

# International HapMap Project Resource For Association Studies of The Future

Gonçalo Abecasis  
Center for Statistical Genetics  
University of Michigan



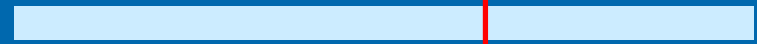
# Genetic Association Studies

- Identify genetic variants with relatively small individual contributions to disease risk
- Require detail measurement of genetic variation
  - > 10,000,000 catalogued genetic variants, so ...
  - Until recently, limited to candidate genes or regions
    - A hit-and-miss approach...
  - Decreasing assay costs allow comprehensive studies
- A few variants could represent all common variants
  - But identifying this subset could be challenging ...

# Linkage Disequilibrium

- Chromosomes are mosaics
- Shaped by
  - Recombination, Mutation, Drift
- Tightly linked markers
  - Reflect ancestral haplotypes
  - Alleles associated

Ancestor

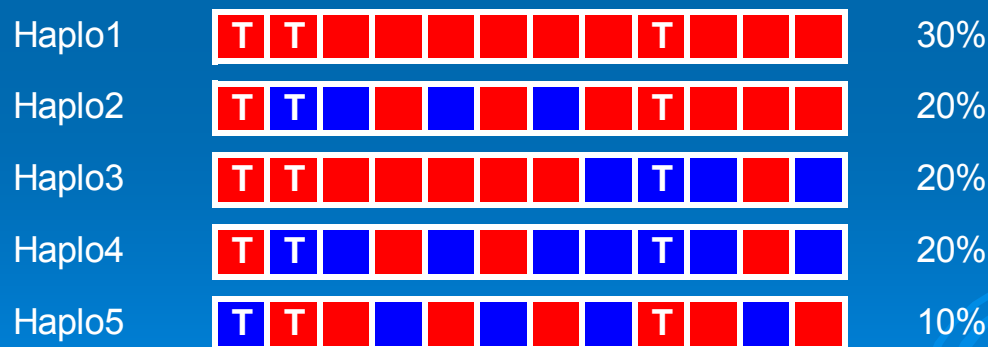


Present-day



# Tagging SNPs

- In a typical short chromosome segment, there are only a few distinct haplotypes
- Carefully selected SNPs can determine status of other SNPs



# HapMap Project

- High-density SNP genotyping across the genome will provide information about
  - SNP validation, frequency, assay conditions
  - Relationship between alleles in the genome
- Public resource to increase efficiency of association for medically relevant traits
- Data is freely available, [www.hapmap.org](http://www.hapmap.org)

# International Effort

- Genotyping being carried out in:
  - Canada (McGill)
  - China (Beijing, Hong Kong, Shanghai)
  - Japan (Riken)
  - United Kingdom (Sanger)
  - United States (Baylor, Illumina, UCSF/WashU, Broad)
- International Working Groups:
  - Analysis, QA/QC, Dataflow, Ethical Issues, ...

# The Hapmap Samples

- European Ancestry
  - 90 CEPH family members (CEU, 30 trios)
- African Ancestry
  - 90 Yoruba from Ibadan, Nigeria (YRI, 30 trios)
- East Asian
  - 45 Japanese from Tokyo, Japan (JAT)
  - 45 Han Chinese from Beijing, China (HCB)

# Current Status

- Data publicly available for >1,000,000 SNPs assayed in 4 populations
  - Including about 700,000 common SNPs
- The first phase provides information on one polymorphic marker per 5kb in each population

# Ongoing Quality Assessment

- Project includes exercises to evaluate and ensure quality of generated data
  - Including evaluating agreement across platforms
- Latest completed round:
  - Data for 1496 markers from public website
  - Repeat data collected with 11 different protocols

