DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
364

ANNE-MAI ILUMÄE

Genetic history of the Uralic-speaking peoples as seen through the paternal haplogroup N and autosomal variation of northern Eurasians

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
364

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
364

ANNE-MAI ILUMÄE

Genetic history of the Uralic-speaking peoples as seen through the paternal haplogroup N and autosomal variation of northern Eurasians

UNIVERSITY OF TARTU
Press

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 the 9th of October 2019 by the Council of the Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu.

Supervisors: Kristiina Tambets, PhD; Vice Director and Senior Research Fellow of Population Genetics, Institute of Genomics, University of Tartu, Estonia

Siiri Rootsi, PhD; Senior Research Fellow of Population Genetics, Institute of Genomics, University of Tartu, Estonia

Richard Villems, Professor of Archaeogenetics, Chair of Evolutionary Biology, Institute of Molecular and Cell Biology, and Research Professor, Institute of Genomics, University of Tartu

Opponent: Beniamino Trombetta, PhD; Associate Professor of Genetics, Department of Biology and Biotechnology "C. Darwin", Sapienza University of Rome, Italy

Commencement: Room No 105, 23B Riia St, Tartu, on the 12th of November 2019 at 10:15.

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

ISSN 1024-6479
ISBN 978-9949-03-203-7 (print)
ISBN 978-9949-03-204-4 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ORIGINAL PUBLICATIONS

I Rootsi S, Zhivotovsky LA, Baldovic M, Kayser M, Kutuev IA, Khusainova R, Bermisheva MA, Gubina M, Fedorova SA, **Ilumäe AM**, Khusnutdinova EK, Voevoda MI, Osipova LP, Stoneking M, Lin AA, Ferak V, Parik J, Kivisild T, Underhill PA, Villems R. 2007. **A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe**. *European Journal of Human Genetics* 15(2):204–211. doi:10.1038/sj.ejhg.5201748

II **Ilumäe AM**, Reidla M, Chukhryaeva M, Järve M, Post H, Karmin M, Saag L, Agdzhoyan A, Kushniarevich A, Litvinov S, Ekomasova N, Tambets K, Metspalu E, Khusainova R, Yunusbayev B, Khusnutdinova EK, Osipova LP, Fedorova S, Utevska O, Koshel S, Balanovska E, Behar DM, Balanovsky O, Kivisild T, Underhill PA, Villems R, Rootsi S. 2016. **Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families**. *American Journal of Human Genetics* 99(1):163–173. doi:10.1016/j.ajhg.2016.05.025

III Tambets K, Yunusbayev B, Hudjashov G, **Ilumäe AM**, Rootsi S, Honkola T, Vesakoski O, Atkinson Q, Skoglund P, Kushniarevich A, Litvinov S, Reidla M, Metspalu E, Saag L, Rantanen T, Karmin M, Parik J, Zhadanov SI, Gubina M, Damba LD, Bermisheva M, Reisberg T, Dibirova K, Evseeva I, Nelis M, Klovins J, Metspalu A, Esko T, Balanovsky O, Balanovska E, Khusnutdinova EK, Osipova LP, Voevoda M, Villems R, Kivisild T, Metspalu M. 2018. **Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations.** *Genome Biology* 19(1):139. doi:10.1186/s13059-018-1522-1

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

Ref.I – performed laboratory work, participated in data analysis, interpretation of results and in preparation of the manuscript.

Ref.II – designed primer sequences, interpreted phylogenetic results, provided main figures and co-wrote the manuscript with other contributing authors.

Ref. III – performed experiments for Y-chromosomal data and admixture dating analysis, participated in interpretation of results and was involved in writing the manuscript.

ABBREVIATIONS

aDNA – ancient DNA
AMH – anti-Müllerian hormone
BCE – before the Common Era
bp – basepair
BP – before present
CE – Common Era
GATK – Genome Analysis Toolkit
HMM – hidden Markov model
IBD – identical by descent
ISOGG – International Society of Genetic Genealogy
KYA – thousand (kilo) years ago
LD – linkage disequilibrium
LGM – last glacial maximum
Mb – megabase
MRCA – most recent common ancestor
MSY – male specific region of Y chromosome
mtDNA – mitochondrial DNA
MY(A) – million years (ago)
N_e – effective population size
NGS – next generation sequencing
NRY – non-recombining region of Y chromosome
PAR – pseudoautosomal region of Y chromosome
PC(A) – principal component (analysis)
SNP – single nucleotide polymorphism
SRY – sex-determining region Y
STR – short tandem repeat
XAR/YAR – X/Y added region
XCR – X conserved region
XDG – X-degenerated region
XTR – X-transposed region
YBP – years before present

1. INTRODUCTION

The study of human genetic variation has continued to fascinate scientists and the general public alike. A scientific field which began by surveying a handful of protein polymorphisms is now able to produce data on hundreds of thousands of markers across the genomes of hundreds of individuals at a reasonable cost within the timeframe of days. This technological development has provided new input for traditional tools of population genetics – mitochondrial DNA (mtDNA) and Y chromosome – by enabling researchers to expose the true variation behind the main uniparental lineages, which was previously shadowed by low level of phylogenetic resolution caused by scarcity of informative polymorphisms. Recombination-free inheritance mode and its substantial length make the human Y chromosome suitable for reconstructing the sequential and time-resolved branching pattern of male lineages, each of them traceable to a single common paternal ancestor and interpretable within a geographic framework. However, the uniparental markers characterise only two genetic loci and may fail to capture the evolutionary history represented throughout the entire genome. It is thus becoming paramount to illuminate the extant autosomal variation with novel methodologies designed for extracting relevant information from the large amount of generated autosomal genotyping data. In complement to data on modern populations, additional layers of human demographic history are now being uncovered by the rapidly developing ancient DNA (aDNA) field.

The Uralic language family is one of the few linguistic families disrupting the homogeneity of the European linguistic landscape dominated by the Indo-European languages. Uralic languages are spoken by approximately 25 million people from populations inhabiting northeastern Europe, the Volga-Uralic region, Western Siberia and, somewhat exceptionally, central Europe. No archaeological consensus on connecting the wide spread of Uralic languages with any known material culture exists and it remains largely unknown whether the dispersal involved a migration of people or was mainly a cultural diffusion. Studies on the uniparental markers have provided contradicting results, with maternally transmitted DNA lineages determined by geography, whereas a large portion of paternal lineages are united though a common haplogroup (hg) N. Hg N reaches prominent frequencies across entire northern Eurasia whereas in the European context, it is primarily featured in the Uralic speakers, which justifies a closer inspection of an assumed connection between them.

The first aim of the present dissertation is to describe the distinctive features of the Y-chromosome in the population genetic framework, to provide a brief insight into research on autosomal genetic variation and give an overview of the current knowledge on the composition of paternal and autosomal gene pool of northern Eurasia. The second aim is to add to this knowledge new information on hg N obtained with a novel approach of initially constructing a time-resolved phylogenetic tree from full Y-chromosomal sequences and then describing the dispersal of revealed clades in a comprehensive set of populations from

northern Eurasia. This is complemented by autosomal genetic research into previously sparsely described Uralic speakers – major carriers of hg N – and an examination of admixture as a likely trail to the presence of a potential counterpart of the Y-chromosomal paternal connection on the level of the entire genome.

2. LITERATURE OVERVIEW

2.1. Overview of the human Y chromosome

2.1.1. Evolutionary development of the mammalian Y chromosome

The unique properties of the human Y chromosome stem from its evolutionary history and are a consequence of the emergence of genetic sex determination system in the mammalian lineage. Known alternatives to genetic determinism present in mammals include haplodiploid system discovered in some insect species, including honeybees (Hasselmann et al., 2008), and environmentally modulated sex determination, found in some species of reptiles and fish (Bachtrog et al., 2014).

Genetic sex determination systems and sex chromosomes have evolved independently many times and represent a fascinating example of evolutionary convergence. However, the processes leading to differentiated sex chromosomes are remarkably similar across different taxa. The consensus model for sex chromosome evolution suggests sex chromosomes originate from ordinary autosomes (Ohno, 1967) that acquire a sex-determining locus followed by recombination suppression between the two initially homologous chromosomes (Bachtrog et al., 2014; Abbott et al., 2017). This triggers a cascade of evolutionary processes, including Muller's Ratchet, genetic hitchhiking and background selection (Figure 1, Wright et al., 2016). Muller's Ratchet refers to the irreversible accumulation of deleterious mutations in finite non-recombining populations owing to stochastic effects. Recombination can recreate mutation-free chromosomes, whereas non-recombining chromosomes that lack back mutation are passed on entirely. In the case of genetic hitchhiking, a new beneficial mutation will simultaneously promote the positive selection of linked deleterious mutations. If the deleterious mutation load becomes too high, purifying selection will eliminate deleterious mutations along with the beneficial ones, causing the non-recombining chromosome to undergo comparatively less adaptive evolution. Eventually, it can become advantageous for the carrier of a non-recombining chromosome to inactivate its maladapted linked genes (Bachtrog, 2013).

Why recombination suppression occurs in the first place remains elusive. The most commonly accepted model predicts a recombination arrest at the loci tightly linked to the sex-determining gene (Wright et al., 2016). This promotes accumulation of sexually antagonistic alleles – mutations that benefit one sex and are detrimental to the other – around the sex-determining locus (Rice, 1987; Wright et al., 2016). Once such alleles begin to accumulate, they are likely to continue crossing over to the homologous chromosome with harmful effects to the opposite sex. An evolutionary approach to counter this would be sex-limited gene expression or recombination suppression near the sex-determining locus, essentially locking the sexually antagonistic genes in the preferred sex (Rice, 1996). However, empirical evidence for the sexual conflict theory as well as

some non-adaptive alternative models have remained inconclusive (Wright et al., 2016).

In mammals, the mechanism for recombination elimination is multiple sequential chromosomal inversions on the proto-sex chromosomes (Lahn and Page, 1999). However, regardless of the exact mechanisms, the degradation of sex-specific sex chromosome reaches a similar end-state for any differentiated sex chromosome: many genes are represented by two doses in one sex and a single dose in the other – the heterogametic sex that is determined by a single sex chromosome. Thus, a gene dosage compensation system is set to evolve. It can be partial and gene-specific or the whole X chromosome can be epigenetically silenced in the somatic cells of females (Graves, 2016b). The latter seems to be a unique approach limited to birth-giving therian mammals that carry the most known and well-studied XY/XX sex-determination system (Graves, 2016b).

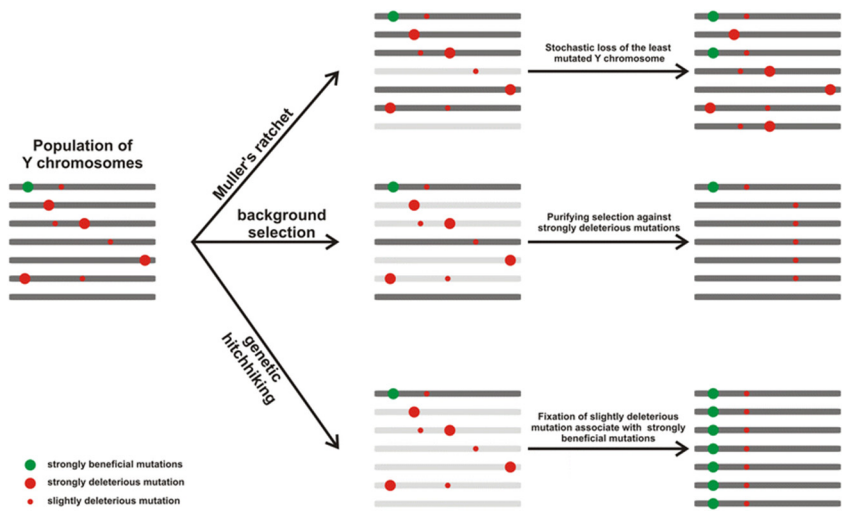

Figure 1. Evolutionary forces driving the degradation of the male-specific region of Y chromosome (MSY) in absence of recombination.
The figure shows a population of seven Y chromosomes that have undergone deleterious and beneficial mutations. In light grey colour: chromosomes that will be lost due to the action of different evolutionary forces. **Muller's ratchet** consists in the stochastic permanent loss of the least mutated chromosomes that cannot be restored within the population. **Background selection** refers to the elimination of strongly deleterious mutations, the reduction of the effective population size and the increase in frequency of slightly deleterious mutations by drift. **Genetic hitchhiking** in which strongly beneficial mutations drag along the initially associated deleterious mutations. Figure reprinted with permission of „Springer Berlin Heidelberg" from Figure 4 in Trombetta and Cruciani, (2017).

The most basal of the four mammalian evolutionary groups, Prototheria, comprises of the order Monotremata, which harbours the platypus and echidna families that both display a complex sex determination systems. The platypus has 10 sex chromosomes, five genetically distinct X that form a chain rather than a pair with five Y chromosomes during male meiosis (Figure 2, Wilson and Makova, 2009). The echidna has nine sex chromosomes with the terminal Y_5 fused onto another Y (Rens et al., 2007). None of the chromosomes show any homology with those of therian mammals (Veyrunes et al., 2008) – a clade that contains Metatheria (marsupials) and Eutheria (placental mammals). This suggests that the XY system present, for example, in humans, originated after the split of monotremes from therians approximately 200 million years ago (MYA) (Luo et al., 2011). The best candidate for monotreme sex determining gene is currently an *AMH* (anti-Müllerian hormone; also involved in sex determination pathway in some other vertebrate species) orthologue that lies on the smallest Y_5 (Cortez et al., 2014). Gene dosage compensation mechanism is currently unclear (Graves, 2016b).

Marsupials and placental mammals share the XX female/XY male sex chromosome system and a male-dominant testis-determining gene *SRY* (Sex-determining region Y), signifying a common origin for sex determination and an evolutionary development time of about 20 million years predating the split of these two lineages (Graves, 2016a). The evolutionary history of now heavily degraded chromosome Y has been retained on its conserved counterpart chromosome X. Comparison of marsupial and eutherian sex chromosomes has revealed added and ancient regions on the X and Y chromosomes (Graves, 1995). The X conserved region (XCR; YCR identified on human Y by mapping the X homologues of Y genes) is preserved in both clades, but is autosomal in monotremes (Figure 2) (Veyrunes et al., 2008). The X added region (XAR/YAR) is autosomal in marsupials, but on the sex chromosomes in all eutherian mammals, indicating a major fusion event within a potentially large pseudo-autosomal region before the X and Y complete differentiation, but after the marsupial split from placental mammals (Figure 2) (Murtagh et al., 2012; Graves, 2016a). Nearly all of the human Y genes derive from this acquired Y-chromosomal region (Waters et al., 2001).

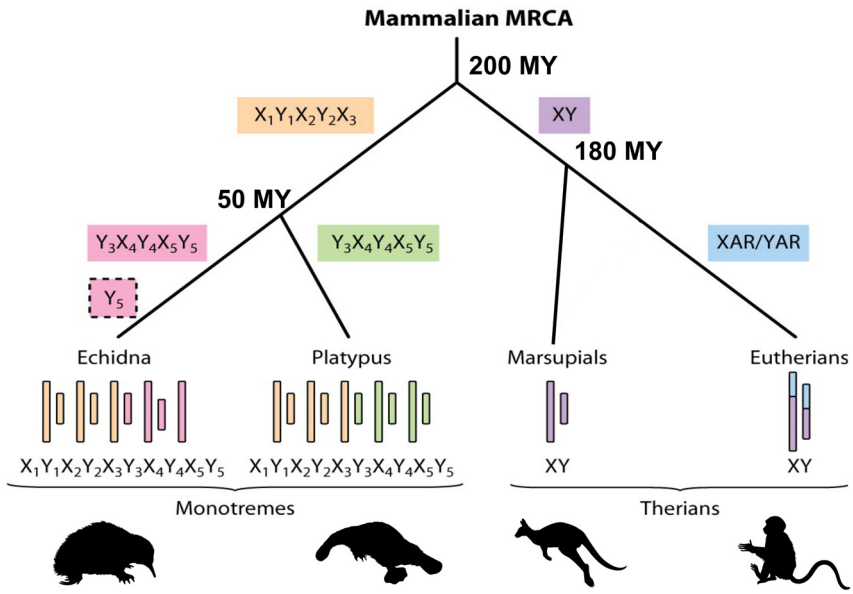

Figure 2. Schematic overview of mammalian sex chromosome evolution.
The figure shows the shared common ancestry of mammalian sex chromosomes in both lineages: monotremes and therians. Coloured boxes without a border identify additions, whereas boxes with a dotted border identify putative loss events. The monotreme and therian sex chromosomes are not homologous. Monotreme sex chromosomes are further distinguished from one another by independent additions along each lineage. Although marsupial and eutherian mammals share a common XY pair, the eutherian sex chromosomes are larger owing to a translocation to both X and Y of a sequence segment, denoted the X- or Y-added regions (XAR/YAR). The most recent common ancestor of mammals is indicated by the label "Mammalian MRCA". Approximate split times are given in million years (MY) according to Cortez et al., (2014). Figure republished and modified according to minor editing privileges with permission of „Annual Reviews, Inc" from Wilson and Makova, (2009); permission conveyed through Copyright Clearance Center, Inc. The vector shilouettes of mammals provided under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). Links to originals provided in the Web Resources section under References.

2.1.1.1. Characteristics of the human Y chromosome

In humans, the two terminal ends of Y chromosome carry pseudoautosomal regions (PARs) that recombine with homologous regions on the X chromosome during male meiosis, ensuring sequence identity between these parts (Rappold, 1993; Skaletsky et al., 2003). The two PARs aggregate about 3 megabases (Mb), whereas the rest of the roughly 57 Mb-sized human Y chromosome sequence (data according to Ensembl GRCh38 release 95) (Zerbino et al., 2017)

escapes recombination, which leads to diversification independent from the X chromosome and uniparental inheritance from father to son, thus being termed the male-specific region of the Y chromosome (MSY). Of this region, approximately half is a variably sized block of highly condensed transcriptionally inert heterochromatin and the remaining is transcriptionally active euchromatin composed of three major sequence classes (Figure 3) (Skaletsky et al., 2003; Jobling and Tyler-Smith, 2017; Skov et al., 2017):

1) the X-degenerate (XDG) region consists of sequence directly descended from the pair that became the sex chromosomes roughly 200 MY ago and that has retained 16 genes homologous to the X genes;
2) the X-transposed region (XTR) is unique to humans and originated by a duplication event from the X chromosome approximately 3–4 MY ago after the human lineage diverged from the human-chimpanzee common ancestor;
3) the ampliconic regions are all highly repetitive and contain large direct repeats and inverted repeats, including eight palindromes with some as large as 1.5 Mb and >99.9% similarity between arms. Genes in these regions tend to be present in multiple copies and are predominantly expressed in the testes.

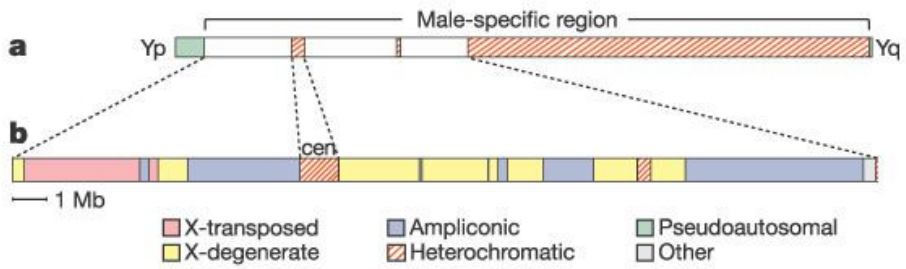

Figure 3. Schematic overview of the human Y chromosome.
a) Schematic representation of the whole chromosome, including the pseudoautosomal and heterochromatic regions**. b)** Enlarged view of a 24-Mb portion of the MSY, extending from the proximal boundary of the pseudoautosomal region on the short arm to the proximal boundary of the large heterochromatic region of the long arm. Shown are three classes of euchromatic sequences (X-transposed, X-degenerate and ampliconic), as well as heterochromatic sequences. A 1-Mb bar indicates the scale of the diagram. Figure republished with permission of „Springer Nature“ from Skaletsky et al., (2003); permission conveyed through Copyright Clearance Center, Inc.

As homologous genes are in a completely different order on the human X and Y chromosomes, Lahn and Page (1999) suggested that large Y-chromosomal inversions had regionally halted the recombination and caused independently diversified regions. Thus, suppression of recombination has occurred in a stepwise progressive manner – approximate sequence regions, termed “evolutionary

strata", stopped to recombine and consequently started to diverge around the same evolutionary time measured by the degree of nucleotide divergence (Lahn and Page, 1999). Additional research has supported this view for the formation of the human strata (Lemaitre et al., 2009). A total of five evolutionary strata have been identified on the human sex chromosomes (Ross et al., 2005). Strata 1–2 are the oldest and most diversified regions located on the XCR/YCR, with stratum 1 also holding the male-defining SRY gene which has been maintained in all therian lineages (Cortez et al., 2014). Strata 3–5 are located on the XAR/YAR and show progressively less X-Y divergence (Ross et al., 2005; Lemaitre et al., 2009; Cortez et al., 2014). Stepwise diversification of the sex chromosomes is not restricted to humans and similar strata have been discovered in species of rodents, birds, fish and plants (Sandstedt and Tucker, 2004; Bergero et al., 2007; Roesti et al., 2013; Wright et al., 2014).

The process of Y chromosome degradation has led to the idea of its eventual disappearance as exemplified by some XX/X0 rodent species with translocated Y-linked genes (Arakawa et al., 2002). However, comparative primate genomics has challenged this view. Comparison of human and rhesus macaque Y chromosome sequences has demonstrated relatively stable gene content maintained over the timespan of the last 25 million years. The oldest four strata are composed of identical genes in both human and the rhesus macaque and some gene loss is evident only from the newest fifth stratum (Hughes et al., 2012). This indicates a nonlinear degeneration that slows over time until stability level is reached (Hughes et al., 2012). Additionally, analytical simulation approaches have demonstrated that intrachromosomal gene conversion, a form of intrachromosomal recombination between palindromes located in the ampliconic regions, can act as an opposing force to Y degeneration, restoring the change-free state of a locus with a deleterious mutation (Rozen et al., 2003; Bachtrog, 2013). Such evidence suggests with high certainty that despite the seemingly unstable nature of its evolution, primate Y chromosome is not on the path to extinction.

2.1.2. Phylogenetic features of the human Y chromosome

The MSY is also known as the non-recombining region of Y chromosome (NRY). It excludes the pseudoautosomal regions, thus being constitutively haploid and confined to the male lineage. Male specificity means that the patterns of diversity accumulated in the MSY reflect male demography in the past and this feature has direct implications for genealogic and phylogenetic studies (Jobling and Tyler-Smith, 2003; Underhill and Kivisild, 2007; Calafell and Larmuseau, 2017). Additionally, MSY plays a defining role in male fertility (McElreavey et al., 2006; Krausz and Casamonti, 2017), has an effect on diseases driven by immune and inflammatory responses (Maan et al., 2017) and is actively used in forensic DNA analysis (Jobling et al., 1997; Kayser, 2017).

2.1.2.1. Uniparental and recombination-free inheritance of male-specific region of the Y chromosome

Absence of meiotic recombination allows the combinations of allelic states of genetic loci to be passed intact to the next generation. Exceptions include a few occasionally recombining loci located on the highly similar non-pseudoautosomal sequences of the X and Y chromosomes (Trombetta et al., 2010, 2014) and occurrence of gene conversion in the duplication-rich ampliconic regions (Rozen et al., 2003; Skaletsky et al., 2003; Skov et al., 2017). Gene conversion is considered inherently different from the conventional meiotic recombination defined as crossing over between two chromosomal homologues, but its abundance renders MSY, rather than NRY, a more appropriate designation for the region free of crossing over events (Jobling and Tyler-Smith, 2003, 2017; Skov et al., 2017). Consequently, despite some exceptions, the MSY can be considered as an intact single locus deprived of the genetic randomisation process. This and a number of additional features make MSY especially suitable for population genetic studies.

2.1.2.2. Sex-specific demography patterns

Assuming equal numbers of males and females in the population, the effective population size (N_e) of the haploid MSY is one-quarter of that of autosomes and one-third of that of the X chromosomes (Jobling and Tyler-Smith, 2003). This makes Y chromosome susceptible to genetic drift, which can be further increased by founder effects endorsed by a variety of sociocultural processes, such as small numbers of men with social prestige benefitting from a greater number of offspring (Zerjal et al., 2003; Xue et al., 2005; Karmin et al., 2015; Poznik et al., 2016) or competition between patrilineal kinship groups (Zeng et al., 2018). Drift accelerates the genetic differentiation of populations and facilitates rapid temporal change of haplotype frequencies.

It has been suggested that about 70% of human populations adhere to patrilocality (Burton et al., 1996) – the custom of wives settling in the same residential area as their husbands' families. This would result in higher level of genetic differentiation between populations for the Y chromosome than for maternally transmitted mtDNA (Heyer et al., 2012). Initial comparisons of MSY and mtDNA variation in matrilocal and patrilocal tribal populations supported this hypothesis (Seielstad et al., 1998; Oota et al., 2001), whereas later research revealed that sex-specific features are often unique to a population and the higher migration rate among patrilocal females does not have a significant effect on the global-scale human variation patterns (Wilder et al., 2004; Balaresque et al., 2006). Latest research utilizing a comprehensive set of global samples, substantially longer stretches of unique MSY and full mtDNA sequences concluded that at the global scale, differences in Y chromosome variation among populations are indeed bigger, although on a lesser level than

previously proposed and with significant regional differences (Lippold et al., 2014).

Long-distance migrations produce an effect opposite of patrilocality and have been predominantly attributed to the male sex. The European expansions during the colonial era have contributed to discordant geographic origins of some populations' maternal and paternal gene pools. Strong introgression of European Y-chromosomal lineages and the retention of indigenous maternal mtDNA lineages have been recorded in the gene pools of Polynesia (Hurles et al., 1998), Greenland (Bosch et al., 2003) and the Americas (Carvajal-Carmona et al., 2000; Abe-Sandes et al., 2004; Hammer et al., 2006a). In Indonesia, historic contacts from Chinese, Indians, Arabs and Europeans form a noticeable fraction of Y chromosome variation, but are not reflected in mtDNA (Tumonggor et al., 2013).

2.1.2.3. Types of genetic polymorphisms

Available genetic polymorphisms for assessing Y-chromosomal global diversity can be broadly categorized as a) biallelic markers and b) multiallelic tandem repeats. The basis for such categorization lays in their mutation rate. The first of the two categories – the slowly mutating biallelic markers, comprised of single nucleotide polymorphisms (SNPs) and certain insertion-deletions, are considered to be unique or near-unique events in human prehistory each representing a monophyletic lineage founded by one man. The second category includes much faster mutating multiallelic microsatellites also known as short tandem repeats (STRs). Ancestral states of unique biallelic markers are determined by comparison of homologous regions in an appropriate outgroup such as the chimpanzee (Hurles and Jobling, 2001). Despite their very low mutation rate, resequencing of hundreds of entire MSY regions has led to the discovery of many recurrent SNP markers that deviate from the uniqueness requirement, appearing on different branches of the Y chromosome tree (The Y Chromosome Consortium, 2002; Karafet et al., 2008; Hallast et al., 2015; Karmin et al., 2015; Poznik et al., 2016). Nevertheless, this does not impair the construction of a robust evolutionary tree as many thousands of additional and reliably unique markers are currently available to support the tree's branching pattern.

The rapidly mutating multiallelic STR markers differ by allele numbers and number of nucleotides in the repetition unit. They have been used to investigate the diversity within a particular Y chromosome haplogroup (see 2.1.3) and provide a possible age for the haplogroup-defining mutation and are part of suspect identification and paternity tests in DNA forensics (de Knijff, 2000; Jobling and Tyler-Smith, 2003; Kayser, 2017). However, the manifold discrepancies between estimated mutation rates proposed in many previous studies have limited the credibility of STR-based dating (Busby et al., 2012; Wei et al., 2013b; Hallast et al., 2015).

2.1.2.4. Caveats in uniparental interpretation of population history

The cost-effectiveness of next-generation sequencing (NGS) platforms has provided a wealth of Y-chromosomal SNP data, which has resulted in reliable NGS-based evolutionary trees constructed with advanced methods and a renewed interest in the phylogeny and **phylogeography** of the Y chromosome – the study of spatial distribution of different haplogroups within a phylogeny (Avise et al., 1987). However, population genetic theory states that a single locus evolutionary tree is a random realization of a stochastic process and depends exclusively on relatedness level of sampled individuals, thus cautioning against over interpretation and necessitating a shift from describing to testing hypotheses about the processes underlying the discovered patterns of genetic variation (Avise, 2000; Edwards and Beerli, 2000; Knowles, 2003; Nielsen and Beaumont, 2009). Coalescent theory first developed by Kingman (1982) and its succeeding theoretical improvements provide a mathematical framework with a focus on the distribution of gene trees in populations, thus bridging the divide between statistical population genetics and primarily tree-based descriptive phylogenetics (Hey and Machado, 2003; Nielsen and Beaumont, 2009). Coalescent-based methods connect a demographic model with a tree, but must account for the uncertainty and stochasticity of genetic trees in populations. In essence, the coalescent approach is based on forming a likelihood function that integrates the probability of data given the tree with the probability of tree given the parameters of the demographic model and considering the set of all possible trees (Nielsen and Beaumont, 2009). This function is not directly solvable with known methods, so simulation approaches are used for approximation, followed by maximum likelihood or Bayesian estimates of the demographic parameters of interest. Such methods are complex and computationally expensive, but affordable generation of data along with falling costs of computing power has facilitated the use and ongoing development of user-friendly software implementing the coalescence-based approaches. For a haploid locus such as the human Y chromosome, congruent reconstruction of global phylogeny based on full sequences has been achieved with both maximum likelihood (Poznik et al., 2016) and Bayesian methods (Karmin et al., 2015), complemented by investigations into historical fluctuation patterns of male and female effective population sizes (Lippold et al., 2014; Karmin et al., 2015) and testing discrete phylogeographic models by Bayesian and maximum parsimony approaches (Scozzari et al., 2014). Establishing novel approaches with software support to discern the information embedded in full genomes (Kelleher et al., 2019; Speidel et al., 2019) is currently one of the most central development trends in modern genomics.

2.1.3. Nomenclature of Y-chromosomal haplogroups in humans

The phylogeny of Y chromosomes has expanded from a single marker typed in 1985 (Casanova et al., 1985) to several thousands of binary polymorphisms discovered 30 years later (Scozzari et al., 2014; Hallast et al., 2015; Karmin et al., 2015; Poznik et al., 2016; Finocchio et al., 2018). New polymorphisms that define additional sublineages are constantly uncovered and a complex and expandable nomenclature needs to be maintained in order to accommodate the growing phylogeny. An initially sufficient and flexible naming system was proposed in 2002, describing the underlying phylogeny in a strictly cladistic manner (The Y Chromosome Consortium, 2002). A monophyletic clade, that constitutes a haplogroup (hg), unites all descendants of a single node that share the derived states of unique polymorphisms (Jobling et al., 2013) and the sequential accumulation of polymorphisms allows for haplogroups (clades) to be connected in a tree-like structure. Major haplogroups are identified by single capital letters beginning with the letter "A" and continuing through the alphabet, sublineages nested within these clades are given numerical suffixes (for example, hg R1 and R2) and this can be continued in alternating alphanumeric manner with lowercase letters and numbers until the most terminal branches have been named (The Y Chromosome Consortium, 2002; Jobling et al., 2013). Contrastingly, some Y chromosomes might share derived states of deep-rooted polymorphisms, belonging to a higher-order clade, but having ancestral states at the lower sublineages' level. Such chromosomes are potentially paraphyletic and new mutations might be discovered that separate them further into new sublineages. In Y nomenclature, sets of these Y chromosomes are called **paragroups** and are highlighted by a "*" (star) symbol (The Y Chromosome Consortium, 2002). Such nomenclature system also supports partial genotyping that does not cover all known variability with the "x" suffix, which signifies "not" (Jobling et al., 2013). For example, if only defining variants for hg R1 but not the R2 are typed, the corresponding paragroup is named R*(xR1) (Jobling et al., 2013).

