Norwegian University
of Life Sciences

Master's Thesis 2021 60 ECTS
Faculty of Chemistry, Biotechnology and Food Science

The GDR:
A novel approach to detect large-scale genomic sequence patterns

Nora B. Bull
Masters of Bioinformatics

Abstract

The emergence of new sequencing technologies over the past two decades has enabled us to dive deeper into the biomolecular aspect of the human recipe. Entire genomes from several hundred thousand people are already accessible, but how to interpretate the connections between the blueprints and the phenotypes are complicated, even for the best developed machine learning algorithms (1). Prediction of the microRNA-mRNA targeting network is a classic example, which is involved with gene regulation of all living cell processes (2-4). These non-coding features make up complex networks of interactions, where microRNAs primarily target 3'UTRs through partial complementary base-pairing. Thus, the challenge to investigate patterns in such large-scaled genomic sequence data requires new approaches.

The Group Diversity Ratio (GDR) metric is presented here as a novel approach to aid in this challenge. The GDR quantifies genome-wide structure in large-scale sequence data with a statistically testable result. Patterns are measured for a group feature that may be related to variation in sequence samples, based on phylogenetic distance estimations. It opens opportunities to quickly gain insights into genomic regions of interests and used to guide further research.

To demonstrate the use of the GDR metric, ethnicity-associated variation patterns in more than 1000 human 3'UTRs was identified with the GDR. The study set was from 1000 Genomes project, which was the first major effort to address the problem of ethnic bias in genetic studies and contained more than 2500 whole-genome sequences from 26 ethnic lineages(5).

In addition to detecting significantly distinct 3'UTR elements for ethnic populations, the key finding of this study was the high potentials of the GDR to facilitate more high-throughput characterization of genomic sequence data.

Sammendrag

Utvikling av ny sekvenseringsteknologi de to siste tiårene har tillatt dypere dykk ned i de biomolekylære aspektene ved menneskets oppskrift. Hel-genom data fra flere hundre tusen mennesker er allerede tilgjengelig, men hvordan den økende mengden informasjon kan settes sammen til meningsfull funksjonell tolkning er komplisert og krever nye metoder. MikroRNA - mRNA interaksjoner utgjør et enormt genreguleringsnettverk som er vanskelig å predikere, selv for dagens beste maskinlæringsalgoritmer(1). Disse ikke-kodende elementene er involvert i omtrent alle cellulære prosesser i mennesket, primært via delvis komplementær baseparing mellom mikroRNA og mRNA, men det er mye vi ikke forstår av dette nettverkets betydning i vår biologi (2-4). Nye metoder er nødvendige for å kunne utforske genetisk variasjon i dette nettverket, som kan gi nye innblikk i hvordan genene våre reguleres.

Her presenteres «The Group Diversity Ratio» (GDR) som en ny målenhet til å møte denne utfordringen. GDR kan kvantifisere evolusjonær struktur av variasjon i store mengder genomisk sekvensdata, med et resultat som kan statistisk valideres. Metoden baserer seg på å måle gruppe-struktur i et distanse-basert fylogenetisk tre av sekvensdata, for forhåndsdefinerte grupper av «blader» i treet. Gruppene representerer en egenskap som kan relateres til sekvensdataen, og det undersøkes til hvilken grad det finnes en sammenheng mellom de to. Metoden kan primært brukes til å raskt skaffe overblikk over store mengder genomisk sekvensdata, som kan gi verdifulle innblikk til videre etterforskning.

For å teste metoden ble GDR brukt til å identifisere variasjon assosiert med etniske populasjoner i 3’UTR data fra «The 1000 Genomes Project» (1KGP). 1KGP var det første store prosjektet som adresserte den etniske skjevheten som nå finnes i genom-databaser, og som utgjør en god grunn til å utforske etnisk genetisk variasjon (5). I tillegg til identifikasjon av mer enn 1000 3’UTR sekvenser som inneholder signifikant etnisitet-spesifikk variasjon, viser dette studiet GDR-metodens høye potensial til å undersøke genetisk variasjon i stor skala.

Acknowledgements

This work is the result of a warm welcome to Simon Rayner's group (computational biology, dept. medical genetics) at Oslo University Hospital, Ullevål, as part of my Master program in Bioinformatics at the Faculty of Chemistry, Biotechnology and Food Sciences (KBM) at the Norwegian University of Life Sciences (NMBU), in 2021.

Thank you, Simon Rayner (supervisor at OUS) for your encouragement, guidance on every detail and ideas, and for the interesting discussions. You, and the rest of the group, have certainly inspired me to continue.

Thank you, Lars-Gustav Snipen (supervisor at NMBU), for all advice and fast responses, and for providing new perspectives to the work.

Thank you

Mom and dad, for your love.
Martin B. Bull, for leading way.
Tante Wenche, for the wise talks.
Ingrid Lian, for always being here.
Family&Friends.

Norway, for 20 years of education.

This would not have been achievable without you all.

List of Abbreviations

RNA	ribonucleic acid
mRNA	messenger RNA
miRNA	micro RNA
ncRNA	non-coding RNA
mtDNA	mitochondrial DNA
SNV	Single nucleotide variant
SNP	Single nucleotide polymorphism
MSA	Multiple sequence alignment
GWAS	Genome Wide Association Study
WGS	Whole genome sequencing
UPGMA	Unweighted pair group method with arithmetic mean
UniFrac	Unique Fraction
1KGP	1000 Genomes project
H3Africa	Human Heredity & Health in Africa
ENCODE	The Encyclopedia of DNA Elements
nt	Nucleotide
bp	base pairs
GDR	Group diversity ratio
5’UTR	5’untranslated region
CPU	Central processing unit
MUSCLE	Multiple Sequence Comparison by Log-Expectation
NJ	Neighbour-Joining
CDF	Cumulative distribution function

Contents

1 Introduction

We are in the middle of a genomic revolution, with high-resolution genomic data availability by virtue of developments in high throughput in sequencing technology. There has been huge progression in human sequence data collection the last two decades, becoming readily available for medical research and development. However, research and databases are heavily biased towards Caucasian populations (6, 7). The 1000 genomes project, initiated in 2008, was one of the first large scale projects to address this issue and collected WGS data from 26 ethnic lineages across the world (5). Several other large-scale sampling projects have followed and are currently underway. Yet, ethnic disparities in the future of precision medicine, where an individual's genetic make-up lay the foundation for directed treatment and diagnosis, remains a problem (8-11). Variation across ethnic populations contain valuable information for scientific discovery, so it is important to evenly include humans of all ancestries in genomic research, and to investigate the impact of population specific variation.

Investigation of the human genome has revealed huge amounts of variants, with the majority of them located in the non-coding regions that are involved with gene regulation in complex networks (12, 13). Of particular interest is the microRNA (miRNA) gene regulatory network, which involves a large group of non-coding RNAs (ncRNA) involved with almost all biomolecular processes in all human living cells (14). We are far from understanding the role of all the detected variation in the human miRNA regulatory network, but recent research couples some of it to evolutionary adaptation mechanisms and to enforce robustness in the cellular system through buffering gene regulation (14-16).

The miRNA-mRNA interactome is a complex system that challenges current experimental and computational approaches. miRNA mainly functions through post-transcriptional repression of mRNA transcripts, where the miRNA primarily targets the 3'untranslated region (3'UTR) of mRNAs through partial complementary base-pairing interactions(14). The non-strict base-pairing rules aids targeting flexibility (17). Also, the miRNA-mRNA interaction

network is highly interconnected, with many miRNAs targeting each mRNA and each miRNA targeting many mRNAs (12). Most studies are focused on identifying "strong" one-to-one connections between a miRNA and a target mRNA that are associated with significant changes in miRNA or mRNA expression, with the goal of detecting a relationship between a miRNA and a potential phenotype (18). However, most targeting interactions are found to be "weak", where the phenotypic outcome of a single interaction is incremental and individually undetectable. For these reasons, the effect of a single variant in many miRNAs may produce an accumulative effect by which multiple small contributing effects can produce an observable change in in a cellular system. Interactions must therefore be studied together to understand how they may regulate genes collectively and pleiotropically.

To investigate the impact of variation in the miRNA regulatory network at the global scale for the human species, we need 1) large databases of human genomic samples and 2) appropriate approaches to detect patterns in this large-scale sequence data. There are still several challenges for both points, and the goal of this thesis is to explore solutions.

Variants in nucleotide sequences emerge with different probabilities that are related to genetic evolutionary mechanisms, and this should be accounted for in sequence comparison for research questions related to evolutionary patterns. Therefore, phylogenetic tree clustering algorithms that incorporate models that consider variation in nucleotide substitutions are a good starting point for this search. Phylogenetic tree clustering is an unsupervised approach, where clusters of input sequence data emerge based on their similarities and differences. However, there are few developed measures to investigate phylogenetic trees further. Usually, trees are visually inspected without any consistent quantification, which rapidly becomes an unfeasible approach as the number of trees grows.

In the search for variation across the whole miRNA - targeting network, we are here presented with a forest of phylogenies. The main quest of this project was therefore to develop and test a metric to quantify structure for predefined groups in each tree in the forest. Therefore, a novel phylogeny-based approach to detect variation patterns in large scale sequence data is developed and tested on ethnicity-associated variation in 3'UTR sequences. Each tree represents an RNA element involved with the miRNA-mRNA network, and each tree-leaf represents a human RNA sample. Specifically, the ultimate goal here is to utilize information from the phylogenetic trees to investigate if sequence variants in a miRNA - targeting network can be coupled to ethnicity, i.e., if the RNA sequences in a network can be

significantly grouped together according to ethnic origin, based on the present sequence variation.

The Group diversity ratio (GDR) was identified and developed to fit this aim. The GDR uses the branch length information in the trees and gives a measure of how well predefined sample groups are clustered together in the tree relative to the other defined groups. For example, in the context of ethnic population groups, one could ask: *are the sequence samples that originate from individuals of the same ethnic populations clustered close together in this tree? Or are they interspersed throughout the tree?* the GDR is presented here as a statistically testable metric that enables concise quantification of groupwise structure in phylogenetic trees.

There are many aspects to consider when introducing a novel methodology. In this work, significant effort was expended in evaluation of how the metric varies under different phylogenetic tree conditions by developing a simulation. Its result gives a clearer view of which criteria the GDR is most dependent on to be both appropriate and informative. Further, the GDR was tested in a real-world application by analysis of 9381 phylogenetic trees, for the set of human genes that are represented in the Reactome pathway databases. Each tree was made up of 2548 human 3'UTR samples that were collected as part of the 1000 Genomes Project and represented 26 distinct ethnic groups. Each 3'UTR corresponds to a specific gene; hence each tree represents a gene. This serves as a starting point for exploring variation in the miRNA - 3'UTR targeting network. Integrated software to calculate GDR for these of trees was developed for this purpose and is available on GitHub: https://github.com/norabbull/Willow2/tree/master/Willow1.0.

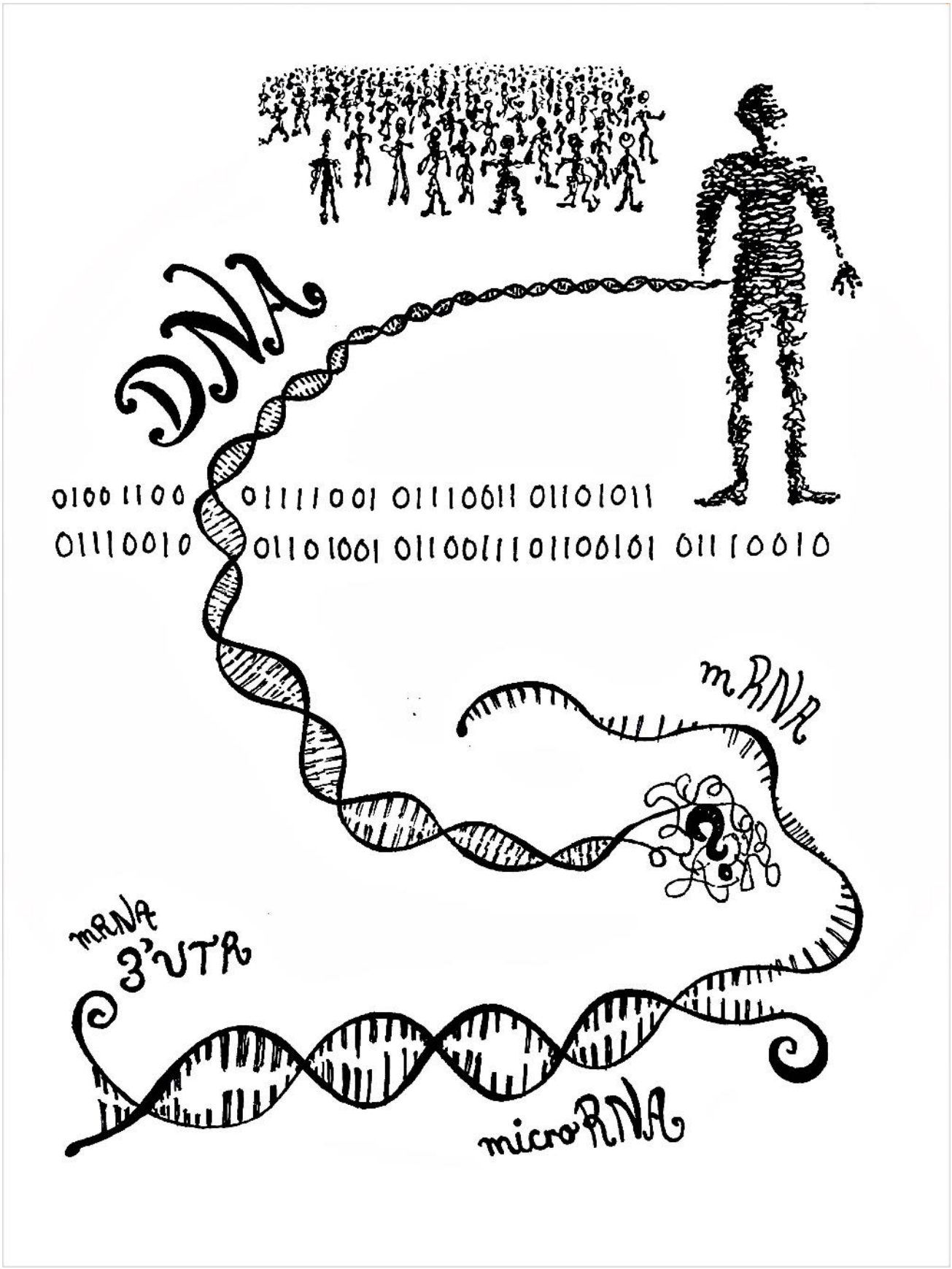


Figure 0: Illustrative figure of humans and miRNA-mRNA targeting interaction

2 Theory

2.1 Human genetic variation

2.1.1 Genetic diversity

All humans are estimated to share approximately 99.9% of the genome, but the remaining 0.1% is shown to encompass large variability (19). High amounts of variants are increasingly being detected in large-scale human genomic studies, with increased focus being given to investigation of the vast amount of discovered single nucleotide variants (SNVs) (20, 21). SNVs are the most common type of genetic variation and describes alteration of a single nucleotide (22). For example, a publication from the TOPmed project (The Trans-Omics for Precision Medicine) from this year (2021), aiming to map **genetic** variation and architecture in a wide range of disorders, reveals more than 400 million SNVs and indels after alignment of 53,831 whole-genome sequence (WGS) samples to the current human reference (GrCH38)(21, 23). 97% were found with frequencies of less than 1%, whereas 53% were singletons, meaning they were only present in one individual. Moreover, a study from 2017 estimated that all human individuals differ 0.5 % from the human reference genome (GrCH38)(24), giving reason to research the scope of human genetic variants in more detail. Thus, every genome is unique, collectively embodying valuable information.

A large proportion of the variation is proven to be significantly distinct on an ethnic level (11). This includes variation at single-nucleotide level to larger structural arrangement. Copy number variation (CNV), for instance, is found to be highly variable across Caucasian and Asians populations (25). Also, the higher genomic diversity found within African ethnic populations compared to other populations in the world reveals complex genetic population structures (26). More than 41,5 million novel SNVs unique to African genomes have been

recently reported, including higher levels of SNV heterozygosity. The recessive allele for sickle cell anaemia is the classic example of a population-specific variant that has been retained despite associated malady, since it gives natural protection against Malaria, compared to the homozygous dominant genotype. In turn, heterozygosity is increased in some African populations (27). “Finnish disease heritage” is another example that refer to the many monogenic diseases that occur exclusively within well-defined Scandinavian lineages (28). Thus, even the rarest SNVs can provide unique value for many aspects of medical research and tend to be population specific as a result of the most recent diversifying events (11, 29). It is argued that individual reference genomes for distinct ancestries should be developed, following the fast-moving efforts seeking to advance medical approaches where an individual's genome is utilized for medical diagnosis and treatment (24, 28, 30).

At a large scale, the gene pool is given resistance through a good mix of climatic and cultural events, where the human species’ insatiable need to trudge around the earth has left clear genomic evolutionary footprints. According to our “ticking” mitochondrial DNA (mtDNA) genetic clock, all human mtDNA sequences coalesce around 200,000 years ago and indicates an out-of-Africa bottleneck (31). Regardless of the exact timing, a genetic bottleneck event corresponds well to the higher diversity found within African populations compared to all other ethnicities (7, 31). Further migration patterns are difficult to both establish and interpret, conjectured as a crossroad of genomic exchange within and between species in the homo genus (31-33). For example, it is inferred that one group of pre-homo sapiens had already left Africa ~600, 000 years ago, migrating to both Sahul (Australian region), northwest (West Asia and Europe) and East, evolving to the distinctly classified species Australo-Papuans, Neanderthals and Denisovans respectively. Another group left ~70,000 years ago and interbred with Neanderthals in West Asia and Europe. A Neanderthal admixture of these ~50,000 years ago were the pre-ancestors of Eurasians, evolving to Europeans, East Asians, Papuans, and Aborigines. While the interbreeding network across time is far from solved, indicators such as the various percentages of neanderthal DNA for distinct modern ethnicities have left traces of our paths to the present genomic variation. (32)

More recent molecular evolutionary events have led to clear ethnic distinctions. The Finnish genomic distinctness arose, amongst others, from a founder effect that allowed a few genes to flourish from a small gene-pool (28). This can be explained by its Nordic location as well as cultural aspects such as religion and language, which isolated a small population and led to a present-day enrichment of specific gene variants. Another example are populations affected

by local adaptation through long cultural traditions of consanguine mating, enriching homozygosity (29, 34). The list of causal events that have occurred through human history is endless. These adaptations influence human genetic evolution and can impact cellular mechanisms in ways that are yet unknown. Currently, rapid globalization over the past century leads us into a new era of genetic exchange across all diversities, uniting the gene pool.

2.1.2 Ethnic bias in genetic Studies

Inequalities in economical, technological, and socio-political development leads to repercussions in the health sector (7). The imperative of catching broader scopes of human variants has important scientific and ethical consequences. The Eurocentric ethnic bias present in genomic database collections and research has gained increased focus over the past decade with argued causes such as convenience in data sampling at the geographical areas of performed studies and economical facets (6, 7). There is, however, rising effort to fill the information gaps and turn future research focus towards more holistic perspectives on human genetics to meet the new horizons of precision medicine (9-11, 30).

Since the human genome was first sequenced in the Human Genome project (HUGO) completed in 2003, sequencing technology has rapidly evolved and become sufficiently cheap and efficient to serve as an informational basis for most genomic studies. Detection of variation in all 3 billion base pairs in the human genome at single nucleotide resolution is enabled through whole-genome sequencing (WGS), and several databases are being built up (35). The Genome Aggregation Database (gnomAD) is the currently largest and most used public database for sequence variation in human populations, overall providing genome data from more than 195,000 individuals including 15,708 whole genomes(36, 37). Data from more than 60 studies is gathered in this common resource, overall performed by more than 100 researchers. Moreover, the UK Biobank recently published 200,000 entire genomes from British national health service (NHS) patients (November 2021) and is aiming to add 300,000 more genomes by 2023 (38). These projects aim to provide WGS and whole exome sequencing (WES) datasets through aggregation and standardization of large-scale sequencing projects. Another type of genetic database is, for instance, The Encyclopaedia of DNA Elements (ENCODE), which focuses on providing functional interpretation of all genomic elements and variational impact(39). Systematic annotation of both coding and noncoding

elements enables detection of rare variants even with low effect sizes, as well as new understanding of regulatory mechanisms in the non-coding regions. In turn, it lays the foundation for new medical research and improvements of human health as a global collaboration.

Despite the rising diversity in available data, focus groups of Caucasian origin are still heavily overrepresented in studies (6, 7, 11). For instance, The gnomAD database (previously named the Exome Aggregation Consortium (ExAC)) released in 2016 aggregated exome sequence data from ~ 60,000 individuals, with the aim to catalogue human genetic diversity, but the data were highly skewed in ancestry distribution (40). Their resulting database comprised 60.9 % Europeans, 13.7 % South Asians, 9.6 % Latino Americans, 8.6 % African Americans and 7.2 % East Asians. Most data were gathered from case-control studies with no thought for inclusion of population-specific sequence data despite under- or overrepresentations. Moreover, Europeans, Asians and Africans represented 81 %, 14% and 3% of all participants in genome-wide association studies (GWAS) respectively in 2016 (7). As the genetic foundation of individuals with European ancestry represents 16 % of the world's population, there is not a harmonious distribution of represented ancestries (6).

When biased genomic databases are utilized for new research, for example to search for patterns in new research contexts, the bias become inherited (11). Data originating from the current GWAS-bias gives a good example(6). In GWAS, a typical pursuit is identification of single nucleotide polymorphisms (SNPs) associated with a phenotype (41). SNPs are defined as single nucleotide variants present in at least 1% of the population (42) (hence, all SNPs can be termed SNVs, but not the other way around). SNPs can, for example, be used as new genetic markers that are further used to train polygenic risk score (PRS) algorithms (8, 43). These PRS algorithms provide a predictive measure of disease susceptibility based upon numerous genetic markers established as risk loci. Thus, with the current bias, the cumulative effect of the individual SNP genetic markers discovered in GWAS studies results in ethnically biased representation of genetic variants. This in turn can lead to incorrect interpretation in under-represented groups and medical treatment becomes optimized for European ancestries.

