**Bioinformatics Exercise 2 | BIOL3120**

In this exercise you will use online tools to build evidence about the variants you identified last week, regarding their pathogenicity. You will then make a preliminary conclusion re: the pathogenicity of variants, based on evidence gathered. Lastly you will check a variant pathogenicity database to see if/how this variant has already been classified.

**Learning objectives:**

At the end of this practical, you should be able to:

* Use gnomAD to find population frequency for variants
* Use Sift and PolyPhen-2 protein prediction algorithms to predict the effect of variants on protein function
* Come up with a preliminary pathogenicity classification for your variants based off the evidence you have
* Use Clinvar to find pathogenicity and further information on variants

**Why do variants need to be classified?**

As discussed in lectures, sequencing technology has gotten to the point where it is very cheap and accessible. Sequencing of tens or hundreds of genes, or even entire exomes/genomes occurs more and more frequently, and such sequencing reveals many variants (i.e. differences compared to the reference genome). A system is needed to identify as clearly as possible which variants are likely to be pathogenic, i.e. cause disease.

The American College of Medical Genetics and Genomics (ACMG) system is the most widely used variant classification tool, and those labs or services not using this system would use something very similar to determine possible pathogenicity of variants. Today’s exercise uses a few categories of evidence from the ACMG system, but the real system includes many other types of evidence in classifying variants.

See <https://www.nature.com/articles/gim201530> for the overview paper of the ACMG system.

**Overview**

In the last practical you found variants in patients’ samples, identifying them using standard naming conventions. For today’s practical, we’ve used the BRCA1 variants, but don’t worry - you can still use CFTR and KCNQ1 variants from last week for practice. Today we will expand our information on these variants using some new tools. In this practical we will:

* Use GnomAD to find the frequency of the identified variants in a healthy population
* Use Sift and PolyPhen-2 to examine if the identified variants are predicted to be pathogenic
* Use this evidence to guestimate if the variants might cause a genetic condition
* Use the Clinvar database to find additional information

**Table 1. Aim of the practical, complete this table**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **BRCA1 Patient** | **DNA variant** | **Effect on Protein (Protein Identification)** | **GnomAD allele frequency** | **Sift** | **polyphen** | **Conclusion from evidence** | **Clinvar pathogenicity rating** |
| Patient 1 | c.563A>T | p.E188V or  p.Glu188Val |  |  |  |  |  |
| Patient 2 | c.4327C>T | p.R1443X or  p.Arg1443Ter |  |  |  |  |  |
| Patient 3 | c.1367\_1368del  or  c.1367\_1368delTT | p.I456RfsX23  or  p.I456Rfs\*23  or  p.Lle456ArgfsTer23 |  |  |  |  |  |
| Patient 4 | c.2719G>C | p.E907Q or  p.Glu907Gln |  |  |  |  |  |

**1 – frequency of variant in the general population**

To investigate if an identified variant may be causing a genetic condition, we can examine how common the variant is in a healthy population. This is called the population frequency data. To generate population frequency data, we will use gnomAD - <https://gnomad.broadinstitute.org/> . This database contains exome and genome data from thousands of people, most of whom have no serious health condition. You can read specific details in the ‘about’ part of the website. There is also a publication that accompanies this, and similar databases like it which gives detailed information, and should be used as a reference when using the database.

* Go to the gnomAD website, using the default gnomAD 2.1.1, search for BRCA1. This will bring up information for all the variants identified in the BRCA1 gene, in the people included in this database. In the table at the bottom of the page, the main columns of interest to us are:
  + **Consequence**: the DNA or protein identification of the variant.
  + **Allele count**: the number of alleles with this variant detected in people included in this database
  + **Allele Number**: the total number of alleles for which they have a sequence covering this part of the gene
  + **Allele frequency**: the number of alleles containing this variant divided by the total allele number. Note that the letter e means “10 to the power of”, so for example the top entry for CFTR has an allele frequency of “3.99e-6”. This means 3.99 x 10-6 or 0.00000399 (with 1 being 100%).

