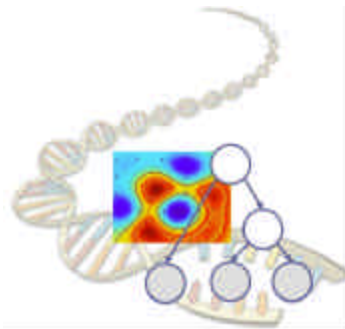


SNPs and Haplotype Inference



10-810, CMB lecture 10---Eric Xing

Polymorphism



- Alleles: Alternative DNA sequences at a locus
- Technical definition: most common variant (allele) occurs with less than 99% frequency in the population
- Also used as a general term for variation
- Many types of DNA polymorphisms, including RFLPs, VNTRs, microsatellites
- 'Highly polymorphic' = many variants

Type of polymorphisms



- Single base mutation (SNP)
 - Restriction fragment length (RFLP)
 - Creating restriction sites via PCR primer
 - Direct sequencing
- Insertion/deletion of a section of DNA
 - Minisatellites: repeated base patterns (several hundred base pairs)
 - Microsatellites: 2-4 nucleotides repeated
 - Presence or absence of Alu segments

Frequency of SNPs greater than that of any other type of polymorphism

Single Nucleotide Polymorphism (SNP)



```
GATC TTCGTAC TGAGT
GATC TTCGTAC TGAGT
GATT TTCGTAC GGAAAT
GATT TTCGTAC TGAGT
GATC TTCGTAC TGAAT
GATT TTCGTAC GGAAAT
GATT TTCGTAC GGAAAT
GATC TTCGTAC TGAAT
```

chromosome 

- “Binary” nt-substitutions at a single locus on a chromosome
 - each variant is called an “allele”

Single Nucleotide Polymorphism (SNP)



- More than 5 million common SNPs each with frequency 10-50% account for the bulk of human DNA sequence difference
- About 1 in every 600 base pairs
- It is estimated that ~60,000 SNPs occur within exons; 85% of exons within 5 kb of nearest SNP

Why SNPs?



- The majority of human sequence variation is due to substitutions that have occurred once in the history of mankind at individual base pairs, SNPs (Patil et al. 2001).
- Markers for pinpointing a disease
- Association study: check for differences in SNP patterns between cases and controls
- There can be big differences between populations!
- <http://snp.cshl.org/about/introduction.shtml>

Linkage disequilibrium

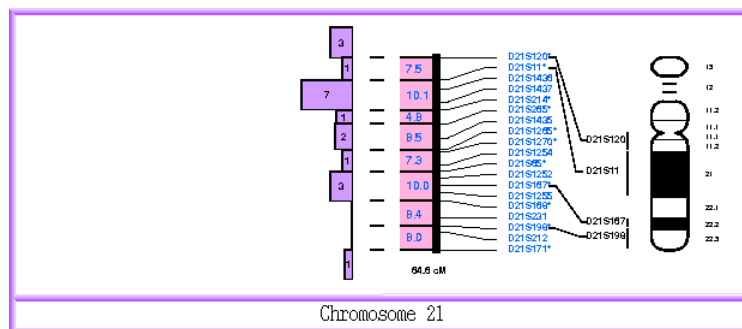


- Relationship between alleles at different loci.
- Alleles at locus A: frequencies p_1, \dots, p_m
- Alleles at locus B: frequencies q_1, \dots, q_n
- Haplotype frequency for $A_i B_j$ equilibrium value = $h_{ij} - p_i q_j$
- Linkage disequilibrium is an allelic association measure (difference between the actual haplotype frequency and the equilibrium value)
- More precisely: **gametic association**

Use of Polymorphism in Gene Mapping



- 1980s – RFLP marker maps
- 1990s – microsatellite marker maps



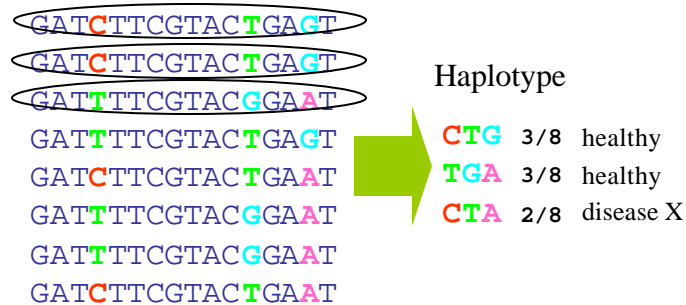
Advantages of SNPs in genetic analysis of complex traits



- Abundance: high frequency on the genome
- Position: throughout the genome (level of influence of type of SNP, e.g. coding region, promoter site, on phenotypic expression?)
- Haplotypic patterns (see later)
- Ease of genotyping
- Less mutable than other forms of polymorphisms
- Allele frequency drift (different populations)

Haplotype

-- a more discriminative state of a chromosomal region



chromosome

- Consider J binary markers in a genomic region
- There are 2^J possible haplotypes
 - but in fact, far fewer are seen in human population
- Good genetic marker for population, evolution and hereditary diseases ...

