

Bayesian Methods for Genetic Inference

Eric P. Xing

MLD, LTI and CSD
School of Computer Science
Carnegie Mellon University

KITP seminar, September 23, 2008

1



SAILING LAB



Laboratory for Statistical Artificial Intelligence & Integrative Genomics

Postdocs




Seyoung Kim
Postdoc (PhD, Postdoc (PhD,
UCI) USydney)

PhD students

 Amir Ahmed LTI	 Wenjie Fu CSD	 Fan Guo CSD	 Steve Hanneke MLD	 Jude Howrylak Comp Bio
 Hetunandan K. CSD	 Maden Kolar LTI	 Andr� Martins LTI / UT Lisboa	 Kriti Punjani LTI	 Pradipta Ray Comp Bio
 Suyash Shringarpure MLD	 Kyung-Ah Sohn CSD			

<http://www.sailing.cs.cmu.edu/>

2



Overview:

Learning and Reasoning under Uncertainty

□ Learning in structured input/output space

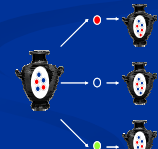


(zen, et al. KDD 07, Zhu, et al. ICML08, Kim, UAI08)

- Semi-supervised and unsupervised maximum margin learning
- Theory and algorithm for optimization, inference and active learning
- Applications in genomics, machine translation, and multi-media analysis

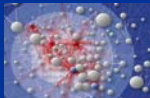
□ Nonparametric Bayesian models for "open worlds"

- Domain-closure, unique name and stationarity assumptions are not always valid:
 - How many clusters/states/objects/relations out there?
 - Ambiguous data association.
 - Birth/death/evolution of possible worlds.
- Infinite-capacity models based on Dirichlet process (Polya urn schemes)
- Applications in genetics and evolution, tracking and email filtering



(Xing, et al. ICML 04,06, Ahmed SDM08)

□ Statistical modeling and inference of relational data



(Guo, et al. ICML 07)

- Modeling the formation, evolution, and dynamics of networks
- Inferring their semantic aspects, missing links, and node attributes
- Biological and social network analysis

3



Overview:

Computational Biology and Statistical Genetics

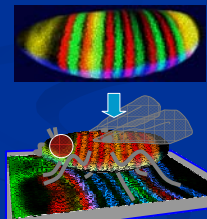
□ Genomics and regulatory evolution



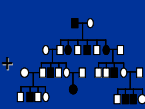
- Statistical models for genome evolution and natural selection
- Functional effects on gene regulation and morphogenesis
- Gene finding and functional prediction via comparative genomic analysis

□ Computation Developmental Biology of Flies

- Image analysis and database
 - Feature processing, segmentation, and pattern representation
 - Recovering 3D structure from 2D images
 - Shape and deformation modeling and categorization
- Spatial-temporal modeling of gene regulation
- Temporal shape evolution and models for morphogenesis
- The genetics of pattern polymorphism and divergence



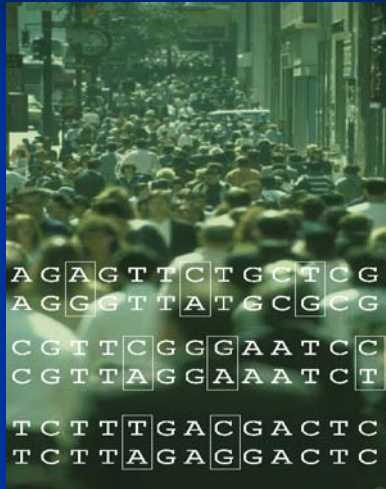
□ Genetic variation and diseases association





- Genealogy/evolution models: how many founders, migration and evolution history...
- Models for linkages between variations and phenotypes
- Clinical and forensic applications

4

Genome Polymorphisms



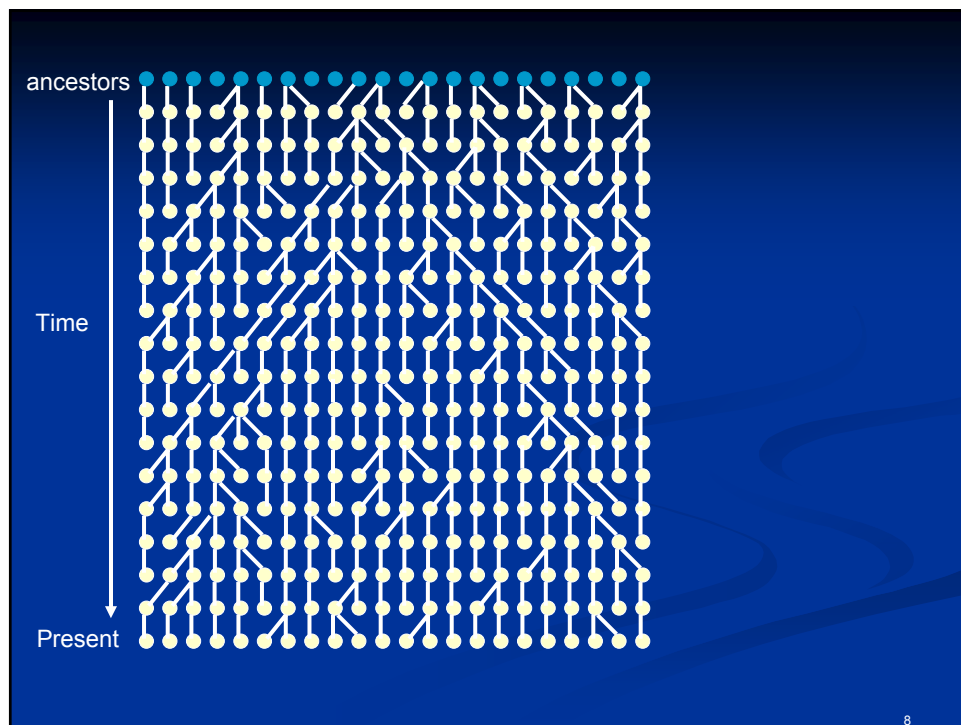
The ABO Blood System				
Blood Type (genotype)	Type A (AA, AO)	Type B (BB, BO)	Type AB (AB)	Type O (OO)
Red Blood Cell Surface Proteins (phenotype)				
	A agglutinogens only	B agglutinogens only	A and B agglutinogens	No agglutinogens
Plasma Antibodies (phenotype)			NONE.	
	Anti-B agglutinin only	Anti-A agglutinin only	No agglutinins	Anti-A and Anti-B agglutinins

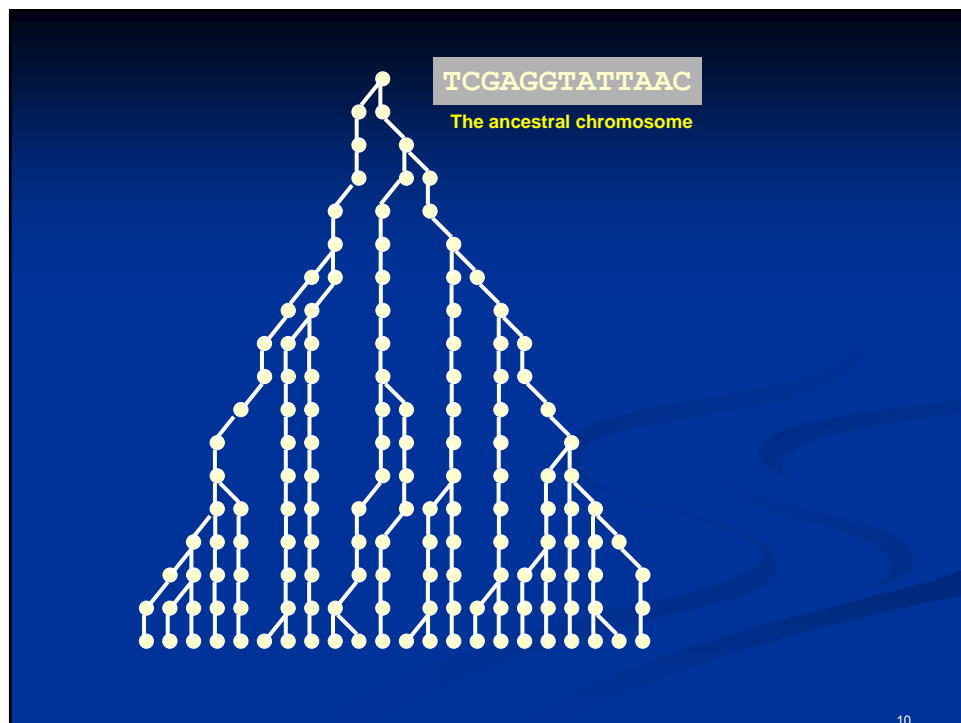
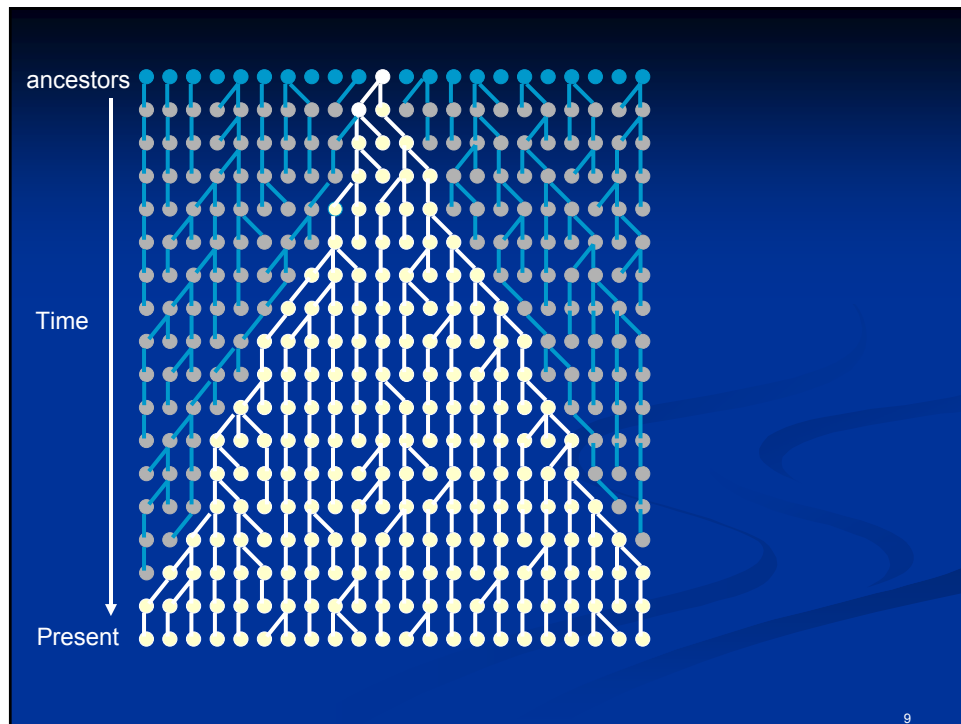


