Bioinformatics, 2016, 1–7
doi: 10.1093/bioinformatics/btw381
Advance Access Publication Date: 26 June 2016
Original Paper

OXFORD

Genome analysis

Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes

Jacob Shujui Hsu[1], Johnny S.H. Kwan[1], Zhicheng Pan[1], Maria-Mercè Garcia-Barcelo[2], Pak Chung Sham[1,3] and Miaoxin Li[1,3,*]

[1]Department of Psychiatry, [2]Department of Surgery and [3]Centre for Genomics Science, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on January 11, 2016; revised on May 20, 2016; accepted on June 14, 2016

Abstract

Motivation: Exome sequencing studies have facilitated the detection of causal genetic variants in yet-unsolved Mendelian diseases. However, the identification of disease causal genes among a list of candidates in an exome sequencing study is still not fully settled, and it is often difficult to prioritize candidate genes for follow-up studies. The inheritance mode provides crucial information for understanding Mendelian diseases, but none of the existing gene prioritization tools fully utilize this information.

Results: We examined the characteristics of Mendelian disease genes under different inheritance modes. The results suggest that Mendelian disease genes with autosomal dominant (AD) inheritance mode are more haploinsufficiency and *de novo* mutation sensitive, whereas those autosomal recessive (AR) genes have significantly more non-synonymous variants and regulatory transcript isoforms. In addition, the X-linked (XL) Mendelian disease genes have fewer non-synonymous and synonymous variants. As a result, we derived a new scoring system for prioritizing candidate genes for Mendelian diseases according to the inheritance mode. Our scoring system assigned to each annotated protein-coding gene ($N = 18\,859$) three pathogenic scores according to the inheritance mode (AD, AR and XL). This inheritance mode-specific framework achieved higher accuracy (area under curve $= 0.84$) in XL mode.

Conclusion: The inheritance-mode specific pathogenicity prioritization (ISPP) outperformed other well-known methods including Haploinsufficiency, Recessive, Network centrality, Genic Intolerance, Gene Damage Index and Gene Constraint scores. This systematic study suggests that genes manifesting disease inheritance modes tend to have unique characteristics.

Availability and implementation: ISPP is included in KGGSeq v1.0 (http://grass.cgs.hku.hk/limx/kggseq/), and source code is available from (https://github.com/jacobhsu35/ISPP.git).

Contact: mxli@hku.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole-exome sequencing has been widely used for the identification of disease-associated genetic variants. For severe and rare Mendelian diseases, it has been proven to be an effective approach (Agha *et al.*, 2014). Exome sequencing may also contribute to the understanding of the etiology of complex diseases, e.g. Schizophrenia (Purcell *et al.*, 2014) and Type 2 Diabetes (Lohmueller *et al.*, 2013). However, the functional annotation of the protein-coding genes is still far from complete. By November 2014, 19 007 protein-coding genes had been registered in the HUGO Gene Nomenclature Committee database (Gray *et al.*, 2013), which includes most of the relevant DNA regions responsible for severe Mendelian diseases and some complex diseases. On the other hand, the data from the 1000 Genomes Project indicate that an apparently healthy person could carry approximately 250–300 loss-of-function (LoF) single-nucleotide variants (SNVs) on average (1000 Genomes Project Consortium *et al.*, 2010; The 1000 Genomes Project Consortium, 2012). Therefore, the accurate estimation of both gene and variant pathogenicity and corresponding etiologic architecture pose a considerable challenge to human genetic community.

A number of bioinformatics tools based on predicted deleteriousness and distinct biological features have been developed to prioritize protein-changing (i.e. non-synonymous) variants in the human protein-coding genes. The tools FATHMM (Shihab *et al.*, 2013), GERP (Cooper *et al.*, 2005), LRT (Sung and Fay, 2009), MutationAssessor (Reva, *et al.*, 2011), SIFT (Ng and Henikoff, 2003), SiPhy (Garber *et al.*, 2009) and PROVEAN (Choi *et al.*, 2012) mainly consider sequence similarities and conservation patterns across various species, whereas CADD (Kircher *et al.*, 2014), MutationTaster (Jana Marie *et al.*, 2014) and Polyphen-2 (Ivan *et al.*, 2010) consider the functional information of the genetic variations from multiple different information sources in their algorithms. Moreover, the mutation significance cutoff uses gene-level dynamic thresholds generated by CADD, PolyPhen-2 or SIFT to reduce false-negative rate (Yuval *et al.*, 2016). These tools are particularly useful in clinical sequencing studies of Mendelian diseases as to exclude a large number of neutral variants. However, inconsistencies in prediction results among all these different variant-based prediction tools have been demonstrated (Li *et al.*, 2012) and results from variant-based functional prediction are hard to be generalized into the gene level.

