

# Forces shaping genetic diversity

- Mutation
- Random Genetic Drift
- Recombination
- Migration
- Natural Selection

# Classes of mutation

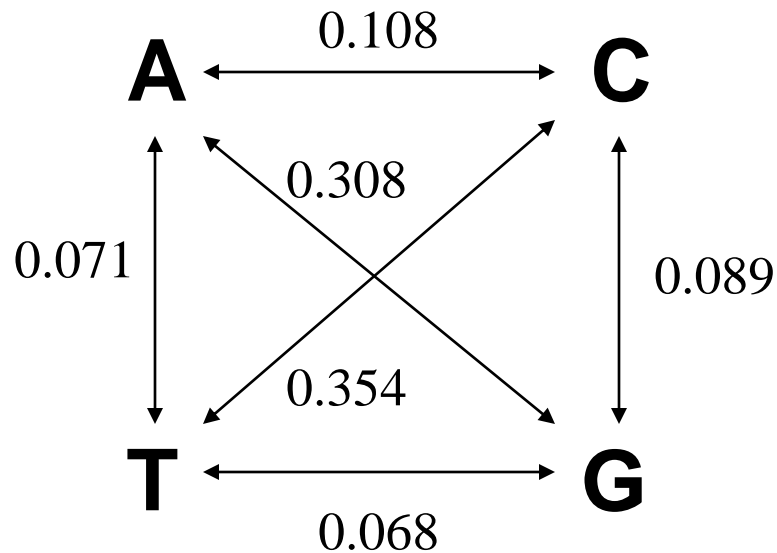
- Germline vs. somatic
- Coding vs. Noncoding
- Missense vs. Nonsense
- Synonymous vs. Nonsynonymous
- Neutral vs. Gain of function vs. Loss of function
- Indels (including Transposable Elements)
- Splicing (cryptic, frameshift, ...)
- Duplication
- Unequal exchange (Philadelphia chromosome)

# Estimation of mutation rate

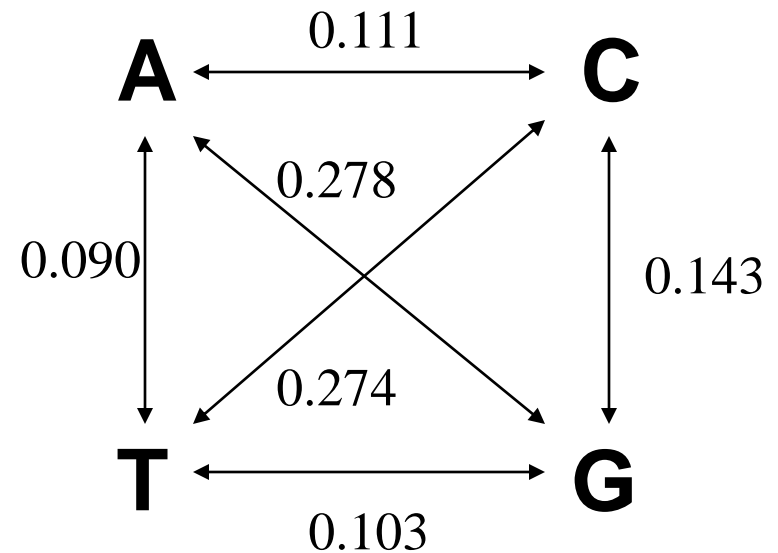
- Inter-species divergence
- Pedigree data
- Extremely rare variants in huge samples
  - Mutation spectrum
- Mutation-accumulation lines

# Mutations as seen through SUBSTITUTIONS and SNPs.

## Human-mouse divergence



## Human polymorphism



Full genome sequence of two parents and two offspring gave a mutation rate estimate of  $1.1 \times 10^{-8}$  per site per gen

Scienceexpress

Report

## Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,<sup>1\*</sup> Gustavo Glusman,<sup>1\*</sup> Arian F.A. Smit,<sup>1\*</sup> Chad D. Huff,<sup>1,2\*</sup> Robert Hubley,<sup>1</sup> Paul T. Shannon,<sup>1</sup> Lee Rowen,<sup>1</sup> Krishna P. Pant,<sup>3</sup> Nathan Goodman,<sup>1</sup> Michael Bamshad,<sup>4</sup> Jay Shendure,<sup>5</sup> Radoje Drmanac,<sup>3</sup> Lynn B. Jorde,<sup>2</sup> Leroy Hood,<sup>1†</sup> David J. Galas<sup>1†</sup>

<sup>1</sup>Institute for Systems Biology, Seattle, WA 98103, USA. <sup>2</sup>Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84109, USA. <sup>3</sup>Complete Genomics, Inc., Mountain View, CA 94043, USA. <sup>4</sup>Department of Pediatrics, University of Washington, Seattle, WA 98195, USA. <sup>5</sup>Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA.

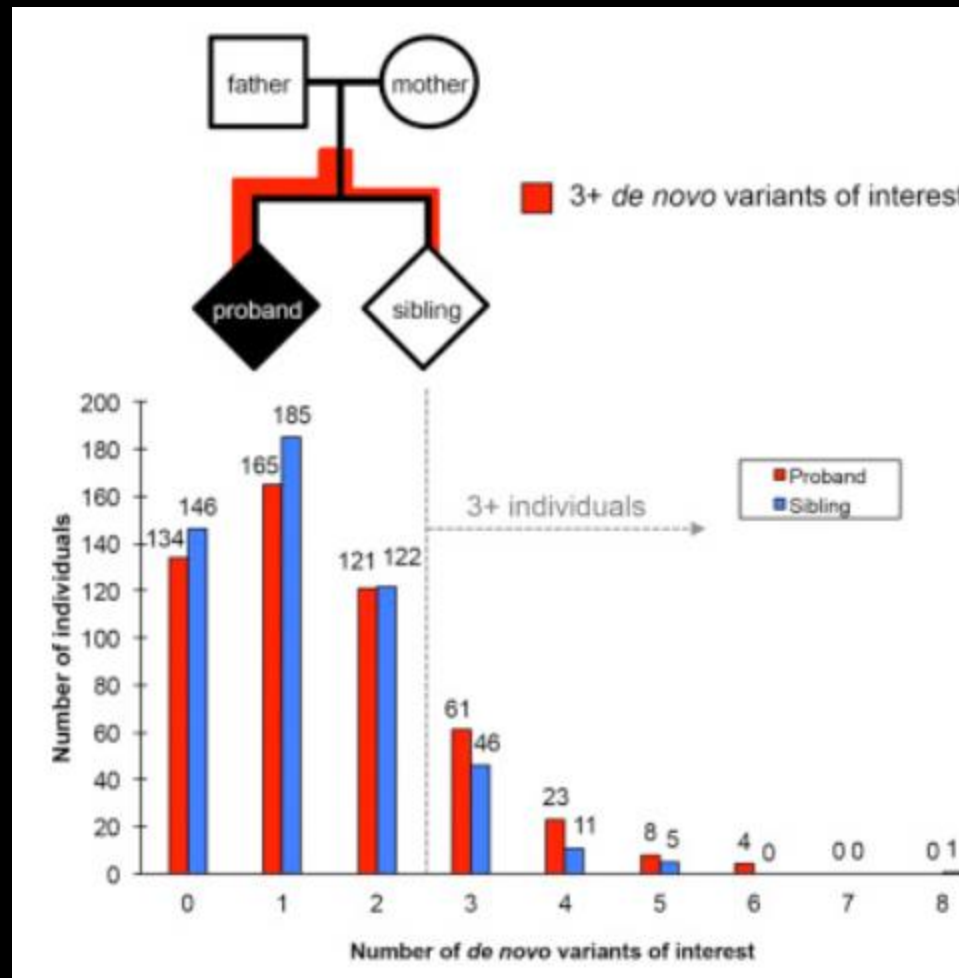
\*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: [dgalas@systemsbiology.org](mailto:dgalas@systemsbiology.org) (D.J.G.), [lh Hood@systemsbiology.org](mailto:lh Hood@systemsbiology.org) (L.H.)

**We analyzed the whole genome sequences of a family of four, consisting of two siblings and their parents. Family-based sequencing allowed us to delineate recombination sites precisely, identify 70% of the sequencing errors (resulting in 99.999% accuracy), and identify very rare**

DNA from each family member was extracted from peripheral blood cells and sequenced by Complete Genomics Inc. with a nanoarray-based short-read sequencing-by-ligation technology (1) including an adaptation of the pairwise end-sequencing strategy (2). Reads were mapped to the NCBI

# Simplex simplex quads - >600 quad families sequenced to 40x

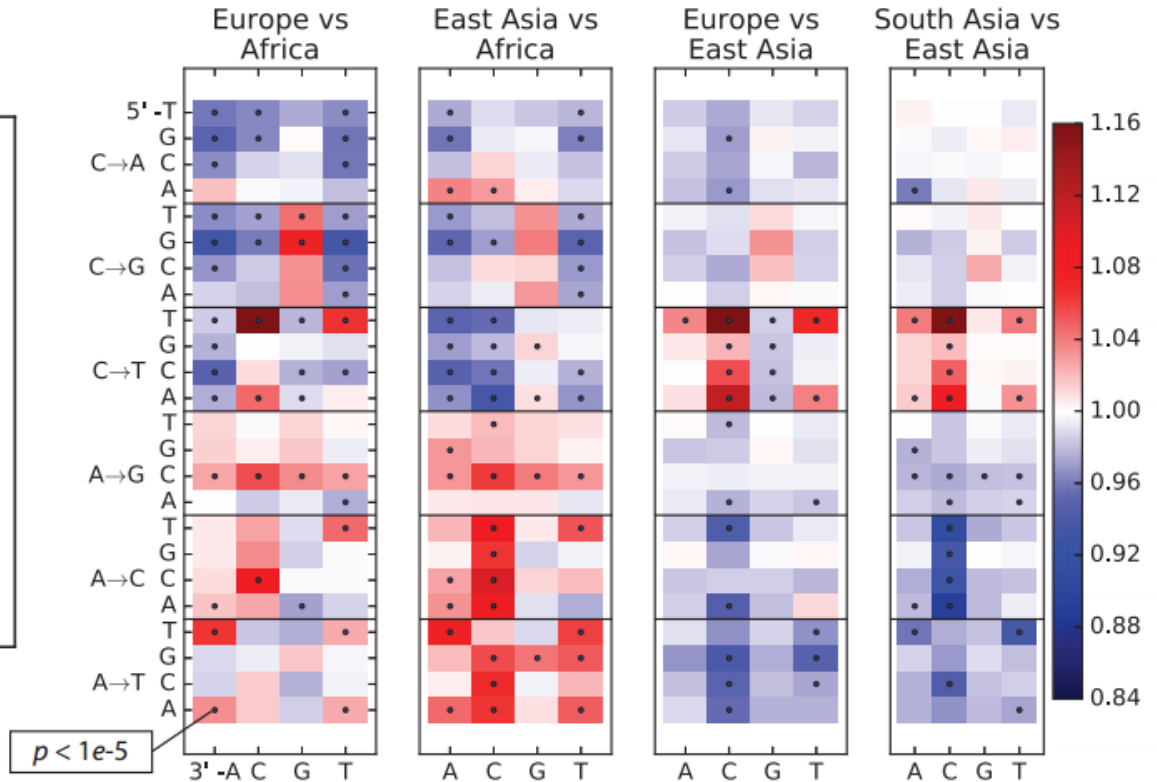
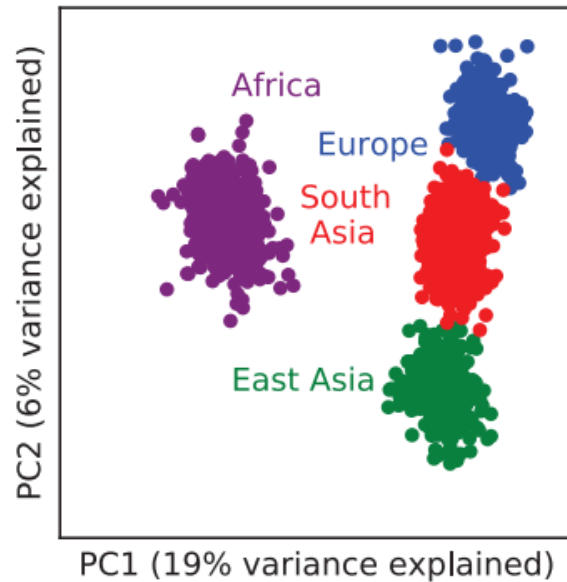


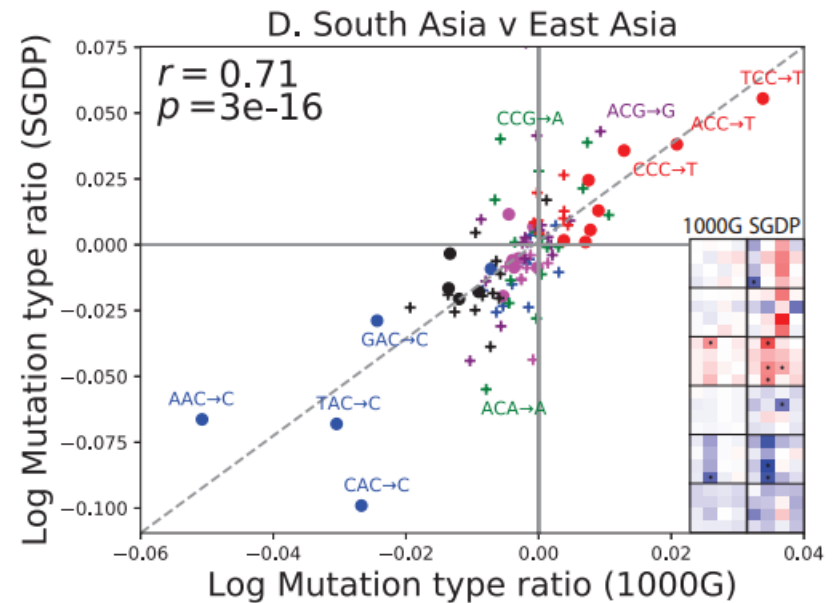
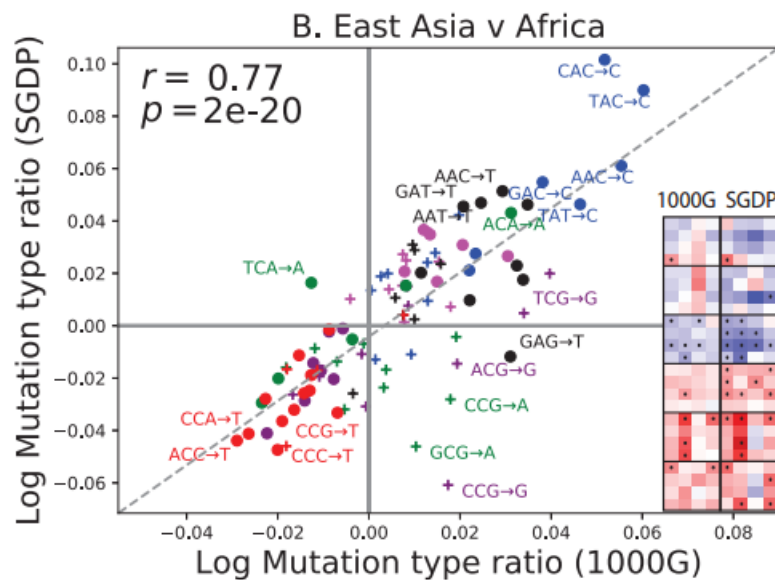
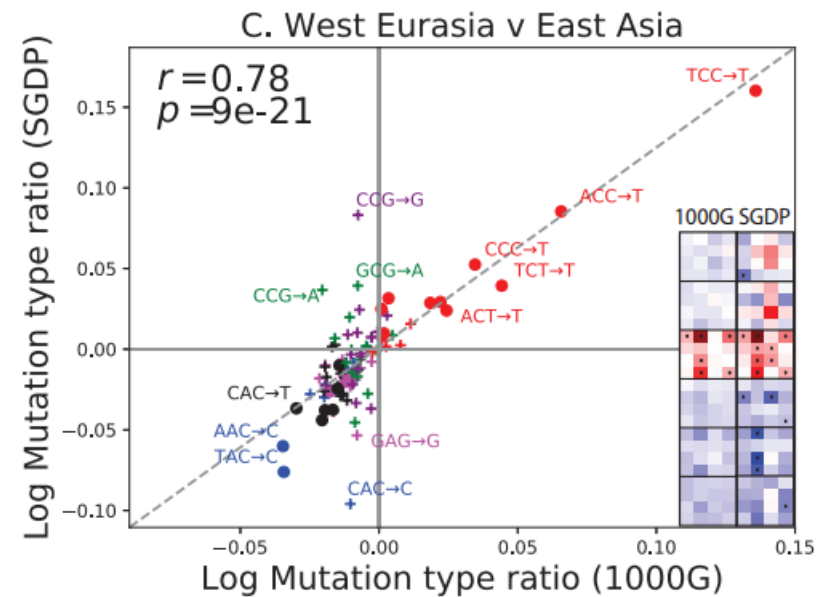
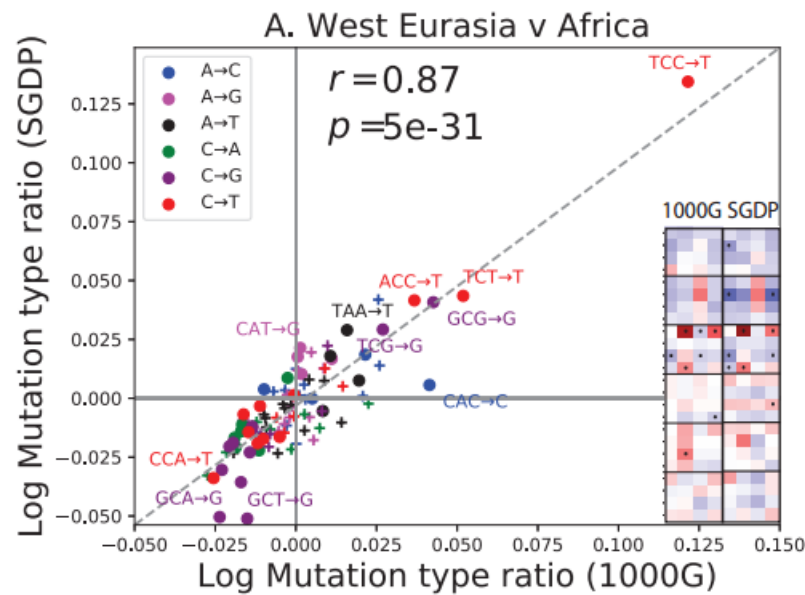
Best estimate yet of human mutation rate –  $1.5 \times 10^{-9}$ .

