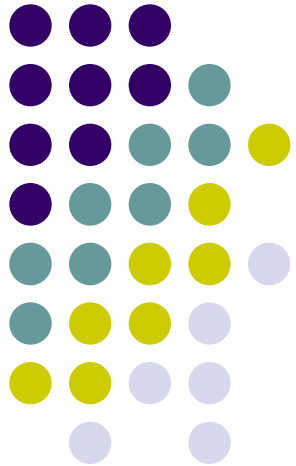
Synthèse bibliographique sur le calcul des proba d'IBD et la classification des haplotypes

Pilar Schneider
GdP-GABI



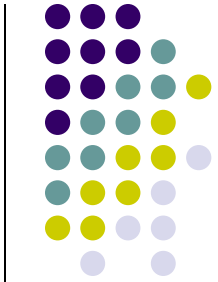
Rules & Tools Workshop

La Rochelle, 23-24 September 2010

Outline

- Introduction
- Summary of different methods used in human and animals genetics
- Comparison among methods
- Conclusions

Haplotypes

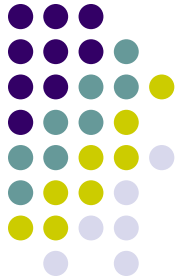


- Haplotypes can provide valuable information in the study of:
  - complex traits
  - diseases
  - population histories
  - evolutionary genetics

➔ current genotyping technologies are unable to resolve the phase of maternal and paternal chromosomes in unrelated individuals

Haplotype: *a sequence of alleles that are on the same physical chromosome (i.e. inherited from the same parent)*

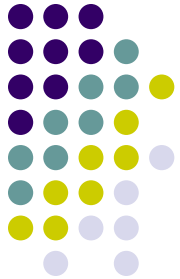
Haplotypes (cont.)



- Haplotypes analysis can provide improved power to detect associations between complex traits and densely spaced genetic markers

- Most methods for multilocus analysis that are suitable for whole-genome association data required phased haplotypes

Haplotype inference



- **Multinomial model for haplotype frequencies**:
  - No prior information about the haplotype frequency distribution
  - Expectation-Maximization (**EM**) algorithm are used to maximize the likelihood (Long et al., 1995; Hawley et al., 1995)

- **Coalescent-based Bayesian method**:
  - It can make predictions about the patterns of haplotypes to be expected in natural populations (**PHASE**, Stephens et al., 2001)
  - Takes into account similarities between and among haplotypes
  - It produces accurate results, but the application is limited for large data sets

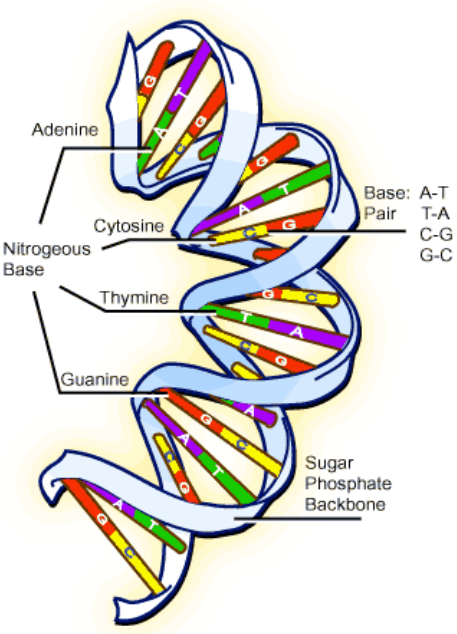
- **Haplotype blocks**
  - Haplotype blocks do not properly explain all the correlation structure between markers, because linkage disequilibrium (**LD**) can extend beyond block boundaries and can have complex patterns within block (Halperin and Eskin, 2004)

- **Methods based on Hidden Markov Model** (**HMM**)

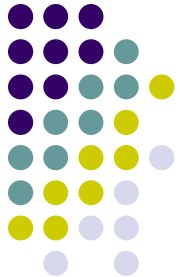
Papers reviewed

- **Human genetics** (5)
  - Unrelated populations, exploit LD information

- **Animal genetics** (3)
  - Exploit family (linkage) & LD information



Human genetics papers



1) To cluster haplotypes and to perform association analysis *(Li et al., 2006)*

2) To infer missing genotype data *(Marchini et al., 2007)*

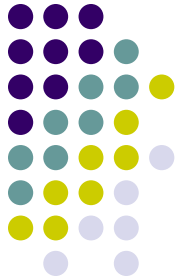
3) To infer haplotype phase and missing genotype data based on clustering algorithms

3.1) *Scheet and Stephens (2006)*

3.2) *Browning and Browning (2007)*

4) To perform an Imputation-based Bayesian regression analysis *(Servin and Stephens, 2007)*

Animal genetics papers

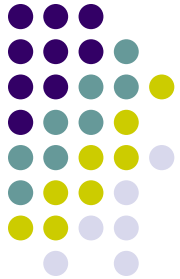


5) To estimate IBD probabilities using linkage (**LA**) and linkage disequilibrium (**LD**) information *(Meuwissen and Goddard, 2001)*

6) To cluster haplotypes based on IBD probabilities, and to perform a variance component analysis to map QTLs *(Druet et al., 2008)*

7) To simultaneously infer haplotypes and missing genotypes and cluster haplotypes combining LA and LD information, and to perform a variance component analysis to map QTLs *(Druet and Georges, 2010)*

Methods used in human genetics



- Genome-wide association studies (**GWAS**): scan the entire genome for variants that are associated with a trait or disease of interest

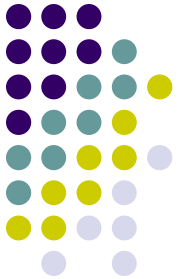
- To improve the power of GWAS different strategies can be adopted (Browning, 2008):

  1) To infer haplotype phase and use haplotype-based methods for association testing in addition to single-maker association

  2) To use missing data imputation to infer genotypes for known reference panels (*e.g.* HapMap)

  3) To combine results across multiple studies, imputing genotypes when SNPs have been genotyped in some, but not all the studies

1) Method to cluster haplotypes and to perform association analysis (Li et al., 2006)



- A distance-based mapping method based on data mining techniques
- The idea: haplotypes carrying trait loci tend to be more similar to each other than haplotypes drawn at random from the population
- The QTL association mapping is based on haplotype information from unrelated individuals
- Clusters are identified using a density-based clustering algorithm
- The method is implemented in the software ***HapMiner***

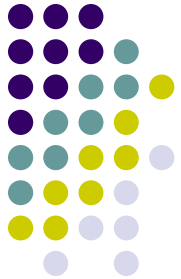
Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Methods

- Phased genotypes (haplotype pairs) have to be provided
- The algorithm scans each marker one by one
- For each marker position, a haplotype segment with certain length centered at the position is considered
- Clusters are identified based on a **similarity measure** via a density-based clustering algorithm
- For each cluster, a **Q-score** based on the *t-statistics* is calculated, representing the deviation of the phenotypic mean of the cluster from the phenotypic mean of all other samples
- The **Q-score** indicates the degree of association between the cluster and the phenotype

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Similarity score (Li & Jiang, 2005)



It combines:

- The length of the shared segments
- The number of common alleles around any marker position of the haplotype (Hamming similarity)

➔ This measure captures both:

- The sharing of haplotype segments due to historical recombination events
- It incorporates recent mutations and/or genotype errors

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Similarity Score (cont.)

e.g.: similarities between haplotypes (h1 and h2) and (h3 and h4):

h1 = (11212) h2 = (12222)

h3 = (11221) h4 = (21222)

- both pairs have three common allele
- h3 and h4 share a longer segment
- Similarity: the length of the longest common interval around the third locus in the middle, then:

  s(h1,h2) = 0
  s(h3,h4) = 2

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Similarity Score (cont.)

