Karolinska Institutet

Thesis for doctoral degree (Ph.D.) 2008

POPULATION GENETIC, ASSOCIATION AND ZYGOSITY TESTING ON PREAMPLIFIED DNA

Ulf Hannelius

Thesis for doctoral degree (Ph.D.)
2008

POPULATION GENETIC, ASSOCIATION AND ZYGOSITY TESTING ON PREAMPLIFIED DNA

Ulf Hannelius

Karolinska Institutet

From the Department of Nutrition and Biosciences
Karolinska Institutet, Stockholm, Sweden

POPULATION GENETIC, ASSOCIATION AND ZYGOSITY TESTING ON PREAMPLIFIED DNA

Ulf Hannelius

Karolinska Institutet

Stockholm 2008

Published by Karolinska Institutet

ISBN 978-91-7409-062-8

Printed by

www.reproprint.se
Gårdsvägen 4, 169 70 Solna

Beware the man of one book. -St Thomas Aquinas

ABSTRACT

New advances in gen etic epidemiological research have lead to the establishment of collaborative biobanks that can be used in large-scale population genetic and association studies. A less explored approach is to use the existing biological repositories, like the Swedish newborn screening registry (SNSR) for similar purposes. The SNSR encompasses about 3 million individual samples in the form of dried blood spots on filter paper. These samples represent all newborns in Sweden since 1975 and the repository grows by 100000 samples each year.

In this thesis I use improved primer extension preamplification (I-PEP-L) and multiple displacement amplification (MDA) to preamplify DNA from dried blood spots and other templates. The methods are used to 1) explore the population genetic substructure in Sweden and Finland, 2) test if variants in the neuropeptide S receptor 1 (*NPSR1*) are associated with an increased risk of respiratory distress syndrome (RDS), and 3) validate a panel of autosomal SNPs for zygosity testing and population genetics.

I show that up to 25-year-old dried blood spots can be used for genetic studies (Study 1) and that haplotypes in NPSR1 associate with an increased risk of RDS (Study 2). Zygosity testing based on 47 unlinked autosomal SNPs is robust and reliable in the presence of population substructure and missing data (Study 3). Historic connections with neighbouring countries, Central Europe and recent immigration in the big cities are evident in Sweden, as well as the presence of regional differences in genetic diversity (studies 4 and 5). The *CCR5* Δ32 variant associated with HIV immunity is more common in the northern parts of Sweden (Study 1). The Finnish substructure is characteristic of an east-west duality congruent with historic political and anthropological borders, and the Swedish speaking part of Ostrobothnia clusters with Sweden (Study 5).

In conclusion, I demonstrate how preamplification can assist in gaining access to samples that would otherwise be incompatible with genetic epidemiological research. Also, due to the possibility of allele dropouts and missing data, quality control should be of very high priority when using these methods.

LIST OF PUBLICATIONS

I. **Hannelius U**, Lindgren CM, Melén E, Malmberg A, von Döbeln U, Kere J. Phenylketonuria screening registry as a resource for population genetic studies. Journal of Medical Genetics, 2005 Oct;42(10):e60

II. Pulkkinen V, Haataja R, **Hannelius U**, Helve O, Pitkänen OM, Karikoski R, Rehn M, Marttila R, Lindgren CM, Hästbacka J, Andersson S, Kere J, Hallman M, Laitinen T. G protein-coupled receptor for asthma susceptibility associates with respiratory distress syndrome. Annals of Medicine, 2006;38(5):357-66.

III. **Hannelius U**, Gherman L, Mäkelä VV, Lindstedt A, Zucchelli M, Lagerberg C, Tybring G, Kere J, Lindgren CM. Large-scale zygosity testing using single nucleotide polymorphisms. Twin Research and Human Genetics, 2007 Aug;10(4):604-25.

IV. Lappalainen T, **Hannelius U**, Salmela E, von Döbeln U, Lindgren CM, Huoponen K, Savontaus M-L, Kere J, Lahermo P. Population structure in Sweden – A Y-chromosomal and mitochondrial DNA analysis. *Submitted*.

V. **Hannelius U**, Salmela S, Lappalainen T, Guillot G, von Döbeln U, Lindgren CM, Lahermo P, Kere J. Population Genetic substructure in Finland and Sweden revealed by a small number of autosomal unlinked SNPs. *Submitted*.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ASPM	Abnormal Spindle-like Microcephaly
BPD	Bronchopulmonary dysplasia
CD/CV	Common disease/common variant
CV/MD	Common variant/multiple disease
CNV	Copy number variation
DOP	Degenerate oligonucleotide PCR
DNA	Deoxyribonucleic acid
DZ	Dizygotic
DBS	Dried blood spot
GWA	Genome wide association
HWE	Hardy-Weinberg equilibrium
HVS	Hyper-variable segment
I-PEP-L	Improved primer extension preamplification long
LCT	Lactase
LD	Linkage disequilibrium
MCPH1	Microcephalin
mtDNA	Mitochondrial DNA
MZ	Monozygotic
MRCA	Most recent common ancestor
MDA	Multiple displacement amplification
NF	Neurofibromatosis
NPSR1	Neuropeptide S receptor 1
PKU	Phenyl ketonuria
PCR	Polymerase chain reaction
PMRCA	Position of most recent common ancestor
PEP	Primer extension preamplification
PCA	Principal components analysis
RDS	Respiratory distress syndrome
RFLP	Restriction fragment length polymorphism
STR	Short tandem repeat
SNP	Single nucleotide polymorphism
TMRCA	Time to most recent common ancestor
TDT	Transmission disequilibrium test
WTCCC	Wellcome Trust Case Control Consortium
WGA	Whole genome amplification

1 BACKGROUND

Only a few years ago in 2001 the first working draft of the three billion base pair long human genome was published as a joint effort between the international Human Genome Project and Celera Genomics in the journals of Nature and Science (Lander et al. 2001; Venter et al. 2001). The final phase took 3-4 years to complete and cost about $300 million. Quite remarkably, this same effort will be achieved in a matter of hours and to a cost of less than $1000 in a foreseeable future (von Bubnoff 2008). Geneticists are already interrogating up to a million genetic variants at a time in thousands of individuals. In the span of only two years, these so called genome-wide association (GWA) studies have resulted in the identification and replication of nearly 100 locations in the genome that influence the risk of developing one of 40 common diseases (Pearson and Manolio 2008).

In 2003, when the initial draft of the genome was refined and published as a nearly complete version (Consortium 2004), researchers could finally estimate with some accuracy the number of protein-coding genes in the genome. An informal bet on this issue was taken in 2000 with estimates ranging from 26000 to over 100000 (Pennisi 2007). The winning bet went to the lowest count of 26000, but today we know that the correct number is maybe even lower than this, namely in the range of 20500 (Clamp et al. 2007). Recent reports are also providing evidence that copy number variations (CNVs) rather than single nucleotide polymorphisms (SNPs) represent the most common source of inter-individual variation in the genome (Redon et al. 2006). Even monozygotic twins (MZ) manifest differences in CNVs, questioning the prevailing belief that identical twins actually are genetically identical (Bruder et al. 2008).

The above-mentioned examples highlight the pace at which genetics and related fields are evolving. These advances create an increasing demand to identify large numbers of phenotypically well-characterized DNA that can be used for genetic studies.

In this thesis, I develop and validate a methodological approach based on preamplification of DNA for using extant large-scale biological repositories for genetic epidemiological studies, including population genetics in Sweden and Finland, association studies and zygsoity testing.

1.1 THE HUMAN GENOMES

The nuclear and the mitochondrial genomes encompass the human hereditary information encoded by deoxyribonucleic acid (DNA). Most genetic research is focused on the nuclear genome, but the much smaller mitochondrial genome plays a very central role in population genetics and forensic medicine where it is used as a marker for human migration and identification. The epigenome, including heritable variations in methylation patterns, histones and chromatin, is not reviewed here but current work is ongoing to characterize this variation and future research will employ tools that consider all genomes jointly (Bird 2007).

1.1.1 The nuclear genome

Approximately three billion pairs of the nucleotides Adenine, Thymine, Guanine and Cytosine represent the human nuclear genome. Stretches of nucleotides form double-stranded DNA molecules that in turn are bound to proteins to form 23 distinct organized structures called chromosomes. Each individual cell has 46 such chromosomes, two copies of the autosomal chromosomes 1-22 plus two X-chromosomes (females) or an X and a Y-chromosome (males). The chromosomes are situated in the nucleus of the cell and are inherited form the parents (Figure 1a). The father and mother each contribute one copy of the autosomal chromosomes; the father contributes either an X or a Y chromosome and the mother an X chromosome.

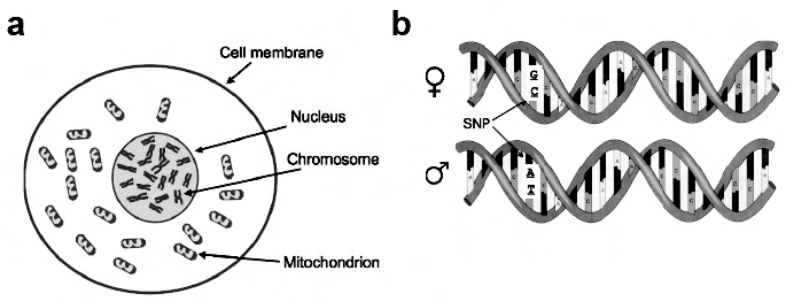

Figure 1. **The genomes and DNA. a** | The nuclear genome consists of 23 distinct chromosomes that reside within the nucleus while the mitochondrial genome is situated within the mitochondria. **b** | A schematic presentation of homologous maternal and paternal DNA sequences. The arrows indicate a SNP, the most common variation in the genome.

The complete genome consists of approximately 20500 protein-coding genes and numerous other regulatory DNA sequences. Each alternative form of these sequences is generally referred to as alleles, and the prevalence of a specific variant is referred to as the allele frequency.

1.1.1.1 Mutation and variation

A major part of the genetic variation between individuals is the result of mutation. Every time a cell divides to form a new daughter cell (mitosis), the DNA must first be replicated. Neither the machinery that catalyses this reaction nor the internal system that is responsible for repairing damage caused by mutagenic chemicals or radiation are foolproof, and the average mutation rate has been estimated to be around 2.5×10^{-8} per base pair and generation (Kumar and Subramanian 2002). There is considerable variation, and one of the most rapidly mutating human genes, Neurofibromatosis 1 (*NF1*), has a mutation rate near 10^{-4} (Stephens et al. 1992). If a mutation takes place in the germ cells (sperms or oocytes) it can be transferred to the following generations and become an inherent part of the variation in the genome.

The earliest human genome linkage maps were based on this variation in the form of single nucleotide polymorphisms (SNPs) and copy-number variations (CNVs), assayed by restriction fragment length polymorphism (RFLP) analysis using southern blotting (Weissenbach et al. 1992). The next generation genetic maps were based solely on microsatellite markers, displaying variable lengths of simple sequence repeats (e.g. stretches of CACACA… or CTGCTGCTGCTG…). These markers were assayed using polymerase chain reaction (PCR) and gel or capillary separation (1992; Gyapay et al. 1994). Finally, the microarray methodology revolutionized the assaying of tens and hundreds of thousands of SNPs, and contemporary genome-wide studies of variation are based on these techniques.

Currently it is estimated that the largest part of variation between two individuals is due to approximately 10 million single nucleotides SNPs (Figure 1b). A single change of a nucleotide can lead to a change in the amino acid composition of a protein. These SNPs are referred to as non-synonymous, while single base substitutions that do not change the amino acid composition are called synonymous. In addition, SNPs may affect the regulatory, non-coding regions of genes thus modifying the gene expression. Due to their large numbers and proven significance in disease, SNPs are the most widely used markers in genetics. Still, with the advent of the microarray technology it has been shown that CNVs are more common than previously believed and they are resurfacing as major polymorphisms in the genome (Feuk et al. 2006).

1.1.1.2 Recombination and haplotypes

Another major part of variation in the nuclear genome is created by recombination. Every time a sperm or an oocyte is formed (meiosis) pieces of DNA are shuffled between the homologous maternal and paternal chromosomes, forming rearranged chromosomes that are then transferred to the offspring. This ensures that the parental DNA sequences are not inherited in their original form but as sequences that encompass parts from both the mother and the father. The closer two alleles are to each other on a chromosome, the tighter they are linked and the less likely it is that they will be separated due to these shuffling events. As a result, this correlation between alleles (linkage disequilibrium; LD) gradually decays over generations. This in turn results in shorter stretches of associated alleles called haplotypes that share a common ancestral

chromosome (Daly et al. 2001; Reich et al. 2001; Gabriel et al. 2002). As an example, short haplotypes are seen in the African populations while longer haplotypes are evident in younger populations like Europeans. A family can be used as an example of a very young population where the haplotypes are long due to the few generations that separate the individuals (Figure 2).

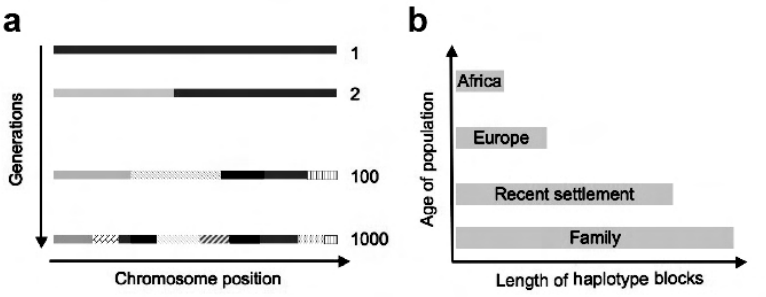

Figure 2. **Recombination and haplotypes. a** | A schematic representation of how recombination changes the haplotype background in a population. **b** | The length of the haplotypes decrease with increasing age of the population.

1.1.1.3 The sex chromosomes

The sex-specific X and Y-chromosomes represent special cases in the nuclear genome. There are approximately 750 genes on the X-chromosome, while the Y-chromosome codes for only 46 unique proteins, many of which are related to sex or fertility. Over the course of history, the X and the Y-chromosomes have almost totally lost the ability to recombine with each other; the two X-chromosomes in females recombine normally, but the single Y-chromosome in males is inherited virtually intact from father to son. Only 5% of the Y-chromosome, the pseudoautosomal regions, exchanges material with the X-chromosome during meiosis.

The lack of recombination in the remaining 95% makes it possible to follow the inheritance of the Y-chromosome from father to son in a population since the only variation present is due to mutation. Based on the variation, phylogenetic trees are constructed consisting of haplogroups that are defined by specific SNPs. If the more mutating microsatellites are used the results are grouped into haplotypes. Haplogroups can be used to trace the original human ancestors while haplotypes have better resolution for genealogical use.