# Mutation spectrum and context dependence of mutation rates

B. Mutation spectrum differences between populations

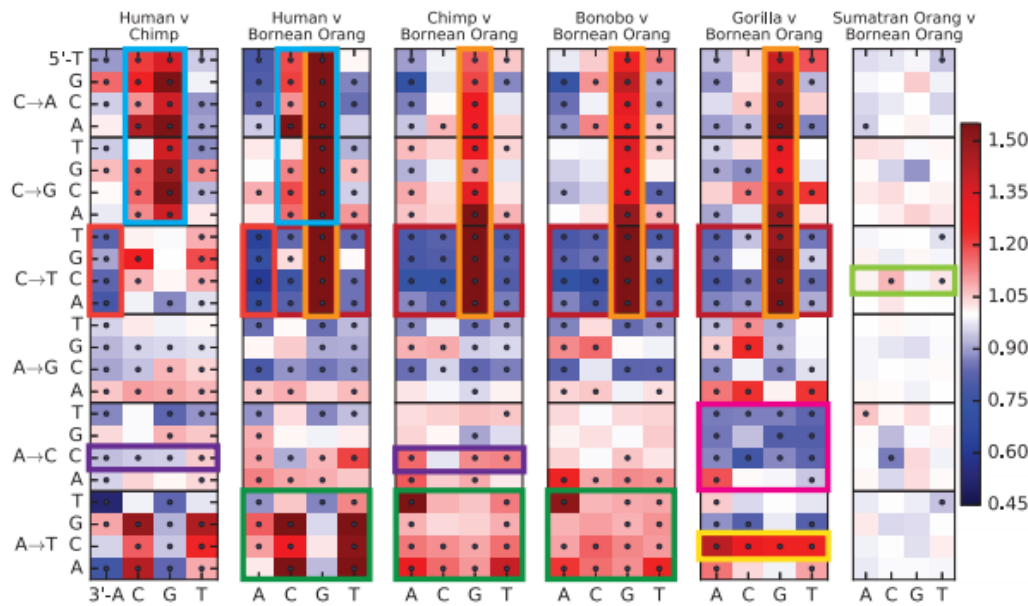
A. PCA of human mutation spectra



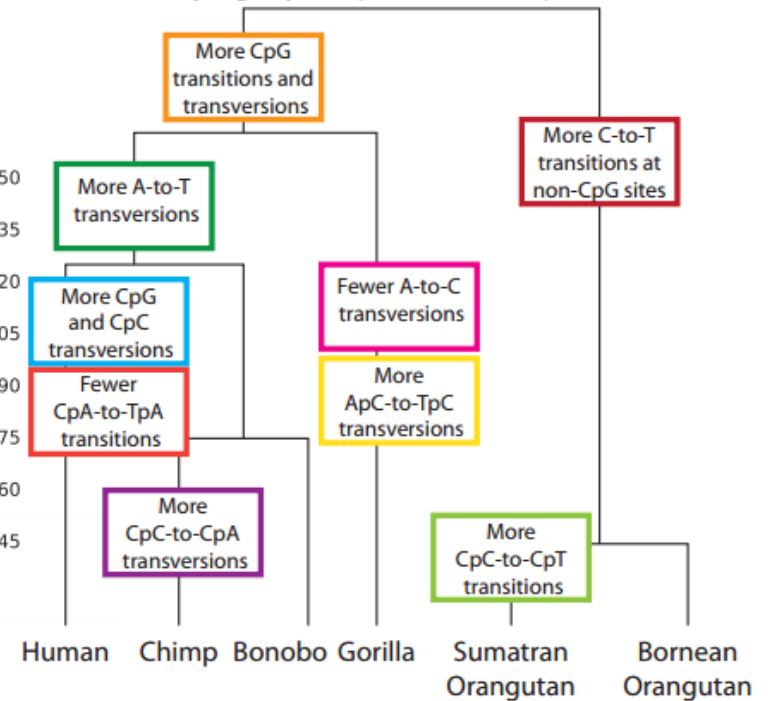




### A. Mutation spectrum contrasts between great ape species



### B. Phylogeny of ape mutation spectra



UUU	Phe	17.1	UCU	Ser	14.7	UAU	Tyr	12.1	UGU	Cys	10.1
UUC	Phe	20.4	UCC	Ser	17.5	UAC	Tyr	15.5	UGC	Cys	12.4
UUA	Leu	7.3	UCA	Ser	11.9	(UAA	STOP)		(UGA	STOP)	
UUG	Leu	12.7	UCG	Ser	4.5	(UAG	STOP)		UGG	Trp	13.0
CUU	Leu	12.9	CCU	Pro	17.3	CAU	His	10.6	CGU	Arg	4.7
CUC	Leu	19.5	CCC	Pro	20.0	CAC	His	15.0	CGC	Arg	10.8
CUA	Leu	7.0	CCA	Pro	16.7	CAA	Gln	11.9	CGA	Arg	6.3
CUG	Leu	40.1	CCG	Pro	7.0	CAG	Gln	34.4	CGG	Arg	11.8
AUU	Ile	15.8	ACU	Thr	12.9	AAU	Asn	16.7	AGU	Ser	12.0
AUC	Ile	21.3	ACC	Thr	19.1	AAC	Asn	19.3	AGC	Ser	19.4
AUA	Ile	7.2	ACA	Thr	14.9	AAA	Lys	24.0	AGA	Arg	11.7
AUG	Met	22.3	ACG	Thr	6.2	AAG	Lys	32.5	AGG	Arg	11.6
GUU	Val	10.9	GCU	Ala	18.6	GAU	Asp	22.1	GGU	Gly	10.8
GUC	Val	14.6	GCC	Ala	28.4	GAC	Asp	25.7	GGC	Gly	22.6
GUA	Val	7.0	GCA	Ala	16.0	GAA	Glu	29.0	GGA	Gly	16.4
GUG	Val	28.7	GCG	Ala	7.6	GAG	Glu	40.3	GGG	Gly	16.4

### Key

N Nondegenerate site

N Two-fold degenerate site

N Four-fold degenerate site

# Diversity of human tRNA genes from the 1000-genomes project

Marc Parisien<sup>†</sup>, Xiaoyun Wang<sup>†</sup>, and Tao Pan<sup>\*</sup>

Department of Biochemistry and Molecular Biology; University of Chicago; Chicago, IL USA

<sup>†</sup>These authors contributed equally to this paper.

**Keywords:** tRNA, isodecoder, SNP, 1000 genomes project

**Abbreviations:** ChIP-seq, crosslinking and immunoprecipitation followed by deep sequencing; SNP, single nucleotide polymorphism

Hg38 has 613 tRNA genes.

1000 genomes data reveals many polymorphisms in both sequence and number.

They found 11 new “isodecoders”.

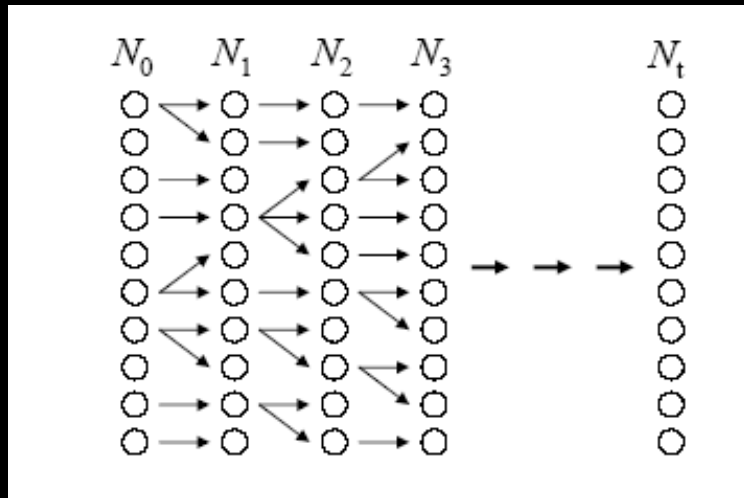
# Mutation practical

- Using the data from Olivier's practical, for all singleton's in the data set, tally the SIX types of changes for each of the 16 flanking contexts.
- Test whether the counts are significantly different across the 16 tables (simple chi-square test).

# Random genetic drift

- Pure drift models are rarely used for inference of natural populations, but model (lab) populations under short time scales can be modeled this way.
- In this circumstance, drift is modeled as a Wright-Fisher process, with recurrent binomial sampling.

# The Wright-Fisher drift model



- Selfing allowed
- Random mating
- Non-overlapping generations
- Constant population size
- No migration
- No selection

## 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:

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

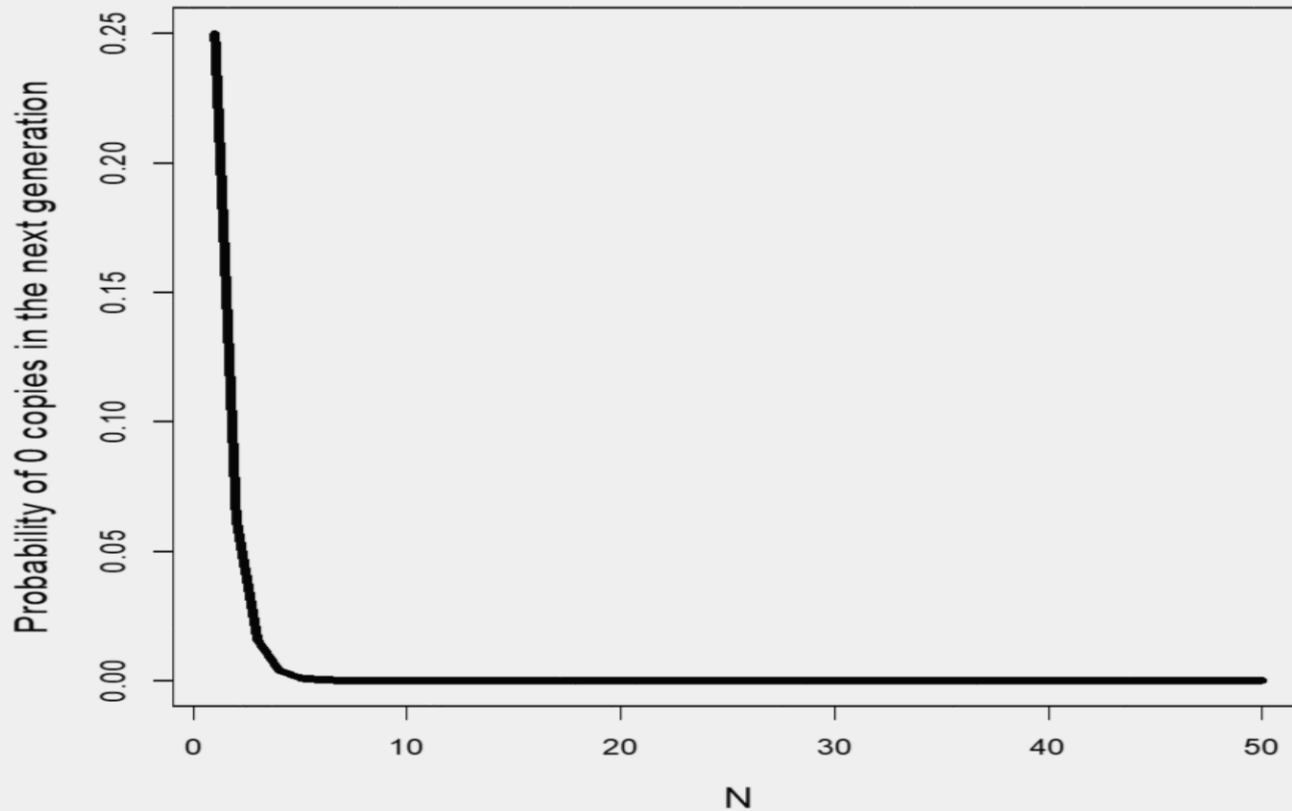
# Binomial sampling

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

- If  $p = q = 1/2$ , 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.