***Use the search bar (top right above the variant list with ‘search variant table’ written in it). To try to find your variants using their protein identification (ex. Patient 2,*** p.Arg1443Ter***). This is found in column 3 of table 1. Write the frequency of your variants in the GnomAD allele frequency column of table 1.***

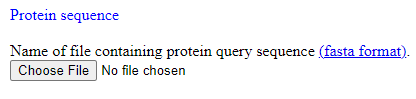
* note that gnomAD uses the three letter amino acid shortcuts, so for example searching for ‘1443’ will greatly reduce the number of variants displayed, and searching for ‘Arg1443’ will hopefully reduce the list even further.
* Your variant may not be in this database - that doesn’t mean it doesn’t exist, just that none of the thousands of people whose sequences are in this database had the variant you are searching for. This information is useful in itself, suggesting that the variant is very rare.
  + However, be careful that you have searched thoroughly before concluding that your variant is not in gnomAD. I recommend manually scrolling through the list and finding where your variant should be (they are in order), before concluding that it is not in gnomAD.
* In general, an allele (variant) with a population allele frequency a cutoff value of 0.005 for recessive, and 0.0005 for dominant conditions is suggested as a general guide for indicating a pathogenic variant (see http://ncbi.nlm.nih.gov/pubmed/31037860 for more info). Therefore, a frequency of 7.17e-3 = 0.00717 would indicate a non-pathogenic variant whilst a score of 7.09e-6 = 0.00000709 would suggest a possible pathogenic variant.

**2 - Missense variant protein function prediction**

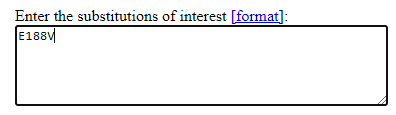
Now you will use protein algorithms to predict whether the identified variants are likely to cause amino acid substitutions are likely to be tolerated (ie the protein can still function even with this change). **Note that these tools are only for substitution of amino acds - frameshift and residue insertion or deletion variants are not appropriate for these algorithms.** You will use two different prediction tools - Sift and polyphen2.

**Sift:**

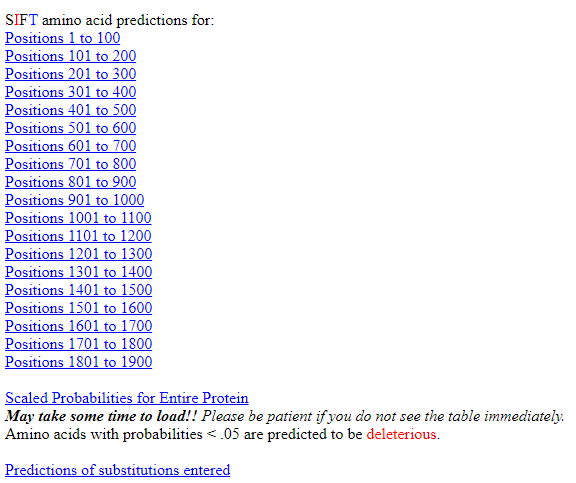
* **Go to the SIFT tool** at <https://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html> This tool requires you provide the protein sequence, as well as the amino acid substitutions you wish to analyse
* To get a reference protein sequence, download the BRCA1 .fasta file from iLearn.
* **Upload the reference amino acid sequence** of your protein of interest



* **In the ‘substitutions of interest’ box, enter the protein variants you wish to investigate in the following format: E88V** (from patient 1)**.** You can enter multiple mutations at the same time.



* Scroll to the bottom of the page and **click submit.** Sift says it can take up to an hour, but this generally only takes about 30 seconds. You can also add your email address above the submit button if you want to be emailed the results.
* On the following results page, click ‘predictions of substitutions entered’ near the bottom of the page.



