



{: style="width:350px; padding-right:50px"}



{: style="width:150px"}

Introduction to Genome Browsers

Anticipated workshop duration when delivered to a group of participants is **4 hours**.

Note that not all the exercises are expected to be completed during the workshop.

For queries relating to this workshop, contact Melbourne Bioinformatics at:
bioinformatics-training@unimelb.edu.au.

Overview

This tutorial will introduce you to the genome browser format and illustrate how some freely available genome browsers can be used to interrogate a variety of data types, such as gene expression, genomic variation, methylation and many more.

Topic

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Statistics and visualisation
- Structural Modelling
- Basic skills

Skill level

- Beginner
- Intermediate
- Advanced

This workshop is designed for participants with no previous experience of using Genome Browsers and no programming experience.

Description

Learn how to make the most of Genome Browsers!

By focusing on gene expression, this hands on tutorial will provide beginners with an introduction to both the UCSC Genome browser and IGV (Integrated Genome Viewer). Tools and public datasets will be used to illustrate how the expression of transcript variants can be investigated in different, tissues and cell types using public data, including human RNAseq data from GTEx and mouse cell type RNAseq data from Tabula Muris, as viewed within the UCSC genome browser. A subset of Single cell RNAseq data from the Allen Brain Atlas Celltax study will also be downloaded from SRA and visualised in IGV. The data and genes used in this workshop are taken from the neuroscience field, however the analysis approaches and tools illustrated can be applied to many research areas.

This tutorial is in three parts:

Section 1: Introduction to the general features of genome browsers.

Section 2: Hands on tutorial of the [UCSC Genome Browser](#).

Section 3: Hands on tutorials of the [Integrative Genomics Viewer](#).

This tutorial was developed for use as part a series of workshops for neuroscience researchers, hence the data and example genes are drawn from neuroscience field and focused on analysis and visualisation of expression data. However, the skills taught in this tutorial are applicable to all areas of research.

Data: [GTEx](#) and [Tabular Muris](#) data as represented in the UCSC Genome Browser, and [Celltax single cell expression atlas data](#) downloaded from [SRA](#).

Tools: [UCSC Genome Browser](#), [Integrative Genomics Viewer \(IGV\)](#).

Workshop instructions

Click [here](#) for a printer friendly PDF version of this workshop.

Learning Objectives

At the end of this introductory workshop, you will:

- Understand the how some types of genomics and expression data are represented in Genome browsers.
 - Understand gene models and identify differences between transcripts variants.
 - Determine the tissue/cell type expression profiles of a gene of interest in mouse and human expression data.
 - Know some basic files types used in Genome browsers and upload and view local BAM files.
 - Use the 'Blat' tool to locate genomic regions with similarity to a sequence of interest.
 - Create custom interactive views with multiple datatypes to share with colleagues and generate images for publications.
-

Requirements and preparation

Attendees are required to provide their own laptop computers.

If delivered as a workshop, participants should install the software and data files below prior to the workshop. Ensure that you provide sufficient time to liaise with your own IT support should you encounter any IT problems with installing software. Unless stated otherwise, recommended browsers are Firefox or Chrome.

Preparing you and your laptop prior to starting this workshop

Required software:

1. Download and install [IGV](#) (Free).
2. Ensure that [Chrome](#) or [FireFox](#) are installed and up-to-date.
3. Create a user account in the [UCSC genome browser](#).

Required Data

No additional data needs to be downloaded prior to this workshop. Required data will be downloaded as part of the tutorial exercises.

Mode of Delivery

This workshop will be run using freely available Web interfaces and free software using graphical user interfaces. See above.

Author(s) and review date

Written by: Victoria Perreau | Melbourne Bioinformatics, University of Melbourne.

Created: October 2020

Reviewed and revised: October 2021

Background

Genome Browser background

Genome browsers are invaluable for viewing and interpreting the many different types of data that can be anchored to genomic positions. These include variation, transcription, the many types regulatory data such as methylation and transcription factor binding, and disease associations. The larger genome browsers serve as data archives for valuable public datasets facilitating visualisation and analysis of different data types. It is also possible to load your own data into some of the public genome browsers.

By enabling viewing of one type of data in the context of another, the use of Genome browsers can reveal important information about gene regulation in both normal development and disease, assist hypothesis development relating to genotype phenotype relationships.

All researchers are therefore encouraged to become familiar with the use of some of the main browsers such as:

- [The UCSC Genome Browser \(RRID:SCR_005780\)](#)
- [ESEMBL Genome Browser \(RRID:SCR_013367\)](#)
- [Epigenome browser at WashU \(RRID:SCR_006208\)](#)
- [Integrative Genomics Viewer \(IGV\) \(RRID:SCR_011793\)](#)

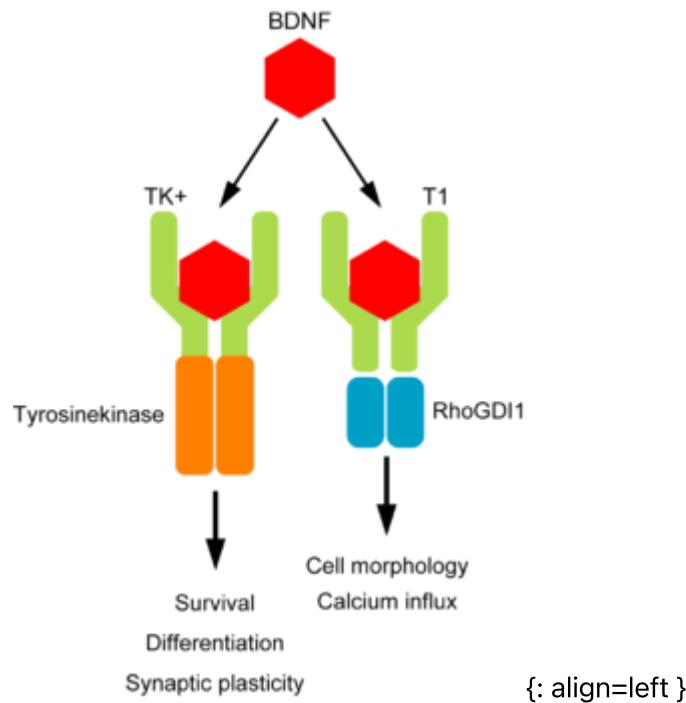
They are designed for use by researchers without programming experience and the developers often provide extensive tutorials and cases studies demonstrating the myriad of ways in which data can be loaded and interpreted to assist in develop and supporting your research hypothesis.

Many large genomic projects also incorporate genome browsers into their web portals to enable users to easily search and view the data. These include:

- [GTEx](#)
- [gnomAD](#)

BDNF and TrkB signalling

This tutorial uses the a well known and important signalling pathway in the central nervous system (CNS) to illustrate some of the Genome browser tools and utility.

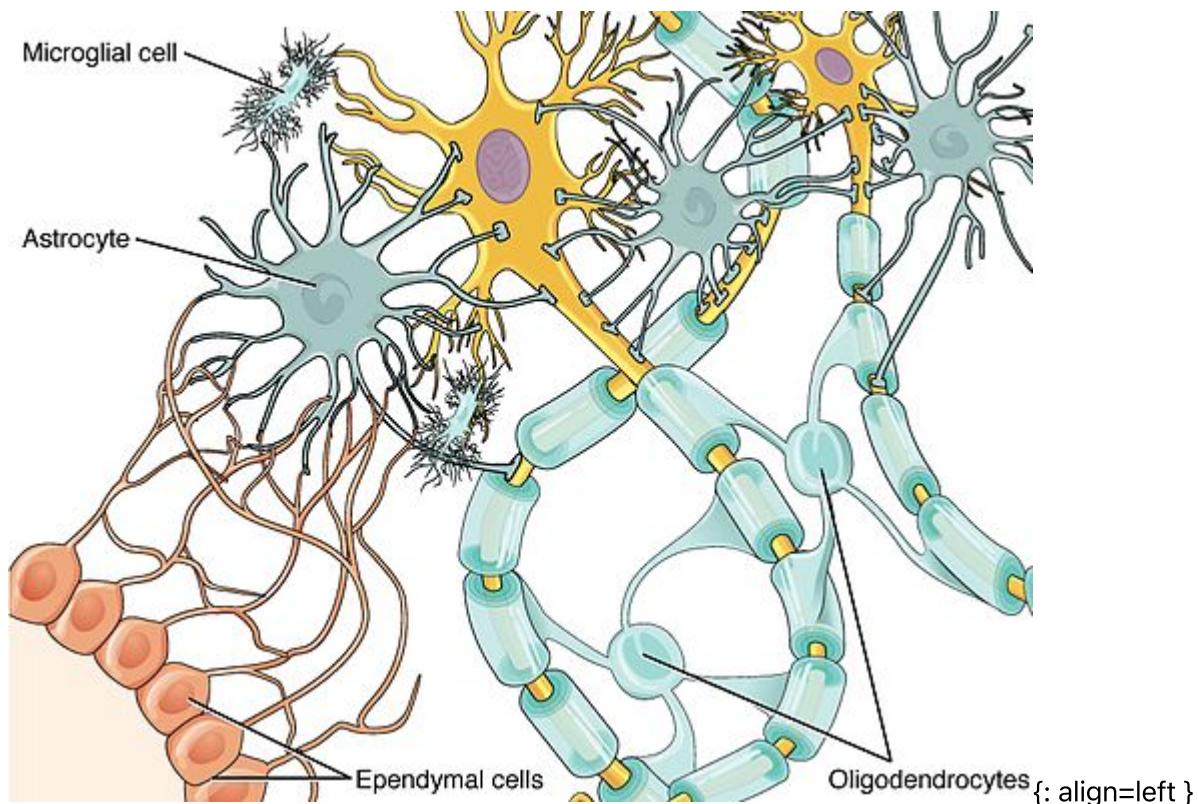


Brain Derived Neurotrophic factor (BDNF) protein is an important neurotrophin responsible for regulating many aspects of growth and development in different cells within the CNS. TrkB is an important receptor that binds extracellular BDNF and propagates the intracellular signalling response via a tyrosine kinase. This TrkB receptor protein is encoded by the NTRK2 gene.

