

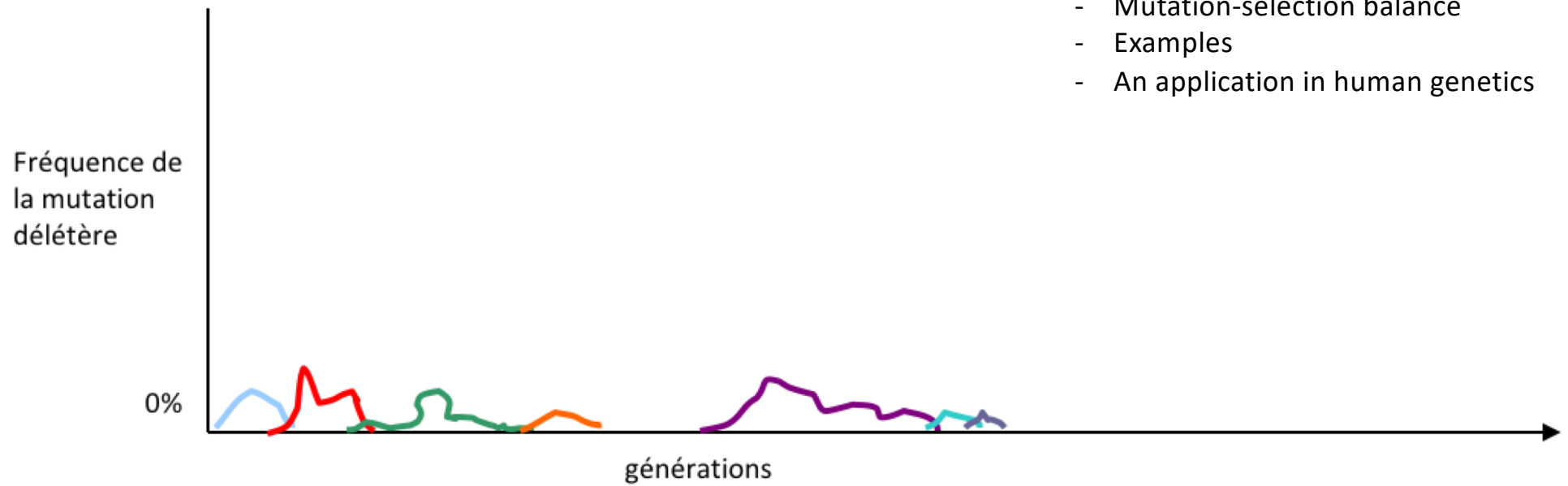
Mutation, sélection naturelle et fréquences des
allèles pathologiques chez l'Homme

