Baysian Haplotype Inference via the Dirichlet Process

Eric Xing, Micheal Jordan, Roded Sharan

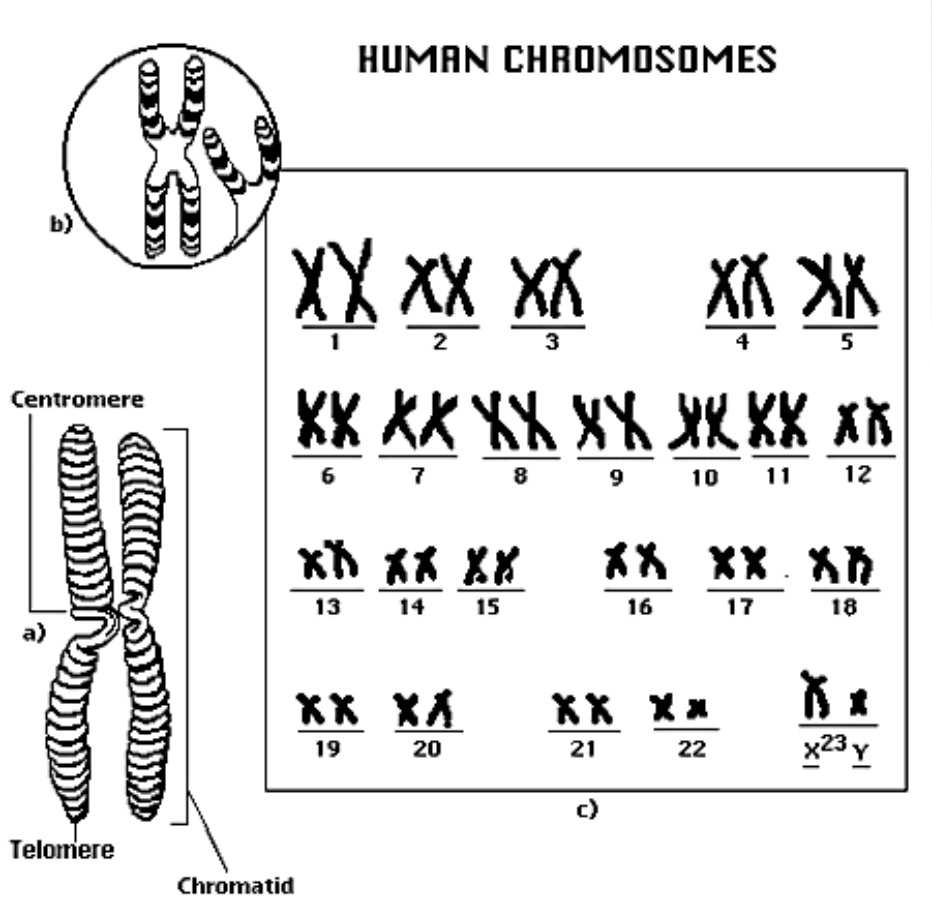
presented by
Amrudin Agovic

Motivation

- 99.9 % of human DNA shared
- 0.1% of DNA makes up for differences
- Need to determine what those 0.1% are
- Find genes responsible for diseases

