POLYMORPHISM AND VARIANT ANALYSIS

Matt Hudson
Crop Sciences
NCSA
IGB
University of Illinois

Outline

- How do we predict molecular or genetic functions using the variants called from sequencing? Two major approaches:
  - Predicting when a coding SNP or SNV is “damaging”
  - Genome-wide association studies

What is a SNP ? And a SNV?

- Single nucleotide polymorphism
- Single nucleotide variant

```
I1:   AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I2:   AACGAGCTAGCGATCGATCGACAACGACTACGAGGT
I3:   AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I4:   AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I5:   AACGAGCTAGCGATCGATCGACAACGACTACGAGGT
I6:   AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I7:   AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I8:   AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
```

Individuals I2 and I5 have a variation (T -> A). This position is both.

Notes on SNPs and SNVs

- A SNV is any old change (e.g. could be a somatic mutation in an individual, or even an artifact)
- To be called a SNP, has to be polymorphic SNV:
  - "Minor" and "Major" alleles
  - Sometimes minor allele frequency (MAF) threshold – e.g 5% at dbSNP
  - "Segregating" sites – germplasm polymorphism in population
- The 1000 Genomes project recorded ~41 Million SNPs by sequencing ~1000 individuals.

Thus, your fields may differ

- If you are a population geneticist doing GWAS, you are generally only interested in SNPs
- If you are a cancer geneticist looking at sequence data from tumors, you are primarily interested in SNVs
- In non-human biology there can be other complications (e.g. polyploidy, HGT etc.).
- Definitions vary by field

Structural variants

- There is increasing interest in the biological consequences of structural variants (such as insertions, deletions, inversions, translocations, etc.).
- Usually, short read "resequencing" only gives very small structural variants, usually indels.
- Long read data is driving research into human structural variation. Still largely a research field, some clinical interest in cancer.

Predicting functional effects

Geneticists often use SNPs as “markers”
But, SNPs and SNVs can cause disease also
How do we know if they are likely to affect protein function?

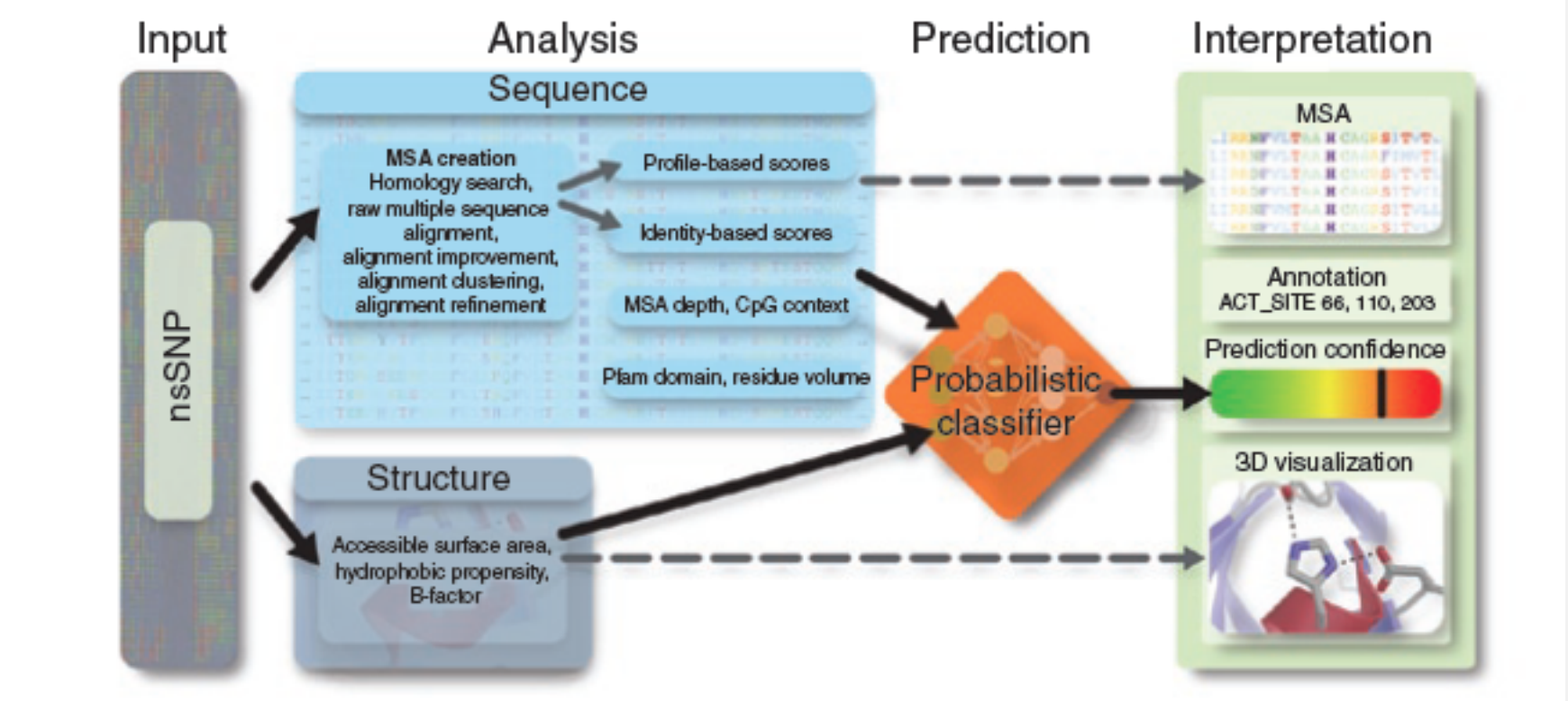
Predicting when coding SNPs are damaging

- Question:
  - I found a SNP inside the coding sequence. Knowing how to translate the gene sequence to a protein sequence, I discovered that this is a non-synonymous change, i.e., the encoded amino acid changes. This is an nsSNP.
  - Will that impact the protein's function?
  - (And I don't quite know how the protein functions in the first place …)

Two popular approaches

- We will discuss one popular software/method for answering the question: PolyPhen 2.0.
  - Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods 7(4):248–249 (2010).
- Another popular alternative: SIFT.
  - Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–81.

PolyPhen 2.0 (current version)



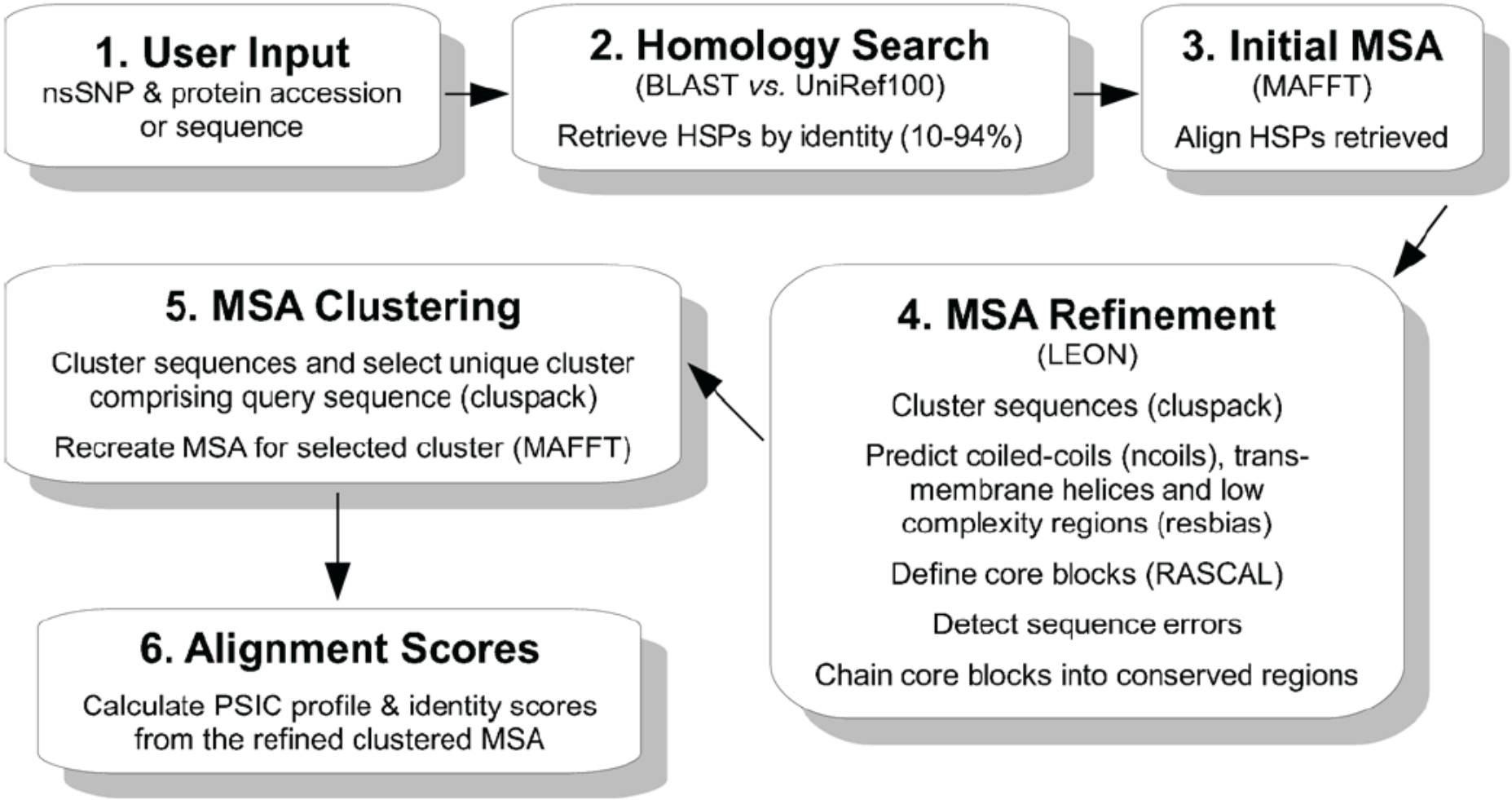
Data for training/evaluation

- HumDiv
  - Damaging mutations from UniProvKB. Look for annotations such as “complete loss of function”, “abolishes”, “no detectable activity”, etc.
  - Non-damaging mutations: differences in homologous proteins in closely related mammalian species

"Features"

