-191G24.1, RNU6-1281P, CALR4P	472.5
chr11	72983350	73370850	175.258	P2RY6, ARHGEF17, RELT, FAM168A, PLEKHB1	RP11-800A3.7, AP002761.1, RP11-809N8.2, RP11-809N8.4, RP11-809N8.6, RP11-809N8.5, HMGN2P38, AP000860.2	387.5
chr12	49812384	50189884	174.685	SPATS2, KCNH3, MCRS1, PRPF40B, FAM186B, FMNL3, TMBIM6, NCKAP5L	RP11-161H23.8, RP11-133N21.10, RNU6-834P, POLR2KP1, RP11-133N21.7, HIGD1AP9, RP11-133N21.12, LSM6P2	377.5
chr2	81633111	81950611	157.622		AC012075.1, AC012075.2, RNA5SP99, AC013262.1	317.5
chr22	46548536	46856036	150.059	PPARA, CDPF1, PKDREJ, TTC38, GTSE1, TRMU, CELSR1	NONP	307.5
chr5	43593018	44048018	149.607	NNT	NNT-AS1, RPL29P12, RP11-8L21.1	455
chr6	121387616	121705116	148.005	TBC1D32	RNU6-1286P, Y_RNA, RP1-276J11.2	317.5
chr1	52415610	52790610	147.496	RAB3B, TXNDC12, KTI12, BTF3L4, ZFYVE9	RNA5SP48, RP11-91A18.1, RN7SL290P, RP11-91A18.4, TXNDC12-AS1, RN7SL788P, RP4-800M22.1, RP4-800M22.2, PDCL3P6, RP4-800M22.4, DNAJC19P7, ANAPC10P1	375
chr1	100410610	100718110	140.666	SLC35A3, HIAT1, SASS6, TRMT13, LRRC39, DBT	RP5-884G6.2, RNU6-750P, RNU6-1318P, RP4-714D9.5, RP4-714D9.2, RP4-714D9.4, RP11-305E17.7, BRI3P1, RP11-305E17.4	307.5

Top ten candidate regions for population GIH

GIH						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr5	43588039	44073039	230.404	NNT	NNT-AS1, CTD-2210P15.2, RPL29P12, RP11-8L21.1, RNU6-381P	485
chr4	106462667	106815167	203.813	ARHGEF38, INTS12, GSTCD	AC004066.3, ARHGEF38-IT1, RP11-311D14.1, RP11-45L9.1	352.5
chr12	49824884	50189884	172.123	SPATS2, KCNH3, MCRS1, PRPF40B, FAM186B, FMNL3, TMBIM6, NCKAP5L	RP11-161H23.8, RP11-133N21.10, RNU6-834P, POLR2KP1, RP11-133N21.7, HIGD1AP9, RP11-133N21.12, LSM6P2	365
chr4	29937667	30175167	166.773		RPS3AP17, RP11-174E22.2	237.5
chr4	29740167	29927667	153.909		EEF1A1P21, AC109351.1, RP11-390C19.1	187.5
chr11	72938350	73355850	150.339	P2RY2, P2RY6, ARHGEF17, RELT, FAM168A	RP11-800A3.4, OR8R1P, RP11-800A3.7, AP002761.1, RP11-809N8.2, RP11-809N8.4, RP11-809N8.6, RP11-809N8.5, HMGN2P38, AP000860.2	417.5
chr7	119623084	119803084	147.573		U1, RP4-742N3.1	180
chr7	119083084	119340584	145.233		AC091320.2, AC091320.1	257.5
chr7	119813084	120140584	138.887	KCND2	RP5-1006K12.1	327.5
chr22	46556691	46856691	138.087	PPARA, CDPF1, PKDREJ, TTC38, GTSE1, TRMU, CELSR1	NONP	300

Top ten candidate regions for population ITU

ITU						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr12	44342384	44904884	316.972	TMEM117, NELL2	RP11-624G19.1, RP11-46I1.1, RP11-46I1.2	562.5
chr6	136485112	137030112	174.316	PDE7B, MTFR2, BCLAF1, MAP7, MAP3K5	RP13-143G15.4, RP3-406A7.1, RP3-406A7.7, RP3-406A7.3, RP3-406A7.5, NDUFS5P1, 7SK, RP3-325F22.5, RP3-325F22.3, RNA5SP219	545
chr11	37928350	38345850	169.991		RP11-159D8.1, RP11-436H16.1	417.5
chr2	81628111	81985611	165.812		AC012075.1, AC012075.2, RNA5SP99, AC013262.1	357.5
chr5	43840539	44033039	152.619		RP11-8L21.1	192.5
chr5	43588039	43820539	145.996	NNT	NNT-AS1, CTD-2210P15.2, RPL29P12	232.5
chr1	51735610	52165610	138.747	RNF11, TTC39A, EPS15, OSBPL9	RP11-275F13.1, RP11-275F13.3, RNU6-877P, RP11-253A20.1, RP11-191G24.1, RNU6-1281P, CALR4P	430
chr20	52982444	53277444	133.471	DOK5	NONP	295
chr22	46558536	46836036	132.641	PPARA, CDPF1, PKDREJ, TTC38, GTSE1, TRMU, CELSR1	NONP	277.5
chr20	30162444	30502444	128.219	ID1, COX4I2, BCL2L1, AL160175.1, TPX2, MYLK2, FOXS1, DUSP15, TTLL9	RNU6-384P, MIR3193, RP11-243J16.7, RP11-243J16.8, RNU1-94P	340

Top ten candidate regions for population PJL

PJL						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr2	81678406	81960906	210.214		AC012075.2, RNA5SP99, AC013262.1	282.5
chr5	43840539	44078039	170.54		RP11-8L21.1, RNU6-381P	237.5
chr5	43593039	43830539	159.641	NNT	NNT-AS1, RPL29P12	237.5
chr12	86117384	86589884	152.781	RASSF9, NTS, MGAT4C	RP13-619I2.2, RP11-18J9.3, RP11-812D23.1	472.5
chr1	87190610	87600610	139.144	SH3GLB1, SEP15, HS2ST1, RP5-1052I5.2	RP4-612B15.2, RP4-604K5.3, RP4-604K5.2, RP11-384B12.2, RP11-384B12.3, LINC01140	410
chr14	63692945	63917945	139.884	RHOJ, PPP2R5E	AL049871.1, RP11-696D21.2, GPHB5	225
chr11	72990850	73370850	134.666	P2RY6, ARHGEF17, RELT, FAM168A, PLEKHB1	RP11-800A3.7, AP002761.1, RP11-809N8.2, RP11-809N8.4, RP11-809N8.6, RP11-809N8.5, HMGN2P38, AP000860.2	380
chr2	194658406	194848406	126.789		RP11-764E7.1	190
chr3	50700297	51455297	125.264	DOCK3, MANF, RBM15B, VPRBP	RP11-804H8.6, MIR4787, RP11-804H8.5, RP11-646D13.1, ZNF652P1, ST13P14	755
chr4	33860167	34370167	119.113		RP11-79E3.3, RP11-79E3.2, RP11-79E3.1, RP11-548L20.1	510

Top ten candidate regions for population STU

STU						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr2	81628302	81958302	176.113		AC012075.1, AC012075.2, RNA5SP99, AC013262.1	330
chr3	96230442	96687942	164.705	MTRNR2L12, EPHA6	RP11-124D9.1, RNU6-1094P, RPL18AP8, AC117444.1, RCC2P5, CDV3P1	457.5
chr7	119178084	119788084	163.904		AC091320.1, RP11-328J2.1, U1, RP4-742N3.1	610
chr22	46546691	46859191	161.901	PPARA, CDPF1, PKDREJ, TTC38, GTSE1, TRMU, CELSR1	NONP	312.5
chr15	45114306	45351806	159.845	C15orf43, SORD	CTD-2008A1.2, CTD-2008A1.1, Y_RNA, RNU1-119P, CTD-2014N11.1, CTD-2014N11.2, RNU6-1108P, RNU6-1332P, CTD-2014N11.3, RNU6-966P, RNU1-78P, RP11-109D20.1, Y_RNA	237.5
chr12	49652384	50187384	153.065	TUBA1C, PRPH, TROAP, C1QL4, DNAJC22, SPATS2, KCNH3, MCRS1, PRPF40B, FAM186B, FMNL3, TMBIM6, NCKAP5L	RP11-977B10.2, RP11-161H23.5, RP11-161H23.9, RP11-161H23.10, RP11-161H23.8, RP11-133N21.10, RNU6-834P, POLR2KP1, RP11-133N21.7, HIGD1AP9, RP11-133N21.12, LSM6P2	535
chr4	29990167	30417667	152.906		RP11-174E22.2	427.5
chr4	29740167	29927667	145.058		EEF1A1P21, AC109351.1, RP11-390C19.1	187.5
chr14	63613154	63918154	140.92	RHOJ, PPP2R5E	AL049871.1, RP11-696D21.2, GPHB5	305
chr6	136605112	136972612	137.612	BCLAF1, MAP7, MAP3K5	RP3-406A7.1, RP3-406A7.7, RP3-406A7.3, RP3-406A7.5, NDUFS5P1, 7SK, RP3-325F22.5, RP3-325F22.3, RNA5SP219	367.5

