SeqAnt 2.0: Whole-Genome Annotation and Natural-Language Searching in the Cloud

Alex V. Kotlar[1], Cristina E. Trevino[1], Michael E. Zwick[1], David J. Cutler[1], and Thomas S. Wingo[1,2,3,*]

[1]Department of Human Genetics, Emory University School of Medicine

[2]Division of Neurology, Atlanta VA Medical Center.

[3]Department of Neurology, Emory University School of Medicine.

*Corresponding author:

Thomas S. Wingo

505K Whitehead Building

615 Michael Street NE

Atlanta, GA 30322-1047

404-727-4905 (office)

404-727-3728 (fax)

thomas.wingo@emory.edu

Abstract Word Count: 70

Main Text Word Count: 1369

Figure Count: 1

Table Count: 1

Abstract

Describing, prioritizing, and selecting alleles from large sequencing experiments remains technically challenging. SeqAnt 2.0 (*https://seqant.emory.edu/)* is the first web application that make these tasks accessible to non-programmers, even for large data sets consisting of thousands of whole-genome samples. It comprehensively describes alleles using public data (e.g., RefSeq, dbSNP, Clinvar, and others), and introduces a natural-language search engine that can locate the alleles of interest in a user's experiment in milliseconds.

Main

While genome-wide association studies (GWAS) and whole-exome sequencing (WES) remain important components of human disease research, the future lies in whole-genome sequencing (WGS), as it inarguably provides more complete data. The central challenge posed by WGS is one of scale. Genetic disease studies require thousands of samples to obtain adequate power, and the resulting WGS datasets are hundreds of gigabytes in size and contain tens of millions of variants. Manipulating data at this scale is difficult. To find the alleles that contribute to traits of interest, two steps must occur. First, the variants identified in a sequencing experiment need to be described in a process called annotation, and second, the relevant alleles need to be identified based on those descriptions in a procedure called variant filtering.

Annotating and filtering large numbers of variant alleles requires specialty software. Existing annotators, such as ANNOVAR[1], VEP[2], GEMINI[3], and SeqAnt 1.0[4] have played an important research role, and are sufficient for small to medium experiments (e.g.,10s to 100s of WES samples). However, they require significant computer science training to use in offline, distributed computing environments, and have substantial restrictions on the maximum size of the data they will annotate online. Existing variant filtering solutions are also limited, with complicated analyses generally requiring researchers to program custom scripts, which can result in errors that impact reproducibility[5]. Therefore, annotation and filtering are not readily accessible to most scientists, and even bioinformaticians face challenges of cost and complexity.

Here we introduce an online, cloud-based application called SeqAnt 2.0 that simplifies variant annotation and filtering. It is the first program capable of annotating sequencing experiments on the scale of thousands of whole-genome samples and tens of millions of variants in a web browser, while also integrating the first natural-language search engine that enables filtering using English phrases. SeqAnt 2.0 makes it possible to efficiently find alleles of

interest without significant computer science training, improves reproducibility, and reduces annotation and filtering costs for large experiments.

SeqAnt's annotation and filtering functions are exposed through a public web application (*https://seqant.emory.edu/)*. Creating an annotation is as simple as registering an account, selecting the genome and assembly used to make the variant call format (VCF)[6] or SNP-format[7] files, and uploading these files from a computer or Amazon S3 bucket. Annotation occurs in the cloud, where instances of the SeqAnt 2.0 annotation engine process the data and send the results back to the web application for storage and display (**Figure 1**).