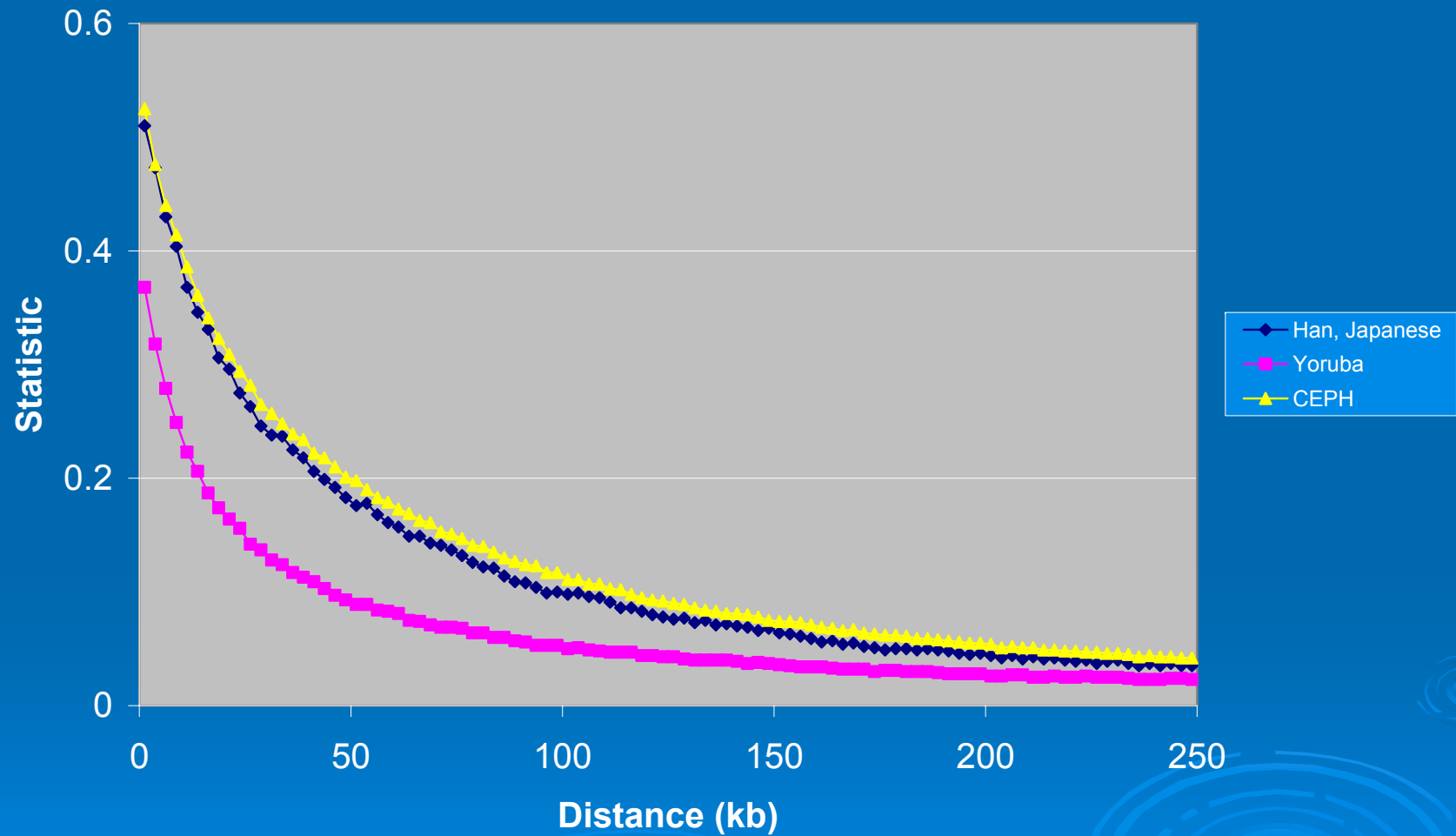
# Consensus Summary

- 116,536 genotypes
- 1,295 polymorphic markers
  - Each Genotyped at 8.8 Centers on Average
- 99.99% completion rate (14 missing genotypes)
- 5 Mendelian inconsistencies
  - In 38,836 trios and 9 parent-offspring pairs
  - Corresponds to "error rate" of 0.00016
  - (But these are probably not errors!)

# Quality Assessment Summary

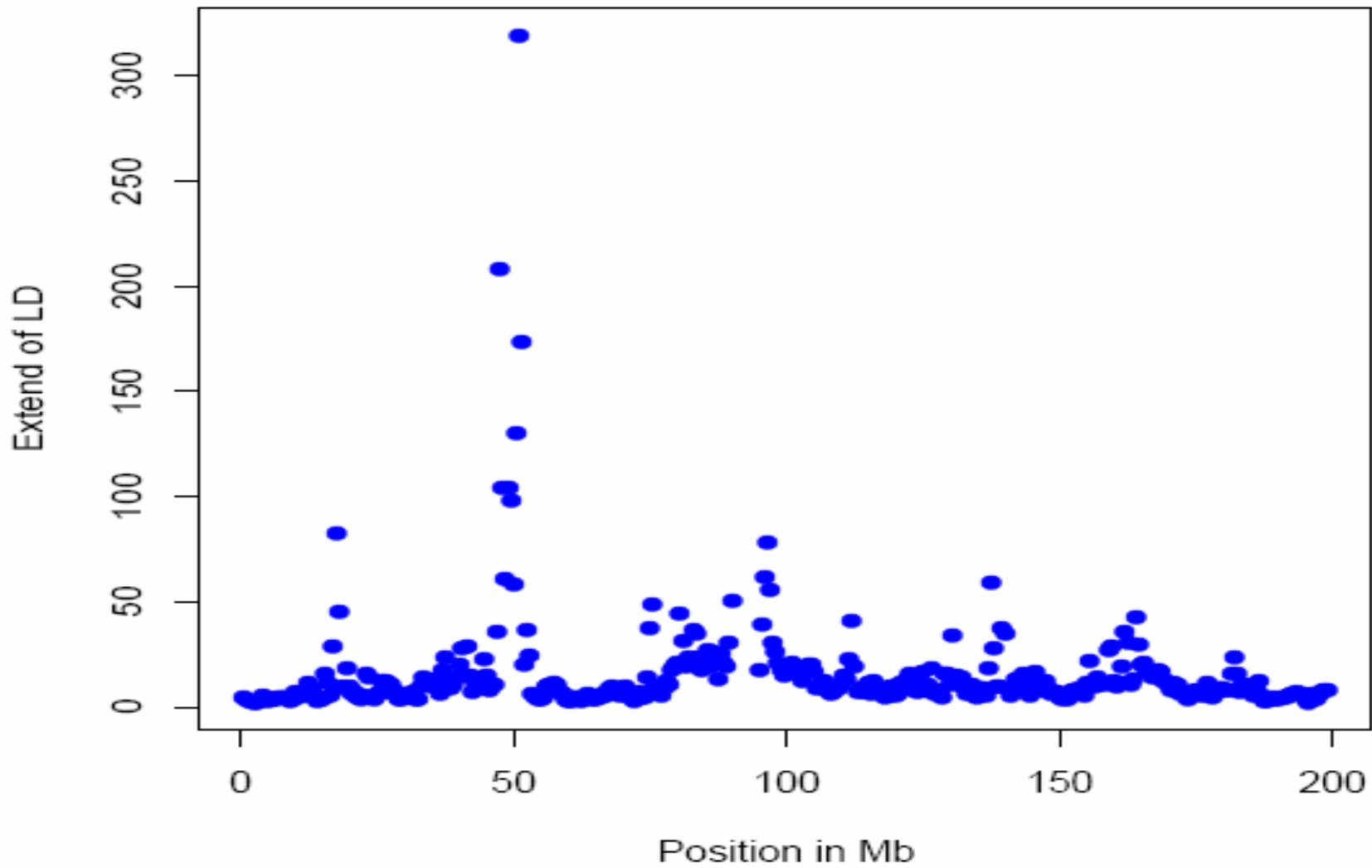
- Error rate is 0.00228 vs original submission
  - Most markers show zero or one differences
  - 3 markers account for most differences
- Error rate seems lower with “internal” checks
  - 0.0006 based on same protocol duplicates
  - 0.0006 based on Mendelian inconsistencies

