Instructions
Practical 2B Use of online data resources for molecular biology

Familial Adenomatous Polyposis Coli (FAP) is an inherited disease with a classical autosomal inheritance pattern, suggesting a monogenetic cause. The APC gene has been associated to this disease. In this practical we will study the gene, the function of the protein it encodes, known genetic variations and mutations in the gene, and their effect on the function and structure of the protein.

This computer lab connects to chapters 3 (genetic variants) and 6 (proteins) of the Reference guide online resources.

Assignment 1: genetic variants in the APC gene in OMIM *(10 minutes)*

▶ *Section 3.8 of the Reference guide.*

Last time you have already used the OMIM database at NCBI, which contains expert-curated information on the links between genes and genetic variants and human disease.

Look up the APC gene in OMIM.

a) What is the OMIM identifier of the APC gene?

611731

b) To which diseases has the gene been linked and with which inheritance pattern (if known)?

Adenoma, periampullary, somatic
Adenomatous polyposis coli – Autosomal dominant
Brain tumor-polyposis syndrome 2 – Autosomal dominant
Colorectal cancer, somatic
Desmoid disease, hereditary – Autosomal dominant
Gardner syndrome – Autosomal dominant
Gastric cancer, somatic
Hepatoblastoma, somatic

Open the Table View of Allelic Variants for the APC gene from the left-side menu.

c) How many mutations are described in OMIM and which types of mutations do they include?

57 mutations are given; these include a few coding and non-coding SNPs, but also many small deletions and insertions; many of the SNPs are terminating (nonsense) mutations.

d) Have all described variants been associated with (a form of) Familial Adenomatous Polyposis Coli (FAP)?

No, many have been but not all.

Select variant 611731.0016 from the table.

e) Which change in the gene and protein and position in the protein are reported? What is the rs number of the SNP?

   C to T change in the gene; arginine (ARG) to stop codon (TER) at position 554 in the protein; the rs number is rs137854573.

Now go back to the main page of the APC gene, and click the OMIM identifier of Familial Adenomatous Polyposis Coli (175100) in the table.

f) Are other genes associated to FAP type 1, as described in OMIM? If so, which one(s)?

   No, only APC.

g) Are other genes associated with any type of FAP, as described in OMIM? If so, which one(s)?

   Yes, MUTYH (FAP type 2); NTHL1 (FAP3); MSH3 (FAP4). Note: these are all recessive forms of FAP.

Assignment 2A: genetic variants in the APC gene in dbSNP *(30 minutes + 10 minutes to discuss)*

▶ *Sections 3.5 till 3.7 of the Reference guide.*

Look up SNP rs137854573 (the one we addressed in OMIM) in dbSNP.

a) Which allelic change and position in RefSeq mRNA NM_000038 and protein NP_000029 are given by dbSNP? Does this correspond to what you have found in OMIM?

   C to T change at position 1660 in the mRNA; arginine (ARG) to stop codon (TER) at position 554 in the protein; this corresponds.

b) Why are there so many names given at the right side of the dbSNP page (sequence identifiers and positions)?

   Because the position (and possibly change) depend on the splice variant or sequence given.

Scroll down to find the gene model for NM_000038 and NP_000029.

c) Which position is given for the variant in the mRNA and the protein? Compare this to your answer in question a. How can you explain your findings?

   Position in the mRNA: 1745; position in the protein: 554; the difference occurs because here the position is given relative to the start of the mRNA, including the 5'UTR; the position given in the name was the position from the start coding (you can verify this: 1660 is three times the position in the protein (+/- the triplet length).

d) The APC gene is encoded by the forward strand. What would happen to the chromosomal alleles given for the SNP (the NC_ code) when the gene would have been on the reverse strand?

They would be complimentary to the SNPs as given by the mRNA. So for a C -> T change in the mRNA, the change on genomic level would be given as G -> A (as this is always based on the reference genome, which gives the sequence of the forward strand).

