GeneGrid Variant Analysis Workshop

Course Tutorial

For more information please contact:

Genomatix Inc.
3025 Boardwalk, Suite 160
Ann Arbor, MI 48108
(734) 205-5990

Phone: (734) 205-5990
Email: grant@genomatix.com
WWW: http://www.genomatix.com

genomatix

Table of Contents

Introduction to GeneGrid

Genomic variants like single nucleotide polymorphisms (SNPs) or small insertions and deletions (InDels) are of major interest to biologists and clinicians alike.

Their impact ranges from determining your eye color to influencing response to medication. They can cause cardiovascular or neurodegenerative diseases, induce cancer or on the other hand trigger resistance to HIV infection. Identifying causal variants is therefore crucial for the understanding of molecular mechanisms and diagnostics of rare or common diseases.

With NGS technology it is possible to detect the millions of variants within an individual genome through a single experiment. One question remains, though: which are the relevant ones?

With GeneGrid, you can quickly reduce millions of variants to the few or even the single relevant one(s). All known and novel SNPs in your results can be annotated using our extensive genome annotation. You can filter the list for those variants of interest to you, perform trio analyses, compare case and control sets (using multiple samples) or identify somatic SNPs.

You can filter by:

- effects on the amino acid sequence (missense, nonsense etc.)
- amino acid substitution and DNA conservation scores
- different allele frequency scores
- associations with diseases (e.g. from the COSMIC or ClinVar databases), tissues, and pathways

Subsequently you can export the lists and associations. You can also switch to the Genomatix Pathway System to look at networks of the affected genes or view the genomic location of a SNP in our Genome Browser.

Small variant analysis: Leber congenital amaurosis and deafness

Introduction

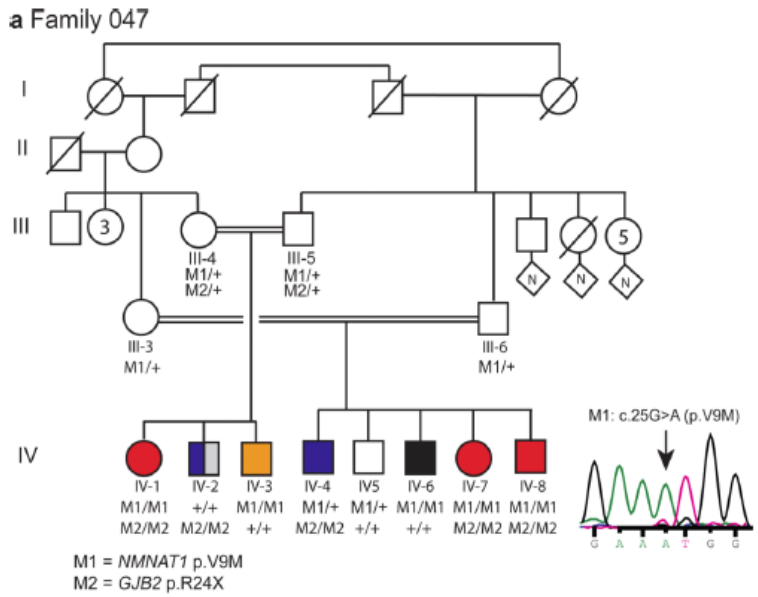
GeneGrid enables you to quickly reduce millions of variants to a few or even the single relevant one(s). All known & novel SNPs and InDels in your results can be annotated using our extensive annotation. This tutorial describes step by step how to annotate and filter the list for those variants of interest to you. You will perform a Trio analysis and easily identify the most likely disease causing variants.

Example for trio analysis

This example will show you how GeneGrid can be applied for a Trio analysis looking for the most likely disease causing variants.

The data in this example is from a whole exome sequencing study (Falk et al., 2012). The authors found a novel homozygous missense mutation (M1=NMNAT1 c.25G>A, p.Val9Met) in *NMNAT1* that likely causes **Leber congenital amaurosis (LCA)** which is a form of inherited retinal degeneration characterized by severe vision loss or blindness.

The following figure shows a large consanguineous Pakistani pedigree, including five children affected with LCA. Exome sequencing also confirmed the presence of a second homozygous mutation (M2=GJB2 c.71G>A, p.Trp24*) in *GJB2*, in children who were affected with **Deafness**.



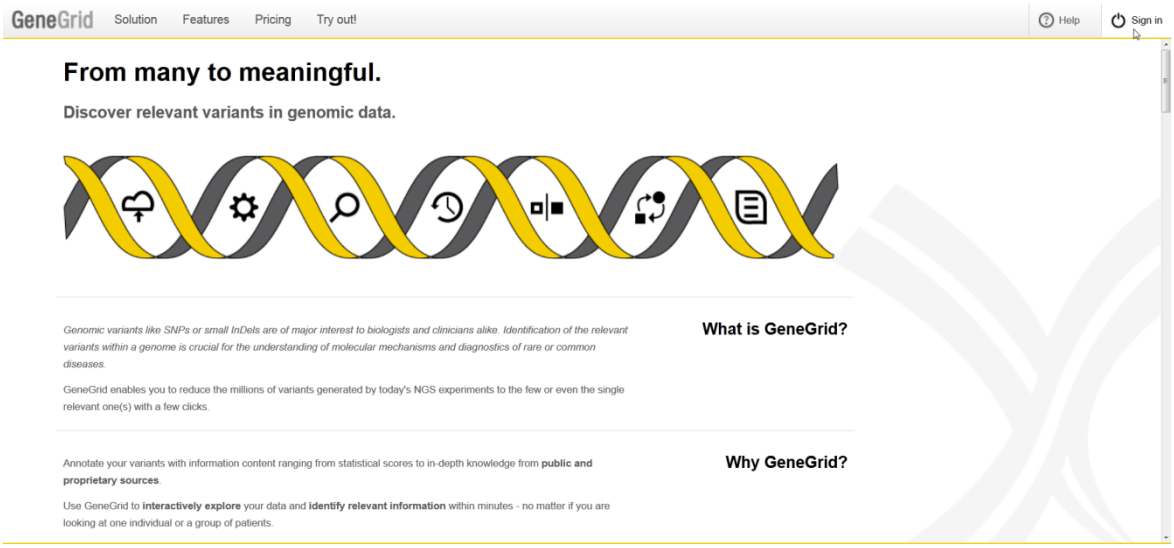
Exome sequence data for the following individuals is available at the NCBI Sequence Read Archive, accession **SRP013517**:

- III-4 mother (M1/+, M2/+)
- III-5 father (M1/+, M2/+)
- IV-1 daughter (M1/M1, M2/M2)
- IV-2 son (+/+, M2/M2)
- IV-3 son (M1/M1, +/+)

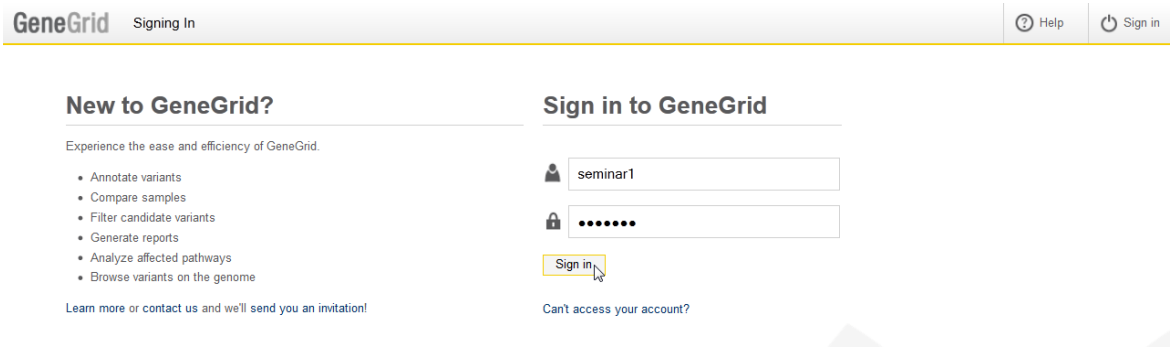
Sequence reads of the individuals IV-1, III-4 and III-5 were mapped to the human reference genome (GRCh38) using the **Genomatix Mining Station (GMS)**. SAMtools (version 1.2, Li et al., Bioinformatics, 2009) was used to call SNVs and InDels jointly with all three samples. The output format of the variant calling step is *VCF* which stands for *Variant Call Format*. It contains all the genomic positions of the variants and the genotypes of the samples and is the required input format to get started with GeneGrid.

The final VCF (Variant Call Format) file (“LCA047_Trio_Demo.vcf.gz”), containing predicted variants (SNPs, small insertions and deletions) in all three individuals, has already been uploaded into a shared project.

Start your browser and open the home page of GeneGrid (https://genegrid.genomatix.com/grid/home). You should see a page like below. Press the *Sign in* button in upper right hand corner to go to the login page.



Enter your user name and password.

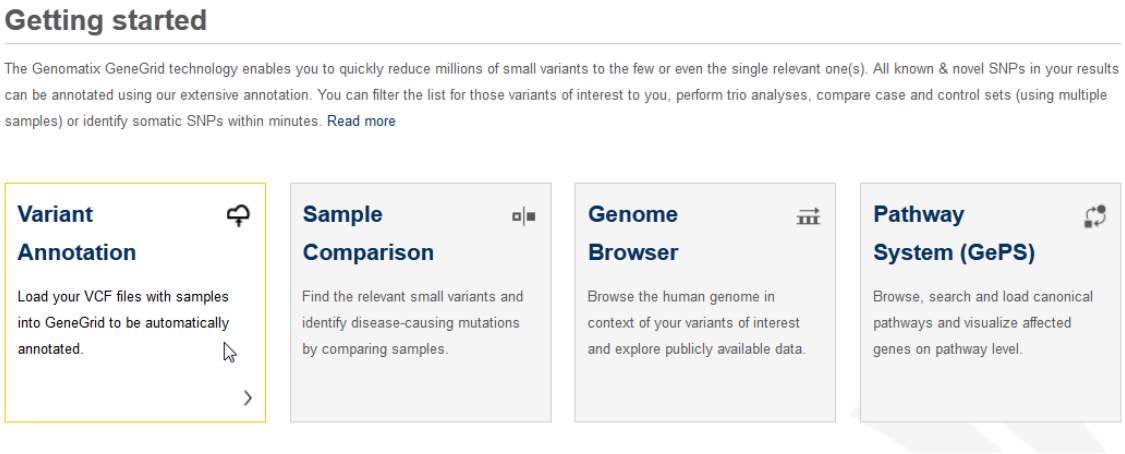


Import sample (demo only)

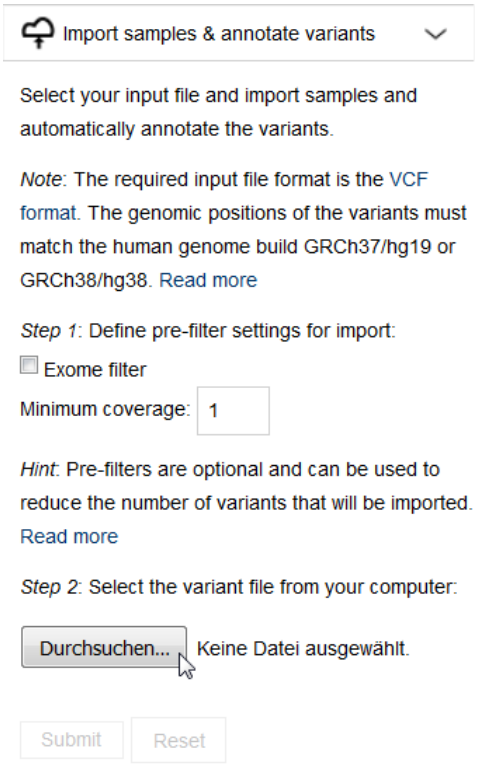
The first step to use GeneGrid is to **import** and **annotate** the predicted variants from the VCF file.

Annotation of variants takes up to one hour, depending on the number of samples and variants that are in the input file. In the course, the necessary data have already been uploaded beforehand into your account in order to save time. The following demonstrates how VCF files can be uploaded.

VCF files are uploaded in the section ‘Variant Annotation’:



VCF files can be imported from the left sidebar. In the first step it is possible to define any pre-filters if it is desired to skip certain variants completely. Defining pre-filters is completely *optional* and based on the default settings all variants having at least a coverage of *one* will be imported. The second step is to select and upload the VCF file from your local computer by clicking on *choose file* or *browse* (the actual label depends on your browser) and selecting the VCF file. The upload starts automatically.

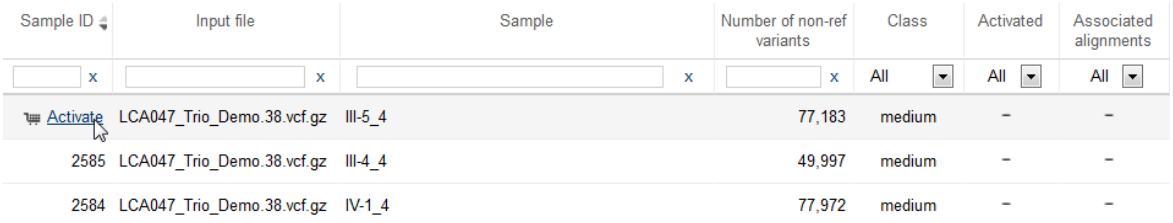


The upload progress is shown on the screen.

