INFORMATICS

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

wKGGSeq: A Comprehensive Strategy-Based and Disease-Targeted Online Framework to Facilitate Exome Sequencing Studies of Inherited Disorders

Mulin Jun Li,[1,2,3] Jiaen Deng,[1,4] Panwen Wang,[1,3] Wanling Yang,[1,5] Shu Leong Ho,[6] Pak Chung Sham,[1,4,7] Junwen Wang,[1,3]* and Miaoxin Li[1,4,8]†

[1]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China; [2]Departments of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China; [3]Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, Guangdong 518057, China; [4]Psychiatry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China; [5]Paediatrics and Adolescent Medicine, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China; [6]Medicine, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China; [7]State Key Laboratory in Cognitive and Brain Sciences, The University of Hong Kong, Hong Kong SAR, China; [8]Centre for Reproduction, Development and Growth, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

Communicated by Christopher Mathew

Received 22 June 2014; accepted revised manuscript 3 February 2015.
Published online 10 February 2015 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22766

ABSTRACT: With the rapid advances in high-throughput sequencing technologies, exome sequencing and targeted region sequencing have become routine approaches for identifying mutations of inherited disorders in both genetics research and molecular diagnosis. There is an imminent need for comprehensive and easy-to-use downstream analysis tools to isolate causal mutations in exome sequencing studies. We have developed a user-friendly online framework, wKGGSeq, to provide systematic annotation, filtration, prioritization, and visualization functions for characterizing causal mutation(s) in exome sequencing studies of inherited disorders. wKGGSeq provides: (1) a novel strategy-based procedure for downstream analysis of a large amount of exome sequencing data and (2) a disease-targeted analysis procedure to facilitate clinical diagnosis of well-studied genetic diseases. In addition, it is also equipped with abundant online annotation functions for sequence variants. We demonstrate that wKGGSeq either outperforms or is comparable to two popular tools in several real exome sequencing samples. This tool will greatly facilitate the downstream analysis of exome sequencing data and can play a useful role for researchers and clinicians in identifying causal mutations of inherited disorders. The wKGGSeq is freely available at http://statgenpro.psychiatry.hku.hk/wkggseq or http://jjwanglab.org/wkggseq, and will be updated frequently.

Hum Mutat 36:496–503, 2015.

KEY WORDS: exome sequencing; Mendelian disease; quality control; variant prioritization; variant annotation

Introduction

With advances in sequencing technologies, exome sequencing has become an important strategy to identify causal mutations of Mendelian diseases [Bamshad et al., 2011; Do et al., 2012] for research and even clinical diagnostics [Rehm, 2013]. Emerging collections of disease-causal genes for clinical usage of disease-targeted genetic testing, such as the NCBI genetic testing registry (GTR) [Rubinstein et al., 2013] are also encouraging more applications of the exome sequencing in clinical practices. While procedures for calling sequence variants from large sequencing raw data are becoming standard, how the causal mutations can be effectively and accurately isolated from the called variants remains a challenge. These procedures require case-specific and even sophisticated combination of multiple analysis functions. Several command-line tools have been developed to flexibly apply different criteria to eliminate irrelevant sequence variants in genetic studies, such as ANNOVAR [Wang et al., 2010b], VAAST [Rope et al., 2011], KGGSeq [Li et al., 2012b], and PriVar [Zhang et al., 2013]. Although those command-line tools provide comprehensive resources and functions for combined filtration and prioritization of causal mutations, they appear complicated for many geneticists and clinicians. While there are several online or graphic interface tools available to analyze exome sequencing data (e.g., VEP [McLaren et al., 2010], GeneTalk [Kamphans and Krawitz, 2012], wANNOVAR [Chang and Wang, 2012], TREAT [Asmann et al., 2012], VAT [Habegger et al., 2012], exomeSuite [Maranhao et al., 2014], and BiERapp [Aleman et al., 2014]), most were originally designed for general filtering and annotation and have weak functions to narrow down exome sequencing variants for a disease in question, especially in filtration based on various

Additional Supporting Information may be found in the online version of this article.

†Correspondence to: Miaoxin Li. E-mail: mxli@hku.hk

*Correspondence to: Junwen Wang. E-mail: junwen@hku.hk

Contract grant sponsors: Hong Kong Research Grants Council GRF (HKU 781511M, HKU 768610M, HKU 776412M, HKU 17121414M, and HKU 777511M); Hong Kong Research Grants Council Theme-Based Research Scheme T12-705/11; NSFC (91229105) of China; European Community Seventh Framework Programme Grant on European Network of National Schizophrenia Networks Studying Gene-Environment Interactions (EU-GEI); the HKU Seed Funding Programme for Basic Research 201302159006 and 201311159090; The University of Hong Kong Strategic Research Theme on Genomics.

inheritance patterns. Moreover, few tools have been developed to integrate disease-gene panels to facilitate clinical diagnosis by exome sequencing strategy for well-studied genetic disorders.

