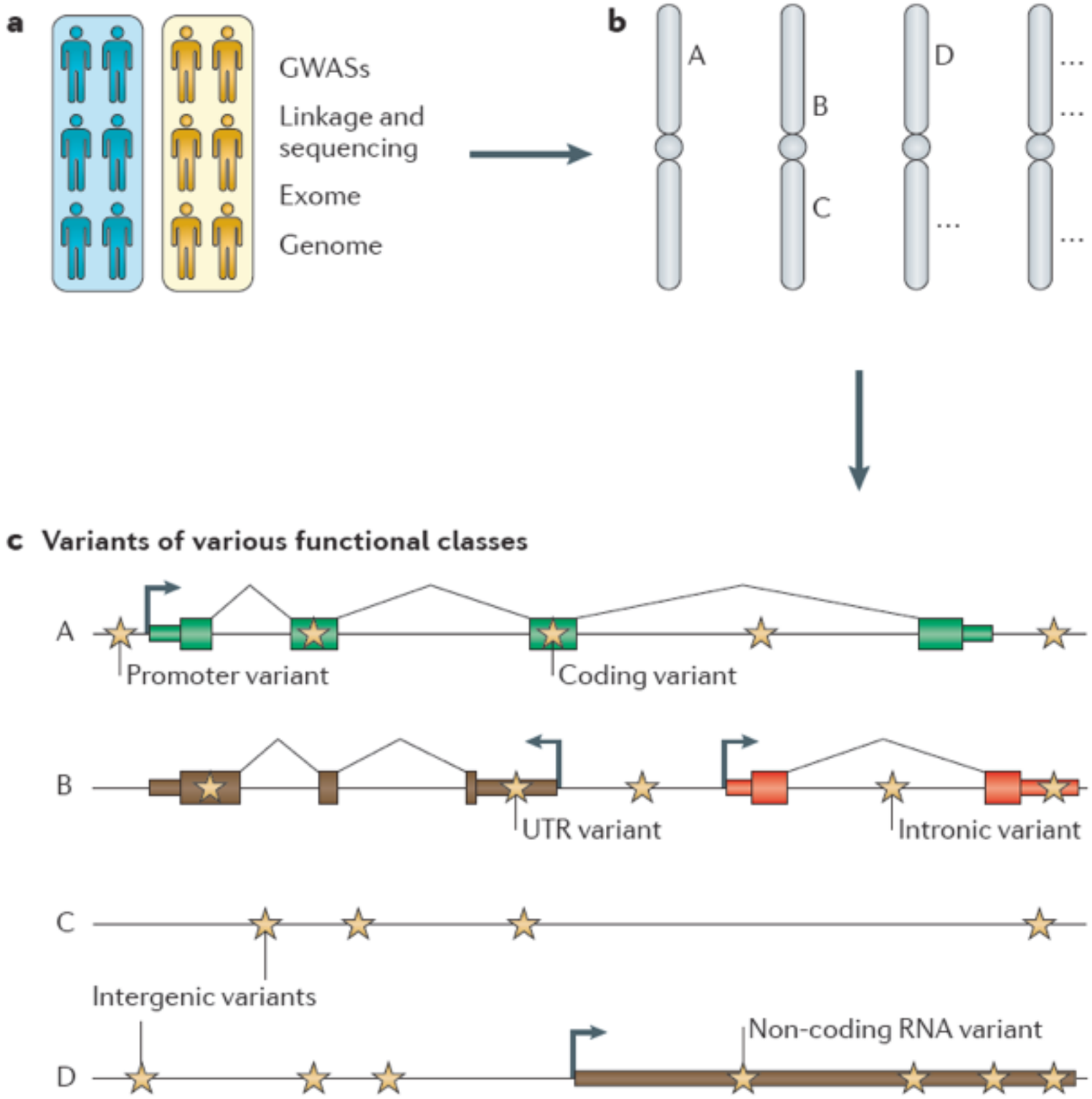
Identification of Causal Variants in Individual Human Genomes

Greg Cooper
HudsonAlpha Institute for Biotechnology
UAB Genomics Immersion Course
October 2015

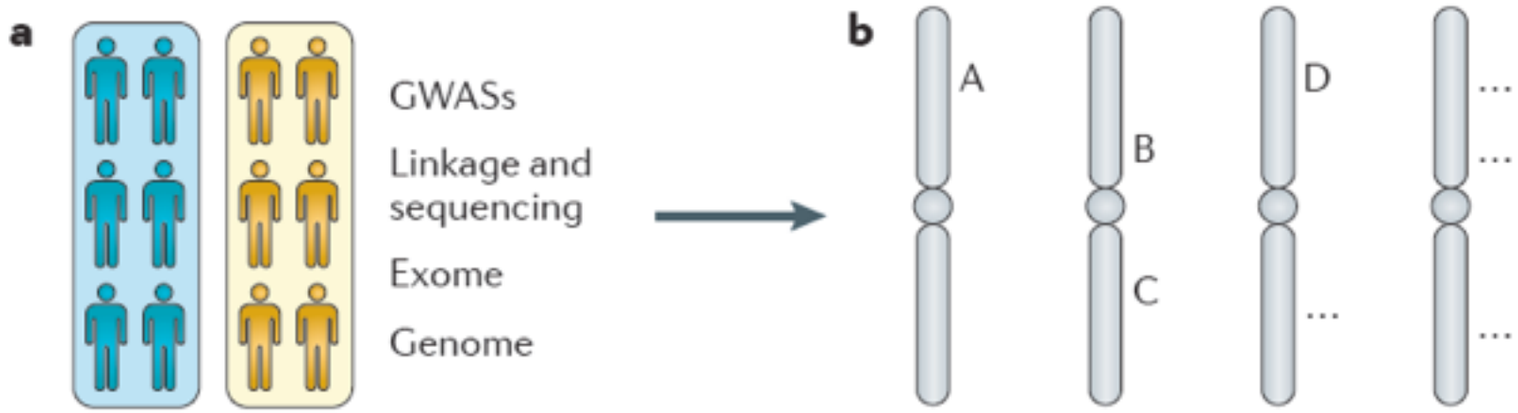
Human Genetics



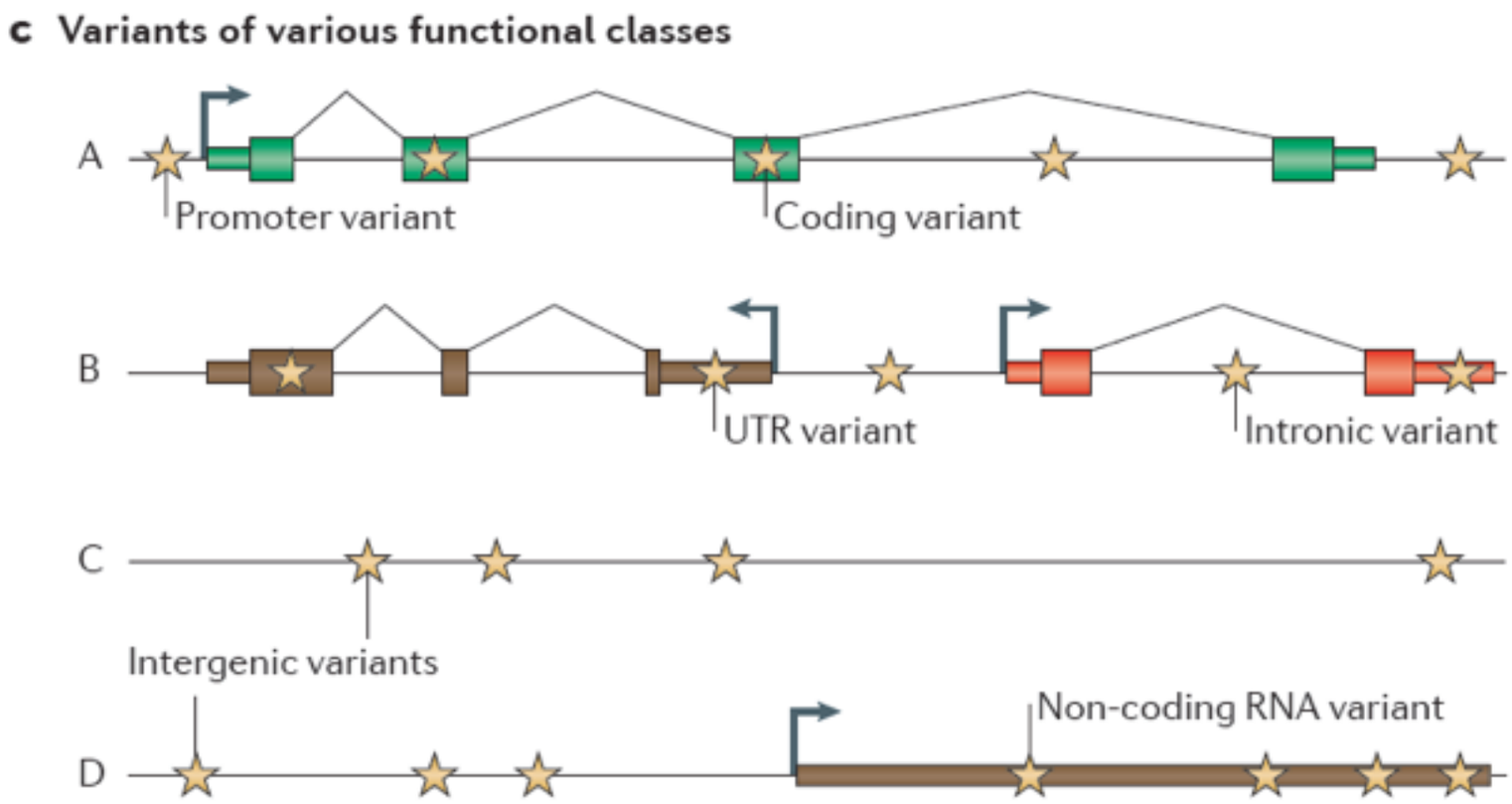
Greg Cooper, UAB Immersion Course, Oct 2015

Cooper and Shendure *NRG* 2011

Human Genetics



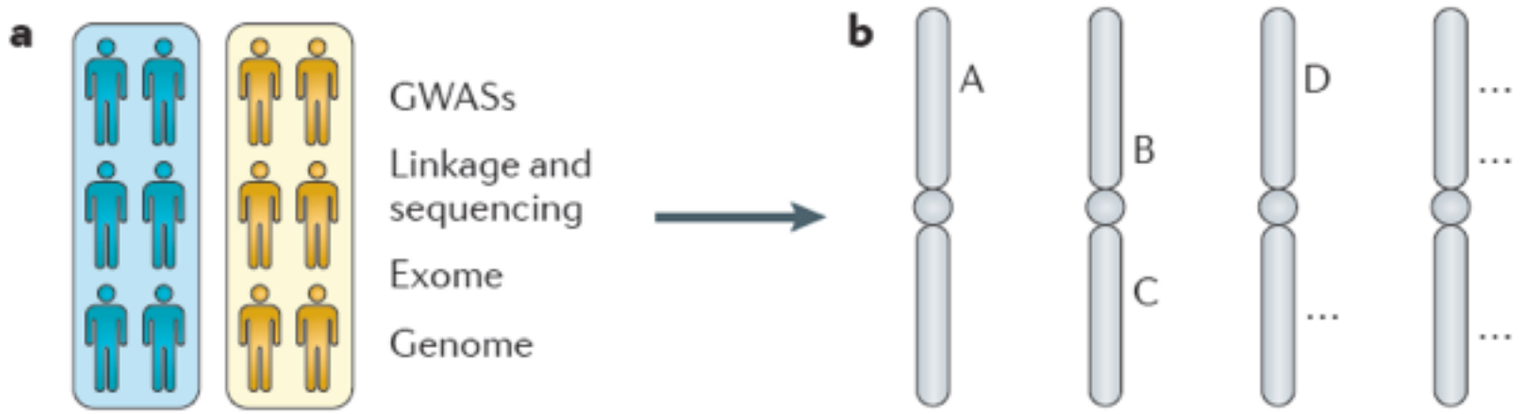
Statistical Power



Greg Cooper, UAB Immersion Course, Oct 2015

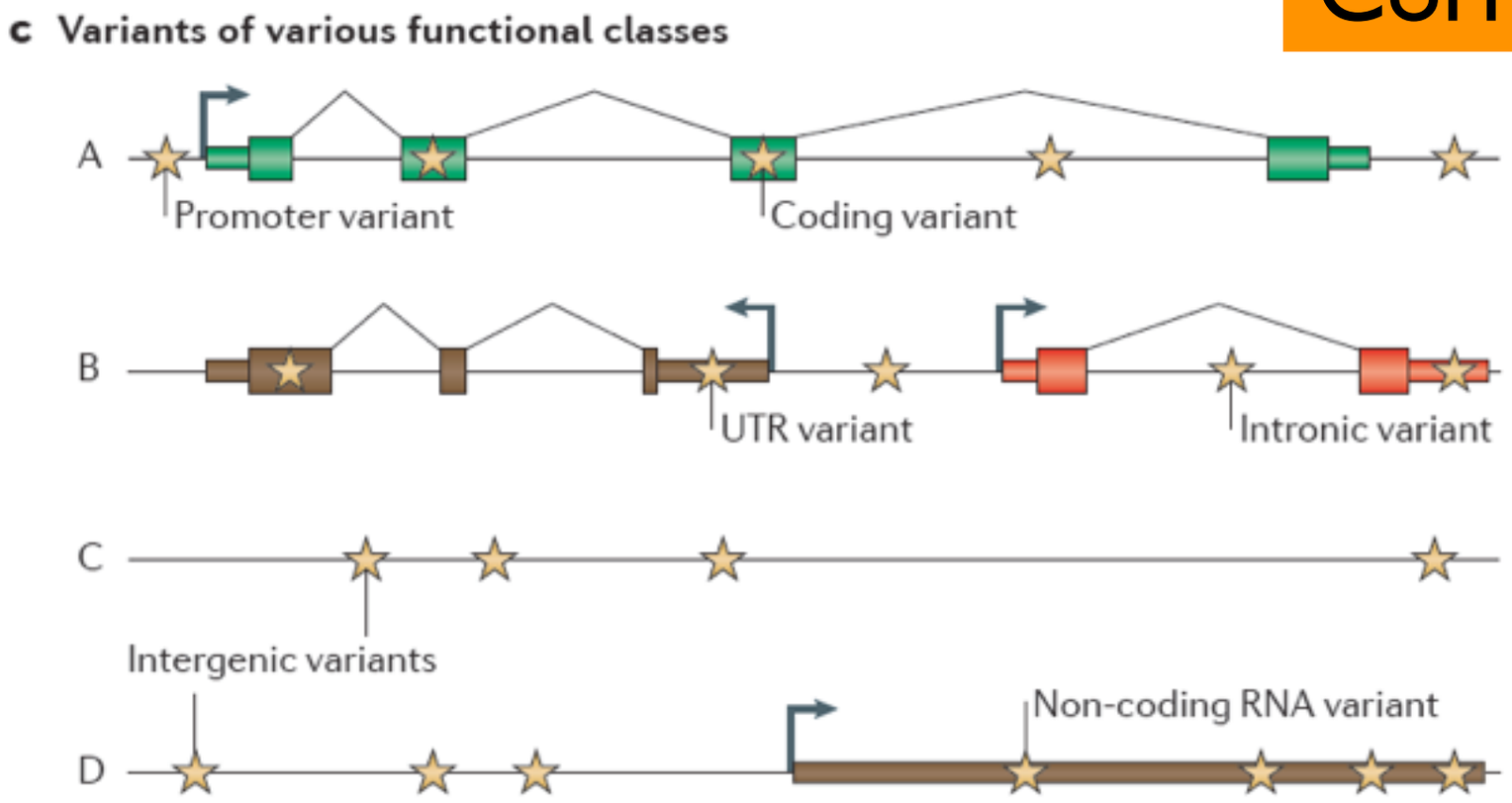
Cooper and Shendure *NRG* 2011

Human Genetics



Statistical Power

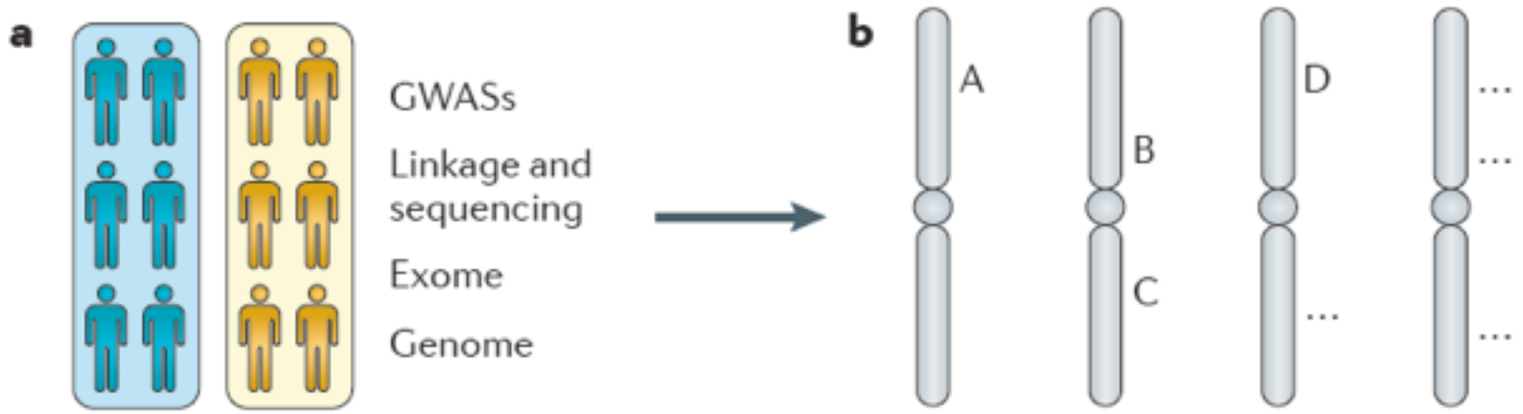
Haplotypic Correlations



Greg Cooper, UAB Immersion Course, Oct 2015

Cooper and Shendure *NRG* 2011

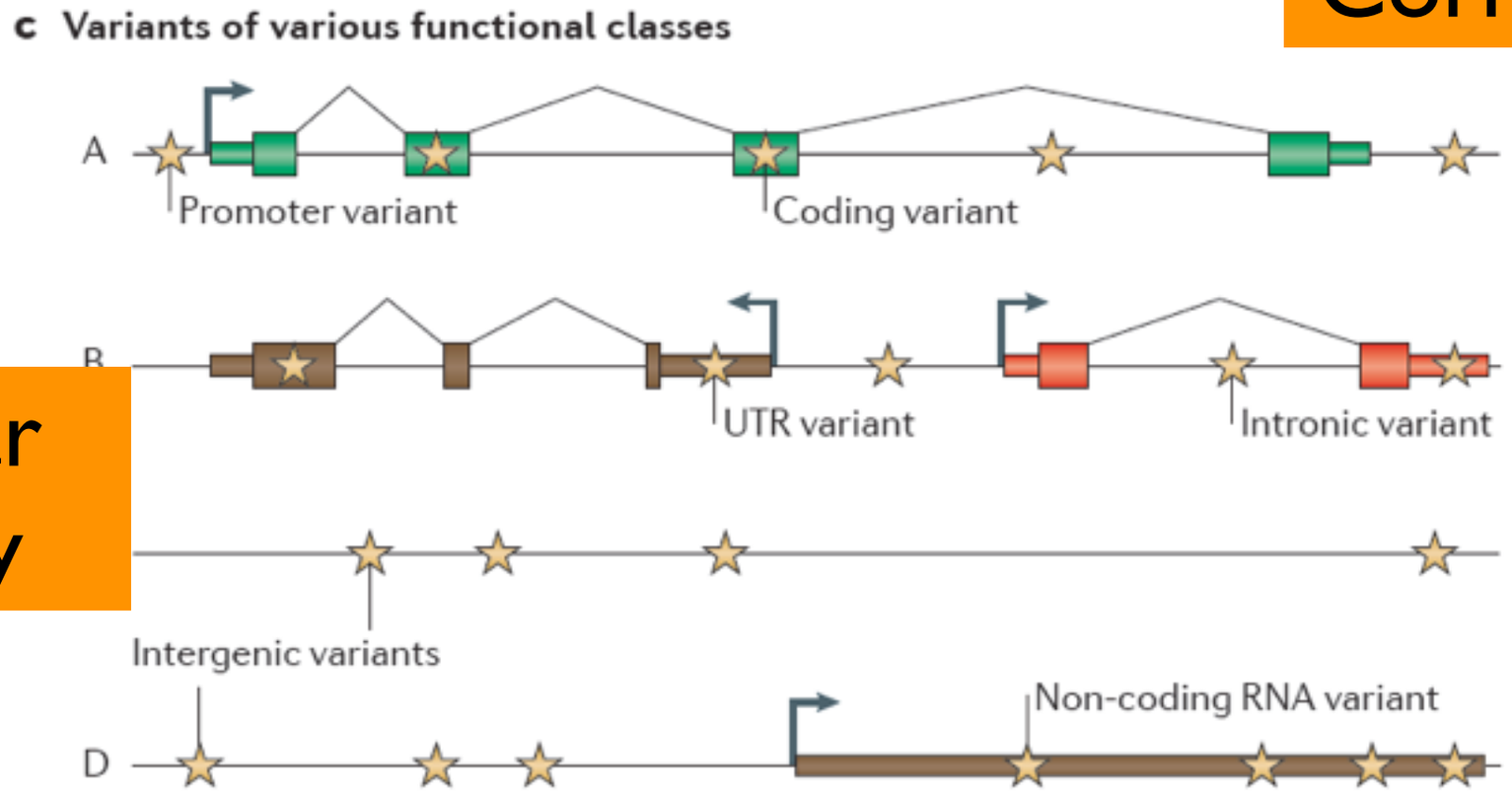
Human Genetics



Statistical Power

Haplotypic Correlations

Molecular Diversity



Cooper and Shendure *NRG* 2011

Needles in Stacks of Needles

•Sequencing and genotyping assays can identify much of the variation present in human genomes

Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants

Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate

Needles in Stacks of Needles

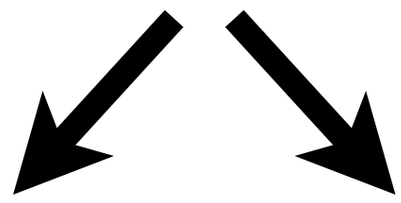
- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate
  - weak or no statistical separation among haplotypically correlated alleles

Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate
  - weak or no statistical separation among haplotypically correlated alleles
- Smart use of information required to prioritize relevant variants

Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate
  - weak or no statistical separation among haplotypically correlated alleles
- Smart use of information required to prioritize relevant variants



Function Evolution

Some Terminology

•Scoring, annotation, ranking, or interpretation schemes generally try to estimate one or more of 4 correlated but distinct variant properties:

Some Terminology

- Scoring, annotation, ranking, or interpretation schemes generally try to estimate one or more of 4 correlated but distinct variant properties:
  - Within a biochemically active "functional" element

Some Terminology

- Scoring, annotation, ranking, or interpretation schemes generally try to estimate one or more of 4 correlated but distinct variant properties:
  - Within a biochemically active "functional" element
  - Molecular function, i.e., has a proximate molecular consequence

Some Terminology

- Scoring, annotation, ranking, or interpretation schemes generally try to estimate one or more of 4 correlated but distinct variant properties:
  - Within a biochemically active "functional" element
  - Molecular function, i.e., has a proximate molecular consequence
  - Causal, i.e., contributing directly to a phenotype

Some Terminology

- Scoring, annotation, ranking, or interpretation schemes generally try to estimate one or more of 4 correlated but distinct variant properties:
  - Within a biochemically active "functional" element
  - Molecular function, i.e., has a proximate molecular consequence
    - Causal, i.e., contributing directly to a phenotype
      - subset of which is disease-causal or pathogenic

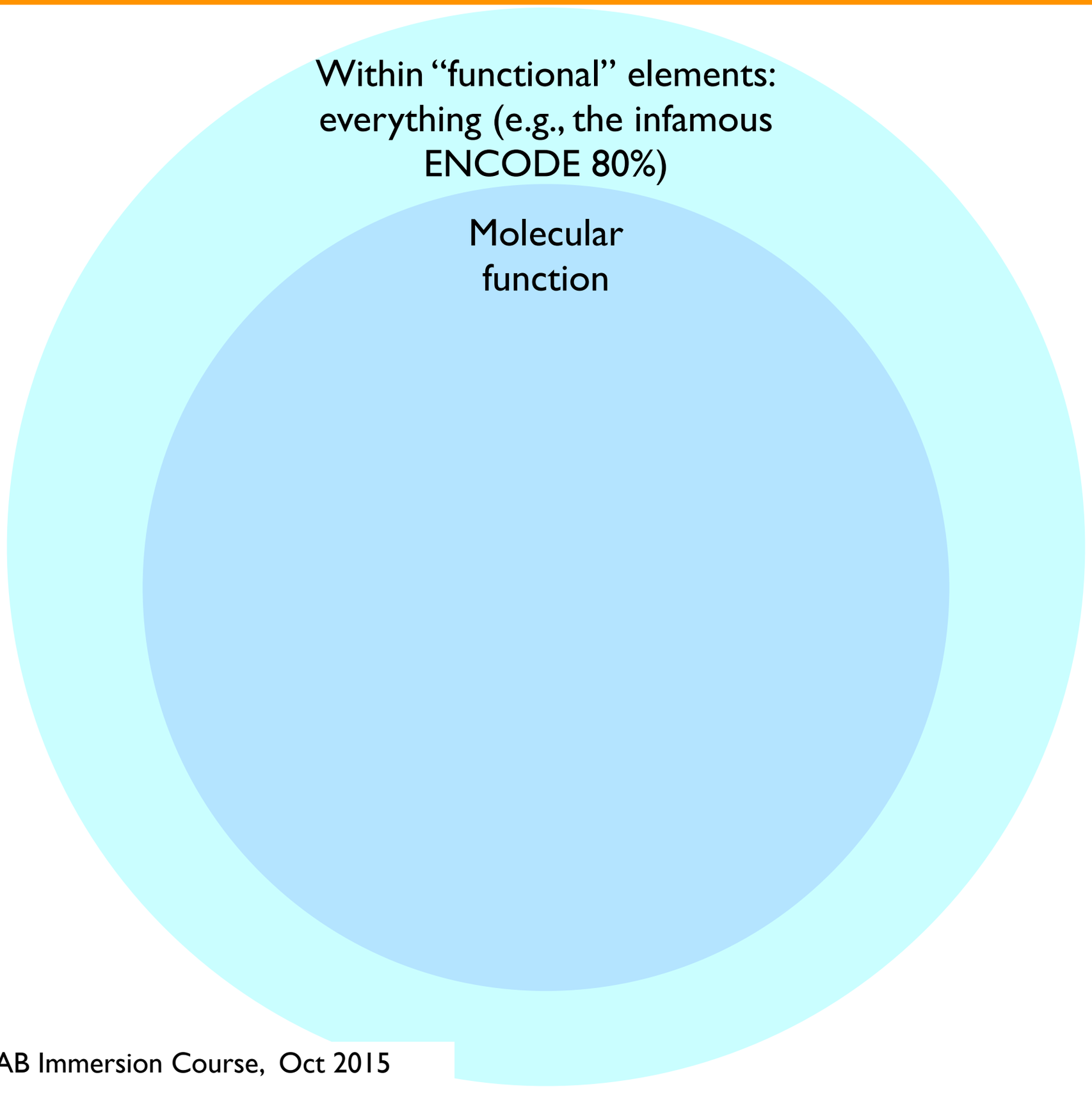
Some Terminology

- Scoring, annotation, ranking, or interpretation schemes generally try to estimate one or more of 4 correlated but distinct variant properties:
  - Within a biochemically active "functional" element
  - Molecular function, i.e., has a proximate molecular consequence
    - Causal, i.e., contributing directly to a phenotype
      - subset of which is disease-causal or pathogenic
  - Deleterious, i.e., reduces survival or reproductive success and imposes fitness costs

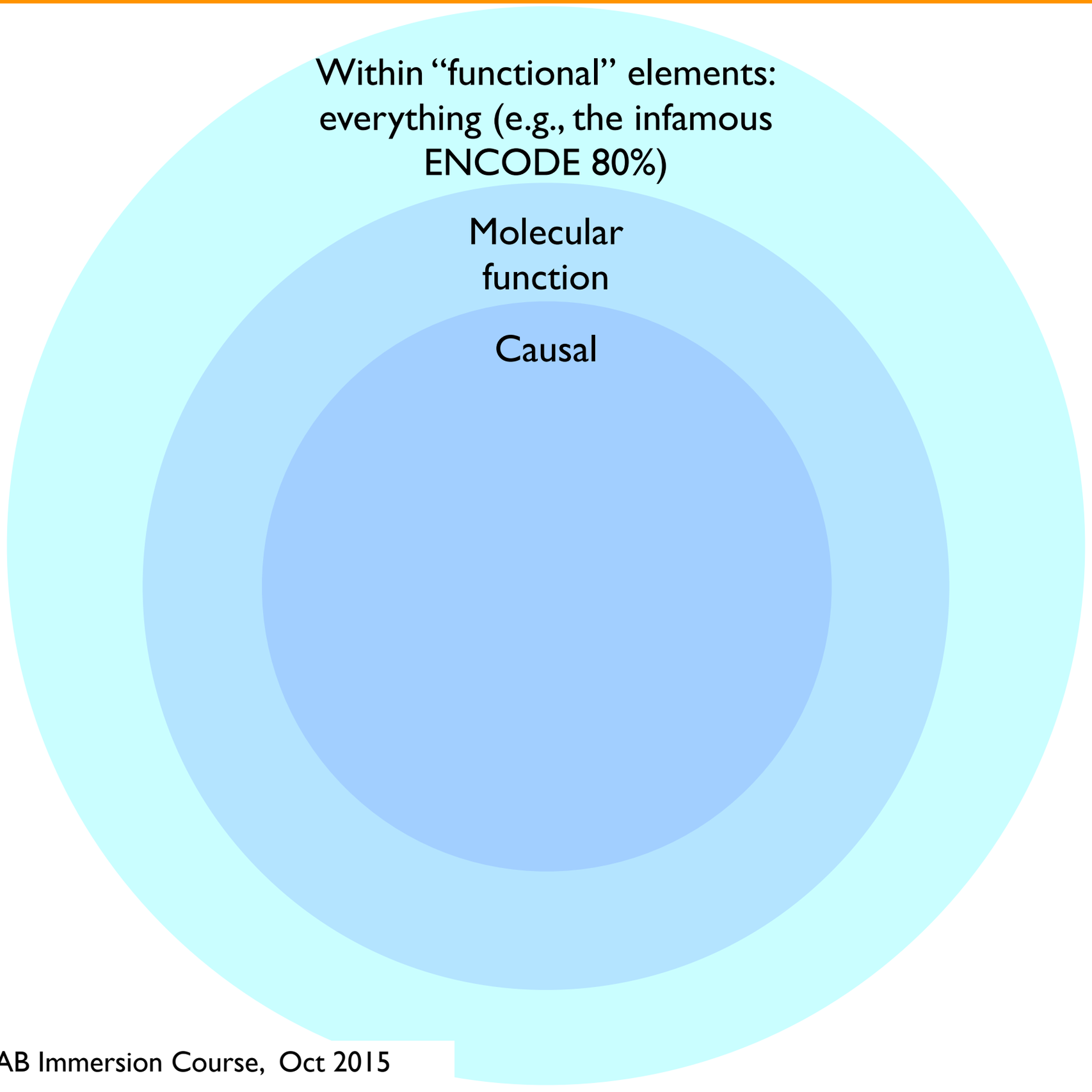
Property Domains

Within "functional" elements:
everything (e.g., the infamous
ENCODE 80%)