Annotated samples are accessible from two different views. The samples view contains a table with a detailed listing of imported samples. From this view it is possible to show details about the sample, open the variants view for the sample or start a comparison analysis. An additional separate page where all the samples can be found is the Result Management which is the interface to rename sample names, edit comments for samples and most importantly to remove samples. The Result Management is also directly accessible from the main menu.

For the next steps we use the first view and load the samples view by selecting Variant Annotation from the main menu. All imported samples will be displayed with information such as source file, sample name, number of non-ref variants, class and activation status.

Newly annotated samples need to be activated in the samples view. They have an *Activate* link in the *Sample ID* column; clicking on it loads the *purchase view* for purchasing the sample with credits.

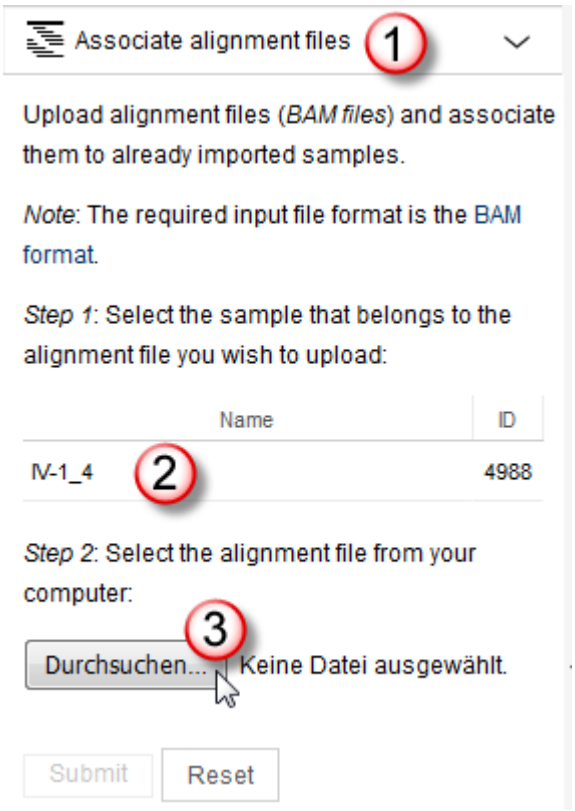


Sample ID	Input file	Sample	Number of non-ref variants	Class	Activated	Associated alignments
x	x	x	x	All	All	All
Activate	LCA047_Trio_Demo.38.vcf.gz	III-5_4	77,183	medium	-	-
2585	LCA047_Trio_Demo.38.vcf.gz	III-4_4	49,997	medium	-	-
2584	LCA047_Trio_Demo.38.vcf.gz	IV-1_4	77,972	medium	-	-

Associate BAM file with uploaded VCF data (demo only)

After a sample has been activated, an alignment file in the BAM format can be associated with an annotated sample. This is optional and can be used later for visualization of the read coverage of variants in the genome browser.

To associate a BAM file, first select the *Associate alignment files* option in the panel on the left. Then drag and drop the sample to which you want to add the BAM file from the sample list into the sample field. Lastly, press the *Browse* button and select the corresponding alignment file from your computer. The file upload will start automatically.



View VCF sample statistics and metadata

For the statistics view, please use the shared data in the **TutorialCRCh38** workspace (check column *Workspace*):

These data have already been activated, and BAM files have been associated. The content of an individual file can be displayed by clicking the *Open* link which appears when hovering the mouse over the *Sample ID* entry. However at this point we are not interested in viewing individual samples by themselves. We'll just have a look at sample metadata.

ID	Input file	Name	Number of non-ref variants	Class	Activated	Associated alignments
	x	x	x	x	All	All	All
2520	LCA047_Trio_Demo.38.vcf.gz	III-5	77,183	medium	✓	✓
2519	LCA047_Trio_Demo.38.vcf.gz	III-4	49,997	medium	✓	✓
2518	LCA047_Trio_Demo.38.vcf.gz	IV-1	77,972	medium	✓	✓

To see sample metadata, click on the corresponding sample name. This will display a tabbed detail view.

In the *Details* tab, you'll find basic information about the sample from the VCF file, such as the sample name or the number of non-reference variants.

Details | Warnings | Alignment file | Annotation distribution (26) | Quality control (26)

Property	Value
Sample ID	2520
Storage ID	3332
Input file	LCA047_Trio_Demo.38.vcf.gz
Name	III-5
Number of variants in database	113,807
Number of total variants	91,428
Number of non-ref variants	77,183

The *Alignment file* tab is available for samples with an associated BAM file. It displays basic information about the BAM file.

Details | Warnings | Alignment file | Annotation distribution (26) | Quality control (26)

Property	Value
File name	SRR504517.group.38.dup.bam
Number of aligned reads	158,659,616
Date	2016-02-12 11:24:09

The *Annotation distribution* tab shows extended variant statistics for the complete sample and for each chromosome.

Details | Warnings | Alignment file | Annotation distribution (26) | Quality control (26)

Chr	Annotated variants in results	Region blocks	Total variants	Reference mismatches	Allele length exceeded	Missing genotypes	Coverage pre-filtered	Exome pre-filtered	Reference calls	Non-ref calls
Total	113,543	0	98,208	0	0	6,780	0	0	14,245	77,183
1	10,932	0	9,784	0	0	632	0	0	1,422	7,730
2	7,799	0	6,780	0	0	427	0	0	857	5,496
3	6,275	0	5,492	0	0	306	0	0	1,005	4,181
4	4,219	0	3,811	0	0	209	0	0	527	3,075
5	6,048	0	4,377	0	0	255	0	0	670	3,452

Finally, the *Quality control* tab lists a number of quality metrics, such as transition / transversion ratio.

Details | Warnings | Alignment file | Annotation distribution (26) | Quality control (26)

Chr	Allele frequencies (%)	dbSNP novelty SNPs (%)	dbSNP novelty indels (%)	Clinical significance	Transitions	Transversions	Transition transversion ratio	Heterozygous	Homozygous	Heterozygous homozygous ratio
Total	93.04	1.81	23.04	3,848	48,293	20,145	2.40	38,549	29,862	1.29
1	93.96	1.52	23.94	328	4,856	1,982	2.45	3,996	2,840	1.41
2	94.18	1.26	18.84	360	3,359	1,472	2.28	2,713	2,115	1.28
3	92.75	1.62	26.00	190	2,560	1,075	2.38	2,132	1,501	1.42
4	94.21	0.86	22.63	131	1,850	819	2.26	1,546	1,123	1.38
5	93.77	0.83	19.23	224	2,177	835	2.61	1,653	1,358	1.22

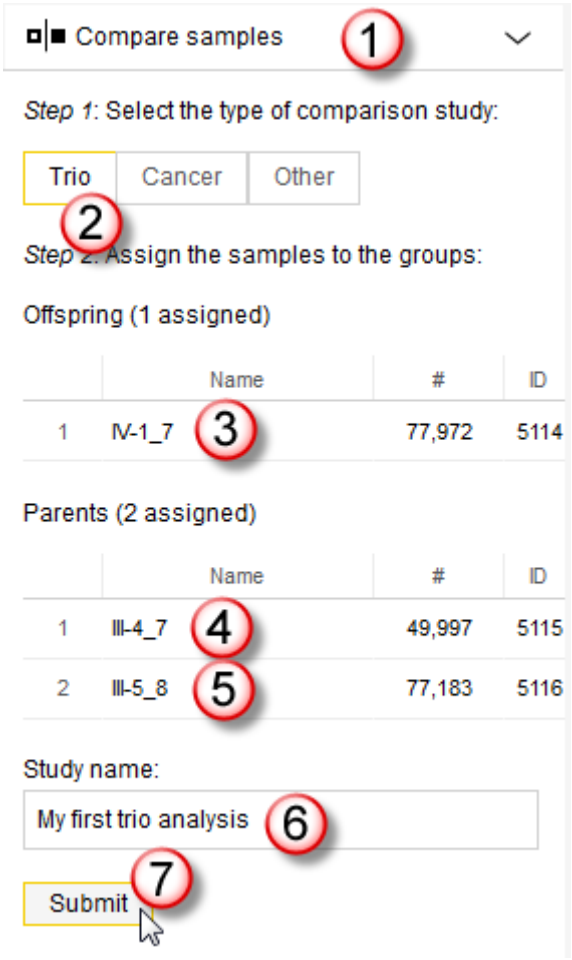
We will continue with comparing the uploaded Trio samples among one other.

Compare samples

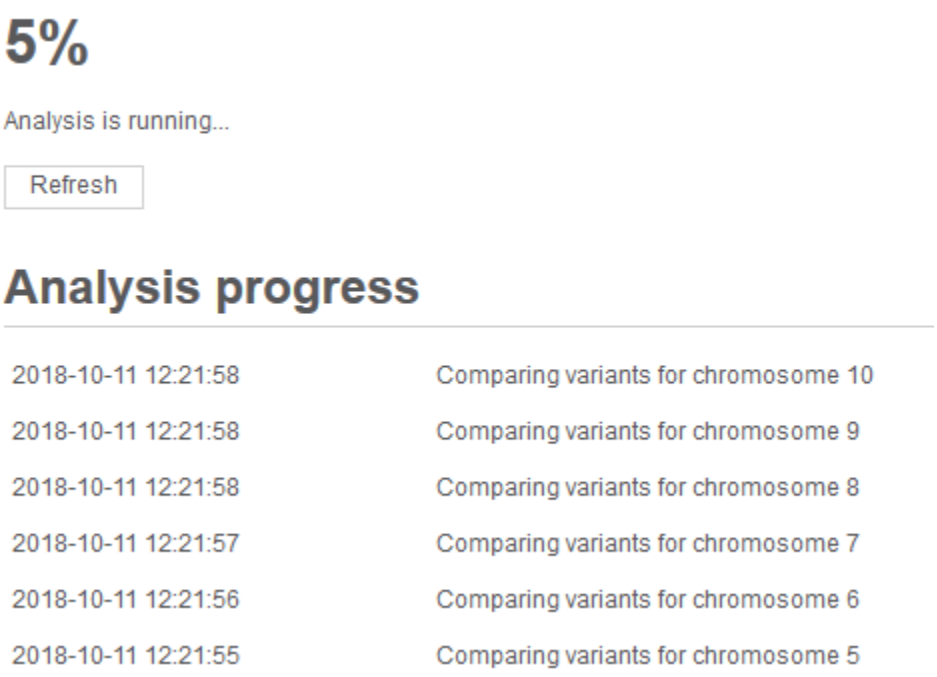
For running a sample comparison analysis, please use the pre-uploaded data data in **your own MyProject** workspace (check columns *Owner* and *Workspace*). Please note that you can run comparisons only on samples you own or on shared samples for which you have write access. During the workshop, the samples you find in shared workspaces are read-only for you, therefore you can only view them.

After importing and annotating variants, we're ready to perform a **sample comparison analysis**. If you are on the main page, you can either click on the Sample Comparison field or again use the main menu select. On the sample view, this time we use the *Compare samples* section in the left sidebar.

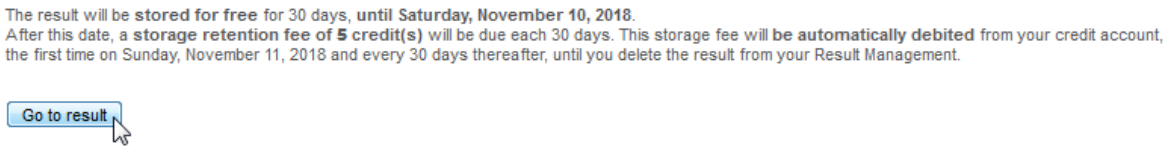
First, we select the type of comparison study we want to perform. Here we select *Trio* and assign our samples to the groups that appear below. Drag the sample *IV-1* (this is the affected daughter) from the table on the right side and drop it in the *Offspring* (affected) group. Repeat this for the sample *III-4* and *III-5* (mother and father) but drop both samples in the *Parents* (not affected) group. Finally, give the analysis a name (*My first trio analysis* is used as title in this example) and hit the *Submit* button.



