# SVI: a simple single-nucleotide Human Variant Interpretation tool for Clinical Use

Paolo Missier[1], Eldarina Wijaya[1], Ryan Kirby[1], and Michael Keogh[2]

[1] School of Computing Science
Newcastle University, UK
{firstname.lastname}@ncl.ac.uk
[2] Institute of Genetic Medicine
Newcastle University, UK
michael.keogh@newcastle.ac.uk

**Abstract.** The rapid evolution of Next Generation Sequencing technology will soon make it possible to test patients for genetic disorders at population scale. However, clinical interpretation of human variants extracted from raw NGS data in the clinical setting is likely to become a bottleneck, as long as it requires expert human judgement. While several attempts are under way to try and automate the diagnostic process, most still assume a specialist's understanding of the variants' significance. In this paper we present our early experiments with a simple process and prototype clinical tool for single-nucleotide variant filtering, called SVI, which automates much of the interpretation process by integrating disease-gene and disease-variant mapping resources. As the content and quality of these resources improve over time, it is important to identify past patients' cases which may benefit from re-analysis. By persistently recording the entire diagnostic process, SVI can selectively trigger case re-analysis on the basis of updates in the external knowledge sources.

## 1 Introduction

### 1.1 Background and Motivation

Whole-exome and whole-genome sequencing (WES, WGS) are increasingly utilised in clinical diagnostics. As the cost of sequencing a human genome continues to decrease [1], and with the number of DNA base pairs sequenced per $ unit reportedly doubling every five months [2], WGS-based genetic testing is poised to become a routine diagnostic technique that can be deployed on a large scale [3]. At the same time, allocating the computation resources needed to process the data is also becoming increasingly affordable. Large initiatives like the 100,000 Genome Project in the UK[3], with specific focus on cancer and rare diseases, promise to deliver genetic testing at population scale within the next few years.

[3] `http://www.genomicsengland.co.uk/`

As genetic diseases affect about 8% of the UK population (5 million people), the potential societal benefits in this country alone are substantial.

The diagnosis of genetic disorders based on WGS data consists of two main stages: variant calling and variant interpretation. Variant calling includes processing the patients genome, or the exome [4,5], using a well-established sequence of computational steps, arranged into a pipeline. This results in a large set of variants, or single-nucleotide mutations and indels. The pipeline incorporates bioinformatics tools chosen from a growing pool of publicly available distributions [6]. The second stage involves analysing the variants based on a clinical hypothesis established from the patients phenotype, with the goal to identify variants that support the hypothesis.

The increasing volume of genomes to be processed, along with the widespread adoption of genetic testing in the clinic, call for scalable solutions for both phases. The *Cloud-e-Genome* project, a collaboration between the Institute of Genetic Medicine and the School of Computing Science at Newcastle University, was funded in 2013[4] to investigate such solutions.

In this paper we focus specifically on the variant interpretation phase, while a separate strand of work is concerned with the exploitation of cloud infrastructure to address scalability of the NGS data processing pipeline [7]. A first scalability issue concerning interpretation is that, although the gap between research and clinical exploitation of genetic diagnostic tools is narrowing, variant interpretation remains a knowledge-intensive decision process, especially for the diagnosis of rare disorders [8]. Diagnosis often requires the expertise of a geneticist, a scarce and expensive resource, for all but the most common cases. This makes the process difficult to scale, as larger number of patients are enrolled for testing.

A second scalability issue is more subtle. Diagnosis relies upon a combination of knowledge, i.e. variant-disease associations, and bioinformatics tools, which compose the exome / genome processing pipeline. Incomplete knowledge and limitations in the tools still result in both false positives and false negatives, or in inconclusive diagnosis, with success rate reported as low as 25% [9]. As both these elements evolve over time, however, there is an expectation that accuracy will improve, suggesting that it may beneficial to periodically revisit certain old cases that may have not been fully solved at the time they were first addressed. The choice of which cases to revisit depends on the combination of knowledge sources and tool selection used to process the original data, and the type of updates that become available, i.e., either in a variant database or in the pipeline. As these cases add to the volume, it is important to ensure that they are chosen accurately.

### 1.2 Goals

With these premises, in this project we explore two hypotheses. Firstly, that it is possible to automate much of the diagnostic process, by capturing its most

[4] Funding for Cloud-e-Genome comes from the NIHR (National Institute for Health and Research) and Biomedical Research Centre in the UK.

common elements into a simple-to-use tool which integrates with a number of external knowledge sources. And secondly, that by recording all details of each patient investigation, from variants to diagnosis, it becomes possible to selectively identify old cases that might benefit from re-analysis, in light of knowledge and/or technology advances.

### 1.3 Contributions

As our first contribution we have studied a cohort of five patients, seen by the Institute of Genetic Medicine (IGM) in Newcastle since 2012, to determine how the temporal evolution of variant-disease associations in the ClinVar[5] variation database affected the ability to diagnose their phenotypes (Sec.2). This small study supports our hypothesis that complete traceability and reproducibility of the diagnostics process is an important requirement, as it enables past patient cases to be selectively revisited based on their original outcome and following updates in the knowledge base.