## Empirical of LD (chromosome 2, $R^2$ )



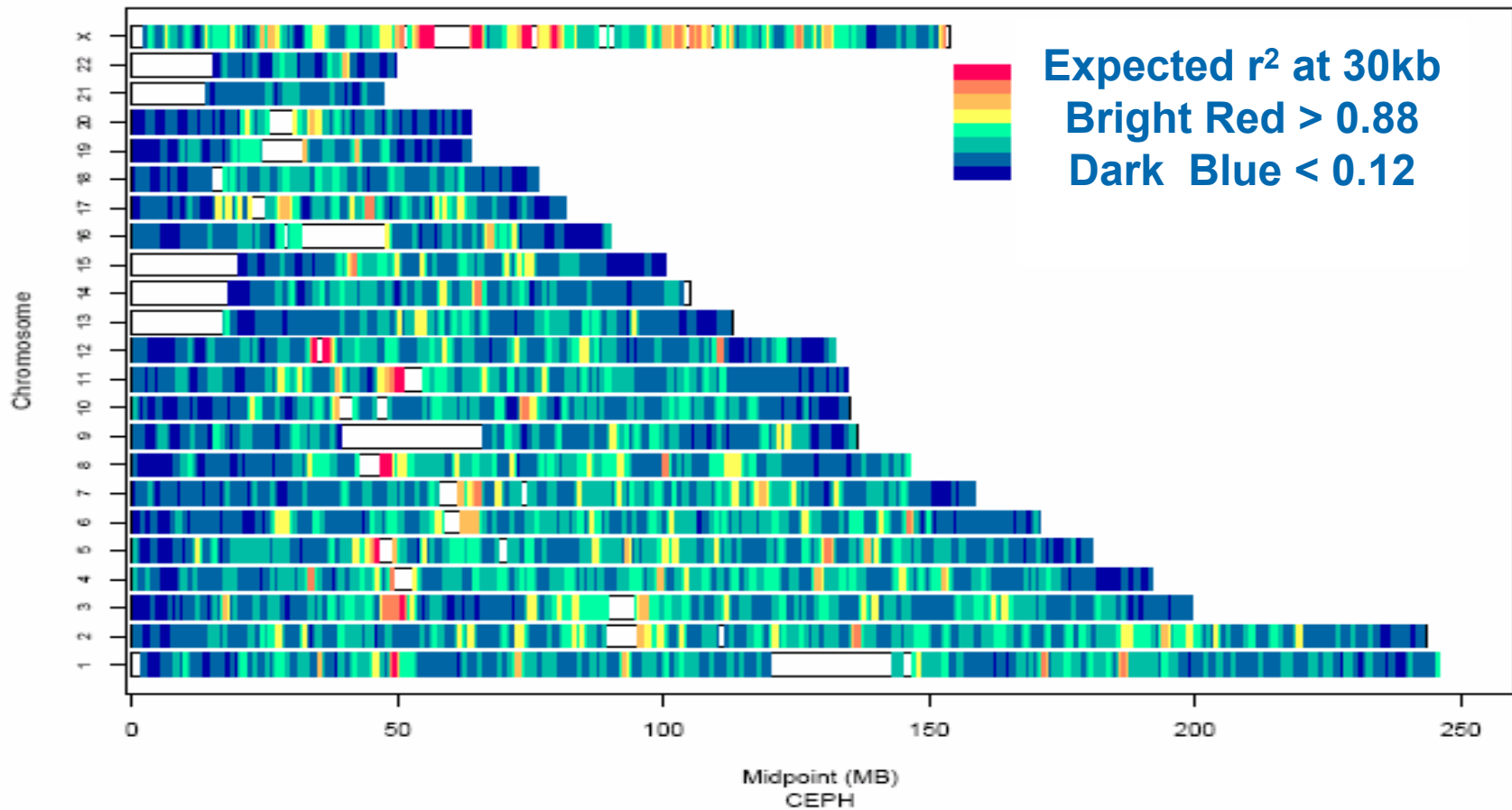
LD extends further in CEPH and the Han/Japanese than in the Yoruba

