GENETIC POPULATION STRUCTURE IN FINLAND:

AN ADVANTAGE OF HAPLOTYPE INFORMATION OVER INDEPENDENT GENETIC MARKERS

SINI KERMINEN

MASTER'S THESIS
BIOTECHNOLOGY
DEPARTMENT OF BIOSCIENCES
UNIVERSITY OF HELSINKI
APRIL 2015

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Biological and Environmental Sciences	Laitos – Institution– Department Department of Bioscience

Tekijä – Författare – Author
Sini Kerminen

Työn nimi – Arbetets titel – Title
Genetic population structure in Finland: an advantage of haplotype information over independent genetic markers

Oppiaine – Läroämne – Subject
Biotechnology

Työn laji – Arbetets art – Level Master’s Thesis	Aika – Datum – Month and year April 2015	Sivumäärä – Sidoantal – Number of pages 70

Tiivistelmä – Referat – Abstract

Studies of population structure are motivated by the need to understand population history and to have well-characterised groups of individuals in studies of genetics of diseases and traits.

A standard method to analyse genetic population structure is principal component analysis (PCA). A disadvantage of PCA is that it can reliably handle only independent genetic markers. This means that the genetic markers that are correlated with other genetic markers have to be excluded from the data. This leads to a loss of information.

In 2012, Lawson et al. published a chromosome painting method that can utilise haplotype information, i.e. information from correlated markers, and thus it can detect more subtle differences in populations than the standard PCA.

This thesis studies two questions. The first question is whether the chromosome painting method can provide more precise genetic clustering of geographically defined Finnish groups than the standard PCA method. The second question is whether the chromosome painting method can reveal new details of population structure in Finland.

The data used in this study are from the FINRISK Study survey of 1997. This cohort includes the genotype data of about 4,000 individuals and the information about individuals’ and their parents’ birthplaces. 345 Individuals were randomly chosen from the cohort in such a way that both of their parents were originated from the same province. Ten provinces of Finland were used as study groups for the method comparison. First, the data were analysed with SmartPCA (a standard PCA method) and ChromoPainter (the chromosome painting method) and the results were compared both visually and quantitatively. Finally, the individuals were assigned to populations based on the ChromoPainter result using FineSTRUCTURE program and these genetic populations were compared to the geographic origin of the individuals.

The results showed that the chromosome painting method clustered seven out of ten groups significantly tighter than the standard PCA. Nevertheless, SmartPCA was faster and easier to use than ChromoPainter. The main population genetic division was found between the eastern and western parts of Finland, which was consistent with earlier studies. All in all, 15 populations were detected and the results revealed that they were geographically clustered. The genetic populations correlated well with the borders of Finnish provinces and counties.

As the first conclusion, the chromosome painting method was able to give more precise results than the standard PCA but the standard PCA is still more suitable for quick preliminary analyses of genetic data. As the second conclusion, the chromosome painting method was able to detect detailed subpopulation structure in Finland and these populations are geographically clustered. Results provide an excellent basis for the future studies of population structure and genetic diseases in Finland.

Avainsanat – Nyckelord – Keywords
population structure, Finland, principal component analysis, PCA, chromosome painting method, ChromoPainter, genome-wide marker data, haplotype, FineSTRUCTURE

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Matti Pirinen and Samuli Ripatti

Säilytyspaikka – Förvaringställe – Where deposited
University of Helsinki, Viikki Campus Library

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty
Bio- ja ympäristötieteellinen tiedekunta

Laitos – Institution– Department
Biotieteiden laitos

Tekijä – Författare – Author
Sini Kerminen

Työn nimi – Arbetets titel – Title
Suomen geneettinen populaatiorakenne: haplotyyppi-informaatioon ja riippumattomiin geenimerkkeihin perustuvien menetelmien vertailua

Oppiaine – Läroämne – Subject
Biotekniikka

Työn laji – Arbetets art – Level
Pro Gradu -tutkielma

Aika – Datum – Month and year
Huhtikuu 2015

Sivumäärä – Sidoantal – Number of pages
70

Tiivistelmä – Referat – Abstract

Populaatiorakenteen tutkimusta motivoi parhaiten ihmisen halu tuntea alkuperänsä. Lisäksi populaatiorakenteen ymmärtäminen on tärkeää geneettisten sairauksien tutkimuksessa. Esimerkiksi tapaus-verrokkitutkimuksissa tutkittavien ryhmien tulisi vastata toisiaan eri muuttujien, myös geneettisen rakenteen, suhteen. Jos ryhmien välillä on otannasta johtuvaa geneettistä eroa, voivat tulokset johtaa väärään johtopäätökseen, jossa tutkittava asia assosioituu perimään virheellisesti.

Yksi käytetyimmistä menetelmistä yksilöiden välisten geneettisten erojen havaitsemiseen on pääkomponenttianalyysi (PCA). PCA:n ongelma on kuitenkin se, että se voi analysoida luotettavasti ainoastaan riippumattomia geenimerkkejä. Tämä tarkoittaa, että geneettisestä aineistosta on poistettava kytkeytyneet geenimerkit ja tietoa yksilöiden välisistä pienistä eroista katoaa.

Vuonna 2012 Lawson ym. kehittivät menetelmän, joka pystyy huomioimaan myös kytkeytyneet geenimerkit. Tämä kromosomin jaottelumenetelmä tutkii yksilöiden genomia haplotyypeittäin. Näin ollen kromosomin jaottelumenetelmä mahdollistaa yksityiskohtaisemman populaatioiden tutkimisen kuin perinteinen pääkomponenttianalyysi.

Tämä tutkielma keskittyy kahteen pääkysymykseen. Ensimmäinen kysymys on, pystyykö kromosomin jaottelumenetelmä erottamaan ja ryhmittämään maantieteellisesti samoilta alueilta peräisin olevat yksilöt paremmin kuin perinteinen PCA-menetelmä. Toinen kysymys on, pystyykö kromosomin jaottelumenetelmä löytämään uusia hienorakenteita Suomen populaatiorakenteesta.

Tutkimuksen käytettiin aineistona suomalaista FINRISKI 1997 -tutkimusta, joka koostuu n. 4 000 yksilön genotyyppiaineistosta sekä yksilöiden ja heidän vanhempiensa syntymäpaikkatiedoista. Aineistosta valittiin satunnaisesti 345 yksilöä, joiden molemmat vanhemmat ovat kotoisin samasta läänistä. Tutkimuksessa oli mukana yksilöitä kymmenestä läänistä ja läänejä käsiteltiin vertailussa omina ryhminään. Ensimmäiseksi aineisto analysoitiin käyttäen SmartPCA (perinteinen PCA-menetelmä) sekä ChromoPainter (kromosomin jaottelumenetelmä) -ohjelmia ja tuloksia vertailtiin sekä visuaalisesti että kvantitatiivisesti. Tämän jälkeen yksilöt jaettiin populaatioihin kromosomin jaottelumenetelmän tulosten perusteella käyttäen FineSTRUCTURE-ohjelmaa. Lopuksi populaatiojakoa verrattiin yksilöiden maantieteelliseen alkuperään.

Työn tulokset osoittivat, että kromosomin jaottelumenetelmä ryhmitteli selkeästi tiiviimmin seitsemän kymmenestä testiryhmästä kuin perinteinen PCA. Kuitenkin SmartPCA oli nopeampi ja helppokäyttöisempi kuin ChromoPainter. Geneettisissä populaatiotuloksissa erottui ensimmäisenä jako Itä- ja Länsi-Suomen välillä, mikä vastaa hyvin aiempien tutkimusten tuloksia. Kaiken kaikkiaan Suomesta löytyi yhteensä 15 populaatiota, joiden maantieteelliset alueet noudattelivat pääosin Suomen entisten läänien ja maakuntien rajoja.

Ensimmäisenä johtopäätöksenä todettiin, että kromosomin jaottelumenetelmä antaa tarkempia tuloksia kuin perinteinen PCA. Perinteinen PCA kuitenkin soveltuu edelleen alustaviin analyyseihin nopeutensa vuoksi. Toisena johtopäätöksenä todettiin, että kromosomin jaottelumenetelmä löysi ennennäkemättömiä hienorakenteita Suomesta ja nämä rakenteet ovat maantieteellisesti ryhmittyneitä. Tulokset luovat erinomaisen pohjan populaatiorakenteen ja geneettisten sairauksien jatkotutkimukselle Suomessa.

Avainsanat – Nyckelord – Keywords
populaatiorakenne, Suomi, pääkomponenttianalyysi, PCA, kromosomin jaottelumenetelmä, ChromoPainter, genominlaajuinen merkkigeeniaineisto, haplotyyppi, FineSTRUCTURE

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Matti Pirinen ja Samuli Ripatti

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin yliopisto, Viikin tiedekirjasto

Muita tietoja – Övriga uppgifter – Additional information

Contents

SUPPLEMENTARY MATERIAL

1 Introduction

Studies of population structure focus on revealing similarities and differences within and between groups of individuals and the processes that are causing these differences. The general motivation for human population studies arises from the need to understand what our roots are. Genetic information offers answers to this question and simultaneously complements the archaeological and historical records. Consequently, several commercial companies are already offering genetic ancestry profiles for individuals but the results typically discover individual's origin only at a continental level. These results may have a sufficient resolution to be interest for such heterogeneous populations as Americans but not for homogenous populations such as Finns. Nevertheless, more accurate methods have recently been developed (Leslie *et al.* 2015).

For scientific research, studies of population structure can provide crucial information for avoiding false interpretations. For example, in genetic studies of complex diseases, the case-control setting is frequently used. In this setting, it is important that the case and the control groups are matched in terms of all relevant variables, including the genetic background. Unknown population structure can result in a spurious genetic association with the trait studied and lead to false positive results.

Genetic markers, especially single nucleotide polymorphisms (SNPs), have been extensively used in studies of population history and structure. The most common method used to analyse these markers and the population structure is principal component analysis (PCA). PCA summarises complex data by creating a visual representation from which it is easy to make interpretations. One limitation of standard PCA is that, in order to obtain reliable results, only unlinked markers should be used. This reduces the amount of information that can be used. To circumvent the problem, Lawson et al. published the chromosome painting method (Lawson *et al.* 2012) that is based on haplotypes and thus can take into account the linked SNPs and linkage information. This new method has been reported to be able to provide finer structure information than standard PCA and distinguish even very young subpopulations (Lawson *et al.* 2012). As Finland has a recent settlement history, it seems likely that the chromosome painting method would be advantageous in the studies of population structure in Finland.

The key idea of the chromosome painting method is to detect common genomic segments between sampled individuals and, with the aid of the FineSTRUCTURE program, assign individuals into genetic populations. In addition to detecting the population structure, the chromosome painting method could also be used to date admixture events at the population and individual levels (Hellenthal *et al.* 2014). By applying these methods to Finnish data, we could give more precise answers to the origin of Finns and their relationships to the neighbouring populations. In turn, the ancestry information could also be used at an individual level. For instance, an application of the chromosome painting method to detect admixture at the individual level would allow a more detailed analysis of genetic ancestry profiles than is currently available. This could revolutionise the genealogy studies of both the academic and private sectors. Another future application of the chromosome painting method would be to compare distribution of the genetic risk scores for complex diseases between populations defined by the method and evaluate how much of the differences in disease risk could be explained by genetics.

In this Master’s thesis, I create a basis for the applications described above by evaluating the usability of the chromosome painting method in a Finnish population cohort. In particular, I answer the following two questions. First, whether the chromosome painting method provides tighter and more precise clustering of geographically defined groups of Finns than PCA based on independent markers. Second, whether the chromosome painting method can reveal new details about population structure in Finland.

2 Background

2.1. Population structure and variation

The term population is defined as a group of organisms that can reproduce with each other. Thus, it includes the fact that organisms need to be in the same region at the same time and are of the same species. The study of population structure provides information about the clustered differences between populations and subpopulations.

Variation between individuals can be observed at different levels, such as phenotypic, protein, chromosomal and genetic levels, which are introduced next. The level of genetic variation is introduced in the section 2.2.1. The phenotypic variation concerns discrete and continuous traits which are easy to detect even without an understanding of genetics. Consequently, phenotypic variation was studied already in the 19th century by the famous scientists Gregor Mendel (1822-1884) and Francis Galton (1822-1911) (Hartl & Clark cop. 2007). Mendel studied visually observable traits such as plant colour, especially in pea plants, and described the segregation laws of dominant and recessive traits. In turn, Galton studied continuous traits, e.g. eye colour and musical ability, and used statistical methods to describe the distributions of the traits in consecutive generations. These men created the basis for genetic studies and thus Mendel is said to be the father of genetics and Galton the founder of biometry (Hartl & Clark cop. 2007).

Protein, enzyme and chromosome level variation can also be studied without a specific knowledge of molecular genetics. For example, the ABO blood group system was the first polymorphism found in humans (found by Landsteiner in 1900) and it was used to study variation of human populations already in 1919 (Hirschfeld & Hirschfeld 1919). Later, more blood group polymorphisms have been found e.g. (Levine & Stetso 1939) and the detection of enzyme polymorphism using the technique of gel electrophoresis quickly increased the amount of information of variations (Smithies 1955). The information about protein and enzyme variation has been used widely in genetic studies, including the studies of population structure (e.g. Menozzi *et al.* 1978). The chromosomal level variations, such as triploidy and large deletions, are rare and typically lead to very severe abnormalities and thus they are not that useful for studies of population structure.

2.2 Human population genetics

The study of population genetics examines changes in allele frequencies of populations and the phenomena behind these changes. Human population genetics studies the genetic processes in species of *Homo sapiens* and its genus.

2.2.1 Genetic information and human genetic variation

The genetic information of humans (and other eukaryotes) is concentrated in the nucleus and the mitochondria of the cell. This information, coded in the base pairs of deoxyribonucleic acid (DNA), describes the biological code of an individual. DNA is constructed of two complementary strands of nucleotides. The strands are paired together with hydrogen bonds and are twisted into a double helix. The pairing happens between the bases of nucleotides, adenine, thymine, cytosine and guanine, in such a way that adenine pairs with thymine and cytosine pairs with guanine. In addition to a base, a nucleotide is composed of a deoxyribose backbone and a phosphate group which attaches nucleotides to each other.

99.9 % of genetic information is shared across all humans according to the Human Genome Project (Check 2005), and the remaining 0.1 % that separates individuals from each other is called variation. When the same variation is observed in several individuals within the same population, it is called a polymorphism. Genetic variation creates new forms of genes and genetic markers. These forms are called alleles and as each individual carries two copies of each chromosome, an individual can have 0, 1 or 2 copies of an allele. If an individual has 2 copies of an allele, i.e. both chromosomes have the same allele, the individual is said to be homozygote. If the copies of the same chromosome of an individual carry different alleles, the individual is said to be heterozygote.

Genetic variation can be divided into small structural changes in a DNA strand, and changes in base pairs. Since the first draft of the human genome was revealed (International Human Genome 2001), several international projects have been identifying DNA variants between human populations around the world (e.g. Sachidanandam *et al.* 2001, Altshuler 2010, McVean 2012). These projects have already identified over 88 million single nucleotide polymorphisms (SNPs) and indels for humans (NCBI bdSNP Build 142, 6.5.2015). As the name indicates, SNPs are (typically) biallelic single nucleotide changes in the individual's genome, and indels are insertions or deletions of one or more base pairs. SNPs and indels are the most used type of variation in the genetic studies as they are evenly distributed throughout the genome and easy to detect. Polymorphisms that occur in less than 1 % of individuals are typically called mutations or rare variants. SNPs are categorised into common and low frequency variants according to their frequency. A SNP is a low frequency variant if its minor allele frequency (MAF) is below 5 % (Altshuler 2010). Nucleotide polymorphisms occur fairly evenly throughout the human genome and cover 90 % of the sequence variation (Collins *et al.* 1998). Most of the SNPs are in regions that do not code a protein but they occur also in protein coding regions and regions that regulate gene expression. These possibly functional SNPs are especially interesting in genetic disease studies but the SNPs in noncoding regions give valuable information for the studies of populations and individuals (Collins *et al.* 1998). The markers used in this study are all SNPs.

Short tandem repeats (STRs) have most actively been used in individual identification and in forensics (Butler 2006). The identification of the individuals is based on the detection of the number of repetitions of the repeat unit of the STR. STRs are also called microsatellites and the length of repetitive unit is normally two to six base pairs long. Longer tandem repeats (10 to 60 bases) are called minisatellites. STRs have a high mutation rate and they are easy to detect with multiplex amplification and fluorescent methods, and these are the main reasons for their utility (Butler 2006).

There are also larger variations in the human genome and these are normally called structural variations. The main types of structural variation are inversions, translocations and copy number variations (CNVs). Inversions are changes where a long part of the genome has inverted while remaining at its position, whereas in translocations, a part of the genome has moved to another locus. CNVs are alterations in a number of repetitions of large genomic regions, such as whole genes, that are usually caused by duplications or deletions. Structural variations have not been commonly used in human population studies but, for example, CNV studies have become popular especially in the studies of complex diseases (Riggs *et al.* 2014).

2.2.2 Population genetic processes

The changes in the genetic composition of populations allow populations to differentiate and eventually lead to the differentiation of species. The main genetic processes that are responsible for these changes are mutation, natural selection, migration, non-random mating, genetic drift and recombination. These processes are introduced next. The features of genetic drift and recombination are emphasised as they are important in this study.

Mutation is the process that leads to a permanent change in genetic code. For example, SNPs, indels and CNVs are mutations when they first appear in a population. Errors in DNA replication and mutagens, such as ionizing radiation, tobacco smoke and free radicals, can cause damage to DNA and thus create new mutations. Mutation is the only process in genetics that creates new variation. The other processes only mix and change the genetic composition of populations. Usually mutation creates novel variation into a population and therefore increases the differences between populations. The estimated mutation rate for humans is 2.5×10^{-8} per nucleotide per generation (Nachman & Crowell 2000).

Natural selection is a force of evolution that changes the species and populations to better fit into their living conditions and environment. The simplified version of the natural selection is that those individuals that are better suited for their environment, have on average more offspring and spread more copies of their alleles to the next generation than other individuals. Thus, natural selection decreases the frequency of harmful variants (negative selection) and increases the beneficial ones (positive selection) by favouring individuals with beneficial traits in reproduction. How natural selection affects a population depends on the environment that the population lives in. For most parts of the genome, natural selection has a minor effect on population structure over a time period of a few generations. Genetic drift, which is introduced below, plays a more important role within the same time period.

In **Migration**, also known as gene flow, the genetic material is transferred between populations by fertile individuals. Migration changes the allele frequencies of both donor and recipient populations and can even introduce new alleles into the recipient population. Therefore, migration reduces the genetic differences between populations.

Non-random mating is the phenomenon where gametes are not fused randomly. Non-random mating occurs when, for example, the partner is chosen according to the ethnic background, physical appearance or cognitive ability. If individuals are favouring individuals genetically similar to themselves, the number of homozygotes is increased; if they favour individuals genetically different from themselves the number of heterozygotes is increased. If different characteristics are favoured in different populations, the situation leads to differentiation of the populations. **Inbreeding** is a form of non-random mating where the mating happens between relatives. Inbreeding is usually observed in small populations and populations with substructure. Inbreeding increases the number of homozygotes and differentiates populations that do not have migration between them.

Genetic drift is the random fluctuation of the allele frequencies over generations. If no other genetic process is involved, the change in frequency is caused by the random sampling of the alleles: the new set of alleles is chosen randomly from the alleles of the current generation and the probability that the allele is chosen is the frequency of the allele in the current generation. Thus, the expected frequency of the allele in the next generation is its frequency in the current generation. Nevertheless, as the process is random and the generation size is limited, there is always some variance for the allele frequency distribution of the next generation. The sampling process is typically modelled with the Wright-Fisher model (Hartl & Clark cop. 2007), closely related to the binomial distribution. The variance of the allele frequency under Wright-Fisher model is inversely proportional to the generation size. This means that the smaller the generation is, the bigger the variance is, and thus the genetic drift affects small generations more than the large ones. The process is analogous to the sampling of red and blue balls from a basket with known frequencies of colours. If you draw only ten balls from the basket, it is much harder to predict the relative frequency of the red and blue balls in the new sample based on the known frequencies, than if you draw one thousand balls.

The example in Figure 1 represents a constant sized population of seven circles throughout four generations. The example shows how genetic drift affects the allele frequencies and finally leads to the fixation of the blue allele. In general, genetic drift leads to the reduction of genetic variation, especially in small populations.

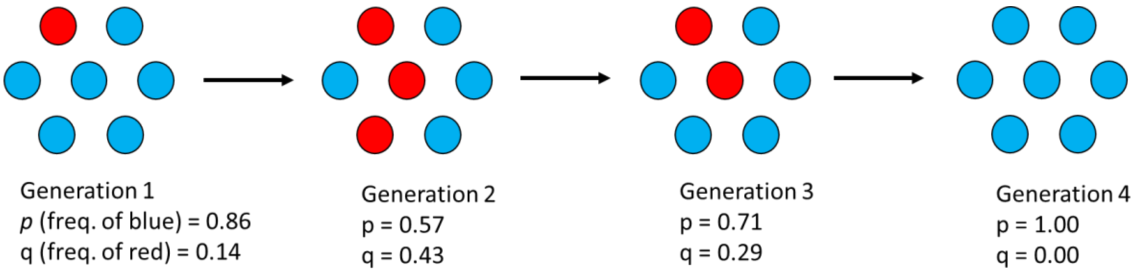


Figure 1 A schematic presentation of the effect of random genetic drift on allele frequencies in four consecutive generations. The colours of the circles represent the alleles and the allele frequencies are presented under each generation. In progressive generations, the red allele is lost and the blue allele becomes fixed due to genetic drift.

Additionally, as genetic drift treats populations of distinct size differently, it also treats common and low frequency variants differently: the frequency of a rare variant can be notably enriched or completely lost while the relative change in the frequency of a common variant remains smaller. Nevertheless, genetic drift treats the harmful and beneficial variants equally.

Sometimes the genetic drift can have a very strong effect on populations. For example, a flood or a fire can dramatically decrease the size of a population and the individuals that survive can be understood as randomly chosen. This kind of a drop in population size is called a bottleneck effect. Generally, the genetic variation is reduced in the population that survives the bottleneck effect but it is possible that some variants that were rare in the original population are enriched. Figure 2 shows how the frequency of red circles actually increases after the bottleneck. Because the size of a population is normally small right after the bottleneck effect, the genetic drift is also strong after the bottleneck. Therefore, the effect of a bottleneck is seen long after it occurred even in fairly big populations.

The second example of strong genetic drift is the founder effect. The founder effect shares the same features as the bottleneck effect but the birth mechanisms of the new population is different. In the founder effect, the old population is not destroyed but a small part of it migrates to a new area. The new area could be, for example, better hunting ground or a new island.

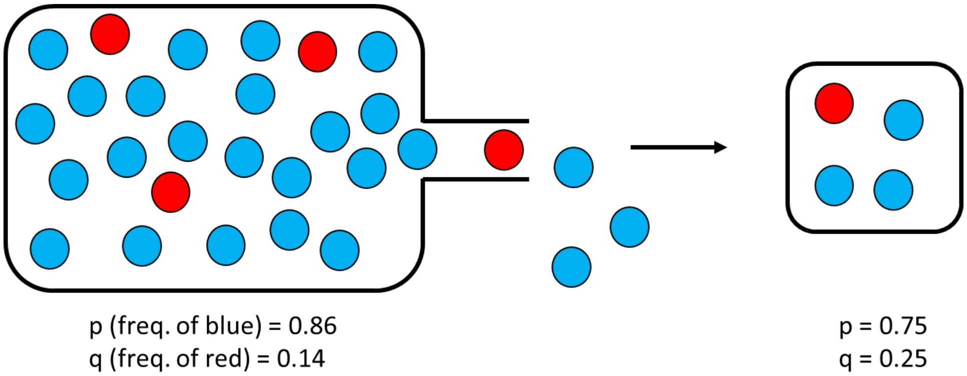


Figure 2 A schematic representation of the bottleneck effect. Because of an accidental event, four random circles from the original population (the left box) survive and establish a new population (the right box) and the allele frequency of the new population differs from the original. The colours represent the alleles and the allele frequencies are shown under the populations.

Recombination is the process of exchange of genetic material between chromosomes in an individual. The most important occurrence of recombination is the mixing of parental chromosomes in meiosis but it is also involved in DNA repairing processes such as in double strand breakage (Alberts *et al.* 2002). As the mixing of parental chromosomes by recombination is a fundamental idea in the main method used in this study, the chromosome painting method, only the recombination in meiosis is discussed here.

In meiosis, recombination takes place after the chromosomes have been duplicated and a chromosomal crossover occurs between parental chromosomes (Figure 3). Crossover exchanges parts of the parental chromosomes and produces chromosomes that are mixes of the original chromosomes. Thus, recombination allows offspring to inherit genetic information from all of the grandparents.

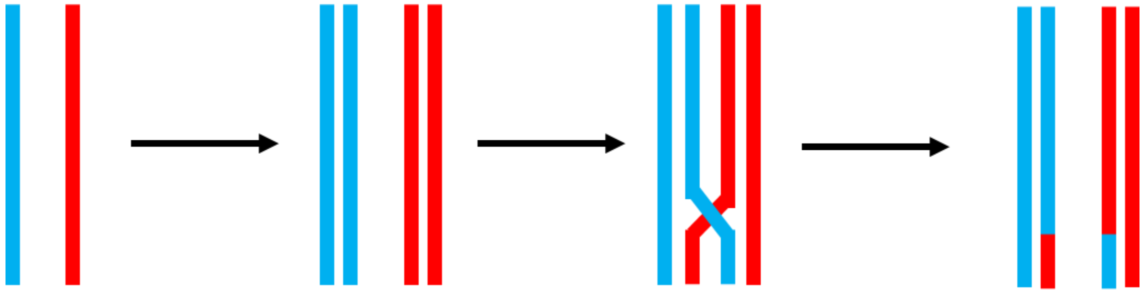


Figure 1 A schematic representation of chromosomal crossover in meiosis. The first step represents the parental chromosomes in a meiotic cell and the second step represents the same chromosomes duplicated. The third step shows a crossover event that exchanges genetic material between homologous chromosomes. The fourth step represents the outcome: two non-recombinant and two recombinant chromosomes are formed. These chromosomes are eventually divided into gametes and carry genetic information on to offspring.

Recombination breaks genetic linkage and thus haplotypes, i.e. segments of chromosome that have been inherited from the same ancestor. Genetic linkage is a phenomenon where specific alleles of two separate genomic positions are inherited together more often than would be expected by chance alone. Genetic linkage is affected by the uneven recombination rate throughout the genome: in some regions of the genome the recombination rate is very low and these regions are frequently inherited together. There are also recombination hotspots where the recombination rate is higher than the average rate. Naturally, recombination happens more often the wider apart the loci are. By studying the probability of recombination between two loci, it is possible to measure genetic distances between two loci in addition to the physical distance. The genetic distance between two loci is measured in morgans (M) according to Thomas Hunt Morgan (1866-1945) who discovered the variation in linkage patterns. This measure represents how often the recombination takes place between loci and thus the genetic distance increases with the amount of recombination.

In addition to the studies of genetic linkage at an individual level, it is also studied within populations. At the population level, the probability of two alleles to be seen together is measured with linkage disequilibrium (LD). LD attempts to measure how often alleles of different loci are inherited together in relation to no linkage in the population. As an example, let us examine two loci A and B with alleles A1, A2 and B1, B2. The frequencies of the alleles are p_{A1}, p_{A2}, p_{B1} and $p_{B2,}$ respectively. The alleles can form four haplotypes A1B1, A1B2, A2B1 and A2B2 whose frequencies in the population are p_{A1B1}, p_{A1B2}, p_{A2B1} and p_{A2B2}. If we observe a situation where, for example, allele B2 is seen only with allele A1 and there is no haplotype A2B2 then there is LD between the loci. This situation arises when there is only allele B1 in the population and the allele B2 appears due to a mutation into a genome with allele A1. The new haplotype A1B2 is passed on to future generations but recombination may also break the haplotype. If recombination happens between haplotypes A1B2 and A2B1, the new haplotype A2B2 arises and the LD is broken down. In the long run, recombination evens out the differences in allele and haplotype frequencies and linkage equilibrium is attained (in case there are no forces generating new LD).

More precisely, in LD, the loci are not independent, i.e. $p_{A1}p_{B1} \neq p_{A1B1}$, while in linkage equilibrium the loci are independent, i.e. $p_{A1}p_{B1} = p_{A1B1.}$ The strength of LD is measured by calculating the difference of the haplotype frequency and the product of allele frequencies and scaling the difference with the covariance between the loci (Pritchard 2001).

2.2.3 Why study population structure?

In general, populations are studied in order to obtain information about their size, growth, behaviour, consumption of resources and their effect on other populations and species. Genetic population studies are most well known in population history analyses but also important in medicine, individual identification, and environmental evaluations. For example in forensics, it is essential to evaluate the specificity of a marker set used in a certain population to minimize the possibility of false positive results (Butler 2006).

Studies of population structure examine the genetic differences in a group of individuals and the genetic processes that have lead the populations to be differentiated. In general, people are very interested in their ancestry and look for answers to the questions of their origin. The genetic studies of population structure and history started about forty years ago with the study of the blood group markers (Menozzi *et al.* 1978, Henn *et al.* 2010). Later, markers in Y chromosome and mitochondrial DNA (mtDNA) became common and this enabled the study of paternal and maternal lineages (Henn *et al.* 2010). These studies have shown that Y chromosomal variation is suitable for studies of distinct populations (Comas *et al.* 2000, Nasidze *et al.* 2003) and, based on studies of mtDNA, the geographical migration of women has been higher than in men (Seielstad *et al.* 1998). During the last decade, the genome-wide marker data, especially SNPs, has proven its potential in the studies of worldwide populations as well as in the detailed studies of small and young populations (Salmela *et al.* 2008, Jakkula *et al.* 2008, Jakobsson *et al.* 2008, Li *et al.* 2008).

The practical motivation for the studies of population structure comes from the case-control setting of association studies of complex diseases. In genetic case-control studies, such as genome-wide association studies, the aim is to detect a genetic locus that differs in cases and control. The locus and its surrounding region is then said to be associated with the disease or trait. Nevertheless, the problem may arise if the cases and controls are selected from the groups that differ also in other ways than the disease status. For example, if the incidence of cardiovascular disease is high in the Eastern parts of Finland, the cases are in a higher probability from Eastern Finland. In turn, the controls can include people evenly distributed from all over Finland. As it is known, people from Eastern Finland differ genetically from people in Western Finland. Thus, if this different genetic background is not taken into account, the association study can show that the alleles that are more frequent in Eastern Finland are associated with cardiovascular disease. In reality, people from Eastern Finland might not have genetic risk for this disease but a cultural habit of eating a lot of *trans* fats. Therefore, it is essential to understand the population structure of the study group.

2.3 Population history in Finland

2.3.1 Populating Finland

The story of the modern human (*Homo sapiens*) started from Africa approximately 150,000 years ago (Mellars 2004, Mellars 2006) and has reached almost all the corners of the world. Based on archaeological and genetic evidence, the human population dispersed from Africa only *ca.* 40,000 years ago – fairly recent, considering the origin of *Homo sapiens* (Mellars 2006). The reasons for the human dispersion from Africa have been debated, but the biggest reasons have probably been climatic and environmental changes, technological and social changes, and dramatic population growth. These reasons drove people first into Europe and Asia (Mellars 2006). According to archaeological evidence, Europe was populated from the Southeast, from the area of modern Turkey (Mellars 2004). It has been suggested that dispersal happened via two routes: one above and along the Danube and the other along the coast of Mediterranean sea. It is also worth mentioning that the spread of agriculture is assumed to have had similar a route 6,000 to 10,000 years before the present (Mellars 2004). The dispersion to the North happened as the ice from the last glacial period receded. The populations naturally moved North to hunt game and gain more living space.

