Intro to population genetics

Shamil Sunyaev

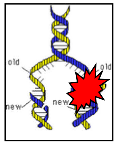
Department of Biomedical Informatics
Harvard Medical School

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Broad Institute of M.I.T. and Harvard

Forces responsible for genetic change

Mutation



μ

Selection

s

Drift

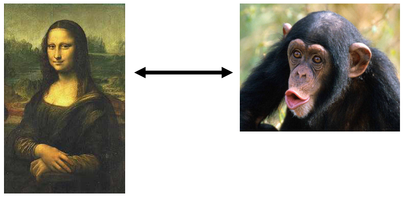
N_e

Population structure

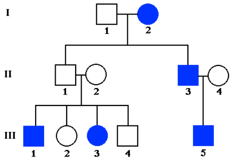
F_{ST}

Mutations

Mutation rate in humans and flies



2.5x10^{-8} (Nachman & Crowell)



Pedigree 2. An idealized pedigree demonstrating the effects of incomplete penetrance.

1.8x10^{-8} (Kondrashov)

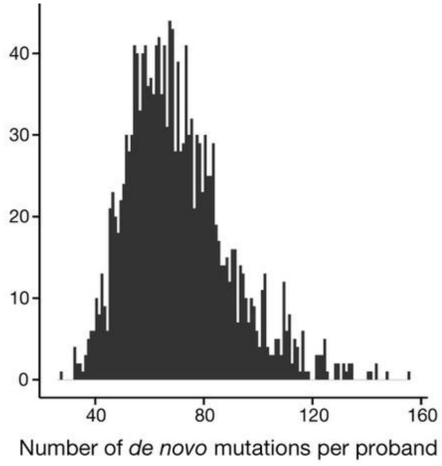
NGS estimates ~*1.2X10^{-8}* per nt changes genome

~*70* per nt changes genome

Other events: indels (10^{-9})

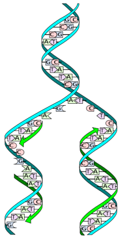
repeat extensions/contractions (10^{-5})

