CSci 8980: Advanced Topics in Graphical Models

Analysis of Genetic Variation

Instructor: Arindam Banerjee

November 26, 2007

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus
  - Nucleotide can be one of $\{A, C, T, G\}$

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus
  - Nucleotide can be one of $\{A, C, T, G\}$
  - Most genetic human variation are related to SNPs

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus
  - Nucleotide can be one of $\{A, C, T, G\}$
  - Most genetic human variation are related to SNPs
  - Each variant is called an *allele*

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus
  - Nucleotide can be one of $\{A, C, T, G\}$
  - Most genetic human variation are related to SNPs
  - Each variant is called an *allele*
- *Haplotype*

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus
  - Nucleotide can be one of $\{A, C, T, G\}$
  - Most genetic human variation are related to SNPs
  - Each variant is called an *allele*
- *Haplotype*
  - List of alleles in a local region of a chromosome

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus
  - Nucleotide can be one of $\{A, C, T, G\}$
  - Most genetic human variation are related to SNPs
  - Each variant is called an *allele*
- *Haplotype*
  - List of alleles in a local region of a chromosome
  - Inherited as a unit, if there is no recombination

Genetic Polymorphism

- Single nucleotide polymorphism (SNP)
  - Two possible kinds of nucleotides at a single locus
  - Nucleotide can be one of $\{A, C, T, G\}$
  - Most genetic human variation are related to SNPs
  - Each variant is called an *allele*
- *Haplotype*
  - List of alleles in a local region of a chromosome
  - Inherited as a unit, if there is no recombination
- Repeated recombinations between ancestral haplotypes

Genetic Polymorphism (Contd.)

- *Linkage disequilibrium* (LD)

Genetic Polymorphism (Contd.)

- *Linkage disequilibrium* (LD)
  - Non-random association of alleles at different loci

Genetic Polymorphism (Contd.)

- *Linkage disequilibrium* (LD)
  - Non-random association of alleles at different loci
  - Recombination decouples alleles, increase randomness, decrease LD

Genetic Polymorphism (Contd.)

- *Linkage disequilibrium* (LD)
  - Non-random association of alleles at different loci
  - Recombination decouples alleles, increase randomness, decrease LD
- Infer chromosomal recombination hotspots

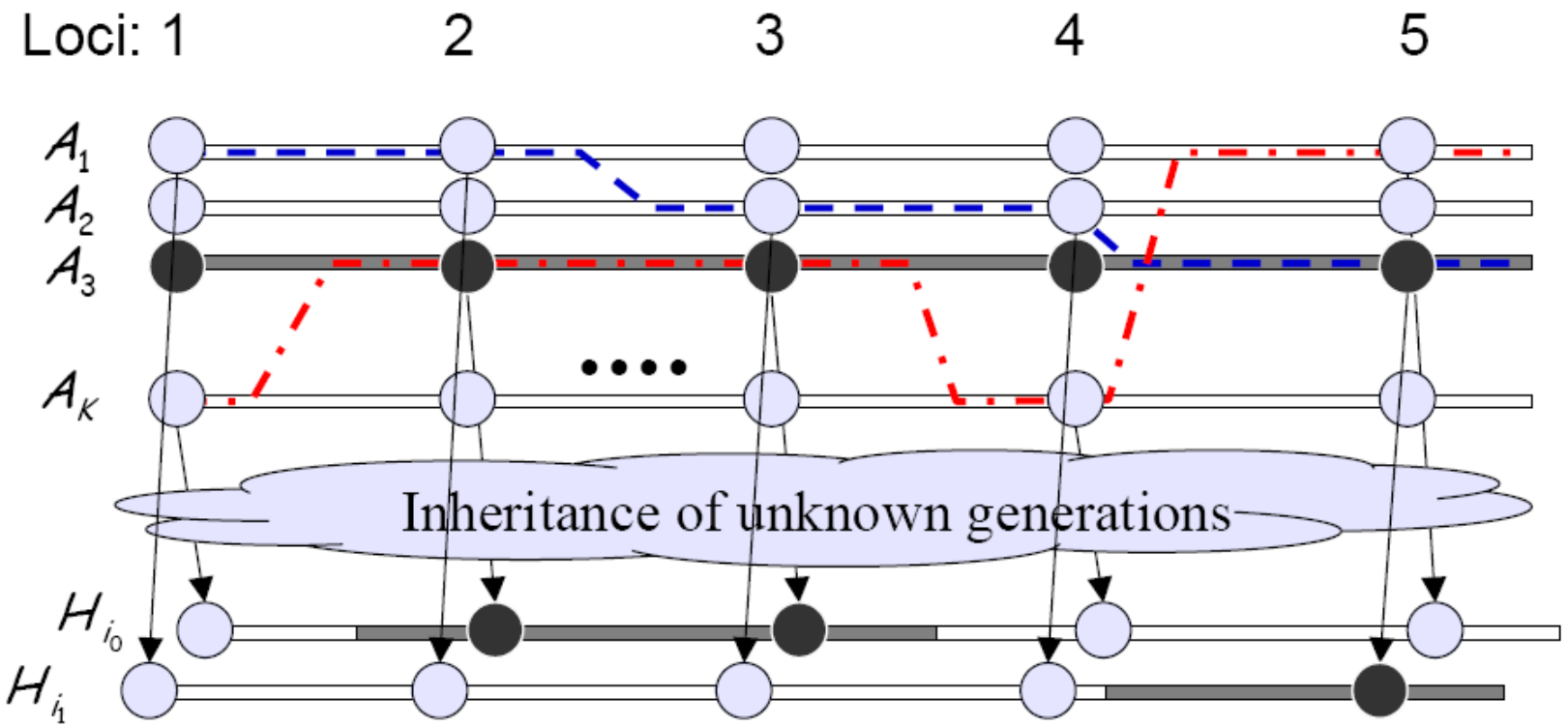
Genetic Polymorphism (Contd.)

- *Linkage disequilibrium* (LD)
  - Non-random association of alleles at different loci
  - Recombination decouples alleles, increase randomness, decrease LD
- Infer chromosomal recombination hotspots
  - Help understand origin and characteristics of genetic variation

Genetic Polymorphism (Contd.)