Background

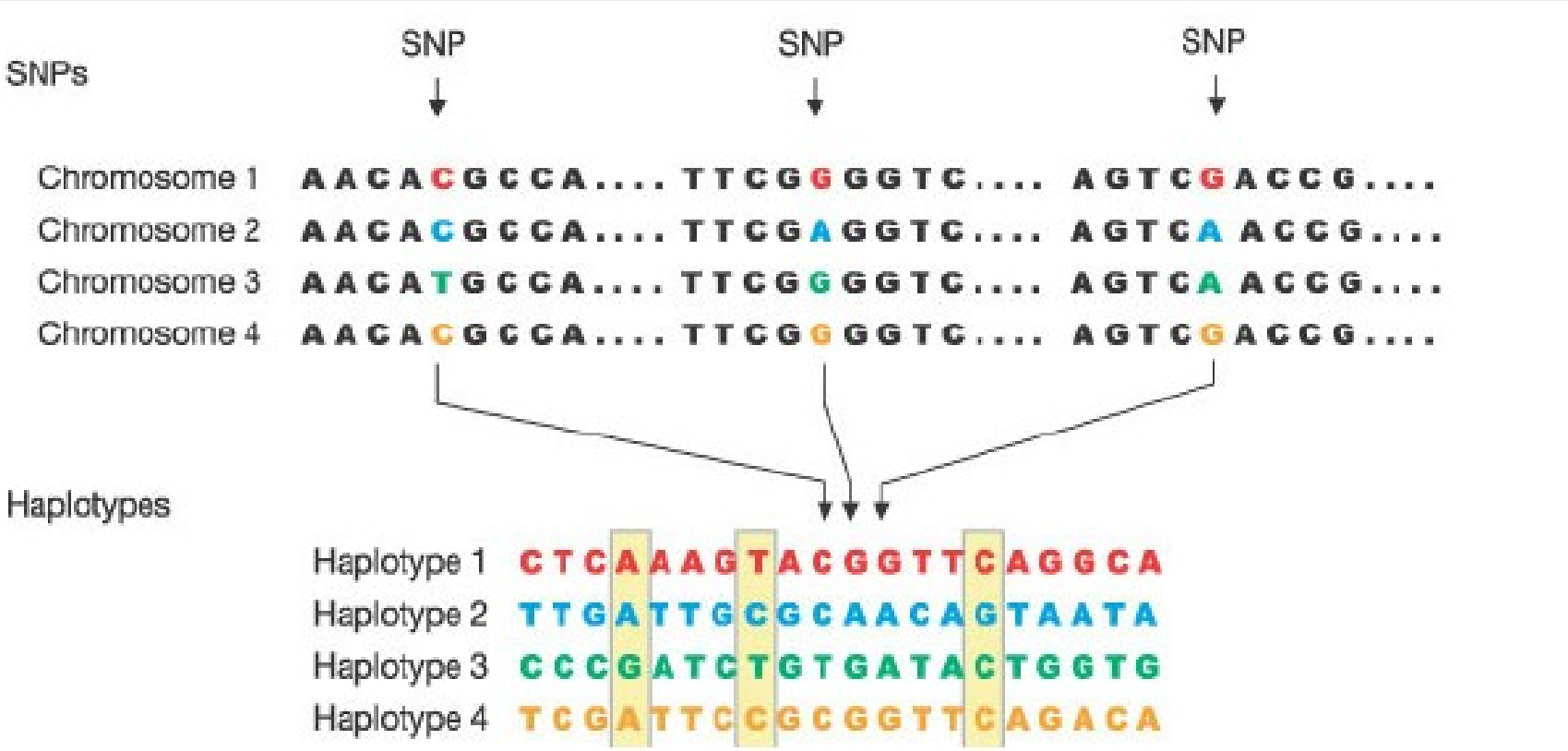
- Humans have 23 pairs of chromosomes in their cells
- 23 come from the father, 23 from the mother
- Certain parts of the genome are inherited unchanged
- Other genetic information gets mixed up

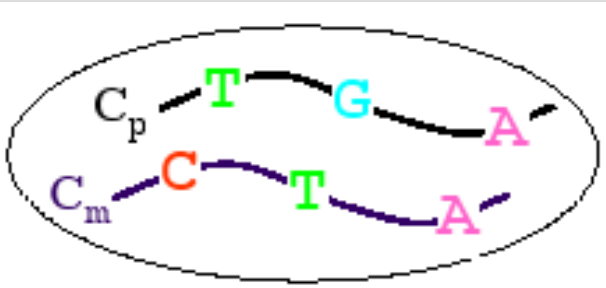


Background

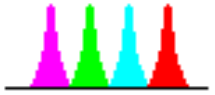
- **Allele:** genetic coding that occupies a position on the chromosome.
- **Genotype:** unordered pairs of Alleles in a region (one from each chromosome)
- **Phase:** Allele Chromosome association (not given)
- **SNP:** Single Nucleotide Polymorphism, difference in one nucleotide (A,C,G,T)
- **Haplotype:** set of associated SNP alleles in a region of a chromosome. A haplotype is inherited as a unit.

