doi:10.1093/nar/gkr1257

A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases

Miao-Xin Li[1,2,3,4,*], Hong-Sheng Gui[1], Johnny S. H. Kwan[1,5], Su-Ying Bao[6] and Pak C. Sham[1,2,3,4,*]

[1]Department of Psychiatry, [2]State Key Laboratory for Cognitive and Brain Sciences, [3]Centre for Reproduction, Development and Growth, [4]Genome Research Centre, [5]Department of Medicine and [6]Department of Biochemistry, University of Hong Kong, Pokfulam, Hong Kong, China

Received August 24, 2011; Revised December 1, 2011; Accepted December 5, 2011

ABSTRACT

Exome sequencing strategy is promising for finding novel mutations of human monogenic disorders. However, pinpointing the casual mutation in a small number of samples is still a big challenge. Here, we propose a three-level filtration and prioritization framework to identify the casual mutation(s) in exome sequencing studies. This efficient and comprehensive framework successfully narrowed down whole exome variants to very small numbers of candidate variants in the proof-of-concept examples. The proposed framework, implemented in a user-friendly software package, named KGGSeq (http://statgenpro.psychiatry.hku.hk/kggseq), will play a very useful role in exome sequencing-based discovery of human Mendelian disease genes.

INTRODUCTION

Identification of mutations underlying all human rare monogenic disorders is far from complete (1–3) and is of substantial interest in understanding disease mechanisms and development of drug targets (4,5). Recent advances in exome sequencing technologies make it possible to reveal the unknown disease mutations (6) and are leading to the discovery of many variants which affect protein function and cause Mendelian diseases (2), compared to traditional positional cloning strategies (7). However, finding the causal mutation(s) for a particular Mendelian disease among millions of variants is as difficult as looking for a needle in the haystack. In addition, it has been noted that most private sequence variants of a person or a pedigree, which are not small in size, are likely to be neutral and do not cause any severe disorders (8). So it is still costly, laborious and challenging to pinpoint the genuine disease mutations even though the price of exome sequencing is now dropping dramatically (9).

A number of software tools can be used to narrow down the list of candidate variants in exome sequencing studies. Some are statistical genetics tools which prioritize genomic regions based on evidence for shared ancestral polymorphisms and/or genetic linkage co-segregation, like BEAGLE, GERMLINE, PLINK IBD and MERLIN (10–12). Meanwhile, a few computational biology tools focus more on predicting degree of deleteriousness of a non-synonymous (NS) single nucleotide variant (SNV) in a protein-coding gene by various computational algorithms using genomic features like amino acid physicochemical properties, protein structure, cross-species conservation, etc (13,14). Recently, a database, named dbNSFP, has complied and standardized the deleteriousness scores derived by five widely used prediction tools [SIFT (15), Polyphen2 (16), LRT (17), MutationTaster (18) and PhyloP (19)] at the NS SNVs of consensus coding sequences (CCDS) regions of human genome to facilitate the process of evaluating functional importance of large amount of NS SNVs in exome sequencing studies (20). Other bioinformatics tools, such as SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation131/) and ANNOVAR (21), focus on comprehensive annotation of variants using information from diverse bioinformatics resources including gene features, genomic conservation, etc. However, these functionalities are scattered in different analytical tools, which means users have to do the time-consuming job of combining their results together. Sometimes, the results from different functional site prediction tools are inconsistent (17), making it difficult to obtain a single list of candidates for follow-up validation. Moreover, other valuable resources, including biological pathways and biomedical literatures, are still not incorporated into the existing tools.

Accordingly, we proposed a comprehensive three-level framework to combine a number of filtration and prioritization functions into one analysis procedure for exome sequencing-based discovery of human Mendelian disease

*To whom correspondence should be addressed. Tel: +852 2819 9559; Fax: +852 2819 9550; Email: mxli@hku.hk
Correspondence may also be addressed to Pak C. Sham. Tel: +852 2819 9557; Fax: +852 2819 9550; Email: pcsham@hkucc.hku.hk

genes. We then evaluate the performance of this framework by a number of synthesized proof-of-concept examples about known causal mutations of Mendelian disorders.

MATERIALS AND METHODS

Construction of a three-level filtration and prioritization framework

The proposed framework is comprised of a series of functions to filter and prioritize variants at three different levels: genetic level, variant-gene level and knowledge level, according to the resources used (illustrated in Figure 1). This framework has been implemented as one of functional modules in our software tool called KGGSeq (a biological Knowledge-based mining platform for Genomic and Genetic studies using Sequence data, http://statgenpro.psychiatry.hku.hk/kggseq). In KGGSeq, these functions can be carried out sequentially or skipped optionally according to various purposes.

