Exercises

dN/dS = 0.34 for a gene. Is this most consistent with positive or negative selection? If 66% of nonsynonymous sites are under strong purifying selection, what is dN/dS for the remaining sites? 34% have dN/dS ~ 0 so remaining must be 1: 0.34 = 0*.66 + X*0.34

Make it more clear if numbers are different..
dN/dS = 0.30 for a gene. If 60% of nonsynonymous sites are under strong purifying selection, what is dN/dS for the remaining sites?

dN/dS = average (strong purifying and remaining)
0.30 = 0.6*0 + 0.4*X
X = 0.3/0.4
X = 0.75

If we use original numbers:
0.34 = 0.66*0 + 0.34*X
X = 1

Exercises

1) Why are GC ending codons at high frequency in highly expressed genes, but low frequency in low expressed genes? Selection for translation accuracy/speed increases with expression level, when expression is low mutation biases dominate

2) Are close or distantly related species better for identifying short functional sequences? distant because pii decreases with substitution rate and influences probability of short sequence

3) What is the most likely function of conserved noncoding sequences? regulatory sequences

4) What causes variation in the mutation rate across the genome? CpG, repeats, biased gene conversion, NOT GC content.

5) Why is the frequency of CpG sites lower than expected? When they are methylated they have higher mutation rate and loss of CpG site

Exercises

6) Why are CpG islands unmethylated? methylation causes silencing, also if they were methylated the CpG sites would be lost through mutation

7) Why do CpG sites violate assumptions of nucleotide substitution models? Mutation rate is dependent on the prior base, sites are not independent, not time-homogenous. Ergocity, independence, time-reversible

8) What is the expected dN/dS ratio for a psuedogene – a duplicate gene that has become non-functional? 1

9) How does recombination influence the substitution rate? Biased gene conversion

10) In humans most coding sequences are (conserved/unconserved) and most conserved sequences are (Coding/Noncoding)

Today's objectives

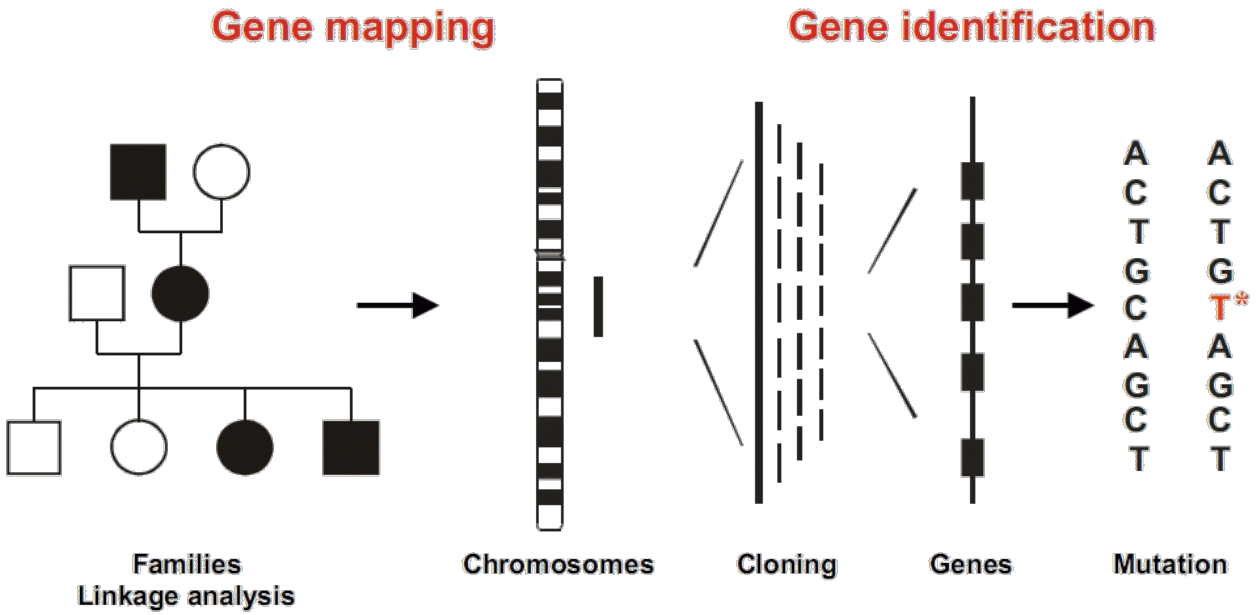
Comparative genomics

- Identifying disease mutations
- Deleterious variants
- Confusion matrix
- ROC curves
- Cancer drivers

Identifying disease mutations

Pre-genomics era

- Linkage mapping – genetic markers physically linked to a disease mutation should co-segregate with it, i.e. correlation between markers and disease
- Positional cloning – isolation of partially overlapping DNA segments from genomic libraries that contain a specific gene



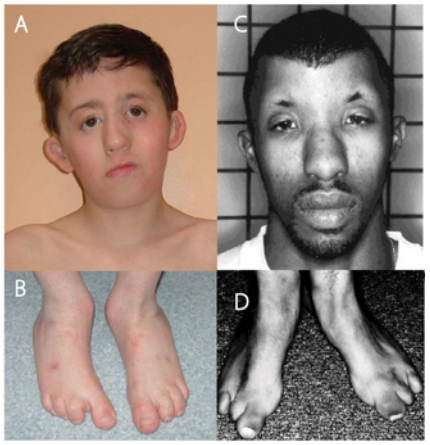
Huntington's Disease
- first disease gene mapped
- chr4 linkage 1983, Venezuela
- HD caused by $CAG_{>40}$ repeat expansion (heterogeneity) 1993
- HTT function is still unclear

Cystic fibrosis
- first (1989) positionally cloned gene (CFTR – ion channel needed for epithelial fluid transport)

Evidence: at least three families, biochemical evidence, not in healthy people

Post-genomics disease mapping

Miller syndrome: rare genetic disorder with craniofacial and limb deformations.
- micrognathia (small jaw), cleft lip and/or palate, hypoplasia (underdeveloped) limbs



Exome sequencing
- four affected individuals in three independent families
- 1 mutation found in four affected individuals within dihydroorotate dehydrogenase 1 (DHODH) involved in pyrimidine biosynthesis
- All four individuals are compound heterozygotes for missense mutations predicted to be damaging by (PolyPhen – a comparative genomics prediction algorithm)

Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)		Kindreds 1+2		Kindreds 1+2+3	
Filter	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362	3,099	1,810	2,654	1,525
Not in dbSNP129	641	102	647	114	369	53	105	25	63	21
Not in HapMap 8	898	123	923	128	506	46	117	7	38	4
Not in either	456	31	464	33	228	9	26	1*	8	1*
Predicted damaging	204	6	204	12	83	1	5	0	2	0

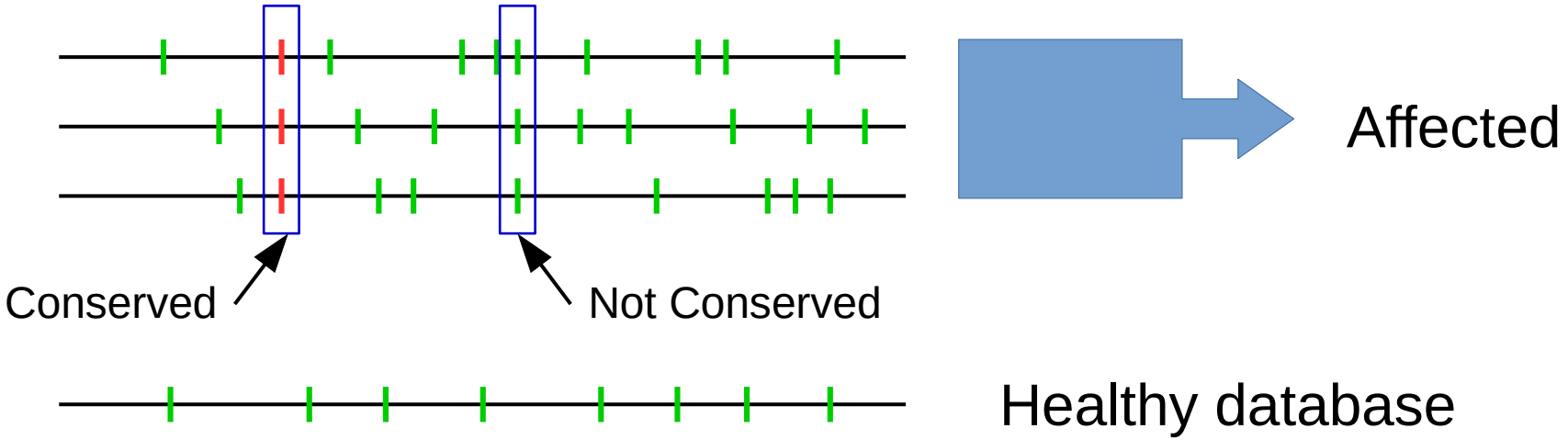
NS = nonsynonymous,
SS = splice site, I = coding indel

Conservation

Additional families

Ng et al. (2010)

Disease causing mutations



- Rare variant
- Causal mutation

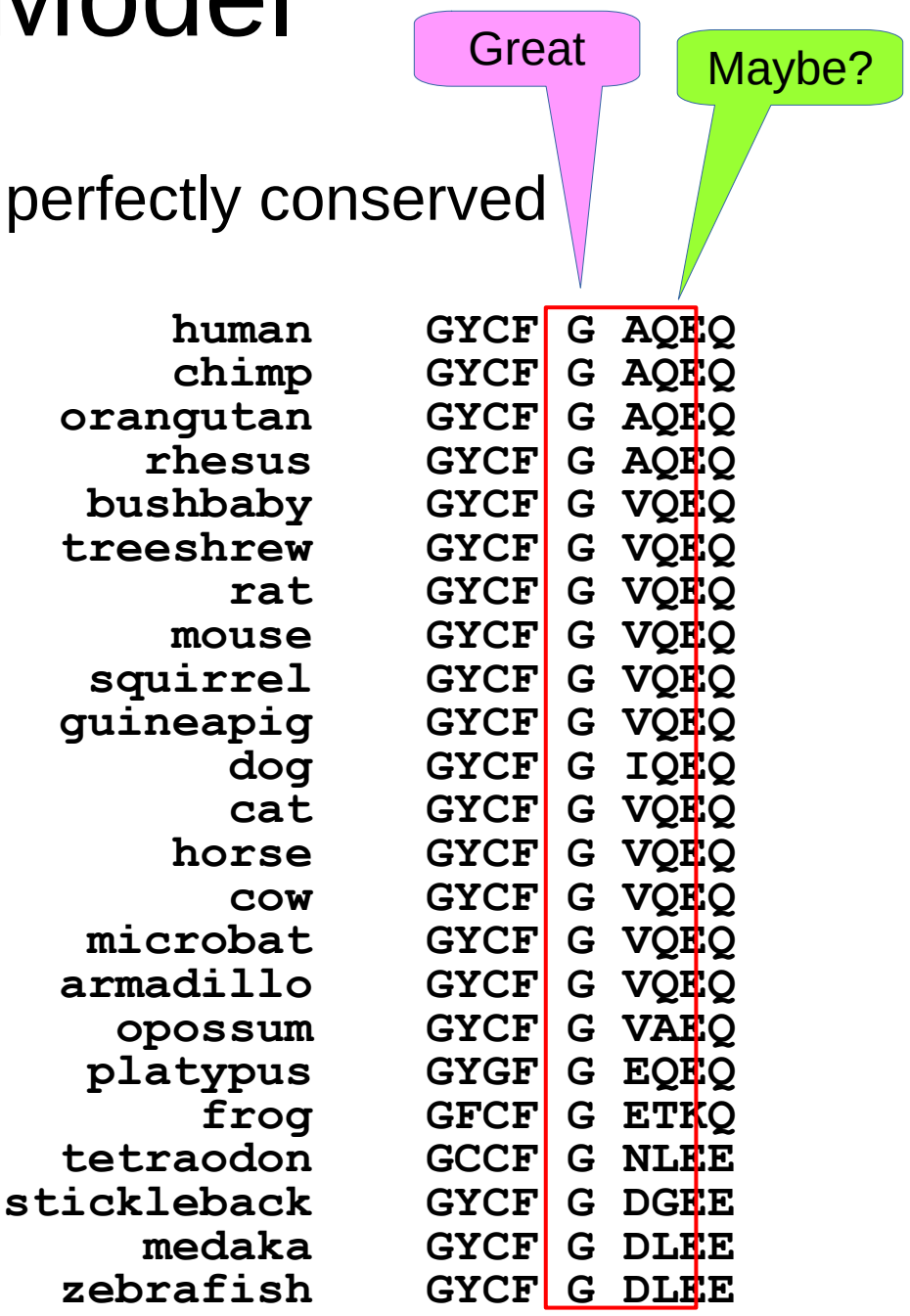
Ideas

- affected: have (same or different) mutations in the same gene
- causal mutation: not present in health individuals
- causal mutation: disrupts conserved position

Simple Model

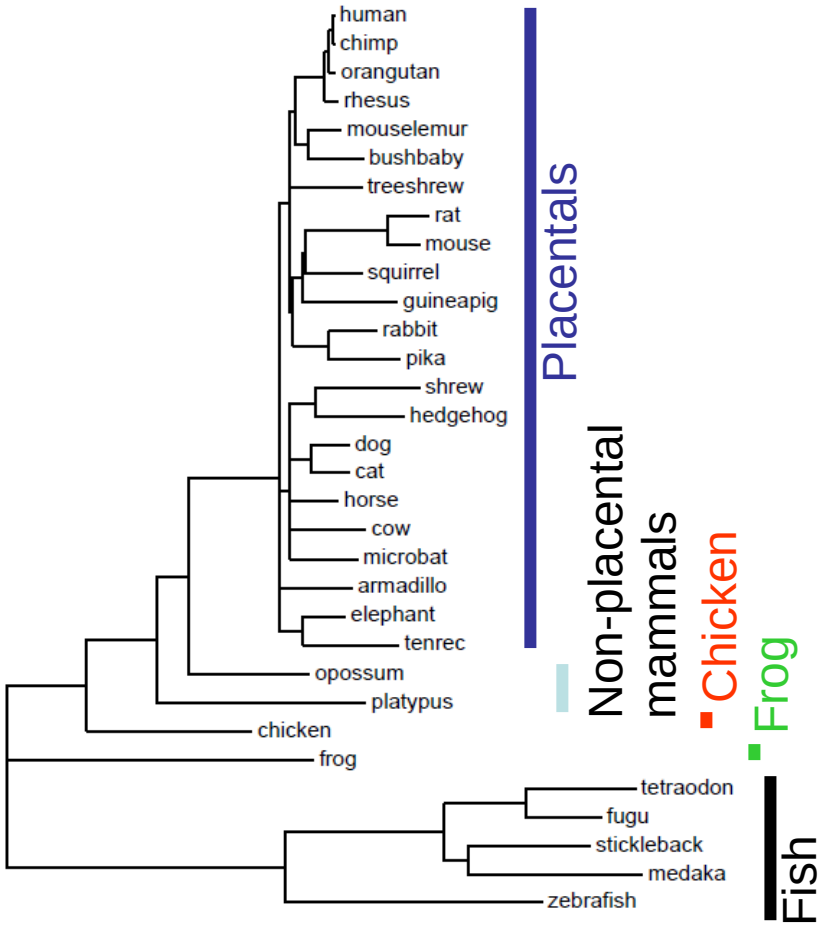
Proposal: Disease mutations disrupt perfectly conserved (100% identity) sites

Problem 1: How many species, which species?