The bias is already identified in cases of medical treatment. For example, Warfarin is one of the most widely prescribed oral blood thinners in the US and is widely used worldwide (9). In a study of genotype-based dose Warfarin prediction, it was shown that exclusion of African American alleles led to significant dosing errors for this population. Also, PRS is already in

use for prediction of several cancer types, diabetes and heart disease, and shows promising results for a wide range of future precision medicine applications (8, 44). However, PRS are found to be significantly more accurate in assessment of European descents than of other ethnicities. This disparity can be hard to offset if treatment is further developed on a European-favoured database.

Fortunately, efforts are being made to counter this problem. The 1000 genomes project was the flagship project addressing the bias and includes genetic diversity across populations from all continents (37, 45). Their first release in 2015 contained 2,504 reconstructed human genomes spread across 26 ethnic distinct populations and was complemented by subsequent updates. The gnomAD database version 3.1 released in 2020 included data from more than 60 populations from all continents, with more than 3,000 samples chosen specifically to increase ethnic diversity (46). Around 52% of the 76,156 included genomes are of Europeans, whereas 27% and 7 % are of African and Asian ethnicities respectively, which represents a significant improvement from their first release. Other large-scale projects to particularly elucidate African and Asian genomics are also in progress. H3Africa (Human Heredity and Health in Africa) launched in 2012 is a broad collaboration envisioning to enhance all types of genomic research by, on and for African populations with particular focus on medical inclusion (47). Correspondingly, the GenomeAsia 100K Project mission is to map 100,000 individual Asian genomes, currently providing a WGS reference dataset from around 1700 humans spread across 219 population groups and 64 countries sampled in Asia (48). Lastly, focus is increasingly given to construct reference genomes for individual population groups world-wide. Amongst others, this is a focus of the TOPmed project, which highlights their goal to include ethnic diversity (21, 23). By this means, collaborations across the world with increased focus on inclusion of all human races gives hope for a future of equality. In the early stages of genomics research, results led to few practical utilities. Now, commercially available genetic testing to serve medical diagnosis and treatment is becoming a reality. Therefore, it is urgent to address the current bias and continue the work of world-wide genomic inclusion, as we explore new depths of genomic information and utilize it for further medical development. In this context, our non-coding genome is a particularly relevant topic, as it includes regulatory regions.

2.2 The human non-coding genome: miRNAs and 3′UTRs

2.2.1 The non-coding genome

Our linear DNA sequence is a human cookbook, where non-coding regions mainly embody the instructions to cook the ingredients (i.e., information necessary to express genes at the right time, place and levels). Every cell of an organism accommodates the full sequential genome, which as a fact clearly remarks the critical roles of the regulatory apparatus that control its active and latent parts. Variations in regulatory elements, mainly located in the non-coding regions, are widely found coupled to complex traits and disease (13, 41, 49, 50). Complexity of their regulatory network is far from fully understood, where most identified variants located in non-coding regions have unknown functions(39). New approaches are needed for better characterization of these regions to further our understanding of human genomic diversity.

"A lions share" of the human genome is non-coding, meaning it does not translate to functional proteins (51). The relevance and importance of this major portion of genetic information was long underestimated, but genetic risk factors found in these regions have been increasingly incorporated to phenotype-prediction studies in the past decade. 93% of disease-associated SNPs are found to be located in non-coding regions, including regulatory RNA regions linked to disease (13, 41) . Functional evaluation of complex traits in non-coding regions is for instance challenged by the "missing heritability" problem, referencing the concept that single genetic variations are generally unable to account for much of the heritability of disease (52-54). Exploring the non-coding regulatory regions is key to improve understanding of how human variants affect the process from genetic blueprint to (dys)functional phenotype.

Noncoding RNA (ncRNA) is a central concept in regulation of gene expression and functionality. In humans, noncoding RNA makes up ~98% of all transcribed material (55), where micro RNAs (miRNA), small nuclear RNAs and long-noncoding RNAs (ncRNA) constitute major groups. Specific functional roles of most ncRNAs are still unknown, yet variation in these molecules is commonly found to be connected to complex traits and

disease(54), such as cardiovascular diseases, cancers and a wide range of neurological disorders(56). Dysregulation of several miRNAs are, for example, shown in blood and brain samples of schizophrenic patients, which is a neuropsychiatric disease involved with multiple abnormal mental traits such as delusions(57). The human miRNA-mRNA interactome, encompassing all regulatory interactions between miRNAs and their mRNAs targets, represents a particularly comprehensive network of gene regulation(12, 13). The miRNAs and 3'untranslated regions (3'UTR) of mRNAs embody core regions of the network(14).

2.2.2 miRNAs

Almost every biomolecular process in all living cells involves miRNAs (14). These typically 20-22 nucleotides-short ncRNA elements have an active presence in the genomic regulatory network and are found to regulate most protein coding genes (3). A hairpin stem-loop structure is characteristic to all miRNAs, with their main function to post-transcriptionally suppress gene expression by inhibiting mRNA translation into a protein. They are found to be crucial in development, cell differentiation and homeostasis, and are often linked to disease (58).

miRNAs are predominantly generated through "the canonical pathway"(3). ~60-100 nt long primary transcripts (pri-miRNA) are processed into ~65 nt pre-miRNAs in the nucleus by the Microprocessor-complex (14, 58). A "classic" pri-miRNA contains single stranded RNA elements at both ends, a stem-sequence of ~34 bp and a terminal loop. These are recognized by the Microprocessor, complex functioning as a gatekeeper that segregates pri-miRNAs from other similar structures (58) . Mismatches are found to be tolerated in the stem loop to various degree, where special motifs can affect the processing. The produced pre-miRNA is transported into the cytoplasm and becomes a substrate for Dicer.

Dicer is a RNAse type III endonuclease that cleaves the pre-miRNA to produce a miRNA duplex (14, 58). Dicer often interacts with cofactor proteins where multiple positive and negative feedback-loops between Dicer and its products are also involved. The duplex is further integrated to a ribonucleotide complex called RISC (RNA induced silencing complex), that contains the Argonaut protein(s) (AGO). One of the duplex miRNA strands is retained in the complex to become the guide miRNA strand, that determines which mRNA the complex

subsequently targets. The other strand is degraded. Which of the strands that is retained as guide is determined by several factors, such as relative thermodynamic stability of the respective duplex strands and, in some cases, by AGO-proteins. Even the least favoured duplex-strand can be selected with a given frequency. The final mature miRNA complex is then ready to target RNAs.

Additional maturation pathways also exist. For example, they can be generated without a Microprocessor or Dicer (14). For instance, the catalytic mechanism is "replaced" by similar functionality of an AGO protein in the cytoplasm (2). Also, other non-coding RNAs than miRNA units, like tRNAs, can bypass Drosha processing and serve as precursors to Dicer. However, many of these components such as Microprocessor, Dicer, Drosha and AGO proteins exhibit other functions than miRNA maturation and are hence not necessarily evolved or optimized solely for this purpose (14). These alternative pathways and functionalities provide flexibility and increase the diversity of the generated miRNAs.

mRNA repression through partial complementary base pairing is the dominant function for a mature miRNA complex (14). The two main mRNA-fates include degradation through cleavage or silencing by destabilization or physical translation inhibition (59). The latter is reversible and is often a result of imperfect base-pairing, since it introduces bulges affecting the nuclease activity of Ago2. A single miRNA can regulate several hundred genes, and vice versa (3).

Thus, genetic variation is allowed through imperfect base-pairing in several of the steps, which introduces a further layer of variation. miRNA biogenesis encompasses complex multi-pathway networks where each step unit introduces flexibility.

2.2.3 3'UTRs

miRNA mainly target mRNAs through partial complementary base-pairing to the mRNA 3'UTR (14). The region delineates the end of an mRNA transcript, which is not translated into a protein, but rather contains information to direct and regulate translation of the mRNA (59). Many deeply conserved motifs and functions have been identified, even across species (59, 60). However, multiple isoforms exists and a large fraction of a 3'UTR is in general found variable (61).

3’UTRs vary in size, on average ~950 nt, but found to vary at different cell stages and for different cell types (59, 62). For example, they are lengthened during cell differentiation. Moreover, genes with tissue-specific expression exhibit in some cases longer 3’UTRs, with more miRNA binding sites, than non-tissue specific 3’UTRs (63). For instance, the average length for neural 3’UTRs is estimated to be ~1300 nt compared to ~700 nt for non-neural tissue (62).

3’UTRs have expanded in size during evolution, where a correlation between higher cellular complex organisms and length is identified. Their regulatory roles are thus suggested coupled to regulation of higher complexity of organisms, of which neural cells as components of the human brain play a central role (59). For example, it is found that long and short 3’UTR isoforms of brain-derived neurotrophic factor (Bdnf) and α-synuclein (Snca) are differentially expressed in activated and resting neurons, which are involved with controlling brain function (64). Whether this impacts miRNA target sites in particular remains unclear, but it has also been shown that miRNAs are found in especially high numbers in neurons, where it has become clear that they play a major role in all stages of neural development and fine-tuning protein expression levels (65).

Moreover, there are clear connections between variants disrupting 3’UTR target sites and phenotypes. To give but one example, in a study of variants in non-coding NGS data related to neurological disorders, miR-215 was found to significantly downregulate expression of the gene ARHGEF39 when its 3’UTR contained the reference allele rs72727021 (‘A’) (18). When this allele is swapped to the alternative allele ‘C’, repression by miR-215 is significantly decreased. The experiment was repeated for multiple cell lines including neural cells and controls by examining expression levels in vivo. The ‘C’ allele is associated with SLI (specific language impairment), where the gene expression level of ARHGEF39 gene is elevated compared to that of the reference allele. The same study identified more variants in miRNA target sites of 3’UTRs associated with complex mental conditions such as schizophrenia, bipolar disorder and autism. The authors of this study further suggest that examination of target sites in 3’UTR regions of candidate genes can help to identify some of the missing heritability of complex traits such as neuropsychiatric disorders. This example shows a direct connection between 3’UTR variants and disease.

Thus, these regions of the non-coding genome may hold potential for elevated understanding of the cellular regulatory network. C. Mayr argues in her recent paper "What 3'UTRs are doing?" that 3'UTRs are understudied and there is a scarcity of methods available to study their functions with regards to their broad scope of regulatory roles (59). To develop such relevant methodologies with regards to the miRNA targeting regions of 3'UTRs, the interactome between miRNAs and mRNAs must be understood in greater detail.

2.3 The miRNA - mRNA interactome

2.3.1 The intricacy of targeting interactions

The challenge of mapping miRNA-mRNA targeting networks is complicated by the many factors governing the flexible targeting rules (2, 3). Prediction through machine learning (ML) and deep learning approaches is at the front end for connecting the dots, but still relies heavily on the accuracy of experimental approaches(1, 4, 66). It is central to map targeting interactions in detail to enable recreation of the network algorithmically, although opportunity to gain understanding of network patterns through novel computational methodologies also exist.

A key element of the miRNA-mRNA interaction is determined by complementary base-pairing between the miRNA 'seed' region (nucleotides 2-8 from its 5'end) and the 3'UTR of mRNA, termed "canonical targeting" (2). However, the targeting regions and base-pairing rules extend far beyond these bounds, in fact, perfect base-pairing is rare in eukaryotes (2, 3). A preliminary (unpublished) analysis showed that approximately ⅓ of ~ 35 000 experimentally verified interactions are "non-canonical". For example, in several cases, base pairing between the 3'UTR and 3' half of the miRNA are found to create stability when seed-pairing interactions are weak (3). Also, binding sequences are found in the mRNA coding sequence and mRNA 5'UTR (49). The seed region is nonetheless shown to be important in stabilizing initial pairing engagement in the majority of targeting events (2, 3)

Moreover, interactions with other components such as proteins, transcription factors, epigenetic modifiers and RNP complexes are shown to be highly influential to the targeting

specificity (67). For example, alterations in AGO protein conformations post base-pairing in the seed region may influence interactions in the miRNA 3'region. Nonetheless, it is argued in the 2017 paper "Rules for functional microRNA targeting" that targeting rules reported in one study are rarely reproducible in another, which may be caused by un-definable targeting rules rather than lack of reproducible interactions between a miRNA and its target (17).

To understand how the miRNAs-mRNA interactome is involved in cellular systems, we need methods to identify the targeting sites with high precision. This challenge represents the hardest nut of all miRNA-related studies and is essentially rooted in the extensiveness of the targeting rule-multiverse (1, 68). Without a clear understanding of targeting rules, prediction of targeting interactions is shown to be a hard task for supervised machine learning approaches that make predictions based on learned data of targeting interactions (66). Despite increasing amounts of high-quality, experimentally verified interactome databases such as miRTarBase (69) and miRecords (70), no machine learning algorithm has demonstrated satisfactory across the full-scoped network (1, 66). Numerous attempts have been made with gradual improvement, where the common challenge remains the identification of a feature space that manage to statue all binding events correctly. In other words, it is difficult to spot a black cat at dawn, especially if the shape of a cat is strange. Speaking of cats, one could also think that targeting interactions for other species than humans would help in the data mining, but inter-species transferability is found to be highly variable and is of little help (1). Supervised deep learning (DL) approaches, which make decisions based upon artificial neural networks designed through trial and error, may aid prediction better since they do not rely on a defined feature space. miRAW, for instance, is a DL-based approach that is at front end of predicting miRNA-targeting interactions (71). Such algorithms could ultimately conquer the need for experimental validation and speed up the network-investigation process immensely when sufficient data become available, which is one of the reasons for this being a main current focus area of the field.

The current experimental approaches differ in accuracy and efficiency (68). The most common per se is co-expression analysis, measuring levels of mRNAs in conjunction with miRNAs over- or under-expression in tissue-cultured cells (1, 68). This method is vague in determination of exact binding sites. Moreover, it is reductionistic since it omits other potential influencers of expression level alterations, such as indirect miRNA regulators. Also, the experimental setting may affect the targeting for unknown reasons. Other methods include

immunoprecipitation- and pull-down approaches, which exhibit higher accuracy, but are still inefficient for technical reasons (68). Pull-down methods that involve creation and detection of biotin-miRNA-mRNA complexes, such as CLIP and CLASH, are argued to be the most promising strategies so far(72) (1, 68).

Emerging insights of miRNA-mRNA interactome moves us towards higher understandings of regulatory mechanisms of the network. With developing approaches to solve the interaction-puzzle both experimentally and computationally, we can further explore how variation in the network functions and influences gene regulation both presently and in a larger, evolutionary perspective.

2.3.2 miRNA-mRNA network variation and regulatory functions

In the early stages of miRNA network research, leading research reported significantly low amounts of SNVs in miRNAs (73). Publication of the 1000 genomes project overturned this findings, as enormous amounts of variation in these regions were discovered (45, 73), supported by other following large-scale projects incorporating ethnic diversity(13, 23, 47, 48). Naturally occurring variation may, for instance, be permitted through the liberal targeting rules of imperfect base-pairing (2). However, at the same time, many microRNAs are found to be deeply conserved (3, 67). Of all human genes, more than 60% embody at least one conserved miRNA target site, which indicates vital regulatory roles across evolution. The duality of the high amounts of variation and necessary conservation in this network, however, raise questions of the functionality of the variation itself, why it is “allowed” into the system and how it may influence the wide span of regulatory mechanisms of miRNAs?

miRNA studies often focus on strong-repression interactions between miRNAs and phenotypes, in search of a clear connection to explain a single trait (13). Most of these studies are narrowly scoped, where candidate genes are preselected and measured in relation to one or a few specific miRNAs, often through expression level analysis (12). This can indeed give valuable medical insights, however, a down-side of these types of analyses seen in the bigger picture is the accumulation of experimental data on single strong-repression interactions. Less experimental data is gathered on the weak-repression interactions, which are in fact most common (15). Weak interactions ensue less obvious phenotypic outcome and would therefore exert functional changes in a more complex fashion (12, 13, 15, 16). It is thus less

straightforward to design studies of weak-repression interactions, requiring new ways of thinking. The skewed pool of experimental data could also lead to erroneous training of prediction algorithms in supervised ML-methods.

A recent study of 1000 genomes project data showed 44% of genomic regions encoding for miRNAs are affected by SNVs(73). SNVs in the highly interconnected miRNAs-target network are presumed to cause pleiotropic effects(12, 13). This means that they cause a fractional difference to the phenotype through, for example, small enhancement or diminution of miRNA expression. This study also estimated a loss-of-function (LOF) SNVs ratio of 0.05 per miRNA gene, 3.5 times higher than of protein-coding regions (73). These facts complicate discrimination between pathogenic and non-pathogenic variants in the targeting network. Therefore, methodologies that incorporate variation at a larger scale, including both weak and strong interactions, can lead to better predictions of how they collectively affect phenotypic outcomes and how it may cofunction at larger scales.

Multiple cellular functionalities are connected to this variation(13, 15). For example, the variation is associated with 'robustness' mechanisms, i.e., creating a broader set of opportunities for new environmental adaptations at the genomic level (15, 16, 74). It is for instance becoming clear that miRNAs are involved in stress-response(13, 74). Variation in miRNA stress-response systems may not present phenotypically under normal conditions but come to light under stressful conditions (13, 74). As described by Jie Du et al. (2019), miRNAs are prevalent in both local regulation and the whole organism for almost all kinds of stress responses including trauma, surgery, burns, radiation and toxic shock. In a similar manner, latent variants that are hidden under normal conditions can become critical during environmental changes, putting molecular selection pressure onto the advantageous variants. In this way, accommodation of variation is proposed as a driving property of evolution, combining stochastic variation and interaction with the environment.

It starts to become clear that variation in the system has its causes **and** effects and perturbs weak interactions that may be connected to driving mechanisms of evolution at the molecular level. To proceed with this research hypothesis, it is necessary to shift from investigating single, strong interaction and map large amounts of variations involving weak interactions. Furthermore, to capture this variation we need data that is representative of global human diversity at the genetic level. Finally, to investigate sequence data from a diverse set of human

populations, it is necessarily to develop new methodologies for large-scale pattern searching and analyses. In turn, finding population-specific variation structure in, for example 3'UTRs, can aid the search of variants that must be carefully considered in development of medical approaches.

2.4 Phylogeny

Phylogeny is an umbrella term describing the study of evolutionary relationships among a group of organisms. These relationships can be based on differences of shared characteristics such as anatomical or genetic. These differences are commonly represented using phylogenetic trees, where more similar organisms are closer together in the tree. In this study, the focus is on molecular phylogeny, which characterises organisms based on differences in their DNA or protein sequence.

2.4.1 The concept of phylogenetic trees

A phylogenetic tree represents an estimation of biological relatedness among a group of studies samples. In this representation, more similar sequences are assumed to have evolved from a more recent common ancestor. Estimating similarity among a group of sequences is not a trivial task. Since ground truth is unknown, unsupervised methods are required. Many different algorithms have been developed on different criteria, but each need to seek a balance between computational requirements and achieving an optimal estimate of phylogeny (75). For example, UPGMA (unweighted pair group method with arithmetic mean) is a distance-based algorithm using agglomerative hierarchical clustering(76). It constructs sets of trees where the most likely tree given the data is selected, i.e., *max P (tree | data)*. However, for large datasets, an exhaustive search of all possibilities, including highly unlikely resemblances, is too CPU- and memory intensive and not feasible. Thus, most estimators are heuristic based. The Maximum Likelihood algorithm identifies in the tree giving the largest probability of observing the input data sequences, i.e., *max P (data | tree)* (77). The Neighbour-Joining (NJ) method is a third, distance-based algorithm, described in more detail below since the forest of trees analysed in this project were generated using this approach.

We assume that our measure of evolutionary distance between two sequences is related to their sequence diversity, but *how* the variation occurred is another story. For example, because of physicochemical properties of the nucleotides, transitions (purine-purine or pyrimidine-pyrimidine mutations) are more energetically pertinent than transversions (purine-pyrimidine mutations), increasing their likelihood to occur (78). Also, more transitions than transversions are more likely since fewer amino acids are changed by the former. A third artifact is reverse mutation. These are impossible to predict precisely but estimated rates may account for their likelihood. All these mechanisms need to be considered to return estimates of evolutionary distance more likely to resemble the truth.

These mechanisms are captured in nucleotide substitution models which are integrated in tree-clustering algorithms (75, 78). UPGMA, for instance, assumes constant rates of evolution, which is often violated and is thus of little use. Kimura distance is a more used substitution model which includes two substitution parameters: transitions and transversions (79). Nucleotide frequencies and substitution rates beyond these are assumed equal. Parsimonious algorithms, on the other hand, postures the simplest explanation as the preferred arrangement (75). For sequence data, it means to assume that the relationship requiring the fewest mutations, best explains their relationship, ignoring other evolutionary mechanisms. Hence, only current-time, observable substitutions are counted, and evolutionary sequence history is omitted. These are also widely used in tree-estimations, but their results are considered less reliable in estimation of evolutionary distance. Depending on the research topic and genomic regions of interest, deliberation of suitable models is an important step to produce sufficient yet reliable phylogenetic trees.

2.4.2 The tree outline

Estimation of tree topology and branch lengths comprise the essence of a distance-based phylogenetic tree building. A tree is branched at several points, called nodes, leading to either more internal nodes or leaf nodes, which represent the clustered samples as tree branch tips, see figure 2.1. A bifurcating tree has a maximum of two branches leading out of each internal node. A tree can either be rooted or unrooted, describing whether all nodes are derived from a single common node or not. The distance between a root node and leaf nodes makes up the largest evolutionary (linear) time distance in in the tree. For unrooted trees, however, no single node necessarily connects all leaf nodes, hence there is no guarantee for an estimated last common ancestor connecting all leaf nodes. The branch length distance between the leaf nodes represents the estimated evolutionary distance between the sequence samples. Branch lengths can be constructed proportional to their estimated evolutionary distance (number of nucleotide substitutions), resulting in scaled trees. Unscaling is the other option, where branch lengths are un-proportional to evolutionary divergence.