Number of de novo mutations per individual

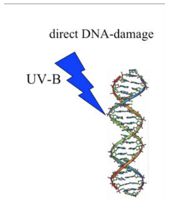


Jonsson et al., *Nature* 2017

Mutation rate is variable along the genome

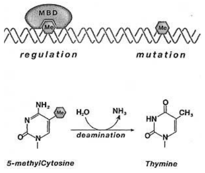


Replication fidelity



DNA damage

DNA repair



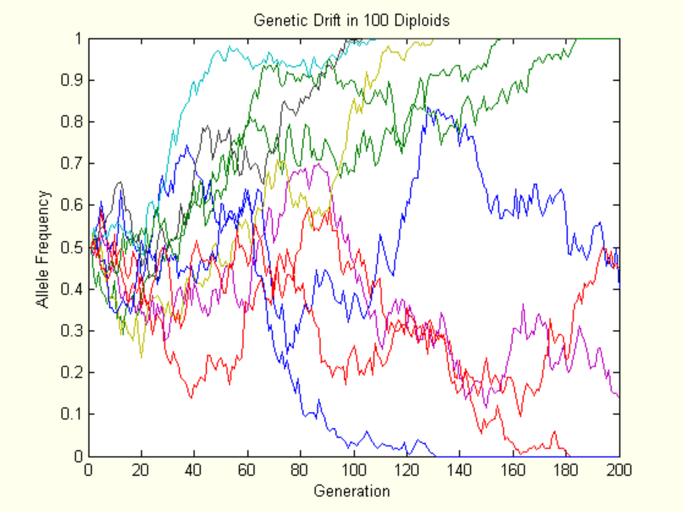
CpG deamination

Regional variation of mutation rate

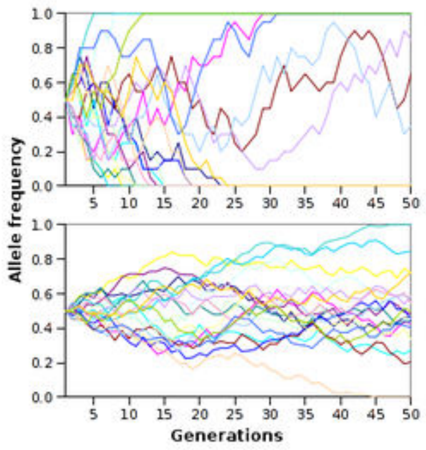
Context dependence of mutation rate

Genetic drift

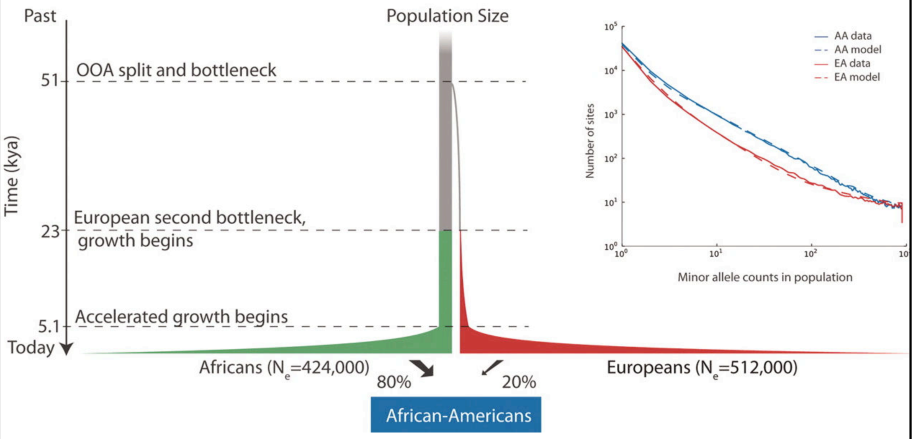
Drift is a random change of allele frequencies



Drift depends on population size



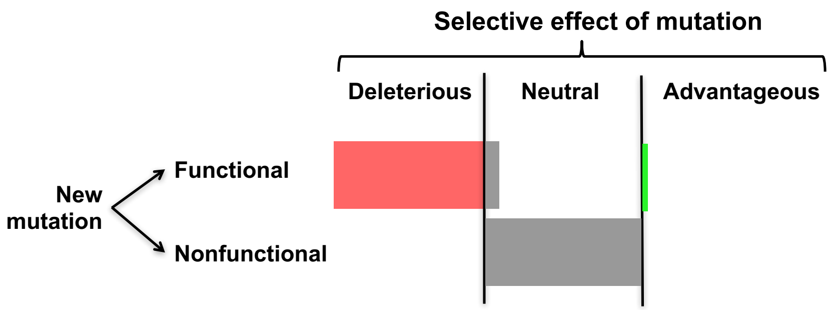
Demographic history



Tennessen et al. *Science* 2012

Selection

Most functional mutations are deleterious

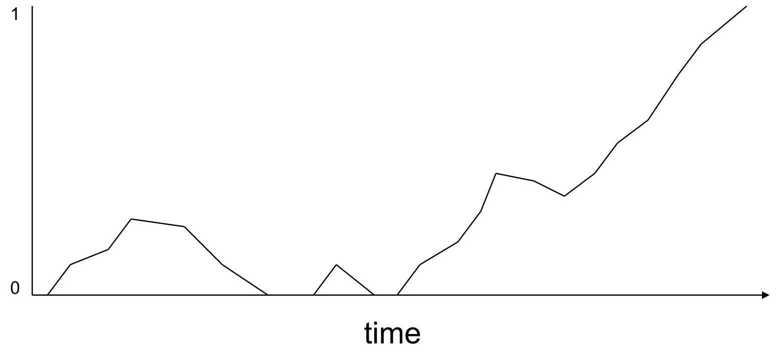


Selection indicates functional mutations, whether or not the tested trait is under selection

Methods of mathematical population genetics

Dynamic of allelic substitution

Mathematically, allele frequency change in a population follows a one-dimensional random walk



Diffusion approximation

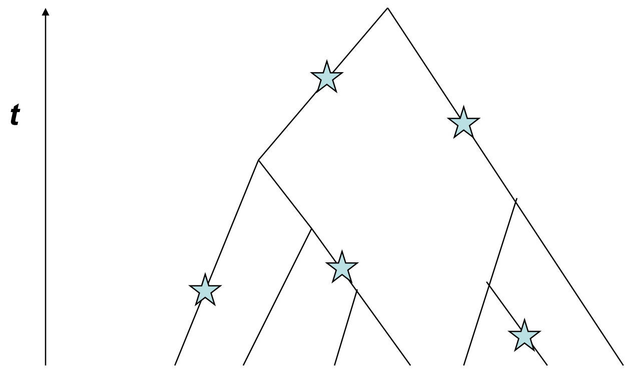
Random walk that does not jump long distances can be approximated by a diffusion process

$$\frac{\partial \phi(x,p,t)}{\partial t} = -\frac{\partial M\phi(x,p,t)}{\partial x} + \frac{1}{2}\frac{\partial^2 V\phi(x,p,t)}{\partial x^2}$$

Coalescent theory

Instead of modeling a population, we can model our sample

Time goes backwards !



