DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
396

LUDOVICA MOLINARO

Ancestry deconvolution of Estonian, European and Worldwide genomic layers: a human population genomics excavation

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
396

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
396

LUDOVICA MOLINARO

Ancestry deconvolution of Estonian, European and Worldwide genomic layers: a human population genomics excavation

UNIVERSITY OF TARTU
Press
1632

Institute of Molecular and Cell Biology, University of Tartu, Estonia

Dissertation was accepted for the commencement of the degree of Doctor of Philosophy in Gene Technology on 27th of April, 2022 by the Council of the Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu.

Supervisors: Luca Pagani, PhD; Associate Professor, Department of Biology, University of Padova, Italy; Senior Research Fellow of Population Genetics, Institute of Genomics, University of Tartu, Estonia

Francesco Montinaro, PhD; Research Fellow, Department of Biology, University of Bari, Italy; Research Fellow of Population Genetics, Institute of Genomics, University of Tartu, Estonia

Mait Metspalu, PhD; Professor of Evolutionary Genomics, Institute of Genomics, University of Tartu, Estonia

Opponent: Priya Moorjani, PhD; Assistant Professor, Department of Molecular & Cell Biology, UC Berkeley, Berkeley, California, United States

Commencement: Room No 105, 23B Riia St, Tartu, on 20th of June 2022, at 16:00

Publication of this thesis is granted by the Institute of Molecular and Cell Biology and the Institute of Genomics, University of Tartu.

ISSN 1024-6479
ISBN 978-9949-03-901-2 (print)
ISBN 978-9949-03-902-9 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES

TABLES

LIST OF ORIGINAL PUBLICATIONS

I. **Ludovica Molinaro**, Francesco Montinaro, Burak Yelmen, Davide Marnetto, Doron M. Behar, Toomas Kivisild & Luca Pagani, West Asian sources of the Eurasian component in Ethiopians: a reassessment, *Scientific Reports*, Volume 9, Article Number 18811, December 2019, https://doi.org/10.1038/s41598-019-55344-y

II. **Ludovica Molinaro**, Davide Marnetto, Mayukh Mondal, Linda Ongaro, Burak Yelmen, Daniel John Lawson, Francesco Montinaro, Luca Pagani, A Chromosome-Painting-Based Pipeline to Infer Local Ancestry under Limited Source Availability, *Genome Biology and Evolution*, Volume 13, Issue 4, April 2021, evab025, https://doi.org/10.1093/gbe/evab025

III. Davide Marnetto, Katri Pärna, Kristi Läll, **Ludovica Molinaro,** Francesco Montinaro, Toomas Haller, Mait Metspalu, Reedik Mägi, Krista Fischer & Luca Pagani, Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals, *Nature Communications*, Volume 11, Article Number 1628, March 2020, https://doi.org/10.1038/s41467-020-15464-w

Author's contributions to the listed articles are as follows:

Ref. I: I participated in drafting the experimental design, performed all analyses, interpreted the results and wrote the manuscript

Ref. II: I participated in all analyses, performing all of them with the exception of constructing the Correlation-Assignment-Score Matrix, I interpreted the results and wrote the manuscript

Ref. III: I performed Local Ancestry deconvolution analyses on all the tested groups

ABBREVIATIONS

1KGs	1000 Genomes (Project)
aDNA	ancient DNA
AD	Ancestry Deconvolution
AT	Admixture Time
CP	ChromoPainter
CRF	Conditional Random Field
DNA	Deoxyribonucleic Acid
DT(s)	Decision Tree(s)
EM	Expectation-Maximisation
F_{ST}	Fixation index
GA	Global Ancestry
GWAS	Genome-wide Association Studies
HMM	Hidden Markov Model
ky(a)	Kilo years (ago)
LA	Local Ancestry
LAI	Local Ancestry Inference
LD	Linkage Disequilibrium
MCMC	Markov Chain Monte Carlo
mtDNA	Mitochondrial DNA
NNLS	Nonnegative Least Squares
PC(A)	Principal Component (Analysis)
RF	Random Forest
SNP(s)	Single Nucleotide Polymorphism(s)
SVD	Singular Value Decomposition
TMRCA	Time of the Most Recent Common Ancestor

1. INTRODUCTION

1.1 Encounter of people

Nowadays, we often assume that our modern way of life is necessarily better than the past ones, and that the more we go back in time, the more the past humans struggled with poor living conditions or unperforming technologies. It is no surprise then that historical and archaeological studies tackle the curiosity of many of us, as with them we are forced to reshape our system of modern beliefs and gasp before how much past populations could in fact achieve.

For example, Neanderthals were often depicted as some kind of monstrosity, able only to abide by the most cruel of animal tendencies, "that half-savage, half-animal past of our race" says H. G. Wells in The Grisly Folk and Their War With Men (in an imaginative anthropological article in the *Saturday Evening Post*, v193 #37, March 12, 1921). The 'caveman' archetype was so instilled in popular culture that it appeared almost impossible to find traces of Neanderthal art (Hoffmann et al. 2018), let alone inbreeding with humans (Sankararaman et al. 2014). However, misconceptions about the past do not stop at other human species.

Around 12.000 years ago groups of nomads in southeast Anatolia built an enormous archaeological structure made of multiple stone pillars up to 20 metres in diameter and 5.5 metres in height, some of them were even elaborately carved (Dietrich et al. 2019). For reference, the primal example that usually comes to mind in terms of megaliths, the Stonehenge, was built around 9000 years later (Pearson et al. 2007). It is still unclear how the monoliths in Turkey were moved or carved, even the purpose of the site is still debated (Banning 2011). Traces of plant food processing and hunting, without the presence of large storage facilities, point to seasonal activity within the site, indicating a cyclical presence of people (Dietrich et al. 2019). Were nomadic groups from 12.000 years ago so organised to meet periodically in one specific location in Turkey to build such an enormous site?

Despite the fact that nowadays long distance travelling is made quite easy, long travels are not a modern prerogative, not even comfortable travel. Athenaeus of Naucratis depicts in his 'The Deipnosophists' the great Syracusia: one of the most large and elegant ships of 240 BC, built with wood that could be sufficient for sixty triremes, with couches, mosaic, ivory doors and a temple dedicated to Aphrodite. We have lost trace of the great Syracusia, but not of Caligula's 'Giant ship', a over 100 metres barge found in Italy and linked to the 800-ton obelisk ship that transported the famous St. Peter's Square obelisk from Egypt.

Should we then be surprised by the large commercial trade that characterised the entire Mediterranean in the Bronze age? Or even more so, if one happened to walk around Tel Kabri, in Israel, should they be surprised to see Aegean (specifically Minoan) style frescos from the 17th century B.C.E (Cline, Yasur-Landau, and Goshen 2011)?

Historical and archaeological records continue to remind us to look at the past with unstagnant eyes. As much as we do nowadays, people in the past were organised, they moved, they travelled, they shared ideas, cultures and technologies: and we can still find traces of their encounters.

To fully comprehend our past history a single set of eyes cannot be sufficient, as no scientific field is fully independent. With this in mind, this thesis aims to look at the residual traces of past encounters through the genetic data.

2. LITERATURE OVERVIEW

2.1 Encounter of genes: layered effects of the admixture event

2.1.1 Encounter of genes

The consequences of an encounter can be traced throughout all the layers of human complexity, from cultural (religion, customs, tales (Bortolini et al. 2017)) to biological ones. When it comes to studying the encounter of people through biology, genetics serves as a great tool to trace such events.

By comparing the genetic differences between populations separated by either space or time we can trace the movement of humans throughout continents and years. Although most differences are shared between individuals regardless of their group, there are still small genetic details that can contribute to our understanding of human evolution.

When people meet and mix, their genetic makeup is passed on to their children, whose DNA becomes a mixture of their parents' DNA. We can follow the path upwards: we can trace an offspring's mother and father, grandmothers and grandfathers and so on. Following one's genealogical tree, we can see that the ancestors double at each generation, so that in k generations there will be 2^k branches that are theoretically related to any given person (Ralph and Coop 2013, Donnelly 1983). However, from a genetic perspective, the number of genetic contributors halves at each past generation, due to the nature of autosomal DNA. Consequently, while finding one specific genealogical ancestor far back in time by leveraging on the genetic features is doomed to fail, such a huge number of branches (2^k), when taken together, can give insight on the average genetic makeup of one's group of origin.

To trace branches through time and space we move away from a mere genealogical point of view, and instead look at individuals as a collection of traces that characterised their population of origin. Moreover, starting from contemporary populations we can trace back to the ones that contributed to their history.

If individuals mix to produce an offspring, populations then go through what is referred as admixture event to create admixed populations. The process of tracing the contributors of an admixture event is called ancestry deconvolution (AD).

2.1.2 A mosaic of ancestries

When using AD approaches we look at the genetic makeup of an admixed entity as a mosaic, where each tile is a genetic block that can be traced back to an ancestry or population that contributed to the admixture event.

Several different forces will act on these blocks, and, in turn, the mosaic pattern will show levels of variation throughout the individuals within the population. Despite tracing their history back to the same sources, no two individuals will look alike.

The major force that shuffles the mosaic tiles is recombination, which is the process by which chromosomes exchange genetic material. This is done by transferring information from one homologous strand to the other either by breaking, shuffling and then repairing the strands, or by copying genetic information from the opposite strand. Such shuffling happens with different intensity during the mitosis and, most importantly in our case, the meiosis. In fact, as the 'Great-reshuffler' (sic (Jobling 2004)), recombination plays a key role in generating variability at each new generation by generating new combinations of loci (or mosaic tiles) (Alves et al. 2017; Stapley et al. 2017).

However, recombination does not impact the entire genome equally: there are loci that are rarely affected by it, and thus rarely separated (Nachman 2002).

The lack of recombination between loci will lead to a correlation between them, favouring a phenomenon named Linkage Disequilibrium (LD): defined as the event by which alleles at two separate genetic loci are found more often together at a population level than would be expected based on their individual allelic frequencies (sic, (Rybicki et al. 2002)). Low recombination rates and physical proximity are among the causes for LD, upon these variables selection, mutations, genetic drift and finally, gene flows may occur (Slatkin 2008).

In an admixed group we will in fact find different types of correlation between SNPs that will cause different LD patterns (Falush, Stephens, and Pritchard 2003). Correlation between SNPs may be due to the variation of ancestry between individuals (named mixture LD) (S Gravel 2012); the admixing sources will contribute with their own LD pattern (background LD) (Pritchard and Przeworski 2001); however, even if the admixing sources show little to none LD, as long as their allele frequencies differ, they will raise the level of LD in the admixed group, that will then show patterns defined admixture LD (Liang and Nielsen 2014; Chakraborty R and Weiss K M 1988).

2.1.3 Global and Local Ancestry approaches

Ancestry Deconvolution (AD) is an approach that allows to analyse the genetic mosaic of an admixed group, trace the admixing contributors and further characterise the admixing event. Ancestry Deconvolution results may yield key information regarding demographic histories, but the approach can be used as a tool to solve broader scientific questions that do not solely revolve around understanding historical events.

Generally, to perform an Ancestry Deconvolution study one relies on Global ancestry (GA) methods and/or Local ancestry (LA) methods. GA methods allow to infer the average proportion of each ancestry that contributed to the admixture event, while LA approaches allow to infer which ancestry falls within a given

inherited locus, appointing the observed mosaic tile (whether it is an allele or a small genomic tract) to its ancestral contributor (Figure 1).

To achieve such a fine level of inferences, most LA approaches leverage on several assumptions and need to compare the target admixed group with a limited number of admixing sources, possibly in large sample size, to best capture the genetic variation within the admixed population. However, most of the time the actual sources of the admixture event are not available in unadmixed form, and one must rely on proxy groups to deconvolute the target population. In addition, genetically similar sources tend to hinder the LA assignment, which poses a great issue when deconvoluting a sub-continental admixture events, as in European populations. Such is the reason why most LA applications regard Latin and African American groups, characterised by divergent sources (Wangkumhang and Hellenthal 2018).

So while performing LA on an admixed population characterised by worldwide ancestries may be a task completely focused on discovering hidden and unknown past demographic events, applying LA on sub-continental admixtures, such as European populations, requires a methodological approach designed to comprehend the LA limitations first. Eventually, leveraging on the knowledge that mostly all human groups are admixed (see next paragraph), it comes naturally to propose AD and, more specifically LA, as approaches useful beyond the realm of demography studies.

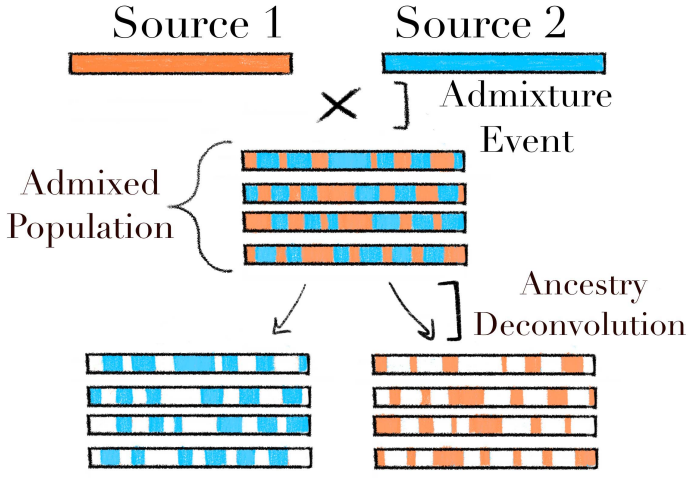

Figure 1. Schematic representation of an admixture event and subsequent Ancestry Deconvolution analysis from a genetic perspective. Source populations are indicated as 'Source 1' and 'Source 2' and their contribution is indicated in orange and blue, respectively.

2.1.4 Admixture throughout human populations

By looking at the genome as the result of past events, we will discover that encounters between populations are far from being rare in human history. Sometimes genomics studies can confirm and reinforce knowledge given by historical records (Ongaro et al. 2019), in other cases they can unravel unknown demographic histories (Hellenthal et al. 2014; Patterson et al. 2012).

For example, genetic data showed that most modern Indians (Reich et al. 2009) are characterised by two very different ancestral layers, indicating that two distinct human groups mixed in the past. One contributor was a population genetically close to Middle Easterners, Central Asians and Europeans. The origin of the second group remains unclear, although it may descend from a trifurcation with Andamanese and East Asian components (Yelmen et al. 2019). Similarly, the Indonesian area (Hudjashov et al. 2017) is genetically characterised mainly by two ancestries: Papuan and Asian. While the first one can be traced back to the first occupation of the area, the second one originated from Taiwan and arrived in the islands relatively recently.

On the other side, the genetic studies of the Americas benefitted from the large historical knowledge and described in finer detail the demographic events that occurred in the continent. As a consequence of the vast migrations and forced relocations, many modern American populations carry multiple ancestries, whose origin and proportions depend on the specific history of the area. For example, the Caribbean area was characterised by large amount of sugar plantations where massive exploitation of slave labour was performed, and nowaday modern populations show a high contribution (~88%) from sub-Saharan African ancestry; one of the largest migration of the 19th century, named the 'italian diaspora', saw over 10 million italians moving to different areas of the continent, event shown through the american populations' DNA as an italian-like component is heterogeneous and at pan-american level (Ongaro et al. 2019).

On the other hand, the African (Busby et al. 2016) continent does not possess large historical written records, thus demographic events can be inferred mainly through archaeological or genetic studies. As seen for many other human groups, signs of recent migrations can be seen in almost all sub-Saharan populations (Currie et al. 2013; Seidensticker et al. 2021) and from Africa into nearby Levantine and Southern Europeans (Moorjani et al. 2011).

Similar examples can be found throughout almost all human populations, as Hellenthal et al describe '*admixture happens to be an almost universal force that shaped modern human populations.*' (Hellenthal et al. 2014).

An essential key factor that allows us to describe in finer detail the old narratives of the past world is ancient DNA (aDNA). One of the most exceptional examples of the impact of aDNA in rephrasing the historical narratives and detangling intricate histories can be seen in Europe.

2.1.5 The Neolithic and post Neolithic expansion in Europe

To an extent unlike any other area, the study of Europe's past demographic history has been extremely favoured by finding and sampling large amounts of ancient DNA, allowing for an understanding of past and recent historical movements deeper than any other continent.

Although the first examples of anatomically modern humans more closely related to present day Europeans than to East Asian are found as early as ~39–36kya thousand years ago (kya) (Fu et al. 2016; Nielsen et al. 2017), their genetic contribution to the contemporary Europeans is minimal (Günther and Jakobsson 2016). Starting from at least 17 kya, before the Bølling–Allerød interstadial, a climatic warming during the last glacial period, a homogeneous Hunter-Gatherer ancestry became dominant in most of Europe (Fu et al. 2016; Bortolini et al. 2021). Such ancestry can be referred to as WHG, Western Hunter-Gatherer, to distinguish it from the Eastern Hunter-Gatherers (EHG) ancestry, found in Eastern European Mesolithic samples (~8kya), and Caucasus Hunter-Gatherers (CHG) ancestry, found in samples from the Caucasus region from the Upper Palaeolithic and Mesolithic period (Lazaridis et al. 2016; Jones et al. 2015).

Starting from 8,800 ya the first farmers expanded from the Anatolian area to Western Europe, while Hunter-Gatherers from Caucasus (CHG) migrated towards the northern steppe (Lazaridis et al. 2016).

While up until 6 kya Neolithic farmers in Europe show little traces of WHG ancestry, successively the two groups must have admixed as the Hunter-Gatherer ancestry increases to 20% in the Neolithic farmers samples found (Mathieson et al. 2018). By the end of the Neolithic period Western Europe was inhabited mainly by the descendents of the first farmers of Anatolian ancestry carrying genetic traces of the local Western Hunter-Gatherers. During the Bronze age another massive migration from the North-East arrived in Western Europe, bringing EHG, Iran and CHG ancestry, altogether identified as the Steppe ancestry.

In turn, the European genetic landscape carried three distinct ancestries: WHG-like, Anatolian-like and Steppe-like ancestry (Lazaridis et al. 2014). These three components still characterise all present day Europeans, although with different proportions depending on their geographic location (Haak et al. 2015).

After the Bronze Age migrations, several gene flows within the European continent reshaped the genetic composition of European sub-groups, decreasing the genetic differences (Günther and Jakobsson 2016). Post Bronze Age demographic events can be still studied with fine-scale analyses (Leslie et al. 2015; Gilbert et al. 2019; Pankratov et al. 2020; Martin et al. 2018; Bycroft et al. 2019; Saint Pierre et al. 2020; Drineas, Lewis, and Paschou 2010). As a testament of the high level of detail that can be obtained from fine-scale analyses see Figure 2, reproposed here from Pankratov et al. 2020. However, in such sub-continental context, AD seems to underperform, so that, to date, it is not possible to extract the ancestral components within European groups, but just characterise them with fine-scale analyses (Günther and Jakobsson 2016, Novembre and Stephens 2008, Lao et al. 2008; Marnetto et al. 2022).

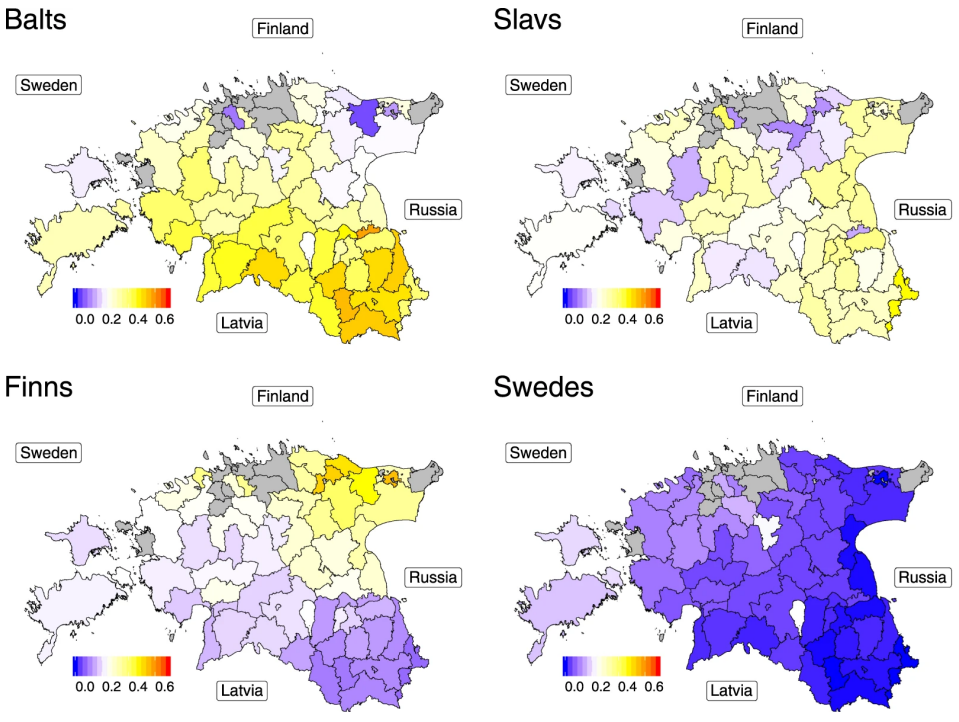

Figure 2. Differential genetic affinities of "Baltic", "Slavic", Finnish and Swedish groups in Estonians. Differential genetic affinities of «Balts» (Latvians and Lithuanians), «Slavs» (Belarusians, Poles, Russians, Ukrainians), Finns, and Swedes obtained with fine-scale approaches by Pankratov V. et al 2020.

2.1.6 The Bronze Age expansions in Eastern Africa

The Neolithic and post Neolithic waves that started from the Near East and expanded throughout all West Eurasia were not limited to the European continent. While from one side the steppe ancestry arrived to Northern India, on the other side the neolithic farmers contributed to the genetic history of the African continent. In fact, while Anatolian farmers during the neolithic period moved northwest towards central Europe, traces of the genetic makeup linked to the neolithic farmers are found in Eastern Africa a few millennia later (Lazaridis et al. 2016).

Initially, links between the Horn of Africa and non-African cultures were found through the studies of uniparental markers (mtDNA (Kivisild et al. 2004) and Y chromosome (Semino et al. 2002). The admixture event was detected also with the autosomal data of several modern Ethiopian groups, who showed two genomic layers: an African component and a West Eurasian (or non-African) component arrived in the area 3 kya (Pickrell et al. 2014; Pagani et al. 2012). An additional admixture event was found tracing back 23 kya (Hodgson et al. 2014).

To date, the ancient individual that best describes the African ancestry of Ethiopians is a 4500 years old hunter-gatherer found in a cave in Mota, Ethiopia (Llorente et al. 2015). The sample was free from any West Eurasian trace, confirming a recent date for the arrival of such a component into East Africa, which therefore could not be older than 4500 ya.

On the other hand, the exact origin of the non-African component has not been yet clarified. The non-African component has been consistently indicated to be genetically closer to the Mediterranean populations (see Figure 3, originally presented in Pagani et al. 2012), however studies indicated either a Sardinian-like ancestry (Llorente et al. 2015; Pickrell et al. 2014), or a similarity with farmers from the Neolithic Levant (Lazaridis et al. 2016). An additional link to the Levant area is given by the presence of the Ethiosemitic language group in Ethiopia, thought indeed to have originated in the Levant area and arrived in the Horn of Africa 3 kya (Kitchen et al. 2009).

Being an Anatolian source or a Levantine one, results point to the source of the backflow being genetically similar to '*the one that fueled the Neolithic expansion into Europe*' (sic, (Llorente et al. 2015)).

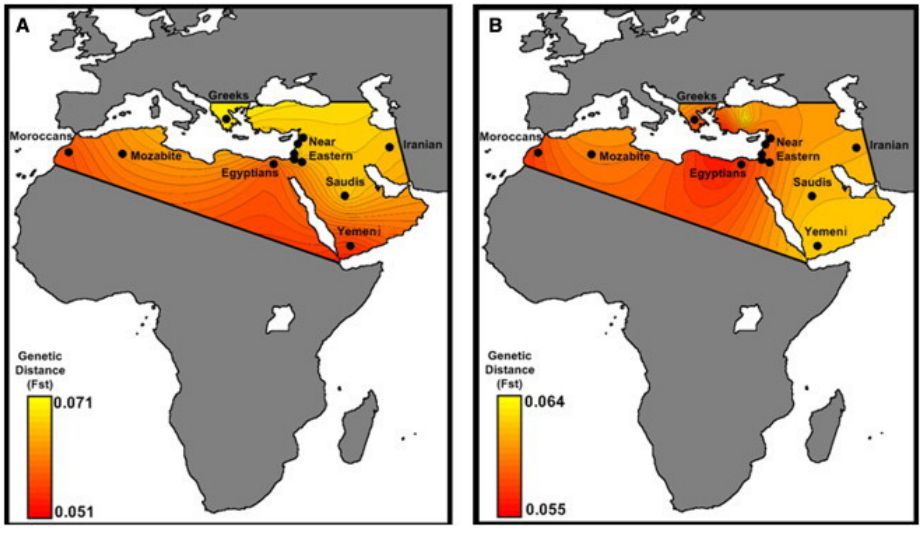

Figure 3. Pairwise F_{ST} between Semitic-Cushitic Ethiopians and Surrounding Populations, presented by Pagani et al 2012. (A) ten haploid genomes from the Semitic-Cushitic Ethiopians, showing that modern Yemeni, Egyptians, and Moroccans are closest to the Ethiopians, and (B) ten haploid non-African genomes from the same groups, showing instead a prevalence of Egyptian and Middle Eastern contributions to the non-African Ethiopian gene pool. By permission of Elsevier.

2.2 Admixture outcomes on the phenotypes

The information carried by 3 billion bases is astonishingly huge, allowing us to ask very broad scientific questions. Up to now, the thesis focused on the demographic and evolutionary consequences of the admixture event. However, along with inheriting the ancestral components from the admixing sources, we do inherit SNPs associated with specific traits (or phenotypes). So, instead of looking at the genome as a mosaic made of ancestral components, we could zoom in and focus in more detail on the biological effect of the inherited alleles. Shifting our mindset from the past to the present, we won't ask anymore how the admixture event unravelled in the distant past, but how admixture affects our traits in the present-days.