# Binomial sampling



$$\text{Pr}(0 \text{ copies in next generation}) = (1-p)^{2N}$$

# 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:

$$\Pr(i \text{ copies to } j \text{ copies}) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

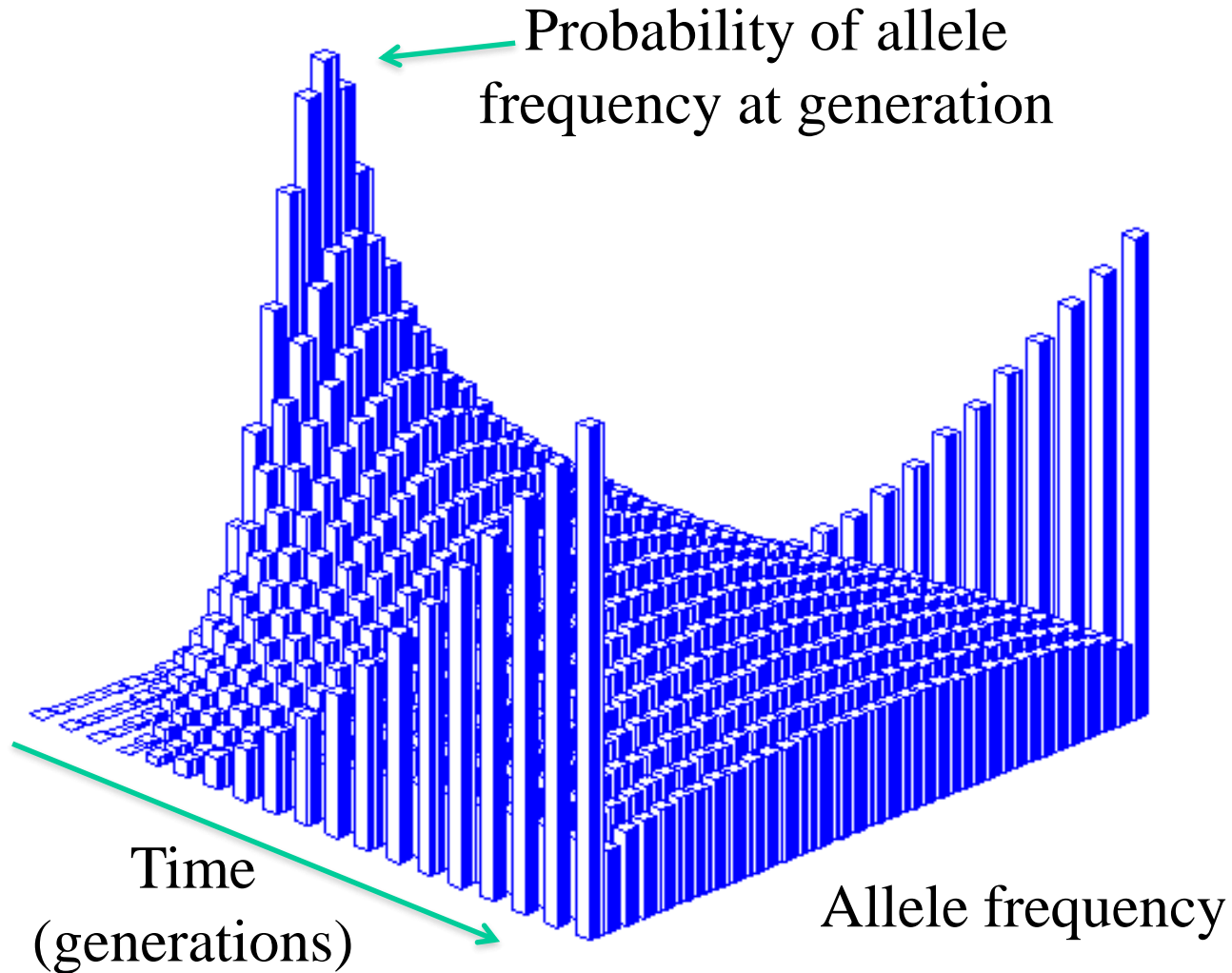
- This specifies a **Transition Probability Matrix** for a Markov chain.

# Wright-Fisher model

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

$$\begin{array}{c} \begin{array}{c} i \\ 0 \\ 1 \\ 2 \end{array} \end{array} \begin{array}{c} \begin{array}{c} j \\ 0 \\ 1 \\ 2 \end{array} \end{array} \begin{bmatrix} 1 & 0 & 0 \\ .25 & .5 & .25 \\ 0 & 0 & 1 \end{bmatrix}$$

# Wright-Fisher model



$$2N = 32$$

# Wright-Fisher practical

- Run the simulations and solve the problems given in the R program: **WFSim.r**

# Drift plus mutation

- Mutation always happens, so the simplest model of drift usually also must incorporate mutation.
- Mutation introduces all variation into the population, and drift always removes it. These two forces come to a balance – the mutation-drift balance.
- Two primary models of this are the infinitely many alleles model (infinite alleles model) and the infinite sites model.

# Infinite alleles model

- Suppose each mutation gives rise to a novel allele.
- Then no new mutant allele is IBD with any previous allele.
- So what is the probability of IBD now?
- And what happens to the probability of IBD over time?

# Equilibrium $F$ under infinite alleles

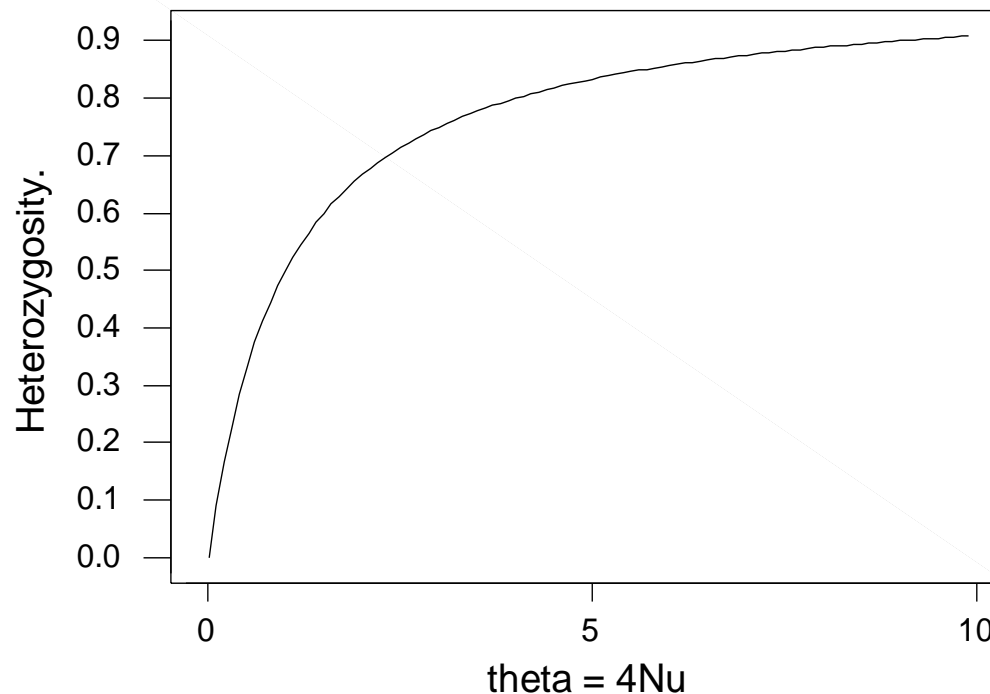
$$F_t = \left[ \frac{1}{2N} + \left( 1 - \frac{1}{2N} \right) F_{t-1} \right] (1 - \mu)^2$$

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

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



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



$$H = \theta / (1 + \theta), \quad \text{where } \theta = 4N_e\mu$$

Infinite alleles model: Expected number of alleles ( $k$ ) given sample size  $n$  and  $\theta$

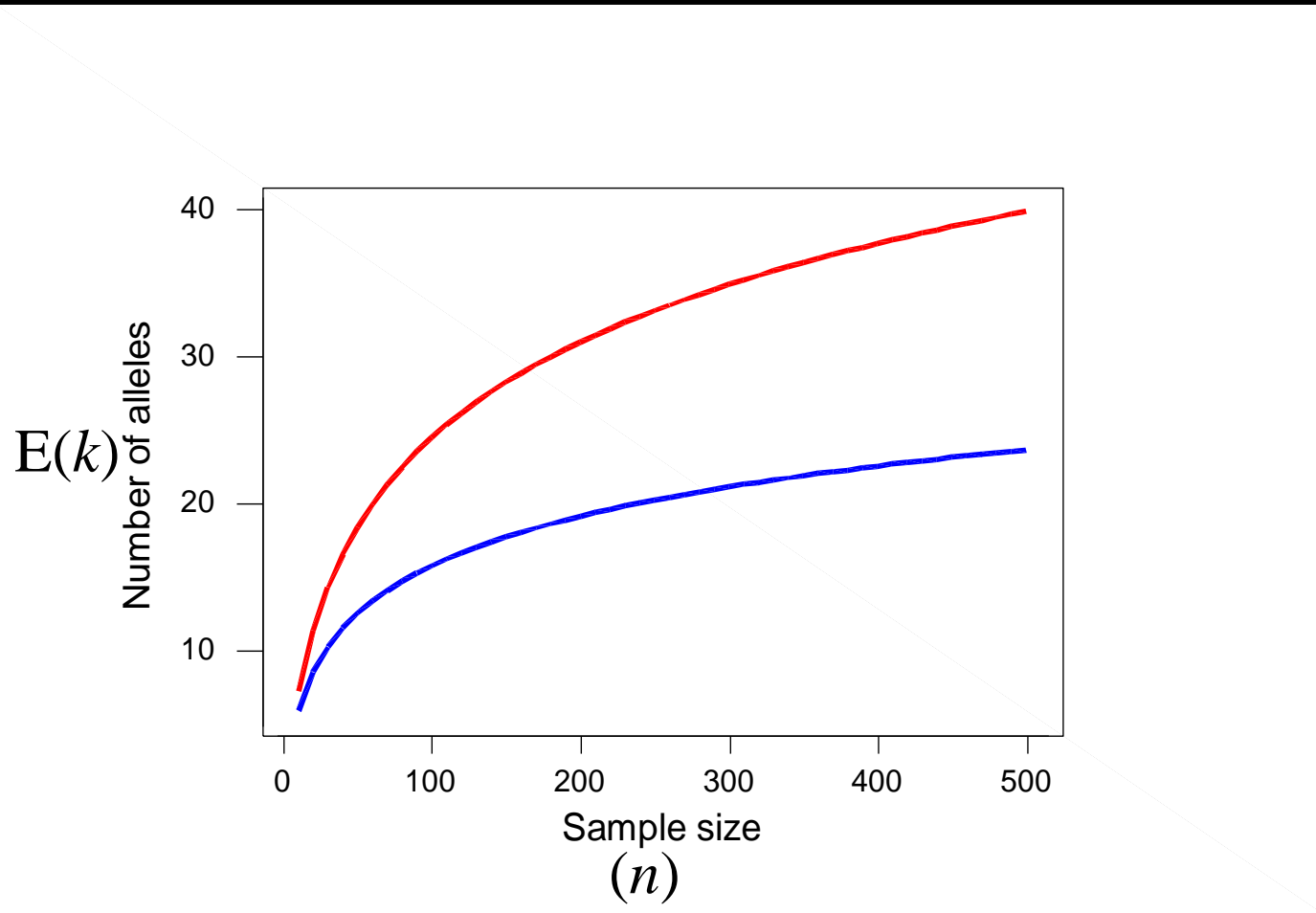
$$E(k) = 1 + \frac{q}{q+1} + \frac{q}{q+2} + \dots + \frac{q}{q+n-1}$$

Ewens 1972

$$\theta = 4N_e\mu$$

Note: assumes no recombination

# Infinite alleles model: Expected number of alleles



$\theta = 10$  and  $\theta = 5$

Infinite alleles model: Expected number of alleles ( $k$ ) given sample size  $n$  and  $\theta$

