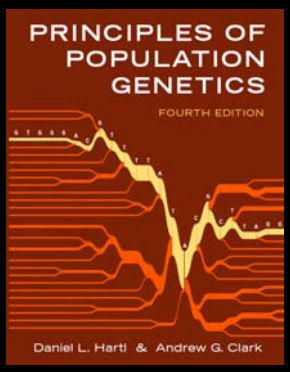


Due to appear January 2007 !

The scope of Population Genetics

- Why are the patterns of variation as they are? (mathematical theory)
- What are the forces that influence levels of variation?
- What is the genetic basis for evolutionary change?
- What data can be collected to test hypotheses about the factors that impact allele frequency?
- What is the relation between genotypic variation and phenotype variation?

Forces acting on allele frequencies in populations

- Mutation
- Random genetic drift
- Recombination/gene conversion
- Migration/Demography
- Natural selection

Genotype and Allele frequencies

Genotype frequency: proportion of each genotype in the population

Genotype	Number	Frequency
B/B	114	114/200 = 0.57
B/b	56	56/200 = 0.28
b/b	30	30/200 = 0.15
Total	200	1.00

Frequency of an allele in the population is equivalent to the probability of sampling that allele in the population.

Let p = freq (*B*) and q = freq (*b*)

$p + q = 1$

p = freq (*B*) = freq (*BB*) + ½ freq (*Bb*)
q = freq (*b*) = freq (*bb*) + ½ freq (*Bb*)

Genotype	Number
B/B	114
B/b	56
b/b	30
Total	200

p = freq (*B*) = freq (*BB*) + ½ freq (*Bb*) = 0.57+0.28/2 =0.71
q = freq (*b*) = freq (*bb*) + ½ freq (*Bb*) = 0.15 + 0.28/2 = 0.29

Gene Counting

p = count of *B* alleles/total = (114 x 2 + 56)/400 = 0.71
q = count of *b* alleles/total = (30 x 2 + 56)/400 = 0.29

Hardy-Weinberg Principle

For two alleles of an autosomal gene, *B* and *b*, the genotype frequencies after one generation

freq(*B*) = *p* freq(*b*) = *q*

freq (*B*/*B*) = p^2
freq (*B*/*b*) = $2pq$
freq (*b*/*b*) = q^2

Gene frequencies of offspring can be predicted from allele frequencies in parental generation

Assumptions of Hardy Weinberg

- Approximately random mating
- An infinitely large population
- No mutation
- No migration into or out of the population
- No selection, with all genotypes equally viable and equally fertile

Graphical proof of Hardy Weinberg Principle

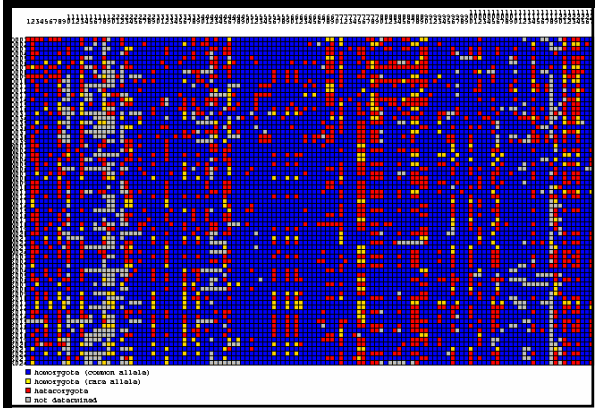
Eggs \ Sperm	*B*	*b*
B	p^2	pq
b	pq	q^2

Freq of alleles in offspring

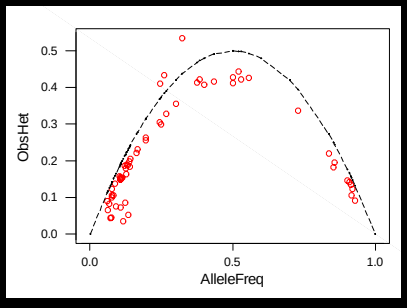
freq (*B*) = $p^2 + ½ (2pq)$ = $p\ (p+q)$ = $p\ (1) = p$

freq (*b*) = $q^2 + ½ (2pq)$ = $q\ (p + q\ )$ = $q(1) = q$

SNPs in the ApoAI/CIII/AIV/AV region of chromosome 11



Hardy-Weinberg tests for Quality Control



Heterozygotes are being under-called (Boerwinkle et al.)

Example from MN blood typing

	MM	M/N	N/N	Total
Num. Individuals	1787	3037	1305	6129
Number M alleles	3574	3037	0	6611
Number N alleles	0	3037	2610	5647
Number M+N	3574	6074	2610	12258

Allele freq of M = 6611/12,258 = 0.53932 = p
Allele freq of N = 5647/12,258 = 0.46068 = q

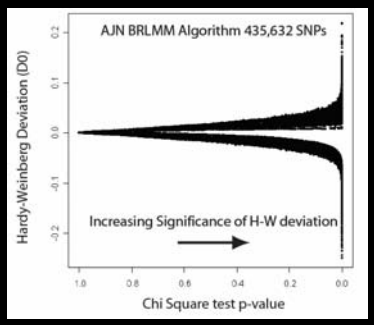
Expected freq	p^2 = 0.29087	2pq=0.49691	q^2 = 0.21222	1.00
Expected # (freq x 6129)	1782.7	3045.6	1300.7	6129

$$\chi^2 = \frac{\sum(\text{observed number} - \text{expected number})^2}{\text{expected number}}$$

$$\chi^2 = \frac{(1787 - 1782.7)^2}{1782.7} + \frac{(3037 - 3045.6)^2}{3045.6} + \frac{(1305 - 1300.7)^2}{1300.7} = 0.04887$$

Df = number of classes of data (3) – number of parameters estimated (1) –1 = 1 df
Probability of a chi-square this big or bigger = .90

Hardy-Weinberg tests on steroids – the Affy 500k chip



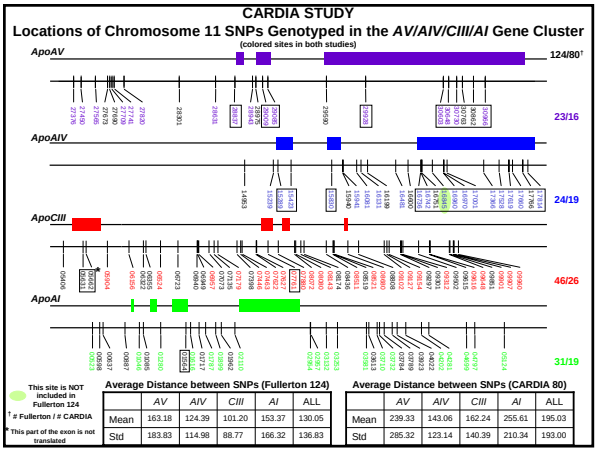
HW deviation = observed – expected heterozygosity

Extensions of the Hardy-Weinberg Principle

- More than two alleles
- More than one locus
- X-chromosome
- Subdivided population

Mutation

- What is the pattern of nucleotide changes?
- Is the pattern of mutations homogeneous across the genome?
- Are sites within a gene undergoing recurrent mutation?