The mutation rate in the Y-chromosome has been estimated to be up to 4.8 times higher than in other nuclear chromosomes (Lindblad-Toh et al. 2005). These mutations also segregate quickly due to genetic drift (random fluctuations, see chapter 1.2.1.2) since there are only 1/4 as many Y-chromosomes in a population compared to autosomal chromosomes and 1/3 as many compared to X-chromosomes. This, and the fact that

the human Y-chromosome compared to other mammals is smaller and less gene-rich, has been used as an argument for the Y-chromosome slowly disappearing (Aitken and Marshall Graves 2002; Graves 2006).

1.1.2 The mitochondrial genome

While the Y-chromosome is specific to the paternal lineage, the mitochondrial genome is inherited exclusively from the mother. Every cell has 10-10000 copies of this genome, compared to only two copies of the nuclear genome. The circular 16569 base pair long genome codes for only 37 genes but accumulating evidence for its implications in neurological disorders, aging and cancer is building (Taylor and Turnbull 2005; Krishnan et al. 2007; Schapira 2008). There are no recombining events taking place between generations, but the rate of mutation is higher than in the autosomes, especially the hyper-variable segments, HVS-I and HVS-II, located in the untranslated control region.

Because of the extensive variation between individuals, mitochondrial DNA (mtDNA) has a widespread use in forensics and anthropology. The multiple copies of the mitochondrial genome in the cells make it easier to extract large amounts of DNA of good enough quality for genotyping or sequencing even from ancient samples (Gilbert et al. 2007). Similar to the non-recombining part of the Y-chromosome, the lack of recombination in the mitochondrial genome facilitates hierarchical grouping of variants into phylogenetic trees based on haplogroups.

1.2 POPULATION GENETICS

It has been estimated that the modern human, *Homo sapiens*, has a single recent origin in East Africa some 150000 years ago. Genetic data suggest that this population expanded out of Africa about 50000 years ago (Reich et al. 2001), but it is debated how the expansion took place, and whether this population replaced or interbred with the other human lineages that coexisted during these times (Stringer 2002; Macaulay et al. 2005; Ramachandran et al. 2005; Ray et al. 2005; Soficaru et al. 2006; Templeton 2007; Li et al. 2008). As the original population has grown and migrated to new areas of the globe, it has been subject to disease, famine, war and other events that have shaped the cultural and genetic characteristics of the 6.5 billion people that inhabit the Earth today. By using the characteristic inheritance patterns and variation in the nuclear and mitochondrial genomes, it is possible to explore both recent and ancient human evolution, thus identifying events that have formed the genetic architecture of disease and other human traits.

1.2.1 Concepts

1.2.1.1 Selection

Under certain circumstances an individual may gain a reproductive advantage over other individuals and the underlying traits tend to undergo positive selection and grow in frequency in the population. Likewise, if the traits are deleterious they may be negatively selected and decrease in frequency over the following generations. Recent work show that in humans, positive selection seems to facilitate regional adaptation of subpopulations by targeting regulatory and amino acid-altering variants in gene regions. Negative selection is a more global phenomenon, ensuring that damaging variants do not reach high frequencies in the overall human population. In general, genes that show signs of either positive or negative selection appear to be more frequently involved in disease than would be expected simply by chance (Barreiro et al. 2008).

1.2.1.2 Genetic drift and inbreeding

When a genetic variant is neutral, i.e, has no effect on the reproductive fitness of the individual, its prevalence in the population evolves by chance alone, so called genetic drift. If a population consists of a very small number of reproducing individuals, the influence of genetic drift becomes large. Variants that are normally targeted by positive selection may therefore totally disappear from one generation to the next and variants under negative selection may increase in prevalence. A small population also increases the frequency of inbreeding, further reducing the genetic variability in the population.

1.2.1.3 Population bottlenecks

Events that significantly reduce the number of reproducing individuals, like natural disasters, epidemics or war, are called population bottlenecks and result in decreased genetic variability. In humans, it has been suggested that a bottleneck event took place around the time of migration out of Africa about 50000 years ago (Reich et al. 2001) and more recent bottlenecks have taken place for example in medieval Europe during the plague epidemics (Stenseth et al. 2008).

1.2.1.4 Migration and isolation

Genetic drift is countered by the effect of migration, also called gene-flow. When individuals migrate and come across new populations they bring with them new genetic variants that contribute to increasing the genetic variation in the combined population. This assumes that individuals mate randomly, something that cannot be taken for granted due to cultural, religious and other differences that often determine how people choose their partners. In today's world people have the means to cross most geographical and political borders, and the amount of migration is consequently much larger than it was only a few decades ago. Historically though, migration has often been difficult and certain populations have because of this lived more or less isolated for longer periods of time. An example of this kind of a population is Iceland that was settled about 1100 years ago by some 8000-20000 individuals and has since then lived

in relative isolation due to the natural barrier of the North Atlantic that has hindered gene-flow (Helgason et al. 2000).

1.2.1.5 Hardy-Weinberg Equilibrium

If a population is infinitely large and evolves under neutrality then the genotype frequencies between generations are constant. This so-called Hardy-Weinberg Equilibrium (HWE) (Hardy 1908) is widely used in genetics to control for errors in generated data and use as a model distribution for different genetic algorithms. The assumptions for HWE are, however, very stringent; 1) infinitely large population, 2) random mating, 3) every one in the population mates, 4) everyone produce the same number of offspring, 5) there is no migration into or out of the population, 6) there is no positive or negative selection and 7) there is no mutation. These criteria are of course impossible to meet in real life but can be used as an approximation of how alleles behave in a population. Deviations from this rule, barring errors in sampling of individuals and generated data (Cox and Kraft 2006), could be due to events that break one or many of these criteria (Ryckman and Williams 2008).

1.2.2 Measures of selection and differentiation

The concepts explained above can be investigated by searching for the specific demographic footprints that they leave behind in the genome of humans and other species. By using different variants or haplotypes from the autosomal chromosomes, X-chromosome, Y-chromosome or mitochondrial genome, it is possible to zoom in on population genetic events that took place at different time points in history. One should bear in mind, however, that none of the events are mutually exclusive (Sabeti et al. 2006). Therefore, to strengthen the conclusions drawn by such studies it is imperative that the genetic findings can be put in relation to historical events. Below I describe some examples of how selection, population differentiation and migration can be measured.

1.2.2.1 Traces of selection

Selection is a local event that affects single variants and the surrounding areas in the genome. Very old events of positive selection can be investigated by calculating the relative amount of function-altering variants compared to neutral variants in different genomic regions in different species. This is because over long time periods the variants in a gene that are beneficial to a species should increase in number and population frequency. Common measures used in this context are the rate ratio of non-synonymous to synonymous substitutions, d_N/d_S, and the McDonald-Kreitman test (McDonald and Kreitman 1991).

An example of two genes that have been reported to show signs of strong positive selection in primate lineages leading to humans are the Abnormal Spindle-like Microcephaly-associated (*ASPM*) gene and microcephalin (*MCPH1*) (Evans et al. 2005), genes that are associated with a developmental disorder and have been suggested to affect brain size (Ponting and Jackson 2005). However, other reports

claim that the evidence is more consistent with modern humans interbreeding with archaic humans (Evans et al. 2006; Hawks et al. 2008) and data show that the genes are not associated with IQ or brain size (Woods et al. 2006; Mekel-Bobrov et al. 2007). Still others suggest that the signs are consistent with human population structure and expansion in the absence of selection (Currat et al. 2006), indicating that the proper methodologies are still not in place to satisfactorily answer this question.

More recent events of selection can be studied by searching for areas in the genome that are more homogeneous than expected under neutral evolution. This is explained by the observation that since variants are inherited as parts of haplotypes, both the variant under selection and the linked variants on the haplotype increase in prevalence (Smith and Haigh 1974). This is called a selective sweep and leads to a pattern of reduced genetic variation in the region that will remain for periods of up to 250000 years before erased by mutation. The size of the area that manifests reduced variation is directly comparable to how strong the selection process was.

Selective sweeps can also be investigated by using information regarding the frequency spectrum (Figure 3). A frequency spectrum is a summary of the allele frequencies of all mutations that segregate within the population. This summary is compared to a model consistent with a population evolving under neutrality, and any deviations from this model might be indications of selection (Figure 3a). The most famous example of a measure of the frequency spectrum is Tajima's D (Figure 3b) (Tajima 1989).

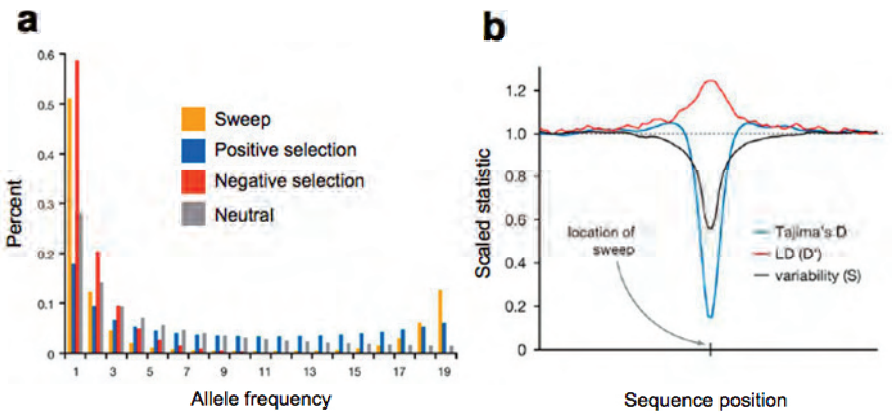

Figure 3. **| The frequency spectrum. a |** An example of the frequency spectrum under neutral evolution and as affected by different selective events. **b** | The effect of a selective sweep on Tajima's D, linkage disequilibrum and genetic variability. (Reprinted, with permission, from the Annual Review of Genetics, Volume 39 ©2005 by Annual Reviews www.annualreviews.org).

Recent selective sweeps also lead to very long-range haplotypes that persist for less than 30000 years before the correlation between nearby alleles decay due to recombination (Sabeti et al. 2006). An example of this kind of pattern can be found

around the lactase gene (*LCT*) that encodes the enzyme that metabolizes lactose. Approximately 80% of northern Europeans share an over 1 million base pair long haplotype that harbours a variant associated to lactase persistence, the condition when *LCT* remains active after weaning. This corresponds to a selective event that took place 5000-10000 years ago, around the same time that dairy farming was introduced (Bersaglieri et al. 2004). It has also been shown that specific biological processes, like host-pathogen interaction and reproduction, are overrepresented in genes that have undergone recent positive selection (Wang et al. 2006).

A feature of selection on variants on the Y-chromosome and in the mitochondrial genome is that the lack of recombination will cause all variants to be swept up together with the selected variant. Selection will consequently reduce the total genetic variation in these DNA sequences.

1.2.2.2 Population subdivision

While natural selection works locally, events like population bottlenecks, genetic drift, migration and isolation have an effect on the whole genome. Together, these events have lead to extensive heterogeneity in haplotype patterns and allele frequencies in different parts of the genome (Weir et al. 2005). To capture the overall genetic differentiation between individuals and populations, data should therefore be compared from several genomic locations.

A common approach is to estimate the genetic distance between populations using randomly chosen variants from the genome. Common measures for the genetic distance are the F-statistics that divide the total genetic variation in a population into variation within individuals relative to the subpopulation (F_{IS}), variation in a subpopulation relative to the total population (F_{ST}) and variation within individuals relative the total population (F_{IT}). Measuring the correlation between geographic and genetic distance tests isolation-by-distance, that is, the notion where the genetic distance between populations is a result of genetic drift due to at least partially isolated populations. These and several other genetic measures are integrated into many software packages, one of the most widely used being Arlequin (Excoffier et al. 1992; Excoffier and Heckel 2006).

Another commonly used approach is principal components analysis (PCA) that facilitates visualization of multidimensional data. Here the number of the measured variables that explain the total variation is reduced by forming a smaller number of synthetic variables, or principal components, that extract as much of the variation as possible from the data. The use of PCA in population genetics was first demonstrated in 1978 when Luca Cavalli-Sforza and co-workers constructed synthetic maps of Europe and East Asia based on data from 10 loci (Menozzi et al. 1978). While the results can be visually appealing and correspond to historical patterns of gene-flow (Seldin et al. 2006), the maps are still difficult to interpret as genetically meaningful. A recent study shows that mathematical artefacts, instead of real migration events, might explain some of the patterns observed in PCA plots (Novembre and Stephens 2008). So, while PCA

is a very useful exploratory tool, it should optimally be used in conjunction with more formal hypothesis tests.

The methods mentioned above require a prior division of individuals into populations before the actual genetic analysis. This can be approached in an opposite way by first dividing individuals into groups based on similarity in genotypes and then test or visualize how these groups correspond to prior knowledge about ancestral origin or other characteristics. The most widely used program that uses this principle is called STRUCTURE (Pritchard et al. 2000a; Falush et al. 2003a; Falush et al. 2003b), an algorithm that assumes that all genotypes in the sampled populations are derived from one or more unobserved populations. Based on assumptions regarding HWE, marker independence, level of admixture and number of unobserved populations, the algorithm tries to divide the individuals into groups that correspond to the assumed genetic distribution. Herein lies a potential weakness since the algorithm is sensitive to the choice of model parameters and assumptions about the data, but the parameter that has the biggest impact on statistical power is the number of markers (Turakulov and Easteal 2003; Rosenberg et al. 2005).

A similar method to STRUCTURE is the model-based algorithm Geneland (Guillot et al. 2005; Excoffier and Heckel 2006). The novel aspect of Geneland is that it makes use of the geographic coordinates to assist in dividing the individuals into genetically similar groups. The logic behind this is that individuals living close to each other are assumed to have more in common genetically than individuals living far apart. The method has successfully been used in ecology to infer the population substructure in situations with very low differentiation (Coulon et al. 2006; Rowe and Beebee 2007; Latch et al. 2008).

1.2.2.3 Maternal and paternal migration patterns

The non-recombining uniparentally inherited Y-chromosome and mitochondrial DNA are widely used to make inferences about historical parental migration patterns and genealogy. Because the mutation rate is higher in Y-chromosomal haplotypes (formed by microsatellites) and mitochondrial variants from the hypervariable regions, these are used to make inferences on the genealogical time-scale, while haplogroups are used to make inferences about demographic events on the population level.

By comparing sequence data between individuals it is possible to construct hierarchical phylogenetic trees, and the root of the trees refer to the ancestral haplogroup representing the most recent common ancestor (MRCA). By using geographical frequency information it possible to estimate the position of the most recent ancestor (PMRCA) and based on estimates of mutation rates it is also possible to infer a time-scale on when each haplogroup diverged, giving an approximation of the time to the most recent ancestor (TMRCA).

Today there are large databases that contain information about haplogroup frequencies in different parts of the world, and by using this information it is possible to derive a

time and direction of when and from where a country was settled. Diverging migration patterns between Y-chromosomal and mitochondrial data can be used to make conclusions about events that have affected maternal and paternal lineages differently. However, as the non-recombining uniparentally inherited sequences only represent one possible evolutionary outcome, autosomal and X-chromosomal data should be used to form a more complete picture of the genetic architecture.