Molly Przeworski

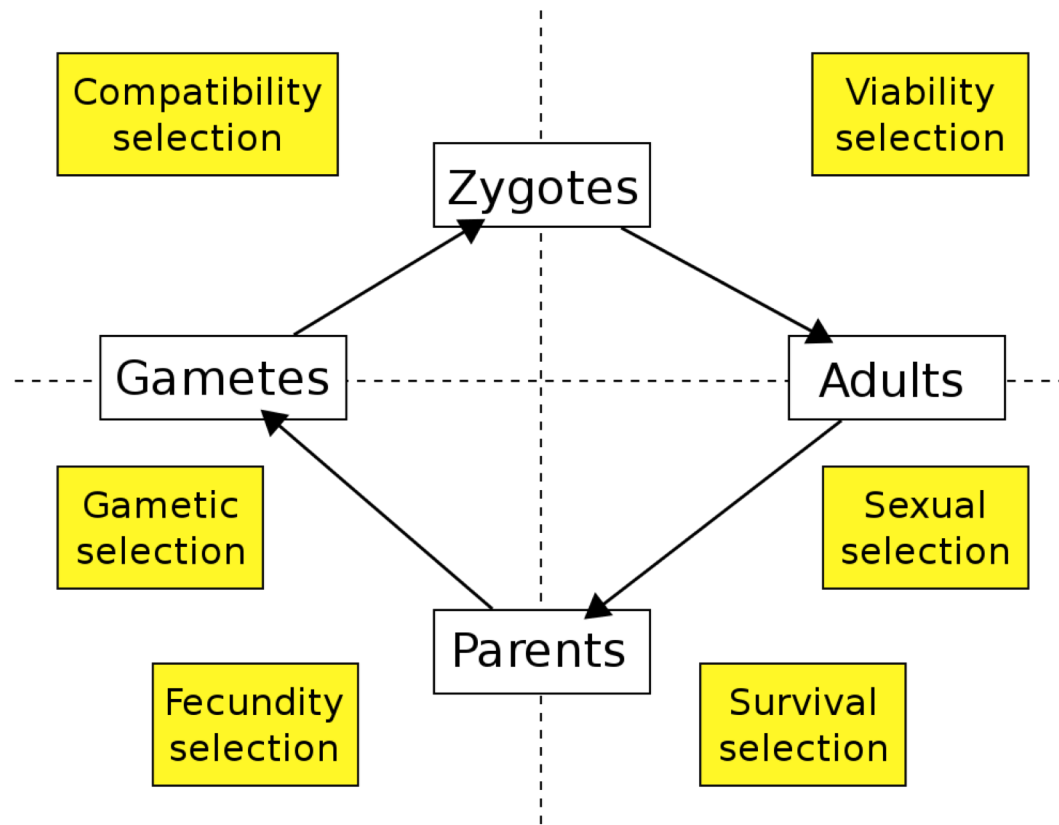
Cours #3

Outline:

- Viability selection model
- Mutation-selection balance
- Examples
- An application in human genetics



See Graham Coop's population genetics notes
https://github.com/cooplab/popgen-notes/blob/master/release_popgen_notes.pdf



+

No genetic drift
Panmixia

https://commons.wikimedia.org/wiki/File:Life_cycle_of_a_sexually_reproducing_organism.svg

Three genotypes

	A_1A_1	A_1A_2	A_2A_2
Absolute no. at birth	Np_t^2	$N2p_tq_t$	Nq_t^2
Fitnesses	W_{11}	W_{12}	W_{22}
Absolute no. at reproduction	$NW_{11}p_t^2$	$NW_{12}2p_tq_t$	$NW_{22}q_t^2$
Relative freq. at reproduction	$\frac{NW_{11}p_t^2}{N\bar{W}} = \frac{W_{11}}{\bar{W}}p_t^2$	$\frac{NW_{12}2p_tq_t}{N\bar{W}} = \frac{W_{12}}{\bar{W}}2p_tq_t$	$\frac{NW_{22}q_t^2}{N\bar{W}} = \frac{W_{22}}{\bar{W}}q_t^2$

- ❖ N is the number of diploid individuals
- ❖ p is the frequency of A1 in the population, q is the frequency of A2 in the population
- ❖ W_{ij} is the probability of a zygote carrying A_iA_j surviving to reproduction
- ❖ Frequencies at reproduction assume random mating
- ❖ $N\bar{W}$ is the total number of individuals who survive to reproduce, i.e., the mean fitness of the population times the number of individuals

See Graham Coop's population genetics notes
https://github.com/cooplab/popgen-notes/blob/master/release_popgen_notes.pdf

Three genotypes

	A_1A_1	A_1A_2	A_2A_2
Absolute no. at birth	Np_t^2	$N2p_tq_t$	Nq_t^2
Fitnesses	W_{11}	W_{12}	W_{22}
Absolute no. at reproduction	$NW_{11}p_t^2$	$NW_{12}2p_tq_t$	$NW_{22}q_t^2$
Relative freq. at reproduction	$\frac{NW_{11}p_t^2}{N\bar{W}} = \frac{W_{11}}{\bar{W}}p_t^2$	$\frac{NW_{12}2p_tq_t}{N\bar{W}} = \frac{W_{12}}{\bar{W}}2p_tq_t$	$\frac{NW_{22}q_t^2}{N\bar{W}} = \frac{W_{22}}{\bar{W}}q_t^2$

$$p_{t+1} = \frac{W_{11}p_t^2 + W_{12}p_tq_t}{\bar{W}}$$

Can write $w_{11} = W_{11}/\bar{W}$, and $w_{12} = W_{12}/\bar{W}$

See Graham Coop's population genetics notes

https://github.com/cooplabor/popgen-notes/blob/master/release_popgen_notes.pdf

Three genotypes

	A_1A_1	A_1A_2	A_2A_2
Absolute no. at birth	Np_t^2	$N2p_tq_t$	Nq_t^2
Fitnesses	W_{11}	W_{12}	W_{22}
Absolute no. at reproduction	$NW_{11}p_t^2$	$NW_{12}2p_tq_t$	$NW_{22}q_t^2$
Relative freq. at reproduction	$\frac{NW_{11}p_t^2}{N\bar{W}} = \frac{W_{11}}{\bar{W}}p_t^2$	$\frac{NW_{12}2p_tq_t}{N\bar{W}} = \frac{W_{12}}{\bar{W}}2p_tq_t$	$\frac{NW_{22}q_t^2}{N\bar{W}} = \frac{W_{22}}{\bar{W}}q_t^2$