Name	Definition	Values with ranges in HumDiv
nt1	wild type allele nucleotide	A,C,G,T
nt2	mutation allele nucleotide	A,C,G,T
site	SITE annotation from UniProt/Swiss-Prot	Yes, No
region	REGION annotation from UniProt/Swiss-Prot	NO, PROPER, SIGNAL, TRANSMEM
phat	PHAT matrix element in the TRANSMEM region	[-8.0, 4.0], mean = -0.04
score1	PSIC score for the wild type allele	[-1.1], mean = 1.07
score2	PSIC score for the mutant allele	[-1.39, 2.64], mean = .166
score_delta	difference of PSIC scores (Score1-Score2)	[-3.23, 4.57], mean = .905
num_observ	number of residues observed at the position of the multiple alignment	[1, 432], mean 69.3
delta_volume	change in residue side chain volume	[-167, 167], mean = -1,93
transv	mutation origin by transversion or transition	Yes, No
CpG	mutation origin in the CpG hypermutable context	Yes, No
pfam_hit	position of the mutation within/outside a protein domain as defined by Pfam	Yes, No
id_p_max	congruency of the mutant allele to the multiple alignment	[0, 95.5], mean = 24
id_q_min	sequence identity with the closest homologue deviating from wild type allele	[1.56, 95.5], mean 68.76
cpgVar1Var2	presence of the CpG context combined with wild type and mutant amino acid types	NO, AA1_AA2
cpg_transition	whether variant happened as transition in CpG context	No, Transition, Transversion
charge_change	change in electrostatic charge	0,1,2
hydroph_change	change in hydrophobicity	[0, 2.85], mean 0.80
ali_ide	sequence identity with the closest homolog with known 3D structure	[0, 1], mean 0.33
ali_len	alignment length with the closest homolog with known 3D structure	[0, 1213], mean 130.0
acc_normed	normalized accessible surface area of amino acid residue	[0, 1.55], mean .35
sec_str	secondary structure	HELIX, SHEET, OTHER
map_region	region of the Ramachandran map	ALPHA, BETA, OTHER
delta_prop	change in accessible surface area propensity	[-2.89, 2.89], mean -0.07
b_fact	crystallographic beta-factor	[-1.85, 5.17], mean 0.0
het_cont_ave_num	average number of contact with heteroatoms	Yes, No
het_cont_min_dist	minimal distance to a heteroatom	Yes, No
inter_cont_ave_num	average number of interchain contacts in a protein complex	Yes, No
inter_cont_min_dist	average minimal interchain distance	Yes, No
delta_volume_new	change in residue volume for buried residues	[-119, 138], mean -0.5
delta_prop_new	change in accessible surface area propensity for buried residues	[-1.83, 2,89], mean 0.0026

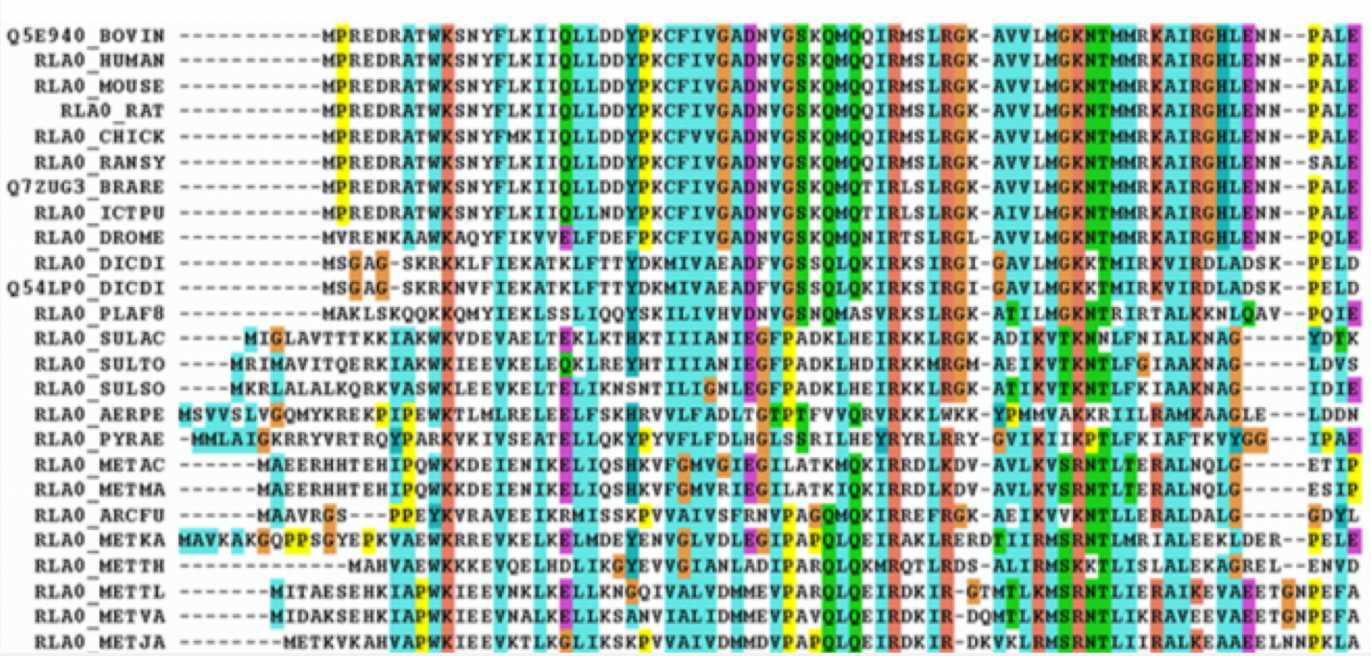
Multiple Sequence Alignment



PMID 10360979

Position Specific Independent Count (PSIC)

- Reflects the amino acid's frequency at the specific position in sequence, given an MSA.

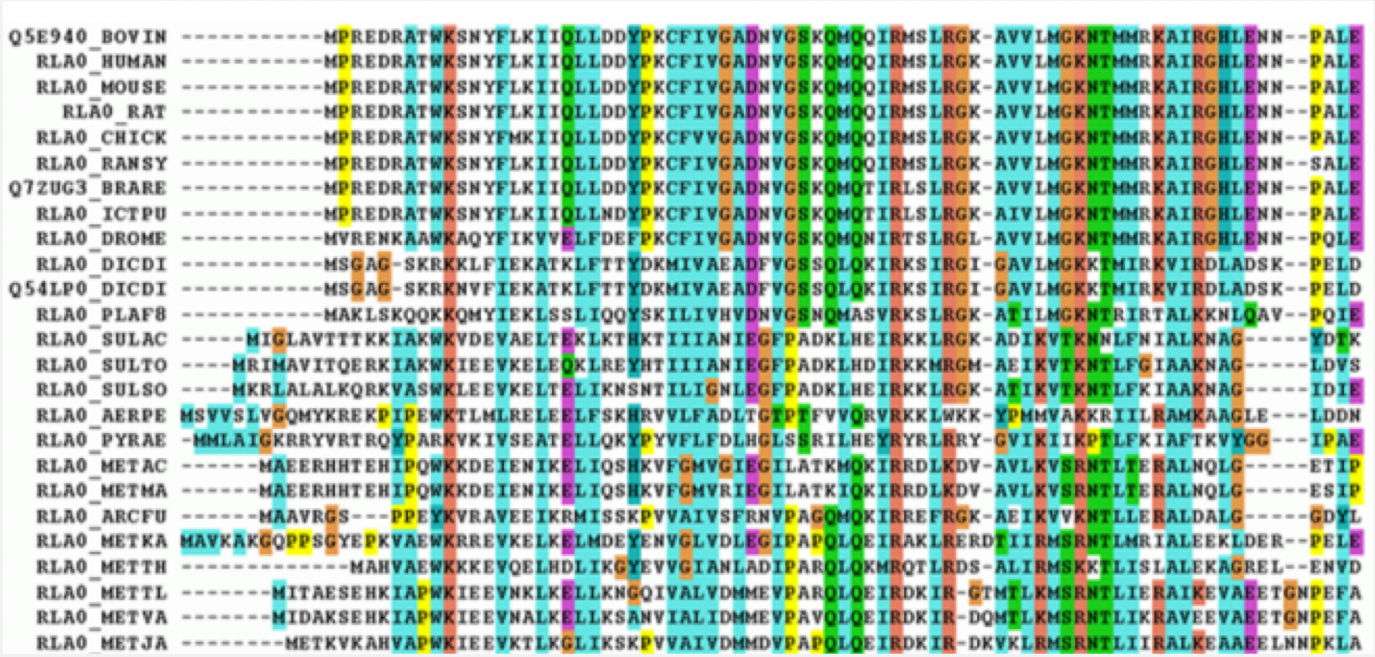


PMID 10360979

PSIC Score

- For each column, calculate frequency of each amino acid:

$$p(a,i) = \frac{n(a,i)_{eff}}{\sum_{b} n(b,i)_{eff}}$$



PMID 10360979

PSIC Score

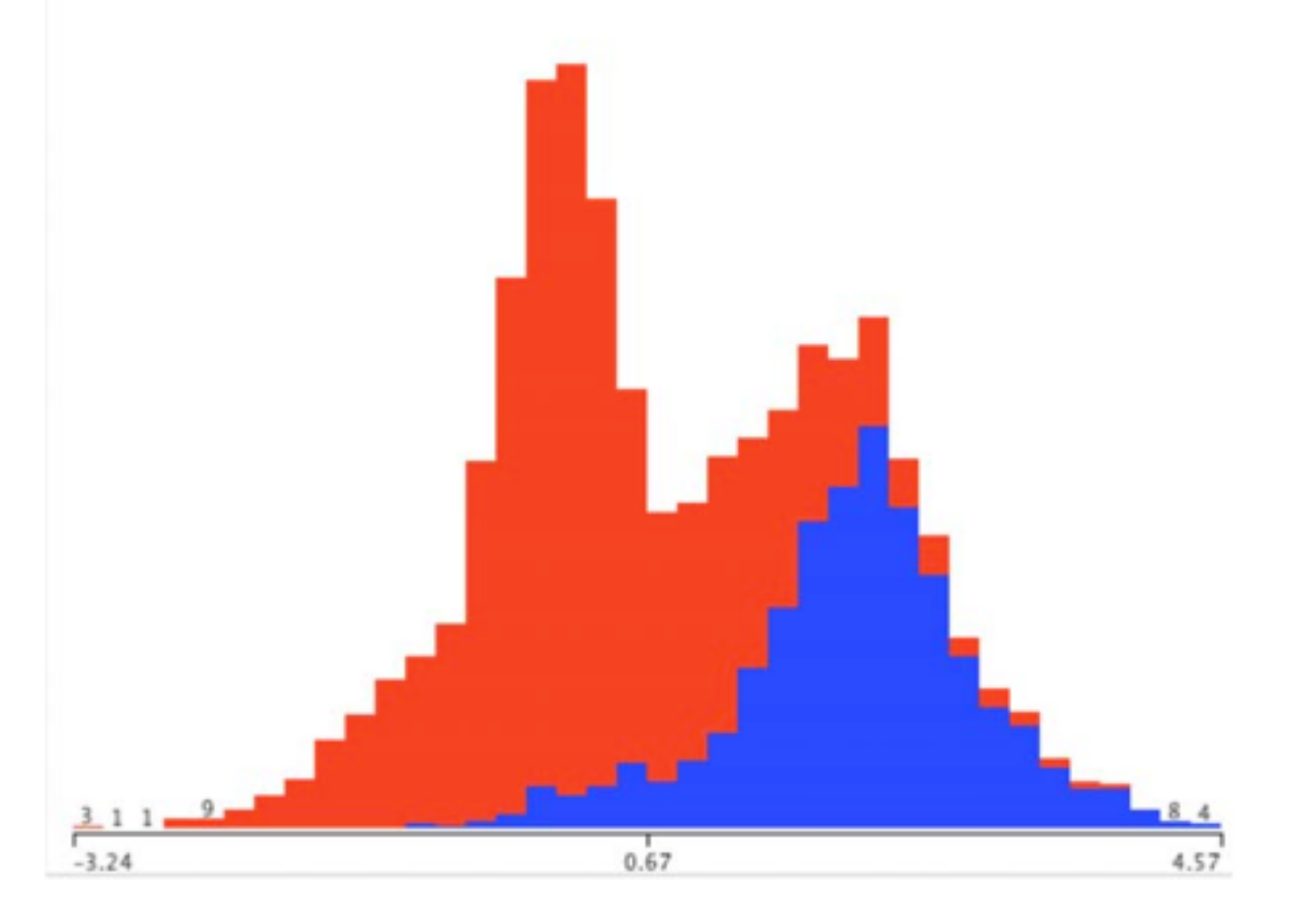
- For each column, calculate frequency of each amino acid:

$$p(a,i) = \frac{n(a,i)_{eff}}{\sum_b n(b,i)_{eff}}$$

- The clever idea: $n(a,i)_{eff}$ is not the raw count of amino acid 'a' at position *i*.
- The raw count *n(a,i)* is adjusted to account for the many closely related sequences present in the MSA.
- PSIC score of a SNP a ➔ b at position *i* is given by:

$$\mathrm{PSIC}(a \to b,i) \propto \ln \frac{p(b,i)}{p(a,i)}$$

PSIC Score histogram from HumDiv



Classification

- Naive Bayes method
- A type of classifier. Other classification algorithms include “Support Vector Machine”, “Decision Tree”, “Neural Net”, “Random Forest” etc.
- Sometimes called “Machine Learning”
- What is a classification algorithm?
- What is a Naive Bayes method/classifier?

Classifiers

- $X_{11}, X_{12}, X_{13}, \ldots, X_{1n},$
- $X_{21}, X_{22}, X_{23}, \ldots, X_{2n},$
- ...
- $X_{i+1,1}, X_{i+1,2}, X_{i+1,3}, \ldots, X_{i+1n},$
- $X_{i+2,1}, X_{i+2,2}, X_{i+2,3}, \ldots, X_{i+2n},$
- ...

+
+
i Positive examples

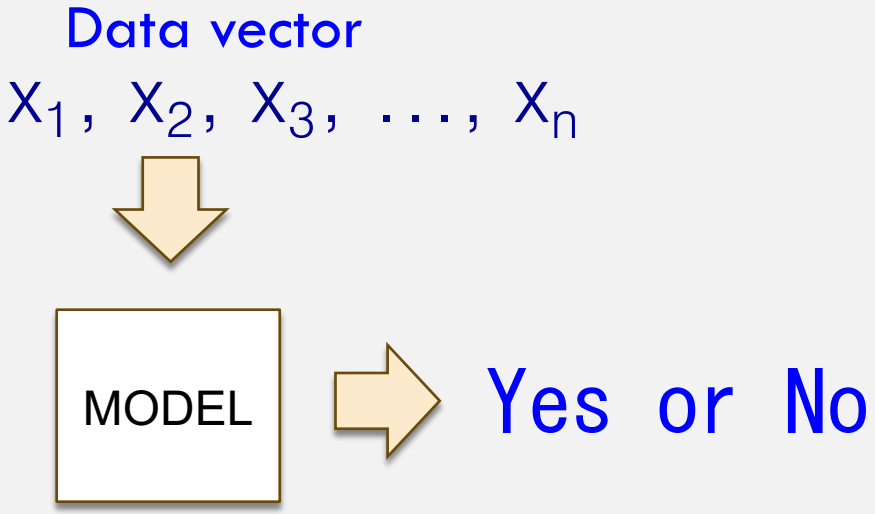
−
−
j Negative examples

"Training Data"

"Supervised learning"

MODEL

Classifiers



Naive Bayes Classifier

"Training Data"

$$\Pr(x_1 \mid +), \Pr(x_1 \mid -),$$
$$\Pr(x_2 \mid +), \Pr(x_2 \mid -), \ldots$$
$$\Pr(x_n \mid +), \Pr(x_n \mid -),$$

Bayesian inference:

Expresses how a subjective assessment of likelihood should rationally change to account for evidence

$$\Pr(+ \mid x_1, x_2, \ldots, x_n) \propto \Pr(x_1 \mid +)\Pr(x_2 \mid +)\ldots\Pr(x_n \mid +)\Pr(+)$$

$$\Pr(- \mid x_1, x_2, \ldots, x_n) \propto \Pr(x_1 \mid -)\Pr(x_2 \mid -)\ldots\Pr(x_n \mid -)\Pr(-)$$

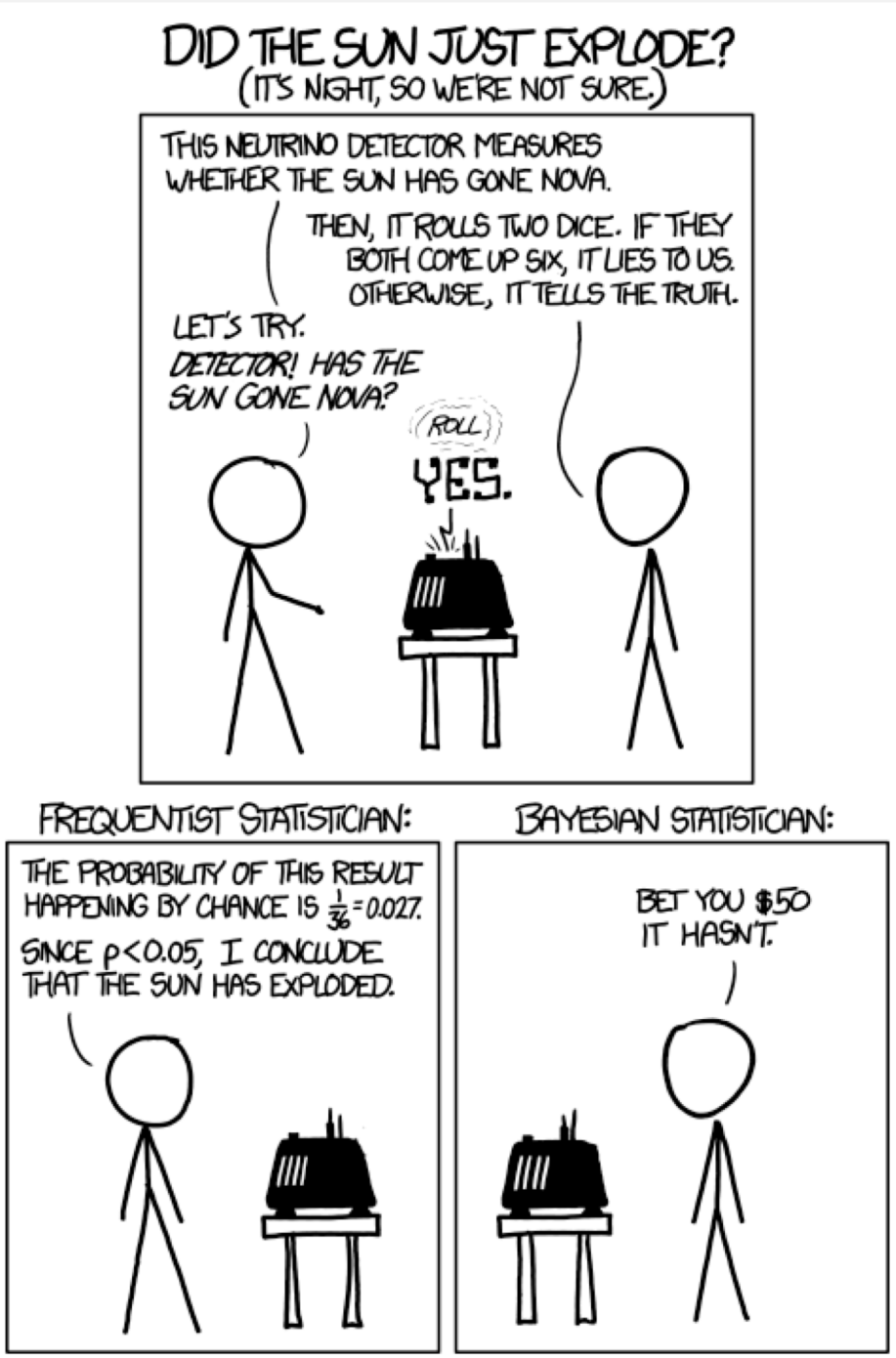
→ + or −

Bayesian probability

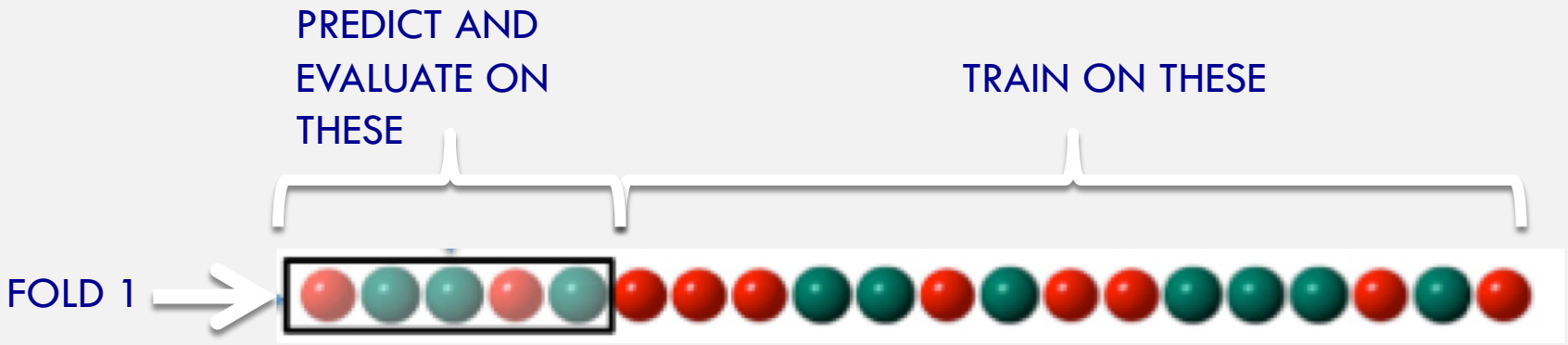
- In statistics, *frequentists* and *Bayesians* often disagree.