In total five most parsimonious phylogenies unifying all revisions of hg topology along with defining SNPs and following the described nomenclature have been published (The Y Chromosome Consortium, 2002; Jobling and Tyler-Smith, 2003; Karafet et al., 2008; Van Geystelen et al., 2013; van Oven et al., 2014). The Y phylogeny roots in Africa, with hgs A and B restricted to African populations (Figure 4) (Hammer et al., 1998; Underhill et al., 2000; Poznik et al., 2016). The succeeding out of Africa migration was followed by a geographic expansion and diversification of all contemporary non-African hgs within the last 100 000 years (Hammer et al., 1998; Underhill et al., 2000; Poznik et al., 2016).

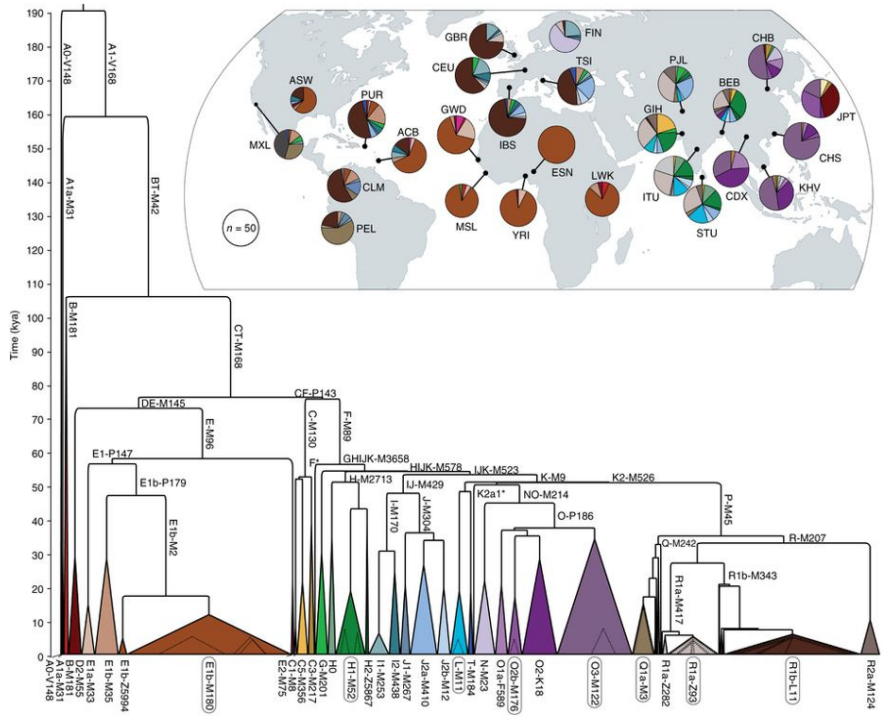

Figure 4. A calibrated Y chromosome phylogenetic tree based on 60,555 SNPs from 1244 present-day chromosomes from the 1000 Genomes Project.
Mutation rate of 0.76×10^{-9} basepair^{-1} year $^{-1}$ was used for calibration (Fu et al., 2014). The labels on the branches and below the triangles are haplogroup names in the form "haplogroup – key defining mutation". An asterisk indicates paragroup. Triangle width represents the frequency of the haplogroup in the sample and height coalescence time. Labels outlined in grey ovals indicate haplogroups that have undergone rapid recent expansions (Poznik et al., 2016). **Inset**, world map indicating, for each of the 26 populations, the geographic source, sample size, and major haplogroup distribution coloured according to the presented tree. Three-letter labels are abbreviated population code-names: ACB – African Caribbeans in Barbados; ASW – Americans of African ancestry in the south-west United States; BEB – Bengali from Bangladesh; CDX – Chinese Dai in Xishuangbanna, China; CEU – Utah Residents (Centre d'Etude du Polymorphisme Humain) with northern and western European ancestry; CHB – Han Chinese in Beijing, China; CHS – southern Han Chinese; CLM – Colombians from Medellin, Colombia; ESN – Esan in Nigeria; FIN – Finns in Finland; GBR – British in England and Scotland; GIH – Gujarati Indian from Texas, USA; GWD – Gambian in the Western Divisions in the Gambia; IBS – Iberian population in Spain; ITU – Indian Telugu from the United Kingdom; JPT – Japanese in Tokyo, Japan; KHV – Kinh in Ho Chi Minh City, Vietnam; LWK – Luhya in Webuye, Kenya; MSL – Mende in Sierra Leone; MXL – Mexican ancestry from Los Angeles, USA; PEL – Peruvians from Lima, Peru; PJL – Punjabi from Lahore, Pakistan; PUR – Puerto Ricans from Puerto Rico; STU – Sri Lankan Tamil from the United Kingdom; TSI – Tuscans from Italy; YRI – Yoruba in Ibadan, Nigeria. Reprinted by permission of "Springer Nature" from Poznik et al., (2016), permission conveyed through Copyright Clearance Center, Inc.

Since the advent of NGS technologies the number of Y-SNPs and their corresponding lineages has increased to the level that makes alphanumeric branch names too long to follow and thus infeasible for practical use (e.g., R1a1a1-b1a1a in hg R tree by International Society of Genetic Genealogy (ISOGG) Y-DNA tree 2019). An increasingly dominant substitute is to use the name of the most informative derived polymorphism in conjunction with its basal haplogroup name, proposed as a potential second naming convention by The Y Chromosome Consortium (2002). This approach was used in the most recent minimal reference version of the Y-chromosomal tree by van Oven et al.(2014). Authors of the latest publications that survey Y chromosome variation with NGS technology have adhered to this "short-hand" notation style (Hallast et al., 2015; Poznik et al., 2016; Finocchio et al., 2018). Karmin et al.(2015) proposed to incorporate time depth of respective haplogroups estimated from sequence data into the alphanumeric naming convention. The alternative naming approach will probably form the basis for the emerging nomenclature standard as the number of sequenced Y chromosome samples and thus the tree itself continue to grow.

2.1.4. Mutation rate of the human Y chromosome

One of the major goals of evolutionary research is to date the events that shaped the present-day phylogeny and phylogeography of a particular species. For this aim, the recognition of a correct mutation rate is paramount. In human Y-chromosomal research, there are three main approaches to estimate the mutation rate and all three of them require certain limitations and assumptions in order to obtain a credible mutation speed.

2.1.4.1. Genealogical mutation rate

The genealogical mutation rate consists of counting mutations in preferably deep-rooted pedigrees or in the simplest case, father-son pairs and dividing by the number or generations (per-generation mutation rate) or number of years (per-year mutation rate) (Figure 5A, Balanovsky, 2017). In literature, this can be called "pedigree mutation rate", "*de novo* mutation rate" or the "**genealogical mutation rate**".

The genealogical approach is the most straightforward one and in addition to a preferably large number of confirmed pedigrees requires sequencing technology able to detect polymorphisms with sufficient precision (Balanovsky, 2017). NGS tends to be error-prone, which makes high coverage and advanced filtering steps crucial in achieving a reliable result. However, the main disadvantage of genealogical mutation rate in the context of evolutionary studies is its indirect application, since it derives from an external dataset of scarcely available deep-rooted pedigrees. For chronological results, a conversion from generation to years is needed and this brings along questions on the generation time that is difficult to

estimate for the entire duration of human history. Presently, a consensus male generation time of 30 years seems to apply to a variety of populations and is supported by a genetic study of human generation time measured since Neanderthal admixture (Fenner, 2005; Moorjani et al., 2016).

A landmark study by Xue et al. (2009) examined a validated Chinese pedigree, which had two male members separated by 13 generations. With the exact birth year of the common ancestor known, the result yielded a mutation rate estimate of 1.00×10^{-9} bp^{-1} $year^{-1}$. The result compares well with the evolutionary mutation rates of 1.24 – 1.50 bp^{-1} $year^{-1}$ deduced from human-chimp comparisons (Table 1), but is limited by a wide confidence interval stemming from low sample size. A larger study by Helgason et al.(2015) investigated 753 Icelandic males and resulted in a mutation rate of 0.89×10^{-9} bp^{-1} $year^{-1}$ for X-degenerate region (Table 1). A statistically significant lower rate of 0.74×10^{-9} bp^{-1} $year^{-1}$ was found for the palindromic region and was explained by gene conversion between paralogous sequences correcting mutations back to their ancestral states (Helgason et al., 2015). A similar mutation rate was obtained by analysing members of a patrilineal Kazakh clan with a common male ancestor, whose date of birth was estimated from written records (Table 1) (Balanovsky et al., 2015).

2.1.4.2. Evolutionary mutation rate

Evolutionary or **phylogenetic** mutation rate involves identifying a chronologically known population founder event and a sufficiently diversified haplogroup associated with this population. Subsequently, the per-year mutation rate is estimated as the average number of mutations along all phylogenetic lineages, divided by the time since the peopling event (Figure 5B, Balanovsky, 2017). A somewhat different approach designated also as the evolutionary mutation rate involves the comparison of chimpanzee-human homologous genetic regions and division by the split time between species as in Thomson et al.(2000) and Kuroki et al.(2006).

Evolutionary approach produces mutation rate measured per basepair per year (bp^{-1} $year^{-1}$), thus avoiding the problematic generation time conversion, but is strongly dependent on the extracted variance, which is often incomplete, and temporal precision of the archaeological calibration point (Balanovsky, 2017). Molecular divergence caused by a demographic increase must coincide with the archaeological peopling date. In reality, the carriers of the lineages could have started to expand substantially earlier or later and this results in either over- or underestimated mutation rate. Applicability of interspecies evolutionary rate obtained through comparison of chimpanzee-human sequences raises concerns in the context of human demographic studies due to uncertainties of the temporal precision of the human-chimpanzee split (Wang et al., 2014a). Secondly, the Y chromosomes of the two species are structurally diverse which obscures precise alignment and exemplifies possibly different selective pressures acting upon the two lineages (Wang et al., 2014a).

Evolutionary mutation rate for Y chromosome was suggested by Poznik et al.(2013), who calibrated the split between two main branches of the Amerindian haplogroup Q-M3 against the archaeological estimate of the peopling of the Americas. Another study by Francalacci et al.(2013) examined the Sardinian-specific subhaplogroup I2a1a and calibrated the discovered variability in the X-degenerate region against the archaeological timeframe of the Neolithic peopling of Sardinia. The calculated mutation rates were 0.82×10^{-9} bp^{-1} $year^{-1}$ for the former and 0.53×10^{-9} bp^{-1} $year^{-1}$ for the latter study. These rates depend on the extracted variability and thus are expected to vary if more mutations are uncovered. In the case of Sardinian study, four of 1204 samples were sequenced with higher coverage (>13x), revealing additional SNPs and a faster rate of 0.65×10^{-9} bp^{-1} $year^{-1}$. The confidence intervals for the three calibration-based evolutionary mutation rates do not overlap (Table 1), demonstrating the challenges of connecting archaeological dates with phylogenetic events (Balanovsky, 2017).

An alternative genealogical approach was used by Mendez et al.(2013), who derived the Y-chromosomal mutation rate using paternal autosomal mutation rates reported on a dataset of parent-offspring trios. The result is one of the slowest of existing mutation rate estimates (Table 1). However, several methodological assumptions have been criticised, including the choice of unreasonable generation time and linear modelling of autosomal and Y-chromosomal mutation rates (Elhaik et al., 2014).

2.1.4.3. Ancient DNA based mutation rate

Ancient DNA (aDNA) based methodology has appeared rather recently in connection with the intensified sequencing of DNA obtained from ancient samples with reliable radiocarbon dates. Ancient DNA has had less time to accumulate mutations compared to modern samples and the number of such "missing mutations" divided by aDNA sample's age provides a per-year mutation rate estimate (Figure 5C).

The aDNA-based approach has the advantage of reliable radiocarbon dates of the ancient samples used as calibration points on the phylogenetic tree. It also provides a direct estimate of the mutation rate as ancient samples can be processed together with all samples without an estimation of the generation time (Francalacci et al., 2016). Currently, main difficulties with this approach lie in possible contamination and poor quality of the extracted DNA causing underestimation of variability due to low coverage sequencing (Balanovsky, 2017) .

Using high-coverage sequences from ancient samples to directly estimate the Y-chromosomal mutation rate was pioneered by Fu et al.(2014). Palaeolithic ~45 000-year-old Ust'-Ishim sample from Siberia placed near the basal node of haplogroup NO with an estimated mutation rate of 0.76×10^{-9} bp^{-1} $year^{-1}$ (Table 1) (Fu et al., 2014). Incorporation of European Mesolithic ~8 000-year-old male sample from Loschbour and the Siberian Ust'-Ishim sample into a phylogeny

constructed with a hundred high coverage sequences from modern male individuals yielded the mutation rate of 0.71×10^{-9} bp^{-1} $year^{-1}$ (Table 1) (Trombetta et al., 2015b). Karmin et al.(2015) placed two ancient samples on a tree constructed with modern MSY samples and accounted for "missing" transversions, as transitions happen frequently due to post-mortem damage, which resulted in the mutation rate of 0.74×10^{-9} bp^{-1} $year^{1}$ (Table 1). In conclusion, the current aDNA-based approach has provided remarkably similar estimates of Y-chromosomal evolutionary mutation rate with overlapping confidence intervals (Table 1).

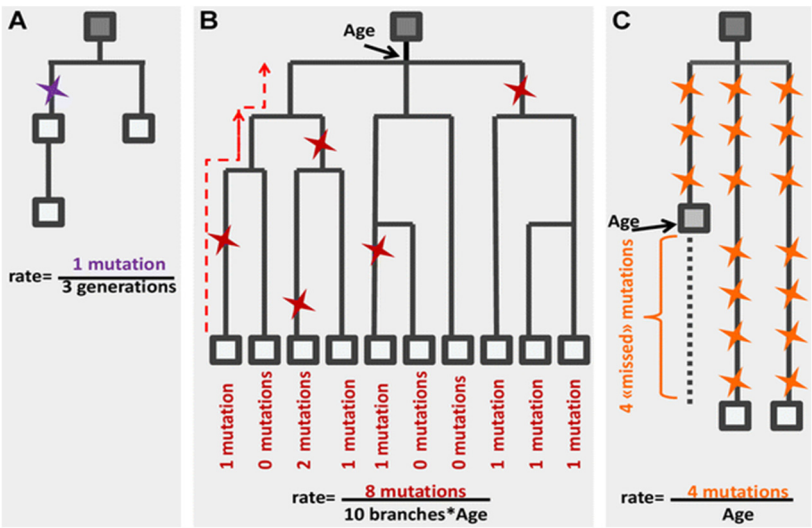

Figure 5. An overview of the three approaches to estimate the mutation rate on the Y chromosome.
a) Genealogical approach. Mutations separating members of the pedigree are counted and divided by the number of generations. **b)** Calibration approach. The average number of mutations from the MRCA to the modern samples divided by the time to the MRCA, which is assumed to coincide with a population event of known date. **c)** Ancient DNA approach. The older the ancient sample, the less time if has had to accumulate mutations. Thus the number of "missed" mutations is proportional to the (radiocarbon) age of the sample. Reprinted by permission of "Springer Nature" from Figure 1 in Balanovsky, (2017); permission conveyed through Copyright Clearance Center, Inc.

2.1.4.4. Comparison of Y-SNP mutation rates

Presently published mutation rates and their corresponding confidence intervals mostly overlap within two groups based on the calculation approach (Table 1). Genealogical rates tend to be approximately 20% faster than evolutionary rates (Balanovsky, 2017), which is in accordance with the notion of a faster, short-term mutation rate and a slower, long-term substitution rate influenced by puri-

fying selection (Ho and Larson, 2006). In addition to the uncertainties of correlating archaeological settlement dates with population expansions, the remarkably slow evolutionary rate by Francalacci et al.(2013) is depended on revealed mutations, and is considered to be limited to the phylogenetic tree on which it was calculated (Francalacci et al., 2016). There is also evidence on mutation rate variation between and within different regions of the Y chromosome and future research needs to take this into account (Helgason et al., 2015; Trombetta et al., 2015b). The question concerning the choice of the published mutation rates remains unsettled. Statistically viable options involve applying the lowest and highest estimated rates or averaging aDNA-based estimates for lower bound and genealogical estimates for upper bound as choosing a single published rate might be regarded biased toward a preferential result (Poznik et al., 2016; Balanovsky, 2017). In the coming years, additional deep-rooting pedigrees and the growing database of precisely dated aDNA samples with entirely sequenced Y chromosomes are expected to alleviate the current confusion and provide more information toward a potential convergence of Y-SNP mutation rates (Wang et al., 2014a; Francalacci et al., 2016; Balanovsky, 2017).

Table 1. Mutation rates for human Y-chromosomal single nucleotide polymorphisms.

Mutation rate bp^{-1} year $^{-1}$ (× 10^{-9})	**95% confidence interval (× 10^{-9})**	**Approach**	**Reference**
1.24	–	Evolutionary Chimpanzee/human	Thomson et al., 2000
1.50	–	Evolutionary Chimpanzee/human	Kuroki et al., 2006
1.00	0.3 – 2.5	Genealogical	Xue et al., 2009
0.61	0.44 – 0.71*	Genealogical Autosomal model	Mendez et al., 2013
0.89	0.80 – 0.99	Genealogical X-degenerated region	Helgason et al., 2015
0.74	0.64 – 0.85	Genealogical Palindromic region	Helgason et al., 2015
0.78	0.62 – 0.94	Genealogical	Balanovsky et al., 2015
0.82	0.72 – 0.92	Evolutionary	Poznik et al., 2013
0.53	0.52 – 0.55	Evolutionary Low coverage	Francalacci et al., 2013
0.65	0.62 – 0.68	Evolutionary High coverage	Francalacci et al., 2013
0.76	0.67 – 0.86	Ancient DNA	Fu et al., 2014
0.74	0.63 – 0.95	Ancient DNA	Karmin et al., 2015
0.71	0.62 – 0.82	Ancient DNA	Trombetta et al., 2015b

** 90% confidence interval*

2.1.4.5. Mutation rate of Y-STRs

In the earlier age of SNP scarceness, the faster mutating STRs were the markers of choice for inferring haplogroup age and dating possible migrations associated with this haplogroup. Over 4 000 STRs are located on the Y chromosome and they remain key components in paternity analysis and forensic identification. A study by Ballantyne et al.(2010) investigated 186 of them in nearly 2000 father-son pairs, resulting in mutation rates from 3.78×10^{-4} to 7.44×10^{-2} per locus per generation. Diverse factors such as repeat tract interruptions, repeat tract length and father's age are found to be influencing the mutability of individual STR loci (Goedbloed et al., 2009; Ballantyne et al., 2010; Willems et al., 2016). In practise, only a subset of STRs are analysed with commercially available kits and variation in the mutation rate of individual STR loci highlights the importance of locus-specific approach in any STR-based dating application. Both genealogical and phylogenetic approaches have been used to estimate the mutation rates of the most frequently genotyped STRs. In some of the largest pedigree studies, over 1000 family samples have been collected to count the mutations in father-son pairs (Gusmão et al., 2005; Hohoff et al., 2007; Ge et al., 2009; Goedbloed et al., 2009; Wang et al., 2016). High sample sizes have produced comparable results with relatively high precision and although some significant discrepancies with smaller scale studies have been found (Decker et al., 2008), a general agreement on mutation rate for STR pedigrees is emerging (Table 2) (Ge et al., 2009; Balanovsky, 2017).

Table 2. Mutation rates for human Y chromosome short tandem repeats (STRs) in evolutionary and genealogical studies with large sample sizes (>1000).

Mutation rate ($\times 10^{-3}$ per locus per generation)	95% confidence interval	Number of father-son pairs	Number of STRs	Approach	Reference
2.0	$1.5 – 2.6 \times 10^{-3}$	3026	17	Genealogical	Gusmão et al., 2005
2.1	$1.5 – 3.0 \times 10^{-3}$	1029	15	Genealogical	Hohoff et al., 2007
2.1	$1.7 – 2.5 \times 10^{-3}$	2913	17	Genealogical	Ge et al., 2009
2.5	$1.6 – 3.4 \times 10^{-3}$	1764	17	Genealogical	Goedbloed et al., 2009
2.6	$1.9 – 3.5 \times 10^{-3}$	1033	17	Genealogical	Wang et al., 2016
6.9	$1.2 – 12.9 \times 10^{-4}$	–	7	Evolutionary	Zhivotovsky et al.,2004

The evolutionary rate of 6.9 x 10^{-4} per STR locus per generation was estimated by Zhivotovsky et al. (2004) using data on STR variation in several populations with documented recent histories (Table 2). The rate was later revised to better suit different STR sets, but remained the same for the most commonly genotyped 17 STR loci (Shi et al., 2010). The at least threefold difference between the genealogical and evolutionary mutation rates has been considered controversial. Probable explanations include genetic drift eliminating the genealogically emerged diversity through time and possible back mutations due to the fast mutation rate of STRs, which all decrease the observed diversity (Zhivotovsky et al., 2006; Wei et al., 2013b). Despite theoretical reasoning at least partially explaining the discrepancy, it has remained unclear which rate should be used in population genetic studies. The accessibility of whole genome sequencing has made it possible to compare time estimates obtained by Y-STR marker variability with ages calculated with the more reliable Y-SNP substitution rates. The general conclusion states that evolutionary rates are more suitable for older haplogroups and the faster genealogical rates work better for younger haplogroups, although exact age boundaries for "young" and "old" vary (Hallast et al., 2015; Karmin et al., 2015; Balanovsky, 2017). Ages calculated with both types of STR-based mutation rates deviate substantially from the SNP-based ages, which are set to become the gold standard in the context of evolutionary dating and haplogroup age estimates, while STR-based approaches will remain restricted to the youngest of lineages.

2.2. Phylogeography of Y-chromosomal haplogroups in northern Eurasia

Northern Eurasia is a vast region stretching from the Fennoscandian fjords to the shores of Beringia. The Arctic Ocean forms a natural northern geographic barrier to the Eurasian landmass, whereas the location of the southern border of northern Eurasia is subject to a somewhat arbitrary choice. In this thesis, northern Eurasia is defined as the region north of the Caucasus Mountains, Black and Caspian Seas and bounded in the south by the present-day political border of Mongolia with China and further southeast by Russian border with China. Naturally, both ethnically and linguistically diverse human populations with a varied history of settlement inhabit such an enormous territory.

Earliest archaeological remains of human colonisation in Fennoscandia – the most western area of northern Eurasia – date to *circa* 10 000 BP and follow the retreating ice sheet formed during the Last Glacial Maximum (LGM) (Bergman et al., 2004). Solid archaeological evidence of human occupation on the north and south coast of the Baltic Sea that predate the last glacial period are lacking. In contrast, the vast majority of northern Eurasian mainland remained ice-free and habitable during the LGM as the north-eastern limit of the Eurasian ice sheet did not extend past the Kara Sea (Svendsen, 2004; Kuzmin, 2008; Ivanovic et al., 2016). Archaeological findings of human hunting activities

within the Siberian Arctic stretch back to 45 000 years BP (Pitulko et al., 2016). Direct evidence of human occupation in northern Eurasia comes from genome sequences of a ~ 45 000-year-old femur (Ust'-Ishim) from a modern human male discovered in western Siberia (Fu et al., 2014) and fragmented milk teeth found at the ~32 000-year-old Palaeolithic site on the Yana river well above the Arctic circle (Pitulko et al., 2004, 2017; Sikora et al., 2018). Archaeological findings suggest that at least southern Siberia was populated by humans even at the height of the LGM (Kuzmin, 2008). The latter notion is supported by sequenced aDNA extracted from ~ 25 000-year-old sample unearthed in Ma'lta in south-central Siberia (Raghavan et al., 2014). A multitude of migrations and demographic events has undoubtedly shaped the long history of human settlement in northern Eurasia and most probably only a fraction remain detectable in the modern genomes. However, genomic studies of contemporary populations cover sample sizes large enough for general inter- and intrapopulation inferences and provide necessary background comparison for the emergent field of aDNA research.

Although earlier publications exploit less markers than are known today and often have provided only partial resolutions of the available data, a sufficient number of studies has been conducted to provide a broad overview of the main haplogroups describing the absolute majority (>95%) of extant male lineages in northern Eurasia, that is geographically divided into continental northern Europe and northern Asia at the Ural Mountain range.

2.2.1. Phylogeny and spread of Y-chromosomal haplogroups in northern Europe

In northern and northeastern Europe, a total of four Y-chromosomal hgs (hgs R1a, R1b, I and N) describe over 75% of male lineages. Among European men, two sublineages of hg R (hgs R1a-M198 and R1b-M269) are the most frequent and are noteworthy for their opposing frequency clines. The hg R1b-M269 itself is characterized by a dichotomy of primarily L23-defined eastern lineages in the Caucasus and Uralic region (Figure 6A) and western lineages defined by M412 polymorphism (Figure 6B). Subclades of hg R1b-M412 reach very high frequencies among Western European men (almost 100% in Western Ireland) and decline towards the east (Myres et al., 2011). In northern Europe, hg R1b-M412 characterizes about 20% of Swedish males, but drops abruptly to about 3% in the Finns (Figure 6B) (Karlsson et al., 2006), but has an exceptional frequency peak (up to 75%) in a northern subpopulation of Bashkirs residing on the southwestern foothills of the Ural Mountain range (Myres et al., 2011). Baskhir paternal lineages are however very heterogeneous and geographically separated subpopulations display differing frequency levels of subclades (Myres et al., 2011; Post et al., 2019). The clinal spread of hg R1b-M412 has been linked with Palaeolithic migrations from population refugiums (Semino, 2000)

and Neolithic demic diffusion (Balaresque et al., 2010), but the latter association was subjected to criticism because of highly uncertain and often criticised STR-based coalescence time calculations (Busby et al., 2012) Novel evidence based on Y-chromosomal sequences acquired through NGS show a shallow coalescence time-depth of 5–7 KYA for the most common European subclade of R1b-M412 (Batini et al., 2015; Hallast et al., 2015; Karmin et al., 2015; Poznik et al., 2016).

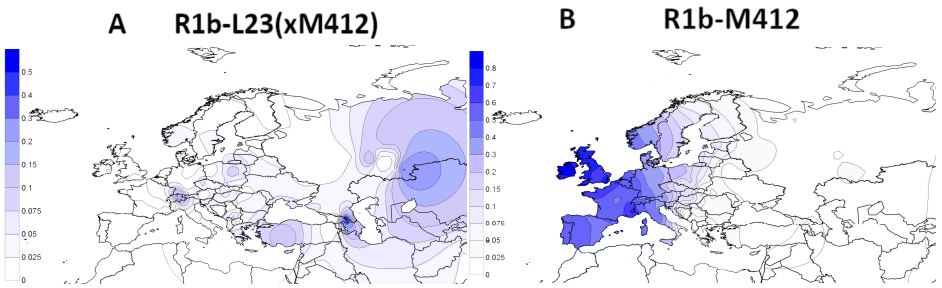

Figure 6. Spatial-frequency distributions of frequency data for human Y-chromosomal haplogroups **a)** R1b-L23 and **b)** R1b-M412.
The frequency data were converted to spatial-frequency maps using Surfer software (version 7, Golden software Inc., Cold Spring Harbor, NY, USA), following the Kriging procedure. Adapted by permission of "Springer Nature" from Figure 1 in Myres et al., (2011); permission conveyed through Copyright Clearance Center, Inc.

Haplogroup R1a has a remarkably wide distribution area extending from eastern Europe to southern Siberia and Central and South Asia (Underhill et al., 2009). It is characterized by a phylogenetic division separating European and Asian sublineages (Figure 7): hgs R1a-Z282 and R1a-Z93, respectively (Underhill et al., 2009; Pamjav et al., 2012). A radiating decline is evident in Europe − R1a-Z282 characterises over 40% of all male lineages in Poland, Ukraine, Belarus and southern Russia, but occurrence rate falls gradually towards Scandinavia, Western Europe, the Balkans and the Volga-Uralic region (Figure 7A) (Kayser et al., 2005; Underhill et al., 2015). The European lineages do not extend to Siberia past the Uralic Mountains and the R1a lineages that are found at frequencies as high as 50% in some Altaian and South Siberian populations form a separate North Asian subclade R1a-Z93(xZ95) (Underhill et al., 2015). High frequencies of hg R1a-Z95 (a sublineage of R1a-Z93) describe Central and South Asian populations (Figure 7B) (Underhill et al., 2015). According to Y-chromosomal sequencing data, the split time of Asian and European R1a lineages dates to approximately 5–6 kya (Poznik et al., 2016), followed by a rapid diversification within a relatively short time span (Batini et al., 2015). Extant basal lineages and high STR diversity within hg R1a suggest Middle

East and present-day Iran as a possible geographic origin of hg R1a (Underhill et al., 2015), but this hypothesis remains to be reinforced by future aDNA research. In aDNA record, extant lineages of R1a and R1b appear in samples from Europe dated to Late Neolithic/Bronze Age and their spread is connected with massive migration from the steppe after ~3000 BCE (Haak et al., 2015). In northeastern Europe, lineages belonging to extant European R1a clades appear with the Corded Ware Culture as an extension of the steppe cultures and continue to dominate through the Bronze Age (Saag et al., 2017, 2019; Mittnik et al., 2018).

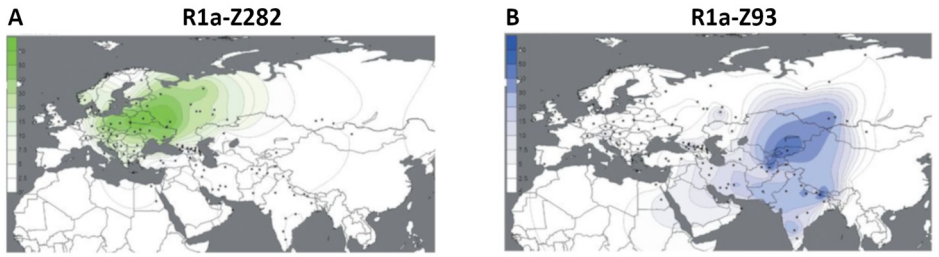

Figure 7. Spatial-frequency distributions of frequency data for human Y-chromosomal haplogroups **a)** R1a-Z282 and **b)** R1a-Z93.
The frequency data were converted to spatial-frequency maps using Surfer software (version 7, Golden software Inc., Cold Spring Harbor, NY, USA), following the Kriging procedure. Adapted by permission of "Springer Nature" from Figure 2 and Figure 3 in Underhill et al., (2015); permission conveyed through Copyright Clearance Center, Inc.