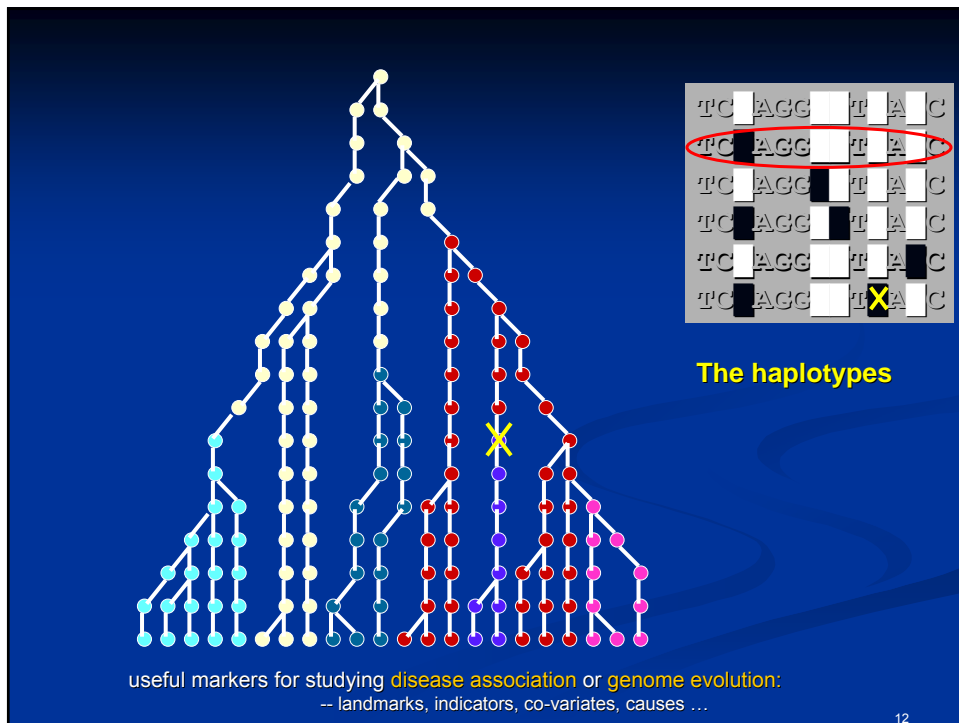
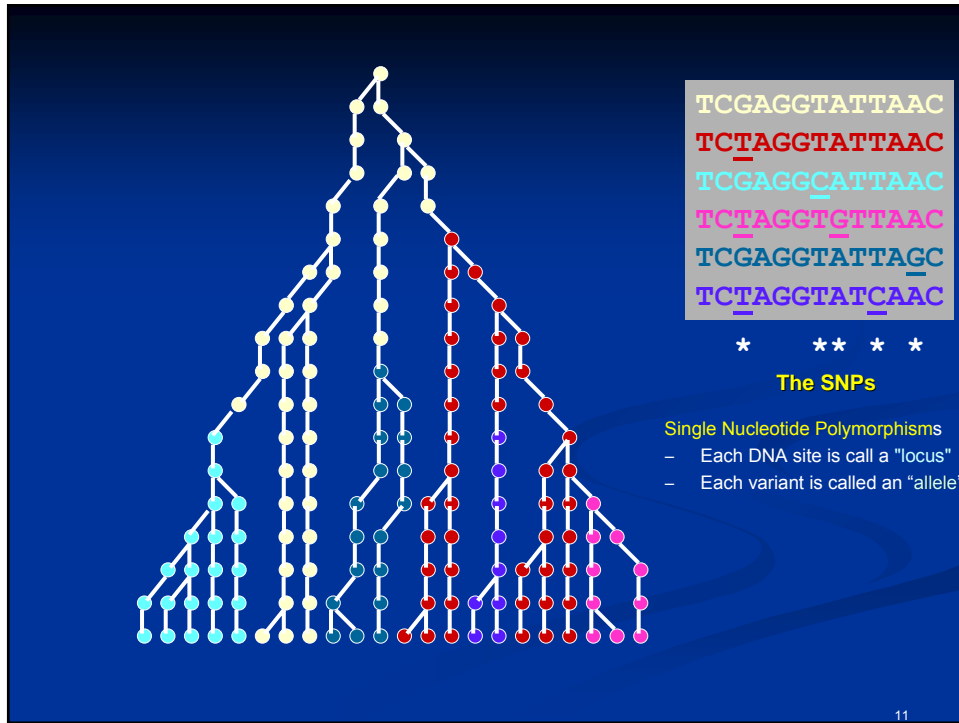
5



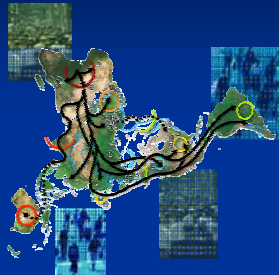
E



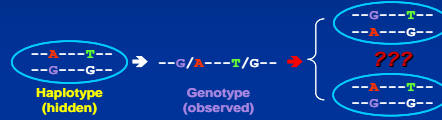




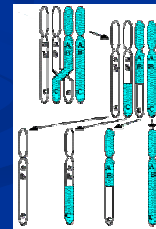
Genetic Inference



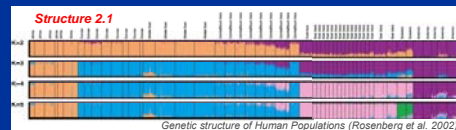
- Determine genetic markers
 - Haplotype inference



- Reveal genome inheritance events
 - Recombination hotspot identification



- Deconvolve population structure
 - Ancestral spectrum analysis



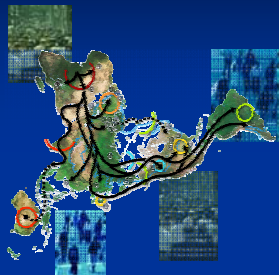
13

Outline

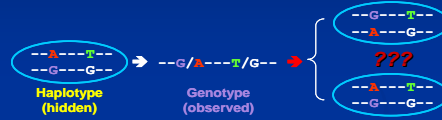
- **Haplotype Inference**
 - Dirichlet Process for phasing single population
 - Hierarchical DP for phasing multiple population
- **Linkage-disequilibrium analysis**
 - Hidden Markov DP for identifying recombination hotspots
- **Population structure analysis**
 - Admixture model
 - HMDP models

14

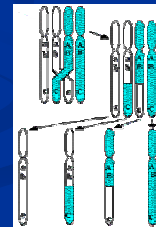
Genetic Inference



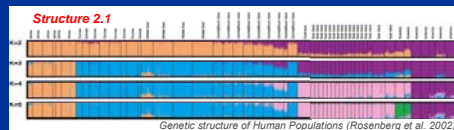
- Determine genetic markers
 - Haplotype inference



