Learning from re-sequencing data: what to do when the $1000 genome arrives?

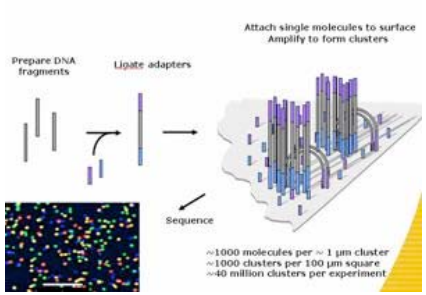
Shamil Sunyaev

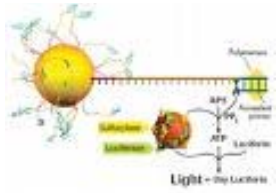
Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

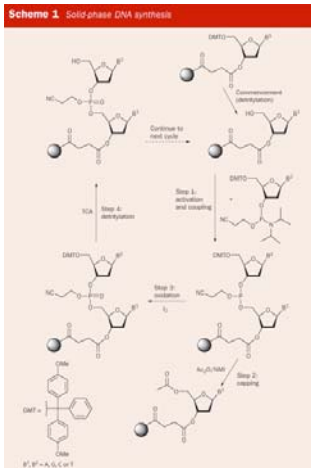
Harvard-M.I.T. Health Sciences & Technology Division

Genomes of many well-phenotyped individuals will be available soon

New sequencing technologies





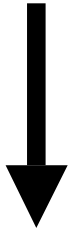


New ways to collect clinical populations

Will this development revolutionize search for genes underlying human phenotypes?

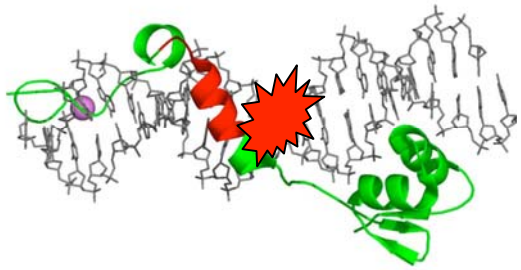
Our approach:

Learn from existing sequencing data



Simulate large sequencing studies

Functional genetic variation



Coding

Non-coding

1) Mutations in protein coding regions

2) Mutations in non-coding regions

Exon capture technology

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

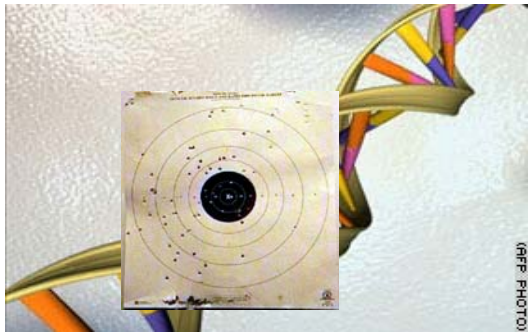
Antagonistic pleiotropy (Trade Off)

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

Mutation/selection balance

Theory 2:

Common diseases are due to multiple rare deleterious alleles in mutation-selection balance



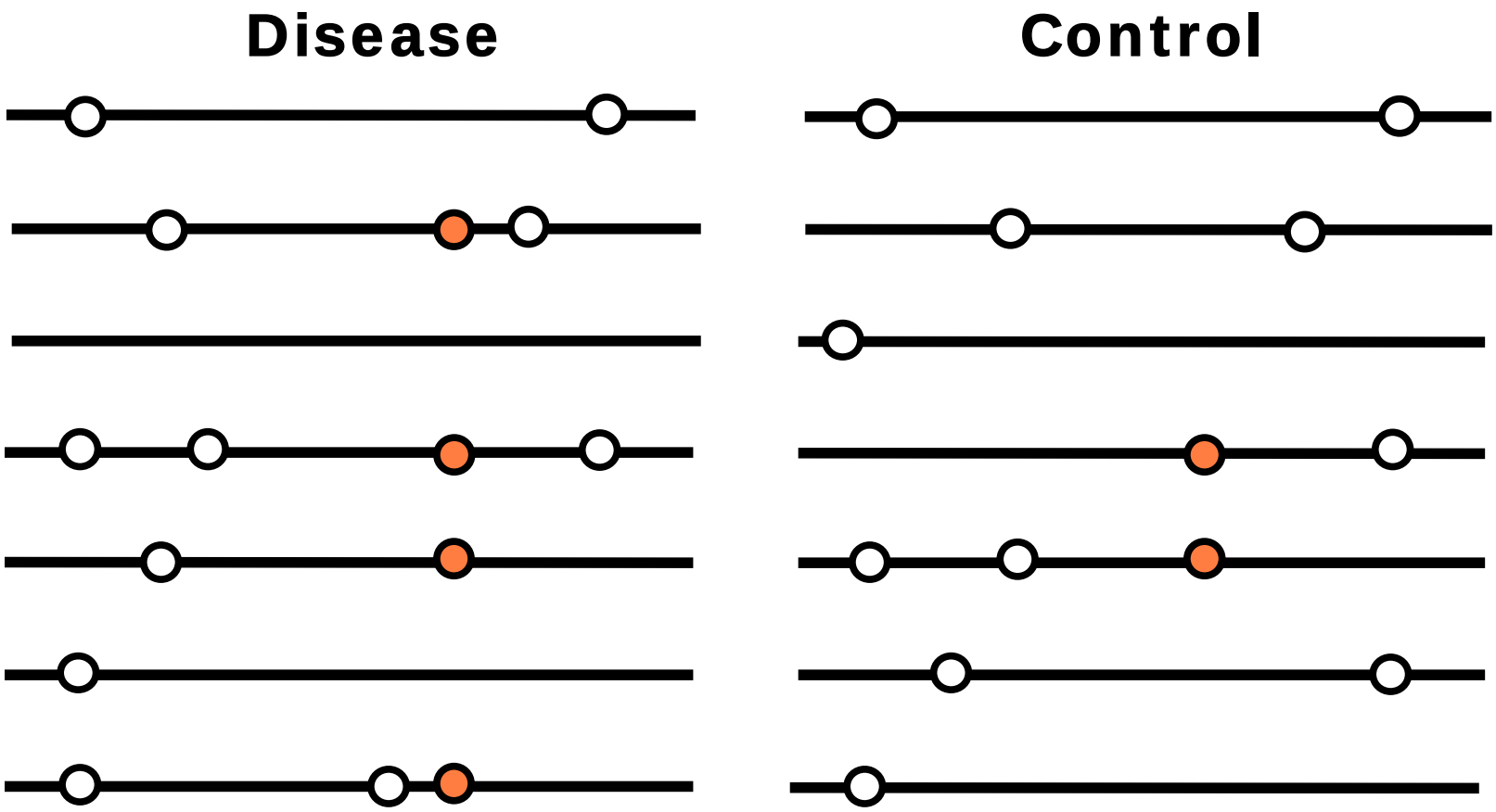
- **Weak selection**
- **High mutation rate**

CURRENT ESTIMATE:

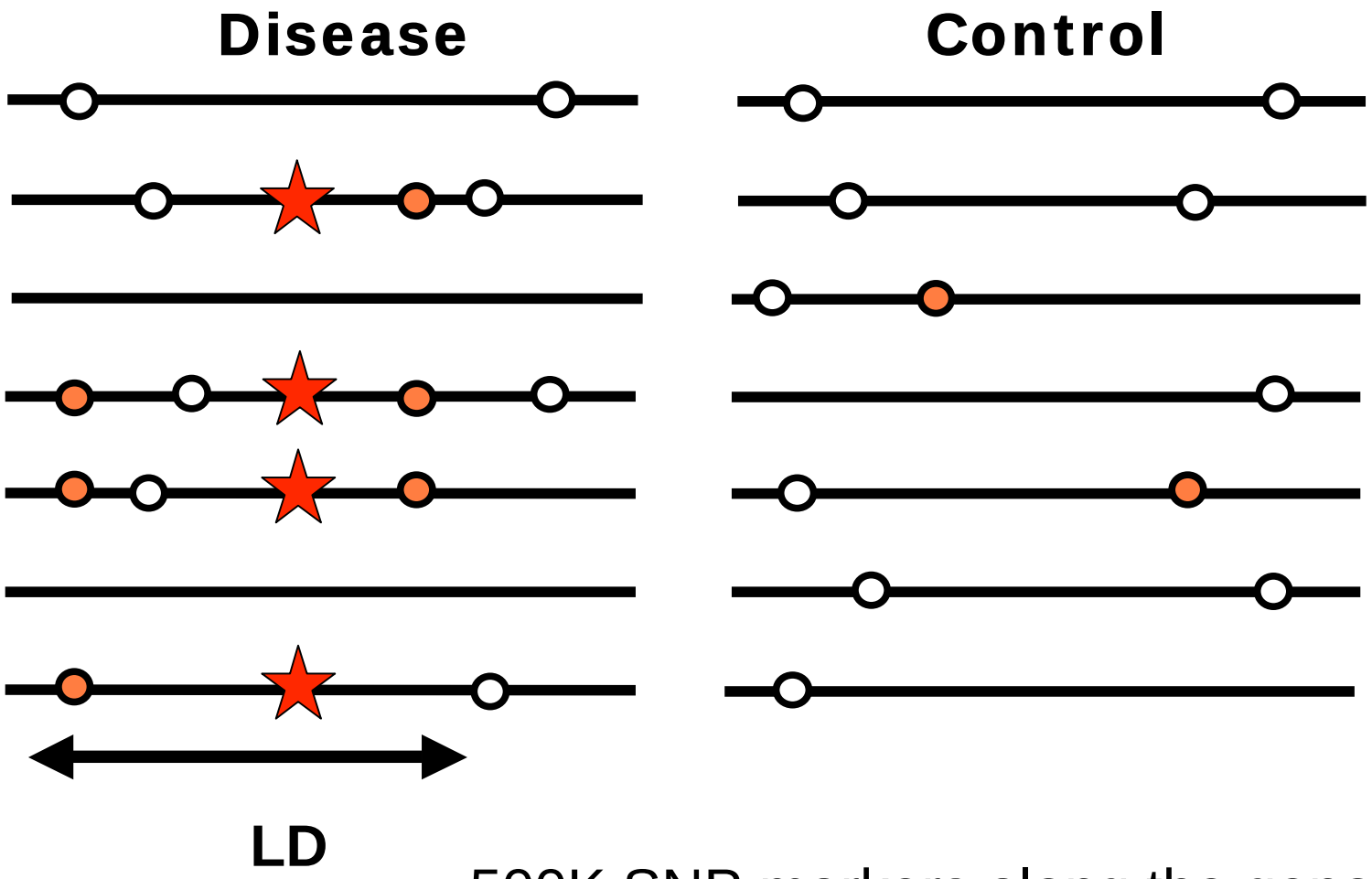
~100 new mutations per genome
~1-2 new amino acid changes per genome

Examples: LDL-C, HDL-C, Triglyceride, Colorectal adenomas

Association studies



Genome Wide Association Studies (GWAS)

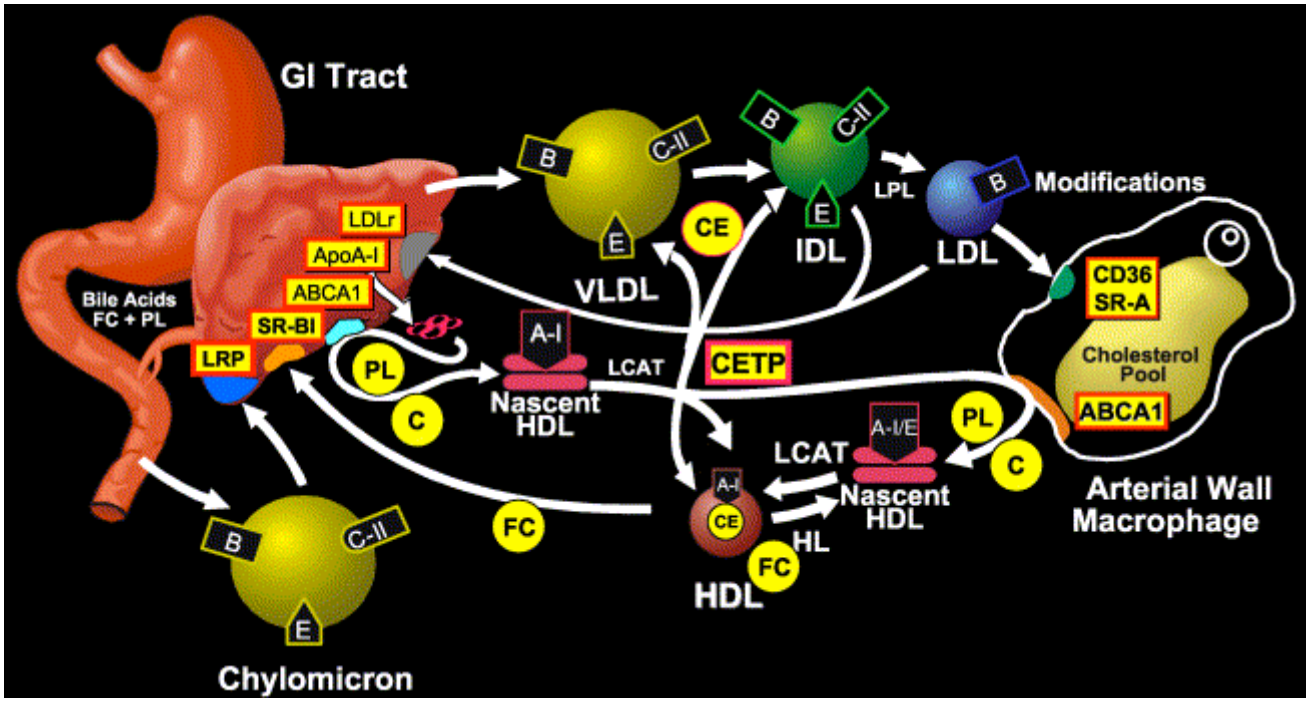


500K SNP markers along the genome

Lessons from Genome-Wide Association Studies (GWAS)

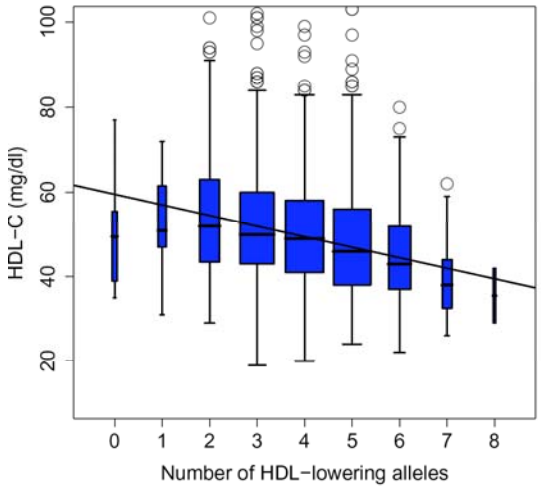
- **Some variants can be identified reproducibly**
- **~10,000 of individuals provide sufficient power to detect SNPs**
- **Some variants make sense, while most look highly surprising**

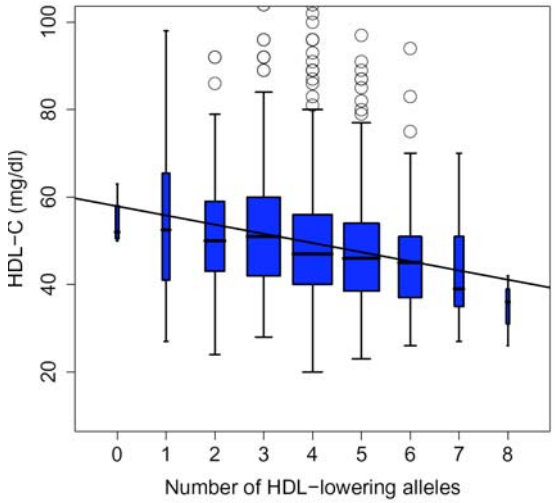
- **In many cases effects are very small**
- **Relative risk is generally very small**
- **Very small fraction of heritable variation can be explained!**

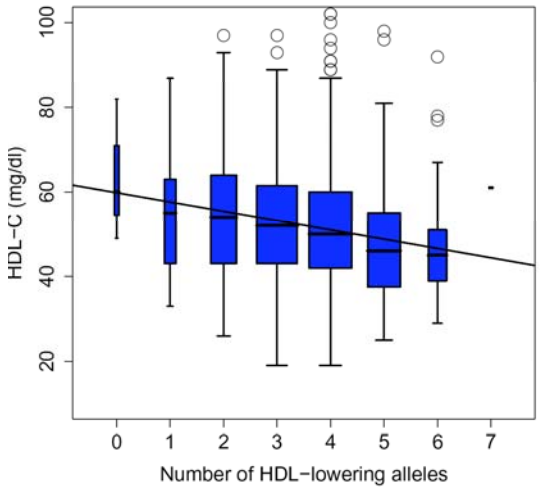


Adopted from Brewer et al., 2003

Effect of four SNPs on HDL-C







3 out of 4 SNPs are non-coding

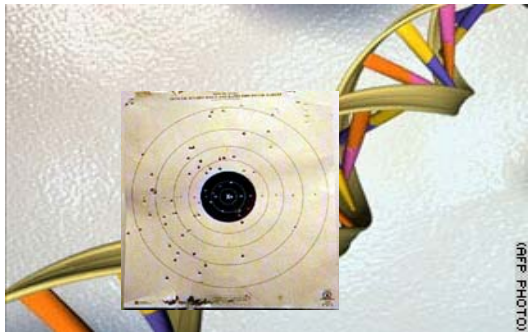
Only 2.2% of variance explained

Spirin, Schmidt et al., *Am. J. Hum. Genet.* 2007

Mutation/selection balance

Theory 2:

Common diseases are due to multiple rare deleterious alleles in mutation-selection balance



