7 Steps for Human Variant Analysis

(and How to Automate Them!)

DNASTAR

In partnership with GENOMENON®

ABOUT DNASTAR

DNASTAR is a pioneer in the field of bioinformatics, offering comprehensive software solutions for molecular biology, genomics, transcriptomics, and protein analysis. Our Lasergene Genomics software enables you to set up complex genomic sequencing projects in mere minutes and automates tasks that typically require extensive manual intervention in other software packages.

About Genomenon

Genomenon is an AI-driven genomics company that organizes the world's genomic knowledge to connect patient DNA to scientific research in the diagnosis and development of treatments for patients with rare genetic diseases and cancer. Genomenon's Mastermind Genomic Search Engine powers variant interpretation by reducing turnaround time, increased diagnostic yield, and lowering the risk of missing a patient diagnosis.

CONTENTS

Introduction

In this ebook, we discuss some of the challenges involved in human variant analysis and explore some of the solutions available for addressing these challenges. Whether you are working with a core facility, a bioinformatics group, or doing variant analysis on your own, this ebook will help you learn about important considerations to keep in mind throughout the process.

Why study variants?

A common workflow in the study of human genetic variation involves the analysis and identification of deleterious variants or of variants associated with a particular population or trait. There are thousands of known variants that cause Mendelian disorders, and thousands more whose molecular basis is yet unknown.

In a typical human variant analysis study, the researcher's goal is to identify which single-nucleotide polymorphisms (SNPs), small insertions and deletions (INDELs), copy number variations (CNVs), or other types of structural variations and rearrangements (SVs) have functional significance. Functionally significant variants are those that cause amino acid changes, abnormal exon splicing, or other protein structure changes that contribute to a diseased state.

Challenges in variant analysis

The path from raw DNA reads to an understanding of the clinical significance of their variants can be long and complex. Each step and application will affect the accuracy and completeness of the results. An additional complication is that the large number of tools available have led to a myriad of pipelines. Just analyzing one data set can require mastery of up to a dozen bioinformatics applications and online databases.

Another challenge is that many open-source tools are created and then abandoned as graduate students finish their bioinformatics dissertations or laboratories lose funding or change focus. Few of these tools provide comprehensive documentation or instruction.

Even supposedly "automated" open source and commercial pipelines may rely on command line utilities at some of the key steps or require you to navigate through multiple "wizards" at each step in the pipeline. These additional steps generate multiple intermediate data files and increase hands-on time and clock time from raw data to the data analysis steps. It is therefore important to assess your ability and willingness to enter a pipeline that may require the use of command line tools and at least intermediate knowledge of bioinformatics.

Chapter 1

Step-by-Step Variant Analysis Procedure

There is no universally agreed upon number of steps involved in variant analysis, and some projects require more steps than others. In this ebook, we have divided the bioinformatics portion of the workflow into **seven steps** based on when you would typically need to switch from one tool to the next to keep advancing through the pipeline.

If you have bioinformatics expertise, you can build these steps into an automated or semi-automated pipeline using scripting and open-source tools. Note that some commercial variant analysis software packages combine two or more of these steps into a single process. If you are working with a core facility, they may have their own pipeline that covers some or all of these steps before delivering the data to you.

If you are new to variant analysis, be sure to do your research and have a plan in place to make sure that the output from one step is readable by the next tool along the path. If a core facility or bioinformatics group is doing some of this work for you, you'll want to understand what data they will be delivering to you so you can plan your next steps accordingly.

Variant Analysis Overview

1. Import & clean up sequencing data
2. Align reads to a reference genome
3. Remove PCR duplicates
4. Call variants and INDELS
5. Filter data to discover important variants
6. Determine how variants affect genes
7. Visualize the impact of the variant on the 3D protein structure

Preliminary Steps

Choose a sequencing strategy and prepare samples

While this ebook focuses primarily on the bioinformatics and data analysis steps involved in studying human variants, there are many important decisions you will need to make before you can dive into the data. How you design and execute your sequencing strategy will greatly impact the quality of your data, so be sure to consider your analysis goals before beginning your experiment.

Before obtaining DNA samples, first determine if you will be doing a germline or somatic analysis. Somatic analysis compares samples from related tissues from a single individual and is often used in cancer research. Germline analysis, by contrast, compares samples from different individuals that may be heterozygous or homozygous for a trait.

Another consideration is whether you will be doing whole-genome, whole-exome, or RNA sequencing. All of these can yield important results and help elucidate the genetic cause of a disease. RNA-Seq is useful when you want to study differences in gene expression rather than simply the presence or absence of a SNP in the DNA sequence. Whole-genome sequencing (WGS) is not typically used in the case of human subjects. One reason is that it is prohibitively expensive for most researchers, although costs have come down substantially. Another concern is that the significance of non-coding variants is typically much more difficult to determine than coding changes and their abundance (~400x more than in coding regions) makes data handling and analysis more cumbersome and inefficient. Technological improvements and cost reductions are making WGS a more viable option when needed.

Because of the issues mentioned here, whole-exome sequencing, which considers only the ~1-2% of DNA that codes for proteins, is much more common than whole-genome sequencing for variant analysis studies. One caveat to be aware of is that exome capturing kits may miss some regions of interest, potentially missing important variants.

Once you have obtained the appropriate human DNA, each sequencing platform (e.g., Sanger, Illumina, Ion Torrent, PacBio, Oxford Nanopore), will have their own wet-lab protocols for sample preparation specific to the technology. For example, short read sample preparation typically requires fragmentation and PCR amplification steps, while long read preparation avoids DNA fragmentation at all costs. If you would like to enhance mapping accuracy and identify structural rearrangements with more clarity, consider a technology that produces long reads, paired-end, or mate-pair reads.

The final preliminary step is the actual sequencing, which is typically performed at a dedicated sequencing facility.

Bioinformatics Steps

After sequencing is complete and you have sequence data in FASTQ format or variant call format (VCF), you are ready to start your analysis. The remaining steps all come down to your bioinformatics software and your ability to utilize it. Keep in mind that some of these steps may have been completed before you receive your data.

Step 1: Import & clean up sequencing data

If your data files are in FASTQ format, this step is where you'll jump in and upload the file(s) into a software application for assessment and possible clean-up.

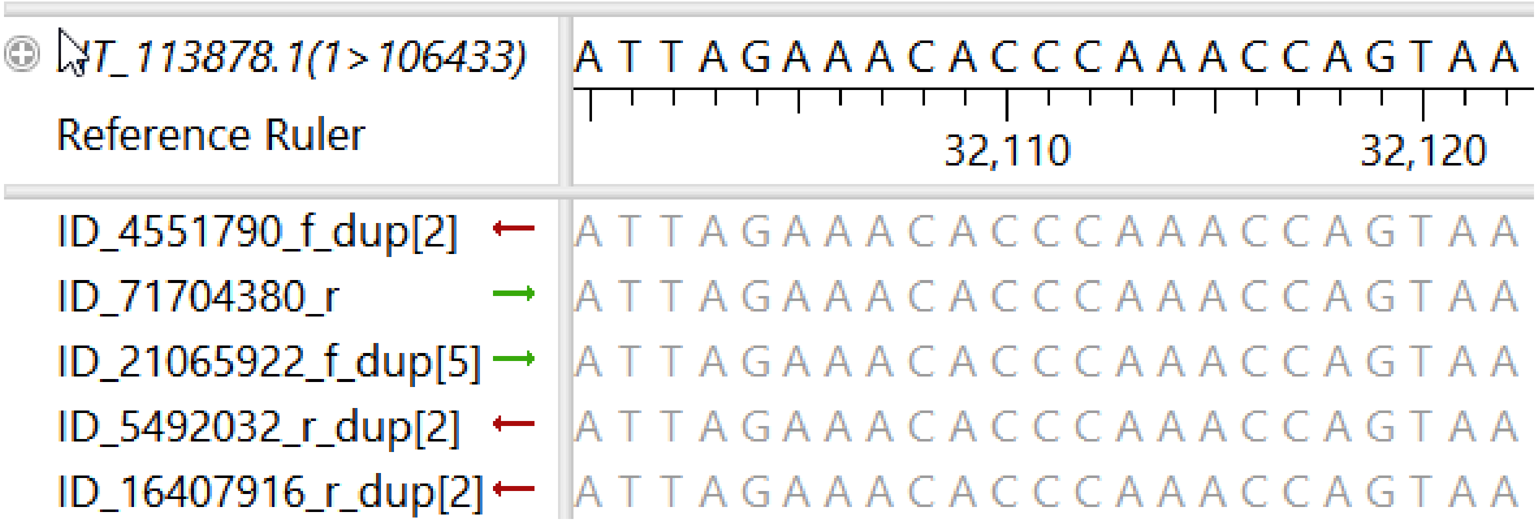
FASTQ files? Here's your starting point

NGS data files (Illumina and Ion Torrent) are typically cleaned up using the pipeline tools associated with the sequencing instrument. This is normally sufficient. However, some output sequence files can benefit from scanning with a third party tool like FastQC.

Sanger data, by contrast, is not typically cleaned up during sequencing. Sanger data usually contains many base calling errors at the 5' and 3' ends where the chromatogram peaks are not high quality. This type of data requires a high-quality software program that can accurately trim the sequence ends,

Step 2: Align reads to a reference genome

This step involves using a computer program to align/map the read sequences to an existing reference genome; this is usually followed by a local realignment. The two most widely used public sources for human reference genomes are the Genome Reference Consortium (GRC), which provides GRCh36 through GRCh38, and the University of Santa Cruz (UCSC), which provides versions hg18 and hg19.



```
NT_113878.1(1>106433)       A T T A G A A A C A C C C A A A C C A G T A A
Reference Ruler                            32,110                   32,120
ID_4551790_f_dup[2]   ←     A T T A G A A A C A C C C A A A C C A G T A A
ID_71704380_r         →     A T T A G A A A C A C C C A A A C C A G T A A
ID_21065922_f_dup[5]  →     A T T A G A A A C A C C C A A A C C A G T A A
ID_5492032_r_dup[2]   ←     A T T A G A A A C A C C C A A A C C A G T A A
ID_16407916_r_dup[2]  ←     A T T A G A A A C A C C C A A A C C A G T A A
```

Some popular tools used for alignment include SOAP, Bowtie/Bowtie2, BWA, and MOSAIK. The results of the alignment are generally saved as BAM or CRAM formatted files.

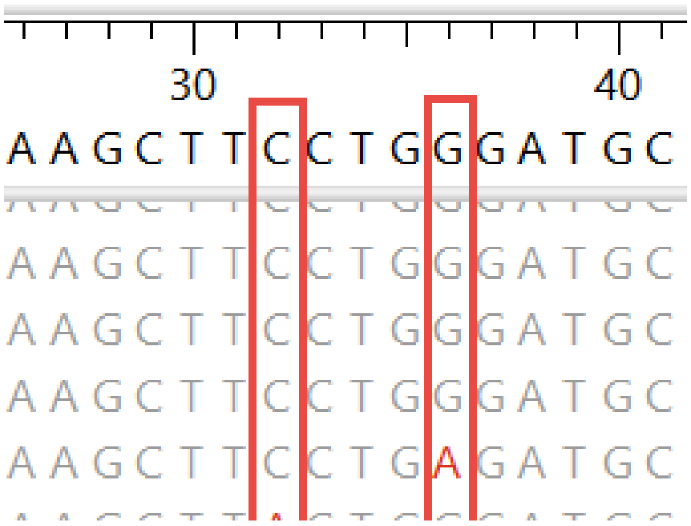
Step 3: Remove PCR duplicates

When performing whole-genome or whole-exome sequencing, PCR duplicates should be removed immediately after the alignment step. This prevents duplicate reads that originate from a single template from interfering with later variant calling statistics. Commercial solutions may automatically detect and remove duplicate reads from the variant caller. One tool that can be used for this purpose is Picard Tools, though it works only with BAM files.

Step 4: Call variants and INDELS

Variant calling is the process of comparing the aligned read sequences to the reference sequence in order to find locations where they disagree. The four main types of variants are:

- Single-nucleotide polymorphisms (SNPs)
- Small insertions and deletions (INDELs)
- Larger indels, known as "structural variations" (SVs)
- Copy number variations (CNVs), in which a section of DNA is repeated multiple times



Scores of variant callers are available and are generally classified into four groups:

- **Germline callers** (including CRISP, SAMtools, GATK), typically used for elucidating the causes of rare diseases.
- **Somatic callers** (Including GATK, SomaticSniper), the choice used in most cancer studies.
- **Copy number variation (CNV) finders** (including CONTRA, CNVnator, RDXplorer). CNVs can be detected in both whole-genome and whole-exome assemblies, which is not true of other structural variations.
- **Non-CNV structural variation (SV) finders** (including ExomeCNV, CONTRA) are used to find inversions, translocations or large INDELs.

It is critically important to choose the right variant caller for your data. Using the wrong variant calling pipeline can lead to missed variant calls. In addition, different variant callers have been shown to work better with different types of sequencing technology.

After alignment and variant calling, the limited number of applications that can perform sample cross-normalization or that can utilize a BED file to filter the variant list will perform those steps now.

The output from GATK and other variant calling pipelines is a Variant Calling Format (VCF) file. VCF is a scalable, tab-delimited text file with specific rules pertaining to the order and contents of columns. Each row represents a single nucleotide variant (SNV), insertion, deletion, or other sequence variation. Each variant is uniquely identified with a unique combination of letters and numbers, and the file may contain identifiers from multiple databases.

Step 5: Filter data to discover important variants

If Steps 1-4 didn't sound familiar or relevant to your experience, that's because they are frequently performed by a core facility or bioinformatics group. This step—which may start with the import of one or more VCF files—is the first one in which most researchers become directly involved.

VCF files? Jump in here!

All sequencing platforms produce sequence reads that contain many base-level errors, and it is crucial to apply filtering that can separate "sequencing noise" from the variant "signal." Much of this baseline filtering can be done in an automated fashion, preferably using statistical confidence measures, but some may require you to run some initial "noise filtering" tools.