A solution: dN/dS for single codons

32 vertebrate species
18,993 alignments
d_S = 12.2 subs/site



$$LLR = \log \frac{L(D|T,\theta,d_N = \hat{C}d_S)}{L(D|T,\theta,d_N = d_S)}$$

```
      human    GYCF G AQEQ
      chimp    GYCF G AQEQ
  orangutan    GYCF G AQEQ
     rhesus    GYCF G AQEQ
   bushbaby    GYCF G VQEQ
  treeshrew    GYCF G VQEQ
        rat    GYCF G VQEQ
      mouse    GYCF G VQEQ
   squirrel    GYCF G VQEQ
  guineapig    GYCF G VQEQ
        dog    GYCF G IQEQ
        cat    GYCF G VQEQ
      horse    GYCF G VQEQ
        cow    GYCF G VQEQ
    microbat   GYCF G VQEQ
   armadillo   GYCF G VQEQ
     opossum   GYCF G VAEQ
    platypus   GYGF G EQEQ
       frog    GFCF G ETKQ
   tetraodon   GCCF G NLEE
stickleback    GYCF G DGEE
     medaka    GYCF G DLEE
   zebrafish   GYCF G DLEE
```

Simple Model

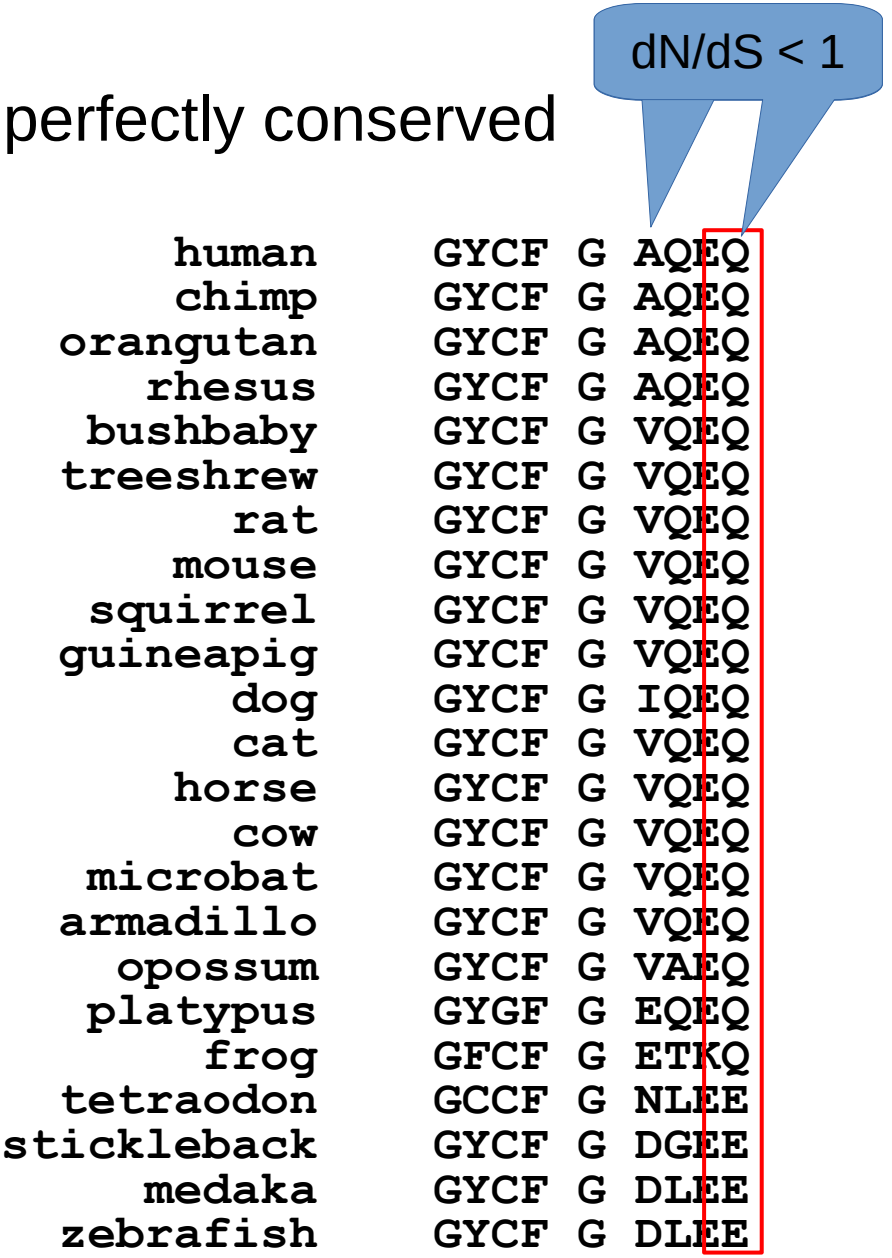
Proposal: Disease mutations disrupt perfectly conserved (100% identity) sites

Problem 1: How many species, which species?

Problem 2: Do all amino acid substitutions cause disease?

Q-E has Blossum62 score of +2
Q-F has Blossum62 score of -2

Need to account for species diversity & types of changes



A solution: Position Specific Scoring Matrix (PSSM)

PAM and BLOSSUM

- PAM matrix indicates the likelihood of a substitution over time
- Likelihood is over MANY sites in MANY proteins
- Base on observed changes and frequency of amino acids

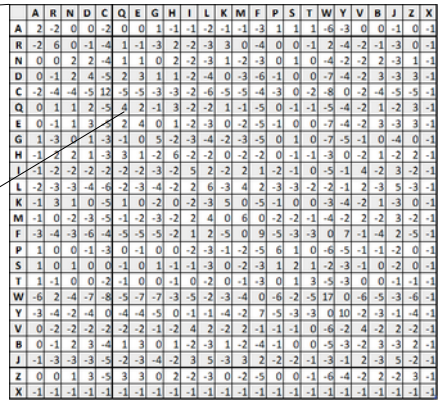
PSSM

- PSSM matrix indicates the likelihood of an amino acid at each position
- Likelihood is over ONE site in ONE protein
- Base on observed frequencies of amino acids

Protein alignment

```
GKFELRGIPPAP
.....N....D.
```

	A	R	N	D	C	Q
A	2	-2	0	0	-2	0
R	-2	6	0	-1	-4	1
N	0	0	2	2	-4	1
D	0	-1	2	4	-5	2
C	-2	-4	-4	-5	12	-5



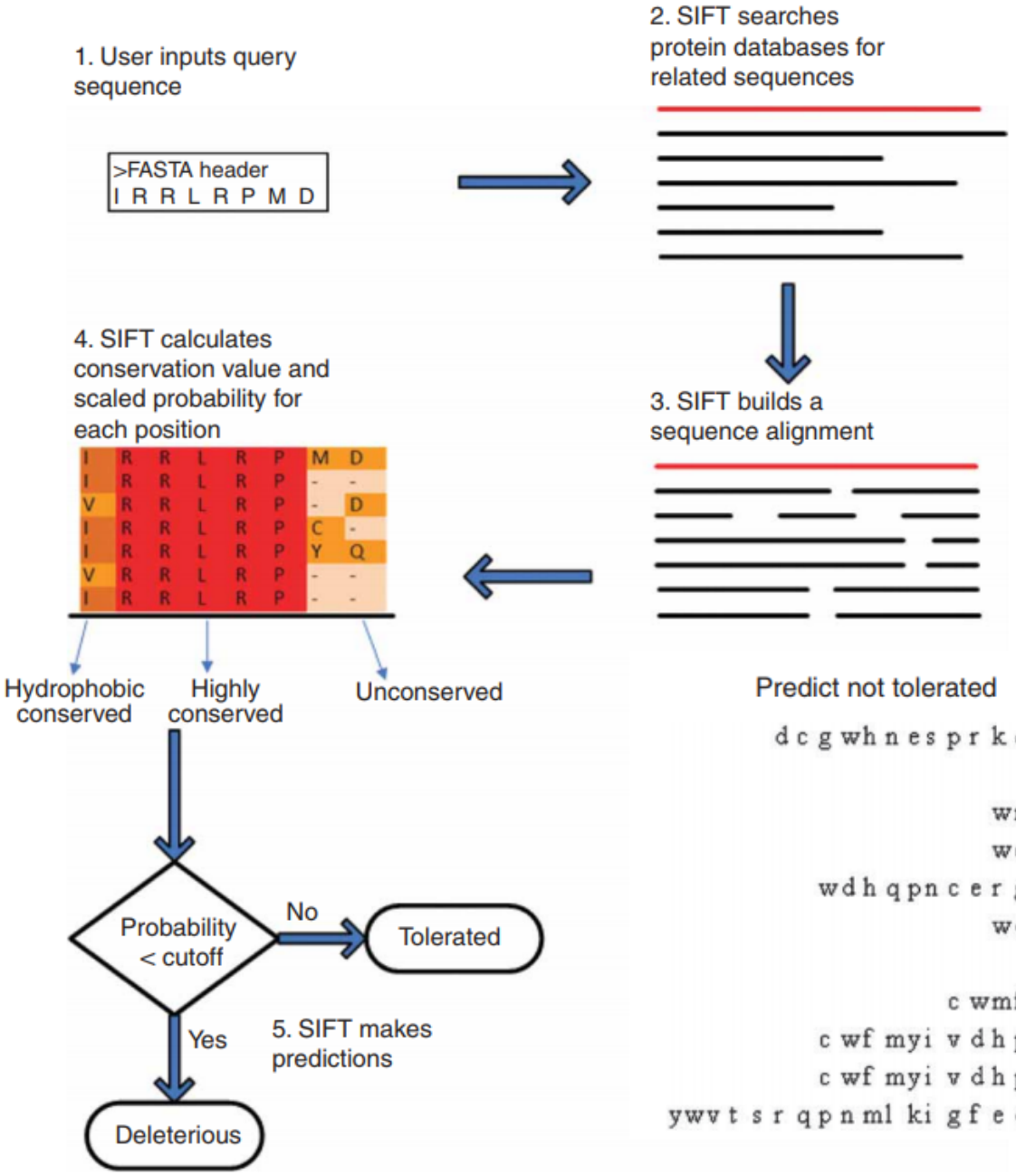
```
frog          GFCF A ETKQ
tetraodon     GCCF A NLEE
stickleback   GYCF D DGEE
medaka        GYCF N DLEE
zebrafish     GYCF A DLEE
```

A	3/5
N	1/5
D	1/5

Comparison

- PSSM requires many homologs (matrix can be sparse)
- PSSM can be more accurate (protein and position specific)

SIFT: sorts intolerant from tolerant substitutions



- Both SIFT and PolyPhen output position specific measures of acceptable vs disease causing mutations based on alignment.
- Similar to PAM score but is protein/position specific

Predict not tolerated	Position	Seq Rep	Predict tolerated
d c g w h n e s p r k q y t	3M	0.60	a f v i ML
w	4A	0.60	c f m y i h v l p r q t n k s e GDA
w m h f	5C	0.80	y i q r e l k d p n v g t a SC
w c m f	6R	0.80	y i h v p l d g n q e t a k SR
w d h q p n c e r g k s	7V	0.80	y t a mF l I V
w d h g	8I	0.80	n c r q p e y k s f m t AvLI
	9N	0.80	w c h p f y Mi q r v g e d t a k s l N
c w m f i d	10R	0.80	p v y g s l t n a e q Hk R
c w f m y i v d h p g l	11R	0.80	n s t a e k Q R
c w f m y i v d h p g l	12R	0.80	n s t a e k Q R
y w v t s r q p n m l k i g f e d c a	13H	0.80	H

SIFT PSSM

SIFT calculates a substitution score similar to PAM, but it is position-specific.

