PLOS ONE

RESEARCH ARTICLE

Scripps Genome ADVISER: Annotation and Distributed Variant Interpretation SERver

Phillip H. Pham[5‡], William J. Shipman[1,2‡], Galina A. Erikson[1,2‡], Nicholas J. Schork[1,2,4,5], Ali Torkamani[1,2,3,5]*

1 Scripps Health, La Jolla, CA 92037, United States of America, **2** The Scripps Translational Science Institute, La Jolla, CA 92037, United States of America, **3** The Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, United States of America, **4** The Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, United States of America, **5** Cypher Genomics, Inc., La Jolla, CA 92037, United States of America

‡ These authors contributed equally to this work.
* atorkama@scripps.edu

Citation: Pham PH, Shipman WJ, Erikson GA, Schork NJ, Torkamani A (2015) Scripps Genome ADVISER: Annotation and Distributed Variant Interpretation SERver. PLoS ONE 10(2): e0116815. doi:10.1371/journal.pone.0116815

Academic Editor: Nancy Lan Guo, West Virginia University, UNITED STATES

Received: May 13, 2014

Accepted: December 1, 2014

Published: February 23, 2015

Data Availability Statement: The methods described herein are available at (genomics.scripps.edu/ADVISER).

Funding: This work was supported by the NHGRI Genome Sequencing Informatics Tools (GS-IT) Program via grant National Institute of Health U01 HG006476 to Ali Torkamani. Further information available at: (http://iseqtools.org). Further support is provided by Scripps Genomic Medicine, a National Institute of Health - National Center for Advancing Translational Sciences Clinical and Translational Science Award (CTSA; 5 UL1 RR025774) to STSI. The funders had no role in study design, data

Abstract

Interpretation of human genomes is a major challenge. We present the Scripps Genome ADVISER (SG-ADVISER) suite, which aims to fill the gap between data generation and genome interpretation by performing holistic, in-depth, annotations and functional predictions on all variant types and effects. The SG-ADVISER suite includes a de-identification tool, a variant annotation web-server, and a user interface for inheritance and annotation-based filtration. SG-ADVISER allows users with no bioinformatics expertise to manipulate large volumes of variant data with ease – without the need to download large reference databases, install software, or use a command line interface. SG-ADVISER is freely available at genomics.scripps.edu/ADVISER.

Introduction

The availability of high-throughput DNA sequencing technologies has enabled nearly comprehensive investigations into the number and types of sequence variants possessed by individuals in different populations. For example, not only is it now possible to sequence a large number of genes in hundreds if not thousands of people, but it is also possible to sequence entire individual human genomes in the pursuit of inherited disease-causing variants or somatic cancer-causing variants [1–5]. The day where whole genome sequencing is a relatively routine procedure lies within the near future, as high-throughput sequencing costs and efficiency continue to improve at a blistering pace.

One particularly vexing problem that has accompanied the development and application of high-throughput sequencing is making sense of the millions of variants identified per genome. For example, recent successes at identifying variants associated with rare disease have generally required large bioinformatics teams—restricting the effective implementation of whole genome sequence-based clinical and research endeavors to large institutions and/or genome centers [2,3]. Similarly, while the GWAS strategy could potentially identify tag-SNPs explaining up to

collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Ali Torkamani has read the journal's policy and the authors of this manuscript have the following competing interests: NJS and AT are founders and stock-holders of Cypher Genomics. Ali Torkamani is a PLOS ONE Editorial Board member. This does not alter the authors' adherence to PLOS ONE Editorial policies and criteria.

half the heritability of common diseases [6,7], sequence-based methods will likely be necessary for the identification of rare variants predisposing to common diseases where variable penetrance, allelic and locus heterogeneity, epistasis, gene-gene interactions, and regulatory variation play a more important yet elusive role. The sensitivity of set-based rare variant analyses to the inclusion of non-causal and exclusion of causal variants indicates a clear role for automated set generation and variant prioritization in these analyses [8,9]. Finally, the recent unveiling of the role of ultra-rare and/or *de novo* variants in the etiology of human disease, especially in idiopathic disease and neuropsychiatric disorders—or the vast number of somatic mutations that can perturb tumor suppressor function in cancer—suggests that reliance upon variants statistically associated with disease for molecular diagnosis at an individual level will be suboptimal in many instances [10–12]. The issue of interpretation of variants of unknown significance can only be expected to worsen as humans continue to postpone reproduction to more advanced age and the number of recently derived deleterious variants continues to explode [13]. Analysis of rare variants in these various scenarios is potentially addressable through holistic and accurate variant annotation.

