Quantitative trait evolution with mutations of large effect

Joshua G. Schraiber and Michael J. Landis

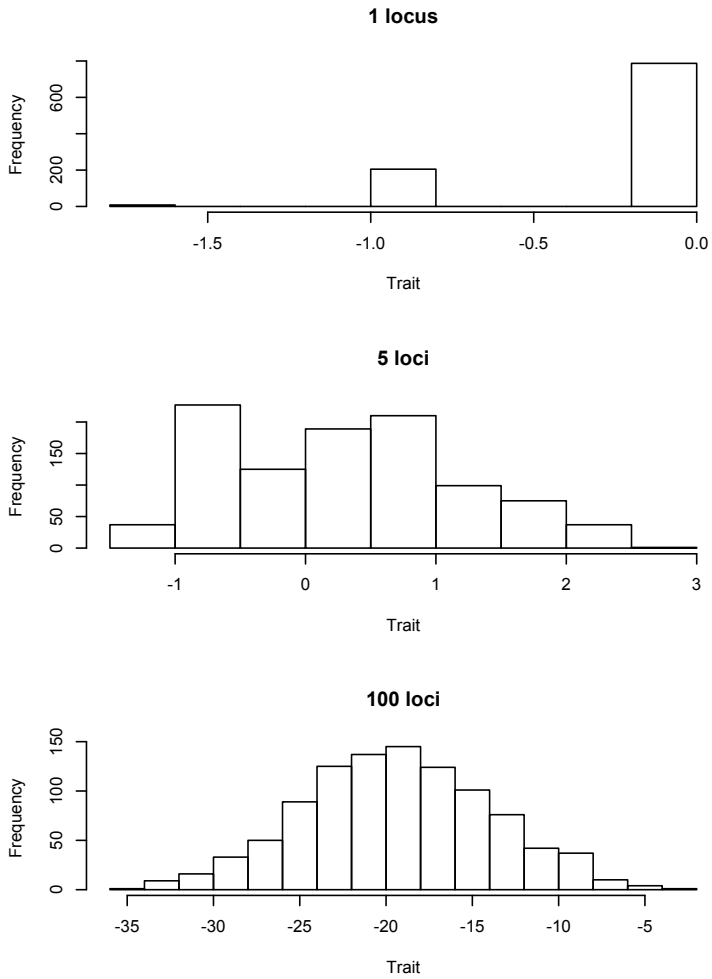
May 1, 2014

Quantitative traits

- Traits that vary continuously in populations
  - Mass
  - Height
  - Bristle number (approx)
- Adaption
  - Low oxygen tolerance
- Disease
  - Obesity
- Agriculture
  - Fruit yield

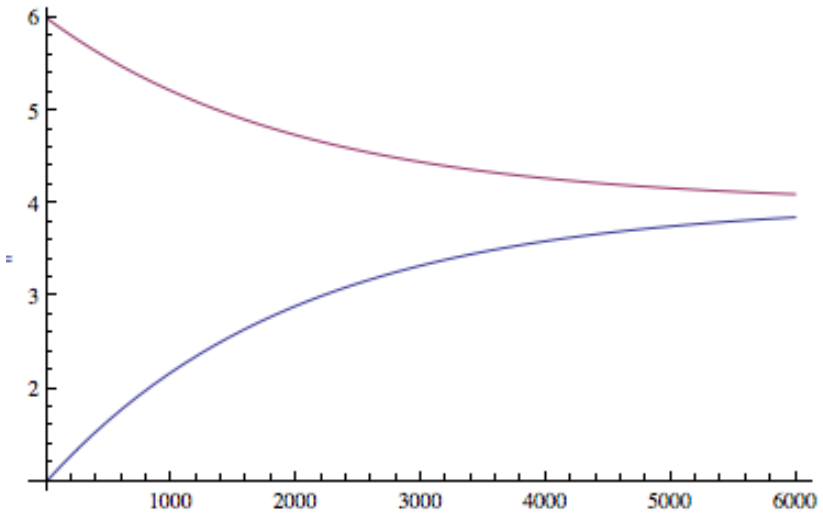
Fisher's quantitative trait model

- n biallelic loci impacting trait
- Allele j at locus i has effect a_{ij}
- Phenotype is $X_k = \sum_{i,j} a_{ij} z_{ijk}$
  - z_{ijk} is number of copies of allele j at locus i in individual k
- $\mathbb{E}(\bar{X}) = 2n\mathbb{E}(p)\mathbb{E}(a)$
- $\mathbb{E}(\mathrm{Var}(X)) = 2n\mathbb{E}(p(1-p))\mathbb{E}(a^2)$
- Environmental variation induces extra variability



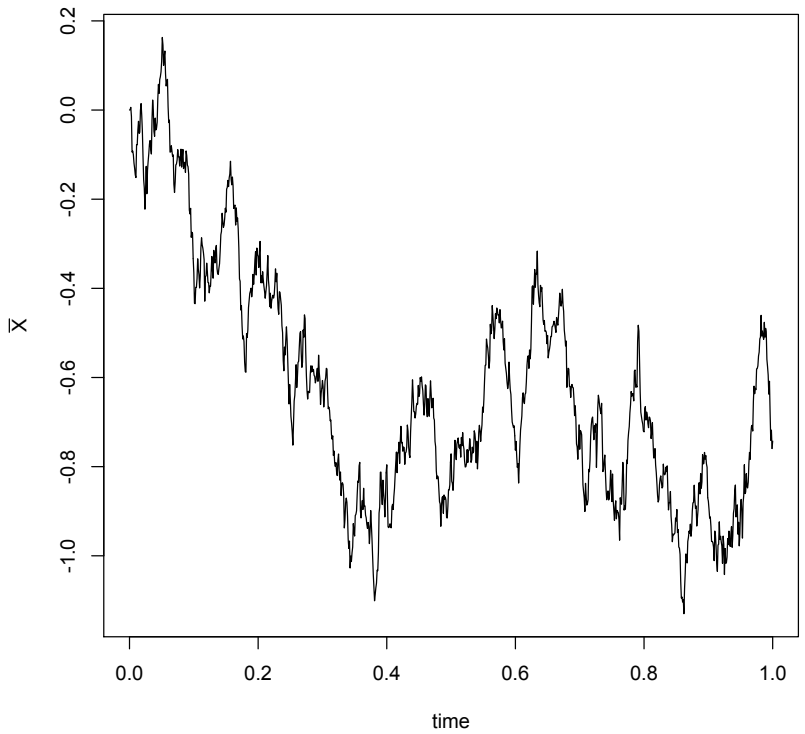
Neutral genetic variance

- Equilibrium genetic variance
  - $\mathbb{E}(V_A(\infty)) = 2NV_m$
- V_m is the additive variance of new mutations
  - $V_m \approx 2\mu\sigma^2$
  - μ: is trait-wide mutation rate
  - σ^2: variance of mutant effect size distribution
- $\mathrm{Var}(V_A(\infty)) \approx 4N\mu\sigma^4$ (Lynch and Hill 1986)