A large fraction of male lineages in northern Europe belong to hg I. It contains two subhaplogroups I1a-M253 and I1b-P37, that exhibit a contrasting frequency gradient (Figure 8) (Rootsi et al., 2004). Latest research involving NGS sequences revealed a more nuanced picture with clade I1b-P37 belonging to a higher lever haplogroup I2'3-M438 that contains several subclades confined mostly to Europe (Karmin et al., 2015). A deep split of approximately 30 kya separates I1-M253 and I2'3-M438 (Karmin et al., 2015; Poznik et al., 2016). Hg I1-M253 (nomenclature according to Karmin et al., (2015)) encompasses almost 45% of the Swedish male population sampled from Scandinavia (Figure 8A) (Rootsi et al., 2004; Karlsson et al., 2006; Lappalainen et al., 2008). It has an equally prominent occurrence in the neighbouring Uralic-speaking Finnish and Saami males and shows a decreasing trend south towards the Baltic countries and east towards the Ural Mountains (Rootsi et al., 2004). Within hg I2'3-M438, clade I2a-M423 has highest (20–40%) frequencies in northwestern Balkans and spreads to eastern Europe, where it is the major hg I subclade in Slavic-speaking Ukrainians and Southern Russians, but is gradually outnumbered by the subclade I1-M253 in Northern Russians (Figure 8) (Balanovsky et al., 2008; Kushniarevich et al., 2015). Both subclades can be found outside their high

frequency areas at a marginal (<5%) level. Lineages of hg I have been found in aDNA from European hunter-gatherer samples, providing evidence for hypothesized autochthonous European origin of the haplogroup (Semino, 2000; Haak et al., 2015; Jones et al., 2015; reviewed in Kivisild, 2017).

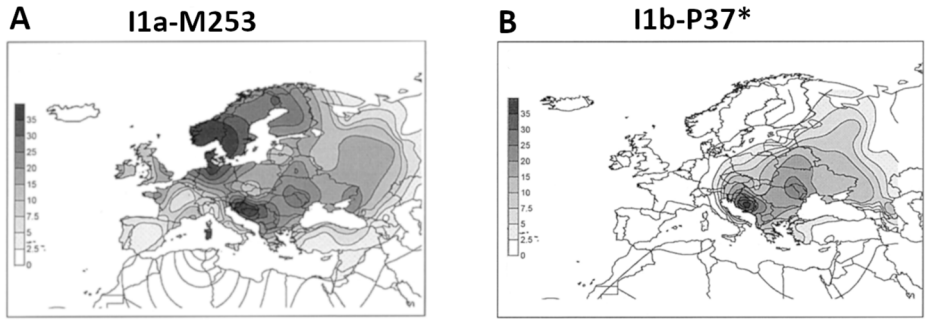

Figure 8. Spatial-frequency distributions of frequency data for human Y-chromosomal haplogroups **a)** I1a-M253 (I1-M253 in Karmin et al., (2015)) and **b)** I1b*-P37 (I2a-M423 in Karmin et al., (2015)).
The frequency data were converted to spatial-frequency maps using Surfer software (version 7, Golden software Inc., Cold Spring Harbor, NY, USA), following the Kriging procedure. Adapted by permission of "Elsevier" from Figure 1 in Rootsi et al., (2004); permission conveyed through Copyright Clearance Center, Inc.

Haplogroup N, although globally distributed at a shallow frequency level, shows a remarkably wide distribution area and high frequency particularly in northern Eurasia (Karafet et al., 2008) (Figure 9A). It transcends the conventional boundary between Asia and Europe by being one of the main haplogroups in populations on both side of the Ural mountains (Zerjal et al., 1997; Rootsi et al., 2000; Tambets et al., 2001; Karafet et al., 2002). It is present in almost all Siberian populations, reaching above 90% in Yakuts, Nganasans and Nenets and falls below 10% in the Han Chinese (Yan et al., 2011; Fedorova et al., 2013). The reduced diversity is a common feature in Siberian populations, with only one major haplogroup describing whole populations, which reflects their small effective population sizes and prolonged isolation (Zerjal et al., 1997; Karafet et al., 2002). In Europe, hg N accounts for a large proportion of Y chromosomes in all Uralic-speaking populations with the exception of Hungarians, who have a marginal level of hg N (Post et al., 2019), whereas its frequency drops abruptly in Indo-European speaking neighbour populations. Only Baltic-speaking Latvians and Lithuanians deviate from this pattern and harbour frequencies of hg N as high as their Uralic-speaking northern neighbour populations (Villems et al., 1998; Rootsi et al., 2000; Tambets et al., 2001, 2004; Pliss et al., 2006; Lappalainen et al., 2008). Hg N is suggested to represent a common "Uralic component", that is exemplified by the significant decline in

frequency from north to south within the Russian population, suggesting a possible language shift in some populations ancestral to contemporary Northern Russians (Balanovsky et al., 2008). None of the Late Neolithic/Bronze Age samples associated with migrations from the east into Europe belong to hg N, signifying hg N to be a later addition to the European paternal gene pool (Saag et al., 2017; Mittnik et al., 2018).

Haplogroups J, G and E form a minor component in the haploid gene pool of men in northern Europe. Hg J is present throughout northern Eurasia at low frequencies. It is one of the most common haplogroups in Anatolia and in regions surrounding the Mediterranean Sea and was initially considered a signal of demic diffusion of male farmers from the Middle East to Europe during the Neolithic (Semino, 2000) (Figure 9B). Later research revealed hg J to consist of surviving ancient sublineages with subtle radiation signs in the Bronze Age (Di Giacomo et al., 2004; Zalloua et al., 2008; Finocchio et al., 2018). Modern populations from Anatolia lack samples that would coalesce to more basal ancestral nodes (Finocchio et al., 2018), but this haplogroup has been found in hunter-gatherers from the Caucasus and Karelia (Jones et al., 2015; Mathieson et al., 2015) and has probably originated north of its present-day main Anatolian distribution area (Kivisild, 2017; Finocchio et al., 2018).

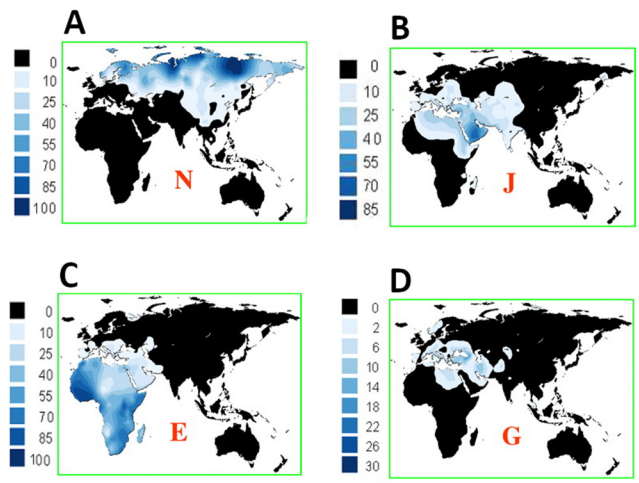

Figure 9. Global phylogeographic distribution maps for human Y-chromosomal haplogroups N, J, E and G.
Adapted with permission of *PNAS* from Figure 2 in Chiaroni et al., (2009).

Additional two lineages associated with Neolithic expansions are hgs E3b-M35 and G-M201. The former is a subclade of haplogroup E, which is more common in southern Europe, but falls below the frequency of 10% in the populations of northern Europe. Hg E is the most common Y chromosome clade in the entire African continent and forms a collection of deep subclades with very different

evolutionary histories, including an association with the dispersal of early pastoralists to south Africa (Cruciani et al., 2004; Henn et al., 2008) (Figure 9C). The broadly defined hg E3b-M35 has now been refined into several deep subclades with only a few of them observed at higher frequencies outside Africa (Cruciani et al., 2007; Trombetta et al., 2015a). Detailed analysis of the phylogeographic distribution of these lineages in northern Eurasia is currently lacking, but a survey in a limited number of European populations has demonstrated the subclade E3b-V13 to account for the majority of European hg E3b-M35 male lineages (Semino, 2000; Cruciani et al., 2007; Trombetta et al., 2015a). This subclade originated approximately 8 kya, which is consistent with the hypothesis of a Neolithic expansion that might have passed through the Balkans, where this lineage encompasses about a quarter of the extant Y chromosome pool (Cruciani et al., 2007; Trombetta et al., 2015a).

In northeastern Europe, hg G accounts for less than 5% of male lineages, but is very common in the Caucasus, Near and Middle East (Rootsi et al., 2012) (Figure 9D). The co-occurrence of two basal lineages – G1 and G2 and high diversity within subhaplogroups suggest eastern Turkey, Armenia or western Iran as potential locations for the geographic origin of hg G. It is the prevalent male haplogroup genotyped in samples of early farmers from Anatolia and Central Europe, signifying a connection to the spread of agriculture in Europe (Rootsi et al., 2012; Haak et al., 2015; Mathieson et al., 2015). The 5 300-year-old Tyrolean Iceman from the Copper Age belongs to a subhaplogroup of G2, currently most frequent in Corsica and Sardinia while nearly absent in mainland Europe, and clusters with modern Sardinians based on autosomal analysis, which lends support to the demic diffusion of Neolithic agriculturalists and their male lineages across Europe (Keller et al., 2012).

2.2.2. Phylogeny and spread of Y-chromosomal haplogroups in northern Asia

Haplogroups N, C, O and Q characterise the majority of Y chromosomes in northern Asia. Hgs O and C are mostly restricted to Asia, whereas hg Q reaches high frequency only in some northern Asian populations, but is the defining male lineage in the Americas. Hg N occupies a wide area across entire northern Eurasia, but is rare in South and East Asia (Figure 9A, see 2.2.3). Hg O, a neighbour clade of hg N, is the dominant male haplogroup across populations in East and Southeast Asia, extending also to eastern India and comprising roughly three quarters of Han Chinese and at least half of Japanese and Koreans (Figure 10A) (Su et al., 1999; Hammer et al., 2006b; Yan et al., 2011). Some of its sublineages are associated with rapid expansion 7–5 kya, a timeline that coincides with the shift to intensive agriculture in East Asia (Yan et al., 2014; Poznik et al., 2016). In northern Eurasia, the haplogroup is dominant in Mongol, Kazakh and the indigenous Nivkh populations while being rare in most other

North Asian populations. In Mongols, the sublineage O3–122 accounts for the majority of Mongol hg O Y chromosomes, reflecting the proposed northward migration from South Asia (Katoh et al., 2005; Shi et al., 2005; Wang et al., 2013).

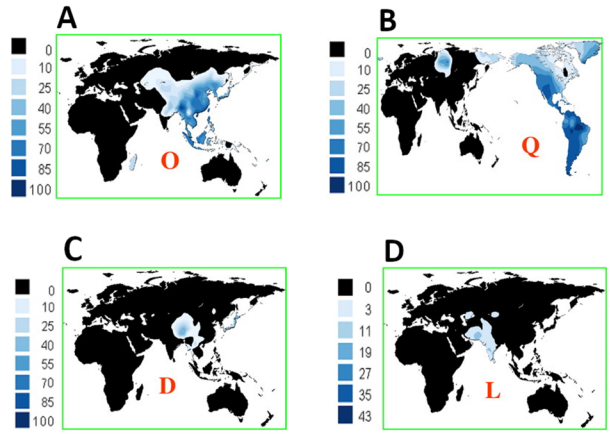

Figure 10. Global phylogeographic distribution maps for human Y-chromosomal haplogroups O, Q, D and L.
Adapted with permission of *PNAS* from Figure 2 in Chiaroni et al., (2009).

One of the most widely distributed Y chromosome haplogroups in East Asia and Oceania is hg C that is characterised by geographic specificity – differentiated subclades are restricted to Japan, Oceania, India and Australian Aborigines (Figure 11) (Hammer et al., 2006b; Kayser et al., 2006; Sengupta et al., 2006; Hudjashov et al., 2007). Subhaplogroup C3-M217 has the broadest spread across Asia, Siberia and the Americas (Zhong et al., 2010). In northern Asia, it accounts for over 50% of male lineages in Mongolic-speaking Buryats, Mongols and Oirats as well as in small indigenous populations of Siberia, such as Evens, Evenks, Nivkhs, Nanai, Negidals and Yukaghirs, whereas it is below 1% in the surrounding Russian population (Karafet et al., 2002; Lell et al., 2002; Malyarchuk et al., 2010; Balinova et al., 2019). Two haplotypes of subhaplogroup C3-M217 account for 3–8% of all sampled men in East Asia and show very rapid spread during last 1000 years based on STR-calculated time to MRCA (Zerjal et al., 2003; Xue et al., 2005). This has led to its associations with the male descendants of Genghis Khan and Qing Dynasty nobility (Zerjal et al., 2003; Xue et al., 2005). However, newer research with the advantage of NGS has linked the "Genghis Khan" cluster to at least two times earlier diffusion of all Mongolic-speaking populations (Wei et al., 2018b). Regardless of speculative historical assignments, the notable increase in frequency of such lineages warrants social selection as a probable explanation (Zerjal et al., 2003; Malyarchuk et al., 2010).

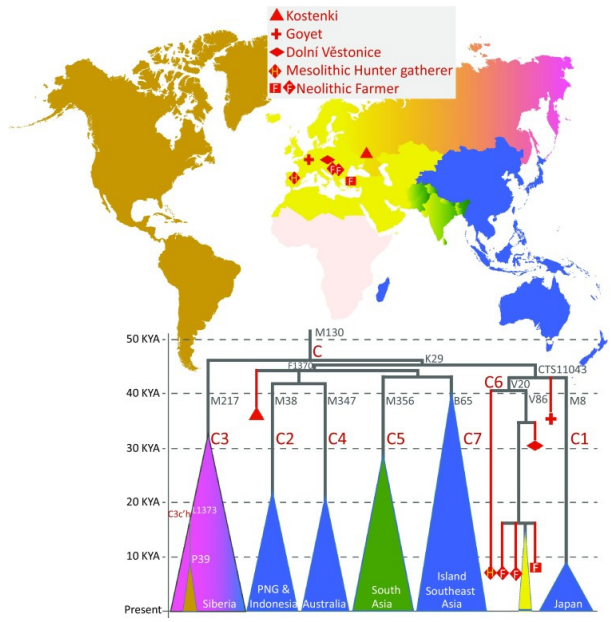

Figure 11. Major subclades of hg C in ancient and present-day populations.
The tip of each triangle is in proportion to the subclade's coalescent time estimated from high coverage genomes in present-day populations. The phylogenetic mapping of ancient Y chromosomes is shown with red symbols. Haplogroup names are shown in brown font and their defining SNP marker names in grey font next to relevant branches. Time scale shown in KYA (thousand years ago). Triangle colours match geographic regions on the map. PNG – Papua New Guinea. Adapted from Figure 4 in Kivisild, (2017) under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Subclades of hg C occur in Palaeolithic and Mesolithic hunter-gatherers in Europe, Early Neolithic hunter-gathers from Lake Baikal region and Neolithic samples from Central Europe, but none of the ancient lineages are found in modern Europe, where nowadays hg C is almost non-existent (Figure 11) (Seguin-Orlando et al., 2014; Mathieson et al., 2015, 2018; Kivisild, 2017; de Barros Damgaard et al., 2018). Haplogroup's extensive spread and present-day limited regional overlap between sublineages suggest that hg C might mark the earliest settlers of Eurasia followed by long-time isolation between its carriers (Zhong et al., 2010; Kivisild, 2017).

Haplogroup Q is scattered across Eurasia at generally low incidence but with frequency peaks in some Siberian populations (Kivisild et al., 2003; Karafet et al., 2008). It represents one of the founder lineages carried by the initial settlers of the American continent – sublineage Q1-L54 accounts for at least 75% of male Native Americans (Karafet et al., 1997; Zegura et al., 2004; Battaglia et al., 2013; Wei et al., 2018a). Same paternal lineage is present at high frequencies in South Siberian populations residing in Altai region, linking indigenous

South Siberians and Native Americans through common paternal ancestry (Figure 10B) (Lell et al., 2002; Dulik et al., 2012). The high frequency of hg Q in Ket and Selkup populations and its low occurrence or near absence in most other Siberian populations may reflect founders effect coupled with genetic drift enhanced by small population sizes (Karafet et al., 2002; Fedorova et al., 2013). Hg Q incidence above 10% in Tadjik and Turkmen populations and the presence of rare early branches in Iran and in the Arab Peninsula suggests West or Central Asia as a possible source area of hg Q (Wells et al., 2001; Balanovsky et al., 2017). Studies on aDNA indicate wide dispersal of hg Q already in late Neolithic and early Bronze Age – Y chromosomes belonging to subhaplogroup Q1a have been identified in ancient remains dated 6500 BCE from Latvia (Mathieson et al., 2018) and in approximately 3000-year old samples from Central Plain area in China (Zhao et al., 2014).

Haplogroups D and L form a rare and minor addition to the paternal pool of lineages in Northern Eurasia. Hg D in one of the main haplogroups among populations in the Japanese archipelago, Tibet, Southeast Asia and the isolated Andaman Islands, whereas hg L is primarily found in southern Eurasia: South Asia, Anatolia, South Caucasus and Europe along the Mediterranean coast (Figure 10CD) (Karafet et al., 2001, 2008; Kivisild et al., 2003; Thangaraj et al., 2003; Cinnioglu et al., 2004; Hammer et al., 2006b; Sengupta et al., 2006; Wang and Li, 2013).

2.2.3. Haplogroup N as a sign of gene flow between Eastern and Western Eurasia

A study by Zerjal et al. (1997) was among the first to identify a frequent SNP shared by a subset of Asian and European populations. The authors pointed out that the majority of European men carrying the derived allele speak a language belonging to the Uralic language family, whereas allele frequency drops abruptly in most neighbouring Europeans speaking Indo-European languages (Zerjal et al., 1997). Authors also suggested a Siberian origin of the polymorphism based on frequency gradient and STR diversity in the background of the mutation identified as Tat (=M46) that defines one subclade in the broader hg N (Zerjal et al., 1997; Karafet et al., 2008). Ensuing research detected additional upstream and downstream polymorphisms that correspondingly define the whole hg N, its neighbour hg O and the subhaplogroups within hg N (Fig 12A), followed by attempts to unify the nomenclature of the emergent phylogenetic tree (Rosser et al., 2000; Semino, 2000; Underhill et al., 2000; The Y Chromosome Consortium, 2002; Jobling and Tyler-Smith, 2003; Cinnioglu et al., 2004; Karafet et al., 2008). Further research revealed additional high incidence in Indo-European speaking Latvians, Lithuanians and Northern Russians and an abrupt frequency drop in Polish and Swedish males (Lahermo et al., 1999; Rootsi et al., 2000; Tambets et al., 2001; Karlsson et al., 2006; Pliss et al., 2006; Balanovsky et al., 2008; Lappalainen et al., 2008). A curious exception are the

modern Hungarians, who contrast with other Uralic-speakers by exhibiting near-absence of hg N (Rootsi et al., 2000; Tambets et al., 2001, 2004). However, latest studies showed common STR haplotypes within hgs R1a and N (Dudás et al., 2019) and a shared hg N lineage, although present at low frequencies, between Hungarian and Volga-Ural/West Siberian men (Post et al. 2019).

A discovered bifurcation within hg N resulted in the subclades N2 and N3 (Figure 12A). Continuous addition of new branches has caused several reassessments of the nomenclature (Figure 12A) and this section follows the classification proposed by Jobling and Tyler-Smith, (2003). Both subclades attain significant level of incidence in Khanty and Mansi populations residing in Northwest Siberia (Karafet et al., 2002; Tambets et al., 2004; Pimenoff et al., 2008). The prominence of hg N in northern Eurasia has led to the question of its geographic origin and timing of the dispersal. Owning to the limited genetic resolution available in the early studies, interpretations of the spread of hg N and its ensuing diversification were primarily dependent on the sampled STR variation and the uncertainties of the STR mutation rate. This led to inconclusive and conflicting results with suggestions of a possible European origin and a subsequent west-to-east migration of hg N carriers or a migratory route that reached northeastern Europe before the Ural Mountain region (Villems et al., 1998; Rootsi et al., 2000; Derenko et al., 2007; Mirabal et al., 2009a, 2009b). South Siberia was proposed as a possible area of origin of subclade N2 based on comparison of STR based TMRCAs (Malyarchuk and Derenko, 2009). However, wider geographic sampling and increased sample size pointed to high genetic drift in small and diffusely located Siberian populations, leading to reduced diversity estimates that confound diversity-based population comparisons (Karafet et al., 2002; Fedorova et al., 2013).

Hgs NO, N*, N3* have been found in Neolithic samples excavated from several archaeological sites in Northern China, whereas currently prevalent hg O starts to appear in sites dated to a more recent time period (Cui et al., 2013; Gao et al., 2015; Zhang et al., 2017). Serial male population changes are reflected in Y chromosome hgs in aDNA samples from the vicinity of Late Baikal, where hg N2 is the main haplogroup during late Neolithic (~5000 BCE), but changes to hg Q in Early Bronze Age (~2000 BCE) (de Barros Damgaard et al., 2018). In northeastern Europe, earliest carriers of hg N appear along with autosomal Siberian ancestry in hunter-gatherer remains dated to ~1500 BCE in the Kola Peninsula (Lamnidis et al., 2018) and in samples from *tarand*-type graves associated with the onset of Iron Age (~800 BCE) in Estonia (Saag et al., 2019). Earlier Estonian Bronze Age samples lack both the hg N and autosomal signs of Siberian ancestry (Saag et al., 2017; Mittnik et al., 2018). Latest research has revealed a total of 5 hg N3* samples out of 29 graves of the Conquerors who arrived into the Carpathian Basin ~900 CE and are considered to be the founders of present-day Hungary and carriers of the Hungarian language (Neparáczki et al., 2019). However, their pool of Y chromosomes was extremely heterogeneous featuring major European hgs R1b, R1a, I1 and I2a, which reflects their multi-ethnic background as an alliance of tribes with poten-

tial connections to Bashkirs, Volga Tatars and other populations in the Volga-Uralic region (Neparáczki et al., 2019).

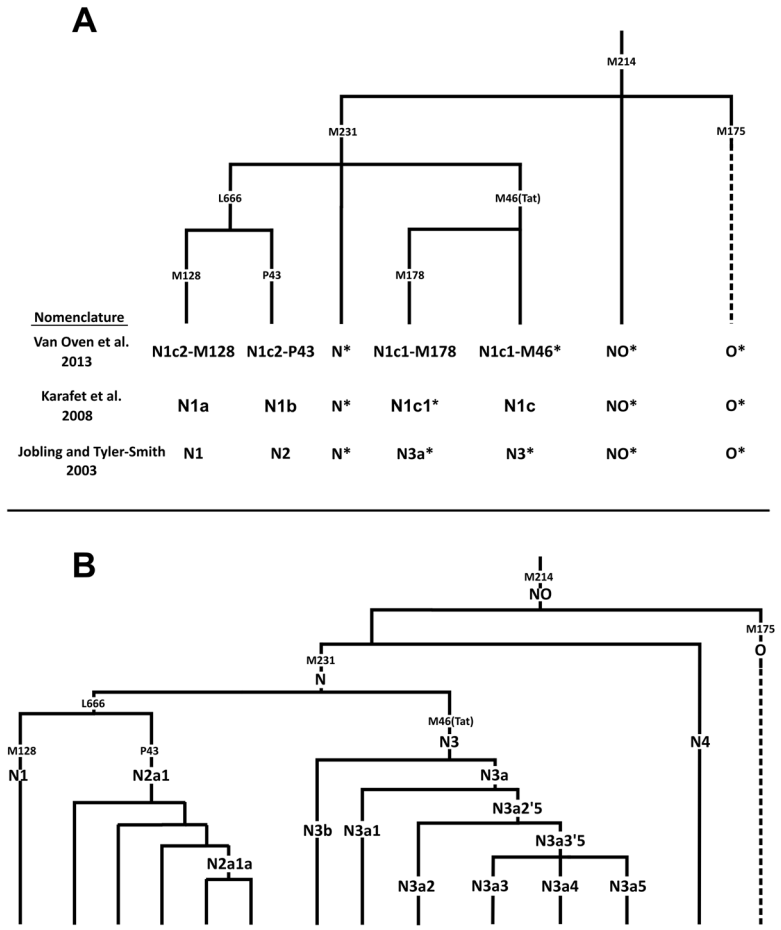

Figure 12. Schematic overview of phylogenetic relationships in the NO clade with the defining SNP markers.
a) nomenclature and tree topology according to the references; **b)** nomenclature and tree topology according to Karmin et al.(2015). Dotted lines indicate hg O lineages.

A breakthrough in understanding the precise inner structure of Y-chromosomal haplogroups, among them also hg N, in a manner free of ascertainment bias along with branch lengths proportional to time came with application of NGS sequencing to long sections of MSY (Francalacci et al., 2013; Wei et al., 2013a; Hallast et al., 2015; Karmin et al., 2015; Poznik et al., 2016; Finocchio et al., 2018). Publications with a sufficient number of hg N samples show that hg N separates into two deep basal clades – hgs N4 (nomenclature according to

Karmin et al.(2015)) and hg N 1'3 that is a large clade formed by males mostly from Siberia and Europe (Figure 12B) (Karmin et al., 2015; Poznik et al., 2016). Both Karmin et al. (2015) and Poznik et al. (2016) used mutation rate derived from aDNA and provided broadly concurrent age estimates of *circa* 40 – 45 ky for NO clade and 19 – 22 ky for haplogroup N. Karmin et al. (2015) were the first to publish a comprehensive refined topology of hg N coupled with geographically broad sampling range and preliminary age estimates of the reported novel clades. Hg N1'3 separates further into two subclades: hg N1'2 that incorporates the previously known subclade N1b-P43 with its formerly undetermined diversification, and hg N3 that is a large clade which includes the initially discovered M46(Tat) polymorphism shared by Siberian and European populations (Zerjal et al., 1997; Karmin et al., 2015). Hg N3 follows a complex diversification pattern with a rather recent (~5 kya) subclade N3a3'5 that unites single sequenced samples from as diverse populations as Finns, Estonians, Mongols and Koryaks (Karmin et al., 2015). However, the reviewed publications focus on discovering unbiased variance underlying the true tree topology and lack broader population-scale approaches for clarifying phylogeography of the discovered subclades.

2.3. Autosomal approaches in studying human genetic diversity

2.3.1. Types of available autosomal data

2.3.1.1. Microarray data

Most demographic histories lead to many variable genealogy trees. Any tree built from a single locus such as the Y chromosome represents one realisation of many equally possible genetic histories for the same population and the variance caused by the random nature of trees can be reduced by considering many nuclear loci simultaneously for estimating genetic diversities, divergence times, admixture and other population parameters of interest (Rosenberg and Nordborg, 2002; Brumfield et al., 2003; Hey and Machado, 2003; Nielsen and Beaumont, 2009). Technological advances have transformed the accumulation of genetic data from surveying a handful of variable loci with restriction fragment length polymorphism to rapid genotyping of hundreds of thousands of SNPs with DNA hybridization microarrays to the massively parallel NGS.

At its core, the microarray approach relies on sequencing a panel of individuals and then using the discovered polymorphisms for microarray-based genotyping in a much larger sample size. Despite being the fastest and most affordable genome-wide genotyping method, inherent limitations of this approach stem from ascertainment bias associated with the selection of individuals for the initial sequencing step and the distorted minor allele frequency spectrum (McTavish and Hillis, 2015). As the polymorphisms are discovered in a limited number of samples, there is a relatively smaller probability of capturing variance with rare

minor allele frequency and the observed variance tends to be skewed towards SNP loci with intermediate and common occurrence (Nielsen, 2004; Lachance and Tishkoff, 2013). Ascertainment bias will affect estimates of any population genetic parameters based on the allele frequency spectrum as well as tests for natural selection and the magnitude of its effect will depend on the exact ascertaining scheme. For example, when the heterozygosity of variants discovered in a European ascertainment panel is assessed in African populations, it will falsely show European populations to harbour a greater amount of variation (Eller, 2001). According to the neutral theory of evolution, population bottlenecks result in deficiency of rare alleles and ascertainment bias can inflate the estimates of population contraction. Comparison of whole genome sequences and SNP array data in African hunter-gatherers show higher values for population differentiation measures, such as F_{ST}, for ascertained array polymorphisms, demonstrating the inflating effect of the ascertainment bias on population differentiation estimates (Lachance and Tishkoff, 2013). Accurately phased haplotype data are more robust to ascertainment bias (Lohmueller et al., 2009; Haasl and Payseur, 2010). Ascertainment bias can also be corrected for by predicting the properties of the missing polymorphisms and several approaches to account for the absent variance have been proposed (Nielsen, 2004; Clark et al., 2005; Albrechtsen et al., 2010; Quinto-Cortés et al., 2018), but not extensively utilised, as they require either resequencing data or precise knowledge of the ascertainment schemes, which are often complex and unclear (Novembre and Ramachandran, 2011; Lachance and Tishkoff, 2013).

2.3.1.2. Next generation sequencing data

It is impossible to completely evade ascertainment bias, as all frequency distributions are drawn from small sample sizes and not the entire populations, but sequencing whole genomes would minimise it significantly. Although the latter is becoming more and more economically feasible, the costs of good quality high-coverage data are still prohibitive for routine use in large sample sizes. The challenges of analysing NGS data come from substantially larger amount of generated information and shorter read lengths with higher error rates relative to traditional Sanger sequencing approach and involve many analytical and technical choices preceding the final set of SNPs subjected to further study (Crawford and Lazzaro, 2012; Schraiber and Akey, 2015). The large amount of data together with different coverage levels, platform-specific error profiles along with artefacts have necessitated the development of a multitude of bioinformatic algorithms for efficient data analysis with computational costs often surpassing those of data production (Mardis, 2010).

The initial key step, upon which any following NGS analysis is based, is the accurate mapping of reads to an already sequenced reference genome. The main challenges lay in mapping reads that differ from the reference genome or originate from a repetitive genomic region that is longer than the read itself

(Pfeifer, 2017). By chance, such reads might map to several locations in the genome, leading to potential biases and errors in the variant calling procedure. The use of paired-end or mate pair reads has enhanced mapping quality in such regions, with determined gaps between reads providing information on placement of one read if another is confidently mapped (Pfeifer, 2017). However, placing reads remains difficult in repetitive or highly polymorphic areas, and the ongoing development of sophisticated mapping methods along with increasing read length and combining multiple technologies is still required to increase the quality of the final alignment (Pool et al., 2010; Pfeifer, 2017).

The first step in generating the final high-quality variant call set involves identifying individual alleles at all sites that differ from the reference sequence. Accurate calling of heterozygous sites requires high-coverage data (>20x) to mitigate effects of random sequencing errors – observing at least 10 reference reads and 10 reads with alternative allele makes a strong case for a truly heterozygous site (Schraiber and Akey, 2015). In the most common study design case of sequencing large number of individuals at a low or medium coverage, probabilistic variant calling approaches that compute genotype likelihoods with additional information (for example, base and alignment quality scores, error profiles of different platforms, read coverage) need to be applied (Nielsen et al., 2011; Pfeifer, 2017). The most commonly used software operating in the probabilistic framework is the Genome Analysis Toolkit (GATK) (McKenna et al., 2010).

After variant calling, a filtering step should remove false positive variant sites to increase specificity. Filtering can be done by considering certain thresholds of variant characteristics, such as location in regions with poorly aligned reads, low quality scores and tight clustering with other variants, prompting custom approaches in sex chromosomes that differ in terms of coverage patterns and rate of homozygous calls. Newer filtering methods use machine learning to achieve higher specificity at low coverage levels (Cheng et al., 2014). It is also common practise to filter out sites that excessively deviate from Hardy Weinberg equilibrium, as this suggests error-prone regions with unreliable variants. Although error-mitigating strategies of NGS technology have substantially improved since its wide introduction in 2005, no gold-standard exists and different sequencing applications, platforms and research goals require custom-tailored approaches to the selection of numerous, often software-specific, parameters and algorithms (Mardis, 2011; Pfeifer, 2017).