- A *frequentist* is a person whose long-run ambition is to be wrong 5% of the time.
- A *Bayesian* is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.

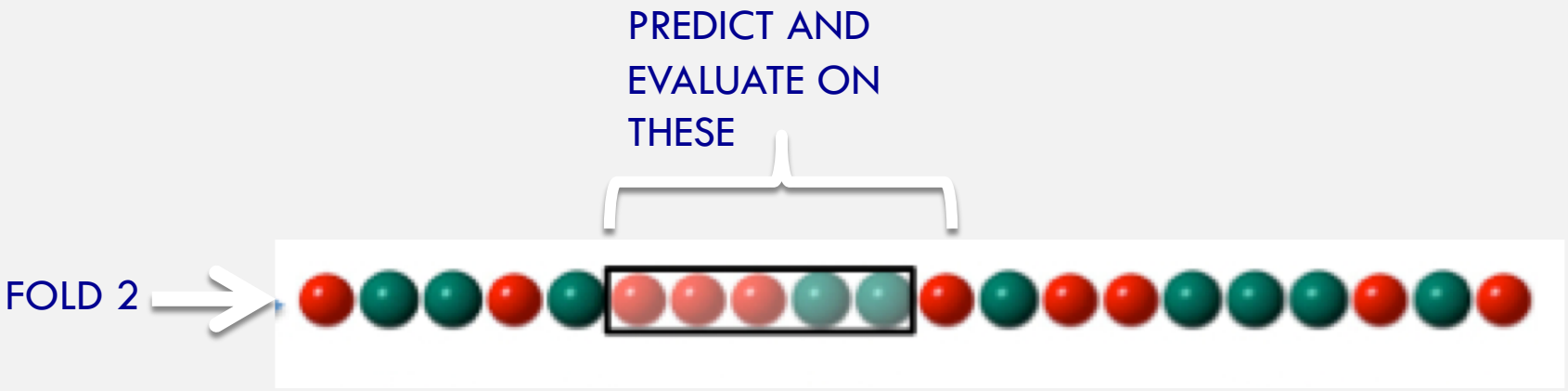
Or…


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

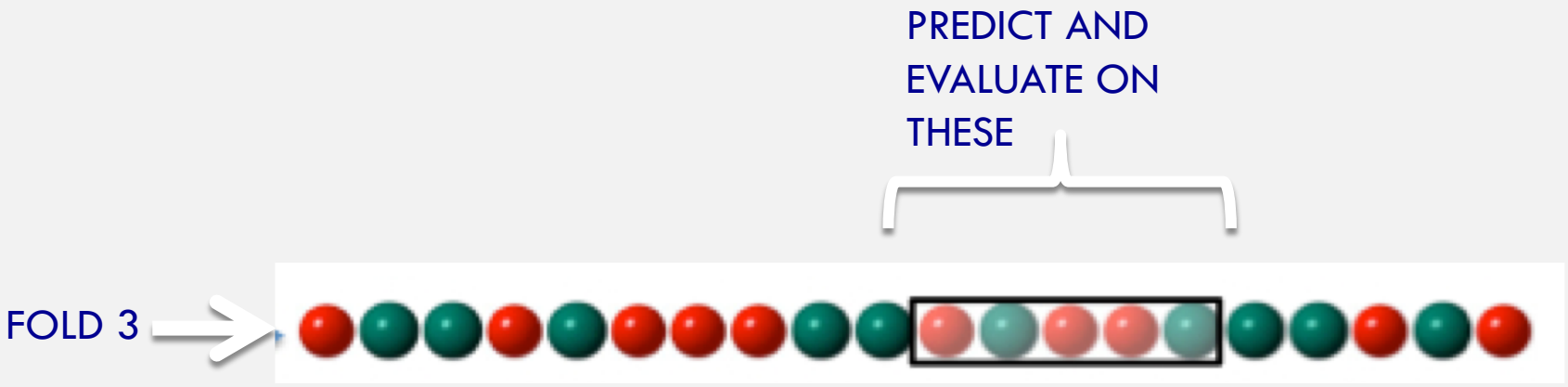
Evaluating a classifier: Cross-validation



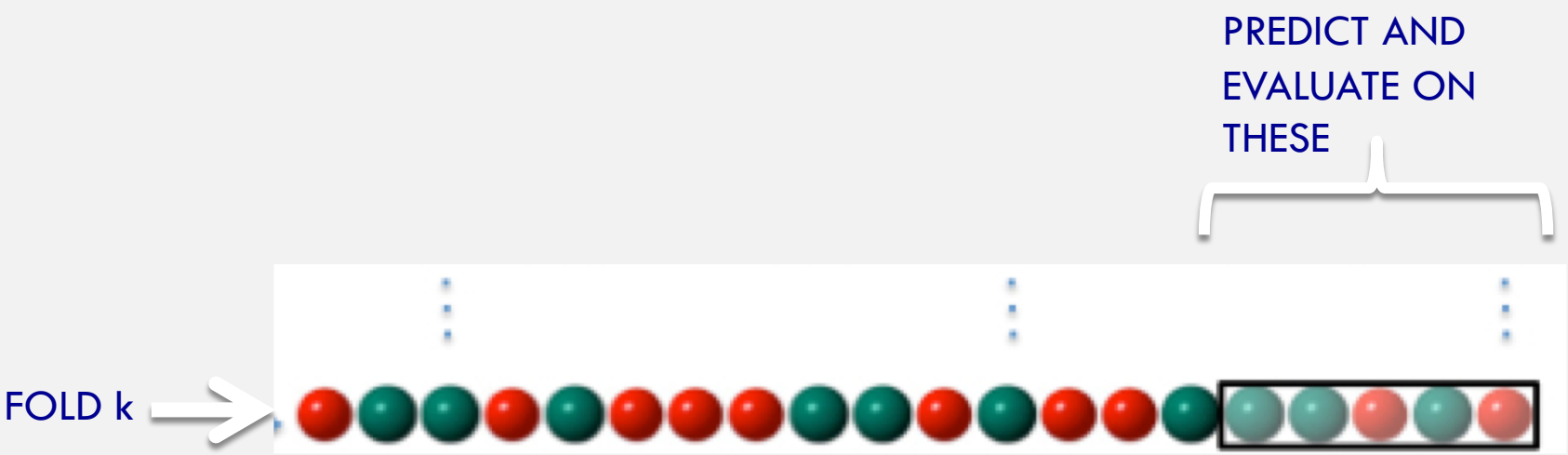
Evaluating a classifier: Cross-validation



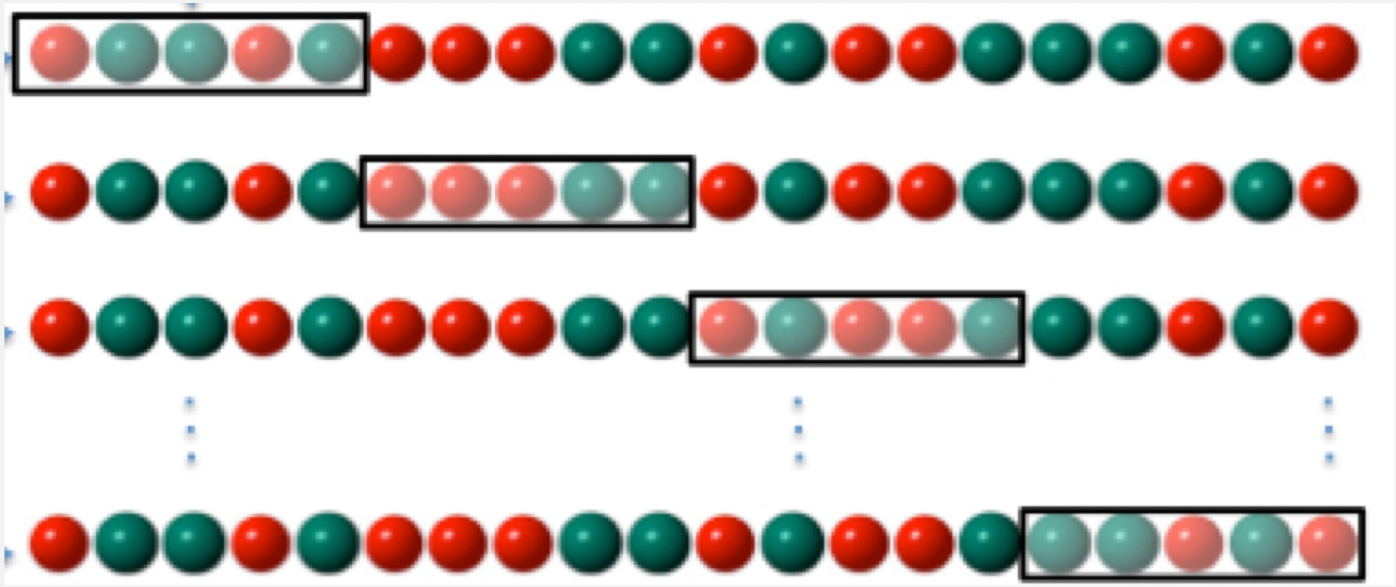
Evaluating a classifier: Cross-validation



Evaluating a classifier: Cross-validation



Evaluating a classifier: Cross-validation



Collect all evaluation results (from k "FOLD"s)

