Mathematical Statistics
Stockholm University

Evaluation of a multipoint method for imputing genotypes using HapMap III

Emil Rehnberg

Examensarbete 2009:5

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2009:5**,
http://www.math.su.se/matstat

Evaluation of a multipoint method for imputing genotypes using HapMap III

Emil Rehnberg*

June 2009

Abstract

The common disease-common variant hypothesis postulates that multiple common genetic variants influence the susceptibility of complex diseases. By genotyping a large number of genetic markers (SNPs) across the genome and performing association studies (GWAS), one can identify regions that harbor such disease susceptability variants. However, statistical power is a continuing obstacle as GWAS require very large sample sizes. By combining study samples from several studies, the statistical power increases and more reliable statistical inference is possible.

The lack of overlapping genetic markers constitutes a problem when combining study samples performed on different platforms. Here, imputation methodology is useful, in order to "fill in" the information for those SNPs that are present on a platform but not the other. However, it is not known whether certain choices of genetic markers are more suitable for imputation than others or if different reference populations are preferred for the imputations. In this thesis we compare two studies, CAPS and CAHRES that were genotyped using chips from Affymetrix and Illumina respectively, and two reference popultations, CEU from HapMap phase 2 and phase 3.

Validation of imputations, carried out on samples from the two studies that use different genetic markers, show that individuals from CAHRES impute better than those from CAPS. Possible explanations for the difference in imputation results are the selection of genetic markers, the quantity of genetic markers and how well the reference population resembles the study sample individuals.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: emil.rehnberg@gmail.com Supervisors: Juni Palmgren, Monica Leu, Keith Humphreys

Contents

1 Introduction **4**

2 A short guide to genetic terminology **5**

3 Data **6**

3.1 CAPS 7

3.2 CAHRES 7

4 The International HapMap Project **7**

5 Statistical Methods **9**

5.1 Imputation 9

5.2 Hidden Markov Models 10

5.3 Imputing genotypes using a Hidden Markov Model 12

5.4 Incomplete data model 12

5.4.1 Transition probabilities 13

5.4.2 Mutation probabilities 15

5.5 Methods for validating imputations 16

5.5.1 PRESS 16

5.5.2 RMSEP 17

5.6 Imputation of CAPS and CAHRES 17

5.7 Validation of imputation on CAPS and CAHRES 18

6 Results **19**

7 Discussion **22**

8 Appendix **24**

8.1 Computational considerations 24

8.2 HapMap phase comparison table 24

8.3 Results plots 25

9 Acknowledgements **29**

1 Introduction

During the past decade efforts have been put into finding genetic factors that contribute to complex diseases such as many cancers, cardiovascular disease, Alzheimer's disease and Crohn's disease. Many of these efforts have been based upon the common disease-common variant (CD-CV) hypothesis [1] which postulates that multiple genetic variants influence the susceptibility of complex diseases. Most of the human genome is the same from person to person. The single-base locations on the DNA sequence that can take multiple forms are called SNPs (Single Nucleotide Polymorphism). The CD-CV hypothesis has driven genetic research to search for SNPs that might be associated with complex diseases. The procedure of reading the pair of alleles constituting the genotypes at certain SNP loci is called "genotyping". To establish whether SNPs are associated with disease, the value of genetic variants is compared across individuals with different phenotype (disease) values. These studies are known as genetic association studies. The case-control study, for example, compares cases (i.e. individuals with the disease) and controls (individuals free from the disease). In recent years it has become possible to genotype a large number (hundreds of thousands of SNPs) across the genome at a reasonable cost, and association studies that utilizes this technology are known as a Genome Wide Association Study (GWAS) [2].

In 2007, the Wellcome Trust Case-Control Consortium published an article based on GWAS, providing conslusive evidence that certain genes are involved in Crohn's disease, type 1 and 2 diabetes, rheumatoid arthritis, bipolar disease, and coronary artery disease [3]. They compare 7 different disease groups each with 2000 cases (seperately) to a single set of 3000 controls. This provided evidence that controls can be shared between GWAS studies. In building up to this success of the GWAS study, there were three crucial developments. First, the international HapMap project, which documents the variation and correlation between known SNPs in the human genome [4]. Second, the evolution of dense genotyping chips made the high throughtput genotyping possible for hundreds of thousands of SNPs, providing good genome coverage. Third, reasonably large and well-characterized sample collections were assembled for multiple common diseases.

Genotyping cost is one of the obstacles when perfoming GWAS. Sharing controls for different studies decreases this burden. Creating control pools would save studies from genotyping controls for every study and focus could be more on the genotyping of cases.

When performing high throughput genotyping for an individual, not all of the known SNPs in the genome are genotyped since SNPs are correlated with each other and there is a lot of redundancy. Costs are proportional to the number of SNPs genotyped. Hence, one genotype a clever selection of SNPs that can capture most of the genomic variation for an individual. This selection of SNPs differs between different high throughput genotyping chips manifactured by genotyping companies. Two of these companies

are Affymetrix [5] and Illumina [6]. The percentage of overlapping SNPs for these two chips is about 20%. Thus individuals from studies genotyped using different genotyping chips are uncomparable and a simple merge of data is not possible. One solution is to impute the missing SNPs (i.e. fill in the missing SNP information) by using the correlation structure between SNPs from completely genotyped populations, and to add the imputed data to the individuals in the control pool. We can estimate the non-genotyped SNPs in the genome using the correlation structures of a suitable HapMap reference population. The correlation structure is a feature on the population level, so to estimate non-genotyped SNPs, one needs a reference population that has close genomic patterns to the population we predict missing genotypes for. Currently there are 2 relevant versions of HapMap, phase 2 and phase 3. Phase 3 includes a large number of populations, and uses a different approach regarding SNP quality controls compared to phase 2.

This thesis focuses on imputations for the purpose of creating a control pool. We ask the following questions: Are there selections of SNPs that are preferable for imputations? Which version of HapMap is most suitable for imputation in our studies? Are certain individual's genotypes hard to predict? Are certain SNPs hard to predict?

In section 2 there is a short genetic summary on genetic terminology used in this thesis. Sections 3 and 4 cover information on our data and HapMap respectively. The statistical methods for imputation and for validation of these imputations are covered in section 5 followed by a results section 6. An overall discussion on the methods used and the results are given in section 7.

2 A short guide to genetic terminology

Humans carry genetic information in the form of double helix string consisting of *nucleotides*. This double helix string is called *DNA* (deoxyribonucleic acid) where the four possible nucleotides form complementary pairs. The whole DNA sequence, also known as the *genome*, is partitioned into 46 pieces called *chromosomes*, 23 pairs with one pair of sex chromosomes. In Figure 1 the letters A, C, G and T illustrates the nucleotides, the double helixes in the top and bottom of the figure make up a part of one chromosome pair. On a chromosome, the nucleotides bind to each other deterministically, A always binds to T and C always binds to G.

Most of the DNA is identical between individuals, but at some *loci* (genetic positions, *locus* in singular) different variants exist called *alleles*. More specifically, one allele represents one locus and is consisting of one nucleotide. In Figure 1 the highlighted G and T make up the two alleles on opposite chromosomes. This pair (G,T) is called a genotype. If the rarer allele has a frequency of at least 1% then the variation at this locus constitutes a *SNP*

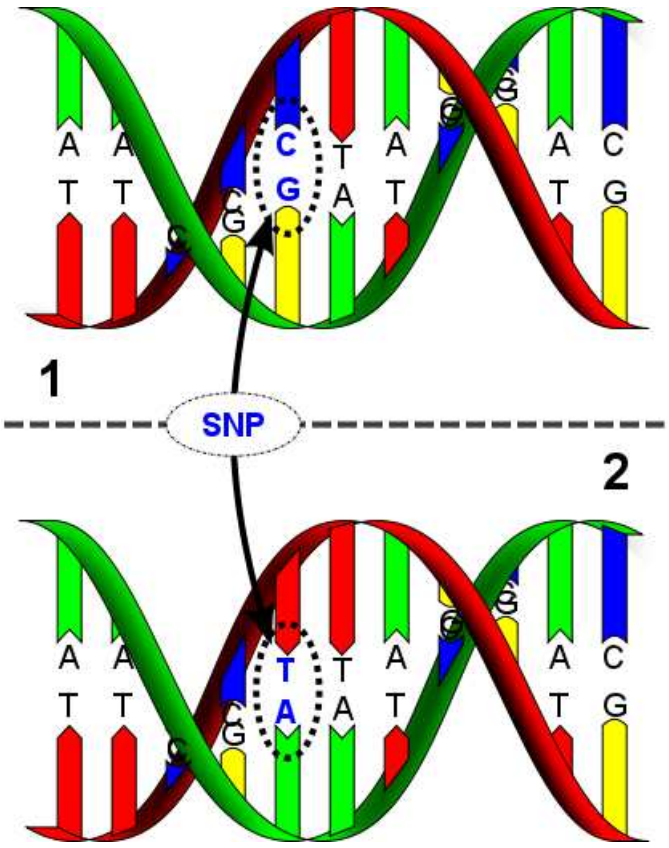

Figure 1: Illustration of the DNA structure

(Single Nucleotide Polymorphism). The ordered combination of alleles along a single chromosome is called a *haplotype*. The two haplotypes in Figure 1 are $TTCCCTAGGTG$ in the area labeled 1, and $TTCCTTAGGTG$ in area 2.