For a pair of haplotypes h_i, h_j, the similarity score with respect to locus 0 is:

$$s_{i,j} = \sum_{k=-l}^{r} w_1(x_k) I(h_i(k), h_j(k)) + \sum_{k=1}^{r'} w_2(x_k) + \sum_{k=-1}^{-l'} w_2(x_k),$$

- x_k = the genetic/physical distance from any locus to locus 0 ($-l \le k \le r$.)
- w_1 and w_2 = weights

The first summation: is a weighted measure of the number of alleles in common between haplotypes h_i and h_j in the region

The remaining summations: form a weighted measure of the longest continuous interval of matching alleles around locus 0

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Weight functions

- **Weight functions:** are based on the distance of a marker to the reference marker
- They are calculated as the linkage disequilibrium coefficient such as D'(x_0,x_k) between a locus *k* and the reference locus 0
- This proposed weight not only captures much information within block, they can also incorporate some moderate LD between blocks
- **HapMiner** automatically calculate the values of D'
- Haplotype segment length is set by the user

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Density-based algorithms

- They are based on the notion of local density:
  - High density areas form clusters and low density areas may be due to random noise

- For the clustering, the **DBSCAN** algorithm (Density Based Spatial Clustering of Applications with Noise) (Ester et al., 1996)
  - DBSCAN examines every haplotype and start to construct a cluster once a core haplotype is found
  - Then, iteratively collects directly reachable haplotypes from a core haplotype, merging clusters when necessary
  - The process terminates when all haplotypes have been examined
  - **Clusters are output and the haplotypes that do not belong to any cluster are regarded as noise**

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Association analysis

Q-score: measure the degree of association of a cluster and the quantitative trait

$$Q = \frac{(\mu_c - \mu_r)\sqrt{m+n-2}}{\sqrt{(\sigma_c^2 + \sigma_r^2)(1/m + 1/n)}}.$$

m = the number of haplotypes in the cluster

n = the number of remaining haplotypes

u_c σ^2_c the sample mean and the variance of the *m* haplotypes within the cluster

u_r σ^2_r the sample mean and the variance of the *n* remaining haplotypes

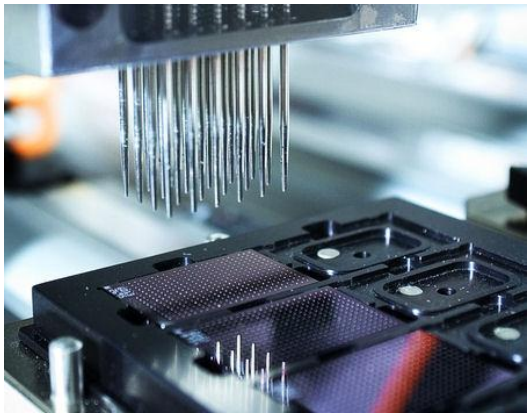
A large Q-score means a strong association between the cluster (i.e. haplotypes within the cluster) and the trait

- It is based on the ***t*-statistics** which assume that the haplotypes in the cluster and the remaining haplotypes are sampled from two different populations

- To assess the significance of the predicted gene position a **permutation test** is used to obtain **empirical *p*-values** (*i.e.* shuffling the phenotypes among the haplotypes)

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Haplotype phase and missing genotype inference



- In human genetics, thousands of individuals are genotyped for hundreds thousands of SNPs
- Although the SNPs data is very large, still a larger proportion of SNPs remain untyped
- Thus, genotype imputation methods have become increasingly popular for recovering untyped genotype data
- The methods rely on Hidden Markov Models (**HMM**)

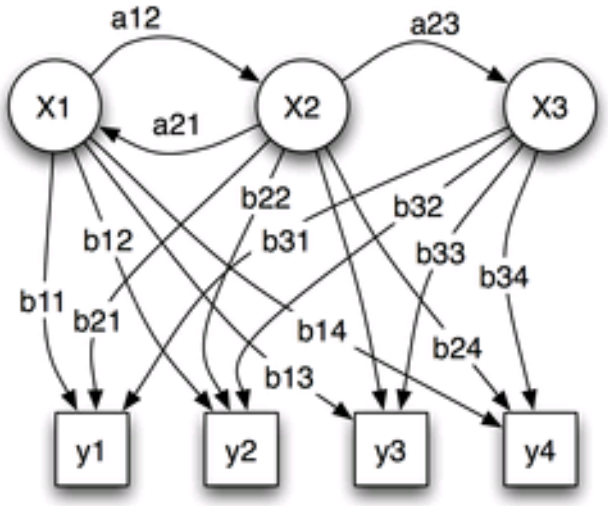
What is a HMM?

A **hidden Markov model** (**HMM**) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved state

An HMM can be considered as the simplest dynamic Bayesian network

-In a **regular Markov model**, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters.

-In a ***hidden* Markov model,** the state is not directly visible, but output, dependent on the state, is visible.



Probabilistic parameters of a hidden Markov model (example)
x — states
y — possible observations
a — state transition probabilities
b — output probabilities

From http://en.wikipedia.org/wiki/Hidden_Markov_model

What is a HMM?

- In a HMM, an underlying hidden (*i.e.* unobserved) states generates the observed data (Rabiner, 1989)

- In the context of haplotype phase and missing genotype inference:
  - the **observed data** are the observed unphased genotypes (errors and/or missing data),
  - while the **hidden state** represents the haplotype phase and true genotypes

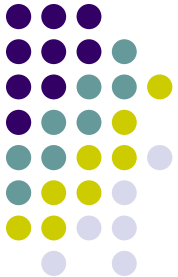
- A Markov Model is applied to the hidden states along the chromosome

- They have a very simple probabilistic structure that results in a relatively parsimonous model and facilitates computation

- The observed data at a marker depend only on the hidden state at the marker (the hidden state is said to “emit” the observed data)

Browning (2008)

2) Method to infer missing genotype data (Marchini et al., 2007)



- A model-based imputation method for inferring genotypes at observed and unobserved SNPs
  - The aim: to improve power in multipoint association mapping
- The main idea: to combine observed data and missing data
  - to predict (or “impute”) the missing data based upon the observed data
- It uses an approximate population genetics model:
  - It gives more weight to genotypes that are consistent with the local patterns of LD
  - It uses information from all the markers in LD with an untyped SNP, but in a way that decreases with genetic distance from the SNP being imputed
  - There is no need to specify the number of markers to be used, how to use them, and the physical distance to define haplotypes for haplotype analysis
- The method is implemented in the software **IMPUTE**

Marchini et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes

Methods

It is based on a HMM of each individual's vector of genotypes, G_i, conditional on H (a set of N haplotypes), and a set of parameters. The model is written as:

$$P(G_i|H, \theta, \rho) = \sum_z P(G_i|Z, \theta), P(Z|H, \rho)$$

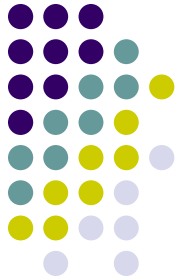
- Where $Z = \{Z_1, .., Z_L\}$ with $Z_j = \{Z_{j1}, Z_{j2}\}$ and $Z_{jk} = \{1, ..,N\}$. The $\boldsymbol{Z_j}$ can be though of as the **pair of haplotypes from the reference panel** at SNP_j that are being copied to form the genotype vector G_i.