The sample comparison will start and you should see the following progress info:

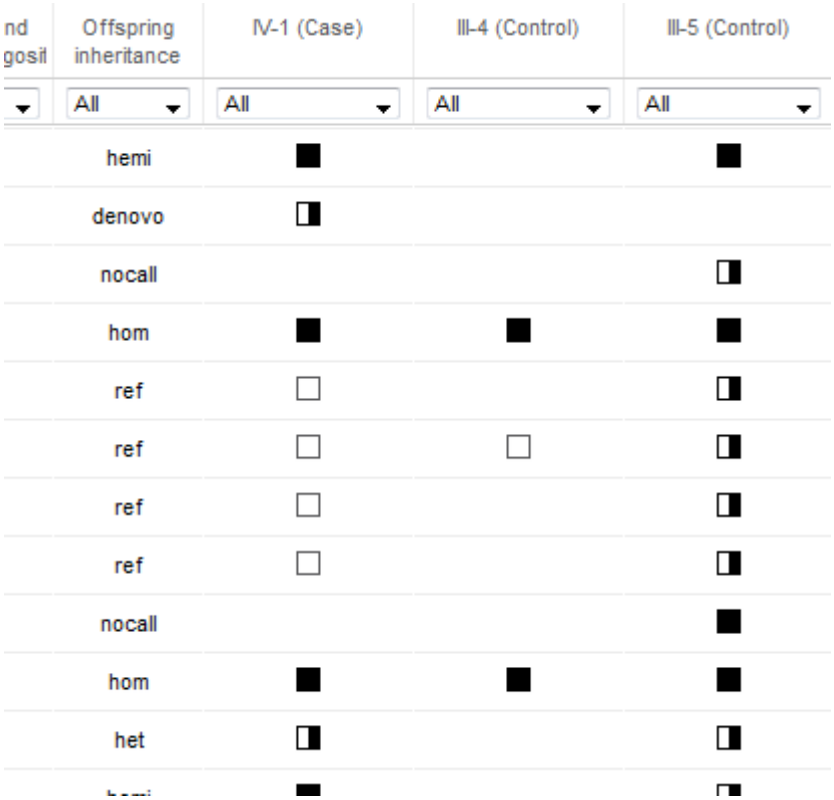


You can see that the analysis has been submitted successfully. The analysis will take a couple of minutes. Once the comparative analysis is finished, you will be redirected automatically to the activation page.



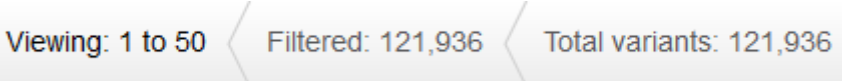
Please click on the *Go to result* button. You can safely ignore the information text here, which is only relevant for paid credit-based accounts.

After proceeding the result table with the **variants** is loaded. This table is our workspace and contains several general columns and an additional column for each sample at the very end. The sample column will display the genotype of each sample using a symbol. A *homozygous* variant call is a filled black square, a *heterozygous* variant call is a half-filled square and a *reference* call is a white square. If GeneGrid has no or low quality information about the genotype of a certain sample, an empty cell is displayed to indicate a *no call*.



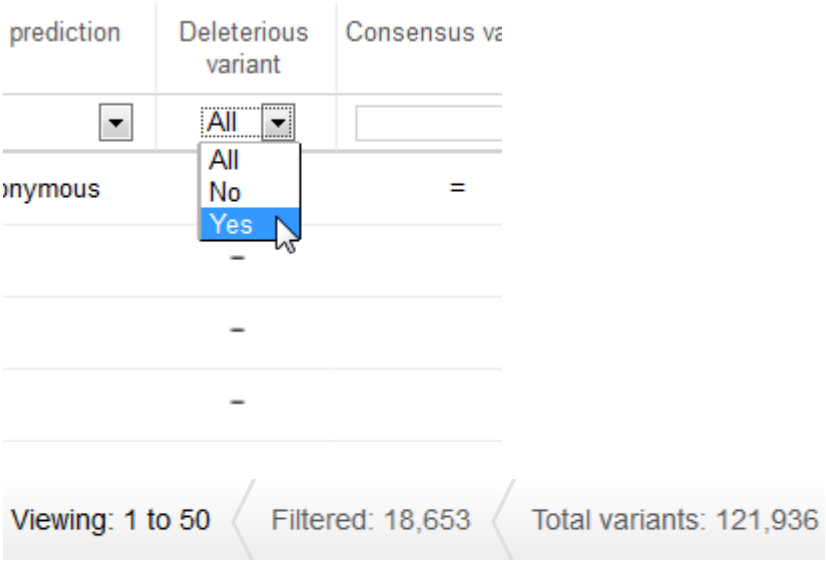
Filter variants

Let's see how we can **filter** for the most likely disease causing variants. The main table of the comparative sample analysis contains **121,936** variants. You find the total number of variants at the bottom right of the table.

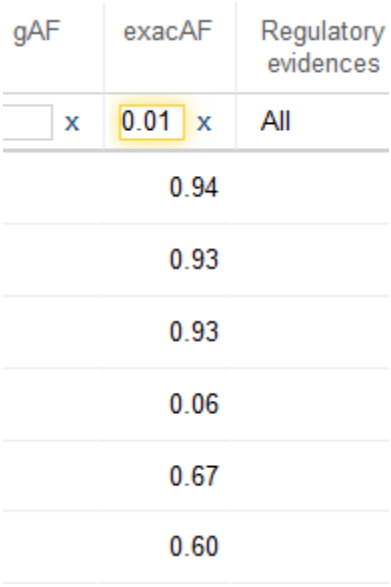


Above the table there is a filter bar, through which users can define filter criteria for each column. In some cases drop-down boxes lists the available search terms that can be filtered for, in other cases typing a free text and clicking enter will activate the filter. If filter criteria for multiple columns are defined, a row has to match all criteria of the filter bar otherwise it will be hidden.

First, we consider only variants that are **deleterious** (e.g. non-synonymous or frameshift mutations) and alter the protein sequence or hit a canonical splice-site. Please select *Yes* from the drop-down box in the *Deleterious variant* column. The number of filtered variants should go down to **18,653**.

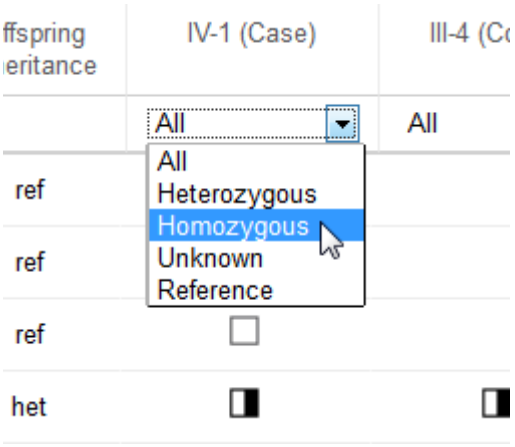


When searching for rare diseases, it is very helpful to compare the variants found in affected individuals against background populations. For each variant within a background population, a **global allele frequency** can be calculated. In this example, we use the allele frequencies from the Exome Aggregation Consortium (ExAC) project. Enter *0.01* in the filter field of the *exacAF* column and press the return key. This filter removes all variants that occurred in more than 1% of the ExAC project. The number of filtered variants should now be **3,402**



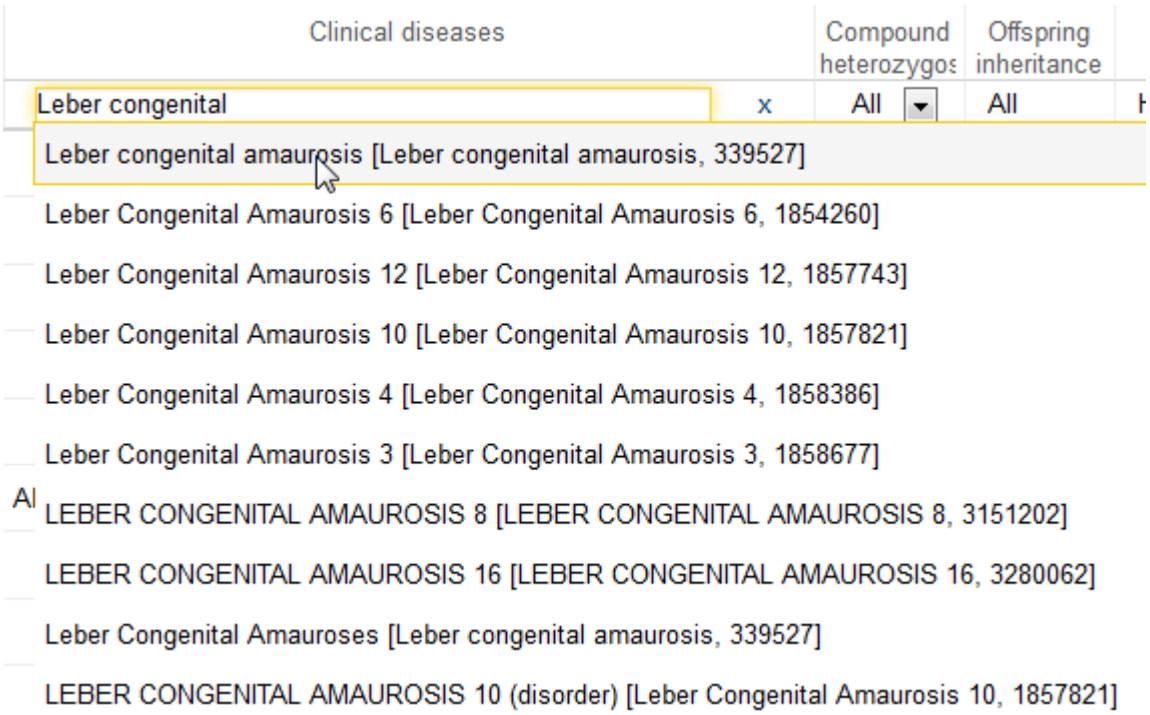
In our trio example, both parents are *unaffected* but the daughter is *affected*. In this case our **genotype** search strategy includes all filters for an *autosomal recessive* disease.

First, we will search for a homozygous mutation in the daughter. Select *Homozygous* in the drop-down box of the *IV-1* column, and *Heterozygous* for both parental genotypes. This reduces the number of variants to **45**.



We will use **gene-disease associations** to filter the variant list and link the remaining variants to the diagnosed disease *Leber congenital amaurosis*. Disease association columns are optional columns and are not visible by default. Click on the settings wheel above the table on the left. Additional annotations (e.g. BLOSUM, SIFT, PhyloP) can be selected for display here.

Select the annotation *Clinical diseases*. A new column is now added to the main table. In the filter bar, enter the disease term *Leber congenital amaurosis*. While you enter the term, a suggestion will pop up.



Choose the very first term and hit enter to start the filter process. **One variant remains**. It is a homozygous missense mutation on chromosome 1 at position 9,972,098 in the gene *NMNAT1*.

Clicking on the variant row, additional details are shown at the bottom. You will find more detailed information on the variant, like the coverage, SIFT, and other protein effect scores. As you can see in the *Sample details* table the position of the *NMNAT1* mutation is well covered in all three samples (>30 reads).

Sample details (3) | Transcript effects (9) | dbSNP (1) | ClinVar (1) | Somatic mutations | COSMIC | Literature diseases (39) | Clinical c

Sample	Variant	Chr	Position (bp)	Variant	Zygosity	Ref	Alt	Alt2 allele	Genotype	Gene symbol	Known gene	Effect prediction	Deleterious	Quality	Coverage
III-4	396	1	9,972,098	snv	het	G	A		◨	NMNAT1	✓	missense	✓	728.00	35
III-5	694	1	9,972,098	snv	het	G	A		◨	NMNAT1	✓	missense	✓	728.00	98
IV-1	709	1	9,972,098	snv	hom	G	A		■	NMNAT1	✓	missense	✓	728.00	116

Clicking on the *Transcript effects* tab, you get additional transcript information.