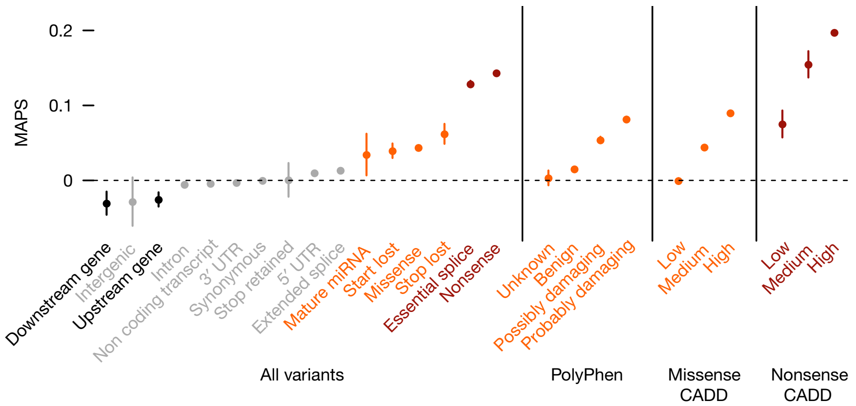
Natural selection in protein coding regions

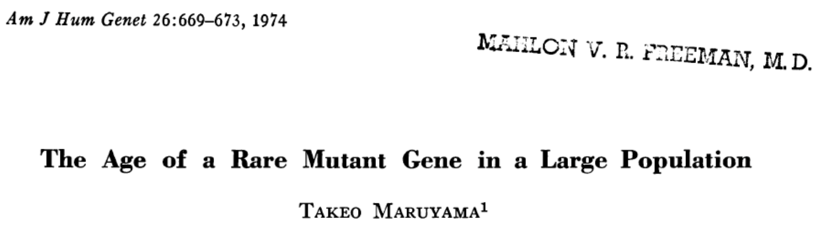
Signatures of purifying selection

Reduced variation

Excess of rare alleles

Diversity and allele frequency





Am J Hum Genet 26:669–673, 1974

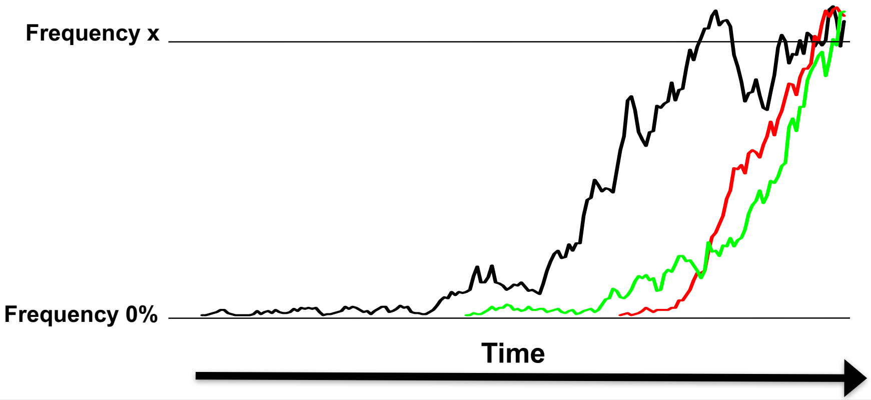
MAHLON V. R. FREEMAN, M.D.

The Age of a Rare Mutant Gene in a Large Population

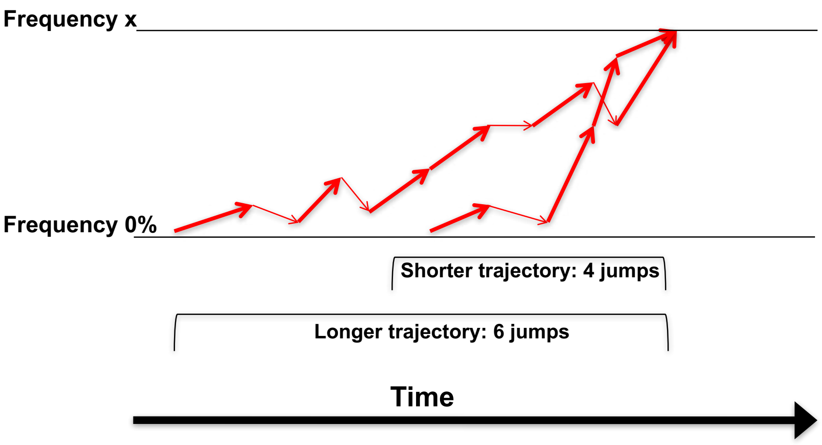
TAKEO MARUYAMA[1]