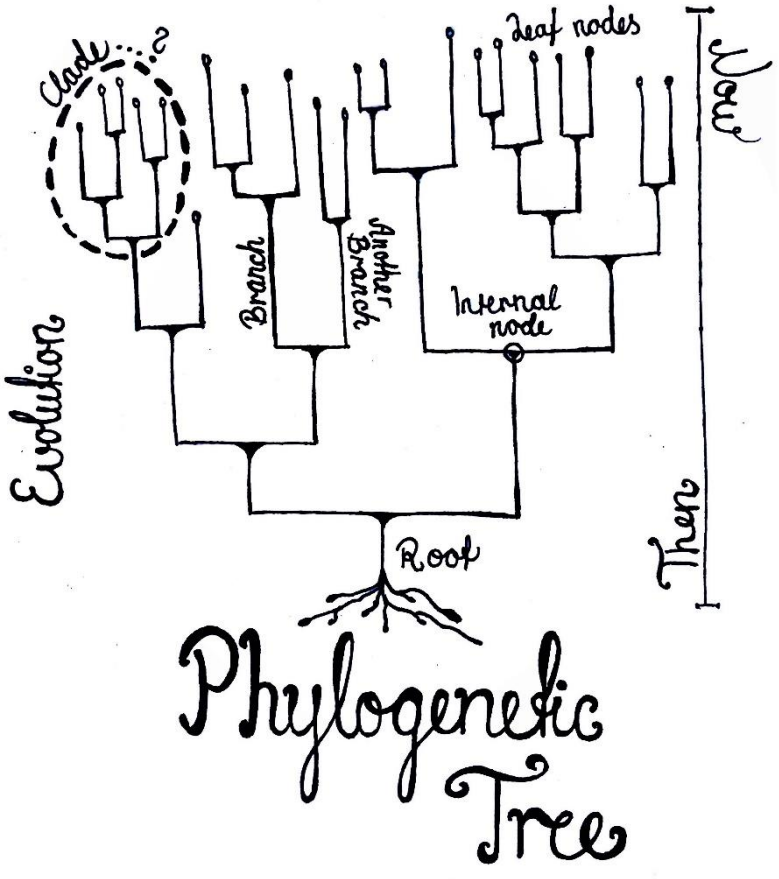


Figure 2.1: Illustration of a bifurcating phylogenetic tree. The tree is rooted with 19 internal nodes and 21 leaf nodes.

2.4.3 From sequences to phylogenetic trees with MUSCLE and NINJA

Multiple sequence alignment (MSA) is the first step towards a phylogenetic tree, casting how comparison of sequence data should proceed. The sequences are assumed to share a common ancestor, which makes them comparable. In general, MSA is a computationally heavy job where optimized algorithms achieve shortcuts in various ways and the result is no guaranteed true solution. MUSCLE (Multiple Sequence Comparison by Log-Expectation) is a currently popular software tool for this procedure (80). It is fast proceeding, being advantageous in alignment of a high number of sequences. Both input and output sequences are delivered in FASTA-file format.

MUSCLE first computes a quick draft alignment based on a *k*-mer distance, that is, the number of common short sub-sequences of length k shared between two sequences (80). These are counted and used to calculate a distance matrix, for which an intermediate binary tree is estimated. This is a graph-based data structure representing the likelihoods of alignment arrangements. The alignment is further improved progressively (sequences are aligned one after the other, starting with the most similar) based on the Kimura distance metric. In this step, the initial distance matrix and the binary tree is re-estimated. Finally, the improved alignment is refined, based on this re-estimation. Accuracy is increased through iterations of this process, including binary tree re-estimation and refinement. The MSA serves as input to the clustering procedure in FASTA-format.

The NJ clustering algorithm transforms pairwise sequence sample variation into a 2-dimensional numerical matrix reflecting pairwise evolutionary distances(81). This can be seen as a conversion of qualitative data into a quantitative measure of sequence relationships. Its benefits is its fast performance but produces trees with low precision with regards to divergence times. NJ allows unequal rates of evolution, resulting in a tree with branch lengths proportional to the amount of sequence variation in the data.

The NJ algorithm is an agglomerative (bottom-up) clustering approach, iteratively clustering the closest nodes together in the tree (81). In an iteration, for example as illustrated in figure 2B, a new internal node I1 is created between leaf nodes N1 and N2, as they share the shortest distance between any pairwise nodes in the tree. Distances between the node I1 and the other nodes in the tree are recalculated before the process starts over again. In the new iteration, node I1 represents the conjunction of N1 and N2, meaning only distances between I1 and the other tree nodes are accounted for in this iteration. Iterations run until all nodes are processed by the algorithm. The NJ algorithm is implemented in the NINJA software program which was used to generate the trees used in this project.

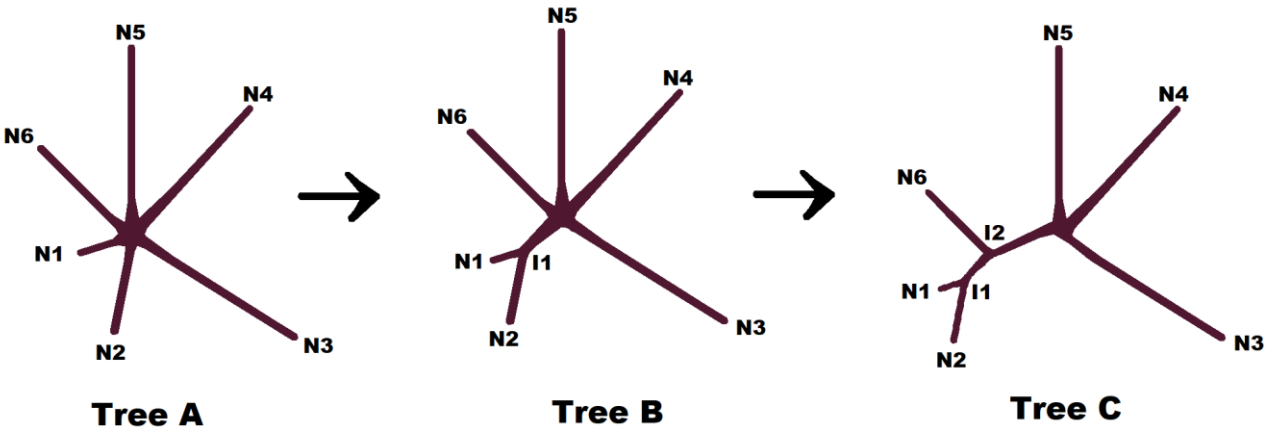


Figure 2.2: Illustration of neighbour-joining algorithm. Arrows illustrate iterations. In the first iteration, the leaf nodes N1 and N2 in tree A are merged to share the new internal node I1, shown in tree B. Branch lengths from I1 to the other leaf nodes are re-calculated, before the next internal node I2 is generated, as shown in tree C.

2.4.4 Phylogenetic inspection and evaluation

Traditionally, phylogenetic trees are used to investigate biological relationships at small scales. The tree-shaped representation enables a visual, intuitive interpretation, which is convenient to get an overview of how each sample/leaf node is evolutionary related to the others. Based on the criteria that samples are clustered close together in the tree, leaf nodes can be delimited into groups. This way, phylogenetic trees aid the *identification* of groups of evolutionary related samples, without pre-knowledge of the relationship between the observations (hence the unsupervised clustering approach).

There is no general rule to how clusters should be defined since interpretation of the cluster result is context dependent. Optimal partitioning of grouped samples in a tree, however, is often thought of as the branch cut-off that provide maximally separated clusters of samples (82). A common approach is to visualize the result with a dendrogram and make a visual branch cut-off that separates leaf nodes into distinct clades, with respect to a meaningful interpretation of the data that is clustered. Another, more mathematical approach, is presented in the paper "Separating Disambiguation from Composition in Distributional Semantics" by D. Kartsaklis et al., where clusters are defined based on the Variance Ratio Criterion (VRC) metric. In brief, VRC measures the compactness of clusters through inter-cluster and intra-cluster variances. Thus, minimizing the VRC score for a tree can facilitate group definitions. The classified groups can then be used to identify features important for categorizing that specific type of data, or the groups themselves can serve as valuable information.

In opposite of the traditional approach, where phylogenies are used to identify groups, some methods have also been developed to measure group structure for pre-defined sample groups in a tree. Within the field of microbiology, phylogenies are widely utilized in this manner through the UniFrac metric. This is explored in more detail in the next section to understand the relationship between the GDR and the UniFrac metrics.

2.4.5 UniFrac

UniFrac (unique fraction) is a metric widely used to determine distinctiveness of sequences in two different biological samples and is based on phylogenies (83). It is often used to distinguish communities of taxa, where samples are gathered from distinct environments, which make up distinct groupings. The target question is then whether the taxa present in these defined groupings are significantly different. If that is the case, they can be classified as distinct communities. The taxa may appear in more than one of the measured environments, and to various degrees. There are two main versions of the measure: weighted and unweighted (84). The weighted UniFrac accounts for sample abundances and is thus less sensitive to sample abundances in the groups, whereas unweighted only considers presence/absence of samples in the groups. Overlaps between groups are allowed in calculation of the UniFrac measure.

Phylogenetic trees are constructed through approaches described above, for example from samples of 16SrRNA in the case where microbial taxa are studied, which is the gold standard genetic marker to confer a measure of evolutionary relatedness of microbial taxa (85). UniFrac measures evolutionary distance between sets of the taxa in a phylogenetic tree, originally described in terms of "the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both", i.e.,

$$(unweighted)\ UniFrac = \frac{sum\ of\ unshared\ branch\ lengths}{sum\ of\ all\ tree\ branch\ lengths}$$

In this manner, the approach quantifies phylogenetic tree-structure based on unique branch lengths leading to respective environmental groups of taxa. It is calculated for pairwise combinations of environmental groups and results in a distance matrix of UniFrac-fractions. P-values for this quantification can be calculated by comparing to the distribution of UniFrac values obtained from the same tree after a randomization process, where the UniFrac metric is calculated for the same phylogeny with shuffled leaf node labels, i.e., the taxonomic group a leaf node belongs to or represents. In this way, the UniFrac measure can reveal statistically significantly distinct groups from a phylogeny.

3 Aim

There are two main goals within the scope of this work. One part involves development of a method, whereas the other involves the utilization and exploration of the potentials of the method in a case-study.

Specific/detailed Aims:

1. Find a metric to quantify large-scaled patterns related to predefined groups in a grand forest of phylogenetic trees. The metric must signal an overall groupwise-tree structure rather than providing a detail-oriented analysis. Thus, the metric aims to facilitate and direct further research, such as functional analysis and interpretation of the detected sequence elements that show groupwise structure. Properties of the metric include:
    a. The metric must be statistically testable
    b. The metric must utilize information from phylogenetic trees, i.e., evolutionary distance estimations conceived as branch lengths in a phylogenetic tree topology
    c. The metric should be applicable for large-scaled genomic sequence data. Minimally, it must be calculated for at least as many trees as provided for the case-study, but preferentially capable of handling more, within reasonable timeframes. The minimum is thus 9381 trees with 2548 leaf nodes each, as investigated in the case-study.
    d. The metric should enable quantification of group-structure related to an optional number of groups.

e. The metric should be robust against a few misclassifications in the phylogenetic tree estimation, to meet prerequisites of the full approach (from sequence data to phylogenetic trees). Since it was aimed to estimate structural patterns in large-scale genomic sequence data, the pre-procedure benefit from utilizing fast sequence alignment tools and fast phylogenetic clustering algorithms. Hence, the metric has to be suited for trees acquired through such methods.

f. Development of software to make the procedure high throughput would be additionally advantageous for the case-study and future research

2. Case-study: Utilize the metric to search for ethnicity-associated patterns in human 3'UTRs.

- The metric was aimed to measure population structure for 9381 phylogenetic trees, each constructed from 2548 human 3'UTRs. Each tree represented one specific 3'UTR element that was segmented from a gene.
- Ethnicity-associated patterns were aimed to be searched at two population-group levels: "Super" and "sub", containing 5 and 26 population groups respectively. Hence, the goal was to search for ethnic patterns in the human genome at two ethnic lineage-scales, which could be couplet to human evolutionary diversifying events at two time-scale levels
- Potentially detected patterns were aimed to be explored in larger contexts of human evolutionary history and adaptation mechanisms in association with 3'UTRs.
- The case-study was aimed to demonstrate the potential of the developed methodology for further investigating the miRNA-3'UTR interactome in particular, in light of its complex network structure
- Moreover, this aim included to address the ethnic bias in genomic databases, by 1) researching a diverse set of ethnic populations and 2) researching ethnicity associated patterns in the non-coding region that would underpin the importance of ethnic genomic inclusion in future studies

4 Data

The initial dataset explored in this study was pre-processed from raw-data as part of another study. In brief, this pre-processing involved extracting 3'UTR sequence data for all population samples in the 1000 genomes project and performing multiple sequence alignment for each 3'UTR as input for estimates of phylogenetic trees. The produced outcome was 9381 phylogenetic trees in Newick-format that was the input data used in this study. These steps are described in more detail below.

4.1 The 1000 genomes project

The 1000 genomes project (1KGP) was the first comprehensive attempt to obtain a catalogue of human genetic variation (45). The project aimed to map sequence variation in the human genome across distinct ethnicities and was completed in 2015, although there continue to be subsequent updates in the form of higher read coverage and additional samples. Its final phase ("phase three") was completed with variant calls for a total of 2548 sequenced individuals, originating from 26 ethnic populations, hereafter termed sub-populations. These sub-populations are further grouped into 5 super-populations based on continental ancestry, see table 1. DNA samples originated from lymphoblastoid cell cultures, established from fresh blood. All individuals were whole genome sequenced (WGS) with mean depth of 7.4 × and 30 × coverage, as well as targeted exome sequencing (WES) with mean depth 65.7 ×. For some individuals, variants and haplotypes were additionally detected through high-density SNP microarrays.

Table 4.1: The number of samples from each population included in the 1000 Genomes project dataset

Super-population	Sub-population	# Samples
Africa (AFR)	Yoruba in Ibadan, Nigeria (YRI)	107
	Luhya in Webuye, Kenya (LWK)	103
	Gambian in Western Divisions in the Gambia (GWD)	113
	Mende in Sierra Leone (MSL)	90
	Esan in Nigeria (ESN)	100
	Americans of African Ancestry in SW USA (ASW)	61
	African Caribbean in Barbados (ACB)	97
	Total:	671
Europe (EUR)	Utah Residents with Northern and Western European Ancestry (CEU)	99
	Toscani in Italia (TSI)	111
	Finnish in Finland (FIN)	105
	Iberian Population in Spain (IBS)	107
	British in England and Scotland (GBR)	100
	Total:	522
East Asia (EAS)	Han Chinese in Beijing, China (CHB)	106
	Japanese in Tokyo, Japan (JPT)	105
	Southern Han Chinese (CHS)	105
	Chinese Dai in Xishuangbanna, China (CDX)	100
	Kinh in Ho Chi Minh City, Vietnam (KHV)	99
	Total:	515
Ad Mixed Americas (AMR)	Mexican Ancestry from Los Angeles USA (MXL)	64
	Puerto Ricans from Puerto Rico (PUR)	104
	Colombians from Medellin, Colombia (CLM)	95
	Peruvians from Lima, Peru (PEL)	85
	Total:	348
South Asia (SAS)	Gujarati Indian from Houston, Texas (GIH)	106
	Punjabi from Lahore, Pakistan (PJL)	96
	Bengali from Bangladesh (BEB)	86
	Sri Lankan Tamil from the UK (STU)	102
	Indian Telugu from the UK (ITU)	102
	Total:	492

The 1KGP's goal was to capture 95% of the common human variants, primarily to uncover information for medical research objectives. Sampling selection was therefore based on examination of divergence within each defined population and focused on large populations rather than smaller, isolated ethnicities. The amount of divergence within the group of samples originating from Europe, East Asia and South Asia was estimated sufficiently low to capture most rare variants in these geographical areas, with an estimated fixation index (F_{ST}) of about 1%. F_{ST} is a measure of heterozygosity reduction in one population compared to that of a population with random mating with the same allele frequencies (86). For Africa, on the other hand, the degree of differentiation amongst sub-populations was estimated too high to have captured the aimed amount of variation, which is more comprehensively researched in H3Africa studies(47). In the Americas, the two populations ASW and CEU were shown to have mixed ancestry from both Europe, Africa and indigenous American, and were relocated post sampling from the American group to the African and European groups, respectively. For this reason, there are only 4 sub-populations in the American super-population, and an additional sub-population in both Europe and Africa. This leads to a slight imbalance in sampling records between the super-populations. Moreover, the number of samples from each sub-population spans from 61 to 113 samples, although most populations comprise more than 90 samples. See table 1.

Originally, the 1KGP was based on the human reference version GRCh37. Variation in phase 3 data was previously called through lift-over approaches via alignments created between the GRCh37 and GRCh38 references by dbSNP (86). A total of more than 88 million variants including single nucleotide variants (SNVs), insertion/deletions (INDELS) and larger structural variants, were called in this set. 81.4 million of these were reported as SNVs, of which 99.6% were reported as biallelic. Over 99% of the SNVs were observed to be present with a frequency of more than 1% for several ancestries. However, because of the uncertainties in this process, such as unmatching mapping regions between the assemblies, "simpler" multi-caller approaches were used to perform de novo calling of SNVs and INDELS on the GRCh38 primary assembly of autosomes and pseudo autosomal regions, performed by Lowy-Gallego E, Fairley S, Zheng-Bradley X et al (87). Their variant call set was highly consistent with the lift-over call set and additionally included biallelic INDELS and mapped variants in areas of GRCh38 that were absent in GRCh37.

The GRCh38 de novo variant call set containing biallelic SNVs and INDELS from the 2548 samples were used for analysis in this project (87). The variant calls were available as VCF

files in two forms: single genome files and chromosome files. The presence of all SNVs for all 2548 individuals were only available in the chromosome files, hence these were utilized. The samples were only defined by sample names in the chromosome files, thus an additional TSV-file containing population information was cross-referenced to the sample names during tree parsing (described below).

4.2 3′UTR extraction

For this work, the primary interest was in non-coding variants for the 2548 1KGP samples connected to the miRNA-regulatory network. The focus was therefore focused on variants occurring within the 21458 3'UTR sequences identified in the ENSEMBL annotation for the GRCh38.102 release. This 3'UTR set was further filtered to only retain entries that were included in the Reactome pathway database, to aid interpretation of any results in a functional context in downstream analyses. Reactome.org is an open access resource providing extensive information on biological processes and pathways. A total of human 2478 pathways were included, referencing 9416 genes at the time of extraction. Pathway information is stored in the systems biology markup language (SBML), and each gene is identified by its UniProt gene ID. UniProt.org is another open access database that provides high quality information related to proteins and their function.

A first step was to obtain the sequences in FASTA format for the 3'UTRs corresponding to these 9416 genes. UniProt does not directly provide location information for entries, but this can be obtained by accessing a map file that provides cross references to related information in other databases. Using this data, the corresponding ENSEMBL entries for each gene were accessed. ENSEMBL is a comprehensive database of vertebrate genome information and includes gene annotations for the human genome and corresponding transcript information (including 3' UTRs). For each UniProtID, the corresponding gene entry in ENSEMBL (release GRCh38.102, as identified in the map file) was obtained (identified by a ENSG ID) together with the matching set of transcripts (identified by their ENSG/ENST IDs). This transcript set was then filtered to select the 3'UTR entries. If multiple entries were found, the longest entry was selected. All 3'UTR annotations had transcript support level (TSL) =1, meaning the annotation had high confidence. The genome locations of this 3'UTR set was

then intersected with the 1KGP VCF files to obtain the SNVs located within these regions which were then used to generate FASTA sequences for each sample (using the standard FASTA reference file for release GRCh38.102). Finally, a sequence set of 9416 x 2548 3'UTR sample sequences were generated, containing called variation information for all sampled individuals in the 1KGP.

4.3 Multiple Sequence Alignment

Multiple sequence alignments (MSA) were built from the 2548 human samples for the 9416 3'UTRs with the MUSCLE alignment software tool. MUSCLE achieves better computational speed and average accuracy than other common aligners (ie. ClustalW and T-Coffee)(80), which was considered important in this context because of the high number of sequences that needed to be aligned. The generated 3'UTR FASTA sequences were given as input and run with default software parameters. The default number of iterations is 16, however, it stops upon reaching convergence (likely to occur after a few iterations for 3'UTR sequences). The generated alignments were considered robust.

4.4 Clustering

Neighbour-Join trees with Kimura substitution model were built from the alignments, resulting in 9381 trees in Newick-format. This methodology was selected for the aim to identify larger structural patterns rather than more detailed evolutionary relationships between individual sequences (i.e., detecting groups of ethnicity-specific samples, rather than smaller variant distinctions within a population). Moreover, using the NJ method had the advantage of increased efficiency compared to more detail-oriented algorithms, which was necessary because of the magnitude of the clustering jobs.

4.5 Meta – data

For each phylogenetic tree, additional information about the unique sequences that the phylogeny was built from was provided. This included which samples (including their ethnic origin at both super- and sub population level) that had the same unique sequence. From this, the number of unique sequences that made up each tree could be counted and the population distribution across the unique sequences could be retrieved.

5 Methods

5.1 Introduction

The core aim was to find a quantitative measure of clustering structure for self-defined sample groups in phylogenetic trees. After evaluation of multiple potential metrics, the group diversity ratio (GDR) was selected and developed. The method is first described in general terms, including how a simulation of it was performed. Then, the pipeline that implemented the GDR to search population-specific variation patterns in 9381 phylogenetic trees, each representing a human 3'UTRs, is introduced. Finally, the Python software "Willow" was developed as an integrated environment for high-throughput calculation of GDR from phylogenetic distance matrices. The program is introduced and described in brief here, and the program and documentation can be found on: https://github.com/norabbull/Willow2/tree/master/Willow1.0.