Sample details (3) | Transcript effects (9) | dbSNP (1) | ClinVar (1) | Somatic mutations | COSMI

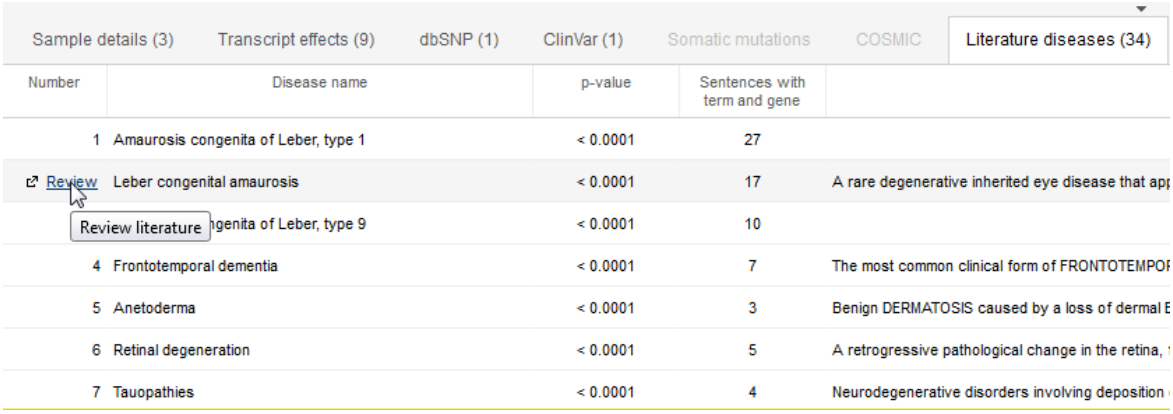
Ref allele	Alt allele	Ref protein sequence	Alt protein sequence	Gene symbol	Transcript accession	Transcript version	Strand	Type	Genomic feature
G	A	V	M	NMNAT1	NM_001297778	1	+	protein coding	cds • exon
G	A	V	M	NMNAT1	NM_022787	3	+	protein coding	cds • exon
G	A	V	M	NMNAT1	NM_001297779	1	+	protein coding	cds • exon
G	A	V	M	NMNAT1	ENST0000037720	5	+	protein coding	cds • exon
G	A	V	M	NMNAT1	ENST0000040319	5	+	protein coding	cds • exon
G	A	V	M	NMNAT1	XM_017002108	1	+	protein coding	cds • exon

This includes the position of the amino acid change in the protein sequence. Also, the variation is predicted as damaging (SIFT < 0.05)

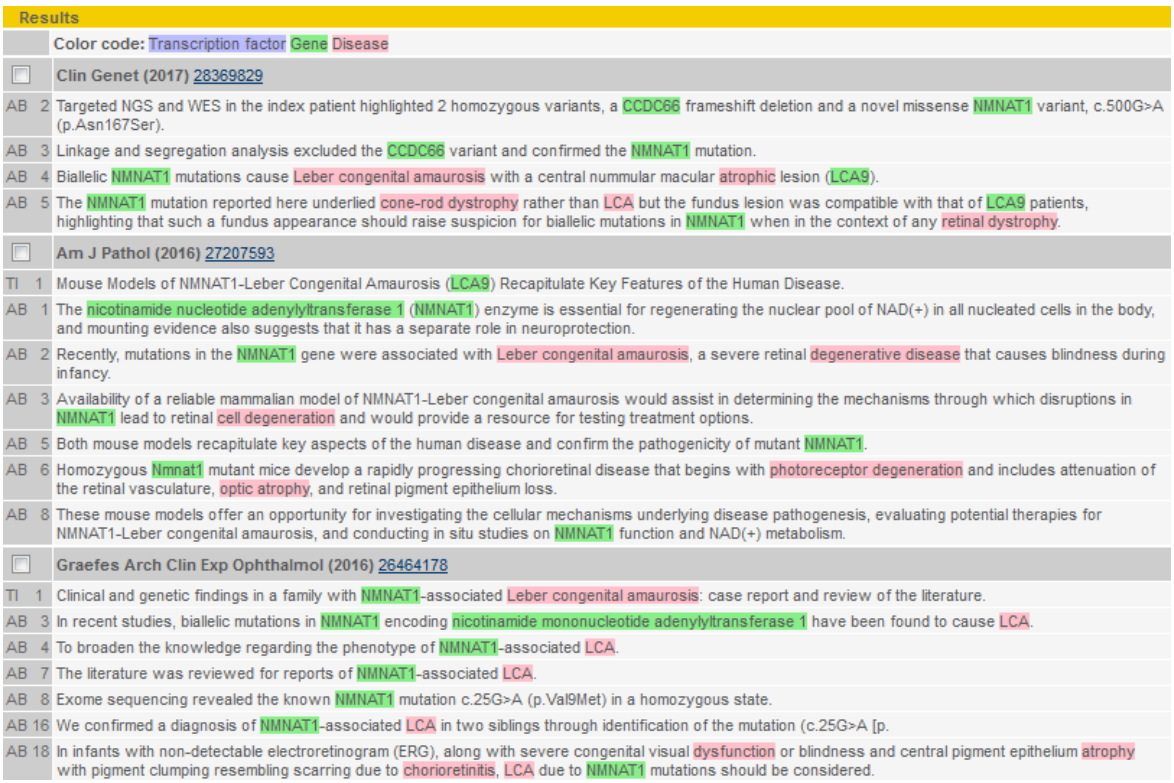
C | Literature diseases (39) | Clinical diseases (2) | Gene details | Citations (2) | Literature tissues (27) | Ca

Effect prediction	Deleterious	Low confidence	Variation coding	Variation protein	Coding position (bp)	Relative coding position	BLOSUM	SIFT score	SIFT damaging
missense	✓	-	25G>A	Val9Met	25	0.03	1	0.01	✓
missense	✓	-	25G>A	Val9Met	25	0.03	1	0.01	✓
missense	✓	-	25G>A	Val9Met	25	0.05	1	0.01	✓
missense	✓	-	25G>A	Val9Met	25	0.03	1	0.01	✓
missense	✓	-	25G>A	Val9Met	25	0.05	1	0.01	✓
missense	✓	✓	25G>A	Val9Met	25	0.06	1		

The tab *Literature diseases* lists all disease associations of that particular gene, including descriptions and links to **Genomatix LitInspector** results. There you can find all publications where the gene disease association has been found.

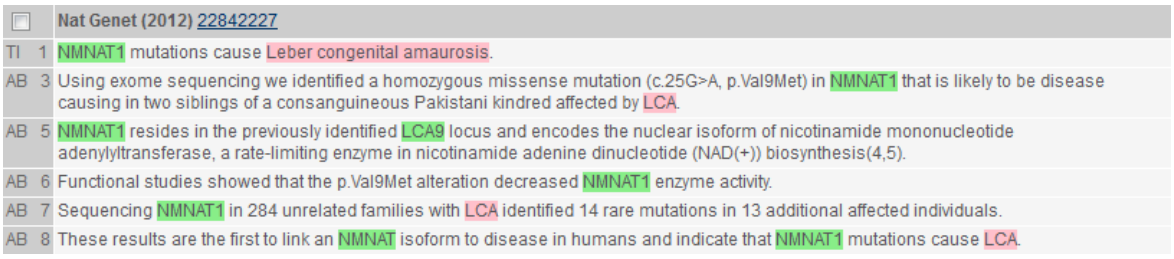


We can also get a list of all publications where the gene has been linked to a given disease. By moving the mouse over the row with *Leber Congenital Amaurosis* a small link called *Review* appears. From there you can jump directly to the evidence listing from *LitInspector*.



Results		
		Color code: Transcription factor Gene Disease
☐		**Clin Genet (2017) 28369829**
AB	2	Targeted NGS and WES in the index patient highlighted 2 homozygous variants, a CCDC66 frameshift deletion and a novel missense NMNAT1 variant, c.500G>A (p.Asn167Ser).
AB	3	Linkage and segregation analysis excluded the CCDC66 variant and confirmed the NMNAT1 mutation.
AB	4	Biallelic NMNAT1 mutations cause Leber congenital amaurosis with a central nummular macular atrophic lesion (LCA9).
AB	5	The NMNAT1 mutation reported here underlied cone-rod dystrophy rather than LCA but the fundus lesion was compatible with that of LCA9 patients, highlighting that such a fundus appearance should raise suspicion for biallelic mutations in NMNAT1 when in the context of any retinal dystrophy.
☐		**Am J Pathol (2016) 27207593**
TI	1	Mouse Models of NMNAT1-Leber Congenital Amaurosis (LCA9) Recapitulate Key Features of the Human Disease.
AB	1	The nicotinamide nucleotide adenylyltransferase 1 (NMNAT1) enzyme is essential for regenerating the nuclear pool of NAD(+) in all nucleated cells in the body, and mounting evidence also suggests that it has a separate role in neuroprotection.
AB	2	Recently, mutations in the NMNAT1 gene were associated with Leber congenital amaurosis, a severe retinal degenerative disease that causes blindness during infancy.
AB	3	Availability of a reliable mammalian model of NMNAT1-Leber congenital amaurosis would assist in determining the mechanisms through which disruptions in NMNAT1 lead to retinal cell degeneration and would provide a resource for testing treatment options.
AB	5	Both mouse models recapitulate key aspects of the human disease and confirm the pathogenicity of mutant NMNAT1.
AB	6	Homozygous Nmnat1 mutant mice develop a rapidly progressing chorioretinal disease that begins with photoreceptor degeneration and includes attenuation of the retinal vasculature, optic atrophy, and retinal pigment epithelium loss.
AB	8	These mouse models offer an opportunity for investigating the cellular mechanisms underlying disease pathogenesis, evaluating potential therapies for NMNAT1-Leber congenital amaurosis, and conducting in situ studies on NMNAT1 function and NAD(+) metabolism.
☐		**Graefes Arch Clin Exp Ophthalmol (2016) 26464178**
TI	1	Clinical and genetic findings in a family with NMNAT1-associated Leber congenital amaurosis: case report and review of the literature.
AB	3	In recent studies, biallelic mutations in NMNAT1 encoding nicotinamide mononucleotide adenylyltransferase 1 have been found to cause LCA.
AB	4	To broaden the knowledge regarding the phenotype of NMNAT1-associated LCA.
AB	7	The literature was reviewed for reports of NMNAT1-associated LCA.
AB	8	Exome sequencing revealed the known NMNAT1 mutation c.25G>A (p.Val9Met) in a homozygous state.
AB	16	We confirmed a diagnosis of NMNAT1-associated LCA in two siblings through identification of the mutation (c.25G>A [p.
AB	18	In infants with non-detectable electroretinogram (ERG), along with severe congenital visual dysfunction or blindness and central pigment epithelium atrophy with pigment clumping resembling scarring due to chorioretinitis, LCA due to NMNAT1 mutations should be considered.

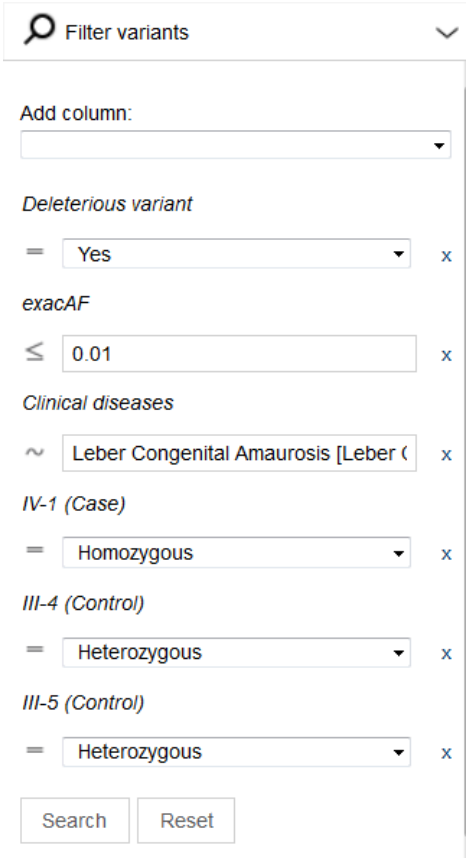
As of October 2018, there are 15 different publications containing both, *NMNAT1* and *Leber congenital amaurosis*, in the same abstract. Scroll down to the last entry.



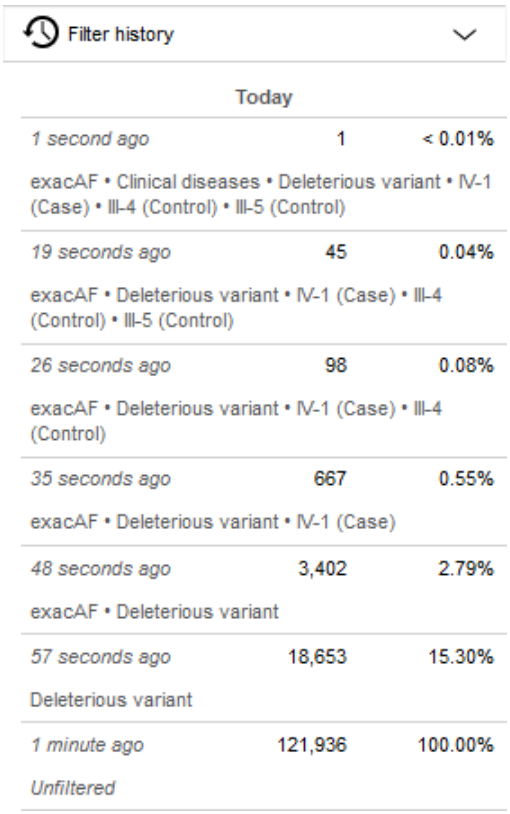
☐		**Nat Genet (2012) 22842227**
TI	1	NMNAT1 mutations cause Leber congenital amaurosis.
AB	3	Using exome sequencing we identified a homozygous missense mutation (c.25G>A, p.Val9Met) in NMNAT1 that is likely to be disease causing in two siblings of a consanguineous Pakistani kindred affected by LCA.
AB	5	NMNAT1 resides in the previously identified LCA9 locus and encodes the nuclear isoform of nicotinamide mononucleotide adenylyltransferase, a rate-limiting enzyme in nicotinamide adenine dinucleotide (NAD(+)) biosynthesis(4,5).
AB	6	Functional studies showed that the p.Val9Met alteration decreased NMNAT1 enzyme activity.
AB	7	Sequencing NMNAT1 in 284 unrelated families with LCA identified 14 rare mutations in 13 additional affected individuals.
AB	8	These results are the first to link an NMNAT isoform to disease in humans and indicate that NMNAT1 mutations cause LCA.