When the DNA is copied during *meiosis* (i.e. the making of sperm and egg cells), the copying process might be broken and continue the copying on the opposite chromosome. This occurence is called a *genetic recombination*, and makes the offspring inherit a combination of the chromosome segments from the two parents.

3 Data

In this thesis we use controls from two studies carried out at the Department of Medical Epidemiology and Biostatistics (MEB). We explore the possibility

of using controls from these two studies as a potential Swedish control pool, for comparing with Swedish cases, of a particular disease.

3.1 CAPS

The CAPS [CAncer Prostate in Sweden] study [7] has GWAS data on 1705 controls. The subjects were selected from the Swedish Population Registry, to geographically match the cases in that study. The controls were men (generally of age 60+) from two regions, where one region represents the north and the other represents the middle of Sweden. Table 1 shows the age statification for the CAPS study controls:

CAPS	# of contols	% of controls
$\leq$59	275	16.1
60-69	739	43.3
$\geq$70	691	40.5

Table 1: The age stratification of the CAPS controls

Due to financial issues, not all individuals could be genotyped. We will use the 1028 controls with genome-wide data in this thesis.

3.2 CAHRES

The other study was a Swedish breast cancer study, CAHRES (CAncer and Hormone REplacementS) [8]. The cases and controls were all Swedish-born women between 50 and 74 years of age, resident in Sweden from 1993 to 1995. The controls were randomly selected from the Swedish Population Registry to geographically match the cases in 5-year age strata (and all the cases were identified at diagnosis through the six regional cancer registries in Sweden). The original study included 3000 controls. Out of these, only 764 women were genotyped due to financial issues.

4 The International HapMap Project

The aim of the International HapMap Project has been to determine the common patterns of DNA sequence variation in the human genome, by characterizing sequence variants, their frequencies, and the correlations between them.

The HapMap project officially started late October 2002 as a collaboration between researchers from Canada, China, Japan, Nigeria, the United Kingdom and the United States. Originally the project studied 4 populations: 30 trios (mother, father and child) from Ibadan, Nigeria (YRI), another 30 trios from US residents with european ancestry (CEU), 44 unrelated people from Tokyo, Japan (JPT) and 45 unrelated individuals from Beijing, China (CHB).

The correlation structure between SNPs is a population feature, so it was necessary to include populations from different parts of the world. One way to study the common patterns of variation in the human DNA is to examine a haplotype map (a HapMap).

Currently there are two completed phases of the HapMap and a third phase draft is released. Phases 1 and 2 analysed 270 individuals from 4 different populations and has had multiple releases. Phase 3 on the other hand has released a first draft where 1115 individuals have been analysed from 11 populations. The expectation is that this new release will give a more accurate HapMap by genotyping more individuals from more populations and by being more stringent with the SNP genotype quality control (QC).
Because of the stricter QC, there are less SNPs in phase 3 than phase 2. The new phase contains about 38% of the SNPs from the previous phase (the remaining 62% were discarded) and 4% of the SNPs in phase 3 are new. Table 2 presents the populations currently in HapMap phase 3.

	Population
ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MEX	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscani in Italia
YRI	Yoruba in Ibadan, Nigeria

Table 2: The populations in HapMap phase 3

Table 3 shows an overview of the number of individuals and SNPs in the

two HapMap phases (and Table 9 on page 24 shows a more detailed comparisons between the two phases). Here, the CEU population is written in bold because of it's importance for this thesis.

Population	# individuals		# SNPs	
	Phase 2	Phase 3	Phase 2 (r19)	Phase 3 (r2)
ASW	NA	71	NA	1 632 186
CEU	**90**	**162**	**3 901 408**	**1 634 020**
CHB	45	82	3 903 524	1 637 672
CHD	NA	70	NA	1 619 203
GIH	NA	83	NA	1 631 060
JPT	45	82	3 902 623	1 637 610
LWK	NA	83	NA	1 631 688
MEX	NA	71	NA	1 614 892
MKK	NA	171	NA	1 621 427
TSI	NA	77	NA	1 629 957
YRI	90	163	3 806 920	1 634 666
Total	270	1115	3 819 322	1 525 445

Table 3: HapMap's populations sizes and number of SNPs over the two phases

5 Statistical Methods

5.1 Imputation

Adequate handling of missing data is a common problem in statistical modeling and inference. One approach is to discard observations that are incomplete. Another approach is to fill in the missing information. Using probability models to solve the puzzle in the latter approach is called imputing.

Imputation is often used to handle incompleteness of data so that an analysis can be performed on "complete" data. Otherwise, discarding observations containing missing information can reduce power and introduce a selection bias.

Figure 2 is an example of imputation for missing genotypes. The genotypes are coded as 0, 1 or 2, while the alleles are coded as 0 or 1 (the coding for a certain allele being 0 or 1 is arbitrary) and haplotypes are strings of 0 and 1. Missing genotypes are denoted as ?. By using haplotypes from a

reference popultaion it is possible to impute the ?:s, and hopefully end up with a table like the one to the right in Figure 2, with no ?:s. However, the table to the right is an over-simplification of the imputation output. The output are probability distributions over the genotypes 0,1 or 2 (as seen in the lower right of the figure), and these are then used to estimate the missing genotypes.

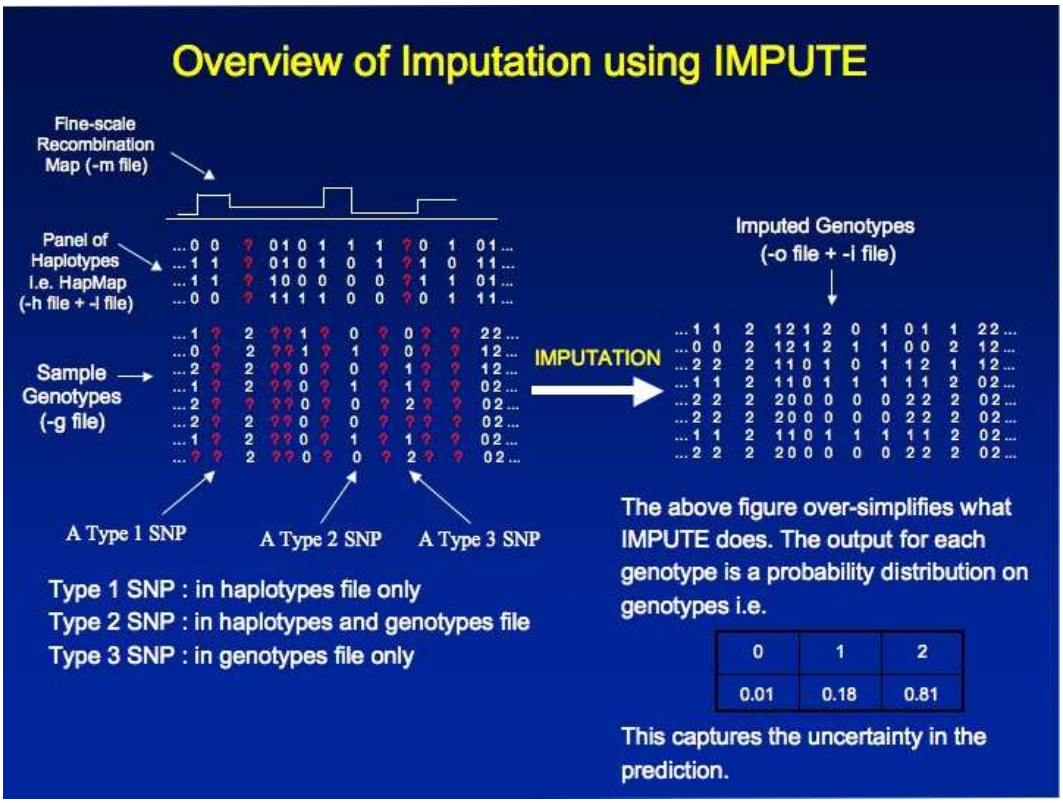

Figure 2: An example of how imputation works. A set of haplotypes is used to complete a set of incomplete genotypes.