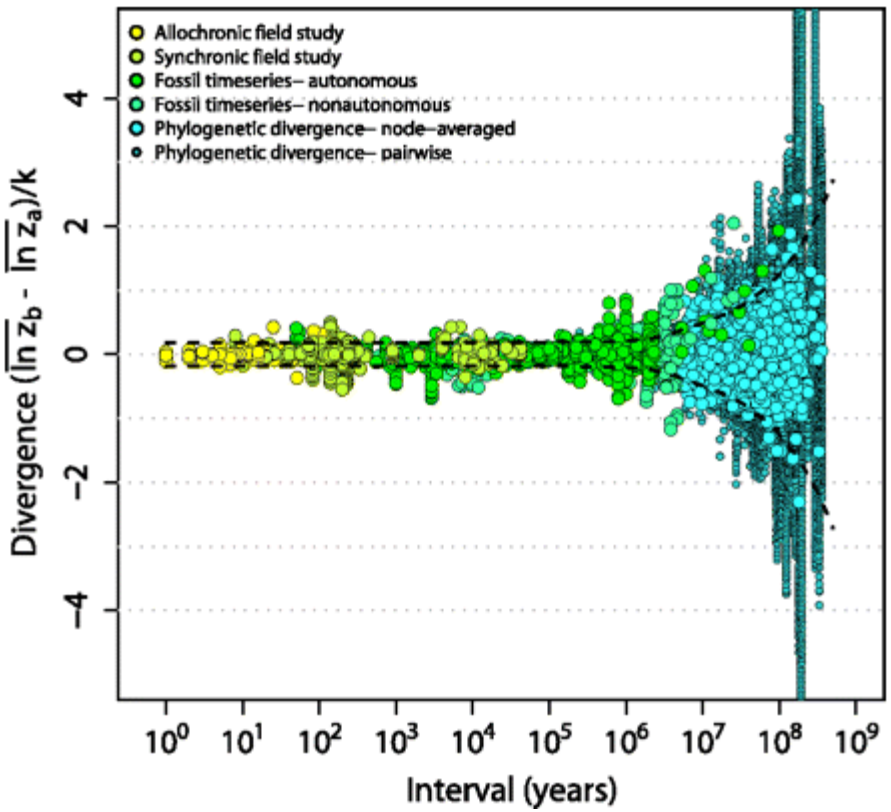


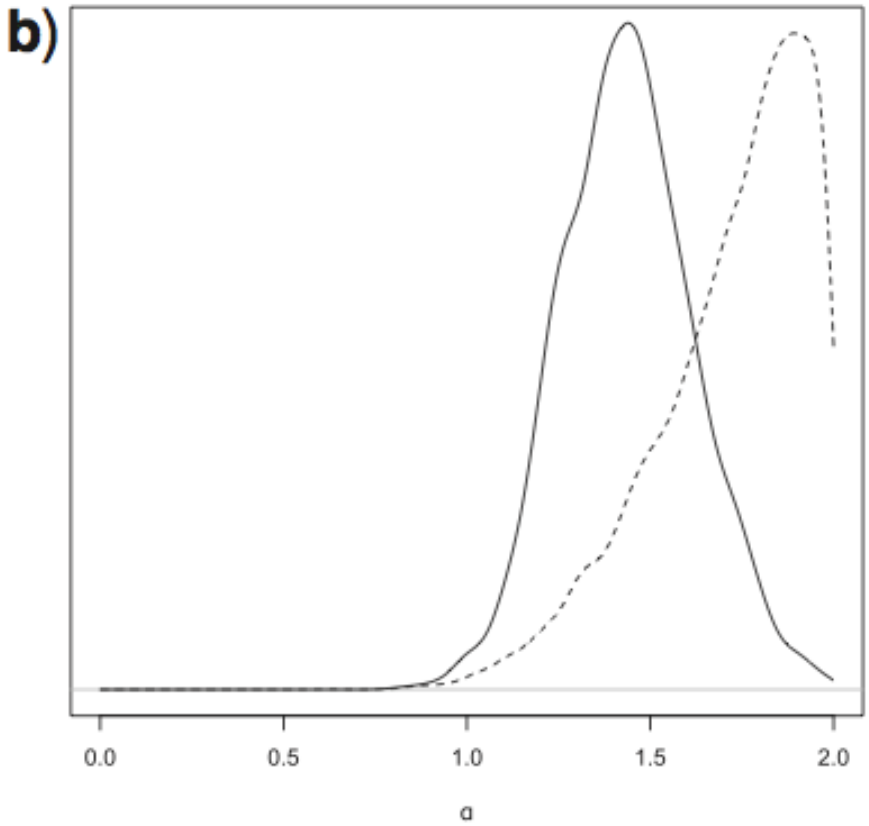
Lande's Brownian motion model

- Approximation to full model
  - Don't keep track of individual loci
  - Assume genetic variance constant in time
  - Mutations have small effect sizes
- Can incorporate selection via Ornstein-Uhlenbeck model
- Extremely influential in comparative biology
  - c.f. Felsenstein: independent contrasts



Evidence for non-Brownian evolution

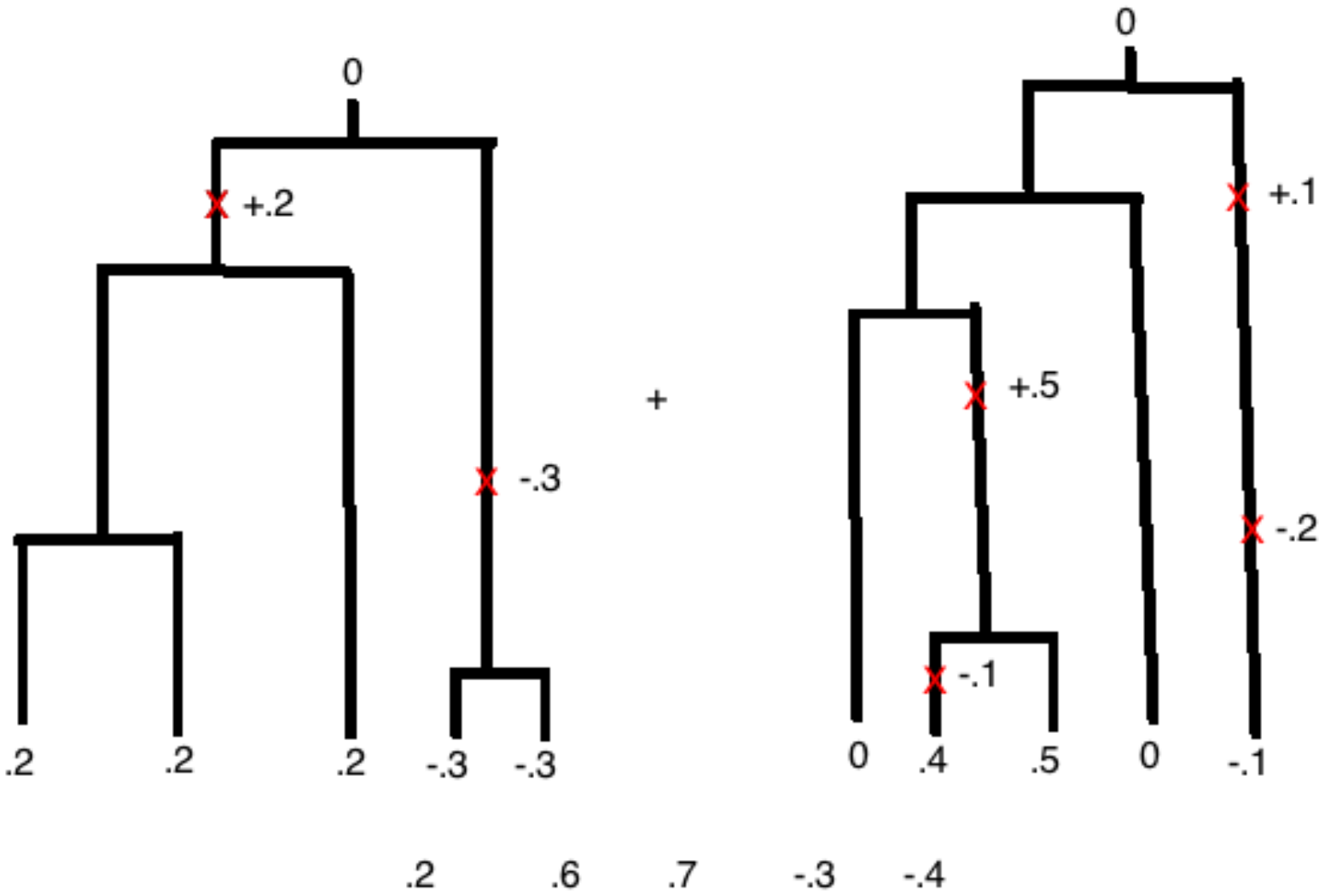




A coalescent model

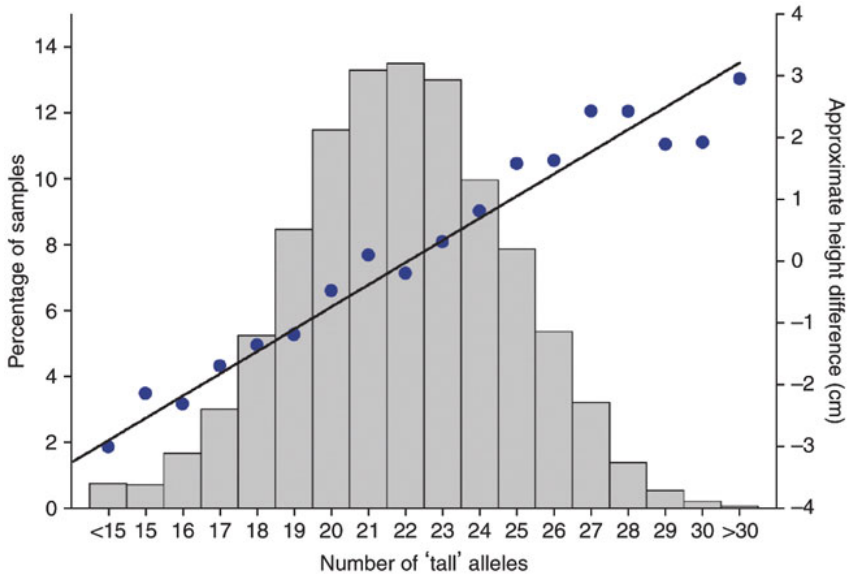
- Haploid population of size N
- Trait governed by n loci
- Each locus has an independent coalescent tree
- Each locus has mutation rate $\frac{\theta}{2}$ (coalescent time units)
- When a mutation happens, effect Y is drawn from distribution with density $p(y)$.