Our second contribution is the design of a process for single-nucleotide variant interpretation, which reflects emerging practice in the research lab while aiming to bridge the knowledge gap between genetic research and clinical diagnosis. The process is described in Sec.3.

Thirdly, based on such process we have been implementing a variant interpretation user tool that simplifies the decision process by integrating multiple external knowledge sources to assist in the diagnosis. The tool, code-named SVI and still currently under development, is described in Sec. 4. SVI currently integrates OMIM[6] and ClinVar as its main external knowledge sources. However, the architecture is designed to accept additional sources of disease-variant associations as those may become available.

The SVI tool is still under active development, in collaboration with researchers at the IGM.

### 1.4 Related work

To the best our knowledge, most of the tools available for variant interpretation cater more to geneticist researchers than to clinicians. One example is the Exomiser [10, 11], which computes variant prioritisation according to a number of user-defined criteria, which partially overlap with those used in SVI. Pathogenicity prediction comes from the dbNSFP database [12]. Although the online tool offers a simple input interface, its output would be difficult for non-specialist clinicians to interpret.

Qiagen's Ingeniuty Variant Analysis is a mature tool that benefits from the HGMD variant-disease association knowledge base[7]. While it purportedly does target variant interpretation in the clinic, it is a commercial product that plays a role in the genetic diagnostics market.

In contrast, Extasy [13] is a research product, derived from the Annotate-it tool [14], which relies on a combination of multiple predictions from different

[5] http://www.ncbi.nlm.nih.gov/clinvar/

[6] http://www.ncbi.nlm.nih.gov/omim

[7] http://www.hgmd.cf.ac.uk/

sources. We see this tool as a possible additional source of predictive knowledge of pathogenicity, which we may try to integrate into SVI in the future. Once again, however, its output is designed to be consumed by specialists.

### 1.5 Recording the diagnostic process

One novel feature of SVI is the tracking of the entire diagnostic process, for each patient case, including human decisions as well as the dependencies amongst the data consumed and produced at each step, from user input to diagnosis (which may be inconclusive). This form of systematic provenance tracking aims to bring a number of additional benefits to users. Firstly, provenance tracking provides a way to fulfill one of our main goals, namely to determine which past cases should be revisited, in view of updates to any of the knowledge bases involved in the process (or when a new one is added).

Secondly, it provides both accountability and the ability to explain the decision process in detail. This is important not only because of the sensitivity of the process domain (clinical diagnostics), but also because of the sensitivity of the process itself. These include, amongst others, the version of external data sources, as well as the parameters used for variant filtering, as briefly described in Sec. 4.

Finally, as the collection of provenance traces grows and it is stored persistently, SVI provides support for a variety of analytical functions that cut across patient cases, different clinicians, and also range over time. For example, one common use case for this capability is to establish associations amongst independent cases, based on commonalities amongst the data involved in each of their processes. In turn, this has the potential to make investigators more efficient by allowing them to selectively share their cases with other group members.

### 1.6 Choosing a primary variant database.

It is broadly accepted within the genetic research community that no single variant database is sufficient to cover a broad range of pathologies. We have chosen to use ClinVar, NCBI‘s human genomic variations database, as our primary source for integration into SVI, on account of its fast growth and good overall coverage, as well as based on availability considerations. While several other variant repositories are available, not all of them are freely accessible (eg HGMD, mentioned earlier, which requires a license), and those that are tend to focus on specific phenotypes, or sub-specialties of clinical practice, or are exposed to false negatives due to incompleteness. Two prominent examples are the family of Locus Specific Mutation Databases (LSDB)[8], hosted on the LOVD (Leiden Open Variation Database) platform[9], and the Decipher project [15].

*LSDB.* As each LSDB is locus-specific, investigations that focus on specific phenotypes require that the appropriate databases be selected within the family. Although their common LOVD interface facilitates integration through programmatic access, their coverage is unpredictable and on a number of cases they

[8] http://grenada.lumc.nl/LSDB_list/lsdbs

[9] http://www.lovd.nl/

have proven unreliably incomplete for the purpose of clinical diagnosis. Consider for instance the NM_020745.3 single nucleotide variant on Gene AARS2 (c.1774C>T). This variant has been described as being highly likely to be pathogenic, as described in the next section. ClinVar records the variant as *Likely Pathogenic* with a known associated condition, which was last evaluated in Aug. 2014, and cites the relevant support literature [16]. Searching for AARS2 variants across the LOVD network returns hits in three additional databases: the LOVD shared installation (LUMC - NL), LOVD at University of Melbourne, and the Mitochondrial Disease MSeqDR-LSDB (Massachusetts). However, of these only MSeqDR-LSDB reports the variant, and it actually cites ClinVar as the source. Other pathogenic variants on AARS2, listed on ClinVar, are missing from the entire network at the time of writing.