2.3.2. Detecting and dating admixture with autosomal data

2.3.2.1. Approaches based on variation of allele frequencies

Finalised variant data set is usually first subject to a population structure analysis. Understanding the genetic structure of sampled individuals is necessary for understanding and modelling evolutionary history, including admixture, of sampled populations, with prior knowledge of population structure often one of the necessary input parameters for further modelling (Figure 13). Principal

component analysis (PCA) is now widely used as the most common method for inferring population structure and its use in conjunction with geography and human genetic variation was pioneered by Cavalli-Sforza et al., (1994). The main idea is to reduce the large dimensionality of genotype data by algebraically identifying principal components that most efficiently represent the variation present within the data without any prior classification of the samples. Analysed samples are then plotted according to their first few principal components and individuals with more similar genotypes will tend to cluster together, whereas more-distantly related individuals will lie further apart and admixed individuals will be positioned relatively in-between their ancestral clusters (Figure 13B) (Brisbin et al., 2012; Schraiber and Akey, 2015). Additionally, PCA can be used as a quality control measure to detect technical sources of variation that might arise when samples are genotyped with different instruments or at different facilities (Schraiber and Akey, 2015). However, in the context of population admixture history and geography, genetic resemblance patterns obtained with PCA should be regarded with caution. Gradient-like patterns that emerge when principal components are projected onto a geographic map might arise from either isolation-by-distance or a directed migration, leading to inconclusive interpretations that require further analysis with additional methods and integration of alternative sources of information such as aDNA or archaeology and historical records (Novembre and Stephens, 2008). However, PCA remains a valuable tool for detecting population (sub)structure within and between populations and correcting for stratification in disease studies (Reich et al., 2008).

Additional approaches for studying genetic ancestry and structure are exemplified by the model-based Bayesian method STRUCTURE – a clustering algorithm first developed by Pritchard et al., (2000). In brief, STRUCTURE and STRUCTURE-like approaches use genotype data to probabilistically assign individuals to *K* number of clusters representing genetically distinct populations, each of which is characterised by a set of allele frequencies at each locus, group individuals that share underlying common allele frequencies and identify admixture proportions from each cluster *K* at the individual level (Figure 13C) (Pritchard et al., 2000; Liu et al., 2013). The initial STRUCTURE implementation is too slow for whole genome data and computationally more efficient maximum-likelihood estimation approaches such as ADMIXTURE (Alexander et al., 2009) or FRAPPE (Tang et al., 2005) are recommended for large-scale datasets. The selection and interpretation of parameter *K* is the key challenge in all of the methods mentioned above. The ADMIXTURE software helps to choose the best *K* that minimises the cross-validation error and has been shown to be faster and more accurate than FRAPPE, making it the preferred software for genetic studies in human populations (Alexander et al., 2009). However, it is advisable to inspect results from a range of *K* values and regard any STRUCTURE-like analysis as the beginning stage of a detailed demographic and admixture analysis with critical assessment of the underlying model and its ramifications (Novembre, 2016; Lawson et al., 2018).

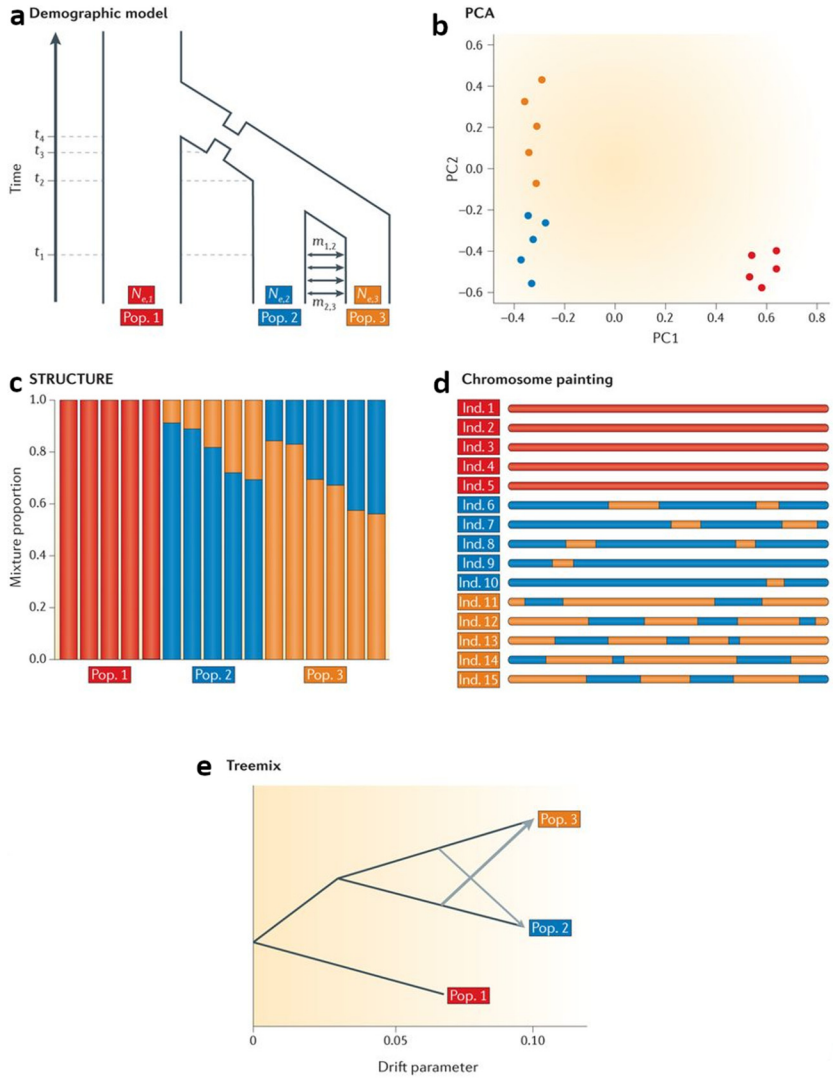

Figure 13. Schematics for the output of various methodological tools for assessment of population history.
a) A simple three-population model with changes in population size and asymmetric gene flow. **b)** Principal component analysis (PCA) qualitatively illustrates population structure and admixture of populations 2 and 3 by the spread of individuals along the first (PC1) and second (PC2) principal components. **c)** STRUCTURE provides estimates of the proportion of each individual genome from populations 1, 2 and 3. **d)** Chromosomal painting shows the specific tracts of sequences inherited from ancestors in each population. **e)** Treemix illustrates asymmetrical migration rates (difference in arrowhead size) between populations 2 and 3. Figure adapted by permission of "Springer Nature" with minor alterations from Figure 2 in Schraiber and Akey, (2015); permission conveyed through Copyright Clearance Center, Inc.

Specific hypothesis of admixture and population history can be formally tested with methods that are developed within a framework of fitting mixture events to allele frequencies observed in multiple populations, allowing exploration of potential population contacts that fit the genetic data (Patterson et al., 2012). Expanding upon these ideas, Reich et al.(2009) introduced several F-statistics and Green et al. (2010) the D-statistic. One of the F-statistics, the F_3-statistic explicitly tests whether a population of interest (C') is directly derived from admixture between two other populations (A' and B') as exemplified by the admixture graph on Figure 14D. The statistic measures the covariance of the difference in allele frequencies between populations C' and A' and populations C' and B'. Positive result (F_3 (C'; A', B') >0) shows that there is no evidence that C' is the result of an admixture event between A' and B', whereas negative result (F_3 (C'; A', B') < 0) indicates that C' might be admixed from A' and B'. A limitation in interpretation might come from high population-specific drift in admixed population C' that increases the value of branch length C-C' and masks the true signal of admixture (Figure 14D) (Patterson et al., 2012).

The F_4-statistic was proposed as measurement of shared drift to validate a proposed unrooted tree topology of four populations (all possible topologies presented in Figure 14 panels A, B, C). Drift is defined as the frequency change of an allele along a branch of a tree – hence drift between populations P_1 and P_2 is a function of differences in the allele frequencies in P_1 and P_2 (Patterson et al., 2012). If the true tree can be formulated as (P_1,P_2 ; P_3,P_4), then $F_4 = 0$, in case of notation (P_1,P_3 ; P_2,P_4) representing the true clusters, the statistic is positive and in case of (P_1,P_4 ; P_2, P_3) being the true relationship the F_4(P_1,P_2 ; P_3,P_4) < 0. Essentially, the value of F_4 measures the length of the internal branch of the tree with the sign indicating the true topology. A false negative result might arise if P_4 is admixed from both P_2 and P_3 in equal proportions, as equal amount of ancestry stems from each source and is opposite in drift direction (Reich et al., 2009; Harris and DeGiorgio, 2017).

The D-statistic is used for testing the presence of gene flow from a third group into either of the two sister groups by validating a proposed rooted four-population tree topology. The initial D-statistic was applied to detect Neanderthal introgression into modern humans (Green et al., 2010). The basis lays in counting alleles that differ between sister groups and that are shared between the putatively introgressing population and one of the sister groups (Harris and DeGiorgio, 2017). Additionally, the outgroup population must share the same allele as the unadmixed sister population, indicating that the derived allele arose on the genealogical branch connecting the donor and the putatively admixed population. The value and sign of the D-statistic depends on the excess pattern of the shared alleles – positive in case of more alleles shared between the second sister population and the admixing population (non-Africans and Neanderthals in Figure 14E), negative in case of more alleles shared between the first sister population and the admixing population (Africans and Neanderthals in Figure 14E) and zero in case of absent excess of shared alleles. The statistic can also be called the ABBA-BABA-statistic, according to the sharing

pattern of ancestral (A) and derived (B) alleles on the branches of the proposed tree topology (Figure 14E). Low sequencing depth can skew true allele counts and obscure relationships between populations. Such is often the case for aDNA studies, where error correction for sample contamination and ensuing calculation modifications within the D-statistic framework need to be taken into account (Raghavan et al., 2014; Soraggi et al., 2018). Significance for the F- and D-statistics is based on Z-score calculated as weighted mean value of the statistic across all blocks of equal length, divided by the standard error (Harris and DeGiorgio, 2017) and both statistics have been shown to be robust to ascertainment bias (Patterson et al., 2012).

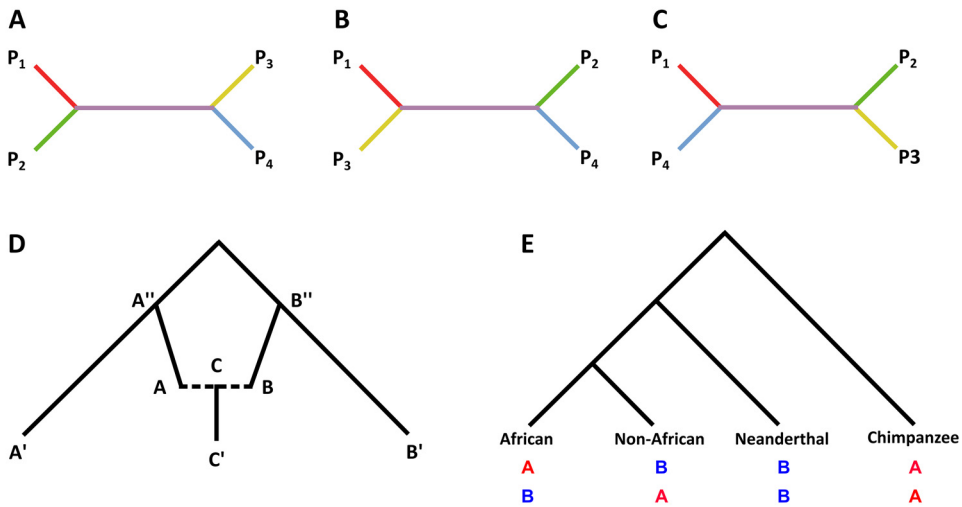

Figure 14. Schematic visualisations for the F_3-, F_4-, D-statistic and Alder.
a-c) Unrooted four-population trees showing the coloured drift paths for $F_4(P_1,P_2 ; P_3,P_4)$ for the three possible topologies. **a)** $F_4(P_1,P_2 ; P_3,P_4) = 0$ **b)** $F_4(P_1,P_2 ; P_3,P_4) > 0$ **c)** $F_4(P_1,P_2 ; P_3,P_4) < 0$. In case of **b)** and **c)** trees the magnitude of F_4 is equivalent to the length of the internal (purple) branch. **d)** Three-population admixture graph showing the relationships between populations A, B and C applicable to F_3-statistic and Alder software. Present-day population C' is descended from an admixture between A and B to form C; populations A' and B' are present-day references. Branch lengths represent drift between ancestral and descendant nodes. **e)** Asymmetric four-population tree rooted to the chimpanzee outgroup to which the D-statistic was originally applied in Green et al., (2010). Ancestral alleles labelled as A, derived alleles as B. ABBA indicates sites where the derived allele is shared between non-African and Neanderthal, BABA sites are those where African and Neanderthal populations share the derived allele (B), while non-African and chimpanzee share ancestral allele (A). Figures 12ABCE adapted from Harris and DeGiorgio, (2017) with permission of Wayne State University Press and 14D adapted by permission of Genetics Society of America from Figure 1A in Loh et al., (2013).

Building upon the reviewed ideas and the available allele frequency data, graph construction methods relating a larger number on populations have been developed. The qpGraph software tests the fit of F-statistics to a user-specified bifurcating tree of *n* populations (Patterson et al., 2012). Developed as an expansion to qpGraph, MixMapper first builds a bifurcating tree of unadmixed populations using the F_3-statistic and incorporates the a priori defined admixed populations (Lipson et al., 2013). TreeMix software utilizes a somewhat different approach by building a directed acyclic graph from allele frequency data through fitting a particular model of population relationship in a maximum likelihood framework (Figure 13E) (Pickrell and Pritchard, 2012). Descendant populations have the same mean allele frequencies as their ancestor and branch lengths reflect the drift measurement descendants share relative to their ancestor. Additionally, TreeMix allows to explore alternative demographic scenarios by evaluating the fit of the data to graphs with and without an arbitrary number of migration events (Harris and DeGiorgio, 2017). Graph construction methods fit patterns in allele frequency correlation across populations. Choosing the most appropriate method relies on the assumed demographic complexity among populations of interest (Harris and DeGiorgio, 2017). This factor is however ultimately unknown, necessitating the evaluation of a vast number of possible tree topologies and potential migration events, making it difficult to assess the extent to which equally likely solutions exists (Mathieson et al., 2018; Wangkumhang and Hellenthal, 2018).

2.3.2.2. Approaches based on identification of shared haplotypes

It is reasonable to assume that an admixed individual should have genomic segments inherited from different ancestral populations and that such segments diminish over time due to the process of recombination. Identifying these segments, their length, distribution and linkage pattern allows obtaining a more refined picture of admixture. Modelling correlation in ancestry along the genome leads to the production of mosaic ancestry maps also knows as local ancestry deconvolutions or chromosome “paintings” (Figure 13D) (Brisbin et al., 2012; Gravel, 2012; Schraiber and Akey, 2015). Approaches like Saber (Tang et al., 2006), LAMP (Sankararaman et al., 2008) and PCAdmix (Brisbin et al., 2012) divide genomes into evenly-sized sliding windows and assign an ancestry label to each of them. The ancestry assignment depends on either majority vote using all overlapping windows (LAMP), PCA-based ancestry identification of variants (PCAdmix) or must be predefined by the user (Saber). Most recent alternative methods such as RFMix (Maples et al., 2013) and EILA (Yang et al., 2013) take a different route by modelling the dependency of the unobserved variables (ancestries) directly as a function of the observed variables (alleles) and have been shown to be faster and more accurate (Padhukasahasram, 2014). These methods do not test an explicit parametric population genetic model, but are valuable for selecting variant reference panels based on ancestry or iden-

tifying recent targets of selection, and allowing more accurate identification of genetic variants associated with disease in admixed populations (Pasaniuc et al., 2011; Brisbin et al., 2012; Jin et al., 2012).

None of the mentioned methods model linkage disequilibrium (LD) in the ancestral populations and thus omit additional information extractable from densely genotyped data. An influential model proposed by Li and Stephens (2003) approximated the coalescent process with recombination in a scalable and computationally efficient manner featuring hidden Markov model (HMM). In the context of admixture, HMM combines observed genotypes at neighbouring markers to provide additional information regarding ancestry of a variant and accounts for uncertainties such as differences in allele frequencies between true ancestors and their modern population proxies (Falush et al., 2003; Tang et al., 2006). HAPMIX was the first haplotype-based method to apply and extend the Li and Stephens model in order to determine the genetic ancestry for each chromosomal position or segment in the genome (Price et al., 2009). Unlike HapMix, subsequent haplotype-based-developments – fineSTRUCTURE (Lawson et al., 2012) and Globetrotter (Hellenthal et al., 2014) – do not require predefined ancestral populations, instead building a coancestry matrix, in which each element estimates the number of haplotypes from a genome of an individual that are most closely related to corresponding haplotypes from another sampled individual. This matrix forms the basis for either performing PCA or for a model-based analysis to identify clusters of individuals forming genetically related populations with similar genetic ancestry profiles (Lawson et al., 2012). FineSTRUCTURE then reconstructs the hierarchical relationship between clusters in the form of a tree (Busby et al., 2015). If there is no close proxy for the ancestral populations, especially likely for ancient admixture or poorly sampled regions, Globetrotter approximates the pattern of haplotype sharing of the admixed population by combining several most suitable donor populations (Hellenthal et al., 2014). Both methods have been extensively used in studies revealing fine-structure and admixture in human populations on both global and local levels (Busby et al., 2015; Leslie et al., 2015; Hudjashov et al., 2018) as well as in populations of other species such as for example wheat (Joukhadar et al., 2017).

An alternative haplotype-based method for detecting admixture involves identifying genomic segments that are identical-by-descent (IBD) between individuals. An IBD tract is usually defined as a contiguous region that descends from a common ancestor without recombination. Lack of recombination means that there is little need to model complex recombination patters, which significantly simplifies the implementation of IBD-based methods (Schraiber and Akey, 2015). The challenge lies in mitigating the excess of false positive segments created by identical by state nucleotides that are counted as IBD. False IBD segments can occur in evolutionary conserved regions or regions with unusually high LD and this issue needs to be addressed with careful filtering criteria that account for both length and frequency of the segments (Browning, 2008). Several methods for detecting IBD tracts from chip genotyping data have been

published (Purcell et al., 2007; Gusev et al., 2008; Browning and Browning, 2013b) and there is an ongoing development of newer approaches that focus on sequence data, while also incorporating info on demography (Browning and Browning, 2013a; Tataru et al., 2014). IBD sharing patterns can be used to estimate populations' historical N_e (Browning and Browning, 2015) and fine-scale substructure on either single-population level or continental between-populations level, showing a decline in pairwise sharing of IBD tracts that correlates with geographic distance (Ralph and Coop, 2013; Martin et al., 2018). For admixture estimation, a test devised to identify deviation from expected distance-based decay of shared IBD tracts was used to detect an elevated level of ancestry between Turkic speakers in Western Eurasia and present-day populations from South Siberia and Mongolia in comparison to their immediate non-Turkic speaking geographic neighbours (Yunusbayev et al., 2015; Pankratov et al., 2016).

2.3.2.3. Approaches for timing admixture

In addition to detecting potential admixture, estimating the timeframe for the admixture event allows for a more nuanced picture of the ancestry of a population. For whole genome data, most widely used dating methods are haplotype-based approaches involving modelling the signature of decay in LD between a pair of variants located on the same chromosome as the distance between the sites increases (Harris and DeGiorgio, 2017). A schematic phylogeny with point admixture is represented on Figure 14D, where admixed population C' retains longer-range LD from a past mixture event involving previously separated populations A' and B' (Loh et al., 2013). Calculating the exponential LD decay pattern shows the number of recombination events that have occurred since the admixture and the latter is reflected in the length of chromosomal segments from the admixed population that can be used to obtain a timeframe in generations (Moorjani et al., 2011). In mathematical terms, for two alleles X and Y drawn from an admixed individual in population C', the probability after n generations that X and Y originate from the same haplotype is e^{-nd}, and observed correlation of alleles can be presented as a function of their genetic distance d – the weighted LD statistic $A(d)$ – that is approximated by the result of exponential decay from the initial state A_0 such that $A(d) \approx A_0 e^{-nd}$ (Harris and DeGiorgio, 2017). The rate of decay will allow estimating the time in generations (n) – steeper decay corresponds to older admixture. ROLLOFF by Moorjani et al., (2011) was the original method to apply this idea for testing and dating African admixture in Europe and the Levant (Figure 15). To avoid confounding background LD, the optimal starting distance between markers should be at least 0.05 cM (Moorjani et al., 2011).

Extending the ideas behind ROLLOFF, Alder derives the formula for A(d) as dependent on the mixture proportion of each admixing lineages and the expected squared allele frequency difference for a randomly drifting neutral allele

drawn from two admixing sources (Loh et al., 2013). In addition to several orders of magnitude faster computational speed, Alder implements several improvements in comparison to ROLLOFF. It can compute time of admixture from a single donor population when the other is unsampled and accounts for shared demography producing spurious false positive results by comparing background LD in both the test and admixing populations and using only sufficiently distanced loci for fitting the LD decay curve (Loh et al., 2013). As the true allele frequencies from past populations are unknown, authors of Alder algorithm show that the genetic drift along the A''A'/B''B' lines (Figure 14D) adds noise to the LD decay curve, but does not change the timeframe estimates of admixture. Alder also provides an LD-based test of admixture that complements the F_3 test and in contrast to F_3 helps to identify admixed populations that have undergone extensive drift, but has limitations to detect admixture events older than roughly 250 generations because of the rapid decay of the LD curve (Loh et al., 2013).

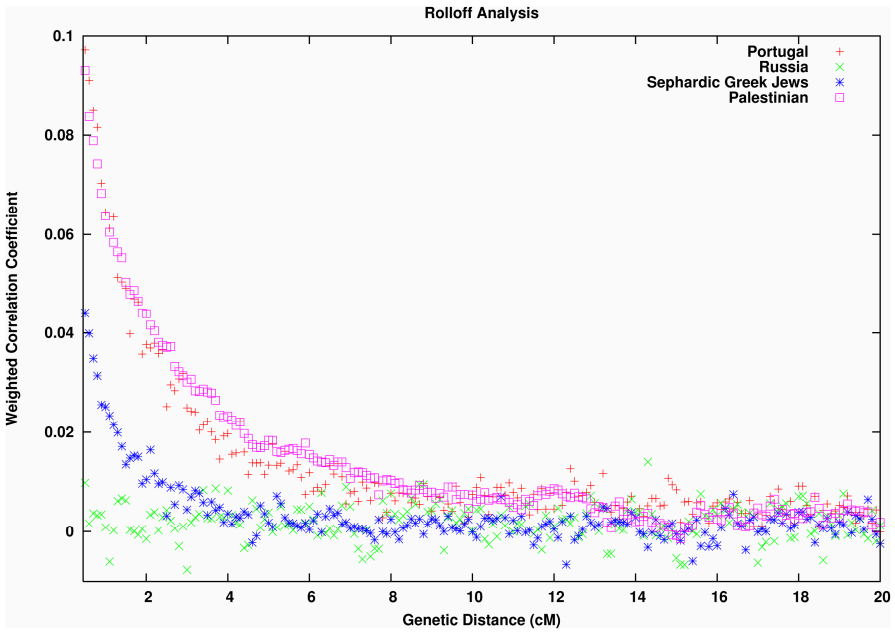

Figure 15. Testing for LD due to African admixture in Portuguese, Russians, Sephardic Greek Jews and Palestinians with ROLLOFF.
Algorithm calculates the LD between pairs of markers in each population, weighted by their frequency difference between African Yoruba and a corresponding West Eurasian population to make the statistic sensitive to admixture LD. A significant exponential curve indicating African admixture is detected in Portuguese, Palestinians and Jewish, but not in Russians. Republished from Figure 3 in Moorjani et al., (2011) under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Globetrotter is the most recent LD-decay based admixture dating software that relies on similar theoretical reasoning that is behind both ROLLOFF and Alder, but gains power by using haplotypes that are shared between populations instead of allele frequencies and modelling source populations as a mixture of several possible "donors" (Hellenthal et al., 2014). Globetrotter also accounts for a possibility of two admixture events by fitting a model with a mixture of LD curves with different decay rates and a different set of source population, estimating timeframe for both admixtures. These improvements provide a more realistic approach to human population history. Admixture simulations with samples from real populations have shown haplotype-based Globetrotter to be more accurate in estimating dates and admixture sources relative to ROLLOFF (Hellenthal et al., 2014), however Globetrotter authors did not complement the comparison with results from Alder.

2.4. Uralic-speaking populations

2.4.1. Dispersal of Uralic languages

The Uralic language family comprises of 47 known languages clustered into 9 subgroups and spoken by about 25 million people across northeastern Europe and Siberia with the exception of Hungarian speakers residing in Central Europe (Figure 16A) (Janhunen, 2009; Syrjänen et al., 2013). By territorial extension it is one of the most widespread languages in northern Eurasia, but its spatial range is uneven with patches of Uralic speakers surrounded by much larger populations speaking Indo-European and Turkic languages. Hungarian (~ 13 million speakers); Finnish (~5 million speakers) and Estonian (~1 million speakers) are the three Uralic languages with the largest number of native speakers and statuses of official state languages (Eberhard et al., 2019). All other Uralic languages are considered to be endangered minority languages spoken in Russia, Finland, Estonia, Sweden and Norway (Abondolo, 2017). (Abondolo, 2017). However, Erzya, Moksha, Mari, Udmurt, and Komi are constitutionally recognised as official regional languages of the Russian Federation and dialects of Sami are official minority languages in several Finnish, Swedish and Norwegian municipalities.

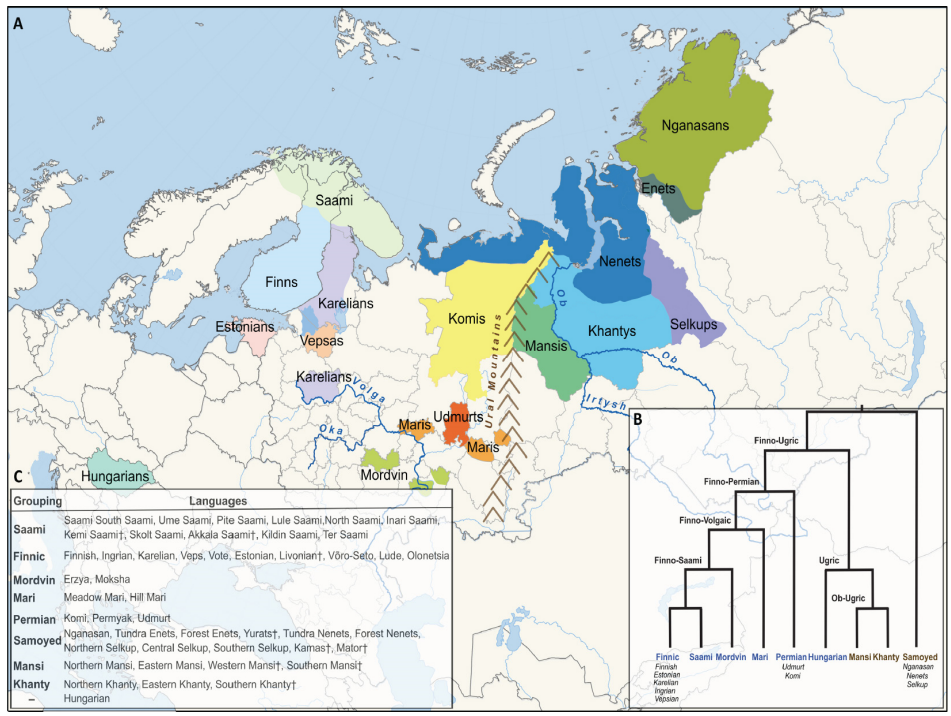

Figure 16. a) Geographic dispersal of Uralic speakers. **b)** Simplified phylogeny of Uralic language family (Korhonen, 1981; Syrjänen et al., 2013). **c)** Uralic languages and their main subgroupings. Extinct languages are marked with † (Syrjänen et al., 2013). Panels CB adapted from Syrjänen et al., (2013) with permission from "John Benjamins Publishing Company" (https://www.benjamins.com/catalog/dia). Distribution of languages according to Carpelan and Parpola, (2001), geographic landmarks by dr. Mait Metspalu.

The traditional lexical view, that was formed already in the early 20th century (Szinnyei, 1910), separates the Uralic languages into two large branches, the more diversified geographically western Finno-Ugric branch and the geographically eastern Samoyedic branch (Figure 16B) (Korhonen, 1981; Janhunen, 2009). Exact branching order and split times within these two clades has been debated (reviewed in Kallio, 2006), but latest methods in linguistics that follow the phylogenetic approaches known from biology have upheld the general classification tree presented in Figure 16B (Syrjänen et al., 2013). Divergence times calculations conducted in a Bayesian framework have dated the first bifurcation of eastern and western branch at ~5300 YBP (years before present), followed by the separation of Finno-Ugric to Finno-Permian and Ugric clades (Figure 17, but also note the wide 95% probability margins for some branches on Figure 17) (Honkola et al., 2013). Within the latter, Hungarian diverged from Khanty and Mansi languages ~3300 YBP (Honkola et al., 2013). Saami languages split from the Finnic clade ~2500 YBP and the latter separated

into northern (Finnish, Karelian, Veps) and southern (Estonian, Livonian) ~1200 YBP (Figure 17) (Honkola et al., 2013).

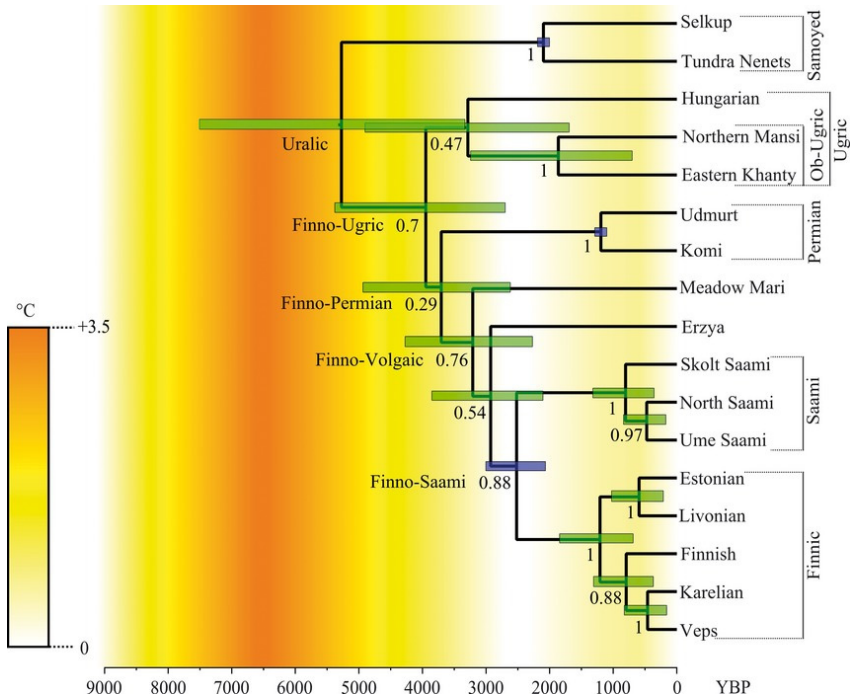

Figure 17. Timing analyses of the Uralic languages.
Green bars represent the 95% highest probability density for the divergence times. Scale values represent years before present (YBP). Values outside the nodes represent posterior probabilities. Calibration points (Samoyed, Permian and Finno-Saami) are labelled with blue bars indicating the uniform prior of the calibration points. Names of different protolanguages are marked on the nodes of the tree, and the names of different subclasses are on the right margins. The colour scale of the picture describes the temperature changes with relation to current temperature (+3.5–0 °C red-white) of the Northeastern Europe/East European tundra. Figure adapted by permission of "John Wiley and Sons" from Figure 2 in Honkola et al., (2013), permission conveyed through Copyright Clearance Center, Inc.

