Oct 21, 2020

UCSC_Genome_Browser_and_BLAST_protocol

Forked from UCSC_Genome_Browser_and_BLAST_protocol

[1], [2]

[1]revamped to go into more detail one GB and BLAST; [2]UCSC

1 Works for me

This protocol is published without a DOI.

UCSC BME 22L

ABSTRACT

In this lab, students will work through BLAST and the UCSC Genome browser to find and analyze information about their genes of interest.

PROTOCOL CITATION

, 2020. UCSC_Genome_Browser_and_BLAST_protocol. **protocols.io** https://protocols.io/view/ucsc-genome-browser-and-blast-protocol-bnpmmdk6

FORK FROM

Forked from UCSC_Genome_Browser_and_BLAST_protocol,

CREATED

Oct 21, 2020

LAST MODIFIED

Oct 21, 2020

PROTOCOL INTEGER ID

43469

DISCLAIMER:

ABSTRACT

In this lab, students will work through BLAST and the UCSC Genome browser to find and analyze information about their genes of interest.

Setup: UCSC Genome Browser

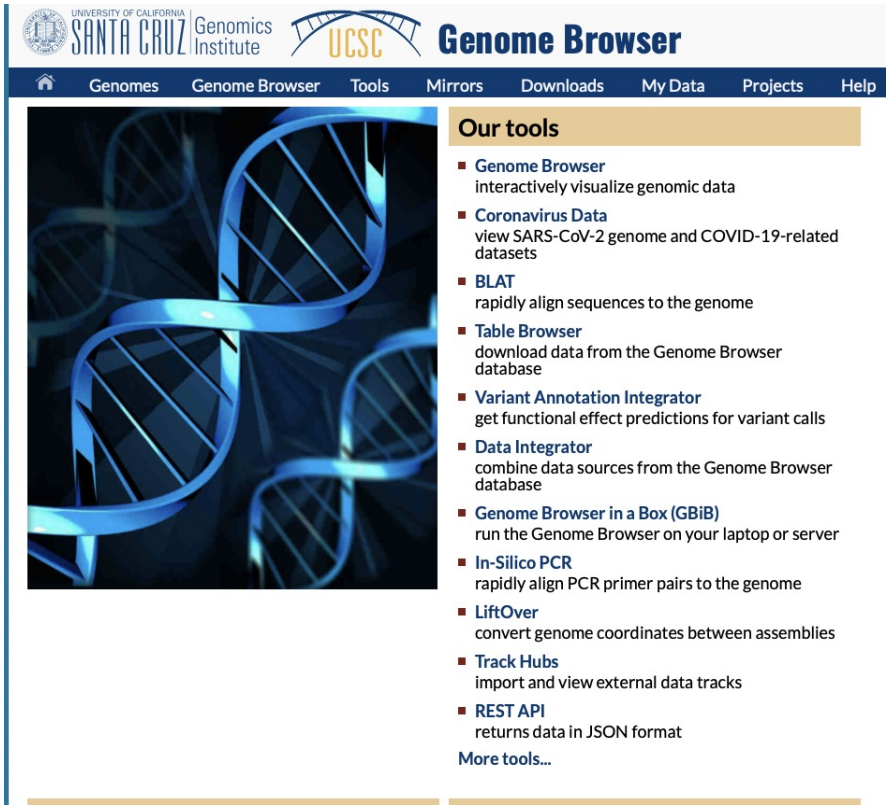
1 **Genome Browser**

The browser offers a variety of interfaces that can be used to explore reference genomic data including the current

reference human genome (Hg38), reference RNA expression databases, and SARS-CoV-2 reference genomes and phylogenies.

Goals + Motivation: Molecular biology, genetics, and genomics all revolve around our ability to acquire, annotate, modify, and compare the nucleic acid sequences that form genomes and transcriptomes. The UCSC Genome Browser is a valuable tool to identify and extract genomic sequences of interest, research genes and their function, and identify contextual genomic features. *This is a building block towards independent design of many molecular experiments.*