A clear need for functional annotation has been recognized since investigators began searching for causal variants linked to GWAS tag-SNPs. Early tools developed for this purpose, built under the assumption that common variants would explain disease predisposition, are limited to databases which provide only information on known SNPs [14–17]. Novel and/or rare, *de novo*, and indel variants, are not accessible within this framework. More recently developed tools sensitive to the importance of undiscovered or more complex variants simply annotate variants based on the known genomic elements they reside and/or restrict functional predictions to pre-computed nonsynonymous variant functional predictions [18–24]. While these tools are immensely useful in their own right, none are capable of producing *predictions* for the near infinite possible variants generated in a sequencing project. We would like to emphasize the distinction between algorithmic prediction rather than simple determination of residence within genomic elements. For example, while missense SNP impact predictions via e.g., Polyphen [25] or SIFT [26], can be precomputed with relative ease [27], it technically impossible to precalculate the algorithmically predicted impact of all possible inframe indels on protein function or on transcription factor binding sites. A powerful webserver interface is required to enable this sort of *de novo* calculation. There is a clear need for a more holistic and integrated annotation tool to both annotate and predict the functional effects of the numerous variant classes produced by whole-genome sequencing projects and allow for the processing of those predictions alongside genotype data. The tool presented here, Scripps Genome ADVISER, aims to fill this role in a manner accessible to research endeavors at all levels of bioinformatics sophistication.

Methods

Overview

SG-ADVISER is a multi-component system (Fig. 1) including: 1) a privacy tool for markedly reducing or eliminating the usefulness of genomic data should it be intercepted in transit to the webserver, 2) a webserver that accepts and returns genomic data and annotations, 3) a variant validation and correction system that accepts and converts various variant file formats, and validates and/or corrects the accuracy of variant information against the reference genome with informative error and corrections reporting and the option to immediately resubmit valid and corrected variants, 4) a high-performance computing system that utilizes both pre-computation databases and parallel computations to produce variant annotations rapidly, and 5) a local client graphical user interface that allows loading of genotype information and the filtration of

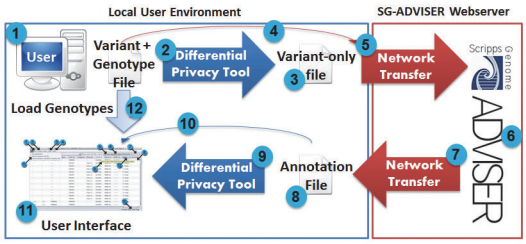

Fig 1. SG-ADVISER Suite and Workflow. This figure depicts the workflow for variant annotation and analysis. Beginning with a user with a file containing variant and genotype information (1), the user can optionally use the privacy tool (2) to generate a variant-only file with (3) with genotypes removed and clinically relevant variants implanted. This file, or the original variant file if desired, is then uploaded to the SG-ADVISER webserver (4,5). The SG-ADVISER webserver performs variant validation and annotation (6). If the file contains errors, the validated variants may be automatically resubmitted to the webserver. At the completion of annotation, annotation information is downloaded from the webserver down to the users local environment (7). The resultant annotation file (8) can then be run in reverse through the privacy tool (9) to remove implanted clinical relevant variants. The resultant file or the original annotation file is then loaded into the user interface (10,11). Finally and optionally, genotype information from the original variant file can be loaded into the user interface (12). The genomic data is ready for downstream analysis.

doi:10.1371/journal.pone.0116815.g001

variants based upon annotations and comparisons of multiple genomes using custom as well as predefined variant filtration strategies. The overall goal is to provide near comprehensive variant annotation without the burden of complex software or intense client-side compute capabilities, while simultaneously maintaining the privacy of user data and avoiding over-simplification of the annotations themselves.

Computational Infrastructure

Annotation proceeds in highly parallel fashion and includes classes of variant annotations that are entirely independent of one another, serially dependent annotations whose execution are dependent upon the completion and status of prior annotations, and synthetic annotations that generate new information through the combination of multiple annotation outputs. In contrast to existing tools which rely upon independent lookup tables, SG-ADVISER can produce a virtually infinite range of different annotation outputs depending upon the nature of the submitted variants. These processes are computed de-novo for any variants not previously observed in any genome, while annotations for previously observed variants are retrieved from a pre-annotation database. De-novo annotations are executed on a cluster of five Dell PowerEdge servers configured with 16 cores and eight terabytes of local disk space per server. Once completed, the new annotations are stored in the pre-annotation database for fast lookup of subsequent occurrences. The pre-annotation database is stored in MongoDB a NoSQL format database divided into separate collections by chromosome and indexed by 5 variant characteristics [start coordinate, end coordinate, variant type, reference allele, alternate allele]. The pre-annotation database currently contains over 220 million variants, consisting mostly of variants observed in the 1,000 Genomes Project, NHLBI exome sequencing project, dbSNP, and the Scripps Wellderly cohort [28–30]. For a detailed description of the computational processes underlying SG-ADVISER see *S1 Text*.