- Reveal genome inheritance events
 - Recombination hotspot identification

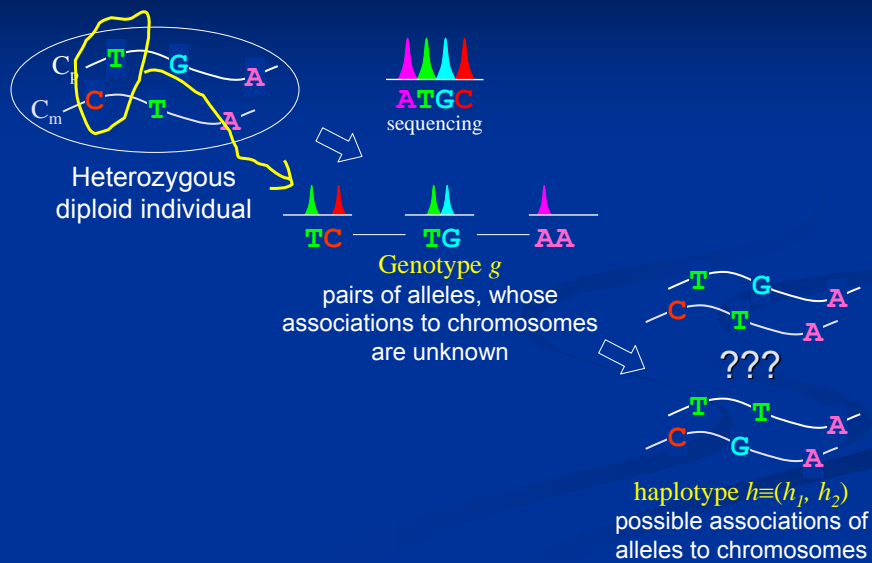


- Deconvolve population structure
 - Ancestral spectrum analysis



15

Haplotype Ambiguity

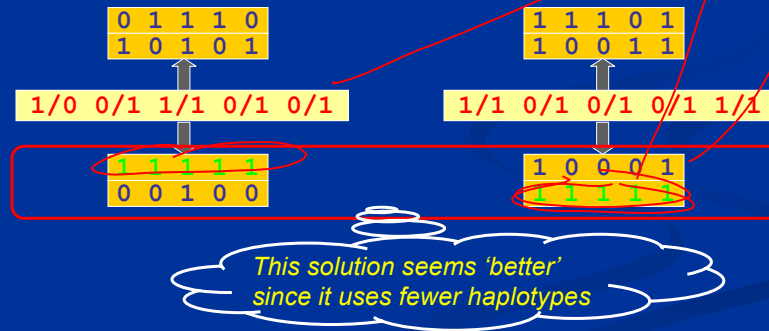


16

Haplotype Inference

The Rationale: parsimony

- Many haplotypes are *shared* in a population
- Data for many individuals allows *phasing* SNP genotypes



17

A Finite (Mixture of) Allele Model

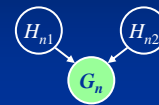
- The probability of a genotype g :

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

Population haplotype pool

Haplotype model

Genotyping model



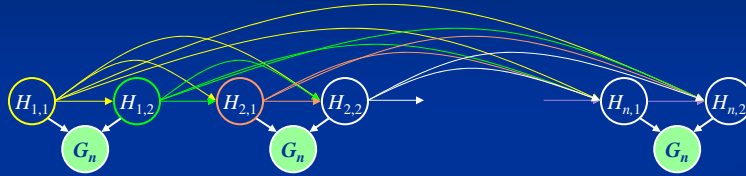
- Standard settings:

- $|\mathcal{H}| = K \ll 2^I$ fixed-sized population haplotype pool
- $p(h_1, h_2) = p(h_1)p(h_2) = f_1 f_2$ Hardy-Weinberg equilibrium

- Problem: $K?$ $\mathcal{H}?$

18

The PAC Model



- The joint probability of all haplotypes h_1, h_2, \dots, h_n :

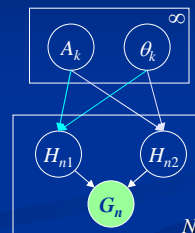
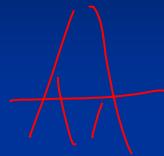
$$p(h_1, h_2, \dots, h_n) = p(h_1) p(h_2 | h_1) p(h_3 | h_1, h_2) \dots p(h_n | h_1, \dots, h_{n-1})$$

- Problem:

- Ordering?
- Ancestor?

19

An Infinite (Mixture of) Allele Model



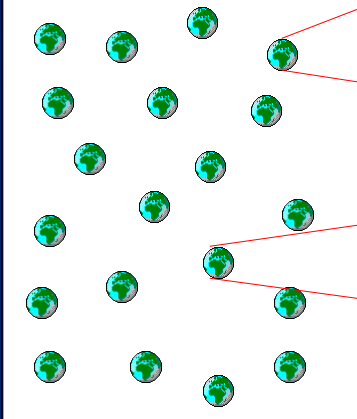
- How?

- Via a nonparametric hierarchical Bayesian formalism !

20

Dirichlet Process

Possible worlds of partitions



- A CDF, G , on possible worlds of random partitions follows a **Dirichlet Process** if for any measurable finite partition $(\phi_1, \phi_2, \dots, \phi_m)$:

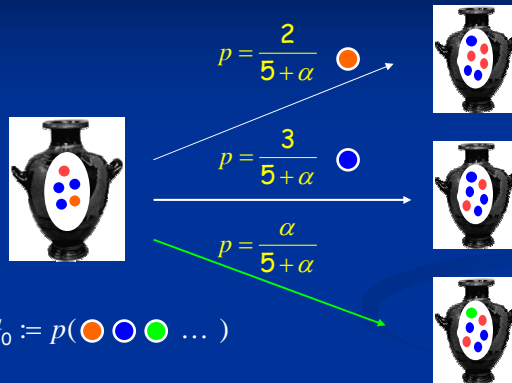
$$(G(\phi_1), G(\phi_2), \dots, G(\phi_m)) \sim \text{Dirichlet}(\alpha G_0(\phi_1), \dots, \alpha G_0(\phi_m))$$

where G_0 is the **base measure** and α is the **scale parameter**

Thus a Dirichlet Process G defines a distribution of distribution

21

DP – a *Pólya* urn Process



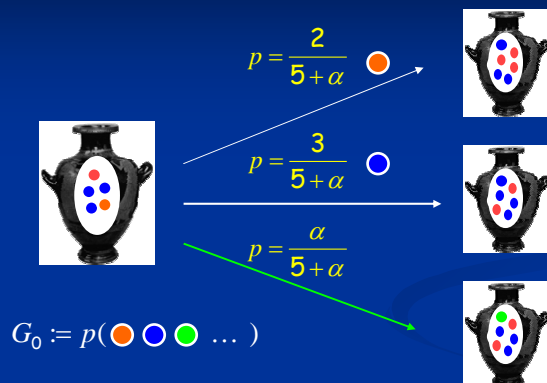
Joint: $G(\text{urn}) \sim DP(\alpha G_0)$

Marginal: $\phi_i | \phi_{-i}, \alpha, G_0 \sim \sum_{k=1}^K \frac{n_k}{i-1+\alpha} \delta_{\phi_k} + \frac{\alpha}{i-1+\alpha} G_0$

- "Infinite"
- Self-reinforcing property
- **exchangeable partition** of samples

22

Clustering and DP Mixture



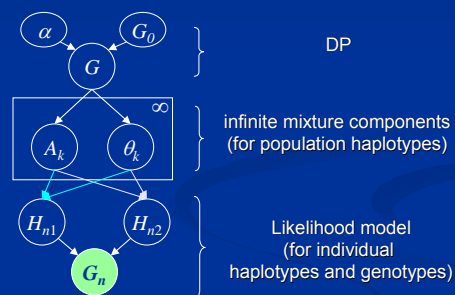
- | | | | |
|---------------------|---------------------|---------------------|-----|
| ① ③ | ② ④ ⑤ | ⑥ | ... |
| $\{a_1, \theta_1\}$ | $\{a_2, \theta_2\}$ | $\{a_3, \theta_3\}$ | ... |
- We can associate ancestors (i.e., mixture components) with the colors in the Pólya urn and thereby define an **infinite clustering** of the haplotypes (i.e., balls)

23

Dirichlet Process Mixture of Haplotypes

(Xing et al. ICML 2004)