$$\Delta p_t = p_{t+1} - p_t = \frac{w_{11}p_t^2 + w_{12}p_tq_t}{\bar{w}} - p_t$$

where $w_{11} = W_{11}/\bar{W}$, and $w_{12} = W_{12}/\bar{W}$

See Graham Coop's population genetics notes

https://github.com/cooplabor/popgen-notes/blob/master/release_popgen_notes.pdf

Three genotypes

	A_1A_1	A_1A_2	A_2A_2
Absolute no. at birth	Np_t^2	$N2p_tq_t$	Nq_t^2
Fitnesses	W_{11}	W_{12}	W_{22}
Absolute no. at reproduction	$NW_{11}p_t^2$	$NW_{12}2p_tq_t$	$NW_{22}q_t^2$
Relative freq. at reproduction	$\frac{NW_{11}p_t^2}{N\bar{W}} = \frac{W_{11}}{\bar{W}}p_t^2$	$\frac{NW_{12}2p_tq_t}{N\bar{W}} = \frac{W_{12}}{\bar{W}}2p_tq_t$	$\frac{NW_{22}q_t^2}{N\bar{W}} = \frac{W_{22}}{\bar{W}}q_t^2$

Rewriting this equation in terms of the marginal fitnesses: $\bar{w}_1 = w_{11}p_t + w_{12}q_t$

$$\bar{w}_2 = w_{12}p_t + w_{22}q_t$$

$$\Delta p_t = \frac{(\bar{w}_1 - \bar{w}_2)}{\bar{w}} p_t q_t$$

See Graham Coop's population genetics notes

https://github.com/cooplab/popgen-notes/blob/master/release_popgen_notes.pdf

Three genotypes

genotype	A_1A_1	A_1A_2	A_2A_2
absolute fitness	W_{11}	$\geq W_{12} \geq$	W_{22}
relative fitness (generic)	$w_{11} = W_{11}/W_{11}$	$w_{12} = W_{12}/W_{11}$	$w_{22} = W_{22}/W_{11}$
relative fitness (specific)	1	$1 - sh$	$1 - s.$

s difference in relative fitness between the two homozygotes

h dominance coefficient

where

$$\Delta p_t = \frac{p_t h s + q_t s (1 - h)}{\bar{w}} p_t q_t,$$

$$\bar{w}_t = 1 - 2p_t q_t s h - q_t^2 s.$$



Deleterious mutations will be eliminated from the population

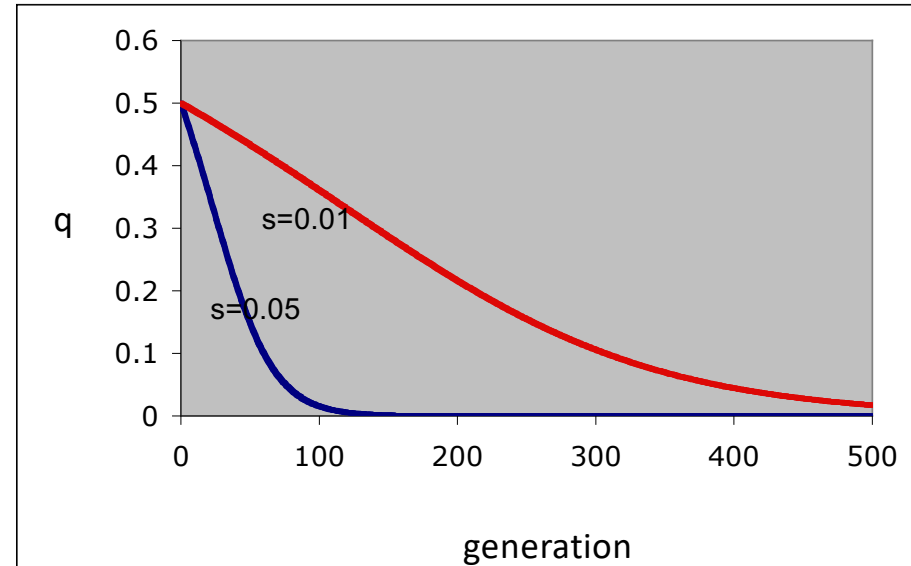
Completely dominant deleterious mutation (so $h=1$)