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

CURRENT ESTIMATE:

~100 new mutations per genome
~1-2 new amino acid changes per genome

Examples: LDL-C, HDL-C, Colorectal adenomas

Effect of new missense mutations

Effect of new mutation may range from lethal, to neutral, to slightly beneficial

NO DELETERIOUS POLYMORPHISM

LOTS OF DELETERIOUS POLYMORPHISM

Mutation rate model

- Mutations causing Mendelian diseases
- Human-chimpanzee divergence
- Systematic re-sequencing datasets

Mutation model

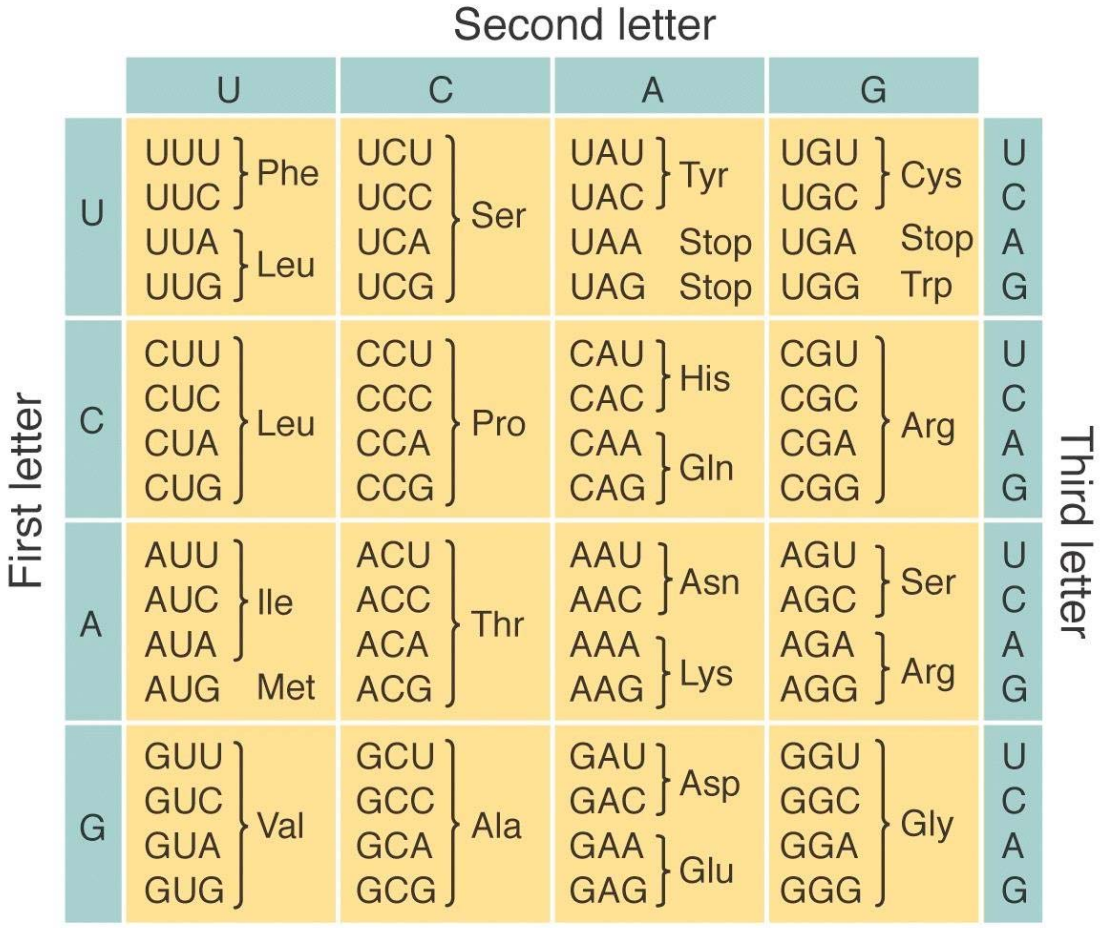
Human	ACCT**TGC**AAAT
Chimpanzee	ACCT**TAC**AAAT
Baboon	ACCT**TAC**AAAT

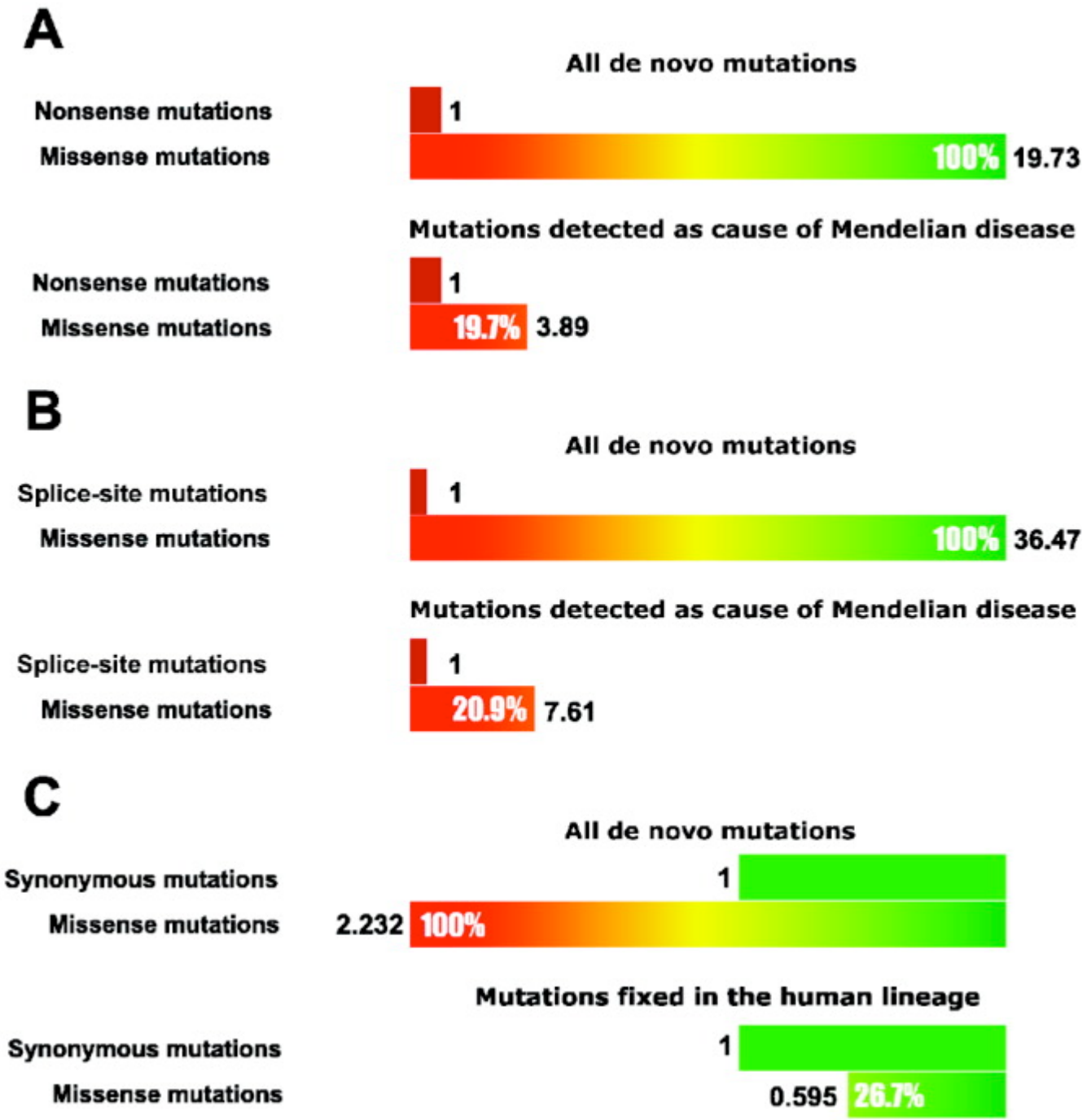
Prob(TAC->TGC) ≠ Prob(TGC->TAC)

Prob(XY_1Z->XY_2Z) 64x3 matrix

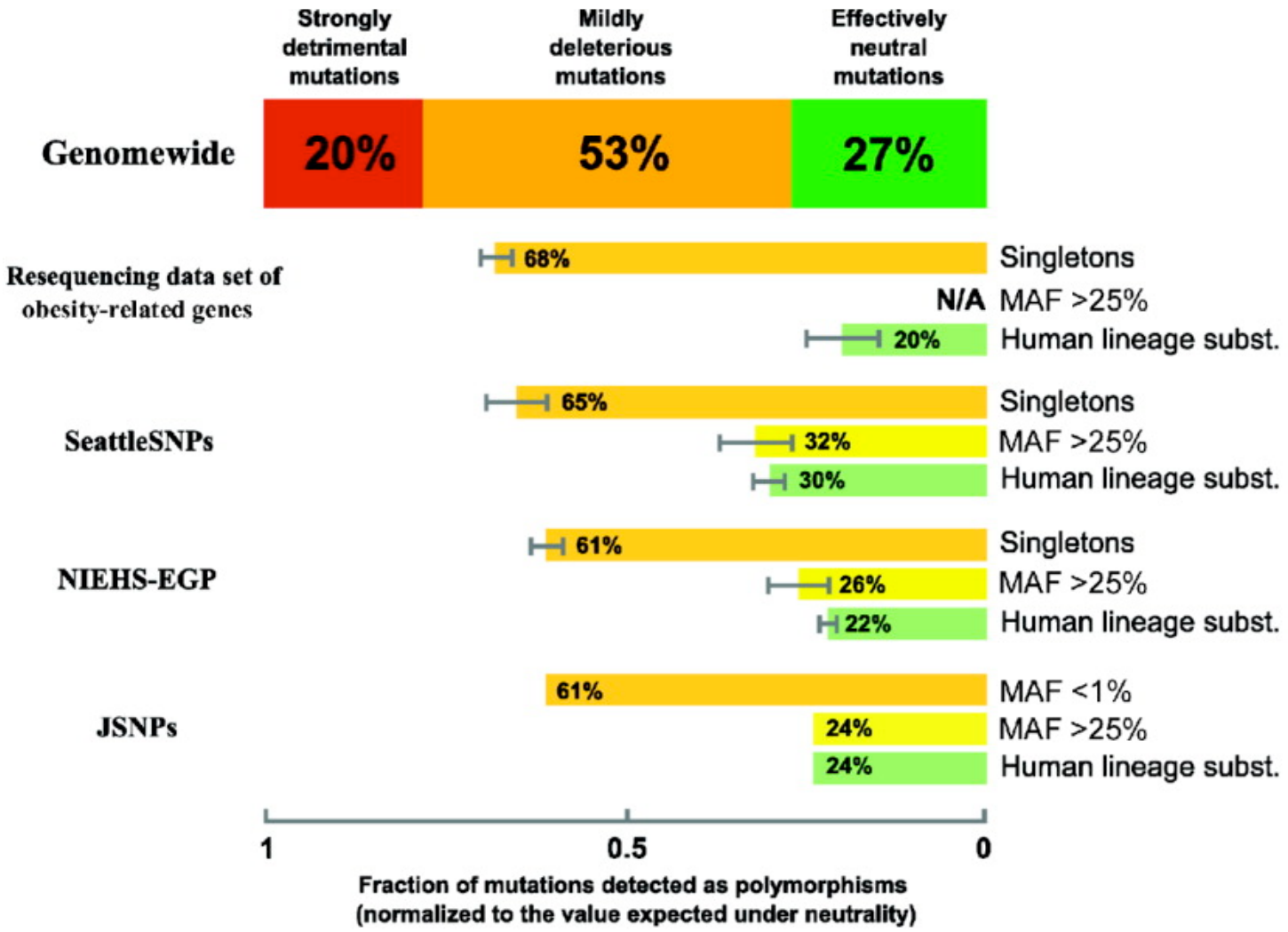
Kryukov et al., *Am. J. Hum. Genet.* 2007

Effect of mutations: protein coding regions




A
All de novo mutations
Nonsense mutations
Missense mutations
1
100%
19.73
Mutations detected as cause of Mendelian disease
Nonsense mutations
Missense mutations
1
19.7%
3.89
B
All de novo mutations
Splice-site mutations
Missense mutations
1
100%
36.47
Mutations detected as cause of Mendelian disease
Splice-site mutations
Missense mutations
1
20.9%
7.61
C
All de novo mutations
Synonymous mutations
Missense mutations
1
2.232
100%
Mutations fixed in the human lineage
Synonymous mutations
Missense mutations
1
0.595
26.7%

Effect of new missense mutations



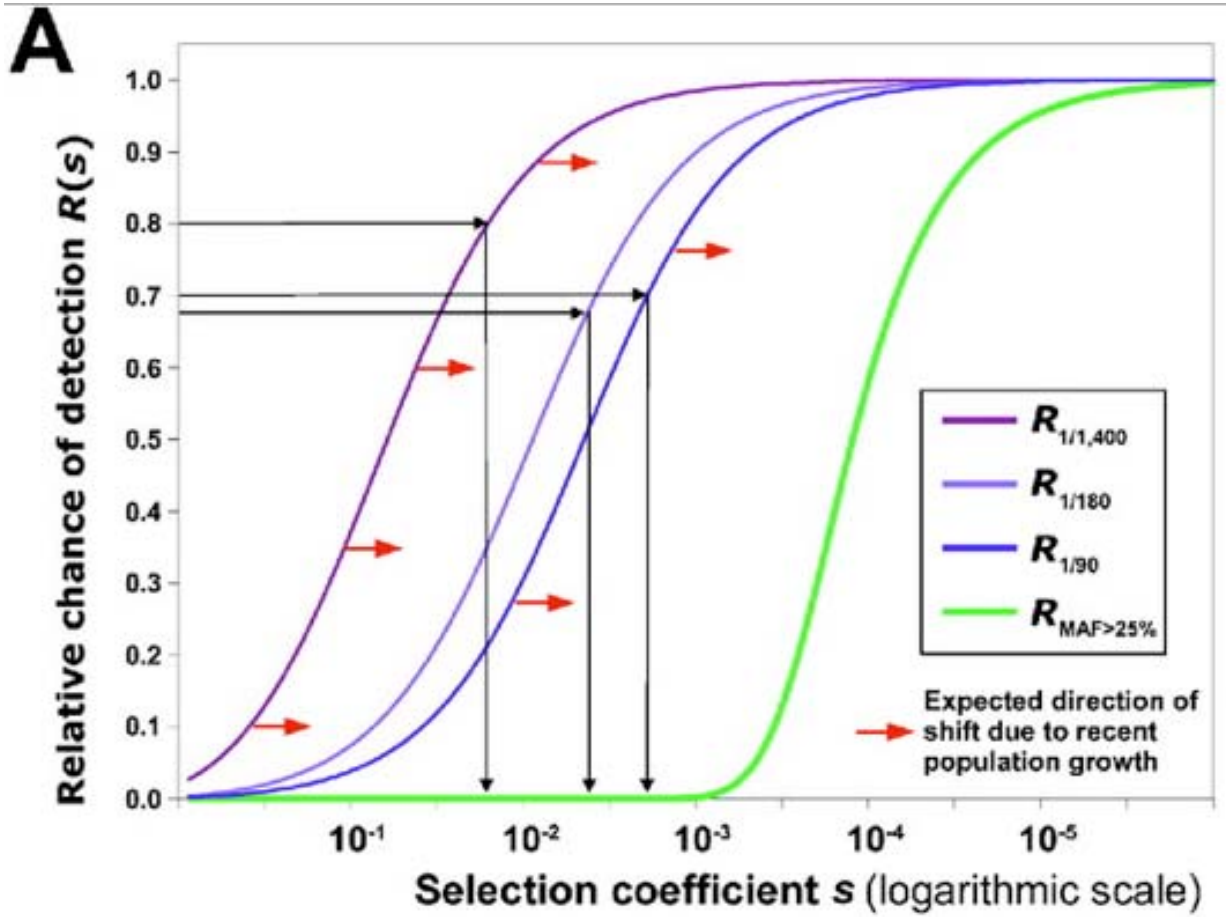
These estimates suggest that…

Table 2. Fraction of Deleterious Substitutions among Rare Missense SNPs

Set	No. of Sequenced Individuals	Percentage of Deleterious SNPs among Missense Singletons[a]
Resequencing data set of obesity-related genes	757	71 ± 8
NIEHS-EGP	90–95	64 ± 1
SeattleSNPs	46–47	52 ± 6

[a] Data are mean ± SE.

Estimating strength of selection



$$F_{singlet}(s)=\int_0^1 \frac{e^{-2N_e s(1-x)}-1}{x(1-x)(e^{-2N_e s}-1)}\left(C_1^m x(1-x)^{m-1}+C_{m-1}^m x^{m-1}(1-x)\right)\ dx$$

$$F_{MAF>0.25}(s)=\int_0^1 \left[\frac{e^{-2N_e s(1-x)}-1}{x(1-x)(e^{-2N_e s}-1)}\sum_{0.25m<j<0.75m} C_j^m x^j(1-x)^{m-j}\right]\ dx$$

We conclude that…

Combined frequency of functional (mildly deleterious) nsSNPs in the average gene is 1%

Mutation-selection balance is a feasible explanation for common human phenotypes

We conclude that…

Majority of low frequency missense variants are functional (mildly deleterious)

“Mutation enrichment” association studies are feasible

Will this development revolutionize search for genes underlying human phenotypes?

Potential: Sequencing will make every gene susceptible for genetic analysis

Most genes do not have a common functional coding variant. However, all genes have rare coding variants.