Top ten candidate regions for population CLM

CLM						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr2	21615551	21943051	159.323		AC067959.1, AC011752.1, AC009411.2, AC009411.1, AC018742.1	327.5
chr2	194660551	195148051	159.494		RP11-764E7.1, AC068135.1, GLULP6, HNRNPA1P47	487.5
chr5	15333039	15575539	151.077	FBXL7	MARK2P5, CTD-2313D3.1	242.5
chr15	44507203	44802203	137.243	CASC4, CTDSPL2	AC073940.1, AC090519.2, AC090519.7, AC090519.6, AC090519.1, AC090519.5, AC090519.4, AC090519.3, RP11-616K22.1, RP11-616K22.2, RP11-516C1.1, RN7SL347P, HNRNPMP1	295
chr1	188745610	188965610	136.491		RP11-316I3.2, LINC01035	220
chr1	27723110	28193110	135.125	WASF2, AHDC1, FGR, IFI6, FAM76A, STX12, PPP1R8, AL109927.1	RP4-752I6.1, RP1-159A19.4, RP1-159A19.3, RP11-288L9.1, RP11-288L9.4, RNU6-949P, CHMP1AP1, RNU6-424P, RP3-426I6.2, RPEP3, RP3-426I6.5, RP3-426I6.6, RNU6-1245P, SCARNA1	470
chr4	13305167	13555167	131.345	RAB28, NKX3-2	HSP90AB2P, LINC01096	250
chr6	43410112	43650112	129.699	ABCC10, DLK2, TJAP1, LRRC73, POLR1C, YIPF3, XPO5, POLH, GTPBP2, MAD2L1BP, RSPH9, MRPS18A	RNU6-1113P, RP3-337H4.9, RP3-337H4.6, SCARNA15, RP3-337H4.10, RP3-337H4.8	240
chr12	45524884	45857384	126.903	ANO6	PLEKHA8P1, RP11-139E19.2, RP11-438E8.2	332.5
chr10	65919165	66376665	123.022		DBF4P1	457.5

Top ten candidate regions for population MXL

MXL						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr1	100410610	100790610	205.92	SLC35A3, HIAT1, SASS6, TRMT13, LRRC39, DBT, RTCA	RP5-884G6.2, RNU6-750P, RNU6-1318P, RP4-714D9.5, RP4-714D9.2, RP4-714D9.4, RP11-305E17.7, BRI3P1, RP11-305E17.4, RP11-305E17.6, MIR553	380
chr10	74926660	75406660	203.628	ECD, FAM149B1, DNAJC9, MRPS16, TTC18, ANXA7, MSS51, PPP3CB, USP54, MYOZ1, SYNPO2L	Y_RNA, EIF4A2P2, DNAJC9-AS1, RP11-152N13.5, RNU6-833P, snoU13, Y_RNA, RP11-537A6.9, RP11-345K20.2, AL353731.1, RP11-137L10.6, RNU6-883P, RP11-137L10.5, RP11-464F9.20, RP11-464F9.22	480
chr10	31454160	31896660	181.056	ZEB1	RP11-192P3.4, ZEB1-AS1, RNA5SP309, SPTLC1P1, RP11-192P3.5, RP11-472N13.2	442.5
chr10	65919160	66304160	161.945		DBF4P1	385
chr11	38005850	38358350	147.295		RP11-436H16.1	352.5
chr17	58443615	58688615	144.012	USP32, C17orf64, APPBP2, RP11-15E18.4, PPM1D	RPL12P38, RP11-15E18.5, RP11-15E18.1, RP11-15E18.3, RP11-15E18.2	245
chr10	74749160	74914160	132.076	P4HA1, NUDT13, ECD	RPL17P50, RP11-344N10.4, RP11-344N10.2, Y_RNA, RP11-344N10.5, RP11-152N13.16, SNORA11	165
chr1	149998110	150188110	128.18	VPS45, PLEKHO1	RP11-458I7.1, RN7SL480P	190
chr22	46558914	46843914	127.261	PPARA, CDPF1, PKDREJ, TTC38, GTSE1, TRMU, CELSR1	NONP	285
chrX	19235939	19523439	117.669	PDHA1, MAP3K15	Y_RNA	287.5

Top ten candidate regions for population PEL

PEL						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr2	82426507	82874007	201.671		AC105761.1, RNU6-685P, Y_RNA, AC010105.1, AC109638.1	447.5
chr3	89715225	90160225	168.113		U3	445
chr6	128550112	128945112	162.482	PTPRK	RP1-86D1.2, RP1-86D1.3, RP1-86D1.5, RP1-86D1.4, EEF1DP5, Y_RNA, snoU13	395
chr3	154365225	154695225	162.431		CTD-2501O3.2, CTD-2501O3.3, RPL9P15, RP11-439C8.1, RP11-439C8.2	330
chr7	145830584	146065584	147.669	CNTNAP2	NONP	235
chr1	248130610	248365610	140.725	OR2L13, OR2L5, OR2L2, OR2L3, OR2M5, OR2M2	OR2L9P, OR2L1P, Y_RNA, OR2L6P, Y_RNA, Y_RNA, OR2T32P, OR2M1P	235
chr15	64424703	65129703	138.865	SNX1, SNX22, PPIB, CSNK1G1, CTD-2116N17.1, KIAA0101, TRIP4, ZNF609, OAZ2, RBPMS2, PIF1	SNORA48, RN7SL595P, RN7SL707P, Y_RNA, RP11-702L15.4, GAPDHP61, RP11-330L19.1, RP11-330L19.2, Y_RNA, RNU6-549P, AC100830.4, AC100830.5, AC100830.3, MIR1272	705
chr17	58491115	58848615	135.623	USP32, C17orf64, APPBP2, RP11-15E18.4, PPM1D, BCAS3	RPL12P38, RP11-15E18.5, RP11-15E18.1, RP11-15E18.3, RP11-15E18.2, RNU6-623P, RN7SL606P, AC111155.1, Y_RNA	357.5
chr16	14129447	14396947	128.207	MKL2	CTA-276F8.2, TVP23CP2, AC040173.1, Y_RNA, RP11-65J21.3	267.5
chr22	46592628	46852628	123.798	PPARA, CDPF1, PKDREJ, TTC38, GTSE1, TRMU, CELSR1	NONP	260

Top ten candidate regions for population PUR

PUR						
CHR	**Start**	**END**	**max LR**	**Coding**	**Noncoding**	**size in kb**
chr2	194680495	195185495	193.372		RP11-764E7.1, AC068135.1, GLULP6, HNRNPA1P47	505
chr5	15328039	15563039	156.748	FBXL7	MARK2P5, CTD-2313D3.1	235
chr2	195202995	195257995	140.672		AC018799.1	55
chr8	32608715	33058715	136.989	NRG1	RP11-1002K11.1, RNU6-663P, RP11-11N9.4, MTND1P6, MTND2P32, RANP9, AC104037.1	450
chr20	58387701	58575201	136.945	PHACTR3, SYCP2, FAM217B, PPP1R3D, CDH26	RNU7-141P	187.5
chr6	75554339	75834339	134.31	COL12A1	RP11-560O20.1	280
chr1	188758110	188958110	123.838		RP11-316I3.2, LINC01035	200
chr17	58578615	58851115	122.751	APPBP2, RP11-15E18.4, PPM1D, BCAS3	RP11-15E18.5, RP11-15E18.1, RP11-15E18.3, RP11-15E18.2, RNU6-623P, RN7SL606P, AC111155.1, Y_RNA	272.5
chr20	20392701	20762701	113.27	RALGAPA2	EIF4E2P1, RP11-23O13.1, RN7SL607P	370
chr12	79032384	79244884	105.654		RP11-123M21.2, RP11-123M21.1	212.5

