
THE UNIVERSITY *of* EDINBURGH

# Recent identity by descent in human genetic data – methods and applications

THE UNIVERSITY
*of* EDINBURGH

Dominik Glodzik

Institute for Genetic and Molecular Medicine

University of Edinburgh

A thesis submitted for the degree of

*Doctor of Philosophy*

2013

The thesis has been composed entirely by me.
The thesis presents my own work unless where I obtained help from group members, as acknowledged.
The work presented in this thesis has not been submitted for any other degree or professional qualification.

Dominik Glodzik

I dedicate this thesis to my loving parents and family,
to Cici,
to all friends I met in Edinburgh
and to the inspiring city.

# Acknowledgements

Most of the work presented in the second and third chapters was published as an article (Glodzik et al., 2013). I developed and implemented the algorithm, applied it to data and wrote the article. I obtained help on data pre-processing, as acknowledged later. A re-sequencing study that used my algorithms was also described in a published article (Joshi et al., 2013).

I would like to thank my supervisor Paul McKeigue and all kind friends who helped me by reading parts of this thesis.

# Abstract

The thesis describes algorithms for detecting regions of recent identity by descent (IBD) from human genetic data and its applications in optimising resequencing studies, genomic predictions and detecting Mendelian subtypes of diseases.

Firstly, we describe the algorithm ANCHAP, which scans pairs of multi-point SNP genotypes for sharing IBD of long haplotypes. A comparison with other methods shows that ANCHAP outperforms them in terms of speed or accuracy. We demonstrate the algorithm on data from population isolates - from Orcades, Croatian islands, and from a population of unrelated individuals. We compare the abundance of IBD segments between cohorts, and identify genetic regions where IBD is most common.

Secondly, we verify the IBD regions detected from array data against exome sequence data. We estimate that where sharing IBD between a pair of individuals is inferred, this is confirmed by exome data in 98% of cases. Correctness of IBD detection varies with settings of ANCHAP, length of IBD segments, and position with respect to segment endpoints. We find that with sample sizes of 1000 individuals from an isolated population genotyped using a dense SNP array, and with 20% of these individuals sequenced, 65% of sequences of the unsequenced subjects can be partially inferred. Implementation of such resequencing strategies requires an IBD-based imputation algorithm, which is outlined.

Thirdly, we use recent IBD to detect carriers of Mendelian subtypes of colon cancer. We show this with the example of Lynch syndrome, which accounts for about 3% of colon cancer patients. We detect IBD

sharing between known and unknown carriers around DNA mismatch-repair genes. Using the IBD relationship, we build and evaluate a model that predicts presence of Lynch Syndrome mutations.

Finally, we discuss whether regions of identity by descent can be used for genomic predictions. We conclude that the utility of the inferred IBD regions depends on accuracy of detection, time to most recent common ancestors and mutation rates since.

# Contents

# List of Figures

# Abbreviations

CC colorectal cancer

GWAS genome-wide association study

HMM hidden Markov model

IBD identity by descent, identical-by-descent

LS Lynch syndrome

MMR mismatch-repair genes

MRCA most recent common ancestor

SNP single nucleotide polymorphism

# Chapter 1

# Introduction

## 1.1 Motivation and aims

We first present an idea that connects the parts of the thesis, and then outline the scientific, technological and historical background behind it.

The idea behind the work presented in this thesis is to tag rare variants in genetic data through long-range haplotypes. Rare genetic variants could explain some of the missing heritability of diseases and quantitative traits. Due to allelic heterogeneity, single nucleotide polymorphisms (SNPs) may not tag rare variants, however the rare variants would be likely captured by long-range haplotypes. Long identical-by-descent (IBD) haplotypes are more likely to occur in isolated populations, where due to geographical isolation and constrained population size any two individuals are likely to have a recent ancestor. Resulting from the relatedness are long haplotypes shared IBD, which can be recovered from SNP data by algorithms.

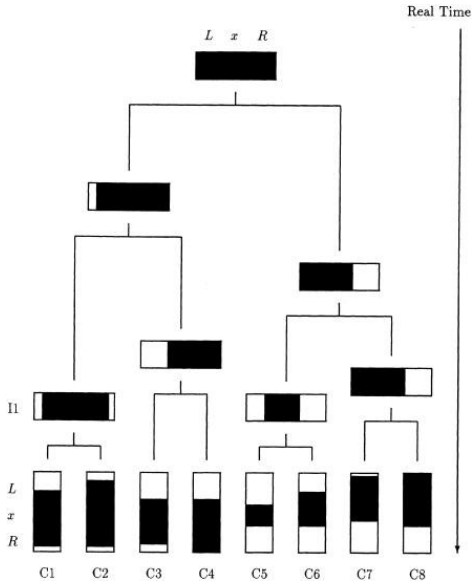



Figure 1.1: An ancestral haplotype shared IBD in current generation. The ancestral haplotype, depicted in black, was shortened by recombination events as it was passed through generations towards bottom of the figure. Where cross-overs did not occur, current generation of individuals (C1-C8) share the haplotype IBD around locus $x$. Source: (Morris et al., 2002)

Genetically isolated populations result from "the founder effect of a small number of individuals as a consequence of some type of bottleneck. They exist in geographical, cultural, or geographical and cultural isolation over many generations without genetic interchange from other subpopulations" (Arcos-Burgos and Muenke, 2002). In isolated populations most individuals share relatively recent common ancestors. If more than one individual inherited the same ancient haplotype in a region, we call them haplotype sharers. Segments of their chromosomes are identical by descent - their haplotypes "descend from a common ancestor without either of them experiencing a recombination" (Powell et al., 2010). The concept of sharing IBD of an ancestral haplotype is depicted in Figure 1.1. While any two copies of an allele are identical by descent with a recent common ancestor in the remote past, where longer stretches have been inherited, the common ancestor is likely to have been more recent. Although there is a continuum between recent and ancient sharing, we distinguish recent IBD from the ancient one based on length of segments shared IBD. While SNP arrays do not directly reveal

gametic phase, haplotype sharers can be identified using computational methods.

Applications of inference of regions of identity by descent include:

1. mapping genetic effects on complex traits,
2. optimization of resequencing studies
3. genomic predictions,
4. studying Mendelian subtypes of diseases.

In the first application, inference of shared haplotypes may make it possible to detect the effects of genes in which functional variants that are rare in the general population have drifted to high frequency in the isolated population. Furthermore, the reduced allelic heterogeneity in an isolate provides an opportunity to detect associations with these otherwise rare variants. In an outbred population conventional GWAS studies may fail to detect associations with rare variants, as these may not be in linkage disequilibrium with SNPs on genotyping arrays that have been optimized to tag common variants (Johnson et al., 2001; Terwilliger and Weiss, 1998). In genetically isolated populations like Iceland or Finland the linkage disequilibrium may be higher. Examples of diseases which were linked to genetic loci in studies of isolates include: myocardial infarction, stroke, type 2 diabetes, atrial fibrillation, prostate cancer, schizophrenia and asthma (Kristiansson et al., 2008). The uncovered ancestral haplotypes can be in even stronger association with rare functional variants and hence improve the power of association tests. The most ambitious attempts to map effects of shared haplotypes reconstruct descent trees, but this approach has been found computationally infeasible (Morris et al., 2002).

In the second application, when SNP genotypes are available and next-generation sequencing is planned, exploiting haplotype sharing between individuals can save resources. With sharing inferred from SNP genotypes, it is possible to choose a minimally redundant subset of individuals to be sequenced, and then to impute sequence data into other subjects with SNP genotype data.

In the third application, IBD segments could be used in models for genomic predictions. They should be most effective when traits being predicted are affected by rare genetic variants.

In the fourth application, IBD segments could be used for studying Mendelian subtypes of diseases. Around each causal mutation the carriers are likely to share haplotypes even in out-bred populations, and depending on the time to the recent common ancestor the shared segments could be long. Carriers of the same Mendelian subtypes could be identified through IBD analysis on SNP data. Identifying carriers of Mendelian subtypes of diseases is a real clinical problem, for example in colon cancer, where patients who carry Mendelian forms of the disease are often not correctly diagnosed and accordingly do not receive the most appropriate treatment.

### 1.1.1 Aims

The aims of the project were:

- to develop an algorithm for detecting recent identity by descent from SNP data,
- to compare it against existing methods,
- to use detected IBD haplotypes for optimisation of resequencing study in an isolated population,
- to develop and test a method for detecting patients of carrying Mendelian subtypes of diseases, such as Lynch syndrome in colon cancer,
- to evaluate utility of IBD segments for genomic predictions.

## 1.2 Genotyping and sequencing technology

Recent developments in genotyping and sequencing technologies present exciting opportunities for understanding genetic cause of diseases, and for verifying population-genetics models against large-scale data sets. SNP genotyping is gradually giving way to next-generation sequencing technologies, which are being employed in increasingly larger studies.

### 1.2.1 SNP arrays

The HapMap project (Consortium, 2007) characterised single nucleotide polymorphisms of 270 individuals from African, Asian and European ancestry. Identified and verified were 3.1 million variants with frequency at least 0.05 (Stranger et al., 2011). Once identified, the SNPs are used to design genotyping arrays. Primers for the SNPs are bound to a chip, such that SNPs can be identified by their positions on a plate (Syvänen, 2001). Segmented and amplified fragments of a DNA segments hybridise to the primers, after which fluorescent-marked nucleotides attach to sample fragments, and different nucleotides are marked with different colours. Decoding genotypes at SNPs on an array involves analysis of an image, in which different colours of light denote different nucleotides.

### 1.2.2 Next-generation sequencing

The development of sequencing technology led to discovery of 10 million new variants in the 1000 Genomes project (Consortium, 2012). Next-generation sequencing, irrespective of specific technology, consists of template preparation, sequencing and imaging (Metzker, 2009; Nekrutenko and Taylor, 2012). Firstly, a sample of DNA is fragmented to produce templates. These are immobilised to a solid surface or support, and clonally amplified through polymerase chain reaction. Many clusters of identical single-stranded templates are produced, and are gathered close together to facilitate sequencing. Sequencing is done for many amplified templates simultaneously to speed up the process. Sequencing is initiated by adding known primers, after which nucleotides hybridise to the templates. These are often dyed (Solexa platform), or their inclusion can be detected through light the inclusion of nucleotides produces (Roche 454). A template is read through analysis of images from cameras attached to sequencing plates. In subsequent analysis template reads are aligned against the reference human genome, and variants from the reference are detected.

## 1.3 Mapping of disease and trait-associated genetic loci

The emergence of technology for reading DNA poses an exciting opportunity for understanding genetic diseases. Such knowledge could be used for understanding their molecular basis or estimating individuals' risk for developing the conditions. The following section presents the methods for identifying genetic loci where variants are associated with traits and diseases.

### 1.3.1 Family-based linkage studies

Linkage analysis is a way of localising genetic loci that predispose to disease with respect to a set of known genetic markers (Teare and Barrett, 2005). Linkage studies require families with multiple affected members, knowledge of genotypes at genetic markers in the family and a pedigree. Estimate of genetic position of a disease-causing locus comes from maximising the LOD score (Morton, 1956). The LOD score is calculated as a logarithm of the ratio of likelihood of the data given a disease locus at a genetic position to the likelihood of the data given no linkage with the locus. A major problem with linkage studies is reduced power when there are many genes associated with a disease and they are of small effect. Nevertheless, genes for disorders affected by several genes have been mapped with linkage analysis, for example BRCA1 and BRCA2 for breast cancer. Another limitation is that resolution of mapping in this way is limited to tens of centiMorgans, depending on the number of informative meioses in a pedigree. Finally, linkage analysis has only been effective for simple diseases with large risk-ratios for different genotypes. Building a genetic model for parametric linkage studies requires knowledge of mode of inheritance of a disease, disease allele frequency and penetrance, some of which may be unknown.

Major gene disorders are amenable to parametric linkage analysis, because the disease allele frequencies and penetrance can be estimated in advance. For complex diseases, non-parametric methods have been used but there have been relatively few successes in discovering disease susceptibility genes. Non-parametric linkage methods involve testing of increased IBD sharing among relatives, in cer-

tain genomic regions. These tests can be pair-wise, between siblings, but also exploit larger pedigrees. Simplest tests would look for excess IBD sharing compared to expected due to relatedness at a single locus at a time, while more complex ones would attempt to map the genes throughout the chromosomes. An advantage of linkage studies in comparison with association mapping is that they are not affected by allelic heterogeneity.

### 1.3.2 Genome-wide association studies

Linkage analysis has little chance of success with complex diseases which could be affected by variants of modest effect in multiple genes, because the power of the method is heavily reduced when little of trait variance is explained (Sham et al., 2000). In order to dissect genetic mechanisms that drive diseases and complex traits, genome-wide association studies analyse genetic data from unrelated individuals who share at most short haplotypes around causal variants. Genome-wide association studies (GWAS) assume that the causal variants can be tagged by single-nucleotide polymorphisms (SNPs), at which the samples are genotyped (McCarthy et al., 2008). In GWAS genotypes at each SNP individually are tested for association with a disease or trait, through linear or logistic regression.

GWAS studies have identified many disease-associated loci, but the associations detected at genome-wide significance level account for only a small proportion of estimated genetic variance, as estimated from phenotypic resemblances between relatives. This is the 'missing heritability', which limits the potential for individual disease risk prediction (McCarthy et al., 2008). Missing heritability could arise from allelic heterogeneity. Multiple causal variants may not be tagged by genotyped SNPs (Terwilliger and Weiss, 1998). Some of the problem with allelic heterogeneity can be dealt with through choice of samples, for example through choosing distantly related familial cases. Finally, the problem of allelic heterogeneity in GWAS could be addressed through increasing coverage of variation on genotyping arrays, so that typed markers are correlated with the causal variants. However, to tag all rare variants, very dense arrays or resequencing would be required.

### 1.3.3 Population isolates

Allelic heterogeneity of complex diseases and traits can be reduced by studying isolated populations (Peltonen et al., 2000). Population isolates may be generated either by prolonged constraint of population size or by a founder effect. Genetic drift and population bottlenecks reduce genetic complexity of individuals. Genetic variants that are rare in general population may become more frequent, and they may be better tagged by SNPs on genotyping arrays. Association studies may be further facilitated for diseases particularly common in an isolate, also because environmental factors and cultural features are more uniform than in general population.

Two widely studied isolated populations are Finland and Iceland. The majority of Finns descended from migration waves 4000 and 2000 years ago, and since then formed several sub-isolates around the country. Iceland was settled around 1000 years ago by immigrants from Norway, Ireland and Scotland, and since then received few newcomers.

Population isolates offer reduced allelic heterogeneity, increased frequency of some diseases and uniform exposure to environmental factors. While reduced genetic complexity of individuals in isolates may facilitate association studies, it also means that the associations found may not generalise to wider population.

### 1.3.4 Linkage disequilibrium studies

Linkage disequilibrium studies assume that affected individuals share a region of a chromosome around the causal variant, because they all co-inherited a haplotype on which the mutation occurred. This is more likely in a population isolate, where due to migration constraints any two carriers are likely to have a recent common ancestor, and where genetic heterogeneity of diseases is reduced. A linkage disequilibrium study in the isolated population of Finland localised a gene associated with diastrophic dysplasia (Hastbacka et al., 1992). Studied were 77 Finnish families where the disease segregated. Initially, 20 markers of chromosome 5 had been shown to be in linkage with the diseases, and the following linkage disequilibrium study focused on haplotypes at these markers. At a 2 consecutive markers 95% of individuals affected by disease shared a haplotype, from which it

was inferred that the causal variant was within as little as 0.06 cM.

The linkage disequilibrium study narrowed down the position of the causal variant from a region of 1.6 cM which had previously been established by linkage analysis. The linkage disequilibrium study took advantage of many more informative recombinations that occurred in the whole isolate, in comparison to within affected families. The approach treated all Finnish carriers as a big, distantly related family.

### 1.3.5 Coalescent model

Linkage disequilibrium between genetic loci arises because of history of mutation, recombination and coalescence of lineages. Alleles on the same haplotypes are statistically dependent because of genetic linkage, and there is dependence between haplotypes due to shared ancestry. The coalescent is a stochastic process that enables modelling history behind genetic polymorphism data (Rosenberg et al., 2002).

The Kingman's coalescent is an approximation which allows a computation of probability of a genealogical tree for each genetic locus. At the bottom of the tree are alleles in samples studied, which coalesce in the past to form lineages. All lineages within a sample coalesce at the time when the most recent common ancestor (MRCA) lived, from which all samples inherited an allele. Mutation at a locus might have occurred at some point in the past, so that all lineages descending from an ancestor for which it occurred, carry the variant. The Kingman's coalescent allows computation of probabilities of ancestral trees. The lineages under study randomly choose other lineages to coalesce with, at times in the past dependent on number of lineages. The rate at which they coalesce depends on the number of samples in the study and number of lineages at each point in the past. Eventually all lineages coalesce to their MRCA.

Recombination events that gave rise to haplotypes in the studied samples also affect coalescent trees. Coalescent trees at neighbouring loci are will be distinct if recombination occurred between neighbouring haplotypes. The extent to which two coalescent trees between neighbouring sites are similar will be affected by recombination rates in the region. When few cross-overs occur, this will manifest

itself as linkage disequilibrium in the genetic region.

The coalescent model can be used for building simulations, or estimating parameters of history of a population studied. With the coalescent it is possible to compare fit of different history models with the genetic polymorphic data. This involves summing over all possible models of the past, which however is computationally very demanding.

### 1.3.6 Shattered coalescent

One of uses of the coalescent model is fine-mapping of disease loci (Morris et al., 2002). Morris et al. described a method for inference of location of a causal disease locus from case-control haplotypes from an implicated region. The idea is the same as in linkage disequilibrium studies, namely that disease haplotypes in the neighbourhood of the causal variant are all the same, and descend from same branches of the coalescent tree.

In an effort for disease fine-mapping, the authors specify a fully Bayesian model of case-control haplotypes, with a prior probability on coalescent trees which connect them. Adjacent neighbouring trees are linked by recombination events, which in a simplified way are also included into the coalescent model. An important innovation is allowing the tree to coalesce to multiple roots rather than one common ancestor, from which the name 'shattered coalescent' is derived. In this way, the model can account for multiple mutations affecting the disease and sporadic occurrence of a diseases, for example due to genes outside of the studied region.

Morris et al. studied 92 control 94 case chromosomes from patients affected by cystic fibrosis, a disease whose genetic background is well understood. 23 markers in a previously implicated region on chromosome 7 entered the analysis. As a result of applying the method, the most likely associated causal locus coincided closely with a variant known to be a causal one. As a further benefit of modelling the coalescent trees, estimated was also time to common ancestor of the cystic fibrosis mutation. However, the computational cost of Bayesian analysis turned out to be very large - analysis with only 23 markers took 2 days of computation on a personal computer. Similar large-scale analysis of more complex disease may

be prohibitively demanding, because computation time scales exponentially with depth of descent tree

### 1.3.7 Fragmentation-coagulation process

Coalescent trees are one way of modelling haplotypes found in genetic data, however other more tractable methods have been developed. The other methods group haplotypes into clusters throughout the genome, such that all haplotypes in a cluster descended from a common ancestor. The methods represent the haplotypes as a mosaic of ancestral haplotypes.

One example of such a mosaic-based model is the Fragmentation-Coagulation Process (Teh et al., 2011), which is illustrated in Figure 1.2. Because this is a Bayesian non-parametric model, the number of haplotype clusters can adjust to data in each genetic region. Haplotypes in study change their cluster memberships throughout genetic regions. In coagulation event all haplotypes from a cluster join another one, and other clusters may fragment into several clusters. Coupling this prior haplotype model with a data-likelihood, various types of inference can be performed, for example allele imputations.

Inference on the Fragmentation-coagulation process requires Monte-Carlo Markov Chain sampling, which is computationally demanding for large data sets. The computational time required by the algorithm is proportional to number of clusters of haplotypes in data, so limiting the maximum number of haplotype clusters at a locus simplifies inference. Several phasing algorithms are based on this idea, typically using hidden Markov models, as outlined in further sections.

## 1.4 Phasing algorithms

The aim of phasing algorithms is to reveal haplotypes in genotype data. Short-range HMM-based methods rely on linkage disequilibrium between neighbouring genetic loci, whereas long-range methods depend on long haplotypes shared between distantly related samples.

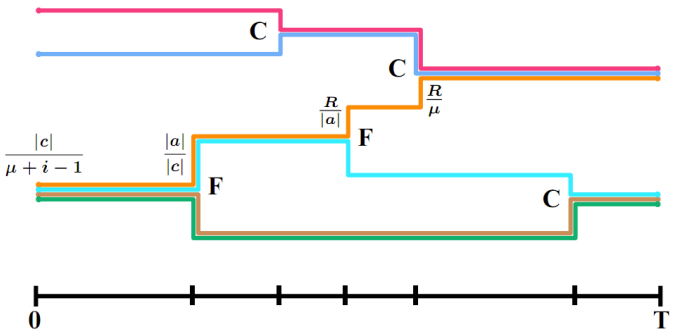



Figure 1.2: Fragmentation Coagulation Process, a Bayesian non-parametric model for haplotypes. Six different haplotypes are modelled in the genetic region between 0 and $T$. At loci marked with C the clusters merge, where marked by F the clusters split. Probabilities of these events depend on cluster sizes. Source: (Teh et al., 2011).

### 1.4.1 Short-range HMM-based methods

All of the discussed short-range phasing methods rely on the the idea that haplotypes of a new sample are noisy copies of haplotypes among other samples (Browning and Browning, 2011*b*). They form a mosaic of the reference haplotypes, as a result of recombination events that occurred since their common ancestor. Additionally, the sample and reference haplotypes can differ at individual loci, which is a trace of mutational process. Because the methods allow for recombinations and mutations between haplotypes, they are called coalescent-based.

The phasing programs connect the observed input genotypes with haplotype models through hidden Markov models (HMM). HMMs are sequential models of data with hidden structure, appropriate for the application because genetic loci are linked by linkage and linkage disequilibrium. Visible states correspond to input genotype data, whereas hidden states correspond to the underlying haplotypes, either to the particular reference haplotypes in some phasing programs, or to modal haplotypes. Transition probabilities between hidden states represent recombination rates, and will likely depend on genetic map in a region. Emission probabilities, which link hidden states and the input genotypes, enforce a match

between inferred haplotypes and the data, and to some extent model mutation between reference and sample haplotypes. The output haplotypes are a result of inference with HMMs, returning the most likely haplotypes given the data. The algorithms typically require several iterations, initially starting with random haplotypes, at each iteration improving the haplotype estimates. The phasing program PHASE is an exception (Stephens et al., 2001), because instead of an HMM it uses an Expectation-Maximisation (EM) algorithm which does not model the local structure of haplotypes. Instead, PHASE uses a coalescent-inspired model of possible haplotypes out of already known ones, and makes haplotypes hidden variables whose most likely values are inferred by the EM algorithm.

The HMM-based phasing programs differ in how they model possible haplotypes. **FastPHASE** models haplotypes as belonging to one of a very limited number of clusters throughout genomic locations (Scheet and Stephens, 2006). For each of the clusters and at each locus, haplotype alleles are assumed to come from a binomial distribution whose parameters are estimated. The number of clusters is chosen by cross-validation, guided by imputation accuracies. The recommended number of clusters is 8, which limits the chance of the cluster to convey long-range dependencies between loci. **MACH** and **IMPUTE**, rather than modelling haplotype clusters, as hidden states take haplotypes that have been estimated for other individuals in previous iterations or from a reference panel (Howie et al., 2009; Li et al., 2010; Marchini et al., 2007). In HMMs large number of hidden states make inference harder, so both MACH and IMPUTE reduce their numbers. As hidden states MACH uses a random subset of haplotypes, whereas IMPUTE chooses haplotypes that are globally most similar to a current estimate of haplotypes for a proband. Instead of using haplotypes from other samples, **Beagle** utilizes the localized haplotype cluster model as a parsimonious empirical LD model (Browning and Browning, 2007). The haplotype model is built from haplotypes reconstructed so far for other samples, and the complexity of the haplotype model through genetic regions can vary with complexity of the haplotypes. Limiting the number of hidden states is very important, since complexity of inference on HMMs scales quadratically with number of hidden states, which becomes prohibitive with sizes of samples currently becoming available. The main advantage of **ShapeIT**, another phasing program is that complexity of

its computation scales linearly with the number of samples in a study (Delaneau et al., 2011). The algorithm represents haplotypes as a graph structure, whose space is limited to a constant for each marker. Based on the graph is an HMM, which despite its state space conveys information about all of the haplotypes. Furthermore, the sampling of haplotypes consistent with the genotypes is also simplified. The method splits the genome into smaller regions containing only a few heterozygous markers, and in each of these regions the allowed haplotypes are enumerated. At the region borders the haplotypes are allowed to switch their state membership. The recently described algorithm of **ShapeIT2** uses a similar representation of haplotypes and inference, however it improves phasing accuracy through an idea borrowed from IMPUTE (Delaneau et al., 2012). When inferring haplotypes for a sample, the algorithm is guided by several haplotypes of other samples globally most similar to a proband. This improves the phasing accuracy and the chance of haplotypes being correct across longer genetic regions. In summary, obtaining scalability of phasing algorithms relies on making compressed representations of haplotypes as hidden states of HMMs.

### 1.4.2 Long-range phasing methods

Long-range phasing methods rely on long identical-by-descent haplotypes inherited from common ancestors of any pair. They do not model or simplify the haplotypes, which often increases phasing accuracy. The first rule-based algorithm for long range phasing was described by Kong et al. (Kong et al., 2008*a*, 2009) and is similar to the method presented later by Hickey et al. (Hickey et al., 2011). The principle behind these methods is explained in Figure 1.3. Both of these algorithms detect IBD sharing only in pre-specified genetic regions, identical for all pairs of compared multi-point genotypes, whereas in reality boundaries of IBD regions can occur anywhere across the genome. Further implementations and improvements were brought by Palin et al. (Palin et al., 2011) through Systematic Long Range Phasing (SLRP).

Systematic Long Range Phasing (SLRP) is a fully probabilistic model for phasing and IBD detection in isolated populations. The Bayesian network in SLRP allows for handling of genotyping errors. Additionally, this framework also

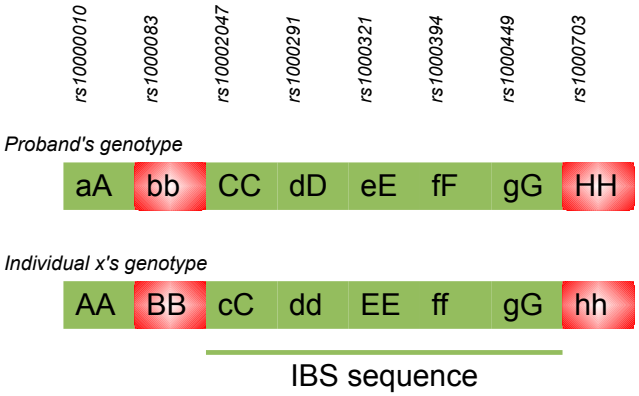



Figure 1.3: Principle by which IBD sharing is inferred from SNP data in the algorithm described by Kong et al.
Multi-point genotypes at consecutive SNPs are shown for two individuals. Minor alleles are denoted by lower-case letters, while major alleles by capital letters. No haplotype could be shared between the two samples at loci with opposing homozygotes, i.e. at the "bb/BB" locus and the "HH/hh" locus. Where there is a long region without opposing alleles, IBD sharing of a haplotype is inferred.

integrates detection of IBD and phasing, so that they are done simultaneously.

Inference of phase and the regions of IBD sharing is done through loopy belief propagation, a special case of the sum-product algorithm for factor graphs (Kschischang et al., 2001). Maximum-a-posteriori configuration is arrived at after many iterations, in each of which nodes adjacent in the Bayesian network send messages one to another. In networks without loops this procedure finds an exact solution, however in networks with loops like SLRP the algorithm is not guaranteed to converge to an optimal solution. The authors limit the number of iterations of the algorithm to 30, and according to their experiments the messages sent always converge before then.

The computational complexity of the loopy belief propagation in SLRP scales quadratically with number of individuals in the study, and linearly with the number of SNPs and iterations. The authors of SLRP took steps to decrease the computational load. The belief propagation algorithm is applied only to such parts of the network where pre-processing identifies possible IBD sharing. The set of putative IBD regions can be pruned, as typically only few haplotype sharers

are required for phasing. Furthermore, the genome can be divided according to the putative IBD regions, so that the algorithm can proceed independently in each of the regions and chromosomes. Finally, the implementation of SLRP had been prepared for parallel processing.

At the time I started the project only the method by Kong et al. was available. SLRP is the most recent, thorough and elegant method for long-range phasing, and it is against this method I will compare my method described in further sections. Another motivation for developing my long-range program was that implementation of Long Range Phasing was not published.

## 1.5 Algorithms for detecting and analysis of IBD segments

For several applications described in this thesis the purpose is not phasing, but rather obtaining IBD segments where two individuals share long haplotypes IBD. Often, like in the case of SLRP algorithm, the IBD segments are a by-product, since they are used for phasing (Palin et al., 2011). Other methods for detection of IBD segments use short-range phasing methods, and then attempt to correct for likely phasing errors.