1.3 GENETIC ARCHITECTURE OF COMMON DISEASES

As we have seen, the genome is under constant evolution due to different population genetic and mutational events. Most rare diseases are due to a broad spectrum of highly damaging recent mutations in single specific genes. The frequency of the deleterious mutations in the general population is low due to negative selection, but if the population is small and isolated these mutations may increase in frequency due to genetic drift or inbreeding, resulting in local high incidences of specific rare diseases. It is estimated that there are about 10000 of these diseases, and approximately 2000 of them have been characterized on the molecular basis (McKusick 2008).

Common diseases, however, that have a much larger impact on public health have proven much more difficult to characterize. For some diseases, it is estimated that up to 100 individual genetic variants, excluding the effect of the environment covers the complete risk spectrum (Kruglyak 2008). It is consequently extremely difficult to extract the underlying predisposing factors from the seemingly infinite pool of interconnected genes and other functional elements. This has lead researchers to propose different hypothesis regarding the genetic architecture, creating frameworks for how to approach the characterization of common diseases. It is only during the past few years, with the advent of genome wide panels of SNPs, that it has been possible to put these hypotheses to test.

1.3.1 The hypotheses

According to the common disease/common variant hypothesis (CD/CV), most of the genetic variability in common disorders is explained by a few major variants that rose to high frequencies due to past bottleneck events. This hypothesis was first suggested in 2001 and has been a prevailing paradigm since (Reich and Lander 2001). Changing environmental conditions could have caused a mismatch between the common variants and their local setting, leading to population specific deleterious effects that cause disease (Di Rienzo and Hudson 2005). Indeed, many of the life-style diseases of today, like asthma and type II diabetes, show population differential incidences that are consistent with environmental risk factors (D'Amato et al. 2005; Misra and Ganda 2007).

An extension of the CD/CV hypothesis has also been suggested (Becker 2004). The common variant/multiple disease (CV/MD) hypothesis incorporates the observation that several common variants are seen to associate with different diseases, emphasising

shared genetic and environmental factors between diseases. While these hypotheses stress the importance of variants with high prevalence in populations, a "competing" hypothesis suggests that the genetic spectra of common diseases are better explained by many rare variants (rare variant/common disease hypothesis) (Pritchard 2001; Pritchard and Cox 2002).

1.3.2 A population genetics perspective

Contemporary studies are now able to empirically test the validity of these hypotheses. Large collaborative efforts like the International HapMap Project (2003; 2005; Frazer et al. 2007) have made data available from millions of SNPs genotyped in multiple worldwide populations. Current reports based on these data are highlighting the importance of rare slightly damaging variants as risk factors for common disease (Gorlov et al. 2008; Lohmueller et al. 2008). These studies argue that slightly damaging SNPs may have a significant impact on common disease because they are widely distributed in the genome and are not totally wiped out by negative selection since their effect on reproductive fitness is only small. Consistent with a bottleneck during the expansion out of Africa, these SNPs are also more common in Europe compared to Africa (Lohmueller et al. 2008).

In other reports, the effect of positive selection is emphasized as a strong force in shaping the population distribution of several variants that are important for morphology and disease (Barreiro et al. 2008; Myles et al. 2008). The difference between continents is further shown in a study where individuals clustered according to ancestral place of origin based on only a few randomly picked SNPs (Allocco et al. 2007), while the difference in prevalence in common non-synonymous SNPs between African Americans and other populations in the United States is significant (Guthery et al. 2007).

1.3.3 Genome-wide association studies

Already over a decade ago it was realized that very large numbers of variants and samples would be needed to investigate the genetics of common complex disorders (Lander 1996; Risch and Merikangas 1996). The International HapMap Project followed the Human Genome Project, with an ambition to characterize all the common variation in the human genome in populations from Africa, Asia and Europe (2003; 2005; Frazer et al. 2007). These projects created a need for high-throughput genotyping platforms, and have lead to the situation today where it is possible to characterize 1 million SNPs in thousands of individuals at a time.

Only a few years ago this was impossible and most studies relied on some prior knowledge of which gene should be tested. These so-called candidate gene approaches had mixed success in robustly identifying predisposing variants to common diseases. Meta-analyses have shown that several of the positive associations are either false or at least lack evidence of replication, most probably owing to statistically underpowered

studies (Ioannidis et al. 2001; Lohmueller et al. 2003; Munafo 2004). Interestingly, it has also been shown that most variants that have been robustly associated to common disease prior to the genome wide association era do not show significant differentiation across populations compared to random variants (Lohmueller et al. 2006).

Today, with the aid of new high-throughput platforms and large sample sizes, new candidate genes for common diseases are reported on a monthly basis. The identified variants are common, but usually have very small effects, explaining only a fraction of the total genetic component (McCarthy et al. 2008). For some diseases, like asthma, where over 100 variants have been implicated based on candidate-gene approaches (Zhang et al. 2008), a single published GWA study did not confirm any of these (Moffatt et al. 2007) very likely due to too sparse SNP map used. Conversely, for diseases like type 1 and 2 diabetes, GWA studies show better success (2007; Hakonarson et al. 2007; Saxena et al. 2007; Scott et al. 2007; Sladek et al. 2007; Todd et al. 2007; Zeggini et al. 2007).

The picture that is emerging with respect to the genetic architecture is consequently a mixed one, reflecting different aetiologies for different diseases. The GWA studies are currently only able to interrogate common variants and it will be interesting to see if the predictions regarding rare variants will turn out to be correct or not. Importantly, GWA studies are not expected to implicate rare variants as they occur on different haplotypic backgrounds. Whole genome sequencing will be required to test multiple rare variants.

1.4 ZYGOSITY TESTING

While population genetic and association studies aim to investigate historical migration patterns and the genomic architecture of common diseases, twin studies are used to estimate the size of the genetic component of a disease. By comparing the incidence of a phenotype between groups of identical and non-identical twins, assuming equal environments for individuals in a twin pair, it is possible to estimate how much of the variation in incidence between the groups is attributable to heredity (Boomsma et al. 2002).

1.4.1 Twinning

One to two percent of all births are twin births and out of this one third represents twins that have developed from the same oocyte, so called monozygotic (MZ) twins (Hankins and Saade 2005). Until recently monozygotic twins were considered to share almost 100 percent of their genomes, but a new study on identical twins has discovered differences in CNVs, challenging the notion of genetic identity (Bruder et al. 2008). Conversely, dizygotic (DZ) twins develop from two separately fertilized oocytes and share on average 50 percent of their parental genetic material. While zygosity can be estimated based on validated questionnaires and chorionicity, the most reliable method

is to make use of the genetic variation in the human genome (Jackson et al. 2001; Ooki et al. 2004; Reed et al. 2005).

1.4.2 Choice of markers for zygosity testing

Much like ordinary fingerprints, it is possible to obtain unique genetic patterns that can be used to discriminate between individuals (Jeffreys et al. 1985). A set of genetic markers is chosen based on criteria that they are highly polymorphic, i.e. show large variation in the population. Often, they are also chosen to be unlinked and neutral. By minimizing the correlation between markers the power of exclusion is maximized. Neutral markers are used to try to ensure that they are not associated to any disease and to minimize the possible differences in population specific allele frequencies (Petkovski et al. 2005; Pakstis et al. 2007).

These panels most commonly consist of microsatellites and there are several robustly validated commercial sets available on the market (Nyholt 2006). However, technological advancements have made the use of SNPs possible and they are gaining ground also within the field of genetic testing (Petkovski et al. 2005; Pakstis et al. 2007). Besides the clear improvements in throughput, they are more robust when it comes to DNA of bad quality or whole genome amplified DNA (Utsuno and Minaguchi 2004; Petkovski et al. 2005; Dixon et al. 2006).

1.4.3 Estimating zygosity

Accurate zygosity assignment is achieved by estimating the likelihoods for the two individuals being either monozygotic or dizygotic given their genotypes. Numerous methods have been developed that probabilistically estimate the relationship between individuals (Risch et al. 1999; Abecasis et al. 2001), including methods that account for genotyping errors (Epstein et al. 2000; Sieberts et al. 2002). The inclusion of an error model is vital since quite small errors in zygosity assignment have been reported to cause false positive results in twin studies (Boomsma et al. 2002; Reed et al. 2005). This is also the reason why non-genetic methods for zygosity assignment are suboptimal since they have an accuracy of only 95-98% (Lichtenstein et al. 2002; Reed et al. 2005). Also, if DNA of lower quality is used, and especially whole genome amplified DNA that can lead to allelic dropouts, the likelihood for errors grows and should be taken into account in the analyses.

1.5 BIOLOGICAL ARCHIVES

Contemporary genetic studies involve very large numbers of individual samples since it has been widely accepted that smaller studies do not have the statistical power to find variants of small effects (Ioannidis et al. 2001; Lohmueller et al. 2003; Munafo 2004). Efforts like the UK Biobank are ongoing to establish repositories that will harbour adequate numbers and quality of material for studies dealing with several different

diseases (Ollier et al. 2005). However, depending on the incidence and age-of-onset of the disease or trait under study, it will take between 6 to 20 years to prospectively identify enough patients for a statistically well-powered study (Ollier et al. 2005; Palmer 2007). This is the case for most cancers as well as for diseases like adult type diabetes and Alzheimer's disease that develop later in life. Also, these estimates are based on the assumption that most disease variants are common, something that cannot be taken for granted given recent reports (Gorlov et al. 2008; Lohmueller et al. 2008).

A currently prevailing approach is to venture into multi-centre studies like the Wellcome Trust Case Control Consortium (WTCCC) that consists of 50 different laboratories working together to characterize the genetic variation in several different diseases. The WTCCC successfully demonstrated the use of 14000 cases of seven common diseases and 3000 common controls in a genome wide association study, paving the way for new large-scale collaborative projects (2007).

Other repositories exist that offer their own advantages, both in scale and quality. For example, archives of biological material built up during decades of routine screening programs, diagnostic efforts and local research projects, or the massive Twin Registries that exist in many countries around the world represent unique sample collections. Here, I will briefly describe the sample collections used within the scope of this thesis, with a focus on the Swedish newborn screening registry that represents the foundation in three of my papers.

1.5.1 The Swedish newborn screening registry

In 1963 Guthrie published data that showed the feasibility of using blood collected on filter paper for Phenylketonuria (PKU) screening on newborns (Guthrie 1992). PKU is a metabolic disease that, if not treated, leads to mental retardation in the patient. Guthrie's discovery bypassed the need for drawing a vial of blood from each newborn, and made mass screening for PKU possible. Since then, so-called Guthrie cards have been routinely used for determining the levels of numerous analytic substances. In Sweden, nation-wide screening on all newborns started in 1965 and all samples dating back to 1975, approximately 3 million individual slips of filter paper, are still stored at the Karolinska University Hospital in Huddinge.

Today, every newborn is routinely tested for five diseases, namely PKU, galactosemia, congenital hypothyreosis, androgenital syndrome and biotinidase deficiency. This repository grows by approximately 100000 samples each year and encompasses the largest single biobank in Sweden.

1.5.2 The Swedish Twin Registry

The Swedish Twin Registry was founded in the late 1950s mainly for research purposes that focused on the effect of smoking and alcohol consumption on cancer and cardiovascular diseases. It has since then grown to encompass over 170000 twins,

corresponding to virtually all twins born between the years 1886 and 2000 (Lichtenstein et al. 2006).

1.5.3 The Oulu RDS cohort

The Oulu RDS cohort encompasses 521 infants born during the years 1998 and 2001 in Oulu University Hospital. Respiratory distress syndrome (RDS) was diagnosed based on a requirement of supplemental oxygen and continuous distending airway pressures for at least 48 hours after birth unless treated with exogenous surfactant in established respiratory failure, diffuse reticulogranular pattern with air bronchograms and ground glass appearance of lung fields in the chest X-rays. Surfactant therapy prior to radiography was considered an exclusion criterion, and neither pneumonia nor transient tachypnoea was diagnosed as RDS. Bronchopulmonary dysplasia (BPD) was diagnosed in infants born before 32 weeks of gestation that required supplemental oxygen at postmenstrual age of 36 weeks (Jobe and Bancalari 2001).

1.5.4 Sample collection for population genetics in Finland

To investigate the population genetic substructure in Finland, 657 individuals have been collected through the Finnish Red Cross. These individuals represent a single generation of males between 40 and 55 years of age. The subjects are sampled so as to represent a rural population, and their geographic locations have been assigned according to grandparental birthplaces.

1.6 EXTRACTION AND PREAMPLIFICATION OF DNA

1.6.1 DNA sources and extraction

To avoid inhibition or disturbance of enzymatic reactions, most current genotyping applications require purified DNA as their starting material. Since DNA is a rather stable macromolecule it can be separated from the cellular proteins and other debris by means of different physical and chemical methods. Which approach should be used is dependent on the tissue, age and type of sample, as well as the requirements concerning cost, throughput, quality and quantity of the extracted DNA. Regardless, the methods involve steps that disrupt the cell, dissolve the phospholipid membranes surrounding the cell, mitochondria and nucleus, and finally remove or degrade proteins and other fragments to make the DNA accessible.

1.6.1.1 Blood and saliva

The most common source of DNA that is used in genetic research is whole blood. As long as the sample has been collected and stored in an adequate way the quality and quantity of the extracted DNA is generally very good (Steinberg et al. 2002). An alternative and highly attractive source is saliva where the epithelial cells lining the inner surface of the oral cavity yield sufficient material for extracting DNA. Since the collection procedure is non-invasive it often leads to a high percentage of subjects willing to participate in the study (Rylander-Rudqvist et al. 2006). It is vital that this so

called response rate is high as it positively influences the power of the study and minimizes possible biases due to non-responsiveness. A non-invasive method is especially important when dealing with young children or people who cannot donate blood due to medical or religious reasons.

Further, the logistic burden of collecting saliva compared to blood is lighter since the study subjects themselves can perform the actual sample collection. The Genographic Project, a population genetic project aimed at charting the human migration patterns throughout history, is a great example of how this works in practice. Here, the sample collection and transportation is organized using ordinary mail, minimizing the cost for personnel and making it possible for anyone around the globe with access to a mailbox to participate in the study (Behar et al. 2007).

The most obvious drawback with using saliva over blood is the large amount of extracted foreign DNA due to the bacterial flora and food particles in the mouth. It is estimated that around 30% of the total amount of extracted DNA from saliva is of bacterial origin (Rylander-Rudqvist et al. 2006). Even if this does not interfere with the downstream assays, the amount of human DNA used for these applications may be underestimated. Also, the variation in DNA concentration, both bacterial and human, between individual samples extracted from saliva is larger compared to whole blood, and because of this it is very important to make sure that ample amounts of DNA are used for any downstream applications. Interestingly, some inflammatory conditions are associated with DNA yield and this may bias studies that ascertain individuals based on the amount of extracted DNA (Alanne et al. 2004). This should be remedied by using whole genome amplification or by sampling another tissue.