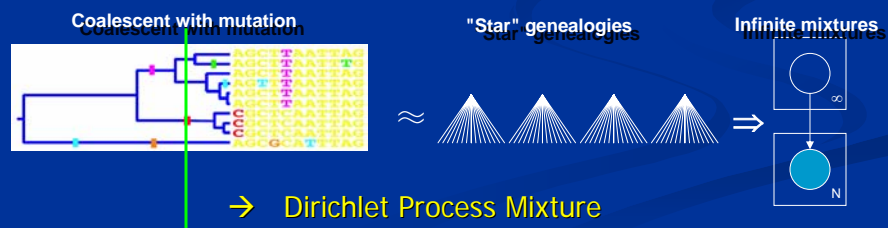
- A Hierarchical Bayesian Infinite Allele model



24

Population Genetic Basis of IAM

- Kingman coalescent process with fixed (large) population size
- New population haplotype alleles emerge along all branches of the coalescence tree at rate $a/2$ per unit length
 - **Ewens Sampling Formula: an exchangeable random partition of individuals**



25

Inheritance and Observation Models

- Single-locus mutation model

$$A_{C_{i_k}} \rightarrow H_{i_k}$$

$$P_H(h_i | a_i, \theta) = \begin{cases} \theta & \text{for } h_i = a_i \\ \frac{1-\theta}{|B|-1} & \text{for } h_i \neq a_i \end{cases}$$

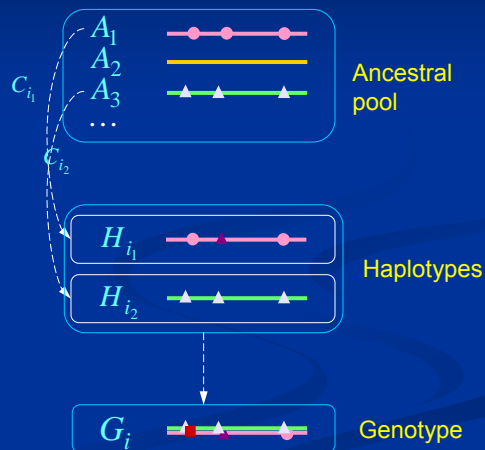
→ $h_i = a_i$ with prob. θ

- Noisy observation model

$$H_{i_1}, H_{i_2} \rightarrow G_i$$

$$P_G(g | h_1, h_2):$$

$$g_i = h_{1,i} \oplus h_{2,i} \text{ with prob. } \lambda$$



26

MCMC for Haplotype Inference

- Gibbs sampling for exploring the posterior distribution under the proposed model
 - Integrate out the parameters such as θ or λ , and sample c_{i_e} , a_k and h_{i_e}

$$p(c_{i_e} = k \mid \mathbf{c}_{[-i_e]}, \mathbf{h}, \mathbf{a}) \propto \underbrace{p(c_{i_e} = k \mid \mathbf{c}_{[-i_e]})}_{\text{Prior}} \times \underbrace{p(h_{i_e} \mid a_k, \mathbf{h}_{[-i_e]}, \mathbf{c})}_{\text{Likelihood}}$$

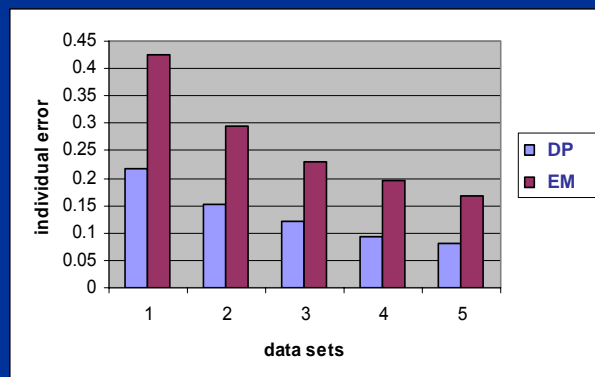
Posterior
Pólya urn
Likelihood

- Gibbs sampling algorithm: draw samples of each random variable to be sampled given values of all the remaining variables

27

Results - HapMap Data

- DP vs. Finite Mixture via EM



28

Extensions of the DP haplotyper



KITP seminar, September 23, 2008

29

Multi-population Genetic Demography



- Inference done separately, or jointly?

30

Multi-population Genetic Demography

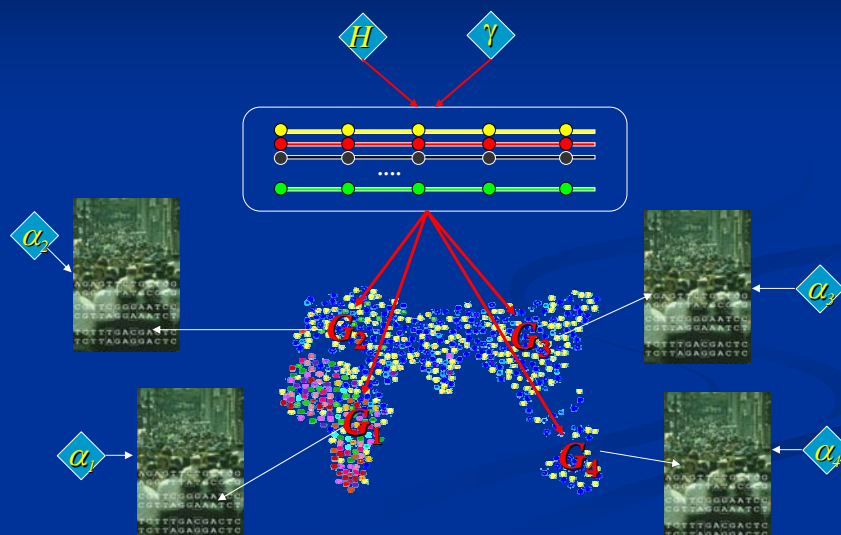


- Pool everything together and solve 1 hap problem?
 - --- ignore population structures
- Solve 4 hap problems separately?
 - --- data fragmentation
- Co-clustering ... solve 4 *coupled* hap problems jointly

31

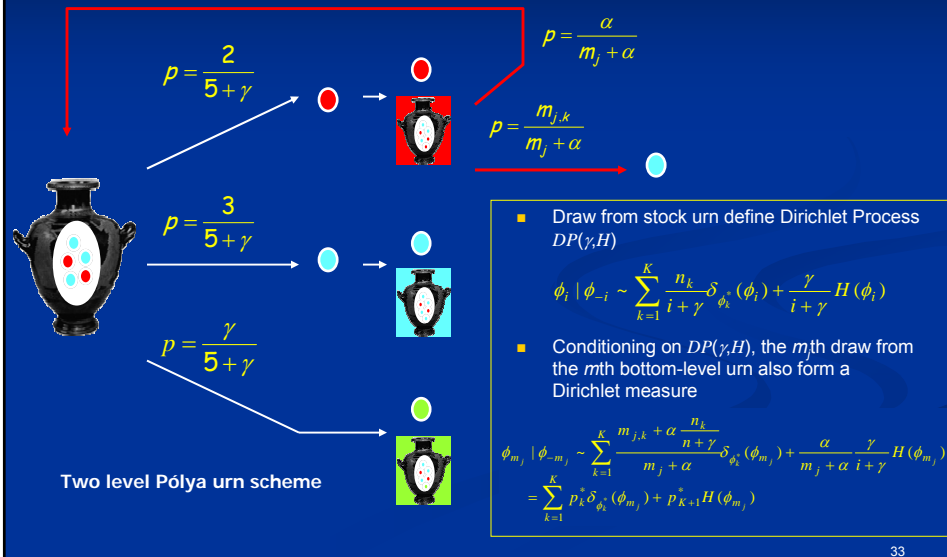
Hierarchical DP Mixture

(Xing et al. ICML 2006)



32

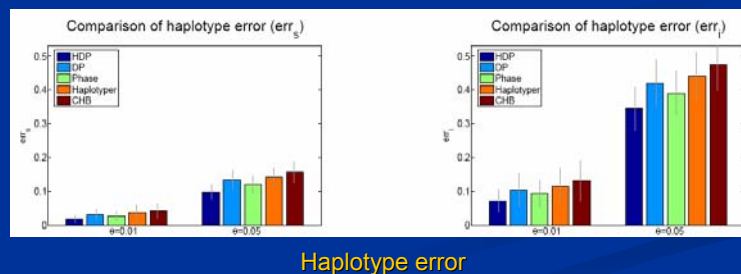
A Hierarchical Pólya Urn Sampler



33

Results - Simulated Data

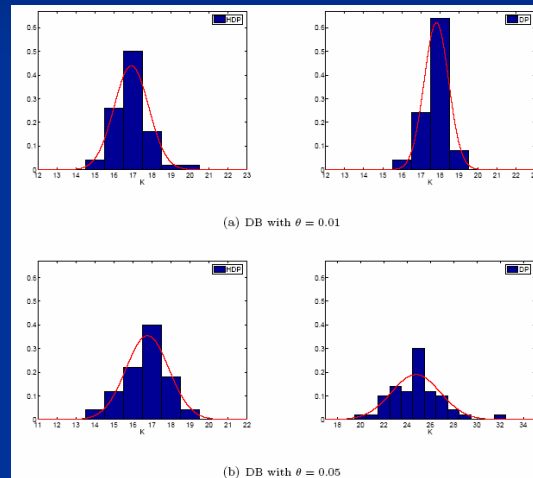
- 5 populations with 20 individuals each (two kinds of mutation rates)
- 5 populations share parts of their ancestral haplotypes
- the sequence length = 10