Even after filtering out sequencing errors, a typical human exome set will still contain thousands of variants. You must now rely on additional criteria, typically imported into the analysis pipeline, to differentiate the uninteresting variants from those that have functional or clinical significance. Some open-source tools that support filtering are VCF Tools and SnpSift.

Many filtering strategies can be used, but here are four examples:

- Filter out **synonymous SNPs.** Since these SNPs encode for the normal amino acid, they will not result in a deleterious effect.

- In the case of a whole-genome study, filter out **variants located in non-coding regions**.

- Filter out **common alleles**. For example, we can reasonably assume that a mutation that causes a rare kidney disease will be rare in the human population. Thus, if a SNP is common (i.e., has a high allele frequency) across a population, it is unlikely to be the cause of the rare disease and can be removed from consideration.

- Some cloud-based tools (e.g., the Integrative Genomics Viewer) and standalone applications (CLC Bio, Lasergene) let you **compare variants from multiple VCF files**. For instance, to find variants related to a particular form of brain cancer, you could filter out variants found in both brain cancer patients and controls and focus on variants found only in the cancer patients.

Step 6: Determine how variants affect genes

The goal of this step is to determine the functional and clinical consequences of variants and how they impact genes. This is best accomplished with data tables that contain both variant and functional information for one or multiple samples and which have the ability to easily verify variant calls by visualizing the assembled sequence reads. This is most often done using genome browsers, which are available in standalone and web-based versions.

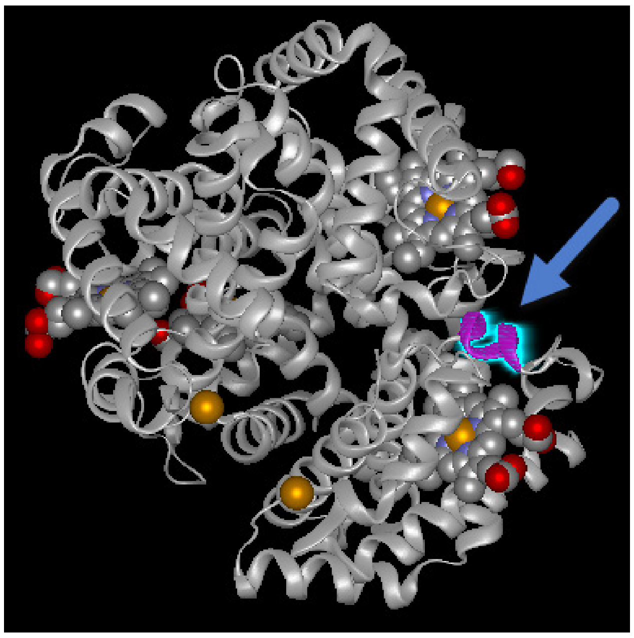
A benefit to **standalone browsers** is that most provide a graphical user interface (GUI) which can support zooming, easier visualization, and the ability to browse interactively. One drawback, however, is that you may be responsible for importing a large number of different data tracks for each of your samples. In addition, cross-comparison and analysis of multiple samples may be challenging, depending on the analysis tools that are available.

One advantage to **web-based genome browsers** (e.g., Ensembl, UCSC Genome Browser, ANNOVAR, AnnTools, VariantAnnotation, NGS-SNP, and snpEff) is that you do not usually need to download or update a human variant annotation database. On the other hand, you must upload your own proprietary data to a remote server, which is not a viable option for all organizations, owing to data privacy policies. Furthermore, some of these tools are command-line driven and can require typing multiple long strings of text into the command-line interface.

At a minimum, a genome browser should allow you to display the aligned reads, see variants clearly (e.g., by displaying them in a different color), view annotation information, and visit the corresponding public database (e.g., Ensembl, the GWAS Catalog, EVA, UniProt, dbSNP, ClinVar) entries online via hyperlinks. Most tools only allow visualization of SNPs, while a few also support viewing CNVs and SVs. Some genome visualization tools allow you to compare sequences from multiple individuals or even multiple organisms.

Once you have tabulated a list of variants, you may identify some for which there is limited data. Most search engines cannot excavate any further than paper titles and abstracts, where 85% of variants are not mentioned. This represents an enormous lost opportunity. While the crowd-sourced ClinVar platform is a useful repository of variant interpretations and related information, it is far from complete. In fact, an average of 31% of submitted references per variant are false-positives, with 30% of submissions having no references at all. Together, the downstream effects of these statistics can be serious, and may result in erroneous conclusions and/or a missed diagnosis. To mitigate these information quality issues, a deeper analysis can be performed using the Mastermind Genomic Search Engine, which is discussed in detail in Chapter 3.

Step 7 (Optional): Visualize the impact of the variant on the 3D protein structure



An optional step in the variant analysis pipeline is to attempt to determine the effect of a given variant on the 3D structure of the protein it codes for. Sometimes, the 3D structure may be minimally affected, if at all. In other cases, the variant can cause a dramatic change in protein structure, vitally impacting protein function.

Tools like I-Mutant and SDM can predict how the variant might affect protein stability, while structure prediction tools like I-TASSER and Phyre2 can predict the effect on protein structure. However, the latter tools will only work if a structural homolog is available. In some cases, you may find that the structure of the variant protein has already been determined through x-ray crystallography and is available for download through the Protein Data Bank (PDB). You can view these structures through various programs, including the LiteMol viewer.

An easier option is to use DNASTAR's Protean 3D application to accomplish all these tasks. Protean 3D's "Protein Design" workflow lets you use a wizard to easily "mutate" one or more residues and then calculates whether the changes are likely to be stabilizing or destabilizing compared to the original structure. NovaFold, a separately-licensed application that runs through the Protean 3D interface and is powered by the award-winning I-TASSER algorithm, can be used to predict the new protein structure caused by the mutated residue(s). Protean 3D lets you view the original and mutated proteins as fully customizable and rotatable 3D structures.

Chapter 2

Using Lasergene Genomics + Mastermind Genomic Search Engine to Streamline Variant Analysis

HOW IT WORKS

With **Lasergene Genomics** and **Mastermind**, just three applications are required to get meaningful variant results. Once you've identified variants of interest, you have the option to do further downstream analysis and visualization in Lasergene.

1 › 2 › 3

SeqMan NGen
Import and assemble reads or VCF files with auomatic variant calling.

ArrayStar
Filter data to discover important variants and determine how variants affect genes.

Mastermind
Determine clinical significance of variants.

> “This is the complete package, from assembly to analysis”
>
> Marjorie Beggs, Arkana Laboratories

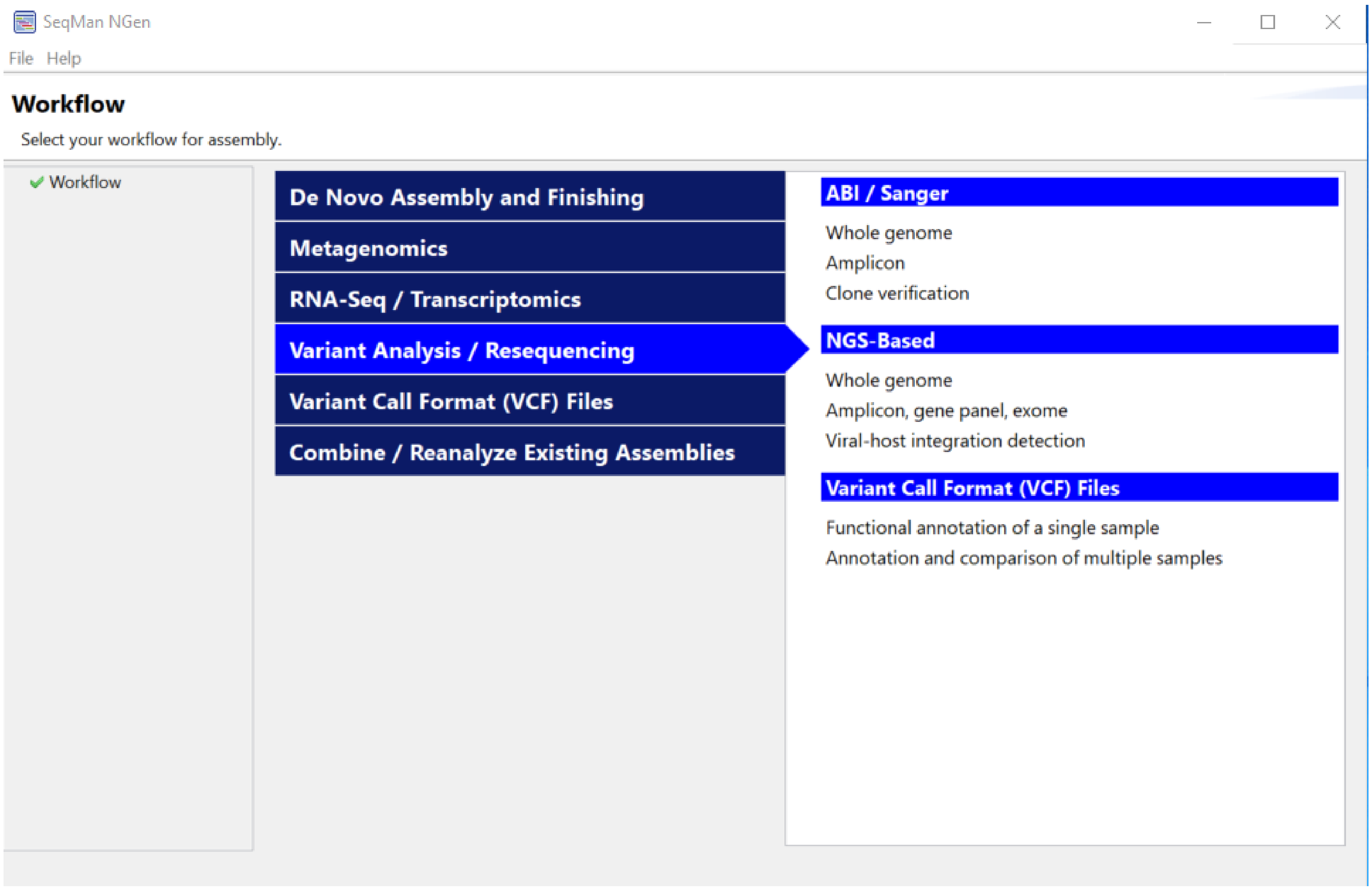
The **Lasergene Genomics** variant analysis pipeline is completely different from the open-source software or other commercial software described in Chapter 1. How?

Instead of using separate tools and requiring multiple manual interventions at each step, Lasergene fully automates most of these steps into one easy wizard-based application: **SeqMan NGen**. The steps below will take longer to read than to perform, as it typically only takes about **2 minutes** to make selections and begin the assembly and variant calling. After this, variant analysis is performed using **ArrayStar and Mastermind**, with optional analysis and visualizations in other fully-integrated Lasergene applications.

Getting Started: Assembly and Variant Calling in SeqMan NGen

After launching **SeqMan NGen**, the Workflow screen (**Figure 1**) prompts you to choose a workflow. Select either the Variant Analysis/Resequencing workflow or the RNA-Seq/Transcriptomics workflow if you have raw sequence read data; or the Variant Call Format (VCF) File workflow if your data has already been assembled and the variants detected and saved to VCF format.

Figure 1. SeqMan NGen's multi-tabbed Workflow screen is the starting point for the project.



Workflow

Select your workflow for assembly.

✓ Workflow
De Novo Assembly and Finishing
Metagenomics
RNA-Seq / Transcriptomics
Variant Analysis / Resequencing
Variant Call Format (VCF) Files
Combine / Reanalyze Existing Assemblies

ABI / Sanger

Whole genome
Amplicon
Clone verification

NGS-Based

Whole genome
Amplicon, gene panel, exome
Viral-host integration detection

Variant Call Format (VCF) Files

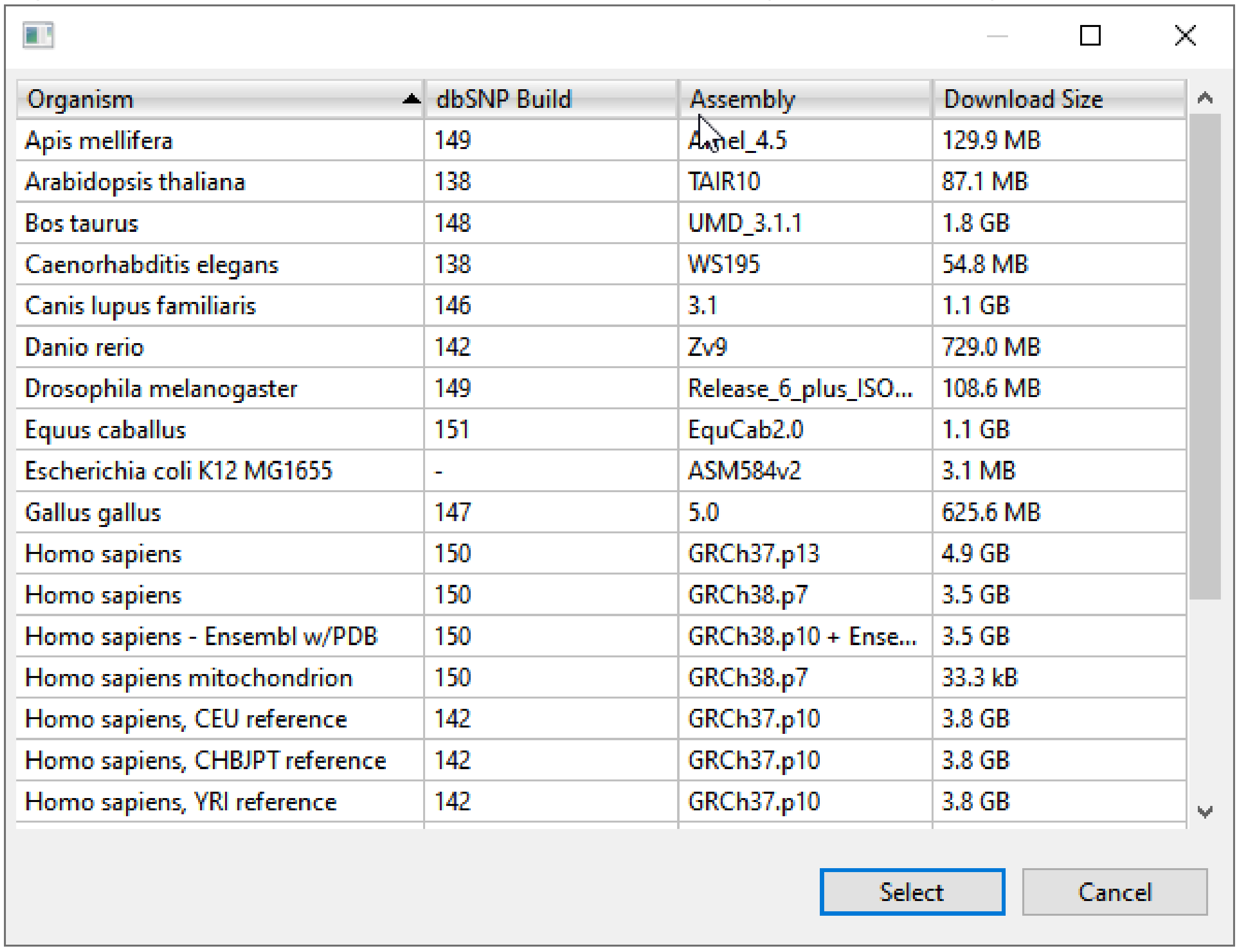
Functional annotation of a single sample
Annotation and comparison of multiple samples

The SeqMan NGen wizard will guide you through the setup for one or multiple samples, auto-grouping and auto-naming FASTQ input files so that assembly (or VCF import), variant detection, variant filtering, sample cross comparison, and variant annotation are all run autonomously. SeqMan NGen supports a wide variety of file types and read technologies.