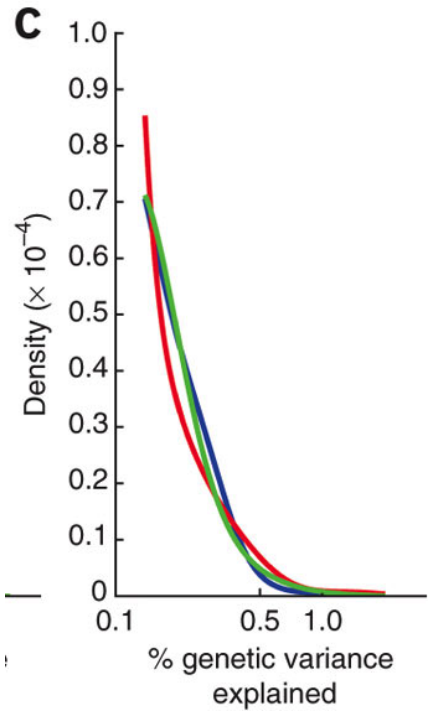
Example



Mutational effects

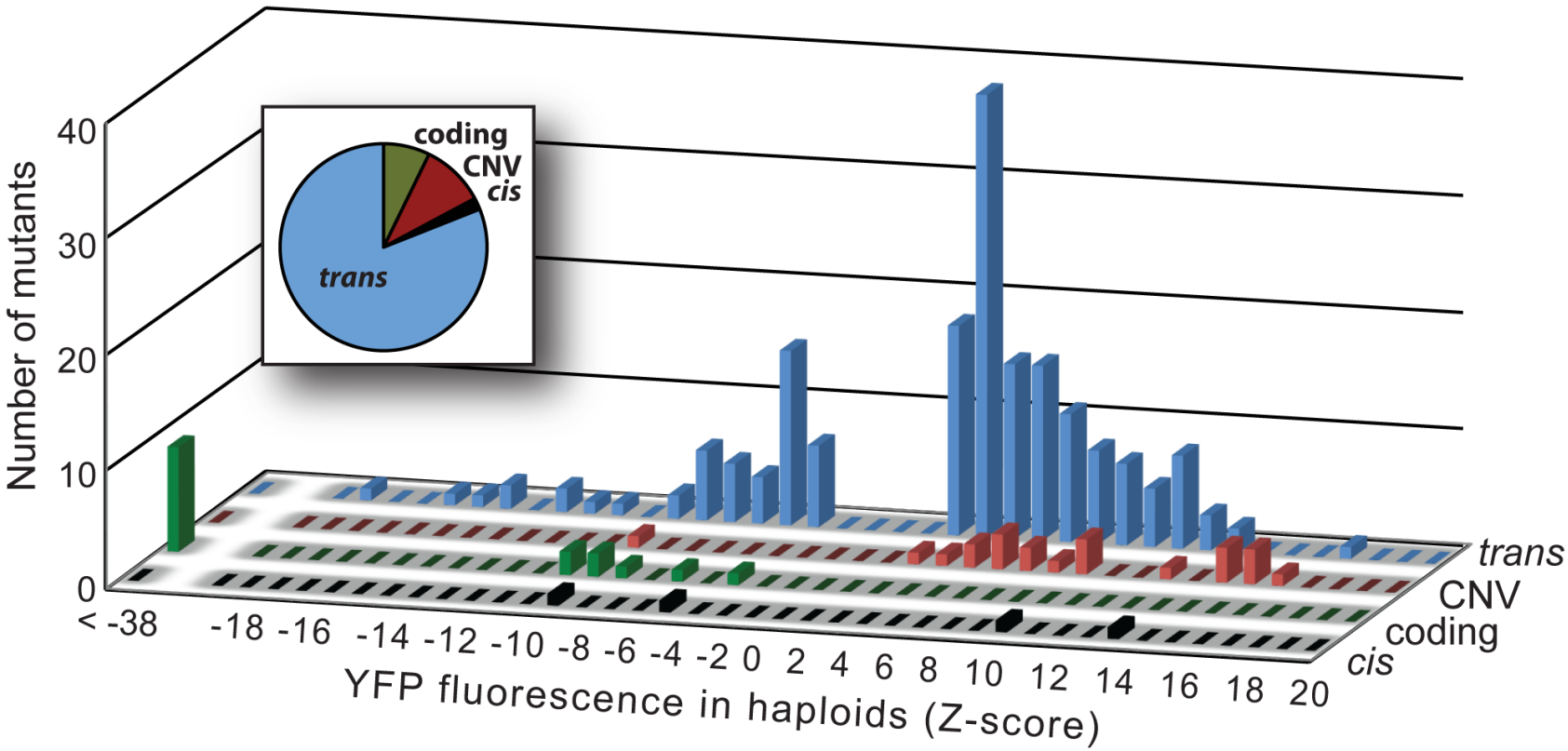
- Common assumption: many loci of very small effect
  - Infinitesimal model





Large effect mutations

- In some cases, large effect mutations may occur
  - Transcription factor binding sites
  - Null mutants upstream in pathways



Characteristic functions

Characteristic function

For a random variable X drawn from the probability measure $\mu(\cdot)$, the function

$$\begin{aligned}\phi_X(k) &= \mathbb{E}(e^{ikX}) \\ &= \int e^{ikx} d\mu(x)\end{aligned}$$

is called the characteristic function

Sums of random variables

If $X_1, \ldots, X_n$ are independent random variables, and $X = \sum_i X_i$ then

$$\phi_X(k) = \prod_i \phi_{X_i}(k)$$

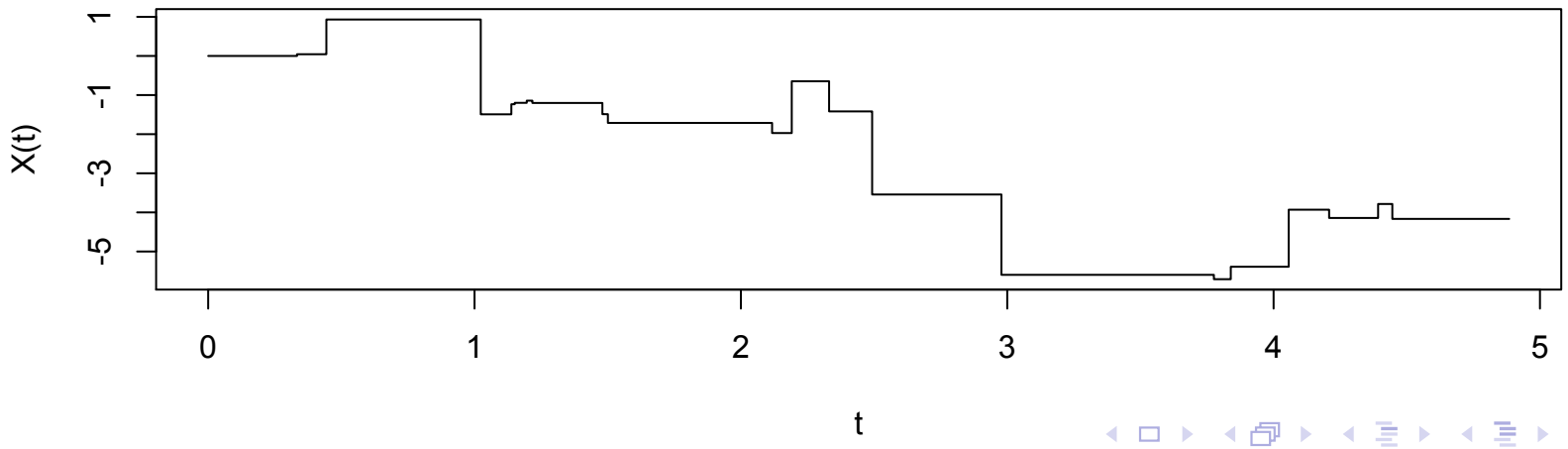
Compound Poisson process

Definition

Let $N(t)$ be a rate λ Poisson process, and $(Y_i, i \geq 1)$ a sequence of independent and identically distributed random variables. Then

$$X(t) = \sum_{i=1}^{N(t)} Y_i$$

is called a compound Poisson process.