34

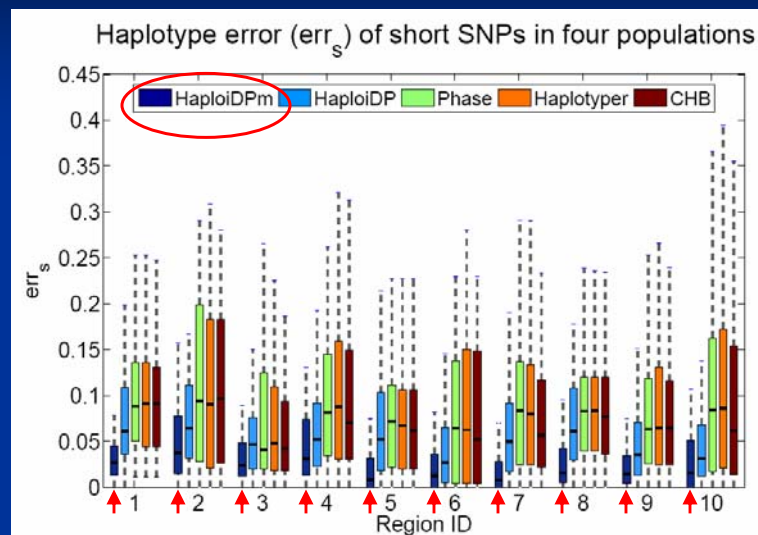
Results on Simulated Data

Estimation of K



35

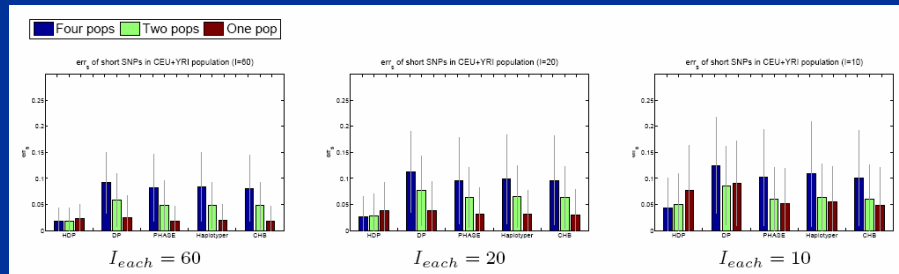
Results - International HapMap DB



36

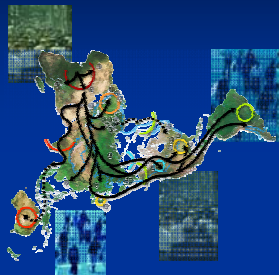
Results - International HapMap DB

- Different sample sizes, and different # of sub-populations



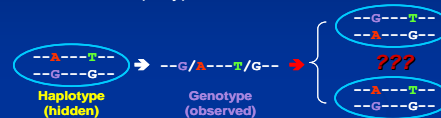
37

Genetic Inference



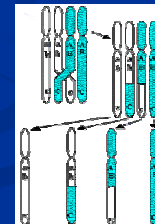
- Determine genetic markers

- Haplotype inference



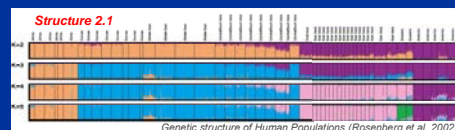
- Reveal genome inheritance events

- Recombination hotspot identification



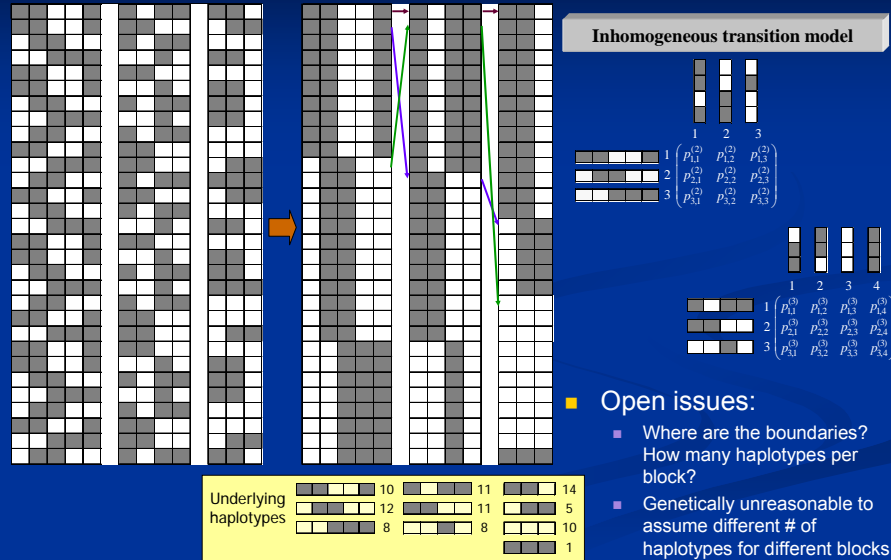
- Deconvolve population structure

- Ancestral spectrum analysis



38

Modeling Haplotype Structure



39

Inheritance Model

Each individual haplotype is a mosaic of ancestral haplotypes

Ancestral chromosomes (K=5)

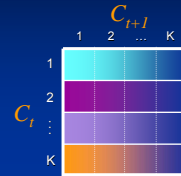
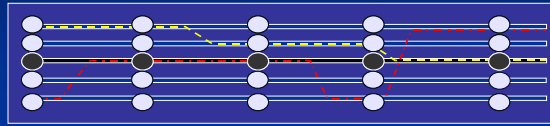


Individual chromosomes



40

The Hidden Markov Model



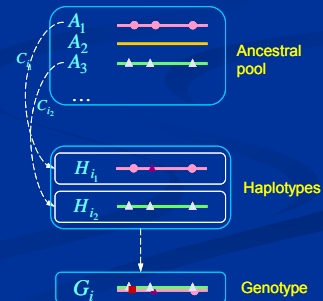
- Transition process: recombination

$$p(c_{i,t+1} = k' | c_{i,t} = k) = e^{-dr} \pi_{k,k'} + (1 - e^{-dr}) \delta(k, k')$$

- Emission process: mutation

$$p(h_{i,t} | a_{k,t}, \theta_k) = \theta_k^{I(h_{i,t} = a_{k,t})} \left(\frac{1 - \theta_k}{|B| - 1} \right)^{1 - I(h_{i,t} = a_{k,t})}$$

How many recombining ancestors?

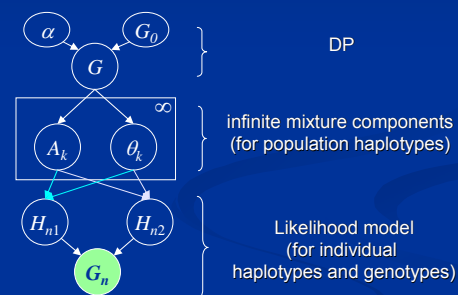


41

Recall DP Mixture

(Xing et al. ICML 2004, 2006)

- A Hierarchical Bayesian Infinite Allele model



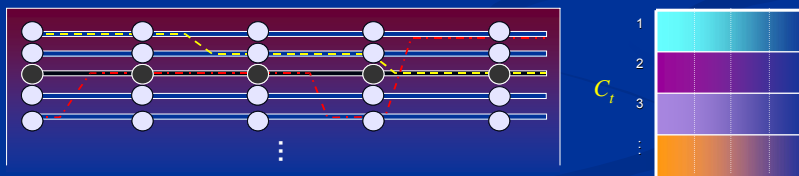
42

Hidden Markov Dirichlet Process

(Xing and Sohn. *Bayesian Analysis*, 2007, Sohn and Xing, *ISMB* 2007)