Here, we introduce a user-friendly analysis online tool, wKGGSeq, to ease comprehensive filtration and prioritization of exome sequencing variants for inherited diseases. Specifically, wKGGSeq offers a strategy-based pipeline to facilitate detection of causal variants for diseases with multiple genetic inheritance patterns and different study designs. Besides, integrating gene panels of GTR, wKGGSeq also provides a disease-targeted pipeline for quickly isolating causal mutations in known disease causal genes for the purpose of aiding exome sequencing-based molecular diagnosis. Furthermore, the system includes many interactive functions for visualization, prioritization, and annotation of disease-causal variants.

Methods

Web Server Framework

The Web server is composed of three main functional Web interfaces (strategy-based, disease-targeted, and parameters-flexible), a job management system, and a backend task engine. The analysis workflow is illustrated in Figure 1. The analysis strategy and/or known candidate gene(s) should be specified first on the Web interface. User then uploads the sequencing data and submits a filtration and prioritization job to the backend tool KGGSeq [Li et al., 2012b] on the server. The analysis results of KGGSeq are visualized and further annotated by the Web server. The wKGGSeq Web server is implemented in Perl and based on the Catalyst Web framework under model–view–controller design pattern, which facilitates the expansibility and efficiency.

Strategy-Based Analysis

We proposed a strategy-based pipeline to reorganize the filtration and prioritization functions of KGGSeq [Li et al., 2012b] for exome sequencing studies with genetic inheritance models under various study designs [Robinson et al., 2011; Gilissen et al., 2012]. It included six different analysis strategies: linkage, runs of homozygosity, double-hit gene, overlapping, de novo mutation, and candidate gene strategies. Each strategy includes both unique and common filtration and prioritization functions to systematically process the sequence variants.

The following are the unique function of each strategy-based analysis:

(1) Linkage strategy: this analysis strategy uses conventional genetic linkage information (including identical by descent [IBD] or linkage regions and inheritance patterns) in either large or small pedigree to prioritize sequence variants. Detailed description can be found in our previous KGGSeq paper [Li et al., 2012b]. Briefly, a region harboring the causal mutation should be shared across the affected family members but not the unaffected family members. Although this strategy is applicable to most types of family structure (even including trios), large pedigrees are more effective for the filtration.

(2) Runs of homozygosity strategy: this strategy specially targets the rare and recessively inherited disorder in a patient whose parents are suspected to be consanguineous and thus stretches of his or her two homologous chromosomes are IBD. wKGGSeq has a simple function to explore the longest surrounding region of consecutive homozygous genotypes among patients for each variant. To avoid technique noises, variants and genotypes with low quality (see more in the below quality control section) will be ignored. Each variant is annotated with the longest runs of homozygosity region (in base pair) and number of supporting variants with homozygous genotypes. A cutoff is used to filter out variants with short runs of homozygosity regions. The cutoff relies on how closely related his or her consanguineous parents are. The closer the relationship is, the larger cutoff should be used due to limited number of recombination in meioses. Given the number of generations since the common ancestor g, the expected length of a runs of homozygosity region follows an exponential distribution with mean equal to $1/(2^*g)$ Morgans [Howrigan et al., 2011]. By default, the wKGGSeq uses 1 cM (or 1 Mb), which corresponds to 50 generations.

(3) Double-hit gene strategy: this strategy is designed for a disorder caused by a gene with compound-heterozygous mutations, in which both copies of the gene on the two homologous chromosomes are damaged by two different mutations. wKGGSeq uses two different types of input data, phased genotypes of a patient, or the unphased genotypes in a trio (including parents and an offspring) to explore compound-heterozygous mutations. Note that wKGGSeq does not infer the phase of genotypes by itself. Instead, phased genotypes produced by population genetics tools (e.g., [Browning and Browning, 2013; Delaneau et al., 2013]) can be inputted into wKGGSeq for detecting variants that damage both copies of a gene. If parents' genotypes are available, wKGGSeq can directly search the genes with compound-heterozygous mutations in an offspring by using unphased genotypes. But the transmitted alleles can be extrapolated when only one of the parents has heterozygous genotypes at different loci of a gene. For example, if (1) the parental and maternal genotypes are heterozygous and homozygous at one variant, (2) their genotypes are homozygous and heterozygous at another variant of the same gene, and (3) the child's genotypes at both variants are heterozygous, it is clear that the gene has compound heterozygous mutations in the child. Moreover, only variants predicted to be pathogenic by KGGSeq [Li et al., 2013b] are taken into account. In principle, the compound-heterozygous mutation pattern is very similar to the recessive pattern in which the two copies of a gene are damaged by the same variant. Therefore, this double-hit gene strategy is also applicable to recessive diseases. All double-hit genes and involved variants will be extracted for further prioritization according to other criteria.

(4) Overlapping strategy: when studying the rare monogenic disorders without genetic heterogeneity, the overlapping strategy on wKGGSeq can be used to efficiently extract common damaging mutations across unrelated patients for the same genetic disorder. The more shared alleles at a variant in multiple patients, the more likely the variant is to be a casual mutation of the disease. However, it may be sometimes too strict to ask for a common allele. So, wKGGSeq can also search a common gene on which each affected subject has at least one predicted pathogenic mutation by using the overlapped gene function.