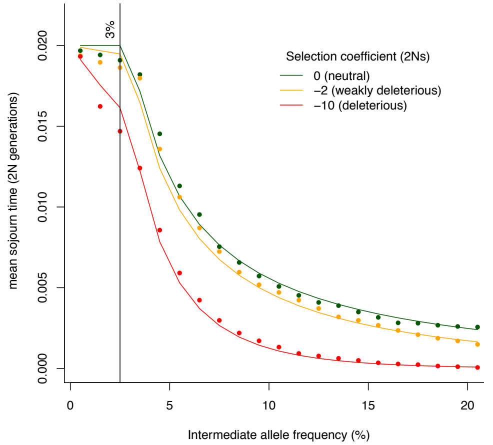
At a given frequency deleterious and advantageous alleles are younger than neutral

Maruyama effect (1974): at any frequency **advantageous** , or **deleterious** alleles are younger than **neutral** alleles



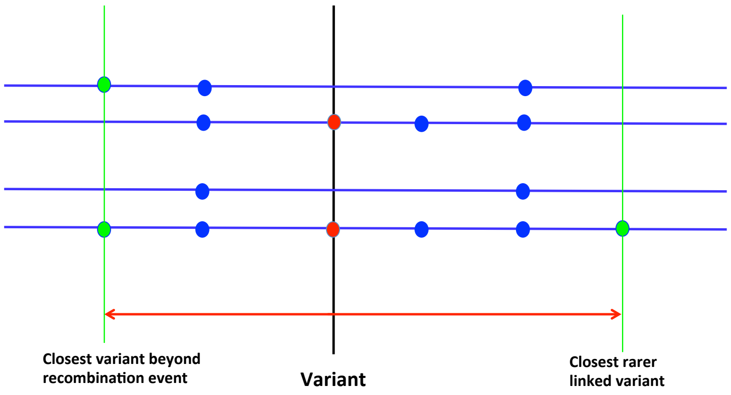
Intuition: shorter trajectories require fewer lucky jumps





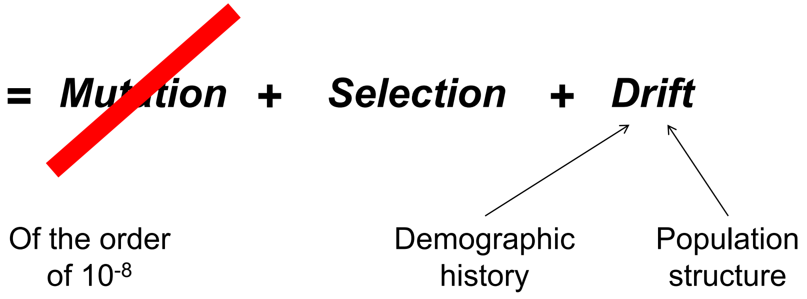
Kiezun et al. *PLOS Genetics* 2013

Neighborhood clock (fuzzy clock)


Closest variant beyond recombination event
Variant
Closest rarer linked variant

Selection inference using frequencies of individual SNPs

Change in allele frequency **=**



= ~~***Mutation***~~ **+** ***Selection*** **+** ***Drift***

Of the order of 10^{-8}

Demographic history

Population structure

Focusing on rare deleterious PTVs

PTV – protein truncating variant (a.k.a. nonsense)

Combine all PTVs per gene – we assume that they have identical effects

Consider each gene as a bi-allelic locus – PTV / no PTV

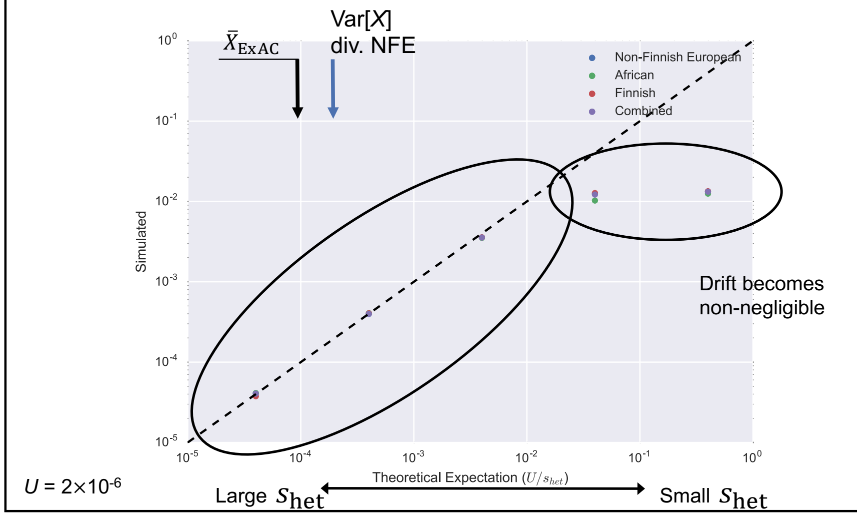
Selection inference using combined frequency of PTVs