There are many ways to perform imputations. In this thesis we will use a Hidden Markov Model for this purpose.

5.2 Hidden Markov Models

In a regular Markov model states can be observed directly, in contrast to the Hidden Markov Model (HMM), were the states are hidden and one observe

stochastic outcomes of these states. As the states have a probability distribution over it's outcomes, the observations of these outcomes thus carry information about the states themselves.

Figure 3 below provides a graphical representation of a HMM. Here $X = \{x_1, x_2, x_3\}$ represents the hidden states and $Y = \{y_1, y_2, y_3, y_4\}$ represents the possible outcomes of X, $\{a_{ij}\}$ are the state transition probabilities (from state i to state j) and $\{b_{ij}\}$ are the probabilities of the outcomes for each state (x_i having the outcome y_j).

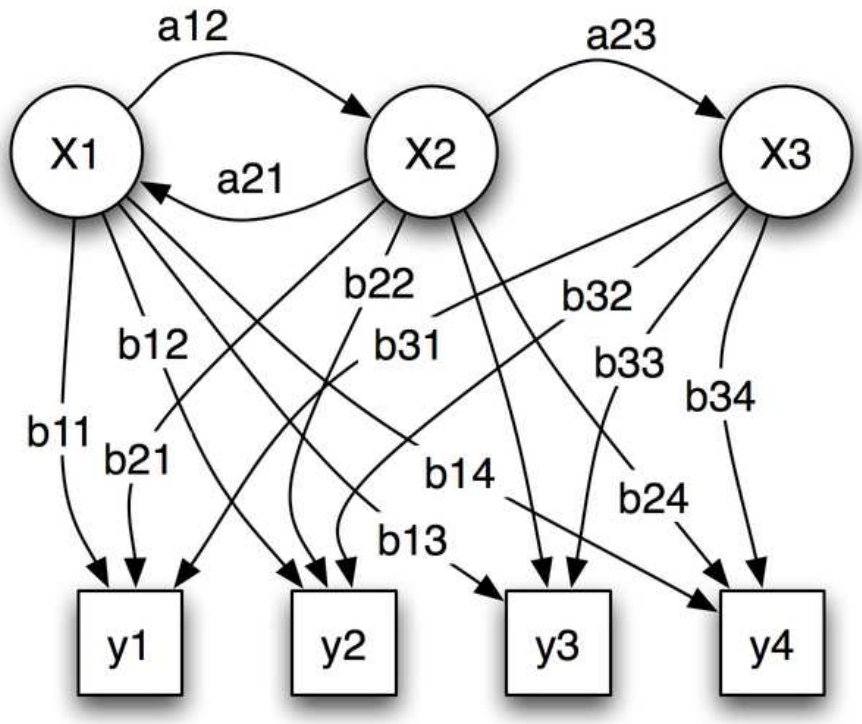

Figure 3: An example of a Hidden Markov Model. The set $\{X\}$ represents a Markov chain with hidden states, the set $\{Y\}$ are the outcomes of an element of $\{X\}$, $\{a_{ij}\}$ are the transition probabilities in the Markov chain and $\{b_{ij}\}$ are the probabilities of the outcomes from an element i of the set $\{X\}$

Returning to the example in Figure 2: the observations are the panel of haplotypes and the hidden states are the sample genotypes. More precisely,

given a locus, the hidden states are the two haplotypes for the two alleles in each genotype, hence each genotype is estimated using it's two haplotypes (at that locus). Due to mutations, the two haplotypes do not determine the genotype deterministically. Here, the transition probabilties are the probabilities of the genotype given the haplotypes at a locus.

5.3 Imputing genotypes using a Hidden Markov Model

Let us denote $H = \{H_1, .., H_N\}$ as a set of N known haplotypes where $H_i = \{H_{i1}, .., H_{iL}\}$ denote the alleles of haplotype i at the L SNP locus sites (in Figure 2, $N = 4$ and $L = 14$). Each H_{ij} is an indicator variable for whether the locus j in haplotype i contains the allele labelled "1" (at each locus there are two possible alleles, one is coded as "1" and the other as "0"). Now let $G = \{G_1, .., G_K\}$ be the random variables denoting genotypes of K individuals (in Figure 2, $K = 8$), where $G_i = \{G_{i1}, .., G_{iL}\}$ and each G_{ij} takes values in the set $\{0, 1, 2, missing\}$. I.e. the G_{ij} denotes the number of the allele coded as 1 at the SNP for individual i at locus j.

The basic idea of the imputation algorithm proposed by Marchini et. al [9] is to consider the N haplotypes in HapMap as "ancestral" and that all haplotypes observed in the current sample are derived (through mutations and recombination events) from this "pool" of ancestral haplotypes.

5.4 Incomplete data model

As we are interested in the distribution of the missing genotypes in the current sample, we partition the G vector into missing and observed genotypes $G = \{G_O, G_M\}$. Index O is for observed and M is for missing. By assuming that the genotype vectors of the K individuals are independent, we can express the required distribution:

$$P(G_M|G_O, H) \propto P(G_M, G_O|H) = P(G|H) = \prod_{i=1}^{K} P(G_i|H) \qquad (1)$$

Each individual's genotype vector, given the haplotypes, $P(G_i|H)$ is a Hidden Markov model with the hidden states G_i and the pair of haplotypes (that are being copied to form the genotypes) $(Z_i^{(1)}, Z_i^{(2)})$, where $Z_i^{(j)} = \{Z_{i1}^{(j)}, .., Z_{iL}^{(j)}\}$ for both alleles $j \in \{1, 2\}$ and each $Z_{il}^{(j)} \in \{1, .., N\}$. That is, $Z_{il}^{(j)}$ denotes which haplotype is observed at allele j and locus l for individual i. There are two alleles $j = 1$ and $j = 2$ for each locus and these are the alleles at opposite chromosomes for a specific locus (e.g. the G and T pointed out in figure 1 at page 6) that constitutes the genotype at that locus.

By applying the law of total probability, we get:

$$P(G_i|H) = \sum_{Z_i^{(1)}, Z_i^{(2)}} P(G_i|Z_i^{(1)}, Z_i^{(2)}, H) P(Z_i^{(1)}, Z_i^{(2)}|H)$$

5.4.1 Transition probabilities

The "hidden" haplotypes for individuals are constructed as combinations of the N different haplotypes in the HapMap pool, with transition probabilities (in the sub-model $P(Z_i^{(1)}, Z_i^{(2)}|H)$) parametrised by population recombination rates. Based on coalescent theory assuming a constant population size N_e (that of the sample to be imputed) and an ancestral set of haplotypes of size N (i.e. HapMap). Li and Stephens suggests the following model for recombination [10]:

Let r_l be the genetic distance between locus l and $l+1$, N_e be the effective population size and the scaled recombination rate is $\rho_l = 4N_e r_l$ [11]. The algorithm proposed by [9] uses the estimate $N_e = 11418$.

The number of recombinations X between two loci l and $l+1$ are assumed to follow a Poisson distribution $Po(\frac{\rho_l}{N})$. So, $P(X=0) = e^{-\frac{\rho_l}{N}} \frac{(\rho_l/N)^0}{0!} = e^{-\frac{\rho_l}{N}}$ and, $P(X \neq 0) = 1 - e^{-\frac{\rho_l}{N}}$.
Also, assume that $P(Z_{i(l+1)}^{(j)} = Z_{il}^{(j)}|X \neq 0) = \frac{1}{N}$. That is, if there is recombination between locus l and $l+1$ then, the loci $l+1$ is placed on any of the haplotypes with equal probability. So each $Z_i^j|H$ is a Markov chain with state space $\{1, .., N\}$, where the lengths of staying on the same haplotype is exponentially distributed (the length of stay at a haplotype is a function of genetic distance between the loci) and uniform transition probabilities over all the haplotypes.