Theory:

$$\mu_{nt} = 2\text{x}10^{-8}$$

$$\mu_{gene} = 2\text{x}10^{-8}\text{x}10^{3} = 2\text{x}10^{-5}$$

$$s = 10^{-3}$$

$$f = \mu_{gene}/s = 0.02$$

Data:

Cumulative frequency of nsSNPs with frequency below 5%

EGP	*2.8%*
SeattleSNP	*2.9%*
Ahituv et al. 2007	*1.5%*

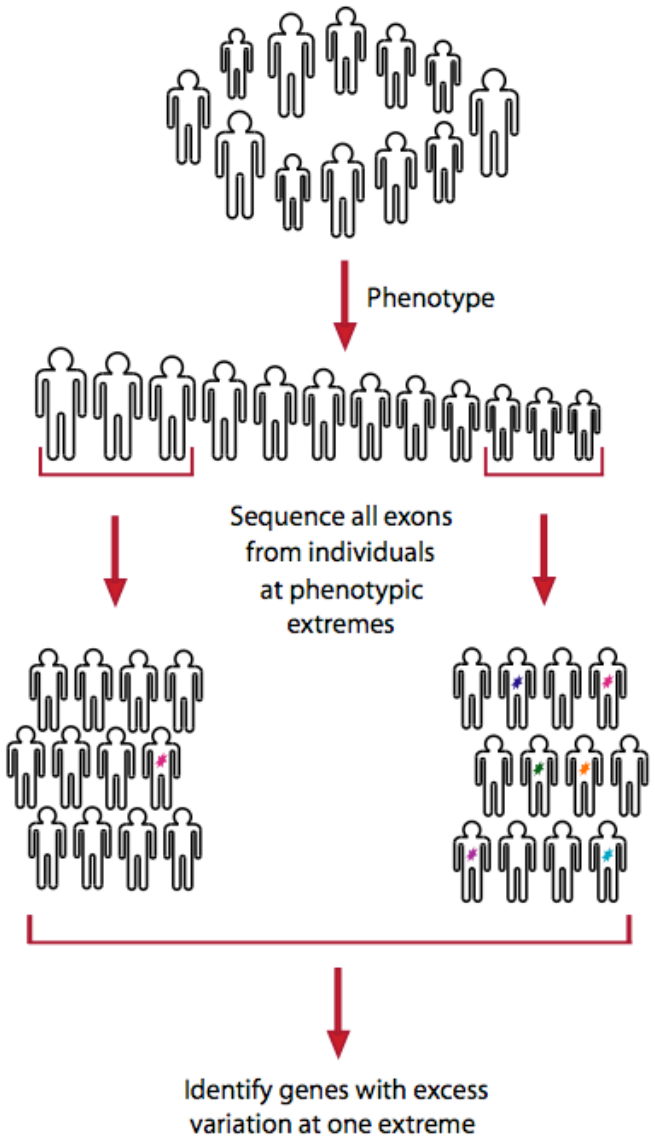
Statistical challenge!

Sequencing will uncover many low frequency variants.

1. Power to detect association with rare variants is reduced.

2. Multiple test correction becomes very stringent

Combine all non-synonymous variants in a single test



Theory:

1) Most new missense mutations are functional (*mutagenesis, population genetics, comparative genomics*)

2) Most new missense mutations are only weakly deleterious (*population genetics*)

3) Most functional missense mutations are likely to influence phenotype in the same direction (*mutagenesis, medical genetics*)

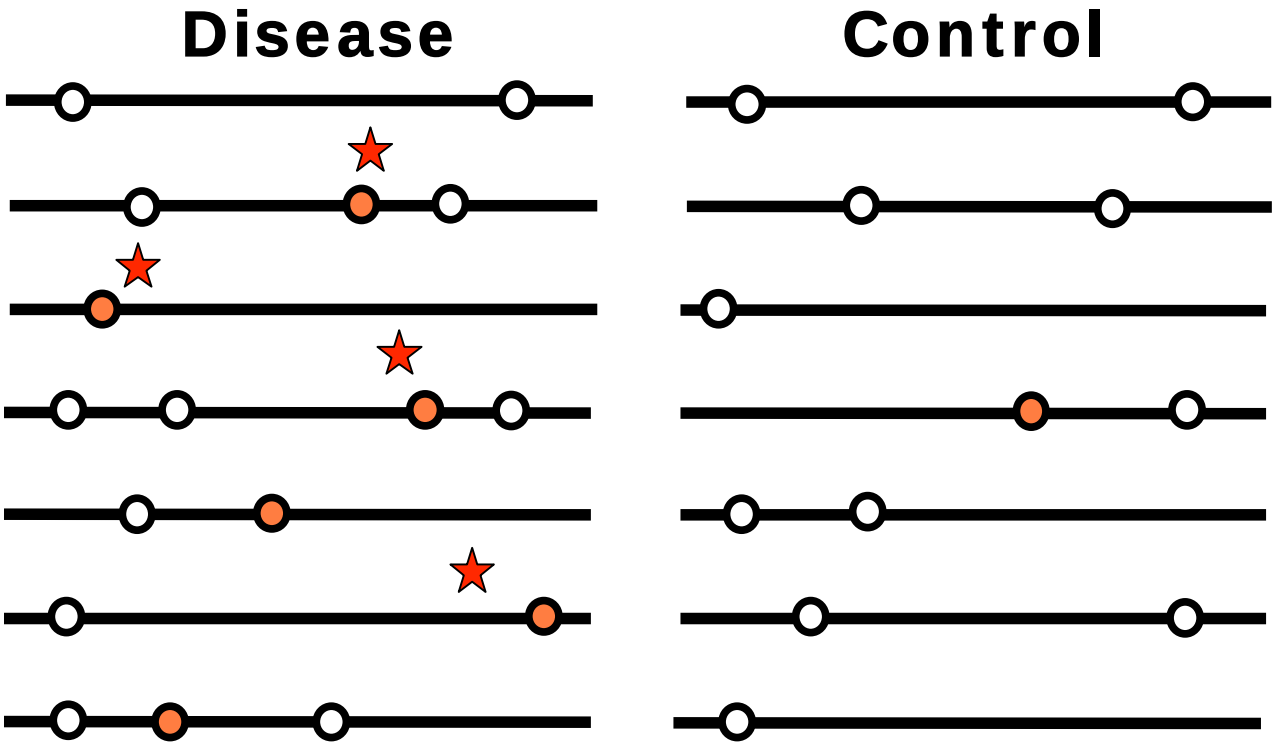
Data:

multiple candidate gene studies

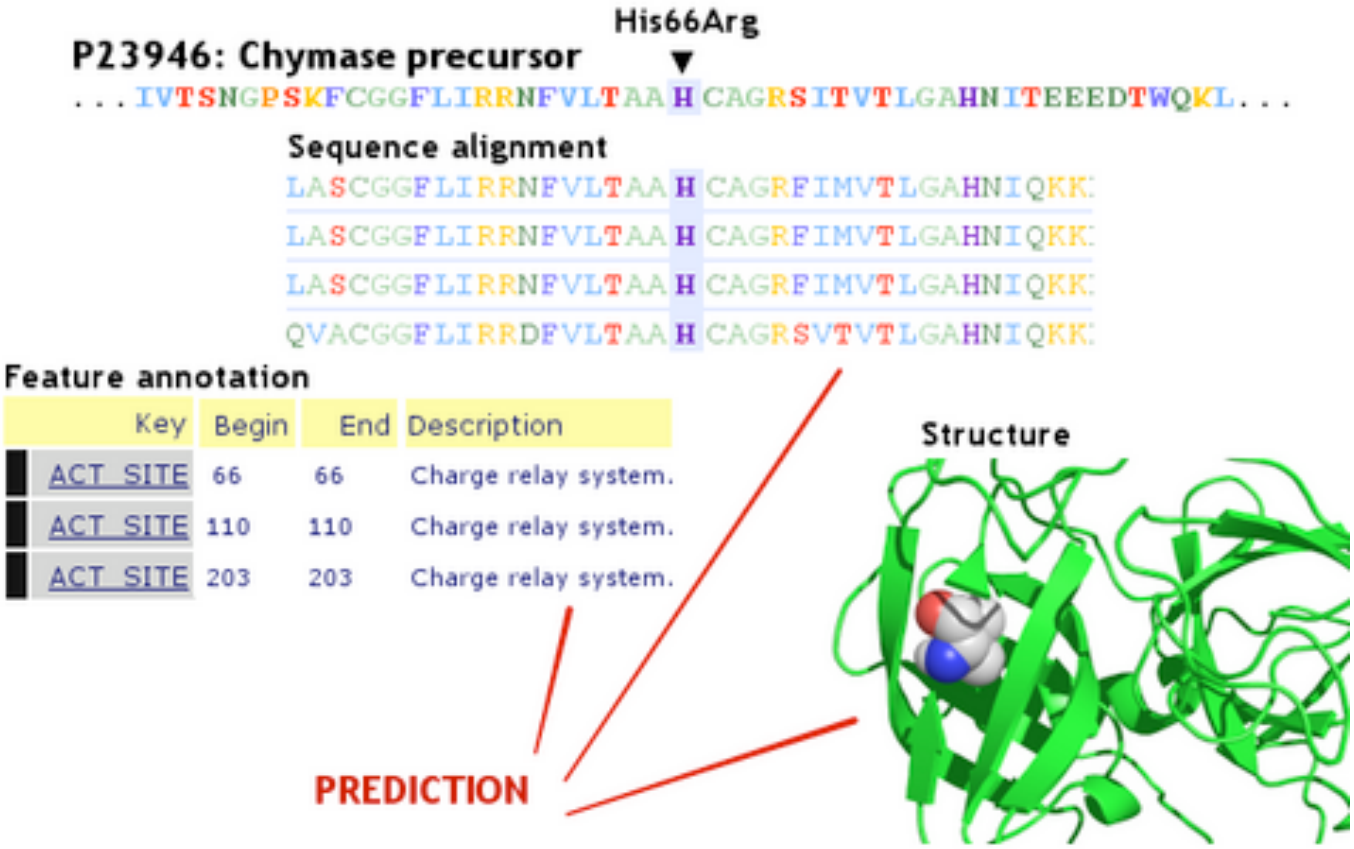
HDL-C, LDL-C, Triglycerides, BMI, Blood pressure, Colorectal adenomas

Mutation enrichment association studies

And if we can predict functional missense variants

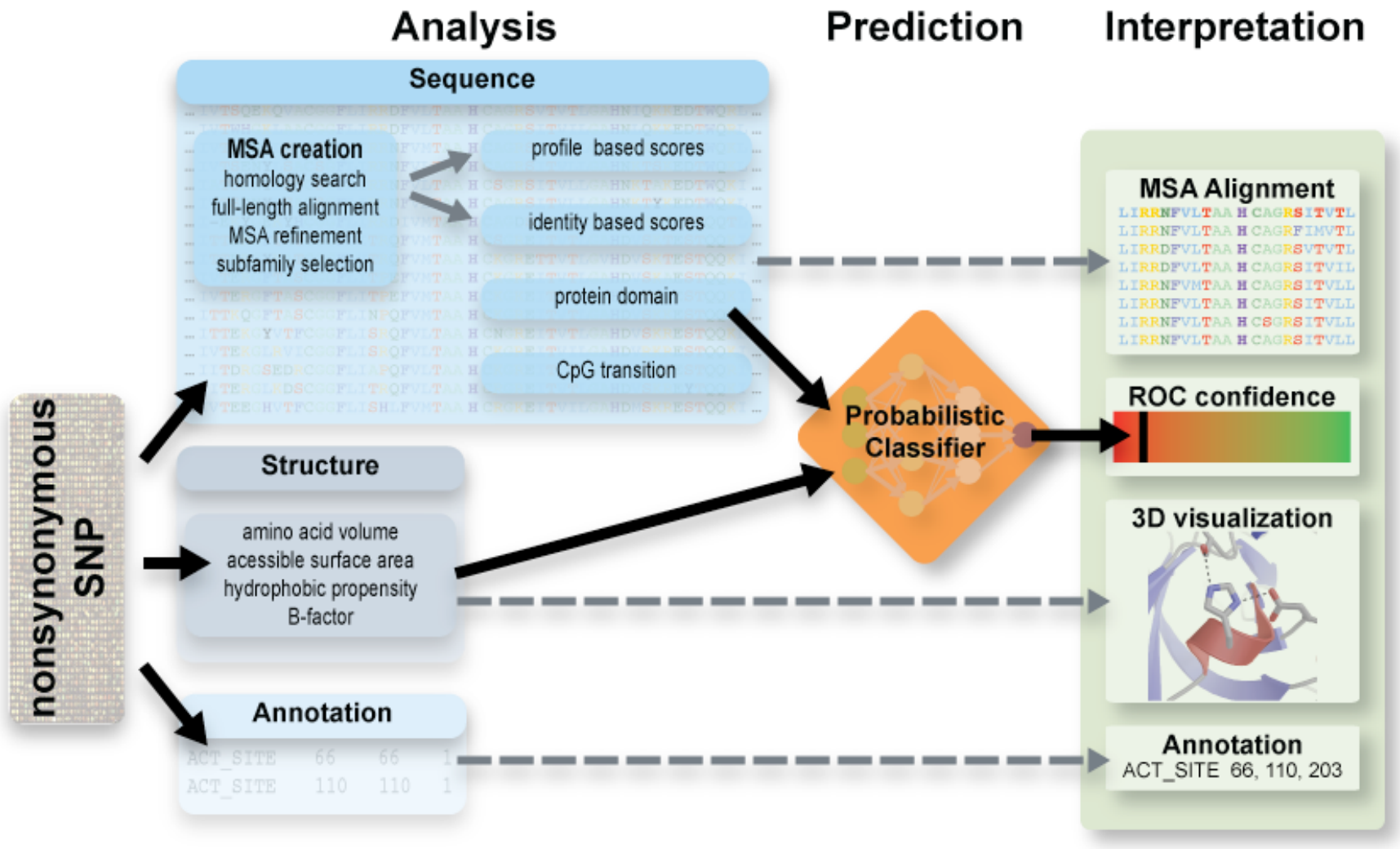


Predicting the effect of nsSNPs



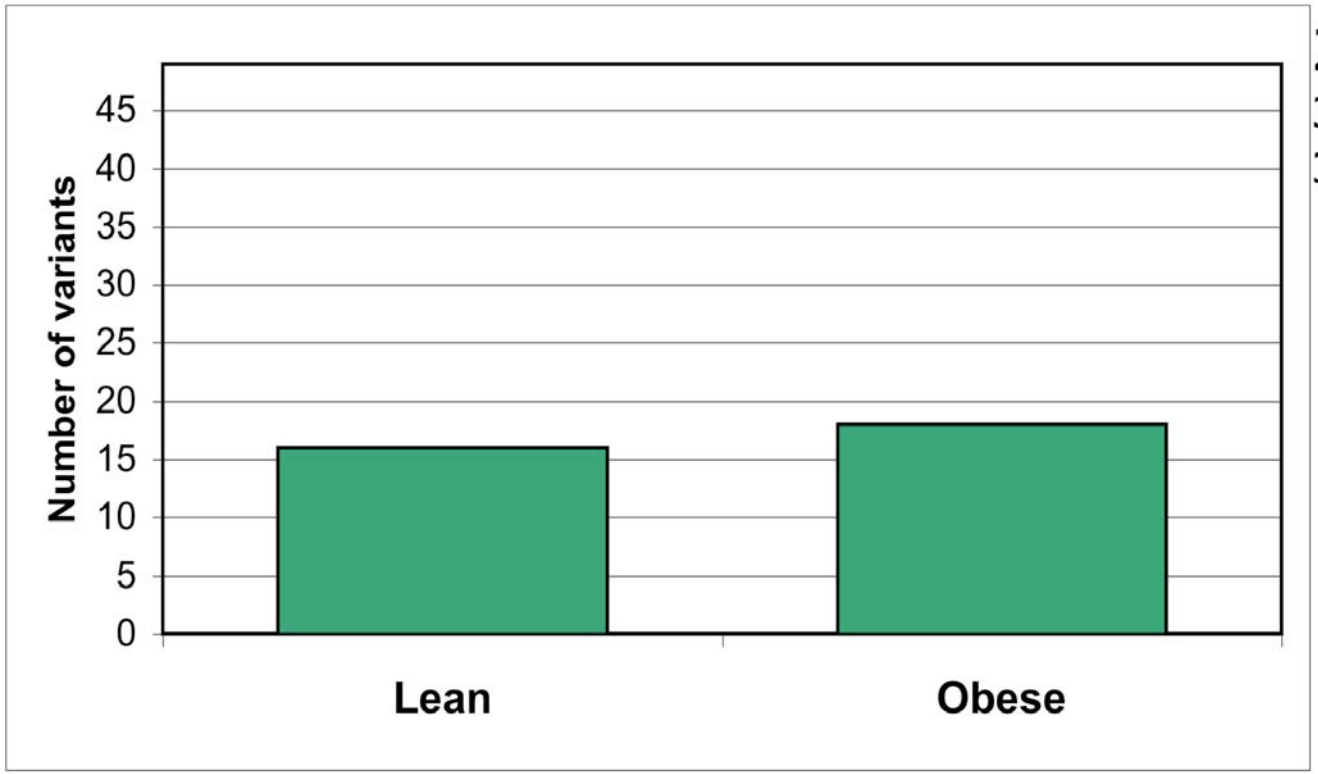
www.genetics.bwh.harvard.edu/pph

Pipeline


Analysis
Prediction
Interpretation
Sequence
MSA creation
homology search
full-length alignment
MSA refinement
subfamily selection
profile based scores
identity based scores
protein domain
CpG transition
Structure
amino acid volume
acessible surface area
hydrophobic propensity
B-factor
Annotation
ACT_SITE 66 66 1
ACT_SITE 110 110 1
nonsynonymous SNP
Probabilistic Classifier
MSA Alignment
ROC confidence
3D visualization
Annotation
ACT_SITE 66, 110, 203

