UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
INSTITUTE OF MOLECULAR AND CELL BIOLOGY
DEPARTMENT OF EVOLUTIONARY BIOLOGY

Dmitry Lomovsky

Estimating identity by descent with GERMLINE software in individuals from the Saami population

BSc thesis

Supervisors: MSc Anne-Mai Illumäe
PhD Georgi Hudjashov

TARTU 2014

TABLE OF CONTENTS 2
ABBREVIATIONS 3
INTRODUCTION 4
1. LITERATURE OVERVIEW 5
  1.1. Variation in the human genome 5
    1.1.1. Linkage disequilibrium 7
    1.1.2. Haplotypic structure of the human genome 8
  1.2. Overview of identity by descent in the human genome 11
    1.2.1. Examples of research involving IBD fragments length in population genetics studies.. 14
    1.2.2. Methods of IBD estimation 18
  1.3. GERMLINE algorithms 21
2. EXPERIMENTAL PART 25
  2.1. Aim of the study 25
  2.2. Materials and methods 25
    2.2.1. Genotyping data 25
    2.2.2. Data processing 25
  2.3. Results and discussion 26
CONCLUSION 31
Resümee 33
REFERENCES 35
USED WEB ADDRESSES 41
ACKNOWLEDGEMENTS 42
SUPPLEMENTARY DATA 43

ABBREVIATIONS

SNP – Single Nucleotide Polymorphism

LD – Linkage Disequilibrium

IBD - Identity By Descent

IBS - Identity By State

CNV – Copy-Number Variation

HMM – Hidden Markov Model

kb – kilobase

SD – Standard Deviation

ya – years ago

INTRODUCTION

High-resolution genome sequencing and genotyping has brought terabytes of whole-genome data and methods for its processing and analysis. Constantly increasing number of complete human genome sequences available allows researchers to detect previously unknown rare variants of alleles and apply this knowledge to improve our understanding of haplotype structure of the human genome.

However, increased data complexity is correlated with increase in computation time, which demands new solutions from modern software developers. This is exemplified by temporary sliding window approach implemented in fastest present-day programs, and also by other novel algorithms for handling multi-aligned and densely genotyped data.

High resolution of modern genetic data provides enough accuracy to estimate the precise length of identical by descent (IBD) segments in genomes of any two individuals. The fact that unrelated individuals may have common identical segments of genome is explained by the shared ancestry broken into pieces by the recombination events and Mendelian laws of inheritance (Browning, 2008). IBD segments' mean length, their distribution along the chromosomes and lack or excess of identity sharing shed light on recent population history and natural selection (Albrechtsen et al., 2009; Purcell et al., 2007; Gusev et al., 2009).

Aim of this study is to give a brief overview on software designed for IBD detection with examples of IBD use for population genetics studies. One of these computational approaches is further applied to detect the lengths of IBD segments in a total of 15 individuals from the population of the Swedish Saami in order to explore the population's IBD background and analyze unreported relatedness (the excess of genome sharing between unrelated individuals compared to the theoretical expectations).

1. LITERATURE OVERVIEW

1.1. Variation in the human genome

A diploid human cell without abnormalities contains 46 chromosomes, one half of which derive from the mother and the other half from the father. Due to the independent assortment of the paternal and maternal chromosomes during meiosis, possible number of variants of different chromosomal sets reaches up to 8,388,608 (Perez, 2007). Even genomes of monozygotic twins will have some discerning point mutations, which occur with random frequency during cell division when genome replicates into two practically identical copies. Using whole genome sequencing data, mean mutation rate has been estimated to be around 1.2 to 1.4×10^{-8} per site per generation (Roach et al., 2010; Scally et al., 2012). Single basepair point mutations are called single nucleotide polymorphisms (SNPs) and due to negative selection are usually found more frequently in non-coding than in coding regions of the genome or, in general, in positions, where such point mutations would be swept to fixation by positive selection (Barreiro et al., 2008). SNPs found in coding segments of the genome, but not affecting protein sequence are called synonymous. Non-synonymous SNPs are could be divided into missense and nonsense types. Missense mutation results in an amino acid change in the protein sequence and nonsense polymorphism results in an appearance of the stop codon (Lodish et al., 2000).

Another important source of diversity is recombination between matching DNA segments of homologous chromosomes during meiosis (crossing-over). When homologous DNA chromatids in gametic cells align along metaphase plate during meiosis, one of them may break and swap genetic material with the other one. Imprecise paring of chromosomes during this process could lead to the emergence of various structural variants - deletions, duplications, inversions or translocations (Figure 1) (Bu and Cao, 2012).

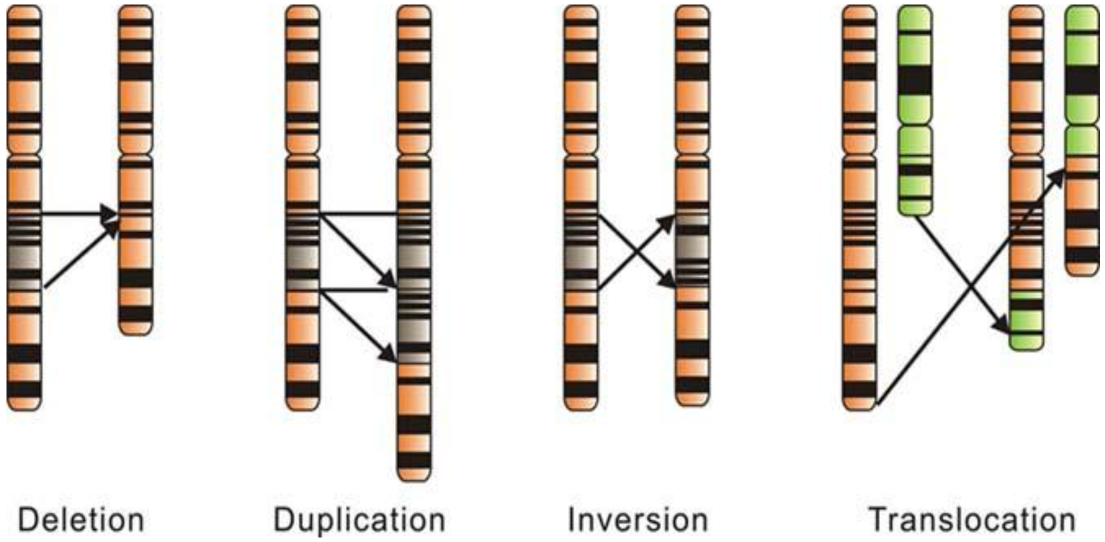

Figure 1. Most important structural variants contributing to the overall variation in human genome (Bu and Cao, 2012).

Human haplotype map for 38 million SNPs, 1.4 million short insertions and deletions, and more than 14,000 larger deletions was validated by the year 2012 (The 1000 Genomes Project Consortium, 2012). Overall amount of SNPs detected by all independent projects and human genome research groups by June 2012 reached 187,852,828 SNPs (NCBI dbSNP build 137 for Homo Sapiens).

Copy-number variant (CNV) is a structural variant of a genome segment, which is present at a variable number of copies in comparison to the reference sequence (Feuk et al., 2006). One of the largest assessment of genomic copy number variation was conducted by Shaikh et al., in 2009 in the context of human genetic variation, disease susceptibility, and clinical molecular diagnostics with 2026 samples of healthy individuals comprised of Caucasians, African-Americans and Asian-Americans, and using high-density SNP-based oligonucleotide microarrays. Altogether 54,462 CNVs were detected, collectively spanning over 551,995,356 unique base pairs, or ~19.4% of the human genome. Most of the detected CNVs (77.8%) were classified as non-unique CNVs, meaning that they were present in more than one unrelated individual. Average number of CNVs per individual was 26.9 (Shaikh et al., 2009).

1.1.1. **Linkage disequilibrium**

Recombination depends on the distance between genetic regions. If two genes are located closely together on one chromosome, the likelihood that the recombination event will separate these two genes is less than if they are located further apart from each other. Linkage describes the tendency of genes to be co-inherited as a result of their close physical position on the same chromosome. Linkage disequilibrium (LD) describes a situation, in which some combinations of alleles occur more or less frequently in a population than it would be expected (Slatkin, 2008).

Linkage disequilibrium can be derived as a function from the difference between the observed frequency of the alleles in a given locus and the frequency of the alleles as if they segregated by the law of the independent assortment (Equation 1), and commonly denoted with a capital D (Ardlie et al., 2002).

Equation 1: $\mathbf{D = P_{AB} – P_A * P_B}$

P_A and P_B are expected frequencies of alleles A, a and B, b in given loci A and B, considering independent assessment during meiosis. P_{AB} is frequency of the haplotype where both loci co-exist together.

Denoting D(0) as an extent of LD in the beginning, we can now infer the D(t) after the t number of generations with *r* as a rate of recombination events (Equation 2) (Ardlie et al., 2002).

Equation 2: $\mathbf{D(t) = (1 - r) * t * D(0)}$

In order to compare LD extent among different loci, D value must be normalized. For these purposes D' and r^2 definitions are often used, where D' is the ratio between D and the maximum possible value of D, given the allele frequencies at the two loci (Equation 3) and r^2 is a statistical correlation between LD of compared loci (Equation 4).

Equation 3: $\mathbf{D' = D / Dmax}$

Equation 4: $\mathbf{r^2 = D^2 / P_A * P_a * P_B * P_b}$

D' ranges from 0, when there is no LD between markers, to 1, when given loci are in perfect LD. Value r^2 has the same range from 0 to 1 but is more likely to be used when the sample number is small in order to reduce random sampling error (Ardlie et al., 2002; Wall and Pritchard, 2003; Mueller, 2004).

Beside genetic factors LD is influenced by the demographic history of the population, natural selection, admixture between populations, recent mutations and also founder effects. Population that has recently expanded is expected to show lower LD values than a population with a history of constant size or a bottleneck event, such as the population of Saami (Laan et al., 1997; Tambets et al., 2004). Migration tends to increase LD due to a usually small group of founders contributing few alleles at each locus to the overall genetic variation among extant populations (Abecasis et al., 2005). European and Asian populations contain longer LD segments in comparison to Africans due to the demographic bottleneck, which has occurred during the out of Africa migration of anatomically modern humans. The fraction of SNP pairs which are separated by less than 1kb and show evidence of historical recombination falls between 14% to 18% in African (Yoruban) and African-American samples and is only between 3% to 6% in European and Asian populations (Gabriel et al., 2002). Migration is usually followed by population growth that increases allele frequencies and decreases LD. Fixation of the allele by genetic drift can be another reason for decrease of LD (Ardlie et al., 2002).