Two examples of such programs are fastIBD and GERMLINE (Browning and Thompson, 2012; Gusev et al., 2009). Both algorithms are based on the same idea. They first use short-range phasing methods to obtain estimates of haplotypes, and then check for match of haplotypes between samples. Next, they extend the matches where possible, correcting for phasing errors. Both of the methods score the likelihood of true IBD sharing in a region. The difference between fastIBD and GERMLINE is that the former uses Beagle haplotyping algorithm internally, whether GERMLINE expects pre-phased haplotypes as input.

Methods for analysis of the inferred IBD segments have also been described. The effectiveness of localising genetic variants with fastIBD was described for case-control studies (Browning and Thompson, 2012). Statistical methods assessed whether cases of a disease share IBD more often than controls, which is reminiscent of non-parametric linkage studies. The power of the approach was

evaluated using simulation, which showed that IBD analysis for populations with a very recent bottleneck is more powerful than SNP association analysis. They conclude that if the founding event for a population studied is very recent, IBD analysis should be preferred over conventional association studies, particularly with large samples sizes where rare variants would accumulate. For older populations without such history the result no longer holds. Using fastIBD, the authors search for IBD associations with type I diabetes, in the Wellcome Trust Case Control data. They find that conventional association studies found more significant associations than the IBD analysis. They conclude that IBD analysis could still be beneficial with larger sample sizes. However, in larger studies it is likely that data will come from several different genotyping platforms, in which case accurate IBD detection would be more challenging.

Another method for analysis of IBD segments is named DASH (Gusev et al., 2011). The method corrects potentially noisy IBD segments with graph-theoretical approach, and checks for associations of IBD clusters with diseases. IBD clusters could be in closer associations with diseases than SNPs, because the latter may not be sufficiently correlated with the functional variants. The authors use identical-by-descent regions as proxies for recent variants. IBD segments are first recovered from genotype data using GERMLINE, and errors are corrected through graph theoretical methods. Transitivity of IBD means that if haplotype A is identical to B and B is identical to C, then A should be identical as C. Although in theory the IBD relationship should be transitive, imperfect detection of recently co-inherited regions may remove this property. The algorithm restructures the IBD relationship so that it becomes transitive, and the graph representing it consists of fully connected components. For each IBD cluster, the method computes a likelihood ratio between it being a true cluster, where all individuals share IBD with each other, and of the cluster being spurious, where IBD sharing between the members of the group is due to noise. They translate the principle of likelihood into density of a graph, and propose an algorithm for finding optimal sub-graphs. The output of the method is clustering of haplotypes, which can potentially be different at each locus. The resulting clusters are checked for associations with quantitative traits, and in particular whether they carry more signal than individual markers typed in the nearby regions.

## 1.6 Outline of the thesis

We have described methods and technology for localising genetic loci associated with traits and diseases. At the core of theses methods is the coalescent model, according to which haplotypes containing disease haplotypes have been inherited from a common ancestor.

In the subsequent chapters we describe a novel long-range haplotyping method and its application to resequencing studies, to understanding diseases with Mendelian subtypes and to genomic predictions.

# Chapter 2

# Inference of identity by descent in genetically isolated populations

Data from large genetic studies of population isolates was generated in my group, with intention of studying effects of rare variants. SNP genotypes became available for around 1000 individuals each from Scottish archipelago of Orkney and from Croatian islands of Vis and Korcula. Long-range phasing was then first described (Kong et al., 2008*b*), but no implementation was available. My aim was to develop an algorithm capable of correctly recovering long-range haplotypes and of detecting regions of haplotypes shared IBD between pairs of samples. During my work other methods for long-range phasing and detecting of recent IBD were released. In this chapter the newly described methods are evaluated.

The work presented here was described in a published article (Glodzik et al., 2013).

**Outline.** We describe ANCHAP, a new long-range algorithm for detection of identical by descent haplotypes in genetically isolated populations. Our method is designed to detect borders of regions of identity by descent precisely, with minimal computation time and with state-of-art sensitivity and false discovery rates. We compare ANCHAP against other long-range methods, a short-range method, and demonstrate an application of the identified IBD regions for optimisation of

sequencing studies.

## 2.1 Background

### 2.1.1 Properties of recent identity by descent

Expected lengths of haplotypes shared IBD can be derived by making assumptions about population history and properties of the recombination process. Haplotypes that are identical by descent originate 'from a common ancestor without either of them experiencing a recombination' (Powell et al., 2010). The length of a common haplotype between a pair of samples depends on the number of recombinations that occurred on the haplotypes since their common ancestor, so also indirectly on the number of generations that carried the haplotype. Assuming that genetic positions of cross-over events follow a Poisson arrival process at rate 1 per Morgan, we can derive the expected length of a shared haplotype and its variance. For two individuals sharing a haplotype inherited from a common ancestor, the lengths of the shared regions are exponentially distributed with mean equal to $(2n)^{-1}$ Morgans, where $n$ is number of generations back to most recent common ancestor (MRCA) (Browning and Browning, 2010; Haldane, 1919). However, the distribution of segment lengths has variance of $(2n)^{-2}$ Morgans, so the correspondence of segment length with time to common ancestor is only approximate. For example, for a pair of individuals with a common ancestor 25 generations ago, the expected length of a shared haplotype segment is 2 cM, with standard deviation of 2 cM.

Furthermore, we expect haplotype sharers from a population isolate to form clusters. As shown in Figure 2.1, where several individuals co-inherited the same haplotype IBD at a locus, they will all share IBD with each other. Haplotype sharing with respect to the gametes of each individual is a transitive relationship. If a haplotype A is IBD with haplotype B, B with C, then A is IBD with C. These characteristics of IBD sharing are assumed in the algorithm for detecting IBD from SNP data, and we can make us of the properties to correct imprecise IBD inference.

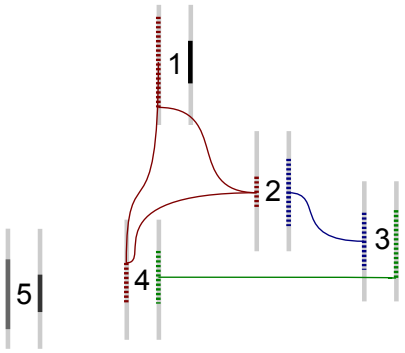



Figure 2.1: Haplotype sharing within a population isolate. Genotyped individuals are identified by numbers 1 to 5. Each individual has two haplotypes, represented by thick light-grey bars. Red, blue, and green dotted lines represent identity by descent of two haplotypes in a genomic region. The dark-grey shaded haplotypes are unique in the sample, and they are not shared between the sampled subjects.

### 2.1.2 Identifying regions of IBD sharing from SNP data

After the first long-range phasing algorithm was described by Kong et al., subsequent ones appeared with implementations (Kong et al., 2008*a*, 2009). Hickey at al. presented a method that was very similar to Kong's in that it also detected IBD in genetic regions that a genome is first divided into (Hickey et al., 2011). Subsequently Systematic Long Range Phasing (SLRP) appeared, which is a more flexible, elegant probabilistic model (Palin et al., 2011). As an alternative to long-range phasing appeared, FastIBD, which uses a HMM-based short-range method. The likely phasing errors are then corrected when identifying IBD sharing between genotypes which were phased. The algorithms are described in more detail in Chapter 1.

In this chapter for comparison we used SLRP and FastIBD, each representing long-range and short-range methods for detecting haplotypes shared IBD between pairs of samples. SLRP exemplifies a long-range method capable of performing probabilistic inference simultaneously for pairs of individuals, at likely high computational cost. FastIBD reduces the computational effort by more concise haplotype modelling. In contrast, the method I developed, ANCHAP, is a long-range but heuristic method.

## 2.2 Methods

### 2.2.1 Populations studied

In our study of ancestral haplotypes we analysed four European cohorts, three of which (ORCADES, CROATIA-VIS, CROATIA-KORCULA) are from isolated island populations and one from a mainland population (SOCCS).

The Orkney Complex Disease Study (ORCADES) is a family-based, cross-sectional study in the isolated Scottish archipelago of Orkney (McQuillan et al., 2008). Genetic diversity in this population is decreased compared to mainland Scotland, consistent with the high levels of endogamy throughout history. Orkney has been inhabited for over 5000 years, but the original population was almost completely replaced by Norse Vikings about 800-900 CE. From about 1300 to 1600 CE there was an influx of mainland Scots (Wilson et al., 2001). For this analysis we used data from 749 participants aged 18-100 years from ten islands, however for the purposes of evaluation of methods we removed parents from genotyped parent-offspring pairs which reduced the cohort size to 597 individuals. Genotyping in the study was done using the Illumina HumanHap300 array with 302379 single nucleotide polymorphisms (SNPs) after quality control.

The CROATIA-VIS study is a family-based, cross-sectional study in the villages of Komiza and Vis on the isolated island of Vis that included 1,056 examinees aged 18-93 years (Vitart et al., 2006). The CROATIA-VIS study genotyping used the Illumina Hap300v1 SNP chip with 301069 SNPs after quality control.

The CROATIA-KORCULA study is a family-based, cross-sectional study in the villages of Lumbarda, Zrnovo and Racisce on the isolated island of Korcula in Croatia (Polasek et al., 2009). The study included 965 examinees aged 18-95 years. The CROATIA-KORCULA study genotyping used the Illumina Hap370CNV SNP chip with 317223 SNPs.

The Study of Colorectal Cancer in Scotland (SOCCS) is a case-control study of prospectively collected colorectal cancer cases from all Scottish hospitals, and matched controls. One thousand participants in each group in the first phase of the study were genotyped with Illumina HumanHap300 array with 306204 SNPs. The participants for the control group were matched by age, sex and region to

cases according to a nearly complete population based register, and then selected at random. We analysed the genotypes from the control group, so as to obtain a sample representative of the Scottish population as a whole.

### 2.2.2 Recent identity by descent in population isolates

In an isolated population such as the Orkney population, founded by Viking settlement about 50 generations ago, and where population size has been constrained over many generations, the time to MRCA is either of the order of 1000 generations ago, during the early settlement of Europe, or less than 50 generations ago. To be able to use IBD sharing to infer sharing of rare variants, taking into account mutation rates (Duret, 2009; Nachman and Crowell, 2000), we restrict the definition of IBD sharing to sharing via a recent common ancestor. In practice, we can only do this by setting a minimum length for the shared region. For this study we set the cut-off at 2 cM, equal to the expected length of sharing given a time to MRCA of 25 generations. We used the high-resolution genetic map from the HapMap project (Myers et al., 2005). Additionally, the cut-off at 2 cM has been suggested in literature as a threshold above which accurate detection from contemporary genotyping arrays can be obtained (Browning and Browning, 2010).

### 2.2.3 Algorithm of ANCHAP

The objective of ANCHAP is to infer recent identity by descent from SNP data with maximum sensitivity and specificity, which means that it should declare IBD only where the haplotype was co-inherited from a recent common ancestor, and find all of such regions.

The algorithm consists of three stages:

1. Stage I. First scan for IBD sharing from comparisons of multi-locus genotypes of all pairs of individuals.

2. Stage II. Splitting haplotype sharers by alignment and phasing. Individuals carrying parts of the individual's "maternal" haplotype are distinguished from ones that carry the "paternal" haplotype. While the actual paternal or

maternal origin of the proband's haplotypes is unknown, haplotype sharers are split into two groups.

3. Stage III. Second scan for haplotype sharing: a more sensitive and specific scan for IBD sharing, by pairwise comparisons of partially uncovered haplotypes.

In Stage I, ANCHAP detects IBD sharing between pairs of multi-locus genotypes, scanning for large regions without opposing homozygotes. Allelic dosages at each SNP are coded as 0, 1 and 2, and opposing homozygotes between two individuals at a SNP are when allelic dosages are 0 for first individual and 2 for the second one, or 2 and 0. Where there are no opposing homozygotes over a long region, a haplotype is likely to be shared IBD. To account for uncertain sharing near the boundaries of a region with no opposing homozygotes, a number of markers at the margins are trimmed, and are not included into the shared region. The parameters of the method are the IBD threshold (the minimum genetic length in centiMorgans of a region without opposing homozygotes between a pair of genotypes) and the number of markers to be trimmed at margins. The threshold values in centiMorgans relate to the time to common ancestor from whom the haplotype was inherited, while values expressed as number of SNPs exclude regions with low SNP density.

Stage II of the algorithm is executed for each proband in a study separately. In a given region, in order to reconstruct phase, the proband's haplotype sharers can be split into two groups by alignment in Stage II. If sharers of proband's haplotypes on both gametes are present, they will form two groups. If sharers of only one gamete are present, or if the proband is homozygous by descent they will form one group. When sharers of each proband's haplotypes are identified, phasing becomes possible. We can recover the proband's haplotypes at each locus where at least one of the haplotype sharers is homozygous, but information about all of the homozygotes among the sharers reduces the number of phasing errors. Because a haplotype is shared, the haplotype sharer's allele at a homozygous locus must be the same as the allele on the proband's haplotype (Kong et al., 2008*a*). Each homozygous SNP among haplotype sharers is a "vote" for consistent phase at proband's heterozygous locus; and votes from all homozygous sharers at a locus

are added up to decide phase. As the method distinguishes groups of individuals sharing each of the proband's haplotypes in a region without recourse to pedigree, the actual paternal or maternal origin of these proband's haplotypes is not known. Haplotype sharers are split into two groups, which could correspond to haplotype origin. Figure 2.2 illustrates Stages I and II of the algorithm.

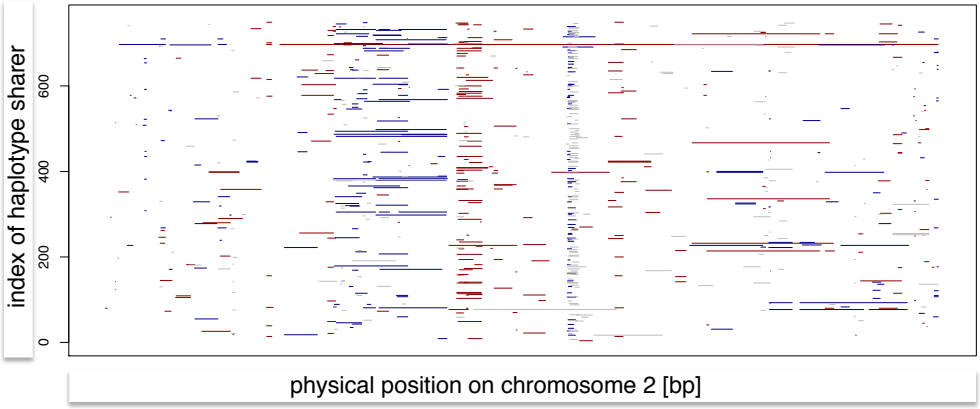



Figure 2.2: Example of IBD detection (Stage I) and alignment of IBD regions (Stage II) for one individual from ORCADES, chromosome 2. First we find his/her haplotype sharers across the genome, and mark the regions of putative IBD sharing as segments. The shared sequences are aligned into two groups, and marked red, and blue accordingly. The grey segments denote misaligned shared sequences. Individual 697 is a full sibling of the proband, with almost the entire chromosome shared and more distant relatives share smaller blocks. Regions of increased IBD are visible, which could arise in parts of genome where we incorrectly infer time to common ancestors from whom haplotypes were inherited.

In Stage III of ANCHAP we make use of the phase information obtained from IBD regions detected earlier. When partially complete haplotypes have been inferred, a second scan for IBD sharing is undertaken, exploiting the additional phase information gained. Phased haplotypes are compared between samples and IBD is declared when they continuously match across a number of markers. The idea for this second scan for haplotype sharing was inspired by the hidden Markov model described in (Genovese et al., 2010). A pair of completely known

haplotypes can have different alleles at any phased locus, while in a pair of unphased genotypes only at loci homozygous for both individuals are indicative of sharing. Partially complete haplotypes carry more information for IBD detection, and thus the second scan can be more informative. To detect IBD from comparisons of nearly complete haplotypes, we can use a smaller threshold in terms of consecutive number of SNPs than the one that delineates IBD sharing from IBS in Stage I, and we no longer need to trim the margins of the shared regions as in Stage I. Recent IBD is declared in a region of consecutively matching alleles between haplotypes that spans a sufficient genetic distance, and when the number of phased markers in the region exceeds the threshold of minimum phase information.

### 2.2.4 Settings required by ANCHAP

ANCHAP requires values for the following settings:

- $T_I$ - IBD threshold (Stage I). Minimum length of a region without opposing homozygotes before it is declared as IBD, expressed in centiMorgans.
- $R_M$ - IBD region margins (Stage I). Number of markers trimmed from margins of IBD segments.
- Alignment parameters: $O_T$ overlap threshold (number of heterozygous SNPs where two segments overlap) and matching threshold $M_T$ (Stage II).
- $P_{III}$ - minimum number of markers phased for both individuals in a putative IBD region (Stage III).
- $T_{III}$ - IBD threshold (Stage III), expressed in centiMorgans.

Using this notation for the parameters, algorithms for Stages I, II and III of ANCHAP are described in Figures 2.3, 2.4, 2.5 and 2.6.

Figure 2.3: Stage I: algorithm for first scan for sharing from unphased genotypes

---

Input: $G$, $T_I, R_M$

$G$ : Genotype matrix: $N \times M$ ($N$: Number of samples, $M$ number of SNPs), with $G_{i,j} \in \{0, 1, 2, \text{NA}\}$ - allele dosage

1. initialise $IBD.segments$ (empty list for storing IBD segments)
2. **for** all pairs of individuals $(i, j)$
   (a) identify segments of genome longer than $T_I$ centiMorgans without opposing homozygotes between multi-point genotypes of $i$ and $j$,
   (b) **for** all segments
      i. trim $R_M$ from each margin of an IBD segment identified in (a)
      ii. add indices of start and end SNPs, indices of individuals $i, j$ to $IBD.segments$

Output: $IBD.segments$

---

Figure 2.4: Stage II: algorithm for alignment of IBD segments.

Input: $G$, $IBD.segments$, $O_T$, $M_T$

1. Initialise $H1$, $H2$ - two matrices to store phased haplotypes of all probands, of same size as $G$. Initialise them at homozygous loci, otherwise leave unknown.
   i.e. $H1(G == 0) = 0$, $H1(G == 1) = NA$, $H1(G == 2) = 1$, similarly for $H2$.

2. Initialise $gam$ - a vector of same size as $IBD.segments$, $gam \in \{1, 2\}$, to store which proband's gamete a segment belongs to

3. **for** all individuals $i$

   (a) $IBD.segments.i$ - IBD segments of individual $i$ with others, sorted by the segment length in descending order

   (b) **for** all segments $s$ in $IBD.segments.i$

      i. check if the genotype of the haplotype sharer in segment $s$ is matching haplotype 1, 2, or none of the proband $i$, with parameters: if there are enough phased loci as specified by $O_T$ and that $i$'s haplotype alleles agree with sharers genotype with tolerance $M_T$, accordingly assign 1 or 2 to $gam[s]$

      ii. based on the new segment and its gamete assignment, update $i$'s haplotypes in $H1$ and $H2$
      $(H1, H2) = phase(i, IBD.segments.i[1 : s], gam[1 : s], G, H1, H2)$ (Figure 2.5)

Output: $H1$, $H2$, $gam$

Figure 2.5: Stage II: algorithm for phasing using the aligned IBD segments

$(H1, H2) = phase(i, IBD.segments.i, gam, G, H1, H2)$

1. initialise vectors counting phase "votes" , at each locus: "votes" for major allele on gamete 1 $votes.H1.major$, and for minor allele on gamete 1 $votes.H1.minor$
2. **for** all genetic loci $l$
   (a) $IBD.segments.i.l$ is a list IBD segments with individual $i$ spanning locus $l$
   (b) identify all homozygotes of haplotype sharers at locus $l$ from $G$
   (c) according to the homozygous genotypes, and their gamete assignments, update phase votes $votes.H1.major$ and $votes.H1.minor$
   (d) If $votes.H1.major \neq votes.H1.minor$, decide on phase at locus $l$ accordingly, i.e. major H1 if $votes.H1.major > votes.H1.minor$

Output: updated $H1$, $H2$

Figure 2.6: Stage III: algorithm for second scan for sharing from partially complete haplotypes

---

Input: $H1$, $H2$, $T_{III}, P_{III}$

1. initialise $IBD.segments.III$(empty list for storing IBD segments)
2. **for** all pairs of individuals $(i, j)$
   and 4 gamete combinations $(g1, g2) \in \{(1,2),(1,1),(2,1),(2,2)\}$
   (a) identify segments of genome without opposing homozygotes between multi-point haplotypes of $i$ and $j$ from gametes $g1$, $g2$, longer than $T_{III}$ centiMorgans, with at least $P_{III}$ phased SNPs
   (b) **for** all found segments
      i. add indices of start and end SNPs, indices of individuals $i, j$, gamete indicators $g1$, $g2$ to $IBD.segments.III$

Output: $IBD.segments.III$

---

### 2.2.5 Comparison of methods

We evaluated methods for detecting recent IBD using genetic data for parent-offspring trios. We compared ANCHAP against SLRP - a fully probabilistic method for long-range phasing, and against fastIBD - a short-range method designed for populations of unrelated individuals. We evaluated their results genome-wide against recent IBD that can be reliably detected by comparison of haplotypes phased using parental genotypes.

Reference haplotype sharing was recovered between individuals whose parents were also genotyped. Among the individuals genotyped in ORCADES, there were 58 individuals with both parents genotyped, and on average 80% of heterozygous loci of such reference individuals were phased. We identified the regions of reference recent IBD sharing between pairs of reference individuals where their haplotypes are identical for at least 2 cM, over at least 100 SNPs. For genotyping arrays used in this study 2 cM corresponds roughly to 200 SNPs, and setting a

lower threshold eliminates only the regions with particularly low SNP density.

Such a definition of reference IBD segments carries possible problems. Firstly, by setting a threshold on length of the segments, we are not setting a precise cut-off for time to common ancestor. Given length of a haplotype shared IBD there is much uncertainty about time to the common ancestor from which the haplotype was inherited. This is further discussed in Chapter 5. Secondly, there may be genotyping errors or incomplete phase in the genotypes of the parents, and therefore the reference segments may be inaccurate. An alternative would be to generate true IBD segments and genotype data through simulations, yet this would involve making many assumptions about the history of the populations. Our reference IBD segments from ORCADES should reflect the population well.

Each of the compared methods was applied to genotype data from the 597 individuals in ORCADES, free from parent-offspring pairs. The results between the 58 reference individuals were evaluated against the regions of reference IBD obtained from their known haplotypes, as also parents of the reference individuals were genotyped. For fairness of the comparison with ANCHAP, segments shorter than 2 cM had been pruned from results of SLRP and fastIBD. Before this simple post-processing of the result, ANCHAP outperformed the other methods. Since the aim is to recover regions of IBD sharing longer than 2 cM, I found this post-processing fair.

The total number of markers in output regions that are also in the reference IBD regions is $TP$ (true positives), in reference regions but not in the output regions is $FN$ (false negatives), not in reference regions but in the output regions is $FP$ (false positives). For each method in the comparison we quote sensitivity defined as the ratio $TP/(TP + FN)$ and false discovery rate $FP/(FP + TP)$.

Parameter tuning for the methods was informed by the following performance metrics:

- sensitivity - $TP/(TP + FN)$,
- false discovery rate - $FP/(FP + TP)$. We used false discovery rate, rather than false negative rate, because it is informative for further analysis of the IBD relationship. It is useful to know that given recent IBD was detected at a locus, how likely a recent haplotype is not shared.

In case of ANCHAP, there are also evaluation measures that express the quality of alignment in Stage II:

- $I_r$ - inconsistency rate - how many of the alleles of haplotype sharers were homozygotes not consistent with homozygotes of the majority of the haplotype sharers, divided by the number of haplotype sharers.

- $P_a$ - percentage of aligned sequences in the first stage of an algorithm. Out of all detected IBD regions in the first round, what proportion of them were aligned into one of the gametes.

### 2.2.6 Parameter tuning

All of the compared methods require setting different parameters. The methods were tuned according to their sensitivity and false discovery rate on a subset of the ORCADES data set from chromosome 2, using the reference individuals phased in parent-offspring trios, as well as by success of alignment in Stage II ($I_r$ and $P_a$)

We attempted to set the IBD threshold at Stage I of ANCHAP ($T_I$) such that the length of falsely assumed IBD regions is reduced (false discovery rate) while recovering as much of the true IBD regions as possible (sensitivity), and thus the phase recovery that uses the IBD segments is most accurate and maximally spread. The margin sizes $R_M$ were set by comparison of margins of IBD regions deduced from genotypes and the reference haplotypes. The setting of the alignment parameters at Stage II aimed at increasing the ratio of the IBD segments aligned into haplotypes ($P_a$), and minimizing the inconsistencies between them ($I_r$), which indicate alignment errors. At Stage III, the minimum number of markers phased for both individuals in a putative IBD region ($P_{III}$) was set using the reference haplotypes and the sensitivity and specificity values.

Also using the values of sensitivity and false discovery rate in data from chromosome 2, we adjusted the parameters of SLRP and fastIBD. SLRP required setting the expected length of IBD regions and expected regions of IBS but not IBD regions. The scale parameter in fastIBD controlled the parsimony of the haplotype model.

### 2.2.7 Data pre-processing

The data sets were pre-processed in PLINK (Purcell et al., 2007) to eliminate low quality markers. We removed markers with call rate of less than 95%, out of Hardy-Weinberg equilibrium ($p < 0.001$), or those with minor allele frequency lower than 1%. We excluded individuals with more than 7% genotype markers missing, and retained only the autosomal SNPs. After pre-processing, the following numbers of samples remained: ORCADES (749 individuals, 302,379 SNPs on 22 chromosomes), CROATIA-KORCULA (945 individuals; 317,223 autosomal SNPs, including 295,574 ORCADES SNPs), CROATIA-VIS (991 individuals; 301,069 autosomal SNPs, including 291,857 ORCADES SNPs), SOCCS (958 individuals; 306,204 autosomal SNPs, including 294,703 ORCADES SNPs). We localized the SNPs on the HapMap genetic map of recombination rates (Consortium, 2007).

## 2.3 Results

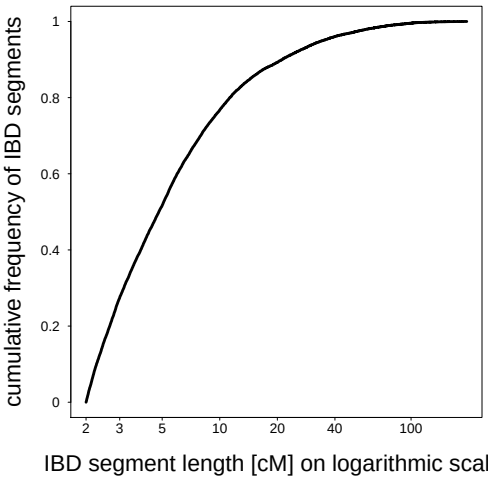
### 2.3.1 Tuning ANCHAP

#### 2.3.1.1 Reference sharing in ORCADES study

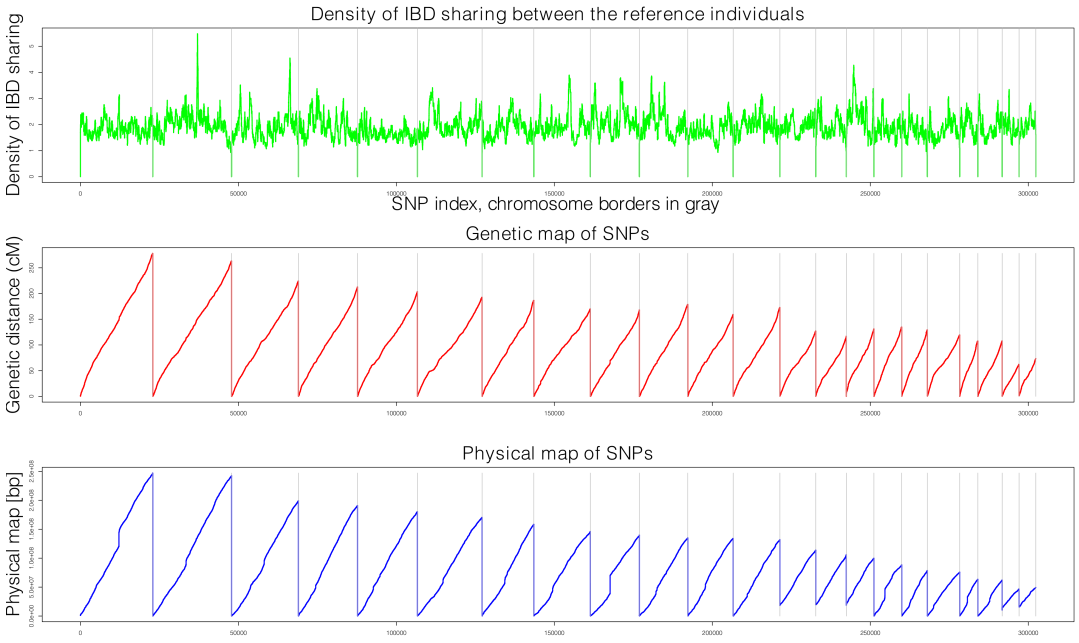
The evaluation of the algorithms was possible thanks to parent-offspring pairs genotyped in the ORCADES study. There are 58 individuals with both parents genotyped, and at 80% of their heterozygous loci they could be phased using their parents' genotypes. There are 160 with at least one parent genotype and they could be phased at 70% of heterozygous loci.

