Bystro: Rapid online variant annotation and natural-language filtering at whole-genome scale

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

Accurately selecting relevant alleles in large sequencing experiments remains technically challenging. Bystro (*https://bystro.io/*) is the first online, cloud-based application that makes variant annotation and filtering accessible to all researchers for terabyte-sized whole-genome experiments containing thousands of samples. Its key innovation is a general-purpose, natural-language search engine that enables users to identify and export alleles of interest in milliseconds. The search engine dramatically simplifies complex filtering tasks that previously required programming experience or specialty command-line programs. Critically, Bystro annotation and filtering capabilities are orders of magnitude faster than previous solutions, saving weeks of processing time for large experiments.

Keywords

Natural-language search, genomics, bioinformatics, annotation, filtering, web, online, cloud, big data

Background

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 selected 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], SeqAnt[2], VEP[3], and GEMINI[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 in terms of performance and the maximum size of the data they will annotate online. Existing variant filtering solutions are even more limited, with most analyses 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 performance, cost and complexity.

Here we introduce an application called Bystro that significantly simplifies variant annotation and filtering, while also improving performance by orders of magnitude and saving weeks of processing time on large data sets. It is the first program capable of handling sequencing experiments on the scale of thousands of whole-genome samples and tens of millions of variants online in a web browser, and integrates the first natural-language search engine that enables real-time, complex variant filtering using simple phrases. Bystro makes it possible to efficiently find alleles of interest in any sequencing experiment without computer science training, improving reproducibility while reducing annotation and filtering costs.

Results

To compare Bystro's capabilities with other recent programs, we submitted 1000 Genomes[6] Phase 1 and Phase 3 VCF files for annotation and filtering (Figure 1). Phase 1 contains 39.4 million variants from 1,092 WGS samples, while Phase 3 includes 84.9 million alleles from 2,504 WGS samples. We first evaluated the online capabilities of the web-based versions of Bystro, wANNOVAR[7], VEP, and GEMINI (running on the Galaxy[8] platform). Bystro was the only program able to complete either 1000 Genomes Phase 1 or Phase 3 online.

When tested with small subsets of Phase 3 ($5x10^4$ – $6.4x10^6$ variants), Bystro was 131-210x faster than GEMINI/Galaxy and provided the most comprehensive whole-genome annotations (Figure 2).

Figure 1 **| Using Bystro online to find alleles of interest in sequencing experiments. A**) After logging in (*https://bystro.io/*), users upload one of more VCF or SNP-format files - containing alleles from a sequencing experiment - from a computer or a connected Amazon S3 bucket. Datasets of over 890GB, containing thousands of samples and tens of millions of variants are supported. The data is rapidly annotated in the cloud, using descriptions from public sources (e.g. RefSeq, dbSNP, Clinvar, and others). The annotated results can be filtered using Bystro's natural-language search engine, and any search results can be saved as new annotations. Annotated experiments and saved results can be viewed online, downloaded as tab-delimited text, or uploaded back to linked Amazon S3 buckets. **B**) An example of using Bystro's natural language search engine to filter 1000 Genomes Phase 3 (*https://bystro.io/public*). To do so, users may type natural phrases, specific terms, numerical ranges, or apply filters on any annotated field. Queries are flexible, allowing misspelled terms such as "earl-onset" to accurately match. Complex tasks, such as identifying *de novo* variants can be achieved by using Boolean operators (AND, OR, NOT), field filters, and user-defined synonymous terms.

Figure 2 | **Online performance comparison of Bystro, VEP, wANNOVAR, and GEMINI**. Bystro, wANNOVAR, VEP, and GEMINI (running on Galaxy) we run under similar conditions. Total processing time was recorded for 1000 Genomes Phase 3 WGS VCF files, containing either the full data set (2,504 samples, 8.49×10^7 variant sites), or subsets (2,504 samples and 5×10^4, 3×10^5, 1×10^6, and 6×10^6 variants). Only Bystro successfully processed more than 1×10^6 variants online: wANNOVAR (not shown) could not complete the smallest 5×10^4 variant subset; VEP could not complete more than 5×10^4 variants; and GEMINI/Galaxy could not complete more than 1×10^6 variants. Online, VEP outputted a restricted subset of annotation data compared to its offline version. GEMINI and Bystro (but not VEP) outputted whole-genome CADD scores, while only Bystro also returned whole-genome PhyloP and PhastCons scores. Bystro was faster than GEMINI/Galaxy by 131x-210x across all time points.

We next tested each program's offline performance on identical servers to gauge performance in the absence of web-related file-size and networking limitations. Bystro was 60x faster than ANNOVAR and 419x faster than VEP, completing Phase 3 in less than 6 hours (Table 1). Furthermore, ANNOVAR was unable to finish either Phase 1 or Phase 3 annotations due to memory requirements (exceeding 60GB of RAM), and VEP annotated Phase 3 at a rate of 10 variants per second, indicating that it would need at least 98 days to complete. Critically, Bystro'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 (Additional File 1;Additional File 2).