A few gene scoring systems have been proposed to estimate the effect of genetic variations at the gene level. First, Haploinsufficiency (HI) score (Huang *et al.*, 2010) combines a list of biological properties by inspecting the copy number variations (CNV) among thousands of healthy individuals. In this system, genes with a higher HI score may imply their potential of causing dominant traits. Second, Recessive (REC) score (MacArthur *et al.*, 2012) uses 213 known LoF tolerant genes and 858 known recessive disease genes to build up a linear discriminant model to estimate the probability that a gene can cause a recessive disease. Third, genic intolerance (RVIS) score (Petrovski *et al.*, 2013) assesses functional genetic variation tolerance ability from profiles of SNVs with allele frequency information. Fourth, network indispensability (NET) score (Khurana *et al.*, 2013) calculates the gene centrality and indispensability in various protein–protein interactions and regulatory networks to dissect the gene importance. Fifth, gene constraint (CONS) score (Samocha *et al.*, 2014) compares the genomic background mutation profile with the observed mutation profile of a gene for interpreting *de novo* mutations. Lastly, human gene damage index (GDI) (Itan *et al.*, 2015) also utilizes the mutation profile in both monogenic disease patients and the general population to prioritize exome variants. However, each of these popular gene scoring systems can only explain part of the genetic architecture from different perspectives. For example, (i) HI estimation does not include the profiles of non-CNV genetic variants; (ii) the REC score does not consider the dominant disease-predisposing genes; (iii) the RVIS score does not deal with variations of allele frequencies in different populations; (iv) the systematic comparison for different known disease-associated genes is missing in NET score study and the sample size of LoF-tolerant and essential genes are moderated; (v) the CONS score is only applicable for the interpretation of *de novo* mutations and (vi) the GDI score only considers mutation profiles. Therefore, more sophisticated analysis and comprehensive prioritization schemes on genes causing human disease are needed.

One of the most important characteristics of a Mendelian disease is that its mode of inheritance can be inferred from the pedigree. However, only a few studies have partially examined the unique biological features of genes causing the different inheritance modes of Mendelian diseases (Huang *et al.*, 2010; MacArthur *et al.*, 2012). Therefore, we developed a new gene-level scoring system, Inheritance mode Specific Pathogenicity Prioritization (ISPP) for the prediction of inheritance-specific pathogenicity. In addition to the consideration of the different inheritance modes, ISPP also combines six existing gene-level prioritization systems with many gene features. The performance of the proposed approach was evaluated systematically by multiple curated disease-associated gene sets.

2 Materials and methods

2.1 Collection of gene features and benchmark datasets

2.1.1 Variant level, gene level and functional-related gene features

Population variant profiles in protein-coding genes were obtained from the dataset of NHLBI GO Exome Sequencing Project (ESP6500), which includes data from more than 200 000 individuals from over 20 clinical sequencing studies. We extracted the variant profiles of all protein-coding genes by using KGGSeq (Li *et al.*, 2012) and counted the number of non-synonymous (Nonsyn_v) and synonymous (Syn_v) variants in each gene with approved gene symbol from the HUGO Gene Nomenclature Committee. Moreover, we retrieved 18 biological gene features from the Ensembl database (Flicek *et al.*, 2014) through the BioMart system. These features include gene length (Length), GC content (GC), total number of RNA isoforms (ALL_tr_count), mean length of all RNA isoforms (ALL_tr_length), mean length of the protein coding RNA isoforms (CO_tr_length), number of protein coding (CO_tr_count) and non-coding (NC_tr_count) RNA isoforms (Kasprzyk, 2011). In addition, we also gathered functional-related gene features (Lawrence *et al.*, 2013), including (i) global expression (Expr) derived from RNA-Seq data and summed across the tens different cell lines, (ii) DNA replication time (Reptime), (iii) the Hi-C statistic (HIc), a measurement of chromatin statuses as well as (iv) the noncoding mutation rate (NC_mut_rate), which measured the mutation rate of intronic regions. Moreover, published gene prioritization scores, HI (Huang *et al.*, 2010), REC (MacArthur *et al.*, 2012), RVIS (Petrovski *et al.*, 2013), NET (Khurana *et al.*, 2013), CONS (Samocha *et al.*, 2014) and GDI (Itan *et al.*, 2015) were downloaded from either dbNSFP (Liu *et al.*, 2013) or their corresponding studies. Tissue-specific expression information (Tissuespe) was obtained from Liu et.al (Liu *et al.*, 2008). See Supplementary Table S1 for details.