- *Linkage disequilibrium* (LD)
  - Non-random association of alleles at different loci
  - Recombination decouples alleles, increase randomness, decrease LD
- Infer chromosomal recombination hotspots
  - Help understand origin and characteristics of genetic variation
- Analyze genetic variation to reconstruct evolutionary history

Haplotype Recombination and Inheritance



Hidden Markov Process

- Generative model for choosing recombination sites

Hidden Markov Process

- Generative model for choosing recombination sites
- Hidden Markov process

Hidden Markov Process

- Generative model for choosing recombination sites
- Hidden Markov process
  - Hidden states correspond to index over chromosomes

Hidden Markov Process

- Generative model for choosing recombination sites
- Hidden Markov process
  - Hidden states correspond to index over chromosomes
  - Transition probabilities correspond to recombination rates

Hidden Markov Process

- Generative model for choosing recombination sites
- Hidden Markov process
  - Hidden states correspond to index over chromosomes
  - Transition probabilities correspond to recombination rates
  - Emission model corresponds to mutation process that give descendants

Hidden Markov Process

- Generative model for choosing recombination sites
- Hidden Markov process
  - Hidden states correspond to index over chromosomes
  - Transition probabilities correspond to recombination rates
  - Emission model corresponds to mutation process that give descendants
- Implemented using a Hidden Markov Dirichlet Process (HMDP)

Dirichlet Process Mixtures

- We know the basics of DPMs

Dirichlet Process Mixtures

- We know the basics of DPMs
- Haplotype modeling using an infinite mixture model

Dirichlet Process Mixtures

- We know the basics of DPMs
- Haplotype modeling using an infinite mixture model
  - A pool of ancestor haplotypes or founders

Dirichlet Process Mixtures

- We know the basics of DPMs
- Haplotype modeling using an infinite mixture model
  - A pool of ancestor haplotypes or founders
  - The size of the pool is unknown

Dirichlet Process Mixtures

- We know the basics of DPMs
- Haplotype modeling using an infinite mixture model
  - A pool of ancestor haplotypes or founders
  - The size of the pool is unknown
- Standard coalescence based models

Dirichlet Process Mixtures

- We know the basics of DPMs
- Haplotype modeling using an infinite mixture model
  - A pool of ancestor haplotypes or founders
  - The size of the pool is unknown
- Standard coalescence based models
  - Hidden variables is prohibitively large

Dirichlet Process Mixtures

- We know the basics of DPMs
- Haplotype modeling using an infinite mixture model
  - A pool of ancestor haplotypes or founders
  - The size of the pool is unknown
- Standard coalescence based models
  - Hidden variables is prohibitively large
  - Hard to perform inference of ancestral features

Dirichlet Process Mixtures (Contd.)

- $H_i = [H_{i,1}, \ldots, H_{i,T}]$ haplotype over T SNPs, chromosome i

Dirichlet Process Mixtures (Contd.)

- $H_i = [H_{i,1}, \ldots, H_{i,T}]$ haplotype over T SNPs, chromosome i
- $A_k = [A_{k,1}, \ldots, A_{k,T}]$ ancestral haplotype, mutation rate θ_k

Dirichlet Process Mixtures (Contd.)

- $H_i = [H_{i,1}, \ldots, H_{i,T}]$ haplotype over T SNPs, chromosome i
- $A_k = [A_{k,1}, \ldots, A_{k,T}]$ ancestral haplotype, mutation rate θ_k
- C_i, inheritance variable, latent ancestor of H_i

Dirichlet Process Mixtures (Contd.)

- $H_i = [H_{i,1}, \ldots, H_{i,T}]$ haplotype over T SNPs, chromosome i
- $A_k = [A_{k,1}, \ldots, A_{k,T}]$ ancestral haplotype, mutation rate θ_k
- C_i, inheritance variable, latent ancestor of H_i
- Generative Model:

Dirichlet Process Mixtures (Contd.)

- $H_i = [H_{i,1}, \ldots, H_{i,T}]$ haplotype over T SNPs, chromosome i
- $A_k = [A_{k,1}, \ldots, A_{k,T}]$ ancestral haplotype, mutation rate θ_k
- C_i, inheritance variable, latent ancestor of H_i
- Generative Model:
  - Draw a first haplotype

$$\begin{aligned} a_1 | DP(\tau, Q_0) &\sim Q_0 \\ h_1 &\sim P_h(\cdot | a_1, \theta_1) \end{aligned}$$

Dirichlet Process Mixtures (Contd.)

- $H_i = [H_{i,1}, \ldots, H_{i,T}]$ haplotype over T SNPs, chromosome i
- $A_k = [A_{k,1}, \ldots, A_{k,T}]$ ancestral haplotype, mutation rate θ_k
- C_i, inheritance variable, latent ancestor of H_i
- Generative Model:
  - Draw a first haplotype

$$\begin{aligned} a_1 | DP(\tau, Q_0) &\sim Q_0 \\ h_1 &\sim P_h(\cdot | a_1, \theta_1) \end{aligned}$$

  - For subsequent haplotypes

$$c_i | DP(\tau, Q_0) \sim \begin{cases} p(c_i = c_j \text{for some } j < i | c_1, \ldots, c_{i-1}) = \frac{n_{c_j}}{i-1+\alpha_0} \\ p(c_i \neq c_j \text{for all } j < i | c_1, \ldots, c_{i-1}) = \frac{\alpha_0}{i-1+\alpha_0} \end{cases}$$

Dirichlet Process Mixtures (Contd.)

- Generative Model (contd)

Dirichlet Process Mixtures (Contd.)

- Generative Model (contd)
  - Sample the founder of haplotype i

$$\phi_{c_i} | DP(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} \text{if} c_i = c_j \text{for some} j < i \\ \sim Q(a, \theta) \text{if} c_i \neq c_j \text{for all} j < i \end{cases}$$

Dirichlet Process Mixtures (Contd.)

- Generative Model (contd)
  - Sample the founder of haplotype i

$$\phi_{c_i}|DP(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} \text{if} c_i = c_j \text{for some} j < i \\ \sim Q(a, \theta) \text{if} c_i \neq c_j \text{for all} j < i \end{cases}$$

  - Sample the haplotype according to its founder

$$h_i|c_i \sim P(\cdot|a_{c_i}, \theta_{c_i})$$

Dirichlet Process Mixtures (Contd.)

- Generative Model (contd)
  - Sample the founder of haplotype i

$$\phi_{c_i}|DP(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} \text{if} c_i = c_j \text{for some} j < i \\ \sim Q(a, \theta) \text{if} c_i \neq c_j \text{for all} j < i \end{cases}$$

  - Sample the haplotype according to its founder

$$h_i|c_i \sim P(\cdot|a_{c_i}, \theta_{c_i})$$

- Assumes each haplotype originates from one ancestor

Dirichlet Process Mixtures (Contd.)

- Generative Model (contd)
  - Sample the founder of haplotype i

$$\phi_{c_i}|DP(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} \text{if} c_i = c_j \text{for some} j < i \\ \sim Q(a, \theta) \text{if} c_i \neq c_j \text{for all} j < i \end{cases}$$

  - Sample the haplotype according to its founder

$$h_i|c_i \sim P(\cdot|a_{c_i}, \theta_{c_i})$$

- Assumes each haplotype originates from one ancestor
  - Valid only for short regions in chromosome

Dirichlet Process Mixtures (Contd.)