Table 1 | **Bystro, VEP, ANNOVAR offline command-line performance**.

Software	Dataset	Samples	Variants	Variants/s	Bystro vs
Bystro	1000G Phase 3 chr1	2504	$1x10^6$	4468 ± 70.0	-
	1000G Phase 3 chr1	2504	$2x10^6$	4481 ± 5.79	-
	1000G Phase 3 chr1	2504	$4x10^6$	4513 ± 30.2	-
	1000G Phase 3 chr1	2504	$6.5x10^6$	4278 ± 3.96	-
	1000G Phase 3	2504	$8.5x10^7$	4189 ± 91.9	-
	1000G Phase 1	1092	$3.9x10^7$	4366 ± 62.0	-
VEP	1000G Phase 3	2504	$8.5x10^7$	10.00 ± 0.00	419x
	1000G Phase 1	1092	$3.9x10^7$	18.67 ± 0.58	234x
ANNOVAR	1000G Phase 3 chr1	2504	$1x10^6$	74.67 ± 0.21	59.8x
	1000G Phase 3 chr1	2504	$2x10^6$	75.32 ± 0.06	59.5x
	1000G Phase 3 chr1	2504	$4x10^6$	75.15 ± 0.39	60.1x
	1000G Phase 3 chr1	2504	$6.5x10^6$	NA	NA
	1000G Phase 3	2504	$8.5x10^7$	NA	NA
	1000G Phase 1	1092	$3.9x10^7$	NA	NA

Bystro, VEP, and ANNOVAR were similarly configured on Amazon i3.2xlarge servers. “Dataset” refers to the VCF file used. “Variants/s” is the average of three trials. VEP performance was recorded after $2x10^5$ sites in consideration of time. In runs of $1x10^6$ or more annotated sites, VEP performance did not deviate from the $2x10^5$ value. ANNOVAR could not complete the full Phase 1, Phase 3, or Phase 3 chromosome 1 datasets due to memory limitations. Thus, ANNOVAR was compared to Bystro on subsets of 1000 Genomes Phase 3 chromosome 1.

Next, we explored the Bystro search engine’s ability to filter the 84.9 million annotated Phase 3 variants. Bystro’s search engine was unique in its natural-language capabilities, and no other tested program could handle the full Phase 3 dataset online (Figure 2). First, we used Bystro’s search engine to find all alleles in exonic regions by entering the term “exonic” (933,343 alleles, 0.030 ± .001 seconds, Table 2). The search engine calculated a transition to transversion ratio of 2.96 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, 0.029 ± 0.025s, Table 2). This search

query could be written using partial words ("pathogen"), possessive nouns ("expert's"), different tenses ("reviews"), and synonyms ("nonsynonymous") without changing the results.

Table 2 | **Online comparison of Bystro and recent programs in filtering 8.49×10^7 variants from 1000 Genomes**

Group	Search query	Time (s)	Variants	Tr:Tv
1	exonic	0.030 ± 0.030	993,343	2.96
2 (a)	cadd > 20 maf < .001 pathogenic expert review missense	0.029 ± 0.009	65	1.71
2 (b)	cadd > 20 maf < .001 pathogenic **expert's** review **non-synonymous**	0.036 ± 0.019	65	1.71
2 (c)	cadd > 20 maf < .001 **pathogen** expert-**reviewed nonsynonymous**	0.044 ± 0.025	65	1.71
3 (a)	early onset breast cancer	0.046 ± 0.029	4,335	2.51
3 (b)	**early-onset** breast cancer	0.037 ± 0.020	4,335	2.51
3 (c)	**Early onset** breast **cancers**	0.033 ± 0.015	4,335	2.51
4 (a)	Pathogenic nonsense Ehlers-Danlos	0.038 ± 0.027	1	NA
4 (b)	**pathogenic** nonsense **E.D.S**	0.078 ± 0.087	1	NA
4 (c)	**pathogenic stopgain eds**	0.040 ± 0.022	1	NA

The full 1000 Genomes Phase 3 VCF file (853GB, 8.49×10^7 variants, 2,504 samples) was filtered in the publicly-available Bystro web application using the Bystro natural-language search engine. VEP, GEMINI, and wANNOVAR (not shown) were also tested, but were unable to annotate this data set or filter it. Bystro's search engine uses a natural language parser that allows for unstructured queries: queries in groups 2, 3, and 4 show phrasing variations that did not affect results returned, as would be expected for a search engine that could handle normal language variation. "Tr:Tv" is the transition to transversion ratio automatically calculated for each query by the search engine. The transition to transversion ratio of 2.96 for the "exonic" query is close to the ~2.8-3.0 ratio expected in coding regions, suggesting that the search engine accurately identified exonic (coding) variants.

