# DNA Sequence- first part

## Locate the DNA sequence within a gene (hint- use BLAST)

Using the sequence in the file DNASequence.txt:

* Is the sequence in a gene? If so, which gene is it in? (hint- this is a human sequence)
  + What version of the reference sequence are you using (if applicable?) Where is the sequence located (relative to that reference sequence)?
* What is the tool doing (briefly)?
* What does the output from the tool mean? (hint- use the documentation)
  + How confident is it? (hint- what is the meaning of the E score and the BitScore (in brief))
  + Query cover and Identity? What do these mean?
    - Have a look at the differences in output for DNA sequences 2 and 3 to help your understanding. Are the query cover and identity different for DNA sequences 2 and 3 when they are compared against the same reference sequence?
* What does this gene do (briefly)?
* Are there any diseases associated with this gene in humans? If so, pick one and make a few notes about it.
  + Hints:
    - Symptoms
    - Inheritance pattern
    - Mechanism of disease (if known)
    - Clinical management of condition

## Resources

BLAST (available through NCBI and EBI)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

<https://www.ebi.ac.uk/Tools/sss/ncbiblast/>

<https://www.ncbi.nlm.nih.gov/books/NBK153387/>

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\_TYPE=BlastDocs&DOC\_TYPE=FAQ#expect

https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

DNASequence.txt

DNASequence2.txt

DNASequence3.txt

# DNA Sequence- second part

## Take a DNA sequence and use exon prediction tools/transcription factor binding site prediction software

Exon prediction tools will be used with prokaryotic data in the phylogenetics competency document. Although software exists to attempt it, gene prediction is very difficult in eukaryotes. To be completely accurate, the software would have to identify the reading frame correctly (more on this in the translation into a protein sequence section below), identify the introns and the exons accurately from many putative splice sites, and know which of the possible alternate transcripts were the correct ones (exons to be included in or excluded from the transcript). This process is not used diagnostically in human genomics. The closest thing that is used in the clinical setting is prediction of disruption to known splice sites using software. This will be covered in the third part of this activity. There are no currently functional available tools to predict transcription factor binding sites through a web interface.

Transcription factor binding site prediction is not currently used diagnostically. Prediction of sites is not sufficiently reliable to be useful. It could potentially be useful to identify if changes to someone’s DNA may alter transcription factor binding and so change the way a gene is regulated (e.g. turned on or off).

1. Do you think that the output of these prediction tools used in this section will be the same as is found *in vivo*?

## Resources

### Exon prediction

#### Hidden Markov Model-based

<http://hollywood.mit.edu/GENSCAN.html>

#### Similarity-based

<http://www.softberry.com/berry.phtml?topic=fgenes_plus&group=programs&subgroup=gfs>

### Splice-site prediction

<http://www.cbs.dtu.dk/services/NetGene2/>

<http://www.fruitfly.org/seq_tools/splice.html>

#### Literature

<https://www.sciencedirect.com/topics/medicine-and-dentistry/gene-prediction>

### Transcription factor prediction

Tool is no longer working

#### Database of transcription factors

<http://jaspar.genereg.net/>

DNAExPred.txt

## Take a fragment of DNA sequence and translate into protein sequence

* Use more than one tool (start with ExPASy). Do the outputs match?
* ExPASy gives multiple outputs (different boxes). What do the different boxes mean?
  + This could be confusing- please ask for help if you need it
* Which gene is it from? (hint- protein blast, hmmer)
* How could we get the known protein if it is available (this could help us to choose between our different predicted protein sequence options output by our translation tools (e.g. different reading frames))? (hint- NCBI database)
  + Retrieve the known protein that best matches the predicted protein sequence
  + Compare the predicted sequence and actual known protein sequence. Is there a difference? If so, what is it? (hint ClustalOmega)
    - Why/Why not?
    - Save a screenshot of the output
  + Save the sequence of the known protein in a text file called “ProteinSequence.txt”. This is for later work.
* What kind of function is it likely to have (hint- hmmer output, try InterPro)
  + Search the predicted sequence and actual known protein sequence in InterPro. Is there a difference in the outputs? If so, what is it?
  + Save a screenshot of the output for both
* What does this gene/protein do (briefly)?
* Are there any diseases associated with this gene in humans? If so, pick one and make a few notes about it.
  + Hints:
    - Symptoms
    - Inheritance pattern
    - Mechanism of disease (if known)
    - Clinical management of condition

## Resources

### Translate Tools

<https://web.expasy.org/translate/>

<https://www.ebi.ac.uk/Tools/st/emboss_transeq/>

### Open reading frame identifier

<https://www.ncbi.nlm.nih.gov/orffinder/>

### Search sequence- identify

<https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>

<https://www.ebi.ac.uk/Tools/sss/ncbiblast/>

<https://www.ebi.ac.uk/Tools/hmmer/>

### Search sequence- compare

<http://www.clustal.org/omega/> (note available to run through EBI and Uniprot websites)

<https://www.ncbi.nlm.nih.gov/books/NBK20261/#_A168_>

### Search database- functional domains

<http://www.ebi.ac.uk/interpro/sequence-search>

SecondDNASequence.txt

# DNA Sequence- third part

## Locate DNA sequences within a gene (pipelines in the genetics service)

### Overview of the analysis pipeline

#### Storage and data processing

* Where are the pipelines and genomic data stored in the human genomics service? Why are they there?
* Genomic data is considered personal information under the GDPR. How is data security maintained around this data in the human genomics laboratory (briefly)?