B.4 LR_{T_3} profile for *COL8A1*, *CMSS1* and *FILIP1L*

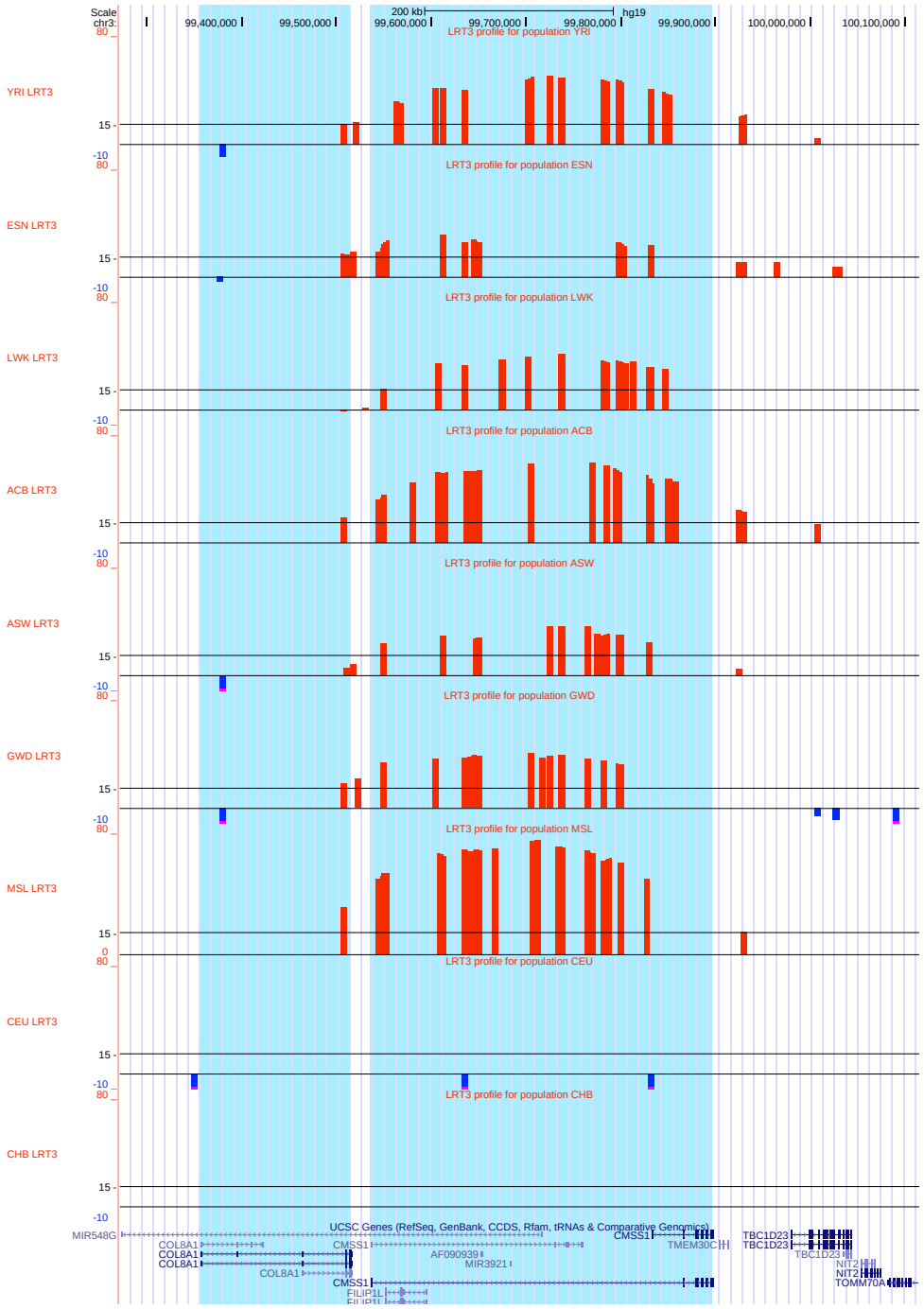


FIGURE B.2: LR_{T_3}-profile for region surrounding the genes *COL8A1*, *CMSS1* and *FILIP1L*, being significant for African populations. The LR_{T_3}-profile is shown for all seven African populations, for comparison reason, LR_{T_3}-profile for one European population CEU and one East Asia population CHB are given. Shown is the chromosomal position chr3:99,270,626-100,114,711. All three genes are highlighted. Illustration via https://genome.ucsc.edu/. Note: Only LR_{T_3}-range from -10 to 80 is shown.

B.5 $\mathbf{LR_{T_3}}$ profile for region containing *ZRANB3*, *LCT*, *MCM6* and *DARS*

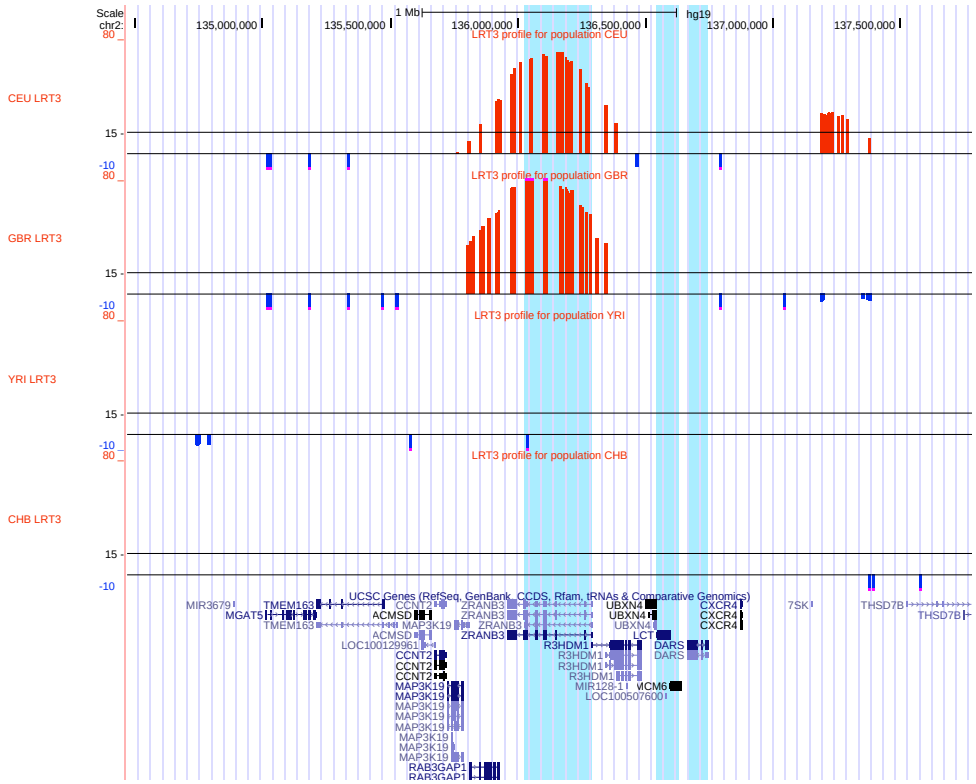


FIGURE B.3: LR_{T_3}-profile for region surrounding the genes *ZRANB3*, *LCT*, *MCM6* and *DARS*. Region contatining gene *ZRANB3* shows significant LR_{T_3} for population CEU and GBR. For comparison reason, LR_{T_3}-profile for YRI and CHB is given. Shown is the chromosomal position chr2:134,467,025-137,779,354. Illustration via `https://genome.ucsc.edu/`. Note: Only LR_{T_3}-range from -10 to 80 is shown.

B.6 GO enrichment Analysis

In the following the top three most significant enriched GO terms (of the top 10 region list) for each European population is shown.

Population IBS

IBS				
IBS - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0042743	hydrogen peroxide metabolic process	8.33E-7	1.26E-2	DUOXA2, DUOXA1, DUOX2, DUOX1
GO:0072593	reactive oxygen species metabolic process	2.25E-6	1.7E-2	DUOXA2, CYB5R4, DUOXA1, DUOX2, DUOX1
GO:0035176	social behavior	4.2E-6	2.12E-2	ANXA7, PPP3CB, DNAJC9, MSS51
IBS - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0016174	NAD(P)H oxidase activity	1.71E-7	7.81E-4	CYB5R4, DUOX2, DUOX1
GO:0050664	oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor	1.4E-6	3.2E-3	CYB5R4, DUOX2, DUOX1
IBS - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0044449	contractile fiber part	4.97E-4	9.49E-1	PPP3CB, SYNPO2L, MYOZ1, LRRC39

Population TSI

TSI				
TSI - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0035176	social behavior	3.58E-7	5.41E-3	ANXA7, PPP3CB, DNAJC9, MSS51, DVL1
GO:0051703	intraspecies interaction between organisms	3.58E-7	2.71E-3	ANXA7, PPP3CB, DNAJC9, MSS51, DVL1
GO:0051705	multi-organism behavior	1.39E-6	6.99E-3	ANXA7, PPP3CB, DNAJC9, MSS51, DVL1
TSI - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
No GO Enrichment Found.				
TSI - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0019866	organelle inner membrane	8.14E-4	1E0	MRPS16, C15orf43, AURKAIP1, ATAD3A, MRPL20, ATAD3B
GO:0031966	mitochondrial membrane	9.61E-4	9.17E-1	[MRPS16, SORD, AURKAIP1, ATAD3A, MRPL20, ATAD3B, WASF1

Population GBR

GBR				
GBR - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0034314	Arp2/3 complex-mediated actin nucleation	3.53E-4	1E0	ARPC1A, ARPC1B
GO:0045010	actin nucleation	9.01E-4	1E0	ARPC1A, ARPC1B
GBR - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
No GO Enrichment Found.				
GBR - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0034314	Arp2/3 complex-mediated actin nucleation	3.53E-4	1E0	ARPC1A, ARPC1B
GO:0045010	actin nucleation	9.01E-4	1E0	ARPC1A, ARPC1B

Population CEU

CEU				
CEU - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
No GO Enrichment Found.				
CEU - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
No GO Enrichment Found.				
CEU - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
No GO Enrichment Found.				