2.2.1 GWAS studies

The approach designed to detect the alleles affecting our phenotypic variation is known as Genome Wide Association Study (GWAS), where the aim is to find a link between one or more genetic variants and a trait within a population. Such a link is found when the allele frequency in specific loci is similar between individuals with a specific trait (Uffelmann et al. 2021).

There are two main trait categories: Mendelian or complex. A Mendelian trait, also indicated as a monogenic trait, is a phenotype caused by a single variant with a very large effect size. On the other side, a complex trait, that can be either oligo- or poly-genic, will be affected by some (oligo) or many (poly) alleles with moderate to small effect sizes. Differently from Mendelian traits, very rare in humans, complex traits are quite common.

For any Mendelian trait, the presence of a single allele will indicate if an individual will either display or not a given trait. Instead, to predict the variants' effect on a complex trait one must weigh the effect of all (sometimes more thousands) loci linked to said trait, where each of these loci will only partially impact the final phenotypic effect.

2.2.2 Polygenic Scores

A widely used method to predict the genetic burden for a given trait is the Polygenic Score (PS) (Visscher et al. 2017; Uffelmann et al. 2021). PS are obtained by summing the contribution of all alleles associated with a trait across the genome, which could either increase or decrease the probability of the phenotypic outcome, weighted by the allele effect size, inferred from the GWASs (Martin et al. 2017).

PS analyses need a validation set, the target set on which the scores are to be estimated, and a GWAS set, also referred to as base set, listing the genotype-phenotype associations (Choi, Mak, and O'Reilly 2020). The two sets should be

independent, as any overlap in samples would cause an over-estimation of associations due to over-fitting (Wray et al. 2013).

Transferring the GWAS associations estimates to the validation set cannot be done directly: there may be differences in the allele effect size between the two sets or differences in LD patterns between the base and validation (Choi, Mak, and O'Reilly 2020). The sets should be thus cleaned from potential biases, by shrinking all SNPs or applying an arbitrary P-value to remove a fraction of the SNPs (Choi, Mak, and O'Reilly 2020).

Beyond samples overlapping and LD patterns, there are several other complications that arise when transferring GWAS estimates. In fact, estimating PSs on populations different from the one the GWAS was based on has shown to be rather complex.

2.2.3 Underrepresentation and Transferability of GWAS studies in admixed groups

GWASs are usually carried out on large cohorts and some groups, such as Europeans (Sirugo, Williams, and Tishkoff 2019; Kim et al. 2018), are more covered than others (Landry et al. 2018), although more and more studies focusing on non-European populations recently started to emerge (Nagai et al. 2017). Besides the inability to fully capture human variability levels throughout world-wide populations, and fueling the ever-lasting bias towards the over-representation of European populations (Need and Goldstein 2009; Petrovski and Goldstein 2016; Bustamante, De La Vega, and Burchard 2011), the main implication of such a lack of variability is that the application of predictive medicine is limited only to the groups that are covered by GWAS analyses (Manrai et al. 2016). In fact, GWAS results are hardly transferable between populations due to the many population-specific forces acting on the genome(Li and Keating 2014).

Many associated SNPs are not in fact causal, but only indirectly linked with the pathway that causes the trait, likely detected thanks to LD with the causal SNPs. We could assume that the associated SNP is indeed an ideal proxy, given its link with the causal marker. However, in distantly related populations, LD patterns may be broken and the proxy SNPs selected would not be in fact in LD with the causal one, thus not actually associated with the trait under study.

GWASs tend to find association between traits and alleles with common or intermediate frequency in the population, therefore failing to recognize low frequency alleles and rare variants showing little sharing among populations (Mathieson and McVean 2012; Simon Gravel et al. 2011). Due to genetic drift, populations will have different allele frequency spectra, and associated SNPs found in one population may present with a different frequency in another. As an example, SNPs associated with height in the Greenlandic Inuit population are found to be highly associated with height in European groups as well. However, such association emerged only in the Inuit groups that carry the allele with a frequency of 0.98, and was not in the European populations, as they carry the

allele in low frequency (0.017), despite their high association with the trait (Fumagalli et al. 2015).

Along with differences in LD patterns and allele frequency spectra, assortative mating, ascertainment bias and environment all will have a population-specific impact.

Consequently, PS estimates, when applied to populations different from the one the GWAS was based on, yield low predictivity (Martin et al. 2017; 2019; Scutari, Mackay, and Balding 2016; Reisberg et al. 2017), relegating the power of predictive medicine to only the groups thoroughly and directly studied with GWAS (Figure 4).

As constantly demonstrated by demographic studies, most human groups show some level of sub-continental structure with differences in terms of LD patterns and allele frequencies that will impact GWASs (Martin et al. 2017). However, given the numerous variables acting upon SNPs-trait associations, it has been suggested that performing GWAS in more diverse cohorts may not be sufficient to reduce discovery bias (De La Vega and Bustamante 2018). To partially overcome the biases a few GWAS studies leveraged on LA, showing indeed an increased statistical power (Martin et al. 2017; Pasaniuc et al. 2011; Szulc et al. 2017). However, LA approaches have not been directly applied to retrieve PS estimations, limiting the promise of predictive medicine to only certain populations.

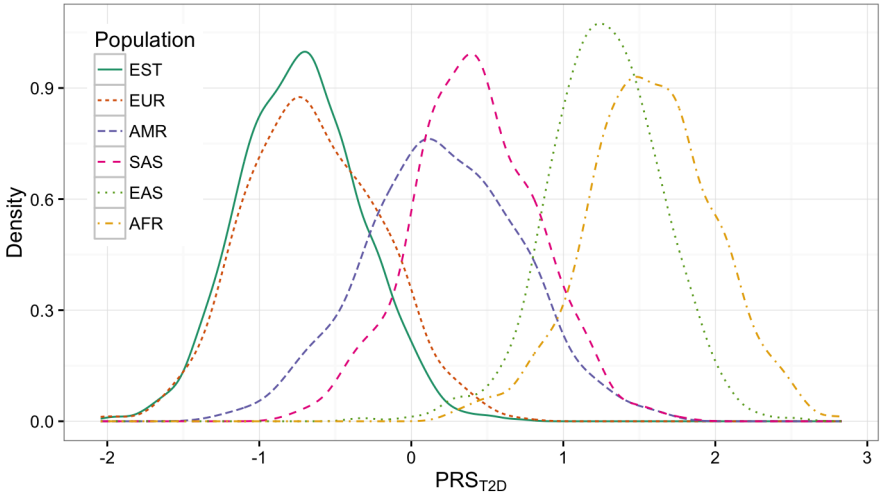

Figure 4. Polygenic Scores relative to Type II Diabetes distributions in different populations presented in Reisberg et al 2017, showing PR transferability bias.

2.3 Studying the admixture events

2.3.1 Explorative analyses

Detecting an admixture event requires first an understanding of the target population structure, along with its genetic relationship with other groups, to find the best proxy sources that contributed to the mixed population. For such preliminary and exploratory analyses, descriptive tools are the best choice, because they can be applied on a dataset without any a priori information of the demographic structure. They will reveal patterns that can be inferred as similarity or dissimilarity between samples and they can recognize admixture events.

2.3.1.1 F_{ST}: the fixation index

One of the most common summary statistics used to detect population structure compares allele frequencies between populations. Between the 1940 and 1950, S. Wright and G. Malecot developed a series of parameters to describe and measure how the genetic diversity is apportioned within and among populations. They computed three different parameters, each one measuring the amounts of heterozygosity at a given level of population structure. Here we will focus only on one: F_{ST}, a measure of both the genetic differentiation between the subpopulation as well as the differentiation within the subpopulation, providing insights into population structure and demographic histories (Holsinger and Weir 2009).

While being generally indicated as $F_{ST} = Vp/p(1-p)$, with Vp being the variance in allele frequency among populations, and p(1-p) the variance in the allelic state for an allele chosen randomly from the entire population, F_{ST} itself has been defined, estimated and applied on datasets in several ways. As mentioned in Bhatia G. et al 2013, Wright defined the F_{ST} as the correlation of randomly drawn gametes from the same population relative to the total population (Bhatia, Patterson, and Sankararaman 2013), however it is not clear what should be defined as 'total population'. A widely used tool to perform genetic analyses, PLINK (Chang et al. 2015), follows Cockerham's definition referring to the 'total population' parameter from an evolutionary point of view, thus as the most recent common ancestral population to the populations considered (Bhatia, Patterson, and Sankararaman 2013).

Given the differences in defining and estimating the parameter, F_{ST} values may differ from study to study, with the additional complication that the type of marker will affect the F_{ST} final estimate (Jakobsson, Edge, and Rosenberg 2013; Auton et al. 2009; The 1000 Genomes Project Consortium et al. 2015; Barreiro et al. 2008; Barbujani and Colonna 2010). Nevertheless, F_{ST} estimates among global human populations are known to vary broadly (The 1000 Genomes Project Consortium et al. 2015). In this document I will use Fst values to indicate fine-scale population structure, commonly defined by Fst values of 0.01 or lower (Novembre and Peter 2016). Such F_{ST} values are generally found within sub-continental groups, for example European populations.

2.3.1.2 Principal Component Analyses

When it comes to analysing the vast amount of information of the genetic data, dimensionality reduction techniques are essential tools. They comprise a set of approaches that aim to understand the relationship between multiple attributes (for example SNPs) of an entity (such as individuals) and assess their relevance in describing the set while minimising the information loss (Jolliffe and Cadima 2016). Among the dimensionality reduction techniques, Principal Component Analysis (PCA) is one of the most used in population genetics since Menozzi et al pioneered its usage to study genetic variation (Novembre and Stephens 2008; Menozzi P., Piazza A., and Cavalli-Sforza L. 1978). Similarly to F_{ST}, PCA does not operate on a priori information, allowing it to be an optimal exploratory analyses to make predictive models on the data.

The base assumption is that genotypes will cluster together in the PC space according to their similarity, so individuals with a recent shared ancestry should fall closer together than more distantly related individuals (Schraiber and Akey 2015). In fact, as demonstrated by McVean 2009 the location of samples on the PC space can be related to the mean time of coalescence between pairs of samples (McVean 2009). Through PCA we can infer past demographic events such as admixture, to an extent where softwares have been developed that discriminate between ancestries within a genome given the PC space (see PCAdmix). In fact, when samples fall along a gradient we can infer they are the results of an admixture event where the putative sources of said event fall at the ends of the gradient (McVean 2009; Patterson, Price, and Reich 2006).

On the other hand, if multiple demographic models have the same effect on mean coalescence times, it is difficult to define what kind of event characterised the population under study (McVean 2009; Novembre and Stephens 2008). For example, given that genetic similarity decays with distance (Novembre et al. 2009; Novembre and Stephens 2008), clines between groups that may look like admixture events might be in fact an effect of isolation-by-distance.

It follows that despite allowing for a wide range of inferences, interpreting past demographic events from PCA patterns is challenging and should be done with caution. PCA projections depend strongly on the sample size, as sample size differences between populations will distort the projection space, but also sampling location and ascertainment of samples may cause biases.

An additional useful feature of PCA is to define the PC space with selected samples and then project the samples of interest. Projecting samples on an already defined PC space avoids potentially skewing the analyses when the sample size of the target group is significantly different from the other set, or when the target samples are characterised by a substantial missing data, as commonly happens when ancient samples are studied. Additionally, projecting samples is useful when they are thought to be admixed, in fact by projecting the admixed individuals in the reference populations defined PC space, it is also possible to identify the admixture proportions (McVean 2009).

Although PCA is a powerful explorative tool to make initial inferences, they should be followed up with further analyses.

2.3.1.3 Clustering analyses

Alternatively to measure the genetic variance (F_{ST}) or conveniently visualise it with PCA, it is possible to summarise the genetic information without a priori information by clustering the target individuals based on their genetic patterns, highlighting population structure within the dataset.

Given a *K* number of clusters, clustering algorithms group together samples based on their similarity. As a result they assign each individual to all clusters with a probability of belonging to that cluster, defined as the membership coefficient. Such assignments occur SNP-wise to account for multiple ancestries within one genome. In this way, each individual has several membership coefficients that summarise the proportion of DNA for which they are most closely related to the other individuals in cluster *K*.

Most of the largely used clustering algorithms (such ADMIXTURE (Alexander, Novembre, and Lange 2009)) do not model correlation between adjacent loci. However, there are several drawbacks linked to the clustering approach itself that go beyond modelling for different patterns of LD.

As indicated for other explorative analyses, different demographic events may cause similar clustering patterns. Multiple non-zero membership coefficients can be due to admixture events, but also bottlenecks, drift, isolation by distance or other evolutionary events (Pritchard, Stephens, and Donnelly 2000; Novembre 2016; Lawson, van Dorp, and Falush 2018). If a sample is assigned to multiple clusters, that does not necessarily imply admixture and thus the *K* components are not indeed representing *K* ancestral populations. However, it is possible to run clustering analyses by indicating *X* populations or groups that are representative of *X* distinctive clusters, so that all other samples' ancestral proportions will be modelled based on the *X* specified groups. In this scenario, the base assumption is that the *X* distinctive clusters are indeed the ancestral groups.

Additionally, many assumptions revolve around choosing the value of *K*. *K* is usually determined a priori and strongly impacts the analyses as the direct consequence is that all individuals are assumed to share from 1 to a maximum of *K* components. While few *Ks* may not be able to properly describe the dataset variability, too many *Ks* will cause overfitting (Novembre 2016). Usually, the analyses are run selecting several numbers of *K* (ie, *K*=2–10), to then select one or two sets of the results obtained. There are several ways to select the most appropriate *K*, such as choosing the *K* with the smallest cross validation error, making use of softwares that parametrize *K* or evaluating the results based on historical knowledge (G. Hellenthal 2019). However, none of these methods can predict the true *K,* given that *K* is usually unlikely to be a biologically meaningful quantity.

2.3.2 Allele frequency analyses

By leveraging allele frequencies differences it is possible to test specific demographic hypotheses, rather than to infer past events from patterns of similarity highlighted by explorative analyses. In my studies I focused on the F-statistics suite to perform most of the allele frequency analyses and to test specific demographic hypotheses.

2.3.2.1 Inferring population structure with allele frequency analyses

Widely used allele frequency methods that analyse population structure are the F-statistics, distinct but related to Wright's F-statistics, introduced by Reich et al 2009 (Reich et al. 2009) and extensively developed by Nick Patterson and collaborators (Patterson et al. 2012; Moorjani et al. 2011), as mentioned in Peter 2016 (Peter 2016).

With F-statistics one explicitly tests a demographic model rather than inferring relationships from the data. The model can be interpreted as a phylogenetic tree and the length of the tree branches relating the groups are a measure of the allele frequency correlations (Peter 2016). Such correlations are interpreted as shared drift and all subsequent inferences are based on the assumption that shared drift implies shared evolutionary history (Peter 2016).

Considering a tree (X;Y,Z), the F-statistics will estimate the allele frequency correlations across X, Y and Z, averaged over many biallelic SNPs. The correlations are expected to change as $E(y|x) = x$ and $E(z|x) = x$, so that the expected allele frequency correlations across the groups can be indicated as: $E(x–y)(x–z) = 0$. The given phylogenetic tree analysed through the F-statistics will serve as the null hypothesis to test $E(x–y)(x–z) = 0$. The hypothesis will be rejected if the drift value is significantly $\neq 0$, measured with Z-Scores. In case of a rejection, the topology tested simply does not reflect the shared drift calculated between the groups, and other evolutionary events apart from drift might have acted upon the tested groups.

F-statistics can be applied on two (*f2-statistics*), three (*f3-statistics)* or four (*f4-statistics*) populations (Figure 5).

F2 is a measure of similarity between pairs of entities, calculated based on shared genetic drift, lower values of F2 are thus expected the more distantly related two populations are, differently from F_{ST}.

Given three populations, F3 can be used either as a formal test for detecting admixture (F3-Admixture) or as a measure of shared genetic drift between two populations given a third one used as an outgroup (F3-Outgroup). F3-Outgroup measures the genetic similarity of population A and B conditioned on the outgroup O. The outgroup allows the statistics to be calculated only on the polymorphic sites of A and B not shared with O. With F3-Outgroup the expected

value of the tree is always >= 0. Instead, F3-Admixture tests whether target population C is admixed with A and B both and in this case, negative values are expected if the admixture did occur. The method is robust, thus a negative value ensures that an admixture event occurred on C. However signals of admixture may be hidden by genetic drift or founder events and in these cases F3-Admixture may result in positive value even if C is admixed.

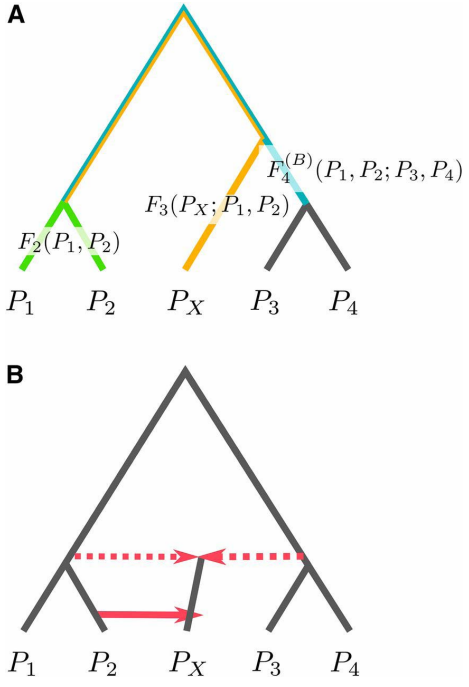

Figure 5. A population phylogeny with (A) branches corresponding to F2 (green), F3 (yellow), and $F4^{(B)}$ (blue), (B) an admixture graph extends a population phylogeny by allowing gene flow (red, solid line) and admixture events (red, dotted line). Peter B., "Admixture, Population Structure and F-Statistics", 2016, Genetics,Volume 202, Issue 4, by permission of Oxford University Press.

F4 is a formal test for a phylogeny (also referred to as treeness), and its rationale is similar to the D statistics (Durand et al. 2011; Green et al. 2010). The D statistic considers four populations: two references (A and B), one target (C) and one outgroup (O), in a phylogeny (A,B,C,O), and the goal is to compare the similarity between C and the references A and B. With D = 0 the built phylogeny do not show excess of shared drift; with D < 0 the statistics indicate a stronger gene flow between C and B; while D > 0 the gene flow is stronger between C and A. The F4-statistics behaves similarly, with positive or negative signals indicating which reference group is closest to the target one.

2.3.2.2 Estimating the admixture proportions using allele frequency analyses

Previously listed F-statistics, while they can detect admixture events, they cannot provide additional information about the gene flow, such as the magnitude. However, by making use of two sets of F4, it is possible to estimate the contribution of one source of the admixed population though the F4-ratio (Reich et al. 2009). For such analysis there is no need for accurate surrogates of the admixture event, but the phylogeny of populations considered must be known.

qpWave and qpAdm allow for a deeper analysis of the admixture event. qpWave detects the minimum number of independent ancestries needed to model the target population, while qpAdm models and estimates the admixture proportions by considering several independent ancestries. Both of these tools require that at least a partial phylogeny is known, as they need a defined set of *M* outgroups (defined 'right' populations, R), a set of *N* putative admixing sources (defined 'left' populations, L) along with the target admixed group, T. qpWave tests whether L and R clades are independent, thus if that gene flow did not occur between the populations of the two sets. Through multiple F4 statistics, qpWave estimates the minimum number of gene flow events occurring between L and R. If all F4s values calculated are 0 (https://uqrmaie1.github.io/admixtools/index.html), the outgroup set R and the reference set L are independent. We can then model the target T as being admixed by L set populations, conditioned on the R set with qpAdm.

Finally, with qpGraph one can test a given phylogeny by building a tree-like graph with N leaves that correspond to real populations and M nodes indicating pseudo-populations. By testing all possible F-statistics between the N leaves, the tool returns all the respective Zscores that may reject ($|Zscores| > 3$) or accept ($|Zscores| < 3$) the proposed topology (Patterson et al. 2012). The tool models also admixture events, for which the best-fitting admixture proportions are calculated, and indicates the amount of drift occurring in each branch. An optimal topology, among the many possible ones, is found when Zscores values for all F-statistics are not significantly different from zero. In such cases the given tree-like graph represents a fitted model for the groups studied, although it cannot be considered as the true topology.

2.3.3 Haplotype data

Many of the analyses listed before consider SNPs to be independent from one another. Such assumption, even if simplistic, allows for computational speed and inferences on past demographic events, even without a priori information. However, modelling the correlation patterns of SNPs along a chromosome, although computationally expensive, allows for fine-detailed inferences (Leslie et al. 2015).

By taking into account combinations of alleles, rather than single SNPs, we are shifting from allele-based to haplotype-based analyses. An haplotype is a

combination of alleles along a set of loci that have been inherited entirely from the maternal or paternal source (McVean and Kelleher 2019). The length of the set of loci considered may vary depending on the study, it may be an entire chromosome, a smaller segment or a block of ancestry. By considering an haplotype as a block of ancestry entirely inherited from one source, we are accounting for correlation between markers and therefore considering Linkage Disequilibrium patterns.

2.3.3.1 Retrieving haplotype data

The haplotype phase can be estimated through either experimental or computational methods. In both cases, the goal is to infer from which parent each allele is inherited. Experimental methods have a very high accuracy, however the cost of generating the sequence data and the requirement of technical expertise limit their application (Browning and Browning 2011). Consequently, several tools have been developed to statistically infer the parental configuration of alleles at heterozygous sites. Especially in cases where related individuals are unavailable, it is necessary to rely on algorithms that estimate the haplotype configurations using a large cohort of samples as reference. Widely used tools that model haplotype phase are based on Hidden Markov Models (see paragraph 2.3.3.1.1) (Howie, Donnelly, and Marchini 2009; Delaneau et al. 2014; Stephens and Scheet 2005; Loh, Palamara, and Price 2016).

Once the haplotype phase is inferred, the resulting genetic information can be used for association studies and inferring demographic events (Browning and Browning 2011).

2.3.3.1.1 Hidden Markov Models

A Hidden Markov Model (HMM) represents a system assuming that: i) it follows a Markov chain where the states of interest are unobservable (or 'hidden states'), the Markov chain itself assumes a sequence of states where the probability of the next state depends only on the present state and not the entire chain of events; ii) there is an observable process whose behaviour is related with the hidden state, thus from the observable states one can learn about the hidden states (Yoon 2009).

To build an HMM, three distribution probabilities are needed: the Initial state distribution $\boldsymbol{P}(z_0)$; the Transition probabilities $\boldsymbol{P}(z_l|z_{l-1})$, indicating the probability of jumping from one hidden state to another; the Emission probabilities $\boldsymbol{P}(x_l|z_l)$, indicating the probability that each hidden state emits the observed state (Wegmann and Leuenberger 2019). The joint distribution of the HMM is equivalent to the product of initial state distribution, transition probabilities and emission probabilities:

$$\boldsymbol{P}(z_{0:L},x_{0:L}) = \boldsymbol{P}(z_0)\prod_{l=1}^{L}\boldsymbol{P}(z_l|z_{l-1})\prod_{l=1}^{L}\boldsymbol{P}(x_l|z_l)$$

A popular HMM that handles genetic data was introduced by Li & Stephens in 2003 (Li and Stephens 2003), and has been widely used to infer the haplotype phase as well as the ancestral state in Local Ancestry analyses. The model allows reconstructing the target haplotypes as an imperfect mosaic of a set of the reference haplotypes. The transition probabilities model how the state of the mosaic tiles change along the chromosome, following the recombination rates (rho), making the model realistic. The emission probabilities allow each target haplotype to differ at some level from the reference haplotypes by accounting for mutation probabilities (θ, also defined as the miscopying parameter) making the model robust to several biases (mutations, genotyping errors).

An issue of such models is that the number of the generated states is generally quite high, for example, when taking into account haplotypes, the number of states will be N^2, where N is the number of haplotypes. Although ideally a straightforward solution would be to calculate the maximum likelihood of the combination of all states, it is not computationally feasible.

There are ways to avoid calculating the maximum likelihood that can still retrieve valid information from the generated states (L. R. Rabiner 1989): with the Forward-Backward algorithm, we can calculate the posterior probabilities $\boldsymbol{P}(z_l \mid x_{0:L})$ of the hidden state per each locus l given the full observed data, $x_{0:L}$ and with the Viterbi algorithm we can find the most likely path through the hidden states. Differently, the Baum-Welch algorithm, a case of the Expectation-Maximisation (EM) algorithm, can be instead applied to infer the model parameters (the transition and emission probabilities).

2.3.3.2 ChromoPainter: a strategy to infer population structure from haplotype data

A widely used haplotype-based approach that models LD is ChromoPainter (CP) (Lawson et al. 2012). CP finds patterns of similarity between haplotypes, laying the foundations for demographic inferences, ancestry proportions, population structure and sample clustering. A strong feature of ChromoPainter is that it does not need a narrow set of pre-specified reference samples to characterise the target haplotype, in contrast to what many other allele-based or haplotype-based approaches do.

In order to find similarity patterns, ChromoPainter considers each target individual as a ‘recipient’ and the rest of the samples as ‘donors’. The recipient individual genome is reconstructed (or painted) using the DNA chunks of the donors, so that the recipient can be seen as a mosaic of several chucks, each of these donated by the suited donor. The most suited donor chunk is found by selecting the closest genetic relative available in the donor set for each haplotype. By

painting the entire chromosome as a series of donated chunks, the painted recipient individual genome will in turn look as a mosaic where each tile is represented by the closest relative of the individual at that locus. The recombination events that break the mosaic tiles at the boundaries of the donated haplotype can be considered as the ancestry switches.