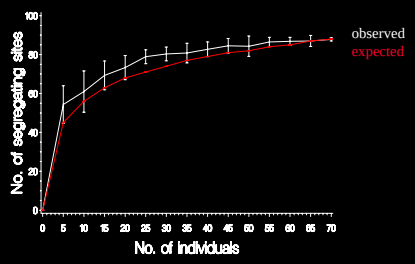
	AV	AIV	CIII	AI	ALL
Mean	163.18	124.39	101.20	153.37	130.05
Std	183.83	114.98	88.77	166.32	136.83

	AV	AIV	CIII	AI	ALL
Mean	239.33	143.06	162.24	255.61	195.03
Std	285.32	123.14	140.39	210.34	193.00

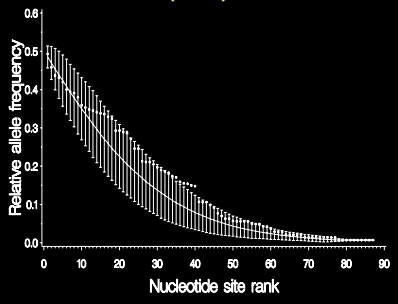
Mutation and Random Genetic Drift

- The primary parameter for drift is N_e.
- Mutation adds variation to the population, and drift eliminates it.
- These two processes come to a steady state in which the standing level of variation is essentially constant.

Observed and expected numbers of segregating sites (Lipoprotein lipase, LPL)



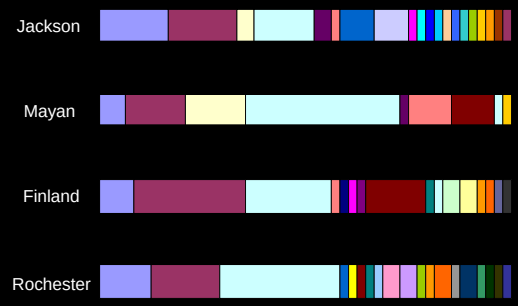
Nucleotide site frequency spectrum (*LPL*)

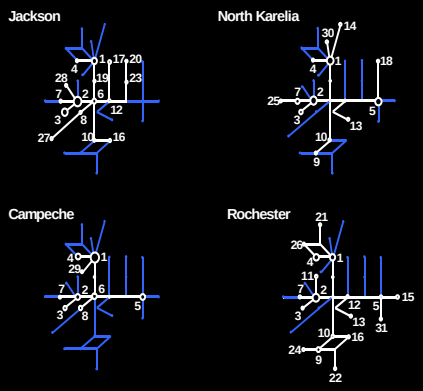


Migration and Population Structure

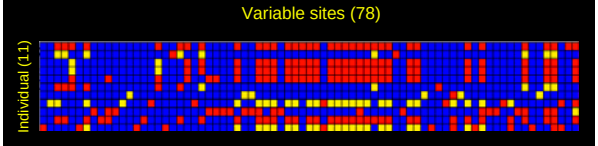
- Does the Hardy-Weinberg principle hold for a population that is subdivided geographically?
- What is the relation between SNP frequency, age of the mutation, and population structure?
- Given data on genetic variation, how can we quantify the degree of population structure?

Population heterogeneity in haplotype frequencies (*ApoE*)





Angiotensin Converting Enzyme (ACE)



AA, *Aa*, *aa*

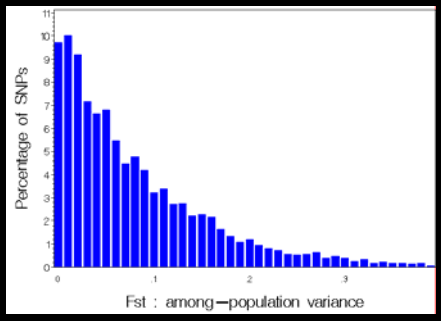
Rieder et al. (1999)

Quantifying population structure

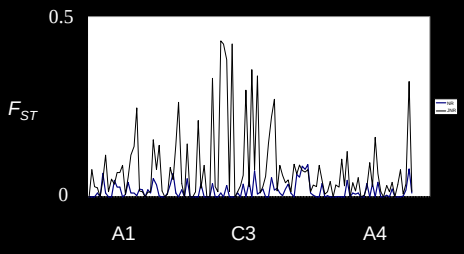
- Suppose there are two subpopulations, with allele frequencies (p_1,q_1) and (p_2,q_2) and average allele frequencies (P and Q).
- $H_T = 2PQ$ = heterozygosity in one large panmictic population
- $H_S = (2p_1q_1+2p_2q_2)/2$ is average heterozygosity across populations
- $F_{ST} = (H_T-H_S)/H_T$

Note – unequal sample sizes require more calculation

Average F_{ST} for human SNPs is 0.08



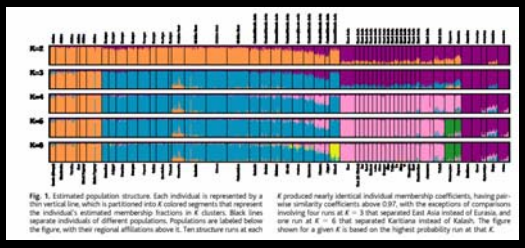
Population differentiation (F_{ST}) Varies among SNPs and genes



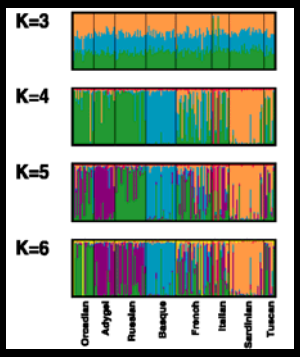
Pritchard et al. method for inferring population substructure

- Specific number of subdivisions.
- Randomly assign individuals.
- Assess fit to HW.
- Pick an individual and consider a swap.
- If fit improves, accept swap, otherwise accept with a certain probability.
- Markov chain Monte Carlo – gets best fitting assignment.

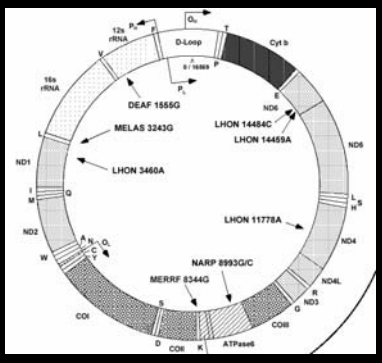
Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002 Genetic structure of human populations. Science. 298:2381-2385.