Genetic level

Genetic information, if used appropriately, can help quickly narrow the candidate regions of interest for Mendelian diseases. The functions at genetic level consider two pieces of information: genomic region shared by multiple affected family members and mode of inheritance of disease. For Mendelian diseases, affected family members usually share the genomic segment harboring the causal mutation(s). Therefore, variants inside the identity-by-descent (IBD) regions found among the affected family members are of primary interest, regardless of the penetrance of the causal mutation(s). KGGSeq can read the IBD regions, estimated by a third-party software tool such as Beagle, PLINK and Merlin, and then highlight variants falling into these regions. It can also read the regions with significant evidence of genetic linkage (or co-segregation) reported by genetic linkage analysis tools like Merlin, SimWalk2, and Allegro in order to filter out regions unlikely covering the causal variants; these tools can utilize the linkage information also in unaffected family members and consider the penetrance of causal mutations through statistical models. The mode of inheritance of disease can also be used to effectively exclude impossible disease-causal variants. Specifically, for rare autosomal recessive disorder KGGSeq excludes sequence variants which have heterozygous genotypes in one or more affected family members; and if unaffected family

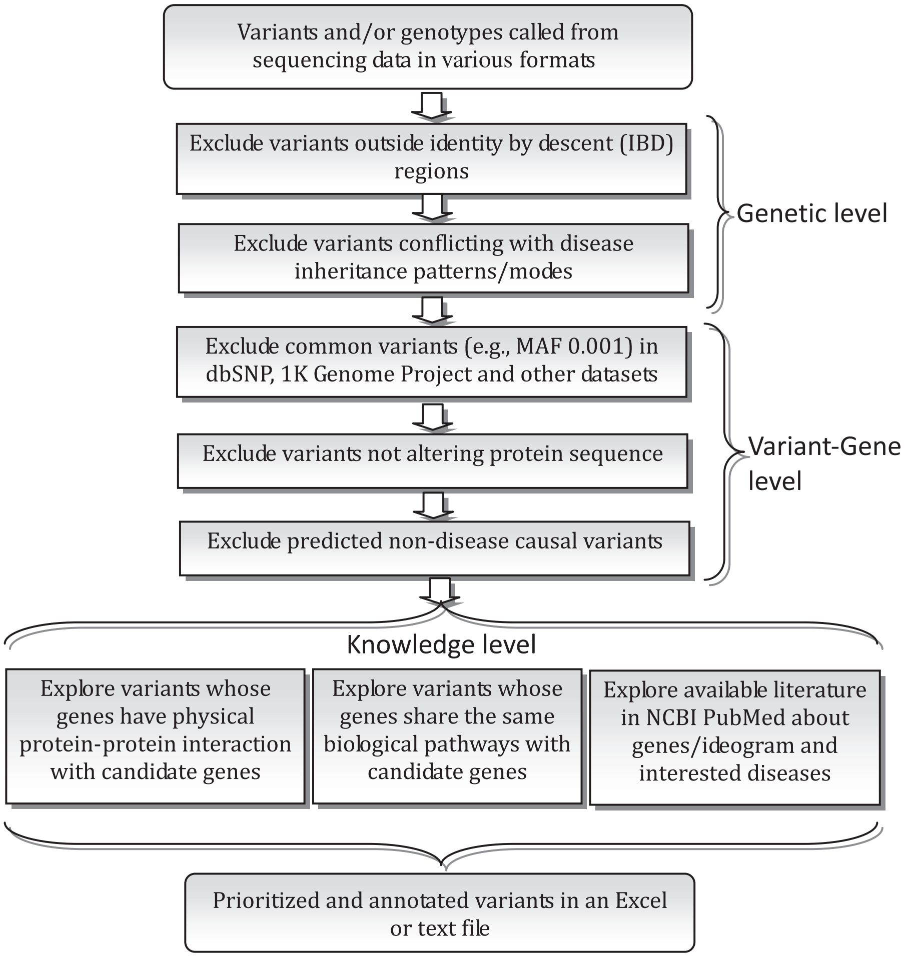
Variants and/or genotypes called from sequencing data in various formats
Exclude variants outside identity by descent (IBD) regions
Exclude variants conflicting with disease inheritance patterns/modes
Genetic level
Exclude common variants (e.g., MAF 0.001) in dbSNP, 1K Genome Project and other datasets
Exclude variants not altering protein sequence
Exclude predicted non-disease causal variants
Variant-Gene level
Knowledge level
Explore variants whose genes have physical protein-protein interaction with candidate genes
Explore variants whose genes share the same biological pathways with candidate genes
Explore available literature in NCBI PubMed about genes/ideogram and interested diseases
Prioritized and annotated variants in an Excel or text file

Figure 1. The three-level filtration and prioritization framework implemented in KGGSeq.

members are also recruited for investigating of a familial early-onset Mendelian disease, KGGSeq can be used to exclude variants which have the same homozygous genotypes in both affected and unaffected family members. For rare autosomal dominant disorders, when it is very unlikely that affected family members without consanguineous mating carry homozygous mutation genotypes, KGGSeq can be used to exclude sequence variants which are homozygous in one or more affected family members; and if unaffected family members are also recruited for investigating of a familial early-onset Mendelian disease, KGGSeq can be used to exclude the bi-allelic variants which are heterozygous genotypes in one or more of the unaffected ones. Note, however, that the inheritance mode-based filtration is proposed under strong assumptions for rare Mendelian diseases with clear inheritance mode. If the inheritance mode is elusive, such filtration is not suggested; otherwise it may lead to the missing of the genuine mutation(s).