To find the closest relative haplotype that best describes the recipient haplotype, ChromoPainter employs a modified version of the Hidden Markov Model proposed by Li and Stephens (described in 2.3.3.1.1)(Li and Stephens 2003). The main difference with Li & Stephens model is that ChromoPainter algorithm considers all donor haplotypes to reconstruct a recipient, instead of ordering the haplotypes based on 'Product of Approximate Conditionals' likelihood, which select only a handful of potential donor haplotypes in the Li & Stephens model.

2.3.3.3 Estimating admixture proportions using haplotype data with ChromoPainter

From the painting process we can obtain the 'copying vectors' that inform from which donor a given recipient copies per each locus. Through linear regression it is possible to make use of ChromoPainter copying vectors to model populations as mixtures of others, and thus calculating the ancestry proportions of each donor group copied by the target group. Such modelling is done by calculating the least squares, by finding the line (combination of donors copying vectors) that minimises the sum of squares residuals calculated from the data points (target copying vector) to said line. From such a combination we can infer the ancestry proportions.

Since the haplotypes are a physical entity, they cannot be defined by a negative number (Chen and Plemmons 2009). In this scenario instead of using least squares, it is more suitable to use a Non-Negative Least Square (NNLS) where the values are constrained to be positive. The NNLS function often employed in CP analyses, described in Hellenthal et al, Leslie et al (Hellenthal et al., 2014; Leslie et al., 2015), is a modification of the Lawson-Hanson NNLS implementation of non-negative least squares function (Lawson and Hanson, 1995).

2.3.4 Dating the admixture event

Along with the ancestry proportions, another feature that can be inferred from an admixed population is assessing when the admixture event occurred. Dating the admixture event is key to contextualising the genetic inferences in relation to historical events. Dating analyses rely on the fact that recombination breaks down the ancestry segments over generations: thus, from the tract length distribution or the number of ancestry switches (Johnson et al. 2011) it is possible to infer the number of generations that occurred since the admixture event, the Admixture Time (AT).

However, the amount of ancestry switches may not always be directly obtained, and thus widely used dating methods rather analyse the exponential decay of admixture LD as a function of genetic distance, which requires fewer parameters and assumptions (ROLLOFF (Moorjani et al. 2011; 2013), ALDER (Loh et al. 2013) and MALDER (Pickrell et al. 2014)). Given a target admixed group and its reference sources, the LD estimates will show an exponential decay over genetic distance, forming a curve whose decay rate indicates the AT (sic, (G. Hellenthal 2019)). Additional exponential functions allow for testing and dating multiple admixture events (Pickrell et al. 2014). The methods can be also used beyond the scope of dating and further characterise the admixture event. For example, by comparing the amplitude of the curves obtained with different reference populations, one can detect the most suited reference groups, which will show the highest amplitude value (Pickrell et al. 2014).

Additionally, it is possible to infer the AT from ChromoPainter analises, as done by the software GLOBETROTTER (Hellenthal et al. 2014) and fast GLOBETROTTER (Wangkumhang, Greenfield, and Hellenthal 2021), using haplotype information rather than allele-frequency to discriminate between the references.

Generally, in case of samples scarcity and ancient admixture events the mentioned methods cannot deliver accurate results, and standard errors are generally quite large when dating events that occurred more than 100 generations ago. However, a recently developed method, DATES, has been shown to reach satisfactory accuracy levels when inferring few, unphased, low-coverage ancient samples (Chintalapati, Patterson, and Moorjani 2022; Narasimhan et al. 2019).

2.3.5 Local Ancestry inferences

Local ancestry inferences (LA) allow us to look in more detail at the effects of the admixture within the genome. Differently from Global ancestry approaches, that allow to estimate the ancestral proportions along the genome, with LA we detect the ancestral contributor at a haplotype or SNP level allowing for inferences on population evolutionary history with a finer level (Yelmen et al. 2019). Additionally, ancestral patterns along the genome impact phenotypic variation and are taken into account also in SNP-trait association studies (Martin et al. 2017), highlighting the relevance of LAI tools not only in evolutionary studies but in clinical ones as well (Pasaniuc et al. 2011).

2.3.5.1 Admixture deciphering key points

LA approaches assign each allele to its respective ancestry by comparing the target and the sources' genetic features. To make such an assignment LA tools are based on several assumptions, therefore several variables should be taken into account to achieve accurate results.

2.3.5.1.1 Input data: the importance of phased data

Since LA approaches assign each allele to the respective ancestry, an unphased dataset would compromise the assignment. While for some human groups phasing might not be an issue, given that the phase might be obtained experimentally or statistically with a large amount of data, such conditions may not be met for poorly sampled human groups or different taxa, for which it could be eventually impossible to perform local ancestry accurately. To overcome such issues some LA tools implemented a phasing step within their inferences (Guan 2014).

2.3.5.1.2 Length of ancestral blocks

Recombination will impact the admixed genome over generations, shortening the inherited segments. Therefore, while recent admixtures, characterised by few recombination events, will show long ancestrals tracts, past admixture events will show short ancestry blocks. Generally, longer tract lengths are easier to detect and to assign correctly to the right ancestry, while small tract lengths will negatively impact LAI accuracy.

It is possible to infer the expected tile length (L) of the ancestry block in a population, given the mixing proportion (m), recombination rate (r) and time since the admixture event (t, in generations), as follows: L=[1−m]r[t−1])^–1 (Racimo et al. 2015).

2.3.5.1.3 The sources of the admixture

LAI tools usually compare the genetic features of the admixed sample to the sources' features to properly assign the admixed components to their respective ancestries. The choice of reference populations is crucial in ancestry deconvolution analyses, as unfit proxy samples will cause misassignation or low rate assignment. Generally the sources' samples are unavailable, or, at best, had gone through extensive genetic drift since the admixture event, so most of the time one relies on proxy samples. Although it is theoretically possible to use ancient DNA as reference, high levels of missingness in the data, low sample size and unphased genotypes will substantially lower the inference accuracy levels.

It should be taken into account, however, that the chosen reference samples most likely do not correspond to the actual admixing source population, and we must be cautious when we link the ancestral component of the admixed group with either modern or ancient labelled populations. Overall, when we link any ancestral component to its source we are referring to the genetic similarity between the component and the putative sources, rather than direct contact (Mathieson and Scally 2020).

2.3.5.1.4 Genetic similarity between sources

LA base their inferences in finding similarity between the target admixed group and the available sources. Similarity between the sources will impact the analyses as the discriminating power of the tools will be lower. In such cases the inherited ancestral tracts will share a large amount of allele frequency or haplotype structure within the target group, which ultimately translates in a difficulty in discerning between the ancestral sources, even with ideal reference samples. Source similarity is an issue encountered mosty when inferring sub-continental admixture events.

2.3.5.2 Local Ancestry inference Methods

LA methods available can be clustered based on several elements: whether they account for LD, whether they can model multi-ways admixtures or based on the type of parameters they require (Geza et al. 2019). Here, I will instead present the three state-of-the-art methodologies that I used throughout my studies given the algorithms they are based on.

2.3.5.2.1 Hidden Markov Model-based approaches

Several LAI tools base their inferences on Hidden Markov Model (HMM) algorithmes (Schraiber and Akey 2015; Hoggart et al. 2004; Tang et al. 2006; Sundquist et al. 2008; Sriram Sankararaman et al. 2008; Price et al. 2009; Baran et al. 2012; Omberg et al. 2012; Guan 2014; Salter-Townshend and Myers 2019). As indicated by Wegmann and Leuenberge (Wegmann and Leuenberger 2019), said algorithm fits perfectly the need of modelling ancestry along the genome, for two reasons: i) Markov Chains (on which HMM are based) do not need to be aware of all states, only the one state previous to the one studied. Along the genome, ancestry segments generally span many loci, so that the knowledge of the ancestry of one locus is sufficient to infer the ancestry of the next if no recombination occurred. ii) The exact sources of the admixture are usually unavailable, so an ancestry cannot be directly observed. Therefore, since the ancestries along the genome are hidden, we can use them as the hidden states.

Among the methods that employ HMM, I will further describe only ELAI (Guan 2014), as its algorithm allows testing for a wide range of different scenarios.

ELAI is based on an extension of HMM, where a two-layer HMM models both sets of LD: one upper-layer accounts for admixture LD while the second lower-layer accounts for background LD.

Within each layer, a K number of clusters are created and labelled to represent the ancestries' alleles, so that, for example, multiple clusters with the same label over adjacent markers indicate an ancestral haplotype. The ancestry switches, thus the loci where cluster labels switch, are recognized as recombination events. This

implies that ELAI does not use pre-specified windows, it can be used on either small or long tract lengths, thus on recent or (relatively) ancient admixture events. Local ancestry is then inferred by condensing the local haplotypes inferred from the lower-layer, and then assigning them probabilistically into the ancestral groups following the upper-cluster labels.

Like many other local ancestry tools, ELAI needs a set of reference populations that are closely related to the real sources of the admixture to efficiently assign the markers to the respective ancestral component. Another costrain is given by the need for specifying the number of generations that elapsed since the admixture event, which may be inferred by other dating software. However, many other parameters that are limitations for other softwares are inferred by ELAI directly from the data: there is no strict need for phasing, as the software will attempt to phase the given unphased data, and recombination rates are inferred from the dataset. Additionally, ELAI can be run without surrogates for one of the admixing groups (Zhou, Zhao, and Guan 2016).

2.3.5.2.2 Principal Component Analysis-based approaches

PCAdmix (Brisbin et al. 2012) is a Principal Component Analyses-based algorithm to perform LAI. PC analyses themselves fit well the needs of Local Ancestry, as they are fast and can separate samples on a continuous space based on the population structure.

The first step of the analysis relies on inferring the PCs via Singular Value Decomposition (SVD) based on the reference samples, to then project the target individual windows upon the PC space created. The method proceeds by analysing short windows of SNPs and assessing the probability that a given window of the target admixed individual comes from each reference population.

To determine how informative each SNP is in classifying the ancestry of a genetic region, the PC loadings (the eigenvectors) for each SNPs are collected from K – 1 PCs, where K is the number of ancestral populations (Brisbin et al. 2012).

To model the posterior probability of the ancestry in each window (Brisbin et al. 2012), PCadmix uses a Forward-Backward algorithm. In this framework the transition probabilities are $q_{i,j}\pi$, the recombination rates (π) and the average ancestry proportion (q) of population *j* in target haplotype *i*, which is estimated by calculating the Euclidean distance in the PC space between *i* and all non-*j* populations. The emission probabilities are defined by a multivariate normal distribution that takes into account, among other parameters, the window loading scores.

This methodology requires pruning the dataset beforehand, removing loci in strong LD, therefore losing some level of genetic information, and a pre-specified set of reference populations. PCAdmix has been tested on a wide range of populations and can handle two-ways and three-ways admixture events.

2.3.5.2.3 Random Forest-based approaches

With generative approaches, such as Hidden Markov Models, we assume that the observable state is linked to the hidden one. However, even if we account for recombination rates, miscopying rates and eventually perfectioning the parameters with an EM-algorithm, real-world observations may be also characterised by other features and dependencies that we cannot control (Wallach 2004). Alternatively to the generative approaches, there are the discriminative approaches, which focus on separating one class from another, rather than generating new data from the observed states (Lasserre and Bishop 2007).

A widely used discriminative approach to perform LA is RFmix (Maples et al. 2013). RFmix applies a Conditional Random Field (CRF), a probabilistic framework for segmenting and labelling sequence data, on windows of the genome of predefined length (Wallach 2004). The CRF is parameterized by a Random Forest (RF) algorithm, which is itself based on Decision Trees (DTs). DTs are a decision support tool based on a tree-like graph where each internal node splits into two events and the leaves are the possible final outcomes of the events.

The RF creates a random set of DTs (a set of decision trees as big as a forest), the observed trait is then passed through all DTs and then the most voted output is chosen.

In RFMix, there are two learning steps: initially, the chromosomes are divided into windows, and in each of them a RF, trained on the reference panel, is used to estimate the posterior probability of the ancestral state; then, the ancestry assignment obtained from the previous state is used to improve the final inference accuracy with an EM step (Maples et al. 2013).

Similar to the majority of the other methods, RFmix relies on a panel of proxies for the source populations, however, the algorithm should be capable of learning to discriminate from the admixed samples themselves, overcoming the issue of limited source availability. Ideally, such a framework makes Rfmix perfect for a dataset with sample scarcity. However, this LAI tool has been mainly tested on recently admixed populations with large block length and availability of samples, such as the American populations.

2.3.5.3 Downstream exploitation of Local Ancestry inferences

Generally, LAI tools will return the posterior probabilities of the ancestral states per each site or per each window. The user will then apply a threshold to remove all windows/sites where the posterior probability of an ancestry does not reach the desired level. Threshold levels may be chosen based on the experimental design, aim of the analysis or amount of windows needed to perform successive analyses (a stringent threshold generally implies removing a larger amount of sites).

Furthermore, it is possible to perform subsequent ancestry-specific analyses by masking out either ancestries in order to retrieve only the SNPs assigned to one ancestral component (Yelmen et al. 2019; 2021).

2.3.5.4 Testing against known scenarios

Since all LAI approaches rely on several assumptions it is possible that some of these assumptions are not perfectly met and errors are accumulated throughout all LAI steps. As a control for the LAI analysis, it is useful to create a simulated set of samples for which all genetic parameters are known and perform LAI on this set as well.

Testing against known scenarios provides insights on the sensitivity and specificity levels of the method developed by comparing the experimental results, the predicted values, with the simulated ones, which are the known truth. By applying several thresholds to build different confusion matrices, we can build a ROC curve, where the true positive rate is defined by the rate of ancestries assigned correctly, and the false positive rate is given by the misassignation.

Many simulation tools are based on the program published by Hudson in 2002, named ms (Hudson 2002) and its legacy. Ms generates a random genealogical history using a coalescent approach. A random set of samples is then drawn and used to investigate the properties of a population evolving under a neutral model (mutation cannot occur twice on the same site (infinite-sites model), no selection is acting upon the samples, generations do not overlap and population size is finite). The neutral model allows for mutations, recombination, gene conversion, symmetric migration among subpopulations, and simple population size changes. Further developments of ms program also allow for crossovers and gene conversion hotspots (Hellenthal and Stephens 2007), and are able to deal with large sample sizes more efficiently then the base version (Kelleher, Etheridge, and McVean 2016).

Similarly, programs that simulate admixture between multiple populations have been developed. Several admixing parameters can be indicated, such as: number of admixed individuals, admixing proportions, sources, generations elapsed since the admixture. The result is an arbitrary number of admixed individuals for which the ancestral contributor at each allele is known.

3. AIM OF THE STUDIES

The three studies I will present aim to discuss the potentialities and limits of the Ancestry Deconvolution (AD) approaches. The Neolithic and post-Neolithic migrations characterised a large area, favouring the encounter of divergent populations over time. Such expansions can be traced to sub-Saharan Africa, highlighting an event of cross-continental admixture. Even more so, the entire European region has been the stage of massive migrations, to such an extent where each European group can be modelled as the result of a series of admixture events.

The AD approaches have been applied on three different topics: studying an admixture event where worldwide, highly divergent populations came together; studying the limitations in deconvoluting fine-scale admixture events, such as in European groups; and finally applying the built knowledge to overcome the limitations in applying Polygenic Score and GWAS estimates on admixed populations.

3.1 Aims of the first study (Ref I)

East Africans are genetically characterised by a combination of a Non-African layer, originating from a wave of migrations from West Eurasia during the Bronze Age, along with an autochthonous African layer. Despite numerous studies describing the non-African layer, a consensus on the origin of such a component has not been reached yet. In fact, Pickrell et al found the layer to be genetically closer to a Sardinian-like ancestry, while Lazaridis et al 2016 found genetic similarity with farmers from the Neolithic Levant (Pickrell et al. 2014; Lazaridis et al. 2016). In both studies, the Ethiopians' ancestral layers have been analysed with global ancestry methods, therefore considering both components together. However, whole-genome inferences may be clouded by much discordant information that the layers carry altogether. Starting from LAI approaches, I aim to study the demographic history of the Ethiopian genetically non-African layer by leveraging on ancestry specific analyses.

3.2 Aims of the second study (Ref II)

LA tools have been applied on admixed populations with divergent ancestries, where the admixing contributors came from different continents (ie, Latin American and African American groups) and for which large amounts of samples per each reference group should be available. This is not the case for any European population, where the post Neolithic migrations waves and subsequent migrations within the continent contributed to shape the genetic makeup of all present-day European populations, causing the admixing sources to be genetically too similar to perform LA accurately.

Secondly, LA inferences require a large amount of samples per each reference population, to best capture the genetic variability of the admixed individuals. However, while such a requirement may be frivolous for some, well-typed and largely available human groups, for other key human groups or different taxa it might not be the case.

I compared different LA tools accuracy levels under different scenarios where the admixing sources show different degrees of similarity and availability of samples, to better understand current state-of-the-art LAI tools limitations. Additionally, I proposed WINC (Window-based ChromoPainter/NNLS) a novel LA tool that leverages on haplotype-painting technique. Such technique has been shown to accurately describe sub-continental population structure and to not be affected by low samples size as the painting step is done at individual levels (Drineas, Lewis, and Paschou 2010; Leslie et al. 2015; Gilbert Edmund et al. 2019; Pankratov et al. 2020; Saint Pierre et al. 2020; Martin et al. 2018; Bycroft et al. 2019).

3.3 Aims of the third study (Ref III)

Polygenic Scores (PSs) summarise the effect of many genetic variants shown to be associated with a phenotype or a disease (Lambert, Abraham, and Inouye 2019; Dudbridge 2013). However, PSs rely on population-dependent contributions of many associated alleles, with limited applicability to understudied populations and recently admixed individuals. We proposed to leverage on LA inferences to extract ancestry-specific SNPs and estimate PS on the partial deconvoluted segments.

4. MATERIALS AND METHODS

4.1 First study (Ref I)

The study focused on 120 whole-genome sequences from five Ethiopian populations: Amhara, Gumuz, Oromo, Ethiopian Somali, Wolayta (Pagani et al. 2015).

Along with the Ethiopian samples, I combined various datasets where both ancient and modern genomes were available, so that the analyses could focus on both geographical and temporal information. Globally, the dataset included worldwide populations data from the 1000 Genomes project, HGDP-CEPH project along with published ancient samples and 100 whole-genome sequences from Egypt (http://reich.hms.harvard.edu/datasets; The 1000 Genomes Project Consortium et al. 2015; Behar et al. 2010; Pagani et al. 2015). After merging, the dataset was downsampled to maximise the number of individuals typed at each SNP, retrieving 1,037,084 markers and 6,382 individuals.

PCA analyses were performed on the dataset comprehensive of all the available populations through EIGENSOFT (Patterson, Price, and Reich 2006) smartpca using lsqproject: YES, defining the PC space with modern individuals and then projecting the target samples along with the aDNA. ADMIXTURE (Alexander, Novembre, and Lange 2009) analyses were performed from k=2 to k=15 after excluding individuals with more than 15% missing data with PLINK (Chang et al. 2015). LA inferences were carried out with PCAdmix (Brisbin et al. 2012), analysing windows of 20 SNPs, and ELAI (Guan 2014) setting 100 admixture generations. All allele-frequency analyses were run using F-Statistics, with POPSTATS (Skoglund et al. 2015) and AdmixTools (Patterson et al. 2012). Dating inferences were carried out with MALDER, with mindis parameter set as 0.005 (Pickrell et al. 2014).

4.2 Second study (Ref II)

I simulated with mspms (Kelleher et al., 2016) a Test Set, composed of 13 populations with the addition of seven sister groups labelled them as “Ghost” (GST). The GST demes were used to create eight two-ways admixed populations and one three-ways admixted group.

I then created the Empirical Set, characterised by all available groups from the 1000 Genome Project and three two-ways and one three-ways simulated admixed groups (The 1000 Genomes Project Consortium et al. 2015). I simulated admixture events between: FIN-TSI, CHB-TSI, YRI-TSI and YRI-CHB-TSI.

All simulated admixed groups were obtained with Admix-Simu software from Williams Lab (https://github.com/williamslab/admix-simu) and were characterised by an admixture event dated 100 generations ago with the sources contribution of 70%–30% for the two-ways admixture and 40%–30%–30% for the three-ways admixture. I simulated 50 individuals per each admixed group.

To test our approach on a real case scenario, I used 61 ASW individuals and 64 MXL (American of African Ancestry in South West and Mexican Ancestry from Los Angeles USA) from the 1000 Genome Project (The 1000 Genomes Project Consortium et al. 2015).

I performed LA analyses on the simulated data with ELAI (Guan 2014), RFMix (Maples et al. 2013) and PCAdmix (Brisbin et al. 2012), adjusting the parameters to account for the number of reference samples (45 and 2 individuals per source) and time since the admixture (100 generations).

To test our framework, I first estimated ChromoPainter nuisance parameters mu (mutation rate) and Ne (effective population size), through the Expectation-Maximisation algorithm. I obtained the mu = 0.0011, and Ne = 2516.3133 for the Test set, and mu = 0.0008281 and Ne = 939.2658 for the Empirical set. I included in the donor panel all available groups within the tested set, but downsampled the proxy sources first to 45 and then to 2 samples. I then splitted ChromoPainter copying vectors in genomic windows of equal lengths. Finally, to perform the ancestry assignment I used the Non-Negative Least Squares (NNLS) approach presented in Hellenthal et al., 2014 on the previously splitted copying vectors.

4.3 Third study (Ref III)

We created five simulated admixed populations using 1000 Genomes Project groups combining deeply divergent populations, so that they traced their ancestry to East Asian (ASN), Africa (AFR) and Europe (EUR), with different proportions and admixture generations. All simulations were carried out using Admix-Simu software from Williams Lab (https://github.com/williamslab/admix-simu). I performed LA analyses with ELAI using CEU to retrieve the EUR ancestry, YRI to retrieve the AFR ancestry and CHB to retrieve the ASN ancestry. I then performed LAI on a subset of phased admixed individuals within UK-BioBank (Bycroft et al. 2018), 220 Ethiopian whole-genome sequences (Pagani et al. 2015) and ASW (The 1000 Genomes Project Consortium et al. 2015) using West Eurasian, African and Asian groups as reference sources.

All LA inferences were carried out using ELAI. The admixture generation parameter was adjusted based on the population under study. We assigned each SNP to the respective ancestry selecting a threshold of 0.9. To evaluate PS predictivity on the deconvoluted segments further analyses were carried out on the Estonian BioBank (Leitsalu et al. 2015) and UK BioBank making use of also SNP-trait association studies of Japan BioBank (Akiyama et al. 2017; 2019) .

5. RESULTS AND DISCUSSION

5.1 Local Ancestry inferences applied on a demographic study: the Ethiopian case

5.1.1 Performing Local Ancestry and masking genomes

Local Ancestry inferences were carried out with PCAdmix on 120 samples from five Ethiopian ethinic groups: Amhara, Ethiopian Somali, Oromo, Wolayta and Gumuz. To retrieve the West Eurasian ancestry I selected as source CEU, Utah residents with ancestry from northern and western Europe; while I used Gumuz, an Ethiopian group previously shown to have close to 0% of Eurasian component, as proxy for the African component. I selected a threshold of 0.9 probability for a window to be assigned to either one layer of ancestry or the other. If a window did not reach the threshold for any component, I then labelled it as unassigned. I masked the Ethiopian genomes by retrieving only the windows that passed the threshold (Figure 6). I thus obtained 240 partial genomic segments characterised by SNPs that were assigned with a 0.9 probability to either the African (AF) or the non-African (NAF) ancestry.

Given the low proportion of NAF ancestry in Gumuz, I did not proceed carrying on allele-frequency analyses on this ethnic group.

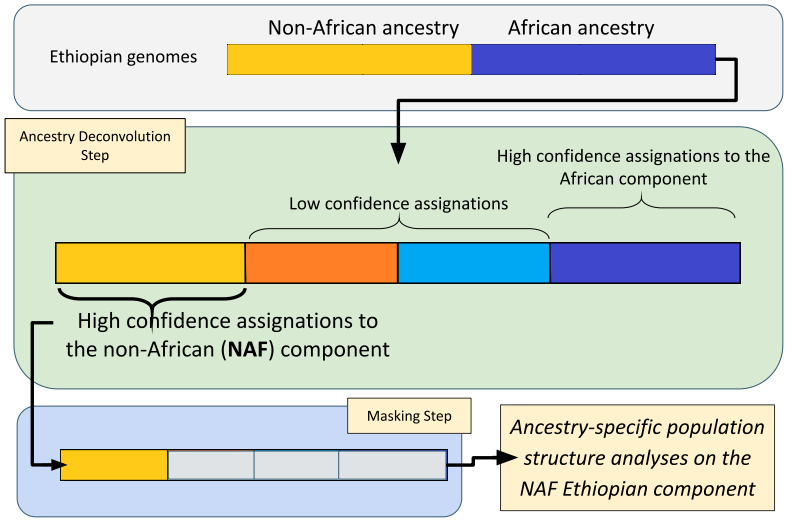

Figure 6. Schematic workflow to perform ancestry-specific analyses. Workflow to perform Ancestry Deconvolution and masking process to obtain the ancestral components within the Ethiopian genomes.

5.1.2 Explorative analyses with masked genomes

PCA shows Ethiopians whole-genomes falling outside the European cline, as expected given the African layer (Figure 7). Masked NAF genomes cluster with other non-African populations from the Mediterranean area, notably: North African Jews and ancient samples from the anatolian area such as 'Minoans_Lasithi' and 'Minoan_Odigitria', samples dated 2210–1680 BCE and linked to a Bronze age civilization from Create, and 'Anatolia_N', samples dated 5500–5000 BCE and linked with the neolithic farming period. The ancient Levantine samples from the Mesolitic (Natufians) to the Eneolithic (Neolithic and Chalcolithic samples) era fall between the Anatolian cluster and the Ethiopian whole-genome sequences (Bar-Yosef and Valla 1990; Pearce 2019). The other main West Eurasian ancestral sources, WHG and CHG, fall at opposite sites outside the modern European cline, as expected.