- Generative Model (contd)
  - Sample the founder of haplotype i

$$\phi_{c_i}|DP(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} \text{if} c_i = c_j \text{for some} j < i \\ \sim Q(a, \theta) \text{if} c_i \neq c_j \text{for all} j < i \end{cases}$$

  - Sample the haplotype according to its founder

$$h_i|c_i \sim P(\cdot|a_{c_i}, \theta_{c_i})$$

- Assumes each haplotype originates from one ancestor
  - Valid only for short regions in chromosome
  - Long regions will have recombination

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space
- Conditioned on each state, sample a DP with the same support

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space
- Conditioned on each state, sample a DP with the same support
- Hierarchical Urns

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space
- Conditioned on each state, sample a DP with the same support
- Hierarchical Urns
  - Stock urn Q_0 with balls of K colors, n_k of color k

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space
- Conditioned on each state, sample a DP with the same support
- Hierarchical Urns
  - Stock urn Q_0 with balls of K colors, n_k of color k
  - HMM-urns $Q_1, \ldots, Q_K$ for prior and transition probabilities

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space
- Conditioned on each state, sample a DP with the same support
- Hierarchical Urns
  - Stock urn Q_0 with balls of K colors, n_k of color k
  - HMM-urns $Q_1, \ldots, Q_K$ for prior and transition probabilities
  - Let $m_{j,k}$ be the number of balls of color k in urn Q_j

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space
- Conditioned on each state, sample a DP with the same support
- Hierarchical Urns
  - Stock urn Q_0 with balls of K colors, n_k of color k
  - HMM-urns $Q_1, \ldots, Q_K$ for prior and transition probabilities
  - Let $m_{j,k}$ be the number of balls of color k in urn Q_j
  - HDPM can be simulated by sampling from the urn hierarchy

Hidden Markov Dirichlet Process

- Nonparametric Bayesian HMM
- Sample a DP to form the support of the infinite state space
- Conditioned on each state, sample a DP with the same support
- Hierarchical Urns
  - Stock urn Q_0 with balls of K colors, n_k of color k
  - HMM-urns $Q_1, \ldots, Q_K$ for prior and transition probabilities
  - Let $m_{j,k}$ be the number of balls of color k in urn Q_j
  - HDPM can be simulated by sampling from the urn hierarchy
- Hierarchical DPM

$$\begin{aligned} Q_0|\alpha, F &\sim DP(\alpha, F) \\ Q_j|\tau, Q_0 &\sim DP(\tau, Q_0) \end{aligned}$$

Hidden Markov Dirichlet Process (Contd.)

- Each color corresponds to ancestor configuration $\phi_k = \{a_k, \theta_k\}$

Hidden Markov Dirichlet Process (Contd.)

- Each color corresponds to ancestor configuration $\phi_k = \{a_k, \theta_k\}$
- For n random draws from Q_0

$$\phi_n | \phi_{-n} \sim \sum_{k=1}^{K} \frac{n_k}{n-1+\alpha} \delta_{\phi_k}(\phi_n) + \frac{\alpha}{n-1+\alpha} F(\phi_n)$$

Hidden Markov Dirichlet Process (Contd.)

- Each color corresponds to ancestor configuration $\phi_k = \{a_k, \theta_k\}$
- For n random draws from Q_0

$$\phi_n|\phi_{-n} \sim \sum_{k=1}^{K} \frac{n_k}{n-1+\alpha}\delta_{\phi_k}(\phi_n) + \frac{\alpha}{n-1+\alpha}F(\phi_n)$$

- Conditioned on Q_0, the marginal configs from Q_j

$$\phi_{m_j}|\phi_{-m_j} \sim \sum_{k} \frac{m_{j,k} + \tau\frac{n_k}{n-1+\alpha}}{m_j - 1 + tau} + \frac{\tau}{m_j - 1 + \tau}\frac{\alpha}{n-1+\alpha}F(\phi_{m_j})$$

Hidden Markov Dirichlet Process (Contd.)

- Each color corresponds to ancestor configuration $\phi_k = \{a_k, \theta_k\}$
- For n random draws from Q_0

$$\phi_n|\phi_{-n} \sim \sum_{k=1}^{K} \frac{n_k}{n-1+\alpha}\delta_{\phi_k}(\phi_n) + \frac{\alpha}{n-1+\alpha}F(\phi_n)$$

- Conditioned on Q_0, the marginal configs from Q_j

$$\phi_{m_j}|\phi_{-m_j} \sim \sum_{k} \frac{m_{j,k} + \tau\frac{n_k}{n-1+\alpha}}{m_j - 1 + tau} + \frac{\tau}{m_j - 1 + \tau}\frac{\alpha}{n-1+\alpha}F(\phi_{m_j})$$

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters
  $F(A, \theta) = p(A)p(\theta)$

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters $F(A, \theta) = p(A)p(\theta)$
- $p(A)$ is assumed uniform, $p(\theta)$ is assumed beta

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters $F(A, \theta) = p(A)p(\theta)$
- $p(A)$ is assumed uniform, $p(\theta)$ is assumed beta
- $C_i = [C_{i,1}, \ldots, C_{i,T}]$ ancestral index for chromosome i

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters $F(A, \theta) = p(A)p(\theta)$
- $p(A)$ is assumed uniform, $p(\theta)$ is assumed beta
- $C_i = [C_{i,1}, \ldots, C_{i,T}]$ ancestral index for chromosome i
- With no recombination, $C_{i,t} = k, \forall t$ for some k

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters $F(A, \theta) = p(A)p(\theta)$
- $p(A)$ is assumed uniform, $p(\theta)$ is assumed beta
- $C_i = [C_{i,1}, \ldots, C_{i,T}]$ ancestral index for chromosome i
- With no recombination, $C_{i,t} = k, \forall t$ for some k
- Non-recombination is modeled by Poisson point process

$$P(C_{i,t+1} = C_{i,t} = k) = \exp(-dr) + (1 - \exp(-dr))\pi_{kk}$$

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters $F(A, \theta) = p(A)p(\theta)$
- $p(A)$ is assumed uniform, $p(\theta)$ is assumed beta
- $C_i = [C_{i,1}, \ldots, C_{i,T}]$ ancestral index for chromosome i
- With no recombination, $C_{i,t} = k, \forall t$ for some k
- Non-recombination is modeled by Poisson point process

$$P(C_{i,t+1} = C_{i,t} = k) = \exp(-dr) + (1 - \exp(-dr))\pi_{kk}$$

  - d is the distance between the two loci

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters $F(A, \theta) = p(A)p(\theta)$
- $p(A)$ is assumed uniform, $p(\theta)$ is assumed beta
- $C_i = [C_{i,1}, \ldots, C_{i,T}]$ ancestral index for chromosome i
- With no recombination, $C_{i,t} = k, \forall t$ for some k
- Non-recombination is modeled by Poisson point process

$$P(C_{i,t+1} = C_{i,t} = k) = \exp(-dr) + (1 - \exp(-dr))\pi_{kk}$$

  - d is the distance between the two loci
  - r is the rate of recombination per unit distance