You can choose to import a custom VCF file and/or BED file to incorporate into your assembly. Adding these files now means the assembled project will include known variants of interest and defined targeted regions, respectively. This information can be used as filtering criteria without requiring you to import additional data after assembly.

Choose to upload a reference sequence or use one of the curated Genome Template Packages provided for common model organisms, including many versions of the human genome (**Figure 2**). The Genome Template Packages include both annotated reference chromosomes and the dbSNP database.

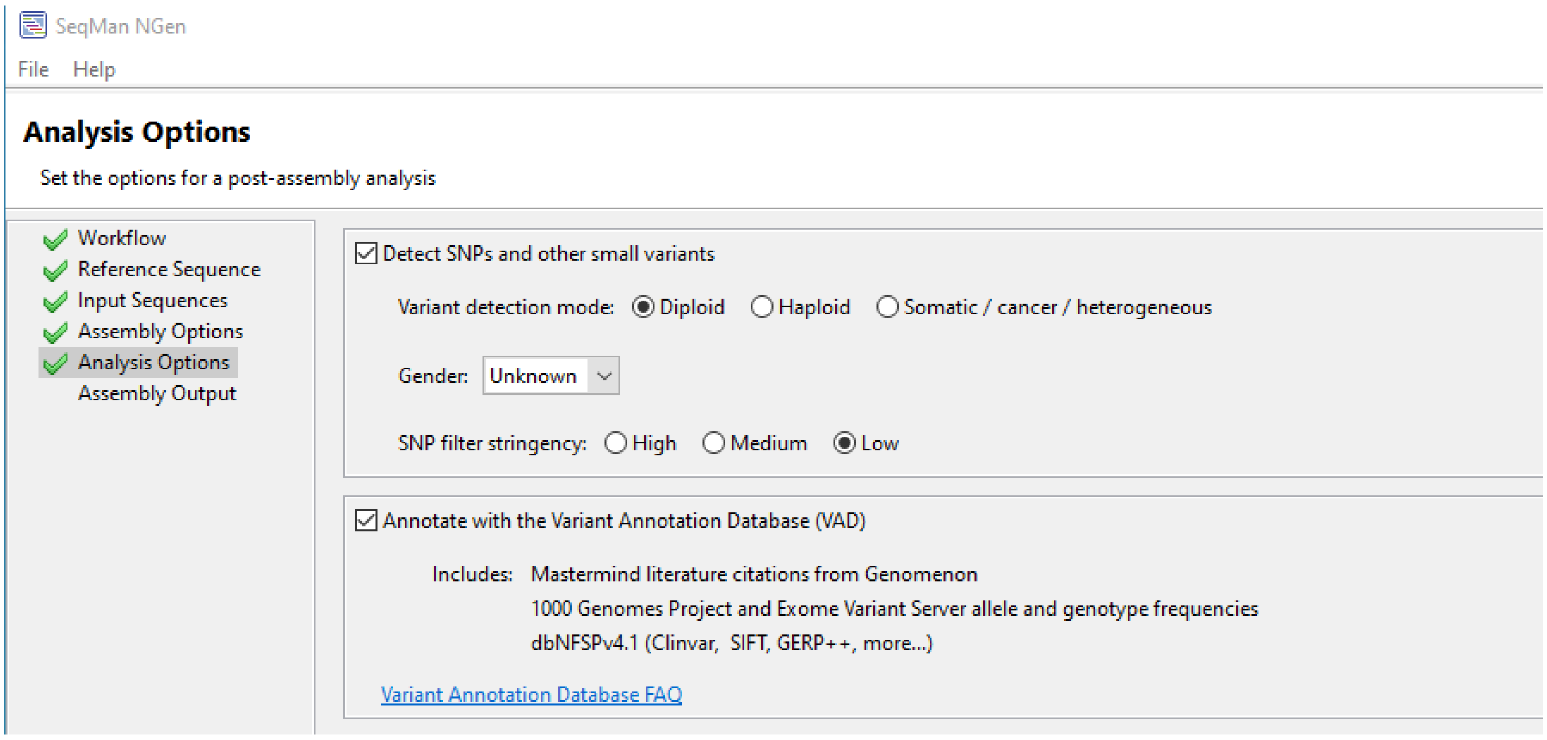
Figure 2. A partial list of SeqMan NGen's available genome packages.



Organism	dbSNP Build	Assembly	Download Size
Apis mellifera	149	Amel_4.5	129.9 MB
Arabidopsis thaliana	138	TAIR10	87.1 MB
Bos taurus	148	UMD_3.1.1	1.8 GB
Caenorhabditis elegans	138	WS195	54.8 MB
Canis lupus familiaris	146	3.1	1.1 GB
Danio rerio	142	Zv9	729.0 MB
Drosophila melanogaster	149	Release_6_plus_ISO...	108.6 MB
Equus caballus	151	EquCab2.0	1.1 GB
Escherichia coli K12 MG1655	-	ASM584v2	3.1 MB
Gallus gallus	147	5.0	625.6 MB
Homo sapiens	150	GRCh37.p13	4.9 GB
Homo sapiens	150	GRCh38.p7	3.5 GB
Homo sapiens - Ensembl w/PDB	150	GRCh38.p10 + Ense...	3.5 GB
Homo sapiens mitochondrion	150	GRCh38.p7	33.3 kB
Homo sapiens, CEU reference	142	GRCh37.p10	3.8 GB
Homo sapiens, CHBJPT reference	142	GRCh37.p10	3.8 GB
Homo sapiens, YRI reference	142	GRCh37.p10	3.8 GB

In the Analysis Options screen (**Figure 3**), choose variant calling options (diploid, haploid, somatic), gender, and initial variant filtering. Check the variant annotation database (VAD) box to automatically annotate any found variants with data from multiple databases, including Mastermind citation counts and hyperlinks to the Mastermind Genomic Search Engine.

Figure 3. SeqMan NGen "Analysis Options" wizard screen.



SeqMan NGen

File Help

Analysis Options

Set the options for a post-assembly analysis

- ✓ Workflow
- ✓ Reference Sequence
- ✓ Input Sequences
- ✓ Assembly Options
- ✓ Analysis Options
- Assembly Output

☑ Detect SNPs and other small variants

Variant detection mode: ● Diploid ○ Haploid ○ Somatic / cancer / heterogeneous

Gender: Unknown

SNP filter stringency: ○ High ○ Medium ● Low

☑ Annotate with the Variant Annotation Database (VAD)

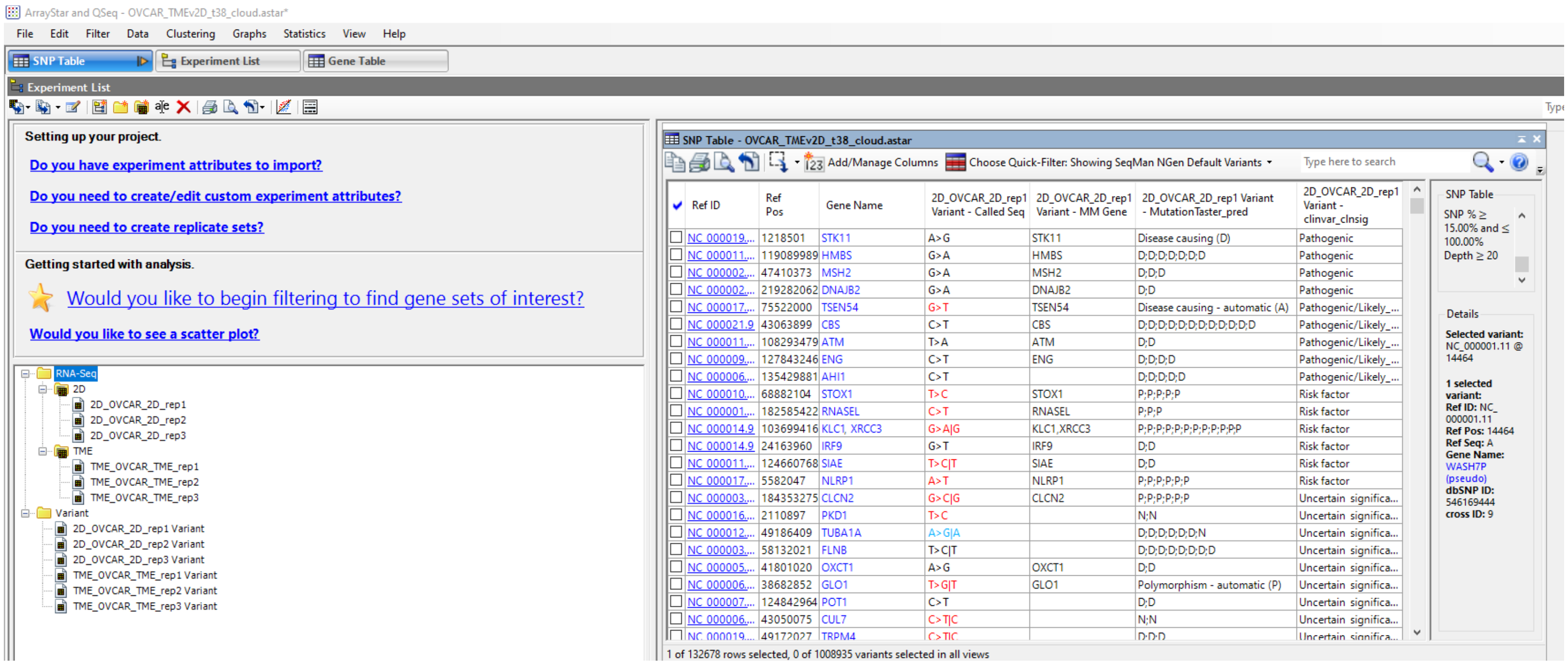
Includes: Mastermind literature citations from Genomenon
1000 Genomes Project and Exome Variant Server allele and genotype frequencies
dbNFSPv4.1 (Clinvar, SIFT, GERP++, more...)

Variant Annotation Database FAQ

You can also opt to detect CNVs and/or SVs. Choosing these options produces an assembly project that includes assembled data, filtered variant calls, CNVs, SVs, and extensive variant annotations.

Follow the guided steps to initiate assembly locally or on the cloud. Once assembly is complete, proceed to open the project in **ArrayStar** to analyze the results.

Figure 4. ArrayStar's "Experiment List" screen.



Setting up your project.

Do you have experiment attributes to import?

Do you need to create/edit custom experiment attributes?

Do you need to create replicate sets?

Getting started with analysis.

Would you like to begin filtering to find gene sets of interest?

Would you like to see a scatter plot?

- RNA-Seq
  - 2D
    - 2D_OVCAR_2D_rep1
    - 2D_OVCAR_2D_rep2
    - 2D_OVCAR_2D_rep3
  - TME
    - TME_OVCAR_TME_rep1
    - TME_OVCAR_TME_rep2
    - TME_OVCAR_TME_rep3
- Variant
  - 2D_OVCAR_2D_rep1 Variant
  - 2D_OVCAR_2D_rep2 Variant
  - 2D_OVCAR_2D_rep3 Variant
  - TME_OVCAR_TME_rep1 Variant
  - TME_OVCAR_TME_rep2 Variant
  - TME_OVCAR_TME_rep3 Variant

SNP Table - OVCAR_TMEv2D_t38_cloud.astar

✓	Ref ID	Ref Pos	Gene Name	2D_OVCAR_2D_rep1 Variant - Called Seq	2D_OVCAR_2D_rep1 Variant - MM Gene	2D_OVCAR_2D_rep1 Variant - MutationTaster_pred	2D_OVCAR_2D_rep1 Variant - clinvar_clnsig
□	NC 000019....	1218501	STK11	A>G	STK11	Disease causing (D)	Pathogenic
□	NC 000011....	119089989	HMBS	G>A	HMBS	D;D;D;D;D;D;D	Pathogenic
□	NC 000002....	47410373	MSH2	G>A	MSH2	D;D;D	Pathogenic
□	NC 000002....	219282062	DNAJB2	G>A	DNAJB2	D;D	Pathogenic
□	NC 000017....	75522000	TSEN54	G>T	TSEN54	Disease causing - automatic (A)	Pathogenic/Likely_...
□	NC 000021.9	43063899	CBS	C>T	CBS	D;D;D;D;D;D;D;D;D;D;D;D	Pathogenic/Likely_...
□	NC 000011....	108293479	ATM	T>A	ATM	D;D	Pathogenic/Likely_...
□	NC 000009....	127843246	ENG	C>T	ENG	D;D;D;D	Pathogenic/Likely_...
□	NC 000006....	135429881	AHI1	C>T		D;D;D;D;D	Pathogenic/Likely_...
□	NC 000010....	68882104	STOX1	T>C	STOX1	P;P;P;P;P	Risk factor
□	NC 000001....	182585422	RNASEL	C>T	RNASEL	P;P;P	Risk factor
□	NC 000014.9	103699416	KLC1, XRCC3	G>A|G	KLC1,XRCC3	P;P;P;P;P;P;P;P;P;P;P;P	Risk factor
□	NC 000014.9	24163960	IRF9	G>T	IRF9	D;D	Risk factor
□	NC 000011....	124660768	SIAE	T>C|T	SIAE	D;D	Risk factor
□	NC 000017....	5582047	NLRP1	A>T	NLRP1	P;P;P;P;P;P	Risk factor
□	NC 000003....	184353275	CLCN2	G>C|G	CLCN2	P;P;P;P;P;P	Uncertain significa...
□	NC 000016....	2110897	PKD1	T>C		N;N	Uncertain significa...
□	NC 000012....	49186409	TUBA1A	A>G|A		D;D;D;D;D;N	Uncertain significa...
□	NC 000003....	58132021	FLNB	T>C|T		D;D;D;D;D;D;D;D	Uncertain significa...
□	NC 000005....	41801020	OXCT1	A>G	OXCT1	D;D	Uncertain significa...
□	NC 000006....	38682852	GLO1	T>G|T	GLO1	Polymorphism - automatic (P)	Uncertain significa...
□	NC 000007....	124842964	POT1	C>T		D;D	Uncertain significa...
□	NC 000006....	43050075	CUL7	C>T|C		N;N	Uncertain significa...
□	NC 000019....	49172027	TRPM4	C>T|C		D;D;D	Uncertain significa...