Obesity

Synonymous substitutions

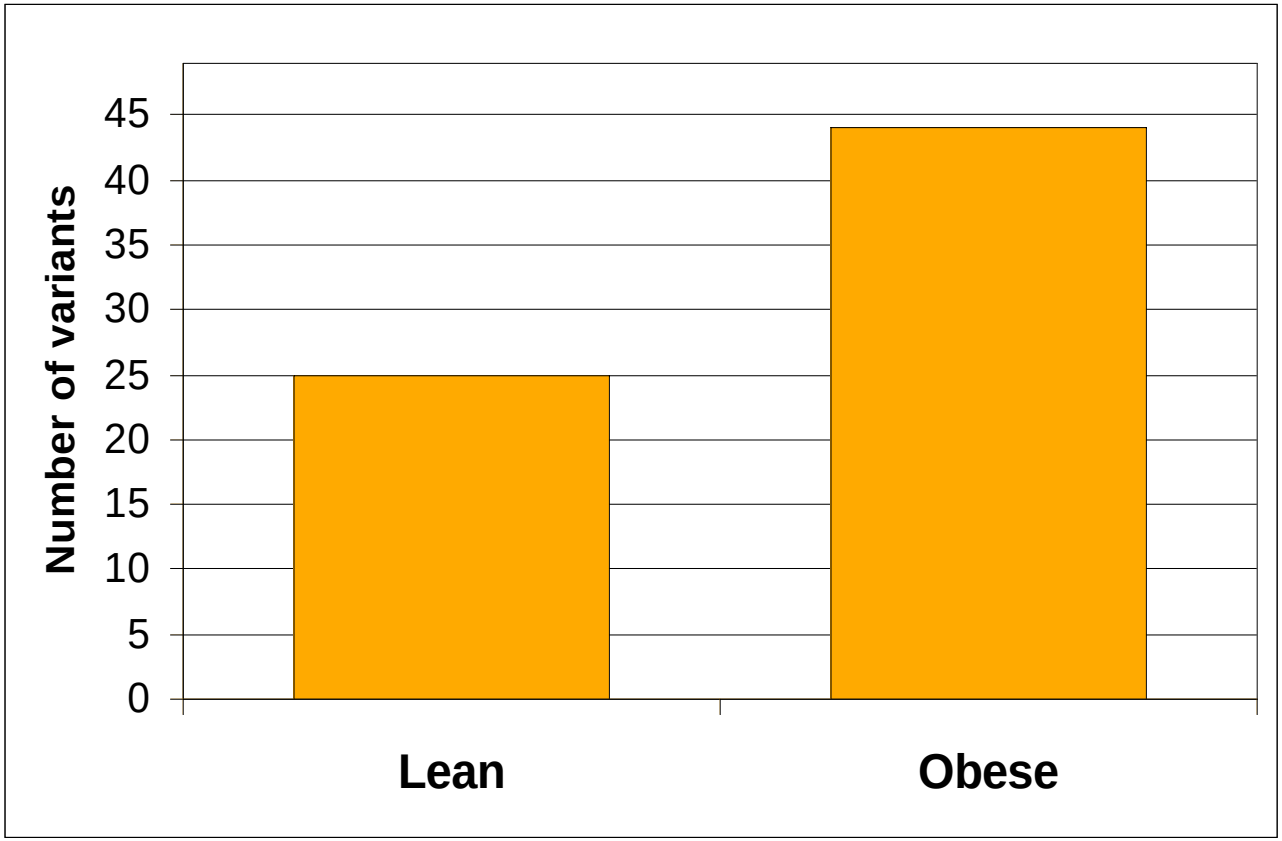


21 genes
379 obese individuals
378 lean individuals

Ahituv et al., *Am. J. Hum. Genet.* 2007

Obesity

Nonsynonymous substitutions

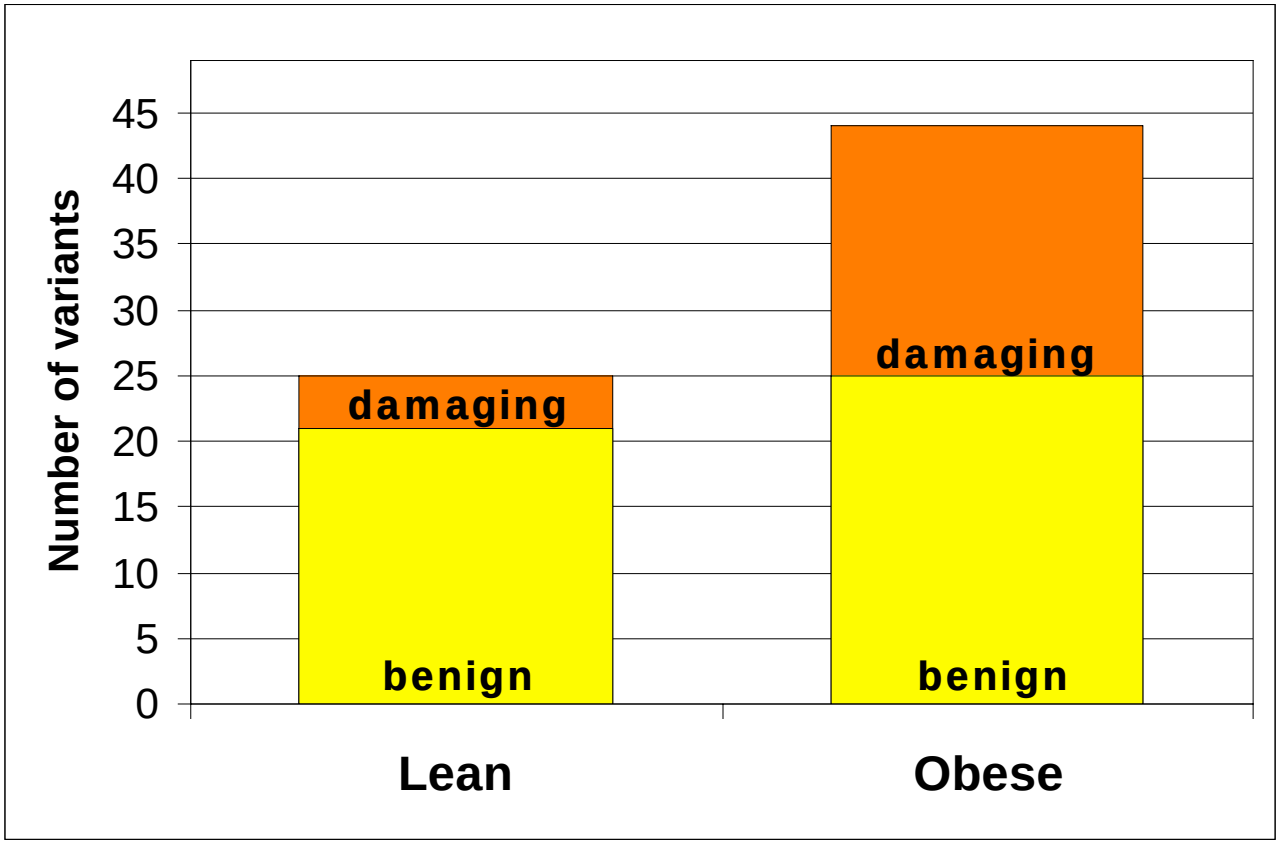


21 genes
379 obese individuals
378 lean individuals

Ahituv et al., *Am. J. Hum. Genet.* 2007

Obesity

Nonsynonymous substitutions



21 genes
379 obese individuals
378 lean individuals

Ahituv et al., *Am. J. Hum. Genet.* 2007

Is it feasible to scale up this approach to the unbiased whole genome gene discovery?

Sequencing will be very cheap very soon…

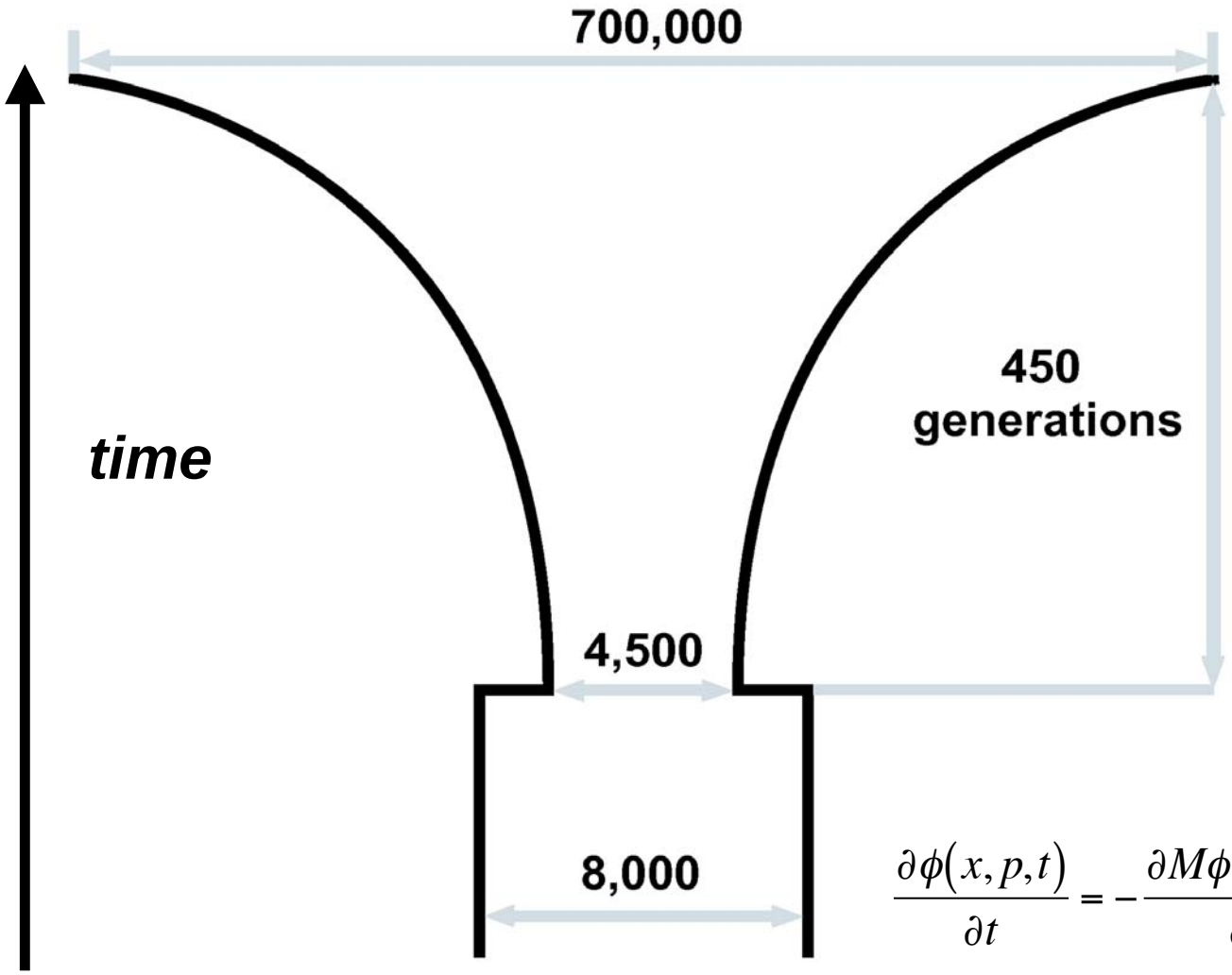
We will soon get large phenotyped populations…

>20,000 genes: with Bonferroni correction we need p-value < $2x10^{-6}$

There are only 6,000,000,000 people on Earth

Is there enough variation in a single gene to guarantee sufficient signal?

Demographic model with four parameters



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

Neutral Wright-Fisher model for variable population size

Diffusion approximation

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N_t} \cdot \frac{\partial^2}{\partial q^2} \{q(1-q)\phi\}.$$

Kimura provided solution for constant population size

$$\phi(q,t|p,N_0) = \sum_{i=1}^{\infty} \frac{(2i+1)(1-(1-2p)^2)}{i(i+1)} \cdot C_{i-1}^{3/2}(1-2p) \cdot C_{i-1}^{3/2}(1-2q) \cdot e^{-\frac{i(i+1)}{4N_0}t},$$

Effective time

$$dt' = (N_0/N_t)dt.$$

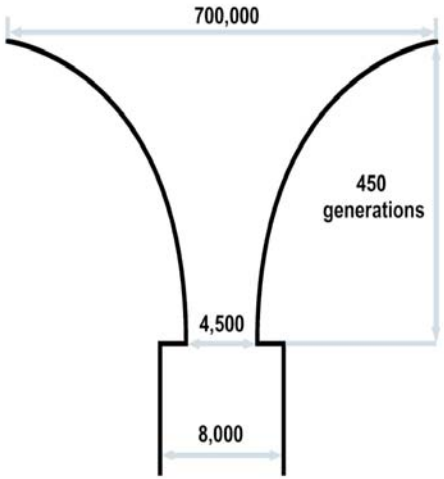
$$\frac{\partial \phi}{\partial t'} = \frac{1}{4N_0} \cdot \frac{\partial^2}{\partial q^2} \{q(1-q)\phi\}.$$

Neutral Wright-Fisher model for variable population size

$$\phi(q, \tau' | p, N_0) = \phi\left(q, \left[\int_0^\tau \frac{N_0}{N_t} dt\right] | p, N_0\right)$$

$$\text{For} \quad N_t = N_0 \cdot e^{\gamma t}, \; \tau' = \int_0^\tau e^{-\gamma t} dt = \frac{1 - e^{-\gamma \tau}}{\gamma}.$$

Summing over epochs



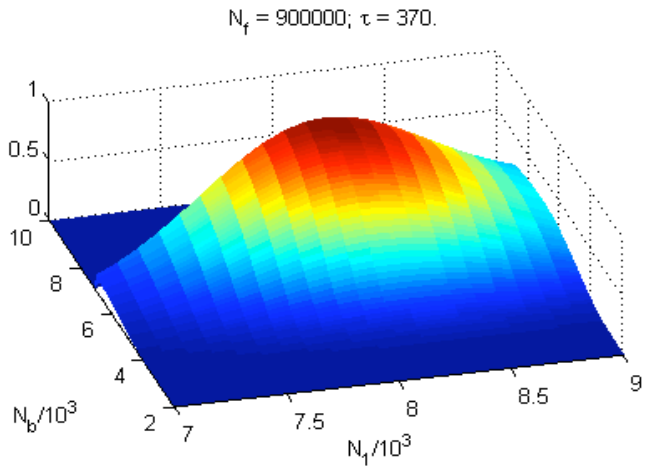
$$f(q) = 4N_1\mu \sum_{i=1}^{2N_b-1} \frac{1}{i} \cdot \phi\left(q, \frac{1-e^{-\gamma\tau}}{\gamma} \middle| \frac{i}{2N_b}, N_b\right) +$$

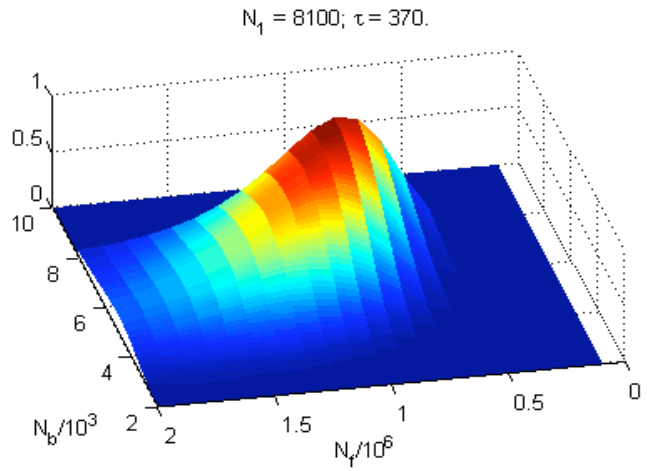
$$2N_b\mu \cdot \sum_{t=1}^{\tau} e^{\gamma t} \phi\left(q, \frac{1-e^{-(\tau-t)\gamma}}{\gamma} \middle| \frac{1}{2N_b e^{\gamma t}}, N_b e^{\gamma t}\right).$$

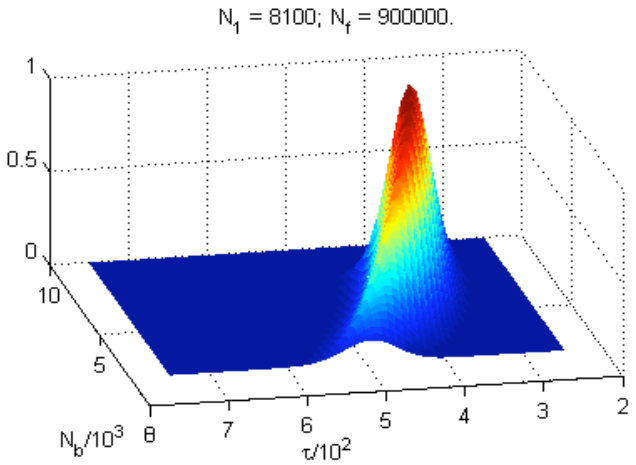
Site frequency spectrum in our sample

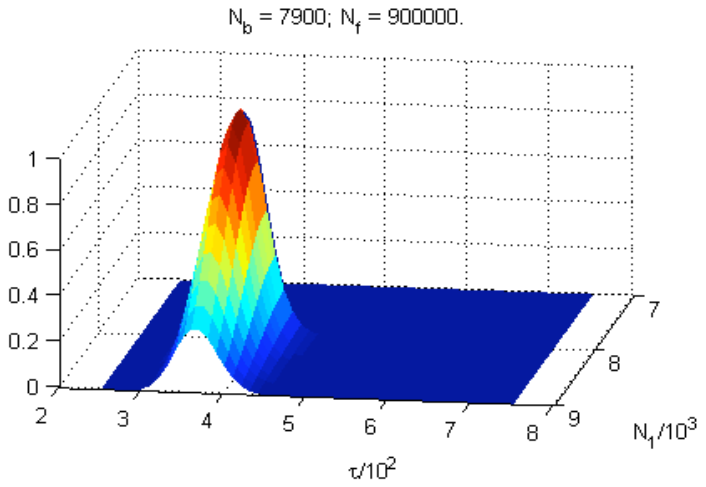
$$F_i = \int_0^1 \binom{N_s}{i} q^i (1-q)^{n-i} f(q) dq.$$