The last glacial period ended and the region of modern day Finland was freed from the ice approximately 10,000 years ago. This allowed Finland to be populated and the first people arrived in the southern coastal regions of Finland already 9,000 years ago (Takala 2004, Pesonen 2005, Tallavaara 2010). Since the first arrivals, Finland has been constantly inhabited but the origin of the first Finns is still an unanswered question. Nevertheless, during the first 5,000 years free from ice, Finland was sparsely populated and the population size was only a few thousand individuals (Tallavaara 2010). The archaeological findings have shown that the first local populations have had several a pre-ceramic cultures (Huurre 2001), a Comb ceramic culture and a Corded ware culture and that the cultures were introduced into Finland mainly from southeast and Estonian (Carpelan 1999).

The population history of Finland is well known for its several migration and bottleneck events. Probably, the first big migration wave has arrived along with the comb ceramic culture 6,000 years ago (Tallavaara 2010, Oinonen 2014). Adoption of agriculture has been a long and complex process that is dated to 2,500 – 4,000 B.P. Agriculture first arrived to eastern and southern parts of Finland and has just relatively recently reached the more northern parts of Finland (Taavitsainen 1998, Tallavaara 2010). The modern understanding of the population history of Finland suggests that there have been several small migrations, such as the immigrants from Sweden during the Middle Ages (~1,000 B.P.) (Pitkänen 2007) rather than single major migrations ("Väestön kehitys esihistoriallisella ajalla", 20.5.2015).

The most significant internal migration event happened in the 16th century when people started to inhabit the eastern and northern parts of Finland in order to gain lower taxation. The king of Sweden, King Gustavus of Vasa, gave lower taxation to people who were willing to move to the wilderness, in order to enlarge his empire. The people from South Savo were the most eager to leave for new areas and the genetic influence of this can be seen even today. The few people that lived in the area blend in with the newcomers or draw back to north (Varilo 1999, Pitkänen 2007).

The population size in Finland has fluctuated during the centuries but has had an increasing trend. For example in 1697, the great famine reduced the size of the Finnish population to about 400,000 individuals (Norio *et al.* 1973). All of Finland was inhabited by the end of 17th century, even though the population remained scattered in small villages (Varilo 1999). By the end of the 19th century the population size of Finland was almost 2 million individuals (Norio *et al.* 1973, Varilo 1999). Today, the population of Finland is 5,5 million ("Suomen väkiluku", 26.2.2015).

2.3.2 Studies of genetic population history and structure in Finland

Because Finland has a small and relatively young population that has gone through several bottlenecks, it has been the focus of several genetic studies. Several migration waves and famines have created a population where some variants, rare elsewhere in the world, have enriched in Finland. This, in addition to a relatively homogenous population structure, has made Finland to be of interest in several medical and population studies. For example, individuals from Finland have been extensively sequenced in the 1000 Genomes project (1000 Genomes 24.11.2014) and Sequencing Initiative Suomi (Sequencing Initiative Suomi 24.11.2014).

The first relevant studies of population structure and genetic features in Finland were initiated in the 1970s by Nevanlinna (1972). Nevanlinna compared the gene frequencies of blood and serum group markers at county, community and village level. In these broad studies, he found clear differences between regions of Finland and built a basis of knowledge about differences between South-Western and North-Eastern Finland. At the same time, the term Finnish disease heritage was created (Norio *et al.* 1973). The term refers to a group of monogenic diseases that are more common in Finland than in other countries. The group involves 35 genetic diseases of which many are geographically clustered. The reasons behind the Finnish disease heritage and the population structure are the same: small population size, migration waves, and bottleneck and founder effects. Thus, it seems likely that there is a correlation between the incidence of these diseases and the subpopulation structure.

The current understanding of genetic features and population structure in Finland is summarised, among others, in the papers of Varilo (1999), Jakkula *et al.* (2008), Lappalainen (2009) and Salmela (2012). Varilo (1999) studied the ages of mutations in Finnish disease heritage and offered a comprehensive review of the population history of Finland at the time. Varilo has compared successfully the ages of mutations with the population historic events. For example, the oldest disease mutations, aspartylglucosaminuria, congenital nephrotic syndrome of Finnish type and infantile neuronal ceroid lipofuscinosis, are dated to 2,000 to 3,000 years ago. Additionally, he has studied the variation of linkage disequilibrium in Finland and its subisolate, Kuusamo, and showed that the presence of LD in Kuusamo is much stronger especially in the X chromosome than in general in Finland. These studies give a good understanding of the population history of Finland, the features of linkage disequilibrium as well as the ages of mutations in Finland. In turn, Jakkula *et al.*, Lappalainen and Salmela have concentrated more on the population structure of Finland using Y chromosomal, mtDNA and genome-wide marker data. The results of these studies and features of Finnish population structure are explained next.

Finns are genetically an outlier population in Europe (Salmela *et al.* 2008, McEvoy *et al.* 2009). According to both antigen frequency based study (Siren *et al.* 1996) and whole genome studies (Salmela *et al.* 2008, McEvoy *et al.* 2009), the allele frequencies and genetic distances between populations show that Finland differs from European and other North European populations more than could be expected based on the geographic location of Finland. Nevertheless, the closest related populations for Finns are Swedes, Estonians and Poles according to Lao *et al*. (2008) and McEvoy *et al.* (2009).

The details of genetic features of Finns have been studied by Y chromosomal, mtDNA and genome-wide markers. The studies of Y chromosomal haplotypes have shown that the diversity of the Y chromosome has decreased compared to other European populations (Hedman *et al.* 2004, Lappalainen *et al.* 2006). In fact, the diversity of Y chromosomal haplotypes is further reduced if the males of Western and Eastern Finland are compared with each other: the diversity in Eastern parts of Finland is smaller (Lappalainen *et al.* 2006). The regional differences of the Y chromosome are not only restricted to its diversity. The Y chromosomal haplotypes show distinct differences in their frequencies between East and West (Lappalainen *et al.* 2006, Lappalainen *et al.* 2008). For example, the Y chromosomal haplogroup N1c (N3) is much more common in Eastern Finland than in Western Finland (Lappalainen *et al.* 2008). In contrast, mtDNA studies have not shown as distinct differences as with the Y chromosome. The diversity of mtDNA haplogroups has been estimated to be similar to European populations and the distribution of mtDNA haplotypes is fairly homogenous (Hedman *et al.* 2007). Only few mtDNA haplogroups (e.g.

haplogroup Z) that resemble Eastern ancestry are found (Meinila *et al.* 2001).Thus, the mtDNA studies have not revealed as strong regional population structure as the Y chromosome studies. The haplotype studies of the Y chromosome and mtDNA reveal only a small part of the genetic features related to population structure. Most of the genetic information lies in the autosomes, and the genome-wide SNP data have become a popular tool for studies of population structure and genetic diseases during the last five years. First of all, the genome-wide SNP data have shown strong support for the existence of eastern and western subpopulations of Finland (Ikäheimo *et al.* 1996, Hannelius *et al.* 2008, Jakkula *et al.* 2008, Salmela *et al.* 2008). Additionally, the studies show how people in Northern Finland are distinguished from the rest of Finland and can even be assigned into subpopulations. For example, the genome-wide data have detected regional differences in the homozygosity and linkage disequilibrium patterns (Jakkula *et al.* 2008). Finally, the high-density marker data have proven their power to detect even regional differences (Jakkula *et al.* 2008, Salmela *et al.* 2008).

2.4 Principal Component Analysis

Principal component analysis (PCA) is a statistical method that summarises large and complex data. The method simplifies data by finding new linear uncorrelated variables that contain as much of the variability of the data as possible. These new variables are called principal components. PCA is mostly used to visualise data, inspect the heterogeneity and clustering of the data. Because of these properties, it is also useful for quality control. In this study, PCA is used to detect population structure from the data of unlinked genotype markers and to visualise the haplotype-based coancestry matrix. Next, the history of PCA in genetics and the basic idea behind the method are introduced. The technical implementation of PCA is discussed in section 4.3.

2.4.1 A brief history of PCA in genetics

PCA was introduced into genetics in the 1970s and it was used to study allele frequencies of just a few polymorphic protein alleles and marker loci (Harpending & Jenkins 1973, Menozzi *et al.* 1978). Harpending and Jenkins defined the method by distinguishing 18 southern African populations with 15 marker loci and comparing the plot of the first two principal components with linguistic and demographic differences (Harpending & Jenkins

1973). They found that the first component distinguished the studied populations into two linguistic groups and the second component correlated with non-African admixture. In turn, Menozzi and Cavalli-Sforza did PCA for only ten loci of 67 populations to study spreading mechanisms of Neolithic farming (Menozzi *et al.* 1978). They constructed "synthetic PCA maps" that corresponded geographically to genetic variation gradients. From these maps, Menozzi and Cavalli-Sforza concluded that the spread of farming was not only diffusion of technology but concrete migration events. PCA has ever since, especially in the 2000s, been used to study population structure and history (Chakraborty & Jin 1993, Stoneking *et al.* 1997, Capelli *et al.* 2006, Sikora *et al.* 2011, Wang *et al.* 2012). Nevertheless, there has been a debate on the interpretation of PCA results and their application. In 2008, Novembre and Stephens reported that some features of geographic PCA maps (Menozzi *et al.* 1978) can be caused by artefacts of the method itself and, thus, the historical interpretation of principal component plots and maps is not straightforward. The archaeological, linguistic and other evidence should always be interpreted simultaneously (Novembre & Stephens 2008). Nevertheless, it should be noted that these problems do not concern studies of the population structure, only the interpretation of historical migration events (Reich *et al.* 2008).

2.4.2 Methods of principal component analysis

As noted above, the main idea of PCA is to simplify complex data so that they are easier to visualise. The example data, shown in Figure 4, consist of 10 individuals from which two variables, x and y, are measured. These variables could represent, for example, weight and height. The data has two dimensions, corresponding to the number of variables. The aim is to reduce the dimensions from 2 to 1 so that as little information is lost as possible. The simplest way to reduce dimensions is to draw a line through the data and to project our individuals onto that line. We want to do this in a way that it retains the variability of the data and this is why we choose the line along which the points have the largest variance. In Figure 4 A, the red line denotes the line on which the points show the largest variance. In Figure 4 B, the individuals are shown in one dimension with the aid of the red line. This line is called the first principal component (PC1) and the differences between individuals can be interpreted using only PC1.

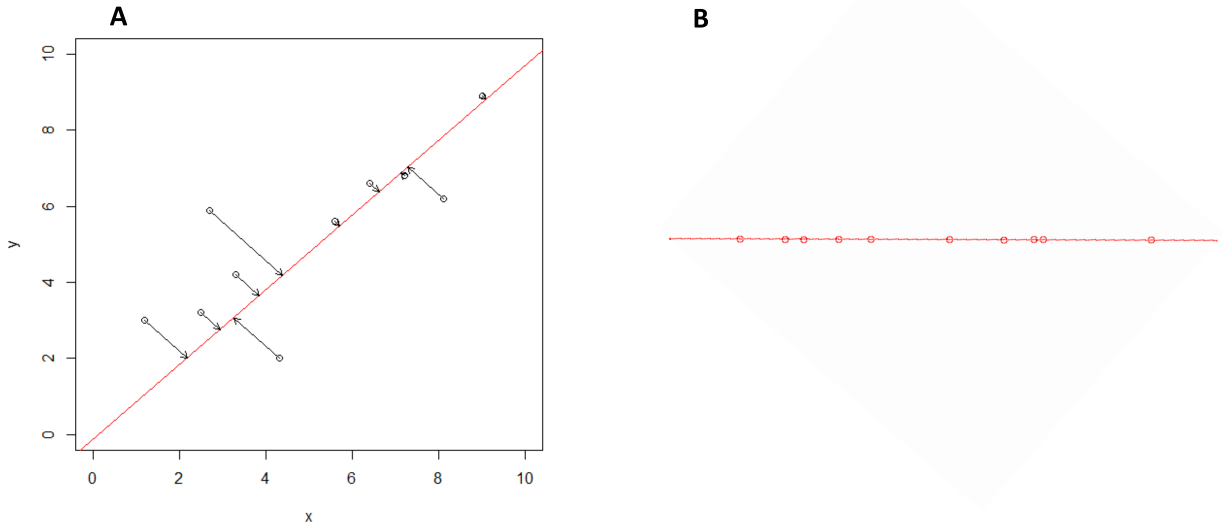


Figure 4 A simplified idea of principal component analysis (PCA). Ten samples are presented with two variables, x and y. PCA reduces the dimension from two to one by finding the line (red) on which the variance of the sample projections is the largest and presents the samples on that line.

The benefits of PCA are easier to understand with data that have several variables and thus several dimensions. For example, the genotype data can include thousands of individuals and hundreds of thousands of markers. Because it is hard to understand figures with 100,000 dimensions, we need a method like PCA. PCA can reduce the dimensions of genotype data in a way which is analogous to the example above. The only difference is that, in addition to PC1, the second, third and further principal components (PCs) can be found. Each new PC contains variance not contained in the previous PCs. Because the PCs are independent and orthogonal to each other, they can be plotted against each other in order to give a visual presentation of the data. In practice, the PCA is done with programs that use eigenvalue or singular value decompositions. The next example demonstrates the interpretation of the PCA results.

The example consists of 2031 individuals and their genotypes. To study the genetic relationship of the individuals, PCA was performed on the genotype data of about 60,000 SNPs and the result is visualised in Figure 5 A. Figure 5 A presents the individuals according to the two most variable principal components, i.e. PC1 and PC2. Figure 5 B represents the individuals on their geographic location of origin. If we first examine only the PCA figure and forget about the colouring, we notice that the individuals are not distributed evenly. The individuals have almost formed a triangle which has denser and

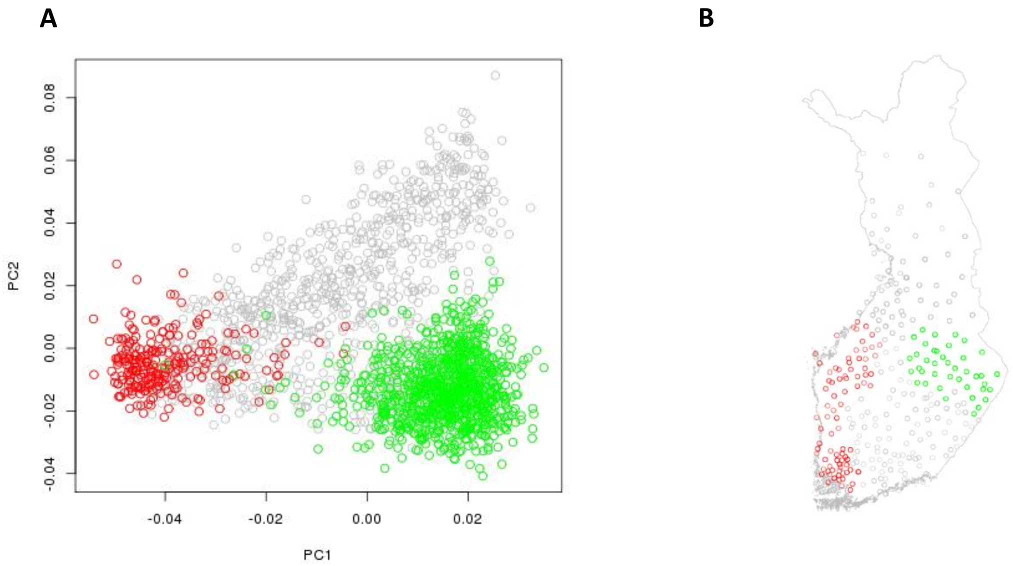


Figure 5 A) An example of PCA that shows population structure in the sample of 2031 individuals from Finland. The individuals are presented with the first two principal components. **B)** The same 2031 individual plotted on the map of Finland according to their birth places. Those individuals that originate from the same municipality are plotted on top of each other. Red colour indicates western individuals and green colour eastern individuals in both figures.

looser areas. Because the PCA analysed the genetic features, we can conclude, based on the PCA figure, that the data have genetic structure and the structure seems to have three main features: the left, right and top corners of the triangle.

Next, we can interpret the reasons behind the genetic structure by comparing the PCA plot and the geographic origin of the individuals (Figure 5). The same individuals have been coloured in both figures. The individuals from the western parts of Finland are coloured in red and the individuals from East are coloured in green. Comparing the figures, we notice that the individuals are distributed geographically in the PCA plot. In fact, PC1 distinguishes the individuals according to the East-West gradient, and PC2 according to the South-North gradient (not shown in the figures). This indicates even more strongly that the data include genetic structure. As seen in the above example, the genetic differences are caused by some external factor. The geographic isolation or distance is normally the strongest factor. Nevertheless, there are also other factors, such as language barriers and socio-economic factors, which can affect the mating behaviour and thus create genetic structures in a population. These smaller effects can be seen in further PCs that summarise the less variable dimensions and are valuable in detailed analysis of population structure.

2.5 Haplotype-based chromosome painting method

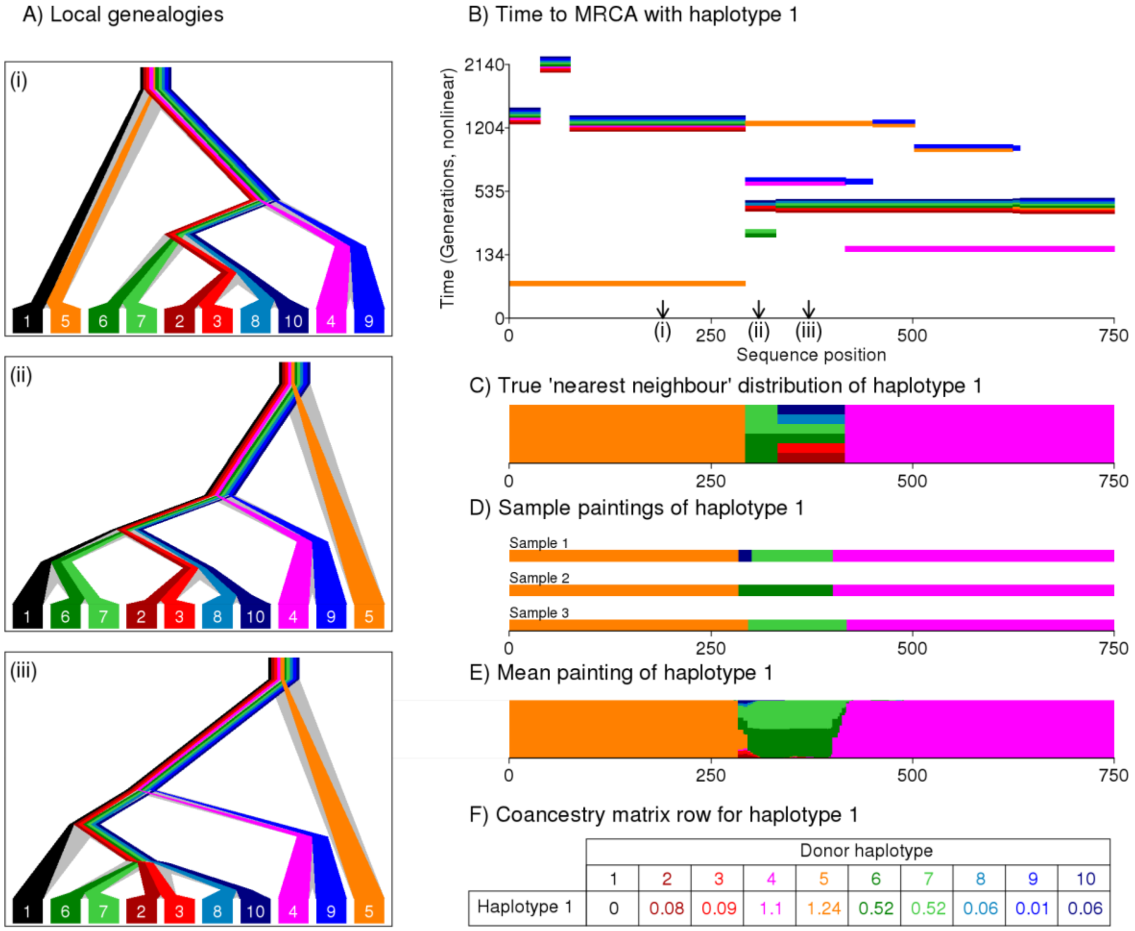
PCA is a useful way to study population structure but to construct a reliable result, only independent markers should be used. SNPs that are in LD with each other need to be excluded because otherwise PCA weighs these regions of the genome more than others (Anderson *et al.* 2010). The PCA plot of linked SNPs does not necessarily resemble the whole genome and can thus lead to false interpretations. The removal of SNPs in LD reduces the amount of information as the complete information in SNPs and LD patterns is not used. In 2012, Lawson *et al*. published a haplotype-based chromosome painting method that takes into account the LD information. The method that summarises the haplotype information is implemented in a program called ChromoPainter and the program that assigns samples into populations is called FineSTRUCTURE (Lawson *et al.* 2012).

2.5.1 ChromoPainter: Summarising genotype data

The aim of the chromosome painting method is to construct a compact representation of the relationships of the sampled individuals that can be further used for PCA or dividing individuals into populations. This compact representation is called the “coancestry matrix” in the ChromoPainter program (Lawson *et al.* 2012). The matrix is an individual by individual matrix whose elements represent the expected number of genomic chunks that the individual “inherits” from the other sampled individuals.

The method is based on the fact that the individual’s genome consists of chunks of its ancestors’ genomes. These chunks are mixed by recombination. The individuals that are closely related share common ancestors and, thus, similar chunks. Therefore, the chunks that take into account the linkage information carry detailed information about the relationships and genealogies between the individuals. The example in Figure 6 is the original illustration of the chromosome painting method by Lawson *et al.* 2012 for a sample of ten individuals. Figure 6 A shows the true genealogies that lie behind the three different loci in the genomes of the individuals. Each individual is examined one at a time by "painting" the haplotype of the individuals. The haplotype is "painted" with the other individuals' haplotypes, i.e. the haplotype chunks and their lengths that are shared between individuals, are estimated. In the following, the individual examined is called "recipient" and other individuals are called "donors". Figure 6 B shows the detailed history of the first individual's haplotype in terms of time to most recent common ancestor with the other individuals' haplotypes. The underlying genealogies would be possible to construct based on this information. Nevertheless, the goal of "chromosome painting" is not to detect the local genealogies but to construct a compact presentation of the data. Thus, the chunks of the closest individuals for individual 1 are gathered in Figure 6 C to see how much genetic material each individual donates for the examined haplotype. Figure 6 C is the "true painting" of the genome of the studied individual, i.e. "the true nearest neighbour distribution of haplotype 1" (Lawson *et al.* 2012).

To find the true painting and the genealogy for each haplotype is often impossible. This is why the chromosome painting method uses an approximation algorithm to perform the task computationally. ChromoPainter uses a Hidden Markov Model (HMM) approximation algorithm that was introduced by Li and Stephens (2003). The algorithm computes several sample paintings (Figure 6 D) and creates a mean painting (Figure 6 E) based on the sample paintings. Finally, the chromosome painting method constructs the coancestry matrix by calculating the number of chunks that are donated to the examined individual in the mean painting. In Figure 6 F, the examined individual's row is presented and it can be seen that the orange and magenta individuals, which cover most of the mean painting, have the largest values in the coancestry matrix. This suggests that the orange and magenta individuals are the closest related individuals to the examined individual in the sample.


A) Local genealogies
(i)
1
5
6
7
2
3
8
10
4
9
(ii)
1
6
7
2
3
8
10
4
9
5
(iii)
1
6
7
2
3
8
10
4
9
5
B) Time to MRCA with haplotype 1
Time (Generations, nonlinear)
2140
1204
535
134
0
0
250
500
750
(i)
(ii)
(iii)
Sequence position
C) True 'nearest neighbour' distribution of haplotype 1
0
250
500
750
D) Sample paintings of haplotype 1
Sample 1
Sample 2
Sample 3
0
250
500
750
E) Mean painting of haplotype 1
0
250
500
750
F) Coancestry matrix row for haplotype 1

	Donor haplotype									
	1	2	3	4	5	6	7	8	9	10
Haplotype 1	0	0.08	0.09	1.1	1.24	0.52	0.52	0.06	0.01	0.06

Figure 6 The original schematic presentation of the chromosome painting method by Lawson *et al.* 2012. **A)** The method is based on the assumption that the genomic data contain information about the site specific local genealogies i-iii. **B)** These genealogies can be examined with the graph which shows the time to the most recent common ancestor (MRCA) between individual 1 and other individuals as a function of the individual's genome. **C)** From this, the true distribution for the closest relatives of the 'nearest neighbours' at each site can be constructed. **D)** To estimate the 'true painting', an approximation algorithm is used to generate sample paintings. **E)** The sample paintings are combined to create the mean painting. **F)** From the mean painting the coancestry matrix of the chunk counts is calculated.

2.5.2 FineSTRUCTURE: Clustering individuals into the populations

Lawson *et al.* (2012) defined a clustering model for assigning individuals into the populations based on the coancestry matrix from their chromosome painting method. The implementation of this method is called FineSTRUCTURE (Lawson *et al.* 2012). FineSTRUCTURE is a model-based method that infers the number of populations and assigns the individuals into them. The algorithm is closely related to that of the STRUCTURE program (Pritchard *et al.* 2000). FineSTRUCTURE does not assume admixture of populations but can still be used for admixed data (Lawson *et al.* 2012). The clustering of individuals into populations is accomplished with a Bayesian approach implemented via the Markov Chain Monte Carlo (MCMC) algorithm (Lawson *et al.* 2012).

The basic idea of algorithm is to use an iterative process for estimating the population assignment. Let us consider very simplified example of the process where we want to estimate only the number of populations, k, and forget about the population assignment. The algorithm starts with a random value (for example $k_0 = 2$) and samples a proposed value based on the starting value and a symmetric proposal distribution. The superiority of the new value (for example $k_{prop} = 3$) is then evaluated against the previous value using a quantity that depends on the likelihood ratio. If the proposed value is accepted, it is assigned to be the new starting value for the next iteration ($k_1 = k_{prop}$). If it is rejected, the next value is the same as the previous value ($k_1 = k_0$). The estimate can be calculated as a mean of the sequence of values $k_0, \ldots, k_n$, obtained by running the algorithm over hundreds of thousands of iterations. The basic idea behind the estimation of the population assignment is similar but the sampling of the new population assignment and its evaluation is more complicated.

After the population assignment, the relationships of the populations are defined by constructing a hierarchical clustering tree. The tree shows the relationships between the populations but does not infer the times of divergence. Thus, it cannot be called a phylogenetic tree. The arrangement of the tree is found by merging the two populations that give the highest probability for the merged group. A more detailed description of the algorithms can be found in section 4 and from Lawson *et al.* (2012).

3 Aims of the study

This study has two aims. First, the standard principal component analysis based on independent markers is compared with the haplotype-based chromosome painting method using data from the FINRISK Study survey of 1997. The first research question is whether the chromosome painting method can provide tighter and more precise clustering than PCA for the geographically defined groups. Second, the Finnish population structure is studied with the FineSTRUCTURE program. The aim is to find out whether the chromosome painting method can reveal new details about population structure in Finland. The answers to these questions form a basis for future studies of population structure and disease genetic in Finland. In Figure 7, the aims of this study are presented together with the workflow.

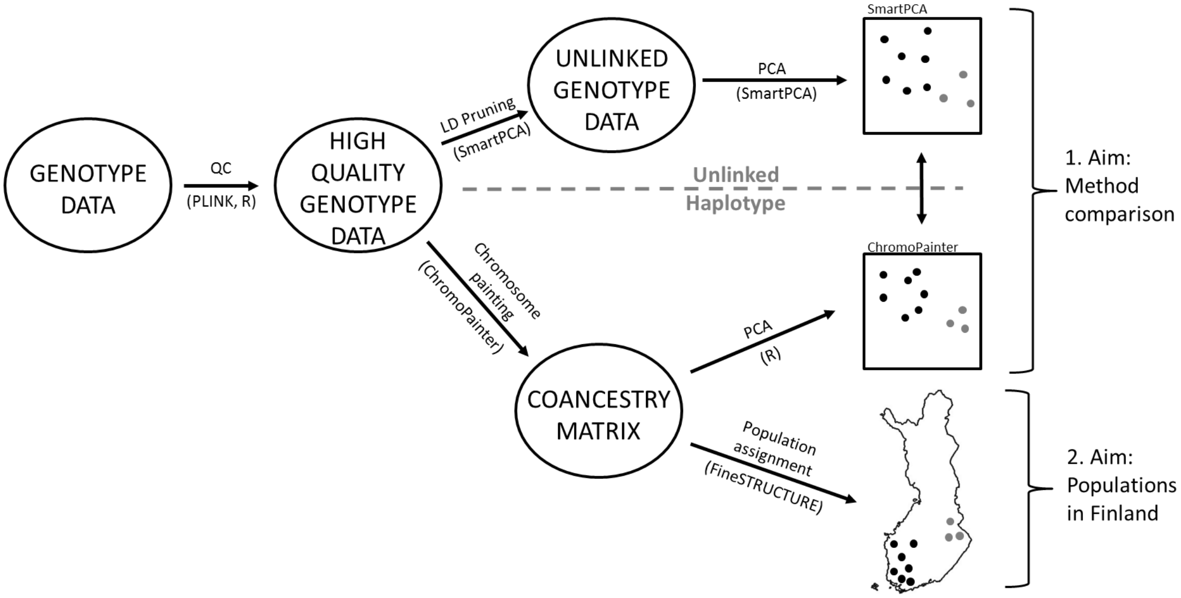


Figure 7 The aims and the work flow of this study. The analyses made are denoted on top of the arrows and the programs used below the arrows. The grey dashed line represents the division between the analyses based on unlinked and haplotype data.

4 Materials and Methods

In this study, I performed analyses of population structure on Finnish genotype data. The analyses included a general quality control, principal component analysis of independent markers with the SmartPCA program, the chromosome painting analysis with ChromoPainter and population assignment with the FineSTRUCTURE program. Below, I refer to the methods and analyses with the name of program, i.e. SmartPCA, ChromoPainter and FineSTRUCTURE. Table 2 (on page 39) summarizes the programs used and their role in this work.

4.1 Data

The data used in this study were obtained from the FINRISK Study survey of 1997 population cohort (Vartiainen *et al.* 1998). The FINRISK Study is a series of national surveys that focus on risk factors of chronic diseases in Finland, especially cardiovascular diseases, and have been conducted every five years since 1972 (Borodulin *et al.* 2013). Permission to use the data was granted by the National Institute for Health and Welfare and the FINRISK Management Group (Permission 2014_55, June 2014). The data included the subset of genotyped individuals and the birthplaces of these individuals and their parents. All the samples were genotyped with Illumina HumanCoreExome-12 BeadChip (547 K SNPs) at the Wellcome Trust Sanger Institute, Cambridge, United Kingdom. The genotypes were determined for each individual by the genotype calling algorithm, zCall (Goldstein *et al.* 2012), at the Institute for Molecular Medicine Finland (FIMM). The positions of the markers were given according to the human reference genome build 37. I performed all my analyses using the FIMM MARS Server.