The NTRK2 gene expresses a number of different transcript variants in different cell types. The most well studied of these is the full length TrkB receptor referred to as TrkB, which is mainly expressed in neuronal cell types. The other transcript variants all express the same exons encoding the extracellular domain of the receptor (shown in the figure here in green) but have truncated intracellular domains, which do not include the tyrosine kinase domain and thus activate different signalling pathways upon binding to BDNF. None of these truncated protein products have been well studied, but the most highly expressed receptor variant is known as TrkB-T1, and is known to be highly expressed in astrocytes.

Since the transcript variants are differently expressed in different cell types within the CNS the NTRK2 gene is a very useful example for exploring cell type specific transcript expression in available public data.

****Major CNS cell types:****



- Neuron (yellow cell in the image)
- Astrocyte
- Oligodendrocyte
- Microglia
- Ependymal

Section 1: Introduction to Genome Browsers

Genome browsers rely on a common reference genome for each species in order to map data from different sources to the correct location. A consortium has agreed on a common numbering for each position on the genome for each species. However, this position will vary based on the version of the genome, as error correction and updates can change the numbering. Therefore it is very important to know which version of the genome your data of interest is aligned to.

The sequence for the human reference genome was accumulated up over many years from sequence data from many different sources and does not represent the sequence of one single person. Instead it is a composite of fragments of the genome from many different people. Also, unlike the human genome which is diploid, the human genome is haploid. That is there is only one copy of each chromosome. It therefore does not reflect the variation on the population, or even the most common variants in the human genome.

Exploring variation within human genome is very important and facilitated by genome browsers but not covered in this workshop.

!!! info "Genome Build version number - further reading" * The Genome reference consortium

* What does the nomenclature mean?

For further info on Human Genome version updates I recommend you look at the updates and [blog pages on the UCSC genome browser] (<https://genome.ucsc.edu/goldenPath/newsrch.html#2019>).

Section 2: The UCSC Genome Browser Interface

In this section we will become familiar with the web interface of the UCSC genome browser and explore some of the tools and public datasets available.

- Explore features of particular chromosomal regions
- Investigate specific genes as well as collections of genes
- Search for locations of sequences and markers
- Retrieve annotation information for specific regions or genome-wide
- View your own data in context of other annotations
- Compare a region of one genome to genomes of other species

Weekly maintenance of the browser is at 5-6 pm Thursdays Pacific time, which is equivalent to 11am-12pm AEST time. During this time the browser may be down for a few minutes. To ensure uninterrupted browser services for your research during UCSC server maintenance and power outages, bookmark one of the mirror sites that replicates the UCSC genome browser.

Accessing the tools: Many of the tools that we will explore can be selected via multiple different routes within the browser interface. One way to access many tools is via from the top toolbar on a pull down list, other tools can be accessed from within the browser window. In the following instructions a series of blue boxes is used to indicate successive lower levels from the pull down menu when starting with the top toolbar. For example, the notation below indicates that you should select 'Genome Browser' from the top tool bar and then click on 'Reset all user settings'.

Toolbar Genome Browser Reset all user settings

Accessing help and training: The UCSC genome browser is supported by a rich training resource which has new material added regularly to the [YouTube channel](#). To access training and develop your skills further go to: Toolbar Help Training

Getting started

1. Open the Browser interface:

- o Navigate to the UCSC genome Browser and sign in if you have an account.

Our tools

- **Genome Browser**
interactively visualize genomic data
- **Coronavirus Data**
view SARS-CoV-2 genome and COVID-19-related datasets
- **BLAT**
rapidly align sequences to the genome
- **Table Browser**
download data from the Genome Browser database
- **Variant Annotation Integrator**
get functional effect predictions for variant calls
- **Data Integrator**
combine data sources from the Genome Browser database
- **Genome Browser in a Box (GBIB)**
run the Genome Browser on your laptop or server
- **In-Silico PCR**
rapidly align PCR primer pairs to the genome
- **LiftOver**
convert genome coordinates between assemblies
- **Track Hubs**
import and view external data tracks
- **REST API**
returns data in JSON format

[More tools...](#)

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the [first complete sequence of the human genome](#).

What's new

Oct 5, 2020 - New dbVar Common Structural Variants track

2. First reset the browser, so that we all see the same screen:

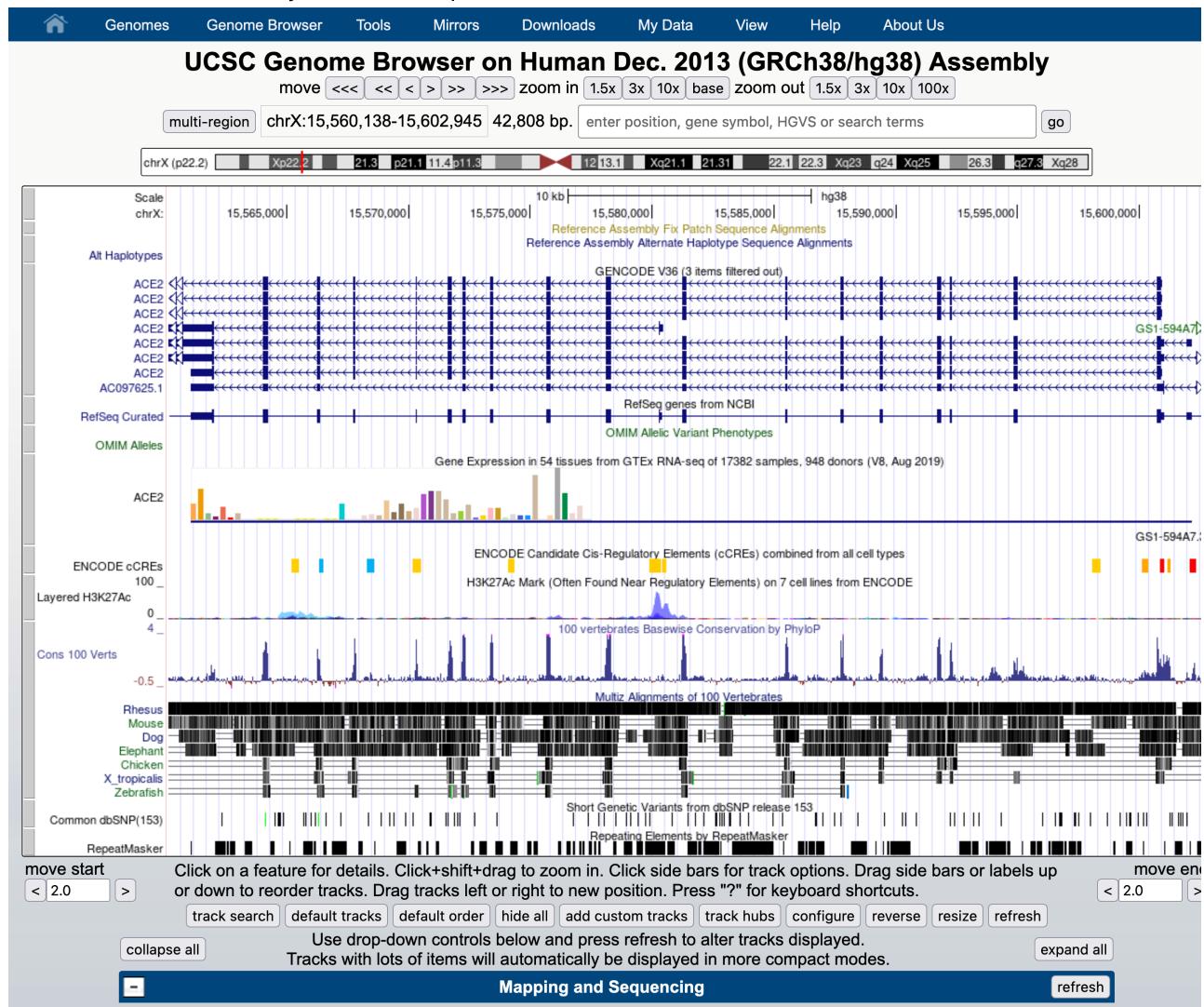
Toolbar Genome Browser Reset all user settings

3. Select and open the human Genome Hg38 at the default position, there are a few different ways to do this

- o Toolbar Genomes (this takes you to the Genome gateway page)
 - Check that GRCh38 is selected in 'human assembly' and click on the blue GO box
- o Toolbar Genomes Human GRCh38hg38 (takes you directly to the genome)

You should see a view of the browser similar to the image below, opening at a position on the X chromosome of Human genome version GRCh38 showing the gene model for the ACE2 gene. Some

of the default tracks may have been updated since this screen shot was made.



