

Using *Population Variation* and novel
Pathogenicity Scores to support diagnostic
interpretation of *variants of unknown clinical
significance*

Matthew Hurles

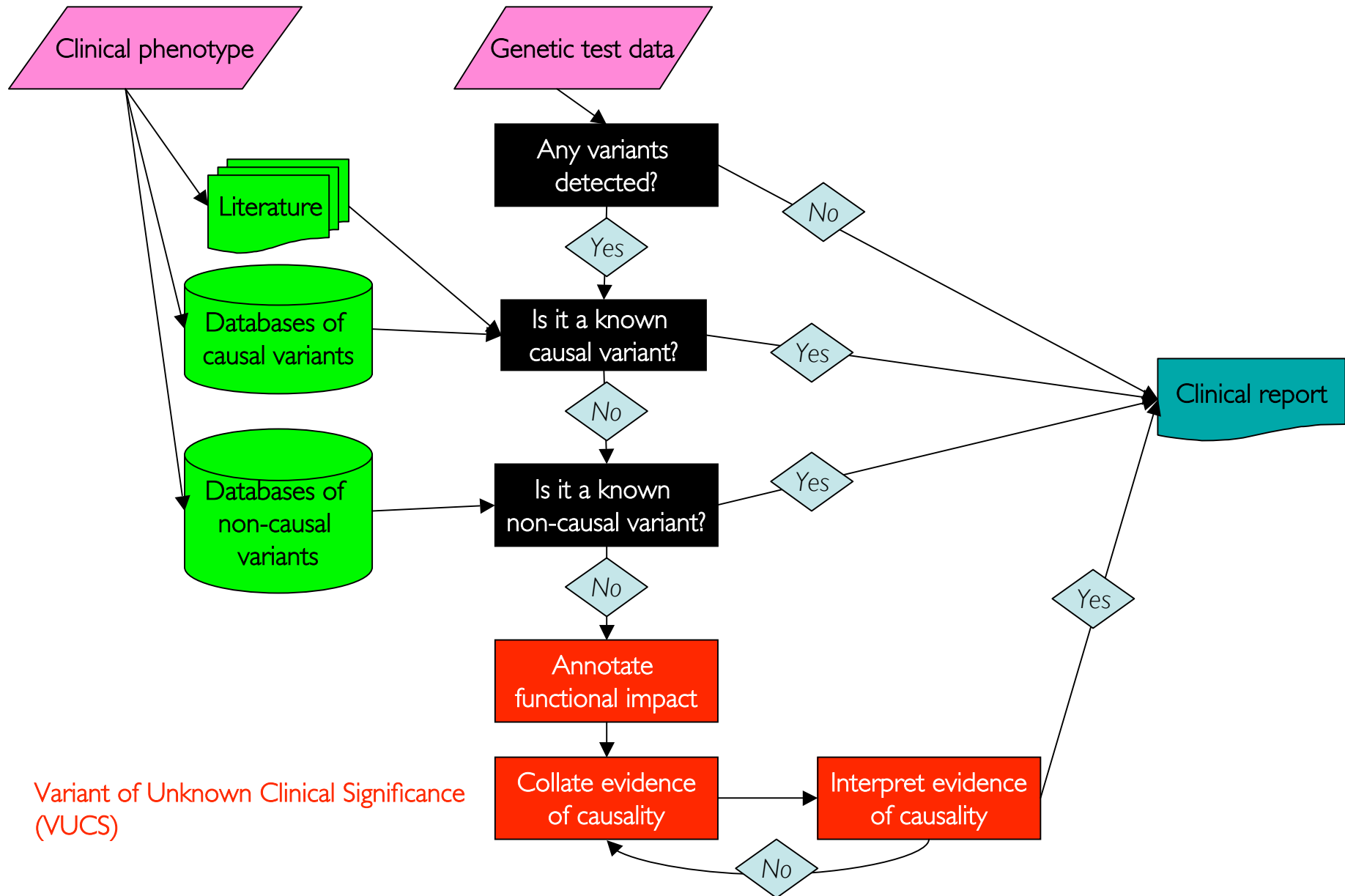


Can we improve the clinical interpretation of Variants of Unknown Clinical Significance?

- Generic framework for assessing the genetic diagnostic hypothesis
 - Integrating pathogenicity scores and population variation
- Specific application to LOF variants (non-recurrent CNVs)
 - Using pathogenicity scores derived from gene-based predictions of haploinsufficiency (dosage-sensitivity)

Generic framework for assessing the genetic diagnostic hypothesis

A generic 'decision tree'



Example scenarios

- My breast cancer patient has a rare missense mutation in BRCA1, which we've not seen before, with POLYPHEN score of X, I only see a variant of this pathogenicity score in 3 out of 400 rare missense variants seen in population controls, is it causal?
- My MR patient has a rare *de novo* deletion that takes out 3 genes of unknown function, is it causal?
- My patient with a CHD and facial dysmorphology has a novel 3Mb duplication, I can't get parental samples, is it causal?

Holy grail: robust interpretation of a variant never seen before

What do we mean by 'causal'?

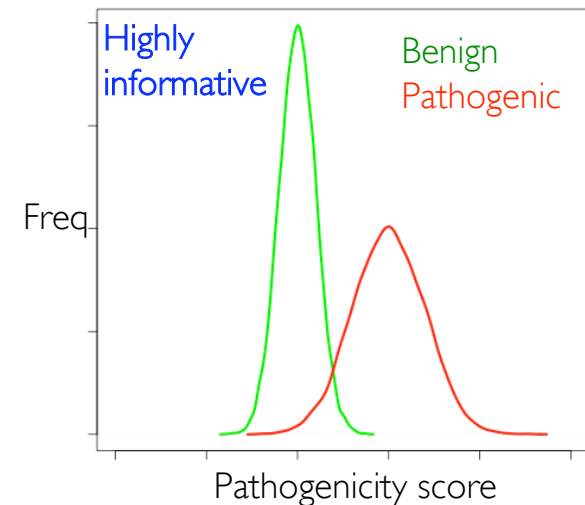
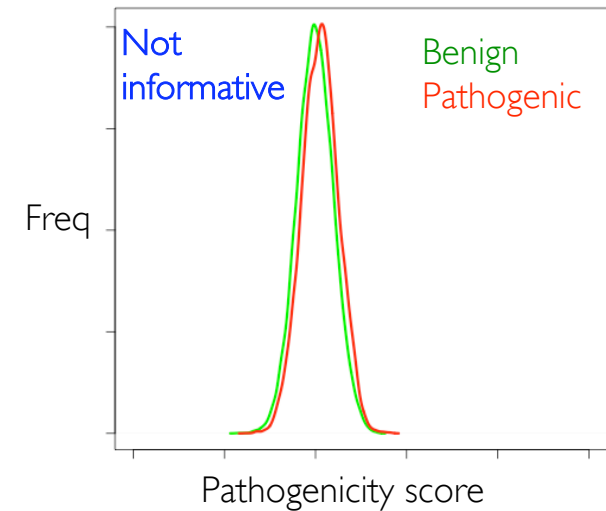
Here:

variant is sufficient to cause patient phenotype
(might have more than one causal variant per genome)

Note: not trying to apportion finite causality among variants in the same genome

What information do we have to inform interpretation?

- Inheritance information
 - *De novo*
 - Inherited from unaffected parent
 - Inherited from affected parent
 - Inheritance status unknown
- Pathogenicity scores
 - Metric that discriminates (imperfectly) between causal and non-causal variants
 - Integrates diverse sources of evidence
 - Genomic, Biochemical, Evolutionary, Functional
 - Trained on available data (causal, non-causal)
 - Specific to types of variation (missense, genic CNV, splicing, etc)
- Population variation

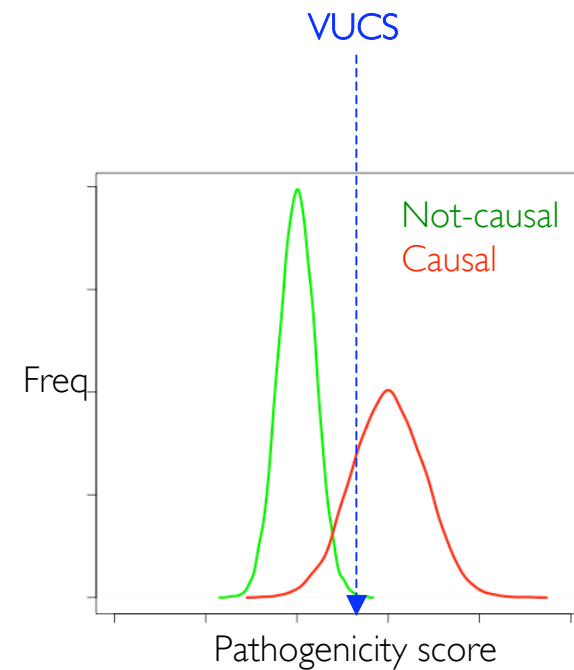


Uses of population variation

- Exclusionary Filter:
 - Variant is too common to be plausibly causal
 - Depends on assumptions about genetic model (dominant/recessive) and penetrance
- Association testing (variants seen in >1 patient):
 - Variant is present in patients at significantly higher frequency than in controls
- How can population variation be used if the variant has not been seen before?
 - Construct distribution of pathogenicity scores of observed variants
 - Having identified a VUCS, use this distribution to estimate:
 - Probability of seeing a variant with this score *by chance from the population*

Pathogenicity Score vs Causality

- Calculate pathogenicity score for a VUCS
- How similar is this score to the score seen in known **causal** and **non-causal** variants?
- What does this tell us?
 - **Prob.(Score | Causal)**
 - **Prob.(Score | Not Causal)**
- But we want to know **Prob.(Causal | Score)**



$\text{Prob.}(\text{Causal} | \text{Score}) \neq \text{Prob.}(\text{Score} | \text{Causal})$

Bayesian diagnosis

- Principle:
 - Revise (clinical) opinion on the basis of the evidence
 - Iterative refinement of the opinion as evidence (a series of tests) accumulates

“clinicians apply bayesian reasoning in framing and revising differential diagnoses without necessarily undergoing, or requiring, any formal training in bayesian statistics.” Gill et al. 2005 BMJ

- In genetic diagnostic setting:
 - Use evidence from genetic assay to revise a PRIOR (pre-test) opinion that a given variant (e.g. rare genic deletion) in a given patient’s genome might be causal

“Admittedly, the definition of pre-test odds of a disease for a given patient is inherently subjective. But the alternative to subjectivity is to exclude clinical judgment (which is all about context) from patient care.” Gill et al. 2005 BMJ

- What is the evidence?
 - Probability of seeing a variant with this score *given it is causal*
 - Probability of seeing a variant with this score *by chance from the population (non-causal)*

Relating $p(\text{Causal} \mid \text{Score})$ to $p(\text{Score} \mid \text{Causal})$

Bayes rule:

$$P(C \mid S) = p(C) \cdot p(S \mid C) / p(S)$$

$$P(\text{not } C \mid S) = p(\text{not } C) \cdot p(S \mid \text{not } C) / p(S)$$

Causal odds:

$$\frac{P(C \mid S)}{P(\text{not } C \mid S)} = \frac{p(C) \cdot p(S \mid C)}{p(\text{not } C) \cdot p(S \mid \text{not } C)}$$

Causal probability = causal odds / (causal odds + 1)

e.g. Causal odds of 19 gives Causal probability of 95%, $0.95 = 19 / (19+1)$

Incorporating additional information

$$\frac{P(C | S)}{P(\text{not } C | S)} = \frac{p(C).p(S | C)}{p(\text{not } C).p(S | \text{not } C)}$$

- How do we include additional categorical information (e.g. inheritance status)
 - Variant is *de novo*
 - Variant rare, and inheritance status unknown

Denote information with F, (Filter):

$$\frac{p(C | S,F)}{p(\text{not } C | S,F)} = \frac{p(C).p(S | C,F).p(F | C)}{p(\text{not } C).p(S | \text{not } C,F).p(F | \text{not } C)}$$

From probabilities to rules

TABLE 3. Proposed Classification System for Sequence Variants Identified by Genetic Testing

Class	Description	Probability of being pathogenic
5	Definitely pathogenic	> 0.99
4	Likely pathogenic	0.95–0.99
3	Uncertain	0.05–0.949
2	Likely not pathogenic or of little clinical significance	0.001–0.049
1	Not pathogenic or of no clinical significance	<0.001

Sequence variant classification and reporting:
 Recommendations for improving the interpretation of cancer susceptibility genetic test results
 Plon et al (2008) Human Mutation 29:1282-1291
IARC Unclassified Genetic Variants Working Group

“We believe that a Bayesian system to generate a posterior probability should ultimately be the standard for all variants.”

Probability	ACMG draft classification	3 tier classification
>0.99	Pathogenic	Pathogenic
0.95-0.99	Uncertain: likely pathogenic	Uncertain
0.5-0.95	Uncertain	
0.001-0.5	Uncertain: likely benign	
<0.001	Benign	Benign

CMGS guidelines on Unclassified Variants

4.13 Value of a Bayesian approach.

A Bayesian approach combining different strands of evidence can be used to produce a final probability of pathogenicity (Goldgar *et al.*, 2004). However there are currently no recognised statistical tools available to diagnostic laboratories for this purpose.

Generic: not inherently monogenic

Causal hypotheses can be varied:

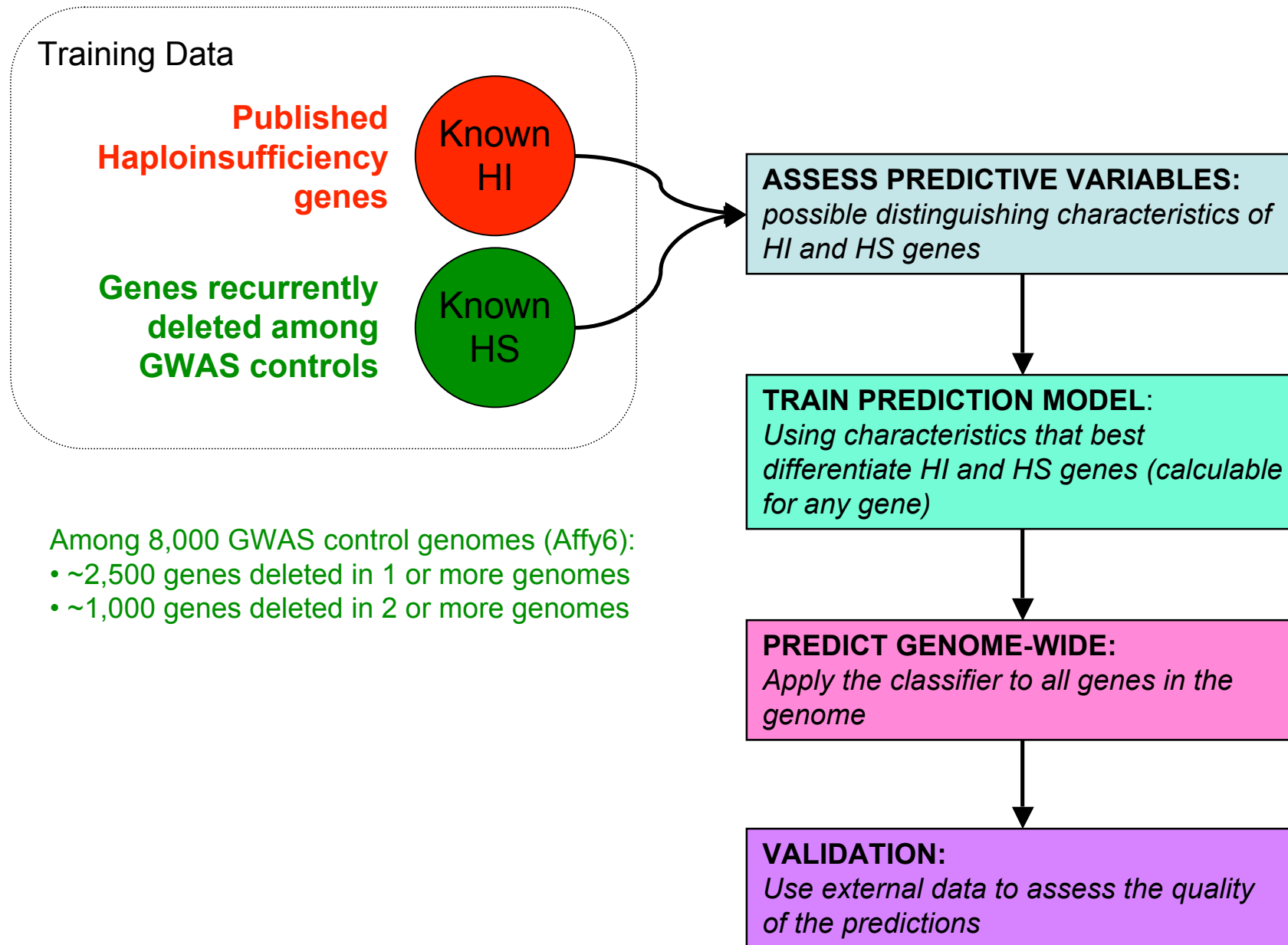
- **Monogenic:**
 - *“Is this variant causal?”*
- **Oligogenic:**
 - *“Is this set of variants causal?”*
 - **Known oligogenic diseases:**
 - Are these three rare variants in the 14 known Bardet-Biedl Syndrome genes causal?
 - **Two-hit CNV hypothesis:**
 - Are these two rare CNVs causal?
 - **LOF burden hypothesis (from exome sequencing):**
 - Are the set of rare LOF variants causal?

Application to non-recurrent CNVs

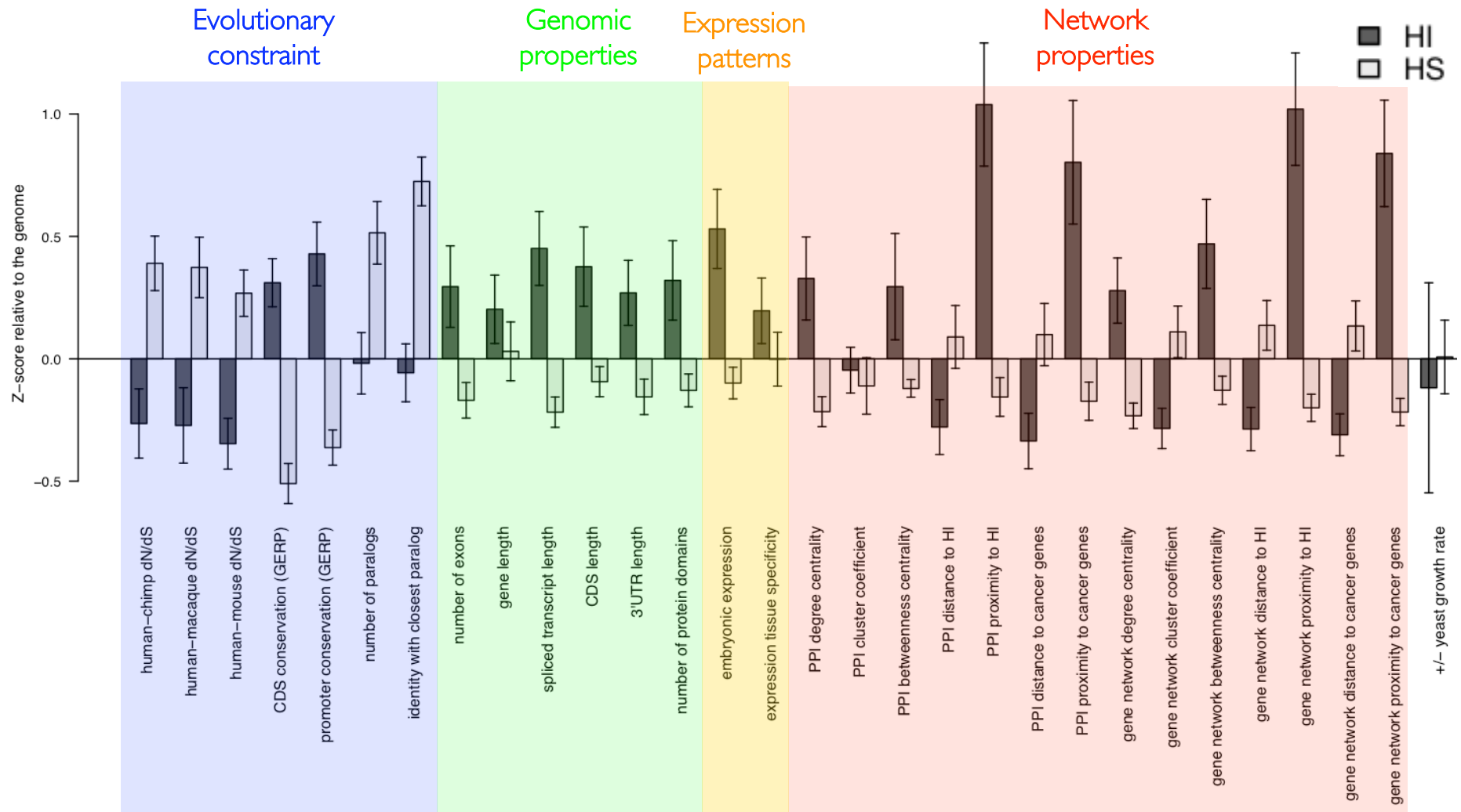
Predicting haploinsufficiency

- Haploinsufficiency: *a single functioning copy of a gene is insufficient to maintain normal function*
 - Major mechanism of dominance (SNPs, indels, CNVs)
 - Predominant pathogenetic mechanism of rare CNVs
- Functional, Genomic and Evolutionary properties of haploinsufficient (HI) and haplosufficient (HS) genes are likely to be different
- CNV maps in GWAS controls identify genomic regions in which haploidy can be tolerated without severe phenotypic consequences
- Mapping CNVs onto gene structures reveals which CNVs cause Loss-Of-Function of overlapping genes:
 - Whole gene deletions, partial gene deletions and duplications

Prediction framework (machine learning)



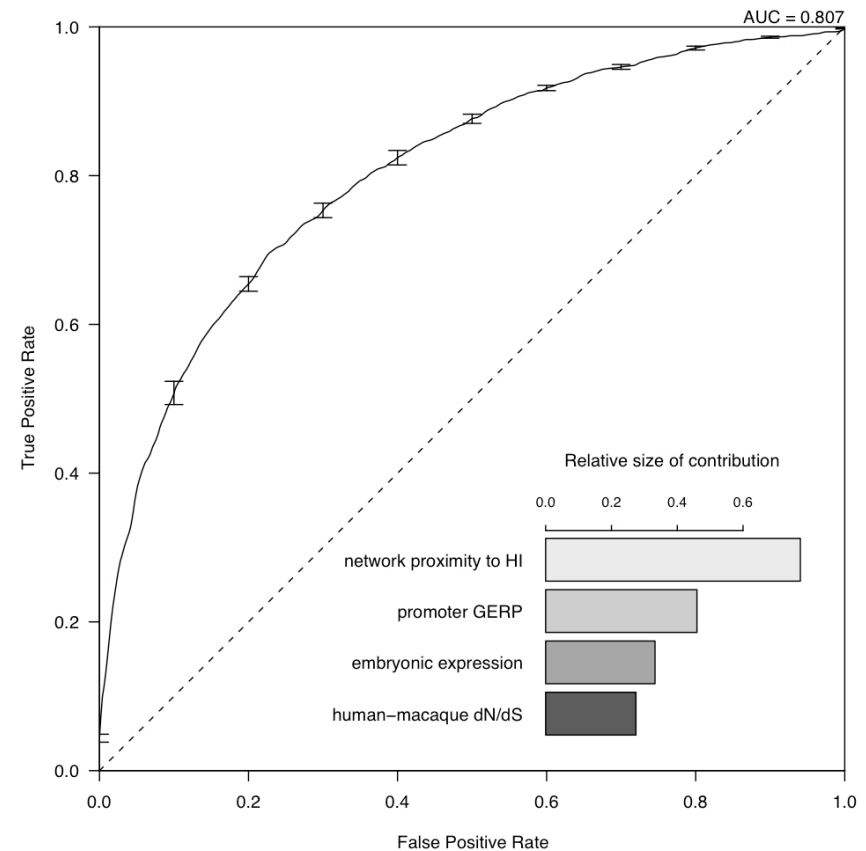
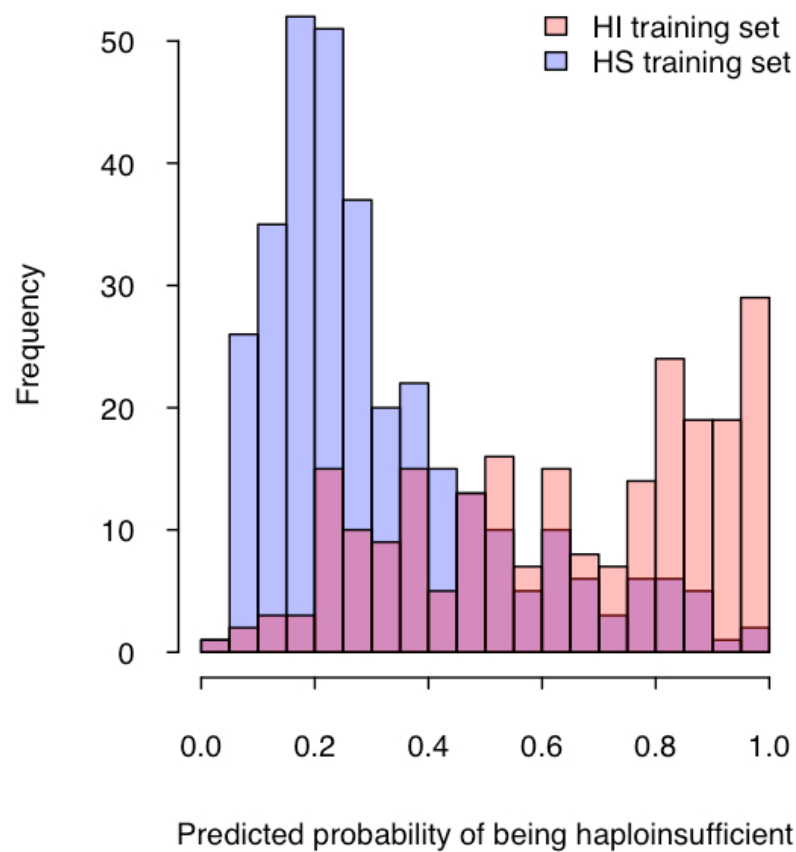
What factors distinguish HI and HS genes?



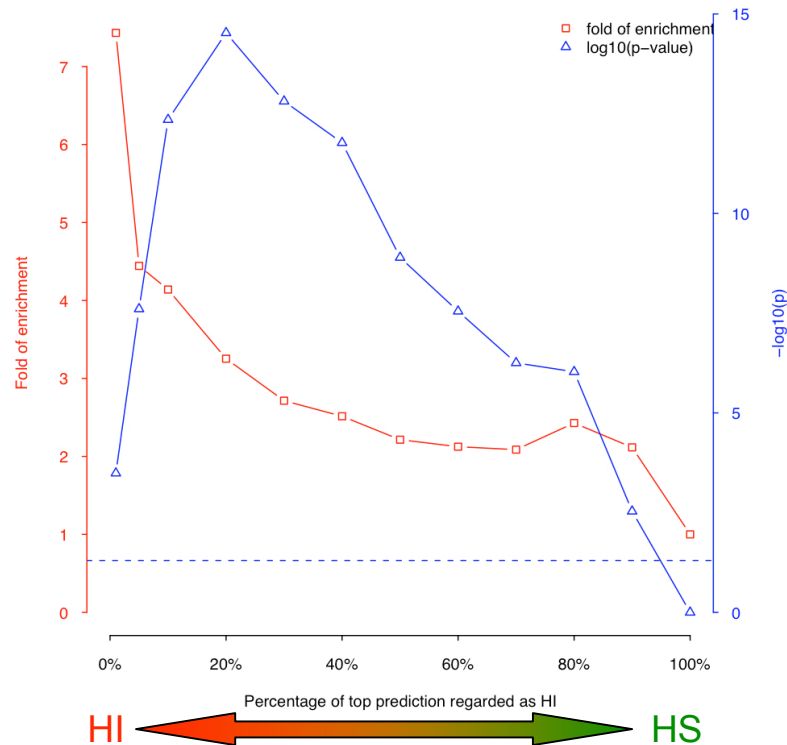
How predictable is haploinsufficiency?

Prediction Model:

1. Protein conservation (human/macaque Ka/Ks)
2. Promoter conservation (GERP score)
3. Embryonic expression
4. Network proximity to known HI gene

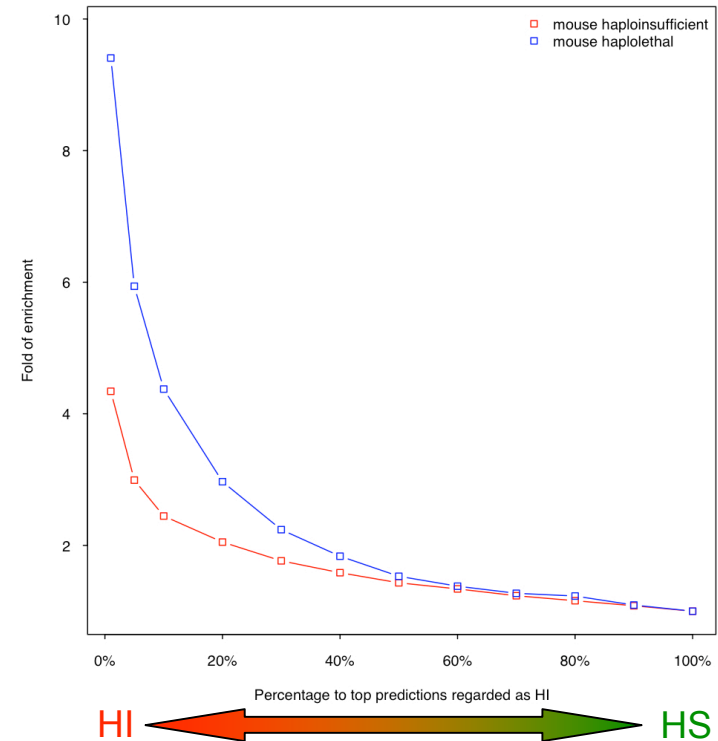


Validating HI predictions



OMIM annotation

>4x enrichment of dominant annotations vs recessive annotations in top 10% p(HI)

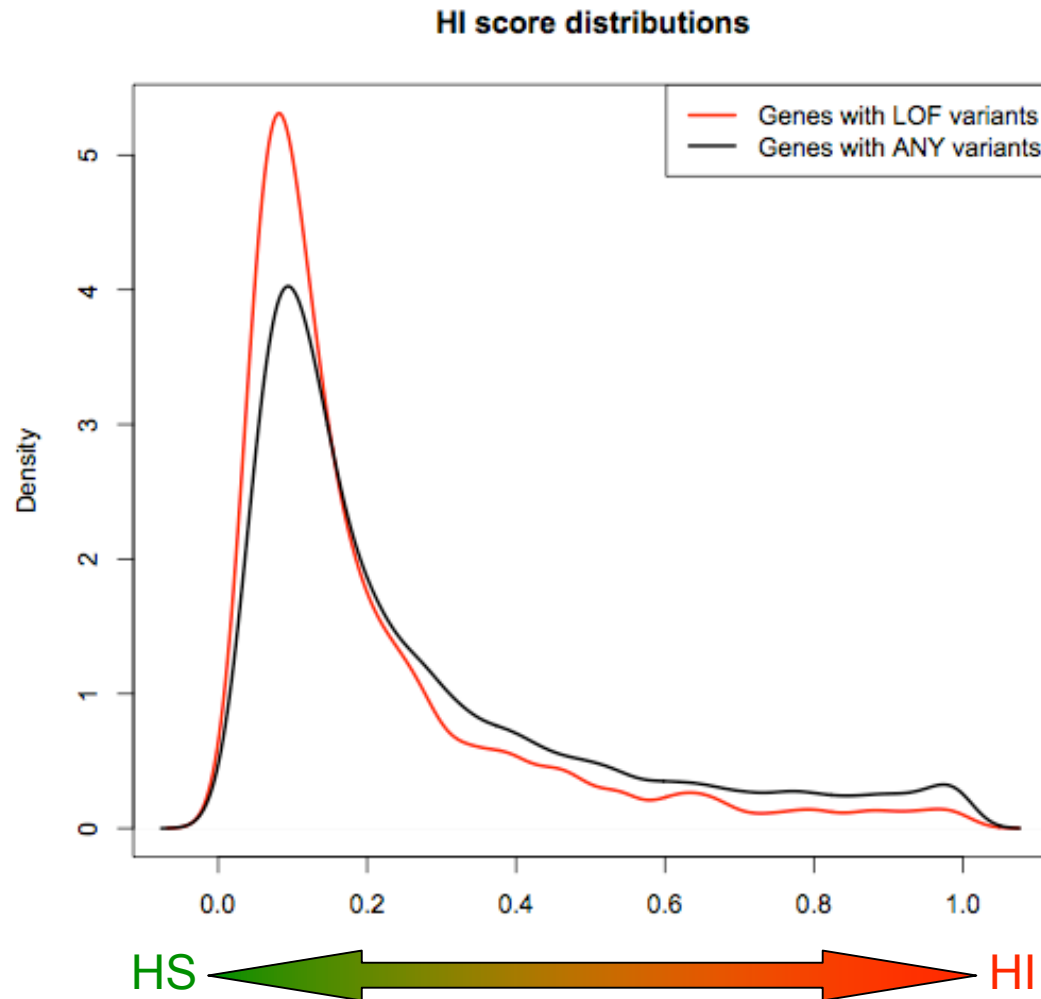


Mouse phenotypes

~2.5X enrichment of genes annotated with phenotypes in heterozygous KOs in top 10%

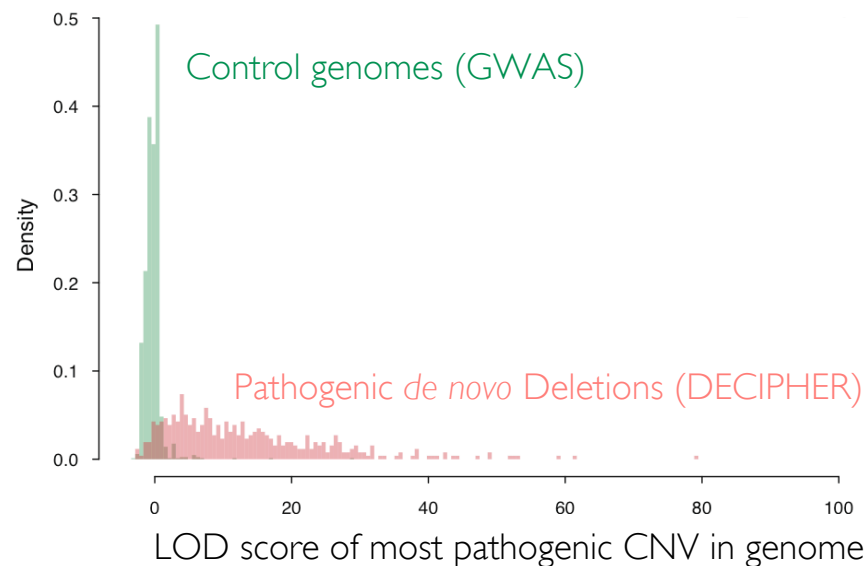
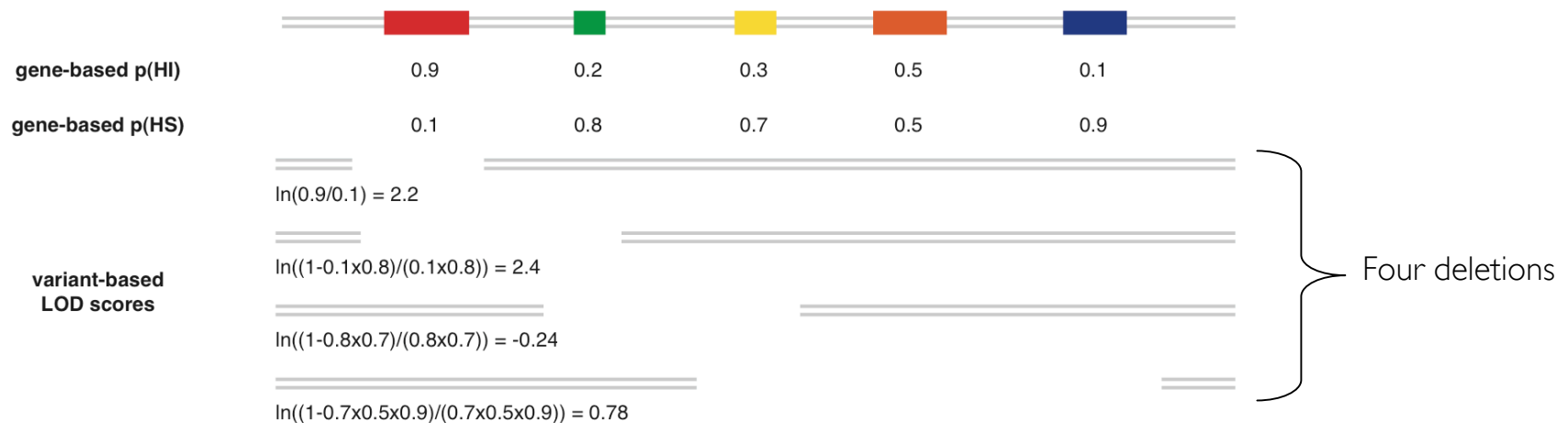
~4.5X enrichment of genes annotated as causing embryonic lethality in top 10%

Genes with LOF variants (nonsense, essential splice sites) in population controls are significantly ($p < 10^{-16}$) less likely to be predicted to be haploinsufficient

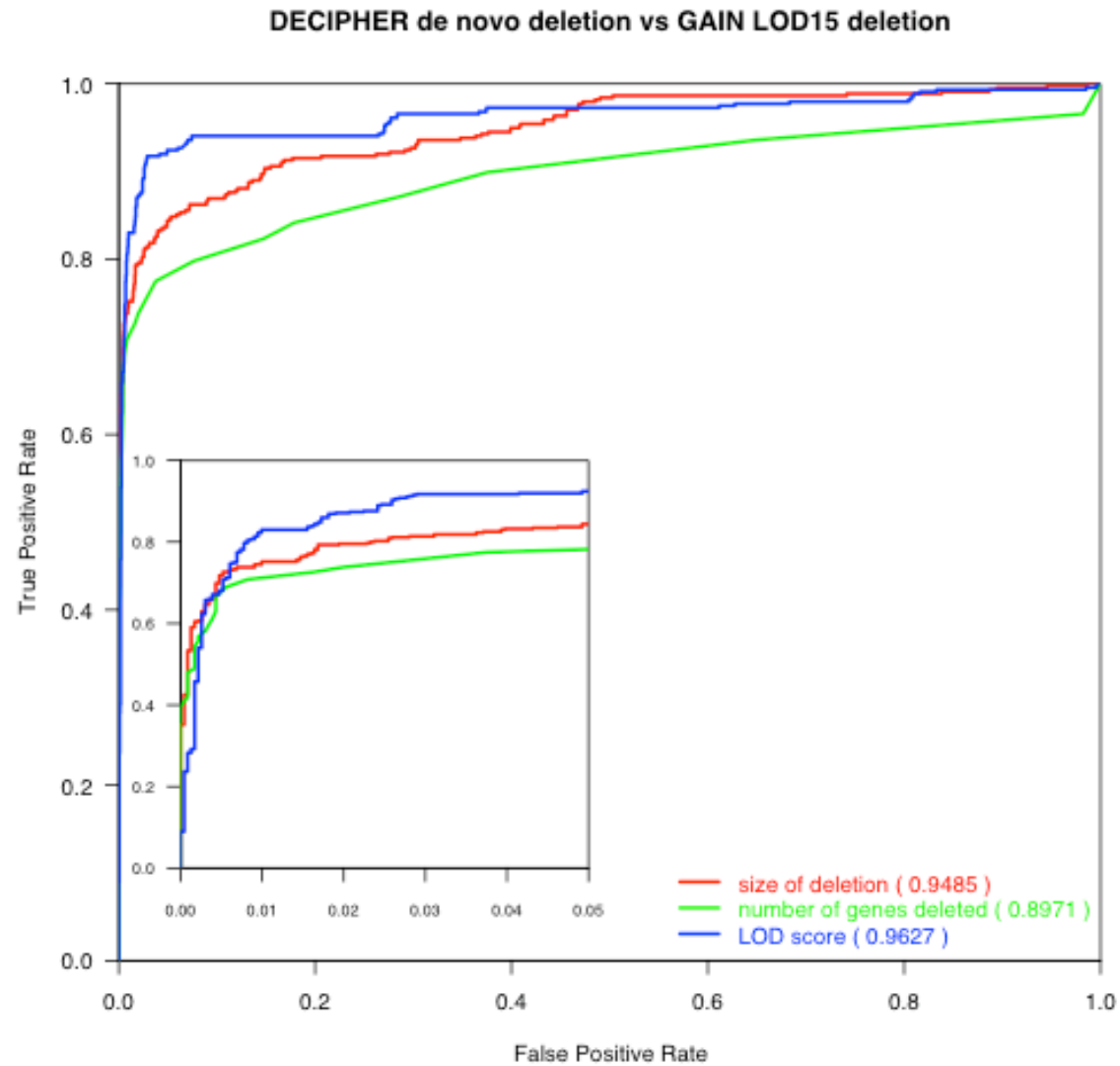


CNV Haploinsufficiency (HI) Score

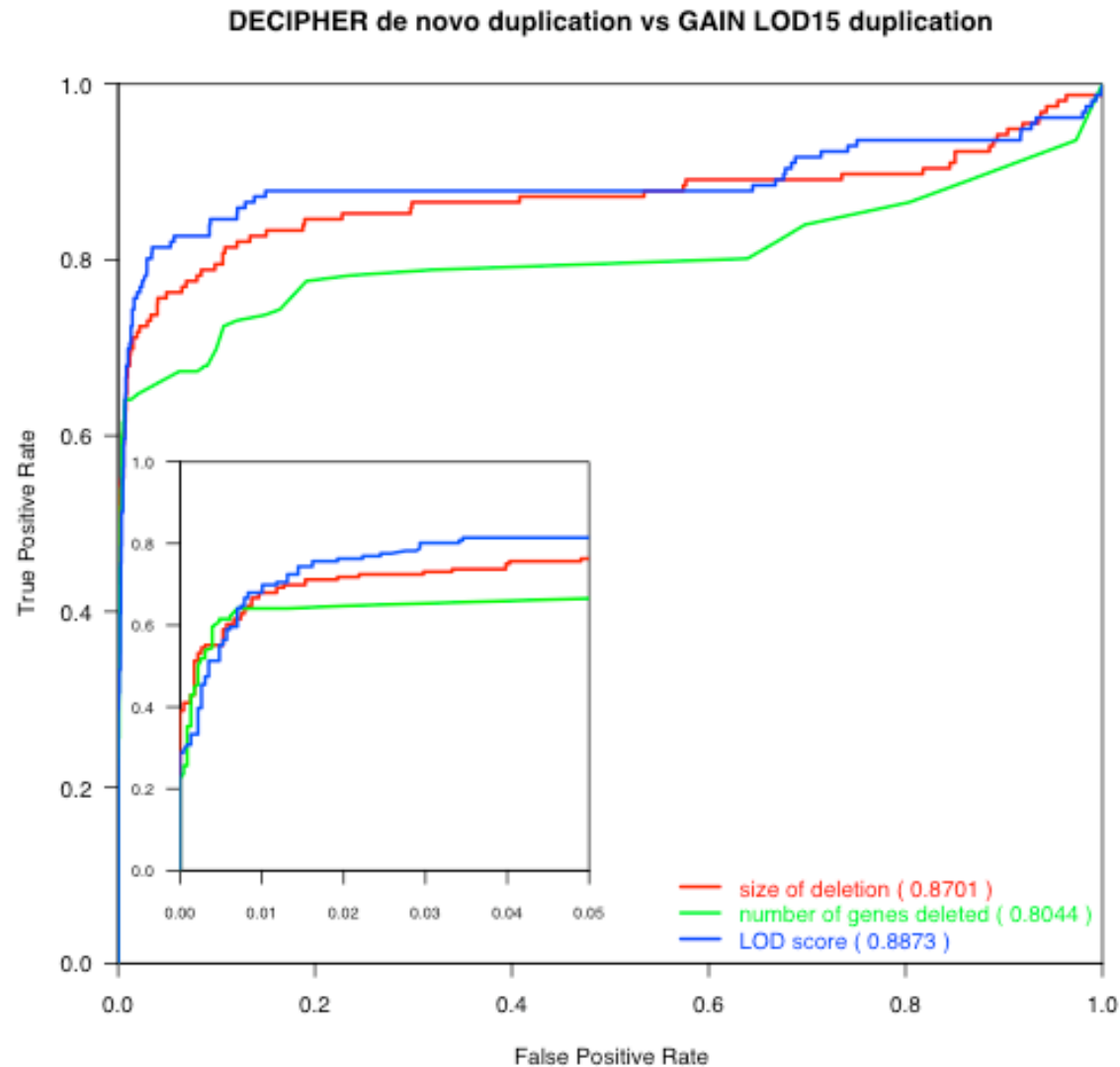
HI odds: $P(\text{HI genes} \geq I) / P(\text{HI genes} = 0)$



Genic deletions: LOD score better than length, number of genes



Genic duplications: Haploinsufficiency \approx Dosage-sensitivity



Can we estimate the key quantities?

F: variant is *de novo* rare genic deletion

$$\frac{p(C | S,F)}{p(\text{not } C | S,F)} = \frac{p(C) \cdot p(S | C,F) \cdot p(F | C)}{p(\text{not } C) \cdot p(S | \text{not } C,F) \cdot p(F | \text{not } C)}$$

$p(C)$: prior of causality: what proportion of patients with this phenotype have this kind of causal variant and how many candidate variants are seen, on average, per patient

$P(\text{not } C)$: $1 - p(C)$

$p(S | C,F)$: estimated from the distribution in known causal variants (e.g. DECIPHER *de novo* deletions)

$p(S | \text{not } C,F)$: estimated from the distribution in known non-causal variants (e.g. approximate: singleton variants in GWAS controls). THOUSANDS AVAILABLE

$p(F | C)$: what fraction of known causal variants are *de novo* (several estimates in literature)

$p(F | \text{not } C)$: what fraction of known non-causal variants are *de novo* (few imprecise estimates in literature)

Causality as a function of DELETION size or LOD score

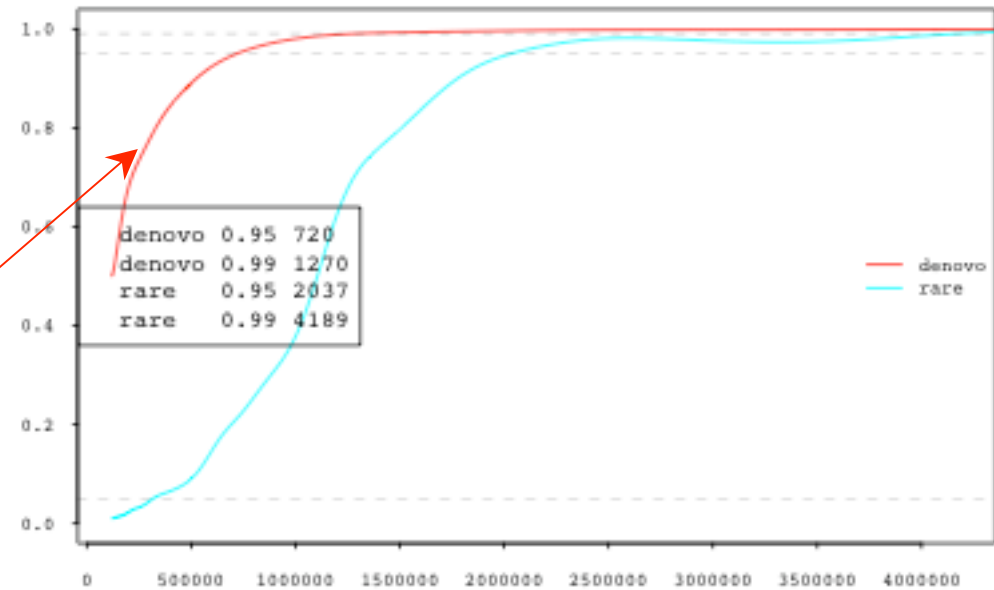
Caveats:

Only look at deletions >100kb to ensure comparability between case and control data

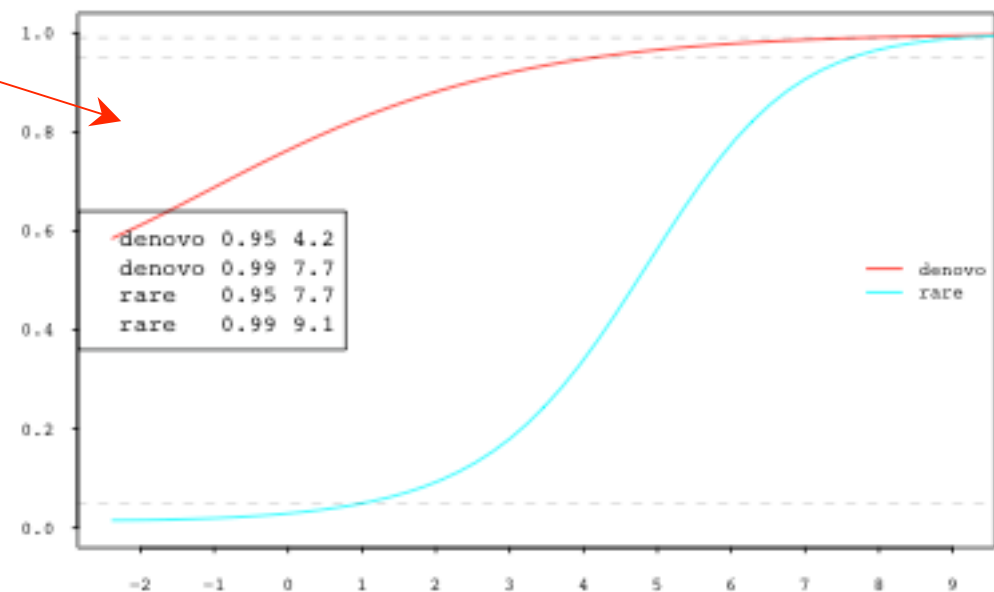
Being *de novo* grants a high probability of a DELETION being causal, but it alone is not sufficient to reach a confidence level of 95%

The probability that a rare DELETION of high pathogenicity but unknown inheritance status is causal, can be as high as a *de novo* DELETION of lower pathogenicity score

Posterior probability as a function of size



Posterior probability as a function of LOD



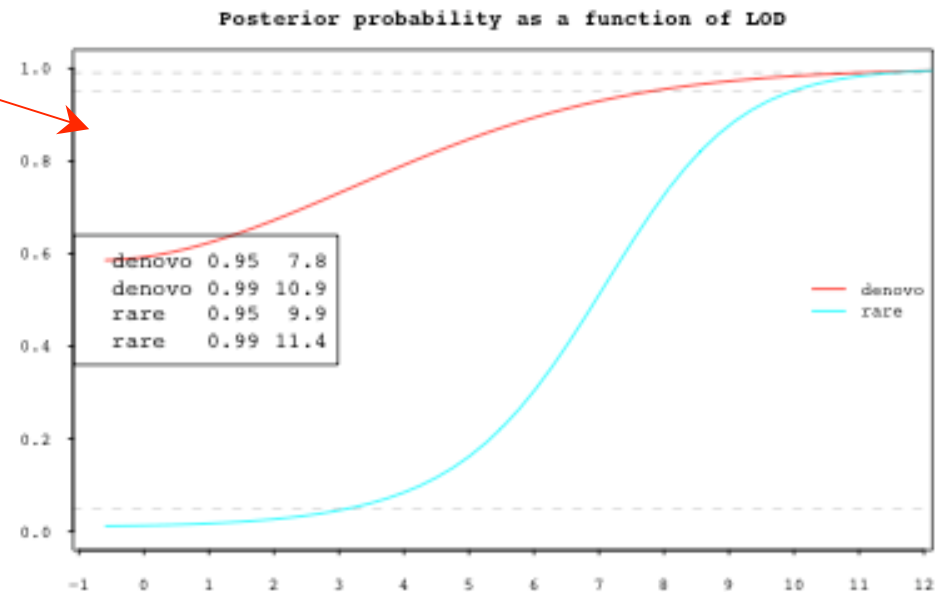
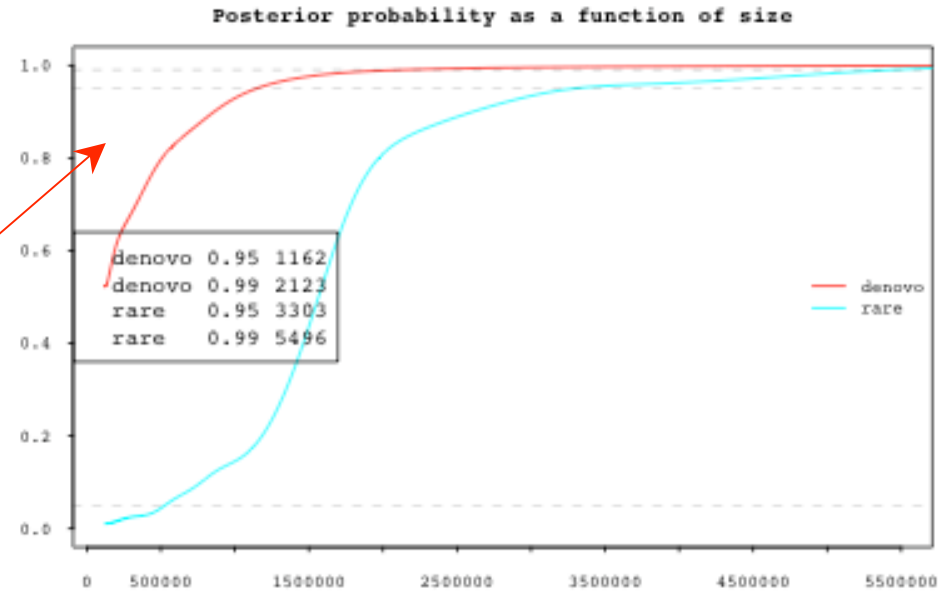
Causality as a function of DUPLICATION size or LOD score

Caveats:

Only look at duplications >100kb to ensure comparability between case and control data

Being *de novo* grants a high probability of a DUPLICATION being causal, but it alone is not sufficient to reach a confidence level of 95%

The probability that a rare DUPLICATION of high pathogenicity but unknown inheritance status is causal, can be as high as a *de novo* DUPLICATION of lower pathogenicity score



Future Work / Wrinkles

- Estimating $p(S | C)$ and $p(S | \text{not } C)$ in different ways gives (slightly) different results
- Ought to integrate uncertainty in input parameters (fully Bayesian)
- Need to benchmark against and implement in large databases of clinical samples (DECIPHER, Signature, BCM, ISCA, etc)
- Not ready for clinical use - but not far off
- Explore added value of more phenotype-specific pathogenicity scores

Challenges:

- Estimating penetrance
- Extend to non-genic variation (need non-gene-based pathogenicity score)

More generally:

Careful thought needed when testing multiple different hypotheses from genome-wide sequencing data (CNV, non-CNV, missense variants, LOF variants, monogenic, oligogenic)

Take home messages

- **Generic Bayesian diagnostic framework aids interpretation of VUCS**
 - Combines clinical judgement and statistical assessment of an explicit causal hypothesis
 - Naturally incorporates rapidly growing population variation data
 - As input information improves, interpretation improves:
 - More knowledge of genetic architecture of a disease => better priors
 - More training data, better algorithms => improved pathogenicity scores
 - Larger databases of causal and non-causal variants => more accurate probabilities
 - Provides simple interpretation of complex data, either Causal Prob. > threshold, or not
- **Need to collate large numbers of causal variants, to estimate $p(S | C)$**
 - Available for CNVs (DECIPHER, ISCA, etc)
 - Not easily accessible for sequence variants (LSDBs)
 - Improved interpretation of sequence variants will lag interpretation of SVs
 - Needs to be accessible to all, competition is good
- **Collating causal variants is not sufficient to make causal inference!**
 - Need large-scale population data (being generated)
 - Sensitivity and specificity of patient and control data need to be well matched
 - Population projects need to make data available in useful formats

Acknowledgements

- *Predicting Haploinsufficiency*
 - Ni Huang, Insuk Lee, Edward Marcotte
- *Inferring CNV causality*
 - Ni Huang
- *Really useful comments:*
 - Nigel Carter, Helen Firth, David FitzPatrick, David Clayton, Chad Shaw

Questions/Comments