MedSavant: An open source platform for personal genome interpretation

Marc Fiume[1], James Vlasblom[2], Ron Ammar[3], Orion Buske[1], Eric Smith[1], Andrew Brook[1], Sergiu Dumitriu[2], Christian R. Marshall[2], Kym M. Boycott[4], Peter Ray[2], Gary D. Bader[1,3,5], Michael Brudno[1,2,*]

[1] Department of Computer Science, University of Toronto, Toronto, ON, Canada
[2] The Hospital for Sick Children, Toronto, ON, Canada
[3] The Donnelly Centre, Toronto, ON, Canada
[4] Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, ON, Canada
[5] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
* corresponding author

A key challenge in the broad deployment of personalized genomic medicine is the processing and interpretation of patients' genomes in the context of their clinical indications. To address this challenge we developed MedSavant, an open-source, app-based platform facilitating the organization, storage, annotation, search, and visualization of patient genomes. MedSavant is accessible to geneticists of all levels of computational expertise, can be installed locally or in a scalable cloud environment, and is freely available (open source) at http://medsavant.org.

The existing paradigm for human genome data analysis typically involves serial processing of flat text files followed by manual inspection of a small variant ser. These workflows require a substantial amount of both genetic and informatics expertise, are non-interactive and time-consuming, and do not easily scale to meet the needs of clinical and other large-scale sequencing applications.

Most freely available tools for the filtration and analysis of genomic variants rely on a command-line interface. For example, GATK[1] is a command-line tool that can be used to filter genetic variants stored in the text file-based variant call format (VCF). Similarly, GEMINI[2] introduced a datastore for genetic variants and an SQL-like language for expressing common queries against it. The few graphical tools do not scale to hundreds of human genomes: VarB[3] and VarSifter[4] are solutions for variant search that run entirely on the desktop, placing limitations on performance and restricting data access to a single computer. A number of related commercial tools are also available, however these are typically closed platforms that are not easily extendable.

To address these issues we have developed MedSavant. MedSavant can be installed on local commodity hardware or in cloud environments, and uses a secure protocol (SSL) to ensure privacy of patient data. A centralized datastore organizes patient (e.g. pedigree, disease, phenotype), genetic (e.g. read alignments, genetic variants) and other information (e.g. cohorts, gene panels, and variant comments). Genetic variants in VCF format can be uploaded via a drag-and-drop interface and are automatically annotated with a configurable set of functional, population, and disease information. MedSavant uses a columnar database technology (Infobright, http://www.infobright.org) which exhibits superior data compression compared to flat files (up to 10X) and faster query times (up to 150X) compared to commonly used command-line utilities (see Supplement, Section 2). Significant advancements in retrieval speed enable interactive exploration of large and richly annotated genomic datasets.

High-throughput DNA sequencing is used both for research (e.g. disease gene mapping) and in the clinic (e.g. diagnostics, incidental findings, pharmacogenetics). To support a wide range of applications, each workflow in MedSavant is implemented as a separate app, using an open Application Programming Interface (API). Tutorials and example code for app development are available on the MedSavant website. A publicly accessible App Store is available to encourage developers of open-source bioinformatics tools to easily produce and distribute MedSavant apps.

MedSavant is shipped with several apps immediately useful for both clinical and research workflows:

The *Discovery* app is a clinically-oriented variant search tool empowering clinicians to quickly identify rare and potentially causal variants in a patient's genome. It automatically filters

variants based on variant quality, harmfulness predictions (e.g. from Polyphen[5]), mutational effect (from Jannovar[6]), gene panels, and allele frequencies from population databases (e.g. 1000 Genome Project[7], 6500 exome NHLBI-ESP, http://evs.gs.washington.edu/EVS/). Included as a default gene panel is the set of 56 incidental genes described by the American College of Medical Genetics and Genomics (ACMG)[8], facilitating rapid incidental finding detection. Variants are annotated with modes of inheritance from ~2,800 genes, as described in the Clinical Genomic Database[9] and linked to Clinvar[10], to accelerate identification of clinically relevant variants. In Figure 1A we use the Discovery app to identify the causal variant in a patient with retinitis pigmentosa. Using default settings, a total of 520 missense and loss-of-function (LOF) mutations were identified that passed quality control thresholds. Using the Human Phenotype Ontology[11] to identify genes associated with decreased central vision, a single missense mutation was observed in the gene *RPE65*, which has been previously implicated in the disease[12].

