**Gemini: Analysis of Mendelian Traits**

**Analysis Under a Variety of Genetic Models to Prioritize Variants**

**Copyrighted © 2017** Imen Chakchouk, Isabelle Schrauwen and Suzanne M. Leal

GEMINI (GEnome MINIng) is a flexible framework for exploring genetic variation that implements human genome annotations. Using the GEMINI framework begins by loading a VCF file (and an optional PED file) into its database. Variant annotation is then performed using either Variant Effect Predictor (VEP) <http://www.ensembl.org/info/docs/tools/vep/index.html> or SnpEff (http://snpeff.sourceforge.net/). GEMINI integrates genetic variation with a diverse and adaptable set of genome annotations (dbSNP, ENCODE, UCSC, ClinVar, KEGG, OMIM, ExAC, gnomAD, etc) into a unified database to facilitate interpretation and data exploration. All data/information is stored in portable SQLite database that allows the user to explore and interpret both coding and non-coding variation. By annotating and filtering using information on inheritance models a short list of variants can be obtained, which can be furthered followed-up, e.g. testing for segregation within the pedigree.

GEMINI provides distinct functionality that is not available with most existing software. One powerful function of Gemini is the built-in Mendelian inheritance tools that support the analysis of multi-generational pedigrees. A pedfile is loaded into the Gemini database, which allow the researcher to prioritize variants based on inheritance and sharing patterns.

Gemini provides a variety of tools and functions. In this exercise, we will introduce the basics and for more information please refer to the website and the article.

<http://gemini.readthedocs.io/en/latest/index.html>

<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003153>

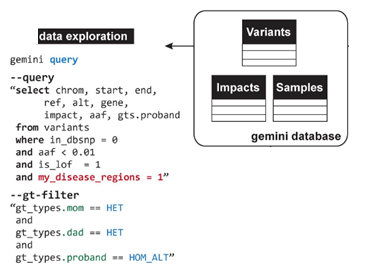
**What can you do with Gemini?**

1- Load a VCF into an “easy to use” database (and an optional PED file): GEMINI imports genetic variants and all samples genotypes from a VCF file into a db database using the load sub-command. If multiple pedigree members have been sequenced then a pedfile can be provided to facilitate analysis using information on genetic modes of inheritance modes, phenotypes, sex and familial relationships.

2- Annotate each variant in the VCF file using different genome annotation sources to facilitate variant exploration. Gemini organizes the annotated data in tables (a database) that uses SQL (Structured Query Language) to “ask” different questions. All variant annotations are stored in the *variants*and *variant\_impacts* tables. The samples information and the loaded PEDfile are kept in *Samples\_table*. Please, visit this link to explore the content of each table in the database that you create <http://gemini.readthedocs.io/en/latest/content/database_schema.html>.

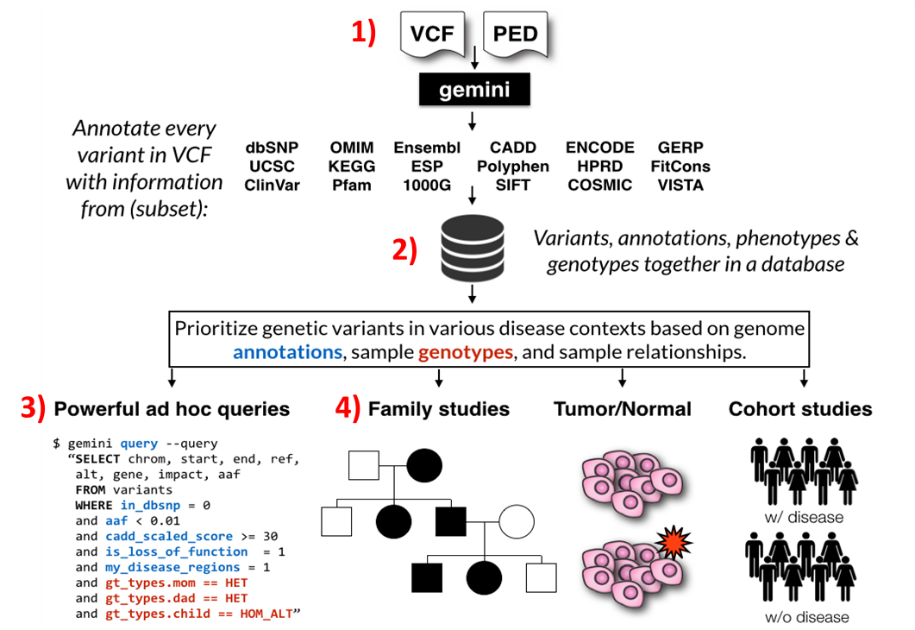
3- Query (fetch data) from database based on annotations or subject(s) genotypes:

Once all data is integrated in a single standardized database, we can use different commands to explore it. A powerful Gemini command is *query*, which allows the researcher to compose queries to the GEMINI database. An example is illustrated below:

****

**Figure 1:** Example of Variant mining using the GEMINI database framework using the query command. Settings of 0=no and 1=yes. In this example, we selected the variants that are not in dbSNP, which have an aaf frequency < 0.01 (minor allele frequency in the gnomAD database) and are loss of function (lof) variants. It should be noted that not all variants in gnomAD are currently in dbSNP. The variant must be homozygous in the probands and heterozygous in both parents. Gemini is also using a disease region, which is set by the user (my\_disease\_region). This would be done when the user has a specific region of interest, e.g. linkage, homozygosity region.

4- Analyze genetic inheritance models: In addition to the *query* command, Gemini provides other tools that address more complex questions without requiring the researcher to write any additional analysis code. Since the PED file is loaded to the sample\_table, all families based studies are in the included tools: de novo, autosomal\_dominant, autosomal\_recessive, comp\_hets and X\_linked.



