A comparison of methods for haplotype inference

Elvira Hyvönen

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki
Finland

May 2013

A comparison of methods for haplotype inference

Master's Thesis

Author:

Elvira Hyvönen

Supervisors:

Mikko Sillanpää
Professor
Department of Biology and
Department of Mathematical
Sciences
University of Oulu

Petri Koistinen
University Lecturer
Department of Mathematics
and Statistics
University of Helsinki

Acknowledgements

The author gratefully acknowledges the major contributions of several colleagues to the author's learning process, analysis of the performance of the software and result presentation, in particular, A. Bouquet for help in understanding the biological and genetic background of the research question, M. Sillanpää and P. Koistinen for providing additional information on the statistical background of the subject, and T. Druet for the provided software and extremely quick and detailed comments on it.

To my parents who have always been very supportive.

Abstract

This study presents some of the available methods for haplotype reconstruction and evaluates the accuracy and efficiency of three different software programs that utilize these methods. The analysis is performed on the QTLMAS XII common dataset, which is publicly available. The program LinkPHASE 5+, rule-based software, considers pedigree information (deduction and linkage) only. HiddenPHASE is a likelihood-based software, which takes into account molecular information (linkage disequilibrium). The DualPHASE software combines both of the above mentioned methods. We will see how usage of different available sources of information as well as the shape of the data affects the haplotype inference.

Contents

Introduction

The genetic information of eukaryotic cells is stored on chromosomes located inside the nucleus. Chromosomes are long double-strand molecules of deoxyribonucleic acid (DNA), associated with proteins that fold and pack fine genetic information into a compact structure (see, e.g. Alberts *et al.*, 2008). Genes can be defined as functional units of heredity. Most of gene functions and locations are still unknown (Human Genome Project and Beyond, 2008), and therefore markers are useful tools to analyze the variation observed in the genome and relate it to phenotypes. A genetic marker is an observable genetically controlled variation that follows a Mendelian pattern of inheritance (Williams, 2005). A main issue when analyzing genome variation concerns the assessment of the diversity and frequencies of alleles in a population and their evolution over time.

Genetic markers can be classified in two main types depending on their informativeness, namely biallelic and multiallelic markers. Biallelic markers present only two alleles segregating at the marker locus. In the study of dairy cattle, scientists work mostly with biallelic markers represented by single nucleotide polymorphisms (SNPs) (Brookes, 1999). Single nucleotide polymorphisms are very frequent on the genome. Current estimates indicate that they occur every 200 base pairs on average (Williams, 2005). Although less informative than multiallelic markers, the very high density of SNPs along the genome offers geneticists the opportunity to track the transmission over generations of very fine chromosomal segments. This should allow improving the accuracy of mapping genes involved in disease or quantitative traits. Another advantage of using SNP markers compared to multiallelic markers is that cheap, fast and very reliable technologies have been developed for their detection.

However, modern genotyping methods do not allow obtaining haplotypes (ordered sequences of alleles on paternal / maternal chromosome) directly. Instead, data are collected routinely in large sequencing projects. Genotypes are obtained as a result of such sequencing. Hence, efficient, accurate and fast

methods are required for inferring haplotypes from genotypes. The methods can be divided into two main groups (cf. Gao *et al.*, 2009): pedigree-based and population-based algorithms. Pedigree-based algorithms reconstruct configurations by minimizing the total number of recombinants in the pedigree data. Population-based methods always assume Hardy-Weinberg equilibrium (Falconer & MacKay, 1996) at individual loci, and they assume linkage disequilibrium among markers.

In the present study both types of methods were studied using three different software programs. All programs were run on the QTLMAS common dataset and their results were compared in terms of accuracy and efficiency.

1 Biological background

The Earth is the planet of great life diversity. It is estimated that there are more that 10 million – perhaps 100 million – living species inhabiting it. All species are different and each reproduces itself so that the progeny belongs to the same species. Indeed, parents transmit genetic information, stored under the form of a deoxyribonucleic acid molecule (DNA), which specifies the characteristic the offspring shall have. This phenomenon of heredity is central to the definition and maintenance of life. Astonishingly, it has been proved during the last century that mechanisms linked to inheritance and gene expression are remarkably conserved between species. In this section we will give a brief description of the location and the structure of the genetic information, and the mechanisms involved in its transmission from generation to generation. In addition, basic concepts used by geneticists to model these processes will be discussed.

1.1 Location of the genetic information

The genetic information of eukaryotic cells is stored on chromosomes located inside the nucleus. Chromosomes are long double-strand molecules of deoxyribonucleic acid (DNA), associated with proteins that fold and pack fine genetic information into a compact structure. The complex of DNA and protein is also called chromatin.

Some eukaryote organisms only contain one set of chromosomes, they are said to be “haploid”. However, most of sexual organisms, such as mammals, are diploid, which means that they contain two copies of each chromosome (Lynch & Walsh, 1998). One copy is of paternal origin, the other of maternal origin. The paternally and maternally inherited chromosomes are defined as homologous. Some species, especially plants, are polypoid, *i.e.* they have genome containing more than two copies of chromosomes in their nucleus. However, those species are not under the scope of this review.

Diploid organisms have two types of chromosomes: sex chromosomes that contain genes participating in sex-deterministic mechanisms (XX and XY in mammals), and autosomes which are all the other chromosomes (Lynch & Walsh, 1998).

Different types of species have different number of chromosomes. For example, a human being has 24 pairs of chromosomes, a cow has 30; some plants might even have thousands of them.

1.2 Organization of the genome

Alberts *et al.* (2008) defines the genome as the totality of genetic information belonging to a cell or an organism. This term is used to describe DNA molecules that carry the information for all the proteins and RNA molecules that the organism will ever synthesize.

1.2.1 Support of the genetic information

DNA is the support and the storage of the genetic information within a cell and its carrier from generation to generation. A DNA molecule consists of two long polynucleotide chains composed of nucleotides (subunits). Each of the chains is known as DNA strand. Nucleotides are made up of two parts:

- a sugar phosphate molecule which allows linking nucleotides together in a chain to form the sugar-phosphate backbone of DNA;
- a base which is specific for each nucleotide.

Four different types of nucleotides exist – namely adenine (A), cytosine (C), guanine (G), thymine (T) – and differ by their base. Specific hydrogen bonds can be tied between the bases of pairs of nucleotides. Adenosine can only be paired with Thymine and Cytosine with Guanine. Those hydrogen bonds allow holding the two DNA strands together. As a result of the base-pairing requirements, both DNA strands are complementary. This property is fundamental for the replication

of DNA which corresponds to the duplication of the genetic information just before a cell division. (Alberts *et al.*, 2008; Deonier *et al.*, 2005).

Due to the chemical and structural properties of nucleotides, the 3-dimensional structure of the DNA molecule is a double helix which confers to it high stability and maximal packing efficiency of the genetic information.

The genome length is very variable between species: the human genome is estimated to contain about 3.2 billions of nucleotide pairs (or base pairs: bp) and the mouse genome about 2.6 billion bp (Human Genome Project and Beyond, 2008).

1.2.2 Organization of the genetic information

Chromosomes carry genes which could be defined as functional units of heredity. A gene is a segment of DNA that contains the instructions for making a particular protein (or a set of closely related proteins), a structural, a catalytic or a regulatory RNA molecule (see e.g., Alberts *et al*,. 2008). The location of a gene on a chromosome is called the locus.