* This will take you to the predictions for your variants entered, with a description of TOLERATED or AFFECT PROTEIN FUNCTION, and a score.

***For each variant analysed in Sift, note the prediction and score in table 1.***

The SIFT score ranges from 0.0 (deleterious) to 1.0 (tolerated). The score can be interpreted as follows:

* 0.0 to 0.05 -- Variants with scores in this range are considered deleterious. Variants with scores closer to 0.0 are more confidently predicted to be deleterious.
* 0.05 to 1.0-- Variants with scores in this range are predicted to be tolerated (benign). Variants with scores very close to 1.0 are more confidently predicted to be tolerated.

**Polyphen:**

* **Go to the PolyPhen-2** **site** at <http://genetics.bwh.harvard.edu/pph2/>
* Similar to Sift, you need to paste the amino acid sequence into the second box. Open the BRCA1 protein .fasta file (using notepad or TextEdit or other text based program that doesn’t have formatting). Copy and paste the exact sequence PoyPhen-2.
* To indicate the protein change you wish to analyse, enter the position (residue number), the original amino acid (AA1) and the amino acid it has changed to (AA2). Click ‘Submit query’ to begin the analysis. Unfortunately only one substation can be analysed per submission.
* The following page shows a summary of all of your requests. The analysis you just submitted will show up under Pending/Running (bottom of table). Click the refresh button at the bottom of the page, and your analysis should move to the ‘Completed’ section above. Note that some queries will take a long time - possibly hours. This seems to occur if the tool is being used by many people, or the protein you are analysing is very large.
* **Click on ‘view’ to see the results of your analysis**. Polyphen presents results using a different system, but again you have a description of the predicted impact on the protein, and some score information.

***Again, for each variant you analysed, note the prediction in the table 1, and the score given.***

The PolyPhen-2 score ranges from 0.0 (tolerated) to 1.0 (deleterious). Variants with scores of 0.0 are predicted to be benign. Values closer to 1.0 are more confidently predicted to be deleterious. The score can be interpreted as follows:

* 0.0 to 0.15 -- Variants with scores in this range are predicted to be benign.
* 0.15 to 1.0 -- Variants with scores in this range are possibly damaging.
* 0.85 to 1.0 -- Variants with scores in this range are more confidently predicted to be damaging.

PolyPhen-2 and SIFT scores use the same range, 0.0 to 1.0, but with opposite meanings. A variant with a PolyPhen-2 score of 0.0 is predicted to be benign. A variant with a SIFT score of 1.0 is predicted to be benign.

* Do the predictions from both Sift and polyphen-2 agree?

**3 - Preliminary conclusion on pathogenicity of your variants**

You now have some, albeit limited, information on all of your variants. The nature of change to the protein, the population frequency of the variant, and for missense (amino acid substation) variants, two protein predictions. Remember that none of these pieces of evidence on their own will give allow a confident conclusion, but the point of this exercise is to understand whether each piece of evidence pushes our conclusion towards benign, pathogenic, or uncertain regarding pathogenicity.

**Effect on protein:** Some protein changes offer insight as to whether the mutation will effect protein function. For example, a frameshift early in the protein is very likely to interrupt protein function, since the majority of the protein is missing. A silent mutation (ie does not change amino acid) seems unlikely to alter protein function.

**Population frequency**: A variant which is seen at a high frequency in a healthy population is unlikely to be pathogenic. A cutoff value of 0.005 for recessive, and 0.0005 for dominant conditions is suggested as a general guide for indicating a pathogenic variant (see http://ncbi.nlm.nih.gov/pubmed/31037860 for more info).

**Protein prediction algorithms:** Sift and polyphen are quite clear in their conclusion, but borderline scores are possible which add little weight to a pathogenicity conclusion.