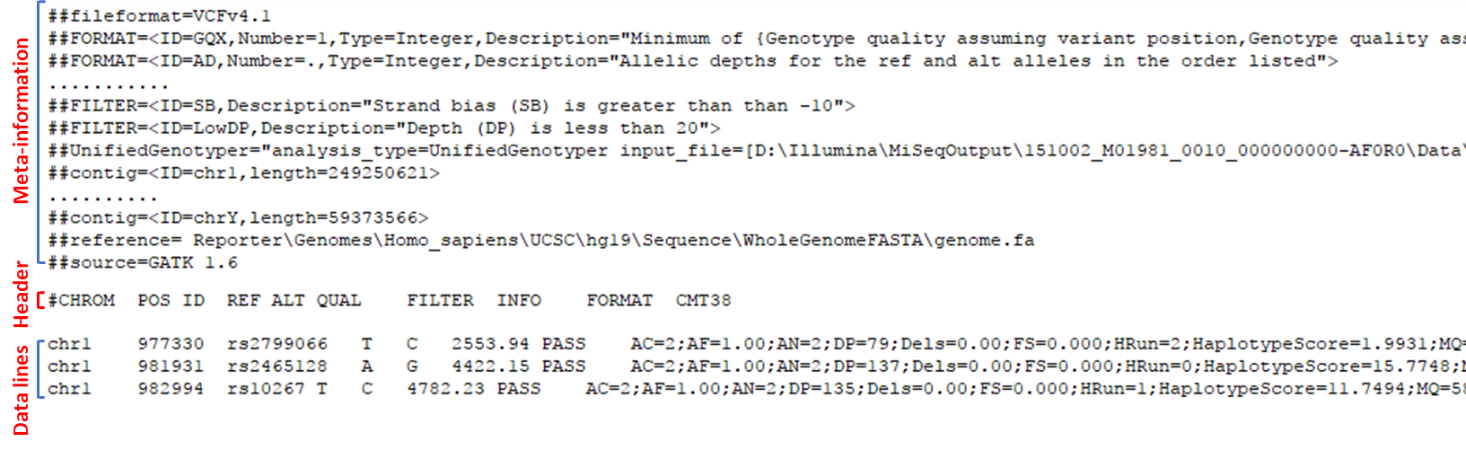
**Figure 2:** Schematic drawing displaying the capabilities of Gemini.

1. **Load VCF and PED file**

# Open terminal in the directory /gemini

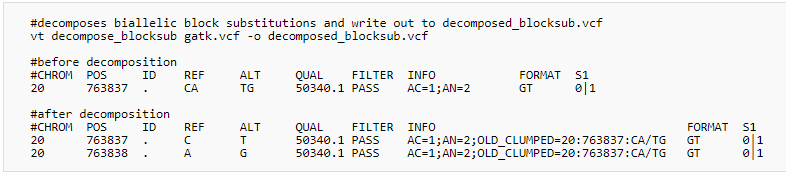
Note that for this first part of the exercise, you are not asked to run the command lines due to the length of time that it takes to decompose and annotate the vcf files.

A Variant Call Format (VCF) is a text file. It contains three main parts: meta-information lines, a header line, and data lines each containing information about a position in the genome.



# In some cases, a VCF file assigns a block of multiallelic successive variants at one genomic position. For easier analysis, each variation should be annotated as one unit instead of a block. In the following example, consecutive SNPs are biallelic block substitutions. In this case, the annotation software will attribute a population frequency to the block instead of the individual variants, which may lead to missing causal variants during analysis. To avoid these problems, we first “decompose” the variants in the vcf file using the command vt decompose before performing annotation. A new vcf file is outputted which contains the decomposed variants. For more examples visit the link <http://genome.sph.umich.edu/wiki/Vt>

$ vt decompose input.vcf -o output.decomposed.vcf



# Gemini currently support annotations produced by either [SnpEff](http://snpeff.sourceforge.net/) or [VEP](http://www.ensembl.org/info/docs/variation/vep/index.html). We recommend using VEP. When using the *gemini load* command we upload the decomposed VCF file and also annotate the data using VEP and create a database (my.db). The –v tells Gemini that a vcf file will be uploaded and –t flag is used before specifying the annotation which will be used, i.e. SnpEff or VEP:

$ gemini load -v output.decomposed.vcf -t VEP my.db

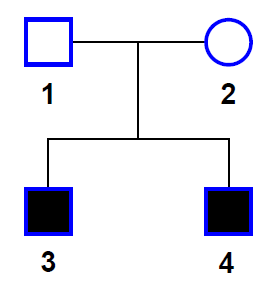
# The Database tables’ details are illustrated in the link below. In summary, there are 5 principal tables: *Variants*, *Variant\_impacts*, *Samples*, *Gene\_detailed*, and *Gene\_summary*. Please see <http://gemini.readthedocs.io/en/latest/content/database_schema.html> for additional information.

# In order to perform inheritance based commands, it is necessary to update the *samples\_table* in the database with the pedigree information by providing the PED file using the *gemini amend* command.

$ gemini amend --sample pedfile.ped my.db

\* PEDFile is tab-delimited, without a header, with the following numeric columns

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Family ID** | **Individual ID** | **Paternal ID** | **Maternal ID** | **Sex** | **Phenotype** |
|  |  |  |  | 1 = male  2 = female  -9= missing | 1 =unaffected  2 = affected  -9 = missing |

Here is an example of a pedigree file for a nuclear family with two affected male children and unaffected parents

1 1 0 0 1 1

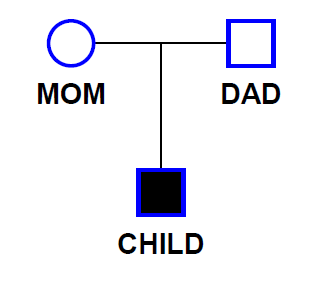
1 2 0 0 2 1

1 3 1 2 1 2

1 4 1 2 1 2

1. **Query the database you created**

For all exercises below, we provide you with databases for which the variants have already been decomposed and annotated with VEP (due to time limitations). When working with your own files you can decompose the variants in your vcf file and make a database using the commands (which are given above) and the appropriate file names.

Let us start by analyzing fam1.db, a database of three samples from a trio family with a child affected with nonsyndromic hearing impairment (NSHI). We will analyze this trio using different modes of inheritance.

**Family 1:** A trio pedigree with unaffected parents and a child with NSHI

**Question 1:** What are the possible modes of inheritance for family 1? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Please make the PED file for family 1 or use the PED file which you have been provided in the exercise folder called family\_1.ped. A database has already been created for family 1, fam1.db. You will first amend the family\_1.ped to the database.

$ gemini amend --sample family\_1.ped fam1.db

# To verify that the PEDfile is correctly loaded in *Samples table*: Query names from the samples table:

$ gemini query -q "SELECT name FROM samples" fam1.db

Check the phenotype of each family member:

$ gemini query -q "SELECT name, phenotype FROM samples" fam1.db

Next, we will check the number of variants available for analysis. It should be noted that for this example we are only analyzing variants from chromosome 2. Note that in Linux code “|” signifies to “pipe” or use the output from the previous command line as an input for the next command. Likewise (wc -l) is a Linux command line, it is used to count lines.