(5) De novo mutation strategy: it is also possible that the disease is caused by a novel mutation. Given a trio sample that consists of an affected offspring and healthy parents, wKGGSeq provides a function to scan de novo mutation(s) in the genome of the affected offspring based on genotypes called by third-party tools such as TrioDenovo (http://genome.sph.umich.edu/wiki/Triodenovo). It applies an inheritance model filtration to remove all sequence variants on which all alleles of a child are present in his or her parents. The assumption is that the

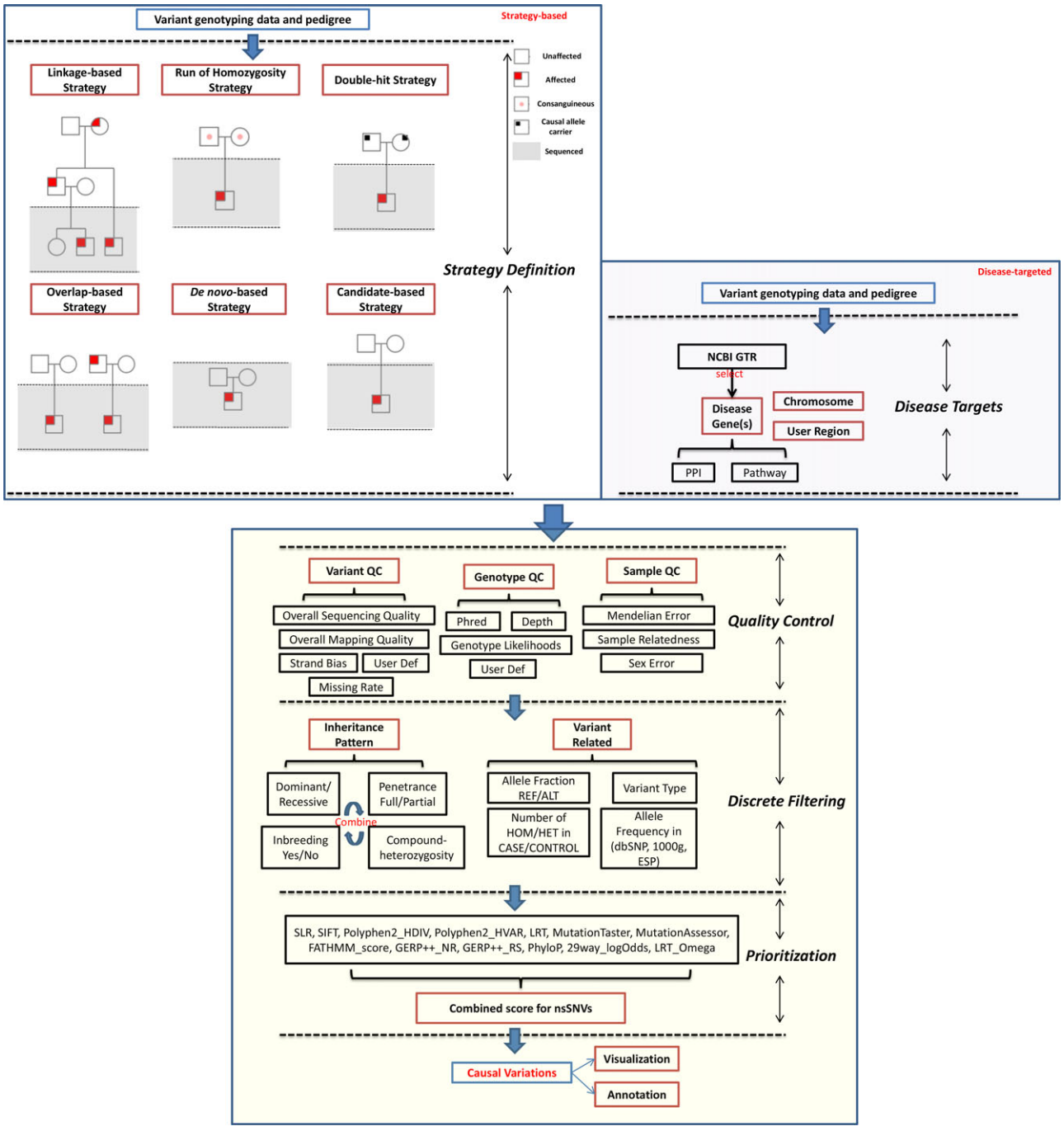

Figure 1. The components and framework of wKGGSeq.

phenotype discordance between the offspring and parents is caused by a de novo mutation. Note that Mendelian errors caused by technique artifacts often confound the analysis. To reduce false de novo mutations, higher coverage (e.g., over 20×) with good sequencing quality (Phred-scaled score over 30) is required.

(6) Candidate gene strategy: when there are a number of known causal genes in mind for a disease, wkGGSeq provides a quick search function to retrieve sequence variants in these candidate genes. This analysis strategy will be very effective to identify novel mutations of known causal genes.