1.1.2. Haplotypic structure of the human genome

Haplotype (derived from “haploid genotype”) is a set of alleles at the locus, which are in (nearly) complete linkage disequilibrium and, therefore, are inherited together. When crossing-over or mutation lead to the appearance of a new allele, a new haplotype emerges (Figure 2). Haplotype may denote a single locus, as well as the whole genome (Ardlie et al., 2002). Taking into the account that alleles belonging the common haplotype are linked, a subset of few carefully selected “tag SNPs” could be used to identify allelic states of other polymorphisms belonging to the same haplotype. Mapping human haplotypes for genome-wide association studies was one of the main goals of the International HapMap Project (The International HapMap Consortium 2003, 2005).

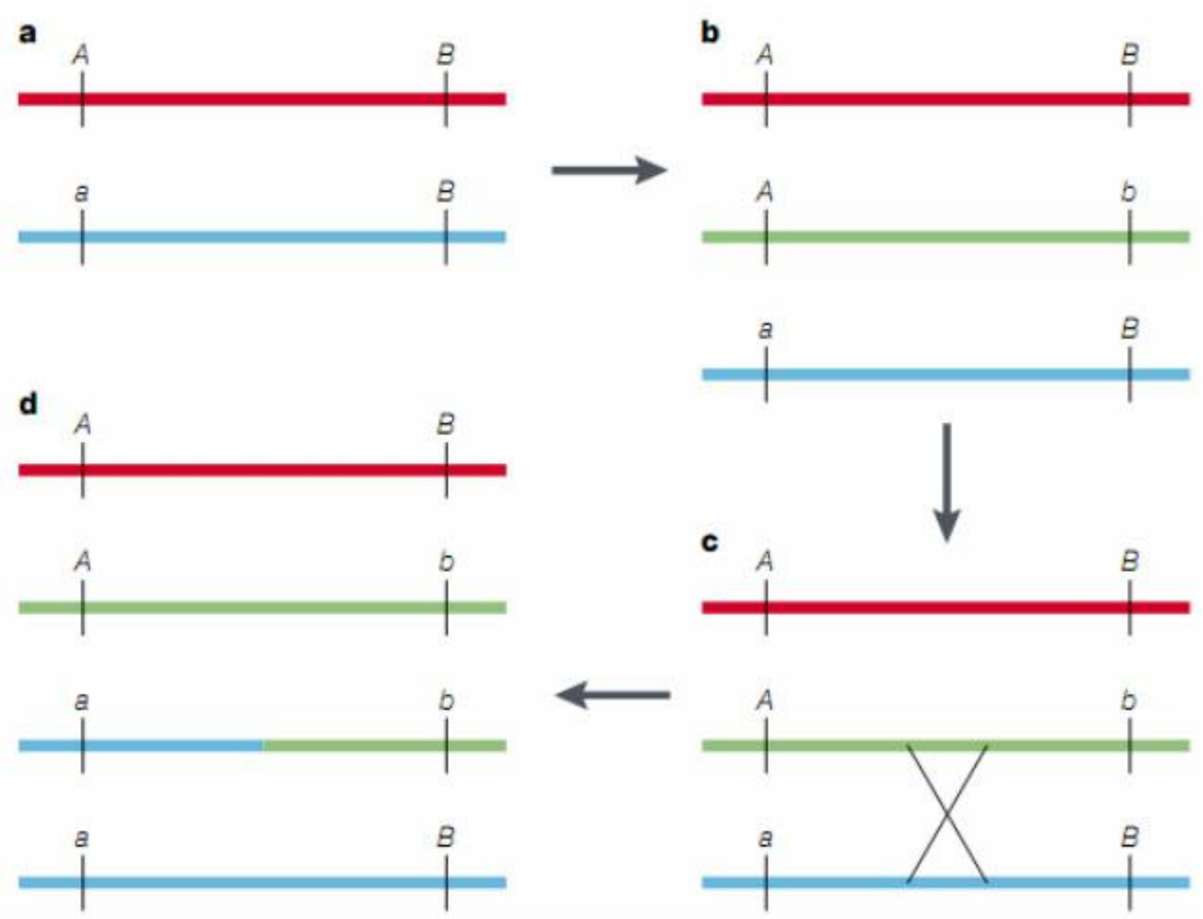

Figure 2. An example how recombination can create a totally new haplotype. (A) At the beginning there are only two haplotypes in the population – **AB** and **aB**.(B) Mutation in **B** creates new **b** allele, resulting in the emergence of new haplotype **Ab**. Three haplotypes are observed: **AB**, **Ab** and **aB**. (C) Homologous recombination occurs between **Ab** and **aB** chromosomes. (D) Recombination event gives rise to the forth haplotype in the population, **ab**. (Ardlie et al., 2002).

Increase in the haplotype diversity can be attributed to recombination in diploid loci and mutation events. On the contrary, genetic drift and natural selection may eliminate the haplotype from the population resulting in the decrease of haplotype diversity. Studies of LD between SNP markers reveal a block-like structure for some regions of the human genome. There is a high level of LD within blocks and low level between them, leading to the increase of haplotype-blocks diversity, compared to the variance across block boundaries (Cardon et al., 2003). Different populations have different haplotypic structure. Studies of haplotype patterns between individuals from Africa, Europe and Asia revealed greatest haplotype diversity in the Yoruban and African-American samples providing evidence for stronger historical recombination in these populations compared to the Europeans and Asians (Figure 3).

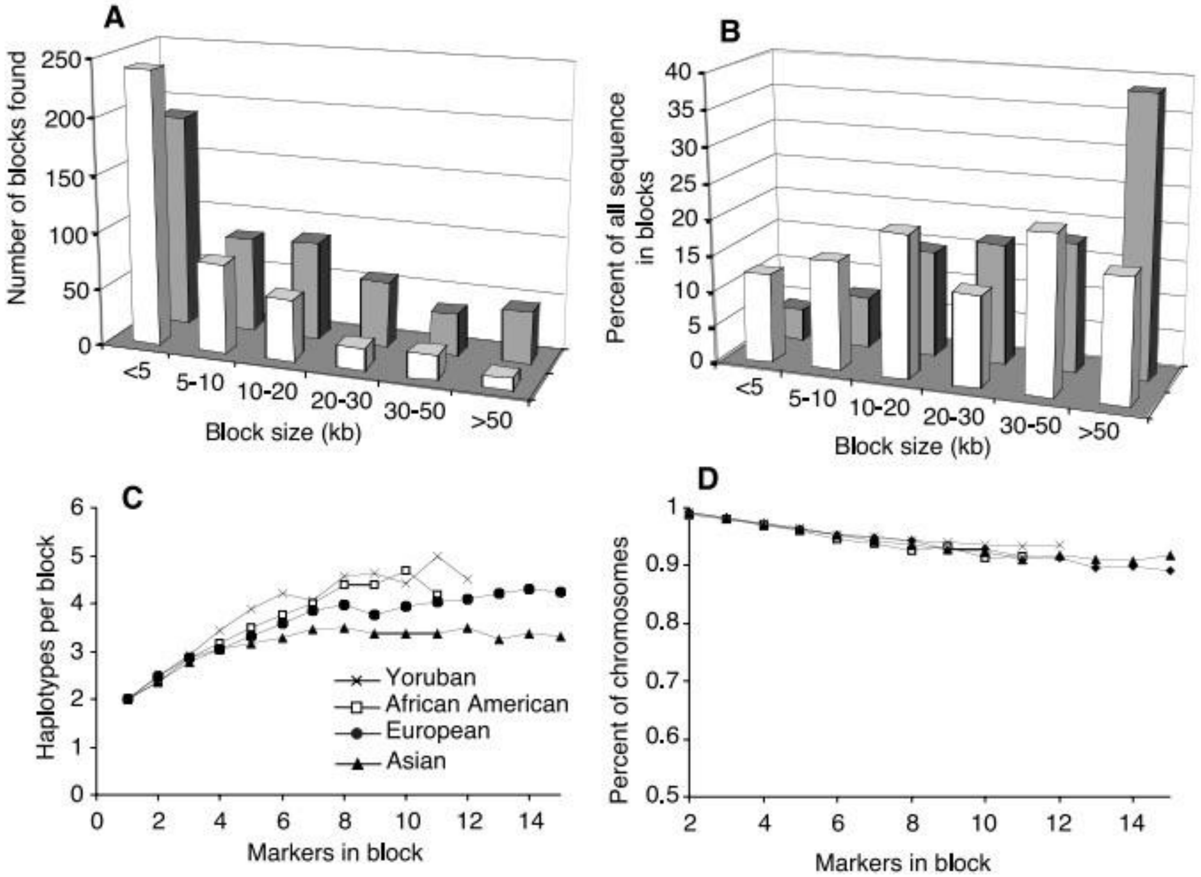

Figure 3. Block statistics among different populations. (A) Distribution of all haplotype blocks found in the analysis by their size in kb. (B) Percentage of genomic sequence spanned by blocks, binned according to their size. (C) The number of common (>5%) haplotypes per block plotted as a function of the number of markers typed in each block. (D) Fraction of all chromosomes with a perfect match to one of the common haplotypes and the same plotting conditions as in C (Gabriel et al., 2002).

Modern models proposed for building haplotypic block structures rely on the fact that genome consists of high LD regions, separated by recombination hot-spots (Gabriel et al., 2002). Some of the models use already known haplotypes (Patil et al., 2001) defining haplotype block as a segment, where rate of recombination exceeds some fixed threshold (Gabriel et al., 2002), while other models rely on inferring haplotypes from genotype data. Haplotype blocks' length is negatively correlated with the rate of recombination and on average spans 5-20kb (Wall and Pritchard, 2003).

1.2. Overview of identity by descent in the human genome

Two genome segments are identical by descent (IBD) if they are inherited from the same common ancestor (Figure 4A). IBD can be defined as two or more alleles that are inherited from the same ancestor without a recombination event (Kong et al., 2008). Identity in nucleotide sequence is called identity by state (IBS) and counted the same as IBD, because the probability of recurrent mutations is too small for being taken into account (Browning, 2008). However, identity may be also caused by the presence of conservative regions in the genome (e.g. regions coding for ribosomal subunits and histons), which have been maintained untouched through time and not counted as IBD (Browning, 2008). Another reason for false positive IBD is the unusually high rate of LD. In addition, it is important to consider haplotype frequencies while estimating IBD, since identical rare variants genotyped from two different persons are perfect signs of recent kinship (Browning, 2008).

Two persons, who are separated in time by N number of meioses, have a probability of $1/2^{(N-1)}$ that any given autosomal segment in their genome is shared due to common ancestry. As more generations pass, the fraction of genome and size of the segments shared between individuals decrease almost exponentially (Figure 4B and 4C). Mean length of IBD segments decreases at a slower pace and can be inferred as N^{-1} M (or $100*N^{-1}$ cM) (Browning and Browning, 2012). An example of IBD sharing in chromosome 1 between fifth cousins (12 meioses), who are expected to share segments with mean length of 8 cM, is presented in Figure 5.