Characteristic function of a compound Poisson process

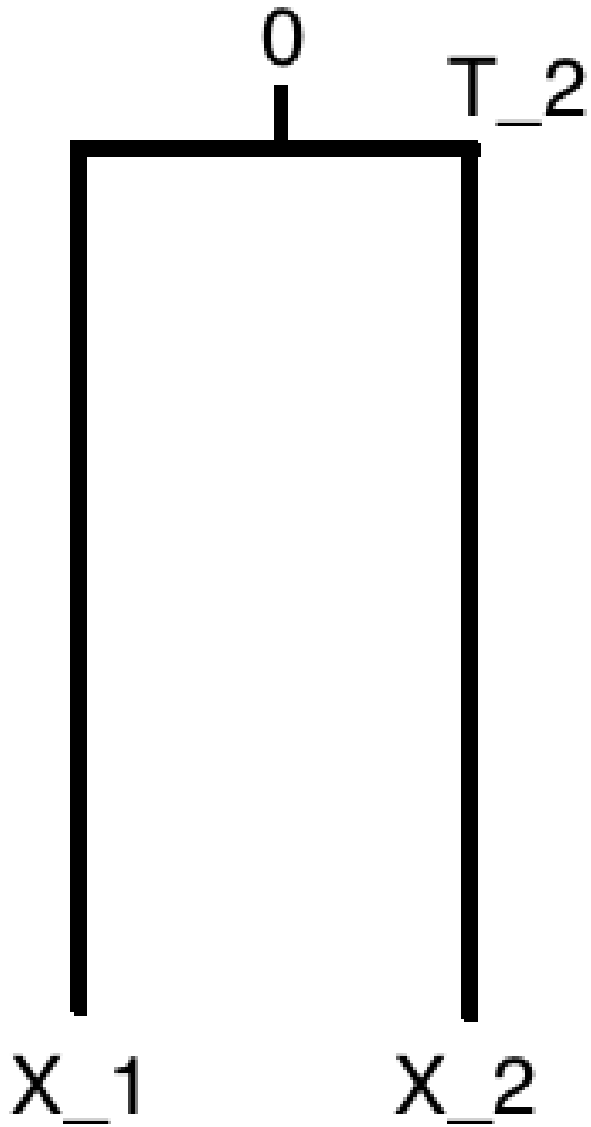
CF of a CP process

If $X(t)$ is a rate λ Compound Poisson process, and $\psi(k)$ is the characteristic function of the jump distribution, then the characteristic function of X is

$$\phi_t(k) = e^{\lambda t(\psi(k)-1)}$$

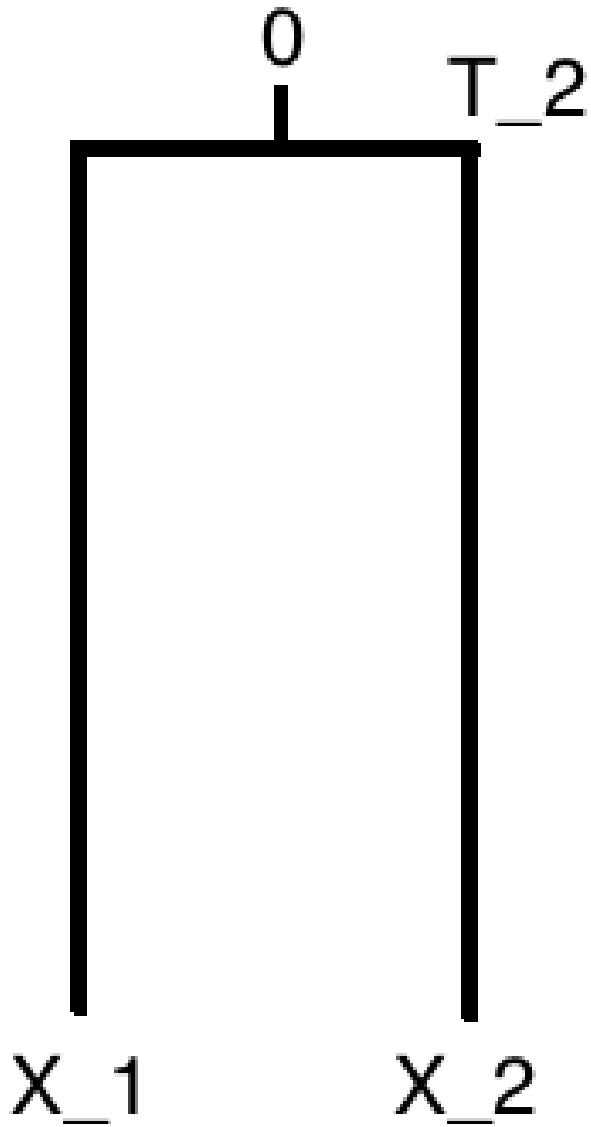
One locus, sample of size 2

- Condition on T_2, the coalescence time
  - X_1 and X_2 are independent $CP(\theta/2)$ processes run for time T_2
  - Joint distribution depends on root value
- Consider $Z = X_2 - X_1$
  - Doesn't depend on root



One locus, sample of size 2 (CF)

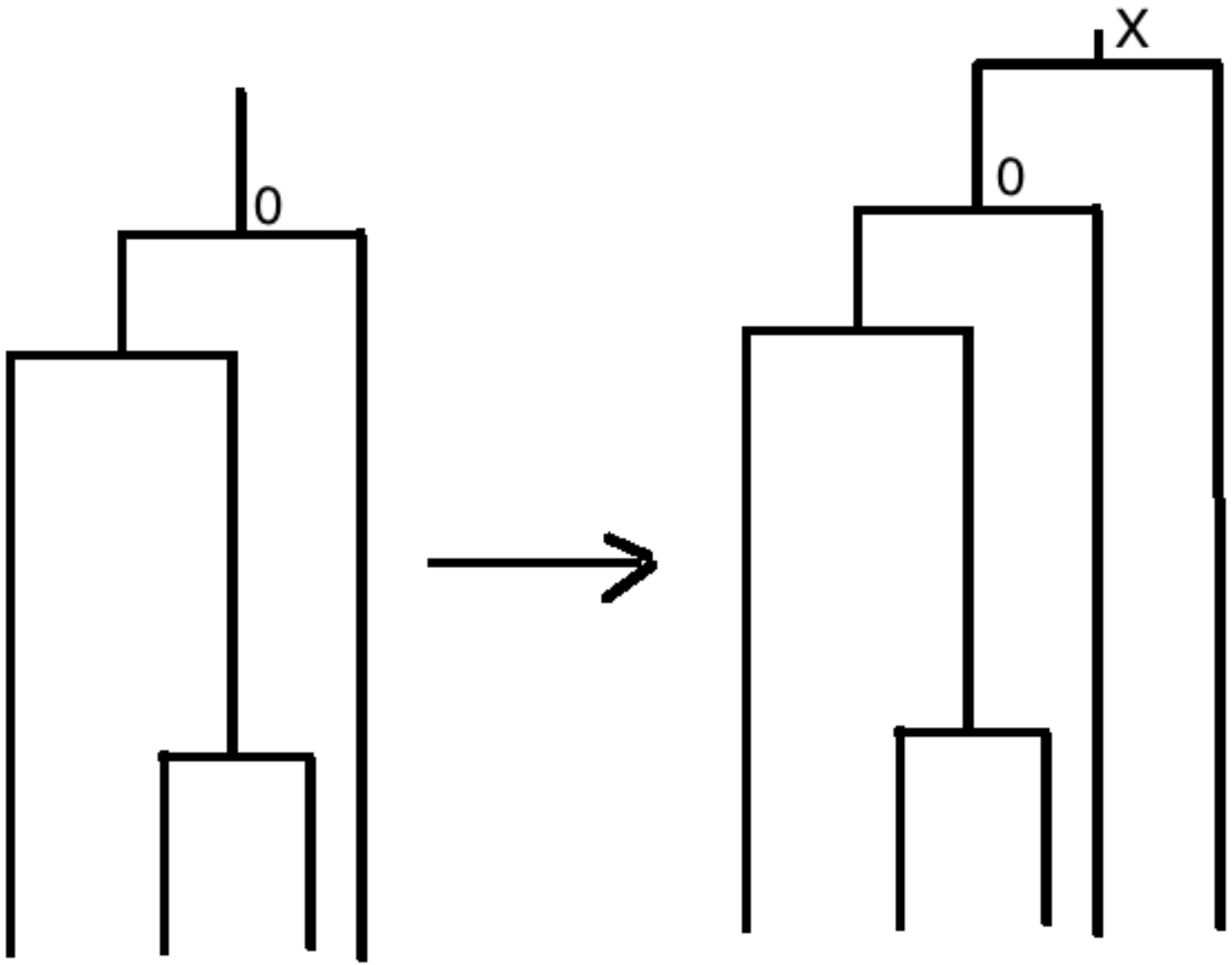
$$
\begin{aligned}
\phi_Z(k) &= \int_0^\infty \mathbb{E}(e^{ikZ})e^{-t}dt \\
&= \int_0^\infty \mathbb{E}(e^{ik(X_2-X_1)})e^{-t}dt \\
&= \int_0^\infty \phi_t(k)\phi_t(-k)e^{-t}dt \\
&= \int_0^\infty e^{\frac{\theta}{2}t(\psi(k)+\psi(-k)-2)}e^{-t}dt \\
&= \frac{1}{1-\frac{\theta}{2}(\psi(k)+\psi(-k)-2)}
\end{aligned}
$$



