The Ensembl and UCSC genome browsers

Antoine van Kampen
Bioinformatics Laboratory
a.h.vankampen@amc.uva.nl
www.bioinformaticslaboratory.nl

Introduction

Genome Browsers contain the *reference sequence* (or working draft assembly) for one or more genomes. Basically, these browsers integrate *annotation* of these genomes such as gene structure, regulatory elements, CpG islands, repeat elements. Genome browser allow you to visualize, explore and use all this information.

In this tutorial you will be introduced to two main genome browsers: **Ensembl** of the European Bioinformatics Institute (EBI; Cunningham, 2015) and **UCSC** of the Genome Informatics Group of the University of California, Santa Cruz (Rosenbloom, 2015).

- Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kahari AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SM, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P (2015) Ensembl 2015. Nucleic acids research 43 (Database issue):D662-669.
- Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ (2015) The UCSC Genome Browser database: 2015 update. Nucleic acids research 43 (Database issue):D670-681.

This tutorial consists of three parts.

- **Part I** of the tutorial will teach you how to use Ensembl and to allow you to explore the information that is provided for the Human genome.
- In **part II** of the tutorial you will be introduced to The Cancer Genome Atlas (TCGA) data repository, which we will use to obtain gene expression (RNAseq) and methylation (Illumina HumanMethylation450 BeadChip) data measured in breast tumors.
- In **part III** we will integrate and visualize this omics data in the UCSC genome browser (note: although this can also be done in Ensembl, UCSC is more convenient and, more importantly, the genomic coordinates of the BeadChip probes are not in agreement with the coordinates used by Ensembl). In addition, the location of the CpG island in Ensembl is different from that in UCSC.

Note

Many public databases are continuously updated and improved. Not only the content, but also the user/web interface. If it turns out that the instructions below are not entirely valid for the latest database version you are going to use, you should still be able to continue the tutorial with some effort from your site (like in real practice!).

Have fun…….

Part I. Ensembl tutorial

Acknowledgements
The exercises in this document are taken and based on the 'Walk_Through_Browser_E75' Ensembl tutorial from the European Bioinformatics Institute (EBI; https://www.ebi.ac.uk), which is no longer available.

Why Ensembl?
Ensembl is one of the major genome databases. It provides, among others, the human DNA sequence including a comprehensive annotation. In this tutorial you will learn how to use Ensembl and discover the types of annotation that are provided.

Important note: this Ensembl tutorial probably takes more than 1 hour to complete but you will get the basic idea after using the browser for an hour. For the second hour continue with Part II and Part III of this tutorial.

1.1 Mutations in the ABCD1 can cause adrenoleukodystrophy

In this worked example we will explore the human ABCD1

Question 1: what is the function of the ABCD1 gene? On which chromosome is ABCD1 located? (Hint: use the GeneCards database; www.genecards.org)

Question 2: what is adrenoleukodystrophy? (Hint: use the MalaCards database; www.malacards.org)

Answer
ABCD1 (ATP-binding cassette, sub-family D (ALD), member 1) gene. The protein encoded by this gene is likely involved in the peroxisomal transport or catabolism of very long chain fatty acids (VLCFAs). Mutations in the *ABCD1* gene can cause **adrenoleukodystrophy**, a rare X-linked disorder that causes a range of clinical phenotypes, often leading to a vegetative state and/or death.

1.2 Ensemble search

Go to the Ensembl homepage (**http://www.ensembl.org/**).

First of all, we have to search for the human *ABCD1* gene.

Select 'Search: Human' and type 'abcd1' in the 'for' text box.

The search result shows the *ABCD1* gene on the top of the list, named ABCD1 (Human Gene), which has been annotated on the reference assembly (X:153724856-153744755).

Click on 'ABCD1 (Human Gene)' on the page with search results.

This leads us to the 'Gene summary' page under the 'Gene' tab.

1.2.1 The Gene tab

'Gene-based displays' are listed in the left-hand side menu.

The 'Gene Summary' page shows general information about the *ABCD1* gene and the transcripts that have been annotated for it as part of the GENCODE gene set (https://www.gencodegenes.org/). Note the information icon (*?*) right to 'Gene summary' opens up a help page.

Question 3: What is the GENCODE project?

Answer

They identify all functional elements in the human genes

Click [Show transcript table]. (if it is not already displayed)

In general, clicking or hovering your mouse over any feature that is shown on a graphical display in Ensembl should result in a pop-up. The pop-up typically contains some basic information about the feature in question and often also one or more links to pages where more detailed information can be found.

Boxes and lines in the transcripts in the graphic display represent exons and introns, respectively. Empty boxes represent untranslated regions (UTRs), while filled boxes represent the coding sequence (CDS).

Click in the graphical display on the three transcripts that have been annotated for the *ABCD1* gene.

The pop-ups show that one of the transcripts (ABCD1-201 / ENST00000218104) is an Ensembl/Havana merge transcript (shown in gold), while the other two only have been annotated by Havana (shown in red,blue).

Question 4: What is a HAVANA transcript?

Answer

The HAVANA group provides the manual annotation of human, mouse, zebrafish and other vertebrate genomes that appear in the Vega browser.

In the transcript table you see that ENST00000218104 is also part of the CCDS (Consensus Coding Sequence;

Question 5: What is the Consensus Coding Sequence (CCDS)?

Answer

https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi) set, which is a set of coding sequences about which Ensembl, Havana, NCBI and UCSC all agree and which therefore are considered to be of high quality.

It is also shown that two transcripts are part of the GENCODE 'Basic Set', that is intended to provide a simplified subset of the GENCODE transcript annotations that will be useful to the majority of users. Transcript ABCD1-302 / ENST00000443684 is not part of this set, because its CDS is incomplete.

The *ABCD1* gene is located on the forward strand of the genome. This can be seen from the arrows next to the transcript names that indicate the direction of transcription and from the fact that the transcript models are shown above the blue bar that represents the genome. Transcripts located on the reverse strand are shown below the blue bar.

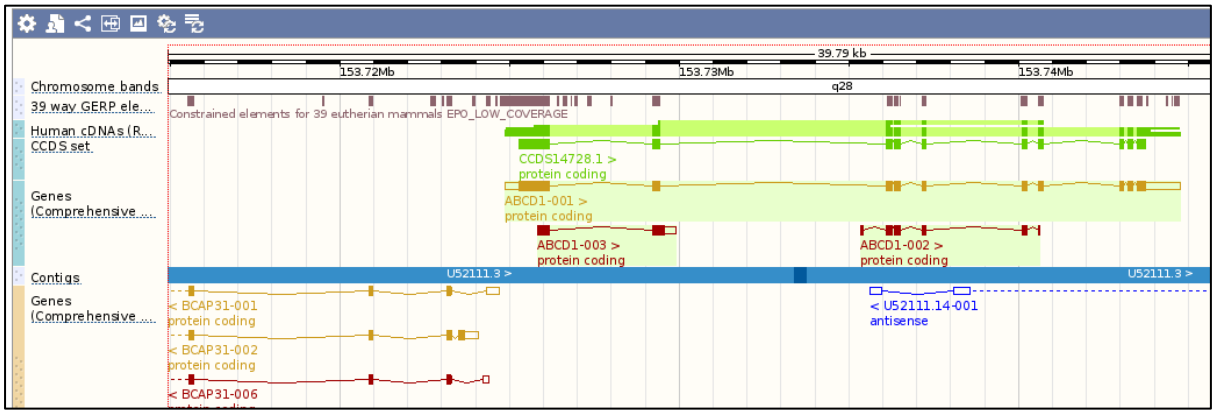
1.2.1.1 Sequences and variants

Click on 'Sequence' in the side menu.

On the 'Sequence' page the sequence of the *ABCD1* gene plus 600 bp upstream and downstream is shown. Note that, because the *ABCD1* gene has multiple transcripts, this page doesn't show the exon-intron structures for the individual transcripts. These can be seen on the 'Exons' pages for the respective transcripts.