Inference of *K*

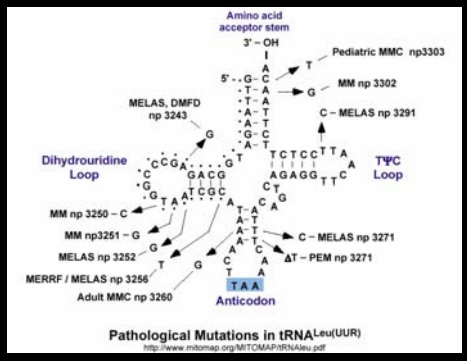


The human mitochondrial genome – 16,659 bp



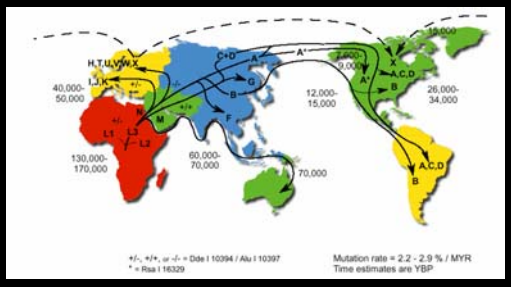
www.mitomap.org

Fine-structure mapping of mitochondrial defects



MELAS: mitochondrial myopathy encephalomyopathy lactic acidosis and stroke

Major human migrations inferred from mtDNA sequences



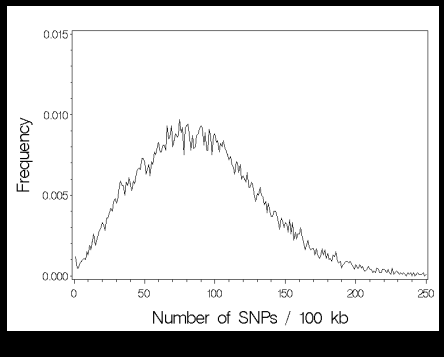
Genome-wide SNP discovery

ACATGCTGACTGACATGCTAGCTGA
GATGCTGACTGACATTCTA
ATGCTGACTGACATTCTAG
TGCTGACTGACATGCTAGC
TGCTGACTGACATGCTAGCT
GCTGACTGACATTCTAGCT
CTGACTGACATGCTAGCTGA

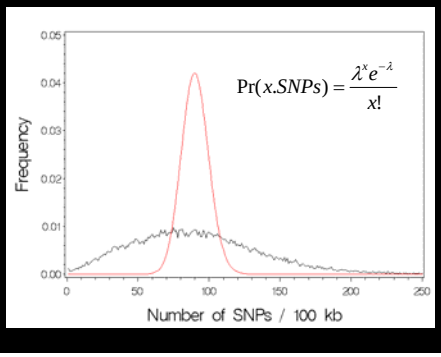
Ssaha SNP

- **S**equence **S**earch and **A**lignment by **H**ashing **A**lgorithm.
- Align reads; apply ad hoc filters to call SNPs
- http://www.sanger.ac.uk/Software/analysis/SSAHA/

Distribution of SNP Density Across the Genome



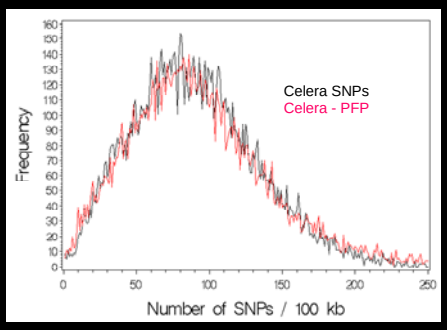
Observed SNP Distribution is not Poisson



Why the Poisson distribution fits badly

- Time to common ancestry for a random pair of alleles is distributed exponentially.
- So the Poisson parameter varies from one region to another.
- Because the time to common ancestry varies widely, the expected number of segregating mutations varies widely as well.
- But variation in ancestry time is not sufficient to explain the magnitude of variation in SNP density.

Similar inference from Celera-only as from Celera vs. public SNPs



Nucleotide diversity (x 10^{-4}) by chromosome

1	7.29	13	7.75
2	7.39	14	7.32
3	7.46	15	7.84
4	7.84	16	8.85
5	7.42	17	7.92
6	7.83	18	7.76
7	8.03	19	9.04
8	8.06	20	7.69
9	8.14	21	8.54
10	8.26	22	8.19
11	7.89	X	4.89
12	7.55	Y	2.82

Mixture models allowing heterogeneity in mutation and recombination can fit the data well

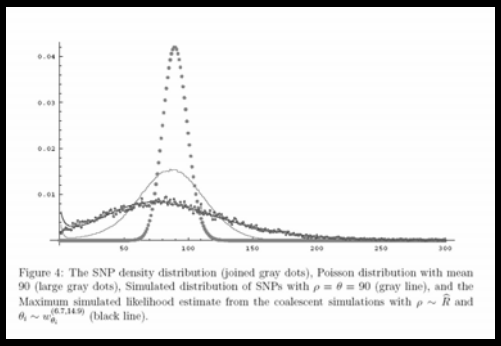


Figure 4: The SNP density distribution (joined gray dots), Poisson distribution with mean 90 (large gray dots), Simulated distribution of SNPs with $\rho = \theta = 90$ (gray line), and the Maximum simulated likelihood estimate from the coalescent simulations with $\rho \sim \hat{R}$ and $\theta_i \sim w_{\theta_i}^{(6.7,14.9)}$ (black line).

Sainudiin et al, submitted

Mutation-drift balance: the null model

- Model with pure mutation
- The Wright-Fisher model of drift
- Infinite alleles model
- Infinite sites model
- The neutral coalescent

Motivation

- Are genome-wide data on human SNPs compatible with any particular MODEL?
- Perhaps more useful -- are there models that can be REJECTED ?
- Models tell us not only about what genetic attributes we need to consider, they also can provide quantitative estimates for rates of mutation, effective population size, etc.

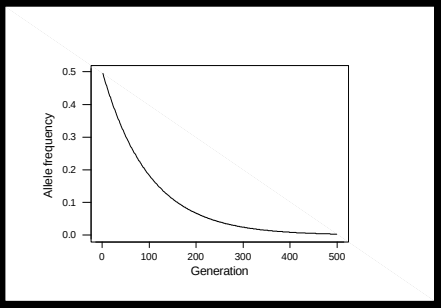
Pure Mutation

- Suppose a gene mutates from *A* to *a* at rate μ per generation. How fast will allele frequency change?
- Let *p* be the frequency of *A*.
- Develop a recursion: $p_{t+1} = p_t(1-\mu)$

Pure Mutation (2)

- What happens over time, if $p_{t+1} = p_t(1-\mu)$?
- $p_{t+2} = p_{t+1}(1-\mu) = p_t(1-\mu)(1-\mu)$
- By induction, $p_t = p_0(1-\mu)^t$
- Eventually, *p* goes to zero.