HMDP for Recombination and Inheritance

- Priors for the conditional model parameters $F(A, \theta) = p(A)p(\theta)$
- $p(A)$ is assumed uniform, $p(\theta)$ is assumed beta
- $C_i = [C_{i,1}, \ldots, C_{i,T}]$ ancestral index for chromosome i
- With no recombination, $C_{i,t} = k, \forall t$ for some k
- Non-recombination is modeled by Poisson point process

$$P(C_{i,t+1} = C_{i,t} = k) = \exp(-dr) + (1 - \exp(-dr))\pi_{kk}$$

  - d is the distance between the two loci
  - r is the rate of recombination per unit distance
- The transition probability to state k' is

$$P(C_{i,t} = k, C_{i,t+1} = k') = (1 - \exp(dr))\pi_{kk'}$$

HMDP for Recombination and Inheritance (Contd.)

- H_i is a mosaic of multiple ancestral chromosomes

HMDP for Recombination and Inheritance (Contd.)

- H_i is a mosaic of multiple ancestral chromosomes
- Model is a time-inhomogenous infinite HMM

HMDP for Recombination and Inheritance (Contd.)

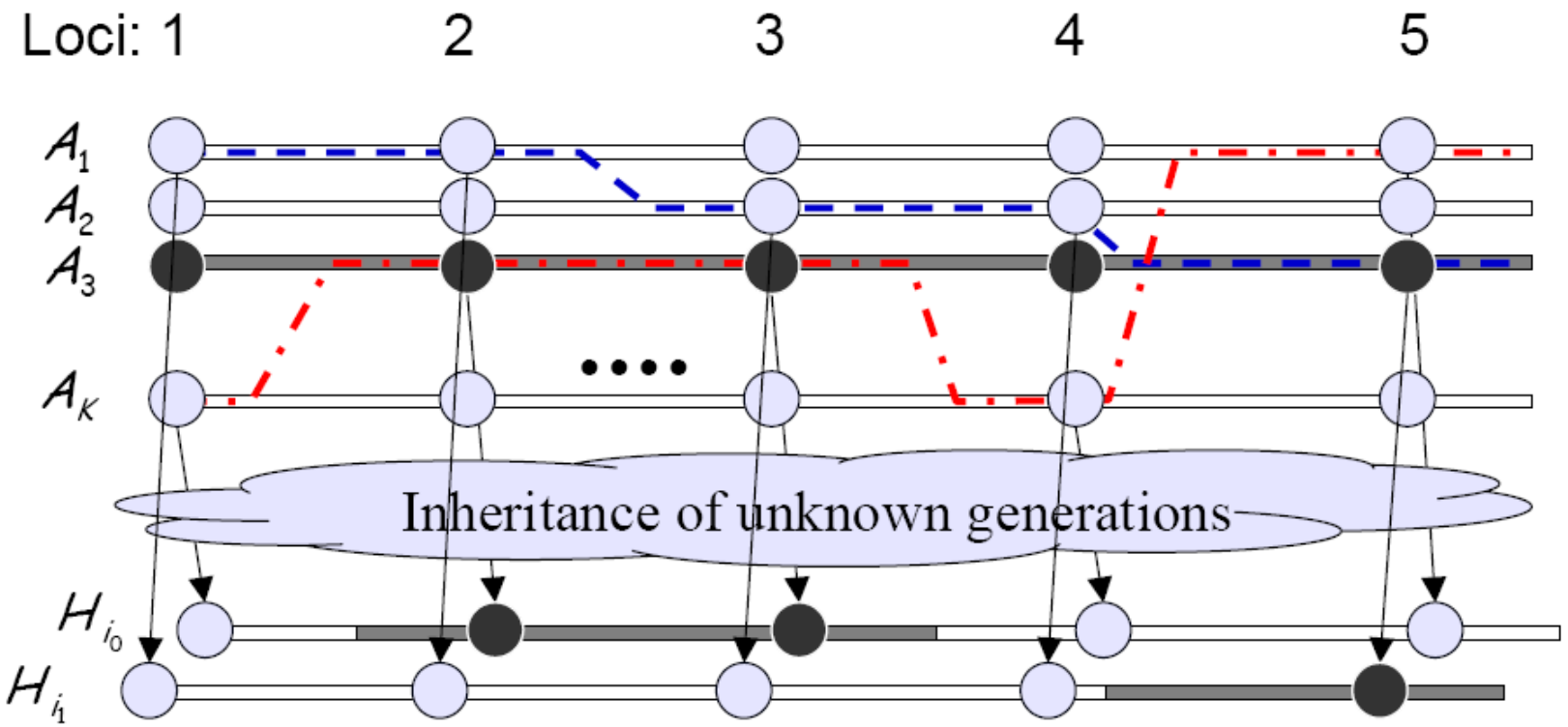
- H_i is a mosaic of multiple ancestral chromosomes
- Model is a time-inhomogenous infinite HMM
- With $r \longrightarrow \infty$, we get stationary HMM

HMDP for Recombination and Inheritance (Contd.)

- H_i is a mosaic of multiple ancestral chromosomes
- Model is a time-inhomogenous infinite HMM
- With $r \longrightarrow \infty$, we get stationary HMM
- Single locus mutation model for emission

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

Haplotype Recombination and Inheritance



HMDP for Recombination and Inheritance (Contd.)

Conditional probability of haplotype list h

$$
\begin{aligned}
p(h|c,a) &= \prod_k \int_{\theta_k} \prod_{i,t|c_{i,t}=k} p(h_{i,t}|a_{k,t},\theta_k) Beta(\theta_k|\alpha_h,\beta_h) d\theta_k \\
&= \prod_k \frac{\Gamma(\alpha_h+\beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)} \frac{\Gamma(\alpha_h+\ell_k)\Gamma(\beta_h+\ell'_k)}{\Gamma(\alpha_h+\beta_h+\ell_k+\ell'_k)} \left(\frac{1}{|B|-1}\right)^{\ell'_k}
\end{aligned}
$$

where

$$
\ell_k = \sum_{i,t} \mathbb{I}(h_{i,t}=a_{k,t})\mathbb{I}(c_{i,t}=k) \quad \ell'_k = \sum_{i,t} \mathbb{I}(h_{i,t}\neq a_{k,t})\mathbb{I}(c_{i,t}=k)
$$