To test the search engine's ability to accurately match variants from full-text disease queries, we first searched "early-onset breast cancer", returning the expected alleles in *BRCA1* and *BRCA2* (4,335 variants, .037 ± .020s, Table 2). Notably, the queried phrase "early-onset breast cancer" did not exist within the annotation, and instead matched closely-related RefSeq transcript names, such as "Homo sapiens breast cancer 2, early onset (BRCA2), mRNA." We next explored Bystro's ability to handle synonyms and acronyms. To test the hypothesis that Bystro could interpret common ontologies, we queried "pathogenic nonsense E.D.S", where "nonsense" is a common synonym for "stopGain" (a term annotated by the Bystro annotation engine), and "E.D.S" is an acronym for "Ehlers-Danlos Syndrome". Bystro successfully parsed this query, returning a single *PLOD1* variant found in 1000 Genomes Phase 3 that introduces an early stop codon in all three of its overlapping transcripts, and which has been reported in Clinvar as "pathogenic" for "Ehlers-Danlos syndrome, type 4" (1 variant, .038s ± .027s, Table 2).

Since no other tested program could load or filter the 1000 Genomes Phase 3 VCF file online, we next compared Bystro to GEMINI (running on the Galaxy platform) on a $1x10^6$ variant subset of 1000 Genomes Phase 3. In contrast with GEMINI's structured SQL queries, which do not easily avail themselves to complex research questions, Bystro enabled shorter and more flexible searches: for instance, GEMINI returned 0 results for "impact = 'nonsynonymous_variant'", while searching for either "missense" or "nonsynonymous" in Bystro returned identical results. Critically, Bystro was also approximately 6,000x to 42,000x faster than GEMINI/Galaxy, enabling real-time filtering (Table 3).

Table 3 | **Online comparison of Bystro and GEMINI in filtering 1x10^6 variants**

Group	Program	Query	Time (s)	Variants
1	**Bystro**	**cadd > 15**	**0.003 ± 0.001**	**29,057**
1	GEMINI	SELECT * FROM variants WHERE cadd_scaled > 15	126 ± 77.9	22,063
2	**Bystro**	**maf < .001 cadd > 15 missense**	**0.006 ± 0.001**	**7,926**
2	GEMINI	SELECT * FROM variants WHERE cadd_scaled > 15 AND (aaf_exac_all < .001 OR aaf_1kg_all < .001) AND impact = 'missense_variant'	36.2 ± 5.98	5,269
3	**Bystro**	**maf < .001 cadd > 15 nonsynonymous**	**0.005 ± 0.001**	**7,926**
3	GEMINI	SELECT * FROM variants WHERE cadd_scaled > 15 AND (aaf_exac_all < .001 OR aaf_1kg_all < .001) AND impact = '**nonsynonymous_variant**'	NA	0

Bystro was compared to GEMINI (running on the Galaxy platform) in filtering the 1x10^6 variant subset of 1000 Genomes Phase 3 (the largest tested file that Galaxy could completely process). GEMINI requires structured SQL queries, while Bystro allows for unstructured, natural-language search. Time represents the number of seconds to return results, averaged from six consecutive repetitions. In queries 2 and 3, Bystro's search engine returns identical results for the synonymous terms "missense" and "nonsynonymous", despite annotating such sites only as "nonsynonymous". In contrast, GEMINI requires its domain-specific query "impact = 'missense_variant'". Comparisons between GEMINI/Galaxy and Bystro are limited, as GEMINI/Galaxy does not provide a natural-language parser, annotation field filters, an interactive result browser, a transition/transversion calculator, or the ability to filter saved search results.

Discussion

The Bystro annotation and filtering capabilities are primarily exposed through a public web application (*https://bystro.io/*), and are also available for custom, offline installation. Creating an annotation online is as simple as selecting the genome and assembly used to make the variant call format (VCF)[9] or SNP[10] format files, and uploading these files from a computer or Amazon S3 bucket, which can be easily linked to the web application. Annotation occurs in the cloud, where distributed instances of the Bystro annotation engine process the data and send the results back to the web application for storage and display (Figure 1).

The Bystro 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 rapidly matches alleles from users' submitted files to descriptions from RefSeq[11], dbSNP[12], PhyloP[13], PhastCons[13], Combined Annotation-Dependent Depletion (CADD), and Clinvar[14]. The annotation engine is aware of alternate splicing, and annotates all variants relative to each alternate transcript. In contrast with current programs that require substantial VCF file pre-processing, Bystro automatically removing low-quality sites, normalizes variant representations, splits multiallelic variants, and checks supplied reference bases for concordance with the reference assembly.