### Chromosome 3

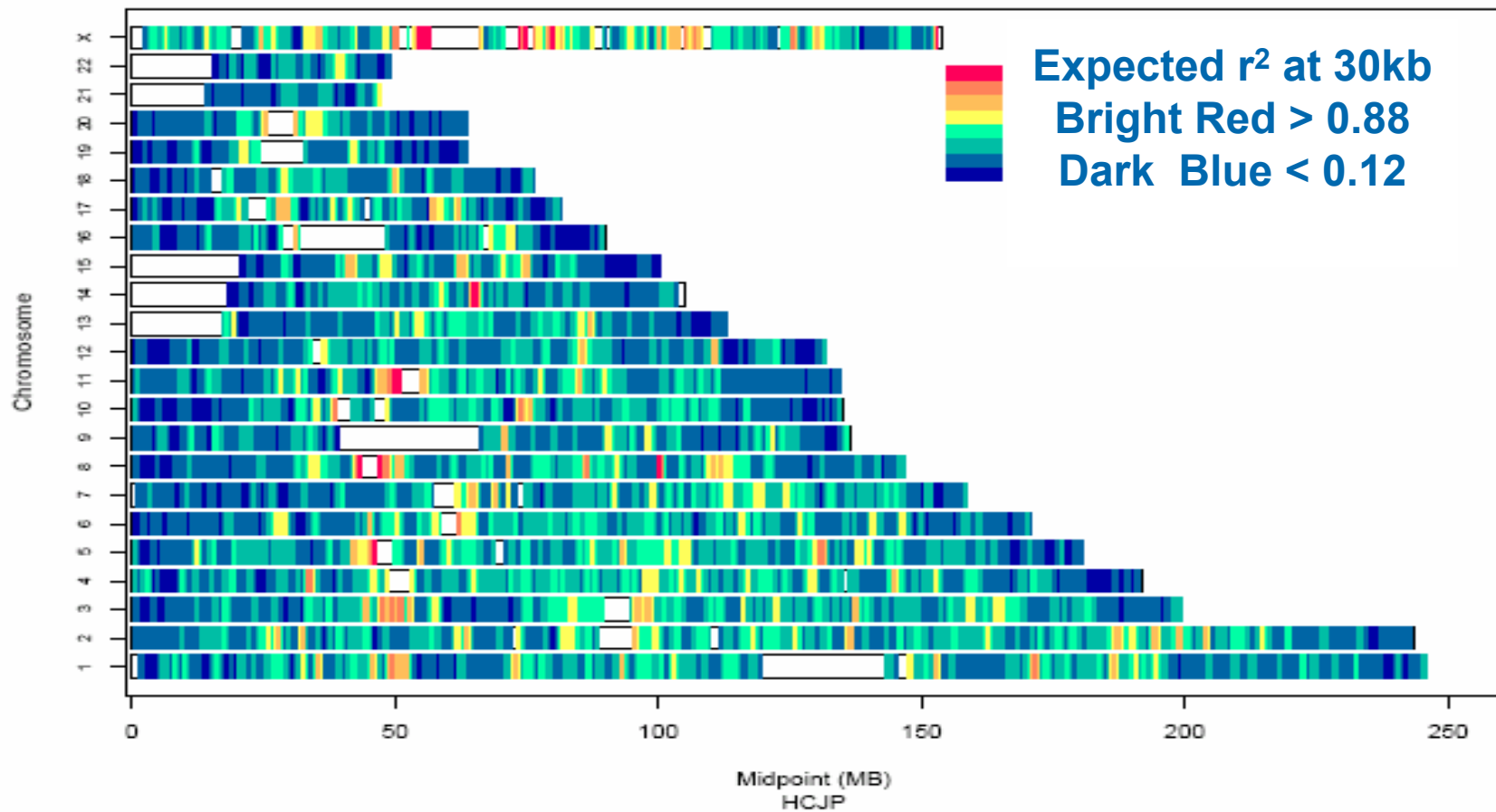


Most chromosomes exhibit regions of extended LD on the megabase scale

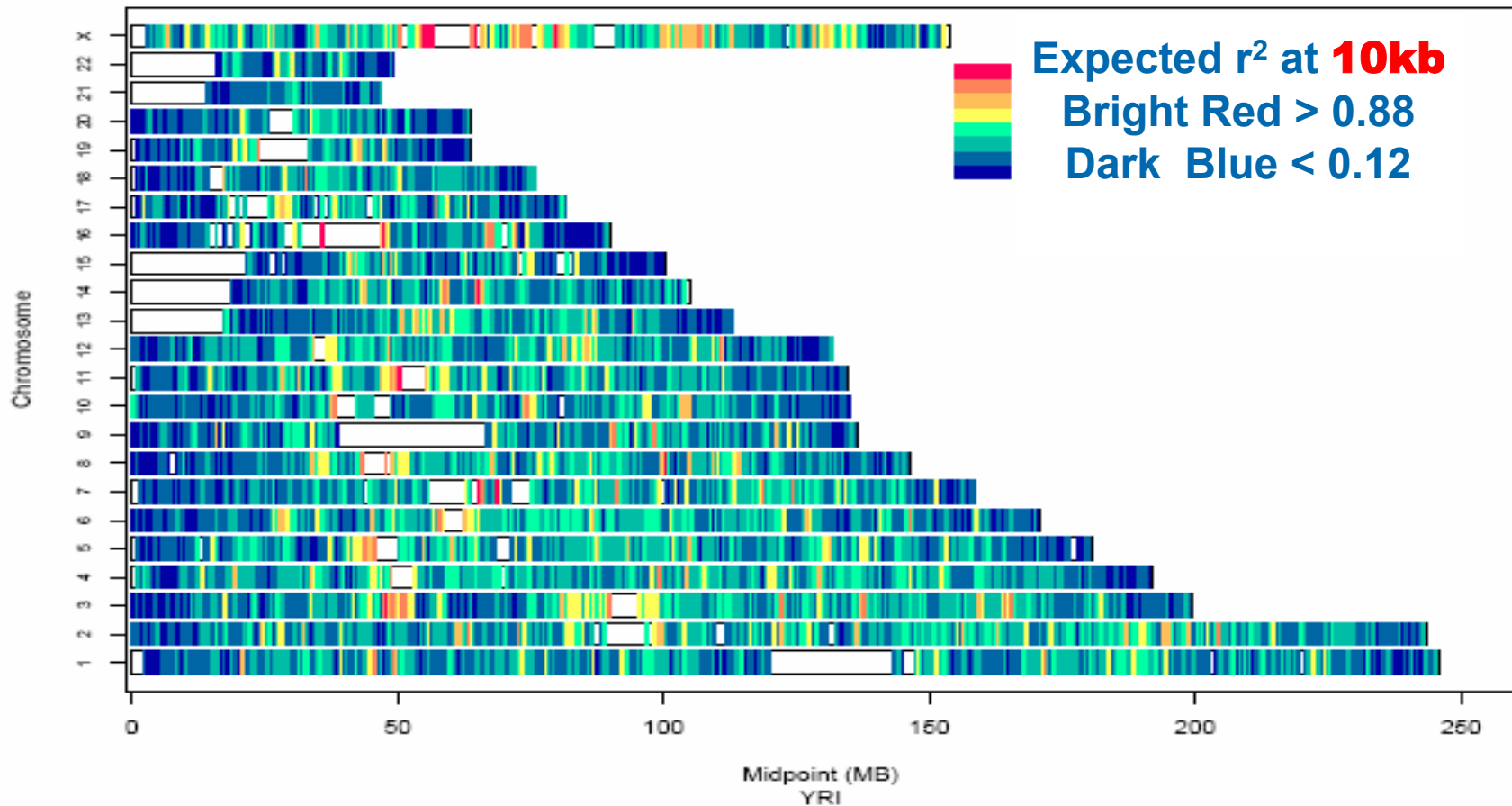
# Genomic Distribution of LD (CEPH)



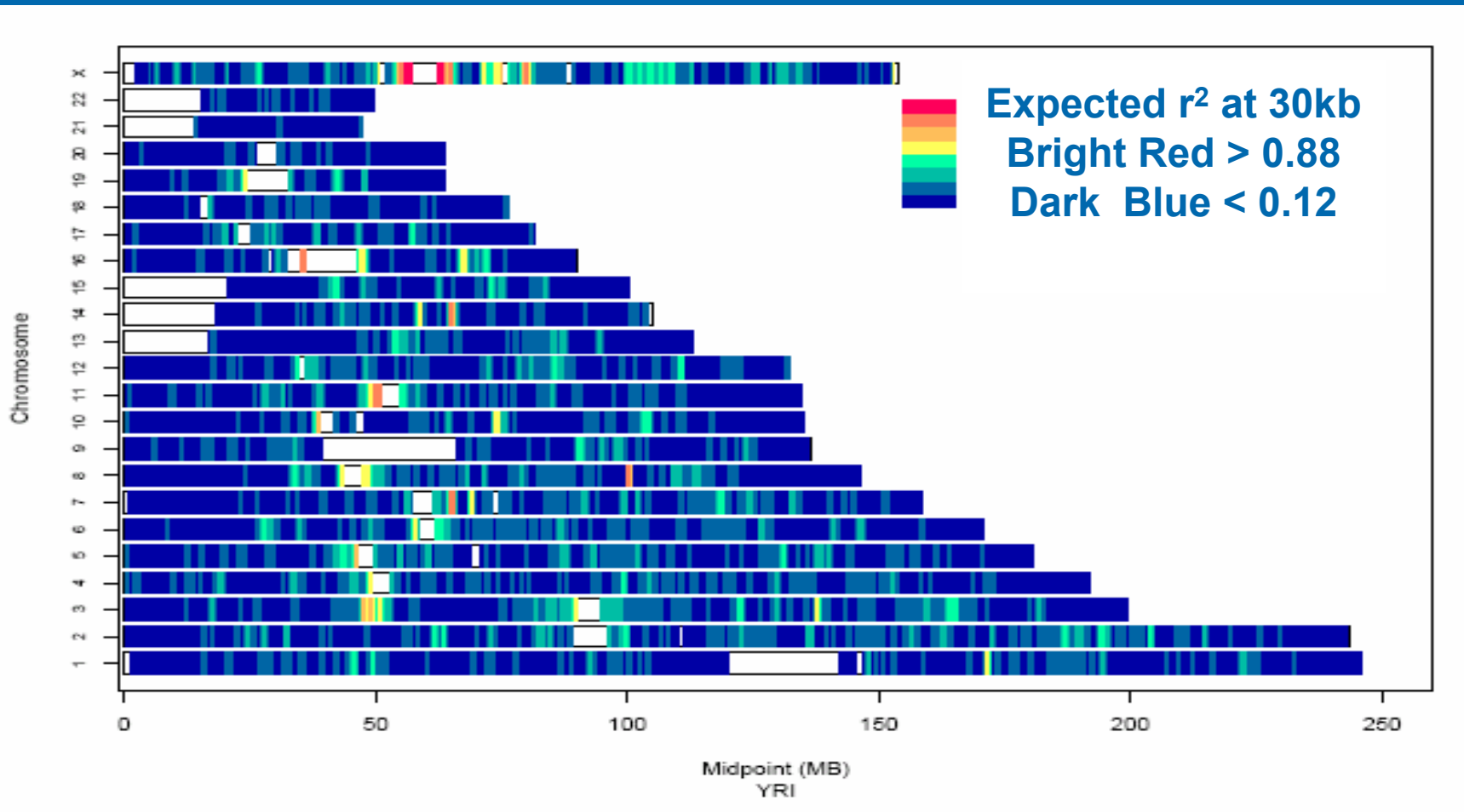
# Genomic Distribution of LD (JPT + HCB)