2.1.2 Benchmark disease-associated gene sets

Based on the OMIM database (URL:http://www.omim.org/Online Mendelian Inheritance in Man), 10 lists of experts-curated disease genes were aggregated as hOMIM gene lists (Blekhman *et al.*, 2008). They are hOMIM_AD [419 autosomal dominant (AD) disease genes], hOMIM_AR [569 autosomal recessive (AR) disease genes], hOMIM_AD_AR (39 both AD and AR disease genes), hOMIM_XL (66 X-linked disease genes), hOMIM_Birth (637 genes for birth-onset diseases), hOMIM_Pre15 (9 genes associated with diseases with onset age before 15), hOMIM_Pre-rep (368 genes associated with diseases with the onset age after 15 and before 40), hOMIM_Post40 (33 genes associated with diseases with onset age after 40), hOMIM_Cancer (548 cancer genes from the COSMIC database), and hOMIM_essential (hOMIM_ess) (1643 genes lethal if knocked-out in mice). We merged hOMIM_Birth and hOMIM_Pre15 into hOMIM_early (646 genes for diseases with the onset age before 15 or birth) since hOMIM_Pre15 only has nine genes. In addition, we collected gene lists about different disease contexts from the RVIS study, including RVIS_HI (Haploinsufficiency, 175 genes), RVIS_AD_Neg (Dominant negative, 364 genes), RVIS_AR (Recessive, 818 genes), RVIS_Denovo (Denovo, 467 genes) and RVIS_Denovo_HI (Denovo plus Haploinsufficiency, 109 genes). Two other gene lists from the Mouse Genome Informatics (MGI) database (Blake *et al.*, 2014), MGI_Lethality (92 genes) and MGI_Seizure (95 genes) genes, were also selected. Furthermore, we constructed a non-essential gene list from dbNSFP dataset. The non-essential gene (Non_ess) must have at least one non-synonymous variant: (i) with a minor allele frequency $q > 3.3\%$ (816 genes), assuming that the frequency of homozygous recessive in the population is $> 0.1\%$; as well as (ii) predicted to be as deleterious by at least four aforementioned variant-based functional prediction algorithms.

Likewise, similar but extended gene lists for diseases with different inheritance modes and onset ages were obtained from the Clinical Genomic Database (CGD) (Solomon *et al.*, 2013): CGD_AD (876 AD disease genes), CGD_AR (1500 AR disease genes), CGD_AD_AR (304 genes with both AD and AR inheritance modes), CGD_XL (182 X-linked Mendelian disease genes), CGD_Adult (69 genes) and CGD_Paediatric (1401 genes). We also collected the genes for both complex and Mendelian disorders characterized by Jin *et al.* (2012). MC (Mendelian and Complex diseases, 525 genes), MNC (Mendelian but Not Complex diseases, 445 genes) as well as CNM (Complex but Not Mendelian diseases, 2594 genes). See Supplementary Table S1 for details.

2.2 Permutation test for exploring pattern-specific model

To assess the biological and functional features of each disease-associated gene set, we computed the mean of each feature for each set and compared them with randomly generated gene sets. For each disease-associated gene set, we (i) computed the observed mean (X_i) of each feature i; (ii) obtained the sampling distribution under null hypothesis and the sampling mean μ_i and variance σ_i of each feature i from selecting 100 000 random gene sets with the same gene number as the observed gene set; (iii) calculated a z-score for feature i of the disease-associated gene set, Z_i, based on the central limit theorem, i.e. $Z_i = (X_i - \mu_i)/\sigma_i$ (Fig. 2, Supplementary Table S2); and (iv) calculated the corresponding P-value of the z score. A Bonferroni-corrected P-value threshold of $0.05/520 \approx 9.615 \times 10\text{E-}05$ was used to declare significance.

2.3 Machine learning model construction from selected gene features

For differentiating gene sets with different inheritance modes and age of onset, we used the random forest algorithm which was implemented in Weka package (http://www.cs.waikato.ac.nz/ml/weka/) to construct the machine learning models with different combinations of gene features. By using CGD database as the training dataset, we constructed all possible combinations of biological features to test the performance of the model. We compared the critical predictors when different feature subsets can achieve similar performance. The 10-fold cross-validation was used to assess the performance of the models. The best model was defined as the one with the largest area under curve (AUC) of Receiver Operating Characteristics (ROC). In the ISPP_XL model, the cross validation which were only sampled from those genes on X chromosome.

3 Results

In this study, we first collected 30 known disease-associated gene lists from independent studies or public databases as the benchmark dataset for building and testing the inheritance-specific disease gene prediction models. While various, these disease-associated gene lists are not necessarily mutually exclusive. We included nine gene groups of different inheritance modes and eight gene groups for disease with different ages of onset. Besides, we collected 14 gene features as well as six widely used gene prioritization scores as predictors in our prediction models (see Section 2 for details). The data collection and analysis framework are summarized in Supplementary Table S1 and Fig. 1.

3.1 Existing gene prioritization scores unsatisfactory distinguish among the disease-associated gene lists

We first compared the performance of six popular gene-based prioritization methods (HI, REC, RVIS, NET, CONS and GDI) which assess gene pathogenicity using different perspectives across the 26 disease-associated gene lists. Spearman correlation analysis suggests that these six methods only have modest correlation with each other ($r < 0.46$) except for the correlation between HI and REC ($r = 0.77$). (Supplementary Figs. S1 and S2, Table S1 and Text S1).