Question 6: To what transcript(s) do the exons belong that are shown in black letters on a peach background?

Answer: to the antisense transcript U52111

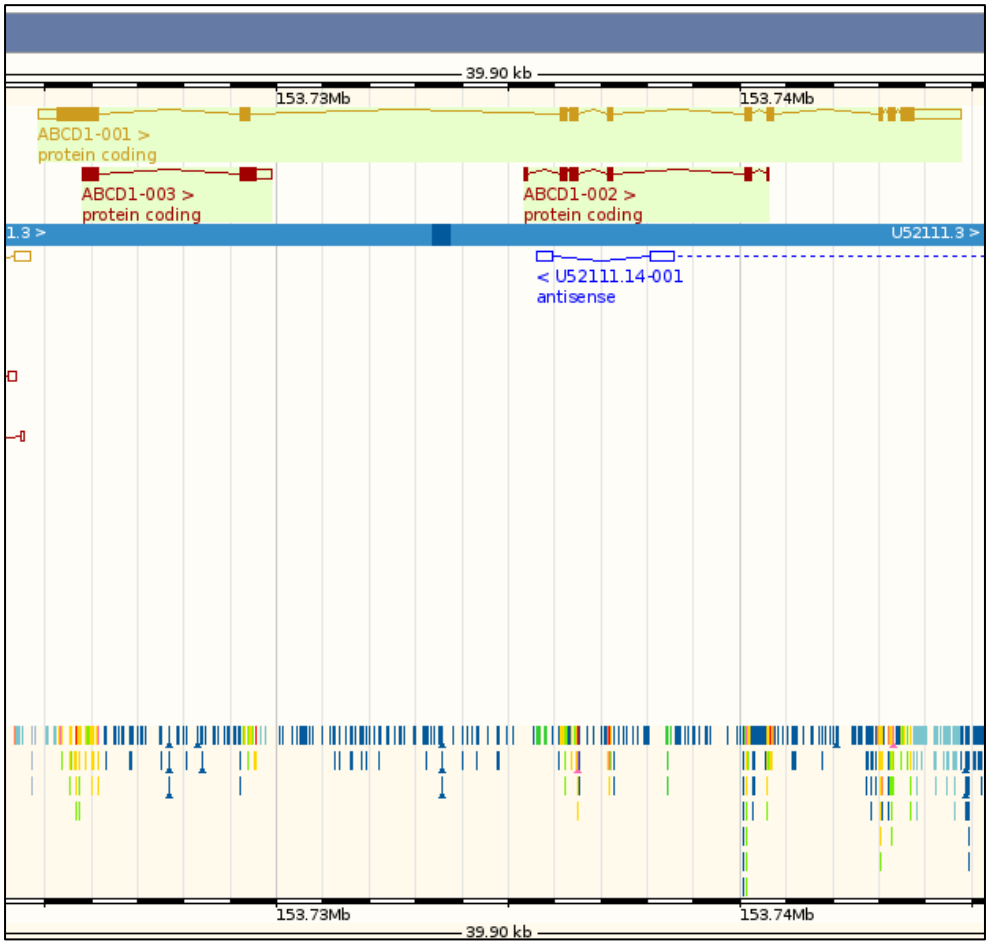


<<<<<<< END ANSWER

Almost all graphical displays in Ensembl can be configured. This is always done using the [Configure this page] button.

Go back to the previous track view (thus disregarding the sequencing).
Click [Configure this page] in the side menu.
Add a track to show the dbSNP variants, and inspect few of the variants.

Answer



<<<<<< END OF ANSWER

Now configure the Ensembl browser to show the variations in the sequence.

The 'Sequence' page has now been reloaded and shows the position of variants as well as links to pages where more detailed information about them can be found. The variants are shown in the sequence in IUPAC notation.

Question 7: What is IUPAC notation?

Answer

(http://en.wikipedia.org/wiki/Nucleic_acid_notation).

This universally accepted notation uses the Roman characters G, C, A, and T, to represent the four nucleotides commonly found in deoxyribonucleic acids (DNA). Given the rapidly expanding role for genetic sequencing, synthesis, and analysis in biology, researchers have been compelled to develop alternate notations to further support the analysis and manipulation of genetic data. These notations generally exploit size, shape, and symmetry to accomplish these objectives.

Symbol[2]	Description	Bases represented				
A	Adenine	A				1
C	Cytosine		C			
G	Guanine			G		
T	Thymine				T	
U	Uracil				U	
W	Weak	A			T	2
S	Strong		C	G		
M	aMino	A	C			
K	Keto			G	T	
R	puRine	A		G		
Y	pYrimidine		C		T	
B	not A (**B** comes after A)		C	G	T	3
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	
V	not T (**V** comes after T and U)	A	C	G		
N or **-**	No idea (not a gap)	A	C	G	T	4

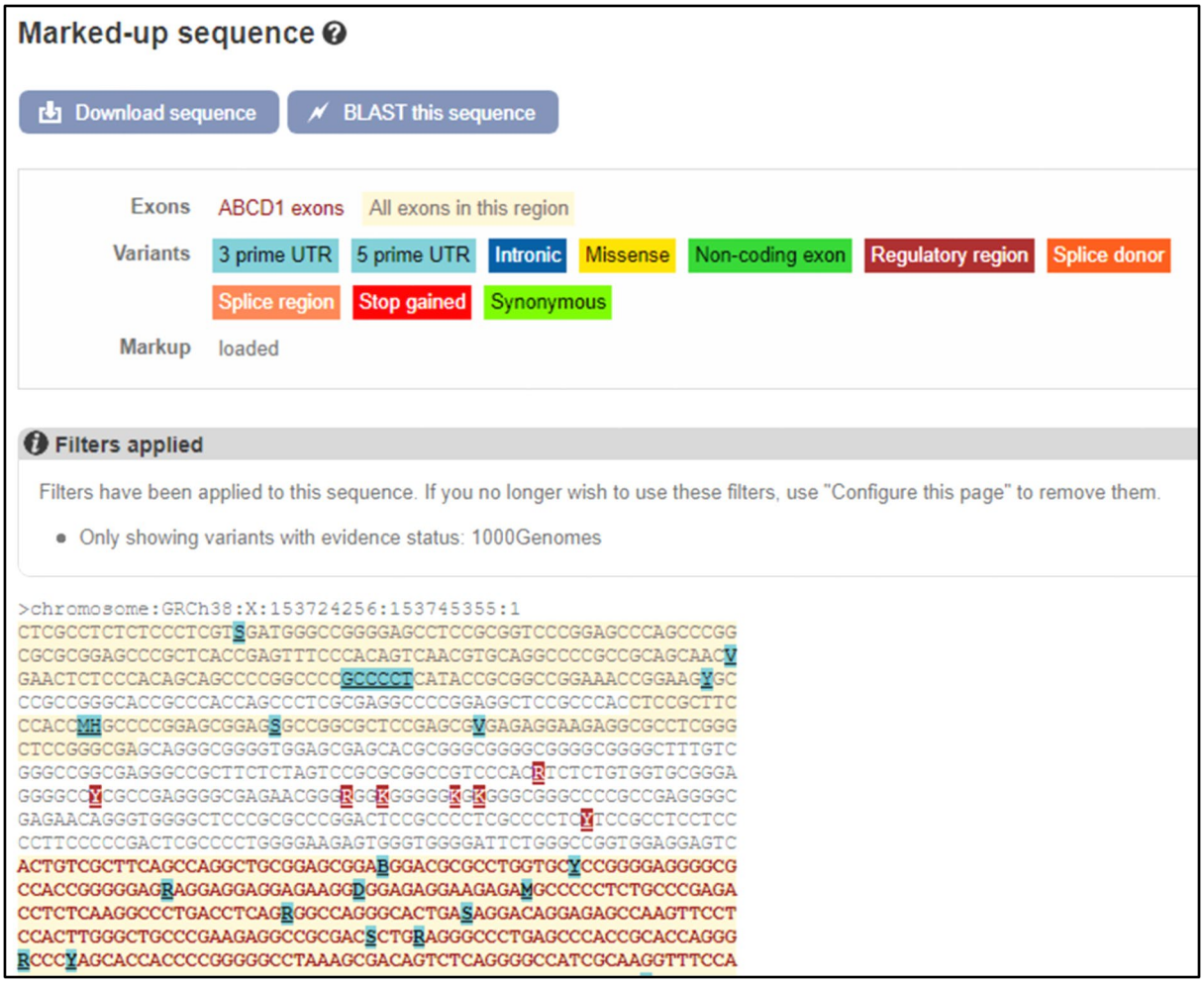
(Table from Wikipedia)

<<<<<<<END ANSWER

Question 8: What is the first mutation (1000 Genomes track) you observe in the first exon of the ABCD1 gene?

