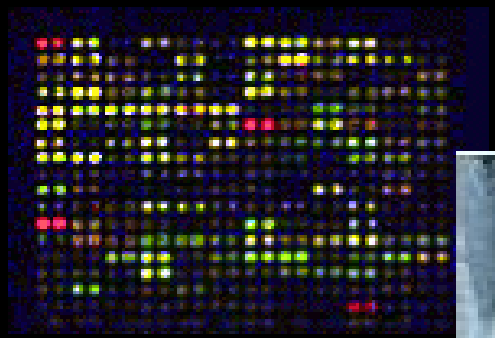
The utility of the HapMap reference samples for clinical populations

(the informatics of sequence variations and haplotypes)



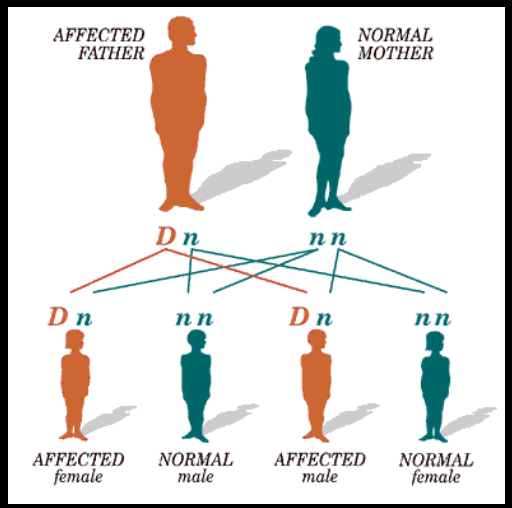
VanBug Seminar
Vancouver, BC, Canada
September 9, 2004

Gabor T. Marth

Department of Biology, Boston College
marth@bc.edu

Why do we care about variations?

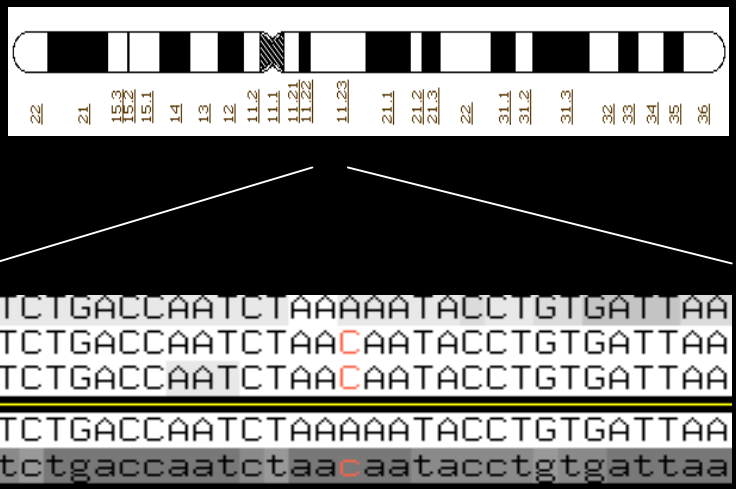
underlie phenotypic differences



cause inherited diseases

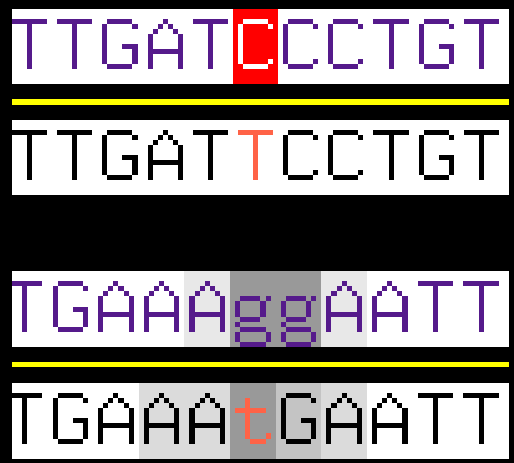
allow tracking ancestral human history

How do we find sequence variations?

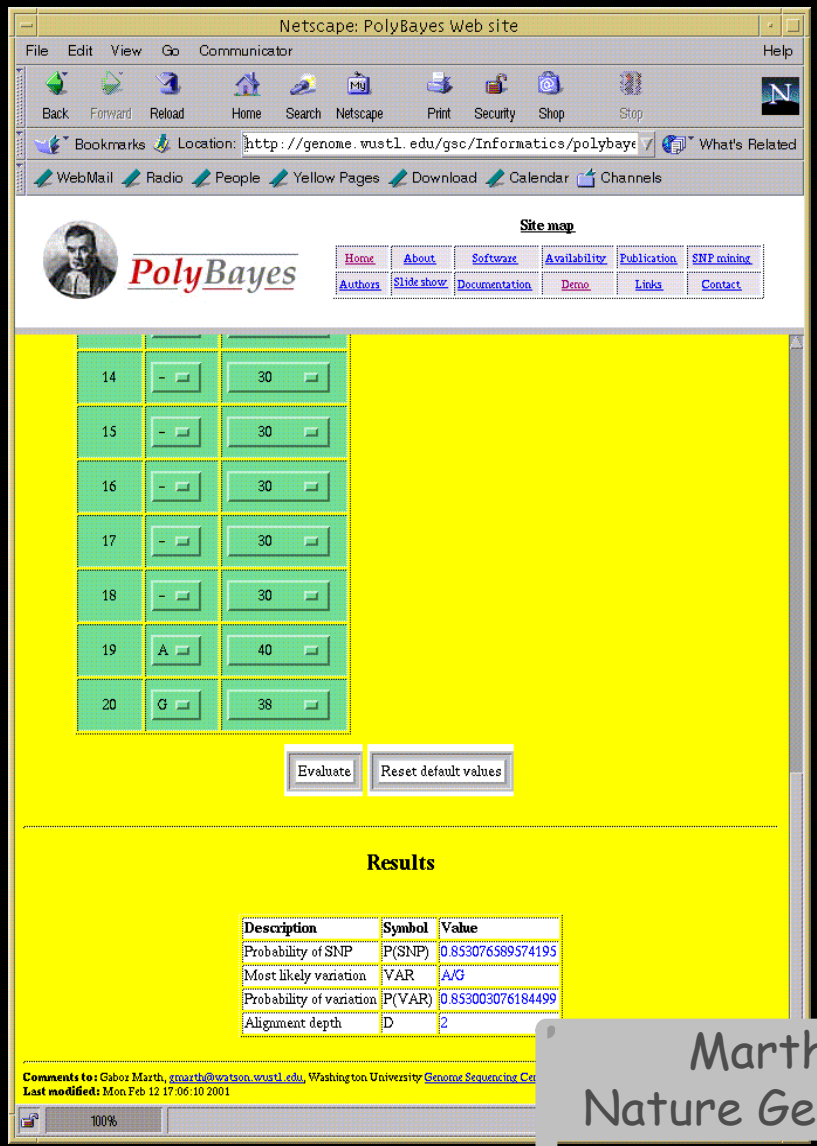


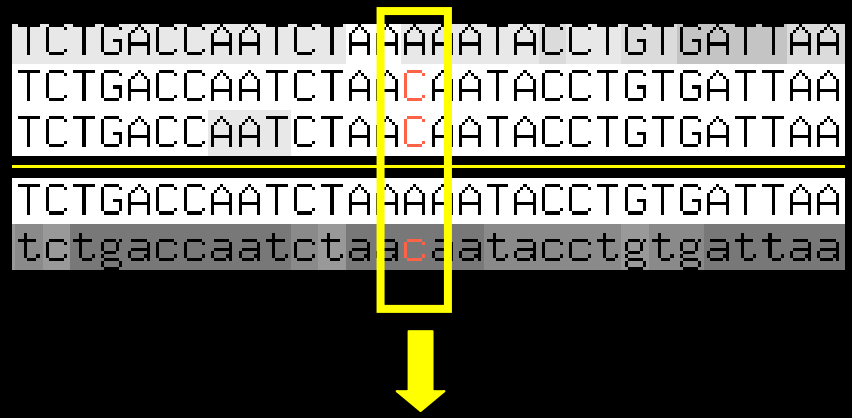
- look at multiple sequences from the same genome region

- use base quality values to decide if mismatches are true polymorphisms or sequencing errors



Automated polymorphism discovery

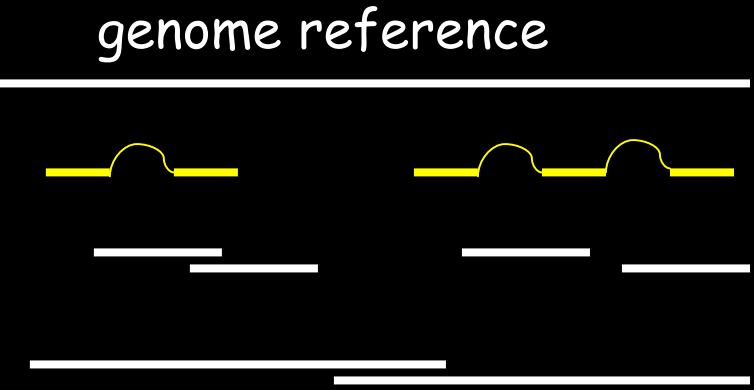


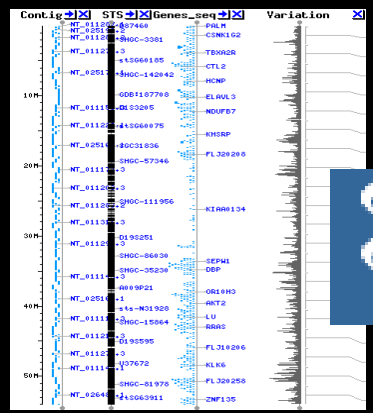


$$P(SNP) = \sum_{all\ variable\ S} \frac{\frac{P(S_1|R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N|R_N)}{P_{Prior}(S_N)} \cdot P_{Prior}(S_1,\ldots,S_N)}{\sum_{S_{i_1} \in [A,C,G,T]} \ldots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1}|R_1)}{P_{Prior}(S_{i_1})} \cdot \ldots \cdot \frac{P(S_{i_N}|R_1)}{P_{Prior}(S_{i_N})} \cdot P_{Prior}(S_{i_1},\ldots,S_{i_N})}$$

Marth *et al.*
Nature Genetics 1999

Large SNP mining projects

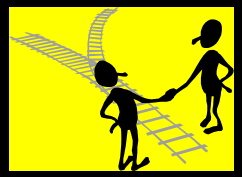




~ 8 million

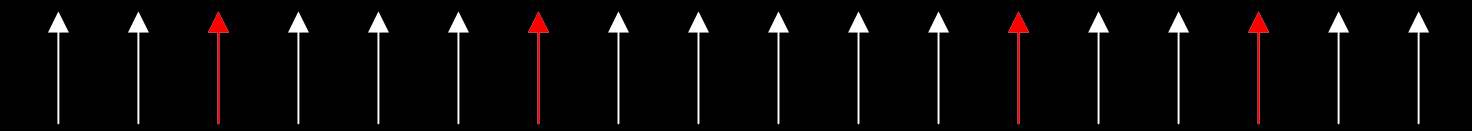
NCBI

Sachidanandam *et al.*
Nature 2001

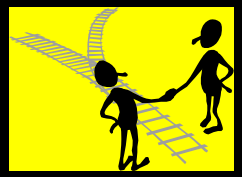


How to use markers to find disease?