1 of 132678 rows selected, 0 of 1008935 variants selected in all views

SNP Table

SNP % ≥ 15.00% and ≤ 100.00%
Depth ≥ 20

Details

Selected variant: NC_000001.11 @ 14464

1 selected variant:
Ref ID: NC_000001.11
Ref Pos: 14464
Ref Seq: A
Gene Name: WASH7P (pseudo)
dbSNP ID: 546169444
cross ID: 9

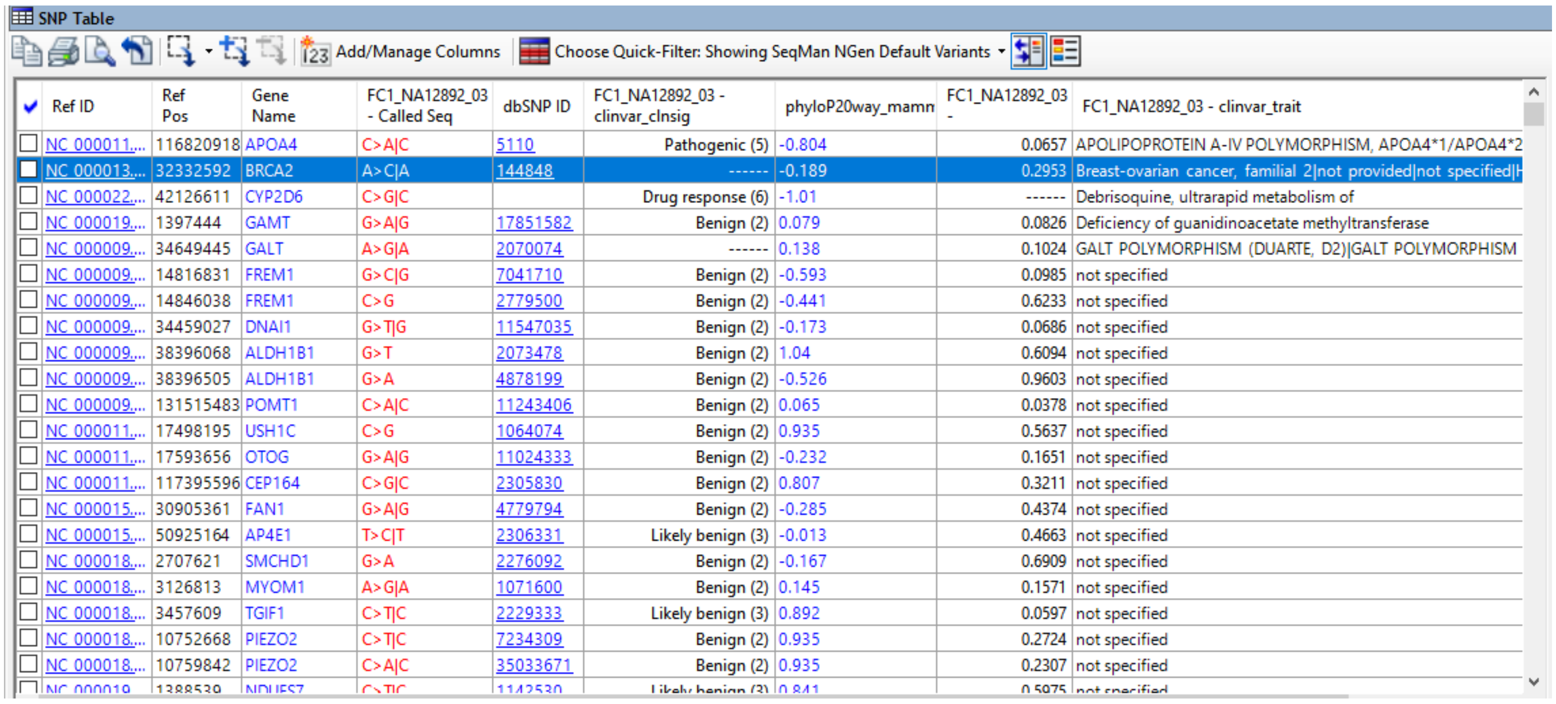
Find important variants with ArrayStar

ArrayStar automates Steps 5-6 with versatile filtering tools and rich graphical and tabular displays showing the variant results in context. Start by using the Advanced Filtering tool to filter out uninteresting variants, leaving only those worthy of further study. See the next chapter for an example of how to apply variant filtering.

Next, move to the Variant Table (**Figure 5**), which is ready for analysis and allows you to effortlessly add (or remove) data columns with information about each variant. If you checked the VAD box in SeqMan NGen, there are hundreds of columns to choose from. Many of these include links to the database entry for the variant in databases like the Mastermind Genomic Search Engine, ClinVar, dbSNP, 1000 Genomes Project, the Gene Ontology Consortium (GO), GERP, MutationTaster, PhastCons, dbNSFP, and many others.

ArrayStar is the only tool we are aware of that additionally performs an automatic cross sample analysis. If a variant is detected in one sample, the application cross checks all other samples at the same position and reports the status (variant, no variant call, no coverage) as a Variant Table column

Figure 5. ArrayStar's Variant Table displaying a selection of customizable data columns.



✓	Ref ID	Ref Pos	Gene Name	FC1_NA12892_03 - Called Seq	dbSNP ID	FC1_NA12892_03 - clinvar_clnsig	phyloP20way_mamn	FC1_NA12892_03 -	FC1_NA12892_03 - clinvar_trait
□	NC 000011....	116820918	APOA4	C>A|C	5110	Pathogenic (5)	-0.804	0.0657	APOLIPOPROTEIN A-IV POLYMORPHISM, APOA4*1/APOA4*2
□	NC 000013....	32332592	BRCA2	A>C|A	144848	------	-0.189	0.2953	Breast-ovarian cancer, familial 2|not provided|not specified|H
□	NC 000022....	42126611	CYP2D6	C>G|C		Drug response (6)	-1.01	------	Debrisoquine, ultrarapid metabolism of
□	NC 000019....	1397444	GAMT	G>A|G	17851582	Benign (2)	0.079	0.0826	Deficiency of guanidinoacetate methyltransferase
□	NC 000009....	34649445	GALT	A>G|A	2070074	------	0.138	0.1024	GALT POLYMORPHISM (DUARTE, D2)|GALT POLYMORPHISM
□	NC 000009....	14816831	FREM1	G>C|G	7041710	Benign (2)	-0.593	0.0985	not specified
□	NC 000009....	14846038	FREM1	C>G	2779500	Benign (2)	-0.441	0.6233	not specified
□	NC 000009....	34459027	DNAI1	G>T|G	11547035	Benign (2)	-0.173	0.0686	not specified
□	NC 000009....	38396068	ALDH1B1	G>T	2073478	Benign (2)	1.04	0.6094	not specified
□	NC 000009....	38396505	ALDH1B1	G>A	4878199	Benign (2)	-0.526	0.9603	not specified
□	NC 000009....	131515483	POMT1	C>A|C	11243406	Benign (2)	0.065	0.0378	not specified
□	NC 000011....	17498195	USH1C	C>G	1064074	Benign (2)	0.935	0.5637	not specified
□	NC 000011....	17593656	OTOG	G>A|G	11024333	Benign (2)	-0.232	0.1651	not specified
□	NC 000011....	117395596	CEP164	C>G|C	2305830	Benign (2)	0.807	0.3211	not specified
□	NC 000015....	30905361	FAN1	G>A|G	4779794	Benign (2)	-0.285	0.4374	not specified
□	NC 000015....	50925164	AP4E1	T>C|T	2306331	Likely benign (3)	-0.013	0.4663	not specified
□	NC 000018....	2707621	SMCHD1	G>A	2276092	Benign (2)	-0.167	0.6909	not specified
□	NC 000018....	3126813	MYOM1	A>G|A	1071600	Benign (2)	0.145	0.1571	not specified
□	NC 000018....	3457609	TGIF1	C>T|C	2229333	Likely benign (3)	0.892	0.0597	not specified
□	NC 000018....	10752668	PIEZO2	C>T|C	7234309	Benign (2)	0.935	0.2724	not specified
□	NC 000018....	10759842	PIEZO2	C>A|C	35033671	Benign (2)	0.935	0.2307	not specified
□	NC 000019	1388539	NDUFS7	C>T|C	1142530	Likely benign (3)	0.841	0.5975	not specified

Further exploration of variants with Mastermind

Also accessible from ArrayStar is **Genomenon's powerful Mastermind Genomic Search Engine,** a comprehensive genomics database designed to be an essential resource for analysts faced with data uncertainty. Mastermind allows you to quickly and easily search and cross-reference variant data from millions of PubMed articles. Additionally, Mastermind provides entries for millions of variants that are not included in databases such as HGMD and differentiates them with much higher precision.

Mastermind enables access to millions of full-text articles encompassing the entire breadth and depth of published genomic evidence. Using patented algorithms, Mastermind allows you to filter and prioritize relevant articles with high specificity, without compromising sensitivity. Additionally, Mastermind is able to draw genetic associations between keywords (e.g., genes, variants, diseases, CNVs, phenotypes, and therapies) and deliver a segmented literature output that has been customized for your particular case.

The successful use of this tool should result in a highly useful yield of evidence surrounding your variant of interest. On the other hand, if your results are limited in Mastermind, you can be assured that limited data exists.

Optional Analysis in Lasergene

After identifying important variants using ArrayStar and Mastermind, you can use other integrated Lasergene applications to visualize your assembly read alignment and coverage and/or model mutations on protein structure (see Step 7 in Chapter 2).

To conclude this section, we have provided a table contrasting the time and number of steps to perform variant analysis with DNASTAR software and with competing software. In the next chapter, we'll show you an example of how to combine the filtering in ArrayStar and Mastermind to home in on variants of interest in clinical sequencing data.

Table 1. Typical steps involved in the variant analysis workflow using DNASTAR versus other software tools, with average elapsed time to complete the workflow, including both hands-on and compute time.

Other Pipelines	Lasergene Genomics and Mastermind
1. Import & clean up sequencing data • Import FASTQ • Assess data & remove flawed portions	**Assembly and Variant Detection in SeqMan NGen**
2. Align reads to a reference genome • Align reads to a reference genome • Local realignment	
3. Remove PCR duplicates	
4. Call variants and INDELs • Call variants • Call INDELS • Remove marginal variants • Import VCF and identify known variants • Import allele and genotype frequencies • Predict amino acid changes • Import variant databases, including Mastermind Genomic Search Engine • Multiple functional predictions	
5. Filter data to discover important variants	**Variant Analysis in ArrayStar and Mastermind**
6. Determine how variants affect genes • Cross-sample analysis • Determine clinical significance using Mastermind and other databases	
7. Visualize the impact of the variant on the 3D protein structure (optional)	**Mutation Modeling in Protean 3D**
Total average time: >4 hours **10+ applications**	**Total average time: 2 hours** **3 integrated applications**

Chapter 3

Use Case with Lasergene Genomics & Mastermind Genomic Search Engine

A recent collaboration between DNASTAR and Resonant Therapeutics used RNA-Seq Illumina reads to study gene expression differences in ovarian carcinoma cells (OVCAR-2 cell line) grown using 2D and 3D cultures. In case these terms are unfamiliar, a 2D culture is where cells are grown in a petri dish, while a 3D culture is grown on a matrix that mimics the microenvironment of a tumor. In this study, we show how to quickly reduce a large data set down to three variants of interest, and then demonstrate the utility of Genomenon's Mastermind Genomic Search Engine for exploring deleterious variants.

Setting up and running the assembly in SeqMan NGen

SeqMan NGen was used to align Illumina RNA-Seq data from multiple replicates to the GRCh38.p2 human genome package. In addition to read alignment, the following steps were performed automatically during the assembly process:

1. SNPs and other small variants were called using the diploid model
2. Variants were annotated using the Variant Annotation Database (VAD)
3. Differential gene expression values were calculated using DESeq2

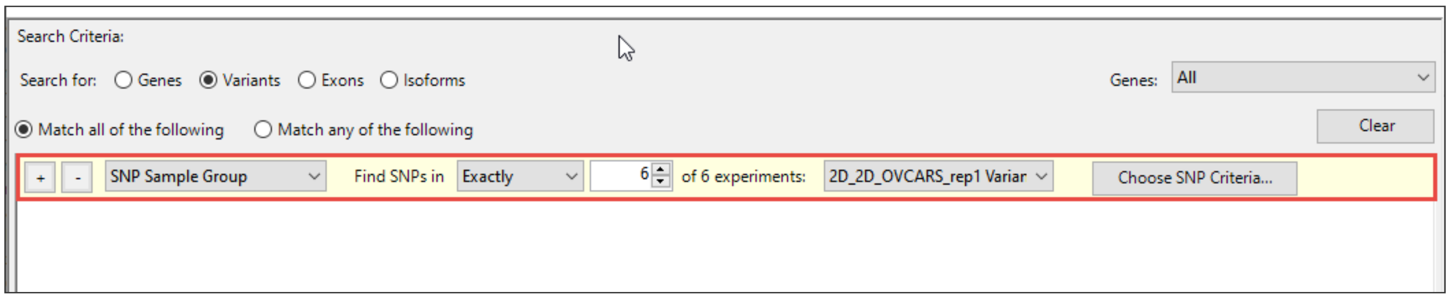
This study involved cancer cells, which understandably have a large number of mutations. During assembly, SeqMan NGen called 965,209 variants.

