INFORMATICS

Human Mutation

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

Prioritizing Disease-Linked Variants, Genes, and Pathways with an Interactive Whole-Genome Analysis Pipeline

In-Hee Lee,[1] Kyungjoon Lee,[2] Michael Hsing,[1] Yongjoon Choe,[1] Jin-Ho Park,[1,3] Shu Hee Kim,[4] Justin M. Bohn,[1] Matthew B. Neu,[1] Kyu-Baek Hwang,[5] Robert C. Green,[6] Isaac S. Kohane,[1,2] and Sek Won Kong[1]*

[1]*Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Department of Medicine, Boston Children's Hospital, Boston, Massachusetts 02115;* [2]*Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115;* [3]*Department of Family Medicine, Seoul National University Hospital, Seoul 110-744, South Korea;* [4]*Humanities and Sciences, Stanford University, Palo Alto, California 94305;* [5]*School of Computer Science and Engineering, Soongsil University, Seoul 156-743, South Korea;* [6]*Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115*

Communicated by George Patrinos

Received 4 October 2013; accepted revised manuscript 23 January 2014.
Published online 29 January 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22520

ABSTRACT: Whole-genome sequencing (WGS) studies are uncovering disease-associated variants in both rare and nonrare diseases. Utilizing the next-generation sequencing for WGS requires a series of computational methods for alignment, variant detection, and annotation, and the accuracy and reproducibility of annotation results are essential for clinical implementation. However, annotating WGS with up to date genomic information is still challenging for biomedical researchers. Here, we present one of the fastest and highly scalable annotation, filtering, and analysis pipeline—gNOME—to prioritize phenotype-associated variants while minimizing false-positive findings. Intuitive graphical user interface of gNOME facilitates the selection of phenotype-associated variants, and the result summaries are provided at variant, gene, and genome levels. Moreover, the enrichment results of specific variants, genes, and gene sets between two groups or compared with population scale WGS datasets that is already integrated in the pipeline can help the interpretation. We found a small number of discordant results between annotation software tools in part due to different reporting strategies for the variants with complex impacts. Using two published whole-exome datasets of uveal melanoma and bladder cancer, we demonstrated gNOME's accuracy of variant annotation and the enrichment of loss-of-function variants in known cancer pathways. gNOME Web server and source codes are freely available to the academic community (http://gnome.tchlab.org).
Hum Mutat 35:537–547, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: whole-genome sequences; variant annotation; disease gene discovery; analysis pipeline

Introduction

The maturation of ultrahigh-throughput sequencing technology has opened a new era of personal genome sequencing [Ashley et al., 2010; Cirulli and Goldstein, 2010; Meyerson et al., 2010; Drmanac, 2011; Tabor et al., 2011; Chang and Wang, 2012; Kidd et al., 2012], and has shifted the researcher's burden from the identification of genetic variants to the interpretation of large numbers of variants in each individual. Although studies using whole-genome sequencing (WGS) in a large disease population might still be a few years away, proof-of-concept studies on WGS and whole-exome sequencing (WES) have already proven the technology to be useful in identifying disease-causing mutations in rare Mendelian disorders [Hoischen et al., 2010; Lalonde et al., 2010; Lupski et al., 2010; Ng et al., 2010a; Ng et al., 2010b; Roach et al., 2010; Bamshad et al., 2011; Klassen et al., 2011]. Moreover, a few studies using the case-control study design also demonstrated the utility of WGS and WES in identifying disease-associated genomic variants for nonrare diseases [Calvo et al., 2010; Pelak et al., 2010; Holm et al., 2011; Rivas et al., 2011].

An individual human genome has 3–4 million variants, or locations that differ from the human reference genome. Because of the large number of variants, it is essential to filter out those weakly associated with the researcher's target phenotype and to reduce these variants down to a manageable number. This can be done by means of various heuristics such as allele frequencies (AFs) and their impacts on protein functions [Cooper and Shendure, 2011; Goldstein et al., 2013]. There is a tremendous need in the biomedical research community for a tool that can filter these millions of variants based on the most up-to-date annotations and utilize the growing arsenal of genome analysis methods.

The number of bioinformatics pipelines for analyzing WGS and WES is rapidly increasing. However, a majority of such tools focus on processing raw sequence data to detect high-confidence genomic variants rather than focusing on downstream analyses such as annotation-based variant filtering and statistical analysis [McKenna et al., 2010; Lam et al., 2012; Pabinger et al., 2013]. Even available downstream analysis tools are limited by their (1) static-filtering methods, (2) insufficient annotation, and (3) absence of multigenome comparison methods [Wang et al., 2010; Ge et al., 2011; Yandell et al., 2011; Cingolani et al., 2012; MacArthur et al., 2012; San Lucas et al., 2012]. Moreover, these tools are difficult to use for most researchers and clinicians due to the lack of an intuitive user interface. To overcome these limitations, we developed

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

*Correspondence to: Sek Won Kong, Informatics Program, Boston Children's Hospital, 300 Longwood Ave. Enders 137, Boston, MA 02115. E-mail: sekwon.kong@childrens.harvard.edu

Contract grant sponsors: NIH (NHGRI U01HG006500); NIMH (P50MH094267, R01MH085143); NRF of Korea (2012R1A1A2039822).

gNOME, an interactive downstream analysis pipeline that combines comprehensive genomic annotation sources with statistical analysis in an expandable framework. We demonstrated the accuracy of annotation using the validated genomic variants from published WES datasets. The pipeline is written in C++, Perl, and SQL, and all source codes are freely available to the academic community (http://gnome.tchlab.org).

Materials and Methods

Overview of gNOME Workflow

The goal of most WGS and WES studies is to find variants that are possibly associated with a phenotype of interest. A common approach toward this goal is to prioritize variants that have deleterious impact on protein function and/or are more frequently observed in cases compared with controls and an ethnicity-matched healthy population [Lim et al., 2013]. Following this strategy, gNOME's streamlined analysis workflow is as follows: (1) creating a project and uploading variant files, (2) annotation, (3) filtering variants using an interactive user interface, (4) statistical analysis, and (5) summarizing the results (Fig. 1). We use a double colon (::) to indicate a *menu*::*submenu* in the gNOME interface and single quotation marks to denote selected values throughout the description.

The first step is to define a project with a corresponding experimental design (i.e., "Case only" or "Case vs. Control") and to upload variant files (Step 1). The pipeline supports the variant call format (VCF) [Danecek et al., 2011], genome variation format (GVF) [Reese et al., 2010], and Complete Genomics' VAR file format. Each variant file should be assigned to groups (either "case" or "control") in a project with a specific reference genome build (i.e., "hg18" and "hg19"). A set of samples in gNOME is distinguished by the project name and group label. The uploaded variant files are placed in the internal queuing system for annotating with 60 different sources of genomic information collected from 17 publicly available databases (Step 2) (see *Materials and Methods* and Supp. Table S1 for details). This step takes at most 30 min for an individual WGS with 4–5 million variants. For efficient handling of dataset with multiple genomes, it is recommended to upload them as multiple sample VCF files, such that gNOME can speed up the annotation step by processing the entire variants—union of variants found in any of genomes in the file—in a single step. Once the annotation is completed, each genome in the multiindividual VCF file is stored individually. For instance, the merged VCF file for 1,097 samples of the 1000 Genomes Project (1KGP) has only 39,706,715 variants, for which gNOME can complete the annotation in 50 min (see *Results*). The resulting annotated variant files are stored in an internal MySQL database. Once the annotation is completed, users will receive a notification e-mail. The summary statistics for all uploaded variant call files are available on *3.Summary*::*Genome Level* (Supp. Fig. S1). This overview is available for a single genome at a time or for multiple genomes. When a group of genomes is selected, gNOME displays the average and range of summary statistics for each variant type.