1.6.1.2 Blood on filter paper

Due to the stability of blood samples dried on filter paper, ease of shipment and the small blood amounts required, they offer a convenient source of DNA and other analytes (Bereczky et al. 2005; Hollegaard et al. 2007; Olshan 2007; Sjoholm et al. 2007). Several different methods, both commercial and non-commercial, can be used to extract DNA from different types of filter paper (Kline et al. 2002; Pachot et al. 2007; Sjoholm et al. 2007).

Very old samples and samples that have been stored under suboptimal conditions, like the earliest batches in the Swedish Newborn Screening Registry, may require more aggressive DNA extraction methods that involve the use of strong detergents, boiling, sonication, vortexing or enzymatic degradation. The reason for this kind of an approach is to maximize the extraction yield since the absolute amount of DNA in the samples may be small or chemically modified and thus difficult to access by standard methods. Optimally these kinds of approaches should be used sparingly since old samples already contain fragmented DNA (Chaisomchit et al. 2005; Sjoholm et al. 2007) and the most attractive genotyping platforms today require DNA of high fidelity. The use of whole genome amplification, either directly on dried blood spots or on extracted DNA is a viable option to increase the usability of these samples.

1.6.2 Whole genome amplification

Pieces of the DNA-replication machinery are used in genetic laboratories to facilitate the enrichment of certain fragments of DNA, an innovation called polymerase chain reaction (PCR) that revolutionized the field in the 1980s and was awarded the Nobel price in Chemistry in 1993. In the early 1990s, the PCR technique was extended to encompass the whole genome, making it possible to elongate the lifetime of individual DNA samples or to investigate the genetic make-up of single cells. Further development has lead to the discovery of novel enzymes that are less prone to cause errors during replication, more efficient and generate considerably longer pieces of DNA than the original methods. As is the case with DNA extraction, however, there is no single omnipotent method since the quality and nature of the starting material often governs which approach should be used.

1.6.2.1 Different WGA methods

Preamplification methods that are based on PCR technology include primer extension preamplification (PEP), Degenerate oligonucleotide PCR (DOP) and the GenomePLEX technology. The PEP method is based on using a mix of randomly assembled primers that are 15 nucleotides long and a combination of low and high stringency temperature cycles (Zhang et al. 1992). The random primers anneal during the low stringency-cycles to arbitrary places in the genome and when the temperature is raised the PCR reaction becomes more specific, amplifying the genome using the random sequences annealed during the first cycles as priming sites. The method has been improved since it was first published and now includes a high-fidelity polymerase (I-PEP-L) (Dietmaier et al. 1999). A modified version of this protocol (modified-I-PEP, or mI-PEP) has also been developed for use with very small amounts of DNA in forensics (Hanson and Ballantyne 2005).

Very similar to PEP, degenerate oligonucleotide PCR (DOP) and its modifications relies on partially degenerate primers that are 22 bases in length and a combination of low- and high-stringency temperature cycles (Telenius et al. 1992; Cheung and Nelson 1996; Kuukasjarvi et al. 1997; Buchanan et al. 2000; Kittler et al. 2002). GenomePLEX is a more recent method that relies on PCR amplification of randomly fragmented DNA and using universal linkers that are attached to the ends of each fragment as priming sites for the enzyme (Little et al. 2006). Another recent development within WGA has lead to techniques that utilize the DNA polymerase from the ϕ29 bacteriophage and random hexamers protected from exonuclease degradation to isothermally amplify DNA (Lizardi et al. 1998; Dean et al. 2002).

1.6.2.2 Utilities and restrictions

Because of the restrictions of the polymerases used in PEP and DOP the resulting preamplified DNA is not comparable to high quality DNA, with average fragment length < 3kB (Silander and Saarela 2008). The short fragments of these methods is a clear disadvantage, as the most high-throughput genotyping methods today require high

molecular DNA as template. Some parts of the genome, like repetitive sequences, are also not amplified by PEP and DOP and the genome is thus not fully covered after preamplification.

GenomePLEX, on the other hand, has been successfully used on genotyping platforms that require DNA of higher quality as well as on DNA of bad quality, making it a very versatile method (Hittelman et al. 2007). MDA is also compatible with the high-throughput platforms, including next-generation sequencers (Montgomery et al. 2005; Paynter et al. 2006; Pinard et al. 2006; Berthier-Schaad et al. 2007). However, it has been shown that MDA performs poorly when the quality of the starting material is bad (Sun et al. 2005; Leanza et al. 2007). Old, fragmented DNA is consequently not suitable and it is here that the PCR-based methods have a clear advantage. MDA has also been associated with the formation of chimeric sequences that are due to secondary structures formed during the amplification reaction (Lasken and Stockwell 2007).

PCR based whole genome amplification methods are especially vulnerable to amplification bias (Dean et al. 2002; Lovmar et al. 2003; Sun et al. 2005). This means that one DNA strand may at times be favoured over the other one, leading to an overrepresentation of the favoured strand and consequent artefacts that may manifest like any other genetic variant or mask an existing variant. This phenomenon is highly dependent on the amount and quality of starting material, and to minimize the bias it is recommended to use a minimum of 10 nanograms of template (one cell contains on average 6 picograms of DNA). While MDA is not as vulnerable to amplification biases, it still remains a potential problem and the same criteria concerning the amount of starting material should be followed, with increased amount of DNA correlating with improved whole genome amplification and genotyping results (Lovmar et al. 2003; Bergen et al. 2005).

1.7 QUALITY CONTROL

Surprisingly small errors can distort the results in a study, consequently leading to conclusions that are based on faulty or unrepresentative data. Because of this, quality control should have the highest priority in any study and form the basis of any well-designed genetic study. In practice this means that the researcher should be vigilant from the very beginning of the study, focusing not only on identifying errors in the genotypic data and concomitant analyzed results, but also paying attention to how the individual subjects are recruited and what information is known about them. Below I discuss examples of how errors can be detected and controlled in different phases of a project.

1.7.1 Internal validity of a study

1.7.1.1 Genotyping errors and missing data

Genotyping errors may in a best-case scenario only introduce noise and reduce the power to find association (Gordon and Finch 2005; Moskvina and Schmidt 2006; Nicodemus et al. 2007). This happens if the error is random, while a non-random genotype error may lead to spurious associations and distorted haplotype inferences (Liu et al. 2006b; Moskvina and Schmidt 2006). Misclassification of phenotypes may also lead to loss of power (Gordon and Finch 2005), and differentially missing genotype data have been shown to increase the likelihood of false positive associations and estimates of haplotype frequencies (Liu et al. 2006a). The power of the TDT (transmission disequilibrium test) algorithm is also reduced when excluding incomplete trios based on missing data for offspring (Guo et al. 2008). Also, if the genotyping error is differential with regards to phenotype, i.e. more cases than controls are affected by the error, the false positive rate of association may be elevated (Clayton et al. 2005), and is more pronounced in family based studies (Hao and Cawley 2007).

1.7.1.2 How to detect errors

The most common way to detect the presence of errors is to use DNA with known genotypes as controls as well to genotype replicates of the same sample to identify inconsistent genotypes (Pompanon et al. 2005). The replicates can be collected independently from the same individual and genotyped in parallel (biological replicate), or one extracted DNA sample can be genotyped repeatedly (technical replicate) (Figure 4). As cost is often a limiting factor the number and type of replicates and controls should be added based on prior knowledge about the quality of the DNA.

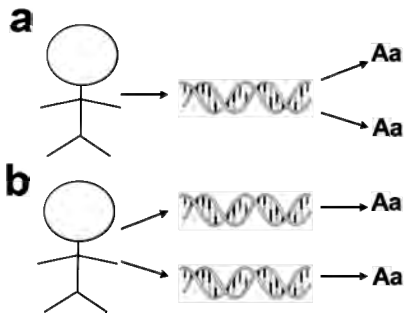

Figure 4. **Replicates. a** | Technical replicate: DNA is extracted once and genotyped in parallel, genotypes are compared. **b** | Biological replicate: DNA is extracted twice in parallel from the same individual and the replicas are genotyped and results are compared.

Deviations from HWE are also used to check for discrepancies. Still, based on simulations the power to detect genotype errors using HWE is quite small (Leal 2005; Cox and Kraft 2006; Moskvina and Schmidt 2006), and deviations may be due to any event that breaks the assumptions of HWE (see chapter 1.2.1.5). Data that are inconsistent with HWE should therefore not be discarded automatically and some scientists even advice against the use of HWE for detecting genotype errors (Zou and

Donner 2006). This might still be a dangerous suggestion as data generated from the genome-wide association panels are impossible to thoroughly check manually and large errors are detected by HWE (Teo et al. 2007).

Genotypes that are inconsistent with Mendelian inheritance are readily detected (O'Connell and Weeks 1998) and using information about neighbouring loci makes it possible to detect errors that are consistent with Mendelian inheritance (Abecasis et al. 2002; Sobel et al. 2002). Still, even if the presence of an error is detected it is sometimes impossible to pinpoint it. In this case the whole family has to be discarded for that specific marker and this reduces the power of the study. It is also difficult to collect large family material and the power of case-control studies have been shown to be larger than for family-based studies (Morton and Collins 1998; McGinnis et al. 2002).

1.7.1.3 Imputation and fuzzy calls

Data are usually considered to be missing at random and consequently disregarded in the analysis. Several algorithms that infer haplotypes, however, impute the missing genotypes by "guessing" based on the genotype distribution from successfully called samples, genotypes from related individuals and/or based on information regarding population substructure (Yu and Schaid 2007). As long as data are missing at random these methods are fairly exact (Sun and Kardia 2008), but problems arise when a specific genotype is overrepresented in the missing data (Liu et al. 2006a).

Instead of discarding or imputing missing data it is also possible to use so-called "fuzzy calls" that allow for the uncertainty in genotypes of lower quality in the association tests. Simply put, a genotype that would otherwise be discarded is assigned probabilities based on signal intensities or similar cluster information. A genotype could then be considered as 30% homozygote and 70% heterozygote in the association analysis. By using this approach it was shown that biases based on differentially missing data were decreased (Plagnol et al. 2007).

1.7.1.4 Population stratification and cryptic relatedness

Although case-control studies are preferred over family-based approaches, they can be confounded due to population stratification or cryptic relatedness (Knowler et al. 1988; Lander and Schork 1994; Devlin and Roeder 1999). When cases and controls represent populations with different ancestral origins the allele frequency differences between these populations may lead to spurious associations. In genetic epidemiological studies, the controls should therefore mirror the genetic background of the cases so that any detected associations are due to overrepresentation of predisposing alleles rather than to population substructure.

Cryptic relatedness refers to the situations when some individuals in a case-control sample are close relatives, resulting in an inflated false positive rate in association tests that do not account for this excess relatedness. It has been shown, though, that the effect of cryptic relatedness is often negligible given that the study is well designed. It

should be considered when the size of the study population is small, the population has undergone recent rapid growth, or when extensive inbreeding is suspected (Voight and Pritchard 2005).

Several approaches exist that test for the presence and account for the effect of population substructure and/or cryptic relatedness. Genomic control utilizes information from null markers (i.e. markers that are believed to be independent of the disease or trait under study) to measure the underlying population substructure. The association statistic is then adjusted based on the detected substructure (Devlin and Roeder 1999; Bacanu et al. 2000; Devlin et al. 2001). A similar approach uses the data from unlinked markers to infer the ancestry and admixture proportions of sampled individuals. Association is consequently tested for within the inferred subpopulations (Pritchard et al. 2000b; Pritchard and Donnelly 2001), but while this controls for population substructure it does not correct for cryptic-relatedness (Voight and Pritchard 2005). Other methods that cluster individuals into groups based on genotype data are also used to correct for population substructure (Price et al. 2006).

It is needless to say that if the population substructure is known from before based on population genetic studies, it is easier to sample cases and controls in a way that minimizes the problem of stratification. It is also important to remember that several environmental factors may confound an analysis and it has been even argued that the impact of population stratification is exaggerated in genetics (Paradies et al. 2007).

1.7.2 External validity of a study

Compared to the internal validity of a study that focuses on the trustworthiness of the generated data, the external validity considers to what extent these data can be generalized to other populations or situations. To evaluate the generalizability and trustworthiness of data, results should be replicated in another population (McCarthy et al. 2008). If this is not achieved the results may either be wrong (low internal validity) or they may be specific to the sampled population. Also, a recent study showed that the strength of the genetic effect may vary by age, and this may in itself lead to situations where it is difficult to replicate an initial association (Lasky-Su et al. 2008).

Another concern lies with the ascertainment bias regarding genetic markers. For example, the HapMap SNPs represent variants identified based on a quite small number of individuals that were then genotyped in different populations (2003). This ascertainment strategy favours SNPs with high frequencies in the initial sample panel and will lead to an overrepresentation of common SNPs. Any population genetic measures that rely on the frequency spectrum (for example Tajima's D and F_{ST}), linkage disequilibrium or heterozygosity are affected (Clark et al. 2005). The amount of heterogeneity between populations may be underestimated since common SNPs are more likely to be shared between populations, and recent population expansions could remain undetected as rare variants are underrepresented in the SNP panel. This should

be taken into consideration when drawing conclusions based on the results or corrected for using algorithms that adjust the frequency of each SNP based on the probability that they were discovered in the initial ascertainment panel (Clark et al. 2005; Rosenblum and Novembre 2007).

2 PRESENT INVESTIGATION

In the previous chapters I have presented some of the recent advances, challenges and opportunities within genetic epidemiology. Population genetic studies strive to explore the events that have shaped the genomes of contemporary human populations, and association studies aim to define the genetic risk factors that predispose these populations to disease. The development of methodological frameworks facilitates these studies, and the basis of this thesis was to provide a methodological approach for using large biobanks for a variety of different applications within genetic epidemiology. More specifically, in the five studies included in this thesis I:

1. Develop and validate of a methodological approach for using the newborn screening registry as a resource for genetic studies (Study 1)
2. Investigate the role of *NPSR1* as candidate gene for RDS using the methodological approach developed in study 1 (Study 2)
3. Develop a panel of SNP markers that can be used for large-scale zygosity testing on a wide range of samples, including preamplified samples (Study 3)
4. Investigate the population genetic substructure in Sweden and Finland using the preamplified samples from the newborn screening registry (Studies 4 and 5)

In the following chapters I will summarize the results from these studies and briefly discuss the implications of the methodological approach with respect to population genetics, association studies and zygosity testing.

2.1 USEFULNESS OF DRIED BLOOD SPOTS AND PREAMPLIFIED DNA (STUDIES 1-5)

Here, based on the five studies, I summarize the observations on the methodological strengths and weaknesses of using dried blood spots and preamplification for different purposes. In our first study we wanted to develop, validate and implement a method for extracting and preamplifying DNA from samples derived from the Swedish New Born Screening registry. DNA was extracted from 3mm^2 filter punches by incubating them in saponin to degrade the haemoglobin and cell membranes before boiling the sample in a buffer mixture containing a cation chelating resin, chelex-100. The extracted DNA was then preamplified using the improved primer preamplification method (I-PEP-L) (Obermann et al. 2003) and genotyped on the Sequenom MALDI-TOF platform.