All program and package/module versions used in this project are listed on:

https://github.com/norabbull/Willow2/tree/master/Willow1.0/packnameList.csv

5.1.1 Terminolgy

The GDR, or group diversity ratio, is a novel cluster-structure measure inspired by the subtype diversity ratio (SDR) metric and its application in the article "Phylogeny and the origin of HIV-1" by Rambaut et al., published in Nature in 2001 (88). They used it in a different manner and to answer another type of research question than investigated here, but the concept is built upon the same mathematical principle. The metric is renamed from SDR to GDR to better reflect its novel application in this work. Thus, the two terms share the same mathematical definition, but GDR is the term used throughout the thesis.

5.2 Phylogenetic cluster dispersion measure

In contrast to traditional approaches of inspecting and interpreting phylogenetic trees visually, it is here aimed to utilize phylogenetic tree information to quantify the distinction between predefined groups of genomic sequence samples. Evolutionary patterns in the sequence data are recognized by the unsupervised clustering algorithm, but to which degree these patterns correspond to the predefined groups is the key question. To clarify, the goal is **not** to **identify** groups; here, groups are aleady specified. See figure 5.1 and 5.2.

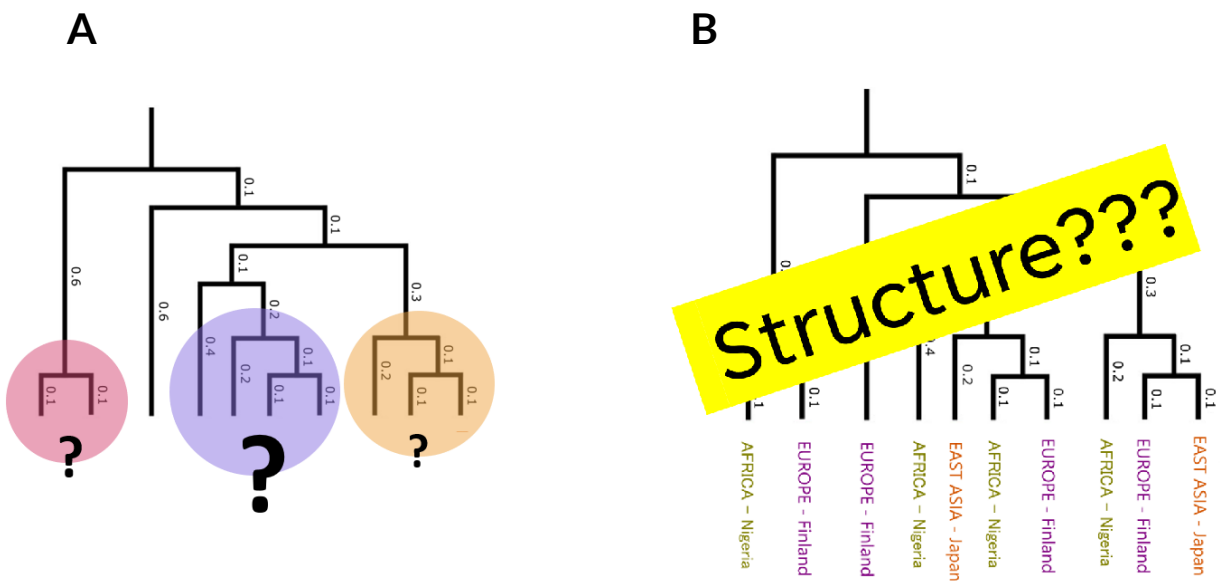


Figure 5.1:

A: In traditional approaches, a phylogenetic tree is inspected visually to identify groups. In this illustration, the coloured circles show groups that could be identified, based on the short evolutionary distance between their leaf node pairs relative to other leaf nodes in the tree.

B: In this study, it is questioned whether samples from predefined groups are significantly more closely clustered in the tree compared to the other predefined groups in the tree. In this illustration, Europe, East Asia and Africa make up predefined sample groups. The GDR can quantify to which degree the groups are clustered together in the tree.

Sample data are clustered through a tree-clustering algorithm, where the resulting phylogeny is based upon the best or most likely arrangement of the data samples according to

thei(evolutionary) similarities and dissimilarities. Defined groups, on the other hand, represent another sample feature, hidden from the clustering algorithm. Thus, the required information for this study is:

1. A phylogeny of clustered genomic sequence data samples
2. Meta-data of all samples, representing one or more grouping feature(s)

Formally, the metric to quantify a tree structure must be suited to utilize phylogenetic data as input, i.e., branch length information, and be applicable for statistical testing. Since a further goal was to make the procedure attainable for large amounts of data, i.e., high number of phylogenetic trees, another criterion was computational efficiency. To meet this criterion, both algorithmic simplicity as well as programmatically optimization had to be considered. The GDR metric was found to meet all criteria and is intuitively and mathematically pertinent. It is described in detail in the "GDR" section below.

5.2.1 A GDR-application example

Quantification of the relationship between phylogenetic tree structure and presumed groups can give a statistically verifiable measure of a potential relationship between the source of the sample variation used for clustering and the group feature. An example scenario where this would be useful is where disease subtypes are suspected to be coupled to variation in the amino acid sequence of a protein. A phylogeny is constructed from a representative number of sequence samples. The disease subtypes are the defined groups, where all sequence samples of the phylogeny can be grouped into one of the subtypes (which must be pre-determined information). The GDR for the defined subtype groups is estimated for the phylogeny and will provide an indication of how well the amino-acid sequence samples are clustered relative to the disease subtypes. If a "strong" groupwise tree-structure is implied, it could be interpreted to support the presence of a relationship between variations in the amino-acid sequence and the disease subtypes. If no structure was found, it would indicate that the amino-acid sequence is unrelated to the disease subtypes.
Moreover, if multiple proteins were suspected to be involved in the disease, one could construct phylogenies for all proteins and measure their tree-structure in relation to the disease subtypes. The search for a significant protein-disease relationship signal for a large number of

proteins could be performed simultaneously. In this way, it would be much faster to both analyse, inspect and interpret a hypothesized relationship between sequence data variation and a group feature. In this approach, phylogenetic trees are utilized in a novel way and their evaluation is heavily simplified by a single metric value rather than one-by-one inspection of each tree in a "phylogeny-forest".

The measure does not suggest what kind of link there may be between sample data and hypothesized groups. Neither does the method inform how the sequence variation was used to group the samples, which requires additional analysis using approaches such as PCA (89) or feature selection. However, the metric can guide further research by supporting a proposed relationship, revealing which of a set of hypotheses is more likely, or identifying specific sequences (e.g., genes) that are associated with the grouping. For the disease subtype example above, to investigate the tree-structure of many proteins, it is possible to identify "significant" proteins, i.e., with GDR values that indicate that the subtype groups occur in separate tree clades. These can then be selected for further investigation, such as finding what type of variation in the sequence data is causing the tree structure, that is, which amino acid variants between the subtypes could be involved with the disease.

5.3 The GDR

The GDR quantify the amount of group-structure in a distance-based clustering result for predefined groups. It is defined as the ratio between the mean distance of all samples within a group, to the mean distance of all samples in distinct groups:

$$\mathbf{GDR} = \frac{mean\ intra{-}group\ pairwise\ distance}{mean\ inter{-}group\ pairwise\ distance}$$

A lower mean intra-group distance to a mean inter-group distance would produce a lower GDR: when the mean intra-group distance approaches 0, GDR approaches 0. Conversely, when the mean inter-group distance approaches 0, GDR approaches infinity. Thus, the lower the GDR, the closer are samples in the same defined group clustered together and vice versa:

GDR	=	0	groups are perfectly separated
GDR	=	1	groups are perfectly intertwined
GDR	>	1	there is on average a higher distance between groups of the same defined group than of distinct groups. There is most certainly not any direct relationship between the sample variation and the defined groups.

Tree A in figure 5.2 shows a tree that would produce a lower GDR compared to Tree B (figure 5.2). Samples from the three defined groups in the trees (Europe, East Asia and Africa) are clustered closer together in tree A than in tree B, i.e., there is a shorter average branch length between samples of the same group compared to the average branch length between samples of different groups in tree A. In tree B, on the opposite, these averages are similar, which would result in a GDR close to 1.

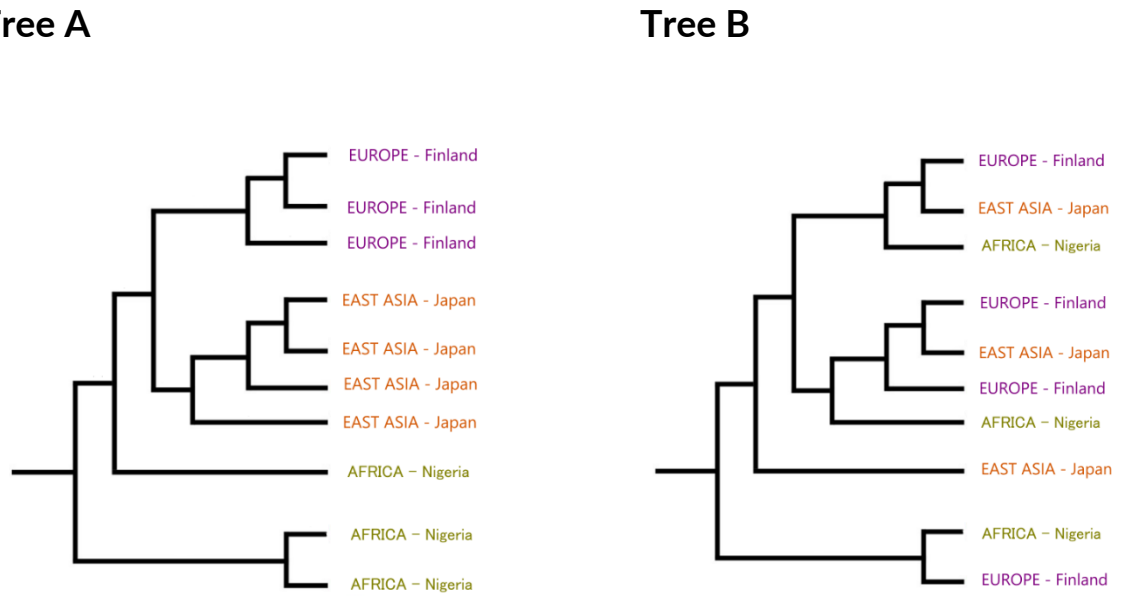


Figure 5.2: Tree A and Tree B illustrates two phylogenetic trees with 3 defined groups: Europe, East Asia and Africa. Samples from the population groups are shown with different colours and labels.

Tree A: The samples from each group are arranged closely together in the tree, indicating a groupwise tree-structure. The GDR would be lower for this tree compared to tree B.

Tree B: The samples from each group are not arranged close together in the tree, hence a groupwise structure is not present. The GDR would be higher for this tree compared to Tree A.

In this manner, the GDR takes advantage of branch length distances for all pairs of samples in a phylogenetic tree and quantifies the diversity with respect to the defined sample groups. The GDR calculation requires a minimum assignment of two groups in a clustering calculation, and at least two samples in each group. However, the derived GDR will require additional samples in each defined group to reach statistical significance.

A. Rambaut et al. used the SDR to compare the genetic diversities of human immunodeficiency virus type 1 (HIV-1), by comparing SDR values calculated for two individual phylogenetic trees (88) . Defined groups were based on HIV-1 strain subtypes, and their goal was to determine whether the two phylogenies differed in strain-type diversity by comparing tree structures through the SDR score. One of the phylogenies used standard nomenclature of HIV-1 strains to allocate the 11 subtypes, whereas a heuristic optimization algorithm minimizing the SDR score (i.e., maximizing the subtype structure) was used to allocate the 11 subtypes for the other phylogeny. In the latter case, multiple numbers of allocated subtypes were tested, producing different SDRs. Their statistical testing approach was directed to test whether the two SDR scores, that represented each of the phylogenies, were significantly different. If this was the case, it would imply a difference in diversity between the HIV-1 subtypes.

In this project, the intention is not to compare diversity between groups, that is, how divergent the groups in different phylogenetic trees are compared to each other. The intention is rather to measure whether a relationship exists between sequence data variation and the defined groups. The GDR estimate for an individual phylogenetic tree can exclusively serve to answer the query, because of the difference in research question it targets here. Therefore, comparable GDRs from multiple phylogenies (the null-distribution set aside) are unnecessary, creating a distinction to how it is implemented in the HIV-1 research. Also, trees in this procedure are not optimized for GDR values, compared to the trees that are created upon an SDR-optimization criteria in A. Rambaut et. al’s study.

5.4 GDR statistics

Evaluating statistical significance of a GDR value requires testing against a GDR null distribution. It is generated by keeping the tree topology constant and randomizing the defined groups, i.e., randomizing which sample belongs to which group, and then calculating the GDR for this randomized sample-group arrangement. The number of groups and number of samples within each group are kept constant and equal to those of the formerly defined groups. In this way, the randomness tested is solely based on group arrangement and tree topology, eliminating other factors. This restricts the number of group-rearrangements, as only a discrete number of random group arrangements are possible. The GDR values thus follow a hypergeometric probability distribution. It is of interest to determine the probability of obtaining a GDR equal to or lower than the observed GDR (describing success, in statistical terms), since a lower GDR also indicates group structure. Therefore, the cumulative distribution function (CDF) is suitable for hypothesis testing.

A significant GDR would imply that the mean intra-group distance is lower than the mean inter-group distance, than expected by chance. In other words, it would insinuate a consensus between grouping predicted by the clustering algorithm and the groups defined by the researcher, which was unlikely to occur if the groups were randomly assigned. Therefore, the essence of the statistical test is whether randomly assigned groups could give rise to an equal or lower GDR in the same cluster result.

H0: The GDR is not lower than expected by chance. No relationship is present between the tree structure and the group feature

H1: The GDR is lower than expected by chance. Defined groups are significantly clustered closer together in the tree, indicating a relationship between the tree structure and the group feature

The required number of GDR values for the CDF to contain the true empirical p-value at chosen significance level can be found through test-calculations of confidence interval (CI). This was achieved by adjusting the *num_perm_samples* variable in the R code listed below,

until the desired upper CI is found (displayed by printing the *high_q* variable in the code below):

```
prob = 0.05
var_p = prob * (1 - prob)
num_perm_samples = 100
var_p_n = var_p / num_perm_samples
sd_p_n = sqrt(var_p_n)

# 0.025 quantile
low_q = max(0, (prob-1.96*sd_p_n))

# 0.0975 quantile of true empirical p-val
high_q = prob + 1.96 * sd_p_n

print(high_q)
```

For example, calculating 100 random GDRs for randomly assigned groups will result in a ~93% confidence that the true empirical value is contained in the interval.

A p-value for observed GDR can then be calculated from its null distribution, and H0 or H1 may be rejected. Multiple testing correction is necessary in case of many tested group arrangements in the same tree cluster, or if many trees are created and a similar study is performed on all tree clusters (i.e., same type of data and subtypes are defined by the same feature).

5.5 Simulation

A simulation of GDR values was performed to investigate the behaviour of the metric and to understand how it varied with the structure of the phylogenetic tree. This was done by estimating GDR values for a fixed number of samples belonging to two defined groups and varying how these were clustered into two clades. The simulation was implemented in Python and is available as part of the code bundle on GitHub.

The simulation imitated a phylogenetic clustering result of N samples divided into two clades, *C1* and *C2*, with $N_{C1} = N_{C2} = N/2$ number of samples in each clade, see figure 5.2 for an example. A distance *d* between the two clades was arbitrarily selected, as it did not influence the GDR result for this tree-constellation. The samples within each clade had distance $d = 0$ to each other, meaning they clustered together perfectly.

Two groups, *G1* and *G2*, were defined. The number of samples within each group was always equal, where $N_{G1} = N_{G2} = N / 2$. Hence, in the situation where *G1* and *G2* were perfectly separated into clade *C1* and *C2* in the tree, each clade contained an equal number of samples, $N_{C1} = N_{C2} = N_{G1} = N_{G2} = N / 2$, as shown in figure 5.2A.

The simulation was performed by calculating the simulated GDR (*simGDR*) for the clustering where, in each round of the simulation, one *G1* sample was "permuted" from *C1* to *C2*. The permutation parameter, *P*, was defined to describe the number of samples in group *G1* that were moved from *C1* to *C2*, i.e., the number of samples that were no longer clustered together in a clade with the group to which it originally belonged, as shown in figure 5.2B. *simGDR* was calculated for all combinations of *P* and *N* in the range N = 2 – 200 and P < N.

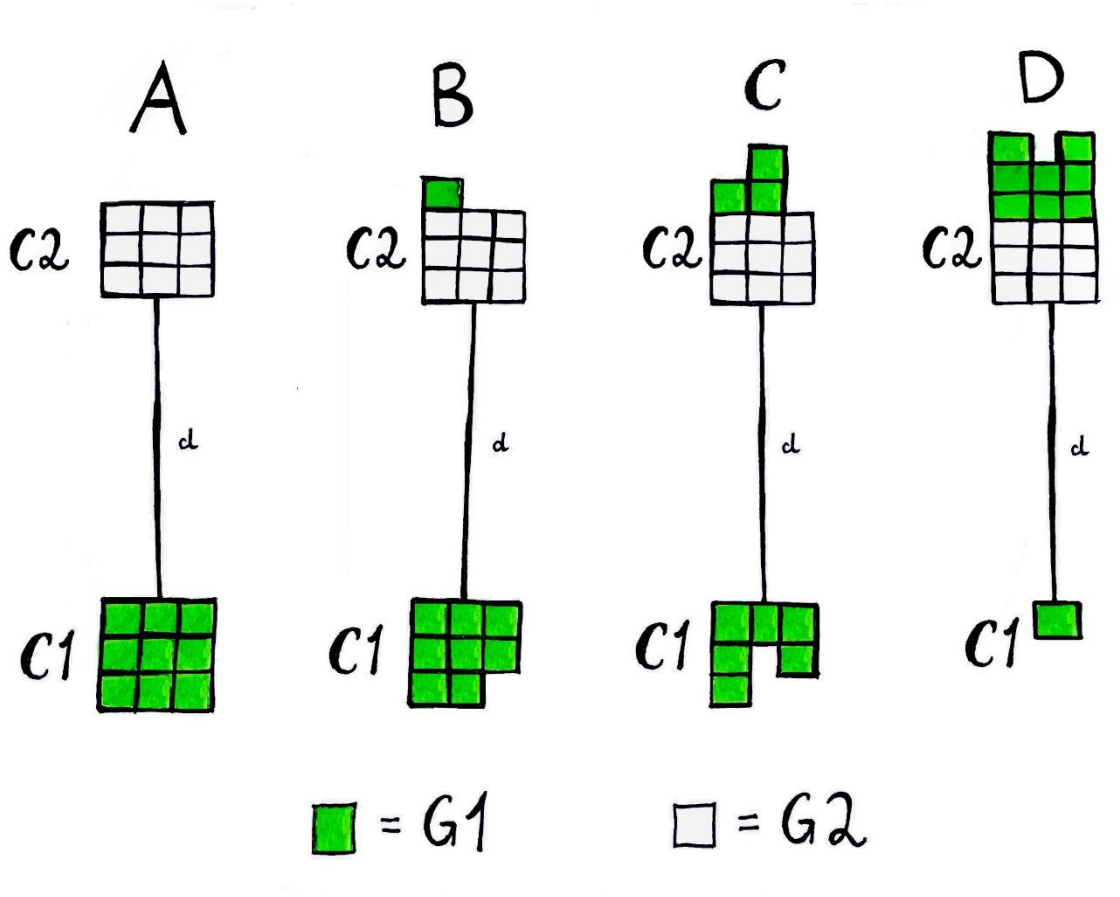


Figure 5.3:

Illustration figures of simulated trees, A, B, C and D, with N = 18 samples and N_{G1} = N_{G2} = 9 samples in each group. Clade 1 and clade 2 in trees are marked with C1 and C2. d = distance between C1 and C2.

In tree A, samples are perfectly separated into two clades, i.e., GDR = 0. In tree B, one sample is permuted from C1 to C2. In tree C, three samples are permuted from C1 to C2. In tree D, one G2 sample remain in C1, hence all samples apart from one are equal and GDR would be close to 1, which indicates no groupwise tree structure.

To further explore the statistical significance of these *simGDR*s, the CDF of each value was calculated. The observed *simGDR* was thereby tested against its CDF to obtain p-values, which were further corrected for multiple testing. The *p.adj* function in the standard R library was used with default parameters, and using the default Holm correction. See GitHub for code details.

5.6 GDR for ethnic populations in 3'UTRs

The GDR was used to measure the degree of structure in 9381 assembled phylogenies of human 3'UTR sequence samples, with ethnic populations as the defined groups. In this way, the relationship between variation in 3'UTRs and ethnicity was investigated. By this means, from a global perspective, it was examined whether evolutionary patterns exist in these non-coding, regulatory regions of the genome, which are primary targets of microRNAs and form part of the miRNA-mRNA interactome. The 9381 trees represented the set of genes that had entries in the Reactome database and were selected from the complete set of human genes defined in the Ensembl GrCh38 reference genome annotation. The trees served as the input data to the Willow analysis pipeline which had the following steps:

1. Calculate the GDR for each gene-tree
2. Select a subset of genes for statistical validation
3. Generate null distributions for selected genes
4. Perform hypothesis testing

Code was written in Python and R. R was mainly used for conversions between specific data formats and for statistical testing, whereas the Willow software was written in Python and used for GDR calculations. The purpose of creating Willow was to (i) automate high-throughput GDR calculation from large numbers of phylogenetic tree distance matrices, and (ii) make the procedure reproducible and open source. Also, the program was developed to allow future expansion to include new applications and functionalities. See https://github.com/norabbull/Willow2/tree/master/Willow1.0 for code and documentation.