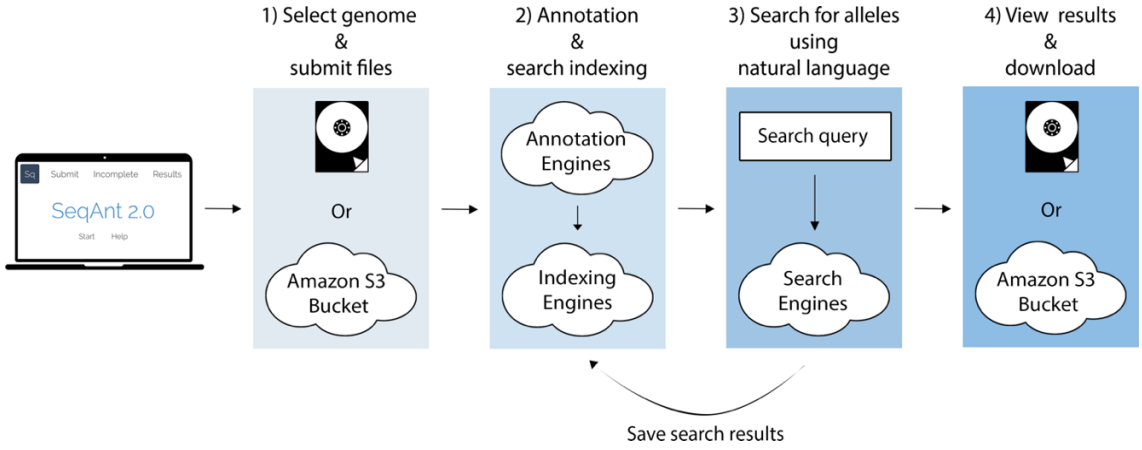

Figure 1. Using SeqAnt 2.0 online to find alleles of interest in sequencing experiments. 1) After registering an account in the SeqAnt 2.0 web application (*https://seqant.emory.edu*), users upload one of more VCF or SNP-format files containing alleles from a sequencing experiment. Data sets containing hundreds of gigabytes of data are supported. 2) The data is processed in the cloud, adding descriptions from public databases (e.g. RefSeq, dbSNP, Clinvar, and others). 3) The annotated results can be filtered using SeqAnt's natural-language search engine, and any search results can be saved. 4) Annotated experiments and saved results can be viewed online, downloaded as tab-delimited text to a local computer, or uploaded to an Amazon S3 bucket.

The SeqAnt 2.0 annotation engine is open source, and supports diverse model organisms including *homo sapiens* (hg19, hg38), *m. musculus* (mm9, mm10), *r.macaque* (rheMac8), *r.norvegicus* (rn6), *d. melanogaster* (dm6), *c. elegans* (ce11), *s. cerevisiae* (sacCer3). To annotate, it matches alleles from users' uploaded files to descriptions in RefSeq[8], dbSNP[9], PhyloP[10], PhastCons[11], Combined Annotation-Dependent Depletion (CADD)[12], and Clinvar[13] databases. It is aware of alternate splicing, and annotates all variants relative to each alternate transcript. The annotation engine is designed to scale to any size experiment, offering the performance of distributed computing solutions such as Hail[14]. To handle many large experiments online, it uses a novel architecture that automatically distributes work throughout the cloud, without user involvement.

When the web application receives a completed annotation, it saves the data and creates a permanent results page. The user may then explore several quality control metrics, including the transition to transversion ratio on a per-sample or per-experiment basis. They may also download the results as tab-delimited text to their computer or Amazon S3 bucket. In parallel with the completion of an annotation, the SeqAnt 2.0 search engine automatically begins indexing the results. Once finished, a search bar is revealed in the results page, allowing users to filter their variants using the search engine.

The SeqAnt 2.0 search engine accepts natural phrases instead of specific terms to provides flexibility similar to Google and Bing. It understands English grammar, matching phrases regardless of capitalization, punctuation, and word tense. For complex queries, it supports Boolean operators, numerical ranges, and regular expressions. The search engine has a built-in dictionary of synonymous terms, for instance equating "missense" and "non-synonymous", and allows users to define their own synonyms, which is helpful for uncovering nuanced relationships. For instance, it is possible to label trios to uncover *de novo* variants or to test allele transmission models. SeqAnt 2.0 also provides search tools, which are small programs, accessible by a single mouse click, that dynamically modify any query to generate

complex result summaries. Some of their functions include identifying compound heterozygotes and ranking genes by allele count.

Most importantly, users can save the results of any search query, which enables multi-step filtering on a single dataset. All saved results are permanent records, which promote reproduce analyses. Multi-step filtering provides functionality similar to custom processing scripts but is more accessible to a broad range of researchers. Furthermore, like the annotation engine, the search engine automatically distributes work throughout the cloud, easily handling the filtering of variants from large sequencing experiments.

To compare SeqAnt's capabilities with other programs, we first submitted 1000 Genomes Phase 1 and Phase 3 VCF files for annotation. Phase 1 contains 39.4 million variants from 1,092 WGS samples, while Phase 3 includes 84.7 million alleles from 2,504 WGS samples. When tested online, SeqAnt was found to be the only program to complete Phase 1 or Phase 3 (**Supplementary Figure 1**). When tested offline to gauge performance in the absence of web-related limitations, SeqAnt was 58x faster than ANNOVAR and 425x faster than VEP, completing Phase 3 in less than 6 hours (**Supplementary Table 1**). By contrast, ANNOVAR was unable to finish either annotation due to memory requirements, and VEP annotated Phase 3 at a rate of 10 variants per second, indicating that it would need at least 98 days to complete that data set. Critically, SeqAnt's run time grew linearly with the number of submitted genotypes, suggesting that it could handle even hundreds of thousands of samples within days. A detailed comparison of the exact settings used is given (**Supplementary Dataset 1** and **Supplementary Dataset 2**).

