




SEAVE

*A Comprehensive Variant Filtration
Platform for Clinical Genomics*



USAGE GUIDE v1.0

Dr. Velimir Gayevskiy

*This guide is for Seave users and describes
how to filter genomic variant data
and use all of Seave's features.*

*It assumes that Seave has been installed,
has variant data loaded in it and
user accounts have been made.*

Contents

About	3
Acknowledgements	3
A Note on Importing Data	4
Logging In	4
Querying Short Variants	4
Databases List	4
Family and Analysis Selection	5
Query Page	6
Results Page	8
Modifying Pedigree Information	10
Database Variants Summary	10
Querying Long Variants	11
Overview	11
GBS Query Page	11
Gene List(s) Analysis Type	11
Sample Overlaps Analysis Type	12
Method Overlaps Analysis Type	12
Genomic Coordinates Analysis Type	13
ROHmer Analysis Type	13
SV Fusions Analysis Type	13

About

Seave is a web-based platform for genetic variant filtration, primarily for use in clinical genomics. It stores short variants in the form of single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels). Long genetic variants such as copy number variants (CNV), structural variants (SV) and losses of heterozygosity (LoH) are stored in the form of genomic blocks of any size. Variants can be filtered using a variety of parameters and the results are annotated with a large number of external annotation databases.

The development of Seave began in January 2015 to provide the Kinghorn Centre for Clinical Genomics (KCCG) at the Garvan Institute a place to store genomic data and make it accessible to clinicians and researchers without bioinformatics backgrounds. It grew out of a simple front end to GEMINI to include many advanced features. It has gone on to be used commercially within Genome.One, a spin-off from KCCG, and for a variety of germline and somatic research projects within and outside the Garvan Institute.

Seave was designed and developed by Dr. Velimir Gayevskiy. Dr. Tony Roscioli contributed early advice for improving inheritance logic and usability for clinicians. Dr. Mark Cowley has overseen development and contributed significantly to feature requests and code reviews.

Acknowledgements

We thank the Translational Genomics group (KCCG) and Genome.One for their comments and questions that lead to new features and bug fixes: Lisa Ewans, Eric Lee, Marie Wong, Kishore Kumar, Clare Puttick and André Minoche.

Seave would not be possible without the existence of GEMINI, a free academic tool created by Dr. Aaron Quinlan at the University of Utah. His work allowed us to rapidly prototype Seave and use GEMINI in any way we desire due to his generous MIT Licensing.

A Note on Importing Data

Short variants (SNVs and Indels) are stored in Seave as **GEMINI** databases. This means that you cannot import a VCF file directly into Seave, you must first load it into a **GEMINI** database. Long variants (CNVs and SVs) called by tools that Seave supports can be directly imported without modification. All data must be imported into Seave by administrator users manually or automatically as part of a bioinformatics pipeline. For more information on importing data, please see the companion Seave Administrator Guide.

Logging In




Navigate to Seave and click the “Log In” link on the top right. Enter your email address and password on the next page to log in. If you forgot your password, Seave does not currently have a forgotten password feature so you will need to contact an administrator to reset your password. Once you log in you will be automatically redirected to the home page.

Querying Short Variants

Databases List

Short variants in Seave are stored in **GEMINI** database files. After logging in, click the “Take me to the data” button on the home page or the “Databases” link on the top left menu. This will take you to a page displaying all databases that you have access to. The table lists each database on a separate row and the columns are as follows:

- **Database** – the database filename, this is typically set to an identifying feature for the database, such as a family ID or research project name.
- **Group** – the group the database belongs to; by definition, you are a member of this group, and you can use this column to quickly identify the databases in different projects that have been assigned separate groups.
- **Sample Names** – a list of all samples in the database, if you mouse over this list a tooltip will appear with all of the samples shown.
- **Samples** – the total number of samples in the database.