The alternative forms of a gene at a locus are called alleles. Since DNA replication is not a perfect process, mutations arise, and as a consequence different versions of a gene can coexist in a population. Therefore, the two inherited "copies" of each gene carried by diploid individuals need not be identical.

Monomorphic loci are loci at which all gene copies are identical. Polymorphic loci exhibit more than one allele (Lynch & Walsh, 1998).

Eukaryotic genes contain exons, the coding sequences of the DNA, and introns which are non-coding sequences that separate the exons. The majority of genes consist of a long string of alternating exons and introns with most of the gene consisting of introns (Alberts *et al.*, 2008; Deonier *et al.*, 2005). According to estimations, genes comprise only about 2% of the human genome and the remainder consists of introns (Human Genome Project and Beyond, 2008). Besides, the human genome is estimated to contain between 25000 and 31000 protein-encoding genes (Baltimore, 2001).

The particular combination of alleles found in a specific individual is called the genotype. At a specific locus, genotype of a diploid organism consists of two alleles, one of which was inherited from the mother and another one from the father. A combination of alleles at different loci inherited as a unit form the mother or from the father is called the haplotype. The difference between a genotype and a haplotype is shown in Figure 1.

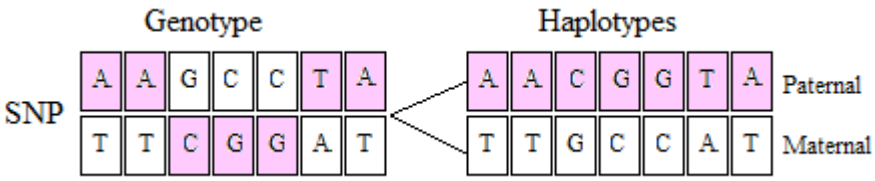


Figure 1: The difference between a genotype and a haplotype.

If the alleles in a locus are identical, the individual is called homozygote. If two different alleles were inherited from parents, the individual is called heterozygote (Lynch & Walsh, 1998).

1.2.3 Ways to mark genes

As most of gene functions and locations are still unknown, markers are useful tools to analyze the variation observed in the genome and to relate it to phenotypes. A main issue when analyzing genome variation concerns the assessment of the diversity and frequencies of alleles in a population and their evolution over time.

A genetic marker is an observable genetically controlled variation that follows a Mendelian pattern of inheritance (definition from Williams, 2005). In other words, markers are used to track inheritance of chromosomal regions in families which satisfy Mendelian laws formulated in 1866:

- the Law of Segregation. This law states that, for any particular trait, its two alleles (alternate versions of the same gene) separate so that only one is passed to the offspring. Which of the two alleles is inherited is left to chance, thus explaining variations within siblings;
- the Law of Independent Assortment. Put simply, this law states that traits are expressed independently of each other. For example,

inheritance and expression of the genes for blue eyes does not directly affect inheritance and expression of the gene for hair color (Sobel & Lange, 1996).

Given that relatives are more likely to carry similar alleles, a polymorphic marker is a very valuable tool to discriminate how related or distant some individuals can be. On the contrary, a monomorphic marker is not informative because all individuals in the population carry the same allele at this locus.

Correlating marker information with phenotypes (disease, quantitative performance for animals like milk production, meat production, etc.) expressed by individuals in a family or in a population allows locating genes involved in those traits relative to the marker positions.

Genetic markers can be classified in two main types depending on their informativeness, namely biallelic and multiallelic markers.

Multiallelic markers are the most informative ones because they present more than two different alleles segregating in the population. The most commonly used multiallelic markers are microsatellites. Microsatellites consist of the repetition of a small DNA sequence, the polymorphism residing in the number of repetition of this sequence. Most of microsatellite markers have generally between five and ten different alleles. Other multiallelic markers exist as reviewed by Williams (2005).

Biallelic markers present only two alleles segregating at the marker locus. Single nucleotide polymorphisms (SNPs) are the most commonly used biallelic markers, which are of the interest in this study. Brookes (1999) defines SNP as a single base pair position in DNA at which two different sequence alternatives (alleles) exist in individuals of some population(s). In the example illustrated by the Figure 2 sequences AAGC**C**TA and AAGC**T**TA differ by one nucleotide, and C / T (or G / A) are alternative alleles.

To avoid having poor accuracy in statistical inferences made from SNP information, SNPs are taken into account in genetic analyses if the least frequent allele has an abundance of 1% or greater.

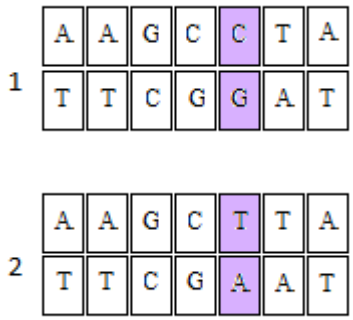


Figure 2: Single nucleotide polymorphism (SNP).

Single nucleotide polymorphisms are very frequent on the genome. Current estimates indicate that they occur every 200 bp on average (Williams, 2005). Although less informative than multiallelic markers, the very high density of SNPs along the genome offers geneticists the opportunity to track the transmission over generations of very fine chromosomal segments. This should allow improving the accuracy of mapping genes involved in disease or quantitative traits. Another advantage of using SNP markers compared to multiallelic markers is that cheap, fast and very reliable technologies have been developed for their detection.

1.3. Transmission of the genetic information across generations

The transmission of genetic information across generation is one of the main features for the survival of species.

1.3.1. Segregation of chromosomes during the meiosis

At the parental level each of the two diploid organisms contains N pairs of chromosomes in its cell nucleus ($2N$ chromosomes). During the reproductive cycle, the germline tissues produce haploid sex cells (gametes) with N chromosomes – ova from a female and spermatozoa from a male. Fusion of the gametes after mating produces a zygote that contains $2N$ chromosomes and is the start for a new diploid organism. The inheritance mechanism is schematically

shown in Figure 3. For more details see, for example, Alberts *et al.* (2008), on page 1090.

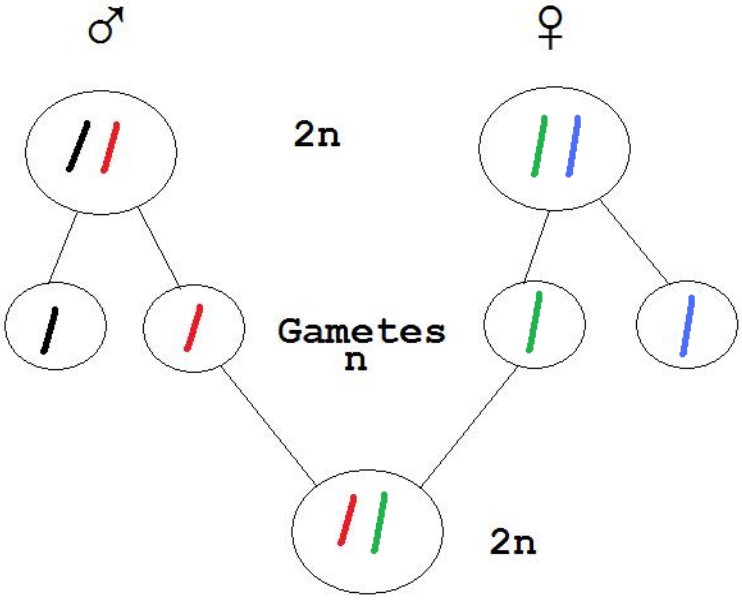


Figure 3: Schematic illustration of inheritance mechanisms without recombination.