Variant-gene level

For rare severe diseases, underlying causal mutations are very unlikely to be common in human population. KGGSeq can filter out common variants deposited in public databases (including the 1000 Genomes Project and NCBI dbSNP) as well as existing in the in-house data sets according to an adjustable allele frequency threshold (1% by default in KGGSeq). In addition, one can use gene features of variants for prioritization. As severe Mendelian disorders are more likely to be caused by NS or splicing or insertion/deletion mutations which change the amino acids in a protein (2), focusing only on these NS or splicing variants often substantially narrows down the number of candidate variants. KGGSeq can map the variants onto Refseq genes according to the coordinates and allow users to exclude variants by their gene features (such as intron and synonymous variants). Moreover, because not all NS SNV contribute equally to affecting functions of coded proteins, KGGSeq incorporated the five deleteriousness scores from various bioinformatics algorithms (20), by logistic regression model to more accurately predict whether a NS SNV is potentially disease-causal or not (see more in the 'Logistic regression prediction model for NS SNVs' below).

Knowledge level

Since the protein products of genes responsible for the same or phenotypically similar disorders tend to physically interact with each other so as to carry out certain biological functions (22), KGGseq incorporates the physical protein–protein interactions (PPIs) from STRING database version 9.0 (23) (http://string-db.org/) and highlights variants located in a gene whose protein product is known to have PPI(s) with the protein products of some user-specified seed genes. These seed genes are often known to cause the exact disease in question or phenotypically similar diseases. Analogously, causative genes of the same (or phenotypically similar) diseases are inclined to distribute within the same biological modules like pathways (24,25). KGGSeq currently incorporates 880 canonical pathways curated by GSEA (26) and is able to highlight variants of a gene sharing the same biological pathway(s) with some user-specified seed genes. Besides, KGGSeq can automatically look up the relevant literature information in NCBI PubMed database (http://www.ncbi.nlm.nih.gov/pubmed) using gene symbol, ideogram location and the disease name(s) as keywords. This feature can be very effective for finding the causal variant (either novel or not) of a disease within known casual genes or published genetic linkage regions.

Logistic regression prediction model for NS SNVs

The logistic regression model was constructed to combine the five deleteriousness scores [SIFT (15), Polyphen2 (16), LRT (17), MutationTaster (18) and PhyloP (19)] in order to give a more accurate prediction of the role of a NS SNV in Mendelian disease. We selected 7296 unique NS SNVs underlying certain human monogenic disorders as cases and 9829 unique NS SNVs with minor allele frequencies (MAF) <0.01 as controls (see more in the 'Data sets' section below) to train and test the prediction model. The 10-fold cross-validation approach was used to assess the performance of the prediction model. The receiver operating characteristic (ROC) curves were used to compare the performance of the proposed model with the individual deleteriousness scores. We used a discrimination cutoff which led to the maximal summation of true positive rate (sensitivity) and true negative rate (specificity) to classify a variant as disease-causal or neutral by the trained logistic regression model.

Data sets

Disease-causal and neutral variants. 9133 unique variants associated with some human diseases in the OMIM database were downloaded and extracted from Galaxy (http://main.g2.bx.psu.edu/library). 59557 unique NS SNVs in the 1000 Genomes Project dataset (released in March 2010 and provided by ANNOVAR, http://www.openbioinformatics.org/annovar/) were also downloaded. The variants from the OMIM database were regarded as disease-causal, after exclusion of variants in the 1000 Genomes Project and/or those associated with complex diseases. The variants from the 1000 Genomes Project were regarded as being neutral. Five types of standardized deleteriousness scores (ranging from 0 to 1) downloaded from the dbNSFP database (20) were used as explanatory variables for each NS SNV in the multiple logistic regression model. Variants with any missing deleteriousness scores were ignored. The numbers of disease-causal, neutral variants with $\mathrm{MAF} < 0.01$ and neutral variants with $\mathrm{MAF} \geq 0.01$ examined are 7296, 9829 and 38 260, respectively.

Synthesized exomes with disease causal variants. We downloaded exome sequence variants of six HapMap subjects [NA12156 and NA12878 (Caucasian) NA18507 and NA19240 (African), NA18956 (Japanese) and NA18555 (Chinese)] from the public domain provided by Ng's group (27). In order to test the effectiveness of KGGseq in prioritizing disease causal variants/genes, we inserted several known causal mutations of monogenic disorders into these exomes so as to make eight synthesized exomes (named S_exome1till S_exome8). The disease causal variants included a missense mutation (in heterozygous form) on MYH3 for Freeman–Sheldon syndrome (FS) (27), a truncating mutation (in homozygous form) on SERPINF1 for Osteogenesis imperfect (OI) (28) and a 1-bp frameshift insertion (in heterozygous form) for Miller's syndrome (2). Each of the six case exomes (S_exome1 ∼ S_exome6) contained the missense mutation for FS; S_exome7 was made up of the OI causal truncating mutation (in homozygous form) and SNV variants from NA18555. One 1-bp frameshift insertion on DHODH for Miller syndrome was merged with 145 indels from NA18555 to form S_exome8.