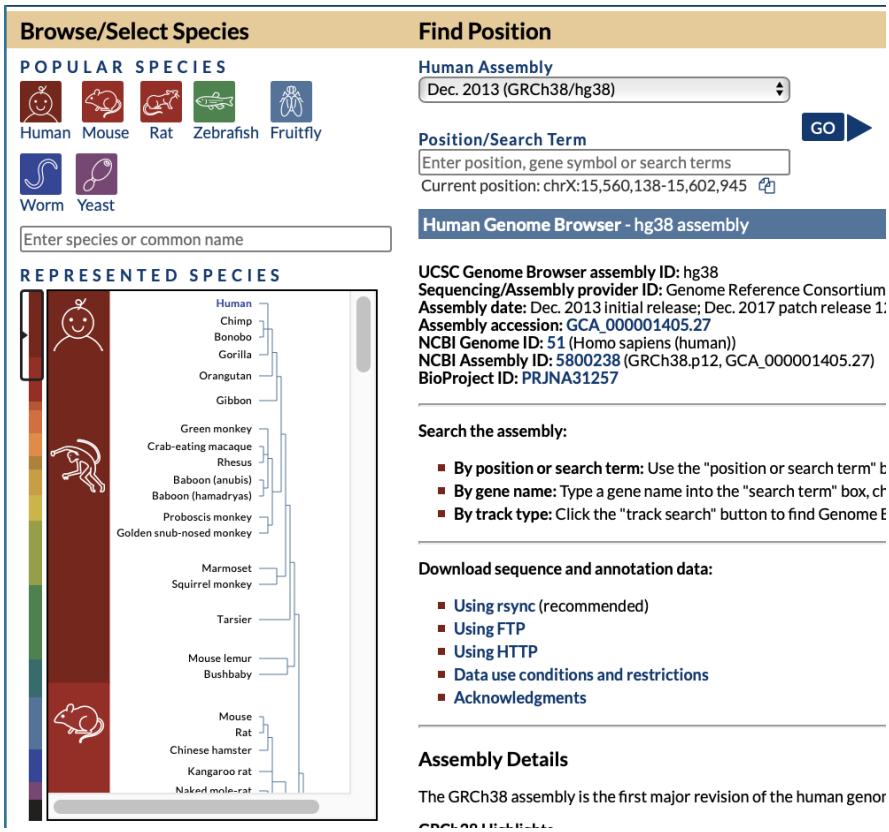
To get started, navigate to genome.ucsc.edu. You'll land at the launch page that links to a variety of tools. This protocol focuses on the ***Genome Browser*.** Follow that link.



2 Now we can select a reference genome to interact with. Notice the tremendous amount of information contained within the Genome Browser repertoire. Dozens of organisms from multiple clades. There are a few versions of the Human Reference Genome, most of them are now legacy versions that lack modern annotations and structure. The two commonly used versions are Hg19 (2009) and Hg38 (2013). Near ~2013-2016 or so, there were probably good reasons to use Hg19 (well understood variation, thoroughly annotated, dominated publications) but at this point, the Hg38 version and its subsequent patches reflect the most advanced and well annotated reference genome available.

If you're interested in human genome sequencing and how we build references, check out the work done by the Nanopore group (Mark, Miten, and Hugh!) alongside Benedict Paten, Karen Miga, and others at UCSC on building the next reference genome.

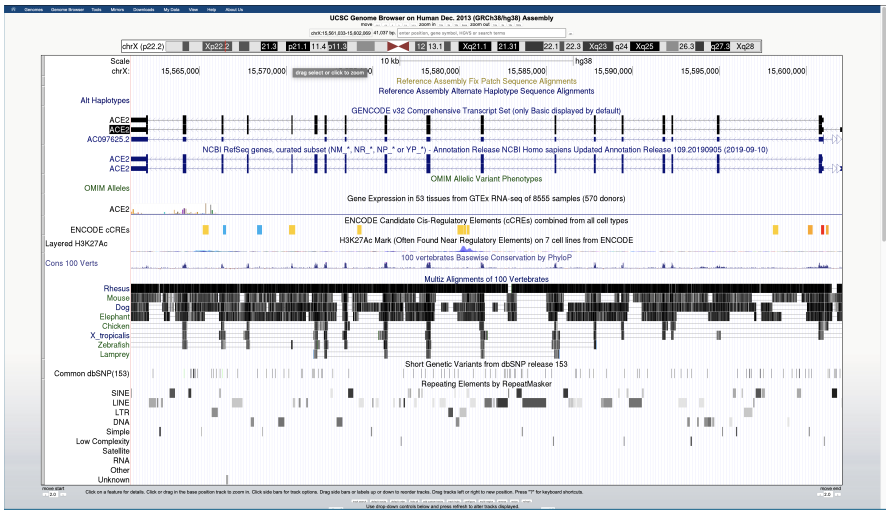
For this protocol, we'll the use the human Gene **ACE2** as our starting point, enter that and hit **GO**



3 Welcome to the genome browser! At first glance this is a ton of information not all of which are relevant to you (yet!). The information in the genome browser is organized in **tracks**, the horizontal slices overlaid atop the vertical blue reference background. In order to help you interpret this, imagine the horizontal plane (x axis) and the genome sequence.

0: Notice the top bar detailing the Chromosome (X in ACE2's case) and position range
1. mouse over the browser interface to highlight individual tracks
2. click and drag to reorder them
3. (cmd/ctrl) click and drag to select sections and zoom
4. Zoom out a fair bit using the buttons above the search bar, get sense of scale in the genome
5. click and drag left and right to move along the genome in linear space