Combination of settlement archaeology, toponymy and proposed linguistic divergence times suggests that Fennoscandia and the territory of Estonia, currently most northwestern Uralic regions, were already inhabited by people speaking unknown language(s) prior to the arrival of Uralic people. Moreover, some linguistic evidence shows that the Saami languages were present in entire Finland before the precursor of modern Finnish, the linguistic carriers of which

gradually expanded northwards, admixing with as well as replacing the Saami-speakers (Aikio, 2012). According to historical documents and archaeological sources, the Hungarian language came to the Carpathian Basin in Central Europe in the end of the 9th century with conquering tribes most probably from Volga Ural region or Western Siberia (Róna-Tas, 1999). Classic principles of linguistic geography place the homeland of languages in the original location of the deepest split within the language family. For Uralic, this might correspond to the Volga river area southwest from the Ural Mountains (Salminen, 1999; Häkkinen, 2009), although alternative theories propose the region between Ob and Yenisei drainage area as a possible birthplace for the Proto-Uralic language (Janhunen, 2009). Regardless of the exact location, most linguists agree that the reconstructions of Proto-Uralic suggest its speakers were foragers who lived far from the sea in a forest environment in the vicinity of the Ural Mountains (Anthony, 2007).

Various theoretical explanations have been constructed to explain the wide spread of the Uralic languages. Archaeological record in the proposed linguistic homeland lacks any clear indication of technological or organisational advantage over any neighbouring community. Position at the frontier of forest and steppe belts or closeness to one of the first metal age cultures in southern Ural region have been suggested as possible strategic advantages facilitating the use, dispersal and subsequent diversification of Uralic languages (Janhunen, 2009). Another hypothesis connects known climate fluctuations during the Holocene and the followed ecological changes with demographic surplus or worsening environmental conditions that facilitated migrations of the language carriers (Honkola et al., 2013). The question remains open to interpretations.

2.4.2. Autosomal genetic diversity of Uralic-speaking populations

Autosomal genetic structure in Europe correlates with geography as demonstrated by large-scale studies involving thousands of genotyped individuals sampled from populations across Europe (Heath et al., 2008; Lao et al., 2008; Novembre et al., 2008; McEvoy et al., 2009; Nelis et al., 2009). The geography-dependent fine-scale genetic structure is evident even between rural European villages (O'Dushlaine et al., 2010). Differing linguistic affiliations have not generated any substantial genetic barriers between populations – for example, the Uralic-speaking Hungarians are genetically closest to their Indo-European-speaking Central European neighbours (Novembre et al., 2008; Novembre and Ramachandran, 2011).

Among Uralic-speaking populations, Finns and Estonians have been extensively studied on the genetic intrapopulation level and both display genetic variation predominantly dependent on geographic distance (Salmela et al., 2008; Nelis et al., 2009). However, in Europe-wide genetic comparison studies Fin-

nish population exhibits an outlier status (Lao et al., 2008; McEvoy et al., 2009; Nelis et al., 2009). The latter has been caused by a demographic history of founder effects and ensuing drift with probably only a marginal level of immigration (Kere, 2001; Wang et al., 2014b). Comprehensive haplotype analysis showed greater cumulative IBD sharing between a pair of individuals from Finland to be on average significantly larger than between individuals from other European countries, corresponding to an additional bottleneck absent in neighbouring populations (Martin et al., 2018). This has also led to unconventionally high or low carrier frequencies of some autosomal recessive diseases, that are extremely rare or, conversely, much more common in other European populations (Peltonen et al., 1995; Kere, 2001; Palo et al., 2009). The reduced diversity is especially prominent in Y chromosomal haplotype data, whereas mtDNA diversity is on the average European level (Sajantila et al., 1996; Palo et al., 2009; Översti et al., 2017). Within the Finnish population, a significant genetic diversification has been found between southwestern and northeastern Finland in studies involving Y-chromosomal and autosomal markers (Kittles et al., 1998; Lappalainen et al., 2006; Salmela et al., 2008). The substructure within Finland is of the same magnitude as between geographically much more separated Swedish and British populations and substantially stronger than that between British and Germans, acting as a cautionary example against assumptions of homogeneity in seemingly isolated and relatively small populations (Salmela et al., 2008). The presence of hidden genetic structure in a population isolate was affirmed by a genome-wide autosomal study of the Finnish Saami that also revealed a minor (below 10%) Asian contribution, which is also present in Eastern Finns and originates either from Saami-Finnish admixture or population migratory history (Salmela et al., 2008; Huyghe et al., 2011).

Genome-wide autosomal study has confirmed the distinctiveness of Northern Russians due to admixture with Finno-Ugric speaking populations and Komi as an additional pole of genetic diversity in geographical northeastern Europe, with the latter population also displaying genetic differentiation between subpopulations residing in different regions (Figure 18) (Khrunin et al., 2013).

Similar to Europe, geography accounts for the majority of observed genetic variation in Siberia, but long-distance migration and admixture have also shaped the Siberian gene pool (Pugach et al., 2016). Genetic differentiation in Siberia is higher than in most European populations (Triska et al., 2017; Wong et al., 2017) and is enhanced by several bottlenecks that are also manifested in elevated levels of genome-wide LD in populations living on the margins of the Siberian landmass, such as Samoyedic-speaking Nganasans and Chukotko-Kamchatkan-speaking Koryaks (Pugach et al., 2016). Samoyedic-speaking Nganasans and Nenets residing on the Taimyr peninsula display some genetic affinities to the Tungusic-speaking Evenks and Evens from central Siberia (Fedorova et al., 2013; Pugach et al., 2016; Wong et al., 2017). Despite their linguistic and geographic closeness, Nganasans are genetically distinct from Nenets and unlike Nenets lack any European ancestry components, indicating a differing genetic history (Pugach et al., 2016). Uralic-speaking Khanty, Mansi

and Nenets in Western Siberia share significant genetic ancestry with ancient North Eurasian individuals represented by 24 000- and 17 000-year-old Siberian aDNA samples (Raghavan et al., 2014) and this creates common genetic substrate with Native Americans who harbour ~40% of ancient North Eurasian ancestry (Raghavan et al., 2014; Wong et al., 2017).

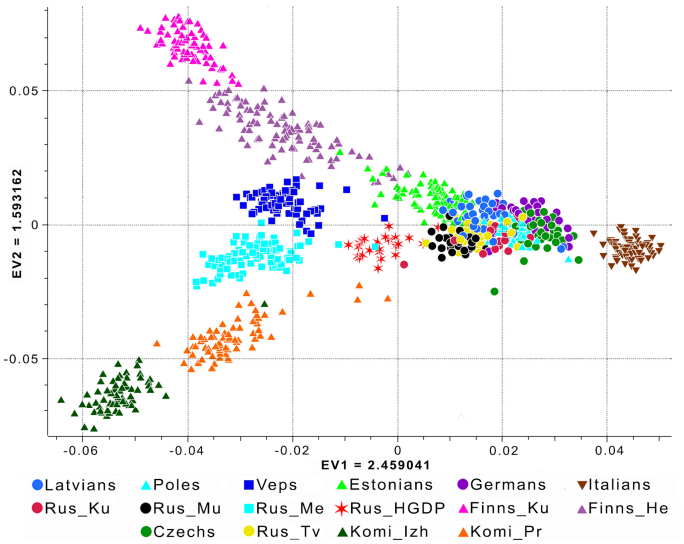

Figure 18. First two principal components of the PCA of the autosomal genotypic data of individuals from Russia and seven European populations.
Colour legend for the predefined populations is indicated within the plot. Population abbreviations: Finns_He – Finns from Helsinki; Fin_Ku – Finns from Kuusamo; Komi_Izh – Izhemski Komi; Komi_Pr – Priluzski Komi; Rus_HGDP – Russians from the Human Genome Diversity Panel; Rus_Me – Russians from Mezen; Rus_Mu – Russians from Murom; Rus_Ku – Russians from Kursk; Rus_Tv – Russians from Tver. Adapted from Figure 3 in Khrunin et al., (2013) under the Creative Commons Attribution License https://creativecommons.org/licenses/by/4.0/.

3. AIMS OF STUDY

Publications included in this study exemplify the developments in the field of human population genetics. The main goal of the first two publications is to provide an in depth phylogenetic analysis of the uniparentally transmitted Y-chromosomal hg N by following the emerging technological and methodological advancements on the levels of data acquisition and analysis. The third publication complements the uniparental approach and investigates genetic affinities and autosomal diversity of Uralic speakers – major carriers of hg N – and their geographic neighbours in northern Eurasia. The goals set in the three publications are outlined below.

Goals of the first reference (**Ref. I**):

- Detailed analysis of the geographic distribution of hg N and its most common subclades.
- Assess the direction of hg N spread.
- Estimate the coalescent times of hg N and its subclades with most up-to-date methods available at the time of the publication.

Goals of the second reference (**Ref. II**):

- Resolve the hg N phylogenetic tree with a total of 94 high-coverage Y chromosomal sequences.
- Calibrate the branching events within the resolved hg N tree using SNPs discovered from the high coverage sequences.
- Survey the geographical distribution of the detected subclades using a comprehensive sampling set of Eurasian populations.

Goals of the third reference (**Ref. III**):

- Examine the autosomal genetic diversity of Uralic speakers within a wider context of populations in northern Eurasia.
- Test for any recent shared genetic ancestry between Uralic-speaking populations.
- Date any potential Eastern gene flow into Western Uralic-speaking populations and their geographic neighbours.
- Examine genetic affinities of Uralic speakers and their geographic neighbours with available ancient samples from Eurasia.

4. SUBJECTS AND METHODS

The origin of the DNA samples analysed in this study is provided in the respective publications or their supplementary materials. The DNA samples were obtained from unrelated volunteers after receiving informed consent in accordance with the guidelines of the ethical committees of the institutions involved.

Experimental and computational methods used in this study are described in detail in the respective publications or their supplementary materials.

To summarise briefly: in **Ref. I**, a total of 5389 of samples was genotyped with either sequencing the polymorphic sites by Sanger method or assaying the variants with restriction-fragment length polymorphism approach. Haplogroup affiliations of additional 2630 samples were acquired from published literature. STRs were studied with the Y-Filer Kit and a median joining network was constructed with software Network 4.1.1.2.

In **Ref. II**, a total of 43 new Y chromosome sequences with >60× coverage were complemented with 54 previously published 40× coverage sequences to generate the phylogenetic tee of hg N using the software BEAST v.1.7.5. A panel of 6521 samples from across Eurasia was assembled and genotyped by Sanger sequencing in a hierarchical manner according to the updated phylogenetic relationships within the hg N. A total of 617 samples were genotyped for Y-STRs with Y-Filer Kit or the PowerPlex 23 Kit and the software Network 4.6.1.1 was used to construct a phylogenetic network.

In **Ref. III**, a total of 135 samples, of which 121 belonged to Uralic-speakers, were genotyped using the Illumina 610K, 650K, 660K or 1M SNPs arrays and analysed with additional data from 1665 samples acquired from published research. Subsequent filtering steps and analyses are described in detail in the respective section of **Ref. III**.

5. RESULTS AND DISCUSSION

Publications that are part of this study are ordered according to their publication dates. The first two articles reflect the accumulation and refinement of scientific knowledge on one of the major north Eurasian haplogroups – hg N. The third publication surveys the autosomal genetic variation of Uralic-speaking populations – major carriers of hg N – and tests whether the shared linguistic and patrilineal affiliation is reflected in their autosomal genetic composition. All three publications additionally examine a comprehensive set of neighbouring populations as human demographic history of any region cannot be reliably studied without placing the data in a relevant geographic context.

The following section summarises the main results and relevant discussion points, without recapturing them word by word. Full overview of results and their interpretations according to the knowledge available at the time of their publication is accessible in the articles. All three publications feature extensive supplementary information that, if presented in full, would by far exceed the space limits allocated for a printed copy of this thesis. Yet for readability and for highlighting certain results, some additional figures from electronic supplementary materials of original publications are included in the Supplementary Information section of this thesis. Links to full supplementary materials published online with each respective article are also provided in the Supplementary Information section of this thesis.

5.1. The north Eurasian dispersal of the Y-chromosomal haplogroup N and its time-resolved phylogeny (Ref. I and II)

Gradual accumulation of phylogenetically informative discovered polymorphisms (Zerjal et al., 1997; Underhill et al., 2000, 2001; Karafet et al., 2002; Cinnioglu et al., 2004) provided an initial toolbox to decipher the topology of hg N inner structure shown in Figure 12A (Section 2.2.3). Ref. I demonstrated hg N to have a widespread distribution, but reaching higher frequencies only in northern Eurasia. Due to the scarcity of diagnostically informative SNPs during the time this study was performed, the phylogenetic resolution remained at a limited level. Paragroup N* along with its precursor paragroup NO* was shown to be present at marginal frequencies in the modern populations of Japan and Southeast Asia and this serves as an indirect evidence for hgs NO, N and O to potentially originate from Mainland Southeast Asia or North China, albeit their current geographic spread patterns overlap only to a limited extent (Supplementary Table 1 in Ref. I). The notion of counter-clockwise northern dispersal of hg N is strengthened by the discovery of hg N in ancient samples from North

China and Lake Baikal (Cui et al., 2013; Gao et al., 2015; Zhang et al., 2017; de Barros Damgaard et al., 2018).

In Ref. II, we used a total of 94 high-coverage hg N samples to resolve the actual inner structure of the hg N and additional 3 hg O samples were included as an outgroup (Figure 1A in Ref. II). Within the hg N phylogeny, the deepest N5-B482 defined lineage was found in a sample of mixed origin and the next deep lineage (N4) is present in China and southeast Asia (Poznik et al., 2013; Lippold et al., 2014; Hallast et al., 2015; Karmin et al., 2015). Due to sampling limitations comprehensive information on frequency distribution of clades N5 and N4 is lacking, but data accessible on the pre-print level shows the presence of hg N among Han and Tibetan populations in China (Hu et al., 2015). Based on only 11 samples of Japanese and Chinese origin, the inner clades of hg N coalesce prior to the far more common subclades N2a and N3 (Figure 1A in Ref. II). This supports the notion of southern origin of hg N. Alternatively, this may suggest that harsher climate conditions with repeated population bottlenecks and prolonged low population size might have eliminated deeper patrilineages in populations that inhabit northern parts of Eurasia (Ref. II).

The STR-based coalescent age estimates in Ref. I were calculated with both evolutionary and pedigree mutation rates and as discussed in Section 2.1.4.5, differ significantly (Table 3 in this thesis). In Ref. II, we used the SNP-based approach to estimate the clade ages and for this purpose, placed the ~ 45 000-year-old aDNA specimen from Ust'-Ishim in western Siberia on the phylogenetic root of the extant NO-branch (Figure 1, inset in Ref. II). The average number of mutations from the tree root to tips along with analysed sequence length yielded a mutation rate of 0.76×10^{-9} substitutions per site per year, which is equal to that in Fu et al., (2014) and similar to Karmin et al., (2015). Age estimates of major hg N clades are given in Table 3 (STR-based ages of all hg N clades are given in Table 1 in Ref. I and the SNP-based ages in Supplementary Table S5 in Ref. II). As discussed in Section 2.1.4, coalescent ages based on SNP-mutation rate are considered the most reliable and are becoming the standard in population genetics field. Table 3 shows that STR-based age calculations with pedigree rate produce severely underestimated results, whereas 95% confidence intervals of evolutionary rate age estimates fall within the lower boundaries of SNP-based ages. However, sample size and sampling scheme significantly influence the evolutionary mutation rate results as demonstrated by the disparate STR-based hg N age estimates and wide confidence intervals of hg N3 (Table 3). In future population genetics studies, STRs will be used mostly for deciphering the inner structure of selected subclades and, to a lesser extent, ages of the youngest clades that have exhibited almost star-like expansion patterns.

Table 3. Age estimates of major hg N clades.
Nomenclature according to Ref.II, except for N2-A and N2-E that denote STR-defined clades in Ref.I. Ref.I estimates given in thousand years, Ref.II estimates in standard years.

Haplogroups	Evolutionary coalscent age estimates (Ref.I)	Pedigree-based coalescent age estimates (Ref.I)	SNP-based age estimates (Ref. II)	Lower 95% boundary (Ref.II)	Upper 95% boundary (Ref.II)
N	19.4 ± 4.8[a]	5.8±1.4[a]	25 313	22 764	27 934
N	14.2±4.0[b]	4.2±1.2[b]	-	-	-
N1'4	-	-	19 937	17 954	21 988
N2-A	6.2±2.0	1.8±0.6	-	-	-
N2-E	6.8±2.9	2.0±0.9	-	-	-
N2a	-	-	9314	7802	10888
N2a1	-	-	4727	4018	5502
N2a2	-	-	4909	3864	6158
N3	11.8±6.8	3.5±2.0	12 989	11 336	14 648
N3a2	-	-	4490	3594	5394
N3a3'6	-	-	4995	4353	5700

a - calculations based on data with Yakuts N3 and N2-E samples

b - calculations based on data without Yakut N3 and N2-E samples

Together with some previously published results (Kayser et al., 2003; Hammer et al., 2006b), our comprehensive sample set covering entire northern Eurasia revealed hg N3 to be the most common hg N subclade in this region, with east-west decline in frequencies per population (Figure 2g in Ref. I). However, it is important to note that the Native Siberian populations are by far smaller in absolute numbers than Finns, Estonians, Latvians and Lithuanians, making the actual number of men carrying hg N3 the highest in eastern Fennoscandia and the Baltics.

Analysis of STR length variation suggested the existence of several potential subclades within hg N3 (Ref. I), but this information remained largely elusive until our significantly deeper phylogenetical analysis, where complete re-sequencing of large set of hg N Y chromosomes was combined with extensive phylogeographic mapping of established subclades, accompanied by examination of STR variation (Ref. II). The phylogenetic resolution presented in Ref. II revealed a complex substructure with individual subclades showing distinct, in many cases surprisingly discrete geographic distribution (Figure 1AB in Ref. II, full phylogeographic survey presented in Table S2 in Supplementary Information of Ref. II). For example, the subclade N3b is restricted to Altai region in Southern Siberia (Figure 1B and Figure 3 in Ref. II). In contrast, hg N3a

subclades exemplify hg N spread across entire northern Eurasia (Figure 2 in Ref. II). The deepest subclade of N3a is N3a1-B211, which is mostly present in the Volga-Uralic region and in Khanty and Mansi populations in western Siberia (Figure 1B and Figure 3 in Ref. II). Next branching involves subclade N3a2-M2118, mostly found in Central Siberia, where over 80% of Sakha (Yakut) males belong to this clade along with the majority of Sakha Dolgans (however, Dolgans from Taymyr Peninsula have larger proportion of subclade N2a1-B478) and linguistically distant Evenks and Evens (Figure 3 and Table S2 in Ref. II). This clade is present at lower frequencies in western Siberia, where a distinct Y-STR pattern suggests an additional internal diversification within N3a2 (Ref. II). Hg N3a2 patrilineage is a prime example of a strong founder effect primarily in central Siberian populations, but has a deep branch represented by samples from Lebanon and China (Figure 1A in Ref. II) and additional data places one Turkish and Bhutan individual into the same subclade (Batini et al., 2015), indicating incidental occurrences in Near East and South Asia.

Hg N3a3'6, a neighbour clade to N3a2, has a remarkable geographical distribution with separate subclades present at significant frequencies in geographically most distant populations – Chukchi, Buryats, and Lithuanians, separated from each other at a distance over 6500 kilometres (Figure 1A in Ref. II). In Europe, N3a3 is present in about a third of present day Estonian, Latvian and Lithuanian men and is the dominant hg N subclade in Belarusians, Ukrainians and Southern Russians (Table S2 in Ref. II). It becomes less common in more northern populations such as Finns, Karelians, Saami, Vepsa and Northern Russians as well as Volga Tatars and Bashkirs all of whom harbour N3a4 as the main hg N clade (Table S2 and Figure 3 in Ref. II). In case of Northern Russians, this provides additional evidence of largely non-Slavic patrilineal genetic substrate among the Northern Russians (Balanovsky et al., 2008). These two clades – N3a3 and N3a4 – are nearly equally present in our Saami sample, suggesting the frequency pattern of the two lineages being shaped by random genetic drift in historically small populations (Table S2 in Ref. II).

In contrast to mostly European hgs N3a3 and N3a4, their neighbour clades N3a5 and N3a6 are clearly restricted to eastern parts of Eurasia (Figure 1B and Figure 3 in Ref. II). The latter of them is the only hg N lineage dominant among the Nanais from Lower reach of the Amur River and reflects a strong founder effect, but denser sampling in eastern Siberia might reveal this lineage in other small indigenous populations living in this region (Ref. II). Hg N3a5-B197 diversified soon after its inception into two distinct lineages: the N3a5-F4205 encompasses the Mongolic-speaking Buryats and Mongols living around Lake Baikal, whereas N3a5-B202 comprises of Chukotko-Kamchatkan-speaking Chukchis and Koryaks in Beringia and eastern Siberia. At present, the carriers of these two subclades are geographically far apart – about 5000 km (Ref. II).

The split of hg N3a3'6 occurred ~5 kya and high frequencies of this clade describe patrilineal pools of populations belonging to Altaic, Uralic, Indo-European, and Chukotko-Kamchatkan language families. Linguists have not agreed

upon a consensus regarding chronological connections between these linguistic phyla, although a recent study has proposed a potential separation of Uralic and Indo-European language families at roughly 10 kya (Kassian et al., 2019), indicating that it is safe to assume that the major language families diversified several thousand years prior to the emergence of hg N3a3'6. The initial spread of hg N3a3'6 clades most likely ignored any potential language barriers, but the succeeding diversification may have occurred further within already linguistically defined populations (Ref. II). Research on aDNA has suggested that some contemporary populations in northeastern Europe show higher levels East Asian influence than expected from their genetic composition featuring ancient northern Eurasian component (Lazaridis et al., 2014). This might hint at an additional Siberian gene flow represented by the westernmost hg N branches that mostly correspond to the present-day linguistic borders of Finno-Ugric languages (Ref. II).

It is currently difficult to pinpoint the precise region where the hg N3a3'6 radiation might have happened – possible regions include the vicinity of the Urals as the middle ground of their current geographic distribution or western Siberia with highest diversity of hg N sublineages (Figure 1B in Ref. II). It is however evident that the spread was remarkably quick and today covers the entire northern Eurasia, in accordance with the initially proposed "counterclockwise northern route" (Ref. I, II). A mid-Holocene warming in Siberia, that broadly coincides with the timeframe of N3a3'6 spread, allowed forests to advance northward (Monserud et al., 1998) and might have facilitated the rapid dispersal, but such parallels remain speculative.

The second widespread subclade of hg N is N2a with the majority of individuals belonging to its subclade N2a1-B523, with frequencies ranging from 10% to 30% in populations residing in western and southern Siberia, the Volga-Uralic region and northeastern Europe (Figure 1B in Ref. II). The initial network of N2 STR haplotypes presumed a bipartite distribution with separate European and Siberian subclusters (Ref. I), denoted respectively N2-E and N2-A, but the 19 re-sequenced hg N2a1 chromosomes revealed three separate subclades (Figure 1 in Ref. II). One consisting of primarily Siberian populations with a separate subbranch of three individuals (of Turkish, Arab and Afghan ethnic backgrounds), that is occasionally found in Mongols, Tatars and European Slavic-speakers (Table S2 in Ref. II). The formerly "European" STR cluster consists of two clades: N2a1-B528, spread more in the southern Volga-Uralic region and N2a1-L1419, spread mainly in northern part of that region (Table S2 in Ref. II).

The least frequent clade within hg N is N1, sparsely distributed in populations from Kazakhstan and Korea and in northern Han Chinese (Supplementary Table 1 in Ref. I). Its position on the resolved phylogenetic tree in Ref. II was verified by the open access database of YFull (YFull YTree v7.07.00), but additional sampling might reveal more regarding the geographic spread and phylogeny of this subclade.

5.2. Traces of common recent autosomal ancestry of most of the Uralic-speaking populations (Ref. III)

To test whether the sign of gene flow between East Eurasia and northeastern Europe detected in shared paternal lineages with a shallow coalescent age, is present in autosomal genetic variation, we first contextualised the autosomal genetic landscape of Uralic speakers and their geographic neighbours. The results of PCA and trees built from F_{ST} distances are primarily dependent on geographic distance, signifying geography as the main predictor of genetic affinity (Figure 2AB in Ref. III). At lower *K* values, ADMIXTURE shows the genetic resemblance of Uralic speakers to their geographic neighbours, but from *K*=9, a genetic component that is primarily, but not exclusively, shared among Uralic-speaking populations starts to appear (K9, magenta in Figure 3A in Ref. III). The spatial frequency of this component decreases rapidly from West Siberia towards east, south and west, with an exceptional peak in the Saami in Northern Scandinavia, who, in terms of proportion of this component, resemble the geographically more distant Finno-Ugric speakers in the Volga Uralic region. However, the westernmost Uralic speakers, Estonians and Hungarians, nearly lack the *K9* component, whereas Turkic speakers from the Volga-Uralic region display a significant portion (~20%) of it (Figure 3A in Ref. III). Correlation analysis between the geographic spread of hg N subclades relevant in the context of Uralic speakers and the *K9* component showed a weak, but significant correlation with subhaplogroups distributed near the Ural Mountains, but not with those that extend to Fennoscandia (Ref. III).

To examine further the tentative common genetic substrate, we calculated the D-statistic for westernmost Uralic speakers in order to assess the sharing pattern of derived alleles and additionally tested whether geographically separated Uralic-speaking groups share more IBD segments than their geographic neighbours, who speak non-Uralic languages. The most evident contrast between westernmost Uralic speakers (Saami, Finns, Estonians, Hungarians) and proximate European populations (Swedes, Poles, French) is an excess of Siberian-derived alleles in Saamis, Finns and, to a lesser extent, Estonians, but not in Hungarians (D-statistic calculated with Swedes presented in Figure S1 in Supplementary Information of this thesis, all D-statistic results presented in additional online files of Ref. III). Finns share more derived alleles with Siberians than Estonians do and the latter do not share more derived alleles with other Finno-Ugric, Saami or Ob-Ugric speaking populations than their southern Indo-European speaking Latvians (Figure S2 in Supplementary Information). However, in contrast to Latvians, Estonians share significantly more derived alleles with the Samoyed-speaking Nganasans and Nenets as well as with several non-Uralic Siberian populations (Figure S2 in Supplementary Information).

The IBD analysis showed Finnic-speakers in northeastern Europe to share more IBD segments with each other and, with the exception of Estonians, with their linguistic relatives in the Volga-Ural region and Siberia (Figure 4A in Ref. III). In case of Saami, Karelians and Vepsians IBD segments are also shared

with non-Uralic populations in Siberia (green dots in Figure 4A in Ref. III). A notable exception are the Estonians, whose IBD sharing level does not extend beyond the Uralic-speakers of northeastern Europe. Uralic-speaking Maris and Udmurts, in contrast to their geographic neighbours Chuvashes, Tatars and Bashkirs, display higher IBD sharing with Saami, Vepsians and northern Russians in the west and specifically only with their linguistic relatives to the east of the Ural Mountains (Figure 4C in Ref. III). No excessive IBD sharing with linguistic relatives or Siberian populations was detected in Hungarians (Figure 4B in Ref. III). These results suggest that there is at least some, although detectable at variable levels in different populations, common genetic substrate among the analysed Uralic populations and it also tends to extend to geographically closest non-Uralic speaking neighbours (Ref. III). The shared genetic component suggests that the spread of Uralic languages was at least to some degree associated with the movement of people. In contemporary populations it is however more pronounced between Uralic speakers from the Volga-Uralic region and northeastern Europe, and between West Siberian Uralic speakers and the Volga-Uralic region, whereas it is virtually non-detectable in Hungarians, Mordovians and Estonians.

The outlier status of Hungarians might reflect a bias in our test group as our Hungarian sample set is drawn from the capital region. A comprehensive sampling across Hungary might reveal existing ties to linguistic relatives as was recently shown in Post et al., (2019). Moreover, studies involving Hungarian aDNA samples from the time period of their arrival into the Carpathian basin have identified hg N3 lineages and East Asian mtDNA haplogroups, testifying of traces of a real migration of people with links to eastern Eurasia (Neparáczki et al., 2017, 2019). Similarly, Early Iron Age aDNA samples from *tarand* graves in Estonia carrying the first found hg N3a lineages from the Eastern Baltic region are associated with a westward migration from the east into the region (Saag et al., 2019).

5.2.1. Globetrotter and ALDER admixture dating analysis

To capture more recent signals of shared demographic history in the Uralic-speaking peoples that are more relevant in the context of language expansions within the last 5000 years, we employed fineSTRUCTURE to cluster individuals according to patterns of haplotype sharing similarity. Formed clusters (presented in Figure 3S in Supplementary Information) were further used in Globetrotter analysis (Figure 5AB in Ref. III). The identified clusters mostly correspond to self-identified ethnic groups and higher hierarchical levels follow geographic proximity. As expected, the identified admixture events in full Globetrotter analysis involve contacts between geographically close source populations and thus, in “regional” analysis, we excluded neighbours from the set of possible donors and allowed copying only from individuals with a different group affiliation (Ref. III).

The “Finnic”, “Saami” and both European clusters show admixture from West Siberian sources (the “Komi” cluster), although it is more prominent (above 3%) in the former two (Figure 5B in Ref. III). In the Volga-Uralic region, Uralic speakers display admixture mostly from “Europe1” cluster (containing primarily Baltic/Russian individuals) and Uralic-speaking Siberian donors (“Khanty-Mansi”, “Samoyed” clusters) (Figure 5B in Ref. III). This contrasts with geographically close Turkic-speakers, who show admixture with diverse European clusters and have received substantial contribution from East Asian/South Siberian groups (“E-Asia/S-Sib” cluster) (Figure 5B in Ref. III). The contacts detected between Uralic speakers show mostly unidirectional east-to-west donating pattern, for example Komis are a major donating component in “Finnic” and “Saami” groups, but none of the latter two contribute much to admixture within the “Komi” group (Ref. III).

Uralic speakers in Western Siberia have a complex admixture pattern with a multitude of donors, but the most distinct difference is the presence of East Asian/South Siberian (“E-Asia/S-Sib” cluster) component specifically in the Samoyedic speakers (“Samoyed” and “Nganassan” cluster), whereas it is not present in the “Khanty-Mansi” group (Figure 5B in Ref. III).

The time depth for admixture events inferred by Globetrotter and ALDER are relatively recent (within the last 2000 years) and broadly coincide (comparison Table S12 available from online additional file 13 of Ref. III). The most statistically significant admixture events identified by ALDER comprise of eastern and western surrogate donor populations and form a notable temporal pattern (Figure 19): admixture in Uralic speakers from northeastern Europe and the Volga-Uralic region is dated as the oldest (800–900 CE or older), this is followed by admixture in Turkic speakers from Volga-Uralic region (~1200 – 900 CE), whereas the timeframe for admixture in Siberian populations is the youngest (>1500 CE). This suggests that the western geneflow into Siberia has been relatively recent, while eastern Eurasian influx into northeastern Europe is much older and has probably taken place within the first millennium CE.