- The term **Pr($Z|H, \rho$)** models how the pair of copied haplotypes changes along the sequence and is defined by a Markov chain in which the switching between states depends on an estimate of the **fine-scale recombination map (ρ)** across the genome.

- The term **Pr($G_i|Z, \theta$)** allows each observed genotype vector to differ through mutation from the genotypes determined by the pair of copied haplotypes and it is controlled with the **mutation parameter θ**.

Marchini et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes

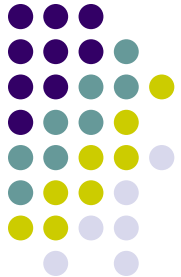
Methods (cont.)



- The **fine-recombination map** (in cM/Mb) is estimated from the phase *II HapMap*
  - It is used as a fixed set of parameters in the models and is scaled by an estimate of the effective population size (N_e) to obtain the population scale recombination map across the region
- The **mutation parameter θ** is fixed internally by the program
- $\boldsymbol{N_e}$ must be set by the user

Marchini et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes

Methods (cont.)



- The exact **marginal probability distribution** of each possible genotype (0, 1, 2) for the missing genotypes that are conditional on the observed genotype in the vector G_i are obtained using a forward-backward algorithm

- It also provides a probability distribution for each called genotype to facilitate correction of genotyping errors

- These **probabilities** can be used to carry out **an association test** at all typed and untyped SNPs

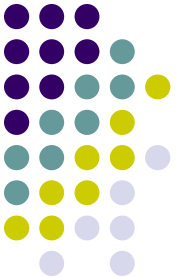
Marchini et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes

Association test

- The imputation model generates probability distributions of untyped genotypes
- Once genotypes have been imputed, a test of association can be carried out at a much larger set of SNPs than the one originally typed
- By testing each SNP in turn, it is assumed that disease variants will be detected based on their marginal effects
- A single-SNP test of association takes this uncertainty into account
  - It involves Bayesian statistics (Bayes Factor)

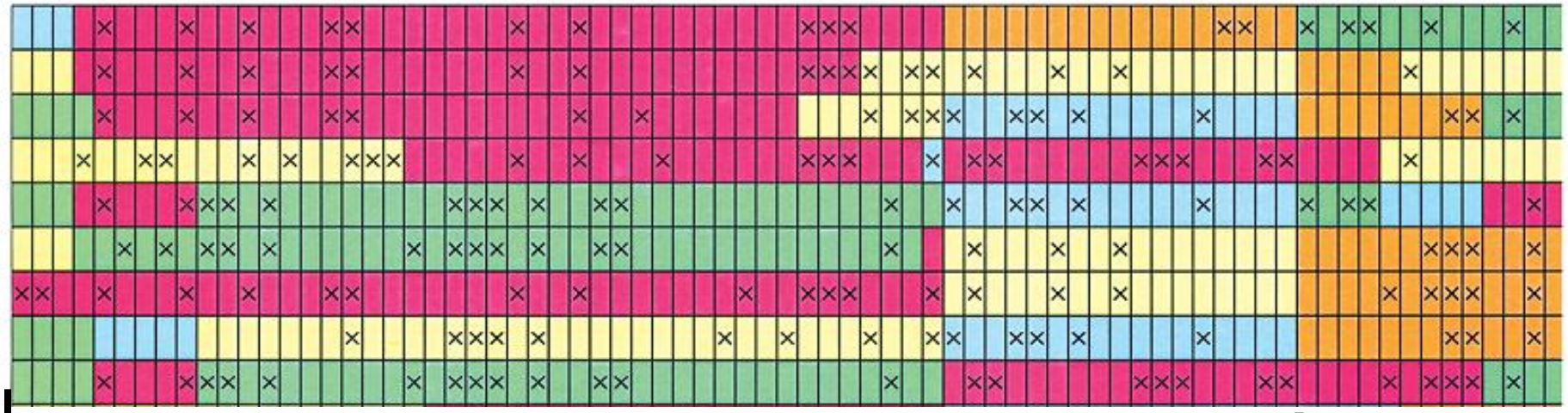
Marchini et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes

3.1) Method to infer haplotype phase and missing genotype by clustering haplotypes (Scheet and Stephens, 2006)



- The idea: over short regions, haplotypes in a population tend to cluster into group of similar haplotypes
- Clustering tends to be local in nature: as a result of recombination, those haplotypes that are closely related to one another and therefore similar will vary as one moves along the chromosome
- The model allows:
  - clusters memberships to change continuously along the chromosome according to a HMM
  - for both “block-like” patterns of LD and a gradual decline in LD with distance
- The method is implemented in the software ***fastPHASE***
- It is applicable to large data sets

Scheet & Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.



Row: estimated pair of haplotypes for successive individuals

Column: a SNP

Colors: estimated cluster membership of each allele, which changes as one moves along each haplotype

-Each cluster can be thought of as (locally) representing a common haplotype, or a combination of alleles,
-The HMM assumption for cluster membership results in each observed haplotype being modeled as a mosaic of a limited number of common haplotypes

Scheet & Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.

Methods

- The model specifies a set of ***K* unobserved states** or **clusters** that represent common haplotypes

- Each individual's genotype data is modeled as a HMM on this state space with transitions between states controlled by a set of parameter (*r*) at each SNP

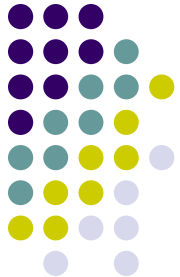
- The probability of G_i is obtained as:

$$P(G_i|\alpha, \theta, r) = \sum_z P(G_i|Z_i, \theta)P(Z_i|\alpha, r)$$

α = a weight that denotes the fraction of haplotypes it contains a site *l*
θ = the associated frequency of allele 1 at each site (for each cluster)
$P(G_i|Z_i,\theta)$ = models how likely the observed genotypes are given the underlying states
$P(Z|\alpha,r)$ = models patterns of switching between states, where states represent clusters

Scheet & Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.

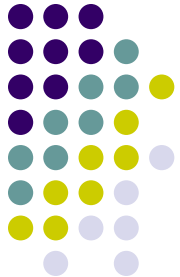
Methods (cont.)



- *Missing-genotype imputation:*
  - This method imputes genotypes marginally and provides **"best guess"** for each genotype
  - It sample from the joint distribution of the missing genotypes given observed-data, *e.g.,* by sampling from the conditional distribution of the haplotypes for all individuals
- *Haplotype inference:*
  - Sampling the pairs of haplotypes of all individuals from their joint distribution given the unphased genotype data

    ➔ it provides a useful way to asses or account for uncertainty in haplotype estimates

Methods (cont.)



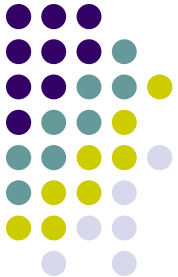
- EM algorithm

- Results across applications of the EM algorithm are averaged
  - It provides much better results than choosing a single best estimate

- Number of clusters:
  - **fastPhase** can choose an optimal number of clusters (5, 10 and 15)
  - For large data sets, between 20 and 30 clusters
  - Computation times increases quadratically with the number of clusters

Scheet & Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.