We found that the REC score had the best performance in predicting disease-predisposing genes among all six gene prioritization scores examined (Fig. 2). However, the REC scores were insensitive to inheritance modes even though the program was designed to carry out the recessive disease causation probability. For instance, both hOMIM_AD (Z score = 23.9) and hOMIM_AR

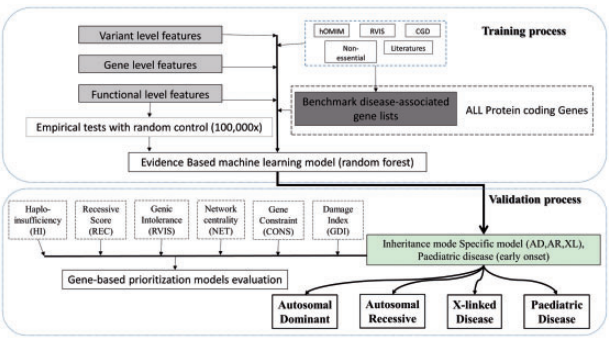

Fig. 1. Schematic diagram illustrating the construction of the inheritance mode-specific gene prioritization model

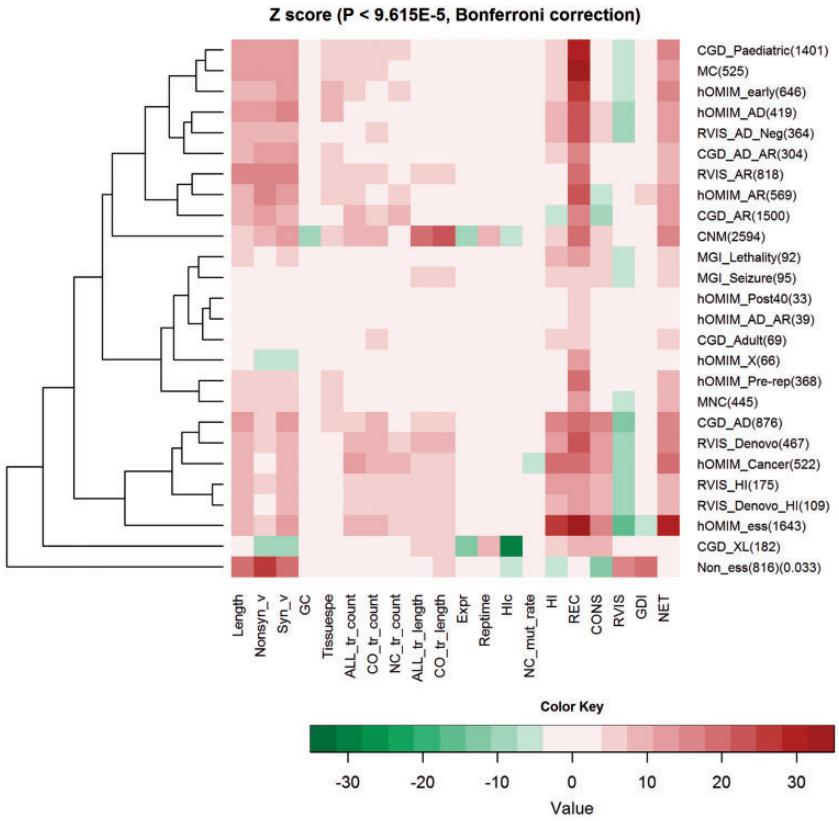

Fig. 2. Gene characteristics versus disease-associated gene lists heat map (with clustering based on the empirical significance levels). Red (solid) colors represent higher value than random gene sets of the same size. The attached dendrogram on the left depicts hierarchical clustering results for the disease-associated gene lists. Gene lists with same mode of inheritance could be clustered together by exhibiting similar gene feature patterns. See Section 2 for details

(Z score = 21.8) had significantly high REC scores compared to the random gene sets in permutation test; also, hOMIM_AD even had a marginally higher REC score (recessive disease causation probability) than hOMIM_AR (Rank sum P-value = 0.061) (Supplementary Fig. S2). Similarly, CGD_AD had significantly higher REC scores than CGD_AR (Wilcoxon Rank-Sum P value = 1.554E-07).

For the HI score, the performance in classifying AR genes from random gene sets was poor. All recessive disease-associated genes in the lists RVIS_AR, hOMIM_AR and CGD_AR, as well as Non_ess genes had no difference or significantly lower HI score compared to random gene sets (Supplementary Table S3 and Text S2). Also, in HI scores, the AUC of CGD_AD and CGD_AR training gene sets were only, respectively, 0.69 and 0.49 (Fig. 3). The NET scores estimated the indispensability of each gene by interaction network with moderate performance on disease-associated genes. The AUC of NET scores can reach 0.7 and 0.62 on CGD_AD and CGD_AR, respectively.

The purpose of the RVIS score is to estimate the tolerance of a gene with functional variations; as a result, most of the disease-associated gene lists had significantly lower RVIS scores than random, especially hOMIM_AD (P = 5.91E-14), hOMIM_early (P = 1.11E-10) and RVIS_Denovo (P = 4.68E-26). However, the RVIS scores failed to separate recessive disease-associated genes (RVIS_AR, hOMIM_AR and CGD_AR), and the AUC only had, respectively, 0.64 and 0.54 in CGD_AD and CGD_AR (Fig. 3). The similar idea was used in GDI score to build a gene damage index from the mutation profile of those known diagnosed monogenic disease subjects. However, the AUC of GDI score only had 0.45 and 0.58 in CGD_AD and CGD_AR, respectively.