$$E(k) = 1 + \frac{q}{q+1} + \frac{q}{q+2} + \dots + \frac{q}{q+n-1}$$

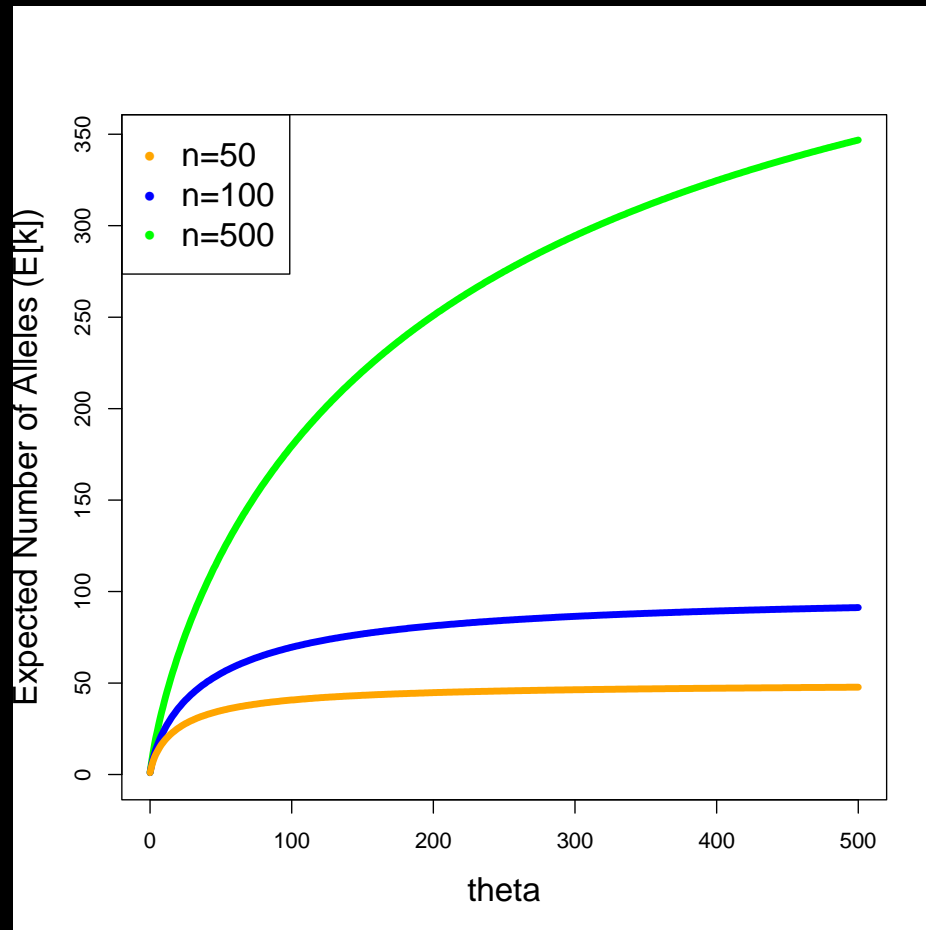
Ewens 1972

$$\theta = 4N_e\mu$$

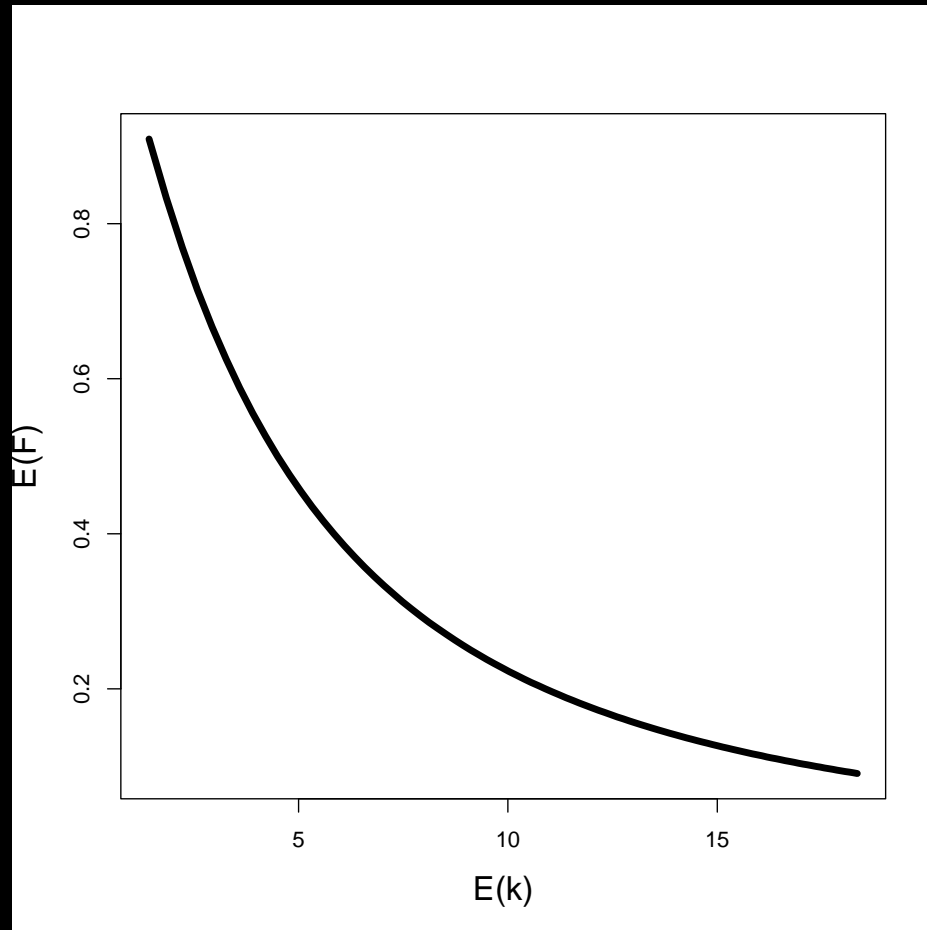
Note: assumes no recombination

# Infinite alleles model:

## Expected number of alleles



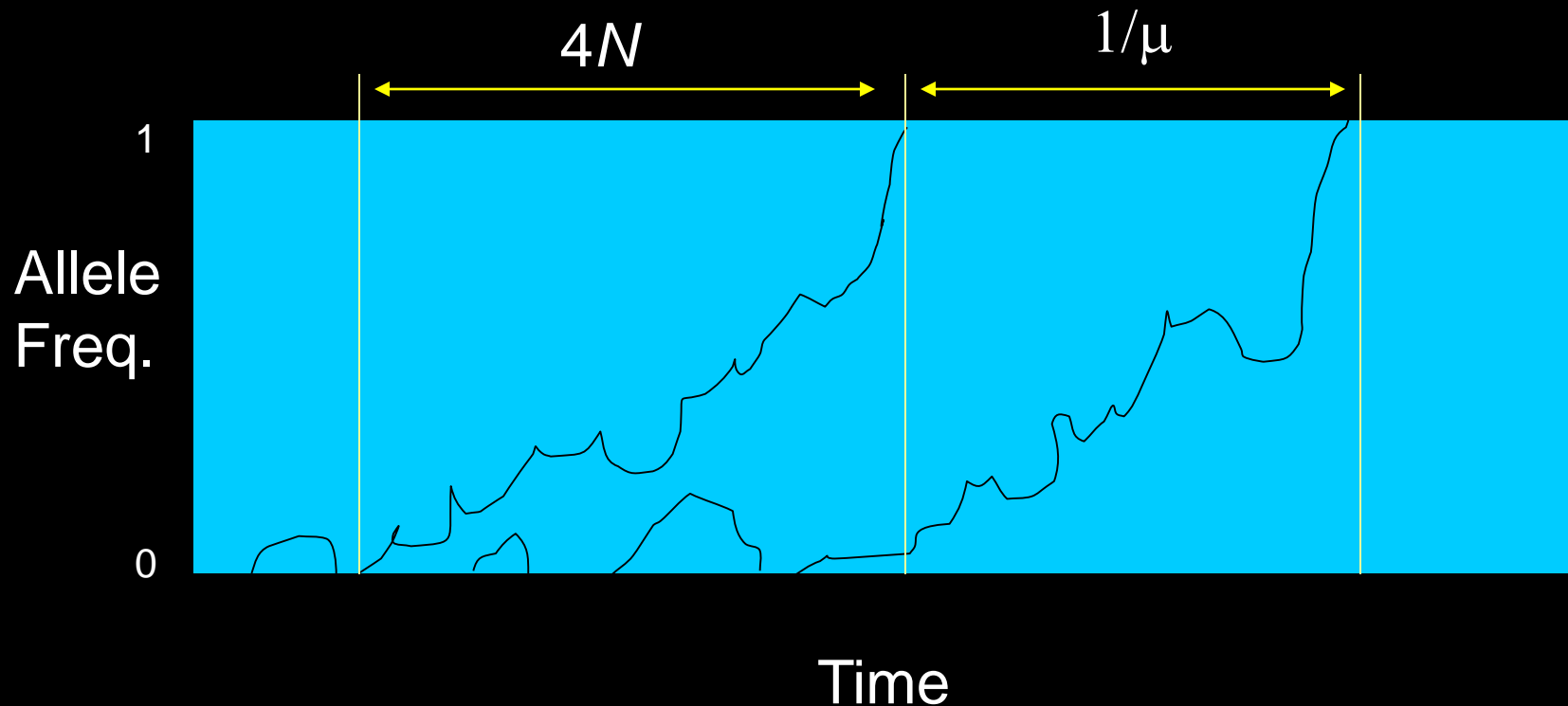
# Infinite alleles model: Expected number of alleles



# The Ewens-Watterson test: a goodness of fit test to the infinite alleles model

- Calculate observed  $F$  (homozygosity) from the data.
- Sample size ( $n$ ) and number of alleles ( $k$ ) are sufficient statistics, under the Ewens' sampling formula, to determine the expected distribution of  $F$ .
  - $F$  will differ based on allelic configuration
- Simulations are run to generate the null distribution of  $F$  given  $n$  and  $k$ .

# Mutation-drift and the neutral theory of molecular evolution (Kimura and Ohta 1969)



Assuming that an allele is eventually fixed...

mean time between origination and fixation =  $4N$  generations

Mean interval between fixations =  $1/\mu$  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  $\mu$  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  $S_2$  as the number of segregating sites in a sample of 2 gene copies.  
(Watterson 1975)

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

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

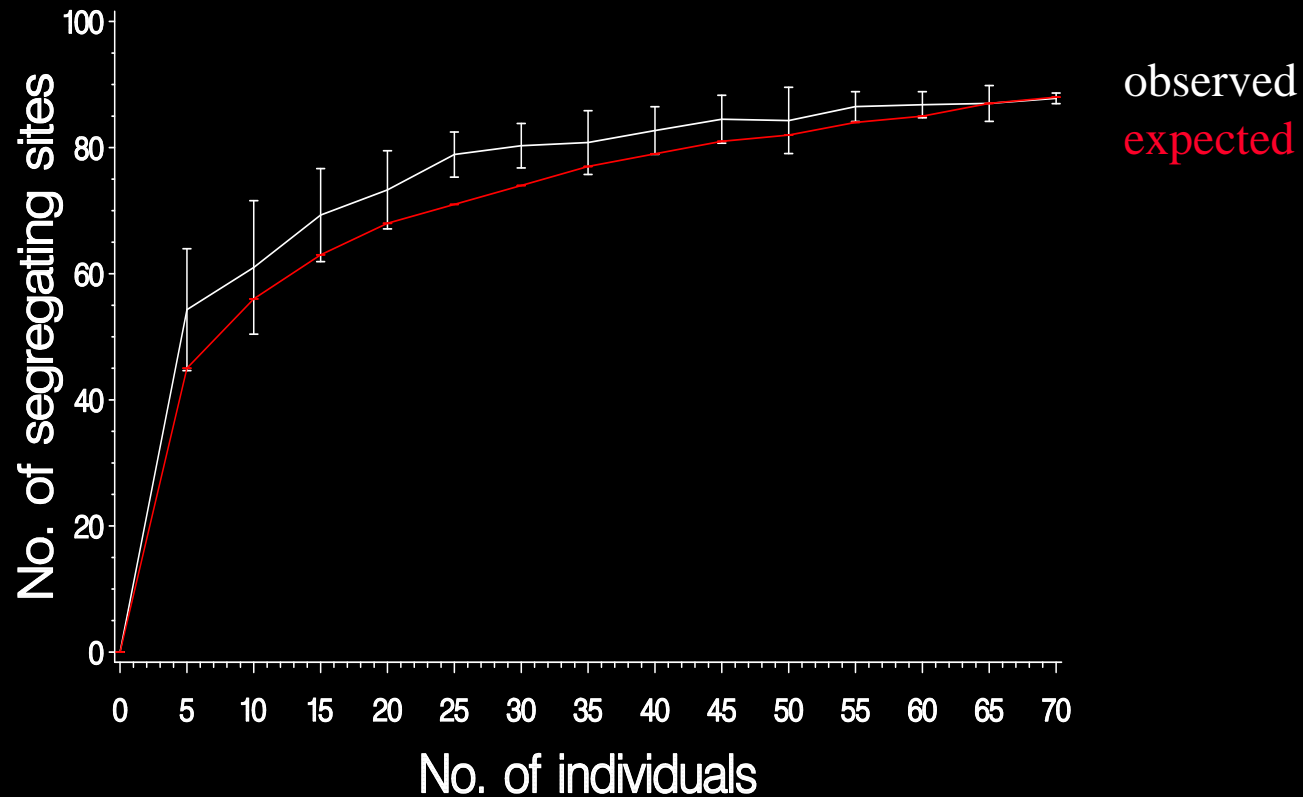
$$\Pr(S_2 = 0) = \left( \frac{1}{\theta + 1} \right)$$

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  $\theta$  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)



# How the site frequency spectrum is generated

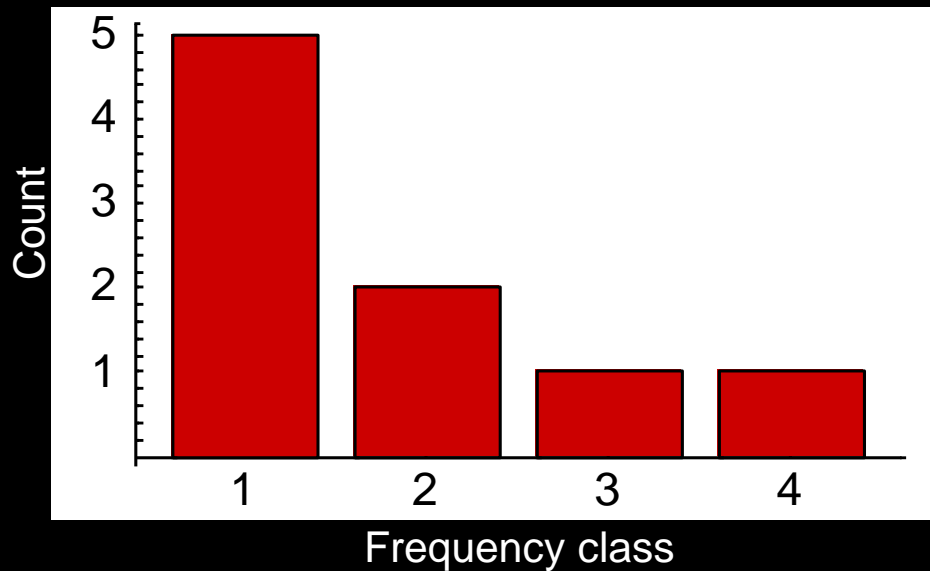
Sequence

A	G	G	C	T	T	A	A	A
A	T	G	C	T	C	G	A	A
G	T	G	T	T	C	A	C	G
A	G	G	C	T	C	A	A	G
A	G	A	C	C	C	G	A	A

Frequency class:

Ancestral       Derived

## The frequency spectrum



## Theoretical expectation for the site frequency spectrum

- Under the infinite sites model, the expected number of (Tajima 1989; Fu and Li 1993)

singletons is  $\theta$

doubletons is  $\theta/2$

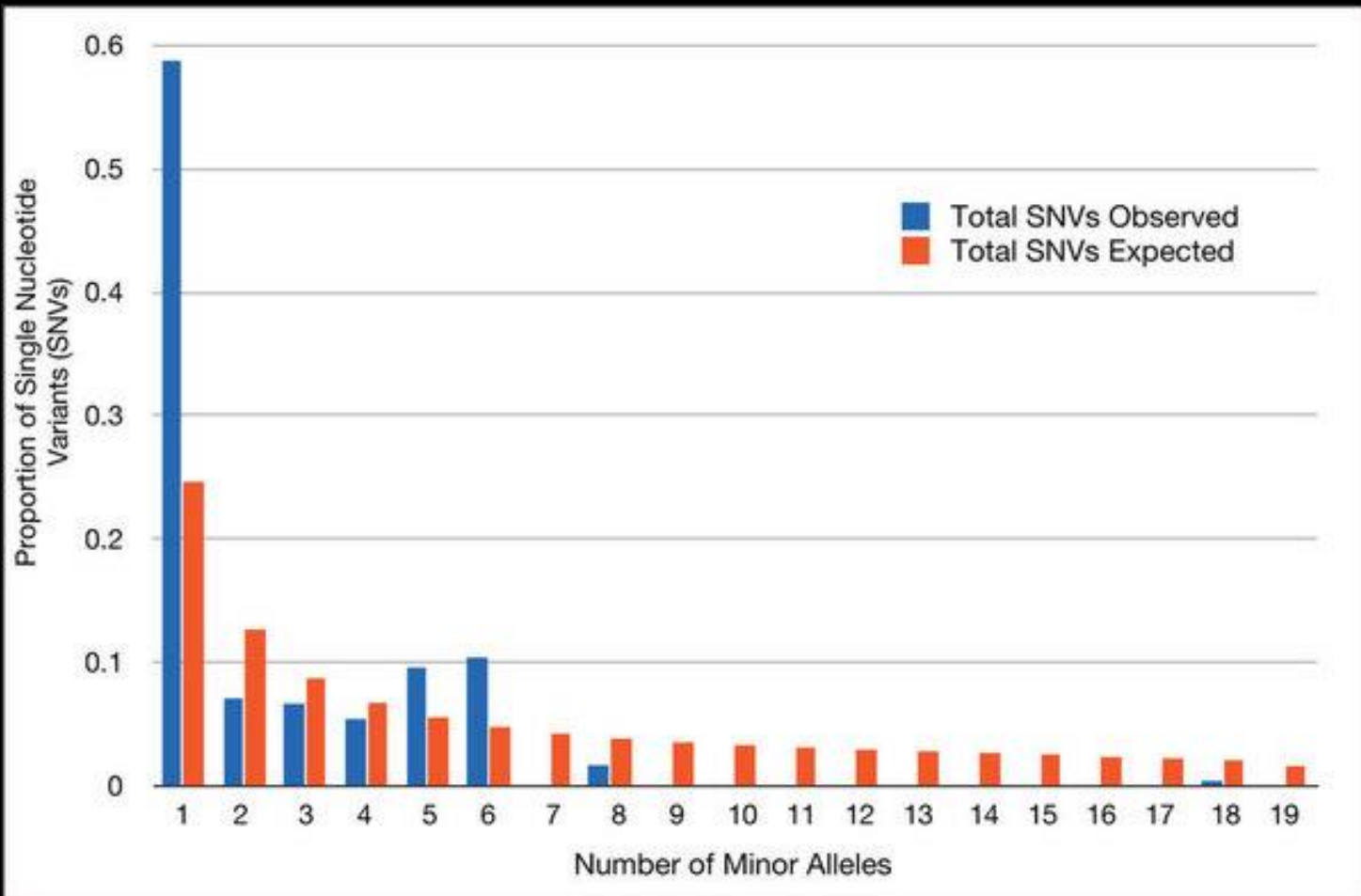
tripletons is  $\theta/3$

...