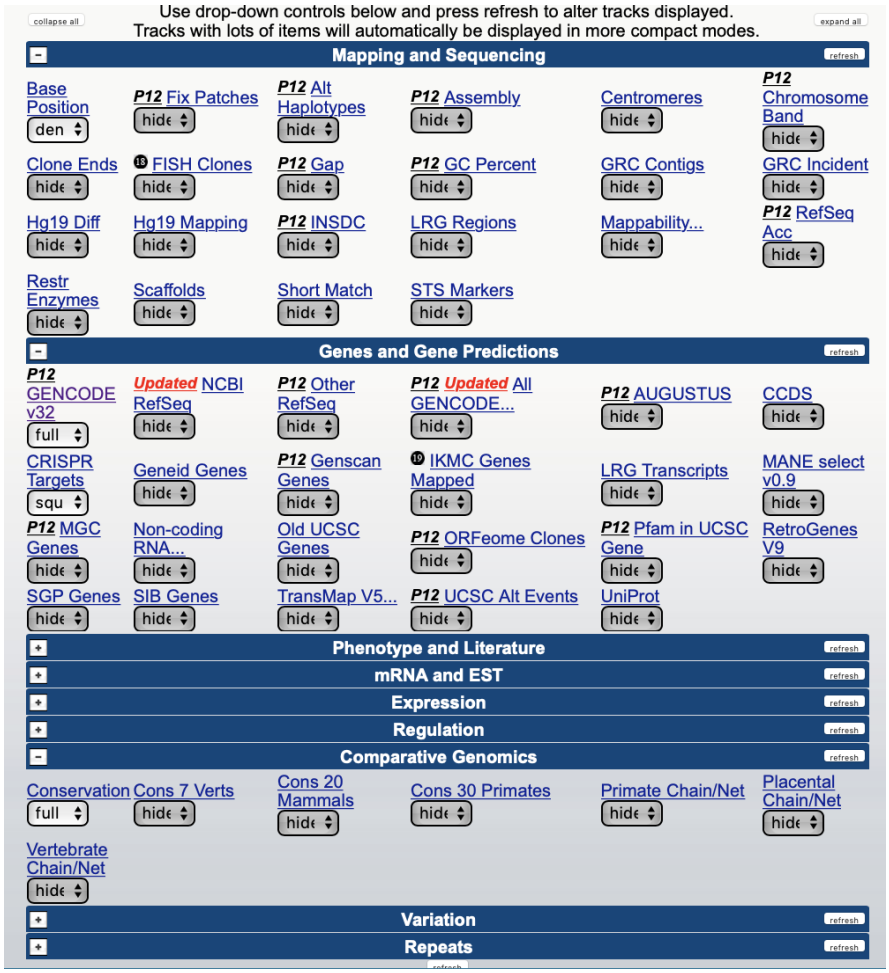
** - If you're unhappy with the size of the text or the browser window go to the top bar menu and select **View > Configure Browser** and use the options at the top to adjust to your liking



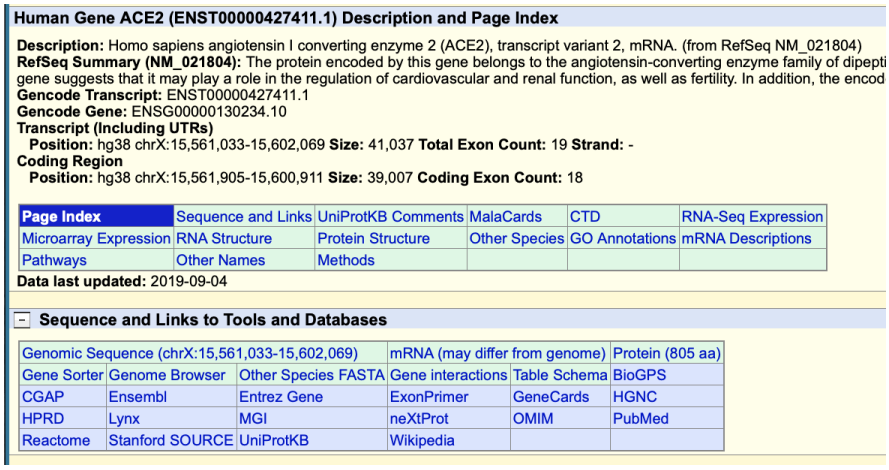
4 The first I like to do when I'm using the browser is create a friendly interface for myself to streamline the information I need to gather. We can pick and choose which tracks to display, and at what detail, at any given time. For now, I suggest we use just a couple:

1. **Mapping and Sequencing**: Base Position (full or pac)
2. **Genes and Gene Predictions**: GENCODE v32 (full), CRISPR Targets (squish to browse, full to use)
3. **Comparative Genomics**: Conservation (full)

From here, you can easily click the **left grey bars** aside each track and make changes to the tracks and investigate their schema.



5 Welcome to **your** genome browser! As you get more comfortable navigating the browser and start asking more questions of it, you now have the basic guid to adding/dropping tracks to suit your interests. Let's explore the main tracks we've got, starting with the **GENCODE** annotation. **GENCODE** identifies genes that are transcribed from the genome, giving us a layer of information that details the functional nature of the genomic space we're looking at. Clicking on the **gene name** on the left end of the browser will take you to a comprehensive collection of data on the gene:



This is where we can start to interact with one of the pillars of molecular biology: **sequences**. There are 3 sequences available from the GENCODE reference reflecting the central dogma biology: genomic (exon, intron, utr, your choice), mRNA (spliced, no introns), protein (translate mRNA sequence). There are also a plethora of resources describing function, structure, relationship to disease and phenotype, and tissue specific abundance.