Inference

- Gibbs sampler proceeds in two steps

Inference

- Gibbs sampler proceeds in two steps
  - Sample inheritance $\{C_{i,k}\}$ given h and a

Inference

- Gibbs sampler proceeds in two steps
  - Sample inheritance $\{C_{i,k}\}$ given h and a
  - Sample ancestors $a = \{a_1, \ldots, a_K\}$ given h, C

Inference

- Gibbs sampler proceeds in two steps
  - Sample inheritance $\{C_{i,k}\}$ given h and a
  - Sample ancestors $a = \{a_1, \ldots, a_K\}$ given h, C
- Improve mixing for sampling inheritance

Inference

- Gibbs sampler proceeds in two steps
  - Sample inheritance $\{C_{i,k}\}$ given h and a
  - Sample ancestors $a = \{a_1, \ldots, a_K\}$ given h, C
- Improve mixing for sampling inheritance
  - By Bayes rule

$$p(c_{t+1} : t+\delta | c_-, h, a) \propto \prod_{j=t}^{t+\delta} p(c_{j+1} | c_j, m, n) \prod_{j=t+1}^{t+\delta} p(h_j | a_{c_j,j}, \ell_{c_j})$$

Inference

- Gibbs sampler proceeds in two steps
  - Sample inheritance $\{C_{i,k}\}$ given h and a
  - Sample ancestors $a = \{a_1, \ldots, a_K\}$ given h, C
- Improve mixing for sampling inheritance
  - By Bayes rule

$$p(c_{t+1} : t+\delta | c_-, h, a) \propto \prod_{j=t}^{t+\delta} p(c_{j+1} | c_j, m, n) \prod_{j=t+1}^{t+\delta} p(h_j | a_{c_j,j}, \ell_{c_j})$$

  - Assume probability of having two recombinations is small

$$p(c_{t+1} : t+\delta | c_-, h, a) \propto p(c_{t'} | c_{t'-1}, m, n) p(c_{t+\delta+1} | c_{t+\delta} = c_{t'}, m, n)$$

Inference (Contd.)

- Assuming d, r to be small, $\lambda = 1 - \exp(-dr) \approx dr$

$$p(c_{t'} = k | c_{t'-1} = k, m, n, r, d) = \begin{cases} \lambda\pi_{k,k'} + (1-\lambda)\delta(k,k') \text{for} k' \in \{1 \\ \lambda\pi_{k,K+1} \quad \text{for} k' = K+1 \end{cases}$$

Inference (Contd.)

- Assuming d, r to be small, $\lambda = 1 - \exp(-dr) \approx dr$

$$p(c_{t'} = k | c_{t'-1} = k, m, n, r, d) = \begin{cases} \lambda\pi_{k,k'} + (1-\lambda)\delta(k,k') \text{for} k' \in \{1 \\ \lambda\pi_{k,K+1} \quad \text{for} k' = K+1 \end{cases}$$

- Terms can be replaced in original equation to get sampler

Inference (Contd.)

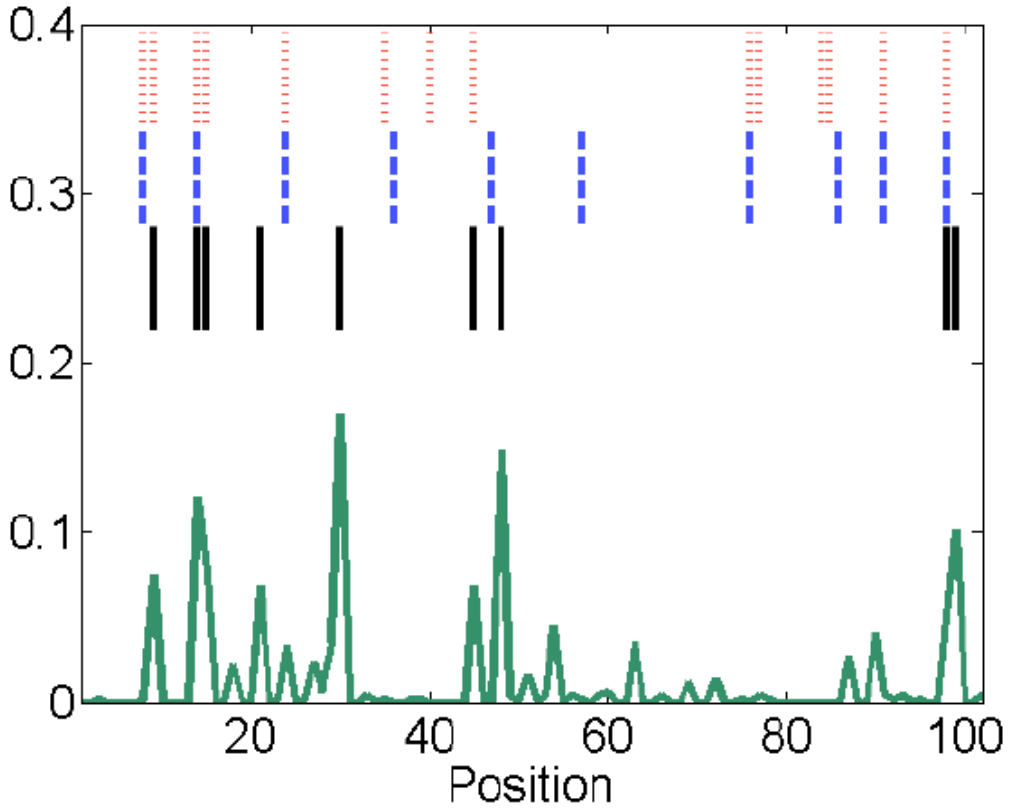
- Assuming d, r to be small, $\lambda = 1 - \exp(-dr) \approx dr$

$$p(c_{t'} = k | c_{t'-1} = k, m, n, r, d) = \begin{cases} \lambda\pi_{k,k'} + (1-\lambda)\delta(k,k') \text{for} k' \in \{1 \\ \lambda\pi_{k,K+1} \quad \text{for} k' = K+1 \end{cases}$$

- Terms can be replaced in original equation to get sampler
- Posterior distribution for ancestors

