**Bioinformatics Exercise 1 | BIOL3120**

In this exercise you will learn how to use Benchling, software which can be used for a range of DNA analyses. Today you will focus on using Benchling to acquire sequences from NCBI’s database, in order to identify variants in patient sequences. You will then name the variants using standard naming conventions.

**Learning objectives:**

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

* Obtain reference sequences from NCBI for the full gene, mRNA, CDS and protein of a gene of interest
* Compare provided sequences with a reference sequence to discover variants
* Name your variants using proper nomenclature

There are questions throughout this exercise - these are aimed at raising your awareness of what you are actually doing in each step of this prac, as well as reinforcing information from lectures. ***The most useful way you can approach bioinformatics work is to be as aware as possible of why you are doing what you are doing at each step*** - the point of this exercise is not to get the answers, but to be aware of what you are doing, and become familiar/comfortable with the process of identifying variants in patient sequences.

*Note also that all sequences used in our bioinformatics work refer to the coding strand, or non-template strand of DNA.*

**Get to know Benchling**

Benchling is a free (for academic use) web-based program, available at <http://benchling.com> - head to this address and register an account, using your Macquarie email address:

* Affiliation - academic
* Company name / phone / team / role are not important
* ‘Join an organization’ can be left blank
* On the ‘welcome to benchling’ screen, click ‘start cloning’ under Molecular bio & CRISPR

The first screen you see is an example sequence of a plasmid. I will help you with this in class, but some main functions you should be familiar with are:

* Split workspace (bottom right) - allows you to have two views open at a time. I recommend working with only one main workspace.
* Projects (the briefcase icon top left) - allows you to organise your work into folders. Click the ‘projects’ text above example projects, to get to the overview of all your projects.

1. **Make a new project called ‘Exercise 1’** - we will organise our work in this project for this week:

* Click the + icon in the left pane -> project
* Name: Exercise 1. The rest can be left as it is, click ‘create project’

1. **Import the human BRCA1 gene sequence into your Exercise 1 project:**

* Click the + icon in the left pane -> DNA sequence -> import DNA sequences
* Click the ‘search external databases’ tab
* In the search box, type ‘BRCA1’ (without the quote marks) and click search. This opens up a lot of new options.
* For genome, select GRCh38 (human reference genome. Note that 37 is the only option for some genes - this is ok too, but use 38 where it is available).
* Leave transcript as the default.
* Leave the ‘import cDNA’ option unticked. Ticking this will import only the mRNA sequence, which we will use in future work. Leaving it unticked will import the entire gene sequence, including introns.
* With ‘Import As’ you can change the name that the sequence will take in your list, for now label it as ‘BRCA1 - full gene’.
* ‘Save to’ will choose which project folder this sequence goes to. Choose your Exercise 1 project folder.
* Click import

1. **Customise your view**

You should now be looking at the sequence map (see tabs) of BRCA1 gene sequence, with various features showing. You can customise which features are displayed by clicking the settings () icon to the top right of the sequence. I recommend having the following options ticked, and leaving the rest unticked:

* Annotations
* Axis
* Translation
* Amino acid indices

This should leave you looking at the BRCA1 gene sequence, with exons visible, and the amino acid sequence of the protein visible above the exons. I recommend you work in this view, because all required information is visible.

You can click on the ‘linear map’ tab to see the entire sequence in one horizontal line, but you will again have to adjust which features are shown (cog icon). In this view I recommend having the following ticked:

* Annotations
* Axis

Note that there is a zoom slider in both views, above the sequence view. Feel free to experiment with any other view options (and Benchling as a whole!). Also note that in both views there is information available in the status bar below the sequence.

**Questions regarding the BRCA1 gene sequence:**

* How long is this gene, in basepairs?
* How many exons are present in this gene?
* What are between the exons?
* What main gene feature is before the first exon?

1. **Importing a transcript (mRNA) sequence**

Do another sequence import as above in part 2, **but this time tick the ‘import cDNA’ option**, and name the sequence as ‘BRCA1 transcript’. Leave the ‘transcript’ choice as it is, but note that there are a range of transcripts you can choose from. cDNA is DNA created from mRNA, so you are now looking at the BRCA1 DNA sequence for those parts that end up being in the mRNA. It isn’t an mRNA sequence - there are still thymine basepairs present.

Note that clicking your cursor on a base pair will show you the number of this base pair on the status bar at the bottom (‘Insert 706’ indicates that your cursor is before basepair 706), and highlighting a region will give you further info in the same place.

**Questions regarding the BRCA1 transcript sequence:**

* There are multiple transcripts available for download - what changes between different transcripts of the same gene?
* What is the size in basepairs of this transcript?
* What gene features are missing in this transcript sequence compared to the full gene sequence?
* Which is the largest exon in this transcript? How big is it?
* At which basepair of this transcript does the coding region (CDS) start?
* What is the first codon and amino acid of the protein made from this transcript? What is this codon referred to as?
* How long (in amino acids/residues - these mean the same thing) is the BRCA1 protein made from this transcript?
* What amino acid is at residue 88 of this protein product?