Background

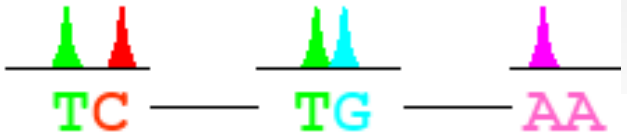

SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCAACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCAACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA



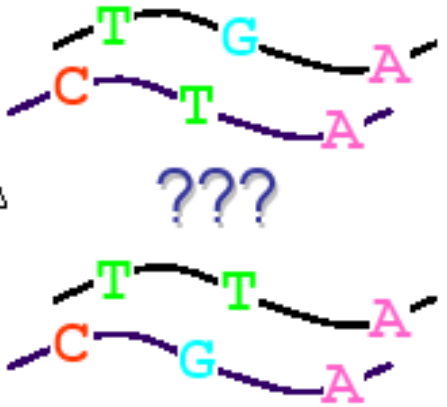
A heterozygous diploid individual



sequencing



The **Genotype:**
pairs of alleles with association of alleles to chromosomes unknown



Haplotype $h \equiv (h_1, h_2)$
possible associations of alleles to chromosome

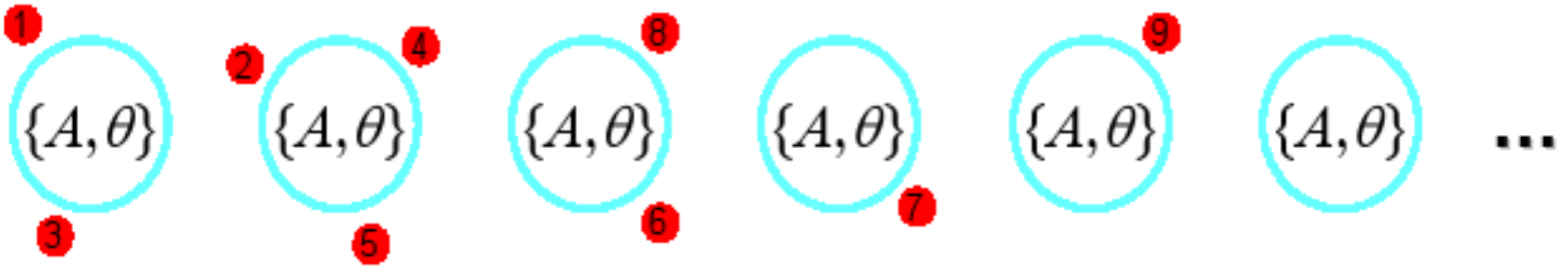
Dirichlet Process Representation

Let

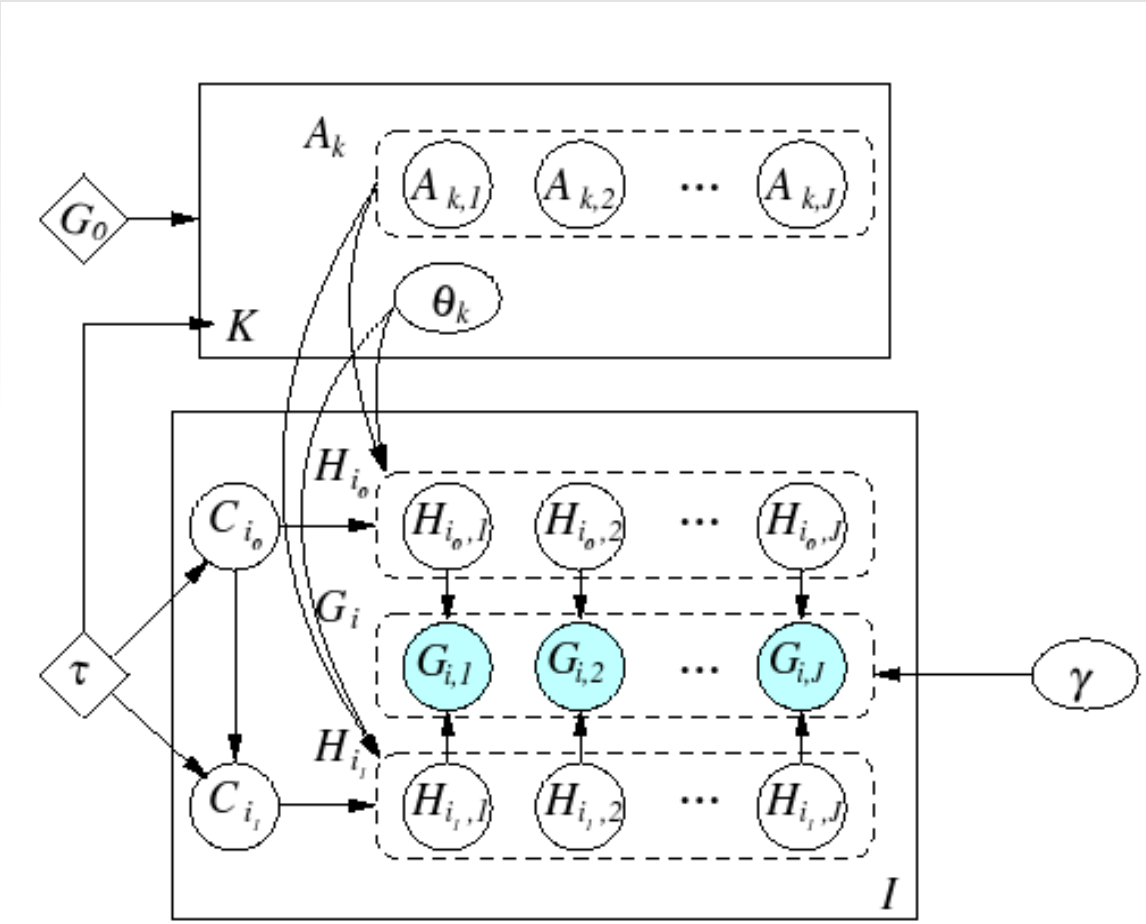
- $G_0(\Phi)$ be a base measure for the dirichlet process
- $A^{(k)} := [A_1^{(k)}, .., A_J^{(k)}]$ be a founding haplotype configuration (ancestral template) at loci t=[1,..,J]
- $\theta^{(k)}$ be the mutation rate of the ancestor
- Φ be the parameter associated with a mixture component. Where $\Phi_k = \{A^{(k)}, \theta^{(k)}\}$

Dirichlet Process Representation

- Use Chinese Restaurant Process
- Associate population haplotype with table
- Sample for each table $\Phi_k = \{A^{(k)}, \theta^{(k)}\}$



The Model


A_k
$A_{k,1}$
$A_{k,2}$
$\cdots$
$A_{k,J}$
G_0
θ_k
K
H_{i_o}
C_{i_o}
$H_{i_o,1}$
$H_{i_o,2}$
$\cdots$
$H_{i_o,J}$
G_i
τ
$G_{i,1}$
$G_{i,2}$
$\cdots$
$G_{i,J}$
γ
H_{i_1}
C_{i_1}
$H_{i_1,1}$
$H_{i_1,2}$
$\cdots$
$H_{i_1,J}$
I

