VariantGrid: Drag and Drop Variant Analysis

David M. Lawrence[1,2], Jinghua Feng[2], Andreas W. Schreiber[1,2], Joel Geoghegan[1]

[1] ACRF Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology, Adelaide, South Australia

[2]School of Molecular and Biomedical Science, University of Adelaide, South Australia

Variants Are Important

Sequencing and variant calling is an increasingly popular means of detecting individual genetic variation, promising to deliver understanding, diagnosis and treatment of many genetic diseases such as inherited disorders and cancer. Variant calling is a leading driver of diagnostic adoption of high throughput sequencing.

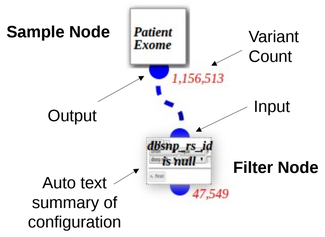
Problem

Individual variation ensures hundreds of thousands to millions of real variants in an exome/genome. Almost all are unrelated to a disorder.

Sorting through this requires expert knowledge of genes, pathways and phenotypes as well as the ability to manage enormous data sets. These skills are rarely found in the same person.

Solution

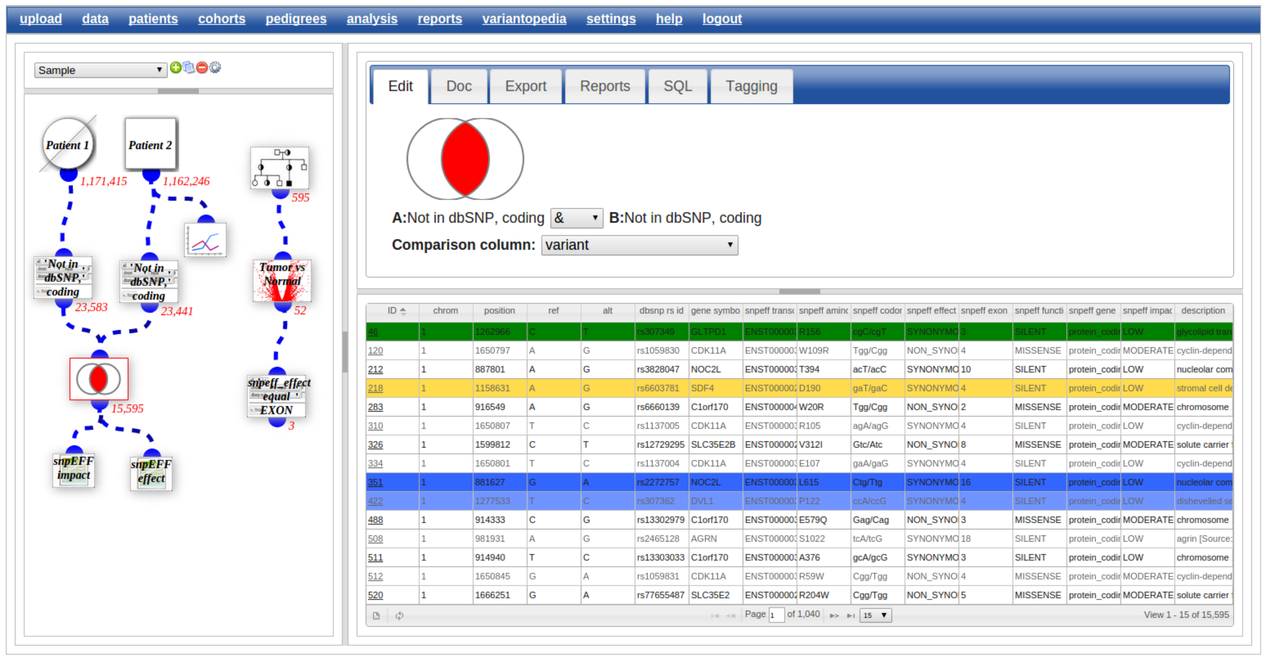
We have built a new user interface specifically designed to allow non-programming experts to search through millions of variants.



Our drag & drop interface allows users to create a Directed Acyclic Graph of variant filtering and analysis operations.

Nodes change to reflect configuration and results, allowing a high level view of the analysis.

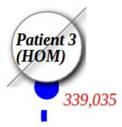
The interface encourages exploration, with any changes cascading down through pipelines automatically. This is handled by a scalable backend which sorts dependencies into parallelisable tasks, distributing them to workers across virtual machines.



VariantGrid analysis. Left panel: Drag and drop analysis nodes representing a custom analysis. Selected node in red. Top right panel, configuration for selected node (Venn Intersection). Bottom right, variants from selected node.

Source Nodes

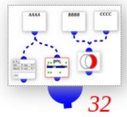
These nodes take no input, and represent a source of variants.