Filtering Variants in ArrayStar

After assembly in SeqMan NGen, we used the Advanced Filtering feature in ArrayStar to identify variants found in all six experiments that met the following criteria (**Figure 6, next page**):

- General tab - Non-synonymous SNPs only
- Statistics tab - Pnotref > 90%, SNP% > 15, Depth > 20
- Pathogenicity tab - Labeled in the ClinVar database as Pathogenic or Likely Pathogenic

Figure 6. Selections made in ArrayStar's "Advanced Filtering" dialog.



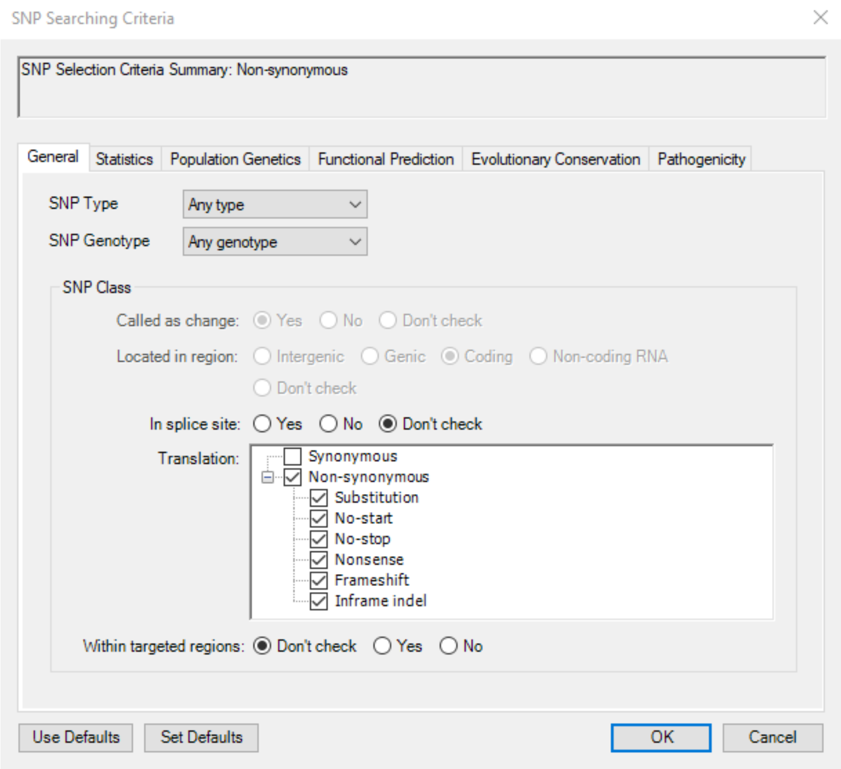
Search for: ○ Genes ● Variants ○ Exons ○ Isoforms

Genes: All

● Match all of the following ○ Match any of the following

SNP Sample Group Find SNPs in Exactly 6 of 6 experiments: 2D_2D_OVCARS_rep1 Variar

Figure 6a. Filtering dialog



SNP Selection Criteria Summary: Non-synonymous

SNP Type: Any type

SNP Genotype: Any genotype

Called as change: ● Yes ○ No ○ Don't check

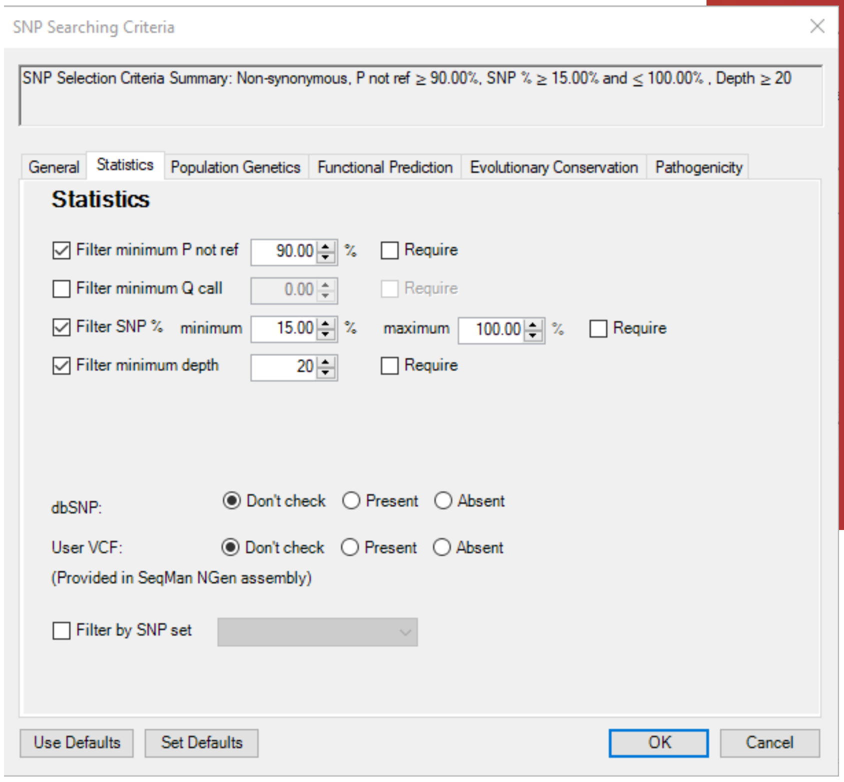
Located in region: ○ Intergenic ○ Genic ● Coding ○ Non-coding RNA ○ Don't check

In splice site: ○ Yes ○ No ● Don't check

Translation:
- □ Synonymous
- ☑ Non-synonymous
  - ☑ Substitution
  - ☑ No-start
  - ☑ No-stop
  - ☑ Nonsense
  - ☑ Frameshift
  - ☑ Inframe indel

Within targeted regions: ● Don't check ○ Yes ○ No

Figure 6b. General filtering options with non-synonymous filter applied.



SNP Selection Criteria Summary: Non-synonymous, P not ref ≥ 90.00%, SNP % ≥ 15.00% and ≤ 100.00% , Depth ≥ 20

☑ Filter minimum P not ref 90.00 % □ Require

□ Filter minimum Q call 0.00 □ Require

☑ Filter SNP % minimum 15.00 % maximum 100.00 % □ Require

☑ Filter minimum depth 20 □ Require

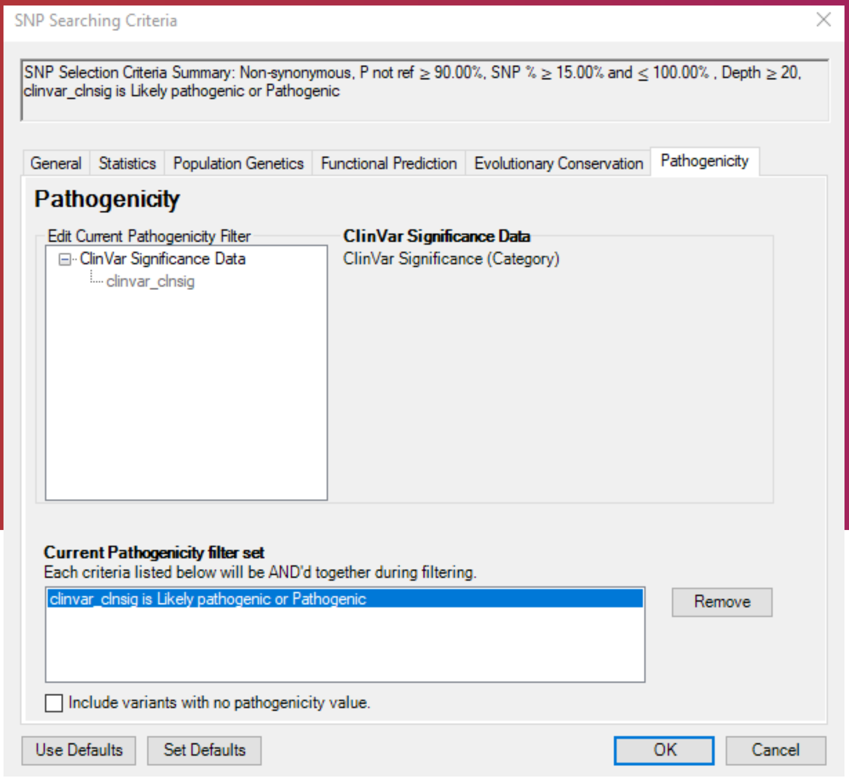
dbSNP: ● Don't check ○ Present ○ Absent

User VCF: ● Don't check ○ Present ○ Absent

(Provided in SeqMan NGen assembly)

□ Filter by SNP set

Figure 6c. Statistical SNP filtering options.



SNP Selection Criteria Summary: Non-synonymous, P not ref ≥ 90.00%, SNP % ≥ 15.00% and ≤ 100.00% , Depth ≥ 20, clinvar_clnsig is Likely pathogenic or Pathogenic

Current Pathogenicity filter set

Each criteria listed below will be AND'd together during filtering.

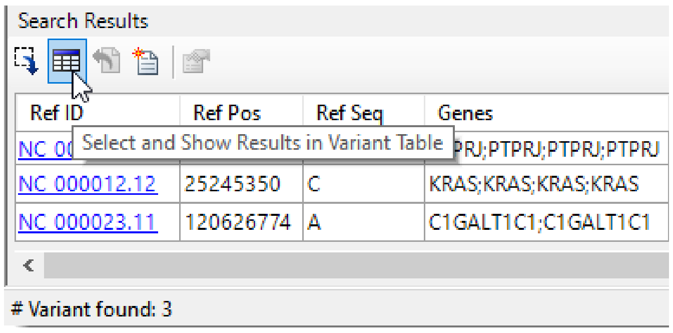
clinvar_clnsig is Likely pathogenic or Pathogenic

□ Include variants with no pathogenicity value.

Figure 6d. Pathogenecity filtering options with ClinVar filter applied.

ArrayStar found **three variants** that fit all criteria **(Figure 7).**

Figure 7. The three variants of interest found after filtering in ArrayStar.



Search Results

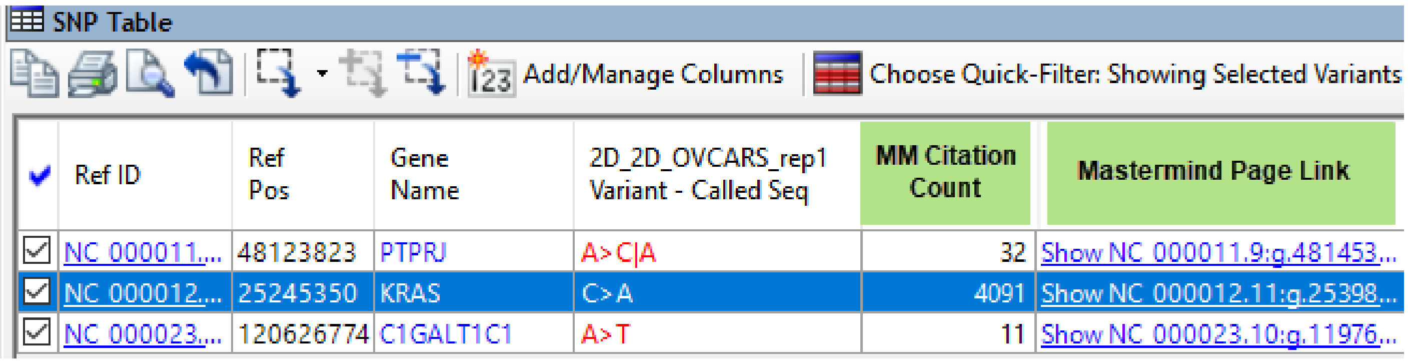
Select and Show Results in Variant Table

Ref ID	Ref Pos	Ref Seq	Genes
NC_00			PRJ;PTPRJ;PTPRJ;PTPRJ
NC_000012.12	25245350	C	KRAS;KRAS;KRAS;KRAS
NC_000023.11	120626774	A	C1GALT1C1;C1GALT1C1