It is also possible to look for all SNPs in a gene. A convenient way of doing so, is looking the gene up in NCBI Gene (see previous computer lab) and then click 'SNP: GeneView' from the menu at the right side of the page. Find the human APC gene in the NCBI Gene database.

e) What is the NCBI Gene identifier of the human APC gene?

324

f) What do we learn from the short description given for this gene? Does this correspond to what we found out in OMIM?

It mentions that many known mutations in the gene are truncating ones; this corresponds; also it mentions that most of those are in the same part of the sequence.

Open the GeneView SNP table by following the link. Verify that by default it only shows the coding SNPs. Change the setting from coding SNPs (cSNP) to all SNPs with a known allele frequency (has frequency), by using the radio buttons on top of the table.

g) Which types of SNPs are given by dbSNP and which colour code is used?

- 5' near gene (upstream) (white)
- 5' UTR (orange)
- Coding missense (red)
- Coding nonsense (truncating) (red)
- Coding synonymous (green)
- Frame shift (blue)
- Intron (pale yellow)
- 3' UTR (orange)
- 3' near gene (downstream) (white)

We will now look at information related to genotype frequencies in different populations. These are often not known for rare pathogenic SNPs as they are difficult to measure in a study population. So we will now focus at another SNP with rs number rs2229992.

h) What type of SNP is the variant with rs number rs2229992? What is its minor allele frequency (MAF)?

A synonymous coding SNP; the MAF is 49.00%.

Open the SNPs page by clicking its rs number. The SNP has been measured in multiple population genetics projects. One of them is the HapMap project with identifier ss5258439. Scroll down to find the data generated by this project.

i) Complete the table below.

Population	Number of individuals	Allele frequencies		Genotype frequencies			HWP Goodness of Fit p value
		C	T	C/C	C/T	T/T	
CEU	112	62.95%	37.05%	33.93%	58.04%	08.04%	0.010
JPT	86	72.67%	27.33%	52.33%	40.70%	06.98%	1.000
YRI	113	10.18%	89.82%	01.77%	16.81%	81.42%	0.403

j) Based on the table, are the allele frequencies similar or different between different populations?

They are very different. The by far most common allele in the Africans from Nigeria is the T allele, whereas the C allele is more common in both other populations.

k) What would the expected genotype frequencies by for the Utah residents of European origin? And what for the inhabitants of Tokyo? Do they correspond to the observed genotype frequencies?

CEU: C/C: 0.6295^2 = 39.63%; C/T: 0.6295*0.3705*2 = 46.65%; T/T: 0.3705^2 = 13.73%

JPT: C/C: 0.7267^2 = 52.81%; C/T: 0.7267*0.2733*2 = 39.72%; T/T: 0.2733^2 = 07.47%

For CEU the observed frequencies deviate quite a bit from the expected ones

For JPT the observed frequencies roughly correspond to the expected ones

When genotype frequencies correspond to the expected ones given the allele frequencies we say that the SNP is in Hardy Weinberg Equilibrium (HWE) or Hardy Weinberg Principle (HWP). This can be tested, and is given in the table as the HWE Goodness of Fit p value.

l) What do the p values tell you for both populations? Does this correspond to your own computation? What does the p value depend on?

For CEU it is significant, which means we reject H0 (that there is no deviation from expected genotype frequencies). For JPT it is totally not significant, so we do not reject H0. This corresponds to your own findings.

The p value depends on the difference between the observed and expected frequencies, but also on the number of individuals measured: the bigger the sample size, the smaller the difference that will be significant.

Assignment 2B: genetic variants in the APC gene in Ensembl *(30 minutes + 10 minutes to discuss)*

▶ *Sections 3.1 till 3.4 of the Reference guide.*

Look up SNP rs137854573 (the one we addressed in OMIM) in Ensembl.

a) Which allelic change and position in transcript ENST00000257430 and protein ENSP00000257430 are given by dbSNP? Does this correspond to what you have found in OMIM?

C to T change at position 1660 in the transcript; arginine (Arg) to stop codon (Ter) at position 554 in the protein; this corresponds.

b) Why are there so many names given for the SNP at the Ensembl page (sequence identifiers and positions)?