A_1A_1 A_1A_2 A_2A_2

1 1-s 1-s

q frequency of A_2

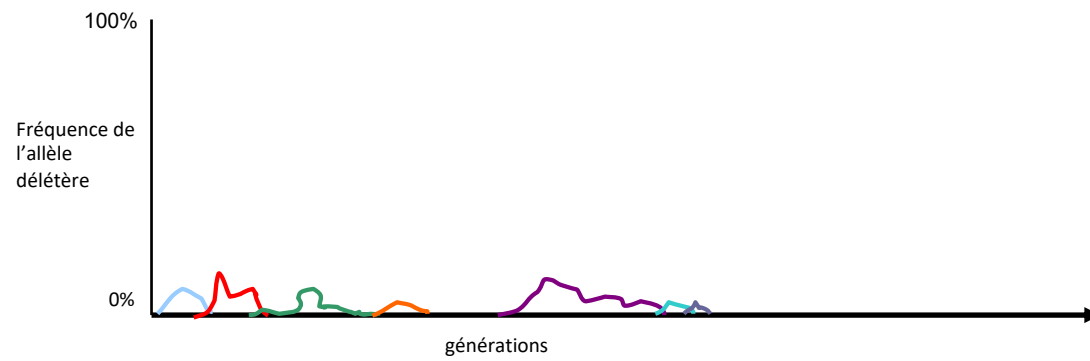
$$\Delta q = -\Delta p = \frac{-sq(1-q)^2}{1-sq(2-q)}$$



The deleterious mutations will be eliminated from the population

Two extreme genetic architectures of diseases

- Mendelian. Mutations in a **single gene sufficient** to produce a fairly predictable disease phenotype (e.g., cystic fibrosis, Huntington's disease and many others)
- Complex (polygenic). Disease phenotype results from **many loci**, as well as **environmental factors** (e.g., diabetes, autoimmune diseases, most psychiatric disorders, etc.)



“Mutation-Selection Balance”

Selection



Mutation

Mutation-Selection Balance

A disease allele will be at a low frequency (q) in the population that is balance between mutation, which generates deleterious mutations, and selection that eliminates them. Let u be the mutation rate to disease alleles.

- If the heterozygote is deleterious:
 - Frequency: $q \approx u/hs$
 - Incidence: $\sim 2q \approx 2u/hs$
- If the heterozygote is *completely* recessive:
 - Frequency: $q \approx \sqrt{u/s}$
 - Incidence: $q^2 \approx u/s$

Spinal muscular atrophy

Caused by loss of function recessive mutations A_2 :

A_1A_1	A_1A_2	A_2A_2
1	1	1-s
1	1	0.1

- ❖ The mutation rate to loss of function alleles is $u = 0.00014$ per generation and $s = 0.9$.
- ❖ The frequency of A_2 is $1:80$ ($=0.0125$), so $\sim 1:40$ people are carriers, and the incidence of the disease is $\sim 1:6000$

Expect disease mutations at frequency = $\sqrt{u/s} = \sqrt{.00014/.9} = 0.0125$
or $1/80$

genotype	A_1A_1	A_1A_2	A_2A_2
absolute fitness	W_{11}	$\geq W_{12} \geq$	W_{22}
relative fitness	$w_{11} = 1$	$w_{12} = 1 - sh$	$w_{22} = 1 - s.$

q frequency of A_2
 μ mutation rate from A_1 to A_2
 A_2 rare

Change due to selection

$$\Delta_S q_t = \frac{\bar{w}_2 - \bar{w}_1}{\bar{w}} p_t q_t \approx -hsq_t$$

$$q_t = q_0(1 - hs)^t$$

Change due to mutation

$$q' = \mu p_t + q_t = \mu(1 - q_t) + q_t$$

$$\Delta_M q_t = q' - q_t = \mu$$

At equilibrium between mutation and selection

$$\Delta_M q_t + \Delta_S q_t = 0$$

$$q_t = \frac{\mu}{hs}$$

For the case with a changing population size,
 see Simons et al. (2014) Nature Genetics

Hutchinson-Gilford Progeria Syndrome

Caused by loss of function dominant mutations A_2 in LMNA gene:

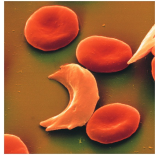
A_1A_1	A_1A_2	A_2A_2
1	1	1-s
1	0	0

❖ The incidence of the disease is 1 in 4 million, i.e., 0.25×10^{-7}

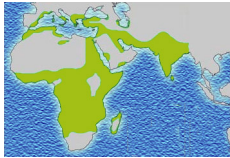
❖ Recall that incidence = $2q = 2u/hs$. Here, $hs=1$ and the mutation rate μ is 1.25×10^{-8} per base pair in humans

=> Have $\mu = Lu$ where L is the mutational target size.