To obtain the reference IBD information, we extracted IBD regions between the 58 reliably phased reference individuals. We required alleles with identical alleles in a region of haplotypes larger than 2 cM and containing at least 100 SNPs. Such regions are highly likely to be IBD since they are based on haplotypes reconstructed from parents' data. The search for IBD regions is more specific when comparing known haplotypes than when comparing genotypes, since the former uses information at all markers, while the latter only at doubly homozygous markers. The length of IBD regions between the reference individuals is

shown in Figure 2.7a. Frequency of IBD sharing across the genome is shown in Figure 2.7b. To verify if the regions of increased IBD sharing coincide with unusual SNP densities in region with respect to physical and genetic maps, we show these below.




(a) Lengths of reference IBD segments in the reference data set for 58 individuals from ORCADES. Most of the shared regions are only slightly larger than 2 cM, giving a median of 5 cM.




(b) Distribution of IBD segments across the genome in the reference data set for 58 individuals. Regions of common IBD sharing could arise because of many false positive detections in a region, for example in regions with poor SNP coverage, or where genetic map does not allow to distinguish recent from ancient IBD sharing. Top: frequency of IBD sharing at a genetic locus, averaged over samples in study. Middle: genetic positions of SNPs. Bottom: physical positions of SNPs. We conclude that there are is no unusually low density of SNPs in the regions of frequent IBD. Around the peak on chromosome 6 there are fewer recombinations than in neighbouring genetic regions.

Figure 2.7: Properties of IBD segments between the 58 reference samples in ORCADES.

#### 2.3.1.2 IBD threshold in Stage I

In Stage I of ANCHAP we would like to phase all heterozygous genotypes with maximum accuracy, and using this principle we set a parameter value for $T_I$. Genotypes would be widely phased if many haplotype sharers are detected throughout the genome. There would be few phasing errors if there is no falsely detected IBD sharing. Therefore the sensitivity and false discovery rate of detecting IBD sharing, as evaluated on the reference phased individuals, are meaningful metrics which will reflect the quality of phasing. The plot of sensitivity and false discovery rates of IBD sharing for different IBD thresholds in Stage I is shown in Figure 2.8. In further experiments we set the threshold $T_I$ to 3 cM.

On the other hand, incorrectly detected IBD in the first stage does not necessarily lead to phasing errors. When there is more than one haplotype sharer, and some falsely detected IBD segments in the region, the alignment stage of ANCHAP will likely eliminate it if true IBD sharers are in majority. This is because the phase of a genotype is decided by voting from genotypes of haplotype sharers (see Figure 2.5).

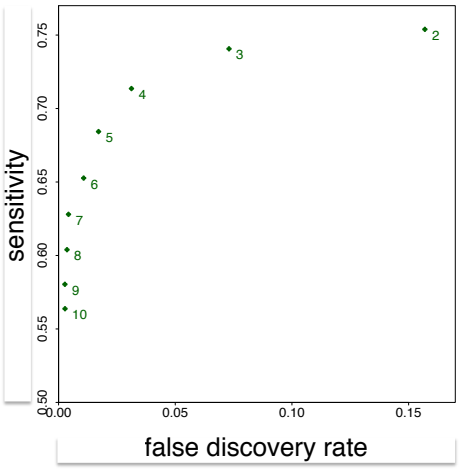



Figure 2.8: Sensitivity and false positive rate of IBD regions as recovered by Stage I of ANCHAP. The numbers in green on the plot are the $T_I$ threshold values expressed in centiMorgans.

#### 2.3.1.3 IBD region margins

At each margin of a putative IBD sharing region we trimmed $R_M = 100$ markers. In the experiments with the reference data, after cutting off 100 markers at each side, 94% of detected sharing regions did extend to where the reference haplotypes were no longer identical.

#### 2.3.1.4 Stage II - alignment parameters

In Stage II of ANCHAP haplotype sharers are split into two groups based match between sharers' genotypes and haplotypes of a proband being reconstructed. The algorithm starts with the longest and therefore most certain IBD regions shared with other individuals, reconstructs a draft of the phase for genotypes of an individual, and then matches the remaining sharers against the preliminarily phased genotypes. Errors may occur in the preliminarily reconstructed haplotypes, and therefore a few inconsistencies between the draft of the haplotypes and the aligned sequences may be allowed.

There are two parameters necessary for this part of the algorithm. The overlap threshold ($O_T$) specifies the minimal number of markers of overlap between the draft of phase of an individual and the new IBD region shared. The matching threshold ($M_T$) specifies how many alleles may be mismatching between the draft of the phase and a genotype of the putative IBD sharer.

Appropriate values of parameters will result in more accurate splits of haplotype sharers into two groups and consequently in lower phasing error, and higher proportions of genotypes will be phased. A higher proportion of the putative IBD sequences would be therefore aligned ($P_a$). The genotypes of IBD sharers who are all classified as sharing the same haplotype, should also be consistent between each other. There should be no opposing homozygotes between genotypes of haplotype sharers of an individual, and therefore the inconsistency rate ($I_r$) should be lower.

In Figure 2.9 we evaluate the impact of different values of the overlap threshold and the matching threshold. For each pair of values, we evaluate the percentage of the putative IBD regions successfully aligned ($P_a$), and the inconsistency ratio ($I_r$).

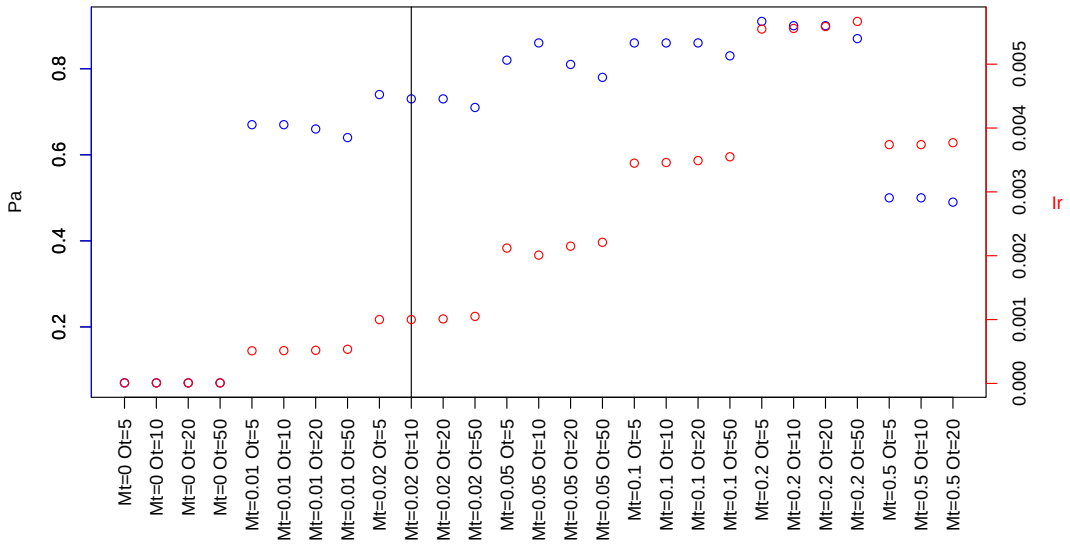



Figure 2.9: Experiments with parameters for Stage II of ANCHAP.
On the left Y-axis is $P_a$ - percentage of aligned sequences in the Stage II of an algorithm. On the right Y-axis is $I_r$ - the inconsistency rate as a consequence of alignment - how many of the alleles of haplotype sharers were homozygotes not consistent with homozygotes of the majority of the haplotype sharers, divided by the number of haplotype sharers.
Intending to maximise $P_a$ while minimising $I_r$, in further experiments we chose parameter values indicated by the vertical line.

#### 2.3.1.5 Stage III parameters

For an individual, where regions of IBD sharing were detected in Stage I and haplotype sharers were assigned accordingly to gamete of origin in Stage II, his haplotypes would have been partially recovered. In Stage III the algorithm searches for haplotypes matching continuously in regions which are at least 2 cM long. In addition we require that both of the compared haplotypes are phased. Another parameter ($P_{III}$) specifies a minimum number of markers phased in both of the haplotypes. The default parameter value is 200 SNPs.

In Table 2.1 we show the accuracy of IBD detection when the $P_{III}$ threshold is varied. With values of the parameter below 100 SNPs sensitivity is not increased, as there are hardly any reference IBD segments with fewer than 100 consecutive and phased SNPs. For larger values of the threshold false discovery rate decreases, because they eliminate regions with low SNP density or unphased genotypes. In order to reduce false discovery rate, in further experiments we set $P_{III}$ to 200 phased markers.

| $P_{III}$ | sensitivity | false discovery rate |
|---|---|---|
| 10 | 0.84 | 0.031 |
| 20 | 0.84 | 0.031 |
| 50 | 0.84 | 0.027 |
| 100 | 0.84 | 0.016 |
| 200 | 0.81 | 0.011 |

Table 2.1: Experiments with values of the parameter for Stage III - $P_{III}$. This parameter specifies how many markers in the region of putative IBD need to be phased in the relevant region of two multi-point genotypes. Below 200 markers noted is increase of false discovery rate. Marked in grey is the value of the parameter used in further experiments.

### 2.3.2 Tuning settings of SLRP

Table 2.2 shows experiments with empirical and default parameter values for SLRP. We compared accuracy of the algorithm in detecting the IBD segments on chromosome 2 when used with the default values and with values obtained empirically. The expected IBD length in centiMorgans was computed from the IBD regions between the reference individuals in ORCADES, after they were phased. The expected IBS but not IBD was calculated from IBS segments between the reference individuals. Because we defined IBD as matching of haplotypes within a region longer than 2 cM, out of the output of SLRP we filtered out the results shorter than this threshold. Table 2.2 shows accuracy of SLRP.

| SLRP setting | ExpectedIBS (cM) | ExpectedIBD (cM) | sensitivity | false discovery rate |
|---|---|---|---|---|
| default | 1 | 10 | 0.76 | 0.0076 |
| empirical | 0.42 | 9.17 | 0.77 | 0.0106 |

Table 2.2: Tuning parameter settings for SLRP. Only counting the IBD regions longer than 2 cM. Sharing between the 58 Orkney individuals was evaluated using data from chromosome 2. Marked in grey is the value of the parameter used for a genome-wide comparison.

### 2.3.3 Tuning settings of fastIBD

In Table 2.3 we show experiments varying the scale parameter required by fastIBD. The scale parameter controls the complexity of haplotype model created. We filtered out regions shorter than 2 cM, in accordance with our definition of IBD.

| scale | | sensitivity | false discovery rate |
|---|---|---|---|
| minimum advised | 1 | 0.270 | 0.000 |
| | 2 | 0.631 | 0.010 |
| | 2.5 | 0.744 | 0.018 |
| | 2.6 | 0.767 | 0.018 |
| | 2.7 | 0.783 | 0.019 |
| | 2.8 | 0.802 | 0.021 |
| | 2.9 | 0.805 | 0.024 |
| | 3 | 0.825 | 0.024 |
| | 3.1 | 0.832 | 0.027 |
| | 3.2 | 0.837 | 0.028 |
| | 3.3 | 0.845 | 0.030 |
| | 3.4 | 0.849 | 0.032 |
| | 3.5 | 0.857 | 0.036 |
| maximum advised | 4 | 0.870 | 0.045 |
| merge 10 runs | 3 | 0.868 | 0.044 |

Table 2.3: Tuning parameter settings for fastIBD, using data from chromosome 2 for the 58 reference individuals from ORCADES. Marked in grey is the value of the parameter used, where sensitivity matches that one of ANCHAP on chromosome 2.

### 2.3.4 Phase propagation in ANCHAP

Table 2.4 shows the gain in sensitivity and the reduction in false discovery rate in detection of recent IBD regions that are obtained at Stage III of our algorithm, as compared to Stage I. On chromosome 2, sensitivity of IBD detection between the 58 reference individuals per pair of individuals per marker grew from 0.75 to 0.81 in the second round. Detection of identity by descent for partially phased haplotypes in the second round helped to reduce the false discovery rate from 0.16 to 0.01.

| method | ANCHAP Stage | ANCHAP Stage III |
|---|---|---|
| sensitivity | 0.75 | 0.81 |
| false discovery rate | 0.16 | 0.01 |

Table 2.4: Experiments with data from chromosome 2, 597 ORCADES individuals with their genotyped parents removed. The identified regions of IBD were evaluated against phased haplotypes of 58 individuals who could be phased using the genotypes of their parents.
Stage III of ANCHAP offers better accuracy in detecting regions of IBD than the first one.

### 2.3.5 Comparison of ANCHAP against other methods

Table 2.5 compares different tuning settings of ANCHAP, SLRP and fastIBD. Using data from chromosome 2, we manipulated parameters of SLRP and fastIBD to match sensitivity and false discovery rate of ANCHAP. Notably, as the sensitivity of fastIBD grows to exceed ANCHAP's 0.81, the false discovery rate of fastIBD reaches 0.024.

| method | parameters and values | sensitivity | false discovery rate |
|---|---|---|---|
| ANCHAP | $T_I$ - IBD threshold Stage I: 3 cM | 0.81 | 0.010 |
| | $T_{III}$ - IBD threshold Stage III: 2 cM | | |
| | $O_T$ - overlap threshold: 10 markers | | |
| | $M_T$ - mismatch tolerance: 2 % | | |
| | $P_{III}$ - minimum phase information: 200 SNPs | | |
| SLRP | default | 0.76 | 0.008 |
| | ExpectedIBS: 1cM | | |
| | ExpectedIBD: 10 cM | | |
| SLRP | empirical | 0.77 | 0.011 |
| | ExpectedIBS: 0.42 cM | | |
| | ExpectedIBD: 9.17 cM | | |
| fastIBD | scale: 1 | 0.27 | 0.000 |
| fastIBD | scale: 2.8 | 0.80 | 0.021 |
| fastIBD | scale: 2.9 | 0.81 | 0.024 |
| fastIBD | scale: 3 | 0.83 | 0.024 |
| fastIBD | scale: 4 | 0.87 | 0.044 |

Table 2.5: Parameter tuning of ANCHAP, SLRP and fastIBD. Experiments with data from chromosome 2, 597 ORCADES individuals with their genotyped parents removed. The identified regions of IBD were evaluated against phased haplotypes of 58 individuals who could be phased using the genotypes of their parents. Highlighted rows indicate parameters used in genome-wide analysis.

Table 2.6 shows the accuracy of IBD detection of ANCHAP against the other methods and their running times. Genome-wide, the methods achieved similar sensitivity of IBD: from 0.75 for SLRP, 0.78 for ANCHAP and 0.82 for fastIBD. Long-range methods, ANCHAP and SLRP resulted in similar false discovery rates of 0.009 and 0.007 respectively, while for fastIBD it is 0.025. Genome-wide inference of IBD with the SLRP model took much longer than for the other methods: the analysis with SLRP took 207 hours, whereas ANCHAP handled the same task in 20 hours and fastIBD in 12 hours.

| method | ANCHAP | SLRP | fastIBD |
|---|---|---|---|
| sensitivity | 0.78 | 0.75 | 0.82 |
| false discovery rate | 0.009 | 0.007 | 0.025 |
| runtime(hours) | 20 | 207 | 12 |

Table 2.6: Comparison of accuracies of methods for IBD detection. ANCHAP is compared to SLRP - a probabilistic method for phasing in isolated populations, and to fastIBD - a method designed for general populations. This genome-wide comparison was run on the subset of 597 individuals from ORCADES, such that their parents were not included. Regions of IBD were also evaluated using parent-offspring trios. Experiments were run on a computer with a 2.0 GHz and 16 GB of RAM.

Additionally, in Figure 2.10 we show frequency of IBD sharing across the genome, as detected by different methods. In Figure 2.11 compared are lengths of IBD segments detected by different methods. As can be seen from the graph, ANCHAP does not detect the longest segments of IBD sharing in one piece. This could be because ANCHAP does not account for possible phasing errors in Stage III.

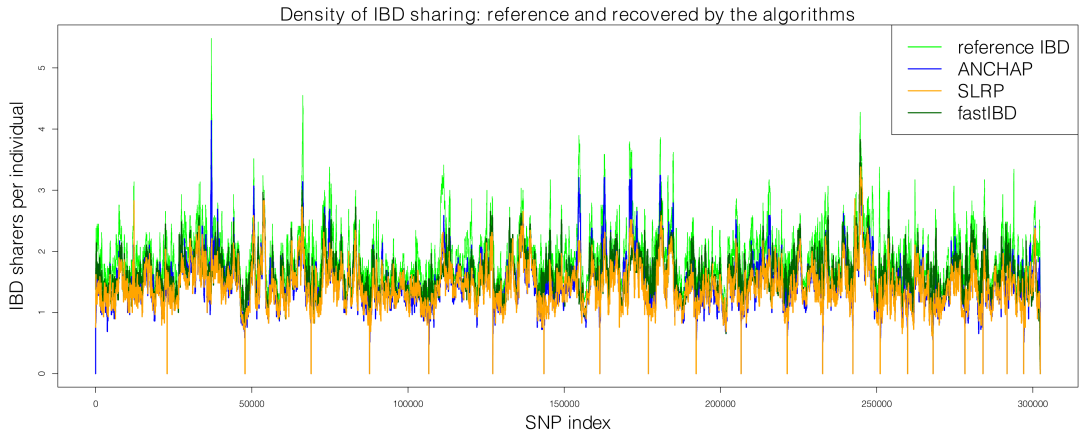



Figure 2.10: Genome-wide view of haplotype sharing as recovered by the compared methods. SLRP and fastIBD are more conservative in IBD detection, and have less apparent IBD peaks.

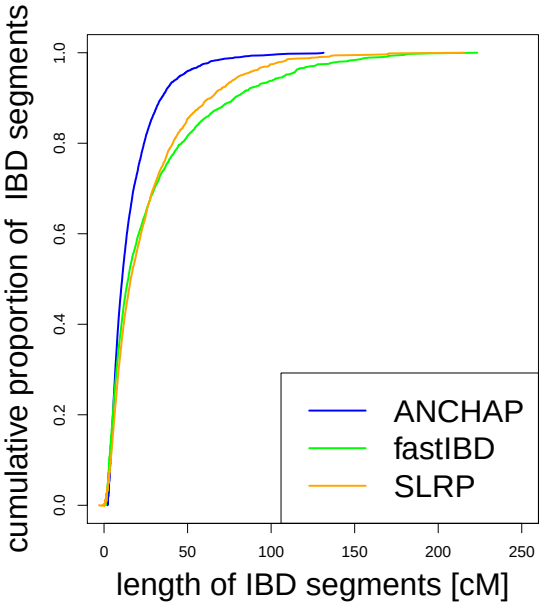



Figure 2.11: Lengths of detected IBD segments [cM]. IBD regions detected by ANCHAP are generally shorter than ones identified by fastIBD or SLRP.

### 2.3.6 Sharing in different cohorts and across the genome

The average number of haplotype sharers per SNP varied from 9.4 in CROATIA-KORCULA, through 12.3 in ORCADES and 12.6 in CROATIA-VIS. In SOCCS which consists of genotypes of nominally unrelated individuals, there were only 0.9 sharers per locus on average.

The frequency of haplotype sharing varies not only between the cohorts, but also across the genome. Figure 2.12 shows average counts of haplotype sharers in different locations across the genome. Drops at the telomeres can be consistently observed, as well as the peaks on chromosomes 2, 6, 8, and 9. In SOCCS particularly notable are the peaks on chromosomes 2 and 6, that also occur in ORCADES and CROATIA-VIS but not in CROATIA-KORCULA.

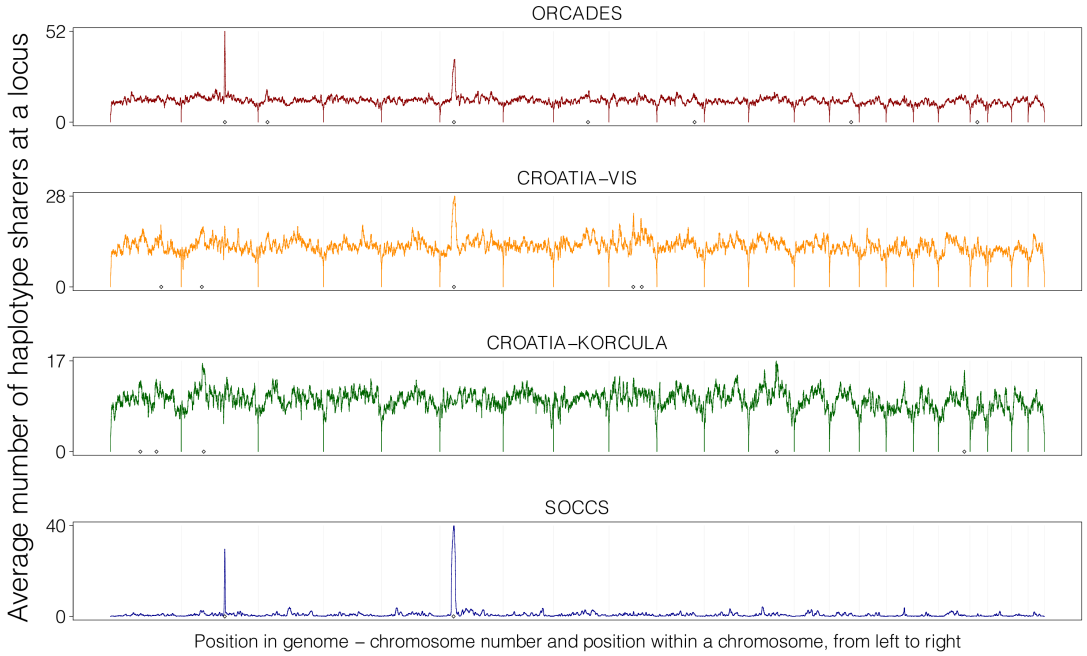



Figure 2.12: Density of haplotype sharers across the genome, in the four cohorts. The horizontal axis shows index of a SNP, and not its physical or genetic location. Genetic positions of the peaks highlighted with diamonds are given in Tables 2.7, 2.8, 2.9, 2.10.

### 2.3.7 Regions of increased frequency of IBD

In Figure 2.12 marked with dots at horizontal axes are regions where recent IBD is particularly common. Tables 2.7, 2.8, 2.9, 2.10 list positions of these peaks in the genome. Next to the locations of peaks, shown are references to studies where the same peaks have been found in different outbred and inbred populations, and interpretation or names of genes present.

| chromosome | position left [kb] | position right [kb] | studies reporting differential selection in region | interpretation |
|---|---|---|---|---|
| 2 | 134144 | 138947 | Moskvina, Han | Immunity (AMSD), lactase (LCT) |
| 3 | 15484 | 24365 | | |
| 6 | 27145 | 33161 | Moskvina, McEvoy, Albrechtsen, Han | HLA region - immunity, zinc fingers |
| 8 | 95306 | 97626 | Albrechtsen, Han | COH1, VPS13B, COX6C |
| 10 | 100639 | 119196 | | |
| 14 | 77965 | 88690 | Albrechtsen | |
| 19 | 18379 | 34464 | | |

Table 2.7: Positions of peaks in frequency of IBD in ORCADES (build 36)

| chromosome | position left [kb] | position right [kb] | studies reporting differential selection in region | interpretation |
|---|---|---|---|---|
| 1 | 186888 | 190805 | Han | RGS1 |
| 2 | 47152 | 59773 | Albrechtsen | |
| 6 | 25952 | 33936 | Moskvina, McEvoy, Albrechtsen, Han | HLA region - immunity, zinc fingers |
| 9 | 80562 | 83295 | | |
| 9 | 101090 | 106669 | | |

Table 2.8: Positions of peaks in frequency of IBD in CROATIA-VIS (build 36)

| chromosome | position left [kb] | position right [kb] | studies reporting differential selection in region | interpretation |
|---|---|---|---|---|
| 1 | 90094 | 101013 | | |
| 1 | 167719 | 177243 | Han | NME7, BLZF1, C1orf114, GPR52, TNR |
| 2 | 54456 | 63368 | | |
| 12 | 77079 | 90001 | | |
| 18 | 64446 | 66196 | | |

Table 2.9: Positions of peaks in frequency of IBD in CROATIA-KORCULA (build 36)

| chromosome | position left [kb] | position right [kb] | studies reporting differential selection in region | interpretation |
|---|---|---|---|---|
| 2 | 134028 | 139092 | Moskvina, Han | Immunity (AMSD), lactase (LCT) |
| 6 | 25535 | 33096 | Moskvina, McEvoy, Albrechtsen, Han | HLA region - immunity, zinc fingers |

Table 2.10: Positions of peaks in frequency of IBD in SOCCS (build 36)

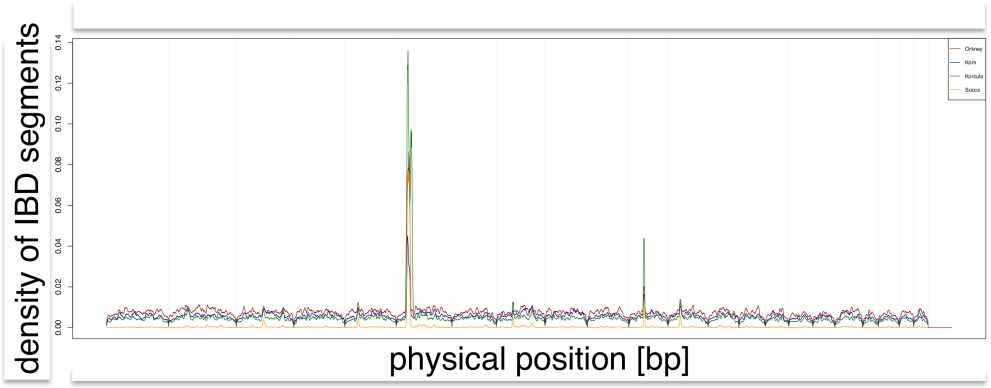



Figure 2.13: Frequency of IBD sharing throughout the genome, when lengths of IBD segments were expressed as a number of SNPs.
ORCADES - red, CROATIA-VIS - blue, CROATIA-KORCULA - green, SOCCS - orange.
IBD peaks on chromosomes 6 and 11 had been apparent before a genetic map was used for measuring length of IBD segments, and indirectly measuring time to common ancestor from whom a haplotype was co-inherited.

Figure 2.13 shows peaks on chromosomes 6 and 11, which we removed by specifying IBD thresholds as minimum length in centiMorgans, rather than as minimum number of SNPs in a region. In Figure 2.13 the peak on chromosome 6 falls in the HLA region, where according to the genetic map recombinations are infrequent, but the density of SNPs per centiMorgan is unusually high. The shared haplotypes recovered around the peaks were probably co-inherited from common ancestors in very remote past. The more ancient the common ancestor, the more contemporary lineages there that carry the haplotype, which would explain the peaks of IBD frequencies. As using the genetic map eliminates the drastically high peaks, we believe that time to common ancestors of haplotype of same lengths can be dated back to similar times in the past throughout the genome.

## 2.4 Discussion

In the previous sections we described ANCHAP - a new heuristic-based algorithm for detecting recent IBD sharing from SNP data. The algorithm was compared against SLRP - a probabilistic long range-phasing method and fastIBD - a short-range program designed for studies containing nominally unrelated individuals. We found that ANCHAP is an order of magnitude faster than SLRP and has lower false discovery rate than fastIBD. We studied recent IBD in three isolated populations and in a study of nominally unrelated individuals. We noted that in some parts of the genome IBD sharing is particularly frequent.

### 2.4.1 Comparison with other methods for IBD detection

Design of the algorithms affects the performance of the methods for IBD inference. Systematic Long Range Phasing (SLRP) is a model-based probabilistic approach for simultaneous IBD detection and phasing. It can simultaneously handle genotyping errors and phase uncertainty, yet this comes at the price of high computational demand. The loopy belief propagation algorithm, which SLRP uses for inference, may not find the optimal solution and is not guaranteed to converge. ANCHAP does not explicitly model genotyping errors; phasing and IBD detection are separate steps, yet in our test detects recent IBD as well as the computationally more expensive SLRP. When a genotyping error gives rise to a pair of opposing homozygotes in a region of IBD sharing, the region of sharing detected by ANCHAP may be shorter, or missed altogether. However, in ORCADES we encountered on average only 1 opposing homozygote per 10,000 markers in genotypes of parent-offspring pairs, so genotyping errors will not prevent most of the shared regions from being detected. FastIBD is a method for IBD detection designed for general populations (Browning and Browning, 2011*a*). It builds a model of haplotypes which can capture only short range allele correlations. This deficiency is then ameliorated by sampling multiple haplotypes for each individual, and checking overlap of such samples between pairs of individuals. FastIBD was more sensitive than ANCHAP or SLRP, but also returned more false discoveries. A possible explanation for why fastIBD yields more false discoveries is that haplotype re-sampling of short blocks may occasionally yield

haplotype matches between samples by chance.

It seems unlikely that the differences in sensitivity and false discovery rate between the methods would seriously affect the uses of detected IBD region in mapping complex traits or IBD-based imputations. Sensitivity approaching 100% would be desirable for downstream applications (Browning and Browning, 2011*b*), but none of the methods achieves sensitivity of IBD detection of more than 81% for the ORCADES data. In case of ANCHAP this probably results from inability to handle the incorrect assignments in Stages I and II, which trigger phasing errors, so that IBD is no longer detected in Stage III. Ability to recover from sporadic phasing errors would certainly improve the sensitivity of IBD detection. For SLRP incomplete IBD detection could be due to limitations of the inference algorithm, conservative approach to declaring IBD, or low tolerance to inconsistencies between the IBD sharing relationship and possibly noisy data. In case of fastIBD, if for a pair of individuals sharing IBD there are few haplotypes which would explain the genotypes, the program may not sample the matching pair. In accordance with this observation, the highest sensitivity we could achieve was by runs of fastIBD with scale parameter set to 4.0, repeated ten times. Together with the sensitivity going up to 89%, the false discovery rate also grew to 7%.