The genotype data were available to me in PLINK format (“PED files”, 14.1.2015) that consists of two files, genotype (.ped) and marker (.map) files. A genotype file includes information about the individuals and their genotypes (example in Figure 8). Each individual is on its own row and the first two columns define individual identity number (ID) and family ID. The following columns contain paternal ID, maternal ID, sex and possible phenotype status, respectively. The remaining columns contain the haploid genotype for each SNP marker. SNPs and their order are defined in a marker file that lists the chromosome, SNP ID, genetic distance (in morgans), and position (in base pairs). Thus, the size of the original genotype data was 4,191 rows times (6 + 2 × 547,000) = 1,094,006 columns, corresponding to 528 megabytes.

```
 A                                                B
0001   0001   0   0   1   -9   G   A   T   …   G     1    rs000001   0.6    750000

0002   0002   0   0   1   -9   A   A   T       A     1    rs000002   1.2    990000

…                                                    …

4000   4000   0   0   1   -9   A   A   C   …   G     22   rs500000   75.4   5000000
```

Figure 8 A) An example of .ped file format. Rows correspond to the individuals and the columns correspond to individual ID, family ID, paternal ID, maternal ID, sex, phenotype status and genotypes (two for each marker). **B)** An example of a .map file. The order of genotypes in .ped file is defined in this file. Columns contain chromosome, SNP ID, genetic distance (in morgans) and position (in base pairs).

4.2 Quality control

The genotypes of individuals are determined by gene chips that measure the intensities of annealed probes (Peterson 2013). The sample DNA is multiplied, denatured into single stranded molecules, cut into small fragments, tagged with fluorescent dye and finally, annealed into a gene chip that contains probes for each allele of the studied SNPs. Those sample fragments that are not annealed into the chip are washed away. The remaining fragments are attached to the probes of particular alleles in particular locations in the chip. The genotypes can then be interpreted based on the fluorescent signal and its location. The probe intensities are converted into genotypes by genotype calling algorithms. The genotyping and genotype calling can include errors and systematic bias. These problems can be reduced with careful quality control (QC) (Anderson *et al.* 2010). QC is typically performed separately at the marker level and at the individual level. Rare markers are usually more prone to genotyping errors (Anderson *et al.* 2010) and therefore the variants that have low minor allele frequency (MAF) are preferably excluded. A strong deviation from Hardy-Weinberg equilibrium (HWE) can indicate genotyping or genotype calling errors and thus is screened for as a step in QC. Some markers may be difficult to genotype and therefore contain more errors. The genotyping success rate is a good measure to detect poorly genotyped markers and the exclusion of markers of low success ensures the homogeneous quality of the data. In addition, it is also good to check the genotyping success rate of individuals and whether some individuals are much more heterozygous than average. A strong deviation from the mean heterozygosity can indicate contamination of the sample.

4.2.1 Quality control on SNPs for SmartPCA

For SmartPCA, I performed QC on the SNPs in the genotype files. At the beginning, there were 528,255 SNPs. First, I extracted the autosomal SNPs because Y and X chromosomes have different population dynamics and their population structure is not considered here. The extraction was performed by generating the list of included SNPs using R (R Core Team 2014) and extracting the SNPs with PLINK version 1.07 (Purcell 2007, Purcell 2009). Next, I filtered the SNPs by MAF, HWE p-value and genotyping success rate. SNPs whose MAF was under 5 %, HWE p-value under 10^{-6} and the success rate under 99 % were excluded from the analysis using PLINK. These thresholds were even stricter than the commonly used ones (Jakkula *et al.* 2008, Leslie *et al*. 2015) which ensured a high quality data. These filtering thresholds left a total of 251,998 SNPs in the data.

To ensure the independence of the SNPs, I calculated the linkage disequilibrium as the square of the pairwise correlation coefficient (r^2) for the SNPs using SmartPCA program version 8000 (Patterson *et al.* 2006). I used a 1 centimorgan window for the calculation of r^2 values for each SNP and removed the SNPs in such a way that for the remaining SNPs the pairwise r^2 values were under 0.2. In addition, the European population includes 24 fairly large genomic regions (> 2 Mb) that are in strong linkage disequilibrium (Price *et al.* 2008). I removed these long-range LD SNPs described in (Price *et al.* 2008). To ensure that PCA treats each SNP equally, I performed a preliminary PCA with SmartPCA program and plotted the SNP weights of the first ten principal components (Figure 9). SNP weights tell how much each SNP is contributing to the principal component (see section 4.3 for formal definition). We want the distribution of the weights to be roughly uniform across the genome so that the results are not driven by certain genomic regions. The plots showed no strong differences between genomic regions and verified the successful exclusion of linked

SNPs. Finally, I examined the presence of the SNPs in the data for ChromoPainter and included only the markers that were also in the data set for ChromoPainter (see next section). This step was performed to ensure reliable comparison between SmartPCA and ChromoPainter analyses. The resulting number of SNPs for the SmartPCA was 60,251. A summary of the quality control steps and their effects on the number of SNPs is presented in Table 1.

Table 1. Quality control steps for the marker data of SmartPCA.

QC step name	**Excluding options**	**SNPs removed**	**SNPs left**
At the beginning			528,255
Extraction of autosomes	X, Y	7,693	520,562
SNP filtering	MAF < 0.05 HWE < 10^{-6} Success < 0.99	268,564	251,998
LD pruning	r^2 > 0.2 in 1 cM region	188,723	63,285
Removing the long-range LD regions	Price *et al.* 2008	1,016	62,269
Extracting SNPs that were present in ChromoPainter data		2,018	60,251

4.2.2 Quality control on SNPs for ChromoPainter

The QC on SNPs for the ChromoPainter analyses included the same steps as for SmartPCA, except for the LD pruning. The SNPs that had a MAF score under 5 %, HWE under 10^{-6} and genotype success rate under 99.9 % were removed from the data set. Note that the success rate was even stricter than in the data set used for SmartPCA. The resulting data set for ChromoPainter included 238,438 SNPs.

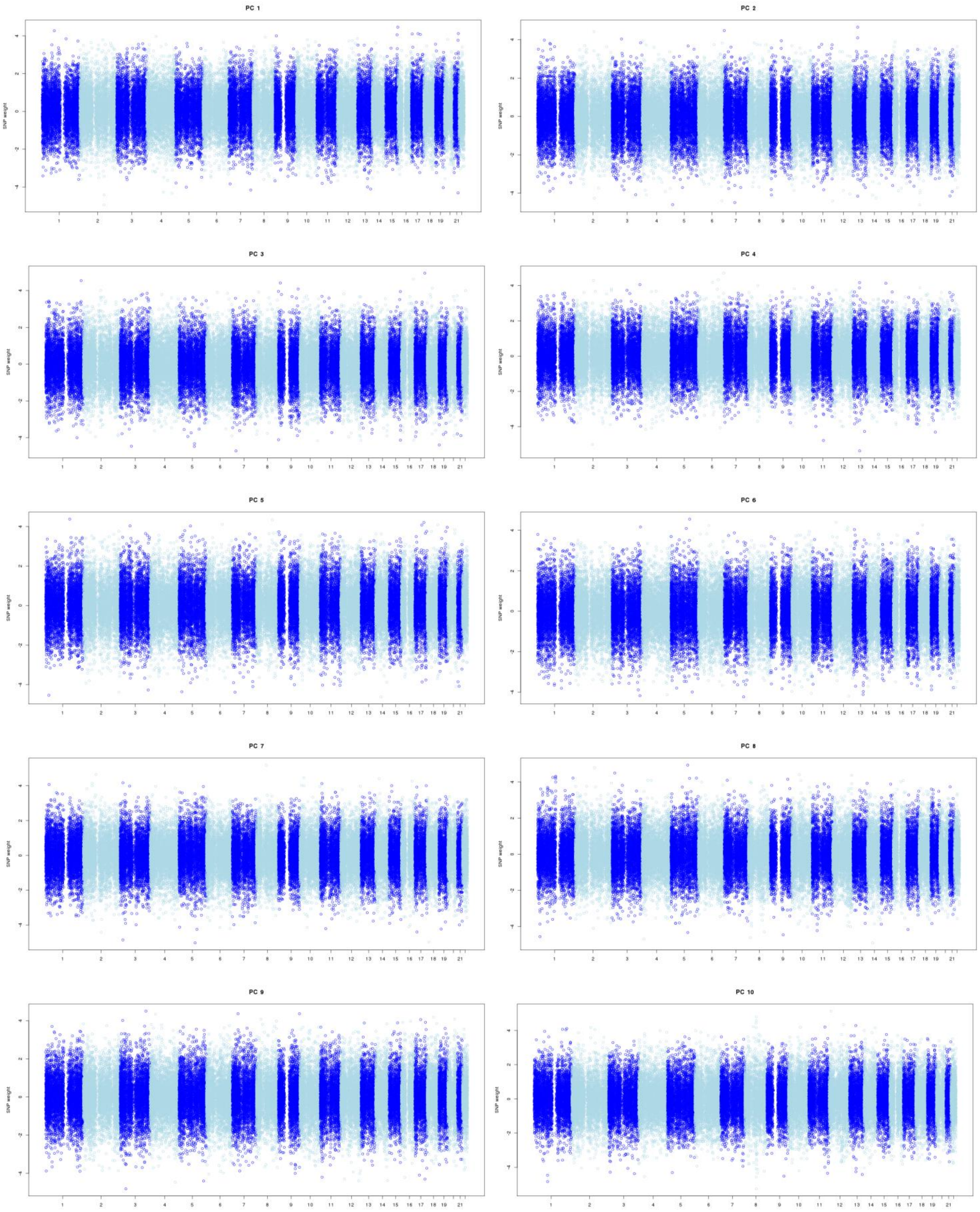


Figure 9 SNP weights of the first ten principal components in SmartPCA. The x-axis represents SNP position and y-axis represents the SNP weight. There are no large deviations from the average weight confirming that PCA treats every part of the genome evenly.

4.2.3 Quality control on individuals

The aim of the QC on individuals was to define high quality set of individuals for SmartPCA, ChromoPainter and FineSTRUCTURE analyses. After quality control of SNPs, I calculated the average heterozygosity and SNP success rates for the individuals using PLINK. I plotted the individuals with respect to these values (Figure 10). I removed the nine individuals (shown in red) as they clearly differed from the rest. To ensure that the individuals are not closely related, I calculated the relatedness coefficients between all individuals using two different methods. The identity by descent values were calculated using PLINK and kinship values using the GCTA program (Yang *et al.* 2011). I removed one individual from each pair of individuals that had one or both of the relatedness values over 0.05. After these steps, there were 3,606 individuals remaining.

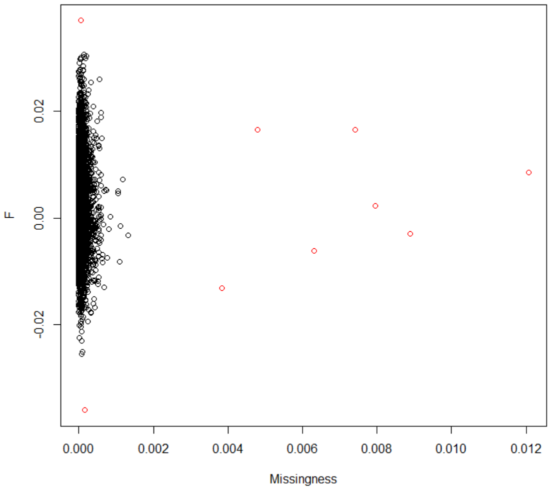


Figure 10 Individuals plotted by SNP missingess (success) rate and heterozygosity value (F). The individuals marked in red were excluded from the analysis.

I defined the final data set by extracting only those individuals whose parents were born in the same geographic region. This decreased the sample size to 2,317 individuals. Then, to rule out the possible effects of population size, I extracted 35 individuals from each of the 10 geographic regions. The regions correspond to the provinces of Finland in 1996 and are as follows: Uusimaa (USM), Province of Turku and Pori (TUP), Province of Häme (HAM), Province of Kyme (KYM), Province of Mikkeli (MIK), Province of Northern Karelia (NOK), Province of Kuopio (KUP), Province of Vaasa (VAS), Province of Oulu (OUL), and Lapland (LAP). The abbreviations are used to refer to the geographic region of the provinces henceforth. The locations of the provinces are presented in Figure 11. It was assumed that the population substructures are seen on the level of provinces as they align with the Finnish subcultural and dialect borders. The provinces in this study cover almost all of contemporary mainland Finland except for the province of Central Finland. The sample size of this province was too small (<35) to be included here.

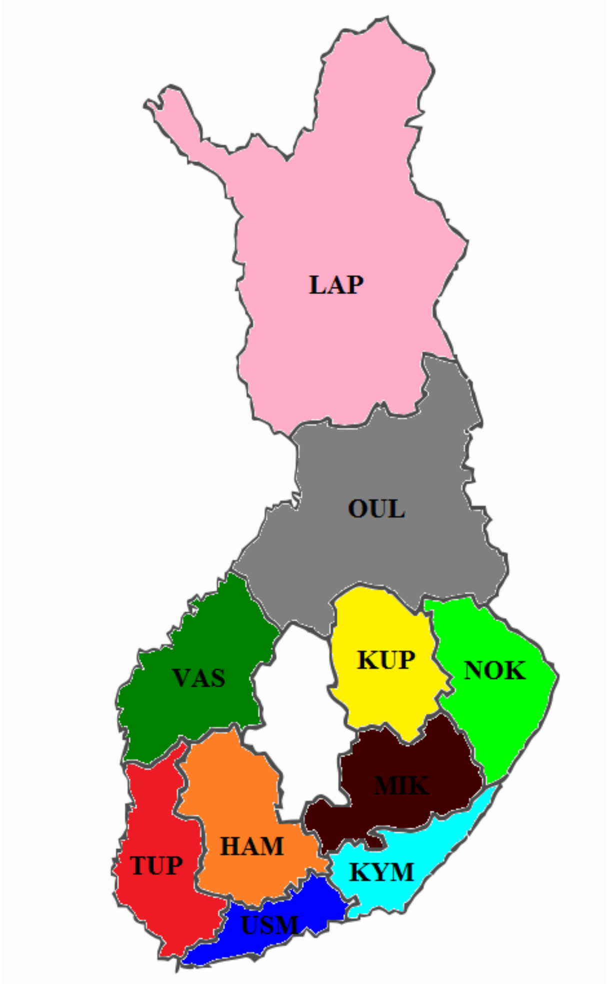


Figure 11 The geographic locations of the provinces: Uusimaa (USM), Province of Turku and Pori (TUP), Province of Häme (HAM), Province of Kyme (KYM), Province of Mikkeli (MIK), Province of Northern Karelia (NOK), Province of Kuopio (KUP), Province of Vaasa (VAS), Province of Oulu (OUL), Lapland (LAP). The province of Central Finland (white) was left out from the study as the number of individuals was too small.

I performed a preliminary analysis with ChromoPainter and it revealed that five individuals stood out from the rest in principal component 6 (Supplementary Figure S1). Since these five individuals did not stand out in preliminary analyses of SmartPCA, I decided to take a closer look at them. First, I examined the relationship of these outlier individuals, but they did not show differences from the rest of the pairwise relatedness values. Next, I examined how the outlier individuals behave in the analysis with more SNPs. It turned out that the individuals stood out even more. Finally, a closer look at the QC files of the genotyping process revealed that the outlier individuals originated from two genotyping plates that had more heterozygosity failures than usual. Therefore, I decided to leave out the outlier individuals to avoid possible contamination. The five individuals were from Uusimaa, Lapland and the province of Vaasa, and two from the province of Kyme. The final sample set consisted of 345 individuals and it was used in SmartPCA, ChromoPainter and FineSTRUCTURE analyses.

4.3 Principal component analysis of independent SNPs

The intuitive idea of PCA was described in the section 2.3.2 and the technical implementation is discussed next. PCA is typically performed on an $n \times m$ matrix, **M**, where n is the number of samples and m is the number of variables. In genotype data, the samples are individuals and the variables are markers, such as in Figure 8 A. The aim of PCA is to find the principal components (PC) of individuals. These PCs are defined as,

$$PC_{ji} = \sum_{k=1}^{m} a_k^j x_{ik}$$

for the *j*th PC, where *i* is the individual, *a* is weight factor of the marker and *x* is *k*th marker of *i*th individual. Thus, the PCs are linear combinations of the markers. The aim of PCA is to find the factors for *j*th PC a_1^j... a_m^j, also known as SNP weights, so that the variance $Var(\sum_m a_m^j x_{im},)$ is maximized with the constraint that this linear combination is orthogonal to the previous PCs (*1,…,j-1*).

In practice, PCA can be performed by using eigen decomposition, i.e. by finding the eigenvalues and eigenvectors for the covariance matrix of **M**. For example Shlens (2014) has shown why this eigen decomposition finds the PCs. This proof is based on the $m \times m$

covariance matrix $\mathbf{C_M} = \frac{1}{n}\boldsymbol{M^T M}$ and on the theorem that states that $\mathbf{C_M}$ can be diagonalised by an orthogonal matrix of its eigenvectors. Because the diagonal of a covariance matrix includes the variance of the variables, diagonalization maximises the variances and minimises covariances which was the ultimate aim of PCA. Nevertheless, genotype data typically include thousands of individuals and hundreds of thousands or even millions of markers. Thus, the approach of $m \times m$ matrix is often impossible. Fortunately, it is possible to find the PCs using eigen decomposition of an $n \times n$ covariance matrix which reduces the amount of computation (Price *et al* 2006).

In this study, I performed the principal component analysis with SmartPCA version 8000 of the Eigensoft package (Patterson *et al.* 2006). SmartPCA performs PCA in three steps. First, it normalises the data to give each SNP equal variance, independent of allele frequency. The normalisation of each element of the genotype matrix is calculated as,

$$\mathrm{M(i,j)} = \frac{\mathrm{C(i,j)} - \mathrm{\mu(j)}}{\sqrt{\mathrm{p(j)(1-p(j))}}},$$

where M(i,j) is the normalised value for the element, C(i,j) is the number of reference alleles for individual *i* in marker *j* (0 or 2 for homozygotes and 1 for heterozygotes), μ(*j*) is the mean number of reference alleles in marker j, p(j) is the estimated allele frequency (Patterson *et al.* 2006). Second, SmartPCA calculates the $n \times n$ covariance matrix $\mathbf{X} = \frac{\mathbf{1}}{\boldsymbol{m}}\boldsymbol{MM^T}$ and third, it computes the eigen decomposition of the covariance matrix. After performing SmartPCA, I calculated the variance contained in each principal component by dividing the particular eigenvalue by the sum of all eigenvalues (Chang 2013). I plotted the principal components by using R (R Core Team 2014).

4.4 Chromosome painting

ChromoPainter aims to capture the individuals' relationships modified by recombination and genealogical processes and it is based on the Hidden Markov Model of Li and Stephens (2003). The algorithm proceeds by considering one individual at a time. The intuition of the algorithm is described at section 2.5. Briefly, the studied individual can be seen as a recipient and the others as donors of genetic material. The donated genetic material is obtained from the common ancestors and the terms “recipient” and “donor” are used as a metaphor.

In the model, the donor sequence of recipient haplotype (below denoted by Y) is first studied by calculating a probability for other individuals serving as a donor for each marker site. The probability distribution for a site is based on the distribution of the previous site, recombination probability (1) and mutation probability (2), defined below. The transition probability for Y, i.e. the probability for donor haplotype transition between sites *l* and *l+1* including recombination is,

$$\Pr(Y_{l+1} = y_{l+1} | Y_l = y_l) = \begin{cases} \exp(-\rho_l) + (1-\exp(-\rho_l)) f_{yl+1} & \text{if } y_{l+1} = y_l \\ (1-\exp(-\rho_l)) f_{yl+1} & \text{otherwise} \end{cases}, \qquad (1)$$

where y_l is the existing donor haplotype state at site *l* and y_{l+1} is the existing donor haplotype state at site *l+1*, $f_{yl+1}$is the copying probability of copying from donor haplotype y_{l+1} and ρ_l is the population-scaled genetic distance: $\rho_l = N_e g_l$, where N_e is a scaling parameter based on the effective population size and g_l is the genetic distance between sites *l* and *l+1*. In other worlds, the upper part of the equation (1) defines the transition probability when the current donor haplotype does not change. This probability is the sum of the probability that recombination does not happen and the probability that recombination happens but between the donor haplotype itself. The lower part of the equation defines the probability that recombination happens between two different donor haplotypes. The probability for observing an allele given the donor haplotype at site *l* is,

$$\Pr(h_{*l} = a | Y_l = y) = \begin{cases} 1.0 - \theta & h_{yl} = a \\ \theta & h_{yl} \neq a \end{cases}, \qquad (2)$$

where h_{*l} is the observed allele of haplotype * and θ is a mutation parameter. This means simply that the probability is the mutation probability when the sites do not match and the probability of no-mutation when the sites match. The whole haplotype is first examined site by site using above equations from left to right and then the probability distributions are completed by updating from right to left. This is so called forward-backward method of

Hidden Markov models (Lawson *et al.* 2012). The number of copied chunks is then determined by detecting the most probable sites for the chunk end based on the site specific probability distributions. The detected numbers of chunks are gathered into a "coancestry matrix" which is the output of ChromoPainter.

I performed the haplotype-based chromosome painting analyses with ChromoPainter 0.0.4 and ChromoCombine 0.0.4 programs (Lawson *et al.* 2012). Because ChromoPainter handles linked information and data as haplotypes, the genotypes have to be assigned into parental chromosomes. This construction of original haplotypes is called phasing. The phasing of genotype data was performed simultaneously for the whole data set by using SHAPEIT version 2 (Delaneau *et al.* 2013), by Antti-Pekka Sarin. Then, I converted the phased data into the ChromoPainter format and created the recombination files that contained the information about the recombination rate per base pair for SNPs based on HapMap phase II build 37 recombination maps ("HapMap phase II", 24.7.2014).

In addition to phased genotype data and recombination maps, ChromoPainter also needs estimates for N_e and θ parameters. Here, I used Watterson's default estimate (Watterson 1975) for the global mutation rate θ, and the population size-based scaling parameter N_e was estimated using ChromoPainter's expectation-maximisation algorithm. The estimation was performed on every tenth individual and on each chromosome using ten iterations. Then, I calculated the average value from the individual results and used it in the final analysis. File format conversions and calculations of parameter averages were performed using Perl scripts found from the program homepage's (Lawson, 24.10.2014). Finally, I performed the linked chromosome painting analysis separately for each chromosome and combined the results with ChromoCombine.

I modified the coancestry matrix obtained from ChromoCombine to ensure that the principal components obtained from the results of ChromoPainter are comparable with those of SmartPCA. The modifications included the addition of the column sums to the diagonal, subtraction of the column means from the elements of the coancestry matrix and symmetrising it by multiplying it with its transpose (Lawson *et al.* 2012). Then, I used the modified coancestry matrix to create principal component analysis to compare the results from ChromoPainter with the results of SmartPCA.

4.5 Quantitative method comparison

In addition to a visual comparison of PCA plots of SmartPCA and ChromoPainter, I compared them quantitatively. I studied the tightness of the clusters of individuals from the same geographic region in the plane defined by the principal components 1 and 2. For each group, I calculated the mean distance of the individuals in that group from the average of the group as,

$$D = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}\,,$$

where n is the number of individuals in the group, x_i and y_i are the PC1 and PC2 values of individual i and $\bar{x}$ and $\bar{y}$ are the group means of PC1 and PC2, respectively. To compare the methods of SmartPCA and ChromoPainter, I scaled the D values of the groups by the scales of PC1 and PC2 of each method. The scaling was performed by sampling 100,000 random sets of individuals and calculating D as above for the sampled sets. The size of the set was the same as the size of the original group. Then, I calculated the ratio between the observed and randomly sampled groups as,

$$\frac{D_{obs}}{D_{samp}}.$$

I plotted the distribution of distance ratios with a density function in R. A small distance ratio indicates tight clustering while a larger value indicates looser clustering. The aim of this test was to discover whether ChomoPainter had in general smaller distance ratios than SmartPCA and thus tighter clustering.

4.6 Population clustering

FineSTRUCTURE aims to assign individuals into populations based on the coancestry matrix and it defines a population with three properties: 1) all the individuals in a population share equal amount of chunks and are thus equally related, 2) all the individuals receive an equal amount of chunks from other populations and, 3) donate an equal amount of chunks for the members of other populations (Lawson *et al.* 2012). The population assignment is evaluated with a likelihood model,

$$F(x|p,q) = \prod_{i=1,j=1}^{N} \left(\frac{P_{qiqj}}{n_{qj}}\right)^{x_{ij}/c},$$

where N is number of individuals, i and j represent the individuals in populations q_i and q_j, n_{qi} is the number of individuals in population q_i, P_{qiqj} is a population level coancestry matrix, x_{ij} is the chunk count in the *ij*-element of the coancestry matrix and c is the effective number of independent chunks. The c-value is defined by ChromoPainter and it models the fact that the chunks are not completely unlinked. The term P_{qiqj}/n_{qj} defines a likelihood for a single chunk being donated from j to i. Therefore, the total likelihood is the multiplication across all individuals. The algorithm of FineSTRUCTURE relies on a Markov chain Monte Carlo (MCMC) method in which the population assignment is iteratively searched and evaluated with the above likelihood. The Dirichlet distribution, which is often used in Bayesian frameworks, was used as a prior for the number and distribution of populations.

I performed the assignment of the individuals into populations with FineSTRUCTURE 0.0.5 (Lawson *et al.* 2012) using 200,000 MCMC iterations from which the first 100,000 rounds were discarded (burn-in) to make sure that the results are based on converged iterations. From the remaining 100,000 iterations I recorded only every 100th iteration to save disk space. From this set of 1,000 population assignments, I constructed the tree structure using the FineSTRUCTURE tree option and 10,000 additional hill-climbing iterations. The above options were used both for the analysis without assumption about the number of the populations, and the analyses of fixed number of populations from 2 to 18. All in all, I carried out 18 FineSTRUCTURE analyses.

To visualise populations and their geographic clustering, I plotted the individuals on a map of Finland and marked them according to which population they belonged into, based on FineSTRUCTURE clustering. I determined the position of the individual on the map as the average of the coordinates of his or her parents' home municipality. If only one of the parents' municipalities was known, the individual was positioned at those coordinates. If only the province of the parent was known, the individual was positioned at the centre of the province. The data contained 15 individuals whose parents' municipality was unknown, 2 individuals whose mother's municipality, and 3 individuals whose father's municipality were unknown. The map of Finland was from the GADM database (http://biogeo.ucdavis.edu/data/gadm2/R/FIN_adm0.RData) and the municipality coordinates are the coordinates of Finnish municipalities in 2011 (http://fba.evvk.com/kuntien_keskipisteet.html).

Table 2 Summary of the programs and their role in this work.

Program	**Used for**	**Reference**
PLINK v1.07	Quality control of the data	Purcell 2009
GCTA	Relatedness estimation	Yang *et al.* 2011
SmartPCA 8000	LD pruning, PCA of independent SNPs	Patterson *et al.* 2006
SHAPEIT version 2	Phasing of the genotype data	Delaneau *et al.* 2013
ChromoPainter	Calculation of coancestry matrix	Lawson *et al.* 2012
ChromoCombine	Combining coancestry matrices of different chromosomes	Lawson *et al.* 2012
FineSTRUCTURE	Population assignment	Lawson *et al.* 2012

5 Results

5.1 Principal component analysis of SmartPCA and ChromoPainter

To compare SmartPCA and ChromoPainter analyses, I performed PCA for both of the analyses using the data set of 345 individuals. The percentage of variance contained in the first ten PCs (see section 4 Materials and Methods) are shown in Table 3. The first PC of SmartPCA contains 0.502 % of the total variance, and the first PC of ChromoPainter contains 0.416 % of the total variance. The rest of the components of both methods contain a gradually decreasing portion of the variance starting from 0.370 % for SmartPCA and 0.330 % for ChromoPainter. The first PCs of ChromoPainter contain less variance than the first PCs of SmartPCA. This could be a result of a larger number of SNPs used in ChromoPainter: the larger amount of information is harder to compress into a small number of dimensions.

Table 3. Percentages of variance explained by the first ten PCs of SmartPCA and ChromoPainter.

Principal component	**SmartPCA (%)**	**ChromoPainter (%)**
PC1	0.502	0.416
PC2	0.370	0.330
PC3	0.364	0.323
PC4	0.356	0.313
PC5	0.352	0.312
PC6	0.351	0.308
PC7	0.350	0.306
PC8	0.349	0.305
PC9	0.348	0.305
PC10	0.347	0.305

The six principal components for both of the methods that contain over 0.35 % of the variance of the SmartPCA analysis are plotted against the first PCs in Figures 12, 13 and 14 where the individuals are coloured according to their province of birth. The first PCs distinguish the individuals strikingly well along the geographic East-West axis in both methods. The second component most clearly separates the individuals from the Province of Vaasa (Figures 12 A and B). Additionally, the ChromoPainter method separates the individuals from Northern Finland including Lapland and the Province of Oulu according to the second PC. Even though the second PC reveals some similarity to the geographic North-South gradient, it is not as evident as the gradient in the first PC. The third PC (Figures 12 C and D) shows again differences in individuals from LAP and VAS. The fourth principal component (Figures 13 A and B) distinguishes single individuals from the VAS. In the fifth PCs, the clear patterns of SmartPCA method start to diminish (Figure 13 C), but the ChromoPainter method successfully separates the individuals from USM, KYM and, TUP (Figure 13 D). The sixth PC of ChromoPainter still clearly distinguishes single individuals from Northern Finland (Figure 14 B) but no clustering can be detected in the rest of the PCs (Figures 14 C and D). According to these results, the first PCs clearly correlate with geographic distances. It seems also evident, that ChromoPainter captures more geographic structure than SmartPCA. In the next section, the results are compared quantitatively.

5.2 Results of quantitative comparison between SmartPCA and ChromoPainter

The numerical comparison between the methods was performed by calculating the ratio of observed and expected average distance from the group mean, for each of the 10 geographically defined groups. The comparison used the plane defined by PC1 and PC2, where expectation was computed by randomly sampling a group of individuals. The density plots of these ratios (Figure 15) and the means and standard deviations (Table 4) show clear differences between SmartPCA and ChromoPainter. The average distance ratios of SmartPCA are from 0.43 to 0.68 and the average distance ratios of ChromoPainter are from 0.26 to 0.57. Table 4 also shows that TUP, NOK and KUP are the most tightly clustered regions for both of the methods. The most loosely clustered regions are VAS and OUL for the SmartPCA and USM, HAM and KYM for ChromoPainter. ChromoPainter standard deviations vary from 0.18 to 0.40 and SmartPCA from 0.33 to 0.51.