Performance

The computational infrastructure underlying SG-ADVISER allows rapid turn-over of a single whole genome variant files. We evaluated the performance of SG-ADVISER by annotating 10 whole Wellderly genomes, sequenced by Complete Genomics, and not previously annotated by

the SG-ADVISER system. At an average of 4,091,804 unfiltered variants per genome, the average time to completion was 110 ± 9 minutes. Exomes complete in considerably less time, at an average of 112,008 unfiltered variants per exome (10 exomes total), variant annotation completed in 24 ± 6 minutes per exome. One caveat to this performance is that only one variant file at a time can be processed, occupying the entire computational cluster, thus real turn around times can be dependent upon user traffic. However, as the number of variants annotated by SG-ADVISER increases, performance and turn-around-time is expected to improve further.

Data Input Formats

SG-ADVISER supports human genome annotation (hg19) only. SG-ADVISER accepts variant files in VCF, Complete Genomics, or plain tab-delimited file formats. For most accurate results variants should be submitted in 0-based coordinates with positive strand nucleotides reported. However, given our experience with the numerous variant input formats provided by early users, the SG-ADVISER validator will attempt to determine whether variants are 0-based or 1-based by evaluating matches to the reference genome, and convert coordinates appropriately. Moreover, the SG-ADVISER validator will attempt to correct reference-alternate allele swaps and/or nucleotides reported relative to the negative strand. While the presence of any incorrectly formatted variants will stop the automated annotation process—a descriptive error file is produced with the option of automatically resubmitting the corrected variants. The nature of the applied correction is provided in the final annotation output. Often times errors can be produced due to conflicting reference genome coordinates—rather than annotating this variants regardless of the reference match, an error is produced but annotation can be continued on all other verified variants. For more information on input formats and error reporting, see: http://genomics.scripps.edu/ADVISER/Input_Desc.jsp.

Data Output Format

Annotations are output in a tab separated file, where the first eight columns contain information about the submitted variant itself, and the rest of the columns are annotations produced by SG-ADVISER. Variants are presented as a single line per variant, yet complete annotations are produced for each individual transcript influenced by a variant, thus the format of each annotation column depends upon whether the annotation is relative to the gene or transcript it impacts or relative to the physical location of the variant. Any column containing annotations produced relative to a gene or transcripts are further subdivided by triple back slashes ("///"). Across annotation columns, "///" separated values correspond to one another—i.e. annotations in the same position relative to "///" separated values within a column influence the same transcript. Annotations not directly relevant to a particular transcript, for example transcription factor binding sites or the conservation of the position, are also "///" separated but that separation corresponds to a related column. For example, transcription factor binding sites influenced by a variant are "///" separated, and the calculation of the impact of the variant on binding of the "///" factor is presented in a separate "///" separated column. When an annotation is not applicable to a variant or transcript, a null value is represented by a "-" character, often in the format of the column. For example, a column where entries are formatted as "Value1 ~ Value2", if null, will receive a value of "- ~ -". This is required due to partially complete outputs, for example where only one of two output values is null. For a more thorough description of the annotation types and output format, see *S1 Text* and http://genomics.scripps.edu/ADVISER/Result_Desc.jsp

Security and Privacy

Data is encrypted during transfer to SG-ADVISER via a Secure Socket Layer (SSL 3.0) to a secure computational cluster maintained by The Scripps Research Institute. Thus, SG-ADVISER is compliant with the dbGaP Security Best Practices for controlled access data. Additionally, variant files uploaded to SG-ADVISER, as well as the resultant annotation file, are destroyed 30-days after variant file upload. To ensure confidentiality of valuable research data, we do not retain any information about the number, identity, or combinations of variants submitted by any user. As mentioned previously, annotations for each individual variant are stored in a pre-computed annotation database to improve the speed of future annotation, but no information beyond the physical location of the variant is retained—no association between variants in the pre-annotation database and the source or additional observations of the variant is preserved.

To facilitate and improve privacy further, a privacy tool is available for download at (http://genomics.scripps.edu/ADVISER/PrivacyTool.jsp). This tool will automatically strip genotype information from VCF files for users without the bioinformatics means to do so. Genotype information is not required for SG-ADVISER annotations—thus, removal of genotype information from uploaded files is suggested for sensitive genomes. However, because we suspect it is nearly impossible to de-identity a genome without information loss we have designed our privacy tool render any transit or server-side data interceptions uninformative [31]. The SG-ADVISER privacy tool will implant known clinically relevant variants into a variant file processed by the tool—making the identification of true vs. implanted clinically informative variants impossible. These variants can then be removed from the annotation file on the client side through the privacy tool by referencing the original VCF file. Thus, overall control of privacy remains in the hands of the end user with the original variant file, which need not ever be transferred to the SG-ADVISER web-server.