RESULTS

Filtration and prioritization in the synthesized exomes

KGGSeq was used to prioritize causal variants/genes in each synthesized exome through the three-level filtration and prioritization framework (the IBD filtering function are ignored because these HapMap subjects are unrelated). Table 1 shows the counts of variants after a step-by-step filtration.

For the FS syndrome (S_exome1-S_exome6), the first two-level filtrations produced a small set of ∼100–150 candidate variants. The FS syndrome is a subtype of Distal arthrogryposis type 2A (DA2); and Distal arthrogryposis type 1A (DA1) is clinically similar to (but less severe than) DA2. Therefore, in the knowledge level prioritization, we used four known causal genes (TNNI2, TNNT3, TPM2 and MYBPC1) for DA1 and DA2 as seed candidate genes for PPI and biological pathway exploration and three terms (Freeman–Sheldon syndrome, Distal arthrogryposis type 2A and Distal arthrogryposis type 1A) for literature mining. The third level prioritization successfully narrowed down the candidate variants to a very small subset variants and even pinpointed the exact mutation, a missense mutation p.672R > H at 17th exon of MYH3, in the S_exome1 and S_exome6. We found the underling causative gene MYH3 had physically PPIs with all of the four seed candidate genes (Figure 2) and also shared two pathways (REACTOME_MUSCLE_CONTRACTION and REACTOME_STRIATED_MUSCLE_CONTRACTION, http://www.reactome.org/entitylevelview/PathwayBrowser.html#DB = gk_current&FOCUS_SPECIES_ID = 48887&FOCUS_PATHWAY_ID = 397014&ID = 397014) with the four genes.

For the recessive disorder (S_ exome7), OI, the filtration functions in the early steps (till common variants exclusion step) effectively reduced the candidate variants from 16 048 to 51. The logistic risk score led to a further removal of 49 predicted non-disease causal variants. Among the two remaining variants, the causal mutation (p. $232Y > X$ of SERPINF1) has a higher predicted score which is related to the probability of being involved in Mendelian disease given the deleteriousness scores compared to the other mutation. Hence, the knowledge-level filtration is ignored and Table 1 has no results at this level. For the Miller syndrome (S_ exome8), the common variants filtration function only removed around 30 indels and the logistic prediction model based on the deleteriousness scores was not applicable to them. To avoid circular reasoning, we did not use the known causal gene DHODH as seed candidate genes but employed four anonymous disease names (Miller syndrome, Postaxial acrofacial dysostosis, Genee–Wiedemann syndrome, Wildervanck–Smith syndrome) to explore the NCBI PubMed for a prioritization. Eventually, eight indels were highlighted. The cytoband regions of seven different indels co-occurred in the abstracts of five published papers with the disease names as PubMed keywords. The causative gene DHODH was mentioned by three papers about Miller syndrome in the PubMed database.

Logistic regression model-based prediction

Figure 4a shows the ROC curves of various prediction methods to differentiate Mendelian disease-causal

Table 1. Counts of SNPs (and genes) after filtrations by functions of the three-level framework in KGGSeq

Steps	S_exome1	S_exome2	S_exome3	S_exome4	S_exome5	S_exome6	S_exome7	S_exome8
	SNV	SNV	SNV	SNV	SNV	SNV	SNV	Indel
Initial	16 120	15 971	19 721	19 518	16 012	16 048	16 048	146
Inheritance pattern[a]	10 180 (Dom)	9929 (Dom)	12 897 (Dom)	12 867 (Dom)	9133 (Dom)	9182 (Dom)	6867 (Rec)	146 (Dom)
Non-synonymous[b]	4705	4582	5837	5833	4171	4163	3089	143
Rare in dbSNP + 1000 Genome[c]	457	508	709	794	410	508	51	116 (113)
Predicted to be disease causal	106 (90)	127 (117)	149 (133)	164 (152)	95 (87)	120 (107)	2 (2)	–
Knowledge-related[d]	1 (1)	7 (7)	4 (4)	6 (6)	6 (5)	1 (1)	–[e]	8 (8)
PPI	1 (1)	2 (2)	1 (1)	1 (1)	1 (1)	1 (1)	–	–
Pathway	1 (1)	5 (5)	2 (2)	4 (4)	2 (2)	1 (1)	–	–
PubMed	1 (1)	3 (3)	3 (3)	4 (4)	4 (3)	1 (1)	–	8 (8)

[a] Dominant mode only considered with variants with heterozygous genotypes and recessive mode only considered with variants with homozygous genotypes; [b] Non-synonymous includes missence, stopgain, stoploss and splicing SNVs and insertions/deletions causing frameshift, non-frameshift, stoploss, stopgain and splicing differences; [c] The rare variants referred to variants with MAF < 0.01 in dbSNP and 1000 Genome; [d] Knowledge-related variants/genes refer to those variants' genes having PPI(s) or sharing pathway(s) with provided candidate gene(s), and those variants fell into region(s) or gene(s) which co-occurred in the titles or abstracts of papers in PubMed database; [e] '—' means the corresponding analyses were not conducted for reasons stated in the 2nd paragraph of the 'Results' section.