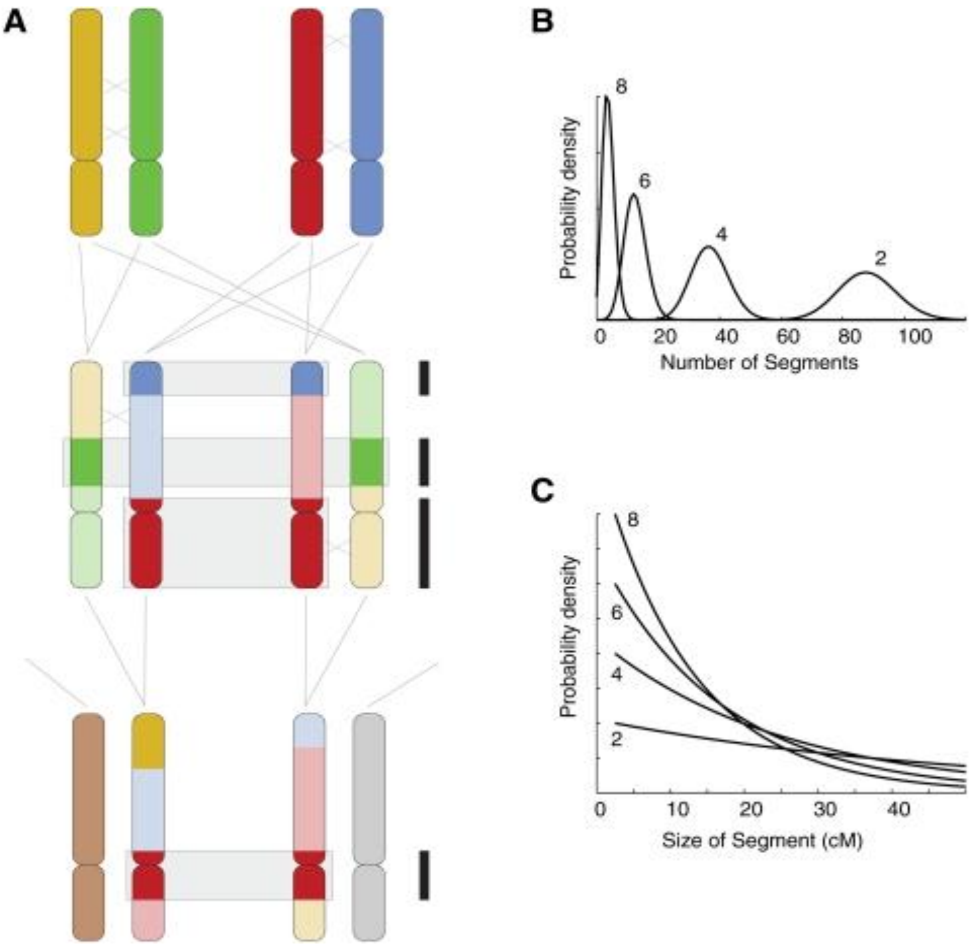

Figure 4. Expected distribution of IBD chromosomal segments across two individuals. (A) The sample pattern of shared IBD regions after two generations with a single recombination event per meiosis. Parental homologous autosomes are shown in color, grey and brown chromosomes are derived from unrelated individuals. Shared IBD segments of siblings in first generation and first cousins in second are shown in boxes and marked by black bars. (B) The number of shared IBD segments across all chromosomes is approximately Poisson-distributed with a mean that depends on the number of meioses separating two individuals, where 2, 4, 6 and 8 denotes number of meiosis between them. (C) The lengths of the IBD segments are approximately exponentially distributed, with mean length depending on the relationship between individuals (theoretical distributions are shown for 2, 4, 6 and 8 meioses separating two individuals) (Huff et al., 2011).

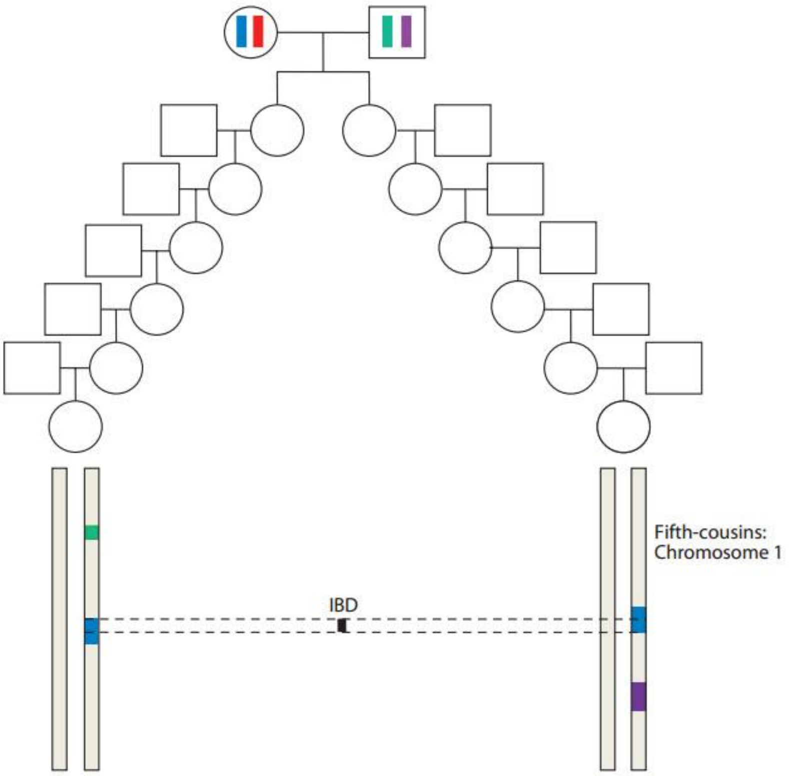

Figure 5. IBD on chromosome 1 for fifth cousins. IBD is shown in black. Regions of chromosome 1 that derive from the common ancestor are shaded in color (Browning and Browning, 2012).

Another example is the Maori population of New Zealand, which went through a population bottleneck ~800 years ago (Sutton, 1994; Browning, 2008) with ~190 founding woman ancestors (Whyte et al., 2005, Browning, 2008). Assuming that there were in total 1000 founding ancestral haplotypes (500 individuals), the probability of detecting two randomly chosen haplotypes that are IBD would be ~1/1000. As 800 years is approximately equal to 30 generations in humans, the mean length of a single-path IBD tract would be 1/60 = 0.017M (Browning, 2008). Besides recent ancestry, mean length of shared IBD segments is higher between individuals from closer geographic regions compared to that from distant ones (Browning and Browning, 2012).

1.2.1. Examples of research involving IBD fragments length in human population genetics studies

Due to the correlation between distribution of IBD segments lengths and time from the common ancestor, the number of genetic common ancestors and other aspects of recent population history can be inferred from the detected IBD structure. While previous methods of studying common ancestry mostly relied on uniparental markers, such as mtDNA and Y chromosome, genome-wide genotyping and re-sequencing datasets provide a much richer picture of human history by considering both maternal and paternal lineages in the entire distribution of ancestors (Ralf and Coop, 2013) and rely on the fact that all individuals are expected to be related over a time scale considering that each of them has 2*n* ancestors from *n* generations ago (Chang, 1999; Rohde et al., 2004).

Recent study of IBD structure in a total of 2257 European samples at a continental scale revealed 1.9 million shared IBD segments with a mean sharing of 10-50 genetic common ancestors from the last 1500 years and exponential increase up to 500 genetic ancestors from the previous 1000 years (Ralf and Coop, 2013).

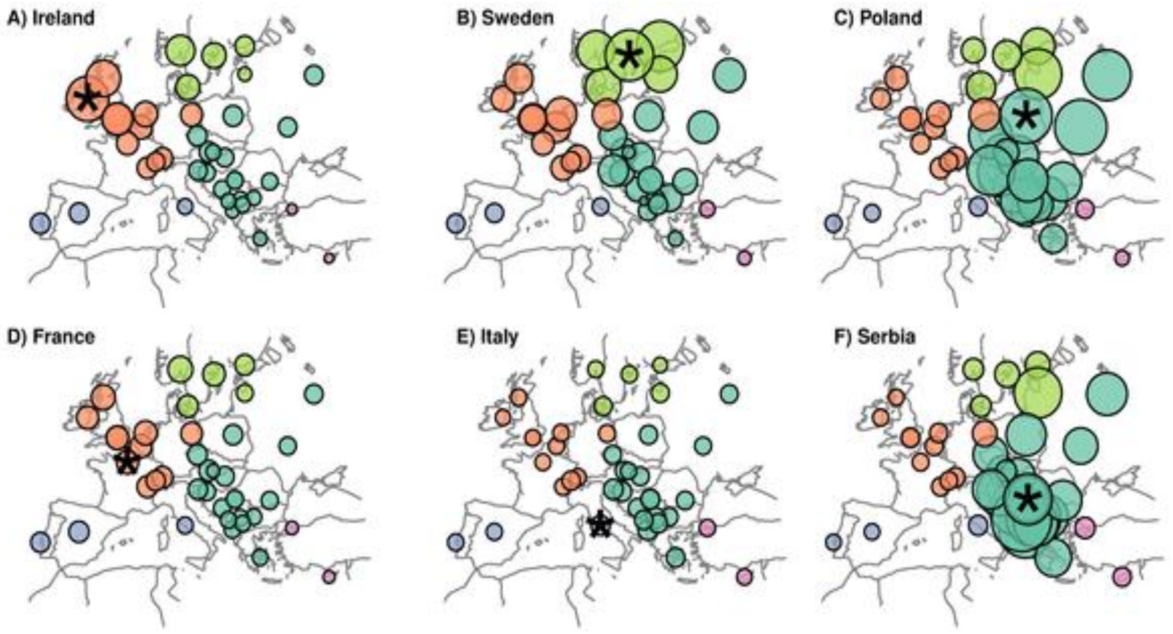

Figure 6. (A–F) Size of the circle for each population depends on the mean number of IBD blocks with length over 1 cM shared between individuals randomly picked from given population (marked with a star) and other European populations (Ralf and Coop, 2013).

Though overall variability among European populations is too complex to make far-reaching conclusions, some evident patterns were identified, such as gradual decrease in rate of IBD sharing as the geographic distance between individuals increases (Figure 6) (Ralf and Coop, 2013), with Italy as an exception which has very low rate of IBD sharing with other populations (Figure 6E). Most of the IBD that Italy shares with other countries is derived from more than 2500 ya. In addition, surprisingly low rates of IBD that Italy shares within itself are found to be as small as with foreign populations. Despite that individuals from the same population tend to share more IBD with members of the same population than with others, this exception reveals significant population substructure in Italy, suggesting that people from different towns share genetically common material as much or as little as from different sides of Europe (Ralf and Coop, 2013).

Besides geographic comparison, distribution of IBD segments length was implemented to infer the number of most recent common ancestors. Shared blocks' size tends to be shorter if their common ancestor lived earlier in time (Pool and Nielsen, 2009; Ralf and Coop, 2013). Numbers of theoretical common ancestors for each of three time periods of 0-500 ya, 500-1500 ya and 1500-2500 ya respectively are shown in Figure 7. IBD segments' size varies from 10 cM and longer during 0-500 ya period, with the highest likelihood for shared segments to belong to individuals from the same population, reaching 4 cM value in the next 500-1500 ya period. In the following 1500-2500 ya time period small shared IBD segments longer than 2 cM are so abundant, that any randomly chosen European individual will have over 50 most recent common ancestors from that period of time (Ralf and Coop, 2013).