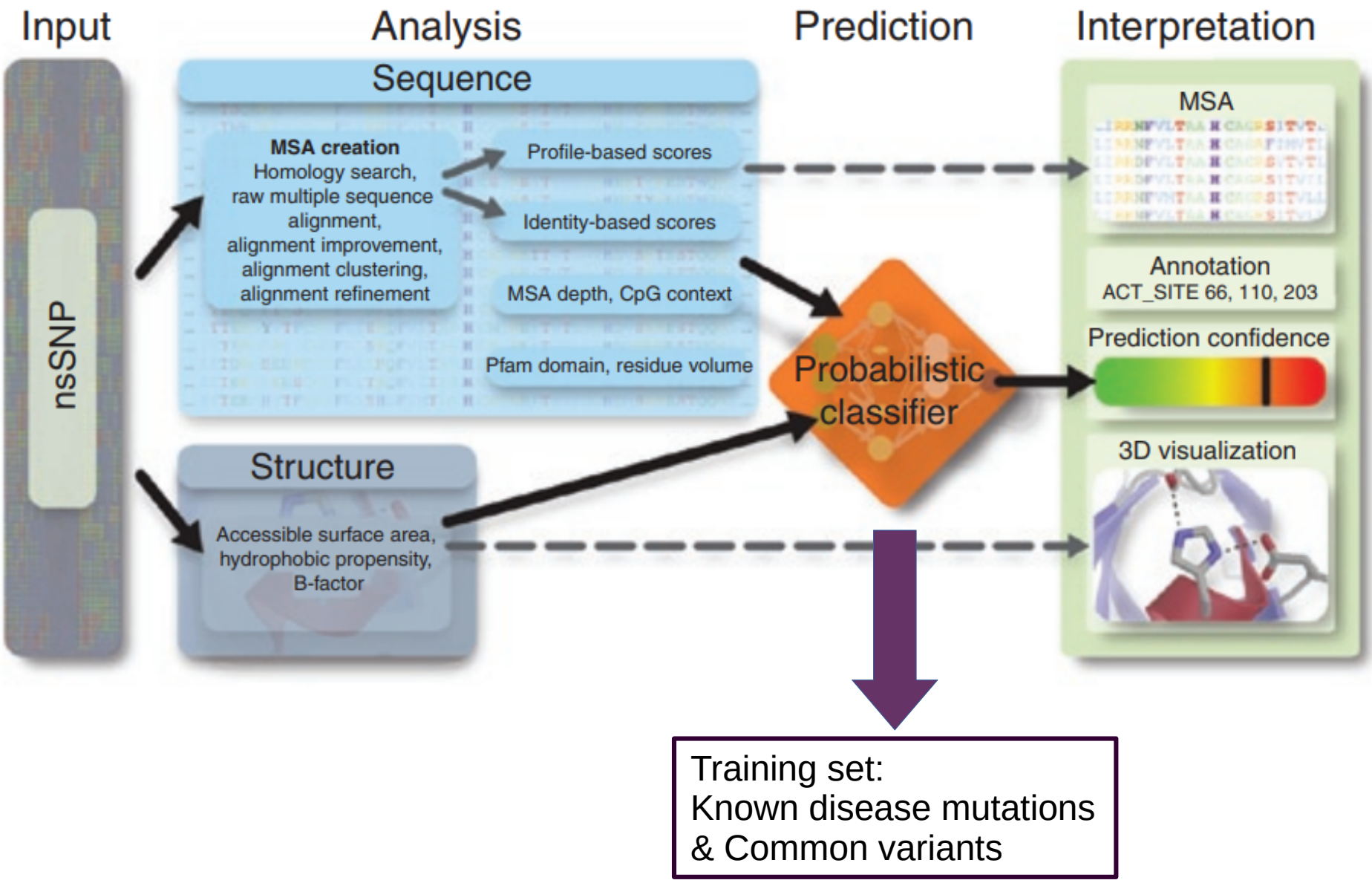
The alignment is converted into a position-specific scoring matrix (PSSM). The score, p_{ca} is a weighted average of the observed amino acid frequencies in the alignment at position c and for amino acid a. When the score < 0.05 it predicts not tolerated.

Alignment	frequency
PS	Position 1 P = 3/5 = 0.6
PS	Position 1 S = 2/5 = 0.4
SS	Position 2 S = 5/5 = 1.0
PS	Position 2 F = 0/5 = 0.0
SS	

SIFT calculates the median information content (conservation score) to measures the diversity of the sequences in the alignment.

SIFT adds pseudocounts based on expected frequencies to account for unobserved amino acids.

PolyPhen


Input
Analysis
Prediction
Interpretation
nsSNP
Sequence
MSA creation
Homology search,
raw multiple sequence
alignment,
alignment improvement,
alignment clustering,
alignment refinement
Profile-based scores
Identity-based scores
MSA depth, CpG context
Pfam domain, residue volume
Structure
Accessible surface area,
hydrophobic propensity,
B-factor
Probabilistic
classifier
MSA
Annotation
ACT_SITE 66, 110, 203
Prediction confidence
3D visualization
Training set:
Known disease mutations
& Common variants

PolyPhen PSSM

PSIC (position-specific independent counts) is used to score conservation:

$$W(a,i) = \ln\left[\frac{p(a,i)}{q_a}\right]$$

The probability p(a,i) of observing amino acid a at the position i after infinitely long evolution, and q_a is the expected frequency of the amino acid in a database of proteins.

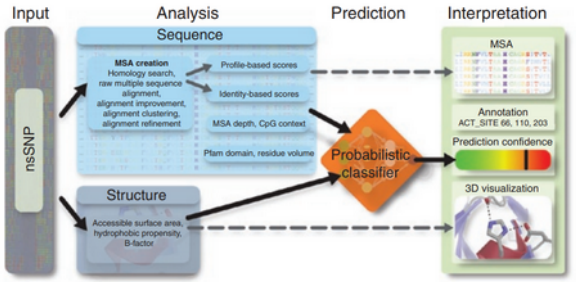
If all sequences were independent, the best estimator for p(a,i) is the raw frequency.

Since the sequences may be strongly dependent, a normalized effective number $n(a,i)_{eff}$ of observations is computed to determine p(a,i)

Human and chimp are closely related, not independent

Sunyaev,S.R., Eisenhaber,F., Rodchenkov,I.V., Eisenhaber,B., Tumanyan,V.G. and Kuznetsov,E.N. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. , 12 , 387 –394.

Evaluation of classification algorithms



- PolyPhen and SIFT are heuristic algorithms, they both make a prediction based on a formula or set of observations rather than a complex likelihood
  - PolyPhen uses structural information and alignment
  - SIFT uses scores that change for each position and each gene
- Problem: we need a means of evaluating how well they work, and which one is better
- Solution: Benchmarking and ROC curves

Benchmarking

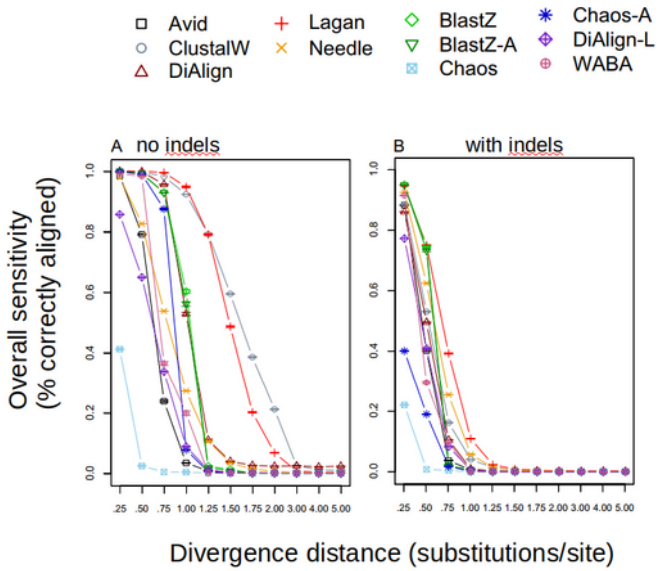
Benchmarking is running a program in order to assess its relative performance, normally through comparison with standards or other programs.

- Performance can be measured in terms of time or memory required to solve a problem.
- Performance can also be measured by how often a program returns a correct result.
- Simulations and datasets with known answers are two ways in which algorithms are commonly benchmarked.

Simulation example

Simulated sequences (true alignment is known)

```
AGCC-GTAC
..T.C...C
```



Evaluate ClustalW & Dialign by testing whether they produce the correct alignment:

ClustalW

```
AG-CCGTAC
..T.....C
```

Dialign

```
AGCC-GTAC
..T.C...C
```

When results can be scored as a binary outcome (true/false), they can be evaluated using a confusion matrix

Confusion matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Data	Truth	SIFT	PPH
1	D	D	D
2	D	B	D
3	D	B	D
4	B	B	D
5	B	B	D
6	B	B	B

SIFT + = deleterious (D)
\- = benign (B)

Predicted \ Truth	+	-
+	1	0
-	2	3

3 true negatives
1 true positive
0 false positive
2 false negative

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power a	b **False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error c	d **True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{1}{\frac{1}{2}\left(\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}\right)}$

$$\text{True positive rate (sensitivity)} = \frac{\text{\# of true positives}}{\text{\# of known positives}} = a / (a+c)$$

$$\text{True negative rate (specificity)} = \frac{\text{\# of true negatives}}{\text{\# of known negatives}} = d / (b+d)$$

$$\text{False discovery rate (FDR)} = \frac{\text{\# of false positives}}{\text{\# of positive predictions}} = b / (a+b)$$

$$\text{False positive rate} = \frac{\text{\# of false positives}}{\text{\# of known negatives}} = b / (b+d)$$

$$\text{Accuracy} = \frac{\text{\# of true positives + \# of true negatives}}{\text{\# of predictions}} = (a+d) / (a+b+c+d)$$

Evaluating disease mutation predictions

To evaluate prediction of disease causing mutations (i.e. not small risk factors) from comparative genomics data, we need:

- True positives - disease mutation database (found without comparative genomics)
- True negatives - common variants in human populations (unlikely to cause disease)

Disease Mutation Databases

Few regulatory mutations known; mostly coding

Table 1 Numbers of different mutations by mutation type present in HGMD Professional 2013.2 and the publicly available version of the database (June 28th 2013)

Mutation type	Total numbers of mutations		
	HGMD Professional	With chromosomal coordinates	Publicly available
Missense substitutions	62,368	61,845	44,933
Nonsense substitutions	15,781	15,574	11,306
Splicing substitutions	13,030	12,538	9,467
Regulatory substitutions	2,751	2,713	1,753
Micro-deletions ≤ 20 bp	21,681	21,134	15,796
Micro-insertions ≤20 bp	8,994	8,721	6,494
Micro-indels ≤20 bp	2,083	2,004	1,459
Gross deletions >20 bp	10,267	0	6,156
Gross insertions/ duplications >20 bp	2,376	0	1,253
Complex rearrangements	1,409	0	946
Repeat variations	421	0	305
Totals	141,161	124,529	99,868

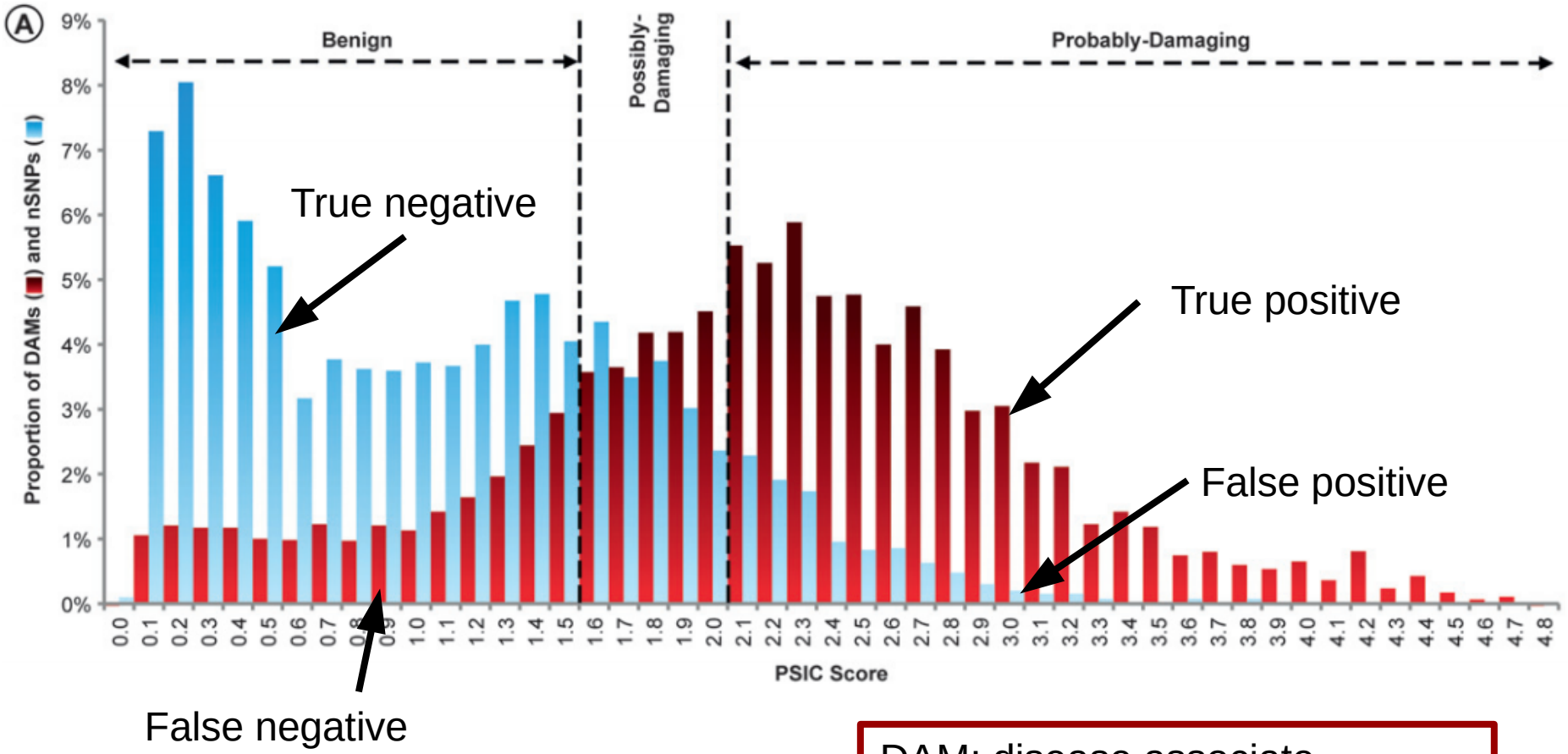
- ClinVar: medically relevant mutations with clinical presentations (583k)
- HGMD: human gene mutation database (62k)
- OMIM: Online mendelian inheritance in man

OMIM

- Total number of phenotypes for which the molecular basis is known: 6,161
- Total number of genes with phenotype-causing mutation: 3,880

PolyPhen PSIC Score

PSIC: Position-Specific Independent Counts



Exercise

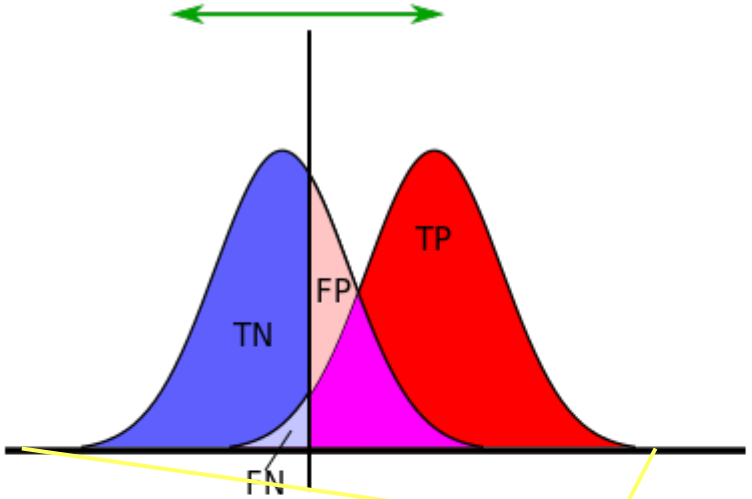
What cutoff would you use to eliminate false positives (maximize specificity)? ~4

What cutoff would you use to maximize sensitivity (maximize true positives)? 0

DAM: disease associate mutations (known positive)