- Hidden Markov Dirichlet process mixtures
 - Extension of HMM model to infinite ancestral space
 - Infinite dimensional transition matrix
 - Each row of the transition matrix is modeled with a DP: $G_i | \alpha, G_0 \sim \text{DP}(\alpha, G_0)$

$$G_0 | \gamma, H \sim \text{DP}(\gamma, H)$$



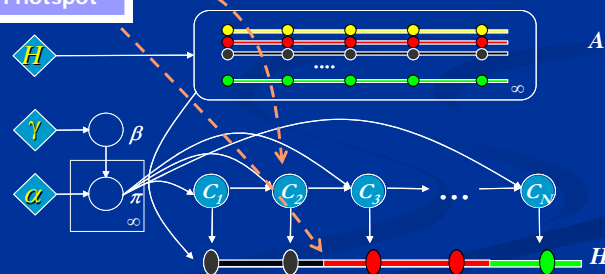
43

HMDP as a Graphical Model

Ancestor allele reconstruction

Inferring population structure

Inferring recombination hotspot



44

MCMC Inference

■ Gibbs sampling

■ Block sampler:

$$\begin{aligned}
 p(c_{t+1}, \dots, c_{t+\tau} \mid c^-, \mathbf{h}, \mathbf{a}) &= \prod_{s=t}^{t+\tau-1} p(c_{s+1} \mid c_s, \mathbf{h}, \mathbf{a}) \\
 &= p(\text{recombination location} \in [t'-1, t'], c_t = k \mid c_{t+\tau}, \mathbf{h}, \mathbf{a}) \\
 &= p(c_{t+1 \dots t'-1} = c_t, c_{t' \dots t+\tau} = k \mid c_{t+\tau}, \mathbf{h}, \mathbf{a})
 \end{aligned}$$

■ Pólya urn sampler: posterior transition probability

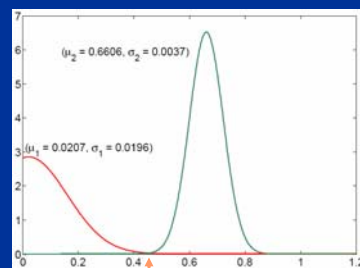
$$p(c_{t+1} = k \mid c_t = j, \mathbf{m}, \mathbf{n}) = \begin{cases} r \times \frac{m_{jk} + \alpha \pi_k}{m_j + \alpha} + (1-r) \times \delta(j, k) & \text{for } j = 1, \dots, K \\ r \times \pi_{c_t + \tau + 1} \end{cases}$$

$$\text{where } \pi = \left(\frac{n_1}{n + \gamma}, \dots, \frac{n_K}{n + \gamma}, \frac{\gamma}{n + \gamma} \right)$$

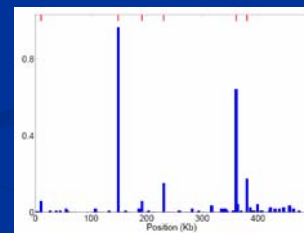
45

Recombination Analysis

Recombination hotspot detection

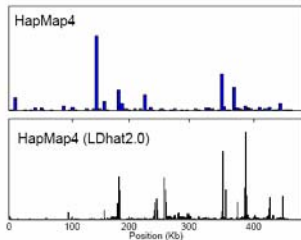
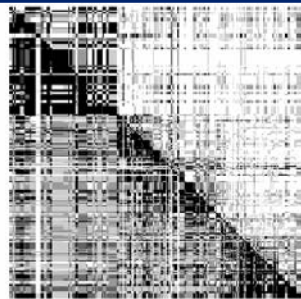


threshold for hotspot detection



46

Recombination Analysis



	w_{tol}	Stepcrum			LDhat 2.0			HMM ($K = 5$)		
		0	± 1	± 2	0	± 1	± 2	0	± 1	± 2
$\omega =$	FPR	0.16	0.11	0.07	0.19	0.09	0.06	0.18	0.12	0.11
3rd quartile	FNR	0.11	0	0	0.22	0.11	0.11	0.33	0.11	0.11
ω S.L.	FPR	0.16	0.11	0.07	0.22	0.11	0.07	0.18	0.12	0.11
FNR ~ FAR	FNR	0.11	0	0	0.22	0.12	0.11	0.33	0.11	0.11

47

Association Mapping as Regression

	Phenotype (BMI)	Genotype
Individual 1	2.5	<div> <div>C</div> <div>T</div> <div>C</div> <div>T</div> </div>
Individual 2	4.8	<div> <div>C</div> <div>A</div> <div>C</div> <div>T</div> </div>
⋮		
Individual N	4.7	<div> <div>G</div> <div>T</div> <div>C</div> <div>T</div> </div>



Benign SNPs



Causal SNP

$$y_i = \sum_{j=1}^J x_{ij} \beta_j$$

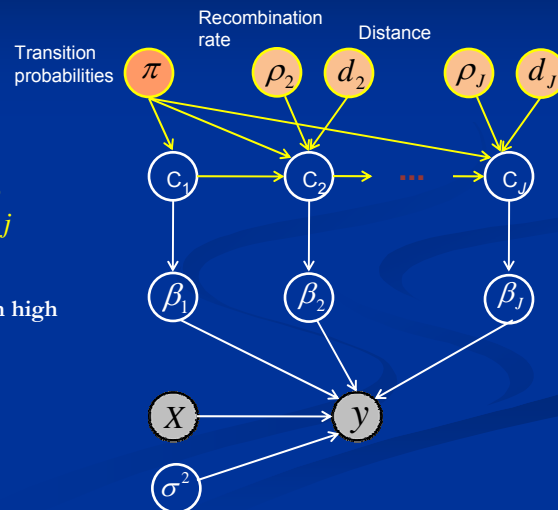
SNPs with large $|\beta_j|$ are relevant

Block-regularized Regression

(Kim and Xing, UAI 2008)

$$\mathbf{y}_i = \sum_{j=1}^J x_{ij} \beta_j$$

Standard LASSO will results in high false positives with very high dimensional X



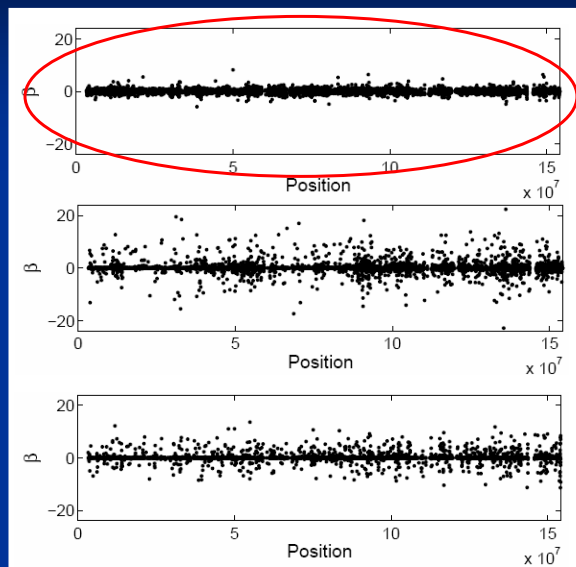
49

Mouse Data (BROAD institute)

Block-regularized regression

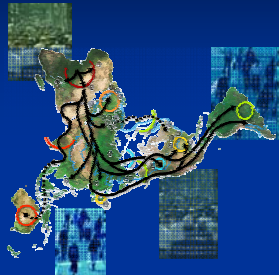
Independent Bernoulli prior

Lasso

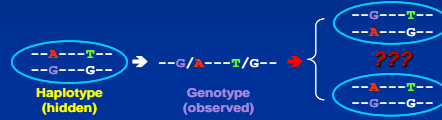


50

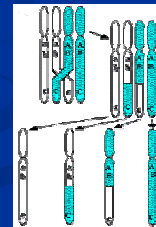
Genetic Inference



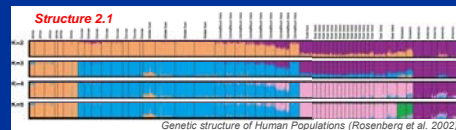
- Determine genetic markers
 - Haplotype inference



- Reveal genome inheritance events
 - Recombination hotspot identification



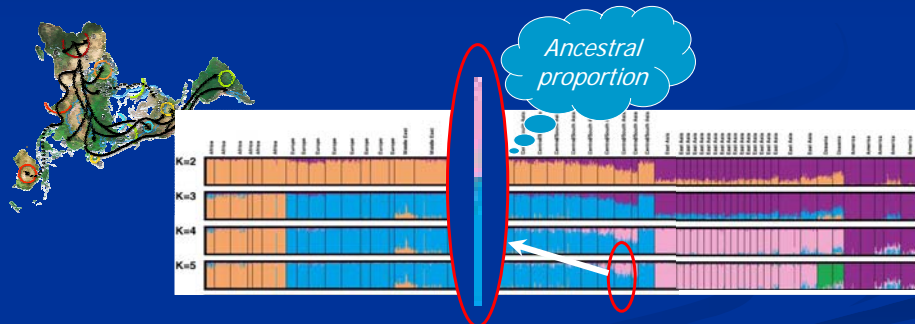
- Deconvolve population structure
 - Ancestral spectrum analysis