The Bystro annotation engine is designed to scale to any size experiment, offering the speed of distributed computing solutions such as Hail[15], but with less complexity. Current well-performing annotators - such as ANNOVAR and SeqAnt - load significant amounts of data into memory to improve performance. However, when these programs use multiple threads to take advantage of multicore CPUs they may exceed available memory (in some cases over 60GB), resulting in a sharp drop in performance or system crash. To solve this, Bystro annotates directly from an efficient memory-mapped database (LMDB), using only a few megabytes per thread, and because memory-mapped databases naturally lend themselves to the caching frequently accessed data, Bystro achieves most of the benefits of in-memory solutions, but

without the per-thread penalties. This approach allows Bystro to take excellent advantage of multicore CPUs, while also enabling it to perform well on inexpensive, low-memory machines. Critically, when multiple files are submitted to it simultaneously, the Bystro annotation engine can automatically distribute the work throughout the cloud (or a user-configured computer cluster), gaining additional performance by processing the files on multiple computers (Figure 1).

When the web application receives a completed annotation, it saves the data and creates a permanent results page. Users 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 upload them to any connected Amazon S3 bucket. In parallel with the completion of an annotation, the Bystro 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 (Figure 1).

Unlike existing filtering solutions, Bystro's Elasticsearch-based natural-language search engine accepts unstructured, "full-text" queries, and relies on a sophisticated language parser to match annotated variants. This allows it to offer the flexibility of modern search engines like Google and Bing, while remaining specific enough for the precise identification of alleles relevant to the research question. The Bystro search engine matches terms regardless of capitalization, punctuation, or word tense, and accurately finds partial terms within long annotation values. For complex queries, it supports Boolean operators, numerical ranges, regular expressions, Levenshtein-edit distance fuzzy matches, and prefix queries. It also has a built-in dictionary of synonyms, for instance equating "stopgain" and "nonsense".

The Bystro search engine also allows users to define their own synonymous terms. Among other uses, this make it is possible to label trios, which can be used to easily identify *de novo* variants and test allele transmission models. Bystro also provides search tools and annotation field filters, which are small programs, accessible by a single mouse click, that

dynamically modify any query to generate complex result summaries or refine search results. Some of their functions include identifying compound heterozygotes, finding per-field (such as per-gene) allele counts, and easily selecting variants that have been reported in association with a specific disease, coding consequence, or any other annotation. Like the annotation engine, the search engine is also exceptionally fast, automatically distributing indexed annotations throughout the cloud, enabling users to sift through millions of variants from large whole-genome sequencing experiments in milliseconds.

Most importantly, users can save and download the results of any search query, which enables multi-step filtering on a single dataset. The saved results are indexed for search, and hyperlinked to the annotations that they were generated from, forming permanent records that can be used to reproduce complex analyses. This multi-step filtering provides functionality similar to custom command-line filtering script pipelines, but is significantly faster, less error prone, and accessible to researchers without programming experience.

While Bystro's annotation and filtering performance is currently unparalleled by any other approach, other software (such as Hail[15]) could achieve similar performance by implementing distributed computing algorithms like MapReduce[16], and spreading annotation workloads across many servers. Bystro demonstrates that these workarounds are unnecessary to achieve reasonable run-times for large datasets online or offline. Additionally, while Bystro's natural-language search engine significantly reduces the difficulty of variant filtering, it does not handle language idiosyncrasies as robustly as more mature solutions like Google's, and may return unexpected results when search queries are very short, since such queries may have multiple correct matches. This is easily avoided by using exact phrases (e.g., by quoting terms), using user-specified synonyms, or applying field filters.

Conclusions

To date, identifying alleles of interest in sequencing experiments has been time-consuming and technically challenging, especially for whole-genome sequencing experiments. Bystro increases performance by orders of magnitude and improves ease of use through three key innovations: 1) a low-memory, high-performance, multithreaded variant annotator that automatically distributes work in cloud or clustered environments; 2) an online architecture that handles significantly larger sequencing experiments than previous solutions; and 3) the first general-purpose, natural-language search engine that simplifies complex variant filtering, and which can return matching variants from whole-genome datasets in milliseconds, enabling real-time data analysis. Bystro's features enable practically any researcher – regardless of computational experience - to analyze large sequencing experiments (e.g thousands of whole-genome samples) within less than a day, and small ones (e.g hundreds of whole-exome samples) in seconds. As genome sequencing continues the march toward ever-larger datasets and becomes more frequently used in diverse research settings, Bystro's combination of performance and ease of use will prove invaluable for reproducible, rapid research.