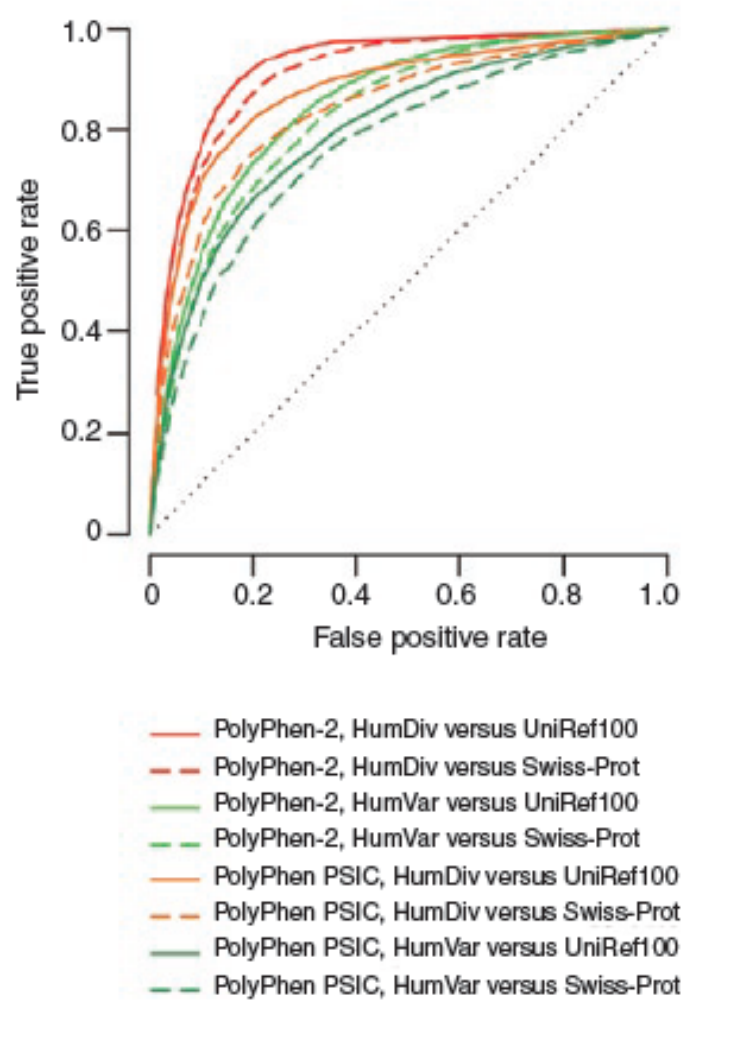
Evaluating classification performance

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Wikipedia

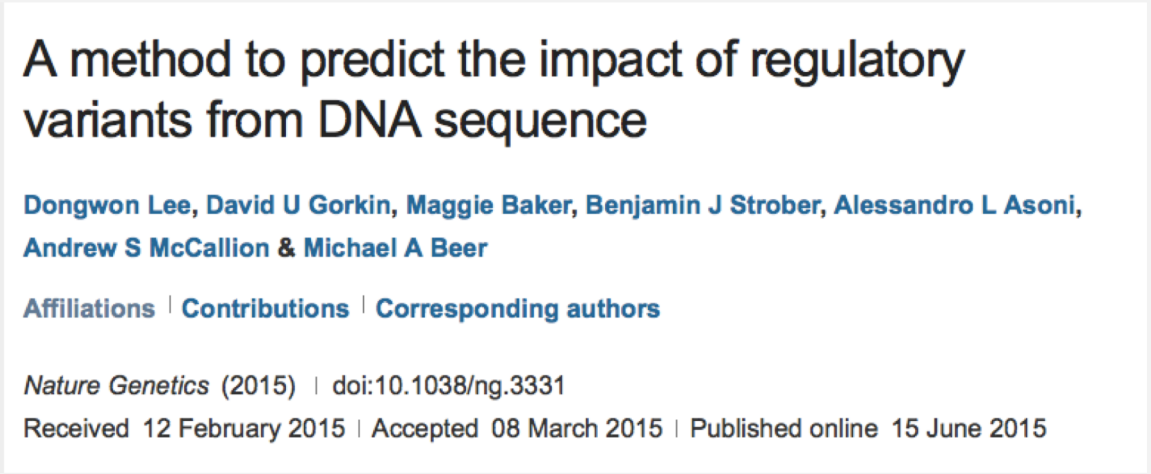
ROC of PolyPhen 2.0 on HumDiv

The Receiver Operating Characteristic (ROC) curve: True +ve vs False +ve



What about SNPs outside coding regions?

- Generally hard enough to predict within coding regions – regulatory sequences notoriously hard to pin down (see ENCODE controversy)
- One interesting new approach uses Support Vector Machine (SVM) classifiers to describe damage to cell-specific regulatory motif vocabularies.



A method to predict the impact of regulatory variants from DNA sequence

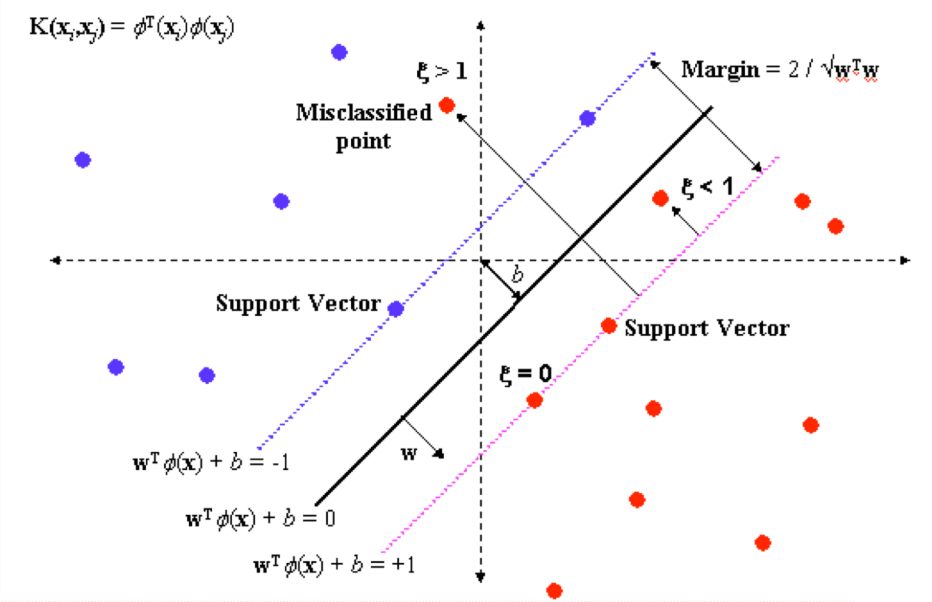
Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion & Michael A Beer

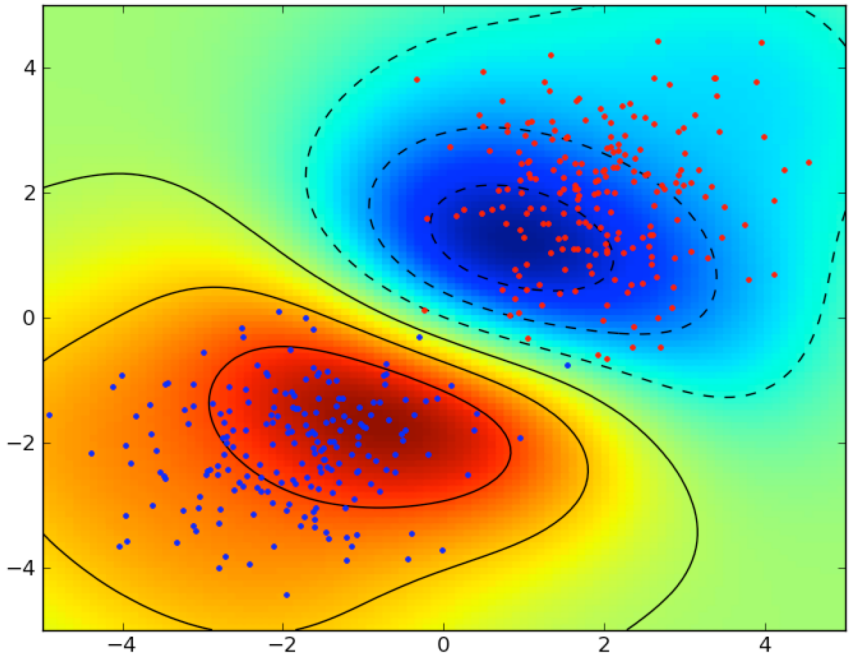
Affiliations | Contributions | Corresponding authors

Nature Genetics (2015) | doi:10.1038/ng.3331

Received 12 February 2015 | Accepted 08 March 2015 | Published online 15 June 2015

Support Vector Machines





Genome-wide Association Studies (GWAS)

http://www.ploscollections.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002828

http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002822

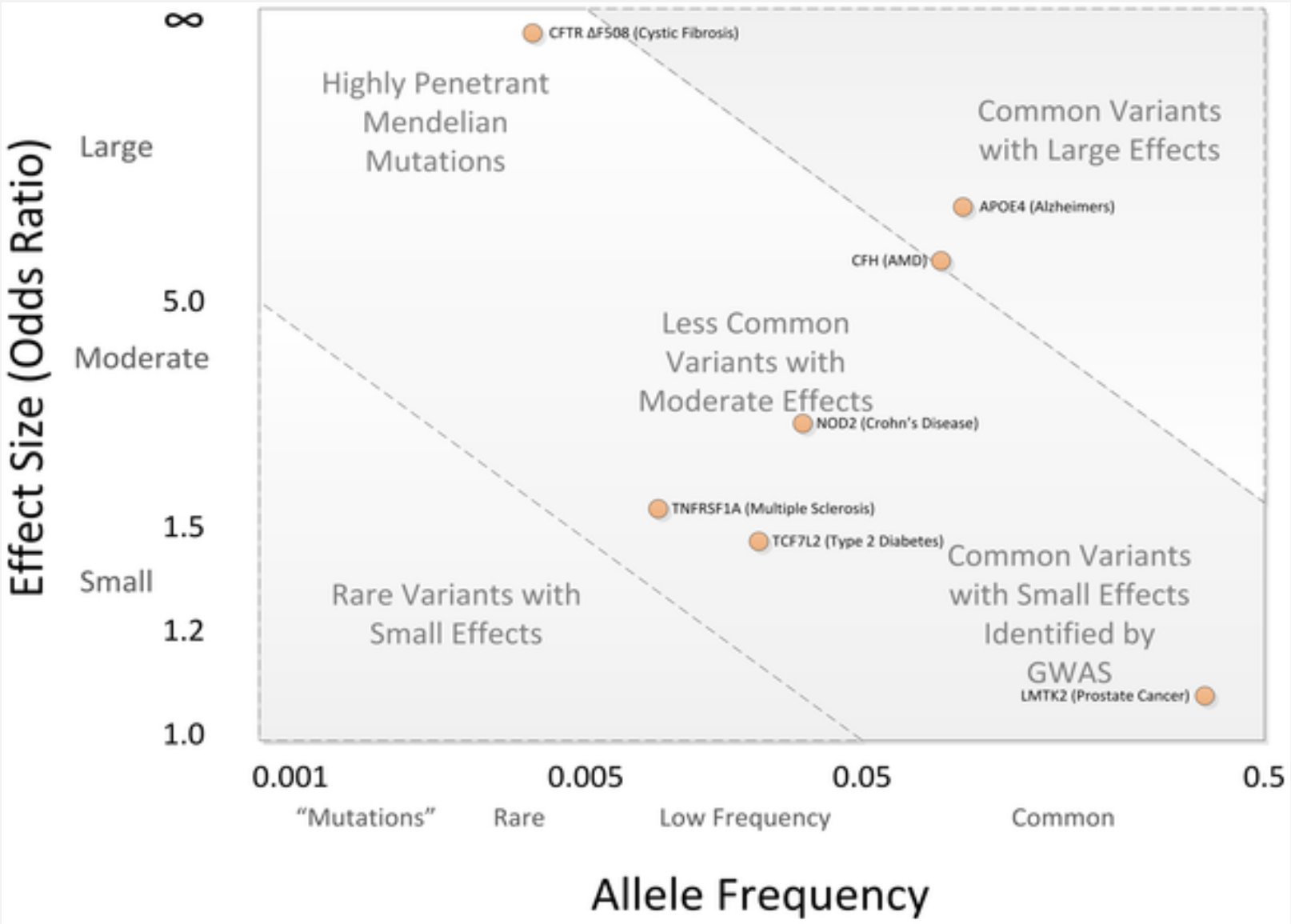
Genetic linkage analysis

- Cystic Fibrosis and the CFTR gene mutations.
- "Linkage analysis"
  - Genotype members of a family (with some individuals carrying the disease)
  - Find a genetic marker that correlates with disease
  - Disease gene lies close to this marker.
- Linkage analysis less successful with common diseases, e.g., heart disease or cancers.