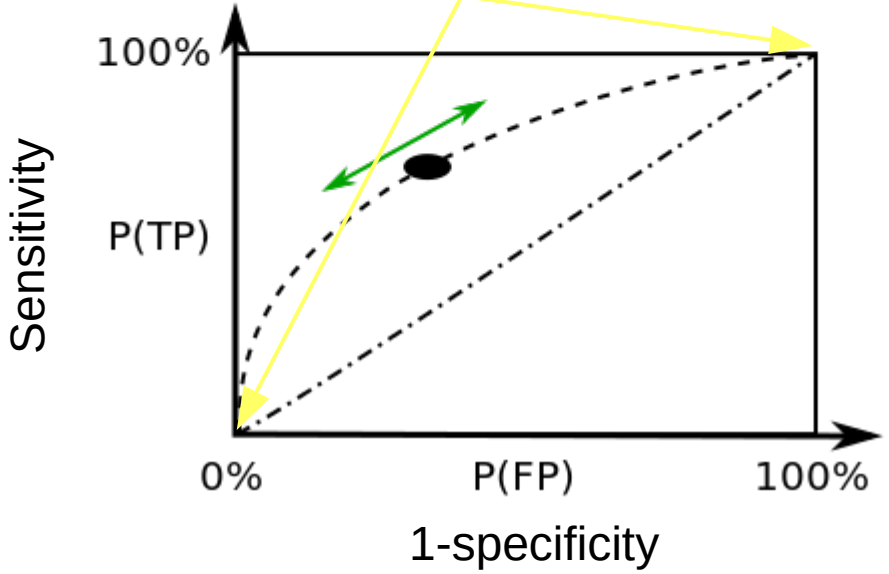
nSNPs: nonsynonymous single nucleotide polymorphisms (known = negative)

Cutoff influence the confusion matrix



TP	FP
FN	TN

Changing the cutoff leads to a trade-off between true positives (TP) and false positives (FP)



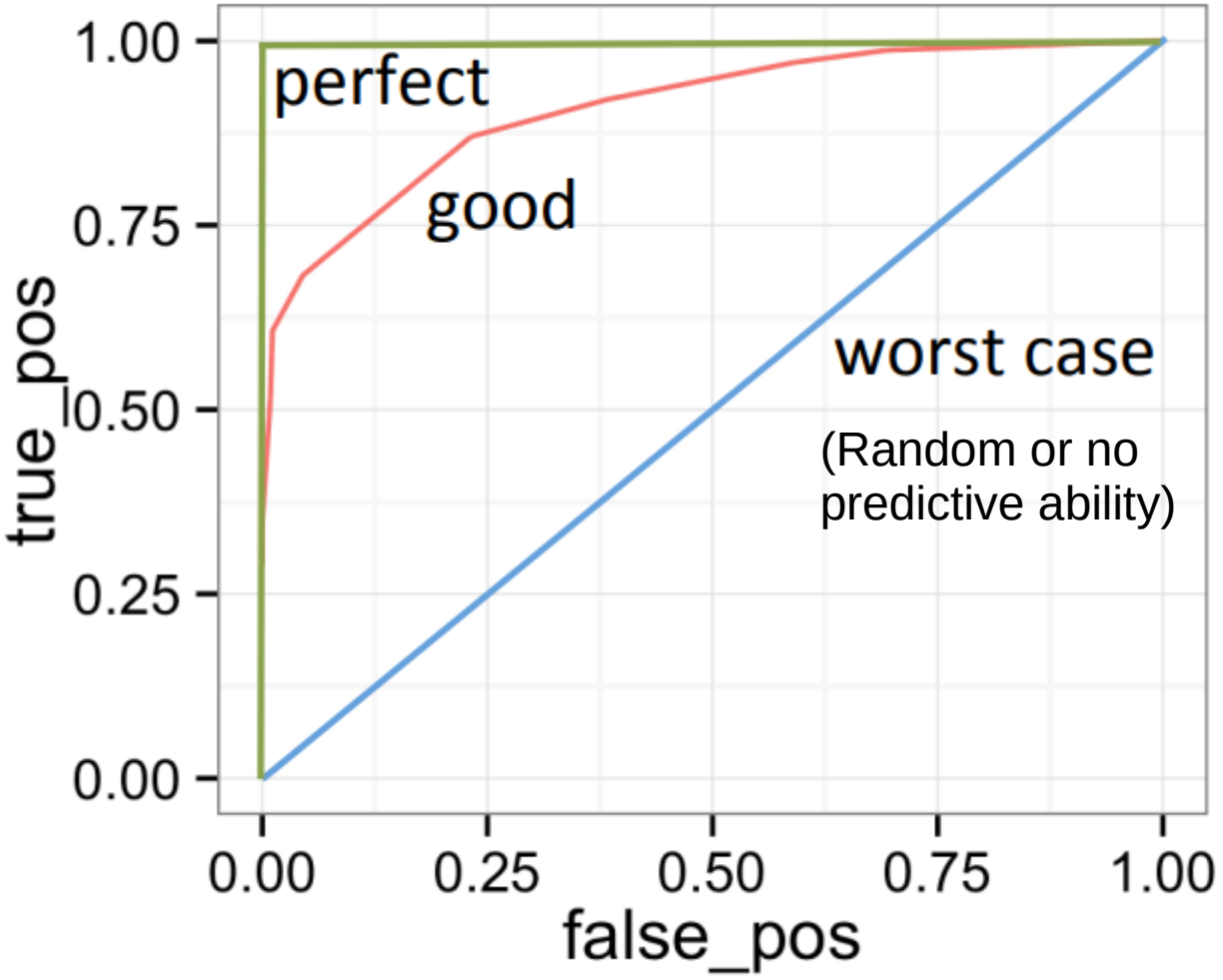
- False positives can be eliminated with a cutoff to the right, but a cost to TP
- False negative can be eliminated by a cutoff to the left, but a cost to FP
- All cutoffs form a curve

ROC Curves

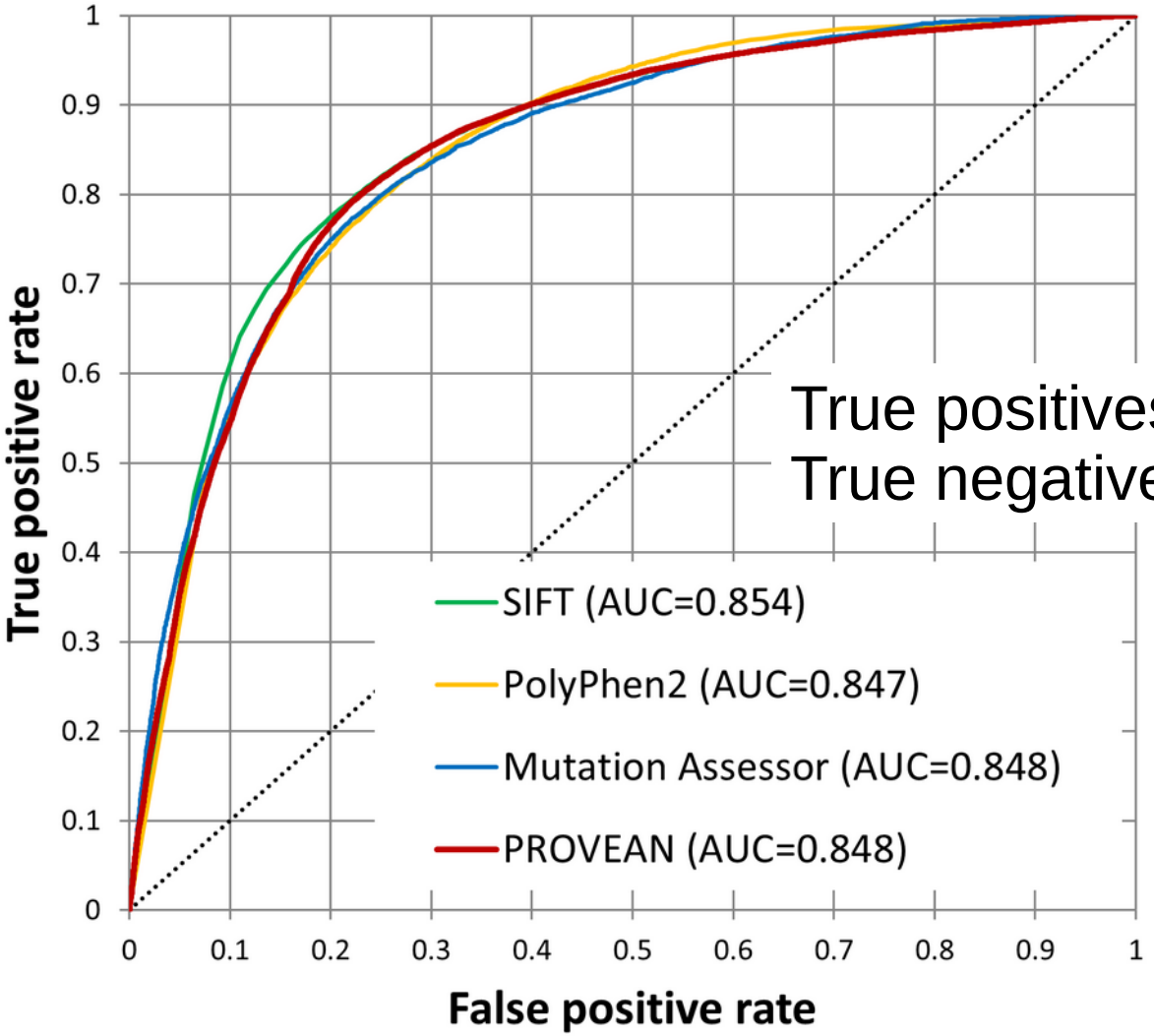
Reciever operator characteristic curve (ROC) analysis had its beginnings in observations made in Britain during World War II when radar receiver operators were being assessed on their ability to differentiate signal (e.g., enemy aircraft) from noise (e.g., flocks of birds).

AUC: Area Under the ROC

- AUC provides a fair comparison between methods across all cutoffs



ROC: Comparison of methods



True positives: HGMD
True negatives: common nSNPs

SIFT: Ng P, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863-874.

PolyPhen: Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov A et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591-597.

Deleterious Mutations in 'healthy' humans

Table 1. Summary of deleterious mutations found in three individuals and the reference genome

Genome	High-quality variants	Tested		Deleterious	
		Number	Heterozygotes (percent)[a]	Number[b]	Heterozygotes (percent)[a]
J. Craig Venter	7534	5645	52	796 (14%)	78
James D. Watson	7353	5417	49	816 (15%)	76
Han Chinese	7462	5707	58	837 (15%)	83
Reference	NA	10,689	NA	838 (8%)	NA

[a]The frequency of heterozygotes was derived from genotype calls in the original publications.
[b]The percentage of tested mutations that are deleterious is shown in parentheses.
NA, Not available.

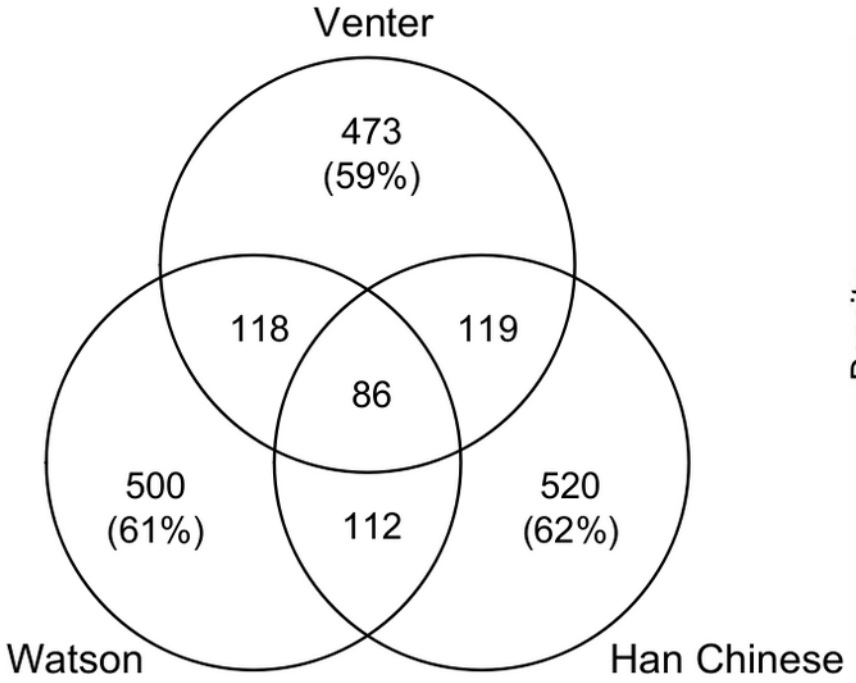
Exercise

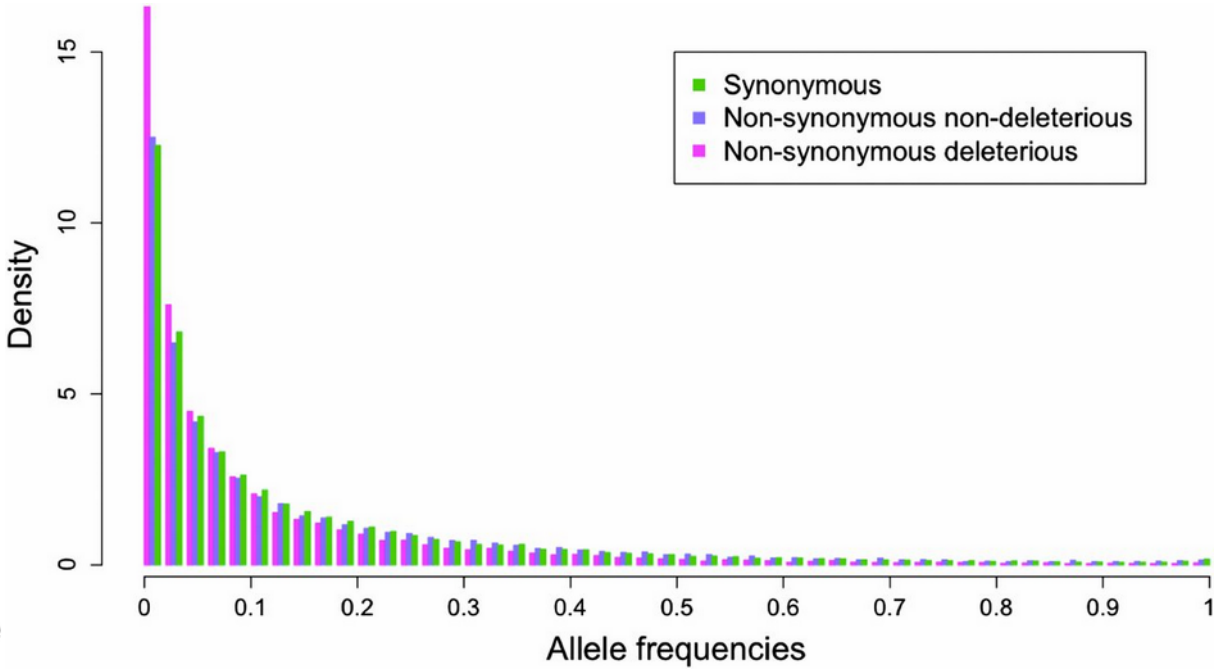
Why are deleterious mutations in 'healthy' humans?