# Genomic Distribution of LD (YRI)



# Genomic Distribution of LD (YRI)



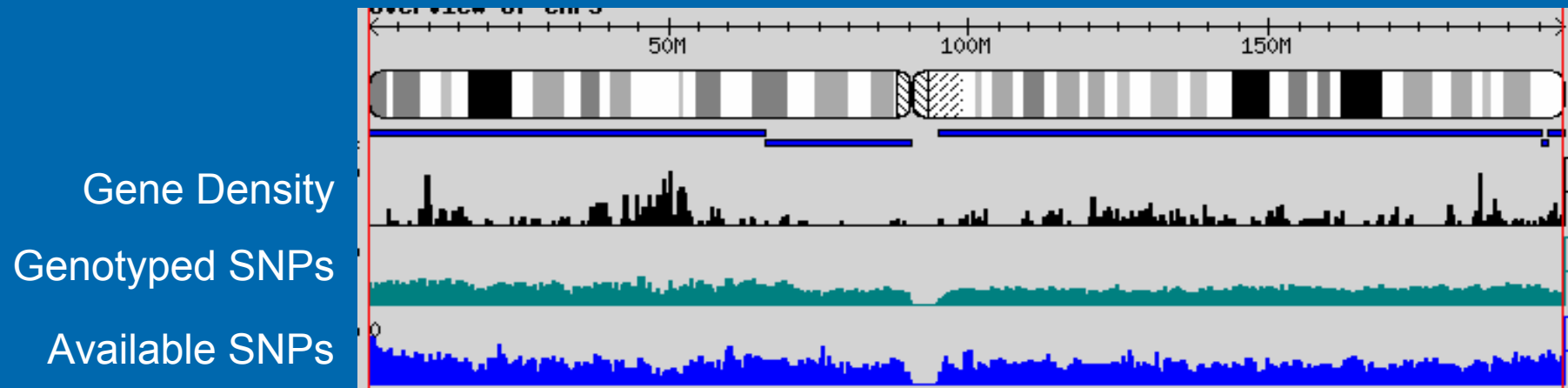
# Big Picture

- Substantial variation in genomic LD
- Global features:
  - More LD on larger chromosomes and X
  - LD is generally low near telomeres
  - LD is generally high near centromeres
  - Relative LD is *usually* similar across samples
- Best predictor of disequilibrium in a region is LD in another population for that region

# Tagging SNPs

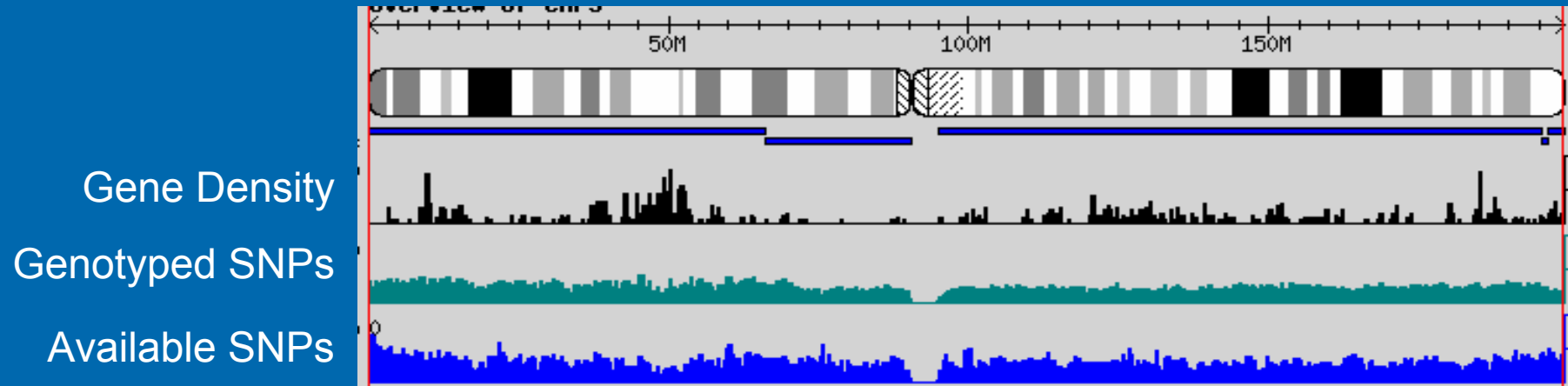
- Fine description of linkage disequilibrium allows identification of SNPs that capture variation in each region
- To illustrate the possibilities, selected set of tagging SNPs using pairwise  $r^2$  criterion
  - SNPs are good surrogates if:
    - Nearly identical frequency
    - Nearly always appear on the same haplotype

# Example: Chromosome 3 Tags



- 45,727 SNPs genotyped (December 2004)
  - Exhaustive tagging:
    - 11,366 tags ( $r^2$  threshold of 0.50)
    - 18,638 tags ( $r^2$  threshold of 0.80)
  - But even fewer tags can be quite useful ...

# Example: Chromosome 3 Tags



- 45,727 SNPs genotyped (December 2004)
  - Exhaustive tagging:
    - 11,366 tags ( $r^2$  threshold of 0.50)
    - 18,638 tags ( $r^2$  threshold of 0.80)
  - Low hanging fruit (the 10% best tags):
    - 1,136 tags capture 21,491 markers ( $r^2$  threshold of 0.50)
    - 1,863 tags capture 18,994 markers ( $r^2$  threshold of 0.80)

# Chromosome 2 Tagging...

## ( $r^2$ threshold of 0.50)

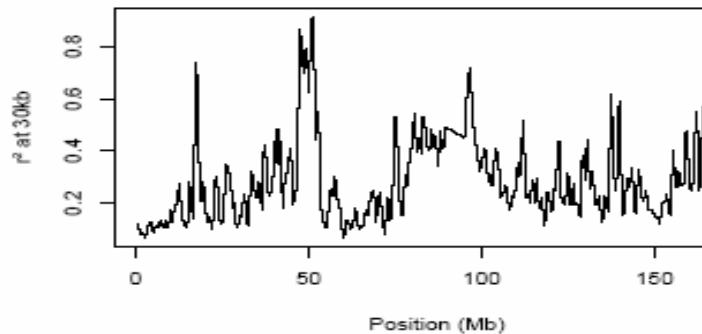
- Exhaustive tagging:
  - CEPH: 13,706 tags for 57,503 common SNPs
  - YRI: 23,710 tags for 53,534 common SNPs
  - HBJT: 11,045 tags for 43,950 common SNPs
- Low hanging fruit (top 10% of all tags)
  - CEPH: 1,370 tags for 27,151 common SNPs
  - YRI: 2,371 tags for 21,724 common SNPs
  - HBJT: 1,104 tags for 20,202 common SNPs
- 10 scans “half-genome scans” focused on low hanging fruit cost about the same as one comprehensive scan
  - But genotype cost may fall too fast to make this necessary!