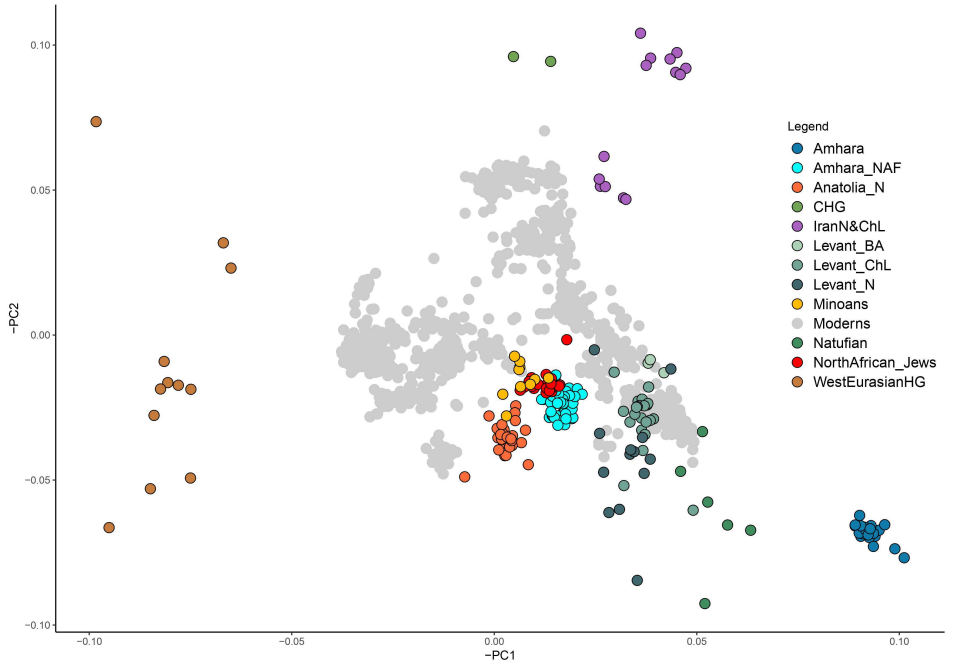

Figure 7. Principal Component Analysis of Amhara group and Amhara Non-African segments (NAF), along with modern West Eurasian populations and ancient samples.

5.1.3 Shared drift estimations

I measured the shared genetic drift between the target samples and the available modern and ancient groups with F3-Outgroup analysis. The analyses were performed on the Ethiopian populations considering either the masked segments (Ethiopian_NAF) or the whole-genome sequences (Ethiopian). A systematic survey has been carried out using the formula: f3(Ethiopian_NAF/Ethiopian, X; Outgroup). I set as X all the available modern and ancient samples from the broad Mediterranean area; while as an outgroup I used the Mbuti population, a hunter-gatherer group from Central Africa. I then compared the results of the masked genomes (on the x axis in Figure 8) against the respective whole-genome population results (on the y axis in Figure 8).

The ancient samples from the Anatolian area (Minoan individuals and Anatolia Neolithic farmers) and modern Jewish populations from North Africa show high values of shared genetic drift with the Ethiopian NAF component. Differently, the whole-genome sequences are genetically closer to the ancient Levantine groups.

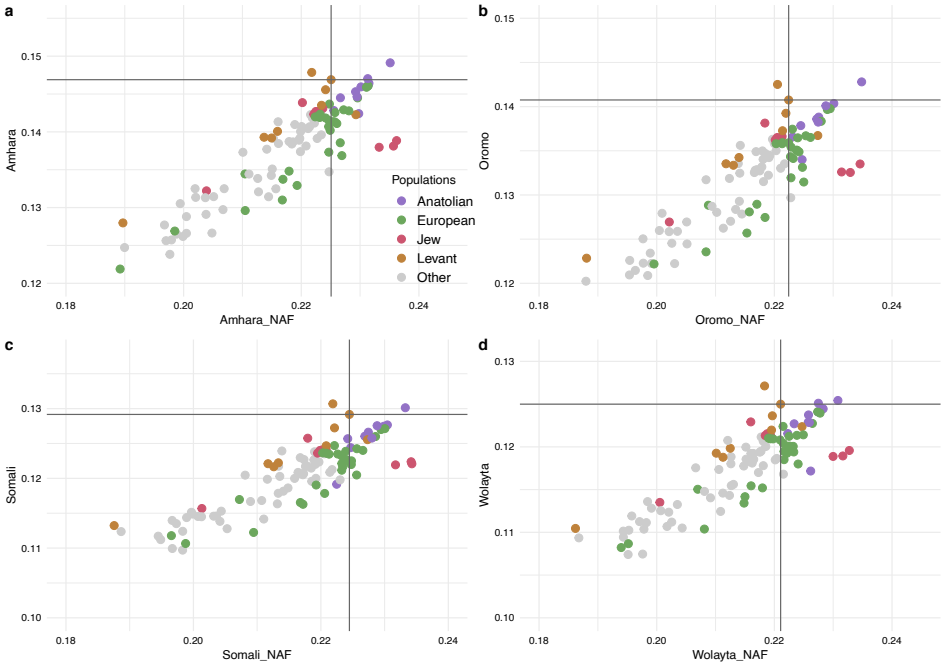

Figure 8. F3-Outgroup analysis of the Ethiopian whole-genome and masked sequences, where masked sequences are presented on the x axis and whole-genome sequences results are presented along the y axis: a) Amhara and Amhara_NAF, b) Oromo and Oromo_NAF, c) Ethiopian Somali and Ethiopian Somali_NAF, listed as Somali in the graph for brevity d) Wolayta and Wolayta_NAF

5.1.4 Modelling multiple ancestral contributors

I aimed to characterise the admixture event in more detail, by modelling the masked and the whole-genome as a mixture of multiple ancestries. I focused on the F3-Outgroup top scoring groups: Anatolian Neolithic (Anatolia_N), Minoan (Minoan_Lasithi and Minoan_Odigitria), Jewish individuals from North Africa and Levantine Neolithic (Levant_N) populations.

I used a custom list of Left populations to test two-ways and three-ways admixture events: the Test population, X and Mota for the two-ways admixture; the Test population, X, Y and Mota for the three-ways admixture. As Test population I included either Ethiopian NAF or Ethiopian whole-genome sequences, as X I selected for each analyses either Anatolia_N, Levant_N or Minoan samples and as Y I selected CHG.

QpWave results indicated that X and Mota and X, Y and Mota represented independent ancestries. QpAdm results for Ethiopian whole and Masked genomes depict the differences already hinted at in previous analyses when comparing whole-sequences genomes and Masked genomes: complete genomes display Mota and Levant_N as most likely ancestries, while NAF Masked genomes show an Anatolian-like and CHG contribution, ~80% and ~20% respectively (Figure 9).

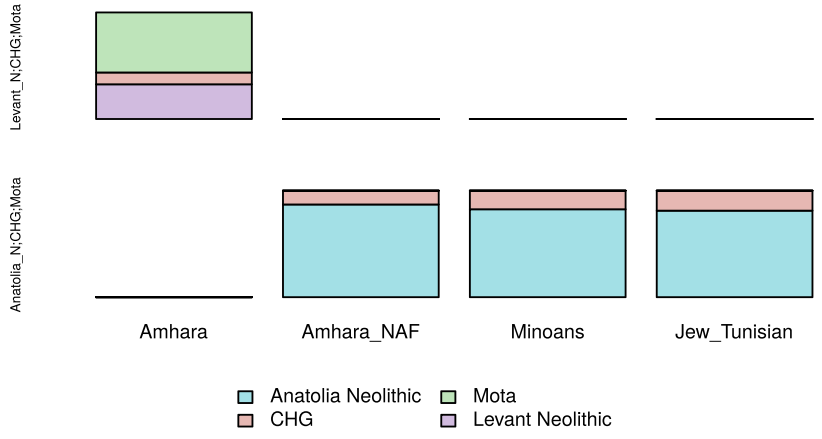

Figure 9. Estimating admixture proportions in Amhara, Amhara_NAF, Minoans and jews from Tunisia, by modelling them as a mixture of Mota, CHG and either Anatolian or Levantine ancestry.

5.1.5 Bias testing

PCAdmix assigns each window to an ancestry with a probability, so that, after the AD, each admixed individual's genomic segments are either: high confidence non African (NAF), low confidence NAF (X), low confidence African (Y) or high confidence African (AF). In the main analyses I discarded the low confidence

segments, however they might bear genetic patterns with different signatures than the ones we found.

I combined the four components and tested their genetic signature through a series of F4 statistics. The low confidence NAF component, when merged together with the high confidence NAF segments, does not qualitatively affect our inferences. Furthermore, I assembled all high confidence components (NAF and AF, referred to as 'Joint') and reiterated the signals of the global whole-genome segments. The results show that the low confidence components are not holding a distinct genetic signature (Figure 10).

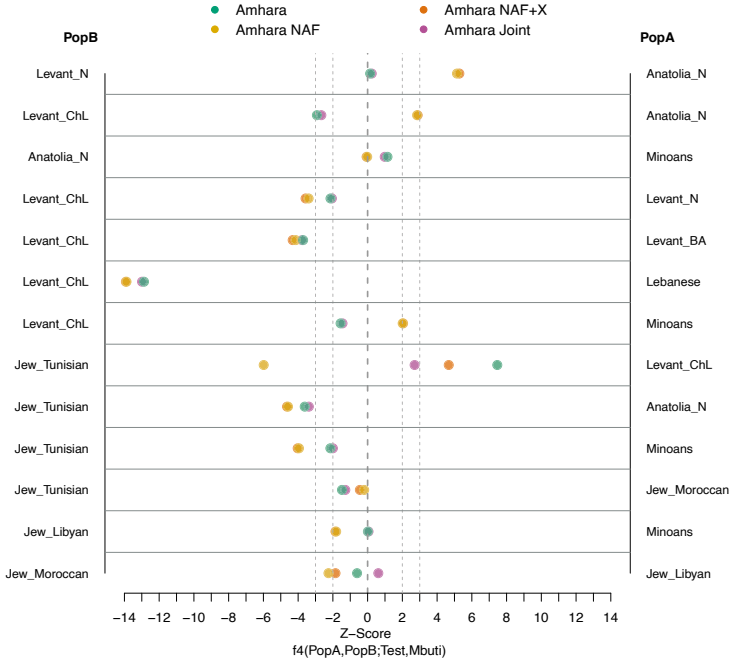

Figure 10. Frequency-based allele-sharing analyses. F4-statistics test on several Amhara ancestral segments: Amhara, where all segments are considered; Amhara_NAF, where only the high confidence Non African segments are retrieved; Amhara Joint, where the high confidence segments, buth African and Non African, are considered and Amhara_NAF+X, where high confidence and low confidence Non African segments are selected

PCAdmix, as many other LAI tools, requires a set of reference populations to detangle the ancestral components of the admixed samples and they are crucial in the analyses.

The analyses shown in the main manuscript are drawn by deconvoluting the Ethiopian ethnic groups with CEU and Gumuz. The CEU population is characterised by European ancestry, which bears additional traces of Anatolian neolithic ancestry and therefore might cause biases against the levantine ancestry. Whereas the Gumuz are an Ethiopian ethnic group characterised by a negligible amount of

Non African ancestry (< 3%) that might assign non African genetic segments to the African component.

We explored potential confounders linked with the choice of CEU and Gumuz by selecting different sets of reference populations. As an alternative to the CEU, we selected the Druze population, a Levantine population showing little signs of recent African admixture (Moorjani et al. 2011), to minimise the distance from the true West Eurasian source. For the alternative African ancestry, we selected YRI, Yoruba from Nigeria, to instead maximise the distance of the African proxy to the population that was likely involved in the admixture event. We also performed the Local Ancestry analyses using a different software, ELAI, and compared the results with PCAdmix. The comparisons were carried out through a series of F4-statistics tests and confirmed the robustness of our approach (Figure 11).

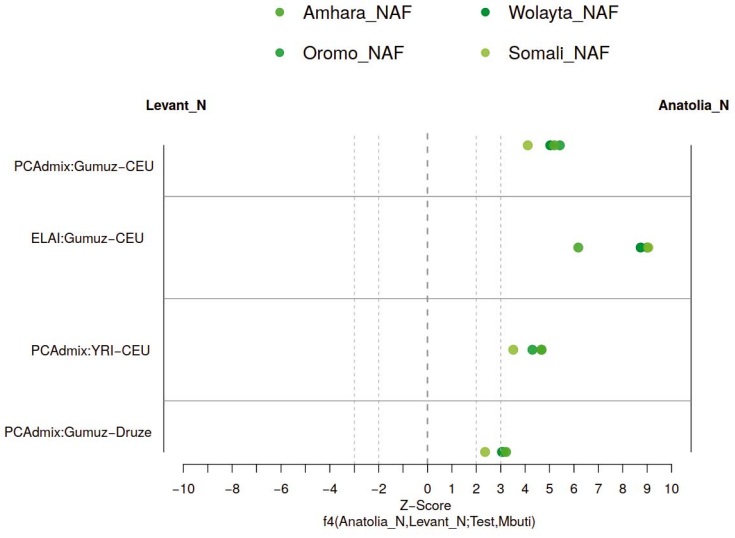

Figure 11. F4 statistics results on masked Ethiopians, using different populations as sources (Gumuz-CEU, YRI-CEU and Gumuz-Druze) **and different LAI tools** (PCAdmix or ELAI) to retrieve the masked segments.

5.2 Understanding and overcoming Local Ancestry inferences limits: WINC

5.2.1 WINC Framework

Our proposed Local Ancestry approach leverages on the ChromoPainter/NNLS framework. Our approach can be divided into three steps: performing ChromoPainter analyses, splitting the copying vectors in genomic windows and analysing each window through NNLS.

In the first step, ChromoPainter should be performed in order to retrieve both the admixed group and the reference populations' copying vectors. In this step both the targeted admixed group as well as the proxy admixing sources should be set as recipients, so that they are all painted by the rest of the donor panel.

In the second step the copying vectors are splitted in genomic windows of the same length. Since the window size depends on the length of the ancestry chunks, an approximate understanding of generations elapsed since the admixture must be available.

Finally, in the third step, WINC compares the painting profile of the admixed sample's genomic window to the references' windows with NNLS analyses and assigns the admixed window to the most representative ancestry. We refer to our approach as Window-based ChromoPainter/NNLS: WINC (Figure 12).

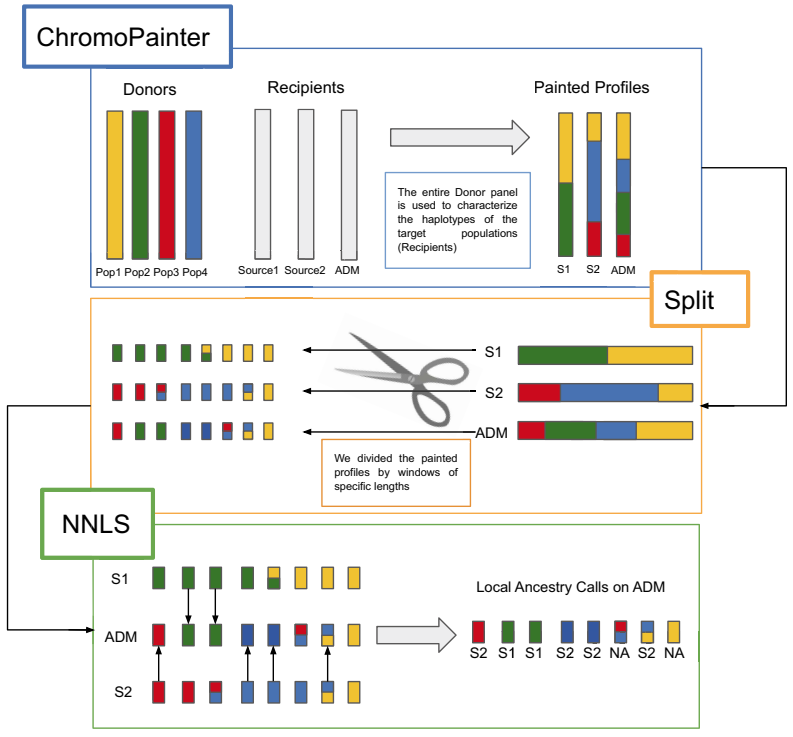

Figure 12. Schematic workflow of WINC, WINC is based on the ChromoPainter/NNLS framework (ChromoPainter step, top part of the figure), with the additional step of splitting the copying vectors resulting from the ChromoPainter (CP) run (Split step, middle part of the figure) before analysing them through the NNLS step (NNLS, bottom part of the figure).

5.2.2 Simulating Test Set

Following a similar model as in Van Dorp et al 2015 (van Dorp et al., 2015), we simulated 13 populations to represent current European (EUR), East Asian (ASN) and African (AFR) groups. We then added seven sister groups (GST), characterised by a divergence time from their sisters of 100 generations (3 kya), for a total of 20 simulated populations (Figure 13). These GST populations were later used to create admixed groups, but were not included in any LA analysis, in order to mimic a real scenario where the actual sources of the admixture are not available. For each group we simulated 50 individuals and phased genomic segments of 250 Mb, mimicking the length of chromosome 1.

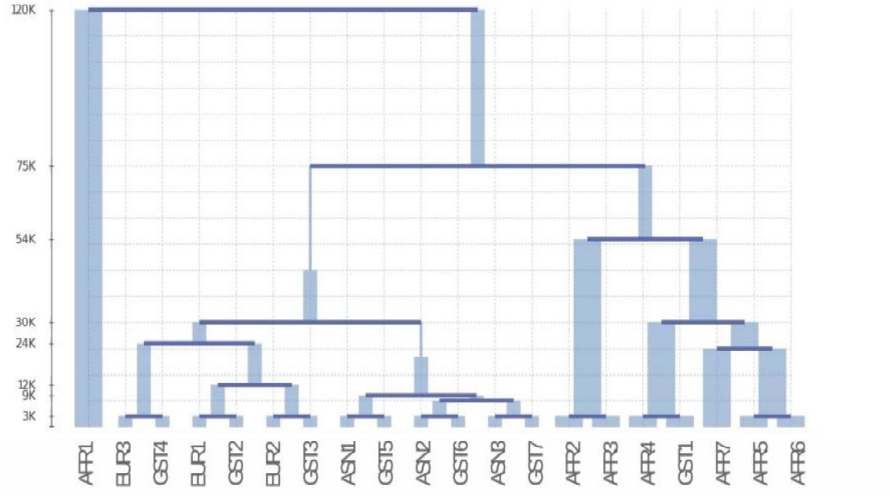

Figure 13. Demography of the simulated populations. The x-axis lists the simulated population labels, y-axis lists the kilo (K) years elapsed from the present.

By combining pairs of simulated GST (Ghost) demes, I generated 8 two-ways admixed and one three-ways admixed populations with 50 individuals each and 100 generations elapsed since the admixture event (Table 1). The admixing GST demes were selected to cover a broad spectrum of divergence times, allowing us to test admixture events with sources progressively more genetically similar. The two-ways admixture are all characterised by ancestral proportions of 70%–30%. The simulated population that was obtained from a three-way admixture is characterised by African-like, European-like and Asian-like contributors, with proportions 40%–30%–30%. The resulting admixed simulated samples were combined with the previously simulated dataset, from which the GST demes were removed. I referred to the obtained dataset as the Test Set.

Table 1. List of the simulated admixed groups using GST demes as sources. In Source 1, Source 2 and Source 3 we listed the GST demes used to create the admixed group and in Proxy 1, Proxy 2 and Proxy 3 we indicated the reference source populations used in the deconvolution analyses. Lastly, in Divergence Time we indicated the divergence time between the sources.

Source 1	Source 2	Source 3	Proxy 1	Proxy 2	Proxy 3	Divergence Time (KYA)
GST1	GST2		AFR4	EUR1		75
GST3	GST7		EUR2	ASN3		30
GST4	GST5		EUR3	ASN1		30
GST7	GST4		ASN3	EUR3		30
GST3	GST4		EUR2	EUR3		24
GST2	GST3		EUR1	EUR2		12
GST5	GST6		ASN1	ASN2		9
GST6	GST7		ASN2	ASN3		7.5
GST1	GST4	GST5	AFR4	EUR3	ASN1	

5.2.3 Testing WINC on different window sizes

Given that all simulated populations showed an AT of 100 generations, the expected length of the ancestry tiles in our dataset is ~1 megabase (Mb).

I compared WINC performances on windows of different lengths, splitting the copying vectors of two admixed populations in genomic tiles of 100 kilo bases (kb), 500 kp and 1 Mb. Results show that different window lengths can affect WINC performances. Notably, a shorter window length causes decrease in performance. Differently, choosing a window length closer to the expected one (1 Mb) does not heavily affect the performance when 50 individuals are used per source. On the other hand, when two individuals are used per source, WINC shows higher accuracy levels when applied on 1 Mb window length.

In the following analyses we maintain as standard window size the length of 500 kb, in order to ensure that each ancestry block analysed falls within the expected window tile of 1 Mb.

5.2.4 Comparison between Local Ancestry tools

We deconvolved the admixed groups of the Test set with several local ancestry tools: ELAI (Guan 2014), RFMix (Maples et al. 2013) and PCAdmix (Brisbin et al. 2012) and WINC. The admixed samples' sources show a degree of similarity based on how long ago they diverged, i.e. if they diverged 75 kya they will show a low level of similarity, while if they diverged 7.5 kya their genetic patterns are expected to be similar. Secondly, the ancestral block length is expected to be short given that 100 generations have elapsed since the admixture. Sources' similarity as well as small block length may interfere with the accuracy levels for all LAI tools analysed here.

I applied a wide range of ancestry assignment thresholds (or ancestry scores, AS), removing all SNPs or windows that did not reach said threshold. I then

compared the resulting assignments with the true ancestral state that was given by the simulation software as a separate output. By comparing the observed ancestry assignment to the known ancestral state I was able to estimate the accuracy levels of the LA tools. Additionally, I estimated the portions of genomic windows available after filtering for the SNPs that did not pass the threshold, expecting that higher thresholds would remove a higher number of SNPs/ windows.

Overall RFMix results show low accuracy levels, independently on how similar the sources' genetic patterns are or how many reference individuals are available. Most likely this is due to the length of the ancestry blocks. On the other hand, ELAI, PCAdmix and WINC all show high accuracy levels (>0.8) where 50 reference individuals are available and the two sources are sufficiently divergent (> 24 kya between sources) (Figure 14). When only two individuals are available as sources, both ELAI and WINC perform with high accuracy levels (> 0.8) when the sources are highly differentiated (diverged 75 kya). WINC maintains accuracy > 0.8 also when the sources are less divergent (30 kya or higher). When the sources diverged less than 24 kya all tools tested show a decrease in accuracy, no matter the number of available reference individuals.

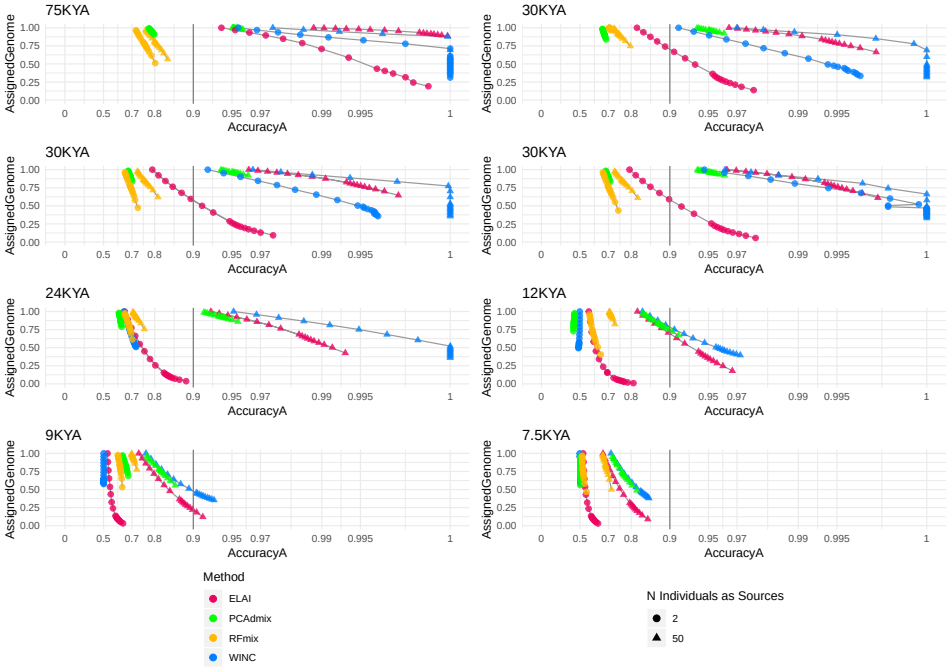

Figure 14. Local Ancestry inferences on two-ways admixture groups from Test Set. Comparing ELAI (red), RFmix (yellow), PCAdmix (green) and WINC (blue) in terms of the window proportion assigned (y axis) and accuracy levels (x axis), under a range of threshold. The eight panels are labelled based on the divergence time of the admixing sources. Local Ancestry inferences were carried out with 50 individuals per sources (triangle) and two (dots)

Overall, results on the three-way admixture event show that ELAI reaches higher performances. Both ELAI and WINC reach high accuracy levels when 50 individuals per source are available, but WINC retrieves fewer genomic windows. The accuracy levels are lower for both methods when performing LA based on only 2 individuals per source, however ELAI retrieves more genomic windows.

5.2.5 C-AS matrix

As seen from the Test set deconvolution results, LAI approaches are expected to perform with higher accuracy when the admixing sources are genetically distant.