$ gemini query -q "SELECT chrom, start, end, ref, alt, gene FROM variants" fam1.db | wc -l

You can also save the columns that you want consult to a text output file. Open it in an excel table and select tab as the only separator option then examine the variants in the file (tip: if you need to see the files in your local machine type scp username@ngs1.statgen.us:/path/to/file /local/path/)

$ gemini query -q "SELECT chrom, start, end, ref, alt, gene, aa\_change, transcript, rs\_ids, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, clinvar\_gene\_phenotype FROM variants" fam1.db > fam1output.txt

**Question 2:** How many variants are available for analysis\_\_\_\_\_\_\_?

We can use *stats*, a gemini command argument, to compute useful variant statistics. For more details visit <https://gemini.readthedocs.io/en/latest/content/tools.html#stats-compute-useful-variant-statistics>

The following command will allow us to compute how many variants are present per sample.

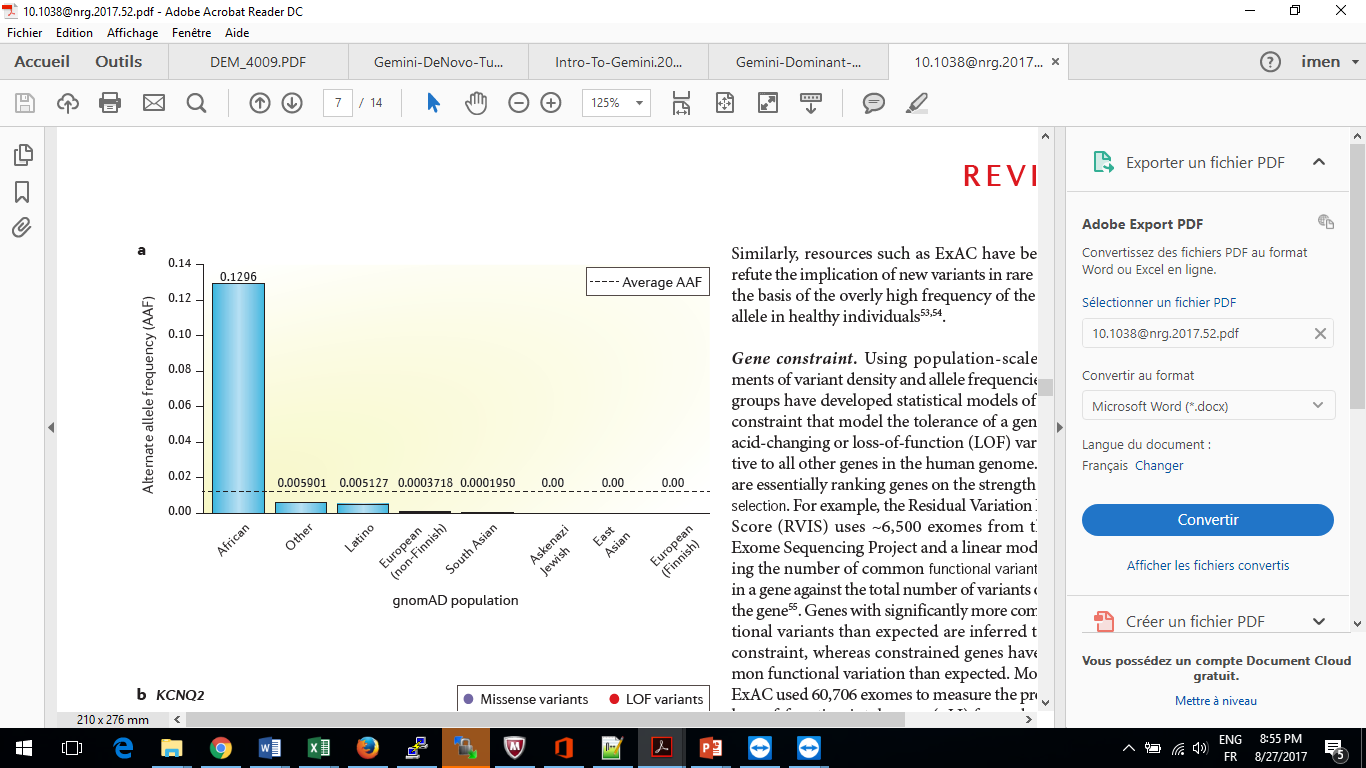
$ gemini stats --vars-by-sample fam1.db

**Questions 3:** How many variants does the mom\_\_\_\_\_, dad\_\_\_\_\_, and child \_\_\_\_\_\_ have available for analysis?

# Let’s start to prioritize variants based on their frequency in control databases (gnomAD in this case): Query the rare variants (max\_aaf\_all <= 0.006):

$ gemini query --header -q "SELECT chrom, start, end, gene FROM variants WHERE max\_aaf\_all <= 0.006" fam1.db | wc -l

Variant allele frequencies can be very different between populations. It is highly unlikely that a pathogenic variant will have a high minor allele frequency (MAF). To select rare variants, we do not want to use the average frequency for all populations in genomAD since a variant, which might not even be present in one population, can have a high frequency in another population (see figure 3). Therefore, we selected variants with an allele frequency less than the specified cut-off in all in gnomAD populations using the command max\_aaf\_all, i.e. the maximum allele frequency observed in any of these populations. Please note when this command is used Gemini automatically uses allele frequencies from the gnomAD database. The variant frequency used is motivated by the disease prevalence and the mode of inheritance. If analysis is being performed under an autosomal dominant mode of inheritance a lower allele frequency is usually used than for an autosomal recessive mode of inheritance since it is unlikely that affected individuals will be included in a database but there can be carriers of autosomal recessive variants. If an autosomal dominant disease has reduced penetrance a higher MAF cut off might be desirable.

****

**Figure 3**: A variant may present an average frequency which is low making it a potential causal variant. But the frequency may be higher in specific subpopulations thus causing doubt on its relevance to rare disease phenotypes. In this example shown, the rs79444516 variant in *USH2A* is in low in European populations but considerably higher in Africans. This variant with a frequency of 12% must therefore be benign. Yet if we take the average frequency across populations it would have a MAF <0.006 and be selected for further analysis.

From Eilbeck Ket al. 2017

**Questions 4:** How many rare variants are available for analysis knowing that the header is also counted as a line with (wc -l) \_\_\_\_\_\_\_\_\_\_\_?

# In some cases, the pathogenic variant in a family may have been previously identified or may be in a known disease gene although the variant itself is novel. To ascertain this information, we can use details from Clinvar which is provided in the VEP annotation. Gemini database provides different Disease phenotype information, detailed below:

* clinvar\_causal\_allele: The allele(s) that are associated or causal for the disease.
* clinvar\_sig: The clinical significance scores for each of the variant according to ClinVar: unknown*,*untested*,*non-pathogenic, probable-non-pathogenic*,*probable-pathogenic, pathogenic*,*drug-response*,*histocompatibility, other
* clinvar\_disease\_name: The name of the disease to which the variant is relevant.
* clinvar\_gene\_phenotype: A list of phenotypes associated with this gene (includes any variant in the same gene in Clinvar not just the studied variant).