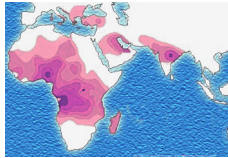
Suggests that $L = 2q/2\mu = 10$ base pairs.



The human sickle-cell hemoglobin polymorphism



Distribution of malaria



of the sickle allele

S is a mutation in β -hemoglobin (there are others)

S allele persists in some populations at the frequency of 10% to 20% despite strong SS disadvantage

AS heterozygotes largely asymptomatic; SS homozygotes have severe anemia (sickle-cell disease)

AS are less susceptible to malaria than AA

Genotype	AA	AS	SS	Freq S allele
Freq. infant	0.66	0.31	0.03	0.186
Freq. adult	0.61	0.38	0.007	0.198
Fitness	0.93	1.23	0.24	
Rel. fitness	0.76	1	0.18	
	(1-s)	1	(1-t)	$q^* = s/(s+t)=0.226$

Pour les maladies récessives
létales,


la fréquence attendue est de \sqrt{u}

soit $\sim 10^{-4}$

11/13/2018 gnomad

gnomAD browser

This is a new version of the gnomAD browser. The old version is available at <http://gnomad-old.broadinstitute.org>



gnomAD


genome aggregation database

Examples - Gene: PCSK9, Variant: 1-55516888-G-GA

The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed [here](#).

All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use [here](#). Sign up for our mailing list for future release announcements [here](#).



<http://gnomad.broadinstitute.org/> 1/1

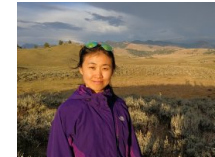
Équilibre mutation-dérive-sélection

417 allèles pour lesquels $h=0$, $s=1$

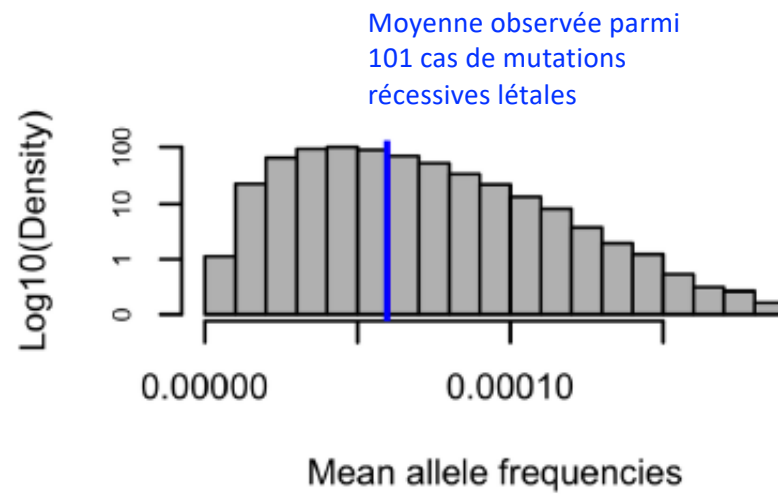
Quatre types de mutations: CpG Ti, CpG Tv, non-CpG Ti, non CpG Tv



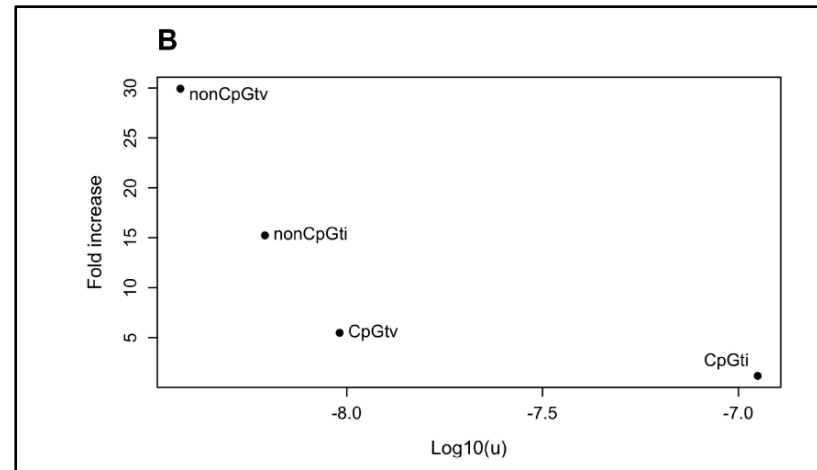
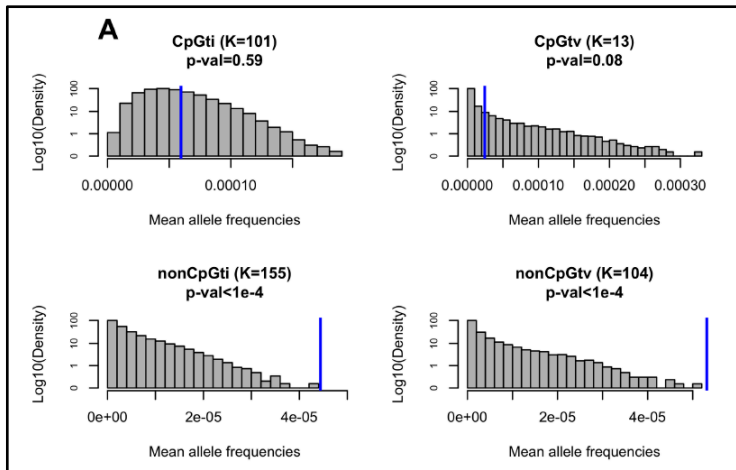
Eduardo Amorim