*Decipher* is a recent project aimed at sharing knowledge of genotype-phenotype associations, following the rationale that "accurate diagnosis of human genetic disorders in a clinical setting requires the identification of other patients that share the same/similar genomic variants and comparison of their phenotypes." [15]. The are two main reasons why Decipher is not a suitable choice for our investigations. Firstly, it is once again focused on specific phenotypes, namely developmental delay disorders in children. Such phenotypes are not common in the clinical setting from which our test cases were obtained, which specialises on rare mitochondrial diseases and degenerative disorders. Secondly, it relies on submission of anonymised patient data. In contrast, privacy and patient consent must be considered before uploading large scale individual genetic data in the clinical or research setting. Decipher remains, however, one of the best examples of international collaborative phenotype-genotype consortia. In the future we may be able to engage with similar initiatives in the area of adult rare disease, such as GEM.app[10].

## 2 A small-scale time-travel experiment

We now present a study on 5 WES patient cases, all of them with the same phenotype (*multiple mitochondrial respiratory complex deficiency*), which were solved by our geneticist researchers in October 2012. The aim of this study is manifold. We want to determine whether or not a diagnosis can be reached using a limited number of external knowledge sources, such as OMIM and Clinvar. We are also interested in tracking, albeit at an anedoctal level, how the diagnostic power of those sources changes over time, and how it compares with a diagnostic process based solely on published literature research. Finally, we have used the study experience to help design the process that forms the basis for our tool.

The study involved "going back in time", in this case to 2012, to see whether the knowledge that was available then was sufficient to produce a diagnosis, either by an expert, who would be using direct research from phenotypic or investigational search terms relevant for each case within literature search databases

[10] `https://genomics.med.miami.edu/`

such as PubMed, or by an automated process using ClinVar. Our findings are summarised in Table 1, while the charts in Fig. 1 give a sense of progress in ClinVar content over time, by reporting on the number of variants of interest available in 2012 and in 2014.

| Patient | Gene Name | Variant | Clinvar 2014 | Date submitted |
|---|---|---|---|---|
| 1 | C12orf65 | Hom c.210delA:p.P70fs | Pathogenic | 22-Nov-13 |
| 2 | RMND1 | Hom. c.1349G>C:p.*450Serext*32 | Pathogenic | 04-Aug-14 |
| 3 | AARS2 | Het c.1774C>T:p.Arg592Trp | Pathogenic | 04-Aug-14 |
| 4 | MTO1 | Hom. c.1232C>T:p.Thr411Ile | Not found | |
| 5 | VARS2 | Het c.1045G>A:p.Ala349Thr | Not found | |

Table 1: Variants identified in Clinvar based on records in November 2014.

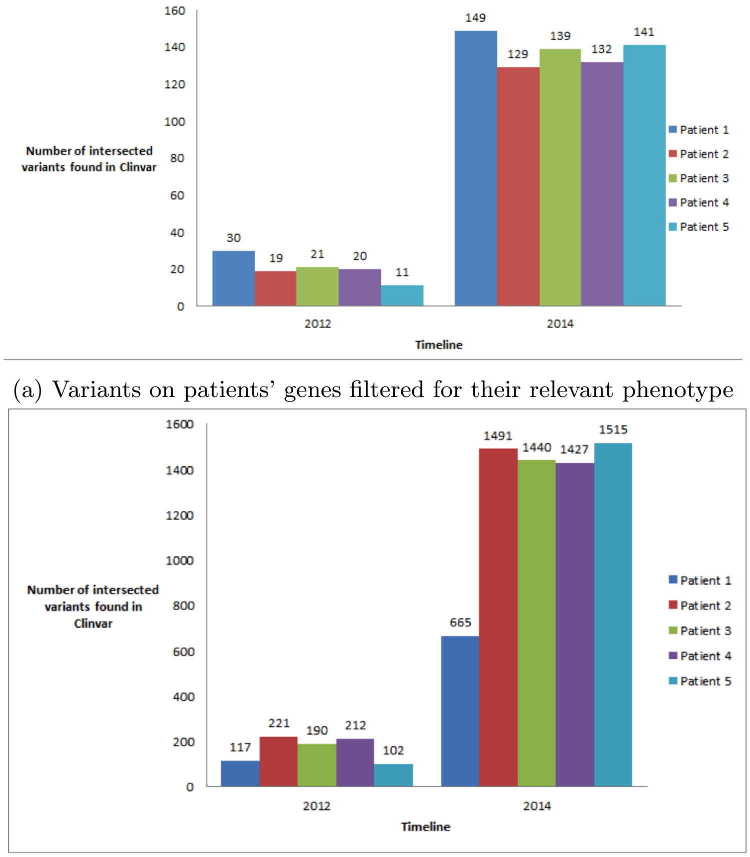
As the cases were indeed solved with a positive diagnosis, we benefit from the ground truth consisting of the actual variants found by the researchers. Our first finding is that none of these variants were recorded in the 2012 version of ClinVar, while only three out of five appear in the 2014 version. When they do appear, their clinical significance is reported as Pathogenic/Likely pathogenic, confirming the early researchers' diagnosis. This seems to support, at least anecdotically, the hypothesis that the relevance of a variant databases like ClinVar does increase over time, complementing and possibly eventually replacing experts' knowledge.