Likelihood surface

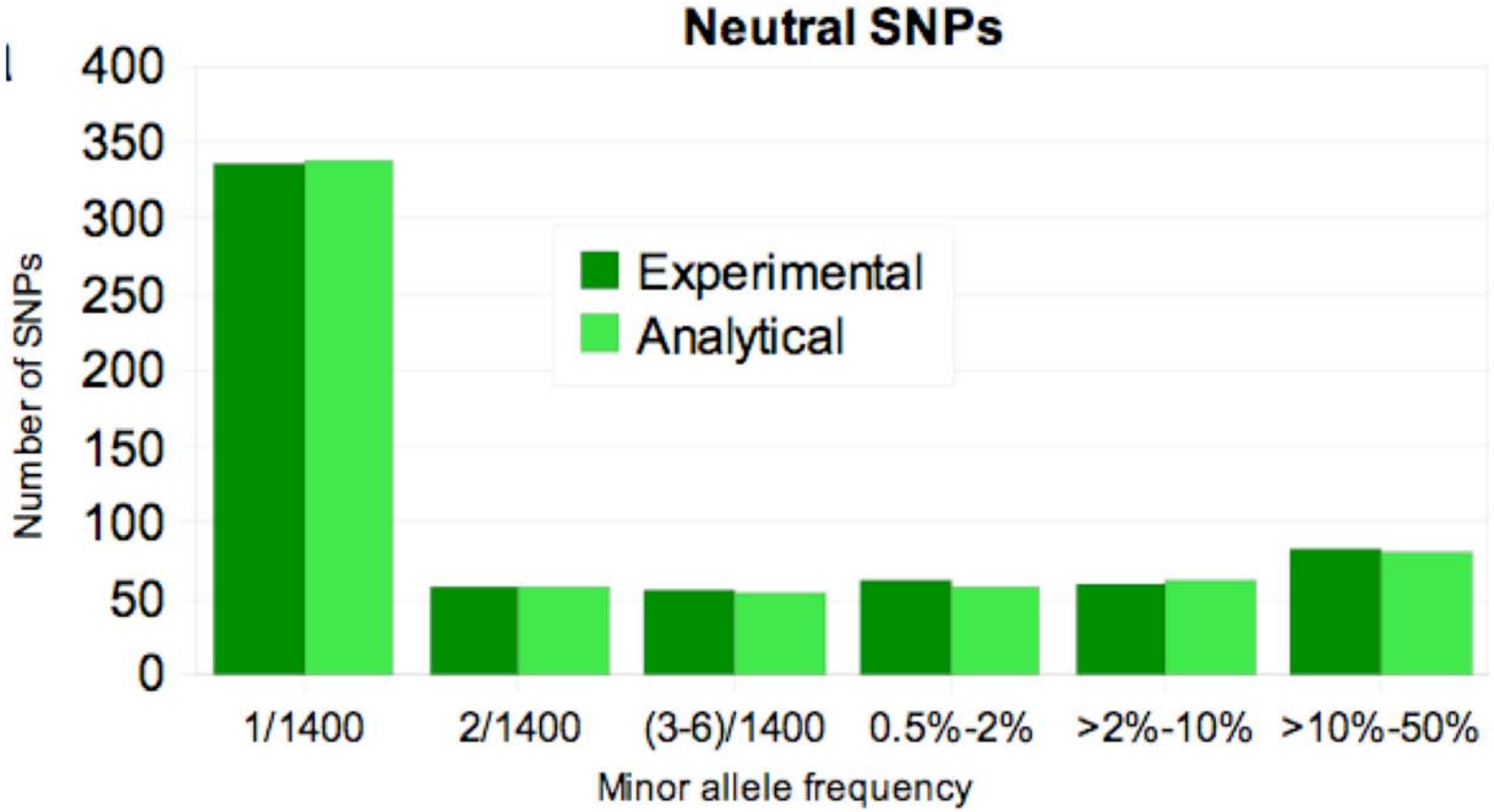




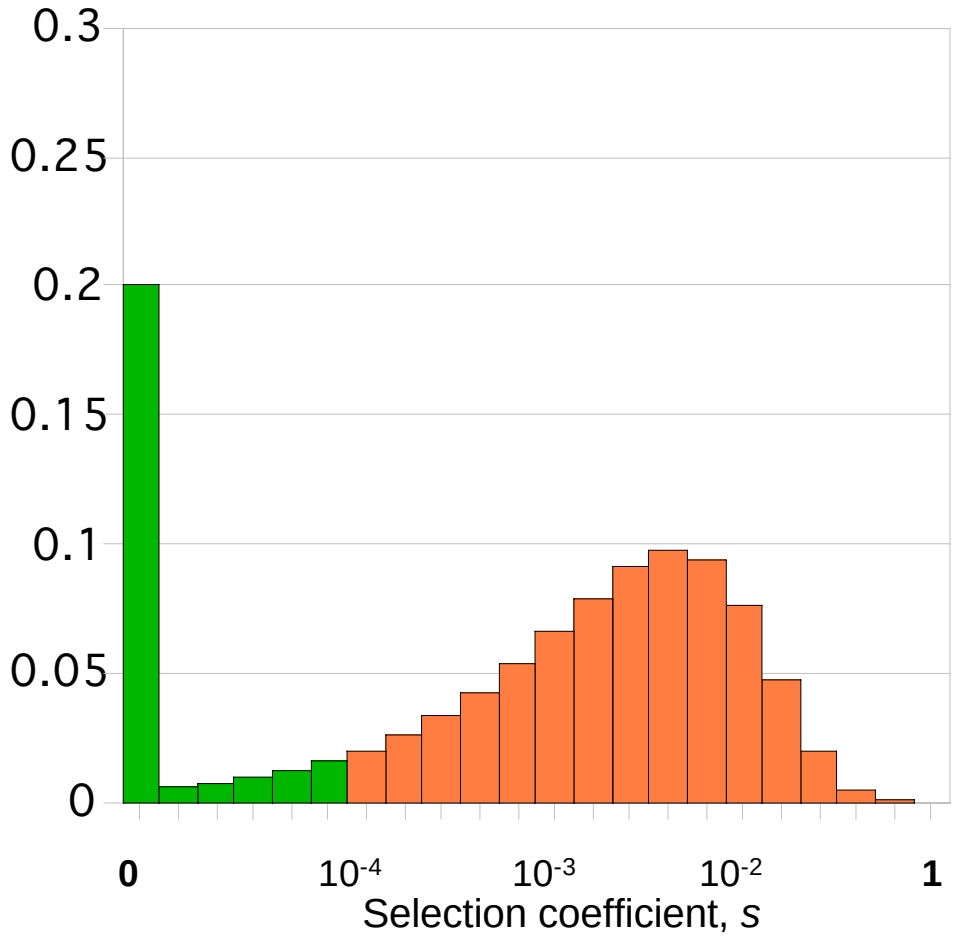




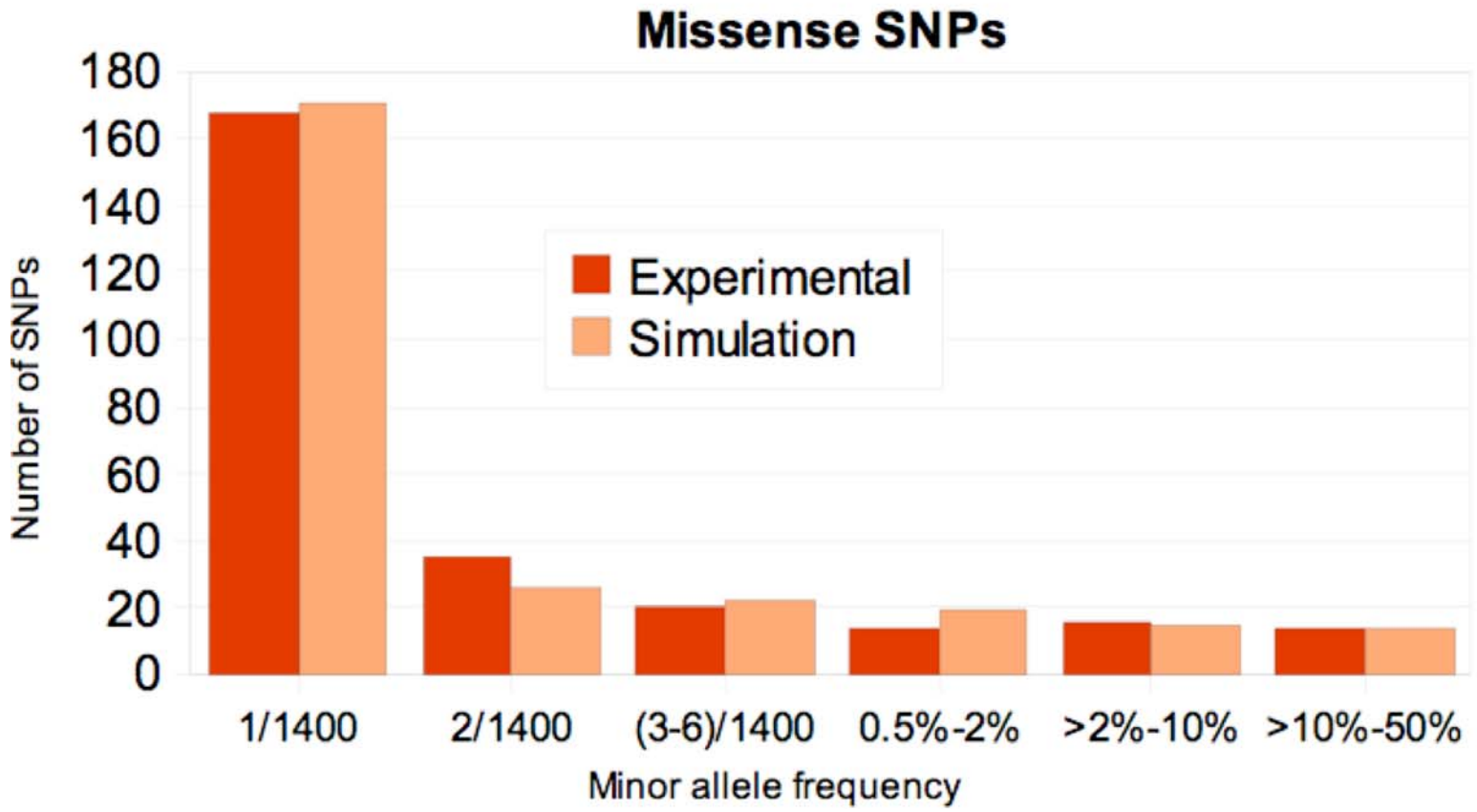
Agreement with the data



Distribution of selection coefficients

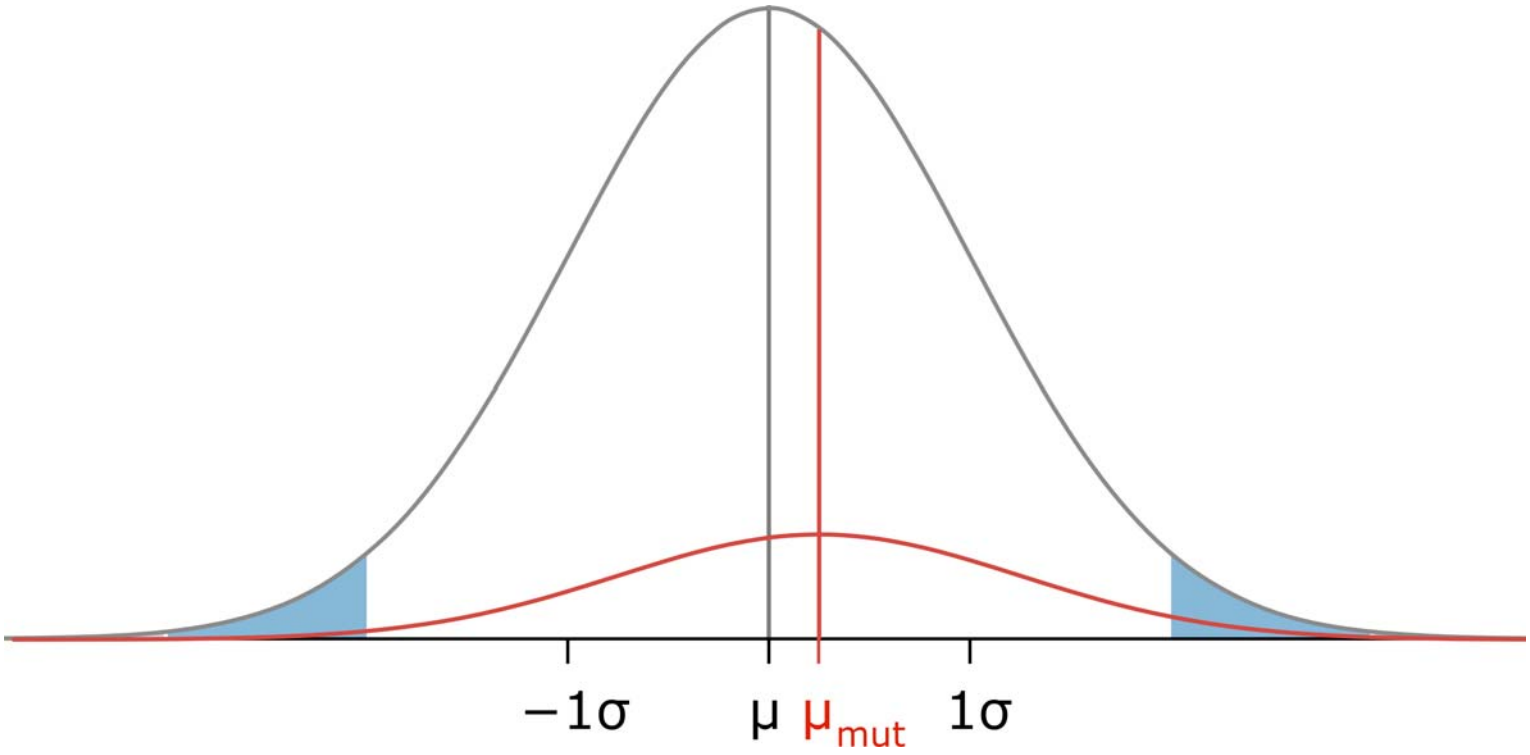


Missense mutations - adding natural selection

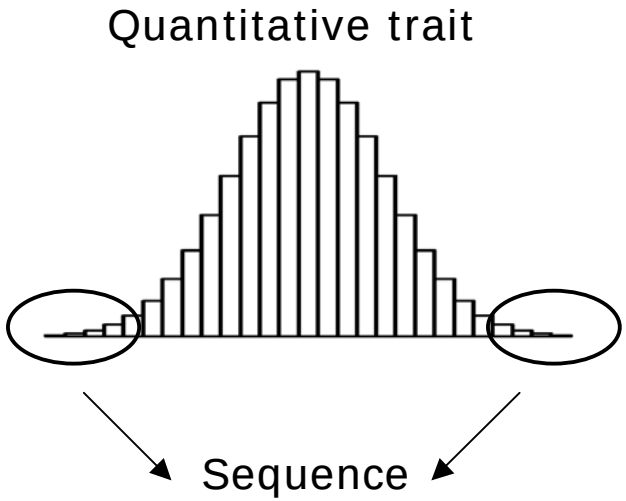


Modeling the effect of mutations on phenotype

We do not assume pre-existing variation with phenotypic effect, we simply rely on mutation rate!



Are whole genome “mutation excess” association studies feasible?