Ziyue Gao



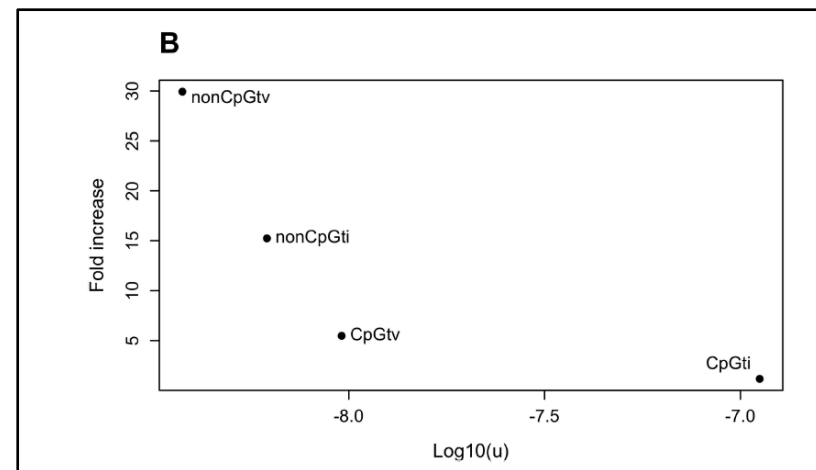
Pour les transitions CpG
 p bilatérale = 0,59



Possible explanations

- (i) Errors in the identification of causal variants
- (ii) Misspecification of the demographic model
- (iii) Misspecification of the mutation rate
- (iv) Heterozygote advantage
- (v) Low penetrance of Mendelian disease alleles (modifiers)

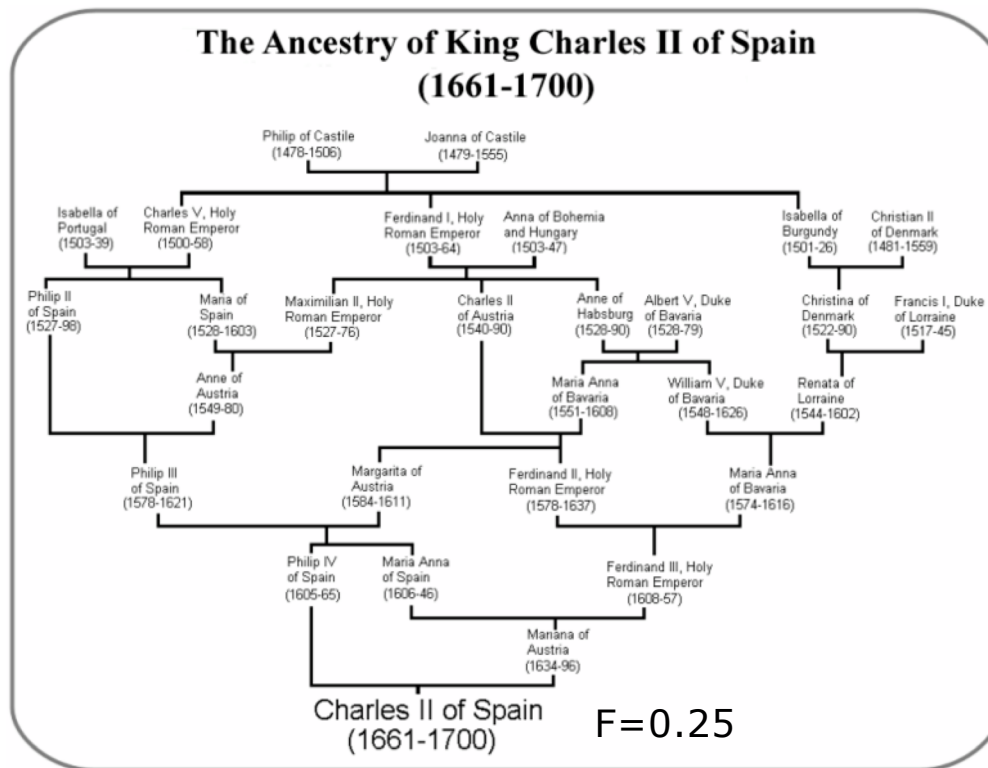
(vi) ascertainment bias in their discovery