genome-wide, dense SNP marker map

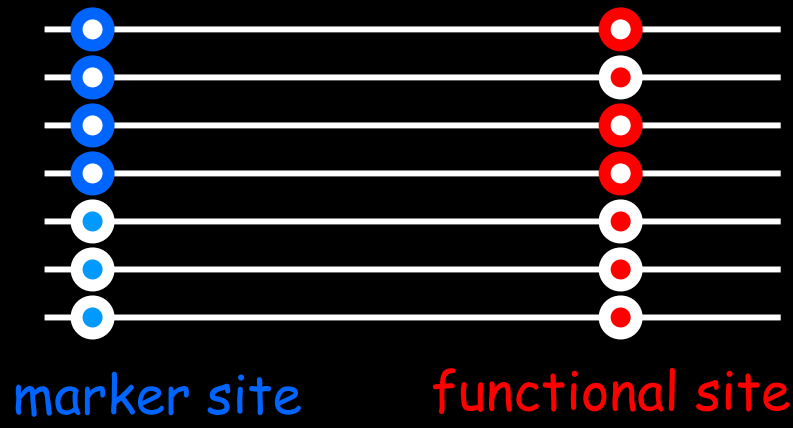


- problem: genotyping cost precludes using millions of markers simultaneously for an association study
- question: how to select from all available markers a subset that captures most mapping information (marker selection, marker prioritization)
- depends on the patterns of allelic association in the human genome



Allelic association

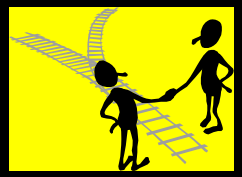
- allelic association is the non-random assortment between alleles i.e. it measures how well knowledge of the allele state at one site permits prediction at another



- significant allelic association between a marker and a functional site permits localization (mapping) even without having the functional site in our collection

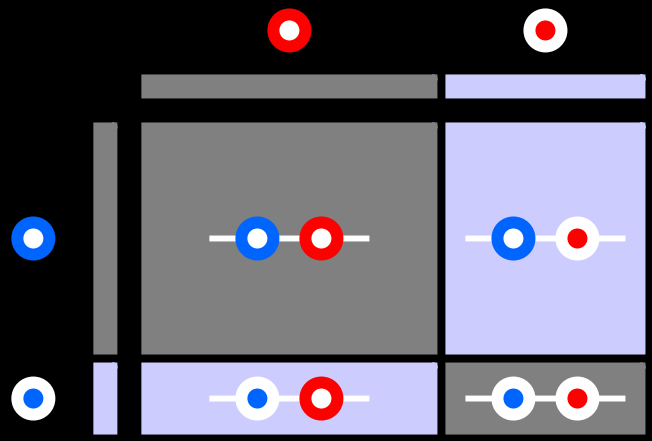
- by necessity, the strength of allelic association is measured between markers

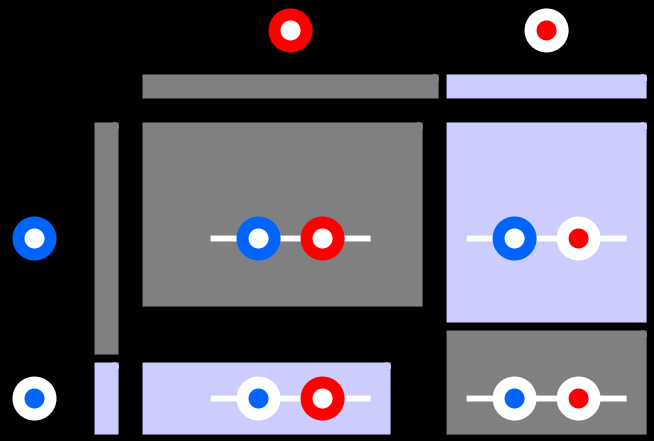
- there are pair-wise and multi-locus measures of association



Linkage disequilibrium

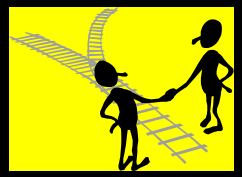
• LD measures the deviation from random assortment of the alleles at a pair of polymorphic sites





D=f(●●) – f(●) x f(●)

• other measures of LD are derived from D, by e.g. normalizing according to allele frequencies (r^2)

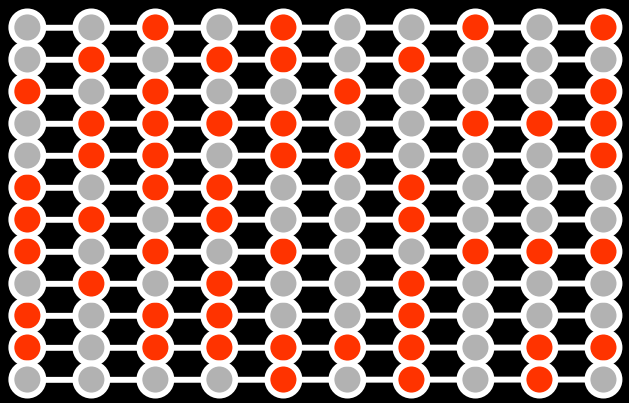


Haplotype diversity

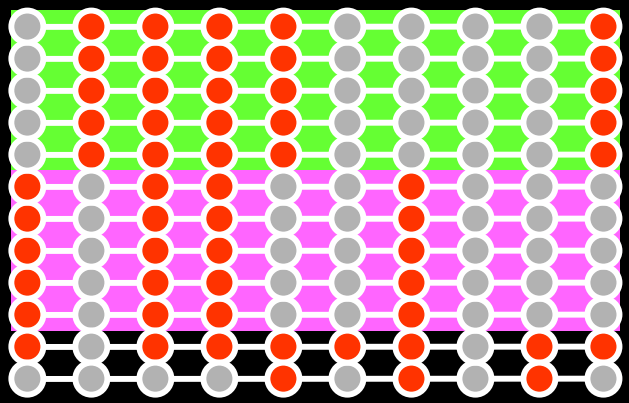
- the most useful multi-marker measures of associations are related to haplotype diversity

n markers

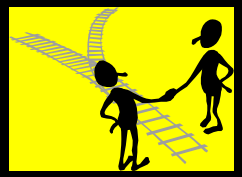
2^n possible haplotypes



random assortment of alleles at different sites

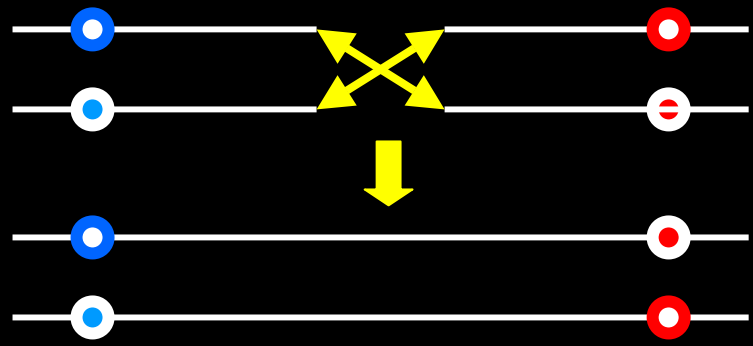


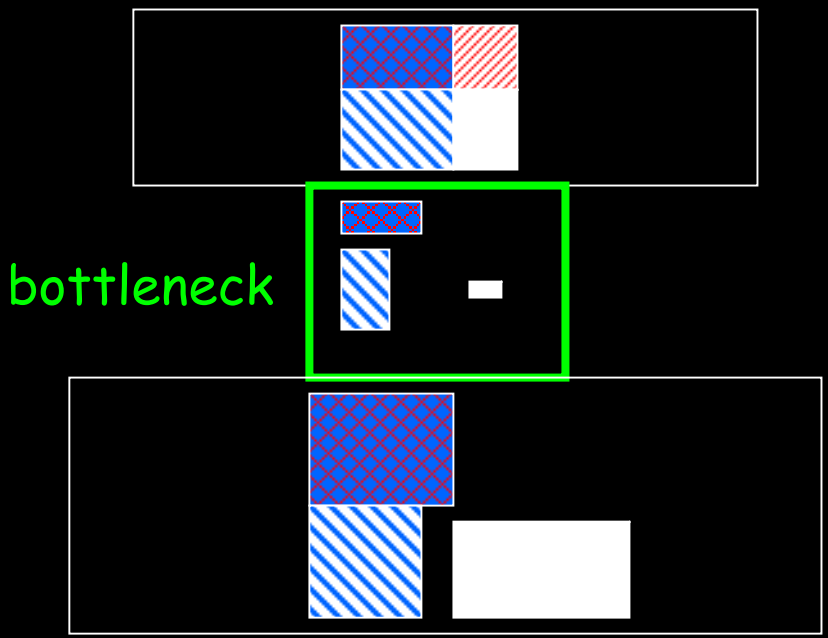
strong association: most chromosomes carry one of a few common haplotypes – reduced haplotype diversity



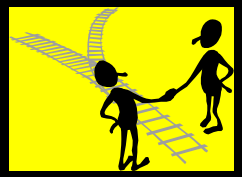
The determinants of allelic association

- recombination: breaks down allelic association by "randomizing" allele combinations

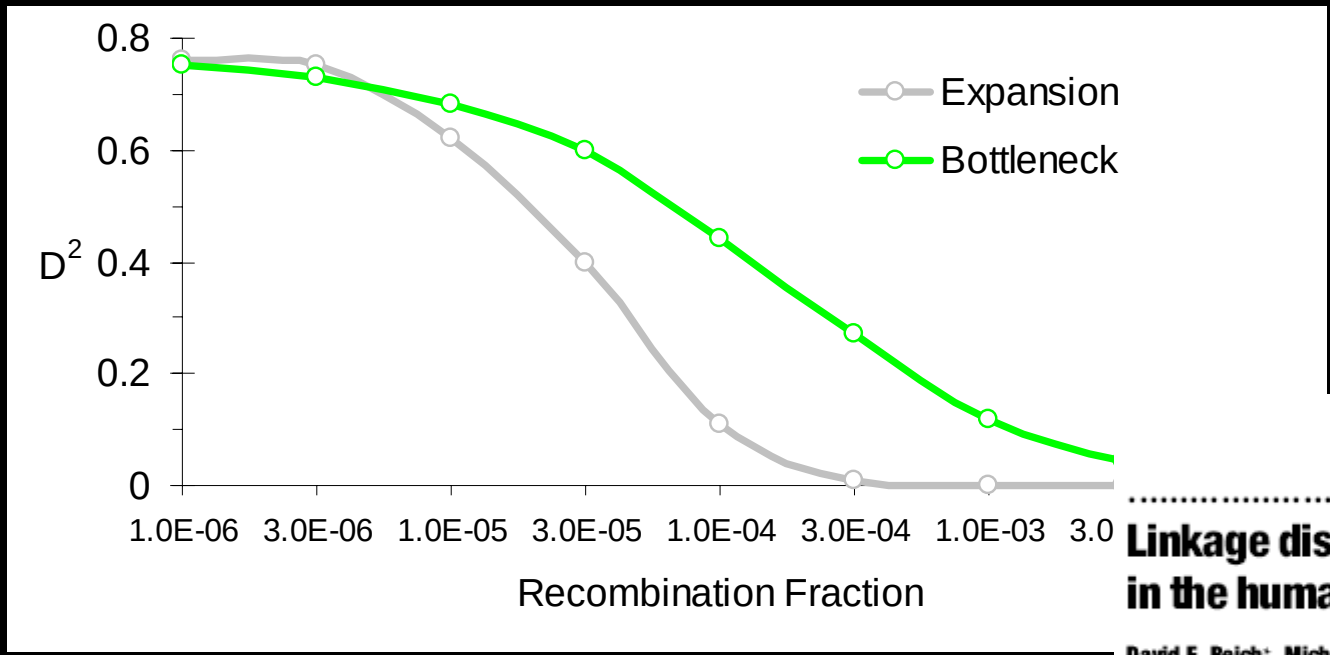




- demographic history of effective population size: bottlenecks increase allelic association by non-uniform re-sampling of allele combinations (haplotypes)