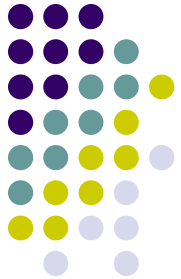
3.2) Method to infer haplotype phase and missing genotype data by clustering haplotypes Browning and Browning (2007)



- It is a novel application of the ***localized haplotype-cluster model*** used for association testing (Browning and Browning, 2007; Browning, 2006)
- The localized haplotype-cluster model is an empirical LD model that adapts to the local structure of the data
- It uses an iterative approach for phasing haplotypes
- It is implemented in the software ***Beagle***:
  - single marker and multilocus association analysis, permutation testing
- It can be applied to large whole-genome data sets

Browning & Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.

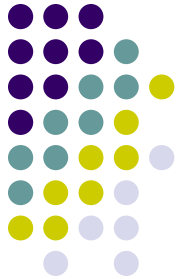
Localized haplotype-cluster model



- It is a special class of **directed acyclic graph** (**DAG**)
- It defines a **Hidden Markov Model** that can be used to sample haplotype pairs or to find the most likely haplotype pair for each individual conditional on the individual's genotypes
- The phasing algorithm involves iteratively sampling haplotypes pairs and building the localized haplotype-cluster model from the sampled haplotype pairs

Browning & Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.

Localized haplotype-cluster model



- It makes use of the localized LD, which empirically models haplotype frequencies on a local scale
  - Correlation between markers is a localized phenomenon, since LD decays with distance
- It clusters haplotypes at each marker to improve prediction of alleles at markers *t*+1, *t*+2, *t*+3,.., given alleles at marker *t*, *t*-1, *t*-2, …on a haplotype
- This is achieved by defining cluster according to a Markov property: given cluster membership at position t, the sequence of alleles at markers *t*, *t*-1, *t*-2, … is irrelevant for predicting the sequence of alleles at marker *t*+1, *t*+2, *t*+3,…

Browning & Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.

Directed acyclic graph (DAG)

- The **edge**, ***e***, represents a cluster of haplotypes consisting of all haplotypes whose path from the initial node to the terminal node of the graph traverses ***e***.

- Haplotypes are defined over the whole chromosome, but haplotypes within a cluster corresponding to an edge at level *m* will tend to have similar patterns of alleles at markers immediately to the right of marker *m*.
  - Each edge defines a localized haplotype cluster that is determined by local LD pattern.

- Recombination between haplotypes is modeled as merging edges

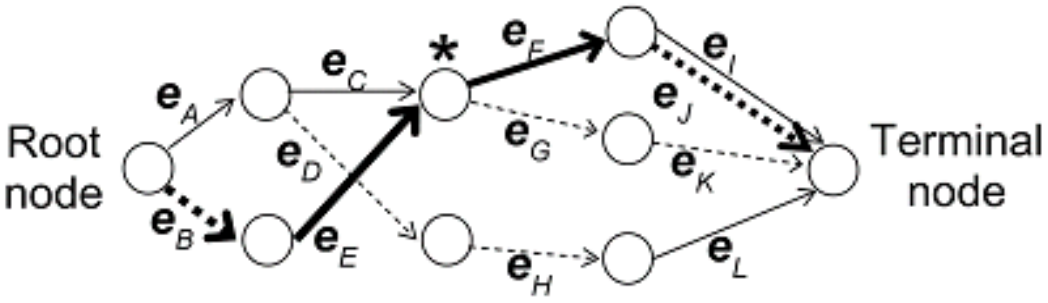
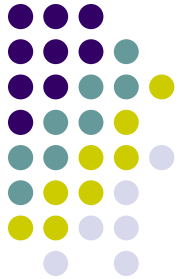


Figure 1. Example of a DAG representing the localized-cluster model for 4 markers. Edge e_F includes haplotypes 1**11**1, 1**11**2, 2**11**1 and 2**11**2. The bold-lines edges from the root to the terminal node represent the haplotype 2**11**2. For each maker, allele 1 is represented by a solid line and allele 2 by a dashed line. The node marked by an asterisk is the parent of node e_F .

Browning & Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.

The induced HMM



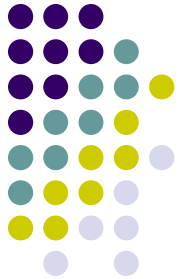
- A localized haplotype-cluster model determines an HMM
  - the states of the HMM are the edges of the localized haplotype-cluster model, and the emitted symbol for each state is the allele that labels the edge of the localized haplotype-cluster model

- HMM
  - emission probabilities
  - initial-state probabilities, and
  - transmission probabilities

  - The initial-state probabilities and the transition probabilities are computed from the edge counts.

Browning & Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.

The Beagle Phasing algorithm

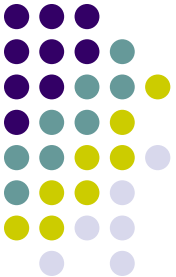


- The phasing algorithm samples from diploid HMM conditional on the observed data by use of a forwards-backwards algorithm

At each iteration of the algorithm:

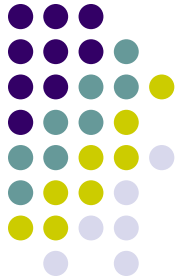
- Phased input data are used to build a localized haplotype-cluster model
- Phased haplotypes for each individual are sampled from the induced diploid HMM conditional on the individual's genotypes
- The sample haplotypes are the input of the next iteration
- At the final iteration, the Viterbi algorithm is used to select the most-likely haplotypes for each individual, conditional on the diploid HMM and the individual's genotype data
- The **most-likely haplotypes** are the output of the phasing algorithm

4) Method to perform an Imputation-based Bayesian regression analysis (Servin and Stephens, 2007)



- It is a new methodology for using imputed values in association testing
- BIM-BAM uses fastPHASE to perform the imputation of genotypes
- Missing data are imputed multiple times, with the imputed values being used in a Bayesian regression approach to test for association
- Implemented in the package ***BIM-BAM*** (Bayesian Imputation-Based Association Mapping)
- It is applicable for whole genome association studies and candidate gene studies

Servin & Stephens (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits



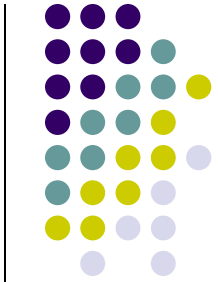
- It has an improvement on standard analysis by exploiting available information on LD among untyped and typed SNPs

- Partial information can be available from:
  - the International HapMap project
  - resequencing data available from public data bases
  - data collected from association study designs

➔ the approach combines this background knowledge of LD with genotypes collected at typed SNPs in the association study, to predict ("impute") genotypes in the study sample at untyped SNPs

Servin & Stephens (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits

Association study design



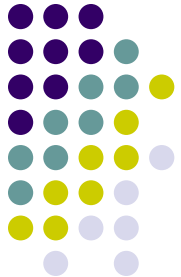
- ***Observed genotypes (G_{obs}):*** genotype data available for a dense set of SNPs on a panel of individuals

- **Tag SNPs:** genotypes available for a subset of these SNPs on a cohort of individuals who have been phenotyped for a quantitative trait

- The strategy:
  - To use patterns of LD in the panel together with the tag SNP genotypes in the cohort to predict the genotypes at all makers for the members of the cohort
  - To analyze the data as if the cohort had been genotyped at all markers (tag and non-tag)

Servin & Stephens (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits

Association study design (cont.)