Using the below command, we will be provided with Clinvar’s full phenotype information on variants and genes (i.e. the genes do not have to be specifically involved in hearing impairment). Please run the following command to select those variants in disease causing genes with a MAF<0.006 in all gnomAD populations.

$ gemini query --header -q "SELECT chrom, start, end, ref, alt, gene FROM variants WHERE max\_aaf\_all <= 0.006 and clinvar\_gene\_phenotype is not NULL" fam1.db | wc -l

**Questions 5****:** How many rare variants that are in disease genes reported in Clinvar are available for analysis \_\_\_\_\_\_\_\_\_?

# Using the below command, we will verify whether a known disease-associated allele is reported in our data. This will list the known alleles present in Clinvar, their clinical significance and the name of the disease to which the variant is relevant. In our table, we will only display the rare variants with at least 7 reads for all associated or causal genes in our data. Including a read depth criteria reduces the chance that a variant is a false positive. Although if you do not find any candidate variants you may wish to relax the criteria and investigate the reads in the BAM file using IGV.

$ gemini query --header -q "SELECT chrom, start, end, ref, alt, gene, clinvar\_causal\_allele, clinvar\_sig, clinvar\_disease\_name, gt\_types FROM variants WHERE max\_aaf\_all <= 0.006 and clinvar\_causal\_allele is not NULL and depth > 6" fam1.db > known\_variants.txt

Note that *gt\_types* is a compressed binary vector of numeric genotype “types” (0: Homozygous Reference, 1: Heterozygous, 2: Homozygous Altered)

**Questions 6:** How many rare variants in our samples are known and listed in Clinvar \_\_\_\_\_\_\_\_? Are there any known NSHI variants\_\_\_\_\_\_\_? Focus on the clinical significance scores and verify if a known variant listed in Clinvar which is considered to be pathogenic is reported in our samples\_\_\_\_\_\_\_\_\_\_\_\_? Describe the variant genotype in mom\_\_\_\_\_\_\_\_\_, dad\_\_\_\_\_\_\_\_\_\_, and child \_\_\_\_\_\_\_\_\_\_? Do we need to go further in the analysis\_\_\_\_\_\_\_?

# Next, we want to determine which variants are novel, i.e. they are neither in dbSNP nor gnomAD. Please note that dbSNP has not yet incorporated all of the variants from gnomAD.

$ gemini query -q "SELECT chrom, ref, alt, gene, rs\_ids FROM variants WHERE max\_aaf\_all == -1 and in\_dbsnp == 0" --header fam1.db | wc -l

**Question 7:** How many novel variants are observed in these samples\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

1. **Analyzing simple genetic models**

To use Mendelian inheritance models command lines in the analysis it is necessary to load a PED file to the database. For each mode of inheritance, we are looking for specific genotype patterns. For example, for autosomal dominant the individuals must be heterozygous. Below are the genotype requirements for each genetic model.

1. **Using the autosomal\_recessive command**

# Gemini provides a command that allows the analysis of autosomal recessive (AR) inheritance without specifying each family member’s genotype. The command line *autosomal\_recessive* is available when a proband and at least one parent has sequence data available. The genotypes requirements are:

* All affected must be hom\_alt
* No unaffected can be hom\_alt
* In the absence of established parent/child relationships in the PED file, GEMINI will issue a WARNING

In the below command, we are selecting rare variants (MAF<= 0.006) for which each member of the trio has a genotype with at least 7 reads (depth>6) and the affected proband is homozygous and both parents are heterozygous (autosomal\_recessive).

$ gemini autosomal\_recessive --columns "chrom, start, end, ref, alt, gene, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, clinvar\_gene\_phenotype " --filter "max\_aaf\_all <= 0.006 and depth > 6 " fam1.db > autos\_recessive\_fam1.txt

Below are examples of some of the bioinformatics and conservation annotations Gemini can provide:

* polyphen\_score: Polyphen scores for the severely affected transcript
* sift\_score: SIFT scores for the predictions
* Impact: The consequence of the most severely affected transcript
* Impact\_severity: Severity of the highest order observed for the variant
* Gerp\_bp\_score: GERP conservation score
* Cadd\_scaled: Combined Annotation Dependent Depletion

Open autos\_recessive\_fam1.txt in an excel table and select tab as the only separator option and answer the following questions:

**Question 8:** How many variants that segregate under an autosomal recessive mode, are observed in these samples\_\_\_\_\_\_\_\_\_\_\_\_\_\_? How many variants effect exons \_\_\_\_\_\_\_\_\_\_? How many variants have a high impact\_severity \_\_\_\_\_\_\_\_\_? Do you suspect any variant to be causal\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

We can also examine the compound heterozygote variants using the following command:

$ gemini comp\_hets --columns " chrom, start, end, ref, alt, gene, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, clinvar\_gene\_phenotype" --filter "max\_aaf\_all <= 0.006 and depth > 6 " fam1.db | wc -l

**Question 9:** How many compound heterozygote variants segregate in these samples \_\_\_\_\_\_\_\_\_\_\_\_\_\_?

1. **Using the autosomal\_dominant command**

# Gemini provides a command line that allows the analysis of the autosomal dominant inheritance mode without specifying each family member’s genotype. *autosomal\_dominant* command line is available to filter for and prioritize potential autosomal dominant mutations.

The genotypes requirements are:

* All affected must be het
* At least 1 affected must have 1 affected parent.
* This tool requires that you identify familial relationships via a PED file when loading your VCF into GEMINI

In this example, the inheritance could be autosomal dominant with reduced penetrance. Therefore, one of the parents carries the variant but does not show the phenotype.

In the below command, we are selecting rare variants (MAF< 0.003; this cut-off might differ depending on the prevalence of your dominant disease of interest) for which each member of the trio has a genotype with at least 7 reads (depth>6) and the affected proband is heterozygous and one of the parents is heterozygous (autosomal\_dominant).

$ gemini autosomal\_dominant --columns "chrom, start, end, ref, alt, gene, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, clinvar\_gene\_phenotype, clinvar\_disease\_name " --filter "max\_aaf\_all <= 0.003 and depth > 6 " fam1.db | wc -l

**Question 10:** How many variants that segregate under an autosomal dominant mode, are observed in these samples\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

1. **Using the de\_novo command**

# De novo mutations may explain genetic disorders in which an affected child has a mutation in every cell in the body but the parents do not, and there is no family history of the disorder.

Gemini provides a *de\_novo* command that allows the analysis of the *de novo* inheritance mode. A parent-child relationship is required in the loaded PED file.