Phenotype at the root

- The root of the tree must be specified
- A new individual could coalesce *more anciently* than the current root
  - For each n, there is a fixed root
  - Ensuring consistency could be hard

The root problem



Getting rid of the root

Normalization

If $X = (X_0, \ldots, X_{n-1})$ are the trait values in the sample, then the random vector

$$\begin{aligned} Z &= (Z_1, \ldots, Z_{n-1}) \\ &= (X_1 - X_0, \ldots, X_{n-1} - X_0) \end{aligned}$$

does not depend on the root value

- "Normalizing" by an arbitrary individual makes n unimportant
  - Intuition: now every sample has to follow the path through the tree to individual 0, so the root doesn't matter

One locus, sample of size 3

Characteristic function for $n = 3$, and symmetric mutational effects

$$\phi_{Z_1,Z_2}(k_1,k_2) = \frac{\frac{1}{1+\theta(1-\psi(k_1))} + \frac{1}{1+\theta(1-\psi(k_2))} + \frac{1}{1+\theta(1-\psi(k_1+k_2))}}{3 - \frac{\theta}{2}(\psi(k_1)+\psi(k_2)+\psi(k_1+k_2)-3)}$$

- Sketch of derivation:
  - Condition on topology (each of 3 topologies equally likely)
  - Compute characteristic function for each topology while conditioning on coalescence times
  - Integrate over coalescence times
  - Take weighted sum of characteristic functions for each topology

Sending the number of loci off to infinity

- We would like some nice limit as the number of loci increases
  - Trivial result: "uniform on $\mathbb{R}$" or $\delta(x)$
- Need to decrease the effect size or mutation rate of each locus
- Three nontrivial limits
  - Mutation rate per locus decreases but effect sizes remain constaint
  - Mutation rate per locus remains constant but effect sizes decrease and effect size distribution does not have fat tails
  - Mutation rate per locus remains constant but effect size decreases and effect size distribution has fat tail

Decreasing mutation rate

Correlated CPP limit

As $n \uparrow \infty$ and $\theta \downarrow 0$ such that $n\theta \to \Theta$,

$$\phi_{Z_1,Z_2}(k_1, k_2) \to e^{\frac{\Theta}{2}(\psi(k_1)+\psi(k_2)+\psi(k_1+k_2)-3)}$$

which is the characteristic function of two correlated compound Poisson processes.

- Perhaps not very biologically relevant
- Will ignore for rest of talk

Small effect mutations

Bivariate Gaussian limit

As $n \uparrow \infty$ and the second moment of the effect distrbution, τ^2, $\downarrow 0$ such that $n\tau^2 \to \sigma^2$,

$$\phi_{Z_1,Z_2}(k_1, k_2) \to e^{-\frac{\theta}{2}\sigma^2(k_1^2+k_1k_2+k_2^2)}$$

which is the characteristic function of a Bivariate Gaussian distribution with mean vector $(0, 0)$ and variance-covariance matrix

$$\Sigma = \theta\sigma^2 \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

Large effect mutations

Bivariate stable distribution

Assume that $p(y) \sim \kappa|y|^{-(\alpha+1)}$ and set $t = \kappa\pi\left(\sin(\alpha\pi/2)\Gamma(\alpha)\alpha\right)^{-1}$. As $n \uparrow \infty$ and $t \downarrow 0$ such that $nt \to c$,

$$\phi_{Z_1,Z_2}(k_1, k_2) \to e^{-\frac{1}{2}\theta c(|k_1|^\alpha + |k_2|^\alpha + |k_1+k_2|^\alpha)}$$

where $c = \tilde{c}\frac{\pi}{\sin\left(\frac{\alpha\pi}{2}\right)}$, which is the characteristic function of a bivariate α-stable distribution

de Finetti's theorem

Theorem

If $(X_1, X_2, \ldots)$ is an infinitely exchangeable random vector, then the probability density of $(X_1 = x_1, \ldots X_n = x_n)$ is a mixture over i.i.d. probability densities. That is,

$$p(x_1, \ldots, x_n) = \int \prod_i p_\theta(x_i) \nu(d\theta)$$

- Suggests that we can find a de Finetti measure such that all our normalized samples are i.i.d.

A guess in the normal case

- Given the mean, the trait is distributed $\mathcal{N}(0, \frac{\theta}{2}\sigma^2)$
- The mean is itself distributed $\mathcal{N}(0, \frac{\theta}{2}\sigma^2)$
- Law of total variance gives $\mathrm{Var}(X) = 2N\mu\sigma^2$
  - Same as classical derivation

$$\begin{aligned}
\phi(k_1, k_2) &= \int \left(\int \prod_{l=1}^{2} e^{ik_l x_l} \frac{1}{\sqrt{\pi\theta\sigma^2}} e^{-\frac{(m-x_l)^2}{\theta\sigma^2}} dx_l \right) \frac{1}{\sqrt{\pi\theta\sigma^2}} e^{-\frac{m^2}{\theta\sigma^2}} dm \\
&= e^{-\frac{\theta}{4}\sigma^2\left(k_1^2+k_2^2\right)} \int e^{im(k_1+k_2)} \frac{1}{\sqrt{\pi\theta\sigma^2}} e^{-\frac{m^2}{\theta\sigma^2}} dm \\
&= e^{-\frac{\theta}{4}\sigma^2\left(k_1^2+k_2^2\right)} e^{-\frac{\theta}{4}\sigma^2(k_1+k_2)^2} \\
&= e^{-\frac{\theta}{2}\sigma^2\left(k_1^2+k_1k_2+k_2^2\right)}
\end{aligned}$$