Answer

The sequence shows a 'B'. This represents a C,G or T. The associated variant identifier is rs=781857838. Retrieving information about this variant shows that C (wild type) is mutated in G or T.



<<<<<< END OF ANSWER

1.2.1.2 Orthologues

Question 9: what is an orthologues gene?

Answer

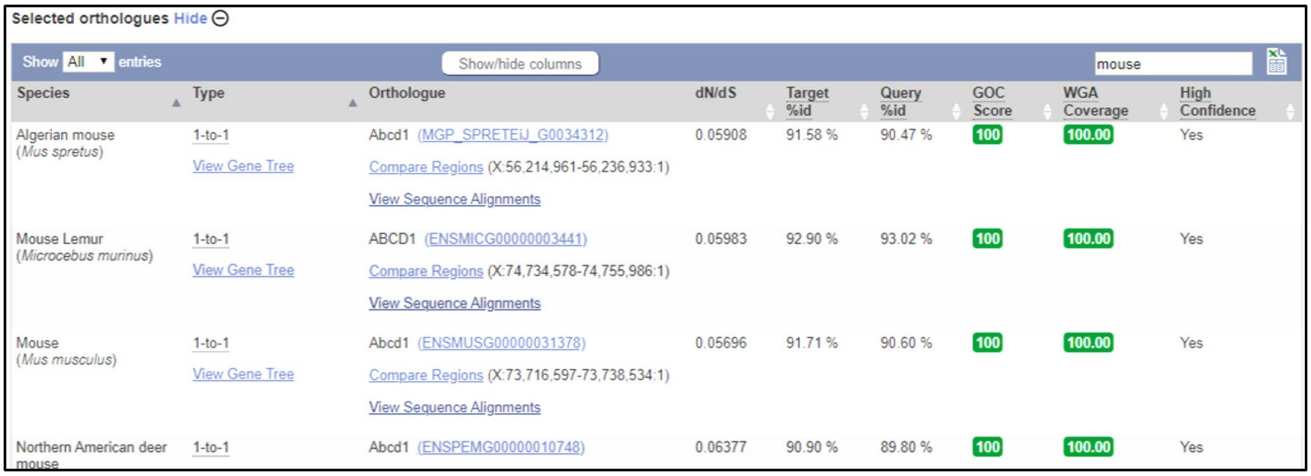
Homologous (i.e. evolutionary related) sequences are orthologous if they are inferred to be descended from the same ancestral sequence separated by a speciation event: when a species diverges into two separate species, the copies of a single gene in the two resulting species are said to be orthologous. Orthologs, or orthologous genes, are genes in different species that originated by vertical descent from a single gene of the last common ancestor.

Click on 'Comparative Genomics- Orthologues' in the side menu.

All *ABCD1* orthologues in other species as identified by Ensembl are shown in the 'Selected orthologues' table on the 'Orthologues' page. Table columns can be hidden using the [Show/hide columns] button. Rows can be (re)ordered using the triangles next to the column headers and filtered using the 'Filter' text box. The table can be exported as an Excel spreadsheet by clicking on the Excel icon.

Type ‘mouse’ in the ‘Filter’ text box.

This results in a table that only shows those rows that contain the term ‘mouse’.


Selected orthologues Hide

Show All entries | Show/hide columns | mouse

Species	Type	Orthologue	dN/dS	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Algerian mouse (*Mus spretus*)	1-to-1	Abcd1 (MGP_SPRETEiJ_G0034312)	0.05908	91.58 %	90.47 %	100	100.00	Yes
	View Gene Tree	Compare Regions (X:56,214,961-56,236,933:1)						
		View Sequence Alignments						
Mouse Lemur (*Microcebus murinus*)	1-to-1	ABCD1 (ENSMICG00000003441)	0.05983	92.90 %	93.02 %	100	100.00	Yes
	View Gene Tree	Compare Regions (X:74,734,578-74,755,986:1)						
		View Sequence Alignments						
Mouse (*Mus musculus*)	1-to-1	Abcd1 (ENSMUSG00000031378)	0.05696	91.71 %	90.60 %	100	100.00	Yes
	View Gene Tree	Compare Regions (X:73,716,597-73,738,534:1)						
		View Sequence Alignments						
Northern American deer mouse	1-to-1	Abcd1 (ENSPEMG00000010748)	0.06377	90.90 %	89.80 %	100	100.00	Yes

For the human *ABCD1* gene, a 1-to-1 orthologue has been identified in mouse, i.e. ENSMUSG00000031378.

1.2.1.3 Other information

Click on External References in the side menu.

This shows matches to the Ensembl gene in other projects and databases. A table that links Ensembl transcripts to UniProt and RefSeq identifiers is found at the bottom of the page.

Click on ‘Phenotypes’ in the side menu.

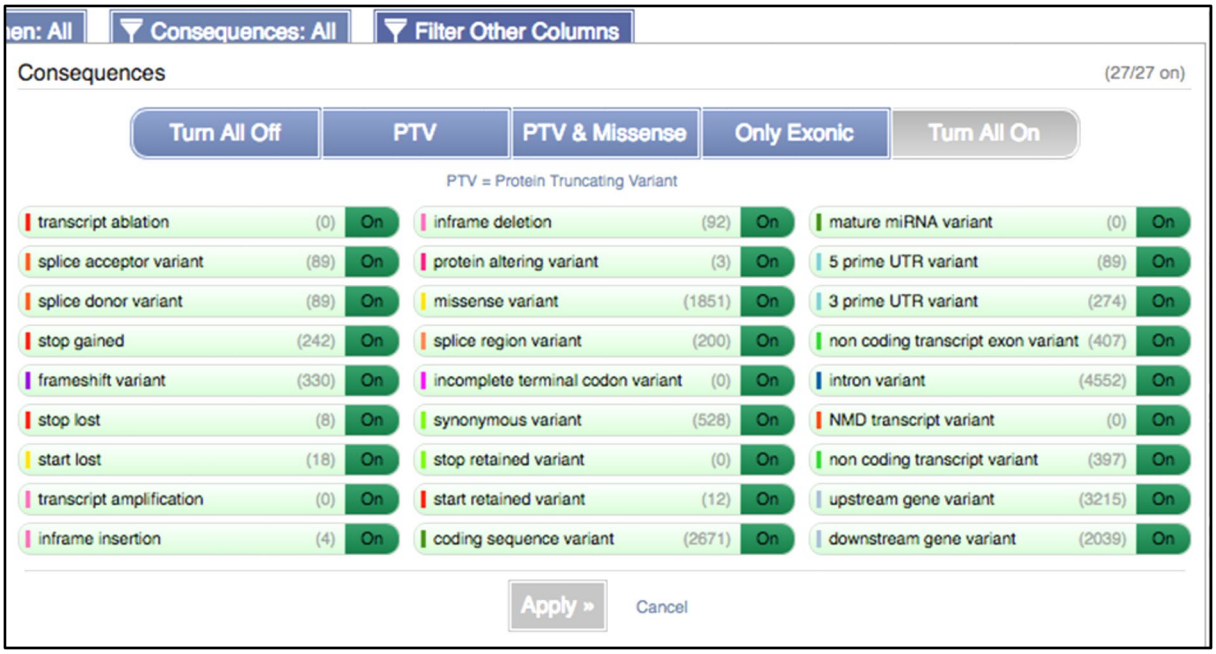
On the ‘Phenotypes’ page phenotypes that have been associated with the *ABCD1* gene as well as with variants associated with the *ABCD1* gene are shown.