Assumptions

- $G_0(A,\theta)=p(A)p(\theta)$
- p(A) uniform distribution over all haplotypes
- p(θ) is Beta(α_h,β_h)

$$
\begin{aligned}
& p(H_{i_t} = h | C_{i_t} = k, \mathbf{A} = \mathbf{a}, \boldsymbol{\theta}) \\
&= \quad p(H_{i_t} = h | A_k = a, \theta_k = \theta) \\
&= \quad \prod_j p(h_j | a_j, \theta),
\end{aligned}
$$

$$
p(h_j | a_j, \theta) = \theta^{\mathbb{I}(h_j = a_j)} \left(\frac{1-\theta}{|B|-1} \right)^{\mathbb{I}(h_j \neq a_j)}
$$

Distributions

Considering for all alleles mutations:

$$p(\mathbf{h}, \boldsymbol{\theta} | \mathbf{c}, \mathbf{a}) \propto \prod_k \theta_k^{m_k + \alpha_h - 1} \left(\frac{1 - \theta_k}{|B| - 1} \right)^{m'_k} \left[1 - \theta_k \right]^{\beta_h - 1}$$

Integrating out theta:

$$p(\mathbf{h}, \boldsymbol{\theta} | \mathbf{c}, \mathbf{a}) \propto \prod_k \theta_k^{m_k + \alpha_h - 1} \left(\frac{1 - \theta_k}{|B| - 1} \right)^{m'_k} \left[1 - \theta_k \right]^{\beta_h - 1}$$

Noisy Observation Model

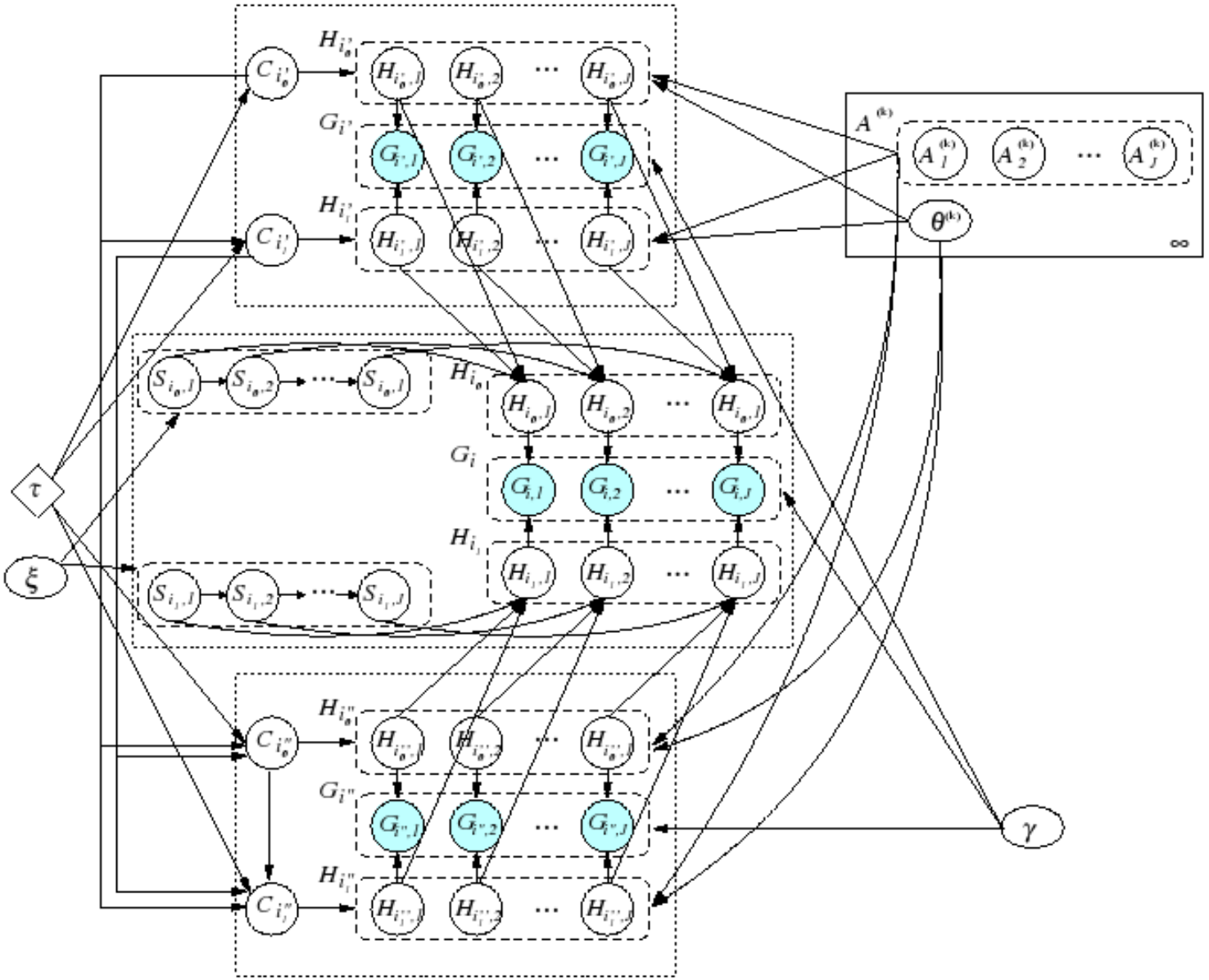
- Observed Genotype at a locus determined by parental and maternal alleles
- If genotype disagrees penalize