Strength of LD in the human genome

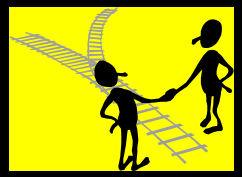


letters to nature

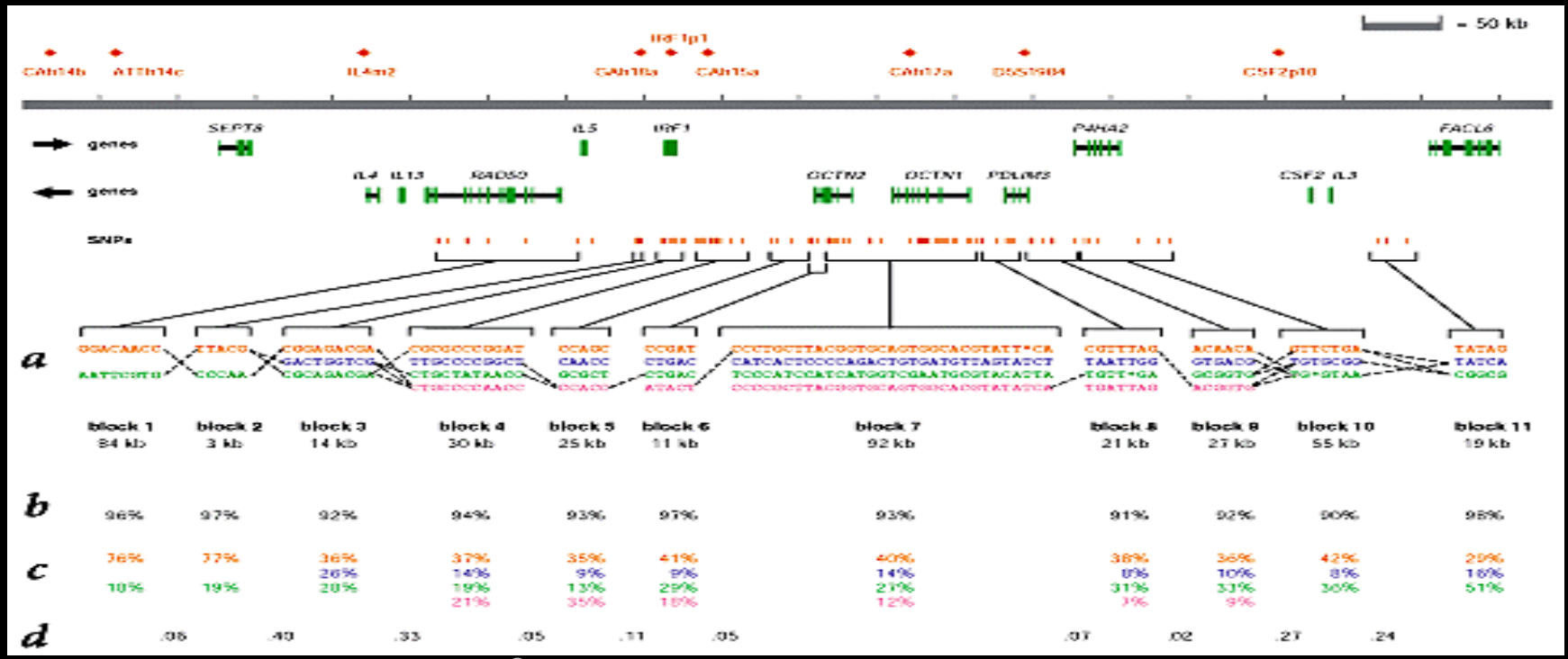
Linkage disequilibrium in the human genome

David E. Reich*, Michele Cargill*†, Stacey Bolk*, James Ireland*, Pardis C. Sabeti‡, Daniel J. Richter*, Thomas Lavery*, Rose Kouyoumjian*, Shelli F. Farhadian*, Ryk Ward‡ & Eric S. Lander*§

- LD is stronger, extends longer than previously thought

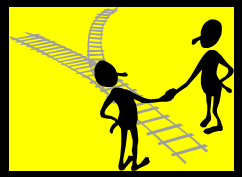


Haplotype blocks

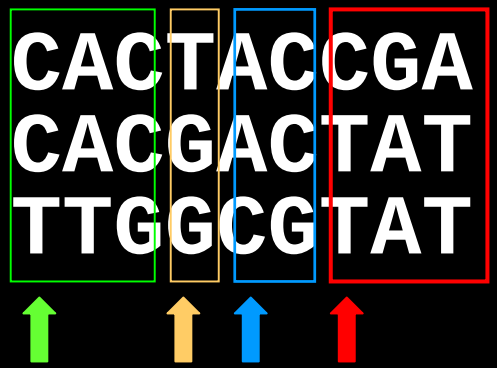


Daly *et al.*
Nature Genetics 2001

- experimental evidence for reduced haplotype diversity (mainly in European samples)



The promise for medical genetics



- within blocks a small number of SNPs are sufficient to distinguish the few common haplotypes → significant marker reduction is possible

- if the block structure is a general feature of human variation structure, whole-genome association studies will be possible at a reduced genotyping cost

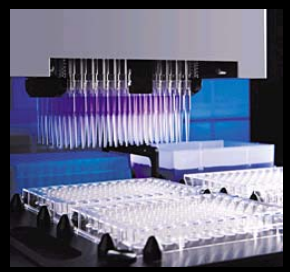
- this motivated the HapMap project

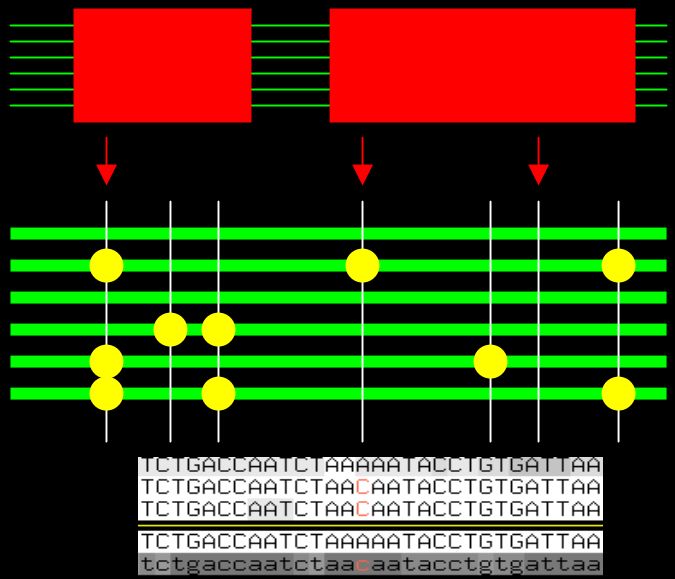
Gibbs *et al.*
Nature 2003

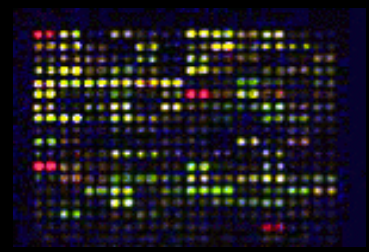
The HapMap initiative

- goal: to map out human allele and association structure of at the kilobase scale

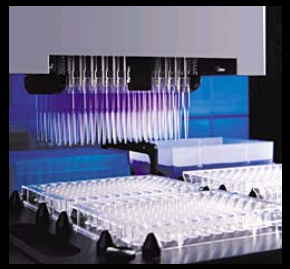
- deliverables: a set of physical and informational reagents



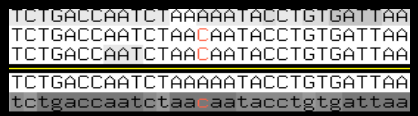




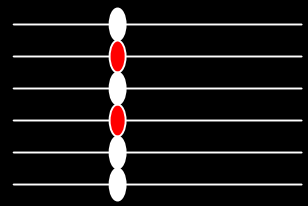
HapMap physical reagents

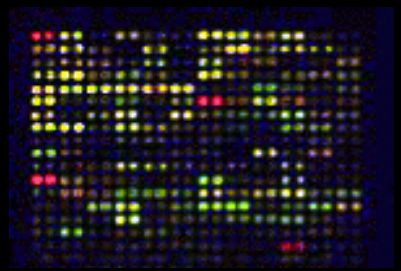


- reference samples: 4 world populations, ~100 independent chromosomes from each



- SNPs: computational candidates where both alleles were seen in multiple chromosomes

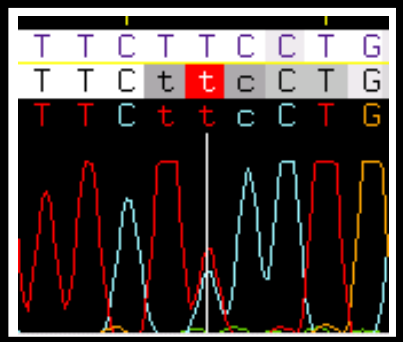


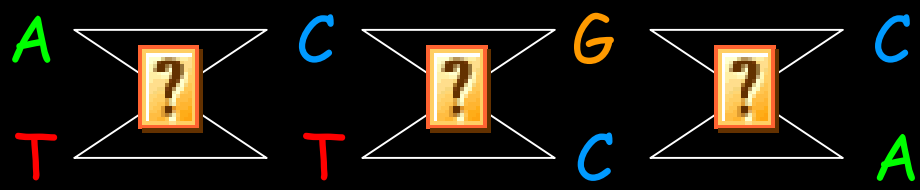


- genotypes: high-accuracy assays from various platforms; fast public data release

Informational reagents: haplotypes

- the problem: the substrate for genotyping is diploid, genomic DNA; phasing of alleles at multiple loci is in general not possible with certainty





- experimental methods of haplotype determination (single-chromosome isolation followed by whole-genome PCR amplification, radiation hybrids, somatic cell hybrids) are expensive and laborious

Computational haplotype inference

- Parsimony approach: minimize the number of different haplotypes that explains all diploid genotypes in the sample