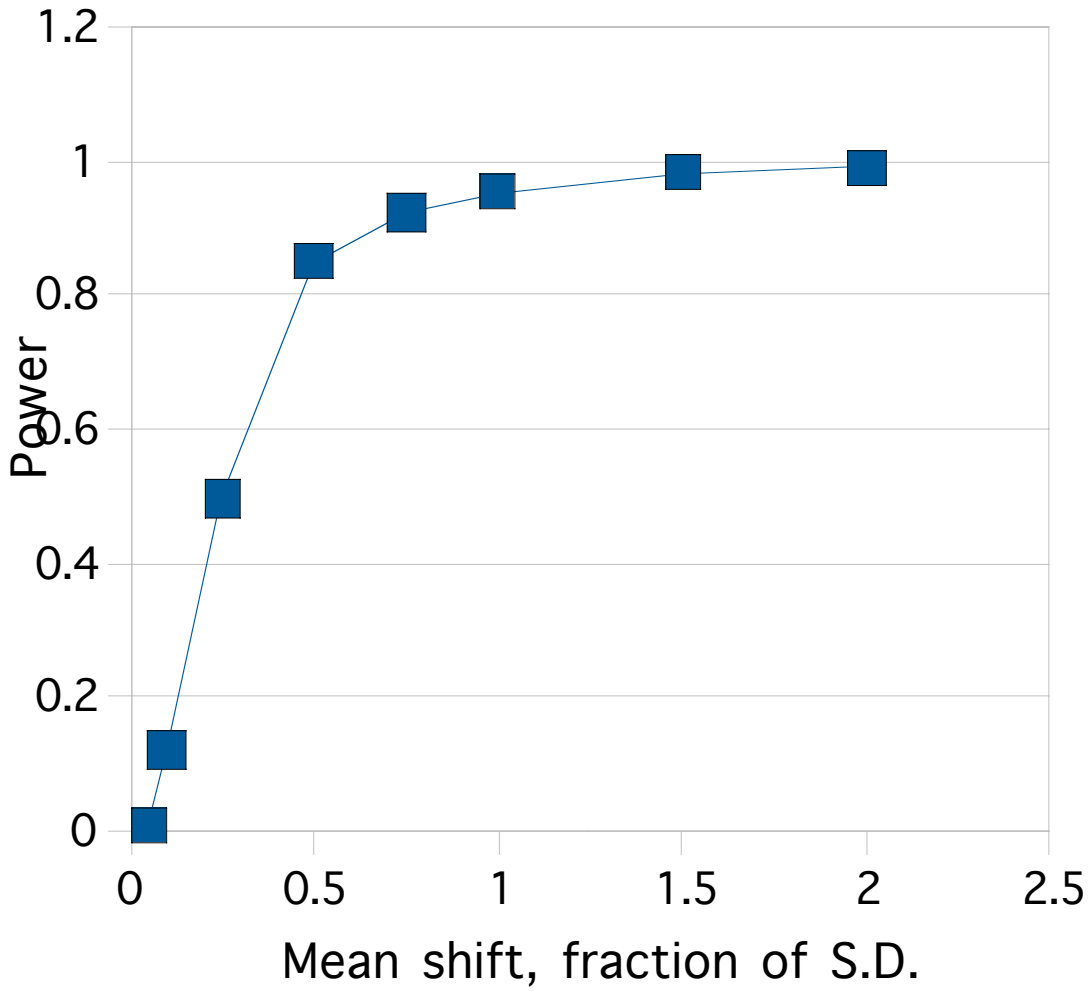
	Gene A		Gene B		Gene C	
	<5% percentile	>95% percentile	<5% percentile	>95% percentile	<5% percentile	>95% percentile
Missense substituions	1	0	0	4	2	0
	7	6	11	19	18	21
	0	3	1	0	0	3
	9	5	0	1	1	5
	1	0	0	3	1	0

>20,000 genes: with Bonferroni correction we need p-value < 2×10^{-6}

Power Table

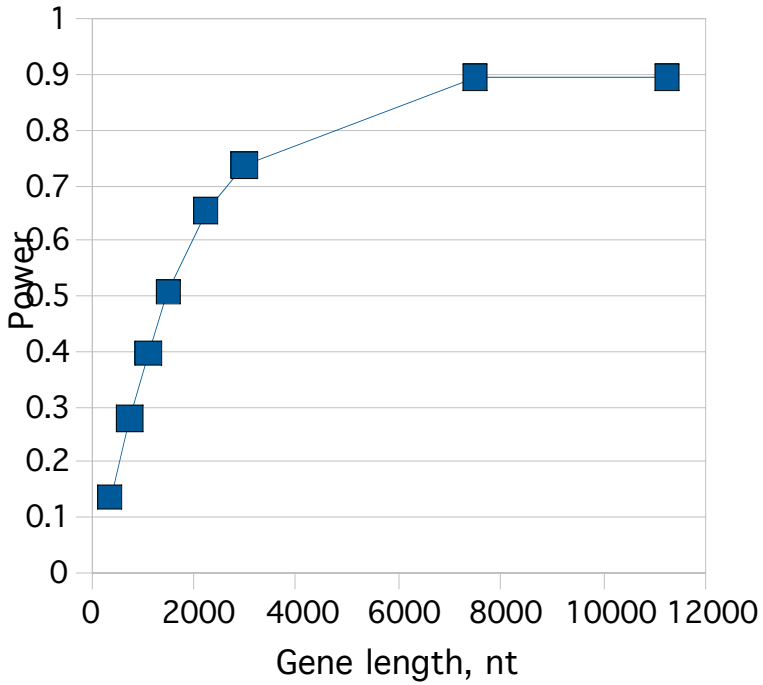
Effect of functional mutations (in fractions of standard deviation)	Number of sequenced individuals	Number of phenotyped individuals				
		12,500	25,000	50,000	100,000	200,000
0.25σ	5,000	0.11	0.18	0.24		
	10,000		0.24	0.31	0.40	
	20,000			0.38	0.51	0.59
0.5σ	5,000	0.36	0.47	0.57		
	10,000		0.56	0.69	0.77	
	20,000			0.76	0.84	0.88

What can we do with smaller sample sizes?



Find genes with larger phenotypic effects

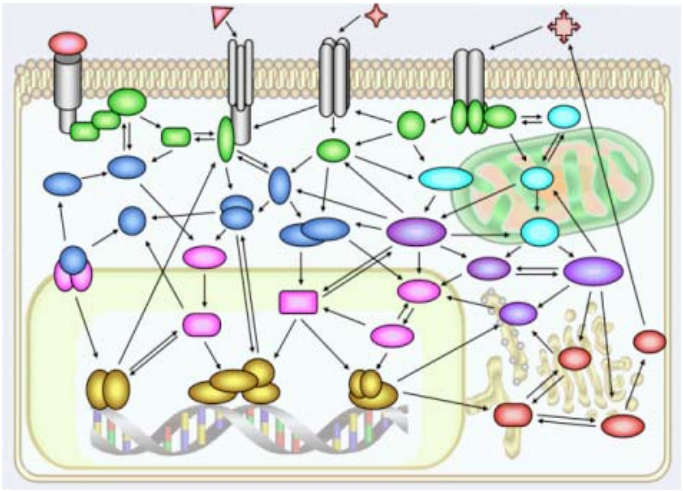
What can we do with smaller sample sizes?



Find longer genes or genes according to pathways:

increase amount of variation;

reduce number of tests



Is this technologically feasible?

- **Sequencing**
  - New sequencing technologies
  - Exon capture is on the way
  - We are approaching to $1,000 per exonome

- **Phenotyping**
  - Current size of clinical cohorts: 10,000-30,000 individuals
  - Well-phenotyped cohorts total 216,000
  - Prospective collection of samples conditional on phenotype

What do we want?

- **Understanding allelic architecture**

- **Search for all variants, coding and non-coding, rare and frequent to explain phenotypic variation in the population**

- **Finding genes**

- **Very deep exon resequencing has a potential of finding relevant genes even if their contribution into population variation is very limited**

- **This approach is analogous to a genetic screen but relies on natural mutations**

Most of the Genome is Non-coding

… and probably is an evolutionary junkyard

However, many genomic regions are highly conserved!

acgtcttcccttaggatc

gcatcttcccttaggcgc

Definition:

Conservation \Con`ser*va"tion\, n. [L. conservatio: cf. F. conservation.] The preservation of a genetic sequence over time due to natural selection.

Population genetics evidence

- Conserved regions are maintained by selection rather than by reduced mutation rate or simply by chance.
- Selective pressure maintaining conserved regions is weak.

Kryukov et al., *Hum. Mol. Genet.* 2005; Drake et al., *Nature Genetics* 2006; Chen et al. *AJHG* 2007

Other reasons to think that some non-coding regions are important:

Medical genetics

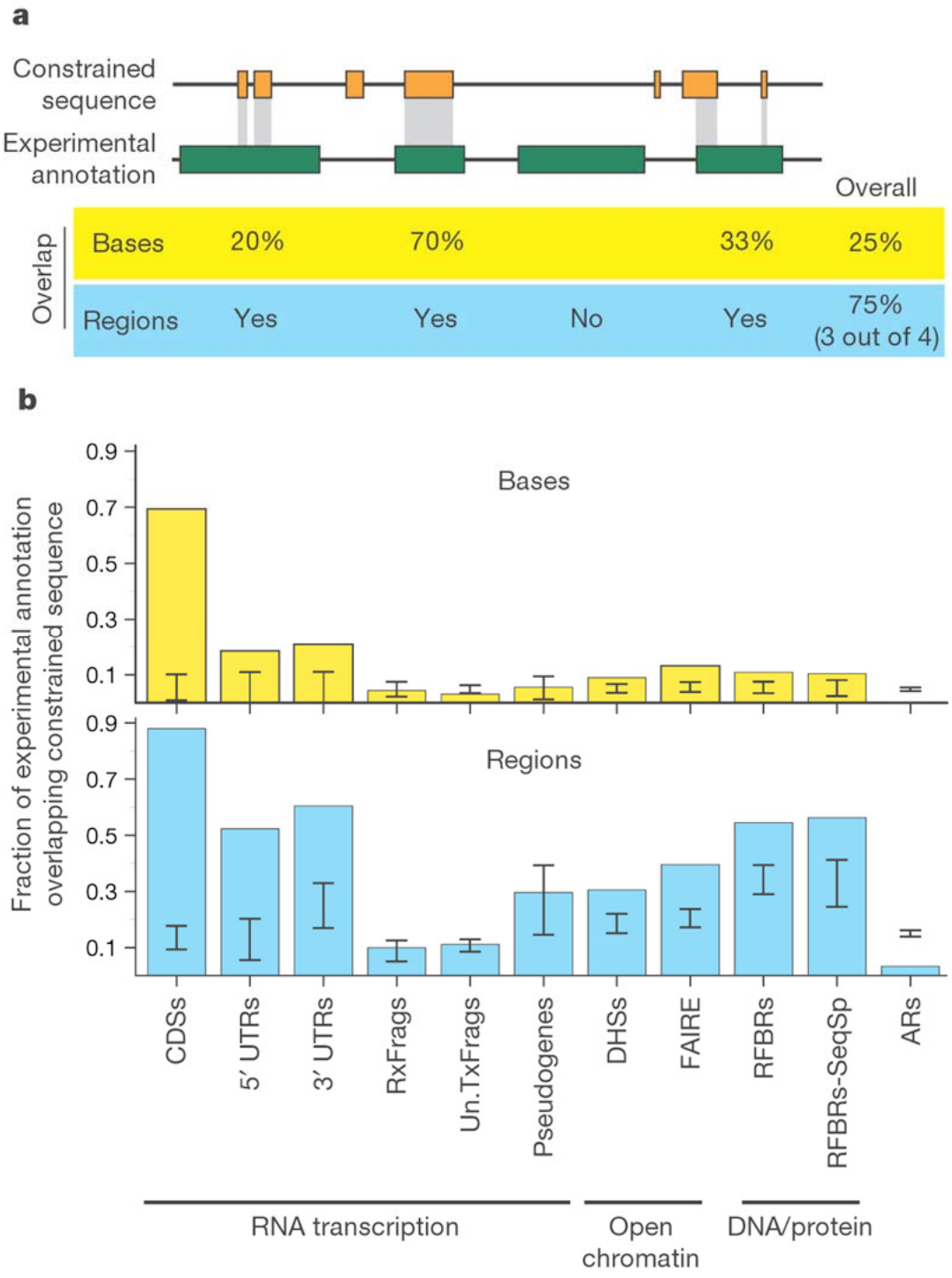
Functional genomics

Medical Genetics

A Common Allele on Chromosome 9 Associated with Coronary Heart Disease

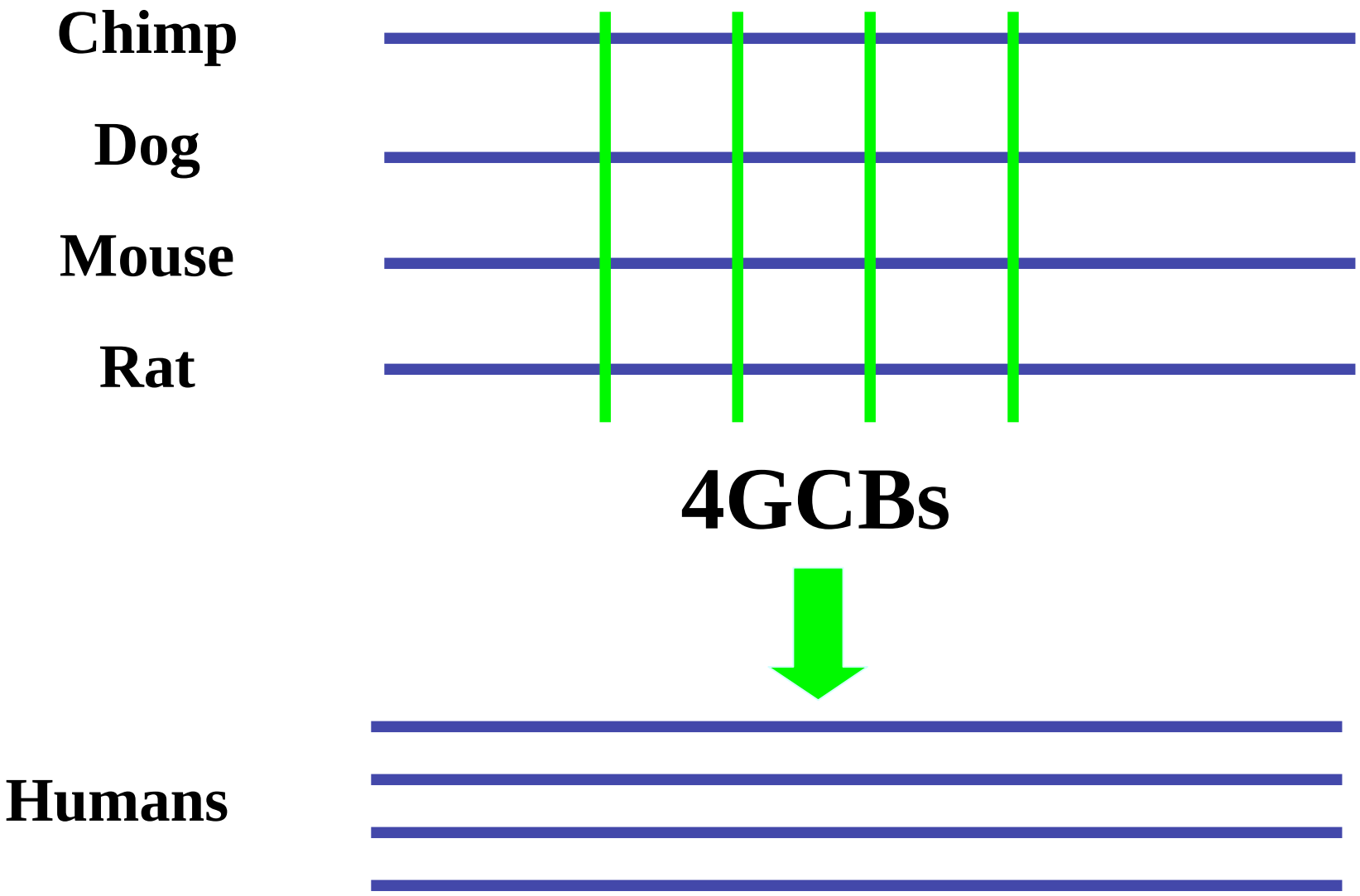
Ruth McPherson,[1*†] Alexander Pertsemlidis,[2*] Nihan Kavaslar,[2] Alexandre Stewart,[2] Robert Roberts,[1] David R. Cox,[3] David A. Hinds,[3] Len A. Pennacchio,[4,5] Anne Tybjaerg-Hansen,[6] Aaron R. Folsom,[7] Eric Boerwinkle,[8] Helen H. Hobbs,[2,9] Jonathan C. Cohen[2,10†]

Coronary heart disease (CHD) is a major cause of death in Western countries. We used genome-wide association scanning to identify a 58-kilobase interval on chromosome 9p21 that was consistently associated with CHD in six independent samples (more than 23,000 participants) from four Caucasian populations. This interval, which is located near the *CDKN2A* and *CDKN2B* genes, contains no annotated genes and is not associated with established CHD risk factors such as plasma lipoproteins, hypertension, or diabetes. Homozygotes for the risk allele make up 20 to 25% of Caucasians and have a ~30 to 40% increased risk of CHD.


a
Constrained sequence
Experimental annotation
Overall
Overlap
Bases
20%
70%
33%
25%
Regions
Yes
Yes
No
Yes
75% (3 out of 4)
b
Fraction of experimental annotation overlapping constrained sequence
Bases
Regions
0.9
0.7
0.5
0.3
0.1
CDSs
5′ UTRs
3′ UTRs
RxFrags
Un.TxFrags
Pseudogenes
DHSs
FAIRE
RFBRs
RFBRs-SeqSp
ARs
RNA transcription
Open chromatin
DNA/protein

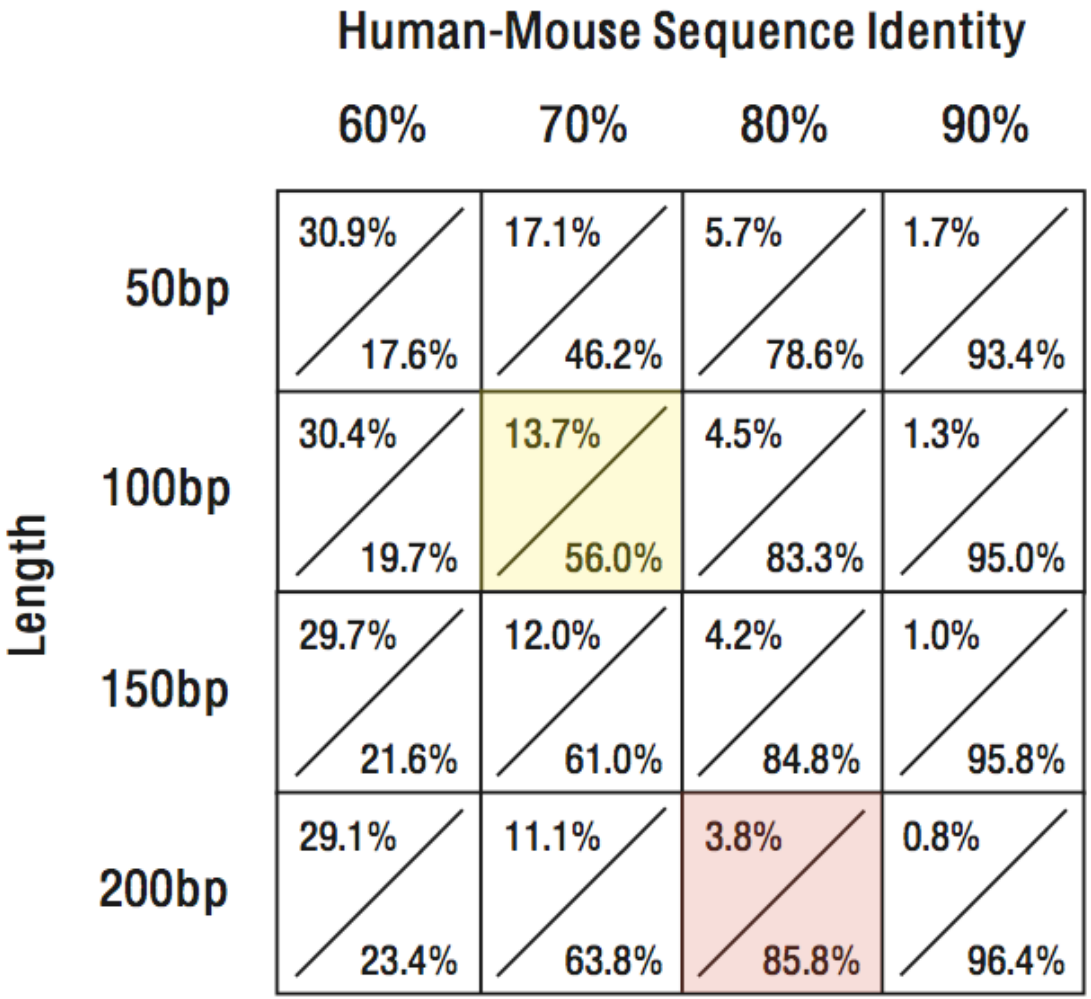
What is in the genome?