Similarly, the more two sources are differentiated at a given genomic window, the easier it should be for the NNLS as well to assign a haplotype to one or the other source population. We leveraged on the similarity between sources to predict whether NNLS has sufficient information to correctly infer the local ancestries.

To assess the similarity between different sources, we computed a Pearson correlation coefficient (rho) between the same window of each pair of source populations from the Test set.

We performed the NNLS analysis on the windows, applying a wide range of thresholds (Assignment Score, AS) and calculated the accuracy of the NNLS assignment.

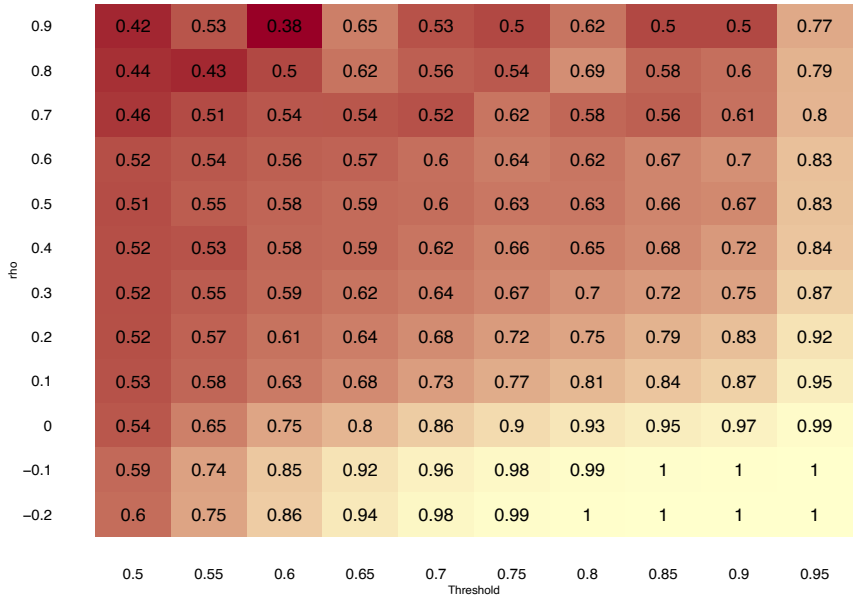

Figure 15. Correlation-AssignmentScore matrix, estimated from the Test set.

In this way, for each window of all pairs of sources, we obtained a correlation coefficient that measured their similarity and the accuracy level reached by the NNLS analyses when assigning the target window to the source under several threshold values.

In doing so, we obtained a Correlation-Assignment Score (C-AS) matrix that, given different values of similarity between sources (correlation) and a threshold (or assignment score (AS)), will inform on the NNLS assignment accuracy values (Figure 15).

5.2.6 Simulating Empirical Set

To test the transferability of the C-AS matrix, obtained from the simulated Test set, I simulated an additional dataset of admixed individuals along with their proxy sources: the Empirical Set. The Empirical Set was characterised by all groups available from the 1000 Genome Project and four admixed groups (The 1000 Genomes Project Consortium et al. 2015).

I simulated two-ways admixture events between an European (TSI, Toscani in Italy) and an African (YRI, Yoruba in Nigeria) population. TSI-YRI; an European (TSI) and Asian (CHB, Han Chinese in Beijing) population, TSI-CHB, and within European populations (TSI and FIN, Finnish in Finland), TSI-FIN. Lastly, I simulated a three-ways admixture event between continents mixing YRI, CHB and TSI with proportions 40%–30%–30% respectively.

I used CEU (Utah residents with European ancestry) as a source population to retrieve TSI fragments, ESN (Esan in Nigeria) for YRI ancestry and CHS (Han Chinese South) for CHB ancestry. To retrieve FIN fragments, we set as source all FIN individuals not used to create the admixed population TSI-FIN.

5.2.7 C-AS Matrix Transferability

We tested the applicability of the C-AS matrix estimated from the Test Set on the Empirical Set. For a given correlation in a specific window, we used the minimum AS threshold needed to obtain the desired accuracy value. We analysed the overall performance and transferability of the C-AS matrix on the Empirical Set and compared it with the results obtained by selecting the windows only by AS thresholds.

Our tool, as well as ELAI, operates with high accuracy values (> 0.9) also on the Empirical Set when the sources are genetically differentiated and 45 or two individuals are available (Figure 16). All LAI tools tested do not reach satisfactory accuracy levels when the source populations are genetically similar, such as TSI-FIN.

For both TSI-YRI and TSI-CHB populations, WINC calibrated with the C-AS matrix performs equally well to WINC alone in terms of accuracy, but retrieves higher portions of the genome, with the additional difference that WINC+C-AS

accuracy is predictable in its outcome. By applying the C-AS matrix to WINC we could in fact assign windows with the desired accuracy, with the only exception being reaching an observed accuracy of ~0.97 when the expected one was set at 0.99. Differently from WINC alone, WINC + C-AS matrix tends to not assign any genomic window of TSI-FIN (maximum 0.1%), when threshold values were set to 0.85 or higher, hence providing an effective way of drastically reducing false positives.

Differently from the Test Set results, ELAI shows lower accuracy levels with respect to both WINC and WINC+C-AS when two individuals are used as source for the LAI analyses on the three-ways admixture group.

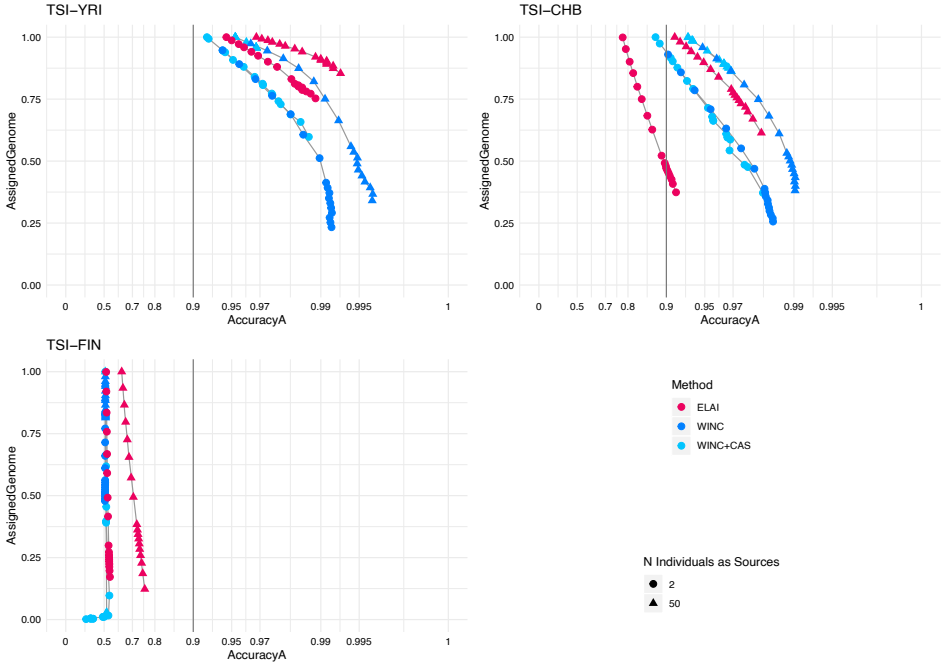

Figure 16. Local ancestry inferences on the Empirical Set using WINC and ELAI, with WINC (blue), WINC+CAS (light blue) and ELAI (red), given 45 (triangles) or 2 (dots) individuals per source.

5.2.8 Real Case scenario

Lastly, we applied the WINC and WINC+C-AS matrix approaches on real genomes: ASW (American of African Ancestry in SW) and MXL (Mexican Ancestry from Los Angeles USA)(The 1000 Genomes Project Consortium et al. 2015). To deconvolute ASW, we used CEU and ESN as reference sources, while for MXL we used CEU, PEL and ESN. Each analysis was composed of either 45 or 2 of source individuals.

Being a real case not resulting from simulations, to assess WINC performances we chose to compare WINC the results with ELAI ancestry assignments using 45 individuals, deemed as the "gold standard". For comparison, we also performed ELAI analyses on ASW and MXL using 2 individuals per source and benchmarked the results against the same ELAI run using 45 reference samples.

Consistently with the previous results on highly divergent simulated populations, WINC shows high accuracy levels deconvoluting ASW, despite the number of reference samples. Discrepancies on the portions of the assigned genome between the real case and the Test Set could be due to the fact that ELAI assigns windows that WINC set as NA, or vice versa.

On the MXL population WINC reaches accuracy of 0.9 or higher when using 45 individuals per source, but unlike ELAI, it does not reach high accuracy levels when inferring the three MXL ancestries when only 2 individuals are used per source (Figure 17).

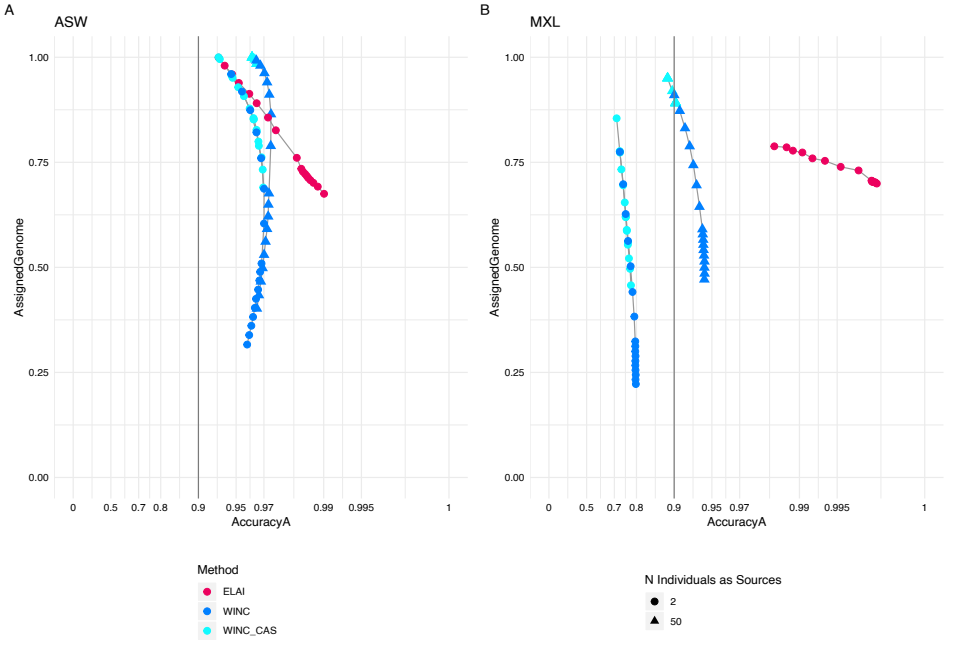

Figure 17. Local Ancestry analyses on ASW and MXL with WINC (blue), WINC+CAS (light blue) and ELAI (red), given 45 (triangles) or 2 (dots) individuals per source.

5.3 Employing Local Ancestry inferences to overcome Polygenic Scores limited transferability

5.3.1 Testing Local Ancestry inferences on a simulated set

We first performed LAI on simulated genomes that mimic the real populations targeted in the study to assess ELAI accuracy when the AT is not precise. I performed several LA runs per each simulated group, indicating different numbers of generations for each run and compared the LAI results with the known truth given by the simulation software. I performed LA given: the exact admixture time, doubled time and half time, an average of the former dates and, finally, a quarter of the exact time. ELAI shows robustness also in cases where the admixture time given is not precise.

5.3.2 Local Ancestry inferences on selected samples from 1000 Genomes project, Ethiopian groups and UK BioBank

We then performed LAI on the UK-BioBank, selected 1000 Genomes project samples, five Ethiopian groups and Egyptian samples using ELAI. All populations were characterised by divergent ancestral components, one of them of West Eurasian origin. To deconvolute the ASW population, Egyptian and Ethiopian whole-genome sequences, we used 72 samples equally distributed among CEU, TSI, IBS to represent the West Eurasian ancestry and GUM (Gumuz), LWK, YRI, to represent the African ancestry. The admixture generations parameter was set as 100 for Ethiopians, 30 for Egyptians, 6 for ASW.

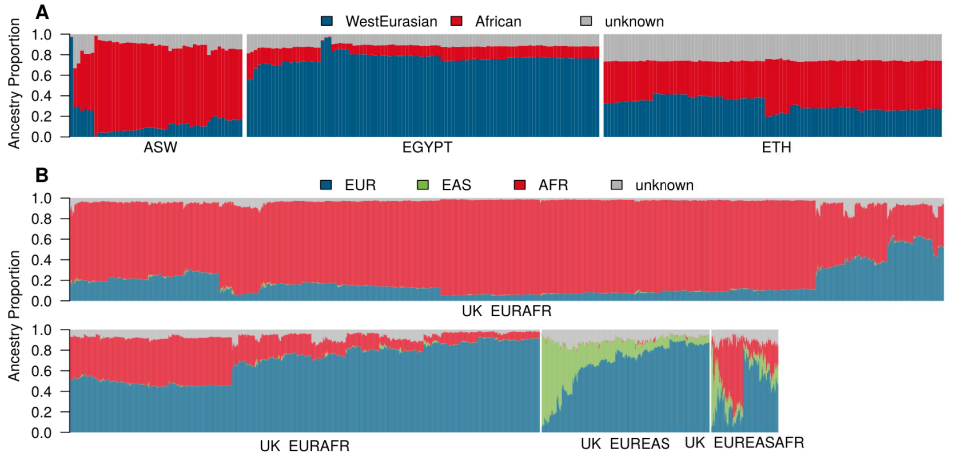

Figure 18. Local Ancestry assignments of the target dataset with a) Ethiopian and Egyptian whole-genome sequences and ASW group from 1000 Genome Project b) admixed samples from UKBB

The UK-BioBank admixed samples were deconvoluted using UK-BioBank individuals to minimise batch effects. To retrieve the West Eurasian ancestry, 100 samples were selected among the UK-BioBank samples that fall near the GWAS training set. For the African and East Asian components we extracted the 100 samples with the highest appropriate ancestry fraction according to ADMIXTURE results. The admixture generations parameter was set as 10 for the admixed UK-BioBank samples.

We selected a threshold of 0.9 for the assignment, so that all SNPs assigned with less than 0.9 probability were removed and labelled as unknown (Figure 18).

5.3.3 Partial Polygenic Scores transferability

The LAI assignments were then used to mask the targeted genomes in order to retrieve European ancestry (Figure 19). Successively, we preceded estimating the polygenic scores (PS) on the masked samples. PS were calculated on the masked genomes accounting for the fraction of the genome available, therefore applying a modified estimation of the PS defined partial polygenic score (pPS). To assess the predictive value of the pPS obtained, a parallel set of masking analyses was carried out on the Estonian BioBank (EstBB).

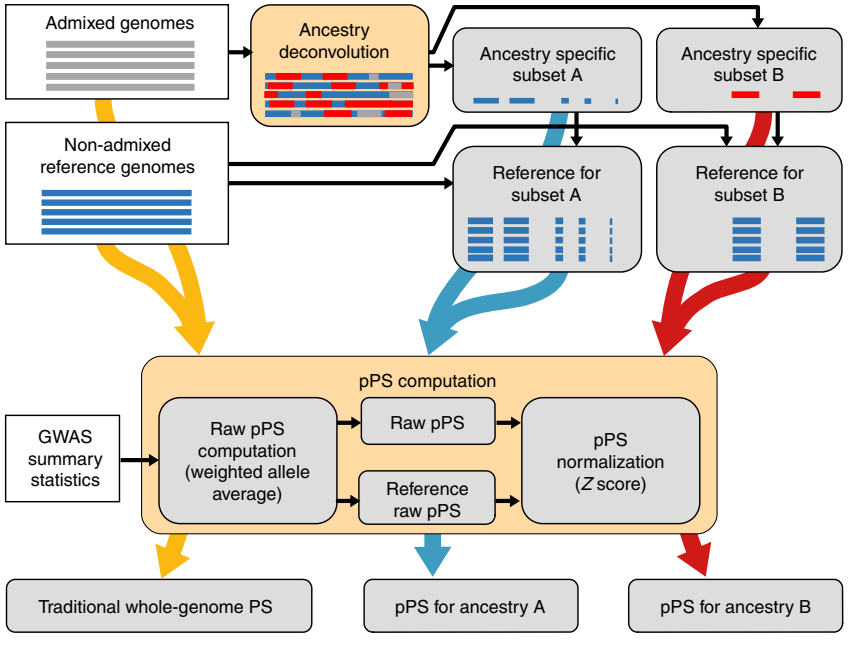

Figure 19. Schematic workflow to obtain partial Polygenic Scores (pPS). A graphical representation of the workflow we adopted to obtain normalised PS and ancestry specific pPS. White boxes represent input data, the two key steps of ancestry deconvolution and partial PS computation have an orange background.

The masked loci obtained from the UK-BioBank were used to mask the EstBB samples, on which the pPS were calculated. Separately, the PS were calculated using the entire, unmasked genomes of the EstBB. We then compared the PS estimates of the EstBB against the pPS estimates of the masked EstBB. Results showed that pPS estimated on only a fraction of the admixed genome, retrieved thanks to LAI, were sufficient to improve prediction estimates. Furthermore, we applied the approach on real admixed genomes showing the effectiveness of the approach also on real-case scenarios. We analysed individuals from the UK BioBank, for which both the East Asian and European ancestrals components could be analysed through pPS, thanks to the availability of the SNP-trait associations studies from the UK BioBank and BioBank Japan. We added together the two ancestry-specific PS and showed that predictivity of the combined-ancestry-specific PR (casPS) outperforms both pPS and PS estimates.

6. CONCLUSIONS

The three studies presented in this thesis aimed to contribute to the current understanding of Ancestry Deconvolution approaches, focusing on Local Ancestry inferences.

Ref I – The West Eurasian component in four Ethiopian ethnic groups (Amhara, Oromo, Ethiopian Somali, Wolayta), when analysed separately, is found to be genetically characterised by the presence of approximately 80% of Anatolia Neolithic ancestry and 20% of CHG ancestry. Such signal is maintained when performing Local ancestry analyses with a non-African source enriched for Levantine-like ancestry and also when accounting for the windows discarded due to their low assignment rate. Differently, when the entire genome is analysed, thus when considering all ancestral layers, the Ethiopian ethnic groups are defined as being admixed with a Levantine Neolithic ancestry and an autochthonous African layer. Our work indicates that when the mixing components are deeply differentiated, such as in the case of contemporary Ethiopians, controlling for all components with ancestry deconvolution may be a supporting tool for allele sharing tests and allow for further insight on past demographic histories.

Ref II – Current LA methods are limited when assessing components that are genetically similar, since they do not have the power to properly distinguish between the admixing sources. Additionally, most of the tools require a large amount of reference samples to reach a satisfactory accuracy level. Our framework was able to reach high accuracy levels when deconvoluting an event whose sources were sufficiently divergent (up to 30 kya) and when only two individuals were used per each source. However, the haplotype-painting framework that was proven successful in describing sub-continental substructures could not yield satisfactory results when applied in a LA framework to deconvolute sources that diverged earlier than 24 kya. Our work allowed us to understand and measure part of the limitations of state-of-the-art tools, along with providing a novel LA methodology able to perform ancestry deconvolution in cases of sample scarcity.

Ref III – Our work showed that Local Ancestry inferences can be employed to estimate Polygenic Scores in admixed individuals. LA allows to select the genomic segments derived from a specific ancestry, shared with the GWAS cohort, and ancestry-specific Polygenic Score can be estimated on the remaining genetic fraction. Our work enables the estimation of Polygenic Scores for individuals of mixed ancestries.

Local Ancestry Inferences can be a viable approach to control for multiple ancestral components and perform ancestry specific analyses. Along with contributing in the understanding of past demographic events, LAI allows to refine GWAS-dependent strategies such as the estimation of PS. Despite the great power

that this approach holds, there are several limitations that should be taken into account, above all the lower accuracy level reachable in case of high similarity between the ancestral components.

Future steps into refining LA's ability of deconvoluting sub-continental admixture will allow us to gain deeper knowledge on the evolutionary history of our species, so that we can continue narrating through past encounters the story of our ancestors.

SUMMARY IN ESTONIAN

Eesti, Euroopa ja üleilmsete inimgenoomide geneetilise päritolu kihtide lahtikaevamine

Tänapäeval eeldame sageli, et meie moodne eluviis on minevikust tingimata parem, ja et mida kaugemale me ajas tagasi läheme, seda enam heitlesid mineviku inimesed kehvade elutingimuste või küündimatu tehnoloogiaga. Pole üllatav, et ajaloolised ja arheoloogilised uuringud äratavad meist paljude uudishimu, kuna nende tõttu oleme sunnitud oma tänapäevast uskumuste süsteemi ümber kujundama ning jahmuma sellest, kui palju mineviku populatsioonid tegelikult saavutada suutsid. Ajaloolised ja arheoloogilised andmed meenutavad meile jätkuvalt, et minevikku tuleb vaadata värske pilguga. Samavõrd nagu tänapäeval olid mineviku inimesed organiseerunud ühiskonnaks, nad liikusid, nad rändasid, nad jagasid ideid, kultuure ja tehnoloogiaid – ja me leiame ikka veel nende kokkupuudete jälgi. Meie mineviku ajaloo täielikuks mõistmiseks ei piisa ühestainsast silmapaarist, kuna ükski teadusvaldkond pole täiesti sõltumatu. Seda silmas pidades on käesoleva doktoritöö eesmärgiks mineviku kokkupuudete jälgede uurimine geneetilistes andmetes.

Kui indiviidid segunevad, et saada järeltulijaid, leiab populatsioonides aset segunemissündmus, mis tekitab segunenud populatsioone. Segunemissündmuse osaliste jälgede ajamise protsessi nimetatakse põlvnemise lahtiharutamiseks (Ancestry Deconvolution, AD). AD on lähenemisviis, mis võimaldab segunenud grupi geneetilist mosaiiki analüüsida, ajades segunemise osaliste jälgi, ja segunemissündmust täpsemalt iseloomustada.

Vaadates genoomi AD abil kui minevikusündmuste tulemit, avastame, et populatsioonide kokkupuuted pole inimajaloos olnud kaugeltki harvad, mõnikord saavad kokku geneetiliselt lähedased populatsioonid (segunemine maailmajao siseselt), mõnikord jällegi segunevad väga erinevad grupid (segunemine üle maailmajao).

AD lähenemisviiside seas on meetodid, mida nimetatakse lokaalse põlvnemise tuletamiseks (Local Ancestry Inferences, LAI), mis võimaldavad tuletada antud päriliku lookuse põlvnemist. Sellistel kõrge lahutusastmega tuletustel on piirangud, mis mõnel juhul LAI tööriistade jõudlust tugevalt piiravad. Tegelikult käsitlevad LAI rakendused peamiselt segunemissündmusi üle maailmajao, kus on segunenud väga erinevad allikad.

Samuti, kui LA rakendamine segunenud populatsioonil, mida iseloomustab põlvnemine eri maailmajagudest, võib olla täielikult peidetud ja tundmatute mineviku demograafiliste sündmuste avastamisele keskendunud ülesanne, nõuab LA kasutamine maailmajao sisestel segunemistel, näiteks Euroopa populatsioonidel, metodoloogilist lähenemist, mis on kujundatud nii, et see võtab kõigepealt arvesse LA piiranguid.

Lõpuks, toetudes teadmisele, et üldiselt on kõik inimrühmad segunenud, on loomulik pakkuda, et AD ja täpsemalt LAI on lähenemisviisidena kasulikud ka

väljaspool demograafilisi uuringuid. Tegelikult peitub 3 miljardis aluspaaris hämmastavalt hiiglaslik hulk informatsiooni, mis võimaldab meil esitada väga laiu teaduslikke küsimusi. Koos segunevate allikate põlvnemiskomponentidega pärime konkreetsete tunnuste (või fenotüüpidega) seotud SNP-d. Teatud fenotüübi väljakujunemise tõenäosust on võimalik osaliselt ennustada polügeensete skooride (Polygenic Scores, PS) hindamise abil. See hinnang saadakse, summeerides kõigi tunnusega seotud alleelide panuse üle kogu genoomi, mis kas suurendaks või vähendaks fenotüübilise tulemuse tõenäosust, kaalutuna alleeli mõju suuruse järgi. Seos SNP-de ja tunnuste vahel tuletatakse ülegenoomsetest assotsiatsiooniuuringutest (Genome-Wide Association Studies, GWAS) (Martin jt 2017). Samas viiakse GWAS uuringuid tavaliselt läbi suurtes kohortides ja mõni grupp, näiteks eurooplased (Sirugo, Williams ja Tishkoff 2019; Kim jt 2018), on teistest rohkem kaetud (Landry et al. 2018). Kuna genoomile mõjuv tohutu hulk muutujaid on populatsioonispetsiifilised, saab GWAS-i tuletusi kasutada ainult nende populatsioonide puhul, mis on GWAS-i aluseks olnud populatsioonile geneetiliselt lähedased, mis takistab PS-i ülekantavust ja rakendatavust vähem uuritud populatsioonidele ja segunenud indiviididele.

Ma esitlen kolme uuringut, mille eesmärgiks on arutleda põlvnemise lahtiharutamise (Ancestry Deconvolution, AD) lähenemisviiside võimalusi ja piiranguid. AD lähenemisviise, täpsemalt LA tuletamist, on rakendatud kolmel erineval teemal: uurides segunemissündmust, milles said kokku väga erinevad populatsioonid eri maailmajagudest; uurides väikesel skaalal segunemissündmuste, näiteks Euroopa rühmade puhul, lahtiharutamise piiranguid; ja lõpuks rakendades kogutud teadmisi, ületamaks polügeensete skooride ja GWAS hinnangute kasutamise piiranguid segunenud populatsioonidel.