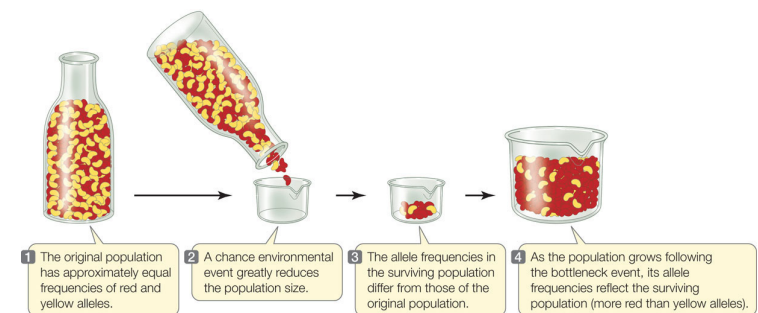
In summary

- The mutation rate per generation is a key parameter determining the frequency at which we expect to find deleterious mutations in populations and hence of the incidence of disease
- In the absence of inbreeding, the incidence of disease is $2q$ for a dominant disease and q^2 for a completely recessive one.

The incidence of recessive disease can be increased by consanguinity and bottlenecks (but there will be fewer diseases)



Heir of Spanish Habsburg dynasty



In summary

- The mutation rate per generation is a key parameter determining the frequency at which we expect to find deleterious mutations in populations and hence of the incidence of disease
- If there are effects in heterozygotes, even subtle, we expect to see deleterious mutations at an average frequency of approximately u/hs , where u is the mutation rate to the disease allele and hs the fitness effect in heterozygotes
- There are good reasons to believe the model applies widely, and in particular it provides a good fit to Mendelian disease alleles at highly mutable sites (CpG transitions). For less mutable types, there may still be an appreciable ascertainment bias in discovery.
- But a subset of cases of deleterious alleles could also persist due to balancing selection. What fraction is unknown.