- **Variants** – the total number of variants in the database.
- **Size** – the file size of the database.
- **Date** – when the database was last modified (i.e. when the database was imported or when the pedigree was last changed).
- **GEMINI** – the [GEMINI](#) version used to create the database.
- **Actions** – contains icons for actions you can perform on the database:
 -  – view the variant summary report (see the [Database Variants Summary](#) section for what this includes).
 -  – view and modify pedigree information, the icon is red when no pedigree information has been set for the database and black when it has (see the [Modifying Pedigree Information](#) section).
 -  – perform a query on GBS data, this icon only appears when one or more samples in the database have data in the GBS (see the [Querying Long Variants](#) section).

To proceed, click anywhere in the row of the database you want to query.

Family and Analysis Selection

If you selected a database with pedigree information set, you will now see a page where you can select families and analysis types, otherwise proceed to the next section.

The database you selected will be displayed under the “Database selected” heading and all families specified on the pedigree page will be visible as boxes to the right of the first “Entire Dataset” box. Click the family you would like to analyse and the “Family information” section below will be automatically populated with all samples for the family, along with their gender and affected status. It is recommended to check this information prior to each query. If you would like to query the entire database, leave the “Entire Dataset” option selected and proceed to the end of this section.

The next section “Select an analysis type” allows you to optionally select an inheritance pattern for restricting your search to variants that match the pattern. This is a powerful tool if you have prior information about the sample(s) in the family, so let’s go through the assumptions and options. First of all, only the affected status of samples will be used for filtration, rather than their familial information (mother, father, child). The reason for this is that it allows you to group any combination of samples into a family where you assume their phenotype is caused by the same variant(s). This

means you can group 2 siblings into a family, a mother/proband, include an affected uncle or even just create a family with an affected singleton. Next, each of the analysis types corresponds to a well-known mechanism of genetic inheritance of variation, but the search is performed in such a way that only affected status matters. For example, the “Heterozygous Dominant” analysis will return all variants where affected individuals are heterozygous and unaffected individuals are not (i.e. can be homozygous reference or homozygous recessive). The heterozygous state is assumed to be disease-causing for this analysis type, it would therefore be wrong to run it on a family where there are 2 unaffected parents and an affected proband. You must use your own knowledge of genetics and inheritance when selecting analysis types, **Seave will not stop you from running an analysis that doesn’t make biological sense**. For more information about the filtration logic for each analysis type see the sidebar on the right of this page, or for even more examples see the “Familial Filters” page that can be found on the top menu.

This next section titled “Return variant information for” will only appear if there is more than one family in the database. It lets you set whether you would like to see variant information for just the family selected or for all samples in the database. The information that will be restricted is sample-specific and includes things like genotype, read depth, VAF, quality, etc.

Click “Proceed to query options” when you have selected what you would like to query.

Query Page

The query page allows you to specify restrictions on variants that will be returned from the database you selected. The first 2 headings at the top show you the database and family you selected (if any).

The “Inclusion genomic location(s)” and “Exclusion genomic location(s)” sections allow you to restrict the genomic search space for which variants will be returned. Options selected in the inclusion section will mean that only variants in the regions/genes selected will be returned whereas options in the exclusion section will mean everything but the selected regions/genes will be returned. They can be used in conjunction with one another and exclusion locations will override inclusion locations. Within each section you can first enter genomic coordinates, these must be in the format of chromosome:start-end and multiple entries must be separated by semi-colons (e.g. chr1:5412543-6232356;chr3:23567-12367437). The next selection is for gene lists, these must be created and managed by Seave administrators and will automatically appear here when added. The name of the list along with the number of genes in it is displayed. Multiple lists

can be selected by Control/Command/Shift clicking multiple entries in the list. Finally, custom gene names can be supplied as a semicolon, comma or space-separated list (e.g. BRCA1;PIK3CA;TP53). If you want to search the whole genome, don't select or enter anything in these sections.

The "Impact" section contains restrictions on the way the variants returned impact on genes. The most common restriction here is the "High & Medium Impact" selection which will disregard all low impact variants. For a list of impacts and their classifications, see here: http://gemini.readthedocs.io/en/latest/content/database_schema.html. Broadly speaking, the vast majority (>99%) of variants identified using whole genome sequencing have a low impact on coding genes and are safely discarded when looking for pathogenic variants. Your mileage may vary depending on your questions, sequencing technology and application so don't treat this as law! Other options for impact include the various typical pathogenicity classification categories such as loss of function and coding variants only. Note that if you select an analysis type on the previous page, you will not be able to search "Coding" or "All impacts" categories due to GEMINI speed limitations. Finally, the "Minimum scaled CADD score" slider below will restrict variants by their scaled CADD score, as annotated by GEMINI itself. A value of 0 on this slider will not use the CADD score as a filter. For more information on CADD scores see here: <http://cadd.gs.washington.edu>.