Population FIN

FIN				
FIN - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0055086	nucleobase-containing small molecule metabolic process	9.69E-4	1E0	NNT, SLC35A3, MBD4, ACOT7, DBT, GPHN
FIN - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
No GO Enrichment Found.				
FIN - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0055086	nucleobase-containing small molecule metabolic process	9.69E-4	1E0	NNT, SLC35A3, MBD4, ACOT7, DBT, GPHN

B.6.1 Top three most significant enriched GO terms: African vs non-African

Here, we investigate once more, if a prinicipal difference can be observed between African and non-African populations, considering biological functions and pathways targeted by selective sweep. One may expect that candidate genes, which are shared between multiple different subpopulations but not Africa, that these adaptations are a result of the Out-Of-Africa migration. For instance genes involved in the adaptation to climatic changes or food supply.

Shared between several African Population				
Shared between several African Population - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0002440	production of molecular mediator of immune response	1.44E-4	1E0	IGKV3D-20, DENND1B, IGKV2D-29, IGKV2D-28, IGKV1D-33, IGKV2D-30, IGKV6D-21, IGKV2D-26
GO:0002377	immunoglobulin production	2.63E-4	1E0	IGKV3D-20, IGKV2D-29, IGKV2D-28, IGKV1D-33, IGKV2D-30, IGKV6D-21, IGKV2D-26
GO:0030449	regulation of complement activation	5.98E-4	1E0	IGKV3D-20, SUSD4, VTN, IGKV2D-28, IGKV1D-33, IGKV2D-30, C8G
Shared between several African Population - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0016509	long-chain-3-hydroxyacyl-CoA dehydrogenase activity	5.74E-4	1E0	HADHB, HADHA
Shared between several African Population - Cellular component				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0005740	mitochondrial envelope	2.11E-4	4.06E-1	HADHB, MAOB

Shared between several Non-African SuperSuperpopulation				
Shared between several Non-African Superpopulation - Biological process				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0006396	RNA processing	3.59E-22	5.44E-18	DHX9, CDK12, AFF2, DDX5, EXOSC10, RBM39, SYF2, AARS, SNORA48, GTF2F2, SNORD37, TSEN2, PAPOLB, SCARNA1 , GTF2H3, GEMIN5, CIRH1A, SCARNA20, THUMPD3, CPSF3, SNORA27, SNORD74, RBM6, RBM5, AICDA, NOL9, NOC4L, MNAT1, EXOSC6, SNRPN, SNORA84, SNORA49, SNORD115-6, RBPMS2, SNORD115-5, SNORA46, SNORD115-12, SNORD115-11, SNORD115-14, SNORD115-8, PDCD7, SNORD115-10, SNORA40, SNORD115-9, SNORD115-15, ISY1, SNORD115-17, SNORD115-16, SNORD115-18, SNORD115-19, SNORD115-20, SNORD115-21, SNORD115-23, PTCD1, SNORA77, SNORD115-29, SNORD115-22, SNORD115-25, SNORD115-33, SNORD115-32, SNORD115-31, CPSF4, SNORD115-30, SNORD115-37, SNORD115-38, SNORA62, SNORD115-35, BUD31, SNORD115-36, SNORD115-34, SNORA51, HNRNPLL, SNORA70F, SNORD116-7, SNORA9, SNORD54, SNORD116-3, SNORA1, RTCA, SNORD116-6, SNORD116-5, SNORD116-11, SNORA66, MTFMT, SNORD116-10, SCARNA11, SNORD116-2, SCARNA16, SNORD116-1, LIN28B, SNORD116-30, SNORD116-14, SNORD116-15, SNORD116-9, SNORD60, SNORD116-8, SCARNA15, SNORD116-23, SNORD116-16, SF3B, SNORD116-13, SNORA31, SNORD116-12, SNORD116-18, SNORD73A, RPP38, SNORA70, SNORD116-24, SNORA24, SNORD116-27, SETX, SNORD21, SNORD64, SNORD116-17, SNORD116-20, CDC40, SNORD115-3, SNORD102, SNORD115-4, SNORA11, SNORD115-2, SNORD116-25 RRP9, INTS12, PRPF40B, PRPF6, SNORD118, AGO3, NOL8, SNORD90, AGO4, RPL5, C7orf60, SNORD116-19, AGO1, PUSL1, PUS7L, UTP3, SNORD115-28, SNORA25, SNORD115-27, SNORD115-24, SNORD115-45, DDX51, PPP1R8, CPSF7, PAF1, RPP40, AURKAIP1, SNORD115-39, SNORD115-40, SNORD116-29, SNORD115-48, SNORD115-43, SNORD115-44 SNORD115-41, SNORD115-42, RBPMS, TRMU, CELF6, PUS1, NAT10, SNORD127 , RNGTT, CPSF3L, RNMT, SNORA7A, SNORA3, SNORD108, MRPS111, SFPQ, RBFOX2, SNORD112, SRSF1, PSPC1, HNRNPA2B1, SRSF2, SNORD87, TRMT13, SNORD115-1, NOL3, SNORD109B, SNORD109A, JMJD6, RPL10A, DHX16, SNORD3A, ECD, RBM22, GRSF1
GO:0035194	posttranscriptional gene silencing by RNA	1.09E-17	8.26E-14	MIR551A, MIR922, MIR550A1, MIR223, MIR422A, MIR135A1, MIR553, AGO3, MIR63, MIR328, MIR320C2, AGO4, MIR875, MIR125B2, AGO1, MIR636, MIR193B , MIR181B2, MIRLET7G, MIR181A, MIR598, MIR211, MIR99A, MIR599, MIRLET7C, TNRC6C, MIR1275, MIR548A3, MIR147Aa, CNOT8, MIR490

GO:0035195	gene silencing by miRNA	1.42E-17	7.18E-14	MIR551A, MIR922, MIR223, MIR550A1, MIR422A , MIR135A1, MIR553, MIR633 , MIR328, MIR320C2, MIR875, MIR125B2, MIR636, MIR193B, MIR181B2, MIRLET7G, MIR598, MIR181A2, MIR599, MIRLET7C, MIR211, MIR99A, MIR1275, TNRC6C, MIR548A3, MIR147A, CNOT8, MIR490
Shared between several Non-African Superpopulation - Molecular function				
GO Term	**Description**	**P-value**	**FDR q-value**	**Genes**
GO:0034987	immunoglobulin receptor binding	2.78E-7	1.27E-3	TRBC2, IGLC1, IGLL5, IGJ, IGLC3, IGLC2, IGLC6, FGR, IGLC7
GO:1903231	mRNA binding involved in posttranscriptional gene silencing	1.1E-6	2.52E-3	MIRLET7G, MIR328, MIR181A2, MIR223, MIRLET7C, MIR181B2, MIR125B2
GO:0046982	protein heterodimerization activity	3.75E-5	5.7E-2	HIST1H3D, HIST1H3E, HIST2H2BF, HIST1H3I, AOC3, ABCG5, HIST1H2BO, HIST2H2BE, ABCG8, HIST1H4D, HIST2H3D, HIST1H4F, SUCLG2, HIST1H3G, HIST1H3J, SMC3, HIST1H2AM, HIST2H2AC, HIST2H2AA3, CTNNA1, HIST2H3C, HIST1H2AL, HIST1H2AC, HIST1H2BE , HIST1H2BF, HIST1H2BI , HIST1H2BH , PVRL1, HIST1H2BG , KCNH5, HIST1H2BN , CREB3L3, P2RY1, MYOD1, PPP3CA, CENPT, ARF1, ZHX1, IKBKB, HIST2H4A, HIST1H4G, HIST1H4L, HIST1H4E , HIST1H4H , HIP1, ATF2 , HIST2H2BD, SNX1, HIST2H2AB, HIF1A, HIST1H2BD, FLOT1, NEUROD2, MICU1, HIST1H2AD, TAF4B, HIST1H2AE, ABTB2, RAF1, DYNLL2, TFAP2E, EGFR, TWIST1, NPAS3, CD3G, TENM4, SYCP2, PPARD, SLC51B, TENM3, TUBB2B, CLCF1, HEXA, HIST1H3F, BCL2L1, TAS1R3, IRAK2 , GPHB5
Shared between several Non-African Superpopulation - Cellular component				