Earliest published occurrence of autosomal Siberian ancestry in Fennoscandia is from ~3500-year-old samples in Bolshoy Oleni Ostrov in eastern Fennoscandia coupled with Y-chromosomal hg N3a3’6 (marker L392) and mtDNA haplogroups that are common in modern Siberia (Lamnidis et al., 2018). In the Baltics, earliest published indications of Siberian ancestry along with hg N3a3 appear considerably later in samples dated ~500 BCE and excavated from the territory of modern Estonia (Saag et al., 2019). This might suggest multiple admixture events that brought Siberian-related ancestry into northeastern Europe at different time periods with dating approaches providing a single averaged estimate – this is also in accord with complex admixture history showed by Globetrotter analysis. Moreover, admixture is not an instantaneous process and might have continuously taken place over several thousands of years. The earliest Siberian component precedes the diversification of existing Uralic languages spoken in Fennoscandia, making it hard to connect this ancestry with carriers of any known Uralic tongue. However, the age of the

first known hg N3a3a samples from the Baltic region coincides with the diversification of the Finnic branch (Figure 17 in Section 2.4.1), which allows to tentatively suggest an association with the dispersal of Uralic languages into the region.

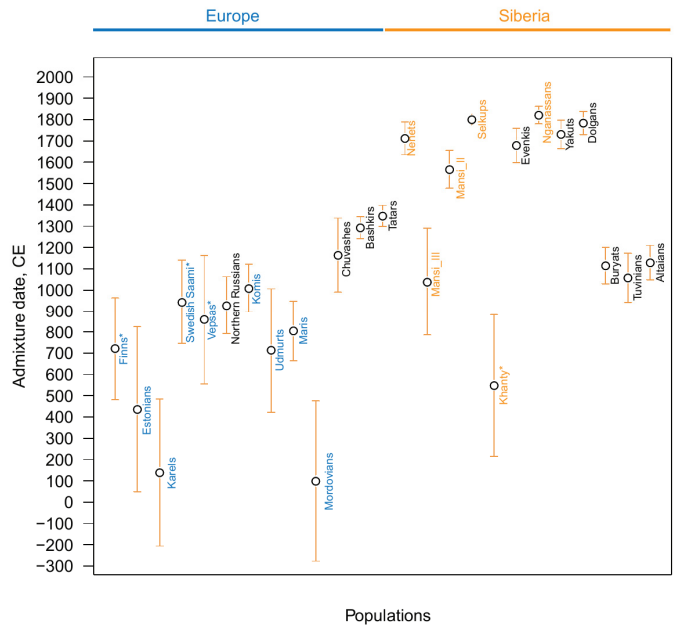

Figure 19. Admixture dates for the eastern and western components of the Uralic-speaking populations (highlighted according to geography blue for Europe and orange for Siberia) in the context of their geographical neighbours on an absolute time scale. Dates are calculated with ALDER according to decay rates of two-reference weighted linkage disequilibrium curve using the generation time of 30 years. Black circles show point estimates and error bars indicate 95% confidence intervals. Admixture dates before Common Era (CE) are shown with a negative sign. (*) indicates admixed populations with inconsistent LD curve decay rates. Adapted from Ref.III, under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

5.2.2. Affinities of the Uralic speakers with ancient Eurasians

We calculated the f3-statistic to measure the extent of shared drift between modern and ancient Eurasians. As reported previously (Lazaridis et al., 2014), northeastern Europeans including Uralic speakers share more drift with any Mesolithic hunter-gatherer group than Central or Western Europeans (Figures S9A-C in online additional file 3 of Ref. III). The results however revealed that the Uralic speakers from the Volga-Ural region and the Saami share more drift with Eastern hunter-gatherers, whereas Finnic speakers share more with Wes-

tern hunter-gatherers (Figure 4S in Supplementary Information). We also tested alternative demographic scenarios that fit the observed genetic diversity of extant Finno-Ugric populations in a model involving four known European ancestral components (Western hunter-gatherers, Eastern hunter-gatherers, early farmers, steppe people of Yamnaya/Corded Ware culture) and a novel fifth Siberian component (Figure 6A in Ref. III). The latter was approximated by the Nganasans who show the least admixture from Western Eurasia. The estimated Siberian contribution was the main component in Western Siberian Uralic speakers and constitutes one third of the genomes in modern Volga-Uralic populations and the Saami, whereas it drops to 5% in Estonians and is almost non-existent in Latvians and Lithuanians (Figure 6B in Ref. III). The only Uralic speakers that did not fit into the tested model with five ancestral components were the Hungarians.

To explain the demographic history of populations in northeastern Europe an inclusion of previously unknown Siberian component is thus needed. The geographic distribution of the main part of this component is likely associated with the extant Uralic speakers, but it is not confined exclusively to them and reflects complex historical ties connecting northeastern Europe and Siberia. The shared ancestry is more pronounced in the current Y chromosome gene pool of hg N and the tested Uralic-speaking populations show marginal, though significant, higher affinity in the autosomal loci to populations with high frequency of the relevant hg N subclade than predicted from their X chromosomal similarity and in comparison to other populations where this subclade is rare or absent (Figure S13 in Supplementary Information in Ref. III). Such sex-specific differences might reflect their shared population history through complex socio-cultural factors amplified by small effective population sizes and genetic drift.

6. CONCLUSIONS

- Sporadic occurrence of Y-chromosomal hg NO* and the presence of deeply diverged subclades in Southeast Asia and Japan in addition to hg N aDNA samples from North China, suggest that hg N, currently a major patrilineage in northern Eurasia, may have originated in North China or Mainland Southeast Asia.
- Chronological phylogeny built on the basis of sequenced high-coverage hg N sequences added several new subclades and resolved previously known subhaplogroups N2 (N1b) and N3 (N1c) into discrete lineages that arose at various times and have distinctive geographical distributions.
- Majority of N2a individuals belong to subhaplgroup N2a1-B523 with a coalescent age ~5 kya, which covers western and southern Siberia, the Taimyr Peninsula and the Volga-Uralic region.
- Within subhaplogroup N3, several subclades are distinctly localised, but the most wide-spread north Eurasian subclade is N3a3'6, coalescing at about 5 kya. It is the dominant hg N subhaplogroup in populations ranging from Baltic-Fennoscandian Europe and Volga-Uralic region, to Mongolia and the Pacific Russian Beringia, exhibiting a nearly star-like spread with distances between different subclades spanning 5000–6500 km. Its carriers became dominant among speakers of major north Eurasian families, such as Chukotko-Kamchatkan, Altaic, Uralic and several Indo-European Balto-Slavic populations.
- Autosomal genetic variation of Uralic speakers is foremost dependent on geography.
- A small, yet significant autosomal genetic component of possibly Siberian origin is shared between most of the Uralic-speaking populations, suggesting a migration of people to be at least to some extent responsible for the current dispersal of Uralic languages. However, this component is not exclusive to Uralic speakers and extends also to some of their geographic neighbours, while being lower in Estonians and virtually absent in Hungarians.
- In order to explain the autosomal genetic composition of contemporary northeastern Europe, a novel Siberian component needs to be invoked. The precise timeframe for its arrival remains elusive, but genetic admixture pattern derived from contemporary populations suggests the recent eastern Eurasian influx into northeastern Europe to be older than that of western influence on the genetic structure of Siberian population.

SUMMARY IN ESTONIAN

Uurali rahvaste geneetiline ajalugu läbi isaliini N ja autosoomse varieeruvuse prisma

Uurali keeli kõneleb maailmas umbes 25 miljonit inimest, arvuliselt suurem osa neist elab Kesk- ja Kirde-Euroopas ning nad erinevad keeleliselt enamikust Euroopa rahvastest, kes kõnelevad indoeuroopa keeli. Vähemaarvulisem uurali keelte kõnelejate hulk elab Lääne-Siberis, sealhulgas ka Taimõri poolsaarel. Arheoloogias puudub konsensus seostamaks uurali keelte laia levikut mõne kindla arheoloogilise kultuuriga ning on jäänud selgusetuks, kas keeled levisid inimeste rände kaudu või oli tegu pigem kultuurilise edasikandega.

Inimese Y-kromosoomi bioloogiline roll on suunata embrüo arenemist meessoost organismiks ja evolutsioon on vorminud Y-kromosoomist ainult isaliinis ehk isalt pojale päranduva geneetilise lookuse. Valdav osa Y-kromosoomist on haploidne, mis, sarnaselt teise inimgenoomi haploidse lookusega – vaid emaliinis päranduva mitokondriaalse DNA-ga, tähendab rekombinatsiooni mittetoimumist ning seega tervikuna edasikandumist järgmisesse põlvkonda. Y-kromosoomi muutlikkust põhjustavad eri tüüpi mutatsioonid, kuid käesoleva dissertatsiooni raames on peamiseks uurimisobjektiks aeglase mutatsioonikiirusega ühe nukleotiidi muutust põhjustavad punktmutatsioonid (inglise keeles *single nucleotide polymorphism*, SNP). Enamik punktmutatsioone nii Y-kromosoomis kui teistes genoomi osades on tekkinud üks kord anatoomiliselt kaasaegse inimese evolutsiooni jooksul.

Võrreldes rahvusvaheliselt kokkulepitud referents-genoomiga, on iga inimese genoomis sadu tuhandeid punktmutatsioone. Kuna inimese Y-kromosoom moodustab vaid ligikaudu 2% kogugenoomist, on erinevuste koguarv Y-kromosoomis arusaadavalt märksa tagasihoidlikum, kuid tänu rekombinatsioonivabale pärandumisele piki meesliini saab samu mutatsioone kandvatest Y-kromosoomidest moodustada fülogeneetiliselt informatiivseid hulkasid ehk nn. haplogruppe – ühe kindla meessoost eellasega ja kõiki selle järglasi hõlmavaid Y-kromosoomi järjestusvariante ehk monofüleetilisi klaade. Haplogrupid võib omakorda siduda jagatud mutatsioonide põhjal fülogeneetilisteks puudeks ning nii ilmnevad ajas edukalt edasikandunud katkematud, ühemõtteliselt klassifitseeritavad isaliinid.

Piisav hulk mutatsioone võimaldab tuvastada haplogrupisiseste alamklaadide tõelist harunemisjärjekorda ning puu korrektse topoloogia selgitamine omakorda hinnata ka iga hargnemise ligikaudset tekkeaega ning tänapäevani jälgitava ekspansiooni algust. See on saanud võimalikuks alles hiljaaegu tänu uudse DNA järjestust määrata võimaldava tehnoloogia leiutamisele.

Kui Y-kromosoomi markerite uurimine annab infot vaid indiviidi isaliini kohta, siis ülegenoomsetes autosoomiandmetes sisaldub teoreetiliselt teave kõikide indiviidi genoomi kunagi panustanud esivanemate kohta. Rekombinatsiooni tõttu ei võimalda autosoomsed punktmutatsioonid samas moodustada selgelt interpreteeritavaid puulaadseid struktuure ning populatsioonide demograa-

filise ajaloo rekonstrueerimiseks autosoomides sisalduva info põhjal on välja töötatud mitmeid uusi statistilisi meetodeid.

Nii mitokondriaalsete emaliinide kui Y-kromosoomi isaliinide jaotus sõltub eelkõige geograafiast: Aasiat ja Euroopat iseloomustavad erinevad haplogrupid. Sellest mustrist erineb selgesti haplogupp N, mis seob Euroopa ja Aasia põhjaaladel elavaid rahvaid ning iseloomustab uurali keelte kõnelejaid Kirde-Euroopast Siberini, haarates lisaks osaliselt ka Hiinat ja Jaapani.

Käesoleva töö üks eesmärk oli selgitada haplogrupi N fülogeneetilise puu sisemist struktuuri, kasutades selleks uusimaid täisgenoomide sekveneerimismeetodeid, määrata alamklaadide ekspansiooniaegu, hinnata meie töös avastatud uute alamklaadide esinemissagedust Põhja-Euraasia rahvaste seas ning võrrelda alamklaadide levikumustrit suurimate Põhja-Euraasia keelkondade levikuga. Töö teiseks eesmärgiks oli uurida uurali keeli kõnelevate rahvaste ülegenoomset mitmekesisust teiste tänapäeva Põhja-Euraasiat asustavate rahvaste kontekstis ning tuvastada potentsiaalset geneetilist ühisosa, hindamaks uurali keelte levikuviisi ning ka -suunda ja -aega.

Käesoleva töö põhilised tulemused ja järeldused on järgmised:

- Põhja-Euraasia üks sagedasemaid meesliine – haplogrupp N – on tõenäoliselt pärit Põhja-Hiinast või mandrilisest Kagu-Aasiast, kus on säilinud sügava koalestsentsiajaga alamklaade ning haplogruppi NO* kuuluvaid isaliine.
- Kõrge katvusega sekveneeritud Y-kromosoomi täisjärjestustest konstrueeritud haplogrupi N fülogeneetiline puu võimaldas tuvastada mitmeid uusi sügava lahknemisajaga alamklaade ning lahutas juba tuntud alamklaadide N2 ja N3 (teised nimevariandid vastavalt N1b ja N1c) sisemise struktuuri mitmeks kindla geograafilise levikuga isaliinihulgaks.
- Enamik haplogrupi N2 kandjaid kuulub N2a1-B523 alamklaadi koalestsentsajaga 5000 aastat. Alamklaad on levinud Lääne- ja Lõuna-Siberis, Taimõri poolsaarel ning Volga-Uurali piirkonnas, kuid ulatub üllatuslikult ka läänemeresoome keelkonda kuuluvate vepsadeni, puududes samas neile keeleliselt ja geograafiliselt lähedastel karjala meestel.
- Haplogruppi N3 (eelnevalt tuntud ka kui hg N1c-Tat) kuulub mitmeid piiritletud levikuga isaliine, milles suurima geograafilise ulatusega alamklaad on N3a3’6 koalestsentsiajaga 5000 aastat. See on põhiline haplogrupi N hulka kuuluv isaliin Läänemere idakaldal, Fennoskandias, Volga-Uurali piirkonnas, Mongoolias ning ka mõnedel Amuuri jõe suudmeala ja Vene Beringia rahvastel. Arvestades N3a3’6 alamklaadide lühikest lahknemisaega oli nende levik kiire ning suure geograafilise ulatusega – tänapäeval elavad osad selle isaliinigrupi eri alamklaadide kandjad teineteisest enam kui 5000 km kaugusel, kuuludes kõikidesse peamistesse Põhja-Euraasia keelkondadesse – altai, mongoli, uurali, tšuktši-kamtšadali, indoeuroopa.
- Uurali keeli rääkivate rahvaste autosoomne mitmekesisus on määratud eelkõige nende geograafilise asukohaga Euraasias.

- Võrreldes naaberrahvastega jagab enamus uurali keeli kõnelejatest piiratud, kuid tuvastatavat hiljutist tõenäoliselt Siberi päritolu geneetilist ühisosa, mis vihjab keelte levikuga kaasas käinud inimeste rändele. Samas on see ühine geneetiline komponent levinud ka osade altai ja indoeuroopa keeli kõnelevate rahvaste seas ning on seotud ulatuslikumate ida poolt lähtunud mõjutustega kaasaegsete Kirde-Euroopa rahvaste geenifondile, mis ei piirdunud ainult uurali keeli rääkinud rahvaste rändega.

Lõpetuseks – käesolevasse doktoritöösse kaasatud uuringute tulemusena on võimalik selgemini formuleerida lahtisi küsimusi kõigi Põhja-Euraasia populatsioonide geneetilise ajaloo kohta. Nii näiteks on perspektiivne haplogrupi N kõrval senisest märksa sügavama geneetilise ulatusega uurida ka teiste isaliinide ajalis-ruumilist jaotust. Teiseks on ilmne, et täisgenoomsete andmete analüüsimeetodid täienevad pidevalt ja juba kogutud andmete interpretatsioon saab olla jätkuvalt süvenev. Kolmandaks tuleb lisada, et vana-DNA uuringud, seda eriti Siberi aladelt, on alles varajase arengu faasis ning võimalike tulemuste tõlgendamise lähtekohaks saab olla vaid adekvaatne teadmine tänapäevaste populatsioonide geneetilisest varieeruvusest.

REFERENCES

Abbott, J.K., Nordén, A.K., and Hansson, B. 2017. Sex chromosome evolution: historical insights and future perspectives. *Proceedings of the Royal Society B: Biological Sciences* 284: 20162806.

Abe-Sandes, K., Silva, W.A.J., and Zago, M.A. 2004. Heterogeneity of the Y chromosome in Afro-Brazilian populations. *Human Biology* 76: 77–86.

Abondolo, D. 2017. Uralic Languages. *Oxford Handbooks Online*.

Aikio, Á. 2012. An essay on Saami ethnolinguistic prehistory. *A Linguistic Map of Prehistoric Northern Europe* 63–117 (Suomalais-Ugrilaisen Seuran Toimituksia 266).

Albrechtsen, A., Nielsen, F.C., and Nielsen, R. 2010. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution* 27: 2534–2547.

Alexander, D.H., Novembre, J., and Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19: 1655–1664.

Anthony, D.W. 2007. The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World. (Princeton: Princeton University Press).

Arakawa, Y., Nishida-Umehara, C., Matsuda, Y., Sutou, S., and Suzuki, H. 2002. X-chromosomal localization of mammalian Y-linked genes in two XO species of the Ryukyu spiny rat. *Cytogenet Genome Res* 99: 303–309.

Avise, J.C. 2000. Phylogeography: the history and formation of species (Cambridge, Mass: Harvard University Press).

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., and Saunders, N.C. 1987. Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics* 18: 489–522.

Bachtrog, D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics* 14: 113–124.

Bachtrog, D., Mank, J.E., Peichel, C.L., Kirkpatrick, M., Otto, S.P., et al. 2014. Sex Determination: Why So Many Ways of Doing It? *PLoS Biology* 12: e1001899.

Balanovsky, O. 2017. Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. *Human Genetics* 136: 575–590.

Balanovsky, O., Gurianov, V., Zaporozhchenko, V., Balaganskaya, O., Urasin, V., et al. 2017. Phylogeography of human Y-chromosome haplogroup Q3-L275 from an academic/citizen science collaboration. *BMC Evolutionary Biology* 17:.

Balanovsky, O., Rootsi, S., Pshenichnov, A., Kivisild, T., Churnosov, M., et al. 2008. Two Sources of the Russian Patrilineal Heritage in Their Eurasian Context. *The American Journal of Human Genetics* 82: 236–250.

Balanovsky, O., Zhabagin, M., Agdzhoyan, A., Chukhryaeva, M., Zaporozhchenko, V., et al. 2015. Deep Phylogenetic Analysis of Haplogroup G1 Provides Estimates of SNP and STR Mutation Rates on the Human Y-Chromosome and Reveals Migrations of Iranic Speakers. *PLOS ONE* 10: e0122968.

Balaresque, P., Bowden, G.R., Adams, S.M., Leung, H.-Y., King, T.E., et al. 2010. A Predominantly Neolithic Origin for European Paternal Lineages. *PLoS Biology* 8: e1000285.

Balaresque, P., Manni, F., Dugoujon, J.M., Crousau-Roy, B., and Heyer, E. 2006. Estimating sex-specific processes in human populations: Are XY-homologous markers an effective tool? *Heredity* 96: 214.

Balinova, N., Post, H., Kushniarevich, A., Flores, R., Karmin, M., et al. 2019. Y-chromosomal analysis of clan structure of Kalmyks, the only European Mongol people, and their relationship to Oirat-Mongols of Inner Asia. *European Journal of Human Genetics.* 27: 1466–1474.

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., et al. 2010. Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *The American Journal of Human Genetics* 87: 341–353.

de Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J.V., Kroonen, G., et al. 2018. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360: eaar7711.

Batini, C., Hallast, P., Zadik, D., Delser, P.M., Benazzo, A., et al. 2015. Large-scale recent expansion of European patrilineages shown by population resequencing. *Nature Communications* 6:7152.

Battaglia, V., Grugni, V., Perego, U.A., Angerhofer, N., Gomez-Palmieri, J.E., Woodward, S.R., Achilli, A., Myres, N., Torroni, A., and Semino, O. 2013. The First Peopling of South America: New Evidence from Y-Chromosome Haplogroup Q. *PLoS ONE* 8: e71390.

Bergero, R., Forrest, A., Kamau, E., and Charlesworth, D. 2007. Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: evidence from new sex-linked genes. *Genetics* 175: 1945–1954.

Bergman, I., Olofsson, A., Hörnberg, G., Zackrisson, O., and Hellberg, E. 2004. Deglaciation and Colonization: Pioneer Settlements in Northern Fennoscandia. *Journal of World Prehistory* 18: 155–177.

Bosch, E., Calafell, F., Rosser, Z.H., Norby, S., Lynnerup, N., Hurles, M.E., and Jobling, M.A. 2003. High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by. *Human Genetics* 112: 353–363.

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J.G., and Bustamante, C.D. 2012. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* 84: 343–364.

Browning, B.L., and Browning, S.R. 2013a. Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. *The American Journal of Human Genetics* 93: 840–851.

Browning, B.L., and Browning, S.R. 2013b. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* 194: 459–471.

Browning, S.R. 2008. Estimation of Pairwise Identity by Descent From Dense Genetic Marker Data in a Population Sample of Haplotypes. *Genetics* 178: 2123–2132.

Browning, S.R., and Browning, B.L. 2015. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *The American Journal of Human Genetics* 97: 404–418.

Brumfield, R.T., Beerli, P., Nickerson, D.A., and Edwards, S.V. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* 18: 249–256.

Burton, M.L., Moore, C.C., Romney, A.K., Aberle, D.F., Barcelo, J.A., Dow, M.M., Guyer, J.I., Kronenfeld, D.B., Levy, J.E., and Linnekin, J. 1996. Regions Based on Social Structure. *Current Anthropology* 37: 87–123.

Busby, G.B.J., Brisighelli, F., Sanchez-Diz, P., Ramos-Luis, E., Martinez-Cadenas, C., et al. 2012. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proceedings of the Royal Society B: Biological Sciences* 279: 884–892.

Busby, G.B.J., Hellenthal, G., Montinaro, F., Tofanelli, S., Bulayeva, K., et al. 2015. The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Current Biology* 25: 2518–2526.

Calafell, F., and Larmuseau, M.H.D. 2017. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Human Genetics* 136: 559–573.

Carpelan, C., and Parpola, A. 2001. Emergence, contacts, and dispersal of Proto-Indo-European, Proto-Uralic and Proto-Aryan in archaeological perpective. In Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations p.55–151, (Suomalais-Ugrilainen Seura).

Carvajal-Carmona, L.G., Soto, I.D., Pineda, N., Ortíz-Barrientos, D., Duque, C., et al. 2000. Strong Amerind/White Sex Bias and a Possible Sephardic Contribution among the Founders of a Population in Northwest Colombia. *The American Journal of Human Genetics* 67: 1287–1295.

Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., Bishop, C., Fellous, M., Purrello, M., Fiori, G., and Siniscalco, M. 1985. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* 230: 1403–1406.

Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. 1994. The history and geography of human genes (Princeton, N.J: Princeton University Press).

Cheng, A.Y., Teo, Y.-Y., and Ong, R.T.-H. 2014. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* 30: 1707–1713.

Chiaroni, J., Underhill, P.A., and Cavalli-Sforza, L.L. 2009. Y chromosome diversity, human expansion, drift, and cultural evolution. *Proceedings of the National Academy of Sciences* 106: 20174–20179.

Cinnioglu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S., et al. 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Human Genetics* 114: 127–148.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15: 1496–1502.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P.D., Grützner, F., and Kaessmann, H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* 508: 488–493.

Crawford, J.E., and Lazzaro, B.P. 2012. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front Genet* 3: 66.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., et al. 2004. Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa. *The American Journal of Human Genetics* 74: 1014–1022.

Cruciani, F., La Fratta, R., Trombetta, B., Santolamazza, P., Sellitto, D., et al. 2007. Tracing Past Human Male Movements in Northern/Eastern Africa and Western Eurasia: New Clues from Y-Chromosomal Haplogroups E-M78 and J-M12. *Molecular Biology and Evolution* 24: 1300–1311.

Cui, Y., Li, H., Ning, C., Zhang, Y., Chen, L., Zhao, X., Hagelberg, E., and Zhou, H. 2013. Y Chromosome analysis of prehistoric human populations in the West Liao River Valley, Northeast China. *BMC Evolutionary Biology* 13: 216.

Decker, A.E., Kline, M.C., Redman, J.W., Reid, T.M., and Butler, J.M. 2008. Analysis of mutations in father–son pairs with 17 Y-STR loci. *Forensic Science International: Genetics* 2: e31–e35.

Derenko, M., Malyarchuk, B., Denisova, G., Wozniak, M., Grzybowski, T., Dambueva, I., and Zakharov, I. 2007. Y-chromosome haplogroup N dispersals from south Siberia to Europe. *Journal of Human Genetics* 52: 763–770.

Di Giacomo, F., Luca, F., Popa, L.O., Akar, N., Anagnou, N., et al. 2004. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Human Genetics* 115: 357–371.

Dudás, E., Vágó-Zalán, A., Vándor, A., Saypasheva, A., Pomozi, P., and Pamjav, H. 2019. Genetic history of Bashkirian Mari and Southern Mansi ethnic groups in the Ural region. *Mol Genet Genomics* 294: 919–930.

Dulik, M.C., Zhadanov, S.I., Osipova, L.P., Askapuli, A., Gau, L., Gokcumen, O., Rubinstein, S., and Schurr, T.G. 2012. Mitochondrial DNA and Y Chromosome Variation Provides Evidence for a Recent Common Ancestry between Native Americans and Indigenous Altaians. *The American Journal of Human Genetics* 90: 229–246.

Eberhard, D.M., Simons, G.F., and Fenning, C.D. 2019. Ethnologue: Languages of the World. (Dallas, Texas, USA: SIL International).

Edwards, S.V., and Beerli, P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54: 1839–1854.

Elhaik, E., Tatarinova, T.V., Klyosov, A.A., and Graur, D. 2014. The 'extremely ancient' chromosome that isn't: a forensic bioinformatic investigation of Albert Perry's X-degenerate portion of the Y chromosome. *European Journal of Human Genetics* 22: 1111–1116.

Eller, E. 2001. Effects of ascertainment bias on recovering human demographic history. *Human Biology* 73: 411–427.

Falush, D., Stephens, M., and Pritchard, J.K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.

Fedorova, S.A., Reidla, M., Metspalu, E., Metspalu, M., Rootsi, S., et al. 2013. Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evolutionary Biology* 13: 127.

Fenner, J.N. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 128: 415–423.

Finocchio, A., Trombetta, B., Messina, F., D'Atanasio, E., Akar, N., Loutradis, A., Michalodimitrakis, E.I., Cruciani, F., and Novelletto, A. 2018. A finely resolved phylogeny of Y chromosome Hg J illuminates the processes of Phoenician and Greek colonizations in the Mediterranean. *Scientific Reports* 8:7465.

Francalacci, P., Morelli, L., Angius, A., Berutti, R., Reinier, F., et al. 2013. Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs European Y-Chromosome Phylogeny. *Science* 341: 565–569.

Francalacci, P., Sanna, D., and Useli, A. 2016. Human Y Chromosome Mutation Rate: Problems and Perspectives. In Anthropology: Current and Future Developments, M.

Lima, A. Ramos, and C. Santos, eds. (BENTHAM SCIENCE PUBLISHERS), pp. 65–91.
Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514: 445–449.
Gao, S.-Z., Zhang, Y., Wei, D., Li, H.-J., Zhao, Y.-B., Cui, Y.-Q., and Zhou, H. 2015. Ancient DNA reveals a migration of the ancient Di-qiang populations into Xinjiang as early as the early Bronze Age: The Origin Of The Tianshanbeilu Population. *Am. J. Phys. Anthropol.* 157: 71–80.
Ge, J., Budowle, B., Aranda, X.G., Planz, J.V., Eisenberg, A.J., and Chakraborty, R. 2009. Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci Int Genet* 3: 179–184.
Goedbloed, M., Vermeulen, M., Fang, R.N., Lembring, M., Wollstein, A., et al. 2009. Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR® Yfiler® PCR amplification kit. *International Journal of Legal Medicine* 123: 471–482.
Gravel, S. 2012. Population Genetics Models of Local Ancestry. *Genetics* 191: 607–619.
Graves, J.A. 1995. The origin and function of the mammalian Y chromosome and Y-borne genes--an evolving understanding. *Bioessays* 17: 311–320.
Graves, J.A.M. 2016a. Did sex chromosome turnover promote divergence of the major mammal groups?: De novo sex chromosomes and drastic rearrangements may have posed reproductive barriers between monotremes, marsupials and placental mammals. *BioEssays* 38: 734–743.
Graves, J.A.M. 2016b. Evolution of vertebrate sex chromosomes and dosage compensation. *Nature Reviews Genetics* 17: 33–46.
Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., et al. 2010. A Draft Sequence of the Neandertal Genome. *Science* 328: 710–722.
Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. 2008. Whole population, genome-wide mapping of hidden relatedness. *Genome Research* 19: 318–326.
Gusmão, L., Sánchez-Diz, P., Calafell, F., Martín, P., Alonso, C.A., et al. 2005. Mutation rates at Y chromosome specific microsatellites. *Human Mutation* 26: 520–528.
Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522: 207–211.
Haasl, R.J., and Payseur, B.A. 2010. Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity* 106: 158.
Häkkinen, J. 2009. Kantauralin ajoitus ja paikannus: perustelut puntarissa. *Journal de la Société Finno-Ougrienne* 48.
Hallast, P., Batini, C., Zadik, D., Maisano Delser, P., Wetton, J.H., et al. 2015. The Y-Chromosome Tree Bursts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades. *Molecular Biology and Evolution* 32: 661–673.
Hammer, M.F., Chamberlain, V.F., Kearney, V.F., Stover, D., Zhang, G., Karafet, T., Walsh, B., and Redd, A.J. 2006a. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic Science International* 164: 45–55.
Hammer, M.F., Karafet, T., Rasanayagam, A., Wood, E.T., Altheide, T.K., Jenkins, T., Griffiths, R.C., Templeton, A.R., and Zegura, S.L. 1998. Out of Africa and back

again: nested cladistic analysis of human Y chromosome variation. *Molecular Biology and Evolution* 15: 427–441.
Hammer, M.F., Karafet, T.M., Park, H., Omoto, K., Harihara, S., Stoneking, M., and Horai, S. 2006b. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *Journal of Human Genetics* 51: 47–58.
Harris, A.M., and DeGiorgio, M. 2017. Admixture and Ancestry Inference from Ancient and Modern Samples through Measures of Population Genetic Drift. *Human Biology* 89: 21.
Hasselmann, M., Gempe, T., Schiøtt, M., Nunes-Silva, C.G., Otte, M., and Beye, M. 2008. Evidence for the evolutionary nascence of a novel sex determination pathway in honeybees. *Nature* 454: 519–522.
Heath, S.C., Gut, I.G., Brennan, P., McKay, J.D., Bencko, V., et al. 2008. Investigation of the fine structure of European populations with applications to disease association studies. *European Journal of Human Genetics* 16: 1413–1429.
Helgason, A., Einarsson, A.W., Guðmundsdóttir, V.B., Sigurðsson, Á., Gunnarsdóttir, E.D., Jagadeesan, A., Ebenesersdóttir, S.S., Kong, A., and Stefánsson, K. 2015. The Y-chromosome point mutation rate in humans. *Nature Genetics* 47: 453–457.
Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. 2014. A Genetic Atlas of Human Admixture History. *Science* 343: 747–751.
Henn, B.M., Gignoux, C., Lin, A.A., Oefner, P.J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S.A., Mountain, J.L., and Underhill, P.A. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proceedings of the National Academy of Sciences* 105: 10693–10698.
Hey, J., and Machado, C.A. 2003. The study of structured populations — new hope for a difficult and divided science. *Nature Reviews Genetics* 4: 535–543.
Heyer, E., Chaix, R., Pavard, S., and Austerlitz, F. 2012. Sex-specific demographic behaviours that shape human genomic variation. *Molecular Ecology* 21: 597–612.
Ho, S., and Larson, G. 2006. Molecular clocks: when timesare a-changin'. *Trends in Genetics* 22: 79–83.
Hohoff, C., Dewa, K., Sibbing, U., Hoppe, K., Forster, P., and Brinkmann, B. 2007. Y-chromosomal microsatellite mutation rates in a population sample from north-western Germany. *International Journal of Legal Medicine* 121: 359–363.
Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J., Syrjänen, K., and Wahlberg, N. 2013. Cultural and climatic changes shape the evolutionary history of the Uralic languages. *J. Evol. Biol.* 26: 1244–1253.
Hu, K., Yan, S., Liu, K., Ning, C., Wei, L.-H., et al. 2015. The dichotomy structure of Y chromosome Haplogroup N. *ArXiv:1504.06463 [q-Bio]*.
Hudjashov, G., Endicott, P., Post, H., Nagle, N., Ho, S.Y.W., et al. 2018. Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles. *Scientific Reports* 8: 1823.
Hudjashov, G., Kivisild, T., Underhill, P.A., Endicott, P., Sanchez, J.J., et al. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proceedings of the National Academy of Sciences* 104: 8726–8730.
Hughes, J.F., Skaletsky, H., Brown, L.G., Pyntikova, T., Graves, T., et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* 483: 82–86.
Hurles, M.E., Irven, C., Nicholson, J., Taylor, P.G., Santos, F.R., Loughlin, J., Jobling, M.A., and Sykes, B.C. 1998. European Y-Chromosomal Lineages in Polynesians: A

Contrast to the Population Structure Revealed by mtDNA. *The American Journal of Human Genetics* 63: 1793–1806.
Hurles, M.E., and Jobling, M.A. 2001. Haploid chromosomes in molecular ecology: lessons from the human Y. *Molecular Ecology* 10: 1599–1613.
Huyghe, J.R., Fransen, E., Hannula, S., Van Laer, L., Van Eyken, E., Mäki-Torkko, E., Aikio, P., Sorri, M., Huentelman, M.J., and Camp, G.V. 2011. A genome-wide analysis of population structure in the Finnish Saami with implications for genetic association studies. *European Journal of Human Genetics* 19: 347–352.
Ivanovic, R.F., Gregoire, L.J., Kageyama, M., Roche, D.M., Valdes, P.J., Burke, A., Drummond, R., Peltier, W.R., and Tarasov, L. 2016. Transient climate simulations of the deglaciation 21–9 thousand years before present (version 1) – PMIP4 Core experiment design and boundary conditions. *Geoscientific Model Development* 9: 2563–2587.
Janhunen, J. 2009. Proto-Uralic—what, where, and when? *The Quasquicentennial of the Finno-Ugrian Society. Suomalais-Ugrilaisen Seuran Toimituksia. Mémoires de La Société Finno-Ougrienne.* 57–78.
Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B., and Jin, L. 2012. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.* 22: 519–527.
Jobling, M., Tyler-Smith, C., Hollox, E., Hurles, M.E., and Kivisild, T. 2013. Human Evolutionary Genetics (Garland Science).
Jobling, M.A., Pandya, A., and Tyler-Smith, C. 1997. The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med* 110: 118–124.
Jobling, M.A., and Tyler-Smith, C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics* 4: 598–612.
Jobling, M.A., and Tyler-Smith, C. 2017. Human Y-chromosome variation in the genome-sequencing era. *Nature Reviews Genetics* 18: 485–497.
Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., et al. 2015. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications* 6:8912.
Joukhadar, R., Daetwyler, H.D., Bansal, U.K., Gendall, A.R., and Hayden, M.J. 2017. Genetic Diversity, Population Structure and Ancestral Origin of Australian Wheat. *Frontiers in Plant Science* 8: 2115.
Kallio, P. 2006. On the absolute chronology of proto-languages of Finnish. *Virittäjä* 110: 24.
Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R.S., Redd, A.J., Zegura, S.L., and Hammer, M.F. 2001. Paternal Population History of East Asia: Sources, Patterns, and Microevolutionary Processes. *The American Journal of Human Genetics* 69: 615–628.
Karafet, T., Zegura, S.L., Vuturo-Brady, J., Posukh, O., Osipova, L., et al. 1997. Y chromosome markers and trans-Bering Strait dispersals. *American Journal of Physical Anthropology* 102: 301–314.
Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L., and Hammer, M.F. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research* 18: 830–838.
Karlsson, A.O., Wallerström, T., Götherström, A., and Holmlund, G. 2006. Y-chromosome diversity in Sweden – A long-time perspective. *European Journal of Human Genetics* 14: 963–970.