The shared filtration, annotation, and prioritization functions of aforementioned analysis strategies are described in the Supplementary Notes. All of those functions have been implemented into wKGGSeq.

Disease-Targeted Analysis

To characterize pathogenic mutations in established causal genes for well-studied diseases, such as hereditary cancers, hearing loss, and cystic fibrosis, we have integrated a collection of causal genes of various human diseases from NCBI GTR into a disease-targeted

filtration and prioritization analysis. NCBI GTR has curated a lot of disease-causal gene panels from GeneReviews, OMIM, and many clinical resources. We have written a bridge program to synchronize the gene and disease list in wKGGSeq with the resource data stored in NCBI GTR. On wKGGSeq, one can easily retrieve known causal genes of a disease by the key disease term(s). The analysis of this module starts with automatically retrieving casual genes of specified diseases and subsequently extracts all variants with good quality (described below) inside those genes. As variants of those genes may have different potentials to cause a disease, wKGGSeq will then perform common variant-level filtration and prioritization, such as elimination of common and evolutionary neutral variants in human populations (see online Supplementary Notes). This analysis module is able to facilitate clinical genetic diagnostics.

Data Management and Security

To ensure efficient and safe data delivery, we designed multiple procedures to help users manage their original input data and analysis results online. Moreover, we also deployed the wKGGSeq framework on a local Linux virtual machine using Oracle VM VirtualBox. When using wKGGSeq on a local server, users can fully control their own data regardless of any institutional policy about the data security (online Supplementary Notes).

Visualization and Annotation

wKGGSeq provides abundant functions to visualize and annotate result of each submitted job. In an overview page, we designed a job report page that summarizes the number of retained variants of each filtration and prioritization analysis step. This page also contains several charts to show distribution of variants in different functional classes. For each candidate mutation in the resulting table, wKGGSeq not only shows original annotations of KGGSeq, it also has a backend MySql database to record extra genomic annotations including microRNA target sites and available genetic association information, as well as external links to UCSC genome browser [Karolchik et al., 2014] and SNVrap [Li et al., 2012a; Li and Wang, 2014].

Real-Exome Sequencing Samples to Evaluate the Framework

We used three in-house real exome sequencing pedigrees to test the performance of wKGGSeq. The first pedigree contained an offspring affected with neonatal-onset Crohn's disease and his healthy parents. The second pedigree only consisted of an aunt and a nephew, which were affected with spinocerebellar ataxia. The third pedigree contained six sequenced subjects in two generations (including a healthy mother, a healthy child, and four children affected with familial spastic paraplegia). The underlying causal mutations have been proposed by their original studies [Mao et al., 2012; Li et al., 2013a; Li et al., 2014]. All subjects were collected in Hong Kong with Institutional Review Board approval and were sequenced by the Illumina Genome Analyzer II or HiSeq 2000 platform. The paired-end short reads produced by the sequencer were aligned and mapped onto the UCSC human reference genome (hg19), by Burrows–Wheeler aligner [Li and Durbin, 2009]. Duplicated reads were marked by Picard. The genome analysis toolkit [DePristo et al., 2011] was then used for quality score recalibration, realignment, and sequence variant calling (by UnifiedGenotyper).

We compared the basic gene feature annotations of wKGGSeq with two popular online tools, wANNOVAR (http://wannovar2.usc.edu/) and VEP (http://asia.ensembl.org/info/docs/tools/vep) and the whole downstream analysis pipeline for these in-house datasets.

Results

wKGGSeq Online Analysis System

The wKGGSeq online downstream analysis system for exome sequencing data is the major result of the present work. It provides a series of functions to facilitate parameter setting, result visualization, and data management. One of the important unique features of wKGGSeq is that it has the strategy-based and disease-targeted pipelines to effectively ease the analysis.

The usage of the system for filtering and prioritizing exome sequencing data is straightforward. One need upload called sequence variants in VCF (http://www.1000genomes.org/node/101) format and subject's phenotype information first. For strategy-based analysis, the next step is to choose an analysis strategy on the Web page. For disease-targeted analysis, known causal genes of a specified disease can be selected on the system. One then only need click a "submit" button on the Web page to submit an analysis job with default settings. Advanced users can also tune the settings such as allele frequency cutoff or variant attribute for hard filtering prior to the submission. Once the job is finished, summary filtration and prioritization results can be visualized in plots and shortlisted variants are shown in tables on different interactive Web pages. Each variant are also linked to external bioinformatics databases for more annotations. More descriptions of Web interface and usage are detailed in the Supplementary Notes.