GO:0005730	nucleolus	2.75E-23	5.25E-20	DHX9 , MAD2L1BP, C9orf3, DDX5, EXOSC10 , ORC1, MKI67IP, SNORA48, DPH6 , TRAIP, OSBP, SNORD37, TSEN2, SCARNA1, CIRH1A, SCARNA20, MOB1B, THUMPD3, MIF4GD, CDC14B, TRERF1, SNORA27, SNORD74, TTF1, NOL9, NOC4L, AGPS, POLD4, accessory subunit, EXOSC6, SNORA84, OXR11, SNORA49, SNORD115-6, SNORD115-5, SNORA46, SNORD115-12, SNORD115-11, BCAS3, SNORD115-14, SNORD115-8, SNORD115-10, SNORA40, PDHA2, SNORD115-9, PDHA1, SNORD115-15, SNORD115-17, SNORD115-16, SNORD115-18, SNORD115-19, SNORD115-20 , FGF1, SNORD115-21, NIP, SNORD115-23, GLI2, SNORA77, SNORD115-29, MXI1, SNORD115-22, SNAPC1, SNORD115-25, SNORD115-33, FBXL22, SNORD115-32, SNORD115-31, MED1, SNORD115-30, SNORD115-37, SNORA62, SNORD115-38, NVL, SNORD115-35, SNORD115-36, SNORD115-34, ZNF655, SNORA51, SNORD116-7, SNORA70F, SNORA9, SNORD54, SNORD116-3, SNORA1, SNORD116-6, SNORD116-5, SNORD116-11, SNORA66, S100A3, SNORD116-10, SCARNA11, SNORD116-2, SCARNA16, SNORD116-1, LIN28B, SNORD116-14, SNORD116-30, SNORD116-15, SNORD116-9, SNORD60, SNORD116-8, SCARNA15, SNORD116-233, SNORD116-16, SF3B4 , SNORD116-13, SNORA31, SNORD116-12, SNORD116-18, SNORD73A, ZNF106, RPP38, SNORA70, SNORD116-244, SNORA24, SNORD116-277, SETX, SNORD116-266, SNORD21, SNORD64, SNORD116-17, SNORD116-200, SNORD115-3, SNORD102, ARFGEF1, SNORD115-4, SNORA11, SNORD115-2, SNORD116-255, S100A16, RRP9, MCRS1 , HMGB2, SNORD118, HN1 , NOL8, FANCD22, SENP5, SNORD90, MAP2, MPHOSPH8, RPL5, C7orf60, SNORD116-19, RPS3A, UTP3, SNORD115-28, SNORA25, SNORD115-27, SNORD115-24, CTSV, VRK1, PAK1IP1, PPP1CA, SNORD115-39, SNORD115-40, SNORD116-299, RPAP2, SNORD115-48 , SNORD115-43, SNORD115-44, PPP1CC, SNORD115-41, VCX3A, SNORD115-42, WDR82, GRWD1, NAT10, SNORD127, ITPR3, MACROD2, RASL11A , SNORA7A, SNORA3, SNORD108, ABHD14B, SDHAF2, SIX1, CTCF, CBFA2T3, SNORD112, SNORD87, SNORD115-1, NOL3, SNORD109B, SNORD109A, JMJD6, SNORD3A, DDX55, EME1, H1FX, GTF3C3
GO:0035068	micro-ribonucleoprotein complex	1.49E-21	1.42E-18	DHX9, MIR551A, MIR922, MIR223, MIR550A1, MIR422A, MIR135A1, MIR553, AGO3, MIR633, MIR328, MIR320C2, AGO4, MIR875, MIR636, MIR125B2, AGO1, MIR193B, XPO5, MIRLET7G, MIR598, MIR181A2, MIR99A, MIR599, MIRLET7C, MIR211, MIR1275, MIR548A3, MIR147A, MIR490

GO:0000786	nucleosome	9.93E-16	6.33E-13	HIST1H3D, HIST1H2BD, HIST1H3E, HIST1H3I, HIST2H2BF, HIST1H2AD , HIST1H2BO, HIST1H2AE, HIST2H2BE, HIST1H1B, HIST1H1E, HIST2H3D, HIST1H4D, HIST1H1D, HIST1H4F, HIST1H3G, HIST1H3J, HIST1H2AM, HIST2H2AC, HIST2H3C, HIST2H2AA3, HIST1H2AL, MPHOSPH8, HIST1H2AC , HIST1H2BE , HIST1H2BF, HIST1H2BI , HIST1H2BH , HIST1H2BG , HIST1H2BN , HIST2H4A , HIST1H4L , HIST1H4E , HIST1H4H , HIST1H3F, H1FX, HIST2H2BD, HIST2H2AB

Bibliography

Abecasis, G. R. et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65.

Akey, J. M. (2009). "Constructing genomic maps of positive selection in humans: where do we go from here?" In: *Genome Res.* 19.5, pp. 711–722.

Akey, J. M. et al. (2002). "Interrogating a high-density SNP map for signatures of natural selection". In: *Genome Res.* 12.12, pp. 1805–1814.

Amos, W. and Hoffman, J. I. (2010). "Evidence that two main bottleneck events shaped modern human genetic diversity". In: *Proc. Biol. Sci.* 277.1678, pp. 131–137.

Ashburner, M. et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". In: *Nat. Genet.* 25.1, pp. 25–29.

Auton, A. et al. (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

Barreiro, L. B. et al. (2008). "Natural selection has driven population differentiation in modern humans". In: *Nat. Genet.* 40.3, pp. 340–345.

Beall, C. M. (2000). "Tibetan and Andean contrasts in adaptation to high-altitude hypoxia". In: *Adv. Exp. Med. Biol.* 475, pp. 63–74.

Becerra, T. A. et al. (2014). "Autism spectrum disorders and race, ethnicity, and nativity: a population-based study". In: *Pediatrics* 134.1, pp. 63–71.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society Series B (Methodological)* 57.1, pp. 289–300. DOI: http://dx.doi.org/10.2307/2346101. URL: http://dx.doi.org/10.2307/2346101.

Bersaglieri, T. et al. (2004). "Genetic signatures of strong recent positive selection at the lactase gene". In: *Am. J. Hum. Genet.* 74.6, pp. 1111–1120.

Blum, M. G. and Francois, O. (2006). "Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance". In: *Syst. Biol.* 55.4, pp. 685–691.

Boero, F. (2015). "From Darwin's Origin of Species toward a theory of natural history". In: *F1000Prime Rep* 7, p. 49.

Botchkarev, V. A. and Fessing, M. Y. (2005). "Edar signaling in the control of hair follicle development". In: *J. Investig. Dermatol. Symp. Proc.* 10.3, pp. 247–251.

Braverman, J. M. et al. (1995). "The hitchhiking effect on the site frequency spectrum of DNA polymorphisms". In: *Genetics* 140.2, pp. 783–796.

Bryk, J. et al. (2008). "Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation". In: *PLoS ONE* 3.5, e2209.

Cai, Z. et al. (2011). "Identification of regions of positive selection using Shared Genomic Segment analysis". In: *Eur. J. Hum. Genet.* 19.6, pp. 667–671.

Campbell, M. C. and Tishkoff, S. A. (2008). "African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping". In: *Annu Rev Genomics Hum Genet* 9, pp. 403–433.

Canali, G. et al. (2018). "Genetic variants in autism-related CNTNAP2 impair axonal growth of cortical neurons". In: *Hum. Mol. Genet.* 27.11, pp. 1941–1954.

Carlson, C. S. et al. (2005). "Genomic regions exhibiting positive selection identified from dense genotype data". In: *Genome Res.* 15.11, pp. 1553–1565.

Carlsten, C. et al. (2014). "Genes, the environment and personalized medicine: We need to harness both environmental and genetic data to maximize personal and population health". In: *EMBO Rep.* 15.7, pp. 736–739.

Charlesworth, B. and Charlesworth, D. (2017). "Population genetics from 1966 to 2016". In: *Heredity (Edinb)* 118.1, pp. 2–9.

Chen, H., Patterson, N., and Reich, D. (2010). "Population differentiation as a test for selective sweeps". In: *Genome Res.* 20.3, pp. 393–402.

Chun, S. and Fay, J. C. (2011). "Evidence for hitchhiking of deleterious mutations within the human genome". In: *PLoS Genet.* 7.8, e1002240.

Danecek, P. et al. (2011). "The variant call format and VCFtools". In: *Bioinformatics* 27.15, pp. 2156–2158.

Darwin, C. and Wallace, A. (1858). "On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection". In: *Zoological Journal of the Linnean Society* 3.9, pp. 45–62. DOI: 10.1111/j.1096-3642.1858.tb02500.x. eprint: /oup/backfile/content_public/journal/zoolinnean/3/9/10.1111_j.1096-3642.1858.tb02500.x/1/j.1096-3642.1858.tb02500.x.pdf. URL: http://dx.doi.org/10.1111/j.1096-3642.1858.tb02500.x.

Davidson, R. et al. (2015). "Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer". In: *BMC Genomics* 16 Suppl 10, S1.

Dayrat, B. (2003). "The roots of phylogeny: how did Haeckel build his trees?" In: *Syst. Biol.* 52.4, pp. 515–527.