Results

Annotation Categories

At its core, SG-ADVISER is an automated computational system for producing known and predicted information about genetic variants—otherwise known as variant annotations. SG-ADVISER produces four major classes of variant annotations including: 1) residence within known or inferred genomic elements (e.g., exons, promoters, conserved elements, transcription factor binding sites, protein domains etc.); 2) annotation and prediction of the functional impact of a variant on genomic elements (prediction of impact on protein function, changes in transcription factor binding strength, splicing efficiency, microRNA binding, etc.); 3) annotation of molecular and biological processes which link variants across genes and/or genomic elements with one another, and 4) annotation of known or predicted population-based, clinical, and/or molecular characteristics of the gene or variant (e.g. population frequency, pharmacogenetic variants, disease associations, eQTLs etc.). Detailed descriptions of the 70+ specific annotations are provided in *S1 Text* and are available at (http://genomics.scripps.edu/ADVISER/Result_Desc.jsp). Key highlights include:

1. SG-ADVISER produces predictions for the functional impact of numerous variant types including; nonsynonymous variants, in-frame variants, truncating variants, splice site variants, microRNA binding site variants, transcription factor binding site variants, and the changes in microRNA targets induced by variants within microRNAs themselves. As previously emphasized—these annotations are not limited to classification as one of the above types of variants or residence within a motif or pre-defined site, but rather classification

plus a *prediction* as to whether the variant functionally impacts the genomic element they resides in.

2. Allele frequency information from the 1000 Genomes Project [28], NHLBI Exomes Project [29], and the Scripps Translational Science Institute Wellderly cohort are disseminated through SG-ADVISER. The Wellderly cohort is composed of individuals over the age of 80 with no common chronic conditions. 400+ individuals have been whole genome sequenced by Complete Genomics. Their allele frequencies are available through SG-ADVISER and will continue to be updated as the cohort continues to be sequenced.
3. Prior knowledge from the Human Gene Mutation Database (HGMD) [32], OMIM [33], Clinvar [34], the Genetic Association Database and GWAS Catalog [35,36], and the Catalogue of Somatic Mutations in Cancer [37] are provided. HGMD license information is required for the return of results from HGMD.
4. A synthesis of the above produces an American College of Medical Genetics-like (ACMG) ADVISER variant classification schema for known and predicted disease associated.

ADVISER Variant Classification

Two different modified American College of Medical Genetics (ACMG) variant classifications are produced, one based upon variants, or variants in genes known to be causally associated with a phenotype (ADVISER Clinical) and a second score which includes genes known to carry genetic variants that are statistically associated risk factors for the development of a disease (ADVISER Research). The ACMG scoring guidelines, with categories 1–6, are modified and expanded to include a 1∗, 2∗ and 4∗ category to provide more granularity to variant stratification, for example by down weighting reported pathogenic variants to category 1∗ based on allele frequency, or by allowing for stratification of variants of the same functional class (e.g. missense variants) across the ADVISER classes based on algorithmic predictions of pathogenicity rather than relegating all nonsynonymous variants unreported as pathogenic to variants of unknown significance [38]. Variants of category 1–2∗ are of most clinical relevance and category 6 contains common risk factors for disease. The details for ADVISER classes are defined in *S1 Text* and will be updated at (http://genomics.scripps.edu/ADVISER/ACMG.jsp). In brief, ADVISER category 1 variants are rare (<1% allele frequency) reported pathogenic variants. Category 1∗ includes more common (1–5% allele frequency) reported pathogenic variants—which tend to be either false positive reports or variants with incomplete penetrance or acting as modifiers. Category 2 contains rare variants in known disease genes, unreported as pathogenic, but predicted to impact gene function by either removing a splice site donor or acceptor, producing an amino acid substitution predicted to functionally impact the protein, or truncating the protein in a damaging manner. Category 2∗ includes rare truncating variants not predicted to damage protein function or uncommon truncating variants predicted to damage protein function. Allele frequencies are determined using the maximum allele frequency across our previously described reference populations.

The performance of the ADVISER classification schema was evaluated by categorizing a set of known high confidence nonsynonymous disease causative and neutral polymorphisms derived from the SWISS-PROT feature table [39]. 16,549 variants classified as disease causative (positive class) and 11,282 variants classified as neutral polymorphisms (negative class) in known disease causative genes were compiled in order to determine how well the SG-ADVISER classifications recapitulated the SWISS-PROT classifications at various SG-ADVISER class thresholds. Variants are considered true positive if a SWISS-PROT disease causative variant

Table 1. ADVISER Class Performance.