In Step 3, users can select multiple criteria for annotation-based filtering through an interactive Web interface. For instance, one can select rare or novel loss-of-function (LoF) variants at highly conserved loci that are exclusively found among cases but neither in controls nor in an ethnicity-matched population dataset. Also, during this step, possible false positives are reduced by filtering out low-quality variants and variants found in repetitive regions. We grouped filtering options into four broad categories in the Web interface: (1) Allele Frequency, (2) Functional Impact, (3) Knowledge Enrichment, and (4) Others (Fig. 2A). The LoF variants at highly conserved

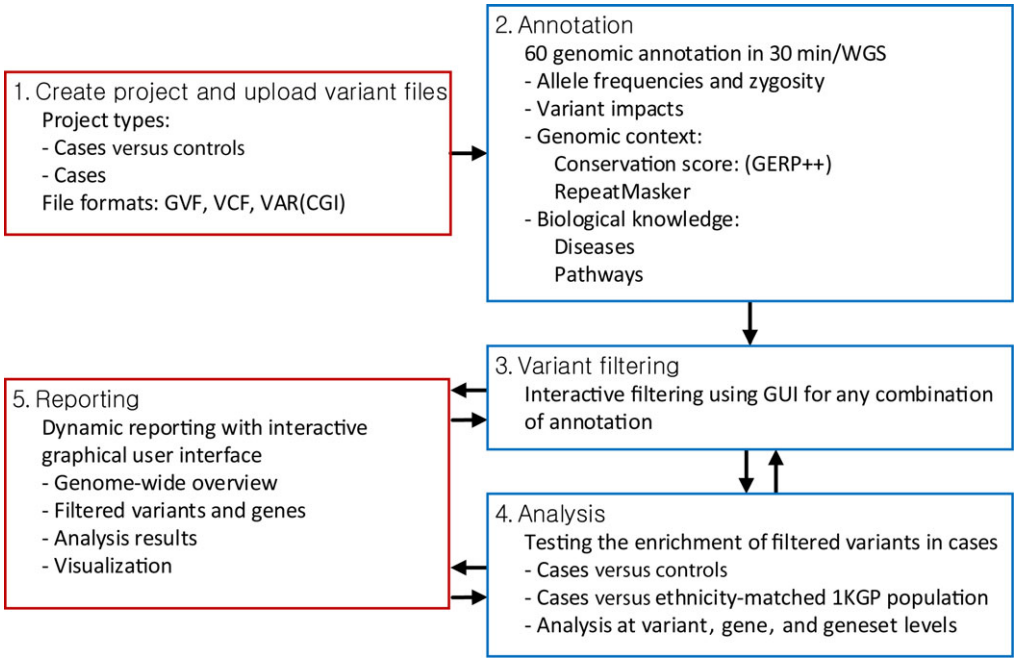

Figure 1. A schematic overview of gNOME. The analysis of whole-genome and whole-exome dataset starts with creating a project and uploading it according to project type (Step 1). The uploaded files are annotated with 60 annotation tracks (Step 2), and annotation-based variant filtering can be interactively performed (Step 3). gNOME supports variant-, gene-, and gene set-level association tests between two groups: case versus ethnicity-matched population data from the 1KGP or cases versus controls (Step 4). Filtering and analysis results are dynamically reported on the Web-based interface (Step 5). Steps 3–5 can be performed iteratively based on different variant-filtering criteria.