Change in allele frequency **=**

= ***Mutation*** **+** ***Selection*** **+** ~~***Drift***~~

Assuming string selection and a very large population, combined frequency of rare deleterious PTVs is expected to be Poisson distributed with $\lambda = U/hs$

Simulations



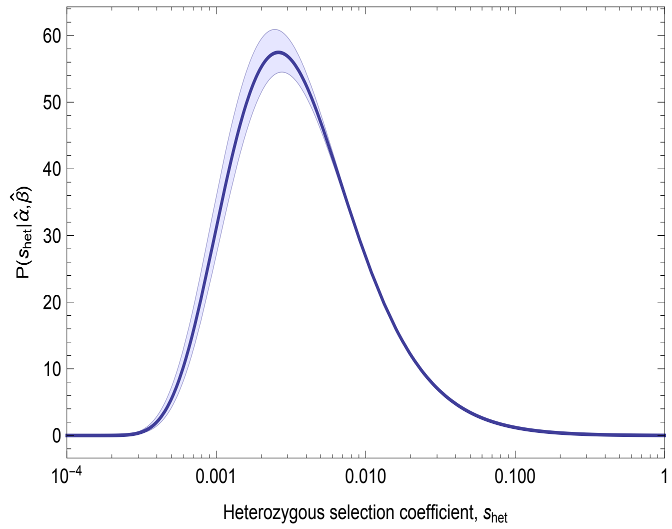
The model

PTV counts in each gene are Poisson distributed but we lack sufficient data to estimate selection coefficients

We can treat selection coefficients as random variables with a distribution to be estimated

$$P(n|\alpha,\beta;\nu) = \int P(n|s_{\mathrm{het}};\nu)\, P(s_{\mathrm{het}};\alpha,\beta)\, \mathrm{d}s_{\mathrm{het}}$$

Distribution of selection coefficients



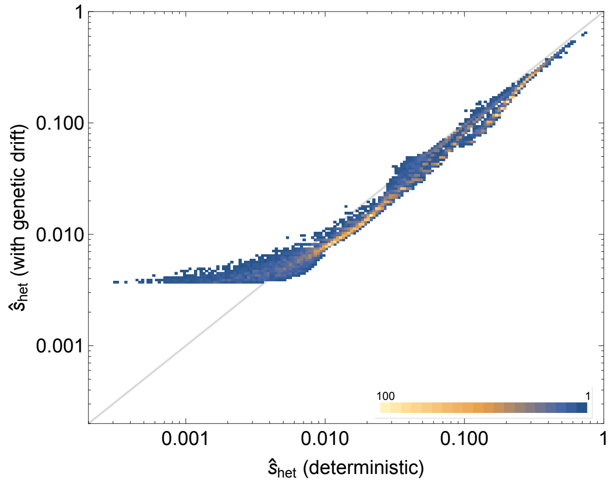
Cassa, Weghorn, Balick, Jordan et al. *Nature Genetics* 2017

Estimates for each gene

The estimated distribution over selection coefficients can be now used as a prior, and per gene estimates from posteriors

$$P\left(s_{\mathrm{het},i}|n_i;\nu_i\right) = \frac{P\left(n_i|s_{\mathrm{het},i};\nu_i\right)P\left(s_{\mathrm{het},i};\hat{\alpha}_t,\hat{\beta}_t\right)}{\int P\left(n_i|s;\nu_i\right)P\left(s;\hat{\alpha}_t,\hat{\beta}_t\right)\,\mathrm{d}s}$$