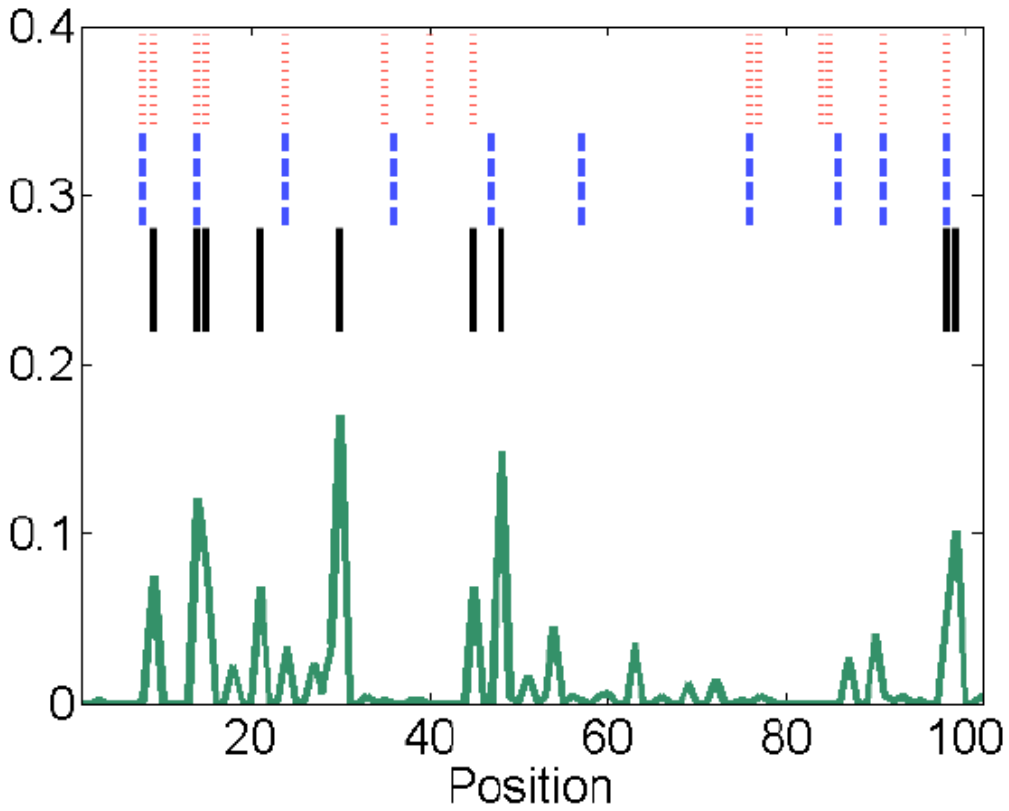
$$p(a_{k,t} | c, h) \propto \frac{\Gamma(\alpha_h + \beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)} \frac{\Gamma(\alpha_h + \ell_{k,t})\Gamma(\beta_h + \ell'_{k,t})}{\Gamma(\alpha_h + \beta_h + \ell_{k,t} + \ell'_{k,t})} \left(\frac{1}{|B| - 1}\right)^{\ell'_{k,t}}$$

Single Population Data

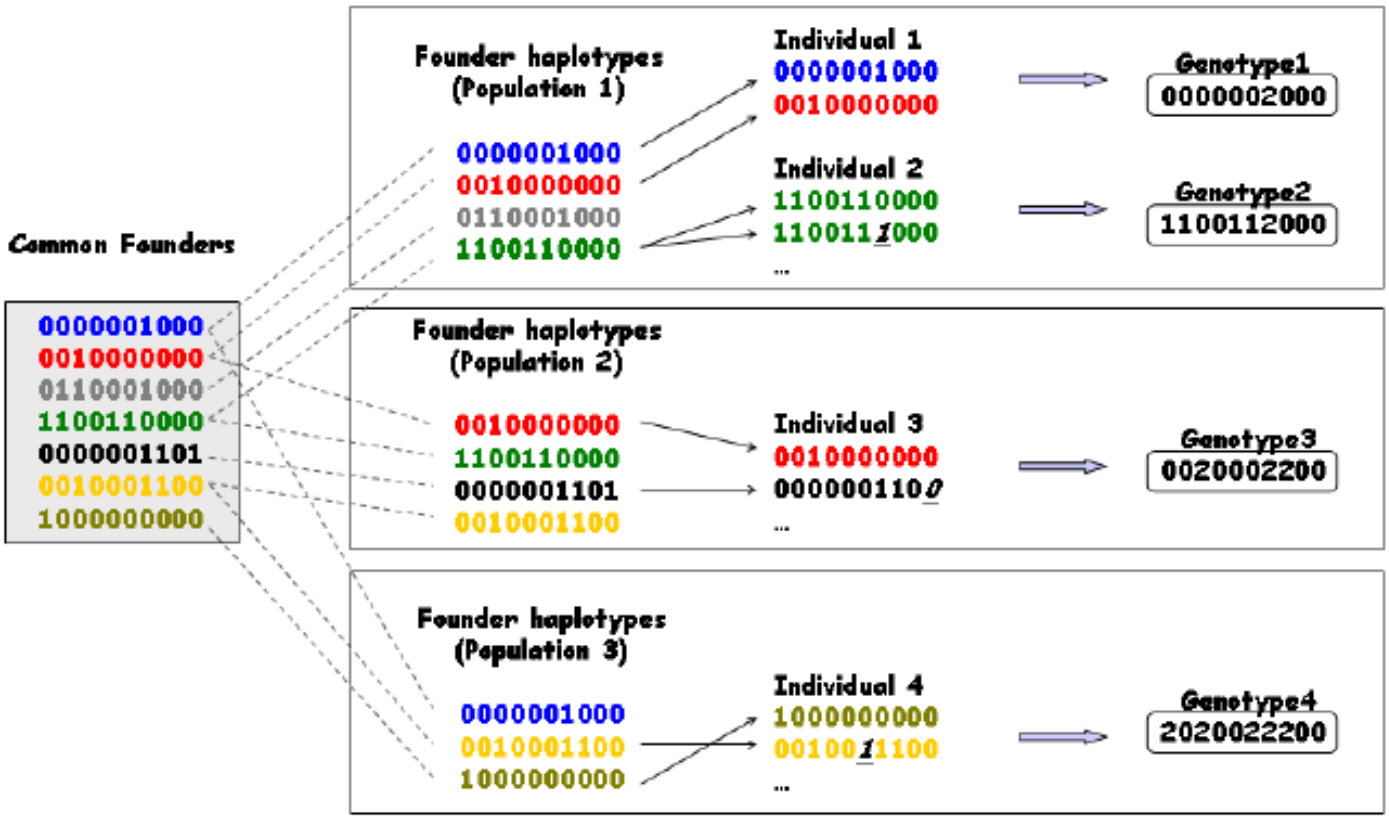


Haplotype block boundaries
HMDP (black solid), HMM (red dotted), MDL (blue dashed)

Two Population Data



Hierarchical DPM for Haplotype Inference



Hierarchical DPM for Haplotype Inference (Contd.)