Tempting interpretation and conjecture

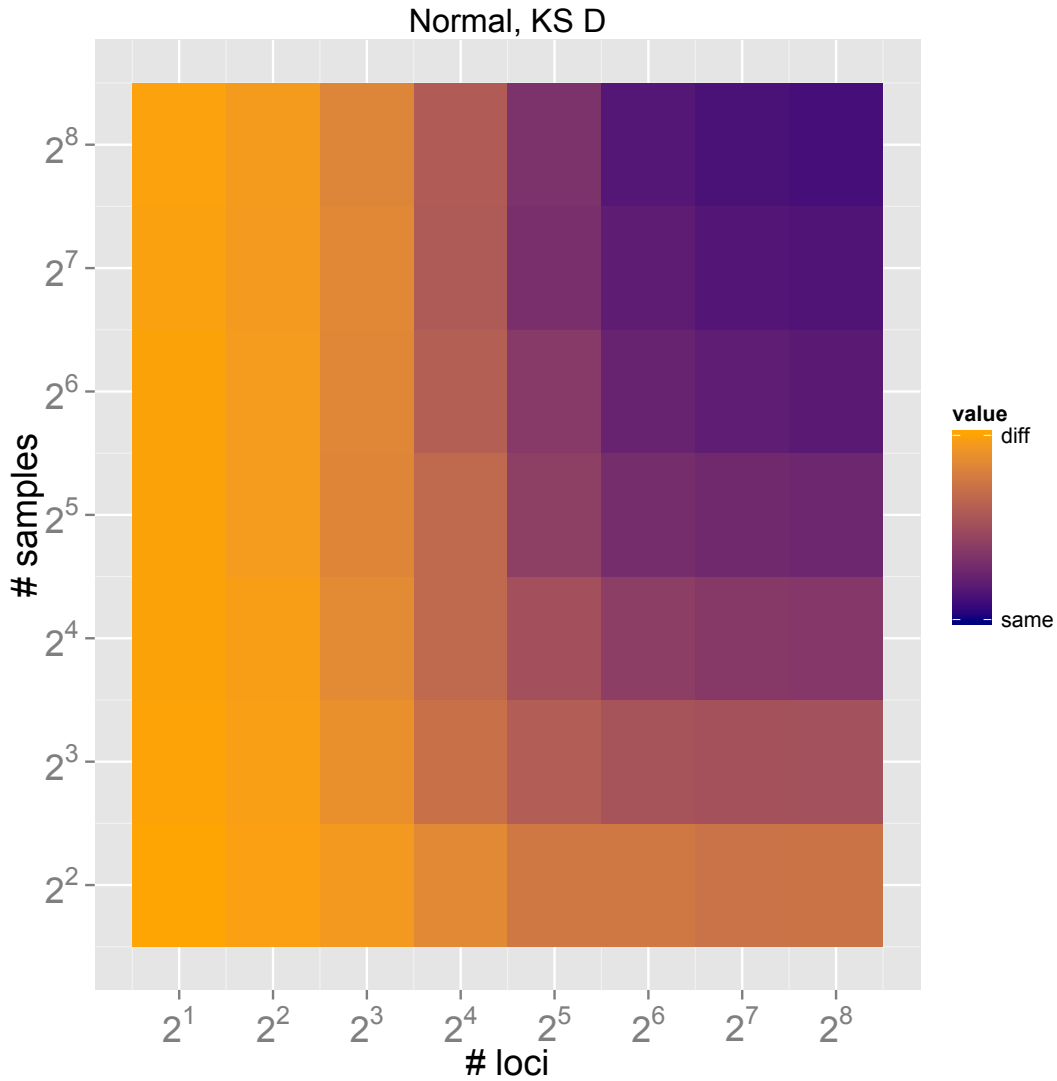
- Conditioning on the amount of evolution "exclusive to sample 0"
  - Integrate over whether sample 1 or sample 2 coalesces with sample 0 first
- Suggests that this can be extended to arbitrary sample sizes

Conjecture about larger samples (Gaussian limit)

When the mutation kernel has only small effect mutations, the limit distribution is normal with variance $\frac{\theta}{2}\sigma^2$ and random mean.

Testing the Gaussian conjecture

- Simulate quantitative traits according to the coalescent model
- Use KS test to assess convergence in the limit



Conjecture for the stable limit

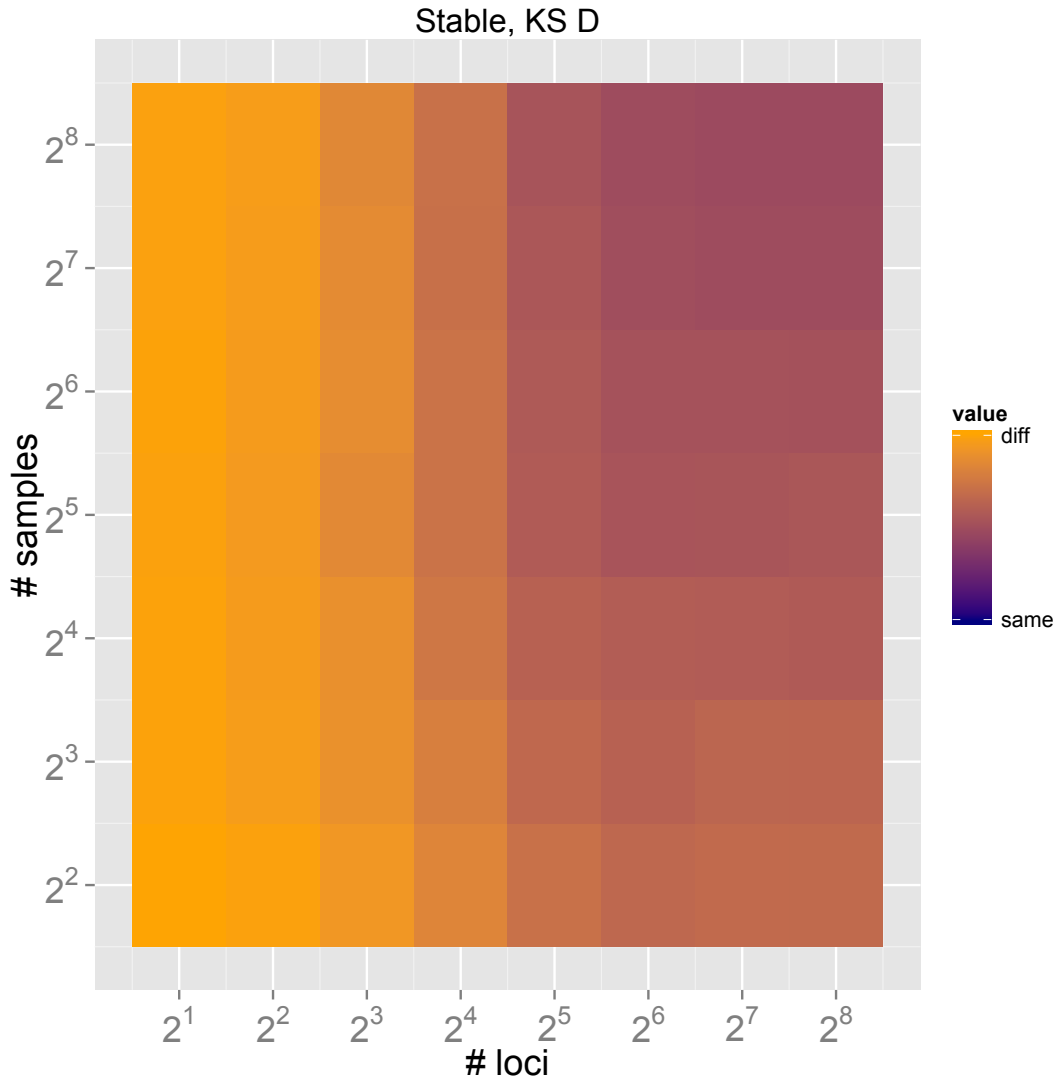
- Bivariate case
  - Independently α-stable with random median
  - Analogous to Gaussian limit

Conjecture about larger samples (Stable limit)