It then follows that,

$$\begin{aligned}
P(Z_{i(l+1)}^{(j)} = Z_{il}^{(j)}) &= \sum_x P(Z_{i(l+1)}^{(j)} = Z_{il}^{(j)}|X=x) P(X=x) \\
&= P(X=0) P(Z_{i(l+1)}^{(j)} = Z_{il}^{(j)}|X=0) \\
&+ P(X \neq 0) P(Z_{i(l+1)}^{(j)} = Z_{il}^{(j)}|X \neq 0) \\
&= e^{-\frac{\rho_l}{N}} \cdot 1 + (1 - e^{-\frac{\rho_l}{N}}) \cdot \frac{1}{N} \\
&= e^{-\frac{\rho_l}{N}} + \frac{1 - e^{-\frac{\rho_l}{N}}}{N}
\end{aligned}$$

and,

$$
\begin{aligned}
P(Z^{(j)}_{i(l+1)} \neq Z^{(j)}_{il}) &= \sum_x P(Z^{(j)}_{i(l+1)} \neq Z^{(j)}_{il} | X = x) P(X = x) \\
&= P(X = 0) P(Z^{(j)}_{i(l+1)} \neq Z^{(j)}_{il} | X = 0) \\
&+ P(X \neq 0) P(Z^{(j)}_{i(l+1)} \neq Z^{(j)}_{il} | X \neq 0) \\
&= e^{\frac{-\rho_l}{N}} \cdot 0 + (1 - e^{-\frac{\rho_l}{N}})(1 - \frac{1}{N}) \\
&= (1 - e^{-\frac{\rho_l}{N}}) \cdot \frac{N-1}{N}
\end{aligned}
$$

Let us evaluate $P(Z^{(1)}_i, Z^{(2)}_i | H)$. Theinitial state (i.e at the first locus $l = 1$) of the Markov Chain is uniformly distributed of the N^2 possible haplotype states:
$P(Z^{(1)}_{i1}, Z^{(2)}_{i1} | H) = \frac{1}{N^2}$

The transition probabilities that the haplotype status changes from locus l to $l + 1$ (i.e. there is a recombination) is given by equation 2 [10].

$$
P(\{Z^{(1)}_{il}, Z^{(2)}_{il}\} \rightarrow \{Z^{(1)}_{i(l+1)}, Z^{(2)}_{i(l+1)}\} | H) = \begin{cases} (e^{-\frac{\rho_l}{N}} + \frac{1-e^{-\frac{\rho_l}{N}}}{N})^2 \text{ if A happens} \\ 2(e^{-\frac{\rho_l}{N}} + \frac{1-e^{-\frac{\rho_l}{N}}}{N})(\frac{N-1}{N}(1 - e^{-\frac{\rho_l}{N}})) \text{ if B happens} \\ (\frac{N-1}{N}(1 - e^{-\frac{\rho_l}{N}}))^2 \text{ if C happens} \end{cases} \tag{2}
$$

With events A,B, and C defined as follows:
A as the case $Z^{(j)}_{il} = Z^{(j)}_{i(l+1)}$ for both $j = \{1, 2\}$. I.e. there are no recombination between the loci (mirrored in the $e^{-\frac{\rho_l}{N}}$ expression) or either loci might have had multiple recombinations and come back to the original haplotype (which corresponds to the $\frac{1-e^{-\frac{\rho_l}{N}}}{N}$ expression).
B as the case $Z^{(j)}_{il} = Z^{(j)}_{i(l+1)}$ for $j = 1$ but not $j = 2$ or vice versa. I.e. there is a change in haplotypes between l and $l + 1$ for one allele (which is the $\frac{N-1}{N}(1 - e^{-\frac{\rho_l}{N}})$ expression), but no change for the other allele (mirrored in the $e^{-\frac{\rho_l}{N}} + \frac{1-e^{-\frac{\rho_l}{N}}}{N}$ expression).
C as the case $Z^{(j)}_{il} \neq Z^{(j)}_{i(l+1)}$ for both $j = \{1, 2\}$. I.e. both alleles change haplotypes between loci l and $l + 1$.

We express the chain of events as:

$$P(Z_i^{(1)}, Z_i^{(2)}|H) = P(Z_{i1}^{(1)}, Z_{i1}^{(2)}|H) \prod_{l=1}^{L-1} P(\{Z_{il}^{(1)}, Z_{il}^{(2)}\} \rightarrow \{Z_{i(l+1)}^{(1)}, Z_{i(l+1)}^{(2)}\}|H) \tag{3}$$

Coming back to Figure 3, the Z's correspond to the hidden states X's in the figure, H correponds to the outcomes of the states Y, the transition probabilities $a_{ij} = 1$ if $j = i + 1$ and 0 otherwise (as the estimation of Z is done in a set order) and the outcome probabilities b_{ij} are the probabilities in equation 3.

5.4.2 Mutation probabilities

We now write out $P(G_i|Z_i^{(1)}, Z_i^{(2)}, H)$. This term is mimicing the effects of mutations as the haplotypes are being copied, so the observed genotype will be close to the haplotypes copied but not always exact. Let $\theta = (\sum_{i=1}^{N-1} \frac{1}{i})^{-1}$ be the mutation parameter [10]. Then $\lambda = \frac{\theta}{2(\theta+N)}$ is the probability of a mutation [10]. Moreover, the number of mutations at a SNP is $\sim Bin(2, \lambda)$. So the probability of there being no mutations is $(1-\lambda)^2$, one mutation at either allele is $2\lambda(1-\lambda)$ and two mutations at a SNP is λ^2. Let the following Table 4 describe $P((H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l}) \rightarrow G_{il})$

		G_{il}		
		0	1	2
$H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l}$	0	$(1-\lambda)^2$	$2\lambda(1-\lambda)$	λ^2
	1	$\lambda(1-\lambda)$	$\lambda^2 + (1-\lambda)^2$	$\lambda(1-\lambda)$
	2	λ^2	$2\lambda(1-\lambda)$	$(1-\lambda)^2$

Table 4: Probability table for the haplotypes → genotype

The rows where the haplotypes sum to 0 and 2 are straight-forward (as the quantity of mutation are $\sim Bin(2, \lambda)$). Though the row where the haplotypes sum to 1 is explained as follows: As the haplotypes sum to 1 we have two different alleles in the haplotypes. Assume that the first allele is 0 (0-allele) and the second is 1 (1-allele). If only the first allele mutates, then the genotype is read as 2 (as the allele mutates from the 0-allele to the 1-allele, you will have two 1-alleles) but if instead only the second allele mutates, then the genotype is read as 0 (as this allele mutates from the 1-allele to the 0-allele, you will have two 0-alleles). So these two scenarios require a certain allele to

mutate, giving both scenarios the probability $\lambda(1-\lambda)$. Now, there are two scenarios for the genotype being read as 1, given that the haplotypes sum to 1. Either no mutation occured or both alleles mutate. So the probability for this scenario is $\lambda^2 + (1-\lambda)^2$.
We end up with:

$$P(G_i|Z_i^{(1)}, Z_i^{(2)}, H) = \prod_{l=1}^{L} P(G_i|Z_{il}^{(1)}, Z_{il}^{(2)}, H) = \prod_{l=1}^{L} P((H_{Z_{il}^{(1)}l} + H_{Z_{il}^{(2)}l}) \rightarrow G_{il})$$

$P(G_i|Z_i^{(1)}, Z_i^{(2)}, H)$ also constitutes a HMM. Refering to Figure 3, the genotypes G are the hidden states X in the figure, while the haplotype indicators Z are the outcomes Y, the transition probabilities $a_{ij} = 1$ if $j = i + 1$ and 0 otherwise (because we estimate the genotypes in a set order) and table 4 shows the outcome probabilities b in the figure.

5.5 Methods for validating imputations

Imputing is perfomed based on Bayes formula for $P(G_M|G_O, H)$ (see (1) on page 12). For each SNP we end up with a probability distribution across genotypes for each individual. To measure the imputations success a validation procedure can be used to evaluate the probability distributions. We consider a subset of the known SNPs as unknown and evaluate how well the imputed distributions for the SNPs correspond to the genotyped values.

5.5.1 PRESS

To quantify the prediction error for the imputed SNPs, we use the prediction error sum of squares (PRESS). Denote $Y = \{y_{km}\}$ as the known genotypes for a set of SNPs for n individuals with k denoting the SNP and $m \in \{0, .., n\}$. Consider a subset of these SNPs $Y_U = \{y_{im}\}$ as unknown, with $\{i\} \subset \{k\}$. Impute the unknown SNPs in Y_U with the help of $Y \setminus Y_U$ (the set of Y not included in Y_U). The imputations give estimated probabilities p_{jkm} (where $j \in \{0, 1, 2\}$) for the values 0, 1 or 2 for each SNP.