$$p(g_{i,j}|h_{i_0,j}, h_{i_1,j}, \gamma) = \gamma^{\mathbb{I}(h_{i,j}=g_{i,j})}[\mu_1(1-\gamma)]^{\mathbb{I}(h_{i,j}\overset{1}{\neq}g_{i,j})}[\mu_2(1-\gamma)]^{\mathbb{I}(h_{i,j}\overset{2}{\neq}g_{i,j})}$$

- γ has Beta prior

$$p(\mathbf{g}, \gamma|\mathbf{h}) = \prod_i p(g_i, \gamma|h_{i_0}, h_{i_1})$$

Pedigree-Haplotyper



Inference - Gibbs Sampling

- γ and θ integrated out
- Sample C_{it} , $A_j^{(k)}$, $H_{it,j}$

1) Given current hidden values of haplotypes sample c_{it} , $a_j^{(k)}$

$$
\begin{aligned}
& p(c_{i_t} = k \mid \mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a}) \\
\propto\ & p(c_{i_t} = k \mid \mathbf{c}_{[-i_t]}) \int p(h_{i_t} \mid c_{i_t} = k, \theta_k, a^{(k)}) p(\theta^{(k)} \mid \{h_{i'_{t'}} \ : \ i'_{t'} \neq i_t, c_{i'_{t'}} = k\}, a^{(k)}) d\theta^{(k)} \\
=\ & p(c_{i_t} = k \mid \mathbf{c}_{[-i_t]}) p(h_{i_t} \mid a^{(k)}, \mathbf{c}, \mathbf{h}_{[-i_t]}) \\
=\ & \begin{cases} \frac{n_{[-i_t],k}}{n-1+\tau} p(h_{i_t} \mid a^{(k)}, \mathbf{m}_{[-i_t],k}) & \text{if } k = c_{i'_{t'}} \text{ for some } i'_{t'} \neq i_t \\ \frac{\tau}{n-1+\tau} \sum_{a'} p(h_{i_t} \mid a') p(a') & \text{if } k \neq c_{i'_{t'}} \text{ for all } i'_{t'} \neq i_t \end{cases}
\end{aligned}
$$