De, A. and Durrett, R. (2007). "Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum". In: *Genetics* 176.2, pp. 969–981.

Disanto, F. and Rosenberg, N. A. (2016). "Asymptotic Properties of the Number of Matching Coalescent Histories for Caterpillar-Like Families of Species Trees". In: *IEEE/ACM Trans Comput Biol Bioinform* 13.5, pp. 913–925.

Eden, E. et al. (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". In: *BMC Bioinformatics* 10, p. 48.

Eiberg, H. et al. (2008). "Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression". In: *Hum. Genet.* 123.2, pp. 177–187.

Elsabbagh, M. et al. (2012). "Global prevalence of autism and other pervasive developmental disorders". In: *Autism Res* 5.3, pp. 160–179.

Enattah, N. S. et al. (2008). "Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture". In: *Am. J. Hum. Genet.* 82.1, pp. 57–72.

Epperson, B. K. (1999). "Gustave Malécot, 1911-1998: Population Genetics Founding Father". In: *Genetics* 152.2, pp. 477–484. ISSN: 0016-6731. eprint: http://www.genetics.org/content/152/2/477.full.pdf. URL: http://www.genetics.org/content/152/2/477.

Eriksson, A., Mahjani, B., and Mehlig, B. (2009). "Sequential Markov coalescent algorithms for population models with demographic structure". In: *Theor Popul Biol* 76.2, pp. 84–91.

Esposito, T. et al. (2013). "Unique X-linked familial FSGS with co-segregating heart block disorder is associated with a mutation in the NXF5 gene". In: *Hum. Mol. Genet.* 22.18, pp. 3654–3666.

Ewens, W. J. (1972). "The sampling theory of selectively neutral alleles". In: *Theor Popul Biol* 3.1, pp. 87–112.

Ewing, G. and Hermisson, J. (2010). "MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus". In: *Bioinformatics* 26.16, pp. 2064–2065.

Ezkurdia, I. et al. (2014). "Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes". In: *Hum. Mol. Genet.* 23.22, pp. 5866–5878.

Fay, J. C. and Wu, C. I. (2000). "Hitchhiking under positive Darwinian selection". In: *Genetics* 155.3, pp. 1405–1413.

Felsenstein, J. (1985). "Confidence limits on phylogenies: An approach using the bootstrap". In: *Evolution* 39.4, pp. 783–791.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates, Inc.

Ferrer-Admetlla, A. et al. (2014). "On detecting incomplete soft or hard selective sweeps using haplotype structure". In: *Mol. Biol. Evol.* 31.5, pp. 1275–1291.

Ferretti, L., Disanto, F., and Wiehe, T. (2013). "The effect of single recombination events on coalescent tree height and shape". In: *PLoS ONE* 8.4, e60123.

Ferretti, L. et al. (2017). "Decomposing the Site Frequency Spectrum: The Impact of Tree Topology on Neutrality Tests". In: *Genetics* 207.1, pp. 229–240.

Ferretti, L. et al. (2018). "The neutral frequency spectrum of linked sites". In: *Theor Popul Biol*.

Fisher, R. (1930). "The genetical theory of natural selection". In: *Oxford University Press*.

Fortuno, C. and Labarta, E. (2014). "Genetics of primary ovarian insufficiency: a review". In: *J. Assist. Reprod. Genet.* 31.12, pp. 1573–1585.

Frazer, K. A. et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164, pp. 851–861.

Fu, Y. X. and Li, W. H. (1993). "Statistical tests of neutrality of mutations." In: *Genetics* 133.3, pp. 693–709.

Fujimoto, A. et al. (2008). "A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness". In: *Hum. Mol. Genet.* 17.6, pp. 835–843.

Gallego Romero, I. et al. (2012). "Herders of Indian and European cattle share their predominant allele for lactase persistence". In: *Mol. Biol. Evol.* 29.1, pp. 249–260.

Gazda, M. A. et al. (2018). "Signatures of Selection on Standing Genetic Variation Underlie Athletic and Navigational Performance in Racing Pigeons". In: *Mol. Biol. Evol.* 35.5, pp. 1176–1189.

Gene Ontology Consortium (2008). "The Gene Ontology project in 2008". In: *Nucleic Acids Res.* 36.Database issue, pp. D440–444.

Gene Ontology Consortium (2017). "Expansion of the Gene Ontology knowledgebase and resources". In: *Nucleic Acids Res.* 45.D1, pp. D331–D338.

Griffiths, R. C. and Marjoram, P. (1996). "Ancestral inference from samples of DNA sequences with recombination". In: *J. Comput. Biol.* 3.4, pp. 479–502.

Grossman, S. R. et al. (2010). "A composite of multiple signals distinguishes causal variants in regions of positive selection". In: *Science* 327.5967, pp. 883–886.

Grossman, S. R. et al. (2013). "Identifying recent adaptations in large-scale genomic data". In: *Cell* 152.4, pp. 703–713.

Haldane, J. B. S. (1927). "A mathematical theory of natural and artificial selection, Part V: Selection and Mutation". In: *Proc. Camb. Phil. Soc.* 23, 838—844.

Haldane, J. B. S. (1940). "The mean and variance of chi-square, when used as a test of homogeneity, when expectations are small". In: *BIOMETRIKA* 31.3/4, pp. 346–355. DOI: https://doi.org/10.2307/2332614.

Haldane, J. B. S. (1957). "The cost of natural selection". In: *J. Genet* 55, pp. 511–524.

Hales, C. M. et al. (2013). "Abnormal gephyrin immunoreactivity associated with Alzheimer disease pathologic changes". In: *J. Neuropathol. Exp. Neurol.* 72.11, pp. 1009–1015.

Han, L. and Abney, M. (2013). "Using identity by descent estimation with dense genotype data to detect positive selection". In: *Eur. J. Hum. Genet.* 21.2, pp. 205–211.

Hartl, D. and Clark, A. (2007). *Principles of Population Genetics, 4th Edition*. Sinauer Associates, Inc. ISBN: ISBN: 978-0-878-93308-2.

Hernandez, R. D. et al. (2011). "Classic selective sweeps were rare in recent human evolution". In: *Science* 331.6019, pp. 920–924.

Hider, J. L. et al. (2013). "Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry". In: *BMC Evol. Biol.* 13, p. 150.

Higasa, K. et al. (2009). "Evaluation of haplotype inference using definitive haplotype data obtained from complete hydatidiform moles, and its significance for the analyses of positively selected regions". In: *PLoS Genet.* 5.5, e1000468.

Hill, W. G. and Robertson, A. (1968). "Linkage disequilibrium in finite populations". In: *Theor. Appl. Genet.* 38.6, pp. 226–231.

Huang, T. et al. (2013). "Noncoding RNAs in cancer and cancer stem cells". In: *Chin J Cancer* 32.11, pp. 582–593.

Hubacek, J. A. et al. (2017). "Frequency of adult type-associated lactase persistence LCT-13910C/T genotypes in the Czech/Slav and Czech Roma/Gypsy populations". In: *Genet. Mol. Biol.* 40.2, pp. 450–452.

Hudson, R. R. (1983). "Properties of a neutral allele model with intragenic recombination". In: *Theor Popul Biol* 23.2, pp. 183–201.

Hudson, R. R. and Kaplan, N. L. (1985). "Statistical properties of the number of recombination events in the history of a sample of DNA sequences". In: *Genetics* 111.1, pp. 147–164.

Hudson, R. R. and Kaplan, N. L. (1988). "The coalescent process in models with selection and recombination". In: *Genetics* 120.3, pp. 831–840.

Hudson, R. (1990). "Gene genealogies and the coalescent process". In: *Oxford surveys in evolutionary biology* 7.1, p. 44.

Itan, Y. et al. (2010). "A worldwide correlation of lactase persistence phenotype and genotypes". In: *BMC Evol. Biol.* 10, p. 36.

Johansson, A. and Gyllensten, U. (2008). "Identification of local selective sweeps in human populations since the exodus from Africa". In: *Hereditas* 145.3, pp. 126–137.

Julian, C. G., Wilson, M. J., and Moore, L. G. (2009). "Evolutionary adaptation to high altitude: a view from in utero". In: *Am. J. Hum. Biol.* 21.5, pp. 614–622.

Jun, L. et al. (2001). "NXF5, a novel member of the nuclear RNA export factor family, is lost in a male patient with a syndromic form of mental retardation". In: *Current Biology* 11.18, pp. 1381 –1391. ISSN: 0960-9822. DOI: https://doi.org/10.1016/S0960-9822(01)00419-5. URL: http://www.sciencedirect.com/science/article/pii/S0960982201004195.

Kamberov, Y. G. et al. (2013). "Modeling recent human evolution in mice by expression of a selected EDAR variant". In: *Cell* 152.4, pp. 691–702.

Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989). "The "hitchhiking effect" revisited." In: *Genetics* 123.4, pp. 887–899. ISSN: 0016-6731. eprint: http://www.genetics.org/content/123/4/887.full.pdf. URL: http://www.genetics.org/content/123/4/887.

Karlsson, E. K. et al. (2013). "Natural selection in a bangladeshi population from the cholera-endemic ganges river delta". In: *Sci Transl Med* 5.192, 192ra86.

Kayser, M., Brauer, S., and Stoneking, M. (2003). "A genome scan to detect candidate regions influenced by local natural selection in human populations". In: *Mol. Biol. Evol.* 20.6, pp. 893–900.

Kelley, J. L. et al. (2006). "Genomic signatures of positive selection in humans and the limits of outlier approaches". In: *Genome Res.* 16.8, pp. 980–989.

Kim, Y. and Nielsen, R. (2004). "Linkage disequilibrium as a signature of selective sweeps". In: *Genetics* 167.3, pp. 1513–1524.

Kimura, M. (1968). "Evolutionary rate at the molecular level". In: *Nature* 217.5129, pp. 624–626.

Kimura, M. (1983). "The neutral theory of molecular selection". In: *Cambridge University Press*.

Kingman, J. (1982a). "The coalescent". In: *Stochastic Processes and their Applications* 13.3, pp. 235 –248.

Kingman, J. (1982b). "On the genealogy of large populations". In: *J. Appl. Prob.* 19A, pp. 27–43.

Kingsmore, S. and his Team from the Rady Children's Institute for Genomic Medicine (RCIGM). *New GUINNESS WORLD RECORDS^{TM} Title Set for Fastest Genetic Diagnosis*. Available at https://www.rchsd.org/about-us/newsroom/press-releases/new-guinness-world-records-title-set-for-fastest-genetic-diagnosis/ (last visited: May 2018).

Kirkpatrick, M. and Slatkin, M. (1993). "Searching for evolutionary patterns in the shape of a phylogenetic tree". In: *Evolution* 47.4, pp. 1171–1181.

Kudaravalli, S. et al. (2009). "Gene expression levels are a target of recent natural selection in the human genome". In: *Mol. Biol. Evol.* 26.3, pp. 649–658.

Lander, E. S. et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921.

Lappalainen, T. et al. (2010). "Genomic landscape of positive natural selection in Northern European populations". In: *Eur. J. Hum. Genet.* 18.4, pp. 471–478.

Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models". In: *Genetics* 49.1, pp. 49–67.

Lewontin, R. C. and Kojima, K. (1960). "The Evolutionary Dynamics of Complex Polymorphisms". In: *Evolution* 14.4, pp. 458–472.

Li, H. (2011). "A new test for detecting recent positive selection that is free from the confounding impacts of demography". In: *Mol. Biol. Evol.* 28.1, pp. 365–375.

Li, H. and Wiehe, T. (2013). "Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation". In: *PLoS Comput. Biol.* 9.5, e1003060.

Li, M. J. et al. (2014a). "dbPSHP: a database of recent positive selection across human populations". In: *Nucleic Acids Res.* 42.Database issue, pp. D910–916.

Li, W. and Freudenberg, J. (2009). "Two-parameter characterization of chromosome-scale recombination rate". In: *Genome Res.* 19.12, pp. 2300–2307.

Li, Y. et al. (2014b). "Structure of Crumbs tail in complex with the PALS1 PDZ-SH3-GK tandem reveals a highly specific assembly mechanism for the apical Crumbs complex". In: *Proc. Natl. Acad. Sci. U.S.A.* 111.49, pp. 17444–17449.

Lionel, A. C. et al. (2013). "Rare exonic deletions implicate the synaptic organizer Gephyrin (GPHN) in risk for autism, schizophrenia and seizures". In: *Hum. Mol. Genet.* 22.10, pp. 2055–2066.

Liu, C. M. et al. (2011). "ANXA7, PPP3CB, DNAJC9, and ZMYND17 genes at chromosome 10q22 associated with the subgroup of schizophrenia with deficits in attention and executive function". In: *Biol. Psychiatry* 70.1, pp. 51–58.

Liu, X. et al. (2013). "Detecting and characterizing genomic signatures of positive selection in global populations". In: *Am. J. Hum. Genet.* 92.6, pp. 866–881.

Lopez Herraez, D. et al. (2009). "Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs". In: *PLoS ONE* 4.11, e7888.

Luo, Y. et al. (2011). "Transcriptome profiling of whole blood cells identifies PLEK2 and C1QB in human melanoma". In: *PLoS ONE* 6.6, e20971.

Maynard Smith, J. and Haigh, J. (1974). "The hitch-hiking effect of a favourable gene". In: *Genet. Res.* 23.1, pp. 23–35.

McVean, G. A. and Cardin, N. J. (2005). "Approximating the coalescent with recombination". In: *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 360.1459, pp. 1387–1393.

Mendel, G. (1865). "Versuche über Pflanzen-Hybriden". In: *Verhandlungen des naturforschenden Vereines in Brünn.* Bd.4 (1865-1866), pp. 3–47. URL: https://www.biodiversitylibrary.org/part/175272.

Mendizabal, I. et al. (2012). "Adaptive evolution of loci covarying with the human African Pygmy phenotype". In: *Hum. Genet.* 131.8, pp. 1305–1317.

Mizuno, H. et al. (2010). "Fine-scale detection of population-specific linkage disequilibrium using haplotype entropy in the human genome". In: *BMC Genet.* 11, p. 27.

Moehle, M. and Sagitov, S. (2001). "A Classification of Coalescent Processes for Haploid Exchangeable Population Models". In: *Ann. Probab.* 29.4, pp. 1547–1562. DOI: 10.1214/aop/1015345761. URL: https://doi.org/10.1214/aop/1015345761.

Montinaro, F. et al. (2015). "Unravelling the hidden ancestry of American admixed populations". In: *Nat Commun* 6, p. 6596.

Moran, P. A. P. (1958). "Random processes in genetics". In: *Proc. Camb. Phil. Soc.* 54, pp. 60–71.

Mughal, M. R. and DeGiorgio, M. (2018). "Localizing and classifying adaptive targets with trend filtered regression". preprint on bioRxiv 320523; doi: http://dx.doi.org/10.1101/320523.

National Human Genome Research Institute (NHGRI). *All About the Human Genome Project*. Available at http://www.genome.gov/10001772 (last visited: May 2018).

Nielsen, R. and Slatkin, M. (2013). *An Introduction to Population Genetics: Theory and Applications*. Sinauer. ISBN: 9781605351537. URL: https://books.google.de/books?id=Iy08kgEACAAJ.

Nielsen, R. et al. (2017). "Tracing the peopling of the world through genomics". In: *Nature* 541.7637, pp. 302–310.

Ober, U. et al. (2013). "The expected linkage disequilibrium in finite populations revisited". preprint on arXiv 13044856v2; https://arxiv.org/pdf/1304.4856.pdf.

Okasha, S. (2016). "Population Genetics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.

Oleksyk, T. K. et al. (2008). "Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations". In: *PLoS ONE* 3.3, e1712.

Palmer, R. F. et al. (2010). "Explaining low rates of autism among Hispanic schoolchildren in Texas". In: *Am J Public Health* 100.2, pp. 270–272.

Parra, E. J. (2007). "Human pigmentation variation: evolution, genetic basis, and implications for public health". In: *Am. J. Phys. Anthropol.* Suppl 45, pp. 85–105.

Peng, Y. et al. (2017). "Down-Regulation of EPAS1 Transcription and Genetic Adaptation of Tibetans to High-Altitude Hypoxia". In: *Molecular Biology and Evolution* 34.4, pp. 818–830. DOI: 10.1093/molbev/msw280. eprint: /oup/backfile/content_public/journal/mbe/34/4/10.1093_molbev_msw280/2/msw280.pdf. URL: http://dx.doi.org/10.1093/molbev/msw280.

Peter, B. M., Huerta-Sanchez, E., and Nielsen, R. (2012). "Distinguishing between selective sweeps from standing variation and from a de novo mutation". In: *PLoS Genet.* 8.10, e1003011.

Pickrell, J. K. et al. (2009). "Signals of recent positive selection in a worldwide sample of human populations". In: *Genome Res.* 19.5, pp. 826–837.

Pitman, J. (1999). "Coalescents With Multiple Collisions". In: *Ann. Probab.* 27.4, pp. 1870–1902. DOI: 10.1214/aop/1022874819. URL: https://doi.org/10.1214/aop/1022874819.

Polimanti, R. et al. (2014). "Human pharmacogenomic variation of antihypertensive drugs: from population genetics to personalized medicine". In: *Pharmacogenomics* 15.2, pp. 157–167.

Pritchard, J. K. and Przeworski, M. (2001). "Linkage disequilibrium in humans: models and data". In: *Am. J. Hum. Genet.* 69.1, pp. 1–14.