4. Familiarise yourself with the main areas of the interface and locate:

- The main Toolbar
- Blue bar track collections (data of similar types are collected together under the same 'Blue bar' heading). Scroll down to see additional data collections and which ones are turned on as default.
- Genome species and version number
- Position box
- Navigation tool buttons
- Chromosome ideogram
- Genome view window
- Pre loaded tracks, track titles:
 - The grey bars on the left of the genome view can be used for selecting and configuring the tracks.
 - You can change the order of the tracks by dragging these grey bars up and down.
- Turn tracks on and off:
 - You can hide tracks by right clicking on the grey bar or by turning them off in the Blue bar collections.
 - You have to click on a 'refresh' button for changes to be reflected in the genome view window.

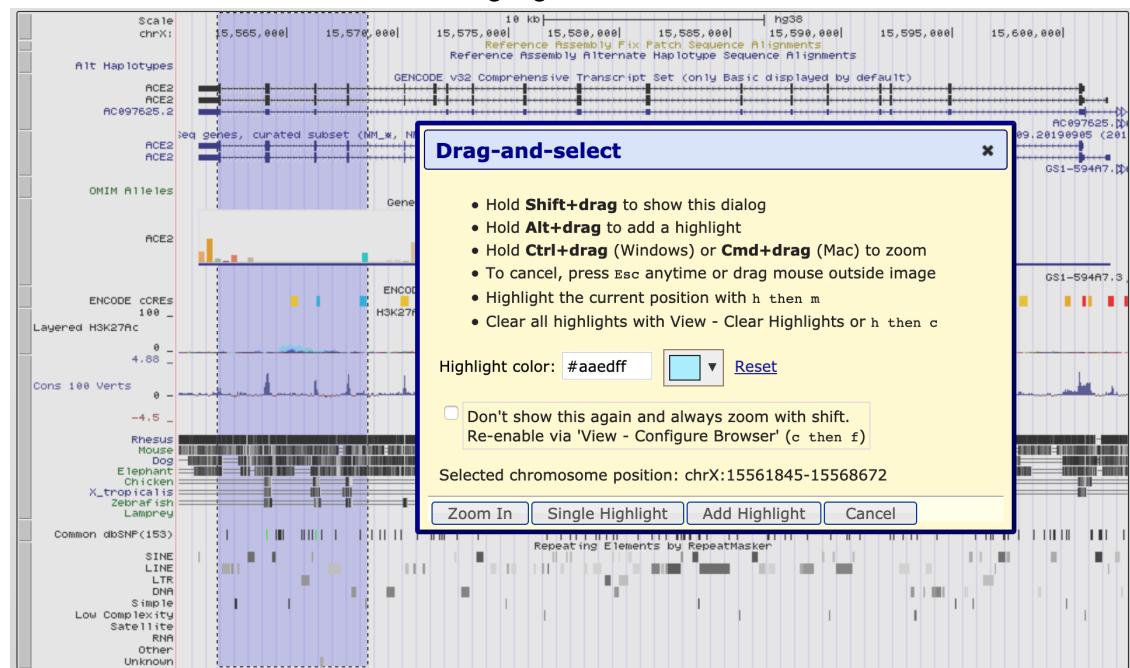
- View the **configuration page** for one of the tracks. The configuration page gives you a lot of information about the data track and its colouring. You can open the configuration page for a track by:
 - clicking on the grey bar for the track or,
 - clicking on the track title in the Blue bar collection. More information and options are usually available by selecting the configuration page for a track via the track title in the Blue Bar collections.
- Select 'resize' button under the genome view to fit the genome view window to your screen

5. Customise your view by using the 'Configure' tool to change the font size to 12. Use either method below to open the Configure tool.

- Toolbar View Configure browser text size 12 submit
- or start by clicking on the configure button below the genome view window.

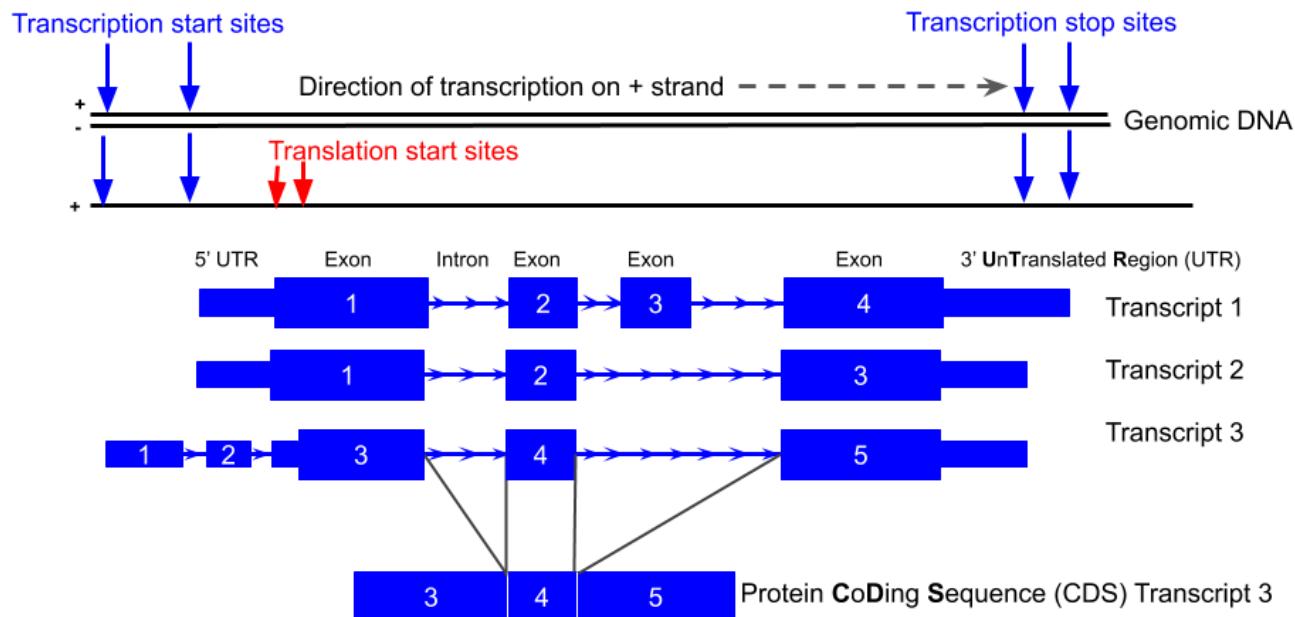
6. Practice navigating around the genome view.

- move left and right- using both the navigation buttons and your mouse
- zoom in and out- using navigation buttons
- zoom in to a region of interest- using '**Drag-and-select**':
 - using your mouse select a region of interest by clicking the ruler (position track) at the very top of the genome view window.
- This is also how to access the '**highlight tool**' which you will use in a later exercise to highlight a region of interest.
 - Click on the down arrow next to the highlight colour to select a different colour.



###Understanding the gene models

Gene model representation

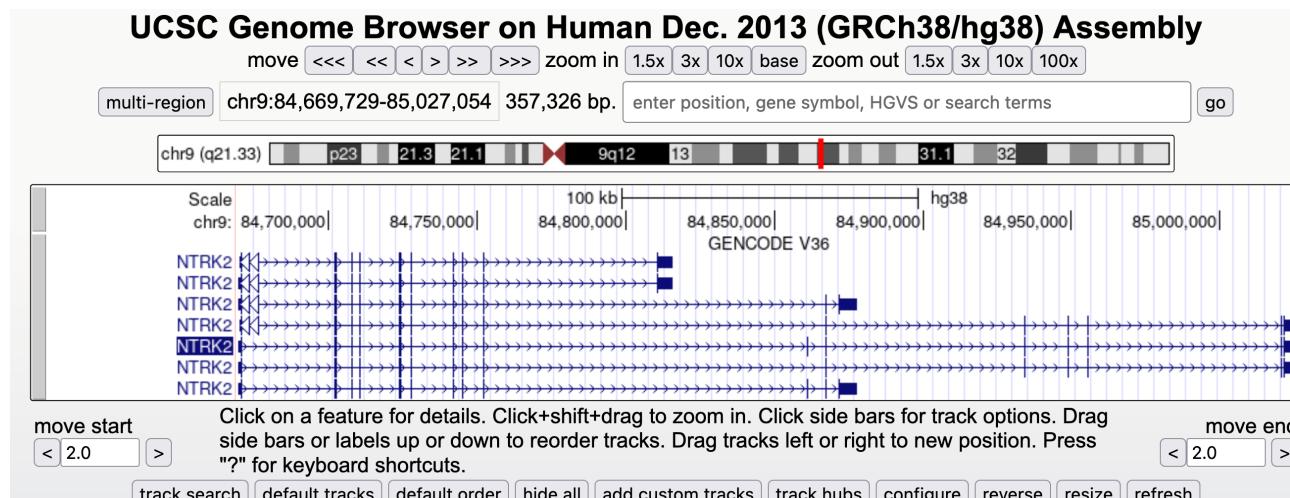


NTRK2

First we are going to familiarise ourselves with the gene model representation of the different transcripts of NTRK2.