Gibbs Sampling

$$p(a_j^{(k)} \,|\{h_{i_t,j} \;:\; c_{i_t} = k\}) =$$

$$\begin{cases} \frac{1}{Z} p(h_{i_t,j}|a_j^{(k)}) & \\ = \left(\frac{\alpha_h}{\alpha_h+\beta_h}\right)^{\mathbb{I}(h_{i_t,j}=a_j^{(k)})} \left(\frac{\beta_h}{(|B|-1)(\alpha_h+\beta_h)}\right)^{\mathbb{I}(h_{i_t,j}\neq a_j^{(k)})} & \text{if } k \text{ is not previously instantiated} \\ \frac{1}{Z} p(\{h_{i_t,j} \;:\; c_{i_t} = k\}|a_j^{(k)}) & \\ = \frac{1}{Z} \frac{\Gamma(\alpha_h+m_{k,j})\Gamma(\beta_h+m'_{k,j})}{\Gamma(\alpha_h+\beta_h+n_k)\cdot(|B|-1)^{m'_{k,j}}} & \\ = \frac{\Gamma(\alpha_h+m_{k,j})\Gamma(\beta_h+m'_{k,j})/(|B|-1)^{m'_{k,j}}}{\sum_{l\in B}\Gamma(\alpha_h+m_{k,j}(l))\Gamma(\beta_h+m'_{k,j}(l))/(|B|-1)^{m'_{k,j}(l)}} & \text{if } k \text{ is previously instantiated,} \end{cases}$$

2) Given ancestral assignment and ancestral pool sample haplotype

$$\begin{aligned} & p(h_{i_t,j}|\mathbf{h}_{[-(i,j)]}, h_{i_{\bar{t}},j}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \\ \propto \;& p(g_i|h_{i_t,j}, h_{i_{\bar{t}},j}, \mathbf{u}_{[-(i,j)]}) p(h_{i_t,j}|a_j^{(k)}, \mathbf{m}_{[-(i_t,j)],k}) \\ = \;& R_g \frac{\Gamma(\alpha_g+u)\Gamma(\beta_g+(u'+u''))}{\Gamma(\alpha_g+\beta_g+IJ)} [\mu_1]^{u'} [\mu_2]^{u''} \times \\ & R_h \frac{\Gamma(\alpha_h+m_{i_t,k,j})\Gamma(\beta_h+m'_{i_t,k,j})}{\Gamma(\alpha_h+\beta_h+n_k)\cdot(|B|-1)^{m'_{i_t,k,j}}}, \end{aligned}$$

Metropolis Hastings

- Long list of loci and uniform prior p(a), leaves probability of sampling new ancestor very small.
- Slow mixing
- Sample ancestor assignment using proposal distribution

$$q(c^*_{i_t} = k | c_{[-i_t]}) = \begin{cases} \frac{n_{[-i_t],k}}{n-1+\tau} & : \text{ if } \; k = c_{i'_{t'}} \text{ for some } i'_{t'} \neq i_t \\ \frac{\tau}{n-1+\tau} & : \text{ if } \; k \neq c_{i'_{t'}} \text{ for all } i'_{t'} \neq i_t \end{cases}$$

Metropolis Hastings

- In acceptance probability, the proposal factor cancels out

$$\begin{aligned}
\frac{q(c_{i_t}|\mathbf{c}_{[-i_t]})}{q(c^*_{i_t}|\mathbf{c}_{[-i_t]})}\frac{\pi(c^*_{i_t})}{\pi(c_{i_t})} &= \frac{q(c_{i_t}|\mathbf{c}_{[-i_t]})}{q(c^*_{i_t}|\mathbf{c}_{[-i_t]})}\frac{p(c^*_{i_t}|\mathbf{c}_{[-i_t]},\mathbf{h},\mathbf{a})}{p(c_{i_t}|\mathbf{c}_{[-i_t]},\mathbf{h},\mathbf{a})} \\
&= \frac{q(c_{i_t}|\mathbf{c}_{[-i_t]})}{q(c^*_{i_t}|\mathbf{c}_{[-i_t]})}\frac{p(c^*_{i_t}|\mathbf{c}_{[-i_t]})p(h_{i_t}|a^{(c^*_{i_t})},\mathbf{c},\mathbf{h}_{[-i_t]})}{p(c_{i_t}|\mathbf{c}_{[-i_t]})p(h_{i_t}|a^{(c_{i_t})},\mathbf{c},\mathbf{h}_{[-i_t]})} \\
&= \frac{p(h_{i_t}|a^{(c^*_{i_t})},\mathbf{c},\mathbf{h}_{[-i_t]})}{p(h_{i_t}|a^{(c_{i_t})},\mathbf{c},\mathbf{h}_{[-i_t]})},
\end{aligned}$$