Pure Mutation (3)



μ = 0.01

For a typical mutation rate of 10^{-8} per nucleotide the "half-life" is 69 million generations

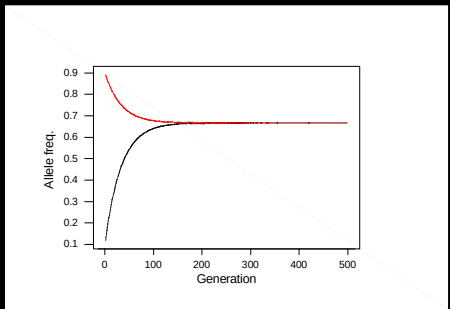
Pure Mutation (4)

- What if mutation is reversible? Let the reverse mutation rate, from *a* back to *A* occur at rate ν.
- $p_{t+1} = p_t(1-\mu) + q_t\nu$
- What happens to the allele frequency now?
- Solve for an equilibrium, where $p_{t+1} = p_t$

Pure Mutation (5)

- $p_{t+1} = p_{t+1}(1-\mu) + q_t\nu$ df
- Let $p_t = p_{t+1} = p^*$, and $q_t = 1-p^*$
- $p_{t+1} = p_t(1-\mu) + q_t\nu$, after substituting, gives
- $p^* = p^*(1-\mu) + (1-p^*)\nu$
- $p^* = p^* - p^*\mu + \nu - p^*\nu$
- $p^*(\nu+\mu) = \nu$
- $p^* = \nu/(\nu+\mu)$

Pure Mutation (6)



μ = 0.01, ν = 0.02, so p* = 2/3

Pure Drift – Binomial sampling

- Consider a population with N diploid individuals. The total number of gene copies is then $2N$.
- Initial allele frequencies for A and a are p and q, and we randomly draw WITH REPLACEMENT enough gene copies to make the next generation.
- The probability of drawing i copies of allele A is:

$$\Pr(i) = \binom{2N}{i} p^i q^{2N-i}$$

Binomial sampling

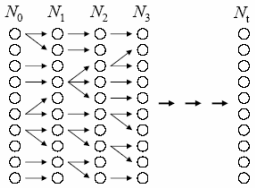
$$\Pr(i) = \binom{2N}{i} p^i q^{2N-i}$$

- If $p = q = ½$, then, for $2N = 4$ we get:

i =	0	1	2	3	4
Pr(i)=	1/16	4/16	6/16	4/16	1/16

- Note that the probability of jumping to p=0 is $(1/2)^{2N}$, so that a small population loses variation faster than a large population.

The Wright-Fisher population model



- Diploid Individuals reproduce by sexual reproduction with possibility of selfing
- Mating is random with respect to location and genotype
- Generations are non-overlapping (everyone reproduces simultaneously)
- The population size is constant of size N ($2N$ alleles)
- There is no migration or selection

Pure Drift: Wright-Fisher model

- The Wright-Fisher model is a pure drift model, and assumes only recurrent binomial sampling.
- If at present there are i copies of an allele, then the probability that the population will have j copies next generation is:

$$\Pr(i\_copies\_to\_j\_copies) = \binom{2N}{j}\left(\frac{i}{2N}\right)^j\left(1-\frac{i}{2N}\right)^{2N-j}$$

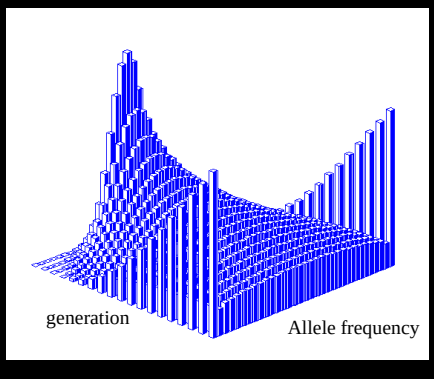
•This specifies a Transition Probability Matrix for a Markov chain.

Wright-Fisher model

- For $2N = 2$, the transition probability matrix is:

		j = 0	1	2
i =	0	1	0	0
	1	.25	.5	.25
	2	0	0	1

Wright-Fisher model

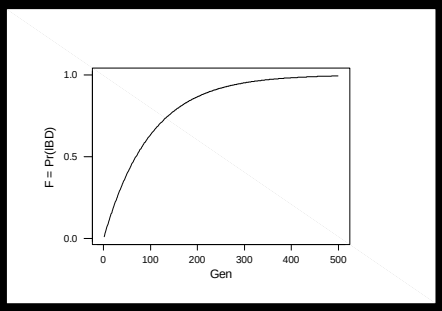


$2N = 32$

Identity by descent

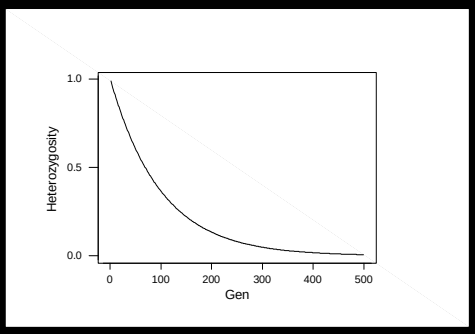
- Two alleles that share a recent common ancestor are said to be Identical By Descent
- Let F be the probability that two alleles drawn from the population are IBD.
- $F_t = 1/2N + (1 - 1/2N)F_{t-1}$ is the pure drift recursion.

F = prob(identity by descent) under pure drift



$F_{t+1} = 1/2N + (1- 1/2N)F_t$ $2N = 100$

Note that heterozygosity, $H = 1-F$



$H_{t+1} = (1- 1/2N)H_t$ $2N = 100$

Conclusions about pure drift models

- All variation is lost eventually.
- When all variation is lost, all alleles are IBD.
- Small populations lose variation faster.
- Heterozygosity declines over time, but the population remains in Hardy-Weinberg equilibrium.
- Large populations may harbor variation for thousands of generations.

Mutation and Random Genetic Drift

- The primary parameter for drift is N_e.
- Mutation occurs at rate μ, but we need to specify how mutations occur:
- Infinite alleles model: each new mutation generates a novel allele.
- Infinite sites model: each new mutation generates a change at a previously invariant nucleotide site along the gene.

Infinite alleles model

- Suppose each mutation gives rise to a novel allele.
- Then no mutant allele is IBD with any preceding allele.
- The recursion for F looks like:

$$F_t = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_{t-1}\right](1-\mu)^2$$

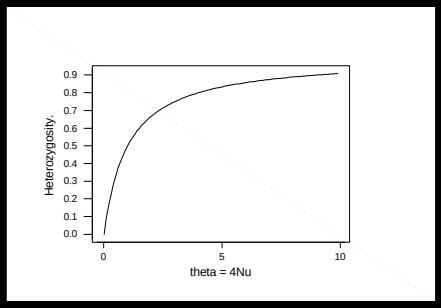
Equilibrium *F* under infinite alleles

$$F_t = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_{t-1}\right](1-\mu)^2$$