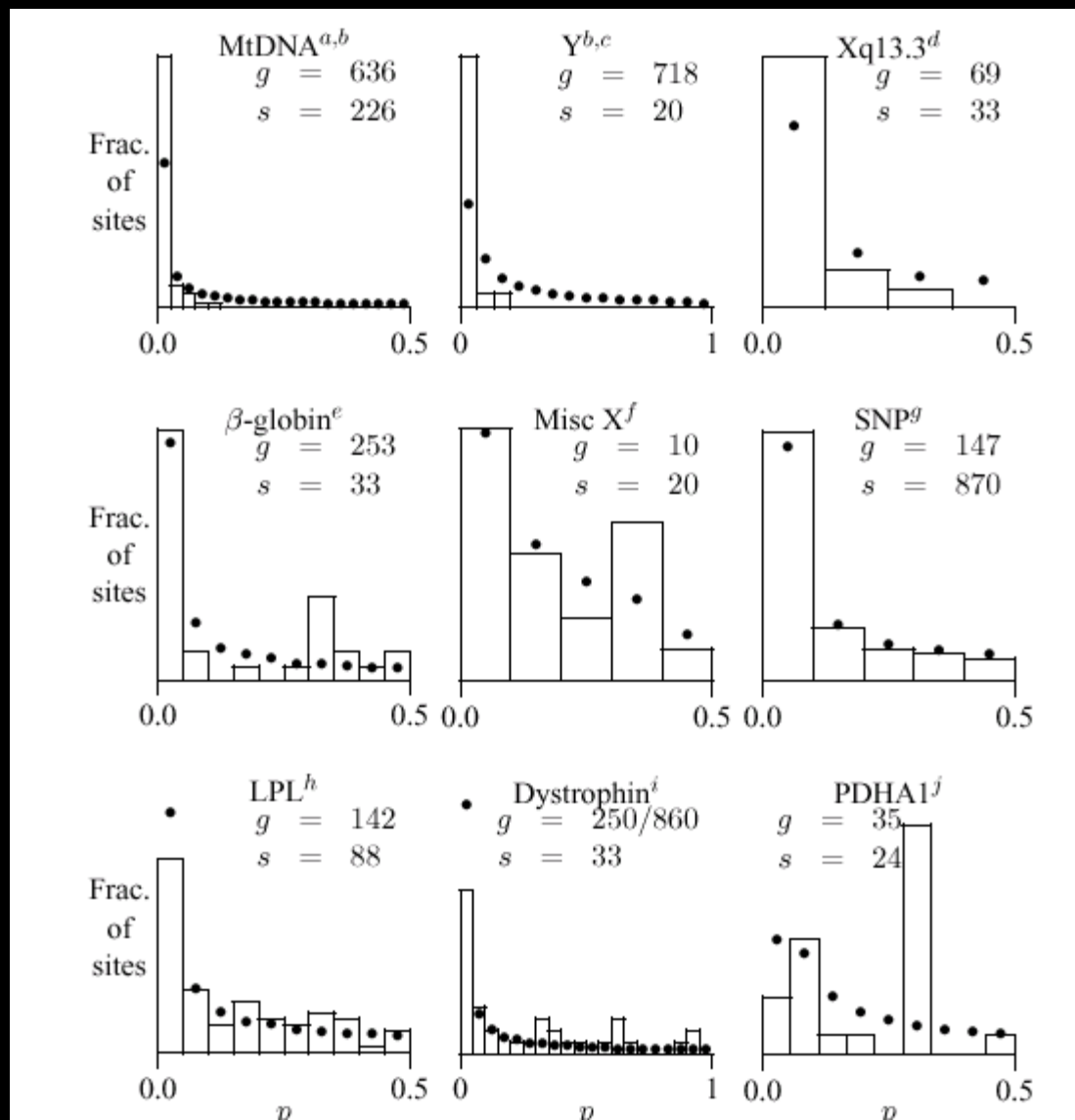
$n$ -pletons is  $\theta/n$

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

# Nucleotide site frequency spectrum



# Some observed human site frequency spectra





# Infinite-sites practical

- Run the simulations and solve the problems in the R program called: **InfiniteSites.r**

# Estimation of recombination ( $\rho = \rho = 4Nc$ )

- Direct estimation of recombination requires counts of meiotic exchanges.
- Even rather huge pedigrees have relatively few meiotic exchanges.
- This means that pedigree-based linkage maps are of limited resolution (10 cM or so).
- Fine-scale recombination is better inferred INDIRECTLY from LINKAGE DISEQUILIBRIUM.

# Linkage disequilibrium measures

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

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

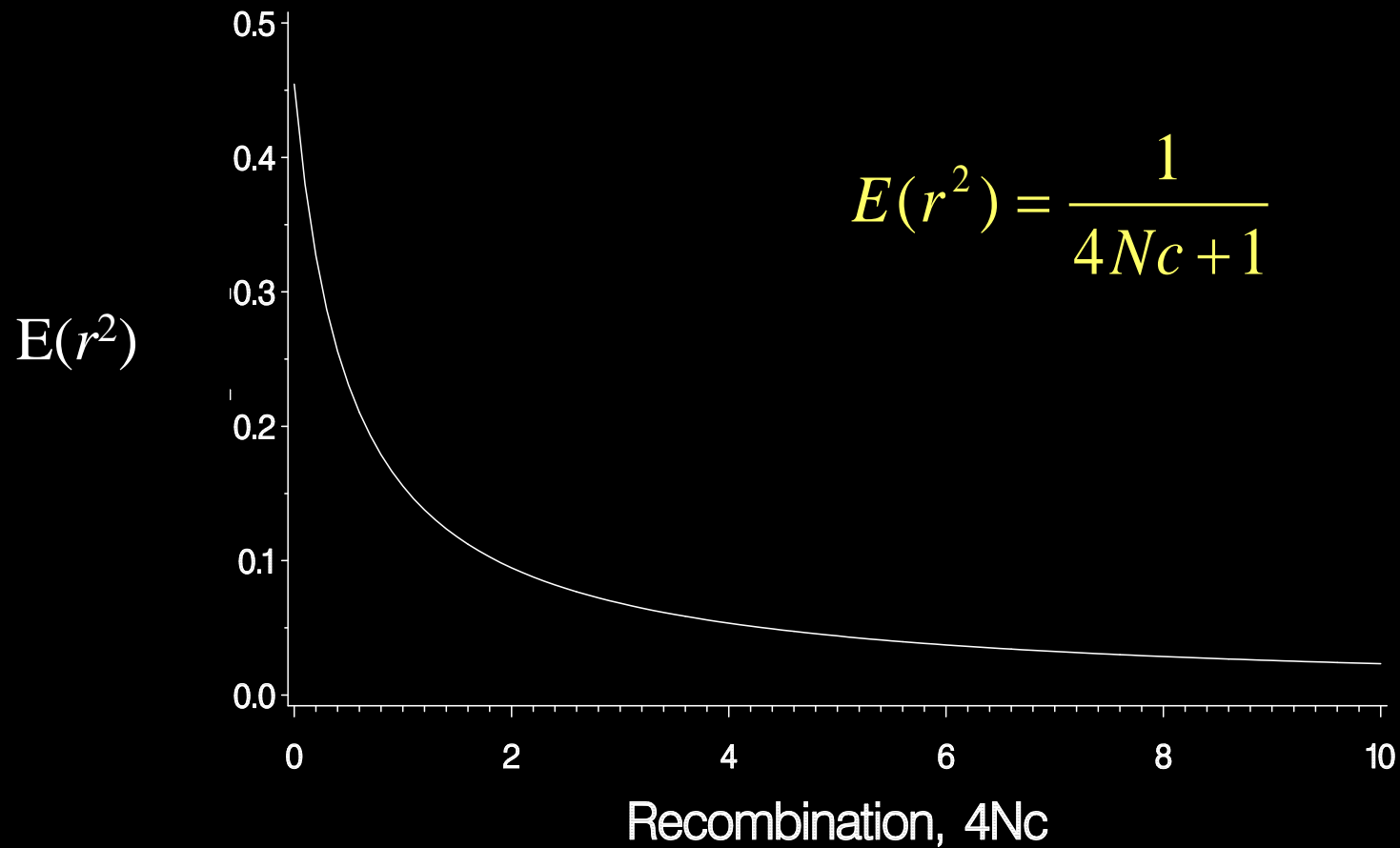
The maximum value  $D$  could ever have is if  $p_{AB} = p_{ab} = 1/2$ . When this is so,  $D = 1/4$ . Likewise the minimum is  $D = -1/4$ .

$D'$  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  $1$ .

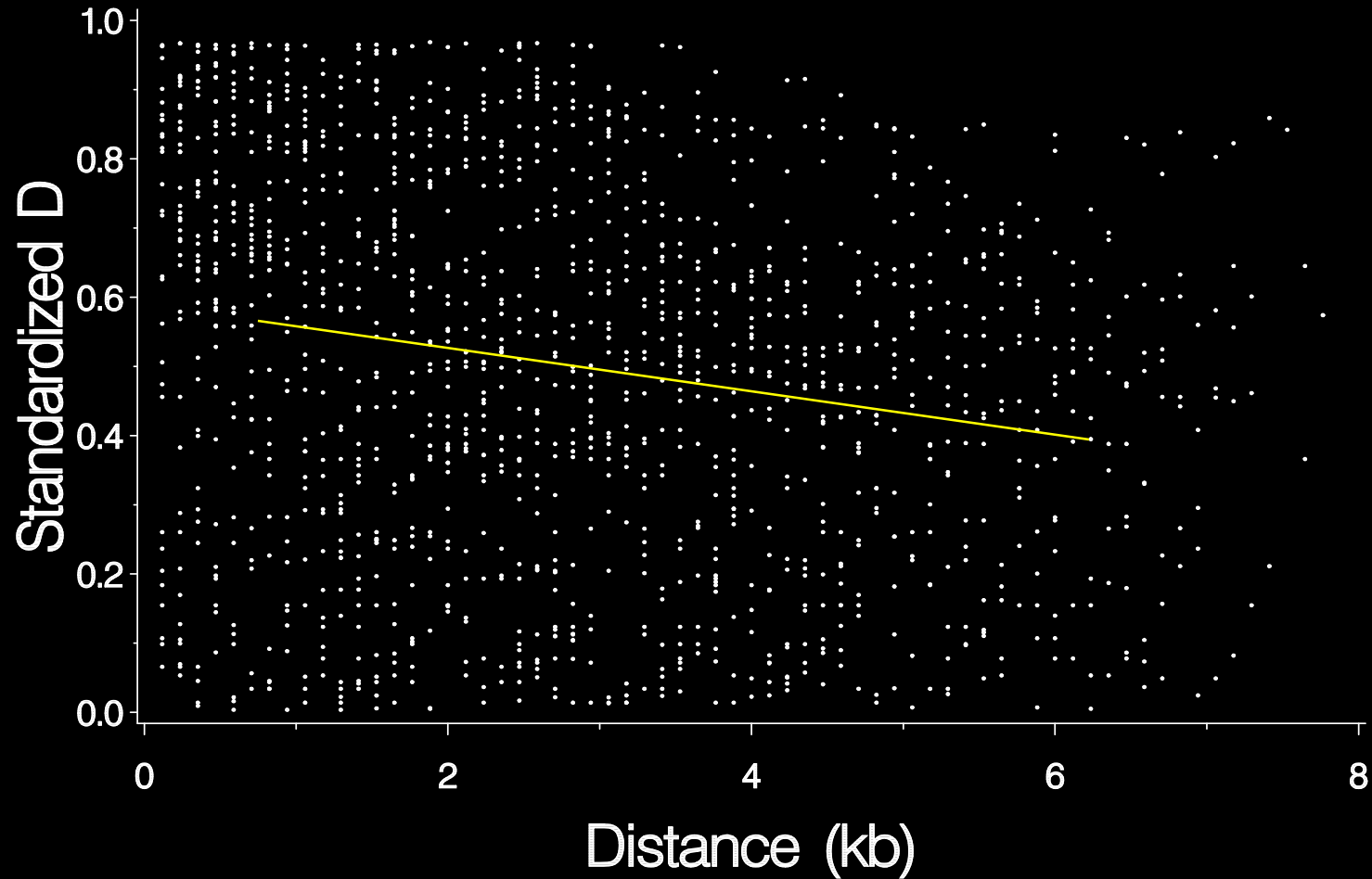
A third measure is the squared correlation coefficient:

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

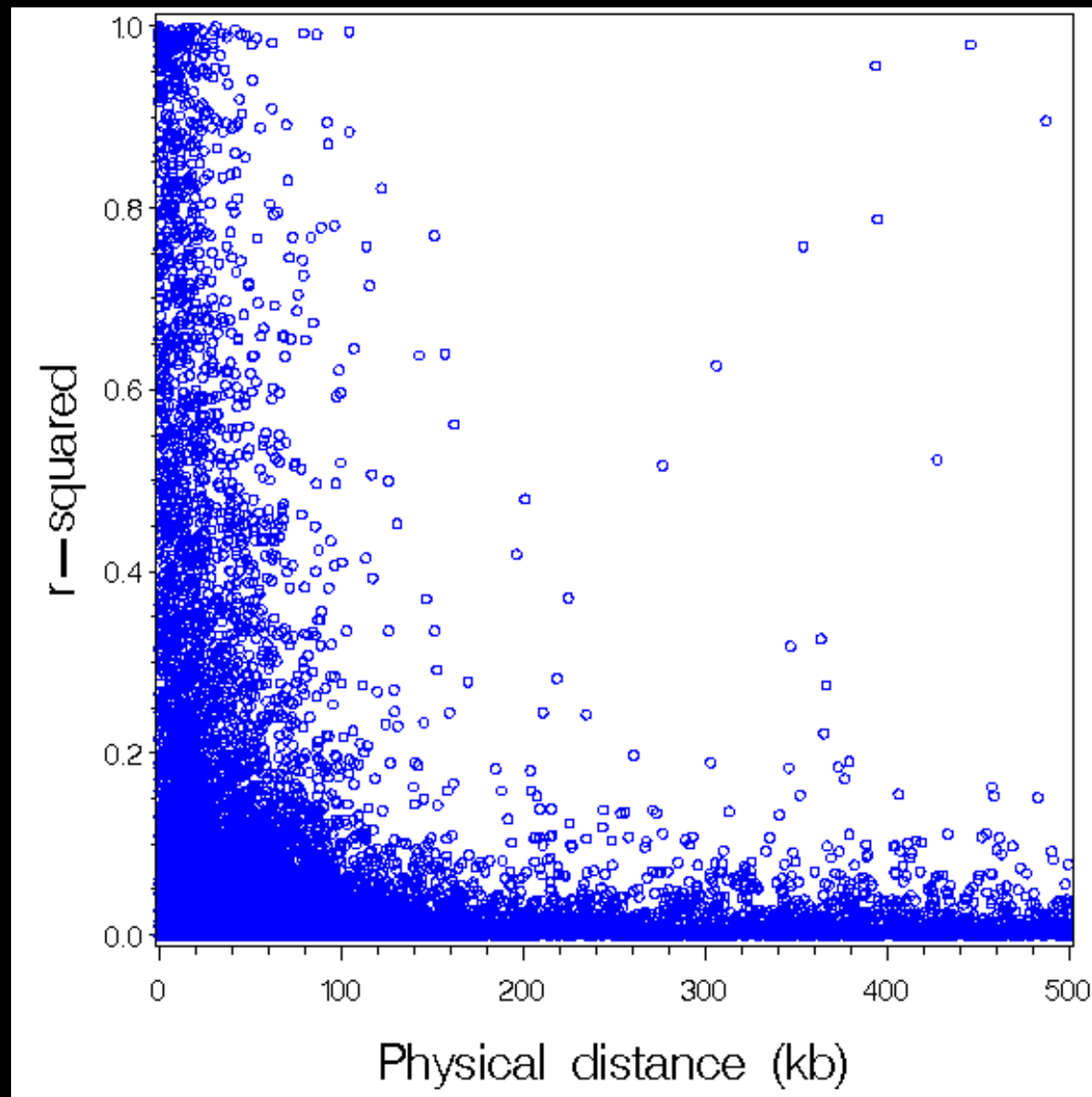
# Equilibrium relation between LD and recombination rate