The other evident pattern is an increased rate of sharing of large number of common ancestors from about 1500 years ago among southeastern Europeans and a relatively high degree of sharing of IBD between pairs of individuals across eastern Europe and mostly Balkan region (Figure 7. S-C, PL, R-B, Bal populations in 555-1500 and 1515-2535 ya timescale), suggesting that eastern European ancestral population could have been relatively small when it expanded over large geographic region about 1,000-2,000 ya, and could be explained by Slavic and Hunnic expansions (Davies, 2010; Barford, 2001; Ralf and Coop, 2013).

Overall results prove the theory that all Europeans are related through a short period of time, supporting models for close genetic relatedness of almost all present-day populations (Rohde et al., 2004, Ralf and Coop, 2013).

These results correlate with some historical events in Europe and show that the modern genome-wide data is now capable of giving explicit view of recent population and subpopulation history back through time over a few thousands years ago (Ralf and Coop, 2013).

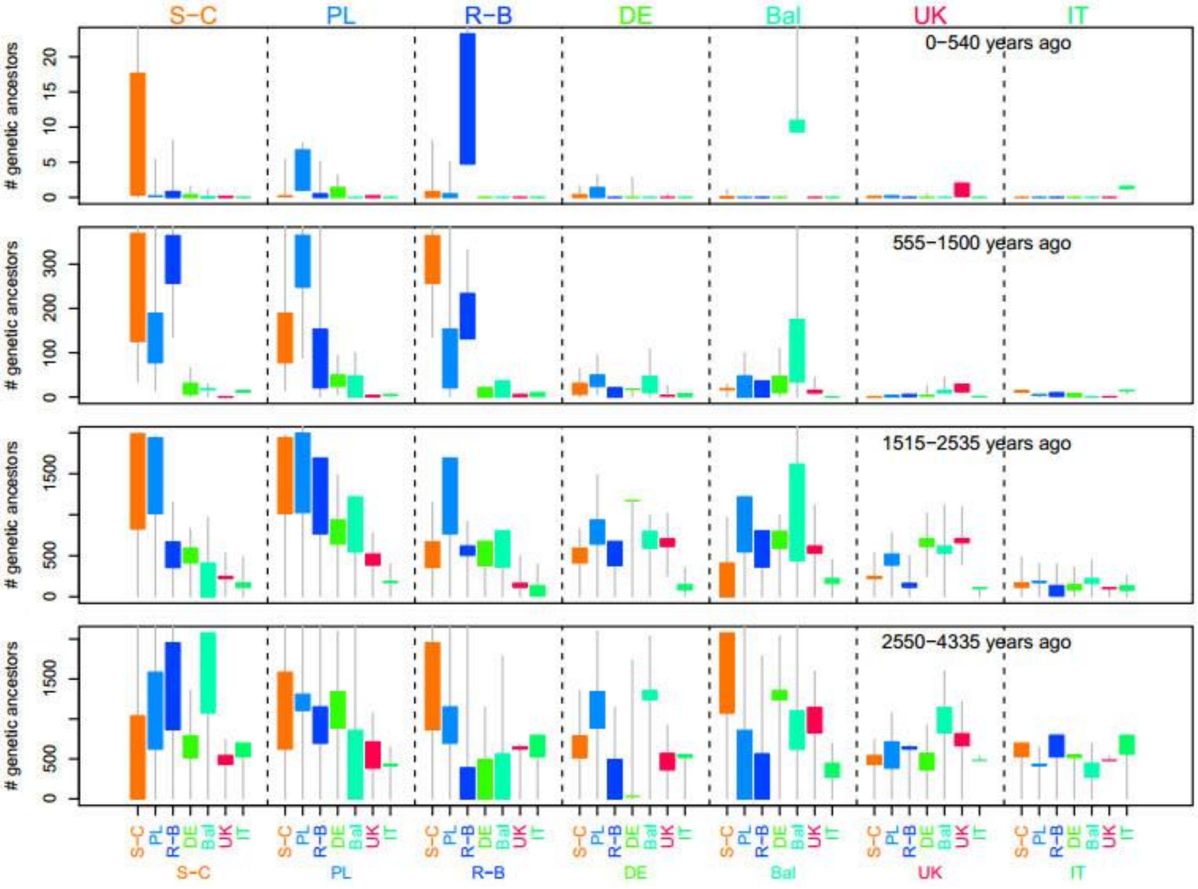

Figure 7. Overall number of estimated genetic common ancestors for different periods of time in POPRES dataset. Some populations were merged to get larger sample sizes: “S-C” means Serbo-Croatian populations in former Yugoslavia, “IT” denotes Italy, “PL” denotes Poland, “R-B” represents Romania and Bulgaria, “DE” denotes Germany, “Bal” denotes Baltic sea region with such combined populations as from Latvia, Finland, Sweden, Norway, and Denmark, ”UK” denotes the United Kingdom. Lower boundaries are always significantly above zero except for the most recent period 0-540 ya (Ralf and Coop, 2013).

Another example of IBD use for revealing recent intrapopulation history is a study by Palamara et al., (2012), where mathematical modeling approach for reconstruction of such important events as founder effect, population shrinking or expansion was inferred using distribution of IBD segments. Method relies on the variation of IBD segments' average length depending on the population size – if population had a small effective size and went through a founder effect very long time ago, then the length of shared IBD segments between pairs of individuals from this population is expected to be short. If population size has only recently decreased, then the length of shared IBD segments tends to be longer. Model accuracy was evaluated using synthetic data from simulated populations with known history and modeling procedure was applied to a real data set of 500 individuals from Ashkenazi Jews (AJ) population and 56 Maasai (MKK) samples from the HapMap data set (Palamara et al., 2012).

Results of demographic inference for AJ population are shown in Figure 8. While considering segments over 5 cM only, simple model of expansion showed a best-fit to the data. However it could not explain the gap in distribution frequency of 2-5 cM IBD segments among the individuals of AJ. More complex model of multiple exponential expansions including founder effect was further proposed and fitted frequency distribution of all detected shorter segments of 1-2 cM length. In overall, the model which fitted best to the real data and frequency distribution of all IBD segments suggested two exponential expansions, separated by a founder event. Population expansion of Ashkenazi Jews started from effective population size of about 2,300 ancestors near 200 generations ago and increasing to 45,000 individuals. After that effective population size dropped to near 270 individuals during the founder event, which took place about 34 generations ago. Subsequently AJ population began expanding rapidly reaching over 4 million individuals at present-day (Palamara et al., 2012).

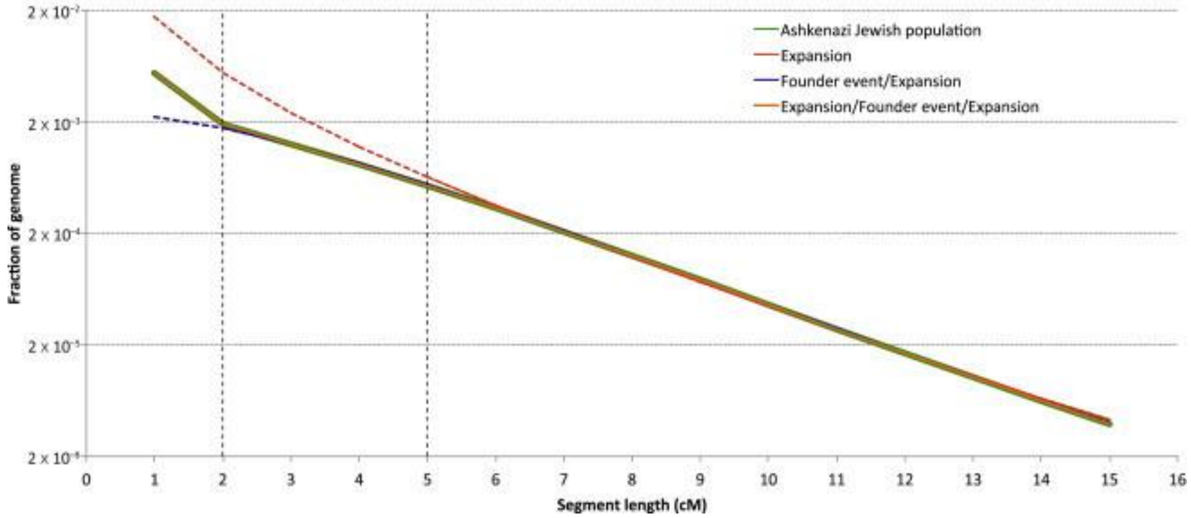

Figure 8. Frequencies of haplotype sharing were used to infer demographic history of population of Ashkenazi Jews by fitting real data derived from HapMap Phase 3 data set with different theoretical models.

Similar approach with Maasai population showed exponential contraction and fit the model where population began with an effective size of about 23,500 individuals and decreasing to present day number of 500 over the past 23 generations. However, these results conflict with studies of MKK by Coast (2001), who proposed slow growth of the Maasai, and with the estimate of about 1 million Maasai people living nowadays (Maasai Assosiation, 2012). This misfit might be explained by the particular population structure, presented by Palamara et al. as a so-called "village effect" model, where large population is divided into few small sub-populations with constant migration events resulting in excess of IBD sharing, or hidden relatedness, which has been previously reported by other studies (Pemberton et al., 2010; Palamara et al., 2012). Best-fit for this model proposed the presence of 44 villages of 485 individuals in each with a constant migration rate of 0.13 (Palamara et al., 2012).

1.2.2. Methods of IBD estimation

First methods for IBD estimation consisted of a simple comparison of the total number of alleles shared across genome between individuals, allowing to detect IBD segments >10 cM long (Weir et al., 2006). Such approaches are suitable for the association studies or for the detection of harmful mutations (Cherny et al., 2001). However, because exponential decrease of length of IBD segments depends on the number of meioses separating the individuals, these methods are suitable for detecting relationships only as distant as third-degree relatives (e.g. aunt or uncle) (Huff et al., 2011). Figure 9 shows the degrees of consanguinity relationships for the reference.

Main toolsets for calculating IBD segments less than 10cM long are BEAGLE, PLINK and GERMLINE (Genetic Error-tolerant regional matching with linear-time extension). While first two are large multi-functional toolsets, GERMLINE is specifically designed for IBD detection. Once haplotypes are inferred, GERMLINE can be used to estimate IBD sharing based on direct matching portions of haplotypes between samples. Advantage of GERMLINE is its computational efficiency. GERMLINE computing time increases linearly with the sample number; this is achieved by holding in memory only a part of the genome while sliding along the homologous chromosomes of given samples. BEAGLE and PLINK algorithms for IBD estimation extend computing time exponentially. This ability of GERMLINE software allows studying large populations with complex datasets simultaneously using genome-wide data (Gusev et al., 2009).