51

Genetic Population Structure

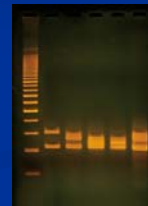
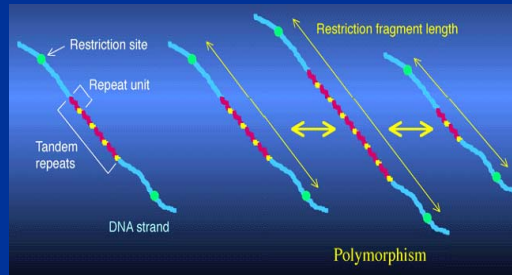
- How to display population structure?
 - *Structure*



Genetic structure of Human Populations (Rosenberg et al. 2002)

52

Variable Number of Tandem Repeats (VNTR) Polymorphism

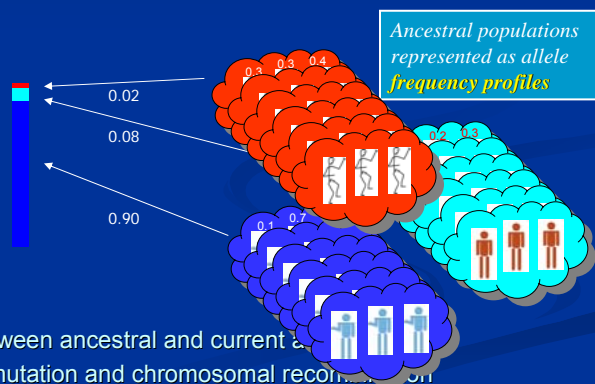


53

The Admixture Model

- Admixture of "ancestral frequency profiles (AP)"

Structure 2.1

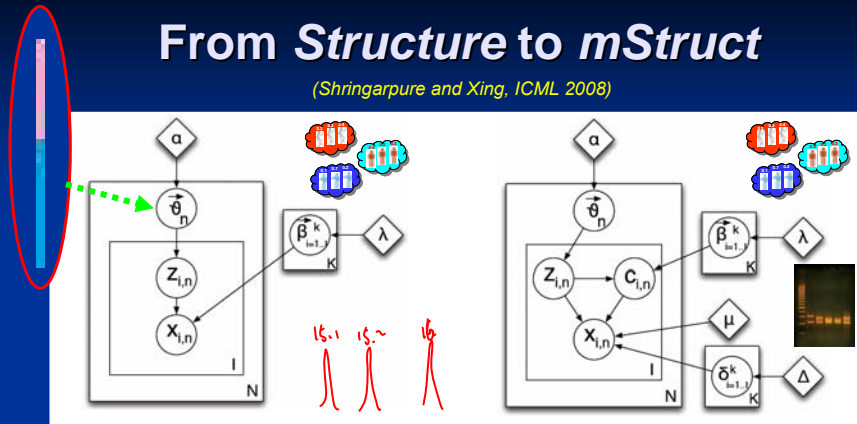


- No distinction between ancestral and current populations
- Does not model mutation and chromosomal recombination

54

From Structure to mStruct

(Shringarpure and Xing, ICML 2008)



- From admixture of APs to admixture of MIMs
 - MiM: population-specific Mixture of Inheritance Models

- The inheritance model:

- Microsatellite:

$$P(b|a) = \frac{1-\delta}{1-\delta^a + \delta} \delta^{b-a}$$

- SNPs:

$$P(b|a) = \delta^{T[b=a]} \times (1-\delta)^{T[b \neq a]}; \quad a, b \in \{0, 1\}.$$

55

Variational Inference

- The joint:

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}, \mathbf{c}, \bar{\theta} | \alpha, \beta, \mu, \delta) &= p(\bar{\theta} | \alpha) \prod_{i=1}^I p(z_i | \bar{\theta}) p(c_i | z_i, \bar{\beta}_i^{k=1..K}) p(x_i | c_i, z_i, \mu_i, \delta_i^{k=1..K}) \\ &= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \left(\prod_{k=1}^K \theta_k^{\alpha_k - 1} \right) \prod_{i=1}^I \prod_{k=1}^K \theta_k^{z_{i,k}} \prod_{l=1}^{L_i} (\beta_{i,l}^k)^{c_{i,l} z_{i,k}} f(x_i | \mu_{i,l}, \delta_i^k)^{c_{i,l} z_{i,k}}. \end{aligned}$$

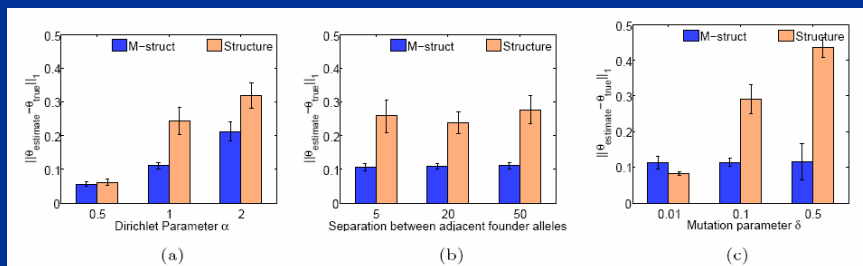
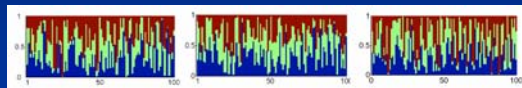
- We can sample z , c , and θ as in *Structure* --- slow
- Alternatively, we approximate $p(z, c, \theta | \mathbf{x})$ by $q(z, c, \theta) = q(z)q(c)q(\theta)$
 - Minimizing $KL(q|p)$:

$$\begin{aligned} q(\bar{\theta}) &\propto \prod_{k=1}^K \theta_k^{\alpha_k - 1 + \sum_{i=1}^I \langle z_{i,k} \rangle} \\ q(c_i) &\propto \prod_{l=1}^{L_i} \left(\prod_{k=1}^K (\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k))^{\langle z_{i,k} \rangle} \right)^{c_{i,l}} \\ q(z_i) &\propto \prod_{k=1}^K \left(e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L_i} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right)^{z_{i,k}} \right) \end{aligned}$$

- Fixed-point iteration ...

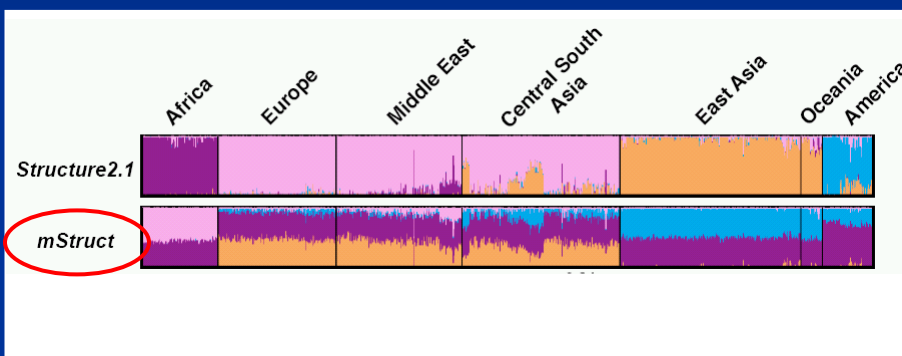
56

Accuracy of Admixing Vector Est.