In the next "Prevalence" section, the maximum prevalence of variants in various frequency control databases can be selected. The three databases of 1000 Genomes, ESP and ExAC are available and a percentage between 0% and 10% can be selected using the sliders. A value of 0% will mean that the database won't be used to restrict the variants returned. As for recommendations, in a rare disease setting, a value of 1% can be used conservatively, whereas in a cancer predisposition setting the value may have to be increased to 5% and for somatic variants a setting of 0.1% may be appropriate.

The "Quality" section allows restriction by sequencing quality parameters. The "Minimum sequencing depth in all samples queried" slider will discard all variants where the sequencing depth in **one or more** samples being queried (i.e. in the family selected or across the whole database depending on settings) is below the value specified. This can be a dangerous option to use as just by chance one of the samples queried could have a lower than expected depth where the others are adequate but the variant would be discarded nonetheless. The "Minimum variant quality" slider will discard variants whose variant quality is below the value in the slider, this quality is from the QUAL column in the VCF used to make the GEMINI database. Finally, the "Exclude Failed Variants" option is enabled by default and will discard all variants that are not PASS in the FILTER column of the VCF used

to make the GEMINI database.

The “Variant type(s)” section allows you to return only SNVs or Indels, but by default Seave will search across both. The “Maximum number of variants to return” section will restrict how many variants will be returned for your query. This can not be set above 2,000 as any number larger than this would place significant stress on your web browser due to the sheer volume of data returned and may also cause Seave to take too long to perform the query. Instead, try to limit your search space and parameters in such a way that you don’t get more than 2,000 variants.

After you have specified your query, click the “Launch query” button to start your query. The time to run it will depend on the size of your database and the complexity of your query. Queries typically take seconds but can take up to a minute if the database contains more than 50 whole genomes of variants. While the query is running, Seave will be unusable for you, just wait for it to take you to the results page without refreshing or trying to click anything else on the page.

Results Page

The results page displays variants returned for the query you set up on the previous pages. The focus of the page is the table containing the variants and, by default, this table will include columns summarising the most important variant information including genomic location, quality, gene impacted, variant type (e.g. SNV, insertion, deletion), impact type (e.g. missense, frameshift), allele frequency in the MGRB control database and an impact summary. The impact summary consists of a number of small boxes always ordered the same way, one per variant annotation. Mouse over each box to see a tool tip of the annotation whose impact is shown. A red box means that the variant or gene it impacts are present in the annotation and are marked as damaging, a white box means it is in the annotation but marked as benign and a grey box means it is not in the annotation.

Make sure to watch out for a warning above the results table indicating that the maximum number of variants you selected to return have been returned. This will usually mean your query produced more variants than your maximum allows, meaning you are missing variants if you ignore it. The simplest solution is to go back and increase the number of variants to return, however, if this is not possible, you may have to restrict your query further or run multiple queries with subsets of the genome.

The results table displays 10 variants at a time by default, but you can change this with the dropdown on the top left of the table. You are able to search the results for any number of keywords using the box on the top

right of the results table. This search will be performed across all table data, not just the columns visible by default.

Any time one of the row entries in the results table is underlined, you can click the text in that cell to be taken to that annotation's source. For example, clicking the row text in the column housing genomic coordinates in the format of `chr1:g.13328864A>G` will take you to the UCSC Genome Browser for those coordinates, allowing you to see the genomic context of each variant. Any annotation for which there is a web entry will have a link.

Under the results table you will find a number of boxes with annotation names, such as "Genotypes", "OMIM" and "CADD Scores". Some of these are already selected by default while others can be clicked to select them. Doing so will dynamically add the columns for the annotation to the results table. You can show as many annotations at a time as you'd like, although the table will quickly fill up with columns if too many are shown at once. The width of the results table can be increased by clicking "Increase/decrease table width" below the table which will give you some more room to play with. To see more information about annotations and their current versions used by Seave, click "Data Sources" on the top menu.