Click on ‘Genetic Variation - Variation table’ in the side menu.

On the ‘Variation table’ page all variants in *ABCD1* gene are shown, grouped by consequence type.

Answer

(Click on the Consequences filter button and you get the summary:

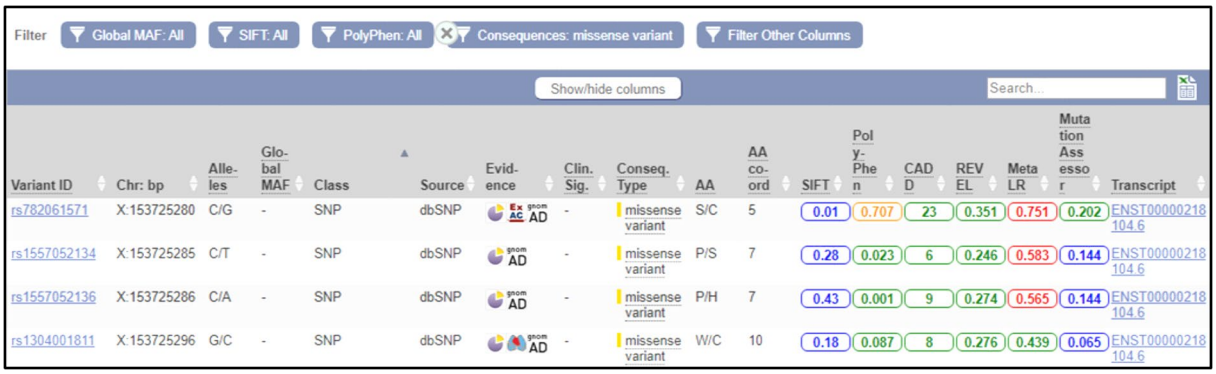


<<<<<<< END ANSWER

Turn of all and then turn on 'Missense variant' and click 'Apply'.

This results in a list of all missense variants. Clicking on the ID of a variant will lead us to the 'Variation' tab, where more information about the variant in question can be found.

Answer



<<<<<<< END ANSWER

<<<<<<< END ANSWER

1.2.2 The Transcript tab

Click on 'ENST00000218104.6' in the transcript table at the top of the page.

This leads us to the 'Transcript summary' page under the 'Transcript' tab.

Note that, because we have moved from the 'Gene' tab to the 'Transcript' tab, the side menu has changed and now shows links to pages with transcript-specific information.

1.2.2.1 Information about exons

Click on 'Sequence - Exons' in the side menu.

On the 'Exons' page the sequence of the unspliced transcript is shown. The coding sequence (CDS) is shown in blue, untranslated regions (UTRs) in red, introns in gray and flanking sequences in green. By default only a small part of the introns and the flanking sequences is shown, but this can be changed on the configuration page.

Show All entries | Show/hide columns | Filter

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						.....actcgcccctggggaagagtgggtggggattctgggccggtggaggagtc
1	ENSE00000868271	153,724,856	153,726,166	-	0	1,311	ACTGTCGCTTCAGCCAGGCTGCGGAGCGGACGGACGCGCCTGGTGCCCCGGGGAGGGGCG CCACCGGGGGAGGAGGAGGAGGAAGGTGGAGAGGAAGAGACGCCCCCTCTGCCCGAGA CCTCTCAAGGCCCTGACCTCAGGGGCCAGGGCACTGACAGGACAGGAGAGCCAAGTTCCT CCACTTGGGCTGCCCGAAGAGGCCGCGACCCTGGAGGGCCCTGAGCCCACCGCACCAGGG GCCCCAGCACCACCCCGGGGGCCTAAAGCGACAGTCTCAGGGGCCATCGCAAGGTTTCCA GTTGCCTAGACAACAGGCCCAGGGTCAGAGCAACAATCCTTCCAGCCACCTGCCTCAACT GCTGCCCCAGGCACCAGCCCCAGTCCCTACGCGGCAGCCAGCCCAGGTGACATGCCGGTG CTCTCCAGGCCCCGGCCCTGGCGGGGGAACACGCTGAAGCGCACGGCCGTGCTCCTGGCC CTCGCGGCCTATGGAGCCCACAAAGTCTACCCCTTGGTGCGCCAGTGCCTGGCCCCGGCC AGGGGTCTTCAGGCGCCCGCCGGGGAGCCCACGCAGGAGGCCTCCGGGGTCGCGGCGGCC AAAGCTGGCATGAACCGGGTATTCCTGCAGCGGCTCCTGTGGCTCCTGCGGCTGCTGTTC CCCCGGGTCCTGTGCCGGGAGACGGGGCTGCTGGCCCTGCACTCGGCCGCCTTGGTGAGC CGCACCTTCCTGTCGGTGTATGTGGCCCGCCTGGACGGAAGGCTGGCCCGCTGCATCGTC CGCAAGGACCCGCGGGCTTTTGGCTGGCAGCTGCTGCAGTGGCTCCTCATCGCCCTCCCT GCTACCTTCGTCAACAGTGCCATCCGTTACCTGGAGGGCCAACTGGCCCTGTCGTTCCGC AGCCGTCTGGTGGCCCACGCCTACCGCCTCTACTTCTCCCAGCAGACCTACTACCGGGTC AGCAACATGGACGGGCGGCTTCGCAACCCTGACCAGTCTCTGACGGAGGACGTGGTGGCC TTTGCGGCCTCTGTGGCCCACCTCTACTCCAACCTGACCAAGCCACTCCTGGACGTGGCT GTGACTTCCTACACCCTGCTTCGGGCGGCCCGCTCCCGTGGAGCCGGCACAGCCTGGCCC TCGGCCATCGCCGGCCTCGTGGTGTTCCTCACGGCCAACGTGCTGCGGGCCTTCTCGCCC AAGTTCGGGGAGCTGGTGGCAGAGGAGGCGCGGCGGAAGGGGGAGCTGCGCTACATGCAC TCGCGTGTGGTGGCCAACTCGGAGGAGATCGCCTTCTATGGGGGCCATGAG
	Intron 1-2	153,726,167	153,729,231			3,065	gtggggcaggttgggggtgccgggca.....tctctgtgtctgtcaccccccgcag
2	ENSE00000677936	153,729,232	153,729,412	0	1	181	GTGGAGCTGGCCCTGCTACAGCGCTCCTACCAGGACCTGGCCTCGCAGATCAACCTCATC CTTCTGGAACGCCTGTGGTATGTTATGCTGGAGCAGTTCCTCATGAAGTATGTGTGGAGC GCCTCGGGCCTGCTCATGGTGGCTGTCCCCATCATCACTGCCACTGGCTACTCAGAGTCA G
	Intron 2-3	153,729,413	153,736,111			6,699	gtgagacccagggctccaagaggat.....cacccacccctctggggctttgcag

<<<<<<< END ANSWER

1.2.2.2 Information about cDNA

Click on 'Sequence - cDNA' in the side menu.

The spliced transcript, as well as its coding sequence and translation are shown with sequence variants drawn. The page can be simplified with 'Configure this page'.