- Solve for equilibrium by letting $F_t = F_{t-1} = F^*$. After some algebra, we get:

$$F^* = \frac{1}{4N\mu + 1}$$

Steady state heterozygosity ($H = 1 - F$) under the infinite alleles model



$H = \theta/(1+\theta)$, where $\theta = 4N_e\mu$

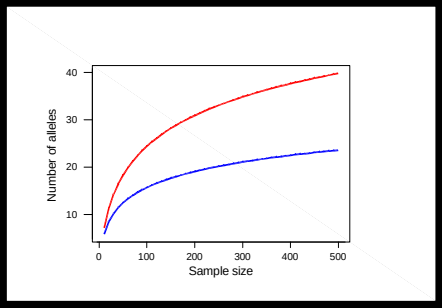
Infinite alleles model: Expected number of alleles (k) given sample size n and θ

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+n-1}$$

$\theta = 4N_e\mu$

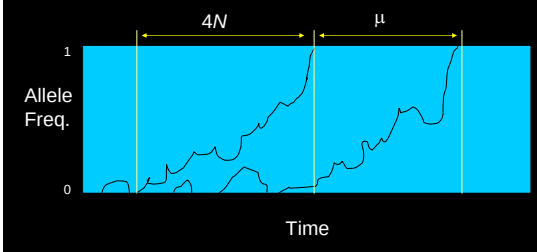
Note: assumes no recombination

Infinite alleles model: Expected number of alleles



θ =5 and θ=10

Mutation-drift and the neutral theory of molecular evolution (Motoo Kimura)



Mean time between origination and fixation = $4N$ generations
Mean interval between fixations = μ generations.

Infinite sites model: each mutation generates a change at a previously invariant nucleotide site

- Drift occurs as under the Wright-Fisher model.
- Mutations arise at rate μ at new sites each time.
- Does this model give rise to a steady state?
- How many sites do we expect to be segregating?
- What should be the steady state frequency spectrum of polymorphic sites?

Infinite sites model

Define S_i as the number of segregating sites in a sample of i genes.

$$\Pr(S_2 = j) = \left(\frac{1}{\theta+1}\right)\left(\frac{\theta}{\theta+1}\right)^j$$

So, the probability that a sample of 2 genes has zero segregating sites is:

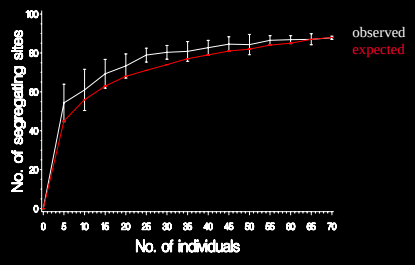
$$\Pr(S_2 = 0) = \left(\frac{1}{\theta+1}\right)$$

Note that $\Pr(S_2=0)$ is the same as the probability of identity, or F.

Infinite sites model: The expected number of segregating sites (S) depends on θ and sample size (n)

$$E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

Observed and expected numbers of segregating sites (Lipoprotein lipase, LPL)



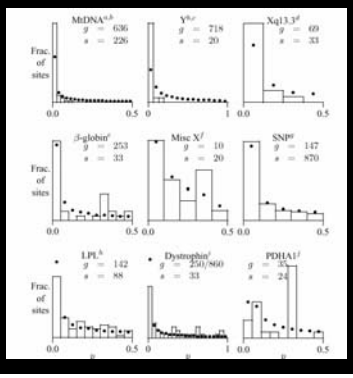
Site frequency spectrum

- Under the infinite sites model, the expected number of

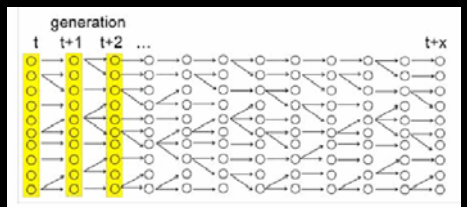
  singletons is θ
  doubletons is θ/2
  tripletons is θ/3
  …
  n-pletons is θ/n

Note that the expected number of singletons is invariant across sample sizes!

Some observed human site frequency spectra

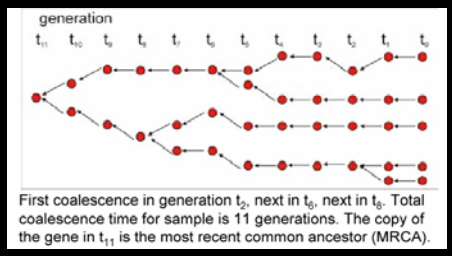


Looking forward in time – the Wright-Fisher model



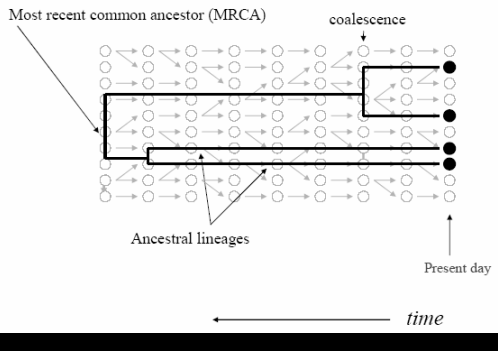
2N = 10 copies of gene in population. (N is the number of diploid individuals.)

Modeling the ancestral history of a sample: The Coalescent

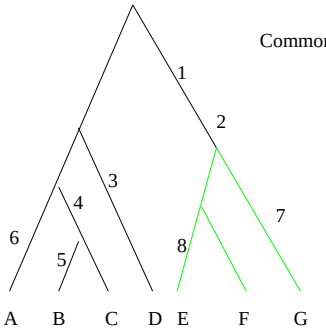


First coalescence in generation t_2, next in t_6, next in t_8. Total coalescence time for sample is 11 generations. The copy of the gene in t_{11} is the most recent common ancestor (MRCA).

The coalescent: samples in populations



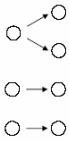
Relating the neutral coalescent to observed sequence data



Common ancestor = 00000000

A: 00000100
B: 00011000
C: 00010000
D: 00100000
E: 11000001
F: 11000000
G: 11000010

The genealogical process for two chromosomes



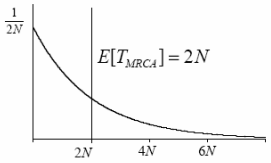
Probability from same parent (coalescence) $= \frac{1}{2N}$

Probability from different parents $= 1 - \frac{1}{2N}$

Probability of coalescence t generations ago $= \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$

Not coalesced for first t-1 generations

Coalesce in next generation



63% of outcomes have $T_{MRCA} < 2N$

Expected time to the next coalescence

- Pr(2 alleles had two distinct parents) = $1 - 1/2N$
- Pr (3 alleles had 3 distinct parents) = $(1 - 1/2N)$(prob 3rd is different)
  $= (1 - 1/2N)(1 - 2/2N)$
- Pr (k alleles had k distinct parents) = $\prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right) \approx 1-\frac{\binom{k}{2}}{2N}$