Karmin, M., Saag, L., Vicente, M., Sayres, M.A.W., Järve, M., et al. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research* 25: 459–466.

Kassian, A., Zhivlov, M., Starostin, G., A Trofimov, A., A Kocharov, P., Kuritsyna, A., and N Saenko, M. 2019. Rapid radiation of the Inner Indo-European languages: an advanced approach to Indo-European lexicostatistics [pre-print accepted in Diachronica 2020].

Katoh, T., Munkhbat, B., Tounai, K., Mano, S., Ando, H., et al. 2005. Genetic features of Mongolian ethnic groups revealed by Y-chromosomal analysis. *Gene* 346: 63–70.

Kayser, M. 2017. Forensic use of Y-chromosome DNA: a general overview. *Human Genetics* 136: 621–635.

Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., et al. 2006. Melanesian and Asian Origins of Polynesians: mtDNA and Y Chromosome Gradients Across the Pacific. *Mol Biol Evol* 23: 2234–2244.

Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P., Oefner, P., Tommaseo-Ponzetta, M., and Stoneking, M. 2003. Reduced Y-Chromosome, but Not Mitochondrial DNA, Diversity in Human Populations from West New Guinea. *The American Journal of Human Genetics* 72: 281–302.

Kayser, M., Lao, O., Anslinger, K., Augustin, C., Bargel, G., et al. 2005. Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Human Genetics* 117: 428–443.

Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., and McVean, G. 2019. Inferring whole-genome histories in large population datasets. *Nature Genetics* 51: 1330–1338.

Keller, A., Graefen, A., Ball, M., Matzas, M., Boisguerin, V., et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications* 3: 698.

Kere, J. 2001. Human population genetics. Lessons from Finland. *Annu. Rev. Genom. Hum. Genet.* 2: 103–128.

Khrunin, A.V., Khokhrin, D.V., Filippova, I.N., Esko, T., Nelis, M., et al. 2013. A Genome-Wide Analysis of Populations from European Russia Reveals a New Pole of Genetic Diversity in Northern Europe. *PLoS ONE* 8: e58552.

Kingman, J.F.C. 1982. The coalescent. *Stochastic Processes and Their Applications* 13: 235–248.

Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D., and Long, J.C. 1998. Dual Origins of Finns Revealed by Y Chromosome Haplotype Variation. *The American Journal of Human Genetics* 62: 1171–1179.

Kivisild, T. 2017. The study of human Y chromosome variation through ancient DNA. *Human Genetics* 136: 529–546.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., et al. 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *The American Journal of Human Genetics* 72: 313–332.

de Knijff, P. 2000. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *The American Journal of Human Genetics* 67: 1055–1061.

Knowles, L.L. 2003. The burgeoning field of statistical phylogeography: Statistical phylogeography. *Journal of Evolutionary Biology* 17: 1–10.

Korhonen, M. 1981. Johdatus lapin kielen historiaan (Helsinki: Suomalaisen Kirjallisuuden Seura).
Krausz, C., and Casamonti, E. 2017. Spermatogenic failure and the Y chromosome. *Human Genetics* 136: 637–655.
Kuroki, Y., Toyoda, A., Noguchi, H., Taylor, T.D., Itoh, T., et al. 2006. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nature Genetics* 38: 158–167.
Kushniarevich, A., Utevska, O., Chuhryaeva, M., Agdzhoyan, A., Dibirova, K., et al. 2015. Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. *PLOS ONE* 10: e0135820.
Kuzmin, Y.V. 2008. Siberia at the Last Glacial Maximum: Environment and Archaeology. *Journal of Archaeological Research* 16: 163–221.
Lachance, J., and Tishkoff, S.A. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it: Prospects & Overviews. *BioEssays* 35: 780–786.
Lahermo, P., Savontaus, M.-L., Sistonen, P., Béres, J., de Knijff, P., Aula, P., and Sajantila, A. 1999. Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. *European Journal of Human Genetics* 7: 447–458.
Lahn, B.T., and Page, D.C. 1999. Four evolutionary strata on the human X chromosome. *Science* 286: 964–967.
Lamnidis, T.C., Majander, K., Jeong, C., Salmela, E., Wessman, A., et al. 2018. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nature Communications* 9: 5018.
Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., et al. 2008. Correlation between Genetic and Geographic Structure in Europe. *Current Biology* 18: 1241–1248.
Lappalainen, T., Koivumäki, S., Salmela, E., Huoponen, K., Sistonen, P., Savontaus, M.-L., and Lahermo, P. 2006. Regional differences among the Finns: A Y-chromosomal perspective. *Gene* 376: 207–215.
Lappalainen, T., Laitinen, V., Salmela, E., Andersen, P., Huoponen, K., Savontaus, M.-L., and Lahermo, P. 2008. Migration Waves to the Baltic Sea Region. *Annals of Human Genetics* 72: 337–348.
Lawson, D.J., van Dorp, L., and Falush, D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications* 9: 3258.
Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. 2012. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics* 8: e1002453.
Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513: 409–413.
Lell, J.T., Sukernik, R.I., Starikovskaya, Y.B., Su, B., Jin, L., Schurr, T.G., Underhill, P.A., and Wallace, D.C. 2002. The Dual Origin and Siberian Affinities of Native American Y Chromosomes. *The American Journal of Human Genetics* 70: 192–206.
Lemaitre, C., Braga, M.D.V., Gautier, C., Sagot, M.-F., Tannier, E., and Marais, G.A.B. 2009. Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in Human Sex Chromosomes. *Genome Biology and Evolution* 1: 56–66.
Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., et al. 2015. The fine-scale genetic structure of the British population. *Nature* 519: 309–314.

Li, N., and Stephens, M. 2003. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* 165: 2213.

Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., Schröder, R., and Stoneking, M. 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. 17.

Lipson, M., Loh, P.-R., Levin, A., Reich, D., Patterson, N., and Berger, B. 2013. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Molecular Biology and Evolution* 30: 1788–1802.

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S.A., Tesfaigzi, Y., and Bruse, S. 2013. Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics* 7: 1.

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. 2013. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics* 193: 1233–1254.

Lohmueller, K.E., Bustamante, C.D., and Clark, A.G. 2009. Methods for Human Demographic Inference Using Haplotype Patterns From Genomewide Single-Nucleotide Polymorphism Data. *Genetics* 182: 217–231.

Luo, Z.-X., Yuan, C.-X., Meng, Q.-J., and Ji, Q. 2011. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* 476: 442–445.

Maan, A.A., Eales, J., Akbarov, A., Rowland, J., Xu, X., Jobling, M.A., Charchar, F.J., and Tomaszewski, M. 2017. The Y chromosome: a blueprint for men's health? *European Journal of Human Genetics* 25: 1181–1188.

Malyarchuk, B., and Derenko, M. 2009. On the origin of Y-chromosome haplogroup N1b. *European Journal of Human Genetics* 17: 1540–1541.

Malyarchuk, B., Derenko, M., Denisova, G., Wozniak, M., Grzybowski, T., Dambueva, I., and Zakharov, I. 2010. Phylogeography of the Y-chromosome haplogroup C in northern Eurasia: Y-chromosome haplogroup C phylogeography. *Annals of Human Genetics* 74: 539–546.

Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. 2013. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* 93: 278–288.

Mardis, E.R. 2010. The $1,000 genome, the $100,000 analysis? *Genome Medicine* 2: 84.

Mardis, E.R. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470: 198–203.

Martin, A.R., Karczewski, K.J., Kerminen, S., Kurki, M.I., Sarin, A.-P., et al. 2018. Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. *The American Journal of Human Genetics* 102: 760–775.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., et al. 2018. The genomic history of southeastern Europe. *Nature* 555: 197–203.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528: 499–503.

McElreavey, K., Ravel, C., Chantot-Bastaraud, S., and Siffroi, J.-P. 2006. Y chromosome variants and male reproductive function. *Int J Androl* 29: 298–303; discussion 304–306.

McEvoy, B.P., Montgomery, G.W., McRae, A.F., Ripatti, S., Perola, M., et al. 2009. Geographical structure and differential natural selection among North European populations. *Genome Research* 19: 804–814.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.

McTavish, E.J., and Hillis, D.M. 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC Genomics* 16:.

Mendez, F.L., Krahn, T., Schrack, B., Krahn, A.-M., Veeramah, K.R., et al. 2013. An African American Paternal Lineage Adds an Extremely Ancient Root to the Human Y Chromosome Phylogenetic Tree. *The American Journal of Human Genetics* 92: 454–459.

Mirabal, S., Regueiro, M., Cadenas, A.M., Cavalli-Sforza, L.L., Underhill, P.A., Verbenko, D.A., Limborska, S.A., and Herrera, R.J. 2009a. Y-Chromosome distribution within the geo-linguistic landscape of northwestern Russia. *European Journal of Human Genetics* 17: 1260–1273.

Mirabal, S., Underhill, P.A., and Herrera, R.J. 2009b. Reply to B Malyarchuk and M Derenko: a need for further investigation of Uralic and Siberian populations in the search for haplogroup N1b's origins. *European Journal of Human Genetics* 17: 1541–1543.

Mittnik, A., Wang, C.-C., Pfrengle, S., Daubaras, M., Zariņa, G., et al. 2018. The genetic prehistory of the Baltic Sea region. *Nature Communications* 9:.

Monserud, R.A., Tchebakova, N.M., and Denissenko, O.V. 1998. Reconstruction of the mid-Holocene palaeoclimate of Siberia using a bioclimatic vegetation model. *Palaeogeography, Palaeoclimatology, Palaeoecology* 139: 15–36.

Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., and Reich, D. 2011. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genet* 7: e1001373.

Moorjani, P., Sankararaman, S., Fu, Q., Przeworski, M., Patterson, N., and Reich, D. 2016. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proceedings of the National Academy of Sciences* 113: 5652–5657.

Murtagh, V.J., O'Meally, D., Sankovic, N., Delbridge, M.L., Kuroki, Y., et al. 2012. Evolutionary history of novel genes on the tammar wallaby Y chromosome: Implications for sex chromosome evolution. *Genome Research* 22: 498–507.

Myres, N.M., Rootsi, S., Lin, A.A., Järve, M., King, R.J., et al. 2011. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *European Journal of Human Genetics* 19: 95–101.

Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., et al. 2009. Genetic Structure of Europeans: A View from the North–East. *PLoS ONE* 4: e5472.

Neparáczki, E., Juhász, Z., Pamjav, H., Fehér, T., Csányi, B., et al. 2017. Genetic structure of the early Hungarian conquerors inferred from mtDNA haplotypes and Y-chromosome haplogroups in a small cemetery. *Mol Genet Genomics* 292: 201–214.

Neparáczki, E., Maróti, Z., Kalmár, T., Maár, K., Nagy, I., et al. 2019. Y-chromosome haplogroups from Hun, Avar and conquering Hungarian period nomadic people of the Carpathian Basin (Genetics).

Nielsen, R. 2004. Population genetic analysis of ascertained SNP data. *Human Genomics* 1: 218.

Nielsen, R., and Beaumont, M.A. 2009. Statistical inferences in phylogeography. *Molecular Ecology* 18: 1034–1047.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* 12: 443–451.
Novembre, J. 2016. Pritchard, Stephens, and Donnelly on Population Structure. *Genetics* 204: 391–393.
Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., et al. 2008. Genes mirror geography within Europe. *Nature* 456: 98–101.
Novembre, J., and Ramachandran, S. 2011. Perspectives on Human Population Structure at the Cusp of the Sequencing Era. *Annual Review of Genomics and Human Genetics* 12: 245–274.
Novembre, J., and Stephens, M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* 40: 646–649.
O'Dushlaine, C., McQuillan, R., Weale, M.E., Crouch, D.J.M., Johansson, Å., et al. 2010. Genes predict village of origin in rural Europe. *Eur J Hum Genet* 18: 1269–1270.
Ohno, S. 1967. Sex chromosomes and sex-linked genes. (Monographs on endocrinology, Vol. 1.). (Berlin, Heidelberg, New York: Springer Verlag.).
Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., and Stoneking, M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics* 29: 20.
van Oven, M., Van Geystelen, A., Kayser, M., Decorte, R., and Larmuseau, M.H. 2014. Seeing the Wood for the Trees: A Minimal Reference Phylogeny for the Human Y Chromosome. *Human Mutation* 35: 187–191.
Översti, S., Onkamo, P., Stoljarova, M., Budowle, B., Sajantila, A., and Palo, J.U. 2017. Identification and analysis of mtDNA genomes attributed to Finns reveal long-stagnant demographic trends obscured in the total diversity. *Sci Rep* 7: 6193.
Padhukasahasram, B. 2014. Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics* 5:.
Palo, J.U., Ulmanen, I., Lukka, M., Ellonen, P., and Sajantila, A. 2009. Genetic markers and population history: Finland revisited. *Eur J Hum Genet* 17: 1336–1346.
Pamjav, H., Fehér, T., Németh, E., and Pádár, Z. 2012. Brief communication: New Y-chromosome binary markers improve phylogenetic resolution within haplogroup R1a1. *American Journal of Physical Anthropology* 149: 611–615.
Pankratov, V., Litvinov, S., Kassian, A., Shulhin, D., Tchebotarev, L., et al. 2016. East Eurasian ancestry in the middle of Europe: genetic footprints of Steppe nomads in the genomes of Belarusian Lipka Tatars. *Scientific Reports* 6: 30197.
Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., 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: e1001371.
Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. 2012. Ancient Admixture in Human History. *Genetics* 192: 1065–1093.
Peltonen, L., Pekkarinen, P., and Aaltonen, J. 1995. Messages from an isolate: lessons from the Finnish gene pool. *Biol. Chem. Hoppe-Seyler* 376: 697–704.
Pfeifer, S.P. 2017. From next-generation resequencing reads to a high-quality variant data set. *Heredity* 118: 111–124.
Pickrell, J.K., and Pritchard, J.K. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics* 8: e1002967.

Pimenoff, V.N., Comas, D., Palo, J.U., Vershubsky, G., Kozlov, A., and Sajantila, A. 2008. Northwest Siberian Khanty and Mansi in the junction of West and East Eurasian gene pools as revealed by uniparental markers. *Eur J Hum Genet* 16: 1254–1264.

Pitulko, V., Pavlova, E., and Nikolskiy, P. 2017. Revising the archaeological record of the Upper Pleistocene Arctic Siberia: Human dispersal and adaptations in MIS 3 and 2. *Quaternary Science Reviews* 165: 127–148.

Pitulko, V.V., Nikolsky, P.A., Girya, E.Yu., Basilyan, A.E., Tumskoy, V.E., Koulakov, S.A., Astakhov, S.N., Pavlova, E.Yu., and Anisimov, M.A. 2004. The Yana RHS Site: Humans in the Arctic Before the Last Glacial Maximum. *Science* 303: 52–56.

Pitulko, V.V., Tikhonov, A.N., Pavlova, E.Y., Nikolskiy, P.A., Kuper, K.E., and Polozov, R.N. 2016. Early human presence in the Arctic: Evidence from 45,000-year-old mammoth remains. *Science* 351: 260–263.

Pliss, L., Tambets, K., Loogvali, E.-L., Pronina, N., Lazdins, M., Krumina, A., Baumanis, V., and Villems, R. 2006. Mitochondrial DNA Portrait of Latvians: Towards the Understanding of the Genetic Structure of Baltic-Speaking Populations. *Annals of Human Genetics* 70: 439–458.

Pool, J.E., Hellmann, I., Jensen, J.D., and Nielsen, R. 2010. Population genetic inference from genomic sequence variation. *Genome Research* 20: 291–300.

Post, H., Németh, E., Klima, L., Flores, R., Fehér, T., et al. 2019. Y-chromosomal connection between Hungarians and geographically distant populations of the Ural Mountain region and West Siberia. *Scientific Reports* 9: 7786.

Poznik, G.D., Henn, B.M., Yee, M.-C., Sliwerska, E., Euskirchen, G.M., et al. 2013. Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females. *Science* 341: 562–565.

Poznik, G.D., Xue, Y., Mendez, F.L., Willems, T.F., Massaia, A., et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature Genetics* 48: 593–599.

Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. 2009. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics* 5: e1000519.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Pugach, I., Matveev, R., Spitsyn, V., Makarov, S., Novgorodov, I., Osakovsky, V., Stoneking, M., and Pakendorf, B. 2016. The Complex Admixture History and Recent Southern Origins of Siberian Populations. *Mol Biol Evol* 33: 1777–1795.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81: 559–575.

Quinto-Cortés, C.D., Woerner, A.E., Watkins, J.C., and Hammer, M.F. 2018. Modeling SNP array ascertainment with Approximate Bayesian Computation for demographic inference. *Scientific Reports* 8: 10209.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505: 87–91.

Ralph, P., and Coop, G. 2013. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biology* 11: e1001555.

Rappold, G.A. 1993. The pseudoautosomal regions of the human sex chromosomes. *Human Genetics* 92: 315–324.
Reich, D., Price, A.L., and Patterson, N. 2008. Principal component analysis of genetic data. *Nature Genetics* 40: 491–492.
Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. 2009. Reconstructing Indian population history. *Nature* 461: 489–494.
Rens, W., O'Brien, P.C.M., Grutzner, F., Clarke, O., Graphodatskaya, D., et al. 2007. The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biology* 8:R243.
Rice, W.R. 1987. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced Recombination between primitive sex chromosomes. *Evolution* 41: 911–914.
Rice, W.R. 1996. Evolution of the Y Sex Chromosome in Animals. *BioScience* 46: 331–343.
Roesti, M., Moser, D., and Berner, D. 2013. Recombination in the threespine stickleback genome – patterns and consequences. *Mol Ecol* 22: 3014–3027.
Róna-Tas, A. 1999. Hungarians and Europe in the early Middle Ages: an introduction to early Hungarian history (Budapest ; New York: Central European University Press).
Rootsi, S., Kivisild, T., Benuzzi, G., Help, H., Bermisheva, M., et al. 2004. Phylogeography of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe. *The American Journal of Human Genetics* 75: 128–137.
Rootsi, S., Kivisild, T., Tambets, K., Adojaan, M., Parik, J., Reidla, M., Metspalu, E., Laos, S., Tolk, H.-V., and Villems, R. 2000. On the phylogeographic context of sex-specific genetic markers of Finno-Ugric populations. In The Roots of Peoples and Languages of Northern Eurasia II and III, (Tartu: University of Tartu, Division of Uralic Languages/Societas Historiae Fenno-Ugricae), pp. 148–164.
Rootsi, S., Myres, N.M., Lin, A.A., Järve, M., King, R.J., et al. 2012. Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. *European Journal of Human Genetics* 20: 1275–1282.
Rosenberg, N.A., and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics* 3: 380–390.
Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., et al. 2005. The DNA sequence of the human X chromosome. *Nature* 434: 325–337.
Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., et al. 2000. Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language. *The American Journal of Human Genetics* 67: 1526–1543.
Rozen, S., Skaletsky, H., Marszalek, J.D., Minx, P.J., Cordum, H.S., Waterston, R.H., Wilson, R.K., and Page, D.C. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423: 873–876.
Saag, L., Laneman, M., Varul, L., Malve, M., Valk, H., et al. 2019. The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Current Biology* 29: 1701–1711.e16.
Saag, L., Varul, L., Scheib, C.L., Stenderup, J., Allentoft, M.E., et al. 2017. Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. *Current Biology* 27: 2185–2193.e6.
Sajantila, A., Salem, A.H., Savolainen, P., Bauer, K., Gierig, C., and Paabo, S. 1996. Paternal and maternal DNA lineages reveal a bottleneck in the founding of the

Finnish population. *Proceedings of the National Academy of Sciences* 93: 12035–12039.
Salmela, E., Lappalainen, T., Fransson, I., Andersen, P.M., Dahlman-Wright, K., et al. 2008. Genome-Wide Analysis of Single Nucleotide Polymorphisms Uncovers Population Structure in Northern Europe. *PLoS ONE* 3: e3519.
Salminen, T. 1999. Euroopan kielet muinoin ja nykyisin. In Pohjan poluilla. Suomalaisten juuret nykytutkimuksen mukaan. Bidrag till kännedom av Finlands natur och folk., (Helsinki: Suomen Tiedeseura), pp. 13–26.
Sandstedt, S.A., and Tucker, P.K. 2004. Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome. *Genome Research* 14: 267–272.
Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. 2008. Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics* 82: 290–303.
Schraiber, J.G., and Akey, J.M. 2015. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics* 16: 727–740.
Scozzari, R., Massaia, A., Trombetta, B., Bellusci, G., Myres, N.M., Novelletto, A., and Cruciani, F. 2014. An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. *Genome Res.* 24: 535–544.
Seguin-Orlando, A., Korneliussen, T.S., Sikora, M., Malaspinas, A.-S., Manica, A., et al. 2014. Genomic structure in Europeans dating back at least 36,200 years. *Science* 346: 1113–1118.
Seielstad, M.T., Minch, E., and Cavalli-Sforza, L.L. 1998. Genetic evidence for a higher female migration rate in humans. *Nature Genetics* 20: 278.
Semino, O. 2000. The Genetic Legacy of Paleolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective. *Science* 290: 1155–1159.
Sengupta, S., Zhivotovsky, L.A., King, R., Mehdi, S.Q., Edmonds, C.A., et al. 2006. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *The American Journal of Human Genetics* 78: 202–221.
Shi, H., Dong, Y., Wen, B., Xiao, C.-J., Underhill, P.A., Shen, P., Chakraborty, R., Jin, L., and Su, B. 2005. Y-chromosome evidence of southern origin of the East Asian–specific haplogroup O3-M122. *The American Journal of Human Genetics* 77: 408–419.
Shi, W., Ayub, Q., Vermeulen, M., Shao, R. -g., Zuniga, S., van der Gaag, K., de Knijff, P., Kayser, M., Xue, Y., and Tyler-Smith, C. 2010. A Worldwide Survey of Human Male Demographic History Based on Y-SNP and Y-STR Data from the HGDP-CEPH Populations. *Molecular Biology and Evolution* 27: 385–393.
Sikora, M., Pitulko, V., Sousa, V., Allentoft, M.E., Vinner, L., et al. 2018. The population history of northeastern Siberia since the Pleistocene. *BioRxiv* 448829.
Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423: 825–837.
Skov, L., Consortium, T.D.P.G., and Schierup, M.H. 2017. Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLOS Genetics* 13: e1006834.

Soraggi, S., Wiuf, C., and Albrechtsen, A. 2018. Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. *G3:Genes|Genomes|Genetics* 8: 551–566.

Speidel, L., Forest, M., Shi, S., and Myers, S.R. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics* 51: 1321–1329.

Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., et al. 1999. Y-Chromosome Evidence for a Northward Migration of Modern Humans into Eastern Asia during the Last Ice Age. *The American Journal of Human Genetics* 65: 1718–1724.

Svendsen, J. 2004. Late Quaternary ice sheet history of northern Eurasia. *Quaternary Science Reviews* 23: 1229–1271.

Syrjänen, K., Honkola, T., Korhonen, K., Lehtinen, J., Vesakoski, O., and Wahlberg, N. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *DIA* 30: 323–352.

Szinnyei, J. 1910. Finnisch-ugrische Sprachwissenschaft (Leipzig: Göschen).

Tambets, K., Rootsi, S., Kivisild, T., Help, H., Serk, P., et al. 2004. The Western and Eastern Roots of the Saami—the Story of Genetic "Outliers" Told by Mitochondrial DNA and Y Chromosomes. *The American Journal of Human Genetics* 74: 661–682.

Tambets, K., Rootsi, S., Kivisild, T., and Villems, R. 2001. The concepts of Richard Indreko about the origin of the Finno-Ugric speakers and the population genetics of the extant north-east European populations. *Journal of the Humanities and Social Sciences* Vol 5: 59–74.

Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. 2006. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. *The American Journal of Human Genetics* 79: 1–12.

Tang, H., Peng, J., Wang, P., and Risch, N.J. 2005. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology* 28: 289–301.

Tataru, P., Nirody, J.A., and Song, Y.S. 2014. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics* 30: 3430–3431.

Thangaraj, K., Singh, L., Reddy, A.G., Rao, V.R., Sehgal, S.C., Underhill, P.A., Pierson, M., Frame, I.G., and Hagelberg, E. 2003. Genetic Affinities of the Andaman Islanders, a Vanishing Human Population. *Current Biology* 13: 86–93.

The Y Chromosome Consortium 2002. A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups. *Genome Research* 12: 339–348.

Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J., and Feldman, M.W. 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences* 97: 7360–7365.

Triska, P., Chekanov, N., Stepanov, V., Khusnutdinova, E.K., Kumar, G.P.A., et al. 2017. Between Lake Baikal and the Baltic Sea: genomic history of the gateway to Europe. *BMC Genet* 18: 110.

Trombetta, B., and Cruciani, F. 2017. Y chromosome palindromes and gene conversion. *Human Genetics* 136: 605–619.

Trombetta, B., Cruciani, F., Underhill, P.A., Sellitto, D., and Scozzari, R. 2010. Footprints of X-to-Y gene conversion in recent human evolution. *Mol Biol Evol* 27: 714–725.

Trombetta, B., D'Atanasio, E., Massaia, A., Ippoliti, M., Coppa, A., et al. 2015a. Phylogeographic Refinement and Large Scale Genotyping of Human Y Chromosome Haplogroup E Provide New Insights into the Dispersal of Early Pastoralists in the African Continent. *Genome Biology and Evolution* 7: 1940–1950.

Trombetta, B., D'Atanasio, E., Massaia, A., Myres, N.M., Scozzari, R., Cruciani, F., and Novelletto, A. 2015b. Regional Differences in the Accumulation of SNPs on the

Male-Specific Portion of the Human Y Chromosome Replicate Autosomal Patterns: Implications for Genetic Dating. *PLOS ONE* 10: e0134646.

Trombetta, B., Sellitto, D., Scozzari, R., and Cruciani, F. 2014. Inter- and Intraspecies Phylogenetic Analyses Reveal Extensive X–Y Gene Conversion in the Evolution of Gametologous Sequences of Human Sex Chromosomes. *Molecular Biology and Evolution* 31: 2108–2123.

Tumonggor, M.K., Karafet, T.M., Hallmark, B., Lansing, J.S., Sudoyo, H., Hammer, M.F., and Cox, M.P. 2013. The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. *Journal Of Human Genetics* 58: 165.

Underhill, P.A., and Kivisild, T. 2007. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* 41: 539–564.

Underhill, P.A., Myres, N.M., Rootsi, S., Metspalu, M., Zhivotovsky, L.A., et al. 2009. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *European Journal Of Human Genetics* 18: 479–484.

Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Mirazón Lahr, M., Foley, R.A., Oefner, P.J., and Cavalli-Sforza, L.L. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* 65: 43–62.

Underhill, P.A., Poznik, G.D., Rootsi, S., Järve, M., Lin, A.A., et al. 2015. The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *European Journal of Human Genetics* 23: 124–131.

Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., et al. 2000. Y chromosome sequence variation and the history of human populations. *Nature Genetics* 26: 358–361.

Van Geystelen, A., Decorte, R., and Larmuseau, M.H.D. 2013. Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs. *Forensic Science International: Genetics* 7: 573–580.