Next, we explored SeqAnt's ability to filter the 84.7 million annotated Phase 3 variants (**Table 1**). No other tested online program could load the data. First, we used SeqAnt's natural-language search engine to find all alleles in exonic regions by entering the term "exonic" (933,330 alleles, 0.15 seconds). The search engine calculated a transition to transversion ratio of 2.92 for the query, consistent with previously observed values in coding regions**.** To refine

results to rare, predicted deleterious alleles, we queried “cadd > 20 maf < .001 pathogenic expert review missense” (65 alleles, 0.13 ± 0.05s). This search query could be written using partial words (“pathogen”), possessive nouns (“expert’s”), different tenses (“reviews”), and synonyms (“nonsynonymous”) without changing the results. The SeqAnt search engine allows for rapid and detailed data exploration. For example, searching for “early-onset breast cancer”

Table 1 | **Seqant 2.0 online search engine performance**

Group	Search query	Time (s)	Variants	Tr:Tv
1	exonic	0.15 ± 0.04	993,330	2.92
2 (a)	cadd > 20 maf < .001 pathogenic expert review missense	0.12 ± 0.02	65	1.71
2 (b)	cadd > 20 maf < .001 pathogenic **expert’s** review **non-synonymous**	0.11 ± 0.03	65	1.71
2 (c)	cadd > 20 maf < .001 **pathogen** expert-**reviewed nonsynonymous**	0.13 ± 0.05	65	1.71
3 (a)	early onset breast cancer	0.04 ± 0.02	4,335	2.48
3 (b)	**early-onset** breast cancer	0.04 ± 0.02	4,335	2.48
3 (c)	**Early onset** breast **cancers**	0.05 ± 0.02	4,335	2.48
4 (a)	Pathogenic nonsense Ehlers-Danlos	0.05 ± 0.03	1	NA
4 (b)	**pathogenic** nonsense **eds**	0.02 ± .003	1	NA
4 (c)	**pathogenic stopgain ehler danlos**	0.04 ± 0.03	1	NA

The 1000 Genomes Phase 3 autosome and X chromosome, consisting of 8.47×10^7 variants and 2,504 samples, were filtered using the SeqAnt 2.0 natural-language search engine online. Queries in groups 2, 3, and 4 contain phrasing differences. “Time” is the number of seconds taken to return matching alleles, averaged from three consecutive repetitions of each query. “Variants” is the number of annotated sites returned from each search query. “Tr:Tv” is the transition to transversion ratio automatically calculated for each query by the search engine.

returned alleles in *BRCA1* and *BRCA2,* and querying "pathogenic nonsense eds" returned a nonsense variant in *PLOD1* that is reportedly associated with Ehlers-Danlos syndrome.

A potential limitation of SeqAnt's comparison to other software is that sophisticated users could implement distributed computing algorithms like MapReduce[15], and spread annotation workloads across multiple servers to improve performance. Additionally, we could have limited the number or types of annotations applied to the data, especially whole-genome annotation sources (e.g., CADD), to improve performance of the other software. SeqAnt demonstrates that these workarounds are unnecessary to achieve reasonable run-times for large datasets online and offline. Finally, while SeqAnt's natural-language search engine significantly reduces the difficulty of filtering annotated variants, it is potentially limited when using partial search terms that match multiple words in the query. This is easily avoided by using exact phrases (e.g., by quoting terms) or using user-specified synonyms.

To date, identifying alleles of interest in large sequencing experiments has been technically challenging. SeqAnt 2.0 simplifies this process by introducing the first online application capable of annotating large whole-genome datasets and then filtering them using a natural-language search engine. It requires no computer science experience to use and is available for many species, including *homo sapiens*, *m. musculus*, *d. melanogaster*, *c. elegans*, and *s. cerevisiae*. As sequencing experiments expand in size to and scope, SeqAnt 2.0's integrated capabilities will prove invaluable for reproducible annotation and filtering.

Author contributions

A.V.K designed, wrote, and tested SeqAnt 2.0 and performed experiments. C.E.T wrote SeqAnt 2.0 documentation and performed quality control. M.E.Z and D.J.C. contributed to the design of SeqAnt 2.0 and experiments. T.S.W. designed and wrote SeqAnt 2.0 and designed and performed experiments. A.V.K. and T.S.W. wrote the manuscript with contributions from all authors.

Acknowledgements

This work was supported by the Molecules to Mankind program (a project of the Burroughs Wellcome Fund and the Laney Graduate School at Emory University), Veterans Health Administration (BX001820), National Institutes of Health (AG025688), and Amazon AWS Educate grants. We thank Kelly Shaw and Katherine Squires for beta testing and design suggestions. We thank Viren Patel and the Emory Integrated Genomics Core (EIGC) for technical support.

References

1. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* **10**, 1556-1566 (2015).
2. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
3. DeFreitas, T., Saddiki, H. & Flaherty, P. GEMINI: a computationally-efficient search engine for large gene expression datasets. *BMC Bioinformatics* **17**, 102 (2016).
4. Shetty, A. et al. SeqAnt: A web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics* **11**, 471 (2010).
5. Sandve, G.K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten simple rules for reproducible computational research. *PLoS Comput Biol* **9**, e1003285 (2013).
6. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
7. Johnston, H.R. et al. PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc Natl Acad Sci U S A* (2017).
8. O'Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).
9. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).
10. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121 (2010).
11. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).
12. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014).

13. Landrum, M.J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862-868 (2016).

14. Ganna, A. et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat Neurosci* **19**, 1563-1565 (2016).

15. Taylor, R.C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics* **11 Suppl 12**, S1 (2010).