We will evaluate the statistic:

$$PRESS_m = \sum_i \sum_{j=0}^{2} p_{jim}(y_{im} - j)^2$$

for each individual m, and the prediction error for each SNP i:

$$PRESS_i = \sum_m \sum_{j=0}^{2} p_{jim}(y_{im} - j)^2$$

High posterior probabilities far from the true genotype contribute to high values of PRESS. Good imputation quality implies a low PRESS for individuals and SNPs, respectively.

5.5.2 RMSEP

$RMSEP$ stands for root mean square error of prediction and is defined as $\frac{PRESS}{n}$, where n is size of the sets of SNPs $\{i\}$ or individuals $\{m\}$. The scaling of the $PRESS$ statistic is needed in order to make a fair comparison between HapMap phase 2 and 3, since the two phases differ in the number of SNPs (see Table 9 at page 24).

5.6 Imputation of CAPS and CAHRES

In this thesis the missing data occur because specific genotyping chips are set to read the genome at predetermined SNPs, thus the missingness is not expected to be informative for the genotype-disease association. The documented correlation structures across the genome (haplotype structure), i.e. the correlation structure between the missing and the observed SNPs will provide the basis for performing imputations.

When imputing for CAPS and CAHRES, we need haplotypes from a reference population, which are matched to be geographically close and thus has a higher likelihood of similar SNP correlation structure (linkage disequilibrium) and allele frequences [12]. We use the CEU population from HapMap (i.e. Utah residents with Northern and Western European ancestry), which is the closest geographically matched population to the Swedish population and it is present in both phase 2 and phase 3 of HapMap.

The CAPS data was genotyped on two different versions of the Affymetrix chip [5]. When the CAPS samples were genotyped, Affymetrix used an unbiased and completely random SNP selection.

The CAHRES data were genotyped using an Illumina chip [6]. Illumina uses a selection of tagSNPs for genotyping. TagSNPs are selected SNPs which are known to have high correlation with other SNPs. By choosing a good selection of tagSNPs, one can hope for good genome coverage.

5.7 Validation of imputation on CAPS and CAHRES

Our purpose is to compare the prediction of SNPs based on data from the two platforms, so we have chosen to validate imputations only for the overlapping SNPs between the two control groups/genotyping platforms. By choosing the overlapping SNPs (for which the genotype is known), we expect a fair comparison between the two studies/genotyping platforms. Table 5 shows the overlap between the SNPs from the studies for each chromosome.

	CAHRES	CAPS	Overlap	CAHRES%	CAPS%
1	39105	34593	6770	17.31	19.57
2	42357	35971	7224	17.06	20.08
3	35301	29702	5919	16.77	19.93
4	31398	27383	5030	16.02	18.37
5	32421	28029	5455	16.83	19.46
6	34341	28727	5683	16.55	19.78
7	28342	23387	4593	16.21	19.64
8	29655	23915	4954	16.71	20.72
9	25276	20267	4347	17.20	21.45
10	27282	24960	4928	18.06	19.74
11	25543	23000	4495	17.60	19.54
12	25356	22034	4548	17.94	20.64
13	19518	16530	3182	16.30	19.25
14	17317	13864	3014	17.40	21.74
15	15542	12449	2744	17.66	22.04
16	15821	13267	2804	17.72	21.14
17	13571	10192	2319	17.09	22.75
18	15759	12815	2638	16.74	20.59
19	9083	6240	1478	16.27	23.69
20	13439	11069	2324	17.29	21.00
21	7825	6226	1313	16.78	21.09
22	7971	5663	1304	16.36	23.03

Table 5: Amount and percetages of SNPs overlapping between CAHRES and CAPS

The validation for each study is made so that we exclude the overlapping SNPs and impute them by using the correlation structures from HapMap on the non-overlapping SNPs. Subsequently, we compare the imputed genotype distributions for the overlapping SNPs with the real genotypes for those SNPs using the RMSEP statistics.

6 Results

The analyses were carried out for three chromosomes, number 8, 17 and 21. Chromosomes 8, 17 were selected due to being repeatedly found associated to various cancers, while chromosome 21 (which is the smallest, non sex chromosome) was chosen as a test chromosome. Table 6 shows the mean RMSEP over SNPs and over individuals (they are identical by definition) and *mean* standard deviation (SD) also over SNPs and individuals for chromosome 8 for both studies and HapMap phases.

	Mean RMSEP	SNP SD	Individual SD
CAHRES phase 2	0.0244	0.0358	0.0029
CAHRES phase 3	0.0231	0.0353	0.0030
CAPS phase 2	0.0558	0.0587	0.0354
CAPS phase 3	0.0587	0.0607	0.0350

Table 6: Chromosome 08 RMSEP means and standard deviations (SD) for the individual and SNP means with HapMap phase 2 and 3

Judging by this table it seems that for the individuals in the CAHRES study the imputation does better than for CAPS (i.e. the RMSEP is generally lower and varies less for CAHRES than CAPS), both for each SNP and each individual. It also appears that for the CAHRES controls, the HapMap phase 3 CEU population seems to be slightly better as a reference population than the corresponding CEU population for phase 2 of HapMap, while the opposite holds for the CAPS controls. In conclusion, neither HapMap phase is consistently superior to the other. The standard deviations (SD) of the mean RMSEP for individuals is smaller than the SDs for SNPs, which indicates that the imputation accuracy is less stable for single SNPs than for single individuals especially with individuals from the CAHRES study (as the SNP SD $\approx 10\times$Individual SD). This is also the case for the CAPS study, though not as convincing. These results are replicated in all of the three tested chromosomes.

Figure 4 shows histograms over the RMSEPs for individuals (to the right) SNPs (to the left), where the CAPS validation with HapMap phase 2 and 3 are placed on the first and second row respectively and the CAHRES validation with HapMap phase 2 and 3 are placed on the third and fourth row respectively. The histograms make a very convincing case for that the missing SNPs in the CAHRES population are imputed better than the same missing SNPs in the CAPS population.

Figure 5 shows the RMSEP correlation for the two HapMap phases and the two studies. The RMSEP plotted in the phase 2 vs 3 plots are for each individual, as only the individual validation is comparable between HapMap phase 2 and 3 within a study. Whilst in the CAPS vs CAHRES plots, it

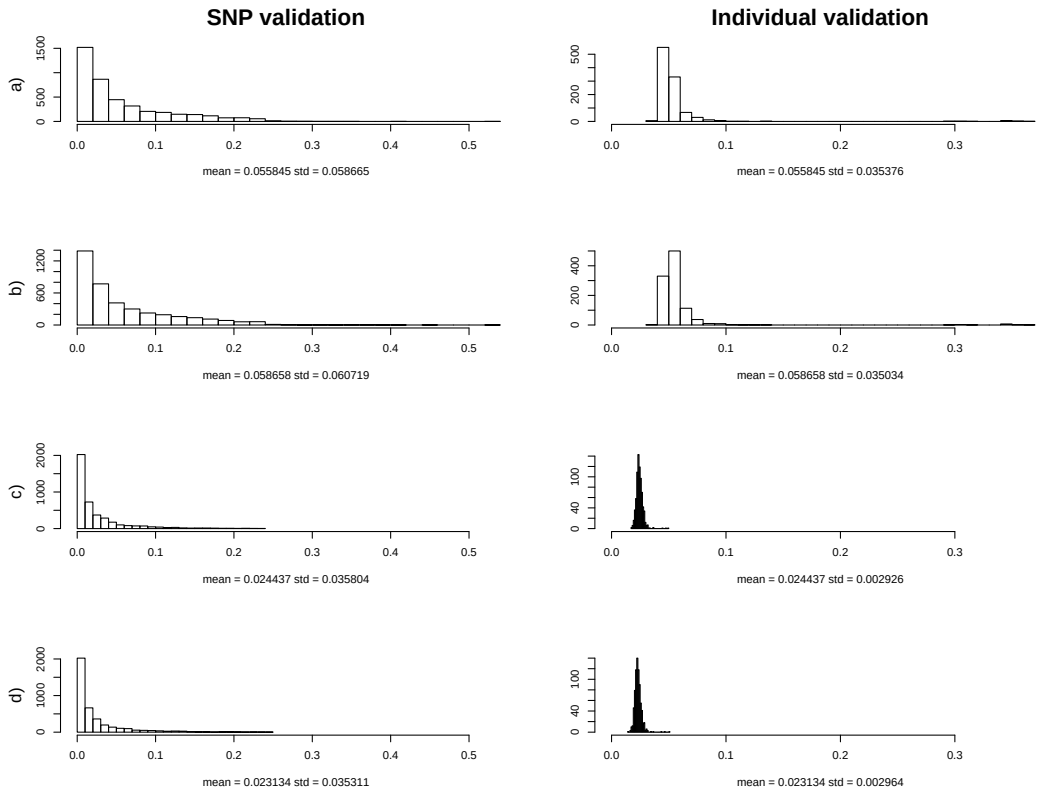

Figure 4: Validation histograms of RMSEP for chromosome 8. SNP validation (left) and individual validation (right). (a) CAPS HapMap phase 2; (b) CAPS HapMap phase 3; (c) CAHRES HapMap phase 2; (d) CAHRES HapMap phase 3

is the RMSEP for each SNP that is plotted, as we have studied the same SNPs in both studies. The correlation between the individual RMSEPs for phase 2 and phase 3 of HapMap is high for both studies, meaning that if an individual has been imputed bad for one phase, he/she will be imputed bad for the other phase as well. Comparing the studies and their respective SNP score vs each other, shows once again that CAHRES generally impute better than CAPS as most SNPs show a worse RMSEP in CAPS compared to their CAHRES counterpart.