- **fastPHASE** is used to generate multiple imputations for the complete genotype data (all individuals at all SNPs) by sampling from $P(G/G_{obs})$

- These imputations are incorporated in the inference which involves adding a step in the MCMC scheme to sample the imputed genotypes from their posterior distribution given all data and averaging relevant calculations over imputations

Servin & Stephens (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits

Standard linear regression

$$y_i = \mu + \sum_j \mathbf{x}_{ij}\beta_j + \varepsilon_i$$

$\boldsymbol{y_i}$ = the phenotype for individual *i*
u **=** the phenotype mean of individuals carrying the "reference" genotype
$\boldsymbol{x_{ij}}$**s** = the elements of the design matrix *X* (which depends on the genotype data)
$\boldsymbol{\beta_j}$**s** = the corresponding regression coefficients
$\boldsymbol{\varepsilon_i}$**s =** independent and identically distributed ~N(0,1/τ), where τ denotes the inverse of the variance

$$y_i|\mu, \mathbf{x_i}, \beta, \tau \sim N(\mu + \sum_j \mathbf{x}_{ij}\beta_j, 1/\tau)$$

$$P(y_i|\mathbf{x_i}, \mu, \beta, \tau) \propto \sqrt{\tau}\exp[-0.5\tau(y_i - (\mu + \sum_j \mathbf{x}_{ij}\beta_j))^2]$$

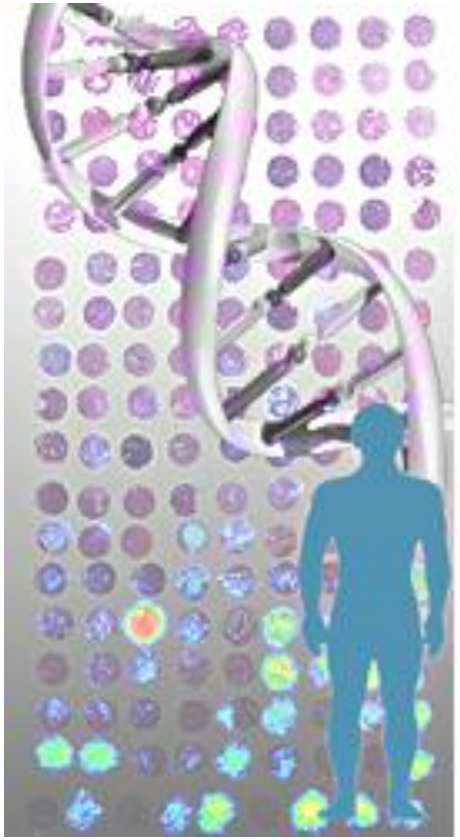
Servin & Stephens (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits

Bayes Factor

- Bayes Factor (**BF**): BF = $P(y|G, H_1) / P(y|G, H_0)$
- $\boldsymbol{H_0}$ denotes the null hypothesis that none of the SNPs is a QTN ($a_j=d_j=0$ for all j)
- $\boldsymbol{H_1}$ denotes the complementary effect (*i.e.* at least one SNP is a QTN)
- Computing BF involves integrating out unknown parameters
- A ***p*-value** for testing H_0 can also be obtained from a BF through permutation

Servin & Stephens (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits

SUMMARY



HapMiner

- It is an efficient method for whole-genome studies given that the sliding window of haplotypes is not very large
- Limitation: required phased haplotypes as input
- Association analysis: **HapMiner** was more robust and achieved better power than the single-marker association analysis under the simulated scenarios
- The effectiveness of the method depends on the similarity measure of haplotype fragments and the clustering algorithm

Beagle & fastPHASE

- Conceptually, both methods are similar

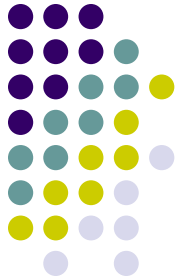
- Difference:
  - **Beagle:** allows the number of clusters vary from one position to another

  - **fastPHASE:** the number of cluster is fixed

In **Beagle**, the graph will have few or many edges in regions in which there is low or high LD, respectively

➔ The number of cluster can vary at each locus and complex recombination patterns can be found in the data ☺

Beagle & fastPHASE (cont.)



Parameter estimation:

- **fastPHASE**: a large number of parameters (definition of haplotype cluster and recombination and mutation rates) need to be estimated
  - For realistic data sets is expected that all parameters to be not well correctly estimated (Scheet and Stephens, 2006)

- **Beagle:** there is not need to estimate parameters such as recombination and mutation rates
  - Thus, the Beagle model is more parsimonious
  - It makes computation faster
  - It seems to have an effect in the number of iterations required compared to **fastPHASE** (Browning, 2008)

BIM-BAM

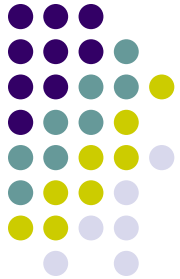
- **BIM-BAM** uses **fastPHASE** to perform the imputation and includes a new approach for using imputed values for test association

- Missing data are imputed multiple times and the imputed values are used in a Bayesian regression analysis

- An interesting point about this approach:
  - It tests variants about which something is known (*i.e.* SNPs that are known to exist and have documented patterns of LD) and exploits this information

  - This idea is more compelling than testing hypothetical untyped variants about which nothing is known, like in **fastPHASE** (Servin and Stephens, 2007)

Comparison among methods



- **Accuracy** (*e.g.* error rate) & **performance** (*e.g.* computing time) of the different software
  - different data sets (*i.e.* number of markers and individuals) & different assumptions

- Sheet and Stephens (2006):
  - fastPHASE was as accurate than PHASE, GERBIL, HaploBlock
  - For haplotype estimation, fastPHASE was slightly less accurate than the other methods, but required a small fraction of computational cost

- Browning and Browning (2007):
  - Beagle outperformed fastPHASE, HaploRec, Hap and 2SNP in terms of speed and accuracy (3,002 individuals genotyped for 490,032 markers with 99% of masked alleles imputed correctly)

- Marchini and Howie (2010):
  - IMPUTE was at least twice as fast as both Beagle and fastPHASE to impute genotypes using a reference panel of 1000 haplotypes

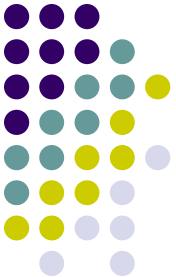
Animal genetics

5) To estimate IBD probabilities using linkage (**LA**) and linkage disequilibrium (**LD**) information (Meuwissen and Goddard, 2001)

6) To reconstruct haplotypes using LA information, cluster haplotypes based on IBD probabilities, and to perform a variance component analysis to map QTL (Druet et al., 2008)

7) To simultaneously infer haplotypes and missing genotypes (combining LA and LD information) and sort haplotypes clusters and to perform a variance component analysis to map QTL (Druet and Georges, 2010)

5- Method to estimate IBD probabilities using linkage (LA) and linkage disequilibrium (LD) information (Meuwissen and Goddard, 2001)



- The method predicts **IBD probabilities** at a given chromosomal location given data on a haplotype of markers spanning that position
- The probabilities that two gametes are IBD at a particular locus increases as the number of markers surrounding the locus with identical alleles increases
- It is based on a simplification of the coalescence process, and assumes that the number of generations since the base population and effective population size is known
- It was developed for the situation where the pedigree of the animals was unknown (*i.e.* all the information come from the maker genotypes), and the situation where *T* generations of unknown pedigree are followed by some generations where pedigree and marker genotypes are known

Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

IBD probabilities: one linked marker

The probability that two haplotypes are IBD at some locus of interest (*e.g.* locus A) is estimated given **one** or **multiple** linked markers

- **One linked marker:** the probability that the alleles at locus A are IBD given the marker data is:

$$P(\text{IBD}|\text{marker}) = P(\text{IBD}|S) = \frac{P(A = \text{IBD} \;\&\; S)}{P(A = \text{IBD} \;\&\; S) + P(A = \text{nonIBD} \;\&\; S)} \qquad (1)$$

S = an indicator of the Alike in State (AIS) situation of the maker alleles:

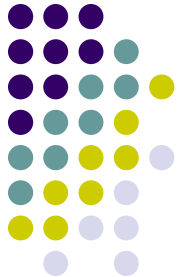
S=1 ➔ alleles are AIS

S=0 ➔ alleles are on AIS

If S=1, the maker locus may be IBD or nonIBD

Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

IBD probabilities: one linked marker



- (ϕ) (character string) = to summarize the IBD status in the region and indicates:
- if locus A is IBD or nonIBD
- if the marker locus is IBD or nonIBD
- if the region between the two loci is IBD:
  - the same common ancestor as the loci (*i.e.* the region in between the makers was inherited as a whole from the same common ancestor without a recombination that split the region)
  - there has been a recombination and, if the two loci are IBD, they are probably IBD due to different common ancestors

Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

The P(S & A) = IBD can be obtained by summing all over possible IBD status, ϕ, with locus A= IBD:

$$P(S \,\&\, A = \text{IBD}) = \sum_{\phi|\phi(1)=1} P(S|\phi) \times P(\phi)$$

$$P(S \,\&\, A = \text{nonIBD}) = \sum_{\phi|\phi(1)=0} P(S|\phi) \times P(\phi)$$

- $\Sigma_{\phi\,/\,\phi\,(1)=1}$ ($\Sigma_{\phi\,/\,\phi\,(1)=0}$) = summation over all possible ϕ vectors where locus A is (non)IBD
- P(S/ ϕ)= the probability of AIS markers denoted by S given the statuses denoted by ϕ
- P(ϕ) have to be calculated for all the status
- The P(S & locus A) with IBD and non IBD locus A are combined in equation (1) to obtain the probability that locus A is IBD given the linked marker haplotype.

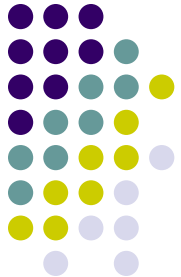
Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

IBD probabilities: multiple linked markers

- Equation (1) remains the same, except that the maker information is now due to several markers
  - S is a (*m* x 1) vector of AIS status indicators, where *m* is the number of marker loci in the haplotype
  - Φ vector is extended by adding two characters for every additional locus
    - one indicating whether the region between this locus and the previous locus was inherited in block from a common ancestor or not
    - one character indicating whether the locus is IBD or nonIBD

Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

Pedigree information



- The information on markers splits the pedigree into 2 parts:
  - i) Generations where neither pedigree nor maker data is available

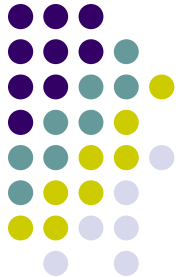
    ➔ This pedigree part results in LD marker haplotypes and locus A in the first generation of the pedigreed population and thus contains the LD information

  - ii) Generations with known pedigree and marker data, although the marker information may be missing on some individuals

    ➔ This pedigree part contains the linkage information, the inheritance of the makers and locus A are traced through the known pedigree and the frequency with which recombinations occur yield information about the linkage between locus A and the markers

Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

Pedigree information (cont.)



- **In livestock pop:** some generations of pedigree recorded but non-genotyped individuals followed by generations of genotyped and pedigree recorded animals

- An approximation to calculate IBD probabilities given marker and pedigree information, in the situation where the pedigree of genotyped animals is known for some generations, but the individuals in the pedigree are not genotyped.

- The approach is analogous to the Wright's F-statistics:
  - markers haplotypes are related due to a finite population size for *T* generations (pedigree part i, Wright's F_{ST})
  - marker haplotypes are related due to relationships in the pedigree (pedigree part ii, Wright's F_{IS}).

Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

The total IBD at locus A given the one generation of marker haplotypes and some ancestral generations of pedigree (analogous to Wright's F_{IT}) is:

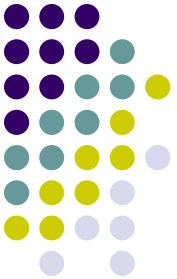
$$P_{IT}(\text{IBD}|\text{marker, pedigree}) = P_{IS}(\text{IBD}|\text{marker, pedigree}) + [1 - P_{IS}(\text{IBD}|\text{marker, pedigree})]P(\text{IBD}|\text{marker})$$

- P_{IS}(IBD|marker, pedigree) = the IBD probability at locus A due to a common ancestor within the pedigree and given the maker information (*i.e* due to recent relationships)

- P(IBD|marker) = the probability that two regions are IBD before they entered the pedigree, *i.e* due to *T* generations of random drift in a population of size N_e

Meuwissen & Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes

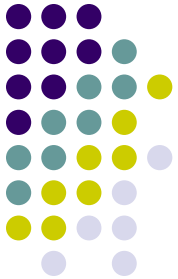
6- Method to reconstruct haplotypes using LA information, cluster haplotypes based on IBD probabilities, and to perform a variance component analysis to map QTL (Druet et al., 2008)



- A fine mapping analysis of QTL affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map
  - IBD probabilities estimation among base haplotypes (Meuwissen and Goddard, 2001)
  - These probabilities were used to group the base haplotypes in different clusters
- A granddaughter design with 17 half-sib families (926 sons with phenotype & genotyped) and pedigree information
- **Linkage analysis** (regression and VC analysis)
- **Linkage and linkage disequilibrium** (**LDLA**)
  - VC mapping method that includes information from LD between base haplotypes in the construction of the relationship matrix among QTL allelic effects

Druet et al. (2008). Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using single-nucleotide polymorphism map

Chromosomes were grouped in different categories: sire chromosome (**SC**) and paternally and maternally inherited chromosomes (**PC** and **MC**) of the sons. **SCs** and **MCs** are considered as base haplotypes. At each tested position:



- Probabilities of transmission are computed to determine to which **SC** a **PC** corresponds. The rules to calculate these probabilities are the same as those computed for the linkage analysis. LD information is not taken into account in this step

- **IBD probabilities** (ϕ_p) are estimated among each pair of base haplotypes conditionally on the IBS status of the neighboring markers using a window of 10 flanking markers (Meuwissen and Goddard, 2001)

- Base haplotypes are grouped with a clustering algorithm with SAS proc CLUST, using ($1 - \phi_p$) as a distance measure

- Base haplotypes were grouped if $\phi_p > 0.5$ (Ytournel et al., 2007). PCs are grouped within clusters if:
  - i) the two SCs of a sire are grouped in the same cluster (the PCs of all his sons are then grouped in this cluster) or
  - ii) a PC can be associated with a base haplotype with a probability > 0.95 (it is grouped to the corresponding cluster)

Druet et al. (2008). Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using single-nucleotide polymorphism map