Typically, the results table is useful to quickly explore a genome, but when you need to see all of the variant annotations and start further prioritising variants, it is useful to export the variants to a spreadsheet. To enable this, Seave lets you download the full results table containing all variants and annotations by clicking "Download query results (.tsv format)" under the results table. You may need to right click it and select Download if clicking the text does not lead to a file download. The .tsv file can be opened in any spreadsheet software such as Microsoft Excel for further exploration and filtration. A useful trick for further sub-filtering the results in Excel is to use the data filter feature to deselect values you are not interested in per annotation column to dynamically remove them from the spreadsheet.

Once you are finished on the results page, if you need to go back and modify your query, click the "Back to query options" button at the bottom of the page. This will take you back to the query page and all of the query parameters will be retained as they were run for your previous query. This lets you rapidly change one parameter while keeping everything else the same. Alternatively, if you need to go back and change the analysis type, click "Back to family and analysis selection" at the bottom of the page. Even if you change the family or analysis type, your query options will still be kept, allowing you to rapidly query multiple families and/or analysis types with the same query settings.

To share your results with someone without access to Seave, you can either


send them the results .tsv file or send them the full URL from the results page. This will allow them to see the same results page and interact with the variants using the results table, but they will not be able to query the database themselves.

Modifying Pedigree Information

Clicking the / icon for a database on the Databases page takes you to a page that allows you to modify the pedigree information for that database. If the icon is red, it means no pedigree information has been set yet for the database, a black icon means it has. Either way, the modify pedigree page lists all samples in the database and allows you to change their family, phenotype and gender. The sample names cannot be changed, but you can group multiple samples into the same family by giving them the same family name. Within each family, the affected status you select can later be used to find variants matching specific inheritance patterns (see the [Family and Analysis Selection](#) section for more information). Gender does not affect Seave queries but is a useful quality control metric to later check the family composition on the family selection page. After you have finished editing the pedigree information, click “Modify pedigree” to save it to the database.

You can download the current pedigree for the database as a .ped file from the modify pedigree page by clicking the “Download current pedigree as PED file” button below all the samples. This file is useful for some downstream tools, or for a Seave administrator to reapply a pedigree setting to a new version of a database.


Database Variants Summary

The database summary page is reached by clicking the  icon from the Databases page. This page contains a table that shows a breakdown of the total number of variants (at the top) into sub-categories depending on their impact. Categories are not mutually exclusive but do provide a useful summary of how variants are distributed within the database, mostly for publication or reporting purposes. Two columns of counts are shown, one for total variant counts (“All Variants”) and another for infrequently observed variant counts (“Rare Variants”). To make it into the latter, the variant must be present at a frequency of 1% or less in ExAC, ESP and 1000 Genomes, or not present at all.

Querying Long Variants

Overview

Long variants such as copy number variants, structural variants and regions of homozygosity are stored within Seave's Genome Block Store (GBS). Each "block" is a discrete event (variant) with a set of coordinates (e.g. chr1:10000-20000), an event type (e.g. duplication), a copy number (e.g. CN=6), the method used to call it (e.g. CNVnator) and any annotations the caller made to the block. A block can have any number of annotations, these are the various important pieces of information a tool will produce to describe and annotate a block. Examples of annotations include the sequencing depth of the block, whether it is rare and whether it passes quality filters.

Block variants are stored by sample name and the method used to call them and are automatically linked to short variants where an existing [GEMINI](#) database contains a sample name with GBS data. If this occurs, a GBS icon () is displayed in the "Actions" column on the databases page. Clicking this icon takes you to the GBS query page for the database.

Please note that the GBS is the newest feature in Seave, and as such, is the least developed with regards to ease of use and data presentation.

GBS Query Page

The GBS query page is similar to the [Family and Analysis Selection](#) page. It first shows you the [GEMINI](#) database you selected for query from the Databases page. You can then select a family within the database for restricting your search to samples within that family (families are set on the [Modifying Pedigree Information](#) page) or select "Entire dataset" to query across all samples with GBS data. If you select a family, you will see the sample names, affected statuses and GBS methods for the samples within it. Some analyses can only be conducted when specific types of GBS data have been imported, this will be indicated by the analysis type box being red and unselectable.