The process of replication and reduction of chromosome numbers from $2N$ to N is called meiosis (Deonier *et al.*, 2005). The separation of the pairs of chromosomes ($2N$) by half allows forming two new cells (gametes), each of which contains N chromosomes with possibly different genetic information. This phenomenon is known as segregation of chromosomes.

1.3.2. Creation of gametes with new allele assortments through recombination

Recombination is a very important phenomenon that sometimes occurs during meiosis. It consists of the exchange of homologous DNA sequences between a pair of homologous chromosomes via the chromosomal crossover. Crossover occurs when homologous chromosomes overlap, break in the points of overlapping and then reconnect, but to the different end piece. As a result of recombination, the allele combinations found on chromosomes in gametes can be different from the combinations found in the parental chromosomes. The scheme of recombination process is shown in the Figure 4.

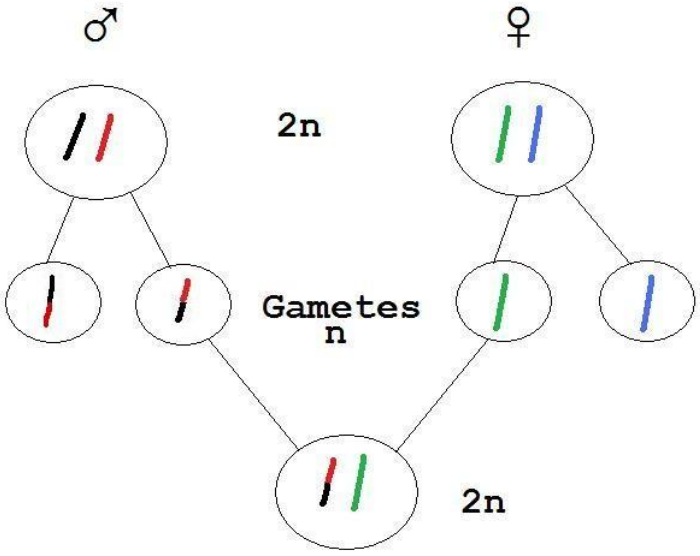


Figure 4: Recombination of the chromosomes in the cell during meiosis. Here, recombination occurred when producing paternal gametes.

Recombination occurs randomly and its rate varies from chromosome to chromosome and within different regions of any particular chromosome.

The further development of zygote is then carried out through a process of cellular multiplication called mitosis. Before the mitotic division of a cell, the DNA of each pair of chromosomes is replicated so that the two new cells formed also contain $2N$ chromosomes with the integrality of the genetic information contained in the original cell.

1.3.3. Mutation as further source of variation

Mutation is another source of variation. According to Alberts *et al.* (2008), mutation is a heritable change in the nucleotide sequence of a chromosome. It happens occasionally due to a failure of the cell's DNA-maintenance processes, which results in a permanent change in the DNA. Mutation can be lethal for an organism if it occurs in a vital position in the DNA sequence. The mutation rate – the rate at which observable changes occur in the genome – is approximately one nucleotide per 10^9 nucleotides each time the DNA is replicated. Non lethal mutations occurring during meiosis are transmitted to the next generations.

1.4. Basic concepts for modeling genetic information

1.4.1 Hardy – Weinberg equilibrium

The first methods to model the evolution of allele and genotype frequencies in a population were proposed at the beginning of the 20th century, *i.e.* a long time before the biological support of genetic information was discovered. Those models were developed supposing the theoretical framework in an ideal population. Such a population is required to be large, panmictic (random mating), not selected, closed (*i.e.* without migration), not affected by mutation and with discrete generations. In this case, effects of evolutionary processes can be ignored (Falconer & MacKay, 1996). Therefore, gene and genotypic frequencies at one locus reach an equilibrium state and remain constant across generations. This phenomenon was first demonstrated by Hardy and Weinberg in 1908 and is known as Hardy-Weinberg equilibrium. Furthermore, when the equilibrium state has been reached, allele frequencies p and q for a biallelic gene *A/a* at a specific locus can be deduced directly from observed genotypes and the following equation holds:

$$p^2 + 2pq + q^2 = 1,$$

where $p^2 = P(AA)$ is the frequency of homozygotes AA, $2pq = P(Aa)$ is the frequency of heterozygotes Aa, $q^2 = P(aa)$ is the frequency of homozygotes aa.

Although simple, this model enabled to analyse the genetic structure of natural populations whose demographic and reproduction parameters do not differ too much from assumptions of the ideal population in Hardy – Weinberg equilibrium. In such a population, equilibrium is reached after 1 generation, but when deviations from ideal population assumptions are larger, this equilibrium takes longer to reach. In most of genetic analyses concerning one locus or independent loci, it is often assumed that Hardy-Weinberg equilibrium has been reached.

1.4.2 Notion of linkage between markers

If two or more loci are located on the same chromosome, it is said that they are physically linked. This is a very important property as the probability that these loci are inherited together is quite high. Thus, a statistical dependence exists between loci located on a same chromosome (Lynch & Walsh, 1998). However, recombination occurring during meiosis may randomly break physical associations between alleles at different loci. As a recombination is more likely to happen between two remote than two close loci on a chromosome, the statistical dependence will decline with increasing distance between loci.

The statistical dependence between close loci, which is perceived as a nonrandom association of alleles at two or more loci, is called linkage disequilibrium (LD). If the association between alleles at different loci is random and no statistical dependence is observed between loci, then those loci are said to be in linkage equilibrium (LE) (Slatkin, 2008).

To illustrate this phenomenon, consider two biallelic loci A and B and let the frequencies of the four gamete types be $P_{A_1B_1}$, $P_{A_1B_2}$, $P_{A_2B_1}$ and $P_{A_2B_2}$, and let allele frequencies be p_{A_1}, p_{A_2}, p_{B_1} and p_{B_2} (example taken from Lynch and Walsh (1998)). In a situation of LE, *i.e.* if allele in locus A is independent of allele state in locus B, we expect that the probability of gamete $P_{A_1B_1} = p_{A_1} \times p_{B_1}$, etc.

Different factors like natural selection, founder effects, migration and assortative mating lead to a situation of linkage disequilibrium, when gamete frequencies depart from expectations based on allele frequencies. A natural measure of LD between loci is calculated by the formula:

$$D_{A_iB_j} = P_{A_iB_j} - p_{A_i} p_{B_j}.$$

This measure is often referred as coefficient of linkage disequilibrium or gametic phase disequilibrium.