For more advanced users, we created the *Variant Navigator* app, a general-purpose variant search engine. Queries can be constructed graphically based on criteria such as patient features (e.g. age, sex, phenotype), genotype features (e.g. quality values, genomic region, functional effect) and annotations (e.g. Gene Ontology[13], OMIM[14], COSMIC[15]). The resulting variants can be visualized using chromosomal heatmaps, searchable spreadsheets, or the integrated genome browser[16]. Detailed information is provided for selected variants and the genes in which they reside, including functionally similar genes as predicted by the GeneMANIA functional interaction network tool[17].

Mendel is a disease gene identification app to guide the resolution of genetic disorders through case-control and pedigree analyses. Given pedigree information, Mendel can perform segregation based on inheritance models. We benchmarked Mendel on 14 rare disorders studied by the FORGE consortium, a Canadian effort aimed at identifying the genes responsible for rare childhood diseases[18]. A MedSavant database was created containing 138,640,418 SNVs and small indels from exome sequencing of 424 FORGE subjects. For each of the chosen disorders, Mendel was used to identify genes having damaging variants that segregated in patients affected by the disorder, compared to the remaining (control) individuals. For example, the query shown in Figure 1B revealed missense and splicing mutations in *C5orf42* for patients affected by Joubert Syndrome[19], with no other results as shown in Figure 1C. In total, the causal gene was independently discovered for 13/14 (93%) of the selected disorders. For the disorder that could not be resolved using Mendel, two of the causal variants did not pass basic quality filters; they were validated with Sanger sequencing in the original study. A summary of Mendel queries used for each disorder is provided in Section 3 of the Supplement.

Growth in the size of sequencing repositories has inspired the development of cloud-based platforms for genome sequence data hosting and access. The MedSavant app for *Google Genomics* is a proof-of-concept that provides a visual interface for listing a large number of publicly accessible read alignment datasets remotely hosted through Google's API (developers.google.com/genomics)[20] and loading them as tracks that can be interactively explored in MedSavant's genome browser. This app demonstrates the potential to efficiently deliver large data collections to researchers with minimal computational infrastructure by leveraging cloud resources. The need to share genomic data has been recognized through initiatives like the Global Alliance for Genomics & Health (GA4GH; genomicsandhealth.org);

future versions will make use of Google and GA4GH APIs to integrate remotely hosted read and variant datasets into other MedSavant apps.

The MedSavant App framework supports realtime cooperation among installed apps. For example, from within the Patient Directory a user can select a patient, further filter the patient's variants using the Variant Navigator, and inspect the results using the Genome Browser or further analyze them with Mendel. The connectivity between apps is exemplified in Figure 2. The process of interpreting genomes within this environment is conceptually different from current approaches, which involve serial processing of data using independent tools. The collaborative system simplifies development and integration of tools, and encourages the growth of an ecosystem that will increase in power as more apps continue being developed, together making the translation of human genome and associated data into research discoveries and clinical applications easier and more efficient.