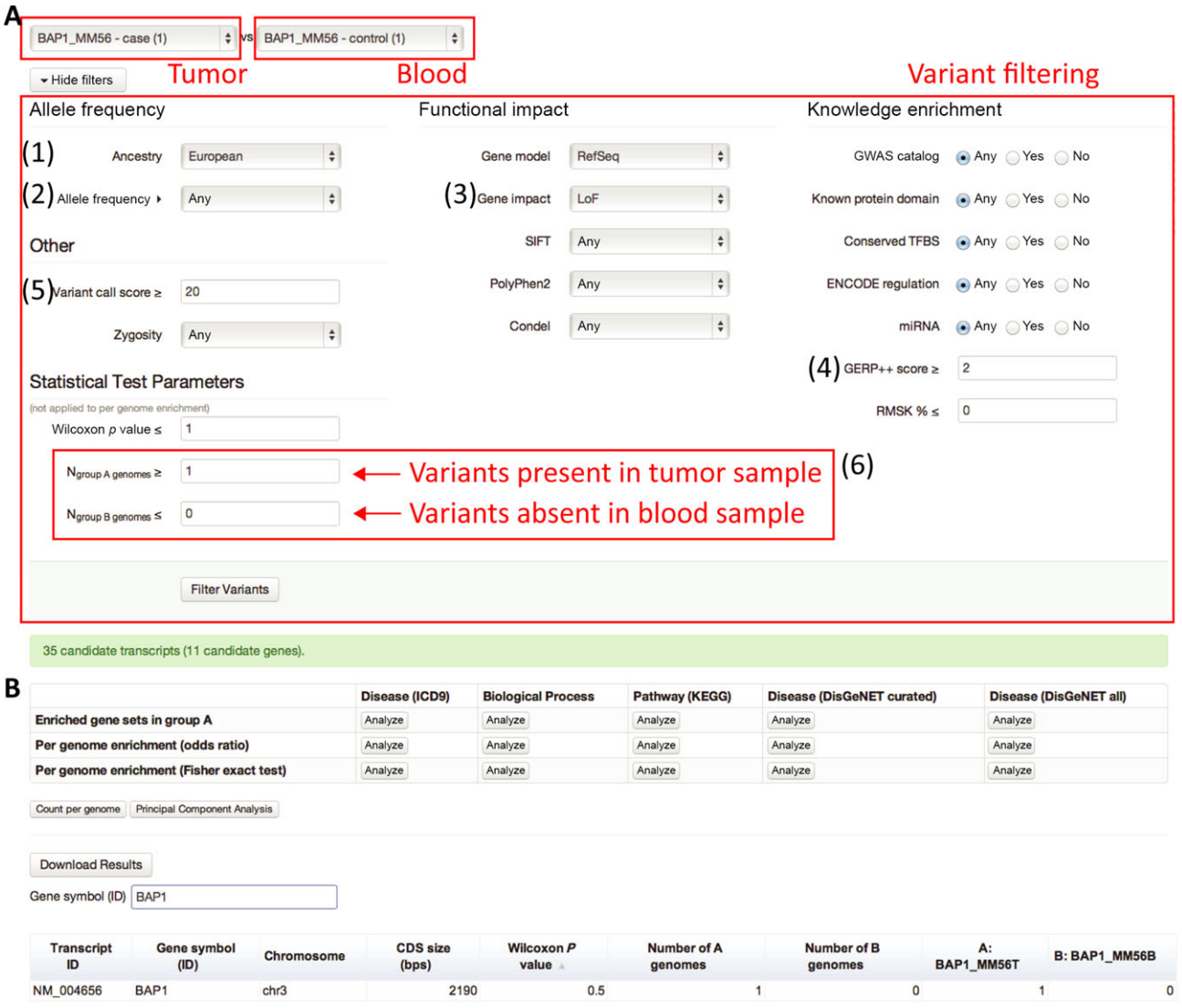

Figure 2. Discovering somatic mutations in tumor–blood paired WESs. **A:** A screenshot for comparing variants from tumor tissue (as "case") and blood sample (as "control"), both of which come from a single patient ("MM56") (see *Finding Somatic Mutations in Uveal Melanoma* and *Materials and Methods* for detail). From both tumor tissue and blood sample, AFs were estimated with (1) European ancestry, and (2) rare or novel (3) LoF variants at (4) highly conserved loci were selected. Low-quality variants were excluded by setting (5) "Variant call score ≥ 20." The potential somatic mutations were selected by choosing variants that were present in tumor sample but not in blood sample (6). **B**: The result from the comparison shown in (**A**). The table can be searched for gene symbol or sorted by the columns. A total of 11 genes including *BAP1* (displayed) met the criteria. gNOME performs a gene set enrichment analysis for five gene set categories with the genes that passed filtering criteria.

loci that are rare or novel in the European population are selected by setting (1) *Allele Frequency*::*Ancestry* "European" (Fig. 2A-1), (2) *Allele Frequency*::*Allele Frequency* "≤1% (rare)" (Fig. 2A-2), (3) *Functional Impact*::*Gene impact* "LoF" (Fig. 2A-3), and (4) *Knowledge Enrichment*::*GERP++ score* ≥ "2" (Fig. 2A-4). We can exclude the variants with low calling quality scores by setting *Other*::*Variant call score* (Fig. 2A-5). The selected variants or genes are displayed in a table that can be sorted by column, and are available for download as a tab-delimited text file, which can be used as an input for the other protein–protein interaction network-based analysis tools such as DAPPLE [Rossin et al., 2011]. *3.Summary::Variant Level* lists the detailed annotations for all variants that passed the filtering criteria, whereas *3.Summary::Gene Level* shows the number of variants that met the criteria for each gene.

Group comparison, Step 4, is one of the unique features that distinguish gNOME from other WGS and WES annotation tools [Wang et al., 2010; Ge et al., 2011; Cingolani et al., 2012]. Group comparison helps to identify a set of variants, genes, and gene sets that are significantly enriched in cases as described in *Materials and Methods*, and is also useful to identify possible false-positive incidental findings such as platform-specific sequencing errors and hypervariable genes and gene sets [Kohane et al., 2012]. These genes can be easily identified in gNOME and filtered out for further analysis if desired (Fig. 2A-6). *4.Analyze::Variants* and *4.Analyze::Genes* (Fig. 2) identify interesting variants enriched in case genomes and genes with such variants. Additionally, in *4.Analyze::Genes*, gNOME can test whether a set of genes with interesting variants are enriched in precompiled gene sets from the Gene Ontology terms, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and the other disease–gene association databases (Fig. 2B). We demonstrated the performance, accuracy, and group comparison features using three publicly available WGS and WES datasets.

Ethnicity-Specific AFs of Known Variants

To calculate ethnicity-specific AFs, we used The NCBI Short Genetic Variations database (dbSNP, http://www.ncbi.nlm.nih.gov/SNP, version 137), the 1KGP [Genomes Project et al., 2012] for European, Asian, and African populations, and the Exome Sequencing Project (ESP, http://esp.gs.washington.edu) [Fu et al., 2013] for European and African populations. The datasets with less than 15 samples in dbSNP were not used due to the inaccuracy in estimating AFs. We categorized AFs into four groups: common (AF $\geq$ 5%), less common (1% $\leq$ AF $<$ 5%), rare (AF $<$ 1%), and novel. The numerical codes −1 and −10 are used to represent a reported variant without a known AF and a novel variant, respectively. If AFs from different data sources were inconsistent, the highest value was used to represent the ethnicity-specific AF. The same rule was applied for mixed ancestries.

Possible Impact on Protein Function

Predicting the functional impact of amino acid changes resulting from nucleotide changes is an important step for prioritizing disease-associated genes since most known disease-causing variants are in protein-coding regions [Choi et al., 2009]. To provide possible consequences of genomic variants in genic regions, we integrated multiple gene models and prediction algorithms as part of gNOME. The Reference Sequence database (RefSeq, http://www.ncbi.nlm.nih.gov/refseq) [Pruitt et al., 2005], Consensus Coding Sequence (CDS) project (CCDS, http://www.ncbi.nlm.nih.gov/CCDS) [Pruitt et al., 2009], Ensembl (http://www.ensembl.org) [Hubbard et al., 2002] and University of California Santa Cruz (UCSC) Known Genes [Hsu et al., 2006] were all implemented in our gene annotation database. The use of multiple transcript models to estimate the functional impact of a variant is essential since possible consequences of a variant can be different across transcript models and an intronic variant in one transcript model can be in the coding region of the other transcript model (see Supp. Fig. S2 for an example). Possible impacts of a variant on each transcript are categorized into synonymous, missense, in-frame insertion, in-frame deletion, splice-site disruption, nonstop, misstart, frameshift, and nonsense. LoF variants were defined to include splice-site disruption, frameshift, and nonsense. A broader category of LoF variants (adding nonstop and misstart) is also provided. We also annotate predicted impacts on protein function using the database for nonsynonymous SNPs' functional predictions (dbNSFP, https://sites.google.com/site/jpopgen/dbNSFP) [Liu et al., 2011] that comprises the predicted impacts estimated using the Sorting Intolerant from Tolerant (SIFT) [Kumar et al., 2009], PolyPhen2 [Adzhubei et al., 2010], MutationTaster [Schwarz et al., 2010], and a likelihood ratio test [Chun and Fay, 2009].

Conservation Scores, Noncoding Elements, and Biomedical Knowledge Enrichment

Conservation scores according to the Genomic Evolutionary Rate Profiling (GERP++, http://mendel.stanford.org/Sidowlab/downloads/gerp) [Davydov et al., 2010] are used to filter variants per locus, and we used an average score of GERP++ for insertions and deletions (indels). Genotyping errors are more frequently observed in repetitive regions, thus excluding the variants in these regions can reduce false-positive findings. We used the RepeatMasker database (http://www.repeatmasker.org) to find any variants in these regions [Smit et al., 1996–2010]. The default value for percent overlap with RepeatMasker regions is set to 0%.

A variant on an important functioning protein domain could have a significant impact on protein function. Known protein domains were collected from the InterPro database [Hunter et al., 2012] and mapped to the reference genome coordinates to facilitate variant annotations. The regulatory regions from the Encyclopedia of DNA Elements (ENCODE) project (http://encodeproject.org/ENCODE) [Consortium, 2011], the conserved transcription factor binding sites from UCSC Table Browser (http://genome.ucsc.edu), and microRNA host genes [Kozomara and Griffiths-Jones, 2011] were included in the annotation database to provide further information for noncoding functioning elements.