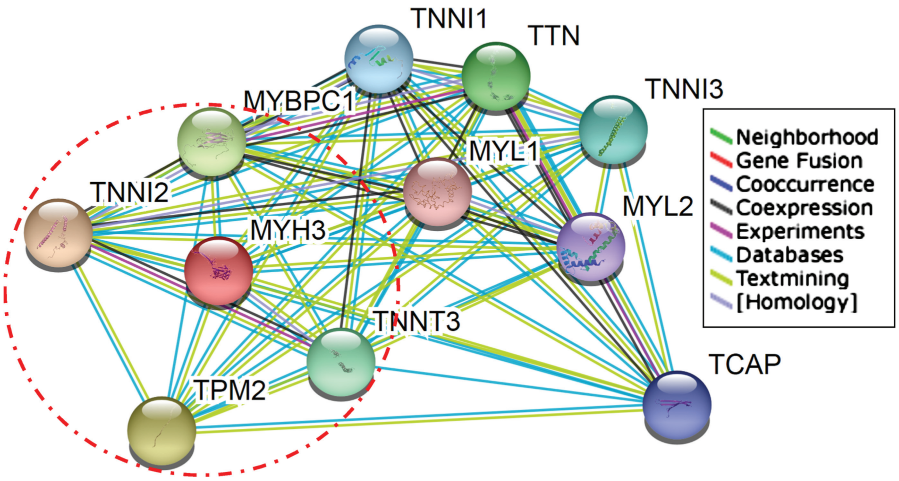

Figure 2. Protein–protein interaction network of MYH3 with four candidate genes. The five involved genes are in dashed circle. Each filled node denotes a gene; edges between notes indicate PPIs between protein products of the corresponding genes. Different edge colors represent the types of evidence for the association. This figure was produced by STRING (V9.0).

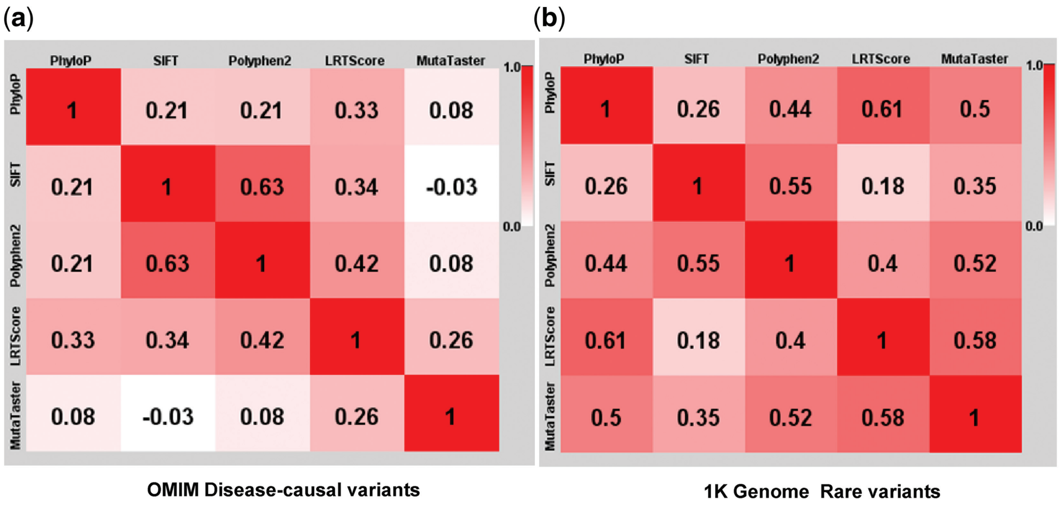

Figure 3. Pair-wise correlation of the five deleteriousness scores in the (a) disease-causal and (b) neutral rare variant sets. The Spearman's rank correlation method was used to calculate the pair-wise correlation coefficients.