Next, we focused on articles that could have been used at different points in time as reference to solve the cases. We recorded the number of papers available at the time of diagnosis, which are related to the patient phenotype, as well as the number papers published before the date of diagnosis. Our findings, reported in Table 2, indicate that of the five cases, only two could have been solved using literature support. One additional case (patient 5) was solved using direct researchers' knowledge of association between the VARS2 gene and the multiple mitochondrial complex deficiency phenotype.

Despite these successes, it is often the case that genetic diagnosis cannot be reached. To illustrate, we have analysed a further patient, which to date is still an unsolved case. Researchers manually identified eight candidate variants for this patient in 2012, however none of those appeared in ClinVar at the time, or could otherwise be confirmed as pathogenic. Using the 2014 version of ClinVar, only one of the variants (c.242G>A:p.Arg81Gln on gene TYMP) was found to be benign, while the others remain unknown. No additional literature has so far emerged (to the best of our knowledge) to support the diagnosis.

## 3 Variant interpretation for genetic diagnosis

We now describe the process of single-nucleotide variant interpretation that underpins our clinical tool, SVI. In a clinical setting, the interpretation process is normally driven by a *disease hypothesis*, specified by the clinician on the basis



(a) Variants on patients' genes filtered for their relevant phenotype

(b) Variants on patients' genes, no phenotype filtering

Fig. 1: ClinVar evolution relative to the variants of interest for sample patients

| Patient | Gene Name | Variant | Pubmed Publications (2014) | Year reference paper published | Pubmed publications (before 2012) | Solvable before 2012? |
|---|---|---|---|---|---|---|
| 1 | C12orf65 | Hom c.210delA:p.P70fs | 9 | 2010 | 2 | Yes |
| 2 | RMND1 | Hom. c.1349G>C:p.*450Serext*32 | 6 | 2012 | 3 | No |
| 3 | AARS2 | Het c.1774C>T:p.Arg592Trp | 4 | 2011 | 1 | Yes |
| 4 | MTO1 | Hom. c.1232C>T:p.Thr411Ile | 48 | 2012 | 29 | No |
| 5 | VARS2 | Het c.1045G>A:p.Ala349Thr | 14 | NA | 11 | No |

Table 2: Number of publications in Pubmed concerning the gene of interest for a specific variant, prior to date of diagnosis in 2012 and in 2014.

of factual observations. The goal of the process is to find variants in the patient's exome, amongst those called by the upstream pipeline, which have either previously been reported to be associated conclusively with similar phenotypes, or conform to the appropriate inheritance pattern, and disease population frequency and occur in genes either known to cause a similar phenotype, or affect similar biological functions. In addition, in silico software tools provide a mechanism of inferring the biological effect of the mutation. The diagnosis is considered inconclusive (on the basis of the variants alone) if no such variants can be found.

Genome variant interpretation has been described as a "needle in the bunch on needles" problem [17], as the target variants are a tiny proportion, typically no more than ten, of the more than 20,000 variants that are detected by a typical pipeline. The vast majority of variants are benign, such as common polymorphisms, which do not affect a patients health. Ideally, the variants of interest lie at the intersection between two subsets of the overall patient's variants, namely (i) deleterious variants, i.e., protein altering and splice site altering mutations, and (ii) variants that are known from the literature to play a role in the target phenotype. As we will see, however, it is not always possible to identify variants that lie precisely in this intersection. Our selection process therefore aims at segregating variants into classes, depending on the amount of available evidence to support the hypothesis that they are indeed the basis for a disease diagnosis. The process consists of three phases, which we describe next: (i) restricting the investigation to a specific set of genes (phenotype and variant scoping), (ii) variant filtering aimed at identifying deleterious variants, and (iii) variant classification. The overall process is depicted in Fig. 2.

### 3.1 Phenotype and variant scoping.

In this phase, user input terms are mapped to genes. Users may specify the disease hypothesis at varying levels of precision, ranging from free text keywords, to terms from the OMIM vocabulary[11] or from the Human Phenotype Ontology [18] (HPO[12]). The latter provides a more precise characterisation of the phenotype (so called *deep phenotyping* [19, 20]). OMIM and HPO both provide standard reference taxonomies of phenotype terms. In addition, we normalise all input formats to OMIM, which also offers phenotype-to-gene mapping. HPO provides a direct mapping to OMIM, and free text keywords are simply mapped to OMIM terms through string matching. The resulting OMIM terms are then mapped to a set of genes, which define the initial scope of the investigation, in the next phase.