The LD phenomenon can as well occur across generations (example adapted from Lynch and Walsh (1998)). Consider an ideal population of effectively

infinite size, in which mating occurs randomly and all of the forces causing alleles at different loci to become statistically associated are absent. In order to obtain gametes of a particular type, chromosomes either have to be transmitted across generations intact or they can recombine and create new gene combinations of that type. Let the frequency of gamete type A_iB_j in generation t be $P_{A_iB_j}$. Then $(1-c)P_{A_iB_j}(t)$ is the frequency that is passed on to the next generation without recombination. Let the proportion of the recombined gametes $p_{A_i}p_{B_j}$ be c. Due to independency of maternally derived A and paternally derived B caused by random mating in the population, c has to contain both A_i and B_j genes. Summing up two terms, we get:

$$P_{A_iB_j}(t+1)=(1-c)P_{A_iB_j}(t)+cp_{A_i}p_{B_j}.$$

After subtracting $p_{A_i}p_{B_j}$ from both sides:

$$D_{A_iB_j}(t+1)=(1-c)D_{A_iB_j}(t).$$

This equation generalizes to

$$D_{A_iB_j}(t)=(1-c)^t D_{A_iB_j}(0).$$

This means that the linkage disequilibrium decays gradually. Even in the case of unlinked genes ($c=0.5$) only 50% of disequilibrium is removed from each generation. If the recombination is less frequent, the time to attain linkage equilibrium ($D=0$) is very long.

The distance between markers is measured in centimorgans. A centimorgan (cM) corresponds to a recombination frequency of 1 %, which means that two markers or genes that appear together on the same chromosome are separated by recombination at a frequency of 0.01 during meiosis (Deonier *et al.*, 2005). If the physical distance between markers is large, the probability that they are recombined during meiosis increases and it is not possible to conclude, if they will be inherited at the same or at a different time.

In other words, in a situation of linkage disequilibrium, some patterns of inheritance are witnessed in the population and knowing the information in one locus it is possible to infer the information in another close locus.

1.5. The need of haplotype inference in animal breeding

Traditionally, the aim of the selective breeding is to improve the genetics of local populations of animals, which led to the development of animals with characteristic phenotypes that could be classified as distinct breeds. The diversity of phenotypes displayed by the various breeds is controlled by a broad genetic diversity, which provides the opportunity for the selection of animals with superior performance in specific desirable traits (for example, growth rate, hair color, milk production and disease resistance).

The use of selective breeding has resulted in dramatic improvements in simple production traits and the level of productivity from the selective improvement of livestock. However, the traditional approach based on selection of phenotypic qualities can lead to a narrowing of the genetic diversity is species, which reduces the genetic variation available for future selection, and also potentially concentrates genetic defects. In order to respond to public demand and develop a sustainable industry, it is necessary to address the potential problems associated with traditional selection approaches by fully exploiting the new technologies available for the selection of genetically superior animals.

To understand genetic variation among animals and to identify correlations between genetic variation and phenotypic variation it is necessary to understand the haplotype structures in populations, which would determine the best choice for the breeding animals.

2 Statistical methods of haplotype inference

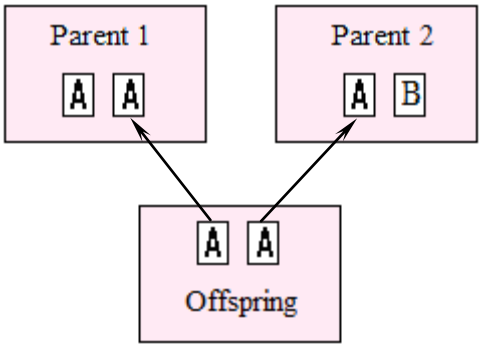
2.1 Pedigree based method

The main ideas of the pedigree-based method were described in Wijsman (1987). The method is based on the deduction rules, which are applied to data sequentially until no new loci can be phased. The logic behind them is the following (Sillanpää, 2004):

1. In case an offspring is homozygous (*AA*), then

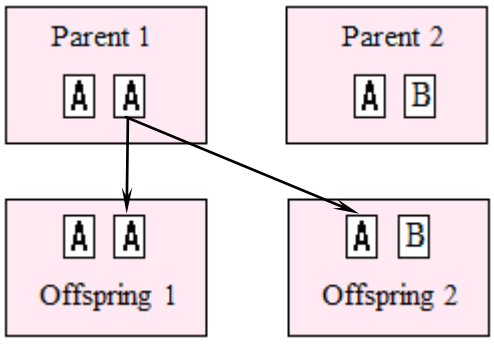
   1.1. The corresponding allele origins can be assigned at random.

   1.2. If there is any uncertainty in the parental genotypes, allele A is assigned to both parents with certainty.



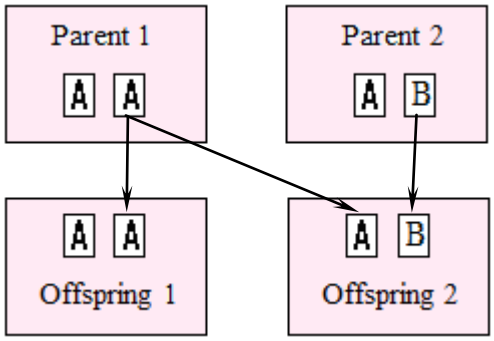
2. In case a parent is homozygous (*AA*),

   2.1. This parent will be the origin of allele *A* in all offspring genotypes.

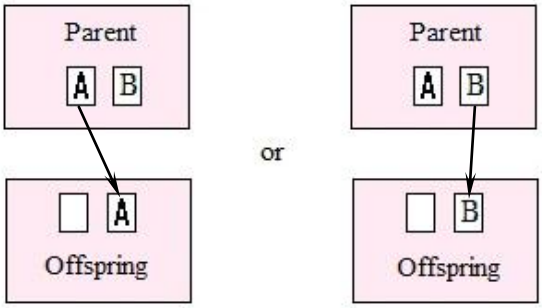


   2.2. If there is any uncertainty in the offspring genotypes, then allele *A* is assigned to all the offspring with certainty.

3. If an offspring allele *B* is not present in the known genotype of parent 1, then the origin of that allele is parent 2.



4. If an offspring has allele *A* with known origin, then the corresponding parent will have that allele in its genotype with certainty.
5. If an offspring has allele *A*, then both parents might have that allele.
6. If a parent is heterozygote (*AB*), then all the offspring will have one of the alleles *A* or *B*.



7. If there is a possibility for a parent to have allele *A*, then such a possibility exists for all the offspring as well.

After loci are phased according to the above rules, parental haplotypes are inferred using the recombination rate. The exact process is described in Druet and Georges (2010). In brief, the probability that marker allele 1 belongs to the "left" (L) homolog is computed as $P^L = L^L / \left(L^L + L^R\right)$ and to the "right" (R) homolog is computed as $P^R = L^R / \left(L^L + L^R\right)$. In these, L^L and L^R correspond to the likelihood that the marker allele belongs to the L (respectively R) haplotype of the parent, conditional on information from flanking anchoring markers. L^L and L^R are

computed respectively as $L_{UP}^{L} \times L_{DN}^{L}$ and $L_{UP}^{R} \times L_{DN}^{R}$, in which L_{UP}^{L} (L_{UP}^{R}) correspond to the likelihood conditional on information from anchoring markers located upstream (UP) (respectively downstream, DN) of the marker to be phased. L_{UP}^{L} (L_{UP}^{R}) and L_{DN}^{L} (L_{DN}^{R}) are computed as

$$\prod_{i=1}^{x}(1-r_i)\prod_{j=1}^{y} r_j,$$

where x (respectively y) is the number of offspring with marker phase in agreement (respectively disagreement) with the tested haplotype configuration of the examined parent and $r_i(r_j)$ is the recombination rate between the tested marker and the nearest (respectively upstream or downstream) anchoring marker informative in offspring i (respectively j).