A complementary question that we were interested in was whether certain genotypes are harder to predict than others. Tables 7 and 8 show the distribution over how the imputed data fits the real data for CAPS and CAHRES. The tables present the real genotype vs the imputed (predicted) genotypes. For example in Table 8 the SNPs with the AA genotype have an average of

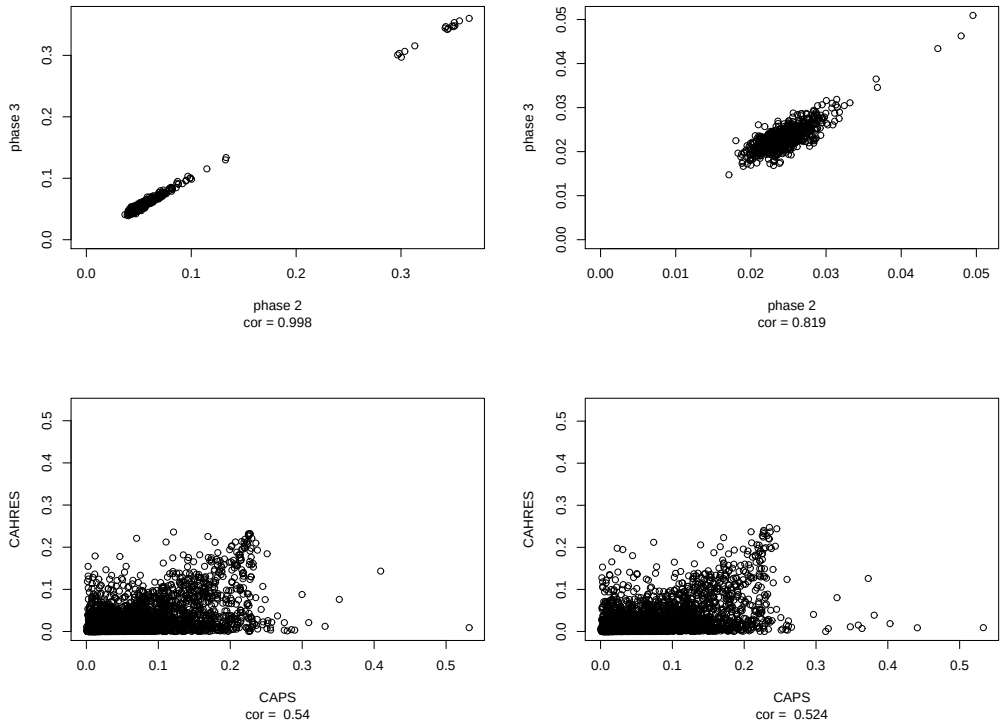

Figure 5: Correlation overview of the RMSEP values for chromosome 8. (top row) Individual RMSEP with HapMap phase 2 vs 3 for CAPS (left) and CAHRES (right); (bottom row) SNP RMSEP with CAPS vs CAHRES for HapMap phase 2 (left) and phase 3 (right)

0.925 for AA, 0.066 for Aa and 0.009 for aa in the imputed genotype probability distribution.

In general it seems like it is slightly harder to impute SNPs with a heterozygote genotype (Aa) than SNPs with either of the homozygote genotypes (AA and aa). It also turns out that this a general result for all chromosomes and HapMap phases in our study.

As seen in all of these analysises, CAHRES impute better than CAPS on basically every level. However, it is very hard to know what this is attributed to.

	Predicted AA	Predicted Aa	Predicted aa
Real AA	0.966	0.032	0.002
Real Aa	0.033	0.936	0.031
Real aa	0.002	0.032	0.966

Table 7: Average prediction distribution table for CAHRES chromosome 8 with HapMap phase 2

	Predicted AA	Predicted Aa	Predicted aa
Real AA	0.925	0.066	0.009
Real Aa	0.075	0.851	0.074
Real aa	0.008	0.067	0.925

Table 8: Average prediction distribution table for CAPS chromosome 8 with HapMap phase 2

7 Discussion

For assessing the accuracy of the imputations we used the overlapping SNPs genotyped in both CAPS and CAHRES data. These SNPs were temporarily treated as missing and then imputed. A more thorough comparison would be to have partitioned the validation so that one would not exclude all the SNPs at once. The optimal solution would have been to leave one out at a time, but computing time consideration made this impossible for this thesis. The reason for why it might not be fair to exclude all the overlapping SNPs at once, is that the imputation is affected by how many genotyped SNPs are close to the SNP that is going to be estimated. Basically, with more surrounding directly genotyped SNPs, one expects better imputation accuracy. Thus, if the overlapping SNPs are more clustered for one study compared to the other, then we are not making a fair comparison. Another issue could be if the percentage of overlapping SNPs for the studies differs a lot. Say that the excluded SNPs make up 10% in one study and 20% in the other, then the imputation results might be drastically worse for the study which had 20% of the SNPs excluded compared to the study where only 10% of SNPs were dropped. The overlapping percentages and the quantity of SNPs differs between the CAPS and CAHRES studies. It is hard to know how much this impacts the results. But if we look at chromosome 8, which we have used as our example in the results section, there are about 30% more SNPs in the CAHRES data set if we take away the overlapping SNPs ($\frac{29655-4954}{23915-4954} \approx 1.30$) so it is not surprising that these individuals have better imputed values.

Another problem with comparing the Illumina and Affymetrix chips is that it is hard to assess the impact of the SNP selection on these two chips on the

imputation quality. We have seen here that CAHRES were better imputed than CAPS. Does this imply that the Illumina SNP selection is better for imputations compared to Affymetrix? It might be that Affymetrix's genotyping platform is less suited for imputation, i.e. the capacity to predict untyped variation is not as good as for Illumina's chip. But it might also be that the CAPS study population resembles the reference population in HapMap (CEU) worse and therefore is harder to impute. Large geographical distances has been shown to mirror differenced in genomic allele frequencies and correlations. The question remains whether there is geographical population structure over smaller distances such as within Sweden and whether the CEU population of HapMap is better reflecting some regions of the country. The CAPS and CAHRES samples have slightly different geographical locations, and if the Swedish population exhibits a genetic population structure, this could affect the accuracy of the imputations.

In order to make a more thorough validation of HapMap 2 and 3 and the Illumina and Affymetrix chips, one could validate all chromosomes. Chromosomes might differ in terms of linkage disequilibrium patterns and SNP density or they might be impacted by specific differences (mutations, inversions, etc) between the sample and reference populations. Thus, even though the results in this study are remarkably similar for all 3 chromosomes, one might extend these investigations to the entire genome.

The statistic $PRESS_m$ that we use for the validation has given equal weight to all SNPs. It might be interesting to weight each SNP contribution by it's Minor Allele Frequency (MAF). It would also be interesting to first assess whether imputing accuracy varies systematically according to MAF.

Whilst imputing SNPs for a control pool with individuals typed on different platforms offers clear benifits (the possibility of increasing the number of controls and thus the statistical power), it is less clear whether, for a pool of controls with GWAS data from the same genotyping platform, there is a potential gain in power from imputing untyped HapMap SNPs to be included in testing genetic association. Published literature [9] claims that it may be useful to impute those parts of the genome with poor tagging (i.e. parts of the genome where no SNPs pick up the variation). So imputing SNPs for direct use in GWAS studies might have been useful a couple of years ago. Nowadays the genotyping chips from Illumina and Affymetrix genotype ~ 1 million SNPs so the tagging across the genome is very good with the modern genotyping chips. Thus studies genotyped with the newer chips will benifit less from imputations for GWAS.