$Q_0(\phi_1, \phi_2, \ldots)\|\gamma, F \sim \mathrm{DP}(\gamma, F)$,	sample a DP of founders for all populations;
$Q_j(\phi_1^{(j)}, \phi_2^{(j)}, \ldots)\|\tau, Q_0 \sim \mathrm{DP}(\tau, Q_0)$,	sample the DP of founders for each population;
$\phi_{i_e}^{(j)}\|Q_j \sim Q_j$,	sample the founder of haplotype i_e in population j;
$h_{i_e}^{(j)}\|\phi_{i_e}^{(j)} \sim P_h(\cdot\|\phi_{i_e}^{(j)})$,	sample haplotype i_e in population j;
$g_i^{(j)}\|h_{i_0}^{(j)}, h_{i_1}^{(j)} \sim P_g(\cdot\|h_{i_0}^{(j)}, h_{i_1}^{(j)})$,	sample genotype i in population j,

Experiments: Hapmap Data

- SNP genotypes from four populations

Experiments: Hapmap Data

- SNP genotypes from four populations
  - CEPH, Utah residents with northern/weatern European ancestry, 60

Experiments: Hapmap Data

- SNP genotypes from four populations
  - CEPH, Utah residents with northern/weatern European ancestry, 60
  - YRI, Yoruba in Ibadan, Nigeria, 60

Experiments: Hapmap Data

- SNP genotypes from four populations
  - CEPH, Utah residents with northern/weatern European ancestry, 60
  - YRI, Yoruba in Ibadan, Nigeria, 60
  - CHB, Han Chinese in Beijing, 45

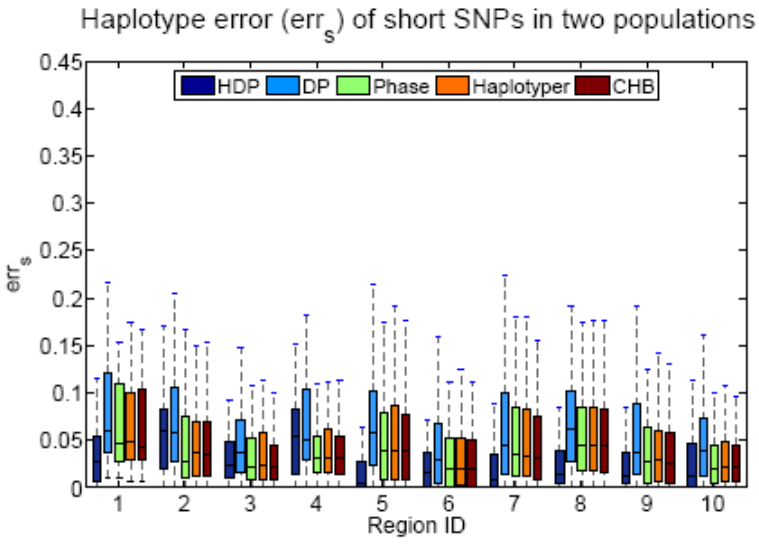
Experiments: Hapmap Data

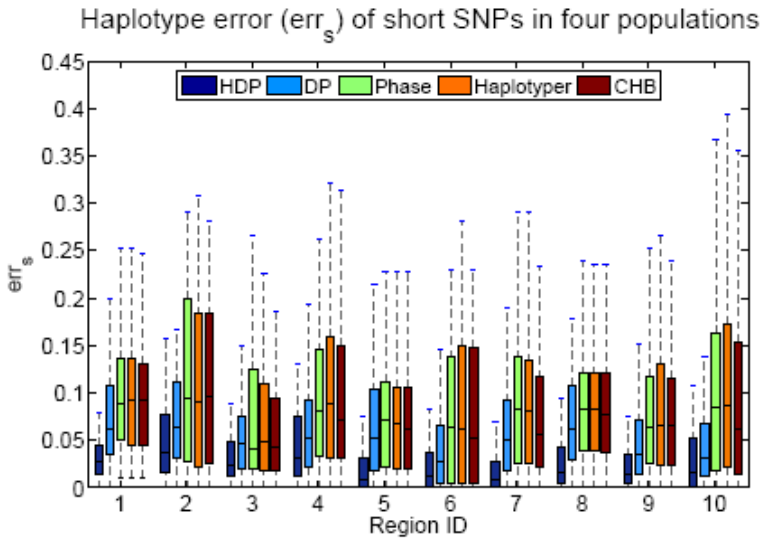
- SNP genotypes from four populations
  - CEPH, Utah residents with northern/weatern European ancestry, 60
  - YRI, Yoruba in Ibadan, Nigeria, 60
  - CHB, Han Chinese in Beijing, 45
  - JPT, Japanese in Tokyo, 44

Experiments: Hapmap Data

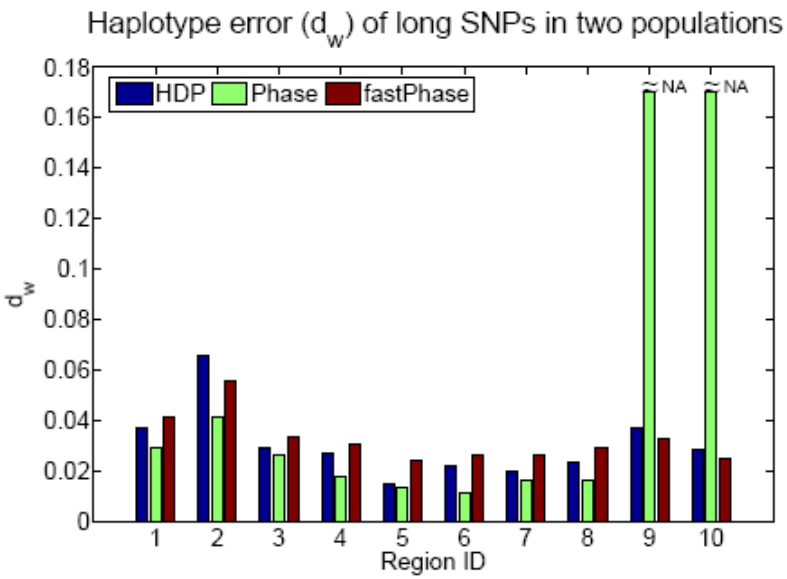
- SNP genotypes from four populations
  - CEPH, Utah residents with northern/weatern European ancestry, 60
  - YRI, Yoruba in Ibadan, Nigeria, 60
  - CHB, Han Chinese in Beijing, 45
  - JPT, Japanese in Tokyo, 44
- Experiments on short (~ 10) and long ($\sim 10^2 - 10^3$) SNPs