- Does the genome consist of protein coding genes, conserved regions and junk?
- Medical genetics and functional genomic data cannot be fully explained by regional conservation.
- Is there anything else?

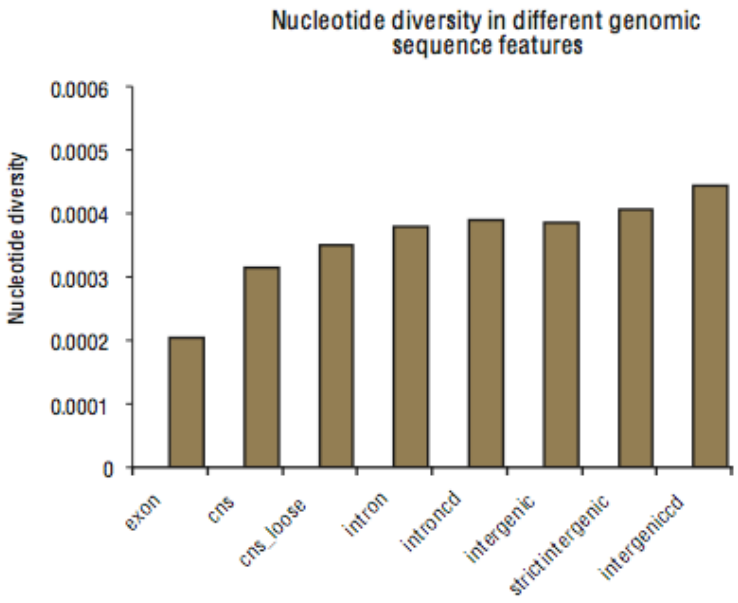


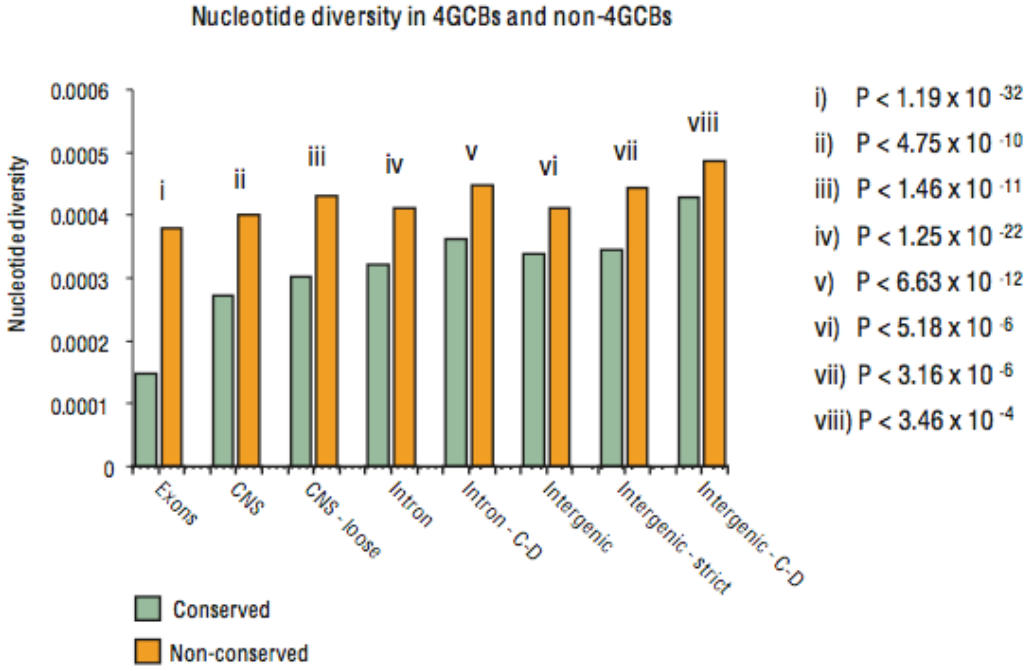
Asthana et al., *PNAS* 2007

4GCBs mostly reside outside of CNSs



Nucleotide Diversity in 4GCBs and non-4GCBs



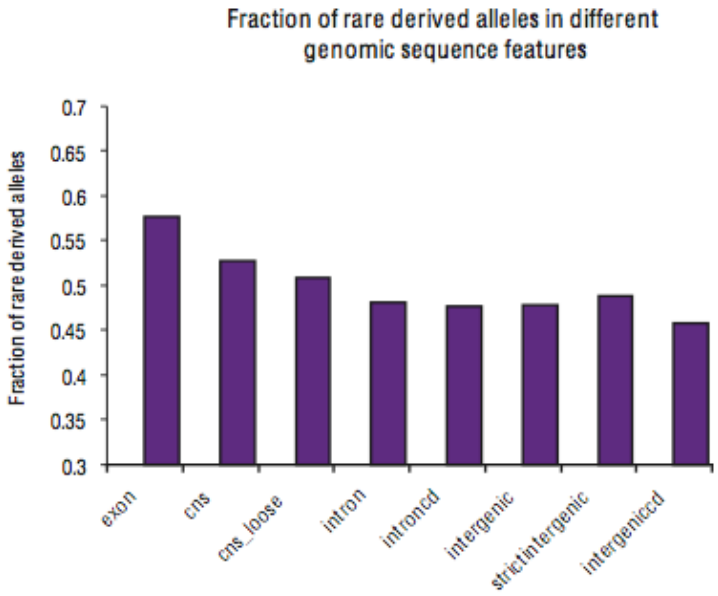


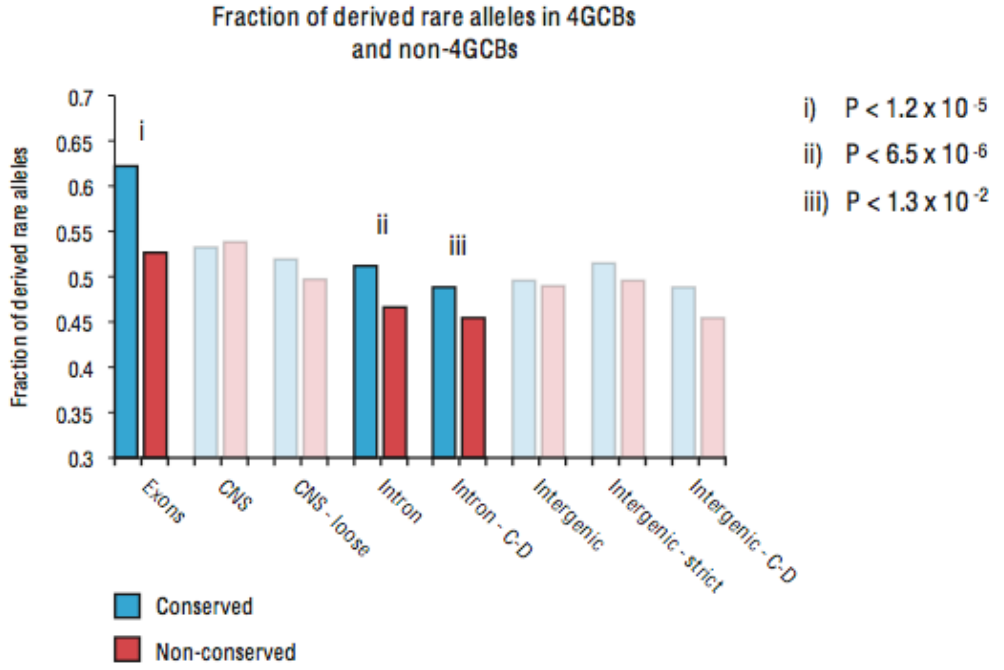
EGP re-sequencing dataset - 567 loci; 90-95 individuals

Is this due to mutation rate heterogeneity?

- Allele frequency distribution
- Polymorphism to divergence ratio

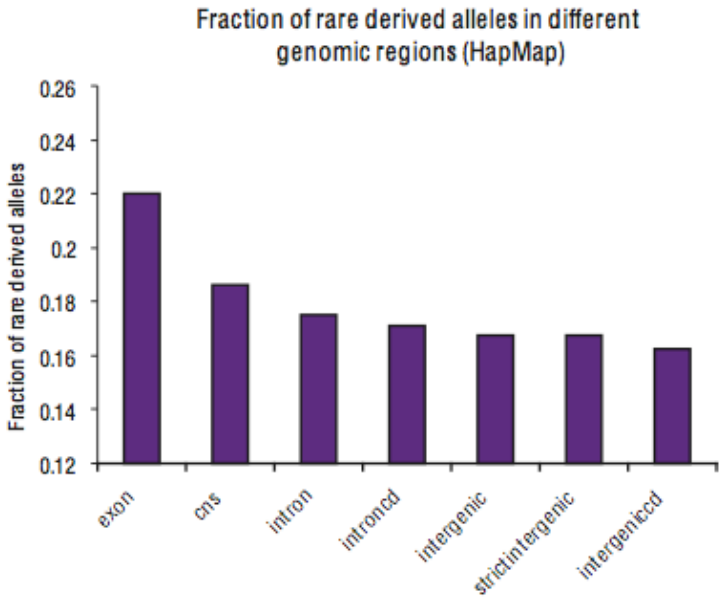
Fraction of rare alleles in 4GCBs and non-4GCBs

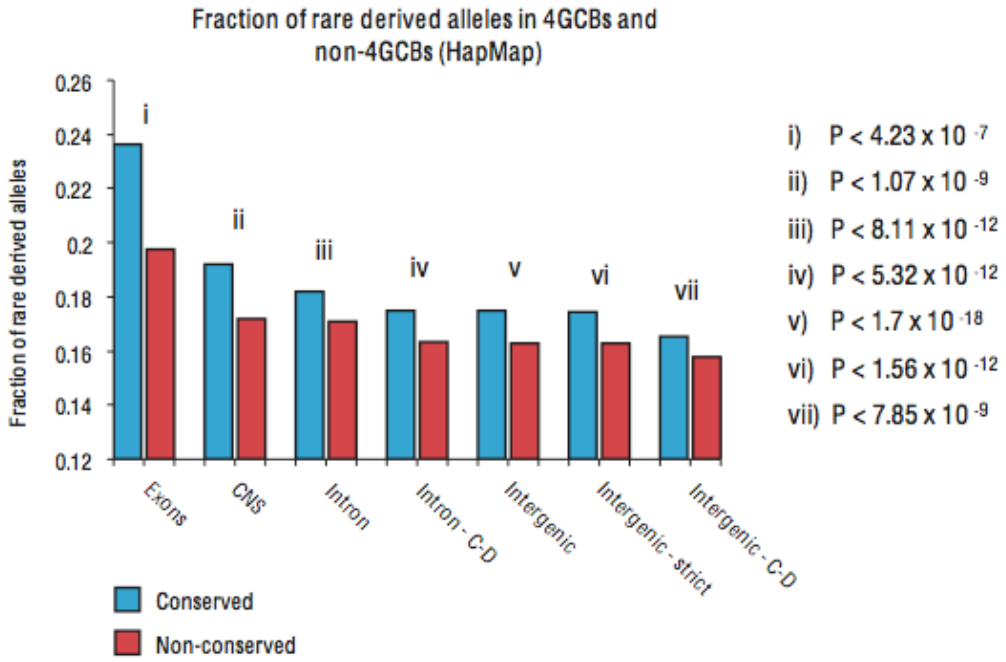




EGP dataset

Fraction of rare alleles in 4GCBs and non-4GCBs





Phase II HapMap dataset

How many functional positions are needed to explain the effect?

Model

- All non-4GCBs are neutral (this is the most conservative assumption)
- 4GCBs are a mixture of neutral and functional sites
- All functional 4GCBs are associated with the same selection coefficient (this is the most conservative assumption)

Fraction of rare neutral alleles

$$F_{neutral}(1\%) = \frac{\int_0^1 \frac{\theta}{x} \cdot \left[mx(1-x)^{m-1} + \frac{m(m-1)}{2} x^2 (1-x)^{m-2} \right] \cdot dx}{\int_0^1 \frac{\theta}{x} \cdot \left(1 - x^m - (1-x)^m\right) dx}$$

$$F_{neutral}(1\%) = \frac{3}{2 \cdot \sum_{i=1}^{m-1} \frac{1}{i}}$$

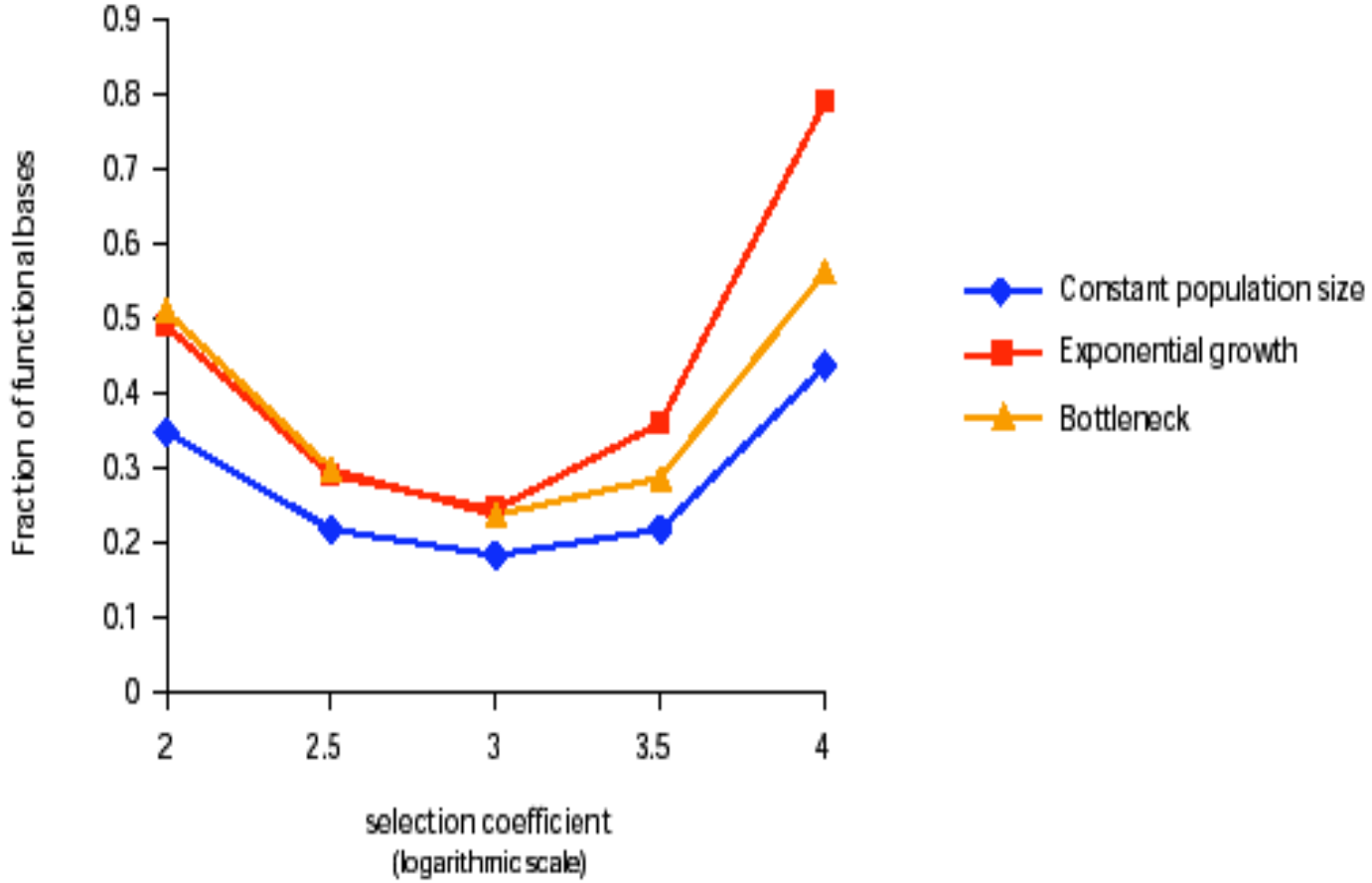
Mixture of neutral and functional sites

$$F_{mixture}(1\%) = \frac{\alpha \cdot n_{functional}(1\%) + \beta \cdot n_{neutral}(1\%)}{\alpha \cdot n_{functional} + \beta \cdot n_{neutral}}$$

$$n_{functional}(1\%) = \int_0^1 \frac{\theta\left(e^{-2N_e s(1-x)} - 1\right)}{x(1-x)\left(e^{-2N_e s} - 1\right)} \cdot \left[mx(1-x)^{m-1} + \frac{m(m-1)}{2} x^2 (1-x)^{m-2} \right] \cdot dx$$

$$n_{functional} = \int_0^1 \frac{\theta\left(e^{-2N_e s(1-x)} - 1\right)}{x(1-x)\left(e^{-2N_e s} - 1\right)} \left(1 - x^m - (1-x)^m\right) \cdot dx$$

How many functional sites are needed to produce observed allele frequency shift?