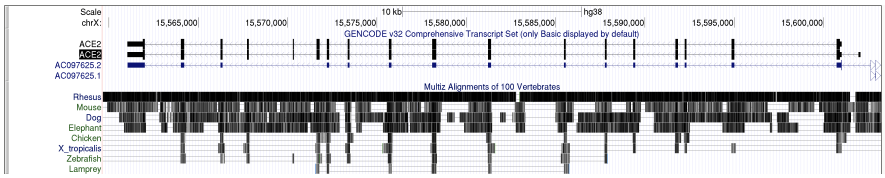
1. look at the genomic sequence of ACE2 **excluding** UTRs and introns
2. break the genomic sequence into exonic subsets, take a look

Now unfortunately, whole gene sequences are not always useful. We're often interested in small, functional regions of genes. Breaking the sequences into exons can help enable more granular analysis of the gene.

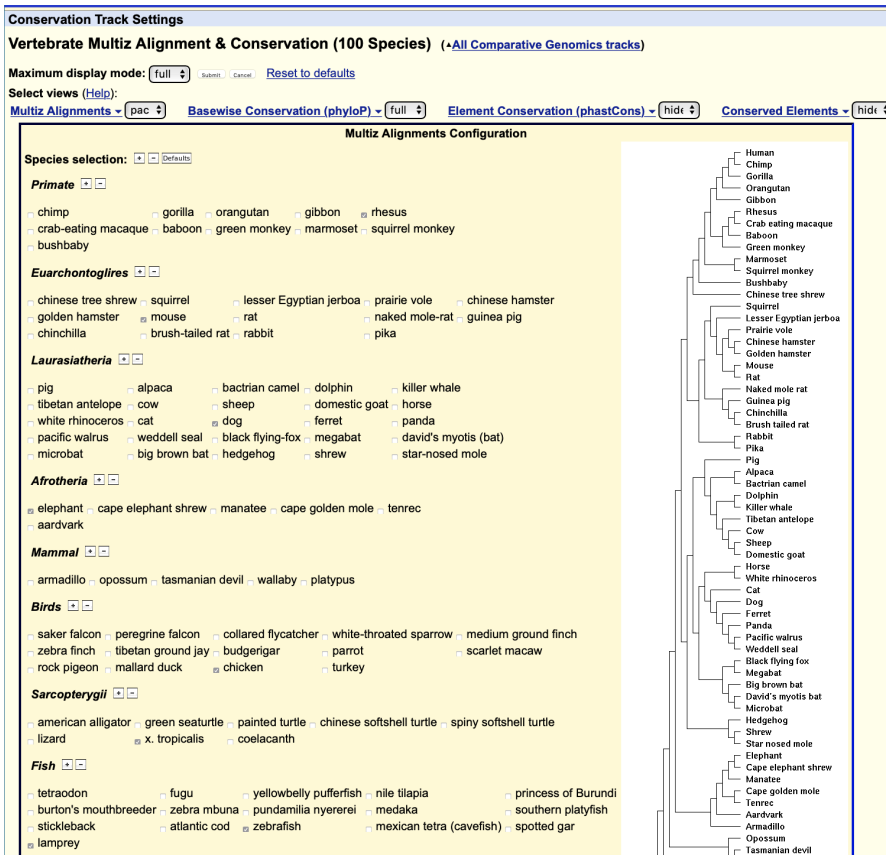
Scroll down a bit, check out the rest of the info:

1. Which tissue has the most detectable ACE2 expressed?
2. What diseases are associated with ACE2?
3. What organisms have orthologs (of those listed)?

6 **Conservation** is a key measurement sequence content that helps us infer both the importance and the tractability of studying certain genes. The display used by the genome browser uses solid blocks to represent strongly similar sequence in the same genomic context, and more lightly colored blocks to represent more distantly similar sequence content.

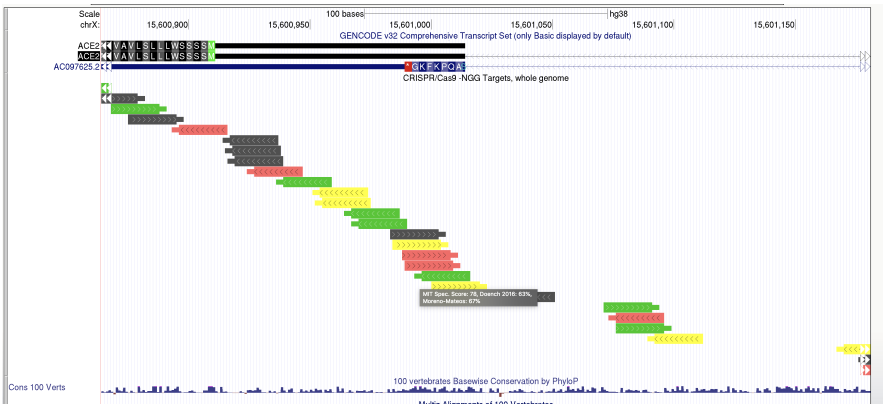


As you might expect, exons are much more conserved in comparison to the non coding intronic regions. We can also add more species or subtract some to better fit our interests:

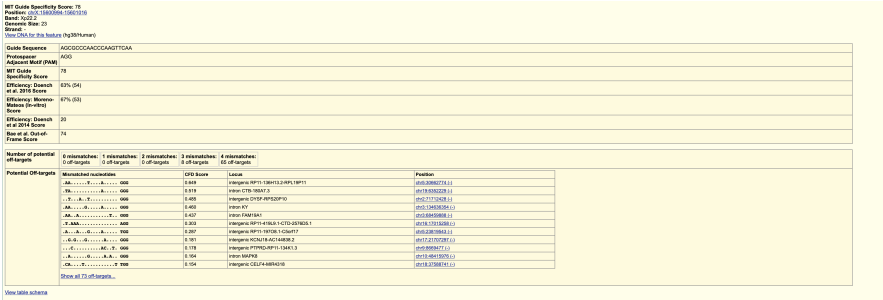


7 Now, speaking of sequences, let's think about designing a CRISPR experiment. The details of CRISPR will be explored further later on in the course, but suffice to say you can't just target the CRISPR machinery *anywhere* in the genome; it requires 1) a protospacer-adjacent-motif (PAM) that it can recognize, and 2) a guide RNA that isn't non-specific. The Genome Browser has a tool built in that annotates all potential CRISPR sites in the genome and scores them based on

a prediction of their specificity and performance. The CRISPR track looks pretty useless when zoomed out, so **zoom all the way in to the transcription start site of ACE2.** If you then set the CRISPR track to 'full', you should see something like this:



Each colored bar represents a potential guide RNA sequence that would direct the CRISPR machinery to operate at the site it terminates at. If we follow the green guide RNA that overlaps with the start site of the top isoform, we can get a detailed summary of predicted CRISPR performance.



The first thing reported? The sequence! This would enable you to order this sequence synthesized as a guide RNA and to deploy it in a CRISPR enabled system.

Sequence Alignment: BLAST

8 NCBI BLAST

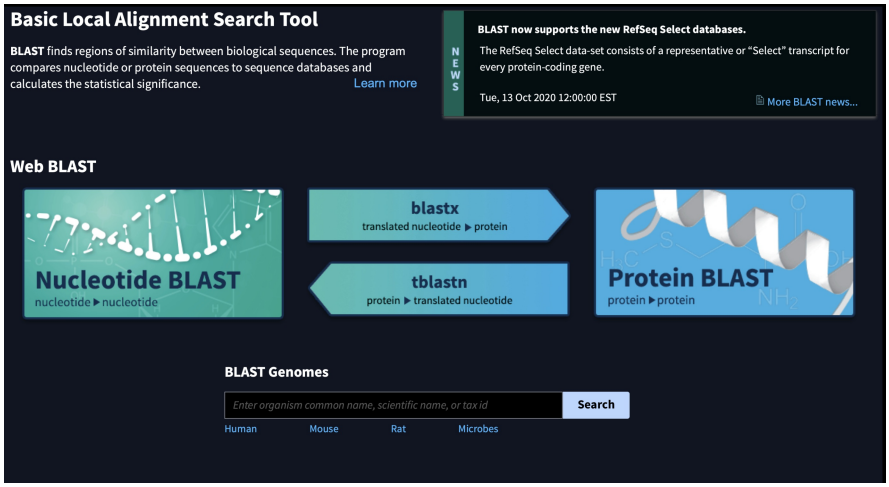
The National Center for Biotechnology Information (NCBI) hosts a robust sequence alignment platform as a browser tool *similar* to the Genome Browser but with a slightly more focused platform. The tools hosted on the website are all implementations of alignment **algorithms**, BLAST{P,N,X,etc} are all tuned to work optimally for the sequences provided: protein, nucleic acid, etc. The Basic Local Alignment Tool (BLAST) implements an algorithmic procedure commonly referred to as **Seed and Extend**. The principle: find subsets of sequence (words or *kmers*), search for these words in many other larger sequences of potential interest, extend in either direction from the matching word and score the similarity of the *local* sequence:

```
# sequence
ACTAGTGTACTGATCG

# words/kmers
ACT
 CTA
  TAG
   AGT
    GTG
     .....

# match the seed in another sequence
TGTACTGTACTGATCACTGATC
```

```
  < --- ACT --- >

# score each local base
        ACTAGTGTACTGATCG
        ACT**T********
TGTACTGTACTGATCACTGATC
   ACT****T**T*A**G
   ACTAGTGTACTGATCG
```

Goals + Motivation: Sometimes you have sequences you know the origin of, maybe you found them on the genome browser. Sometimes, you have sequences that you don't recognize, maybe don't align to Hg38, and remain un-aligned. BLAST allows us to perform a "cheap" (in terms of time and computational resources) alignment to a variety of references. It can help us classify unknown sequences, and do a lot of the stuff that the genome browser also handles like phylogenies, conservation, relationships. I usually like to think of BLAST as a tool to figure out where sequences come from, be it an organism, a plasmid, a virus. There are far more robust alignment algorithms for more targeted, global alignment problems that we will discuss later in the course.