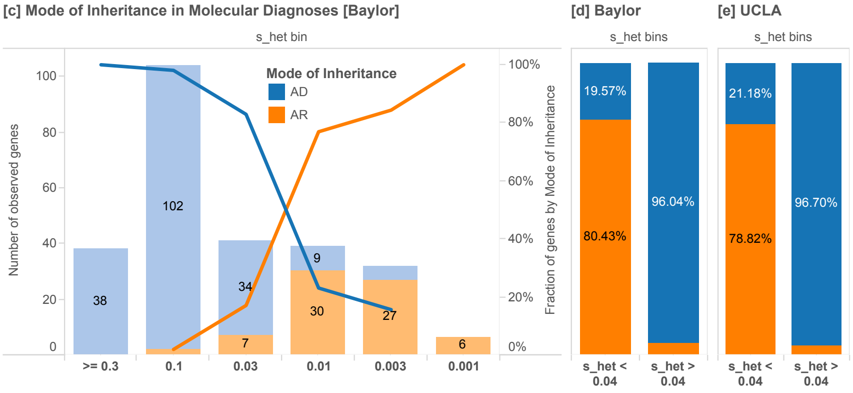
What happens if we incorporate drift?



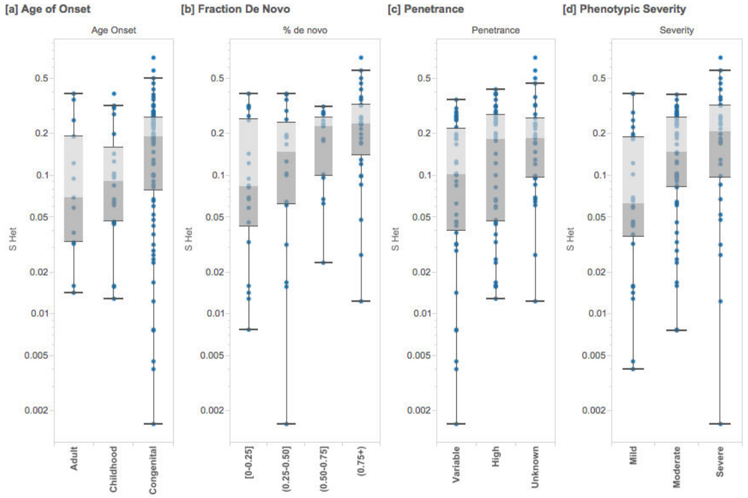
What happens if we incorporate drift?

1) The approach fails if selection is weak
2) The approach fails if mutational target is small
3) These considerations are important for regional constraint scores
4) Overall, the approach is non-informative in case of recessivity