The CONS score aims at estimating the *de novo* mutation effect of a gene by using the *de novo* mutation rate of the genomic background as reference. A CONS score higher than that of the genomic background suggests that the gene under consideration is sensitive to *de novo* mutations, hence more prone to be pathogenic. The gene lists hOMIM_AD and CGD_AD tend to have significantly higher CONS scores than random (P = 2.62E-10 and P = 4.97E-51, respectively), implying that these genes are more vulnerable to *de novo* mutations (Supplementary Text S2), and this is consistent with the definition of dominant diseases. However, the AUC of CONS score only had 0.65 and 0.43 in CGD_AD and CGD_AR, respectively (Fig. 3).

3.2 Each disease-associated gene list had its own unique pattern of biological and functional characteristics

We observed that most disease gene lists examined consistently had significantly more non-synonymous variants, synonymous variants and a longer gene length than random in the permutation test (Fig. 2, Supplementary Table S3). However, there was no clear variant profile for classifying disease genes according to just the different inheritance modes. The only exception was the X-linked disease-associated genes. Both hOMIM_XL and CGD_XL with average gene length (P = 5.9E-02, 5.6E-04, respectively; Bonferroni-corrected P-value ≈ 9.615E-05) had significantly fewer non-synonymous variants (P = 2.79E-10, 1.36E-10, respectively) and synonymous variants (P = 3.72E-25, 1.09E-25, respectively). We also found that recessive disease-associated genes (hOMIM_AR, RVIS_AR and CGD_AR) had significantly much more RNA

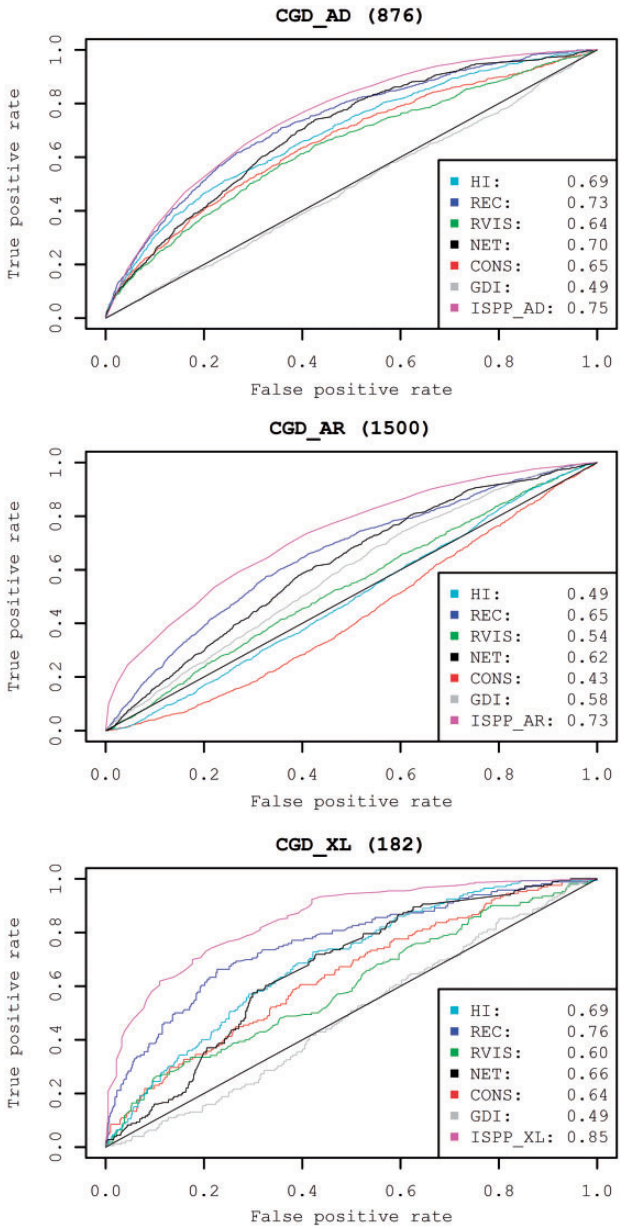

Fig. 3. Performance comparisons of different gene prioritization methods. ROC curves of seven gene prioritization approaches' capacity to predict the corresponding independent disease-associated gene lists. ISPP, 10-fold cross-validation of corresponding machine learning model

non-coding isoforms ($P \leq 2.70E-8$). Moreover, for hOMIM, RVIS and CGD studies, all those recessive disease-associated genes had more non-synonymous variants than that corresponding to dominant disease-associated genes ($P \leq 5.69E-40$) (Supplementary Text S3 and S4). Besides, genes in hOMIM_AR and CGD_AR had significantly smaller CONS scores than random gene sets ($P = 9.82E-8$ and 1.24E-15, respectively), suggesting that the recessive disease-associated genes may be less sensitive to *de novo* mutations. Surprisingly, there was no statistical difference for CONS score among CGD_Paediatric, MC and hOMIM_early, suggesting that *de novo* mutations may have a benign effect on these early-onset disease-associated genes. Furthermore, CGD_AD, Genic_Denovo, hOMIM_Cancer and hOMIM_essential had higher *Z* scores for HI and CONS scores than CGD_Paediatric and hOMIM_early, implying that dominant disease inheritance mechanisms are more likely to be haploinsufficiency or *de novo* mutation sensitive.