5.6.1 GDR calculation in Willow

The total branch length distances between leaf node pairs in the trees were retrieved from the Newick format phylogeny files with the R function `cophenetic()` from the ape package (version 5.5). This outputs a square triangular symmetric distance matrix containing all total branch length values between the leaf nodes. All 9381 gene-trees were run, and the

matrices were saved as .CSV files. Code details are found in GitHub: https://github.com/norabbull/Willow2/tree/master/Willow1.0/R_scripts/newickToDistmat.R

The next step was to utilize the population subtype- and branch length information to calculate the GDR for each tree. As described in the data-section, groups were defined at two category levels: super-populations *(super)* and sub-populations *(sub)*, hence GDR was calculated twice per tree (one per level).
This is a multi-step process in Willow. The program requires the input files described in table 5.1. Sample files are available as part of the GitHub repository. See https://github.com/norabbull/Willow2/tree/master/Willow1.0/src for source code scripts.

Table 5.1: Overview of Willow1.0 program requirements

File type	**Format**	**Content**
Configuration file	yaml	Main parameters to specify are: `func`: function to run. Main options: - `calcGDR` – to calculate GDR values - `calcGDRrandom` – to calculate random GDR values `group_categories`: specify group categories `num_ranodm_values`: number of random values to calculate for null distribution For all parameter values to specify for the user and descriptions, see GitHub: https://github.com/norabbull/Willow2/tree/master/Willow1.0/Documentation/main_config_parameters.csv
Group definition file	TSV	A tab delimited file containing information on the defined groups at the *super* and *sub* population level. The file

		contains 3 columns: `ClassificationType` (CT), `ClassificationName` (CN) and `ClassificationDescription` (CD). CT contains group level information *(super* or *sub*). The CN column contains population groups described with a 3-letter acronym, consistent with the abbreviations used by the 1000 Genomes Project. CD contains additional population information (to provide optional notes to the user). Example row in file: `SUB  FIN  Finnish in Finland`
Gene-tree distance matrix file	CSV	Contains the total branch length distances between each sample pair according to the estimated phylogenetic tree. All matrix entries are comma-separated. Samples are represented as rows and columns in the matrix, on the form "SUP___SUB___*sample info*", where SUP and SUB represent the 3-letter acronyms, matching a *super* or *sub*, population group in the group definition file. Sample info is optional. Example: *(for a sample from Finland in Europe)* `EUR___FIN___HG00178`

Willow calculates GDR for a gene-tree as follows:

1. The function `run_calcGDR()` is the top-level method that run all functions necessary for GDR calculations at both group levels simultaneously, based on the specified tree distance matrix files specified in the input configuration file. All GDRs for a group level are written to a single file. Hence, two GDR CSV-files (*super* and *sub*), will be generated for the 3'UTR study, each containing two comma-separated columns: *gene* and *GDR* value. The gene is written in the form "ENSEMBLE-

REFERENCE___GENE", i.e., including both Ensembl gene id and gene name for the gene that the 3'UTR was extracted from.
Example of resulting CSV-file:

```
gene,GDR
ENSG00000003400___CASPA,0
```

2. In `run_calcGDR()`, a class object of TreeMetrics is instantiated. A configuration file is given as input argument and contains file path information to the other input files required (*subtype definition file* and *gene-tree distance matrix file*). Information related to the distance matrix, group definitions and sample information is retrieved from the files and stored as variables in the TreeMetrics-object.

3. The function `calcGDR()` is further called and is the core function to fetch information from the distance matrix and to calculate total distances within- and between group samples in the tree, respectively. Since *GDR = mean within-subtype pairwise distance / mean between-subtype pairwise distance*, all pairwise distances between all sample pairs are included in the calculation. Since the matrix has triangular symmetry, the function iterates the upper triangular matrix only. Moreover, since the GDR were to be calculated once for each population level, i.e., twice for each tree, the code was written to handle multiple group levels in the same iteration. In this way, Willow can handle an any number of grouping levels at a time, however, groups must be specified correctly in the input group definition file and all group levels must be specified in all sample labels. (See table 5.1 for detailed format description) .
4. The function `calcGroupDists()` computes the total *within*- and *between* group pairwise distances.
5. The function `calcMeanGroupDists()` computes the mean *within*- and *between* group pairwise distances.
6. Finally, the function `calcGDR()` divides mean *within*-group pairwise distance by mean *between*-group pairwise distance, to obtain the GDR.

5.6.2 Gene selection

Statistical testing of GDR values required generation of null distributions for each gene at each group level. To be 90% certain that the true empirical p-value was contained in the generated null distribution, it was necessary to calculate at least 73 calculated random GDR values. This number was obtained by testing different numbers of random GDR samples (the `num_perm_samples` variable in the code below), until the upper quantile (`high_q`) reached ~ 0.1:

```
prob = 0.05
var_p = prob * (1 - prob)
num_perm_samples = 73
var_p_n = var_p / num_perm_samples
sd_p_n = sqrt(var_p_n)

# 0.0975 quantile of true empirical p-val
high_q = prob + 1.96 * sd_p_n
print(high_q)
```

However, to obtain statistical tests with null distributions for all genes at 90% significance level, it is required to calculate a total of (9800 genes x 73 GDRs x 2 levels =) 1.430.800 random GDR values. This was estimated to take at least 8 weeks on the largest available server. Therefore, only a selection of ~1000 genes at both group levels were tested, which was considered feasible with respect to time and to gain a sufficient result for the aims of this work.

The ~1000 genes were selected as follows:

1. A table of all genes and their GDR values combined for both population group levels was created
2. The table was sorted by GDR values
3. The top 1600 genes with the lowest GDR values in the combined table were selected (i.e., the 1600 absolute lowest GDR values across both group levels)

4. The duplicated genes in the filtered list were removed (as a result of having both a low *super-GDR* and *sub-GDR* in the initial table used to select for lowest GDRs).

There were 1054 genes remaining in the list after this filtering process.

5.6.3 Null distribution

A minimum of 73 random GDR values needed to be calculated for each of the 1054 selected genes. For this process, the population information for each sample was shuffled randomly, so that each sample in a tree belonged to both a randomly selected *super*- and *sub*-population. The random GDR could then be calculated for both levels simultaneously. For instance, a sample originating from Finland in Europe could receive the label Peruvian from South Asia. This way, the group sizes as well as number of samples belonging to each group was kept constant at both levels, but the group labels were random.

In Willow, this procedure includes calling the function `shuffleGroupSamples()` to perform a priori GDR calculation, but otherwise following the same steps as described for the GDR calculation above. The parameter `num_random_values` specified the number of required random GDR values for each gene. Here, `num_random_value`*s* = 1000, to obtain as many random GDR values as possible, with the goal of calculating a minimum of 73 random GDR values for each gene-tree.

The calculation was performed on one of the Norwegian Sequencing Centre servers (high-performance cluster), using 5G per CPU and 128 CPUs per task, which took approximately 2 weeks. Since the process was restarted several times during the weeks due to technical issues, the number of calculated random GDR values for each gene exceeded the minimum required and most genes attained ~117 random GDRs. Also, because of earlier test-calculations of null distributions, a few of the gene-trees attained > 700 random GDR values in their null distributions, however, the slightly increased significance level achieved for these did not appear to make a substantial difference to the result. The null distributions, sufficient to provide a 90% CI, were then ready for hypothesis testing.

5.6.4 Hypothesis testing

Hypothesis testing of the observed population-GDR values (*popGDR*) with respect to their respective null distributions was performed in R.

The hypothesis tested:

H0: *popGDR* >= random GDRs values, i.e., reject hypothesis of population-specific structure in 3'UTR phylogenetic trees

H1: *popGDR* < random GDR values, i.e., accept hypothesis of population-specific structure in 3'UTR phylogenetic trees

The function `calcAndSave_pval()` was created to streamline hypothesis testing for all GDR values, outputting the p-value for a *popGDR* for each gene to a CSV-file:

1. The null distribution values for a gene was given as argument to the function `ecdf()` from the stats package (version 3.6.2, which returns the empirical cumulative distribution function (eCDF).
2. *popGDR* is given as argument to the eCDF-function, which calculates and returns its p-value.

Further, all p-values were corrected for multiple testing with the function `p.adjust()`, using the Holm-correction (as default).

5.6.5 Inspection of individual genes

The genes KRA22, NDUS5 and HAUS4 were inspected individually in terms of population distributions across the unique 3'UTR sequences they were built from. KRA22 was inspected due to its low resulting GDR values at both super-and sub population level. The other genes were selected based on their high, significant resulting GDR values, to inspect

how a GDR value were related to its population-distributions over unique sequences for such instances.

5.6.6 Calculating the fraction of non-zero pairwise distances for each tree

The fraction of how many equal samples each tree contained (i.e., the number of samples with 0 distance to each other) was computed to gain a brief overview of the data. This is referred to as the non-zero distances for a tree. It was calculated from each tree's distance matrix, containing all total branch length values between the leaf nodes, as the percentage of the matrix values that were not equal to zero. To calculate these values for all trees, the function `calcNonZeroPhydists()` (contained in the treeMetrics-class) was called, where its calculation procedure is implemented. See GitHub for code details.

6 Results

6.1 Simulation of GDR values for hypothetical phylogenetic trees

To investigate the behaviour of the GDR metric in the simplest possible phylogenetic tree arrangements, GDR values for 19.899 simulated phylogenetic trees were calculated. As presented in the methods, each tree contained two clades (*C1* & *C2*), two defined sample groups (*G1* & *G2*) and a total of *N* samples. The smallest tree contained $N = 4$ samples, whereas the largest tree had $N = 400$ samples. The GDR was calculated for all possible pairwise combinations of the two groups, with group sizes $N_G = N/2$ ($= N_{G1} = N_{G2}$) in the range 2-200 and number of permutations *P* (describing the number of group-switchovers between clades) where $P < N_G$.

The resulting simulated GDRs are plotted in figure 6.1. The x-axis represents the number of *G2* samples that are switched over from *C2* to *C1* (i.e., number of permutations) and the number of samples in each group, N_G, are shown on the y-axis. The GDR value for each arrangement is displayed with colour grades from dark purple to yellow, where yellow points of the samples close to the diagonal correspond to those which exhibit high GDR values (i.e., close to or equal to 1). The plot illustrates an increased resolution of GDR values along the y-axis, i.e., there are more attainable GDR values for each tree as group size increase, since the number of tree arrangements that causes different GDR values increases. This increased resolution corresponds to a smoother colour gradient, as seen in the upper part of the plot, compared to those attained for the lower group sizes.

When group size is small, a few permutations of *G2* samples from *C2* to *C1* cause a large shift in GDR value. When group size is large, on the other hand, a single permutation only results in a fractional change in the GDR value. For example, a single permutation when group size =

5, from P = 2 to P = 3, leads to a 0.25 change in the GDR value (i.e., GDR shifts from 0.5 to 0.75), compared to a group size = 195 when the single permutation produces a 0.0052 shift in the GDR value (i.e., GDR shifts from 0.0103 to 0.0155).

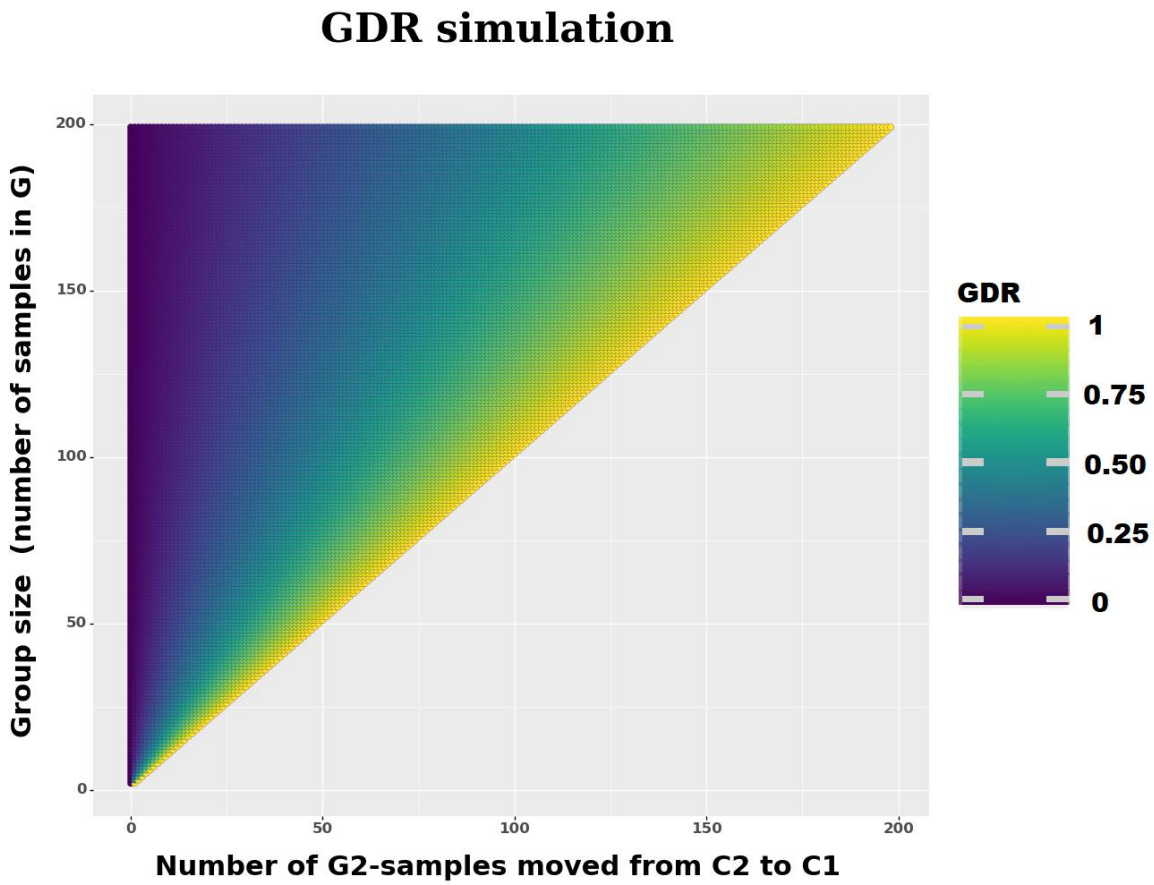


Figure 6.1: Simulated GDR values. Each point in the plot represents the resulting GDR value for one "tree". The "tree" arrangements for each GDR can be read from the X-and Y-axes: The number of samples in each group ($N_{G1} = N_{G2} = N_G$) is shown on the Y-axis, whereas the number of samples that are moved from C2 to C1 (i.e., number of permutations) are shown on the X-axis.

The GDR value is represented by colours, as shown with the GDR legend. Yellow points indicate high GDRs (maximum 1), where no group-specific structure Is present for G1 and G2. The darker colours indicate lower GDRs, closer to 0 (minimum 0). These points indicate that the "tree" show group-specific structure for G1 and G2, i.e., groups are well separated in the simulated "tree".

6.2 Statistical test result of simulated values

Statistical significance for all simulated GDR values was estimated at the ~10% level. For this procedure, a null distribution of 100 random GDR values for each GDR was calculated. The estimated p-values are shown in figure 6.2:

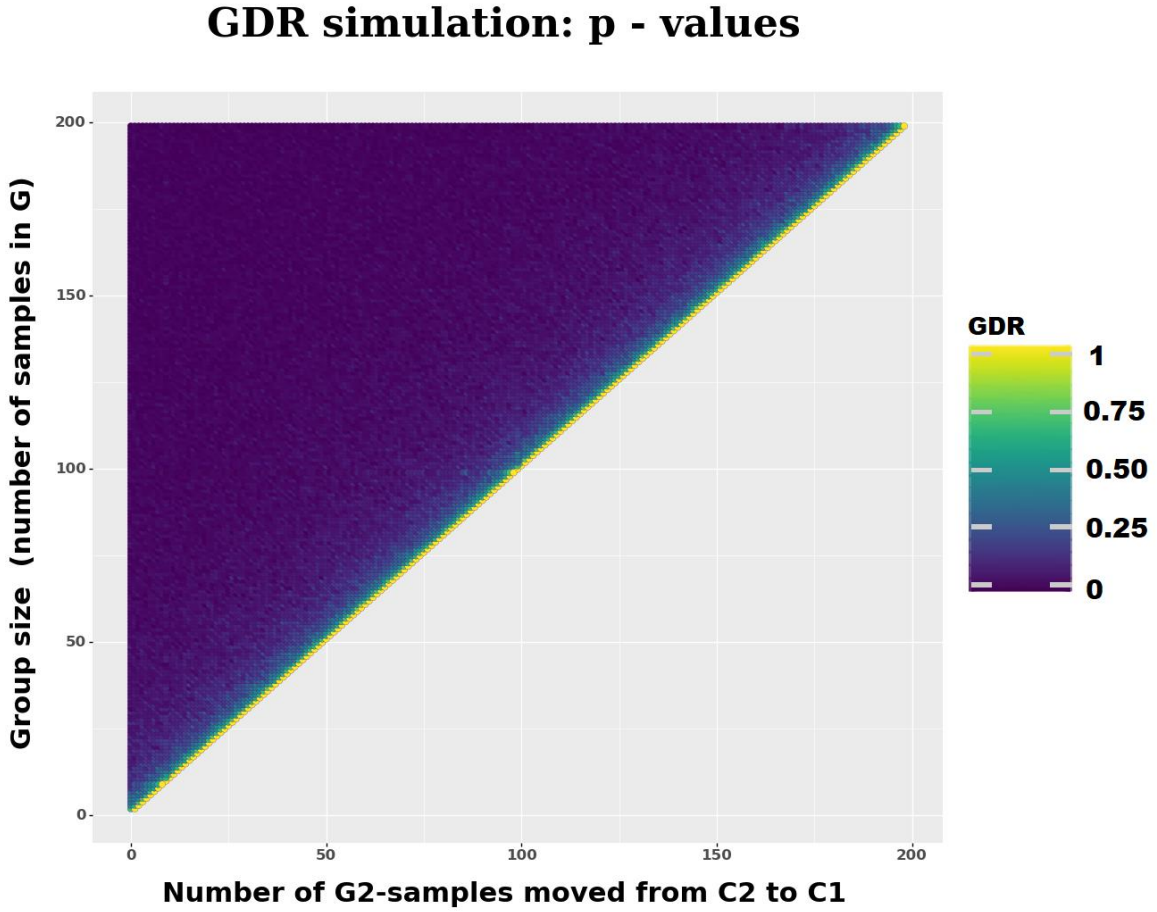


Figure 6.2: Simulated p – values for all simulated GDR values. Each point in the plot represents a p – value that corresponds to the GDR value for the same "tree" arrangements as shown in plot 6.1. See plot 6.1 for detailed description of X- and Y - axes.

The p - value is represented by colours, as shown with the "p – val" legend. Yellow points indicate p-values close to 1, i.e., the group arrangements that resulted in insignificant GDRs, hence showing no group-specific structure for G1 and G2. The darker colours indicate p - values closer to 0, hence these simulated tree-arrangements show statistically significant group-specific structure for G1 and G2, i.e., groups are significantly separated.

Most values in the plot are close or equal to 0, indicating that most simulated trees had a significant groupwise structure.

The p-values in figure 6.2 correspond to the GDR values in figure 6.1. Thus, the plot visualizes the probabilities of obtaining the observed GDR values, or smaller than the observed, for the distinct tree arrangements. Most p-values appear to be less than 0.25 with higher values occurring closer to the diagonal of the plot. This is reasonable, as the diagonal corresponds to trees where most samples are contained in a single clade. There is, however, a predominance of p-values close to 0. It therefore seems like most group arrangements result in a significant GDR, close to 0.

Another story is disclosed after correcting for multiple testing, shown in figure 6.3 below. Most adjusted p-values (p.adj) appear insignificant and equal to 1, whereas the remainder have value 0 with no values in between. Notably, the plot shows that there is a minimum group size (of approximately 50, where the red line crosses the y-axis in figure 6.3) (below), which no adjusted p-values are significant i.e., both clades in the tree need a certain minimum number of samples for the p.adj to reach significance. Additionally, for group sizes > 50, the plot shows that at least ~ 25% of all tree samples (shown with the red line in plot 6.3) need to be contained in C2 for the GDR to reach significance when all samples in C2 belong to G2 (i.e., one group only), when as many as ~20.000 trees are tested at once.

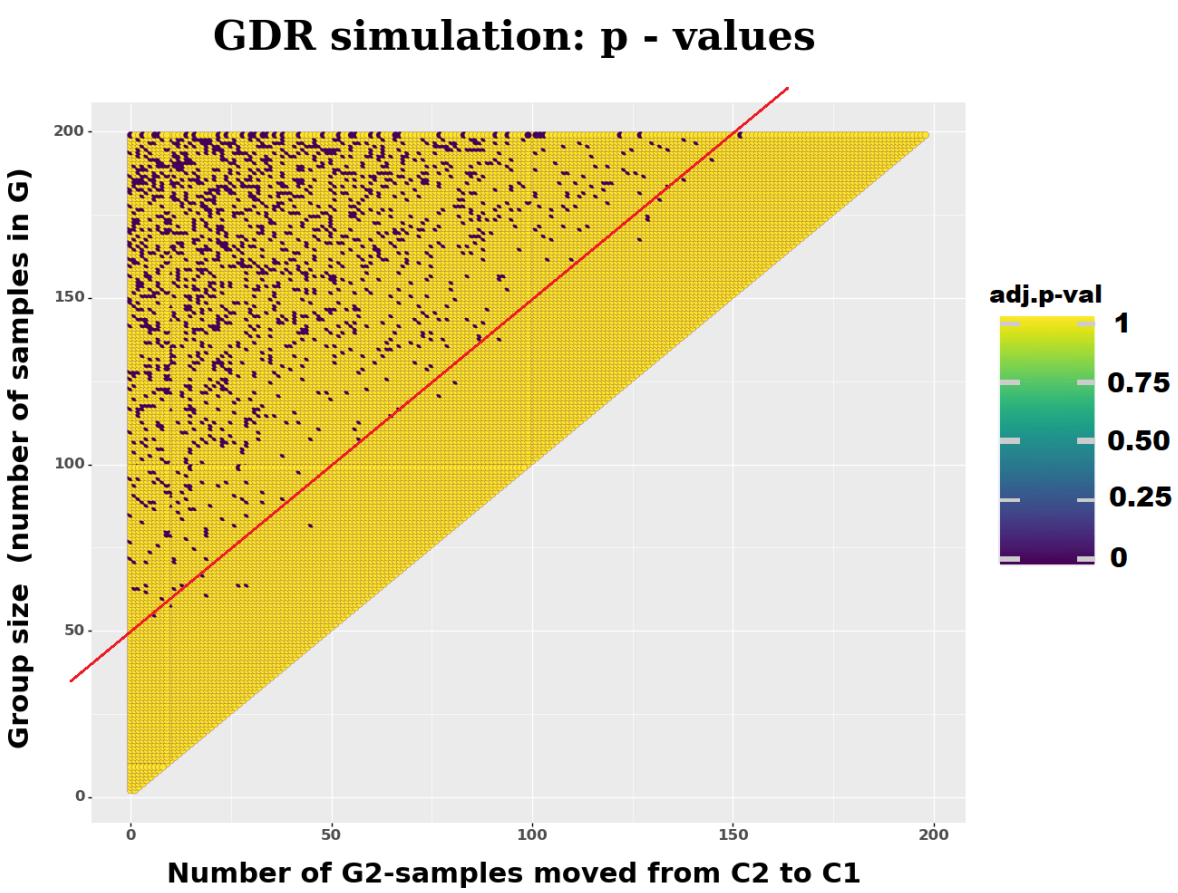


Figure 6.3: Simulated adjusted p – values for all simulated GDR values. Each point in the plot represents an adjusted p – value that corresponds to the GDR value for the same "tree" arrangements as shown in plot 6.1. See plot 6.1 for detailed description of X- and Y- axes.

p.adj value is represented by the colours, as shown with the "adj.p-val" legend, similarly to the representation of p – values in plot 6.2. Most values in the plot are equal to 1 and hence represents insignificant GDRs. Hence, most simulated trees show no significant group-specific structure after correcting for multiple testing.