Because the position (and possibly change) depend on the splice variant or sequence given.

c) The APC gene is encoded by the forward strand. What would happen to the chromosomal alleles given for the SNP (the NC_ code) when the gene would have been on the reverse strand?

They would be complimentary to the SNPs as given by the mRNA. So for a C -> T change in the mRNA, the change on genomic level would be given as G -> A (as this is always based on the reference genome, which gives the sequence of the forward strand).

Open the phenotype data section.

d) To which phenotypes has the SNP been associated according to Ensembl?

Susceptibility to colorectal cancer; Familial Adenomatous Polyposis 1; Hereditary cancer-predisposing syndrome.

Open the Genes and regulation section.

e) What is the difference between the ‘Position in transcript’ and ‘Position in CDS’ given?

The position in transcript is the position in the complete mRNA, including 5’ UTR; the position in CDS (coding sequence) is the position from the start codon.

f) To which gene(s) is the SNP associated and what is/are the Ensembl identifier(s)?

Only to the APC gene; its Ensembl identifier is ENSG00000134982.

Now open the page of the APC gene in Ensembl. From there it is possible to open a table with all SNPs linked to the gene, by opening the Variant table.

g) Which types of SNPs are given by Ensembl and which colour code is used?

- upstream gene variant (grey)
- 5 prime UTR variant (pale green-blue)
- intron variant (blue)
- splice region variant (orange)
- splice donor variant (orange)
- coding sequence variant (green)
- missense variant (yellow)
- synonymous variant (bright green)
- stop gained (red)
- frameshift variant (purple)
- inframe insertion (pink)
- inframe deletion (pink)
- 3 prima UTR variant (pale green-blue)
- downstream gene variant (grey)
- (and some others)

(Note: you can see the classes more easily, by clicking the 'Consequences' button at the top of the table)

Using the 'Consequences' button at the top of the table, select only truncating (PTV) and missense variants to filter the table. Then sort the table by rs number, by clicking the column header. Find the C/G SNP with identifier rs577466163.

h) Why is the SNP present in the table more than once?

Because it is given once for each transcript (splice variant).

i) What are the SIFT and PolyPhen scores of this SNP? What do these scores mean? How are they computed and why are they only given for missense SNPs?

They range between 0.01 and 0.05 for SIFT, and between 0.987 and 0.999 for PolyPhen. Both indicate a 'deleterious' or 'probably damaging' consequence on the protein. SIFT and PolyPhen are tools to predict the effect of amino acid substitutions on the functionality of a protein. SIFT does this by looking at conservation of the residue only, PolyPhen adds annotation such as domains, and if available 3D structural information to this. They are only given for missense SNPs, as for others no predictions can be made by these tools.

We will now look at information related to genotype frequencies in different populations. Since the SNP we are looking at is very rare, we will now focus at another SNP with rs number rs2229992, which has a higher minor allele frequency (MAF). Look this SNP up in Ensembl.

j) What type of SNP is the variant with rs number rs2229992? What is its minor allele frequency (MAF)?

A synonymous coding SNP; the MAF is 49%.

Open the Population genetics section; the SNP has been measured in multiple population genetics projects. One of them is the HapMap project with identifier ss5258439. Scroll down to find the data generated by this project.

k) Complete the table below.

Population	Number of individuals	Allele frequencies		Genotype frequencies		
		C	T	C/C	C/T	T/T
CEU	112	62.95%	37.05%	33.93%	58.04%	08.04%
JPT	86	72.67%	27.33%	52.33%	40.70%	06.98%
YRI	113	10.18%	89.82%	01.77%	16.81%	81.42%

Note: the values in this table are given in more decimal than Ensembl does, as I copied the table from question 2A h). The number of individuals can be computed by summing the number of alleles and dividing by two.

l) Based on the table, are the allele frequencies similar or different between different populations?

They are very different. The by far most common allele in the Africans from Nigeria is the T allele, whereas the C allele is more common in both other populations.