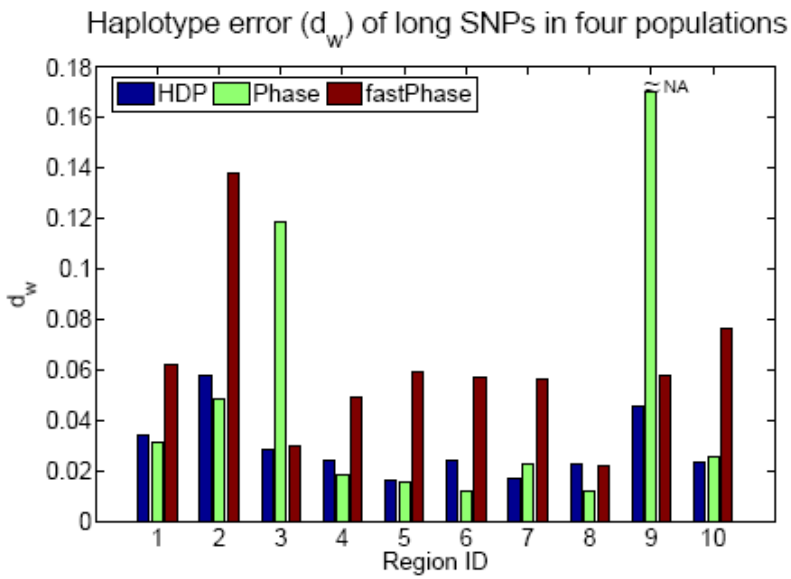
Short SNP Sequences



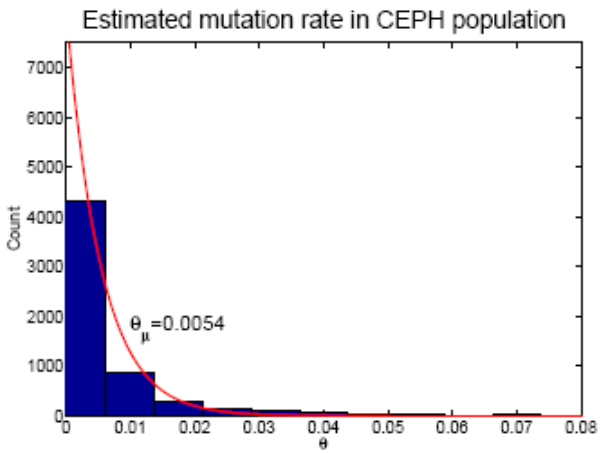


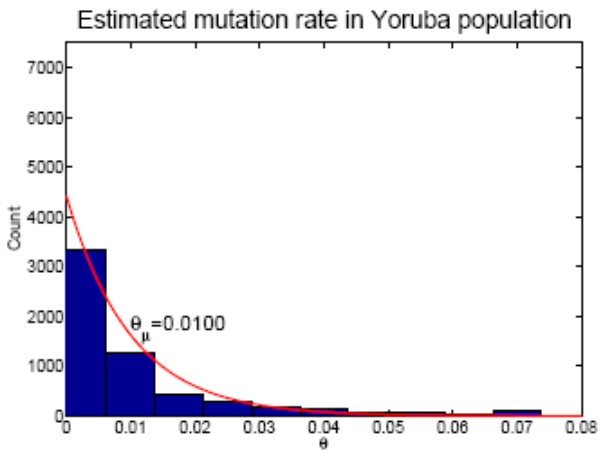
Long SNP Sequences



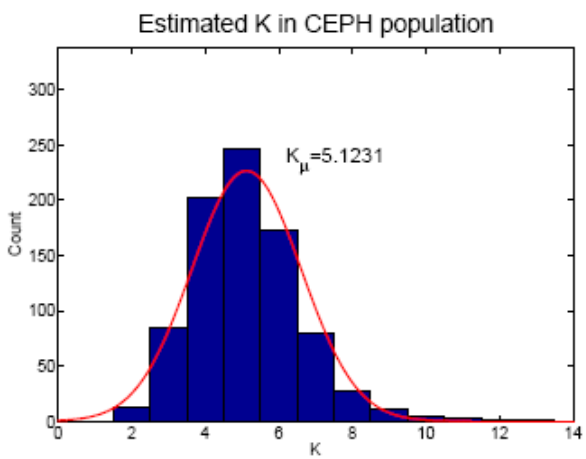


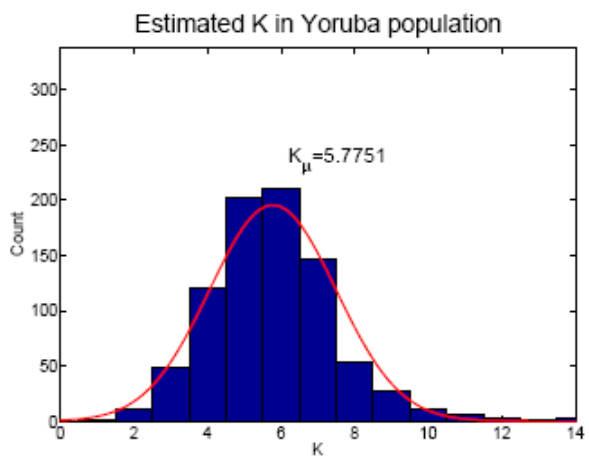
Mutation Rates and Diversity



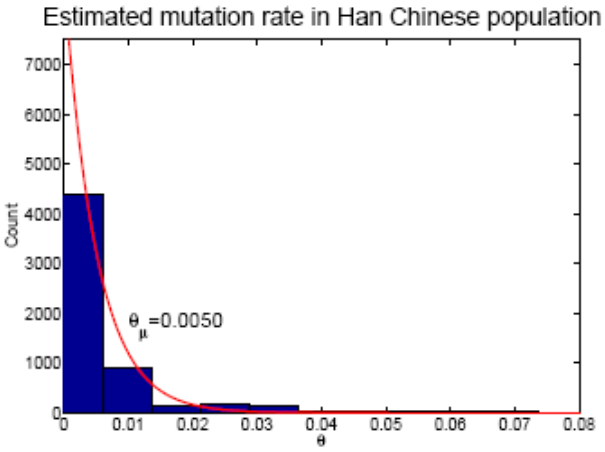


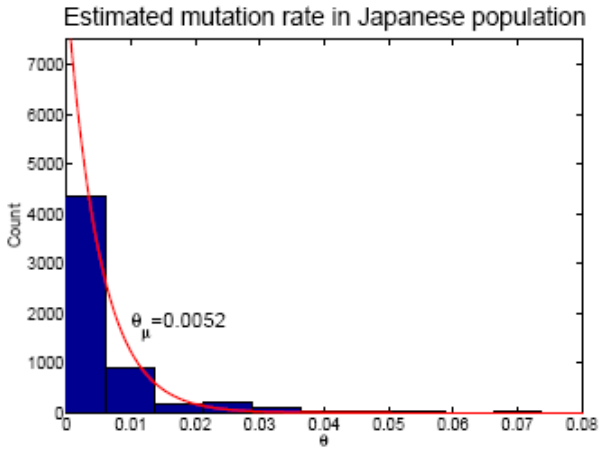
(a) CEPH (b) Yoruba





Mutation Rates and Diversity (Contd.)





(c) Han Chinese

(d) Japanese