variants from neutral variants with MAF < 0.01. Among the five algorithms of the deleteriousness scores, the MutationTaster outperforms the other four. However, the combined prediction by logistic regression model can still improve the overall performance a little bit and is more accurate when the true positive rate (or sensitivity) is over 70%. We also found that the individual deleteriousness scores were only in weak or moderate correlation (Spearman's rank correlation, Figure 3) and four of them are statistically significant in the multiple logistic regression model (Table 2) despite the fact that these tools use some common resources (such as cross-species conservation) to derive these deleteriousness scores. The combination of multiple scores may take advantage of the possible complementarities between different tools to allow more accurate prediction. Figure 4b shows the ROC curves of various prediction methods to differentiate Mendelian disease-causal variants from neutral variants with MAF ≥ 0.01. As expected, it is easier for all prediction tools to classify the relatively common variants and disease causal variants. However,

Table 2. Summary results of multiple logistic regression of five deleteriousness scores

Deleteriousness scores	Beta (±SD)	Z statistic	Pr(>|z|)
PhyloP	0.18 (±0.08)	2.13	0.033
SIFT	1.9 (±0.12)	15.33	<2e − 16
Polyphen2	1.00 (±0.06)	16.73	<2e − 16
LRTScore	0.10 (±0.12)	0.85	0.39
MutationTaster	2.34 (±0.06)	39.62	<2e − 16

The disease causal variants and neutral rare variants (MAF < 0.01) were used for model fitting in the logistic regression model.

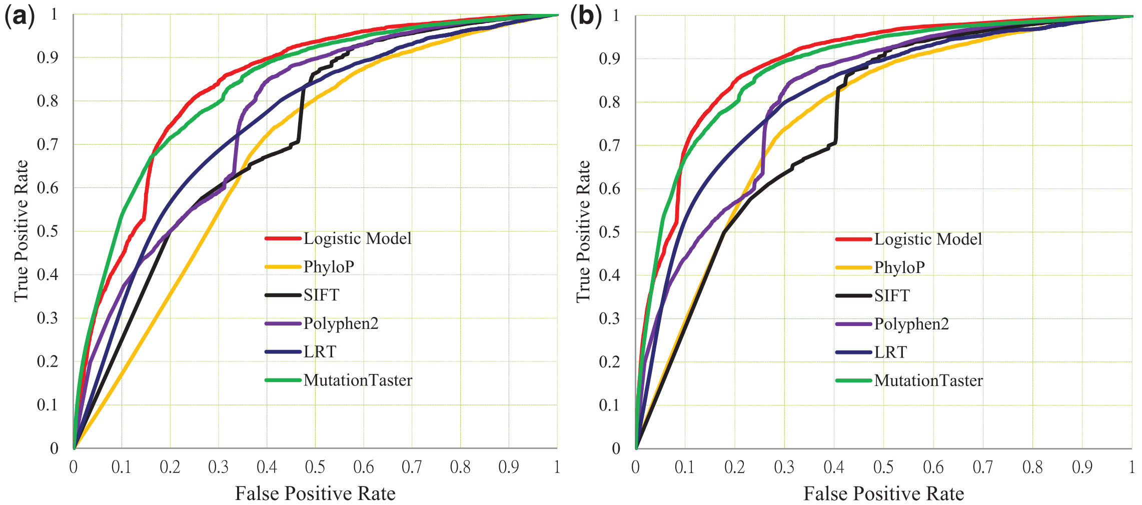

Figure 4. Receiver operating characteristic (ROC) curves of various methods. (**a**) The control (neutral) variants are rare (MAF < 0.01); (**b**) the control (neutral) variants have MAF ≥ 0.01. The true positive rate (sensitivity) and false positive rate (1-specificity) of logistic model were obtained by 10-fold cross validation procedure. Logistic model: performance of conventional multiple logistic regression model when combining the five deleterious scores (PhyloP, SIFT, PolyPhen2, LRT and MutationTaster). The true positive rate and false positive rate of the five different prediction methods were generated by varying the threshold scores for prediction in the entire data set.

since common variants (MAF ≥ 0.01) can be straightforwardly excluded using common variants in human populations, KGGSeq adopted the logistic regression model trained and tested on the data set made up of neutral variants with MAF < 0.01 and OMIM disease mutations to prioritize NS SNVs for Mendelian diseases. Given the deleteriousness scores, a probability-like value can be calculated by the logistic regression formula. The cutoff of the probability-like value which results in the maximal summation of average true positive rate and true negative rate was 0.5 in the 10-fold cross validation procedure. The corresponding average true positive rate and true negative rate are 81.4 and 74.2%, respectively.

KGGSeq platform

KGGSeq provides a user-friendly command line interface for users to utilize functions in the three-level filtration and prioritization framework to process large amount of exome sequencing data easily. It can recognize the variants data inputted in various formats, including the Variant Call Format (VCF, http://vcftools.sourceforge.net/specs.html). It outputs a list of prioritized and annotated variants in a flat text file or an excel file (see more at website of KGGSeq http://statgenpro.psychiatry.hku.hk/kggseq). Resource data of KGGSeq can be automatically updated from the website of KGGSeq or from their original sources.