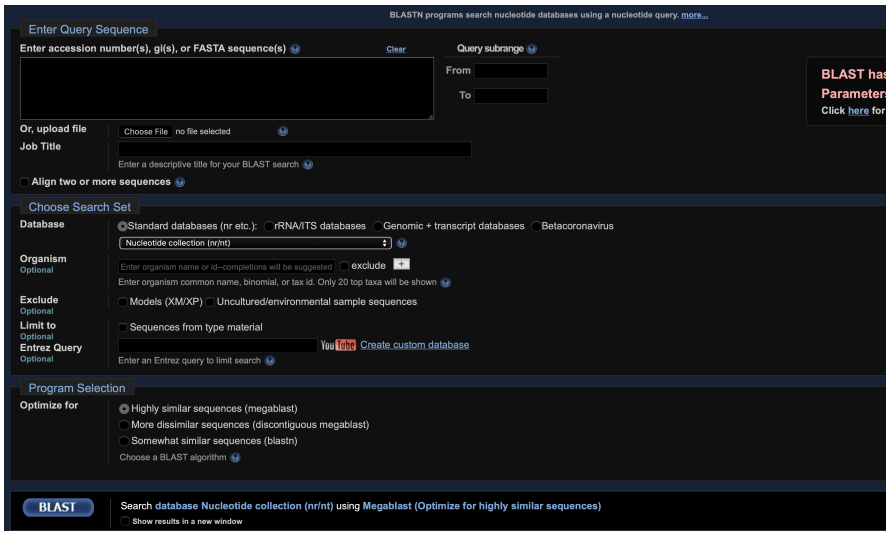
head to the NCBI website and notice how many options there are!! Alignment is not a one-size-fits-all problem, different sequences have different modalities of similarity: amino acids have varying similarities in terms of charge, size, etc and this is accounted for in **BLASTP** using a **BLOSUM** matrix



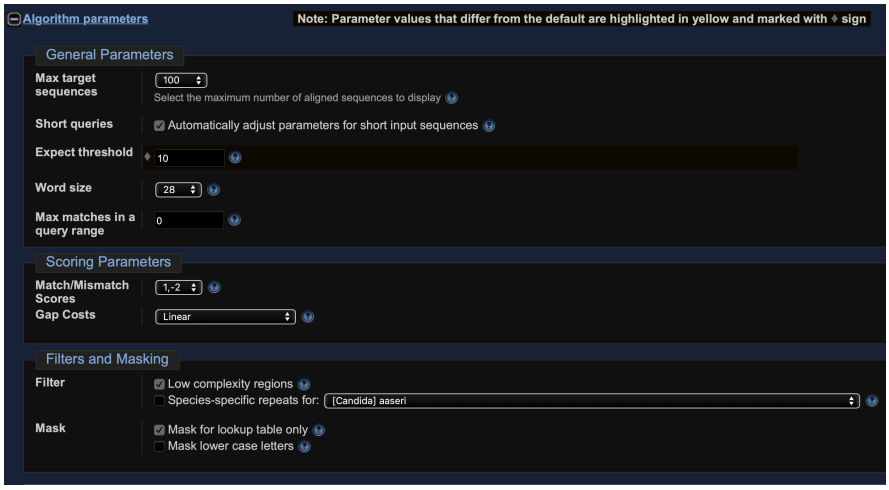
9 Head into the Nucleotide BLAST (**BLASTN**), the main interface you'll use is the sequence entry and reference selection. This is the meat of the tool, and you can get reliable results *for most situations* by just pasting in your sequence and hitting BLAST

of note:

1. The database will determine **where** you search for matching sequences. You can exclude organisms, search only patented sequences, only RNA sequences, so on and so forth. *This will probably be the most useful set of parameters to optimize for a given problem*



10 However, you're not always doing something straightforward...and in those cases you can adjust the **parameters** of the algorithm to optimize them for a specific queries



However, it is important to remember that these parameters are optimized for the vast majority of alignments and adjust according to the input sequence provided

11 Let's query some sequences:

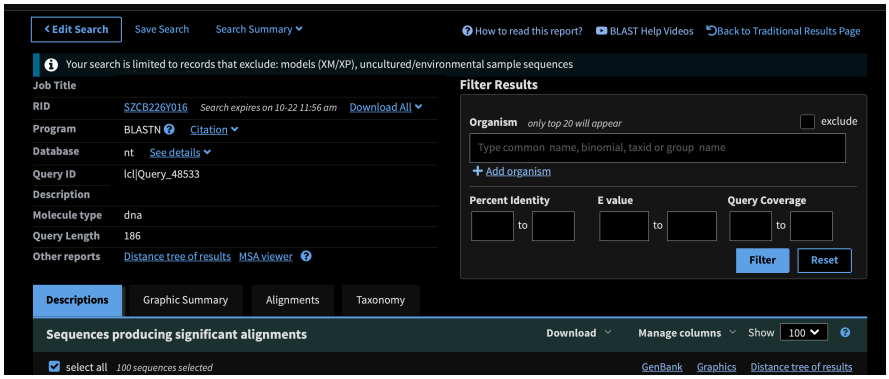
* this is FASTA format, the header is delimited by '>' and until we hit another '>', everything below is considered sequence belonging to that header *

for now, copy these in one at a time to make understanding the output a little easier