When comparing the design of the algorithms, SLRP seems more elegant and flexible than ANCHAP. The SLRP algorithm not only handles genotyping errors, but also discovers IBD regions and haplotyping simultaneously. However, inference on the Bayesian network in SLRP is computationally very expensive, and its implementation relies on simplifications that likely impede performance of SLRP. ANCHAP, on the contrary, has the advantage of simplicity in implementations which, according to the results of the experiments, does not impair the performance.

The comparison of methods as well as parameter tuning are based on presence of parent-offspring trios among the ORCADES samples. Genomic locations of endpoints of reference sharing regions as determined by parent-offspring phasing are only as accurate as the SNP density allows. The reference regions may still have false endpoints because a recombination may not be detectable from SNP alleles. However, the endpoints should not affect the results of the comparison, as they will be small compared to the regions themselves; long matching haplotypes

that do not descent from a common ancestor are unlikely. In the absence of parent-offspring trios, several types of inconsistencies between the genotype data and IBD relationships recovered could indicate errors. For example, the multipoint genotypes of haplotype sharers, which are recognized to carry one of the probands haplotypes, cannot have opposing homozygotes with respect to not just the proband, but also each other.

### 2.4.2 Genetic maps and peaks of IBD

ANCHAP requires setting a minimum length of a segment between two multipoint genotypes without opposing homozygotes, and above this threshold the algorithm infers that two samples share a haplotype IBD. The threshold is expressed in centiMorgans, with respect to a genetic map (Consortium, 2007; Myers et al., 2005). By expressing the threshold on a haplotype segment shared IBD in centiMorgans, we indirectly limit the times to common ancestors from whom the haplotypes were co-inherited. Such a correction is evident in the SOCCS data, where using the HapMap genetic map markedly reduces the size of the peaks for apparent IBD sharing on chromosomes 6 and 11, as shown in Figure 2.13.

Still, the genetic map may not fully account for population history, for example the extent of linkage disequilibrium in DNA of isolate founders in different parts of the genome, or selection pressure that favoured some variants in relevant parts of the genome. In regions of extended linkage disequilibrium haplotypes may be very similar to each other, which may be confusing for ANCHAP. If haplotypes are very similar to each other, and we observe only unphased SNP genotypes, our method may declare IBD incorrectly even when two individuals do not share a recent common ancestor and their full sequences are not identical. If selection acted on some part of the genome, it might have increased frequencies of some haplotypes. ANCHAP would detect the selection signature as increased frequency of IBD sharing in a region. Because of selection, times to common ancestors from whom the haplotypes were inherited would date further back than elsewhere in the genome. This gives rise to regions of increased frequency of IBD sharing.

Even when using the genetic map, we can observe regions of excess IBD sharing also in SOCCS, a cohort composed mainly of unrelated individuals from

across Scotland, around the two peaks on chromosome 2 and 6. These peaks also occur in ORCADES and partially in the Croatian cohorts. Any two individuals in SOCCS are likely to have only very remote common ancestors. Because selection pressure or high linkage disequilibrium in the peak regions are not taken into account, ANCHAP confuses the ancient common ancestors with more recent ones.

We thus believe that while in general by using the genetic map we limit the inference of IBD to segments with recent common ancestors from whom haplotypes were co-inherited, in some parts of the genome we may not account for extensive linkage disequilibrium among isolate founders or selection events.

### 2.4.3 Identity by descent and positive selection

Albrechtsen et al. have shown in simulations that positive selection gives rise to excess IBD sharing (Albrechtsen et al., 2010). They also detect where recent positive selection might have acted in the 11 populations from HapMap. They find peaks on chromosome 6 in the HLA region and on chromosome 8, which contains the defensin gene.

A new method for IBD detection based on a hidden Markov model is described (Han and Abney, 2012). The model gives probabilities of all 9 IBD states between pairs. The model is employed in a search for positive selection in genotype samples from Kenya. A simulation study estimates the accuracy of the method. The authors quantify IBD rates across the genome, and exclude the possibility that the peaks are a reflection of increased linkage disequilibrium. In the increased regions, they search for evidence for positive selection in literature. Around 50 signals were found, half of which are novel. Only the literature search is presented as validation of the results, and no evaluation of significance is provided. They also find peaks in the HLA region, and one on chromosome 11 which contains clusters of olfactory receptors.

The HLA region is known for extensive conservation of haplotypes spanning it (Ferreira et al., 2012). HLA molecules are expressed on many human cells, and more than 100 SNPs in the HLA region have been implicated in autoimmune and inflammatory conditions. Many of the described associations are with hap-

lotypes, as otherwise the mapping efforts have been thwarted by the high linkage disequilibrium.

### 2.4.4 Detection of positive selection through allele frequencies

Genes under positive selection would reveal themselves through excess IBD, but as a consequence also through altered allele frequencies. If several haplotypes carry a variant under selections in a population, this would affect the allele frequencies in the region. Moskovina et al. (Moskvina et al., 2010) analyse differences in allele frequencies between participants of a schizophrenia study from Bulgaria, Ireland, Scotland, Sweden and Portugal. They detect SNPs where allele frequencies are significantly different in the populations. They list 11 top regions and annotate them with gene ontology software.

Equivalently, traces of natural selection can be seen by computing Wright's fixation index $F_{ST}$ (McEvoy et al., 2009). They identify 11 peak regions, annotate them with genes they contain, and provide interpretation. Furthermore, they demonstrate haplotype analysis of the HLA peak. They demonstrate a commonly shared haplotype longer than 3 cM.

### 2.4.5 Possible improvements to the algorithm

To use ANCHAP in new studies, the the program would benefit from improvements like explicit handling of genotyping errors and parallel execution.

#### 2.4.5.1 Explicit handling of genotyping errors

At the moment the genotyping errors in array data are not explicitly accounted for. I estimated in data from ORCADES that opposing homozygotes due to genotyping errors occur only once per 10,000 SNPs. Because of the heuristics used and the amount of data, the genotyping errors should be irrelevant. In the first Stage of the algorithm, the inference relies on presence of opposing homozygotes, which are unlikely to occur by chance. At Stage II, phase is decided based on

genotypes of multiple sharers, most of whom are likely to be correct. Any single genotyping error is unlikely to affect the haplotypes being phased.

However, explicit modelling of genotyping errors will be important for detection of IBD segments from next-generation sequencing data. In next-generation sequencing data, SNP calls carry uncertainty depending on read depth. Handling of genotype errors could be achieved through introducing hidden Markov models, where hidden states correspond to whether a haplotype is shared, and observed variables are possibly noisy genotypes.

#### 2.4.5.2 Implementation

The algorithm has been implementing as an R package, for ease of prototyping and sharing. String matching at Stages I and III was implemented in C++ for efficiency. Further improvement would be achieved if more of the code was translated into C++, for example whole of Stage II.

While for cohorts with 1000 samples the running times are less than one day on a compute cluster, running it on sets with 10000 individuals requires running several smaller jobs in parallel. This can be achieved by dividing the genome into chromosomes, as processing of chromosomes is independent. Furthermore, in Stage II all individuals are processed independently of one another, so these computing jobs could be run in parallel.

### 2.4.6 Conclusions

We have described methods for detecting regions of haplotypes shared IBD from SNP data. We now proceed to applications of inferred IBD regions: optimisation of resequencing studies, studying diseases with Mendelian subtypes and genomic predictions.

# Chapter 3

# Optimisation of resequencing studies in population isolates based on identity by descent

## 3.1 Background

The aim of resequencing studies in population isolates is to identify effects of variants which outside of an isolate could be very rare. While cost of next generation sequencing is still significant, we can reduce costs of resequencing by exploiting widespread identity by descent between the subjects. The reduction of cost can be obtained by sequencing a subset of individuals whose haplotypes are representative of study, and using identity by descent to impute genotypes for non-sequenced samples. Imputations that rely on identity by descent (IBD) are now recognized to increase the power of sequencing studies in population isolates (Zeggini, 2011).

The aim of this chapter is to assess whether recent identity by descent discovered in array data can be useful for optimising resequencing studies. This could be achieved by optimal selection of individuals for resequencing, as well as by more accurate imputations. We investigate accuracy of regions of identity by descent inferred by ANCHAP, which has implications for imputations and the algorithm of ANCHAP. Later we describe an algorithm for optimising the design

of resequencing studies in isolated populations. Finally, we discuss the design of an algorithm for imputations informed by identity by descent.

### 3.1.1 Methods for optimisation of resequencing studies

Imputation models can be categorised into short and long-range ones, with the former relying on linkage disequilibrium and the latter on longer regions of recent identity by descent between samples.

#### 3.1.1.1 Short-range imputation methods

Most short-range imputation methods use hidden Markov models (HMMs). HMMs represent genotypes as mosaics of haplotypes from remaining individuals. The visible states correspond to genotype data, and the hidden states correspond to a haplotype mosaic. The models typically require parameters that represent recombination and mutation rates, and rely on haplotype blocks of 10-100 kb (Daly et al., 2001). Where there are also longer, more recent IBD segments between samples, short-range programs will not necessarily make use of them, because they do not prioritise using longest haplotype segments as templates. In contrast, long-range methods use recent IBD segments, however they will not work at all where there is no recent IBD.

Three examples of short-range imputation methods are IMPUTE2 (Howie et al., 2009), MACH (Li et al., 2010) and Beagle (Browning and Browning, 2007). MACH represents haplotypes in study samples as a mosaic of reference haplotypes from the imputation reference panel only, for example HapMap haplotypes (Gibbs et al., 2003). MACH first creates a haplotype mosaic for a subject using SNPs typed in both study and reference samples, and accordingly imputes the genotypes not typed in study samples. In contrast, IMPUTE2 uses also haplotypes from other study samples in addition to the reference haplotypes. In IMPUTE2 the sampled mosaics are used for phasing only at SNPs typed for both study and reference samples, and haplotypes at these universally typed SNPs are then compared with reference haplotypes, assuming the phasing to be correct. In addition to reference haplotypes, IMPUTE2 also allows use of unphased genotype reference samples. Beagle uses a similar imputation procedure, however it

employs a more parsimonious model of haplotypes. All of the short-range imputation programs first perform phasing at universally typed SNPs, and subsequently use the haplotypes for imputations.

#### 3.1.1.2 Long-range imputation methods

Imputations could exploit long-range identity by descent, rather than using only short co-inherited haplotypes. Where recent IBD is found, such imputations could be more accurate, particularly for recent mutations, which create rare variants. No algorithm has been described which would use IBD to perform sequence imputations.

IMPUTE2 does use the analogous concept of 'surrogate family', however not to find segments of IBD, but rather to speed up the algorithm. For a proband, its 'surrogate family' consists of samples most similar to proband's genotype and haplotypes, in terms of Hamming distance in the region considered (Howie et al., 2011). In phasing, focusing on 'surrogate family' members reduces the number of hidden states of the hidden Markov models. For imputation, only haplotypes of 'surrogate family' members are used from the reference samples. As each sample is imputed using his custom set of reference samples, this allows using very large reference sets. Because of the approximation IMPUTE2 may often make use of long identical-by-descent haplotypes, but this is not explicit or guaranteed.

#### 3.1.1.3 Optimisation of resequencing studies using recent IBD

Optimisation of resequencing studies was demonstrated in an extremely isolated population of the Pacific island Kosrea in Micronesia (Gusev et al., 2012). They proposed both an algorithm for choosing samples from resequencing, and a way to evaluate regions of IBD inferred from array data against sequence data. In the pilot study, they chose the seven sequencing samples, according to SNP genotypes previously available. The selection of individuals is driven by their algorithm called INFOSTIP, which exploits presence of IBD segments between individuals in a study. It is a greedy algorithm for optimising resequencing studies, which is described here in detail before I describe my modifications.

In INFOSTIP (Gusev et al., 2012), the choice of next individual for sequencing

is driven by its utility given the previously selected individuals. Let:

- $P$ be the set of individuals in a population,
- $Q$ - the subset of individuals already sequenced,
- $i$ - index of an individual,
- $k$ - a locus in the genome,
- $R(i, q)$ - set of IBD segments between individuals $i$ and $q$,
- $G$ - the set of SNPs in genome.

Chosen for resequencing each time is an individual that maximises the utility of sequencing given $Q$, $U(i, Q)$. The utility is defined with respect to total information content ($TIC$) of set $\{Q, i\}$ about the whole population $P$, and by information content of the set $Q$ about $P$.

$$U(i, Q) = TIC(P, \{Q, i\}) - TIC(P, \{Q\}) \tag{3.1}$$

Total information content expresses how much information about genotypes of total cohort with individuals $P$ is known if individuals in $Q$ are sequenced, however it is not a measure coming from the field of information theory. Rather, it is fraction of all SNP genotypes in data for $P$ that would be known through either sequencing individuals in $Q$ or ones that could be imputed from $Q$ to $P$, and the number of all SNP genotypes in $P$. The number of SNP genotypes that could be imputed for non-sequenced individuals is $\Sigma_{i \in P \setminus Q} L(i, Q)$.

$$TIC(P, Q) = \frac{|Q|G + \Sigma_{i \in P \setminus Q} L(i, Q)}{|P|G} \tag{3.2}$$

For an individual $i$ not in $Q$, the amount of information that can be imputed is obtained by summing over genetic loci. This uses an indicator whether a genotype of individual $i$ can be imputed at locus $g$ from sequenced samples in $Q$.

$$L(i, Q) = \Sigma_{g \in G} I(i, g, Q) \tag{3.3}$$

Lastly, the indicator is defined with respect to the IBD segments $R$. A genotype can be imputed for individual $i$ at locus $k$ if there exists ($\exists$) an individual $q$, such that $i$ and $q$ share IBD in a region containing $k$.

$$I(i, k, Q) \quad = \quad \begin{cases} 1 & \exists q \in Q(\exists (l, r) \in R(i, q)(l < k \wedge r > k)) \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Therefore, when selecting a new individual, the algorithm considers all regions of genotypes where they share IBD with sequenced sample, as known. In practice, they would be only known when there is a haplotype sharer for both gametes of a proband, and the genotypes of a sequenced samples are homozygous, through a similar argument like for long-range phasing algorithms in Chapter 2.

The algorithm chooses individual $i$ that maximises $U(i, Q)$, and adds $i$ to $Q$. This is a greedy procedure, which is a necessary approximation given that in general the maximal coverage problem is NP-hard. An important part of INFOSTIP is the data structure for storing IBD regions, such that they can be efficiently queried for locations.

The algorithm was evaluated through the quality of imputations at deliberately concealed SNPs. The authors did not propose an IBD-based imputation algorithm, but rather used Beagle, a short-range imputation method. It was shown that random selection of individuals gave imputations of considerably lower quality of imputations than when prioritised according to INFOSTIP.

Furthermore, the authors also evaluated quality of IBD segments against the whole sequence data for seven selected samples. For a pair of samples that are IBD in a region, there should be no opposing homozygote genotypes in sequence data. If a an individual carries two copies of a rare allele at SNP, and his haplotype sharer carriers no copies of the rare alleles at the SNP, this is not consistent with a haplotype being shared. The authors counted all opposing homozygotes between pairs in regions previously inferred as identical by descent. They also

counted all loci where such inconsistencies could have happened, namely where both samples are homozygous at a SNP in a region that the two individuals share IBD, and when at least one genotype has two copies of the rare allele. Let $I$ be the number of opposing homozygotes between pairs of IBD sequences and $A$ number of mutually homozygous genotypes between pairs of IBD sequences, with at least one being non-reference allele. Concordance $C$ is defined as the ratio of non-opposing homozygote genotypes in exome data $(A - I)$ over $A$:

$$C = \frac{A - I}{A} \tag{3.5}$$

Opposing homozygotes suggest incorrectly detected IBD or low quality of sequence genotype calls. The authors of INFOSTIP take the concordance rate as a measure of imputation accuracy when only one individual from the pair was sequenced. By chance match of alleles, the concordance will not be zero even at non-IBD segments, therefore a background concordance of sequence SNPs is calculated for reference.

This method of optimising resequencing studies and evaluating IBD segments against sequence data was a starting point for my analysis of data from the ORCADES study.

## 3.2 Methods

### 3.2.1 Array data

Array data has been merged from ORCADES data set, obtained with a Illumina Human Hap300 array with 293,687 SNPs, and from the Orcadian multiple sclerosis study, where a Omni1 array was used with around 1 million SNPs. After merging and quality control, there were 171,755 genotypes for 908 individuals, from which IBD segments were identified.

### 3.2.2 Exome data

Individuals for exome sequencing were chosen from ones who had previously been genotyped. The subset of individuals was selected to minimise relatedness between samples and maximise representation of haplotypes, using my algorithm described later in this chapter.

Whole exome sequences were generated using the Agilent SureSelect All Exon 50 Mb kit. Average depth of reads was 29.5 x. Read alignment was done with reference to human genome build 19, using Stampy (Lunter and Goodson, 2011). GATKs genotyper generated genotype calls, using default parameters (McKenna et al., 2010), and identified and 217,015 variants. Alignment of reads and SNP calling were performed by Ross Fraser (Joshi et al., 2013).

The resulting genotypes underwent rigorous quality control, since in further analysis we would like to assume their full correctness. Kept were only such variants with phred-scaled quality of more than 40, called in at least 50% of subjects, and with minor allele frequency more than 0.75%. Further criteria included Hardy-Weinberg equilibrium test with p-value greater than $< 10^{-4}$, and mapping to homologous regions. After quality control, there were 100 samples with 159594 SNPs. Retained were only the sites with quality exceeding 40, and genotypes with quality at least 20. Available for analysis was exome sequence data for 99 individuals from the ORCADES study, previously genotyped with GWAS array. There were 7730 SNPs in common between the array and called exome variants, which we used to assess the match between these two data types. We excluded seven samples for whom the correlations between the exome and array genotypes was less than 0.5, because we inferred that genotyping and sequencing were done on different samples and the sample identities were misleading. For the remaining 92 average correlation between exome and array genotypes was 0.92. The imperfect match between array and sequencing genotypes could result from errors in areas of lower sequencing depth. The resultant exome genotypes were used for evaluating accuracy of inferred IBD segments.

### 3.2.3 Evaluation of identity by descent established from array data

Identity by descent inferred from array data should also hold for untyped loci in the region. We can thus confirm postulated IBD-regions between two samples using SNPs in exome sequence data. Even though the latter is unphased and possibly noisy, absence or presence of opposing sequence homozygotes between putative haplotype sharers can cast light on quality of the inferred IBD regions.

As a measure of consistency between sequence genotypes in the IBD segments detected from array data we used concordance as defined in Section 3.1.1.3. This enables comparisons with earlier work, and concordance also has the interpretation as the probability of correct imputation from sequence of IBD sharers. Finally, since the computation involves only pairs of genotypes where at least one is homozygous on the non-reference allele, it is approximately independent of allele frequencies.

The procedure of evaluating IBD regions with sequence data thus involves iterating over all IBD segments identified from array data, and counting opposing homozygotes between sequences of the pairs. In this way we can see if the opposing sequence homozygotes occur more frequently in some regions of the genome, whether they occur more often in tails of such segments, and whether the second round of ANCHAP indeed improves the quality of IBD inference. The procedure is described in Figure 3.1.

We repeated the procedure for control IBD segments. The control segments are the same as the ones detected from data, but the identities of the individuals from whom they come were randomly permuted. As long IBD between a pair of random samples is unusual, such control segments are unlikely to be in IBD regions, but have same lengths as the IBD segments detected. The control segments allow as computing background concordance rates in non-IBD regions.

Figure 3.1: Evaluation of IBD segments against exome SNPs: computing concordance $C$

---

Input: $E$, $IBD.segments$

$E$: Exome SNP matrix: $N \times M$ (number of samples by number of exome SNPs), with $E_{i,j} \in \{0, 1, 2, \text{NA}\}$ - allele dosage

$IBD.segments = \{(\text{id1}, \text{id2}, \text{start.SNP}, \text{end.SNP})\}$ (list of IBD segments from array data, each with identities of sharers, and indices of start and end SNPs)

1. initialise $A \coloneqq 0$, $I \coloneqq 0$
2. **loop** through $IBD.segments$
   (a) access the pair of exome SNPs in $E$ relevant to an IBD segment
   (b) count opposing homozygotes between the pair of exome segments: $i_i$
   (c) count mutual homozygotes in the exome segments: $a_i$
   (d) update to summary variables: $I \coloneqq I + i_i$, $A \coloneqq A + a_i$
3. $C = \frac{A-I}{A}$ (Calculate concordance of exome SNPs in IBD regions)

   Output: $C$

---

### 3.2.4 Description of algorithm for selection of samples in resequencing studies

The uncovered IBD sharing within a cohort can be used for efficient selection of individuals to resequence, with a view to using them as a reference for imputation. Selection of individuals for resequencing is based on maximizing representation of haplotypes and minimizing multiple resequencing of the same haplotypes. As

the first individual for resequencing we choose the one whose haplotypes have the most copies in the rest of the cohort. After excluding regions that have been covered by sharing with individuals already chosen, we repeat the procedure of selecting the individual with most copies until a target level of coverage has been achieved (Figure 3.2).

The algorithm is very similar to one described in (Gusev et al., 2012) - it is also greedy and picks individuals for sequencing to maximise information content. However, the algorithm presented here also takes into account parent-of-origin of recent IBD. To consider a segment of unsequenced genotypes imputable, our algorithm requires at least one sharer of each of the two haplotypes. We therefore modify the total information content, by a factor of 2 to reflect that now we wish to impute an allele on each of the two haplotypes of a proband:

$$TIC'(P,Q) \quad = \quad \frac{2 \times |Q|G + \Sigma_{i \in P \setminus Q} L'(i,Q)}{2 \times |P|G}, \tag{3.6}$$

where

$$L'(i,Q) = \Sigma_{g \in G} I'(i,g,Q). \tag{3.7}$$

We also modify the indicator $I'$ of whether a genotype is imputable to indicate how many alleles of a proband can be inferred: 2, 1 or 0. To do so, we augment the description of IBD segments. The sets of IBD segments of proband $i$ with individual $q$ such that $i$ and $q$ share IBD on the first and second gametes are denoted by $R'_1(i,q)$ and $R'_2(i,q)$. Both alleles can be imputed if there exist haplotype sharers for both gametes (hence logical 'and' denoted by $\wedge$), one allele can be imputed if there is a haplotype sharer of either gamete (hence logical exclusive 'or' denoted by $\veebar$), 0 otherwise, as shown in Equation 3.8.

$$I'(i,k,Q) \;=\; \begin{cases} 2 & \exists q1 \in Q(\exists (l,r) \in R'_1(i,q)(l<k \wedge r>k)) \\ & \wedge \exists q2 \in Q(\exists (l,r) \in R'_2(i,q)(l<k \wedge r>k)) \\ 1 & \exists q1 \in Q(\exists (l,r) \in R'_1(i,q)(l<k \wedge r>k)) \\ & \veebar \exists q2 \in Q(\exists (l,r) \in R'_2(i,q)(l<k \wedge r>k)) \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

My work on this algorithm had been completed before publication of (Gusev et al., 2012), and we had not been aware of the competing algorithm.

Figure 3.2: Choosing an optimal subset of individuals for re-sequencing

Input: $IBD.segments, targetN$

$IBD.segments = \{(\text{id1}, \text{id2}, \text{start.SNP}, \text{end.SNP})\}$ (list of IBD segments from array data)

$target.n$ - target number of individuals for resequencing

1. initialise $I$ (matrix $|P| \times |G|$ for storing the indicator variables $I$ if a genotype is imputable)
2. initialise $Q$ (empty set of individuals for resequencing)
3. **for** $n \in 1 \ldots target.n$
   (a) **for** $i \in P \setminus Q$ (for all non-selected individuals)
      i. compute $U'(i, Q)$, based on $IBD.segments$ and $I$, according to Equation 3.1
   (b) choose an individual $i.picked = \max_i U'(i, Q)$, add to $Q$
   (c) update $I$

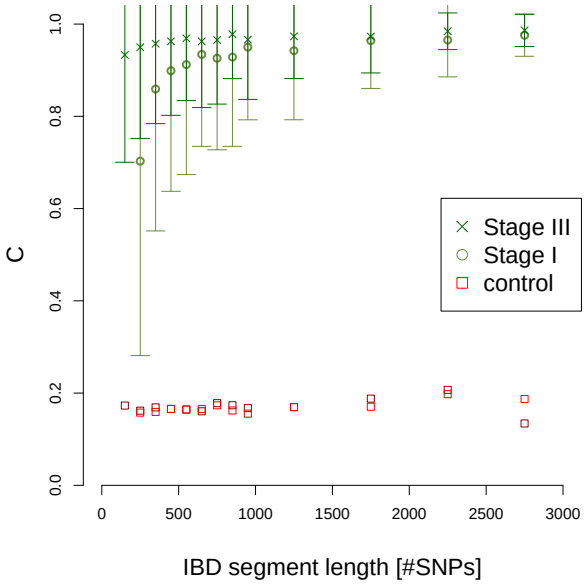
Output: $Q$

## 3.3 Results

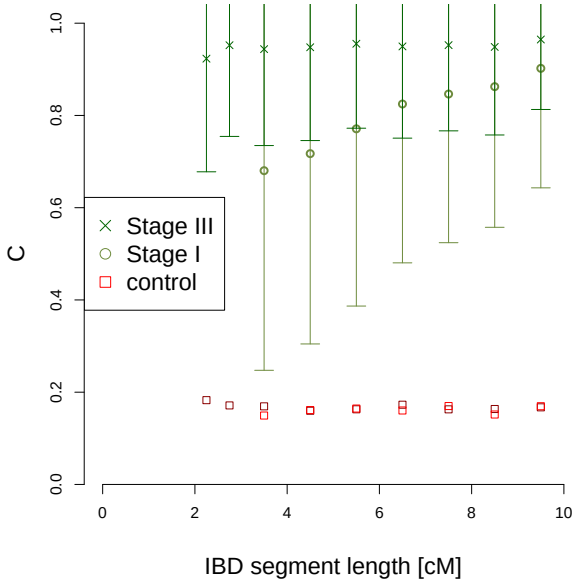
### 3.3.1 IBD inferred from array data against the exome SNPs

Between genotypes of the verified individuals, in Stage I ANCHAP found 33982 IBD segments, and 32868 in Stage III. At an average locus, a sequenced individual

shared IBD with 0.8 other sequenced individuals. Between pairs of IBD sequences from Stage I, there were 8430 homozygotes of opposing alleles in exome sequence data, and 1922 in Stage III. Respective concordance (C) scores were 0.87 and 0.96. The moderately low number of opposing homozygotes in sequence data in IBD regions suggests that the inference of IBD regions is accurate, but because most IBD sequences fall outside of the exons, no verification there is possible. However, the concordance score (C) accounts for the bias of sequence data towards exons.




(a) Concordance (C) of exome sequences in IBD segments of different length, when length is expressed as number of array SNPs a segment contains. The longer a region, the higher the concordance. Irrespective of length, concordance in detected IBD segments is much higher than in control segments.




(b) Concordance (C) of exome sequences in IBD segments of different length, when length is expressed in centiMorgans.

Figure 3.3: Concordance (C) of exome sequences in IBD segments identified from array data, depending on length of the segments.