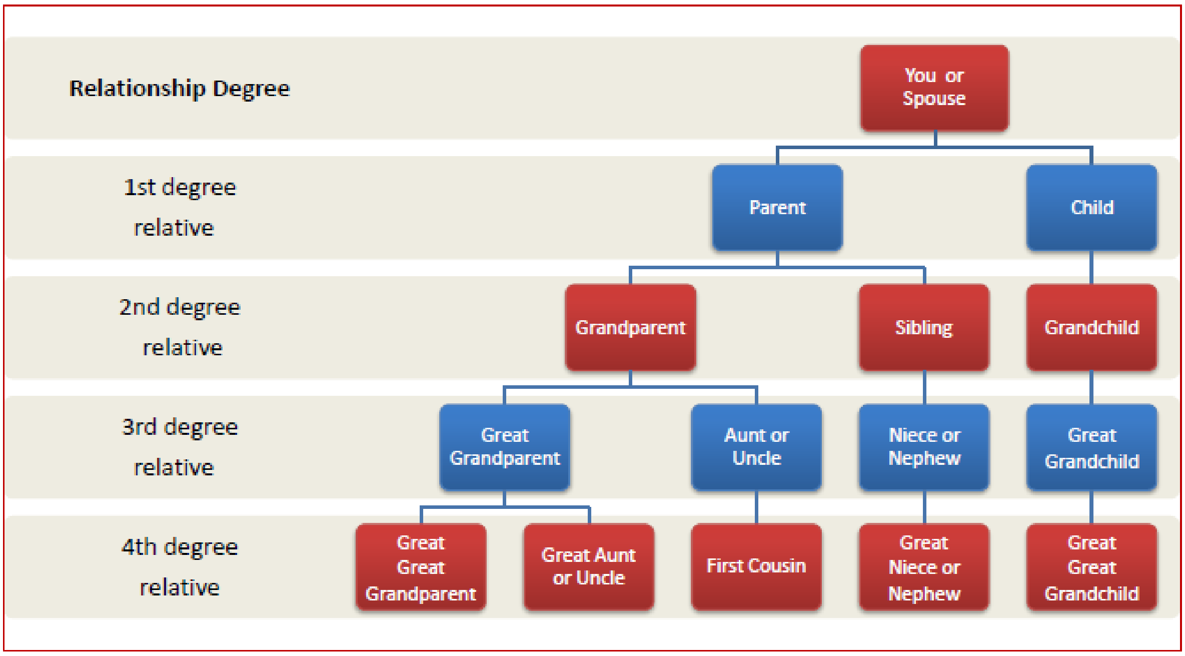

Figure 9. Degrees of consanguinity relationships (Missouri Ethics Commission).

In contrast to GERMLINE which uses haplotype length, PLINK and BEAGLE IBD algorithms are based on haplotype frequencies with built-in Hidden Markov Model (HMM) for calculating the probabilities of IBD status. Prior to building HMM, data must be cleared from independent SNPs in order for the rest of SNP markers set to be in approximate equilibrium in the population (Purcell et al., 2007). This is achieved by setting a threshold for haplotype frequencies to avoid false positive results (FPR). This is especially important in case of the PLINK algorithm, which is very sensitive for background LD and requires pruning of the markers in strong LD beforehand. Because of the SNP data thinning, PLINK is less accurate than BEAGLE (Browning and Browning, 2011).

BEAGLE solves the problem of background LD by implementing a comprehensive LD model, incorporating all local markers. BEAGLE's LD model is based on localized haplotype cluster model, which clusters haplotypes in order to improve prediction of alleles at the next marker after the given one (Browning, 2008). Because of the localization of clustering, haplotypes that belong to the same cluster are expected to be in the next cluster with a certain probability (Browning, 2008).

LD model is incorporated together with IBD model into a single HMM applied in the BEAGLE software package. IBD detection relies on a pairwise score S(H1,H2,m1,m2) between every pair of haplotypes H1 and H2 with any interval of SNPs m1<m2

(Browning and Browning, 2011). In the case of identical HMM sequence for a pair of haplotypes, their pairwise score is the frequency of this shared HMM sequence. Frequencies of HMM sequences for S_m, S_{m+1}, …, S_{m+k} of a given marker *m* can be calculated as a product of state and transition probabilities (Browning and Browning, 2011):

$$P(s_m,\ s_{m+1},\ldots,\ s_{m+k}) = P(s_m)\prod_{j=1}^{k} P\left(s_{m+j}\middle|s_{m+j-1}\right)$$

If two haplotypes are not identical in their sequence of HMM states, then pairwise score is replaced with 100 at each marker for which the HMM states differ. This allows to penalize the pairwise score by inflating the haplotype frequency of shared segments. Smaller score means lower frequency of the shared haplotype and a higher probability for the haplotype to be a true positive IBD segment (Browning and Browning, 2011). For unphased data multiple instances of phased haplotypes are created; switching between them penalizes the total score (Browning and Browning, 2007).

For the purposes of improving computing efficiency, BEAGLE authors have developed the so-called fastIBD approach, which includes both HMM and an opportunity to work with unphased genotypes (Browning and Browning, 2011). While HMM models for whole-genome data demand a lot of computing resources, fastIBD algorithm reduces computing time in a similar manner to GERMLINE software. FastIBD's sliding windows are resized in real-time, depending on the complexity of the haplotypic structure of the genome (Browning and Browning, 2011).

Power to detect IBD is in positive correlation with the mean length of IBD segments and total number of SNP markers. Whole-genome sequencing data provides greater numbers of SNP genotypes for the same segment length than microarray-based genotype data, resulting in different statistical power for IBD detection. Su et al. (2012) showed that fastIBD algorithm is capable of detecting IBD tracts of 0.2 cM with a power of 62.9% using high coverage sequence data (Complete Genomics) (Table 1), while GERMLINE reports 66.5% power to estimate IBD segments of size 0.2 cM for the same type of data. The difference is caused by the high rate of false positives in small segments of IBD detected with GERMLINE (Su et al., 2012). Power falls significantly for both algorithms while using low-density microarray genotype data (WTCCC), but in general, fastIBD shows greater power of IBD detection for low-density genotype data (Table 1).

Table 1. Comparison of statistical power of IBD detection using fastIBD and GERMLINE for different sets of genetic material (Su et al., 2012).

FastIBD					**GERMLINE**			
Segment Size(cM)	**WTCCC**	**HapMap**	**1000g**	**complete**	**WTCCC**	**HapMap**	**1000g**	**complete**
0.2	0.126	0.251	0.562	0.629	0.043	0.344	0.665	0.645
0.4	0.327	0.518	0.781	0.801	0.099	0.551	0.836	0.806
0.6	0.495	0.649	0.864	0.874	0.135	0.617	0.901	0.907
1	0.767	0.840	0.904	0.899	0.231	0.794	0.941	0.944
2	0.909	0.918	0.935	0.919	0.389	0.905	0.982	0.992

A very small number of working data (a total of 15 samples) can produce unreliable haplotype frequencies and skew the potential HMM. Consequently, despite the predicted accuracy of fastIBD algorithm, only GERMLINE was implemented in this study for IBD segment detection.

1.3. GERMLINE algorithms

GERMLINE allows the study of IBD among large cohorts with linear increase of computing time in relation to the sample number. GERMLINE searches for straight matches between portions of haplotypes in given samples without estimating haplotype frequencies, which makes it suitable for low-sample studies.

As an input data structure, GERMLINE takes the 2n x *s* matrix **M** with 2n rows of phased haplotypes from n individuals and *s* columns corresponding to the number of SNP markers. Each slot of the array **H** [*i*,*j*] is defined as a binary vector, with value of 0 if haplotype *i* doesn't have the minor allelic SNP *j*, and 1 otherwise (Gusev et al., 2009).

Identical haplotypes have the same values in **H** and are consequently recorded to the **D** array, which is the so-called haplotype dictionary and is later used for building IBD segments (Gusev et al., 2009).

D has a hash-table data structure, which allows for real-time inserting and searching by the key of a binary vector. Each slot of **D** represents a set of individual haplotypes with identical rows. Every two rows in **D** that have the same key are a match. Set of those matches **M(H)** can be obtained by Algorithm 1 - MATCH (Gusev et al., 2009):

```
Algorithm 1. MATCH(H):
define set M
for $h_i$ in H do D.INSERT ($h_i \rightarrow i$)
for h in D do
  for i in D(h) do
    for i'!=i in D(h) do
      M.add(i,i')
return M
```

As long as IBD segments are not likely to cover all available SNPs in given haplotypes, a threshold for the minimum length of shared identical segment must be defined. In practice, L_{IBD} can be found as the expected length of IBD segment for the most distantly related individuals there is an interest to detect (Gusev et al., 2009). If we slice columns of matrix **H** into distinct “letters”, then a pair of individuals sharing IBD of more than one SNP would have the same “words”, starting in column *j* and ending in *j'*. Thus IBD can be recorded into a set of **M'** where each entry is a quartet (*i,i',j,j'*), where haplotypes *i* and *i'* share the same segment in SNP region [*j..j'*] if L(*j,j'*) exceeds $L_{IBD.}$ Extending complementary matches in neighbouring “letters” is managed by Algorithm 2 - EXTEND (Gusev et al., 2009):

```
Algorithm 2. EXTEND($M_{k-1}$ ', $M_k$ ):
let  $M_k$ ' := $M_k$
for $m_k$ in $M_k$ ' do
i:=  $m_k$ .INDIVIDUAL[1]
i ' :=  $m_k$ .INDIVIDUAL[2]
if $M_{k-1}$ '.CONTAINS(i , i' )
then
$m_{k-1}$ := $M_{k-1}$ '[i , I']
 m.MATCH-START =   $m_{k-1}$ .MATCH-START
$M_{k-1}$'.REMOVE ($m_{k-1}$ )
return $M_w$'
```

Besides L_{IBD}, error rate E per SNP must be considered as long as modern genotyping accuracy is not yet 100% (Paynter et al., 2006). Assuming that E is random and independent, number of mismatches between pair of IBD segments with length δ is Poisson distributed with parameter $\lambda = 2\delta E$. Consisting of number of SNPs as S_{IBD} with Length(S_{IBD}) = L_{IBD}, these segments will have expected number of correct matches according to Equation 5 (Gusev et al., 2009):

Equation 5 :

$$E(N_{\mathrm{IBD}}) = \left\lfloor \frac{s_{\mathrm{IBD}}}{\delta} \right\rfloor e^{-2\varepsilon\delta}$$

Nearly identical matches are handled by the MERGE-PARTIAL algorithm (not presented in this study). MATCH, EXTEND and MERGE-PARTIAL algorithms integrate into a single module Algorithm 3 - HAPLOTYPE-IBD (Gusev et al., 2009):

```
Algorithm 3
HAPLOTYPE-IBD($\mathbf{H}$): given S

given S, $S_{IBD}$, $h_{len}$
let $\mathbf{M}_0$ = MATCH ($\mathbf{H}_0$)
let $\mathbf{M}_0'$ = $\mathbf{M}_0$
define $\mathbf{M}'$
for $k$ in 1 → (S / $h_{len}$) – 1 do
   let $\mathbf{M}_k$ = MATCH ($\mathbf{H}_k$)
   let $\mathbf{M}_k'$ = EXTEND($\mathbf{M}_{k-1}'$, $\mathbf{M}_k$)
$\mathbf{M}$: = EXTEND-PARTIAL ($\mathbf{M}_{k-1}'$, $\mathbf{M}_k$)
for $m_{k-1}$ in $\mathbf{M}_{k-1}'$ do
if LENGTH ($m_{k-1}$) >= $h_{IBD}$
then $\mathbf{M}'$. add ($m_{k-1}$)
return $\mathbf{M}'$
```

Testing GERMLINE software on samples from various populations from HapMap data identified short “gaps” of unusually low IBS with different nature (Gusev et al., 2009) .First type of gaps are considered to be explained by phasing errors, due to the phasing inconsistency occurring from incorrect orientation of haplotypes at heterozygous sites and referred to as “switch” errors (Gusev et al., 2009; Lin et al., 2002; Marchini et al., 2006). The second type of gaps are considered to be structural variants and are validated by searching for overlapping deletion regions in the Database of Genomic Variants (Gusev et al., 2009; Iafrate et al., 2004).