$ gemini de\_novo --columns "chrom, start, end, ref, alt, gene, aa\_change, max\_aaf\_all, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, clinvar\_gene\_phenotype, clinvar\_disease\_name " --filter "max\_aaf\_all <= 0.003 and depth > 6 and impact\_severity != 'LOW'" fam1.db > denovo\_fam\_results.txt

**Question 11:** How many variants that segregate under a *de novo* mode, are observed in this trio\_\_\_\_\_\_\_\_\_\_\_\_\_\_? Is this variation affecting a known NSHI gene \_\_\_\_\_\_\_\_\_\_?

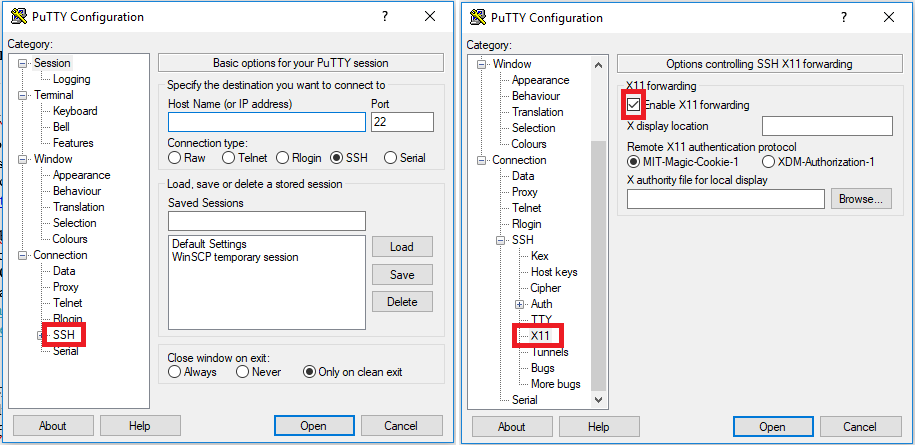
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Position | Ref allele | Alt allele | Impact | aa\_change | clinvar\_gene\_phenotype | max\_aaf\_all |
|  |  |  |  |  |  |  |  |

**Question 12:** What is the genotype of this variant observed in MOM\_\_\_\_\_\_\_\_\_\_, DAD\_\_\_\_\_, CHILD\_\_\_\_\_\_\_\_\_\_?

# We may want to verify variants that we have observed by examining the BAM file. This can be done by using visualizing the variant using IGV. This is particularly important for *de novo* variants since most “*de novo*” variants which are observed are errors.

We can examine a BAM file using IGV. This can be done on your desktop computer or by using Xming and using IGV directly on the cluster.

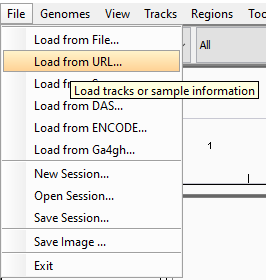
First you have to make sure to check “Enable X11 forwarding” in the X11 section of putty. Please follow the screen shots below:



**Figure** **4:** Enable X11 option when opening PuTTY

Then on the terminal use the command line: **igv .sh**

To open BAMs select File > Load from URL > Fam1BAM files > CHILDchr2.bam, DADchr2.bam, MOMchr2.bam. Then type or copy and paste the chromosome position of the selected variant. Verify the family members genotype of the potential *de novo* variant.



**Figure 5:** IGV screen shot showing how to open the BAM files on the cluster by selecting Load from URL

**Question 13:** Is the variant *de novo* or likely a false variant call\_\_\_\_\_\_\_?

1. **Using the x\_linked command**

# X-linked mutations may explain genetic disorders as well. Genes on the X chromosome can inherit following an autosomal recessive or dominant mode of inheritance or can be *de novo.* X-linked recessive disorders are expressed in females only when there are two copies of the mutation present (one on each X chromosome). However, for males, there needs to be only one copy of an X-linked recessive gene for the trait or disorder to be expressed. Note that in rare cases of X-linked recessive inheritance a female might still express a phenotype as heterozygote due to skewed X-inactivation. Alternatively, a female might be spared from a dominant X-linked disease due to skewed X-inactivation as well, or have a variable phenotype depending on the amount of skewing. The X-linked *de novo* mutations may explain genetic disorders when there is no family history of the disorder. In this case the affected child has a mutation on the X chromosome but none of the parents have it.

**Question 14:** In this example, is the X linked dominant inheritance mode possible \_\_\_\_\_\_\_\_? In which case\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

$ gemini x\_linked\_dominant --columns "chrom, start, end, ref, alt, gene, aa\_change, max\_aaf\_all, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, clinvar\_gene\_phenotype, clinvar\_disease\_name " --filter "max\_aaf\_all <= 0.003 and depth > 6 and impact\_severity != 'LOW'" fam1.db | wc -l

**Question 15:** How many variants that segregate under an x\_linked\_dominant mode, are observed in this trio\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

$ gemini x\_linked\_recessive --columns "chrom, start, end, ref, alt, gene, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, max\_aaf\_all, clinvar\_gene\_phenotype, clinvar\_disease\_name " --filter "max\_aaf\_all <= 0.003 and depth > 6 and impact\_severity != 'LOW'" fam1.db > x\_linked\_rec\_fam\_results.txt

**Question 16:** How many variants that segregate under an x-linked\_recessive mode, are observed in this trio\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Position | Ref allele | Alt allele | Impact | Conservation | Cadd\_scaled | max\_aaf\_all |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

**Question 17:** Based on the bioinformatic predictions, which variant is more likely to be causal \_\_\_\_\_\_\_\_\_\_\_\_\_\_?

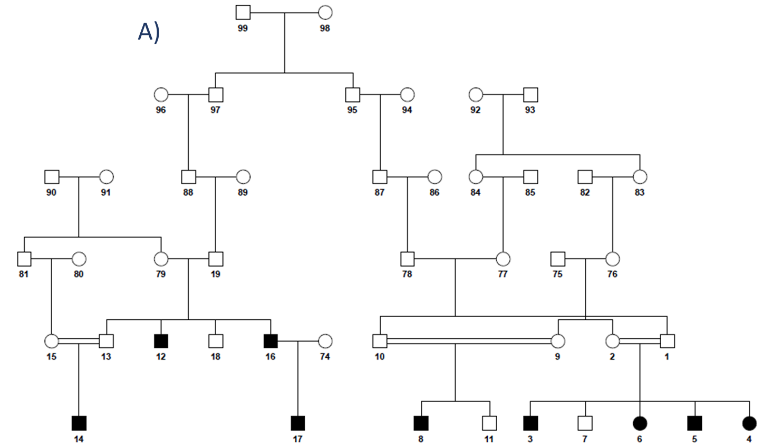
Examine the BAM file using IGV (like the previous example): open BAMs select File > Load from URL > Fam1BAM files > CHILDchrX.bam, DADchrX.bam, MOMchrX.bam.