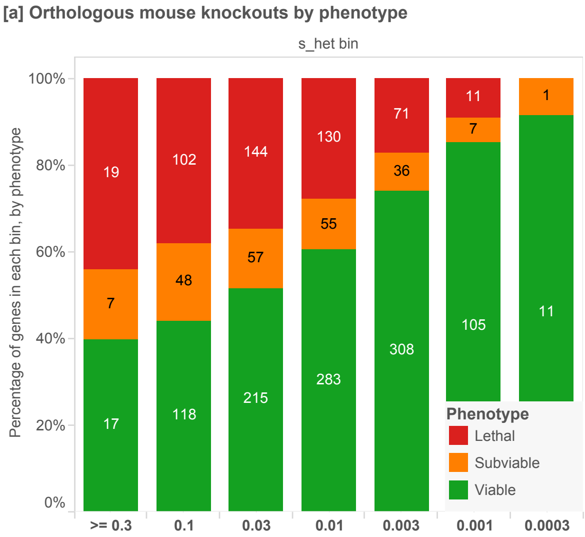
AD and AR Mendelian genes



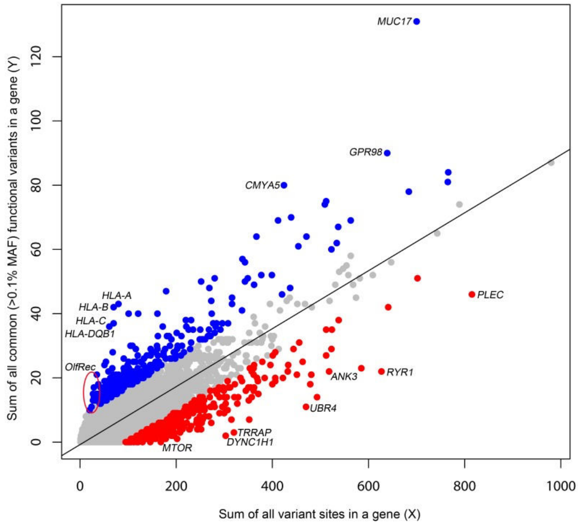
Age of onset, penetrance and severity



Concordance with mouse knockout data



RVIS



Petrovski et al. *PLOS Genetics* 2013

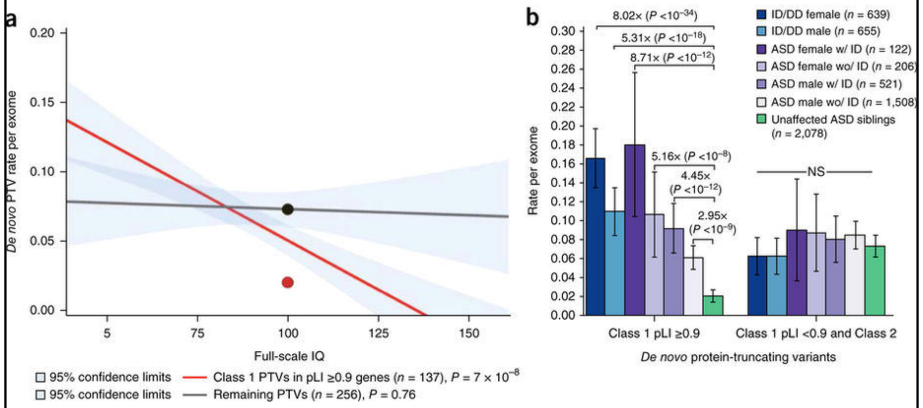
pLI

$$PTV_i \mid Z_i = c \sim Pois(N\lambda_c)$$

$$p(Z_i = c \mid \pi_c, PTV_i) = \frac{Pois(PTV_i \mid N\lambda_c)\pi_c}{\sum_c Pois(PTV_i \mid N\lambda_c)\pi_c},$$

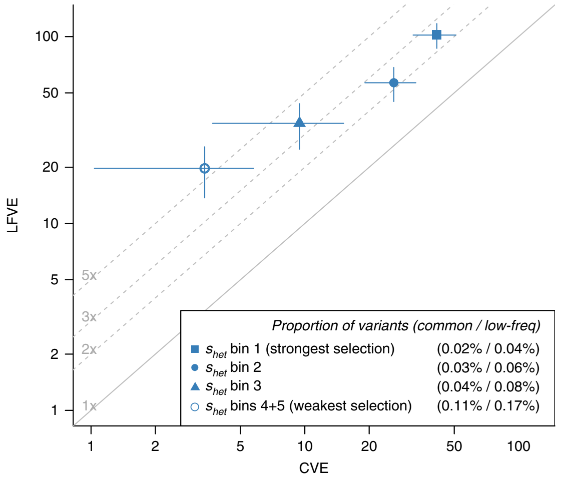
Lek et al. *Nature* 2016

De novo mutations in ASD



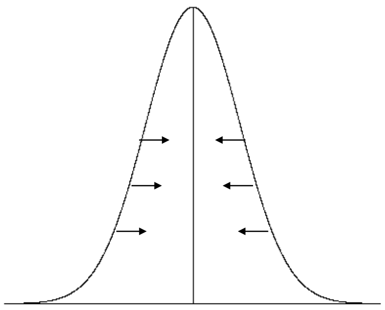
Kosmicki et al. *Nature Genetics* 2017

Heritability enrichment



Gazal et al. *Nature Genetics* 2018

Stabilizing selection is the most common type of selection on a quantitative trait



Stabilizing selection

Selection may be related or unrelated to the trait

Technically, non-neutral genetic variation should not exist!

Forces to maintain variation:

Selection

Mutation

Why does a common genetic disease exist?

From evolutionary perspective common genetic disease should not exist: natural selection should remove disease-causing alleles from the population

Theory 1: MEDICALLY detrimental polymorphisms are not EVOLUTIONARY deleterious

- **Disease late onset (after the reproductive age)**
- **Changed environment and lifestyle (Selection direction reversal)**
- **Compensatory positive effect**

Balancing selection

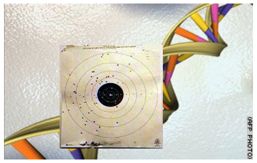
Frequency dependent selection

Antagonistic pleiotropy (Trade Off)

Examples: *APOE* (Alzheimer's disease)**, *AGT*** (Hypertension)**, *CYP3A*** (Hypertension)

Mutation/selection balance

Theory 2: Common diseases are due to multiple deleterious alleles in mutation-selection balance