As genetic testing in clinics tends to specialise on specific disorder areas, the scope can be further restricted to a set of genes that are known to be implicated in phenotypes in that area. Thus, when using the tool the clinician may also optionally provide a more precise characterisation of the scope of the investigation, by directly specifying a list of target genes of interest. This process, depicted on the top left in Fig. 2, produces a final set of *genes in scope*. Only the subset of

[11] `http://www.omim.org/`

[12] `http://www.human-phenotype-ontology.org/`

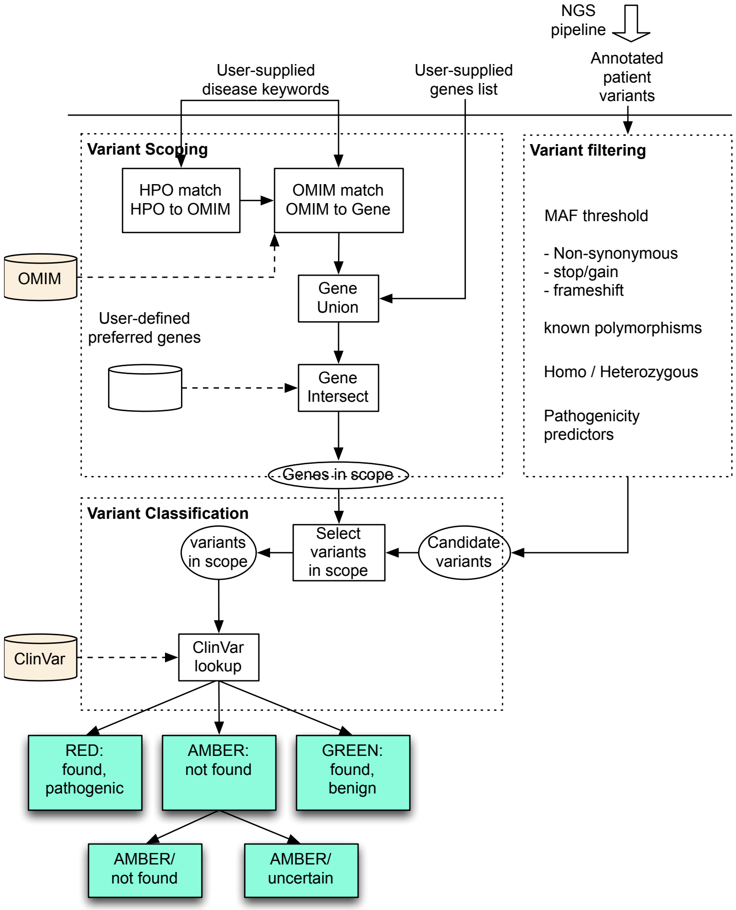


Fig. 2: Variant interpretation process as implemented in SVI.

candidate variants found in phase II (variant filtering), which lie on the scope genes, will be considered for classification.

### 3.2 Variant filtering for identification of deleterious variants

This phase relies on variant annotations, provided by the well-known Annovar annotation service [21]. SVI implements an extensible set of filters, reflecting emerging pratice in the lab. Currently, variants are filtered according to the following conditions.

- Identification of polymorphisms. Variants that are recorded as polymorphisms in the dbSNP database are excluded, as these are common mutations which occur at higher frequencies than the disease phenotype in the population, and are known to be non-deleterious.

- Coverage test. We check that variants are called at 30x fold or more, as this is a de facto standard for confidence in a read. Also, we check the exome coverage percentage (ie what fraction of the exome is covered to 30 fold), and distribution of % coverage across the exome, if this information is available.
- *Synonymous* variants are removed, as those are non-protein altering or splice site altering. Only non-synonymous, stop/gain, frameshift mutations are retained.
- Variants with MAFs (Minor Allele Frequency) greater than 0.01 are also discarded. Ideally, MAF should be checked separately against international controls as well as local control patients. For instance, harmless mutations that are rare within the general population (low MAF) may be incorrectly included, although a localised patients control database would reveal a higher frequency in the patients area of origin. No such localised databases are currently available to us, however.
- When performing trio genetic testing (typically involving parents and affected child), remove all variants which do not conform to pedigree, i.e., remove potentially pathogenic heterozygous variants due to their observation in an unaffected parent, and the detection of de novo variants. Also, determine whether the presence of the same variants is consistent with Mendelian inheritance, as indicated for instance in [22].
- User-defined thresholds on a variety of individual or aggregate pathogenicity predictors [23], including PolyPhen[13] and others that are available through Annovar annotations.

The outcome of this phase is denoted as *candidate variants* in Fig. 2.

### 3.3 Variant classification.

At this stage we have isolated variants with the following properties: (i) they are likely to be deleterious, and (ii) they lie in genes that are broadly related to the target phenotype, via OMIM mapping. The uncertainty associated with the filtering process, combined with the broad nature of OMIM disease-gene mapping, suggest that these conditions are still too weak to provide conclusive evidence in support of the hypothesis. Indeed, at this stage hundreds of variants are still under consideration, mostly false positives.