1. Navigate to the NTRK2 gene position in GRCh38 and view the gene models

- You can navigate to a different region by typing in the position box.
 - If you know the specific location you are interested in type in the location using the format "chr#:1234-1234".
 - If you have a gene of interest you can type in the gene name (eg: NTRK2). Note the autocompleted suggestions that appear when you start typing. You can select from one of the suggestions or click go and select from a wider range of options.
 - Type (or copy and paste) **NTRK2** or **chr9:84,665,760-85,030,334** into the position box.
- Hide all tracks by selecting the button below the genome view.
- Turn on only the **Gencode V36 Genemodels** in 'full' viewing mode by selecting from the blue bar group labelled 'Genes and Gene predictions'.
- Turn on the 'Conservation' track to 'full'
 - Don't forget to click refresh.
- When you have navigated to the NTRK2 gene, zoom out until you can view all of the 5' UTRs and 3' UTRs for all transcript variants for this gene. Then drag the view left and right to center (like in Google maps) or 'drag and select' the region to center the gene in the Genome view. You should see something like the image below.



!!! question "Which strand is the gene encoded on / transcribed from? (+ or - strand)"

- Identify the exons, introns and UTRs

!!! question "Do regions of conservation only occur where there are coding regions?"

!!! question "How many different transcripts variants are there for this gene?"

!!! question "How do they differ?"

- Select a coding region (full height boxes) towards the 3'UTR of the gene.

- zoom in to the region until you can see the letters of the amino acid sequence.

!!! question "Why are some amino acid boxes red or green?"

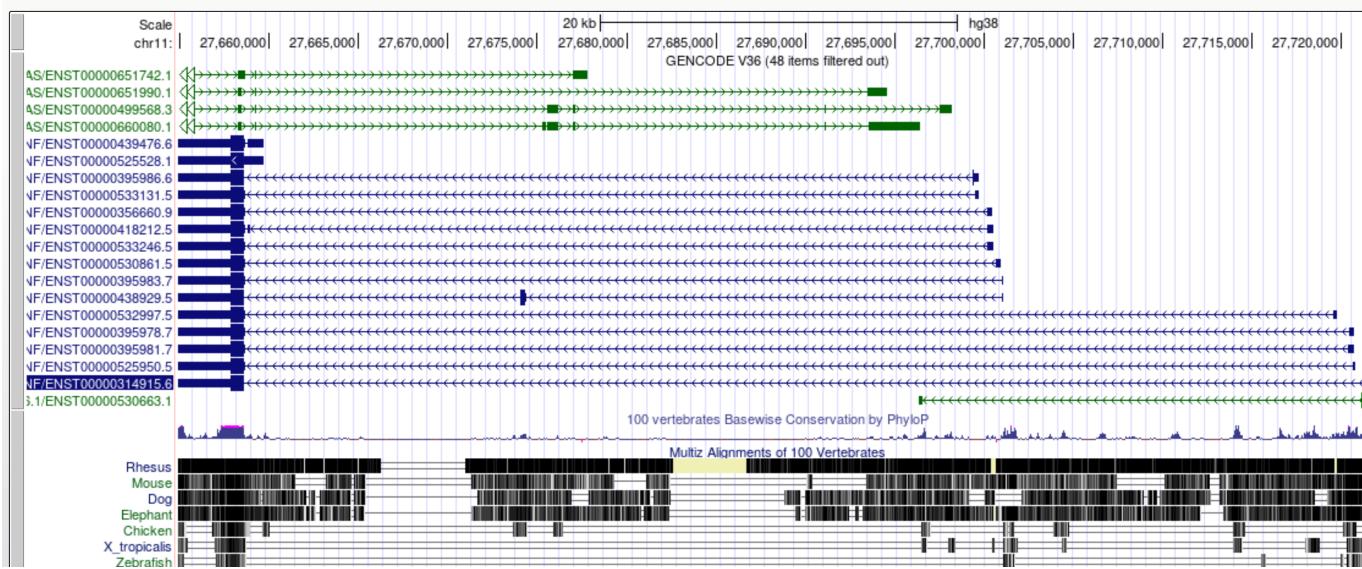
- Zoom in again until you can see each amino acid number.

!!! question "Why do different transcripts have different amino acid numbers?"

- Note that one of the transcript names is in white text with a black background, this is the transcript you selected from the autocompleted list or the search results.
- Change the '**view settings**' for the track. Switch between dense squish pack full to see how it changes the representation of the models.
 - Right click on the track grey bar in the left of the genome window to access view settings.
- Go to the configuration page for the **GENCODE V36** track and change check the box to also reveal the 'Ensembl ID' in the label.
- The transcript names are now too long to fit on the screen. You can use the configuration page (like you did to change the font size at the beginning of the workshop) and change the number of characters in the label so that you can see the entire transcript label.
- Test your understanding of gene model representation by attempting this 6 questions in this [quiz](#).

Now we look at the gene model for BDNF in the same genome. There are some differences that enable us to demonstrate some more tools.

1. Navigate to the BDNF gene position in GRCh38 and view the gene models



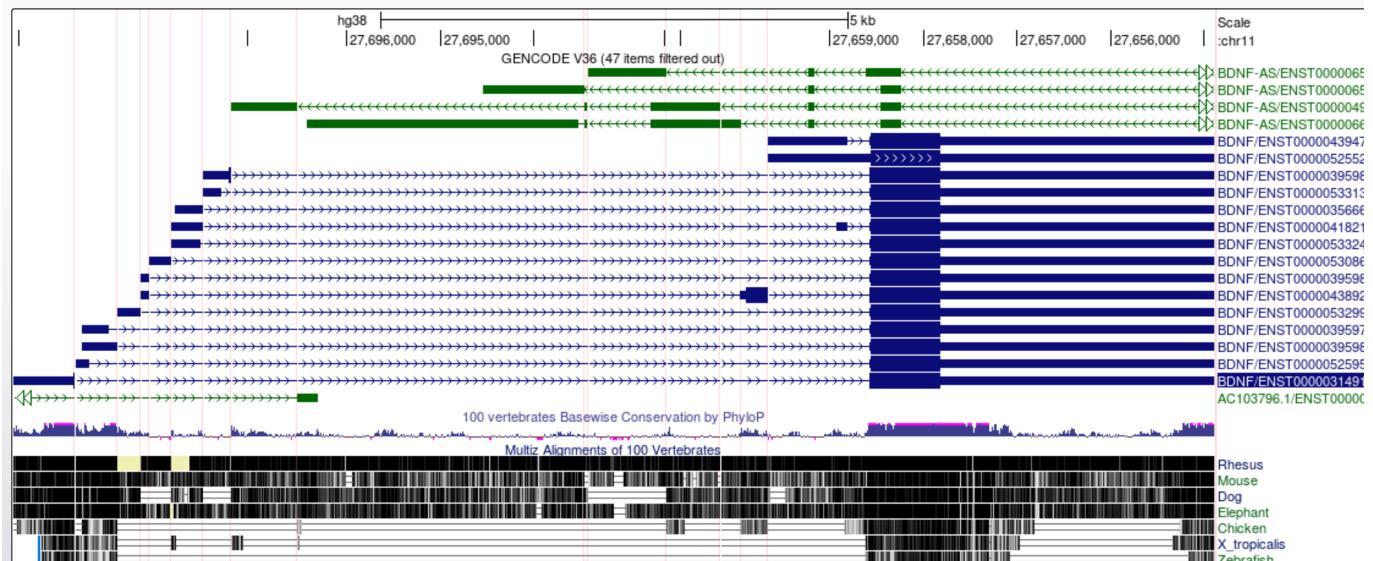
- Note that there are blue transcript models encoded on the - strand and green BDNS-AS transcript models on the + strand. BDNF-AS is the antisense gene.
- Colouring info is specific for each track and can be obtained from the configuration page, below is the colouring legend for the GENCODE V36 track.

Coloring for the gene annotations is based on the annotation type:

- **coding**
- **non-coding**
- **pseudogene**
- **problem**
- **all 2-way pseudogenes**
- **all polyA annotations**

- Since the convention is to display genes in the 5' to 3' orientation it can be useful for our own interpretation, and also for presentation purposes, to flip the orientation of a gene when viewing it in a Genome Browser.
 - Use the reverse button under the genome view window to flip the orientation of the gene.
- When a gene has many large introns taking up a lot of white space in an image it can be difficult to see if exons in different transcript models or other data tracks align. The '**Multi-Region**' view tool can be used to fold the intronic regions out of the view like a concertina. The Broswer selects which region to fold out based on the gene model track(s) that you have turned on at the time.
 - Apply the Multi-Region view from the main tool bar view options.