Clark
Mol Biol Evol 1990

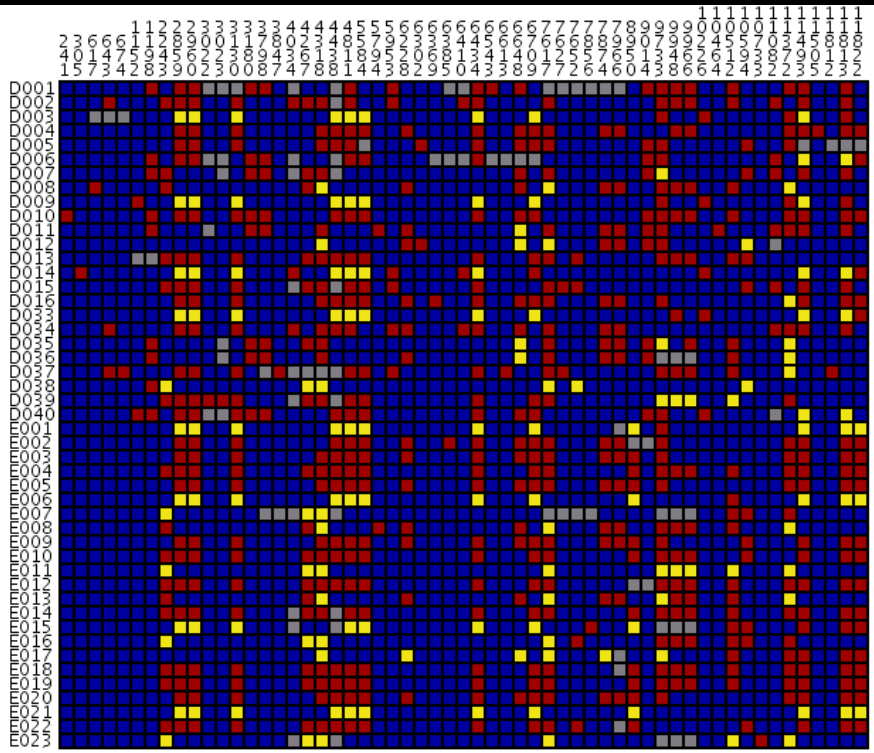
- Maximum likelihood approach: estimate haplotype frequencies that are most likely to produce observed diploid genotypes

Excoffier & Slatkin
Mol Biol Evol 1995

- Bayesian methods: estimate haplotypes based on the observed diploid genotypes and the a priori expectation of haplotype patterns informed by Population Genetics

Stephens *et al.*
AJHG 2001

Haplotype inference

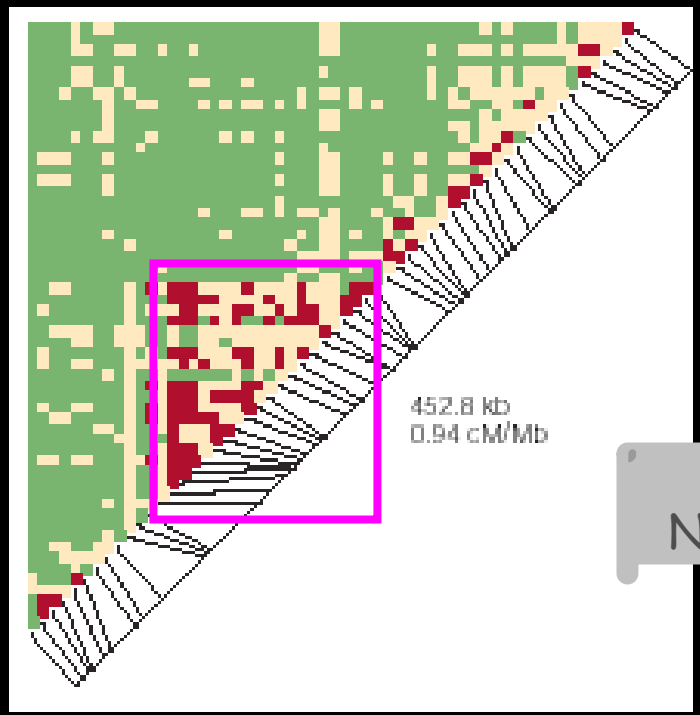


VG2 of: ikbbe.prettybase.txt

Homozygote-Common Allele
Homozygote-Rare Allele
Heterozygote
Undetermined
Rare Allele Percentage: 0.0
Clustered By: NOTHING
Population: null

http://pga.gs.washington.edu/

Haplotype annotations – LD based



Wall & Pritchard
Nature Rev Gen 2003

- Pair-wise LD-plots

- LD-based multi-marker block definitions requiring strong pair-wise LD between all pairs in block

Annotations – haplotype blocks

- Dynamic programming approach

Zhang *et al.*
AJHG 2001

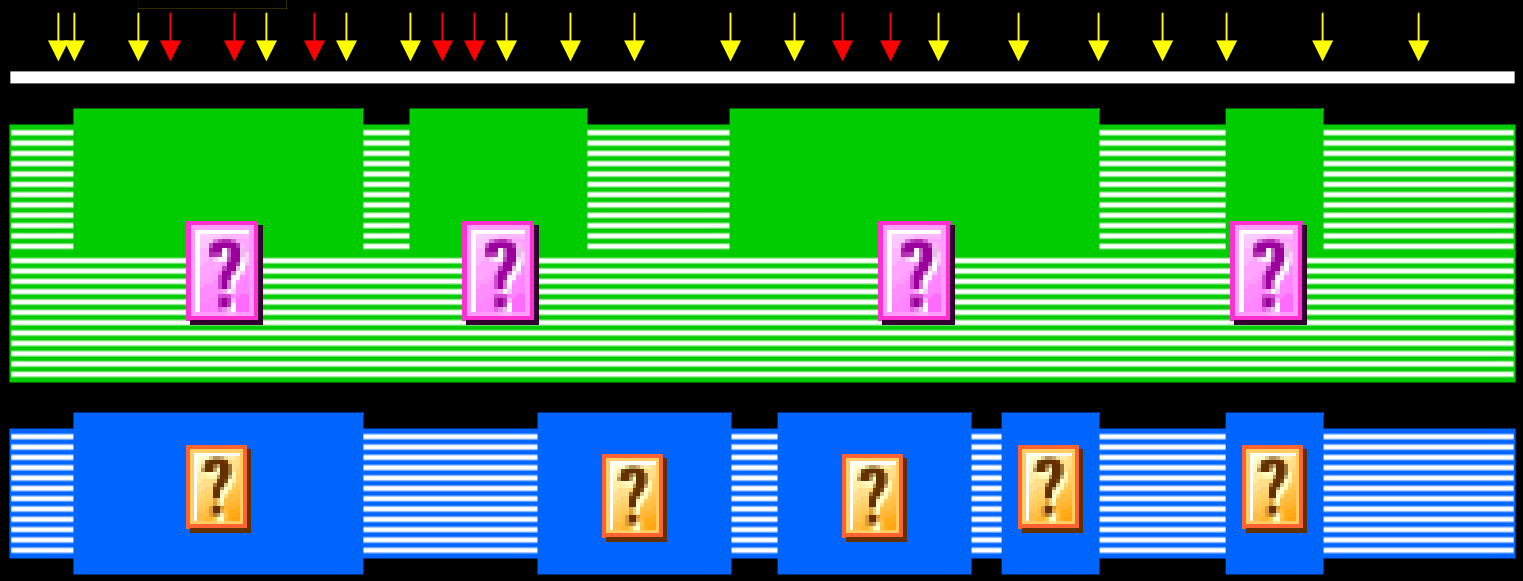
```
B-1     1-1-1-1-1-1-1-1-1|1-0-1-0-0-0-0-1|1-1-1-1
B-2     1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-1-1-0
B-3     1-1-1-1-1-1-1-1-1|1-0-1-0-0-0-0-1|1-1-1-1
B-4     1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-1-1-0
B-5     0-0-0-0-0-0-0-0-0|0-0-0-1-0-1-1-0|0-0-0-0
B-6     1-1-1-1-1-1-1-1-1|1-0-1-0-0-0-0-1|1-1-1-1
B-7     1-1-1-0-1-1-1-0-1|0-0-0-1-0-1-1-0|0-0-0-0
B-8     1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-1-1-0
B-9     1-1-1-0-0-0-0-0-0|0-1-0-0-1-0-0-0|1-0-1-0
B-10    0-0-0-0-0-0-0-0-0|0-0-0-1-0-1-1-0|0-0-0-0
B-11    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-0-1-0
B-12    0-0-0-0-0-0-0-0-0|0-1-0-0-1-0-0-0|0-0-0-0
B-13    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|0-0-1-0
B-14    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-1-1-0
B-15    1-1-1-0-0-0-0-0-0|0-1-0-0-1-0-0-0|1-0-1-0
B-16    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-0-1-0
B-17    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-1-1-0
B-18    1-1-1-0-1-1-1-0-1|1-0-1-0-0-0-0-1|1-1-1-1
B-19    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-1-1-0
B-20    1-1-1-1-1-1-1-1-1|1-0-1-0-0-0-0-1|1-1-1-1
B-21    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-0-1-0
B-22    1-1-1-1-1-1-1-1-1|1-0-1-0-0-0-0-1|1-1-1-1
B-23    1-1-1-0-1-1-1-0-1|0-0-1-0-0-0-0-0|1-1-1-0
```

3 3 3

1. meet block definition based on common haplotype requirements
2. within each block, determine the number of SNPs that distinguishes common haplotypes (htSNPs)
3. minimize the total number of htSNPs over complete region including all blocks

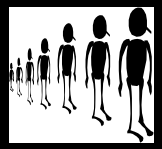
```
block1: left=1 right=9 length=9 htSnps=3 haplotypes=[000000000 (17), 111011101 (13),
 111111111 (10), 000000001 (3), 111000000 (2), 000011111 (1)]
block2: left=10 right=17 length=8 htSnps=3 haplotypes=[10100001 (12), 00100000 (11),
 01001000 (10), 00010110 (10), 00101000 (3)]
block3: left=18 right=21 length=4 htSnps=3 haplotypes=[0000 (15), 1111 (12), 1010 (1
2), 1110 (7)]
blocks=3 htSnps=9
extractHapBlocks.pl completed
```

Questions about the HapMap



- is structure constant with sample size?
- completion, sufficient density?
- haplotype structure across populations?