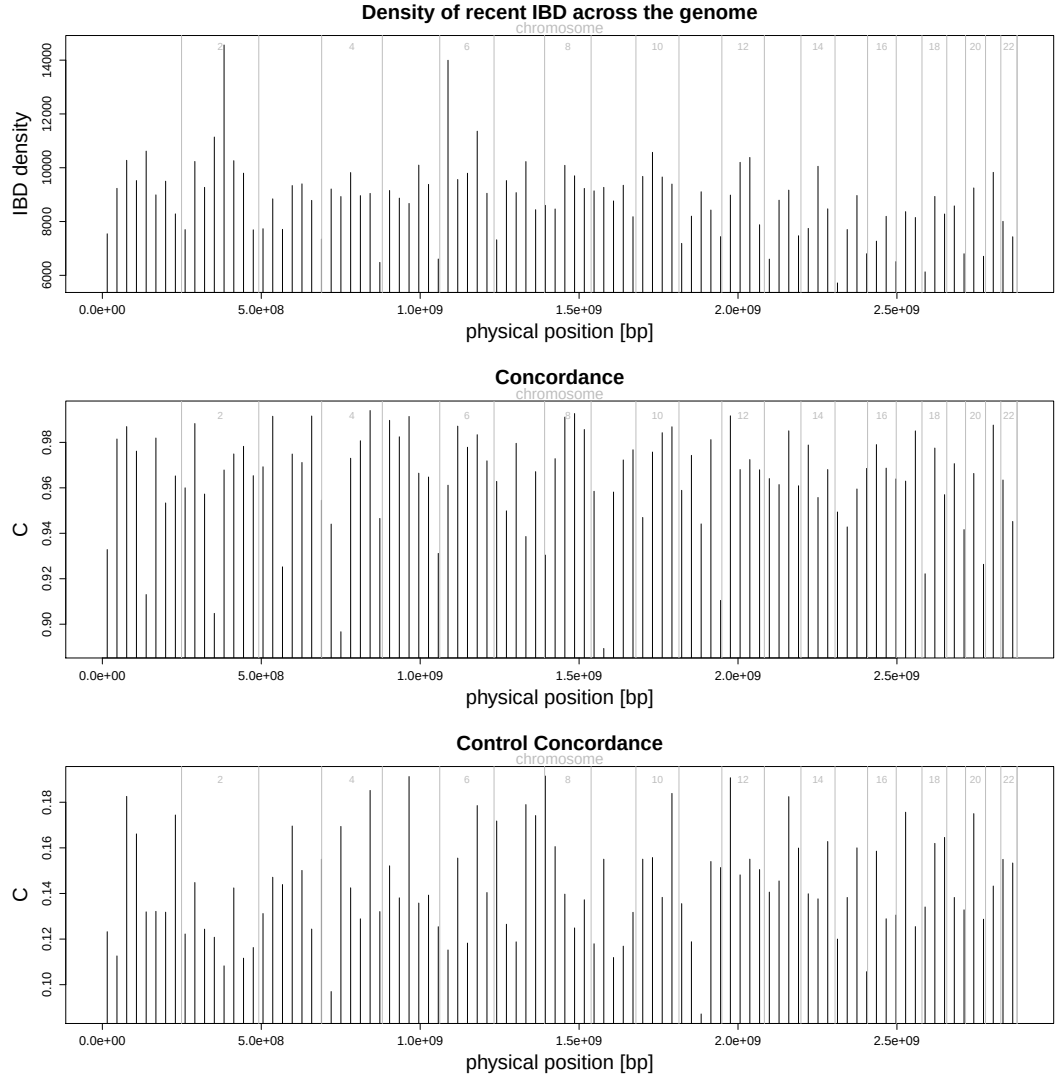



Figure 3.4: Genome-wide view of concordance (C) of exome sequences in IBD segments identified from array data. Note that in regions where sharing IBD is more frequent than elsewhere (chromosomes 2, 6), concordance is typical, within one standard deviation from genome-wide mean and within standard deviation of concordance computed for control segments. This implies that the IBD segments recovered in these regions are not artefacts of the inference method.

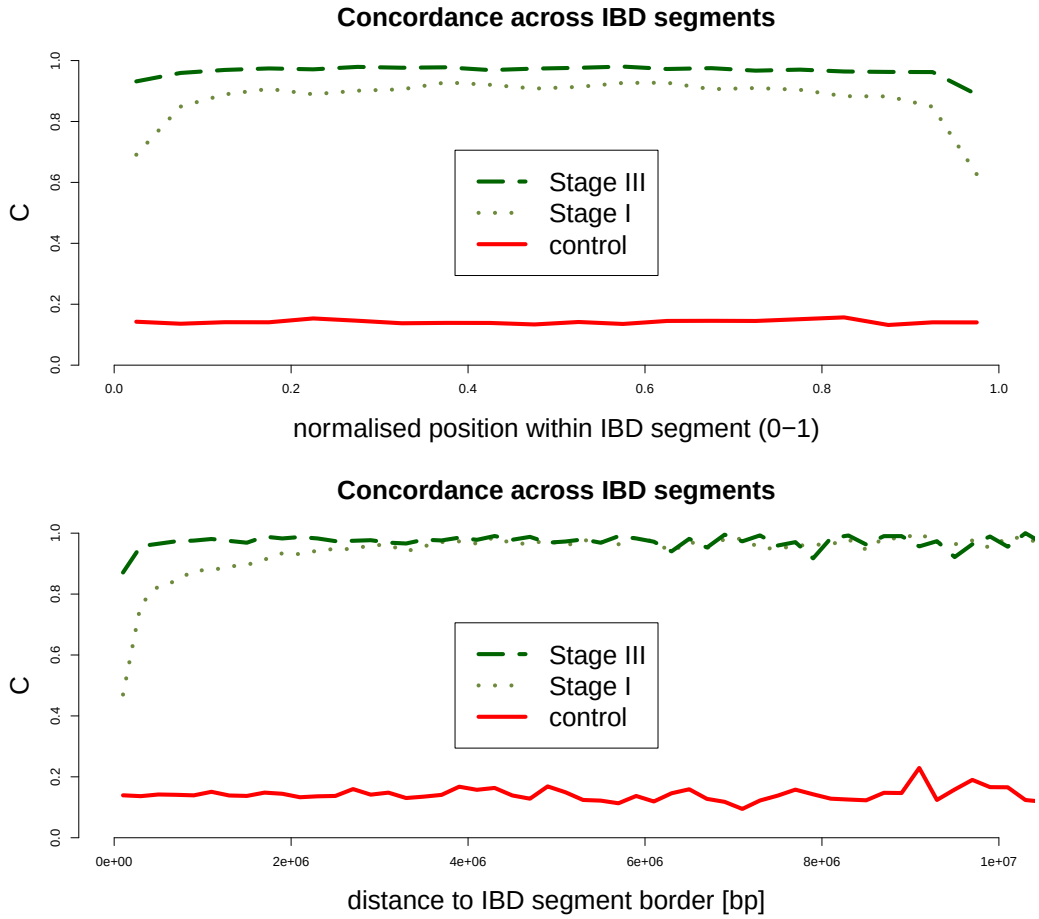



Figure 3.5: Concordance (C) of exome sequences in IBD segments identified from array data, in Stages I and III of ANCHAP. Top: normalised position within IBD segments. Bottom: position with respect to IBD segment borders. Control IBD segments generally give concordance of 0.17, much less than in the IBD segments declared by ANCHAP. As expected, the concordance for the control segments is uniform throughout the segments. Stage III of ANCHAP generally gives better concordance. There is poor concordance close to borders of IBD segments identified in Stage I.

Figure 3.3 shows concordance (C, defined in Equation 3.5) between exome SNPs in regions declared as IBD in array data. IBD regions contain fewer opposing homozygotes than control regions, as depicted in Figures 3.3a and 3.3b. In control regions, the concordance per SNP is 0.18, whereas in IBD regions 0.99 for round 2 in Figure 3.3a. Among the shorter IBD regions, there are some where the exome data suggest they are not IBD, as shown by large standard deviation of error rates, however mean concordance remains at 0.95 even for shortest segments. When IBD threshold is expressed as a number of SNPs in a segment (Figure 3.3a), opposing homozygotes occur only in the shortest segments, but they do occur more often even in longer segments as measured in centiMorgans (Figure 3.3b).

Stage III of ANCHAP reduces not only the mean number of mismatches between sequence segments, but number of outlier segments with very large number of mismatches. Concordance decreases most rapidly with segment length expressed as number of array SNP. Segments longer than 300 markers, which also are longer than 2 cM in round 2, give concordance of 0.98.

Figure 3.4 shows concordance in different parts of the genome, and compares it with the IBD density. On this dataset, in which SNPs are sparser than the one used in Chapter 2, we can still observe peaks of IBD on chromosomes 2 and 6. Interestingly, in the peak on chromosomes 2 and 6 the concordance is typical, within one standard deviation from the mean observed genome-wide. This implies that the peaks of IBD on chromosomes 2 (lactase region) and 6 (HLA region) are genuine, rather than being artefacts of IBD detection.

Figure 3.5 shows concordance with respect to positions within the IBD segments. We could expect to see more of the inconsistencies towards edges of the segments where IBD status is less certain. This can be observed in results for both Stages I and III of ANCHAP. The results justify trimming 50-100 SNPs (5-10e7 bp) at segment borders in round 1, and 25 (2.5-10e7 bp) in round 2. However, in round 2, concordance is high even close to the segment borders.

### 3.3.2 Resequencing optimization

The inferred shared haplotypes in an isolated population can be exploited to increase the efficiency of a sequencing study given a fixed budget. One possible strategy is to identify an optimal subset of individuals for resequencing at high coverage so as to obtain accurate sequence data, then to impute these sequences into the other cohort members with whom they share IBD. For the selection of individuals, our algorithm favours individuals who share the largest regions IBD with individuals who were not chosen for resequencing. We have examined the strategy based on resequencing an optimal 20% of individuals from ORCADES, which would reduce the cost fivefold, with 65% of the unsequenced diploid genomes sharing with the sequenced individuals. Had we chosen the individuals randomly, the IBD coverage of unsequenced haplotypes would have been 61 %, as shown in Figure 3.7.

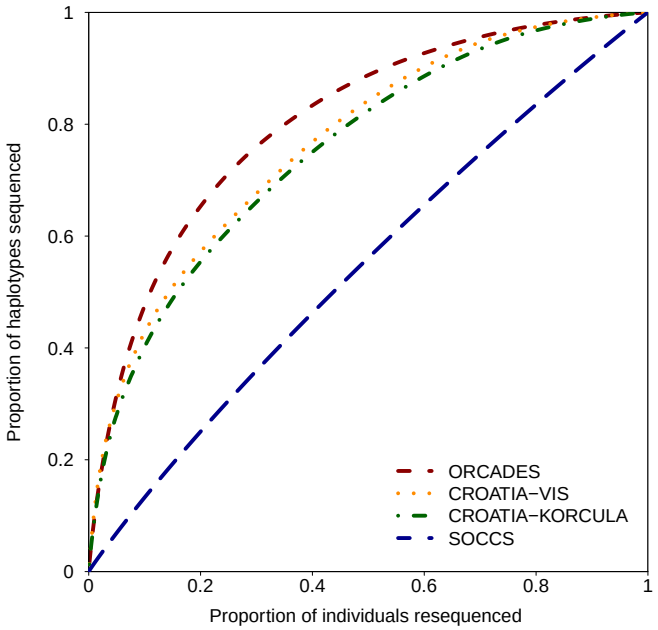



Figure 3.6: Proportion of haplotypes sequenced with respect to proportion of samples sequenced. The curves are affected by relatedness of individuals in a population, and fraction of samples that were added to a study.

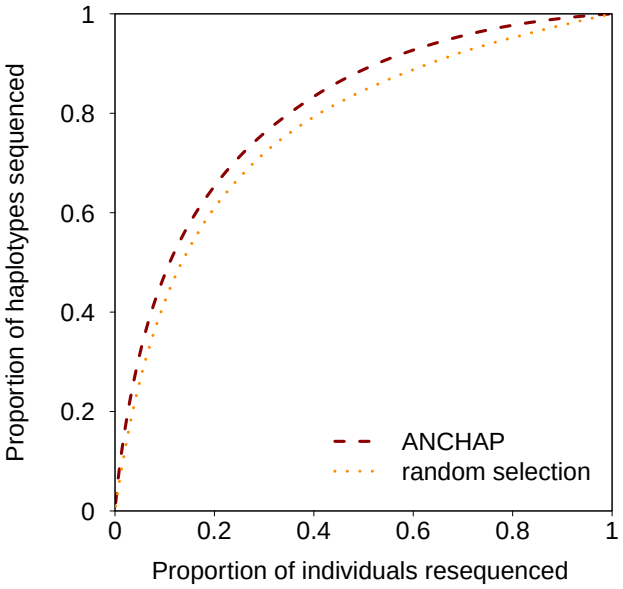



Figure 3.7: Comparison of the algorithm for selecting individuals against random choice - proportion of haplotypes sequenced with respect to proportion of samples sequenced in ORCADES. If samples are selected according to our procedure, the proportion of haplotypes sequenced is always higher than when samples are chosen randomly. For example, when resequencing 20 % of individuals in a study, coverage of haplotypes would be 65 % if individuals are chosen using our method and 61 % if they are chosen randomly.

## 3.4 Discussion

### 3.4.1 Implications for the algorithm of ANCHAP

Exome sequence data show that the IBD segments inferred by the Stage III of ANCHAP are more accurate than ones from Stage I. Trimming borders of the segments in Stage I is definitely necessary, and may still remove errors in IBD segments from Stage III.

Concordance rates (C) in the IBD segments vary throughout the genome. This could be a statistical variation arising from a limited number of sequenced samples. Alternatively, the variation of concordance could be a reflection of features that vary throughout the genome. For example, if in some regions the detected IBD segments date to more ancient ancestors than in other, more mutations might have accumulated. Alternatively, regions of genome with higher concordance could be ones where IBD is falsely detected, for example in regions of poor SNP coverage, or where SNPs are not informative, for example where the minor allele frequencies are very low.

### 3.4.2 Exome sequence data evaluated against IBD segments from array data

Shorter inferred IBD segments show lower concordance with sequence data than longer ones. This could be because shorter IBD segments are more difficult to detect accurately, as they contain fewer array SNPs. Alternatively, for shorter segments the common ancestor was more ancient and mutations were more likely, which reveals itself in opposing homozygotes between sequence genotypes. The relation between length of IBD segments, time to the common ancestor and mutation rates is further commented on in the Discussion of this thesis.

### 3.4.3 IBD-based imputations

The example in Figure 3.8 shows a sample imputation for an individual with three sequenced haplotype sharers. IBD sharing between array SNPs 1 and 2 had been established between the individual and its three haplotype sharers based on the

array data. If any of the sharers carries an a polymorphism in one locus, there is 50% chance the child will carry it too. If any of the haplotype sharers is homozygous on some rare variant, there is 100% chance the child would carry it too. If there are few haplotype sharers, we could calculate probabilities of alleles that the individual carries, based on sharers' genotypes. The resulting sequence-SNP genotype will not be phased, and would contain probabilities of having certain rare variants.

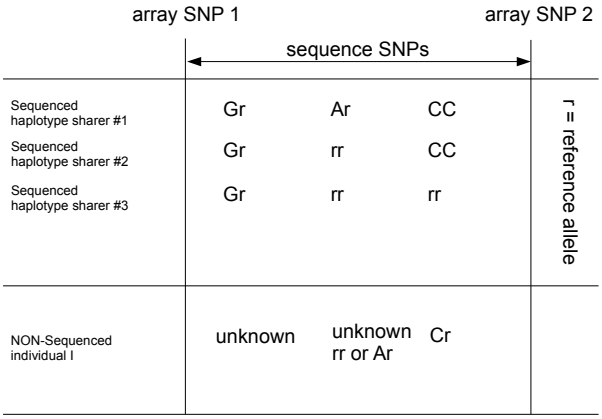



Figure 3.8: Imputation for an individual with three sequenced haplotype sharers

When we know the phase of the SNP genotypes, there is more information for IBD-based imputations. Haplotype sharers of an individual can be divided into two groups corresponding to the two proband's gametes, and often this reduces the number of imputation possibilities. Figure 3.9 shows imputation using phase information for SNP data. With the phase information, the middle locus which couldn't be imputed before, can be imputed.

A number of factors determine the accuracy of imputations based on IBD, as described above. Firstly, only correctly detected IBD would result in correct imputations. Secondly, a variant could be correctly imputed only if it is older than the most common ancestor from whom the haplotype was co-inherited. Longer haplotypes shared IBD should allow for more accurate imputation than with shorter segments for which the common ancestor was much more ancient. Thirdly,

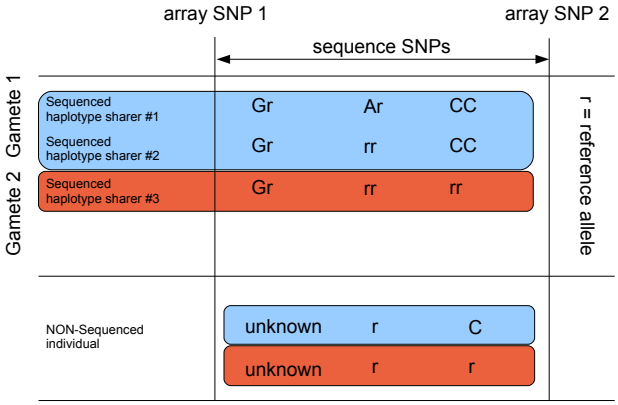



Figure 3.9: Imputation is facilitated when the haplotype sharers are split into two groups (see the middle locus).

imputation relies on the presence of homozygotes among the haplotype sharers, and often more than one sharer is needed, which is not taken into account in the selection algorithm (Algorithm 3.2). Finally, the ease of sequence imputations with the IBD sharing information depends on whether the sequence data can be phased, and whether it is possible to overlay the array and sequence haplotypes. There is work in progress both on algorithmic and laboratory methods for phasing of sequences (Browning and Browning, 2011*b*). If the sequences are not phased, the IBD sharing can still be exploited for imputation, but this would result in more uncertainty about imputed alleles.

For some rare, very recent variants, their carriers should have a recent common ancestor, and all share IBD in the region with each other. If this is identified by ANCHAP, and a haplotype of an unsequenced individual also belongs to such a cluster, he almost certainly also carries the variant. This approach would rely on detecting IBD clusters, and ANCHAP revealing recent IBD alongside parent-of-origin for each segment.

The imputation approaches described so far rely on identifying homozygous minor allele genotypes, which may be very rare, and therefore such approaches may be infeasible. For example, in the exome data from the ORCADES study, less than 3% genotypes are homozygous on minor allele. With the approach that relies

on IBD clusters, we could make use of heterozygous genotypes among sequenced individuals. If a group of genotypes are all IBD in a region, and they all share at least one copy of the minor allele, it is very likely that the shared haplotype contains the rare variant. A similar approach to detecting carriers of Mendelian variants is presented in Chapter 5. Success of this approach to IBD-informed mutations would depend on whether the carriers of rare variants among sequenced individuals indeed share IBD with each other in relevant genomic regions, and on how old a mutation is.

### 3.4.4 Accuracy of short-range imputations

Accuracy of the short-range imputation program IMPUTE2 was analysed on the same data set (Joshi et al., 2013). Accuracy was measured by $r^2$ correlation between sequence genotypes and ones obtained by imputations using data from 1000 genomes project (Consortium, 2012). The accuracy varied heavily based on allele frequencies. The $r^2$ values were as follows:

- Minor Allele Frequency $1\% - 3.2\%$: 0.753
- MAF $3.2\% - 10\%$: 0.867
- MAF $10\% - 32\%$: 0.931
- MAF $> 32\%$: 0.944

The performance is generally very good, however accuracy could be improved for SNPs with low minor allele frequencies. This is because either the model in IMPUTE2 is not capable of capturing long-range haplotypes, or the rare variants can be absent from the reference panel altogether. Imputations based on IBD could help with imputing rare variants, which are possibly newer than others, and can only be distinguished by long-range haplotypes. Because samples were selected for resequencing using our algorithms, the imputation accuracy was likely better that if they had been selected at random.

### 3.4.5 Alternative resequencing strategies

An alternative strategy would be to resequence the entire cohort at low coverage, then exploit IBD sharing to combine data from all individuals who share a haplotype IBD to infer an accurate sequence for this haplotype. For instance, current resequencing methods typically require at least 40 x coverage for accurate sequence imputation. With an average of four haplotype sharers for each gamete, it would be sufficient to type all of the individuals at 10 x coverage.

### 3.4.6 Utility of IBD-informed optimisation of resequencing studies

We showed that samples in the individuals in the ORCADES study share large fragments of haplotypes IBD, and the inferred IBD segments are accurate when evaluated against sequence data. In this section we argue whether the inferred IBD contributes to optimising resequencing studies and imputations.

Gusev et al. showed vast amounts of recent IBD in Kosrea, and many elsewhere unknown variants (Gusev et al., 2012). Also, much recent IBD was identified in data from ORCADES. In IBD segments, the concordance of sequence data is nearly perfect for longer IBD segments, is lower at region boundaries, and varies moderately throughout the genome. Where IBD segments are found, they are accurate, and there is good potential for imputations.

The identified identity by descent is useful for optimising the design of resequencing studies. We have shown it through increased IBD-coverage compared to random selection or one informed by genomic relatedness. Gusev et al. also showed this by improved imputation performance in simulations, using Beagle imputation software. In ORCADES, when resequencing 20% of samples from an original study, we could cover 65% of haplotypes when selecting samples based on their IBD sharing. Had we chosen them randomly, we would cover on average 61%. One could argue our selection algorithm brings only a small improvement, whereas Gusev at al. showed more impressive improvement in imputation quality. In the isolated population of Kosrea, 60% of variant alleles could be imputed with sequencing data from random 1.7% of cohort, or 1.3% if the samples were chosen by the algorithm. In case of the latter study, the good improvement could

be because only seven samples were selected, and it is important that they contain several popular haplotypes. In the case of ORCADES, 20% of the original study corresponds to 180 samples. Even if these are selected randomly, there is a good chance they will cover haplotypes popular in the isolate. Our algorithm is capable of picking individuals who carry moderately popular haplotypes, which however has less contribution to the overall performance score.

Once samples for resequencing are selected and sequence data is available, we focus on accurate imputations. The accuracy of exome imputations using IMPUTE2 in the individuals from ORCADES is generally good, but could be improved for less common and rare variants. It could be improved with IBD-based imputations, thanks to the abundance of IBD identified from array data and its high concordance with sequence data. However, naive IBD-imputation algorithms would fail to deliver the promised improvement in imputation accuracy.

An IBD-based imputation algorithm would need to extract all available information from the sequence data. An algorithm that imputes a rare variant in an individual only when his haplotype sharer is homozygous for a rare allele will fail, since genotypes will be very rarely homozygous for rare alleles. If for example we decide to sequence n optimal 20% of individuals from ORCADES, 65% of haplotypes would be sequenced either directly (20%) or through a sharer (45%). However, if we assume there is exactly one sharer of the haplotype, and we use allele frequencies from exome data, only 3% of the 45% could be imputed. A IBD-based imputation algorithm would need to use not just homozygous genotypes among the sequenced individuals, but also heterozygous ones, as described in Chapter 4.

Before such an IBD-based algorithm is developed, standard imputation strategies, which do not utilise full potential of the data, may be used. In Chapter 4 we demonstrate that IBD-based imputations of rare variants are important for identifying carriers of Mendelian subtypes of diseases.

# Chapter 4

# Identity by descent for identifying Mendelian subtypes of diseases - colorectal cancer

## 4.1 Introduction

Lynch syndrome (LS, hereditary nonpolyposis colorectal cancer) is a Mendelian form of colorectal cancer (CC), caused by loss-of-function variants in DNA mismatch repair (MMR) genes: MLH1, MSH2, MSH6, and PMS2 (Lynch et al., 2009). It is important to detect LS carriers among new CC patients because it has implications for their clinical management. More extensive resection and more intensive follow-up screening is indicated in LS carriers, because of the increased risk of new primary tumours in the unresected colon and in other organs (Vasen et al., 2007). Relatives of Lynch syndrome carriers who share the disease-causing variant also require screening and follow-up to detect cancer at an early stage (Lynch and de la Chapelle, 1999), (Moreira et al., 2012). LS is difficult to diagnose "as there are no specific clinical or histo-pathological features" (Lynch et al., 2009). Current methods for detecting LS carriers rely on clinical information and tumour biomarkers, however they have serious limitations (Barnetson et al., 2006), (Moreira et al., 2012), (Aaltonen et al., 1998).

It is likely that a high proportion of the LS mutations arose only once in history

(Lynch and de la Chapelle, 1999), thus they lie on ancestral haplotypes and the patients that inherited them are distantly related. Thanks to genotype data from modern SNP arrays and algorithms for detecting recent identity-by-descent (IBD) (Kong et al., 2008*b*), carriers of same LS mutations can be identified. Furthermore, we can detect IBD in risk regions between known LS carriers and new colon cancer patients, which is indicative of the latter carrying LS mutations. Where several unrelated colon cancer patients share IBD at MMR genes, they could carry a novel LS mutation.

The novel methodology involves first computing similarity measures of SNP genotypes in regions containing known LS genes through inference of recent IBD. Secondly, a predictive model takes into account possible inaccuracies in the detection of IBD segments.

**Outline** We describe methods for detecting haplotypes co-inherited from recent common ancestors from SNP data, and build a predictive model for LS carrier status among CC patients. We show feasibility of the approach, quoting accuracy of Lynch syndrome prediction, and conclude that the method is promising especially when extensive amounts of genotype data will be available in biobanks.

## 4.2 Materials and Methods

### 4.2.1 Collection of genotypes of patients with Lynch syndrome - MOMA

Genotype data for Lynch syndrome carriers was collected in the MOMA (Modifier of MMR alleles) study, whose aim is to identify the modifiers of mutant alleles of DNA mismatch repair. The main inclusion criterion was that participants must be carriers of pathogenic mutations in one of the DNA mismatch repair genes, and samples in Phase 1 were partly selected for extremes of phenotype and age of onset. Individuals were recruited to the study either because of family history of colon cancer with early onset at the time they or their relatives were diagnosed with the disease, or from prospective studies.

For all of them, genes involved in MMR were sequenced, so the MMR mutations are known independently of array data. The records of mutations the individuals carried follow the standard nomenclature for description of sequence variations (Den Dunnen et al., 2000) . For example, "c.116G>T" denotes that 116th nucleotide in coding sequence of MLH1 was T instead of G.

Among three platforms used in this study, most MLH1 mutation carriers in Phase 1 were genotyped with the Illumina HumanHap660W array, and therefore we focused on this part of the data. The subset of samples we study here consists of 511 individuals (184 from Scotland, 57 from Melbourne, Australia, 136 from Newcastle, Australia, 136 from the Netherlands), 456 of which had mutations in MLH1.

### 4.2.2 Collection of genotypes of patients with colon cancer - SOCCS

The Study of Colorectal Cancer in Scotland (SOCCS) is a case-control study of 3,400 prospectively collected colorectal cancer cases from all Scottish hospitals, and 3400 matched controls. In the first phase of the study 976 early-onset cases and 1,002 matched controls were genotyped with the Illumina HumanHap550 array (Tenesa et al., 2008). The patients had no known family history of colon cancer.

The majority of the samples (more than 80%) underwent the procedure for identifying Lynch syndrome mutations in germline DNA (Barnetson et al., 2006). To detect mutations, 16 exons of MLH1 were analysed with denaturing high-performance liquid chromatography analysis to detect single-base substitutions, insertions and deletions (Wagner et al., 1999). Variants noted there were sequenced, as were MLH1 exons 8, 12, and 15 in every sample. Additionally, in most samples the MMR genes were checked for large deletions using multiplex ligation-dependent probe amplification. When the procedure was positive, samples were removed from SOCCS study, so in theory there should be no remaining LS carriers. For majority of patients also information on tumour biomarkers is available, eg. micro-satellite stability and immunochemistry tests.

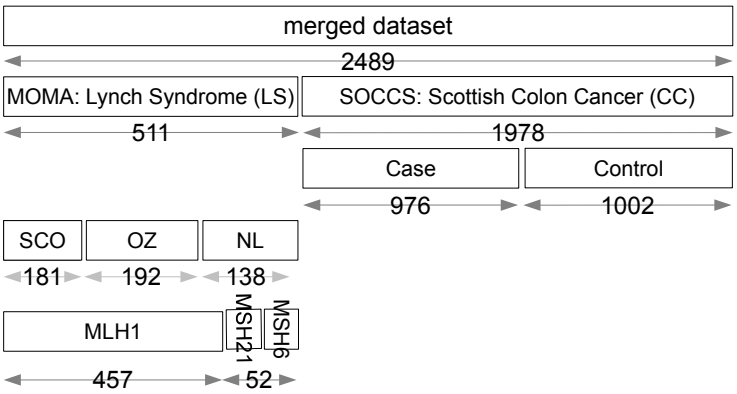



Figure 4.1: A diagram showing the data set merged of MOMA and SOCCS studies.

### 4.2.3 Merging the dataset

Genotypes from the two sources were merged, and markers missing more often than 1/10 were removed from the compiled set. Strand inconsistencies were removed using Plink software. This resulted in a MOMA-SOCCS dataset with 508854 markers and 2489 individuals, summarised in Figure 4.1.

### 4.2.4 Inference of IBD from multi-locus SNP genotypes

In order to infer sharing of LS alleles, identity-by-descent around MMR genes between multi-locus genotypes was detected using software package ANCHAP (Glodzik et al., 2013). In the first stage of ANCHAP's algorithm, large genomic regions without opposing homozygotes are detected, in the second stage shared regions are assigned to one of two gametes, and they are used for phasing of the genotypes. In the last stage, all of the resulting haplotypes are compared to find regions identical by descent with greater accuracy. In stage I, we searched for matching genotype sequences longer than 2 cM. Alignment in stage II was with standard parameters. In stage II, the program identified identical haplotypes longer than 1 cM and containing 200 phased alleles. This is a shorter genetic region than standard settings (2 cM), but since data available had more markers than in the article describing the method, we utilised it and set a threshold for number consecutive SNPs with phased alleles in a segment to 200 (originally 100).

ANCHAP was our preferred algorithm over GERMLINE (Gusev et al., 2009) because the genetic data in the study is unphased, and over fastIBD (Browning and Browning, 2010), whose performance we evaluated and found it matches ANCHAP only when its parameters are tuned. We applied fastIBD with various values of the scale parameter which controls complexity of haplotype model, and show the results for the smallest and larger recommended values of the parameter and for the best performing one.

We visualised the IBD covering whole of each MMR genes by graphs, where nodes represent individuals and links signify sharing IBD between them (Gansner and North, 2000).