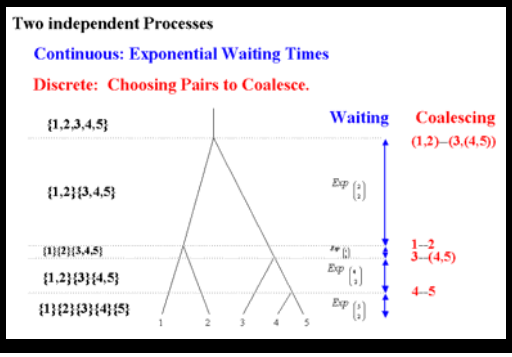
Pr(k alleles had k lineages for t generations, then k-1 lineages at $t+1$ generations ago)

= Pr(k lineages)^t × [1-Pr(k lineages had k parents)]

$$= \frac{\binom{k}{2}}{2N}\left(1-\frac{\binom{k}{2}}{2N}\right)^t \approx \frac{\binom{k}{2}}{2N} e^{-\frac{\binom{k}{2}}{2N}}$$

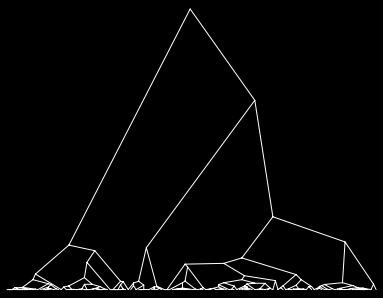
If time is rescaled in units of $2N$ generations, this is simply the exponential distribution, with parameter (k choose 2)$^{-1}$.

Simulation of coalescent trees: Branch lengths and topology



Simulation of gene genealogies:

n = 142, *S* = 88



OMIM: Online Mendelian Inheritance of Man

- Over 9000 traits have been identified and the chromosome location for more than six thousand of these genes has been determined
- Victor McKusick from Johns Hopkins University and colleagues compiled a catalog of human genetic traits
- Each trait is assigned a catalog number (called the OMIM number).
- 94% of traits are autosomal, 5% are X-linked, .4% are Y-linked, and 0.6 % are mitochondrial

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Balance between mutation and selection

- Suppose mutations occur from the normal (*A*) to the mutant (*a*) form at rate μ.
- Suppose the trait is recessive and has a reduction in fitness of *s*.
- The fitness of genotypes:

AA	*Aa*	*aa*
1	1	1-*s*

Ignore mutation for a moment….

- If zygotes have frequencies $p^2 : 2pq : q^2$, then after selection the frequencies are $p^2 : 2pq : q^2(1\text{-}s)$.
- Recall that q = ½ freq (*Aa*) + freq(*aa*)
- This means:

$$q' = \frac{pq + q^2(1-s)}{p^2 + 2pq + q^2(1-s)}$$

Now add mutation back in

- Mutations increase the frequency of a according to the equation $q' = q + p\mu = q + (1\text{-}q)\mu$.
- This yields:

$$q' \approx \frac{pq + q^2(1-s)}{p^2 + 2pq + q^2(1-s)} + (1-q)\mu$$

Balance between mutation and selection

- This looks messy, but at equilibrium, the solution is simple:

$$\hat{q} \approx \sqrt{\frac{\mu}{s}}$$

Crude estimation of mutation rate from mutation-selection balance

- The incidence of cystic fibrosis is about 1/2000.
- It is autosomal recessive, so if this is in HW, then q^2 = 0.0005, or q = 0.0224.
- Apply the equilibrium equation:

$$\hat{q} \approx \sqrt{\frac{\mu}{s}}$$

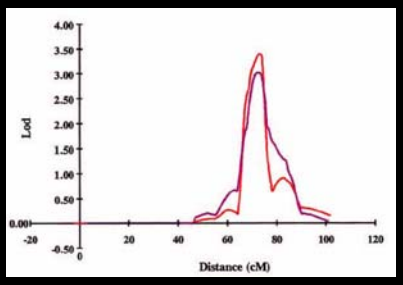
- Letting s=1, so $0.0224 = \sqrt{\mu}$

We get $\mu = 0.0005$. This is awfully high….

Linkage disequilibrium and HapMap

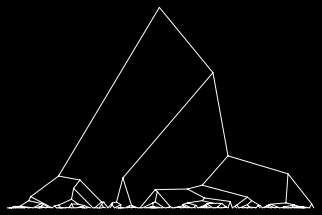
- The Problem – how to map to finer resolution than pedigrees allow.
- Definition of Linkage Disequilibrium.
- Some theory about linkage disequilibrium.
- Patterns of LD in the human genome
- The HapMap project.

The Limit to Resolution of Pedigree Studies



The typical resolution in mapping by pedigree studies is shown above--the 20 centiMorgan peak width is about 20 Megabase pairs….

Possible solution



Sampling from a POPULATION (not just families) means that many rounds of recombination may have occurred in ancestral history of a pair of alleles. Maybe this can be used for mapping….

Theory of Two Loci

- Consider two loci, *A* and *B*, each of which has two alleles segregating in the population.
- This gives four different HAPLOTYPES: *AB*, *Ab*, *aB* and *ab*.
- Define the frequencies of these haplotypes as follows:

p_{AB} = freq(*AB*)

p_{Ab} = freq(*Ab*)

p_{aB} = freq(*aB*)

p_{ab} = freq(*ab*)

Linkage equilibrium

- Suppose the frequencies of alleles *A* and *a* are p_A and p_a. Let the frequencies of *B* and *b* be p_B and p_b.
- Note that $p_A + p_a = 1$ and $p_B + p_b = 1$.
- If loci *A* and *B* are independent of one another, then the chance of drawing a gamete with *A* and with *B* is $p_A p_B$. Likewise for the other gametes:

$$p_{AB} = \text{freq}(AB) = p_A p_B$$

$$p_{Ab} = \text{freq}(Ab) = p_A p_b$$

$$p_{aB} = \text{freq}(aB) = p_a p_B$$

$$p_{ab} = \text{freq}(ab) = p_a p_b$$

- This condition is known as LINKAGE EQUILIBRIUM

Linkage DISequilibrium

- LINKAGE DISEQUILIBRIUM refers to the state when the haplotype frequencies are not in linkage equilibrium.
- One metric for it is *D*, also called the linkage disequilibrium parameter.

$$D = p_{AB} - p_A p_B$$

$$-D = p_{Ab} - p_A p_b$$

$$-D = p_{aB} - p_a p_B$$

$$D = p_{ab} - p_a p_b$$

- The sign of *D* is arbitrary, but note that the above says that a positive *D* means the *AB* and *ab* gametes are more abundant than expected, and the *Ab* and *aB* gametes are less abundant than expected (under independence).

Linkage disequilibrium measures

From the preceding equations for *D*, note that we can also write:

$$D = p_{AB}p_{ab} - p_{Ab}p_{aB}$$

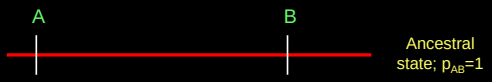
The maximum value *D* could ever have is if $p_{AB} = p_{ab} = ½$. When this is so, $D = ¼$. Likewise the minimum is $D = -¼$.