Veyrunes, F., Waters, P.D., Miethke, P., Rens, W., McMillan, D., et al. 2008. Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Research* 18: 965–973.

Villems, R., Adojaan, M., Kivisild, T., Metspalu, E., Parik, J., Pielberg, G., Rootsi, S., Tambets, K., and Tolk, H.-V. 1998. Reconstruction of Maternal lineages of Finno-Ugric speaking people and some remarks on their Paternal inheritance. In The Roots of Peoples and Languages of Northern Eurasia I, (Turku: Societas Historiae Fenno-Ugricae), pp. 180–198.

Wang, C.-C., Gilbert, M.T.P., Jin, L., and Li, H. 2014a. Evaluating the Y chromosomal timescale in human demographic and lineage dating. *Investigative Genetics* 5: 12.

Wang, C.-C., and Li, H. 2013. Inferring human history in East Asia from Y chromosomes. *Investigative Genetics* 4: 11.

Wang, C.-C., Yan, S., Qin, Z.-D., Lu, Y., Ding, Q.-L., et al. 2013. Late Neolithic expansion of ancient Chinese revealed by Y chromosome haplogroup O3a1c-002611. *Journal of Systematics and Evolution* 51: 280–286.

Wang, S.R., Agarwala, V., Flannick, J., Chiang, C.W.K., Altshuler, D., et al. 2014b. Simulation of Finnish Population History, Guided by Empirical Genetic Data, to Assess Power of Rare-Variant Tests in Finland. *The American Journal of Human Genetics* 94: 710–720.

Wang, Y., Zhang, Y., Zhang, C., Li, R., Yang, Y., Ou, X., Tong, D., and Sun, H. 2016. Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han

population from Guangdong Province, Southern China. *Forensic Science International: Genetics* 21: 5–9.
Wangkumhang, P., and Hellenthal, G. 2018. Statistical methods for detecting admixture. *Current Opinion in Genetics & Development* 53: 121–127.
Waters, P.D., Duffy, B., Frost, C.J., Delbridge, M.L., and Graves, J.A. 2001. The human Y chromosome derives largely from a single autosomal region added to the sex chromosomes 80–130 million years ago. *Cytogenet. Cell Genet.* 92: 74–79.
Wei, L.-H., Wang, L.-X., Wen, S.-Q., Yan, S., Canada, R., et al. 2018a. Paternal origin of Paleo-Indians in Siberia: insights from Y-chromosome sequences. *European Journal of Human Genetics* 26: 1687–1696.
Wei, L.-H., Yan, S., Lu, Y., Wen, S.-Q., Huang, Y.-Z., et al. 2018b. Whole-sequence analysis indicates that the Y chromosome C2*-Star Cluster traces back to ordinary Mongols, rather than Genghis Khan. *European Journal of Human Genetics* 26: 230–237.
Wei, W., Ayub, Q., Chen, Y., McCarthy, S., Hou, Y., Carbone, I., Xue, Y., and Tyler-Smith, C. 2013a. A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Research* 23: 388–395.
Wei, W., Ayub, Q., Xue, Y., and Tyler-Smith, C. 2013b. A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. *Forensic Science International: Genetics* 7: 568–572.
Wells, R.S., Yuldasheva, N., Ruzibakiev, R., Underhill, P.A., Evseeva, I., et al. 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences of the United States of America* 98: 10244–10249.
Wilder, J.A., Kingan, S.B., Mobasher, Z., Pilkington, M.M., and Hammer, M.F. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nature Genetics* 36: 1122.
Willems, T., Gymrek, M., Poznik, G.D., Tyler-Smith, C., and Erlich, Y. 2016. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. *The American Journal of Human Genetics* 98: 919–933.
Wilson, M.A., and Makova, K.D. 2009. Genomic Analyses of Sex Chromosome Evolution. *Annual Review of Genomics and Human Genetics* 10: 333–354.
Wong, E.H.M., Khrunin, A., Nichols, L., Pushkarev, D., Khokhrin, D., et al. 2017. Reconstructing genetic history of Siberian and Northeastern European populations. *Genome Research* 27: 1–14.
Wright, A.E., Dean, R., Zimmer, F., and Mank, J.E. 2016. How to make a sex chromosome. *Nature Communications* 7: 12087.
Wright, A.E., Harrison, P.W., Montgomery, S.H., Pointer, M.A., and Mank, J.E. 2014. Independent stratum formation on the avian sex chromosomes reveals inter-chromosomal gene conversion and predominance of purifying selection on the W chromosome. *Evolution* 68: 3281–3295.
Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology* 19: 1453–1457.
Xue, Y., Zerjal, T., Bao, W., Zhu, S., Lim, S.-K., et al. 2005. Recent Spread of a Y-Chromosomal Lineage in Northern China and Mongolia. *The American Journal of Human Genetics* 77: 1112–1116.

Yan, S., Wang, C.-C., Li, H., Li, S.-L., Jin, L., et al. 2011. An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. *European Journal Of Human Genetics* 19: 1013.

Yan, S., Wang, C.-C., Zheng, H.-X., Wang, W., Qin, Z.-D., et al. 2014. Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers. *PLoS ONE* 9: e105691.

Yang, J.J., Li, J., Buu, A., and Williams, L.K. 2013. Efficient inference of local ancestry. *Bioinformatics* 29: 2750–2756.

Yunusbayev, B., Metspalu, M., Metspalu, E., Valeev, A., Litvinov, S., et al. 2015. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLOS Genetics* 11: e1005068.

Zalloua, P.A., Platt, D.E., El Sibai, M., Khalife, J., Makhoul, N., et al. 2008. Identifying Genetic Traces of Historical Expansions: Phoenician Footprints in the Mediterranean. *The American Journal of Human Genetics* 83: 633–642.

Zegura, S.L., Karafet, T.M., Zhivotovsky, L.A., and Hammer, M.F. 2004. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol Biol Evol* 21: 164–175.

Zeng, T.C., Aw, A.J., and Feldman, M.W. 2018. Cultural hitchhiking and competition between patrilineal kin groups explain the post-Neolithic Y-chromosome bottleneck. *Nature Communications* 9:1

Zerbino, D.R., Achuthan, P., Akanni, W., M Ridwan Amode, Vullo, A., et al. 2017. Ensembl 2018. *Nucleic Acids Research* 46: D754–D761.

Zerjal, T., Dashnyam, B., Pandya, A., Kayser, M., Santos, F.R., et al. 1997. Genetic Relationships of Asians and Northern Europeans, Revealed by Y-Chromosomal DNA Analysis. *The American Journal of Human Genetics* 60: 1174–1183.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., et al. 2003. The Genetic Legacy of the Mongols. *The American Journal of Human Genetics* 72: 717–721.

Zhang, Y., Li, J., Zhao, Y., Wu, X., Li, H., Yao, L., Zhu, H., and Zhou, H. 2017. Genetic diversity of two Neolithic populations provides evidence of farming expansions in North China. *J Hum Genet* 62: 199–204.

Zhao, Y.-B., Zhang, Y., Li, H.-J., Cui, Y.-Q., Zhu, H., and Zhou, H. 2014. Ancient DNA evidence reveals that the Y chromosome haplogroup Q1a1 admixed into the Han Chinese 3,000 years ago. *American Journal of Human Biology* 26: 813–821.

Zhivotovsky, L.A., Underhill, P.A., Cinnioğlu, C., Kayser, M., Morar, B., et al. 2004. The Effective Mutation Rate at Y Chromosome Short Tandem Repeats, with Application to Human Population-Divergence Time. *The American Journal of Human Genetics* 74: 50–61.

Zhivotovsky, L.A., Underhill, P.A., and Feldman, M.W. 2006. Difference between Evolutionarily Effective and Germ line Mutation Rate Due to Stochastically Varying Haplogroup Size. *Molecular Biology and Evolution* 23: 2268–2270.

Zhong, H., Shi, H., Qi, X.-B., Xiao, C.-J., Jin, L., Ma, R.Z., and Su, B. 2010. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. *Journal of Human Genetics* 55: 428–435.

Web Resources

Ensemble	https://www.ensembl.org/index.html
YFull YTree	https://www.yfull.com/tree/
ISOGG	https://isogg.org/

Mammalian silhouettes:

https://creazilla.com/nodes/2556-echidna-silhouette
https://creazilla.com/nodes/580-monkey-silhouette
https://creazilla.com/nodes/2145-kangaroo-silhouette
https://creazilla.com/nodes/1478-platypus-silhouette

ACKNOWLEDGEMENTS

This journey took longer than I had expected. But this ship would have never reached its destination without the navigational skills and experience of my supervisors, helping to guide it through the Bay of Publications, the Bureaucratic Triangle and the bottomless Sea of Scientific Literature. I am grateful for your knowledge and experience that were and will remain invaluable. Kristiina, thank you for the positive attitude that fuelled this somewhat long process. Siiri, I think your PhD defence was the first one I had witnessed, so I would credit you for providing the initial spark of inspiration to instigate this voyage. Richard, there is not much to say that has not been said before, but you truly are the grandmaster in this game of scientific inquiry, always knowing where to place your bets. Maybe I learned something from it.

My gratitude goes out to all of the co-authors of the articles included in this thesis and especially to all the sample collectors and sample donors – your silent contribution is one of the pillars upholding this science.

During the course of my studies, *Eesti Biokeskus* has mutated, evolved and fused, yet it has always kept its friendly, open-minded atmosphere oriented towards only the exquisitely novel science and indisputably the most optimal approaches. To all the EBC oldies and newbies – it has always been a pleasure and a privilege to interact with you inside or outside the lab. To the most recent common trailblazers in our evolutionary biology department at the University of Tartu – Monika, Lehti, Erwan (and to all others from here and there whose PhD defences I have witnessed over the years – Gyaneshwer, Chandana, Triinu, Hiie, Mart, Erge, Maris, Marge, to name only a few) – you have been an inspiration to just do it! Helen and Ajai, some last rowing strokes and you will be just done as well, keep it going a bit more. Katri, you have witnessed the last miles of this journey from close proximity. I wish you smooth sailing through the fickle weather with a myriad of captains aboard. To all others from our rather spacious backroom (and other rooms as well) – just some hard work, then a little bit of luck, and you'll arrive at the imaginary la dolce vita just in time, too, no worries. Mait, thank you for the attitude, digitalised rivers and mountains. Jüri, your stories never fail to reveal how pampered we all are today and how much we rely on the work done by those before us. Anu, your friendly reminders concerning the essential stuff like what, when and to whom are always on time, thank you. To all of the administrative staff, former and present, without your firm support this vessel would have been on the rocks long ago. Thank you for knowing your trade.

I am thankful to the Archimedes Foundation and EBC for supporting my attendance of several courses. I am still convinced that they were quite useful when just standing at the gateway of the PhD experience and especially when in the midst of a never-ending technological transition.

I am grateful to my family for always supporting my somewhat unconventional decisions, despite the confusion regarding what exactly it is all about.

Agur, thank you for taking care of the one-of-a-kind ponycathorseprincess – princesses demand unfaltering perpetual admiration day and night and you have been able to provide exactly that, while I am busy writing these lines. Olen südamest tänulik teie abi ja toetuse eest. Nendeta poleks olnud võimalik käesolevat tööd lõpetada. Aitäh!

PUBLICATIONS

SUPPLEMENTARY INFORMATION

Full Supplementary Information for:

Ref. I
https://www.nature.com/articles/5201748#Sec8

Ref. II
https://www.sciencedirect.com/science/article/pii/S0002929716301604?via%3Dihub#app3

Ref. III
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1522-1#Sec25

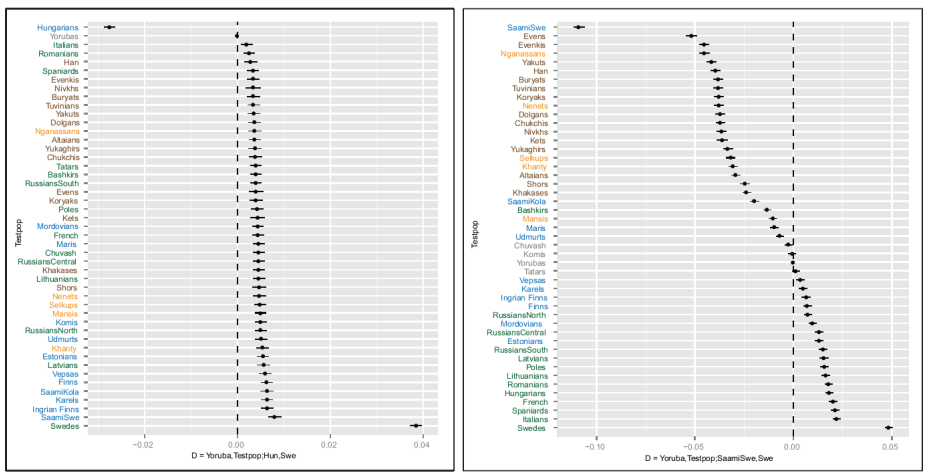

Figure S1. D-statistics calculated for the tree model in the form of D(outgroup, test population; Uralic speaking population, non-Uralic speaking population). Yorubas were used as an outgroup. Uralic speaking populations fixed in the tested model are the Saami from Sweden (SaamiSwe), Finns (Fin), Estonians (Est) and Hungarians (Hun) from Europe. The non-Uralic speaking populations fixed in the tested model are French, Swedes and Poles, but only results with Swedes are presented here. The values on the Y-axis are sorted by D value. Colour codes of populations showing significant deviations from D=0 (Z score ≥3) correspond to linguistic affinities of tested populations: blue – European Uralic speaking populations; green – European non-Uralic speaking populations; orange – West Siberian Uralic speaking populations; brown – Siberian and East Asian non-Uralic speaking populations. Grey coloured labels indicates D=0 (Z score < 3), standard errors to the point estimates are shown with the black bars. Adapted from Ref.III, under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

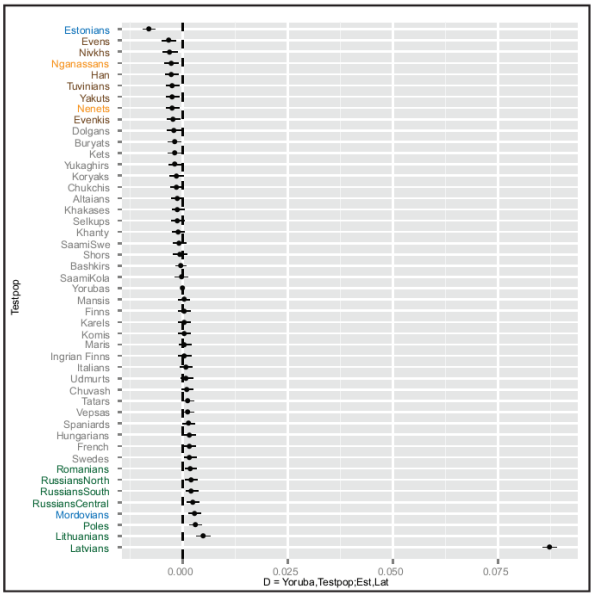

Figure 2S. D-statistics calculated for the tree model in the form of D(outgroup, test population; Estonians, Latvians). Yorubas were used as an outgroup. Colour codes of populations showing significant deviations from D=0 (Z score ≥3) correspond to linguistic affinities of tested populations: blue – European Uralic speaking populations; green – European non-Uralic speaking populations; orange – West Siberian Uralic speaking populations; brown – Siberian and East Asian non-Uralic speaking populations. Grey coloured labels indicates D=0 (Z score < 3), standard errors to the point estimates are shown with the black bars. Adapted from Ref.III, under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

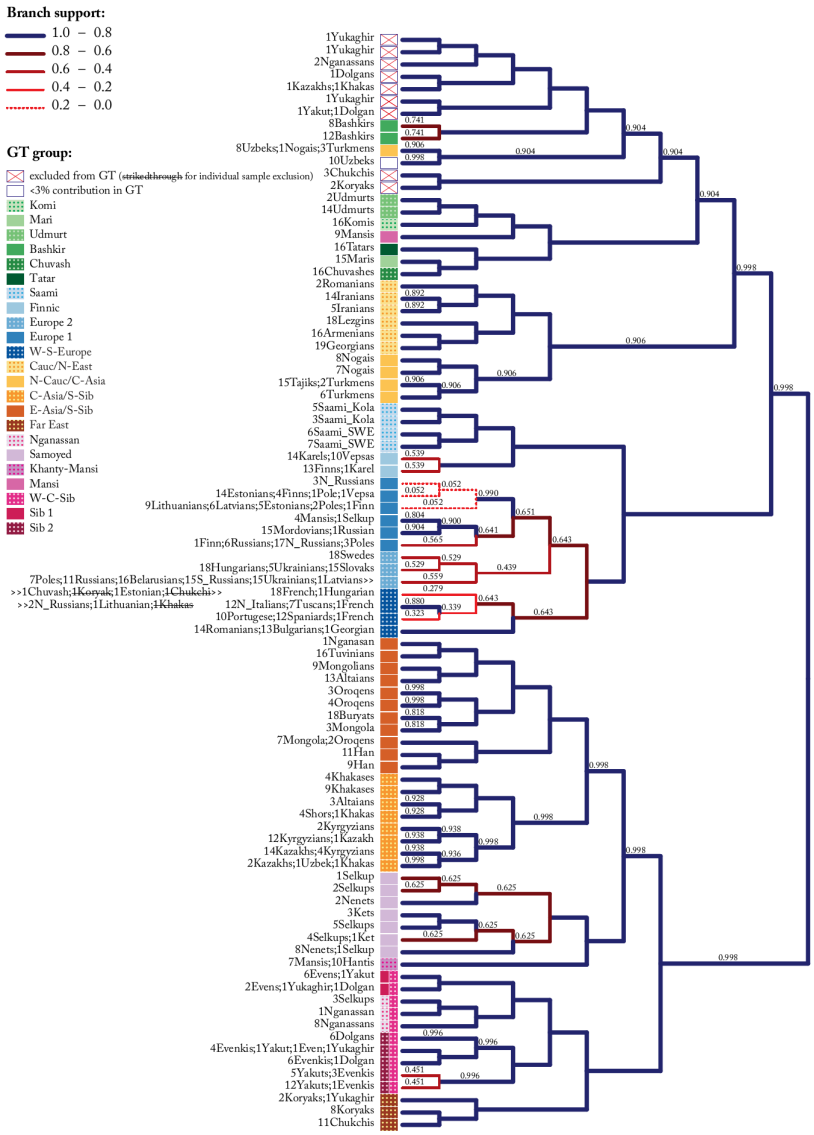

Figure 3S. Clustering of individual samples from the comparative dataset, as inferred by fineSTRUCTURE. The tree clusters individuals with similar copying vectors. Labels identify how many and which samples are included in each cluster. Individual tips were manually inspected, grouped and colour-coded for further admixture analysis in Globetrotter. Colour legend is given on the left-hand side. fineSTRUCTURE populations (crossed out boxes) and individual samples (strikethrough font) which show unusually high levels of admixture were excluded from further analysis. Line thickness of individual branches indicates statistical bootstrap support; legend is given in the top left corner. Adapted from Ref.III, under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

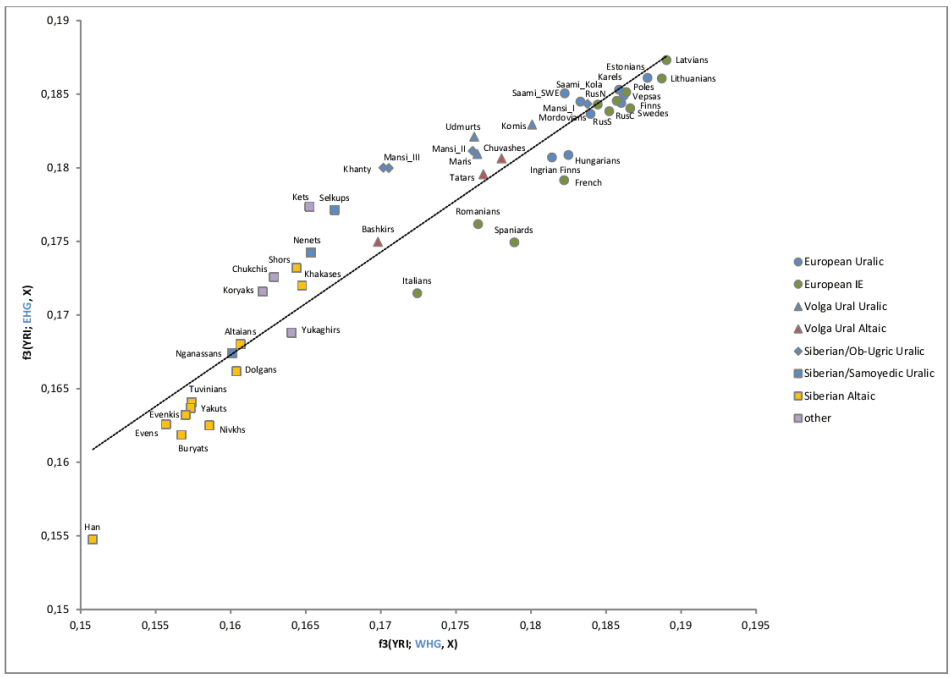

Figure 4S. Outgroup f3-statistics' results in the form f3 (Yorubas; ancient Eurasian population, modern Eurasian population) plotted pairwise against each other. f3 of Eastern hunter gatherer (EHG) is plotted against f3 of Western hunter gatherer (WHG). Adapted from Ref.III, under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

CURRICULUM VITAE

Name: Anne-Mai Ilumäe
Date of Birth: November 21, 1983
Family: partner, 1 child
Contacts: University of Tartu, Institute of Genomics, Riia 23b-306b, 51010 Tartu, Estonia, +372 737 5050
E-mail: ami@ut.ee

Education:

2010–2019 Doctoral studies, University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology

2006–2007 Kansai Gaidai University (Japan), Asian studies 1-year program for foreign students

2005–2008 *MSc*, University of Tartu, Faculty of Science and Technology, Institute of Molecular and Cell Biology, Chair of Evolutionary Biology, supervisors Siiri Rootsi, Ene Metspalu, "Development and implementation of SNaPshot multiplex genotyping for investigating human Y-chromosomal variation in Eastern European populations"

2002–2005 *BSc*, University of Tartu, Faculty of Biology and Geography, Institute of Molecular and Cell Biology, Chair of Evolutionary Biology

Professional employment:

2018–... Junior Researcher, University of Tartu, Institute of Genomics, Research group of mitochondrial DNA and Y chromosome

2013–2017 Junior Researcher, Estonian Biocentre, Research group of population genetics

2008–2011 Programme Manager of International Summer University, Lifelong Learning Centre, University of Tartu

2008 Sales manager at Solis BioDyne

Scholarships, courses and conferences:

2015 Poster presentation in *9th International Society for Applied Biological Sciences (ISABS) Conference on Forensic and Anthropologic Genetics and Mayo Clinic Lectures in Individualized Medicine* in Bol, Croatia

2013 Archimedes Foundation Kristjan Jaak Scholarship for Short-time Visit Participation in *Introduction to Genomic Data Analysis using HapMap and 1000 Genomes Projects* course organised by Transmitting Science and University of Barcelona in Spain

2012 Archimedes Foundation Kristjan Jaak Scholarship for Short-time Visit Participation in *Workshop on Genomics* organised by Broad Institute (USA), Washington University (USA), Lund University (Sweden) and Swedish Research Council in Czech Republic

Teaching, advising and scientific outreach activities:

2019 Lecturer in Estonia's National University 100th anniversary scientific outreach project in high schools across Estonia

2013–2014 Instructor of Molecular Biology Workshop for high school students

2014–... Preparation of laboratory sessions and giving lectures in undergraduate courses on *Population Genetics* and *Evolutionary Biology* at the University of Tartu, supervision of bachelor level students.

Publications:

Tambets K, Yunusbayev B, Hudjashov G, **Ilumäe AM**, Rootsi S, Honkola T, Vesakoski O, Atkinson Q, Skoglund P, Kushniarevich A, Litvinov S, Reidla M, Metspalu E, Saag L, Rantanen T, Karmin M, Parik J, Zhadanov SI, Gubina M, Damba LD, Bermisheva M, Reisberg T, Dibirova K, Evseeva I, Nelis M, Klovins J, Metspalu A, Esko T, Balanovsky O, Balanovska E, Khusnutdinova EK, Osipova LP, Voevoda M, Villems R, Kivisild T, Metspalu M. (2018). **Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations.** *Genome Biology* 19(1):139

Ilumäe AM, Reidla M, Chukhryaeva M, Järve M, Post H, Karmin M, Saag L, Agdzhoyan A, Kushniarevich A, Litvinov S, Ekomasova N, Tambets K, Metspalu E, Khusainova R, Yunusbayev B, Khusnutdinova EK, Osipova LP, Fedorova S, Utevska O, Koshel S, Balanovska E, Behar DM, Balanovsky O, Kivisild T, Underhill PA, Villems R, Rootsi S. (2016). **Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families.** *American Journal of Human Genetics* 99(1):163-173

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, Rootsi S, **Ilumäe AM**, Mägi R, Mitt M, Pagani L, Puurand T, Faltyskova Z, Clemente F, Cardona A, Metspalu E, Sahakyan H, Yunusbayev B, Hudjashov G, DeGiorgio M, Loogväli EL, Eichstaedt C, Eelmets M, Chaubey G, Tambets K, Litvinov S, Mormina M, Xue Y, Ayub Q, Zoraqi G, Korneliussen TS, Akhatova F, Lachance J, Tishkoff S, Momynaliev K, Ricaut FX, Kusuma P, Razafindrazaka H, Pierron D, Cox MP, Sultana GN, Willerslev R, Muller C, Westaway M, Lambert D, Skaro V, Kovačevic L, Turdikulova S, Dalimova D, Khusainova R, Trofimova N, Akhmetova V, Khidiyatova I, Lichman DV, Isakova J, Pocheshkhova E, Sabitov Z, Barashkov NA, Nymadawa P, Mihailov E, Seng JW, Evseeva I, Migliano AB, Abdullah S, Andriadze G, Primorac D, Atramentova L, Utevska O, Yepiskoposyan L, Marjanovic D, Kushniarevich A, Behar DM, Gilissen C, Vissers L, Veltman JA, Balanovska E, Derenko M, Malyarchuk B, Metspalu

A, Fedorova S, Eriksson A, Manica A, Mendez FL, Karafet TM, Veeramah KR, Bradman N, Hammer MF, Osipova LP, Balanovsky O, Khusnutdinova EK, Johnsen K, Remm M, Thomas MG, Tyler-Smith C, Underhill PA, Willerslev E, Nielsen R, Metspalu M, Villems R, Kivisild T. (2015). **A recent bottleneck of Y chromosome diversity coincides with a global change in culture.** *Genome Research* 25(4): 459−466

Kovacevic L, Tambets K, **Ilumäe AM**, Kushniarevich A, Yunusbayev B, Solnik A, Bego T, Primorac D, Skaro V, Leskovac A, Jakovski Z, Drobnic K, Tolk HV, Kovacevic S, Rudan P, Metspalu E, Marjanovic D. (2014). **Standing at the Gateway to Europe – The Genetic Structure of Western Balkan Populations based on Autosomal and Haploid Markers.** *PLoS ONE* 9 (8):e105090

Chaubey G, Singh M, Crivellaro F, Tamang R, Nandan A, Singh K, Sharma VK, Pathak AK, Shah AM, Sharma V, Singh VK, Selvi Rani D, Rai N, Kushniarevich A, **Ilumäe AM**, Karmin M, Phillip A, Verma A, Prank E, Singh VK, Li B, Govindaraj P, Chaubey AK, Dubey PK, Reddy AG, Premkumar K, Vishnupriya S, Pande V, Parik J, Rootsi S, Endicott P, Metspalu M, Lahr MM, van Driem G, Villems R, Kivisild T, Singh L, Thangaraj K. (2014). **Unravelling the distinct strains of Tharu ancestry.** *European Journal of Human Genetics* 22(12):1404−1412

Rootsi S, Zhivotovsky LA, Baldovic M, Kayser M, Kutuev IA, Khusainova R, Bermisheva MA, Gubina M, Fedorova SA, **Ilumäe AM**, Khusnutdinova EK, Voevoda MI, Osipova LP, Stoneking M, Lin AA, Ferak V, Parik J, Kivisild T, Underhill PA, Villems R. (2007). **A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe.** *European Journal of Human Genetics* 15(2): 204−211

ELULOOKIRJELDUS

Nimi: Anne-Mai Ilumäe
Sünniaeg: 21. november 1983
Perekond: elukaaslane, 1 laps
Kontakt: Tartu Ülikool, Genoomika Instituut Riia 23b-306b, 51010 Tartu, Eesti, +372 737 5050
E-post: ami@ut.ee

Hariduskäik:

2010–2019 doktoriõpingud geenitehnoloogias, Tartu Ülikooli loodus- ja täppisteaduste valdkond, molekulaar- ja rakubioloogia instituut

2006–2007 Kansai Gaidai Ülikool (Jaapan), 1-aastane Aasia õpingute programmi

2005–2008 geenitehnoloogia magister, Tartu Ülikooli bioloogia-geograafiateaduskond, molekulaar- ja rakubioloogia instituut, evolutsioonilise bioloogia õppetool. Juhendajad Siiri Rootsi, Ene Metspalu, magistritöö "*SNaPshot multiplex genotüpiseerimismeetodi väljatöötamine ja juurutamine inimese Y-kromosomaalse varieeruvuse uurimiseks Ida-Euroopa populatsioonides*"

2002–2005 geenitehnoloogia bakalaureus, Tartu Ülikooli bioloogia-geograafiateaduskond, molekulaar- ja rakubioloogia instituut, evolutsioonilise bioloogia õppetool

Töökogemus:

2018–... nooremteadur, Tartu Ülikooli genoomika instituut, mitokondriaalse DNA ja Y-kromosoomi uurimisrühm

2013–2017 nooremteadur, Eesti Biokeskus, populatsioonigeneetika uurimisrühm

2008–2011 rahvusvahelise suveülikooli programmijuht, Tartu Ülikooli elukestva õppe keskus

2008 müügiosakond, OÜ Solis BioDyne

Stipendiumid, kursused ja konverentsid:

2015 Posterettekanne *9th International Society for Applied Biological Sciences (ISABS) Conference on Forensic and Anthropologic Genetics and Mayo Clinic Lectures in Individualized Medicine* Horvaatias

2013 Sihtasutuse Archimedes Kristjan Jaagu välislähetuse stipendium osalemiseks kursusel *Introduction to Genomic Data Analysis using HapMap and 1000 Genomes Projects,* organisaatorid Transmitting Science ja Barcelona Ülikool Hispaanias

2012 Sihtasutuse Archimedes Kristjan Jaagu välislähetuse stipendium osalemiseks kursusel *Workshop on Genomics,* organisaatorid Broad

Instituut (USA), Washingtoni Ülikool (USA), Lundi Ülikool (Rootsi) ja Rootsi Teadusagentuur Tšehhi Vabariigis

Teenistuskäigu lisainfo:

2019 Loengud eri Eesti gümnaasiumides Tartu Ülikooli projekti Rahvusülikool 100 raames

2013–2014 Gümnaasiumide geneetika õppepäevade juhendaja Tartus

2014–... Loengud, praktikumide ning üliõpilaste juhendamised Tartu Ülikooli bakalaureuse- ja magistritasemekursustel: evolutsiooniline bioloogia ja populatsioonigeneetika

Teaduspublikatsioonid:

loetletud ingliskeelse CV rubriigis *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. **Kaasik, Krista**. 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 Usküla**. 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.