Methods

Accessing Bystro

For most users, we recommend the Bystro web application (*https://bystro.io*), as it gives full functionality, supports arbitrarily large datasets, and provides a convenient interface to the natural-language search engine. Users with computational experience can download the Bystro open-source package (*https://github.com/akotlar/bystro*). Using the provided installation script or Amazon AMI image, Bystro can be easily deployed on an individual computer, computational cluster, or any Amazon Web Services (AWS) EC2 instance. Bystro has very low memory and CPU requirements, but benefits from fast SSD drives. As such we recommend at AWS

instances with provisioned I/O EBS drives, RAID 0 non-provisioned EBS, or i2/i3-class EC2 instances.

Bystro comparisons with ANNOVAR, wANNOVAR, VEP, and GEMINI/Galaxy

Bystro Database

Bystro databases were created using the open-source package (*https://github.com/akotlar/bystro*). The hg19 and hg38 databases contains RefSeq, dbSNP, PhyloP, PhastCons, Combined Annotation-Dependent Depletion (CADD), and Clinvar fields, as well as custom annotations (Additional File 3). A complete listing of the original source data are enumerated in the Git repository (*https://github.com/akotlar/bystro/tree/master/config*). Other organism databases contain a subset of these sources, based on availability. Pre-built, up-to-date versions of these databases are publicly available (*https://github.com/akotlar/bystro*).

WGS Datasets

Phase 1 and Phase 3 autosome and chromosome X VCF files were downloaded from *http://www.internationalgenome.org/data/*. Phase 1 files were concatenated using bcftools[17] "concat" function. Phase 3 files were concatenated using a custom Perl script (*https://github.com/wingolab-org/GenPro/blob/master/bin/mergeSnpFiles*). The Phase 1 VCF file was 895GB (139GB compressed), and the Phase 3 data was 853GB (15.6GB compressed). The larger size of Phase 1 can be attributed to the inclusion of extra genotype information (the genotype likelihood). The full Phase 3 chromosome 1 VCF file ($6.4x10^6$ variants, 1.2GB compressed), and $5x10^4$-$4x10^6$ variant allele subsets (8-655MB compressed) were also tested. All Phase 1 and Phase 3 data correspond to the GRCh37/hg19 human genome assembly. All data used are available (Additional File 4).

Online annotation comparisons

Bystro, VEP online, wANNOVAR (the online version of ANNOVAR), and GEMINI (within Galaxy, which wraps command-line programs) were tested with the full 1000 Genomes Phase 1 and Phase 3 VCF files, unless they were unable to upload the files due to file size restrictions. Bystro was found to be the only program capable of uploading and processing the full Phase 1 and Phase 3 data sets.

To conduct Bystro online annotations, a new user was registered within the public Bystro web application (*https://bystro.io/*). Phase 1 and Phase 3 files were submitted in triplicate, one replicate at a time, using the default database configuration (Additional File 1). Indexing was automatically performed by Bystro upon completion of each annotation. The Phase 3 annotation is publicly available to be tested (*https://bistro.io/public*).

The public Bystro server was configured as an Amazon i3.4xlarge EC2 instance. The server supported 8 simultaneous users. Throughout the duration of each experiment, multiple users had concurrent access to this server, increasing experiment variance, and limiting observed performance.

Online Variant Effect Predictor (VEP) submissions were done using the VEP web application (*http://www.ensembl.org/info/docs/tools/vep/index.html*). VEP has a 50MB (compressed) file size limit. Due to gateway timeout issues and this file size limit, data sets larger than $5x10^4$ variants failed to complete (Additional File 1).

Online ANNOVAR submissions were handled using the wANNOVAR web application. wANNOVAR could not accept the smallest tested file, the $5x10^4$ variant subset of Phase 3 chromosome 1 (8MB compressed) due to file size restrictions (Additional File 1).

Galaxy submission was made using the public Galaxy servers. Galaxy provides ANNOVAR, but its version of this software failed to complete any annotations, with the error "unknown option: vcfinput". Annotations on Galaxy were therefore performed using GEMINI, which provides annotations similar to Bystro's. Galaxy has a total storage allocation of 250GB (after requisite decompression), and both Phase 1 and Phase 3 exceed this size. Galaxy was therefore tested with the full $6.4x10^6$ variant Phase 3 chromosome 1 VCF file. Galaxy's FTP server was able to upload the file, however, Galaxy was unable to load the data into GEMINI, terminating after running for 36 hours, with the message "This job was terminated because it ran longer than the maximum allowed job run time" (Additional File 1). Subsets of Phase 3 chromosome 1 containing $5x10^4$, $3x10^5$, and $1x10^6$ variants were therefore tested. Three repetitions of the $5x10^4$ variant submission were made. In consideration of the duration of execution, two repetitions were made of the $3x10^5$ and $1x10^6$ variants submissions. Since Galaxy does not record completion time, QuickTime was used to record each submission.