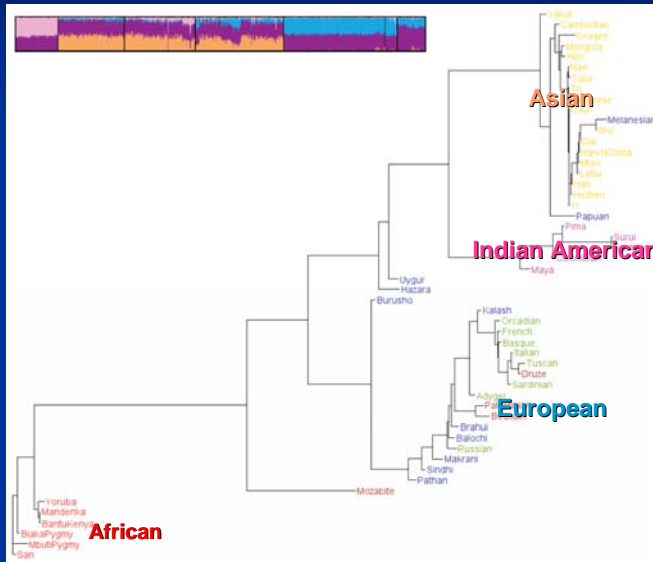
57

Maps under mStruct and Structure



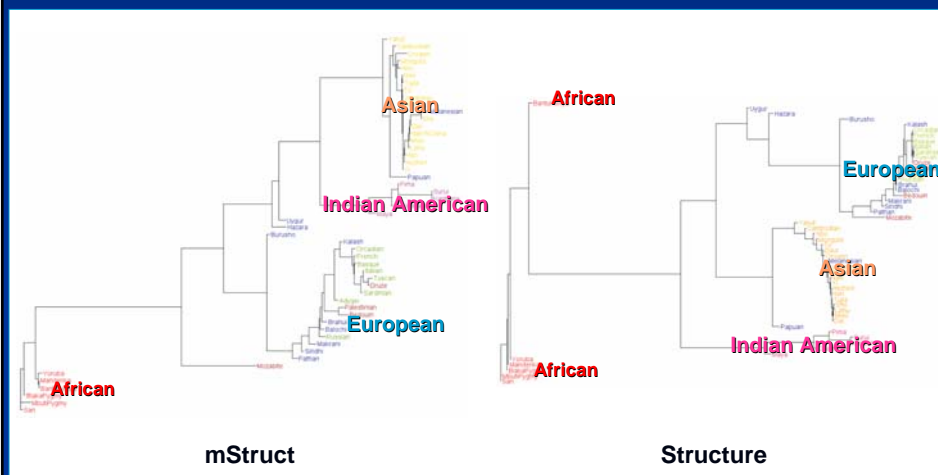
58

Phylogenetic Trees from the Structural Maps



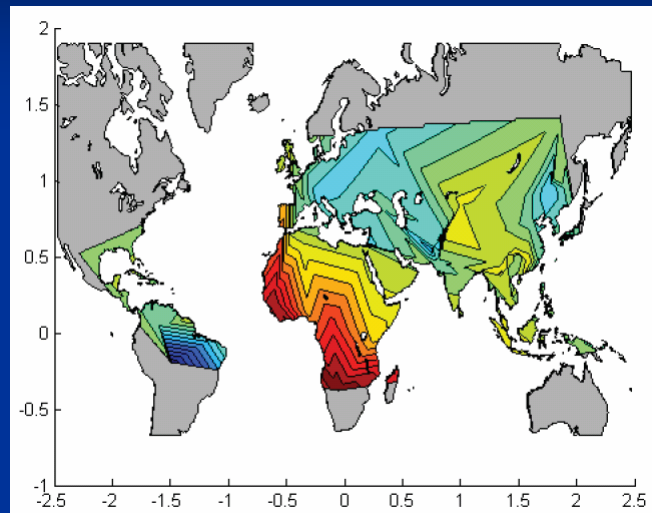
59

Phylogenetic Trees from the Structural Maps



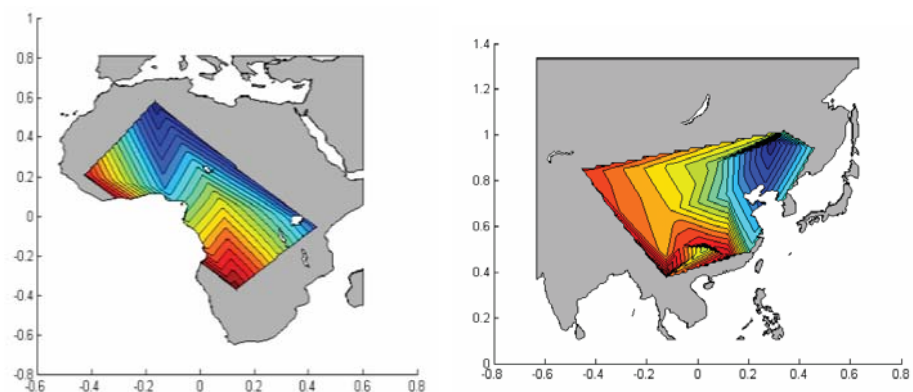
60

Contour of Mutation Rates



61

Contour of Mutation Rates



62

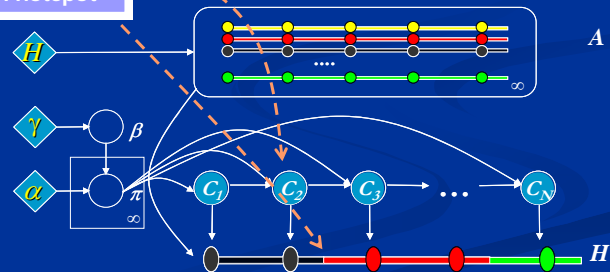
Recombination and Structure

(Sohn and Xing, ISMB, 2007)

Ancestor allele reconstruction

Inferring population structure

Inferring recombination hotspot

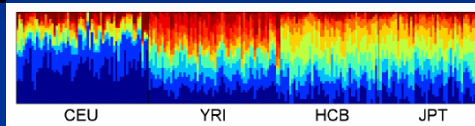


63

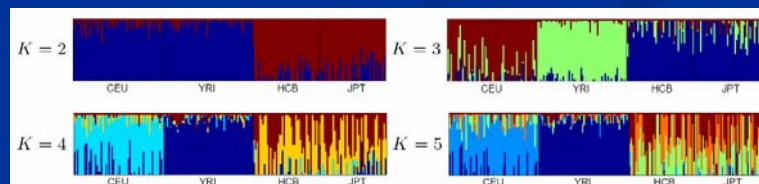
Population Structure Analysis

(Sohn and Xing, ISMB 2007)

Spectrum



Structure 2.1

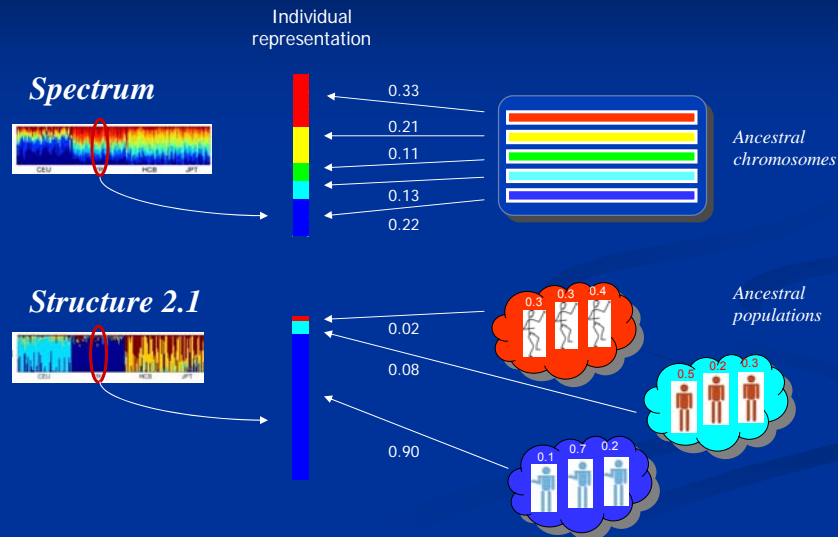


HapMap four population data

CEU: Utah residents with European ancestry
YRI: Yoruba in Ibadan, Nigeria
HCB: Han Chinese in Beijing
JPT: Japanese in Tokyo

64

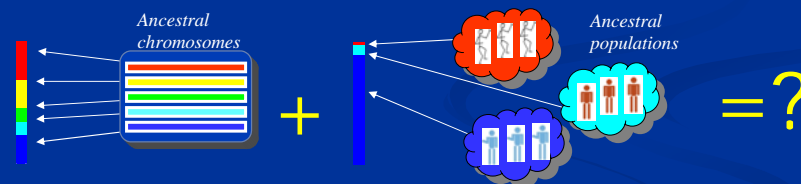
Representational Difference



65

Next Step ...

- A new Bayesian method, *Spectrum*, for jointly modeling recombination and population structure
- Combination of *Spectrum* and *Structure*?



- Computational issues
 - *split-merge MCMC*

66

Summary

■ Haplotype Inference

- Nonparametric Bayesian Models (Dirichlet Process) for phasing single and multiple population

■ Linkage-disequilibrium and GWS

- Hidden Markov DP for identifying recombination hotspots
- Block-Lasso for QTL mapping

■ Population structure analysis

- Bayesian Admixture model

67

Spectrum

Goal

Joint inference of

- Population structure and
- Recombination hotspots and
- Haplotypes
- under unified statistical framework for genetic inheritance process of recombination and mutation
- among an unspecified number of founding alleles

Some open issues:

- Coalescent rate estimation
- Modeling selection, drift, migration, etc.
- Scalability

68

Acknowledgments

- The CMU SAILING group
 - Seyoung Kim
 - Kyung-Ah Sohn
 - Suyash Shringarpure
 - ...

- Pennsylvania Department of Health's Health Research Program No. 2001NF-Cancer Health Research Grant ME-01-739.

- NSF CCF-0523757
- NSF IIS-0713379
- NSF CAREER Award Grant No. DBI-054694
- Sloan Foundation