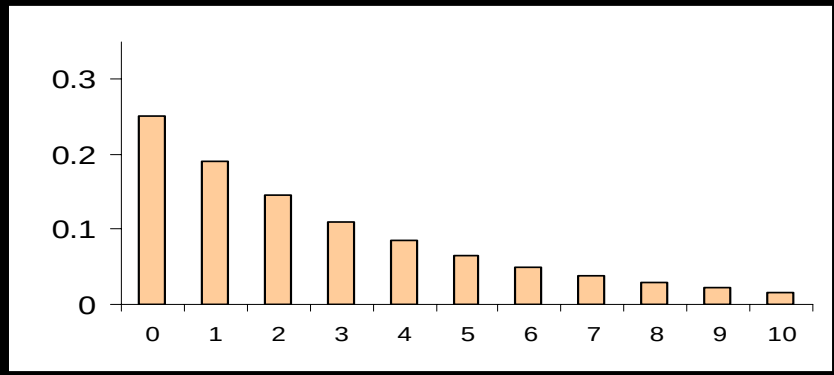
- Explore human allele structure with a Population Genetic modeling and data fitting technique

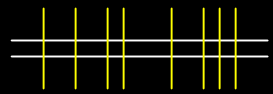


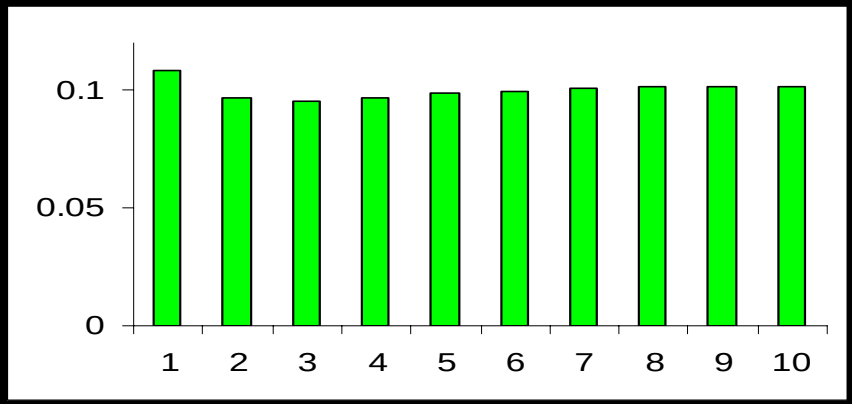
Data: polymorphism distributions

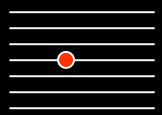
1. marker density (MD): distribution of number of SNPs in pairs of sequences

Clone 1	Clone 2	# SNPs
AL00675	AL00982	8
AS81034	AK43001	0
CB00341	AL43234	2

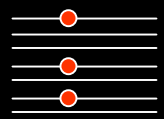








"rare"



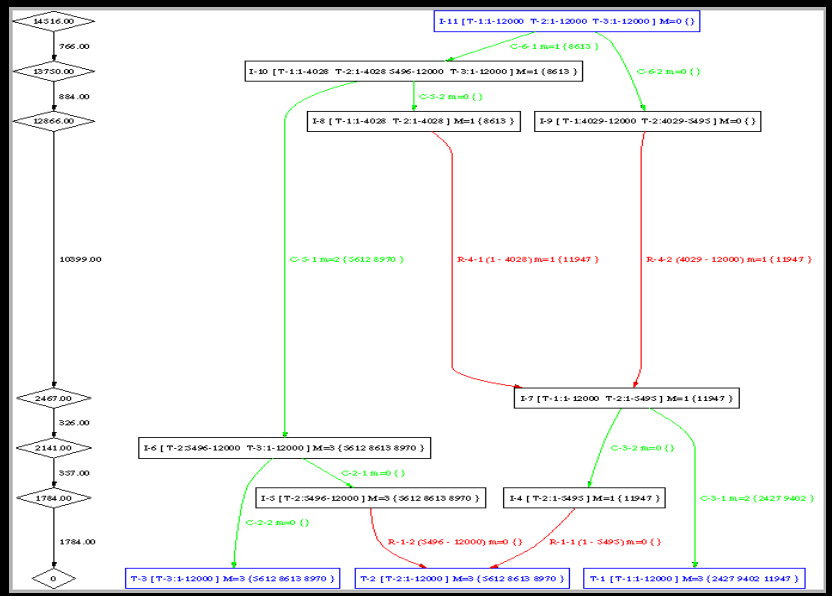
"common"

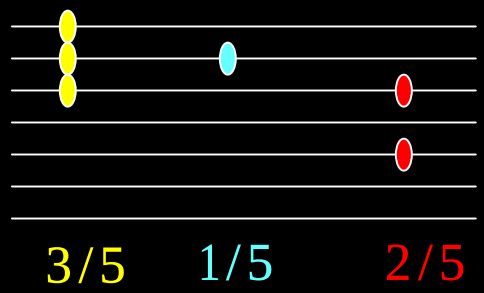
2. allele frequency spectrum (AFS): distribution of SNPs according to allele frequency in a set of samples

SNP	Minor allele	Allele count
A/G	A	1
C/T	T	9
A/G	G	3

Model: processes that generate SNPs

simulation procedures



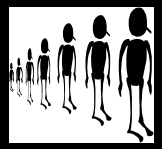


computable formulations

$$P(k)=\frac{1}{1+\theta_1 L}\left(\frac{\theta_1 L}{1+\theta_1 L}\right)^k\left\{1-e^{-(1+\theta_1 L)\Delta\tau_1}\left[1+\sum_{i=1}^{k}\frac{[(1+\theta_1 L)\Delta\tau_1]^i}{i!}\right]\right\}$$

$$+\frac{1}{1+\theta_2 L}\left(\frac{\theta_2 L}{1+\theta_2 L}\right)^k e^{-(1+\theta_1 L)\Delta\tau_1}\left\{\left[1+\sum_{i=1}^{k}\frac{\left[(1+\theta_2 L)\frac{\theta_1}{\theta_2}\Delta\tau_1\right]^i}{i!}\right]-e^{-(1+\theta_2 L)\Delta\tau_2}\left[1+\sum_{i=1}^{k}\frac{\left[(1+\theta_2 L)\left(\frac{\theta_1}{\theta_2}\Delta\tau_1+\Delta\tau_2\right)\right]^i}{i!}\right]\right\}$$

$$+\frac{1}{1+\theta_3 L}\left(\frac{\theta_3 L}{1+\theta_3 L}\right)^k e^{-(1+\theta_1 L)\Delta\tau_1-(1+\theta_2 L)\Delta\tau_2}\left\{1+\sum_{i=1}^{k}\frac{\left[(1+\theta_2 L)\left(\frac{\theta_1}{\theta_3}\Delta\tau_1+\frac{\theta_2}{\theta_3}\Delta\tau_2\right)\right]^i}{i!}\right\}$$



Models of demographic history

past

history

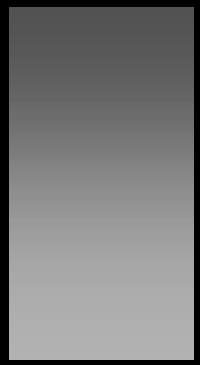
present

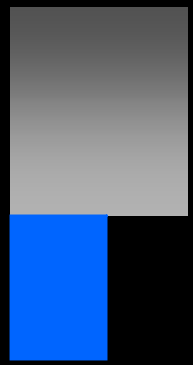
stationary

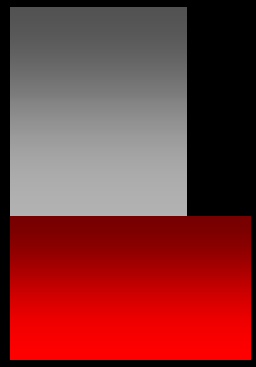
collapse

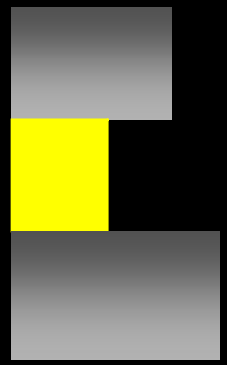
expansion

bottleneck

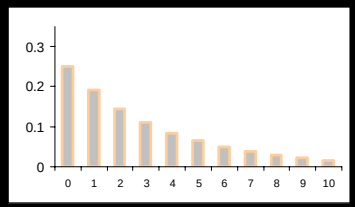


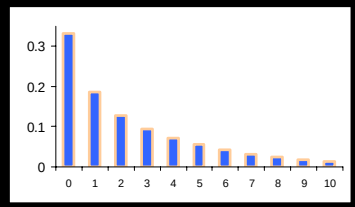


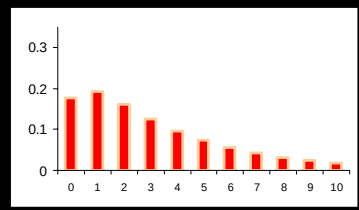


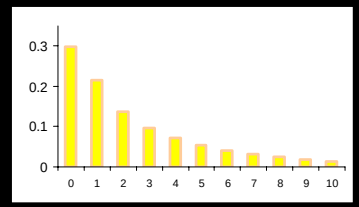


MD
(simulation)

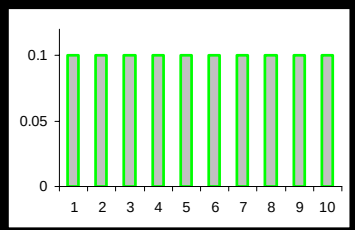


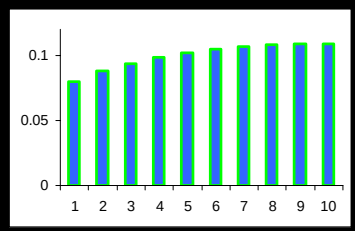


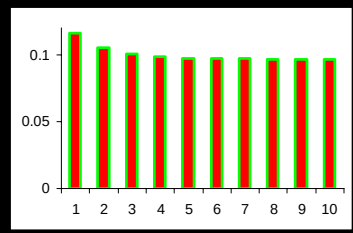


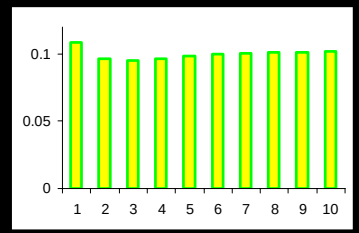


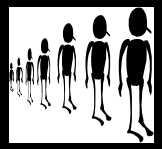
AFS
(direct form)