# So far ...

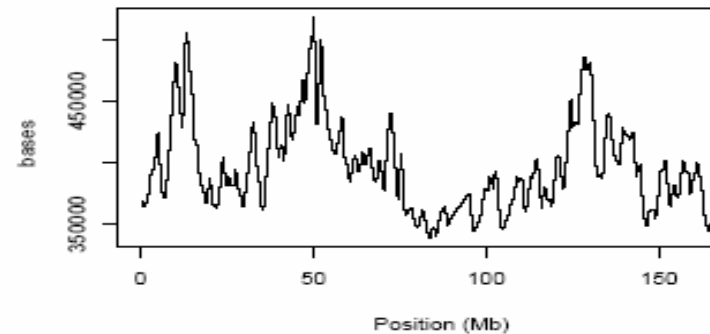
- Description of extensive genomic variation in linkage disequilibrium
  - Similarities and differences between populations
  - Ability to find SNPs that capture variation in any region
- Reflects the main purpose of the project,
  - Aid association studies for medically important traits
- However, data also contain information about interesting and unusual patterns of variation ...

# Does Sequence Variation Predict Linkage Disequilibrium?

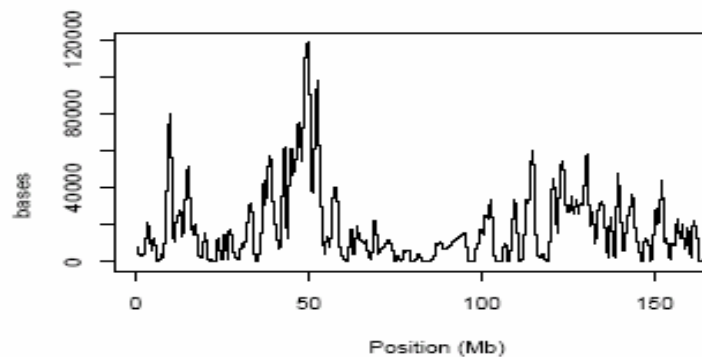
Extent of Linkage Disequilibrium



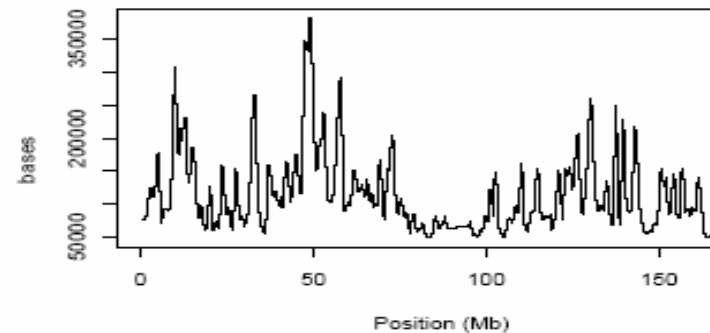
GC Content



Known Exons



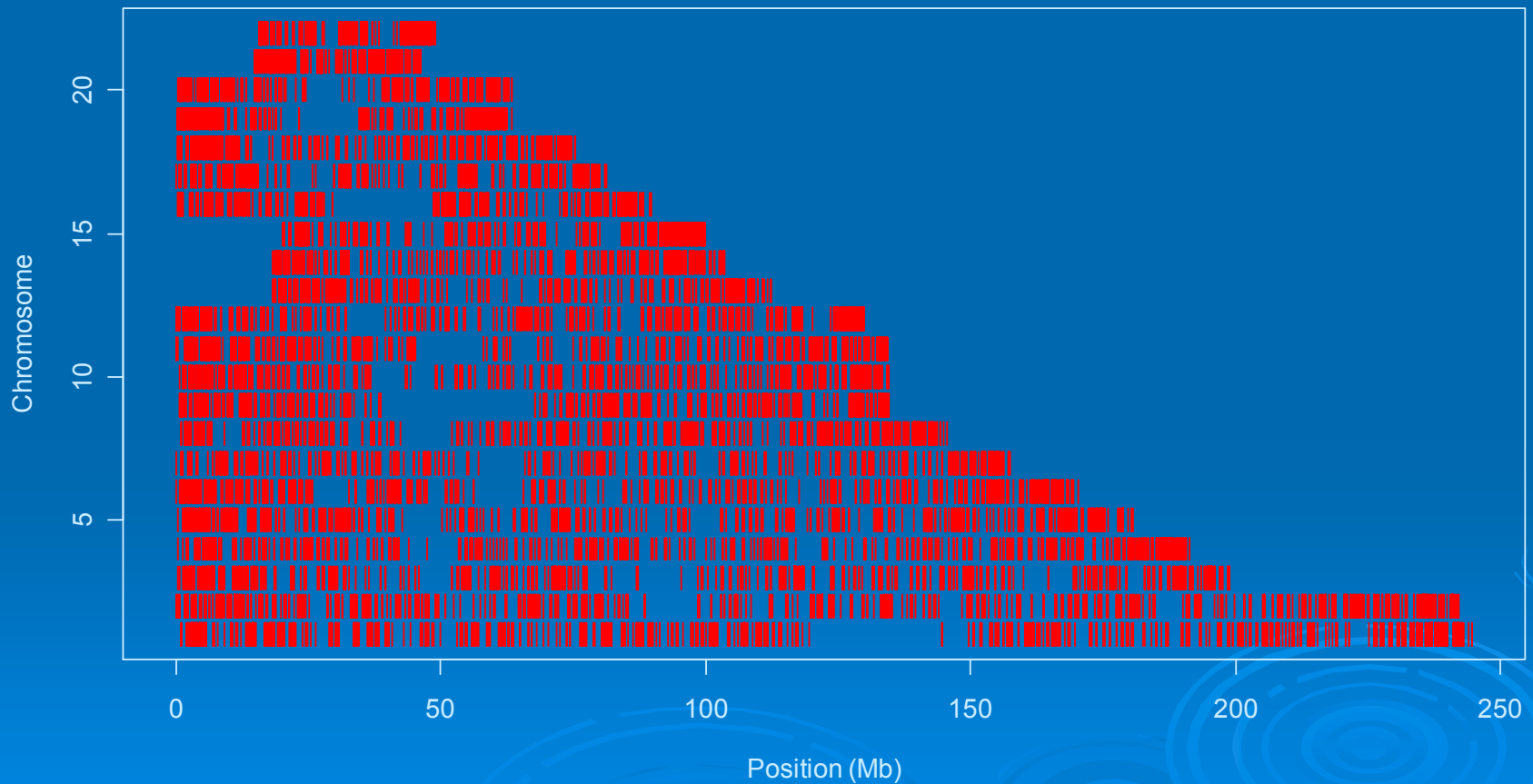
SINE repeats



# Some Very Hot Spots

- High recombination regions
  - Pair of markers within 10kb
  - 5 markers within 40kb to the left
  - 5 markers within 40kb to the right
  - None of 25 pairs shows  $r^2 > 0.10$
- Compared to unselected regions
  - As above, no constraint on LD