- Toolbar View Multi-Region select 'Show exons using GENCODE V36'



It is now a lot easier to view a number of interesting features in the BDNF transcript models:

- The transcript variants for the BDNF vary mostly in the genomic position of the 5'UTR.
- The noncoding AS-BDNF gene transcript includes a region that would be antisense to the coding BDNF transcript.

You may find that using the multi-region tool facilitates visualisation and interpretation of gene expression data later in the workshop.

####Blat tool exercise The **Blat tool** is a sequence similarity tool similar to Blast. It can quickly identify region(s) of homology between a genome and a sequence of interest. Due to the presence of **orthologs and paralogs** a target sequence may have similarity to more than one region in the genome. In this exercise you will use Blat to map the sequences of two different expression probes to their target regions and determine which gene transcripts the probes are likely to detect in an expression study.

Microarray expression data is not commonly used now, but some of the data generated from large well orchestrated studies still provide valuable information to researchers. Microarray probes, like *in situ* hybridisation probes, target a small region of the RNA and do not measure the whole RNA transcript. If you are measuring gene expression it is important to know exactly which region of the gene you are detecting. In this exercise we will employ the blat tool to determine which region of the NTRK2 gene the microarray probes in the following study are detecting.

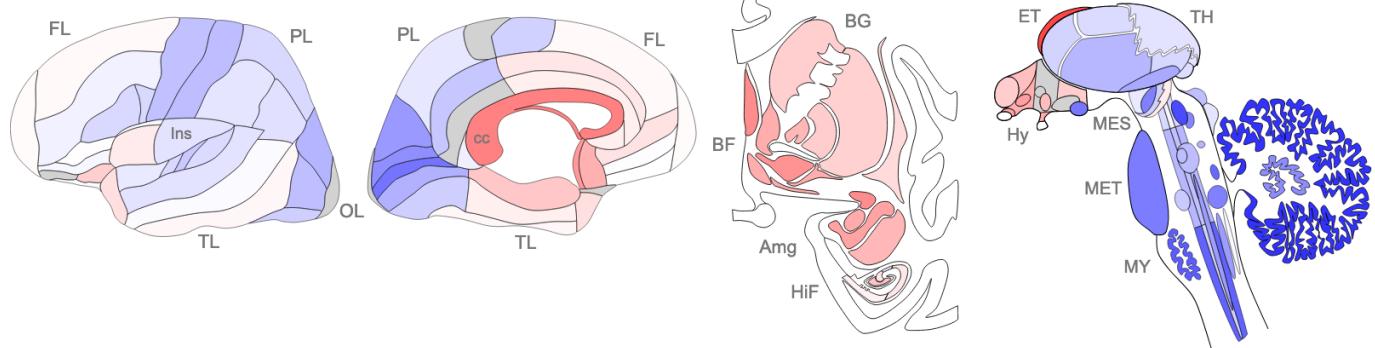
The study was [the Human Brain gene expression atlas](#) generated by the [Allen Institute](#). Below are sequences of two hybridisation probes that were use in a microarray used to detect expression of the gene NTRK2. These two probes result in very different hybridisation and expression patterns across different regions of the brain. As we observed in the exercise above NTRK2 has a number of different transcript variants. The question we have is whether these probes are detecting different or multiple transcripts of NTRK2, and if so which ones?

NTRK2 Probe A_23_P216779 sequence:

TTCTATACTCTAATCAGCACTGAATTCACTGAGGGTTGACTTTTACATCTAACACAGTG

Z score of expression level in Human brain (blue = low expression, red = high expression)

Donor: H0351.2002, 39 yrs, M, Black or African American

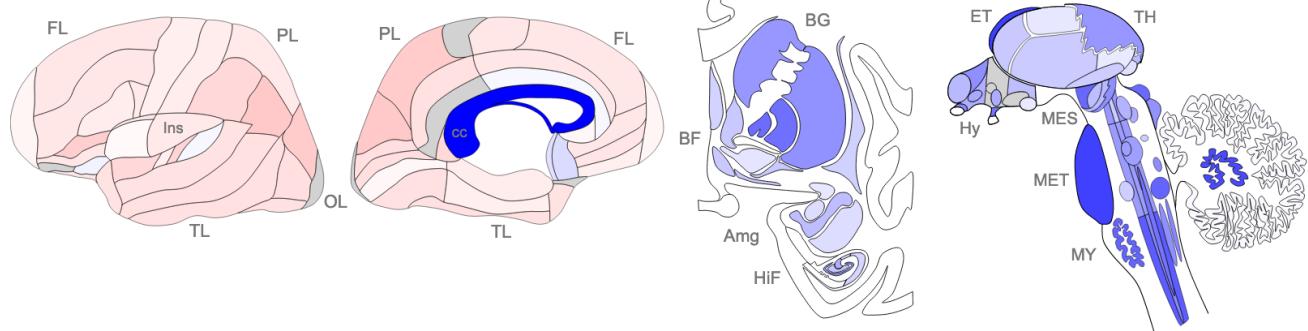


NTRK2 Probe A_24_P343559 sequence

AAGCTGCTCTCCTTCACTCTGACAGTATTAACATCAAAGACTCCGAGAAGCTCTCGAGGG

Z score of expression level in Human brain (blue = low expression, red = high expression)

Donor: H0351.2002, 39 yrs, M, Black or African American



The images above are of one of the six donors included in the atlas, and typical of the expression pattern for NTRK2. These images are taken from the [NTRK2 gene page of Human Brain Atlas](#).

Most obvious in the images above is the high level of expression signal using Probe A_23_P216779 and low level for A_24_P343559 in the corpus callosum (CC) which is a region of white matter in the brain with relatively few neurons and relatively high proportion of myelinating oligodendrocytes. This expression profile is reversed in the the cortical regions, eg. frontal lobe (FL) and parietal lobe (PL), which have a relatively high density of neuronal cells.

1. Use Blat tool to find region of homology

- Select: Toolbar tools blat
- Copy the sequence of the first probe above and paste into the search box
- Select the human GRCh38 for 'Assembly' and click Submit
 - A_23_P216779 returns 2 hits for different chromosomes. One of these has 100% homology over the whole 60 base sequence, the other has 87% homology over a 24 base region.
- Copy and paste the probe name to use as the label for the '**Custom track name**' and '**Custom track description**' and click on Build a custom track with these results. It's not necessary to build a custom track, you could simply click on the browser link to view the results, but creating

a 'custom track' from the blat result enables you to give it a unique name. This is important if you are doing multiple blat searches as you won't be able to tell which one is which.

- Select browser option for the hit with the highest homology to view the result. Observe which region of the NTRK2 gene this probe will target. Zoom out for context.
- Note that a new 'Blue bar' heading has appeared for your custom tracks.
- Repeat for the other probe sequence.

2. Use the 'highlight' tool to keep track of region of interest in the Genome view. It is easy to lose track of a region you are investigating when navigating around the genome in a browser. So we are going to highlight each region of probe homology within the NTRK2 gene, using a different colour for each probe. Highlight is also useful if you have lots of different tracks loaded and you want to check that a feature on one track lines up with another.

- Using your mouse select in the position track at the top to activate 'Drag-and-select'. Select only the region of homology for each probe within the NTRK2 gene and use a different highlight for each region. Then zoom out to view the whole gene again.

!!! question "Do the probes detect coding regions of the NTRK2 gene?"

!!! question "Are the probes likely to detect different transcripts?"

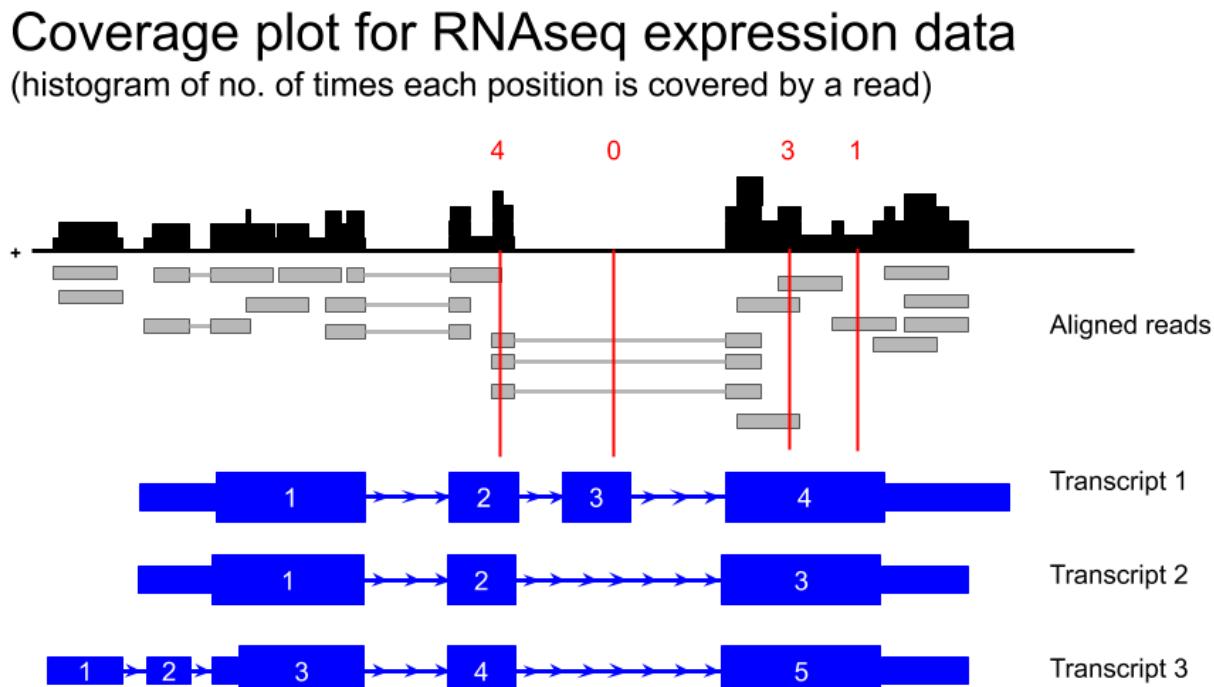
3. Use 'Multiregion view' to make it easier to compare coding regions of different transcripts