D' is a scaled LD measure, obtained by dividing *D* by the maximum value it could have for the given allele frequencies. This means that *D*' is bounded by –1 and 1.

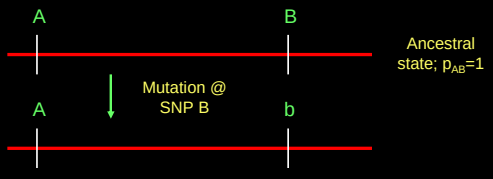
A third measure is the squared correlation coefficient:

$$r^2 = \frac{(p_{AB}p_{ab} - p_{Ab}p_{aB})^2}{p_A p_a p_B p_b}$$

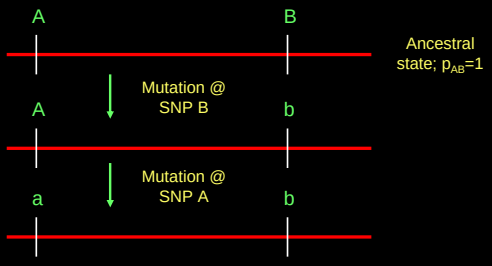
No recombination: only 3 gametes



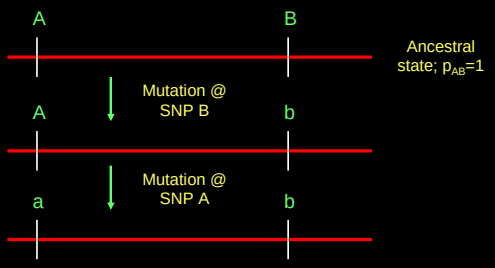
No recombination: only 3 gametes



No recombination: only 3 gametes



No recombination: only 3 gametes

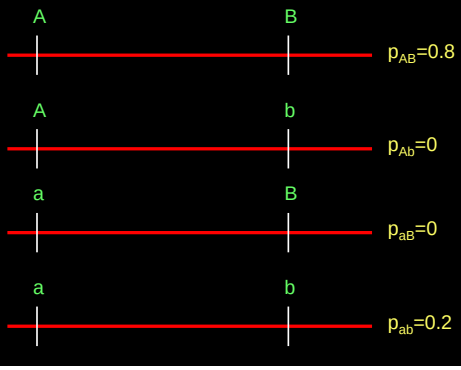


The aB gamete is missing!

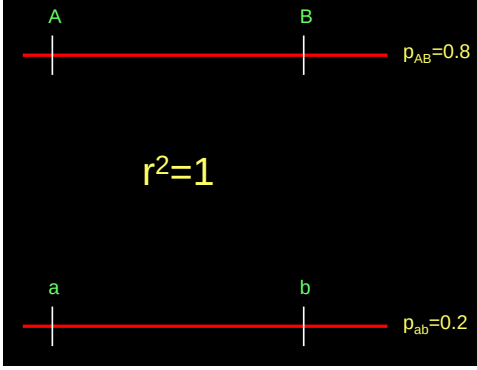
No recombination: only 3 gametes

- Under infinite-sites model: will only see all four gametes if there has been at least one recombination event between SNPs
- If only 3 gametes are present, D'=1
- Thus, D' <1, indicates some amount of recombination has occurred between SNPs

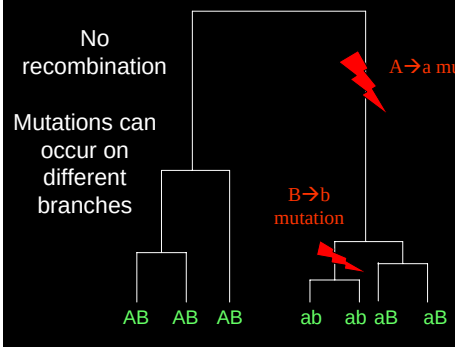
r^2 measures correlation of alleles



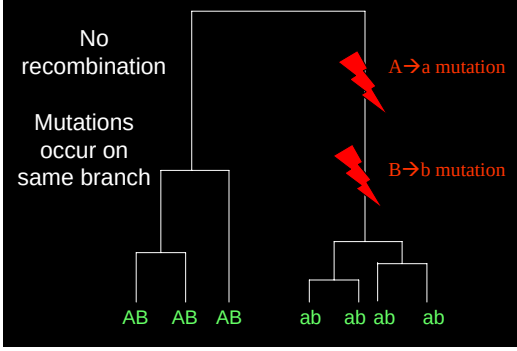
r^2 measures correlation of alleles



Genealogical interpretation of D'=1



Genealogical interpretation of r^2=1



Statistical significance of LD

Notice that the statistics for quantifying LD are simply measures of the amount of LD. They say nothing about the probability that the LD is statistically significantly different from zero.

To test statistical significance, note that the counts of the 4 haplotypes can be written in a 2 x 2 table:

	B	*b*
A	n_{AB}	n_{Ab}
a	n_{aB}	n_{ab}

To test significance, we can apply either a chi-square test, or a Fisher Exact test.

Recursion with no mutation or drift

There are four gametes (*AB*, *Ab*, *aB* and *ab*), and 10 genotypes.

Considering all the ways the 10 genotypes can make gametes, we can write down the frequency of *AB* the next generation:

$$p_{AB}' = p_{AB}^2 + p_{AB}p_{Ab} + p_{AB}p_{aB} + (1\text{-}r)p_{AB}p_{ab} + rp_{Ab}p_{aB}$$

$$= p_{AB} - rD$$

$$p_{Ab}' = p_{Ab} + rD$$

$$p_{aB}' = p_{aB} + rD$$

$$p_{ab}' = p_{ab} - rD$$

How does linkage disequilibrium change?

Note that $D' = p_{AB}'p_{ab}' - p_{Ab}'p_{aB}'$

Substituting we get:

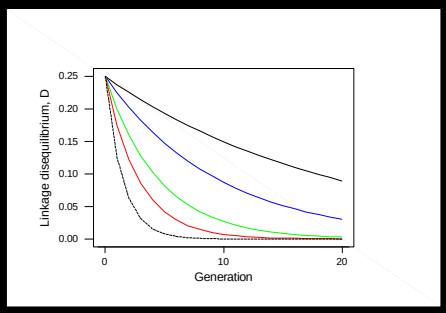
$$D' = (p_{AB} - rD)(p_{ab} - rD) - (p_{Ab} + rD)(p_{aB} + rD)$$

$$= (p_{AB}p_{ab} - p_{Ab}p_{aB}) - rD(p_{AB} + p_{ab} + p_{Ab} + p_{aB})$$

$$= D - rD$$

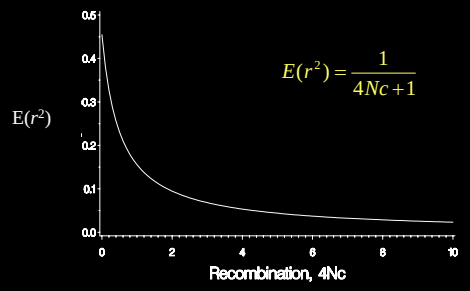
$$= (1 - r)\,D$$

Decay of LD over time.