- Deleterious alleles are recessive
- Deleterious alleles may be small effect (Ns << -1 are conserved)

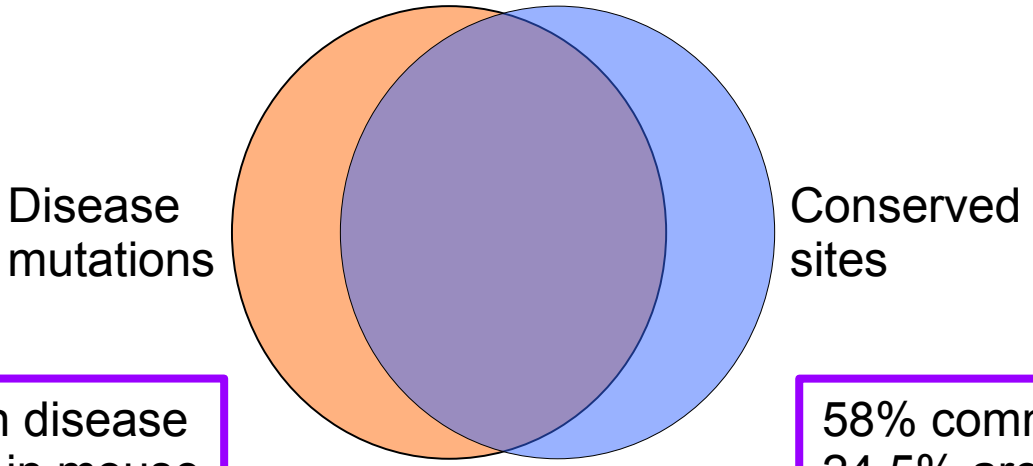
Chun and Fay (2009)

Most Deleterious SNPs are Rare (personal)





Conserved vs Disease Mutation

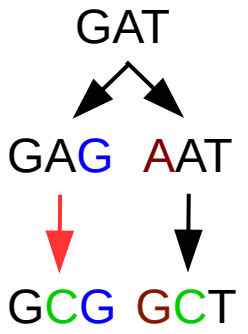


2.2% of human disease alleles are WT in mouse

58% common (>5%)
24.5% are >50%

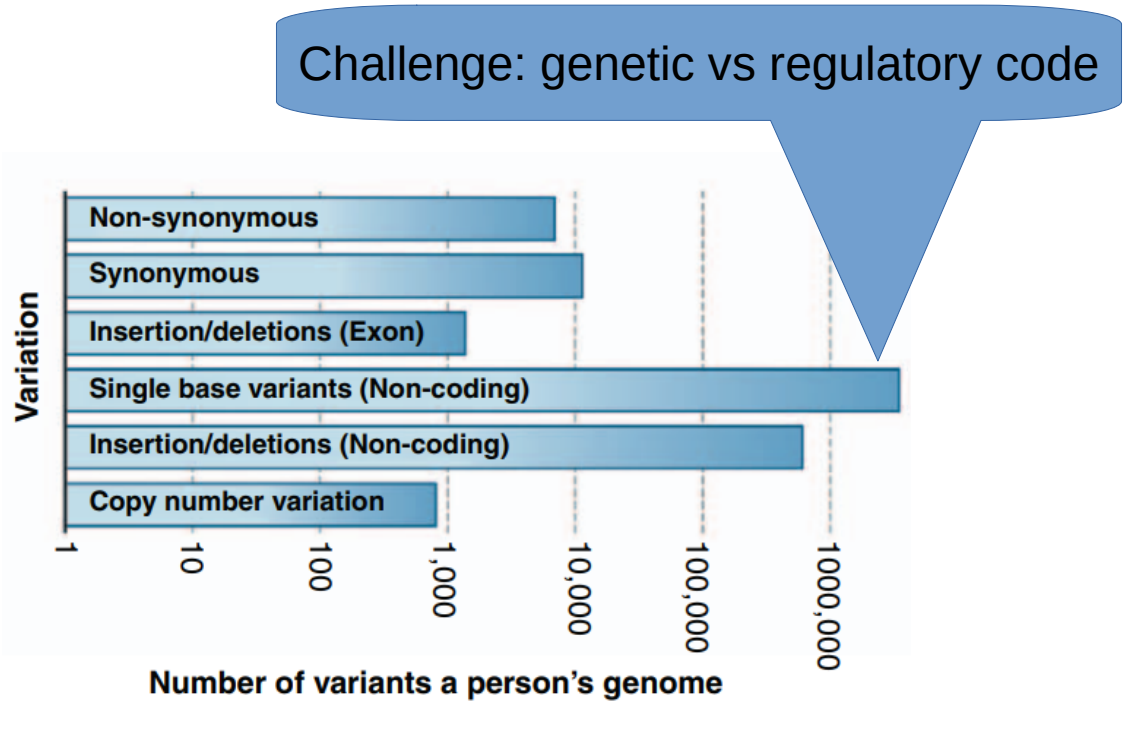
- Epistasis (background)
- Fitness differs across species (diet)
- Disease doesn't affect fitness (late-onset)

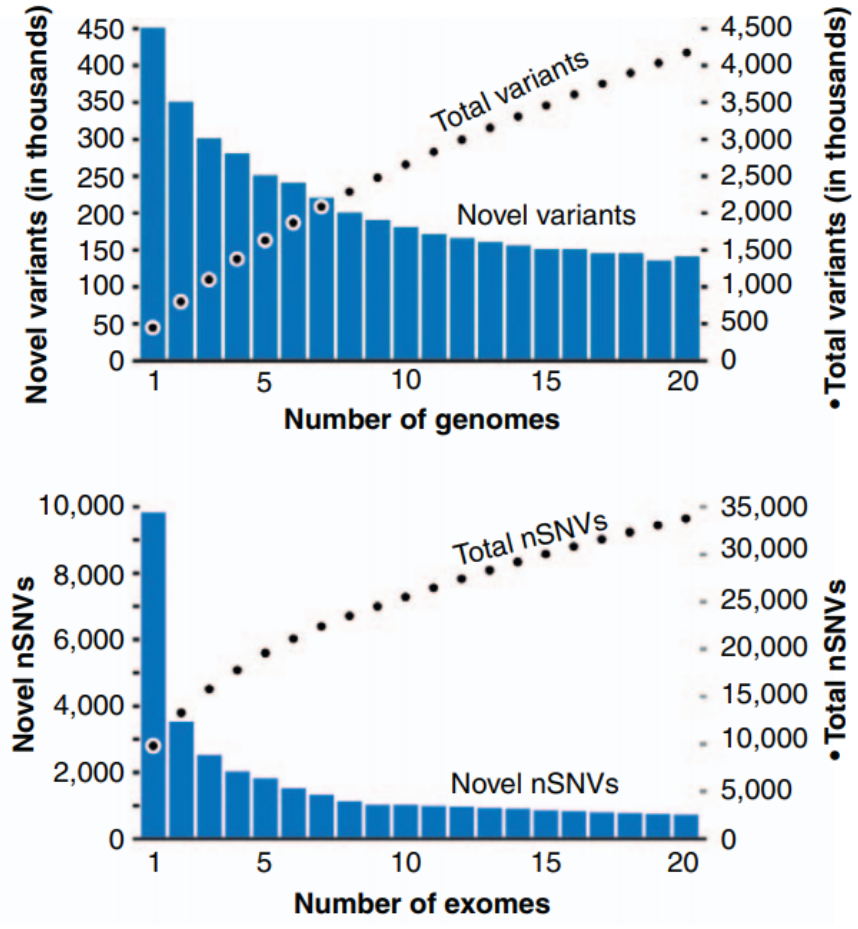
- Duplicate genes
- Ns is small ~10
- Affects fitness (deleterious) but not a disease, e.g. speed, disease resistance



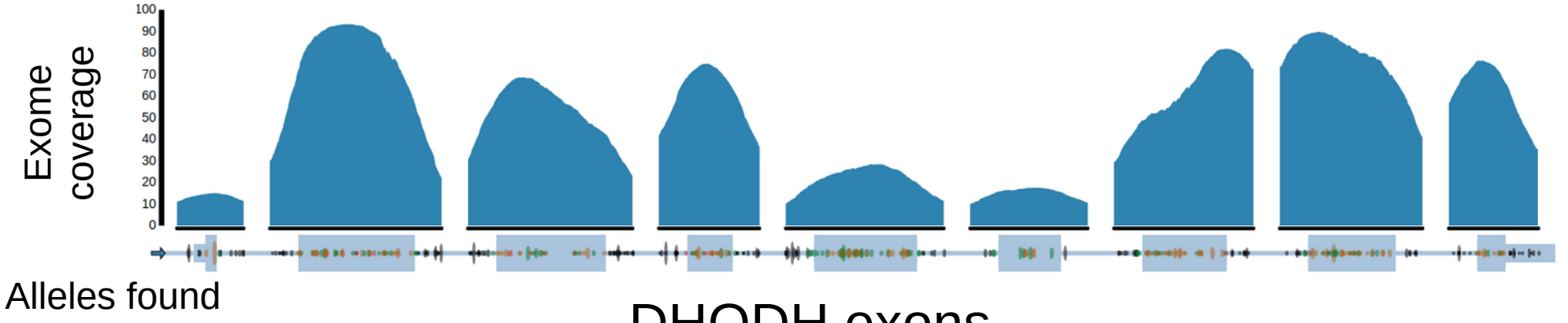
A to C has no effect in GAG, but is deleterious in AAT

Variants of unknown function: Challenge to Phylomedicine

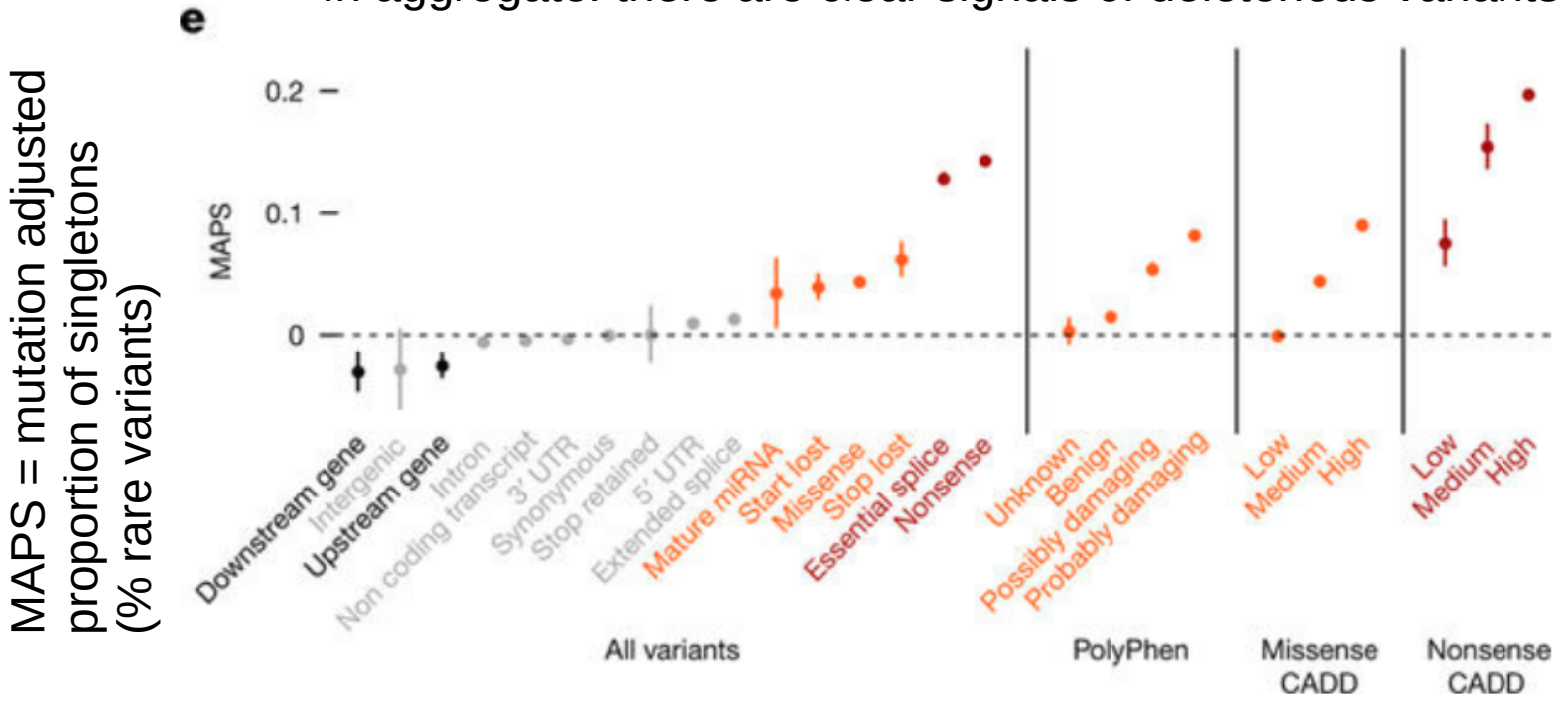




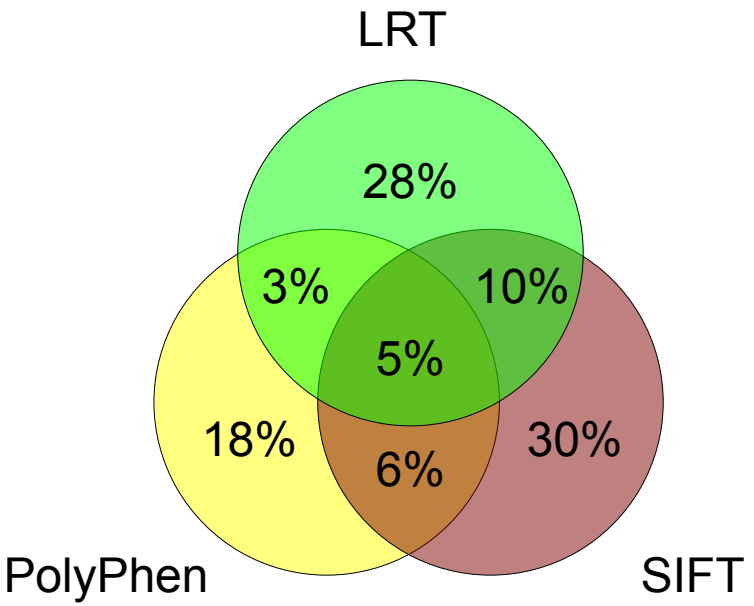
ExAC 60k exomes from humans
Exome Aggregation Consortium