The pipeline includes gene sets for diseases, biological processes, and canonical pathways. A total of 1,253 disease-associated gene sets were compiled through the gene-to-disease mapping using the literature abstracts annotated with Medical Subject Heading (MeSH) terms and NCBI Genes [Mitchell et al., 2003]. The known disease-associated genes and variants from the catalog of genome-wide association studies by National Human Genome Research Institute (http://www.genome.gov/gwastudies/) [Hindorff et al., 2012], Online Mendelian Inheritance in Man (OMIM, http://www.omim.org) [McKusick, 2007], ClinVar (http://www.ncbi.nlm.nih.gov/clinvar) [Riggs et al., 2013], and DisGeNet (http://ibi.imim.es/DisGeNet/web/v02/home) [Bauer-Mehren et al., 2010] were integrated to the annotation database; the Human Gene Mutation Database (HGMD, http://www.hgmd.org) can be used if a user has the license. In addition, we collected 828 biological process gene sets based on the Gene Ontology (GO) terms and 186 KEGG pathways from the Molecular Signatures Database (MSigDB, http://www.broadinstitute.org/gsea/msigdb, c5.bp.v3.0 and c2.cp.v3.0, respectively) [Subramanian et al., 2005]. The original data sources and processing scripts are available on the gNOME Website, and the annotation database will be updated with the latest annotation information every 6 months.

Integration of Population-Scale Individual WGS Data

The purpose of using ethnicity-matched population datasets as a comparison group is twofold. First, the false-positive incidental findings can be identified and reduced as previously described [Kohane et al., 2012]. Second, the genetic burden due to a set of interesting variants in the ethnicity-matched general population can be estimated and compared with study individuals. The comparison of uploaded data with population-scale data from the 1KGP is one of the unique features of gNOME. Following the categorization of the 1KGP, we included 18 different population categories: four by ancestry—European, Asian, African, and admixed American—and 14 by geographical regions.

Statistical Comparison at Variant, Gene, and Gene Set Levels

A set of variants that met a user's annotation filtering criteria can be tested for enrichment in a group. Given two groups of samples (i.e., cases and controls), the interesting variants that are overrepresented in cases can be identified as follows. Supposing the existence of N individuals and a total of M LoF variants, we defined M as the number of unique LoF variants across N individuals and set two groups as G_0 (for instance, case group) and G_1 (for instance, noncase group or ethnicity-matched population from the 1KGP).

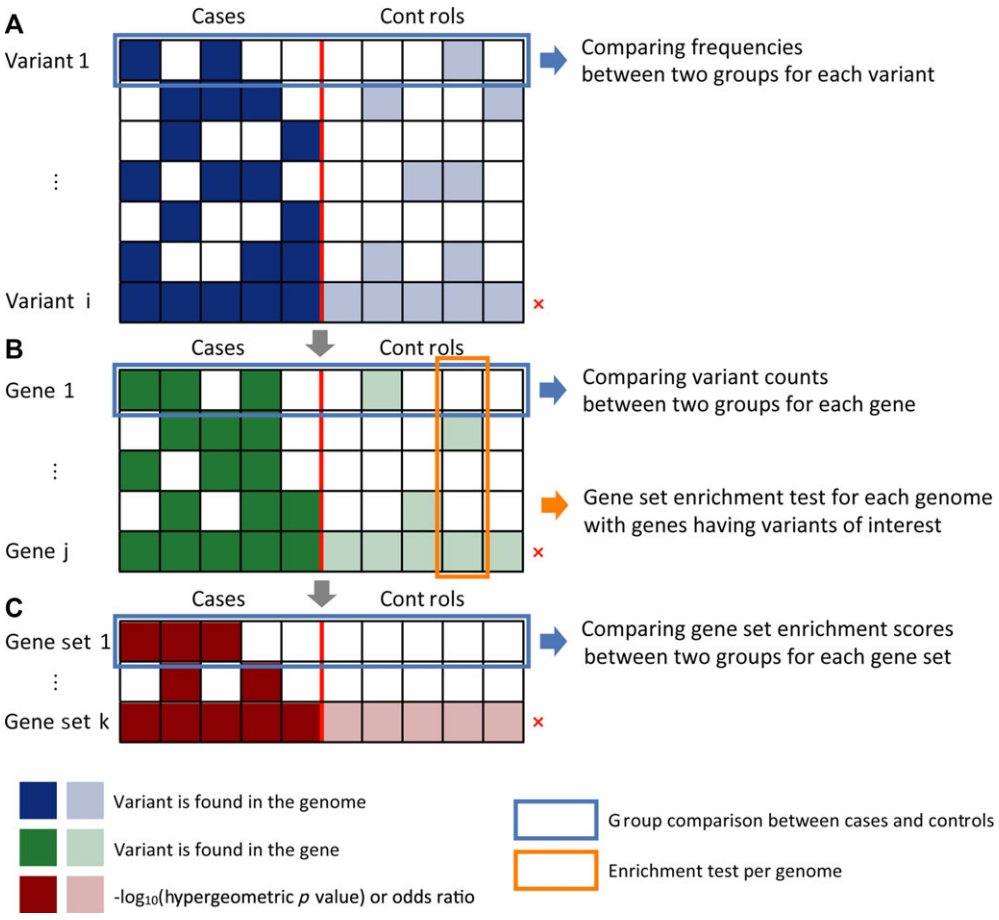

Figure 3. Association tests for variants, genes, and gene sets between two groups. The small number of variants that remain after the annotation-based filtering can be associated with a phenotype in three ways. First, we can test whether a specific variant presents more frequently in cases compared with controls or an ethnicity-matched population (**A**). Second, an association test can be performed at the gene level when each case individual may have different variants on the same gene (**B**). Third, we can expand a gene-level aggregation to gene set level to find the gene set overrepresented with interesting variants among cases (**C**). The rows marked by x (red) denote "hypervariable" variants, genes, or gene sets that frequently have variants in both cases and controls (see *Materials and Methods* for details).

Then, an M-by-N matrix can be expressed as

$$\mathbf{V} = \{v_{i,j}\}_{1\le i\le M, 1\le j\le N}$$

where

$$v_{i,j} = \begin{cases} 1 & i\text{–th variant found in } j\text{–th individual} \\ 0 & \text{otherwise} \end{cases}$$

The matrix V is illustrated as the colored grid box with five cases and five noncases in Figure 3A. Row-wise hypergeometric tests find the variants that are more frequently found in G_0. Alternatively, the number of individuals with interesting variants can be set for each group as shown in Figure 2A-6. For instance, by setting $N_{\text{group A genomes}} \ge 1$ and $N_{\text{group B genomes}} \le 0$, a set of interesting variants will be further filtered to a smaller set of variants that are exclusively present in Group A.