2.1.1 Method validation

In our initial validation we could show that it was possible to extract and preamplify DNA from up to 25-year-old PKU samples. While the difference in DNA yield was non-significant between the different age groups for both extracted and preamplified the PCR success rate for different amplicon lengths decreased by increasing age of the sample, something observed also in other studies (Chaisomchit et al. 2005; Sjoholm et

al. 2007). However, all samples of different age-groups were successfully amplified for 100bp amplicons, suggesting that it is possible to genotype even the oldest samples in the Swedish PKU registry on platforms that utilize short amplicons (Figure 5).

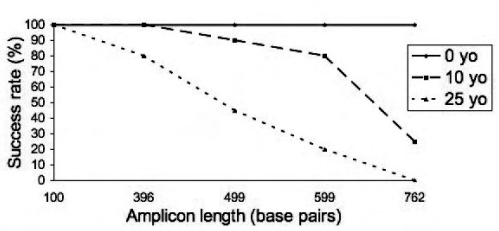

Figure 5. **PCR amplification success rate stratified by age of sample.** DNA was extracted and preamplified from dried blood spots and amplified using specific PCR primers yielding amplicons of different sizes. The success rate was calculated based on 20 samples per age group.

We further validated the method by blindly genotyping a set of 94 samples including an unknown number of patients with one or both of two PKU mutations. It has repeatedly been shown that preamplification methods are prone to allele dropouts due to bad quality or low amounts of template (Dean et al. 2002; Lovmar et al. 2003; Sun et al. 2005). This was not evident here as all patients were correctly identified with no false positive results. Additionally, our results showed no discrepancies between extracted and preamplified DNA when genotyping 90 samples twice for three SNPs.

We extended the method to extract and preamplify 2132 PKU samples. These represent all births from one week in December 2003 plus an additional 89 samples from the northern counties in Sweden. This sample collection was used in this study to investigate the distribution of three disease-associated variants, *PPARγ* Pro12Ala, *APOγ4* and *CCR5* Δ32, as well as for studying the population structure using mitochondrial, Y-chromosomal and autosomal markers (studies 4 and 5). The population genetic results are discussed in chapter 2.4.

2.1.2 Quality control

While there were no apparent problems with the extracted and preamplified DNA in our first validation study, we did see signs of possible biases in the following studies that used the 2132 samples or the developed extraction method.

2.1.2.1 Possible bias in haplotype inference

We used the extraction and preamplification method on filter paper samples from the Oulu RDS cohort to investigate if NPSR1 was a candidate gene for RDS. Here we used an EM (Estimation Maximation) algorithm to infer the haplotypes (i.e. probabilistically estimate which alleles belong to the same haplotype) from pooled genotypes from both cases and controls. After the initital genotyping of seven gene-tagging SNPs we observed a large number of unrecognized haplotypes. We re-genotyped both samples with unrecognized haplotypes and randomly chosen samples. The results indicated that allele dropout was not a significant problem in the overall sample set but we observed a potential bias in the results obtained from the oldest sample cohort.

Because of this possible bias we decided to exclude the samples representing the oldest cohort (n=63) and 25 other samples that had low haplotype inference probabilities. It has been shown that genotyping error may have a marked impact on association tests using haplotypes (Liu et al. 2006b) and it is therefore imperative that proper quality control measures are taken when DNA of suboptimal quality is used.

2.1.2.2 Low success rate and possible heterozygote deficiency

In our following studies the 2132 extracted and preamplified PKU samples were used to investigate the population substructure in Sweden (Studies 4 and 5). In total, 23 Y-chromosomal and 37 mitochondrial SNPs were genotyped in study 4. As Y-chromosomal and mitochondrial genotypes only come as homozygotes, i.e. only one allele is present, allele dropout is not as severe a problem as it is for genotypes consisting of several alleles. We did observe reduced success rates when genotyping Y-chromosomal SNPs, leading us to exclude nine markers from the final analyses. This is indicative of the lower quality DNA due to preamplification.

In accordance with the higher number of mitochondrial genomes, we did not observe similar problems with the mitochondrial markers, but two markers out of 37 were still excluded due to quality issues. Discrepancies in the phylogenetic tree of the mitochondrial DNA were detected for 7.7% of the samples, but these genotypes were concordant between the Sequenom genotyping and RFLP genotyping. Because of this and since normal quality control procedures were followed we believe that at least part of these discrepancies are due to recurrent mutations rather than allele dropouts caused by the preamplification procedure.

In our fifth study we genotyped 34 highly multiplexed autosomal SNPs to further investigate the population substructure in both Sweden and Finland. We observed a larger amount of missing data for the preamplified Swedish samples (approximately 20% missing data) compared to the Finnish genomic DNA samples (approximately 10% missing data) and signs pointing to possible non-random genotyping errors. The increased number of markers not corresponding to HWE (7 out of 34) and possible deficit in heterozygotes due to inflated F_{IS} and F_{IT} lead us to investigate this further by simulating scenarios (Figure 6) that corresponded to non-random genotyping error as well as to different levels of hidden population structure (influence of individuals of unknown ancestry on population genetic measures).

Based on the simulations either scenario was consistent with our results (Figure 6a and 6b). Still, this underlines the notion that meticulous quality control is needed when working with samples of suboptimal quality. Our study was not designed to detect the type of genotyping error present, but this would be possible although cumbersome to do and would further strengthen the conclusions drawn from the study (Pompanon et al. 2005).

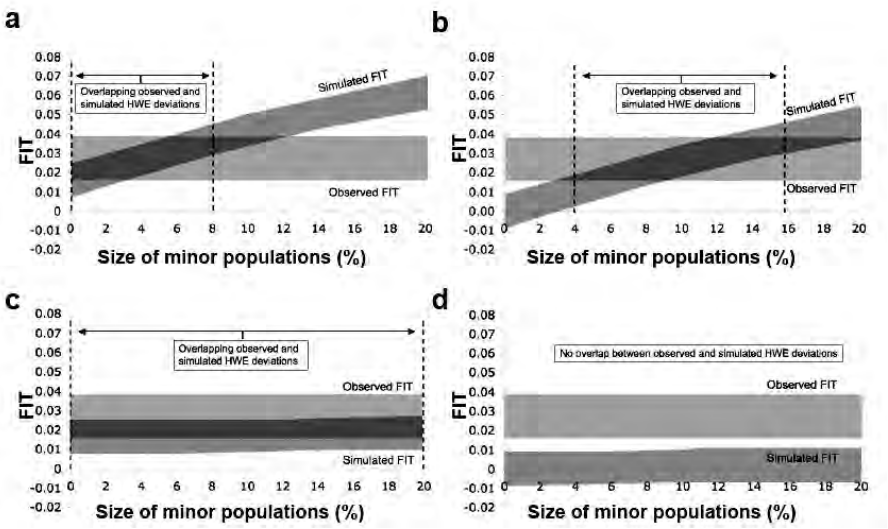

Figure 6. **Simulation results.** Genotyping errors were simulated based on observed errors. Non-European hidden population substructure was simulated based on HapMap frequencies from the African and Chinese populations. European hidden populations substructure was simulated based on observed Finnish allele frequencies. **a** | Non-random error and hidden non-European substructure, **b** | random error and hidden non-European substructure, **c** | non-random error and hidden European substructure, **d** | random error and hidden European substructure

2.1.3 Other DNA sources

In our third study we developed and validated a marker panel for zygosity testing. The panel was evaluated on 11 genomic and preamplified DNA samples extracted from whole blood, saliva, and filter paper. Our results showed that the quality of genotypes from preamplfiied DNA using MDA was comparable or even superior to genotypes from DNA extracted from whole blood and saliva. DNA extracted from filter paper yielded the worst results, while the preamplified DNA worked well.

The samples used here were fresh and consequently of good quality. It should be remembered though that some studies have shown that MDA is not optimal on fragmented samples (Park et al. 2005; Sun et al. 2005), and therefore not compatible with older PKU samples. Other studies report differently (Sorensen et al. 2007), however, and large-scale use of PKU samples preamplified using MDA is on-going (Eising et al. 2007).

2.1.4 Conclusions

In conclusion, we showed that even 25 year-old PKU samples could be used for SNP genotyping and that the Swedish Newborn Registry is a potential resource for large-scale genetic epidemiological studies. This is supported by other studies that make use

of similar samples for both genetic and proteomic purposes (Eising et al. 2007; Hollegaard et al. 2007; Sjoholm et al. 2007). Still, since allele dropouts or differential success rates may generate biases in the data, rigorous quality control is recommended.

2.2 *NPSR1* AS A RISK FACTOR FOR RDS (STUDY 2)

In our second study we extracted and preamplified DNA from 521 infants including 176 preterm infants diagnosed with respiratory distress syndrome and 37 with bronchopulmonary dysplasia. We wanted to assess the role of *NPSR1* as a candidate gene for RDS by genotyping seven tag-SNPs in the gene and testing for association using both haplotypes and single markers.

2.2.1 Respiratory distress syndrome

RDS is a life-threatening developmental disease of newborns. It is caused by an insufficiency of surfactant and structural lung immaturity, and the symptoms are characterized by breathing dysfunctions. RDS affects 0.5-1.5% of all neonates, with preterm birth being the major risk factor. Development of surfactant replacement therapy has decreased the mortality from 100% to 10%, but RDS is still the leading cause of death during the first month of life. Children with sever RDS are predisposed to chronic lung disease, including bronchopulmonary dysplasia.

Estimates of the heritability of RDS vary considerably, ranging from 20 to 80%, and polymorphisms of surfactant proteins are associated with an increased risk of RDS and BPD (Hallman and Haataja 2006).

2.2.2 Neuropeptide S receptor 1 (*NPSR1*)

The neuropeptide S receptor 1 gene (*NPSR1*) is one of over 100 genes reported to be associated to asthma and related phenotypes (Zhang et al. 2008). It was characterized using a candidate-gene approach where the genome was scanned in 86 Finnish families from the Kainuu region for areas that were linked with a high level of IgE. Statistical evidence for linkage was found on chromsosome 7 and was replicated in an independent sample set of Finnish families as well as in a French population for the asthma phenotype. Asthma and high IgE levels were then tested for association to markers in the area, identifying four haplotypes significantly overrepresented in the cases compared to controls in all three populations.

More detailed studies have revealed the presence of two genes, *AAA1* and *NPSR1*, and here *NPSR1* was differentially expressed in bronchial epithelial cells and smooth muscle cells from asthmatic individuals compared to controls, suggesting a role for *NPSR1* in the pathogenesis of asthma and related disorders (Laitinen et al. 2004).

2.2.3 *NPSR1* associated with RDS

Haplotype H1 has in previous studies shown to be both a risk and a protective haplotype for different phenotypes related to allergy and asthma (Kormann et al. 2005; Melen et al. 2005). In this study, it was underrepresented in the RDS cases (OR = 0.5; 95% CI 0.3-0.8; P = 0.01), while the H4/H5 haplotype increased the risk of RDS in infants with a gestational age > 32 weeks (OR = 2.6; 95% CI 1.2-5.5; P = 0.01). Also, the expression of the *NPSR1-B* isoform in the smooth muscle cells of the large bronchi in RDS and BPD was up regulated in preterm children with RDS or BPD. Taken together, these results indicate a role of *NPSR1* in the development of respiratory distress syndrome in preterm infants.

2.3 VALIDATION OF A SNP PANEL FOR ZYGOSITY TESTING (STUDY 3)

In our third study we wanted to develop a panel of SNPs that was amenable for high-throughput zygosity testing using DNA from a wide range of sources. Based on a previous panel of SNPs (Petkovski et al. 2005) we identified 47 markers that were in linkage equilibrium with each other (independently inherited) and polymorphic in a European population, and compared these to a panel of 11 STRs. The markers were genotyped on both genomic and whole genome amplified DNA, yielding results that favoured the SNPs based on higher average success rates and lower variability.

2.3.1 Comparison of a SNP and an STR panel

We then carried out zygosity assignment of 99 twin pairs from the Swedish Twin Registry using Eclipse v1.1 (Sieberts et al. 2002) and assuming a 1% genotyping error. The zygosity of the twins had previously been tested using a validated questionnaire (Lichtenstein et al. 2002). We could show that both the SNP and the STR panels were 100% concordant, with two pairs conflicting with previously assigned zygosities. This is in accordance with a 98% estimated accuracy of questionnaire based zygosity assignment (Lichtenstein et al. 2002). While the qualitative results were concordant, the STRs had a slightly higher failure rate than the SNPs. This, combined with a higher workload and cost, speaks in favour of using SNPs for large-scale zygsoity testing.

2.3.2 Estimation of false positive rates

We further evaluated the SNP panel by simulation studies, estimating the false positive rate of zygosity assignment for the Eclipse algorithm. The simulations were based on 10000 monozygotic and 10000 dizygotic twin pairs assuming genotyping error (0, 1 and 5%), missing data (0, 10 and 20%) and population stratification (0, 10 and 20% difference between assumed and real allele frequencies). The largest error in zygosity was seen for DZ twins that were in 0.79% (79 pairs out of 10000) wrongly identified as MZ. In this case the assumed allele frequencies differed by 20% compared to the real ones and 20% of the genotypes were missing. The largest error observed for MZ twins was 0.02% (2 pairs out of 10000), reflecting a scenario with a 1% genotyping error, 20% allele frequency shifts and 20% missing data.

2.3.3 Conclusions

In conclusion, our results show that SNPs fare better than STRs when it comes to large-scale zygosity testing, mostly due to a slightly higher success rate and lower workload. We also show that our SNP panel is robust in the presence of genotyping error, missing data and for twin pairs that have a different genetic background than assumed based on Swedish allele frequencies. As others have observed, today's world is highly cosmopolitan and marker panels used for zygosity testing or similar genetic fingerprinting purposes should be amenable for individuals with a broad range of ancestral origins (Pakstis et al. 2007). As already presented in chapter 2.1.3 the SNPs are also robust when genotyped on a wide range of different templates. This is imperative if the panel is being used for forensic use or large-scale projects that have sampled DNA from other sources than whole blood.

2.4 POPULATION SUBSTRUCTURE IN SWEDEN AND FINLAND (STUDIES 4 AND 5)

The last glacial period ended approximately 12000 to 15000 years ago, making human settlement possible in the areas previously covered by a thick layer of inland ice. The first signs of inhabitation in Sweden are dated at 12000 BC and in Finland 8000 BC, and according to prevailing theory these early settlers are the ancestors of present-day Finns and Swedes (Norio 2003). Sweden later grew up to be a dominating power during the Middle Ages and the following several centuries, ruling over Finland and large areas in the Baltic region. In the 14th century Finland was divided between Sweden and Russia with many battles being fought on Finnish territory.

2.4.1 Sweden

Relatively little is known about the Swedish genetic substructure. Based on Y-chromosomal studies, there are strong connections to Central Europe, Denmark and Norway, but no clear genetic substructure has been observed within the country (Holmlund et al. 2006; Karlsson et al. 2006). The northern parts of Sweden show admixture with Saami and Finnish people, and local population isolates exhibit higher incidences of several monogenic diseases (Einarsdottir et al. 2007).