8 Appendix

8.1 Computational considerations

The software "impute" [9] was used to impute the missing genotypes. Using a computer with a Intel(R) Xeon(R) CPU X7350 @ 2.93GHz processor and around 100 Gb of RAM, an imputation of chromosome 8 for 1000 individuals takes approximately 48 hours. "PLINK" [14] was used for GWAS data management.

8.2 HapMap phase comparison table

	New SNPs in phase 3		Deleted Phase 2 SNPs	
Chr	% new SNPs	# new SNPs	% deleted SNPs	# deleted SNPs
1	5.47	7,120	61.41	195,612
2	3.25	4,238	62.14	206,964
3	3.89	4,193	60.69	159,958
4	3.82	3,726	62.84	158,608
5	3.52	3,473	62.52	158,983
6	3.92	4,030	64.48	179,323
7	4.40	3,726	63.09	139,045
8	3.10	2,626	63.00	139,875
9	2.85	2,039	63.16	119,159
10	2.71	2,236	62.94	136,285
11	3.93	3,121	63.60	133,358
12	3.80	2,929	63.09	126,932
13	2.16	1,268	64.46	104,229
14	3.80	1,931	61.37	77,652
15	4.32	2,039	58.86	64,550
16	5.61	2,791	58.22	65,455
17	8.03	3,427	57.25	52,568
18	1.94	891	63.50	78,160
19	12.58	3,714	55.84	32,627
20	3.01	1,208	68.38	84,061
21	2.58	559	60.32	32,245
22	5.09	1,155	62.92	36,565
Total	4.11	66,418	62.42	2,482,214

Table 9: Chromosome wise comparison between phase 2 and 3 of HapMap

8.3 Results plots

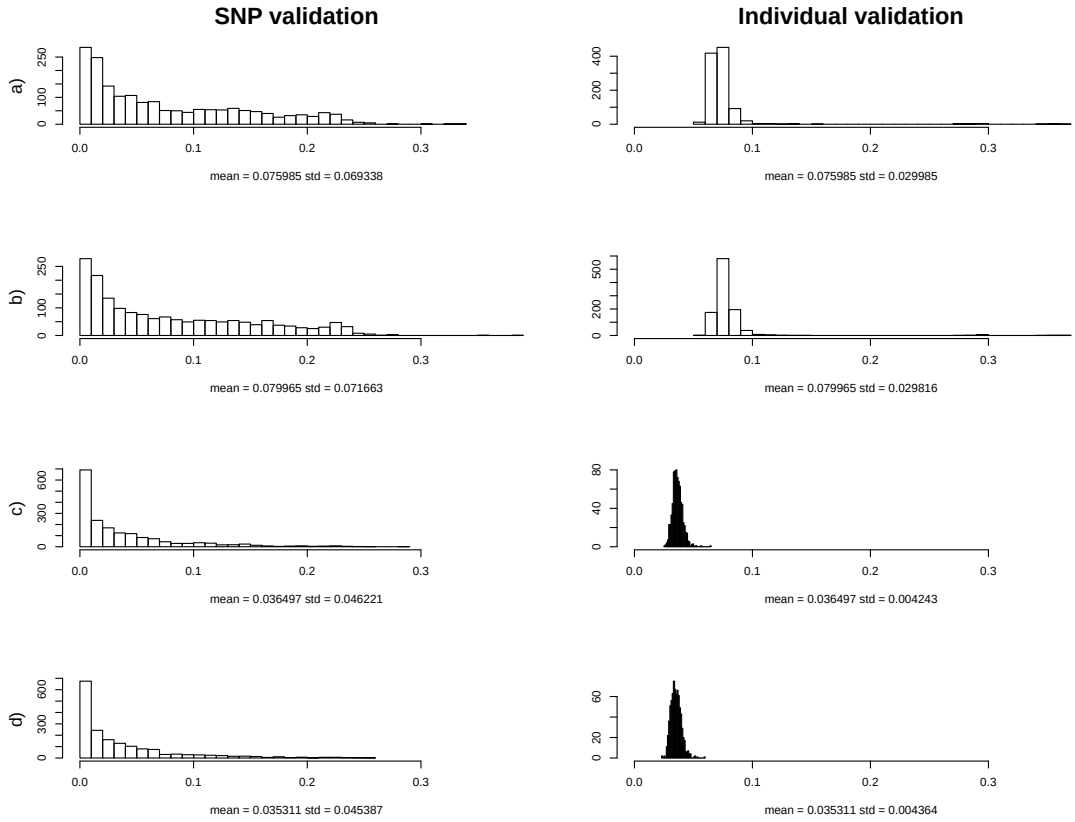

Figure 6: Validation histograms of RMSEP for chromosome 17. SNP validation (left) and individual validation (right). (a) CAPS HapMap phase 2; (b) CAPS HapMap phase 3; (c) CAHRES HapMap phase 2; (d) CAHRES HapMap phase 3

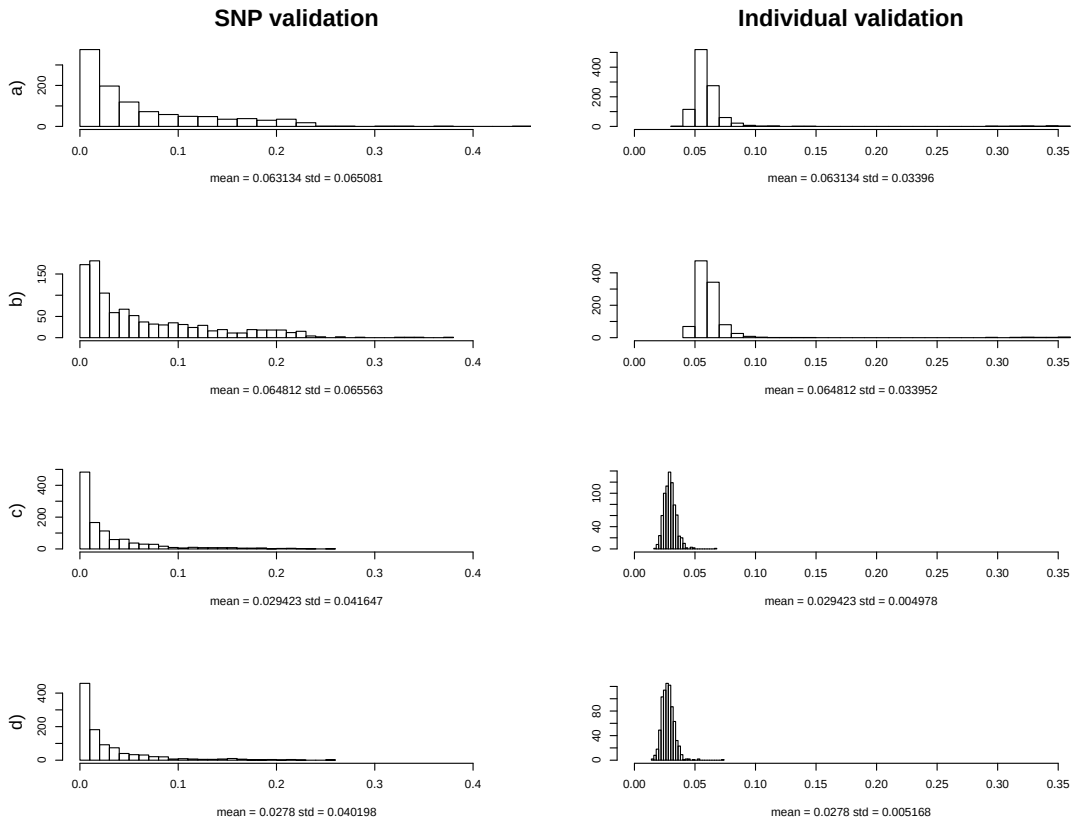

Figure 7: Validation histograms of RMSEP for chromosome 21. SNP validation (left) and individual validation (right). (a) CAPS HapMap phase 2; (b) CAPS HapMap phase 3; (c) CAHRES HapMap phase 2; (d) CAHRES HapMap phase 3