With the exception of a few Mendelian disorders, the likelihood of finding the same disease-linked variants across the patients is low [McClellan and King, 2010]. Instead, multiple rare LoF and missense variants in the same gene or the same pathway could alter disease risks. Burden tests and kernel-based tests compare the cumulative effects of such variants. In burden tests, each variant is weighted differently according to AF, the impact on protein function, and conservation scores [Madsen and Browning, 2009; Han and Pan, 2010; Morris and Zeggini, 2010; Price et al., 2010; Zawistowski et al., 2010]. The most burden tests assume that all variants of interest contribute to the phenotype in the same direction, whereas kernel-based tests such as sequence kernel association test (SKAT) [Wu et al., 2011] and C-alpha [Neale et al., 2011] combine both protective and deleterious effects as well as variant–variant interactions. In our proposed pipeline, we implemented a burden test with the equal weights for all variants selected for specific criteria. The genes with compound heterozygous variants where each variant met user-defined filtering criteria can be prioritized using *4.Analyze::Genes* (Fig. 2). gNOME aggregates interesting variants by counting the number of variants in each gene, and perform a gene-level association test. Without loss of generality, we can assume that the M variants are linked to the total of P genes and the membership of each variant to genes can be represented as

$$P\text{-by-}M \text{ matrix } \mathbf{G} = \{g_{k,i}\}_{1\le k\le P, 1\le i\le M},$$

where

$$g_{k,i} = \begin{cases} 1 & i\text{–th variant is linked to } k\text{–th gene} \\ 0 & \text{otherwise} \end{cases}$$

The matrix multiplication, $\mathbf{B} = \mathbf{GV} = \{b_{k,j}\}_{1\le k\le P, 1\le j\le N}$, gives us the number of variants in a gene for each individual (Fig. 3B). The

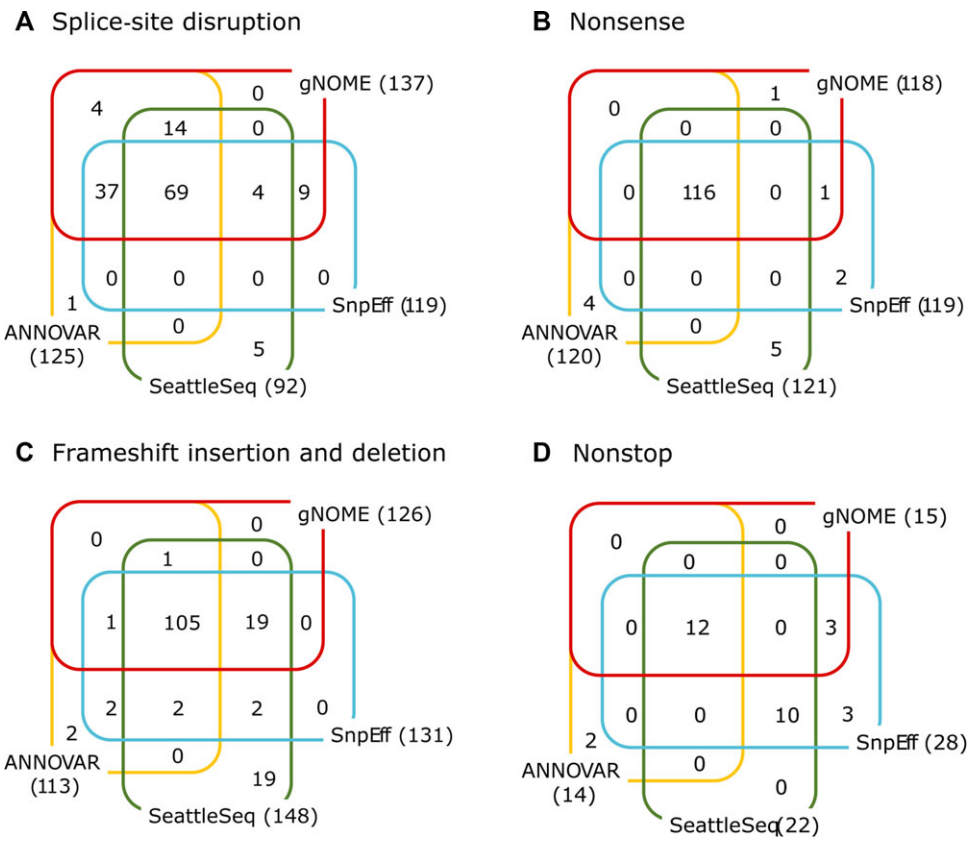

Figure 4. Comparison of annotation results from four software tools. For SSD (**A**), nonsense (**B**), frameshift insertion and deletion (**C**), and nonstop (**D**) variants, we compared the annotation results from four different software tools by comparing genomic coordinate, alternative allele, and reported functional impact for each variant. The numbers next to tool names represent the total number of annotated variants in that category, and four-way Venn diagrams show the concordant and discordant annotation results. Overall, the annotation results are comparable to each other between tools; however, SSD has the most discordant results (**A**). ANNOVAR reports as frameshift even if such variants are found in canonical splice sites; however, gNOME and SeattleSeq report them as SSD variants. Supp. Table S3 lists the details on the discordant results.

genes with a significantly different number of variants between two groups are ranked by a row-wise two group comparison test of B. If a gene is hypervariable [Kohane et al., 2012], it would consistently have nonzero value ($b_{k,j} > 0$ for all j) and be less significant in group comparison test (marked with a red x in Fig. 3B).

A gene-level association test can be further expanded to a set of genes that are functionally related or physically interacting with each other. Even if an individual gene showed weak association in genewise test, those genes can collectively contribute to a specific phenotype. Our analysis pipeline provides several options for gene set-level association tests. First, one can perform a gene-set enrichment test for the genes with interesting variants in each case ("Enriched Gene Sets in group A" function in Fig. 2B). Second, for each individual in G_0 and G_1, we can prioritize the gene sets that are more frequently observed as significantly enriched among the individuals in G_0. We construct a contingency table $T = (t_{11}, t_{12}, t_{21}, t_{22})$ for each gene set per individual with the number of member genes with the selected variants (t_{11}), member genes without the selected variants (t_{12}), nonmember genes with the selected variants (t_{21}), and nonmember genes without the selected variants (t_{22}), where the genes in a given gene set are defined as member genes. The relationship between genes and gene sets is defined as

$$Q\text{-by-}P \text{ matrix } \mathbf{S} = \{s_{l,k}\}_{1\le l\le Q, 1\le k\le P}$$

where

$$s_{l,k} = \begin{cases} 1 & k\text{–th gene belongs to } l\text{–th gene set} \\ 0 & \text{otherwise} \end{cases}$$

Then, we can collapse the values in matrix **B** into 0 or 1 to indicate whether an individual has interesting variants in the gene: $\dot{B} = \{\dot{b}_{k,j}\}_{1\le k\le P, 1\le j\le N}$ where $\dot{b}_{k,j} = 1$ if and only if $b_{k,j} > 0$. The four values in T are defined as:

$$\begin{aligned} t_{11} &= \sum_k s_{l,k}\dot{b}_{k,j} \\ t_{12} &= \sum_k s_{l,k}(1-\dot{b}_{k,j}) \\ t_{21} &= \sum_k (1-s_{l,k})\dot{b}_{k,j} \\ t_{22} &= \mathcal{G} - t_{11} - t_{12} - t_{21} \end{aligned} \quad (1)$$

with the total number of genes ($\mathcal{G}$). The enrichment scores from the contingency tables constitute the Q-by-N matrix $\mathbf{C} = \{c_{l,j}\}_{1\le l\le Q, 1\le j\le N}$ ($c_{l,j}$= the enrichment score of lth gene sets in jth individual) (Fig. 4C). The rowwise two group comparison tests on C are performed either by using a nonparametric test with odd ratios or hypergeometric P values or by comparing the proportion of individuals that passed a user-defined statistical threshold. Using

the latter option, hypervariable gene sets with interesting variants can be identified (marked with a red x in Fig. 3C).