2.4.1.1 CCR5 Δ32 more common in northern Sweden

In study 1, we characterized the frequency of three disease-associated variants in the Swedish population. The variants were *APOε4*, *PPARγ* Pro12Ala, and *CCR5* Δ32, variants associated to Alzheimer's disease, type II diabetes and HIV infection respectively. These variants were genotyped on 2132 samples collected through the Swedish Newborn screening registry in December 2003 (Figure 7). While we did not observe any regional differences in *APOε4* or *PPARγ* Pro12Ala frequencies, we did see a north south trend in the *CCR5* Δ32 frequencies. As previously observed in Europe as

a whole (Libert et al. 1998), the 32 base pair deletion was more common in the northern parts of the country.

There are several theories of how the geographic differentiation in *CCR5* Δ32 frequency has evolved. Some reports link it to the plague or smallpox epidemics in Europe during the Middle Ages, where individuals with the deletion would have been immune to these diseases (Stephens et al. 1998; Galvani and Slatkin 2003). Others have

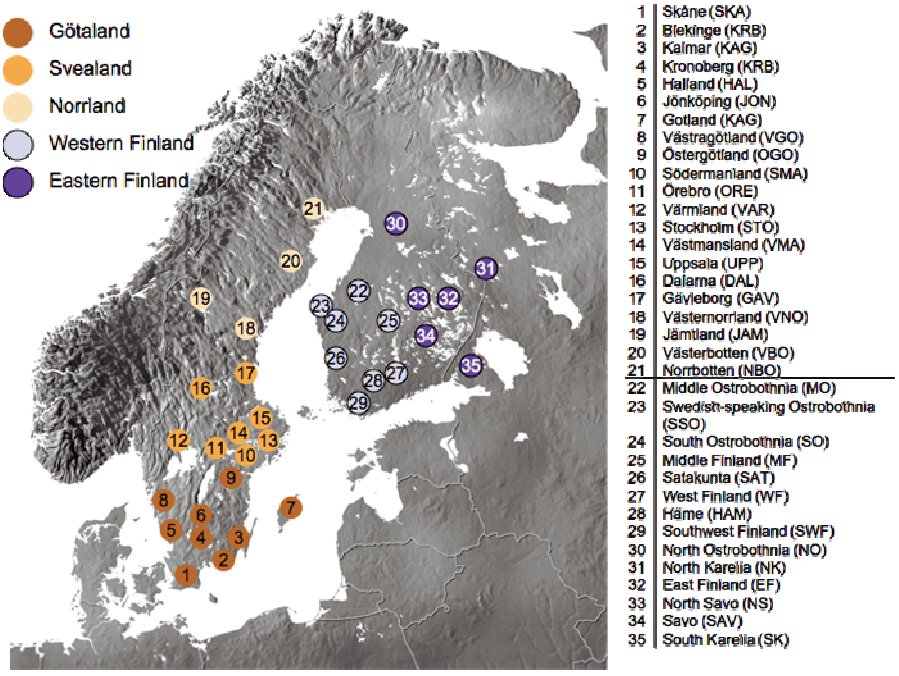

Figure 7. **Geographic location of the Swedish and Finnish population samples.** The individual samples are stratified according to county and region in the analyses, except for the Geneland analyses that use individual coordinates.

suggested that the geographic distribution of the variant is consistent with the Viking trade routes (Lucotte and Dieterlen 2003), while still others show that the variant and the surrounding LD structure do not show signs of selection and are more congruent with neutral evolution (Sabeti et al. 2005; Hedrick and Verrelli 2006).

2.4.1.2 Y-chromosomal and mtDNA markers

In study 4, we investigated the Swedish population by genotyping 14 Y-chromosomal and 35 mitochondrial SNPs on the 2132 newborn screening samples. The Y-chromosomal and mtDNA results confirmed the previously observed connections to the neighbouring countries and to central Europe. As our sampling scheme was non-

selective with regards to ancestry, we could also, with both Y-chromosomal and mtDNA haplogroups detect increased genetic diversity in the big cities that was due to recent immigration.

Genetic and geographical distances were highly correlated for the mtDNA haplogroups, but this was not seen with the Y-chromosomal SNPs. This could be due to the lower resolution for Y-chromosomal SNPs to detect recent migration. Immigration might also have introduced paternal genetic variation that has obscured historical patterns of migration. We also saw signs of bottlenecks, observed as reduced genetic variation in several southern as well as some northern counties.

2.4.1.3 Autosomal SNPs

The 32 autosomal unlinked SNPs genotyped in study 5 on the 2132 newborn screening samples did not detect any substructure within Sweden. As already discussed in chapter 2.1.2.2, we did observe a deficiency of heterozygote genotypes that could be due to the high percentage of individuals with foreign background. However, our simulation studies showed that the same phenomenon could be caused by a heterozygote specific genotyping error. While both Y-chromosomal and mtDNA SNPs indicated a clear contribution of immigrant haplogroups in the big cities, we could not observe anything similar with the autosomal SNPs. As the SNPs used in this study are common they might not detect recent migration patterns.

2.4.2 Finland

Compared to Sweden, the genetic substructure of the Finnish population has been extensively studied. The incidences of nearly 40 monogenic diseases are higher in Finland than elsewhere in the world, indicating strong founder effects during the settlement of Finland. Geographic clustering within Finland for several of the diseases indicate later bottlenecks that have taken place during the settlement of the eastern parts in the 16th century. Several reports have observed a clear east-west division in Finland (Figure 8) that is congruent to the historical political and anthropological borders as well as with settlement history (Norio 2003). In agreement with this, several studies have also observed differences in Y-chromosomal haplotypes between eastern and western Finland (Kittles et al. 1998; Hedman et al. 2004; Lappalainen et al. 2006).

2.4.2.1 Observed east-west duality in Finland

In study 5, we also used the 32 autosomal unlinked SNPs to investigate the population substructure in Finland. The markers were genotyped on 657 samples that were chosen to represent the rural counties of Finland as seen two generations ago (Figure 7). In agreement with previous studies we observed a clear east-west duality as seen on the individual level based on the Geneland algorithm (Figure 9a) and with PCA (Figure 10a) and. A panel of 30 STRs, genotyped on 465 of the samples, confirmed the PCA results (Figure 10b), but these markers were unable to cluster the individuals using Geneland (data not shown). Likewise, the borderline isolation-by-distance seen by the SNPs ($r=0.30$, $P=0.06$) was not supported by the STRs ($r=0.001$, $P>0.2$).

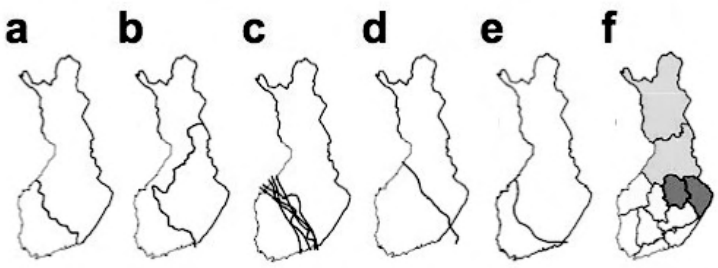

Figure 8. **East-West differences in Finland. a** | Anthropological boundary, **b** | dialect groups, **c** | folkloristic differences, **d** | 1323 national boundary between Sweden and Russia, **e** | eastern boundary of the 4000 year old Battle Axe culture, **f** | mortality differences in coronary heart disease. (Figure adapted from Finnish Disease Heritage II: population prehistory and genetic roots of Finns, Reijo Norio, Human Genetics, 2003)

The SNPs and STRs represent different parts of the frequency spectrum; the SNPs are ascertained for zygosity testing and have minor allele frequencies of 20% or more, while the STRs are chosen based on allele frequencies below 5% in the Finnish population. Therefore, the resolution to capture recent and ancient population genetic events differs between the two panels. Also, the SNPs were genotyped on a larger number of individuals, increasing the power to detect substructure with the SNPs compared to the STRs.

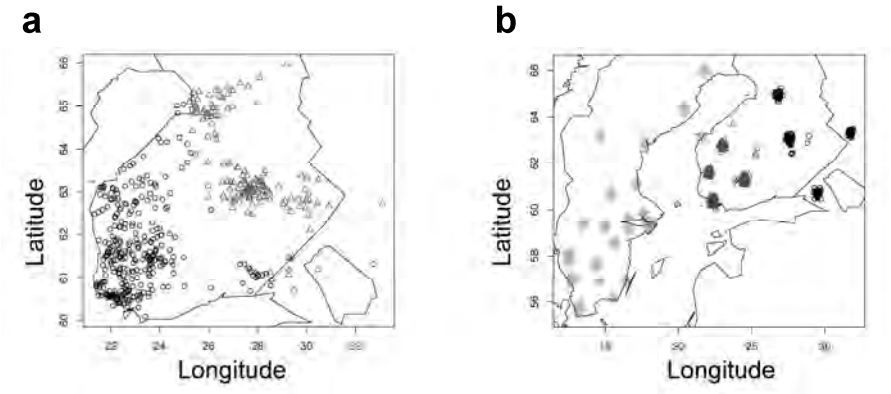

Figure 9. **Cluster analysis using Geneland. a** | A clear east-west duality was observed when the Finnish individuals were clustered using Geneland. **b** | Individuals from the Swedish speaking part of Ostrobothnia clustered with Sweden when a joint-analysis was performed on Swedish and Finnish autosomal genotypes.

2.4.3 Joint analysis of Sweden and Finland

When we analysed the Swedish and Finnish SNP genotypes jointly in Geneland, we observed the same east-western duality within Finland while the Swedish individuals clustered as one (Figure 9b). Similar patterns were observed in the PCA analysis (Figures 10c and 10d). Interestingly, Geneland clustered the Swedish speaking part of Ostrobothnia with Sweden. Swedish settlers inhabited this area in the 13th century, and previous studies have also observed allele frequencies that are in-between Swedish and Finnish frequencies (Virtaranta-Knowles et al. 1991).

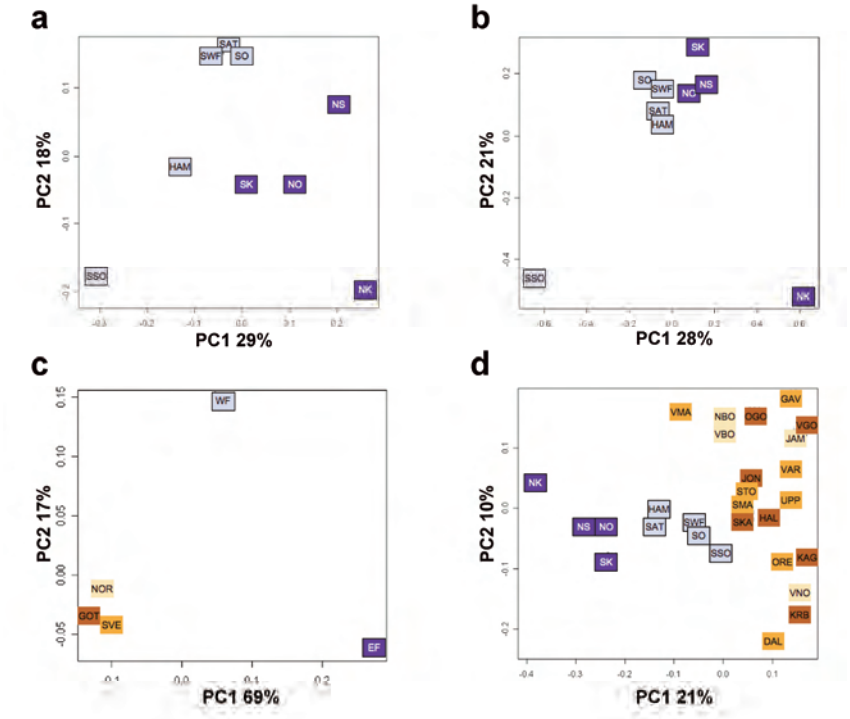

Figure 10. **Principle components analysis a** | Finnish counties analysed based on autosomal SNPs, **b** | Finnish counties analysed based on autosomal STRs, **c** | joint analysis of Finnish and Swedish regions using autosomal SNPs, **d** | joint analysis of Finnish and Swedish counties using autosomal SNPs. The x- and y-axis label correspond to the amount of variation explained by the first two principle components. Colours and county abbreviations correspond to Figure 7.

2.4.4 Conclusions

Taken together, the population substructure in contemporary Sweden seems to lack clear borders. There is evidence of genetic drift and isolation on the regional level that could have been caused by epidemics or war, and patterns of recent immigration is evident in the big cities. In contrast, we confirmed the east-west duality in Finland that is congruent with historical settlement and political borders, and we showed the applicability of including spatial information using the Geneland algorithm in clustering individuals according to their genetic variation.

3 CONCLUSIONS AND FUTURE PROSPECTS

In the previous chapters I have demonstrated how whole genome amplification can facilitate the effective use of biological repositories. The need of very large numbers of individual samples will grow as more advanced study designs are developed that consider multiple interacting factors in conjunction with genome wide association or sequencing. Since it can take decades to collect these cohorts it would be advisable not to only focus on prospectively collecting material but to also investigate the possibility to use the material that already exists in different biobanks around the world.

By making use of these repositories and linked *in silico* registries, it would be possible to design very large studies that could be implemented on rather short notice. These studies could be conducted in a similar manner as we did with the samples from the newborn screening samples; anonymized samples linked to clinical and/or geographical information that is necessary for the study question at hand. Findings from these screenings could then be used to design studies that are more targeted and hypotheses driven.

Preamplification is unfortunately burdened with issues concerning quality, but as long as this is acknowledged it is possible to designing studies that are optimized to discover possible errors. In our study, where we explored the population substructure in Sweden using autosomal SNPs, we simulated these errors to see if the observed results were possible artefacts due to genotyping errors. It is not often that information about quality is reported or used in a systematic way to test the validity of the results, but I argue that this should be done much more often. Most studies and algorithms that do account for errors still consider them to be random, which in some situations might be overly optimistic.

In conclusion, the studies presented in this thesis serve as an example of some applications that are possible by using preamplification. There are some obvious weaknesses inherit in the existing methods, but as long as this is accepted they can be controlled for to generate trustworthy data. Whole genome amplification can optimally be used to facilitate the design and implementation of very large-scale screening projects that use existing repositories of biological material not otherwise suitable for high throughput genetic analyses.

4 SAMMANFATTNING PÅ SVENSKA

Dagens genetikforskning utvecklas med oerhörd hastighet. För mindre än ett årtionde sedan var det möjligt att analysera en genetisk variant åt gången, men idag körs upp till en miljoner reaktioner samtidigt på ett och samma DNA prov. Inom en snar framtid kommer de tre biljoner baspar som utgör människans arvsmassa att analyseras inom loppet av en dag, en bedrift som tog flera år att utföra och kostade över 1000 miljoner kronor för mindre än tio år sedan.