Mixed model

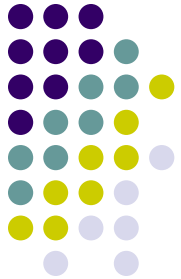
$$y = Xb + Zu + Z_h h + e$$

- ***y*** = the phenotype
- ***b*** = vector of fixed effects
- ***h*** = vector of random QTL effects corresponding to the *haplotype clusters*
- $\boldsymbol{Z_h}$ = design matrix relating phenotypes to corresponding haplotype cluster
- ***u*** = vector of random individual polygenic effects
- ***e*** = vector of individual error term

- Genetic parameters are estimated an AI-REML approach
- Likelihood ratio tests (**LRT**) are used to confirm the QTL presence at the studied position

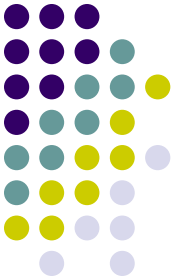
Druet et al. (2008). Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using single-nucleotide polymorphism map

Source of information



- In human genetics ➔ phasing methods usually exploit population information (LD)
- In animal genetics ➔ the primary source of information is familial (Mendelian segregation and linkage) provided by extended pedigree usually available in livestock populations
  - A proportion of genotypes are left unphased, especially for the less connected individuals
  - The use of high density SNPs makes the computation of pairwise IBD probabilities (Meuwissen and Goddard (2001)) to be used for haplotype reconstruction a difficult and limiting task

7- Method to simultaneously infer haplotypes and missing genotypes, to cluster haplotypes combining LA and LD information, and to perform a variance component analysis to map QTL (Druet and Georges, 2010)



- An approach based on HMM that can simultaneously phase and sort haplotypes clusters that can be directly be used for mapping or other purposes
  - It exploits simultaneously both familial information and population information
    - LD information: Beagle and fastPHASE to infer haplotype phase and missing genotypes
  - It assigns reconstructed haplotypes to hidden states that are shown to correspond to clusters of genealogically related chromosomes
  - Cluster states can be directly used to fine map QTL
  - It can handle large data sets based on high-density SNP panels

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

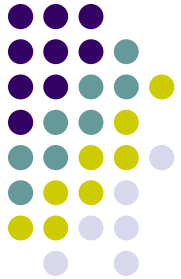
Haplotype reconstruction and clustering

In livestock populations extended pedigrees are available. For the haplotype reconstruction:

- split the pedigreed to consider only relationships between genotyped parents and genotyped offspring
- In the ensuing subpedigrees:
  - individuals can be parent only
  - offspring only
  - or parent and offspring

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

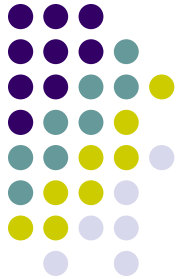
Phasing of heterozygous SNPs



- Step1: Mendelian segregation
- Step 2: Linkage information (offsprings and parents)
- Step 3: Linkage disequilibrium
  - To complete haplotype reconstruction, LD is exploited using algorithms developed either in fastPHASE or in Beagle with modifications

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

Phasing of heterozygous SNPs (cont.)



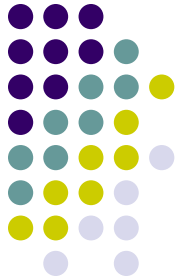
- **Step1: Mendelian segregation**

- Markers alleles of heterozygotes offspring are assigned to the paternal and maternal homolog following Mendelian segregation rules

- In offspring, this leaves only SNPs unphased for which parents have the same heterozygous genotypes as the offspring

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

Phasing of heterozygous SNPs (cont.)



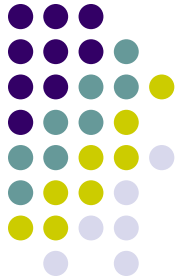
- **Step 2: Linkage**

- ***Parents:*** parental phases are completed on the bases of allelic cosegregation in the offspring (Druet et al., 2008)
  - This process requires heterozygous “anchoring” markers whose alternate alleles define the padumnal *vs* madumnal homolog of the parent

- ***Offspring:*** Heterozygotes markers than remain unphased in offspring can be further treated conditional on the known parental phase (determined in Step 1 and 2) , according to Step 4 in Druet et al. (2008)

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

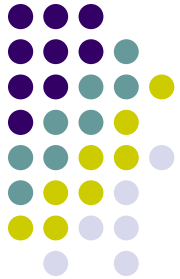
Phasing of heterozygous SNPs (cont.)



- **Step 3: Linkage disequilibruim**
  - To complete haplotype reconstruction, LD is exploit using algorithms developed either in fastPHASE or in Beagle with some modifications

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

Step 3: LD information



- The ***fastPHASE probability model*** (**FPM**):
  - The observed haplotypes are modeled as mosaics of *K* hidden states (**HS**), with *K* (number of clusters) held constant throughout the genome.

- The ***Beagle probability model***:
  - It uses a localized haplotype clustering model (**LHCM**), which can be interpreted as a special class of HMM.
  - The number of *K* of HS is allowed to vary across the genome and it is determined by the number of edges (at the corresponding marker position) of a DAG summarizing all haplotypes encountered in the population

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

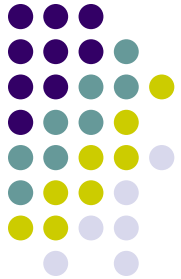
Both HMM are applied in two stages:

1) The models are trained on a haploid training set consisting of the partially phased base haplotypes obtained after Steps 1 and 2.

fastPHASE: generates EM parameter estimators
Beagle: generates an optimal DAG

2) The actual haplotype reconstruction and clustering is completed by running a diploid HMM on the complete data

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping



- Knowledge of HS status in base and descendent individuals allows:
  - Phasing of the markers that remained unresolved after Steps 1 and 2, and
  - imputation of genotypes at missing marker positions
- The corresponding data augmentation is achieved by sampling unresolved phases and missing genotypes according to their respective probabilities computed form the allele-specific emission probabilities of the constituent HS

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

For the QTL mapping, to test the presence of a QTL at a given map position, the following mixed model was used

$$Y = Xb + Z_h h + Z_u u + e$$

b = vector of fixed effects

h = is the vector of random QTL effects corresponding to the ***K* defined HS**

u = is vector of random individual polygenic effects and *e* is vector of individual error terms.

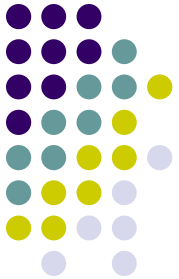
The co-variance between individuals polygenic effects correspond to twice the coefficient of coancestry times the additive genetic variance σ^2_A.

The co-variance between different HS effects was assumed to be zero, hence modeling QTL with a finite number of alleles.

VC were estimated with a REML analysis

The presence of a QTL at a given map position was tested by a LRT

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping



$$Y = Xb + Z_h\, h + Z_u\, u + e$$

Druet et al (2008): $Z_h\, h$ = *cluster*

Druet & Georges (2010): $Z_h\, h$ = *hidden states*

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

Software

- ***PHASEBOOK***
  - **LinkPHASE**: runs Steps 1 (Mendelian) and 2 (Linkage)
  - **HiddenPHASE**: runs Step 3 (LD) with FPM
  - **DualPHASE:** runs Step 1, 2 and 3 with FPM
  - **DAGPHASE:** runs Step 3 with LHCM

Druet & Georges (2010). A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping

Summary

- **Meuwissen and Goddard (2001):**
  - It has been widely used in livestock population studies
  - Drawback : with increasing number of markers and genotyped animals the computation of IBD probabilities becomes limiting
  - This method only provides IBD probabilities
  - No software is provided for this method

- **Druet et al. (2008):**
  - The method use the IBD probabilities to cluster haplotypes and then use the clusters to fit them in a mixed model to map QTLs
  - The clustering was done by a SAS proc Procedure
  - Combining LA and LD information provided a better location of the QTL than the analysis based on LA only

Summary (cont.)