In aggregate: there are clear signals of deleterious variants



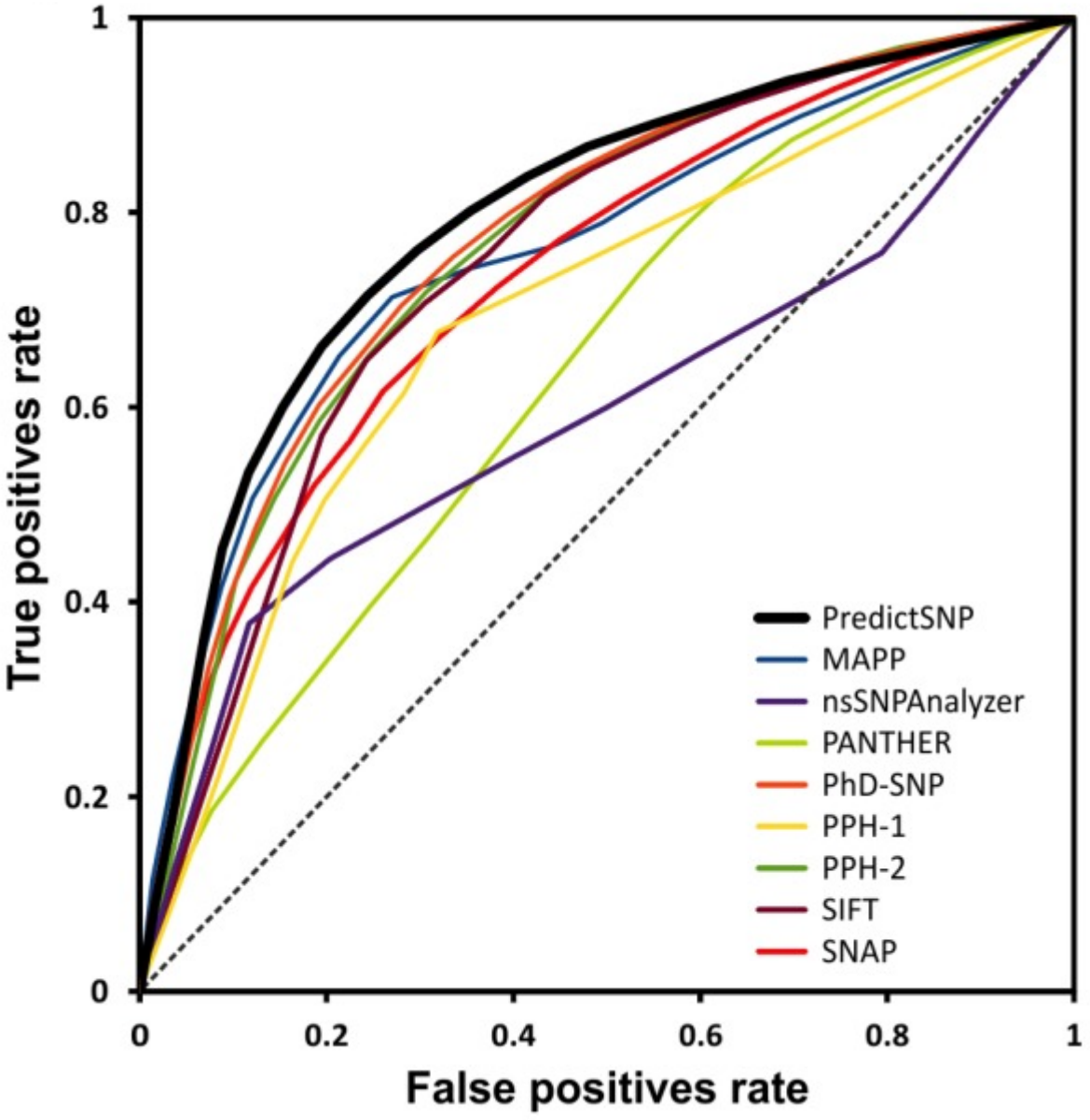
Problem: Low Overlap Among Three Prediction Methods



Solution:
Machine learning that uses consensus among multiple methods

7,534 NSN SNPs in Venter Genome
1,735 SNPs predicted deleterious by any one of the three methods

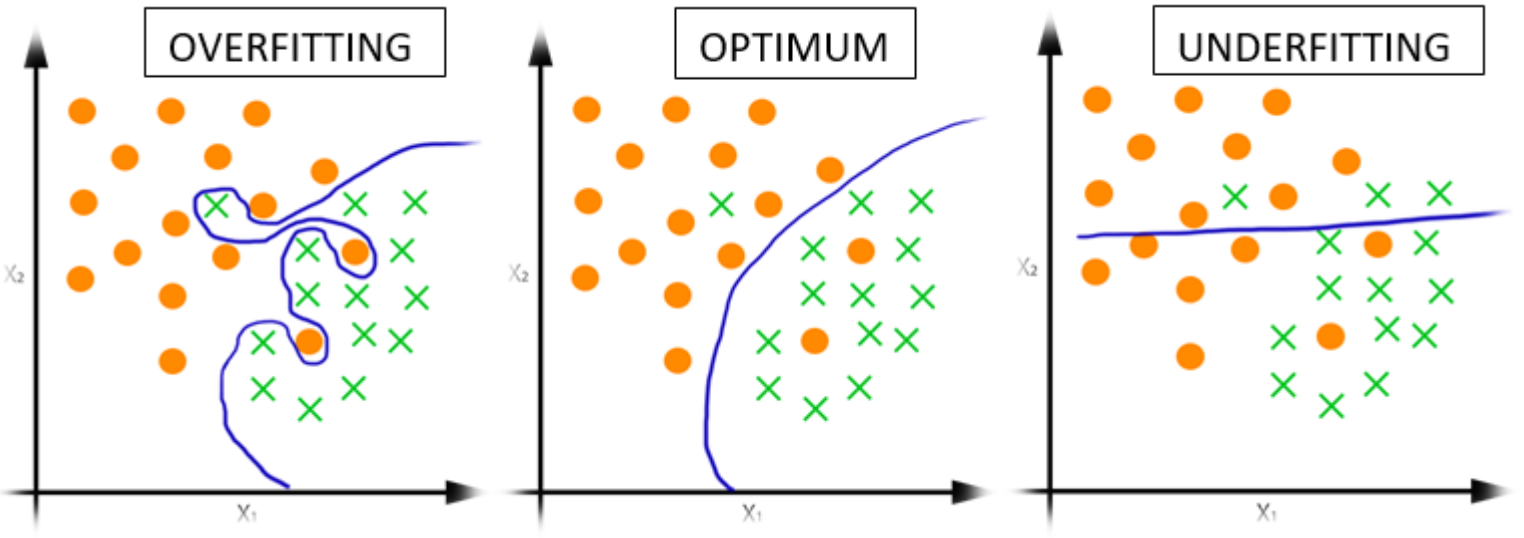
Consensus classifier PredictSNP



- Consensus classifiers use information (predictions) from multiple algorithms to get the best prediction possible.
- Overfitting is a significant concern
- Over-fitting occurs when there are many features use to predict outcomes and performance on new data is not as good as data used to train a model

Bendle et al. (2014)

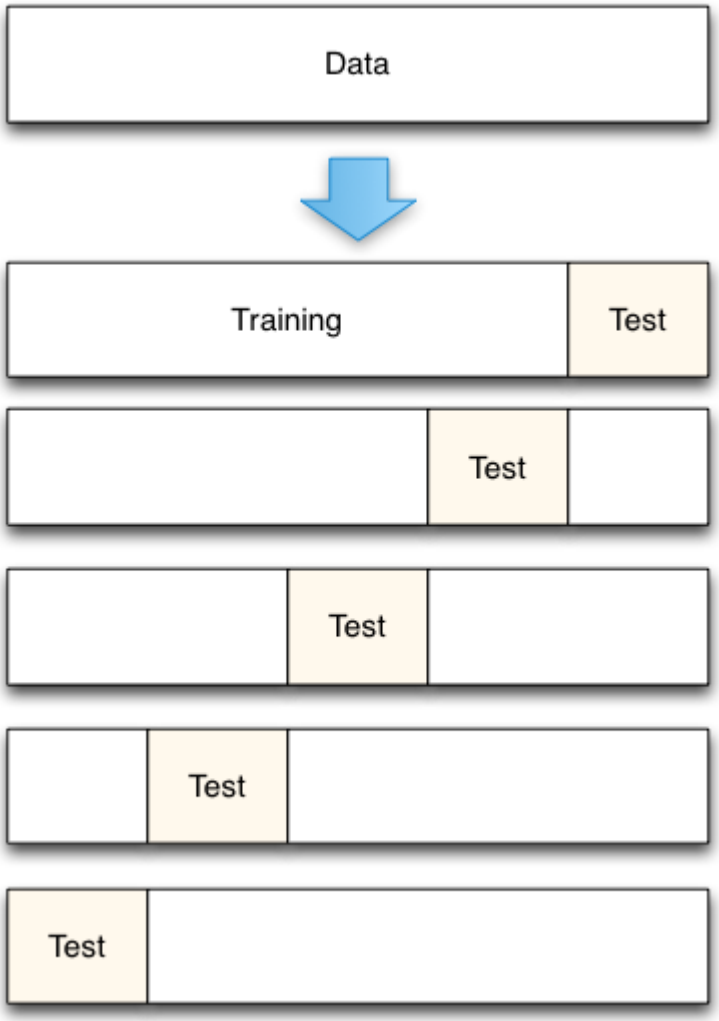
Over-fitting



overfitting is the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit future observations reliably

Cross validation

Cross validation is used to estimate how accurately a predictive model will perform in practice, using new data. It can identify problems due to overfitting or selection bias.



k-fold cross validation

- the original sample is randomly partitioned into k equal sized subsamples
- Of the k subsamples, a single subsample is retained as the validation data for testing the model
- remaining are used for training

Leave-one-out cross-validation

Cancer genomics

Cancer Biology

- Overproliferation of cells, either by loss of function of genes that inhibit cell division (tumor suppressor genes), or activation of genes that promote division (oncogenes).
- Genomes of tumors vs matched non-tumor controls
- Goal: Driver vs passenger mutations, Tumor evolution

Cancer treatment

- Many non-small cell lung cancers have activating mutations in epidermal growth factor receptor (EGFR) (2004)
- Targeted drugs (e.g. gefitinib) are available that counter the effect of some of those mutations.
- Personalized therapy (10% patients responsive)

Cancer genomics

Genomes of tumors vs matched non-tumor controls

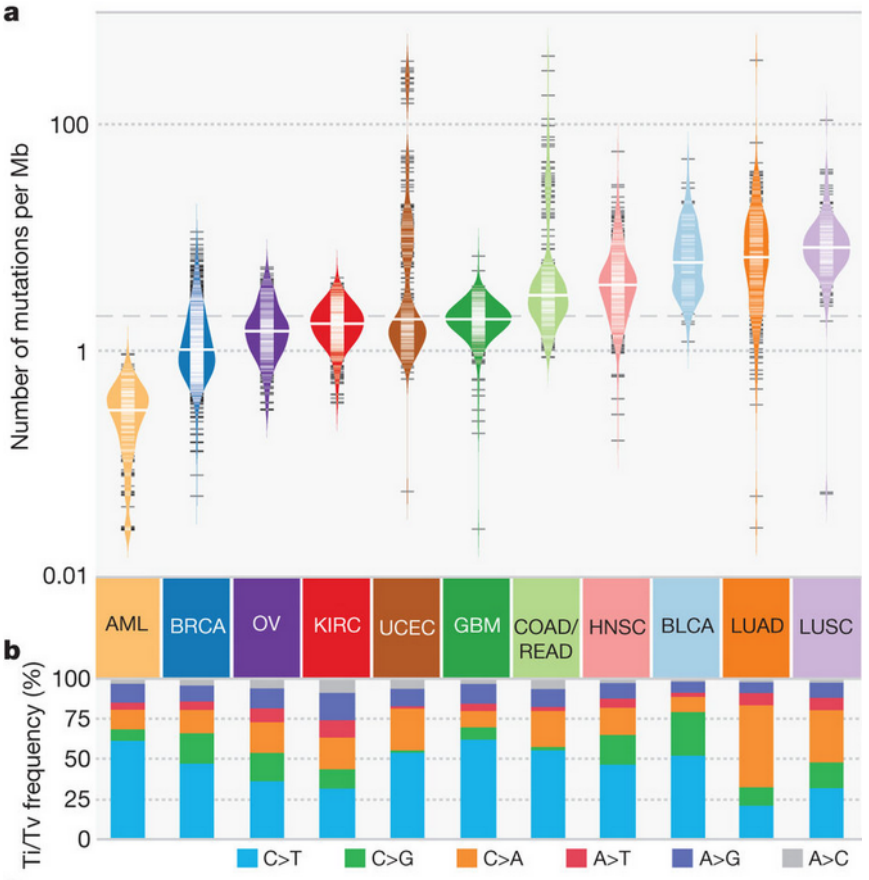
- Which mutation(s) are causal
- Probability of same gene (or codon) being hit (mutated) in different individuals with the same type of cancer
- Mutation rate