Haplotype analyses



- Linkage disequilibrium assessment
- Disease-gene discovery
- Genetic demography
- Chromosomal evolution studies

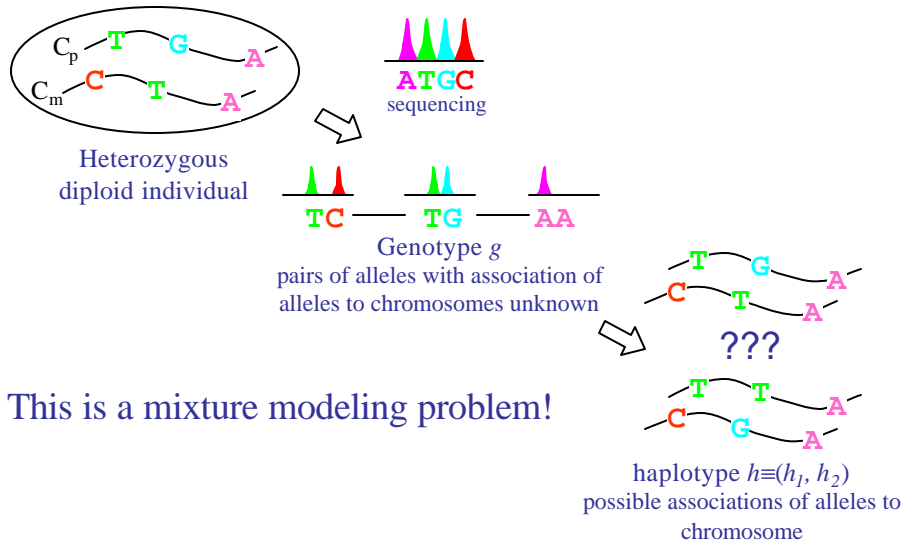
Why Haplotypes



- Haplotypes are more powerful discriminators between cases and controls in disease association studies
- Use of haplotypes in disease association studies reduces the number of tests to be carried out.
- With haplotypes we can conduct evolutionary studies
- Haplotypes are necessary for linkage analysis

Phase ambiguity

-- haplotype reconstruction for individuals



Inferring Haplotypes



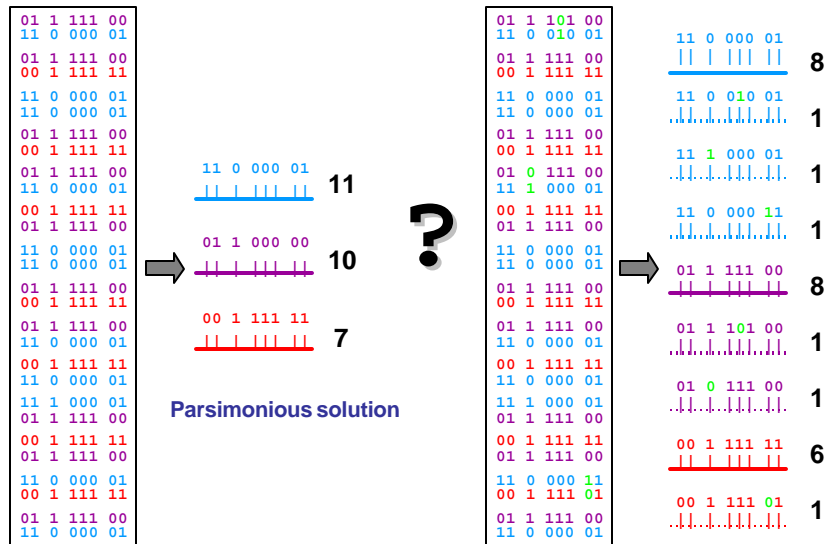
- Genotype: AT//AA//CG
 - Maternal genotype: TA//AA//CC
 - Paternal genotype: TT//AA//CG
 - Then the haplotype is AAC/TAG.
- Genotype: AT//AA//CG
 - Maternal genotype: AT//AA//CG
 - Paternal genotype: AT//AA//CG
 - Cannot determine unique haplotype
- **Problem:** determine Haplotypes without parental genotypes

Identifiability



Genotype representations	Genotypes of 14 individual
0/0 → 0	21 2 222 02
1/1 → 1	02 1 111 22
0/1 → 2	11 0 000 01
	02 1 111 22
	21 2 222 02
	02 1 111 22
	11 0 000 01
	02 1 111 22
	21 2 222 02
	22 2 222 21
	21 1 222 02
	02 1 111 22
	22 2 222 21
	21 2 222 02