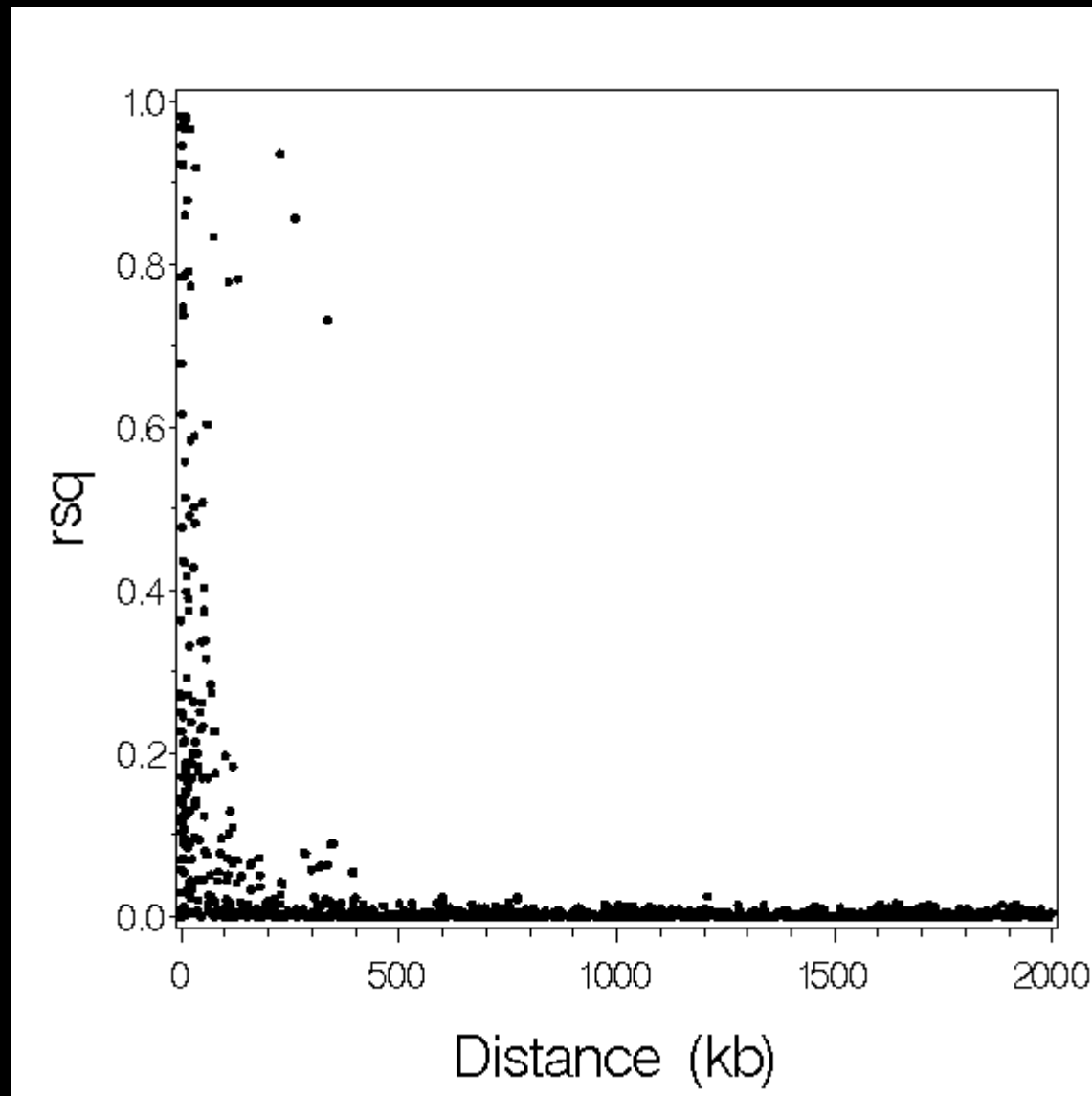
# Linkage disequilibrium decays across a gene



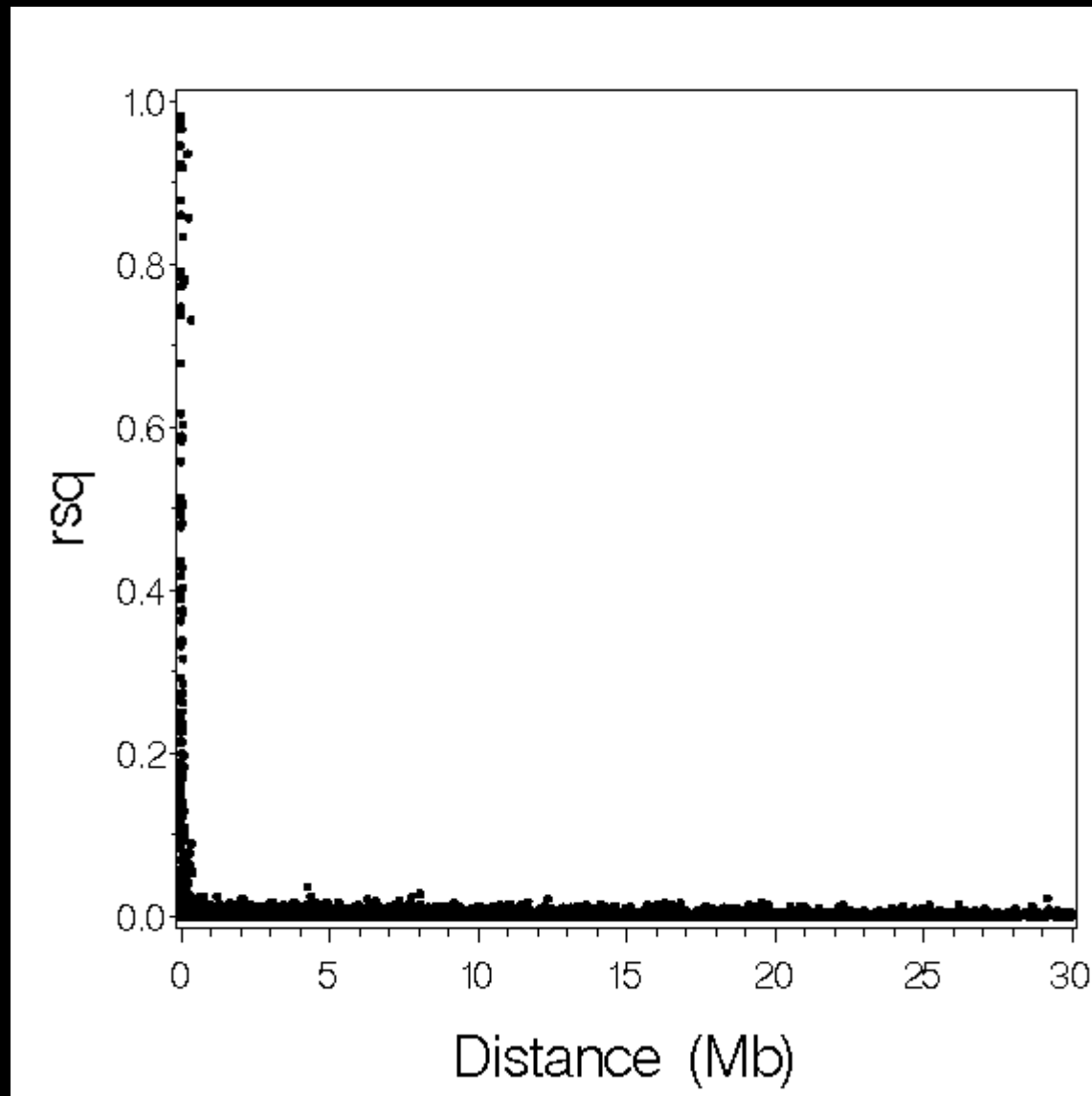
Linkage disequilibrium is rare beyond 100 kb or so



Beyond 500 kb, there is almost zero Linkage disequilibrium

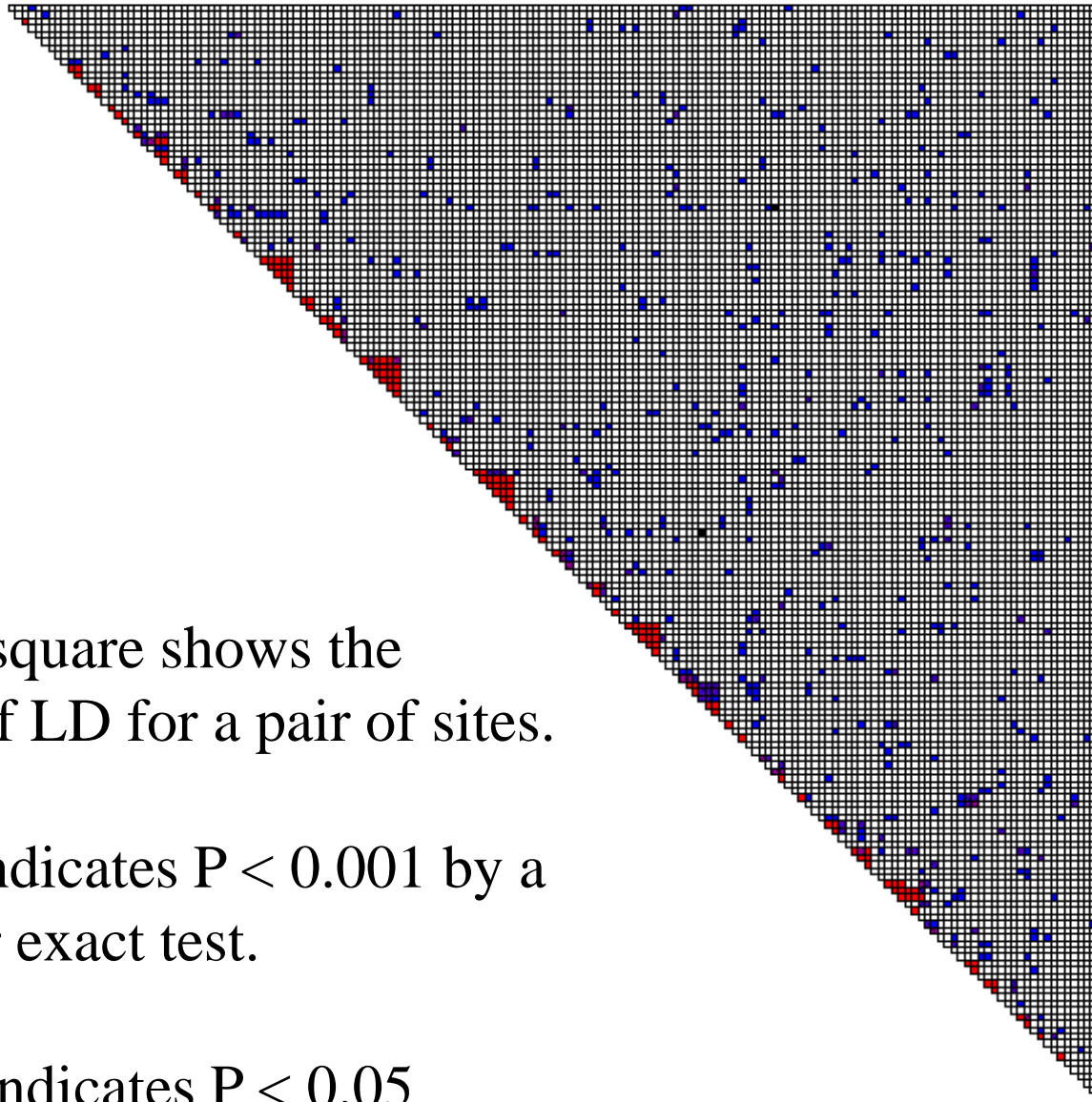


...so observing LD means the sites are likely to be close together





Patterns of LD can be examined by testing all pairs of sites

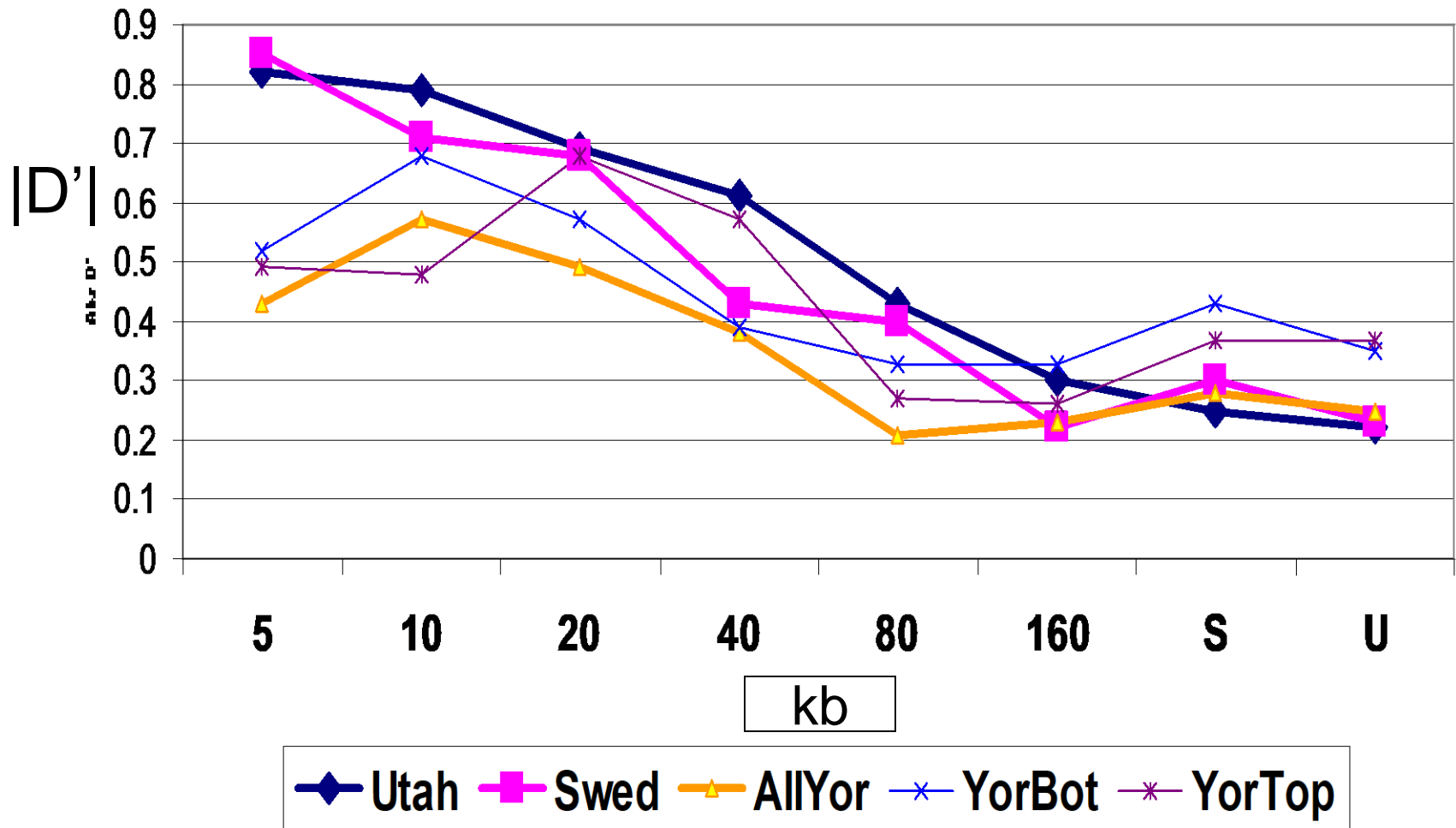


Each square shows the  
Test of LD for a pair of sites.

Red indicates  $P < 0.001$  by a  
Fisher exact test.