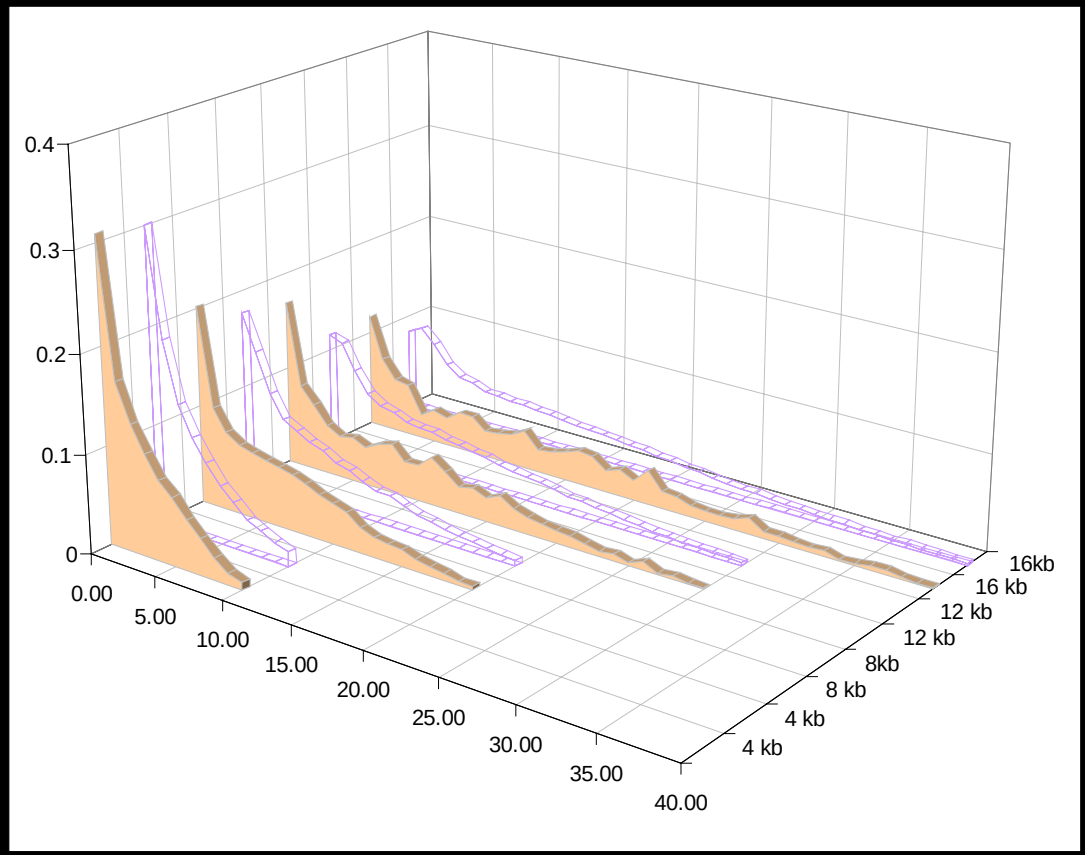




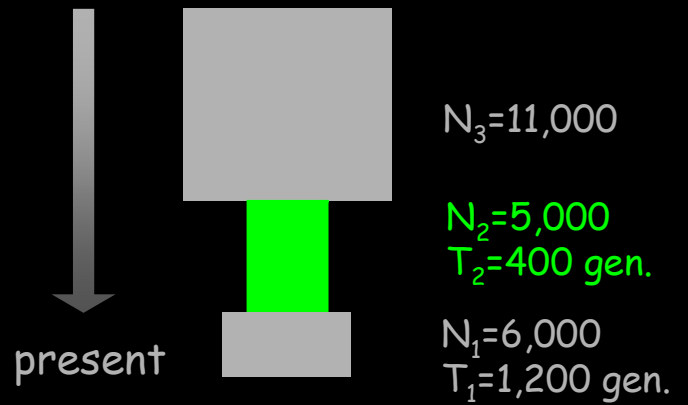




Data fitting: marker density



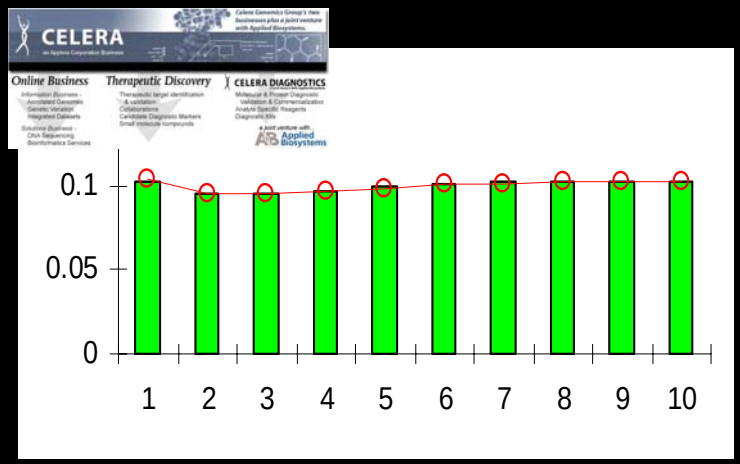
- best model is a bottleneck shaped population size history

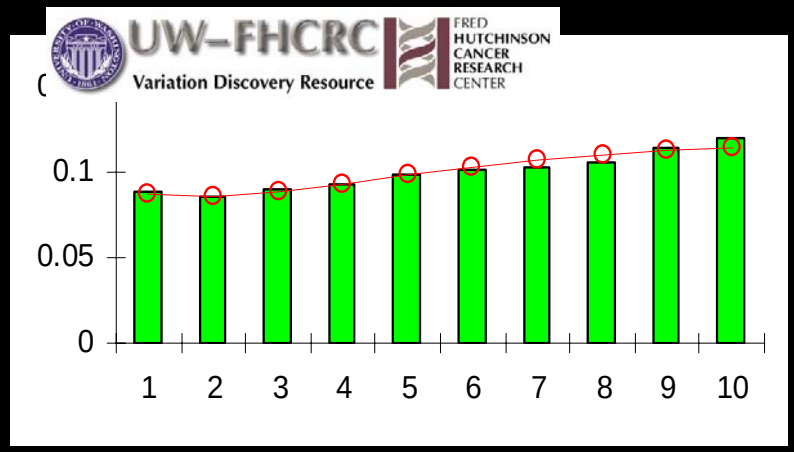


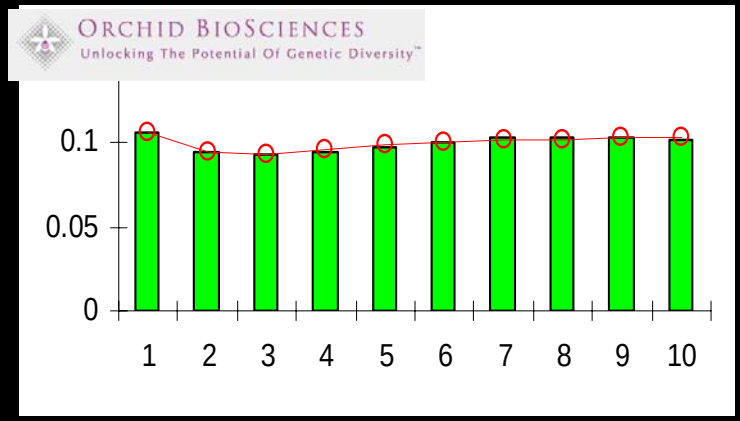
Marth *et al.*
PNAS 2003

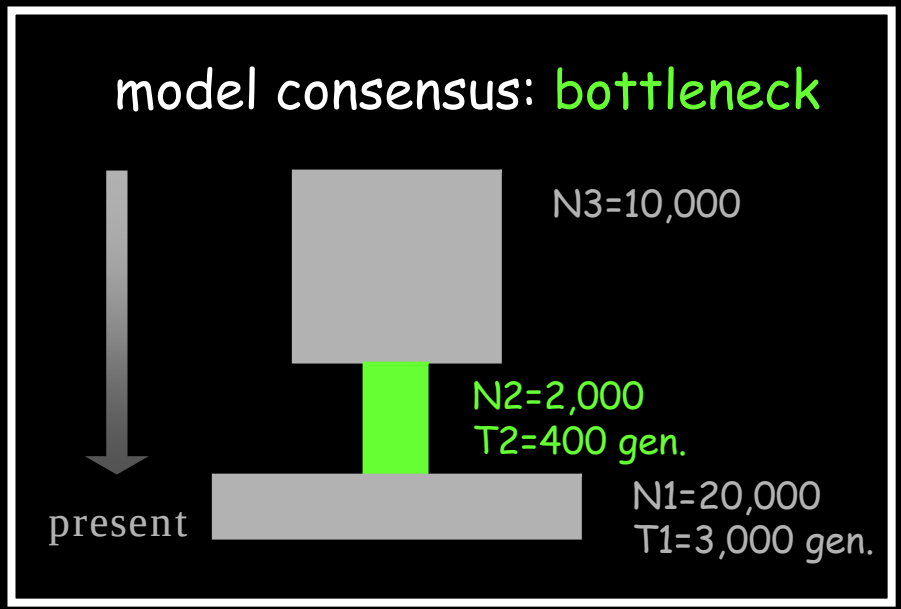
- our conclusions from the marker density data are confounded by the unknown ethnicity of the public genome sequence we looked at allele frequency data from ethnically defined samples

Data fitting: allele frequency



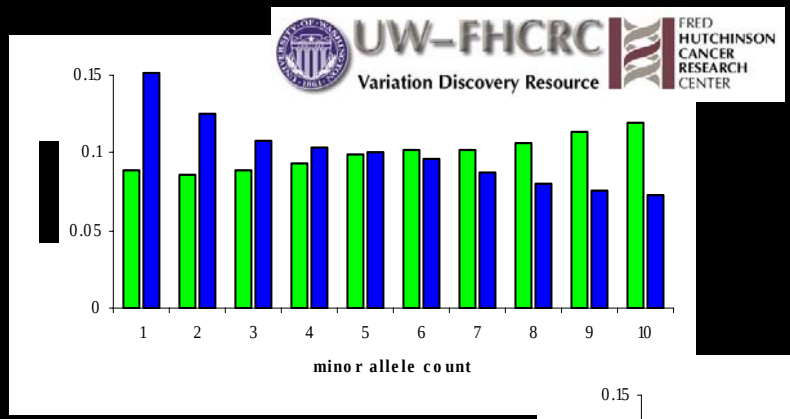






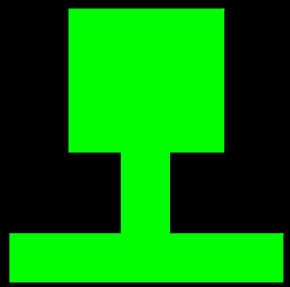
- Data from other populations?