Answer

<<<<<<< END ANSWER

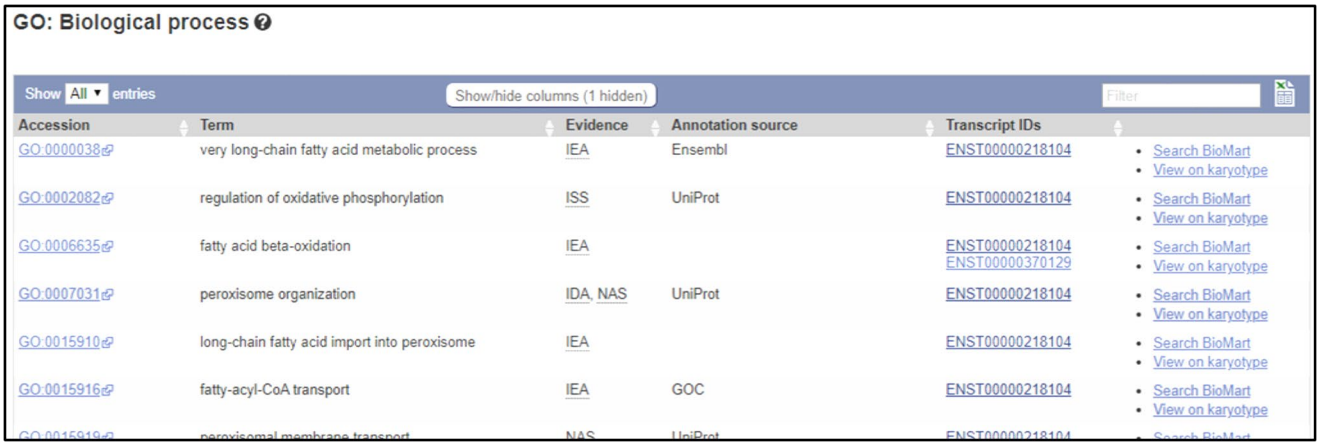
1.2.2.3 Information from Gene Ontology

Click on 'Ontologies – GO:Biological process' in the side menu of the 'Gene' tab.

Question 10: What is the Gene Ontology (GO)?

Answer:

Gene Ontology (GO) terms (http://www.geneontology.org) are used to annotate gene products. Biological process, molecular function and cellular component terms describe what a gene product does, how it does that and where it does that, respectively.



<<<<<<< END ANSWER

The Biological process terms indicate that the ABCD1 protein plays a role in fatty acid transport and catabolism. The Cellular component terms indicate the ABCD1 protein is located in the peroxisomal membrane.

1.2.3 The Location tab

Click on the 'Location' tab.

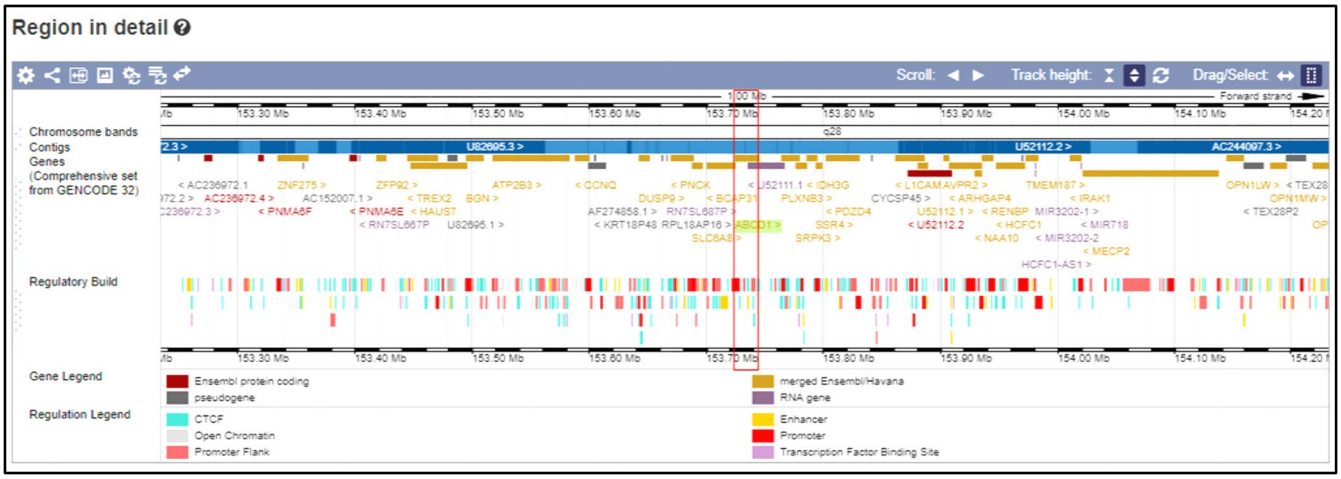
This leads us to the 'Region in detail' page under the 'Location' tab.

The 'Region in detail' page shows the genomic neighbourhood of the *ABCD1* gene. It consists of three parts.

First, the complete chromosome.

Second, the 1 Mb region around the region of interest.

This display is scrollable, by either using the 'Scroll' arrows or by clicking and dragging the image in the same way as e.g. in Google Maps. Zooming into a region is possible by clicking the 'Drag/Select' icon, selecting the region of interest and subsequently clicking 'Jump to region' in the resulting pop-up.



Third, the region of interest. In our case this is the *ABCD1* gene.

There are several ways to navigate this display:

- zooming in and out by using the [+/-] slider

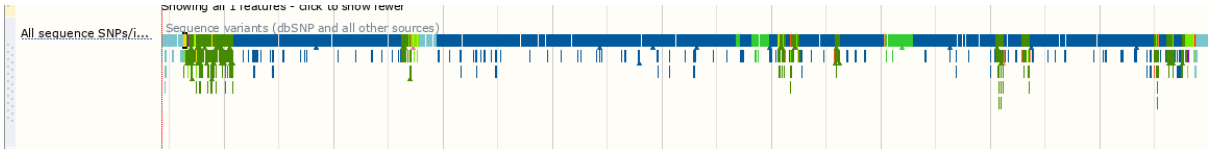
- zooming in by drawing a box around the region of interest and subsequently clicking ‘Jump to region’ in the resulting pop-up
- moving up- and downstream with the single and double arrows next to the [+/-] slider.
- going to a particular region by changing the coordinates in the ‘Location’ text box or by searching for a gene using the ‘Gene’ text box (which has auto completion)

Sets of data are called ‘tracks’. These can be added to the display using [Configure this page]. On the configuration page all available tracks are grouped in the left-hand menu. It is also possible to search for tracks using the ‘Find a track’ text box.

For example, to add all sequence variants to the display:

Click [Configure this page] in the side menu.
Type ‘sequence variants’ in the ‘Find a track’ text box.
Select ‘Sequence variants (dbSNP and all other sources) - Normal’.
Click (✓).

A new track, ‘Sequence variants (dbSNP and all other sources)’, has now been added to the display.



Tracks can be moved by clicking on the bar in front of the track name and dragging the track to the desired location.

At the top and bottom of the display several icons are shown, some of which can also be found on other displays:

- Configure this image: add/delete tracks (same as [Configure this page] button in the side menu).
- Manage your custom tracks: add your own data (same as [Add your data] button in the side menu)
- Share this image: create a URL that can be shared with others without the need to tell them how to configure the page
- Resize this image: resize the image
- Export this image: export the image in various formats (PDF, PNG etc.)

1.3 Sequence search

Finally, we will do a sequence search. For this we need some sequence. To this end, we will take the sequence of the *ABCD1* gene (so, in fact we are cheating as we of course already know what this sequence represents).

Click [Export data] in the side menu.
Click [Next>].
Click on 'Text'.