3.3 Development of a disease inheritance mode-specific gene prioritization system, incorporating the unique patterns of biological and functional characteristics

Here, we have built a unified framework by using the random forest approach to combine multiple biological features and prioritization scores for predicting pathogenic genes with specific inheritance modes. The training datasets included CGD_AD (876 genes), CGD_AR (1500 genes) and CGD_XL (182 genes). Ten-fold cross-validation was adopted to compare the performance of the prediction models, which was quantified by the AUC (see Section 2). For dominant diseases, our combined model can achieve slightly better performance in CGD_AD (with an AUC of 0.75) compared to the individual prediction systems. Surprisingly, the REC scores performed much better than HI scores in separating dominant disease genes from random genes with AUC of 0.73 and 0.69, respectively (Fig. 3). For recessive diseases, our combined model had much better performance than other prediction algorithms in CGD_AR with an AUC of 0.73. Moreover, the performance of our model was considerably superior to that of the other methods regarding the prioritization of X-linked disease-associated genes with an AUC of 0.85 in CGD_XL. However, the AUC of our combined model was similar to the AUC values obtained by REC scores in hOMIM_early and CGD_Paediatric (Supplementary Fig. S3). Finally, we have annotated 18 859 protein-coding genes and provided a set of pathogenic scores on each gene (Supplementary Table S4) for dominant, recessive and X-linked (ISPP_AD, ISPP_AR and ISPP_XL scores).

3.4 Validation of the new gene prioritization system

We validated the performance of our combined prediction model by a series of independent gene sets. Venn diagrams (Supplementary Fig. S4A) show the degree of overlap among various training data (CGD_AD, CGD_AR and CGD_XL) and other reported recessive disease-associated gene lists (hOMIM_AR, REC recessive and RVIS_Recessive). There were, respectively, 128, 107 and 206 recessive disease genes which were reported independently but not be included in our training sets. We further compared the discrepancies of their ISPP_AD and ISPP_AR scores by Wilcoxon rank sum test. In these three recessive gene sub-sets, the ISPP_AR scores from the combined model are significantly higher than the of ISPP_AD scores for all of the non-overlapping sub-sets ($P \leq 2.206E-5$) (Supplementary Fig. S4B).

For each inheritance mode, we demonstrated the ISPP scores for five novel disease-causal genes which recently published by independent next-generation sequencing studies. None of these genes were included in our positive training gene sets. As listed in Table 1, all of the genes have the highest pathogenic prediction scores in the corresponding inheritance modes and the scores are consistent with the mode of inheritance initially described per each gene. Additionally, we validated the results in one of the recent largest X-exome sequencing study on intellectual disability disease (Hu *et al.*, 2015). The study reported seven novel and validated X-linked Intellectual Disability (XLID) genes (*CLCN4, CNKSR2, FRMPD4, KLHL15, LAS1L, RLIM* and *USP27X*) as well as two novel candidate XLID genes (*CDK16* and *TAF1*). Seven of nine genes are on the top 32% of the ISPP_XL prediction model in ISPP_XL score system. *TAF1* is in the CGD_XL training set and has highest score on ISPP_XL score among other inheritance modes (Supplementary Table S4). Even though *RLIM* and *USP27X* have relatively low predicted ISPP_XL score, they still have highest predicted score among

Table 1. Examples of published candidate genes with mode of inheritance and corresponding ISPP scores

Gene	ISPP_AD	ISPP_AR	ISPP_XL	Ref
AD				
APP	0.17(5%)	0.01(69%)	—	Conidi *et al.* (2015)
IRS1	0.12(6%)	0.04(29%)	—	Rong *et al.* (2015)
RET	0.17(5%)	0.01(69%)	—	Figlioli *et al.* (2012)
NCOR1	0.23(5%)	0.01(80%)	—	Fozzatti *et al.* (2011)
EGR2	0.20(5%)	0.01(76%)	—	Lupski *et al.* (2010)
AR				
KMT2B	0.01(48%)	0.03(43%)	—	Agha *et al.* (2014)
APOB	0.11(6%)	0.25(8%)	—	Hammer *et al.* (2013)
GJB2	0.02(37%)	0.07(17%)	—	Nikolay *et al.* (2011)
FECH	0.01(46%)	0.18(8%)	—	Balwani *et al.* (2013)
ATF6	0.02(28%)	0.1(12%)	—	Ansar *et al.* (2015)
X-linked				
POLA1	0.05(13%)	0.01(74%)	0.34(19%)	NIH
OGT	0.02(34%)	0.00(80%)	0.28(21%)	Niranjan *et al.* (2015)
GLUD2	0.01(57%)	0.00(80%)	0.18(24%)	Cukier *et al.* (2014)
BRCC3	0.00(63%)	0.02(61%)	0.16(26%)	Huang *et al.* (2015)
ZMYM3	0.00(63%)	0.00(80%)	0.25(22%)	Philips *et al.* (2014)
CLCN4	0.02(32%)	0.00(80%)	0.21(22%)	Hu *et al.* (2015))
CNKSR2	0.00(63%)	0.00(80%)	0.12(32%)	Hu *et al.* (2015)
FRMPD4	0.00(63%)	0.04(34%)	0.19(24%)	Hu *et al.* (2015)
KLHL15	0.03(27%)	0.00(80%)	0.19(24%)	Hu *et al.* (2015)
LAS1L	0.02(34%)	0.01(66%)	0.13(30%)	Hu *et al.* (2015)
RLIM	0.00(63%)	0.00(80%)	0.05(47%)	Hu *et al.* (2015)
USP27X	0.00(63%)	0.00(80%)	0.03(57%)	Hu *et al.* (2015)
CDK16	0.00(63%)	0.01(74%)	0.15(28%)	Hu *et al.* (2015)