# Some Very Hot Spots



# Sequence in Very Hot Intervals

Feature	Candidates	CEU	YRI	JPT, HCB	Association
<b>N of Intervals</b>	*varies	10155	17144	12224	
<b>Centromere (5Mb)</b>	0.047	0.033	0.032	0.031	---
<b>Telomere (5Mb)</b>	0.036	0.079	0.091	0.078	+++
<b>GC</b>	0.407	0.437	0.439	0.437	+++
<b>CpG Island</b>	0.006	0.008	0.008	0.008	+++
<b>Known Exons</b>	0.026	0.021	0.022	0.021	---
<b>Repeat -- LINE</b>	0.168	0.136	0.135	0.134	---
<b>Repeat -- SINE</b>	0.117	0.144	0.143	0.143	+++
<b>Repeat -- Simple</b>	0.008	0.010	0.010	0.010	+++
<b>Segmental Duplications</b>	0.014	0.010	0.011 --	0.010	---

\* Intervals evaluated: 611255 (HCB, JPT), 670675 (CEPH) 696575 (YRI)

# Big Picture

- Best predictor of LD is information about a different sample
- Multiple genomic features appear to be associated with LD
  - GC, repeat types, exons, conserved sequences
- We are in the process of comparing the relationship between gene function and LD

# Linkage Analysis with Markers in Linkage Disequilibrium

Gonçalo Abecasis  
University of Michigan



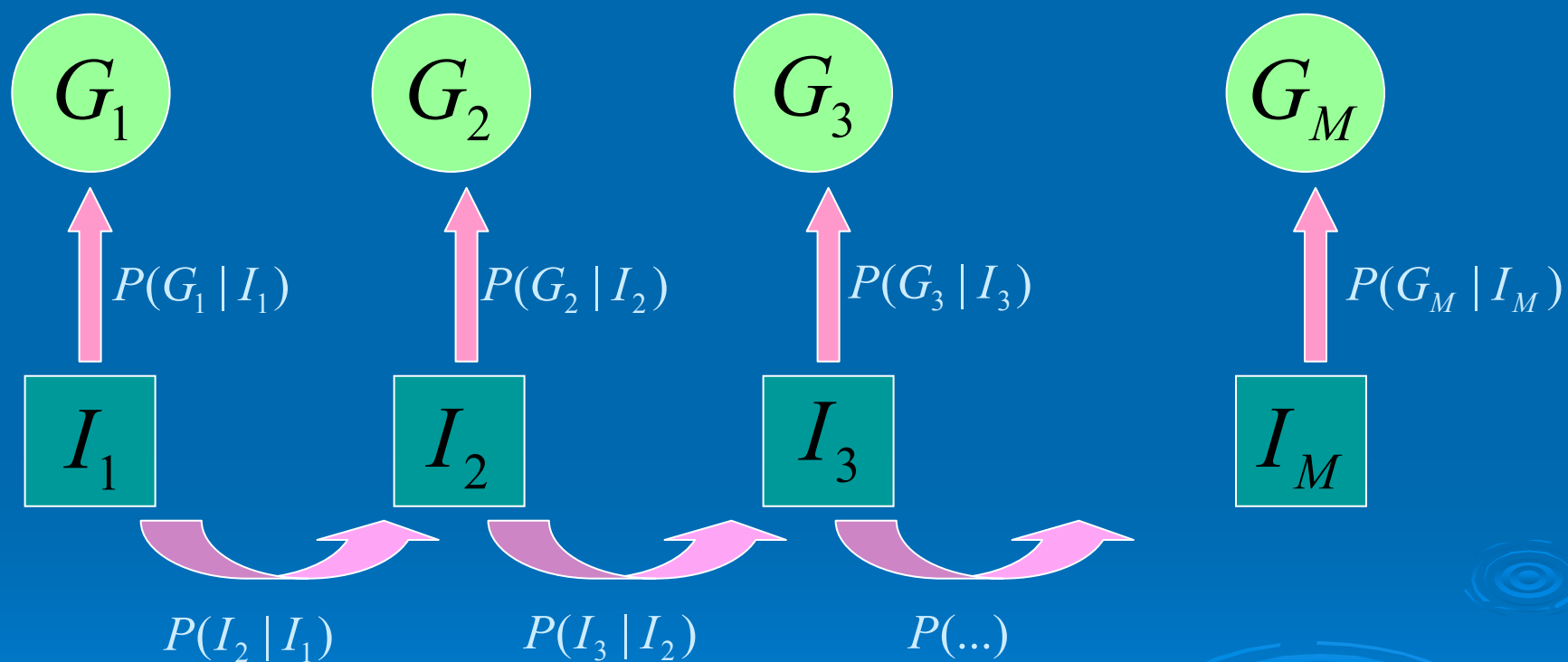
# SNPs

- Abundant diallelic genetic markers
- Amenable to automated genotyping
  - Fast, cheap genotyping with low error rates
- Rapidly replacing microsatellites in many linkage studies

# The Problem

- Linkage analysis methods assume that markers are in linkage equilibrium
  - Violation of this assumption can produce large biases
- This assumption affects ...
  - Parametric and nonparametric linkage
  - Variance components analysis
  - Haplotype estimation