In a testing of the synthesized exome (S_exome5), it took 5 min to reduce the number of variants from 16,012 to 95 on a Linux machine with Intel XEON 2 CPU 2.93 GHz. Memory usage was <1 GB RAM. However, it spent additional 15 min in remotely accessing the NCBI PubMed database to explore relevant literatures for the 95 variants. This step was slowed down deliberately because too frequent connection to PubMed database would be blocked by NCBI.

DISCUSSION

The proposed three-level framework has great potential to pinpoint causal mutations of monogenic diseases in massive amount of exome sequencing data. To our knowledge, KGGSeq is the first tool which efficiently combines multiple diverse resources into a single analysis framework for exome-sequencing-based discovery of human Mendelian disease gene. We have conceptually demonstrated its efficiency and power for prioritization in a number of synthesized data sets of three monogenic diseases (FS syndrome, OI and Miller syndrome), in which it dramatically reduced thousands of variants to a very small candidate variant list for follow-up replication.

We used a logistic regression model to combine multiple deleteriousness scores to predict whether a rare variant (MAF < 0.01) is disease-causal or not. In our testing examples, the prediction model correctly excluded vast majority of benign NS SNVs and even directly pinpointed the causal mutations of the autosomal recessive disease, OI. This suggests that the prediction function may be very effective for dealing with autosomal recessive diseases. In the study, we also found that the conventional logistic regression model could be more accurate than Condel WAS (29) which is a method recently proposed to combine multiple deleteriousness scores, in many scenarios (M.X. Li *et al.*, unpublished data). Condel WAS relied on using prior sensitivity and specificity as weights to adjust each deleteriousness score individually. A possible reason for our observation is that the prior sensitivity and specificity used in Condel WAS are only optimized locally for the individual deleteriousness scores but not globally when all five scores were considered. In addition, the logistical model is widely used and has solid theoretical foundation; it lends itself to flexibly combine more deleteriousness scores or genomic features as we have done for the five deleteriousness scores.

Currently, we combined deleteriousness scores by five different prediction algorithms for a more comprehensive prioritization of NS SNVs. We found these scores were only in weak or moderate correlation although some algorithms used common source data to produce the scores. In the performance evaluation in our training and testing dataset, MutationTaster outperformed the other four prediction algorithms individually and even performed better than a combined prediction of the four by the logistic regression according to the ROC curve (M.X. Li *et al.*, unpublished data). Probably, MutationTaster considered more valuable information for the prediction and/or its naive Bayes classifier trained under different amino acid change models (18) has better performance than the other four computational algorithms. Anyhow, a combined prediction by all of the five deleteriousness scores had better performance than individual scores as well as combined prediction by part of the deleteriousness scores; it has smaller false negative rate when the false positive rate is over 16% for rare NS SNVs (Figure 4a). In the filtration and prioritization procedure of exome sequence variants, it is acceptable to allow a reasonably larger false positive rate at this step and reduce the chance of missing true causative NS SNVs because one often has additional criteria to exclude the false positive variants.

Our use of the knowledge data to filter and prioritize exome sequence variants is unique to other existing tools which can be used to prioritize exome sequence variants. When there is sufficient knowledge about the disease or the underlying causal genes, this level analysis will be very powerful for genetic mapping. In the testing experiments the PPI and pathway information straightforwardly linked the four provided candidate genes to the underlying causal gene MYH3 of FS syndrome. The relationship between these genes may also contribute to our understanding of pathogenic mechanism of the disorder. Its PubMed literature searching-based prioritization function will be very effective for diseases studied by previous independent genetic linkage studies or even sequencing studies. In an real example of our exome sequencing project, KGGSeq successfully pinpointed a novel NS SNV mutation at a gene recently reported responsible for the same monogenic disorder named Spinocerebellar ataxia (M. X. Li *et al.*, submitted for publication).

The synthesized exomes may not completely represent exomes of real patients with monogenic disorders. So the above analysis may not sufficiently illustrate that causal mutation(s) for a rare Mendelian disease can be easily detected by KGGSeq to process the sequencing data of only one subject. Anyhow, these results suggest the three-level filtration and prioritization procedure can help dramatically reduce the number of candidate variants to a very small subset that is human-manageable. In reality, more stringent MAF thresholds (say, 0.005 or even 0.0) can be applied to autosomal dominant Mendelian disorders and more in-house data sets can be used to exclude additional common variants or rare benign sequence variants. The kinship information, if available, can also be used to remove variants in regions that are not shared by affected family members and those that are shared by discordant family members through KGGSeq. All these additional analysis can further reduce the number of candidate variants. Once the subset of highlighted variants is available, conventional Sanger sequencing can be feasibly employed to validate the variants in other subjects.

The knowledge level for filtration and prioritization may be not straightforward for diseases seldom studied. However well-studied diseases (and their causal genes) with similar syndrome or clinical phenotypes to the disease in question can be used as a ‘bait’ to fish the underlying disease genes because causative genes for the same (or phenotypically resembling) diseases tend to distribute within the same biological modules (24,30). In the testing experiment, we provided the known causal genes of DA2 for DA1 and observed the underlying causal gene had PPI and shared the same pathways with the known causal genes of DA2. Anyway, as our knowledge about human diseases and their pathogenesis are growing exponentially, this obstacle is gradually diminishing.