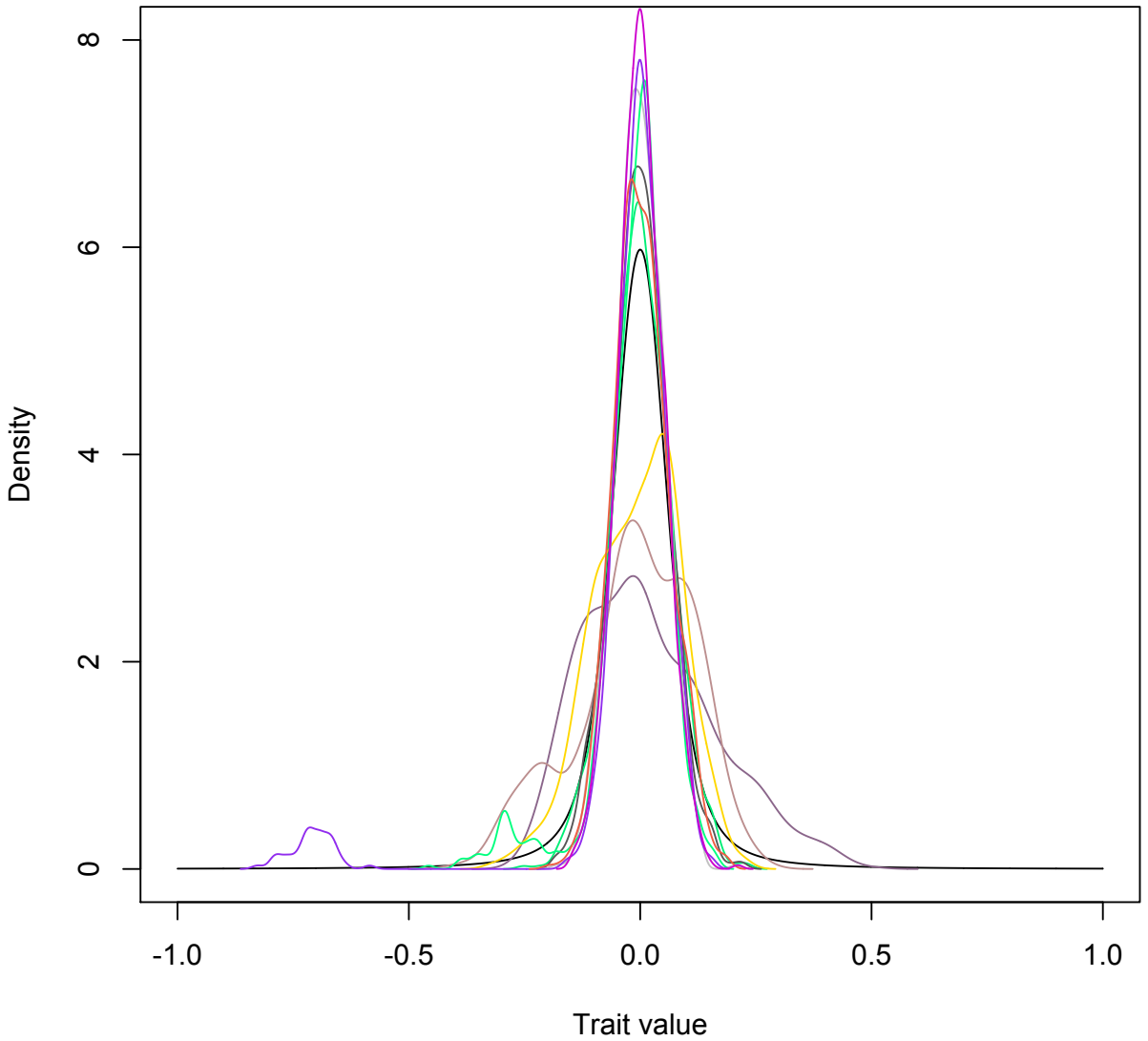
When the mutation kernel has large mutational effects, the limit distribution is α-stable with scale $\left(\frac{\theta}{2}c\right)^{1/\alpha}$ and random median.

Testing the stable conjecture

- Simulate quantitative traits according to the coalescent model
- Use KS test to assess convergence in the limit



More randomness than expected



Stable limit for sample of size 4

- Tedious computation
  - Need to average over 18 trees.
- Conjecture would require

$$\phi_{Z_1,Z_2,Z_3}(k_1,k_2,k_3) \to e^{-\frac{\theta}{2}c(|k_1|^\alpha+|k_2|^\alpha+|k_3|^\alpha+|k_1+k_2+k_3|^\alpha)}$$

Trivariate characteristic function

For a sample of size three, under the same conditions as the bivariate limit,

$$\begin{aligned}\phi_{Z_1,Z_2,Z_3}(k_1,k_2,k_3) \to \exp\{-\frac{\theta}{3}c(&|k_1|^\alpha+|k_2|^\alpha+|k_3|^\alpha \\ &+\frac{1}{2}|k_1+k_2|^\alpha+\frac{1}{2}|k_1+k_3|^\alpha+\frac{1}{2}|k_2+k_3|^\alpha \\ &+|k_1+k_2+k_3|^\alpha)\}\end{aligned}$$

Extremely weak conjecture for large effect mutations

Conjecture (Stable limit)

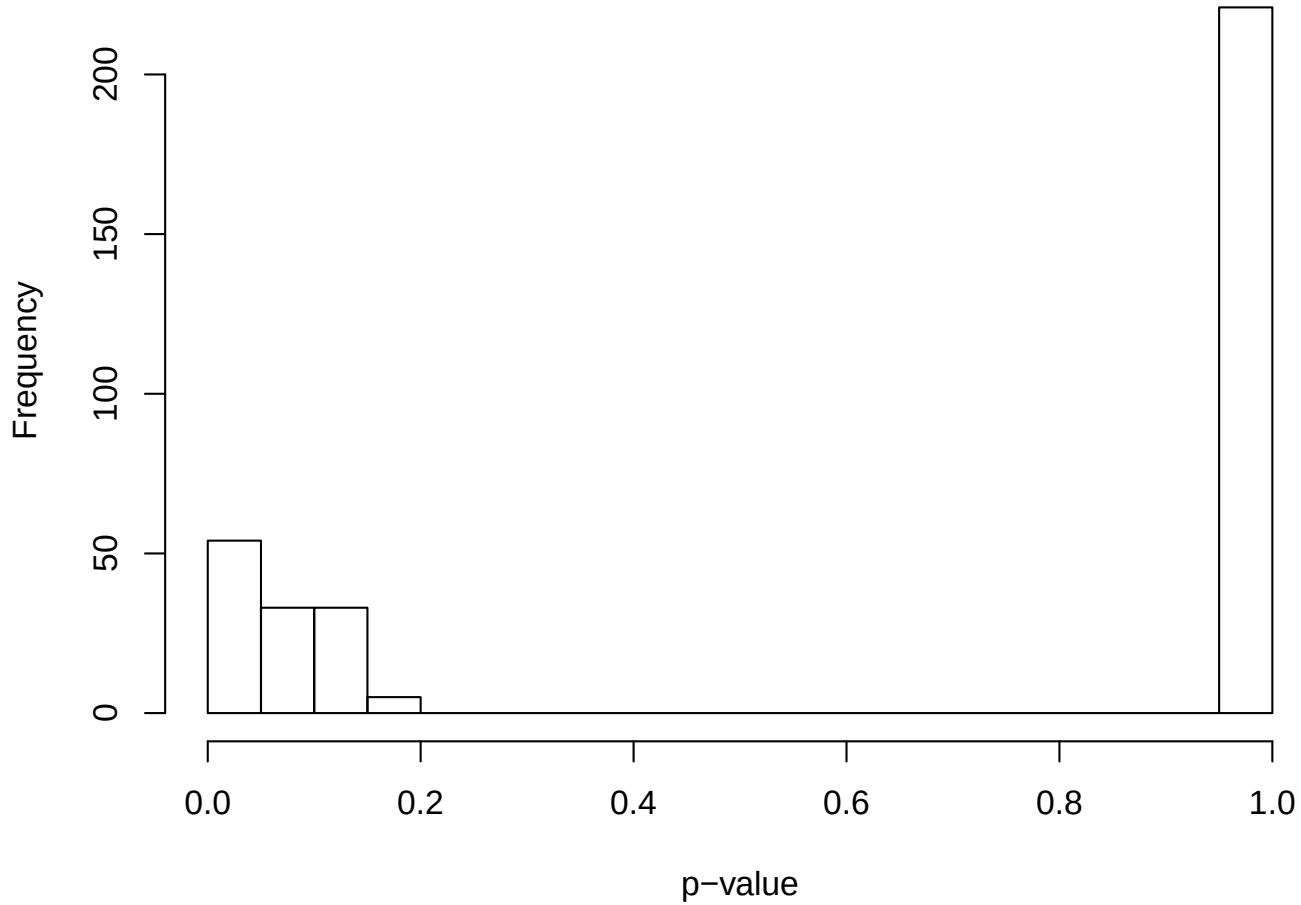
When the mutation kernel has large mutational effects, the limit distribution is α-stable with random parameters

- Not even sure that this is true
  - A mixture of stable distributions?