### 4.2.5 Predictive model for carrying LS mutations

The predictive model allows to compute probabilities of carrying LS mutations, by considering IBD shared by a proband with all known LS carriers. In this model, the probability of sharing a LS mutation, given the length of a segment IBD, is parametrised. The longer the IBD segment, the more recent the common ancestor was, and the more likely that the DNA in the shared region is identical. We define $\phi_{i,j}$ as probability that samples $i$ and $j$ are IBD in a genetic risk region. $\text{IBD}_{i,j}$ is length of identical haplotypes [cM] spanning the region declared by ANCHAP. $a$, $b$ and $c$ are real-valued parameters to be learned from the data.

$$\phi_{i,j} = c \times \text{logistic}(a + b \times \text{IBD}_{i,j}) = \frac{c}{1 + e^{-(a+b\times\text{IBD}_{i,j})}}$$

We build a predictive model on the intuition that a proband carries a LS mutation if he shares the disease haplotype in MLH1 gene with at least one LS carrier, and does not share the disease haplotype with controls. Some haplotypes are particularly common, and may be shared by individuals with and without the LS mutation present. To eliminate their impact, we use IBD between a new patient and SOCCS controls. If the patient shares IBD with several LS carriers, as well as SOCCS controls, this casts doubt as to whether the patient also carries LS mutations. Let $\text{L}_i = 1$ denote that proband $i$ carries a LS mutation, C a set of LS carriers in training set, and O set of SOCCS controls in training set.

$$Pr(\mathrm{L}_i = 1) \propto \left(1 - \prod_{c \in \mathrm{C}} (1 - \phi_{i,c})\right) \left(\prod_{o \in \mathrm{O}} (1 - \phi_{i,o})\right) \triangleq \alpha \quad (4.1)$$

$$Pr(\mathrm{L}_i = 0) \propto \prod c \in \mathrm{C}\,(1 - \phi_{i,c}) \triangleq \beta \quad (4.2)$$

$$Pr(\mathrm{L}_i = 1) = \frac{\alpha}{\alpha + \beta}$$

### 4.2.6 Experimental design

In order to reduce the dependency of the model on close relatives, and ensure it generalises on new patients, who are likely to be unrelated to known LS carriers, when predicting whether a proband carriers a LS mutation, we ignored his close relatives. In products in Equation 4.1, for each predicted patient, we consulted data on only such patients in the training set whose genetic relatedness with the predicted patient is less than a threshold. In experiments we evaluated effects of different thresholds for genetic relatedness. Furthermore, we decided to exclude non-Scottish carriers from the data used for model learning, as otherwise the model might learn to predict Dutch or Australian ancestry, rather than LS carrier status. We trained and evaluated the predictive model on the set of 181 Scottish MLH1 LS carriers and 976 SOCCS controls who we assume do not carry LS mutations.

### 4.2.7 Computation of genetic relatedness matrix

Genetic relatedness was computed as a matrix of dot products of normalised genotype vectors (Yang et al., 2010). Accordingly, average relationship between pairs is 0 and average relationship of an individual with himself is 1. In Equation 4.3 $A_{j,k}$ denotes genetic relatedness of individuals $j$ and $k$, $N$ is the number of SNPs in each genotype vector, and $\mathbf{x}_j$ and $\mathbf{x}_k$ are the genotype vectors for individuals $j$ and $k$, where allele dosages had been normalised with respect to allele frequencies for each SNP.

$$A_{j,k} = \frac{1}{N}\mathbf{x}_j \cdot \mathbf{x}_k \tag{4.3}$$

### 4.2.8 Cross-validation

To train and evaluate the model we used a cross-validation procedure. We run cross-validation with 10 folds, in each choosing different 1/10 of data to be the test set, and the rest to be training set. Based on this, we make predictions on test data in each fold, and finally evaluate the overall procedure. This rigorous cross-validation procedure should ensure that the predictions made on unseen data, SOCCS cases, are of similar quality.

### 4.2.9 Learning

We selected model parameters $(a, b, c)$ to maximise likelihood on training data, in each fold. We optimised the likelihood using a constrained active set optimisation algorithm as implemented in Matlab, with starting parameters set randomly in the region of high-likelihood ($-2 > a > -50$, $2 < b < 50$, $0.01 < c < 0.99$), chosen from a prior visual inspection of the likelihood surface.

### 4.2.10 Predictions on colon cancer patients and verification

Finally, we make predictions of LS status for new CC patients, whose LS status is unknown. We do this using the predictive model optimised using all training data. By reference to the haplotype sharers of a proband, the algorithm also highlights a likely mutation and its location in the MLH1 gene.

### 4.2.11 Verification of the suspected patients by Sanger sequencing

For top indicated patients, the predictions were verified by targeted dideoxy Sanger sequencing in indicated exons. As positive controls we used samples from the MOMA study which had been confirmed to carry the same mutations as the suspected patients. Additionally, we looked up the mutations identified earlier in the original sequencing of the MLH1 gene in SOCCS study. This work was carried out by Susan Farrington.

## 4.3 Results

### 4.3.1 Lynch syndrome carriers share IBD around the MLH1 gene

Figure 4.2 illustrates identity-by-descent of long haplotypes between individuals with MLH1-related Lynch syndrome. Throughout MLH1 LS carriers share IBD around 9 times more often than SOCCS controls, and 4.5 times more often than the same LS carriers elsewhere on chromosome 3. Because on chromosome 3 there are no other regions of increased IBD sharing like at MLH1, frequent IBD sharing at MLH1 between LS carriers cannot be explained by relatedness alone, but rather by sharing the disease haplotype.

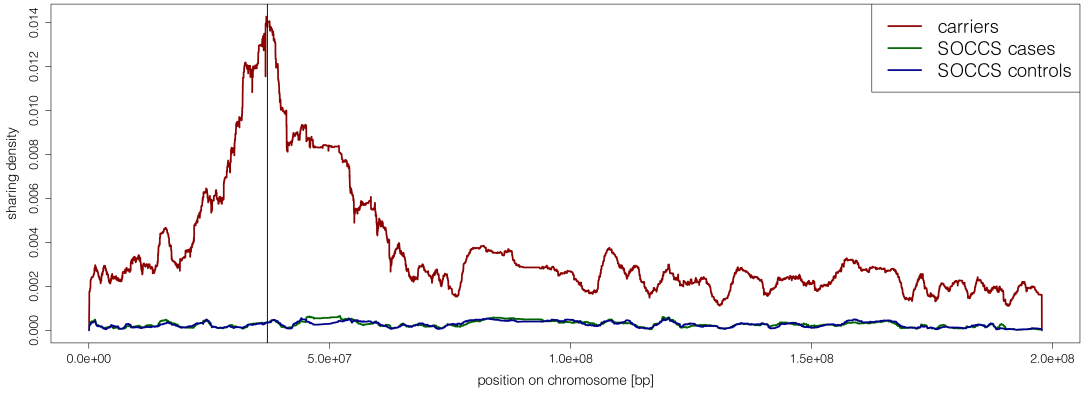



Figure 4.2: Frequency of IBD sharing on chromosome 3, with MLH1 marked by the vertical line. The horizontal axis shows genetic position on chromosome 3, and the vertical axis shows sharing density, or the probability that a pair of individuals share IBD at a locus. Only sequences longer than 2 cM are plotted. LS carriers share IBD around the MLH gene more often than elsewhere on the chromosome. There is almost no sharing between controls.

The IBD relationships in the data at the MLH1 gene were visualised in Figure 4.3.

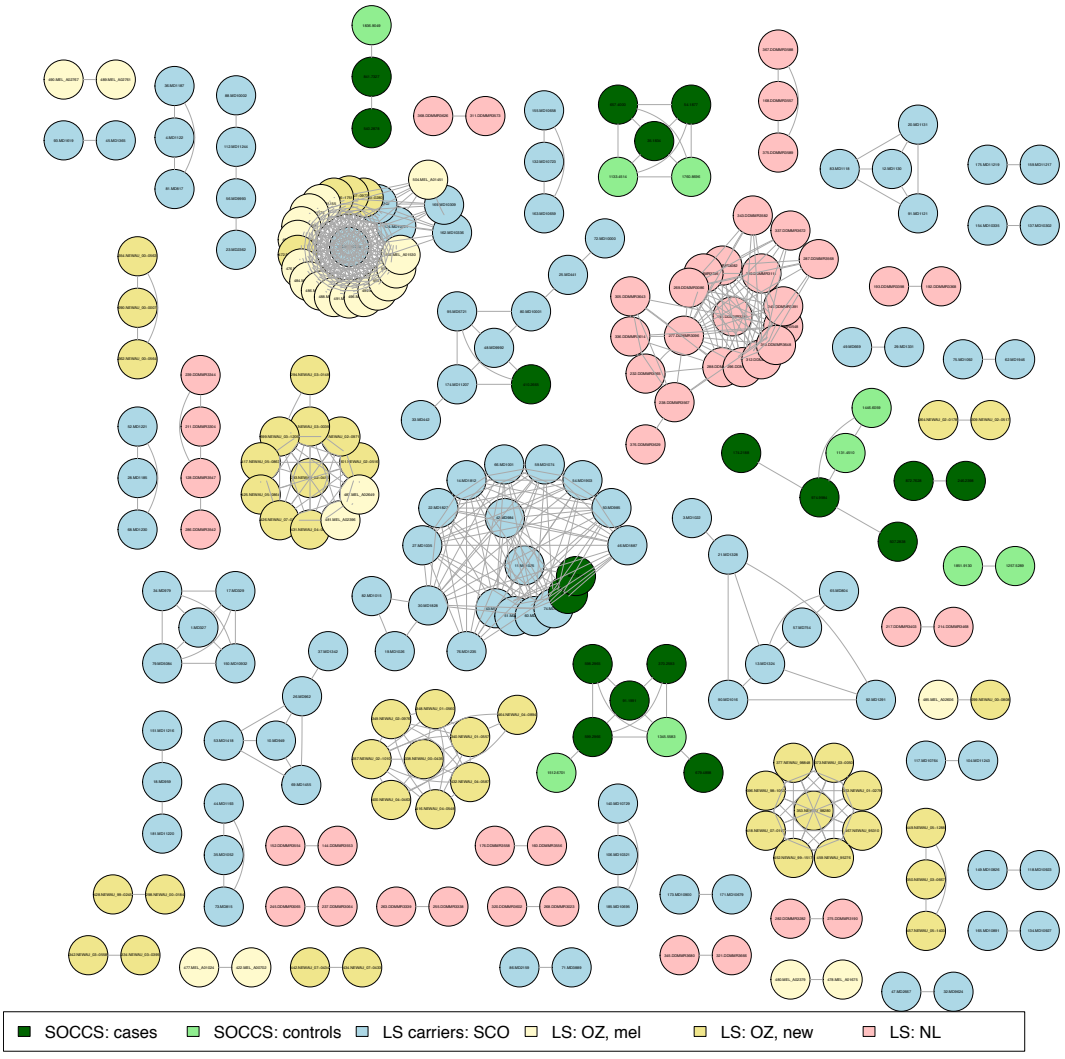



Figure 4.3: IBD graph for the MLH1 region, where edges denote IBD relationship longer than 3 cM. Each node in this figure represents a patient or Lynch syndrome carrier, and shown are only individuals who share a haplotype with at least one other sample. Green nodes represent Scottish colon cancer patients and controls. Other colours represent Lynch syndrome carriers from different geographic locations. Where two nodes are connected, this indicates that the two patients share region including MLH1 gene IBD. Clusters of patients (identified with labels) indicate groups that carry same LS mutations. Where a Scottish colon cancer patient falls into a cluster of LS carriers, we infer the patient could be an unsuspected LS carrier.

### 4.3.2 Evaluation of recent IBD against mutation information

Further validation is provided by resequencing data available for the LS carriers in the MOMA study, through which mutations in MLH1 had been identified. We defined sensitivity as a ratio of pairs who were found to share IBD at MLH1 to number pairs of individuals who share the same LS mutation:

$$\frac{\text{number of pairs that share the same LS mutation and share IBD at MLH1}}{\text{number of pairs that share the same LS mutation of MLH1}}$$

False discovery rate is the ratio of number of pairs of individuals that share IBD at MLH1 despite not having the same LS mutations to the number of pairs of individuals that were found to share IBD at MLH1:

$$\frac{\text{number of pairs that share IBD at MLH1 and have different LS mutations}}{\text{number of pairs that share IBD at MLH1}}$$

IBD sharing between a pair that carry two different LS mutations could arise if the two individuals share the non-disease haplotype IBD.

We evaluated whether pairs identified to share IBD around MLH1 also carry the same LS mutations, and present the result in Figure 4.4. For sequences longer than 3 cM, sensitivity and false discovery rate are 0.51 and 0.1, for sequences longer than 2 cM they are 0.60 and 0.13 respectively. Below this threshold more errors occur, such that for all sequences longer than 1 cM the sensitivity reaches 0.65 and false discovery rate 0.28. Using the same reference data we compared the performance of ANCHAP against fastIBD, another algorithm for detecting recent IBD, and found that it matches the performance of ANCHAP only with most optimal non-standard settings. Overall, there is agreement between the IBD segments and the mutation annotation, but also considerable amount of uninformative IBD sharing that the predictive model needs to deal with.

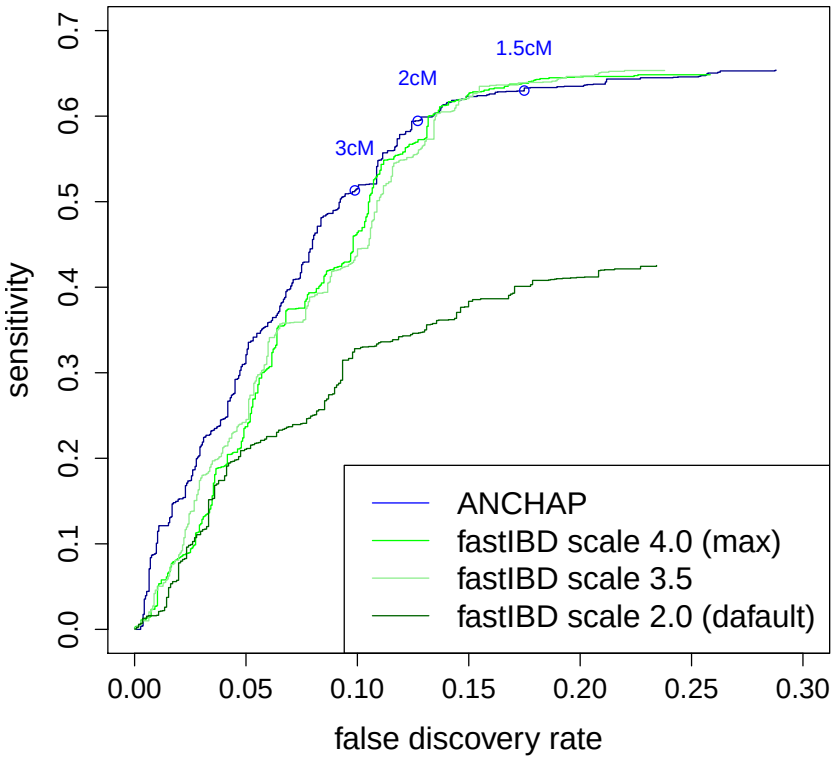



Figure 4.4: Evaluation of IBD detection between carriers of LS mutations in MLH1 against mutation information obtained from sequencing of the gene. We checked whether pairs that share IBD at MLH1 also carry the same mutations. With decreasing length of IBD segments, concordance with mutation information drops. In the comparison of methods ANCHAP performed as well as fastIBD with the optimal scale settings.

### 4.3.3 Relatedness and length of IBD segments

New patients diagnosed with colon cancer are likely to be related with known LS carriers only distantly. To be able to use IBD with LS carriers as a diagnostic, we should be able to find also unrelated pairs who share IBD in the risk region. Figure 4.5 shows this is the case. Even though there are many closely related pairs of Lynch syndrome carriers, there are also nearly unrelated patients who share large segments (8 cM) that spans MLH1. IBD can be observed among pairs where one individual is a LS carrier and second is a SOCCS case, and among LS-SOCCS control pairs. Longest segments (>3 cM) are between LS carriers and SOCCS cases, however also present are shorter segments (<2.5 cM) of haplotype identity between LS carriers and SOCCS controls. In summary, at MLH1 there are long haplotypes shared IBD between nominally unrelated individuals, as required by out method for detecting carriers of LS mutations.

### 4.3.4 Quality of predictions of Lynch syndrome

In order to detect unsuspected LS carriers among new colon cancer patients, we learned parameters for a predictive model. Predictions that are made by the model on test data are illustrated in Figure 4.6. The model identifies the majority of LS carriers in the MOMA set. Performance of the predictive model is summarised in Figure 4.7. When all, also related individuals from the training data, are used for making predictions on test data individuals, the area under the ROC curve is 0.91, and it drops to 0.83 when only individuals whose genetic relatedness with a tested patient is less than 0.05 are used. Predictions are very accurate, however the performance does depend on whether close relatives are used in the analysis.

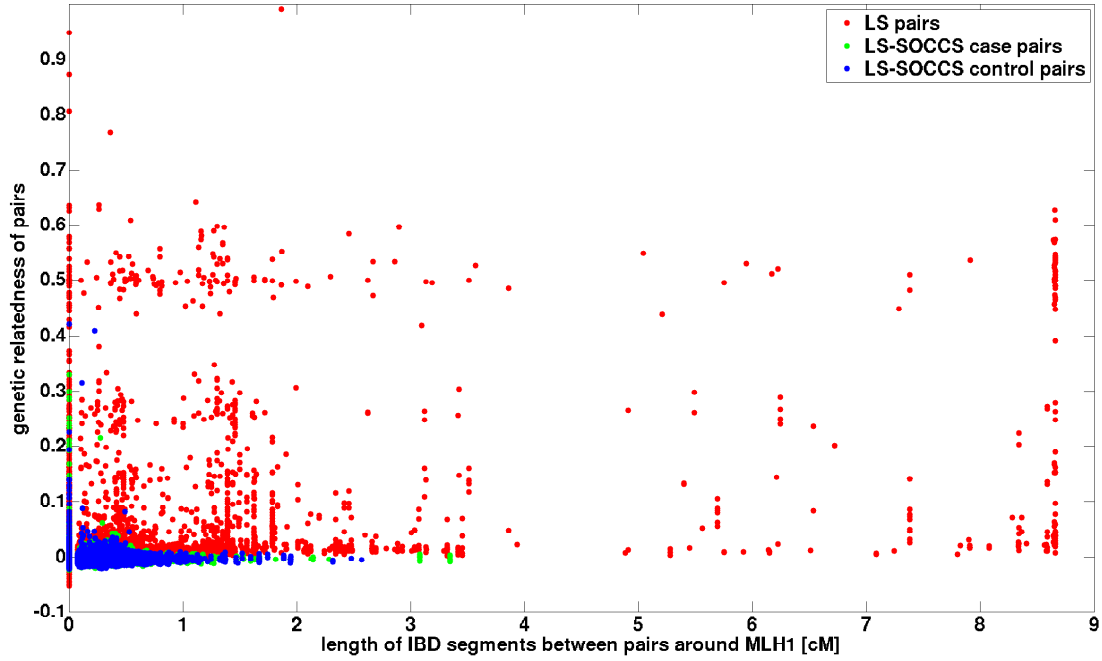



Figure 4.5: Genomic relatedness and the length of segments shared IBD around MLH1. Some pairs of LS carriers (LS pairs) share long regions IBD at MLH1, even though their genome-wide relatedness is low ($< 0.05$). Accordingly, we can expect to find long IBD regions between unrelated pairs of known LS carriers and new colon cancer patients.

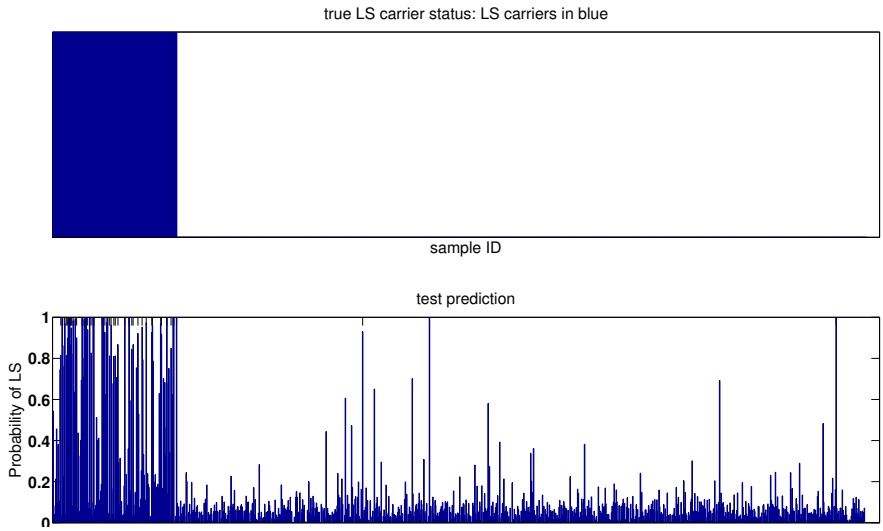



Figure 4.6: Predictions on test data using top model.
Top figure: true LS carrier status for samples, where blue denotes that mutation present.
Bottom figure: predictions for LS carrier status made by the model
The model detects many of true LS carriers, and assigns low probabilities of LS to SOCCS controls, who have not developed colon cancer. The good quality of predictions is summarised by the ROC curve, area under which is 0.85.

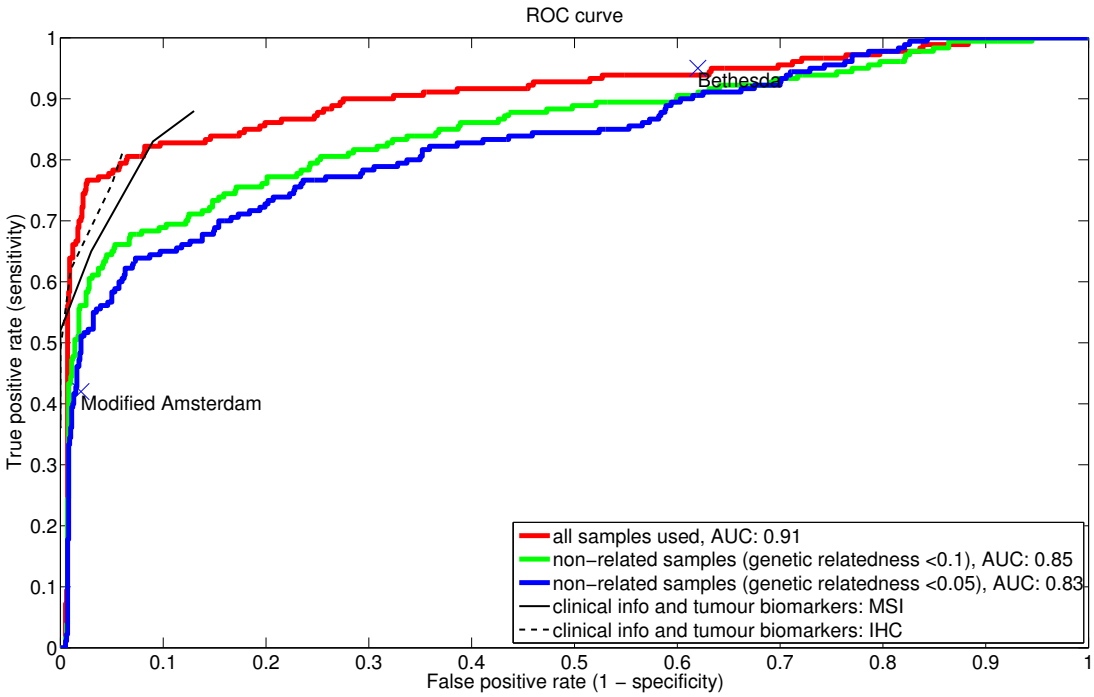



Figure 4.7: ROC curves for the predictive model, on test data. Also plotted is performance of Bethesda and Modified Amsterdam criteria, which need clinical and family information and are currently used in clinics. With all samples used, the performance of our model is similar to one that uses clinical data and tumour information.

### 4.3.5 Comparison against currently used diagnostics

For comparison, Figure 4.7 shows performance of other available diagnostics. At the extremes of our ROC curves we see the Modified Amsterdam and Bethesda criteria, which are based on the disease history in the family and clinical information like presence of microsatellite instability in the tumour. With all samples used, the performance of our model is similar to performance of predictive model that uses clinical information and tumour biomarkers.

### 4.3.6 LS predictions for colon cancer patients

Finally, we used the predictive model to rank colon cancer patients in SOCCS study as unsuspected cases of Lynch syndrome. Table 4.1 shows a list of individ-

uals to whom the model assigned probabilities for carrying LS mutations more than 0.8. Indeed, the three individuals there, ones identified as: "7335", "1863", "2665", could also be identified visually from Figure 4.3, because they share IBD around MLH1 with Scottish samples who share IBD also between each other. The remaining five candidate individuals that were predicted to carry LS mutations shared shorter haplotype segments, and therefore they were not shown in the Figure.

| patient ID | Number of IBD sharers, LS carriers segment length | | | | | LS mutations of IBD sharers | Number of IBD sharers, controls segment length | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 3 cM | 2cM | 1.5cM | 1cM | 0.5cM | | 3cM | 2cM | 1.5 cM | 1cM | 0.5cM |
| 7335 | 6 | 6 | 6 | 15 | 16 | c.116G> $T$ | 0 | 0 | 0 | 1 | 3 |
| 1863 | 6 | 6 | 6 | 6 | 18 | c.116G>T | 0 | 0 | 0 | 0 | 3 |
| 1873 | 0 | 0 | 0 | 0 | 20 | c.116G>T | 0 | 0 | 0 | 1 | 3 |
| 5021 | 0 | 0 | 0 | 0 | 16 | c.116G>T | 0 | 0 | 0 | 1 | 3 |
| 2665 | 3 | 3 | 4 | 5 | 10 | c.1190_1191delT | 0 | 0 | 0 | 1 | 8 |
| 7008 | 0 | 0 | 3 | 3 | 4 | H264R | 0 | 0 | 0 | 0 | 2 |
| 1808 | 0 | 0 | 0 | 2 | 4 | EX1del | 0 | 0 | 0 | 0 | 0 |
| 2631 | 0 | 0 | 2 | 2 | 3 | c.116 +1G>A | 0 | 0 | 0 | 0 | 0 |

Table 4.1: Top colon cancer patients suspected of carrying the LS mutations, their IBD sharing with known LS carriers and controls, and the mutations in MLH1 they could carry.

### 4.3.7 Search for novel Lynch syndrome mutations

Figure 4.8 shows a cluster of SOCCS patients that all share IBD with each other, without sharing with any known LS carriers. We suspect that these individuals carry novel LS mutations. On chromosome 3, the individuals share IBD only around MLH1. The cluster also includes on SOCCS control who may also be carrying the mutation, which is not fully penetrant.

A cluster of SOCCS cases sharing haplotypes IBD longer than 3 cM with each other around MLH1.

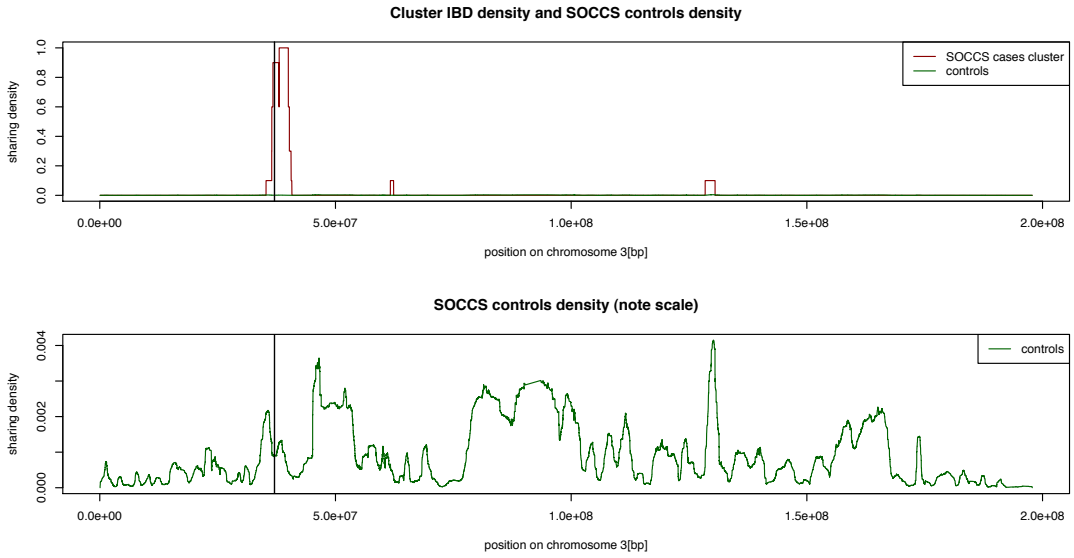



Figure 4.8: A demonstration of detecting possible carriers of unknown mutations in MMR genes among colon cancer patients. On chromosome 3, four unrelated SOCCS patients and one control share IBD only around the MLH1 gene. Almost all pairs of cluster members share IBD between each other around MLH1, which implies sharing a common haplotype. IBD sharing of long haplotypes between unrelated samples is unlikely, and indeed this does not happen outside the vicinity of MLH1, as shown by the middle plot. Bottom plot shows that background rates of sharing such long haplotypes IBD among the controls are negligible. The five samples are therefore suspected of carrying a novel LS mutation.