As an example of using GERMLINE for IBD detection, Gusev et al., 2009 analyze the entire adult population of the Island Kosrae, Micronesia. Phasing was made with BEAGLE and then GERMLINE was implemented to scan for shared segments over 10 cM in length in compliance with consensus database (Duffy, 2006) of standard genetic maps (Lien et al., 2006; Kong et al., 2004). GERMLINE's efficiency and ability to detect IBD in agreement with theoretical expectations for individuals up to 4 meioses apart is shown in Figure 10. Excess of sharing in real data compared with what is expected in 5-7 meiosis range might be explained by the presence of unreported relatedness due to the small size of the ancestral population of the Island Kosrae (Gusev et al., 2009).

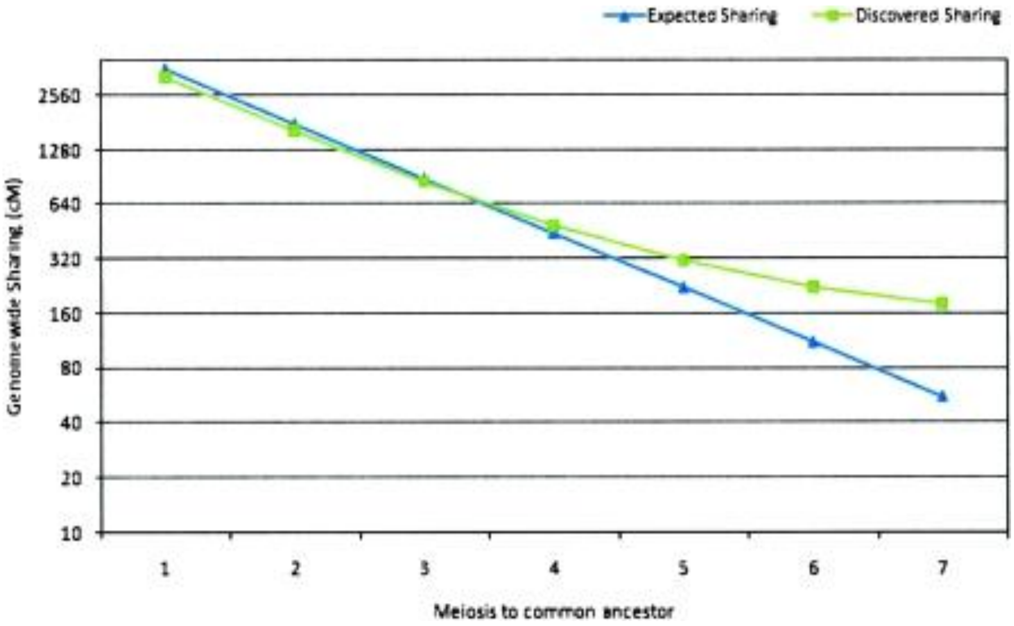

Figure 10. Expected and detected genome-wide sharing in Kosrae Cohort for IBD segments over 10 cM. Expected genome-wide sharing is shown in blue, detected genome-wide sharing is shown in green (Gusev et al., 2009).

2. EXPERIMENTAL PART

2.1. Aim of the study

The main goal of the current study is to test for unreported relatedness using whole-genome genotyping data from 15 Swedish Saami individuals and to exclude samples of close relatives from future use in population genetics research. In addition, we want to explore genetic IBD background in the Saami population. Current experimental work is based on implementing BEAGLE software for genotype phasing and GERMLINE software for IBD segment detection.

2.2. Materials and methods

2.2.1. Genotyping data

A total of 15 new Swedish Saami samples genotyped using Illumina 610K bead array (~500,000 SNPs) are reported here for the first time (see Supplementary Table 1 for additional sample details). Genotype to haplotype reconstruction (phasing) of all data was performed with BEAGLE phasing algorithm.

2.2.2. Data processing

All data processing and computation was performed remotely on the University of Tartu and Estonian Biocentre taevas.hpc.ut.ee computing cluster. The analysis algorithm and step-by-step workflow was as follows:

a) Genotype data from 22 chromosomes was extracted from master-file in PLINK
b) 1.07 software (Purcell et al., 2007) for all 15 samples. Physical distance (Mb) was converted into genetic distance (cM) in accordance with HapMap phased data build 1.36;
c) The output file was further separated into 22 files for each autosomal chromosome;
d) fcgene 1.02 software was used to convert each file into BEAGLE format;

e) BEAGLE software was used to phase genotype data with default settings for each chromosome separately;

f) fcgene 1.02 software was further used to convert each phased chromosome back into PLINK format;

g) GERMLINE 1.5.1 was implemented for each phased chromosome with default settings except for the -bits parameter, which determines the number of SNP markers per one sliding window. Lowering this parameter results in more accurate IBD estimation and slower computing time. -bits value was to 40 in accordance with the recommendation to keep this parameter close to the number of SNPs per 0.2 cM (Browning and Browning, 2013). Other parameters were set to default as follows: -min_m = 3 (length threshold for matching segments to be used for imputation); -err_hom = 2 (amount of mismatching homozygous markers allowed to be present in matching segments); -err_het = 3 (amount of mismatching heterozygous markers allowed to be present in matching segments).

h) Output was compiled into a single csv file and analyzed with R statistical software v.2.15.1, Open Office Calc v.4.1.0 and on-line services. Frequency distribution histograms for mean length of IBD segments were drawn with R v2.15.1 and Open Office Calc v.4.1.0.

2.3. Results and discussion

Distribution of all detected shared segments among a total of 15 Saami individuals is exponential, which is consistent with theoretical expectations (Figure 11). Overall number of detected segments is 5667, with the maximum value of 278.02 cM. Mean length of all segments is estimated as 7.99 cM, while median value is 5.31 cM, first quartile(25) is 3.85 cM and third quartile is 7.95 cM respectively (see Supplementary Table 2 for GERMLINE output data).

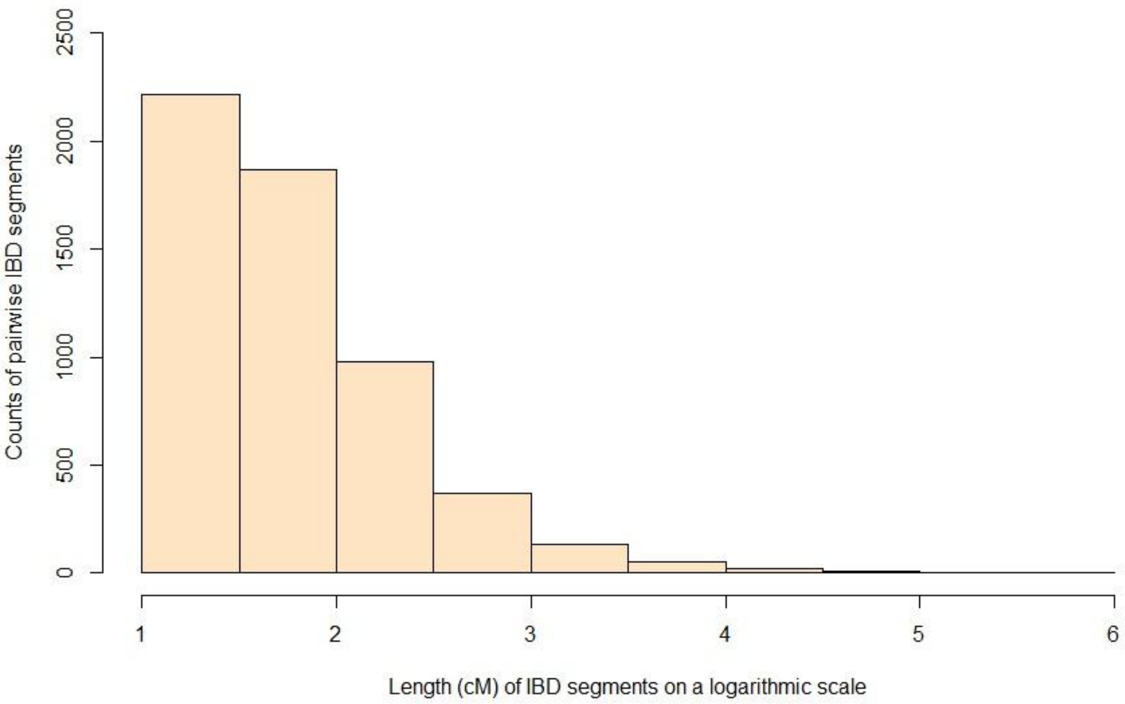

Figure 11. Counts of IBD segments length in 15 Saami samples.

Mean pairwise length of IBD segments is presented in Table 2. Mean length of IBD segments is probably overestimated due to the given threshold of 3 cM for counting segments as true IBD. On the other hand, IBD segments less than 3 cM relay the presence of relationships far back in time between many common ancestors. The goal of this work is to shed light on more recent kinship. In addition, elimination of such small segments reduces the number of false positives. The generally accepted level of kinship for samples used in population genetics studies does not exceed third-degree relatives (8 meioses), who are expected to share mean length of IBD segments around 12.5 cM, which stems from the theoretical distribution of shared ancestry with a mean of N^{-1} M (or $100*N^{-1}$ cM), where N is the number of meioses (Browning and Browning, 2012) (this does not take into account other factors affecting mean length of IBD). Candidates for third degree cousins or closer relationships are marked with yellow in Table 2. The rest of the pairs show relatively high sharing as well, with only 4 pairs from a total of 105 having less than 5 cM in common, corresponding to 20 meioses. While close relatives can be reliably detected, more distant levels of kinship show very high levels of uncertainty (Ralf and Coop, 2013; Henn et al., 2012).