Heterozygous markers that remain unphased in offspring can be further inferred conditionally on the known parental phase (Druet *et al.*, 2008). Using information from flanking markers phased in both parent and offspring, the probability that a marker allele belongs to the R (respectively L) homolog of the offspring can be computed as a function of intermarker recombination rates. A marker considered to be phased if one of the two possible configurations has a probability exceeding a chosen threshold.

2.2 Population-based method

Population-based statistical model allows describing patterns of genetic variation in samples of unrelated individuals. The model (Scheet & Stephens, 2006), which was implemented in the software for our study, is based on the idea that, over short regions, haplotypes in a population tend to cluster into groups of similar haplotypes. This clustering tends to be local in nature because due to recombination those haplotypes that are closely related to one another will vary as one moves along the chromosome.

The efficiency and effectiveness of this method is determined by the usage of the theory of hidden Markov models (HMM), described in Rabiner (1989) and Durbin *et al.* (1998). Observed haplotypes are modeled as mosaics of K hidden states (in other words, ancestral haplotypes), with K held constant throughout the genome. In the present study the number of hidden states is fixed and equals to 25.

Parameters of the HMM are (Druet & Georges, 2010):

- Population frequencies of each hidden state, which may differ between marker positions. The population frequencies of the hidden state at the first marker position define the initial-state probabilities.
- Hidden state-specific allele frequencies at each marker position, which define the emission (or observation) probabilities.
- Recombination rates for each marker interval.

Suppose that we observe unphased genotypes $g = (g_1, \ldots, g_n)$ on n diploid individuals. Let g_{im} denote the genotype at marker m in individual i, which will code as the sum of its alleles. We should now assume that the two haplotypes that make up each multilocus genotype are independent and identically distributed, which means that the population is in Hardy-Weinberg equilibrium. Under this assumption, if $z_{im}^{\bullet}$ denotes the unordered pair of clusters from which genotype g_{im} originates, then $z_i^{\bullet} = (z_{i1}^{\bullet}, \ldots, z_{iM}^{\bullet})$ form a Markov chain with initial state probabilities

$$p\left(z_{i1}^{\bullet} = \{k_1, k_2\}\right) = \begin{cases} \left(\alpha_{k_1 1}\right)^2, k_1 = k_2 \\ 2\alpha_{k_1 1}\alpha_{k_1 2}, k_1 \neq k_2 \end{cases}, \qquad (2.2.1)$$

where k_i is the number of cluster, that for simplicity is assumed to be known; α_{k_i} denotes the relative frequency of cluster k_i.

Transition probabilities are computed as a function of the recombination rate and the population frequencies of the different hidden states at the next marker position (2.2.2). As more than one allele may have nonzero frequency at a given marker position, hidden state in effect define clusters of related haplotypes.

$$p_m\left(\{k_1,k_2\}\to\{k_1',k_2'\}\right)= \\ =\begin{cases} p_m\left(k_1\to k_1'\right)p_m\left(k_2\to k_2'\right)+p_m\left(k_1\to k_2'\right)p_m\left(k_2\to k_1'\right), \\ \text{where } k_1\neq k_2 \text{ and } k_1'\neq k_2', \\ p_m\left(k_1\to k_1'\right)p_m\left(k_2\to k_2'\right), \text{otherwise,} \end{cases} \tag{2.2.2}$$

where $p_m\left(k\to k'\right)$ is defined by

$$p_m\left(k\to k'\right):=p\left(z_{im}=k'\middle|z_{i(m-1)}=k,\alpha,r\right):=\begin{cases} e^{-r_m d_m}+\left(1-e^{-r_m d_m}\right)\alpha_{k'm}, & k'=k \\ \left(1-e^{-r_m d_m}\right)\alpha_{k'm}, & k'\neq k \end{cases}$$

where for $m=2,\ldots M$, d_m is the known physical distance between markers $m-1$ and m, and $r=\left(r_2,\ldots r_M\right)$ and $\alpha=(\alpha_{km})$ are unknown parameters to be estimated.

Given the clusters of origins $z_i^{\bullet}$ it is again assumed that alleles are drawn independently from the relevant cluster allele frequencies:

$$p\left(g_i\middle|z_i^{\bullet},\theta\right)=\prod_{m=1}^{M}p\left(g_{im}\middle|z_{im}^{\bullet},\theta\right),$$

where

$$p\left(g_{im}\middle|z_{im}^{\bullet}=\{k_1,k_2\},\theta\right)=\begin{cases} \left(1-\theta_{k_1m}\right)\left(1-\theta_{k_2m}\right), & g_{im}=2 \\ \theta_{k_1m}\left(1-\theta_{k_2m}\right)+\theta_{k_2m}\left(1-\theta_{k_1m}\right), & g_{im}=3 \\ \theta_{k_1m}\theta_{k_2m}, & g_{im}=4 \end{cases} \tag{2.2.3}$$

Since $z_i^{\bullet}$ is unknown, the probability of g_i is obtained by summing over all the possible values:

$$p\left(g_i\middle|\alpha,\theta,r\right)=\sum_{z_i^{\bullet}}p\left(z_i^{\bullet}\middle|\alpha,r\right)p\left(g_i\middle|z_i^{\bullet},\theta\right), \tag{2.2.4}$$

where $p(z_i^{\bullet}|\alpha, r)$ is determined by equations (2.2.1) and (2.2.2). We assume that α can vary across markers, but is fixed across individuals.

HMM is applied in two stages:

1. The model is trained on a haploid set consisting of the partially phased base haplotypes. Parameter estimation $\vartheta = (\theta, \alpha, r)$ was done using the expectation-maximization (EM) algorithm (Stephens *et al.*, 2001; Dempster *et al.*, 1977).

2. The actual haplotype reconstruction and clustering by running a diploid HMM on the complete data set. The diploid HMM simultaneously models two independent chains, corresponding respectively to "left" and "right" homolog of the individual. The number of hides state combinations at each marker is $K_L \times K_R$, where K_L and K_R characterize "left" and "right" HMM, respectively.

At first, the diploid HMM is applied on the base individuals (genotyped individuals without genotyped parents). Second, the same HMM ($K_L = K_R = K$) was used to model "left" and "right" homolog with estimated EM parameters as estimated on the haploid training set. For each base individual the most likely hidden state composition of haplotypes is determined using the Viterbi algorithm (Forney, 1973).

Next, the descendent individuals were subsequently treated using a modified HMM. Haplotypes derived from genotyped parents were modeled as a mosaic of two parental haplotypes ($K = 2$), thus the number of hidden state combinations at a given marker position was 2×2 for individuals with two descendent haplotypes, and $2 \times K$ for individuals with one descendent and one base haplotype.

2.3 Combination of pedigree and population based method

This method is the combination of the previous two methods, which are iteratively applied one after another. At first, loci are phased based on the deduction rules, described in section 2.1, with the probability 1. At this point the modified diploid HMM applied on descendent individuals extracts linkage information only (offspring with two genotyped parents) or linkage and linkage disequilibrium information jointly (offspring with one genotyped parent). After that the population-based method (section 2.3.) is applied, assuming that all genotypes are independent and identically distributed, *i.e.* "population-wide" hidden state status of the base haplotypes is projected on their descendent haplotypes.