### 4.3.8 Verification of Lynch syndrome carrier status of suspected patients through targeted sequencing

None of the suspected eight patients was found to carry the predicted mutations in MLH1. Some other variants in MLH1 were observed, for example IVS 14-19 SA a>g, IVS 13+14 SD G>A or I219V, as shown in Table 4.2. IVS 14-19 SA A>G stands for A to G mutation of 19th last nucleotide of intron 14 of MLH1 (Den Dunnen et al., 2000), IVS 13+14 SD G>A stands for G to A mutation of 13th nucleotide of intron 13, and I219V denotes change of isoleucine to valine at 219th amino-acid of the MLH1 protein. The SOCCS patients who all share IBD with each other carry variants are also listed in Table 4.2.

| patient ID | LS mutation suspected | mutation confirmed | other variants in MLH1 |
|---|---|---|---|
| Patients suspected because of IBD with known LS carriers | | | |
| 7335 | c.116G> *T* Scotland | No | - |
| 1863 | c.116G>T Scotland | No | IVS 14-19 SA a>g |
| 1873 | c.116G>T Scotland | No | - |
| 5021 | c.116G>T Scotland | No | I219V, IVS 14-19 SA a>g |
| 2665 | c.1190_1191delT Scotland | No | I219V, IVS 14-19 SA a>g |
| 7008 | H264R Oxford | Not screened | |
| 1808 | EX1del Scotland | No | IVS 13+14 SD g>a |
| 2631 | c.116 +1G>A Scotland | No | - |
| Patients suspected because their share IBD with each other | | | |
| 2965 | unknown | | **I219V**, IVS 14-19 SA a>g and IVS 9+10 SD a>g |
| 2593 | unknown | | **I219V**, IVS 14-19 SA a>g |
| 1991 | unknown | | **I219V** |
| 2966 | unknown | | **I219V**, IVS 14-19 SA a>g |

Table 4.2: Resequencing of MLH1 from top suspected SOCCS patients was not confirmed in any case. The patients in the IBD cluster at MLH1 all carry the I219V mutation.

## 4.4 Discussion

### 4.4.1 Spectrum and inheritance of Lynch syndrome mutations

For new LS carrier detected with our method, the suspected patient has to carry the same mutation as at least one known LS carrier, and share a long haplotype with him. Carriers of private mutations cannot be detected in this way.

It is probable that most LS mutations are not private variants but are shared, based on frequencies of mutations in MOMA study. Firstly, the samples in MOMA studies come from only 3 countries, where the pool of mutations may be limited. Secondly, as discussed in the introduction, LS mutations are unlikely to affect reproductive fitness and thus are not removed by selection. To check these assumptions, we analysed the frequencies of LS mutations in the MOMA study, as shown in Figure 4.9. Only 20-24% of mutations occur only once, and some mutations are particularly common.

Many participants of the MOMA study had been recruited because they had affected family members, and therefore we may have a biased view of the proportion of mutations that are shared. In building the predictive models we took steps to reduce the effects of such a design of the study, by ignoring close family members based on genetic relatedness. The reported accuracy of the model may still be unrepresentative of how it would perform in general, as the proportion of shared to private mutations may not be general for the whole population. This problem would be ameliorated if a larger proportion of LS carriers were sequenced.

### 4.4.2 Assumption of high penetrance of Lynch syndrome variants

Our validation study assumes that the control individuals in SOCCS study do not carry any LS mutations, however in reality mutations of low penetrance could occur even among individuals who did not develop colon cancer. In such case LS carriers could occur even among SOCCS controls, and our experimental design would be inappropriate. We hoped to sequence the three SOCCS control individu-

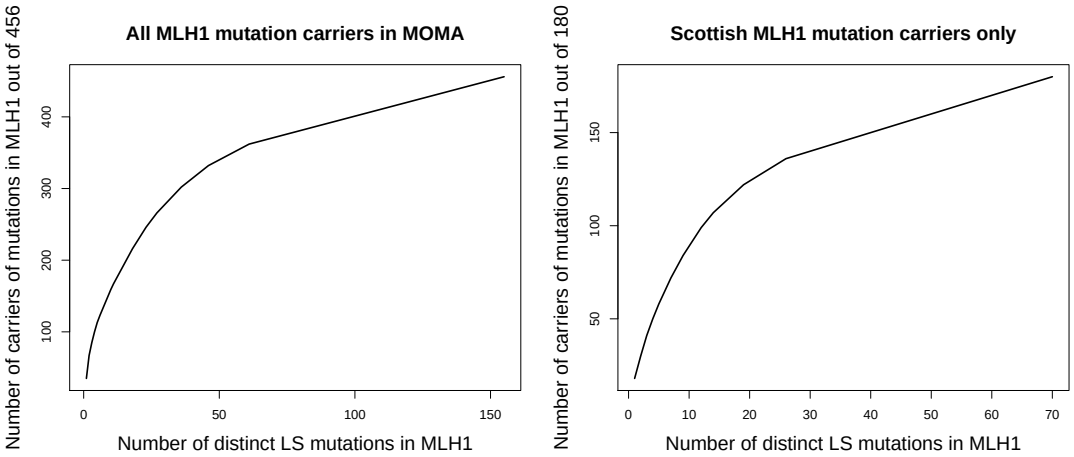



Figure 4.9: Ordered frequencies of Lynch syndrome mutations in MLH1. Left: whole of MOMA study, right: Scottish carriers only. In both cases there exist some very common mutations. In the whole of MOMA study 20% of carriers carry singleton mutations, and among the Scottish carriers 24%. The remaining mutations are shared between two or more individuals, either due to their high frequency or due to design of the MOMA study.

als whom the model indicated that they could carry LS mutations. Unfortunately the sequencing of these samples could not be done due to time constraints of our collaborators.

### 4.4.3 Accuracy of IBD detection

Accuracy of detected recent identity-by-descent is crucial for predictions of LS status. For this application, the identical regions should be co-inherited from a common ancestor more recent than ancestor in whom the LS mutation first arose. In Chapter 2 as important for accurate detection of recent IBD shown were accuracy of genetic map and accounting increased linkage-disequilibrium in some parts of the genome. It is possible that some of the segments we identify as IBD in reality were inherited from their common ancestor that was earlier than the ancestor in whom the disease mutation first arose. It has been reported that dating common ancestors based on length of IBD segments may be imprecise (Ralph and Coop, 2012). Because unexpectedly we detect some IBD segments even between LS carriers and SOCCS controls, we may be detecting ancient

haplotypes that are unusually long. Furthermore, some of the inaccuracies may result from the border position of the MLH1 gene in many IBD segments. Most haplotypes shared IBD that span the MLH1 end shortly after the gene. This could be either because of a recombination hotspot downstream of the MLH1, or increased linkage disequilibrium upstream of the gene. We showed in Chapter 3 that towards the ends of detected IBD segments, genotypes of sequences match less often. For pairs where the recent IBD segments cover MLH1 gene only marginally, LS mutations might not have been inherited. On the other hand, our predictive model appears to cope with occasionally imprecise inference of IBD, given its high predictive accuracy on test data.

### 4.4.4 Sequencing the suspected patients in search for variants in MLH1

The usefulness of work presented here depends on whether the predictive accuracy generalises to new colon cancer patients. We resequenced the top colon patients suspected of carrying LS mutations, as highlighted by our model. Sequencing of the patients suggested by the algorithm, on the contrary, did not identify any of Lynch syndrome carriers. Even though the model was very accurate in predicting LS status on test data, it failed on new colon cancer patients. Possible reasons for this include:

1. incorrect detection of IBD segments,
2. that the detected IBD segments between are so ancient, that the common ancestor precedes the time when a given Lynch syndrome mutation arose,
3. sample mishandling: different samples used for SNP genotyping and resequencing,
4. high-performance liquid chromatography not revealing all mutations.

Incorrect detection of IBD segments is unlikely for long shared haplotypes. IBD segments for the top suspected patient with identifier 7335 are shown in Figure 4.10. There are hardly any missing genotypes in the genotypes of known

LS carriers and the suspected ones. I verified that the haplotypes recovered by ANCHAP for the LS carriers and suspected patients are indeed identical at the typed SNPs. There are no unusual allele frequencies of SNPs around MLH1. Additionally, another program fastIBD returned IBD segments very similar to ones given by ANCHAP.

The most likely source of error is in the assumptions we made. We showed earlier that the method correctly identifies carriers of the same mutations for known LS carriers, as in Figure 4.4. However, it does not imply that the method would work equally well between the LS carriers and new CC patients. IBD segments are generally shorter for the later pairs. Between the known LS carriers and new CC patients, we may be detecting sharing IBD of the ancient, shorter haplotypes that pre-date the LS mutations. This point can be verified from the data: in Figure 4.11 - pairs of known carriers of the same LS mutation mostly share longer IBD segments with each other, than with the top 4 SOCCS patients suspected of carrying the same mutation. Additionally, between known LS carriers false positive detection of IBD is less probable, as there are fewer pairs for detecting IBD sharing.

Another reason why we find shared haplotypes between known LS carriers and new CC patients, but not the same mutations, is that the haplotypes of MLH1 themselves are more prone to mutations, and so increase the risk of colon cancer. This would explain the lack of indicated mutations among the new CC patients. However, we have not found any such haplotypes among the known LS carriers, as they mostly share the mutations when they share IBD around the MLH1 gene.

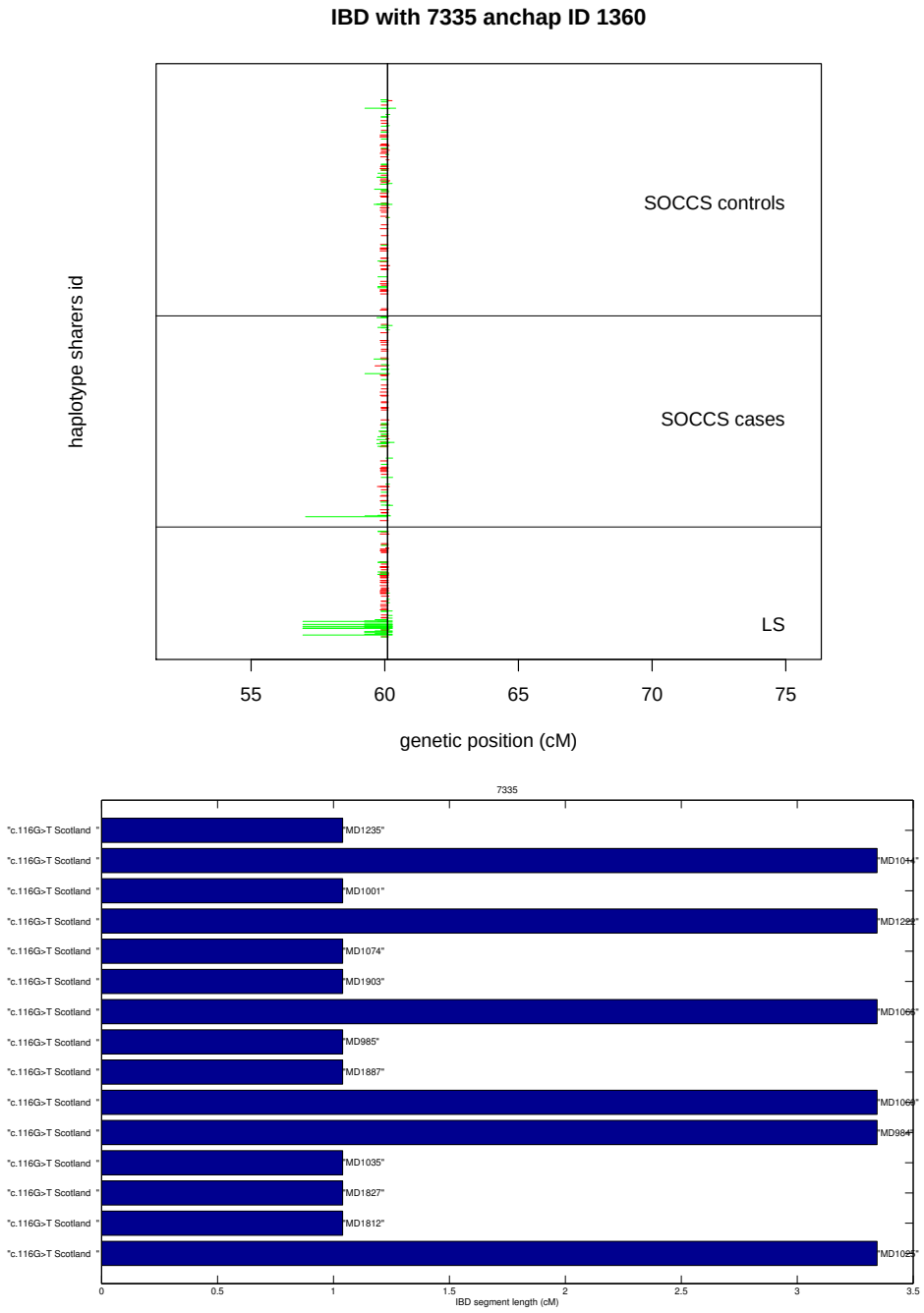



Figure 4.10: Illustration of IBD sharing between top suspected patient 7335 from the SOCCS study, and samples with diagnosed Lynch syndrome. Top figure: position of IBD segments (X-axis, MLH1 marked with a vertical line), shared with other samples in the study (Y-axis), ordered from bottom: known LS carriers, SOCCS cases, SOCCS controls. Bottom figure: length of IBD segments with known LS carriers, and the mutations in MLH1 they carry. As a result, the sample 7335 was predicted to carry the same mutation c.116 G> T with high confidence.

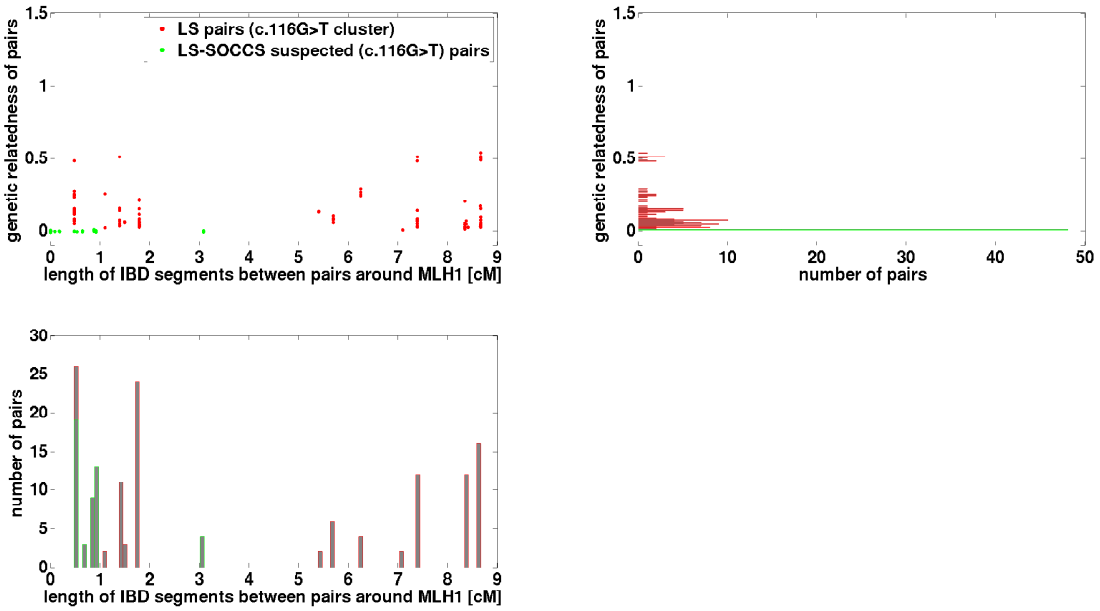



Figure 4.11: Lengths of IBD segments between the pairs of patients with Lynch syndrome with mutation c.116 G>T in MLH1, and between them and four SOCCS patients suspected of carrying the same LS mutation. Known LS pairs share mostly much longer segments IBD, and are much closely related between each other.

The sample mix-up might have occurred due to the sheer number of them handled in the SOCCS study. Identity of sequenced and genotyped samples could be evaluated if larger portions of the MLH1 gene exons had been fully sequenced.

### 4.4.5 Predictive model learning

The fact that MOMA samples were recruited from families could not only distort the results for predictive accuracy of our model, but also affect learning. The model could learn to depend only on very long stretches of IBD, which in reality would not be encountered between unrelated LS carriers and a new colon cancer patient. We attempted to reduce the dependencies on close relatives by restricting the data set to unrelated individuals.

The IBD segments between known LS pairs are mostly much longer than with unsuspected LS mutation carriers. From the $\phi$ function learnt, as shown

in Figure 4.12, we see that the model relies on IBD segments 1.5 cM long as much as on segments 5 cM long. For the data available, it seemed valid, as the predictive accuracy of the model on test data was very good. On data where real LS carriers are not from MOMA study, the performance of the model may not generalise, and longer IBD segments might be required for accurate predictions. This could be due to different sample sizes or to the partially family-based design of MOMA. To ensure a better match between test and unseen data, test data should be composed of LS carriers identified in a prospective study.

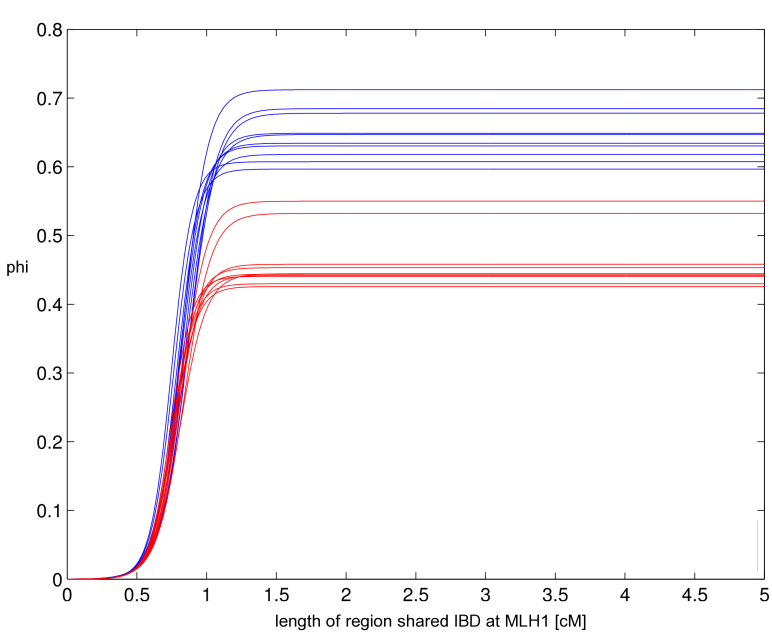



Figure 4.12: Phi functions learnt across folds of the model training. Horizontal axis: length of IBD segments. Vertical axis: value of $\phi$. From the $\phi$ function learnt, we see that the model relies on IBD segments 1.5 cM long as much as on segments longer than 5 cM.

### 4.4.6 Suggestions for repeating the experiments

Most of the SOCCS patients had been screened for Lynch syndrome using high-performance liquid chromatography, sequencing of three exons for everyone and other exons if the initial screening revealed variants (Barnetson et al., 2006). Measured were also biomarkers for most of the sample tumours: micro-satellite

instability and immuno-histochemistry. We aimed to detect Lynch syndrome carriers missed in this procedure, but very few cases would have been missed. Out of the 976 cases, we would expect around 30 LS carriers (Lynch et al., 2009), and among these we expect 11 to carry de-activating mutations in MLH1 (Moreira et al., 2012). In SOCCS, previously reported had been 14 (Barnetson et al., 2006), they had been moved from SOCCS to MOMA, so the chance that any undiagnosed carriers remain is low.

In order to properly evaluate accuracy of the predictive algorithm, we could try to reconstruct the original SOCCS dataset with the 14 prospective LS carriers in it. The 14 individuals who turned out to be LS carriers could be then taken as reference LS carriers. However, the 14 individuals are a too small number for a formal evaluation. I looked into the data on the 14 individuals that were moved from SOCCS to MOMA after LS was detected. Unfortunately, using our method only 2 carriers could be detected because they share IBD with other known LS carriers. A further 4 could be detected, but they have very close relatives in MOMA as well, probably due to design of the MOMA study. The rest of LS carriers either have private mutations, or they share IBD regions that are too short to be picked up by the model from noise.

In conclusion, in order to show the power of the algorithms presented here, one would require a large study covering a larger sampling fraction of cases. With higher sampling fraction, many of the singleton mutations in MMR genes might turn out to be shared.

### 4.4.7 Detecting novel Mendelian subtypes

Sharing long haplotypes IBD in disease risk regions between unrelated individuals could be indicative of unknown mutations causing Mendelian variants of the disease. When a number of patients all share recent IBD with each other, this is unlikely to happen by chance. We accessed the initial sequencing information on the exons of the MLH1 gene from patients in the cluster. We found that the only variant that they share affects the MLH1 protein by substituting isoleucine to valine at amino-acid 219 (I219V), or c.655 A>G in nucleotide notation. This variant is common enough that it was assigned a number rs1799977, and the

frequency of the G allele is 0.4 in European population. There is no conclusive evidence on impact of this mutation on the DNA mismatch-repair mechanism.

With this data set, we could not demonstrate that the approach is able to detect unsuspected LS carriers. This approach could still be examined if a higher proportion of colon cancer patients are genotyped, giving a chance that the unsuspected carriers share larger IBD segments with each other.

### 4.4.8 Possible improvements to the predictive model

The algorithm we presented is a prototype designed for simplicity, which could be further improved. Other types of $\phi$ functions, which we only require are monotonically increasing, could improve predictive accuracy. Further improvement could be brought by explicit modelling of clusters of identity by descent, through graph theoretic algorithms, as for example DASH (Gusev et al., 2011).

### 4.4.9 Prospects for using the method in future

The chance of identifying novel Mendelian subtypes grows with number of samples genotyped. With increasing proportions of population in biobanks for which genotype data is available, more recent IBD will be detected, improving prediction accuracy. The number of colon cancer patients in our study is only about 3% of patients diagnosed in Scotland per 10 years (Scotland, 2013). When a higher fraction of colon cancer patients appears in biobanks, the approach we demonstrate here is likely to perform better for detecting Mendelian subtypes of diseases such as colon cancer.

# Chapter 5

# Applications to Genomic Predictions and Discussion

The original motivation for the work described in this thesis was that identical-by-descent haplotypes capture rare variants. In order to establish utility of IBD segments inferred from SNP data, we have focused on their use in optimising resequencing studies, in studying Mendelian subtypes of diseases, and in genomic predictions. For these applications, utility of the detected IBD sharing depends on:

- how accurate and complete detection of IBD regions is,
- whether time to common ancestor for a shared haplotype can be estimated,
- how likely are mutations that accumulate on haplotypes since the common ancestors,
- whether long-range phasing is more accurate than traditional short-range methods.

We first present preliminary work on genomic predictions that use regions of recent IBD, since the rare variants that they capture were thought to be the reason behind missing heritability of many complex traits. In this section we also discuss the factors that decide on utility of the IBD segments for all of the mentioned applications. Finally, we predict the future of IBD analysis when genetic data for a large proportion of the population is available in biobanks.

## 5.1 Genomic predictions utilising recent identity by descent

Missing heritability of quantitative traits could be explained with rare variants of large effect. Alternatively, common SNPs could explain a large proportion of genetic variants of traits if the traits are very polygenic and the effects sizes are too small to reach significance levels in genome-wide association studies (Yang et al., 2010).

We present preliminary results of our work on genomic predictions here because more work should be done to complete the experiments. The aim of the work presented here is to compare predictive models of quantitative traits based on the two hypotheses, in particular to test whether models that take into account rare variants predict traits better. One predictive models uses common SNPs only, and another one uses ancestral haplotypes shared IBD, together with rare variants that they carry. Building models for genomic predictions is important because, among others, it would allow for predicting risk for diseases and their early prevention.

By comparing the two predictive models, one based on common SNPs only and one based on haplotypes, we evaluate the additional predictive power rare variants could have. If the aim was to develop most accurate predictions of traits and diseases from genetic data, the predictive models would have to become more complex, for example by learning which genetic variants are important for each trait or disease.

In the project on genomic predictions I wrote programs to compute genomic kernels, whereas Athina Spiliopoulou set up the predictive model and experiments.

### 5.1.1 Polygenic model

Known associated SNPs which exceed genome-wide significance threshold can explain 10 % of phenotypic variance for height (Allen et al., 2010), 1.45 % for BMI (Speliotes et al., 2010) and 12.1 % for HDL cholesterol (Teslovich et al., 2010). SNPs that do not reach the threshold could explain some of the remaining

variance (Yang et al., 2010).

Yang et al. set up a polygenic model to estimate the variance explained by common SNPs. The authors utilise a linear additive model of quantitative traits:

$$y_j = \mu + g_j + e_j$$

where $y_j$ is a measurement of a phenotype, $\mu$ is mean, $g_j$ is a genetic component and $e_j$ is the environmental component. The genetic component is a sum of effects of individual variants:

$$g_j = \Sigma_{i=1}^{m} z_{i,j} u_i$$

where $m$ is the number of variants affecting the trait, $z_{i,j}$'s are allele dosages normalised with respect to the allele frequencies, and $u_i$ are the variant effects. Genetic variance is recovered from this equation, using restricted maximum likelihood approach, or equivalently by regressing genomic relatedness of pairs onto the square of difference of the phenotypic values.

$$var(\mathbf{y}) = \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2$$

where $\mathbf{G}$ is the genomic relationship matrix between pairs of individuals at causal loci, for which they obtain an estimate by computing dot products between all available SNP genotypes for pairs of individuals.

With this approach Yang et al. estimate that 0.45 of genetic variance for human height is explained by common SNPs. Not all of variance is explained, because some of the rare causal variants are not in LD with the SNP data available. The authors conclude that the remaining missing heritability is due to rare variants, some of them in weak linkage with genotyped SNPs. However, the heritability computed in this way accounts for only such variants that influence the traits independently and linearly. Zuk et al. (Zuk et al., 2012) claim that many of the variants interact in pathways, and therefore estimates of heritability from relatives and computed as above are inconsistent.

We build a model for genomic predictions borrowing from the polygenic model, which serves us our baseline for predictive accuracy from common SNPs.

### 5.1.2 Kernel-based genomic predictions

The estimate of $\mathbf{G}$, which Yang et al. compute from pair-wise dot-products of all SNP genotypes, is an example of a kernel. We propose another kernel, one based on haplotype IBD sharing which should convey sharing of rare variants. We also describe kernel ridge-regression, with which the kernels can be used for genomic predictions.

Kernel functions provide a measure of similarity between two items (Hofmann et al., 2005), in our case between two SNP genotypes. Valid kernels represent original observations in a different, possibly infinite-dimensional, space. This is equivalent to the kernel matrix, a result of evaluating the kernel function between all pairs of items, being positive definite. Valid kernels can be defined not only between real-valued vector items, but also between other types of data, for example strings.

As a baseline we used a linear kernel, borrowed from Yang et al. If genotype vectors of two individuals are denoted as $\mathbf{x}_i$,$\mathbf{x}_k$, where the allele dosages have been normalised with respect to frequencies for each SNP, then the kernel function for the two takes the following form:

$$k(\mathbf{x}_i, \mathbf{x}_k) = \mathbf{x}_i^\top \mathbf{x}_k$$

.

We compare this with an alternative kernel based on haplotypes, which implicitly conveys the extent to which haplotypes of a pair of individuals are identical by descent. We compute string kernels on the haplotypes, weighting them so that sharing a long segment of a haplotype contributes to similarity more than sharing several sub-strings. Given two haplotypes strings $(\mathbf{h}_A, \mathbf{h}_B)$, the IBD kernel we propose is computed as follows:

$$k(\mathbf{h}_A, \mathbf{h}_B) \quad = \quad (\Sigma_{s \in S} \text{length}(s)^p)^{\frac{1}{p}} \tag{5.1}$$

where $S$ is a set of identical sub-strings between the two haplotype strings, with their length being the number of SNPs in an identical region, and $p$ is a

parameter dictating how much more weighted are longer continuously matching sub-strings. The kernels between haplotypes are then combined to give a similarity measure between two individuals. If genotype $\mathbf{x}_i$ consists of haplotypes $\mathbf{h}_{i,1}$ and $\mathbf{h}_{i,2}$, and $\mathbf{x}_k$ accordingly, the kernel between the individuals is a sum over all four combinations of haplotypes:

$$k(\mathbf{x}_i, \mathbf{x}_k) = k(\mathbf{h}_{i,1}, \mathbf{h}_{k,1}) + k(\mathbf{h}_{i,1}, \mathbf{h}_{k,2}) + k(\mathbf{h}_{i,2}, \mathbf{h}_{k,1}) + k(\mathbf{h}_{i,2}, \mathbf{h}_{k,2})$$

I have proved that this kernel is positive definite.

In order to build a model for genomic predictions, we plugged both of the kernels into kernel ridge regression. Ridge regression is a linear regression model, which can cope with low number of samples compared to number variables thanks to L2 penalty on sum of squared weights, denoted by $\lambda$. It can also be used with kernel functions, which implicitly represent SNP data in another space. In our experiments, for training of the model and the predictions we used implementation of Multiple Kernel Learning, which as a special case includes kernel ridge regression (Bach et al., 2004). We trained and evaluated the models using cross-validation. We evaluated quality of phenotypic predictions on the test data by computing Pearson's correlation between predicted and measured phenotypic values.