Definite evidence can only be provided by research on specific disease-variant associations. As mentioned, we have chosen ClinVar as our initial reference source, with the intention to extend the knowledge base to other sources in the future. To each known single nucleotide variant, ClinVar associates a clinical significance that is simple to interpret (Likely benign / Likely pathogenic / Uncertain) along with the condition associated with a pathogenic significance (using OMIM terms). Importantly for us, in view of the tracking capabilities of our tool, ClinVar also provides metadata about the review status of the entry, with timestamps of the latest update. As shown in the bottom part of Fig. 3, we

[13] `http://genetics.bwh.harvard.edu/pph2/`

exploit the ClinVar output to create a simple separation of the candidate variants, into three classes: Red, Amber, and Green, using a “traffic light” metaphor that clinicians are likely to find simple and useful.

**Red** variants are those that are recorded as pathogenic in ClinVar. Considering the prior filtering and scoping, these provide conclusive evidence for a positive diagnosis.

**Amber** variants are those that are in scope but either not known to ClinVar, denoted Amber/unknown, or recorded in ClinVar with Uncertain significance (Amber/uncertain). Variants `c.4132A>G:p.Ser1378Gly` on gene LRPPRC and `c.842G>A:p.Gly281Asp`on PARK2 are examples of Amber/unknown and Amber/Uncertain variants, respectively. These variants provide weaker evidence than the Red ones, yet they cannot be dismissed, as absence from ClinVar may simply mean that research is still be ongoing or that curation efforts have not yet brought recently published research into database.

Finally, **Green** variants are those that are found in ClinVar, reported as likely benign.

This simple user output is designed to reduce the clinician’s decision process, by separating the “easy” cases which reveal Red variants, from all others. Cases where Amber but no Red variants are found can be referred to specialist researchers for further investigation.

In SVI, these are the prime candidates for re-analysis when updates to ClinVar become available, or when new variant databases are integrated.

## 4 A provenance-aware diagnostic tool

We have implemented the process into SVI, a Web-based user tool designed to be used by clinicians. Evaluation of the tool is still ongoing, both in terms of effectiveness of the variant filtering, and in terms of usability. We define effectiveness as the ability to reproduce benchmark diagnostics decisions obtained by experts. While our results are still preliminary, as an example we report the effect of filtering on the five test patients used in the study described in Sec. 2. In all cases, from generic user input expressing the patients’ common phenotype (*multiple mitochondrial respiratory complex deficiency*) SVI identified between 7 and 11 Red variants, as indicated in Fig. 3 and in Table 3. In all cases, the Red variants include those listed in Table 1 on page 6.[14]

In addition to supporting the filtering process, SVI provides complete tracing of the process itself. The underlying data model (implemented using the MongoDB DBMS) is centred around the main concept of an *Investigation* (Fig.4). An investigation is part of a case about a patient. A case is owned by an investigator (the clinician/user), and it may consists of multiple investigations, each containing full details of one individual search. These details include a reference to the patient, user input (keywords, HPO, OMIM terms) along with their mapping to genes, the variants selected at each stage in the process, and the “traffic light” classification of each variant. Annotations made by the user in support

[14] Experts were not available to confirm whether any of the other Red variants had also been detected.

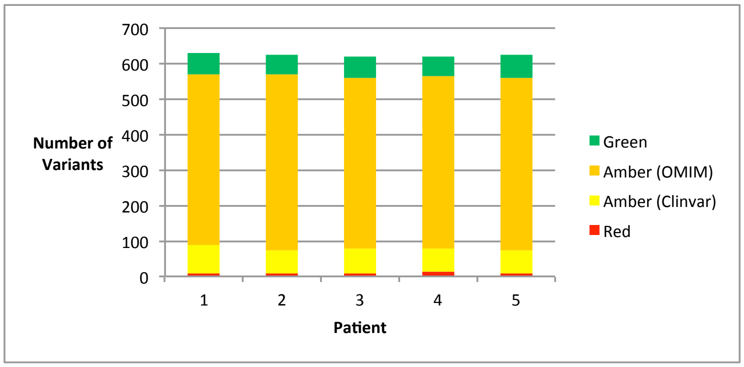


Fig. 3: Distribution of variants amongst the Red/Amber/Green classes for out patients sample

| Patient | Candidate variants | Present in ClinVar | Red | Amber (uncertain) | Amber (unknown) | Green |
|---|---|---|---|---|---|---|
| 1 | 631 | 149 | 10 | 77 | 482 | 62 |
| 2 | 625 | 129 | 7 | 65 | 496 | 57 |
| 3 | 622 | 139 | 7 | 69 | 483 | 63 |
| 4 | 618 | 132 | 11 | 67 | 486 | 54 |
| 5 | 627 | 141 | 8 | 65 | 486 | 68 |

Table 3: Effect of variant filtering in SVI  for a specific phenotype

of a decision, at the level of individual variants, are also captured. Finally, an investigation records the versions of all external data sources used for filtering.