Ramirez-Soriano, A. et al. (2008). "Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination". In: *Genetics* 179.1, pp. 555–567.

Ramnitz, M. S. and Lodish, M. B. (2013). "Racial disparities in pubertal development". In: *Semin. Reprod. Med.* 31.5, pp. 333–339.

Rees, M. I. et al. (2003). "Isoform heterogeneity of the human gephyrin gene (GPHN), binding domains to the glycine receptor, and mutation analysis in hyperekplexia". In: *J. Biol. Chem.* 278.27, pp. 24688–24696.

Reich, D. E. et al. (2001). "Linkage disequilibrium in the human genome". In: *Nature* 411.6834, pp. 199–204.

Roach, J. C. et al. (2010). "Analysis of genetic inheritance in a family quartet by whole-genome sequencing". In: *Science* 328.5978, pp. 636–639.

Rosenberg, N. A. and Kang, J. T. (2015). "Genetic Diversity and Societally Important Disparities". In: *Genetics* 201.1, pp. 1–12.

Ryu, B. J. et al. (2011). "Regulation of the female rat estrous cycle by a neural cell-specific epidermal growth factor-like repeat domain containing protein, NELL2". In: *Mol. Cells* 32.2, pp. 203–207.

Sabeti, P. C. et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure". In: *Nature* 419.6909, pp. 832–837.

Sabeti, P. C. et al. (2006). "Positive natural selection in the human lineage". In: *Science* 312.5780, pp. 1614–1620.

Sabeti, P. C. et al. (2007). "Genome-wide detection and characterization of positive selection in human populations". In: *Nature* 449.7164, pp. 913–918.

Schlebusch, C. M. et al. (2013). "Stronger signal of recent selection for lactase persistence in Maasai than in Europeans". In: *Eur. J. Hum. Genet.* 21.5, pp. 550–553.

Schrider, D. R. and Kern, A. D. (2016). "S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning". In: *PLoS Genet.* 12.3, e1005928.

Schrider, D. R. and Kern, A. D. (2017). "Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome". In: *Mol. Biol. Evol.* 34.8, pp. 1863–1877.

Schrider, D. R. and Kern, A. D. (2018). "Supervised Machine Learning for Population Genetics: A New Paradigm". In: *Trends Genet.* 34.4, pp. 301–312.

Schweinsberg, J. (2000). "Coalescents with Simultaneous Multiple Collisions". In: *Electron. J. Probab.* 5, 50 pp. DOI: 10.1214/EJP.v5-68. URL: https://doi.org/10.1214/EJP.v5-68.

Segurel, L. and Bon, C. (2017). "On the Evolution of Lactase Persistence in Humans". In: *Annu Rev Genomics Hum Genet* 18, pp. 297–319.

Smedley, D. et al. (2015). "The BioMart community portal: an innovative alternative to large, centralized data repositories". In: *Nucleic Acids Res.* 43.W1, W589–598.

Stoletov, K. et al. (2018). "Quantitative in vivo whole genome motility screen reveals novel therapeutic targets to block cancer metastasis". In: *Nat Commun* 9.1, p. 2343.

Storz, J. F., Payseur, B. A., and Nachman, M. W. (2004). "Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa". In: *Mol. Biol. Evol.* 21.9, pp. 1800–1811.

Sturm, R. A. and Duffy, D. L. (2012). "Human pigmentation genes under environmental selection". In: *Genome Biol.* 13.9, p. 248.

Sved, J. A. (1971). "Linkage disequilibrium and homozygosity of chromosome segments in finite populations". In: *Theor Popul Biol* 2.2, pp. 125–141.

Szollosi, G. J. et al. (2015). "The inference of gene trees with species trees". In: *Syst. Biol.* 64.1, pp. 42–62.

Tajima, F. (1983). "Evolutionary relationship of DNA sequences in finite populations". In: *Genetics* 105.2, pp. 437–460.

Tajima, F. (1989a). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism". In: *Genetics* 123.3, pp. 585–595.

Tajima, F. (1989b). "The effect of change in population size on DNA polymorphism". In: *Genetics* 123.3, pp. 597–601.

Tamaki, T. et al. (2017). "A novel transmembrane protein defines the endoplasmic reticulum stress-induced cell death pathway". In: *Biochem. Biophys. Res. Commun.* 486.1, pp. 149–155.

Tan, J. et al. (2013). "The adaptive variant EDARV370A is associated with straight hair in East Asians". In: *Hum. Genet.* 132.10, pp. 1187–1191.

Tang, K., Thornton, K. R., and Stoneking, M. (2007). "A new approach for using genome scans to detect recent positive selection in the human genome". In: *PLoS Biol.* 5.7, e171.

Templeton, A. (2002). "Out of Africa again and again". In: *Nature* 416.6876, pp. 45–51.

Tishkoff, S. A. et al. (2007). "Convergent adaptation of human lactase persistence in Africa and Europe". In: *Nat. Genet.* 39.1, pp. 31–40.

Utsunomiya, Y. T. et al. (2013). "Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods". In: *PLoS ONE* 8.5, e64280.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). "Detecting Natural Selection in Genomic Data". In: *Annual Review of Genetics* 47.1. PMID: 24274750, pp. 97–120. DOI: 10.1146/annurev-genet-111212-133526. eprint: https://doi.org/10.1146/annurev-genet-111212-133526. URL: https://doi.org/10.1146/annurev-genet-111212-133526.

Voight, B. F. et al. (2006). "A map of recent positive selection in the human genome". In: *PLoS Biol.* 4.3, e72.

Vy, H. M. and Kim, Y. (2015). "A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data". In: *Genetics* 200.2, pp. 633–649.

Wacker, M. and Holick, M. F. (2013). "Sunlight and Vitamin D: A global perspective for health". In: *Dermatoendocrinol* 5.1, pp. 51–108.

Wagh, K. et al. (2012). "Lactase persistence and lipid pathway selection in the Maasai". In: *PLoS ONE* 7.9, e44751.

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Greenwood Village: Roberts & Company Publishers. ISBN: 0-9747077-5-9.

Waldman, Y. Y. et al. (2011). "Selection for translation efficiency on synonymous polymorphisms in recent human evolution". In: *Genome Biol Evol* 3, pp. 749–761.

Wang, E. T. et al. (2006). "Global landscape of recent inferred Darwinian selection for Homo sapiens". In: *Proc. Natl. Acad. Sci. U.S.A.* 103.1, pp. 135–140.

Watterson, G. A. (1975). "On the number of segregating sites in genetical models without recombination". In: *Theor Popul Biol* 7.2, pp. 256–276.

Williamson, S. H. et al. (2007). "Localizing recent adaptive evolution in the human genome". In: *PLoS Genet.* 3.6, e90.

Wilson, B. A. et al. (2016). "The population genetics of drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens". In: *Mol. Ecol.* 25.1, pp. 42–66.

Wirtz, J., Rauscher, M., and Wiehe, T. (2018). "Topological linkage disequilibrium calculated from coalescent genealogies". preprint on bioRxiv 286393; doi: https://doi.org/10.1101/286393.

Wright, S. (1931). "Evolution in Mendelian Populations". In: *Genetics* 16.2, pp. 97–159.

Yang, Z. and Rannala, B. (2012). "Molecular phylogenetics: principles and practice". In: *Nat. Rev. Genet.* 13.5, pp. 303–314.

Yang, Z. et al. (2018). "Detecting Recent Positive Selection with a Single Locus Test Bipartitioning the Coalescent Tree". In: *Genetics* 208.2, pp. 791–805.

Zhang, C. et al. (2006). "A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations". In: *Bioinformatics* 22.17, pp. 2122–2128.

Zhong, M. et al. (2010). "A powerful score test to detect positive selection in genome-wide scans". In: *Eur. J. Hum. Genet.* 18.10, pp. 1148–1159.

Eidesstattliche Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie -abgesehen von unten angegebenen Teilpublikationen- noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt.
Die von mir vorgelegte Dissertation ist von Prof. Dr. Thomas Wiehe betreut worden.

Pre-print als Co-Autor über Themen, die in der Dissertation behandelt werden:

Wirtz J., Rauscher M., Wiehe T. (2018).
Topological linkage disequilibrium calculated from coalescent genealogies.
(preprint) on bioRxiv 286393; doi: https://doi.org/10.1101/286393.

Pre-Print als Co-Autor über Themen, die NICHT in der Dissertation behandelt werden:

Jabbari K., Wirtz J., Rauscher M., Wiehe T. (2018).
Interdependence of linkage disequilibrium, chromatin architecture and compositional genome organization of mammals.
(pre-print) bioRxiv 293837; doi:https://doi.org/10.1101/293837.

Publikation als Co-Autor über Themen, die NICHT in der Dissertation behandelt werden:

Schiffer P., Gravemeyer J., Rauscher M., Wiehe T. (2016).
"Ultra large gene families: a matter of adaptation or genomic parasites.
Life 6(3),32.