This is the paper Falk et al. (2012) from our example describing the exact same missense mutation in *NMNAT1* (c.25G>A, p.Val9Met).

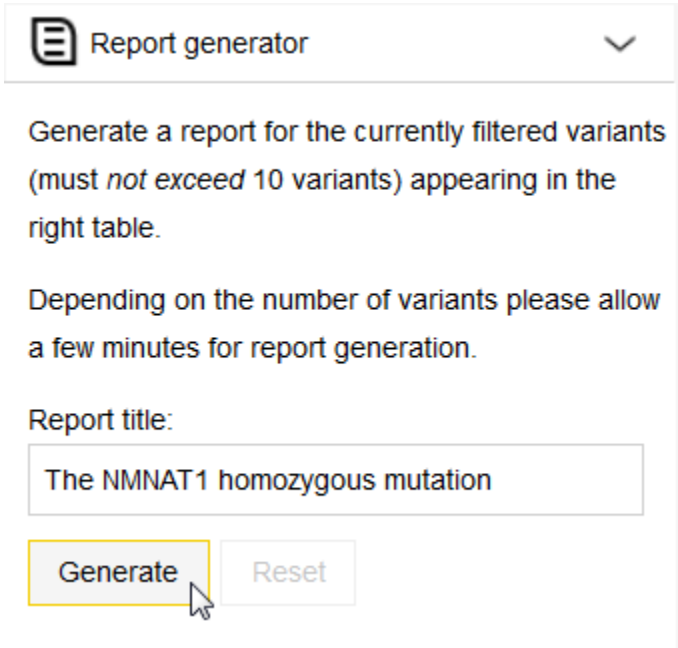
Back in our *variants view* we can take a look at our current filter definition shown in the left sidebar.



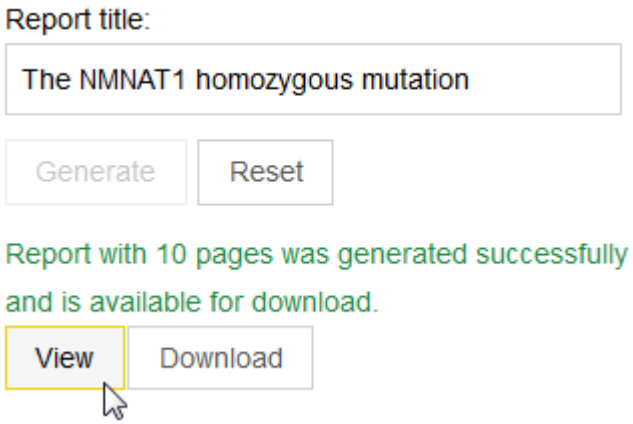
Another convenient tool in the sidebar is the *filter history*. It lists in reverse chronological order the filter steps we have performed so far up until we filtered down the variants to the single one.



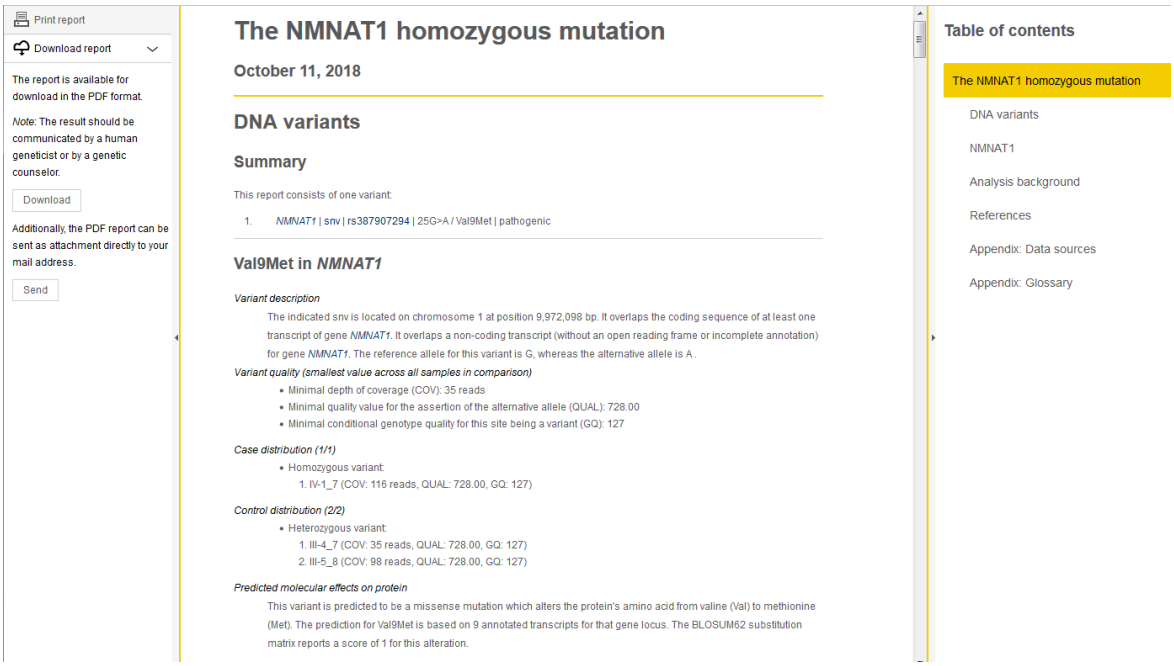
On the left side of the main table you will find a *Report generator* tab. It allows for generating reports for up to 10 filtered variants. Type in a title and hit the *Generate* button and wait for the PDF file.



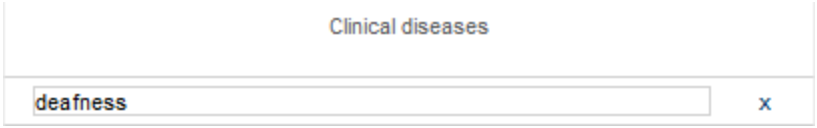
Press the *View* button to display the report in the browser.



You can also use the *Download* button to save the report as a PDF document or the *Send* button on the report page to send the report directly to your e-mail address.



In our example, the daughter is also affected with **Deafness**. Enter the disease term *deafness* into the filter field of *Clinical diseases*.

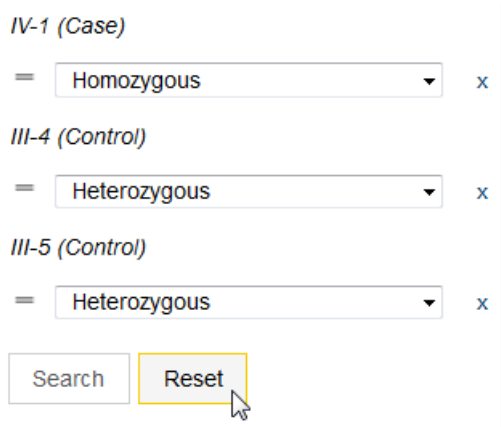


Two variants remain in the list. One of them is a homozygous nonsense mutation in the gene *GJB2* which has also been described in the publication.

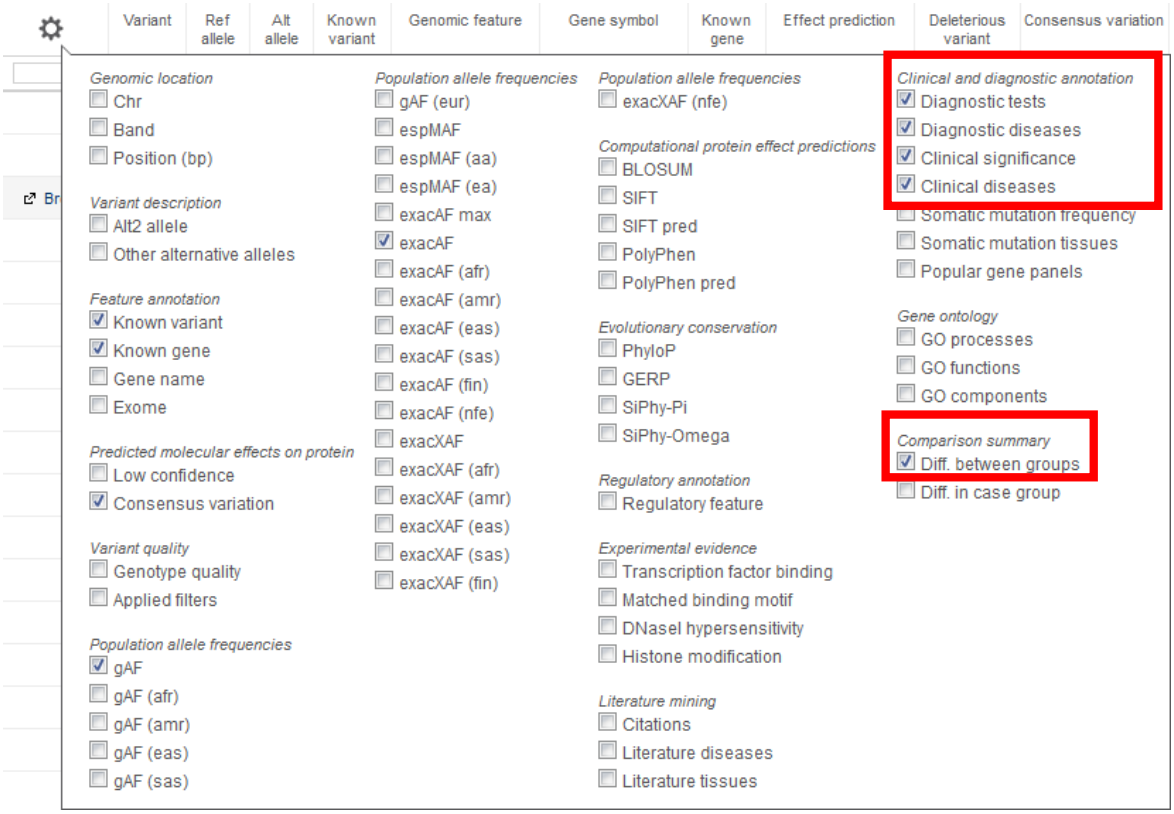
Variant	Ref allele	Alt allele	Known variant	Genomic feature	Gene symbol	Known gene	Effect prediction	Deleterious variant	Consensus variation
All	x	x	All	All	x	All	All	Yes	x
snv	C	T	✓	cds • exon	GJB2	✓	nonsense	✓	Trp24Ter

Advanced usage

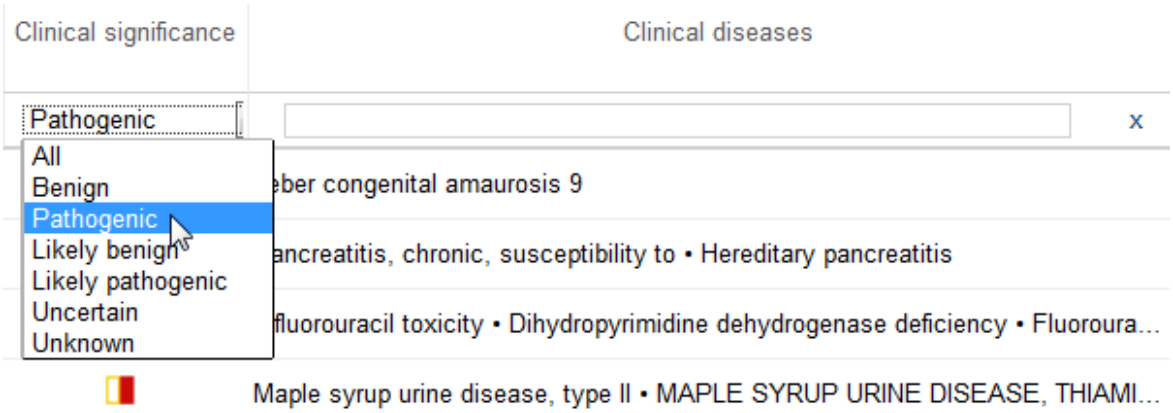
Let's try a completely different strategy where we are mainly interested to see if we have variants that had been discovered already by other researchers. To restart our filters, you can use the *Reset* button on the left sidebar.



We are going select different columns from the optional column settings. Please select the first 4 columns from the section *Clinical and diagnostic annotation* and the column *Diff. between groups* from the section *Comparison summary*.

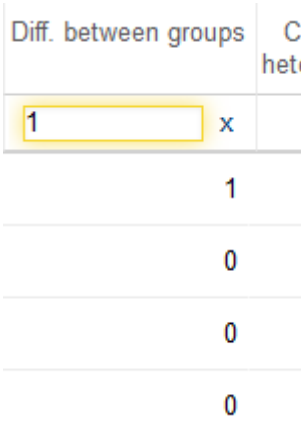


The *Clinical significance* can be used to check whether there exists any annotation in ClinVar at the genomic positions of our variants regardless if the actual variants in our tables are synonymous, missense, or any other kind of effect category. Select *Pathogenic* from the *Clinical significance* drop-down list to filter the variant list.