Variant filtering comparisons

After Bystro completed each annotation, it automatically indexed the results for search. The time taken to index this data was recorded. Once this was completed, the Bystro web application's search bar was used to filter the annotated sequencing experiments. The query time, as well as the number of results and the transition to transversion ratio for each query, were automatically generated by the search engine and recorded. Query time did not take into account network latency between the search server and the web server. All queries were run six times and averaged. The public search engine, which processed all queries, was hosted on a single Amazon i3.2xlarge EC2 instance.

Since VEP, wANNOVAR, and Galaxy/GEMINI could not complete Phase 1 or Phase 3 annotations, variant filtering on these data sets could not be attempted. For small experiments

VEP and GEMINI can filter based on exact matches, while wANNOVAR provides only pre-configured phenotype and disease model filters. VEP could annotate and filter at most only $5x10^4$ variants and was therefore excluded from query comparisons. Galaxy/GEMINI was tested with subsets of 1000 Genomes Phase 3 of $1x10^6$ variants (the largest tested data set that Galaxy could handle), with the described settings (Additional File 1). Since Galaxy does not report run times, Quicktime software was used to record each run, and the query time was calculated as the difference between the time the search submission entered the Galaxy queue, to the time that it was marked completed. Galaxy/GEMINI queries were each run more than 6 times. Because run times varied by more than 17x, the fastest consecutive 6 runs were averaged to minimize the influence of Galaxy server load.

All comparisons with the Bystro search engine are limited, because no other existing method provides natural-language parsing, and either rely on built-in scripts or require the user to learn a specific language (SQL).

Offline annotation comparisons

To generate offline performance data, Bystro, VEP, and ANNOVAR were each run on separate, dedicated Amazon i3.2xlarge EC2 instances. Each instance contained 4 CPU cores (8 threads), 60GB RAM, and a 1920GB NVMe SSD. Each instance was identically configured. All programs were configured to as closely match Bystro's output as possible, although Bystro output more total annotation fields (Additional File 2). Each data set tested was run 3 times. The annotation time for each run was recorded, and averaged to generate the mean variant per second (variant/s) performance.

VEP version 36 was used, as the most recent version at the time of writing, version 37, failed to run with the GRCh37 human assembly. VEP was configured to use 8 threads and to

run in "offline" mode to maximize performance, as recommended[3]. In each of three recorded trials, VEP was set to annotate from RefSeq and CADD, and to check the reference assembly (Additional File 2). Based on VEP's observed performance, adding PhastCons annotations was not attempted. VEP's performance was measured by reading the program's log, which records variant/second performance every $5x10^3$ annotated sites. In consideration of time, VEP was stopped after at least $2x10^5$ variants were completed, and the $2x10^5$ variants performance was recorded.

The latest ANNOVAR version at the time of writing, 2016-02-01, was used. ANNOVAR was configured to annotate RefSeq, CADD, PhastCons 100way, PhyloP 100way, Clinvar, avSNP, and ExAc version 03 (Additional File 2). ANNOVAR's avSNP database was used in place of dbSNP, as recommended. We configured ANNOVAR to report allele frequencies from ExAc, because it does not do so from either avSNP or dbSNP databases. When annotating Phase 1, Phase 3, or Phase 3 chromosome 1, ANNOVAR crashed by exceeding the available 60GB of memory. It was therefore tested with the subsets of Phase 3 chromosome 1 that contained $1x10^6$ – $4x10^6$ variants.

Bystro was configured to annotate descriptions from RefSeq, dbSNP 147, CADD, PhastCons 100way, PhyloP 100way, Clinvar, and to check the reference for each submitted genomic position (Additional File 2).

Availability of data and materials

The Bystro web application is freely accessible at *https://bystro.io/*, and features detailed interface documentation (*https://bystro.io/help*). The Bystro annotator, search indexer, distributed queue servers, and database builder are freely available in the Github repository (*https://github.com/akotlar/bystro*, DOI: *10.5281/zenodo.834960*), under the Apache 2 open-

source license. The software is written in Perl and Go programming languages and runs on Linux and Mac operating systems. Detailed documentation for Bystro software is provided at *https://github.com/akotlar/bystro/blob/master/README.md*. The datasets supporting the conclusions of this article are publicly available in the Amazon S3 repository, *https://s3.amazonaws.com/1000g-vcf/* (Additional File 4).

Additional Files

Additional File 1: Description of online comparison settings (.xlsx, 859KB)

Additional File 2: Description of online comparison settings (.xlsx, 40KB)

Additional File 3: Species supported at time of writing, and their configurations (.xslx, 36KB)

Additional File 4: URLs of 1000 Genomes Phase 1 and 3 VCF files used (.xslx, 47KB)

Declarations

Author contributions

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

Acknowledgements

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.