Comparisons of the Basic Gene Feature Annotation of wKGGSeq with wANNOVAR and VEP

According to McCarthy et al. (2014), it remains a challenge to accurately annotate all sequence variants with gene features. Therefore, we specifically asked how consistent are these gene feature annotations between wKGGSeq and two other popular tools (wANNOVAR and VEP). Table 1 shows the comparison results with wANNOVAR and VEP, respectively, in our familial spastic paraplegia pedigree based on RefSeq gene database (probably the most widely-used one). The results in the other two pedigrees are similar. First of all, the annotation consistency between wKGGSeq and wANNOVAR is very high, ~98.2% out of ~20,000 variants in exonic regions. However, wKGGSeq predicts 22, 5, and 199 more stoploss, stopgain, and missense variants, respectively, than wANNOVAR (Table 1). Around 90% of these wKGGSeq additional variants are predicted to be "exonic_unknown" in wANNOVAR. No stoploss, stopgain, and missense and splicing variants of wANNOVAR are missed by wKGGSeq. wANNOVAR maps a variant onto RefSeq cDNAs. The exonic_unknown variants often occur in the RefSeq cDNAs mapped human reference genome with short insertions or deletions (e.g., NM_000363). On the other hand, wKGSeq corrects the coordinate shifts because of the insertions or deletions when mapping a variant onto human reference genome. Besides, some of these wKGGSeq additional variants have multiple different features in alternative transcripts but wANNOVAR provides only one annotation. More descriptions of gene feature annotation can be seen in the Supplementary Notes and in Supp. Figure S1. Most of these wKGGSeq unique variants have matched annotation in dbSNP (version 142). It should be noted that variants without matched

Table 1. Gene Feature Annotation by wKGGSeq, wANNOVAR, and VEP Based on RefGene Database in an Exome Sequencing Dataset of 66,272 Called Variants

	wKGGSeq versus wANNOVAR			wKGGSeq versus VEP		
	wKGGSeq unique (#dbSNP)	Overlapped	wANNOVAR unique (#dbSNP)	wKGGSeq unique (#dbSNP)	Overlapped	VEP unique (#dbSNP)
Stoploss	22(14)[a]	13	0	22(11)	13	2(1)
Stopgain	4(3)	75	1(0)[b]	15(11)	64	7(1)
Missense	199(128)	9,210	0	1,550(1,345)	7,856	114(43)
Synonymous	112(77)	10,129	16(12)	1,637(1,204)	8,604	92(26)
Splicing	3(1)	81	0	49(20)	35	1,564(123)

[a]This includes some variants which have reversed reference and alternative alleles.
[b]This is actually a stopgain mutation by an 1-bp deletion frameshift, so it could be a frameshift as well.
#dbSNP: the number of variants supported by annotations in dbSNP. However, it should be noted that variants without support from dbSNP are not necessarily incorrect.

annotation in dbSNP are not necessarily incorrect as some variants have not been registered yet and some are mapped onto different alternative transcripts of the same gene. Unexpectedly, compared with VEP, wKGGSeq predicts many more nonsynonymous variants based on RefSeq gene database (Table 1). Again, most of these wKGGSeq unique variants have matched annotation in dbSNP (version 142). Among the 1,550 wKGGSeq unique missense variants, 74% are annotated as intron, upstream, and downstream variants by VEP. In the present study, we did not do the comparison in other reference databases such as GENCODE and KnownGenes because this is beyond the main scope of the present paper and the general conclusion of no perfect annotation tools would be similar to a recent review [McCarthy et al., 2014]. In summary, wKGGSeq showed an elevated concordance with gene feature annotation in dbSNP, compared with wANNOVAR and VEP.

Comparisons of wKGGSeq with wANNOVAR and VEP

wKGGSeq has an intuitive, well-organized, and easy-to-use Web interface to facilitate criteria setting, data submission, and result interpretation under the strategy-based and disease-targeted frameworks (see description and Web server usage on the online Supplementary Notes, Supp. Figs. S2 and S3, and the Supplementary Manual). As listed in Table 2, wKGGSeq has more functions of the three tools for filtration and prioritization of exome sequencing data. Because simulation is difficult to convincingly mimic functional features of variants and genes on a whole exome, we therefore used three real exome sequencing pedigrees as proof-of-principle examples to demonstrate the usefulness and effectiveness of the strategy-based and disease-target framework. Due to limited genetic information in those small samples, pinpointing the causal mutations is usually difficult. At the same time, we also compared the performance of wKGGSeq in terms of filtration and prioritization with wANNOVAR and VEP.

Strategy-based analysis framework for neonatal-onset Crohn's disease

The neonatal-onset Crohn's disease sample contained three sequenced subjects, a male patient and his unaffected parents. The estimated IBD proportions of child–father, child–mother, and

Table 2. Key Functions of the Three Tools for Filtration, Prioritization, and Annotation

	wKGGSeq	wANNOVAR (version 2)	VEP
Quality control	Systematic QC on genotype, variant, and subject levels	Only by depth and sequencing quality	No
Use disease mode	Recessive, dominant compound heterozygous, de novo, and runs of homozygosity	Too simple, cannot directly take genotypes of controls; it only works for recessive and dominant modes	No
Variants reference databases	dbSNP, 1000 Genomes Project, and ESP	Similar to wKGGSeq	Similar to wKGGSeq
Gene annotation	RefGene GENCODE UCSC knownGene	RefGene GENCODE UCSC knownGene ENSEMBL	ENSEMBL RefGene
Functional prediction	SLR SIFT Polyphen2_HDIV Polyphen2_HVAR LRT MutationTaster MutationAssessor FATHMM_score CADD_score GERP++_NR GERP++_RS PhyloP100way_vertebrate 29way_logOdds	Same as wKGGSeq	SIFT Polyphen2_HDIV Polyphen2_HVAR
PPI and pathway	Yes	No	No
Literature	PubMed	No	No
Management of input data	Yes	No	No
Disease-targeted prioritization	Yes	No	No