Table 2**.** Pairwise comparison of mean IBD segment length (cM) estimated with GERMLINE. Candidates for third degree cousins or closer are marked in yellow.

Sample ID	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
1	6,81	6,53	7,03	6,63	5,35	6,14	4,57	5,87	6,65	7,04	5,75	6,04	9,25	5,96	
2	6,28	6,07	5,35	5,27	9,69	6,78	5,39	6,44	7,68	5,9	6,99	5,08	5,69		
3	8,18	7,9	6,65	7,33	6,24	5,97	5,3	5,98	6,94	13,99	6,04	7,26			
4	8,54	7,55	8,98	8,85	6,01	5,42	4,82	5,07	5,33	7,64	5,79				
5	5,83	6,09	6,5	5,94	6,71	24,35	6,22	6,52	6,54	5,65					
6	7,42	8,23	9,68	6,65	6,04	5,6	4,88	5,64	5,99						
7	6,15	6,74	6,6	5,23	8,63	7,8	5,76	12,4							
8	5,66	5,61	5,74	5,7	6,94	6,63	5,89								
9	5,15	4,89	5,46	5,15	5,35	5,45									
10	5,32	5,13	5,64	5,53	7										
11	5,39	5,94	6,32	5,02											
12	73,53	6,84	6,69												
13	7,33	9,57													
14	7,54														
15															

Histogram and Q-Q plot of pairwise mean IBD sharing are shown in Figure 12. Mean IBD sharing of the sample set approximately follows the normal distribution, but is slightly skewed to the right (Figure 12A).

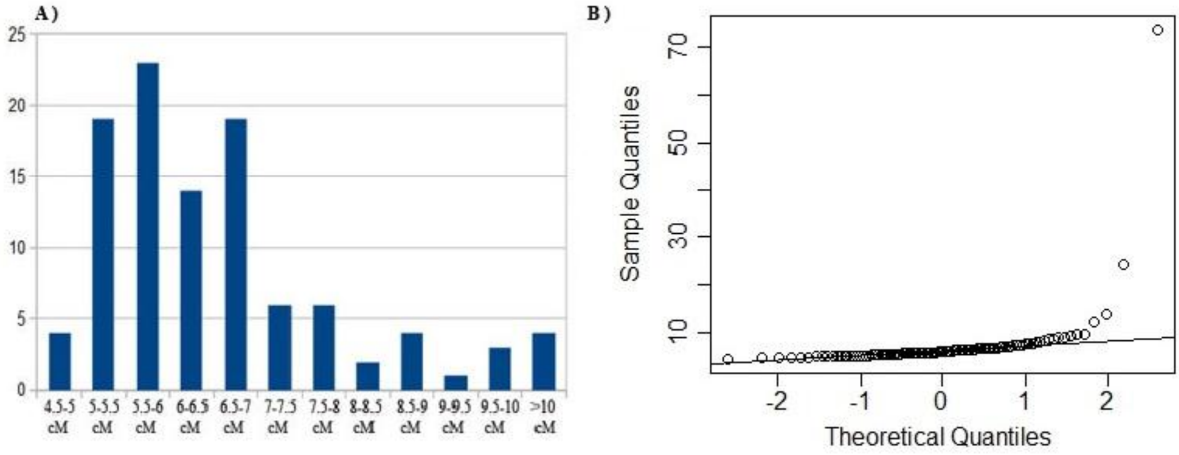

Figure 12. Distribution of pairwise mean IBD length among the Saami sample set. (A) Distribution of frequencies. (B) Q-Q plot of the distribution.

Box-plot with Tukey whisker extents is shown in Figure 13. All pairs marked yellow in Table 2 are outliers on the box-plot, supporting their higher than normal IBD sharing status in the current dataset. One sample from pairs 12/15, 5/10, 6/3 should be excluded from further population genetic studies. In comparison to their pairmates, samples number 12, 3 and 5 show higher levels of IBD sharing with the rest of the sample set and are recommended for exclusion. Other samples that are designated as outliers on the boxplot are below the set threshold of 12.5 cM.

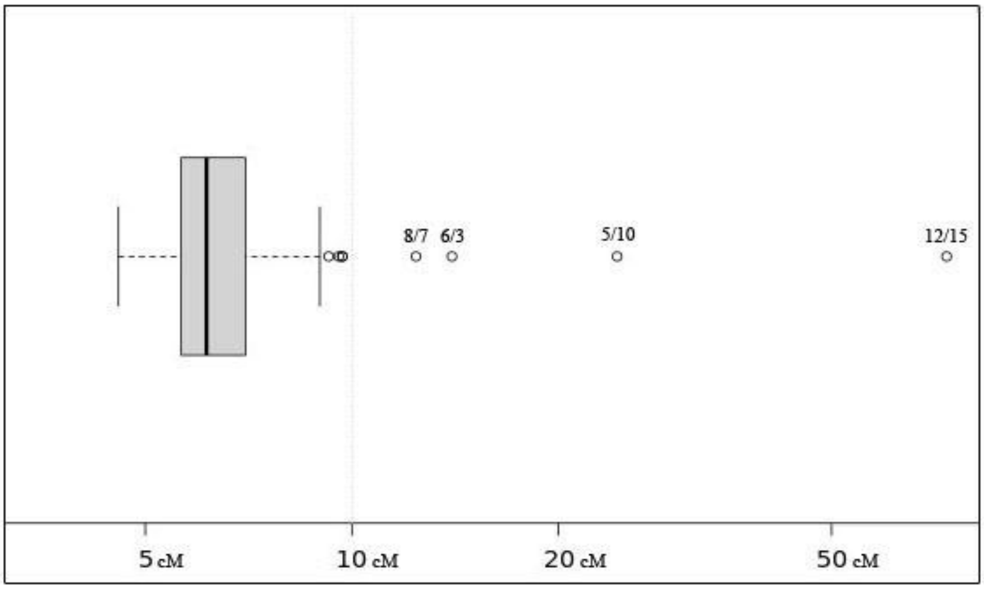

Figure 13. Box-plot of pairwise mean IBD sharing distribution with Tukey whisker extents.

Results by Henn et al.(2012) were used for comparison of IBD background levels between populations. The fraction of sample pairs that share at least one IBD segment greater than or equal to 7 cM, denoted as $F_{IBD,}$ was calculated for the Saami and compared with data from Henn et al.(2012) (Table 3). F_{IBD} resulted in the highest value of 1 and is comparable with such endogamous populations as Surui or Kalash. This also supports retaining the sample pair 8/7, as its excess of IBD sharing could be explained by the generally high IBD background rather than by unreported recent kinship. Such high level of hidden relatedness might be caused by the small effective size and an endogamic population structure of the Saami population.

Table 3. Comparative analysis of Saami samples with the results of Henn et al. (2012). Data from original study is shown in black, current study data shown in red, F_{IBD} denotes the fraction of sample pairs sharing at least a single IBD segment greater than or equal to 7 cM.

Population	**Sample size**	**F_{IBD}**	**Source**
Surui	8	1	Henn et al. (2012)
Karitiana	14	0,88	Henn et al. (2012)
Kalash	23	1	Henn et al. (2012)
Yakut	25	0,92	Henn et al. (2012)
Biaka Pygmies	21	0,96	Henn et al. (2012)
Maya	21	0,47	Henn et al. (2012)
Sardinian	28	0,38	Henn et al. (2012)
Tuscan	8	0,43	Henn et al. (2012)
Ashkenazi	847	0,85	Henn et al. (2012)
Finland	149	0,53	Henn et al. (2012)
Yoruba	21	0,06	Henn et al. (2012)
Canada	373	0,04	Henn et al. (2012)
Han Chinese	44	0,01	Henn et al. (2012)
Italy	386	0,01	Henn et al. (2012)
Saami	15	1	This study

Comparison of sampled data with artificial genetic diversity generated by simulating populations with the same effective size, but under different demographic scenarios, such as shrinking, expanding or retention of constant size could be an avenue for further research. For better interpretation of the effect of demographic history on the extant genetic diversity of the Saami and other small populations, further work should also include additional samples and integration of our knowledge about the lifestyle and history of such isolated human groups.

CONCLUSION

Recombination is an important source of genetic diversity and, if properly detected, reveals net of relationships among people within population. Recombination results in the fragmentation of genome. Alleles within linked genomic fragments, also known as haplotypes, are co-inherited until the new recombination event or mutation will give rise to a new variant of a haplotype. As more generations pass, portion of genetic material inherited from a particular ancestor and its length decreases nearly exponentially. Segments, which remained untouched by recombination and were identical between individuals, are considered identical by descent (IBD) and reflect a certain degree of kinship. As the geographic distance between individuals increases, rate of IBD sharing decreases gradually. This study is based on implementing modern software on whole-genome genotyping data for IBD detection and aimed at researching the distribution of IBD segments lengths among 15 individuals of Swedish Saami descent. Phasing of data was performed with BEAGLE software. GERMLINE was used for detection of IBD segments and PLINK software was used for general data handling operations. Results were further tested for unreported and hidden relatedness by analyzing levels of shared IBD background. Unreported relatedness was detected on a background of excess IBD sharing, which sheds light on the recent demographic history of the Saami and reveals candidate pairs for relationships closer than 3rd degree relatives.

From our results we can conclude that:

1. A sample set of 15 Swedish Saami individuals shares more IBD segments between themselves than expected under the assumption of large effective population size and a population sample consisting only of unrelated individuals.

2. Samples 12 and 15 are close relatives (up to first cousins), 10 and 5 are related with degree up to second cousins or closer (Figure 9), and samples 6 and 3 show degree of relatedness up to sixth cousins. One sample from each of these pairs must be excluded from further population studies.

3. High levels of IBD background revealed in the sample set of Swedish Saami individuals result in the pairwise sharing of at least a single IBD segment of 7 cM or greater length among all sample pairs. This may be caused by a very low effective population size of the Saami or its recent decrease in comparison with other populations or the previously described “village effect”.

Identse päritolu hindamine GERMLINE tarkvaraga
Saami populatsiooni kuuluvate indiviidide näitel

Dmitri Lomovski

Resümee

Uued genoomide sekveneerimistehnoloogiad on kaasa toonud ka uued meetodid andmetöötluses, mille eesmärgiks on leida olulisi inimgenoomide ülesehitust ja inimkonna demograafilist ajalugu ühendavaid seaduspärasusi. Inimestevahelised erinevused genoomis tekivad nii mutatsioonide kui rekombinatsiooni tulemusena. Rekombinatsioon lõhub genoomi väiksemateks osadeks, mis omavahel ristudes moodustavad üha uusi geneetilisi kombinatsioone ja pärandatakse järgnevatele põlvkondadele kuni järgmise rekombinatsioonini, mis lõhub eelnevalt tekkinud genoomset segmenti. Kui kahel inimesel avastatakse suur hulk sarnaseid teatud pikkusega haplotüüpe, siis vihjab see vähemalt ühe ühise esivanema olemasolule lähiminevikus. Sugulusaste sõltub päritolult identsete segmentide arvust ja pikkusest, mis kahaneb aja möödudes ja meiooside hulga kasvades eksponentsiaalselt. Lisaks on näidatud ka, et päritolult identse geneetilise materjali hulk väheneb korrelatsioonis võrreldavate populatsioonide geograafilise kaugusega.