Common disease common variant

- Hypothesis that common diseases are influenced by genetic variation that is "common" in the population
- Implications:
  - Any individual variation (SNP) will have relatively small correlation with disease
  - Multiple common alleles *together* influence the disease phenotype
- Argument for population-based studies versus family based studies. (Think about it!)

Figure 1. Spectrum of Disease Allele Effects.



Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 8(12): e1002822. doi:10.1371/journal.pcbi.1002822
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002822

GWAS: Genotyping methodology

- Microarray technology to assay 0.5 – 1 million or more SNPs, e.g. Affymetrix and Illumina
- One population may need more SNPs to be put on the chip than another population
- Increasingly, people are using whole-genome sequencing. But LD limits utility, arrays still have advantages.

GWAS: Phenotyping methodology

- Case/control vs. quantitative
  - Quantitative (e.g. blood pressure, LDL levels)
  - Case/control (qualitative, disease vs. no disease)

- Possible to look at more than one phenotype? Electronic medical records (EMR) for phenotyping?

GWAS: a very simple idea

- Case/control:

		Disease?
I1:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	–
I2:	AACGAGCTAGCGATCGATCGACAACGACTACGAGGT	+
I3:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	–
I4:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	–
I5:	AACGAGCTAGCGATCGATCGACAACGACTACGAGGT	+
I6:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	–
I7:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	–
I8:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	–

GWAS Gotchas

- Before we start on the stats, some gotchas:
  - Correlation is not causation
  - Population structure (see later)
  - Linkage disequilibrium (see later)
  - Phenotyping

- Also, even if it all works, can be hard to interpret
  - Say a SNP correlates well with heart disease
    - Could be a direct biochemical link
    - Could be behavioral (makes you like bacon…)

GWAS statistics: case vs control

	Has ‘A’	Has ‘T’
Case	3	1
Control	1	9

- The Fisher Exact test

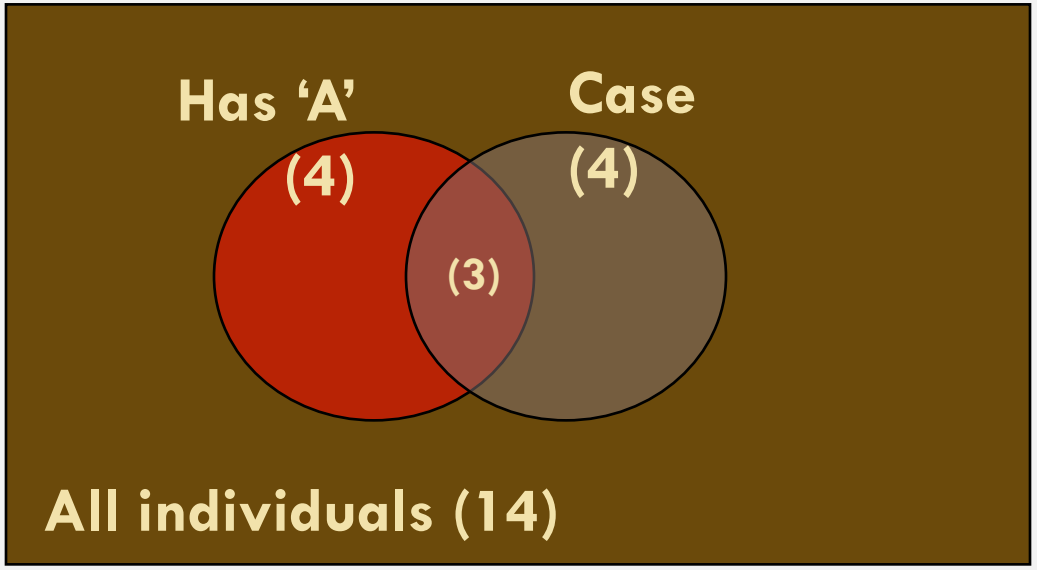
```
I1:     AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I2:     AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I3:     AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I4:     AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I5:     AACGAGCTAGCGATCGATCGACAACGACTACGAGGT  +
I6:     AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I7:     AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I8:     AACGAGCTAGCGATCGATCGACAACGACTACGAGGT  +
I9:     AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I10:    AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  +
I11:    AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I12:    AACGAGCTAGCGATCGATCGACTACGACTACGAGGT  -
I13:    AACGAGCTAGCGATCGATCGACAACGACTACGAGGT  -
I14:    AACGAGCTAGCGATCGATCGACAACGACTACGAGGT  +
```

GWAS statistics: case vs control

- The Fisher Exact test

	Has 'A'	Has 'T'
Case	3	1
Control	1	9

p-value < 0.05



GWAS statistics: case vs control

- Instead of the Fisher Exact test, can use the "Chi Squared test".
- Do this test with EACH SNP separately. Get a p-value for each SNP.
- The smallest p-values point to the SNPs most associated with the disease

Association tests: Allelic vs Genotypic

- What we saw was an “allelic association test”. Test if ‘A’ instead of ‘T’ at the position correlates with disease
- Genotypic association test: Each position is not one allele, it is two alleles (e.g, A & A, T & T, A & T).
- Correlate genotype at that position with phenotype of individual

Genotypic association tests

- Various options
- Dominant model

	AA or AT	TT
Case	?	?
Control	?	?

Genotypic association tests

- Various options
- Recessive model

	AA	AT or TT
Case	?	?
Control	?	?

Genotypic association tests

- Various options
- 2 x 3 table

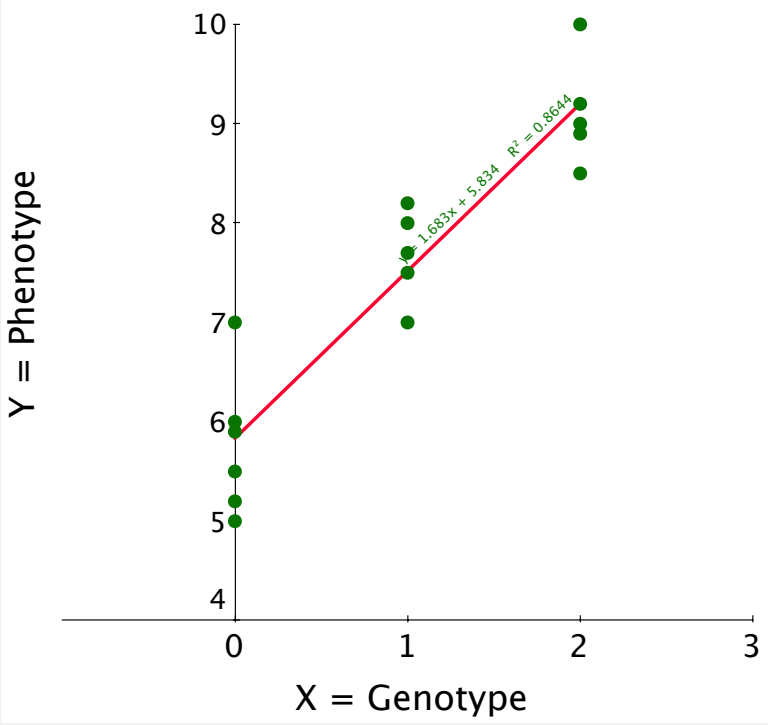
	AA	AT	TT
Case	O_{11}	O_{12}	O_{13}
Control	O_{21}	O_{22}	O_{23}