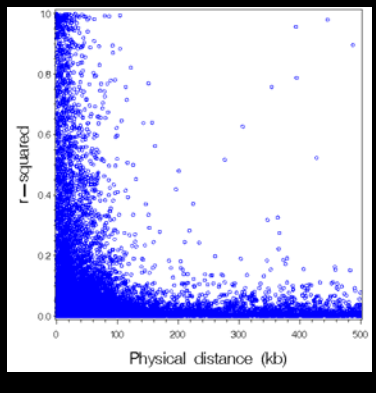
Top to bottom: r = .05, 0.1, 0.2, 0.3, 0.5

Equilibrium relation between LD and recombination rate

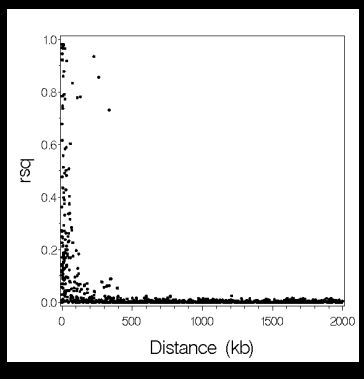


$$E(r^2) = \frac{1}{4Nc + 1}$$

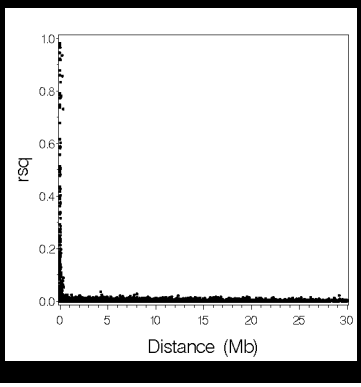
Linkage disequilibrium is rare beyond 100 kb or so

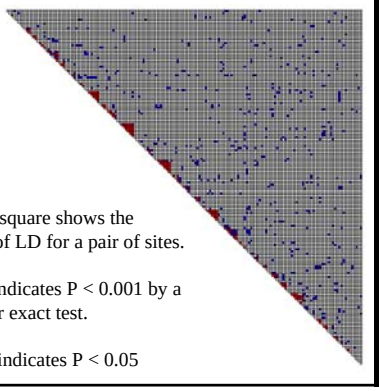


Beyond 500 kb, there is almost zero Linkage disequilibrium



…so observing LD means the sites are likely to be close together



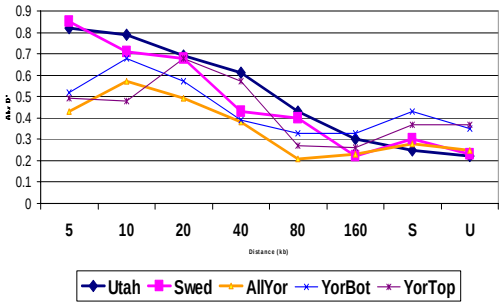


Each square shows the Test of LD for a pair of sites.

Red indicates P < 0.001 by a Fisher exact test.

Blue indicates P < 0.05

Different human populations different levels of LD

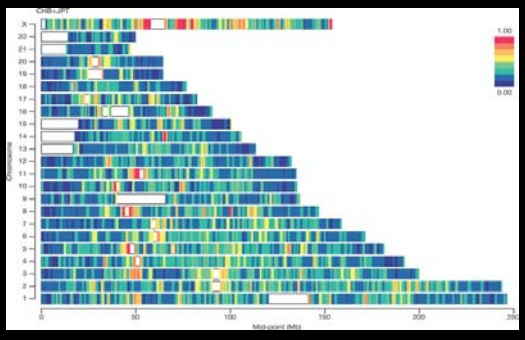


Reich et al. 2001 Nature 411:199-204.

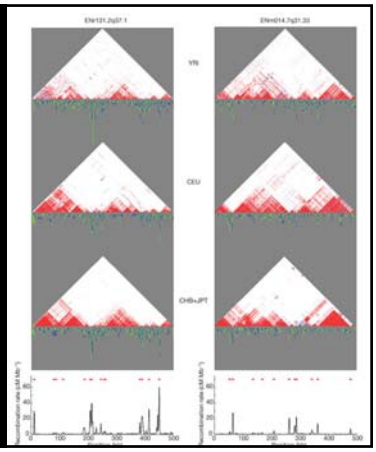
www.hapmap.org

- NIH funded initiative to genotype 1-3 millions of SNPs in 4 populations:
  - 30 CEPH trios from Utah (European ancestry)
  - 30 Yoruba trios from Nigeria (African ancestry)
  - 45 unrelated individuals from Beijing (Chinese)
  - 45 unrelated individual from Tokyo (Japanese)

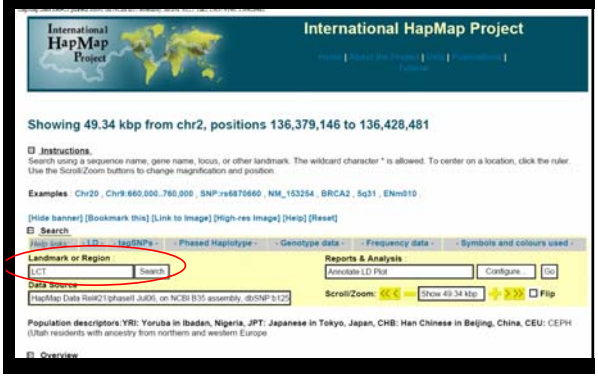
LD across the genome



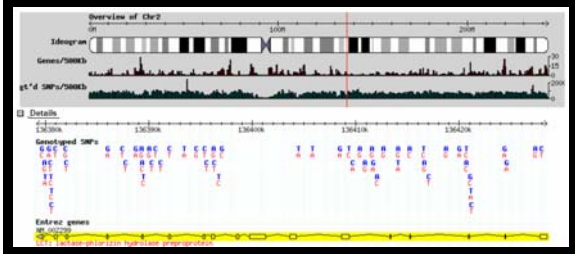
LD blocks can be broken by recombination hotspots



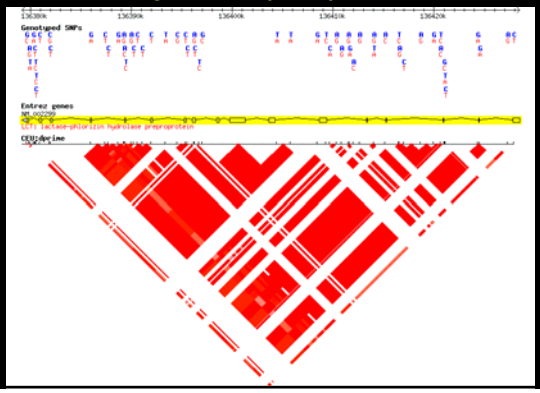
Using the HapMap website



Using the HapMap website



Using the HapMap website



Using the HapMap website