After applying this filter the list of variants shrinks to merely **31**. This filter depends vastly on the content of ClinVar which is steadily increasing and if one of our variants had not been reported in ClinVar we would have missed it at that point. Nevertheless it is a valid strategy to quickly overlap for known variants in the ClinVar set.

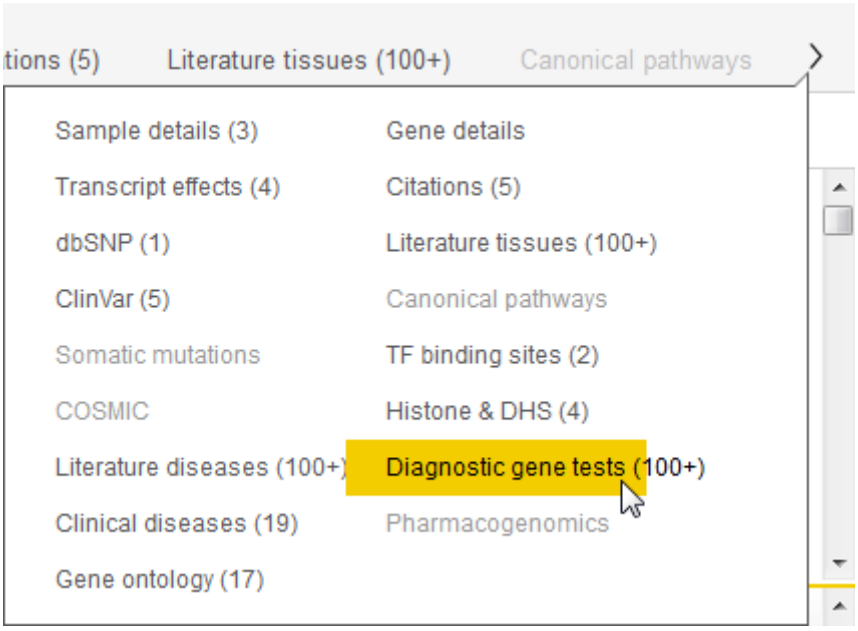
As second filter setting we make sure that our affected sample has to have a different genotype that the unaffected parents. We set *1* as number of *Diff. between groups* which is a very general column to filter for the number of samples that are at different between both groups.



We are now at **7** variants which is already a feasible number of variants to go through individually. Adding the genotype filter *Homozygous* for the affected

sample *IV-1* reduces the list further to **4** variants. This list still includes both previously indicated variants **p.Val9Met** in *NMNAT1* and **p.Trp24Ter** in *GJB2*.

Further inspection gives additional valuable information. *GJB2* has more than 100 diagnostic tests available in Genetic Testing Registry (GTR). The details are accessible in the *Diagnostic gene tests* tab in the details view.

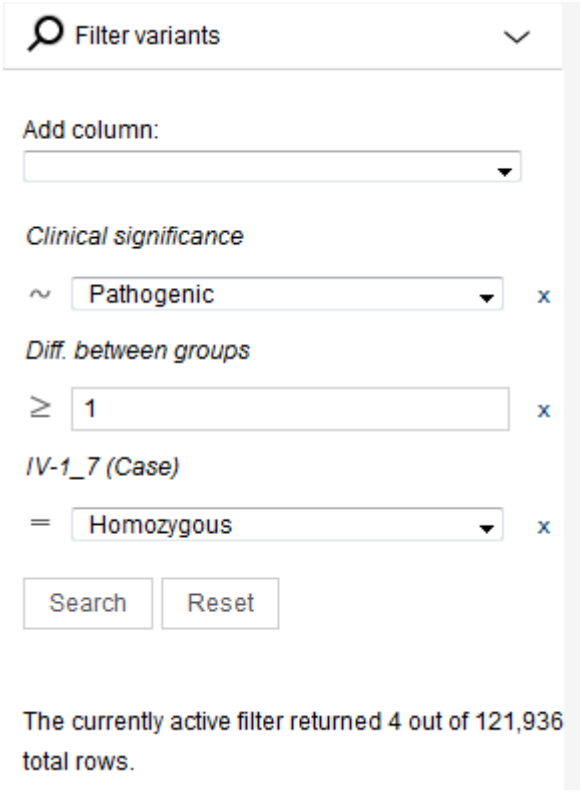


The majority of tests relate to some type of deafness.

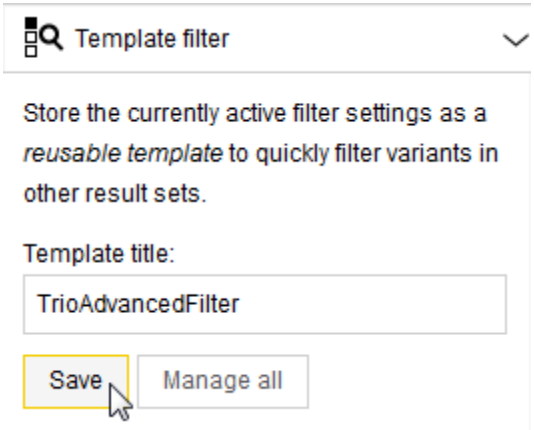
Sample details (3) Transcript effects (4) dbSNP (1) ClinVar (5) Somatic mutations COSMIC Literature diseases (100+) Clinical diseas

Number	Disease	Term ID	Number of tests	Countries
1	Deafness, autosomal recessive 1A	C2673759	91	United States • Spain • Germany • Portugal • Canad...
2	Deafness, autosomal dominant 3a	C2675750	50	United States • Germany • Spain • Austria • Canada ...
3	Keratitis-ichthyosis-deafness syndrome, autosomal dominant	C1835678	32	United States • Germany • Austria
4	Keratoderma palmoplantar deafness	C1835672	26	United States • Germany • Austria
5	Hystrix-like ichthyosis with deafness	C1865234	26	United States • Germany • Austria
6	Mutilating keratoderma	C0265964	23	United States • Germany • Austria
7	Knuckle pads, deafness AND leukonychia syndrome	C0266004	20	United States • Germany • Austria

You can save filter settings for later use in another comparison of the same type. To do this for the current settings, please make sure that your settings are as shown below.



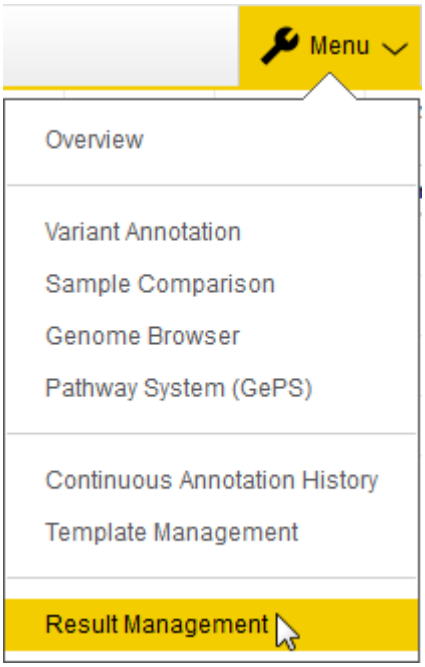
Then open the *Template filter* section in the control panel, provide a name for your filter, e.g. *TrioAdvancedFilter*, and click on *Save*. We will use it in the next step.



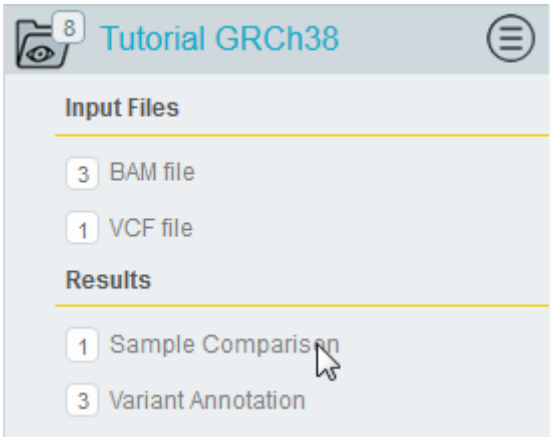
View samples in Genome Browser

Common practice is to examine variants in context of the underlying alignments. A sample can be associated with a BAM file containing all the alignments.

The samples in the comparison that you generated yourself have no BAM files associated. However there is already a pre-calculated copy of this comparison in a shared workspace which you can use for this purpose. From the taskbar menu, select *Result Management.*



Click on the *Sample Comparison* entry in the *Tutorial CRCh38* project



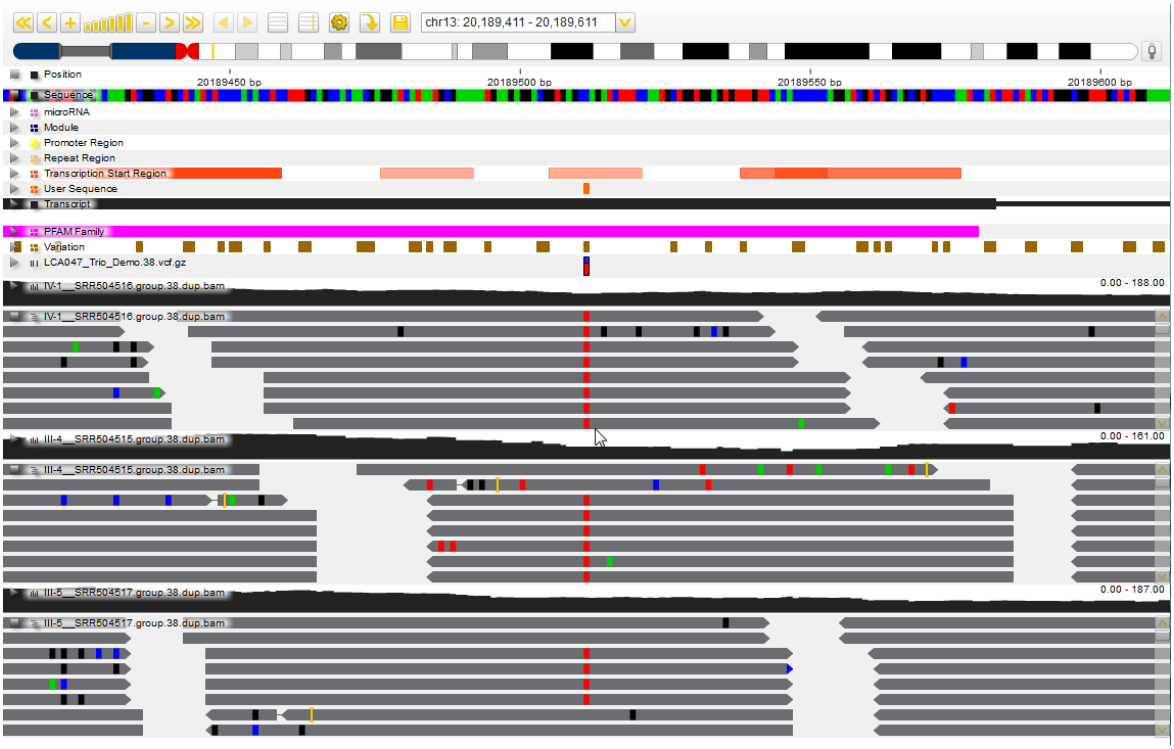
In the comparison list, double-click on the entry *Tutorial_Trio_Analysis_Workflow*.

Open the Template filter section once more, and click on the *TrioAdvancedFilter* entry to apply it to the current comparison.

To directly access the Genome Browser from the variant list just move the mouse over a variant row and the hit the *Browse* button.

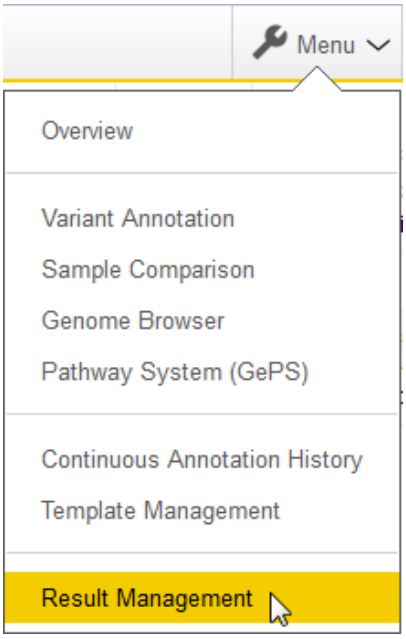
	Variant	Ref allele	Alt allele	Known variant	Genomic feature	Gene symbol
	All	x	x	All	All	x
776	snv	G	A	✓	cds • exon	NMNAT1
69968	snv	C	T	✓	cds • exon	ACTN3
Browse	snv	C	T	✓	cds • exon	GJB2
119762	snv	G	C	✓	cds • exon • intro...	ARSA

The associated BAM files are automatically loaded in the browser.