1. **Creating a CDS reference sequence from the transcript sequence**

* Within the transcript sequence, highlight the DNA sequence which spans the length of the protein product (perhaps most easy by clicking the cursor before the first basepair, then scrolling and shift clicking after the last basepair).
* Right click your highlighted sequence -> Create DNA sequence. Tick the following options: Annotations, translations and primers, Tags
* Select your Exercise 1 folder as the destination.
* You will now have a new DNA sequence in your list. Right click -> rename, and rename this sequence to BRCA1 CDS
* Note that in this process you may lose the first few exons from the sequence - this is ok.
* Highlight the entire sequence (control + a or command + a on macs), right click the highlighted sequence -> create translation -> forward.

You should now have the DNA and amino acid sequence for the CDS of the BRCA1 gene visible. The point of parts 4 and 5 was to create this reference CDS for the gene, to use in alignments with patient samples to identify variants.

***Note that you want to be aligning to find variants off of the CDS, rather than the transcript/mRNA sequence, or full gene sequence.***

1. **Importing patient sequences and aligning against the CDS reference**

* I have provided a file which contains four BRCA1 patient sequences on iLearn. Download the BRCA1 patient files from ilearn onto your computer (trying to drag files directly from ilearn never seems to go well).
* Working from the view of the BRCA1 CDS sequence, click the alignment button ( ) on the right sidebar -> create new alignment.
* You are now in the alignment setup window. Below under Sequences, your BRCA1 CDS should be marked as the template.
* Click ‘Choose file(s)’, and find the BRCA1 patient files you saved onto your computer. Highlight this file and click open.
* Click ‘create alignment’. This may take a short time to align. Notice that alignments are accessed through the top menu (next to sequence map, etc).

You are now shown the alignment of the patient sequences against the reference CDS sequence. You will notice there is an overview down the bottom, with black bars representing your aligned sequences, and red lines on these bars showing where the patient sequences differ from the reference CDS (ie, a variant is present).

1. **Identifying variants in the patient sequences**

* To jump to mismatches, click the arrows next the ‘Find mismatches’ above the alignment. This will jump to each mismatch in turn.
* Each mismatch is highlighted, and your cursor will be in front of the mismatch. Identify the location of the basepair(s) that have changed in each patient, and what exactly has changed at this point, in the table on the next page.
  + For example, patient one has a T at basepair 563, when the reference sequence has an A at this point. So I have put c.563A>T to indicate that in this sequence, an A was expected but a T was reported at basepair 563.
* To determine the change to the protein sequence, you need to create translations in this alignment. Select the entire sequence of any of the patients (double click the sequence, or click in it and press control-A or command-A for macs), right-click the highlighted sequences and select Create translation-> Forward (all sequences).
  + Now you can see the protein (amino acid) sequences which would be produced from each of these DNA sequences. If this isn’t displaying for you, make sure translations are being displayed (click the cog icon, make sure translations is ticked).
  + In patient one, at the location of the basepair substitution, we can see that where the reference sequence amino acid shows at E (glutamic acid) at residue 188, in patient 1 the change in DNA sequence causes this residue to change to a V (valine). Therefore I write in the table below, for the protein change: p.E188V
  + While you only need to identify the changes by the 1-letter amino acid code, it may be useful to know which exact amino acids these are, and what type of amino acid they are. See the image after the table for details on all amino acids.

Identify variants in the remaining patient sequences. The following two naming guides were provided with lecture 9 -Nucleotide mutations, but are provided again for your convenience:

<http://atlasgeneticsoncology.org/Educ/NomMutID30067ES.html>

<https://www.hgvs.org/mutnomen/recs.html>

|  |  |  |
| --- | --- | --- |
| **Patient** | **DNA variant** | **Effect on Protein** |
| Patient 1 | c.563A>T | p.E188V |
| Patient 2 |  |  |
| Patient 3 |  |  |
| Patient 4 |  |  |

**Further practice:**

You have been provided patient sequences for two other genes - KCNQ1 and CFTR. Repeat steps 4-7 for CFTR and KCNQ1 to identify the variants. Also find brief clinical information regarding these genes - the condition they can cause, mode of inheritance, and any information / treatment / management options for someone who is found to have disease causing variant(s).

|  |  |  |
| --- | --- | --- |
| **KCNQ1 Patient** | **DNA variant** | **Effect on protein** |
| Patient A | c.332A>G | Y111C |
| Patient B | c.162\_184del  cgcgcccggcgccccaggtcccg | G57Cfs\*222 |
| Patient C |  |  |
| Patient D | c.914G>A | W305X W305Ter |
| Patient E |  |  |

**KCNQ1 Clinical information:**

|  |  |  |
| --- | --- | --- |
| **CFTR Patient** | **DNA variant** | **Effect on protein** |
| Patient A |  |  |
| Patient B |  |  |
| Patient C | c.1571\_1572insAC | C524X |
| Patient D |  |  |
| Patient E |  |  |

**CFTR Clinical information:**

* Question for CFTR sequences: one of these patients has two mutations - how would you determine if these are both on the same copy, or each on separate copies of the CFTR gene?