- Toolbar View Multi-Region select 'Show exons using GENCODE V36'

I have created a 'public session' of the Blat NTRK2 exercise you can view it from the link in the sessions

Toolbar My data Public session search for "**hg38_NTRK2_blat_probes**".

###Gene expression data



1. Human tissue specific expression data from the [GTEx project](#) is available in UCSC genome browser

- Gene level expression data from GTEx V6 (570) donors, and GTEx V8 (948) donors can be turned on from Blue bar title for more detailed configuration page. These are displayed as coloured bar plots.
- Transcript level data is also available for GTEx V6, this is also displayed as bar plots.
- Transcript level expression data for GTEx V6 is available as coverage plots and is accessed from the '**Track hubs**'.
 - click on the 'track hubs' under the genome view window or access from the toolbar under 'My data'.
 - scroll down and select 'GTEx RNA-seq Signal Hub'. This can be viewed in either hg19 or hg38 so can be compared with a wide variety of other datasets.

Connect	GTEX Analysis Hub	Genotype Tissue Expression (GTEx) Project Analysis Hub	hg19, hg38
Connect	GTEX RNA-seq Signal Hub	RNA-seq read coverage in 53 tissues from GTEx V6 (7572 samples)	hg19, hg38

- The default settings are:

- that all the available data from one individual only is loaded. Data from other subjects in the study can be loaded as desired. For example you could load all available samples for one tissue region only.
- the data is 'autoscale to data view' with a track height of 12 pixels for each sample. You can change the height of the track or add a data transformation.
- you can also change it from a points plot into a density plot by clicking on the track title.

Female Donors Female 20-49 Yrs Track Settings

GTEx RNA-seq read coverage for female donor samples (20-49 years)

Track collection: [▲ GTEx RNA-seq from 199 Female Donors](#)

Description

All tracks in this collection (3)

Display mode: [full](#) [Reset to defaults](#)

Type of graph: [bar](#)

Track height: pixels (range: 8 to 64)

Data view scaling: [auto-scale to data view](#) Always include zero: [OFF](#)

Vertical viewing range: min: max: (range: 0 to 127)

Transform function: Transform data points by: [NONE](#)

Windowing function: [mean+whiskers](#)

Smoothing window: [OFF](#) pixels

Negate values:

Draw y indicator lines: at y = 0.0: [OFF](#) at y = [OFF](#)

[Graph configuration help](#)

- Using the selection matrix for female donors aged 2-49 years, deselect the default samples and select only Brain cortex and Pancreas samples. Then navigate to the location for the gene MYRF. You can increase the height of the datatrack to improve visualisation.

!!! question "Can you locate an exon in the MYRF gene that is present in transcripts expressed in the brain but not in the pancreas?"

!!! question "Does this alternative splicing event result in a frame shift of the coding sequence?"

!!! question "How many amino acids are there in the protein products for each MYRF transcript?"

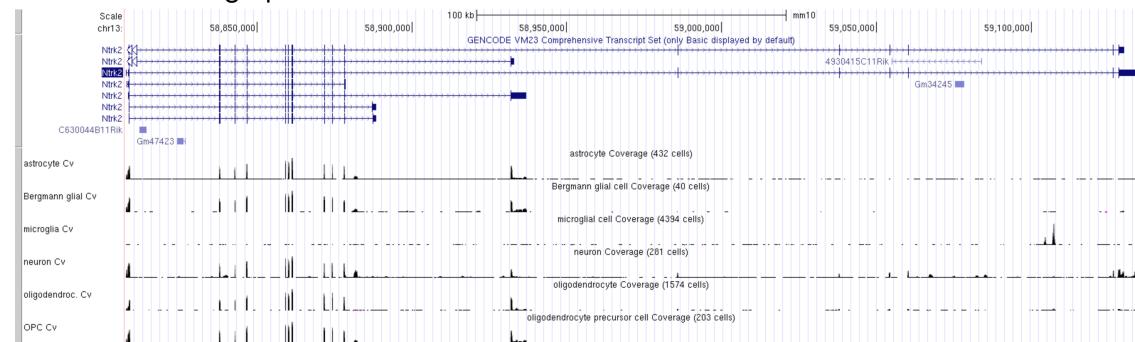
2. The FACS derived data from the **Tabular Muris** cell type data can be visualised as a coverage plot

- Start at the view of the NTRK2 gene in the human genome and navigate to the Ntrk2 gene in the mouse genome using the 'View in other genomes tool'.
 - Toolbar View In Other Genomes (Convert)
 - select New Genome:Mouse , New Assembly:GRC38/mm10, click on 'Submit'
 - select the region with the greatest homology
- Click on the bar chart icon for the Tabular muris data in the default view to see a summary bar chart of the cell type data.
- To see a coverage plot of the expression data we have to configure the Tabular Muris track by selecting it from the blue bar collection.
 - Hide 'Cell expression'
 - Select 'Genome coverage' to full
 - Select 'submit'

This can look a bit overwhelming as there are very many tracks and the default track height is set very high. But it's easy to simplify it by focusing on a few cell types of interest.
- Right click on the grey bar to 'configure the track set'.
 - Change **Track height:** to 30
 - for **Data view scaling:** select 'group auto-scale'
 - clear all the subtracks and then manually select only these few cell types of interest:
 - astrocyte Cv
 - Bergmann glial Cv
 - microglia Cv
 - neuron Cv
 - oligodendrocyte Cv
 - OPC Cv

!!! question "Which cell type has the highest level expression of Ntrk2 in this dataset?"

- Change the 'Data view scaling' to autoscale to dataview.
 - Export a PDF image of the genome view: Toolbar View PDF/PS select 'Download the current browser graphic in PDF'