#### Pipeline Overview

Draw a schematic of the major steps of a bioinformatics pipeline used in a human genomics laboratory. Include the main file types and an example of the software used at each step.

Hints:

* Which tool do we typically use in human genomics to locate DNA sequences within the genome (such as those stored within the fastq file)?
* Which tool do we typically use in human genomics to locate variation within the genome (such as those that can be seen in the bam file if you load it into IGV)
* Which tool do we typically use in human genomics to annotate variation with useful metadata (such as could be useful in clinically interpreting a variant)?
* For all of the above, name the kind of tool and give one example of this software

### The output/results from the data in more detail

#### Single nucleotide variants

1. What file type is the standard format for storing information about the variants which were found in a sample?

Working on the Apple computer on the Trainee account, Look at the file null\_18M01315\_VariantReport.txt

* Were there any variants identified in the sample?
  + What kinds of variants are there in this file? (e.g. SNV, CNV, indel (insertion and/or deletion))
* Find the variant 11:65319712G>A and the variant 2:227875202A>G in the file
  + Is the sample homozygous or heterozygous for the variants?
  + Are they in genes? If so, which genes are they in (if applicable)?
    - How did we get this information to add to our file when we were at the variant annotation stage of the analysis pipeline?
  + If they are relevant to a disorder, what is its inheritance pattern (e.g. autosomal dominant)?
  + What is the position of the variants relative to the transcript? (note the transcript used and the position in HGVS nomenclature)
  + How might we visualise the variants graphically using our own data generated by the pipeline?
    - What software could we use? Which file types from our sample do we need to load in? Which complementary file do we need to load in or select to give a context to our data?
    - Take a screenshot

#### Copy Number Changes

A large copy number change (>1kb) is an example of a structural variant. Structural variants may affect more than one gene. Usually when we know we are looking for a large copy number change we use a different technique to Next Generation Sequencing (NGS) called Array CGH.

* What is a copy number change?
* What sort of phenotypes might be associated with a large copy number changes (hint- what is the referral criteria for array CGH testing, as this is a robust method for detecting them)?
  + The process for deciding whether the genetic change is relevant to the patient’s phenotype is similar to the process for smaller changes (followed in the three clinical cases competency). Although, the repositories of information that are looked may be different. Have a quick look at the databases of structural variants (see resources section).

Open the file 181116\_D00501\_0252\_AHML5YBCX2\_CNV.bam

* Got to the position 6:39873684-39873975
  + What can you see in the coverage plot at the centre of this position? Take a screenshot.
    - What might this coverage change mean?

Open the file 190301\_D00501\_0293\_BHVK3TBCX2\_18M01315\_cnv.vcf

* Are there any copy number changes detected in the sample we have from our NGS?
* Which software was used to call these copy number changes?
  + This is important to know as we can then look at the documentation for this tool, find out which methods and thresholds are being used to call these CNVs

#### Exon Boundaries

* Choose any gene that has a variant called in the sample and identify the exon boundaries for the longest transcript of the gene
  + Hint- look up the gene in a genome browser and take screenshots
    - Which version of the genomic reference sequence are you using? What is the transcript identifier (include NCBI identifier as well as Ensembl identifier if using Ensembl)?
      * Why is this relevant information?
    - How do you know it is the longest transcript?
    - Why is the longest transcript often used in human genomics?

#### Population frequency of a variant

Locate variant 13:111102718C>T in the sample (in the file null\_18M01315\_VariantReport.txt and in IGV- take a screenshot of the variant in IGV). What is the population frequency of this variant? (hint- use GnomAD through the web interface). Take a screenshot.

* + What does this population frequency mean?
    - Is this variant likely to be associated with a rare disease phenotype? Why or why not?

#### Splicing and Splicing tools

Locate variant 11:6638385C>T in the sample (in the file null\_18M01315\_VariantReport.txt and in a genome browser).

* Is it in a protein coding region of the genome?
* What output do we get from running splicing tools on the variant 11:6638385C>T? What might this output mean? (Hint- Alamut is easiest to understand but might still be difficult without help)

## Resources

### Genome Annotation (VEP)

<https://www.ensembl.org/info/docs/tools/vep/index.html>

### Visualising our bam files (IGV)

<http://software.broadinstitute.org/software/igv/>

### Copy Number Changes

<http://dgv.tcag.ca/dgv/app/home>

<https://www.ncbi.nlm.nih.gov/dbvar/content/faq/>

### Genome Browser

<http://grch37.ensembl.org/Homo_sapiens/Info/Index>

### Population frequency (gnomAD database)

<https://gnomad.broadinstitute.org/>

### Splice site prediction

<https://www.interactive-biosoftware.com/doc/alamut-visual/2.6/splicing.html> (this software is available on the genetics computers. Just in case all licenses are in use, screenshots of this variant from this software have been provided).

<http://www.cbs.dtu.dk/services/NetGene2/>

<http://www.fruitfly.org/seq_tools/splice.html>

null\_18M01315\_VariantReport.txt

190301\_D00501\_0293\_BHVK3TBCX2\_18M01315.bam

190301\_D00501\_0293\_BHVK3TBCX2\_18M01315.bai

190301\_D00501\_0293\_BHVK3TBCX2\_18M01315\_cnv.vcf

181116\_D00501\_0252\_AHML5YBCX2\_CNV.bam

181116\_D00501\_0252\_AHML5YBCX2\_CNV.bai