Druet and Georges (2010):

- The simultaneously extraction of LD and familial information improved the accuracy of phase reconstruction and provided accurate genotype imputation

- Comparison this novel approach with the standard LDLA - QTL mapping approach (Druet et al., 2008):

  - The mapping results obtained from the FPM and LHCM were comparable to those obtained with the standard approach based on the pairwise IBD probabilities

  - Significant differences in computing time:
    - to phase and cluster haplotypes:
      - LHCM model (DAGPHASE) : **47 minutes**
      - FPM model (DualPHASE) : **966 minutes**

    - To compute IBD probabilities using already phased genotypes : **9133 minutes!**

Conclusions

- **Beagle** has two advantages:
  - It models patterns of LD with a flexible number of clusters
  - not many parameters need to be estimated, which makes computation faster

- **Beagle**, **fastPHASE**, **IMPUTE**: imputed accurately missing genotypes and haplotype phase reconstruction for large human data sets, containing thousand of individuals and markers

  - In the context of dairy cattle populations these methods may work properly given the size of the data available for this population

  - In pig populations, *e.g*. data from DELISUS project (*i.e.* small half-sib families with 10, 30 and 50 sires)

  These methods will inference missing genotype and haplotype phases accurately?

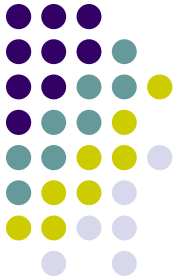
For small populations should we look for other methods (e.g. PHASE) ?

Conclusions (cont.)

- Methods that used reference panels to impute missing genotypes for association analysis (**BIM-BAM** and **IMPUTE**)
  - There is a gain of information that can improve imputing accuracy
  - Including imputed genotypes increase the power of test associations

What about the availability of reference information in livestock populations?

Conclusions (cont.)



- **HapMiner** seems to be a “simple” method to cluster haplotypes for association analysis

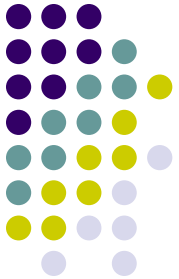
- Disadvantages:
  - Phase haplotypes need to be provided
  - Haplotype length (not too large)

  - *Thus, patterns of LD might not be properly modeled as in the other methods?*

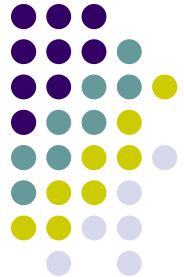
Conclusions (cont.)

- Druet and Georges's approach is very interesting
  - They managed to combine both population and familial information, using method developed in human genetics to exploit the population information (LD)
  - The method accurately imputed missing genotypes and inferred phase haplotypes in a large data base
  - Software is available from the authors

These methods will work properly in small populations?



Merci !



What is a HMM?

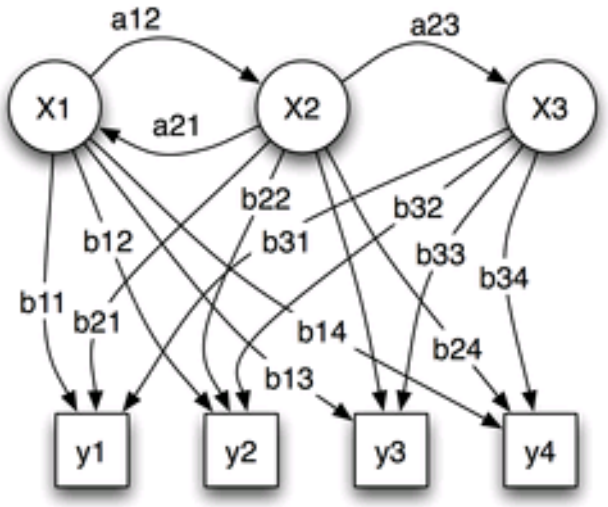
A **hidden Markov model** (**HMM**) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved state

An HMM can be considered as the simplest dynamic Bayesian network

-In a **regular Markov model**, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters.

-In a ***hidden* Markov model,** the state is not directly visible, but output, dependent on the state, is visible.

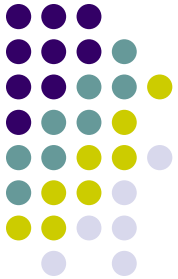
Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; even if the model parameters are known exactly, the model is still 'hidden'



Probabilistic parameters of a hidden Markov model (example)
x — states
y — possible observations
a — state transition probabilities
b — output probabilities

From http://en.wikipedia.org/wiki/Hidden_Markov_model

Directed acyclic graph (DAG)



- It is supposed a sample of haplotypes for *M* markers and the haplotypes have no missing alleles. A localized haplotype-cluster model is a DAG:
- **Root node (initial):** represents all the haplotypes before any markers are processed. It has not incoming edges. Terminal node: represents all the haplotypes after all markers are processed. It has not outgoing edges.
- **Levels of the graph:** *M* +1. Each node *A* has a level *m*. All incoming edges to *A* have the parent note at level *m*-1, and all outgoing edges for A have the child node at level *m*+1. The root node has level 0 and the terminal node level *M*.
- For each *m*=1,2,..,*M*, each edge with the child node at level *m* is labeled with an allele for the *m*th marker.
- For each haplotype in the sample, there is a path from the root node to the terminal node, such the *m*th allele of the haplotype is the label of the *m*th edge of the path. Each edge of the graph has at least one haplotype in the sample whose path traverses the edge.

Browning & Browning (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.

Similarity Score (cont.)

e.g.: Similarities between haplotypes (h1 and h2) and (h3 and h4):

h1 = (11212) h2 = (12222)

h3 = (11221) h4 = (21222)

- Genotyping error or a point mutation from the ancestor haplotype at the 2nd position of h1

➔ the similarity of h1 and h2 will be underestimated

Li et al. (2006). Haplotype-based quantitative trait mapping using a clustering algorithm

Missing-genotype imputation

- For any genotype g_{im} that is unobserved ("missing"), it is straightforward to compute the probability $g_{im} = x$ (x= 0, 1, 2), given all observed genotypes g and parameter values v by the use of:

$$p(g_{im} = x|g,\nu) = \sum_{k_1=1}^{K} \sum_{k_2=k_1}^{K} p(g_{im} = x|z_{im}^{\cdot} = \{k_1,k_2\},\nu)$$

$$\times\ p(z_{im}^{\cdot} = \{k_1,k_2\}|g_i,\nu)\ .$$

The **first term** is the sum given in the previous equation and the **second term** is the conditional distribution of the hidden variables in the HMM

This method imputes genotypes marginally and provides **"best guess"** for each genotype

It sample from the joint distribution of the missing genotypes given observed-data, *e.g.,* by sampling from the conditional distribution of the haplotypes for all individuals

Scheet & Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.

Haplotype inference

Two aspects are considered:

- Sampling the pairs of haplotypes of all individuals from their joint distribution given the unphased genotype data

  ➔provides a useful way to asses or account for uncertainty in haplotype estimates

- Construction of point estimates of the haplotypes carried by each individual

Scheet & Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.