### 5.1.3 Genotype and phenotype data

In order to demonstrate the predictive model, we used genetic and phenotypic data from three Croatian populations. Two of the populations have been isolated on the islands of Korcula and Vis, and the third group of samples comes from the mainland city of Split. Overall, 2,186 individuals were genotyped with SNP arrays, such that after standard quality control the intersection of SNPs contained 267,912 SNPs. Available were measurements of the following phenotypes we wish to predict from genotype data: height, body-mass index (BMI), HDL cholesterol. Individuals with poor genotyping or phenotypic measurements standing out from the mean by four standard deviations were removed. HDL cholesterol measurements were log-transformed since after the transformation the distribution of values resembled normal distribution. All phenotypic measurements were nor-

malised to a mean of 0 and standard deviation of 1.

### 5.1.4 Results

On test data, in experiments with height, HDL cholesterol and BMI, we could not establish with statistical significance that the kernel encoding identity by descent outperformed the linear kernel. Figure 5.1 shows correlations between predicted and measured genotypes in a 5-fold cross-validation on validation data. Because on the validation data we also choose optimal value of the parameter $\lambda$, this experiment could return biased results, performing better than with unseen data.

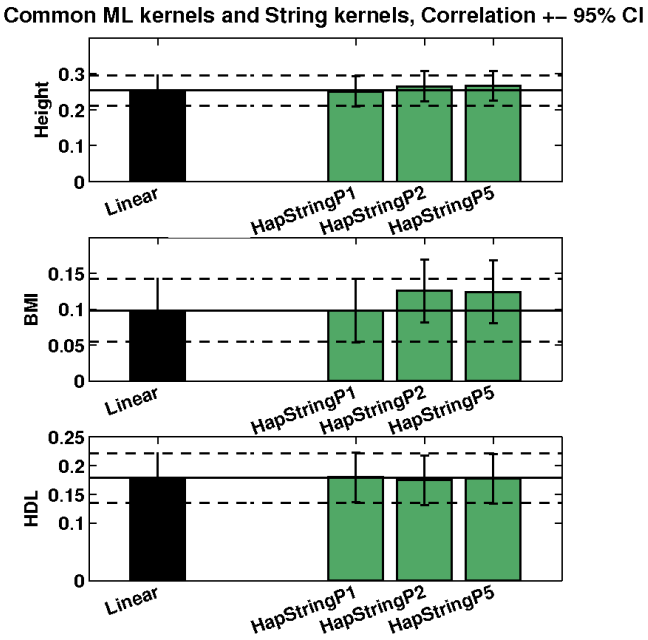



Figure 5.1: Correlations between real and predicted phenotype measurements on validation data. P1, P2, P5 correspond to p in Equation 5.1 taking values 1, 2 and 5 respectively. Kernels based on IBD are not predicting the phenotypes significantly better than the linear kernel.

### 5.1.5 Conclusions

We were not able to answer conclusively whether rare variants as captured by recent identity by descent matter for genomic predictions. Either the kernel constructions we proposed are not better at capturing the rare variants than the linear kernel, or our experiments had too few samples to show this.

Different quantitative traits may be influenced by common SNPs or rare variants to different extent. Both height and BMI are thought to be highly polygenic, so the results of our experiments should not be surprising. Fewer variants have been associated with HDL cholesterol, so it is for this trait that we might have expected to see an improvement in predictive accuracy with an IBD-based kernel.

A possible problem with the presented experiments is that we give performance measures on validation data, which as result could be better than for unseen data. This problem could be tackled by setting up a fully nested cross-validation procedure, where performance would be measured on test data rather than validation data. However, since there is only one real-valued parameter tuned on the validation data, $\lambda$, predictive accuracy on the validation data will likely generalise to unseen test data.

High relatedness of the samples from the isolated populations could also affect our results. On one hand, this ensures that there are more segments of IBD which we could utilise. On the other hand, our model may learn to depend on close relatives only, rather than to rely on any meaningful, causal genetic variations. A possible way to check whether this is the case is to remove pairs of very close relatives from the data and then again assess the predictive accuracy.

Finally, the predictive accuracy of our models could be further optimised. The models could learn to depend on SNPs or haplotype sharing in the predictive parts of the genome only. Such models could learn the regional dependencies from genetic and phenotypic data only, but could also utilise genomic annotations like information on chromatin structure.

## 5.2 Limitations of inference of identity by descent

We could not show the advantages of IBD analysis in genomic predictions over a simple linear SNP-based kernel. This, and the other presented application depends on whether the shared haplotypes also imply sharing rare variants. The shared haplotypes will only capture rare variants when they are recent enough, which could be estimated from the size of regions shared IBD.

### 5.2.1 Relationship between lengths of IBD segments and time to common ancestor

Length of IBD segments given time to a common ancestor can be modelled, from which we may learn about the reverse problem of dating back segments of IBD. If we assume that locations of crossover on gametes follow exponential distribution at each meiosis, and if $n$ is the number of generations to a common ancestor, then the expected length of a IBD haplotypes is $(2n)^{-1}$ Morgans, with variance of $(2n)^{-2}$.

When trying to estimate time to a common ancestor to an IBD segment, the problem is the opposite. It may be misleading to date IBD segments based on their length and the exponential distribution above (Ralph and Coop, 2012). For example, expected length of an IBD segment given a common ancestor 50 generations ago is 1 cM, but if an IBD segment is 1 cM long, then an ancestor more ancient than 50 generations ago is likely. This is because as time to a common ancestor increases, the number of lineages descending from that common ancestor grows fast. There are many individuals sharing ancient haplotypes, and some of them may be unusually long. As a consequence the distribution of time to the common ancestor given length of a haplotype shared IBD is heavily right-skewed towards older segments.

### 5.2.2 Mutation rates in old IBD segments

Older IBD segments that we detect from array data might have accumulated mutations in the time since their common ancestors, so that apparent IBD sharers do not share all variants. This would impair utility in such IBD segments in the applications such as optimising resequencing studies, identifying Mendelian subtypes of diseases and genomic predictions.

Mutations rates per generation were recently estimated in an Icelandic study of 78 parent-offspring trios utilising whole-genome next-generation sequencing (Kong et al., 2012). Average de novo mutation rate was found to be $1.20 \times 10^{-8}$ per nucleotide per generation at parental gametes, and less than third of at maternal gametes. The rate heavily varied with father's age, increasing 4.3 % per year.

The estimates indicate that mutations in IBD segments that we detect from array data are unlikely. With a common ancestor 50 generations ago, the expected length of an IBD segment is 1 cM, which on average corresponds to 1 Mb. If we accept Kong's estimate for mutation rate, the rate of mutation within 50 generations would be still less than $10^{-6}$ per nucleotide. We can therefore rely even on shortest detectable IBD segments, as their sharers share likely all of the variants.

A different situation may arise in Chapter 4. The Lynch Syndrome mutations are generally very rare, so their respective mutations might have occurred in near past. In our method for detecting unsuspected Lynch syndrome carriers we know that a new patient share IBD with a known carrier who inherited the mutation, so in other words we condition on the presence of a rare variant. Taking the low general mutation rate per nucleotide may be misleading for inference of carrier status this case.

## 5.3 Advantages of long-range methods for detecting recent IBD

Already existing short-range phasing might have been used for extracting regions of IBD. Algorithms like Beagle could output haplotypes, and when they continuously match between samples, we could declare them identical by descent.

Whether such strategy could be accurate depends on phasing accuracy of short-range methods.

Phasing accuracy of short- and long-range methods has been compared. Palin et al. computed haplotype accuracy of Beagle and Mach1, short-range phasing algorithms, and of SLRP, a long-range algorithm, and quote switch error rates on data from the ORCADES cohort (Palin et al., 2011). The long range method (SLRP) achieved phasing accuracy of 0.036-0.038 phase switches per centiMorgan, whereas Beagle 0.233-0.625 and Mach1 0.172-0.233 switches per centiMorgan. The accuracy for all methods dropped as the relatedness of samples decreased.

These results impact accuracy of inferring IBD from resulting haplotypes. If the switch errors follow the Poisson process with rate 0.625 cM (Beagle), the probability of a 2 cM region free from switch errors is 0.29. To detect IBD region of size 2 cM both haplotypes have to be switch error free, the probability of which is 0.08. When detecting IBD using short-range methods without accounting for switch errors, many segments would be missed. It is therefore clear that when searching for longer segments using short-range methods possible switch errors need to be taken into account. Long-range methods, such as SLRP, are more appropriate for finding IBD segments.

## 5.4 IBD analysis in future

Rise and growth of biobanks that store genetic data may be the driving force for development of new methods for analysis, for example using IBD segments. Only with high sampling fraction of the population can we find many relatively recent common ancestors. Expected number of haplotype sharers is proportional to average kinship in population and number of samples. The population history, size and number of Iceland is an ideal situation: out currently living 316000 inhabitants of the isolated population, 36000 were genotyped. Similar efforts are now undertaken in the UK, where similar biobanks are being built: for example Generation Scotland or the announced sequencing of whole-genomes from 100000 individuals in Britain (Prime Minister's Office, 2013).

As sizes of the biobanks grow, more important become scalability of algorithms for data analysis. The algorithm of ANCHAP we presented, as well as

many other algorithms require computation time of the order $n^2$, where $n$ is the number of samples. Our algorithm, and many others, could be optimised to linear complexity by using heuristics. For each individual, the analysis could include only a fixed number of closest related individuals. Closest-related individuals could be extracted from genetic relatedness computed as correlations of genome-wide SNP genotypes. Computation of such a relatedness matrix would still have to be optimised as computed naively, the time required still scales quadratically.

Another technological change is the shift from SNP genotyping technology to next-generation sequencing. The advantages of next generation sequencing is that they reveal all variants in DNA, and imputations would be no longer necessary. Because next-generation reads are from one chromosome each, this information can be also used to help phasing (Menelaou and Marchini, 2013). While the costs of next generation sequencing is still significant, combining array data with low-coverage sequencing data could be a cost-efficient option. As genotyping errors in next-generation sequencing are more common than in array data, the algorithms for detecting IBD will necessarily have to handle genotyping errors.

We can hypothesise what we could do had next-generation sequencing been available for the colon cancer patients and Lynch syndrome carriers in Chapter 4. Detecting unsuspected carriers of Lynch syndrome would involve only searching for known mutations in the known mismatch-repair genes. However, for discovering novel mutations the IBD algorithms would still be useful. When we see a novel variant, and the person carrying it shares IBD at a mismatch-repair gene with other affected carriers, the variant is likely very harmful as the co-inheritance of the same haplotype by chance is unlikely. The same idea can be used when detecting new genes associated with the syndrome: frequent IBD sharing of affected patients at a locus is unlikely by chance.

There are also other possible applications of IBD segments not mentioned in the thesis, one of which is searching for genes associated with idiosyncratic drug reactions. Because such reactions can only be detected among patients who were given a drug, the inheritance pattern is often missed. As a consequence, linkage studies would not be appropriate for identifying the responsible variants. With IBD analysis, we may detect distant relatives affected. We may attempt to map the susceptibility genes by noting loci in the genome where haplotype sharing is

more common than elsewhere. A key advantage of this method is that it would allow prediction of the severity of the disease phenotype, based on phenotypes of the haplotype sharers.

In summary, IBD analysis will require development of large biobanks, algorithm improvement and adjustment to data from next-generation sequencing. When these become available, opportunities for other applications of IBD analysis arise.

# References

Aaltonen, L. A., Salovaara, R., Kristo, P., Canzian, F., Hemminki, A., Peltomäki, P., Chadwick, R. B., Kääriäinen, H., Eskelinen, M., Järvinen, H. et al. (1998), 'Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease', *New England Journal of Medicine* **338**(21), 1481–1487. 81

Albrechtsen, A., Moltke, I. and Nielsen, R. (2010), 'Natural selection and the distribution of identity-by-descent in the human genome', *Genetics* **186**(1), 295–308. 53

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S. et al. (2010), 'Hundreds of variants clustered in genomic loci and biological pathways affect human height', *Nature* **467**(7317), 832–838. 110

Arcos-Burgos, M. and Muenke, M. (2002), 'Genetics of population isolates.', *Clinical Genetics* **61**(4), 233–247. 2

Bach, F. R., Lanckriet, G. R. and Jordan, M. I. (2004), Multiple kernel learning, conic duality, and the smo algorithm, *in* 'Proceedings of the twenty-first international conference on Machine learning', ACM, p. 6. 113

Barnetson, R., Tenesa, A., Farrington, S., Nicholl, I., Cetnarskyj, R., Porteous, M., Campbell, H. and Dunlop, M. (2006), 'Identification and survival of carriers of mutations in dna mismatch-repair genes in colon cancer', *New England Journal of Medicine* **354**(26), 2751–2763. 81, 83, 106, 107

Browning, B. L. and Browning, S. R. (2011*a*), 'A fast, powerful method for detecting identity by descent.', *American Journal of Human Genetics* **88**(2), 173–182. 50

Browning, S. and Browning, B. (2007), 'Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering', *American Journal of Human Genetics* **81**(5), 1084. 13, 57

Browning, S. R. and Browning, B. L. (2010), 'High-resolution detection of identity by descent in unrelated individuals.', *American Journal of Human Genetics* **86**(4), 526–539. 20, 23, 85

Browning, S. R. and Browning, B. L. (2011*b*), 'Haplotype phasing: existing methods and new developments.', *Nature Reviews Genetics* **12**(10), 703–714. 12, 51, 77

Browning, S. R. and Thompson, E. A. (2012), 'Detecting rare variant associations by identity-by-descent mapping in case-control studies', *Genetics* **190**(4), 1521–1531. 16

Consortium, . G. P. (2012), 'An integrated map of genetic variation from 1,092 human genomes', *Nature* **491**, 1. 5, 78

Consortium, I. H. (2007), 'A second generation human haplotype map of over 3.1 million snps.', *Nature* **449**(7164), 851–861. 5, 33, 52

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. and Lander, E. S. (2001), 'High-resolution haplotype structure in the human genome.', *Nature Genetics* **29**(2), 229–232. 57

Delaneau, O., Marchini, J. and Zagury, J. (2011), 'A linear complexity phasing method for thousands of genomes', *Nature Methods* **9**(2), 179–181. 14

Delaneau, O., Zagury, J. and Marchini, J. (2012), 'Improved whole-chromosome phasing for disease and population genetic studies', *Nature Methods* **10**(1), 5–6. 14

Den Dunnen, J. T., Antonarakis, S. E. et al. (2000), 'Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion', *Human Mutation* **15**(1), 7–12. 83, 99

Duret, L. (2009), 'Mutation patterns in the human genome: more variable than expected.', *PLoS Biol* **7**(2), e1000028. 23

Ferreira, R. C., Pan-Hammarström, Q., Graham, R. R., Fontán, G., Lee, A. T., Ortmann, W., Wang, N., Urcelay, E., Fernández-Arquero, M., Núñez, C. et al. (2012), 'High-density snp mapping of the hla region identifies multiple independent susceptibility loci associated with selective iga deficiency', *PLoS Genetics* **8**(1), e1002476. 53

Gansner, E. R. and North, S. C. (2000), 'An open graph visualization system and its applications to software engineering', *Software - Practice and Experience* **30**(11), 1203–1233. 85

Genovese, G., Leibon, G., Pollak, M. R. and Rockmore, D. N. (2010), 'Improved ibd detection using incomplete haplotype information.', *BMC Genetics* **11**, 58. 25

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Chang, L.-Y., Huang, W., Liu, B., Shen, Y. et al. (2003), 'The international hapmap project', *Nature* **426**(6968), 789–796. 57

Glodzik, D., Navarro, P., Vitart, V., Hayward, C., McQuillan, R., Wild, S. H., Dunlop, M. G., Rudan, I., Campbell, H., Haley, C. et al. (2013), 'Inference of identity by descent in population isolates and optimal sequencing studies', *European Journal of Human Genetics* . iii, 19, 84

Gusev, A., Kenny, E., Lowe, J., Salit, J., Saxena, R., Kathiresan, S., Altshuler, D., Friedman, J., Breslow, J. and Pe'er, I. (2011), 'Dash: a method for identical-by-descent haplotype mapping uncovers association with recent variation', *The American Journal of Human Genetics* **88**(6), 706–717. 17, 108

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M. and Pe'er, I. (2009), 'Whole population, genome-wide mapping of hidden relatedness', *Genome Research* **19**(2), 318–326. 16, 85

Gusev, A., Shah, M. J., Kenny, E. E., Ramachandran, A., Lowe, J. K., Salit, J., Lee, C. C., Levandowsky, E. C., Weaver, T. N., Doan, Q. C., Peckham, H. E., McLaughlin, S. F., Lyons, M. R., Sheth, V. N., Stoffel, M., De La Vega, F. M., Friedman, J. M., Breslow, J. L. and Pe'er, I. (2012), 'Low-Pass Genome-Wide Sequencing and Variant Inference Using Identity-by-Descent in an Isolated Human Population', *Genetics* **190**(2), 679–689. 58, 65, 66, 79

Haldane, J. (1919), 'The combination of linkage values and the calculation of distances between the loci of linked factors', *Journal of Genetics* **8**(29), 299–309. 20

Han, L. and Abney, M. (2012), 'Using identity by descent estimation with dense genotype data to detect positive selection', *European Journal of Human Genetics* . 53

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. and Lander, E. (1992), 'Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in finland.', *Nature Genetics* **2**(3), 204–211. 8

Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N. and van der Werf, J. H. J. (2011), 'A combined long-range phasing and long haplotype imputation method to impute phase for snp genotypes.', *Genetics Selelection Evolution* **43**, 12. 14, 21

Hofmann, T., Schölkopf, B. and Smola, A. J. (2005), 'A tutorial review of rkhs methods in machine learning'. 112

Howie, B., Donnelly, P. and Marchini, J. (2009), 'A flexible and accurate genotype imputation method for the next generation of genome-wide association studies', *PLoS Genetics* **5**(6), e1000529. 13, 57

Howie, B., Marchini, J. and Stephens, M. (2011), 'Genotype imputation with thousands of genomes', *G3: Genes, Genomes, Genetics* **1**(6), 457–470. 58

Johnson, G., Esposito, L., Barratt, B., Smith, A., Heward, J., Di Genova, G., Ueda, H., Cordel, H., Eaves, I. and Dudbridge, F. (2001), 'Haplotype tagging for the identification of common disease genes', *Nature Genetics* **29**(2), 233–237. 3

Joshi, P. K., Prendergast, J., Fraser, R. M., Huffman, J. E., Vitart, V., Hayward, C., McQuillan, R., Glodzik, D., Polašek, O., Hastie, N. D. et al. (2013), 'Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies', *PLOS ONE* **8**(7), e68604. iii, 62, 78

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A. et al. (2012), 'Rate of de novo mutations and the importance of father/'s age to disease risk', *Nature* **488**(7412), 471–475. 117

Kong, A., Masson, G., Frigge, M., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P., Ingason, A., Steinberg, S., Rafnar, T. et al. (2008*a*), 'Detection of sharing by descent, long-range phasing and haplotype imputation.', *Nature Genetics* **40**(9), 1068–1075. 14, 21, 24

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T. et al. (2008*b*), 'Detection of sharing by descent, long-range phasing and haplotype imputation', *Nature Genetics* **40**(9), 1068–1075. 19, 82

Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K., Jonasdottir, A. et al. (2009), 'Parental origin of sequence variants associated with complex diseases.', *Nature* **462**(7275), 868–874. 14, 21

Kristiansson, K., Naukkarinen, J. and Peltonen, L. (2008), 'Isolated populations and complex disease gene identification.', *Genome Biol* **9**(8), 109.
**URL:** *http://dx.doi.org/10.1186/gb-2008-9-8-109* 3

Kschischang, F. R., Frey, B. J. and Loeliger, H.-A. (2001), 'Factor graphs and the sum-product algorithm', *Information Theory, IEEE Transactions on* **47**(2), 498–519. 15

Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010), 'Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes', *Genet Epidemiology* **34**(8), 816–834. 13, 57

Lunter, G. and Goodson, M. (2011), 'Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads', *Genome research* **21**(6), 936–939. 62

Lynch, H., Lynch, P., Lanspa, S., Snyder, C., Lynch, J. and Boland, C. (2009), 'Review of the lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications', *Clinical Genetics* **76**(1), 1–18. 81, 107

Lynch, H. T. and de la Chapelle, A. (1999), 'Genetic susceptibility to non-polyposis colorectal cancer', *Journal of Medical Genetics* **36**(11), 801–818. 81, 82

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007), 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nature Genetics* **39**(7), 906–913. 13

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J. and Hirschhorn, J. (2008), 'Genome-wide association studies for complex traits: consensus, uncertainty and challenges', *Nature Reviews Genetics* **9**(5), 356–369. 7

McEvoy, B. P., Montgomery, G. W., McRae, A. F., Ripatti, S., Perola, M., Spector, T. D., Cherkas, L., Ahmadi, K. R., Boomsma, D., Willemsen, G., Hottenga, J. J., Pedersen, N. L., Magnusson, P. K. E., Kyvik, K. O., Christensen, K., Kaprio, J., Heikkilä, K., Palotie, A., Widen, E., Muilu, J., Syvänen, A.-C., Liljedahl, U., Hardiman, O., Cronin, S., Peltonen, L., Martin, N. G. and Visscher, P. M. (2009), 'Geographical structure and differential natural selection among north european populations.', *Genome Research* **19**(5), 804–814. 54

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. et al. (2010), 'The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data', *Genome research* **20**(9), 1297–1303. 62

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., Macleod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., Campbell, H. and Wilson, J. F. (2008), 'Runs of homozygosity in european populations.', *American Journal of Human Genetics* **83**(3), 359–372. 22

Menelaou, A. and Marchini, J. (2013), 'Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold', *Bioinformatics* **29**(1), 84–91. 119

Metzker, M. L. (2009), 'Sequencing technologies—the next generation', *Nature Reviews Genetics* **11**(1), 31–46. 5

Moreira, L., Balaguer, F., Lindor, N., de la Chapelle, A., Hampel, H., Aaltonen, L. A., Hopper, J. L., Le Marchand, L., Gallinger, S., Newcomb, P. A. et al. (2012), 'Identification of lynch syndrome among patients with colorectal cancerlynch syndrome and colorectal cancer', *JAMA: The Journal of the American Medical Association* **308**(15), 1555–1565. 81, 107

Morris, A. P., Whittaker, J. C. and Balding, D. J. (2002), 'Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies.', *American Journal of Human Genetics* **70**(3), 686–707. 2, 3, 10

Morton, N. E. (1956), 'The detection and estimation of linkage between the genes for elliptocytosis and the rh blood type', *American Journal of Human Genetics* **8**(2), 80. 6

Moskvina, V., Smith, M., Ivanov, D., Blackwood, D., Stclair, D., Hultman, C., Toncheva, D., Gill, M., Corvin, A., O'Dushlaine, C., Morris, D. W., Wray, N. R., Sullivan, P., Pato, C., Pato, M. T., Sklar, P., Purcell, S., Holmans, P.,

O'Donovan, M. C., Owen, M. J. and Kirov, G. (2010), 'Genetic differences between five european populations.', *Human Heredity* **70**(2), 141–149. 54

Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005), 'A fine-scale map of recombination rates and hotspots across the human genome', *Science* **310**(5746), 321–324. 23, 52

Nachman, M. W. and Crowell, S. L. (2000), 'Estimate of the mutation rate per nucleotide in humans.', *Genetics* **156**(1), 297–304. 23

Nekrutenko, A. and Taylor, J. (2012), 'Next-generation sequencing data interpretation: enhancing reproducibility and accessibility', *Nature Reviews Genetics* **13**(9), 667–672. 5

Palin, K., Campbell, H., Wright, A. F., Wilson, J. F. and Durbin, R. (2011), 'Identity-by-descent-based phasing and imputation in founder populations using graphical models.', *Genet Epidemiology* **35**(8), 853–860. 14, 16, 21, 118

Peltonen, L., Palotie, A., Lange, K. et al. (2000), 'Use of population isolates for mapping complex traits', *Nature Reviews Genetics* **1**(3), 182–190. 8

Polasek, O., Marusic, A., Rotim, K., Hayward, C., Vitart, V., Huffman, J., Campbell, S., Jankovic, S., Boban, M., Biloglav, Z., Kolcic, I., Krzelj, V., Terzic, J., Matec, L., Tometic, G., Nonkovic, D., Nincevic, J., Pehlic, M., Zedelj, J., Velagic, V., Juricic, D., Kirac, I., Kovacevic, S. B., Wright, A. F., Campbell, H. and Rudan, I. (2009), 'Genome-wide association study of anthropometric traits in korcula island, croatia.', *Croatian Medical Journal* **50**(1), 7–16. 22

Powell, J. E., Visscher, P. M. and Goddard, M. E. (2010), 'Reconciling the analysis of ibd and ibs in complex trait studies.', *Nature Reviews Genetics* **11**(11), 800–805. 2, 20

Prime Minister's Office, . D. S. (2013), 'Dna tests to revolutionise fight against cancer and help 100,000 nhs patients'.
**URL:** *https://www.gov.uk/government/news/* 118

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. and Sham, P. C. (2007), 'Plink: a tool set for whole-genome association and population-based linkage analyses.', *American Journal of Human Genetics* **81**(3), 559–575. 33

Ralph, P. and Coop, G. (2012), 'The geography of recent genetic ancestry across europe', *ArXiv preprint ArXiv:1207.3815* . 101, 116

Rosenberg, N., Nordborg, M. et al. (2002), 'Genealogical trees, coalescent theory and the analysis of genetic polymorphisms', *Nature Reviews Genetics* **3**(5), 380–390. 9

Scheet, P. and Stephens, M. (2006), 'A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase', *The American Journal of Human Genetics* **78**(4), 629–644. 13

Scotland, I. (2013), 'Cancer statistics', http://www.isdscotland.org/Health-Topics/Cancer/Cancer-Statistics/Colorectal/. 108

Sham, P. C., Cherny, S. S., Purcell, S. and Hewitt, J. K. (2000), 'Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data.', *Am J Hum Genet* **66**(5), 1616–1630.
**URL:** *http://dx.doi.org/10.1086/302891* 7

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R. et al. (2010), 'Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index', *Nature Genetics* **42**(11), 937–948. 110

Stephens, M., Smith, N. and Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *The American Journal of Human Genetics* **68**(4), 978–989. 13

Stranger, B. E., Stahl, E. A. and Raj, T. (2011), 'Progress and promise of genome-wide association studies for human complex trait genetics', *Genetics* **187**(2), 367–383. 5

Syvänen, A.-C. (2001), 'Accessing genetic variation: genotyping single nucleotide polymorphisms', *Nature Reviews Genetics* **2**(12), 930–942. 5

Teare, M. and Barrett, J. (2005), 'Genetic linkage studies', *The Lancet* **366**(9490), 1036–1044. 6

Teh, Y. W., Blundell, C. and Elliott, L. (2011), Modelling genetic variations using fragmentation-coagulation processes, *in* J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger, eds, 'Advances in Neural Information Processing Systems 24', pp. 819–827. 11, 12

Tenesa, A., Farrington, S. M., Prendergast, J. G., Porteous, M. E., Walker, M., Haq, N., Barnetson, R. A., Theodoratou, E., Cetnarskyj, R., Cartwright, N. et al. (2008), 'Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21', *Nature Genetics* **40**(5), 631–637. 83

Terwilliger, J. D. and Weiss, K. M. (1998), 'Linkage disequilibrium mapping of complex disease: fantasy or reality?', *Current Opinion Biotechnology* **9**(6), 578–594. 3, 7

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J. et al. (2010), 'Biological, clinical and population relevance of 95 loci for blood lipids', *Nature* **466**(7307), 707–713. 110

Vasen, H. F., Möslein, G., Alonso, A., Bernstein, I., Bertario, L., Blanco, I., Burn, J., Capella, G., Engel, C., Frayling, I. et al. (2007), 'Guidelines for the clinical management of lynch syndrome (hereditary non-polyposis cancer)', *Journal of Medical Genetics* **44**(6), 353–362. 81

Vitart, V., Biloglav, Z., Hayward, C., Janicijevic, B., Smolej-Narancic, N., Barac, L., Pericic, M., Klaric, I. M., Skaric-Juric, T., Barbalic, M., Polasek, O., Kolcic, I., Carothers, A., Rudan, P., Hastie, N., Wright, A., Campbell, H. and Rudan, I. (2006), '3000 years of solitude: extreme differentiation in the island isolates of dalmatia, croatia.', *European Journal of Human Genetics* **14**(4), 478–487. 22

Wagner, T., Stoppa-Lyonnet, D., Fleischmann, E., Muhr, D., Pages, S., Sandberg, T., Caux, V., Moeslinger, R., Langbauer, G., Borg, A. et al. (1999), 'Denaturing high-performance liquid chromatography detects reliably¡ i¿ brca1¡/i¿ and¡ i¿ brca2¡/i¿ mutations', *Genomics* **62**(3), 369–376. 83

Wilson, J. F., Weiss, D. A., Richards, M., Thomas, M. G., Bradman, N. and Goldstein, D. B. (2001), 'Genetic evidence for different male and female roles during cultural transitions in the british isles.', *Proceedings of the National Academy of Sciences* **98**(9), 5078–5083. 22

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010), 'Common snps explain a large proportion of the heritability for human height', *Nature Genetics* **42**(7), 565–569. 86, 110, 111

Zeggini, E. (2011), 'Next-generation association studies for complex traits.', *Nature Genetics* **43**(4), 287–288. 56

Zuk, O., Hechter, E., Sunyaev, S. R. and Lander, E. S. (2012), 'The mystery of missing heritability: genetic interactions create phantom heritability', *Proceedings of the National Academy of Sciences* **109**(4), 1193–1198. 111