By comparing ISPP scores of each gene, the possible inheritance mode of the gene could be estimated (Supplementary Text S6). The value in parentheses indicates the percentage among all in the inheritance-specific model. NIH indicates that this X-linked disease candidate gene has not been published yet but has been funded for conducting the research by NIH. Project no.:1R56AI113274-01.

other ISPP scores. This suggests that the consideration of inheritance mode can enhance the pathogenicity prediction.

4 Conclusions

By examining a range of well-studied and independent disease-associated gene lists, our results suggest that most of the existing methods cannot satisfactorily distinguish gene pathogenicity and inheritance mode together among all the various gene lists. In this study, we investigated the characteristics of Mendelian disease genes with different inheritance modes. We found that AD disease-associated genes tend to be *de novo* mutation sensitive and are prone to haploinsufficiency, whereas AR disease-associated genes are likely to have more non-synonymous variants and non-coding RNA isoforms. Moreover, the X-linked disease-associated genes have significantly fewer non-synonymous and synonymous variants. This framework is splendid to combine RNA level information with current gene prioritization methods for predicting pathogenic protein coding gene under different inheritance modes.

The correlation analysis indicated that the overall expression level, Hi-C status and GC content had a positive correlation with each other (Supplementary Text S3). Moreover, almost all of the disease-associated genes had much more RNA isoforms than randomly selected gene sets (Supplementary Text S4). Even though the information bias might be one of the confounding factors, some disease-associated gene lists are not as significant as others such as hOMIM_AD and CGD_XL. Additionally, the recessive disease-associated genes, in particular, had more complicated regulatory processes and genetic variant tolerant ability, and the early-onset disease-associated genes may have *de novo* mutations with merely neutral effect. Apart from GDI scores, none of the gene prioritization scores examined had a strong correlation with the gene features we investigated (Supplementary Texts S1 and S3). This suggests that the gene features we included have no predominant effect on the combined machine learning prediction model. Interestingly, we found that the REC score can well predict all of the disease-associated gene lists with significantly higher recessive disease causation probability (REC, $P \leq 1.99E\text{-}8$). However, it seemed no power for REC to discriminate dominant and recessive disease genes.

Figure 2 shows the dendrogram of cluster analysis based on the empirical Z scores for comparing gene feature patterns among 26 disease-associated gene lists (Supplementary Text S5). All recessive disease-associated gene lists from three independent studies were clustered together, but there are only 165 overlapping genes in these gene lists. All of these three lists had no difference or significantly lower CONS, RVIS and HI scores (Fig. 2 and Supplementary Table S3). On the other hand, based on the cluster analysis, dominant disease-associated gene lists (including CGD_AD, RVIS_Denovo, hOMIM_Cancer and hOMIM_essential gene sets) had significantly higher Z scores for various transcriptional features, and HI and CONS scores. Besides, it is also fair and complementary to use other prediction methods genome-wide regardless the mode of inheritance. Each prediction has its own proposed usage which might not be fully illustrated by ISPP scores only. In conclusion, our inheritance mode-specific prediction model was able to assess the potential pathogenicity of a gene accurately. It also reflects the various biological characteristics behind genes for diseases with different inheritance modes. These information could provide valuable information for gene annotation.

Acknowledgements

The authors acknowledge all public available datasets and their contribution in this study as well as all of the resources provided from Li Ka Shing Faculty of Medicine, The University of Hong Kong.

Funding

This work was supported by the Hong Kong Research Grants Council (GRF HKU 768610M, HKU 776412M and HKU 777511M); the Hong Kong Research Grants Council Theme-Based Research Scheme (T12-705/11); the European Community Seventh Framework Programme Grant on European Network of National Schizophrenia Networks Studying Gene-Environment Interactions; the Hong Kong Health and Medical Research Fund (01121436, 01121616 and 02132236); the HKU Seed Funding Programme for Basic Research (201302159006, 201311159090 and 201411159172) and The University of Hong Kong Strategic Research Theme on Genomics.

Conflict of Interest: none declared.

References

Agha,Z. *et al.* (2014) Exome sequencing identifies three novel candidate genes implicated in intellectual disability. *PLoS One*, **9**.