Population specific demographic history

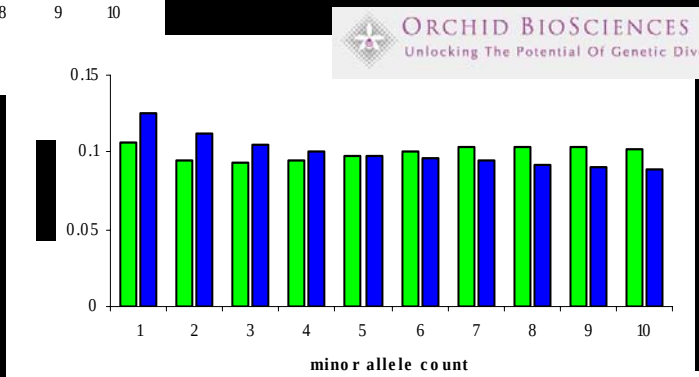


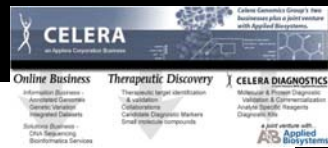
European data

African data

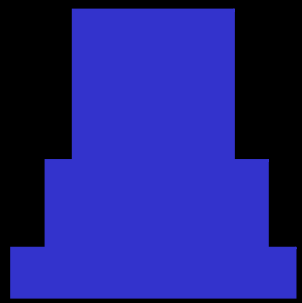


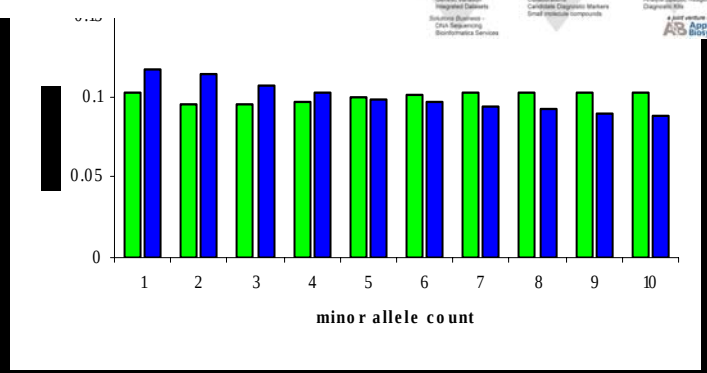
bottleneck



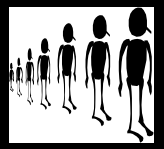


modest but uninterrupted expansion

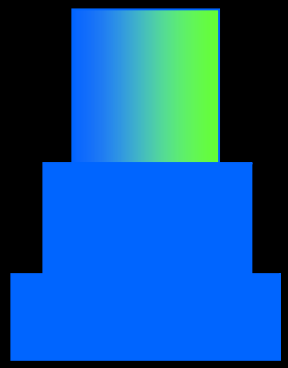


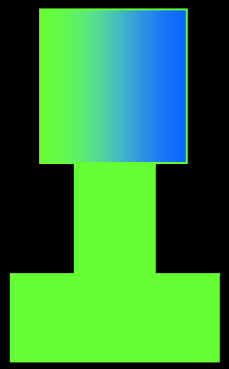


Marth *et al.*
Genetics 2004

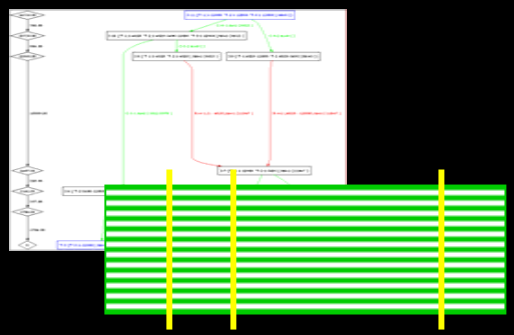


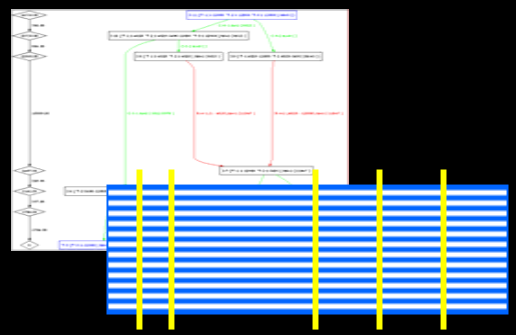
Model-based prediction





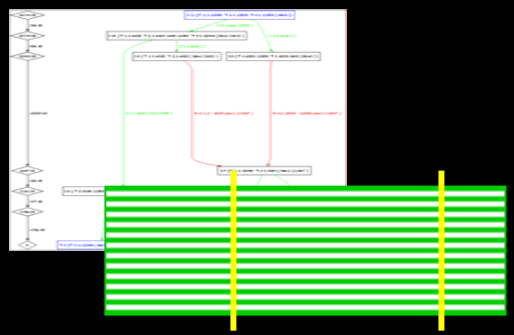
computational model encapsulating what we know about the process

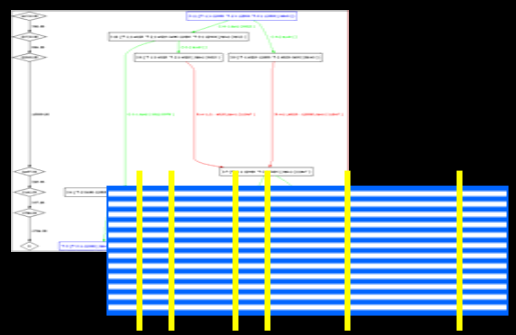




genealogy + mutations

allele structure





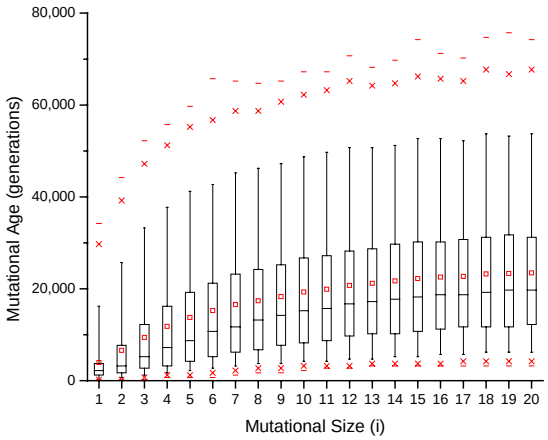
arbitrary number of additional replicates