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

Chi-squared test

Quantitative phenotypes

- Y_i = Phenotype value of Individual i
- X_i = Genotype value of Individual i



$Y = a + bX$

If no association, $b \approx 0$

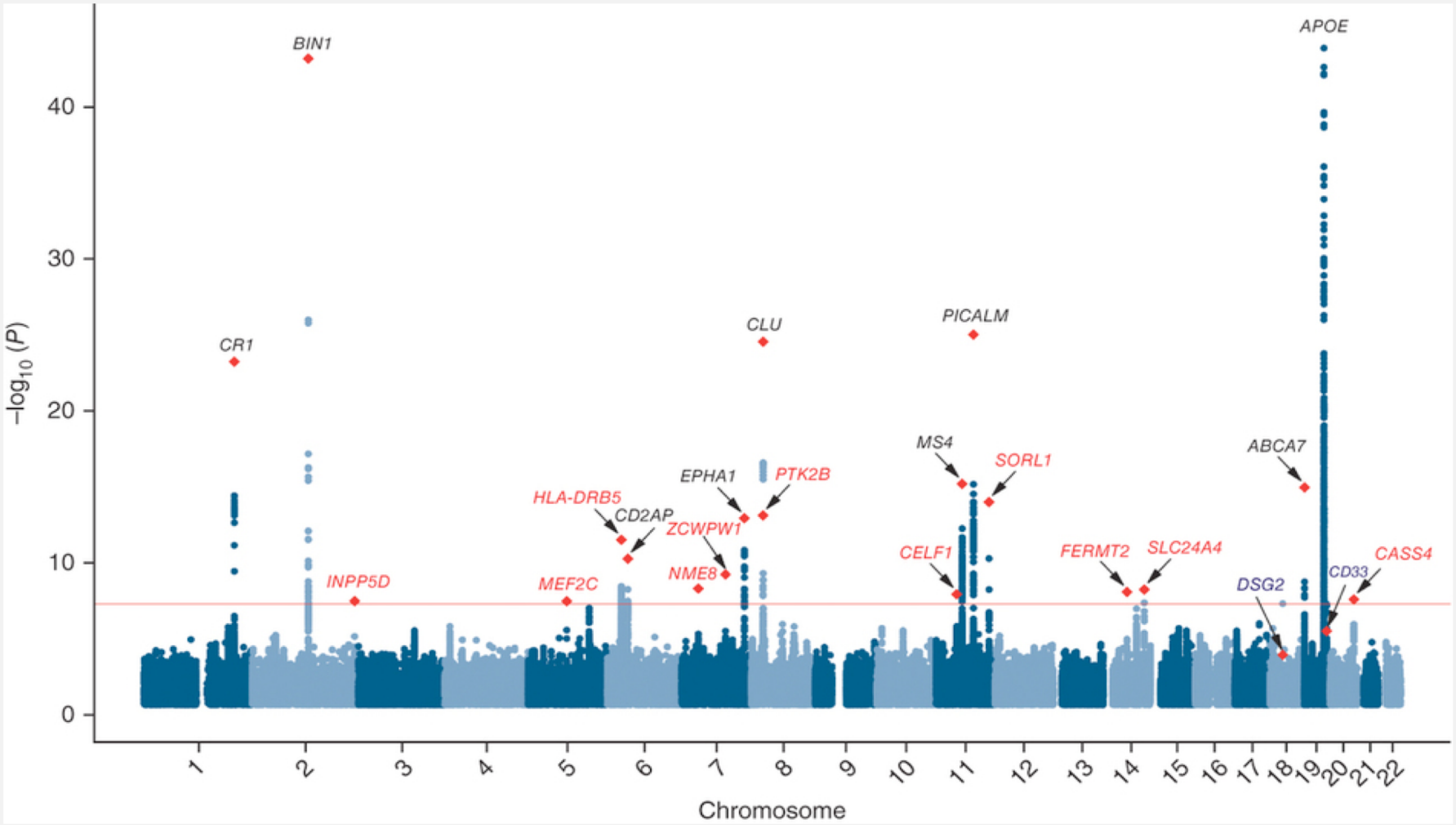
The more b differs from 0, the stronger the association

This is called "linear regression"

Quantitative phenotypes

- Another statistical test commonly used on such GWAS matrices is “ANOVA” (Analysis of Variance)
- Statistical models for GWAS can get quite involved – can give refs on request.

Manhattan plot

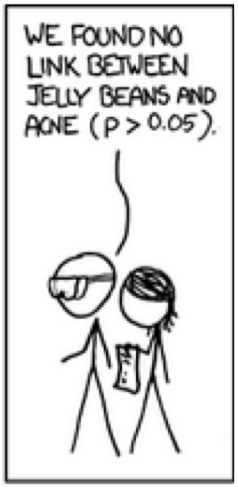


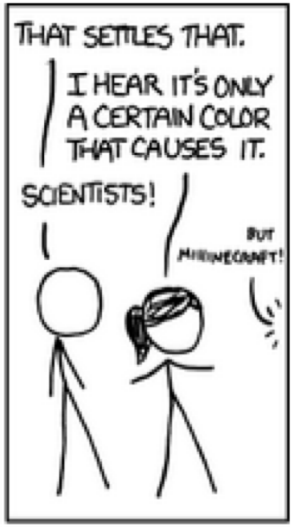
Lambert et al., 2013: Nature Genetics 45, 1452

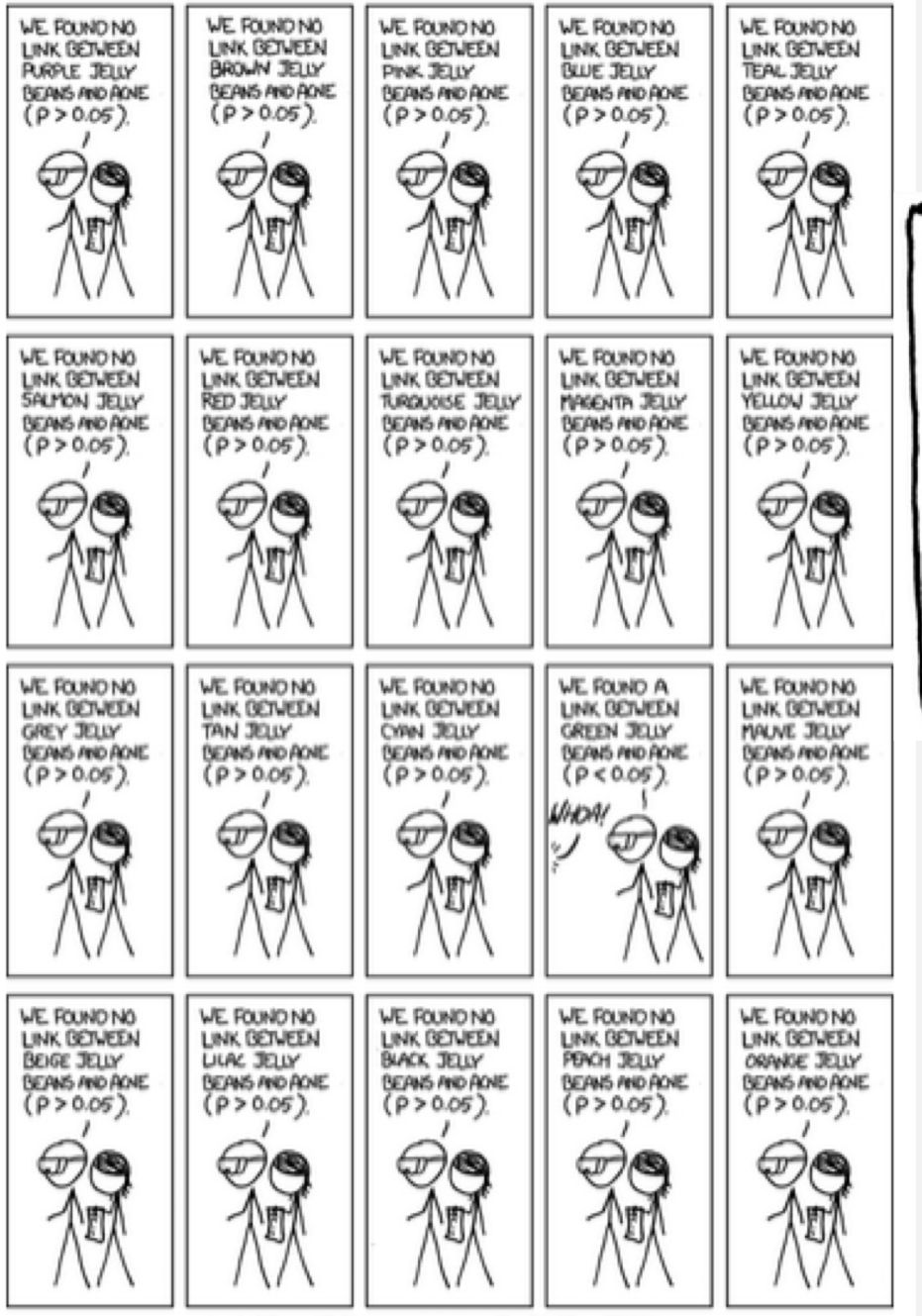
Multiple hypothesis correction

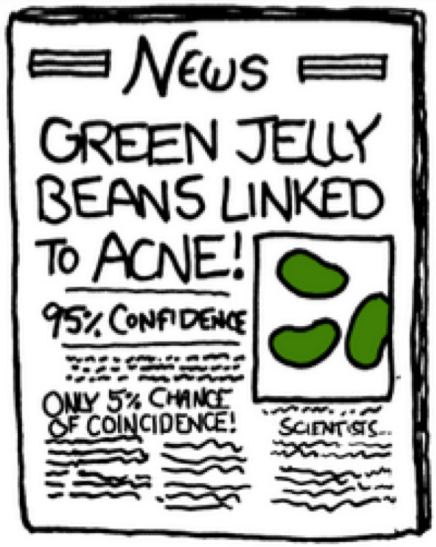
- What does the "p-value of an association test = 0.01" mean ?
- It means that the observed correlation between genotype and phenotype has only 1% probability of happening just by chance. Pretty good?
- But if you repeat the test for 1 million SNPs, 1% of those tests, i.e., 10,000 SNPs will show this level of correlation, *just by chance (and by definition).*
- http://xkcd.com/882/


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.


WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).


THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Bonferroni correction

- Multiply p-value by number of tests.
- So if the original test on a particular SNP gave a p-value of p, define the new p-value as $p' = p \times N$, where N is the number of SNPs tested (1 million ?)
- With $N = 10^6$, a p-value of 10^{-9} is downgraded to $p' = 10^{-9} \times 10^6 = 10^{-3}$. This is quite good.

False Discovery Rate

- Bonferroni correction will “kill” most reported associations (reduced statistical power)
- Too stringent for most applications (although good if it works). Need to balance false positive rate with false negative rate
- False Discovery Rate (FDR) is an alternative procedure to correct for multiple hypothesis testing, which is less stringent.

False Discovery Rate

- Given a threshold α (e.g., 0.05):
- Sort all p–values (N of them) in ascending order:
- $p_1 \leq p_2 \leq \ldots \leq p_N$
- Count for each group of N, p from 1 to i:

$$p_i^{'} = p_i \times \frac{N}{i}$$

- Require p' < α
- This ensures that the expected proportion of false positives in the reported associations is < α

Beyond single locus associations?

- We tested each SNP separately
- Recall that our "common disease, common variant" hypothesis meant each individual SNP carries only a small effect.
- Maybe two SNPs together will correlate better with phenotype.
- So, methods for 2-locus association study.
- Main problem: Number of pairs ~ N^2

Beyond the probed SNPs?

- The SNP-chip has a large number of probes (e.g., 0.5 – 1 Million). But still, way fewer SNPs than WGS.

- But there are many more sites in the human genome where variation may exist. Are we going to miss any causal variant outside the panel of ~1 Million?

- Not necessarily.