An investigation provides a persistent provenance trace of each user execution. We are currently in the process of implementing a number of added value features on top of this provenance database. These include:

- The ability to selectively trigger new analysis of old cases, when changes occur anywhere in the knowledge sources (or indeed in the pipeline upstream). Specifically, when an Amber variant in an investigation appears or changes status in ClinVar, it moves from the Amber class to either the Green or the Red class, possibly resulting in the case being revisited by the clinician. This process can be automated through a simple *diff* process whenever a new version of ClinVar becomes available.
- Analyse historical investigations to determine possible implicit associations between independent cases. For instance, cases that exhibit a substantial overlap in the gene scope or the variant scope may be linked, so that whenever a problem/solution is found in one, the other can be flagged up for further consideration.
- Query the investigation database across multiple dimensions (patients, phenotype, investigator, time). Examples of queries include: "find all patients annotated with shared HPO terms, who also share variants or have vari-

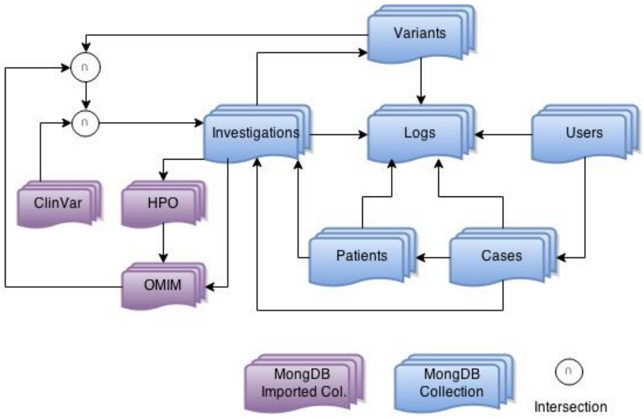


Fig. 4: Data model centred on investigations, designed for provenance support. The arrows indicate one-to-many or many-to-many relationships

ants on the same genes", and "determine how many patients with the same variant have the same HPO matching terms".

Most importantly, the provenance database provides accountability over the entire decision process. This is important not only for audit purposes, but also to allow third party clinicians, who have not been involved in the case, to fully understand how the investigator reached important decisions, which potentially affects a patient's quality of life.

## 5 Conclusions and current work

NGS-based genetic diagnosis is rapidly coming of age. As NGS technology matures, the new bottleneck is likely to be the clinical interpretation of the lists of human variants extracted from the raw WGS data, which remains a knowledge-intensive activity requiring expert human judgement. Making sure that the diagnostic process scales with the increasing volume of patient cases requires automation of this activity. In this paper we have presented an initial attempt at addressing this issue. We have been experimenting with a simple variant filtering process and tool, code-named SVI, which automates most of the process by relying on integration of variant databases. In this initial effort, we have chosen ClinVar as the exemplar variant database, as its content and curation appear to progress rapidly, increasing the chances to identify relevant pathogenic variants. The tool includes full traceability of the diagnostic process.

Our work is progressing in several directions. Firstly, we are now evaluating the effectiveness of SVI in terms of false positives/negatives relative to the expert judgment on a testbed of real patient cases. Secondly, we are working to integrate additional sources of variant-disease associations, such as those on the LOVD platform. Finally, as the number of investigations increases, we expect to be able to perform interesting analysis on the provenance database.