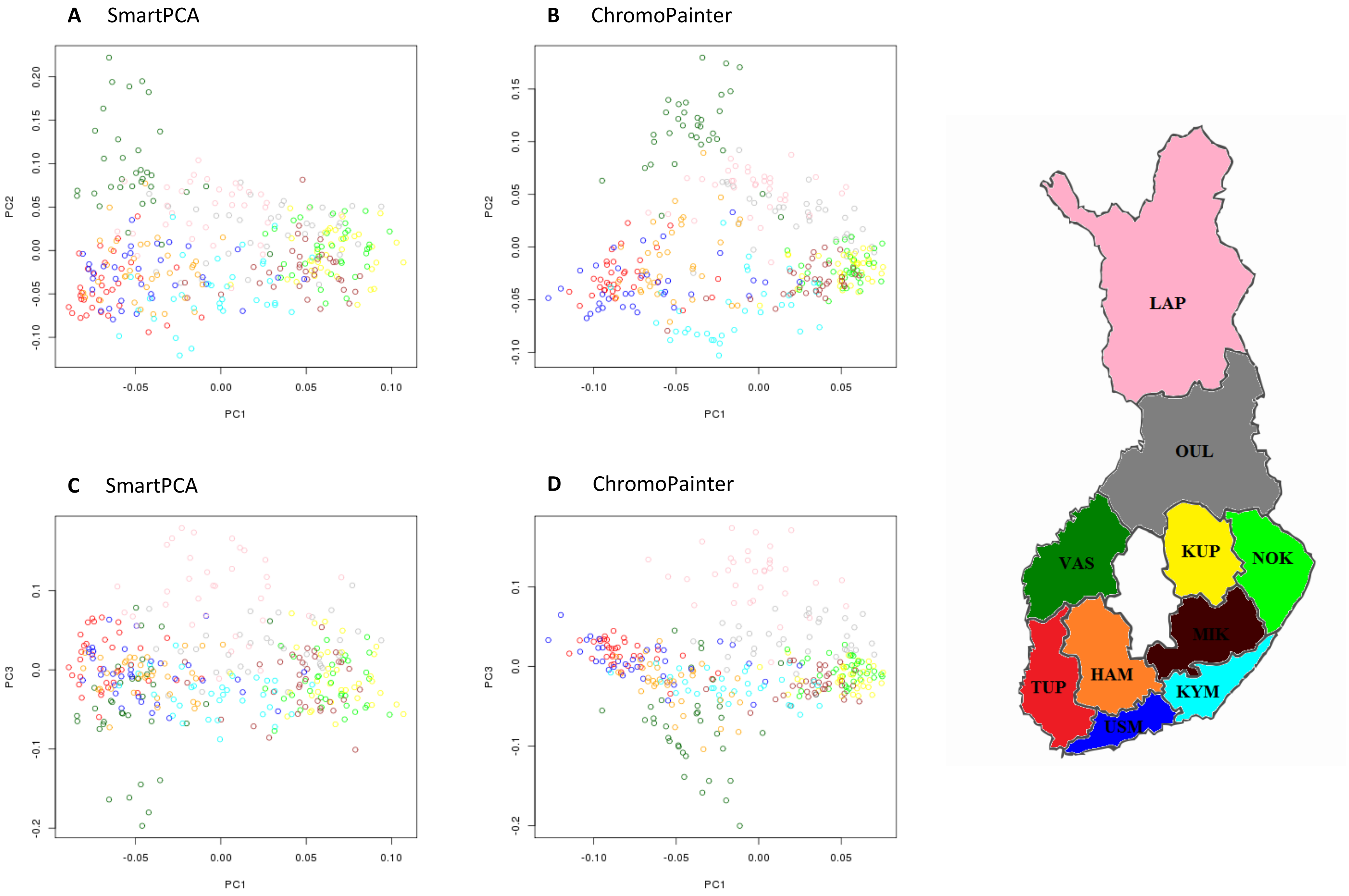


Figure 12 Principal component plots for SmartPCA and ChromoPainter. Individuals are coloured according to their region of origin shown at right. **A)** SmartPCA PC1 and PC2 **B)** ChromoPainter PC1 and PC2 **C)** SmartPCA PC1 and PC3 **D)** ChromoPainter PC1 and PC3.

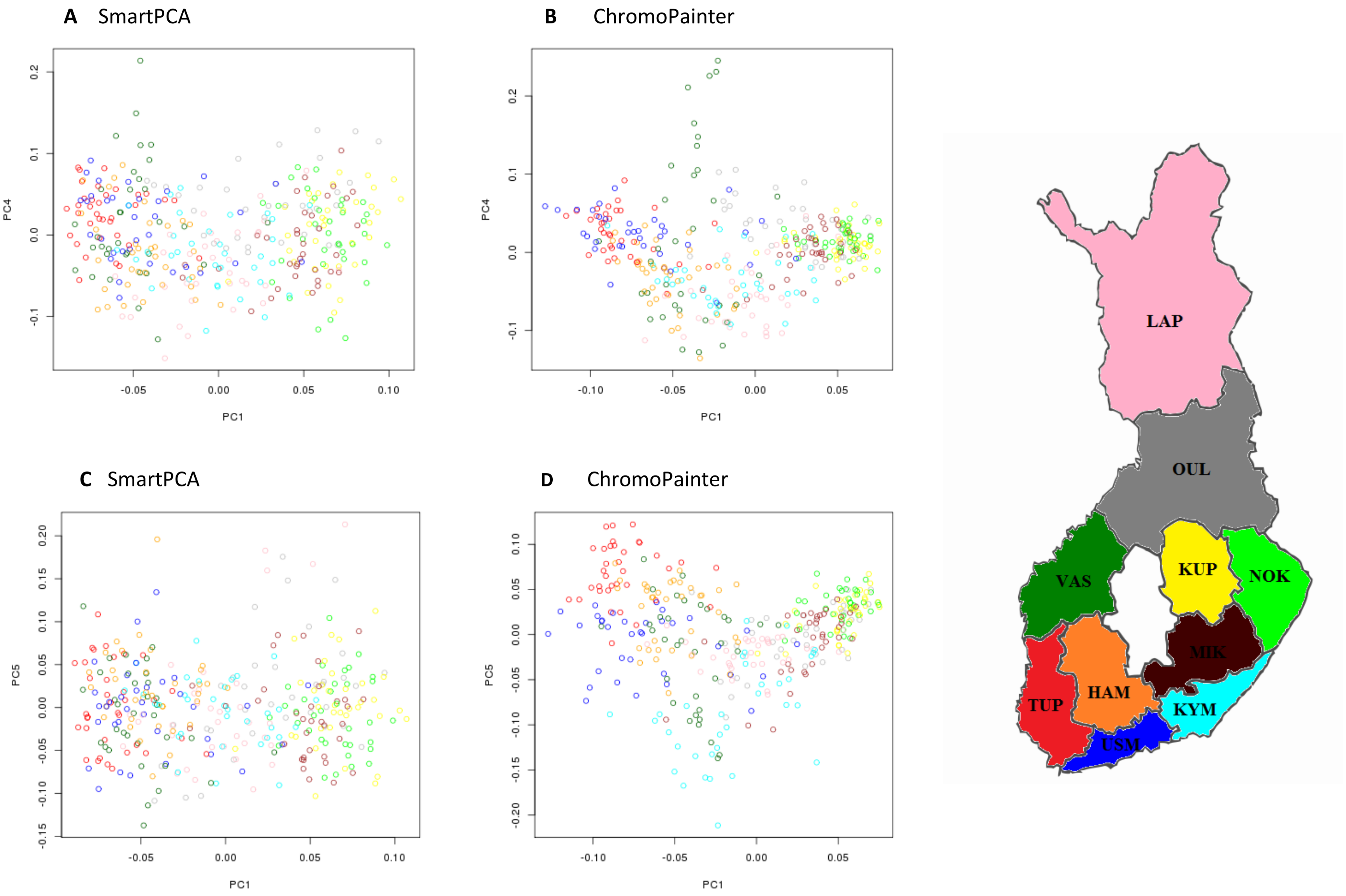


Figure 13 Principal component plots for SmartPCA and ChromoPainter. Individuals are coloured according to their region of origin shown at right. **A)** SmartPCA PC1 and PC4 **B)** ChromoPainter PC1 and PC4 **C)** SmartPCA PC1 and PC5 **D)** ChromoPainter PC1 and PC5.

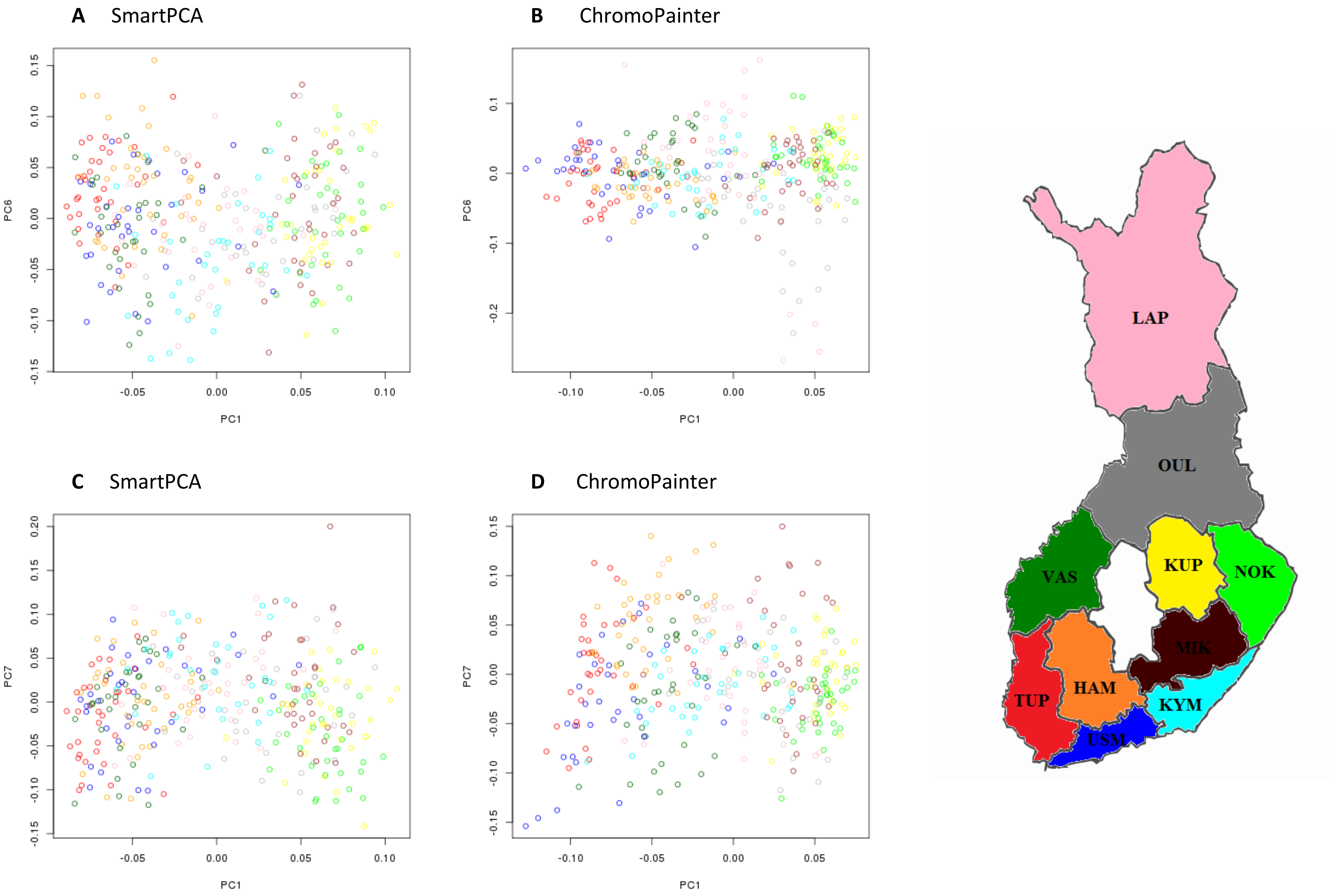


Figure 14 Principal component plots for SmartPCA and ChromoPainter. Individuals are coloured according to their region of origin shown at right. **A)** SmartPCA PC1 and PC6 **B)** ChromoPainter PC1 and PC6 **C)** SmartPCA PC1 and PC7 **D)** ChromoPainter PC1 and PC7.

The distance ratio of SmartPCA is smaller than that of ChromoPainter only in one group, HAM. Nevertheless, the distributions in this group are very close to each other. The distributions in USM and KYM are also similar between methods, while in other groups they are clearly separated. The reasons for ChromoPainter not clustering individuals significantly tighter than SmartPCA in these three groups could be related to the overall large distance ratios of these groups. Possible reasons for this are further discussed in section 6, Discussion.

Table 4 Means and standard deviations of the density plots in Figure 15.

	Mean SmartPCA	**Stdev. SmartPCA**	**Mean ChromoPainter**	**Stdev. ChromoPainter**
USM	0.56	0.042	0.53	0.037
TUP	0.44	0.033	0.33	0.023
HAM	0.50	0.038	0.52	0.036
KYM	0.59	0.044	0.57	0.040
MIK	0.45	0.033	0.32	0.022
NOK	0.44	0.033	0.26	0.018
KUP	0.43	0.033	0.26	0.018
VAS	0.68	0.051	0.47	0.033
OUL	0.63	0.047	0.49	0.034
LAP	0.57	0.043	0.41	0.028

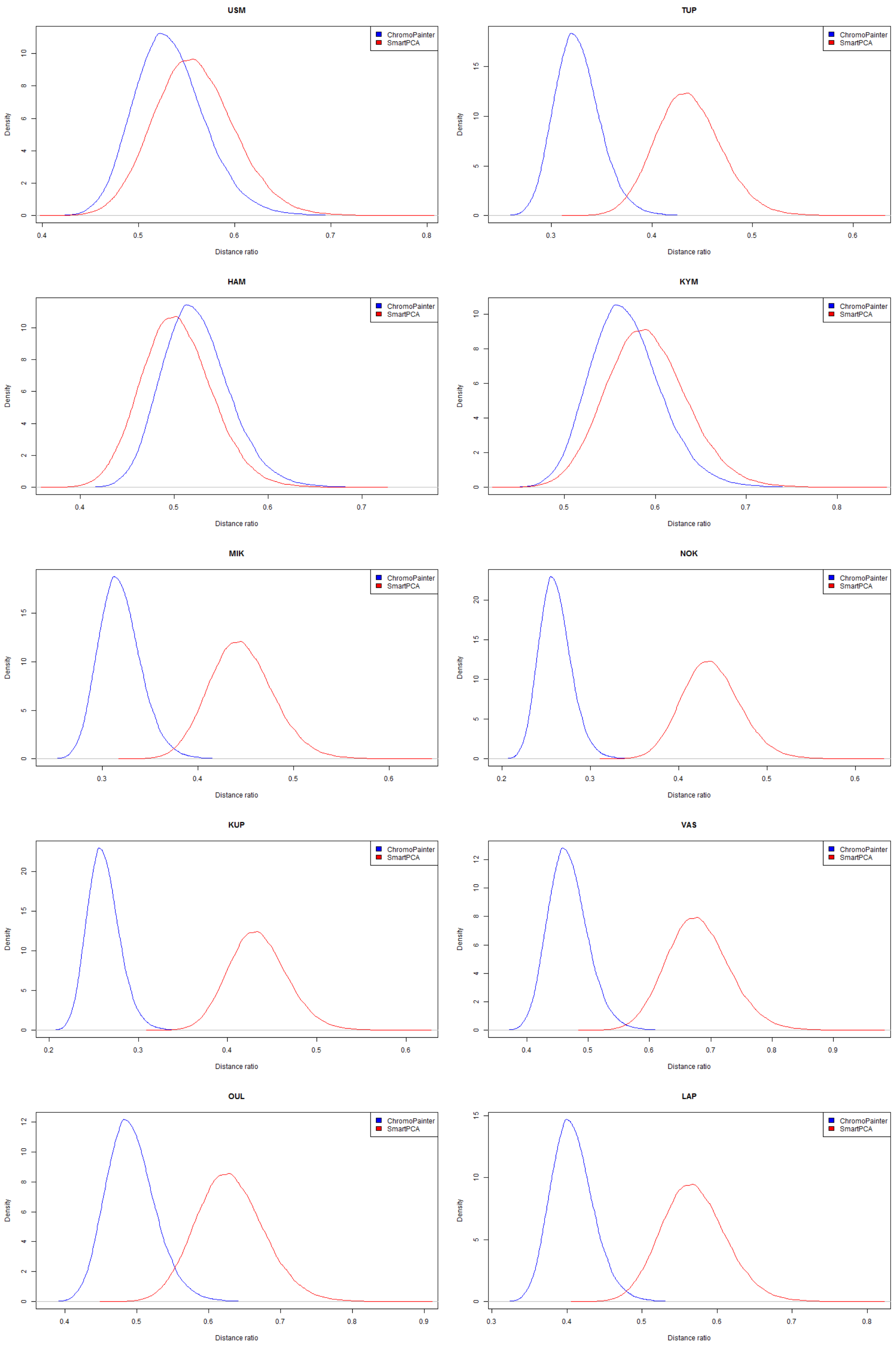


Figure 15 Density plots of the ratio of observed and randomly sampled grouping for principal components 1 and 2 (Figure 12 A and B). A striking difference in methods is seen in seven groups while in three groups (USM. HAM and KYM) the difference is not as clear.

5.3 Results of FineSTRUCTURE clustering

After comparison of SmartPCA and ChromoPainter I concentrated on studying population structure in Finland. Population assignment was performed using FineSTRUCTURE on a coancestry matrix, the output of ChromoPainter. FineSTRUCTURE was first run without an assumption of the number of populations and then with a fixed number of populations. In this section, the results of the population assignment are introduced. First, the visualisation of the coancestry matrix is presented. Second, the FineSTRUCTURE population assignment and the probability matrix are presented for the analysis without an assumption of the number of populations. Third, the same populations are presented on a map of Finland. Finally, the population assignments of the analyses with fixed number of populations are presented on maps of Finland.

ChromoPainter's coancestry matrix calculates the number of genomic chunks received from a donor individual and therefore it also presents the relationships between individuals. In Figure 16, the heat map of the coancestry matrix of this study is presented. The rows of the heat map represent the recipient individuals and columns represent the donor individuals. Dark colour denotes a high chunk count and light colour a low chunk count. Even though the heat map does not contain population assignment and the individuals are organised according to their birth provinces, we can see that there are similar patterns within individuals from the same region. For example, the individuals from VAS share the same pattern with each other and clearly distinguish from the rest. Similar pattern can be seen for the groups of LAP and KYM. Additionally, we can see that the individuals from MIK, NOK, KUP and OUL differ from the other provinces but show similar pattern with each other. This can already be seen as a sign of genetic differences between eastern and western parts of Finland.

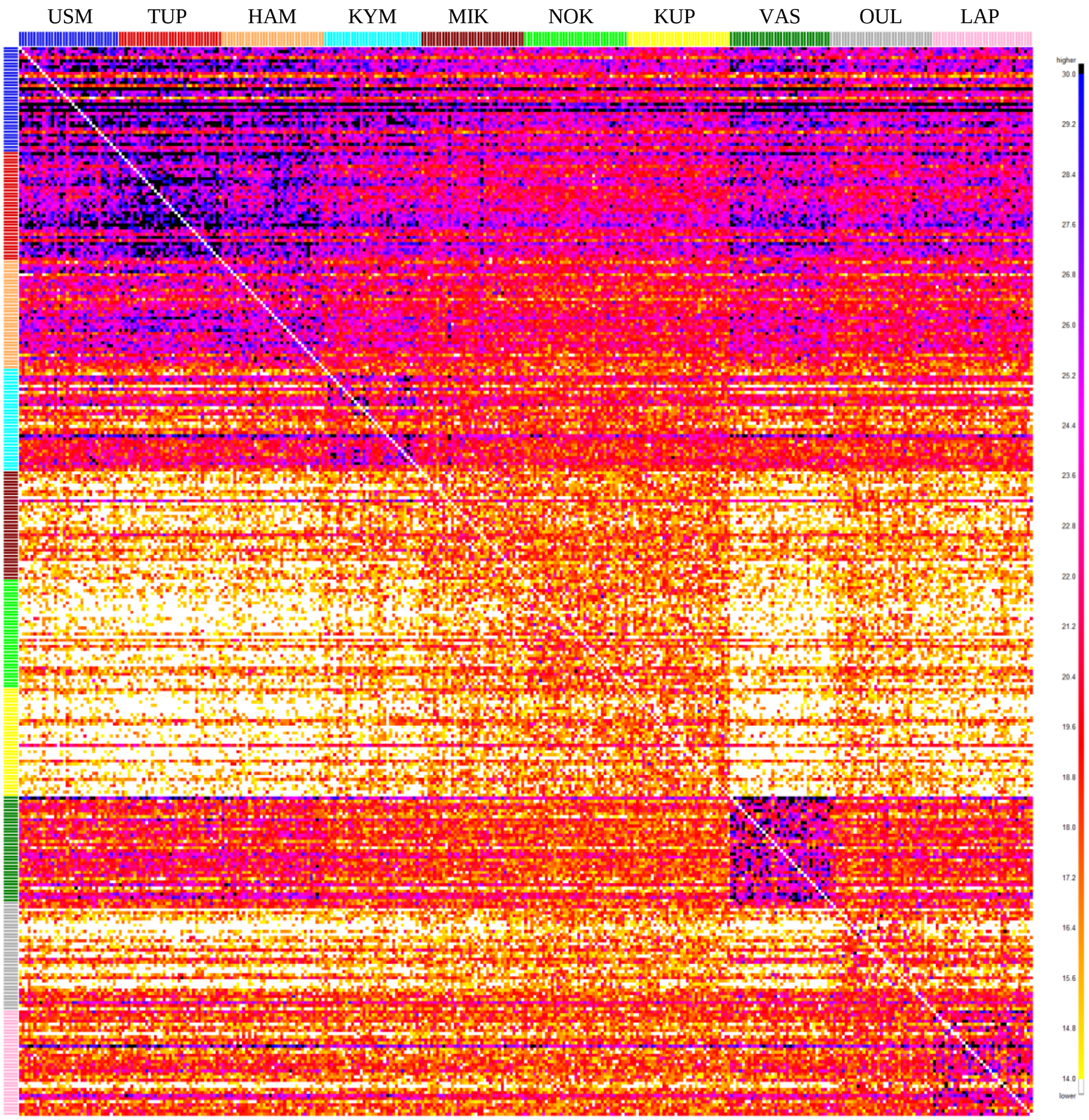


Figure 16 The heat map of ChromoPainter's coancestry matrix. Each row corresponds to the amount of chunks that is copied from the individuals on columns. Darker colours represent larger number of chunks copied. The labels of individuals correspond to the colouring in Figure 11. (Figure visualisation based on FineSTRUCTURE GIU Lawson *et al.* 2012)

The individuals were assigned into populations based on the coancestry matrix by FineSTRUCTURE and the assignment with no assumption of the number of populations is presented in Figure 17 A. In this analysis FineSTRUCTURE identified 15 populations and the populations were named to reflect their geographic location as shown in Figure 17 A. To ease the interpretation of the results I have written the name of the populations defined by FineSTRUCTURE in cursive. For example, Ostrobothnia means the geographic region and in cursive, *Ostrobothnia,* means the genetic population defined by FineSTRUCTURE.

From the assignment we can see that the individuals from the same regions have notably clustered into the same populations. The hierarchical tree structure of the FineSTRUCTURE clustering shows that the first division has happened between the south-western and north-eastern parts of Finland. The South-Western cluster includes four subpopulations of Ostrobothnia in one branch and four other populations in the other branch. The North-Eastern cluster includes two branches. The first one consists of one big population of individuals from NOK and KUP and the other includes six sub clusters of people from MIK, KYM, OUL and LAP.

In Figure 17 B, the probability matrix of the population assignment with no assumption of the number of the populations is shown. The matrix shows the probability of a pair of individuals to belong in the same population. The probability is calculated based on the MCMC iterations and the maximum a posteriori clustering (population assignment) that has the highest overall probability is presented. Therefore, the presented population assignment is not necessary the "best assignment" in which every individual is in the population where it most likely belongs. Thus, some individual can be assigned to one population even though he/she would belong to some other population with higher probability. For example, there are some individuals (red arrows in Figure 17 B) that should be included in another population according to their individual assignment probabilities. Thus, the probabilistic nature of the method should always be considered when interpreting the clustering results. In general, we see that each population has strong and smooth colouring supporting the population structure of 15 populations. The most improbable populations are small populations of the individuals from TUP and HAM (*Pirkanmaa*) and individuals from LAP (*Tornio*) as their blocks show approximately only 0.60 probability. The individuals of these populations have a considerable probability for belonging to other populations as well, which demonstrates how closely related these populations are to their neighbouring populations. The same close relatedness can also be seen in North-Eastern populations but not between South-Western and North-Eastern populations. The most surprising result is the strong support for the small populations within the individuals from VAS. This is further discussed in the section 6.

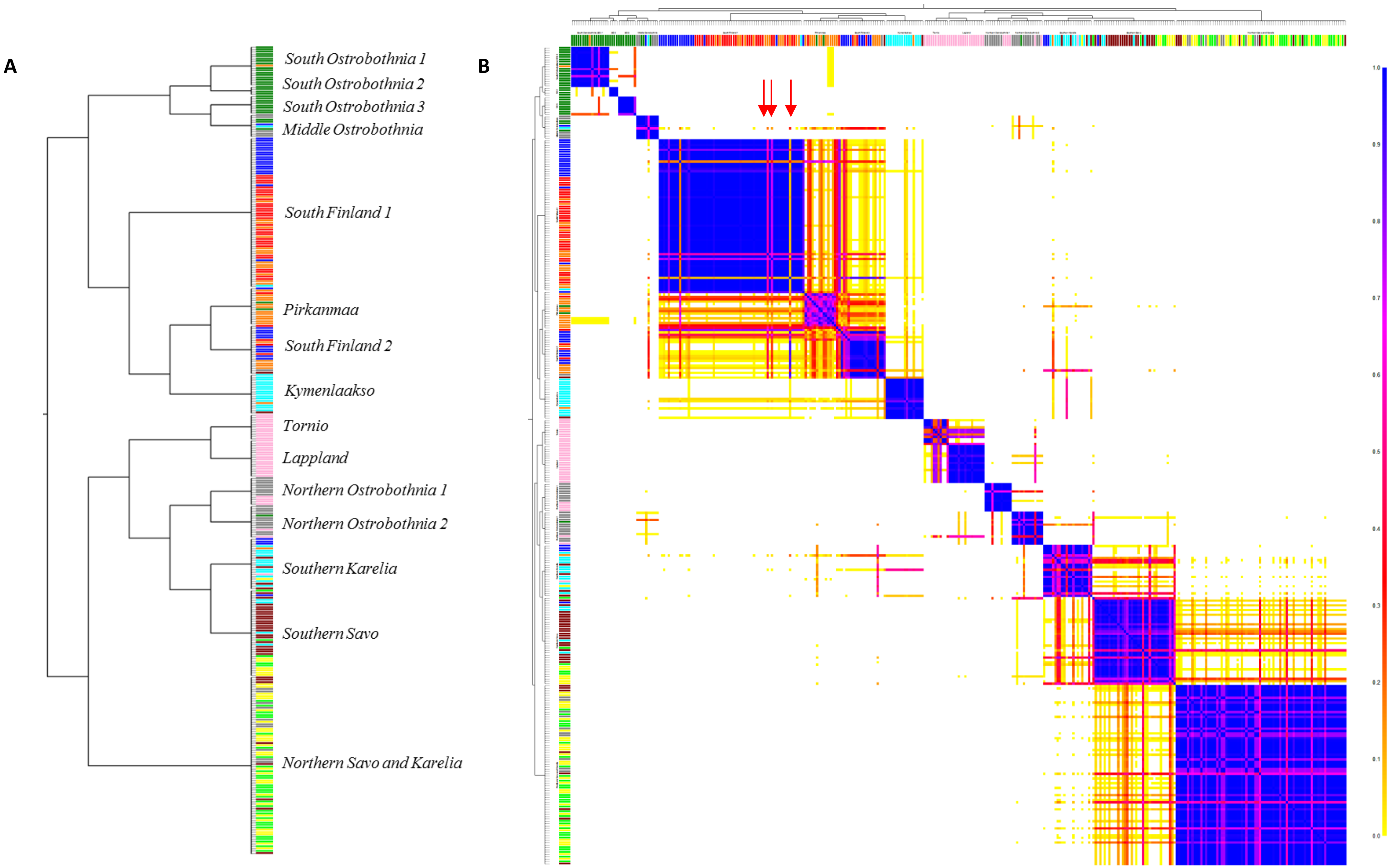


Figure 17 A) The FineSTRUCTURE clustering and **B)** the probability matrix for the analysis with no a priori assumption of the number of populations. The individual label colours (on top and on left) indicate the region of origin shown in Figure 11. The matrix colours (as described at the right) depict the probability that a pair of individuals belong to the same population. The dark colour at diagonal blocks shows that the population structure has high probability. The red arrows point to individuals that could be assigned into other populations as well.

To discern the FineSTRUCTURE population clustering geographically, the individuals were plotted according to their parents' birth places on the map of Finland (Figure 18). I point out the following five details. First, the populations of *Ostrobothnia 1, 2* and *3* are geographically near each other and the *Middle Ostrobothnia* is farther away even though they are genetically close. Second, the South coastal region of Finland, containing the Provinces of USM, TUP and HAM, is divided into three genetic populations, but only population of *Pirkanmaa* can be geographically distinguished. Third, the division of KYM into two genetically distinct populations, *Kymenlaakso* and *Southern Karelia,* is clearly motivated by the geography and the regional borders. The individuals from the region of Kymenlaakso are clustered into South-Western populations and the individuals from South Karelia are part of the North-Eastern populations. Fourth, the OUL is divided into two genetic populations that locate in the southern and northern parts of OUL. Finally, Figure 18 shows that there is no evident geographically motivated segregation of the populations in LAP even though the smaller population, *Tornio*, is clustered a little bit more to the Western Lapland, near the city of Tornio.

Although, the populations are geographically well clustered, there are a few individuals that clearly depart from it. The population of *Southern Karelia,* for example, includes individuals that originate from KUP, HAM and LAP. According to the probability matrix (Figure 17), the individual from HAM could also be clustered into the population of *South Finland 2*. The probability for other individuals to be included into the population of *Southern Karelia* seems to fluctuate a bit but does not explain the exceptional individuals. With this kind of study it is possible to detect individuals whose genetic background does not match their or their parents' birthplaces. Nevertheless, to verify these results of individual history we would need more information about the family history of the individuals.

As described in the section 3, FineSTRUCTURE was also run with the option that assumes a fixed number of populations. The results of clustering individuals into 2, 4, 6, 15 and 18 populations are shown on maps of Finland (Figure 19). Next, the geographic clustering and the special features of these maps are pointed out. The additional maps and the probability matrices are presented in the supplementary materials (Figures S2-S21).