Kaasaegne genotüpiseerimisandmete töötluseks mõeldud tarkvara võimaldab tuvastada genotüüpide identseid järjestusi kahe indiviidi vahel ja hinnata nii tõenäolist otsest sugulusastet kui ka niinimetatud „varjatuid sugulussidemeid“, mis eriti iseloomustavad endogaamseid populatsioone, kujutades endast suurt hulka lühikesi identseid segmente, mis on päritud paljudelt ühistelt esivanematelt. Käesoleva töö eesmärgiks oli hinnata päritolult identsete genoomisegmentide jaotumist Rootsi saamide populatsioonist pärit 15 indiviidi näitel. Genotüpiseerimisandmete faasimiseks kasutati BEAGLE tarkvara, sarnaste segmentide tuvastamiseks GERMLINE programmi. Eelnevat sisendandmete kvaliteedikontrolli teostati programmi PLINK abil. Leitud segmentidega teostati statistiline analüüs ja tulemused viitasid kõrgele päritolusamasuse taustale uuritavasse populatsiooni kuuluvate indiviidide vahel.

Osad indiviidid jagasid omavahel väga pikki identseid segmente, mis vihjab nende lähedasele veresugulusele ja nende genotüpiseerimistulemusi pole soovitatav kasutada järgnevates populatsioonigeneetilistes uuringutes. Kuna mistahes sõltumatult valitud indiviidid võivad omavahel omada rohkem kui ühte ühist esivanemat, siis sugulusastme täpne määramine tuginedes vaid haplotüüpide samasuse keskmisele pikkusele on mõnevõrra raskendatud, kuid annab piisavat tõestust lähedaste (nt nõbude kui 3. sugulusastme esindajate)

veresugulussidemete olemasolule. Kaugemate sugulusastmete korral võib sarnane lähenemine põhjustada võrreldavaid indiviide viimasest ühisest esivanemast lahutava meiooside arvu ülehindamisele. Identsete segmentide keskmistatud pikkust kahe indiviidi vahel tuleb võrrelda populatsiooni üldise geneetilise samasuse taustal, mille kõrge tase Rootsi saamidel vihjab kas väga väikesele efektiivse populatsiooni suurusele või elukorraldusele, kus suurem populatsioon jaguneb väiksemateks alampopulatsioonideks (nn „küla efekt“), mille piires toimub sarnaste haplotüübide fikseerumine kiiremini kui terves populatsioonis. Täpsemate põhjuste väljaselgitamiseks on vajalikud edasised populatsioongeneetilised uuringud, mis hõlmavad suuremal või vähemal määral isoleeritud väikerahvad.

REFERENCES

A) Journals

(2003). The International HapMap Project. Nature 426(6968): 789-796.

(2005). A haplotype map of the human genome. Nature 437(7063): 1299-1320.

Abecasis, G.R. et al., (2005). Linkage disequilibrium: ancient history drives the new genetics. Hum Hered 59(2):118-124.

Albrechtsen, A. et al., (2009). Relatedness mapping and tracts of relatedness for genome wide data in the presence of linkage disequilibrium. Genetic Epidemiology 33 (3): 266–274.

Ardlie, K.G. et al., (2002). Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3(4):299-309.

Barreiro, L. B. et al., (2008). Natural selection has driven population differentiation in modern humans. Nature Genetics 40(3): 340–345.

Browning, S.R. (2008). Estimation of Pairwise Identity by Descent From Dense Genetic Marker Data in a Population Sample of Haplotypes. Genetics 178(4): 2123–2132.

Browning, B.L. and Browning, S.R. (2012). Identity by descent between distant relatives: detection and applications. Annu Rev Genet 46:617-33.

Browning, B.L. and Browning, S.R. (2013). Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics 194(2): 459–471.

Browning, B.L. and Browning, S.R. (2011). A Fast, Powerful Method for Detecting Identity by Descent. Am J Hum Genet 88(2): 173–182.

Browning, S.R. (2006). Multilocus Association Mapping Using Variable-Length Markov Chains. Am J Hum Genet 78(6): 903–913.

Browning, B.L. and Browning, S.R. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am J Hum Genet 81(5):1084-97.

Bu, Y. and Cao, D. (2012). The origin of cancer stem cells. Front Biosci S4: 819-830.

Cardon, L. R. and Abecasis, G. R. (2003). Using haplotype blocks to map human complex trait loci. Trends Genet 19(3): 135-140.

Chang, J.T. (1999). Recent common ancestors of all present-day individuals. Advances in Applied Probability 31(4): 855-1154.

Cherny, R.A. et al., (2001). Treatment with a copper-zinc chelator markedly and rapidly inhibits beta-amyloid accumulation in Alzheimer's disease transgenic mice. Neuron 30(3):665-76.

Coast, E. (2001). Maasai demography. PhD Thesis, University of London, University College London.

Duffy D.L.(2006). An integrated genetic map for linkage analysis. Behav Genet 36(1):4-6.

Feuk, L. et al., (2006). Structural variation in the human genome. Nat Rev Genet 7(2):85-97.

Gabriel, S.B. et al., (2002). The Structure of Haplotype Blocks in the Human Genome. Science 296(5576):2225-9.

Gusev, A., et al., (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res 19(2):318-26.

Henn, B.M. et al (2012). Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. PLoS ONE 7(4): e34267.

Huff, C. et al., (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res 21(5): 768–774.

Iafrate, A.J. et al., (2004). Detection of large-scale variation in the human genome. Nat Genet 36(9):949-51.

Kong, A. et al., (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genetics 40: 1068–1075.

Kong, X. et al., (2004). A combined linkage-physical map of the human genome. Am J Hum Genet 75(6):1143-8.

Laan, M. and Pääbo, S. (1997). Demographic history and linkage disequilibrium in human populations. Nat Genet 17(4):435-8.

Lien, S. et al., (2006). Evidence for heterogeneity in recombination in the human pseudoautosomal region: High resolution analysis by sperm typing and radiation-hybrid mapping. Am J Hum Genet 66(2): 557–566.

Lin, S. et al., (2002). Haplotype inference in random population samples. Am J Hum Genet 71(5):1129-37.

Marchini, J. et al., (2006) A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78(3):437-50.

Marth, G.T. et al., (2004). The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. Genetics 166(1): 351–372.

Mueller, J.C. (2004). Linkage disequilibrium for different scales and applications. Brief Bioinform 5(4):355-64.

Palamara, P.F. et al., (2012). The architecture of longrange haplotypes shared within and across populations. Mol Biol Evol 29(2):473-86.

Patil, N. et al., (2001). Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. Science 294(5547):1719-23.

Paynter, R.A. et al., (2006). Accuracy of multiplexed Illumina platform-based single-nucleotide polymorphism genotyping compared between genomic and whole genome amplified DNA collected from multiple sources. Cancer Epidemiol Biomarkers Prev 15(12):2533-6.

Pemberton, T.J. et al., (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. Am J Hum Genet 87(4):457-64.

Pool, J.E. and Nielsen, R. (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181(2): 711–719.

Purcell, S. et al., (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81(3): 559–575.

Ralf, P. and Coop, G. (2013). The geography of recent genetic ancestry across Europe. PLoS Biology 11(5): e1001555.

Roach, J.C. et al., (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science 328(5978):636-9.

Rohde, D.L.T. et al., (2004). Modeling the recent common ancestry of all living humans. Nature 431(7008):562-566.

Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. Nat Genet 13(10):745-53.

Shaikh, T.H. et al., (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res 19(9):1682-90.

Sindi, S. et al., (2009). A Geometric Approach for Classification and Comparison of Structural Variants. Bioinformatics 25(12):i222-30.

Slatkin, M.(2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9(6):477-85.

Su, S. et al., (2012). Detection of identity by descent using next-generation whole genome sequencing data. BMC Bioinformatics 13:121.

Tambets, K. et al., (2004). The Western and Eastern Roots of the Saami—the Story of Genetic “Outliers” Told by Mitochondrial DNA and Y Chromosomes. Am J Hum Genet 74(4):661-82.

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422): 56–65

Wall, J.D. and Pritchard, J.K. (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73(3):502-15.

Weir, B.S. et al.(2006). Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet 7(10):771-80.

Whyte, A. L. et al., (2005). Human evolution in Polynesia. Hum Biol 77(2):157-77.

B) Books

Barford, P.M. (2001). The early Slavs: culture and society in early medieval Eastern Europe. Cornell Univ Press.

Davies, N. (2010). Europe: A History. Random House.

Lodish, H. et al., (2000). Molecular cell biology, 4th edition. New York: W. H. Freeman.

Sutton, D. G. (1994). The Origins of the First New Zealanders. Auckland University Press.

USED WEB ADDRESSES

GERMLINE's homepage http://www.cs.columbia.edu/~gusev/germline/
Maasai Assosiation website http://maasai-association.org
Missouri Ethics Commission http://www.mec.mo.gov/
NCBI dbSNP build 137 for Homo Sapiens http://ncbi.com
On-line service for generating box-plot http://boxplot.bio.ed.ac.uk
On-line service for generating Q-Q plot http://scistatcalc.blogspot.com/2013/11/q-q-plotter-for-gaussian-distribution.html
Perez, Nancy. "Meiosis". http://www.web-books.com/MoBio/Free/Ch8C.htm

ACKNOWLEDGEMENTS

I would like to express gratitude to my supervisor Anne-Mai Ilumäe for all the support, help and experience she have shared with me.

SUPPLEMENTARY DATA

Supplementary Table 1. Origin of samples used in current study.

Sample ID	Population	State	Country	Language	Source	Chip
saami1	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami2	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami3	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami4	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami5	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami6	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami7	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami8	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami9	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami10	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami11	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami12	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami13	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami14	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K
saami15	Saami	Norrbotten	Sweden	U/F-U/F *	This study	610K

- Herein U/F-U/F denotes for Uralic/Finno-Ugric/Finnic

Supplementary Table 2 including detailed data for all IBD segments detected with GERMLINE in this study can be downloaded from the following address: https://www.dropbox.com/s/wtflo29g6co85pv/SupplementaryTable2.xls

Non-exclusive licence to reproduce thesis and make thesis public

I, Dmitry Lomovsky (date of birth: 11.01.1986),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Estimating identity by descent with GERMLINE software in individuals from the Saami population

supervised by Anne-Mai Ilumäe and Georgi Hudjashov.

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 26.05.2014