SG-ADVISER			Ingenuity Variant Analysis		
ADVISER Class	Sensitivity	Specificity	Metrics	Sensitivity	Specificity
1	83%	95%	Clinical Assessment	91%	83%
1–2	94%	86%	Clinical Assessment or Damaging by SIFT and Polyphen	92%	80%
1–3	98%	72%	Clinical Assessment or Damaging by SIFT or Polyphen	94%	72%

SG-ADVISER Classification performance. The Ingenuity Clinical Assessment considers variants classified as Known Pathogenic or Likely Pathogenic by Ingenuity. The default Ingenuity allele frequency threshold of 3% plus damaging predictions by SIFT and/or Polyphen were used to simulate less confident variant classification tranches in Ingenuity.

doi:10.1371/journal.pone.0116815.t001

achieves a threshold ADVISER class or better (as delineated in Table 1). True negative variants are SWISS-PROT neutral polymorphisms not achieving the threshold ADVISER class or better. For example, a variant classified as disease causative in SWISS-PROT and achieving an ADVISER class of 2 would be considered a false negative at the ADVISER class 1 threshold, true positive at the ADVISER class 1–2 threshold, and true positive at the ADVISER class 1–3 threshold. None of the previously described mentioned annotation tools [18–24] produce overall variant categorizations, therefore, performance was compared to a popular commercial platform for variant analysis, Ingenuity Variant Analysis, under its default settings. As can be seen in Table 1, the SG-ADVISER schema provides a superior and more useful way of capturing potential disease associated variants in a manner that is tuned to relevant use cases. That is, while SG-ADVISER's overall balanced accuracy (mean of sensitivity and specificity) is significantly but not dramatically superior, the specificity-sensitivity profile fulfills the actual requirements for practical use cases with dramatically superior specificity for the high confidence pathogenic categories and much more sensitive results for the lower confidence categories. In other words, when producing known or expected disease causative mutations (ADVISER class 1 and 2), SG-ADVISER's superior specificity reduces false positive disease associations in a context where false positive results are unacceptable—for example when performing predictive molecular diagnosis in the absence of a disease phenotype. Similarly, in a less conservative scenario (ADVISER class 3), SG-ADVISER's accuracy profile is more heavily weighted towards sensitivity, or inclusiveness of potential disease causative variants without unduly introducing false positive results—dramatically boosting negative predictive value. This accuracy profile is more useful in the case where a molecular diagnosis is to be made for an already present phenotype. Overall, the sensitivity-specificity profile of SG-ADVISER summary determinations are superior and address end-user needs in a more meaningful way by transitioning appropriately from conservative, high confidence, disease associations to comprehensive, high coverage, variant reports while maintaining superior accuracy overall.

Comparison to Other Methods

The ADVISER class performance evaluation described above considers only nonsynonymous variants, yet, the accuracy and comprehensiveness of SG-ADVISER annotations extend beyond to other important variant classes (Table 2). Truncating variants (nonsense or frameshift) are not evaluated any further by all available tools, yet it is known that the proximal and distal ends of genes are enriched in presumably neutral truncating variants [40]. Therefore, an algorithmic method to prioritize truncating variants based on the percentage of the conserved portion of the protein removed by the truncating variant after adjustment for alternative start

Table 2. Annotation Tool Comparison.

Tool	Local Install vs. Web Service	Encryption (web)	Graphical User Interface	Variant Filtration	Summary Pathogenicity Score	Truncating variant effect prediction	TFBS variant effect prediction	In-frame variant effect prediction	Splice-site variant effect prediction	microRNA variant effect prediction	Custom Annotation
ANNOVAR	Both	No	No	Yes	Reported only	No	No	No	No	No	Yes
AnnTools	Local	N/A	No	No	Reported only	No	No	No	No	No	Yes
VEP	Both	No	Yes	Yes	Reported only	No	Yes	No	No	No	Yes
SeattleSeq	Both	No	Yes	Yes	Reported only	No	No	No	No	No	No
SeqAnt	Web	No	Yes	Yes	No	No	No	No	No	No	No
SVA	Local	N/A	Yes	Yes	Reported only	No	No	No	No	No	Yes
SG-ADVISER	Web	Yes	Yes	Yes	Reported and Predicted	Yes	Yes	Yes	Yes	Yes	No
SnpEff	Local	N/A	No	Yes	Reported only	Yes	No	No	No	No	Yes
VARIANT	Web	No	Yes	Yes	Reported only	No	No	No	No	No	No

Prediction refers to a determination of functional effect of a specific variant type—not simply whether the variant belongs to that type. Splice-site variant effect prediction refers only to non-donor/acceptor nucleotide predictions.