Funding

This work was supported by the AWS Cloud Credits for Research program, the Molecules to Mankind program (a project of the Burroughs Wellcome Fund and the Laney Graduate School

at Emory University), Veterans Health Administration (BX001820), and the National Institutes of Health (AG025688, MH101720, NS091859).

Competing interests

The authors have no competing interests to declare.

Ethics approval and consent to participate

Not applicable

References

1. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.
2. Shetty AC, Athri P, Mondal K, Horner VL, Steinberg KM, Patel V, Caspary T, Cutler DJ, Zwick ME: **SeqAnt: a web service to rapidly identify and annotate DNA sequence variations.** *BMC Bioinformatics* 2010, **11:**471.
3. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F: **The Ensembl Variant Effect Predictor.** *Genome Biol* 2016, **17:**122.
4. DeFreitas T, Saddiki H, Flaherty P: **GEMINI: a computationally-efficient search engine for large gene expression datasets.** *BMC Bioinformatics* 2016, **17:**102.
5. Sandve GK, Nekrutenko A, Taylor J, Hovig E: **Ten simple rules for reproducible computational research.** *PLoS Comput Biol* 2013, **9:**e1003285.
6. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.
7. Chang X, Wang K: **wANNOVAR: annotating genetic variants for personal genomes via the web.** *J Med Genet* 2012, **49:**433-436.

8. Goecks J, Nekrutenko A, Taylor J, Galaxy T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11:**R86.

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

10. Johnston HR, Chopra P, Wingo TS, Patel V, International Consortium on B, Behavior in 22q11.2 Deletion S, Epstein MP, Mulle JG, Warren ST, Zwick ME, Cutler DJ: **PEMapper and PECaller provide a simplified approach to whole-genome sequencing.** *Proc Natl Acad Sci U S A* 2017.

11. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res* 2016, **44:**D733-745.

12. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308-311.

13. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies.** *Genome Res* 2010, **20:**110-121.

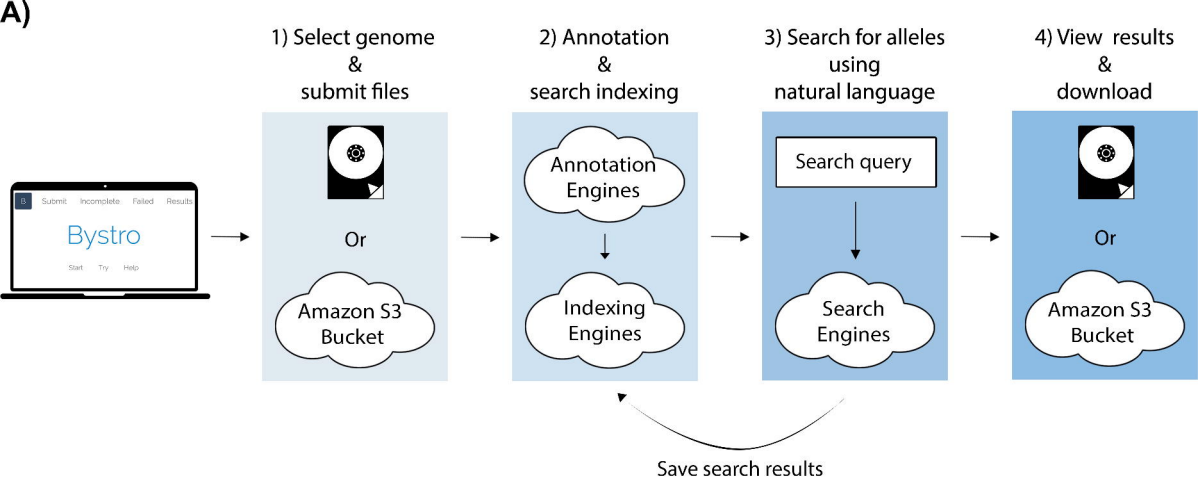
14. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al: **ClinVar: public archive of interpretations of clinically relevant variants.** *Nucleic Acids Res* 2016, **44:**D862-868.

15. Ganna A, Genovese G, Howrigan DP, Byrnes A, Kurki MI, Zekavat SM, Whelan CW, Kals M, Nivard MG, Bloemendal A, et al: **Ultra-rare disruptive and damaging mutations influence educational attainment in the general population.** *Nat Neurosci* 2016, **19:**1563-1565.

16. Taylor RC: **An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics.** *BMC Bioinformatics* 2010, **11 Suppl 12:**S1.

17. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27:**2987-2993.