Need to account for

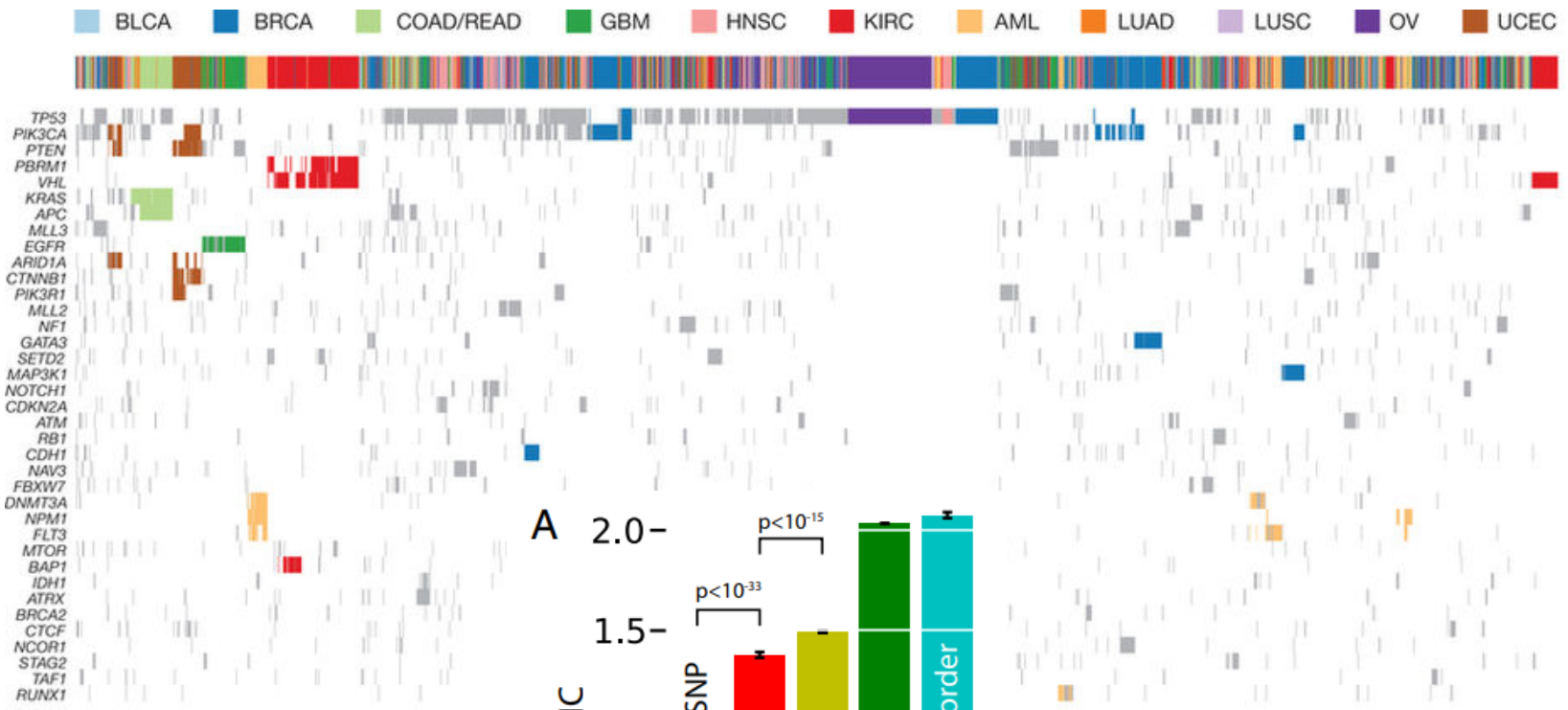
- Mutation rates/biases
- Different cancer types
- Conservation scores
- Chance (passenger mutations)

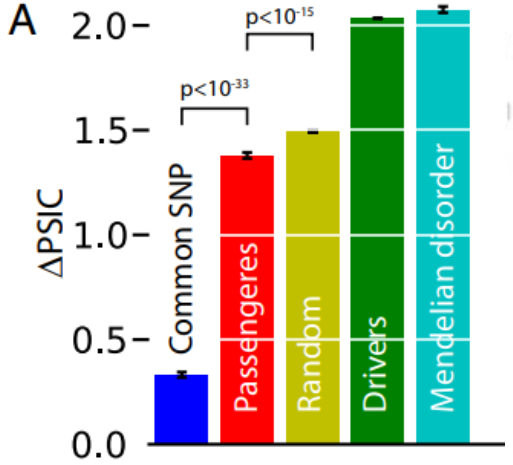
Cancer mutations ~ predictions

Mutation rate and type depend on cancer type



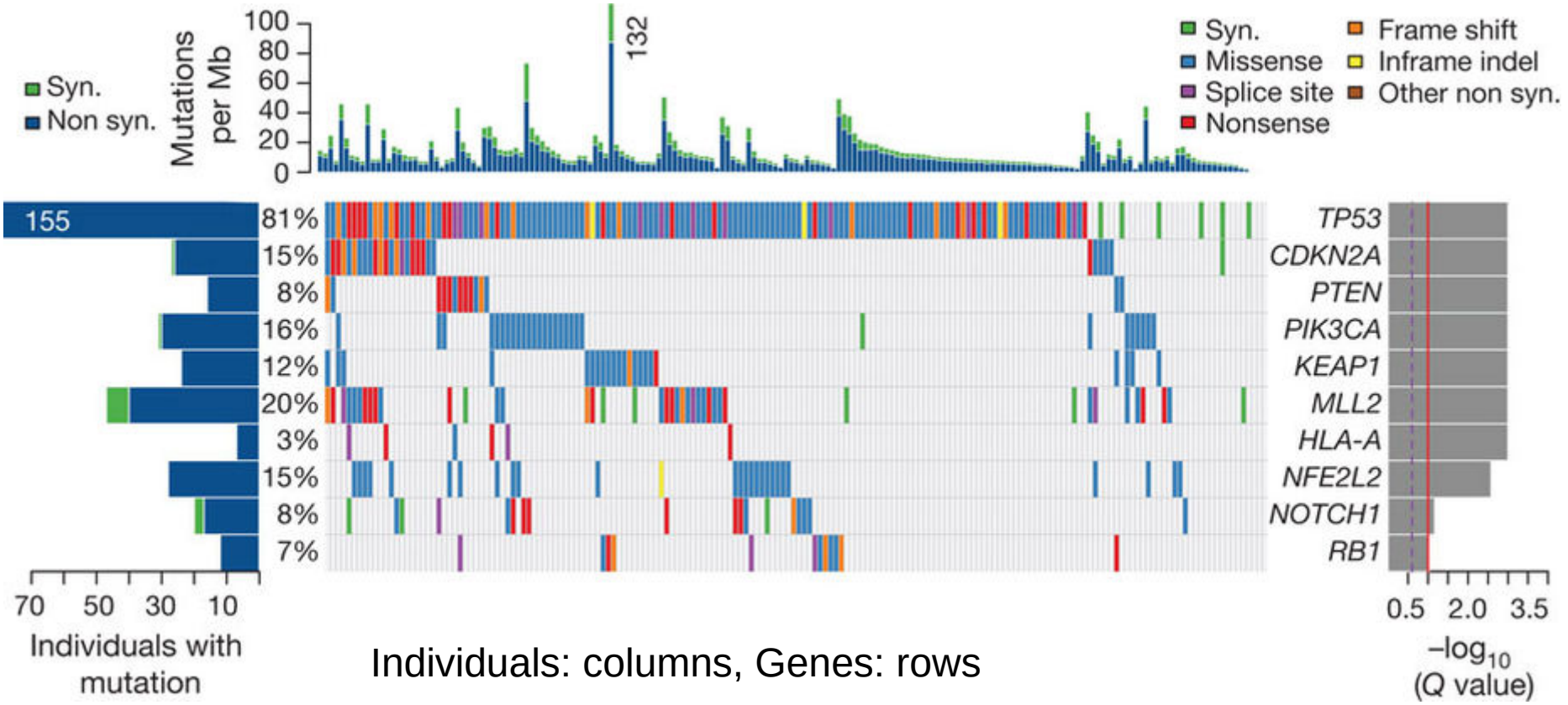
Most cancer mutations are of unknown significance





Drivers:
Parallel evolution
(Mutation adjusted)
e.g. TP53

Top mutated genes squamous cell lung cancers have elevated N/S



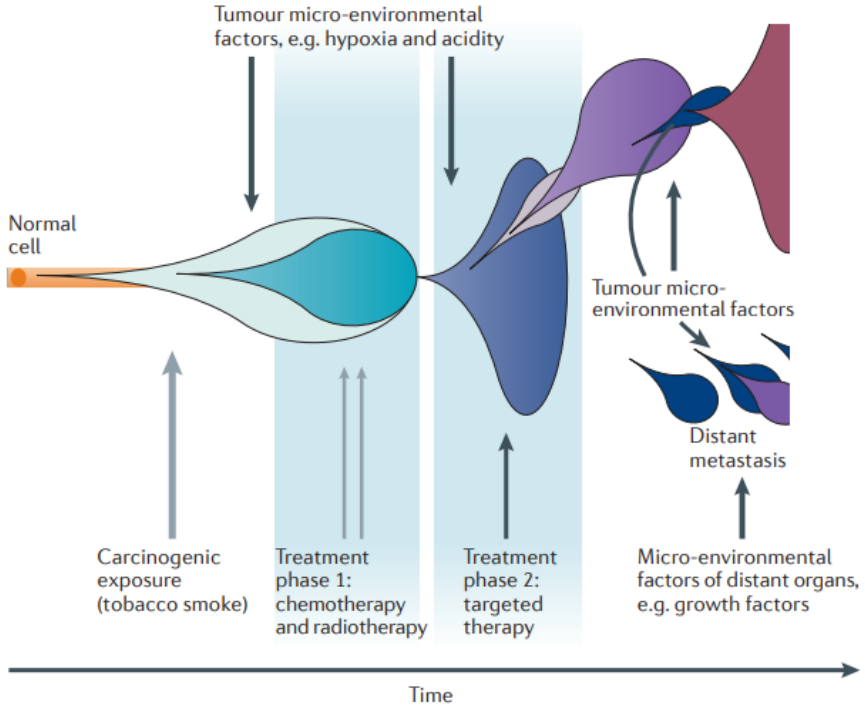
Individuals: columns, Genes: rows

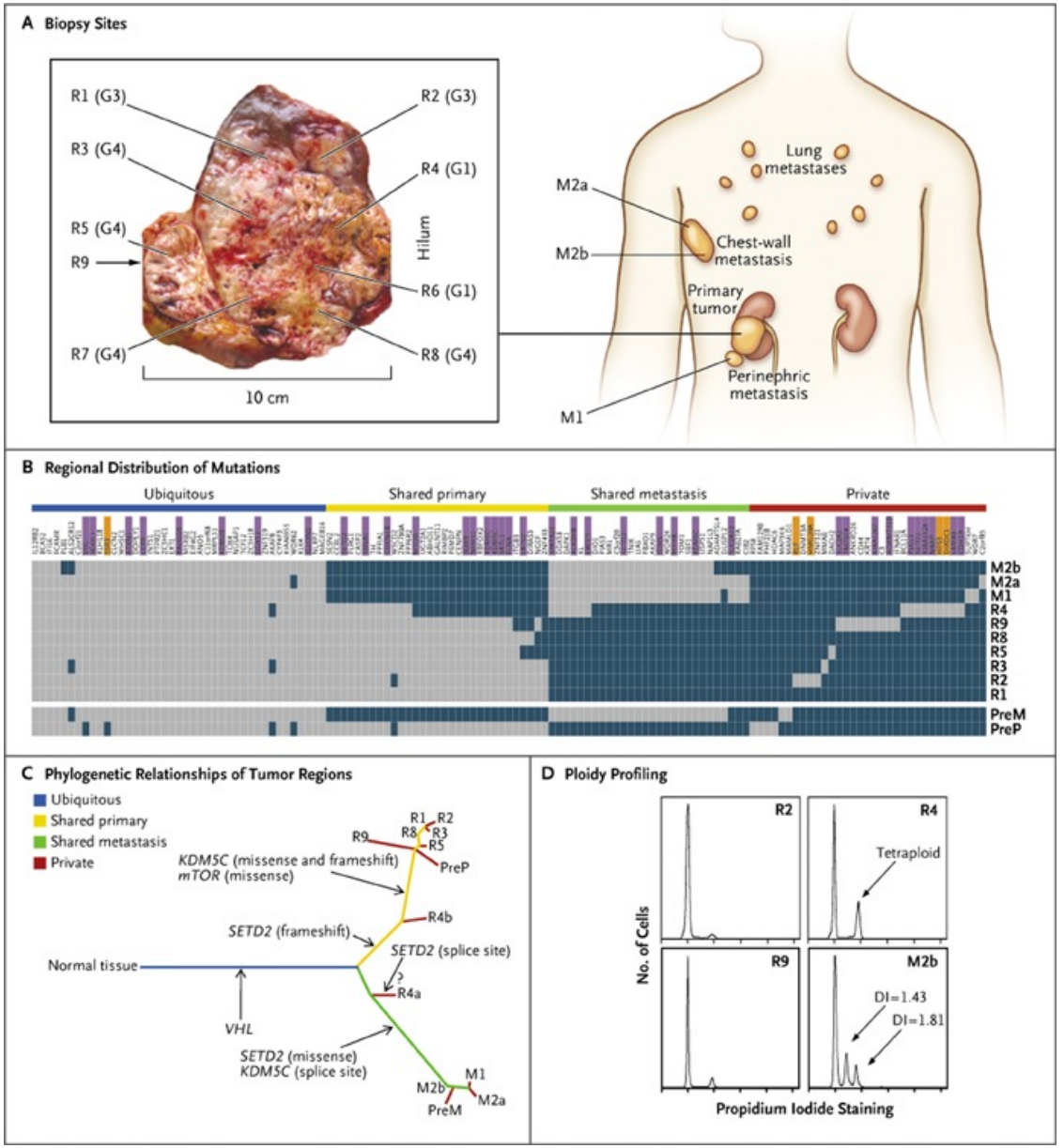
N/S >> 2

Q value (false discovery rate) from MutSig: obs vs exp mutation rate (CpG, genome position, etc)