Variant found: 3

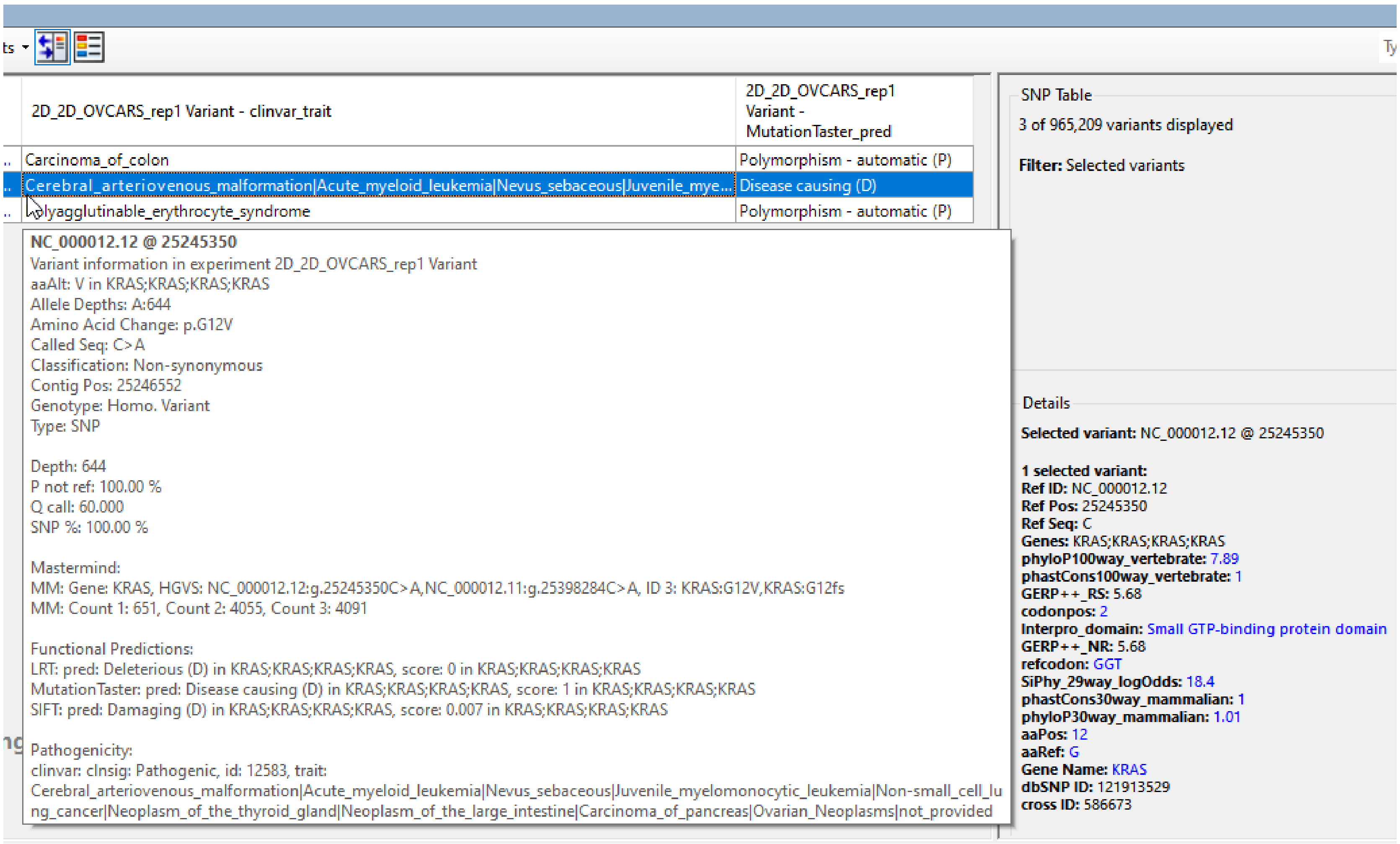
Looking at these results in the Variant Table, we were able to see the SNP base and reference bases, the number of articles found in Mastermind, links to that evidence, the clinvar_trait, and the MutationTaster_pred. For legibility in this ebook, the Variant Table has been split in half to create two figures (**Figures 8 and 9**).

Figure 8. The leftmost columns in ArrayStar's Variant Table.



✓	Ref ID	Ref Pos	Gene Name	2D_2D_OVCARS_rep1 Variant - Called Seq	MM Citation Count	Mastermind Page Link
☑	NC_000011....	48123823	PTPRJ	A>C|A	32	Show NC_000011.9:g.481453...
☑	NC_000012....	25245350	KRAS	C>A	4091	Show NC_000012.11:g.25398...
☑	NC_000023....	120626774	C1GALT1C1	A>T	11	Show NC_000023.10:g.11976...

Figure 9. The rightmost columns in ArrayStar's Variant Table, along with a tooltip for the selected variant and a view of the Details panel.



2D_2D_OVCARS_rep1 Variant - clinvar_trait	2D_2D_OVCARS_rep1 Variant - MutationTaster_pred
Carcinoma_of_colon	Polymorphism - automatic (P)
Cerebral_arteriovenous_malformation|Acute_myeloid_leukemia|Nevus_sebaceous|Juvenile_mye...	Disease causing (D)
olyagglutinable_erythrocyte_syndrome	Polymorphism - automatic (P)

NC_000012.12 @ 25245350
Variant information in experiment 2D_2D_OVCARS_rep1 Variant
aaAlt: V in KRAS;KRAS;KRAS;KRAS
Allele Depths: A:644
Amino Acid Change: p.G12V
Called Seq: C>A
Classification: Non-synonymous
Contig Pos: 25246552
Genotype: Homo. Variant
Type: SNP

Depth: 644
P not ref: 100.00 %
Q call: 60.000
SNP %: 100.00 %

Mastermind:
MM: Gene: KRAS, HGVS: NC_000012.12:g.25245350C>A,NC_000012.11:g.25398284C>A, ID 3: KRAS:G12V,KRAS:G12fs
MM: Count 1: 651, Count 2: 4055, Count 3: 4091

Functional Predictions:
LRT: pred: Deleterious (D) in KRAS;KRAS;KRAS;KRAS, score: 0 in KRAS;KRAS;KRAS;KRAS
MutationTaster: pred: Disease causing (D) in KRAS;KRAS;KRAS;KRAS, score: 1 in KRAS;KRAS;KRAS;KRAS
SIFT: pred: Damaging (D) in KRAS;KRAS;KRAS;KRAS, score: 0.007 in KRAS;KRAS;KRAS;KRAS

Pathogenicity:
clinvar: clnsig: Pathogenic, id: 12583, trait:
Cerebral_arteriovenous_malformation|Acute_myeloid_leukemia|Nevus_sebaceous|Juvenile_myelomonocytic_leukemia|Non-small_cell_lung_cancer|Neoplasm_of_the_thyroid_gland|Neoplasm_of_the_large_intestine|Carcinoma_of_pancreas|Ovarian_Neoplasms|not_provided

SNP Table
3 of 965,209 variants displayed

Filter: Selected variants

Details

Selected variant: NC_000012.12 @ 25245350

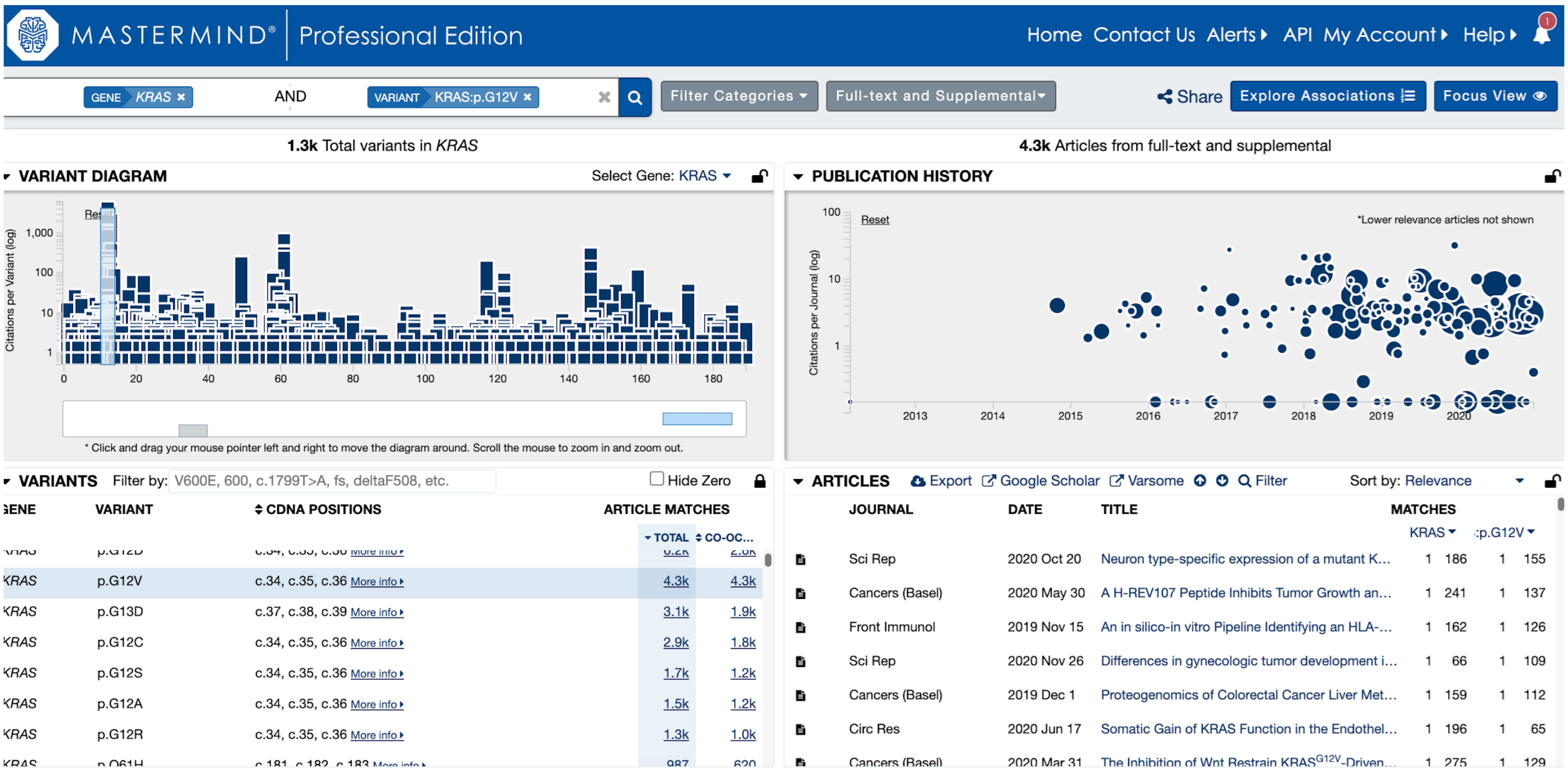
1 selected variant:
Ref ID: NC_000012.12
Ref Pos: 25245350
Ref Seq: C
Genes: KRAS;KRAS;KRAS;KRAS
phyloP100way_vertebrate: 7.89
phastCons100way_vertebrate: 1
GERP++_RS: 5.68
codonpos: 2
Interpro_domain: Small GTP-binding protein domain
GERP++_NR: 5.68
refcodon: GGT
SiPhy_29way_logOdds: 18.4
phastCons30way_mammalian: 1
phyloP30way_mammalian: 1.01
aaPos: 12
aaRef: G
Gene Name: KRAS
dbSNP ID: 121913529
cross ID: 586673

Exploring Genomic Associations with Mastermind

Note that the variant in the KRAS gene (middle of the three variants) has over 4,300 Mastermind citations. This well-known variant, located on chromosome 12 at position 25245350, causes the G12V change in the KRAS protein and is the key driver mutation behind the tumors being studied. In addition to the two-star status of this SNP in ClinVar, notice that all three functional prediction methods present in the VAD also flag it as deleterious. As shown near the bottom of the tooltip, this KRAS gene mutation has been associated with a wide variety of cancers.

ArrayStar includes a direct link to Mastermind, where we can view the entry for known KRAS mutations (**Figure 10**).

Figure 10. The Mastermind Professional Edition landing page for the KRAS G12V variant.


MASTERMIND®
Professional Edition
Home
Contact Us
Alerts ▸
API
My Account ▸
Help ▸
GENE KRAS ×
AND
VARIANT KRAS:p.G12V ×
Filter Categories ▾
Full-text and Supplemental ▾
Share
Explore Associations
Focus View
1.3k Total variants in KRAS
4.3k Articles from full-text and supplemental
VARIANT DIAGRAM
Select Gene: KRAS ▾
Citations per Variant (log)
1,000
100
10
1
0
20
40
60
80
100
120
140
160
180
* Click and drag your mouse pointer left and right to move the diagram around. Scroll the mouse to zoom in and zoom out.
PUBLICATION HISTORY
Reset
*Lower relevance articles not shown
Citations per Journal (log)
100
10
1
2013
2014
2015
2016
2017
2018
2019
2020

VARIANTS Filter by: V600E, 600, c.1799T>A, fs, deltaF508, etc. □ Hide Zero

GENE	VARIANT	CDNA POSITIONS	ARTICLE MATCHES TOTAL	ARTICLE MATCHES CO-OC...
KRAS	p.G12D	c.34, c.35, c.36 More info ▸	6.2k	2.8k
KRAS	p.G12V	c.34, c.35, c.36 More info ▸	4.3k	4.3k
KRAS	p.G13D	c.37, c.38, c.39 More info ▸	3.1k	1.9k
KRAS	p.G12C	c.34, c.35, c.36 More info ▸	2.9k	1.8k
KRAS	p.G12S	c.34, c.35, c.36 More info ▸	1.7k	1.2k
KRAS	p.G12A	c.34, c.35, c.36 More info ▸	1.5k	1.2k
KRAS	p.G12R	c.34, c.35, c.36 More info ▸	1.3k	1.0k
KRAS	p.Q61H	c.181, c.182, c.183 More info ▸	987	620