Identifiability



Three Problems



1. Frequency estimation of all possible haplotypes
2. Haplotype reconstruction for individuals
3. How many out of all possible haplotypes are plausible in a population

given a random sample of multilocus genotypes at a set of SNPs

Haplotype reconstruction: Clark (1990)



- Choose individuals that are homozygous at every locus (e.g. TT//AA//CC)
 - Haplotype: TAC
- Choose individuals that are heterozygous at just one locus (e.g. TT//AA//CG)
 - Haplotypes: TAC or TAG
- Tally the resulting known haplotypes.
- For each known haplotype, look at all remaining unresolved cases: is there a combination to make this haplotype?
 - Known haplotype: TAC
 - Unresolved pattern: AT//AA//CG
 - Inferred haplotype: TAC/AAG. Add to list.
 - Known haplotype: TAC and TAG
 - Unresolved pattern: AT//AA//CG
 - Inferred haplotypes: TAC and TAG. Add both to list.
- Continue until all haplotypes have been recovered or no new haplotypes can be found this way.

Problems: Clark (1990)



- No homozygotes or single SNP heterozygotes in the sample
- Many unresolved haplotypes at the end
- Error in haplotype inference if a crossover of two actual haplotypes is identical to another true haplotype
- Frequency of these problems depend on avg. heterozygosity of the SNPs, number of loci, recombination rate, sample size.
- Clark (1990): algorithm "performs well" even with small sample sizes.

Finite mixture model



The probability of a genotype g :

$$p(g) = \sum_{h_1, h_2 \in H} p(h_1, h_2) p(g | h_1, h_2)$$

Population haplotype pool

Haplotype model

Genotyping model

Standard settings:

- $p(g|h_1, h_2) = \mathbf{1}(h_1 \oplus h_2 = g)$ noiseless genotyping
- $p(h_1, h_2) = p(h_1)p(h_2) = f_1 f_2$ Hardy-Weinberg equilibrium, multinomial
- $|H| = K$ fixed-sized population haplotype pool

$$p(g) = \sum_{\substack{h_1, h_2 \in H \\ h_1 \oplus h_2 = g}} f_1 f_2$$

EM algorithm: Excoffier and Slatkin (1995)



Numerical method of finding maximum likelihood estimates for parameters given incomplete data.

1. Initial parameter values: Haplotype frequencies: f_1, \dots, f_h
2. **Expectation step:** compute expected values of missing data based on initial data
3. **Maximization step:** compute MLE for parameters from the complete data
4. Repeat with new set of parameters until changes in the parameter estimates are negligible.

Beware: local maxima.

EM algorithm efficiency



- Heavy computational burden with large number of loci? (2^L possible haplotypes for L SNPs)
- Accuracy and departures from HWE?
- Error between EM-based frequency estimates and their true frequencies
- Sampling error vs. error from EM estimation process

Bayesian Haplotype reconstruction



- Bayesian model to approximate the posterior distribution of haplotype configurations for each phase-unknown genotype.
- $G = (G_1, \dots, G_n)$ observed multilocus genotype frequencies
- $H = (H_1, \dots, H_n)$ corresponding unknown haplotype pairs
- $F = (F_1, \dots, F_M)$ M unknown population haplotype frequencies
- EM algorithm: Find F that maximizes $P(G|F)$. Choose H that maximizes $P(H|F^{EM}, G)$.

Gibbs sampler



Initial haplotype reconstruction $H^{(0)}$.

- Choose an individual i , uniformly and at random from all ambiguous individuals.
- Sample $H_i^{(t+1)}$ from $P(H_i|G, H_{-i}^{(t)})$, where H_{-i} is the set of haplotypes excluding individual i .
- Set $H_j^{(t+1)} = H_j^{(t)}$ for $j=1, \dots, i-1, i+1, \dots, n$.

HAPLOTYPER: Bayesian Haplotype Inference (Niu et al.2002)



- Bayesian model to approximate the posterior distribution of haplotype configurations for each phase-unknown genotype.
- Dirichlet priors $\beta=(\beta_1, \dots, \beta_M)$ for the haplotype frequencies $F=(f_1, \dots, f_M)$.
- Multinomial model (as in EM algorithm) for individual haplotypes:
 - product over n individuals,
 - and multilocus genotype probabilities are sums of products of pairs of haplotype probabilities.