Evaluation Datasets

We used three datasets to evaluate gNOME's performance and to compare with other programs. The first dataset consists of 97 patients with transitional cell carcinoma (TCC) [Gui et al., 2011]. For each patient, paired tumor–blood samples were sequenced using WES. We downloaded the raw sequence files from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) [Leinonen et al., 2011] (accession number SRA038181), and aligned them to the reference human genome (hg19) using Burrows–Wheeler Aligner (BWA) 0.6.2 [Li and Durbin, 2009]. The potential variants were called by the Genome Analysis Toolkit (GATK) 2.3–4 following the guidelines in Best Practice v4 from GATK's Website (http://www.broadinstitute.org/gatk). The second WES dataset consisted of tumor–blood pairs from two patients with uveal melanoma [Harbour et al., 2010]. The raw sequences (accession number SRA062369) were processed with the same alignment and variant calling procedure as described in the original paper. This dataset was used to compare the accuracy and performance of gNOME's annotation procedure with the other software tools. Finally, we used all variants from a single individual (NA12889) in 1KGP as well as the entire Phase I integrated call set from the 1KGP to evaluate the scalability with data size.

Comparison with the Existing Software Tools

We compared the performance and annotation results of gNOME with published software tools. The list of software tools included ANNOVAR (http://www.openbioinformatics.org/annovar, latest downloaded on May 27, 2013) [Wang et al., 2010], SnpEff (http://snpeff.sourceforge.net, version 3.2) [Cingolani et al., 2012], and SeattleSeq annotation server (http://snp.gs.washington.edu/SeattleSeqAnnotation137, version 137). We downloaded the latest version and installed on the same workstation if source codes were available, or uploaded the same VCF file to the Web server. We used the same RefSeq gene model (downloaded on May 20, 2013 from UCSC Table Browser) for ANNOVAR, SnpEff, and gNOME. However, SeattleSeq annotation server was running with a slightly different version of RefSeq model at the time of experiment. We measured the wall clock time to complete annotation for functional consequences of variants using a WGS variant file and 39.7 million variants from 1KGP.

Results

Comparison of Performance and Annotation Results from the Other Software Tools

The annotation speed of gNOME was compared with those of ANNOVAR [Wang et al., 2010] and SnpEff [Cingolani et al., 2012] using variants from a single genome (NA12889) and the integrated variants from 1,092 individuals of the 1KGP [Genomes Project et al., 2012]. The annotation for functional consequences of variants based on the RefSeq gene model [Pruitt et al., 2005] was performed 10 times for each tool. All three tools finished the annotation procedure in a reasonably short amount of time. For a single genome, SnpEff showed the best performance (192.2 ± 4.47 sec; mean ± standard deviation); however, with 39.7 million variants of 1KGP, gNOME completed the annotation in 484.2 ± 0.92 sec compared with 3,263.0 ± 65.94 and 1,606.8 ± 43.28 sec for ANNOVAR and SnpEff (Table 1). The annotation engine of gNOME—gSearch [Song et al., 2012]—was optimized to handle a larger dataset, whereas the processing time for ANNOVAR and SnpEff increased linearly with the number of variants. It should also be noted that genomes uploaded to gNOME's Web interface will be annotated using four popular gene models, taking four to five times longer than reported in Table 1; at most 30 min for a single genome and 50 min for 1,092 genomes from 1KGP.

Table 1. Comparison of Annotation Performance

Variant file	ANNOVAR	gNOME	SnpEff
A single genome[a]	395.2 (4.78)	351.4 (3.53)	192.2 (4.47)
1,092 genomes[b]	3,263.0 (65.94)	484.2 (0.92)	1,606.8 (43.28)

[a] All variants found in one individual (NA12889) from CEPH/Utah family (4,581,619 variants).
[b] All variants in the Phase 1 integrated call set from the 1KGP (39,706,715 variants).
We used the variant files from a WGS and all concatenated variants of 1,092 individuals from the 1KGP. For each tool, we repeated the annotation procedure 10 times to calculate the average time required to complete the annotation of variant consequences using RefSeq gene model (the standard deviations are shown in parentheses). All three tools perform reasonably quickly. The annotation time of ANNOVAR and SnpEff linearly increases with the number of variants; however, only 37.8% more processing time is required to complete the annotation of 8.7× larger variant file using gNOME.

For the comparison of annotated functional impacts of variants, we included SeattleSeq, a Web-based annotation server, as well as ANNOVAR and SnpEff. For this comparison, we concatenated four variant files from two tumor–blood pairs in the uveal melanoma dataset using the same gene model—RefSeq gene model (downloaded on May 20, 2013 from UCSC Table Browser)—for most tools to preclude discordant annotations due to different gene models (see *Materials and Methods*). Since each tool uses a different set of terminology to describe functional impacts, we mapped the various description terminologies used by separate tools, as listed in Supp. Table S2. Overall, the annotation results from different tools were similar. However, there were categories of functional impacts that were not reported by ANNOVAR and SeattleSeq (Table 2). For instance, a loss of start codon—categorized as misstart in gNOME—was not reported in ANNOVAR and SeattleSeq, and in-frame insertion and deletions were not listed in SeattleSeq.

We found differences in the sets of variants for each functional impact category across four tools (Fig. 4 shows Venn diagrams for splice-site disrupting [SSD], frameshift, nonsense, and nonstop variants). The differences can be explained partly by the differences in RefSeq versions between SeattleSeq and the other three tools (SeattleSeq used genes in September 2012 version from National Center for Biotechnology Information [NCBI], whereas we used genes in May 2013 version from UCSC Table Browser), by annotation errors in all programs, and by a discrepancy between tools in describing the same variants. Of 92–137 SSD variants that were found by four tools, 69 were discovered in common. SeattleSeq missed 37 SSD variants that were found by three other tools, most of which were suspected from differences in gene models. ANNOVAR and gNOME took different approaches in describing the SSD variants due to indels. In gNOME, SSD has priority over other functional consequences if indels were found in canonical splice sites, and vice versa in ANNOVAR. In 21 cases of ambiguous SSD variants, for example, insertions at exon–intron junctions, gNOME classified all ambiguous cases as SSD, whereas other programs report only part of them. The proportion of discordant annotations among programs was the smallest for nonsense variants. Among the nonsense variants that were not reported by gNOME, we found

Table 2. Comparison of Annotation Results from Various Programs

Category	Variant consequences	ANNOVAR	gNOME	SnpEff	SeattleSeq
Single nucleotide changes in coding sequence	Disrupt	125	137	119	92
	Missense	11,687	11,729	11,825	12,042
	Misstart	NA	22	22	NA
	Nonsense	120	118	119	121
	Nonstop	14	15	28	22
	Synonymous	12,970	13,051	13,105	13,206
Short insertions and deletions in coding sequence	Frameshift	113	126	131	148
	In-frame insertion	78	78	79	NA
	In-frame deletion	127	127	126	NA
Variants outside of coding sequence	5′-UTR	3,677	3,941	3,941	3,879
	3′-UTR	8,659	9,008	9,010	8,931
	Intron	262,678	267,803	267,867	253,821
	Intergenic	255,085	265,926	265,941	264,470

The functional consequences for all variants found from two tumor–blood pairs of WES of uveal melanoma samples are compared (see *Materials and Methods*). The functional impacts are based on RefSeq gene definitions, and description terms are compared according to the Supp. Table S2.
NA, not available.

four erroneous annotations (two from SnpEff, one from ANNOVAR and SeattleSeq each). There were three frameshift indels annotated as nonsense by ANNOVAR but as frameshift by the others. The four nonsense variants found only by SeattleSeq were due to an outdated gene model. For nonstop variants, 11 out of 15 discordant annotations between gNOME and other tools resulted from the annotation for possible selenocysteine, which was recognized only in gNOME. The other four were annotation errors from SnpEff and ANNOVAR. Discordant annotations for frameshift variants were more complex (see Supp. Table S3 for details); however, a majority of discordant annotations was due to the different approaches in classifying functional impacts. The details on discordant annotations between programs are summarized in Supp. Table S3. The functional impact of each variant must be evaluated by experts; however, there is a tremendous need for a standard method to describe variants with complex consequences.