$$\xi(c^*_{i_t}, c_{i_t}) = \min\left[1, \frac{p(h_{i_t}|a^{c^*_{i_t}},\mathbf{c},\mathbf{h}_{[-i_t]})}{p(h_{i_t}|a^{c_{i_t}},\mathbf{c},\mathbf{h}_{[-i_t]})}\right]$$

Experiments

- Simulated Data: Haplotypes randomly paired to form genotypes.
- Performance compared to PHASE

	DP(MH)			PHASE			EM
#individuals	err_s	err_i	d_s	err_s	err_i	d_s	err_i
10	0.060	0.216	0.051	0.046	0.182	0.054	0.424
20	0.039	0.152	0.039	0.029	0.136	0.046	0.296
30	0.036	0.121	0.038	0.024	0.101	0.027	0.231
40	0.030	0.094	0.029	0.019	0.071	0.026	0.195
50	0.028	0.082	0.024	0.019	0.072	0.025	0.167
Average	**0.039**	**0.133**	**0.036**	**0.027**	**0.112**	**0.036**	**0.263**

Experiments

- Two real data sets: 129 individuals, 90 individuals from 4 populations

Dataset 1:

		DP(Gibbs)			DP(MH)			PHASE			HAP	HAPLOTYPER
block id.	length	err_s	err_i	d_s	err_s	err_i	d_s	err_s	err_i	d_s	err_s	err_s
1	14	0.223	0.485	0.229	0	0	0	0.003	0.030	0.003	0.007	0.039
2	5	0	0	0	0.007	0.026	0.007	0.007	0.026	0.007	0.036	0.065
3	5	0	0	0	0	0	0	0	0	0	0	0.008
4	11	0.143	0.262	0.128	0	0	0	0	0	0	0.015	-
5	9	0.020	0.066	0.020	0.011	0.033	0.011	0.011	0.033	0.011	0.027	0.151
6	27	0.071	0.191	0.074	0.005	0.043	0.005	0	0	0	0.018	0.041
7	7	0.005	0.018	0.005	0.005	0.018	0.005	0.005	0.018	0.005	0.068	0.214
8	4	0	0	0	0	0	0	0	0	0	0	0.252
9	5	0.029	0.097	0.029	0.012	0.032	0.012	0.012	0.032	0.012	0.057	0.152
10	4	0.007	0.025	0.007	0.007	0.025	0.007	0.008	0.025	0.008	0.042	0.056
11	7	0.010	0.034	0.005	0.005	0.017	0.005	0.011	0.034	0.011	0.033	0.093
12	5	0.010	0.037	0.020	0	0	0	0	0	0	0	0.077
Average	**8.58**	**0.043**	**0.101**	**0.043**	**0.004**	**0.016**	**0.004**	**0.005**	**0.017**	**0.005**	**0.025**	**0.104**

Experiments

Dataset 2:

- Small sample size, tougher data set
- Haplotyper outperforms PHASE

		DP(MH)			PHASE		
region	length	err_s	err_i	d_s	err_s	err_i	d_s
16a	13	0.185	0.480	0.141	0.174	0.440	0.130
1b	16	0.100	0.250	0.160	0.200	0.450	0.180
25a	14	0.135	0.353	0.115	0.212	0.588	0.212
7b	13	0.105	0.278	0.066	0.145	0.444	0.092
Average	**14**	**0.131**	**0.340**	**0.121**	**0.183**	**0.481**	**0.154**

Conclusions

- Algorithm outperform PHASE on two data sets With a big margin on one of them.
- Strength of proposed approach in flexibility
- Can be extended to incorporate aspects of evolutionary dynamics and other things
- Illustrated example: Pedigree information