doi:10.1371/journal.pone.0116815.t002

sites, is incorporated in SG-ADVISER [41]. Similarly, in-frame indels are often considered neutral or not stratified in anyway by other tools, yet important disease causative in-frame indels, such as F508del-CFTR—the most common cause of cystic fibrosis—are well established. SG-ADVISER annotations algorithmically prioritize inframe variants [42]. This approach is amenable to, and will be extended to, the annotation of phased combinations of variants as phased genomes gain in prominence [43]. Finally, approximately 40% of known disease causative variants in HGMD that influence splicing do not impact the conserved splice-donor and acceptor nucleotides—yet, there is no way to prioritize variants nearby intron-exon junctions in available annotation tools. SG-ADVISER annotations prioritize these variants appropriately [44]. These differences extend beyond coding variants to the prediction of changes in transcription factor binding site affinity, via calculation of the change in score for a mutated sequence using position-specific scoring matrices, microRNA binding strength, and altered targets due to variants in microRNAs themselves via recalculation of targets, albeit at lower confidence than the above described predictions. No previously described methods offer these predictions.

A number of tools described as variant annotation tools exist. Table 2 provides a comparison of SG-ADVISER functionalities with similar tools [45–47]. These tools generally predict variant effect by simply identifying overlap with pre-defined bins. Where tools, such as VEP, SnpEff, and ANNOVAR [18–20] incorporate algorithmic predictions, they do so through the inclusion of precalculation tables—thus practically limiting annotations to what can be precalculated (for example SIFT and Polyphen predictions), but allowing for more efficient expansion to other organisms. Similarly other tools, such as BEDTools [21], TREAT [22], SeqAnt [48], and AnnTools [23] simply allow for the overlap of variant coordinates with reference genes or intervals. Simple filters can be executed against the resultant annotations, but again, these tools rely upon the download of large pre-annotation databases and cannot be extended to more complex scenarios. Finally, tools such as GEMINI [49], Annotate-it [50], and VAR-MD [51] provide capabilities for more complex filtration strategies utilizing the basic annotations described previously. SG-ADVISER combines basic annotations, more complex annotations that require on the fly calculation, and complex filtration strategies enabled through the user interface.

User Interface

The SG-ADVISER user interface allows the user to load in an annotation results file, load in the genotypes for the annotated variants from the file submitted to the SG-ADVISER webserver (or from the original variant file passed through the SG-ADVISER privacy tool), and apply a wide variety of custom and pre-defined filters. The user interface is available at (http://genomics.scripps.edu/ADVISER/downloads.jsp)—and is built in Java to support cross operating system use. An annotated screenshot is displayed in Fig. 2. The user interface functionalities include: 1) basic sorting on any column, 2) basic filtration on any column, 3) advanced filters allowing specification of multiple columns linked by AND/OR operators, 4) capability to undo and redo actions, 5) application of custom pre-defined filters including inheritance based filters for family-based studies, 6) export of filtered files to be manipulated further by external tools, and 7) the calculation of summary statistics providing the number and rate of a wide variety of variant classes before or after the application of filters. The UI can load and process queries against a genome nearly in real-time: loading of exome data variant annotations for a trio, total of $\sim$145,000 variants takes $\sim$2 seconds, loading of the genotype data from a VCF file to be manipulated alongside annotations takes $\sim$4 seconds, and the execution of filters completes in less than 5 seconds for even the most complex queries. A standalone user interface has a few benefits: 1) whole genome variant filtration is impractical within a webserver, 2) on-the-fly

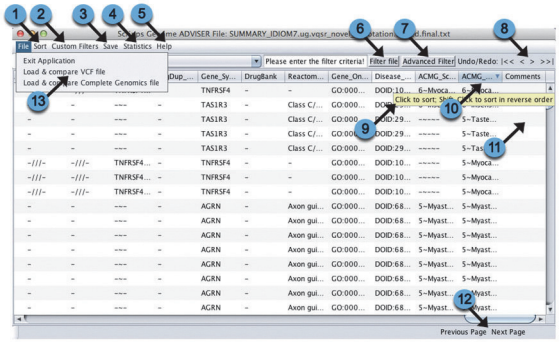

Fig 2. User Interface. The SG-ADVISER user interface provides a number of useful functionalities including: (1) sort the current view by any column; (2) 14 pre-defined custom filters, for example extraction of rare coding variants—for a list of custom filters see http://genomics.scripps.edu/ADVISER/downloads.jsp; (3) post-filtered files can be saved for manipulation outside of the UI; (4) calculation of variant type counts and frequency; (5) a help menu; (6) simple user-defined filter on a single column; (7) advanced multi-column user defined filtering; (8) the capability to move forward and backward through executed filters; (9) extensive tool tips; (10) sorting by clicking the column header; (11) the capability to add and save comments; (12) scrolling through the multiple pages of variants (1000 variants per page); (13) the ability to load in genotype data from the original variant file.

doi:10.1371/journal.pone.0116815.g002

computations such as variant summary statistics can be performed after the execution of customized filters, 3) genotype information can remain in the clients possession, and 4) variant filtration can be executed and saved for later processing. For a more detailed description of SG-ADVISER UI functionality, see *S1 Text*.