Finding Somatic Mutations in Uveal Melanoma

To ensure the annotation accuracy of gNOME, we analyzed a published WES dataset from patients with uveal melanoma (MIM #155720). Harbour et al. (2010) sequenced two cases of matched tumor and peripheral blood samples using WES to find tumor-specific somatic mutations on chromosome 3. They found an inactivating mutation in each tumor sample on *BAP1*. One patient (MM56) had a nonsense mutation (p.W196X), whereas the other (MM70) had an 11-bp deletion (p.Q322fsX100) on the same gene. We processed the downloaded data as described in *Materials and Methods*, and uploaded the variant files from tumor tissues as cases and those from blood as controls to our pipeline and analyzed as depicted in Figure 2. The variant-level analysis of gNOME for MM56 revealed 670 possible somatic mutation candidates in protein-coding regions including p.W196X in *BAP1*. After filtering out nonrare (AF > 1% in European population, the same ethnic group as the patients) or synonymous variants, 171 candidate variants were found. Of these, four nonsynonymous variants—three missense and one nonsense (p.W196X in *BAP1*)—were found on chromosome 3. Similarly, in MM70, the 11-bp frameshift deletion in *BAP1* was the only high-impact tumor-specific nonsynonymous variant on chromosome 3. Interestingly, gene-level analysis of gNOME found 11 genes—*CEP89, FAM135A, GNAQ, HECTD4, HEXA, KCTD20, RAD17, TAS2R31, THBS3, TTLL1*, and *WSCD1*—that contain possible somatic mutations in both MM56 and MM70. Of these genes, frequent somatic mutations in *GNAQ* from patients with uveal melanoma were previously reported [Van Raamsdonk et al., 2009], and the increased protein expression of HEXA was found in metastatic uveal melanoma [Linge et al., 2012].

Discovering Somatic Mutations and Enriched Pathways in TCC

Gui et al. (2011) sequenced nine tumor–blood pairs from the patients with TCC (MIM #109800), and found 465 predicted somatic mutations. Several genes such as *ARID1A* and *CREBBP* had different somatic mutations in tumor samples of different patients. Additionally, tumor–blood sample pairs from 88 patients with TCC—37 nonmuscle invasive (NMI) TCC and 51 muscle invasive (MI) TCC—were sequenced to find frequently mutated genes in MI TCC and NMI TCC. The downloaded data were processed using hg19 (see *Materials and Methods*). First, all nine tumor samples were uploaded onto gNOME as cases, and nine blood samples as controls. We compared the somatic mutation candidates that were identified in gNOME using *4.Analyze::Variants* ($N_{\text{group A genomes}} \geq 1$ and $N_{\text{group B genomes}} \leq 0$) with the list from the original paper (Supplementary Table 3 in the original paper). Of the 208 somatic substitutions that were validated by genotyping or Sanger sequencing, 195 variants were accurately called and annotated using gNOME, except for 13 variants that were not called by our variant calling approach with hg19. One variant (g.chr1:22186113T>G on hg19) was found in the blood sample of B17 individual, but not in B17's tumor sample. Next, we tested whether the somatic mutation candidates were enriched for any gene sets using 97 tumor–blood pairs. For this analysis, we selected the LoF variants with AF ≤ 1% in Asian population, and *P* value threshold ≤0.05, $N_{\text{group A genomes}} \geq 1$, and $N_{\text{group B genomes}} \leq 97$ in Statistical test parameters in *4.Analyze::Genes*. Ninety six genes with the variants that met the criteria were enriched for the cancer-related KEGG pathways such as cell cycle (adjusted *P* value 0.0014), prostate cancer (adjusted *P* value 0.0017), pathways in cancer (adjusted *P* value 0.010), arrythmogenic right ventricular cardiomyopathy (adjusted *P* value 0.008), and bladder cancer (adjusted *P* value 0.013) (Supp. Table S4).

Discussion

gNOME (http://gnome.tchlab.org) enables users to interactively filter a large number of variants down to a small number of disease/phenotype-linked variants dynamically reported at three different levels—variants, genes, and gene sets—simultaneously.

Additionally, gNOME applies nonparametric statistical tests to variant- and gene-level counts of filtered variants between two groups, as well as to gene set enrichment analysis for biological pathways and known disease-linked genes. With the Web interface for interactive annotation-based filtering and statistical tests, we demonstrated our streamlined analysis procedure using two tumor–blood paired datasets—uveal melanoma and TCC. All validated genomic variants in the uveal melanoma dataset and 93% of 208 validated variants in the TCC dataset were accurately identified with gNOME, and new candidate variants and genes were found. Additionally, the cancer-related pathways were found to be enriched with LoF variants that were exclusively found in tumor samples.

Precise identification of genomic variants with high accuracy and the transparent annotation and filtering procedure are essential for clinical sequencing [Gargis et al., 2012]. A combination of the version control system for annotation database and graphical user interface allows gNOME to successfully reproduce results. Furthermore, gNOME can be installed locally as a stand-alone analysis pipeline with a secure storage device behind a firewall since the genome sequence information is confidential. Communications with other servers are not necessary once variant files are transferred. Moreover, enabling the other encryption and security features on MySQL database will make gNOME compatible with the Clinical Laboratory Improvement Amendments.

To further evaluate the annotation accuracy of our proposed pipeline compared with the other genome annotation software tools, we analyzed the same VCF file from the uveal melanoma dataset using ANNOVAR, SnpEff, and SeattleSeq. The results were similar in general, but the number of discordant results varied across functional categories. Aside from cases due to differences in gene models or programming errors, a majority of discordant cases came from variants whose functional consequences can be classified into multiple categories. For instance, the frameshift variants overlapping with canonical splice sites were reported either as frameshift in ANNOVAR or as SSD in gNOME and SeattleSeq. The ontology for describing the functional consequences of sequence variants and a consensus approach to describe the variants with multiple functional consequences should reduce the number of discordant annotations between tools [Eilbeck et al., 2005].

There are a few limitations of gNOME. First, due to the licensing issue, known disease-associated variants in the HGMD were not integrated into gNOME, although the integration itself was straightforward. Second, only a limited set of statistical tests were implemented. Adding diverse genetic burden tests such as the methods implemented in PLINK-SEQ [Neale et al., 2011], Efficient and Parallelizable Association Container Toolbox (EPACTS) [Kang et al., 2010], and SKAT [Wu et al., 2011] will improve the flexibility of gNOME. Finally, the scalability of gNOME would be much improved if run on a computing cloud. The current Web-based version can process medium-sized datasets of up to 1,000 individuals, and a local stand-alone version can be set up to accommodate with a larger dataset or to preserve the confidentiality. However, to analyze a larger dataset (tens of thousands individuals), the gNOME pipeline must be run on a computing cloud. A few WES/WGS tools do support cloud environment as backend: VAT [Habegger et al., 2012] for variant annotation and visualization, Crossbow [Langmead et al., 2009] for variant calling, SIMPLEX [Fischer et al., 2012] for WES analysis, and Galaxy [Goecks et al., 2010] and Taverna [Hull et al., 2006] as general workflow framework. However, these are either difficult to use or do not cover all streamlined process provided by gNOME.