**Question 18:** Is the variant *de novo* or likely a false variant call\_\_\_\_\_\_\_?

1. **Custom genotype filtering by gene:**

In large families with multiple affected individuals, exome sequencing parents and a proband is not the best strategy. Usually, distant probands are selected for sequencing. In this case, we cannot use *autosomal\_recessive* command line as was performed above.

***Family A:***



For autosomal recessive (AR) families particularly if consanguineous, we can usually identify causal variants that are homozygous using a single exome. This is also true for AR families which are outbred: a single exome can be used to identify homozygous variants or potentially compound heterozygous variants, although the variants will have to be checked using DNA samples from the parents to ensure that they do not lie on the same haplotype. If there are multiple branches within an AR pedigree and funds are available, an affected from each branch can be sequenced if it is suspected that there is locus heterogeneity. On the other hand, for dominant families (ADNSHI) we would usually generate sequence data using DNA samples from at least two-family members selecting the most distantly related, affected family members to generate sequence data. This strategy can minimize the number of rare damaging variants that are identified as potentially causal which need to be followed-up by testing segregation in all family members using for example Sanger sequencing.

In previous examples, we saw how to query based on variants and on inheritance mode. In this case, we suspect that there is one homozygous coding variant. We will query by gene (*gene\_wise*) with using the custom genotypes filter (*--gt-filter*). With this tool, multiple --gt-filters can be specified for example besides specifying that a variant must be homozygous we could also search for heterozygous variants.

For family A two affected members were sequenced. This is a family with multiple branches and locus heterogeneity is possible within this family. However, we will first analyze the family assuming that there is locus homogeneity and determine which rare homozygous variants are shared by the two affected family members (individuals L1 and L2) using the --gt-filter filter. This filter allows us to select the genotype of each sequenced member in the database. We have provided the database ARn2.vep.db for analysis.

$ gemini gene\_wise ARn1.vep.db --gt-filter "(gt\_types.L1 == HOM\_ALT) and (gt\_types.L2 == HOM\_ALT)" --filter "max\_aaf\_all <0.006" --columns "chrom, start, end, ref, alt, gene, polyphen\_score, sift\_score, impact, impact\_severity, gerp\_bp\_score, cadd\_scaled, clinvar\_gene\_phenotype, clinvar\_disease\_name" > AR\_results1.txt

The result table comes with a standard set of columns:

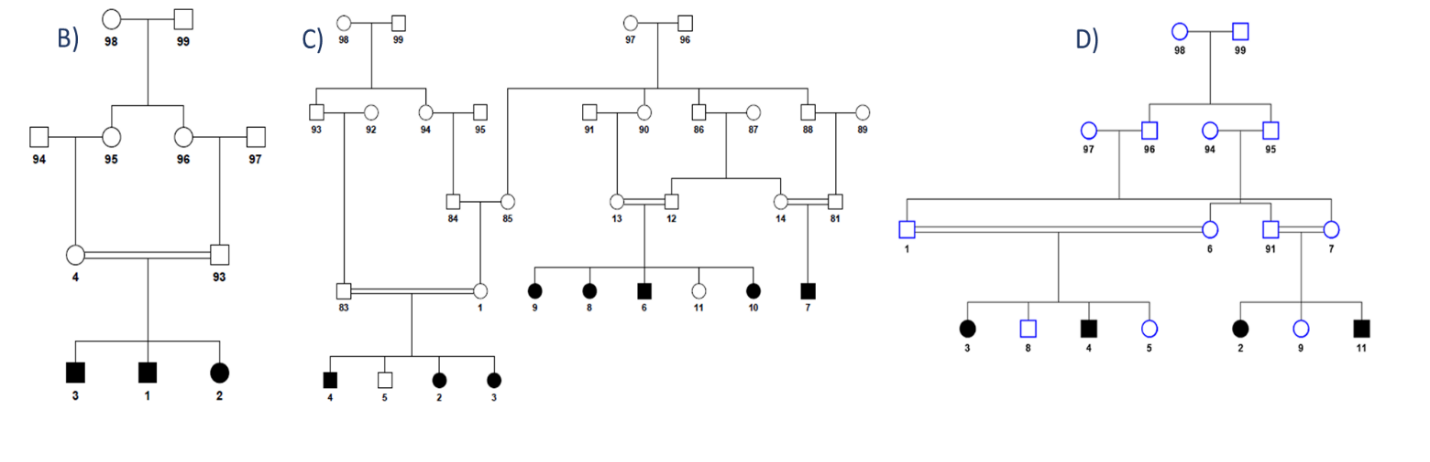
* The *variant\_filters* column shows which filters were passed by the variant. (in this case we have --gt-filter equal one)
* The *n\_gene\_variants* column shows how many variants in the gene are being reported.
* The *gene\_filter* column shows which filters in the gene passed by any variant.

Additional columns may be selected:

* *clinvar\_gene\_phenotype:* list of phenotypes associated with this gene (includes any variant in the same gene in clinvar not just the current variant).
* *clinvar\_disease\_name:* The name of the disease to which the variant is relevant

**Question 19:** How many rare homozygote variants are shared by both samples \_\_\_\_\_\_\_\_\_\_\_\_\_\_? Which gene(s) contain these homozygous variants \_\_\_\_\_\_\_\_\_\_\_? Is this a known NSHI gene \_\_\_\_\_\_\_\_\_\_\_?

***Families B, C and D***



The hearing impairment in three families B, C and D map to the same chr16 region (table 2 below), however they each segregate unique SNP haplotypes within the region, suggesting that they may have different variants in the same gene.

Table 2: Homozygous regions identified in the three-autosomal recessive NSHI families

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Family* | *Exome* | *DB database* | *LOD score* | *Mapped Region* |
| B | AR1 | AR | 1.9 | chr16:69.7-82.0 |
| C | AR2 | 6.0 | chr16:63.6-79.7 |
| D | AR3 | 3.7 | chr16:55.1-80.5 |

# We can query by gene (*gene\_wise*) using the custom genotypes filter (*--gt-filter*). We can filter homozygous altered variants affecting probably the same gene in the sequenced families members. The variant that we are looking for is not necessarily common between 3 samples. So, we query by gene, from ARn1.vep.db the HOM\_ALT variations affecting AR1 sample, the HOM\_ALT variants affecting AR2 and the HOM\_ALT variants affecting AR3

$ gemini gene\_wise ARn2.vep.db --gt-filter "(gt\_types.AR1 == HOM\_ALT)" --gt-filter "(gt\_types.AR2 == HOM\_ALT)" --gt-filter "(gt\_types.AR3 == HOM\_ALT)" --filter "max\_aaf\_all <0.006" --columns " vcf\_id, chrom, start, end, ref, alt, gene, impact, impact\_severity, max\_aaf\_all, gts, clinvar\_gene\_phenotype " > AR\_results2.txt

Open AR\_results.txt in an XL table and fill the following table:

Note that *gts* column shows the genotypes of the samples. The genotypes are organized in the same order of the samples listed in --gt-filter. Sample1: AR1, Sample2: AR2, Sample3:AR3.