Ref I – ida-aafriklasi iseloomustab geneetiliselt mitte-Aafrika kihistus, mis pärineb pronksiaegsest migratsioonide lainest Lääne-Euraasiast, ja autohtoonne Aafrika kihistus. Hoolimata paljudest mitte-Aafrika kihistust kirjeldanud uuringutest pole selle komponendi päritolu osas veel konsensusele jõutud. Pickrell jt väitsid, et kihistus on geneetiliselt lähedasem Sardiinia tüüpi põlvnemisele, aga Lazaridis jt 2016 tuvastasid geneetilise sarnasuse neoliitilise Levandi põlluharijatega (Pickrell jt 2014; Lazaridis jt 2016). Mõlemas uuringus analüüsiti etiooplaste põlvnemiskihistusi globaalsete põlvnemismeetoditega, seega käsitledes mõlemat komponenti koos. Kuid ülegenoomseid tuletusi võib varjutada suur hulk kokkusobimatut informatsiooni, mis on kihistustes olemas. Alustades LAI lähenemisviisidega, on minu eesmärgiks uurida Etioopia mitte-Aafrika kihistuse demograafilist ajalugu, kasutades põlvnemisspetsiifilisi analüüse.

Ref II – LAI tööriistu on rakendatud erinevate põlvnemiskomponentidega segunenud populatsioonide puhul, kui segunemise osalised pärinevad erinevatest maailmajagudest (nt ladina-ameerika ja aafrika-ameerika grupid) ja iga referentsgrupi kohta on olemas suur hulk proove. See ei kehti ühegi Euroopa populatsiooni kohta, kuna kõigi tänapäeva Euroopa populatsioonide geneetilise koostise kujundamisse panustasid neoliitikumijärgsed migratsioonilained ja järgnevad ränded maailmajao sees, mistõttu on segunemise allikad LAI täpseks läbiviimiseks geneetiliselt liiga sarnased.

Teiseks on LAI tuletusteks vaja suurt arvu proove igast referentspopulatsioonist, et segunenud indiviidide geneetilist varieeruvust kõige paremini kirjeldada. Kuid ehkki selline nõue võib mõne põhjalikult uuritud ja laialdaselt kättesaadava inimrühma puhul olla triviaalne, ei pruugi see nii olla teiste oluliste inimrühmade või teiste taksonite puhul.

Võrdlesin erinevate LAI tööriistade täpsuse taset erinevate stsenaariumite puhul, kus segunemise allikatel oli erinev sarnasuse määr ja proovide kättesaadavus, et paremini mõista LAI tööriistade piirangute praegust olukorda. Lisaks pakkusin välja uue LAI tööriista WINC (Window-based ChromoPainter/NNLS), mis kasutab haplotüüpide „värvimise“ tehnikat. On näidatud, et selline tehnika kirjeldab maailmajao sisest populatsioonistruktuuri täpselt ja seda ei mõjuta väike proovide arv, kuna „värvimise“ etapp viiakse läbi indiviidi tasemel (Drineas, Lewis ja Paschou 2010; Leslie jt 2015; Gilbert Edmund jt 2019; Pankratov jt 2020; Saint Pierre jt 2020; Martin jt 2018; Bycroft jt 2019).

Ref III – polügeensed skoorid (PS-d) summeerivad paljude geneetiliste variantide mõju, millel on näidatud seos mõne fenotüübi või haigusega (Lambert, Abraham ja Inouye 2019; Dudbridge 2013). Samas põhinevad PS-d paljude assotsieerunud alleelide populatsioonist sõltuvatel panustel ning nende rakendatavus vähem uuritud populatsioonidele ja hiljuti segunenud indiviididele on piiratud. Tegime ettepaneku kasutada LAI-d põlvnemisspetsiifiliste SNP-de eraldamiseks ja hinnata PS osaliste lahtiharutatud segmentide kohta.

REFERENCES

Akiyama, Masato, Kazuyoshi Ishigaki, Saori Sakaue, Yukihide Momozawa, Momoko Horikoshi, Makoto Hirata, Koichi Matsuda, et al. 2019. "Characterizing Rare and Low-Frequency Height-Associated Variants in the Japanese Population." *Nature Communications* 10 (1): 4393. https://doi.org/10.1038/s41467-019-12276-5.

Akiyama, Masato, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, et al. 2017. "Genome-Wide Association Study Identifies 112 New Loci for Body Mass Index in the Japanese Population." *Nature Genetics* 49 (10): 1458–67. https://doi.org/10.1038/ng.3951.

Alexander, D H, J Novembre, and K Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64. https://doi.org/10.1101/gr.094052.109.

Alves, Isabel, Armande Ang Houle, Julie G Hussin, and Philip Awadalla. 2017. "The Impact of Recombination on Human Mutation Load and Disease." *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1736): 20160465. https://doi.org/10.1098/rstb.2016.0465.

Auton, A, K Bryc, A R Boyko, K E Lohmueller, J Novembre, A Reynolds, A Indap, et al. 2009. "Global Distribution of Genomic Diversity Underscores Rich Complex History of Continental Human Populations." *Genome Research* 19 (5): 795–803. https://doi.org/DOI 10.1101/gr.088898.108.

Banning, E B. 2011. "So Fair a House: Göbekli Tepe and the Identification of Temples in the Pre-Pottery Neolithic of the Near East." *Current Anthropology* 52 (5): 619–60. https://doi.org/10.1086/661207.

Baran, Yael, Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, et al. 2012. "Fast and Accurate Inference of Local Ancestry in Latino Populations." *Bioinformatics (Oxford, England)* 28 (10): 1359–67. https://doi.org/10.1093/bioinformatics/bts144.

Barbujani, G, and V Colonna. 2010. "Human Genome Diversity: Frequently Asked Questions." *Trends in Genetics* 26 (7): 285–95. https://doi.org/S0168-9525(10)00078-8 [pii] 10.1016/j.tig.2010.04.002.

Barreiro, Luis B, Guillaume Laval, Hélène Quach, Etienne Patin, and Lluís Quintana-Murci. 2008. "Natural Selection Has Driven Population Differentiation in Modern Humans." *Nature Genetics* 40 (3): 340–45. https://doi.org/10.1038/ng.78.

Bar-Yosef, O., and F. Valla. 1990. "The Natufian Culture and the Origin of the Neolithic in the Levant." *Current Anthropology* 31 (4): 433–36.

Behar, Doron M., Bayazit Yunusbayev, Mait Metspalu, Ene Metspalu, Saharon Rosset, Jüri Parik, Siiri Rootsi, et al. 2010. "The Genome-Wide Structure of the Jewish People." *Nature* 466 (7303): 238–42. https://doi.org/10.1038/nature09103.

Bhatia, Gaurav, Nick Patterson, and Sriram Sankararaman. 2013. "Estimating and Interpreting F ST : The Impact of Rare Variants." https://doi.org/10.1101/gr.154831.113.

Bortolini, Eugenio, Luca Pagani, Enrico R Crema, Stefania Sarno, Chiara Barbieri, Alessio Boattini, Marco Sazzini, et al. 2017. "Inferring Patterns of Folktale Diffusion Using Genomic Data." *Proceedings of the National Academy of Sciences* 114 (34): 9140–45. https://doi.org/10.1073/pnas.1614395114.

Bortolini, Eugenio, Luca Pagani, Gregorio Oxilia, Cosimo Posth, Federica Fontana, Federica Badino, Tina Saupe, et al. 2021. "Early Alpine Occupation Backdates Westward Human Migration in Late Glacial Europe." *Current Biology* 31 (11): 2484–2493.e7. https://doi.org/10.1016/j.cub.2021.03.078.

Brisbin, Abra, Bryc, Katarzyna, Byrnes, Jake, Zakharia, Fouad, Omberg, Larsson, Degenhardt, Jeremiah, Reynolds, Andrew, Ostrer, Harry, Mezey, Jason G., and Bustamante, Carlos D. 2012. "PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations." *Human Biology* 84 (4): 343–64. https://doi.org/10.3378/027.084.0401.

Browning, Sharon R., and Brian L. Browning. 2011. "Haplotype Phasing: Existing Methods and New Developments." *Nature Reviews Genetics* 12 (10): 703–14. https://doi.org/10.1038/nrg3054.

Busby, George BJ, Gavin Band, Quang Si Le, Muminatou Jallow, Edith Bougama, Valentina D Mangano, Lucas N Amenga-Etego, et al. 2016. "Admixture into and within Sub-Saharan Africa." Edited by Joseph K Pickrell. *ELife* 5 (June): e15266. https://doi.org/10.7554/eLife.15266.

Bustamante, Carlos D., Francisco M. De La Vega, and Esteban G. Burchard. 2011. "Genomics for the World." *Nature* 475 (7355): 163–65. https://doi.org/10.1038/475163a.

Bycroft, Clare, Ceres Fernandez-Rozadilla, Clara Ruiz-Ponte, Inés Quintela, Ángel Carracedo, Peter Donnelly, and Simon Myers. 2019. "Patterns of Genetic Differentiation and the Footprints of Historical Migrations in the Iberian Peninsula." *Nature Communications* 10 (1): 551. https://doi.org/10.1038/s41467-018-08272-w.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–9. https://doi.org/10.1038/s41586-018-0579-z.

Chakraborty R and Weiss K M. 1988. "Admixture as a Tool for Finding Linked Genes and Detecting That Difference from Allelic Association between Loci." *Proceedings of the National Academy of Sciences* 85 (23): 9119–23. https://doi.org/10.1073/pnas.85.23.9119.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (1). https://doi.org/10.1186/s13742-015-0047-8.

Chen, Donghui, and Robert J. Plemmons. 2009. "Nonnegativity Constraints in Numerical Analysis." In *The Birth of Numerical Analysis*, 109–39. WORLD SCIENTIFIC. https://doi.org/10.1142/9789812836267_0008.

Chintalapati, Manjusha, Nick Patterson, and Priya Moorjani. 2022. "Reconstructing the Spatiotemporal Patterns of Admixture during the European Holocene Using a Novel Genomic Dating Method." *BioRxiv*, January, 2022.01.18.476710. https://doi.org/10.1101/2022.01.18.476710.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F. O'Reilly. 2020. "Tutorial: A Guide to Performing Polygenic Risk Score Analyses." *Nature Protocols* 15 (9): 2759–72. https://doi.org/10.1038/s41596-020-0353-1.

Cline, Eric H, Assaf Yasur-Landau, and Nurith Goshen. 2011. "New Fragments of Aegean-Style Painted Plaster from Tel Kabri, Israel." *American Journal of Archaeology* 115 (2): 245–61.

Currie, Thomas E., Andrew Meade, Myrtille Guillon, and Ruth Mace. 2013. "Cultural Phylogeography of the Bantu Languages of Sub-Saharan Africa." *Proceedings of the Royal Society B: Biological Sciences* 280 (1762): 20130695. https://doi.org/10.1098/rspb.2013.0695.

De La Vega, Francisco M., and Carlos D. Bustamante. 2018. “Polygenic Risk Scores: A Biased Prediction?” *Genome Medicine* 10 (1): 100. https://doi.org/10.1186/s13073-018-0610-x.

Delaneau, Olivier, Jonathan Marchini, Gil A. McVean, Peter Donnelly, Gerton Lunter, Jonathan L. Marchini, Simon Myers, et al. 2014. “Integrating Sequence and Array Data to Create an Improved 1000 Genomes Project Haplotype Reference Panel.” *Nature Communications* 5 (1): 3934. https://doi.org/10.1038/ncomms4934.

Dietrich, Laura, Julia Meister, Oliver Dietrich, Jens Notroff, Janika Kiep, Julia Heeb, André Beuger, and Brigitta Schütt. 2019. “Cereal Processing at Early Neolithic Göbekli Tepe, Southeastern Turkey.” *PLOS ONE* 14 (5): 1–34. https://doi.org/10.1371/journal.pone.0215214.

Donnelly, Kevin P. 1983. “The Probability That Related Individuals Share Some Section of Genome Identical by Descent.” *Theoretical Population Biology* 23 (1): 34–63. https://doi.org/10.1016/0040-5809(83)90004-7.

Drineas, Petros, Jamey Lewis, and Peristera Paschou. 2010. “Inferring Geographic Coordinates of Origin for Europeans Using Small Panels of Ancestry Informative Markers.” *PLOS ONE* 5 (8): e11892. https://doi.org/10.1371/journal.pone.0011892.

Dudbridge, Frank. 2013. “Power and Predictive Accuracy of Polygenic Risk Scores.” *PLoS Genetics* 9 (3): e1003348–e1003348. https://doi.org/10.1371/journal.pgen.1003348.

Durand, Eric Y., Nick Patterson, David Reich, and Montgomery Slatkin. 2011. “Testing for Ancient Admixture between Closely Related Populations.” *Molecular Biology and Evolution* 28 (8): 2239–52. https://doi.org/10.1093/molbev/msr048.

Falush, Daniel, Matthew Stephens, and Jonathan K Pritchard. 2003. “Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies.” *Genetics* 164 (4): 1567–87. https://doi.org/10.1093/genetics/164.4.1567.

Fu, Qiaomei, Cosimo Posth, Mateja Hajdinjak, Martin Petr, Swapan Mallick, Daniel Fernandes, Anja Furtwängler, et al. 2016. “The Genetic History of Ice Age Europe.” *Nature* 534 (7606): 200–205. https://doi.org/10.1038/nature17993.

Fumagalli, Matteo, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E. Jørgensen, Thorfinn S. Korneliussen, et al. 2015. “Greenlandic Inuit Show Genetic Signatures of Diet and Climate Adaptation.” *Science (New York, N.Y.)* 349 (6254): 1343–47. https://doi.org/10.1126/science.aab2319.

Geza, Ephifania, Jacquiline Mugo, Nicola J Mulder, Ambroise Wonkam, Emile R Chimusa, and Gaston K Mazandu. 2019. “A Comprehensive Survey of Models for Dissecting Local Ancestry Deconvolution in Human Genome” 20 (April 2018): 1709–24. https://doi.org/10.1093/bib/bby044.

Gilbert, Edmund, Seamus O’Reilly, Michael Merrigan, Darren McGettigan, Veronique Vitart, Peter K. Joshi, David W. Clark, et al. 2019. “The Genetic Landscape of Scotland and the Isles.” *Proceedings of the National Academy of Sciences* 116 (38): 19064–70. https://doi.org/10.1073/pnas.1904761116.

Gilbert Edmund, O’Reilly Seamus, Merrigan Michael, McGettigan Darren, Vitart Veronique, Joshi Peter K., Clark David W., et al. 2019. “The Genetic Landscape of Scotland and the Isles.” *Proceedings of the National Academy of Sciences* 116 (38): 19064–70. https://doi.org/10.1073/pnas.1904761116.

Gravel, S. 2012. “Population Genetics Models of Local Ancestry.” *Genetics* 191 (2): 607–19. https://doi.org/10.1534/genetics.112.139808.

Gravel, Simon, Brenna M. Henn, Ryan N. Gutenkunst, Amit R. Indap, Gabor T. Marth, Andrew G. Clark, Fuli Yu, Richard A. Gibbs, The 1000 Genomes Project, and Carlos D. Bustamante. 2011. "Demographic History and Rare Allele Sharing among Human Populations." Edited by David L. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, et al. *Proceedings of the National Academy of Sciences* 108 (29): 11983–88. https://doi.org/10.1073/pnas.1019276108.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22. https://doi.org/10.1126/science.1188021.

Guan, Yongtao. 2014a. "Detecting Structure of Haplotypes And" 196 (March): 625–42. https://doi.org/10.1534/genetics.113.160697.

Günther, Torsten, and Mattias Jakobsson. 2016. "Genes Mirror Migrations and Cultures in Prehistoric Europe—a Population Genomic Perspective." *Current Opinion in Genetics & Development* 41: 115–23. https://doi.org/10.1016/j.gde.2016.09.004.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe." *Nature* 522 (7555): 207–11. https://doi.org/10.1038/nature14317.

Hellenthal, G. 2019. "Population Structure, Demography and Recent Admixture." In *Handbook of Statistical Genomics*, 247–74. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119487845.ch8.

Hellenthal, Garrett, George B. J. Busby, Gavin Band, James F. Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. 2014. "A Genetic Atlas of Human Admixture History." *Science* 343 (6172): 747–51. https://doi.org/10.1126/science.1243518.

Hellenthal, Garrett, and Matthew Stephens. 2007. "MsHOT: Modifying Hudson's Ms Simulator to Incorporate Crossover and Gene Conversion Hotspots." *Bioinformatics* 23 (4): 520–21. https://doi.org/10.1093/bioinformatics/btl622.

Hodgson, Jason A., Connie J. Mulligan, Ali Al-Meeri, and Ryan L. Raaum. 2014. "Early Back-to-Africa Migration into the Horn of Africa." *PLOS Genetics* 10 (6): 1–18. https://doi.org/10.1371/journal.pgen.1004393.

Hoffmann, D L, C D Standish, M García-Diez, P B Pettitt, J A Milton, J Zilhão, J J Alcolea-González, et al. 2018. "U-Th Dating of Carbonate Crusts Reveals Neandertal Origin of Iberian Cave Art." *Science* 359 (6378): 912–15. https://doi.org/10.1126/science.aap7778.

Hoggart, C J, M D Shriver, R A Kittles, D G Clayton, and P M McKeigue. 2004. "Design and Analysis of Admixture Mapping Studies." *American Journal of Human Genetics* 74 (5): 965–78. https://doi.org/10.1086/420855.

Holsinger, K E, and B S Weir. 2009. "Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting F(ST)." *Nat Rev Genet* 10 (9): 639–50. https://doi.org/10.1038/nrg2611.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies." *PLOS Genetics* 5 (6): e1000529. https://doi.org/10.1371/journal.pgen.1000529.

Hudjashov, Georgi, Tatiana M. Karafet, Daniel J. Lawson, Sean Downey, Olga Savina, Herawati Sudoyo, J. Stephen Lansing, Michael F. Hammer, and Murray P. Cox. 2017. "Complex Patterns of Admixture across the Indonesian Archipelago." *Molecular Biology and Evolution* 34 (10): 2439–52. https://doi.org/10.1093/molbev/msx196.

Hudson, Richard R. 2002. “Generating Samples under a Wright–Fisher Neutral Model of Genetic Variation.” *Bioinformatics* 18 (2): 337–38. https://doi.org/10.1093/bioinformatics/18.2.337.

Jakobsson, Mattias, Michael D Edge, and Noah A Rosenberg. 2013. “The Relationship Between F ST and the Frequency” 193 (February): 515–28. https://doi.org/10.1534/genetics.112.144758.

Jobling, M.A. 2004. *Human Evolutionary Genetics: Origins, Peoples and Disease (1st Ed.)*. Garland Science.

Jolliffe, Ian T, and Jorge Cadima. 2016. “Principal Component Analysis: A Review and Recent Developments.” *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2065): 20150202–20150202. https://doi.org/10.1098/rsta.2015.0202.

Jones, Eppie R., Gloria Gonzalez-Fortes, Sarah Connell, Veronika Siska, Anders Eriksson, Rui Martiniano, Russell L. McLaughlin, et al. 2015. “Upper Palaeolithic Genomes Reveal Deep Roots of Modern Eurasians.” *Nature Communications* 6 (1): 8912. https://doi.org/10.1038/ncomms9912.

Kelleher, Jerome, Alison M Etheridge, and Gilean McVean. 2016. “Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes.” *PLOS Computational Biology* 12 (5): e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

Kim, Michelle S., Kane P. Patel, Andrew K. Teng, Ali J. Berens, and Joseph Lachance. 2018. “Genetic Disease Risks Can Be Misestimated across Global Populations.” *Genome Biology* 19 (1): 179. https://doi.org/10.1186/s13059-018-1561-7.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa, and Connie J. Mulligan. 2009. “Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age Origin of Semitic in the Near East.” *Proceedings of the Royal Society B: Biological Sciences* 276 (1668): 2703–10. https://doi.org/10.1098/rspb.2009.0408.

Kivisild, Toomas, Maere Reidla, Ene Metspalu, Alexandra Rosa, Antonio Brehm, Erwan Pennarun, Jüri Parik, Tarekegn Geberhiwot, Esien Usanga, and Richard Villems. 2004. “Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears.” *The American Journal of Human Genetics* 75 (5): 752–70. https://doi.org/10.1086/425161.

L. R. Rabiner. 1989. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE* 77 (2): 257–86. https://doi.org/10.1109/5.18626.

Lambert, Samuel A, Gad Abraham, and Michael Inouye. 2019. “Towards Clinical Utility of Polygenic Risk Scores.” *Human Molecular Genetics* 28 (R2): R133–42. https://doi.org/10.1093/hmg/ddz187.

Landry, Latrice G., Nadya Ali, David R. Williams, Heidi L. Rehm, and Vence L. Bonham. 2018. “Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice.” *Health Affairs* 37 (5): 780–85. https://doi.org/10.1377/hlthaff.2017.1595.

Lao, Oscar, Timothy T. Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, et al. 2008. “Correlation between Genetic and Geographic Structure in Europe.” *Current Biology* 18 (16): 1241–48. https://doi.org/10.1016/j.cub.2008.07.049.

Lasserre, Julia, and Christopher M. Bishop. 2007. “Generative or Discriminative? Getting the Best of Both Worlds.” In *Bayesian Statistics 8*, edited by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 8:3–24. Oxford University Press.

Lawson, D J, G Hellenthal, S Myers, and D Falush. 2012. “Inference of Population Structure Using Dense Haplotype Data.” *PLoS Genet* 8 (1): e1002453. https://doi.org/10.1371/journal.pgen.1002453.

Lawson, Daniel J., Lucy van Dorp, and Daniel Falush. 2018. “A Tutorial on How Not to Over-Interpret STRUCTURE and ADMIXTURE Bar Plots.” *Nature Communications* 9 (1): 1–11. https://doi.org/10.1038/s41467-018-05257-7.

Lazaridis, Iosif, Dani Nadel, Gary Rollefson, Deborah C. Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, et al. 2016. “Genomic Insights into the Origin of Farming in the Ancient Near East.” *Nature* 536 (7617): 419–24. https://doi.org/10.1038/nature19310.

Lazaridis, Iosif, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, et al. 2014. “Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans.” *Nature* 513 (7518): 409–13. https://doi.org/10.1038/nature13673.

Leitsalu, Liis, Toomas Haller, Tõnu Esko, Mari-Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, et al. 2015. “Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu.” *International Journal of Epidemiology* 44 (4): 1137–47. https://doi.org/10.1093/ije/dyt268.

Leslie, Stephen, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, et al. 2015. “The Fine-Scale Genetic Structure of the British Population.” *Nature* 519 (7543): 309–14. https://doi.org/10.1038/nature14230.

Li, Na, and Matthew Stephens. 2003. “Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data.” *Genetics* 165 (4): 2213–33. https://doi.org/10.1093/genetics/165.4.2213.

Li, Yun R., and Brendan J. Keating. 2014. “Trans-Ethnic Genome-Wide Association Studies: Advantages and Challenges of Mapping in Diverse Populations.” *Genome Medicine* 6 (10): 91. https://doi.org/10.1186/s13073-014-0091-5.

Liang, Mason, and Rasmus Nielsen. 2014. “The Lengths of Admixture Tracts” 197 (July): 953–67. https://doi.org/10.1534/genetics.114.162362.

Llorente, M. Gallego, E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, et al. 2015. “Ancient Ethiopian Genome Reveals Extensive Eurasian Admixture in Eastern Africa.” *Science* 350 (6262): 820–22. https://doi.org/10.1126/science.aad2879.

Loh, Po-Ru, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K Pickrell, David Reich, and Bonnie Berger. 2013. “Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium.” *Genetics* 193 (4): 1233–54. https://doi.org/10.1534/genetics.112.147330.

Loh, Po-Ru, Pier Francesco Palamara, and Alkes L Price. 2016. “Fast and Accurate Long-Range Phasing in a UK Biobank Cohort.” *Nature Genetics* 48 (7): 811–16. https://doi.org/10.1038/ng.3571.

Manrai, Arjun K., Birgit H. Funke, Heidi L. Rehm, Morten S. Olesen, Bradley A. Maron, Peter Szolovits, David M. Margulies, Joseph Loscalzo, and Isaac S. Kohane. 2016. “Genetic Misdiagnoses and the Potential for Health Disparities.” *New England Journal of Medicine* 375 (7): 655–65. https://doi.org/10.1056/NEJMsa1507092.

Maples, Brian K, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. 2013. “RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference.” *American Journal of Human Genetics* 93 (2): 278–88. https://doi.org/10.1016/j.ajhg.2013.06.020.

Marnetto, Davide, Vasili Pankratov, Mayukh Mondal, Francesco Montinaro, Katri Pärna, Leonardo Vallini, Ludovica Molinaro, et al. 2022. “Ancestral Genomic Contributions to Complex Traits in Contemporary Europeans.” *Current Biology* 32 (6): 1412–1419.e3. https://doi.org/10.1016/j.cub.2022.01.046.

Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2017. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.” *The American Journal of Human Genetics* 100 (4): 635–49. https://doi.org/10.1016/j.ajhg.2017.03.004.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. “Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities.” *Nature Genetics* 51 (4): 584–91. https://doi.org/10.1038/s41588-019-0379-x.

Martin, Alicia R., Konrad J. Karczewski, Sini Kerminen, Mitja I. Kurki, Antti-Pekka Sarin, Mykyta Artomov, Johan G. Eriksson, et al. 2018. “Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland.” *The American Journal of Human Genetics* 102 (5): 760–75. https://doi.org/10.1016/j.ajhg.2018.03.003.

Mathieson, Iain, Songül Alpaslan-Roodenberg, Cosimo Posth, Anna Szécsényi-Nagy, Nadin Rohland, Swapan Mallick, Iñigo Olalde, et al. 2018. “The Genomic History of Southeastern Europe.” *Nature* 555 (7695): 197–203. https://doi.org/10.1038/nature25778.

Mathieson, Iain, and Gil McVean. 2012. “Differential Confounding of Rare and Common Variants in Spatially Structured Populations.” *Nature Genetics* 44 (3): 243–46. https://doi.org/10.1038/ng.1074.