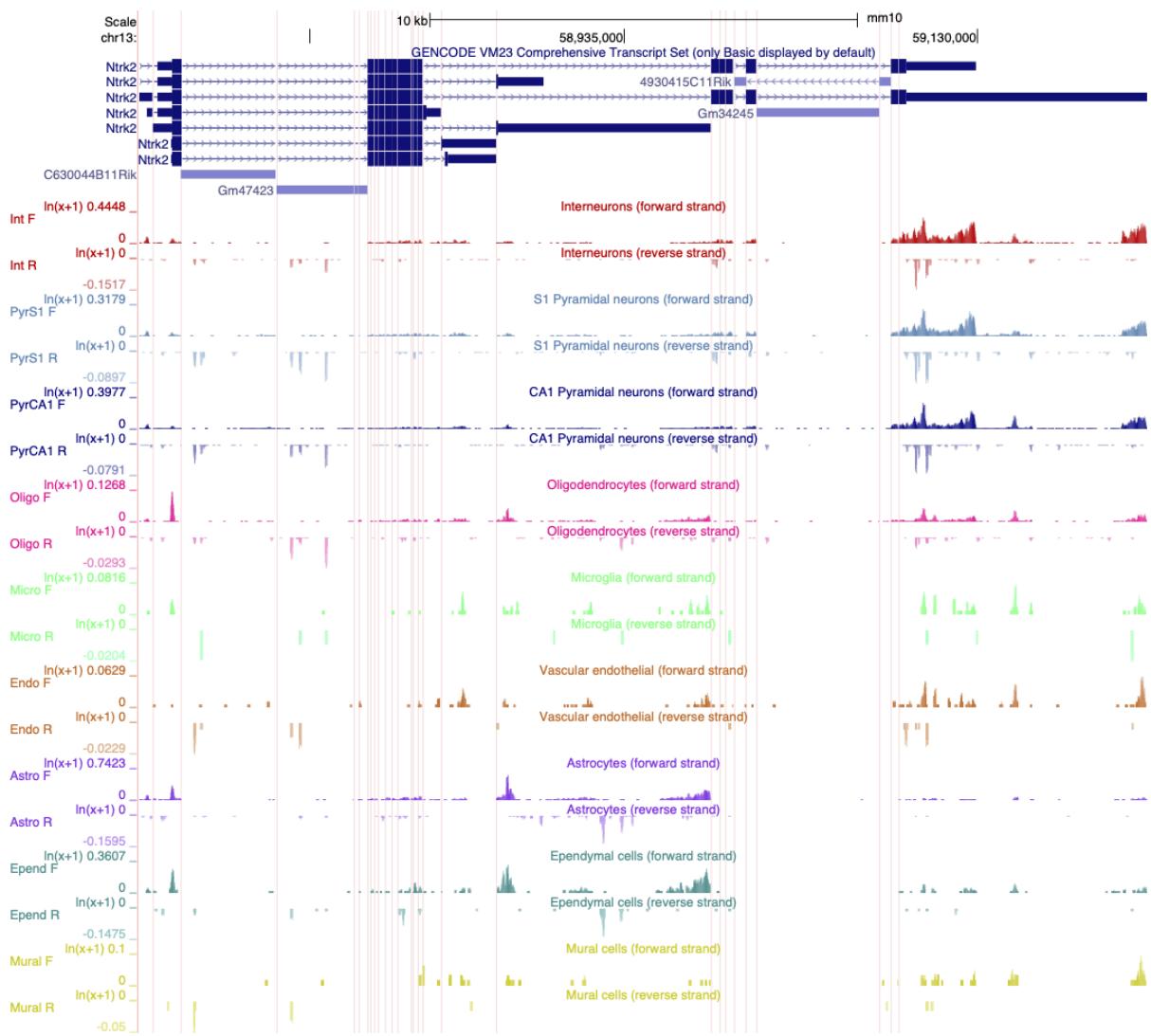


!!! question "Which cell type(s) express the long and short transcripts of Ntrk2?"

3. Mouse CNS cell type expression data can also be validated using an independent single cell dataset of mouse cortex from the [Linnarsson lab](#).

- The data that is publicly available for viewing in the UCCS genome browser but is not housed in the UCSC genome browser. You must first access it from the the Linnarsson lab data page.
- This RNAseq data is stranded, meaning you can see if the transcript data is from the + or - strand.
- Click here for the [Linnarsson lab public data page for this dataset](#) where you can search for cell expression profiles for individual genes.
- Click on the 'Browse the genome' blue text near the bottom of the page.
- This loads 18 different tracks, one for each cell type investigated. The default setting for expression range is quite high and most gene expression is not observed with these settings. Each track must be configured individually rather than as a group, which takes a lot of time. I have created a version of this data as a public session in the UCSC genome browser. Where each track is autoscaled which can make it quicker to determine which expression range would be ideal for visualising the expression of an individual gene. The data is also viewed using 'Multi-Region' which hides the introns in the gene models. The session is illustrated in the screen shot below and you can access this custom track set by clicking on the title in blue text. It is also a publically shared session called **mm10 Linnarsson Celltype autoscale** which you can access from the database of public sessions.

[Linnarsson lab mouse cortex single cell data as autoscaled datatracks](#)



- As an exercise, select 2 or three cell types and adjust the scale to best reflect differences in gene expression of Ntrk2 between these cells. Save this session and share it.

Section 3: IGV

In this section we will download a BAM file of gene expression data from SRA, and view it in the Integrated Genome Viewer (IGV). BAM files must first be sorted and indexed before they can be loaded into genome viewers and IGV has tools to do this without having to use the command line.

The expression data we are using for this exercise is from the mouse [Celltax single cell expression atlas](#) published by the Allen Brain Institute. The cell tax vignette has an expression browser that displays gene level expression as a heat map for any gene of interest. The readsets (fastq files) and aligned data (BAM files) for 1809 runs on single cells are also available for download from SRA.

The SRA study ID for this study is [SRP061902](#) and individual runs from this study are easily selected by viewing the samples in the 'RunSelector'. If you wish to identify particular cell types of interest. For this exercise I have already identified a few samples that we will download in order to illustrate navigating in IGV by looking at the expression of NTRK2 in the same cell types we have discussed in earlier exercises.

For each cell type we will down load a .BAM file containing only the reads from the chromosome of interest.

For each SRA run in the table below open the link to the run to down load the data. Not many raw data sets in SRA have aligned data available for down load but this data set does.

Cell type	SRA run	Vignette Cell ID
astrocyte	SRR2138962	D1319_V
astrocyte	SRR2139935	A1643_VL
neuron	SRR2139989	S467_V4
neuron	SRR2140047	S1282_V

1. Download BAM files from SRA

- Click on the 'Alignment' tab
- Note that the data is aligned to the mouse GRCm38 genome (mm10).
- Select the chromosome of interest. For NTRK2 in mouse it is chr13
- For 'Output this run in:' select **BAM** and click on 'format to:' **File**
- Rename the downloaded file to include the cell type, to avoid confusion. eg:
SRR2138661_astrocyte_chr13.bam

2. Use IGV tools to SORT and INDEX the BAM files

Store sorted BAM files and index files in the same folder.

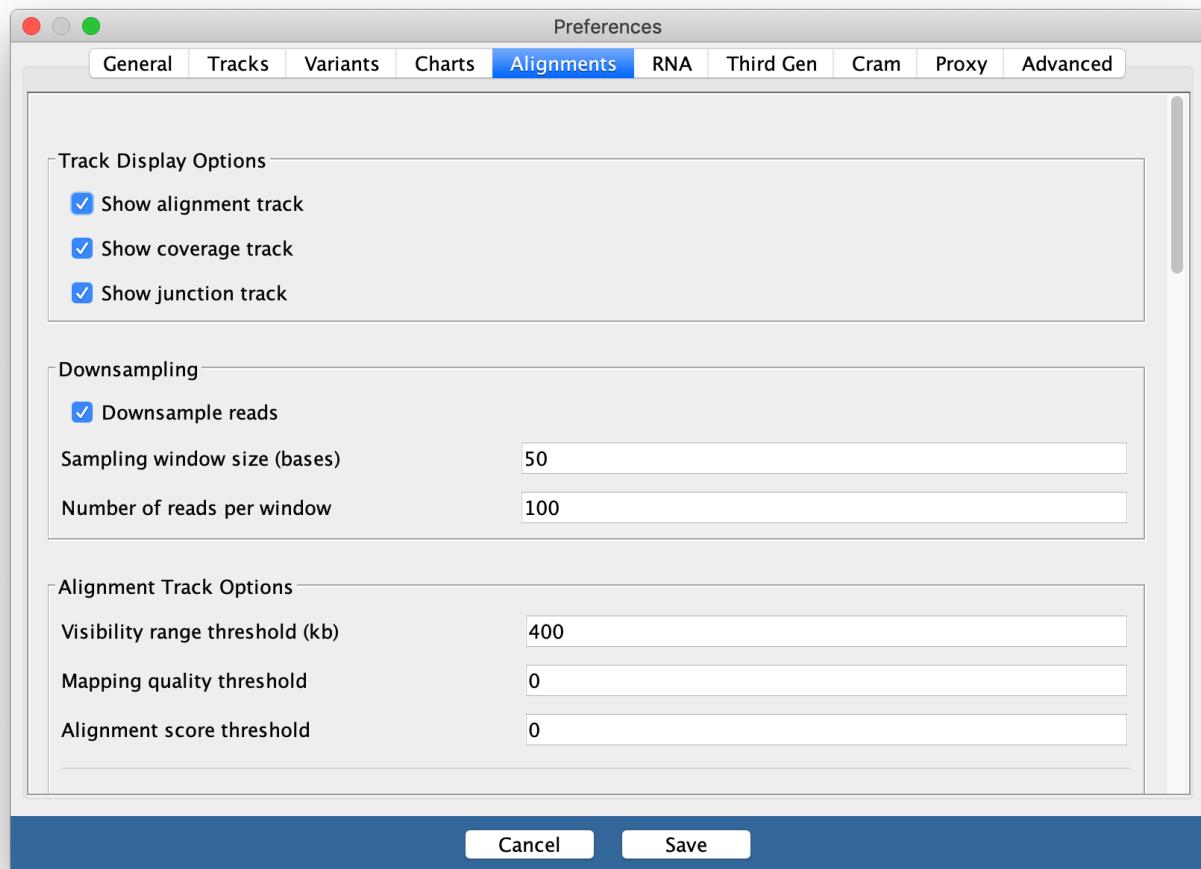
- Open IGV and select Tools / Run igvtools... from the pull down menus.
- Select '**Sort**' from the Command options and use the brows options to select the BAM file you just downloaded and click '**Run**'
- Without closing the igvtools window now select the command '**Index**' and Browse to find the BAM file you just sorted. It will have the same file name with 'sorted' added to the end. eg
SRR2138661_astrocyte_chr13.sorted.bam
- The resulting index file will have the file name : **SRR2138661_astrocyte_chr13.sorted.bam.bai**
 - It is essential that the index file for a BAM file has the same name and is located in the same folder as its BAM file. If not the IGV software will not be able to open the BAM file.