Empirical data

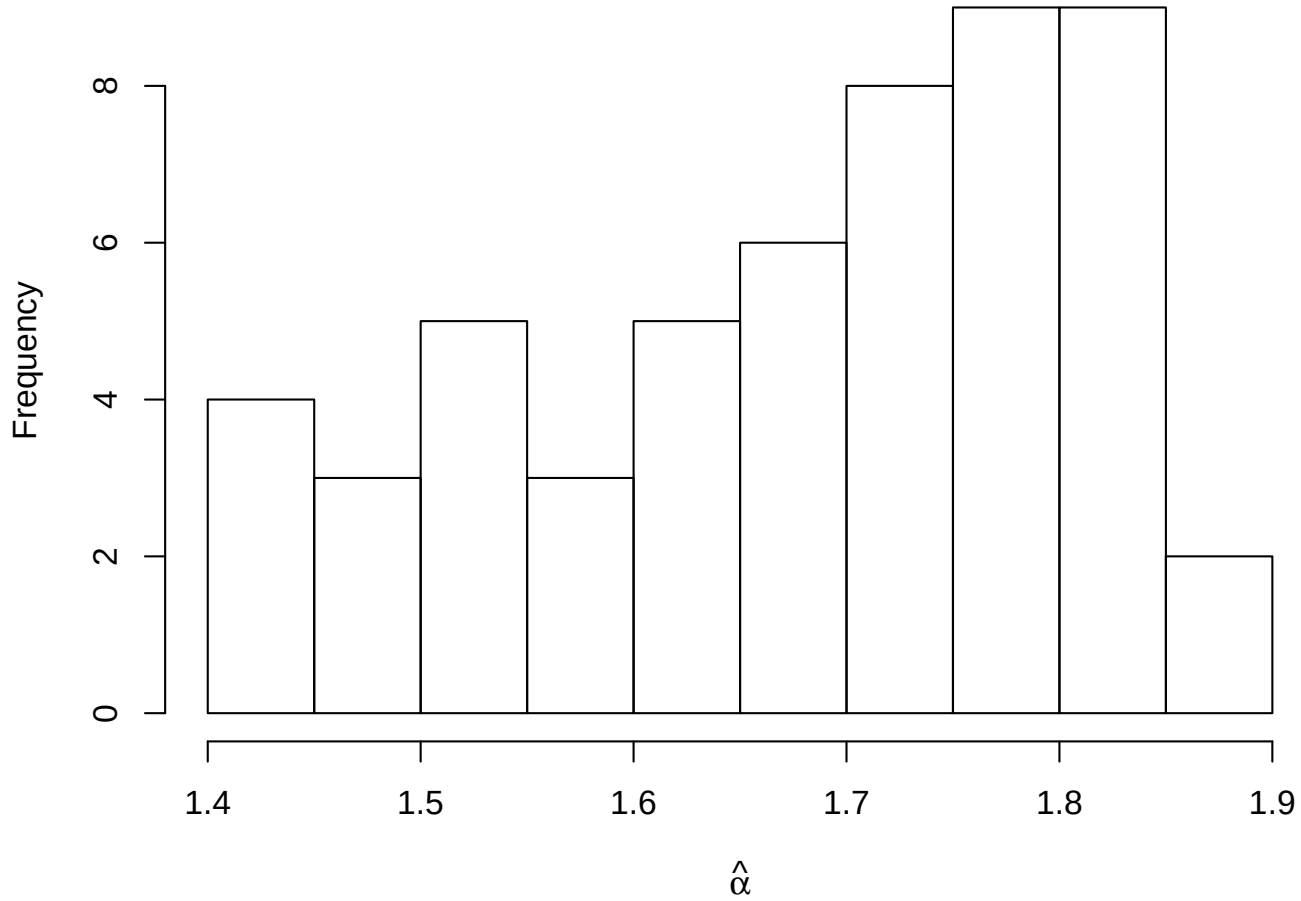
- *Neurospora crassa* mRNA-seq
  - Ellison et al (2011), PNAS
  - Restrict to genes with FPKM > 1
- For each gene fit normal distribution, α-stable distribution
- Bootstrap likelihood ratio test ($H_0 : \alpha = 2$ vs. $H_1 : \alpha < 2$)
- Preliminary!
  - Only analyzed 346 genes

Distribution of p-values

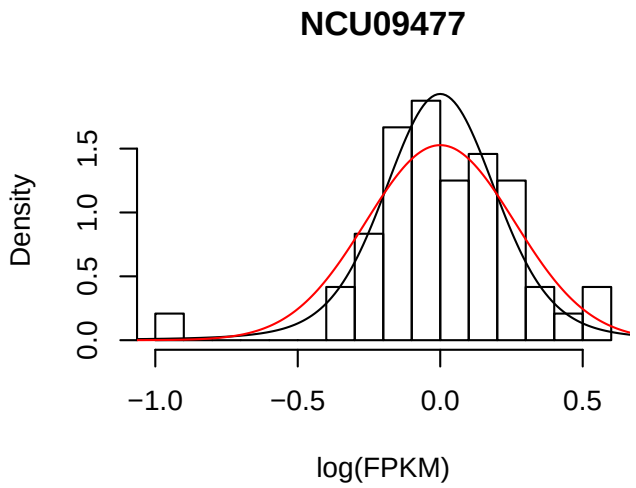


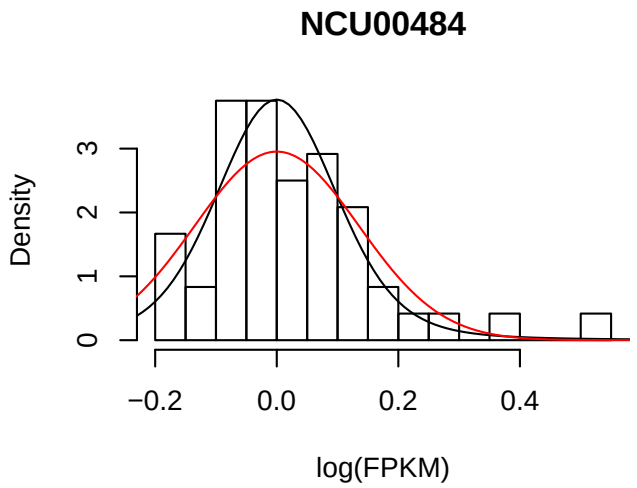
- 54 genes significant at FDR of 32%

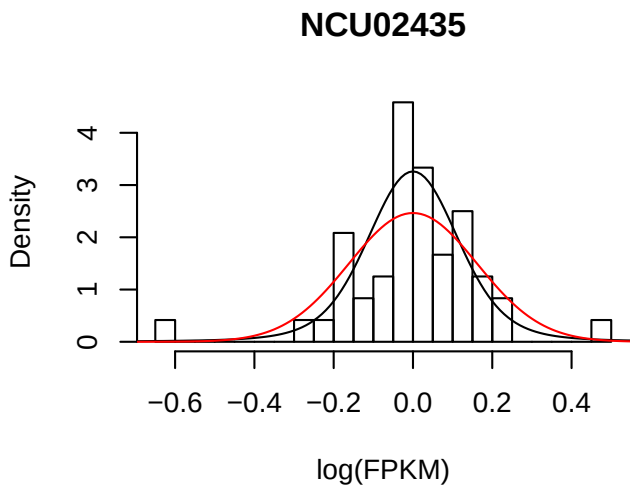
Distribution of α in significant genes

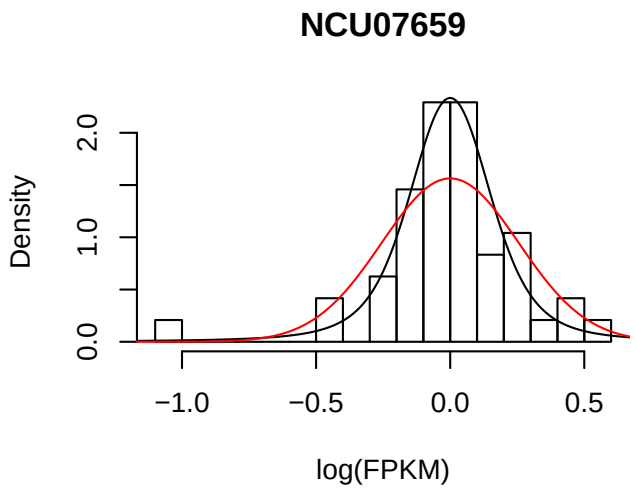


Cherry-picked examples









Conclusions

- Introduced a coalescent model of quantitative trait evolution
  - Provided exact formulas for the characteristic functions for samples of size 2, 3, 4
  - Found limiting distributions as the number of loci became large in each case
  - Conjectured about extending these limits to larger samples
- Extensions
  - Samples not contemporaneous
  - Population structure
  - Diploids
- Why a neutral model?
  - Analytically tractable calculations
  - Intuition for what to expect with weak selection
  - Null model