A)



B)

phase3.vcf **(completed)**
Created on: **Jun 25, 2017 4:25:45 PM**
Notes: **(Click to add a note)**

Search this file
earl-onset pathogenic breast cancer

Sort Tools Size

Search Results Summary

Found **278 results** in **0.027s**
Showing page 1 (10 results per page)

Transitions & Transversions

Tr:Tv Ratio: **2.366**
Transitions: **194**
Transversions: **82**

Filter Search Results

Genes
Exonic Allele Function
RefSeq Site Type
Chromosome

Expand all

BRCA2

chr13 : 32,929,053

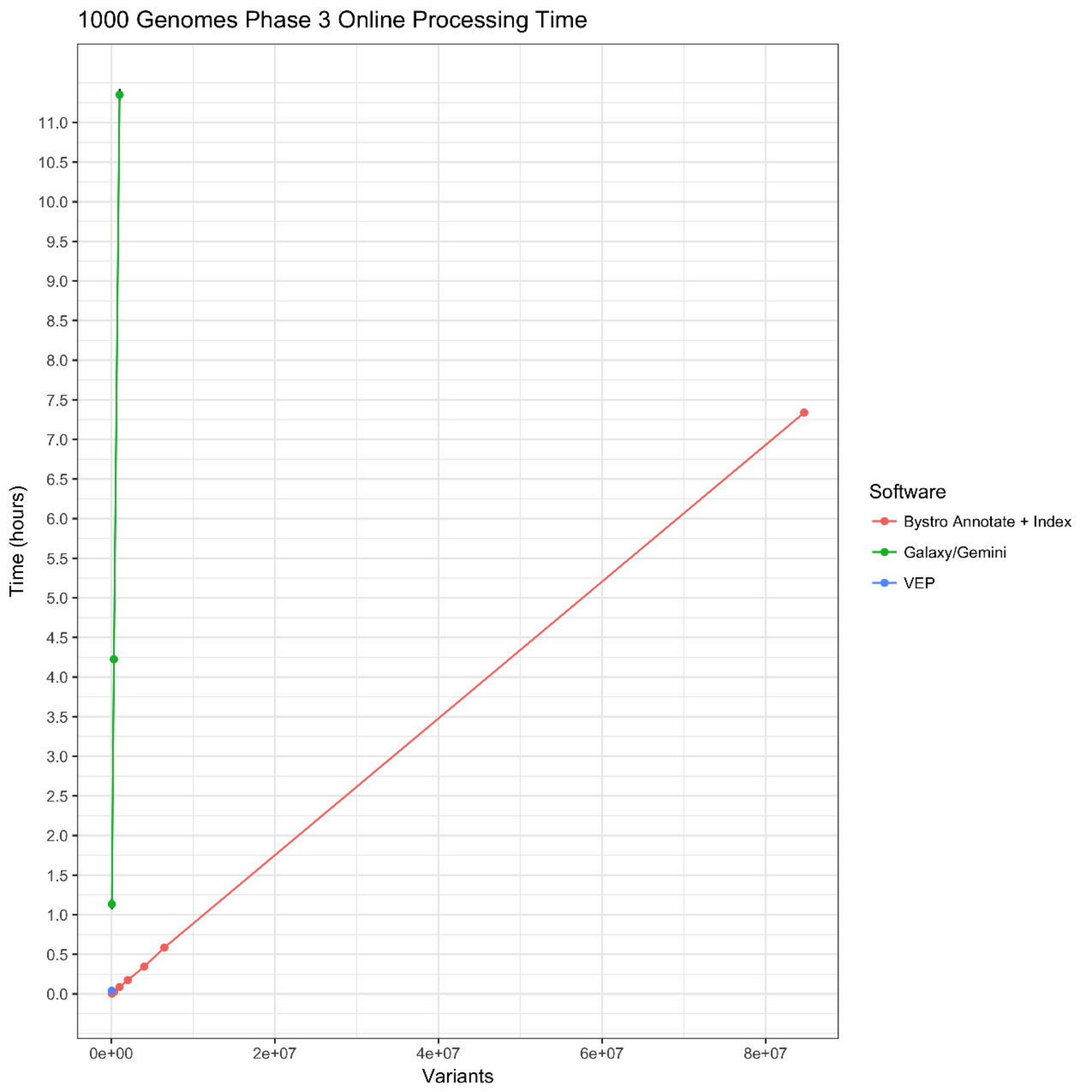
E2355*

Cadd: **43** PhyloP: **1.17** PhastCons: **0.97**

Less Detail

RefSeq Transcripts

Name: **NM_000059**
spDisplayID: **BRCA2_HUMAN**
spID: **P51587**
mRNA: **NM_000059**
protAcc: **NP_000050**
Site Type: **exonic**
Description: **Homo sapiens breast cancer 2, early onset (BRCA2), mRNA.**
Strand: **+**
Function: **stopGain**
Codon Number: **2355**


1000 Genomes Phase 3 Online Processing Time
Time (hours)
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0e+00
2e+07
4e+07
6e+07
8e+07
Variants
Software
Bystro Annotate + Index
Galaxy/Gemini
VEP