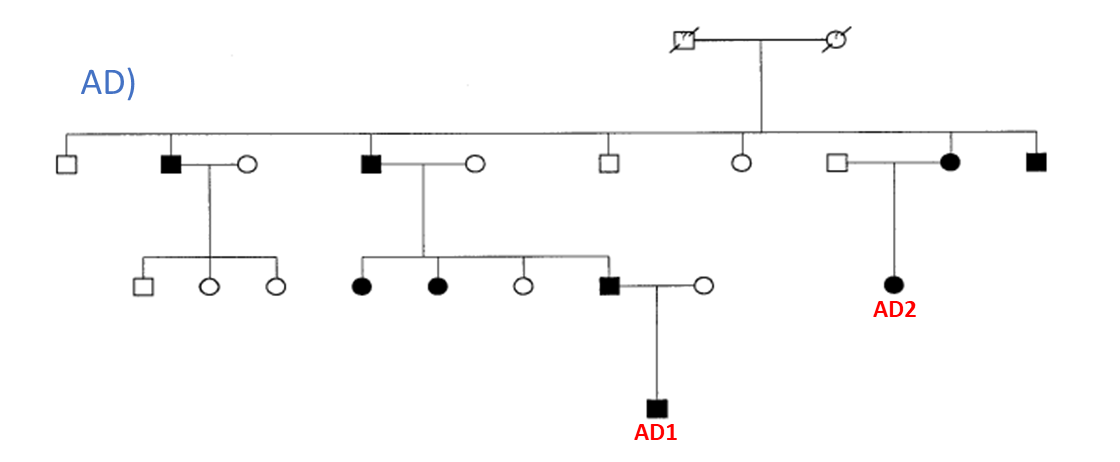
The *variant\_filters* column shows which filters were passed by the variant (--gt-filter 1, 2 or 3)

**Question 20:** Open AR\_results.txt in an excel table and fill the following table: Is this a known NSHI gene \_\_\_\_\_\_\_\_\_\_\_? Verify the phenotype in the gts column

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Family* | *Exome* | *Gene* | *Position* | *Genotype* | *Impact* | *Max\_aaf\_all* |
| B | AR1 |  |  |  |  |  |
| C | AR2 |  |  |  |  |  |
| D | AR3 |  |  |  |  |  |

***Family AD***

Next, we will analyze an autosomal dominant NSHI family of European ancestry (pedigree AD). In order to minimize the number of rare damaging variants, two DNA samples for NGS from two hearing-impaired relatives (first-cousins-once removed) AD1 and AD2 underwent exome sequencing. In this case the *autosomal\_dominant* command line is not executable, so we use gene\_wise with additional filters. The vcf was decomposed and annotated with VEP, we provide AD.vep.db



# We can use *gene\_wise* function to query variants from sequencing data of both probands. We can query variants with custom genotypes using --gt-filter tool. A shared heterozygous variant between both members of the AD family is probably causal.

$ gemini gene\_wise AD.vep.db --min-filters 1 --gt-filter "(gt\_types.AD1 == HET) and (gt\_types.AD2 == HET)" --filter "max\_aaf\_all <0.003" --columns "vcf\_id, chrom, start, end,ref, alt, gene, impact, impact\_severity, max\_aaf\_all, gts, clinvar\_gene\_phenotype, clinvar\_disease\_name, clinvar\_sig " > AD\_results.txt

**Question 21:** Open AD\_results.txt in an excel table and fill the following table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Family* | *Exome* | *Gene* | *Position* | *Genotype* | *Impact* | *Max\_aaf\_all* |
| AD | AD1 |  |  |  |  |  |
| AD2 |  |  |  |

**Question 22:** Is the gene that we found a known NSHL gene \_\_\_\_\_\_\_\_\_\_\_? Is the variant that we found already reported in Clinvar \_\_\_\_\_\_\_\_\_\_\_? What is the clinical significance \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

Based on the combination of filters used, there is only one variant selected, i.e. a known variant rs28937893 in an ADNSHI gene *WFS1.* This *WFS1* p.Ala716Thr variant segregates with low-frequency ADNSHI in the AD family (Bespalova et al. 2001; Lesperance et al. 2003).

**Answers**

**2-Query the database you created**

**Question 1:** What are the possible modes of inheritance for family 1?

-***Autosomal recessive***

***-Autosomal dominant with reduced penetrance***

***-Compound heterozygous***

***-De Novo***

***-X-linked recessive***

***-X-linked dominant (This can happen if the mother shows skewed X-inactivation and carries the mutation)***

***-X-linked De Novo***

gemini query -q "SELECT name FROM samples" fam1.db

***CHILD***

***DAD***

***MOM***

gemini query -q "SELECT name, phenotype FROM samples" fam1.vep.db

***MOM 1***

***DAD 1***

***CHILD 2***

**Question 2**: How many variants are available for analysis ***64307***?

**Questions 3:** How many variants does the mom ***44580***, dad ***28544***, and child ***34003*** have available for analysis?

sample total

***MOM 44580***

***DAD 28544***

***CHILD 34003***

**Questions 4:** How many rare variants are available for analysis knowing that the header is also counted as a line with (wc -l) ***6241***?

**Questions 5:** How many rare variants that are in disease genes reported in Clinvar are available for analysis ***867***?

**Questions 6:** How many rare variants in our sample are known and listed in Cinvar ***12***? Are there any known NSHI variants\_***2: 223066536 C > CA affecting the PAX3 gene is associated with Waardenburg\_syndrome and Craniofacial\_deafness\_hand\_syndrome***? Focus on the clinical significance scores and verify if the known variant listed in Clinvar is considered to be pathogenic \_\_***This variant has an uncertain significance***? Describe the variant genotype in mom***\_\_\_1 A/CA***, dad\_\_\_***0 A/A***, and child \_\_\_1 A/CA? Do we need to go further in the analysis\_***YES because the significance of this variant is uncertain***?