3 Comparison of methods of haplotype inference

3.1 Materials and methods

3.1.1 QTLMAS XII dataset

The simulation of the QTLMAS XII common dataset is described in detail in Lund *et al.* (2009) and Calus *et al.* (2009). This is a publicly available simulated dataset, which consists of 5865 individuals from seven generations. There are 6000 loci evenly distributed over six chromosomes (1000 markers per chromosome), with 0.1 cM between markers.

The dataset is provided in two files:

1. the phenotype file that contains six columns (Animal ID, sire ID, dam ID, Sex (male = 1, female = 2), Generation, Trait value);
2. the genotype file that contains the genotype of each animal in the pedigree, which is described in the phenotype file. The genotype file contains one line for each individual in the pedigree and 12001 columns (Animal ID; marker 1 allele 1 , marker 1 allele 2; marker 2 allele 1 , marker 2 allele 2; ... ; marker 6000 allele 1 , marker 6000 allele 2).

The data is haplotyped meaning that allele 1 of the biallelic marker comes from the father and allele 2 is inherited from the mother with 100% certainty. This information was used for the validation of the mistakenly phased loci while comparing the software outputs with the original dataset.

The amount of individuals per generation varies within the dataset. The ancestor generation (generation 0) consists of 165 individuals. Generations 1 – 3 are made up by 1500 individuals each. The last three generations 4 – 6 contain only 400 animals per generation meaning that not all of the animals from the generation 3 participated in the reproduction process, so selection took place.

3.1.2 Software for haplotype inference

Three different types of software for haplotyping (LinkPHASE 5+, HiddenPHASE, DualPHASE (Druet & Georges, 2010)) were tested. They are based on three different methods for haplotype inference. LinkPHASE 5+ utilizes pedigree-based method. It uses pedigree information (deduction and linkage) for inferring haplotypes. This software considers linkage information from sires with six or more offspring only. The software HiddenPHASE, which is based on theory of hidden Markov models (HMM) uses molecular information (linkage disequilibrium, LD) to reconstruct haplotypes and represents the second group (population-based method). Basically, HiddenPHASE is the implementation in the Fortran language of another commonly used method for haplotype reconstruction, which is called *fast*PHASE (Scheet & Stephens, 2006). Finally, DualPHASE utilizes both pedigree and molecular information.

3.1.3 Comparison criteria

The performance of the software for haplotyping was evaluated in terms of accuracy and efficiency. The accuracy was measured by heterozygous switch error rate and point-wise error rate. The efficiency was characterized by the elapsed time. Every reconstructed haplotype of each individual generated by the three types of software was compared to the correct haplotype provided by the original data.

The heterozygous switch error rate is the proportion of mistakenly inferred loci out of total number of heterozygous loci. Unphased loci were ignored during computation of this error.

The point-wise error rate is calculated allele-by-allele along each haplotype, yielding the overall score showing the difference between the generated haplotype and the original true haplotype taking each locus into account (Li & Li, 2007).

Heterozygous switch error and point-wise error rates were calculated separately for the obtained datasets of whole population, three last generations,

and two last generations, per generation in the whole population and per generation in the three last generations.

3.2 Aggregation of the data

The performance of the software was tested on three datasets, which were obtained from the original QTLMAS XII data. For the interest of the present study the original dataset was reduced from six chromosomes to the size of one chromosome (5865 individuals × 1000 markers), which was required by the design of the software. The subsets of 3 last generations (1200 individuals × 1000 markers) and 2 last generations (800 individuals × 1000 markers) were extracted from the dataset for one chromosome using R software (R Development Core Team, 2007). For the further evaluations the dataset of the whole population was divided into generations and six additional files were created: three generations 1 – 3 of the size (1500 individuals × 1000 markers), three generations 4 – 5 of the size (400 individuals × 1000 markers).

The aggregated data was obtained from the correct data for one chromosome. The process is shown step-by-step in the Figure 4. At first, the values of alleles of each marker were summed up creating aggregated genotypes. Next, markers were reconstructed according to the scheme: if the value of an aggregated genotype was 2, both allele values of the marker were set up as 1; if the aggregated sum was equal to 3 then the first allele became equal to 1 and the second allele became equal to 2; in case the number was 4, both allele values we set up as 2.

The file containing reconstructed genotypes had a fixed format: one line per individual with individual number that takes 6 positions (1 – 5865), two alleles, each takes 2 positions, per marker. Alleles were coded with 1 and 2. The subsets of 3 last generations (1200 individuals × 1000 markers) and 2 last generations (800 individuals × 1000 markers) were extracted from the reconstructed dataset.

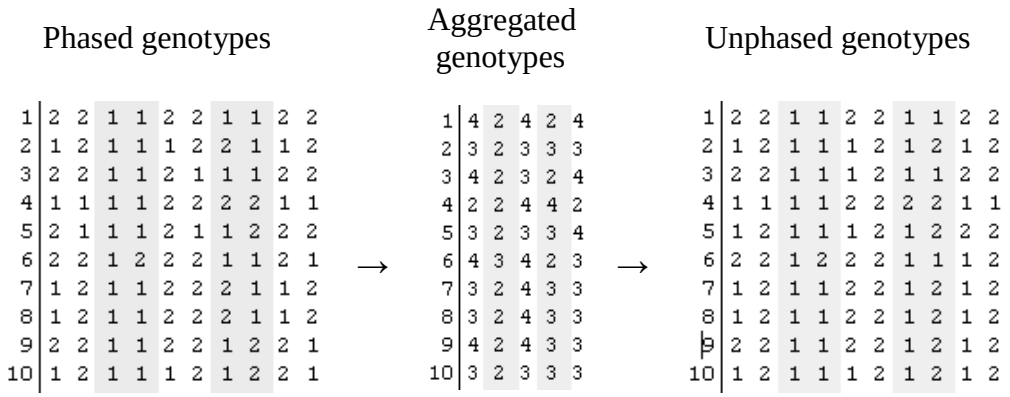


Figure 4: *Aggregation of the data step-by-step.*

The identity numbers in the original dataset contained some missing values. When that file was used as the input file all the types of software created some extra animals to fill-in the gaps. In order to avoid this, the identity numbers were modified so that they would correspond to the number of an animal in the dataset (1 – 5865).

Three additional input files required by the software were created manually:

Pedigree file contained the identity number of an animal, identity number of male parent (sire), identity number of female parent (dam). This file was obtained from the provided phenotype file by extracting the required information out of it. The total number of animals was 5685.

Marker file had a fixed format and was created according to the original data description: 1000 lines, one line per marker. Each line contained marker number (1 – 1000), marker name (1 – 1000) and marker position (0.1 – 100 cM with increment of 0.1 cM).

File with known haplotypes was created empty as no such information was available but the software requested for it.

After a run of each type of the software, output files containing phased loci were of the same format: two lines per individual with the same individual number (6 positions), origin of haplotype (1 – paternal, 2 – maternal), a space and then the haplotypes. The further described analysis was conducted for the output files of each type of the software.