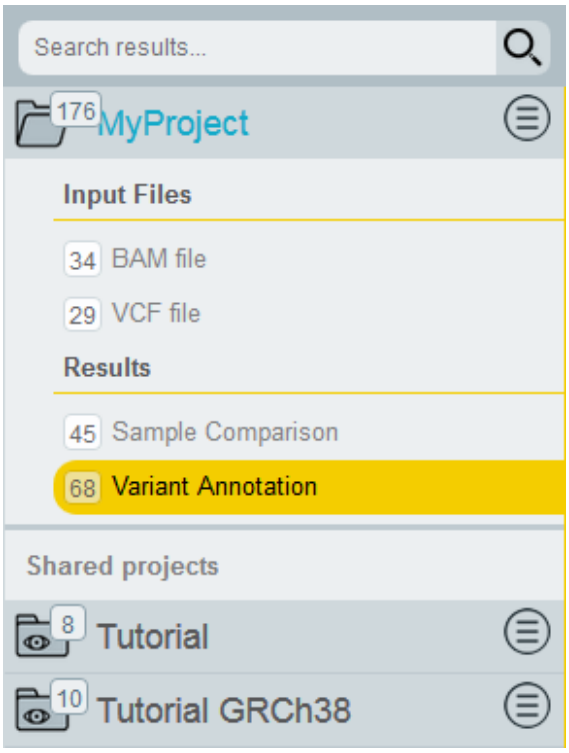


Manage results

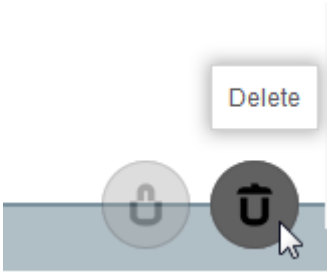
Imported samples and generated comparison analyses can be administered in the Result Management. It can be opened through the main menu:



In the result interface, there are two main sections in your own project space and in any other project space the administrator has shared for you. The *Input Files* section contains your uploaded VCF and BAM files. In the *Results* section, you find the variant annotations (automatically generated when you upload a VCF file) and your generated sample comparisons.



You can also directly open either a sample or a comparison by double clicking on the title. Click twice on the name or comment of an entry to rename the sample or comparison or edit the comment or description for the result.



The waste bin symbol in the lower right hand corner lets you delete any sample or comparison analysis from the server. If you don't like to keep those data you can easily delete them here.

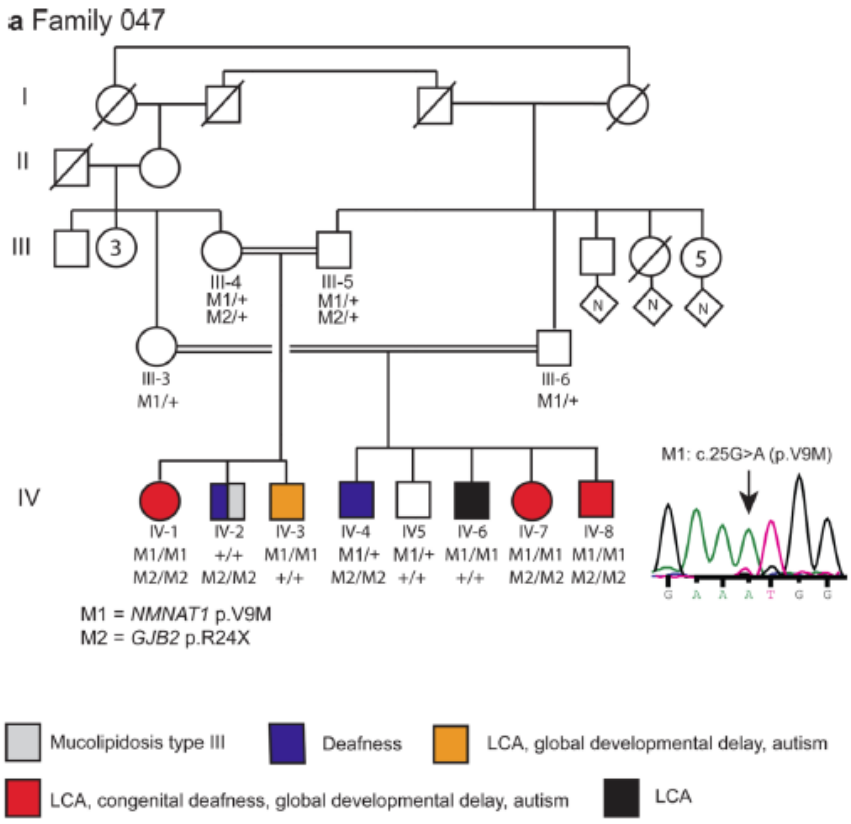
Additional analysis exercises

With the following exercises, you can apply what you just learned to a number of additional examples.

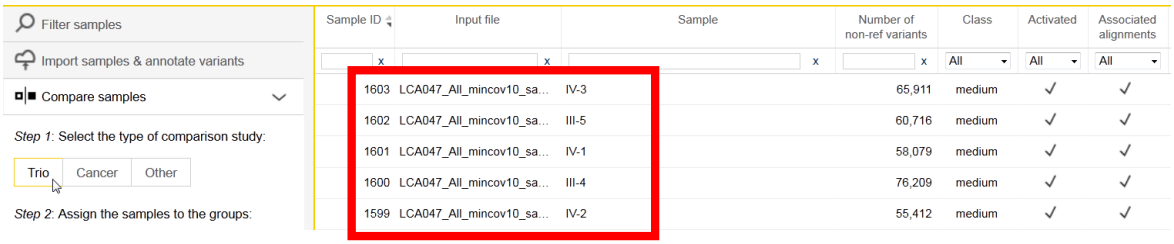
Family analysis

The first few of them will use data from the same family as above, including the siblings of the daughter affected by Leber Congenital Amaurosis.

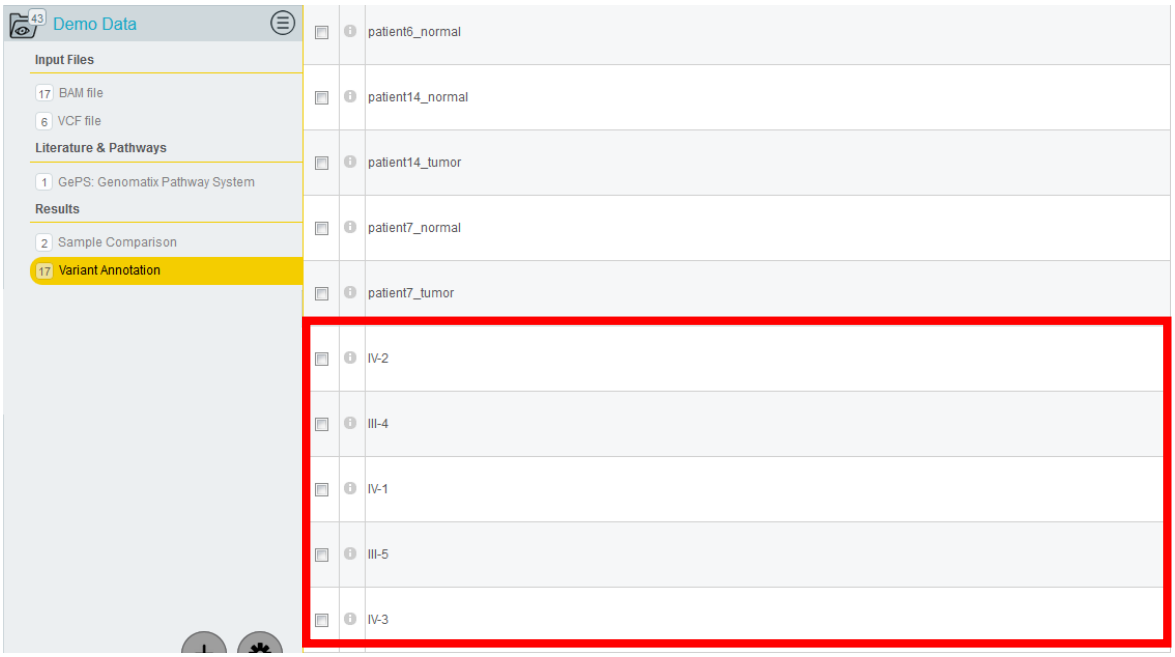
Here is once more the family tree from page 4 with information on the observed phenotypes in generation IV.



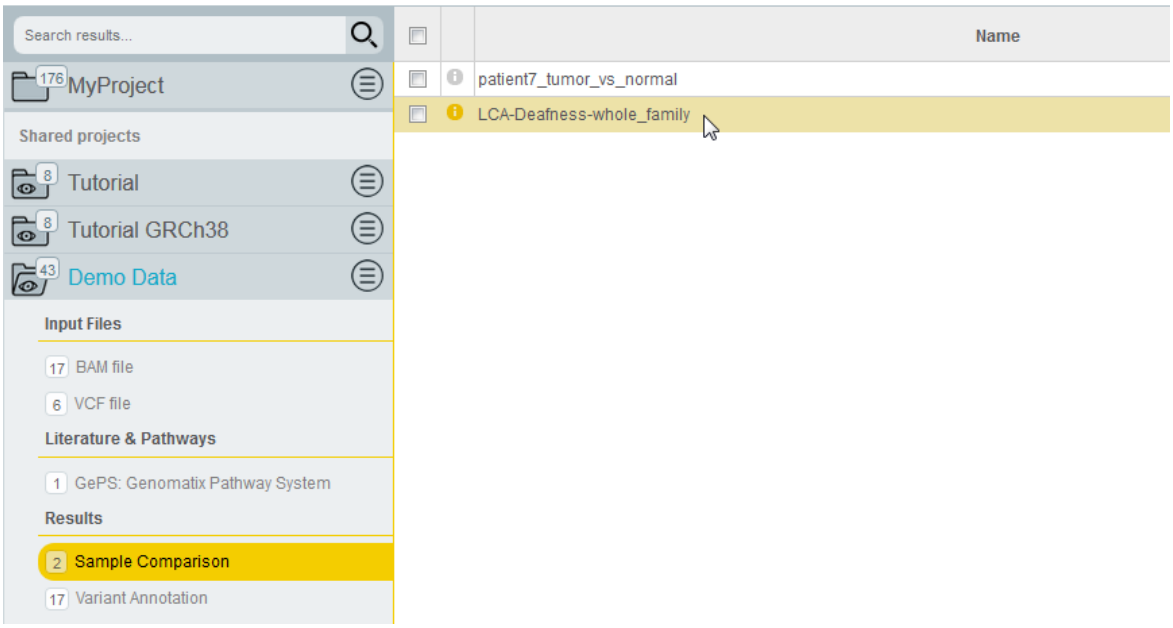
Please open the *Sample Comparison* view from the menu. You'll find the samples for the parents in generation III (III-4 and III-5), as well as the samples for their three children (IV-1, IV-2, and IV-3) in the sample list as shown below if you filter the list using the *Shared by* column for ‘demo’.



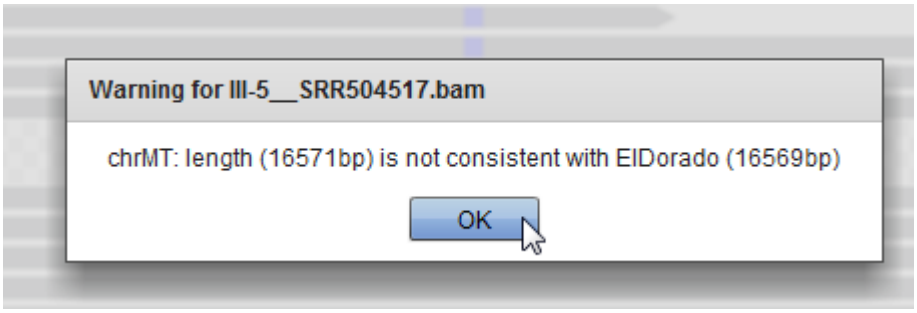
Please note that this set of samples is based on a different genome build (GRCh37) and was generated with a different version of the SAMtools SNP calling algorithm and with different parameter settings than the samples used in the first example. Therefore, the numbers of variants are different, and coverage data for specific variants may also differ. You can find these data also in your result management in the shared project *Demo Data*.



As you can't generate your own comparisons from read-only shared data, please use the Sample Comparison *LCA-Deafness-whole_family* from the *Demo Data* project. This is an *Other* type case-control comparison, with the children in the *Case* group, and the parents in the *Control* group



Note: if you see warning messages like this when you jump to the genome browser using the browse link in the *Sample Comparison* view, please press O.K. until the dialog is hidden. You can ignore the warning for the analysis, as the mitochondrial chromosome was not included in the variant calling.



Exercise 1: LCA in sibling IV-3

As the first exercise, please try to find the mutation which is probably responsible for the LCA phenotype in sibling IV-1 also in sibling IV-3. To achieve this, you will want to use corresponding similar filter settings as you used in the trio analysis. Please note that you can't use template filters generated in a trio analysis for case-control analyses (and vice versa).