**Question 7:** How many novel variants are observed in these samples\_***2445***?

**3-Analyzing simple genetic models**

**a- Using the autosomal\_recessive command**

**Question 8:** How many variants that segregate under an autosomal recessive mode, are observed in these samples***\_41***? How many variants effect exons ***\_NONE***? How many variants have a high impact\_severity \_***NONE***? Do you suspect any variant to be causal\_***NONE of the variants are reported as loss of function, or change an amino-acid, all informatic predictions are low but we should keep these variants in mind***?

**Question 9:** How many compound heterozygote variants segregate in these samples ***\_NONE***?

1. **b- Using the autosomal\_dominant command**

**Question 10:** How many variants that segregate under an autosomal dominant mode, are observed in these samples\_***NONE***.? ***This is because none of the parents in this trio are indicated as affected. If you suspect reduced penetrance, you will have to indicate the parent with reduced penetrance as affected.***

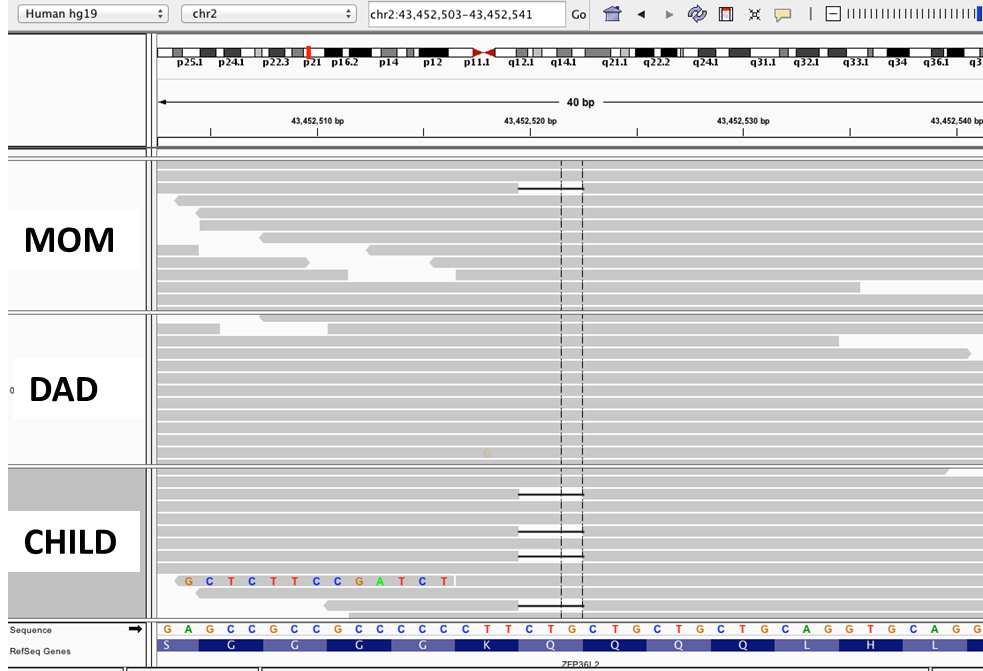
1. **c- Using the de\_novo command**

**Question 11:** How many variants that segregate under a *de novo* mode, are observed in this trio\_\_\_\_\_***\_1***? Is this variation affecting a known NSHI gene \_\_\_\_***\_NO***?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Position | Ref allele | Alt allele | Impact | aa\_change | clinvar\_gene\_phenotype | max\_aaf\_all |
| ZFP36L2 | Chr2: 43452522 | TCTG | T | inframe\_deletion | Q/- | inframe\_deletion | 0.00098056694598 |

**Question 12:** What is the genotype of this variant observed in MOM\_\_\_\_***TCTG/TCTG***, DAD\_***TCTG/TC***TG , CHILD\_***TCTG/T*** ?

**Question 13:** Is the variant *de novo* or likely a false variant call\_***The mother does have one read with the same indel, suggesting that this is a false call. Would most likely want to follow it up with another technology*** ?



1. **d-Using the x\_linked command**

**Question 14:** In this example, is the X linked dominant inheritance mode possible ***\_\_\_YES***? In which case\_\_\_\_ ***skewed X-inactivation in the mother carrying the mutation (i.e. the mother primarily expresses the normal allele in her cells)***?

**Question 15:** How many variants that segregate under an x\_linked\_dominant mode, are observed in this trio\_***NONE***?

**Question 16:** How many variants that segregate under an x\_linked\_recessive mode, are observed in this trio\_\_***\_2***?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Position | Ref allele | Alt allele | Impact | Conservation | Cadd\_scaled | max\_aaf\_all |
| *HMGN5* | Chrx: 80373977 | A | C | missense\_variant | 2.20000004768 | 22.7 | -1 |
| *TEX13A* | chrX: 104463794 | C | T | missense\_variant | -3.90000009537 | 0 | 0.00030061626334 |

**Question 17:** Based on the bioinformatic predictions, which variant is more likely to be causal ***The HMGN5 variant is more likely to be the causal variant***?

**Question 18:** Is the variant de novo or likely a false variant call\_\_ ***This variation is likely to be a x-linked recessive variant. The mutant allele carried by the mother was transmitted to the child (male). This male child is hemizygous for the variant***?



**4- Custom genotype filtering by gene:**

***Family A:***

**Question 19:** How many rare homozygote variants are shared by both samples \_\_\_\_1? Which gene(s) contain these homozygous variants \_\_\_***ADCY1***? Is this a known NSHI gene \_\_***YES***?

***Families B, C and D***

**Question 20:** Open AR\_results.txt in an excel table and fill the following table: Is this a known NSHI gene \_***KARS***? Verify the phenotype in the gts column:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Family* | *Exome* | *Gene* | *Position* | *Ref nucleotide* | *Genotype* | *impact* | *Max\_aaf\_all* |
| B | AR1 | *KARS* | 75670401 | A | G/G | missense\_variant | 0.00022 |
| C | AR2 | *KARS* | 75665623 | C | T/T | missense\_variant | -1 |
| D | AR3 | *KARS* | 75670401 | A | G/G | missense\_variant | 0.00022 |

***Family AD:***

**Question 21**: Open AD\_results.txt in an excel table and fill the following table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Family* | *Exome* | *Gene* | *Position* | *Genotype* | *Impact* | *Max\_aaf\_all* |
| AD | AD1 | *WFS1* | 6303668 | G/A | missense\_variant | 0.000032535137949 |
| AD2 | G/A | missense\_variant | 0.000032535137949 |

**Question 22:** Is the gene that we found a known NSHL gene \_***YES***? Is the variant that we found already reported in Clinvar ***YES***? What is the clinical significance \_ ***pathogenic for WFS1 related\_disorders and likely pathogenic for Diabetes\_mellitus\_AND\_insipidus\_with\_optic\_atrophy\_AND\_deafness*** ?