We believe the combination of holistic annotations and predictions provided by SG-ADVISER, plus the power to utilize those annotations alongside genotype information in the SG-ADVISER UI provides a powerful tool for the up-to-date processing of whole genome sequence information by individuals with little to no computational experience.

Discussion

To our knowledge, SG-ADVISER is the most comprehensive and accurate annotation and variant filtration tool available. The overall goal of the SG-ADVISER suite of tools is to put computational power and bioinformatics expertise into the hands of individuals with little to no computational proficiency, but with the biological and/or clinical expertise to interpret genetic results when appropriately filtered, while protecting the privacy of study subjects. The annotations and filtration strategies enabled by the SG-ADVISER suite have been successfully used in the molecular genetic diagnosis of numerous idiopathic disease cases at The Scripps Translational Science Institute [52]. We hope to enable these sorts of investigations outside of the major genomics centers.

Furthermore, it is clear that sequence-based investigation into common disease will require the ability to accurately parse and prioritize regulatory variants. Therefore, we have placed some emphasis on building tools to not only determine whether a TFBS or miRNA binding site contains a variant, but whether that variant changes the function of that binding site in any meaningful way. Given the known sensitivity of set-based rare variant analysis methods to the inclusion of non-causal variants indicates, it is clear that automated set generation will require variant prioritization in order to achieve maximal power [8,9].

SG-ADVISER will continue be updated and expanded to provide access to new annotations/predictions as necessary. Questions and requests for specific annotations can be made on the Biostar forum http://www.biostars.org/.

Supporting Information

S1 Text. Detailed information about annotation types, annotation processes, and user interface functionality are provided.
(DOC)

Acknowledgments

This work was supported by the NHGRI Genome Sequencing Informatics Tools (GS-IT) Program via grant National Institute of Health U01 HG006476 to Ali Torkamani. Further information available at: http://iseqtools.org. Further support is provided by Scripps Genomic Medicine, a National Institute of Health—National Center for Advancing Translational Sciences Clinical and Translational Science Award (CTSA; 5 UL1 RR025774) to STSI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the experiments: NJS AT. Performed the experiments: PHP WJS GAE AT. Analyzed the data: PHP WJS GAE AT. Contributed reagents/materials/analysis tools: PHP WJS GAE AT. Wrote the paper: PHP WJS GAE NJS AT.

REFERENCES

1. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636–639. doi: 10.1126/science.1186802 PMID: 20220176
2. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13: 255–262. doi: 10.1097/GIM.0b013e3182088158 PMID: 21173700
3. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, et al. (2011) Whole-genome sequencing for optimized patient management. Sci Transl Med 3: 87re83. doi: 10.1126/scitranslmed.3002243 PMID: 21677200
4. Flaherty KT, Puzanov I, Kim KB, Ribas A, McArthur GA, et al. (2010) Inhibition of mutated, activated BRAF in metastatic melanoma. N Engl J Med 363: 809–819. doi: 10.1056/NEJMoa1002011 PMID: 20818844
5. Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, et al. (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 359: 1757–1765. doi: 10.1056/NEJMoa0804385 PMID: 18946061
6. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569. doi: 10.1038/ng.608 PMID: 20562875
7. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88: 294–305. doi: 10.1016/j.ajhg.2011.02.002 PMID: 21376301
8. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311–321. doi: 10.1016/j.ajhg.2008.06.024 PMID: 18691683
9. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773–785. doi: 10.1038/nrg2867 PMID: 20940738
10. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242–245. doi: 10.1038/nature11011 PMID: 22495311
11. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485: 237–241. doi: 10.1038/nature10945 PMID: 22495306

12. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, et al. (2010) A de novo paradigm for mental retardation. Nat Genet 42: 1109–1112. doi: 10.1038/ng.712 PMID: 21076407
13. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, et al. (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493: 216–220. doi: 10.1038/nature11690 PMID: 23201682
14. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, et al. (2010) SCAN: SNP and copy number annotation. Bioinformatics 26: 259–262. doi: 10.1093/bioinformatics/btp644 PMID: 19933162
15. Lee PH, Shatkay H (2008) F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res 36: D820–824. PMID: 17986460
16. Li S, Ma L, Li H, Vang S, Hu Y, et al. (2007) Snap: an integrated SNP annotation platform. Nucleic Acids Res 35: D707–710. PMID: 17135198
17. Ge D, Zhang K, Need AC, Martin O, Fellay J, et al. (2008) WGAViewer: software for genomic annotation of whole genome association studies. Genome Res 18: 640–643. doi: 10.1101/gr.071571.107 PMID: 18256235
18. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070. doi: 10.1093/bioinformatics/btq330 PMID: 20562413
19. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, et al. (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. doi: 10.4161/fly.19695 PMID: 22728672
20. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164. doi: 10.1093/nar/gkq603 PMID: 20601685
21. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. doi: 10.1093/bioinformatics/btq033 PMID: 20110278
22. Asmann YW, Middha S, Hossain A, Baheti S, Li Y, et al. (2012) TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. Bioinformatics 28: 277–278. doi: 10.1093/bioinformatics/btr612 PMID: 22088845
23. Makarov V, O'Grady T, Cai G, Lihm J, Buxbaum JD, et al. (2012) AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics 28: 724–725. doi: 10.1093/bioinformatics/bts032 PMID: 22257670
24. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272–276. doi: 10.1038/nature08250 PMID: 19684571
25. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249. doi: 10.1038/nmeth0410-248 PMID: 20354512
26. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863–874. PMID: 11337480
27. Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 34: E2393–2402. doi: 10.1002/humu.22376 PMID: 23843252
28. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi: 10.1038/nature11632 PMID: 23128226
29. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64–69. doi: 10.1126/science.1219240 PMID: 22604720
30. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311. PMID: 11125122
31. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y (2013) Identifying personal genomes by surname inference. Science 339: 321–324. doi: 10.1126/science.1229566 PMID: 23329047
32. Stenson PD, Ball E, Howells K, Phillips A, Mort M, et al. (2008) Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet 45: 124–126. doi: 10.1136/jmg.2007.055210 PMID: 18245393
33. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–517. PMID: 15608251

34. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42: D980–985. doi: 10.1093/nar/gkt1113 PMID: 24234437

35. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36: 431–432. PMID: 15118671

36. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367. doi: 10.1073/pnas.0903103106 PMID: 19474294

37. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, et al. (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res 38: D652–657. doi: 10.1093/nar/gkp995 PMID: 19906727

38. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, et al. (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med 10: 294–300. doi: 10.1097/GIM.0b013e31816b5cae PMID: 18414213

39. Bairoch A, Apweiler R (1997) The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 25: 31–36. PMID: 9016499

40. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823–828. doi: 10.1126/science.1215040 PMID: 22344438

41. Hu J, Ng PC (2012) Predicting the effects of frameshifting indels. Genome Biol 13: R9. doi: 10.1186/gb-2012-13-2-r9 PMID: 22322200

42. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics 20: 1006–1014. PMID: 14751981

43. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ (2011) The importance of phase information for human genomics. Nat Rev Genet 12: 215–223. doi: 10.1038/nrg2950 PMID: 21301473

44. Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11: 377–394. PMID: 15285897

45. Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, et al. (2011) SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics 27: 1998–2000. doi: 10.1093/bioinformatics/btr317 PMID: 21624899

46. Medina I, De Maria A, Bleda M, Salavert F, Alonso R, et al. (2012) VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. Nucleic Acids Res 40: W54–58. doi: 10.1093/nar/gks572 PMID: 22693211

47. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46: 310–315. doi: 10.1038/ng.2892 PMID: 24487276

48. Shetty AC, Athri P, Mondal K, Horner VL, Steinberg KM, et al. (2010) SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. BMC Bioinformatics 11: 471. doi: 10.1186/1471-2105-11-471 PMID: 20854673

49. Paila U, Chapman BA, Kirchner R, Quinlan AR (2013) GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLoS Comput Biol 9: e1003153. doi: 10.1371/journal.pcbi.1003153 PMID: 23874191

50. Sifrim A, Van Houdt JK, Tranchevent LC, Nowakowska B, Sakai R, et al. (2012) Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease. Genome Med 4: 73. doi: 10.1186/gm374 PMID: 23013645

51. Sincan M, Simeonov DR, Adams D, Markello TC, Pierson TM, et al. (2012) VAR-MD: a tool to analyze whole exome-genome variants in small human pedigrees with mendelian inheritance. Hum Mutat 33: 593–598. doi: 10.1002/humu.22034 PMID: 22290570

52. Chen YZ, Friedman JR, Chen DH, Chan GC, Bloss CS, et al. (2014) Gain-of-function ADCY5 mutations in familial dyskinesia with facial myokymia. Ann Neurol.