Table 3. The Number of Retained Sequence Variants by wKGGSeq in the Neonatal-Onset Crohn's Disease Sample

Functions	#Sequence variants	
	De novo mutation	Double-hit gene
Initial	1,196,282	
Quality control on genotype and variants levels	68,992	
Inheritance pattern	197[a]	–
Rare in dbSNP + 1000 Genome + ESP[b]	133	1,325
Not altering amino acid[c]	7	253
Superduplicate regions[d]	1	228
Predicted to be nonpathogenic	0	171
Hitting both copies of a gene	–	3[e]

[a]The de novo mutation-based inheritance model filtration removes all sequence variants at which all alleles of a child are present in his or her parents.
[b]The rare variants referred to variants with MAF≤3% in the datasets.
[c]This category includes missense, stopgain, stoploss, and splicing single-nucleotide variants and insertions/deletions causing frameshift, nonframeshift, stoploss, stopgain, and splicing differences.
[d]Variants in putative genomic duplications defined in a dataset (genomicSuperDups) from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz), which have higher genotyping error rate (http://blog.goldenhelix.com/?p = 1153).
[e]Double-hit gene filtration function is to only retain variants at which each parent has transmitted at least one mutation in the same gene to the child.

Table 5. The Number of Retained Sequence Variants by wKGGSeq in a Dominant Mode

Functions	Spinocerebellar ataxias	Familial spastic paraplegia
Initial	1,417,935	1,017,018
Quality control on genotype and variants levels	82,543	84,119
Inheritance pattern[a]	21,122	937
Filter all variants in reference database	268	74
Rare in dbSNP + 1000 Genome + ESP[b]	248	69
Protein altering variants[c]	39	4
Superduplicate regions[d]	36	4
Over four variants within the same gene	28	4
Predicted to be nonpathogenic	28	4

[a]Dominant mode only considered variants in heterozygous genotypes and with shared alleles between the two patients.
[b]The rare variants referred to variants with MAF≤1% in the datasets.
[c]This category includes missense, stopgain, stoploss, and splicing single nucleotide variants and insertions/deletions causing frameshift, nonframeshift, stoploss, stopgain, and splicing differences.
[d]Variants in putative genomic duplications defined in a dataset (genomicSuperDups) from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz), which have higher genotyping error rate (http://blog.goldenhelix.com/?p=1153).

father–mother pairs (0.50, 0.50, and 0) on wKGGSeq were identical to the expected values, suggesting no cross-sample contamination and no consanguineous mating. According to their disease statuses, two analysis strategies are potentially applicable for this pedigree: de novo mutation, and double-hit gene strategies. Table 2 lists the number of variants passing each filtration function sequentially. It turned out that the double-hit gene strategy produced the best candidate list in which only three variants retained and the true compound heterozygous mutations' gene, *IL10RA*, was highlighted by protein–protein interactions (PPIs) and shared pathways with some known risk genes of Crohn's disease (Table 3). In addition, *IL10RA* was suggested as causal genes by two studies [Glocker et al., 2009; Begue et al., 2011] prior to the original study of this pedigree [Mao et al., 2012]. wANNOVAR used 13 filtering functions to generate a final list of two variants (Supp. Table S1). However, the true causal mutations were removed by "variants found in user-supplied controls" due to inconsistent inheritance model. The shortest list that contained the causal mutations has 53 variants (Table 4). In contrast, VEP has very weak filtration function and produces a final candidate list of 4,232 variants, which is unfeasibly too large for further validation (Supp. Table S2; Table 4).

Strategy-based analysis framework for Spinocerebellar ataxias

According to the pedigree diagram (Fig. 1 of Li et al. [2013a]), the Spinocerebellar ataxias pedigree was very likely to have an autosomal-dominant mode. The estimated IBD proportion between the patients on wKGGSeq, 0.276, was close to the expected value 0.25, suggesting the sequencing data were free of cross-contamination. We chose linkage analysis strategy to filter and prioritize the sequence variants with the assumption of a dominant mode on wKGGSeq. As shown in Table 5, this pipeline reduced the number of candidate variants from 1,418,751 to 28. The PubMed mining function automatically found a previous paper [Wang et al., 2010a] (prior to this data's original publication [Li et al., 2013a] that suggested that *TGM6* was a causal gene of the same disease. wANNOVAR also effectively produced a final list of 17 variants (Supp. Table S1). However, the causal variant was not in the final list, which was filtered out at the 13th step ("the prioritized genes from Phenolyzer"). The size of the final list retaining the causal mutation