Name
SRR2138661_astrocyte_chr13.bam
SRR2138661_astrocyte_chr13.sorted.bam
SRR2138661_astrocyte_chr13.sorted.bam.bai
SRR2138962_astrocyte_chr13.bam
SRR2138962_astrocyte_chr13.sorted.bam
SRR2138962_astrocyte_chr13.sorted.bam.bai
SRR2139989_neuron_chr13.bam
SRR2139989_neuron_chr13.sorted.bam
SRR2139989_neuron_chr13.sorted.bam.bai
SRR2140047_neuron_chr13.bam
SRR2140047_neuron_chr13.sorted.bam
SRR2140047_neuron_chr13.sorted.bam.bai

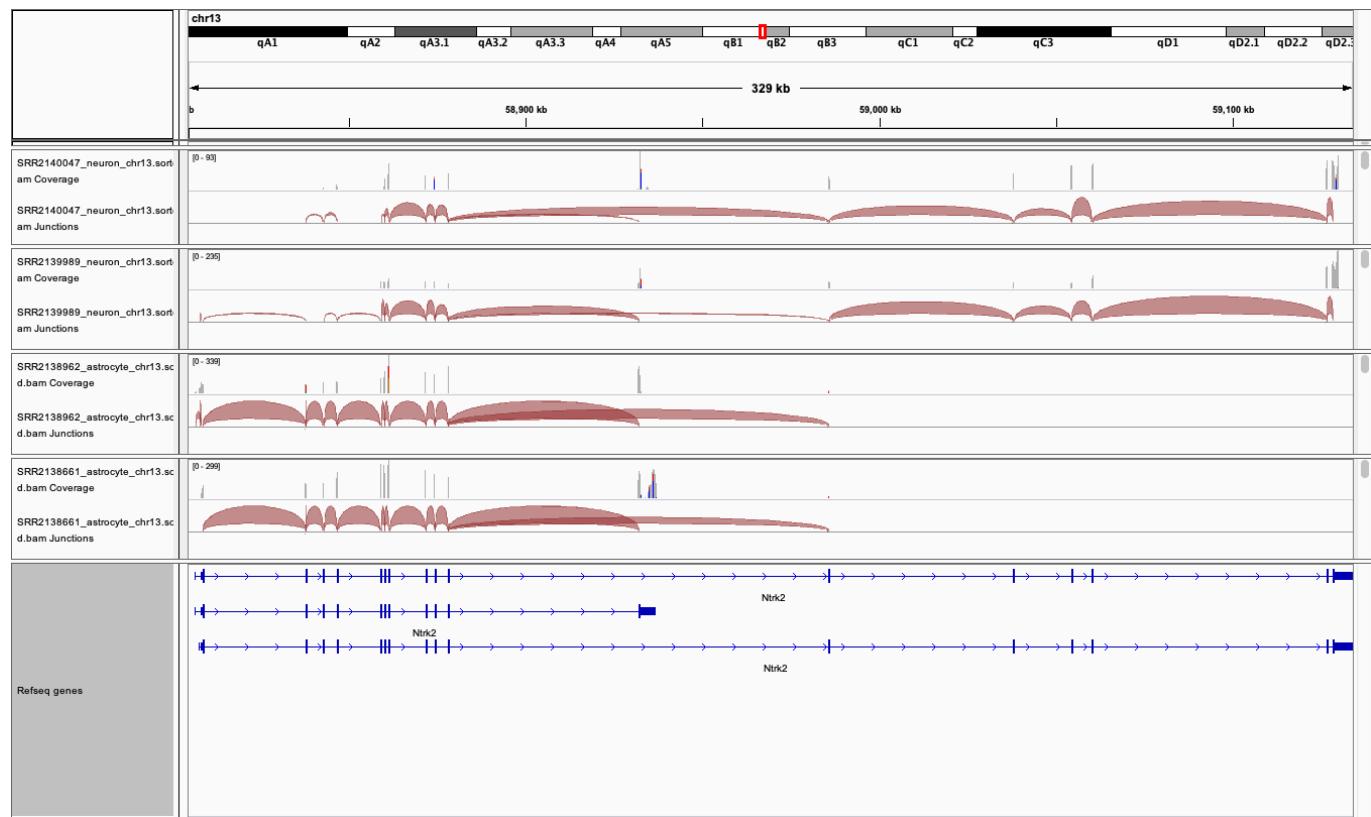
{: style="width:450px"}

3. View the BAM files in IGV

- Select the Mouse (mm10) genome from the genome box in the top right hand corner.
- Select File / Load from File... and select all 4 '_chr13.sorted.bam' files only (use command to select more than one file at a time).
- select open - but don't expect to see any data yet. The genome view window opens on a whole chromosome view as default but it wont show any data until the view region is small enough to show all data in the current view.
- Type the gene name 'NTRK2' into the search window.
- Expand the Refseq gene model track by right clicking it to see all the splice variants
- The gene and thus the genome view is 328kb and the default setting for viewing data is only 100kb. So unless you have already changed your settings alignment data will not get be showing.
- zoom into the region of a coding exon by selecting in the numbered location track at the top of the genome view.
- To see the whole gene in the genome window at the same time you may need to change the preferences.
- Go to View / Preferences and select the 'Alignments tab'.
- Change the visibility range threshold to 400kb.

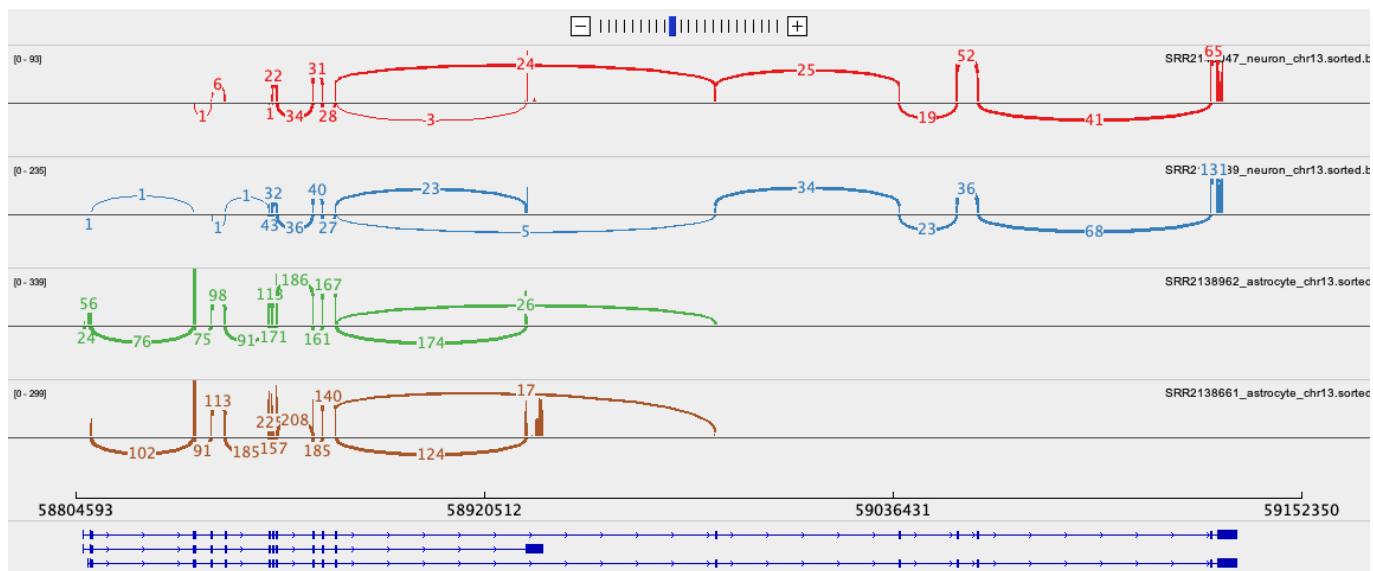


You may need to change this back to a smaller range in the future if you are working with large datasets and/or small amounts of memory on your computer.



4. Export images

- The Genome view above can be exported by selecting 'File / Save image...' from the tool bar.
- To export the Sashimi plot below:
 - Right click on one of the junction tracks and select 'Sashimi Plot' from the poll down menu.
 - Select the tracks you want in your final image.
 - There are some data filtering and style adjustments you can make to the Sashimi plot. Right click on each track to access the menu options. Some changes apply to each track individually and some to all tracks.



5. Download and install the Gencode gene model annotation track

- The refseq gene model track is not as comprehensive as GeneGode gene models. For both Human and Mouse the Gencode gene model gtf annotaion files can be downloaded form [Gencode](#). If you wish to do this be aware that it takes a little time and is not done as part of a workshop.
 - Create a folder called 'annotations/Mouse' in the main 'igv' folder that was installed on your computer when you downloaded IGV.
 - Download the GTF file from the link above and save it in this folder.
 - Unpack and then **SORT** and **INDEX** the .gtf file using igvtools.
 - in IGV, before you load you data files, load this annotation file and it will replace the refseq one.

Additional reading

IGV

https://rockefelleruniversity.github.io/IGV_course/presentations/singlepage/IGV.html