Gibbs sampler



- Haplotypes H are “missing:”

$$P(G, H | F) \sim \prod_{i=1, \dots, n} ? f_{h_{i1}} f_{h_{i2}} \prod_{j=1, \dots, n} ? f_j^{b_j - 1}$$

- Sample h_{i1} and h_{i2} for individual i :

$$P(h_{i1} = g, h_{i2} = h | F, G_i) = \frac{f_g f_h}{\sum_{g' h' \in G_i} f_{g'} f_{h'}}$$

- Sample H given H^{updated} Improving efficiency (Niu et al.)

Gibbs sampler



- **Predictive updating (Gibbs sampling):**

- (N(H))=vector of haplotype counts)

$$P(G,H) \sim \mathcal{G}(\beta+N(H)) / \mathcal{G}(\beta+N(H))$$

- Pick an individual i , update haplotype h_i :

$$P(h_i=(g,h)|H_{-i},G) \sim (n_g + \beta_g)(n_h + \beta_h)$$

(n_g =count of g in H_{-i})

- **Prior Annealing:**

- use high pseudo counts at the beginning of the iteration and progressively reduce them at a fixed rate as the sampler continues.

HAPLOTYPER Discussions



- **Missing marker data:**

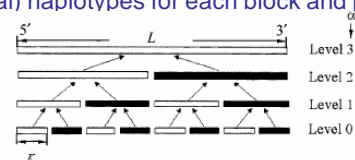
- PCR dropouts ->absence of both alleles,
 - one allele is unscored
 - Gibbs sampler adapts nicely

- **Ligation**

- Problem: large number of loci.
 - Partition L loci into blocks of 8 and carry out block level haplotype reconstruction.
 - Record the B most probable (partial) haplotypes for each block and join them

- **Progressive ligation.**

- **Hierarchical ligation.**



$$L = K \times 2^\alpha$$

Phase

coalescence-based Bayesian Haplotype inference:
Stephens et al (2001)



- What is $P(H_i | G, H_{-i}^{(t)})$?
- For a haplotype $H_i = (h_{i1}, h_{i2})$ consistent with genotypes G_i : $P(H_i | G, H_{-i}) \sim P(H_i | H_{-i}) \sim p(h_{i1} | H_{-i}) p(h_{i2} | h_{i1}, H_{-i})$
- $p(\cdot | H)$ = conditional distribution of a future sampled haplotype given previously sampled haplotypes H .
- r = total number of haplotypes, r_a = number of haplotypes of type a , θ = mutation rate, then a choice for

$$\pi(a | H) = (r_a + \theta \mu_a) / (r + \theta),$$

where μ_a = prob. of type a .

PHASE, details



- This is not working when the number of possible values H_i is too large: 2^{J-1} , J = number of loci at which individual i is heterozygous. Alternatively,

$$\pi(h | H) = \sum_{\alpha \in E} \sum_{s=0}^{\infty} \frac{r_{\alpha}}{r} \left(\frac{\theta}{r + \theta} \right)^s \frac{r}{r + \theta} (P^s)_{\alpha h}$$

where E = set of types for a general mutation model, P = reversible mutation matrix.

- I.e. future haplotype h is obtained by applying a random number of mutations, s (sampled from geometric distribution), to a randomly chosen existing haplotype, r_{α} (**coalescent**).
- Problems: estimation of θ , dimensionality of P ($\dim P = M$, the number of possible haplotypes).

PHASE Discussion



- Key: unresolved haplotypes are similar to known haplotypes
- HWE assumption, but robust to “moderate” levels of recombinations
- More accurate than EM, Clark’s and Haplotyper algorithms
- Provides estimates of the uncertainty associated with each phase call
- Problem (of both Bayesian model): dimensionality

Summary: Algorithms



Clark’s parsimony algorithm:

- simple, effective,
- depends on order of individuals in the data set,
- need sufficient number of homozygous individuals,
- Disadvantage: individuals may remain phase indeterminate, biased estimates of haplotype frequencies

EM algorithm:

- accurate in the inference of common haplotypes
- Allows for possible haplotype configurations that could contribute to a phase-unknown genotype.
- Cannot handle a large number of SNPs.

Summary: Algorithms



Haplotyper:

- Bayesian model to approximate the posterior distribution of haplotype configurations
- Prior annealing helps to escape from local maximum
- Partitions long haplotypes into small segments: block-by-block strategy
- Gibbs sampler to reconstruct haplotypes within each segment. Assembly of segments.
- <http://www.people.fas.harvard.edu/~junliu/index1.html#ComputationalBiology>

Summary: Algorithms



PHASE:

- Bayesian model to approximate the posterior distribution of haplotype configurations
- based on the coalescence theory to assign prior predictions about the distributions of haplotypes in natural populations,
- may depend on the order of the individuals,
- pseudo posterior probabilities (-> pseudo Gibbs sampler),
- lacks a measure of overall goodness.
- <http://www.hgmp.mrc.ac.uk/Registered/Option/phase.html>