Selective constraints in non-coding regions of the genome

- Selectively constrained bases are diffusely distributed along the genome rather than condensed to highly conserved regions
- At least ~20% of 4GCBs are electively constrained (2% of the genome sequence)
- Probably additional constrained positions in non-alignable regions

Regions selected for the ENCODE project have 22 mammalian species sequenced

… and a lot of functional genomics data

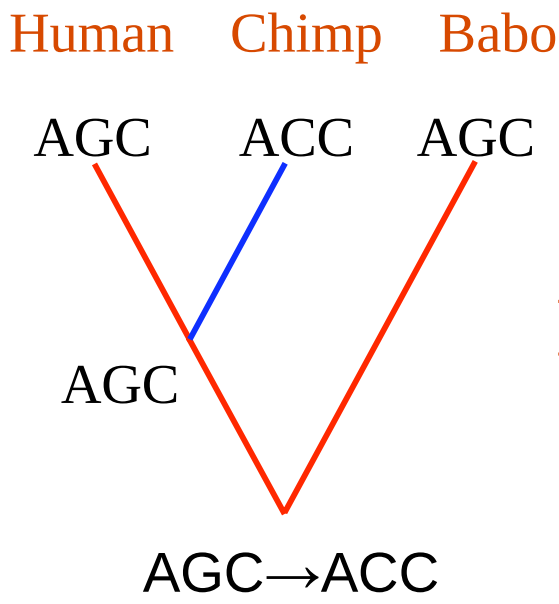
SCONE (Sequence CONservation Evaluation)

Instantaneous rate matrix of transitions Q

$$P(t) = e^{Qt}$$

- Ignores mutation rate heterogeneity along the genome
- Assumes uniformity between species
- Computes Bayesian estimate of evolutionary rate at the site
- Computes *p*-value via simulations

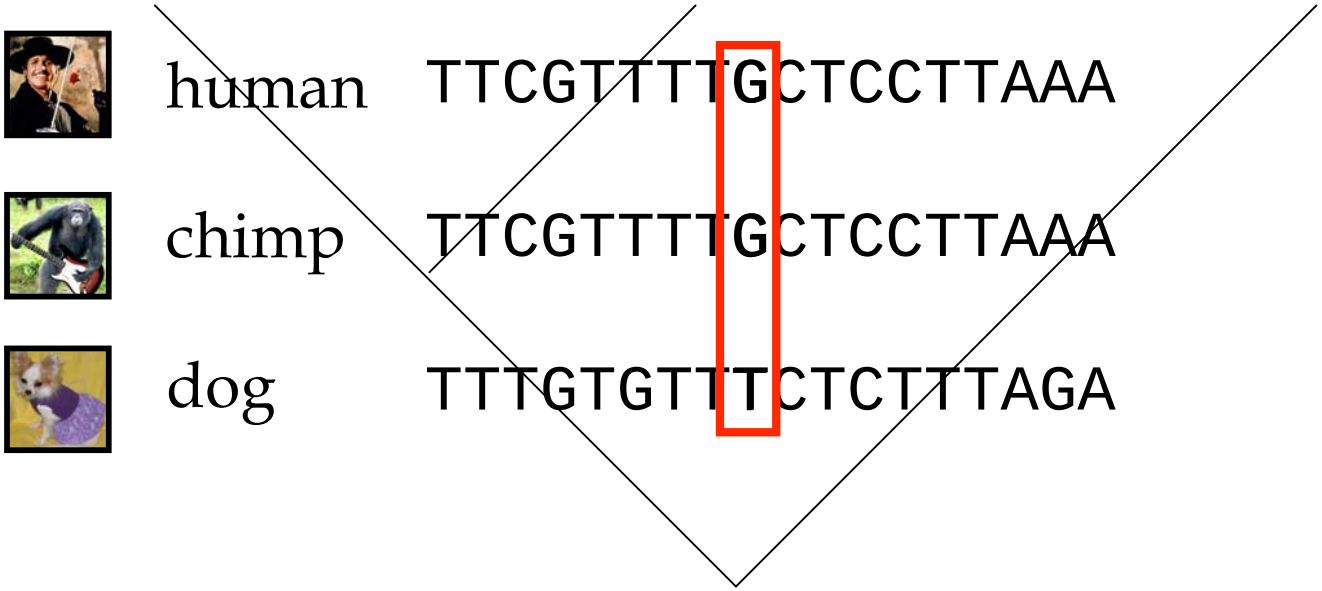
Asthana et al., *PLoS CB* 2007



Mutation rates are modeled as asymmetric and context specific.

The model incorporates insertions and deletions

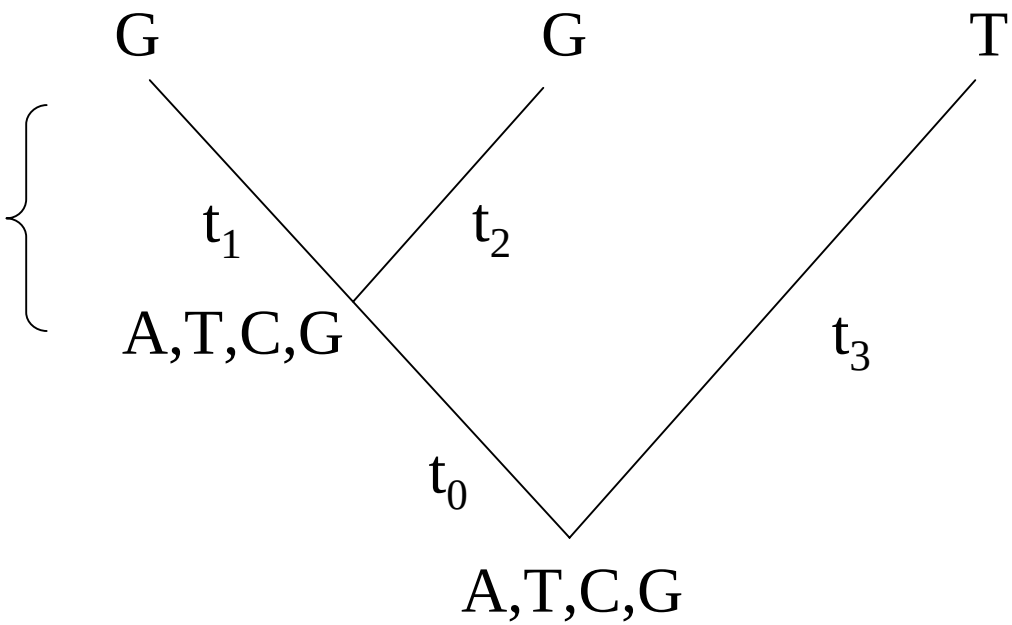
Estimating conservation



Likelihood

$$F(i) = p(i \to G_h, t_1) \cdot F(G_h) \cdot p(i \to G_c, t_2) \cdot F(G_c)$$

$$L(G,G,T) = \sum_{i \in A,T,G,C} \pi_i \cdot F(i)$$



Estimation of substitution rate

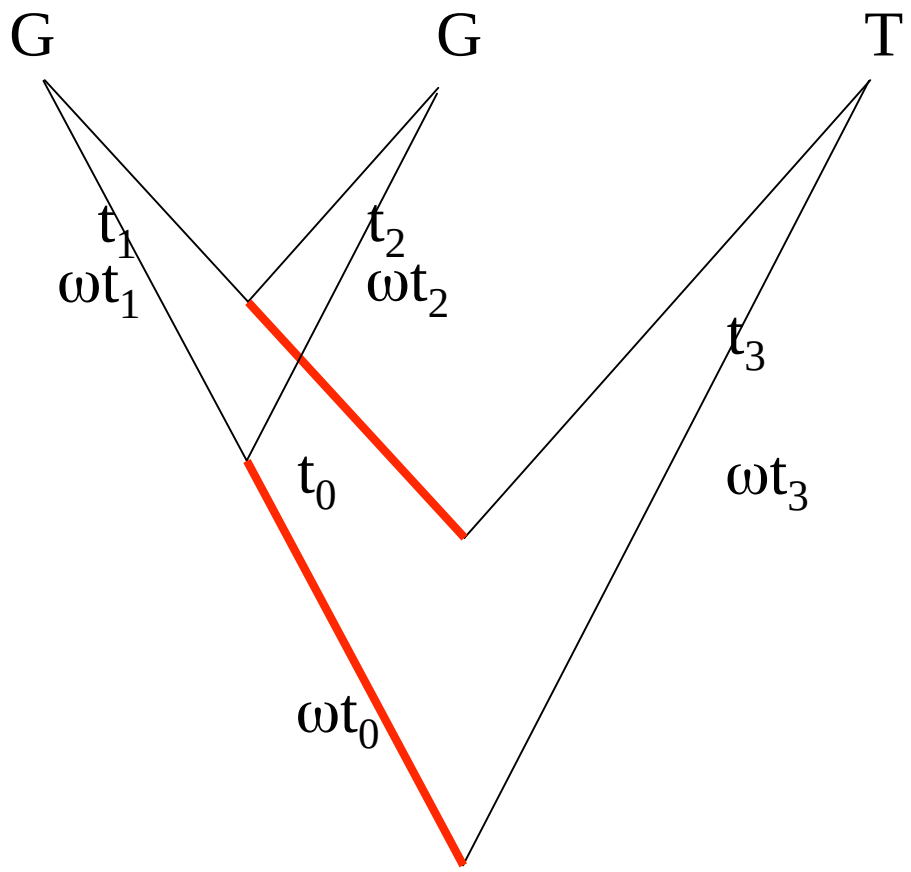
$$F(i,\omega) =$$
$$p(i \to G_h, \omega t_1) \cdot F(G_h) \cdot p(i \to G_c, \omega t_2) \cdot F(G_c)$$

$$L(G,G,T,\omega) = \sum_{i \in A,T,G,C} \pi_i \cdot F(i,\omega)$$

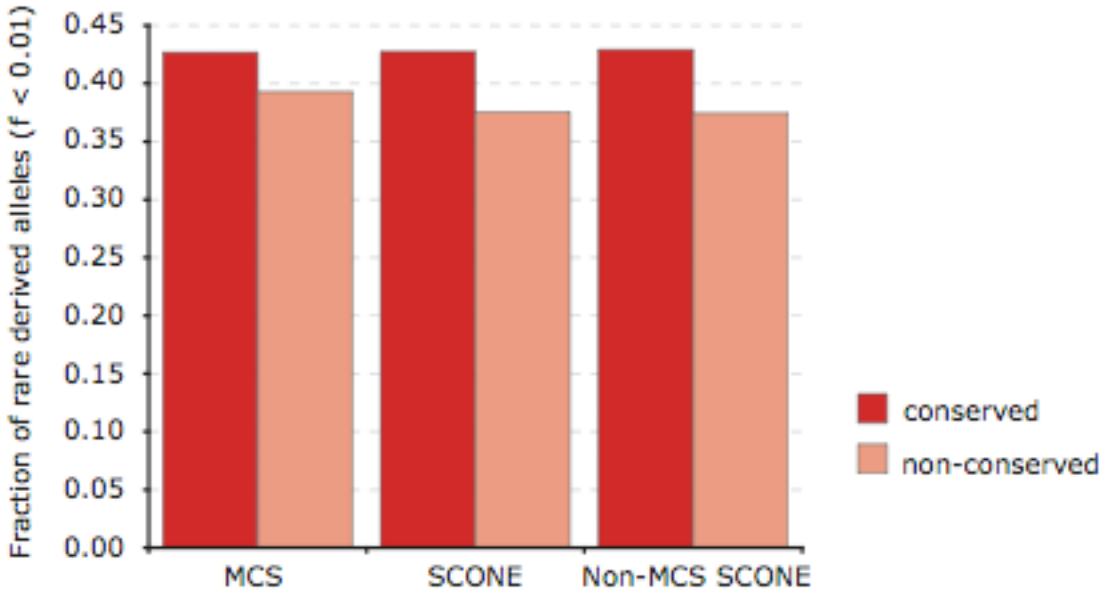
$$\omega_{\max} = \arg\max_{\omega} L(G,G,T,\omega)$$

We also use Bayesian estimate of ω

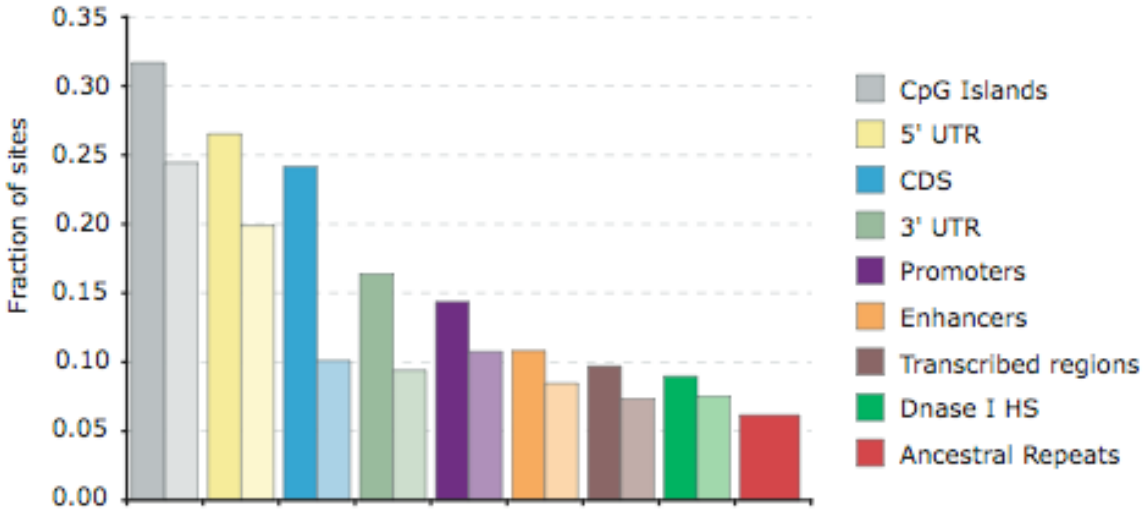
P-value can be computed via simulations



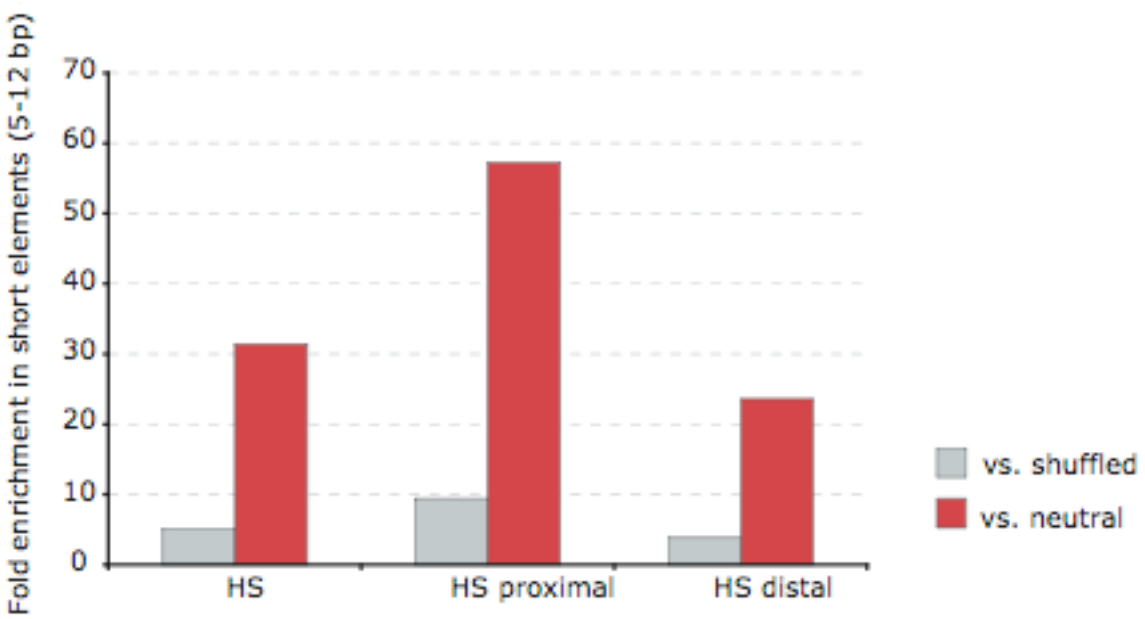
SCONE vs. ENCODE SNPs



Conservation of functional features



Clustering of conserved positions



Non-coding nucleotides

•Analysis of available sequence data suggests that most of selectively constrained nucleotides in the genome are non-coding.

• However, on average, the effect of non-coding mutations is much weaker.

Acknowledgments

The lab: Gregory Kryukov, Alex Shpunt, Ivan Adzhubey, Saurabh Asthana, Victor Spirin, Steffen Schmidt

University of Washington:

John Stamatoyannopoulos, William Noble

Berkeley:

Nadav Ahituv, Len Pennacchio

UT Southwestern:

Jonathan Cohen, Alex Pertsemlidis

NIH, Pfizer