This gives us the sequence of the *ABCD1* gene in FASTA format.

```
>X dna:chromosome chromosome:GRCh38:X:153724856:153744755:1
ACTGTCGCTTCAGCCAGGCTGCGGAGCGGACGGACGCGCCTGGTGCCCCGGGGAGGGGCG
CCACCGGGGGAGGAGGAGGAGGAGAAGGTGGAGAGGAAGAGACGCCCCCTCTGCCCGAGA
CCTCTCAAGGCCCTGACCTCAGGGGCCAGGGCACTGACAGGACAGGAGAGCCAAGTTCCT
CCACTTGGGCTGCCCGAAGAGGCCGCGACCCTGGAGGGCCCTGAGCCCACCGCACCAGGG
GCCCCAGCACCACCCCGGGGGCCTAAAGCGACAGTCTCAGGGGCCATCGCAAGGTTTCCA
GTTGCCTAGACAACAGGCCCAGGGTCAGAGCAACAATCCTTCCAGCCACCTGCCTCAACT
GCTGCCCCAGGCACCAGCCCCAGTCCCTACGCGGCAGCCAGCCCAGGTGACATGCCGGTG
CTCTCCAGGCCCCGGCCCTGGCGGGGGAACACGCTGAAGCGCACGGCCGTGCTCCTGGCC
CTCGCGGCCTATGGAGCCCACAAAGTCTACCCCTTGGTGCGCCAGTGCCTGGCCCCGGCC
AGGGGTCTTCAGGCGCCCGCCGGGGAGCCCACGCAGGAGGCCTCCGGGGTCGCGGCGGCC
AAAGCTGGCATGAACCGGGTATTCCTGCAGCGGCTCCTGTGGCTCCTGCGGCTGCTGTTC
CCCCGGGTCCTGTGCCGGGAGACGGGGCTGCTGGCCCTGCACTCGGCCGCCTTGGTGAGC
CGCACCTTCCTGTCGGTGTATGTGGCCCGCCTGGACGGAAGGCTGGCCCGCTGCATCGTC
CGCAAGGACCCGCGGGCTTTTGGCTGGCAGCTGCTGCAGTGGCTCCTCATCGCCCTCCCT
GCTACCTTCGTCAACAGTGCCATCCGTTACCTGGAGGGCCAACTGGCCCTGTCGTTCCGC
```

Subsequently we do a sequence search with this sequence.

Select and copy the sequence.
Go back to the browser.
Click on the 'BLAST/BLAT' link on the toolbar.
Paste the sequence in the text box at the top of the page.
Select 'Homo_sapiens' as the species to search against.
Select 'BLAT' as the search tool.

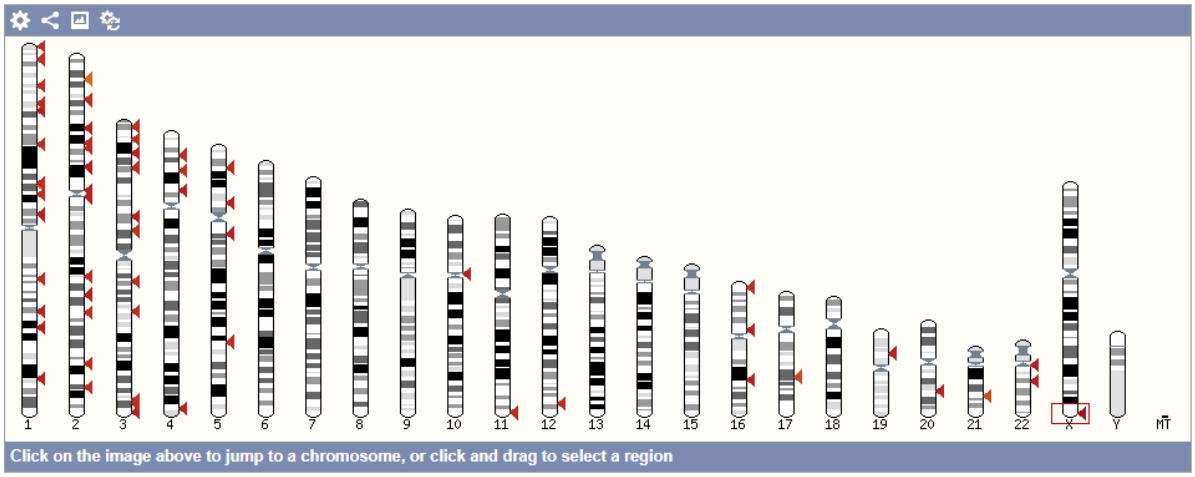
We use BLAT because we are expecting an exact match (as a rule of thumb use BLAT when you are searching against the same species your query sequence is from and use BLAST when you are searching against another species).

Click [RUN>], wait for the job to end and click "View results"

The results page consists of three parts.

The second part ("*HSP distribution on genome*") shows all hits in relation to the genome. Hits are shown as red triangles, the best hit has a box around it. In our case the best hit is located on the end of the X chromosome.

HSP distribution on genome



The first part ("*Results table*") shows all hits in tabular format. In our case there is only one hit that is 100% identical to the genome sequence. This also happens to be the longest hit. Note the E (expect) value, a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.

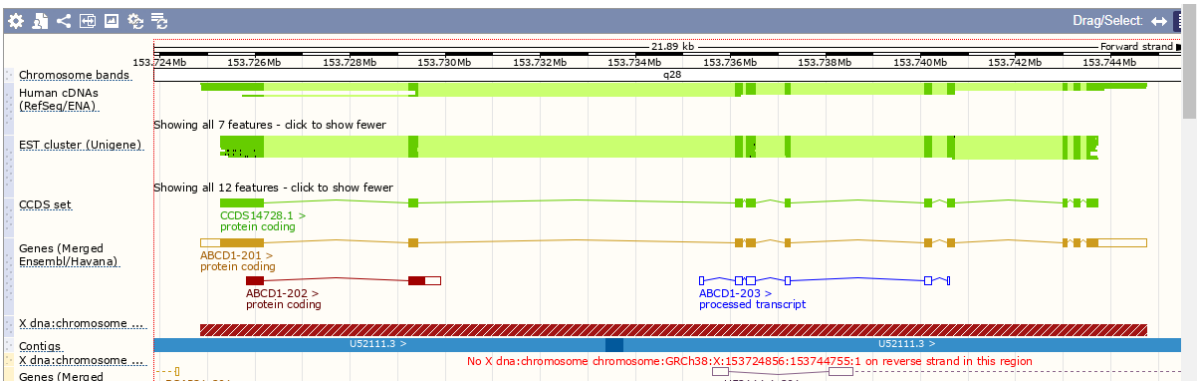
Results table

Show All entries | Show/hide columns (2 hidden) | Filter

Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val	%ID
X:153724856-153744755 [Sequence]	ABCD1, U52111.1	Forward	1	19900	19900 [Sequence]	39266.0	0.0e+00	100.00 [Alignment]
2:91840951-91844160 [Sequence]	ABCD1P5	Forward	16030	19233	3211 [Sequence]	5903.0	0.0e+00	95.42 [Alignment]
22:16388033-16391239 [Sequence]	ABCD1P4	Forward	16030	19233	3212 [Sequence]	5832.0	0.0e+00	94.99 [Alignment]
2:89764321-89767024 [Sequence]		Reverse	16030	18728	2707 [Sequence]	4784.0	0.0e+00	93.76 [Alignment]
KI270728.1:351687-354065 [Sequence]		Forward	16029	18403	2379 [Sequence]	4354.0	0.0e+00	95.00 [Alignment]
KI270728.1:624845-627223 [Sequence]		Forward	16029	18403	2379 [Sequence]	4354.0	0.0e+00	95.00 [Alignment]
16:32475534-32477911 [Sequence]	ABCD1P3	Reverse	16030	18403	2378 [Sequence]	4349.0	0.0e+00	94.95 [Alignment]
10:38603169-38604968 [Sequence]	ABCD1P2	Forward	17436	19233	1801 [Sequence]	3286.0	0.0e+00	95.06 [Alignment]
10:38601760-38603167 [Sequence]	ABCD1P2	Forward	16030	17433	1408 [Sequence]	2539.0	0.0e+00	94.46 [Alignment]
KI270728.1:354078-354743 [Sequence]		Forward	18413	19075	666 [Sequence]	1228.0	0.0e+00	95.35 [Alignment]
KI270728.1:627236-627901 [Sequence]		Forward	18413	19075	666 [Sequence]	1228.0	0.0e+00	95.35 [Alignment]
16:32474856-32475521 [Sequence]	ABCD1P3	Reverse	18413	19075	666 [Sequence]	1219.0	0.0e+00	94.89 [Alignment]

Click on the genomic location of the best hit.