When genotype frequencies correspond to the expected ones given the allele frequencies we say that the SNP is in Hardy Weinberg Equilibrium (HWE) or Hardy Weinberg Principle (HWP).

m) What would the expected genotype frequencies by for the Utah residents of European origin? And what for the inhabitants of Tokyo? Do they correspond to the observed genotype frequencies?

CEU: C/C: $0.6295^2 = 39.63\%$; C/T: $0.6295*0.3705*2 = 46.65\%$; T/T: $0.3705^2 = 13.73\%$

JPT: C/C: $0.7267^2 = 52.81\%$; C/T: $0.7267*0.2733*2 = 39.72\%$; T/T: $0.2733^2 = 07.47\%$

For CEU the observed frequencies deviate quite a bit from the expected ones

For JPT the observed frequencies roughly correspond to the expected ones

Assignment 3: the APC protein in UniProt *(20 minutes)*

▶ *Section 6.1 of the Reference guide.*

Now we turn to the UniProt and RCSB PDB website, to find more information about the function and structure of the protein encoded by the APC gene.

First search the human APC protein in the UniProt database.

a) What is the UniProt identifier of the human APC gene?

P25054

The page starts off with a nice concise description of the function of the protein, supported by references (verify this). Then for APC, some generic information on the types and positions of the mutations is given.

b) Does the information on mutations given by UniProt correspond to what you have seen in NCBI/Ensembl?

Yes, it does (many truncating, clustered in mutations cluster region or MCR).

Then information on the GO annotations of the protein is given.

c) What is the evidence type associated with the annotation to cell cycle arrest? (Hint: click the Source to find out)

It is 'Inferred from direct assay' (click some others to get an idea which types we can have).

d) Which general subcellular locations are given by UniProt annotation and which by GO annotation?

UniProt: plasma membrane, cytoskeleton; GO: Cytoskeleton, Cytosol, Nucleus, Plasma membrane.

Then we find an overview of involvement of mutations in the protein in human disease.

e) Where in the sequence do we find the most disease-causing mutations?

In the part from about one third to one half of the sequence.

We will not go into mutations further for now, as we have already studied this extensively). After the disease-associated mutations, a section is presented on processing of the protein (*e.g.* cleaving, amino acid modifications).

f) Which is the most common modification present for this protein? What is this used for?

Phosphorylation; this is commonly used for activation of the protein; in this case we can also read (below the table) that the GSK3B protein is known to phosphorylate ACP.

Also, scrolling further down, information on interaction partners of the protein is given.

g) With which protein does APC most strongly interact? Which other interacting protein do this protein and APC have in common?

   Strong interaction with CTNB1; AMER1 as common interactor.

Finally, we will have a look at the structural information given in UniProt.

h) Which secondary protein structural elements do exist (in general)?

   The most common ones are: alpha helices, beta strands/sheets (a sheet consists of multiple parallel or anti-parallel strands), and turns.

i) Which secondary structural element is most common for the APC protein, and where in the protein does is mostly occur?

   The helix is by far the most common; helices are present in the first one third of the protein.

j) Which methods have been used to experimentally determine 3D structure of (part of the) protein? What is the common range of the resolutions? Hint: check the list of structures given.

   Either X-ray crystallography or NMR spectroscopy; resolution commonly ranges from 1.5 to 3.1 Ångström (estimate based on the structures available for APC).

In the next assignment, we will study one of the experimentally identified structures, 3NMX, which covers the part of the sequence from amino acids 407 till 751.

Assignment 4: the APC protein in RCSB PDB *(20 minutes)*

▶ *Section 6.2 of the Reference guide.*

Finally, we turn to the RCSB PDB website, to find more information about mainly structure of the protein encoded by the APC gene. Most experimentally characterised protein structures are only parts of the protein under investigation. They may or may not be bound to other proteins or molecules.