Blue indicates  $P < 0.05$

Rho can be estimated from the rate of decay of LD



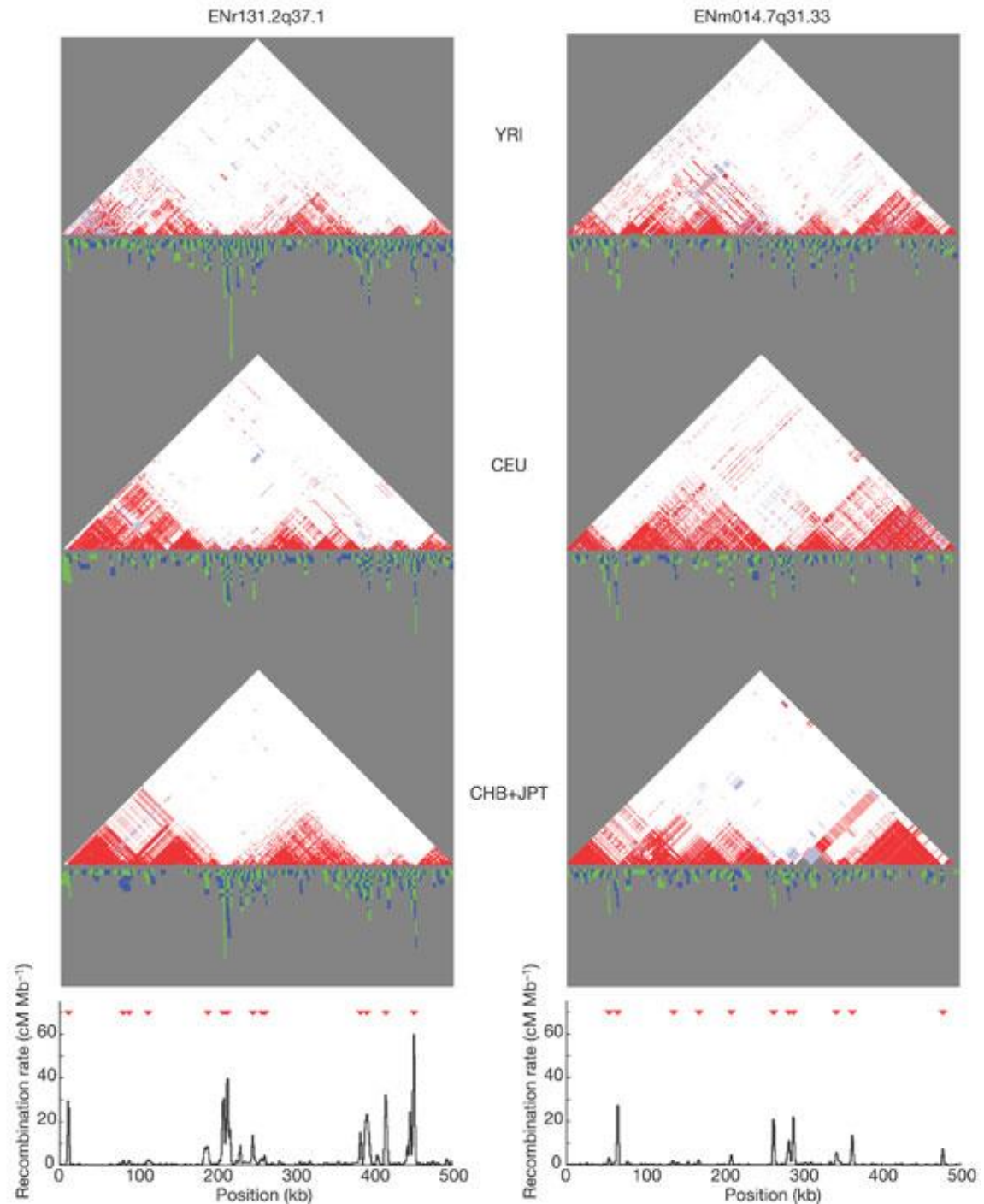
Reich et al. (2001 *Nature* 411:199-204)

# Terminology

- **Recombination Rate** refers to the proportion of recombinants between a pair of sites in a collection of gametes.
- **Recombination Intensity** refers to the recombination rate per unit of physical distance along the chromosome. It is expressed in cM/Mbp.
- **Population Recombination Rate ( $\rho$ )** is to recombination rate as  $\theta$  is to mutation rate.

Average  
recombination  
intensity in  
humans is  
1 cM/Mbp.

Hotspots are up  
to 100 cM/Mbp.



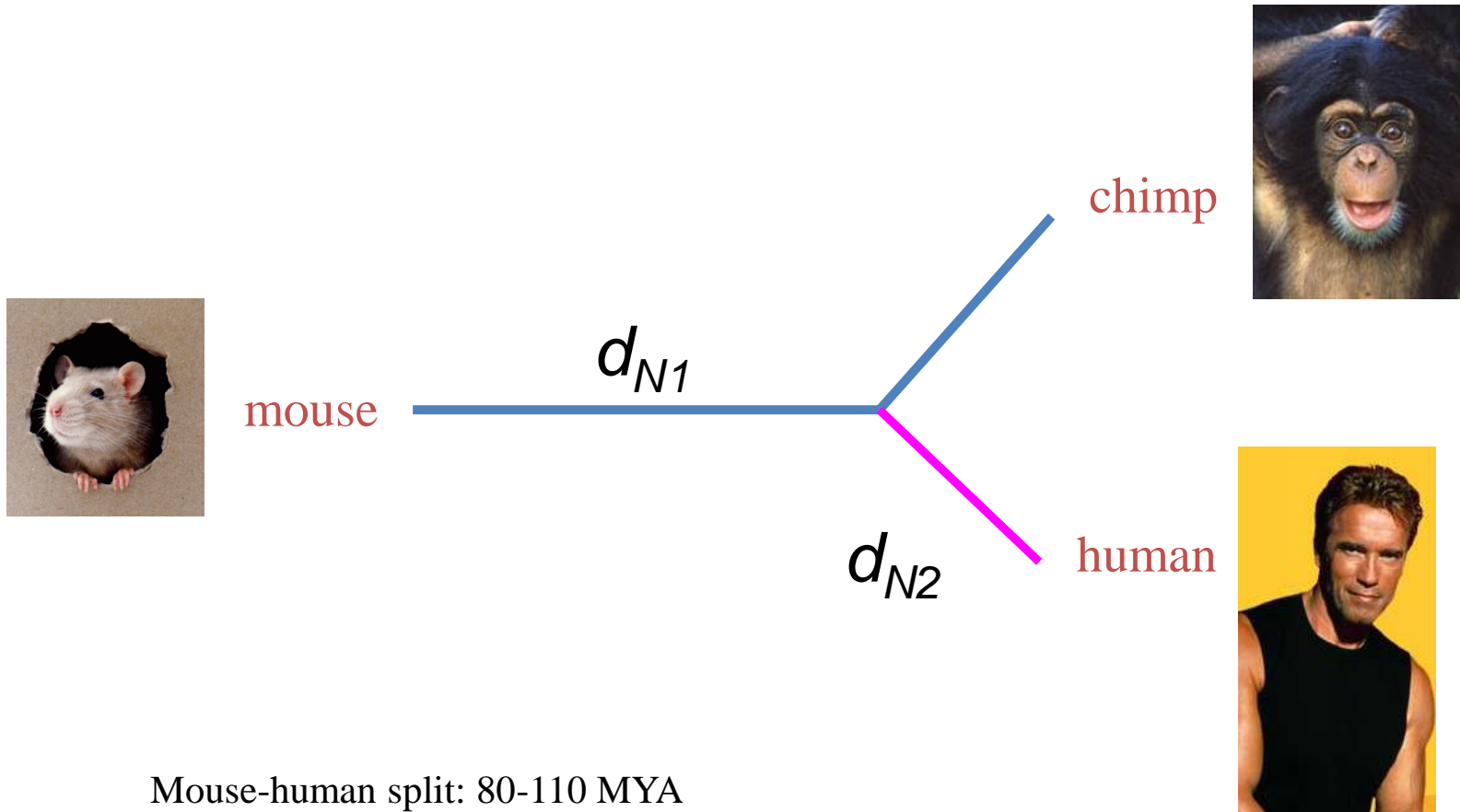
# LD practical

- Install the package “genetics” in R and use it to estimate pairwise LD for the first 20 SNPs in Olivier’s data set.

# Estimation of selection

- Polymorphism vs. Divergence (Aida: HKA, MK)
- Departure from expectation under a pure drift model
- Departure of SFS from neutral (Pascale: Tajima D)
- Excess IBD block size (Pascale: EHH)
- With allele frequency dynamics – departure from expected dynamics
- With detailed sampling: Selection Components Analysis

# First genome-wide selection test: $d_N/d_S$ test with PAML



Mouse-human split: 80-110 MYA  
Human-chimp split: 4.6 – 6 MYA

# Estimation of selection from allele frequency dynamics

- For each generation transition, record the allele frequency and the change.
- Wright-Fisher model implies binomial variance of allele frequency.
- Can formulate a likelihood ratio test (LRT) for goodness-of-fit of Wright-Fisher.
- The only parameter is  $N_e$ , which should be jointly estimated across all SNPs.



# Allele frequency dynamics practical

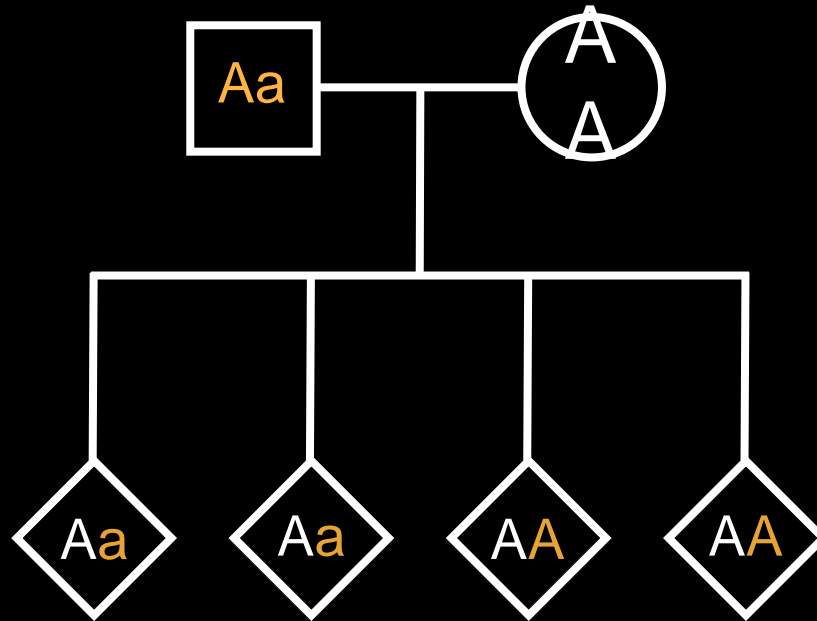
- Run the simulations and solve the problems in the R program: [AlleleFreqDyn.r](#)

# Selection Components Analysis

## SCA: Gametic selection

- $H_0$ : Heterozygotes produce equal proportions of the two gametes (typically scored in offspring).
- Test: Chi-square, or likelihood models.
- Parameter: explicit ML estimates of segregation probabilities.

Gametic selection: do heterozygotes transmit both alleles equally frequently?



$H_0$ : equal frequency of  $Aa$  &  $AA$  offspring

# A full likelihood approach for testing for gametic selection

Male genotype	Female genotype	# kids	# kids w/genotype		
			AA	Aa	aa
Aa	AA	$N_1$	$C_{11}$	$C_{12}$	0
Aa	aa	$N_2$	0	$C_{32}$	$C_{33}$
Aa	Aa	$N_3$	$C_{21}$	$C_{22}$	$C_{23}$
AA	Aa	$N_4$	$D_{11}$	$D_{12}$	0
aa	Aa	$N_5$	0	$D_{32}$	$D_{33}$

# A full likelihood approach for testing for gametic selection

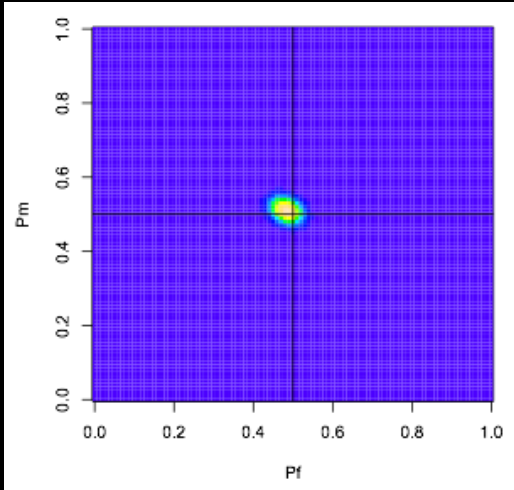
Male genotype	Female genotype	# kids	# kids w/genotype		
			AA	Aa	aa
Aa	AA	$N_1$	$C_{11}$	$C_{12}$	0
Aa	aa	$N_2$	0	$C_{32}$	$C_{33}$
Aa	Aa	$N_3$	$C_{21}$	$C_{22}$	$C_{23}$
AA	Aa	$N_4$	$D_{11}$	$D_{12}$	0
aa	Aa	$N_5$	0	$D_{32}$	$D_{33}$

$p_m$  = probability male transmits allele  $A$

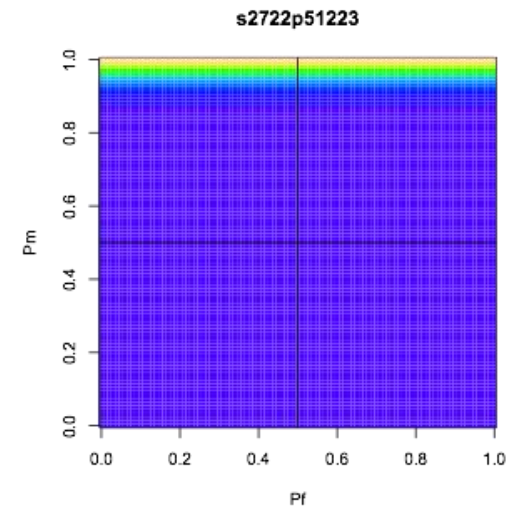
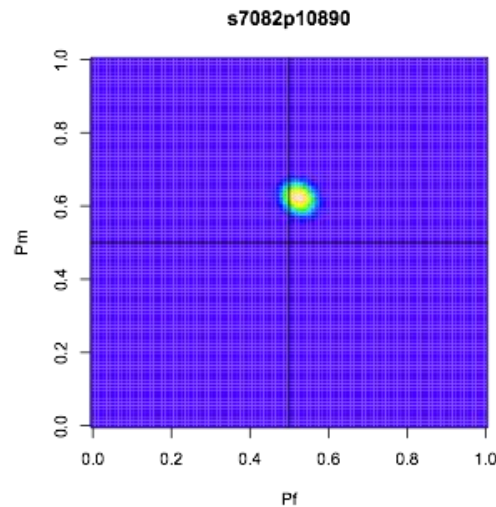
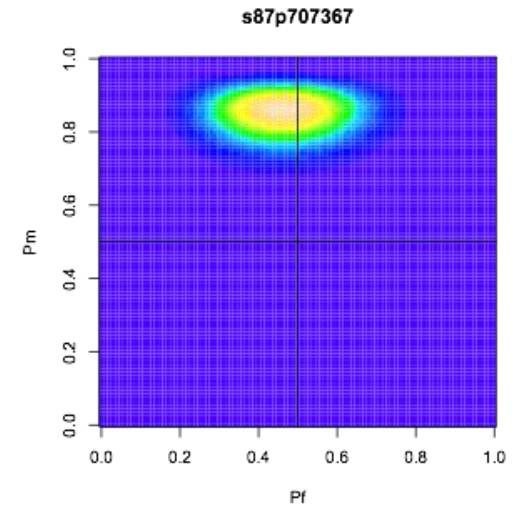
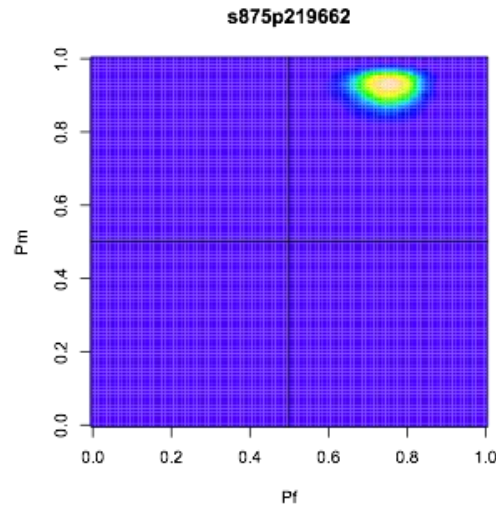
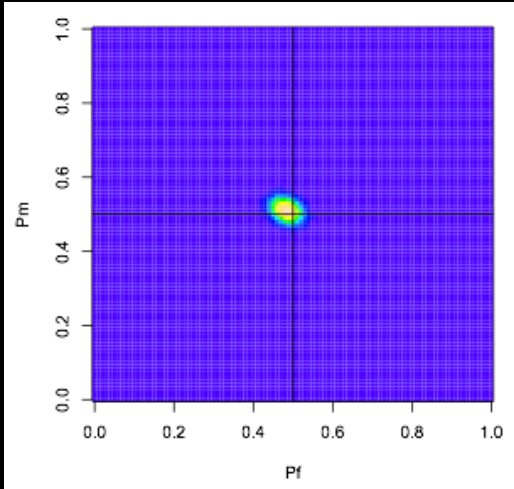
$p_f$  = probability female transmits allele  $A$