Förutom de tekniska framstegen har även vetskapen om hur olika faktorer påverkar den individuella risken att insjukna i tex diabetes eller astma ökat markant. Man har uppskattat att upptill 100 genvariationer, specifika för olika sjukdomar, tillsammans med miljöfaktorer bildar den totala riskbilden. Varje enskild faktor har följaktligen en relativt liten effekt på den totala risken och för att hitta dessa krävs studier som omfattar tusentals individer.

I Sverige undersöks varje nyfödd för fem olika sällsynta sjukdomar. I samband med detta har det Svenska PKU registret sedan 1975 sparat över tre miljoner individuella blodprov på filterpapper. Givet de omfattande studier som krävs för att karakterisera genetiska riskfaktorer har jag i min avhandling utvecklat en metod som gör det möjligt att utvinna och mångfördubbla DNA från dessa filterpappersprover. Jag visar att upp till 25 år gamla filterpapper kan användas för genetisk forskning. Med hjälp av metoden påvisar jag även att variationer i genen *NPSR1* är kopplade till en ökad risk för nyfödda barn att insjukna i ”respiratory distress syndrome”, den vanligaste livshotande komplikationen hos nyfödda.

Genom mina studier har jag likaså utvecklat en robust panel av genetiska varianter med vilken man kan urskilja identiska tvillingar från icke-identiska (zygositetsbestämning). Tvillingsstudier utgör en viktig del av den genetiska forskningen och exakt information om zygositet är nödvändigt för att undvika felaktigheter i dessa studier. Med hjälp av över 2000 anonyma individuella prover från det Svenska PKU registret samt ytterligare över 600 prover från olika regioner i Finland har jag utforskat den genetiska strukturen i de båda grannländerna. Sverige präglas av genetiska kopplingar till sina grannländer samt centrala Europa. Vissa lokala skillnader i genetisk variation ses i delar av södra samt norra Sverige, möjligen som följd av historiska epidemier, krig och folkförflyttningar. Även karateristiska mönster av de senaste årtiondenas immigration kan observeras i storstäderna.

I Finland har man på basen av genetiska varianter tidigare påvisat en öst-västlig indelning som motsvarar historiska gränser mellan Sverige och Ryssland, befolkningshistoria samt skillnader i dödlighet i olika sjukdomar. Genom att använda mig av ett fåtal genetiska markörer kunde jag påvisa en liknande uppdelning. Då jag analyserade Finland och Sverige tillsammans kunde en genetisk koppling mellan Sverige och individer från svenskspråkiga Österbotten påvisas.

Överlag visar mina resultat på de möjligheter som biobanker likt det Svenska PKU registret utgör för dagens genetikforskning. Genom att använda anonyma prover ur dessa register och koppla ihop dem med databaser som innehåller information om diverse sjukdomar vore det möjligt att på en relativt kort tid samla ihop material för mycket stora studier, något som annars skulle kunna ta upp till ett par årtionden att utföra.

5 ACKNOWLEDGEMENTS

A good book has no ending. -R.D. Cumming

I would like to offer my sincere appreciation to everyone who has contributed, directly or indirectly to this work. I realize now that I should not have waited until a few hours before the thesis will be printed to write the acknowledgements… There are too many that I really would like to thank personally.

First and foremost I am deeply grateful to my main supervisor Professor **Juha Kere** with whom I have had the great pleasure to work for the last seven years. You have been a wonderful mentor and provided me and everyone else in the group with a positive atmosphere that is hard to find anywhere else.

My co-supervisor, Dr **Cecilia Lindgren** for your encouragement and great support during the first years, I am really thankful.

My co-authors who I have had the pleasure to collaborate with over the years. Docent **Ulrika Von Döbeln** and the personnel at the Swedish newborn screening registry, Docent **Gunnel Tybring, Loreana Gherman** and **Camilla Lagerberg** at the Swedish Biobank, **Ville Pulkkinen** and **Ritva Haataja** and the RDS-team, **Päivi Lahermo, Tuuli Lappalainen, Elina Salmela** and **Gilles Guillot** for teaching me about population genetics and patiently revising the final manuscript. My co-authors from within the group, **Erik Melén**, **Marco Zucchelli**, **Ville-Veikko Mäkelä** and **Astrid Fungmark** and everyone who has taken the time to read or listen to my more or less scientific theories.

All former and present group members at CBT and MAF. There are too many to mention by name, but it has been great to work with you all, I really really mean it. You have all contributed to the fantastic atmosphere that I have had the pleasure to enjoy all these years. Thank you!

All friends here in Sweden and all over the world for the golf, beer, heavy lifting, violent computer games and entrepreneurial discussions. I have enjoyed all of the above-mentioned activities immensely and hope to continue doing so with all of you for many years to come.

My Armenian sunshine, Gayane and her equally sunny family and friends.

My mother, father and brother for your support and love.

And once again to all of you for making this thesis possible! If I ever write another best seller I will make sure to start with the acknowledgements…

This work was supported by Karolinska Institutet and the Swedish Research Council.

6 REFERENCES

My sources are unreliable, but their information is fascinating. – Ashleigh Brilliant

(1992) A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group. Science 258(5079): 67-86.

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

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

(2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145): 661-678.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2001) GRR: graphical representation of relationship errors. Bioinformatics 17(8): 742-743.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30(1): 97-101.

Aitken RJ, Marshall Graves JA (2002) The future of sex. Nature 415(6875): 963.

Alanne M, Salomaa V, Saarela J, Peltonen L, Perola M (2004) DNA extraction yield is associated with several phenotypic characteristics: results from two large population surveys. J Thromb Haemost 2(11): 2069-2071.

Allocco DJ, Song Q, Gibbons GH, Ramoni MF, Kohane IS (2007) Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. BMC Genomics 8: 68.

Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66(6): 1933-1944.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40(3): 340-345.

Becker KG (2004) The common variants/multiple disease hypothesis of common complex genetic disorders. Med Hypotheses 62(2): 309-317.

Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S et al. (2007) The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3(6): e104.

Bereczky S, Martensson A, Gil JP, Farnert A (2005) Short report: Rapid DNA extraction from archive blood spots on filter paper for genotyping of Plasmodium falciparum. Am J Trop Med Hyg 72(3): 249-251.

Bergen AW, Qi Y, Haque KA, Welch RA, Chanock SJ (2005) Effects of DNA mass on multiple displacement whole genome amplification and genotyping performance. BMC Biotechnol 5: 24.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6): 1111-1120.

Berthier-Schaad Y, Kao WH, Coresh J, Zhang L, Ingersoll RG et al. (2007) Reliability of high-throughput genotyping of whole genome amplified DNA in SNP genotyping studies. Electrophoresis 28(16): 2812-2817.

Bird A (2007) Perceptions of epigenetics. Nature 447(7143): 396-398.

Boomsma D, Busjahn A, Peltonen L (2002) Classical twin studies and beyond. Nat Rev Genet 3(11): 872-882.

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S et al. (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82(3): 763-771.

Buchanan AV, Risch GM, Robichaux M, Sherry ST, Batzer MA et al. (2000) Long DOP-PCR of rare archival anthropological samples. Hum Biol 72(6): 911-925.

Chaisomchit S, Wichajarn R, Janejai N, Chareonsiriwatana W (2005) Stability of genomic DNA in dried blood spots stored on filter paper. Southeast Asian J Trop Med Public Health 36(1): 270-273.

Cheung VG, Nelson SF (1996) Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. Proc Natl Acad Sci U S A 93(25): 14676-14679.

Clamp M, Fry B, Kamal M, Xie X, Cuff J et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104(49): 19428-19433.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15(11): 1496-1502.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD et al. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37(11): 1243-1246.

Consortium. IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011): 931-945.

Coulon A, Guillot G, Cosson JF, Angibault JM, Aulagnier S et al. (2006) Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. Mol Ecol 15(6): 1669-1679.

Cox DG, Kraft P (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. Hum Hered 61(1): 10-14.

Currat M, Excoffier L, Maddison W, Otto SP, Ray N et al. (2006) Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science 313(5784): 172; author reply 172.

D'Amato G, Liccardi G, D'Amato M, Holgate S (2005) Environmental risk factors and allergic bronchial asthma. Clin Exp Allergy 35(9): 1113-1124.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29(2): 229-232.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 99(8): 5261-5266.

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55(4): 997-1004.

Devlin B, Roeder K, Bacanu SA (2001) Unbiased methods for population-based association studies. Genet Epidemiol 21(4): 273-284.

Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21(11): 596-601.

Dietmaier W, Hartmann A, Wallinger S, Heinmoller E, Kerner T et al. (1999) Multiple mutation analyses in single tumor cells with improved whole genome amplification. Am J Pathol 154(1): 83-95.

Dixon LA, Dobbins AE, Pulker HK, Butler JM, Vallone PM et al. (2006) Analysis of artificially degraded DNA using STRs and SNPs--results of a collaborative European (EDNAP) exercise. Forensic Sci Int 164(1): 33-44.

Einarsdottir E, Egerbladh I, Beckman L, Holmberg D, Escher SA (2007) The genetic population structure of northern Sweden and its implications for mapping genetic diseases. Hereditas 144(5): 171-180.

Eising S, Svensson J, Skogstrand K, Nilsson A, Lynch K et al. (2007) Type 1 diabetes risk analysis on dried blood spot samples from population-based newborns: design and feasibility of an unselected case-control study. Paediatr Perinat Epidemiol 21(6): 507-517.

Epstein MP, Duren WL, Boehnke M (2000) Improved inference of relationship for pairs of individuals. Am J Hum Genet 67(5): 1219-1231.

Evans PD, Mekel-Bobrov N, Vallender EJ, Hudson RR, Lahn BT (2006) Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage. Proc Natl Acad Sci U S A 103(48): 18178-18183.

Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR et al. (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309(5741): 1717-1720.

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet 7(10): 745-758.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131(2): 479-491.

Falush D, Stephens M, Pritchard JK (2003a) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4): 1567-1587.

Falush D, Wirth T, Linz B, Pritchard JK, Stephens M et al. (2003b) Traces of human migrations in Helicobacter pylori populations. Science 299(5612): 1582-1585.

Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 Spec No 1: R57-66.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164): 851-861.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J et al. (2002) The structure of haplotype blocks in the human genome. Science 296(5576): 2225-2229.

Galvani AP, Slatkin M (2003) Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. Proc Natl Acad Sci U S A 100(25): 15276-15279.

Gilbert MT, Tomsho LP, Rendulic S, Packard M, Drautz DI et al. (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science 317(5846): 1927-1930.

Gordon D, Finch SJ (2005) Factors affecting statistical power in the detection of genetic association. J Clin Invest 115(6): 1408-1418.

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82(1): 100-112.

Graves JA (2006) Sex chromosome specialization and degeneration in mammals. Cell 124(5): 901-914.

Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. Genetics 170(3): 1261-1280.

Guo CY, Cupples LA, Yang Q (2008) Testing informative missingness in genetic studies using case-parent triads. Eur J Hum Genet.

Guthery SL, Salisbury BA, Pungliya MS, Stephens JC, Bamshad M (2007) The structure of common genetic variation in United States populations. Am J Hum Genet 81(6): 1221-1231.

Guthrie R (1992) The origin of newborn screening. Screening 1: 5-15.

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C et al. (1994) The 1993-94 Genethon human genetic linkage map. Nat Genet 7(2 Spec No): 246-339.

Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE et al. (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448(7153): 591-594.

Hallman M, Haataja R (2006) Surfactant protein polymorphisms and neonatal lung disease. Semin Perinatol 30(6): 350-361.

Hankins GV, Saade GR (2005) Factors influencing twins and zygosity. Paediatr Perinat Epidemiol 19 Suppl 1: 8-9.

Hanson EK, Ballantyne J (2005) Whole genome amplification strategy for forensic genetic analysis using single or few cell equivalents of genomic DNA. Anal Biochem 346(2): 246-257.

Hao K, Cawley S (2007) Differential dropout among SNP genotypes and impacts on association tests. Hum Hered 63(3-4): 219-228.

Hardy GH (1908) Mendelian Proportions in a Mixed Population. Science 28(706): 49-50.

Hawks J, Cochran G, Harpending HC, Lahn BT (2008) A genetic legacy from archaic Homo. Trends Genet 24(1): 19-23.

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 Sci Int 142(1): 37-43.

Hedrick PW, Verrelli BC (2006) "Ground truth" for selection on CCR5-Delta32. Trends Genet 22(6): 293-296.

Helgason A, Sigureth ardottir S, Gulcher JR, Ward R, Stefansson K (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. Am J Hum Genet 66(3): 999-1016.

Hittelman A, Sridharan S, Roy R, Fridlyand J, Loda M et al. (2007) Evaluation of whole genome amplification protocols for array and oligonucleotide CGH. Diagn Mol Pathol 16(4): 198-206.

Hollegaard MV, Sorensen KM, Petersen HK, Arnardottir MB, Norgaard-Pedersen B et al. (2007) Whole genome amplification and genetic analysis after extraction of proteins from dried blood spots. Clin Chem 53(6): 1161-1162.

Holmlund G, Nilsson H, Karlsson A, Lindblom B (2006) Y-chromosome STR haplotypes in Sweden. Forensic Sci Int 160(1): 66-79.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29(3): 306-309.

Jackson RW, Snieder H, Davis H, Treiber FA (2001) Determination of twin zygosity: a comparison of DNA with various questionnaire indices. Twin Res 4(1): 12-18.

Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. Nature 316(6023): 76-79.

Jobe AH, Bancalari E (2001) Bronchopulmonary dysplasia. Am J Respir Crit Care Med 163(7): 1723-1729.

Karlsson AO, Wallerstrom T, Gotherstrom A, Holmlund G (2006) Y-chromosome diversity in Sweden - a long-time perspective. Eur J Hum Genet 14(8): 963-970.

Kittler R, Stoneking M, Kayser M (2002) A whole genome amplification method to generate long fragments from low quantities of genomic DNA. Anal Biochem 300(2): 237-244.

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA et al. (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62(5): 1171-1179.

Kline MC, Duewer DL, Redman JW, Butler JM, Boyer DA (2002) Polymerase chain reaction amplification of DNA from aged blood stains: quantitative evaluation of the "suitability for purpose" of four filter papers as archival media. Anal Chem 74(8): 1863-1869.

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43(4): 520-526.

Kormann MS, Carr D, Klopp N, Illig T, Leupold W et al. (2005) G-Protein-coupled receptor polymorphisms are associated with asthma in a large German population. Am J Respir Crit Care Med 171(12): 1358-1362.

Krishnan KJ, Greaves LC, Reeve AK, Turnbull D (2007) The ageing mitochondrial genome. Nucleic Acids Res 35(22): 7399-7405.

Kruglyak L (2008) The road to genome-wide association studies. Nat Rev Genet 9(4): 314-318.

Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A 99(2): 803-808.

Kuukasjarvi T, Tanner M, Pennanen S, Karhu R, Visakorpi T et al. (1997) Optimizing DOP-PCR for universal amplification of small DNA samples in comparative genomic hybridization. Genes Chromosomes Cancer 18(2): 94-101.