Mathieson, Iain, and Aylwyn Scally. 2020. “What Is Ancestry?” *PLOS Genetics* 16 (3): e1008624. https://doi.org/10.1371/journal.pgen.1008624.

McVean, G. 2009a. “A Genealogical Interpretation of Principal Components Analysis.” *PLoS Genet* 5 (10): e1000686. https://doi.org/10.1371/journal.pgen.1000686.

McVean, Gil, and Jerome Kelleher. 2019. “Linkage Disequilibrium, Recombination and Haplotype Structure.” In *Handbook of Statistical Genomics*, 51–86. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119487845.ch2.

Menozzi P., Piazza A., and Cavalli-Sforza L. 1978. “Synthetic Maps of Human Gene Frequencies in Europeans.” *Science* 201 (4358): 786–92. https://doi.org/10.1126/science.356262.

Moorjani, Priya, Nick Patterson, Joel N. Hirschhorn, Alon Keinan, Li Hao, Gil Atzmon, Edward Burns, Harry Ostrer, Alkes L. Price, and David Reich. 2011. “The History of African Gene Flow into Southern Europeans, Levantines, and Jews.” *PLOS Genetics* 7 (4): e1001373. https://doi.org/10.1371/journal.pgen.1001373.

Moorjani, Priya, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po-Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh. 2013. “Genetic Evidence for Recent Population Mixture in India.” *The American Journal of Human Genetics* 93 (3): 422–38. https://doi.org/10.1016/j.ajhg.2013.07.006.

Nachman, Michael W. 2002. “Variation in Recombination Rate across the Genome: Evidence and Implications.” *Current Opinion in Genetics & Development* 12 (6): 657–63. https://doi.org/10.1016/S0959-437X(02)00358-1.

Nagai, Akiko, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, et al. 2017. “Overview of the BioBank Japan Project: Study Design and Profile.” *SI: Overview of BBJ Cohort* 27 (3, Supplement): S2–8. https://doi.org/10.1016/j.je.2016.12.005.

Narasimhan, Vagheesh M, Nick Patterson, Priya Moorjani, Nadin Rohland, Rebecca Bernardos, Swapan Mallick, Iosif Lazaridis, et al. 2019. “The Formation of Human Populations in South and Central Asia.” *Science (New York, N.Y.)* 365 (6457): eaat7487. https://doi.org/10.1126/science.aat7487.

Need, Anna C., and David B. Goldstein. 2009. “Next Generation Disparities in Human Genomics: Concerns and Remedies.” *Trends in Genetics : TIG* 25 (11): 489–94. https://doi.org/10.1016/j.tig.2009.09.012.

Nielsen, Rasmus, Joshua M Akey, Mattias Jakobsson, Jonathan K Pritchard, Sarah Tishkoff, and Eske Willerslev. 2017. “Tracing the Peopling of the World through Genomics.” *Nature* 541 (7637): 302–10. https://doi.org/10.1038/nature21347.

Novembre, John. 2016. “Pritchard, Stephens, and Donnelly on Population Structure.” *Genetics* 204 (2): 391–93. https://doi.org/10.1534/genetics.116.195164.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, R Adam, Adam Auton, Amit Indap, et al. 2009. “Genes Mirror Geography within Europe” 456 (7218): 98–101. https://doi.org/10.1038/nature07331.Genes.

Novembre, John, and Benjamin M Peter. 2016. “Recent Advances in the Study of Fine-Scale Population Structure in Humans.” *Current Opinion in Genetics & Development* 41: 98–105. https://doi.org/10.1016/j.gde.2016.08.007.

Novembre, John, and Matthew Stephens. 2008. “Interpreting Principal Component Analyses of Spatial Population Genetic Variation.” *Nature Genetics* 40 (5): 646–49. https://doi.org/10.1038/ng.139.

Omberg, Larsson, Jacqueline Salit, Neil Hackett, Jennifer Fuller, Rebecca Matthew, Lotfi Chouchane, Juan L Rodriguez-Flores, Carlos Bustamante, Ronald G Crystal, and Jason G Mezey. 2012. “Inferring Genome-Wide Patterns of Admixture in Qataris Using Fifty-Five Ancestral Populations.” *BMC Genetics* 13 (June): 49–49. https://doi.org/10.1186/1471-2156-13-49.

Ongaro, Linda, Marilia O. Scliar, Rodrigo Flores, Alessandro Raveane, Davide Marnetto, Stefania Sarno, Guido A. Gnecchi-Ruscone, et al. 2019. “The Genomic Impact of European Colonization of the Americas.” *Current Biology* 29 (23): 3974-3986.e4. https://doi.org/10.1016/j.cub.2019.09.076.

Pagani, Luca, Toomas Kivisild, Ayele Tarekegn, Rosemary Ekong, Chris Plaster, Irene Gallego Romero, Qasim Ayub, et al. 2012. “Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool.” *The American Journal of Human Genetics* 91 (1): 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015.

Pagani, Luca, Stephan Schiffels, Deepti Gurdasani, Petr Danecek, Aylwyn Scally, Yuan Chen, Yali Xue, et al. 2015. “Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians.” *The American Journal of Human Genetics* 96 (6): 986–91. https://doi.org/10.1016/j.ajhg.2015.04.019.

Pankratov, Vasili, Francesco Montinaro, Alena Kushniarevich, Georgi Hudjashov, Flora Jay, Lauri Saag, Rodrigo Flores, et al. 2020. “Differences in Local Population History at the Finest Level: The Case of the Estonian Population.” *European Journal of Human Genetics* 28 (11): 1580–91. https://doi.org/10.1038/s41431-020-0699-4.

Pasaniuc, Bogdan, Noah Zaitlen, Guillaume Lettre, Gary K. Chen, Arti Tandon, W. H. Linda Kao, Ingo Ruczinski, et al. 2011. "Enhanced Statistical Tests for GWAS in Admixed Populations: Assessment Using African Americans from CARe and a Breast Cancer Consortium." *PLOS Genetics* 7 (4): 1–15. https://doi.org/10.1371/journal.pgen.1001371.

Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. "Ancient Admixture in Human History." *Genetics* 192 (3): 1065–93. https://doi.org/10.1534/genetics.112.145037.

Patterson, Nick, Alkes L Price, and David Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics* 2 (12): e190. https://doi.org/10.1371/journal.pgen.0020190.

Pearce, Mark. 2019. "The 'Copper Age'—A History of the Concept." *Journal of World Prehistory* 32 (3): 229–50. https://doi.org/10.1007/s10963-019-09134-z.

Pearson, Mike Parker, Ros Cleal, Peter Marshall, Stuart Needham, Josh Pollard, Colin Richards, Clive Ruggles, et al. 2007. "The Age of Stonehenge." *Antiquity* 81 (313): 617–39. https://doi.org/10.1017/S0003598X00095624.

Peter, Benjamin M. 2016. "Admixture, Population Structure, and f-Statistics." *Genetics* 202 (4): 1485–1501. https://doi.org/10.1534/genetics.115.183913.

Petrovski, Slavé, and David B. Goldstein. 2016. "Unequal Representation of Genetic Variation across Ancestry Groups Creates Healthcare Inequality in the Application of Precision Medicine." *Genome Biology* 17 (1): 157. https://doi.org/10.1186/s13059-016-1016-y.

Pickrell, Joseph K., Nick Patterson, Po-Ru Loh, Mark Lipson, Bonnie Berger, Mark Stoneking, Brigitte Pakendorf, and David Reich. 2014. "Ancient West Eurasian Ancestry in Southern and Eastern Africa." *Proceedings of the National Academy of Sciences* 111 (7): 2632–37. https://doi.org/10.1073/pnas.1313787111.

Price, Alkes L, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. 2009. "Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations." *PLoS Genetics* 5 (6): e1000519–e1000519. https://doi.org/10.1371/journal.pgen.1000519.

Pritchard, Jonathan K., and Molly Przeworski. 2001. "Linkage Disequilibrium in Humans: Models and Data." *The American Journal of Human Genetics* 69 (1): 1–14. https://doi.org/10.1086/321275.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–59. https://doi.org/10.1093/genetics/155.2.945.

Racimo, Fernando, Sriram Sankararaman, Rasmus Nielsen, and Emilia Huerta-Sánchez. 2015. "Evidence for Archaic Adaptive Introgression in Humans." *Nature Reviews. Genetics* 16 (6): 359–71. https://doi.org/10.1038/nrg3936.

Ralph, Peter, and Graham Coop. 2013. "The Geography of Recent Genetic Ancestry across Europe." *PLOS Biology* 11 (5): 1–20. https://doi.org/10.1371/journal.pbio.1001555.

Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh. 2009. "Reconstructing Indian Population History." *Nature* 461 (7263): 489–94. https://doi.org/10.1038/nature08365.

Reisberg, Sulev, Tatjana Iljasenko, Kristi Läll, Krista Fischer, and Jaak Vilo. 2017. "Comparing Distributions of Polygenic Risk Scores of Type 2 Diabetes and Coronary Heart Disease within Different Populations." *PLOS ONE* 12 (7): 1–9. https://doi.org/10.1371/journal.pone.0179238.

Rybicki, Benjamin A., Sudha K. Iyengar, Trent Harris, Rachael Liptak, Robert C. Elston, Mary J. Maliarik, and Michael C. Iannuzzi. 2002. "Prospects of Admixture Linkage Disequilibrium Mapping in the African-American Genome." *Cytometry* 47 (1): 63–65. https://doi.org/10.1002/cyto.10036.

Saint Pierre, Aude, Joanna Giemza, Isabel Alves, Matilde Karakachoff, Marinna Gaudin, Philippe Amouyel, Jean-François Dartigues, et al. 2020. "The Genetic History of France." *European Journal of Human Genetics* 28 (7): 853–65. https://doi.org/10.1038/s41431-020-0584-1.

Salter-Townshend, Michael, and Simon Myers. 2019. "Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups." *Genetics* 212 (3): 869–89. https://doi.org/10.1534/genetics.119.302139.

Sankararaman, S, S Mallick, M Dannemann, K Prufer, J Kelso, S Paabo, N Patterson, and D Reich. 2014. "The Genomic Landscape of Neanderthal Ancestry in Present-Day Humans." *Nature* 507 (7492): 354–57. https://doi.org/10.1038/nature12961.

Sankararaman, Sriram, Gad Kimmel, Eran Halperin, and Michael I Jordan. 2008. "On the Inference of Ancestries in Admixed Populations." *Genome Research* 18 (4): 668–75. https://doi.org/10.1101/gr.072751.107.

Schraiber, Joshua G, and Joshua M Akey. 2015. "Methods and Models for Unravelling Human Evolutionary History." *Nature Publishing Group*, no. November. https://doi.org/10.1038/nrg4005.

Scutari, Marco, Ian Mackay, and David Balding. 2016. "Using Genetic Distance to Infer the Accuracy of Genomic Prediction." *PLOS Genetics* 12 (9): 1–19. https://doi.org/10.1371/journal.pgen.1006288.

Seidensticker, Dirk, Wannes Hubau, Dirk Verschuren, Cesar Fortes-Lima, Pierre de Maret, Carina M. Schlebusch, and Koen Bostoen. 2021. "Population Collapse in Congo Rainforest from 400 CE Urges Reassessment of the Bantu Expansion." *Science Advances* 7 (7): eabd8352. https://doi.org/10.1126/sciadv.abd8352.

Semino, Ornella, A. Silvana Santachiara-Benerecetti, Francesco Falaschi, L. Luca Cavalli-Sforza, and Peter A. Underhill. 2002. "Ethiopians and Khoisan Share the Deepest Clades of the Human Y-Chromosome Phylogeny." *American Journal of Human Genetics* 70 (1): 265–68. https://doi.org/10.1086/338306.

Sirugo, Giorgio, Scott M. Williams, and Sarah A. Tishkoff. 2019. "The Missing Diversity in Human Genetic Studies." *Cell* 177 (1): 26–31. https://doi.org/10.1016/j.cell.2019.02.048.

Skoglund, Pontus, Swapan Mallick, Maria Cátira Bortolini, Niru Chennagiri, Tábita Hünemeier, Maria Luiza Petzl-Erler, Francisco Mauro Salzano, Nick Patterson, and David Reich. 2015. "Genetic Evidence for Two Founding Populations of the Americas." *Nature* 525 (7567): 104–8. https://doi.org/10.1038/nature14895.

Slatkin, Montgomery. 2008. "Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future." *Nature Reviews Genetics* 9 (6): 477–85. https://doi.org/10.1038/nrg2361.

Stapley, Jessica, Philine G D Feulner, Susan E Johnston, Anna W Santure, and Carole M Smadja. 2017. "Recombination: The Good, the Bad and the Variable." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. https://doi.org/10.1098/rstb.2017.0279.

Stephens, Matthew, and Paul Scheet. 2005. “Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation.” *American Journal of Human Genetics* 76 (3): 449–62. https://doi.org/10.1086/428594.

Sundquist, Andreas, Eugene Fratkin, Chuong B Do, and Serafim Batzoglou. 2008. “Effect of Genetic Divergence in Identifying Ancestral Origin Using HAPAA.” *Genome Research* 18 (4): 676–82. https://doi.org/10.1101/gr.072850.107.

Szulc, Piotr, Malgorzata Bogdan, Florian Frommlet, and Hua Tang. 2017. “Joint Genotype- and Ancestry-Based Genome-Wide Association Studies in Admixed Populations.” *Genetic Epidemiology* 41 (6): 555–66. https://doi.org/10.1002/gepi.22056.

Tang, Hua, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. 2006. “Reconstructing Genetic Ancestry Blocks in Admixed Individuals.” *American Journal of Human Genetics* 79 (1): 1–12. https://doi.org/10.1086/504302.

The 1000 Genomes Project Consortium, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, James Stalker, Michael Quail, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. “Genome-Wide Association Studies.” *Nature Reviews Methods Primers* 1 (1): 59. https://doi.org/10.1038/s43586-021-00056-9.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. “10 Years of GWAS Discovery: Biology, Function, and Translation.” *American Journal of Human Genetics* 101 (1): 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

Wallach, Hanna M. 2004. “Conditional Random Fields: An Introduction.”, University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-04-21.

Wangkumhang, Pongsakorn, Matthew Greenfield, and Garrett Hellenthal. 2021. “An Efficient Method to Identify, Date and Describe Admixture Events Using Haplotype Information.” *BioRxiv*, January, 2021.08.12.455263. https://doi.org/10.1101/2021.08.12.455263.

Wangkumhang, Pongsakorn, and Garrett Hellenthal. 2018. “ScienceDirect Statistical Methods for Detecting Admixture.” *Current Opinion in Genetics & Development* 53: 121–27. https://doi.org/10.1016/j.gde.2018.08.002.

Wegmann, Daniel, and Christoph Leuenberger. 2019. “Statistical Modeling and Inference in Genetics.” In *Handbook of Statistical Genomics*, 1–50. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119487845.ch1.

Wray, Naomi R., Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard, and Peter M. Visscher. 2013. “Pitfalls of Predicting Complex Traits from SNPs.” *Nature Reviews. Genetics* 14 (7): 507–15. https://doi.org/10.1038/nrg3457.

Yelmen, Burak, Davide Marnetto, Ludovica Molinaro, Rodrigo Flores, Mayukh Mondal, and Luca Pagani. 2021. “Improving Selection Detection with Population Branch Statistic on Admixed Populations.” *Genome Biology and Evolution* 13 (4): evab039. https://doi.org/10.1093/gbe/evab039.

Yelmen, Burak, Mayukh Mondal, Davide Marnetto, Ajai K Pathak, Francesco Montinaro, Irene Gallego Romero, Toomas Kivisild, Mait Metspalu, and Luca Pagani. 2019. “Ancestry-Specific Analyses Reveal Differential Demographic Histories and Opposite Selective Pressures in Modern South Asian Populations.” *Molecular Biology and Evolution* 36 (8): 1628–42. https://doi.org/10.1093/molbev/msz037.

Yoon, Byung-Jun. 2009. “Hidden Markov Models and Their Applications in Biological Sequence Analysis.” *Current Genomics* 10 (6): 402–15. https://doi.org/10.2174/138920209789177575.

Zhou, Quan, Liang Zhao, and Yongtao Guan. 2016. “Strong Selection at MHC in Mexicans since Admixture.” *PLOS Genetics* 12 (2): e1005847. https://doi.org/10.1371/journal.pgen.1005847.

ACKNOWLEDGEMENTS

I am incredibly grateful for the amazing group of people at the Institute of Genomics. I do not believe I will ever be able to write words as big as the kindness and support I have encountered in the last years. So forgive me if I'll use a simple five-letter word to thank all of you, I hope you will understand the size of my gratitude between the lines.

First and foremost, I would like to thank my supervisors, Professor Metspalu, Professor Pagani and Dr Montinaro, none of this would have been possible without their guidance and support.

My deepest gratitude goes to Linda. There simply are no words that encompass the importance of your friendship and how essential your support has been. I am so grateful that our paths have crossed, I feel so lucky to call you my friend.

The past few years would not be the same without Saoni, whose kindness and empathy simply cannot be matched, and Mayukh, whose mix of brightness and kindness will inspire me for a long time.

Katri and Tina, for the support we keep showing each other since we started the PhD together. I want to thank Freddi, Vasili, Mathilde, Stefania, and Ajai, for making the office (and lunch room) days full of interesting discussions, fun and exciting. Thanks to Davide and Rodrigo, two humans with the amazing skill of making anything fun, even more fun. Many thanks to Merilin and Mariza, they never stop welcoming me with a smile, not even at the hundredth silly question.

Thanks to Hovik, Monika, Bayazit, Lena, Bianca, Anne-Mai and Erwan, for contributing in making me feel at home, so far away from home, with your friendliness and smiles. And to Juri for his stories and beautiful pictures.

I would like to thank Mari for her help with the Estonian translations, and Professor Triin Laisk for so kindly agreeing to be the internal reviewer.

I have no doubt that in the future I will think about my PhD with a bittersweet nostalgia thanks to all the people I have met along the way.

PUBLICATIONS

CURRICULUM VITAE

Name: Ludovica Molinaro
Date of birth: March 15, 1994
Nationality: Italian
Address: University of Tartu, Institute of Genomics, Riia 23b, 51010 Tartu, Estonia
E-mail: ludovica.molinaro@ut.ee

Education:

2018–2022	Doctoral studies, University of Tartu, Faculty of Science and Technology, Institute of Genomics, Chair of Evolutionary Biology
2016–2018	MSc, cum laude, University of Padova, Department of Biology, supervisor Professor Luca Pagani "Bronze Age echoes in modern human whole genome sequences from Northeast Africa"
2013–2016	BSc, University of Ferrara, Department of Biology, supervisor Professor Guido Barbujani "Comparing genetic and linguistic diversity in Eurasia"

Professional employment:

2018–2022	University of Tartu, Institute of Genomics, Junior Researcher

Teaching:

2019	Teaching assistant, University of Tartu, Bachelor in Science and Technology, Chair of Evolutionary Biology, Evolution and the Natural World course
2019–2020	Teaching assistant, University of Tartu, Communication of science

International Courses and Conferences:

2021	CodeInPlace, programming course held by University on Stanford
2019	Oral presentation at AAI conference, Padua, Italy
2019	Poster presentation at Centenary of Human Population Genetics, Moscow, Russia
2019	CodeRefinery, workshop on automated testing, software development and module code development, held by The Nordic e-Infrastructure, Tartu, Estonia
2018	Analyses of genotyping and sequencing data in medical and population genetics course, Copenhagen, Denmark

Awards

2020	Dora Pluss Scholarship
2019	Best Talk, Italian Association of Anthropologists Conference, Padua, Italy

2019 Best Talk, Institute of Genomics & Institute of Molecular and Cell Biology Conference, Tartu, Estonia
2018 Ermenegildo Zegna Scholarship

Publications

Santander, Cindy, **Ludovica, Molinaro,** Giacomo, Mutti, Felipe I, Martinez, Jacinto, Mathe et al. 2022, " Genomic variation in baboons from central Mozambique unveils complex evolutionary relationships with other *Papio* species", *BMC Ecology and Evolution, 44 (22).* https://doi.org/10.1186/s12862-022-01999-7

Aneli, Serena, Tina Saupe, Francesco Montinaro, Anu Solnik, **Ludovica Molinaro**, Cinzia Scaggion, Nicola Carrara, et al. 2022. "The Genetic Origin of Daunians and the Pan-Mediterranean Southern Italian Iron Age Context." *Molecular Biology and Evolution* 39 (2): msac014. https://doi.org/10.1093/molbev/msac014.

Marnetto, Davide, Vasili Pankratov, Mayukh Mondal, Francesco Montinaro, Katri Pärna, Leonardo Vallini, **Ludovica Molinaro**, et al. 2022 "Ancestral Genomic Contributions to Complex Traits in Contemporary Europeans." *Current Biology*. https://doi.org/10.1016/j.cub.2022.01.046.

Ongaro, Linda, **Ludovica Molinaro**, Rodrigo Flores, Davide Marnetto, Marco R. Capodiferro, Marta E. Alarcón-Riquelme, Andrés Moreno-Estrada, et al. 2021. "Evaluating the Impact of Sex-Biased Genetic Admixture in the Americas through the Analysis of Haplotype Data." *Genes* 12 (10). https://doi.org/10.3390/genes12101580.

Ongaro, Linda, Mayukh Mondal, Rodrigo Flores, Davide Marnetto, **Ludovica Molinaro**, Marta E Alarcón-Riquelme, Andrés Moreno-Estrada, et al. 2021. "Continental-Scale Genomic Analysis Suggests Shared Post-Admixture Adaptation in the Americas." *Human Molecular Genetics* 30 (22): 2123–34. https://doi.org/10.1093/hmg/ddab177.

Yelmen, Burak, Davide Marnetto, **Ludovica Molinaro**, Rodrigo Flores, Mayukh Mondal, and Luca Pagani. 2021. "Improving Selection Detection with Population Branch Statistic on Admixed Populations." *Genome Biology and Evolution* 13 (4): evab039. https://doi.org/10.1093/gbe/evab039.

Molinaro, Ludovica, Davide Marnetto, Mayukh Mondal, Linda Ongaro, Burak Yelmen, Daniel John Lawson, Francesco Montinaro, and Luca Pagani. 2021. "A Chromosome-Painting-Based Pipeline to Infer Local Ancestry under Limited Source Availability." *Genome Biology and Evolution* 13 (4): evab025. https://doi.org/10.1093/gbe/evab025.

Marnetto, Davide, Katri Pärna, Kristi Läll, **Ludovica Molinaro**, Francesco Montinaro, Toomas Haller, Mait Metspalu, Reedik Mägi, Krista Fischer, and Luca Pagani. 2020. "Ancestry Deconvolution and Partial Polygenic Score Can Improve Susceptibility Predictions in Recently Admixed Individuals." *Nature Communications* 11 (1): 1628. https://doi.org/10.1038/s41467-020-15464-w.

Molinaro, Ludovica, Francesco Montinaro, Burak Yelmen, Davide Marnetto, Doron M Behar, Toomas Kivisild, and Luca Pagani. 2019. "West Asian Sources of the Eurasian Component in Ethiopians: A Reassessment." *Scientific Reports* 9 (1): 18811. https://doi.org/10.1038/s41598-019-55344-y.

Molinaro, Ludovica, and Luca Pagani. 2018. "Human Evolutionary History of Eastern Africa." *Genetics of Human Origins* 53: 134–39. https://doi.org/10.1016/j.gde.2018.10.002.

ELULOOKIRJELDUS

Nimi: Ludovica Molinaro
Sünniaeg: March 15, 1994
Kontakt: Tartu Ülikool, Genoomika Instituut Riia 23b, 51010 Tartu, Eesti
E-post: ludovica.molinaro@ut.ee

Hariduskäik:

2018–2022 Doktoriõpe, Tartu Ülikool, loodus- ja täppisteaduste valdkond, genoomika instituut, evolutsioonilise bioloogia õppetool.

2016–2018 Magistrikraad, cum laude, Padova Ülikool, Bioloogia osakond, juhendaja Professor Luca Pagani *"Bronze Age echoes in modern human whole genome sequences from Northeast Africa"*

2013–2016 Bakalaureusekraad, Ferrara Ülikool, Bioloogia osakond, juhendaja Professor Guido Barbujani *"Comparing genetic and linguistic diversity in Eurasia"*

Töökogemus:

2018–2022 Tartu Ülikool, genoomika instituut, nooremteadur.

Õpetamiskogemus:

2019 Õppeülesannete täitja, Tartu Ülikool, loodusteaduste ja tehnoloogia rahvusvaheline bakalaureuseõppekava (Science & Technology), evolutsioonilise bioloogia õppetool, kursus Elusloodus ja evolutsioon

2019–2020 Abiõppejõud, Akadeemilise väljendusoskuse keskus, teaduse kommunikatsioon

Rahvusvahelised kursused ja konverentsid:

2021 CodeInPlace, programmeerimise kursus, Stanfordi Ülikool

2019 Suuline ettekanne, AAI konverents, Padua, Itaalia.

2019 Poster, 100 aastat inimese populatsioonigeneetikat, Moskva, Venemaa

2019 CodeRefinery, automaattestimise, tarkvaraarenduse ja mooduli-koodiarenduse töötuba, mille korraldas The Nordic e-Infrastructure, Tartu, Eesti

2018 Doktorantide suvekursus "Genotüpiseerimis- ja sekveneerimis-andmete analüüsid meditsiinilises ja populatsioonigeneetikas", Kopenhaagen, Taani.

Auhinnad:

2020 Dora Pluss Stipendium

2019 Parim suuline ettekanne, Itaalia antropoloogide ühingu konverents, Padova, Itaalia

2019 Parim suuline ettekanne, Molekulaar- ja rakubioloogia instituudi ja genoomika instituudi ühine aastakonverents, Eesti
2018 Ermenegildo Zegna Stipendium

Publikatsioonid:
Loetletud inglisekeelse CV rubriigis publikatsioonid (‘Publications’)

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida*. Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand**. The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro*. Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous grassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan**. Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.
41. **Sulev Ingerpuu**. Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.