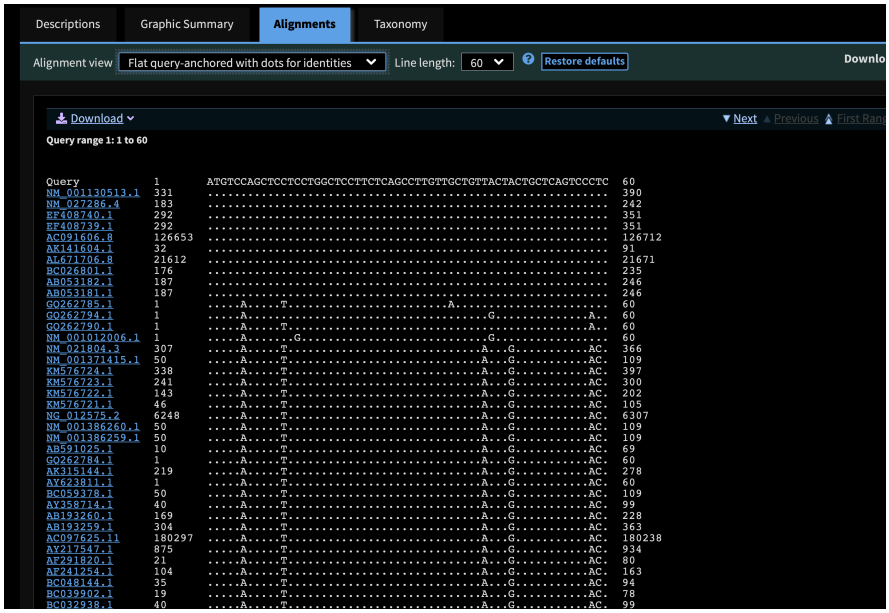
```
>seq_1
ATGTCCAGCTCCTCCTGGCTCCTTCTCAGCCTTGTTGCTGTTACTACTGC
TCAGTCCCTCACCGAGGAAAATGCCAAGeACATTTTTAAACAACTTTAATC
AGGAAGCTGAAGACCTGTCTTATCAAAGTTCACTTGCTTCTTGGAATTAT AATACTAACATTACTGAAGAAAATGCCCAAAAGATG

>seq_2
ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTT
TGGTGGACCCTCAGATTCAACTGGCAGTAACCAGAATGGAGAACGCAGTG
GGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCG
TCTTGGTTCACCGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTCCC
```

```
TCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAA

>seq_3
ATGTCAAGCTCTTCCTGGCTCCTTCTCAGCCTTGTTGCTGTAACTGCTGC
TCAGTCCACCATTGAGGAACAGGCCAAGACATTTTTGGACAAGTTTAACC
ACGAAGCCGAAGACCTGTTCTATCAAAGTTCACTTGCTTCTTGGAATTAT AACACCAATATTACTGAAGAGAATGTCCAAAACATG
```

after each successful search, check out your results. They'll be ranked by **E Value** which is a statistical measurement of the likelihood that an alignment would occur given many random sequences thus lower E value == better confidence in accurate alignment. The main thing I want you to look at here are the organisms producing the alignments, you'll often get a TON of results and may want to use the filter functionality at the top.



One of my favorite ways to view the alignments is with the matrix alignment view provided under the **Alignments** tab and selected with **Flat query-anchored with dots for identities**. It gives a nice overview of how your query matches up with the database sequences and where variance starts to show up

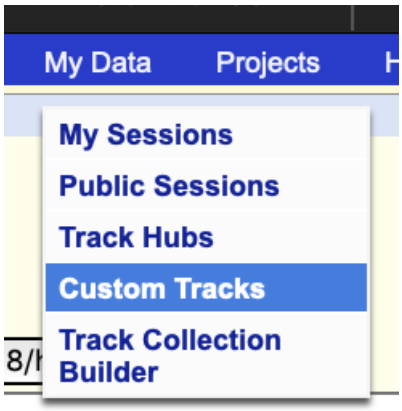


Notice that run of A's ... probably a sign that we're aligning to a different organism. You can find out by following the links!