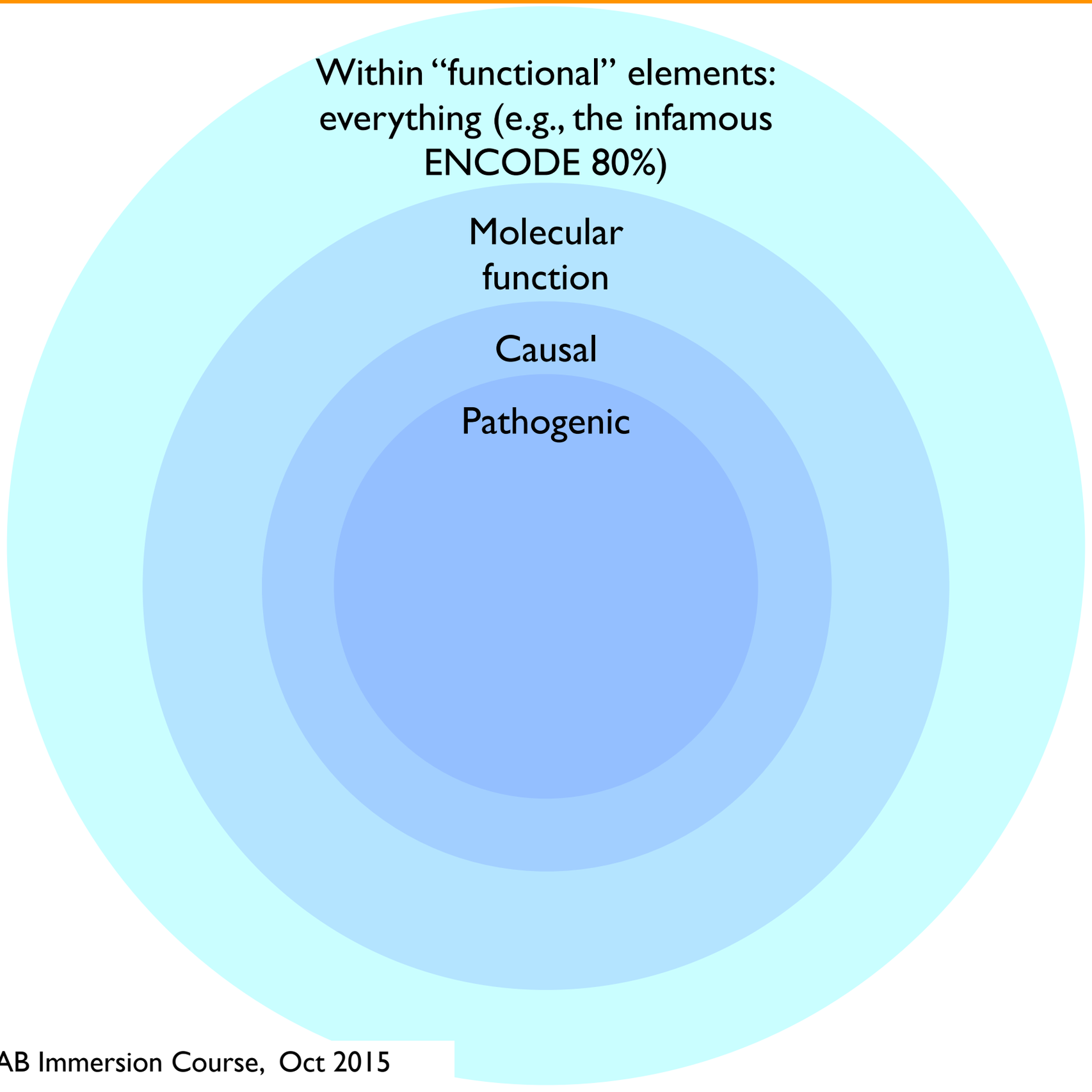
Property Domains



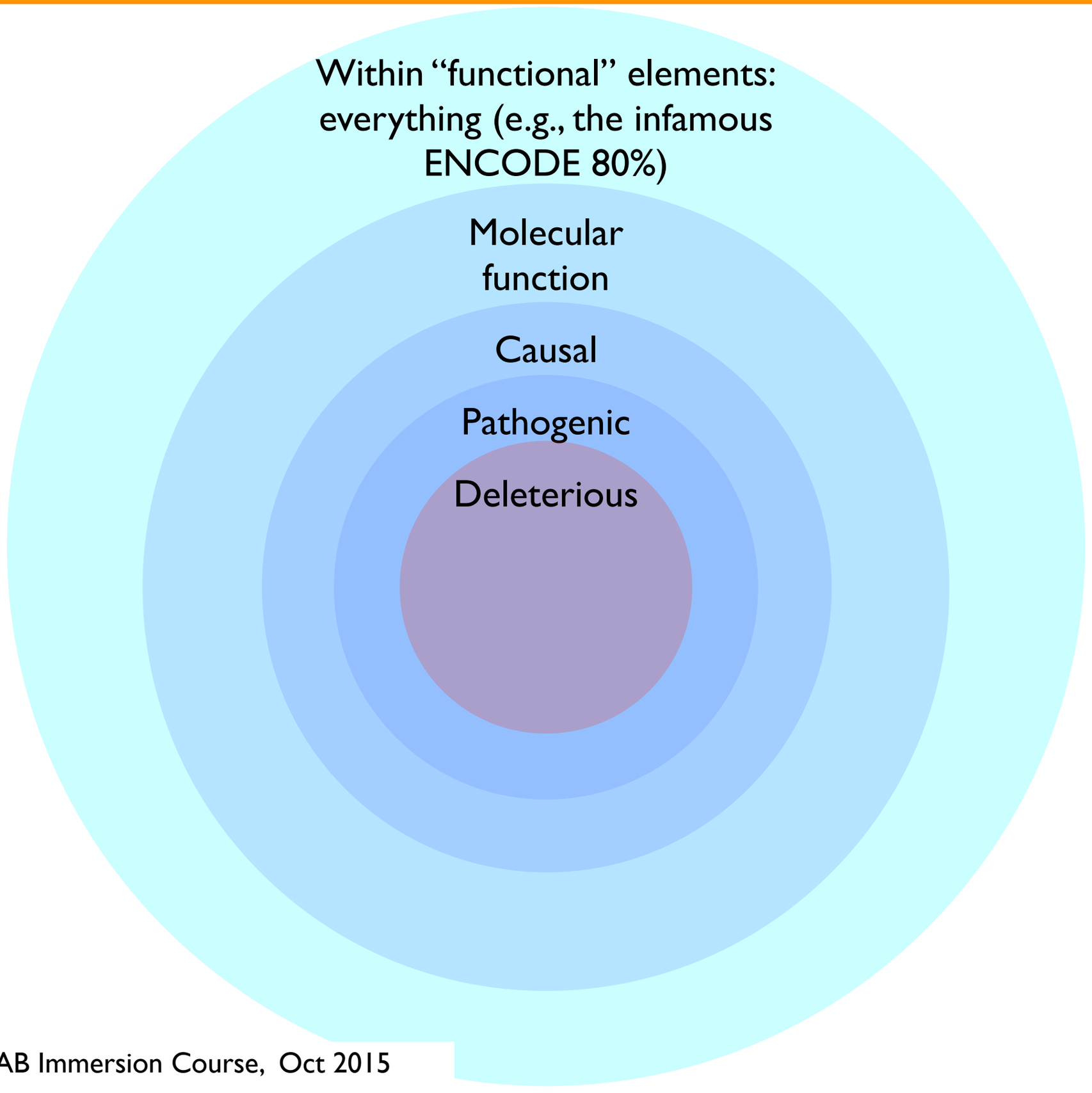
Property Domains



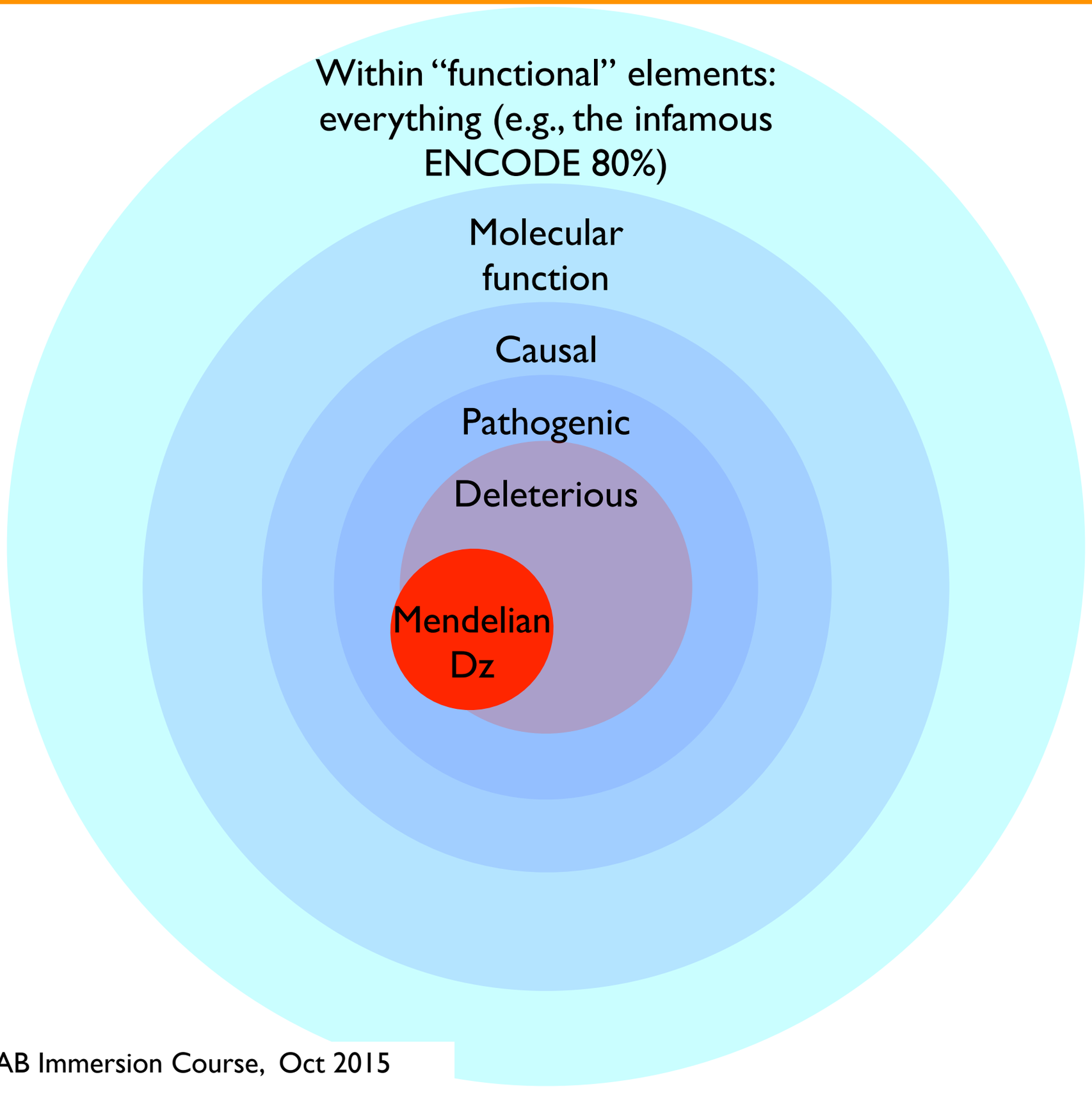
Property Domains



Property Domains

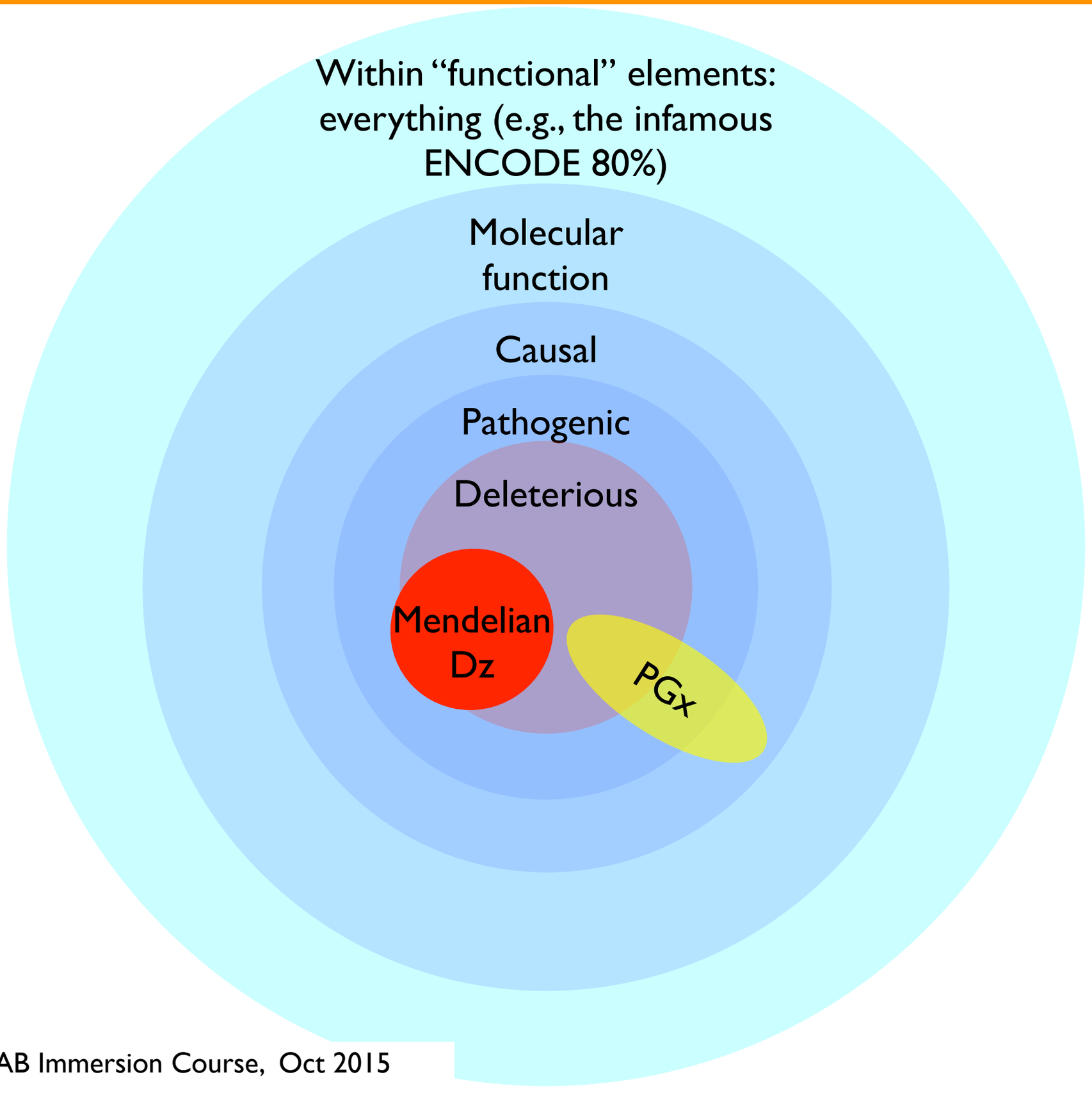


Property Domains

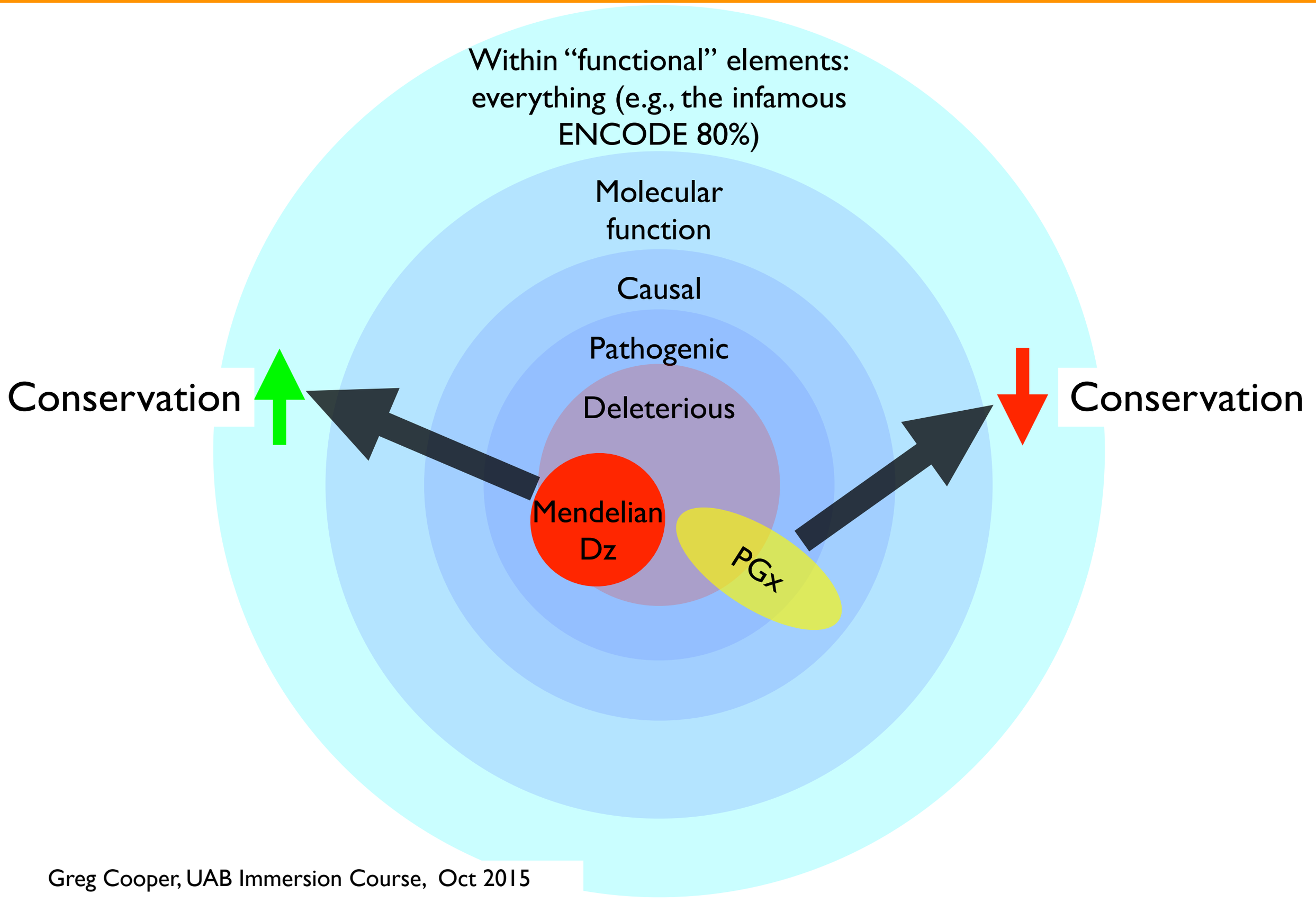


Greg Cooper, UAB Immersion Course, Oct 2015

Property Domains

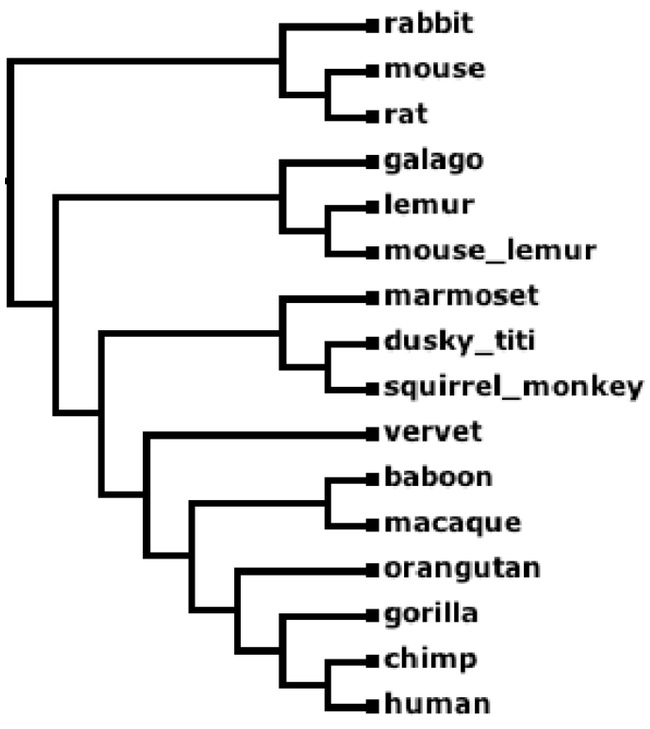


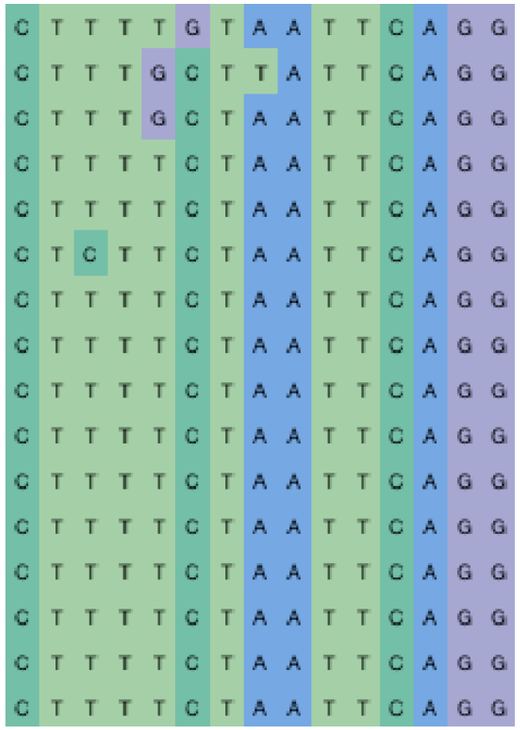
Property Domains

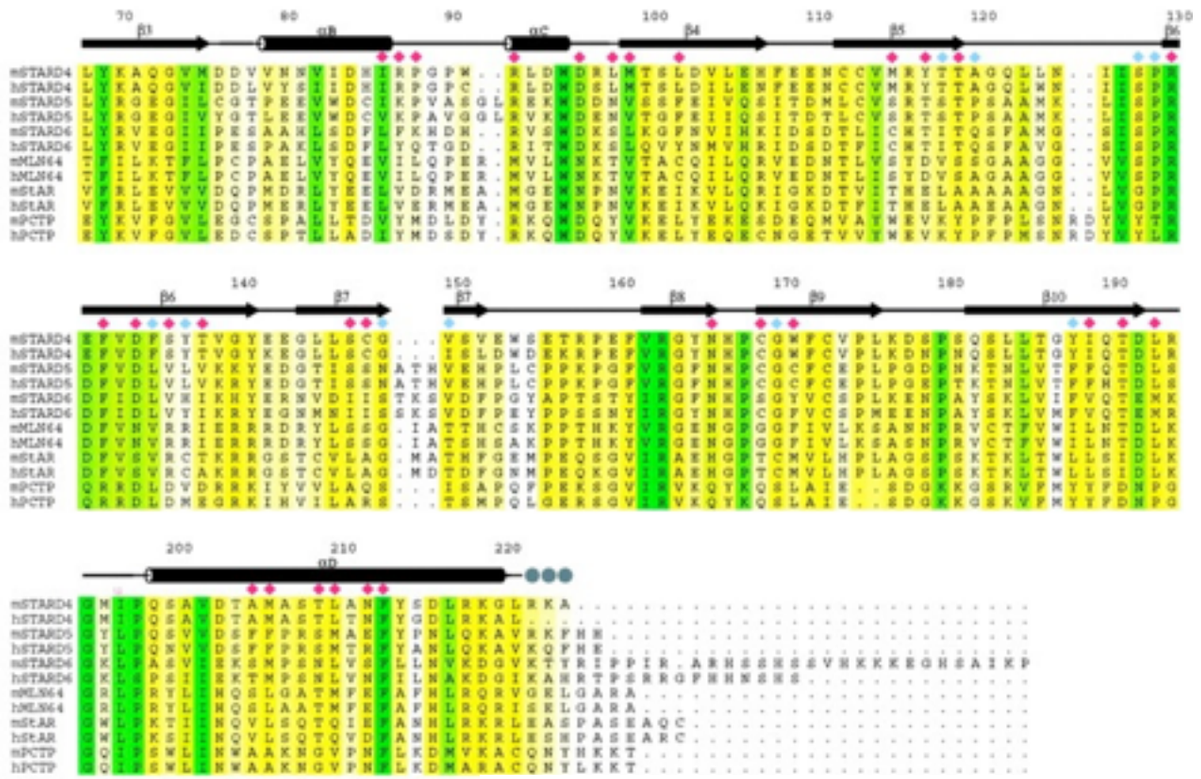


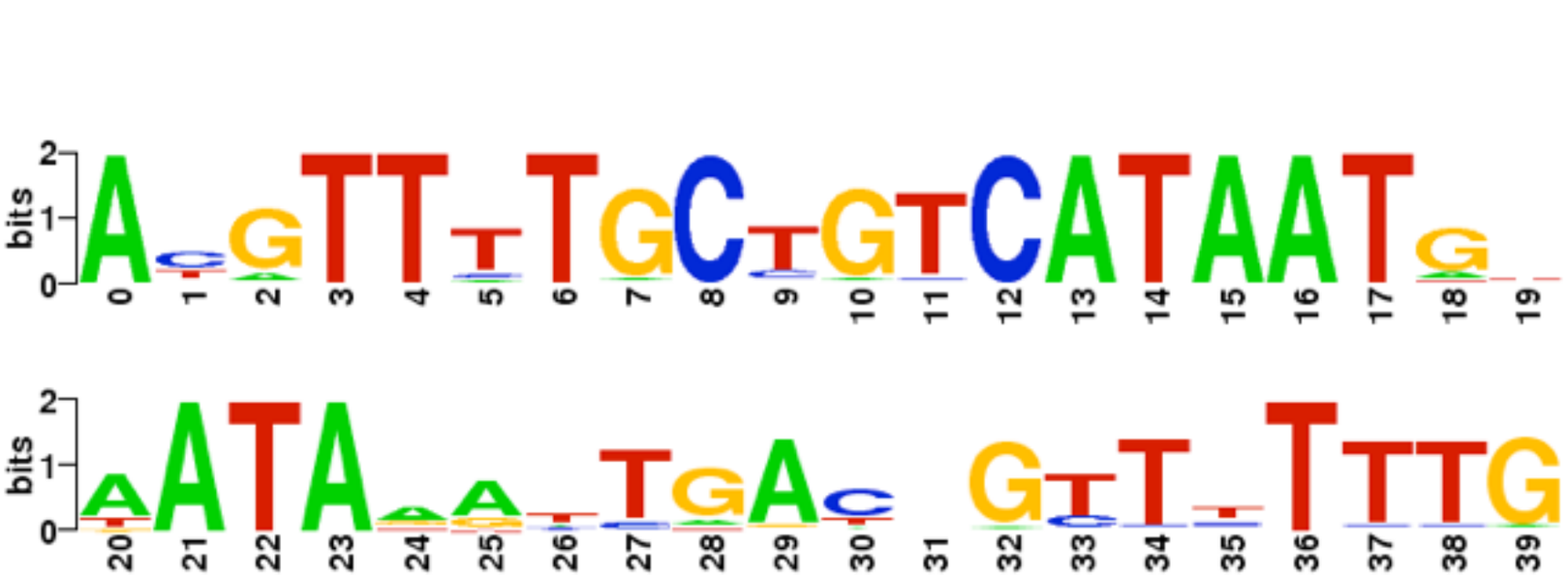
Greg Cooper, UAB Immersion Course, Oct 2015

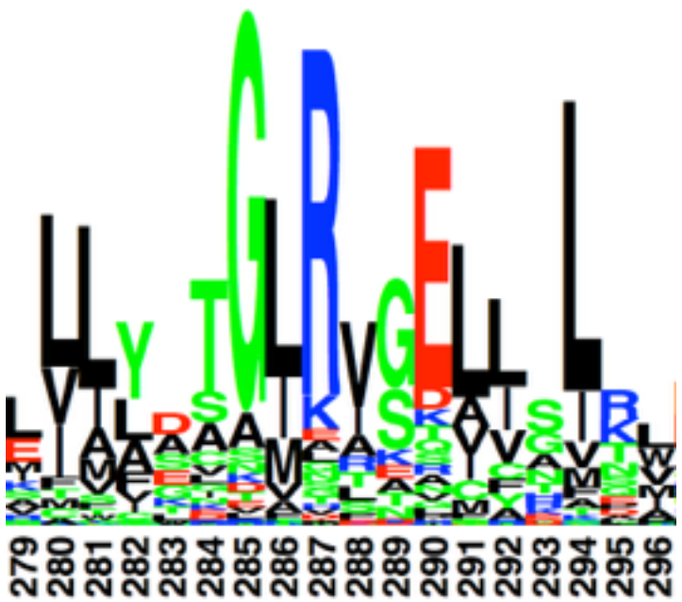
Sequence Comparisons











Greg Cooper, UAB Immersion Course, Oct 2015

Information from Evolution

- Purifying selection, or evolutionary constraint, reduces rates of evolution at functional sites

Information from Evolution

- Purifying selection, or evolutionary constraint, reduces rates of evolution at functional sites

  - Sites in genes or genomes that show higher levels of sequence "conservation" are of potential functional importance

Information from Evolution

- Purifying selection, or evolutionary constraint, reduces rates of evolution at functional sites
  - Sites in genes or genomes that show higher levels of sequence "conservation" are of potential functional importance
  - This relationship is quantitative and useful within many phylogenetic scopes and for many types of function

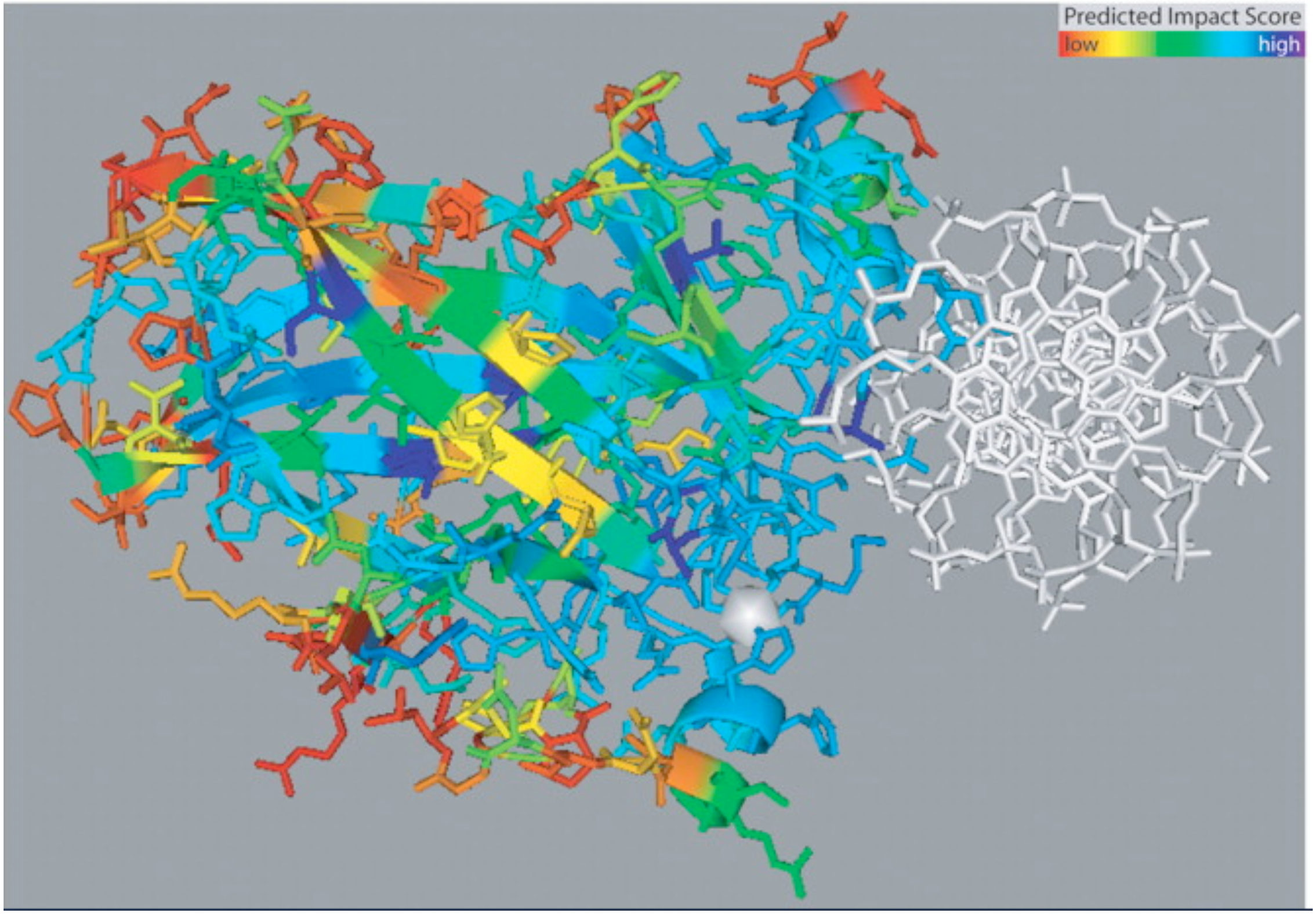
Information from Evolution

- Purifying selection, or evolutionary constraint, reduces rates of evolution at functional sites
  - Sites in genes or genomes that show higher levels of sequence "conservation" are of potential functional importance
    - This relationship is quantitative and useful within many phylogenetic scopes and for many types of function
  - Shared sequences like motifs, domains, etc are also informative

Information from Evolution

- Purifying selection, or evolutionary constraint, reduces rates of evolution at functional sites
  - Sites in genes or genomes that show higher levels of sequence "conservation" are of potential functional importance
  - This relationship is quantitative and useful within many phylogenetic scopes and for many types of function
- Shared sequences like motifs, domains, etc are also informative
- Sequence similarity measures crucial to all methods of variant annotation and to both coding and non-coding variants

Constraint and Structure

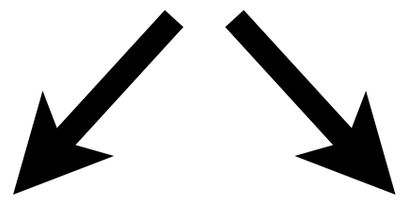


Greg Cooper, UAB Immersion Course, Oct 2015

Stone and Sidow 2005 *Genome Research*

Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate
  - weak or no statistical separation among haplotypically correlated alleles
- Smart use of information required to prioritize relevant variants

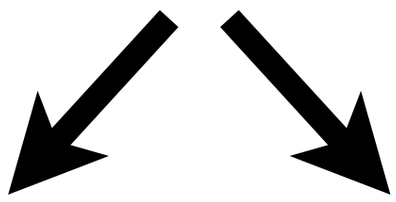


Function Evolution

Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate
  - weak or no statistical separation among haplotypically correlated alleles
- Smart use of information required to prioritize relevant variants

Non-Synonymous Variants



Function Evolution

Annotations for Non-Synonymous Variants

Focus on non-synonymous variants because:

- Enriched for functional effects
- Enriched for disease effects, especially severe and Mendelian
- Stop codons and splice-site disruptions in special category
- However, most observed non-synonymous variants likely have little or no effect on phenotype

Non-Synonymous SNP Annotation

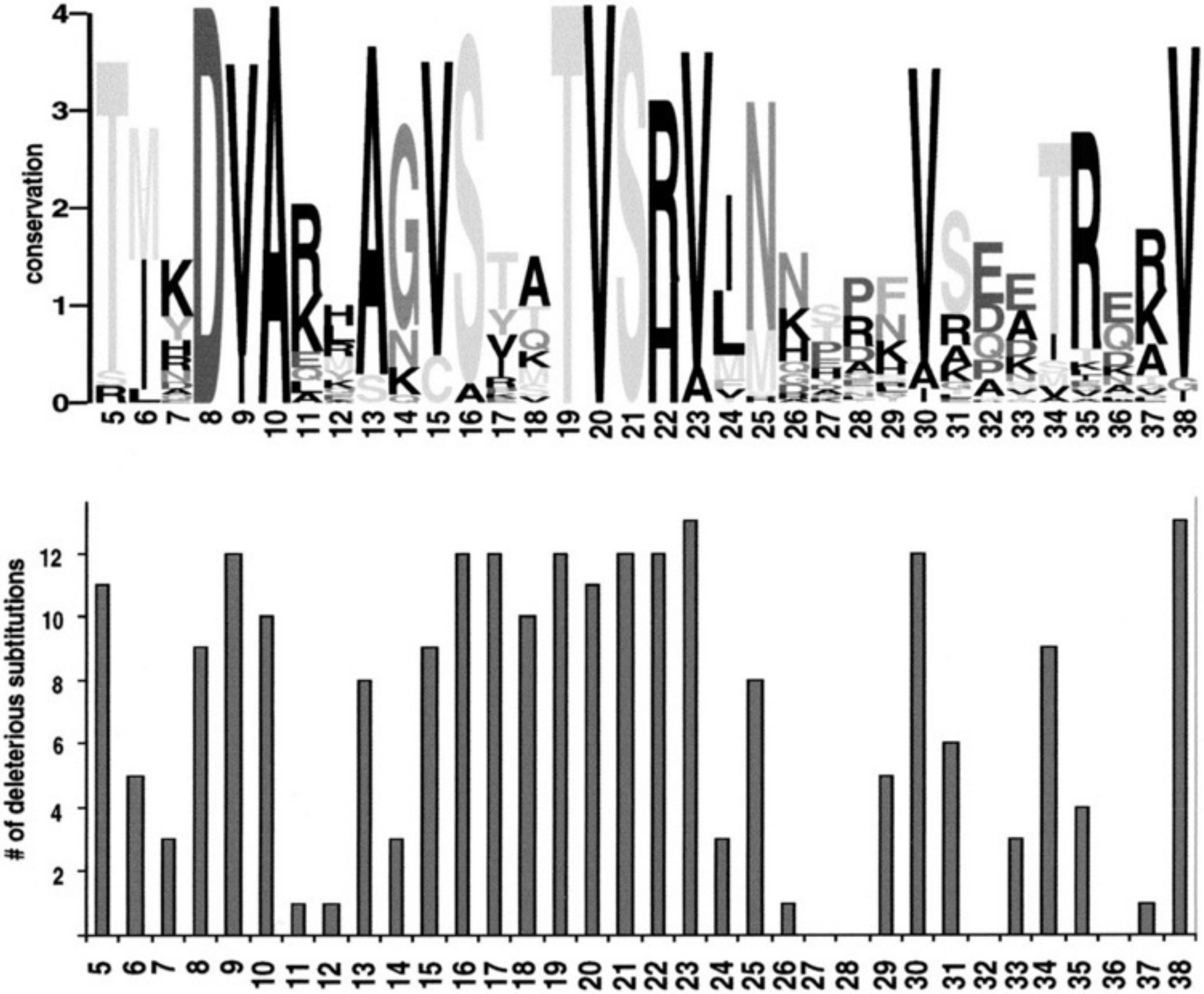
Table 1 | **Tools for protein-sequence-based prediction of deleteriousness**

Name	Type	Information	URL	Refs
MAPP	Constraint-based predictor	Evolutionary and biochemical	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	27
SIFT	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://sift.bii.a-star.edu.sg/	39
PANTHER	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://www.pantherdb.org/	41
MutationTaster*	Trained classifier	Evolutionary, biochemical and structural	http://www.mutationtaster.org/	40
nsSNP Analyzer	Trained classifier	Evolutionary, biochemical and structural	http://snpanalyzer.uthsc.edu/	44
PMUT	Trained classifier	Evolutionary, biochemical and structural	http://mmb2.pcb.ub.es:8080/PMut/	38
polyPhen	Trained classifier	Evolutionary, biochemical and structural	http://genetics.bwh.harvard.edu/pph2/	35
SAPRED	Trained classifier	Evolutionary, biochemical and structural	http://sapred.cbi.pku.edu.cn/	42
SNAP	Trained classifier	Evolutionary, biochemical and structural	http://www.rostlab.org/services/SNAP/	36
SNPs3D	Trained classifier	Evolutionary, biochemical and structural	http://www.snps3d.org/	51
PhD-SNP	Trained classifier	Evolutionary and biochemical (indirect)	http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html	37

*Also makes predictions for synonymous and non-coding variant effects: for example, splicing. MAPP, Multivariate Analysis of Protein Polymorphism; polyPhen, polymorphism phenotyping.

Cooper and Shendure 2011 *NRG*

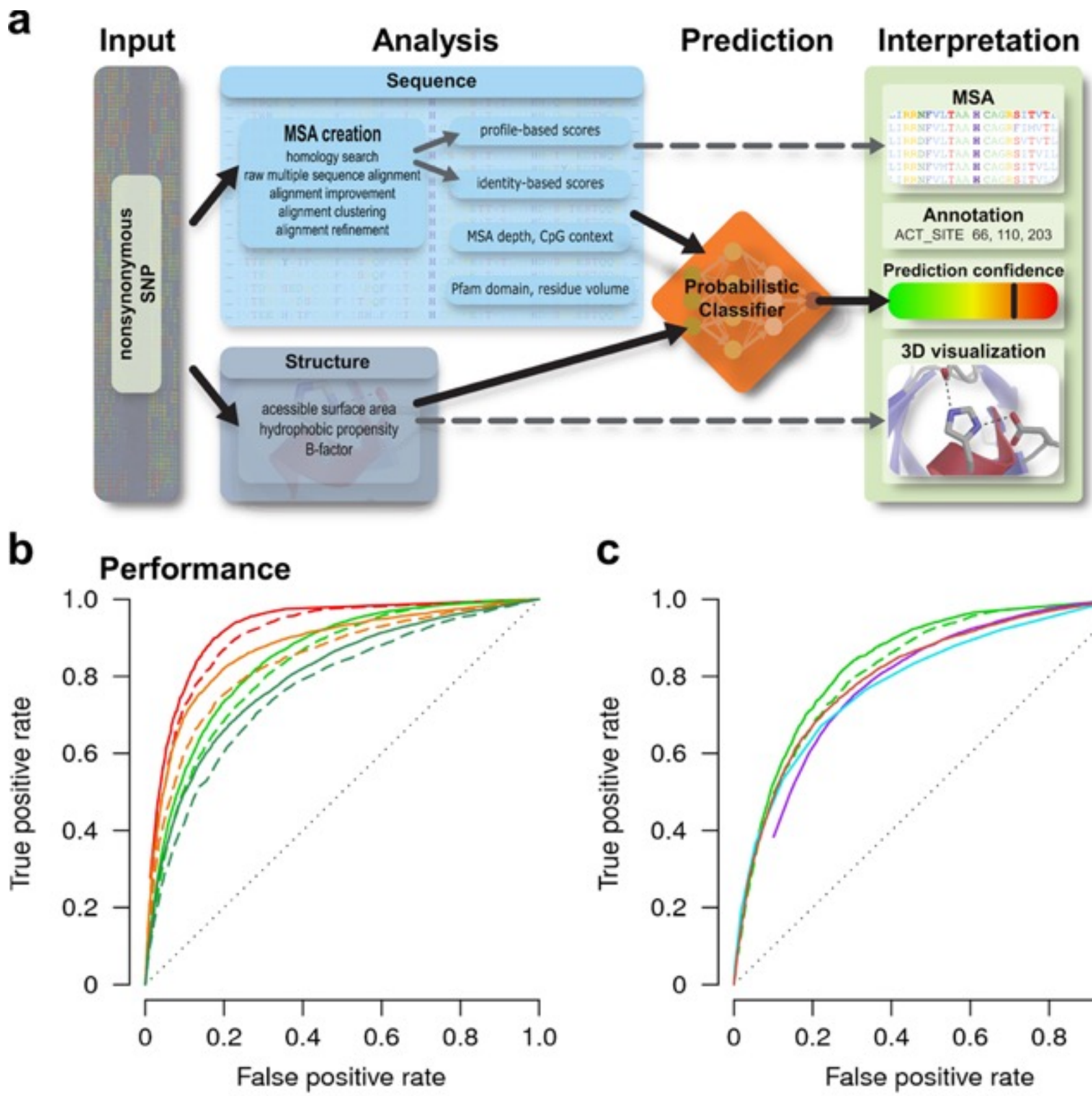
SIFT -- Sequence Conservation



Greg Cooper, UAB Immersion Course, Oct 2015

Ng and Henikoff 2001 *Genome Research*

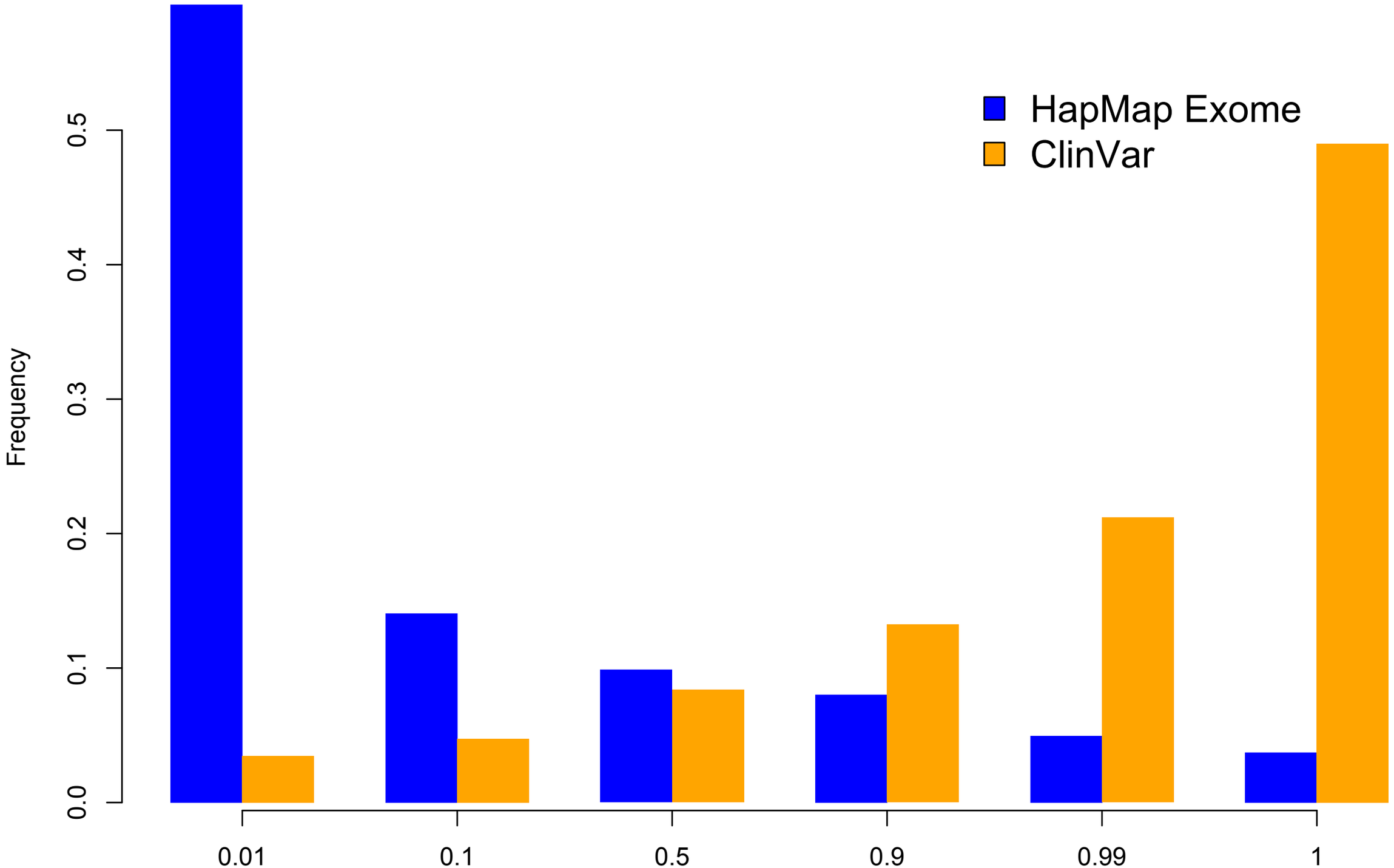
Polyphen -- Trained Classifier on Many Features



Greg Cooper, UAB Immersion Course, Oct 2015

Adzhubei *et al* 2010 *Nature Methods*

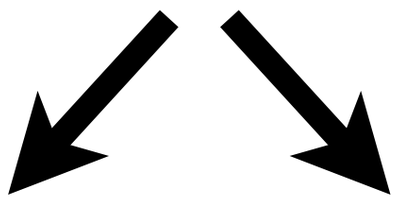
PolyPhen Enrichment for Pathogenic Mutations



Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate
  - weak or no statistical separation among haplotypically correlated alleles
- Smart use of information required to prioritize relevant variants

Non-Synonymous Variants

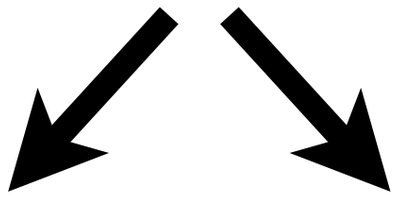


Function Evolution

Needles in Stacks of Needles

- Sequencing and genotyping assays can identify much of the variation present in human genomes
- Genetics alone often insufficient to identify many causal variants
  - among millions of mostly irrelevant variants, very low prior probability for any given candidate
  - weak or no statistical separation among haplotypically correlated alleles
- Smart use of information required to prioritize relevant variants

Non-Synonymous Variants



Function Evolution

What about the other 99%?

Greg Cooper, UAB Immersion Course, Oct 2015

Non-Coding Variant Annotation

Two main sources of data for annotating non-coding variants:

- Comparative genomics
  - sites under strong selection more likely to result in a deleterious effect when mutated
  - typically based on mammalian genomic comparisons, in contrast with protein-level annotations
- Functional genomics:
  - transcription factor binding, open chromatin, etc
  - chromatin modifications
  - proximity to known functional features like promoters, exons

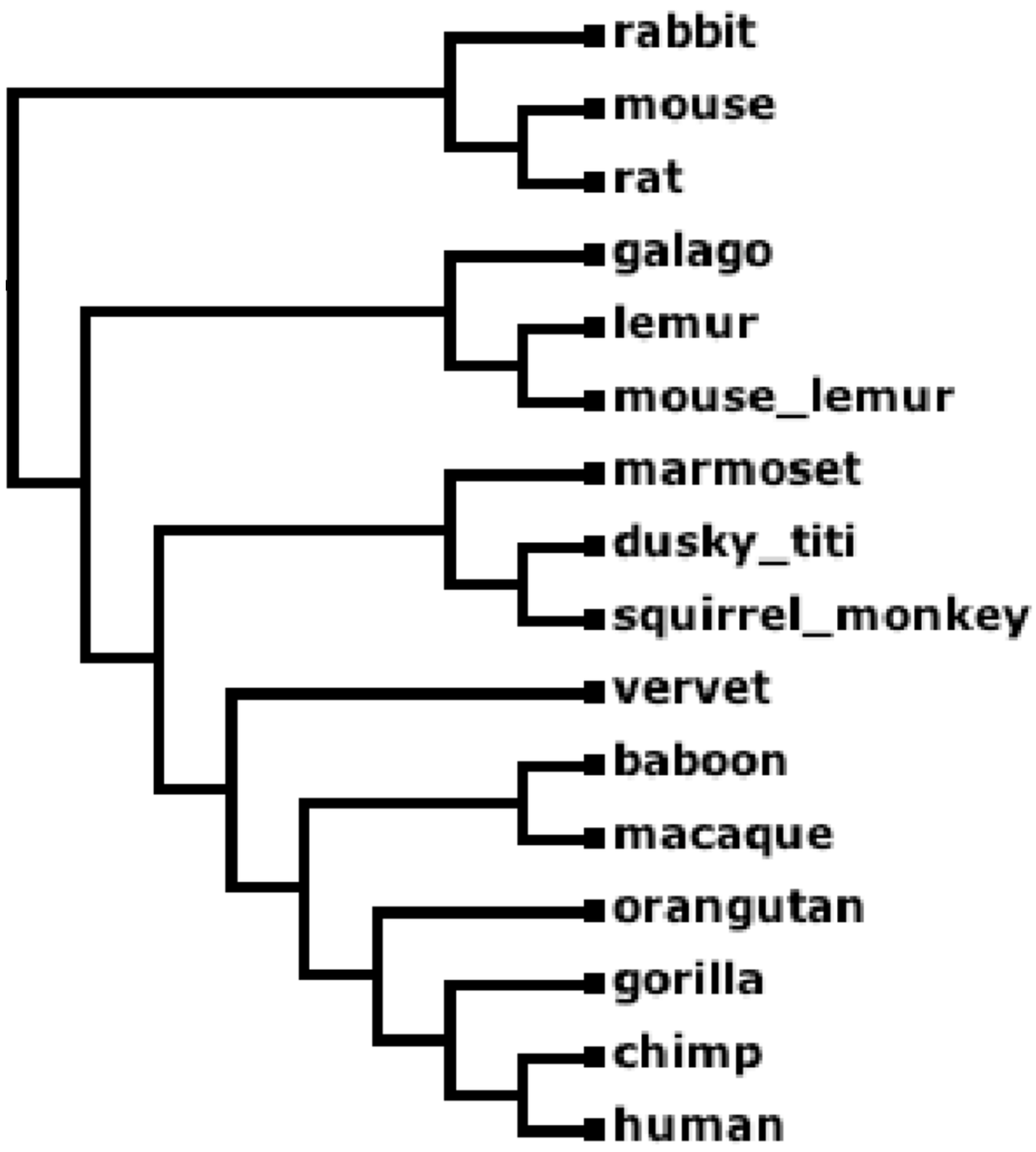
Nucleotide-Level Annotations From Comparative Genomics

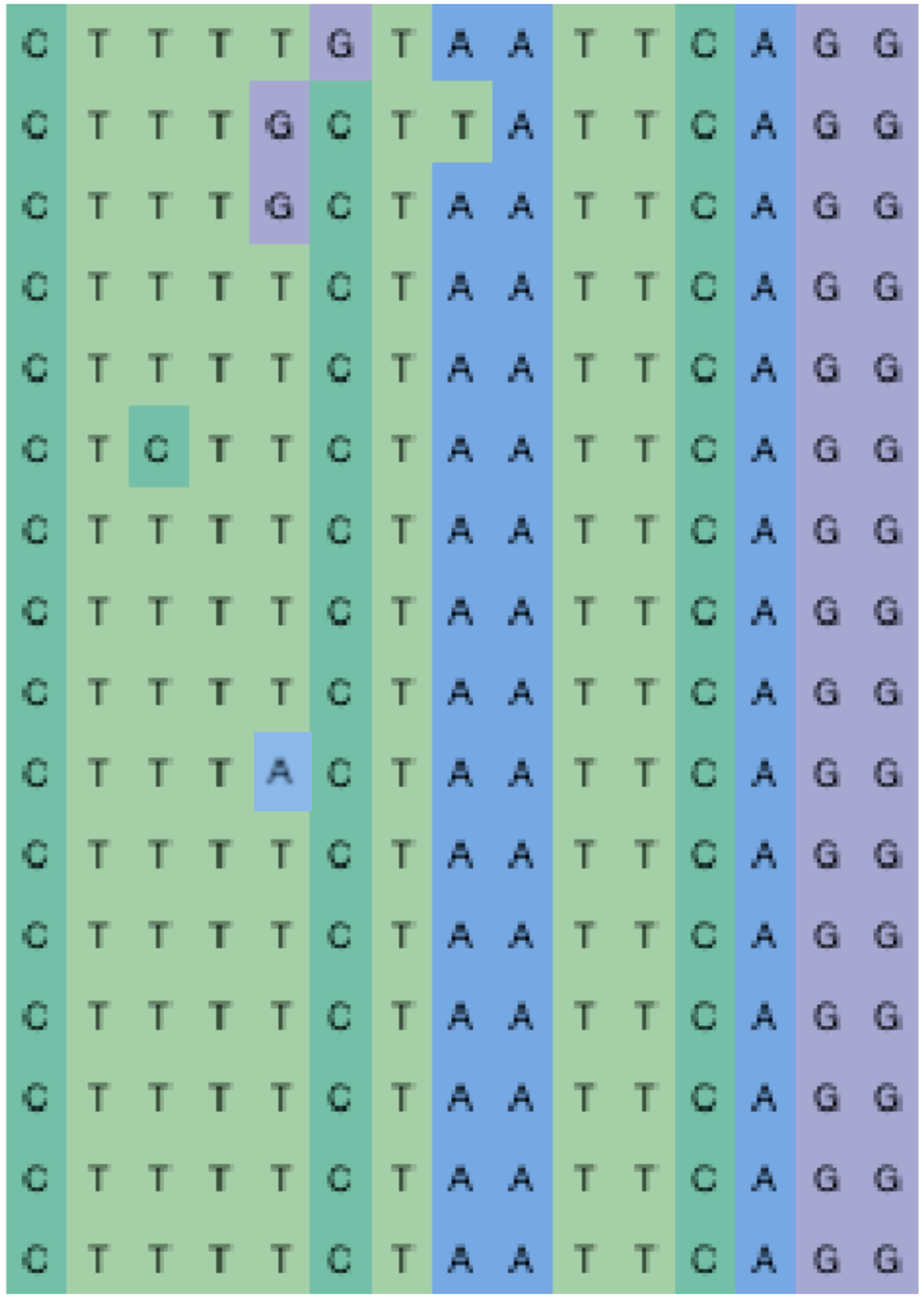
Table 2 | **Tools for nucleotide-sequence-based prediction of deleteriousness**

Name	Type	Information	URL	Refs
phastCons	Phylogenetic HMM	Evolutionary	http://compgen.bscb.cornell.edu/phast/	60
GERP	Single-site scoring	Evolutionary	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	67
Gumby	Single-site scoring	Evolutionary	http://pga.jgi-psf.org/gumby/	21
phyloP	Single-site scoring	Evolutionary	http://compgen.bscb.cornell.edu/phast/	66
SCONE	Single-site scoring	Evolutionary	http://genetics.bwh.harvard.edu/scone/	68
binCons	Sliding-window scoring	Evolutionary	http://zoo.nhgri.nih.gov/binCons/index.cgi	69
Chai Cons	Sliding-window scoring	Evolutionary and structural	http://research.nhgri.nih.gov/software/chai	71
VISTA	Visualization tool (various scores)	Evolutionary	http://genome.lbl.gov/vista/index.shtml	70

GERP, Genomic Evolutionary Rate Profiling; HMM, hidden Markov model; SCONE, Sequence Conservation Evaluation.

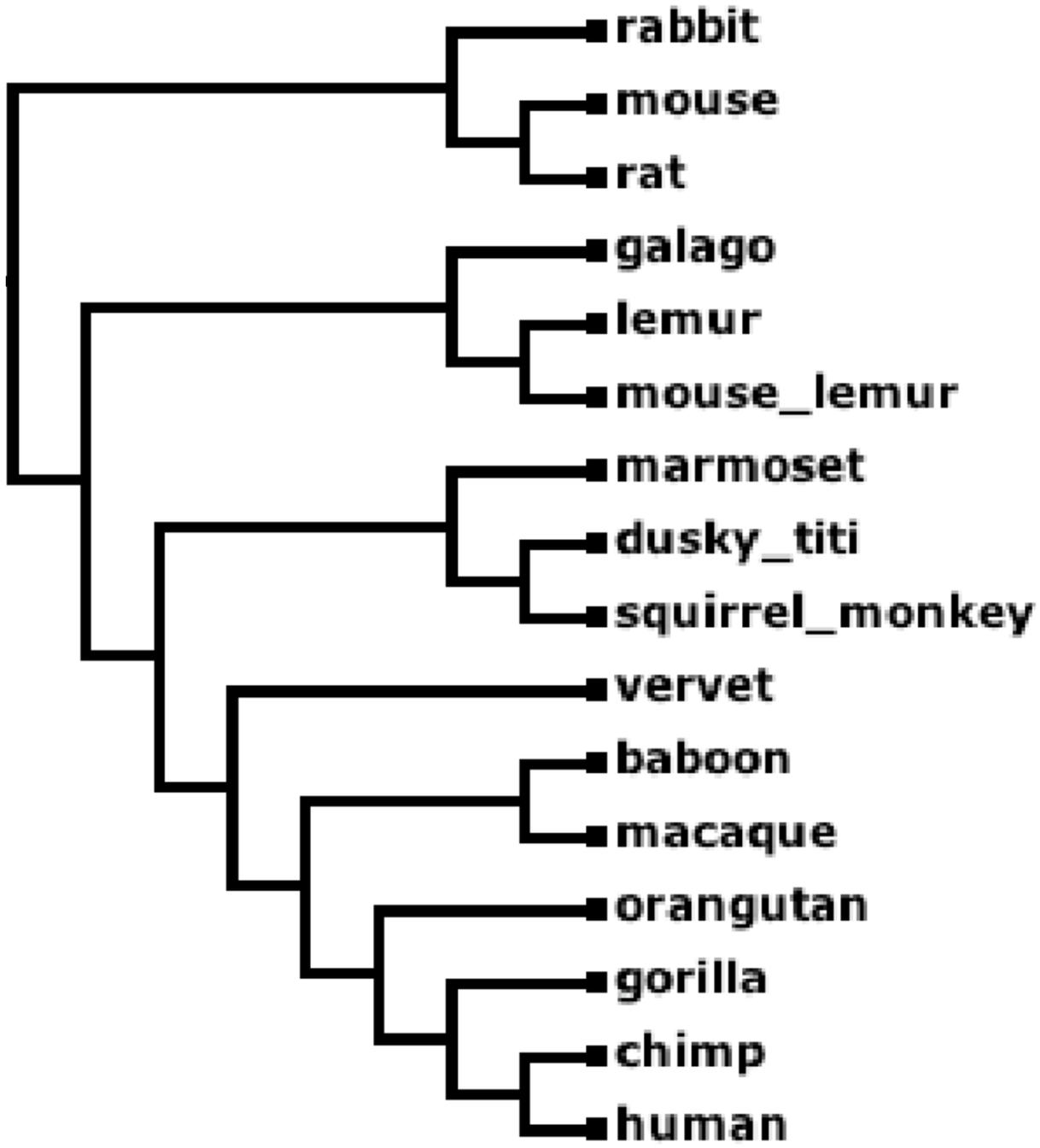
Genome Sequence Comparisons

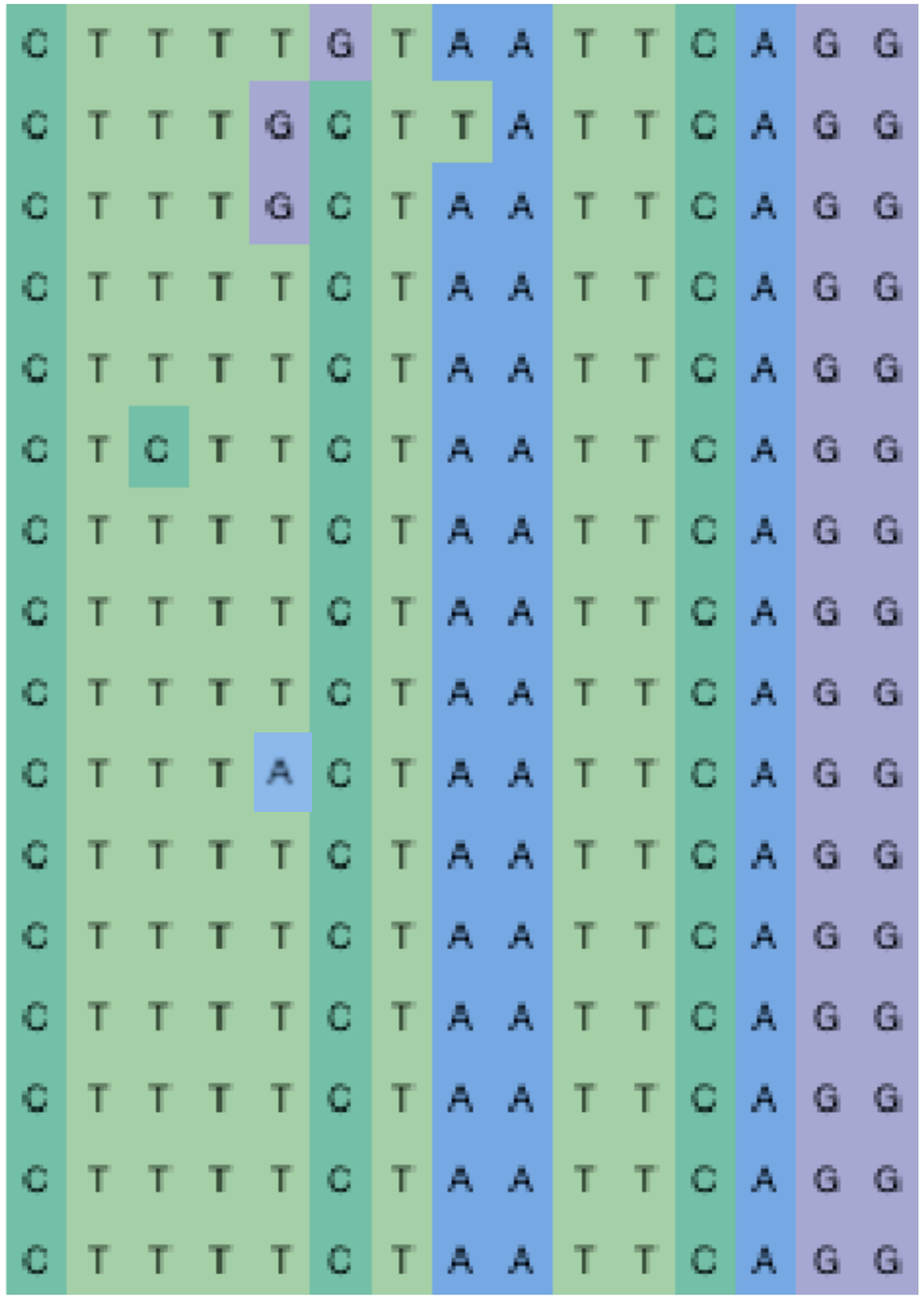




Cooper *et al* 2005 *Genome Research*

Genome Sequence Comparisons

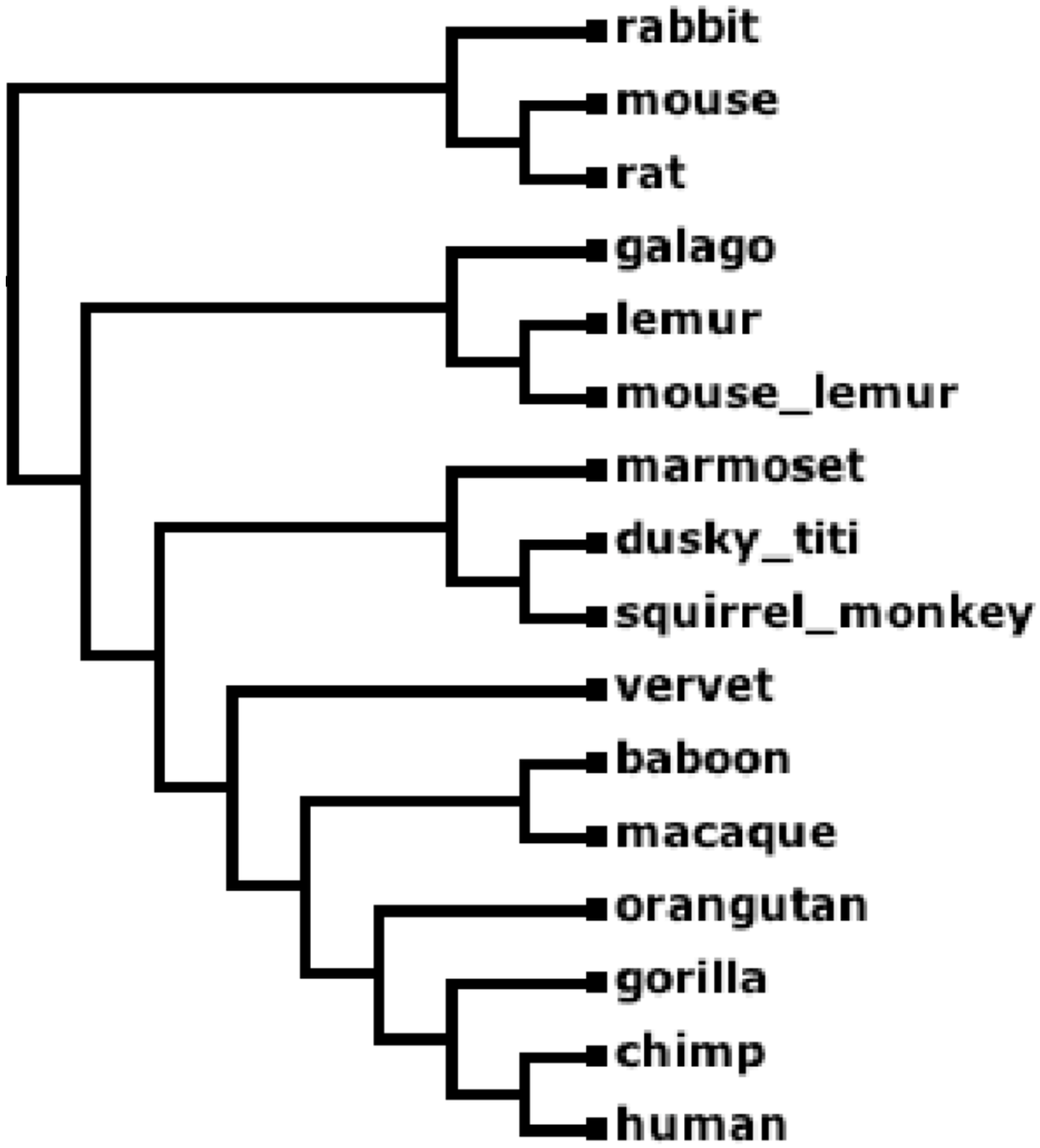


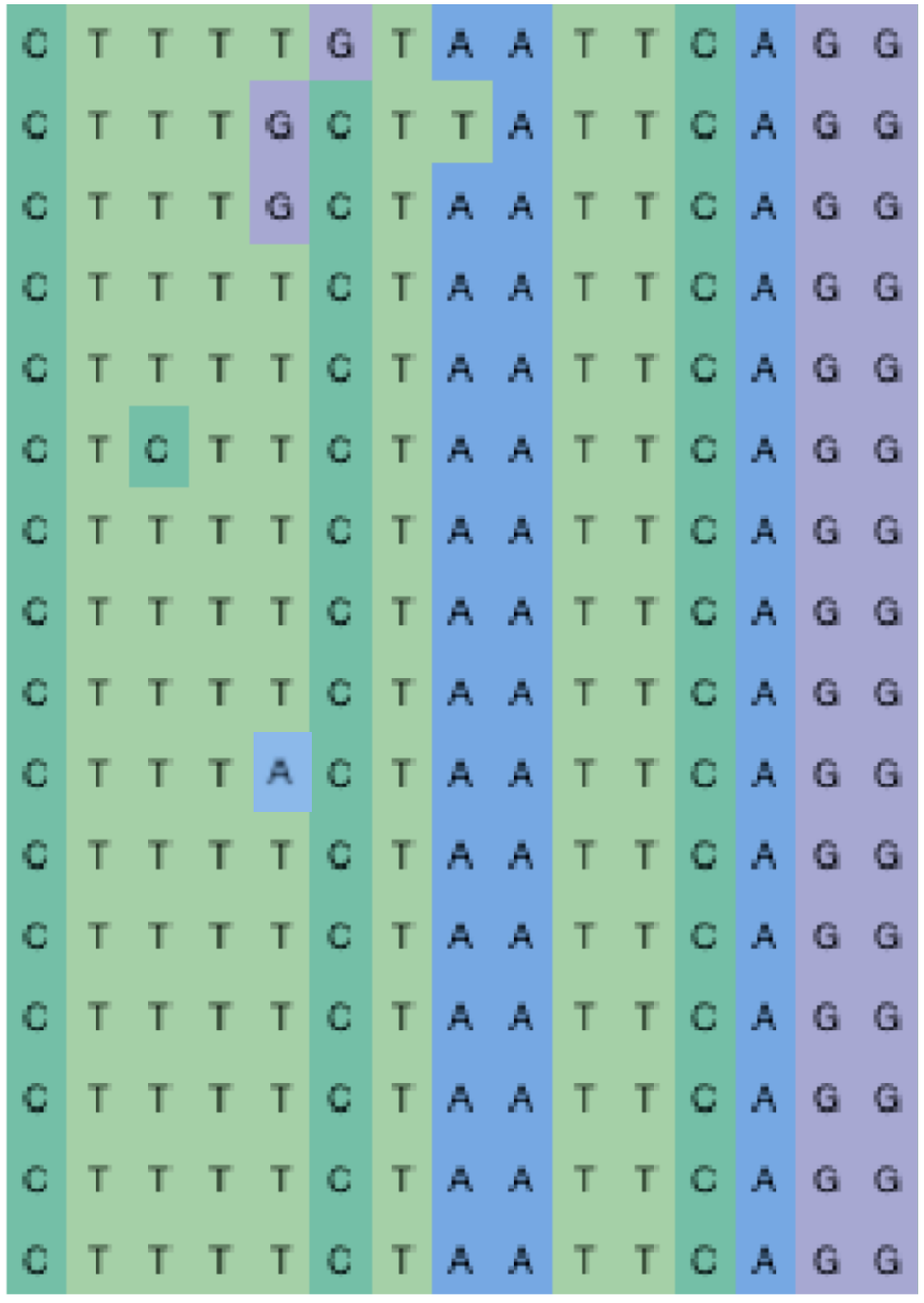


Neutral Average
~5 subs/site

Cooper *et al* 2005 *Genome Research*

Genome Sequence Comparisons



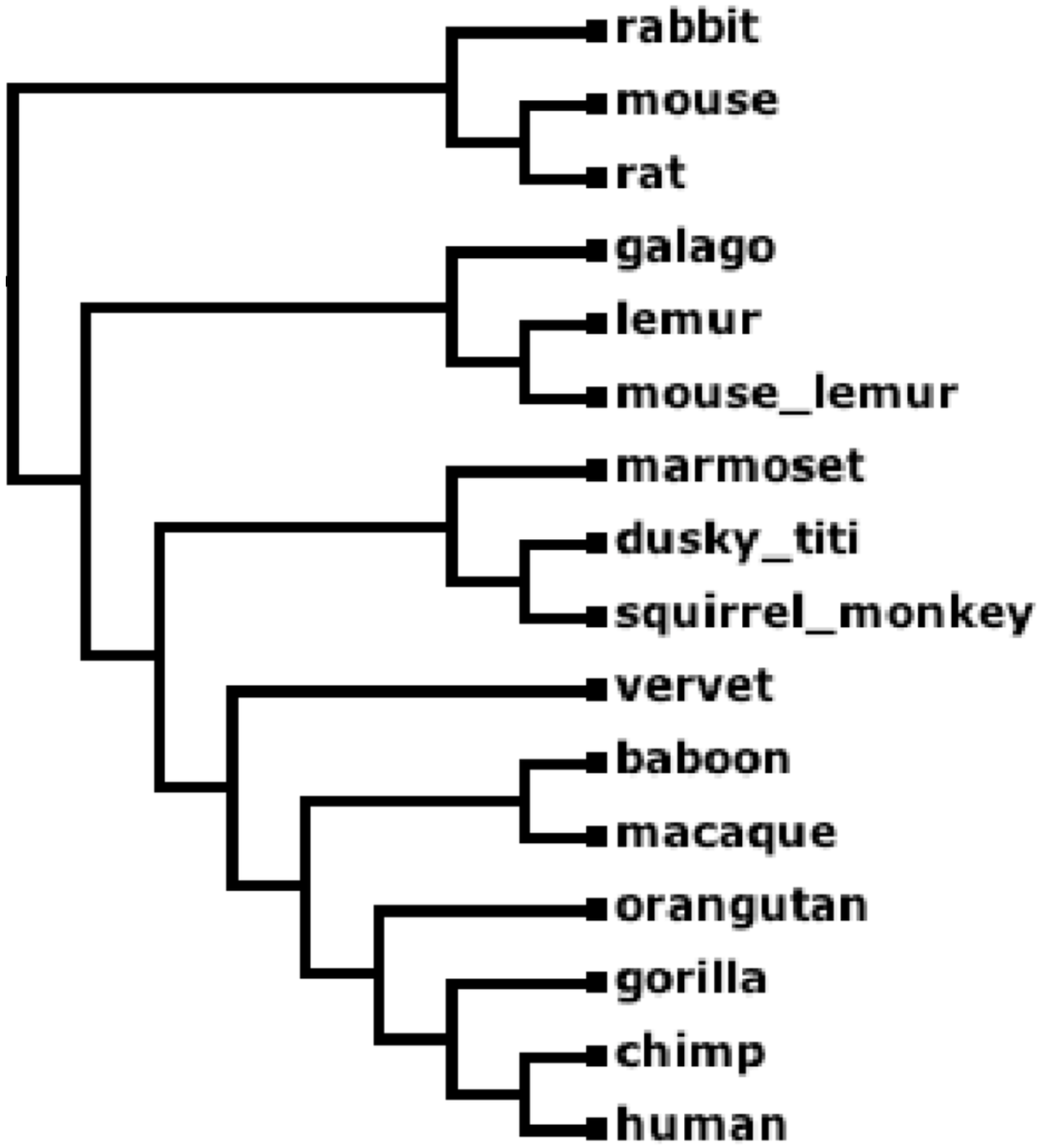


Neutral Average
~5 subs/site

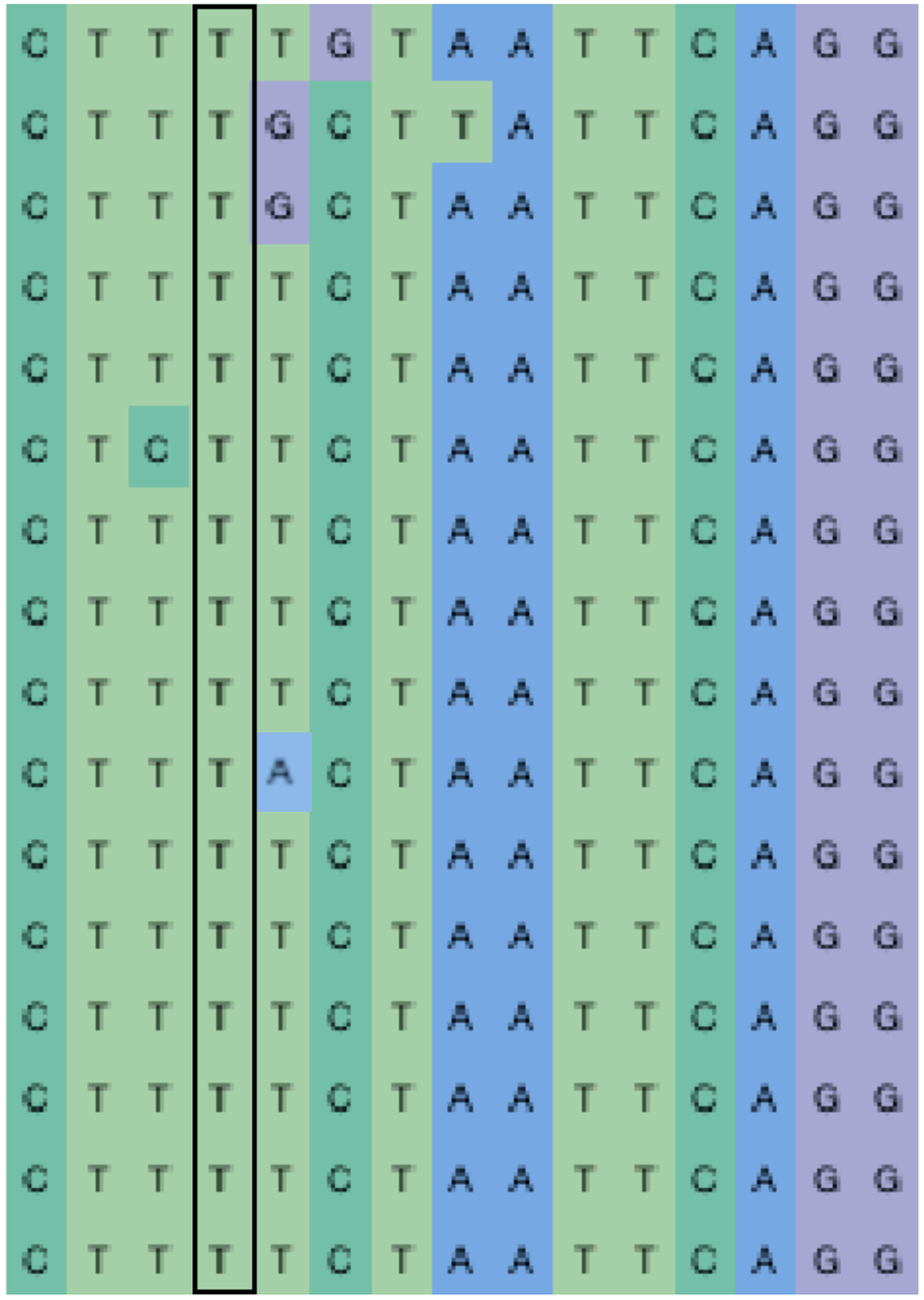
'Rej Subs':

Cooper *et al* 2005 *Genome Research*

Genome Sequence Comparisons



Neutral Average
~5 subs/site

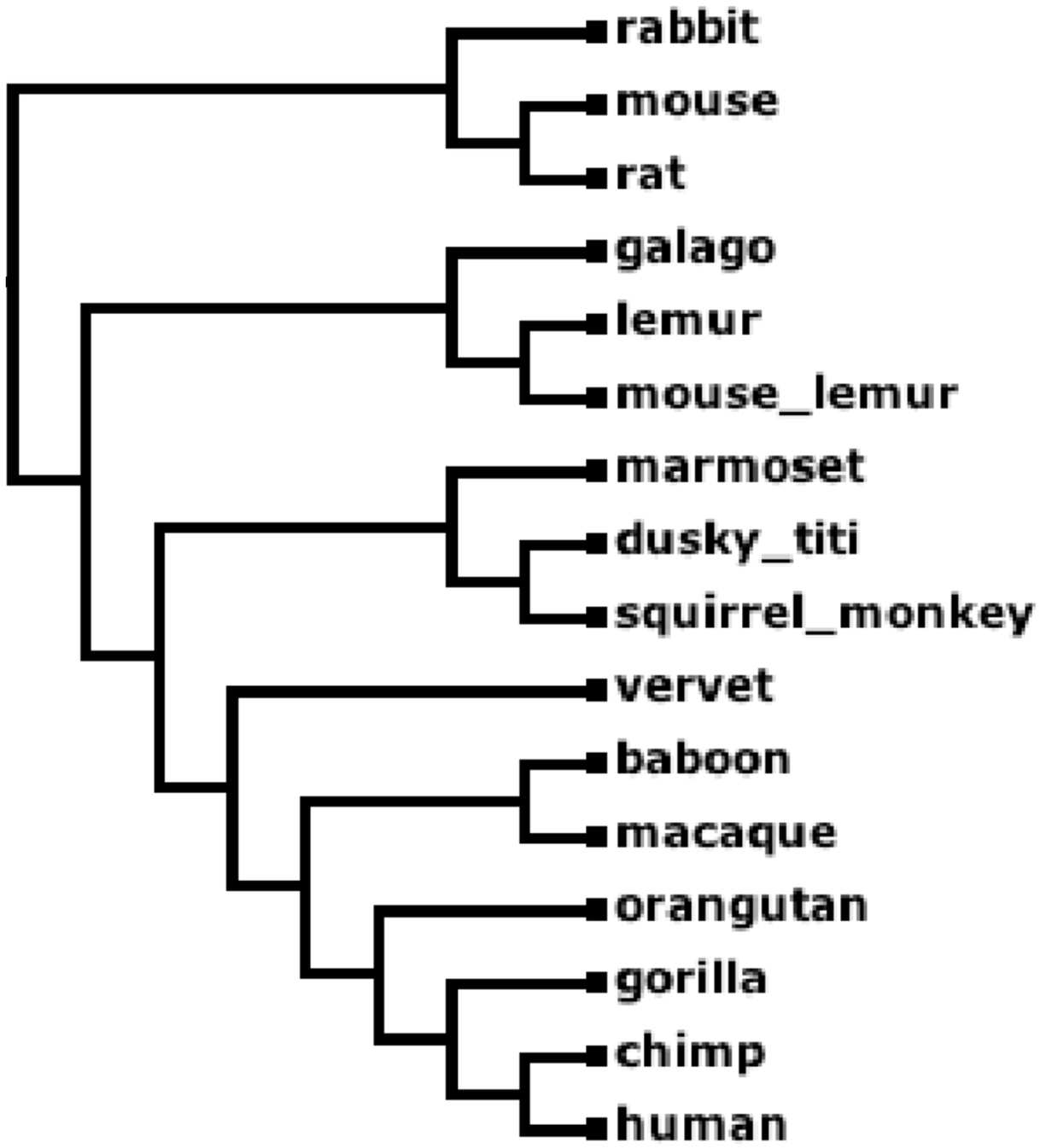


'Rej Subs':

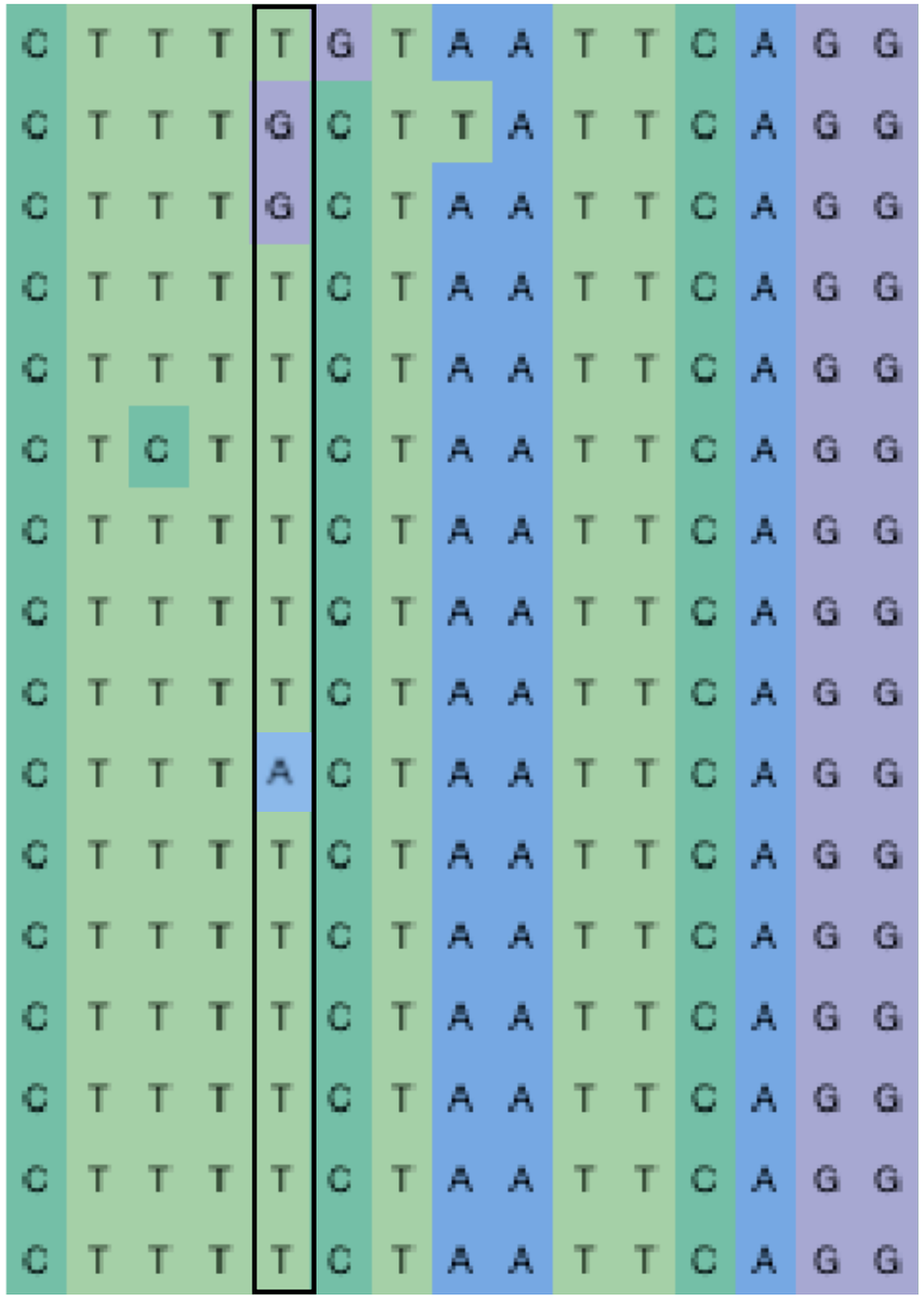
+5

Cooper *et al* 2005 *Genome Research*

Genome Sequence Comparisons



Neutral Average
~5 subs/site

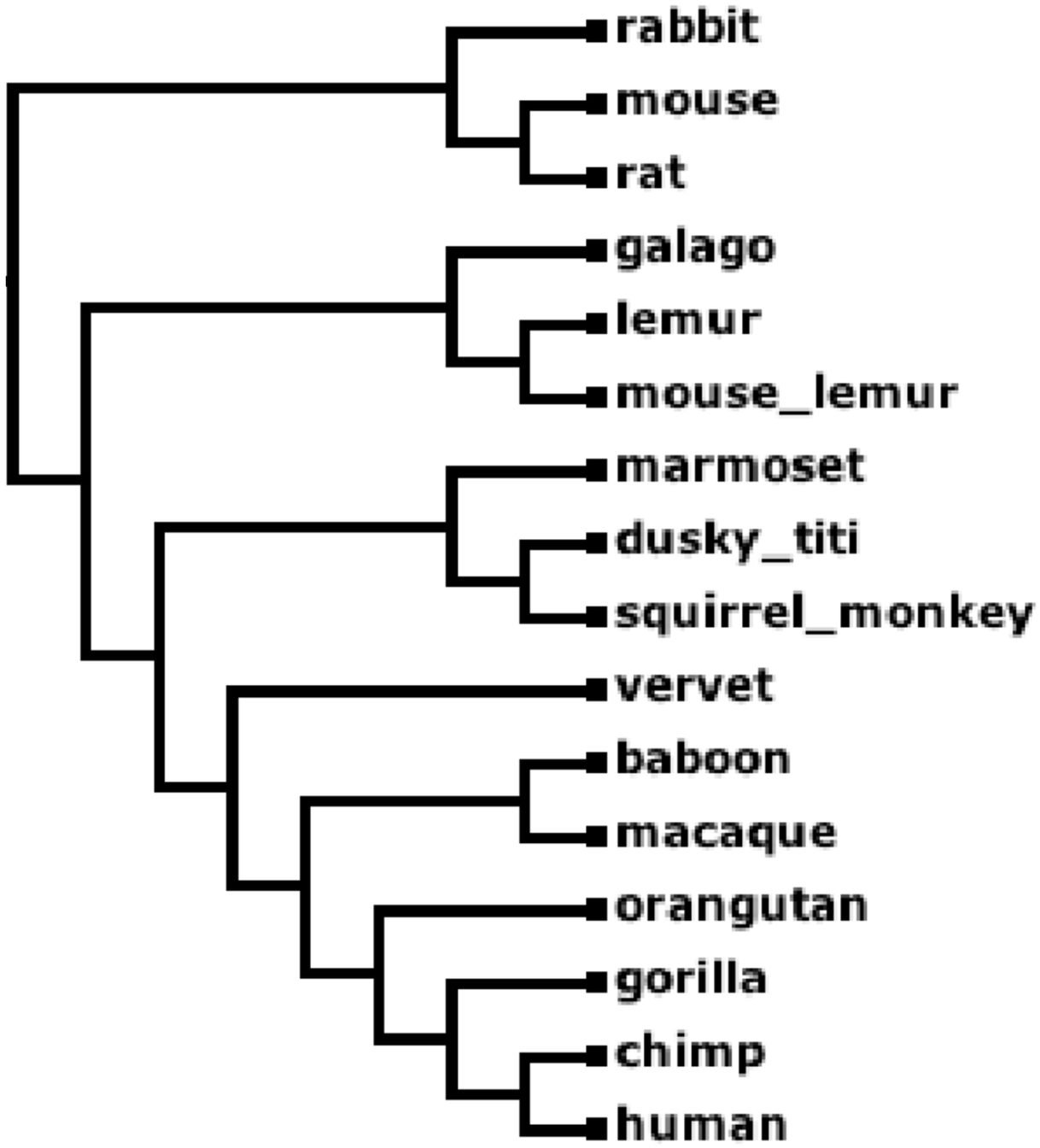


'Rej Subs':

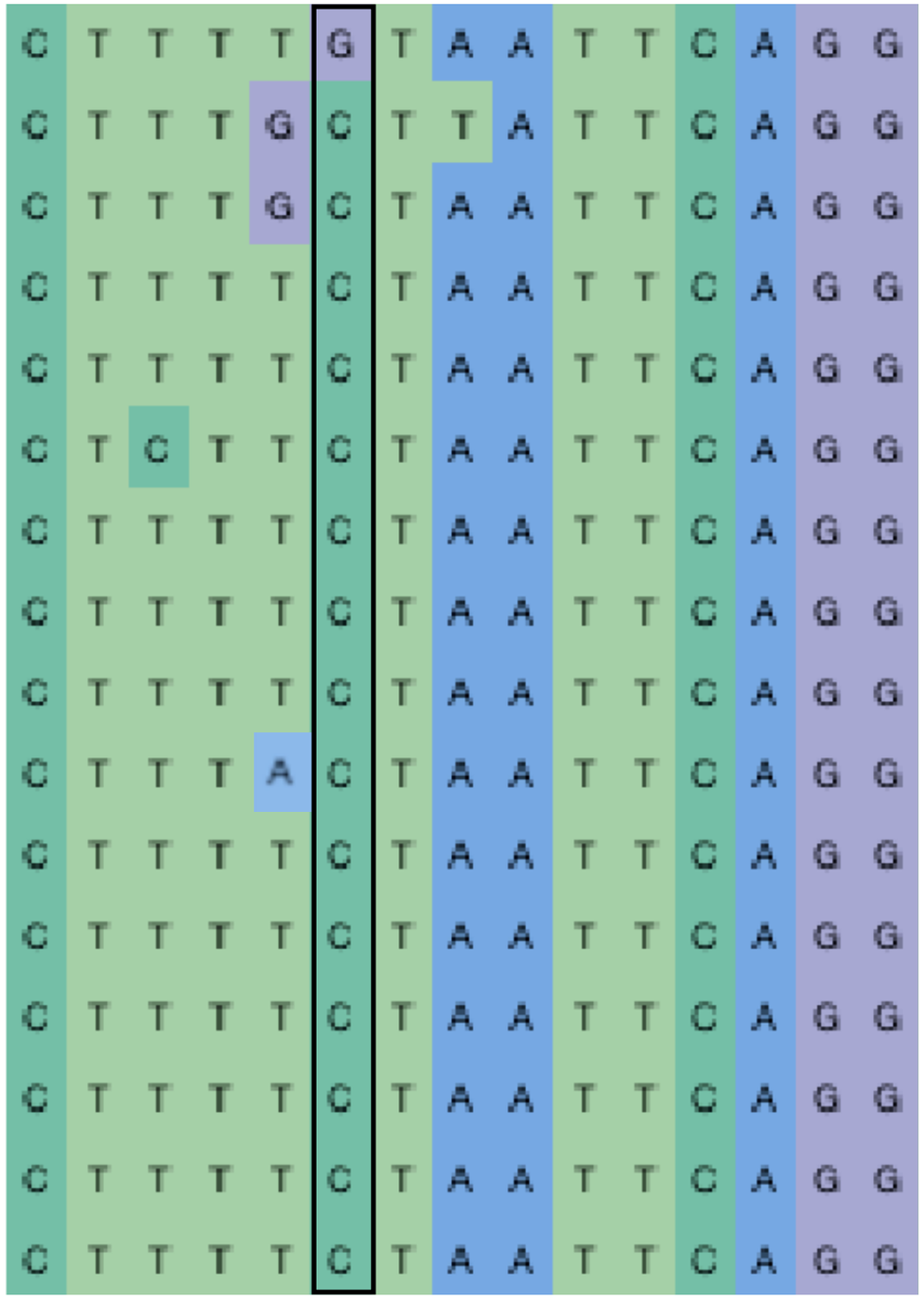
+3

Cooper *et al* 2005 *Genome Research*

Genome Sequence Comparisons



Neutral Average
~5 subs/site

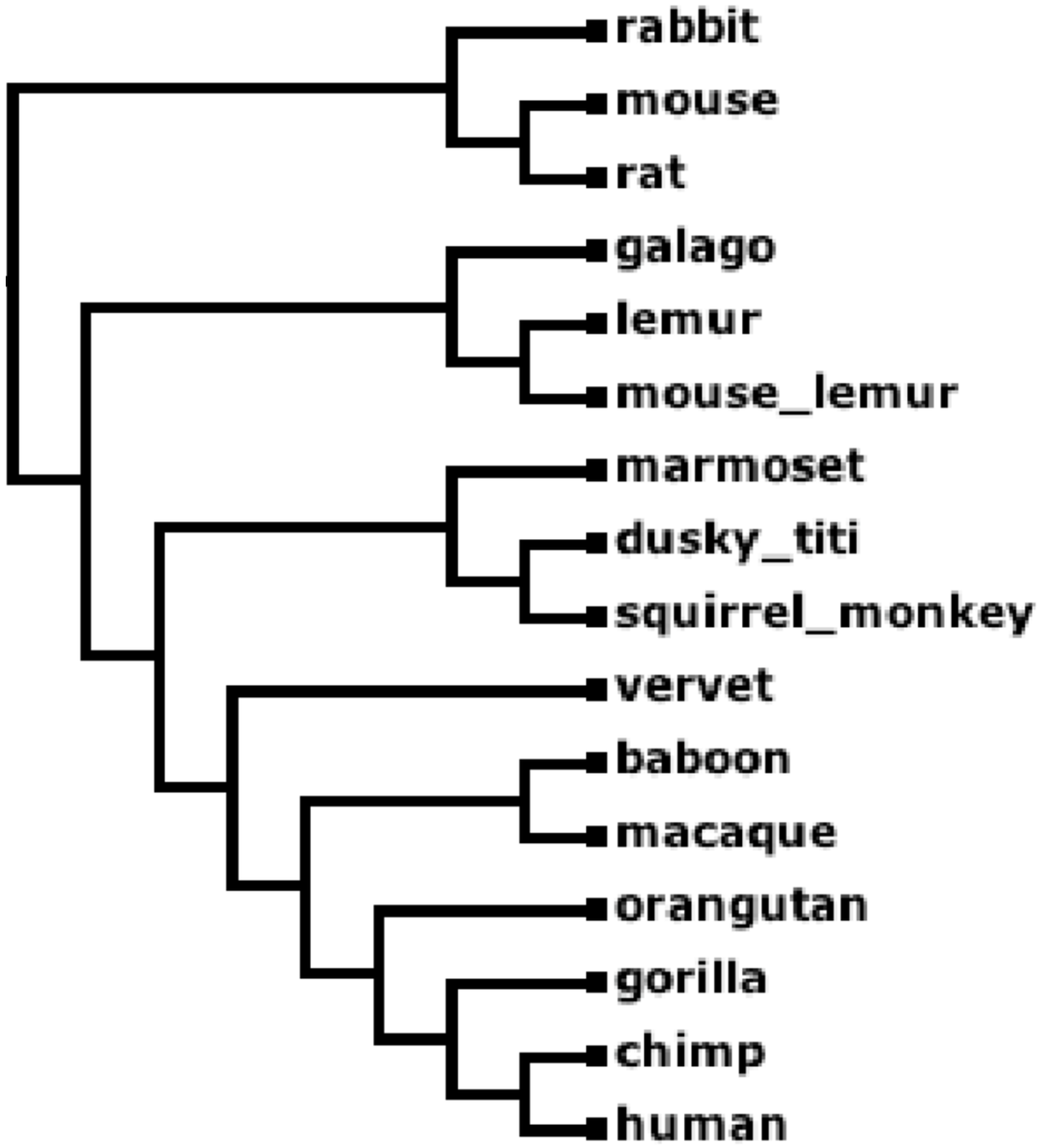


'Rej Subs':

+4

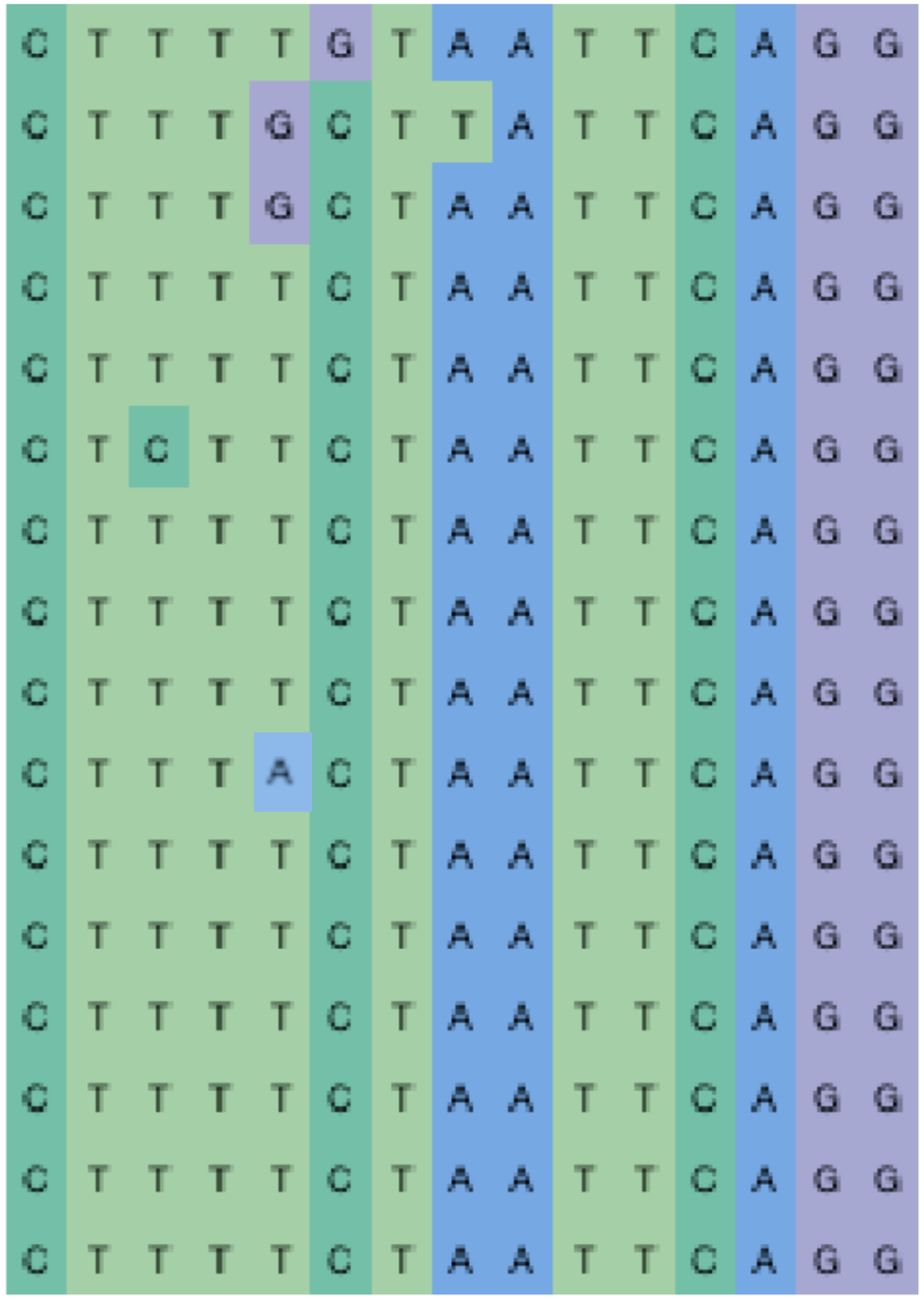
Cooper *et al* 2005 *Genome Research*

Genome Sequence Comparisons



Neutral Average
~5 subs/site

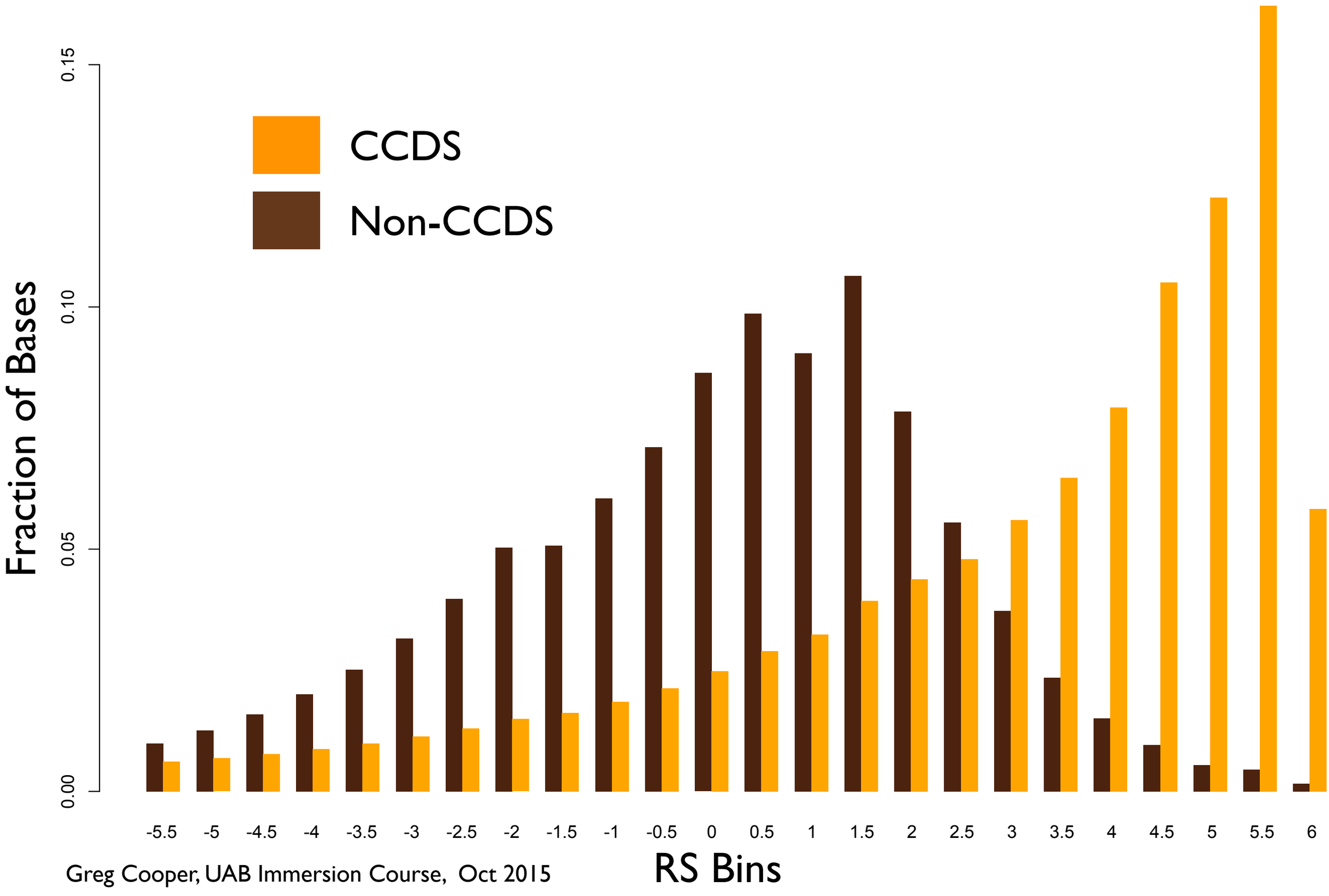
‘Rej Subs’:



...

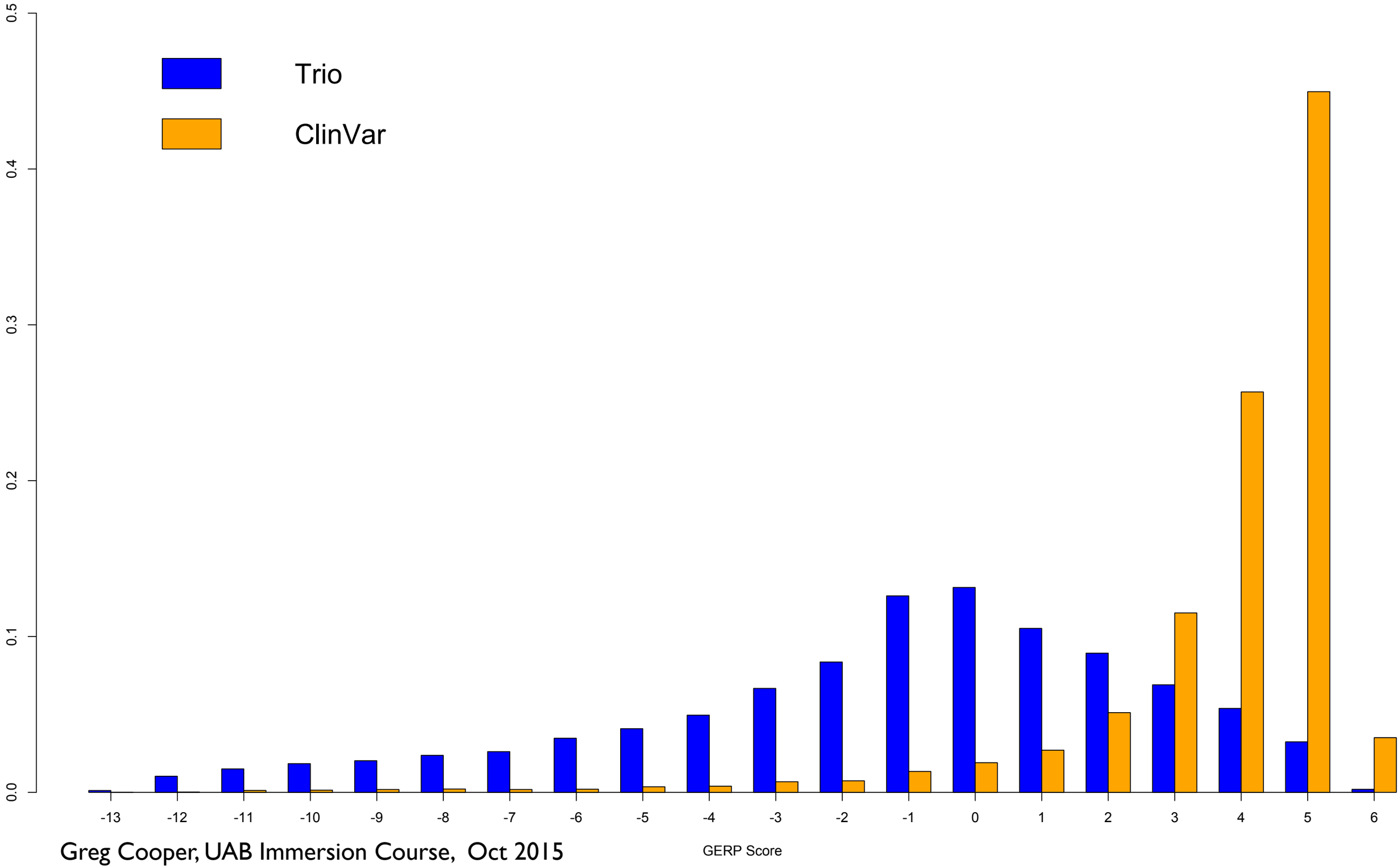
Cooper *et al* 2005 *Genome Research*

Exome vs Genome



Greg Cooper, UAB Immersion Course, Oct 2015

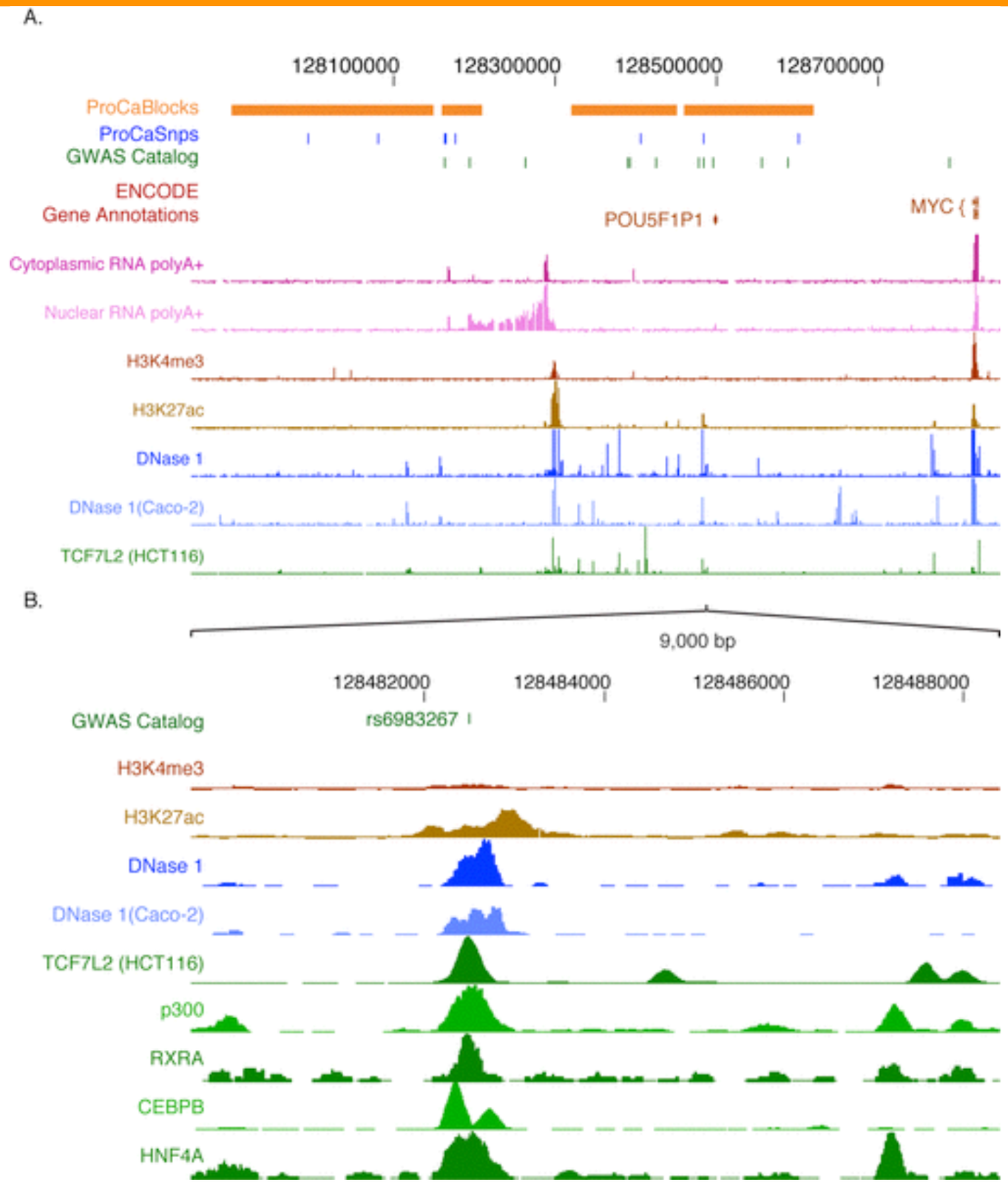
Sequence Conservation of Known Pathogenic Mutations



Nucleotide-Level Functional Annotations

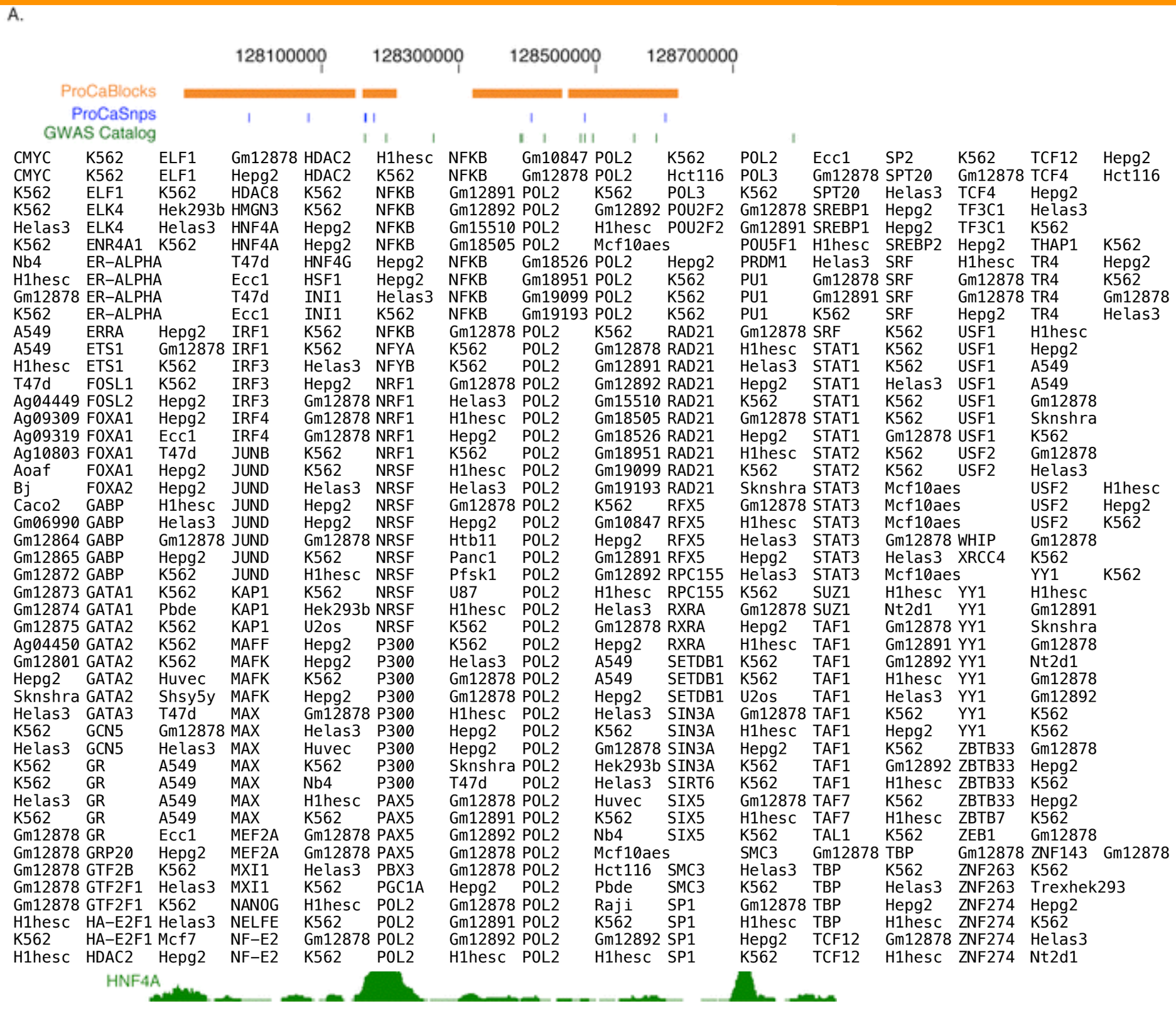
- Functional analysis of non-coding elements has historically received lesser focus than proteins
- Rapidly improving for humans and model organisms
- Key types of available information:
  - Transcription-factor binding events via ChIP-seq
  - Regions of open chromatin via DNase HS/FAIRE/others
  - Chromatin modifications
  - Methylation status
  - Identification of promoters, enhancers, silencers, and insulators via integrative analysis of molecular annotations

Nucleotide-Level Functional Annotation



The ENCODE Project Consortium 2011. *PLoS Biology*

Nucleotide-Level Functional Annotation



The ENCODE Project Consortium 2011. *PLoS Biology*

Annotation Relevance to Disease

•Functional and evolutionary information are useful to predict disease relevance, but:

Annotation Relevance to Disease

- Functional and evolutionary information are useful to predict disease relevance, but:
  - What is the relative value of evolutionary vs functional information?

Annotation Relevance to Disease

- Functional and evolutionary information are useful to predict disease relevance, but:
  - What is the relative value of evolutionary vs functional information?
  - What is the relative importance of various functional categories?

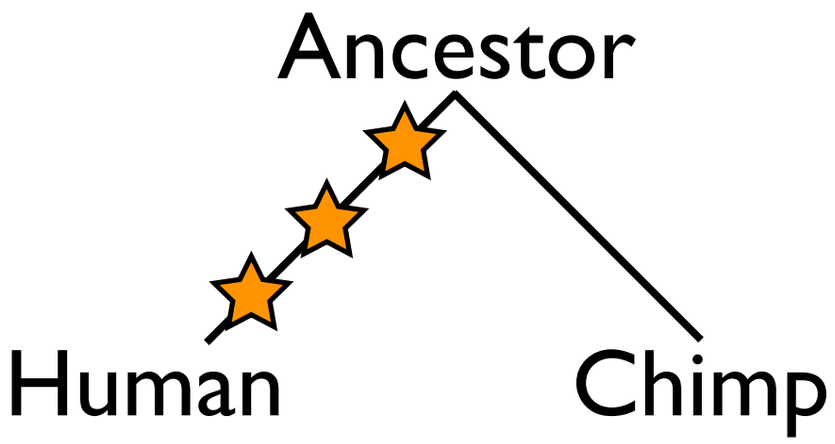
Annotation Relevance to Disease

- Functional and evolutionary information are useful to predict disease relevance, but:
  - What is the relative value of evolutionary vs functional information?
  - What is the relative importance of various functional categories?
  - What is the value of any specific combination of annotations?

Annotation Relevance to Disease

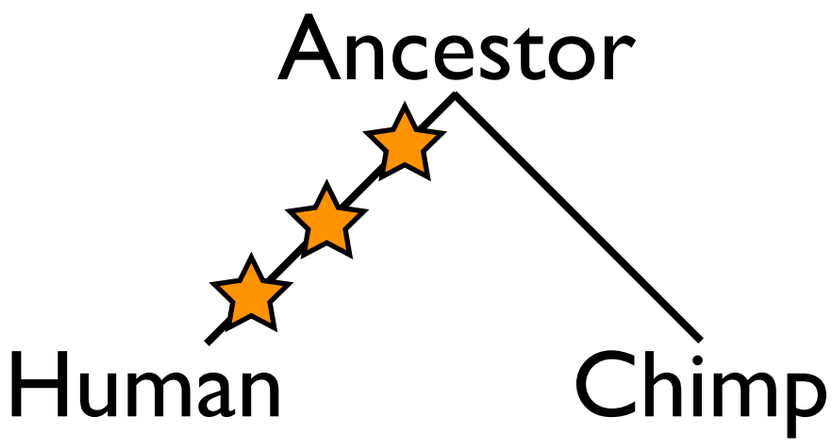
- Functional and evolutionary information are useful to predict disease relevance, but:
  - What is the relative value of evolutionary vs functional information?
  - What is the relative importance of various functional categories?
  - What is the value of any specific combination of annotations?
  - How does one cope with the thousands of partially correlated genomic annotations that are available (e.g. ENCODE)?

Annotation Dependent Depletion



Vs Simulated Mutations

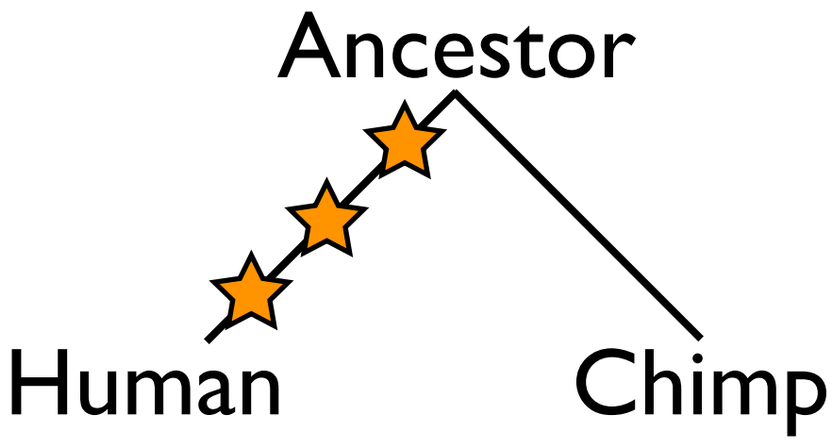
Annotation Dependent Depletion



Vs Simulated Mutations

Annotation	Number Observed	Number Simulated	Observed/ Expected

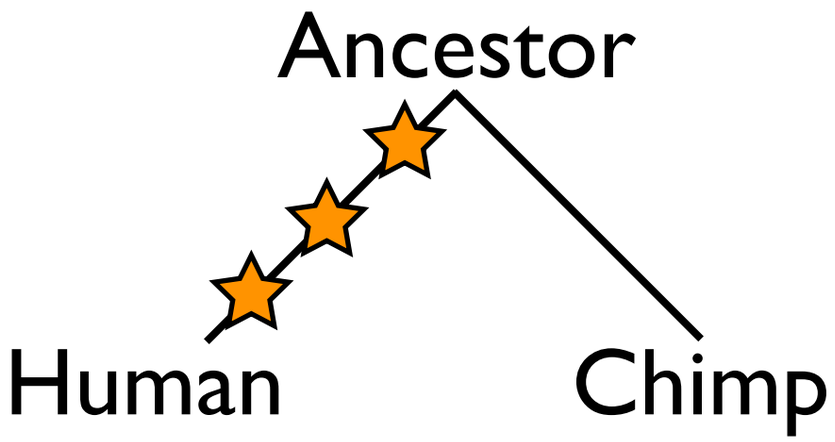
Annotation Dependent Depletion



Vs Simulated Mutations

Annotation	Number Observed	Number Simulated	Observed/ Expected
Stop	183	6,749	0.027

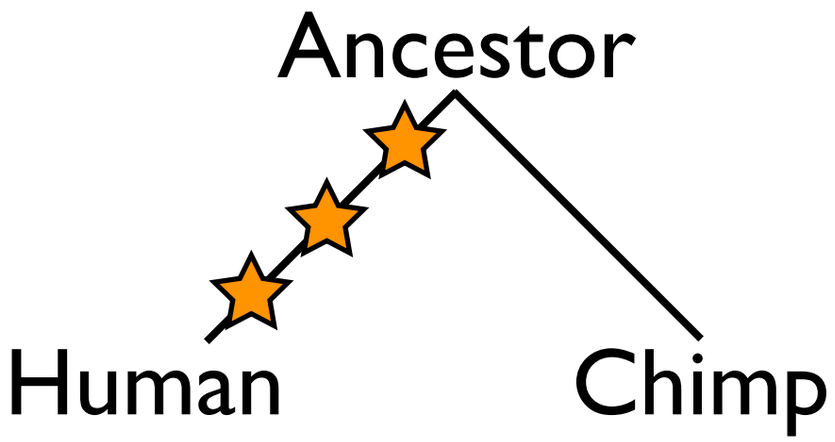
Annotation Dependent Depletion



Vs Simulated Mutations

Annotation	Number Observed	Number Simulated	Observed/ Expected
Stop	183	6,749	0.027
Regulatory	1,142,020	1,291,684	0.884

Annotation Dependent Depletion



Vs Simulated Mutations

Annotation	Number Observed	Number Simulated	Observed/ Expected
Stop	183	6,749	0.027
Regulatory	1,142,020	1,291,684	0.884
…	…	…	“ADD"

Combined Annotation Dependent Depletion (CADD)

1. Generate 63 distinct annotations, including conservation scores, ENCODE data (summarized at various levels), gene body annotations, etc.

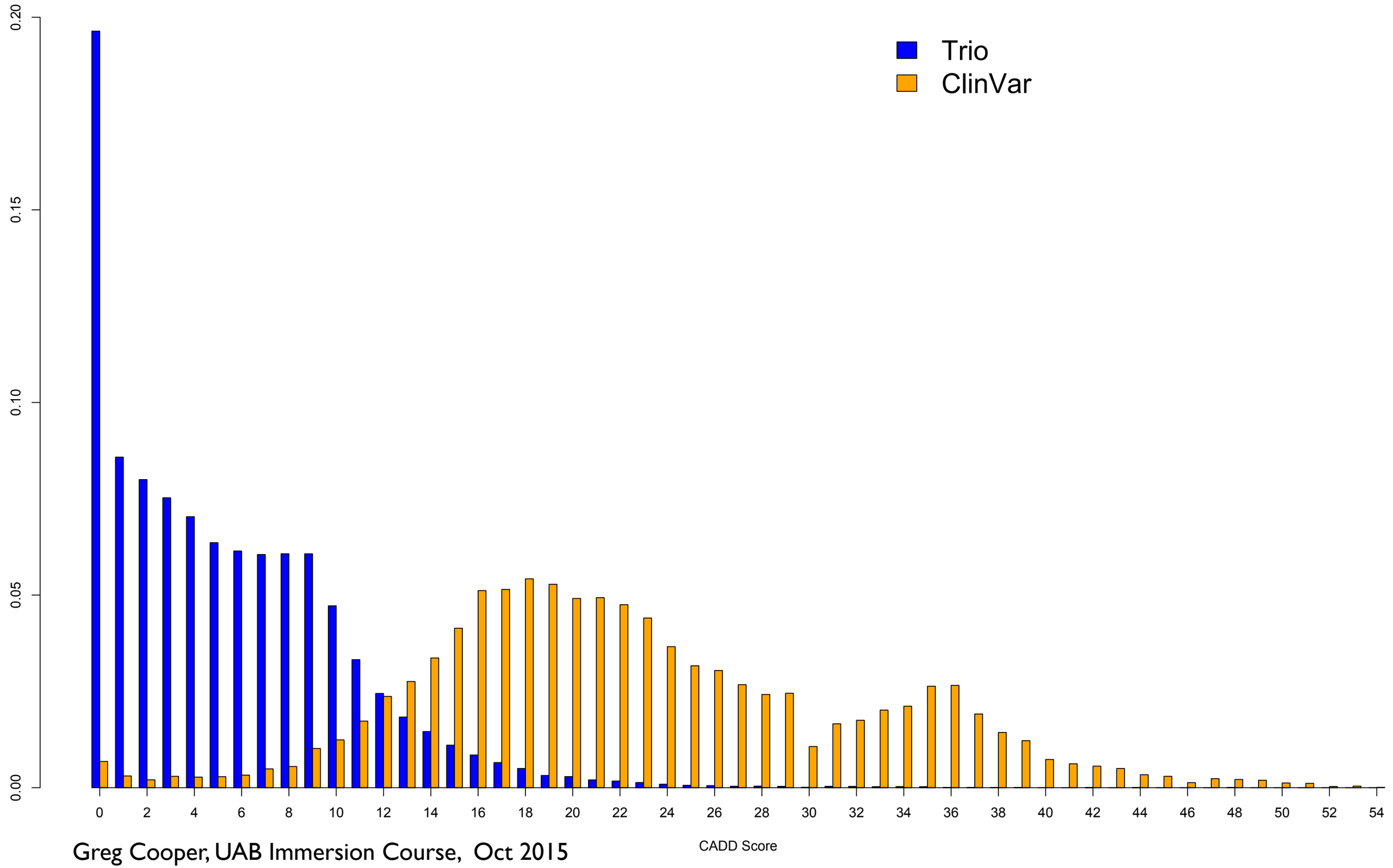
2. Build a support vector machine (SVM) that estimates, for a given variant, whether it is likely to be observed or simulated, based on its combined annotation profile

    Domain-specific information (e.g. polyPhen) via "missing" indicators

    Some interaction terms, like STOP * relCDSpos, also included

3. Score all possible ~8.6 billion possible SNVs of hg19

CADD Scores for Known Pathogenic Mutations



Greg Cooper, UAB Immersion Course, Oct 2015

"Brute Force" Resolution of Mendelian Disease

- In 2009, Ng et al (*Nature*) sequenced exomes from:

"Brute Force" Resolution of Mendelian Disease

- In 2009, Ng et al (*Nature*) sequenced exomes from:
  - 4 individuals from separate families with a rare disease, Freeman-Sheldon syndrome (FSS)

"Brute Force" Resolution of Mendelian Disease

- In 2009, Ng et al (*Nature*) sequenced exomes from:
  - 4 individuals from separate families with a rare disease, Freeman-Sheldon syndrome (FSS)
  - 8 individuals with diverse ancestry (HapMap samples)

"Brute Force" Resolution of Mendelian Disease

- In 2009, Ng et al (*Nature*) sequenced exomes from:
  - 4 individuals from separate families with a rare disease, Freeman-Sheldon syndrome (FSS)
  - 8 individuals with diverse ancestry (HapMap samples)
- Then looked for genes in which:

"Brute Force" Resolution of Mendelian Disease

- In 2009, Ng et al (*Nature*) sequenced exomes from:
  - 4 individuals from separate families with a rare disease, Freeman-Sheldon syndrome (FSS)
  - 8 individuals with diverse ancestry (HapMap samples)
- Then looked for genes in which:
  - 1 or more FSS-affected individuals harbored a "rare" variant, i.e. was not present in dbSNP or present in one of the 8 HapMap samples, AND…

"Brute Force" Resolution of Mendelian Disease

Number of genes in which each affected has at least one...	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4: FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768

Ng *et al* 2009 *Nature*

"Brute Force" Resolution of Mendelian Disease

Heterogenous to Single Gene

Number of genes in which each affected has at least one...	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4: FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768

Ng *et al* 2009 *Nature*

"Brute Force" Resolution of Mendelian Disease

Heterogenous to Single Gene

Function and Frequency Filters

Number of genes in which each affected has at least one...	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4: FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768

Ng *et al* 2009 *Nature*

"Brute Force" Resolution of Mendelian Disease

Heterogenous to Single Gene

Function and Frequency Filters

Number of genes in which each affected has at least one...	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4 FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119

Ng *et al* 2009 *Nature*

"Brute Force" Resolution of Mendelian Disease

Heterogenous to Single Gene

Function and Frequency Filters

Number of genes in which each affected has at least one…	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4 FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160

Ng *et al* 2009 *Nature*

"Brute Force" Resolution of Mendelian Disease

Heterogenous to Single Gene

Function and Frequency Filters

Number of genes in which each affected has at least one…	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4: FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (*MYH3*)	22

Ng *et al* 2009 *Nature*

"Brute Force" Resolution of Mendelian Disease

Heterogenous to Single Gene

Function and Frequency Filters

Number of genes in which each affected has at least one…	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4 FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (*MYH3*)	22

~30-50% of suspected Mendelian diseases can be traced to causal genes with just a handful of exomes!

Ng *et al* 2009 *Nature*

Variant Annotations and Disease Prediction

Number of genes in which each affected has at least one...	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4: FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (*MYH3*)	22
...And predicted to be damaging	160	10	2	1 (*MYH3*)	3

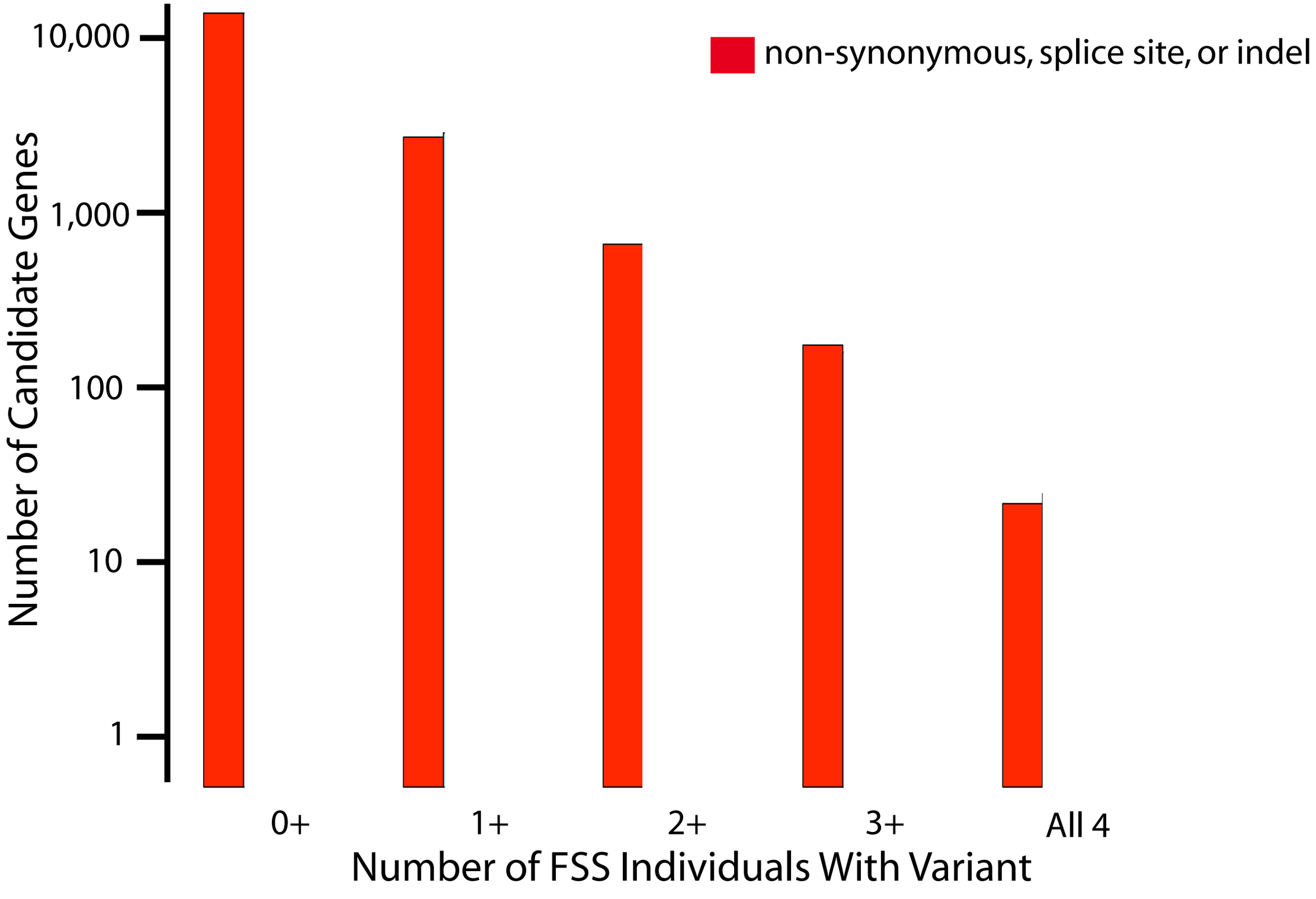
Ng *et al* 2009 *Nature*, Cooper *et al* 2010 *Nature Methods*

Variant Annotations and Disease Prediction

Number of genes in which each affected has at least one...	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	Any 3 of 4: FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (*MYH3*)	22
...And predicted to be damaging	160	10	2	1 (*MYH3*)	3

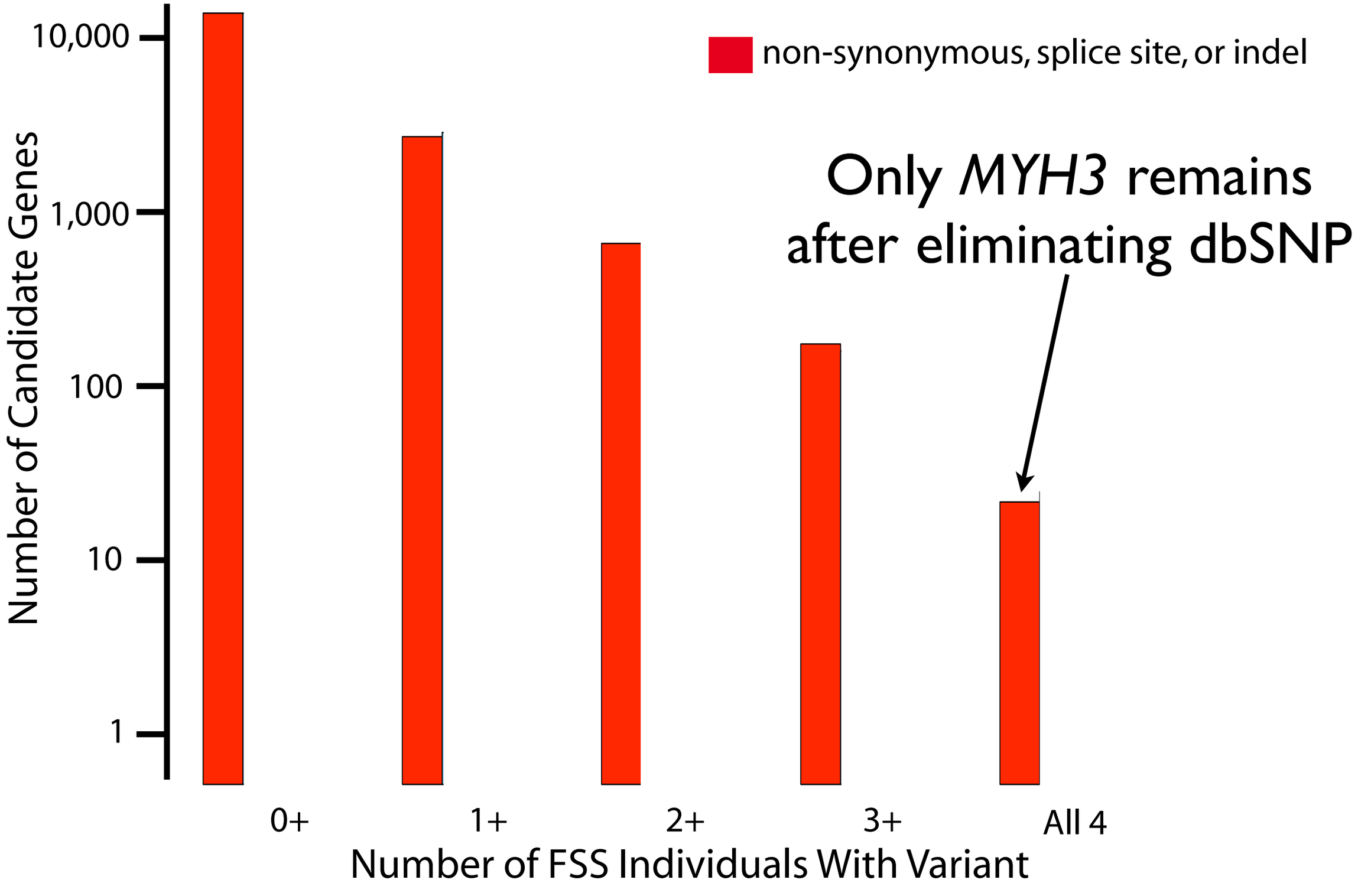
Ng *et al* 2009 *Nature*, Cooper *et al* 2010 *Nature Methods*

Variant Annotations and Disease Prediction



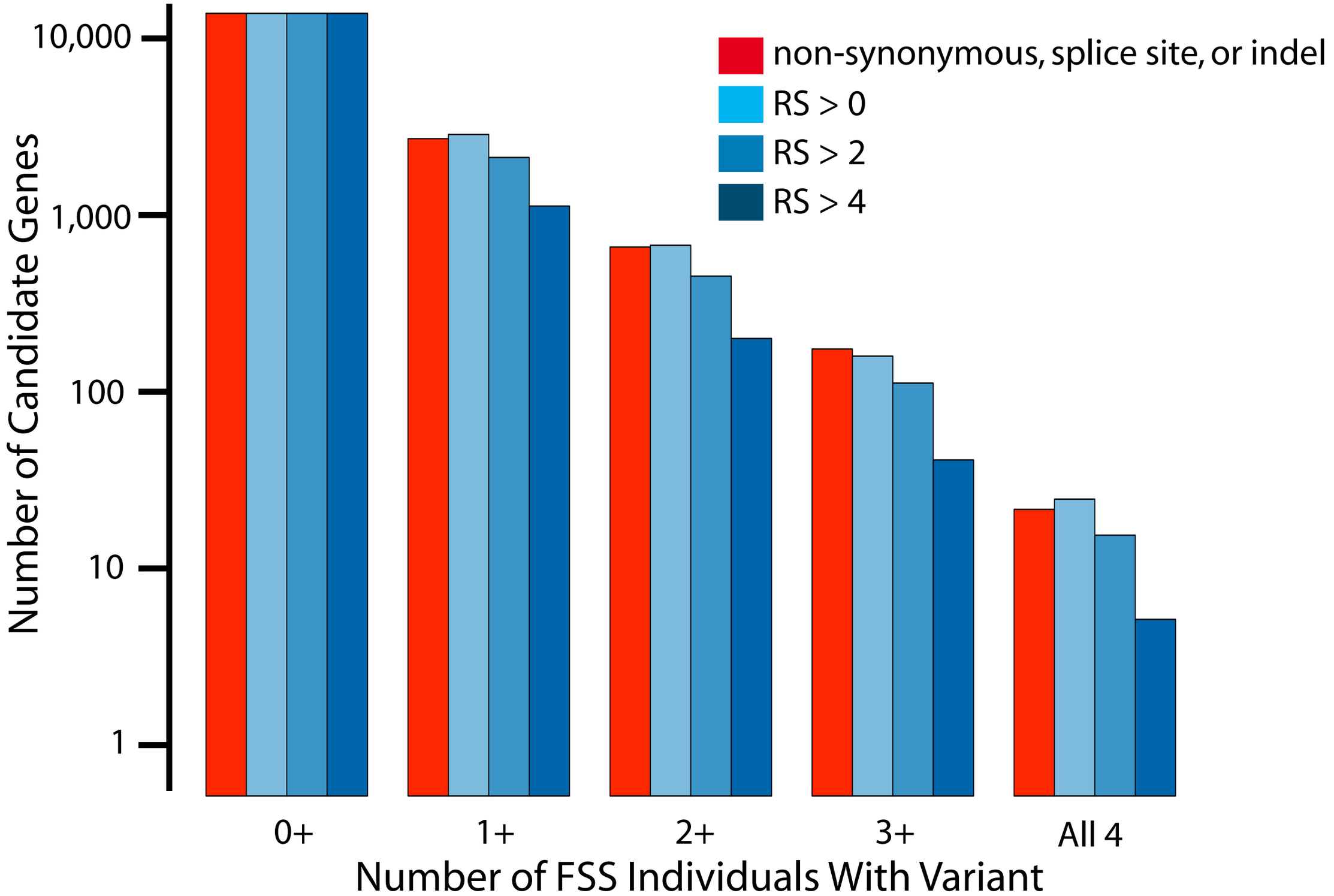
Ng *et al* 2009 *Nature*, Cooper *et al* 2010 *Nature Methods*

Variant Annotations and Disease Prediction



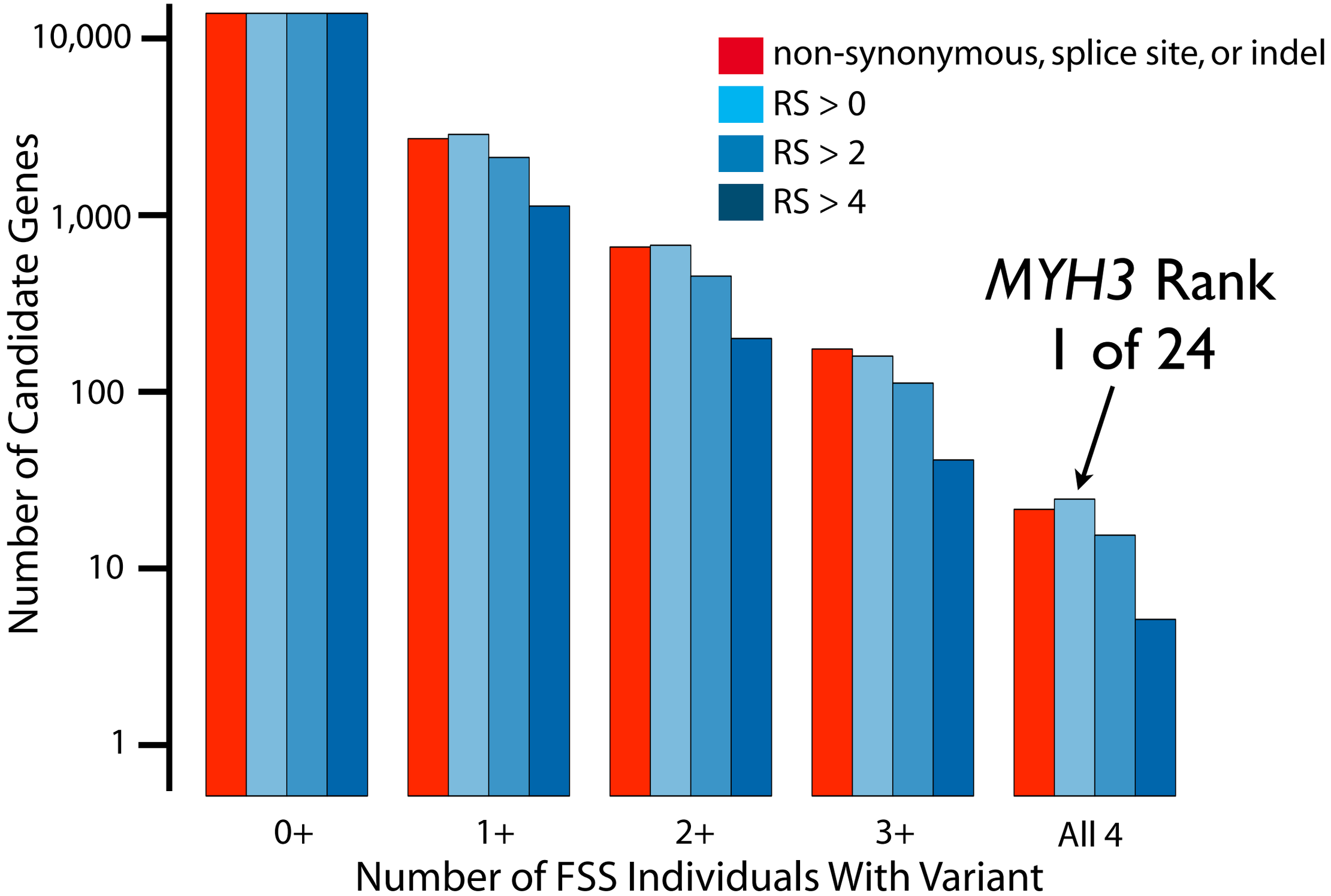
Ng *et al* 2009 *Nature*, Cooper *et al* 2010 *Nature Methods*

Variant Annotations and Disease Prediction



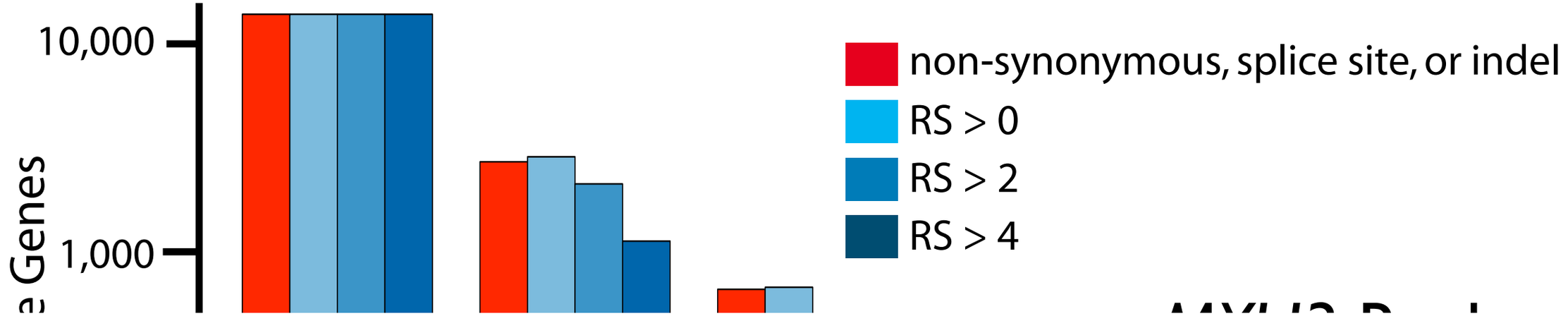
Ng *et al* 2009 *Nature*, Cooper *et al* 2010 *Nature Methods*

Variant Annotations and Disease Prediction

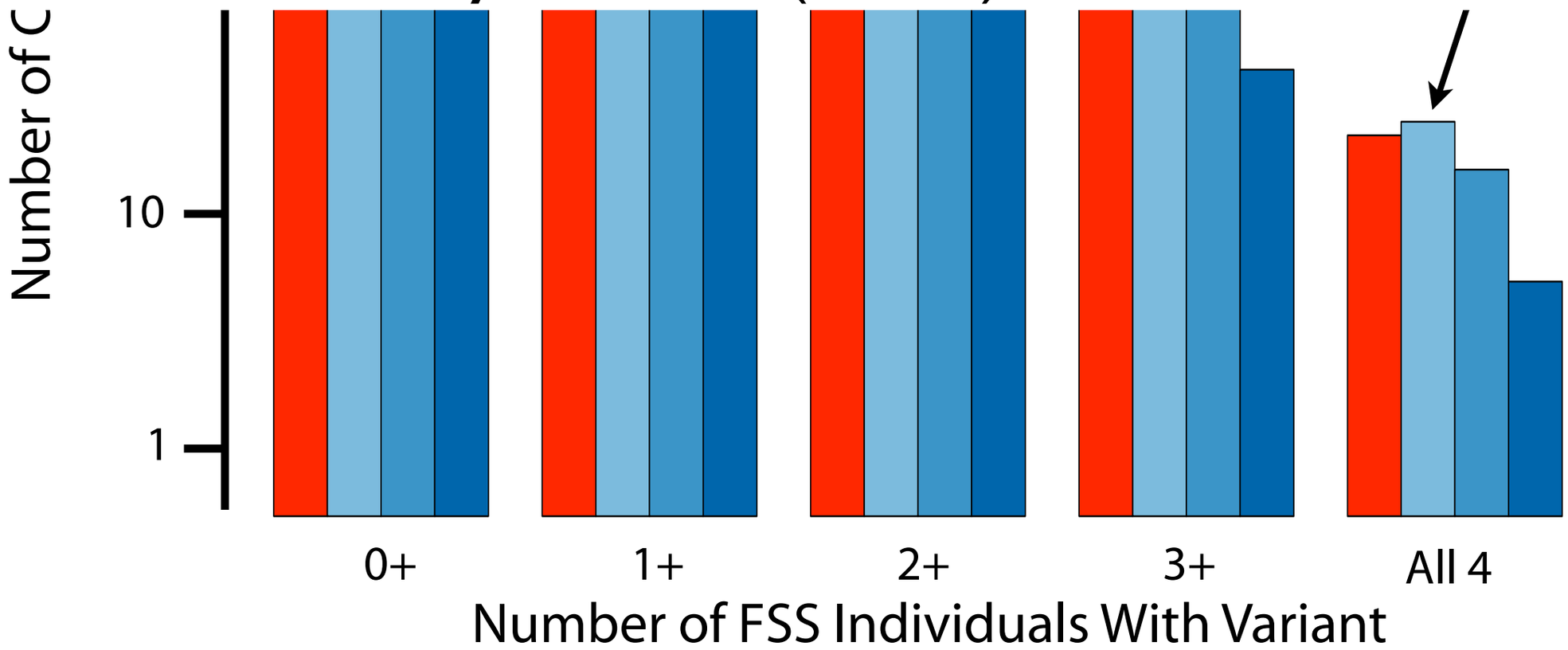


Ng *et al* 2009 *Nature*, Cooper *et al* 2010 *Nature Methods*

Variant Annotations and Disease Prediction



Similar results for Miller syndrome (*DHODH*), Kabuki syndrome (*MLL2*), others



Ng *et al* 2009 *Nature*, Cooper *et al* 2010 *Nature Methods*

Error Rates in Genomic Data

All genomics datasets harbor both random and structured errors:

- Many sources of error, *e.g.* depth of coverage, chemistry issues, PCR mistakes, software limitations, etc.

- All institutions, machines, technicians, protocols, samples, etc introduce varying degrees and types of subtle yet systematic and highly "significant" sources of error

- Even when errors are rare at large, they can be proportionally enriched in subsets of variants

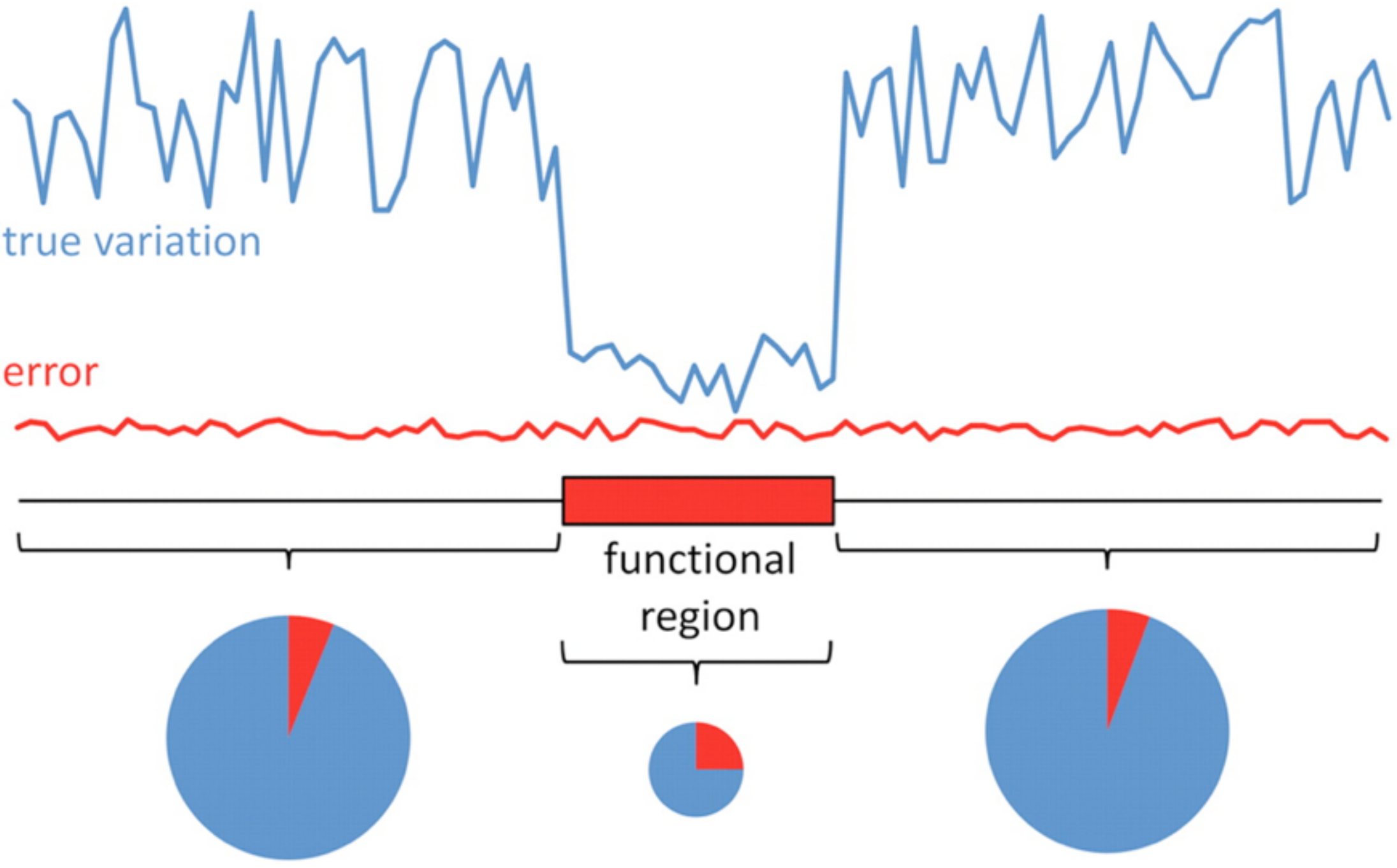
Functional Variants are Proportionally Error-Enriched

Functional Variants are Proportionally Error-Enriched

Irony of variant annotation is that the more "interesting" a variant is, the more likely it is to be the result of an error

false positive rate ≠ false discovery rate

Functional Variants are Proportionally Error-Enriched



Greg Cooper, UAB Immersion Course, Oct 2015

MacArthur and Tyler-Smith 2010 *Hum Mol Gen*

De Novo Mutations and Errors

Greg Cooper, UAB Immersion Course, Oct 2015

De Novo Mutations and Errors

Important example is searching for *de novo* non-synonymous variants by sequencing proband-parent trios. Assuming:

De Novo Mutations and Errors

Important example is searching for *de novo* non-synonymous variants by sequencing proband-parent trios. Assuming:

- Mutation rate of 2.5×10^{-8}

De Novo Mutations and Errors

Important example is searching for *de novo* non-synonymous variants by sequencing proband-parent trios. Assuming:

- Mutation rate of 2.5×10^{-8}
- 20 Mbp of exome prone to non-synonymous variants:

De Novo Mutations and Errors

Important example is searching for *de novo* non-synonymous variants by sequencing proband-parent trios. Assuming:

- Mutation rate of 2.5×10^{-8}
- 20 Mbp of exome prone to non-synonymous variants:
- Sequencing false positive rate (false heterozygote) of 1×10^{-6} (note this is a specificity of 99.9999%)

De Novo Mutations and Errors

Important example is searching for *de novo* non-synonymous variants by sequencing proband-parent trios. Assuming:

- Mutation rate of 2.5×10^{-8}
- 20 Mbp of exome prone to non-synonymous variants:
- Sequencing false positive rate (false heterozygote) of 1×10^{-6} (note this is a specificity of 99.9999%)

Then we expect:

De Novo Mutations and Errors

Important example is searching for *de novo* non-synonymous variants by sequencing proband-parent trios. Assuming:

- Mutation rate of 2.5×10^{-8}
- 20 Mbp of exome prone to non-synonymous variants:
- Sequencing false positive rate (false heterozygote) of 1×10^{-6} (note this is a specificity of 99.9999%)

Then we expect:

- ~0.5 actual *de novo* non-synonymous variants per proband, and 20 false positives, *i.e.* FDR = 97.6%

De Novo Mutations and Errors

Important example is searching for *de novo* non-synonymous variants by sequencing proband-parent trios. Assuming:

- Mutation rate of 2.5×10^{-8}
- 20 Mbp of exome prone to non-synonymous variants:
- Sequencing false positive rate (false heterozygote) of 1×10^{-6} (note this is a specificity of 99.9999%)

Then we expect:

- ~0.5 actual *de novo* non-synonymous variants per proband, and 20 false positives, *i.e.* FDR = 97.6%
- Excludes false negative variants in parents

De Novo Mutations and Errors

Table 1: Overview of all variants detected per proband and impact of the prioritization steps for selecting candidate non-synonymous *de novo* mutations

Trio	1	2	3	4	5	6	7	8	9	10	Average
High-confidence variant calls	20,810	21,658	21,338	22,647	17,694	22,333	21,369	22,658	24,085	22,962	21,755
After exclusion of nongenic, intronic and synonymous variants	5,556	5,665	5,691	5,991	4,607	5,567	5,716	5,628	5,985	5,994	5,640
After exclusion of known variants	165	159	157	155	120	136	120	149	96	171	143
After exclusion of inherited variants	4	7	3	7	7	2	2	6	6	7	5

Vissers *et al* 2010 *Nature Genetics*

De Novo Mutations and Errors

Table 1: Overview of all variants detected per proband and impact of the prioritization steps for selecting candidate non-synonymous *de novo* mutations

Trio	1	2	3	4	5	6	7	8	9	10	Average
High-confidence variant calls	20,810	21,658	21,338	22,647	17,694	22,333	21,369	22,658	24,085	22,962	21,755
After exclusion of nongenic, intronic and synonymous variants	5,556	5,665	5,691	5,991	4,607	5,567	5,716	5,628	5,985	5,994	5,640
After exclusion of known variants	165	159	157	155	120	136	120	149	96	171	143
After exclusion of inherited variants	4	7	3	7	7	2	2	6	6	7	5

- After aggressive QC/false positive control, 51 candidate de novo variants in 10 probands
- 9 of which validated
- FDR = 82%
- In genomes, typically observe many hundreds of plausible candidates, of which <100 are real

Vissers *et al* 2010 *Nature Genetics*

Annotation Summary

- Annotations are a powerful source of information for:

Annotation Summary

- Annotations are a powerful source of information for:
  - QC of individual variants and whole datasets

Annotation Summary

- Annotations are a powerful source of information for:
  - QC of individual variants and whole datasets
  - Increased discovery power in genetic studies of many types

Annotation Summary

- Annotations are a powerful source of information for:
  - QC of individual variants and whole datasets
  - Increased discovery power in genetic studies of many types
  - Mechanistic hypotheses

Annotation Summary

- Annotations are a powerful source of information for:
  - QC of individual variants and whole datasets
  - Increased discovery power in genetic studies of many types
  - Mechanistic hypotheses
- Annotations are neither necessary nor sufficient for causality

Annotation Summary

- Annotations are a powerful source of information for:
  - QC of individual variants and whole datasets
  - Increased discovery power in genetic studies of many types
  - Mechanistic hypotheses
- Annotations are neither necessary nor sufficient for causality
  - Genetic information always paramount (i.e., consistent phenotypes observed among mutation carriers at a statistically non-random level)

Annotation Summary

- Annotations are a powerful source of information for:
  - QC of individual variants and whole datasets
  - Increased discovery power in genetic studies of many types
  - Mechanistic hypotheses
- Annotations are neither necessary nor sufficient for causality
  - Genetic information always paramount (i.e., consistent phenotypes observed among mutation carriers at a statistically non-random level)
  - Biologically rare annotations are ALWAYS enriched for errors, sometimes dramatically so

Annotation Summary

- Annotations are a powerful source of information for:
  - QC of individual variants and whole datasets
  - Increased discovery power in genetic studies of many types
  - Mechanistic hypotheses
- Annotations are neither necessary nor sufficient for causality
  - Genetic information always paramount (i.e., consistent phenotypes observed among mutation carriers at a statistically non-random level)
  - Biologically rare annotations are ALWAYS enriched for errors, sometimes dramatically so
  - Rigorous statistical analysis is important, but manual review of both genetic and non-genetic information is a critical step in evaluating mutations in both clinical and research projects

Greg Cooper, UAB Immersion Course, Oct 2015

Other General Comments on Finding Causal Variants

- Discrete filters are currently standard but sub-optimal

Other General Comments on Finding Causal Variants

- Discrete filters are currently standard but sub-optimal
  - for example, don't simply filter by presence in dbSNP; allele frequency and its implications for penetrance and disease prevalence is the pertinent bit of data

Other General Comments on Finding Causal Variants

- Discrete filters are currently standard but sub-optimal

  - for example, don't simply filter by presence in dbSNP; allele frequency and its implications for penetrance and disease prevalence is the pertinent bit of data

  - "conserved" vs "neutral" ("damaging" vs "not damaging", *etc.*) a dichotomy that ignores valuable, quantitative information

Other General Comments on Finding Causal Variants

- Discrete filters are currently standard but sub-optimal
  - for example, don't simply filter by presence in dbSNP; allele frequency and its implications for penetrance and disease prevalence is the pertinent bit of data
  - "conserved" vs "neutral" ("damaging" vs "not damaging", *etc.*) a dichotomy that ignores valuable, quantitative information
- Most causal variants, even for rare and monogenic disease, are unlikely to be non-synonymous or obvious protein LOF variants

Other General Comments on Finding Causal Variants

- Discrete filters are currently standard but sub-optimal
  - for example, don't simply filter by presence in dbSNP; allele frequency and its implications for penetrance and disease prevalence is the pertinent bit of data
  - "conserved" vs "neutral" ("damaging" vs "not damaging", *etc.*) a dichotomy that ignores valuable, quantitative information
- Most causal variants, even for rare and monogenic disease, are unlikely to be non-synonymous or obvious protein LOF variants
  - Less than 50% of rare, likely Mendelian diseases to which exome sequencing has been applied has yielded good hits
  - Even when a gene is known for a disease, protein-coding alterations often explain only a subset of cases

Greg Cooper, UAB Immersion Course, Oct 2015

Developmental Delay and Intellectual Disability

1 - 2% of kids are born with one or more of:

- intellectual disability
- developmental delay
- heart defects
- craniofacial and skeletal abnormalities
- severe autism
- seizures

Greg Cooper, UAB Immersion Course, Oct 2015

Developmental Delay and Intellectual Disability

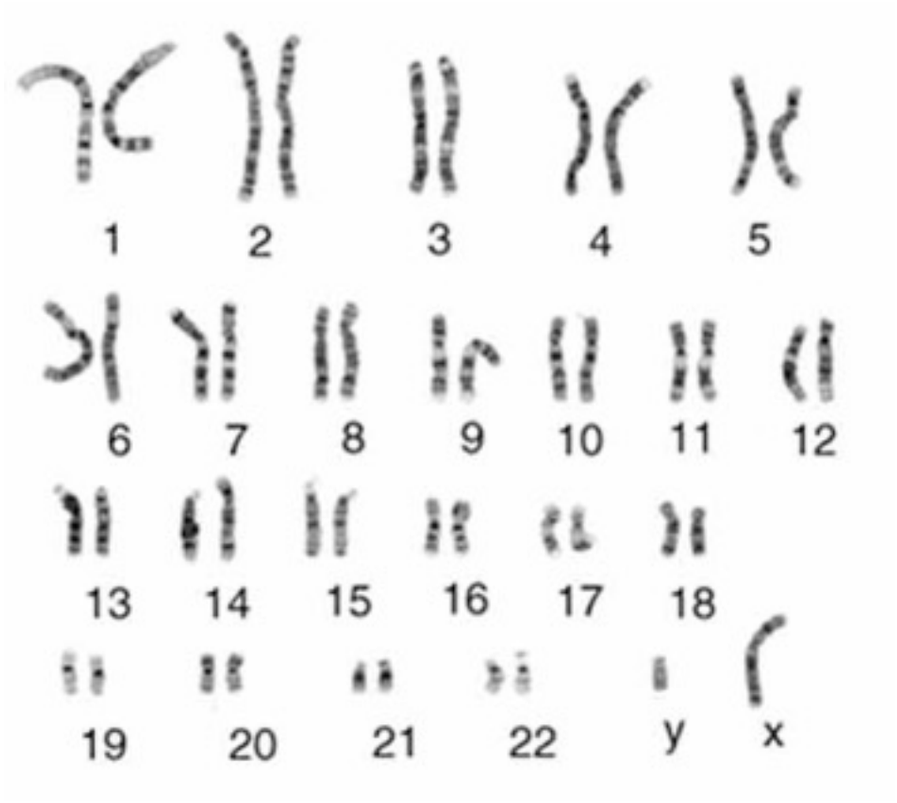
1 - 2% of kids are born with one or more of:

- intellectual disability
- developmental delay
- heart defects
- craniofacial and skeletal abnormalities
- severe autism
- seizures

Most of these problems have genetic causes

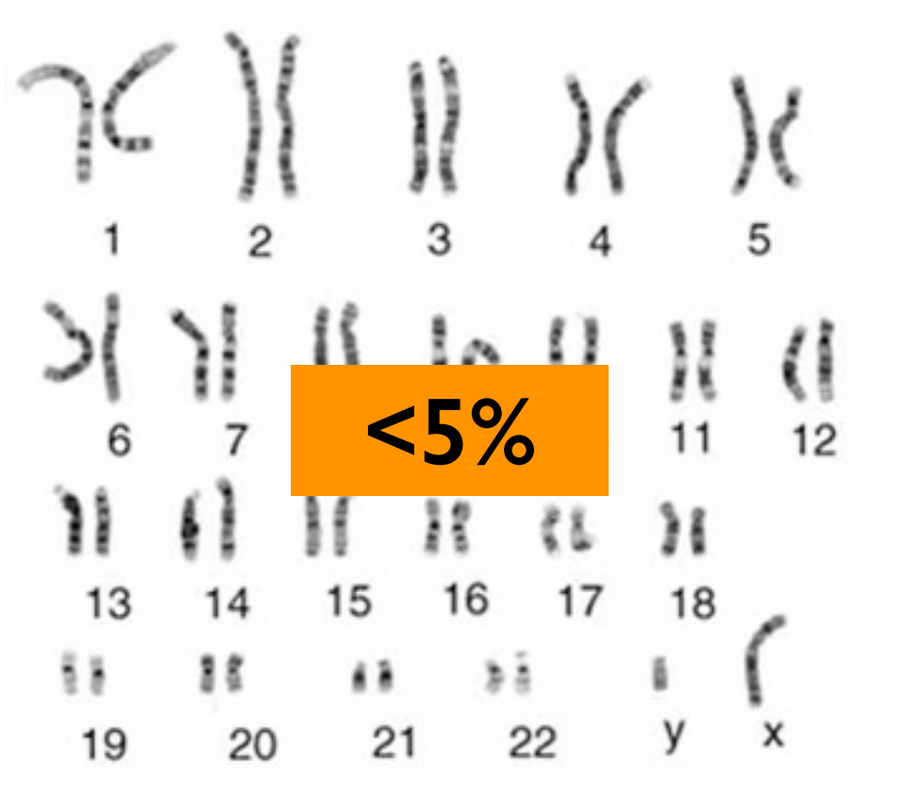
Greg Cooper, UAB Immersion Course, Oct 2015

Genomic Diagnosis of DD/ID



1960 - 1990

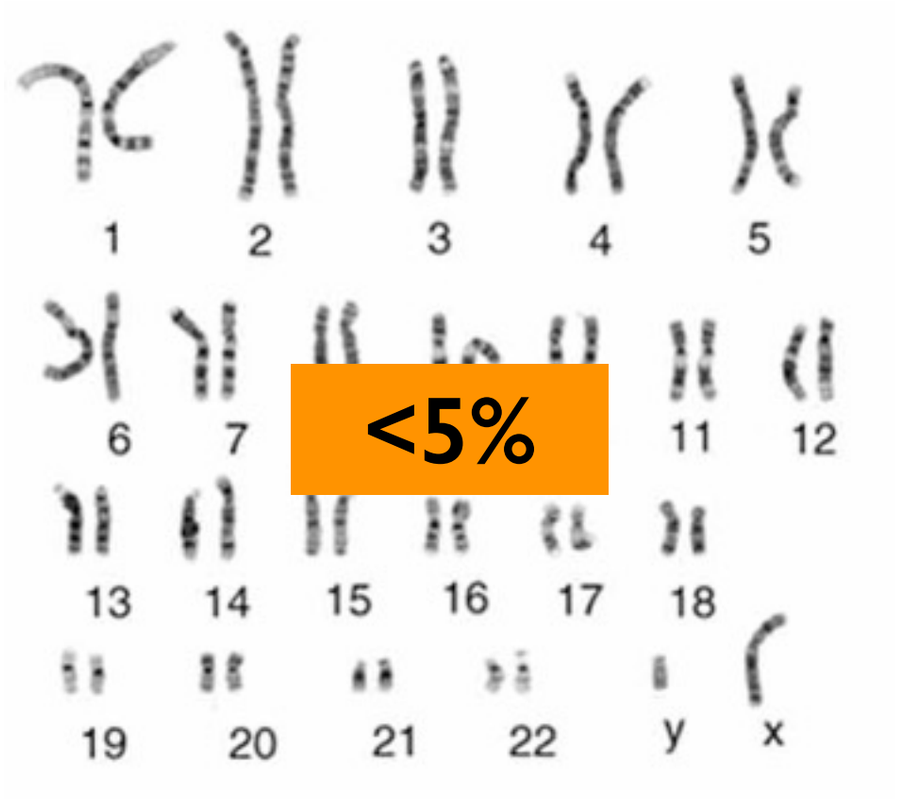
Genomic Diagnosis of DD/ID



<5%

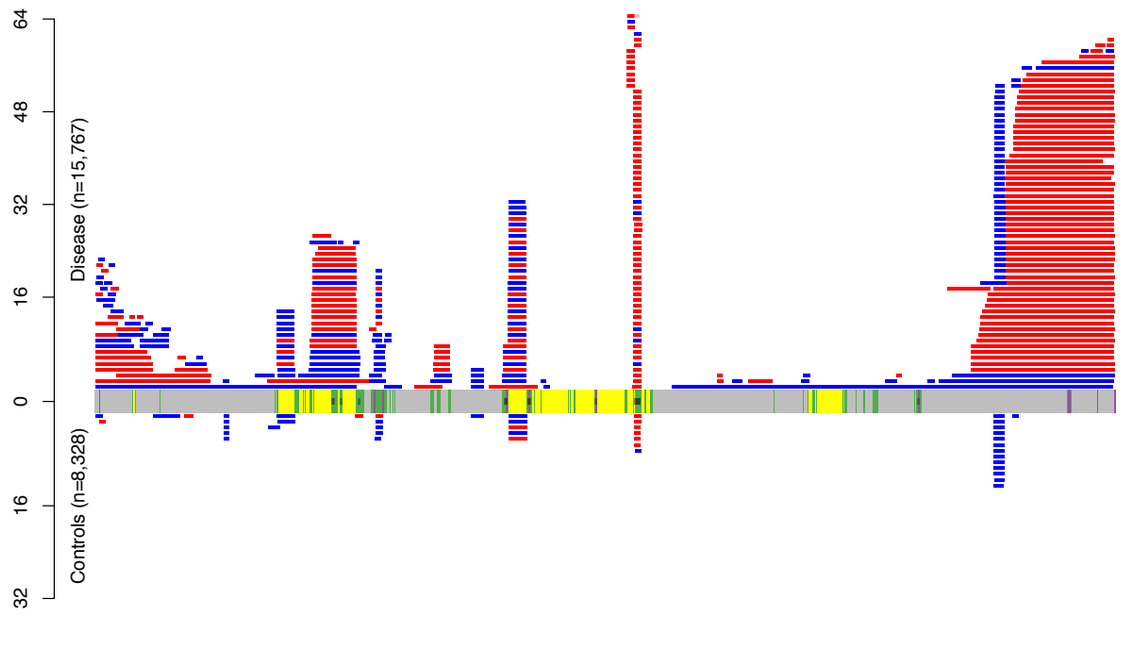
1960 - 1990

Genomic Diagnosis of DD/ID



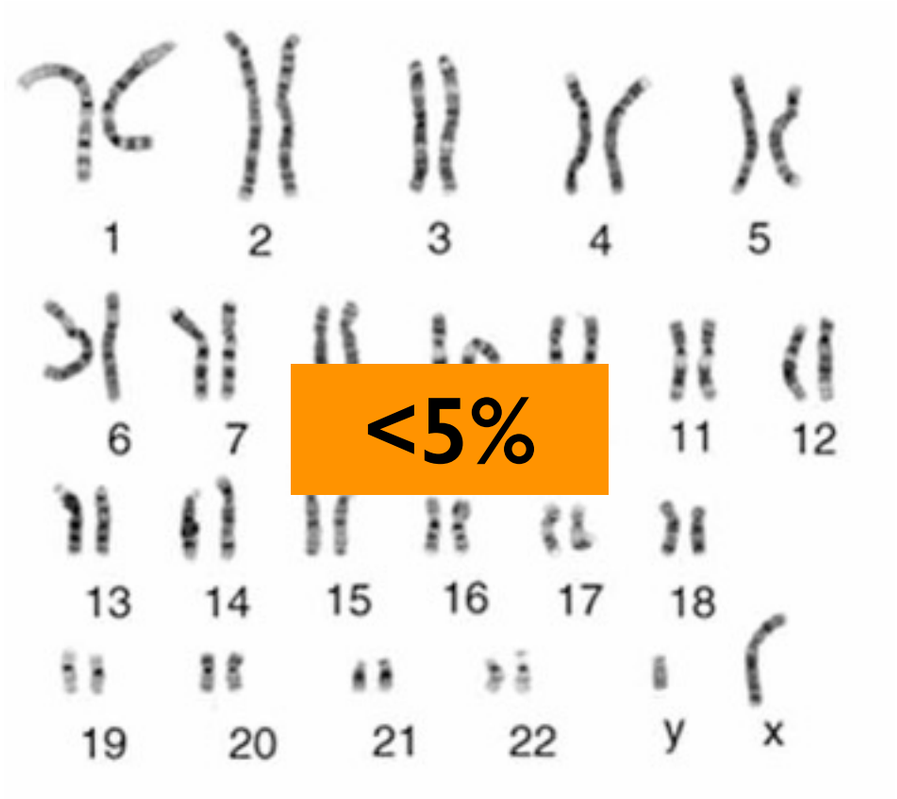
<5%

1960 - 1990



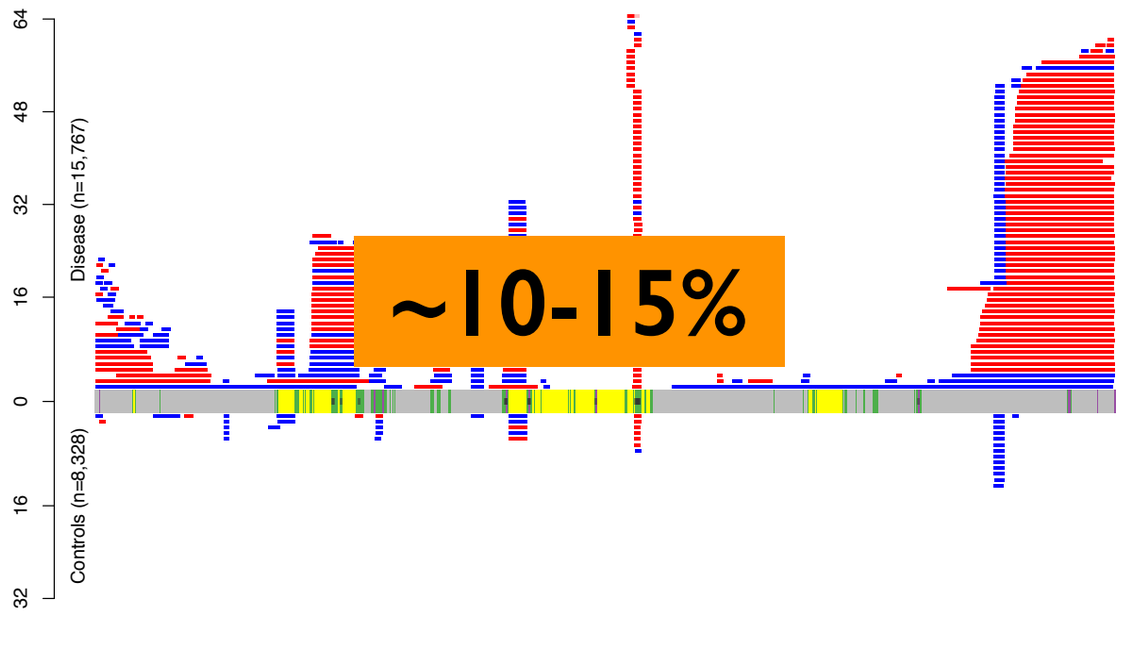
1990 - 2010

Genomic Diagnosis of DD/ID



<5%

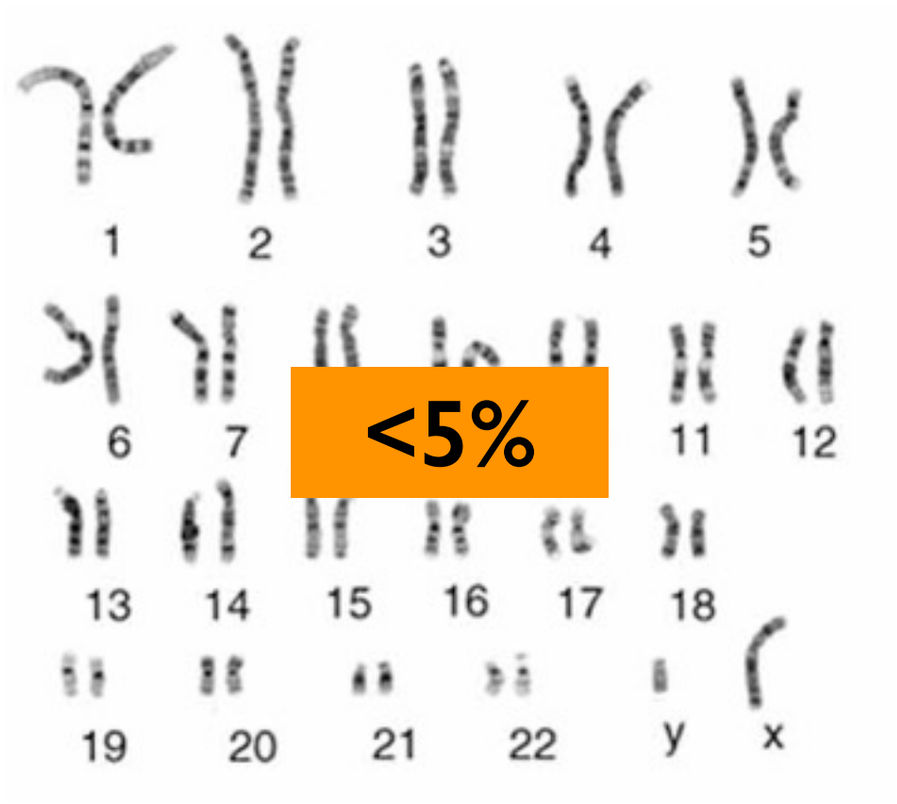
1960 - 1990



~10-15%

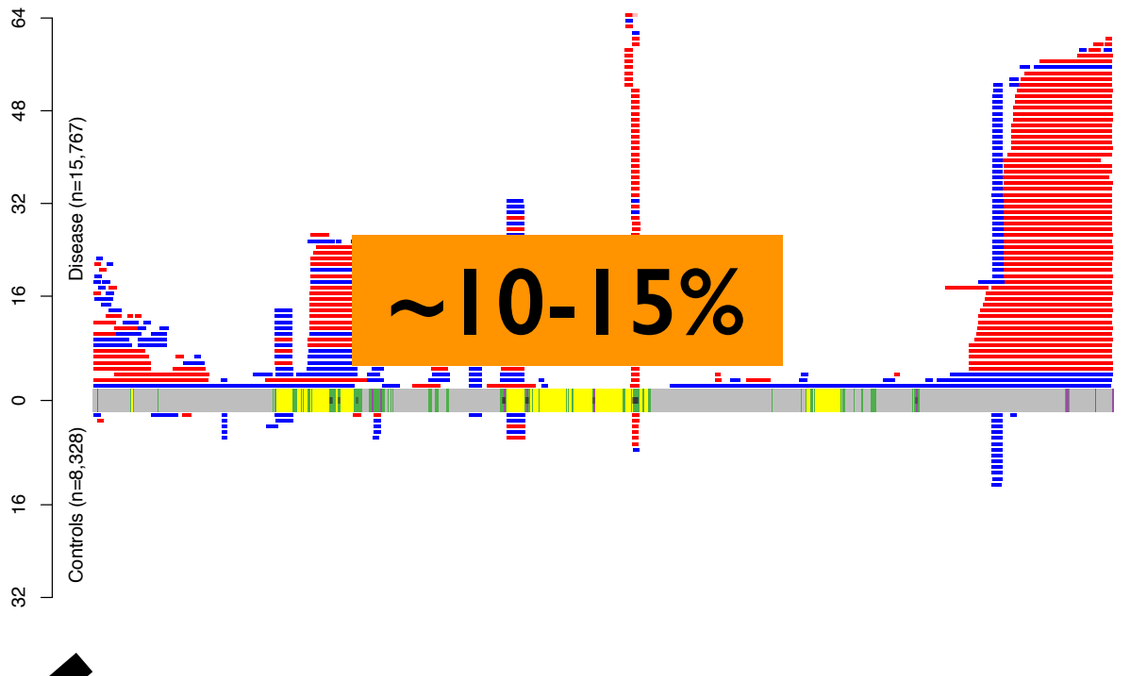
1990 - 2010

Genomic Diagnosis of DD/ID



<5%

1960 - 1990



~10-15%

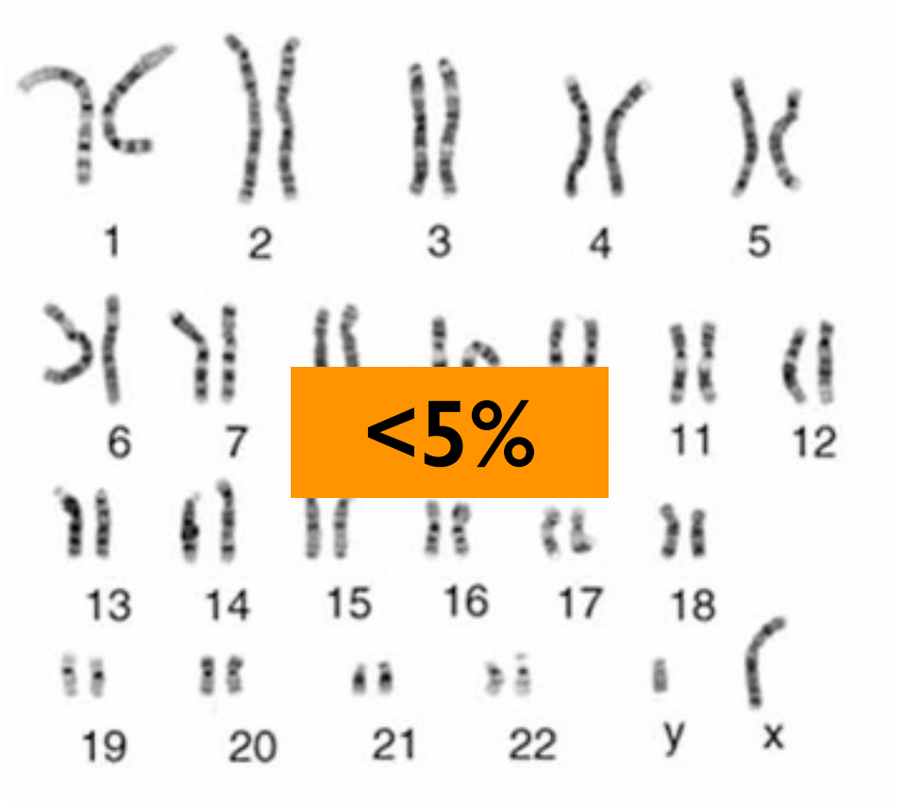
1990 - 2010

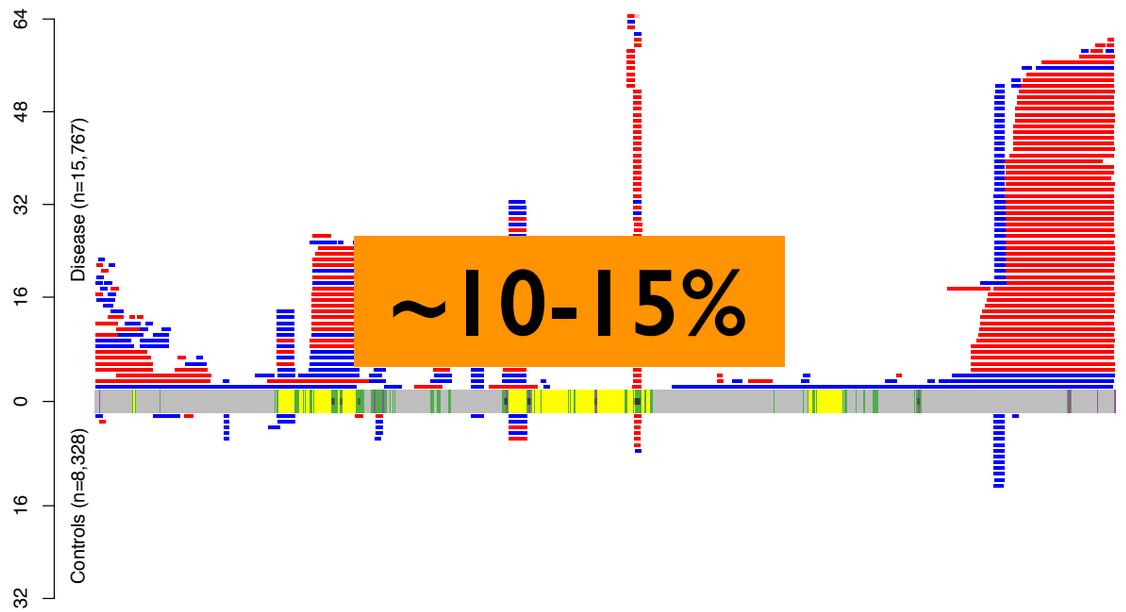
CGTATACCGGGTCATGCACGTGTAGAGCGAG
TTAGCTCGCTGGCTAAAGAGGGTCGACATCC
GCGAGTTTATGAGGAAGAATCGGCAGCTTGA
CCGAAGAGGCGTGGTAAGACCCGTTAGGGAT

Now

Greg Cooper, UAB Immersion Course, Oct 2015

Genomic Diagnosis of DD/ID





1960 - 1990

1990 - 2010

CGTATACCGGGTCATGCACGTGTAGAGCGAG
TTAGCTCGCTGGCTAAAGAGGGTCGACATCC
GCGAGTTTATGAGG? %GAATCGGCAGCTTGA
CCGAAGAGGCGTGGTAAGACCCGTTAGGGAT

Now

Greg Cooper, UAB Immersion Course, Oct 2015

HudsonAlpha Pediatric Genomic CSER Project

HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY

pediatric genomics

1. Recruit 450 parent-offspring trios over 4 years with a DD/ID affected proband without a specific diagnosis

2. Conduct exome (now whole genome) sequencing to identify variants that are DD/ID causal and therefore diagnostic

3. Return pathogenic variants to families

4. Evaluate overall impact of genetic information return on families' health and well-being

Variant Processing

- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:

Variant Processing

- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)

Variant Processing

- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)
  - identify conservation scores and other annotations

Variant Processing

- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)
  - identify conservation scores and other annotations
  - CADD (primary metric for prioritizing)

Variant Processing

- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)
  - identify conservation scores and other annotations
  - CADD (primary metric for prioritizing)
  - **Genetics**

Variant Processing

- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)
  - identify conservation scores and other annotations
  - CADD (primary metric for prioritizing)
- **Genetics**
  - Allele frequency in reference populations crucially important for rare disease variant interpretation

Variant Processing

- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)
  - identify conservation scores and other annotations
  - CADD (primary metric for prioritizing)
- **Genetics**
  - Allele frequency in reference populations crucially important for rare disease variant interpretation
    - e.g., consider implications of a causal variant at 0.1% frequency

Variant Processing

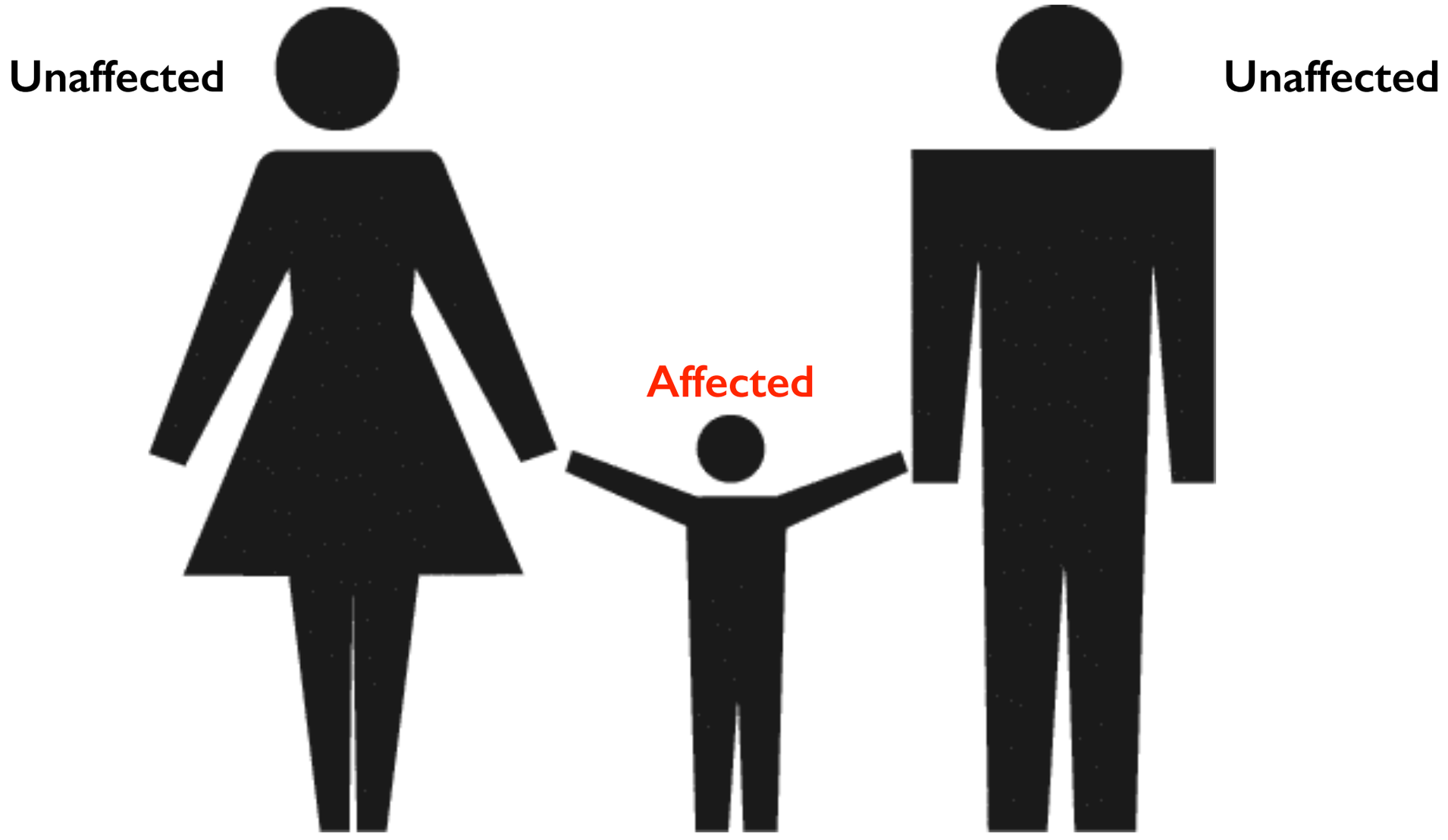
- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)
  - identify conservation scores and other annotations
  - CADD (primary metric for prioritizing)
- **Genetics**
  - Allele frequency in reference populations crucially important for rare disease variant interpretation
    - e.g., consider implications of a causal variant at 0.1% frequency
  - Previous evidence for similar genotypes in individuals with similar phenotypes

Variant Processing

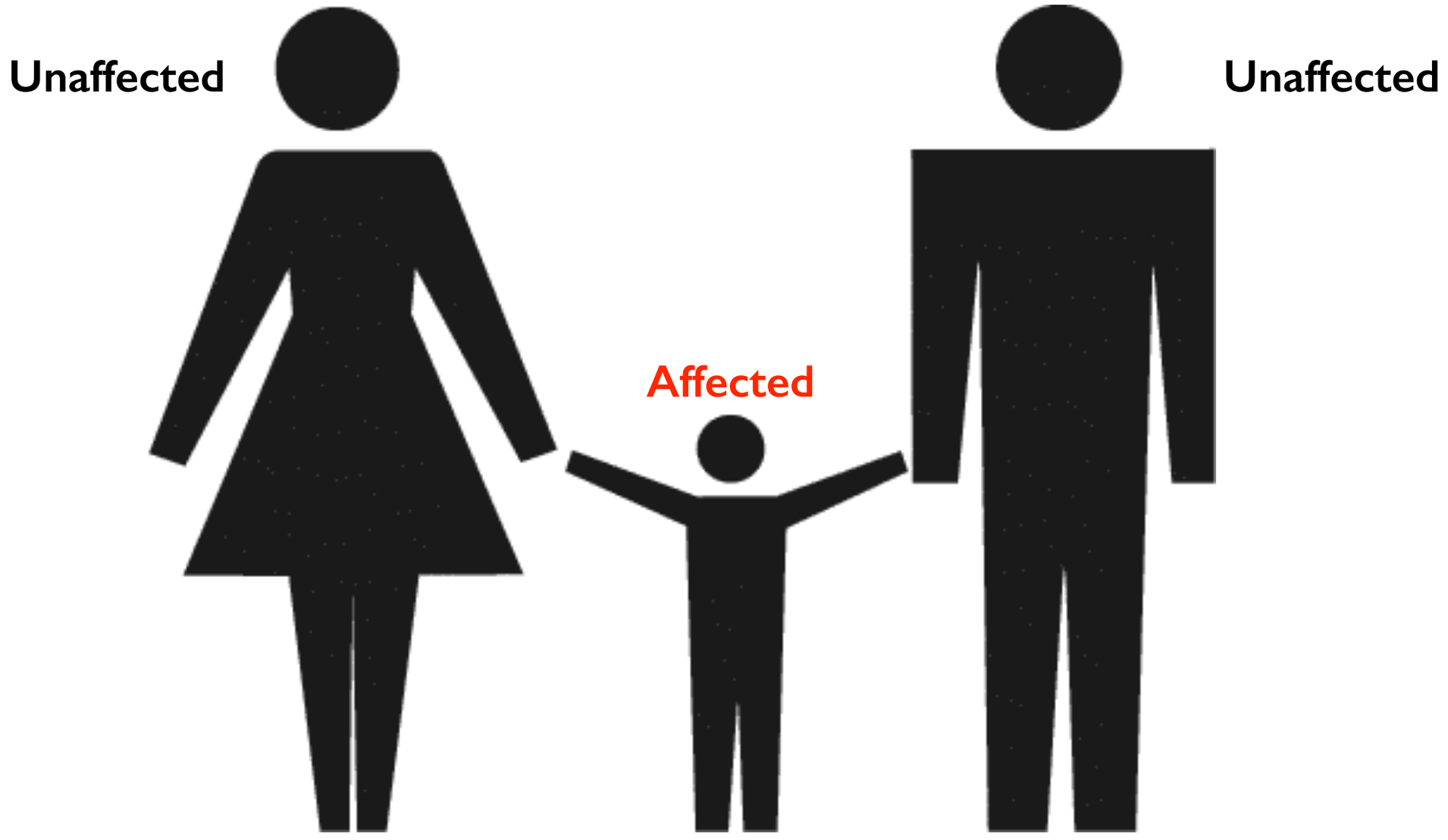
- Typically produce tens of thousands of variants per trio, but the vast majority are not disease relevant, so we:
  - identify gene body overlaps and consequences (e.g., missense)
  - identify conservation scores and other annotations
  - CADD (primary metric for prioritizing)
- **Genetics**
  - Allele frequency in reference populations crucially important for rare disease variant interpretation
    - e.g., consider implications of a causal variant at 0.1% frequency
  - Previous evidence for similar genotypes in individuals with similar phenotypes
  - Familial inheritance also crucial

Greg Cooper, UAB Immersion Course, Oct 2015

Familial Variant Analysis Paradigms

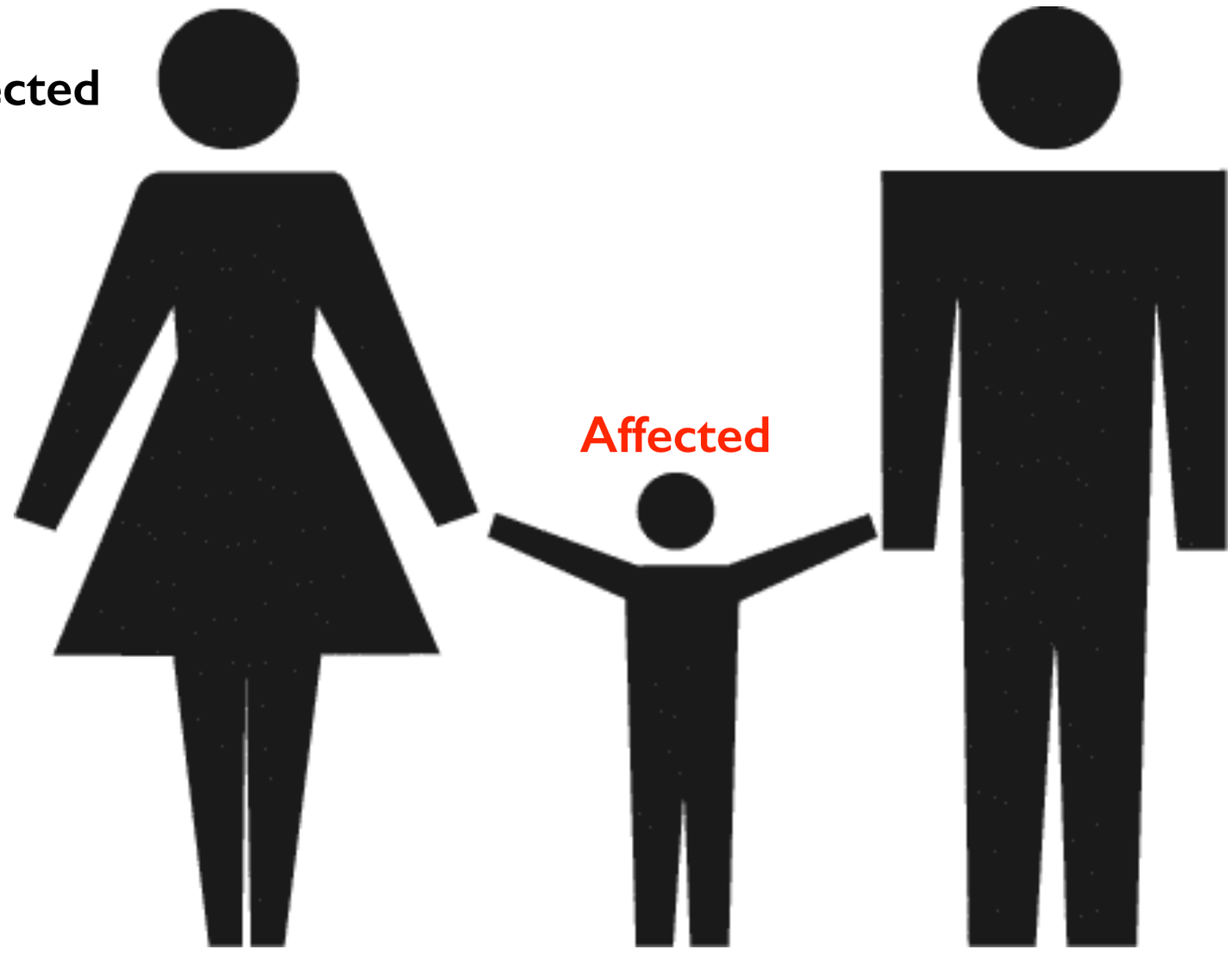


Familial Variant Analysis Paradigms



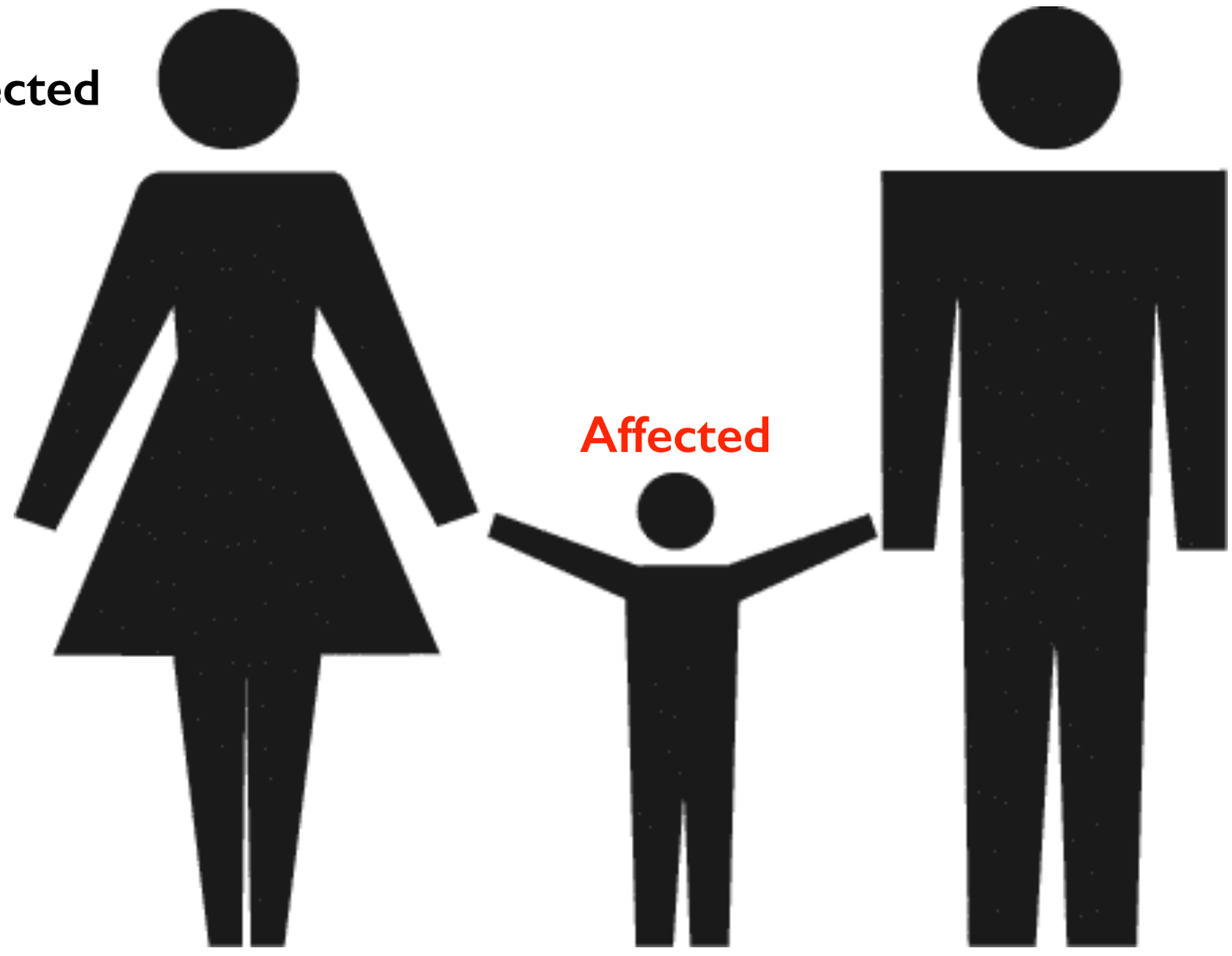
De Novo

Familial Variant Analysis Paradigms



De Novo	+/+	+/-	+/+

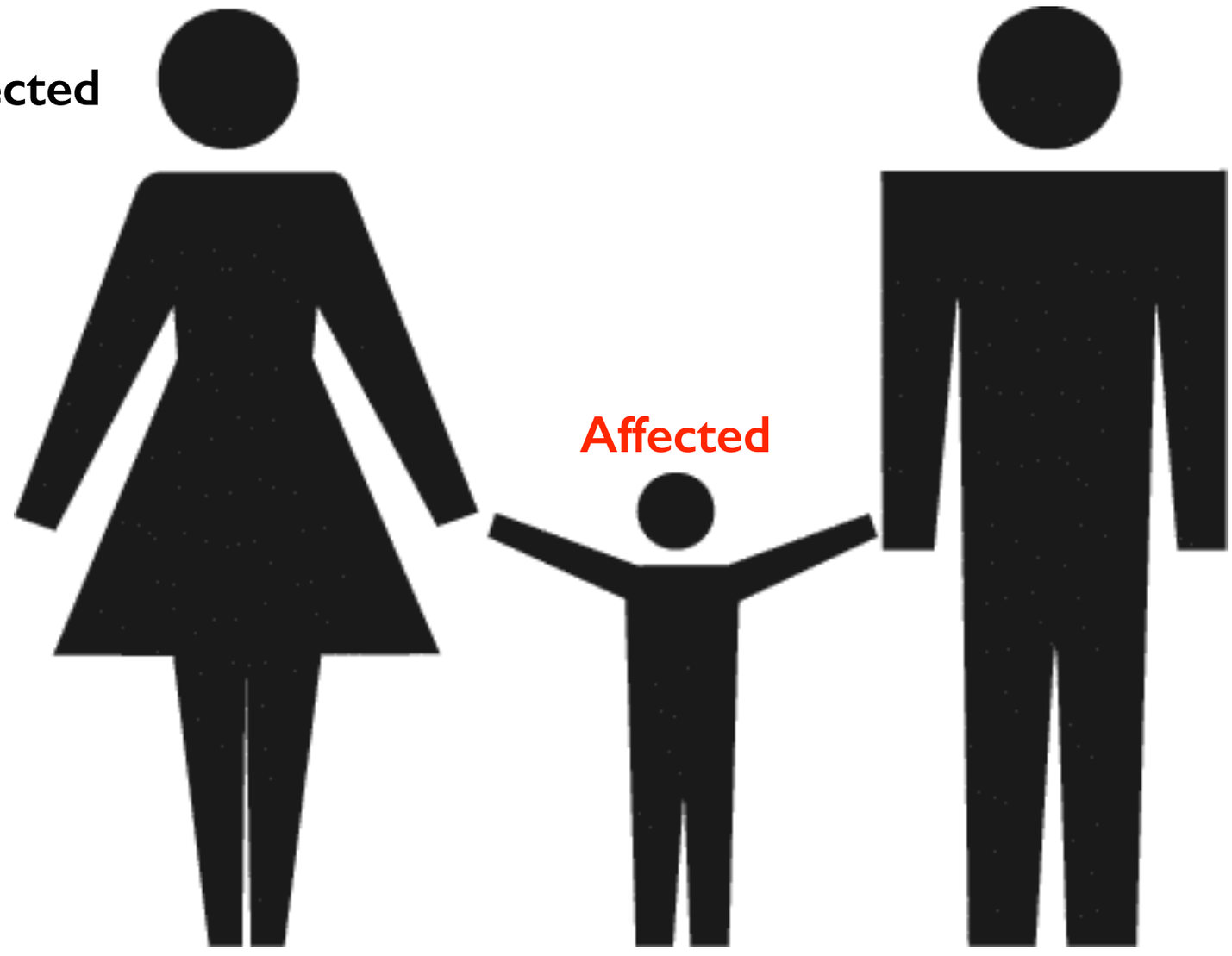
Familial Variant Analysis Paradigms



De Novo +/+ +/- +/+

Comp-Het or Recessive

Familial Variant Analysis Paradigms



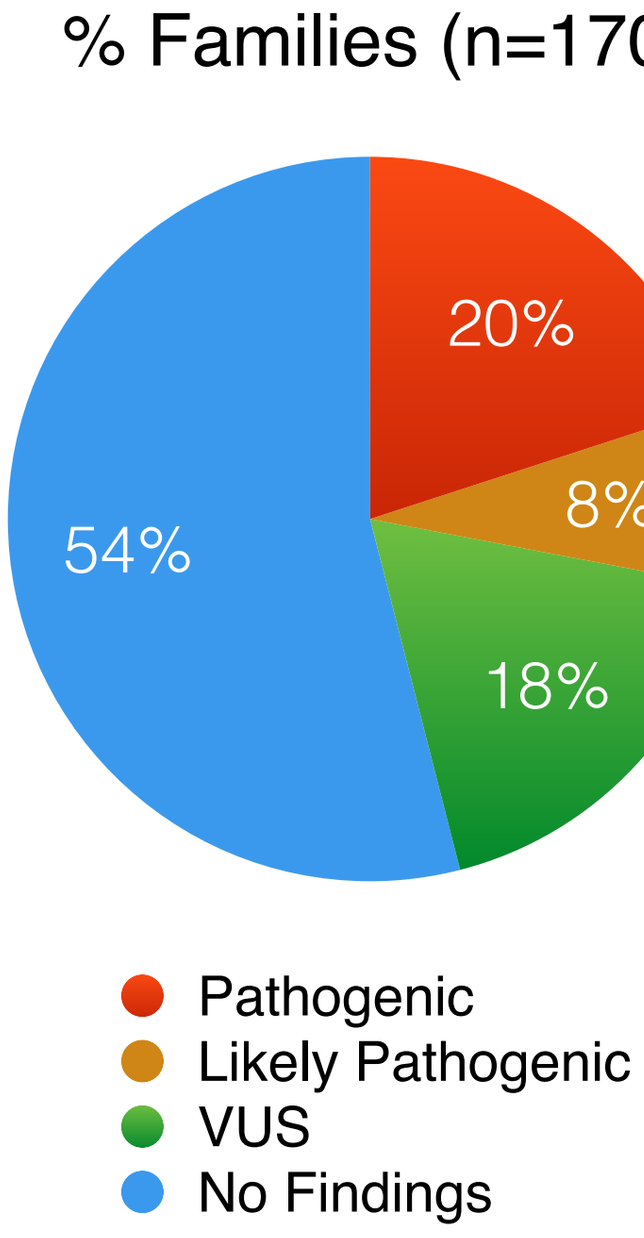
De Novo	+/+	+/-	+/+
Comp-Het or Recessive	+/-	-/-	+/-

Variant Review

- After filtering based on population frequencies, genomic annotations, and inheritance patterns, variants are manually reviewed by at least two genomic analysts
- Lists typically span a few to dozens of variants
- Goal is to evaluate the totality of evidence, integrating resources like OMIM, ClinVar, published literature, genomic annotations, etc., to make a determination as to the medical relevance of any given mutation(s)
  - genetics and correlations with previous cases is paramount

Greg Cooper, UAB Immersion Course, Oct 2015

Diagnostic Finding Overview



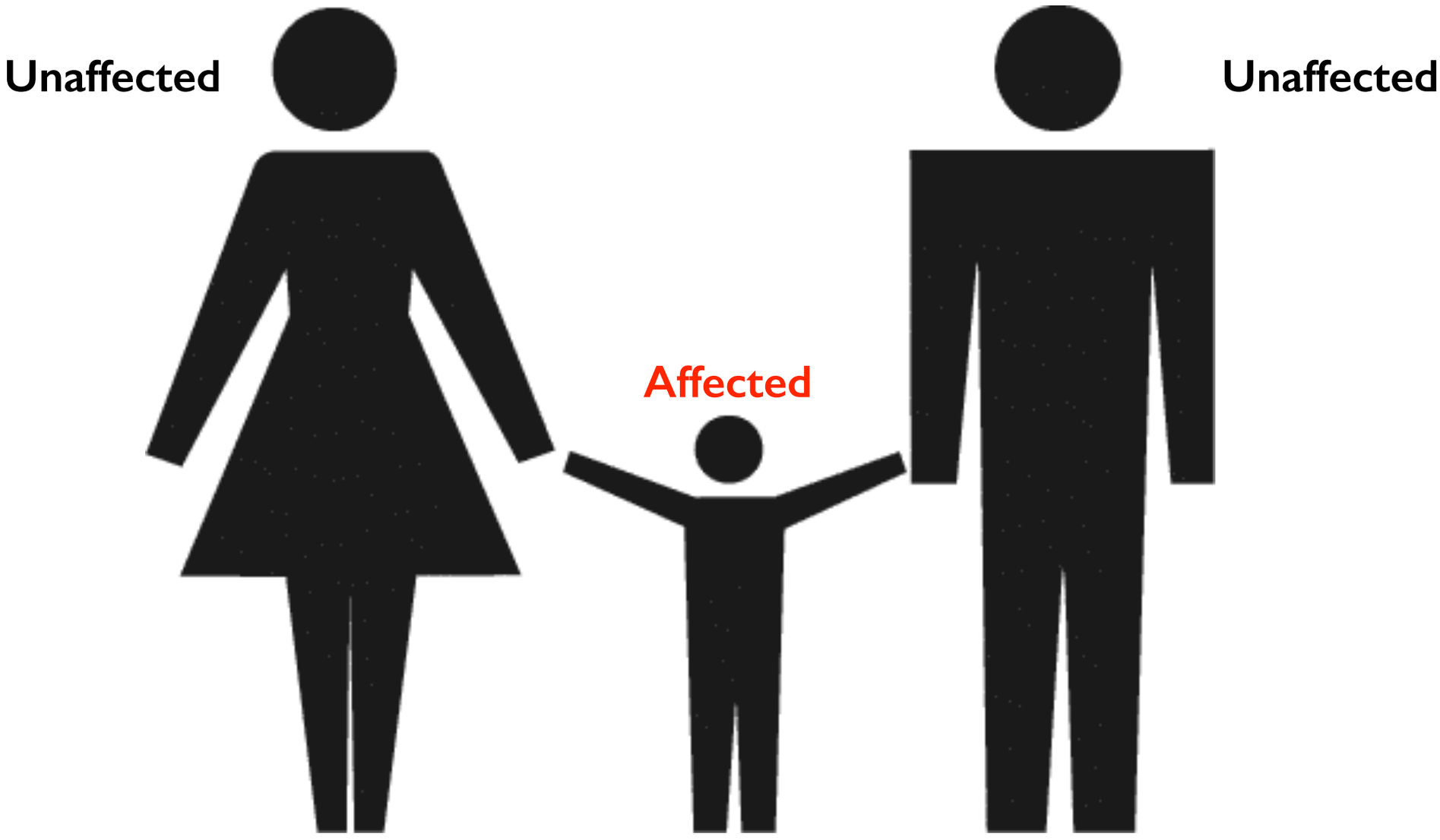
- Genetic diagnosis for 28% of families
  - Pitt-Hopkins syndrome
  - Dravet syndrome
  - Rett syndrome
  - Rubinstein-Taybi syndrome
  - Noonan-like syndrome

- Variants of uncertain significance identified in 18% of families; may be diagnostic in the future

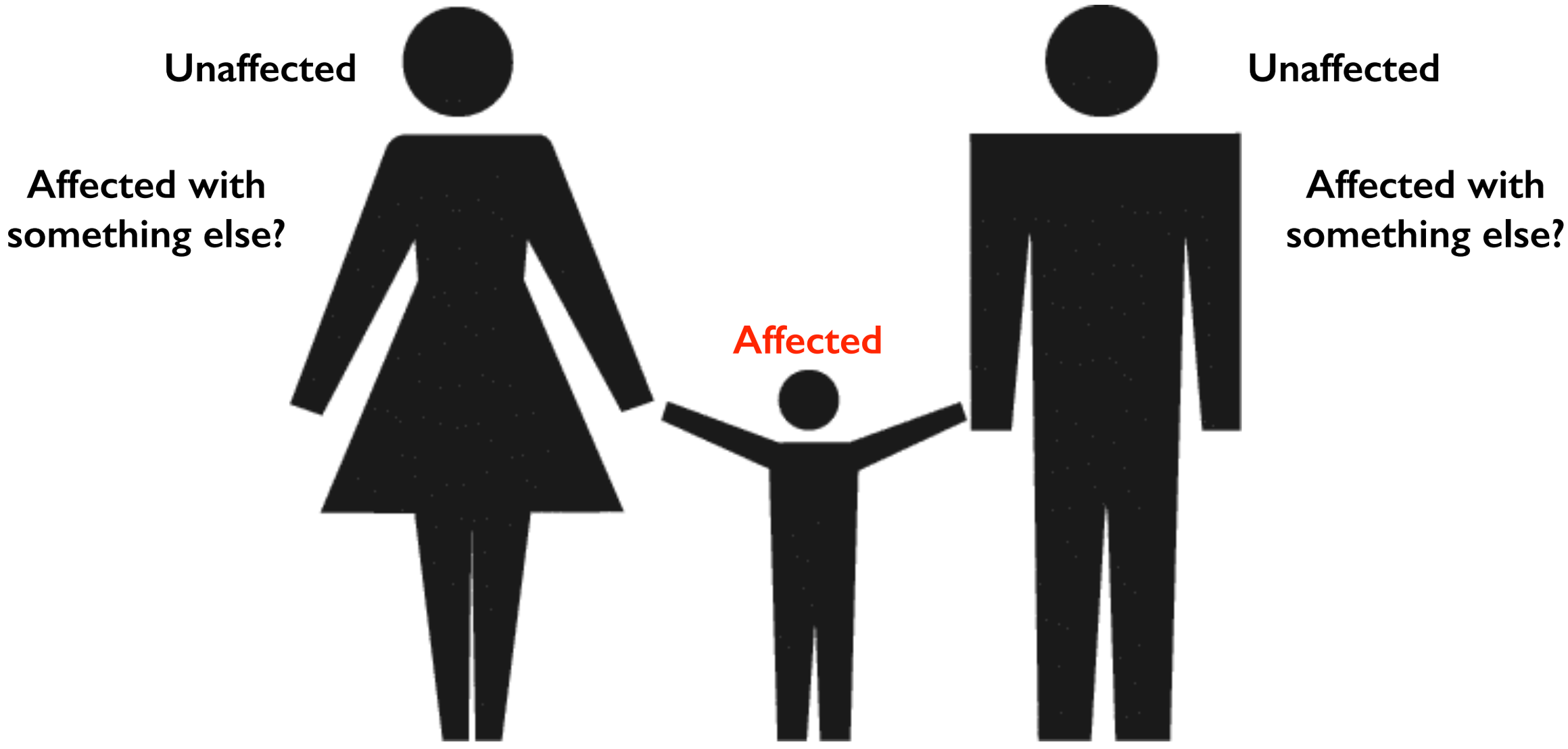
- Re-analysis of negative exomes using updated information identified three diagnostic variants (e.g. MTOR, DDX3X and CLPB)

Greg Cooper, UAB Immersion Course, Oct 2015

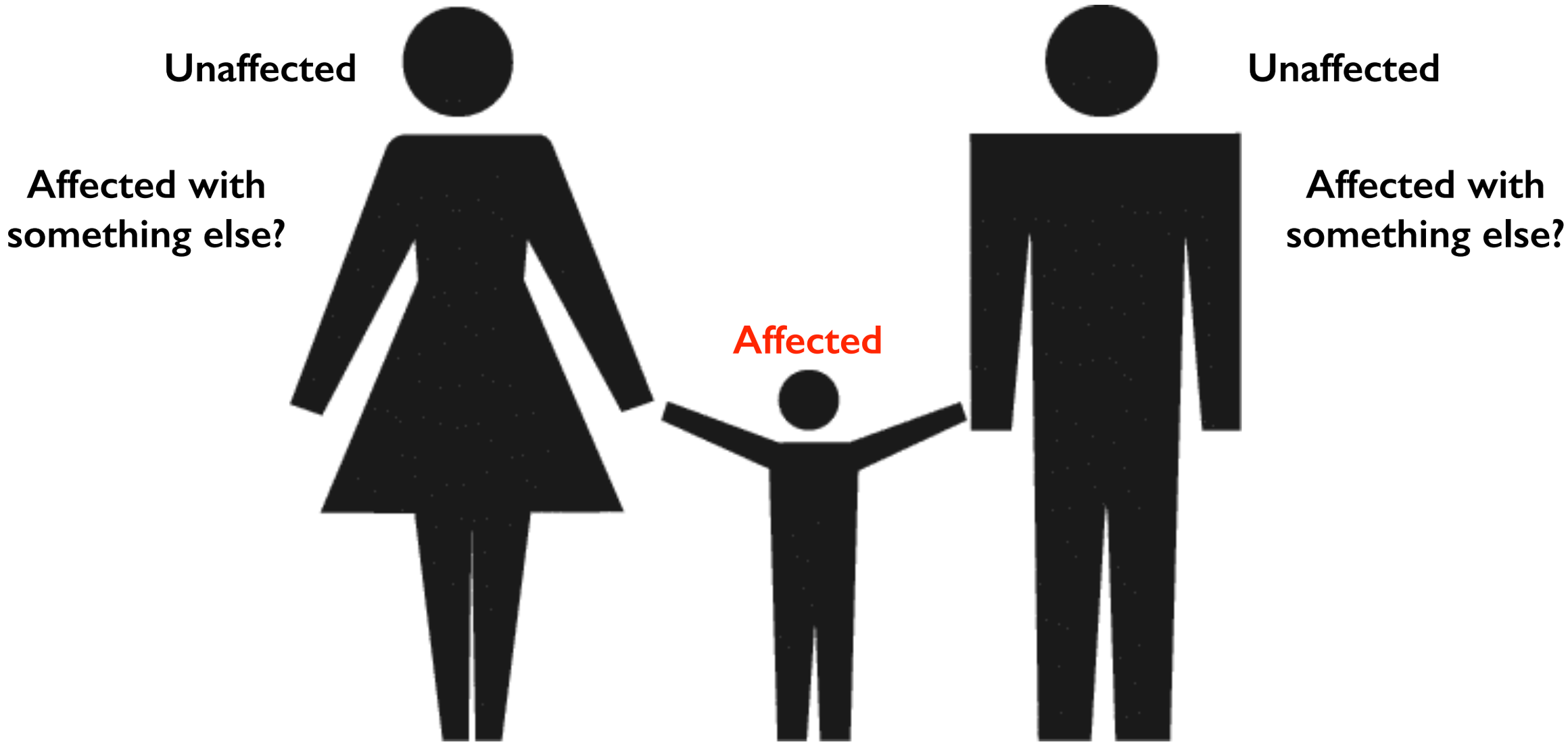
“Secondary” or “Incidental” Findings



“Secondary” or “Incidental” Findings



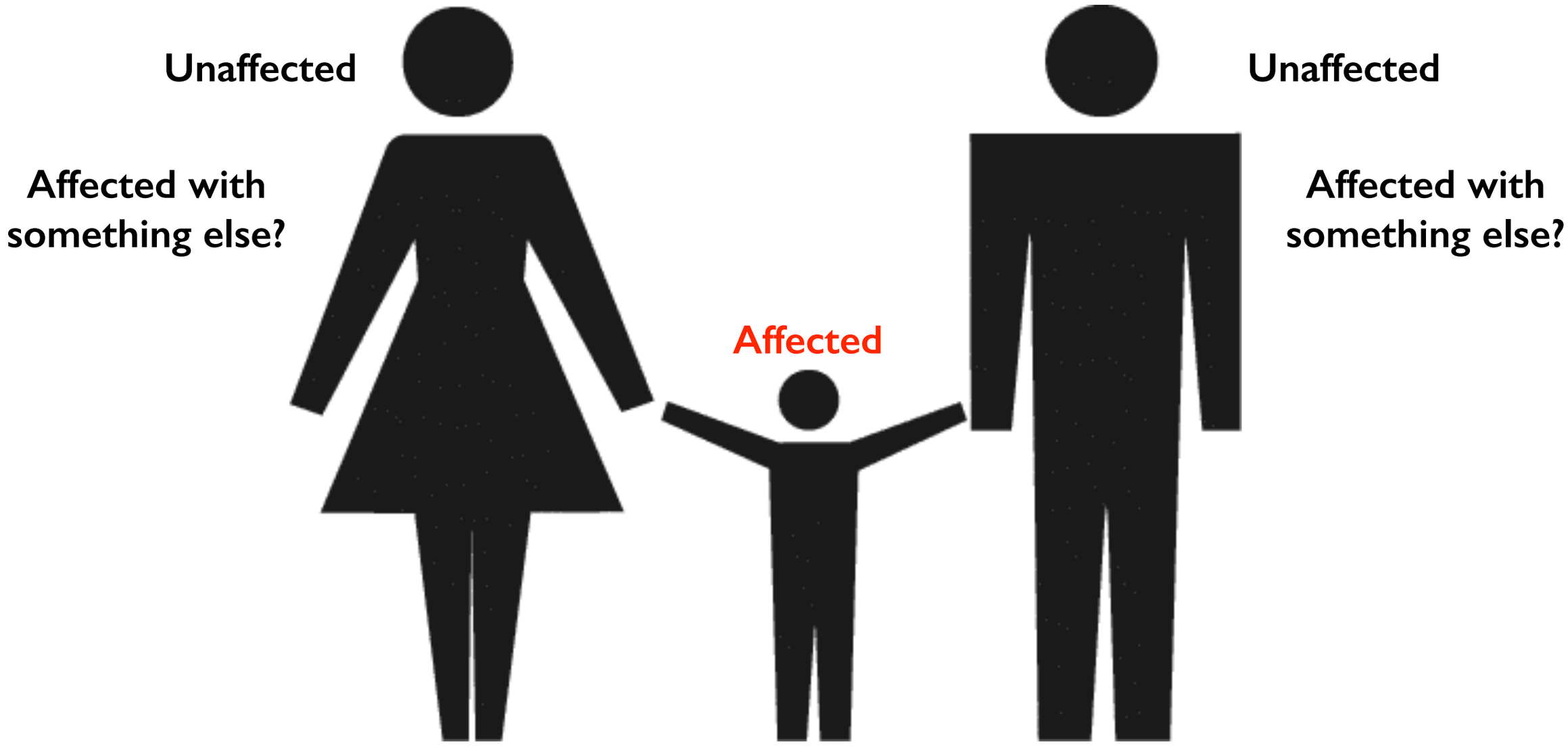
“Secondary” or “Incidental” Findings



ACMG gene list (57 genes), mostly dominant mutations that result in medically “actionable” conditions

Greg Cooper, UAB Immersion Course, Oct 2015

“Secondary” or “Incidental” Findings

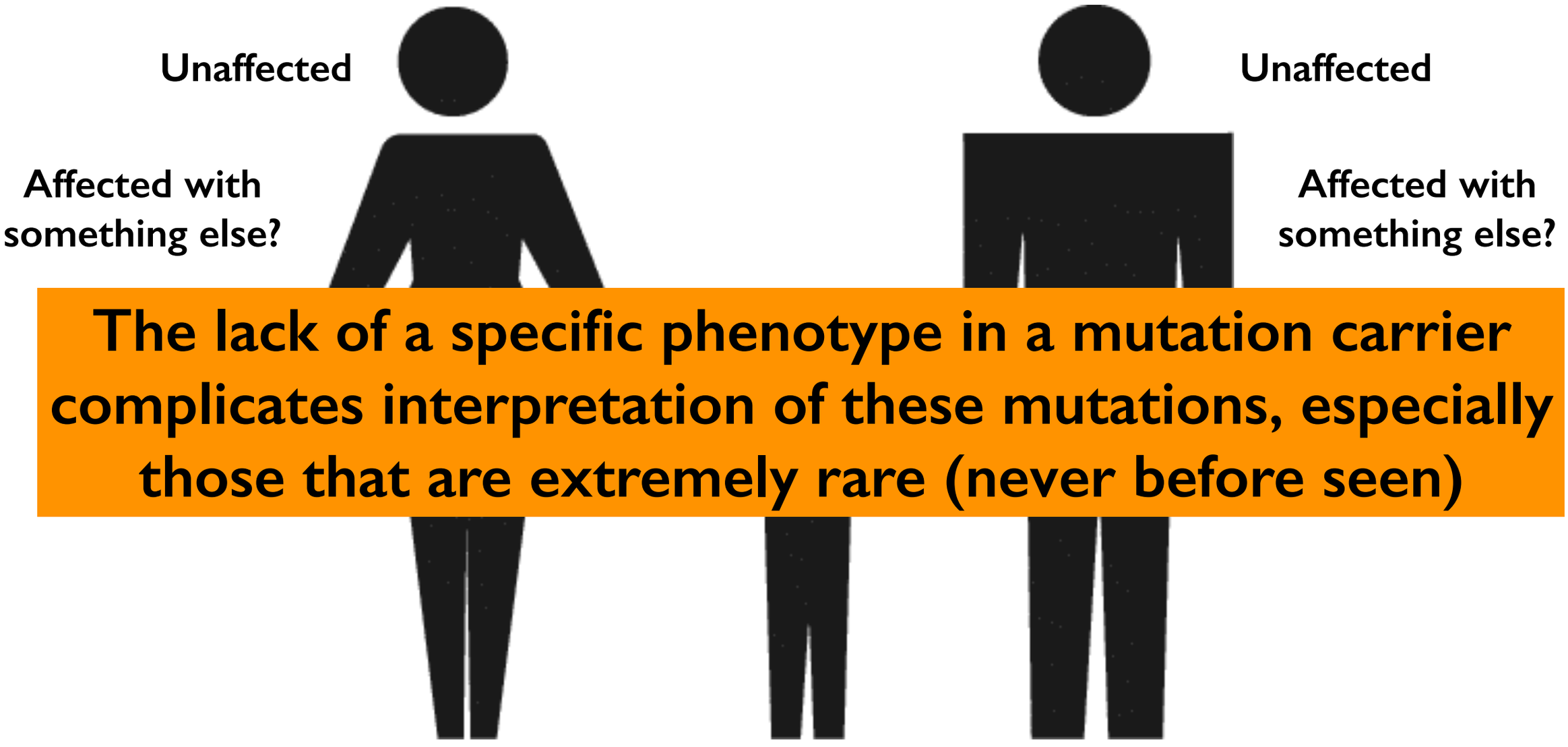


ACMG gene list (57 genes), mostly dominant mutations that result in medically “actionable” conditions

OMIM gene lists for recessive disease risk in future children

Greg Cooper, UAB Immersion Course, Oct 2015

“Secondary” or “Incidental” Findings



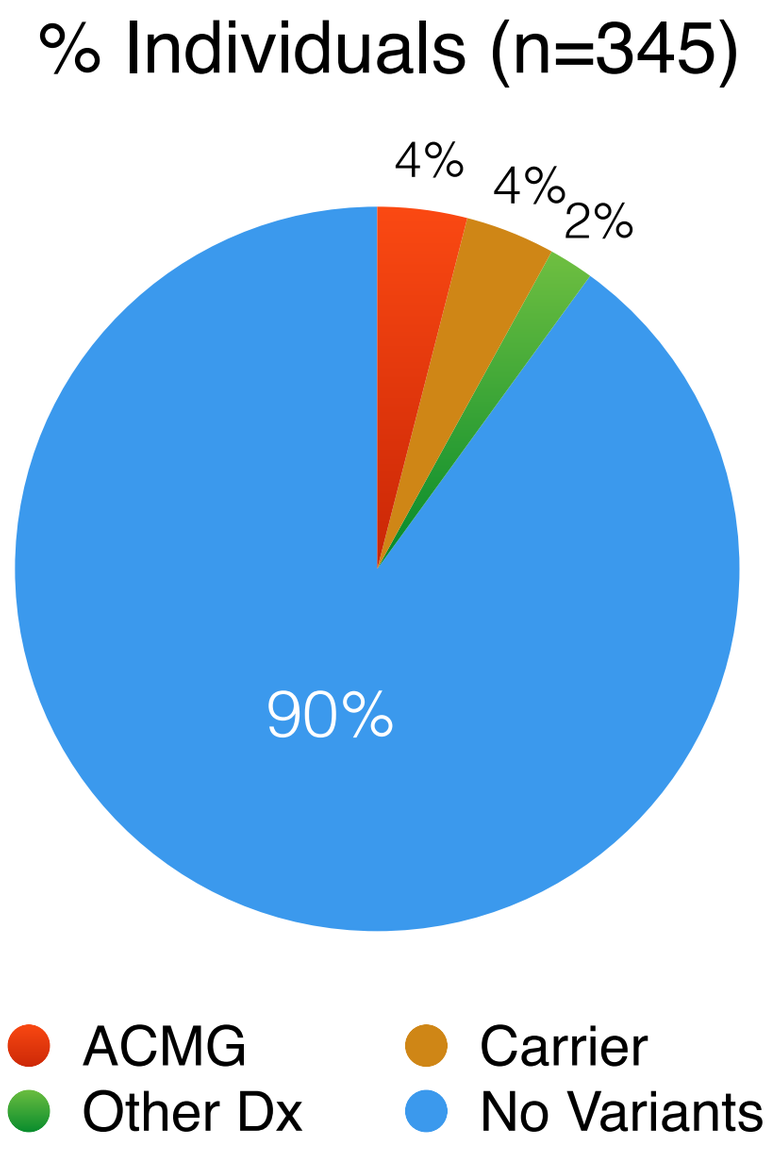
The lack of a specific phenotype in a mutation carrier complicates interpretation of these mutations, especially those that are extremely rare (never before seen)

ACMG gene list (57 genes), mostly dominant mutations that result in medically “actionable” conditions

OMIM gene lists for recessive disease risk in future children

Greg Cooper, UAB Immersion Course, Oct 2015

Secondary Findings Overview



• Identified secondary variants are linked to many different diseases/conditions

- Cystic Fibrosis (carrier)
- Tay-Sachs disease (carrier)
- Sickle cell disease (carrier)
- Breast cancer
- Colorectal cancer
- Cardiomyopathy
- Malignant hyperthermia susceptibility
- Carnitine deficiency
- Albinism

- Return or not driven by preference
- Uncertain genetics and risk/benefit to interventions lead to difficult decisions

Greg Cooper, UAB Immersion Course, Oct 2015

General Conclusions

•Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:

General Conclusions

- Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:
  - Better diagnostics

General Conclusions

•Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:

- •Better diagnostics
- •New gene discoveries

General Conclusions

•Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:

- •Better diagnostics
- •New gene discoveries

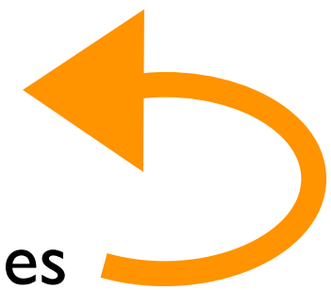
General Conclusions

- Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:
  - Better diagnostics
  - New gene discoveries

- New and better genomic annotations are needed, especially for non-coding variation

General Conclusions

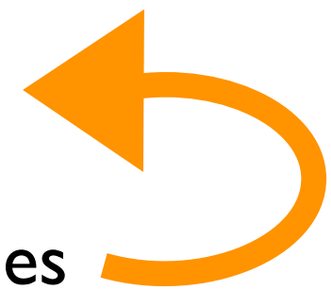
- Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:
  - Better diagnostics
  - New gene discoveries



- New and better genomic annotations are needed, especially for non-coding variation
- Data sharing is another area that needs improvement

General Conclusions

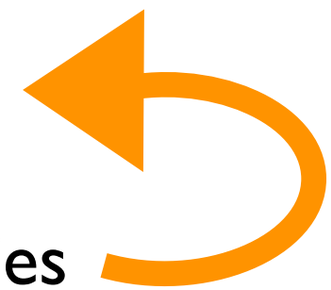
- Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:
  - Better diagnostics
  - New gene discoveries



- New and better genomic annotations are needed, especially for non-coding variation
- Data sharing is another area that needs improvement
  - best evidence is more genotype-phenotype correlation, which can only emerge from vast data networks for rare variants and rare diseases

General Conclusions

- Large-scale sequence analysis of rare diseases in patient populations is producing a wave of new data, new discoveries, and new possibilities:
  - Better diagnostics
  - New gene discoveries



- New and better genomic annotations are needed, especially for non-coding variation
- Data sharing is another area that needs improvement
  - best evidence is more genotype-phenotype correlation, which can only emerge from vast data networks for rare variants and rare diseases
- "Proving" variant causality is very hard, and often impossible, but for clinical decisions certainty may or may not be necessary (i.e., all possible error types and consequences must be evaluated)

Greg Cooper, UAB Immersion Course, Oct 2015