Below is a description of each analysis type and the results you can expect.

Gene List(s) Analysis Type

This analysis type allows you to select one or more gene lists and/or enter custom genes to determine whether any genes are within (i.e. overlap with) long variants stored in the GBS. An overlap is defined as an event where a

GBS block shares one or more bases with one of the genes you selected. For example, a duplication of a whole chromosome will mean that any gene you search for on that chromosome will be returned as a separate row, all referencing the same duplication block event.

As the genes are the focus of this analysis type, a row will be returned for **each** gene-block overlap. Results will also contain the sample name that the block was called for, the event type (e.g. deletion, duplication), the method used to call the block, the block's coordinates, the block size in basepairs and the predicted copy number (if relevant). If you download the query results as a .tsv file by clicking "Download query results (.tsv format)" under the table, you will also get all annotations for each block. These are very useful to determine if the block is real, such as whether it passed quality filters and was found with a high enough level of support.

Sample Overlaps Analysis Type

The sample overlaps analysis is only available when there are 2 or more samples with GBS data *from the same method* for the family/database selected. This analysis will find all blocks where the coordinates of a block for one sample called by a method overlap by one or more bases with the coordinates of a block for another sample for the same method. If the block for one sample is large (e.g. a whole chromosome), and the other sample contains a number of blocks within that chromosome, each pair will be returned. As such, this analysis can return large amounts of rows and should be used sparingly.

As the overlapping blocks must come from the same method, this is the first column returned in the results. Next is the pair of samples that share the overlapping blocks, the genomic coordinates of overlap, the overlap size in basepairs and then all of the information about the 2 overlapping blocks as a concatenated string.

Method Overlaps Analysis Type

This analysis type is very similar to the [Sample Overlaps Analysis Type](#) above. The difference being this analysis can be run when one or more samples in the family selected have 2 or more methods. It will find all overlapping blocks *for a sample* where a block for one method overlaps by one or more bases with a block from another method. This makes this method useful for determining where multiple tools call the same regions, but be careful as there can be quite a few pairwise overlaps between two methods called across a whole genome.

As the overlapping blocks must come from the same sample, this is the first column returned in the results. Next is the pair of methods that share the overlapping blocks, the genomic coordinates of overlap, the overlap size in basepairs and then all of the information about the 2 overlapping blocks as a concatenated string.

Genomic Coordinates Analysis Type

This analysis type allows you to enter in semicolon-delimited genomic coordinate sets in the format of, e.g. chr1:10000-20000;chr2:30000-60000, and find all blocks that overlap these coordinates by one or more bases. The query coordinates are the key to the search, so the results first show the query coordinates followed by the sample, event type, method, block coordinates, block size, copy number and block annotations.

ROHmer Analysis Type

This analysis is for regions of homozygosity called by the method ROHmer. It will only be available if one or more samples in the family selected has blocks from this method. The goal of the analysis is to determine regions of homozygosity in the genome that are shared by affected individuals and not present in any unaffected individual. This is typically performed when a family is known to be consanguineous to determine regions where to look for homozygous variants causing the phenotype.

The results are organised where each row is a separate set of coordinates shared by affected individuals but not present in unaffected individuals. Following this is the size of this overlap and then a column for each affected sample with their contribution to the overlap.

SV Fusions Analysis Type

Structural variants such as inversions and translocations are stored as 2 linked blocks in the GBS, one for each break point. The assumption is that the internal genomic code is not affected by the event, but only the break points themselves. This analysis will return all structural variant blocks where one or both of the break points of the event are inside a gene. You can select one or more gene lists and enter a custom gene list to restrict your search to specific genes, or leave the query options blank to search the entire genome.

Results first show the event type, then information for the first break point block ("Block1 ..." columns), followed by information for the second break

point block ("Block2 ..." columns). The gene columns let you quickly see which gene or genes have been affected by the event and the annotations for each block can be found in the .tsv download file for determining whether the event is real.