Figures

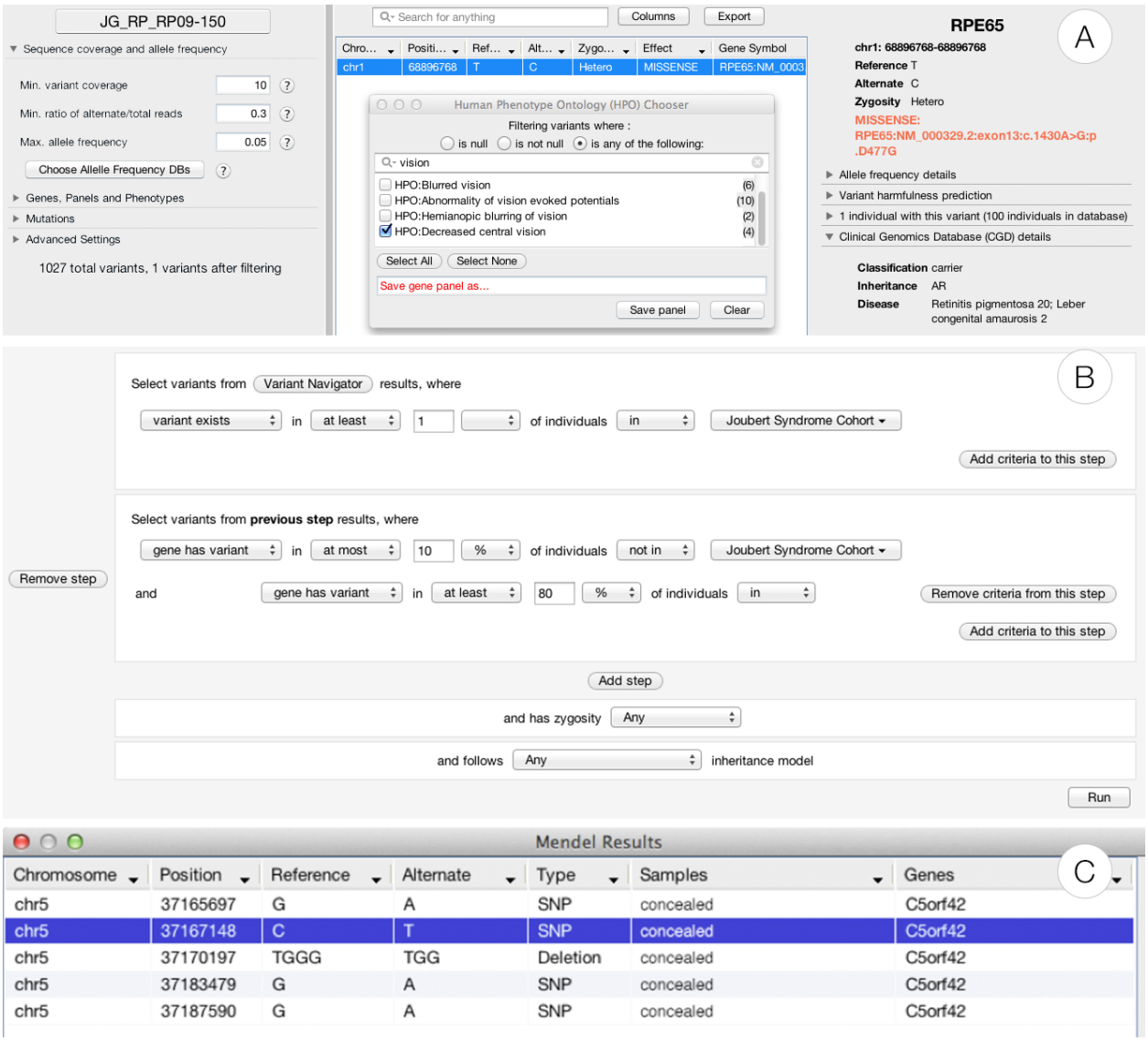

Figure 1: (A) Single patient analysis using Discovery, identifying a missense mutation causing retinitis pigmentosa. Mutations are filtered to low-frequency harmful, and then intersected for the set of genes previously implicated in decreased vision based on the Human Phenotype Ontology. (B) Mendel query used to identify harmful variants in a cohort of 9 French-Canadians with Joubert syndrome. Starting from a list of exonic, splicing, or UTR mutations with Allele Frequency <.05 and Quality > 30 (entered using the Variant Navigator) the query yields (C) only five variants, all in the *C5orf42* gene, suggesting that it is causal.