To perform the analysis the file formats were modified. First, the identity numbers from the correct genotypes were removed by means of the R software. Second, the file format was modified to resemble the format of the output file with phases using Octave software (Eaton, 1997). As we dealt with biallelic markers, the correct haplotypes were obtained from genotypes in a way that each odd allele $(2n-1)$ belonged to paternal haplotype and even allele $(2n)$ belonged to maternal haplotype. The phases file was changed as well: the first two columns containing ID and origin of haplotype were removed using R software.

As software LinkPHASE 5+ was not able to phase all of the loci, it was important to calculate the amount of such gaps (marked as 0 in the output file). To be able to do that, the zeros were substituted with 9 to distinguish the unphased and incorrectly phased markers.

After the above preparations, modified matrix of the correct haplotypes was subtracted from the matrix with the output phases and the absolute value of the algebraic sum was taken. As a result a matrix, which contained elements 0 for the correctly phased allele, 1 for the heterozygous switches and 7 or 8 for the unphased allele, was obtained. These actions were applied to each of the three datasets (whole population, 3 generations and 2 generations).

The subtraction matrices for the whole population and 3 last generations were then divided using the R software by generations, creating nine new files (six generations in the whole population (three first of the size 1500 individuals × 1000 markers and three last of the size 400 individuals × 1000 markers) and three last generations of the size 400 individuals × 1000 markers). Generation number 0 was not included in the study of the error per generation as no pedigree information was available for it. Based on this result the further statistical evaluations were made separately for the obtained datasets (whole population, three last generations, two last generations, per generation in the whole population (excluding generation 0) and per generation in the three last generations).

3.3 Practical considerations and challenges

It is important to mention that the size of the data files was very large, that is why this caused additional challenges for analysis. For instance, the size of the original dataset with six generations was about 120 GB. This made it impossible to use the R software for the mathematical operations during the analysis, as it reserves the memory for the output, and in our case the memory to be reserved was supposed to be 120 GB that cannot be supported by most of the computers. On the contrary, the Octave software does the same computations iteratively, which makes the manipulations with such a big dataset feasible.

However, presenting the dataset as a data frame in the R software allowed doing such operations as, for example, ID number removal and extracting the subsets from the original data extremely quick, using only one command line. These operations would have been more awkward with the Octave software.

These reasons justify the use of two statistical software for performing computations on large datasets.

4 Results

4.1. Computing time

The elapsed time per analysis is presented in Table 1. It can clearly be seen that the LinkPHASE 5+ outperformed other software. HiddenPHASE was the slowest of all, though its running time is quite comparable with that of the program DualPHASE.

Table 1. ***Elapsed time.***

	Running time (sec)		
Software	Whole population	3 last generations	2 last generations
LinkPHASE 5+	60	9	5
HiddenPHASE	42780	18900	18900
DualPHASE	46560	15960	16920

These results show that the use of linkage, which is thought to reduce computing time, is not so efficient. Another aspect to consider is that the software LinkPHASE 5+ is used to phase only part of the loci, leaving quite a significant part of them undetermined. That is why DualPHASE is more preferable for haplotype inference to obtain complete haplotypes.

4.2. Point-wise error rates

Point-wise error rates are presented in Table 2, accounting for data of the whole population (generations 1 – 6), three last generations (generation 4 – 6) and two last generations (generations 5 – 6), respectively. To produce more accurate results the threshold value for LinkPHASE 5+ was set up to 1, meaning that haplotypes were inferred with probability 1 using deduction rules. This left a large number of loci unphased (1411282, 553004, 385672 respectively) and contributed

to the error rate, which is about two times higher than that for the other programs. The number of unphased loci was included in the computation of the error rate.

Table 2. ***Point-wise error rate for the whole population, three last generations and two last generations.***

Software	Whole population	3 last generations	2 last generations
LinkPHASE 5+ *	0.120314	0.230418	0.241045
HiddenPHASE	0.044668	0.187369	0.183619
DualPHASE **	0.043871	0.184635	0.185074

* Threshold = 1; ** Threshold = 0.

Point-wise error rates were also estimated for each generation in the population (Table 3), when whole population was genotyped. The increase in the error rates for the generations 4 – 6 is explained by the structure of the data, *i.e.* low number of offspring per parent in each on the last three generations.

Table 3. ***Point-wise error rate per generation in the whole population, whole population has been genotyped.***

Software	Generation 1	Generation 2	Generation 3	Generation 4	Generation 5	Generation 6
LinkPHASE 5+ *	0.076678	0.081045	0.081045	0.246605	0.267930	0.234552
HiddenPHASE	0.003081	0.001853	0.001853	0.166580	0.201763	0.180391
DualPHASE **	0.000139	0.000240	0.003249	0.165620	0.194075	0.180311

* Threshold = 1; ** Threshold = 0.

The situation when the data are obtained only for two or three last generations in the population occurs very frequently in the real life. This was the reason for including such a case in the present study. Point-wise error rates estimated on the dataset consisting of the last three generations are summed up in the Table 4.

Table 4. ***Point-wise error rate per generation in the last three generations, only three last generations have been genotyped.***

Software	Generation 4	Generation 5	Generation 6
LinkPHASE 5+ *	0.243150	0.230767	0.217337
HiddenPHASE	0.084621	0.003060	0.003185
DualPHASE **	0.192447	0.184986	0.176471

* Threshold = 1; ** Threshold = 0.

4.3. Heterozygous switch error rates

Heterozygous switch error rates are presented in Table 5 accounting for data of the whole population (generations 1 – 6), three last generations (generation 4 – 6) and two last generations (generations 5 – 6), respectively.

Table 5. ***Heterozygous switch error rate for the whole population, three last generations and two last generations.***

Software	Whole population	3 last generations	2 last generations
LinkPHASE 5+ *	0.068755	0.001358	0.000416
HiddenPHASE	0.089328	0.003704	0.003719
DualPHASE **	0.087733	0.006497	0.000482

* Threshold = 0; ** Threshold = 0.

To compare the error rate of each software, threshold was set up 0 for LinkPHASE 5+ in order to phase the maximum number of loci. The number of the unphased loci became fixed when the threshold was around 0 and this indicated the absence of the pedigree information for those loci (108980, 27339 and 39836, respectively). For DualPHASE threshold was chosen as 0 meaning that the probability of double recombination was bigger than 0 (the detailed instructions are provided in the software manual).

Table 6 presents heterozygous switch error rate per generation in the whole population. The rates differ significantly between generations 1 – 3 and 4 – 6 due to the higher number of offspring per family in generations 1 – 3 (“bottle neck” shape of the data). Generations 4 – 6 were formed under selection process, meaning that not all of the animals of generation 3 participated in the reproduction process. It is a well-known fact that it is much easier to deduce the haplotype of a parent, which has a lot of offspring. The opposite does not hold. This causes the increase in the error rates in generations 4 – 6.

Table 6. ***Heterozygous switch error rate per generation in the whole population, whole population has been genotyped.***

Software	Generation 1	Generation 2	Generation 3	Generation 4	Generation 5	Generation 6
LinkPHASE 5+ *	0.000272	0.000416	0.001358	0.236535	0.294083	0.292038
HiddenPHASE	0.006137	0.003719	0.003704	0.331618	0.403721	0.362568
DualPHASE **	0.000276	0.000482	0.006497	0.329707	0.388338	0.362407

* Threshold = 0; ** Threshold = 0.

For this case DualPHASE proved to be the fastest and the most accurate software for haplotyping.