We will explore the 3NMX structure in PDB; look this up by either searching for it in PDB or from the structures overview in UniProt by setting the radio button to 'RCSB PDB' and clicking the 3NMX structure. The 'structure summary' tab will open.

a) Which molecule(s) are in this structure according to its title?

   APC and Asef, to which it is bound in this structure; for both only a part is present.

b) Which method has been used to determine this structure and in which year has it been submitted? Is the resolution more or less than the length of a single carbon-carbon covalent bond (~1.4Å)?

X-ray diffraction; submitted in 2010; the resolution is 2.3 Å, so less precise than a C-C covalent bond.

Scrolling down you find information about the 'Macromolecules'. This indicates which molecules are present and in which chains (amino acid chain fragments)?

c) Which molecules and domains are present, in which chains, and how long are the fragments? Are any small molecules present?

APC, chains A, B, C; 354 amino acids; Armadillo repeats domain; present three times.

Rho guanine nucleotide exchange factor 4 (ARHGEF4); chains D, E, F; residues (amino acids) 170-194; three times.

There are no small molecules present (otherwise, these would be given as well on this page; note: water will always be present, but is not mentioned).

Open the 3D view tab.

d) Which secondary structure element does this structure mostly contain? Does this correspond to the information you had found in UniProt?

Alpha helices; yes it corresponds to UniProt, which had many helices annotated to this part of the protein.

Change the 3D display, by setting color to 'by chain' and style to 'Spacefill'.

e) How do the two fragments relate to each other?

The ARHGEF4 fragment lies embedded in a sort of groove of the APC fragment.

Open the Sequence tab. The tab tells you that the structure has in total six chains, which are represented by two unique entities. By default it only shows each unique chain once. Verify all this. Also verify the annotated secondary structure elements, corresponding to the 3D view.

Add SNP information by selecting 'Single Nucleotide Polymorphism (SNP)' from the 'Add an annotation' drop down box.

f) One of the annotated SNPs is rs137854567. What change does this SNP cause and where in the fragment does it occur? Tip: you may have to make the window a bit smaller, to make sure that the graphical display is below the SNP table, to allow you to hover over the SNP annotation track in the graphical display using your mouse.

It changes an arginine (R) residue to a cysteine (C) residue; it is at position 17 in the fragment, corresponding to position 414 in the APC protein.

g) Which types of side chain can amino acids have? What does this mean in the context of SNP rs137854567?

Generically, the side chain can be positively charged, negatively charged, polar, or apolar (hydrophobic); for this SNP, a positively charged amino acid is replaced by a non-charged amino acid.

h) What makes cysteine special?

Cysteine has a sidechain with a sulphur atom that can form a sulphur bond with another cysteine unit in the protein or in another protein.

Assignment 5: genetic variants and their consequences in other cancer-related genes *(remaining time and when preparing for the case)*

When preparing for you cases and relating to other cancer-related genes/proteins, consult NCBI Gene/dbSNP/OMIM, Ensembl, UniProt, RCSB-PDB, and other resources to find information. You can try to answer questions like (suggested databases are given in brackets):

a) In which disease is the gene involved? Or starting from a disease: which genes or genetic variants have been linked to the disease? (OMIM)

b) What is the summarised function of the protein? In which pathways or GO biological processes is the gene involved? What known interactions does the protein have? (UniProt, Ensembl, NCBI Gene)

c) What are the official name, full name, and identifier of the gene and which exonic structure and splice variants does it have? (Ensembl, NCBI Gene)

d) What can you find out about genetic variants of interest, with respect to their alleles and their frequencies in different populations, the type of coding or non-coding SNP, the expected impact of the change, the known connections to (disease) phenotypes? (Ensembl, NCBI dbSNP)

e) What structural elements does the protein have? How does the 3D structure of (domains of) the protein look like? (RCSB-PDB, UniProt)

f) Also, you can consult additional sources from computer lab 2A, such as metabolite or drug databases. *To illustrate this with an example, find for yourself that when searching for FAP (as an indication), DrugBank gives two drugs: Celecoxib and Sulindac. You could find out more about their mode-of-action by further consulting this data resource.*