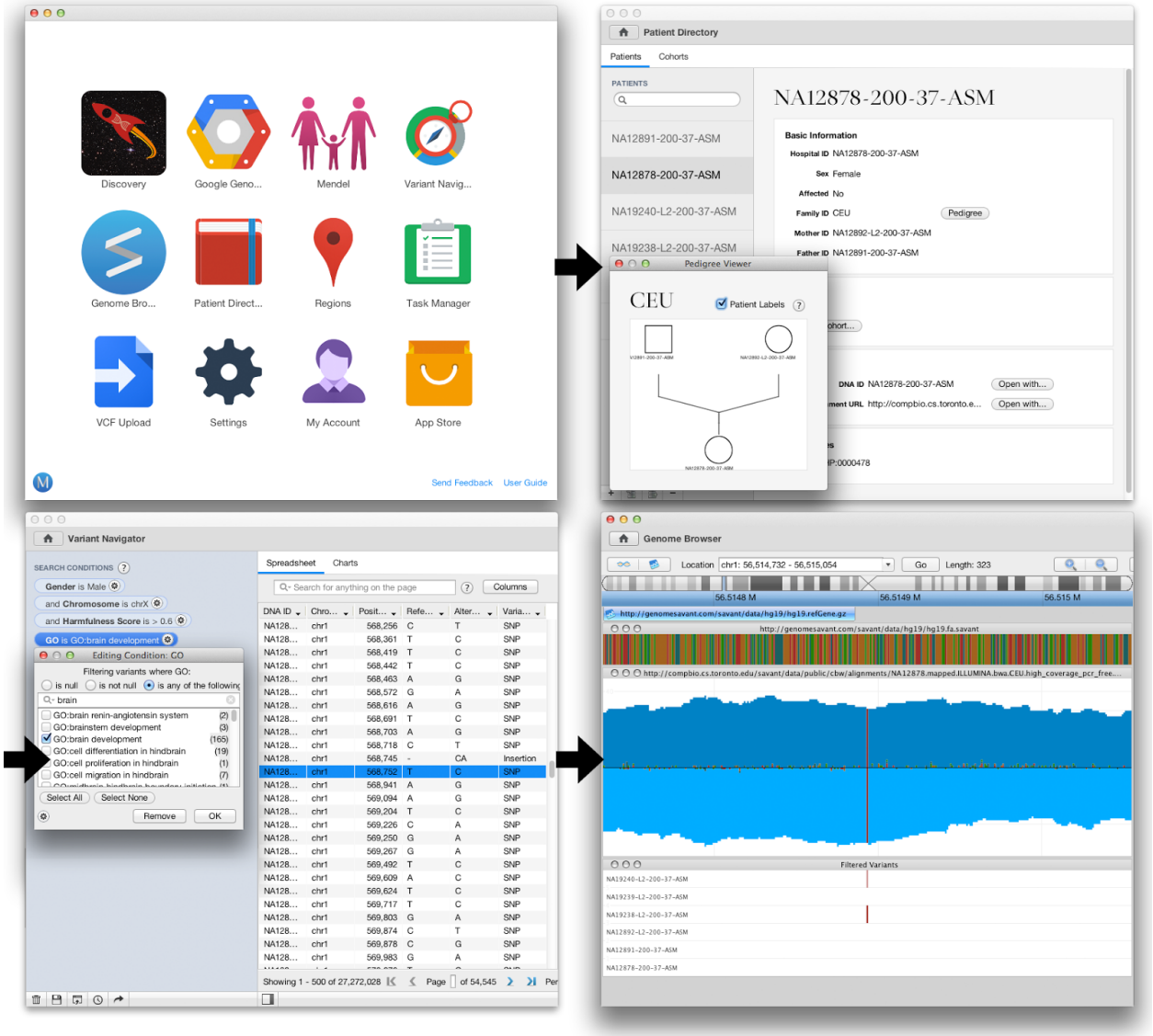

Figure 2: MedSavant dashboard (top-left) and apps. The Patient Directory (top-right) displays patient, family, and cohort information. Genetic variants for individual or groups of patients can be flexibly searched in the Variant Navigator (bottom-left). Search criteria are configured in the left panel, and results are shown in the spreadsheet to the right. Conditions can be constructed based on patient (e.g. sex), genetic (e.g. chromosome), annotation (e.g. harmfulness), and ontology (e.g. Gene Ontology) information. Variants of interest can be visualized using the Savant Genome Browser (bottom-right).

References

1. McKenna, A. *et al. Genome Res.* 20, 1297–1303 (2010).
2. Paila, U. *et al. PLoS Comput. Biol.* 9, e1003153 (2013).
3. Preston, M. D. *et al. Bioinformatics* 28, 2983–2985 (2012).
4. Teer, J. K., *et al. Bioinformatics* 28, 599–600 (2011).
5. Adzhubei, I. A. *et al. Nat. Methods* 7, 248–249 (2010).
6. Jäger, M. *et al. Hum. Mutat.* 35, 548–555 (2014).
7. Siva, N. *Nat. Biotechnol.* 26, 256 (2008).
8. Green, R. C. *et al. Genet. Med.* 15, 565–574 (2013).
9. Solomon, B. D. *et al. Proc. Natl. Acad. Sci. U. S. A.* 110, 9851–9855 (2013).
10. Landrum, M. J. *et al. Nucleic Acids Res.* 42, D980–5 (2013).
11. Robinson, P. N. *et al*. *Am. J. Hum. Genet.* 83, 610–615 (2008).
12. Roman, A. J. *et al. Invest. Ophthalmol. Vis. Sci.* 54, 1378–1383 (2013).
13. Ashburner, M. *et al. Nat. Genet.* 25, 25–29 (2000).
14. Hamosh, A., T. *Hum. Mutat.* 15, 57-61(2001).
15. Bamford, S. *et al. Br. J. Cancer* 91, 355–358 (2004).
16. Fiume, M. *et al. Nucleic Acids Res.* 40, W615–21 (2012).
17. Warde-Farley, D. *et al. Nucleic Acids Res.* 38, W214–20 (2010).
18. Beaulieu, C. L. *et al. Am. J. Hum. Genet.* 94, 809–817 (2014).
19. Srour, M. *et al. Am. J. Hum. Genet.* 90, 693–700 (2012).
20. Gruber, K. *Nat. Biotechnol.* 32, 508 (2014)