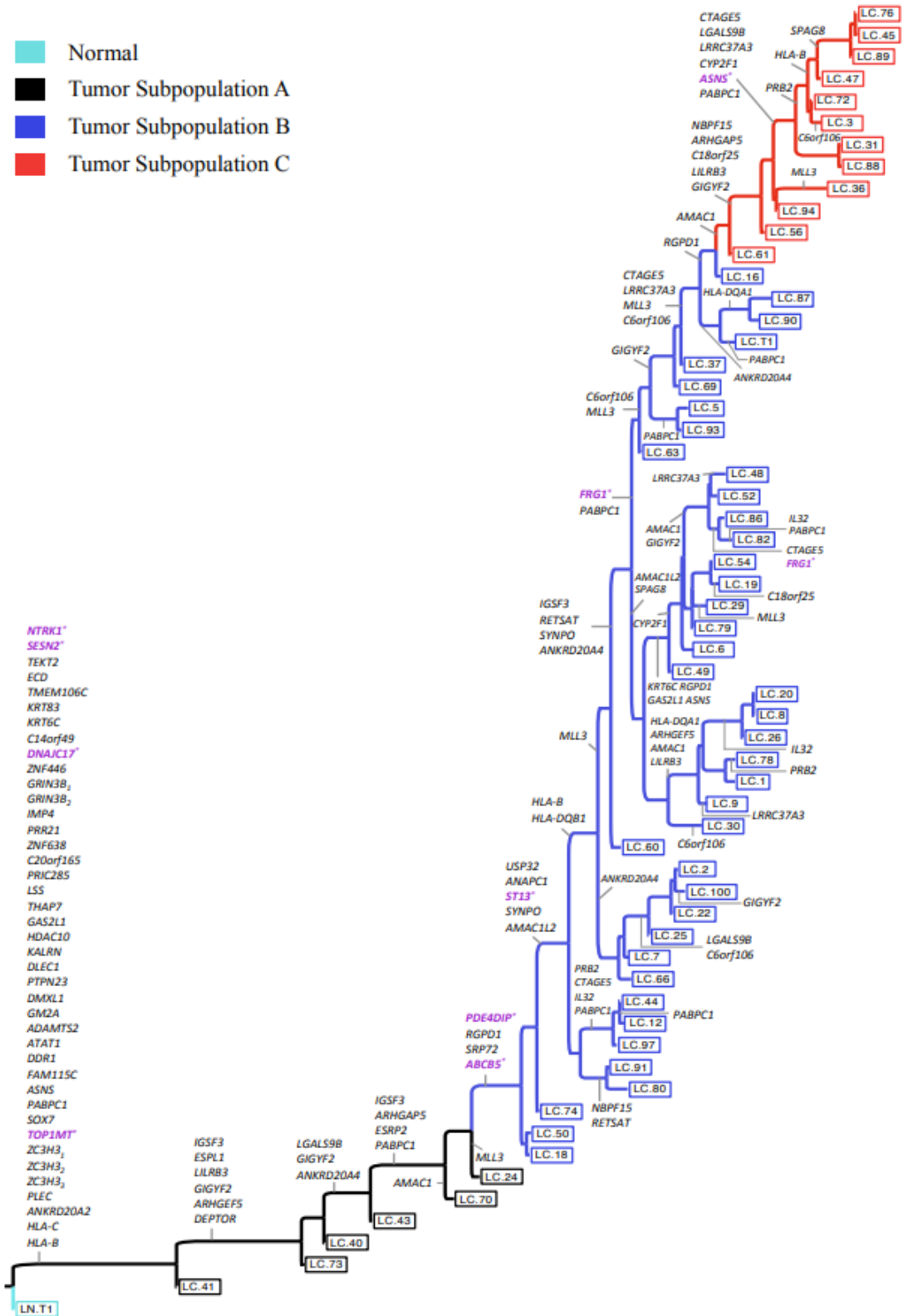
Nature volume489, pages519–525 (2012)

Tumor evolution





- Most somatic mutations are not shared
- Intra-tumor heterogeneity fosters tumor adaptation (convergent loss of function) and therapeutic failure through Darwinian evolution



Resolving tumor phylogeny through single-cell sequencing

Handling single cell errors:
D = missing data
G = true genotypes

$$Pr(D_{i,j}|G_{i,j}) = \begin{cases} 1-\alpha & \text{if } D_{i,j}=0, G_{i,j}=0 \\ \beta & \text{if } D_{i,j}=0, G_{i,j}=1 \\ \alpha & \text{if } D_{i,j}=1, G_{i,j}=0 \\ 1-\beta & \text{if } D_{i,j}=1, G_{i,j}=1 \end{cases}$$

Likelihood of tree (T), based on data (D), and error rates (Theta)

$$\mathscr{L}(\mathscr{T}, \boldsymbol{\theta}) = Pr(D|\mathscr{T}, \boldsymbol{\theta}) = \prod_{i=1}^{n} Pr(D_i|\mathscr{T}, \boldsymbol{\theta}),$$

Zafar et al. 2016

Exercise

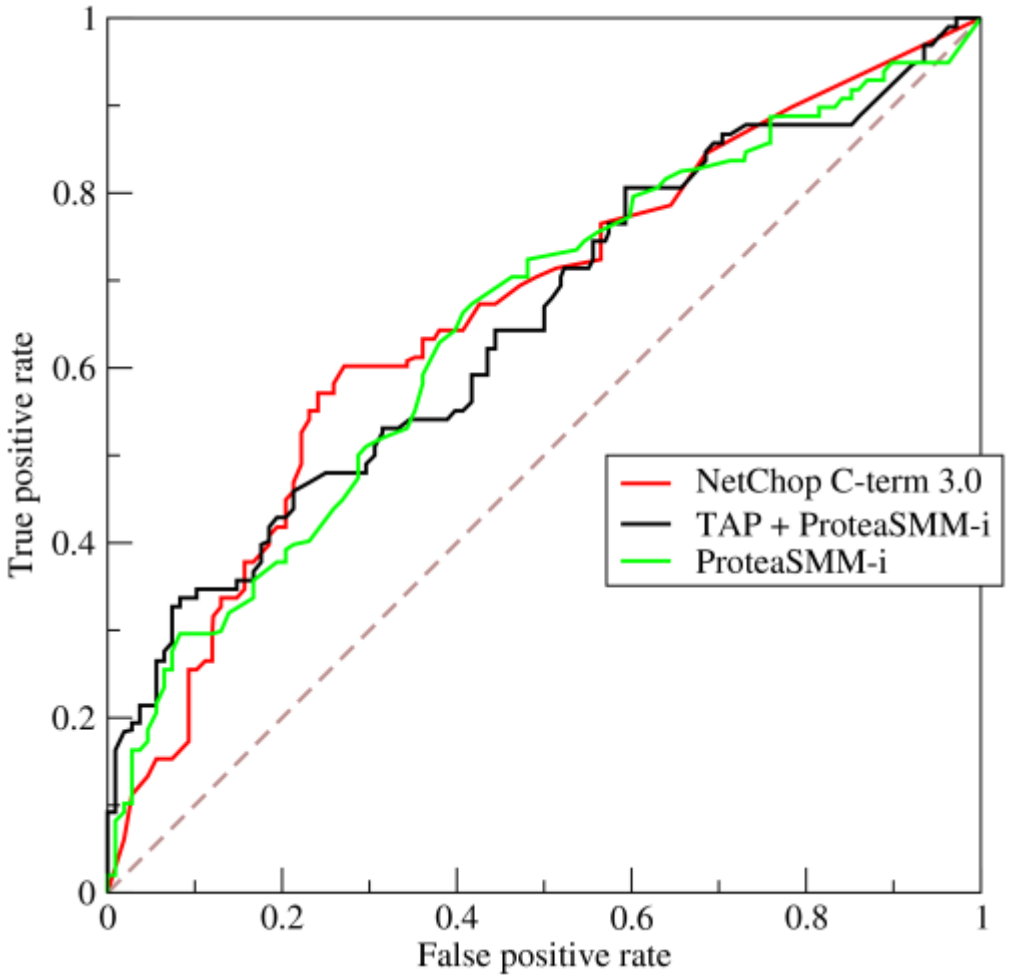
1. Below is a table showing the predictions of a method compared to known (true) outcomes.

What is the sensitivity, specificity, false discovery rate, and accuracy of the method?

	True positive	True negative
Predicted positive	33	10
Predicted negative	15	12

Exercises

2. Would the following methods be better or worse than the three methods on the left?



Method 1:
TP = 66 FN = 28
FP = 60 TN = 33

Method 2:
TP = 88 FN = 120
FP = 20 TN = 120

TP = true positives
FN = false negatives
FP = false positives
TN = true negatives

Exercises

1) In binary prediction methods (e.g. disease mutations), changing the cutoff to increase # TP causes what effect (increase/decrease) on specificity, TPR, FPR?
2) What can explain: a) disease mutations that occur in unconserved sites, b) conservation at sites that don't cause disease.
3) Predicting disease mutations relies on conservation across species (T/F) on the type of amino acid change (T/F)?
4) Disease mutations can be predicted at sites that are not identical across species (T/F)?

Exercises

5) How are PSIC/PSSM scores related to PAM scores and how are they different?

6) What does the mutation adjusted proportion of singletons tell you about the likelihood a site is function?

7) Which are relevant to identifying driver mutations in cancer? Mutational bias (types), nonsynonymous vs synonymous rates, parallel/convergent changes

8) How do you prevent over-fitting when developing a model to predict disease mutations?

9) A ROC curve is more informative about performance than a confusion matrix (T/F)?

Exercise

1. Below is a table showing the predictions of a method compared to known (true) outcomes.

What is the sensitivity, specificity, false discovery rate, and accuracy of the method?

	True positive	True negative
Predicted positive	33	10
Predicted negative	15	12

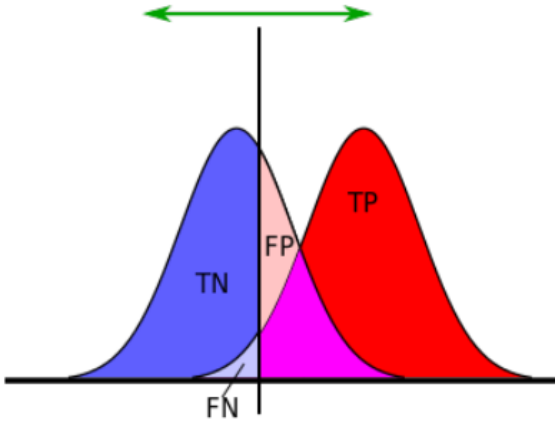
True positive rate = TP / (TP+FN) = 33/(33+15) = 0.69

True negative rate = TN / (TN + FP) = 12/(12+10) = 0.55

False discovery rate = FP / (FP + TP) = 10/(10+33) = 0.23

Accuracy = TP + TN / Total = (33+12)/70 = 0.64

Exercises



1) In binary prediction methods (e.g. disease mutations), changing the cutoff to increase #TP causes what effect (increase/decrease) on specificity, TPR, FPR? specificity (down), TPR (up), FPR (up)

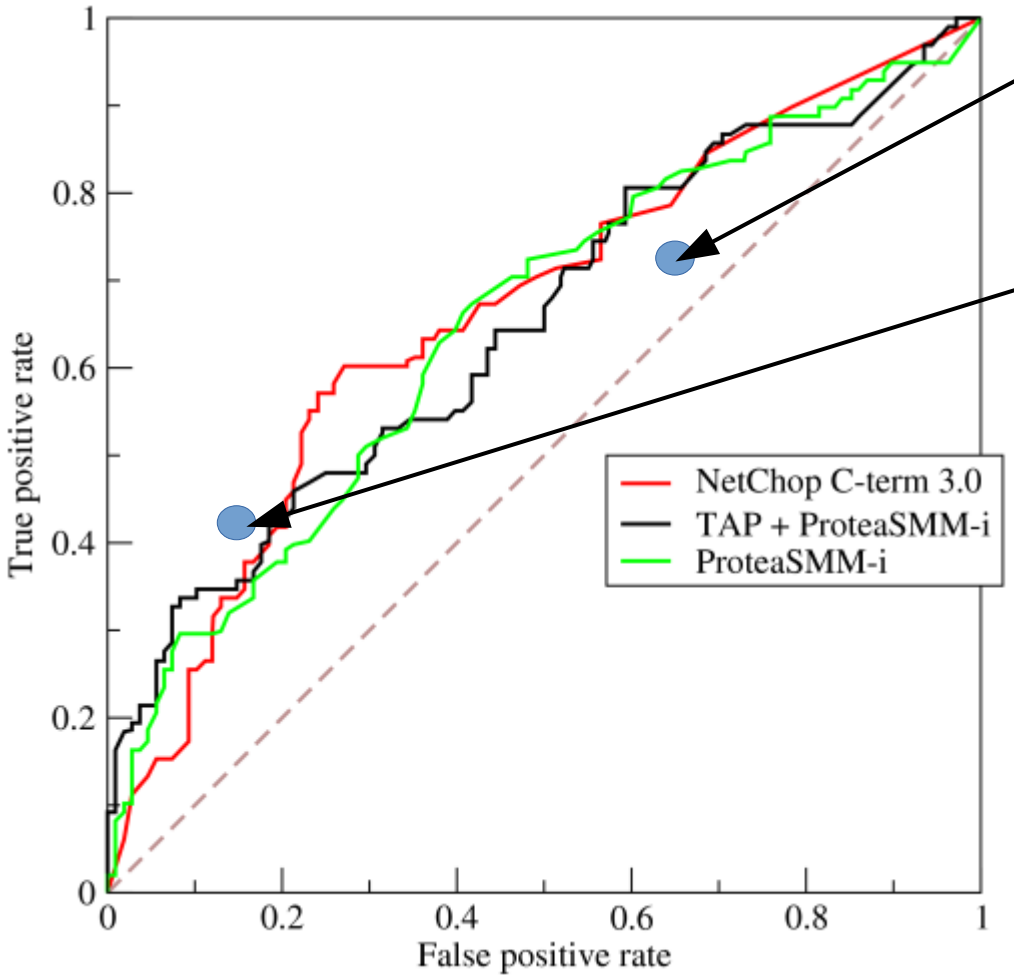
2) What can explain: a) disease mutations that occur in unconserved sites, b) conservation at sites that don't cause disease. a) epistasis, fitness difference btw species, disease doesn't affect fitness, b) duplicated genes, Ns is small, long term fitness

3) Predicting disease mutations relies on conservation across species (T/F) on the type of amino acid change (T/F)?

4) Disease mutations can be predicted at sites that are not identical across species (T/F)?

Exercises

2. Would the following methods be better or worse than the three methods on the left?



Method 1:
TP = 66 FN = 28 TPR = 66/94 = 0.70
FP = 60 TN = 33 FPR = 60/93 = 0.65

Method 2:
TP = 88 FN = 120 TPR = 88/208 = 0.42
FP = 20 TN = 120 FPR = 20/140 = 0.14

TP = true positives
FN = false negatives
FP = false positives
TN = true negatives

Exercises

5) How are PSIC/PSSM scores related to PAM scores and how are they different? PAM is not gene or position specific, both are based on observed relative frequency of different types of amino acid changes. PSIC/PSSM examine one position in alignment, PAM examines one type of substitution across many alignments/positions.

6) What does the mutation adjusted proportion of singletons tell you about the likelihood a site is function? Important sites should be more rare, ie more singletons

7) Which are relevant to identifying driver mutations in cancer? Mutational bias (types), nonsynonymous vs synonymous rates, parallel/convergent changes

8) How do you prevent over-fitting when developing a model to predict disease mutations? Use cross-validation

9) A ROC curve is more informative about performance than a confusion matrix (T/F)?