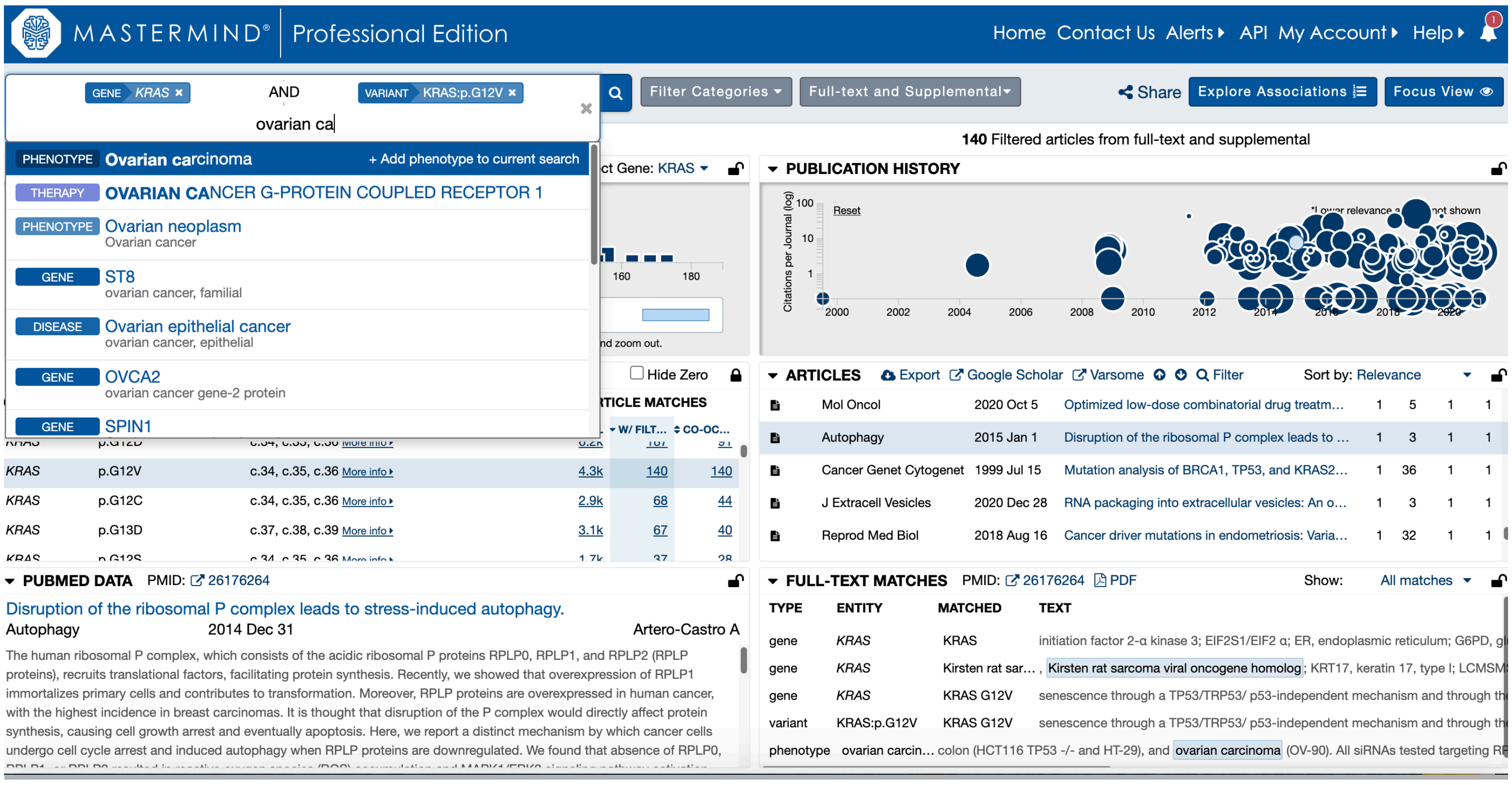
ARTICLES Export Google Scholar Varsome Filter Sort by: Relevance

JOURNAL	DATE	TITLE	MATCHES KRAS		MATCHES :p.G12V	
Sci Rep	2020 Oct 20	Neuron type-specific expression of a mutant K...	1	186	1	155
Cancers (Basel)	2020 May 30	A H-REV107 Peptide Inhibits Tumor Growth an...	1	241	1	137
Front Immunol	2019 Nov 15	An in silico-in vitro Pipeline Identifying an HLA-...	1	162	1	126
Sci Rep	2020 Nov 26	Differences in gynecologic tumor development i...	1	66	1	109
Cancers (Basel)	2019 Dec 1	Proteogenomics of Colorectal Cancer Liver Met...	1	159	1	112
Circ Res	2020 Jun 17	Somatic Gain of KRAS Function in the Endothel...	1	196	1	65
Cancers (Basel)	2020 Mar 31	The Inhibition of Wnt Restrain KRAS^{G12V}-Driven...	1	275	1	129

Optimize output with prioritized sorting and filtering

The image on the previous page shows the full list of over 4,300 publications citing this variant. The literature base can then be focused by adding phenotype keywords. In this case, when “ovarian carcinoma” is applied, the reference number is reduced to 140 articles (**Figure 11**).

Figure 11. Mastermind KRAS G12V with ovarian carcinoma phenotype applied.



MASTERMIND® Professional Edition
Home Contact Us Alerts API My Account Help
GENE KRAS AND VARIANT KRAS:p.G12V
Filter Categories
Full-text and Supplemental
Share
Explore Associations
Focus View

ovarian ca

PHENOTYPE **Ovarian carcinoma** + Add phenotype to current search
THERAPY **OVARIAN CA**NCER G-PROTEIN COUPLED RECEPTOR 1
PHENOTYPE Ovarian neoplasm
Ovarian cancer
GENE ST8
ovarian cancer, familial
DISEASE Ovarian epithelial cancer
ovarian cancer, epithelial
GENE OVCA2
ovarian cancer gene-2 protein
GENE SPIN1

ct Gene: KRAS
160
180
nd zoom out.

☐ Hide Zero

				TICLE MATCHES W/ FILT…	CO-OC…
KRAS	p.G12V	c.34, c.35, c.36 More info	4.3k	140	140
KRAS	p.G12C	c.34, c.35, c.36 More info	2.9k	68	44
KRAS	p.G13D	c.37, c.38, c.39 More info	3.1k	67	40

140 Filtered articles from full-text and supplemental

PUBLICATION HISTORY

ARTICLES Export Google Scholar Varsome Filter Sort by: Relevance

Mol Oncol	2020 Oct 5	Optimized low-dose combinatorial drug treatm…	1	5	1	1
Autophagy	2015 Jan 1	Disruption of the ribosomal P complex leads to …	1	3	1	1
Cancer Genet Cytogenet	1999 Jul 15	Mutation analysis of BRCA1, TP53, and KRAS2…	1	36	1	1
J Extracell Vesicles	2020 Dec 28	RNA packaging into extracellular vesicles: An o…	1	3	1	1
Reprod Med Biol	2018 Aug 16	Cancer driver mutations in endometriosis: Varia…	1	32	1	1

PUBMED DATA PMID: 26176264

Disruption of the ribosomal P complex leads to stress-induced autophagy.
Autophagy 2014 Dec 31 Artero-Castro A

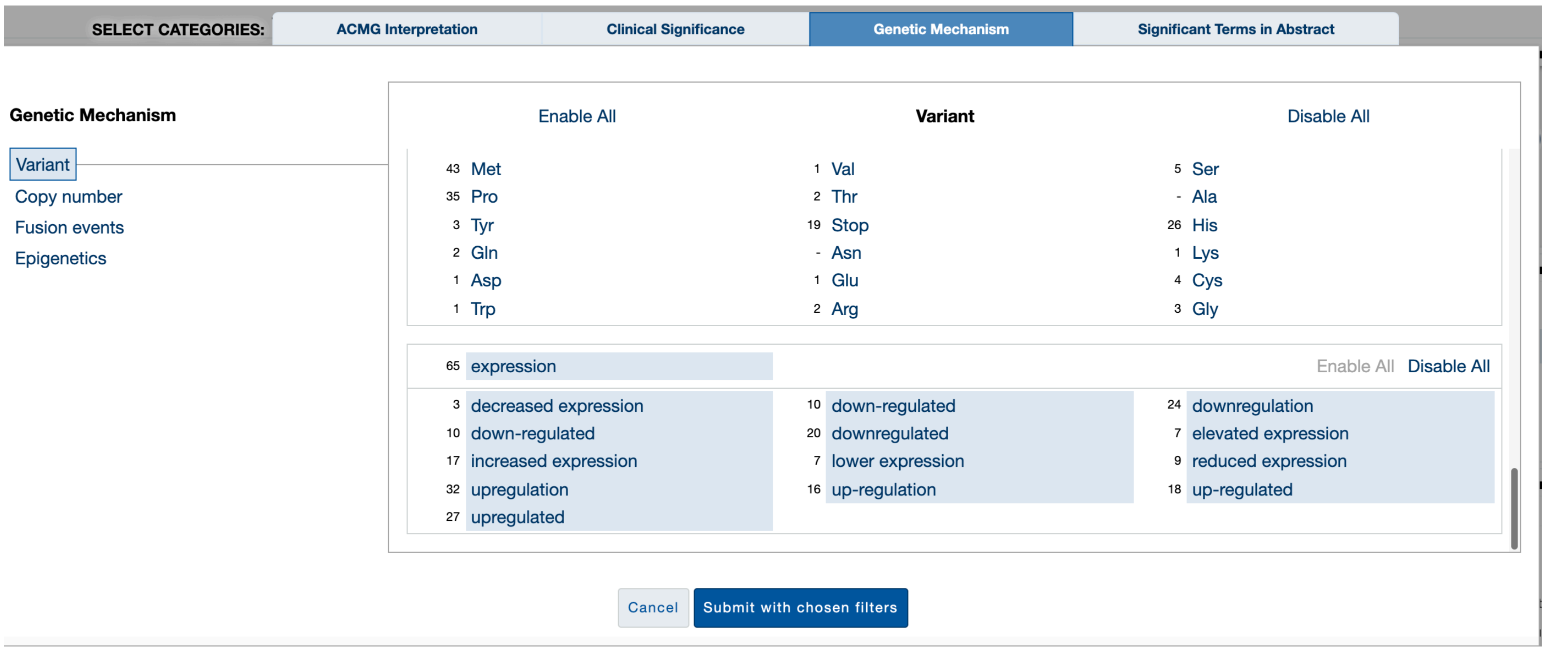
The human ribosomal P complex, which consists of the acidic ribosomal P proteins RPLP0, RPLP1, and RPLP2 (RPLP proteins), recruits translational factors, facilitating protein synthesis. Recently, we showed that overexpression of RPLP1 immortalizes primary cells and contributes to transformation. Moreover, RPLP proteins are overexpressed in human cancer, with the highest incidence in breast carcinomas. It is thought that disruption of the P complex would directly affect protein synthesis, causing cell growth arrest and eventually apoptosis. Here, we report a distinct mechanism by which cancer cells undergo cell cycle arrest and induced autophagy when RPLP proteins are downregulated. We found that absence of RPLP0,

FULL-TEXT MATCHES PMID: 26176264 PDF Show: All matches

TYPE	ENTITY	MATCHED	TEXT
gene	KRAS	KRAS	initiation factor 2-α kinase 3; EIF2S1/EIF2 α; ER, endoplasmic reticulum; G6PD, gl
gene	KRAS	Kirsten rat sar…	, Kirsten rat sarcoma viral oncogene homolog; KRT17, keratin 17, type I; LCMSM
gene	KRAS	KRAS G12V	senescence through a TP53/TRP53/ p53-independent mechanism and through th
variant	KRAS:p.G12V	KRAS G12V	senescence through a TP53/TRP53/ p53-independent mechanism and through th
phenotype	ovarian carcin…		colon (HCT116 TP53 -/- and HT-29), and ovarian carcinoma (OV-90). All siRNAs tested targeting RP

Other filters may be used to further refine this search, including expression filters in the “Genetic Mechanism” tab (**Figure 12**).

Figure 12. Mastermind Filters including expression



SELECT CATEGORIES: ACMG Interpretation | Clinical Significance | Genetic Mechanism | Significant Terms in Abstract

Genetic Mechanism

Variant
Copy number
Fusion events
Epigenetics

Enable All **Variant** Disable All

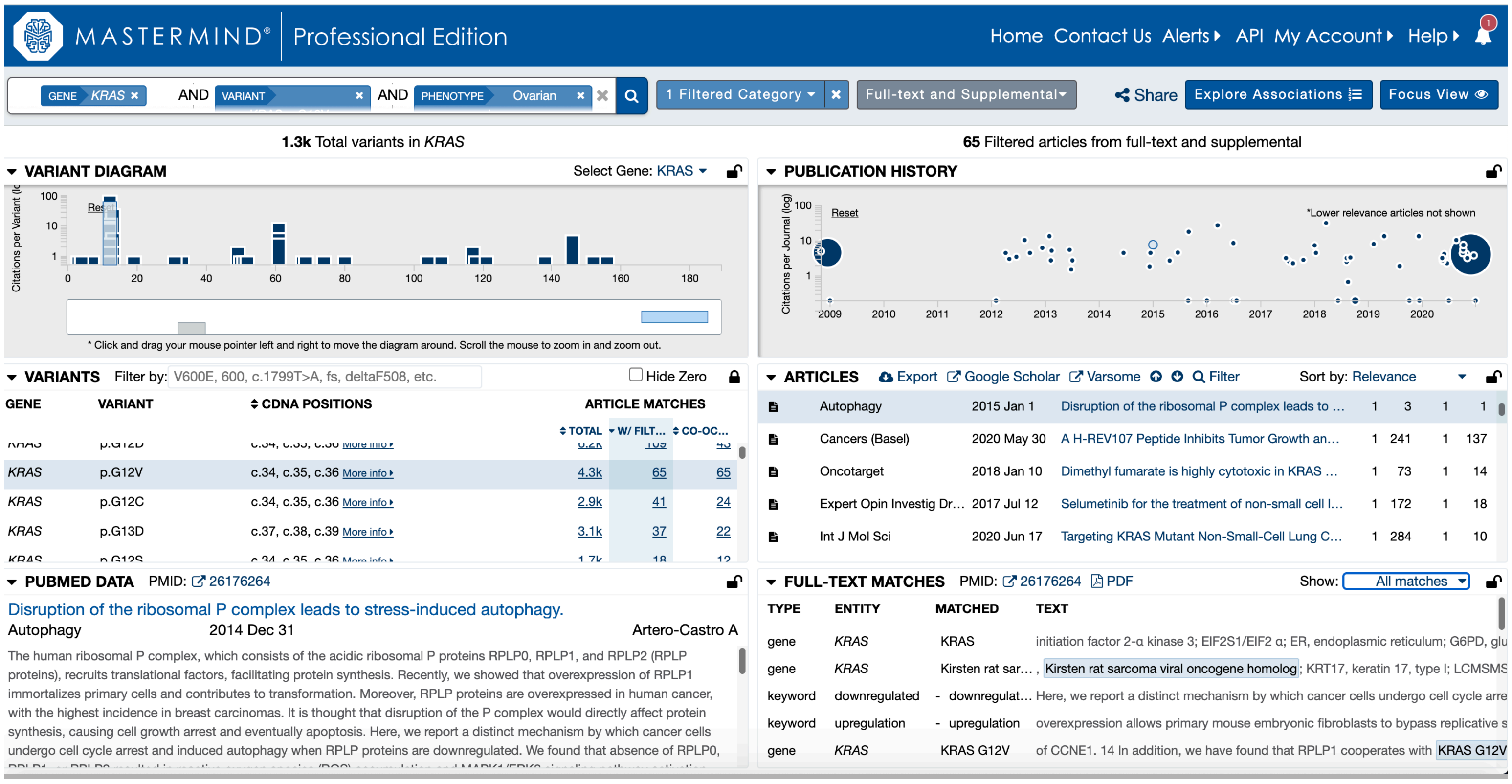
43	Met	1	Val	5	Ser
35	Pro	2	Thr	-	Ala
3	Tyr	19	Stop	26	His
2	Gln	-	Asn	1	Lys
1	Asp	1	Glu	4	Cys
1	Trp	2	Arg	3	Gly

65 expression Enable All Disable All

3	decreased expression	10	down-regulated	24	downregulation
10	down-regulated	20	downregulated	7	elevated expression
17	increased expression	7	lower expression	9	reduced expression
32	upregulation	16	up-regulation	18	up-regulated
27	upregulated				

This focuses the results on **65 highly relevant articles** (**Figure 13**).