## References

1. Wetterstrand, K.: DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP) (2015)
2. Stein, L.D.: The case for cloud computing in genome informatics. Genome biology **11**(5) (January 2010) 207
3. Xuan, J., Yu, Y., Qing, T., Guo, L., Shi, L.: Next-generation sequencing in the clinic: promises and challenges. Cancer letters **340**(2) (2013) 284–295
4. Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D., Bonacci, B.B., Decker, B., Serpe, J.M., Dasu, T., Tschannen, M.R., Veith, R.L., Basehore, M.J., Broeckel, U., Tomita-Mitchell, A., Arca, M.J., Casper, J.T., Margolis, D.A., Bick, D.P., Hessner, M.J., Routes, J.M., Verbsky, J.W., Jacob, H.J., Dimmock, D.P.: Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genetics in medicine : official journal of the American College of Medical Genetics **13**(3) (March 2011) 255–62
5. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., Hardison, M., Person, R., Bekheirnia, M.R., Leduc, M.S., Kirby, A., Pham, P., Scull, J., Wang, M., Ding, Y., Plon, S.E., Lupski, J.R., Beaudet, A.L., Gibbs, R.A., Eng, C.M.: Clinical whole-exome sequencing for the diagnosis of mendelian disorders. The New England journal of medicine **369**(16) (October 2013) 1502–11
6. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z.: A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in bioinformatics (January 2013) bbs086–
7. Cala, J., Xu, Y.X., Wijaya, E.A., Missier, P.: From scripted HPC-based NGS pipelines to workflows on the cloud. In: Procs. C4Bio workshop, co-located with the 2014 CCGrid conference, Chicago, IL, IEEE (2013)
8. Shashi, V., McConkie-Rosell, A., Rosell, B., Schoch, K., Vellore, K., McDonald, M., Jiang, Y.H., Xie, P., Need, A., Goldstein, D.B., Goldstein, D.G.: The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genetics in medicine : official journal of the American College of Medical Genetics **16**(2) (February 2014) 176–82
9. Atwal, P.S., Brennan, M.L., Cox, R., Niaki, M., Platt, J., Homeyer, M., Kwan, A., Parkin, S., Schelley, S., Slattery, L., Wilnai, Y., Bernstein, J.A., Enns, G.M., Hudgins, L.: Clinical whole-exome sequencing: are we there yet? Genetics in medicine : official journal of the American College of Medical Genetics **16**(9) (September 2014) 717–719
10. Zemojtel, T., Kohler, S., Mackenroth, L., Jager, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., Oien, N.C., Schweiger, M.R., Kruger, U., Frommer, G., Fischer, B., Kornak, U., Flottmann, R., Ardeshirdavani, A., Moreau, Y., Lewis, S.E., Haendel, M., Smedley, D., Horn, D., Mundlos, S., Robinson, P.N.: Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Science Translational Medicine **6**(252) (September 2014) 252ra123–252ra123

11. Robinson, P.N., Köhler, S., Oellrich, A., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., Gilissen, C., Haendel, M., Smedley, D.: Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome research **24**(2) (February 2014) 340–8
12. Liu, X., Jian, X., Boerwinkle, E.: dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation **32**(8) (August 2011) 894–899
13. Sifrim, A., Popovic, D., Tranchevent, L.C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J.R., Aerts, J., De Moor, B., Moreau, Y.: eXtasy: variant prioritization by genomic data fusion. Nature methods **10**(11) (November 2013) 1083–4
14. Sifrim, A., Van Houdt, J.K., Tranchevent, L.C., Nowakowska, B., Sakai, R., Pavlopoulos, G.A., Devriendt, K., Vermeesch, J.R., Moreau, Y., Aerts, J.: Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease. Genome medicine **4**(9) (January 2012) 73
15. Swaminathan, G.J., Bragin, E., Chatzimichali, E.A., Corpas, M., Bevan, A.P., Wright, C.F., Carter, N.P., Hurles, M.E., Firth, H.V.: DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. Human molecular genetics **21**(R1) (October 2012) R37–44
16. Götz, A., Tyynismaa, H., Euro, L., Ellonen, P., Hyötyläinen, T., Ojala, T., Hämäläinen, R.H., Tommiska, J., Raivio, T., Oresic, M., Karikoski, R., Tammela, O., Simola, K.O.J., Paetau, A., Tyni, T., Suomalainen, A.: Exome sequencing identifies mitochondrial alanyl-tRNA synthetase mutations in infantile mitochondrial cardiomyopathy. American journal of human genetics **88**(5) (May 2011) 635–42
17. Cooper, G.M., Shendure, J.: Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature reviews. Genetics **12**(9) (September 2011) 628–40
18. Robinson, P.N., Mundlos, S.: The human phenotype ontology. Clinical genetics **77**(6) (2010) 525–534
19. Hennekam, R.C.M., Biesecker, L.G.: Next-generation sequencing demands next-generation phenotyping. Human Mutation **33**(5) (2012) 884–886
20. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C.M., Brown, D.L., Brudno, M., Campbell, J., FitzPatrick, D.R., Eppig, J.T., Jackson, A.P., Freson, K., Girdea, M., Helbig, I., Hurst, J.A., Jähn, J., Jackson, L.G., Kelly, A.M., Ledbetter, D.H., Mansour, S., Martin, C.L., Moss, C., Mumford, A., Ouwehand, W.H., Park, S.M., Riggs, E.R., Scott, R.H., Sisodiya, S., Van Vooren, S., Wapner, R.J., Wilkie, A.O.M., Wright, C.F., Vulto-van Silfhout, A.T., de Leeuw, N., de Vries, B.B.A., Washingthon, N.L., Smith, C.L., Westerfield, M., Schofield, P., Ruef, B.J., Gkoutos, G.V., Haendel, M., Smedley, D., Lewis, S.E., Robinson, P.N.: The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research **42**(Database issue) (January 2014) D966–74
21. Wang, K., Li, M., Hakonarson, H.: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research **38**(16) (September 2010) e164–e164
22. Keogh, M.J., Daud, D., Chinnery, P.F.: Exome sequencing: how to understand it. Practical neurology (June 2013) 1–9
23. Thusberg, J., Olatubosun, A., Vihinen, M.: Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation **32**(4) (April 2011) 358–368