We will keep on refining this framework in the future. More resources [such as more valuable deleteriousness scores of NS SNVs, pseudogene and dispensable genes (21)] after careful evaluation will be incorporated into this framework. Other improvement may include advanced algorithms and statistical models to analytically prioritize the variants. Moreover, we will also look into the effectiveness of this framework (or an improved version) for the prioritization of rare variants responsible for complex diseases/traits.

ACKNOWLEDGEMENTS

We thanks some the institutes and authors of the resource data used by KGGSeq. This includes the annotation database of UCSC, ANNOVA, 1000 Genomes Project, dbNSFP, etc.

FUNDING

Funding for open access charge: Hong Kong Research Grants Council (GRF HKU 774707, HKU 768610M); European Community Seventh Framework Programme Grant on European Network of National Schizophrenia Networks Studying Gene-Environment Interactions; Small Project Funding (HKU 201007176166) and University of Hong Kong Strategic Research Theme on Genomics.

Conflict of interest statement. None declared.

REFERENCES

1. Antonarakis,S.E. and Beckmann,J.S. (2006) Mendelian disorders deserve more attention. *Nat. Rev. Genet.*, **7**, 277–282.
2. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.
3. Chen,J.M., Ferec,C. and Cooper,D.N. (2010) Revealing the human mutome. *Clin. Genet.*, **78**, 310–320.
4. McCarthy,M.I. and Hattersley,A.T. (2008) Learning from molecular genetics: novel insights arising from the definition of

genes for monogenic and type 2 diabetes. *Diabetes*, **57**, 2889–2898.
5. McCarthy,M.I. (2009) Exploring the unknown: assumptions about allelic architecture and strategies for susceptibility variant discovery. *Genome Med.*, **1**, 66.
6. Choi,M., Scholl,U.I., Ji,W., Liu,T., Tikhonova,I.R., Zumbo,P., Nayir,A., Bakkaloglu,A., Ozen,S., Sanjad,S. *et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 19096–19101.
7. Boehnke,M. (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *Am. J. Hum. Genet.*, **55**, 379–390.
8. Majewski,J., Schwartzentruber,J., Lalonde,E., Montpetit,A. and Jabado,N. (2011) What can exome sequencing do for you? *J. Med. Genet*, **48**, 580–589.
9. Maxmen,A. (2011) Exome sequencing deciphers rare diseases. *Cell*, **144**, 635–637.
10. Browning,B.L. and Browning,S.R. (2011) A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.*, **88**, 173–182.
11. Rodelsperger,C., Krawitz,P., Bauer,S., Hecht,J., Bigham,A.W., Bamshad,M., de Condor,B.J., Schweiger,M.R. and Robinson,P.N. (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. *Bioinformatics*, **27**, 829–836.
12. Abecasis,G.R., Cherny,S.S., Cookson,W.O. and Cardon,L.R. (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97–101.
13. Ng,P.C. and Henikoff,S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, **7**, 61–80.
14. Teng,S., Michonova-Alexova,E. and Alexov,E. (2008) Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Curr. Pharm. Biotechnol.*, **9**, 123–133.
15. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
16. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
17. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
18. Schwarz,J.M., Rodelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
19. Siepel,A., Pollard,K. and Haussler,D. (2006) New methods for detecting lineage-specific selection. *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*. Venice, Italy, pp. 190–205.
20. Liu,X., Jian,X. and Boerwinkle,E. (2011) dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. *Hum. Mutat*, **32**, 894–899.
21. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
22. Lim,J., Hao,T., Shaw,C., Patel,A.J., Szabo,G., Rual,J.F., Fisk,C.J., Li,N., Smolyar,A., Hill,D.E. *et al.* (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, **125**, 801–814.
23. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A. and Simonovic,M. (2009) STRING 8¡Xa global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412.
24. Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.
25. Wood,L.D., Parsons,D.W., Jones,S., Lin,J., Sjoblom,T., Leary,R.J., Shen,D., Boca,S.M., Barber,T., Ptak,J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
26. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
27. Ng,S.B., Turner,E.H., Robertson,P.D., Flygare,S.D., Bigham,A.W., Lee,C., Shaffer,T., Wong,M., Bhattacharjee,A., Eichler,E.E. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
28. Becker,J., Semler,O., Gilissen,C., Li,Y., Bolz,H.J., Giunta,C., Bergmann,C., Rohrbach,M., Koerber,F., Zimmermann,K. *et al.* (2011) Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. *Am. J. Hum. Genet.*, **88**, 362–371.
29. Gonzalez-Perez,A. and Lopez-Bigas,N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.
30. Wu,X., Jiang,R., Zhang,M.Q. and Li,S. (2008) Network-based global inference of human disease genes. *Mol, Syst. Biol.*, **4**, 189.