Table 4. The Summary Filtration and Prioritization Results of Three Tools in Three Pedigrees

	wKGGSeq	wANNOVAR	VEP
		Neonatal-onset Crohn's disease	
Initial variants	1,196,282	68,992[a]	68,992[a]
Retained variants when the causal mutations were kept finally	3	53	4,232
Variant hit by PPIs, pathways, or PubMed search	-	-	-
Additional evidence to highlight the causal mutations	PPI+Pathway+PubMed	-	-
		Spinocerebellar ataxias	
Initial variants	1,417,935	82,465[a]	82,465[a]
Retained variants when the causal mutations were kept finally	28	29	6,501
Variant hit by PPIs, pathways, or PubMed search	3	-	-
Additional evidence to highlight the causal mutations	Pathway+PPI	-	-
		Familial spastic paraplegia	
Initial variants	1,017,018	63,207[a]	63,207[a]
Retained variants when the causal mutations were kept finally	4	7	5,109
Variant hit by PPIs, pathways, or PubMed search	3	-	-

[a]As wANNOVAR cannot effectively map variants on VCF data, KGGSeq was used to do the basic quality control on VCF data of affected samples.

was 29 for wANNOVAR. In contrast, the VEP has a rather weak automatic filtration function and fails to generate a short validation list (Table 4; Supp. Table S2).

Strategy-based analysis framework for familial spastic paraplegia

The familial spastic paraplegia pedigree was also very likely to follow an autosomal-dominant mode according to the pedigree diagram (Fig. 1 of Li et al. [2014]). So we chose linkage analysis strategy to filter and prioritize the sequence variants with the assumption of a dominant mode on wKGGSeq. Compared with the above Spinocerebellar ataxias pedigree, this pedigree has four more subjects sequenced and two of which are closely related healthy members. As expected, making use of the additional genetic information, wKGGSeq produced a much shorter candidate variant list for this pedigree than the above Spinocerebellar ataxias pedigree. wKGGSeq's pipeline successfully reduced the number of candidate variants from 1,017,018 to four (Table 5). The reported causal variant (p.R268Q of *ATP2B4*) had the highest pathogenic scores among the four variants. wANNOVAR also successfully produced a short list of seven variants, which is comparable to wKGGSeq (Supp. Table S1). Both tools have reported causal mutation retained in the final list. However, the VEP has very weak automatic filtration functions and can only produce a candidate list with 5,109 variants in this pedigree (Table 4; Supp. Table S2).

Disease-Target Prioritization Analysis in Three Real Samples

Making use of the three real samples, we also imitated to conduct a clinical diagnosis on wKGGSeq. For the neonatal-onset Crohn's disease, we selected all known causal genes of different inflammatory bowel diseases on the disease-target prioritization page of wKGGSeq. By far, there have been 10 genes of nine different Crohn's diseases in total. We did not manually set any other parameters exception for specifying the local path of the 1,196,911 variants in a VCF file and a pedigree file. In about 5 min, the prioritization job finished and accurately pinpointed the two compound heterozygous mutations with a lot of annotations (including population allele frequencies, pathogenic prediction [Li et al., 2013b], and OMIM annotation). The spinocerebellar ataxias are much more heterogeneous. On wKGGSeq, 34 known causal genes of 33 different spinocerebellar ataxias were retrieved. The prioritization also quickly pinpointed the causal mutation in *TGM6* gene out of 1,418,751 variants. For the familial spastic paraplegia, wKGGSeq retrieved 34 candidate genes by the keyword "spastic paraplegia." The newly reported causal gene, *ATP2B4*, was not included. As expected, no potential pathogenic variants were reported by wKGGSeq for the clinical diagnosis.

As we failed to find any alternative tools to do the disease-targeted prioritization, there was no comparison for this module.

Discussion and Conclusion

Compared with existing tools, wKGGSeq provides a much more comprehensive and convenient framework to effectively annotate, filter, prioritize, and visualize sequence variants for exome sequencing studies of inherited diseases. An important novel feature of wKGGSeq is that it is equipped with the strategy-based and disease-targeted analysis procedures to ease a systematic downstream analysis of exome sequencing data and to facilitate clinical diagnosis of well-studied diseases. These procedures have been implemented in Web server with a user-friendly graphic interface. The running process of the filtration and prioritization analysis on the server can instantly be shown. The analysis results are presented by various formats (including plots and tables) to facilitate data quality inspection and results interpretation.

In the present paper, we only compared wKGGSeq with two tools (wANNOVAR and VEP) in three pedigrees. Although it would be more impressive to compare more tools in more real pedigrees, the two tools are almost the most popular tools at present and these pedigrees are difficult cases due to their small sample sizes as well as the limited genetic information for prioritization. As shown in Table 4, wKGGSeq produced shorter candidate lists than wANNOVA and VEP. These final lists have reasonable sizes for follow-up validation regardless of the pedigree sizes. In addition, neither wANNOVAR nor VEP can directly use pedigree information for a deeper filtration. To use the genotypes of unaffected family members for filtration on wANNOVA and VEP, one has to do it manually. In contrast, wKGGSeq allows users to load the pedigree information for a strategy-based filtration and prioritization in an automatic manner. Moreover, the disease-target analysis is a unique function compared with other tools, which will be useful for genetic diagnoses. More and more tools (including the academic ones [Habegger et al., 2012; Aleman et al., 2014] and commercial ones (Ilumina VariantStudio and ingenuity variant analysis) are being developed to facilitate exome sequencing studies of Mendelian diseases. It is almost impossible to exhaustively compare all tools and to describe the comparison results in the present paper. Systematic comparison of the existing tools for guiding choice in practice will be an important work in the future.