This leads us back to the 'Region in detail' page, to which now a new track named 'BLAT/BLAST hits' has been added, that shows the best hit in bright red.



Part II. The Cancer Genome Atlas (TCGA).

1.4 Introduction

The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. The overarching goal of TCGA is to improve our ability to diagnose, treat and prevent cancer.

1.5 TCGA Primary Identifiers (barcodes)

Historically, the Biospecimen Core Resource (BCR) received participant samples and their associated metadata from a Tissue Source Site (TSS). The BCR then assigned human-readable IDs, referred to as **TCGA barcodes**, representing the metadata of the participants and their samples.

TCGA barcodes were used to tie together data that spans the TCGA network, since the IDs uniquely identify a set of results for a particular sample produced by a particular data-generating center.

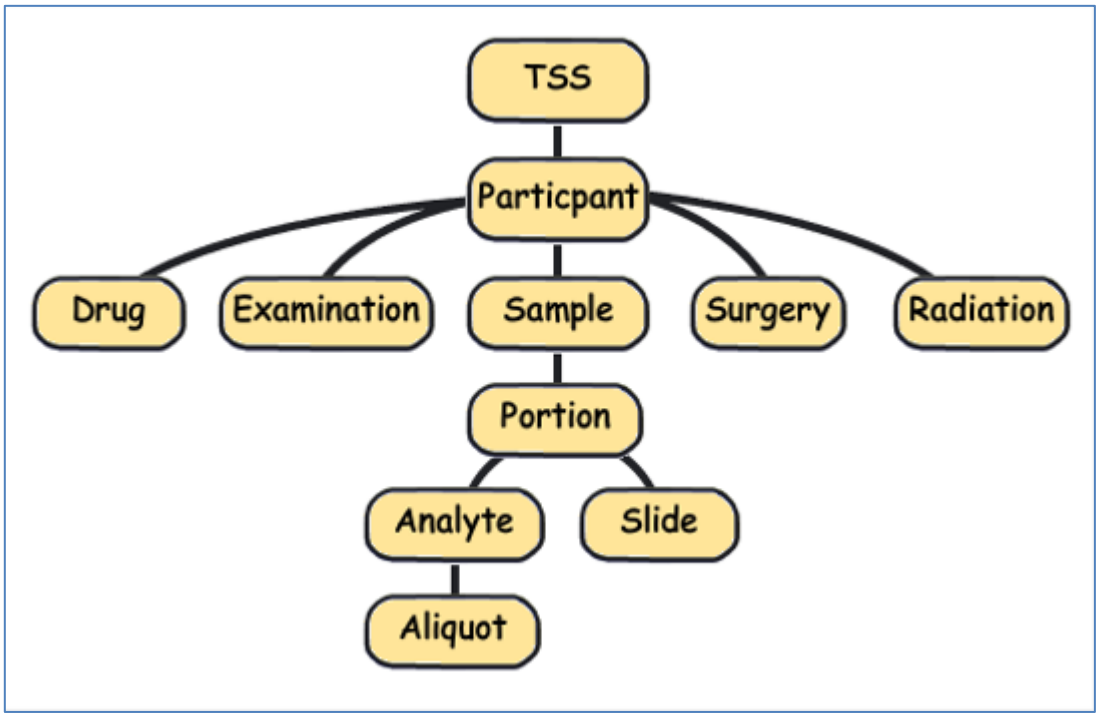


Figure. Hierarchy of biospecimen elements. Barcodes (see next figure) are used to represent all biospecimen elements in this diagram

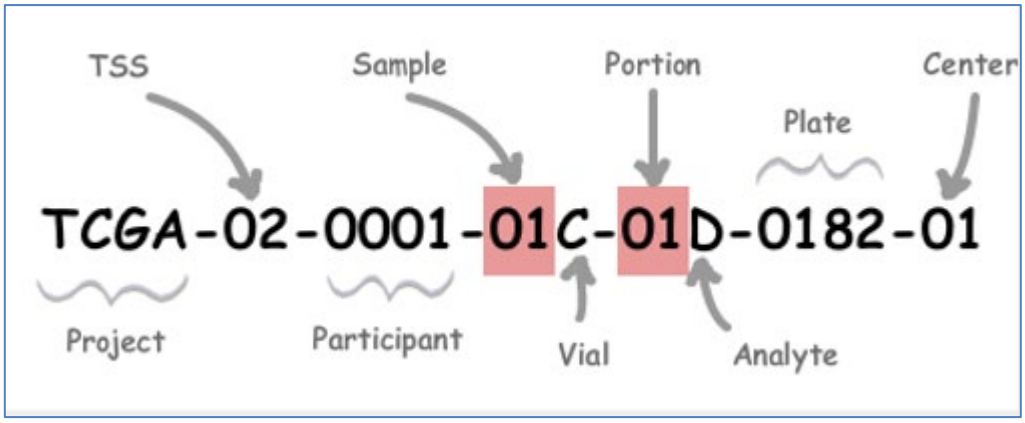


Figure. This figure of an aliquot barcode shows how it can be broken down into its components and translated into its metadata.

1.6 TCGA Data Access

TCGA Data can be accessed in various ways:

- Data Portal:
  - https://gdc.cancer.gov/
  - Data Matrix
  - Bulk Download
- **Using R**
  - **TCGA-assembler** (Zhu, Qiu and Ji (2014), Nature Methods. 11(6):599-600.) http://www.compgenome.org/
  - **TCGABiolinks** (Colaprico et al. (2015). *Nucleic Acids Research*. 44(8):e71)

In this part of the tutorial we have already downloaded files for you using the **TCGA-assembler**.

1.7 Case study: Methylation of the Claudin 1 Promotor

In a recent paper (Di Cello et al (2013) PLoS One. 8(7):e68630; available from course website) the relation between gene expression of Claudin 1 (CLDN1) and methylation of its promotor in breast cancer was investigated.

Downregulation of the tight junction protein claudin 1 (CLDN1) is a frequent event in breast cancer and is associated with recurrence, metastasis, and reduced survival, suggesting a tumor suppressor role for this protein. Tumor suppressor genes are often epigenetically silenced in cancer. Downregulation of claudin 1 via DNA promoter methylation may thus be an important determinant in breast cancer development and progression. To investigate if silencing of claudin 1 has an epigenetic etiology in breast cancer the authors compared gene expression and methylation data from 217 breast cancer samples and 40 matched normal samples available through the Cancer Genome Atlas (TCGA).

They found that methylation of the claudin 1 promoter CpG island is relatively frequent in estrogen receptor positive (ER+) breast cancer and is associated with low claudin 1 expression. In contrast, the claudin 1 promoter was not methylated in most of the ER- breast cancers samples and some of these tumors overexpress claudin 1. Their results indicate that DNA promoter methylation is causally associated with downregulation of claudin 1 in a subgroup of breast cancer that includes mostly ER+ tumors, and suggest that epigenetic therapy to restore claudin 1 expression might represent a viable therapeutic strategy in this subtype of breast cancer.

We will not reproduce their entire analysis but show how you can retrieve the data from the TCGA database and visualize the methylation data in the UCSC genome browser.

We will download clinical, gene expression (RNAseq) and methylation (illumina HumanMethylation450 BeadChip) data for an ER positive and an ER negative patient. In Part III of this tutorial we will visualize the methylation data in the UCSC genome browser to explore if the data of the selected patients is in agreement with their conclusion.