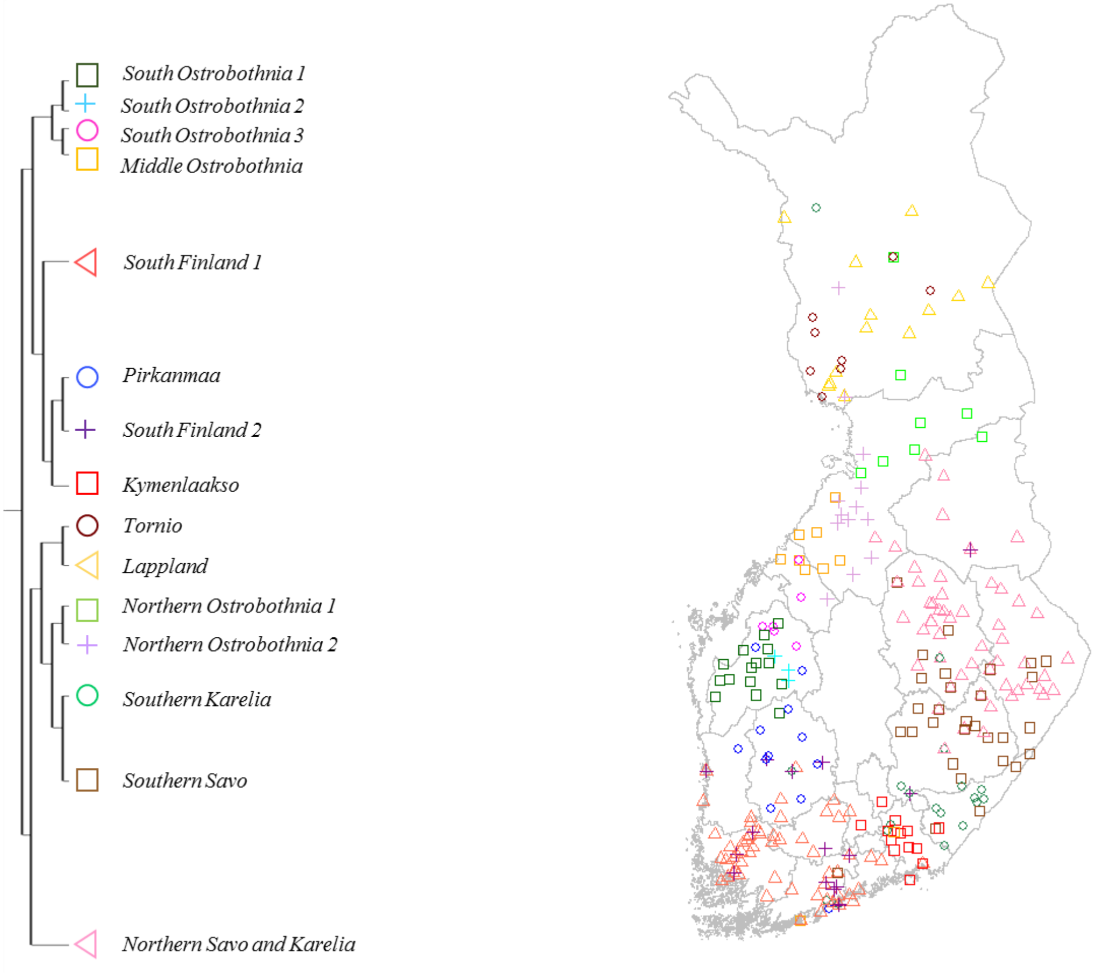


Figure 18 The population division by FineSTRUCTURE analysis with no a priori assumption of the number of populations. The individuals are plotted on the map according to the average coordinates of their parents' birth places. The provinces from which the individuals were chosen are shown in Figure 11. Individuals are marked with the label of the population in which they belong according to the analysis. The population names (given by myself), hierarchy and labels are shown at left.

Figure 19 A shows bimodal genetic population structure of Finland. The individuals are distinguished into South-Western and North-Eastern populations. Surprisingly, the individuals from LAP are clustered into the South-Eastern population. Nevertheless, the clustering of individuals from LAP into the South-Eastern population can be questioned based on the probability matrix (Figure S5). The probability of these individuals to belong into the South-Western population is typically around 0.50 so they could have almost as well been included in the North-Eastern population. This shows that the model of two populations for Finland does not describe the genetics of LAP well.

The uncertainty of the border individuals of the South-Western and North-Eastern populations can also be noticed in Figure 19 B. There, the third population consists of the individuals from the border regions of the East-West division. For this reason I call this population as *Border population*. It is also reasonable that the uncertain individuals from LAP are in *Border population.* In addition, the south-western parts of Finland have been divided into two genetic populations. The individuals from VAS are distinguished into its own population and the individuals from USM, TUP and HAM into the other. I call these populations respectively as the populations of *Ostrobothnia* and *Southern Finland*.

Figure 19 C shows the geographic localisation of six genetic populations. The hierarchy of these populations (Figure S9) shows that the border population observed in Figure 19 B has further divided into Western and Eastern populations. These populations are not clustered geographically and the individuals of these populations can be found both from the Western coast of Finland and the eastern border of Finland. Nevertheless, the individuals from LAP have been clustered into their own population and hierarchically they belong into the Eastern populations.

The division of the individuals into a fixed number of populations converge to the pattern shown in Figure 18 as the number of populations increases. In Figure 19 D, the assumption of fifteen populations is used and the similarity to the Figure 18 is striking. Very similar populations of *South Ostrobothnia 1, 2* and *3, Middle Ostrobothnia*, *South Finland 1 and 2, Pirkanmaa, Kymenlaakso, Lapland, North Ostrobothnia 1 and 2, Southern Karelia, Southern Savo* and, *Northern Savo and Karelia* are found. The only exception is that the analysis with the assumption 15 populations did not find the second cluster in Lapland, the population of *Tornio*. Instead, it had marked the three individuals from HAM to be an additional population.

The FineSTRUCTURE was also run with the assumption of more than 15 populations. As can be seen in Figure 19 E with 18 populations, the new populations are small, including two to four individuals, and located in OUL. The population of four individuals comes from the region of Kainuu but the other new populations do not have as clear geographic localisation. Due to these features and the small sample size, I decided not to divide the sample into any larger number of populations.

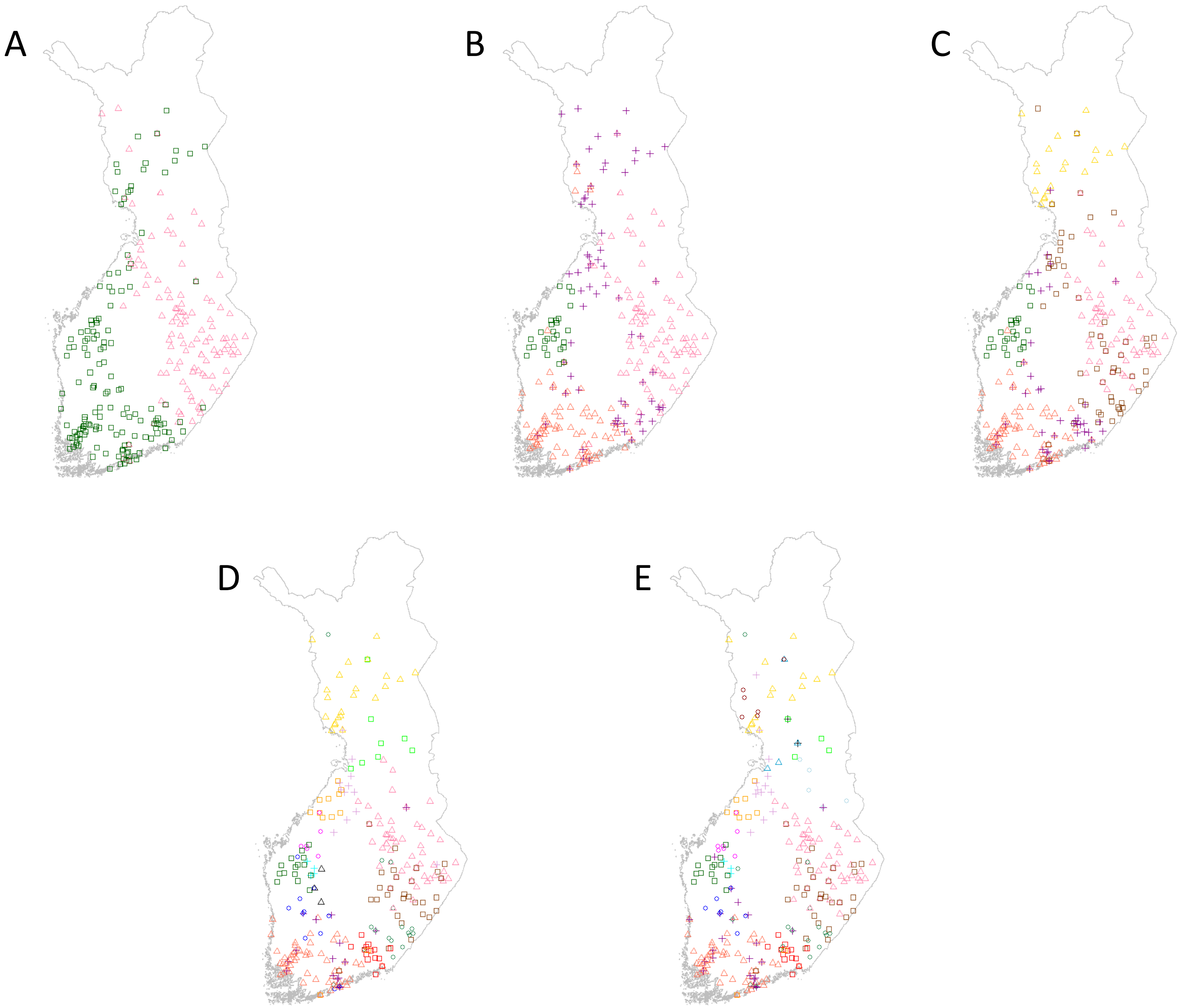


Figure 19 FineSTRUCTURE analysis based on fixed number of populations. Individuals are plotted on the map as in Figure 17. Analysis assumed **A)** 2 populations, **B)** 4 populations, **C)** 6 populations, **D)** 15 populations and **E)** 18 populations.

6 Discussion

6.1 Chromosome painting outperforms the standard PCA in details but loses in usability

The first aim of this study was to compare two methods for studying population structure. These methods are the standard principal component analysis based on independent SNPs, using the SmartPCA program, and the haplotype-based chromosome painting method, using the ChromoPainter program. As the advantage of ChromoPainter is its ability to use more markers than SmartPCA, the result is not based only on the algorithms but also on the information gained by using linked markers. Both of the methods were able to analyse my data set of 345 individuals easily. In fact, both of the methods can analyse thousands of individuals and hundreds of thousands of biallelic markers. For example, ChromoPainter has been used for analysing a data set of 938 individuals and 641,000 markers (Lawson *et al.* 2012) and PCA has been used for 2,051 individuals and 296,553 markers (McEvoy *et al.* 2009).

The time spent on the analysis restricts the size of the data set the most. In this study, SmartPCA took only a minute to perform the whole-genome analysis while the ChromoPainter analysis took from half an hour to almost four hours per chromosome depending on the chromosome length. Additionally, ChromoPainter requires the ChromoCombine program to merge the coancestry matrices of different chromosomes into one matrix which takes a few minutes. ChromoPainter also needs more preliminary analyses than SmartPCA. The input format of ChromoPainter is unique and thus the data must always be converted into this format. Luckily, the PaintMyChromosomes webpage ("PaintMyChromosomes.com", 24.10.2014) included ready-made scripts that can be used for data conversion. ChromoPainter's additional analyses include the estimation of the scaled global mutation rate and the scaling parameter N_e. Additionally, the output of ChromoPainter needs further analyses, in my case PCA and FineSTRUCTURE, and all together, the total analysis time for ChromoPainter based analyses is close to one day with this data set. Additionally, the algorithm and the features of the chromosome painting method are quite complex compared to the well-known PCA that feels simpler and easier to start with.

The PCA plots in Figures 12, 13 and 14 reveal that both of the methods capture the same basic features of the Finnish population. Nevertheless, ChromoPainter reveals geographic details even when standard PCA shows mostly noise. For example, the groups of USM and TUP (Figure 13 D) and NOK and KUP (Figure 14 D) can partially be distinguished from each other in the ChromoPainter analysis. ChromoPainter also separates more single individuals (for example Figure 14 B) than SmartPCA, which can lead to interesting results concerning an individual's ancestry.

The quantitative analyses showed that the clustering of ChromoPainter is much tighter than that of the standard PCA in most regions studied. This confirms that ChromoPainter captures the information of linked SNPs, and that this method is a very promising way to widen our understanding about the history and the structure of populations. Nonetheless, ChromoPainter did not do a significantly tighter clustering in all of the groups. In the groups of USM, HAM, and KYM, SmartPCA produced as tight or even tighter clustering than ChromoPainter. These regions have been inhabited the longest in Finland (Varilo 1999) and therefore the genetic background of the individuals might be more variable than that of individuals in other regions. If the genetic background of the individuals studied is not homogeneous, then we would not expect the individuals to cluster together this tightly. For example, in KYM, ChromoPainter together with FineSTRUCTURE detected two distinct genetic populations. Thus, the quantitative comparison shows not only differences between the two methods but also reveals that there are genetic population structures within the provinces of Finland.

6.2 Finnish people are divided into Western and Eastern populations

As has been demonstrated in earlier studies, the main genetic division of the Finns is into two subpopulations, *South-Western* and *North-Eastern* populations (Lappalainen *et al.* 2006, Hannelius *et al.* 2008, Jakkula *et al.* 2008, Lappalainen *et al.* 2008, Salmela *et al.*

2008). This division has been reported to be so strong that the genetic distance between people from East and West of Finland is larger than between some pairs of European populations that are geographically even further from each other (Hannelius *et al.* 2008, Jakkula *et al.* 2008, McEvoy *et al.* 2009). This genetic division is typically explained by the different population histories and internal migration (Salmela 2012), especially by the population expansion from the Southern Savo region to the eastern and northern parts of Finland in the 16th century. As the new villages were established by a small number of individuals and the population size remained small for long time, founder effect and random genetic drift have played a key part in the origin of the East-West population structure of Finland. The results of this study (e.g. Figure 19 A) are consistent with the previous studies as the East-West structure is the first division detected by FineSTRUCTURE.

The only exception in the FineSTRUCTURE result, shown in Figure 19 A, to the East-West division is that most of the individuals from Lapland seem to be clustered into the *South-Western* population. To understand why this is, we should look more carefully at the population assignment probabilities (Figure S5). Most of the individuals from LAP are clustered into the *South-Western* population, and also have a relatively high probability to be included into the *North-Eastern* population. Thus, these individuals do not belong to either of the two populations with high confidence. However, an explanation for these individuals to be assigned into *South-Western* population could be that the coast of the bay of Bothnia was inhabited relatively early, already before the 16th century from the South (Varilo 1999) and thus there might be Western influence there. Nevertheless, in more refined population assignments, the individuals from LAP belong to their own populations that belong to the Eastern populations according to the population hierarchies (see Figures S9-S22) and the result in Figure 19 A can be interpreted as a crude approximation of the genetic background of individuals from LAP.

The border between the *South-Western* and *North-Eastern* populations resembles significantly the border of the Treaty of Nöteborg in 1323. The treaty was an agreement between Sweden and Novgorod (a historical republic located in modern day Russia) and it defined a border and economic rights, such as taxation, for the participants (Korpela 2002) in the treaty. The border started from the Viborg castle, at the northern corner of the Gulf of Finland, and continued approximately through the Karelia region along the Sestra and Volchya rivers to Ostrobothnia and Pyhäjoki River. The exact border is still under debate. It is suggested that the border has never existed physically elsewhere than on a map (Korpela 2002, Katajala 2012). Thus, it has been claimed that the border did not affect lives of the people within the border regions and the border was not a barrier to marriages (Korpela 2002, Katajala 2012). Still, such a clear correlation between the treaty and the genetic border seems unlikely to be a coincidence. The genetic evidence seen here supports ideas that the border of the treaty might have had a cultural role during the internal migration and the population growth afterwards.

6.3 Finnish subpopulations are geographically clustered

Previous population studies of Finland have focused on the genetic population structure between South-Western and North-Eastern Finland and the relationship of the Finnish people to the other populations in Europe and worldwide. The genome-wide studies have only just recently allowed the more detailed study of population structure of founder populations (Jakkula *et al.* 2008, Wang *et al.* 2014). The results of this study (Figures 18 and 19) demonstrate that Finland is divided into several small genetic populations that are geographically clustered and, additionally, the geographic borders of these genetic populations closely resemble the borders of the provinces or the counties of Finland. As far as I know, a study of similar detail has not been carried outside Northern Finland (Jakkula *et al.* 2008). Next, the most important features of the genetic populations of this work are interpreted.

One of the first populations that stood out from the analysis was the population of Southern Ostrobothnia and its subpopulations (Figure 19 B). The strong separation of this population from the rest of Finland was a small surprise as the province of Vaasa (VAS) is not geographically isolated. Nevertheless, the cultural identity of the province is strong even today and the borders of the genetic population closely resemble the current dialect borders

("Suomen murrealueet", 2.12.2014). The strong population identity of Southern Ostrobothnia has also been supported by the study of multiple sclerosis (MS) (Tienari *et al.* 2004). The study suggests that the increased prevalence of multiple sclerosis in South Ostrobothnia is caused by the founder effect in the 13th and 16th centuries. The same effect may lie behind the genetic population structure as well. Additionally, Lappalainen (2009) also detected that the people from VAS are genetically distinguished from the people in other parts of South-Western Finland. The FineSTRUCTURE results (Figure 18) also showed subpopulation structure among the individuals from VAS. The population of *Middle Ostrobothnia* is a mixture of individuals from Northern VAS and Southern OUL and is clearly explained by the geographic clustering. In turn, the populations of *Ostrobothnia 2* and *3* are geographically mixed with the population of *Ostrobothnia 1*. The reason for these individuals being different from each other could be explained by linguistic, socio-economic or religious features. Unfortunately, I do not have that kind of data available and therefore cannot speculate on the results of this structure any further here.

South-Western Finland consists of three populations, *South Finland 1*, *2* and *Pirkanmaa*, of which only the last one is geographically distinct. South-Western Finland and the coastal regions are the earliest settlements in Finland (Varilo 1999). The people have had time to admix and therefore it is understandable that there are no strong geographic borders for genetic populations. This kind of a broad genetic background of the people of South-Western Finland is also captured in Lappalainen (2009) and it has been explained by the long population history and by the old capital of Finland, the city of Turku, being located in the region. The two geographically mixed populations, *South Finland 1* and *2*, are not too clearly distinguished from each other according to Figure 17. The reasons are probably subtle and complex and it would need additional information about the origin of individuals as in the case of Ostrobothnia. Nevertheless, the third population (*Pirkanmaa*) is clearly formed inside the borders of the county of Pirkanmaa and is most probably explained by the geography. The population of *Pirkanmaa* has not been detected in previous studies.

The northern parts of Finland, including LAP and OUL, showed several genetic populations. First, most of the individuals from LAP are clustered into one genetic population even though Jakkula et al. (2008) have shown that Lapland has distinct regional patterns of linkage disequilibrium and heterozygosity. To figure out whether Lapland has a finer subpopulation, it would be useful to study this region with a larger and more densely sampled population data. The small population of *Tornio* (Figures 18 and 19 E) would support the idea of subpopulations in this area. Second, the individuals of OUL are mainly divided into two populations that are geographically clustered into the southern and northern parts of OUL. Despite these quite distinct main populations, there are three smaller genetic populations in Figure 19 E. The simplest explanation is that these populations are just noise of overly detailed clustering. However, the population of five individuals (light blue circles in Figure 19 E) are clustered into the county of Kainuu, where different hereditary diseases have been detected. For example, lysinuric protein intolerance and congenital chloride diarrhoea are clustered in Kainuu (Norio *et al.* 1973). These disease features and the internal migration during the 16th century would support the theory that the northern parts of Finland might have even more detailed population structure and that the populations seen in Figure 19 E are not just noise. It would also be interesting to analyse the northern data of Jakkula *et al.* (2008) with this new method.

The genetic populations of *Southern Savo*, *Southern Karelia* and *Kymenlaakso* are geographically well defined, especially the border between the populations of *Southern Karelia* and *Kymenlaakso*. This border correlates so well with the Treaty of Nöteborg that it could well be the main reason for genetic population division. However, it was unexpected that the population of *Northern Savo and Karelia* is that uniform. It could have been possible that the founder effect of the internal migration in the 16th century (Varilo 1999) might have created a more scattered subpopulation structure. A more detailed analysis of this region would need more markers, especially rare variants.

Lastly, it would be interesting and very important to see in which populations the individuals from the province of Central Finland would be assigned. This region has many features in common with the Eastern populations, including dialect features and settlement history, but the admixture effect may also be strong. To my knowledge, the province of Central Finland has not been included in detailed studies of population history before and it was not included in the analyses of this study because there were not enough individuals from this province in this data set.

6.4 Possible sources of error and improvements

As this study is strongly based on the algorithms and implementations that have been tested earlier elsewhere, the problems that may affect the results are related to input data, user defined options and parameter estimation. As could be seen from Figure 18, the sample set covers well and evenly the regions that were studied. Nevertheless, the sample density was low in Lapland and the Province of Central Finland was missing entirely from the analysis. An increase of sample size, especially in these regions, would increase the accuracy and give more precise information about the population structure. Additionally, the number of SNPs could have been larger and the rare variants could have been included. The low frequency and rare variants would attain even more detailed differences between closely related populations. Nonetheless, the exclusion of the rare variants ensured the high quality of the data.

The estimates for the scaling parameter N_e and the default value for the mutation parameter θ were considerably smaller than those estimated from the British population (Leslie *et al.* 2015). This was eventually interpreted as a sign of strong genetic drift (Hellenthal, personal communication, 27.1.2015). In future analyses, the mutation parameter should also be estimated as the scaling parameter even though the effect is expected to be small.

As the individuals were assigned into the populations with the iterative MCMC algorithm, it is valid to ask whether 100,000 iterations were enough. I ensured that the number of iterations used is sufficient by comparing the probability matrix of 15 populations between two runs with 100,000 and 1,000,000 iterations, respectively (Figure S22). Although the comparison showed that the individual probabilities are a bit more precise when more iterations are used, the overall probability pattern and population division were exactly the same.

An improvement to the assignment of individuals into populations could be achieved by studying the probability matrix and reassigning those individuals whose probability for the current populations is lower than for some other population. For example, the third individual in Figure 17 has probability of around 0.1 for *South Finland 1* while the probability for *South Finland 2* is around 0.9. Thus, this individual could be reassigned into *South Finland 2*. Nevertheless, since the procedure would have made only minor changes that do not affect the broad conclusions and the developers of the method have not recommended it, it was not performed.

6.5 Future work

This study offers an exciting basis for future studies of the Finnish population both on a national and individual level. As the usefulness of the chromosome painting method has now been shown and the pipeline for running it is now in place, the next step is to improve the current study by increasing the sample size. Including more individuals and SNPs will increase the accuracy, geographic coverage, and provide us with good population reference data for future genetic studies in Finland.

It could also be possible to paint the chromosomes of an individual whose origin is unknown to us with the reference data and approximate the source proportions of different regional ancestry in the individual's genome (Hellenthal *et al.* 2014). With this kind of a detailed population structure, the ancestry might be approximated even on the level of Finnish counties. This application would definitely be interesting for the general public carrying out genealogy studies.

Another future application could be to compare the Finnish population structure to the genetic risk scores of complex diseases. The comparison of the population structure and the risk scores could answer to what extent genetics can explain regional differences in disease incidences. For example, the cardiovascular risk factors have shown differences between Eastern and Western Finland (Vartiainen *et al.* 2010) but it has not been consistently studied using detailed population structure.

Finally, chromosome painting could be used to compare Finns to the surrounding populations by studying and timing admixture events (Hellenthal *et al.* 2014).

7 Conclusion

In conclusion, the use of haplotype information as implemented through a chromosome painting method was able to provide a tighter and more precise clustering of Finnish genetic data than standard PCA that uses independent markers. Therefore, the chromosome painting method has proven its usefulness in detecting detailed population structure and should be the method of choice in future analyses of relatively homogenous populations where standard PCA fails to find substructure. Nevertheless, the standard PCA is still useful for a quick preliminary analysis that precedes more precise analyses.

The chromosome painting method was able to find new details about the population structure in Finland, such as the genetic populations of *Pirkanmaa* and *Kymenlaakso*. The results verify that the genetic populations in Finland are geographically clustered and several of them are found at the levels of provinces and counties.

8 Acknowledgements

First and foremost, I would like to thank my supervisors Matti Pirinen and Samuli Ripatti. Thank you, Matti, for your practical advice, help in troubleshooting and quick feedback in the writing process. Thank you, Samuli, for your professional help and for offering me the possibility to work with this project. I would also like to thank my family, Johann, Paula, Anne and Jyrki. Thank you for giving me motivation and especially support both in this project and throughout my studies. A big thank belongs also to Garrett Hellenthal who has written ChromoPainter. Thank you for visiting us in Helsinki and giving expert advice and exhilarating motivation for future projects. I would also want to thank anonymous participants of the FINRISK Study and the researchers of the FINRISK Study for permission to use the cohort in this study. Thank you, Aki Havulinna and Veikko Salomaa, for the interest towards the study and useful tips of plotting individuals on the map. In addition, thank you, Antti-Pekka Sarin, for your kind help with practical problems, data management, problem solving etc. Lastly, I would like to thank the Human Genomics people at FIMM, especially room F312a1, for the excellent working atmosphere and help.

9 References

Alberts, B., Johnson, A. & Lewis, J. 2002: General Recombination. *Molecular Biology of the Cell*. 4th edition ed. Garland Science New York.

Altshuler, D. M. 2010: Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52-58.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. & Zondervan, K. T. 2010: Data quality control in genetic case-control association studies. *Nature Protocols* 5: 1564-1573.

Borodulin, K., Saarikoski, L., Lund, L., Juolevi, A., Grönholm, M., Helldán, A., Peltonen, M., Laatikainen, T. & Vartiainen, E. 2013: *Kansallinen FINRISKI 2012 - terveystutkimus - Osa I: Tutkimuksen toteutus ja menetelmät*. THL, Tampere.

Butler, J. M. 2006: Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing. *Journal of Forensic Sciences (Wiley-Blackwell)* 51: 253-265.

Capelli, C., Redhead, N., Romano, V., Calì, F., Lefranc, G., Delague, V., Megarbane, A., Felice, A. E., Pascali, V. L., Neophytou, P. I., Poulli, Z., Novelletto, A., Malaspina, P., Terrenato, L., Berebbi, A., Fellous, M., Thomas, M. G. & Goldstein, D. B. 2006: Population Structure in the Mediterranean Basin: A Y Chromosome Perspective. *Annals of Human Genetics* 70: 207-225.

Carpelan, C. 1999: Käännekohtia Suomen esihistoriassa aikavälillä 5100–1000eKr. Pohjan poluilla. Suomalaisten juuret nykytutkimuksen mukaan. Bidrag till kännedom av Finlands natur och folk 153. Finska Vetenskaps-Societeten, Helsinki

Chakraborty, R. & Jin, L. 1993: A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. *EXS* 67: 153-75.

Chang, C. 2013: Documentation of SmartPCA program, Eigensoft, https://github.com/chrchang/eigensoft/tree/master/POPGEN, 30.10.2014

Check, E. 2005: Human genome: Patchwork people. *Nature* 437: 1084-1086.

Collins, F. S., Brooks, L. D. & Chakravarti, A. 1998: A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research* 8: 1229-1231.

Comas, D., Calafell, F., Bendukidze, N., Fañanás, L. & Bertranpetit, J. 2000: Georgian and kurd mtDNA sequence analysis shows a lack of correlation between languages and female genetic lineages. *American journal of physical anthropology* 112: 5-16.

Delaneau, O., Zagury, J. & Marchini, J. 2013: Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 10: 5-6.

Goldstein, J. I., Crenshaw, A., Carey, J., Grant, G. B., Maguire, J., Fromer, M., O'Dushlaine, C., Moran, J. L., Chambert, K., Stevens, C., Swedish Schizophrenia Consortium, ARRA Autism Sequencing Consortium, Sklar, P., Hultman, C. M., Purcell, S., McCarroll, S. A., Sullivan, P. F., Daly, M. J. & Neale, B. M. 2012: zCall: a rare variant caller for array-based genotyping. *Bioinformatics* 28: 2543-2545.

Hannelius, U., Salmela, E., Lappalainen, T., Guillot, G., Lindgren, C. M., von Döbeln, U., Lahermo, P. & Kere, J. 2008: Population substructure in Finland and Sweden revealed by the use of spatial coordinates and a small number of unlinked autosomal SNPs. *BMC Genetics* 9:54.

“HapMap phase II Build 37 recombination maps”, NCBI, http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37/, 24.7.2014

Harpending, H. & Jenkins, T. 1973: Genetic Distance Among Southern African populations. In: Crawford,M,Workman,P (ed.), *Method and theory in anthropological genetics.* 177-199. University of New Mexico Press.

Hartl, D. L. & Clark, A. G. cop. 2007: *Principles of population genetics*. Sinauer Associates, Sunderland, Mass.

Hedman, M., Pimenoff, V., Lukka, M., Sistonen, P. & Sajantila, A. 2004: Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineages. *Forensic Science International* 142: 37.

Hedman, M., Brandstätterb, A., Pimenoffa, V., Sistonenc, P., Paloa, J. U., Parsonb, W. & Sajantilaa, A. 2007: Finnish mitochondrial DNA HVS-I and HVS-II population data. *Forensic science international* 172: 171.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D. & Myers, S. 2014: A Genetic Atlas of Human Admixture History. *Science* 343: 747-751.

Henn, B. M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S. & Bustamante, C. D. 2010: Fine-scale population structure and the era of next-generation sequencing. *Human Molecular Genetics* 19: R221-R226.

Hirschfeld, L. & Hirschfeld, H. 1919: Serological differences between the blood of different races. the result of researches on the macedonian front. *Lancet* 194: 675-679.

Huurre, M. 2001: *Kivikauden Suomi*. Otava, Helsinki.

Ikäheimo, I., Silvennoinen-Kassinen, S. & Tiilikainen, A. 1996: HLA five-locus haplotypes in Finns. *European Journal Of Immunogenetics* 23: 321-328.

International Human Genome, S. C. 2001: Correction: Initial sequencing and analysis of the human genome. *Nature* 412: 565.

Jakkula, E., Rehnström, K., Varilo, T., Pietiläinen, O. P. H., Paunio, T., Pedersen, N. L., deFaire, U., Järvelin, M., Saharinen, J., Freimer, N., Ripatti, S., Purcell, S., Collins, A., Daly, M. J., Palotie, A. & Peltonen, L. 2008: The Genome-wide Patterns of Variation

Expose Significant Substructure in a Founder Population. *American Journal of Human Genetics* 83: 787-794.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, R., VanLiere, J. M., Fung, H., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A. & Singleto, A. B. 2008: Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003.

Katajala, K. 2012: Drawing Borders or Dividing Lands?: the peace treaty of 1323 between Sweden and Novgorod in a European context. *Scandinavian Journal of History* 37: 23-48.

Korpela, J. 2002: Finland's eastern border after the treaty of Nöteborg: An ecclesiastical, political or cultural border?. *Journal of Baltic Studies* 33: 384-397.

Lappalainen, T., Koivumäki, S., Salmela, E., Huoponen, K., Sistonen, P., Savontaus, M. & 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. & Lahermo, P. 2008: Migration Waves to the Baltic Sea Region. *Annals of Human Genetics* 72: 337-348.

Lappalainen, T. 2009: *Human genetic variation in the Baltic Sea region : features of population history and natural selection*. University of Helsinki, Institute for Molecular Medicine Finland & Department of Biological and Environmental Sciences, Helsinki.

Lao, O., Lu, T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A. G., Gieger, C., Wichmann, H-E., Rüther, E., Schreiber, S. & Becker, C. 2008: Correlation between Genetic and Geographic Structure in Europe. *Current Biology* 16: 1241–1248.