The red line delimits where 25% of all G2 samples remain in a separate cluster (C2), and 75% of all G2 samples are moved to C1. For example, when group size = 200, C1 contains 200 G1-samples and 150 G2-samples, whereas C2 contain 50 G2-samples.

6.3 GDR for 3'UTR phylogenies

From the 9381 initial 3'UTR trees, 8782 trees returned one GDR value per population-group level (i.e., each tree obtained one super-population GDR and one sub-population GDR). The remaining 599 trees appeared to return a "not a number" (NAN) value because their distance matrices were filled with zeros, meaning there were no evolutionary distances between any samples in the trees (i.e., the sequences were identical, so there was no variation to allow phylogenetic estimation). Thus, they were automatically filtered out by Willow. The mean, minimum and maximum GDR values for super- and sub population levels is shown in table 6.1.

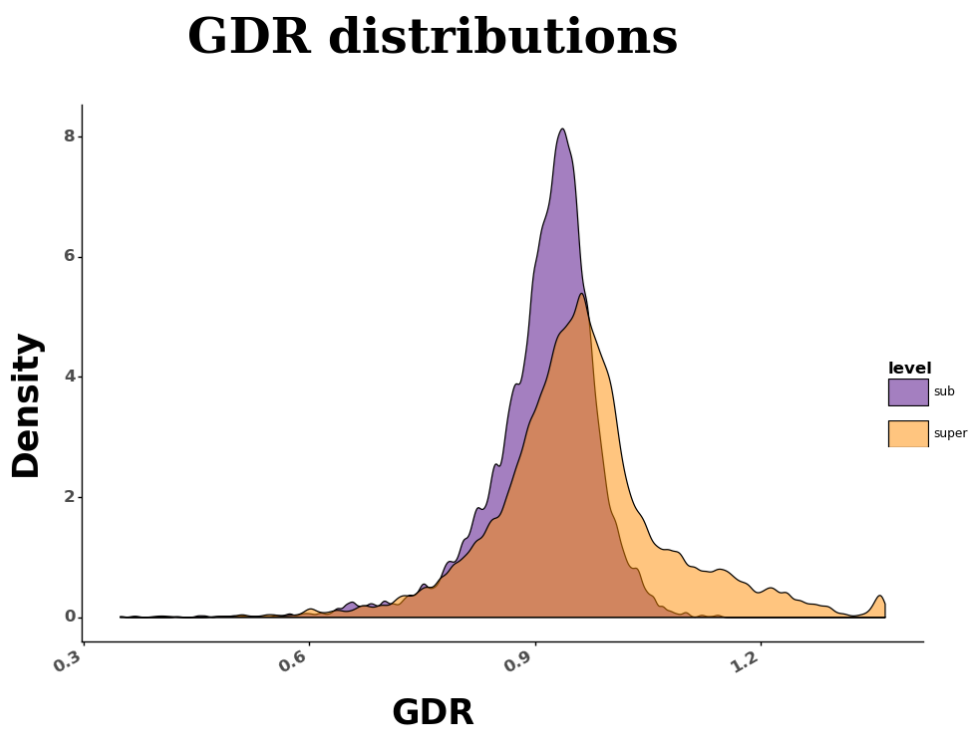


Figure 6.4: GDR distributions for all GDRs at both population group levels, shown as a density plot. The density plot for the super-GDRs are shown with orange colour and sub-population GDRs are shown in purple. The X-axis represents the GDR value and the Y-axis represents amount/density of each GDR value.

At super-population level, most GDRs are shown to be in the range 0.8 – 1.1, with many values exceeding 1. Hence, many 3'UTRs are shown to not exhibit super population-specific variation. Most GDRs at sub-level are in the range 0.8-1.0, hence it appears that more 3'UTR elements contain some ethnicity-associated variation structure at the sub-population level.

The super-population GDR distribution seems to be more spread across a broader range of different GDR values compared to the sub-population GDRs, but the two distributions are similar for GDR values < 0.8.

GDR value distributions for super- and sub populations are shown in figures 6.4 and 6.5. These show that both distributions exhibit long tails in both directions with a slightly shorter tail towards higher values in the sub-population distribution. The boxplots included for each distribution in figure 6.5 a) clearly show the many extreme values of GDR values in both directions:

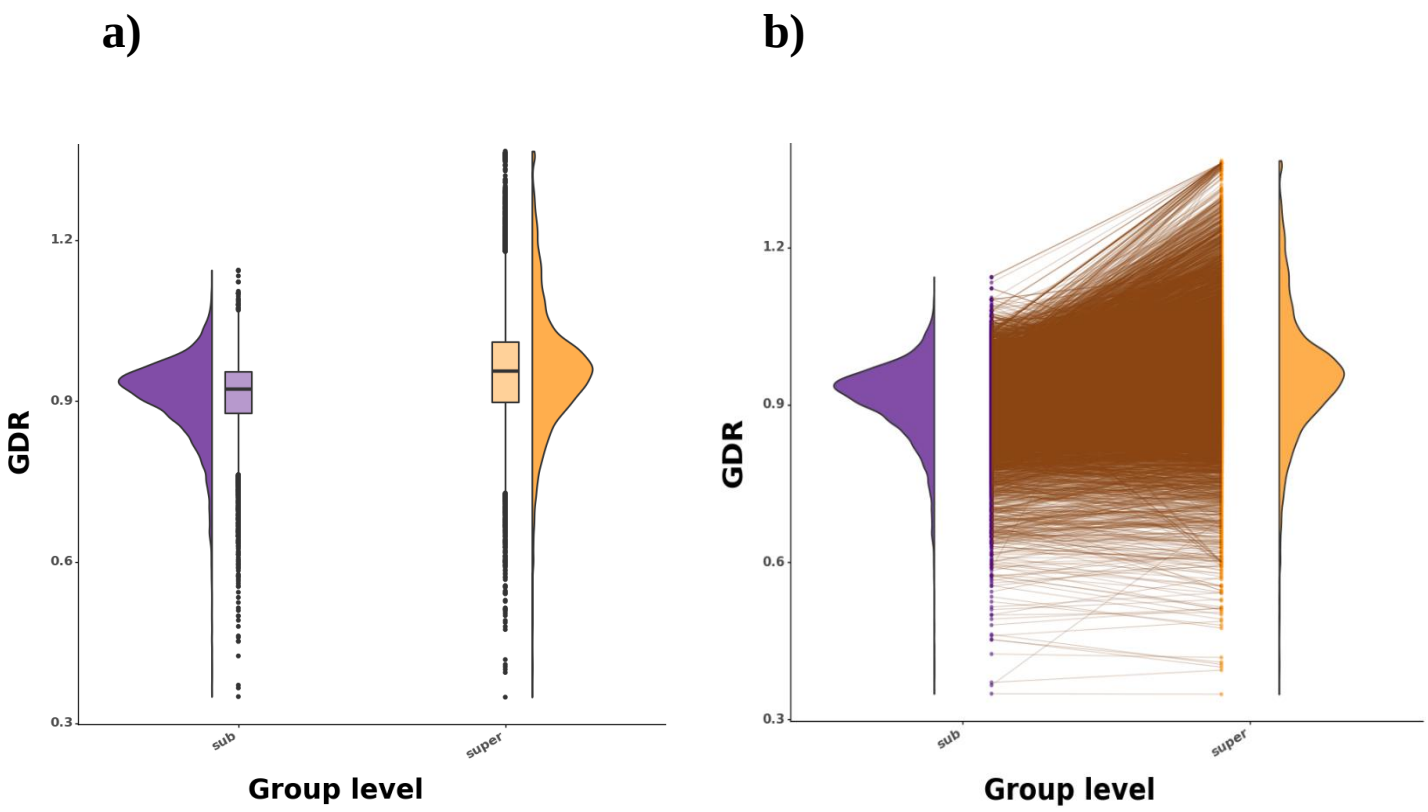


Figure 6.5: GDR distributions for all GDRs at both population group levels. The X-axis represents the population group levels and the Y-axis represents GDR values. Super-population GDR distributions are in orange and sub-population GDRs are shown in purple.

a) GDR super- and sub-population distributions plotted as a violin- and boxplots. The plots show the many extreme values of each distribution in both directions, but with fewer extreme values in the upper range of GDRs (> 1.0) at the sub-group level.

b) Comparison of super- and sub- population GDR distributions. Plots are as in plot a), but with the brown lines joining the corresponding points between the GDRs for each 3'UTR-tree for the two population-group levels. Most 3'UTR elements have similar GDR values for both group-levels in the lower range of GDRs (< 0.8). Hence, the trees with low GDRs show similar degree of population specific structure at either group level. In the upper ranges, the 3'UTR trees show a trend to obtain higher GDRs at super-level compared to at the sub-level.

Figure 6.5 b) shows association between the GDRs pair (i.e., super- and sub-population group) calculated for each gene. Most lines in the mid ranges of GDR values are too overlapping to show clearly in the plot, but a general trend of equally low GDRs for super- and sub-population levels are shown for the trees in lower GDR ranges, with a few notable exceptions. At the opposite range of GDR values, trees trend to keep a stronger population-specific structure for sub-populations than for super-populations. This trend includes ranges where many of the resulting GDR values are most certainly random due to no tree structure for any population groups. However, the observed trend could also indicate that there is, in general, more tree structure at the sub-population level compared to no detectable tree structure at the super-population level.

Table 6.1: Basic statistics for GDR result values

	super	**sub**
Count	8782	8782
Mean GDR	0.959	0.910
Min GDR	0.348	0.349
Max GDR	1.365	1.144

6.4 Statistical test result of GDRs from 3'UTR trees

6.4.1 Hypothesis tests result

Hypothesis tests of the GDRs estimated for a subset of 1050 trees were performed to verify that tree-structure relative to the defined population groups were significant. The 1050 trees were filtered on the criteria that they obtained the low GDR values at either the super- or sup population group level, i.e., they showed strongest population-specific structure.

To perform the test, null distributions were calculated for each tree at each population level. Combining the random calculated GDR values for all trees, a total of 139,184 random GDRs

were obtained in the range 0.857 - 1.123. A distribution plot of 30,000 of these values are shown in figure 6.6 (all values were too comprehensive to plot).

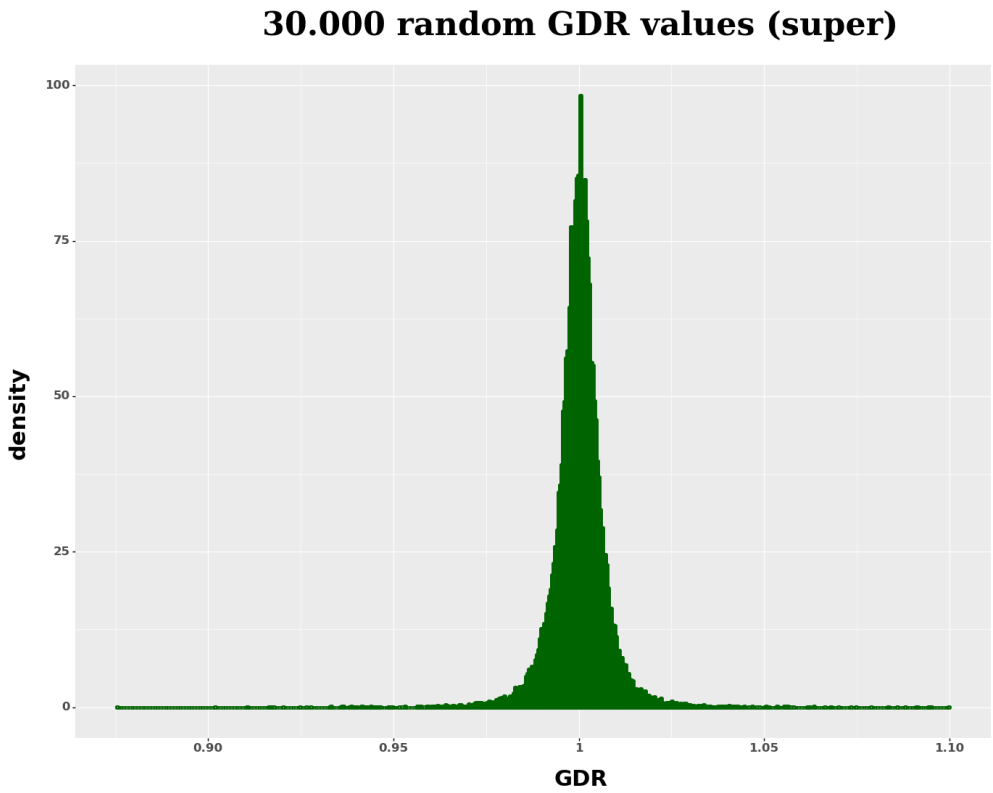


Figure 6.6: Density plot of 30.000 calculated random GDRs across a random selection of 3'UTR trees at the super-population level. The plot shows that most random GDRs are close to 1 for all trees.

The plot show that most random GDR values are close or equal to 1, indicating no tree structure. Hence, it appears unlikely that a tree will achieve population-specific structure by chance.

The Cumulative Distribution Function (CDF) was used to test each GDR against its null distribution. Plot 6.7 shows an examples of an estimated CDF, with the observed GDRs marked for super-populations. The CDF for the NCHL1 gene-tree shows that its GDR value of 0.976 (i.e., close to 1, indicating no tree structure) can return a significant result, where most calculated random GDRs are contained in a small range around 1.

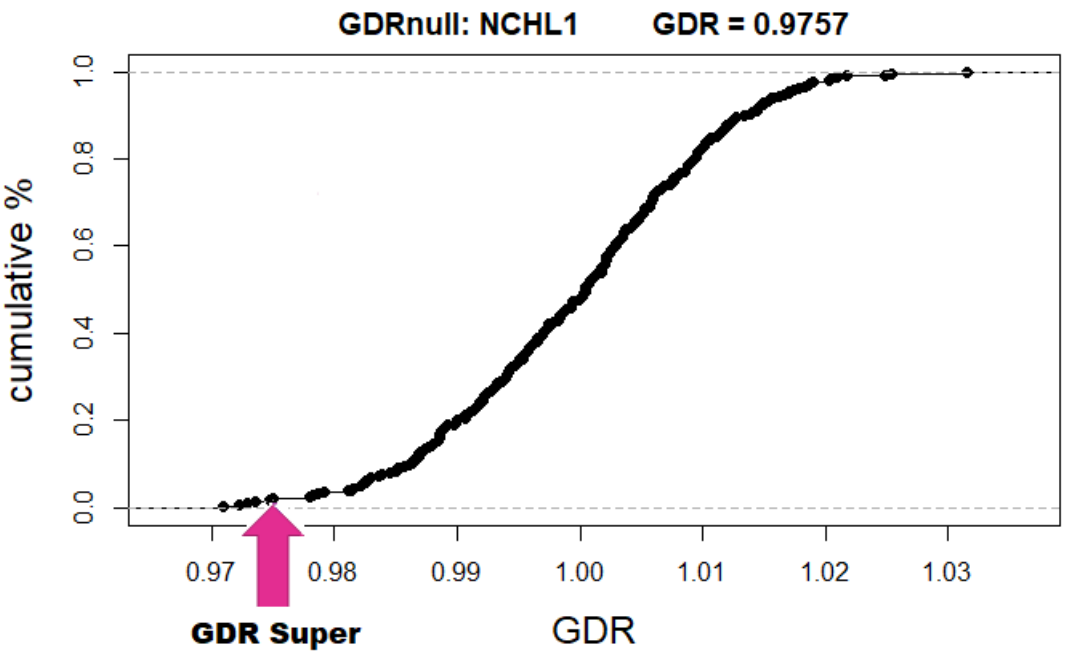


Figure 6.7: Plot of the random cumulative distribution for the NCHL1 - 3'UTR at super-level. The GDR is shown on the X-axis whereas the percent of cumulative probability is shown on the Y-axis. Most random GDRs are contained in the interval 0.98 – 1.02. The observed GDR for the gene is 0.976, marked with the pink arrow in the plot.

Of the 1055 genes that were calculated, 1005 reached statistical significance after correction for multiple testing. All values were corrected to 0 (significant) or 1 (insignificant). Hence, most of the tested GDRs resulted from population-specific structure in their respective trees at both population-group levels.

6.4.2 Inspection of percentage of non – zero values in distance matrices

The percentage of non – zero values in the distance matrices for each tree were inspected to gain an overall overview of zero – distances in the dataset for data-optimization reasons, shown in figure 6.8. It is shown that slightly more than 5 % of all distance matrices had 0 % non – zero distance values, meaning they were filled with zeros, hence there were no distances between any samples in the tree (i.e., all 3'UTRs for all human samples were equal). It is also shown that a total of ~25 % of all distance matrices contained > 80 % non-zero values, hence there must have been many unique 3'UTR sample variants for these genes, resulting in distance values unequal to zero between a large proportion of the samples.

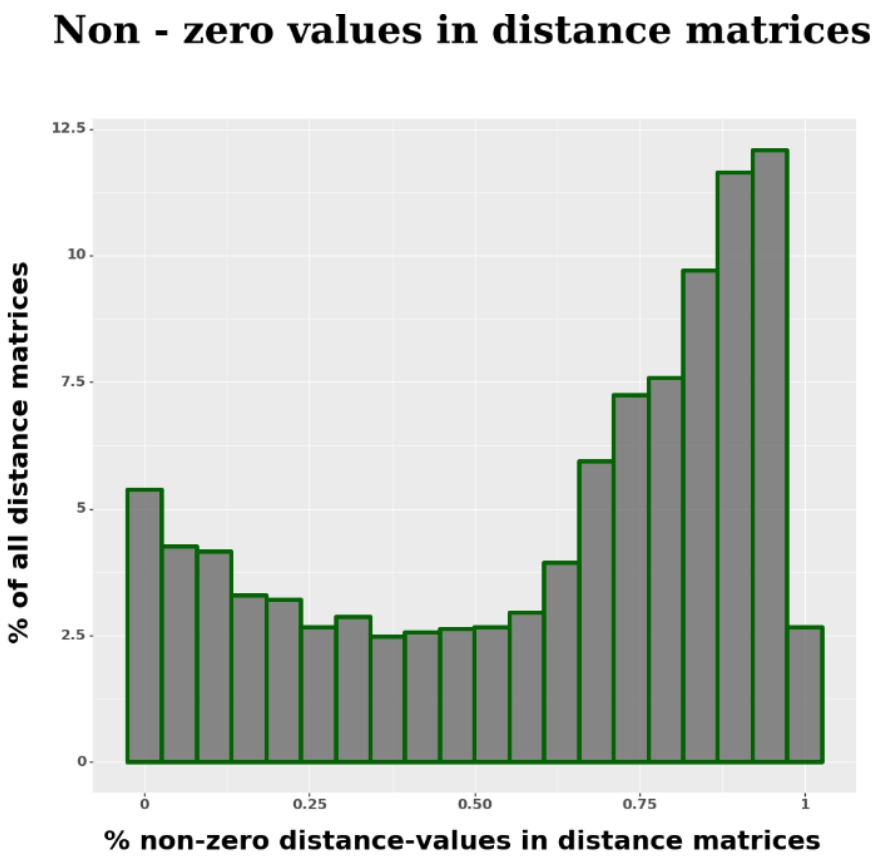


Figure 6.8. An overview of the amount of non – zero values in all the input distance matrices (which contain distance values between all sample pairs in a tree). The percentage of distance matrices is displayed on the y – axis whereas the percentage of non-zero values in a matrix is shown on the x- axis. Approximately 5% of distance matrices contained no pairwise distances (i.e., their 3'UTR was conserved across all samples).

6.4.3 Inspection of single genes

The population-distributions across the unique variants of the KRA22 gene at super- and sub population levels, respectively, is shown in figure 6.9 and 6.10. The same type of plots for the HAUS4 and NDUS5 genes at super-population levels are shown in figure 6.11 and 6.12. See figure captions for detailed descriptions.