Figure 13: Mastermind KRAS G12V Filtered by phenotype, expression and relevance



1.3k Total variants in *KRAS*

65 Filtered articles from full-text and supplemental

* Click and drag your mouse pointer left and right to move the diagram around. Scroll the mouse to zoom in and zoom out.

GENE	VARIANT	CDNA POSITIONS	ARTICLE MATCHES TOTAL	W/ FILT…	CO-OC…
KRAS	p.G12D	c.34, c.35, c.36 More info ▸	6.2k	109	43
KRAS	p.G12V	c.34, c.35, c.36 More info ▸	4.3k	65	65
KRAS	p.G12C	c.34, c.35, c.36 More info ▸	2.9k	41	24
KRAS	p.G13D	c.37, c.38, c.39 More info ▸	3.1k	37	22
KRAS	p.G12S	c.34, c.35, c.36 More info ▸	1.7k	18	12

Autophagy	2015 Jan 1	Disruption of the ribosomal P complex leads to …	1	3	1	1
Cancers (Basel)	2020 May 30	A H-REV107 Peptide Inhibits Tumor Growth an…	1	241	1	137
Oncotarget	2018 Jan 10	Dimethyl fumarate is highly cytotoxic in KRAS …	1	73	1	14
Expert Opin Investig Dr…	2017 Jul 12	Selumetinib for the treatment of non-small cell l…	1	172	1	18
Int J Mol Sci	2020 Jun 17	Targeting KRAS Mutant Non-Small-Cell Lung C…	1	284	1	10

PUBMED DATA PMID: 26176264

Disruption of the ribosomal P complex leads to stress-induced autophagy.

Autophagy 2014 Dec 31 Artero-Castro A

The human ribosomal P complex, which consists of the acidic ribosomal P proteins RPLP0, RPLP1, and RPLP2 (RPLP proteins), recruits translational factors, facilitating protein synthesis. Recently, we showed that overexpression of RPLP1 immortalizes primary cells and contributes to transformation. Moreover, RPLP proteins are overexpressed in human cancer, with the highest incidence in breast carcinomas. It is thought that disruption of the P complex would directly affect protein synthesis, causing cell growth arrest and eventually apoptosis. Here, we report a distinct mechanism by which cancer cells undergo cell cycle arrest and induced autophagy when RPLP proteins are downregulated. We found that absence of RPLP0,

FULL-TEXT MATCHES PMID: 26176264

TYPE	ENTITY	MATCHED	TEXT
gene	*KRAS*	KRAS	initiation factor 2-α kinase 3; EIF2S1/EIF2 α; ER, endoplasmic reticulum; G6PD, glu
gene	*KRAS*	Kirsten rat sar…	, Kirsten rat sarcoma viral oncogene homolog; KRT17, keratin 17, type I; LCMSMS
keyword	downregulated	- downregulat…	Here, we report a distinct mechanism by which cancer cells undergo cell cycle arre
keyword	upregulation	- upregulation	overexpression allows primary mouse embryonic fibroblasts to bypass replicative s
gene	*KRAS*	KRAS G12V	of CCNE1. 14 In addition, we have found that RPLP1 cooperates with KRAS G12V

The articles displayed are sorted by relevance, with the larger circles in the publication history pane denoting the most relevant articles. Relevance is determined by several factors, including, but not limited to:

- The number of times in which the gene (“KRAS”) and variant (“G12V”) names are mentioned in each article,
- Whether these terms are in the title, abstract and/or full text or supplemental information, and
- Proximity to phenotype and other keywords applied through the filtering process.

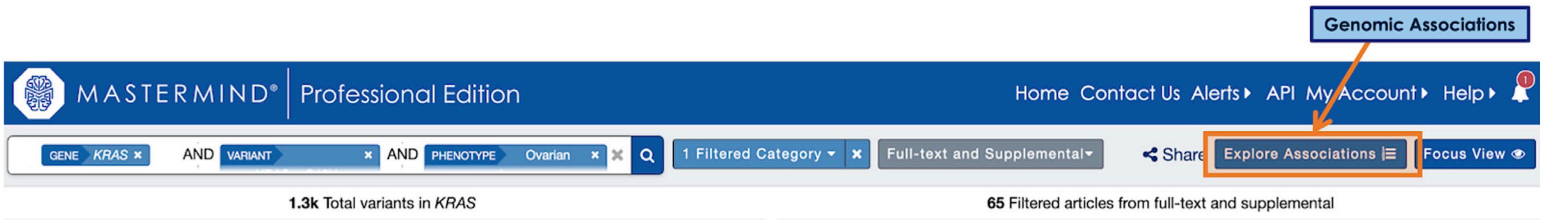
All of these elements can be viewed in the “Full Text Matches” section by viewing all matches being highlighted in the sentence fragments.

Uncover data connections with genomic associations

Due to the highly complex nature of genetic disorders and the variety of factors that contribute to their pathogenicity, the opportunity to create clarity around how your search criteria of interest relate with each other is exceedingly beneficial. The Genomic Associations functionality within Mastermind allows you to explore significant genetic associations that may otherwise be missed or unrecognized, giving you the ability to uncover key data that may lead to a new hypothesis, a research discovery, or even solving a clinical case.

To expand on the example, suppose that you are interested in exploring treatment options for ovarian carcinoma related to KRAS p.G12V mutations. Clicking on the blue “Explore Associations” box (**Figure 14**) directs you to a flexible interface that connects your search keywords with all associated genetic factors in the medical evidence.

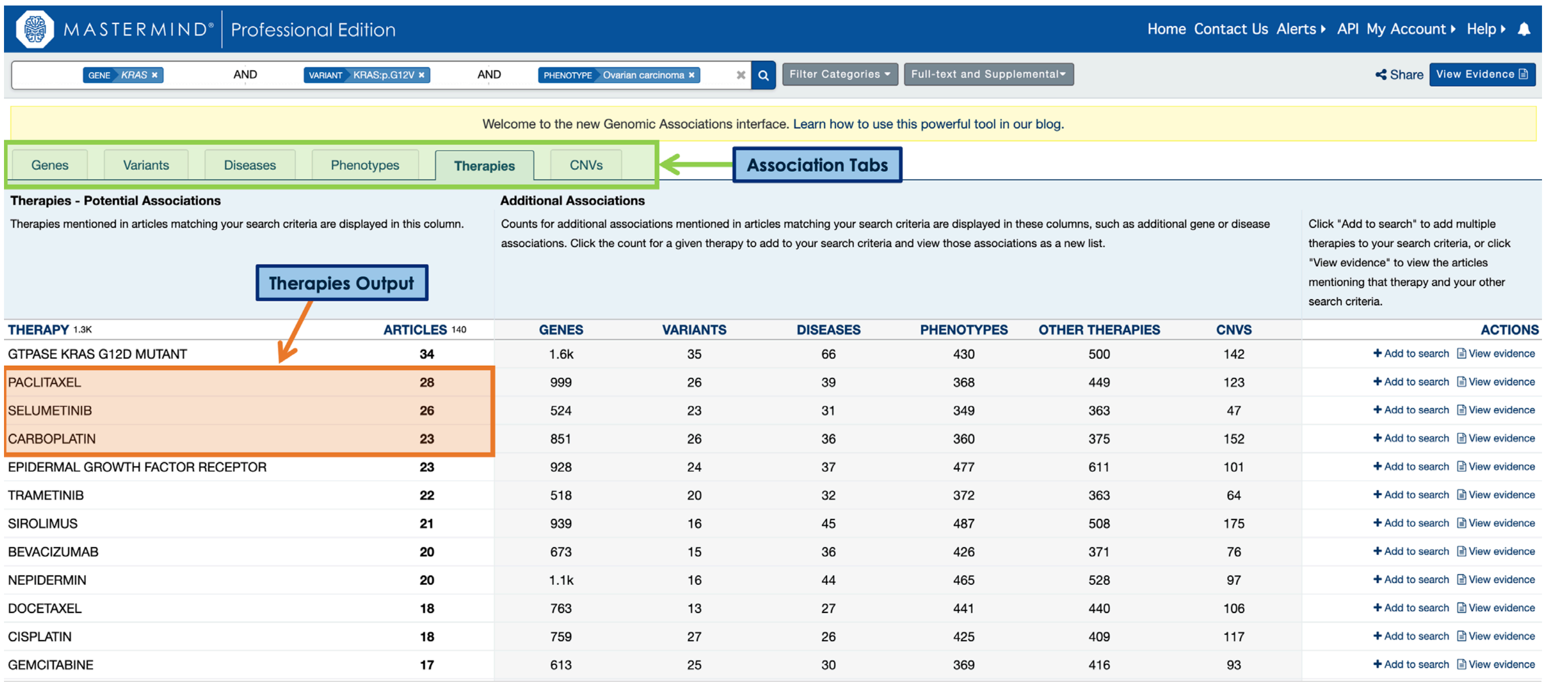
Figure 14. The Explore Associations box on the Mastermind page.


Genomic Associations
MASTERMIND® Professional Edition
Home Contact Us Alerts ▸ API My Account ▸ Help ▸
GENE KRAS ×
AND
VARIANT ×
AND
PHENOTYPE Ovarian ×
1 Filtered Category ▾
Full-text and Supplemental ▾
Share
Explore Associations
Focus View
1.3k Total variants in KRAS
65 Filtered articles from full-text and supplemental

In this case, the “Therapies” tab reveals prioritized articles covering therapeutic intervention for KRAS p.G12V-related ovarian carcinoma. As seen in Figure 15, in terms of article quantity, the top three pharmacotherapies are:

1) Paclitaxel
2) Selumetinib
3) Carboplatin

Figure 15. The “Therapies” tab with the top three pharmacotherapies boxed in orange.



Welcome to the new Genomic Associations interface. Learn how to use this powerful tool in our blog.

Therapies - Potential Associations

Therapies mentioned in articles matching your search criteria are displayed in this column.

Additional Associations

Counts for additional associations mentioned in articles matching your search criteria are displayed in these columns, such as additional gene or disease associations. Click the count for a given therapy to add to your search criteria and view those associations as a new list.

Click "Add to search" to add multiple therapies to your search criteria, or click "View evidence" to view the articles mentioning that therapy and your other search criteria.

THERAPY 1.3K	ARTICLES 140	GENES	VARIANTS	DISEASES	PHENOTYPES	OTHER THERAPIES	CNVS	ACTIONS
GTPASE KRAS G12D MUTANT	**34**	1.6k	35	66	430	500	142	+ Add to search View evidence
PACLITAXEL	**28**	999	26	39	368	449	123	+ Add to search View evidence
SELUMETINIB	**26**	524	23	31	349	363	47	+ Add to search View evidence
CARBOPLATIN	**23**	851	26	36	360	375	152	+ Add to search View evidence
EPIDERMAL GROWTH FACTOR RECEPTOR	**23**	928	24	37	477	611	101	+ Add to search View evidence
TRAMETINIB	**22**	518	20	32	372	363	64	+ Add to search View evidence
SIROLIMUS	**21**	939	16	45	487	508	175	+ Add to search View evidence
BEVACIZUMAB	**20**	673	15	36	426	371	76	+ Add to search View evidence
NEPIDERMIN	**20**	1.1k	16	44	465	528	97	+ Add to search View evidence
DOCETAXEL	**18**	763	13	27	441	440	106	+ Add to search View evidence
CISPLATIN	**18**	759	27	26	425	409	117	+ Add to search View evidence
GEMCITABINE	**17**	613	25	30	369	416	93	+ Add to search View evidence

Awareness of therapies surrounding a particular genetic variant creates informed focus during hypothesis formation and study design. Most importantly, however, this invaluable insight significantly impacts clinical decision making by adding precision to patient care.

Finally, although these patient-facing applications are compelling, they represent a small fraction of this technology’s complete potential: as evidenced by the other five association tabs, there’s much more for Lasergene and Mastermind users to discover.

Conclusion

Next-generation sequencing (NGS) of whole genomes, exomes, and RNA-Seq data is a powerful tool in biomedical research and diagnostics. However, sifting through the vast amount of NGS data to find the most relevant variants can be like looking for a polar bear in a snowstorm. Incorrect or deficient annotations can cause you to overlook potentially disease-causing variants. Open source and commercial tools have been developed to address this issue, but often, a dozen or more unrelated tools must be mastered in order to perform the analysis.

In response, DNASTAR has developed Lasergene Genomics, a fully integrated variant analysis and annotation pipeline with an intuitive, easy-to-use interface. Lasergene Genomics and its Variant Annotation Database greatly simplify variant analysis so that you can get accurate answers without needing a PhD in bioinformatics. With the recent addition of the powerful Mastermind Genomic Search Engine, the once time-consuming job of variant analysis is now faster, easier, and more accurate than ever before.

CONTACT US

To see a demo of the software, or to learn more about how Lasergene Genomics and Mastermind can fit into your research, email or call us at one of the numbers listed below.

www.dnastar.com
Email: sales@dnastar.com
Phone: 608.258.7420
Toll Free in the U.S. and Canada:
866.511.5090
Call free from the U.K.: 0.808.234.1643
Call free from Germany: 0.800.182.4747