To summarize, we have developed a downstream analysis pipeline for WGS/WES datasets that can perform accurate and reproducible annotation with a graphical user interface for annotation-based filtering. A group comparison is one of the unique features that help to reduce possible false-positive findings. We have provided a population scale WGS dataset as a part of the pipeline, which enables users to identify variants specific to cases compared with ethnicity-matched generally healthy population. With these strengths, gNOME will be of use to the biomedical research community.

Acknowledgments

Dr. Kohane is a member of the scientific advisory board of the SynapDx (Lexington, MA).

Disclosure statement: The authors declare no conflict of interest.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, et al. 2010. Clinical assessment incorporating a personal genome. Lancet 375:1525–1535.

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.

Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. 2010. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics 26:2924–2926.

Calvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, Burtt NP, Rivas M, Guiducci C, Bruno DL, Goldberger OA, Redman MC, Wiltshire E, et al. 2010. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet 42:851–858.

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

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096–19101.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. Genome Res 19:1553–1561.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425.

Consortium EP. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9:e1001046.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12:628–640.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025.

Drmanac R. 2011. The advent of personal genome sequencing. Genet Med 13:188–190.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6:R44.

Fischer M, Snajder R, Pabinger S, Dander A, Schossig A, Zschocke J, Trajanoski Z, Stocker G. 2012. SIMPLEX: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data. PLoS ONE 7:e41948.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220.

Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, et al. 2012.

Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 30:1033–1036.
Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP, Zhu M, Singh A, et al. 2011. SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics 27:1998–2000.
Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.
Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86.
Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, Sunyaev S. 2013. Sequencing studies in human genetics: design and interpretation. Nat Rev Genet 14:460–470.
Gui Y, Guo G, Huang Y, Hu X, Tang A, Gao S, Wu R, Chen C, Li X, Zhou L, He M, Li Z, et al. 2011. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nat Genet 43:875–878.
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.
Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70:42–54.
Harbour JW, Onken MD, Roberson ED, Duan S, Cao L, Worley LA, Council ML, Matatall KA, Helms C, Bowcock AM. 2010. Frequent mutation of BAP1 in metastasizing uveal melanomas. Science 330:1410–1413.
Hindorff L, MacArthur J, Morales J, Junkins H, Hall P, Klemm A, Manolio T. 2012. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies.
Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, et al. 2010. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 42:483–485.
Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A, Stefansdottir H, Gretarsdottir S, et al. 2011. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 43:316–320.
Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. Bioinformatics 22:1036–1046.
Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, et al. 2002. The Ensembl genome database project. Nucleic Acids Res 30:38–41.
Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. 2006. Taverna: a tool for building and running workflows of services. Nucleic Acids Res 34(Web Server issue):W729–W732.
Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, et al. 2012. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40(Database issue):D306–D312.
Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354.
Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, Degenhardt JD, Brisbin A, Sheth V, Chen R, McLaughlin SF, Peckham HE, et al. 2012. Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. Am J Hum Genet 91:660–671.
Klassen T, Davis C, Goldman A, Burgess D, Chen T, Wheeler D, McPherson J, Bourquin T, Lewis L, Villasana D, Morgan M, Muzny D, et al. 2011. Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. Cell 145:1036–1048.
Kohane IS, Hsing M, Kong SW. 2012. Taxonomizing, sizing, and overcoming the incidentalome. Genet Med 14:399–404.
Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39(Database issue):D152–D157.
Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081.
Lalonde E, Albrecht S, Ha KC, Jacob K, Bolduc N, Polychronakos C, Dechelotte P, Majewski J, Jabado N. 2010. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum Mutat 31:918–923.
Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, Snyder M. 2012. Detecting and annotating genetic variations using the HugeSeq pipeline. Nat Biotechnol 30:226–229.
Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. 2009. Searching for SNPs with cloud computing. Genome Biol 10:R134.
Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. 2011. The sequence read archive. Nucleic Acids Res 39(Database issue):D19–D21.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, Neale BM, Kirby A, Ruderfer DM, Fromer M, Lek M, Liu L, et al. 2013. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron 77:235–242.
Linge A, Kennedy S, O'Flynn D, Beatty S, Moriarty P, Henry M, Clynes M, Larkin A, Meleady P. 2012. Differential expression of fourteen proteins between uveal melanoma from patients who subsequently developed distant metastases versus those who did not. Invest Ophthalmol Vis Sci 53:4634–4643.
Liu X, Jian X, Boerwinkle E. 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32: 894–899.
Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, et al. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362:1181–1191.
MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828.
Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5:e1000384.
McClellan J, King MC. 2010. Genetic heterogeneity in human disease. Cell 141: 210–217.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303.
McKusick VA. 2007. Mendelian inheritance in man and its online version, OMIM. Am J Hum Genet 80:588–604.
Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696.
Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM. 2003. Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc:460–464.
Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193.
Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. PLoS Genet 7:e1001322.
Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, et al. 2010a. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42:790–793.
Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2010b. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42:30–35.
Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. 2013. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform Advance Access published Jan 21, 2013, doi:10.1093/bib/bbs086.
Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, Heinzen EL, Need AC, et al. 2010. The characterization of twenty sequenced human genomes. PLoS Genet 6:e1001111.
Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838.
Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, et al. 2009. The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19:1316–1323.
Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33(Database issue):D501–D504.
Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K. 2010. A standard variation file format for human genome sequences. Genome Biol 11:R88.
Riggs ER, Wain KE, Riethmaier D, Savage M, Smith-Packard B, Kaminsky EB, Rehm HL, Martin CL, Ledbetter DH, Faucett WA. 2013. Towards a Universal Clinical Genomics Database: the 2012 International Standards for Cytogenomic Arrays Consortium Meeting. Hum Mutat 34:915–919.
Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, et al. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43:1066–1073.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636–639.

Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, Cotsapas C, Daly MJ. 2011. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet 7: e1001273.

San Lucas FA, Wang G, Scheet P, Peng B. 2012. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics 28:421–422.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7:575–576.

Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0. Available at: http://www.repeatmasker.org.

Song T, Hwang KB, Hsing M, Lee K, Bohn J, Kong SW. 2012. gSearch: a fast and flexible general search tool for whole-genome sequencing. Bioinformatics 28: 2176–2177.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102:15545–15550.

Tabor HK, Berkman BE, Hull SC, Bamshad MJ. 2011. Genomics really gets personal: how exome and whole genome sequencing challenge the ethical framework of human genetics research. Am J Med Genet A 155A:2916–2924.

Van Raamsdonk CD, Bezrookove V, Green G, Bauer J, Gaugler L, O'Brien JM, Simpson EM, Barsh GS, Bastian BC. 2009. Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. Nature 457:599–602.

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

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82–93.

Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG. 2011. A probabilistic disease-gene finder for personal genomes. Genome Res 21:1529–1542.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87:604–617.