Lawson, D., Hellenthal, G., Myers, S. & Falush, D. 2012: Inference of population structure using dense haplotype data. *PLoS Genetics , 8 Article e1002453.*

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Lawson, D., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P. & Walter Bodmer, W. 2015: The fine-scale genetic structure of the British population. *Nature* 519: 309.

Levine, P. & Stetso, R. E. 1939: An unusual case of intra-group agglutination. *Journal of the American Medical Association* 113: 126.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M. & Cavalli-Sforza, L. L. 2008: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319: 1100-1104.

Li, N. & Stephens, M. 2003: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165: 2213.

McEvoy, B. P., Montgomery, G. W., McRae, A. F., Ripatti, S., Perola, M., Spector, T. D., Cherkas, L., Ahmadi, K. R., Boomsma, D., Willemsen, G., Hottenga, J. J., Pedersen, N. L., Magnusson, P. K. E., Kyvik, K. O., Christensen, K., Kaprio, J., Heikkilä, K., Palotie, A., Widen, E., Muilu, J., Syvänen, A., Liljedahl, U., Hardiman, O., Cronin, S., Peltonen, L., Martin, N. G. & Visscher, P. M. 2009: Geographical structure and differential natural selection among North European populations. *Genome research.* 19: 804-814.

McVean, G. A. 2012: An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.

Meinila, M., Finnilä, S. & Majamaa, K. 2001: Evidence for mtDNA admixture between the Finns and the Saami. *Human Heredity* 52: 160-170.

Mellars, P. 2004: Neanderthals and the modern human colonization of Europe. *Nature* 432: 461-465.

Mellars, P. 2006: Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proceedings of the National Academy of Science USA* 103: 9381.

Menozzi, P., Piazza, A. & Cavalli-Sforza, L. L. 1978: Synthetic maps of human gene frequencies in Europeans. *Science 92* 201: 786-92.

Nachman, M. W. & Crowell, S. L. 2000: Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics* 156: 297.

Nasidze, I., Sarkisian, T., Kerimov, A. & Stoneking, M. 2003: Testing hypotheses of language replacement in the Caucasus: evidence from the Y-chromosome. *Human Genetics* 112: 255-261.

NCBI dbSNP Build 142, http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi?view+summary=view+summary&build_id=142, 5.6.2015

Nevanlinna, H. R. 1972: The Finnish population structure A genetic and genealogical study. *Hereditas* 71: 195-235.

Norio, R., Nevanlinna, H. R. & Perheentupa, J. 1973: Hereditary diseases in Finland; rare flora in rare soul. *Annals of clinical research* 5: 109-41.

Novembre, J. & Stephens, M. 2008: Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* 40: 646-649.

Oinonen, M., Pesonen, P., Alenius, T., Heyd, V., Holmqvist-Saukkonen, E., Kivimäki, S., Nygrén, T., Sundell, T. & Onkamo, P. 2014: Event reconstruction through Bayesian chronology: Massive mid-Holocene lake-burst triggered large-scale ecological and cultural change. *The Holocene* 24: 1419-1427.

“Paintmychromosomes.com”, http://www.paintmychromosomes.com/, 24.10.2014

Patterson, N., Price, A. L. & Reich, D. 2006: Population structure and eigenanalysis. *PLoS Genetics* 2: 2074-2093.

“PED files”, PLINK Whole genome association analysis toolset, http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped, 14.1.2015

Pesonen, P. 2005: Sarvingin salaisuus – Enon Rahakankaan varhaismesoliittinen ajoitus. *Muinaistutkija* 2: 2-13.

Peterson, L. E. 2013: *Classification analysis of DNA microarrays*. Wiley-IEEE Computer Society Press, Ringgold Inc.

Pitkänen, K. 2007: Suomen väestön historialliset kehityslinjat. In Koskinen S, Martelin T, Notkola IL et al: Suomen väestö. Gaudeamus Helsinki University Press, Tampere

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. 2006: Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 38: 904-909.

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., Goldstein, D. B. & Reich, D. 2008: Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* 83: 132-5.

Pritchard, J. K., Stephens, M. & Donnelly, P. 2000: Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155: 945.

Pritchard, J. K. 2001: Linkage Disequilibrium in Humans: Models and Data. *American Journal of Human Genetics* 69: 1.

Purcell, S. 2007: PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* 81: 559-575.

Purcell, S. 2009: PLINK 1.07, http://pngu.mgh.harvard.edu/purcell/plink/

R Core Team. 2014: R: A language and environment for statistical computing.

Reich, D., Price, A. L. & Patterson, N. 2008: Principal component analysis of genetic data. *Nature genetics* 40: 491-492.

Riggs, E. R., Ledbetter, D. H. & Martin, C. L. 2014: Genomic Variation: Lessons Learned from Whole-Genome CNV Analysis. *Current Genetic Medicine Reports* 2: 146-150.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., Altshuler, D. & International SNP Map Working Group. 2001: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928.

Salmela, E., Lappalainen, T., Fransson, I., Andersen, P. M., Dahlman-Wright, K., Fiebig, A., Sistonen, P., Savontaus, M., Schreiber, S., Kere, J. & Lahermo, P. 2008: Genome-Wide Analysis of Single Nucleotide Polymorphisms Uncovers Population Structure in Northern Europe (SNP Variation in North Europe). *PLoS ONE 3*: e3519.

Salmela, E. 2012: *Genetic structure in Finland and Sweden: aspects of population history and gene mapping*. Department of Medical Genetics, University of Helsinki, Helsinki.

Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. 1998: Genetic evidence for a higher female migration rate in humans. *Nature Genetics* 20: 278-280.

Shlens, J. 2014: A Tutorial on Principal Component Analysis. *ArXiv 1404.1100v1*

Sikora, M., Laayouni, H., Calafell, F., Comas, D. & Bertranpetit, J. 2011: A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. *European Journal of Human Genetics* 19: 84-88.

Siren, M. K., Sareneva, H., Lokki, M. L. & Koskimie, S. 1996: Unique HLA antigen frequencies in the Finnish population. *Tissue Antigens* 48: 703-707.

Sequencing Initiative Suomi, http://www.sisuproject.fi, 24.11.2014

Smithies, O. 1955: Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. *Biochemical Journal* 61: 629-41.

Stoneking, M., Fontius, J. J., Clifford, S. L., Soodyall, H., Arcot, S. S., Saha, N., Jenkins, T., Tahir, M. A., Deininger, P. L. & Batzer M A. 1997: Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Research* 7: 1061-1071.

"Suomen murrealueet", Kotimaisten kielten keskus, http://www.kotus.fi/index.phtml?s=368#alku, 2.12.2014

"Suomen väkiluku", Väestörekisterikeskus, http://vrk.fi/default.aspx?docid=169, 26.2.2015

Taavitsainen, J-P., Simola, H. & Grönlund, E. 1998: Cultivation History Beyond the Periphery: Early Agriculture in the North European Boreal Forest. *Journal of World Prehistory* 2: 199-253.

Takala, H. 2004: Archeological research in the former Jurisdictional district of Äyräpää and excavations at the Telkkälä site in Muolaa. *Museoviraston arkeologian osaston julkaisu no. 10: 117-123.* Fenno-ugri et slavi. Helsinki: Museovirasto

Tallavaara, M., Pesonen, P. & Oinonen, M. 2010: Prehistoric population history in eastern Fennoscandia. *Journal of Archaeological Science* 37: 251-260.

Tienari, P. J., Sumelahti, M. L., Rantamäki, T. & Wikström, J. 2004: Multiple sclerosis in western Finland: evidence for a founder effect. *Clinical Neurology & Neurosurgery* 106: 175-179.

Wang, C., Zöllner, S. & Rosenberg, N. A. 2014: Ancestry estimation and control of population stratification for sequence-based association studies. *Nature Genetics* 46: 409.

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., Branham, K. E., Heckenlively, J., The FUSION Study, Fulton, R., Wilson, R. K., Mardis, E. R., Lin, X., Swaroop, A., Zöllner, S. & Abecasis, G. R. 2012: A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS Genetics* 8: 1-16.

Varilo, T. 1999: The age of the mutations in the Finnish disease heritage a genealogical and linkage disequilibrium study. *National Public Health Institute*, Helsinki.

Vartiainen, E., Laatikainen, T., Peltonen, M., Juolevi, A., Männistö, S., Sundvall , J., Jousilahti, P., Salomaa, V., Valsta, L. & Puska, P. 2010: Thirty-five-year trends in cardiovascular risk factors in Finland. *International journal of epidemiology* 39: 504-518.

Vartiainen, E., Jousilahti P., Juolevi A., Sundvall J., Alfthan G., Salminen I. & Puska P. 1998: Finriski 1997. Tutkimus kroonisten kansantautien riskitekijöistä, niihin liittyvistä elintavoista, oireista ja terveyspalvelujen käytöstä. Tutkimuksen toteutus ja perustaulukot. *Kansanterveyslaitoksen julkaisuja B1/1998.* Paintmedia, Helsinki 1998.

”Väestön kehitys esihistoriallisella ajalla”, Museovirasto, http://www.nba.fi/fi/kansallismuseo/opetus/opetuspaketit/esihistoria/tietoa/vaesto, 20.5.2015

Watterson, G. A. 1975: On the number of segregating sites in genetical models without recombination. *Theoretical population biology* 7: 256-276.

Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. 2011: GCTA: A Tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics* 88: 76-82.

1000 Genomes, A Deep Gatalog of Human Genetic Variation, www.1000genomes.org, 24.11.2014

Supplementary Material

Figure S1

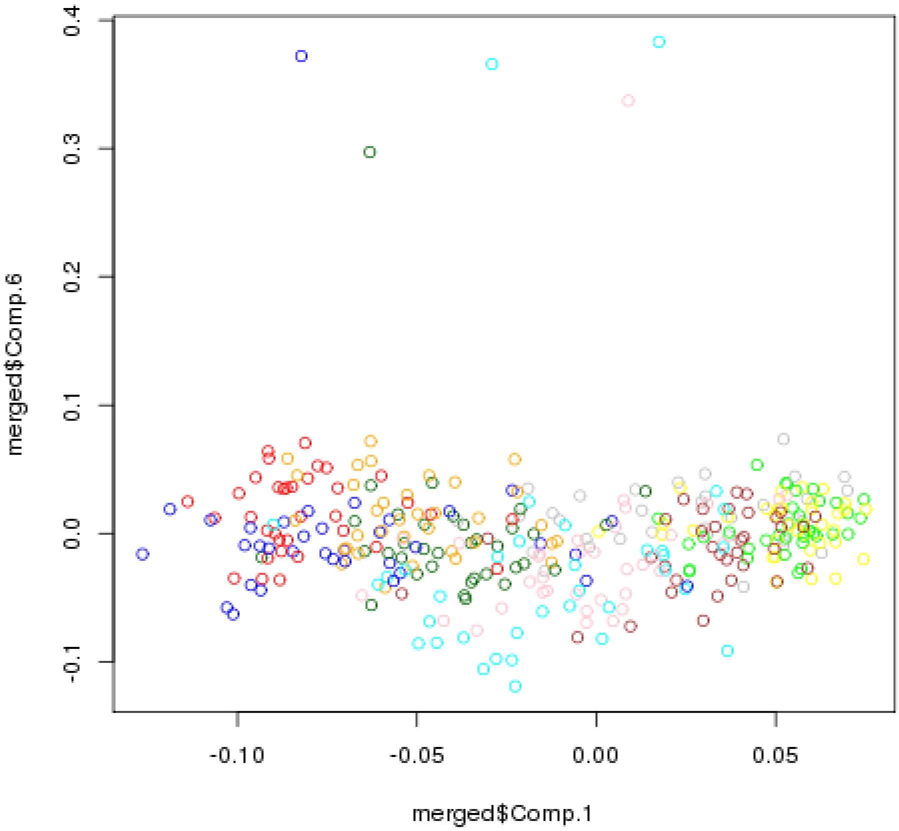


Figure S1 Principal components 1 and 6 of ChromoPainter. The 350 individuals from 10 provinces of Finland are coloured according to Figure X. The five individuals that are separated from the rest were removed from the further analysis.

Figure S2

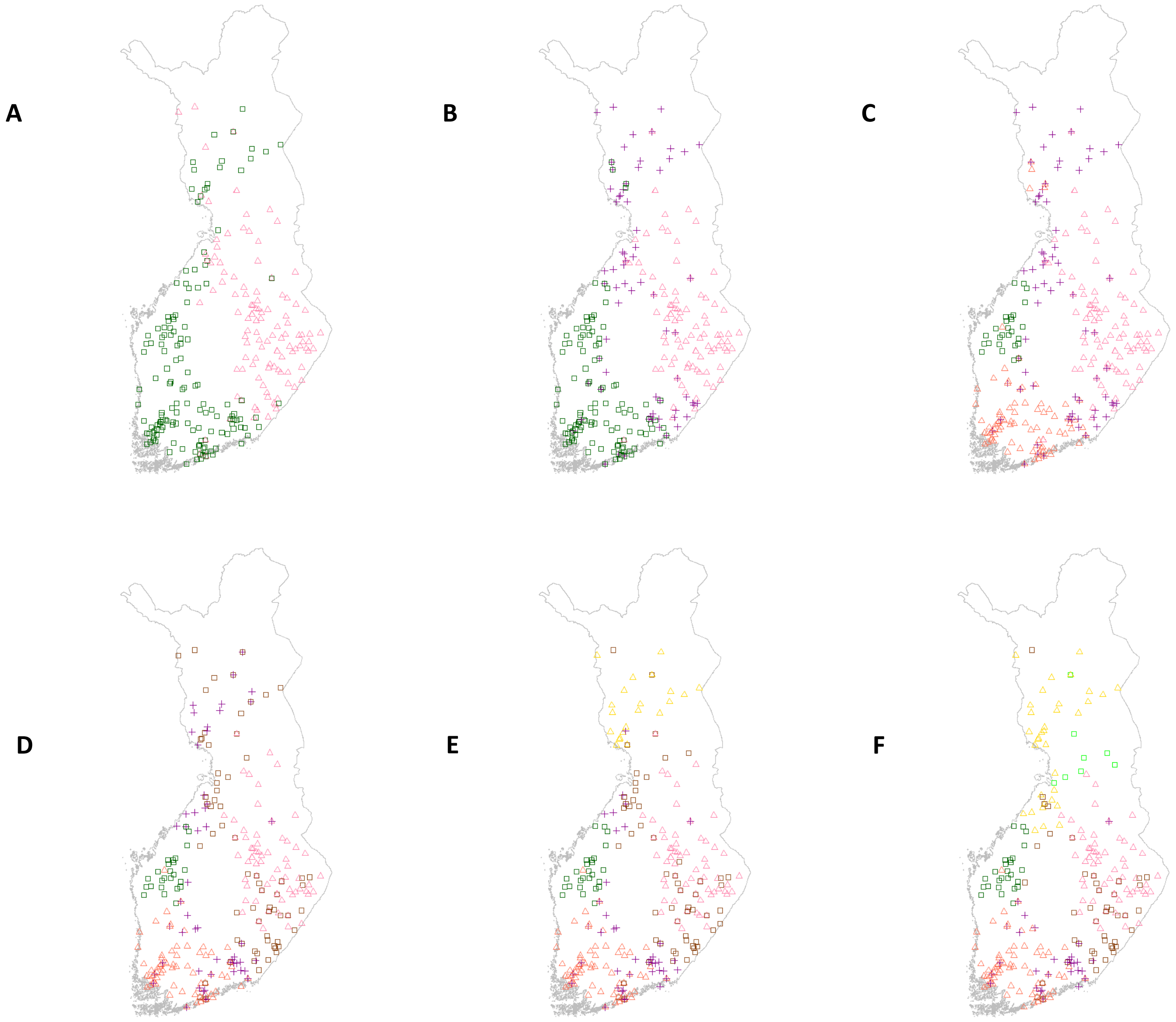


Figure S2 FineSTRUCTURE analysis based on the assumption of population numbers. A) 2 Populations B) 3 Populations C) 4 Populations D) 5 Populations E) 6 Populations F) 7 Populations

Figure S3

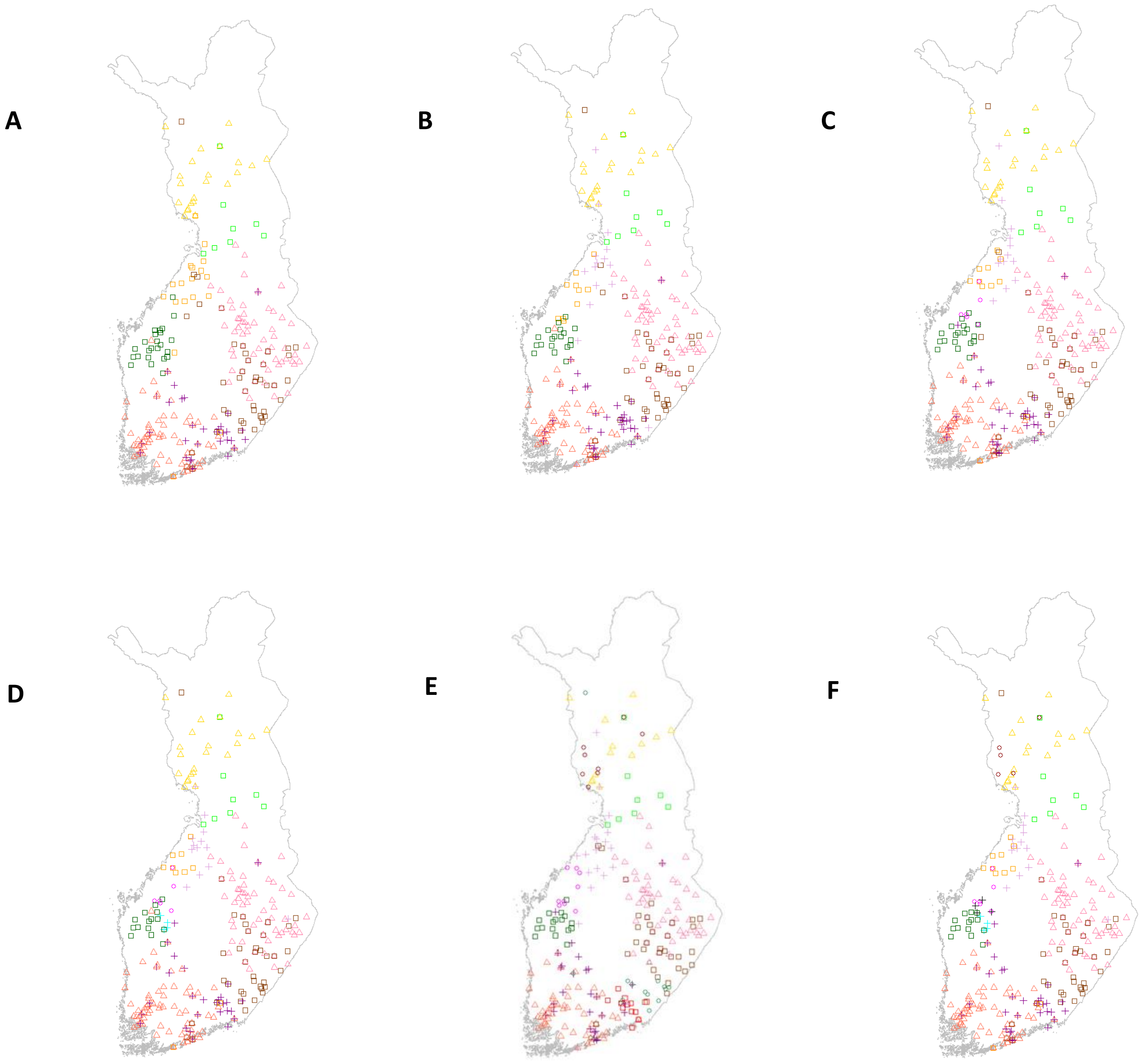


Figure S3 FineSTRUCTURE analysis based on the assumption of population numbers. A) 8 Populations B) 9 Populations C) 10 Populations D) 11 Populations E) 12 Populations F) 13 Populations

Figure S4

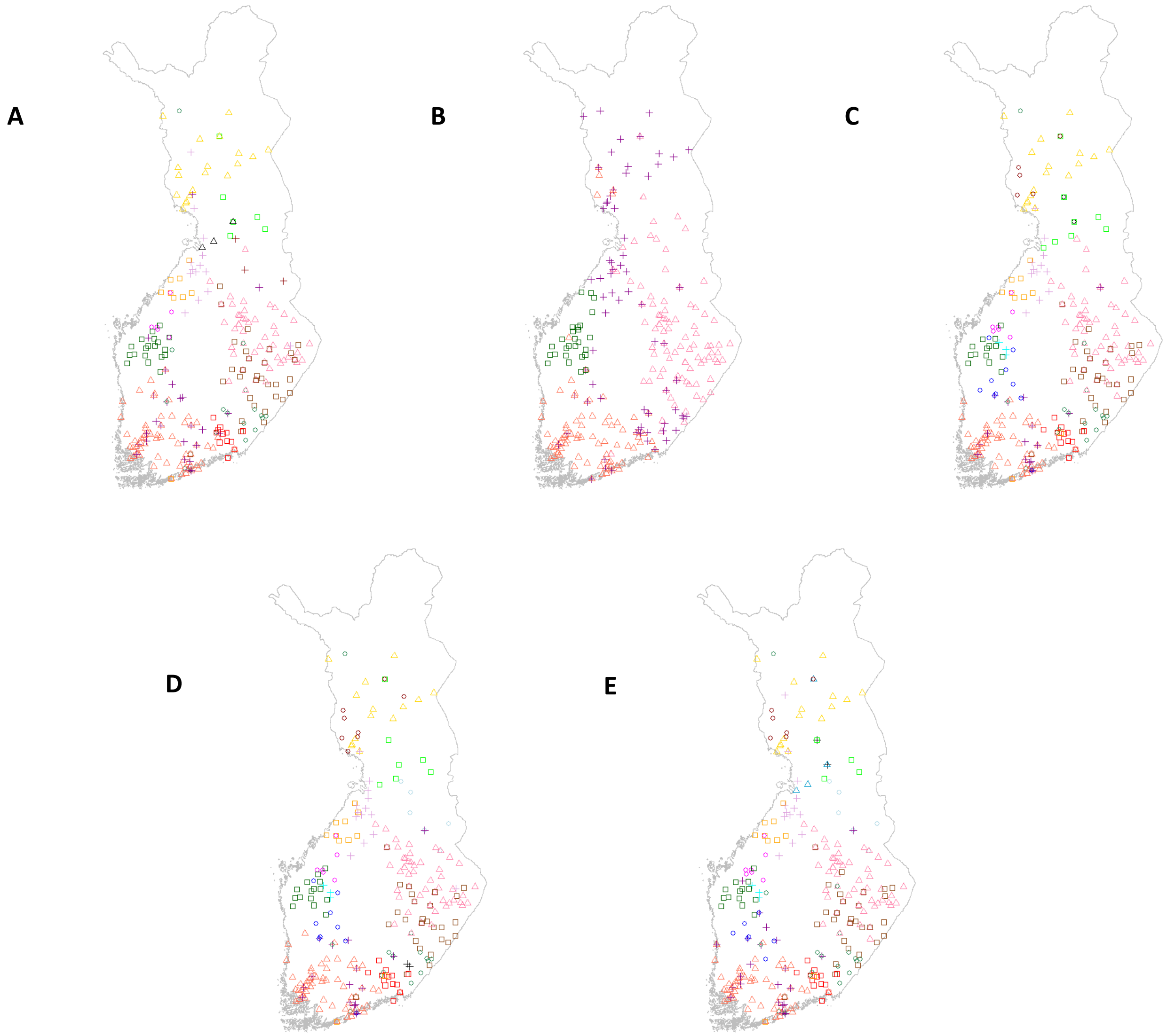


Figure S4 FineSTRUCTURE analysis based on the assumption of population numbers. A) 14 Populations B) 15 Populations C) 16 Populations D) 17 Populations E) 18 Populations

Figure S5

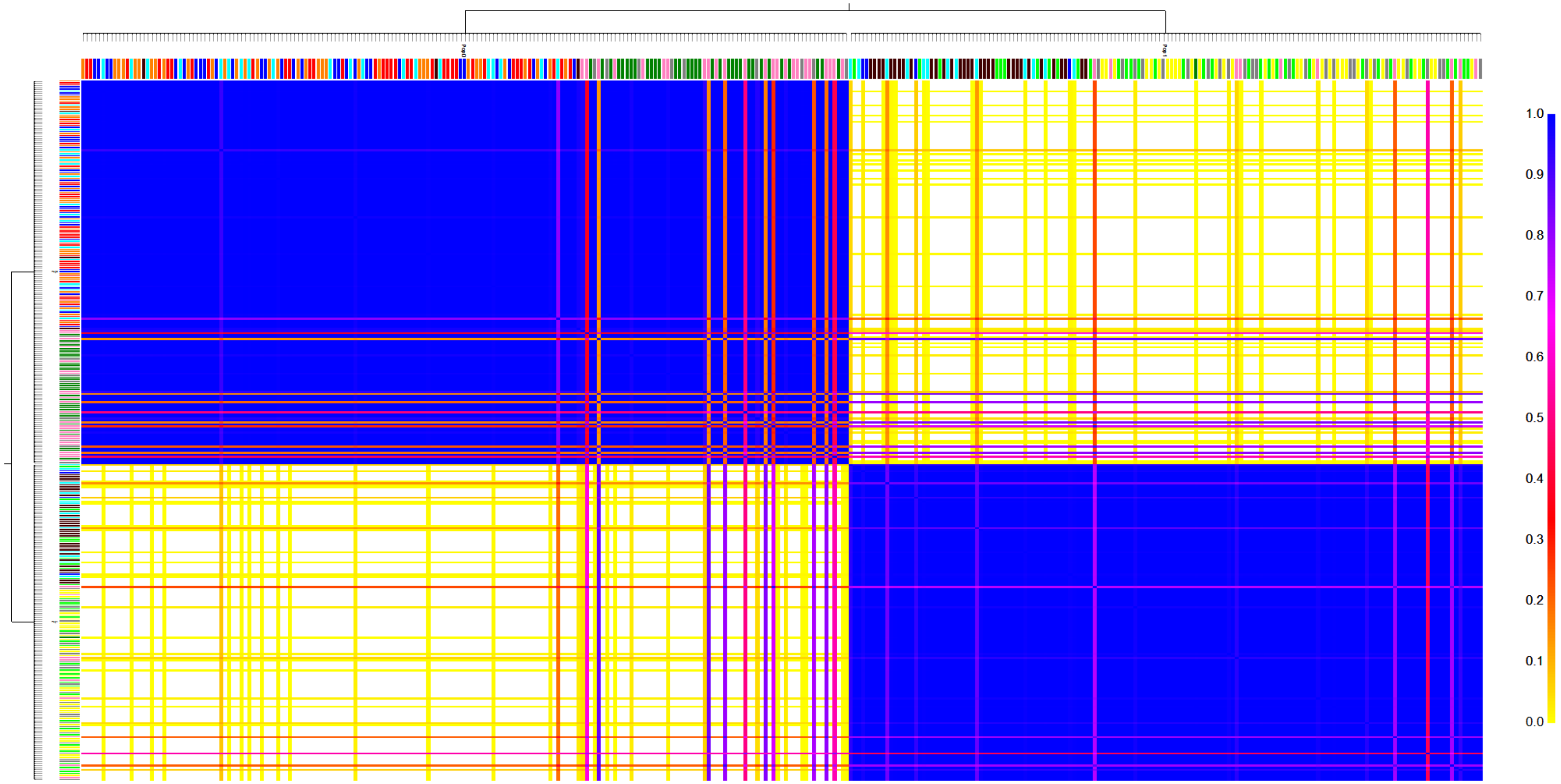


Figure S5 The probability matrix for the assignment of individuals into 2 populations. Note that most of the individuals labelled in pink could also be assigned into the right most populations.

Figure S6

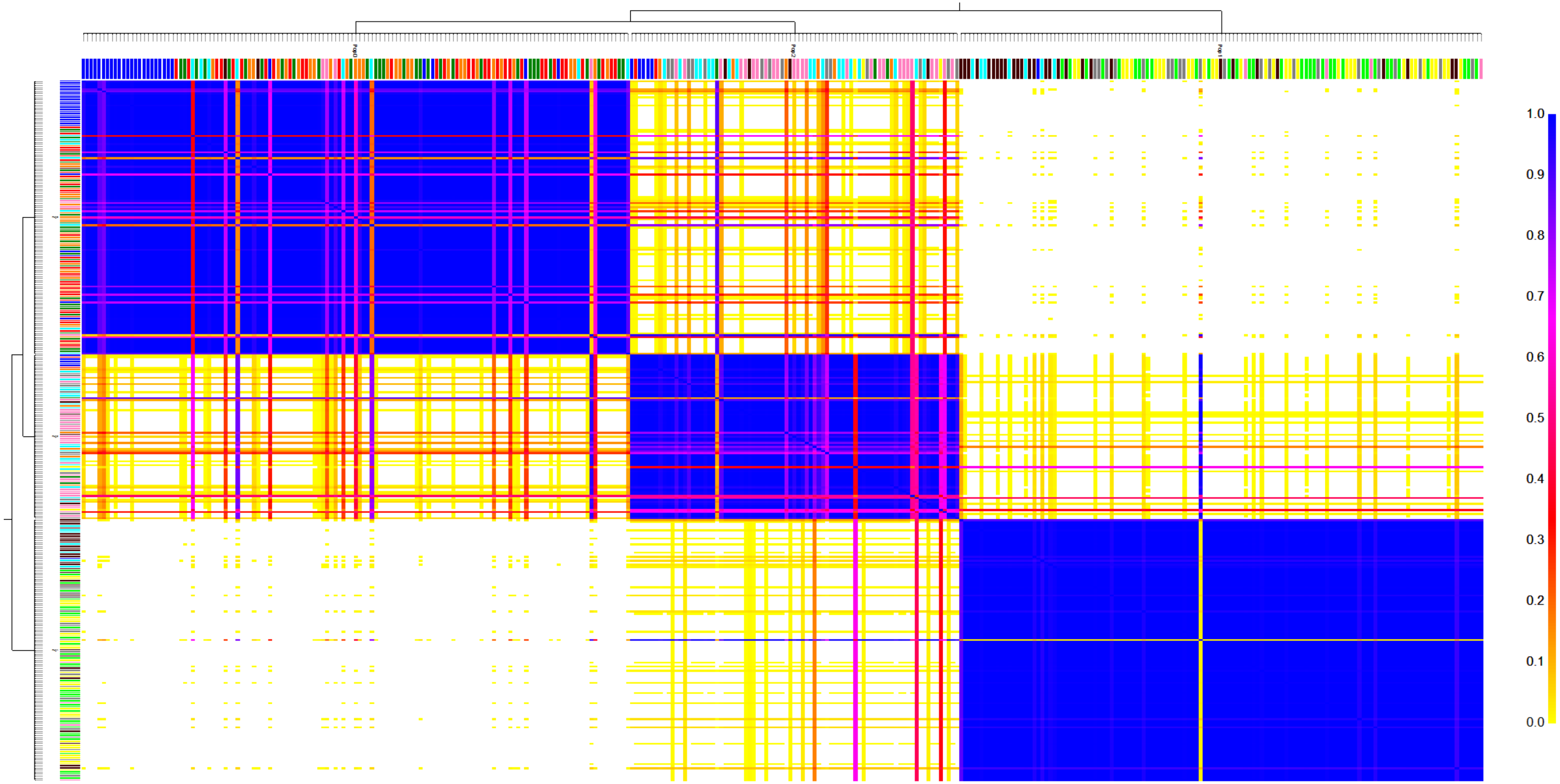


Figure S6 The probability matrix for the assignment of individuals into 3 populations.