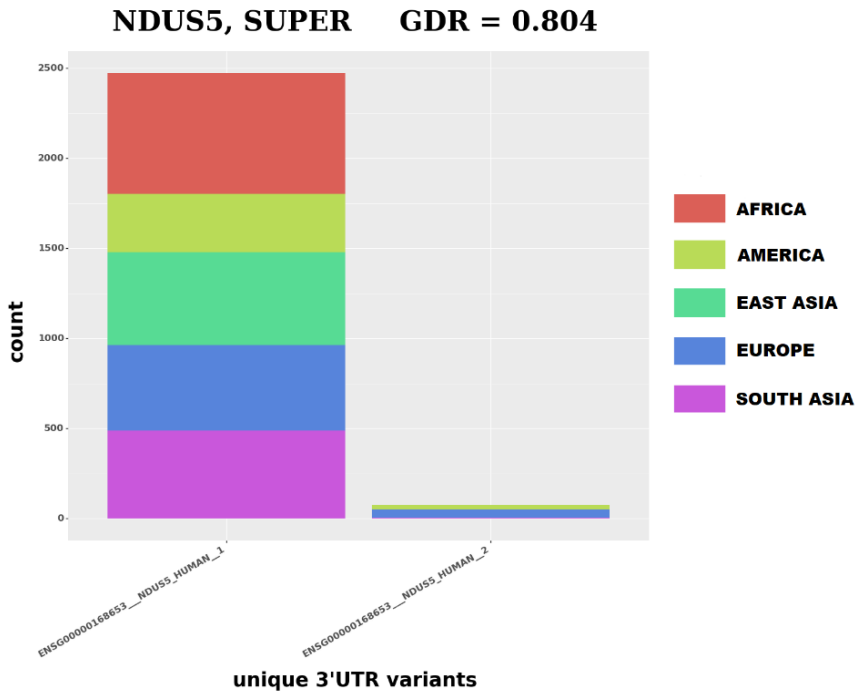


Figure 6.9: Unique sequence map of super populations for the NDUS5 gene 3'UTR, displayed as stacked histogram bars. Super populations are displayed with various colours. Two unique 3'UTR sequence variants are present, where most humans share the same variant (variant 1). The other variant is primarily present in a few European- and American samples, as shown by the blue and green colours in the bar for variant 2.

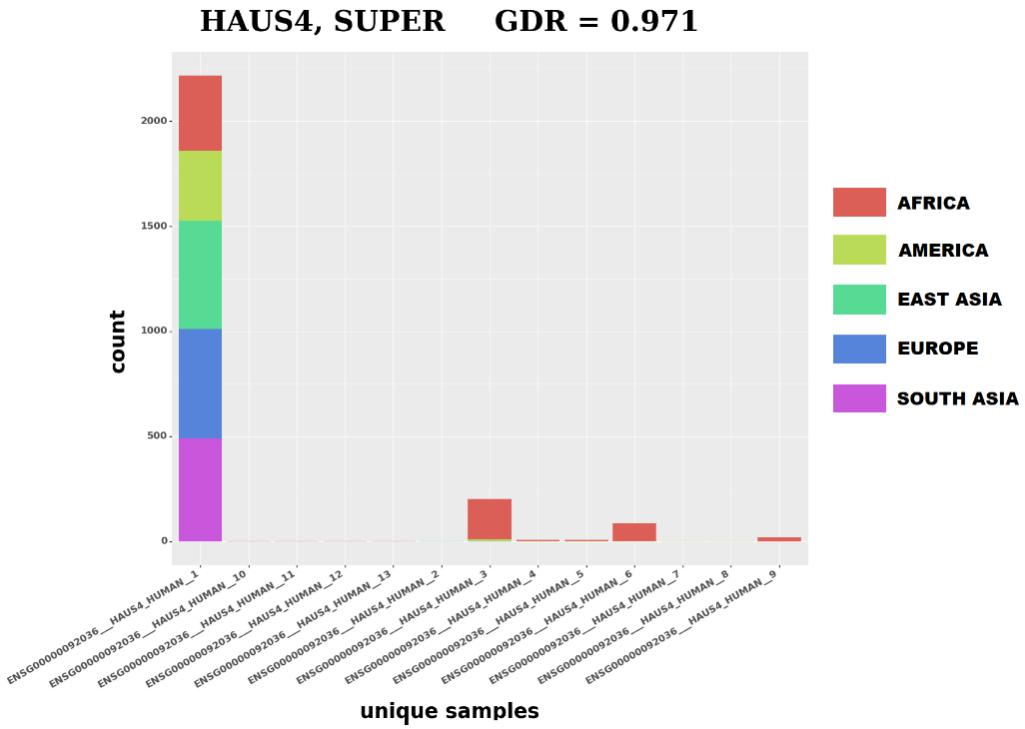


Figure 6.10: Unique sequence map of super populations for the HAUS4 gene 3'UTR, displayed as stacked histogram bars. Super populations are displayed with various colours. 13 unique 3'UTR sequence variants are present, where most humans share the same variant (variant 1). Apart from variant 1, 3 and 6, the remaining 10 variants are only present in one or a few samples, mostly from the African population, shown in red. Variants 3 and 9 are mostly contained in the African population.

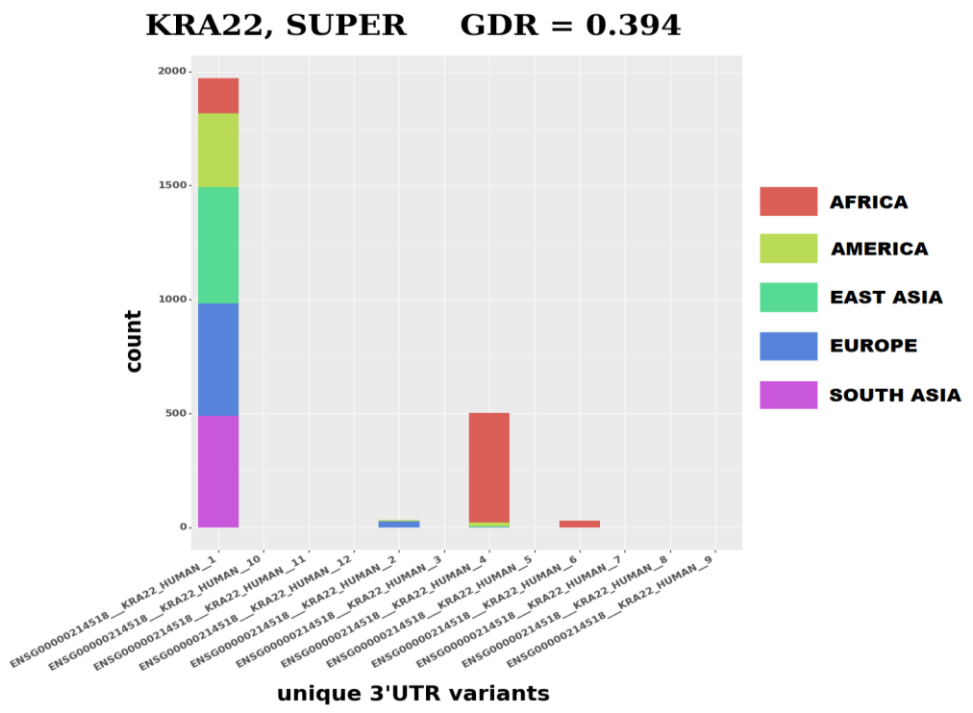


Figure 6.11: Unique sequence map of super populations for the KRA22 gene 3'UTR, displayed as stacked histogram bars. Super populations are displayed with various colours. 12 unique 3'UTR sequence variants are present, where most humans share the same variant (variant 1). Apart from variant 2, 4 and 6, the remaining 9 variants are only present in one or a few samples (too few to be shown in the plot). Variants 4 and 6 are mostly contained in the African population and a few Americans have variant 4. Variant 2 is mainly present in Europeans.

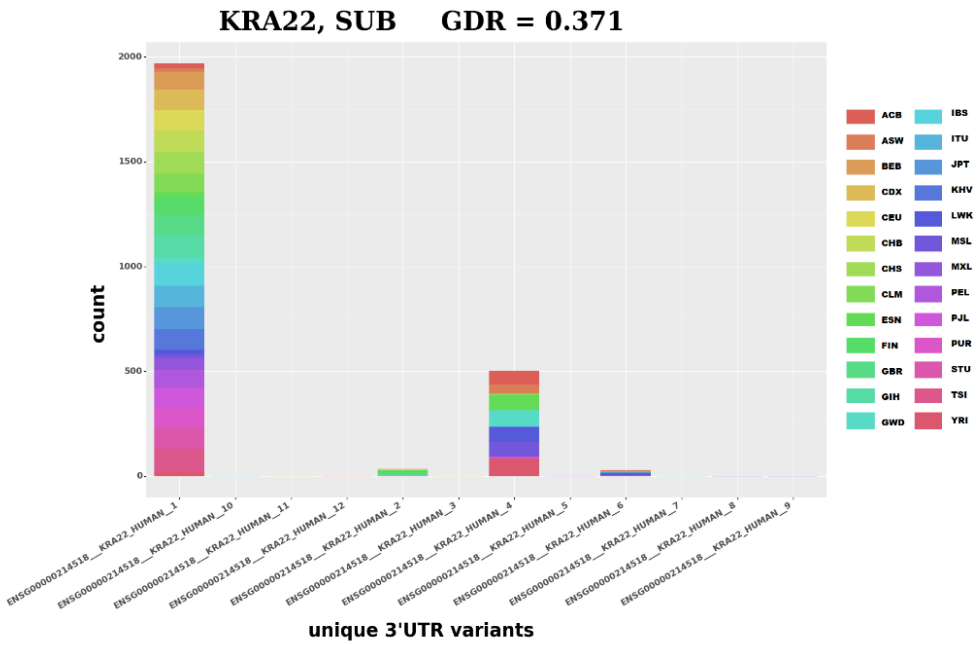


Figure 6.12: Unique sequence map of sub populations for the KRA22 gene 3'UTR, i.e., the same gene as displayed in figure 6.11, but for sub-populations (see figure 6.11 capture for details). Sub-population acronyms are as listed in table 4.1 in data section. Compared to Figure 6.12, variant 4 seem to be spread across African sub-populations. Variant 2 appears mostly present in the Finnish population (a closer inspection of these numbers showed that 19/34 of variant 2-samples originate from the Finnish population).

7 Discussion

The availability of large numbers of whole genome sequence data provides new opportunities for investigating the functional consequences of genetic variation within specific ethnicities. Human migration to all kinds of earthly environments, as well as sociocultural divergence, has required adaptation and caused radial evolution since our last common ancestor. In this study, this is shown to be reflected in the 3'UTR sequences of more than 1000 of our genes. This result underpins the importance of including all ethnic lineages in genomic studies to fully understand our human blueprint. 3'UTRs are involved in gene regulation in multiple ways, including being targeted and repressed by miRNAs. As non-coding regions are under reduced selection pressure compared to coding regions, ethnicity specific mutations in these regions may be associated with environmental adaptation through regulatory mechanisms. However, given the size and complexity of this data, deeper investigation requires new analysis tools that can perform screening and processing in a high throughput manner. The GDR metric introduced in this work shows promising results to aid in this challenge.

The GDR quantifies the groupwise structure in a phylogenetic tree. That is, it can indicate whether a group of samples are evolutionarily more closely related to each other than to the other grouped samples. This quantification can indicate a relationship between groupwise sequence variation and the group-criteria feature. An understanding of how the GDR behaves as a metric in a simple tree consisting of two clades and two groups is achieved in the simulation, giving a clearer view of when the GDR metric is an appropriate approach. The metric is tested on a genome-scale dataset of 3'UTR samples, which shows its potential for analysing large amounts of sequence data through phylogenetic tree clustering. However, the absence of a tree-reliability measure in this study set suggests the findings should be interpreted with a degree of caution. Interpretations of the results are thus mostly discussed in general terms, such as how the GDR potentially can aid to disclose evolutionary patterns in the miRNA-regulatory network.

7.1 Data - phylogenetic trees

The procedure of transforming information from physical DNA samples to a GDR-metric requires many individual, error-prone processing steps. The whole data sampling procedure from DNA extraction to assembly and gene annotation was performed by the 1000 Genomes project, and while the annotation was reviewed here as much as possible, it was nevertheless trusted to be high-quality data. Pre-processing of the raw data, i.e., extraction of 3'UTRs, MSA and clustering was evaluated and considered to be sufficient for the aim of this project.

Since the main goal was to find and develop a suitable methodology to measure group-related tree structure in phylogenetic trees, less consideration was given to the details of the clustering process that was performed to generate the phylogenetic trees that were the primary starting point for this study. It is well known that the Neighbor-Joining algorithm does not provide the most reliable trees, but it captures the main patterns arising from sample variation (81). However, it is unlikely it would make a substantial difference to the final GDR estimate and significance if the trees were generated with another clustering algorithm. Nevertheless, it could, for example, be interesting to compare GDR results for the same data for trees generated with the Maximum Likelihood approach.

A considerable limitation of the trees, on the other hand, is the lack of a concrete reliability measure, such as bootstrap values. Thus, if the tree structure is not robust, a significant GDR could be a false positive. Hence, the assumption of reliable tree structures to measure GDR is potentially violated in the presented case study.

Finally, the type of variation in the 3'UTRs (in a cellular functional context) is not identified. Detailed interpretations of the findings would require additional downstream analysis. i.e., the GDR results direct the researcher to which 3'UTRs may be interesting for further exploration, for example, in terms of functional interpretation. Thus, the result is mainly reviewed in the context of exploring potentials of the GDR metric, but also examined in the context of evolutionary mechanisms in the miRNA-targeting network.

7.1.1 Data and software practicalities

The 9381 CSV-formatted phylogenetic distance matrices required approximately 300GB disk space in unzipped format. Storage required is thus small enough to fit on a general laptop (at least in zipped or binary format), however, storage could be further optimized by collapsing samples comprising zero-distances in the matrix. The overview of nonzero-distances in the trees in figure 6.8 shows that many of the sample pairs in each tree had zero-distance, i.e., the matrices were filled with zeros. More importantly, this solution could increase computational efficiency. In the calculation procedure, memory usage could be reduced since a smaller data matrix is read and stored in the Willow-tree object at each timepoint. Also, since the GDR metric requires an all-versus-all sample comparison, the computational time increases quadratically ($O(n^2)$ for an increased number of samples. By collapsing zero-distance samples, the program could skip arithmetic operations between these pairwise combinations and reduce CPU requirement per processed tree.

To do this, a common "leaf-node" reference that would represent the collapsed zero-distance samples could be created and referred to through a list, for instance. This attribute is planned to be included in future updates of Willow.

7.2 The GDR dispersion metric

Even though the GDR can be used to search for variation patterns in large-scale sequence data, there is one particular issue that should be considered. The GDR calculation requires a phylogenetic tree, which is a more computationally comprehensive task compared to other clustering algorithms. Phylogenies are advantageous for biologically related topics mainly because of their inclusion of substitution models, which account for the emergence of the observed sequence variation across time. This way, the GDR of a phylogeny describes the amount of tree structure with respect to evolutionary divergences of the defined groups. If, on the other hand, the research is targeted towards detecting genotype - phenotype relationships, the inclusion of substitution probability parameters in the phylogenetic estimation makes no sense: as the state of sequence variants in a measured phenotype is already fixed. In this case,

sequences should be clustered based on their current state, which can further be coupled to a phenotype (e.g., GWAS). Then, more efficient clustering algorithms could be used, such as K-means clustering, which clusters the data in a higher dimensional space. Hence, the researcher should ask: *Am I interested in groupwise evolutionary relationships, or present-day groupwise variation?*

7.2.1 GDR vs. UniFrac

UniFrac is conceptually comparable to the GDR, as both methods utilize branch-length information to estimate distinctiveness between groups in phylogenetic trees from sequence data. Hence, it could potentially have served a similar purpose; instead of letting leaf nodes represent taxa, and defined groups represent environments they originate from, leaf nodes can represent 3'UTR sequence data and the defined groups could represent ethnic population groups. However, the metrics exhibit considerable distinctions that made UniFrac unsuitable to meet the aims here.

Firstly, the metrics target different questions: UniFrac is traditionally utilized to identify groups, (e.g., microbial communities), whereas the GDR aims to measure the degree of group structure for predefined groups. To identify clusters as unique, UniFrac measures the amount of sample (i.e., taxa) overlap between the environments as the fraction of a tree that is unique to one of two groups. It is then tested if the overlap is significantly small to confidently classify the groups of samples as distinct, or, in taxa-contexts: classify environments as separate communities. Group overlaps are not permitted in the current implementation of GDR, which makes up a major distinction.

Secondly, the distinct target questions are served by distinct ways of how the branch length information are exploited to measure group structure. The UniFrac is based on a shared/unshared-branch length distance-principle. The location of a sample relative to another sample in the tree topology affects the resulting UniFrac value. For example, in unweighted UniFrac, one sample can determine whether a branch counts as "shared" between groups in the UniFrac estimate, even though several other samples stemming from the same branch are unshared. Weighted UniFrac, on the other hand, accounts for relative sample abundances through weights given to each branch in a tree and is hence less sensitive to few sample

differences. However, they both are distinct from the GDR, where branch lengths between pairs of samples are measured respectively, unaffected by how other sample's branch lengths influence the group-distance estimations.

Thirdly, the result of the UniFrac method is inconvenient to utilize for large-scaled phylogenetic forests. With the UniFrac, groups are compared in a pairwise manner and must be interpreted accordingly. For multiple groups, it results in a distance matrix of values between all compared groups/environments. To analyse and interpret UniFrac-distance matrices for 9381 phylogenetic trees twice, for 26 and 5 population groups respectively, would be infeasible. The aim to obtain a metric suitable for large-scaled analysis was hence not met with UniFrac, in opposite of GDR which gives a single value for each tree.

7.3 GDR Simulation

By comparing the simulated results to the 3'UTR results, it should be noted that the simulation is highly reduced in terms of tree topology complexity. Since the number of clades and number of groups were constant in the simulation, only the relationship between sample size and clade size is explored. The distance between the two clades thus became irrelevant, since the distance describing the relationship between all perturbations, either within or between groups, is equal or zero. Hence the simulation is reduced to a binary problem: equal or unequal samples. In real contexts, there are usually more than two sample variants measured relative to each other, which increases complexity. How GDR behaves in more complex situations could be further researched, but some insight can gained from inspecting the GDR values for the 3'UTR trees described below.

The plot of simulated GDR values in figure 6.1 illustrates how the GDR metric varies relative to the number of samples/leaf nodes in the phylogenetic tree, where GDR resolution increases with sample size. The span of possible GDR values is defined by the possible arrangements of distinct group samples into distinct clades in a tree, thus it depends on the total sample size, the number of defined groups and the number of distinct clades in a tree. The more arrangements possible, the more distinct GDR values can be attained, i.e., it achieves higher resolution. The maximum number of possible arrangements in a tree is obtained when the

tree's clades contain an equal number of samples. The number of possible arrangements in a tree affects how probable each outcome is, hence, the probability of an observed GDR is related to the number of groups defined, the number of leaf nodes and the number of distinct clades of a tree.

If the sample size is small, a single permutation may cause a large change in the GDR value. This is consistent with statistical theory – a larger sample size leads to a more robust result, since the change of a single sample instance does not affect the outcome drastically. When sample size is low, however, it is clearly illustrated that permutation of a single or few samples has a large effect on the resulting GDR value. Therefore, a calculated GDR value must be seen in relation to the number of samples contained in the analysed phylogenetic tree.

For example, for a simulated tree constructed from a total of 20 samples with 5 permuted samples, i.e., these samples are incorrectly clustered from their "true" group relative to the hypothesized ground truth, the GDR is 0.56. For the same tree with one more permuted sample (6 permuted samples in total), the GDR is 0.67, hence a single perturbation causes 0.11 difference to the GDR. This large shift introduced by a single sample permutation demonstrates that the tree-structure must be reliable for the GDR to be correctly informative. If the tree-structure exhibits high uncertainty, this uncertainty is inherited by the GDR measure. Therefore, if the GDR is enforced for phylogenetic trees with a small number of samples/leaf nodes, a reliability measure such as bootstrap values or jack-knifing is essential, and the GDR is not a suitable metric for unreliable, small phylogenetic trees.
This elucidates where traditional approaches and the GDR metric are valuable respectively: the GDR is a useful metric for large-scale sequence analysis, whereas traditional approaches can be more profitable for small-scale analysis (for example, a study of a small number of samples and one or two genes). Even though small, less reliable trees are inappropriate for GDR estimation, they can aid researchers to visually gain an overview and to present evolutionary relationships, without declaring a definite metric or trusting every sample to be clustered correctly. Visual inspection of a tree constructed from 2548 samples, on the other hand, is difficult to interpret visually. Labels become unreadable and there are very many branches to interpret relative to each other. A computer, on the other hand, is better suited for this task.

On the other hand, when the sample size is large, a permutation of one sample from one cluster to the other does not affect the GDR to a large degree, as seen in figure 6.1. In turn, this relaxes somewhat the necessity for reliable tree construction. While the tree should be constructed in the most trustworthy manner, the GDR will not be affected drastically for a few wrongly clustered samples, as long as the major tree-structure patterns are captured. Therefore, less accurate but faster tree-clustering algorithms can be considered sufficient. NJ with NINJA is thus acceptable despite lower accuracy than other more computationally expensive tree-building approaches such as UPGMA. It captures patterns that will be caught by the GDR metric in an informative, reliable manner, despite a few misclassifications.

7.4 Simulation of p-values

At first glance of the simulated p-values from plot 6.2, it appears as if most group constellations are significant. Based on this, one could think it is likely to obtain a significant result. However, it is important to keep in mind that the simulation was performed in a way where the "observed" GDR for each constellation of group sizes and permutations was the least likely arrangement possible, since all group members of Clade 2 purely belonged to a single group for all calculated GDRs. In the randomization procedure, where random GDR values were calculated to obtain a null distribution for the observed GDR values, all resulting randomised GDR values would hence turn out equal to or larger than the observed GDR, i.e., all samples in Clade 2 were a random mix of Group 1 and Group 2 samples. It would therefore be expected to find many low p-values for the observed GDRs.