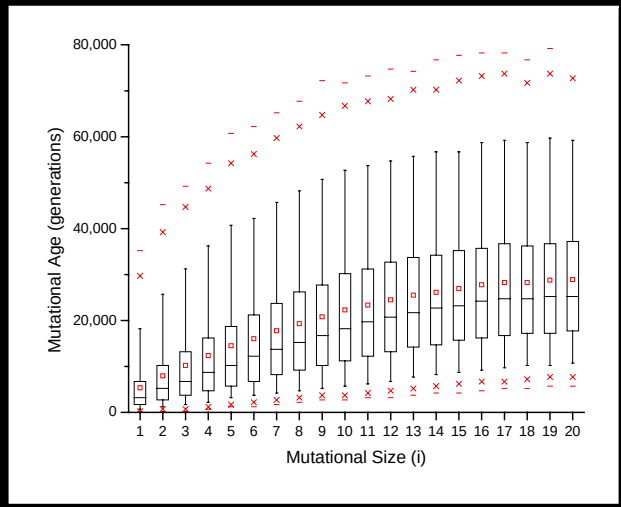
Prediction – allele frequency and age

European data

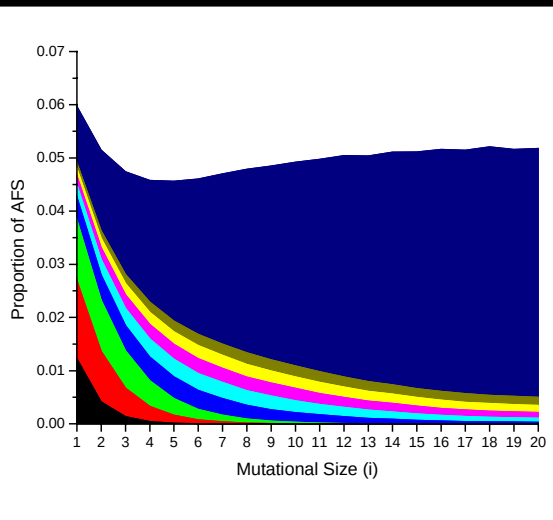
African data

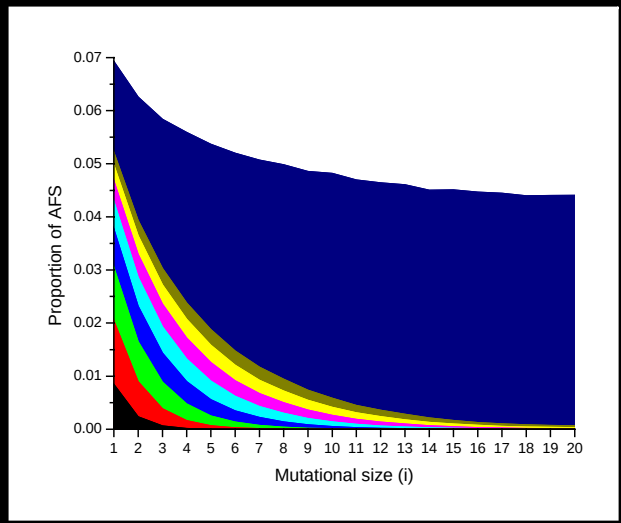
average age of polymorphism



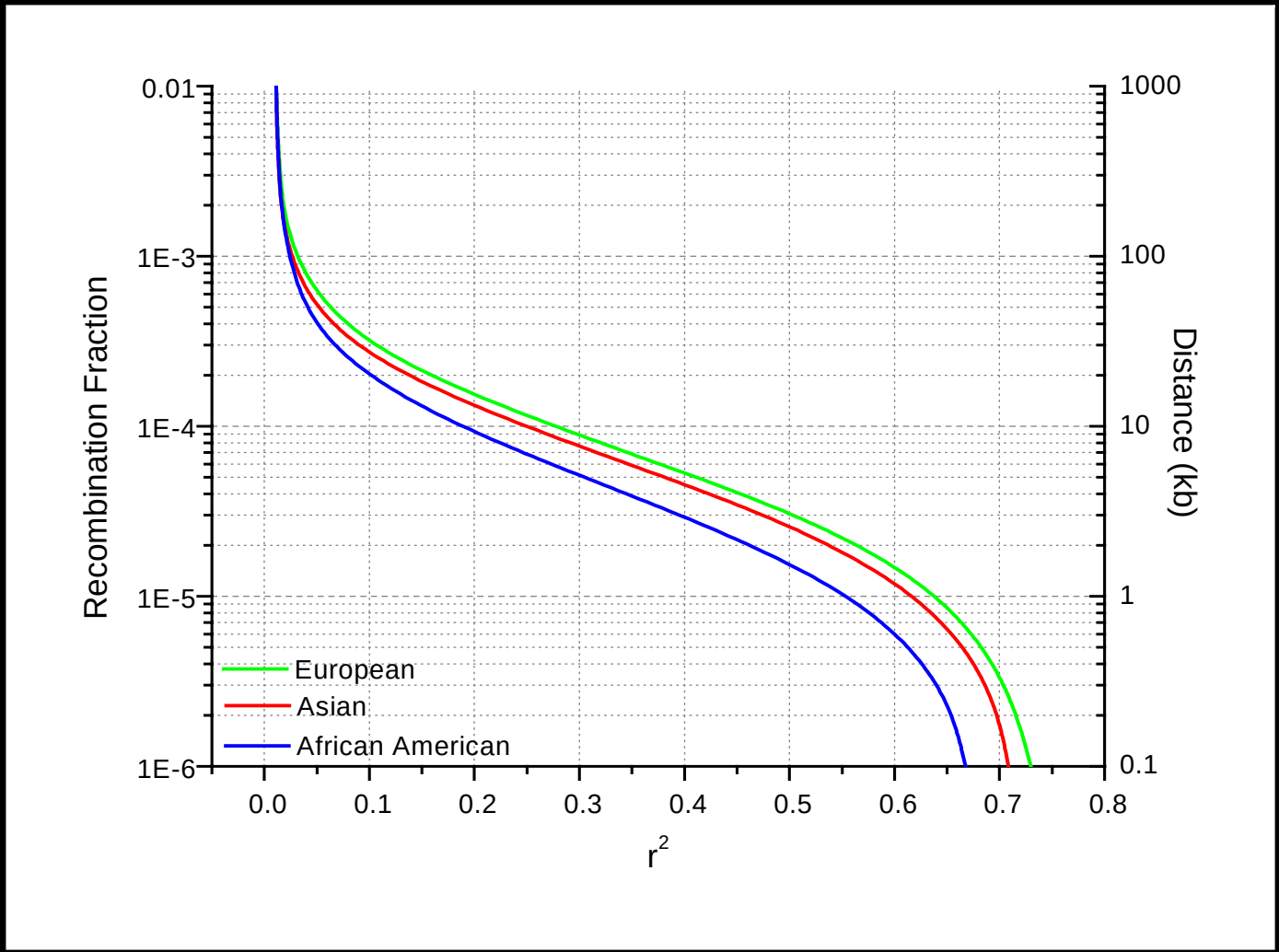


contribution of the past to alleles in various frequency classes

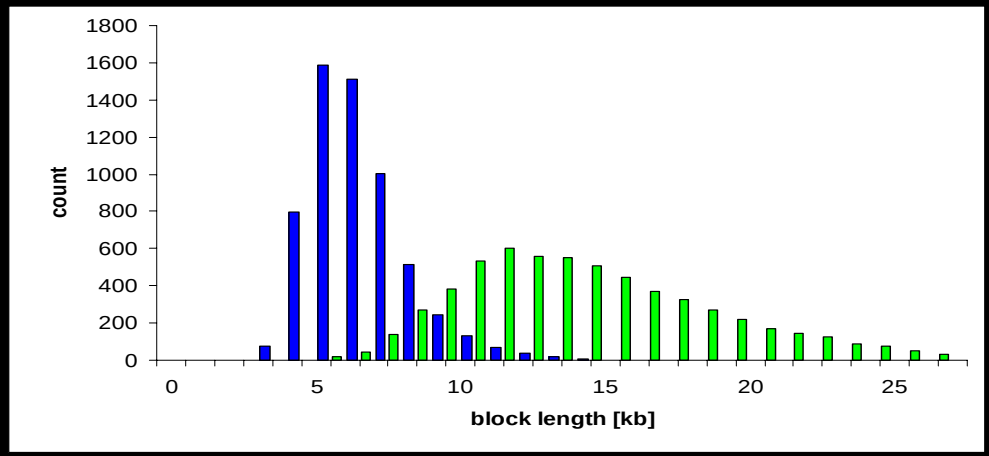




Prediction – extent of LD

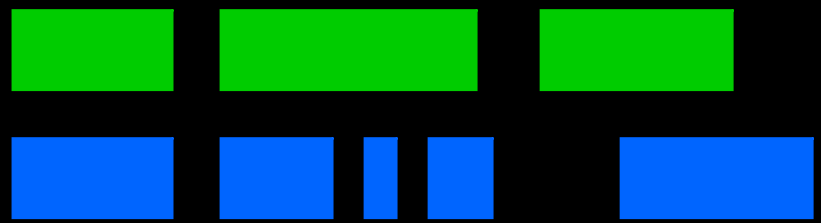


Prediction – haplotype structure

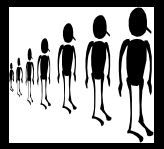


- our models predict shorter blocks in African samples than in Europeans

- what is the spatial relationship between blocks?

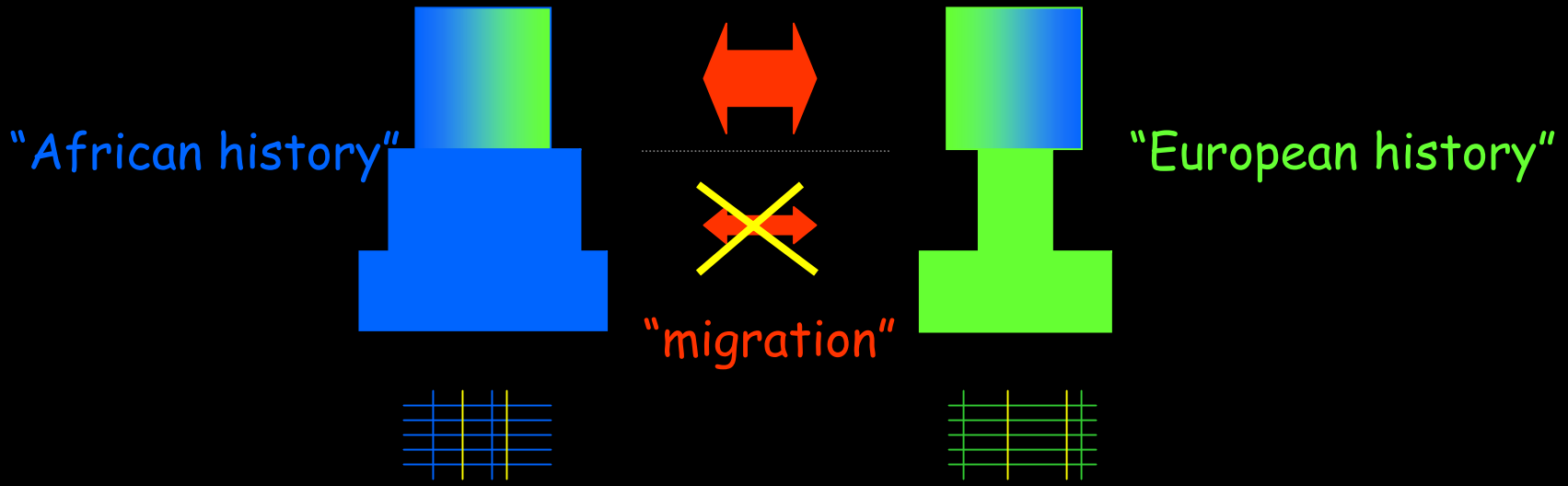


- we must connect the polymorphism structure of different human populations

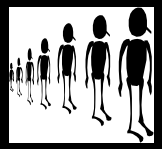


Modeling joint allele structure

- The "true" history of all human populations is interconnected
- We study these relationships with models of population subdivision



- The genealogy of samples from different populations are connected through the shared part of our past
- Polymorphic markers (some shared, some population-specific) and haplotypes are placed into a common frame of reference



Joint allele frequencies

European / African	monomorphic	rare	common
monomorphic	0.0 % 0.0 %	19.9 % 13.2 %	2.3 % 1.0 %
rare	43.4 % 43.7 %	11.5 % 11.0 %	4.6 % 7.4 %
common	10.2 % 4.2 %	4.4 % 6.0 %	6.6 % 13.4 %

observation in UW PGA data

SNPs private to European samples

shared SNPs

SNPs common in both populations

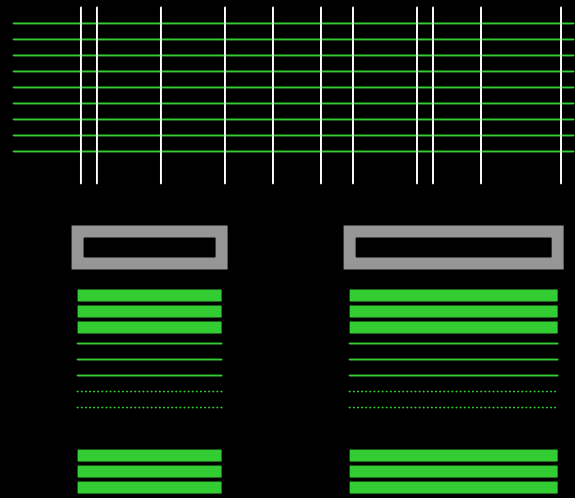
SNPs private to African samples

- our simple model of subdivision captures the qualitative dynamics
- we now have the tools to analyze joint allele structure

Generality for future samples?

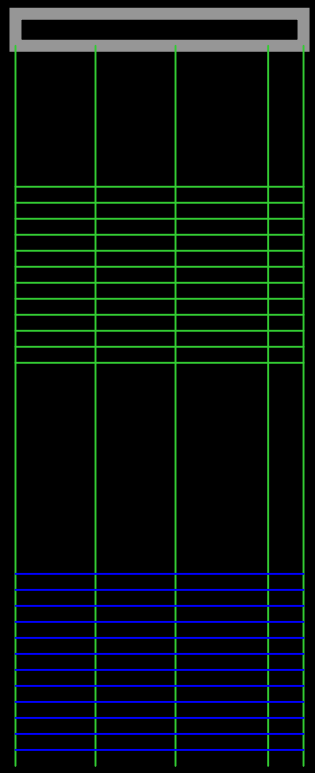
- The haplotype map resource is a collection of reagents

1. reference samples
2. common markers
3. blocks
4. list of haplotypes
5. frequent haplotypes



- How relevant are the reference reagents to future clinical samples (drawn from the same or different population)?

Reference haplotypes



reference haplotypes

same population

99.4%

different population

87.5%

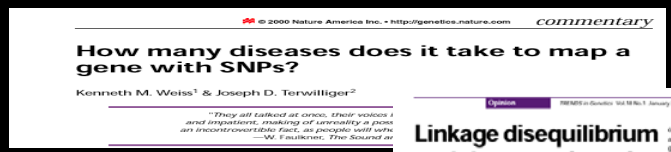
(74.9%, 65.0% at lower minimum marker allele frequency)

- these computational studies inform us about the global, genome-average properties of the HapMap reagents

- what can we say about linkage in specific local regions?

Utility for association studies?

- No matter how good the resource is, its success to find disease causing variants greatly depend on the allelic structure of common diseases, a question under debate


commentary

How many diseases does it take to map a gene with SNPs?

Kenneth M. Weiss[1] & Joseph D. Terwilliger[2]


Linkage disequilibrium and the mapping of complex human traits

Kenneth M. Weiss and Andrew G. Clark

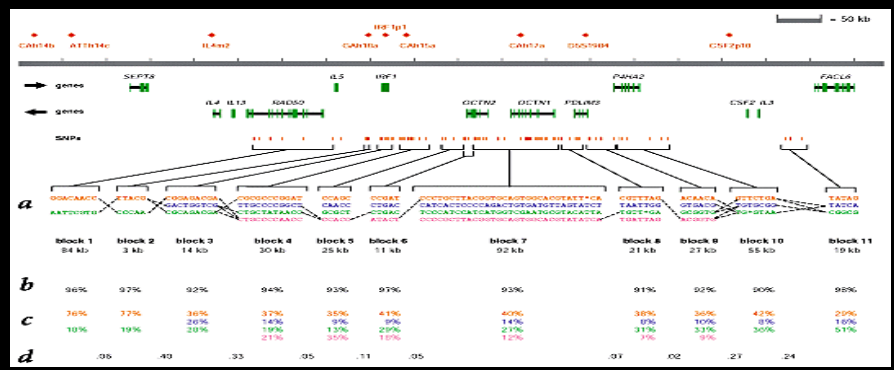

Are Rare Variants Responsible for Susceptibility to Complex Diseases?

Jonathan K. Pritchard


On the allelic spectrum of human disease

David E. Reich and Eric S. Lander

- Regardless of how we describe human association structure, many questions remain about the relative merits of single-marker vs. haplotype-based strategies for medical association studies



Acknowledgements

Steve Sherry
Eva Czabarka
Janos Murvai
Alexey Vinokurov
Greg Schuler
Richa Agarwala
Stephen Altschul

Eric Tsung

Aravinda Chakravarti (Hopkins)
Andy Clark (Cornell)
Pui-Yan Kwok (UCSF)
Henry Harpending (Utah)
Jim Weber (Marshfield)

marth@bc.edu

http://clavius.bc.edu/~marthlab/MarthLab