Figure S7

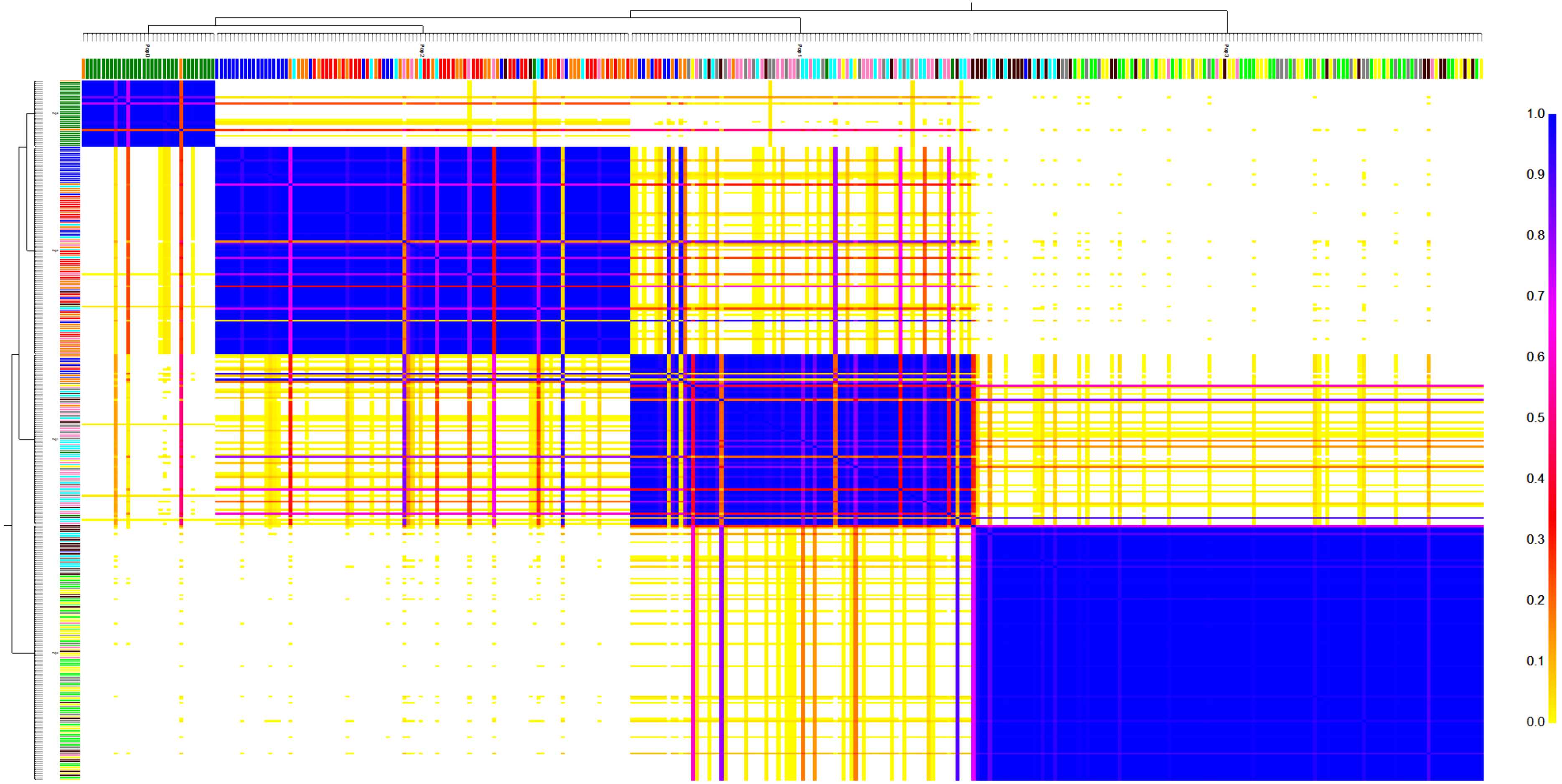


Figure S7 The probability matrix for the assignment of individuals into 4 populations.

Figure S8

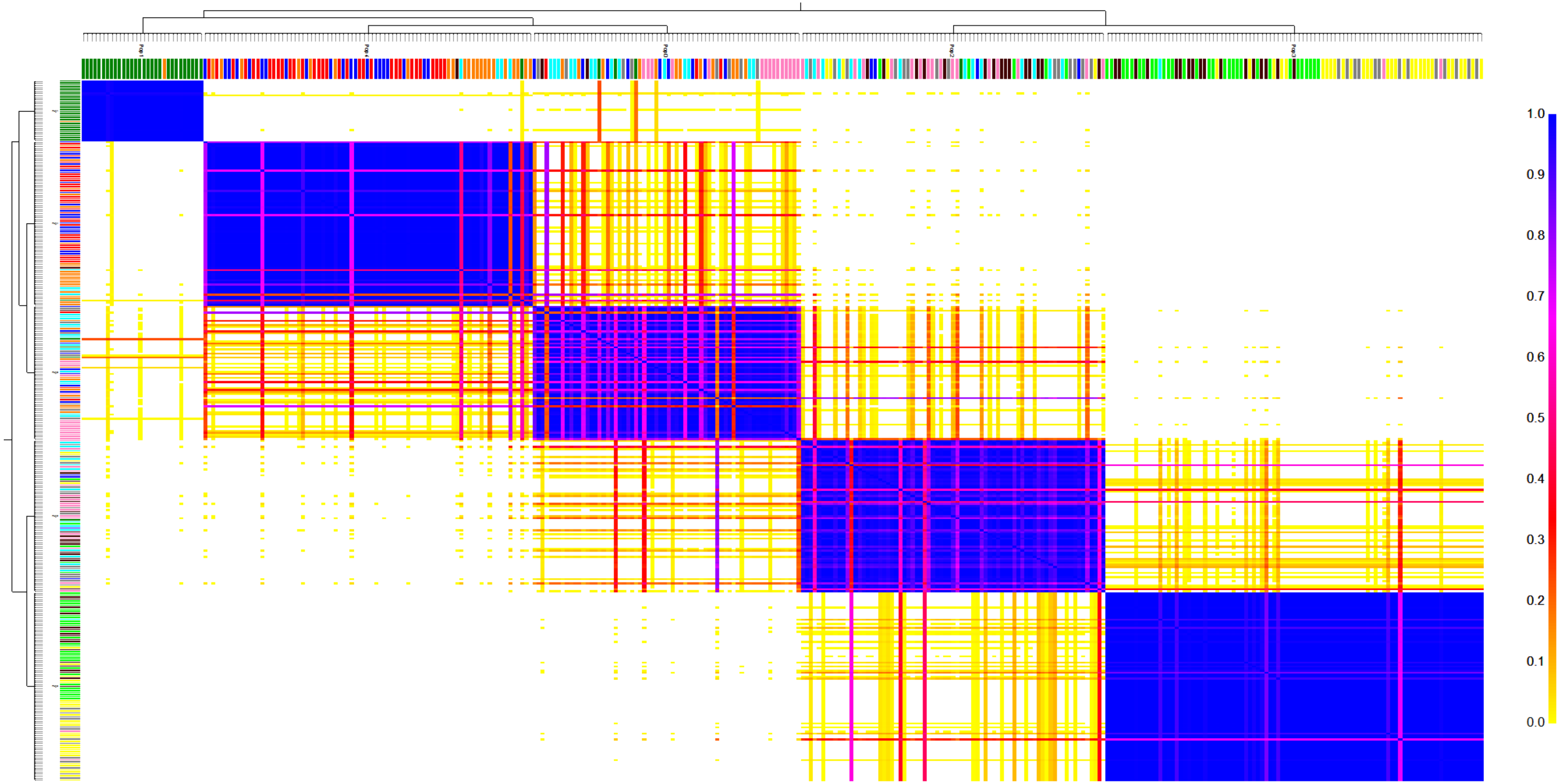


Figure S8 The probability matrix for the assignment of individuals into 5 populations.

Figure S9

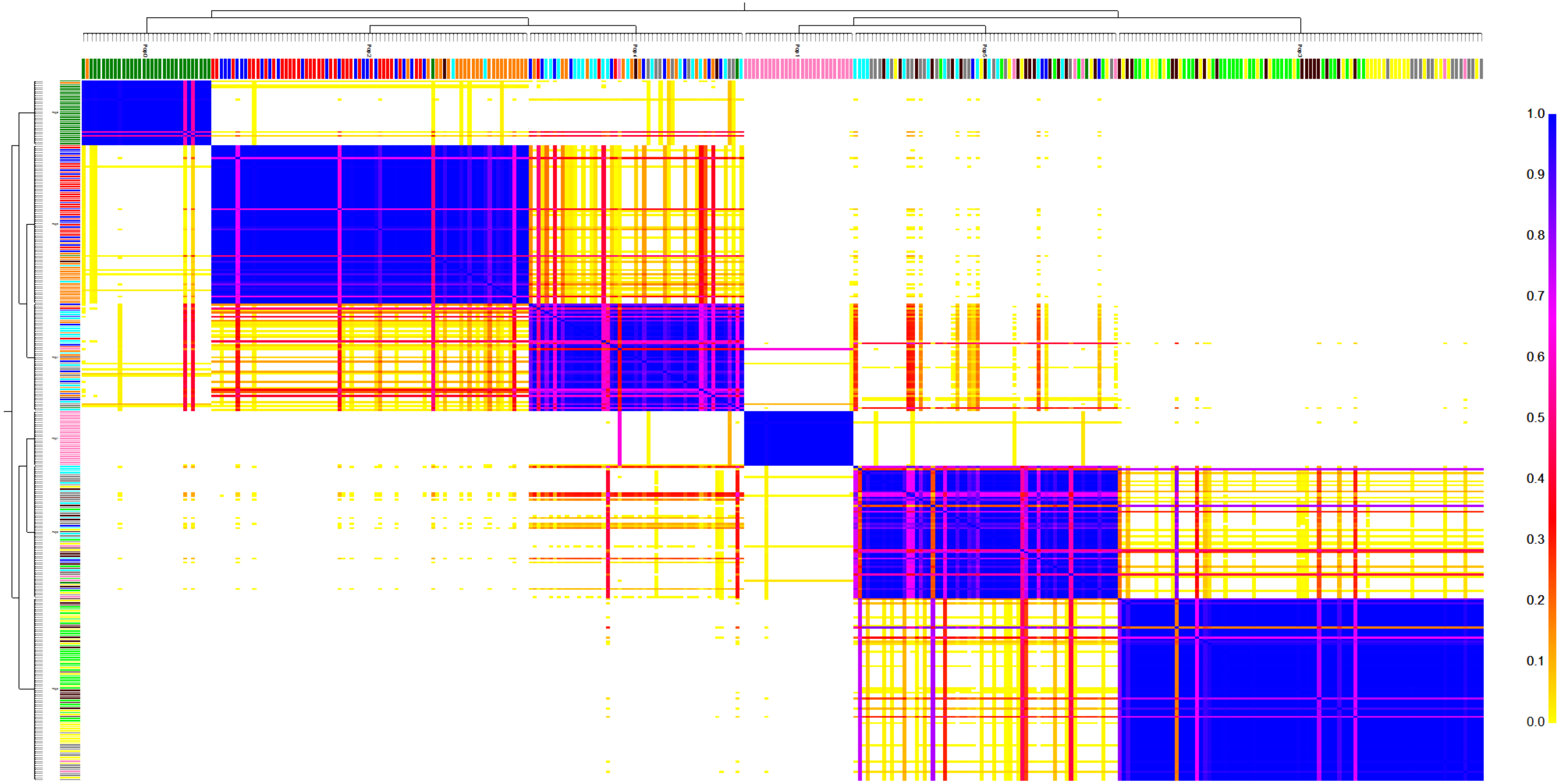


Figure S9 The probability matrix for the assignment of individuals into 6 populations.

Figure S10

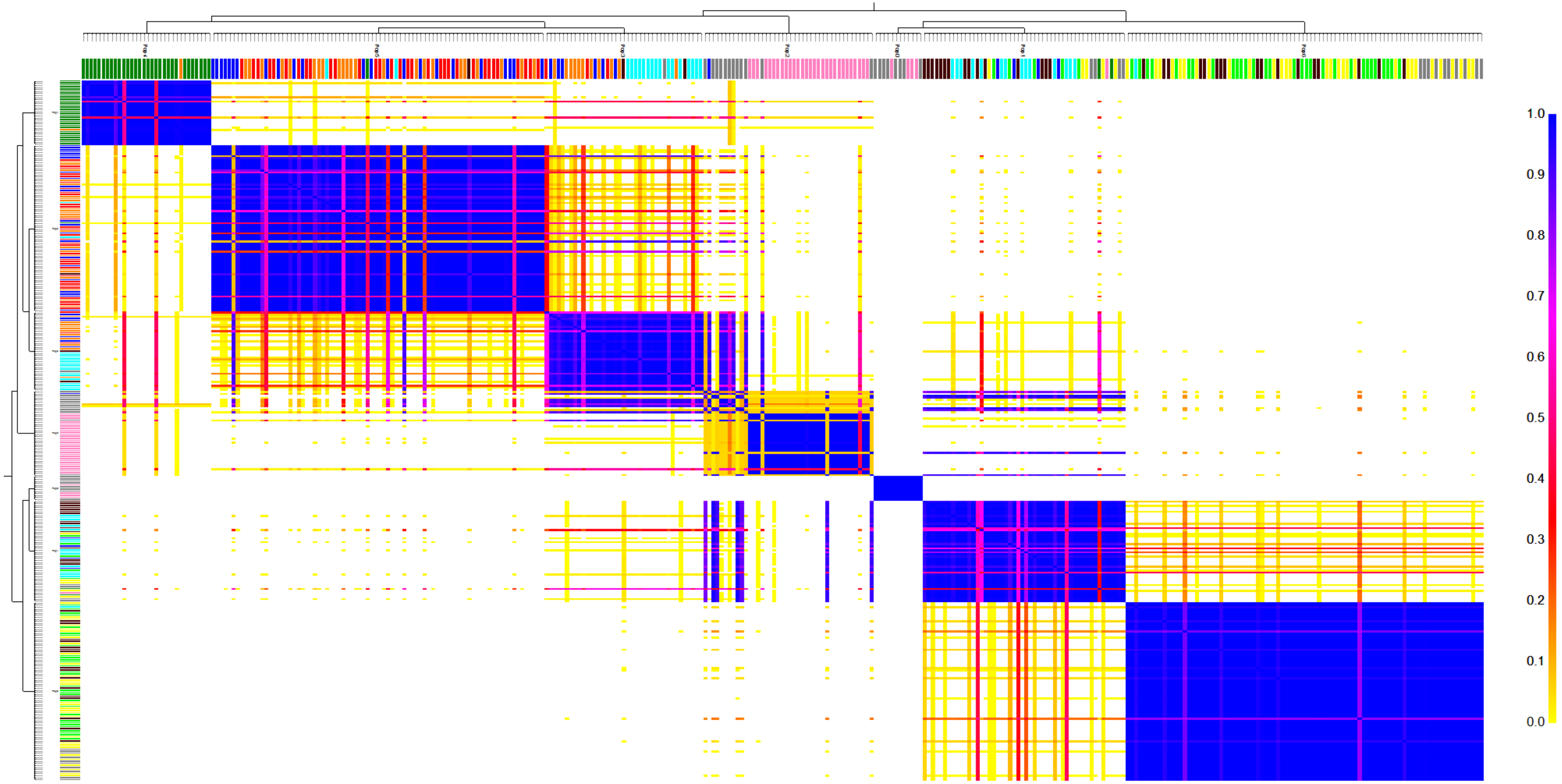


Figure S10 The probability matrix for the assignment of individuals into 7 populations.

Figure S11

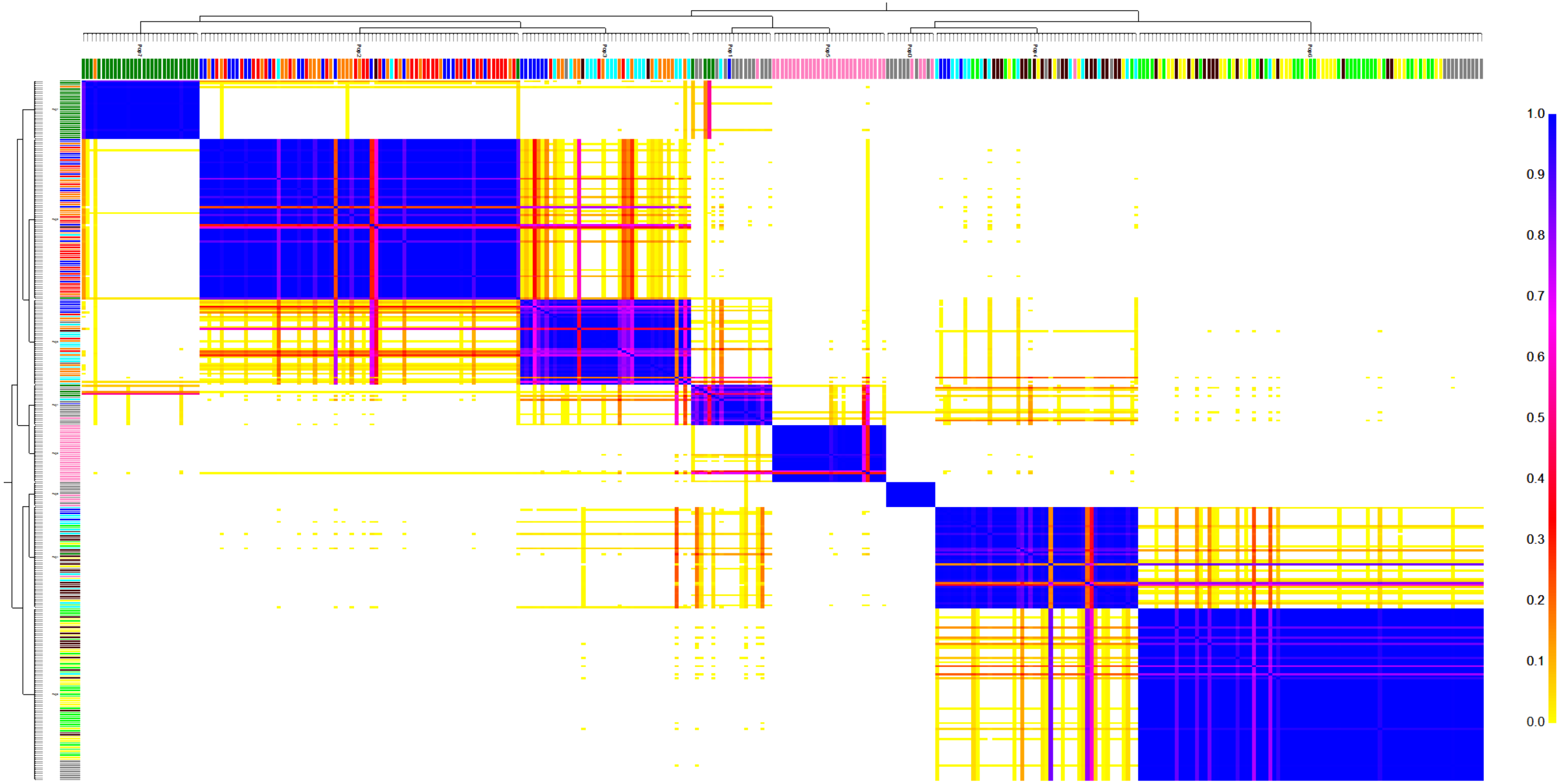


Figure S11 The probability matrix for the assignment of individuals into 8 populations.

Figure S12

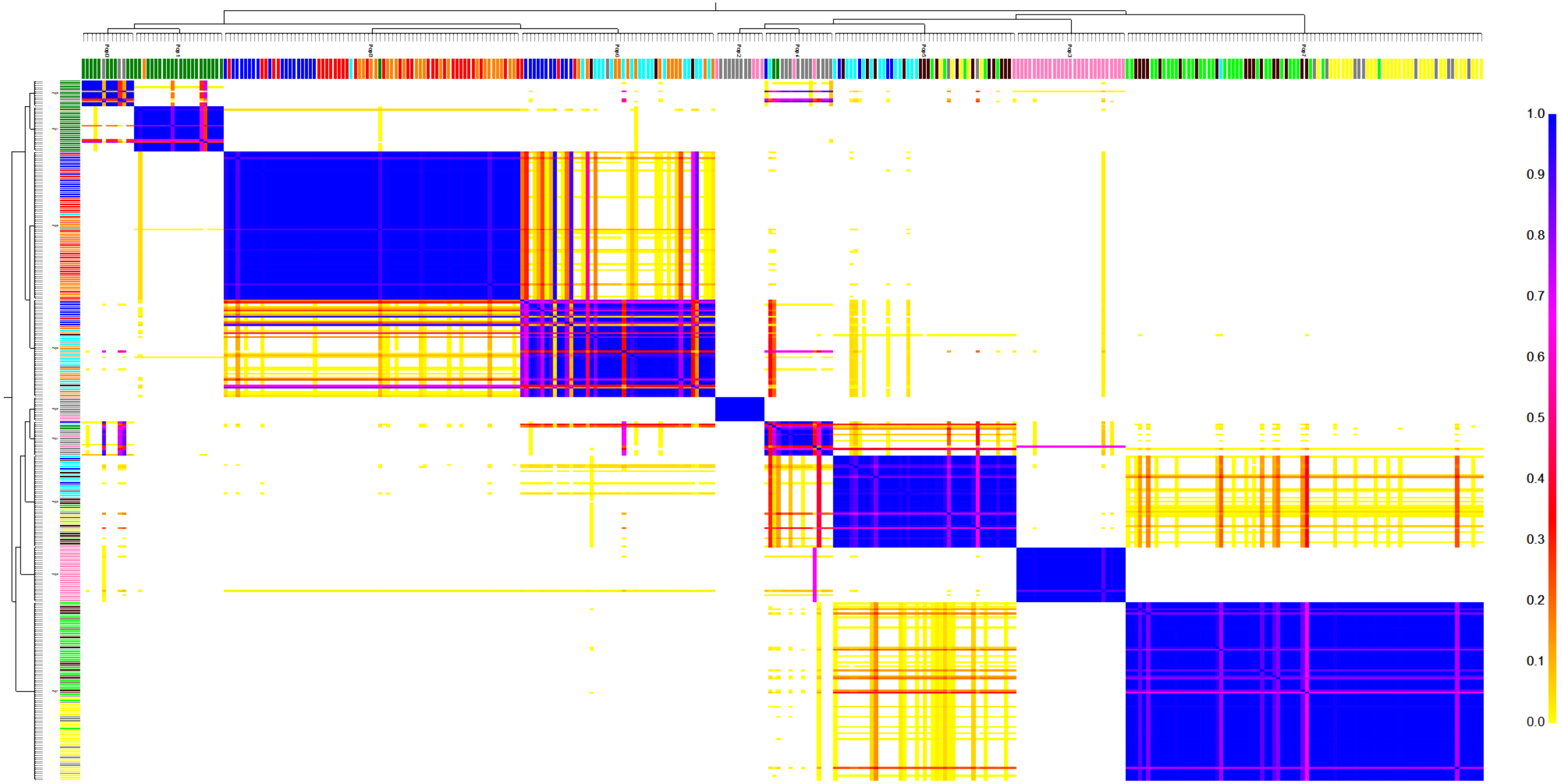


Figure S12 The probability matrix for the assignment of individuals into 9 populations.

Figure S13

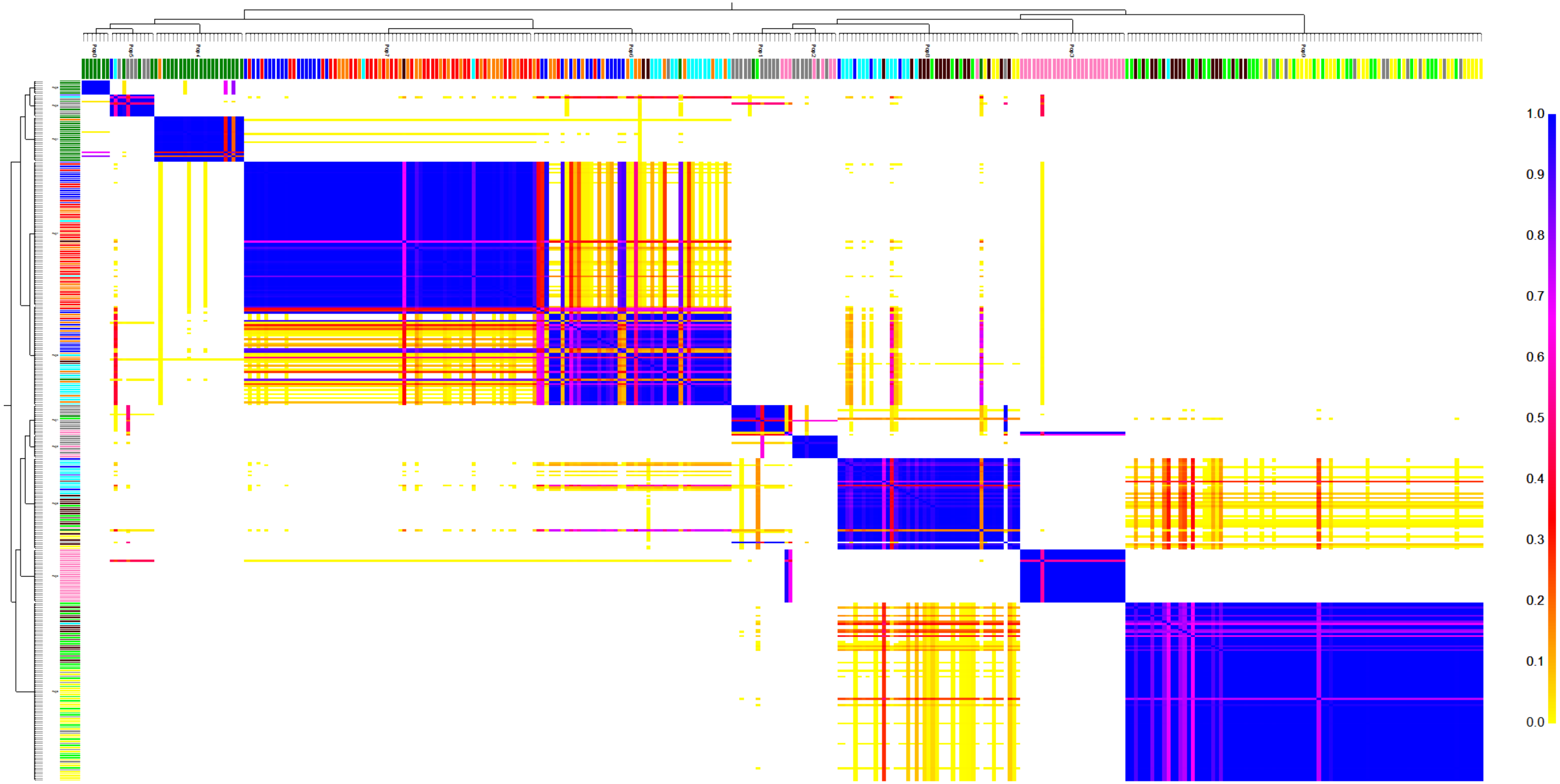


Figure S13 The probability matrix for the assignment of individuals into 10 populations.

Figure S14

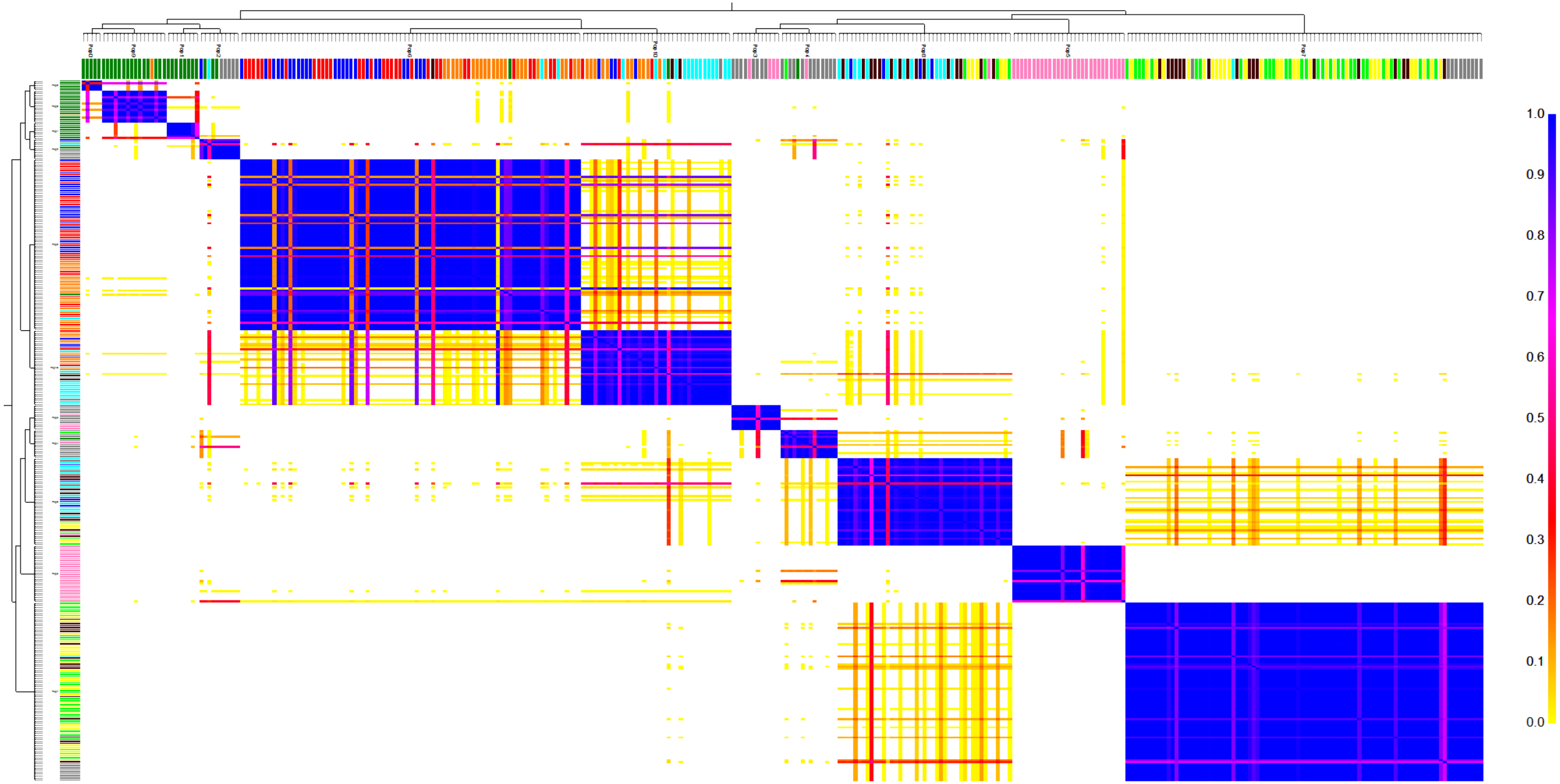


Figure S14 The probability matrix for the assignment of individuals into 11 populations.