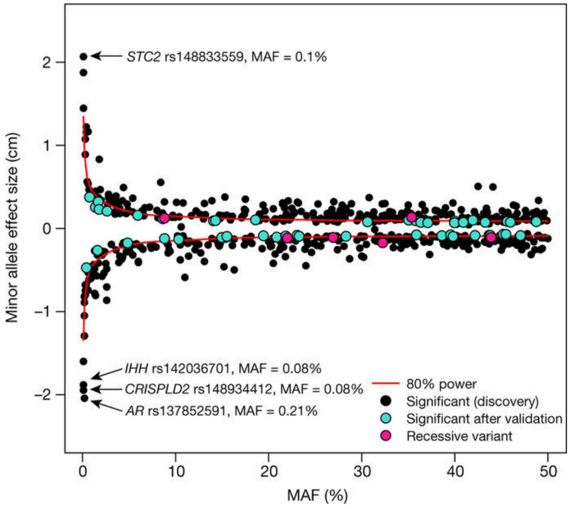
- **Weak selection**
- **High mutation rate**

CURRENT ESTIMATE:

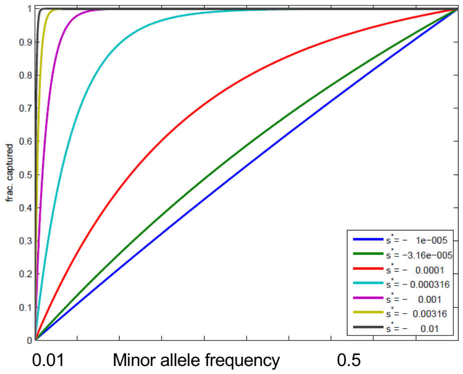
~70 new mutations per genome

~1 new coding mutation per genome

Rare coding alleles have larger effect sizes



Heritibility by allele frequency

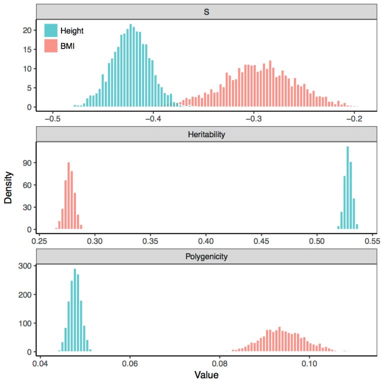


Effective population size: N=10,000

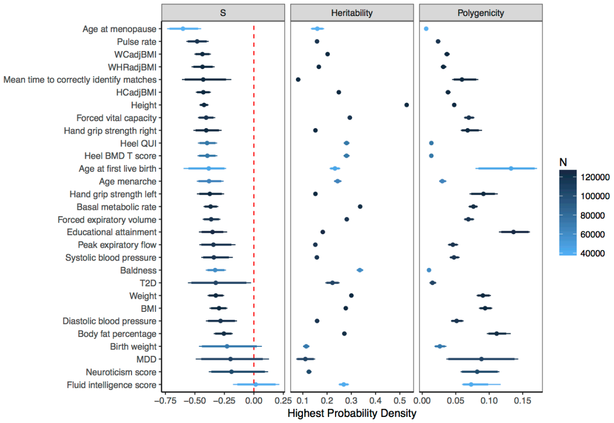
Evidence in favor of the highly polygenic model

$$\beta_j \sim N\left(0, \left[2p_j(1-p_j)\right]^S \sigma_\beta^2\right)\pi + \phi(1-\pi)$$

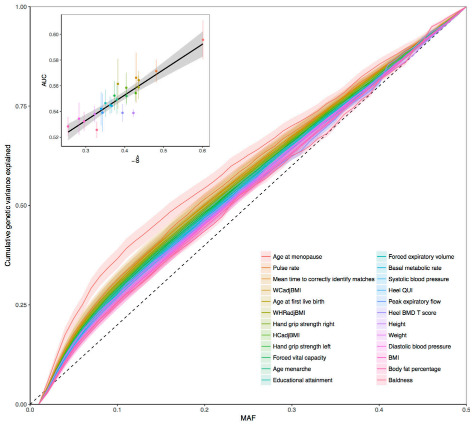
Evidence in favor of the highly polygenic model



Evidence in favor of the highly polygenic model



Evidence in favor of the highly polygenic model


Cumulative genetic variance explained
MAF
AUC
$-\hat{S}$
Age at menopause
Pulse rate
Mean time to correctly identify matches
WCadjBMI
Age at first live birth
WHRadjBMI
Hand grip strength right
HCadjBMI
Hand grip strength left
Forced vital capacity
Age menarche
Educational attainment
Forced expiratory volume
Basal metabolic rate
Systolic blood pressure
Heel QUI
Peak expiratory flow
Heel BMD T score
Height
Weight
Diastolic blood pressure
BMI
Body fat percentage
Baldness