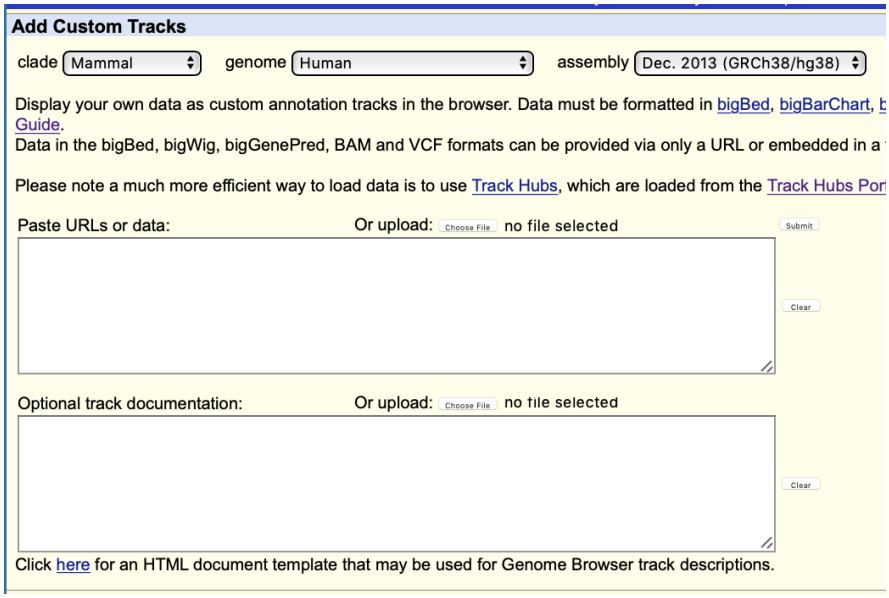
UCSC Genome Browser as an interface -- alignments

12 We've explored data and references hosted by the genome browser but it can also help you contextualize your own data. In this case, we're going to continue thinking about **alignment**: I'm hosting some **RNA-sequencing** data that you can load into the browser and visualize. The reason this works is that we've taken the RNA detected in the

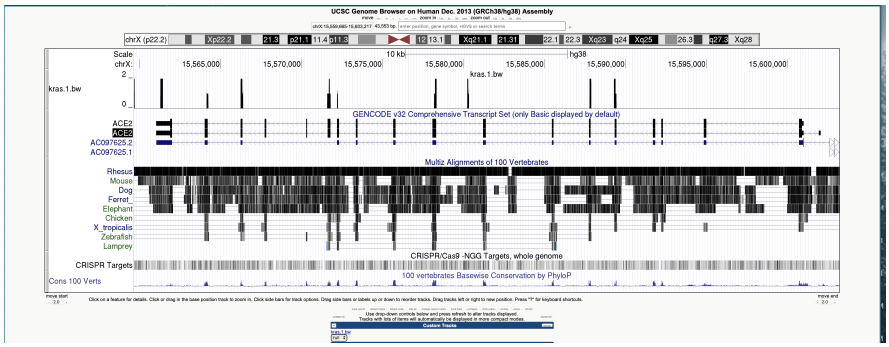
sequencing data and **aligned** it to Hg38. The procedure is a little different than BLAST, but the principle is the same, we will discuss this in more detail later in the quarter. In order to link the data, we're going to create a **custom track** by navigating to the top menu bar and following **My data > Custom Tracks**



from here we can paste in the link to the hosted **bigWig** file. A **bigWig** file is a special format that takes alignment information (position of the alignment in the genome, number of alignments at the position) and makes it usable in the browser. It allows us to look directly at the abundance of reads at specific locations -- primarily exons within genes link: http://public.gi.ucsc.edu/~rreggiar/kras.1.bw --> place it into the first box and hit submit (top right corner)



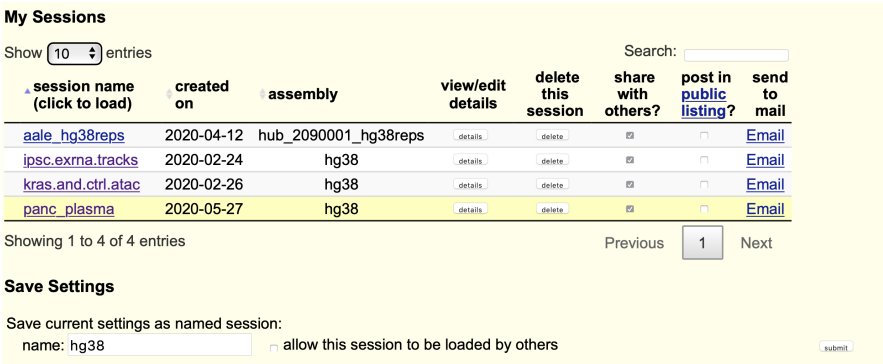
This will bring you back to the main browser window, you'll have a new track option and will see a new track at the top of the window, set it to **full** to see **peaks** corresponding to **aligned sequences** detected in the **RNA sequencing data**



For reference, these are lung epithelial cells...don't seem to have much endogenous ACE2 expression. Check out a hallmark gene like ACTB1 or KRAS -- notice how the scale changes dramatically to capture the different abundances of the mRNA in the dataset.

13 WRAP UP

If you want to save your sweet genome browser setup, head to the top bar **My Data > My Sessions**. From there, you can save your setup as a **session** that you can return to again and again (notice I have a few different sessions that I use for various projects)



Conclusion: These protocols should provide you with a fundamental toolkit to explore genomic sequence content, identify the origin of nucleic acid and protein sequences, and import sequencing data into the UCSC genome browser. These are essential **building blocks** to working on molecular biology projects; integrating genomic sequences and context in your processes will enable you to start engineering experiments and apply the wet lab tools you acquire in this course.