***For each variant in your table, classify each piece of evidence as whether it suggests the variant is pathogenic, benign, or uncertain. Colour evidence accordingly, with red suggesting pathogenic, green suggesting benign, and black uncertain.***

***Using the information present in your table, for each variant write a preliminary conclusion of benign, possibly benign, uncertain, possibly pathogenic, or pathogenic.***

**4 - Pathogenicity classification and further info from ClinVar**

The last tool used in this work is ClinVar, a database of variants with pathogenicity classifications, and further information. ClinVar is not curated, and allows submissions without evidence. Likewise, there would be much information generated which does not make it to ClinVar.

* Go to the NCBI front page - <https://www.ncbi.nlm.nih.gov/>, **select ‘ClinVar’ from the drop down menu, and search for your gene BRCA1**. This will give a list of all reported variants with their pathogenicity rating. For BRCA1 there are thousands of results, so searching using the bar at the top is easiest.
* Clinvar can be quite picky, so I suggest you try multiple ways of searching for variants. Below are some examples based on searching for BRCA1 patient 2’s first mutation - c.4327C>T, p.R1443X:
  + Searching by DNA variant is most precise, but more likely to yield no result. Note that ‘BRCA1[gene]’ should precede your search, for example ‘BRCA1[gene] 1443’ will shorten your list significantly, and will go straight to a result if only one exists. This should yield 10 results, which you will need to scan to determine which is the correct entry (in this case, the 5th entry)
  + Lastly you can manually scroll through results, which are again in order (some genes have results displayed in reverse order, as they are on the complementary strand of DNA), though many genes have thousands of entries, so this can take some time.
  + Not all variants will be on ClinVar
* When you do find an entry of the variant you are searching for (again using the c.1624G>T example), it contains useful information:
  + A pathogenic interpretation: the overall pathogenicity rating after taking into account all submissions to clinvar for this variant. In this case: Pathogenic.
  + Review status: Gives a sense of the confidence of the pathogenicity rating from one to four stars. In this case, 4 stars - practice guideline. You can click ‘practice guideline’ to better understand what this rating means
  + The ‘variant details’ tab contains genetic information, including links to an entry for this variant on other databases.
  + Below this is the ‘submitted interpretations and evidence’, which is a list of all evidence which has led to the pathogenicity rating. A list of citations for the variant are below.

***Note the pathogenicity rating in the appropriate table below, include the review status.***

* Does the clinvar classification agree with your own preliminary conclusion? Why/why not?

**For practice:**

**Below I have provided tables for you to do the above work on variants identified last week in the BRCA1 and KCNQ1 genes. Answers are available on iLearn.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **KCNQ1 Patient** | **DNA variant** | **Effect on protein** | **Gnomad freq** | **sift** | **polyphen** | **Conclusion from evidence** | **Clinvar pathogenicity rating** |
| Patient A | c.332A>G | p.Y111C |  |  |  |  |  |
| Patient B | c.162\_184del cgcgcccggcgccccaggtcccg  or  c.162\_184del | p.I54IfsX222  or  p.I54Ifs\*222 |  |  |  |  |  |
| Patient C | c.858C>T | p.D286=  or p.D286D |  |  |  |  |  |
| Patient D | c.914G>A | p.W305X or p.W305\* or p.W305Ter |  |  |  |  |  |
| Patient E | c.746G>C | p.R249T |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CFTR Patient** | **DNA variant** | **Effect on protein** | **Gnomad freq** | **sift** | **polyphen** | **Conclusion from evidence** | **Clinvar pathogenicity rating** |
| Patient A | c.1521\_1523delCTT | p.F508del |  |  |  |  |  |
| Patient B | No mutation | - |  |  |  |  |  |
| Patient C | c.1571\_1572insAC | p.C524\* or X or Ter |  |  |  |  |  |
| Patient D | c.3790G>A | p.E1264K |  |  |  |  |  |
| Patient E | c.1624G>T  and  c.3790G>A | p.G542\* or X or Ter  and  p.E1264K |  |  |  |  |  |