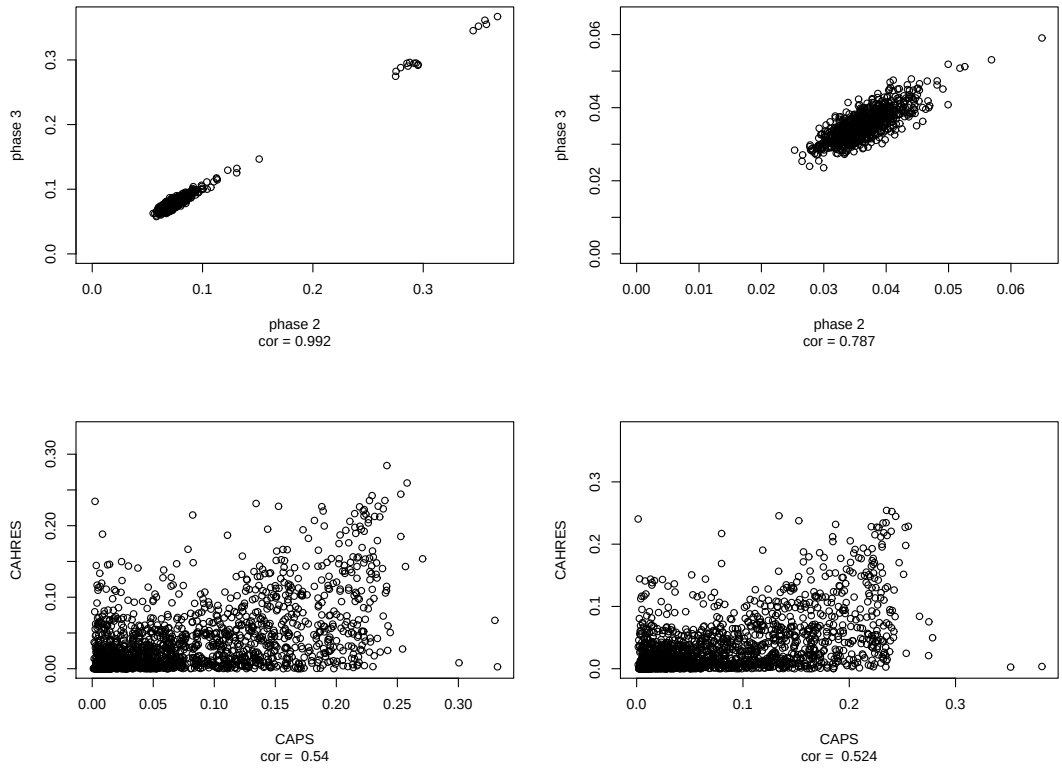

Figure 8: Correlation overview of the RMSEP values for chromosome 17. (top row) Individual RMSEP with HapMap phase 2 vs 3 for CAPS (left) and CAHRES (right); (bottom row) SNP RMSEP with CAPS vs CAHRES for HapMap phase 2 (left) and phase 3 (right)

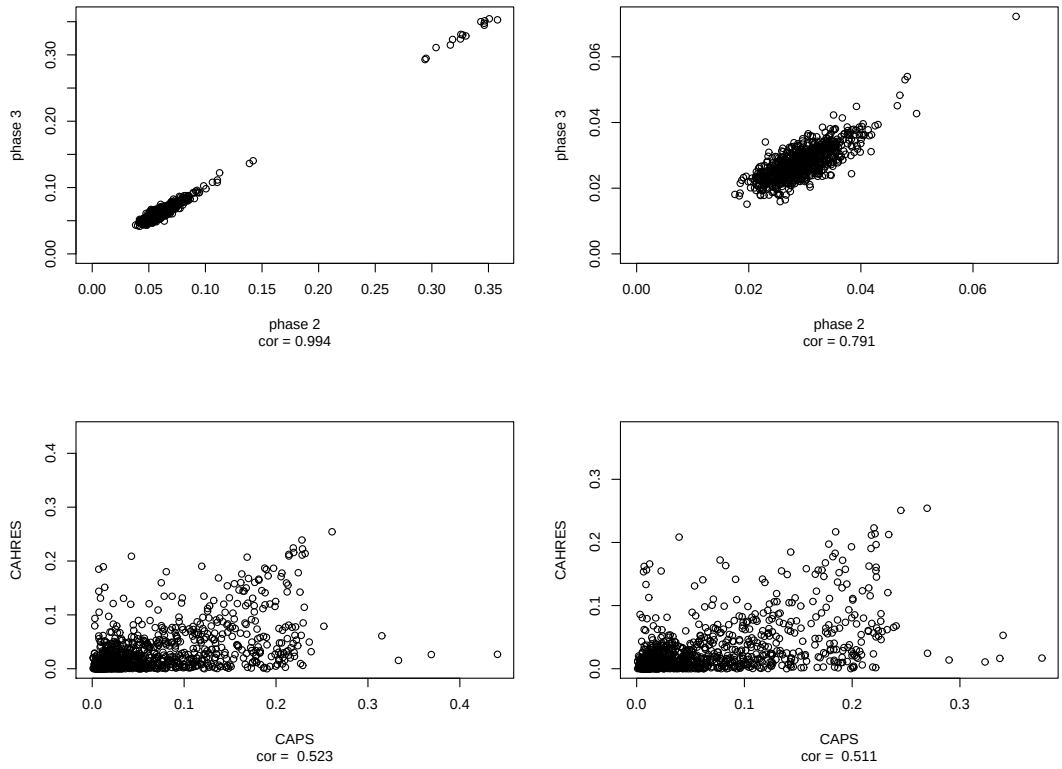

Figure 9: Correlation overview of the RMSEP values for chromosome 21. (top row) Individual RMSEP with HapMap phase 2 vs 3 for CAPS (left) and CAHRES (right); (bottom row) SNP RMSEP with CAPS vs CAHRES for HapMap phase 2 (left) and phase 3 (right)

9 Acknowledgements

This constitutes a master's thesis for 30 credits in mathematical statistics at Stockholm's University. It was written at the department of Medical Epidemiology and BioStatistics (MEB) at Karolinska Institute as a project regarding topics in missing genomic data, how to impute this incomplete data and evaluation of the imputations.

I would like to take the oppurtunity thank my supervisors, Juni Palmgren for discussion, formalizing the problem and having an overview of the project, Monica Leu for help on genetic, programming and data issues and Keith Humphreys for writing input, as well as statistical and genetics discussion.

Also a thank you goes to Ola Hössjer for valuable theoretical help and input, Alex Ploner for various programming help, Rolf Sundberg for problem formulations, Hatef Darabi for help on theoretical and programming issues and Arvid Sjölander for missing data discussion.

References

[1] Doris PA (Ferbruary 2002). Hypertension genetics, single nucleotide polymorphisms, and the common disease:common variant hypothesis *Hypertension* **39 Pt 2**; 323 - 331

[2] Risch N and Merikangas K (September 1996). The Future of Genetic Studies of Complex Human Diseases. *Science* **Vol. 273.**; 1516 - 1517

[3] Wellcome Trust Case Control Consortium (June 2007). Genome-wide association study of 14,000 cases of cases of seven common diseases and 3,000 shared controls. *Nature* **447**; 661-678

[4] The International HapMap Consortium (December 2003). The International HapMap Project. *Nature* **426**; 789 - 796

[5] http://www.affymetrix.com/

[6] http://illumina.com/

[7] Lindström S, Wiklund F, Adami H-O, Augustsson Balter K, Adolfsson J, and Grönberg H (November 2006). Germ-Line Genetic Variation in the Key Androgen-Regulating Genes Androgen Receptor, Cytochrome P450, and Steroid-5-a-Reductase Type 2 Is Important for Prostate Cancer Developmen *Cancer Research* **66**; 11077 - 11083

[8] Magnusson C, Baron J A., Correia N, Bergström R, Adami H-O and Persson I (November 1999). Breast-cancer risk following long-term ostrogen and oestrogen- progestin-replacement therapy *International Journal of Cancer* **Volume 81 Issue 3**; 339 - 344

[9] Marchini J, Howie B, Myers S, McVean G Donnelly P (July 2007). A new multipoint method for genome-wide association studies by imputation of genotypes *Nature Genetics* **39**; 906 - 913

[10] Li N and Stephens M (December 2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165(4)**; 2213-2233

[11] Hartl D L., Clark A G.(October 1997). Principles of Population Genetics. *Sinauer Associates* **3rd edition**; Chapter 7

[12] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko A R., Auton A, Indap A, King K S., Bergmann S, Nelson M R., Stephens M Bustamante C D. (November 2008). Genes mirror geography within Europe *Nature* **456**; 98 - 101

[13] Pardo L, Bochdanovits Z, de Geus E, Hottenga J J, Sullivan P, Posthuma D, Penninx B WJH, Boomsma D and Heutink P (January 2009). Global similarity with local differences in linkage disequilibrium between the Dutch and HapMap-CEU populations *European Journal of Human Genetics* **advance online publication**

[14] http://pngu.mgh.harvard.edu/∼purcell/plink/