Table 1 | Metrics of gene essentiality for human population sequence data

Score	Residual variation intolerance score (RVIS)	EvoTol	Missense Z-score	Probability of haplo-insufficiency (Phi)	Probability of loss-of-function intolerance (pLI)	LoFoot	Selective effects for heterozygous protein-truncating variants (s _h)
Sample size	6,503 exomes (ESP), updated to 60,706 on their website	1 million variants from dbSNP	1,000 Genomes Project data to calculate the baseline mutational rate of de novo mutation; ESP for the observed variant counts	10,500 exomes (ESP + UK10K + ClinVar) autosomes only, updated for 60,706 exomes	60,706 exomes from ExAC	60,706 exomes from ExAC	60,706 exomes from ExAC
Method	Residuals derived from the linear regression of the number of common functional variants on the total number of variants	Residuals derived from the linear regression of the number of common functional variants on the total number of variants	Z-Score to quantify the difference between the observed missense variants and the expectation based on a mutation model	Posterior probabilities from a two-state Poisson mixture model	Posterior probabilities from a three-state Poisson mixture model	Heuristic, builds on EvoTol ²⁸ and mutation model from the Missense Z-score ²⁸	Bayesian estimation of the selection coefficient on heterozygous loss of function
Distinctive feature	Introduces the concept of applying human population genetics to assess function, pathogenicity and essentiality	Combines intraspecies and interspecies information in a similar framework to RVIS	Uses a neutral mutation model as a baseline	Uses a mixture model to estimate haplo-insufficiency	Unprecedented sample size	Non-parametric combination of mutation rates, functional predictions and variation data	Direct estimation of the selection acting on the gene
AF filter	>0.1%	NA	Singleton	<1%	<0.1%	NA	<0.1%
Variant class	SNVs only; missense, stop-gain, stop-loss, prediction of splice-site effects as provided by ESP	SNVs only; considered to be functional based on Pathway prediction	Missense	Stop-gain or frameshift (StopLEI)	Stop-gain, splice, frameshift (VEP-LOFTEE)	Stop-gain, splice, frameshift (VEP-LOFTEE)	Stop-gain, splice (VEP-LOFTEE)
Reported precision against OMIM genes (ROC/AUC)	58% against OMIM disease genes, 78% against OMIM haploinsufficient genes or carrying de novo variants, 80% against de novo OMIM haploinsufficient genes	ROC shown, no AUC reported, outperforms RVIS	87% against de novo OMIM haploinsufficient genes	76% against de novo OMIM haploinsufficient genes	Not reported	86% against de novo OMIM haploinsufficient genes	OMIM not reported, 33% to identify inheritance mode on 450 genes involved in clinical disease
Functional and clinical correlates	High fraction of developmental genes in the top quartile	Nuclear receptors and metabolic genes are intolerant to predicted damaging variants	Higher Z-score in genes with de novo loss-of-function mutations in autism spectrum disorders or intellectual disability cases	Fewer paralogs; more central in protein-protein interaction networks; enrichment in genes with lethal mouse knockout phenotype; more likely to be in protein complexes; enriched in genes for which loss of function leads to loss of cell viability; more conserved	Highly expressed genes; depleted in eQTL hits; enriched in CWA5 hits; more protein-protein interactions; gene-set enrichment analysis results (lipidome, ribosome and proteasome); 50% of assessed human orthologues of mouse genes with conditional lethal knockout phenotype are essential	Enriched in genes expressed preferentially in the brain	Developmental pathways and transcriptional regulators are enriched in high s _h values. Positive correlation with protein-protein interaction count
Interpretation	Lower value = more intolerant	Lower value = more intolerant	Higher value = more intolerant	Higher value = more intolerant	Higher value = more intolerant	Lower value = more intolerant	Higher value = more intolerant
Thresholds for pathogenicity	Standard normal; the authors evaluate the lowest quartile	Percentiles; the authors evaluate the lowest quartile	Standard normal	Binomial on [0,1]; the authors evaluate the set <0.95	Binomial on [0,1]; the authors evaluate the set <0.9	Percentile; the authors evaluate the lowest quartile	Inverse Gaussian
Refs	19	20	21	23	15	22	24

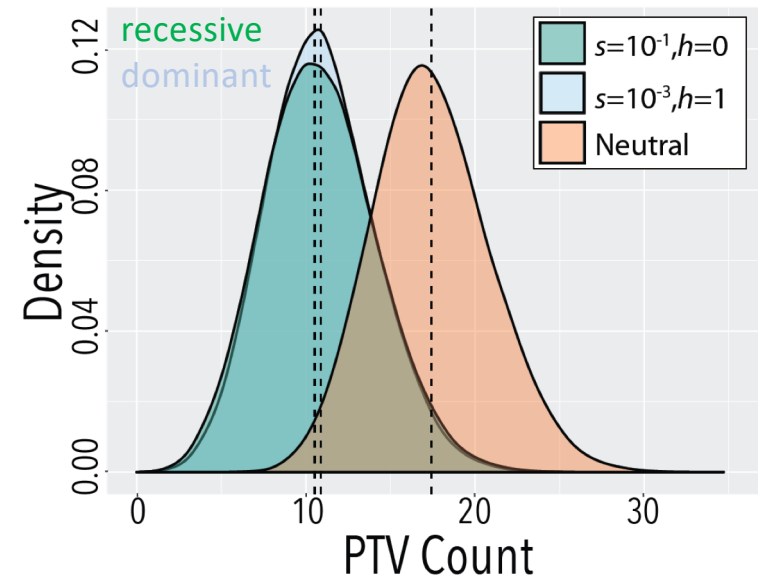
AUC, area under the curve; AF, allele frequency; CoLaus, Cohort Study of Lausanne; dbSNP, The Single Nucleotide Polymorphism database; eQTL, expression quantitative trait locus; ESP, Exome Sequencing Project; ExAC, Exome Aggregation Consortium; Pathway, functional analysis through hidden Markov models; CWA5, genome-wide association study; NA, not applicable; OMIM, Online Mendelian Inheritance in Man; ROC, receiver operating characteristic; SNV, single-nucleotide variant; UK10K, The United Kingdom 10,000 Genomes Project; VEP-LOFTEE, Loss-Of-Function Transcript Effect Estimator plugin for the Ensembl Variant Effect Predictor (VEP).

<https://www.nature.com/articles/nrg.2017.75#t1>

Measures of « intolerance to mutation »



Zach Fuller



= # of disrupting mutations observed

These measures reflect h s, the strength of selection acting on heterozygotes, not h (dominance) and s separately

Fuller et al. 2018 BioRxiv