Ansar,M. *et al.* (2015) Mutation of ATF6 causes autosomal recessive achromatopsia. *Hum. Genet.*, **134**, 941–950.

Balwani,M. *et al.* (2013) Loss-of-function ferrochelatase and gain-of-function erythroid-specific 5-aminolevulinate synthase mutations causing erythropoietic protoporphyria and X-linked protoporphyria in North American patients reveal novel mutations and a high prevalence of X-linked protoporphyria. *Mol. Med.*, **19**, 26–35.

Blake,J.A. *et al.* (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, **42**, D810–D817.

Blekhman,R. *et al.* (2008) Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.*, **18**, 883–889.

Choi,Y. *et al.* (2012) Predicting the functional effect of amino acid substitutions and indels (functional impacts of amino acid variants). *PLoS One*, **7**, e46688.

Conidi,E.M. *et al.* (2015) Homozygous carriers of APP A713T mutation in an autosomal dominant Alzheimer disease family. *Neurology*, **84**, 2266–2273.

Cooper,G.M. *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901.

Cukier,H.N. *et al.* (2014) Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Mol. Autism*, **5**, 1.

Figlioli,G. *et al.* (2012) Medullary thyroid carcinoma (MTC) and RET proto-oncogene: mutation spectrum in the familial cases and a meta-analysis of studies on the sporadic form. *Mutat. Res.*, **752**, 36–44.

Flicek,P. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.

Fozzatti,L. *et al.* (2011) Resistance to thyroid hormone is modulated in vivo by the nuclear receptor corepressor (NCOR1). *Proc. Natl. Acad. Sci. USA*, **108**, 17462–17467.

Garber,M. *et al.* (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–i62.

Gray,K.A. *et al.* (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.*, **41**, D545–D552.

Hammer,M.B. *et al.* (2013) Exome sequencing: an efficient diagnostic tool for complex neurodegenerative disorders. *Eur. J. Neurol.*, **20**, 486–492.

Hu,H. *et al.* (2015) X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Mol. Psychiatry.*, **21**, 133–148.

Huang,N. *et al.* (2010) Characterising and predicting haploinsufficiency in the human genome (predicting haploinsufficiency in the human genome). *PLoS Genet.*, **6**, e1001154.

Huang,D. *et al.* (2015) BRCC3 mutations in myeloid neoplasms. *Haematologica*, **100**, 1051–1057.

Itan,Y. *et al.* (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci. USA*, **112**, 13615.

Ivan,A.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248.

Jana Marie,S. *et al.* (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, **11**, 361.

Jin,W. *et al.* (2012) A systematic characterization of genes underlying both complex and Mendelian diseases. *Hum. Mol. Genet.*, **21**, 1611–1624.

Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*.

Khurana,E. *et al.* (2013) Interpretation of genomic variants using a unified biological network approach (impact of genomic variants in a unified network). **9**, e1002886.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

Lawrence,M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer genes. *Nature*, **499**, 214–218.

Li,M.X. *et al.* (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.*, **40**, e53.

Liu,X. *et al.* (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**.

Liu,X.M. *et al.* (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mut.*, **34**, E2393–E2402.

Lohmueller,K.E. *et al.* (2013) Whole-exome sequencing of 2,000 Danish Individuals and the role of rare coding variants in type 2 diabetes. *Am. J. Hum. Genet.*, **93**, 1072–1086.

Lupski,J.R. *et al.* (2010) Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.*, **362**, 1181–1191.

MacArthur,D.G. *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828.

Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812.

Nikolay,A.B. *et al.* (2011) Autosomal recessive deafness 1A (DFNB1A) in Yakut population isolate in Eastern Siberia: extensive accumulation of the splice site mutation IVS1 + 1G>A in GJB2 gene as a result of founder effect. *J. Hum. Genet.*, **56**, 631.

Niranjan,T.S. *et al.* (2015) Affected kindred analysis of human X chromosome exomes to identify novel X-linked intellectual disability genes. *PLoS ONE*, **10**.

Petrovski,S. *et al.* (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.*, **9**.

Philips,A. *et al.* (2014) X-exome sequencing in Finnish families with intellectual disability - four novel mutations and two novel syndromic phenotypes. *Orphanet J. Rare Dis.*, **9**.

Purcell,S.M. *et al.* (2014) A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, **506**, 185–190.

Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.

Rong,J. *et al.* (2015) A rare co-segregation-mutation in the insulin receptor substrate 1 gene in one Chinese family with ankylosing spondylitis. *PLoS One*, **10**.

Samocha,K.E. *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, **46**, 944–950.

Shihab,H.A. *et al.* (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum. Mut.*, **34**, 57–65.

Solomon,B.D. *et al.* (2013) Clinical genomic database. *Proc. Natl. Acad. Sci.*, **110**, 9851–9855.

Sung,C. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes.(Report). *Genome Res.*, **19**, 1553–1561.

The 1000 Genomes Project Consortium *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **419**, 56–65.

Yuval,I. *et al.* (2016) The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat. Methods*, **13**, 109.