Laitinen T, Polvi A, Rydman P, Vendelin J, Pulkkinen V et al. (2004) Characterization of a common susceptibility locus for asthma-related traits. Science 304(5668): 300-304.

Lander ES (1996) The new genomics: global views of biology. Science 274(5287): 536-539.

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265(5181): 2037-2048.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al. (2001) Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.

Lappalainen T, Koivumaki S, Salmela E, Huoponen K, Sistonen P et al. (2006) Regional differences among the Finns: a Y-chromosomal perspective. Gene 376(2): 207-215.

Lasken RS, Stockwell TB (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol 7: 19.

Lasky-Su J, Lyon HN, Emilsson V, Heid IM, Molony C et al. (2008) On the replication of genetic associations: timing can be everything! Am J Hum Genet 82(4): 849-858.

Latch EK, Scognamillo DG, Fike JA, Chamberlain MJ, Rhodes OE, Jr. (2008) Deciphering ecological barriers to North American river otter (Lontra canadensis) gene flow in the Louisiana landscape. J Hered 99(3): 265-274.

Leal SM (2005) Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet Epidemiol 29(3): 204-214.

Leanza SM, Burk RD, Rohan TE (2007) Whole genome amplification of DNA extracted from hair samples: potential for use in molecular epidemiologic studies. Cancer Detect Prev 31(6): 480-488.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866): 1100-1104.

Libert F, Cochaux P, Beckman G, Samson M, Aksenova M et al. (1998) The deltaccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. Hum Mol Genet 7(3): 399-406.

Lichtenstein P, De Faire U, Floderus B, Svartengren M, Svedberg P et al. (2002) The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. J Intern Med 252(3): 184-205.

Lichtenstein P, Sullivan PF, Cnattingius S, Gatz M, Johansson S et al. (2006) The Swedish Twin Registry in the third millennium: an update. Twin Res Hum Genet 9(6): 875-882.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069): 803-819.

Little SE, Vuononvirta R, Reis-Filho JS, Natrajan R, Iravani M et al. (2006) Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. Genomics 87(2): 298-306.

Liu N, Beerman I, Lifton R, Zhao H (2006a) Haplotype analysis in the presence of informatively missing genotype data. Genet Epidemiol 30(4): 290-300.

Liu W, Zhao W, Chase GA (2006b) The impact of missing and erroneous genotypes on tagging SNP selection and power of subsequent association tests. Hum Hered 61(1): 31-44.

Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC et al. (1998) Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet 19(3): 225-232.

Lohmueller KE, Mauney MM, Reich D, Braverman JM (2006) Variants associated with common disease are not unusually differentiated in frequency across populations. Am J Hum Genet 78(1): 130-136.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33(2): 177-182.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. Nature 451(7181): 994-997.

Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S, Syvanen AC (2003) Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. Nucleic Acids Res 31(21): e129.

Lucotte G, Dieterlen F (2003) More about the Viking hypothesis of origin of the delta32 mutation in the CCR5 gene conferring resistance to HIV-1 infection. Infect Genet Evol 3(4): 293-295.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D et al. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308(5724): 1034-1036.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5): 356-369.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328): 652-654.

McGinnis R, Shifman S, Darvasi A (2002) Power and efficiency of the TDT and case-control design for association scans. Behav Genet 32(2): 135-144.

McKusick VA (2008) Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

Mekel-Bobrov N, Posthuma D, Gilbert SL, Lind P, Gosso MF et al. (2007) The ongoing adaptive evolution of ASPM and Microcephalin is not explained by increased intelligence. Hum Mol Genet 16(6): 600-608.

Melen E, Bruce S, Doekes G, Kabesch M, Laitinen T et al. (2005) Haplotypes of G protein-coupled receptor 154 are associated with childhood allergy and asthma. Am J Respir Crit Care Med 171(10): 1089-1095.

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201(4358): 786-792.

Misra A, Ganda OP (2007) Migration and its impact on adiposity and type 2 diabetes. Nutrition 23(9): 696-708.

Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448(7152): 470-473.

Montgomery GW, Campbell MJ, Dickson P, Herbert S, Siemering K et al. (2005) Estimation of the rate of SNP genotyping errors from DNA extracted from different tissues. Twin Res Hum Genet 8(4): 346-352.

Morton NE, Collins A (1998) Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci U S A 95(19): 11389-11393.

Moskvina V, Schmidt KM (2006) Susceptibility of biallelic haplotype and genotype frequencies to genotyping error. Biometrics 62(4): 1116-1123.

Munafo M (2004) Replication validity of genetic association studies of smoking behavior: what can meta-analytic techniques offer? Nicotine Tob Res 6(2): 381-382.

Myles S, Tang K, Somel M, Green RE, Kelso J et al. (2008) Identification and analysis of genomic regions with large between-population differentiation in humans. Ann Hum Genet 72(Pt 1): 99-110.

Nicodemus KK, Luna A, Shugart YY (2007) An evaluation of power and type I error of single-nucleotide polymorphism transmission/disequilibrium-based statistical methods under different family structures, missing parental data, and population stratification. Am J Hum Genet 80(1): 178-185.

Norio R (2003) Finnish Disease Heritage II: population prehistory and genetic roots of Finns. Hum Genet 112(5-6): 457-469.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40(5): 646-649.

Nyholt DR (2006) On the probability of dizygotic twins being concordant for two alleles at multiple polymorphic loci. Twin Res Hum Genet 9(2): 194-197.

O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63(1): 259-266.

Obermann EC, Junker K, Stoehr R, Dietmaier W, Zaak D et al. (2003) Frequent genetic alterations in flat urothelial hyperplasias and concomitant papillary bladder cancer as detected by CGH, LOH, and FISH analyses. J Pathol 199(1): 50-57.

Ollier W, Sprosen T, Peakman T (2005) UK Biobank: from concept to reality. Pharmacogenomics 6(6): 639-646.

Olshan AF (2007) Meeting report: the use of newborn blood spots in environmental research: opportunities and challenges. Environ Health Perspect 115(12): 1767-1779.

Ooki S, Yokoyama Y, Asaka A (2004) Zygosity misclassification of twins at birth in Japan. Twin Res 7(3): 228-232.

Pachot A, Barbalat V, Marotte H, Diasparra J, Gouraud A et al. (2007) A rapid semi automated method for DNA extraction from dried-blood spots: application to the HLA-DR shared epitope analysis in rheumatoid arthritis. J Immunol Methods 328(1-2): 220-225.

Pakstis AJ, Speed WC, Kidd JR, Kidd KK (2007) Candidate SNPs for a universal individual identification panel. Hum Genet 121(3-4): 305-317.

Palmer LJ (2007) UK Biobank: bank on it. Lancet 369(9578): 1980-1982.

Paradies YC, Montoya MJ, Fullerton SM (2007) Racialized genetics and the study of complex diseases: the thrifty genotype revisited. Perspect Biol Med 50(2): 203-227.

Park JW, Beaty TH, Boyce P, Scott AF, McIntosh I (2005) Comparing whole-genome amplification methods and sources of biological samples for single-nucleotide polymorphism genotyping. Clin Chem 51(8): 1520-1523.

Paynter RA, Skibola DR, Skibola CF, Buffler PA, Wiemels JL et al. (2006) Accuracy of multiplexed Illumina platform-based single-nucleotide polymorphism genotyping compared between genomic and whole genome amplified DNA collected from multiple sources. Cancer Epidemiol Biomarkers Prev 15(12): 2533-2536.

Pearson TA, Manolio TA (2008) How to interpret a genome-wide association study. Jama 299(11): 1335-1344.

Pennisi E (2007) Genetics. Working the (gene count) numbers: finally, a firm answer? Science 316(5828): 1113.

Petkovski E, Keyser-Tracqui C, Hienne R, Ludes B (2005) SNPs and MALDI-TOF MS: tools for DNA typing in forensic paternity testing and anthropology. J Forensic Sci 50(3): 535-541.

Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR et al. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7: 216.

Plagnol V, Cooper JD, Todd JA, Clayton DG (2007) A method to address differential bias in genotyping in large-scale association studies. PLoS Genet 3(5): e74.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6(11): 847-859.

Ponting C, Jackson AP (2005) Evolution of primary microcephaly genes and the enlargement of primate brains. Curr Opin Genet Dev 15(3): 241-248.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8): 904-909.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69(1): 124-137.

Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theor Popul Biol 60(3): 227-237.

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11(20): 2417-2423.

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155(2): 945-959.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67(1): 170-181.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A 102(44): 15942-15947.

Ray N, Currat M, Berthier P, Excoffier L (2005) Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. Genome Res 15(8): 1161-1167.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH et al. (2006) Global variation in copy number in the human genome. Nature 444(7118): 444-454.

Reed T, Plassman BL, Tanner CM, Dick DM, Rinehart SA et al. (2005) Verification of self-report of zygosity determined via DNA testing in a subset of the NAS-NRC twin registry 40 years later. Twin Res Hum Genet 8(4): 362-367.

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17(9): 502-510.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC et al. (2001) Linkage disequilibrium in the human genome. Nature 411(6834): 199-204.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281): 1516-1517.

Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D et al. (1999) A genomic screen of autism: evidence for a multilocus etiology. Am J Hum Genet 65(2): 493-507.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK et al. (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1(6): e70.

Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. J Hered 98(4): 331-336.

Rowe G, Beebee TJ (2007) Defining population boundaries: use of three Bayesian approaches with microsatellite data from British natterjack toads (Bufo calamita). Mol Ecol 16(4): 785-796.

Ryckman K, Williams SM (2008) Calculation and use of the Hardy-Weinberg model in association studies. Curr Protoc Hum Genet Chapter 1: Unit 1 18.

Rylander-Rudqvist T, Hakansson N, Tybring G, Wolk A (2006) Quality and quantity of saliva DNA obtained from the self-administrated oragene method--a pilot study on the cohort of Swedish men. Cancer Epidemiol Biomarkers Prev 15(9): 1742-1745.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P et al. (2006) Positive natural selection in the human lineage. Science 312(5780): 1614-1620.

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B et al. (2005) The case for selection at CCR5-Delta32. PLoS Biol 3(11): e378.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316(5829): 1331-1336.

Schapira AH (2008) Mitochondria in the aetiology and pathogenesis of Parkinson's disease. Lancet Neurol 7(1): 97-109.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316(5829): 1341-1345.

Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J et al. (2006) European population substructure: clustering of northern and southern populations. PLoS Genet 2(9): e143.

Sieberts SK, Wijsman EM, Thompson EA (2002) Relationship inference from trios of individuals, in the presence of typing error. Am J Hum Genet 70(1): 170-180.

Silander K, Saarela J (2008) Whole genome amplification with phi29 DNA polymerase to enable genetic or genomic analysis of samples of low DNA yield. Methods Mol Biol 439: 1-18.

Sjoholm MI, Dillner J, Carlson J (2007) Assessing quality and functionality of DNA from fresh and archival dried blood spots and recommendations for quality control guidelines. Clin Chem 53(8): 1401-1407.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130): 881-885.

Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23(1): 23-35.

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet 70(2): 496-508.

Soficaru A, Dobos A, Trinkaus E (2006) Early modern humans from the Pestera Muierii, Baia de Fier, Romania. Proc Natl Acad Sci U S A 103(46): 17196-17201.

Sorensen KM, Jespersgaard C, Vuust J, Hougaard D, Norgaard-Pedersen B et al. (2007) Whole genome amplification on DNA from filter paper blood spot samples: an evaluation of selected systems. Genet Test 11(1): 65-71.

Steinberg K, Beck J, Nickerson D, Garcia-Closas M, Gallagher M et al. (2002) DNA banking for epidemiologic studies: a review of current practices. Epidemiology 13(3): 246-254.

Stenseth NC, Atshabar BB, Begon M, Belmain SR, Bertherat E et al. (2008) Plague: past, present, and future. PLoS Med 5(1): e3.

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW et al. (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62(6): 1507-1515.

Stephens K, Kayes L, Riccardi VM, Rising M, Sybert VP et al. (1992) Preferential mutation of the neurofibromatosis type 1 gene in paternally derived chromosomes. Hum Genet 88(3): 279-282.

Stringer C (2002) Modern human origins: progress and prospects. Philos Trans R Soc Lond B Biol Sci 357(1420): 563-579.

Sun G, Kaushal R, Pal P, Wolujewicz M, Smelser D et al. (2005) Whole-genome amplification: relative efficiencies of the current methods. Leg Med (Tokyo) 7(5): 279-286.

Sun YV, Kardia SL (2008) Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. Eur J Hum Genet 16(4): 487-495.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3): 585-595.

Taylor RW, Turnbull DM (2005) Mitochondrial DNA mutations in human disease. Nat Rev Genet 6(5): 389-402.

Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA et al. (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics 13(3): 718-725.

Templeton AR (2007) Genetics and recent human evolution. Evolution Int J Org Evolution 61(7): 1507-1519.

Teo YY, Fry AE, Clark TG, Tai ES, Seielstad M (2007) On the usage of HWE for identifying genotyping errors. Ann Hum Genet 71(Pt 5): 701-703; author reply 704.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39(7): 857-864.

Turakulov R, Easteal S (2003) Number of SNPS loci needed to detect population structure. Hum Hered 55(1): 37-45.

Utsuno H, Minaguchi K (2004) Influence of template DNA degradation on the genotyping of SNPs and STR polymorphisms from forensic materials by PCR. Bull Tokyo Dent Coll 45(1): 33-46.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al. (2001) The sequence of the human genome. Science 291(5507): 1304-1351.

Virtaranta-Knowles K, Sistonen P, Nevanlinna HR (1991) A population genetic study in Finland: comparison of the Finnish- and Swedish-speaking populations. Hum Hered 41(4): 248-264.

Voight BF, Pritchard JK (2005) Confounding from cryptic relatedness in case-control association studies. PLoS Genet 1(3): e32.

von Bubnoff A (2008) Next-generation sequencing: the race is on. Cell 132(5): 721-723.

Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103(1): 135-140.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome Res 15(11): 1468-1476.

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J et al. (1992) A second-generation linkage map of the human genome. Nature 359(6398): 794-801.

Woods RP, Freimer NB, De Young JA, Fears SC, Sicotte NL et al. (2006) Normal variants of Microcephalin and ASPM do not account for brain size variability. Hum Mol Genet 15(12): 2025-2029.

Yu Z, Schaid DJ (2007) Methods to impute missing genotypes for population data. Hum Genet 122(5): 495-504.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316(5829): 1336-1341.

Zhang J, Pare PD, Sandford AJ (2008) Recent advances in asthma genetics. Respir Res 9: 4.

Zhang L, Cui X, Schmitt K, Hubert R, Navidi W et al. (1992) Whole genome amplification from a single cell: implications for genetic analysis. Proc Natl Acad Sci U S A 89(13): 5847-5851.

Zou GY, Donner A (2006) The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. Ann Hum Genet 70(Pt 6): 923-933.