42. **Veljo Kisand**. Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa**. Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa**. Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik**. Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo**. Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo**. Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots**. Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero**. Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees**. Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks**. Cholecystokinin (CCK) – induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild**. Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva**. Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna**. Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro**. Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane**. Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm**. Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg**. Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild**. The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu**. Studies of the TOL plasmid transcription factor XylS. Tartu, 2000, 88 p.
61. **Dina Lepik**. Modulation of viral DNA replication by tumor suppressor protein p53. Tartu, 2000, 106 p.
62. **Kai Vellak**. Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu, 2000, 122 p.

63. **Jonne Kotta**. Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu, 2000, 160 p.
64. **Georg Martin**. Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000, 139 p.
65. **Silvia Sepp**. Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira**. On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000, 96 p.
67. **Priit Zingel**. The role of planktonic ciliates in lake ecosystems. Tartu, 2001, 111 p.
68. **Tiit Teder**. Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu, 2001, 122 p.
69. **Hannes Kollist**. Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu, 2001, 80 p.
70. **Reet Marits**. Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu, 2001, 112 p.
71. **Vallo Tilgar**. Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002, 126 p.
72. **Rita Hõrak**. Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002, 108 p.
73. **Liina Eek-Piirsoo**. The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002, 74 p.
74. **Krõõt Aasamaa**. Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002, 110 p.
75. **Nele Ingerpuu**. Bryophyte diversity and vascular plants. Tartu, 2002, 112 p.
76. **Neeme Tõnisson**. Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002, 124 p.
77. **Margus Pensa**. Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003, 110 p.
78. **Asko Lõhmus**. Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003, 168 p.
79. **Viljar Jaks**. p53 – a switch in cellular circuit. Tartu, 2003, 160 p.
80. **Jaana Männik**. Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003, 140 p.
81. **Marek Sammul**. Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003, 159 p
82. **Ivar Ilves**. Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003, 89 p.
83. **Andres Männik**. Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003, 109 p.

84. **Ivika Ostonen**. Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003, 158 p.
85. **Gudrun Veldre**. Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003, 199 p.
86. **Ülo Väli**. The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004, 159 p.
87. **Aare Abroi**. The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004, 135 p.
88. **Tiina Kahre**. Cystic fibrosis in Estonia. Tartu, 2004, 116 p.
89. **Helen Orav-Kotta**. Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004, 117 p.
90. **Maarja Öpik**. Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004, 175 p.
91. **Kadri Tali**. Species structure of *Neotinea ustulata*. Tartu, 2004, 109 p.
92. **Kristiina Tambets**. Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004, 163 p.
93. **Arvi Jõers**. Regulation of p53-dependent transcription. Tartu, 2004, 103 p.
94. **Lilian Kadaja**. Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004, 103 p.
95. **Jaak Truu**. Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004, 128 p.
96. **Maire Peters**. Natural horizontal transfer of the *pheBA* operon. Tartu, 2004, 105 p.
97. **Ülo Maiväli**. Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004, 130 p.
98. **Merit Otsus**. Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004, 103 p.
99. **Mikk Heidemaa**. Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004, 167 p.
100. **Ilmar Tõnno**. The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004, 111 p.
101. **Lauri Saks**. Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004, 144 p.
102. **Siiri Rootsi**. Human Y-chromosomal variation in European populations. Tartu, 2004, 142 p.
103. **Eve Vedler**. Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover**. Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 126 p.
105. **Helen Udras**. Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005, 100 p.

106. **Ave Suija**. Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus**. Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips**. Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **Krista Kaasik**. Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005, 121 p.
110. **Juhan Javoiš**. The effects of experience on host acceptance in ovipositing moths. Tartu, 2005, 112 p.
111. **Tiina Sedman**. Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005, 103 p.
112. **Ruth Aguraiuja**. Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005, 112 p.
113. **Riho Teras**. Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 106 p.
114. **Mait Metspalu**. Through the course of prehistory in India: tracing the mtDNA trail. Tartu, 2005, 138 p.
115. **Elin Lõhmussaar**. The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006, 124 p.
116. **Priit Kupper**. Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006, 126 p.
117. **Heili Ilves**. Stress-induced transposition of Tn*4652* in *Pseudomonas Putida*. Tartu, 2006, 120 p.
118. **Silja Kuusk**. Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006, 126 p.
119. **Kersti Püssa**. Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006, 90 p.
120. **Lea Tummeleht**. Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006, 94 p.
121. **Toomas Esperk**. Larval instar as a key element of insect growth schedules. Tartu, 2006, 186 p.
122. **Harri Valdmann**. Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers**. Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli**. Gata3 and Gata2 in inner ear development. Tartu, 2007, 123 p.
125. **Kai Rünk**. Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007, 143 p.

126. **Aveliina Helm**. Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007, 89 p.
127. **Leho Tedersoo**. Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007, 233 p.
128. **Marko Mägi**. The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007, 135 p.
129. **Valeria Lulla**. Replication strategies and applications of Semliki Forest virus. Tartu, 2007, 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007, 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007, 171 p.
132. **Tatjana Krama**. Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007, 112 p.
133. **Signe Saumaa**. The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida*. Tartu, 2007, 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007, 96 p.
135. **Priit Kilgas**. Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007, 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007, 95 p.
137. **Kärt Padari**. Protein transduction mechanisms of transportans. Tartu, 2008, 128 p.
138. **Siiri-Lii Sandre**. Selective forces on larval colouration in a moth. Tartu, 2008, 125 p.
139. **Ülle Jõgar**. Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008, 99 p.
140. **Lauri Laanisto**. Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008, 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008, 105 p.
142. **Birgot Paavel**. Bio-optical properties of turbid lakes. Tartu, 2008, 175 p.
143. **Kaire Torn**. Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg**. Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd**. Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.
146. **Lauri Saag**. Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.

147. **Ulvi Karu**. Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm**. Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks**. Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu**. Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš**. Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko**. Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast**. Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats**. Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova**. The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik**. Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver**. Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja**. Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast**. Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak**. Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.
162. **Triinu Remmel**. Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe**. Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe**. Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.
165. **Liisa Metsamaa**. Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.

166. **Pille Säälik**. The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil**. Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik**. Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark**. Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap**. Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan**. Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe**. Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi**. Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Velda Lauringson**. Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts**. Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.
176. **Mari Nelis**. Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov**. Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster**. Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap**. Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar**. Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül**. Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.
182. **Arto Pulk**. Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü**. Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla**. Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.
185. **Gyaneshwer Chaubey**. The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.
186. **Katrin Kepp**. Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.

187. **Virve Sõber**. The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro**. The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold**. Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert**. Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu**. Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik**. ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.
193. **Siim Sõber**. Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper**. Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak**. Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo**. Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel**. Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus**. Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius**. Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv**. Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk**. Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.
202. **Arno Põllumäe**. Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht**. Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus**. Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.
206. **Kessy Abarenkov**. PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.

207. **Marina Grigorova**. Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar**. The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild**. Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag**. Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla**. Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve**. Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar**. Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp**. Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero**. Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram**. Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents**. Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.
221. **Katrin Männik**. Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous**. Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu**. Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.
224. **Nele Tamberg**. Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.

225. **Tõnu Esko**. Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula**. Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu**. Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem**. The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen**. Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv**. The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi**. Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais**. Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja**. Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis**. Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme**. Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla**. Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke**. Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan**. Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm**. Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.
240. **Liina Kangur**. High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik**. Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski**. The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja**. Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus**. Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.
245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.

246. **Mario Plaas**. Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov**. Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks**. Arbuscular mycorrhizal fungal diversity patterns in boreonemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina**. The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau**. Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg**. Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis**. Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja**. Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg**. Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon**. A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.
259. **Andrei Nikonov**. RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets**. Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and interannual patterns. Tartu, 2014, 115 p.
263. **Külli Lokko**. Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.

265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun**. Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick**. Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme**. The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm**. Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve**. Genetic diversity and taxonomy within the genus *Rhinanthus.* Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov**. The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook**. Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi**. Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu**. Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.
276. **Hedi Peterson**. Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.
277. **Priit Adler**. Analysis and visualisation of large scale microarray data, Tartu, 2015, 126 p.
278. **Aigar Niglas**. Effects of environmental factors on gas exchange in deciduous trees: focus on photosynthetic water-use efficiency. Tartu, 2015, 152 p.
279. **Silja Laht**. Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices. Tartu, 2015, 100 p.
280. **Martin Kesler**. Biological characteristics and restoration of Atlantic salmon *Salmo salar* populations in the Rivers of Northern Estonia. Tartu, 2015, 97 p.
281. **Pratyush Kumar Das**. Biochemical perspective on alphaviral nonstructural protein 2: a tale from multiple domains to enzymatic profiling. Tartu, 2015, 205 p
282. **Priit Palta**. Computational methods for DNA copy number detection. Tartu, 2015, 130 p.
283. **Julia Sidorenko**. Combating DNA damage and maintenance of genome integrity in pseudomonads. Tartu, 2015, 174 p.
284. **Anastasiia Kovtun-Kante**. Charophytes of Estonian inland and coastal waters: distribution and environmental preferences. Tartu, 2015, 97 p.

285. **Ly Lindman**. The ecology of protected butterfly species in Estonia. Tartu, 2015, 171 p.
286. **Jaanis Lodjak**. Association of Insulin-like Growth Factor I and Corticosterone with Nestling Growth and Fledging Success in Wild Passerines. Tartu, 2016, 113 p.
287. **Ann Kraut**. Conservation of Wood-Inhabiting Biodiversity – Semi-Natural Forests as an Opportunity. Tartu, 2016, 141 p.
288. **Tiit Örd**. Functions and regulation of the mammalian pseudokinase TRIB3. Tartu, 2016, 182. p.
289. **Kairi Käiro**. Biological Quality According to Macroinvertebrates in Streams of Estonia (Baltic Ecoregion of Europe): Effects of Human-induced Hydromorphological Changes. Tartu, 2016, 126 p.
290. **Leidi Laurimaa**. *Echinococcus multilocularis* and other zoonotic parasites in Estonian canids. Tartu, 2016, 144 p.
291. **Helerin Margus**. Characterization of cell-penetrating peptide/nucleic acid nanocomplexes and their cell-entry mechanisms. Tartu, 2016, 173 p.
292. **Kadri Runnel**. Fungal targets and tools for forest conservation. Tartu, 2016, 157 p.
293. **Urmo Võsa**. MicroRNAs in disease and health: aberrant regulation in lung cancer and association with genomic variation. Tartu, 2016, 163 p.
294. **Kristina Mäemets-Allas**. Studies on cell growth promoting AKT signaling pathway – a promising anti-cancer drug target. Tartu, 2016, 146 p.
295. **Janeli Viil**. Studies on cellular and molecular mechanisms that drive normal and regenerative processes in the liver and pathological processes in Dupuytren's contracture. Tartu, 2016, 175 p.
296. **Ene Kook**. Genetic diversity and evolution of *Pulmonaria angustifolia* L. and *Myosotis laxa sensu lato* (Boraginaceae). Tartu, 2016, 106 p.
297. **Kadri Peil**. RNA polymerase II-dependent transcription elongation in *Saccharomyces cerevisiae*. Tartu, 2016, 113 p.
298. **Katrin Ruisu**. The role of RIC8A in mouse development and its function in cell-matrix adhesion and actin cytoskeletal organisation. Tartu, 2016, 129 p.
299. **Janely Pae**. Translocation of cell-penetrating peptides across biological membranes and interactions with plasma membrane constituents. Tartu, 2016, 126 p.
300. **Argo Ronk**. Plant diversity patterns across Europe: observed and dark diversity. Tartu, 2016, 153 p.
301. **Kristiina Mark**. Diversification and species delimitation of lichenized fungi in selected groups of the family Parmeliaceae (Ascomycota). Tartu, 2016, 181 p.
302. **Jaak-Albert Metsoja**. Vegetation dynamics in floodplain meadows: influence of mowing and sediment application. Tartu, 2016, 140 p.
303. **Hedvig Tamman**. The GraTA toxin-antitoxin system of *Pseudomonas putida*: regulation and role in stress tolerance. Tartu, 2016, 154 p.

304. **Kadri Pärtel**. Application of ultrastructural and molecular data in the taxonomy of helotialean fungi. Tartu, 2016, 183 p.
305. **Maris Hindrikson**. Grey wolf (*Canis lupus*) populations in Estonia and Europe: genetic diversity, population structure and -processes, and hybridization between wolves and dogs. Tartu, 2016, 121 p.
306. **Polina Degtjarenko**. Impacts of alkaline dust pollution on biodiversity of plants and lichens: from communities to genetic diversity. Tartu, 2016, 126 p.
307. **Liina Pajusalu**. The effect of CO_2 enrichment on net photosynthesis of macrophytes in a brackish water environment. Tartu, 2016, 126 p.
308. **Stoyan Tankov**. Random walks in the stringent response. Tartu, 2016, 94 p.
309. **Liis Leitsalu**. Communicating genomic research results to population-based biobank participants. Tartu, 2016, 158 p.
310. **Richard Meitern**. Redox physiology of wild birds: validation and application of techniques for detecting oxidative stress. Tartu, 2016, 134 p.
311. **Kaie Lokk**. Comparative genome-wide DNA methylation studies of healthy human tissues and non-small cell lung cancer tissue. Tartu, 2016, 127 p.
312. **Mihhail Kurašin**. Processivity of cellulases and chitinases. Tartu, 2017, 132 p.
313. **Carmen Tali**. Scavenger receptors as a target for nucleic acid delivery with peptide vectors. Tartu, 2017, 155 p.
314. **Katarina Oganjan**. Distribution, feeding and habitat of benthic suspension feeders in a shallow coastal sea. Tartu, 2017, 132 p.
315. **Taavi Paal**. Immigration limitation of forest plants into wooded landscape corridors. Tartu, 2017, 145 p.
316. **Kadri Õunap**. The Williams-Beuren syndrome chromosome region protein WBSCR22 is a ribosome biogenesis factor. Tartu, 2017, 135 p.
317. **Riin Tamm**. In-depth analysis of factors affecting variability in thiopurine methyltransferase activity. Tartu, 2017, 170 p.
318. **Keiu Kask**. The role of RIC8A in the development and regulation of mouse nervous system. Tartu, 2017, 184 p.
319. **Tiia Möller**. Mapping and modelling of the spatial distribution of benthic macrovegetation in the NE Baltic Sea with a special focus on the eelgrass *Zostera marina* Linnaeus, 1753. Tartu, 2017, 162 p.
320. **Silva Kasela**. Genetic regulation of gene expression: detection of tissue- and cell type-specific effects. Tartu, 2017, 150 p.
321. **Karmen Süld**. Food habits, parasites and space use of the raccoon dog *Nyctereutes procyonoides*: the role of an alien species as a predator and vector of zoonotic diseases in Estonia. Tartu, 2017, p.
322. **Ragne Oja**. Consequences of supplementary feeding of wild boar – concern for ground-nesting birds and endoparasite infection. Tartu, 2017, 141 p.
323. **Riin Kont**. The acquisition of cellulose chain by a processive cellobiohydrolase. Tartu, 2017, 117 p.
324. **Liis Kasari**. Plant diversity of semi-natural grasslands: drivers, current status and conservation challenges. Tartu, 2017, 141 p.

325. **Sirgi Saar**. Belowground interactions: the roles of plant genetic relatedness, root exudation and soil legacies. Tartu, 2017, 113 p.
326. **Sten Anslan**. Molecular identification of Collembola and their fungal associates. Tartu, 2017, 125 p.
327. **Imre Taal**. Causes of variation in littoral fish communities of the Eastern Baltic Sea: from community structure to individual life histories. Tartu, 2017, 118 p.
328. **Jürgen Jalak**. Dissecting the Mechanism of Enzymatic Degradation of Cellulose Using Low Molecular Weight Model Substrates. Tartu, 2017, 137 p.
329. **Kairi Kiik**. Reproduction and behaviour of the endangered European mink (*Mustela lutreola*) in captivity. Tartu, 2018, 112 p.
330. **Ivan Kuprijanov**. Habitat use and trophic interactions of native and invasive predatory macroinvertebrates in the northern Baltic Sea. Tartu, 2018, 117 p.
331. **Hendrik Meister**. Evolutionary ecology of insect growth: from geographic patterns to biochemical trade-offs. Tartu, 2018, 147 p.
332. **Ilja Gaidutšik**. Irc3 is a mitochondrial branch migration enzyme in *Saccharomyces cerevisiae.* Tartu, 2018, 161 p.
333. **Lena Neuenkamp**. The dynamics of plant and arbuscular mycorrhizal fungal communities in grasslands under changing land use. Tartu, 2018, 241 p.
334. **Laura Kasak**. Genome structural variation modulating the placenta and pregnancy maintenance. Tartu, 2018, 181 p.
335. **Kersti Riibak**. Importance of dispersal limitation in determining dark diversity of plants across spatial scales. Tartu, 2018, 133 p.
336. **Liina Saar**. Dynamics of grassland plant diversity in changing landscapes. Tartu, 2018, 206 p.
337. **Hanna Ainelo**. Fis regulates *Pseudomonas putida* biofilm formation by controlling the expression of *lapA*. Tartu, 2018, 143 p.
338. **Natalia Pervjakova**. Genomic imprinting in complex traits. Tartu, 2018, 176 p.
339. **Andrio Lahesaare**. The role of global regulator Fis in regulating the expression of *lapF* and the hydrophobicity of soil bacterium *Pseudomonas putida*. Tartu, 2018, 124 p.
340. **Märt Roosaare**. *K*-mer based methods for the identification of bacteria and plasmids. Tartu, 2018, 117 p.
341. **Maria Abakumova**. The relationship between competitive behaviour and the frequency and identity of neighbours in temperate grassland plants. Tartu, 2018, 104 p.
342. **Margus Vilbas**. Biotic interactions affecting habitat use of myrmecophilous butterflies in Northern Europe. Tartu, 2018, 142 p.
343. **Liina Kinkar**. Global patterns of genetic diversity and phylogeography of *Echinococcus granulosus* sensu stricto – a tapeworm species of significant public health concern. Tartu, 2018, 147 p.

344. **Teivi Laurimäe**. Taxonomy and genetic diversity of zoonotic tapeworms in the species complex of *Echinococcus granulosus* sensu lato. Tartu, 2018, 143 p.
345. **Tatjana Jatsenko**. Role of translesion DNA polymerases in mutagenesis and DNA damage tolerance in Pseudomonads. Tartu, 2018, 216 p.
346. **Katrin Viigand**. Utilization of α-glucosidic sugars by *Ogataea* (*Hansenula*) *polymorpha*. Tartu, 2018, 148 p.
347. **Andres Ainelo**. Physiological effects of the *Pseudomonas putida* toxin grat. Tartu, 2018, 146 p.
348. **Killu Timm**. Effects of two genes (DRD4 and SERT) on great tit (*Parus major*) behaviour and reproductive traits. Tartu, 2018, 117 p.
349. **Petr Kohout**. Ecology of ericoid mycorrhizal fungi. Tartu, 2018, 184 p.
350. **Gristin Rohula-Okunev**. Effects of endogenous and environmental factors on night-time water flux in deciduous woody tree species. Tartu, 2018, 184 p.
351. **Jane Oja**. Temporal and spatial patterns of orchid mycorrhizal fungi in forest and grassland ecosystems. Tartu, 2018, 102 p.
352. **Janek Urvik**. Multidimensionality of aging in a long-lived seabird. Tartu, 2018, 135 p.
353. **Lisanna Schmidt**. Phenotypic and genetic differentiation in the hybridizing species pair *Carex flava* and *C. viridula* in geographically different regions. Tartu, 2018, 133 p.
354. **Monika Karmin**. Perspectives from human Y chromosome – phylogeny, population dynamics and founder events. Tartu, 2018, 168 p.
355. **Maris Alver**. Value of genomics for atherosclerotic cardiovascular disease risk prediction. Tartu, 2019, 148 p.
356. **Lehti Saag**. The prehistory of Estonia from a genetic perspective: new insights from ancient DNA. Tartu, 2019, 171 p.
357. **Mari-Liis Viljur**. Local and landscape effects on butterfly assemblages in managed forests. Tartu, 2019, 115 p.
358. **Ivan Kisly**. The pleiotropic functions of ribosomal proteins eL19 and eL24 in the budding yeast ribosome. Tartu, 2019, 170 p.
359. **Mikk Puustusmaa**. On the origin of papillomavirus proteins. Tartu, 2019, 152 p.
360. **Anneliis Peterson**. Benthic biodiversity in the north-eastern Baltic Sea: mapping methods, spatial patterns, and relations to environmental gradients. Tartu, 2019, 159 p.
361. **Erwan Pennarun**. Meandering along the mtDNA phylogeny; causerie and digression about what it can tell us about human migrations. Tartu, 2019, 162 p.
362. **Karin Ernits**. Levansucrase Lsc3 and endo-levanase BT1760: characterization and application for the synthesis of novel prebiotics. Tartu, 2019, 217 p.
363. **Sille Holm**. Comparative ecology of geometrid moths: in search of contrasts between a temperate and a tropical forest. Tartu, 2019, 135 p.

364. **Anne-Mai Ilumäe**. Genetic history of the Uralic-speaking peoples as seen through the paternal haplogroup N and autosomal variation of northern Eurasians. Tartu, 2019, 172 p.
365. **Anu Lepik**. Plant competitive behaviour: relationships with functional traits and soil processes. Tartu, 2019, 152 p.
366. **Kunter Tätte**. Towards an integrated view of escape decisions in birds under variable levels of predation risk. Tartu, 2020, 172 p.
367. **Kaarin Parts**. The impact of climate change on fine roots and root-associated microbial communities in birch and spruce forests. Tartu, 2020, 143 p.
368. **Viktorija Kukuškina**. Understanding the mechanisms of endometrial receptivity through integration of 'omics' data layers. Tartu, 2020, 169 p.
369. **Martti Vasar**. Developing a bioinformatics pipeline gDAT to analyse arbuscular mycorrhizal fungal communities using sequence data from different marker regions. Tartu, 2020, 193 p.
370. **Ott Kangur**. Nocturnal water relations and predawn water potential disequilibrium in temperate deciduous tree species. Tartu, 2020, 126 p.
371. **Helen Post**. Overview of the phylogeny and phylogeography of the Y-chromosomal haplogroup N in northern Eurasia and case studies of two linguistically exceptional populations of Europe – Hungarians and Kalmyks. Tartu, 2020, 143 p.
372. **Kristi Krebs**. Exploring the genetics of adverse events in pharmacotherapy using Biobanks and Electronic Health Records. Tartu, 2020, 151 p.
373. **Kärt Ukkivi**. Mutagenic effect of transcription and transcription-coupled repair factors in *Pseudomonas putida.* Tartu, 2020, 154 p.
374. **Elin Soomets**. Focal species in wetland restoration. Tartu, 2020, 137 p.
375. **Kadi Tilk**. Signals and responses of ColRS two-component system in *Pseudomonas putida.* Tartu, 2020, 133 p.
376. **Indrek Teino**. Studies on aryl hydrocarbon receptor in the mouse granulosa cell model. Tartu, 2020, 139 p.
377. **Maarja Vaikre**. The impact of forest drainage on macroinvertebrates and amphibians in small waterbodies and opportunities for cost-effective mitigation. Tartu, 2020, 132 p.
378. **Siim-Kaarel Sepp**. Soil eukaryotic community responses to land use and host identity. Tartu, 2020, 222 p.
379. **Eveli Otsing**. Tree species effects on fungal richness and community structure. Tartu, 2020, 152 p.
380. **Mari Pent**. Bacterial communities associated with fungal fruitbodies. Tartu, 2020, 144 p.
381. **Einar Kärgenberg**. Movement patterns of lithophilous migratory fish in free-flowing and fragmented rivers. Tartu, 2020, 167 p.
382. **Antti Matvere**. The studies on aryl hydrocarbon receptor in murine granulosa cells and human embryonic stem cells. Tartu, 2021, 163 p.
383. **Jhonny Capichoni Massante**. Phylogenetic structure of plant communities along environmental gradients: a macroecological and evolutionary approach. Tartu, 2021, 144 p.

384. **Ajai Kumar Pathak.** Delineating genetic ancestries of people of the Indus Valley, Parsis, Indian Jews and Tharu tribe. Tartu, 2021, 197 p.
385. **Tanel Vahter.** Arbuscular mycorrhizal fungal biodiversity for sustainable agroecosystems. Tartu, 2021, 191 p.
386. **Burak Yelmen.** Characterization of ancient Eurasian influences within modern human genomes. Tartu, 2021, 134 p.
387. **Linda Ongaro.** A genomic portrait of American populations. Tartu, 2021, 182 p.
388. **Kairi Raime.** The identification of plant DNA in metagenomic samples. Tartu, 2021, 108 p.
389. **Heli Einberg.** Non-linear and non-stationary relationships in the pelagic ecosystem of the Gulf of Riga (Baltic Sea). Tartu, 2021, 119 p.
390. **Mickaël Mathieu Pihain.** The evolutionary effect of phylogenetic neighbourhoods of trees on their resistance to herbivores and climatic stress. Tartu, 2022, 145 p.
391. **Annika Joy Meitern.** Impact of potassium ion content of xylem sap and of light conditions on the hydraulic properties of trees. Tartu, 2022, 132 p.
392. **Elise Joonas.** Evaluation of metal contaminant hazard on microalgae with environmentally relevant testing strategies. Tartu, 2022, 118 p.
393. **Kreete Lüll.** Investigating the relationships between human microbiome, host factors and female health. Tartu, 2022, 141 p.
394. **Triin Kaasiku.** A wader perspective to Boreal Baltic coastal grasslands: from habitat availability to breeding site selection and nest survival. Tartu, 2022, 141 p.
395. **Meeli Alber.** Impact of elevated atmospheric humidity on the structure of the water transport pathway in deciduous trees. Tartu, 2022, 170 p.