Sample

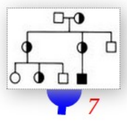
A sample, usually a single genotype (patient, cell or organism) with a set of variants. The selected sample can also be filtered by zygosity.

Linked Analysis



Connect to a previous analysis. This allows reuse of existing pipelines, encouraging standardisation and hiding complexity.

Pedigree



A collection of family samples, filtered by genotype according to inheritance models. Pedigrees can be from imported .ped file, or built on the Pedigrees page.

Cohort

A collection of related samples, eg "control group" or "poor responders" Variants are counted across the samples, allowing filters such as “show the variants which appear in at at least 10 out of 15 samples”

Result Nodes

Processing variants into a result, these nodes have no output.

Column Summary

Retrieve counts and percentages for all unique values for a user selected column

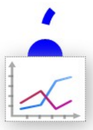
	Counts	Percent
protein_coding	2278	31.904762
	1877	26.288515
processed_transcript	1038	14.537815
nonsense_mediated_decay	667	9.341737
retained_intron	422	5.910364
unprocessed_pseudogene	286	4.005602
processed_pseudogene	244	3.417367
lincRNA	89	1.246499
miRNA	58	0.812325

SNP Matrix

	A	C	G	T
A	0.0	4.1	17.0	3.1
C	4.2	0.0	4.7	17.0
G	16.9	4.7	0.0	4.2
T	3.1	16.9	4.1	0.0

Calculate a matrix of changes observed in SNPs. This can be viewed as counts or a percentage.

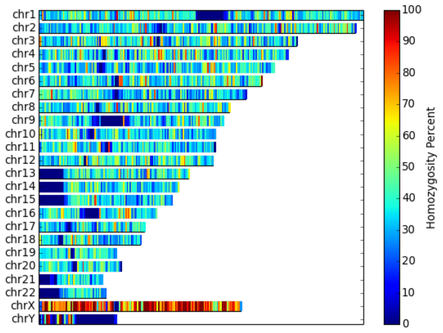
Graph Nodes



Use variants as input for a user defined graph.

SNP density plots can help identify mutation hot spots, and the example to the right shows homozygous regions.

Can you identify the sex of this patient?



Filter Nodes

Variant filtering operations, with both input and output.

Venn Intersection

Filter based on set intersections operations between parent nodes. For instance the image on the left represents “in the right parent, but not the left”. Set intersections can be made for variants, genes and other attributes.

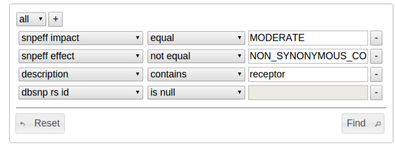
Gene List

Filter to variants where either transcript ID or gene symbol match a gene list. Lists can be uploaded on the upload page.

Gene lists can be automatically applied on a per user level, to enforce patient confidentiality and prevent coincidental findings.

Column Filter

Construct filters based on values from multiple columns.

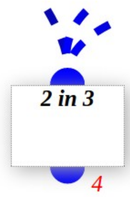


Tag Filter

Filter to variants that have had certain tags applied to them by users.

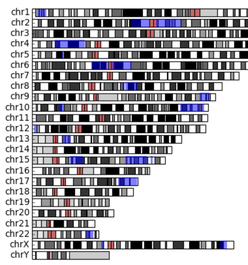
Ontologies can be created, with colours assigned. Variants can then be tagged inside the grid in a structured way supporting searching and collaboration.

Common Count



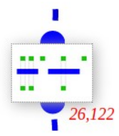
Filter variants to those that appear in at least N of M parents.

Genomic Intervals

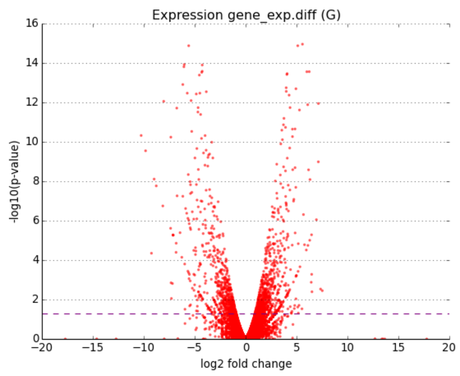


Filter based on intersection with genomic ranges.

Bed files can be uploaded on the “upload” page.



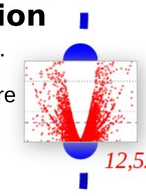
Gene Expression



Filter variants by uploaded gene or transcript expression.

An example is a minimum cutoff to exclude genes that are not expressed in the tissue of interest.

Differential expression can also be used, for example to find genes that are upregulated in tumors.



variantgrid.com
david.lawrence@health.sa.gov.au

Centre for Cancer Biology · SA Pathology · The University of Adelaide · Australian Cancer Research Foundation · Government of South Australia · SA Health