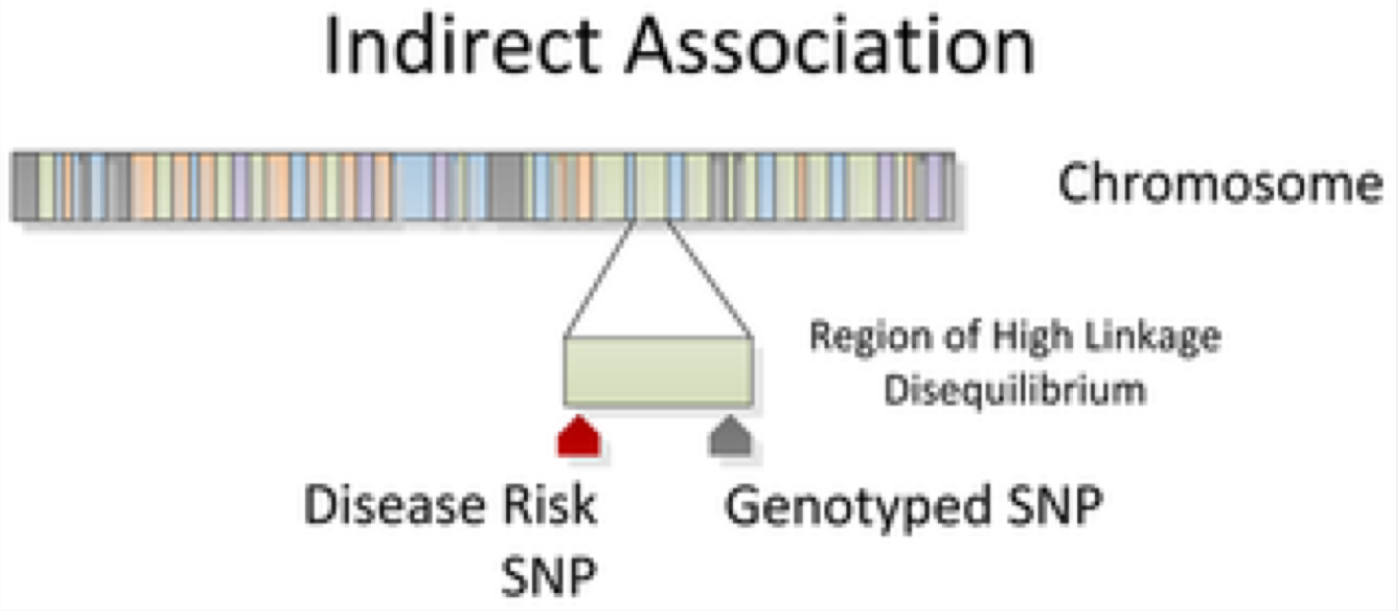
Linkage disequilibirum

- Two sites close to each other may vary in a highly correlated manner. This is linkage disequlibrium (LD).
- Not enough recombination events have happened to make the inheritance of those two sites independent.
- If two sites are in a segment of high LD, then one site may serve as a "proxy" for the other.

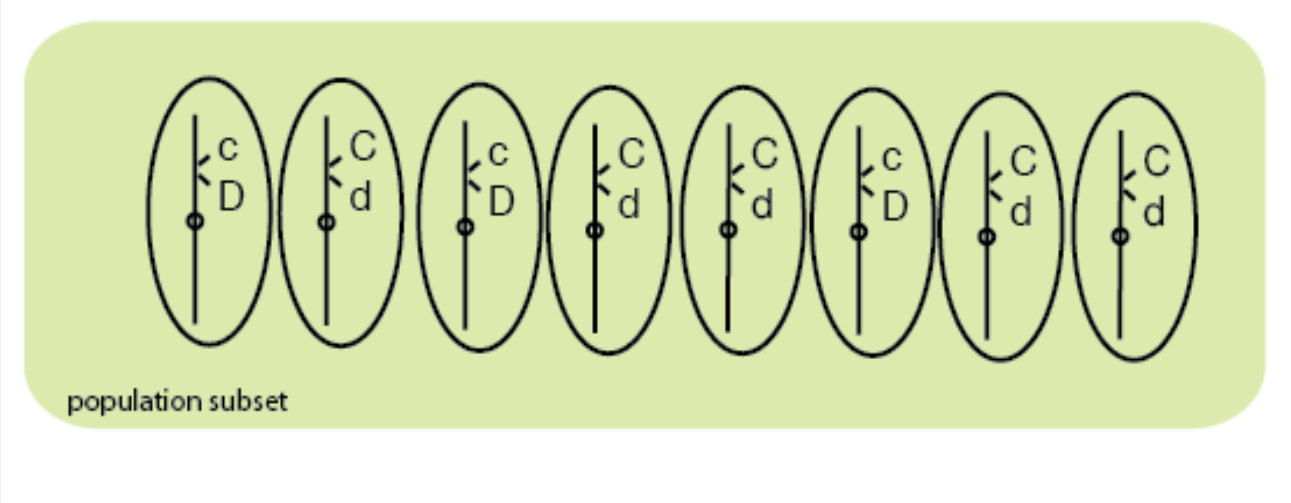
LD and its impact on GWAS

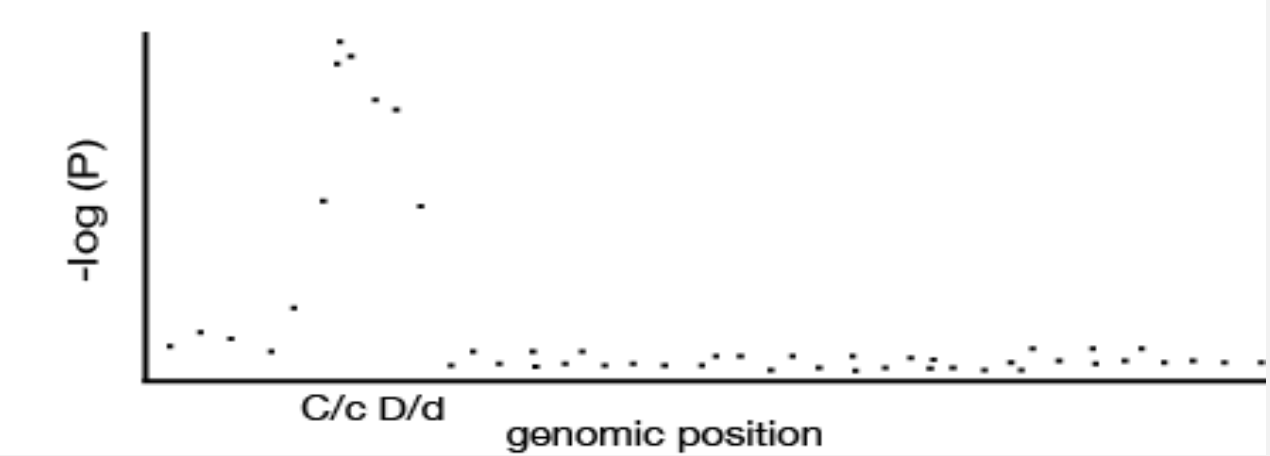
- If sites X & Y are in high LD, and X is on the SNP-chip, knowing the allelic form at X is highly informative of the allelic form at Y.
- So, a panel of 0.5 – 1 Million SNPs may represent a larger number, perhaps all of the common SNPs.
- But this also means: if X is found to have a high correlation with disease, the causal variant may be Y, and not X

LD and its impact on GWAS

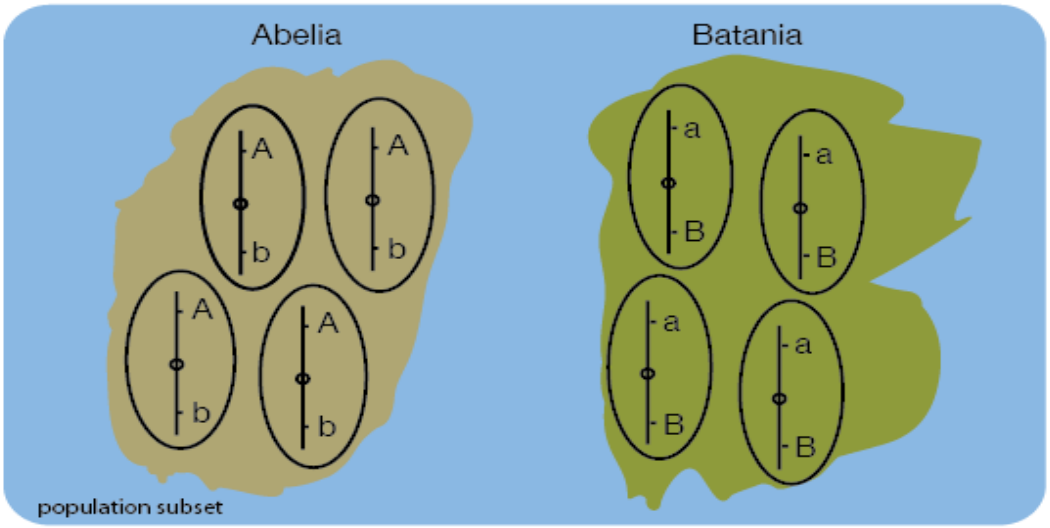


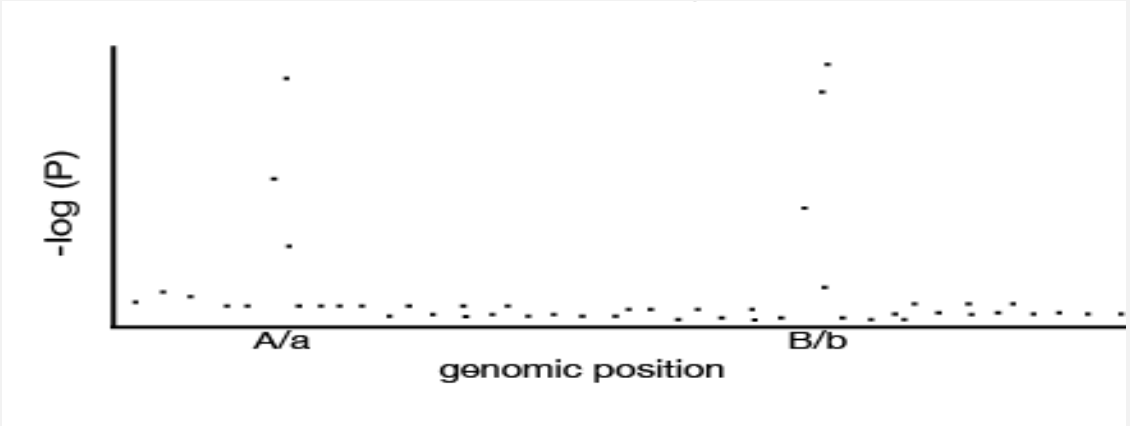
LD impact


c
D
C
d
c
D
C
d
C
d
c
D
C
d
C
d
population subset


-log (P)
C/c D/d
genomic position

Population structure





Discussions

- In many cases, able to find SNPs that have significant association with disease. Risk factors, some mechanistic insights.
- GWAS Catalog : http://www.genome.gov/26525384
- Yet, final predictive power (ability to predict disease from genotype) is limited for complex diseases.
- “Finding the Missing Heritability of Complex Diseases” http://www.genome.gov/27534229

Discussions

- Increasingly, whole-exome and even whole-genome sequencing used for variant detection
- Taking on the non-coding variants. Use functional genomics data as template
- Network-based analysis rather than single-site or site-pairs analysis
- Complement GWAS with family-based studies