$$\begin{aligned}
 L = & \binom{N_1}{C_{11}} p_m^{C_{11}} (1 - p_m)^{C_{12}} \cdot \binom{N_2}{C_{32}} p_m^{C_{32}} (1 - p_m)^{C_{33}} \cdot \\
 & \binom{N_3}{C_{21} + C_{22}} (p_m p_f)^{C_{21}} [p_m(1 - p_f) + (1 - p_m)p_f]^{C_{22}} [(1 - p_m)(1 - p_f)]^{C_{23}} \cdot \\
 & \binom{N_4}{D_{11}} p_f^{D_{11}} (1 - p_f)^{D_{12}} \cdot \binom{N_5}{D_{32}} p_f^{D_{32}} (1 - p_f)^{D_{33}}
 \end{aligned}$$

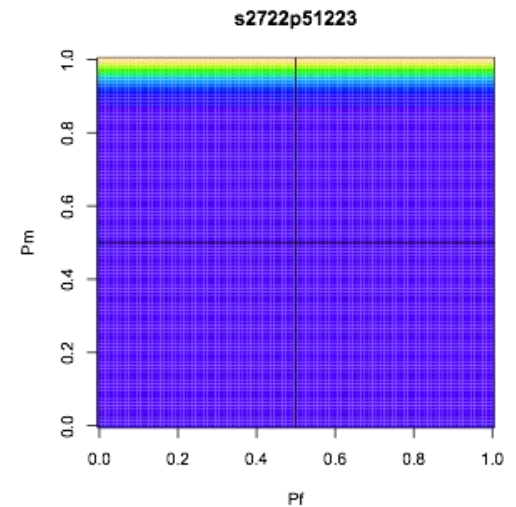
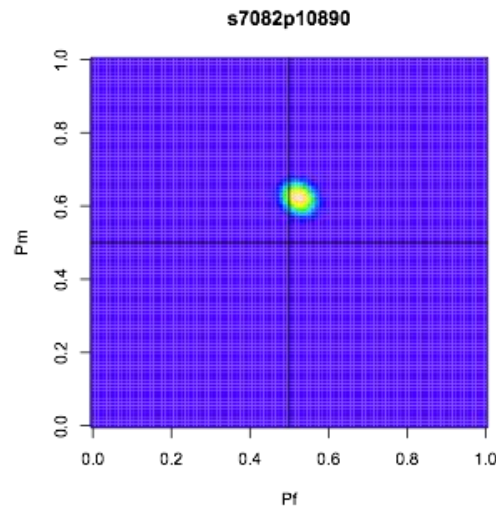
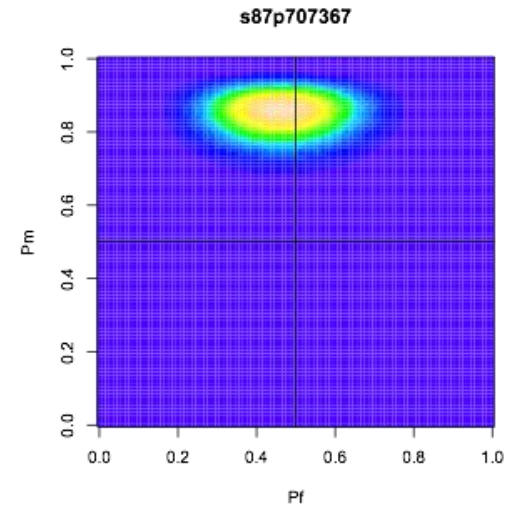
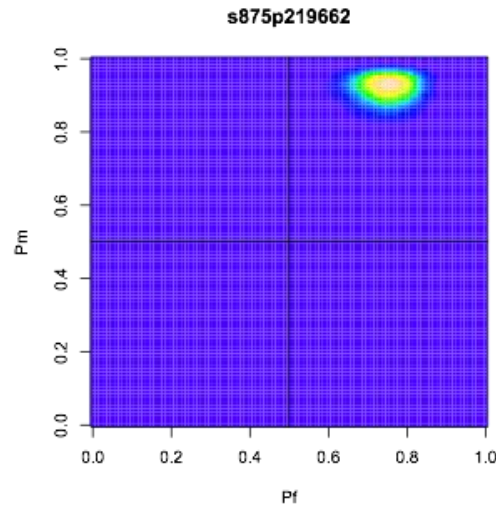
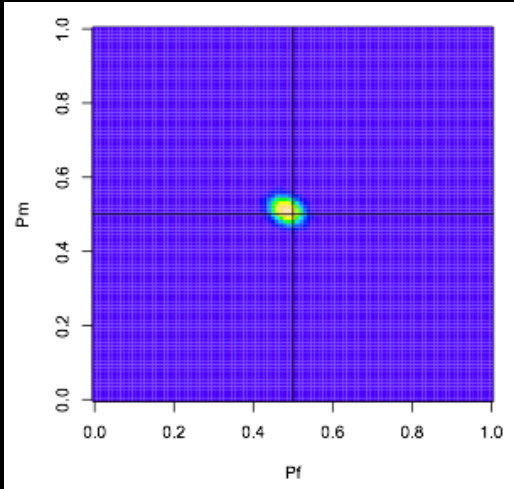
# A full likelihood approach for testing for gametic selection



# A full likelihood approach for testing for gametic selection



# A full likelihood approach for testing for gametic selection



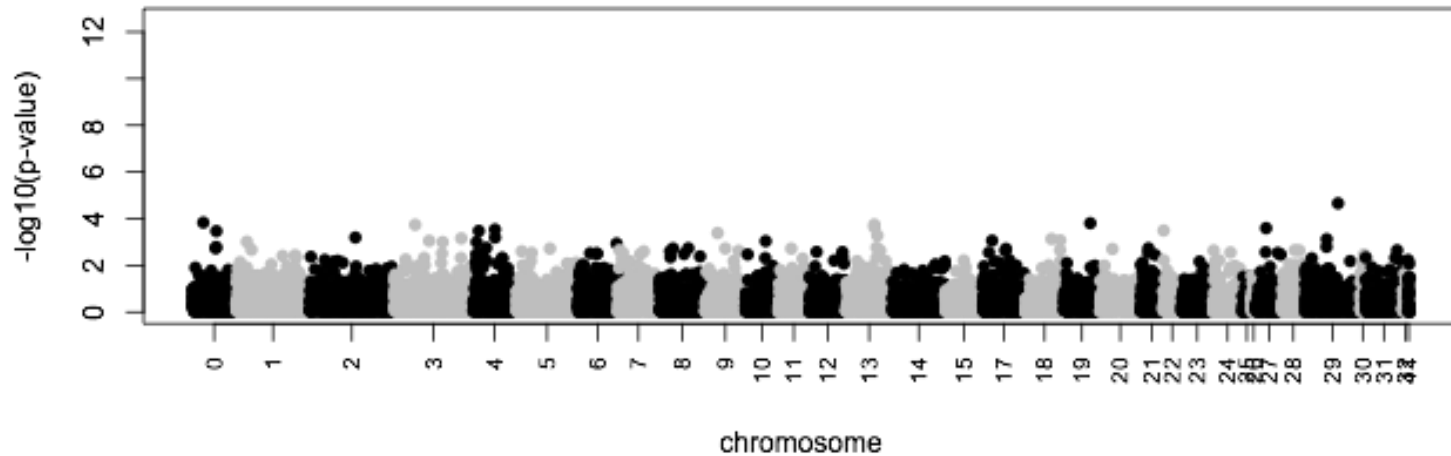
$$LRT_f = -2 \ln \frac{L(0.5, p_{mMax})}{L(p_{fMax}, p_{mMax})}$$

$$LRT_m = -2 \ln \frac{L(p_{fMax}, 0.5)}{L(p_{fMax}, p_{mMax})}$$

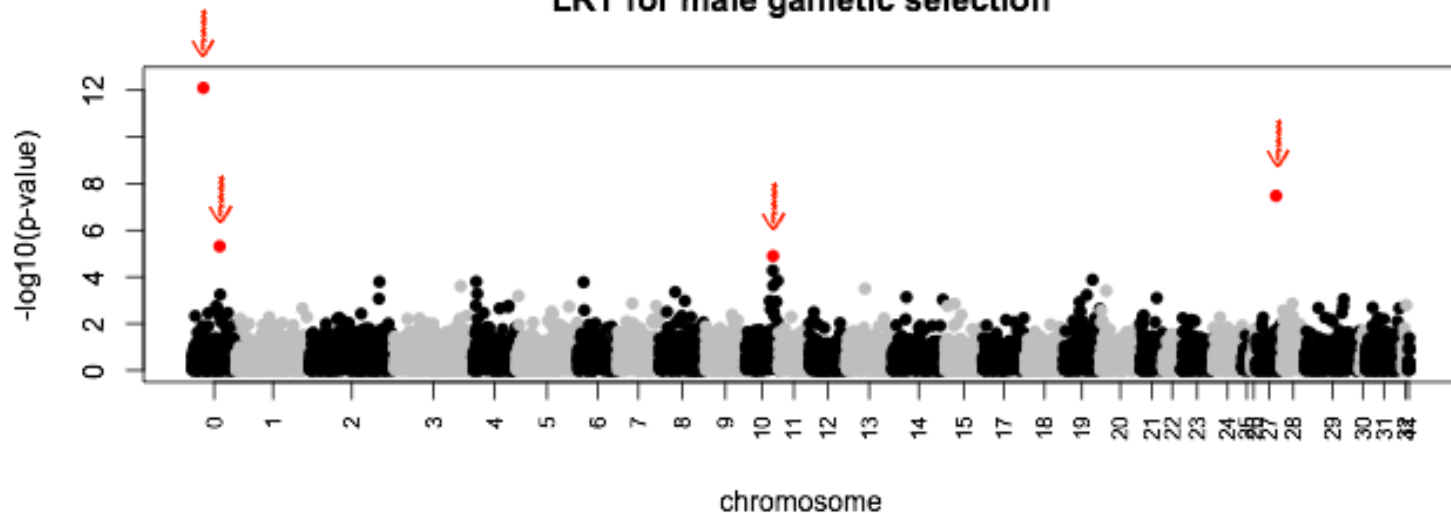


# Four male gametic selection hits

LRT for female gametic selection



LRT for male gametic selection

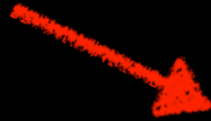


# SCA: Viability

- $H_0$ : All genotypes have the same probability of surviving from zygote to reproductive adult.
- (A bit more subtle – all genotypes have the same survivorship curve)
- Test: General linear models, accounting for other factors.
- Parameters for predictive models are standard fitnesses inferred from the linear models.

# General linear model for viability and fecundity

survival  
breeding status  
clutch size

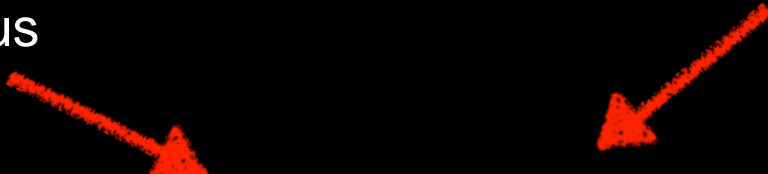


$$y = X\beta + Zu + e$$

# General linear model for viability and fecundity

survival  
breeding status  
clutch size

individual ID  
year  
natal nest  
parent ID


$$y = X\beta + Zu + e$$

# General linear model for viability and fecundity

survival  
breeding status  
clutch size

individual ID  
year  
natal nest  
parent ID

$$y = X\beta + Zu + e$$

age  
inbreeding coefficient  
hatch date  
nestling  
morphometrics  
juvenile morphometrics  
new breeder?  
immigrant status  
territory acquisition

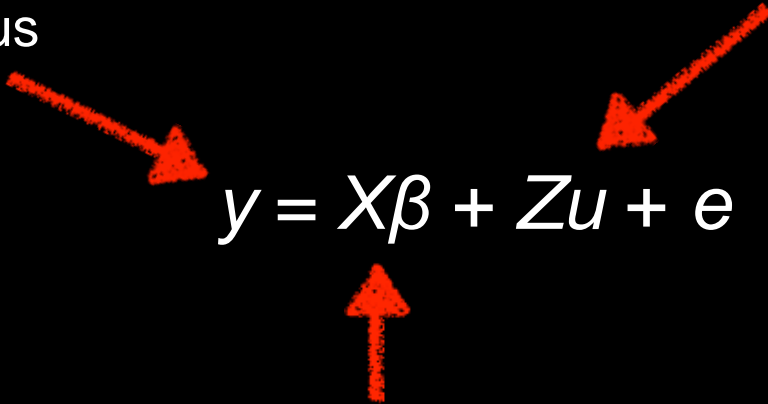
# helpers  
territory size  
time since fire  
habitat composition  
pair experience  
female breeder age  
male breeder age  
incubation date

density  
breeding vacancies  
rainfall  
drought index  
temperature  
acorn abundance  
etc.

# General linear model for viability and fecundity

survival  
breeding status  
clutch size

individual ID  
year  
natal nest  
parent ID


$$y = X\beta + Zu + e$$

age  
inbreeding coefficient  
hatch date  
nestling  
morphometrics  
juvenile morphometrics  
new breeder?  
immigrant status  
territory acquisition

# helpers  
territory size  
time since fire  
habitat composition  
pair experience  
female breeder age  
male breeder age  
incubation date

density  
breeding vacancies  
rainfall  
drought index  
temperature  
acorn abundance  
etc.  
SNP genotype

# SCA: Fecundity

- $H_0$ : All genotypes produce the same number of offspring
- Test: General linear models, accounting for non-genetic influences, such as food availability, pathogen load, etc.
- Parameters for predictive models are typically inferred from linear models (assuming no G x E)

# SCA: Sexual selection

- $H_0$ : All genotypes that survive to adulthood have equal chance of becoming a reproductive adult.
- Requires records of lifetime survival and reproduction.
- Test: General linear models, accounting for non-genetic influences, such as local density, food availability, etc.
- Parameters for predictive models are typically inferred from linear models (assuming no G x E)



# Forces shaping genetic diversity

- Mutation
- Random Genetic Drift
- Recombination
- Migration
- Natural Selection