In summary, wKGGSeq is a useful tool for genetic investigators and even clinicians without advanced computing skills to carry out professional and comprehensive exome sequencing data analysis. This tool may substantially relieve them from learning complex computing skills to pay more attention to analysis design and results interpretation.

Acknowledgments

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Aleman A, Garcia-Garcia F, Salavert F, Medina I, Dopazo J. 2014. A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. Nucleic Acids Res 42:W88–W93.

Asmann YW, Middha S, Hossain A, Baheti S, Li Y, Chai HS, Sun Z, Duffy PH, Hadad AA, Nair A, Liu X, Zhang Y, et al. 2012. TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. Bioinformatics 28:277–278.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745–755.

Begue B, Verdier J, Rieux-Laucat F, Goulet O, Morali A, Canioni D, Hugot JP, Daussy C, Verkarre V, Pigneur B, Fischer A, Klein A, et al. 2011. Defective IL10 signaling defining a subgroup of patients with inflammatory bowel disease. Am J Gastroenterol 106:1544–1555.

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459–471.

Chang X, Wang K. 2012. wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet 49:433–436.

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10:5–6.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, et al. 2011. A framework

for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

Do R, Kathiresan S, Abecasis GR. 2012. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet 21:R1–R9.

Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20:490–497.

Glocker EO, Kotlarz D, Boztug K, Gertz EM, Schaffer AA, Noyan F, Perro M, Diestelhorst J, Allroth A, Murugan D, Hatscher N, Pfeifer D, et al. 2009. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. New Engl J Med 361:2033–2045.

Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A, Rozowsky J, Clarke D, Snyder M, Gerstein M. 2012. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 28:2267–2269.

Howrigan DP, Simonson MA, Keller MC. 2011. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC Genomics 12:460.

Kamphans T, Krawitz PM. 2012. GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics 28:2515–2516.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, et al. 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42:D764–D770.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li M, Ho PW, Pang SY, Tse ZH, Kung MH, Sham PC, Ho SL. 2014. PMCA4 (ATP2B4) mutation in familial spastic paraplegia. PLoS One 9:e104790.

Li M, Pang SY, Song Y, Kung MH, Ho SL, Sham PC. 2013a. Whole exome sequencing identifies a novel mutation in the transglutaminase 6 gene for spinocerebellar ataxia in a Chinese family. Clin Genet 83:269–273.

Li MJ, Sham PC, Wang J. 2012a. Genetic variant representation, annotation and prioritization in the post-GWAS era. Cell Res 22:1505–1508.

Li MJ, Wang J. 2014. Current trend of annotating single nucleotide variation in humans—a case study on SNVrap. Methods. [Epub ahead of print]

Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. 2012b. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic acids research 40:e53.

Li MX, Kwan JS, Bao SY, Yang W, Ho SL, Song YQ, Sham PC. 2013b. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet 9:e1003143.

Mao H, Yang W, Lee PP, Ho MH, Yang J, Zeng S, Chong CY, Lee TL, Tu W, Lau YL. 2012. Exome sequencing identifies novel compound heterozygous mutations of IL-10 receptor 1 in neonatal-onset Crohn's disease. Genes Immun 13:437–442.

Maranhao B, Biswas P, Duncan JL, Branham KE, Silva GA, Naeem MA, Khan SN, Riazuddin S, Hejtmancik JF, Heckenlively JR, Riazuddin SA, Lee PL, Ayyagari R. 2014. exomeSuite: Whole exome sequence variant filtering tool for rapid identification of putative disease causing SNVs/indels. Genomics 103:169–176.

McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, Donnelly P. 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Med 6:26.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069–2070.

Paulsen J, Lien TG, Sandve GK, Holden L, Borgan O, Glad IK, Hovig E. 2013. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. Nucleic Acids Res 41:5164–5174.

Rehm HL. 2013. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet 14:295–300.

Robinson PN, Krawitz P, Mundlos S. 2011. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet 80:127–132.

Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM, Carey JC, Opitz JM, et al. 2011. Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. Am J Hum Genet 89:28–43.

Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, Ovetsky M, Hem V, Gorelenkov V, Song G, Wallin C, Husain N, Chitipiralla S, et al. 2013. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. Nucleic Acids Res 41:D925–D935.

Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, Jiang H, Zhang P, Shen L, Guo JF, Li N, Li YR, et al. 2010. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 133:3510–3518.

Wang K, Li M, Hakonarson H. 2010b. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164.

Zhang L, Zhang J, Yang J, Ying D, Lau YL, Yang W. 2013. PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. Bioinformatics 29:124–125.