Information (e.g., annotation of probes) of the BeadChip can be found at:
http://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit.html

1.8 Preparation: RStudio and tutorial script

All files that you need to complete this tutorial are available from the course web-site:
https://bioinformaticslaboratory.eu/2022/02/introduction-to-bioinformatics-arcaid/

Install the course R script

1. Start RStudio from the startmenu
2. Download the file "**Claudin1.R**" from the course web-site and save it to the Desktop.
3. Create a new directory on the Desktop with a name of your choice and put the "**Claudin1.R**" script in it.
4. Go back to RStudio and open "**Claudin1.R**" via "File -> Open File". This will open the script in the left-upper window of RStudio
5. Carefully read the script, both comments and commands. Comments in the script are preceded by (multiple) "#". Code can be executed by putting the cursor on the line and pressing "Ctrl + R" or selecting the "Run" button. You can also copy-paste the code to the lower left (i.e. "Console") window and press "Enter".

Note: in this tutorial we will not explain how to download all the necessary data, but use pre-processed files. For more information see the vignettes of the various R packages, i.e. TCGAAssembler, TCGABiolinks, etc.

1.9 TCGA data

To download the TCGA data load the Claudin1.R script in Rstudio and follow the instruction in this script.

Inspect the downloaded files that are in the designated directory to see what the files actually look like and what (kind of) information they contain. This can be done in various ways, i.e. in R, using a text-editor or in Excel. If you use Excel, be sure to first start Excel and load the data via the "Data -> Import" menu to prevent Excel from automagically converting your precious data into an undesired format…

If you completed the Claudin1.R script then you have produced a density plot based on 25 ER positive and 25 ER negative patients in which you indicated the expression values for the two patients you selected (blue, red dashed line, see figure below).

This only gives an impression of expression differences between ER neg and ER positive samples. To determine if the expression of Claudin1 is different between ER neg and ER pos further statistical analysis is required. This is beyond the scope of this tutorial.

Question: Could the results in your density plot be in agreement with the results from Di Cello?

Answer

Di Cello observed:

(1) Methylation of CLDN1 promotor CpG island is frequent in ER+ breast cancer (but not all ER+ are methylated in promotor CpG). Methylation is associated with low CLDN1 expression

(2) In most of the ER- samples, CLDN1 was not methylated. Some of these tumors overexpress CLDN1

This is in agreement with the density plot (see next page).

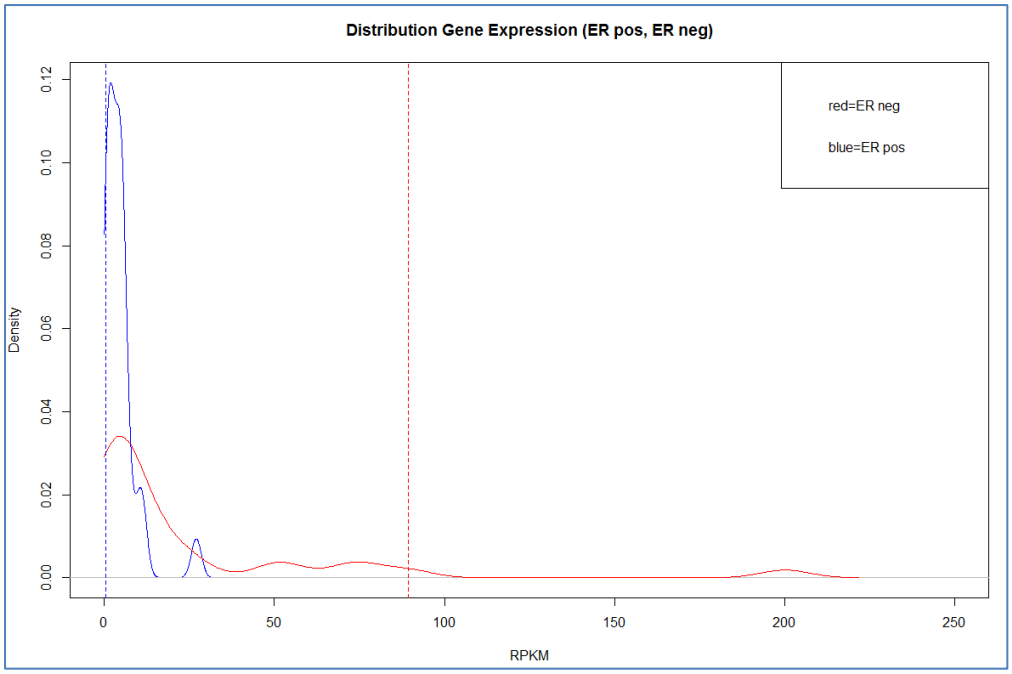


Figure. Density plot of expression of CLDN1 of 25 ER positive and 25 ER negative samples. The dashed lines indicate the expression levels of two selected patients.

Part III. Visualizing DNA methylation data in the UCSC Genome Browser

1.10 Introduction

The UCSC genome browser contains information similar to the Ensembl genome browser although its user interface is different. However, you should be able to use this browser without to much effort.

1.11 Visualization of methylation data

The Ensembl and UCSC genome browsers allow you to integrate your own data (e.g., methylation, gene expression, gene annotation, etc) for visualization. In the next steps we will visualize the methylation data obtained for the ER+ en ER- samples which you have selected in Part II of this tutorial.

Construction of BED file

1. The TCGA Assembler has produced a file 'BRCA_humanmethylation450.txt' (and is in the RawData.zip you downloaded via the 'Claudin.R' script). Open this file with Excel.

2. This file contains the methylation values for each of the 13 probes of the CLDN1 gene for the samples you have selected. Copy this subset of data to a new excel file. Save this file. We will use this file to construct two BED files and bedGraph files in the next steps.

3. To visualize the subset of methylation data we need to convert this file to a format that can be handled by the UCSC browser. These formats are described here: http://genome.ucsc.edu/FAQ/FAQformat.html. We will use the BED format and use only the first 5 columns of this format (chromosome number, chrom start/end, probe name, and methylation score). Note, that these BED files can also be read by Ensembl.
    a. The methylation score should be a value between 0-1000. Therefore, make sure to multiply the methylation values in your excel sheet with 1000.
    b. Make one BED file for your ER+ sample, and another file for your ER- sample.
    c. Use the UCSC documentation to construct these bed files from your Excel file. Make sure to save the BED file as a tab-delimited text (.txt) file.

4. Add one of the following line as the first line in your BED file (this will take you directly to the CLDN1 gene in the browser), depending on the reference genome used (resp. GRCh37/hg19 or GRCh38/hg38):
    browser position chr3:190006744-190056981
    browser position chr3:190305701-190322446

5. The second line in your BED file should be a 'track' line. Make sure to use visibility=2, useScore=1 in this line in addition to other arguments that you may supply.

Visualization of BED file in UCSC browser

1. Go to the UCSC genome browser web-site http://genome.ucsc.edu/ and go to 'Genome browser'.

2. Click on the button ‘add custom track’. This will allow you to upload both of your BED files for visualization.

3. After uploading both BED files, click the button ‘go to genome browser’. Now you should be able to see your methylation data for both of your samples in two separate tracks together with the annotation of the CLDN1 gene.

4. The methylation values are displayed as grades of grey (or another color if you specified another color with ‘color=’ in the track line). This makes it difficult to see the differences. Instead you can create bedGRAPH files which display the data as a histogram. Just add ‘type=bedGraph’ to your track line and upload your new files to the genome browser. This makes the comparison between the ER+ and ER- methylation easier.

Further configuration of the UCSC browser

1. The genome browser also shows the annotation of the CpG island in the CLDN1 promotor. If it does not show the CpG island, then turn it on in the ‘Regulation’ section.

Question: Are the results of your visualization in agreement with the Di Cello paper?

Answer

- Methylation of CLDN1 promotor CpG island is frequent in ER+ breast cancer (but not all ER+ are methylated in promotor CpG)
- In most of the ER- samples, CLDN1 was not methylated