Figure S15

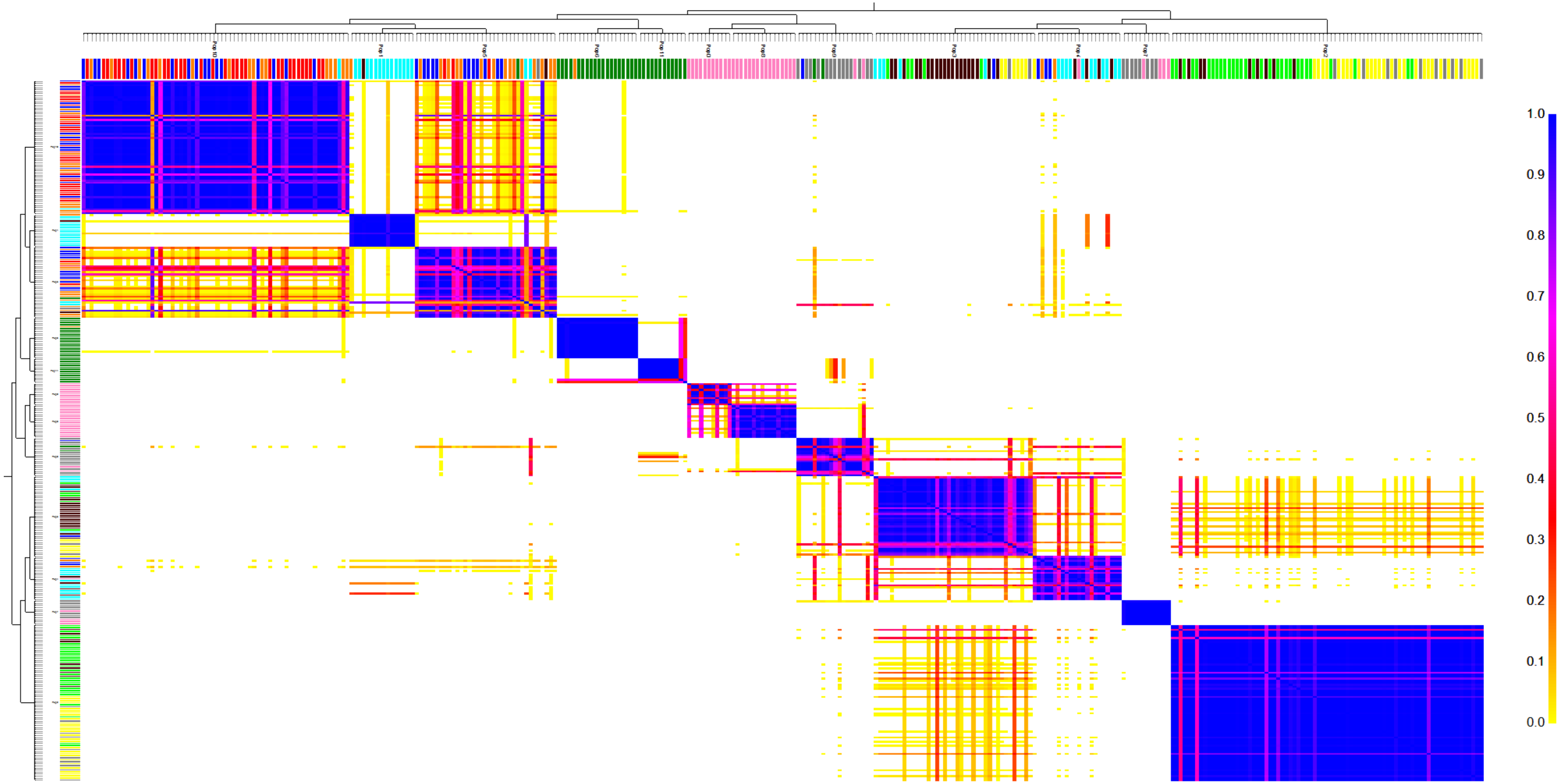


Figure S15 The probability matrix for the assignment of individuals into 12 populations.

Figure S16

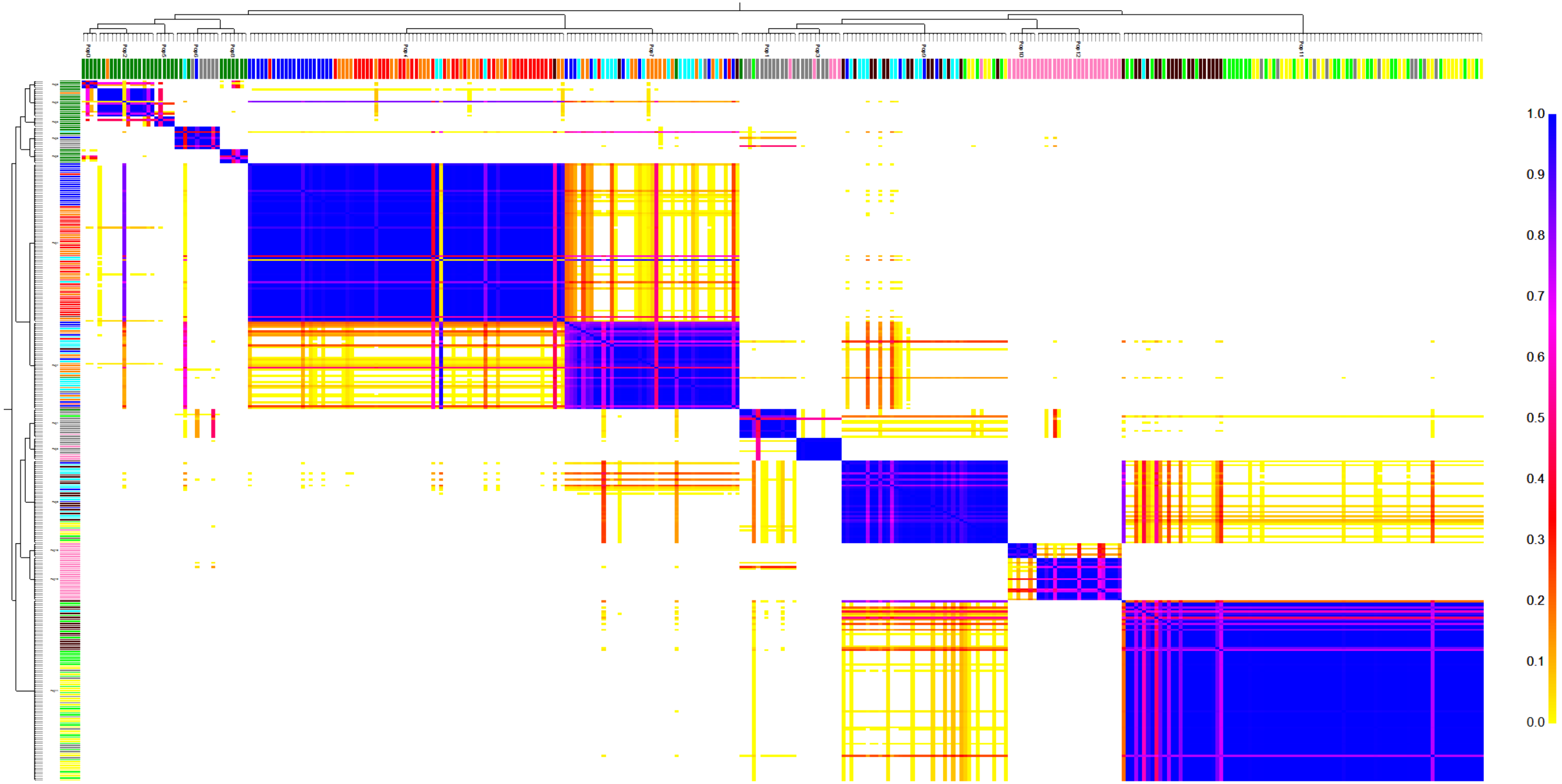


Figure S16 The probability matrix for the assignment of individuals into 13 populations.

Figure S17

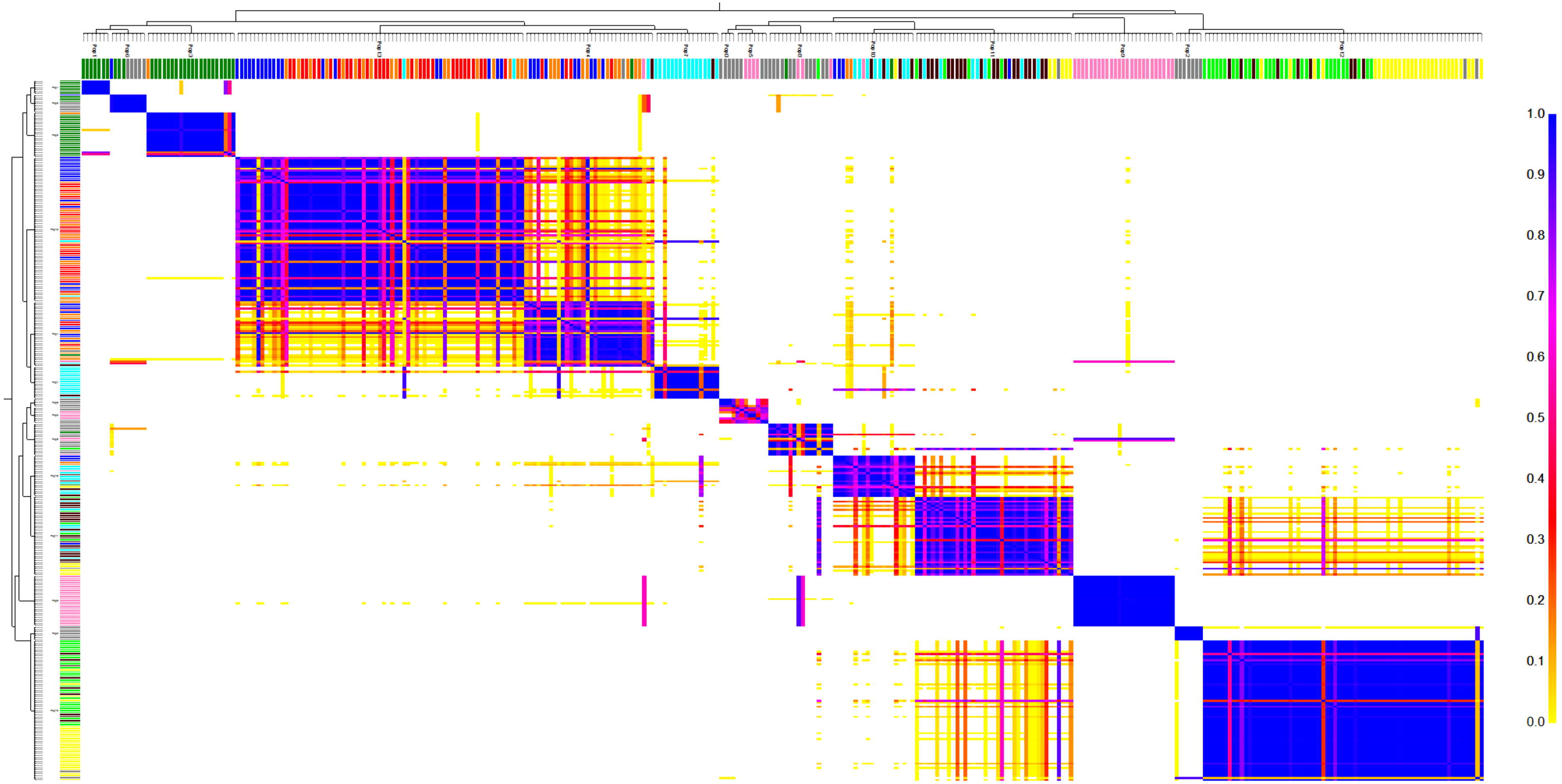


Figure S172 The probability matrix for the assignment of individuals into 14 populations.

Figure S18

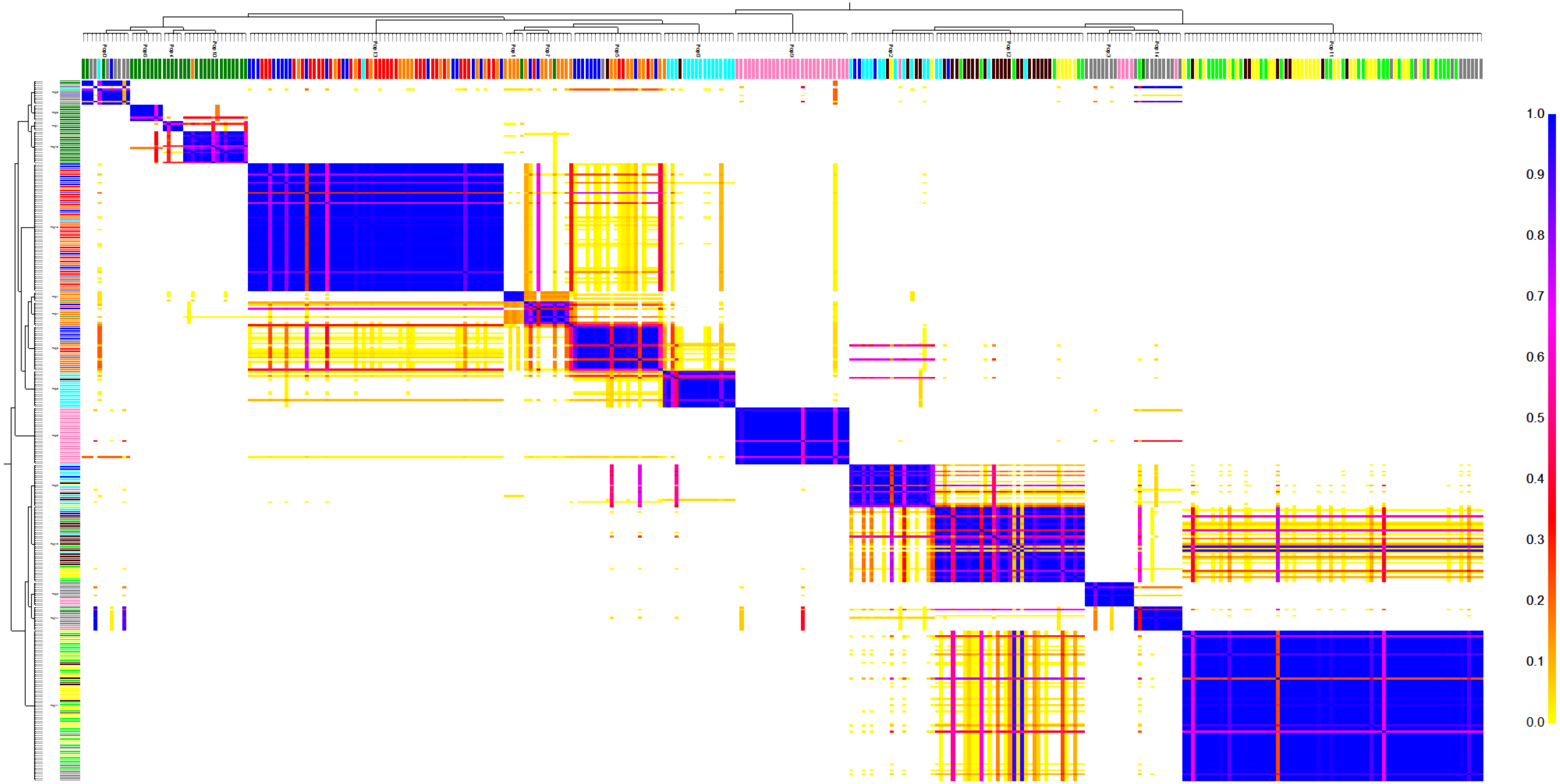


Figure S18 The probability matrix for the assignment of individuals into 15 populations.

Figure S19

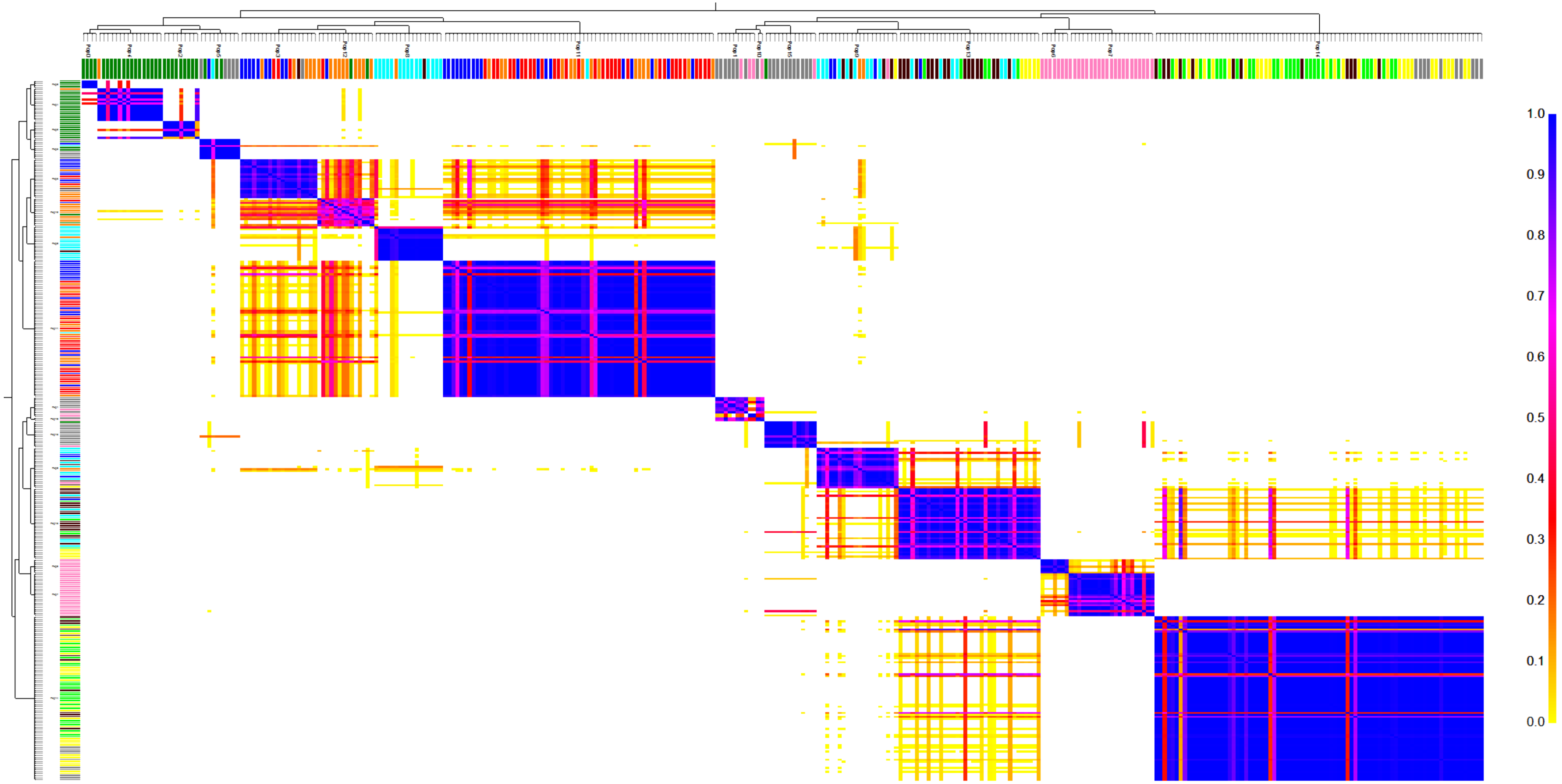


Figure S19 The probability matrix for the assignment of individuals into 16 populations.

Figure S20

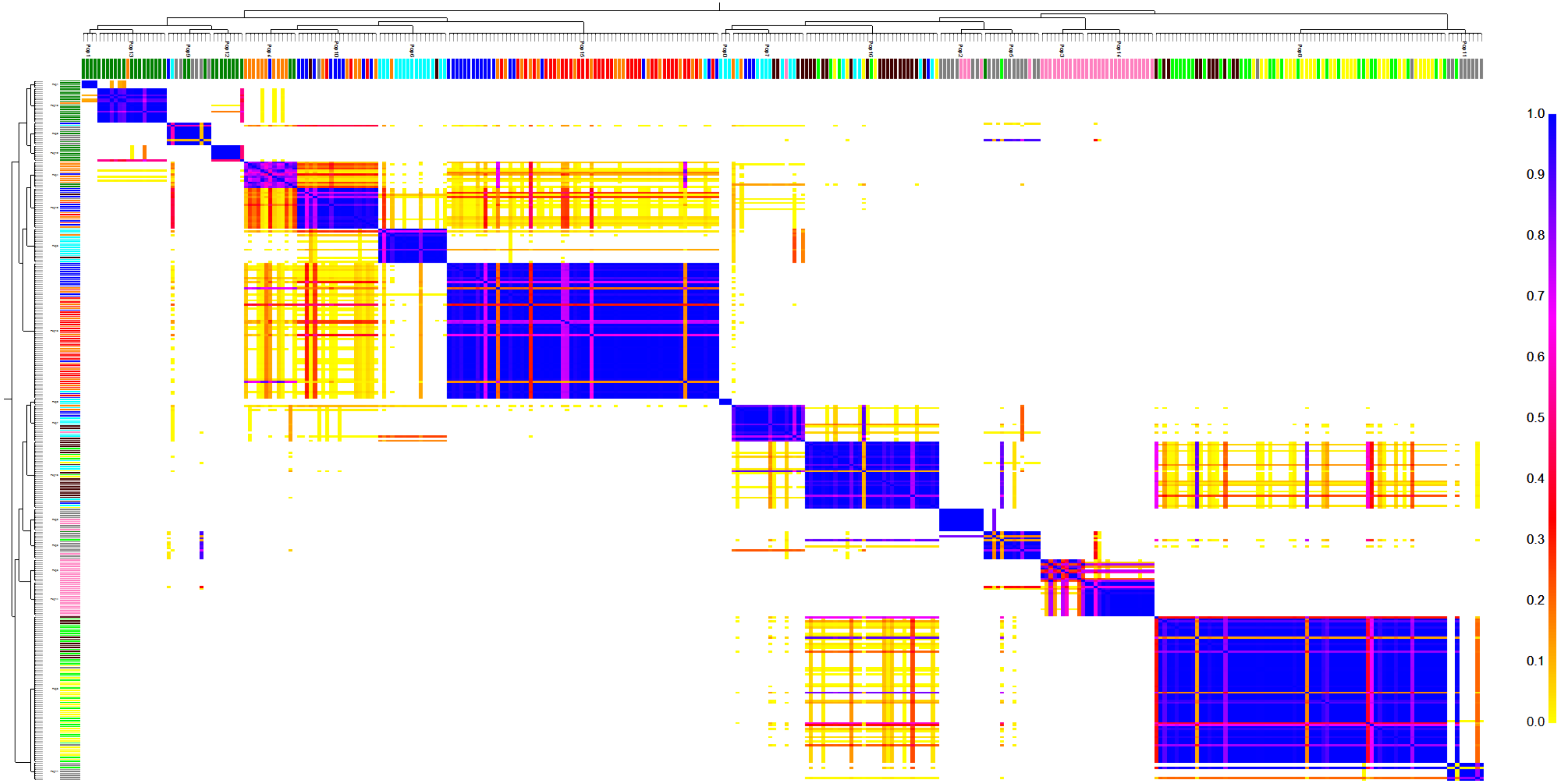


Figure S20 The probability matrix for the assignment of individuals into 17 populations.

Figure S21

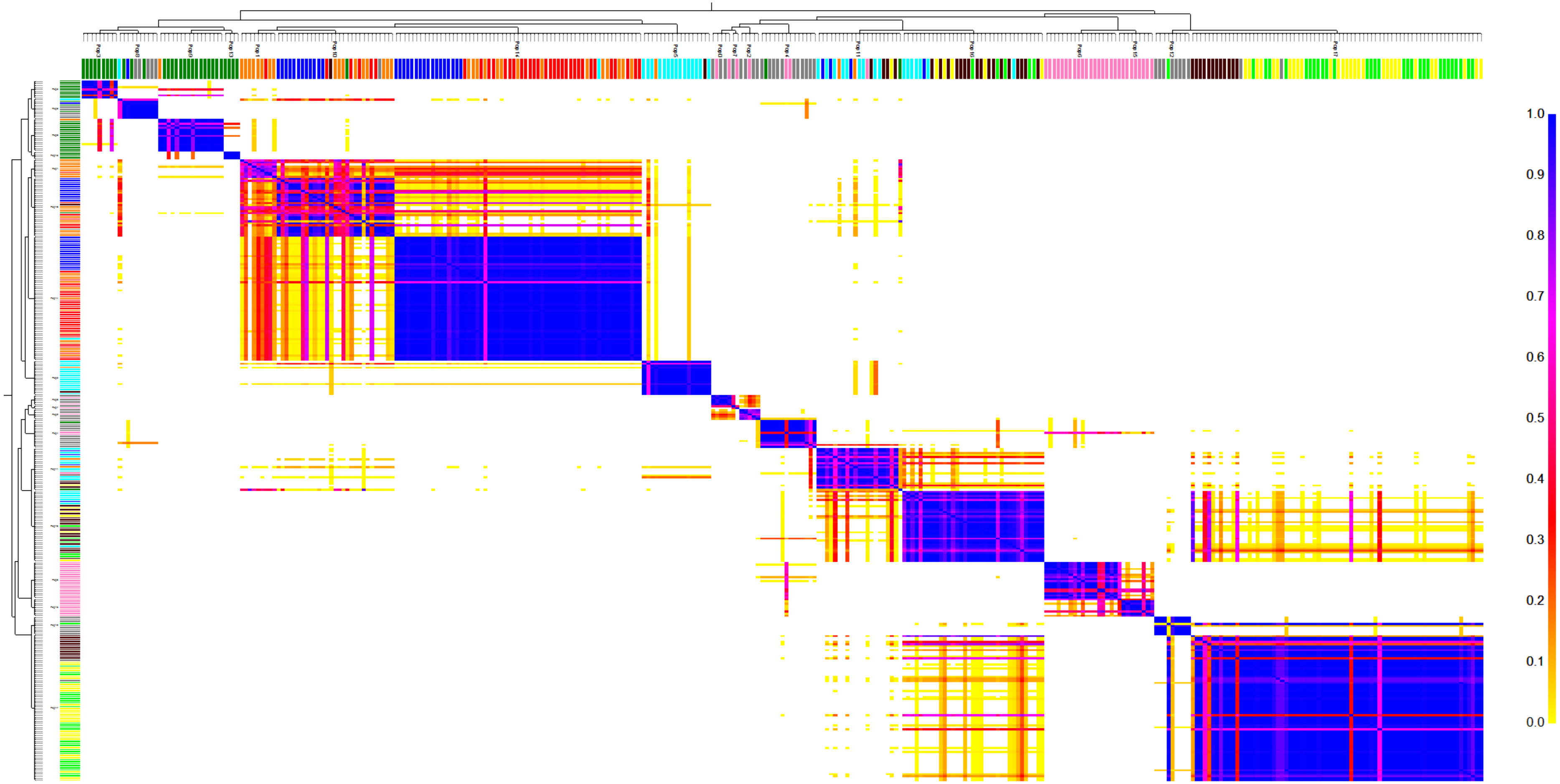


Figure S21 The probability matrix for the assignment of individuals into 18 populations.

Figure S22

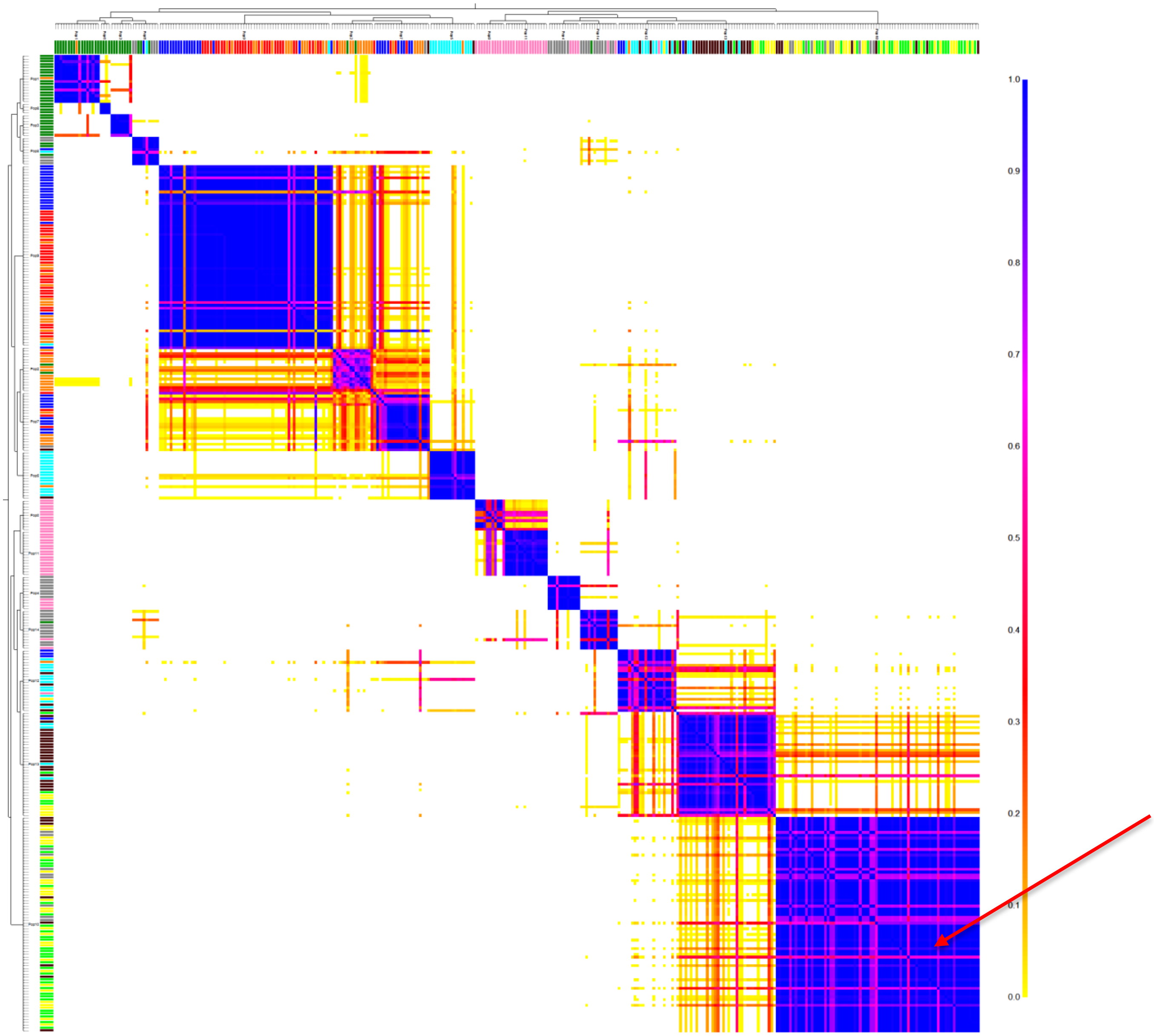


Figure S22 Comparison of the probability matrices of MCMC iterations. The lower triangle is the result of 100,000 iterations and the upper triangle is the result of 1,000,000 iterations. The differences of the analyses are very subtle which indicates that the 100,000 iterations are enough. The red arrow points one of the small differences in individual probability.