Heterozygous switch error rates per generation in the case of three last generations were computed and the summary is represented by the Table 7. Threshold values were chosen the same as in the previous case.

Table 7. ***Heterozygous switch error rate per generation in the last three generations, only three last generations have been genotyped.***

Software	Generation 4	Generation 5	Generation 6
LinkPHASE 5+ *	0.088248	0.096323	0.091867
HiddenPHASE	0.045136	0.001626	0.001692
DualPHASE **	0.102242	0.098278	0.093754

* Threshold = 0; ** Threshold = 0.

It can clearly be seen that the program HiddenPHASE outperformed the others, but the computation time required to run the program was more than five hours.

5 Discussion

The program LinkPHASE 5+, based on pedigree information, was not initially designed to phase all the loci, that is why the output files contain gaps (marked as 0). There are two criteria that have to be fulfilled to obtain the haplotypes using this program: the number of progeny is one of the thresholds that is a natural requirement (two or more progeny need to be informative to phase the marker) was set up inside the program; threshold for probability (between 0 and 1) that we set up manually during the computational process.

Different values for the threshold were tested for LinkPHASE 5+: 1, 0.95, 0.5, and 0 (the results are not shown). The value 1 was recommended by the author of the software. In this case program inferred haplotypes based on the deduction rules and phased loci with probability 1 utilizing the pedigree information, which led to the number of unphased loci was significantly larger (1411282, 553004 and 385672 for the whole population, three last generations and two last generations respectively) and contributed to the point-wise error rate. In the situation, when the threshold equals to 0 the probability to accept false haplotype remarkably rises as we ignore the recombination rate values, but in this case it is possible to detect the loci without origin or missing markers. The number of unphased loci becomes constant and does not change when the threshold value is equal or less than 0.5: 108980, 27339 and 39836 for the whole population, three last generations and two last generations respectively. These values were left out when computing the heterozygous switch error rate. Threshold was set up 0 to make the heterozygous switch error rate of the rule-based software more comparable with that of the other three programs.

Our results indicate that point-wise error rate produced by DualPHASE was smaller than that of the other types of software in the cases of the whole population, three last generations, two last generations and per generation in the whole population. However, point-wise error rate per generation in the three last generations (the closest case to the real situation geneticists deal with) produced by the software HiddenPHASE was the smallest. This can be explained by that

HiddenPHASE does not take into account pedigree information and treats all of the loci as independent. Thus the number of markers becomes significantly smaller compared to the size of the whole population containing six generations.

In general, for the small number of markers the software for haplotyping works rapidly and accurately. As the number and complexity of the system increases haplotype inferring becomes more laborious. This can also explain why the computation time was the same for the datasets of three and two generations.

Heterozygous switch error rate was computed for all the software outputs. However, LinkPHASE 5+ stands out from the rest of software and forms a separate group. It was designed to infer partial haplotypes. That is why the heterozygous switch error rate as well as the elapsed time for this software is hardly comparable.

The elapsed time of the DualPHASE in the cases of 3 and 2 generations was considerably smaller than that of the HiddenPHASE, taking into account that it utilizes both molecular and pedigree information. Computation took about an hour less compared to the time required by HiddenPHASE (about 5 hours 15 minutes). However, in the case of the whole population HiddenPhase performed faster. Our research clearly shows that doing haplotype inference with the help of DualPHASE and HiddenPHASE is highly time consuming, but brings quite accurate results. If the time is a significant factor that needs to be taken into account, it is recommended to use some other type of software, e.g. DAGphase (developed by the same author, T. Druet).

Hence, the results show that different types of the tested software can be used in different situations. In the situation, when a dataset is rather small, dense and the computational time makes little difference, HiddenPHASE or DualPHASE programs are more preferable to use as they are based on quantitative statistical methods, do not require any prior specification, and the error rates they produce are small. Software LinkPHASE 5+ should be used for preliminary analysis to infer partial haplotypes with probability 1 based on deduction rules to make the further computations go faster.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. 2008. Molecular Biology of the Cell. Fifth edition. New York, USA & Abingdon, UK: Garland Science, Taylor & Francis Group, LLC. pp. 1268.

Baltimore, D. 2001. Our genome unveiled. Nature 409: 814 – 816.

Brookes, A.J. 1999. The essence of SNPs. Gene 234: 177 – 186.

Calus, M.P.L., de Roos, A.V.P., Veerkamp, R.F. 2009. Estimating genomic breeding values from the QTLMAS Workshop data using single SNP regression and the haplotype/IBD approach. BMC Proceedings 3(Suppl 1): S10.

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1): 1 – 38.

Deonier, R.C., Tavaré, S. & Waterman, M.S. 2005. Computational Genome Analysis. An Introduction. New York, USA: Springer Science+Business Media, LCC. pp. 534.

Druet, T., Fritz, S., Boussaha, M, Ben-Jemaa, S., Giullaume, F., Derbala, D., Zelenika, D., Lechner, D., Charon, C., Boichard, D., Gut, I.G., Eggen, A. & Gautier, M. 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178: 2227 – 2235.

Druet, T. & Georges, M. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184: 789 – 798.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G.1998. Biological sequence analysis. Probabilistic models of proteins and nucleic acids. Cambridge, United Kingdom: Cambridge University Press. pp. 356.

Eaton, J.W. GNU Octave. 1997. A high-level interactive language for numerical computations. Edition 3 for Octave version 2.1.x. pp. 356.

Falconer, D.S. & Mackay, T.F.C. 1996. Introduction to Quantitative Genetics. Fourth edition. Essex, England: Longman Group Ltd. pp. 464.

Forney, Jr., G.D. 1973. The Viterbi algorithm. Proceedings of the IEEE 61(3): 268 – 278.

Gao, G., Allison, D.B. & Hoeschele, I. 2009. Haplotyping methods for pedigrees. Human Heredity 67: 248 – 266.

Human Genome Project and Beyond. 2008. http://www.ornl.gov/hgmis/publicat/primer. U.S. Department of Energy Genome Research Programs, visited 12.6.2010.

Li, X. & Li, J. 2007. Comparison of haplotyping methods using families and unrelated individuals on simulated rheumatoid data. BMC Proceedings I (Suppl 1): S55.

Lund, M.S., Sahana, G., de Koning, D.-J., Su, G. & Carlborg, Ö. 2009. Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC Proceeding 3 (Suppl 1): S1.

Lynch, M. & Walsh, B. 1998. Genetics and Analysis of Quantitative Traits. Sunderland, Massachusetts, USA: Sinauer Associates, Inc. pp. 980.

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 2 (77): 257 – 286.

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

Sillanpää, M.J. 2004. Multimapper / OUTBRED reference manual. Bayesian QTL mapping software for outbred offspring data. Version 1.1 / for a backcross and F2 – full-sib family. Rolf Nevanlinna Institute. pp. 18.

Slatkin, M. 2008. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9: 477 – 485.

Sobel, E. & Lange, K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. American Journal of Human Genetics 58(6): 1323 – 1337.

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

Wijsman, E.M. 1987. A deductive method of haplotype analysis in pedigrees. American Journal of Human Genetics 41: 356 – 373.

Williams, J.L. 2005. The use of marker-assisted selection in animal breeding and biotechnology. Revue Scientifique et Technique e l Office International des Epizooties 24 (1): 379 – 391