# Standard Hidden Markov Model

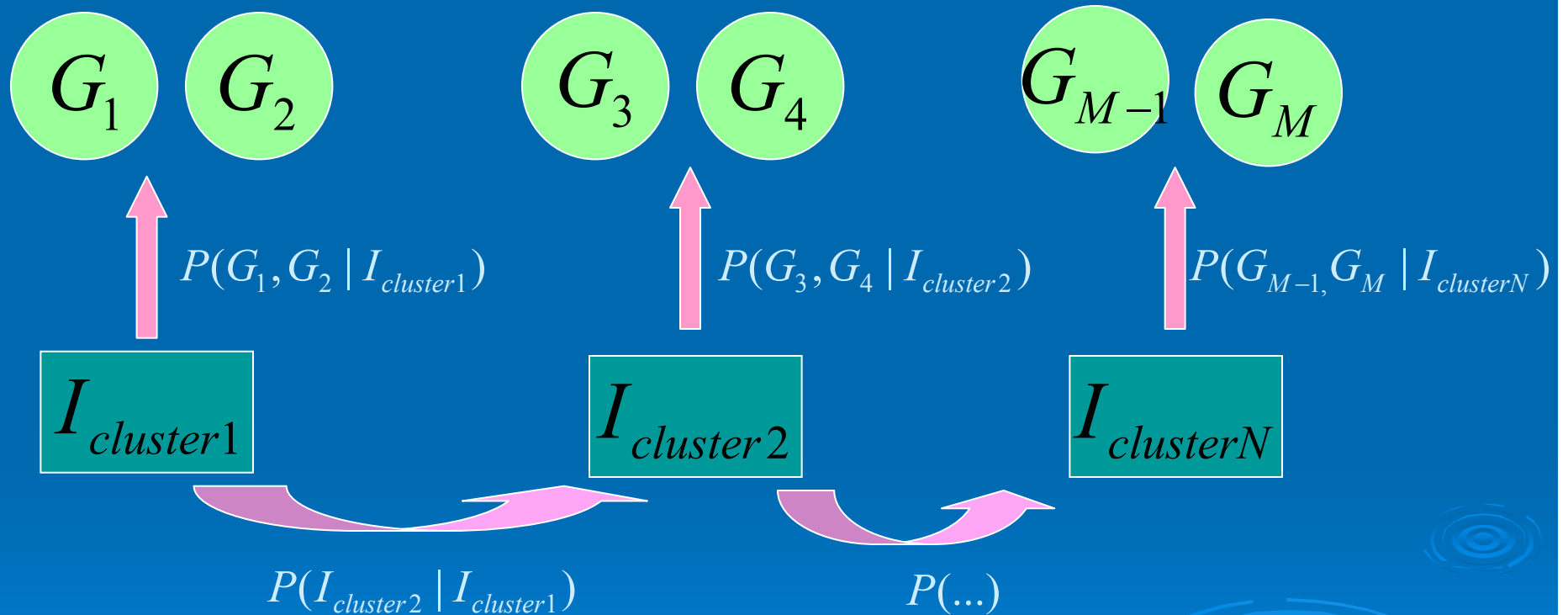


Observed Genotypes Are Connected Only Through IBD States ...

# Our Approach

- Cluster groups of SNPs in LD
  - Assume no recombination within clusters
  - Estimate haplotype frequencies
  - Sum over possible haplotypes for each founder
- Two pass computation ...
  - Group inheritance vectors that produce identical sets of founder haplotypes
  - Calculate probability of each distinct set

# Hidden Markov Model



Example With Clusters of Two Markers ...

# Practically ...

$$\begin{aligned} P(G_1 \dots G_C | f_1 \dots f_h, \nu) &= \sum_{H_1=1}^h \dots \sum_{H_{2f}=1}^h \Pr(G_1 \dots G_C | H_1 \dots H_{2f}, \nu) \Pr(H_1 \dots H_{2f} | f_1 \dots f_h) \\ &= \sum_{H_1=1}^h \dots \sum_{H_{2f}=1}^h \Pr(G_1 \dots G_C | H_1 \dots H_{2f}, \nu) \prod_{i=1}^{2f} \Pr(H_i | f_1 \dots f_h) \end{aligned}$$

- Probability of observed genotypes  $G_1 \dots G_C$ 
  - Conditional on haplotype frequencies  $f_1 \dots f_h$
  - Conditional on a specific inheritance vector  $\nu$
- Calculated by iterating over founder haplotypes

# Computationally ...

- Avoid iteration over  $h^{2f}$  founder haplotypes
  - List possible haplotype sets for each cluster
  - List is product of allele graphs for each marker
- Group inheritance vectors with identical lists
  - First, generate lists for each vector
  - Second, find equivalence groups
  - Finally, evaluate nested sum once per group

# Simulations ...

- 2000 genotyped individuals per dataset
  - 0, 1, 2 genotyped parents per sibship
  - 2, 3, 4 genotyped affected siblings
- Clusters of 3 markers, centered 5 cM apart
  - Used Hapmap to generate haplotype frequencies
    - Clusters of 3 SNPs in 100kb windows
    - Windows are 5 Mb apart along chromosome 13
    - All SNPs had minor allele frequency > 5%
  - Simulations assumed 1 cM / Mb

# Average LOD Scores (Null Hypothesis)

Analysis Strategy	Average LOD		
	Ignore LD	Model LD	Independent SNPs
<b>No parents genotyped</b>			
... 2 sibs per family	1.516	-0.010	-0.005
... 3 sibs per family	2.695	-0.002	-0.002
... 4 sibs per family	2.412	-0.002	-0.005
<b>One parent genotyped</b>			
... 2 sibs per family	0.571	0.008	0.005
... 3 sibs per family	0.660	-0.021	-0.028
... 4 sibs per family	0.522	0.002	0.007
<b>Two parents genotyped</b>			
... 2 sibs per family	0.030	-0.003	-0.005
... 3 sibs per family	0.014	-0.016	-0.021
... 4 sibs per family	0.014	-0.007	-0.006

# 5% Significance Thresholds (based on peak LODs under null)

Analysis Strategy	Significance Threshold		
	Ignore LD	Model LD	Independent SNPs
<b>No parents genotyped</b>			
... 2 sibs per family	11.90	1.18	1.18
... 3 sibs per family	18.06	1.38	1.29
... 4 sibs per family	15.52	1.24	1.22
<b>One parent genotyped</b>			
... 2 sibs per family	5.52	1.54	1.36
... 3 sibs per family	5.84	1.25	1.20
... 4 sibs per family	4.79	1.42	1.36
<b>Two parents genotyped</b>			
... 2 sibs per family	1.64	1.47	1.36
... 3 sibs per family	1.51	1.41	1.31
... 4 sibs per family	1.41	1.34	1.26

# Empirical Power

Analysis Strategy	Power		
	Ignore LD	Model LD	Independent SNPs
<b>No parents genotyped</b>			
... 2 sibs per family	0.097	0.304	0.247
... 3 sibs per family	0.133	0.540	0.405
... 4 sibs per family	0.196	0.874	0.680
<b>One parent genotyped</b>			
... 2 sibs per family	0.088	0.158	0.161
... 3 sibs per family	0.213	0.562	0.430
... 4 sibs per family	0.397	0.782	0.608
<b>Two parents genotyped</b>			
... 2 sibs per family	0.141	0.163	0.145
... 3 sibs per family	0.436	0.435	0.401
... 4 sibs per family	0.779	0.779	0.701

Disease Model,  $p = 0.10$ ,  $f_{11} = 0.01$ ,  $f_{12} = 0.02$ ,  $f_{22} = 0.04$

# Conclusions from Simulations

- Modeling linkage disequilibrium crucial
  - Especially when parental genotypes missing
- Ignoring linkage disequilibrium
  - Inflates LOD scores
  - Both small and large sibships are affected
  - Loses ability to discriminate true linkage

# Example

- Psoriasis data of Stuart et al (2005)
- 274 families
  - 2 or more affected individuals
  - Up to 21 genotyped individuals per family
- Initial microsatellite scan, additional fine-mapping markers added to chromosome 17 candidate region