By further comparing the plot of p-values in figure 6.2 to figure 6.3 of p-values corrected for multiple testing, it is clear that even the least likely sample arrangements result in an insignificant GDR value after all. When many test-settings are performed at once, as here, the observed GDR values are likely to occur by chance. The denser area of dark purple dots towards the upper left corner of the plot correspond to the tendency of larger group size and fewer misclassifications (i.e., where groups are better classified into separate clades in the tree) leading to statistical significance. Moreover, the plot shows that no simulated instances with group size below ~50 reach significance. Considering the highly reductional simulation

setting, this is not unexpected, but it shows how the GDR is a more suitable metric for phylogenetic trees with larger sample sizes and less reliable for smaller trees.

As the GDR metric was made for large-scale data analysis, these results highlight the necessity of performing correction for multiple testing to obtain statistically valid results. Correcting for multiple testing in a scenario where GDR is calculated for many group constellations in the same tree is also needed.

7.5 GDR population distributions in 3'UTRs

The GDR distribution plots give an overview of population group-structure in > 9000 3'UTR-trees at two group levels. 5 and 26 population groups were defined at each level respectively. The overall mean of the sub-GDRs is slightly lower compared to the overall mean of the super-GDRs, indicating overall there is more population-specific structure in 3'UTRs at the subpopulation level. Also, the distribution of super GDRs is more dispersed than the sub- GDR distribution, with more values above 1, indicating no population-specific structure. Since the defined sub-populations all are smaller fractions of the super-population groups, these numbers indicate a subtle pattern: overall, the more "detailed" group definitions result in stronger tree structure. In turn, it indicates that samples stemming from the same sub-population are clustered closer together relative to each other, whereas continent-wise variation is less profound. If this were to be interpreted in an evolutionary perspective, one could say that adaptational patterns in 3'UTRs are better shown at smaller ethnic levels than the continental groupings of human ancestries. In turn, this may indicate higher degree of ethnicity-specific gene regulation adaptations for smaller ethnic lineages.

Of the 1055 GDR values that were statistically tested, 1005 were significant at ~90% level. A significant value does not necessarily imply a strong group structure, but the probability of observing even the upper range of significant GDR values (0.9-1.0) if no group structure was actually present, is small. 59 and 8 of the significant super- and sub GDRs were higher than 0.9, respectively, which imply limited structure overall, but the detected structure can be connected to the ethnic groupings.

Interpretation of the GDR value distributions here is a complicated task since the values represent population specific structure for unknown type of variation in 3'UTR sequences across the whole genome, for both significant and insignificant trees (where most are untested and unknown). How much valuable information can be retrieved from the statistically unverified trees is questionable, especially since reliability of tree structure is also unknown. Thus, further interpretations are highly hypothetical, but can nevertheless give insights of how the metric can be used to explore population structure in 3'UTRs.

Firstly, it should be noted that the distributions do not contain the trees where all sample distances were equal to zero. These were filtered out during GDR calculation, since it was already known that there would be no population specific structure in these trees. However, after reflecting upon how the GDR metric can give an overview of population-specific variation for all analysed genes, it is wrong to filter these before inspection of GDR distributions. These trees should rather obtain a GDR value equal to 1, indicating no tree structure. Hence, Willow has been modified to include these trees.

If it is assumed that most tree structures are reliable, it seems as if minor population-specific variation is present in most 3'UTRs, i.e., those having a GDR < 1.0, with means of 0.96 and 0.92 for super- and sub populations respectively. Overall, this could be interpreted as if some evolutionary patterns are developed in these genomic regions since separation of human lineages, to such an extent that it can be detected at large scale GDR estimation of ~9000 3'UTRs. The 3'UTRs with lower GDR values show stronger population-specific signals, and filtering for these for further investigation could be an appropriate next step, also being the criteria used for selecting genes for statistical testing. This way, GDR as a metric shows its core area of application: to gain overview and perform a first filtering of genomic sequences that are related to the predefined groups.

Setting a GDR-limit to filter for "interesting" genomic regions is context-dependent and must be evaluated individually for each research setting. For example, in this setting, to select for 3'UTRs that clearly showed human adaptation patterns between all populations since separation of all human lineages, a stricter limit (i.e., lower GDR value) would be appropriate, as this would only select trees showing strong structure for more than one or a few populations. On the other hand, to select 3'UTRs that only show some population-specific

variation for one or a very few populations, the cut-off for the GDR upper value could be relaxed. Thus, the metric must be understood in relation to the research interest.

For example, the phylogeny of 3'UTRs from gene *NDUS5* obtained a GDR = 0.804 at super-level. By inspecting how population samples are distributed across the unique sequences upon which this phylogeny is based shown in figure 6.9, it can be seen there are only 2 unique sequences, where most human of one of the variants originate from Europe and the Americas, whereas all other humans keep another variant. Hence, the variant appears to be mostly present in people from the two continents, which causes the modestly indicated structure through the GDR measure.

Another situation is shown with the HAUS4 gene (at super-level), also estimated as significant despite its high GDR of 0.971, see plot 6.10. It comprises 12 unique 3'UTR sequence variants, but most are only present in one or a few human samples and most human samples are equal (i.e., most humans from all continents contain the same variant). Most of the unique variants apart from this are present in the African population, where variant 3 and 6, are present in around 100 and 200 African samples, respectively. These are likely to have caused the significant outcome despite a high GDR. This way, signals of subtle population-specific structure is detected through the GDR metric and statistical validation.

Since the case study here is a rough estimation of all population groups added and analysed together, many of the significant GDRs in the range 0.8-1.0 are possibly caused by situations like these. Thus, to distinguish unique sequences for single populations, it could be interesting to further map and analyse which populations demonstrate a higher tendency of embodying unique, distinct sequences for the 3'UTRs.

7.6 A closer look at 3'UTRs with low GDR

When the first filtering of significant trees with low GDR is made, it could be interesting to look more closely into which populations are clustered more closely together in the trees. The GDR could also be used for this purpose, but in a slightly different way. Instead of calculating an overall GDR for population-structure in a tree, the GDR could be calculated for each single defined population. By modifying the GDR calculation to only average across the intra-group pairwise distance for one population, divided by the inter-group pairwise distances between samples of that population and all other samples, the ratio of mean within- and between group-sample distances is obtained for the single population only.
By calculating the "single-group GDRs" for each population in a tree, the populations can be ranked according to how close their samples are clustered in a tree, relative to all other samples. One could not imply a statistically significant difference between the ranked single-group GDRs, but since it is already known that the tree structure is significant and reliable (through the pre-filtering GDR analysis), this measure would a valid indicator.

The population-distribution across the unique 3'UTR sequences that represent the KRA22-gene is shown in figure 6.11 and 6.12. This gene was selected for inspection because of its low GDR values, being 0.394 for super-population and 0.371 for sub-populations. For super populations, the plot clearly show that African ancestries embody one sequence variant, which is rarely present in populations of other continental ancestries. By inspecting the sub-population plot, it also seems as if another unique variant is more common in the Finnish population (variant 2, shown in plot 6.12). The type of variation and functional interpretation of these variants could be interesting to investigate in more depth. Using this approach, the GDR has aided identification of a 3'UTR that shows population specific variation at both the super- and sub- population level.

A next step, for example, could be to predict miRNAs that target this gene, i.e., through either of the many targeting prediction tools (such as miRAW as it has demonstrated better performance and can be integrated into automated high throughput analyses). Furthermore, the GDR calculation procedure could be run for the predicted miRNAs (assuming miRNA sample data is available for the same ethnic populations as explored here, in sufficient

quantities). In turn, this could show if the miRNA(s) that were predicted to target the KRA22 gene also encompasses distinct variants for the African population.

In the imagined scenario of a successful hit, multiple opportunities for further exploration become possible. For example, the targeting interaction could be verified experimentally (if not already available). In this way, the GDR would have directed targeting-interaction research. Moreover, it could be investigated if the discovered variants affect the targeting interaction, and subsequently protein expression, for example through expression analysis. This way, a more detailed functional interpretation and impact of the distinct African variant of KRA22 – miRNA regulation could be made and potentially aid medical research.

Even though this hypothetical example of "success" requires many steps from a first population-specific signal to medically applicable information, it shows how analysis of large-scale sequence data analysed with the GDR method can direct research. As more sequence data becomes available as sequencing machines are running hot in this very moment, data availability is prospected to increase within few years. Yet, there are many genomic regions remain to be explored and interpreted, where this kind of overview gives the researcher a eagles' eye rather than searching for prey at the ground level.

8 Conclusions

The GDR metric is suitable for analysis of evolutionary patterns in large scale sequence data. However, it relies on several pre-processing steps that require careful consideration. Computational demands for null-distribution calculation, necessary to perform statistical testing provide potential limits, but future code optimizations in Willow are conceivable to speed up calculation.

Nevertheless, the GDR metric provides novel research angles and can give statistically testable quantification of predefined groupwise phylogenetic structure, allowing characterization of many more trees than would be possible using traditional approaches.

The GDR simulation shows that the GDR, as a discrete metric, achieves higher resolution for more sequence samples in the tree, and this should be considered when interpreting results. The number of distinct clades in a tree and the number of groups defined will also influence the significance of a GDR value, since the options and probabilities of GDR values in a tree depend on the possible group arrangements.

The approach was developed as a starting point to investigate ethnic population-specific structure in the miRNA-regulatory network. The integrated Python software Willow was developed to streamline further utilization of the approach. Here, it was first tested on 3'UTR sequence data involved in the miRNA interactome. More than 9000 phylogenetic trees, each containing >2500 clustered 3'UTR samples, were analysed for population-specific evolutionary patterns, of which >1000 trees were statistically tested. Ethnicity specific evolutionary patterns for populations at both continent-level and smaller ethnic population-levels were revealed for ~1000 3'UTR sequences, but this result must be taken with a pinch of salt since trees lacked any form of reliability metric. However, it could be speculated the observed ethnicity-associated patterns are related to human evolutionary adaptation through the miRNA – 3'UTR regulatory network. Regardless, the observation underpins the

importance of including human individuals of all ethnic origins in genomic studies, to capture the holistic picture of human evolution and variants to support future development of precision medicine.

The GDR metric is hereby shown to be a relevant, convenient approach to perform a first stage filtering of genomic sequence regions that show evolutionary patterns related to ethnic populations. It further holds potential to benefit other research of group-related evolutionary patterns in genomic sequence data, such miRNA sequences.

9 Further research

The approach is here shown to aid research on large-scale patterns in genomic sequence data. The analysis of 3UTR sequences can be replicated with inclusion of bootstrap values in the phylogenetic trees and results can be analysed in more details. For example, the genes can be investigated with a Reactome-search, to see which parts of the cellular machinery they may contribute to and gain insights to cellular functions that may be affected by population-specific variation.

Further, the approach can be used to measure ethnicity-associated patterns from other sequence elements in the miRNA-targeting network, to search for distinct human regulatory patterns. Data from other large-scale databases such as H3Africa project, TOPmed project, GenomeAsia 100K Project and UK Biobank project can be included in the analysis, to gain even deeper understating of ethnic variations. Findings can potentially aid in the search of functional interpretation of all the discovered variation in the miRNA interactome.

There are many potentials for what can be researched with the GDR metric in genomic sequence data, beyond ethnicity-associated patterns. For example, groups can be defined as age, sex or disease subtypes. Variation in all kinds of genomic regions that are suitable for phylogenetic tree-construction can be explored with the GDR, as long as the aspect of evolution is important in the target question.

Willow is intended to be improved in terms of user-friendliness and more integrated statistical analysis. All current R-code will be converted and contained within the software for a fully integrated pipeline.

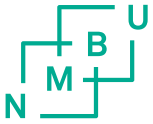
10 References

1. Ben Or G, Veksler-Lublinsky I. Comprehensive machine-learning-based analysis of microRNA-target interactions reveals variable transferability of interaction rules across species. BMC Bioinformatics. 2021;22(1):264.
2. Chipman LB, Pasquinelli AE. miRNA Targeting: Growing beyond the Seed. Trends Genet. 2019;35(3):215-22.
3. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Frontiers in Endocrinology. 2018;9(402).
4. Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. Genome Biol. 2019;20(1):18.
5. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.
6. Whose genomics? Nat Hum Behav. 2019;3(5):409-10.
7. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4.
8. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91.
9. Drozda K, Wong S, Patel SR, Bress AP, Nutescu EA, Kittles RA, et al. Poor warfarin dose prediction with pharmacogenetic algorithms that exclude genotypes important for African Americans. Pharmacogenet Genomics. 2015;25(2):73-81.
10. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016;375(7):655-65.
11. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. 2016;17(1):157.
12. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer. 2010;10(6):389-402.
13. Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, et al. SNPs in human miRNA genes affect biogenesis and function. RNA. 2009;15(9):1640-51.
14. Bartel DP. Metazoan MicroRNAs. Cell. 2018;173(1):20-51.
15. Ferro E, Enrico Bena C, Grigolon S, Bosia C. microRNA-mediated noise processing in cells: A fight or a game? Comput Struct Biotechnol J. 2020;18:642-9.
16. Ebert MS, Sharp PA. Roles for microRNAs in conferring robustness to biological processes. Cell. 2012;149(3):515-24.
17. Kim D, Chang HR, Baek D. Rules for functional microRNA targeting. BMB Rep. 2017;50(11):554-9.
18. Devanna P, Chen XS, Ho J, Gajewski D, Smith SD, Gialluisi A, et al. Next-gen sequencing identifies non-coding variation disrupting miRNA-binding sites in neurological disorders. Mol Psychiatry. 2018;23(5):1375-84.
19. Genetics vs. Genomics Fact Sheet: National Human Genome Research Institute; 2018 [Available from: https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics.
20. Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. Genome Biol. 2017;18(1):36.
21. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-9.
22. Zou H, Wu LX, Tan L, Shang FF, Zhou HH. Significance of Single-Nucleotide Variants in Long Intergenic Non-protein Coding RNAs. Front Cell Dev Biol. 2020;8:347.

23. Burgess DJ. The TOPMed genomic resource for human health. Nature Reviews Genetics. 2021;22(4):200-.
24. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Research. 2017;27(5):677-85.
25. Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, et al. Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. PLoS One. 2009;4(11):e7958.
26. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020;586(7831):741-8.
27. Bunn HF. The triumph of good over evil: protection by the sickle gene against malaria. Blood. 2013;121(1):20-5.
28. Kaariainen H, Muilu J, Perola M, Kristiansson K. Genetics in an isolated population like Finland: a different basis for genomic medicine? J Community Genet. 2017;8(4):319-26.
29. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature. 2017;544(7649):235-9.
30. Njolstad PR, Andreassen OA, Brunak S, Borglum AD, Dillner J, Esko T, et al. Roadmap for a precision-medicine initiative in the Nordic region. Nat Genet. 2019;51(6):924-30.
31. Rito T, Vieira D, Silva M, Conde-Sousa E, Pereira L, Mellars P, et al. A dispersal of Homo sapiens from southern to eastern Africa immediately preceded the out-of-Africa migration. Sci Rep. 2019;9(1):4728.
32. Shar N. Evolutionary Patterns in the Genus Homo. INTERNATIONAL JOURNAL OF HUMAN GENETICS. 2021;21.
33. Fatima R, Shar N. Evolutionary Patterns in the Genus Homo. 2021.
34. Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, et al. Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. Am J Hum Genet. 2006;78(5):889-96.
35. Human Genome Project Results genome.gov: National Human Genome Research Institute; 2018 [Available from: https://www.genome.gov/human-genome-project/results.
36. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.
37. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: lessons from gnomAD. Human Mutation. 2021;n/a(n/a).
38. Kaiser J. 200,000 whole genomes made available for biomedical studies. Science. 2021;374(6571):1036.
39. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
40. Exome Aggregation C, Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 2016:030338.
41. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190-5.
42. SNPs National Institutes of Health: National Cancer Institute; [Available from: https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/snp.
43. Igo RP, Jr., Kinzy TG, Cooke Bailey JN. Genetic Risk Scores. Curr Protoc Hum Genet. 2019;104(1):e95.
44. Kullo IJ, Jouni H, Austin EE, Brown SA, Kruisselbrink TM, Isseh IN, et al. Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates: Effect on Low-Density Lipoprotein Cholesterol Levels (the MI-GENES Clinical Trial). Circulation. 2016;133(12):1181-8.
45. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

46. Grace Tiao JG. gnomAD v3.1 New Content, Methods, Annotations, and Data Availability gnomAD news: gnomAD; 2020 [Available from: https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability/.
47. Mulder N, Abimiku A, Adebamowo SN, de Vries J, Matimba A, Olowoyo P, et al. H3Africa: current perspectives. Pharmgenomics Pers Med. 2018;11:59-66.
48. GenomeAsia KC. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019;576(7785):106-11.
49. Rolle K, Piwecka M, Belter A, Wawrzyniak D, Jeleniewicz J, Barciszewska MZ, et al. The Sequence and Structure Determine the Function of Mature Human miRNAs. PLoS One. 2016;11(3):e0151246.
50. Perenthaler E, Yousefi S, Niggl E, Barakat TS. Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. Front Cell Neurosci. 2019;13:352.
51. Zhen Y, Andolfatto P. Methods to detect selection on noncoding DNA. Methods Mol Biol. 2012;856:141-59.
52. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med. 2018;50(8):1-8.
53. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
54. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446-50.
55. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2001;2(11):986-91.
56. Bhatti GK, Khullar N, Sidhu IS, Navik US, Reddy AP, Reddy PH, et al. Emerging role of non-coding RNA in health and disease. Metab Brain Dis. 2021;36(6):1119-34.
57. Ghafouri-Fard S, Eghtedarian R, Taheri M, Beatrix Bruhl A, Sadeghi-Bahmani D, Brand S. A Review on the Expression Pattern of Non-coding RNAs in Patients With Schizophrenia: With a Special Focus on Peripheral Blood as a Source of Expression Analysis. Front Psychiatry. 2021;12:640463.
58. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014;15(8):509-24.
59. Mayr C. What Are 3' UTRs Doing? Cold Spring Harb Perspect Biol. 2019;11(10).
60. Chen K, Rajewsky N. Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes. Cold Spring Harb Symp Quant Biol. 2006;71:149-56.
61. Liu G, Zhang R, Xu J, Wu CI, Lu X. Functional conservation of both CDS- and 3'-UTR-located microRNA binding sites between species. Mol Biol Evol. 2015;32(3):623-8.
62. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N. Cell-type-specific signatures of microRNAs on target mRNA expression. Proc Natl Acad Sci U S A. 2006;103(8):2746-51.
63. Lee Y, Jeon K, Lee JT, Kim S, Kim VN. MicroRNA maturation: stepwise processing and subcellular localization. Embo j. 2002;21(17):4663-70.
64. Lau AG, Irier HA, Gu J, Tian D, Ku L, Liu G, et al. Distinct 3'UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). Proc Natl Acad Sci U S A. 2010;107(36):15945-50.
65. Schratt G. Fine-tuning neural gene expression with microRNAs. Curr Opin Neurobiol. 2009;19(2):213-9.
66. Ding J, Li X, Hu H. TarPmiR: a new approach for microRNA target site prediction. Bioinformatics. 2016;32(18):2768-75.
67. Shah V, Shah J. Recent trends in targeting miRNAs for cancer therapy. J Pharm Pharmacol. 2020;72(12):1732-49.
68. Li J, Zhang Y. Current experimental strategies for intracellular target identification of microRNA. ExRNA. 2019;1(1):6.
69. Chou C-H, Chang N-W, Shrestha S, Hsu S-D, Lin Y-L, Lee W-H, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Research. 2016;44(D1):D239-D47.

70. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Research. 2009;37(suppl_1):D105-D10.
71. Pla A, Zhong X, Rayner S. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. PLoS Comput Biol. 2018;14(7):e1006185.
72. Subramanian M, Li X, Hara T, Lal A. A Biochemical Approach to Identify Direct MicroRNA Targets. Methods in molecular biology (Clifton, NJ). 2015;1206:29-37.
73. Carbonell J, Alloza E, Arce P, Borrego S, Santoyo J, Ruiz-Ferrer M, et al. A map of human microRNA variation uncovers unexpectedly high levels of variability. Genome Med. 2012;4(8):62.
74. Du J, Li M, Huang Q, Liu W, Li WQ, Li YJ, et al. The critical role of microRNAs in stress response: Therapeutic prospect and limitation. Pharmacol Res. 2019;142:294-302.
75. Weng JF, Thomas DA, Mareels I. Maximum parsimony, substitution model, and probability phylogenetic trees. J Comput Biol. 2011;18(1):67-80.
76. Weiß M, Göker M. Chapter 12 - Molecular Phylogenetic Reconstruction. In: Kurtzman CP, Fell JW, Boekhout T, editors. The Yeasts (Fifth Edition). London: Elsevier; 2011. p. 159-74.
77. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach.
78. Arenas M. Trends in substitution models of molecular evolution. Front Genet. 2015;6:319.
79. Rokas A, Charlesworth D. Molecular Evolution and Phylogenetics. By M. Nei and S. Kumar. Oxford University Press. 2000. ISBN: 0-19-513584-9 (hbk); 0-19-513585-7 (pbk). xiv+333 pages. Price: £65 (hbk); £32.50 (pbk). Genetical Research. 2001;77(1):117-20.
80. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5(1):113.
81. Wheeler TJ. Large-Scale Neighbor-Joining with NINJA. Algorithms in Bioinformatics. 5724. Berlin: Springer; 2009.
82. Balaban M, Moshiri N, Mai U, Jia X, Mirarab S. TreeCluster: Clustering biological sequences using phylogenetic trees. PLoS One. 2019;14(8):e0221068.
83. Lozupone C, Hamady M, Knight R. UniFrac – An online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics. 2006;7(1):371.
84. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012;28(16):2106-13.
85. Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. 2007;73(1):278-88.
86. Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M. MAXIMUM-LIKELIHOOD ESTIMATION OF POPULATION DIVERGENCE TIMES AND POPULATION PHYLOGENY IN MODELS WITHOUT MUTATION. Evolution. 1998;52(3):669-77.
87. Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. Wellcome Open Res. 2019;4:50.
88. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC. Phylogeny and the origin of HIV-1. Nature. 2001;410(6832):1047-8.
89. Konishi T, Matsukuma S, Fuji H, Nakamura D, Satou N, Okano K. Principal Component Analysis applied directly to Sequence Matrix. Scientific Reports. 2019;9(1):19297.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway