CHAPTER 4

CHARACTERIZING AND PREDICTING HAPLOINSUFFICIENCY IN THE HUMAN GENOME

4.1 Introduction

Haploinsufficiency, wherein a single functional copy of a gene is insufficient to maintain the normal phenotype of a diploid organism, is a major cause of human dominant diseases.

Dominance and recessiveness are fundamental concepts of Mendelian genetics. They describe the relationship between a pair of alleles of a gene of a diploid organism with respect to the phenotype they manifest. An allele, A, is dominant to another allele, a, if the corresponding phenotype of Aa is different from aa but indistinguishable from AA. A mutation can be described as dominant or recessive if it is dominant or recessive to the wildtype allele. The majority of observed naturally occurring (deleterious) mutations are recessive. While Fisher explained this as the result of selection for modifier genes that increase the fitness of heterozygotes [105], Wright viewed it as simply a physiological consequence of metabolic pathways [106]. Experimental and theoretical work over the years suggested Wright's explanation is more plausible. Kacser and Burns [107] established an excellent math-

ematical framework for understanding dominance/recessiveness at the molecular level and they showed that recessiveness emerges naturally from the kinetic properties of multi-enzyme system when most enzymes are far from being saturated.

The dominant mutations and the genes that harbor these mutations, though being the minority, contribute to a disproportionate ∼48% (965/2006) of human autosomal Mendelian disorders with known molecular basis recorded to date [108]. Wilkie categorized the molecular mechanisms of dominance into eight types [109], including haploinsufficiency, increased gene dosage, ectopic or temporally altered expression, increased or constitutive protein activity, dominant negative effect, altered structural protein, toxic protein alterations and new protein function. Among those types, haploinsufficiency is especially interesting, since (*i*) it is a relatively common mechanism for dominant diseases as a variety of mutations can lead to heterozygous loss-of-function; (*ii*) the ascertainment of loss-of-function mutations is relatively easy compared to gain-of-function mutations; (*iii*) the direct impact is solely through dosage reduction, which is easier for functional interpretation than other types of dominant mutation; (*iv*) it can be regarded as a property of a gene as the mutant allele is always defunct irrespective of the specific mutation. From a theoretical perspective, Veitia showed that haploinsufficiency is more likely to occur in systems that require the physical interaction of distinct macromolecules such as transcription regulation and assembly of protein complexes, in which the total output of the system is a sigmoid function of the dosage of each single entity [110]. From a more biological perspective, Wilkie suggested that genes encoding structural proteins are required in large quantities in specific tissues, and that subunits of protein complexes assembled under strict stoichiometry and regulatory proteins working close to a threshold level for different actions are more likely to be haploinsufficient [109]. Examples of these types include type 1 collagen [111], ribosomal proteins [112] and members of the *Hox* gene family [113].

Around three hundred genes have been reported haploinsufficient in human so far and Dang *et al* showed that they are less likely, compared to the rest of the genes, to be located in genomic regions susceptible to structural rearrangements [14]. This is expected, as large genomic deletions, a frequent consequence of structural rearrangements, are a major type of loss-of-function (LOF) mutation. Deletions en-

compassing the entire length of a gene unambiguously reduce the number of its functional copies. Partial deletions can also cause LOF, if key elements involved in the initialization of transcription, splicing and translation, such as promoter, splicing signals and start codon, are affected. Even if those elements are intact, premature stop codons could be introduced by frame-shifting deletions or simple truncating deletions, which likely subject the transcripts to nonsense-mediated decay, by which these transcripts are digested rather than translated into mutant proteins [114]. Indeed, large deletions have been found to be causal for diverse dominant developmental disorders, which, in turn, has led to the discovery of a number of haploinsufficient genes (HI genes), for example the discovery of the CHARGE syndrome gene, *CHD7* [115].

However, not all LOF mutations are deleterious. It is clear from sequenced genomes [116], exomes [117] and CNV surveys [12] that every genome, including those of apparently healthy individuals studied as controls in disease studies, harbors tens of unambiguous LOF mutations, including large genomic deletions. Some LOF mutations can be even advantageous [118]. Genes deleted in apparently healthy individuals seem not to be haploinsufficient, at least not to the point that carriers of heterozygous LOF mutations in these genes are kept from being recruited as controls for disease studies. Besides these haplosufficient (HS) genes, and the currently known HI genes, the dosage sensitivity of the majority of the genome remains elusive. Previous studies have shown that sets of HI genes, such as genes implicated in dominant diseases, have biased evolutionary and functional properties with respect to the rest of the genome [119–121]. However, there has not been a direct and systematic investigation of differences in properties between known HI genes and haplosufficient (HS) genes and it is unknown which properties are most informative in predicting dosage sensitivity.

With array-based copy number detection and the current generation of sequencing technologies, our ability to discover genetic variants in patients is running far ahead of our ability to interpret their functional impact and there is a pressing need to distinguish between benign and pathogenic variants. Computational methods have been developed to predict the molecular impact of non-synonymous point mutations. Some totally depend on sequence conservation at the site of the mutation,

such as SubPSEC [122], Align-GVGD [123] and SIFT [124]. Some also consider structural and biochemical properties of the protein (stability, solubility, active sites, etc), such as SNPs3D [125] and PolyPhen [126]. The output of these algorithms is often a continuous score or a category label indicating how damaging the mutation is to the encoded protein. Although, these outputs have been shown to be useful in identifying pathogenic mutations for Mendelian diseases [127], their power to predict impact on fitness at individual level might still be limited, especially in the case of heterozygous mutation wherein one of the alleles still functions normally, as they do not distinguish between the heterozygous and homozygous genotypes of a variant. Computational tools for predicting the functional impact of large copy number variants are still in their infancy [128]. The problem differs from non-synonymous point mutations in that large CNVs can affect multiple genes as well as non-coding regions simultaneously, and thus their interpretation requires the integration of different functional annotations to maximize the information on all affected entities.

Application of such computational interpretative tools in clinical settings requires careful consideration, as these tools are usually trained on collated sets of known damaging and benign mutations that could well be a biased representation of the true spectrum of causal mutations found in real patients or in the general population. The scores or classifications generated by these computation tools are rarely calibrated to diagnostic outcomes, and only infrequently are the distributions of such scores compared between patients and population controls. Characterizing the distribution of such scores in patient and population cohorts has become more feasible in recent years with the growth in databases of pathogenic variants [129, 130] as well as of variants found in large population surveys [12, 34, 77, 78, 131, 132]. Additionally, pathogenicity scores are often just one of the many different types of evidence that influence diagnostic interpretation and needs to be integrated with the other evidence in a sensible way. Most current genetic diagnostic practices adopt a decision-tree-like procedure [133, 134]. A probabilistic process would be desirable which could give every diagnosis a level of confidence. Goldgar *et al* suggested a naïve Bayesian framework to integrate different, typically uncorrelated, types of information and demonstrated its application to the interpretation of variants of unknown clinical significance in the *BRAC1* and *BRAC2* genes [135].

In the work described in this chapter, I first explored the genomic, functional and evolutionary characteristics of HI genes and then I developed a computational approach to predict which genes might exhibit haploinsufficiency. I then investigated the utility of the gene-based HI predictions to measure pathogenicity of large copy number variants, both deletions and duplications. Finally, I proposed a probabilistic diagnostic framework that integrates population distributions of pathogenicity scores, with additional evidence to generate a level of confidence for the diagnosis of causal CNVs, and potentially other forms of genetic variants.

4.2 Materials and methods

4.2.1 Control data

The controls include a set of 6,000 UK individuals recruited as common controls in GWAS of 13 disease conditions undertaken by Wellcome Trust Case Control Consortium 2 (WTCCC2), of which 3,000 samples are from the 1958 British Birth Cohort and 3,000 samples are from the UK Blood Service Control Group. Another set of 2,421 US control individuals, 1,442 of which have European ancestry and the rest with African-American ancestry, are from a control cohort used in GWAS of Schizophrenia and Bipolar disease undertaken by Genetic Association Information Network (GAIN). Samples were previously genotyped on Affymetrix genome-wide human SNP array 6.0. Affymetrix 6.0 CEL files were obtained from Wellcome Trust Case Control Consortium 2 for WTCCC2 controls and from the Database of Genotype and Phenotype (dbGaP) through accession number phs000017 and phs000021 for GAIN controls.

4.2.2 Asserting of loss of function genes

To identify protein-coding genes disrupted in a LOF manner, CNV calls made by the calling pipeline described in Chapter 2 were compared to gene annotation provided by EnsEMBL [136]. Four scenarios were considered LOF to a protein-coding transcript:

1. deletion of over 50% of coding sequence
2. deletion of the start codon or the first exon
3. deletion-disrupted-splicing
4. deletion-caused frame-shift

A gene was considered LOF if all of its transcripts were LOF. Under these criteria, CNVs were identified in GWAS control individuals with a LOF impact on 2,677 genes. I defined haplosufficient genes as being those observed as LOF genes in two or more GWAS control individuals.

4.2.3 Preparing possible predictor variables

4.2.3.1 Genomic properties

The length of gene, spliced transcript, 3'UTR and coding sequence and the number of exons were calculated on the basis of gene annotation downloaded from EnsEMBL. The number of protein domains was retrieved from EnsEMBL build 50.

4.2.3.2 Evolutionary properties

dN/dS data was downloaded from EnsEMBL. Genomic Evolutionary Rate Profiling (GERP) [137] score was downloaded from EBI. Two summed GERP values, one for coding sequence and the other for promoter region, defined as bases within $\pm$100bp of the transcription start site, were then calculated for all human protein-coding transcripts according to EnsEMBL annotations and summarized by gene using the median values. A third summed GERP value for conserved noncoding elements around genes was calculated as the sum of GERP scores of all bases of annotated conserved noncoding elements within an interval $\pm$50kb of the gene. To derive the list of conserved noncoding elements, I retrieved a list of conserved elements throughout placental mammals from the UCSC genome browser (28-Way

Most Cons track) and removed elements overlapping with exons according to EnsEMBL gene annotation. The number and identity of paralogs were downloaded from EnsEMBL.

4.2.3.3 Functional properties

Gene expression profiles in human were obtained from the GNF Atlas [138]. Total expression levels were normalized across genes and the standard deviation of expression across normal tissue types of each gene was used to indicate its tissue specificity of expression. Genes over-expressed by at least 8 fold in human embryonic stem cells [139], fetal tissues [138] and mouse fetal tissues [140] were collectively treated as genes expressed at embryonic stage. A binary coding was used to represent this property in which genes expressed at embryonic stage were labeled 1 and the rest were labeled 0.

4.2.3.4 Network properties

Two interaction networks were used. One is a binary protein-protein interaction network integrated from a number of sources [141–145]. Proteins were mapped to their coding genes and interactions were not counted repeatedly if multiple proteins were mapped to a single gene. This network included 70,632 interactions among 11,077 genes. The other is a probabilistic gene interaction network (a network of 470,217 links among 16,375 human genes calculated using methods previously described for yeast [146] and worm [147] and derived from 22 publicly available genomics datasets including DNA microarray data, protein-protein interactions, genetic interactions, literature mining, comparative genomics, and orthologous transfer of gene-gene functional associations from fly, worm, and yeast, where the weight of a link is the log likelihood score of the interaction [146]. Measures of centrality (degree, betweenness) and modularity (cluster coefficient) were calculated using MCL [148]. Shortest path distance and sum of weight of interactions [147] were calculated as measures of proximity to a group of 'seed' genes.

4.2.3.5 Other properties

A list of 300 genes implicated in cancer was downloaded from the COSMIC database [149]. Growth rate of yeast heterozygous deletion strains were from Deutschbauer *et al* [150].

4.2.4 Comparing predictor variables between HI and HS genes

For continuous variables, the two-tailed Mann-Whitney U test was performed to assess if positive (haploinsufficient) and negative (haplosufficient) training data have the same median value for potential predictor variables. For two-class categorical features, Fisher's exact tests were performed. Statistical tests were performed using R (`http://www.r-project.org`).

4.2.5 Feature selection for the predictive model

I assessed different potential sets of predictor variables for input into the predictive model using the following criteria: (*i*) they allow prediction for at least half the genes in the genome, (*ii*) the Spearman correlation ρ^2 between all pairs of predictor variables is less than 0.05, (*iii*) they are drawn from different broad categories (genomic, evolutionary, functional and network) if possible, and (*iv*) achieve best performance in model assessment.

4.2.6 Assessing model performance

The *sensitivity* of the prediction was plotted against $1 - specificity$ and the area under the ROC curve (AUC) [151] was used as quantitative measure of the performance of the model, where $sensitivity = TP/(TP + FN)$, and $specificity = TN/(TN + FP)$. The other measure used is the Matthews correlation coefficients (MCC) [152], defined as:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

To avoid over-fitting, the sensitivity and specificity were calculated using 10-fold cross-validation. To overcome the variability caused by random partition involved in 10-fold cross-validation, each such assessment was repeated 30 times and the mean values were reported.

4.2.7 Multiple imputation

Multiple imputation was used to fill in ('impute') the missing values for predictor variables incorporated in the model, namely 'dN/dS ratio between human and macaque', 'promoter conservation (GERP)', and 'gene network proximity to HI genes', except for 'embryonic expression' of which the genomic coverage is 100%. Since 'gene network proximity to HI genes' and 'promoter conservation (GERP)' are the top two predictive variables, genes missing both values were removed. To achieve better imputation, I included three additional gene properties, namely 'CDS conservation (GERP)', 'spliced transcript length' and 'gene network betweenness centrality' in the imputation process. Twenty independent imputations of 20 iterations were undertaken. In each iteration, imputation for each predictor variable was in the order of increasing number of missing values using the predictive mean matching method. The computation was done using the R package MICE [153].

4.2.8 Parameter estimation for the Bayesian diagnostic framework

The prior probability of a CNV being causal ($\mathrm{p}(C)$) was estimated as the average number of CNVs found per individual divided by the current diagnostic rate for CNVs. Diagnostic rate and average number of CNVs found per individual were taken from Buysee *et al* [134], which found on average 0.86 deletions and 0.73 duplications per individual and achieved a diagnostic rate of 0.1 using BAC array and Agilent 44K array CGH.

The probability of a causal CNV being rare (population frequency $< 1\%$) ($\mathrm{p}(F|C), F =$ rare) was set at 1. The probability of a causal CNV being *de novo* ($\mathrm{p}(F|C), F =$ *de novo*) was also taken from Buysee *et al* [134] in which 73% of the causal CNVs found were *de novo*. The distribution of pathogenicity scores of *de novo* CNVs in DE-

CIPHER [129] was used to approximate that of causal CNVs. The probability of a causal and rare (or *de novo*) CNV having a pathogenicity score equals to x was taken as the empirical estimation of probability density of the distribution of pathogenicity scores of causal CNVs at x.

The probability of a benign CNV being rare (population frequency $< 1\%$) ($\mathrm{p}(F|\bar{C}), F =$ rare) was estimated as the fraction of WTCCC2 and GAIN control CNVs with a carrier frequency $< 1\%$. The probability of a benign CNV being *de novo* ($\mathrm{p}(F|\bar{C}), F =$ *de novo*) was also taken from Itsara *et al* [131] in which 0.44% of the CNVs found in children of apparently healthy trios were *de novo*. The distribution of pathogenicity scores of benign CNVs was generated using WTCCC2 and GAIN control CNVs after excluding CNVs at known pathogenic loci recorded in DECIPHER. The probability of a benign and rare (or *de novo*) CNV having a given pathogenicity score equals to x was taken as the empirical estimation of probability density of the distribution of pathogenicity scores of benign CNVs with a carrier frequency $< 1\%$ (or with an occurrence of 1, *i.e.* singletons) at x. Since WTCCC2 and GAIN control CNVs were discovered using arrays of considerably higher resolution than the CNVs discovered by Buysee *et al* and the CNVs recorded in DECIPHER, deletions <180kb and duplication <330kb were excluded prior to the above calculation in order to match the number of CNVs discovered per individual.

4.2.9 Text mining through PubMed abstracts

The title and abstract of publications that contain the keyword 'haploinsufficiency' or 'haploinsufficient' were retrieved from PubMed on Aug 2010, using the search term 'haploinsufficient[Title/Abstract] OR haploinsufficiency[Title/Abstract] AND humans[MeSH Terms]'. After cleaning the text, a word frequency table was compiled from all titles and abstracts. A dictionary that maps gene names and synonyms to gene symbols was downloaded from HGNC [154]. For each title and abstract, the sentence containing the keyword 'haploinsufficiency' or 'haploinsufficient' was extracted and parsed by the GENIA tagger [155] to break the sentence into chunks and tag the part-of-speech of each chunk. The chunk immediately before the keyword, the noun chunk in front of a verb and a preposition in front of the keyword were extracted. These chunks were first examined by GENIA tagger to identify the named biomedical entity. If this failed, the noun in the chunk that appeared

fewer than 10 times as recorded in the frequency table and contained numbers or capital letters, or followed immediately by ‘gene’, ‘protein’ or ‘transcript’ was kept as potential gene name. These potential gene names and named entities identified by the GENIA tagger were looked up in the gene name dictionary to convert into unique HGNC gene symbols.

4.3 Results

4.3.1 Characteristics of haploinsufficient genes

I first compiled a list of known human HI genes and a catalog of HS genes. Known HI genes were collated from literature [14, 156]. The catalog of HS genes was generated from genes disrupted in a loss-of-function manner in control individuals used in genome-wide association studies by CNVs detected in data from the Affymetrix 6.0 chip (see Methods). I identified 2,676 putative HS genes seen in any control individuals and 1,079 seen in two or more controls (Figure 4.1), and used the latter set in most downstream analyses. Thus the final list of HI and HS genes contains 301 and 1,079 genes respectively.

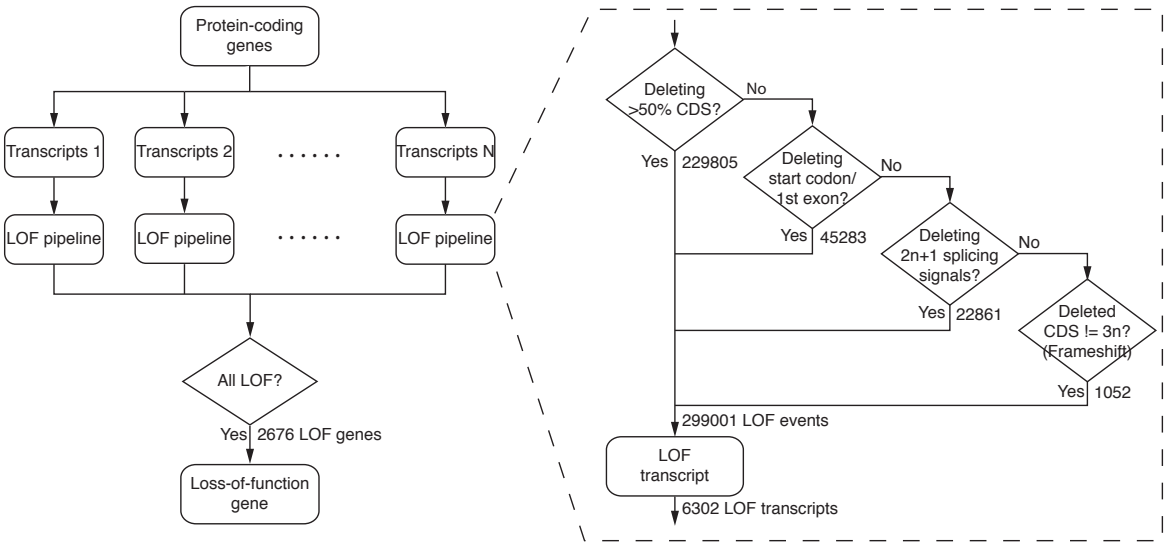

Figure 4.1: Procedure for LOF calling. The flow chart shows the pipeline used to identify LOF genes. A gene with all its transcripts disrupted under any of the four considered LOF scenarios is regarded as LOF. On the right, the numbers under each scenario denotes the number of detected LOF events meeting that criterion. A LOF event is defined as loss of function of one transcript in one individual.

To systematically assess the difference in properties between HI and HS genes, I gathered a large number of annotations describing the evolutionary, functionary and interaction properties of genes (see Methods) and examined the distribution of each individual property in HI and HS genes. I found that HI genes have consistently a more conserved coding se-

quence (human-macaque dN/dS, p = 3.12×10^{-26}), a less mutable promoter (p < 1×10^{-30}), paralogs with lower sequence similarity (p = 1.84×10^{-9}), a longer spliced transcript (p < 1×10^{-30}), a longer 3'UTR (p = 2.63×10^{-12}), higher expression during early development (p = 1.10×10^{-15}), higher tissue specificity in expression (p = 2.29×10^{-6}), more interaction partners in both a protein-protein interaction network (p < 1×10^{-30}) and a gene interaction network (p < 1×10^{-30}) and higher chances of interacting with other known HI genes (p < 1×10^{-30}) and cancer genes (p < 1×10^{-30}) (Figure 4.2). Interestingly, the growth rate of yeast heterozygous deletion strains does not seem to differ between their HI human homologs and HS human homologs, probably reflecting the vast functional differences between the majority of yeast and human genes, except those involved in highly conserved cellular processes.

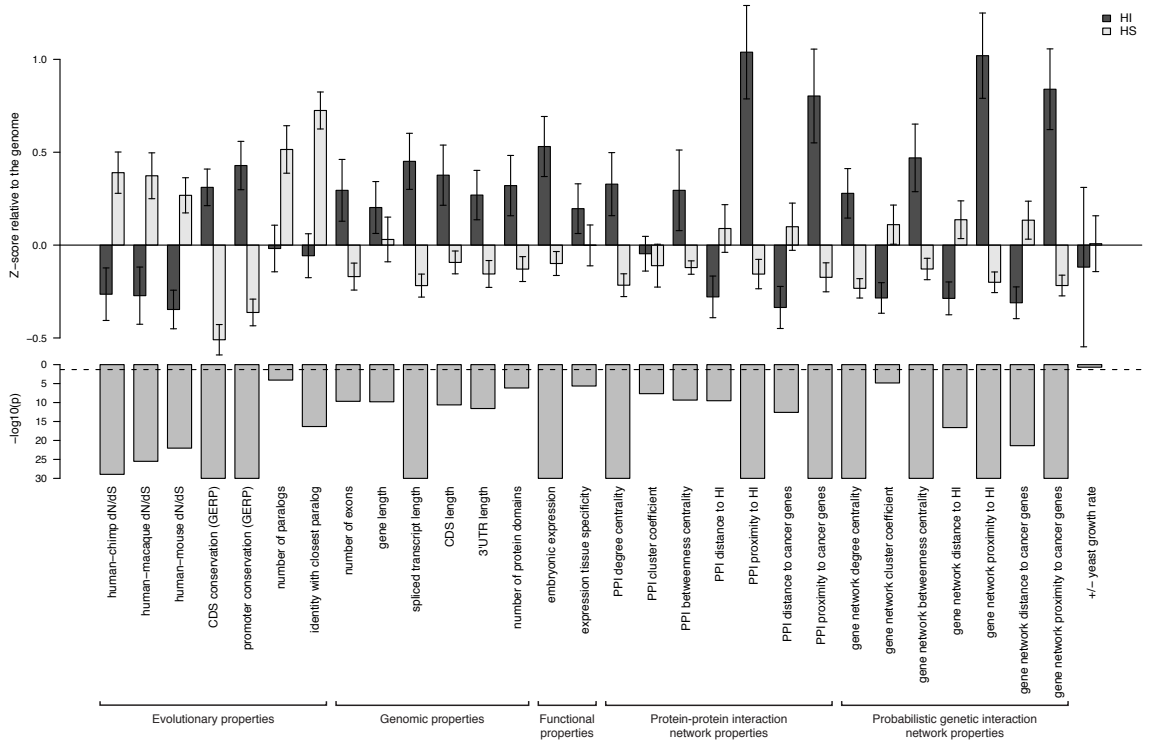

Figure 4.2: Properties that distinguish HI genes from HS genes. The upper part of the figure shows the comparison of the mean of each individual property between HI genes and HS genes. The values are transformed to z-scores relative to the genome average. The error bars represent two times the standard error of the mean. The bars in the middle part present the transformed p value (-$\log_{10}$p) of the Mann-Whitney test on each property. The dashed line marks a p value of 0.05.

Table 4.1: Genomic coverage of gene properties

Property	#Genes	Genomic coverage*
Human-chimp dN/dS	15,084	79.50%
Human-macaque dN/dS	15,025	79.20%
Human-mouse dN/dS	14,386	75.80%
Coding sequence GERP	17,164	90.50%
Promoter GERP	16,807	88.70%
Number of paralogs	11,066	58.30%
Identity of closest paralog		
Number of exons		
Length of gene		
Length of spliced transcript	17,700	93.30%
Length of coding sequence		
Length of 3'UTR		
Number of domains	14,722	88.50%
Embryonic expression†	18,962 (2421)	100% (12.8%)
Tissue specificity of expression	13,950	73.60%
PPI network properties‡	11,077	58.40%
Genetic network properties‡	14,664	77.30%
+/- Yeast growth rate	3,352	17.70%

* Calculated relative to the number of EnsEMBL annotated protein-coding genes that can be uniquely mapped to HGNC symbol.

† Since this is a binary factor where every gene is classified as either over-expressed or not in embryo tissue, the coverage is 100%. The number and fraction of genes over-expressed in embryo is listed in parenthesis.

‡ Including degree, cluster coefficient, betweenness, distance to known HI/cancer genes, proximity to known HI/cancer genes.

4.3.2 Training a model to classify HI and HS genes

The highly significant differences in genomic, evolutionary, functional and network properties between HI and HS genes suggest some combination of these properties may be predictive of haploinsufficiency. I used linear discriminant analysis (LDA) as the supervised classifier, which, given multi-dimensional data and class labels, finds the linear combination of the given dimensions (linear discriminant) that maximizes the inter-class variance. I trained the classifier using various sets of gene properties to obtain a classification model and applied the model to estimate a probability of being HI (p(HI)) for all protein-coding genes in the genome for which all the selected predictor variables were available. Finally, I validated the predictions using external data sets.

The final result is presented below and is followed by discussion of more detailed questions: (*i*) which gene properties should be incorporated (Section 4.3.2.1) ? (*ii*) which training dataset should be used (Section 4.3.2.2) ? (*iii*) does a more sophisticated classifier perform better (Section 4.3.2.3) ? Section 4.3.2.4 presents the validation of prediction. Section 4.3.2.5 described some further improvements of the prediction of which the outcome is not included below as they were undertaken at a later stage.

After assessing various different sets of predictor variables (see Methods, and below) my initial classifier was trained with four predictor variables: dN/dS between human and macaque, promoter conservation, embryonic expression and network proximity to known HI genes. The model was obtained by training on 234 HI genes and 326 HS genes for which the predictor variables were available. All predictor variables were scaled to the same variance before entering LDA so that their contribution can be measured by the coefficients of the resulting linear discriminant. Proximity to known HI genes provided the most predictive power. The model achieved an AUC of 0.81 and a MCC of 0.50 in ten-fold cross-validation (Figure 4.3). I applied the model to estimate a probability of being HI for all 12,443 protein-coding genes in the genome for which all four selected predictor variables were available. The distribution of the predicted p(HI) is clearly bimodal, with a large peak near 0.2 and a much smaller peak at 1 (Figure 4.4 left). The distributions of p(HI) for the HI and HS training sets differ significantly ($p < 1 \times 10^{-30}$, Mann-Whitney test or Kolmogorov-Smirnov test) (Figure 4.4 right).

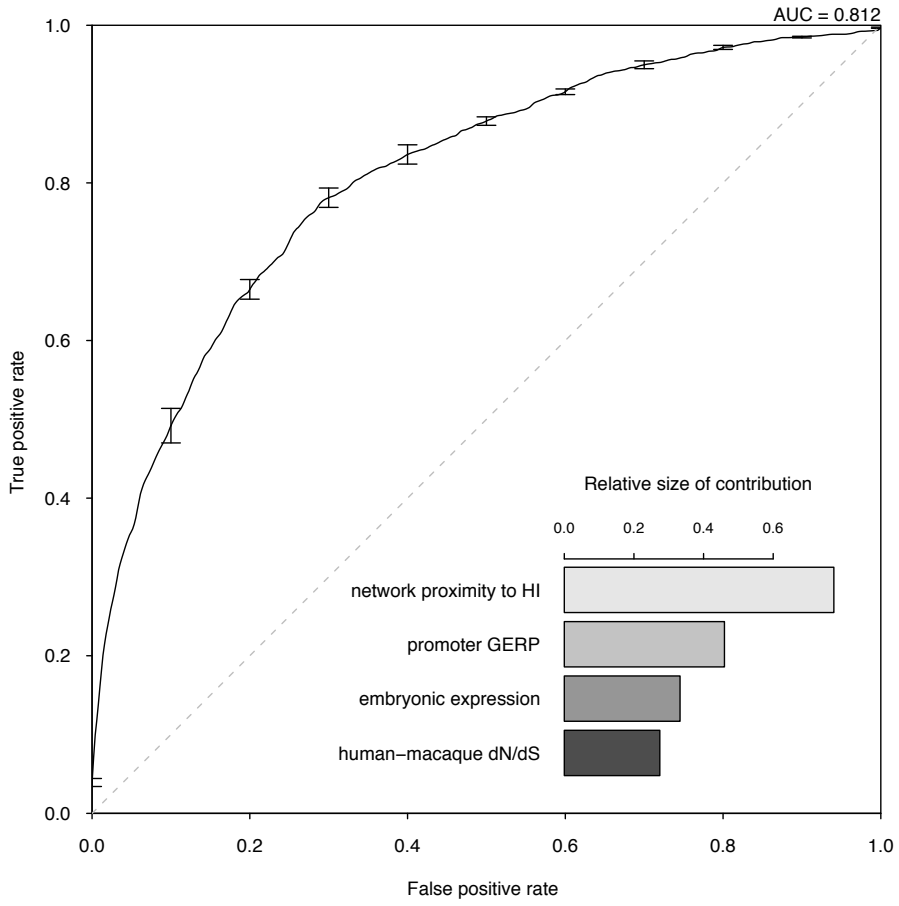

Figure 4.3: Assessment of model performance. The ROC curve demonstrates the performance of the model evaluated by 10-fold cross-validation. The lower right part shows the relative contribution of each predictor variable to the prediction model measured by the absolute value of the scaling factor of each predictor variable constituting the linear discriminant.

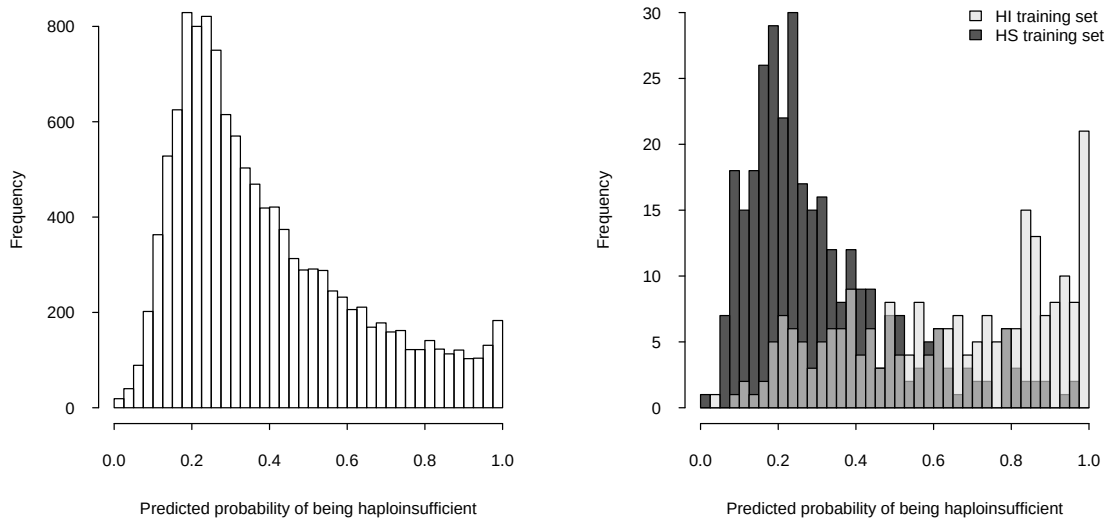

Figure 4.4: Predicted probability of being haploinsufficient. The histogram on the left shows the distribution of the predicted probability of being haploinsufficient (p(HI)) of all 12,443 predictable genes. The histogram on the right shows the distribution of the predicted p(HI) of the HI training set (light grey) and the HS training set (dark grey).

4.3.2.1 Integrating information from multiple 'orthogonal' predictor variables improves classification

To assess the marginal utility of using more than one predictor variable, I trained separate LDA models from the same set of genes (known HI genes plus HS genes) using only one predictor variable at a time and compared the cross-validation performance with using all predictor variables. The latter out-performs models using single predictor variable (max AUC = 0.78 for network proximity to known HI genes whereas the integrated model achieves 0.81) (Figure 4.5), indicating that combining the predictor variables together generated a more predictive model than considering any of the individual predictor variables in isolation.

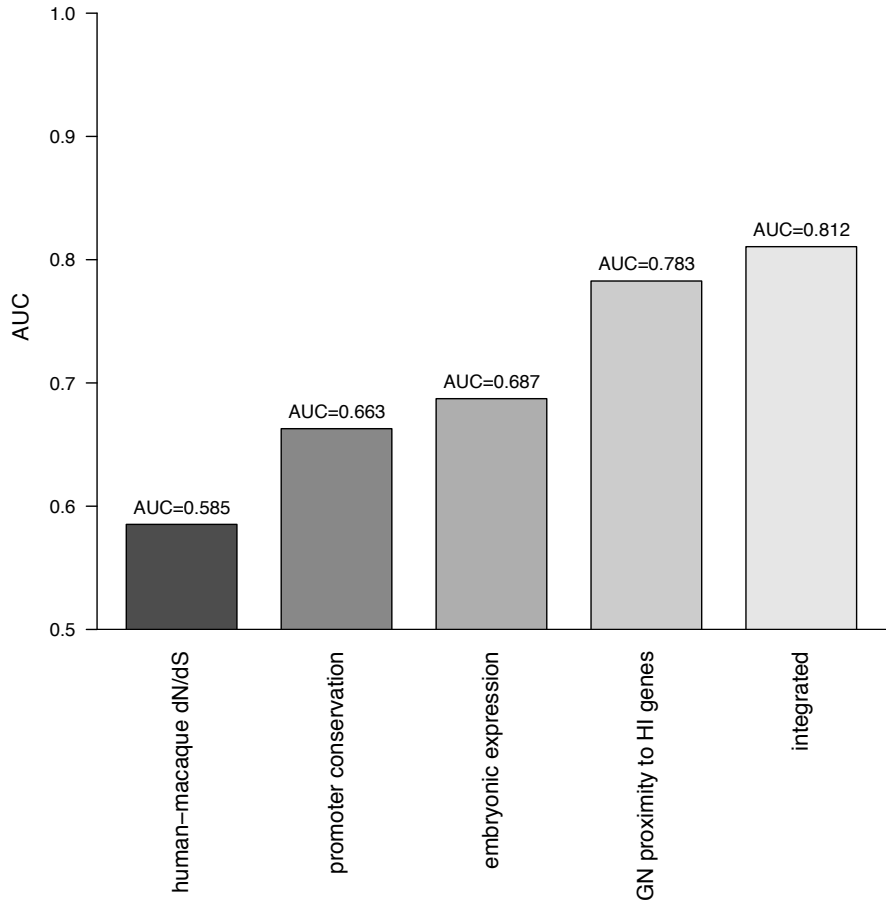

Figure 4.5: Prediction performance of single predictor variable and integrated model. Mean AUC of each model in 10-fold cross-validation repeated 30 times are shown as vertical bars with the actual values label at the top.

Since each gene property annotation is only available for a fraction of genes in the genome (Table 4.1), there is a trade-off between the possible increase in prediction performance by considering more gene properties as predictor variables and the decrease in the coverage of genes one could predict. Therefore, I aimed to select a small number of most predictive properties that are relatively 'orthogonal' in the kind of information they provide (see Methods).

After evaluating a number of possible combinations of predictor variables, which all had similar performance (Figure 4.6), I selected a model comprising of 'dN/dS between human and macaque', 'promoter conservation', 'embryonic expression' and 'network proximity to known HI genes'.

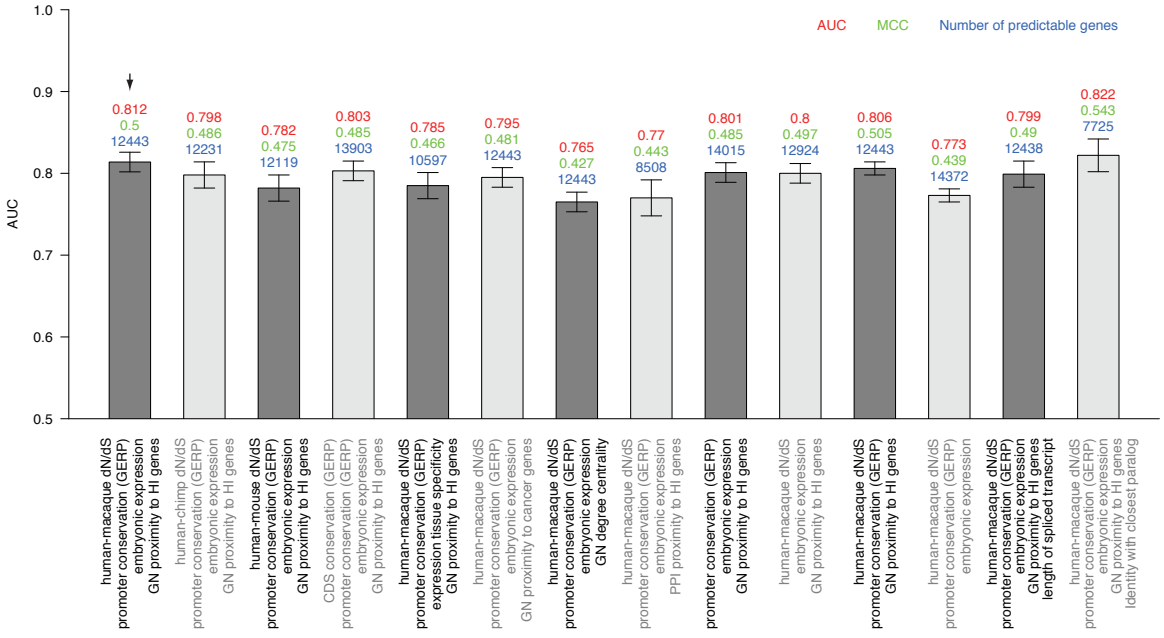

Figure 4.6: Comparison of model performance. The AUCs of each combination of predictor variables in 10-fold cross validation repeated 30 times are shown as vertical bars with error bars represent 2 times standard deviation. The mean AUC (red), mean MCC (green) and the overall gene coverage (blue) are labeled on top of each bar. The bar pointed by the black arrowhead is the chosen combination of predictor variables.

4.3.2.2 Using HS genes as negative training set improves classification

Previous studies [119–121] have compared HI-related gene sets against the rest of the genome to describe their characteristics. I investigated how the choice of negative training set influences the performance of my prediction model. I generated gene sets of different sizes randomly sampled from non-HI genes with complete predictor variable information and compared the cross-validation performance (AUC) resulting from the use of these gene sets as the negative training set to the use of the HS gene set as the negative training set (Figure 4.7). The use of a judiciously selected HS gene set is clearly advantageous.

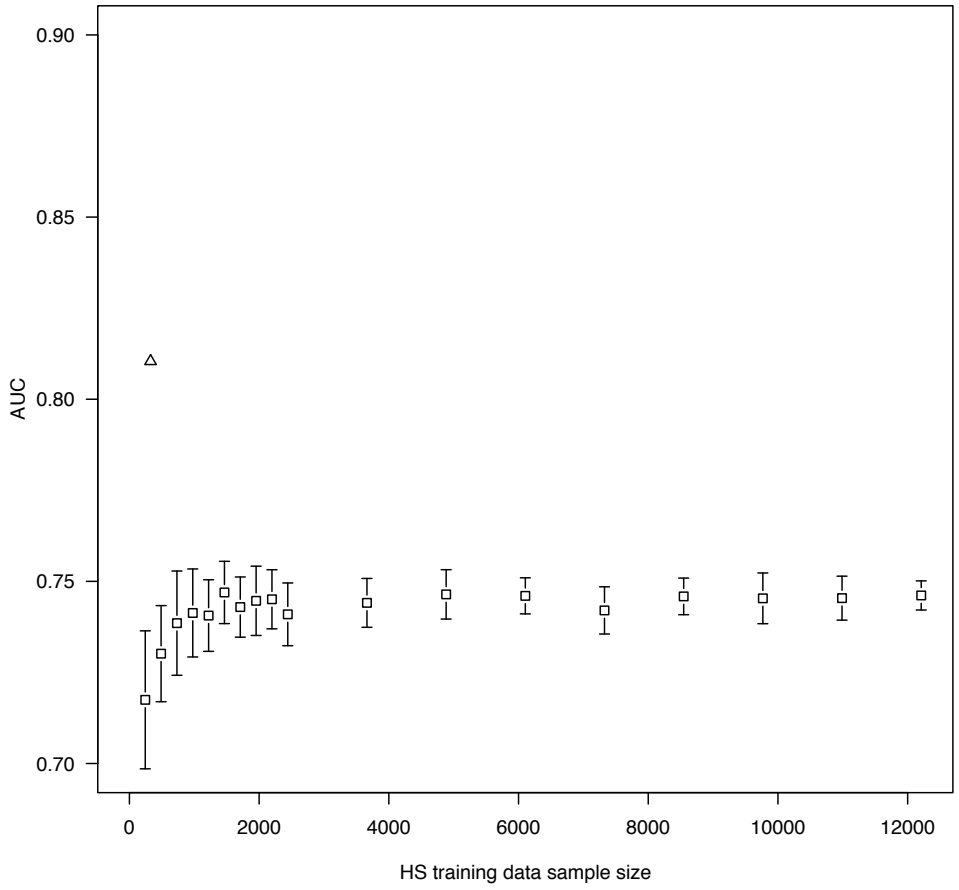

Figure 4.7: Prediction performance of using HS and genome background as negative training set. The plot compares the cross-validation performances resulted from using different gene sets as negative training set. The triangle represents HS gene set generated from CNV data. The squares represent different sizes of random gene sets sampled from the genome after excluding known HI genes. For each size, the gene set was sampled 20 times and the standard deviation of the resulting performances is shown as error bar.

I further investigated if our model performance is sensitive to the CNV discovery and filtering parameters, which determines the stringency of the HS gene set. I examined the influence on cross-validation performance of using different confidence thresholds (Birdseye LOD score) in CNV discovery and population frequency when generating HS gene set. A greater LOD score indicates higher confidence and thus a more stringent CNV set. Similarly, the more frequently a gene is found LOF in apparently healthy individuals, the more likely it is haplosufficient, and thus the negative training set is more stringent. I found that the LOD score threshold has little influence on the model performance, within the range I assessed (Figure 4.8). The use of recurrent LOF genes exhibits an apparent improvement of performance over the use of all LOF genes under most LOD thresholds. Further increase in stringency by requiring higher frequency results in further reduction of the size of negative training set, but little if any increase in performance of the prediction model. Therefore, I adopted the negative training set generated under 'LOD $>$ 10' and 'found in at least two individuals' in further analysis.

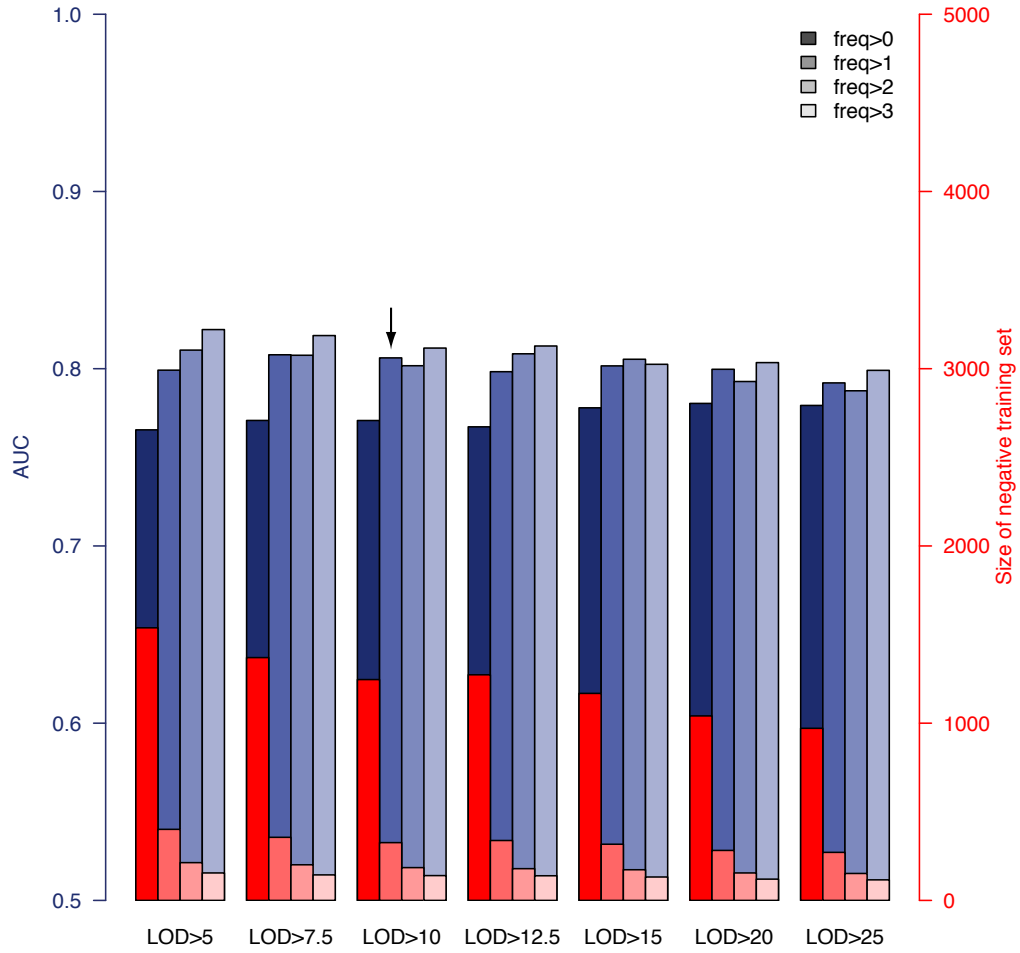

Figure 4.8: Prediction performance under different parameters used in generation of negative training set. The cross-validation performance (AUC) resulted from using negative training sets generated with different parameters are represented by blue vertical bars with axis on the left. The sizes of these negative training sets are represented by red vertical bars with axis on the right. Bars are grouped by the CNV calling parameters, LOD score, and within each group the darkness of coloring represent different frequency threshold used to define HS as shown in the legend. The bar pointed by the black arrowhead represents parameters and corresponding negative training set adopted in further analysis.

4.3.2.3 LDA achieves similar classification performance compared to a more sophisticated classifier

I investigated if the use of support vector machine (SVM), a more sophisticated machine learning method, as classifier would improve prediction performance. An SVM model was trained on the same training set as LDA with optimized parameters (gamma = 0.1, cost = 1) and class weights. The performance was examined by self-validation, leave-one-out

cross-validation and 10-fold cross-validation. Despite being more sophisticated and computational expensive, SVM exhibits no appreciable improvement over LDA (Figure 4.9).

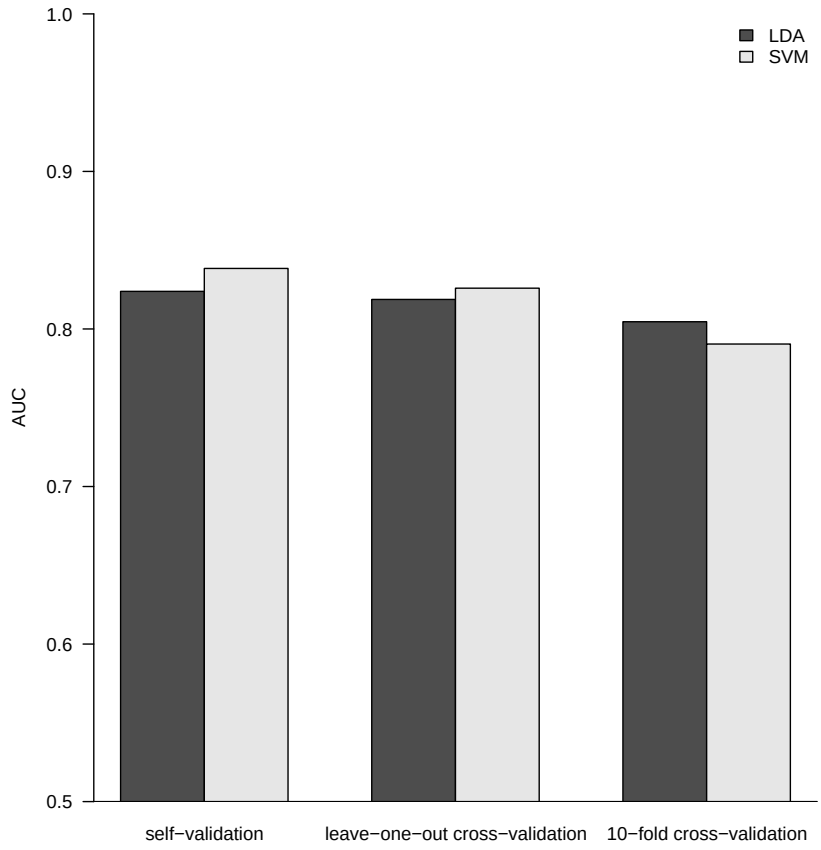

Figure 4.9: Comparing the prediction performance of LDA and SVM. The plot shows the comparison of prediction performance between LDA (dark bar) and SVM (light bar) using three approaches (from left to right): self-validation, leave-one-out cross-validation and 10-fold cross-validation. In the first two comparisons, SVM exhibits only very marginal improvement over LDA, whereas in the third LDA is marginally better.

4.3.2.4 Validating haploinsufficiency predictions using external datasets

It is not possible to assess how well-calibrated the predicted probabilities of being HI are, as the fraction of human genes that exhibit HI is not known. I therefore sought to validate these predictions using indirect approaches that examined the distribution of p(HI) in independent gene sets enriched for HI. As there is no credible estimation of the number of human HI genes, in some of the following validation analyses I arbitrarily labeled the genes in the

top 10% of p(HI) as being predicted HI genes. However, the results were robust against this threshold being varied by at least a factor of at least 2.

First, I asked if genes implicated in human dominant diseases were enriched in our predicted HI genes relative to recessive-disease-causing genes. I retrieved 571 and 772 genes implicated in dominant and recessive disease from the OMIM and hOMIM[119] database, respectively, with no information regarding haploinsufficiency (and thus not included in our training data), and compared the distribution of predicted p(HI) against each other. The HI status could be predicted for 392 dominant genes and 606 recessive genes, of which 87 and 39 were predicted as being HI, respectively. This 4.14 fold enrichment of genes predicted to be HI within the dominant disease gene set is highly significant ($p = 4.46\times10^{-13}$, Fisher's exact test). Simply comparing the distribution of p(HI) values for these dominant and recessive genes also shows a highly significant shift towards high p(HI) values in dominant relative to recessive genes ($p = 4.44\times10^{-16}$, Mann-Whitney U test) (Figure 4.10).

Second, I asked if heterozygous knockouts of the orthologs of predicted human HI genes are more likely to cause severe phenotypic abnormalities in mice. For this purpose, I extracted a list of 1,523 mouse genes whose heterozygous knockout cause various abnormal phenotypes from the MGI database, mapped them onto orthologous genes in humans, removed orthologs to genes in our training gene sets and extracted the predicted p(HI) for the remainder. HI status could be predicted for the orthologs of 1,063 of these genes and 260 (24.5%) of them were predicted HI, indicating a 2.45 fold enrichment ($p < 1\times10^{-30}$, Fisher's exact test) (Figure 4.11). If focusing on those genes of which the heterozygous LOF phenotypes involve prenatal lethality (MP:0002080), the fold of enrichment increased to 4.38 ($p = 3.60\times10^{-12}$, Fisher's exact test) (28 predicted as HI out of 64 that could be predicted).

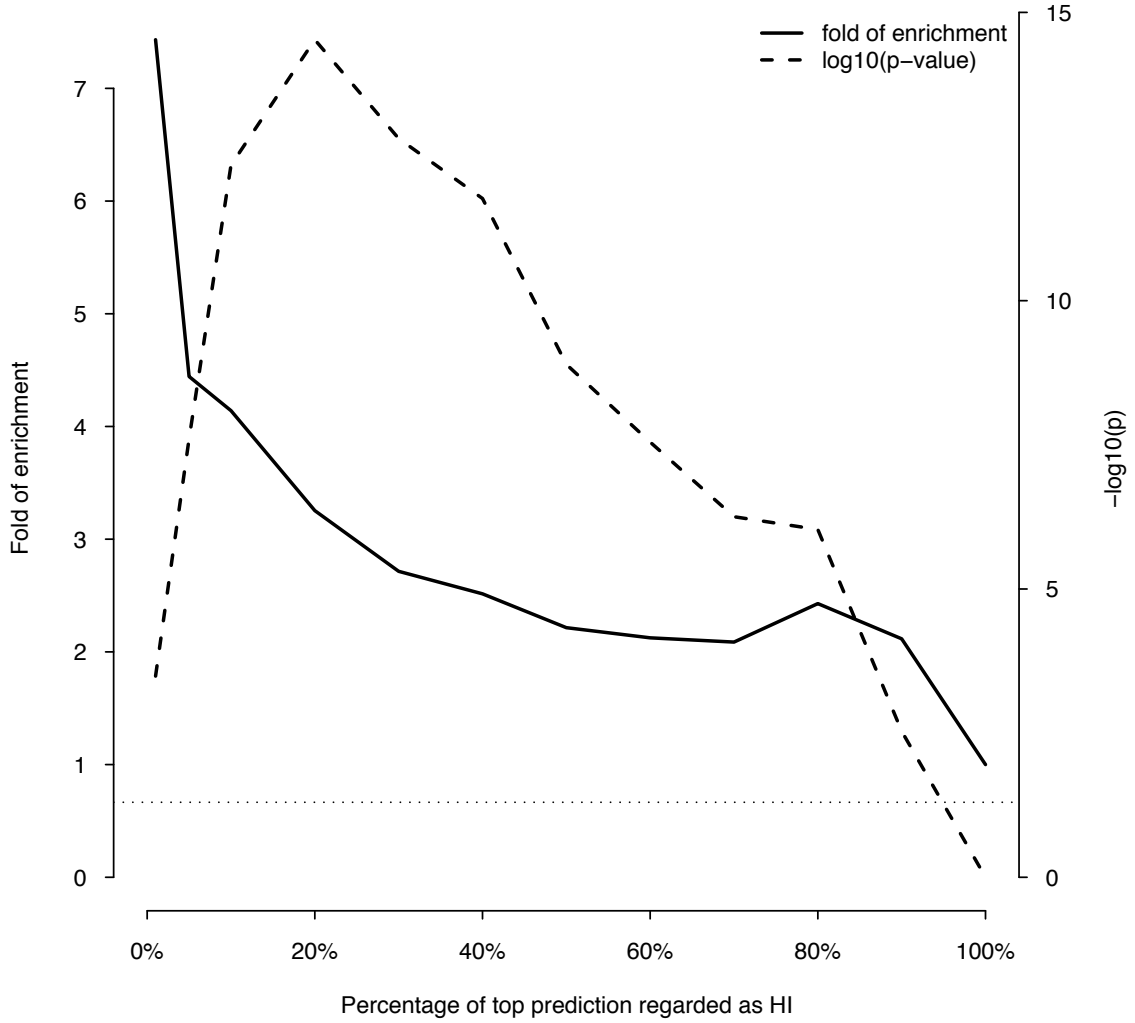

Figure 4.10: Enrichment of predicted HI genes in dominant genes relative to recessive genes. This plot shows the fold of enrichment of predicted HI genes in dominant genes relative to recessive genes (thick solid line) as a function of the proportion of predictions labeled as being haploinsufficient. Also plotted is the transformed p value (-log10p) of the corresponding Fisher's exact test (thick dashed line). The horizontal dashed line marks the p value of 0.05.

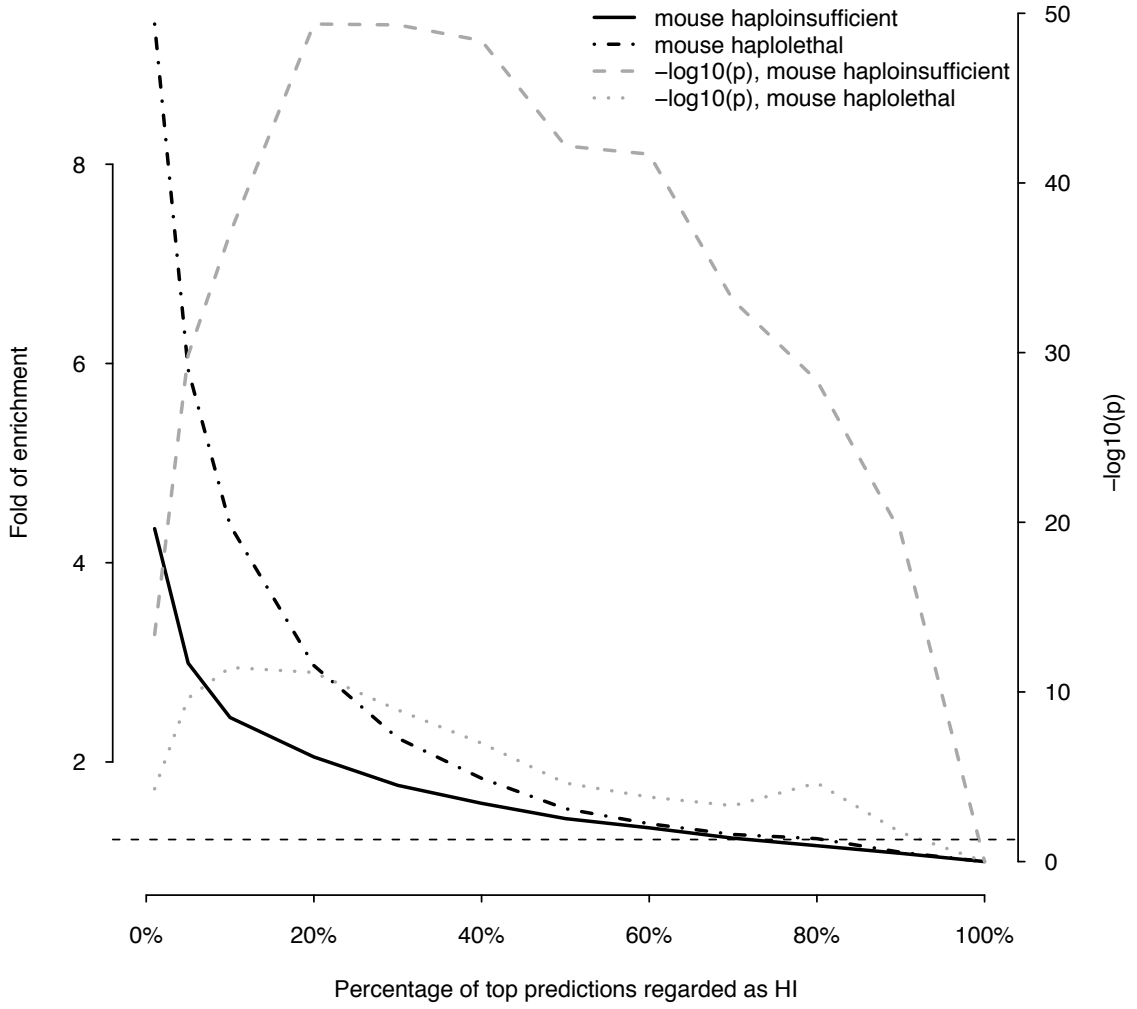

Figure 4.11: Enrichment of predicted HI genes in orthologs of mouse haploinsufficient genes and mouse haplolethal genes. This plot shows the fold of enrichment of predicted HI genes in human orthologs of mouse haploinsufficient genes (black solid line) and mouse haplolethal genes (black dashed line) relative to the genome average as a function of the proportion of predictions labeled as being haploinsufficient. The two lines in grey show the transformed p values of the corresponding Fisher's exact test. The horizontal dashed line marks the p value of 0.05.

4.3.2.5 Improving prediction with expanded training data and improved predictor variables

Having achieved reasonable performance with my initial predictive model of gene haploinsufficiency and shown that neither changing the classifier nor how the HS gene training data are filtered, I explored different potential strategies to improve upon the performance of this predictive model. In this section I describe two, potentially complementary strategies: (*i*) Including new and improved predictor variables into the predictive model, and (*ii*) using improved positive control training data (*i.e.* known HI genes).

4.3.2.5.1 Inclusion of new and improved predictor variables

In the light of the emerging role of conserved noncoding elements in regulation of gene expression, especially of developmental genes known to be dosage sensitive, I investigated several variables that summarize the extent of conserved noncoding sequence within and flanking a gene. I settled on the sum of GERP scores of all bases of conserved non-coding elements within an interval $\pm$50kb of the gene as a candidate predictor variable. This property differs significantly between HI and HS genes ($p = 4.0 \times 10^{-54}$, Mann-Whitney U test).

The coverage of the protein-protein interaction network was also expanded from 11,077 genes and 70,632 interactions to 16,390 genes and 1,240,972 interactions by incorporating data from the STRING database [157]. As a result, the number of genes predictable with the same predictor variables as selected in Section 4.3.2.1 increased to 13,030 (+5%) without imputation or, if using the predictor variables optimized for the updated gene properties as described in Section 4.3.2.5.2, increased to 16,017 (+29%). I also updated the gene property annotations to EnsEMBL 53.

4.3.2.5.2 Improved HI training set through literature mining and manual curation

The known HI genes used as positive training set was initially taken from Dang *et al* and Seidman *et al*, which reflected the current knowledge in Nov 2007. I performed a literature searching on Aug 2010 to include more, newly discovered HI genes. Through text mining of PubMed abstracts (see Methods), 138 genes were added to the HI set, resulting in a combined set of 439 genes. 358 of these genes for which a PubMed abstract is available were

manually curated. After curation of the entire set, 40 genes were removed, 55 were labeled as with weak evidence (see Appendix A). 72 genes that are involved in cancer [149] were also removed, since seemingly dominant inheritance could be the result of somatic loss of heterozygosity instead of truly genetic haploinsufficiency in the case of cancer.

To evaluate the expanded and manually curated HI set, the model was re-trained both with and without genes with weak evidence using the updated version of the same predictor variables as Section 4.3.2.1 and the performance were measured by two approaches: (*i*) the cross-validation AUC (AUC_{CV}) and (*ii*) the AUC for classifying pathogenic and benign CNVs using model-prediction-based LOD scores (AUC_{LOD}) . The model trained with the more stringent set exhibited higher cross-validation AUC than the model trained with more relaxed set (0.77 vs 0.75) and the two had the same variant classification AUC (0.98). Whereas the more stringent model achieved the same cross-validation AUC as the model trained on the initial training set after removing cancer genes, both all were noticeably lower than the model trained on the initial training set with the earlier predictor variables (0.81). Therefore, I explored if other combinations of predictor variables perform better with the updated annotations and training set. A comparison of performance statistics is shown in (Table 4.2). Based on both cross-validation AUC and variant classification AUC, I selected the model that incorporates the predictor variables: 'GERP score of conserved non-coding elements', 'median size of spliced transcripts', 'identity to closest paralog' and 'embryonic expression' and 'proximity to other known HI genes in protein-protein interaction network', and I trained this model using the more stringent updated known HI gene set. The new predictive model achieved higher cross-validation AUC (0.86 vs 0.81) and similar variant classification AUC (0.96 vs 0.96) to the un-updated model, while improving prediction coverage without imputation (16,017 vs 12,443). However, when testing if genes found with LOF substitutions and indels in sequenced exomes have lower p(HI) than the genome background as did in Section 4.3.3.4, the difference was less significant despite still being in the same direction (0.13 vs 0.21, p = 2.2×10^{-12}, Mann-Whitney test). The difference was even smaller when comparing p(HI) of genes found with LOF substitutions in a larger exome-sequencing dataset that consisted of $\sim$300 apparently healthy individuals (0.18 vs 0.21, p = 4.5×10^{-3}, Mann-Whitney test). Thus although the cross-validation seems to indicate improved performance from this later model, the comparisons with external datasets of different types, does not back this up.

Table 4.2: Performance comparison of prediction models

HI training set	Predictors	#HI training	#Predictable	AUC_{CV}	AUC_{LOD}
Initial*	CNC_GERP PPI_LLS2HI	237	16,017	0.866	0.915
	CNC_GERP PPI_LLS2HI TRANS_SIZE PARALOG_DIST EARLY_DEV	237	16,017	0.869	0.945
	MACAQUE_DNDS PROMOTER_GERP EARLY_DEV GGI_LLS2HI	237	13,030	0.765	0.97
Expanded	CNC_GERP PPI_LLS2HI	312	16,017	0.864	0.934
	CNC_GERP PPI_LLS2HI TRANS_SIZE PARALOG_DIST EARLY_DEV	312	16,017	0.864	0.964
	MACAQUE_DNDS PROMOTER_GERP EARLY_DEV GGI_LLS2HI	312	13,030	0.765	0.975

* Cancer genes removed

4.3.3 Using HI gene predictions to assess pathogenicity of deletions

4.3.3.1 Defining a genomic-interval-based pathogenicity score

I investigated how my gene-based predictions of haploinsufficiency might be used to discriminate between benign and pathogenic genic deletions. I considered that a natural way to score the probability of a deletion of a genomic interval causing a haploinsufficiency phenotype is to generate a LOD (log-odds) score comparing the probability that none of the genes covered contained in the interval will cause haploinsufficiency with the probability that at least one of the genes will cause haploinsufficiency, as shown schematically in Figure 4.12. This LOD score is calculated using the formula below:

$$LOD = \ln\left(\frac{1-\prod(1-\mathrm{p(HI)})}{\prod(1-\mathrm{p(HI)})}\right)$$

, and assumes that there is no statistical interaction between the genes. Worked examples of this calculation are shown in the figure below. Higher LOD scores indicate deletions are more likely to be pathogenic as a result of haploinsufficiency.

4.3.3.2 Discriminating benign and pathogenic deletions

I then considered how these deletion-based haploinsufficiency scores might be used to assess whether a genic deletion observed in a patient might cause their disease. One way of framing probabilistically this intuitively simple question is to estimate the opposing probability, that the deletion is unrelated to the patient's disease status. This can be equated to the probability of drawing an individual at random from a healthy control population with a deletion at least as pathogenic as the deletion in the patient. This probability can be estimated empirically as the proportion of healthy controls with a genic deletion having the same or greater haploinsufficiency LOD score.

To test this approach, and to avoid circular reasoning, I retained a subset (2,322 GWAS controls used in studies of schizophrenia and bipolar disease) of the 8,458 apparently healthy individuals from which the HS genes in the original training data were derived and generated a new set of p(HI) by training on the reduced HS gene set identified from the rest of apparently healthy individuals using the same method as described in Section 4.3.2. After imputation of predictor variables (see Methods), this new training set contains 287 HI genes

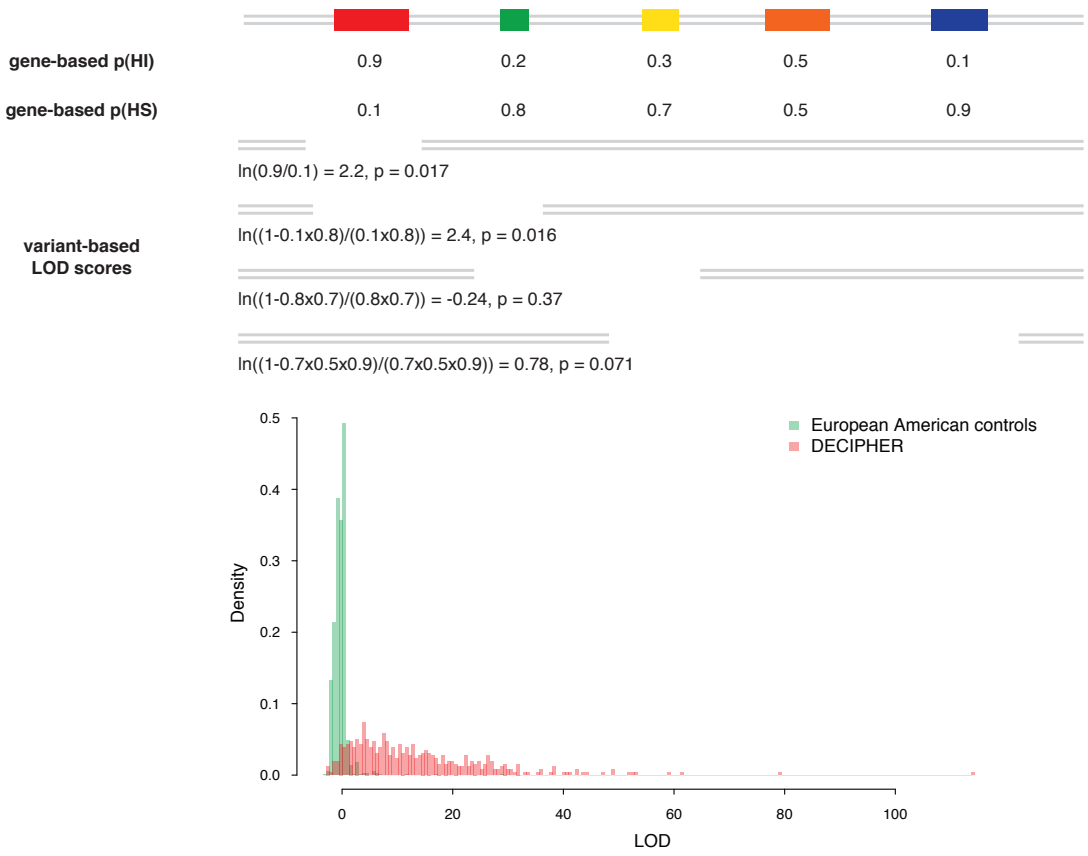

Figure 4.12: Calculation of deletion-based LOD scores and the distribution of LOD score of control individuals and pathogenic *de novo* deletions. The upper portion of the figure is a schematic demonstration of the calculation of the deletion-based LOD score. The contribution of genes with high p(HI) is accordingly weighted in a probabilistic way. The deletion with the largest LOD score in each individual is recorded and their distribution is shown in the lower portion of the figure. The distribution of maximal LOD scores of 2,322 control individuals are shown in green and the distribution of LOD scores of 487 pathogenic *de novo* deletions from DECIPHER are in red. Using the control distribution as the null, the probability a deletion is pathogenic can be assessed.

and 594 HS genes (234 HI genes and 270 HS genes before imputation). The model trained from this reduced training set achieved a similar AUC and MCC in 10-fold cross-validation as the model trained from the original training set (after imputation: AUC = 0.84, MCC = 0.55; before imputation: AUC = 0.81, MCC = 0.50).

The resulting predictions are also highly consistent with the original predictions (correlation between p(HI) is 0.99 both before and after imputation). I used the predictions based on the dataset that includes imputed predictor variables to allow the more reliable assertion of haploinsufficiency of a genomic interval from the vast majority of the genes affected by its deletion (17,456 genes with p(HI) after imputation as opposed to 12,443 before imputation). Based on these predictions I determined the distribution of the maximal deletion haploinsufficiency scores for the retained subset of 2,322 apparently healthy individuals.

To compare this distribution of 'most pathogenic' deletions discovered in apparently healthy individuals with truly pathogenic deletions, I collected 487 *de novo* deletions identified from array-based CNV detection and classified as being putatively pathogenic in the DECIPHER database [129]. I focused exclusively on deletions known to be *de novo* variants, as I infer that their pathogenicity has been ascribed primarily on the basis of their inheritance status, and not their gene content. The distributions of maximal LOD scores in GWAS controls and LOD scores of pathogenic DECIPHER deletions are shown in Figure 4.12. The pathogenic deletions have strikingly significantly higher LOD scores than deletions observed in GWAS controls ($p < 1 \times 10^{-30}$, Mann-Whitney U test). I observed that for 92% of the pathogenic deletions there was a probability of less than 5% of drawing an individual at random from our control population with a genic deletion of equal or greater LOD score, and for 83% of pathogenic deletions there was a less than 1% probability.

I computed ROC curves to compare three different approaches for discriminating between pathogenic deletions and deletions seen in controls: (*i*) LOD scores, (*ii*) the length of the deletion, and (*iii*) the number of genes in the deletion (Figure 4.13). These ROC curves clearly show that the haploinsufficiency LOD score is the best metric of the three for discriminating between pathogenic deletions in patients and deletions seen in controls.

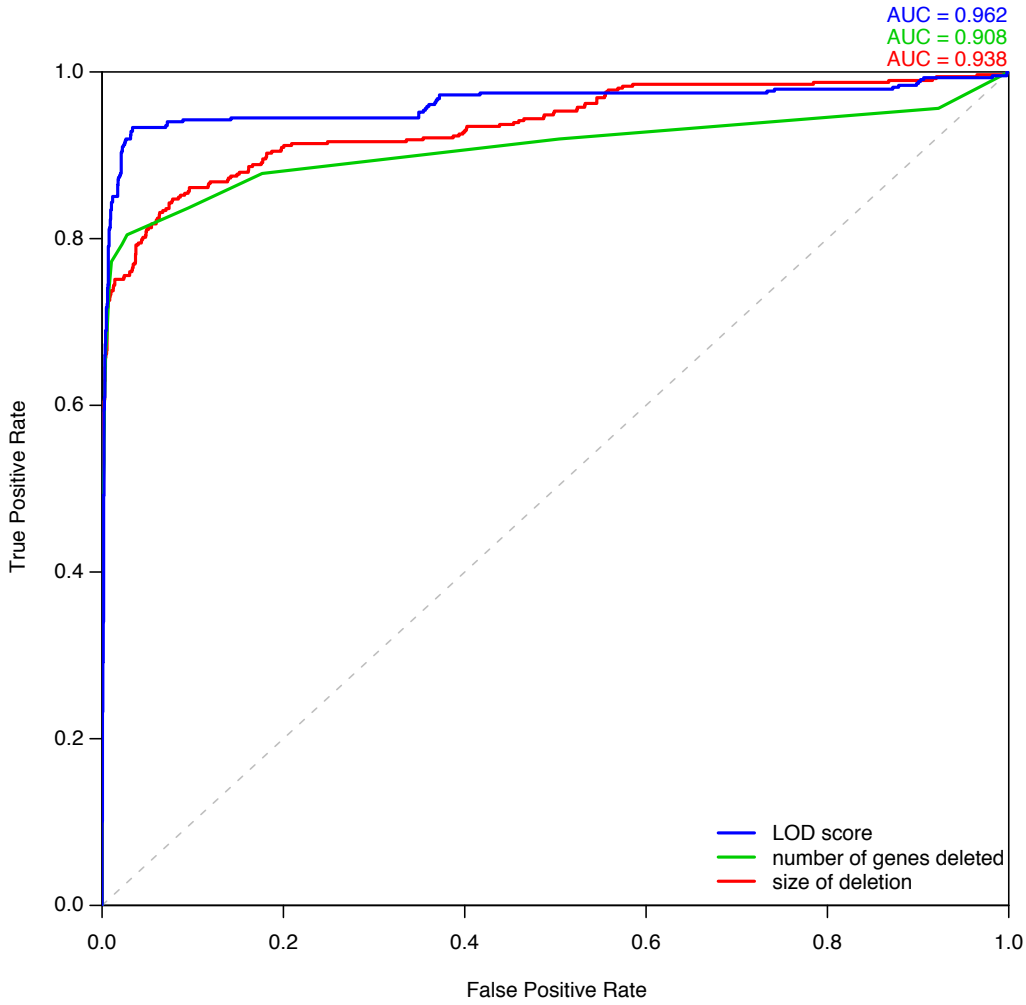

Figure 4.13: Comparison of different metrics for assessing deletion pathogenicity. Three ROC curves represent the performance of three different methods for distinguishing between pathogenic deletions from DECIPHER and the most pathogenic deletions observed in control individuals. The blue curve denotes using LOD score calculated from predicted probability of exhibiting haploinsufficiency as the metric of pathogenicity. The green curve denotes using the number of deleted genes as the metric, in which case the most pathogenic deletion per individual is the one containing greatest number of genes in that individual. The red curve denotes using the size of deletion as the discriminating metric.

I investigated whether the distribution of maximal LOD scores is significantly different between 1,433 European-Americans (EA) and 889 African-Americans (AA) GWAS controls, which, if true, might suggest the necessity of using ethnicity matched population pathogenicity score distributions. I observed that there was not a significant difference in median haploinsufficiency scores in EA and AA populations ($p = 0.71$, Mann-Whitney U test). The EA

controls have a slightly longer tail of more pathogenic deletions (*e.g.* a higher proportion of EA controls have deletions with LOD scores in the top 1% in the pooled distribution, Table 4.3), which is consistent with the previous suggestion that purifying selection is more efficient in African populations due to their larger effective population sizes [158, 159]. However, this difference is again not significant (p = 0.24, Fisher's exact test).

Table 4.3: Population-specific properties of LOF CNVs

Population	Average #(LOF CNV) per individual	Average #(predictable LOF gene) per individual	Average #(LOF gene) in CNV with max LOD per individual	Average of max LOD per individual	Proportion with max LOD $\geq$ 99% of the pooled population
European American	7.41	10.5	2.85	-0.36	1.18%
African American	7.35	10.1	2.74	-0.38	0.79%

4.3.3.3 Extension to duplications

Since the probability of a gene being haploinsufficient partly reflects its general dosage sensitivity, it might be reasonable to expect abnormally increased dosage of at least some HI genes could also be pathogenic, as exemplified by the *PMP22* gene contained in the lim1.5Mb region at 17p11.2 of which duplication causes Charcot-Marie-Tooth syndrome type 1A and deletion causes Hereditary Neuropathy with Liability to Pressure Palsies. Therefore, I investigated if the interval-based haploinsufficiency LOD score could also be applied to classifying the pathogenicity of duplications. All computational procedures were identical to those for deletions, except the slight difference that the LOD scores for duplications were calculated from the p(HI) of genes contained in a genomic interval instead of LOF genes. I again compared the ROC curves of using LOD scores, the length of the duplication, and the number of genes in the duplication (Figure 4.14). The LOD score exhibit similar performance to the size of duplication in discriminating between pathogenic duplications and duplications seen in controls. Both LOD score and size performed better than the number of genes.

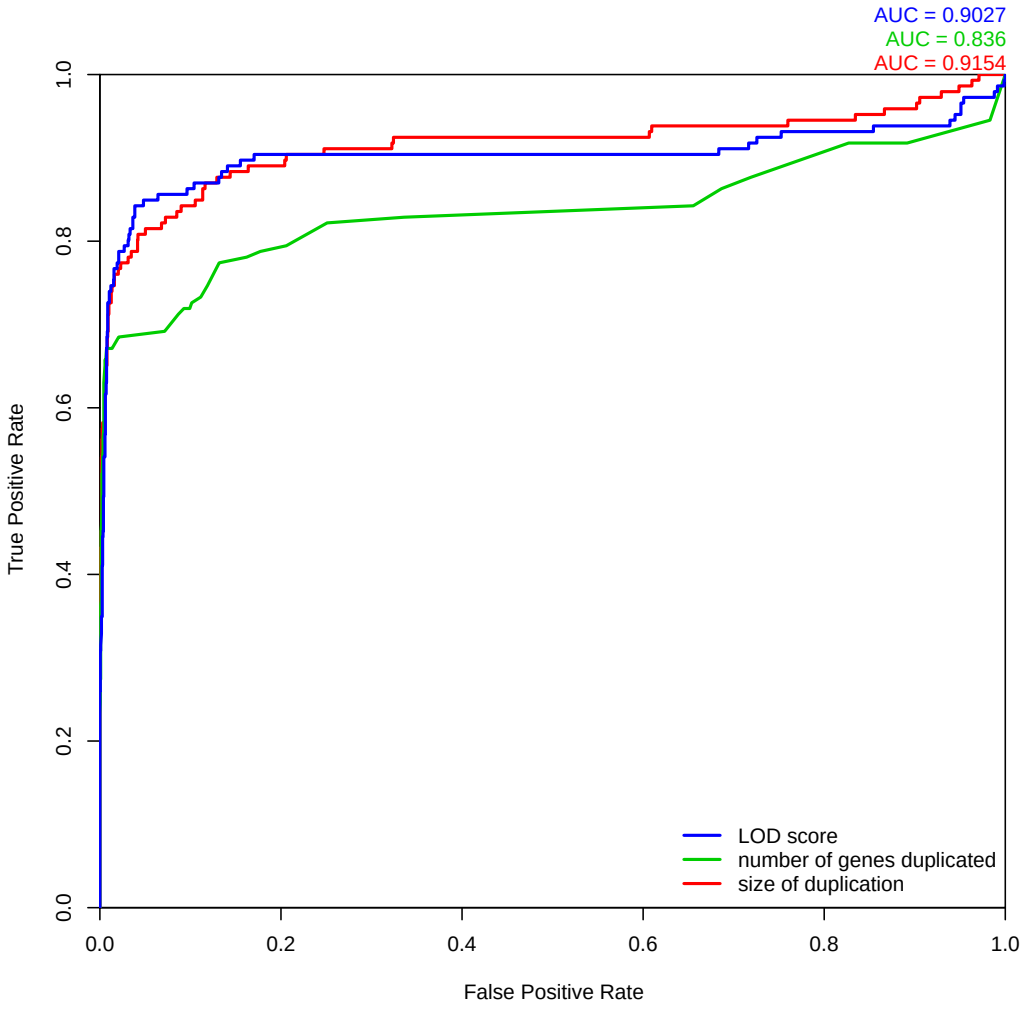

Figure 4.14: Comparison of different metrics for assessing duplication pathogenicity. Three ROC curves represent the performance of three different methods for distinguishing between pathogenic duplications from DECIPHER and the most pathogenic duplications observed in control individuals. The blue curve denotes using LOD score calculated from predicted probability of exhibiting haploinsufficiency as the metric of pathogenicity. The green curve denotes using the number of duplicated genes as the metric, in which case the most pathogenic duplication per individual is the one containing greatest number of genes in that individual. The red curve denotes using the size of duplication as the discriminating metric.

4.3.3.4 Extension to other forms of genetic variation

I investigated whether the gene-based probabilities of haploinsufficiency that I have generated are of general utility across different forms of genetic variation. If this is indeed the case then I should expect that genes harboring loss-of-function substitutions or small in-

dels in apparently healthy individuals should not have a high p(HI). I identified 349 genes as having LOF substitutions and indels in 12 recently sequenced exomes [116], of which I could estimate p(HI) for 176 that were not also in the HS training set (and thus represent a fair set for independent comparisons). These genes are highly significantly enriched among genes with low probabilities of exhibiting haploinsufficiency ($p = 1.06 \times 10^{-20}$ when comparing to the genome, and $p < 1 \times 10^{-30}$ when comparing to known HI genes, Mann-Whitney U test). This result implies that there are not substantial differences between genes that tolerate whole gene deletions and those that tolerate smaller loss-of-function variants.

Moreover, by utilizing a large gene-resequencing dataset that contains 47,576 SNPs found by direct resequencing of 11,404 protein-coding genes in 35 individuals (20 European-Americans (EA) and 15 African-Americans (AA)) [160], I studied the allele frequency spectrum of different types of genic variants with respect to p(HI) of the genes. I hypothesized that genes under stronger negative selection should exhibit an enrichment of rare alleles in their allele frequency spectrum relative to genes under less selective constraint. There are 14,420 nonsynonymous SNPs and 16,213 synonymous SNPs in the dataset found within genes with predicted p(HI). I examined their derived allele frequency (DAF) spectrum as a function of p(HI) of the genes in which they are located (Figure 4.15).

Regardless of population composition, the DAF spectrum of nonsynonymous SNPs are significantly more skewed towards rare variants in gene sets with higher p(HI) than in those with lower p(HI), as assessed by a one-sided Mann-Whitney U test comparing the median of the allele frequency spectrum of nonsynonymous variants in genes with p(HI) in the top 20% with that of nonsynonymous variants in genes with p(HI) in the bottom 80%. The p value for this test in EA was 3.95×10^{-3}, and in AA was 2.85×10^{-7}. As a control, the difference in DAF of synonymous SNPs between high p(HI) genes and low p(HI) genes was not significant (EA p = 0.127, AA p = 0.057). These results suggest greater selective constraint on genes predicted to exhibit haploinsufficiency.

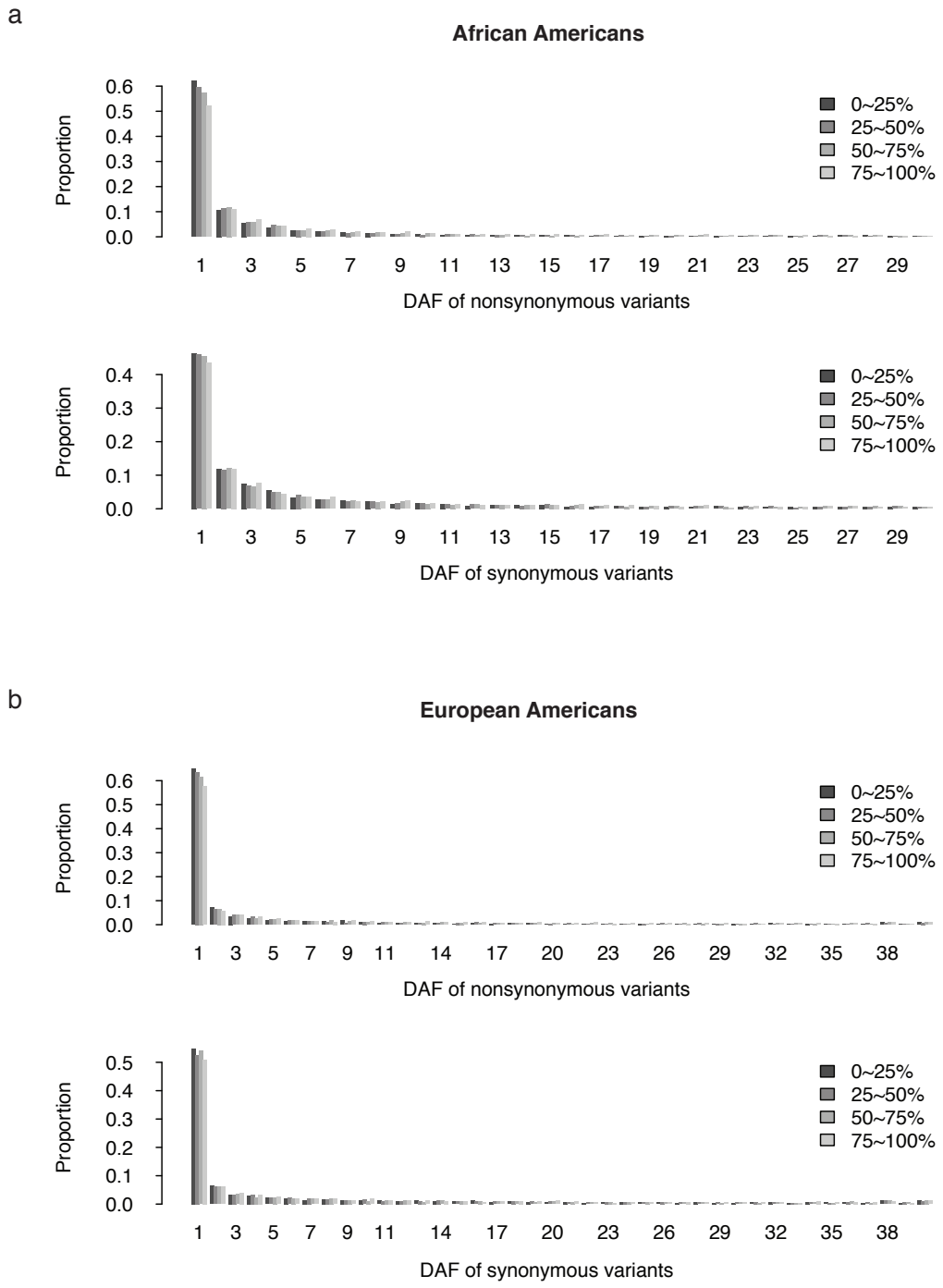

Figure 4.15: Derived allele frequency spectrum of variants in different gene sets. This figure shows the spectrum of derived allele frequency (DAF, represented here as counts of derived allele in the population) of nonsynonymous SNPs and synonymous SNPs discovered by resequencing of human genes in a) 15 African Americans and b) 20 European Americans. In each plot, DAF of variants located in genes of different p(HI) are compared side by side, where bars of decreasing darkness represent quantiles of decreasing p(HI), such that the 0–25% quartile is that with the highest probability of being haploinsufficient.

4.3.4 Probabilistic CNV diagnosis

In Section 4.3.2 and 4.3.3, I demonstrated the usefulness of gene-based p(HI) and its derivative, the interval-based haploinsufficiency LOD score in discriminating between benign and pathogenic deletions by showing that known pathogenic deletions have a LOD score distribution significantly higher than that of even the 'most deleterious' deletions found in apparently healthy individuals. However, in clinical diagnostics the primary question is how likely a variant is pathogenic/causal given all sources of evidence (*e.g.* pathogenic score). This is a typical Bayesian problem of which the answer is affected by both prior belief and evidence. Naturally, I modeled this problem using a Bayesian framework and tried to put it in the context of the general diagnostic process. I applied this framework to CNV diagnostics and examined two frequently encountered scenarios in clinical diagnostics wherein (*i*) the inheritance status of the variant is unknown or (*ii*) the variant is known to have arisen *de novo*.

4.3.4.1 A Bayesian framework for CNV diagnostics

The diagnostic question: 'is this variant, in this patient, sufficient to explain their clinical phenotype?' can be answered by assessing the posterior probability that this variant is causal given all the available evidence, $\mathrm{p}(C|E)$, where C denotes that the variant is causal and E denotes all available evidence. This probability is difficult to measure directly. Instead, the probability to observe such evidence given the variant is causal (and not causal), $\mathrm{p}(E|C)$ (and $\mathrm{p}(E|\bar{C})$), can be estimated directly from medical or population data and can be used to derive $\mathrm{p}(C|E)$ according to the Bayes Rule:

$$\mathrm{p}(C|E) = \frac{\mathrm{p}(C)\mathrm{p}(E|C)}{\mathrm{p}(E)} = \frac{\mathrm{p}(C)\mathrm{p}(E|C)}{\mathrm{p}(C)\mathrm{p}(E|C) + \mathrm{p}(\bar{C})\mathrm{p}(E|\bar{C})}$$

, where $\mathrm{p}(C)$ is the prior probability a variant is causal. Evidence involved in diagnosis of genetic variants includes both dichotomous or categorical conditions and continuous measurements. The former are often used as filters, such as 'overlapping with known disease-causing genes' and 'inherited from similarly affected parents'. The latter can be transformed into filters with defined thresholds, such as the division of common and rare variants based on population frequency thresholds, or used directly as numeric variables, such as pathogenic scores. Therefore, the space of evidence can be split into S, denoting that the variant has a measure of pathogenicity equal to x, and F, representing all other pieces

of evidence that can be used as filters. In this way, the posterior probability and the Bayes factor becomes $\mathrm{p}(C|S,F)$ and $\mathrm{p}(S,F|C)$, respectively. The latter can be further expanded to $\mathrm{p}(F|C)\mathrm{p}(S|C,F)$, so that

$$\mathrm{p}(C|S,F) = \frac{\mathrm{p}(C)\mathrm{p}(F|C)\mathrm{p}(S|C,F)}{\mathrm{p}(C)\mathrm{p}(F|C)\mathrm{p}(S|C,F) + \mathrm{p}(\bar{C})\mathrm{p}(F|\bar{C})\mathrm{p}(S|\bar{C},F)}$$

, or in its likelihood ratio form,

$$LR = \frac{\mathrm{p}(C|S,F)}{\mathrm{p}(\bar{C}|S,F)} = \frac{\mathrm{p}(C)\mathrm{p}(F|C)\mathrm{p}(S|C,F)}{\mathrm{p}(\bar{C})\mathrm{p}(F|\bar{C})\mathrm{p}(S|\bar{C},F)}$$

$$\mathrm{p}(C|S,F) = \frac{LR}{1+LR}$$

$\mathrm{p}(F|C)$ (or $\mathrm{p}(F|\bar{C})$) is the probability the variant passes this filter F given the variant is causal (or benign), and $\mathrm{p}(S|C,F)$ (or $p(S|\bar{C},F)$) is the probability of the variant having a measure of pathogenicity equals to x given it is causal (or benign) and passes the filter F.

$\mathrm{p}(F|C)$ can be estimated as the proportion of causal variants discovered in large patient studies that pass the filter, and $\mathrm{p}(S|C,F)$ can be estimated as the proportion of causal variants passing the filter that have a pathogenic measure equal to x. $\mathrm{p}(F|\bar{C})$ and $\mathrm{p}(S|\bar{C},F)$ are best estimated from all benign variants, from both patients and healthy individuals. In practice, benign variants are usually not reported and recorded in patient studies, and depending on the particular filter, F, such information is sometimes not collected for variants found in population-based or control studies (*e.g.* whether a variant is *de novo* or not). Therefore, $\mathrm{p}(F|\bar{C})$ and $\mathrm{p}(S|\bar{C},F)$ often have to be estimated from approximate distributions. Variants found in control individuals should be similar enough to all benign variants provided the sample size of the control cohort is large. For certain filters, the set of variants that pass them may be obtained through proxy properties. After the approximate variant sets are constructed, $\mathrm{p}(F|\bar{C})$ and $\mathrm{p}(S|\bar{C},F)$ can be estimated as for causal variants. With different F, these components need to be estimated from different sets of variants and the posterior probability changes accordingly. Below I consider two categories of possibly causal variant that are frequently encountered in clinical diagnostics: (*i*) the variant can be shown to be rare, but is of known inheritance status, and (*ii*) the variant can be shown to be *de novo*.

Table 4.4: Estimated parameters of the diagnostic framework

Variant type	Size range	F	$p(C)$	$p(F\|C)$	$p(F\|\bar{C})$
deletion	>180k	rare	0.12	1	0.34
deletion	>180k	*de novo*	0.12	0.73	0.0044
duplication	>330k	rare	0.14	1	0.38
duplication	>330k	*de novo*	0.14	0.73	0.0044

4.3.4.2 The variant is rare, and of unknown inheritance status

Under this scenario, often the only information on the variant is that it is not already known to be pathogenic and is not commonly seen in the population, therefore F denotes the filter that requires variants to be rare as defined by having a population frequency <1%. The estimated value of the parameters: $p(C)$, $p(F|C)$ and $p(F|\bar{C})$ were listed in Table 4.4 (see Methods). I considered either the LOD score or the variant size as the measure of pathogenicity. The distribution of LOD scores and variant sizes for rare casual and benign CNVs, from which $p(S|C,F)$ and $p(S|\bar{C},F)$ can be calculated, were shown in Figure 4.16–4.19. For both deletions and duplications, the resulting posterior probability $p(C|S,F)$ increases as the LOD score, or the size of the variant, becomes greater. In order to achieve a confidence level of 95%, a rare deletion of unknown inheritance status needs to be larger than 2.1Mb or have a LOD score greater than 7.2, and a rare duplication needs to be larger than 3.2Mb or with a LOD score greater than 15.5.

4.3.4.3 The variant is *de novo*

The *de novo* rate of causal and benign CNVs is even harder to obtain as confirming the *de novo* status would require the genotype information of both the parents and the child, *i.e.* the 'trio', and reaching a reasonable estimate requires genotyping a large number of such trios. There are a few studies that have reported CNV diagnosis in hundreds to more than a thousand patients including parents in which low-resolution array-CGH were used to detect large CNVs and *de novo* status were confirmed where possible [134, 161]. These studies are arguably the best sources from which one can estimate the *de novo* rate of causal CNVs. However, even with this data the number of *de novo* CNV from any one study is

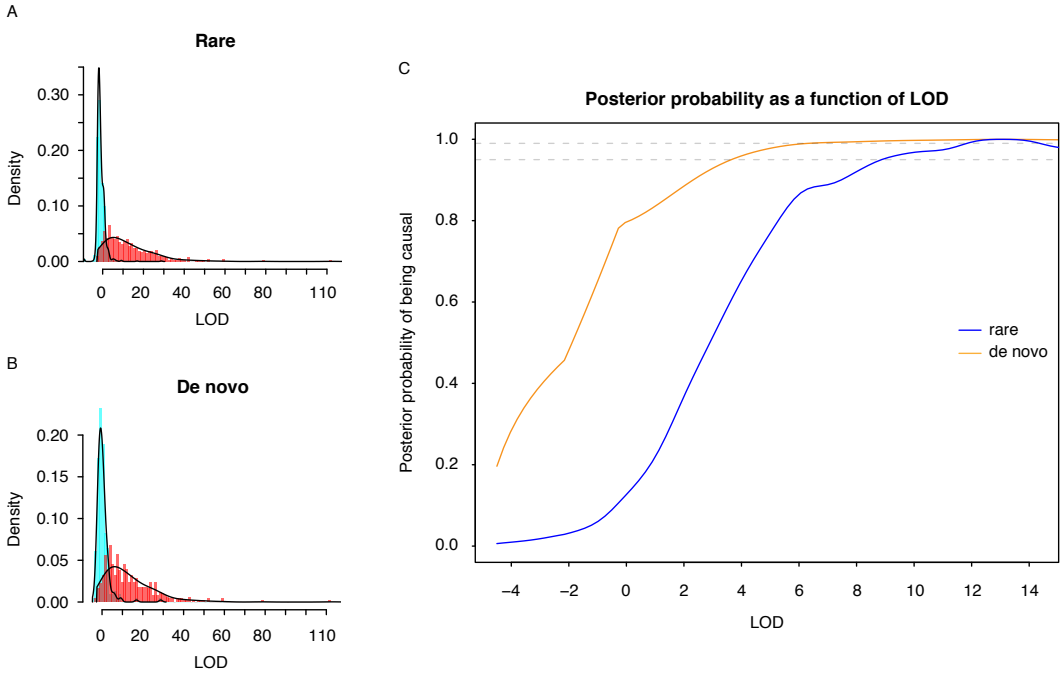

Figure 4.16: The posterior probability of a deletion being causal as a function of pathogenicity score. The distribution of pathogenicity score for causal (red) and benign (green) deletions are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

too small to generate a meaningful distribution of measure of pathogenicity. Therefore, the distribution of pathogenicity measures for *de novo* CNVs was approximated using known *de novo* causal CNVs recorded in DECIPHER. Studies reporting *de novo* CNVs discovered in apparently healthy individuals are even scarcer. I took the benign *de novo* rate from Itsara *et al*, which investigated the rate of *de novo* CNVs in 772 transmissions in pedigrees without neurocognitive disease genotyped on median- to high-resolution SNP genotyping arrays and I approximated the distribution of pathogenicity scores for benign *de novo* CNVs using singleton CNVs found in WTCCC2 and GAIN controls.

The estimated values of the parameters are show in Table 4.4 and the causal and benign distributions of measure of pathogenicity are shown in Figure 4.16–4.19. As expected, for both deletions and duplications, the size or the LOD score required for a variant to have a probability of being causal greater than 0.95 is much smaller than that required for a variant of which the inheritance status is unknown. However, being *de novo* alone does not guarantee pathogenicity as the probability of being a causal variant is still not convincingly high when the variant is small (0.8 at size = 500kb) or with very low LOD score (0.7 at LOD = -2).

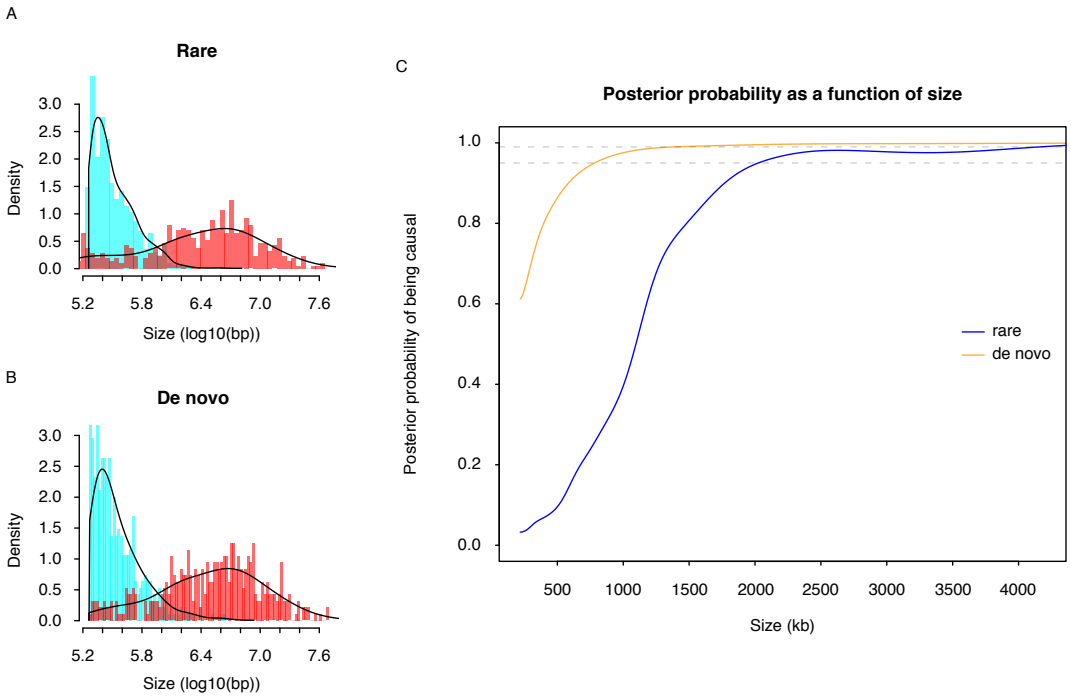

Figure 4.17: The posterior probability of a deletion being causal as a function of size. The distribution of pathogenicity score for causal (red) and benign (green) deletions are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

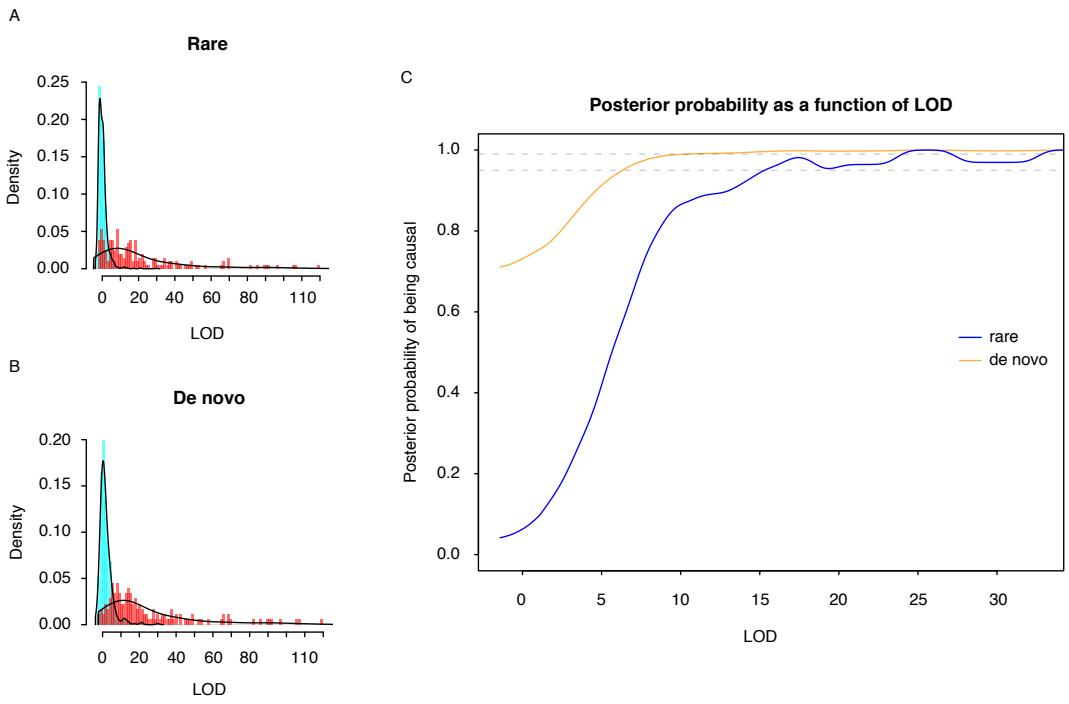

Figure 4.18: The posterior probability of a duplication being causal as a function of pathogenicity score. The distribution of pathogenicity score for causal (red) and benign (green) duplications are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

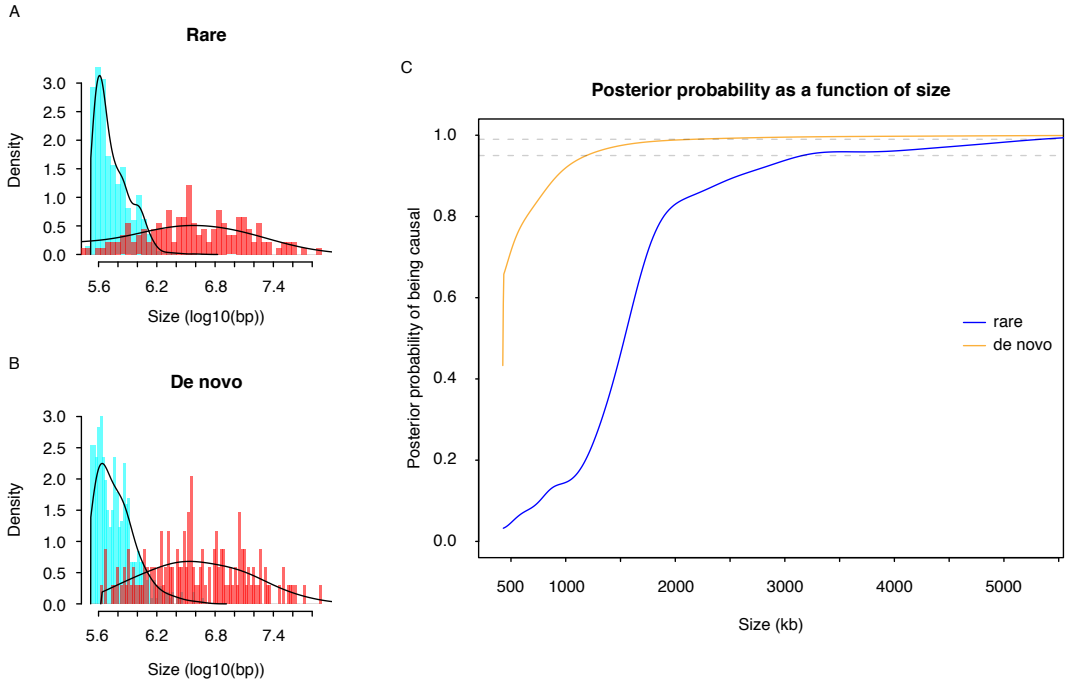

Figure 4.19: The posterior probability of a deletion being causal as a function of size. The distribution of pathogenicity score for causal (red) and benign (green) deletions are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

4.4 Discussion

In this chapter, I described the collection of human HI genes and HS genes, their differences in genomic, evolutionary, functional and network properties, and a computational method that distinguishes the two and predicts the probability of exhibiting haploinsufficiency for human protein-coding genes of unknown dosage sensitivity. A measure of pathogenicity for large genic copy number variants was developed on the basis of the HI predictions. A probabilistic diagnostic framework was designed to transform evidence of pathogenicity of a patient variant into confidence of diagnosis by taking into account the population variance of that measure of pathogenicity.

The traditional view that recessiveness is the norm of deleterious mutations is supported by earlier mutagenesis screen of model organisms [162]. In human, the $\sim$300 known HI genes only account for $\sim$1.5% of the protein-coding genome. However, haploinsufficiency, like most concepts in Mendelian genetics, is a qualitative, rather than quantitative, description based on a phenotype-specific definition of insufficiency, Insensitive or incomplete phenotyping or diagnosis could lead to underestimation of the proportion of the genome that is actually dosage sensitive. In genetics studies of model organisms, it is common that only the most prominent phenotypic consequence of a mutation or traits that are in relation with certain prior expectation are examined and reported. Abnormalities that are subtle and require specially designed tests to reveal or occur in completely unexpected tissues or cells can often be overlooked. Even in human, wherein measurements of physiological and morphological abnormalities is thought to be much more sensitive and thorough, complete phenotyping is never guaranteed. For example, the mutant allele of the gene *GJB2*, which is causal for the most frequent form of recessive congenital hearing loss, was recently found responsible for increased epidermal thickness in a dominant or semi-dominant manner [163, 164]. Thickened epiderm is obviously a less prominent trait that could not be detected without skin ultrasonography or similar technologies. In this chapter, the definition of haploinsufficiency has focused on severe clinical phenotypes (broadly-defined) as sufficiency relates to being qualified to be recruited as an apparently healthy control in a study of common disease susceptibility. With more complete phenotyping and hence a more stringent definition of sufficiency, the haploinsufficient/dosage-sensitive proportion of the genome might grow larger. In addition, most early work of Fisher, Wright and others that emphasized the dominance of the wildtype allele focused on metabolic enzymes. We now know that metabolic enzymes are less likely to be haploinsufficient whereas transcription factors, structural pro-

teins and subunits of protein complexes are more likely to be haploinsufficient due to the kinetic properties of the respective molecular system in which they function [107, 165, 166]. As transcription factors alone account 5–10% of the human protein-coding genome [167], the currently ∼300 known human HI genes is likely just a tip of the iceberg.

Not surprisingly, the known HI genes were found to be larger in size, which is a general characteristic of disease genes [168, 169], though it might be attributed to ascertainment bias, as, all things being equal, it is easier to find multiple families with causal mutations in the same gene if the gene is larger. HI genes were found to be more conserved in their coding sequence than HS genes, which is consistent with previous comparison between dominant and recessive disease genes [119]. In addition, the promoter sequences of HI genes are more conserved as well, which might suggest transcription regulation of these genes, as a part of dosage control mechanism, is under greater purifying selection, although this needs to be confirmed by human variation data. HI genes were found to have fewer paralogs and/or paralogs with lower sequence similarity than HS genes. This is consistent with a yeast study [150] which reported that HI genes tend not to have paralogs and suggested having a close paralog may provide a buffer against the effects of haploinsufficiency, but contradicts another report by Kondrashov *et al* [120] that found human dominant disease genes tend to have more paralogs than recessive disease genes and argued that such is the result of positive selection. However, the latter finding is not strictly comparable to this study, since homozygous LOF mutation of recessive disease genes can cause severe phenotypic defects and are hence under selection and less likely to be found in large genomic deletions, from which the HS gene set used in this study are collected. Indeed, there is a significant underrepresentation ($p = 0.0023$) of recessive disease genes in the HS gene set. The strong enrichment of olfactory receptor genes in the HS set (13% compared to 2% genome-wide, $p < 2.2\times10^{-16}$) could also affect the result. With respect to their spatiotemporal expression patterns, HI genes are more tissue specific and active during early development, which is expected since many of the haploinsufficient transcription factors play vital and tissue specific role in early developmental processes such as patterning, morphogenesis and organ development [156]. As for network properties, HI genes are found to be more central and closer to one another. The latter may support the view that haploinsufficiency tend to occur in certain molecular systems (early-development-related signaling and transcription regulation pathways, protein complexes), but may also be confounded by the ascertainment bias that search for novel disease genes tends to follow interaction partners of known disease genes.

The prediction of HI was implemented by training a statistical classifier on known HI and HS genes using gene properties that best distinguish the two as predictor variables. This is not a strictly mechanism-based approach, but an approach that exploits the correlation between haploinsufficiency and other gene properties. Though the performance of the prediction, as assessed by cross-validation using the training data, is moderately good (AUC = 0.81 without imputation of predictors and 0.84 with imputation; when requiring 80% sensitivity, the version without imputation has 70% specificity and version with imputation has 75% specificity), it is better than using any single gene property alone and has been validated to be able to prioritize potential real genes. Proximity to other known HI genes within gene or protein networks was found to be the most predictive property of which the contribution to performance cannot be fully explained by sequence conservation, tissue-specificity of expression, or other gene properties. Incomplete coverage of all genes in the genome by gene-gene and protein-protein networks is therefore also the major factor limiting the genome coverage of these predictions. The predictions should be substantially improved in both accuracy and coverage with the future generation of more complete and accurate human genetic interaction networks.

Although haploinsufficiency can be regarded as the property of a single gene, phenotypic manifestation of any genetic mutation, including heterozygous LOF mutation, is, strict speaking, the output of a perturbed multi-layer system of interacting molecules, cells and organs. Consequently, dosage sensitivity of a gene could vary across different genetic backgrounds. For example, heterozygous deletion of *Tbx5* causes embryonic lethality in 129S mice, but produces viable mice on B6 background [170]. In humans, patients carrying a second large CNV in addition to the micro-deletion at 16p12.1 exhibit much severer developmental delay than those having the 16p12.1 micro-deletion alone [91]. Therefore, the ideal prediction of haploinsufficiency should come from a system biology approach that models all interacting genes and biochemical reactions in a cell mathematically similar to that of Kacers and Burns [107] in which haploinsufficiency could be determined by numeric simulation and single component sensitivity analysis.

The measure of pathogenicity of a CNV was defined as the log of odds that at least one affected gene is haploinsufficient. As the likelihoods of being haploinsufficient of individual genes are combined in such a probabilistic way, its application is not limited to individual genomic intervals. For example, one can measure the genome-wide pathogenic burden of an individual by calculating the odds that at least one gene is haploinsufficient out of all genes affected by any CNV, or other LOF variants, in this individual's genome. However,

there are also obvious caveats of such measure. First, the measure can only be applied to CNVs affecting protein-coding genes for which a prediction is available. Second, potential functions of intergenic sequences are ignored. Third, the effects of each gene are assumed to be independent. To tackle the first two limitations, one could consider features that are not bound to genes, such as the density of repeat elements or the number of conserved non-coding elements. However, these properties need to be combined with the likelihoods of genes exhibiting haploinsufficiency in a meaningful way. For the third caveat, the idea solution would again be a system biology approach as described above, substituting the single-component sensitivity analysis with a multiple-component sensitivity analysis.

The probabilistic diagnostic framework provides a natural way to integrate both qualitative and quantitative measures of pathogenicity and produces quantified confidence of diagnosis by considering the population variance of the quantitative measure of pathogenicity. Being a Bayesian method, it has the advantage of not naively assuming that different measures are independent, but at the same time it requires knowledge of the conditional distribution of the quantitative measure of pathogenicity, which is not always readily available. In its application to rare and *de novo* CNVs in Section 4.3.4, the patient and control distribution of pathogenicity score under the condition that the CNVs are *de novo* were unavailable and were substituted with approximated distributions. Another problem, which is common for all Bayesian inferences, is the requirement of a proper prior. The prior probability of a variant being causal can be affected by a number of factors, for example, the specific type of disease and the filters or tests having been applied before the application of this framework. As for CNVs, since different CNV discovery platforms vary vastly in their sensitivity and resolution, which could have profound impact on the population distribution of the quantitative measure of pathogenicity, the prior should be estimated from the same or similar platform that the population distribution of pathogenicity scores is generated.

Previously, *de novo* CNVs discovered in patients were highly likely to be diagnosed as being a causal variant in clinical practice. As early CNV discovery technologies, such as cytogenetic methods and low-resolution array CGH could only find very large events, those diagnoses might largely hold correct. However, in recent years, with improved CNV discovery technology and accumulated CNV datasets, it is known that *de novo* CNVs, especially smaller ones, arise at an appreciable rate (estimates ranging from 1×10^{-2} to 3×10^{-2} CNVs per haploid genome per generation [89, 131, 171]) in healthy individuals. Therefore, there is growing recommendation for not relying solely on the *de novo* status in the interpretation of variant causality [133, 172]. My application of this diagnostic framework to *de novo* CNVs

not only supported this view, but also provides a quantitative level of confidence as a function of the size of the variant or its pathogenicity score. However, these quantitative values should be interpreted with caution at this stage, and are not mature enough for clinical implantation, for several reasons. First, as the distribution of CNVs and functional sequences is uneven across the genome whereas the size or the pathogenicity score of CNVs are locus-independent measures. In addition, these results are highly dependent on the CNV discovery platform and the prior. Furthermore, the use of approximate conditional distributions of pathogenicity scores has introduced additional uncertainty. With the increasing application of array CGH, high-resolution genotyping array and medical sequencing, and hence ascertainment of a more complete spectrum of variants in patient genomes, this diagnostic framework is expected to produce a more accurate estimation of confidence to aid the diagnosis of novel, rare variants for which detailed locus-specific information is unavailable.