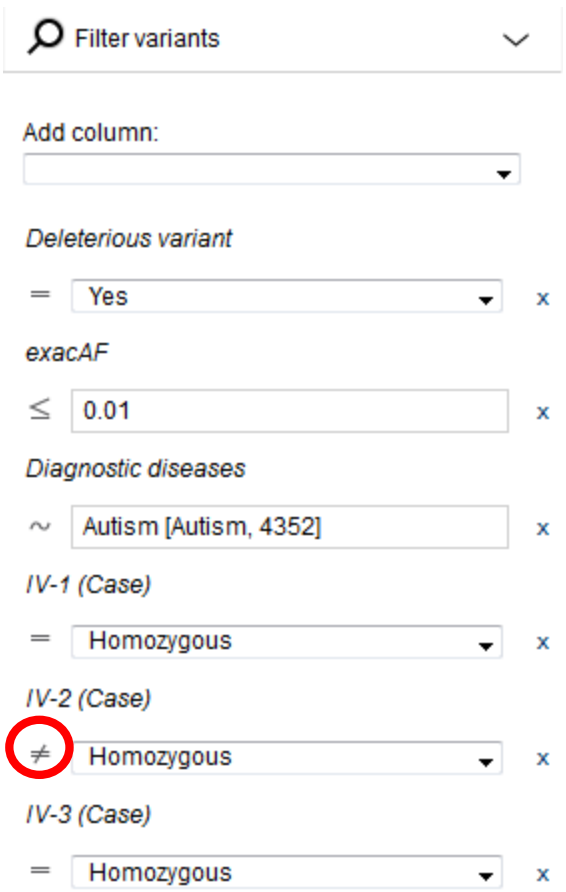
Exercise 2: deafness in sibling IV-2

Next, please try to find candidate mutations that could be causative for the observed deafness in sibling IV-2. Alternatively use annotation in the columns *Literature Diseases* and *Clinical Diseases* as one of the filtering criteria. Compare the results to each other, and to the corresponding results you get for sibling IV-1.

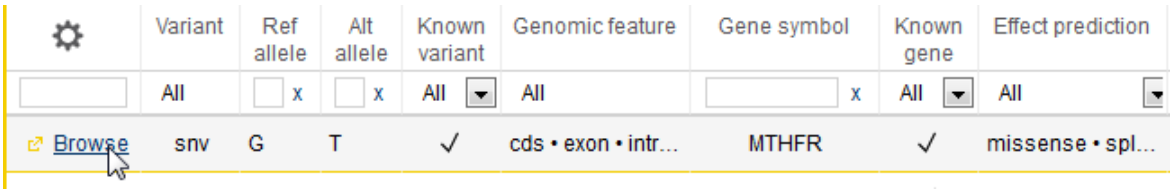
Exercise 3: autism in siblings IV-1 and IV-3

This exercise is somewhat less straightforward then the first two. The siblings IV-1 and IV-3 have been diagnosed with autism, while their brother, sibling IV-2, is not affected. A way to find candidate variants is to filter for rare, deleterious variants. You can set the genotypes for the samples of the affected siblings to homozygous. Try filter combinations including a general term like *Autism* or *Autism Spectrum Disorders* in *Diagnostic Diseases*, and look for variants that are homozygous in both affected persons, but not in the unaffected ones.

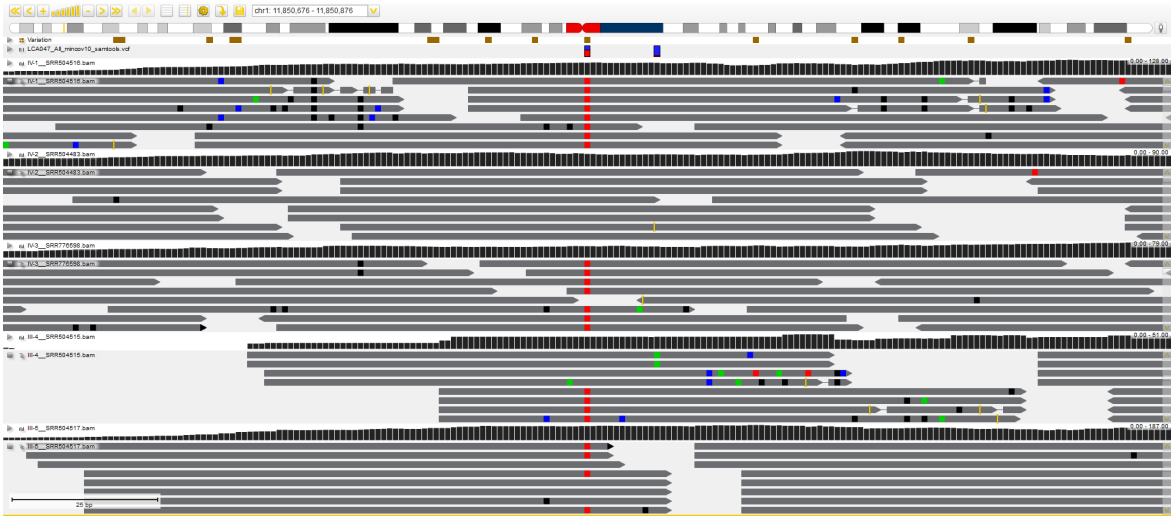
Note that the filter settings in the left hand panel allow you to change the operators for the filters, e.g. for the genotype to *is not equal to*.



The genotype of the parents may be inconclusive, so avoid setting the heterozygous filter for the parents in this case. You may want to check if the genome browser view can give you more information.



Here you can see directly if there are any reads supporting a variant genotype. You may need to scroll down in a read track to see the relevant reads, e.g. those supporting a variant genotype in sample III-4 (the mother) for the variant in the MTHFR locus above, which was not called by the variant caller.

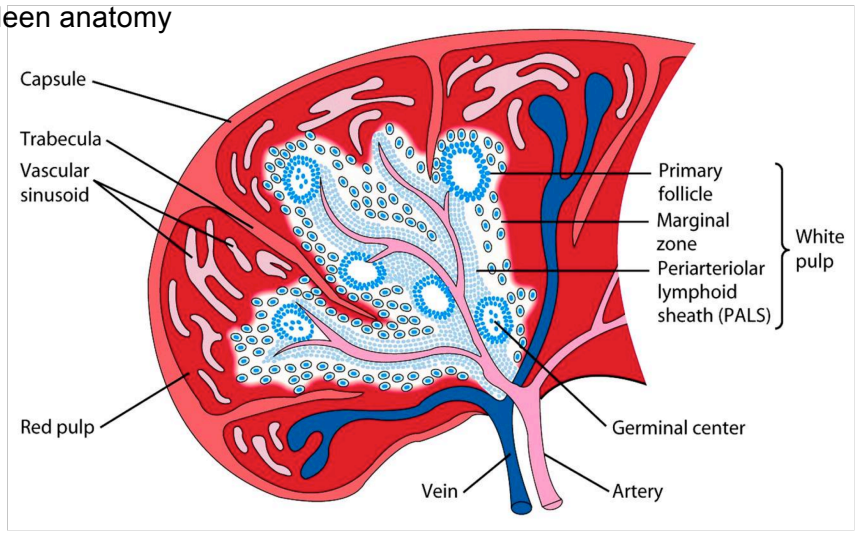


Cancer analysis

The next exercises use a set of splenic marginal zone lymphoma (SMZL) data from Martínez et al. (2014). The sequence data were downloaded from the Sequence Read Archive (SRA, study SRP033125) and mapped using the Genomatix Mining Station; variants were called with SAMtools.

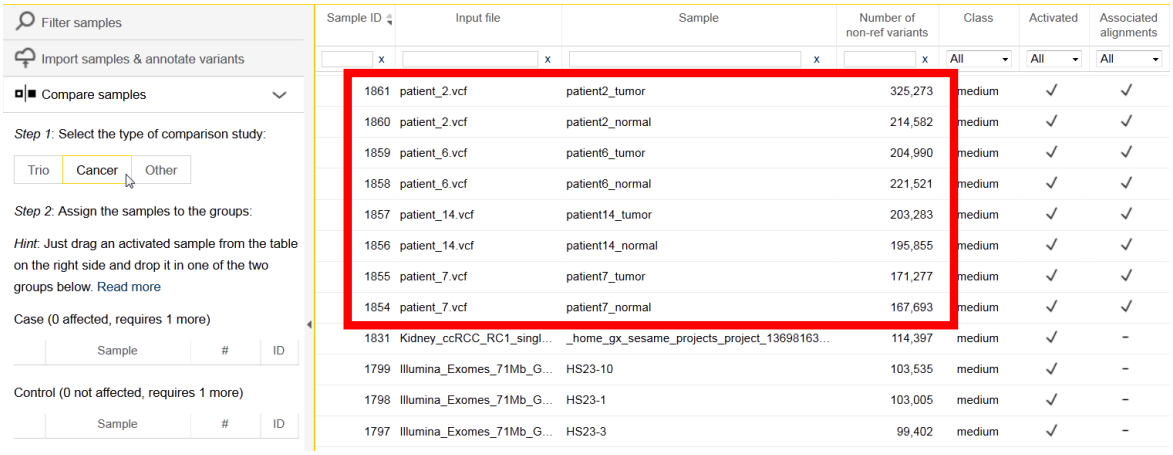
SMZL is a lymphoma made up of B-cells that replace the normal architecture of the white pulp of the spleen. The neoplastic cells are both small lymphocytes and larger, transformed blasts, and they invade the mantle zone of splenic follicles and erode the marginal zone, ultimately invading the red pulp of the spleen.

Spleen anatomy



Source: Northern Arizona University http://www2.nau.edu/~fpm/immunology/spleen1.jpg

The corresponding files are also available in the sample list:



There are annotated variant data from tumor and normal spleen tissue samples from four different patients (patient 2, 6, 7, and 14).

Exercise: somatic mutations in patient 7

From the menu, please open the *Result Management*. There, open the comparison *patient7_tumor_vs_normal* from the *Demo Data* project.

This is a *Cancer* type comparison with the patient 7 tumor sample in the *Case* group, and the patient 7 normal sample in the *Control* group. The resulting *Sample Comparison* will contain a filter column *Somatic*, indicating if the genotypes in the *Case* and *Control* samples are different.

Filter for rare (*exacAF Score*) *deleterious somatic* variants with good quality indicators. You can try combinations of *Quality* (a score indicating the reliability of the variant call), *Genotype Quality* (indicating the reliability of the genotype call), and *Coverage* (counting only high-quality base calls). *Quality* and *Genotype Quality* are Phred-like scores, i.e. 20 indicates 99% reliability, 30 indicates 99.9% reliability etc.. Add a suitable annotation term, for example, in the *Literature Diseases* column filter, the term *B-Cell Lymphoma*. You should be able to find variants in MYD88 and TP53. Find additional evidence for a variant being cancer-related by clicking on the row in the main table and reviewing the information in the detail table that will open for the entry (e.g. in the tabs *Transcript effects*, *Somatic mutations*, and *COSMIC*). View regions containing a variant in the genome browser to see overlaps with genomic annotation, e.g. PFAM protein domains.

Literature

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SM, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P: Ensembl 2015. Nucleic Acids Res. Jan;43(Database issue):D662-669 (2015).

Falk MJ, Zhang Q, Nakamaru-Ogiso E, Kannabiran C, Fonseca-Kelly Z, Chakarova C, Audo I, Mackay DS, Zeitz C, Borman AD, Staniszewska M, Shukla R, Palavalli L, Mohand-Said S, Waseem NH, Jalali S, Perin JC, Place E, Ostrovsky J, Xiao R, Bhattacharya SS, Consugar M, Webster AR, Sahel JA, Moore AT, Berson EL, Liu Q, Gai X, Pierce EA: NMNAT1 mutations cause Leber congenital amaurosis. Nat Genet 44(9), 1040-1045 (2012).

Gregg J, Fraizer G: Transcriptional Regulation of EGR1 by EGF and the ERK Signaling Pathway in Prostate Cancer Cells. Genes Cancer 2(9), 900-909 (2011).

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16), 2078-2079 (2009).

Martínez N, Almaraz C, Vaqué JP, Varela I, Derdak S, Beltran S, Mollejo M, Campos-Martin Y, Agueda L, Rinaldi A, Kwee I, Gut M, Blanc J, Oscier D, Strefford JC, Martinez-Lopez J, Salar A, Sole F, Rodriguez-Peralto JL, Diez-Tascón C, García JF, Fraga M, Sebastián E, Alvés J, Menárguez J, González-Carreró J, Casado LF, Bayes M, Bertoni F, Gut I, Piris MA. Whole-exome sequencing in splenic marginal zone lymphoma reveals mutations in genes involved in marginal zone differentiation. Leukemia. 28(6), 1334-1340 (2014).

Vinuesa CG, Tangye SG, Moser B, Mackay CR: Follicular B helper T cells in antibody responses and autoimmunity. Nature Reviews Immunology 5, 853-865 (2005).

List of resources available on the web:

Gene Expression Omnibus:
http://www.ncbi.nlm.nih.gov/geo/

Further reading:
http://www.genomatix.de/expertise/publications.html

This tutorial was compiled for GeneGrid release January 2018.

Please note that depending on the program versions and database releases used slight variations in results (e.g. gene numbers) may occur.