



Oct 21, 2020

# UCSC\_Genome\_Browser\_and\_BLAST\_protocol

Forked from [UCSC\\_Genome\\_Browser\\_and\\_BLAST\\_protocol](#)

1, 2

<sup>1</sup>revamped to go into more detail one GB and BLAST; <sup>2</sup>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-bnppmdk6>

## FORK FROM

Forked from [UCSC\\_Genome\\_Browser\\_and\\_BLAST\\_protocol](#),

## LICENSE

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

## CREATED

Oct 21, 2020

## LAST MODIFIED

Oct 21, 2020

## PROTOCOL INTEGER ID

43469

## DISCLAIMER:

### DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to [protocols.io](#) is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with [protocols.io](#), can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

## 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

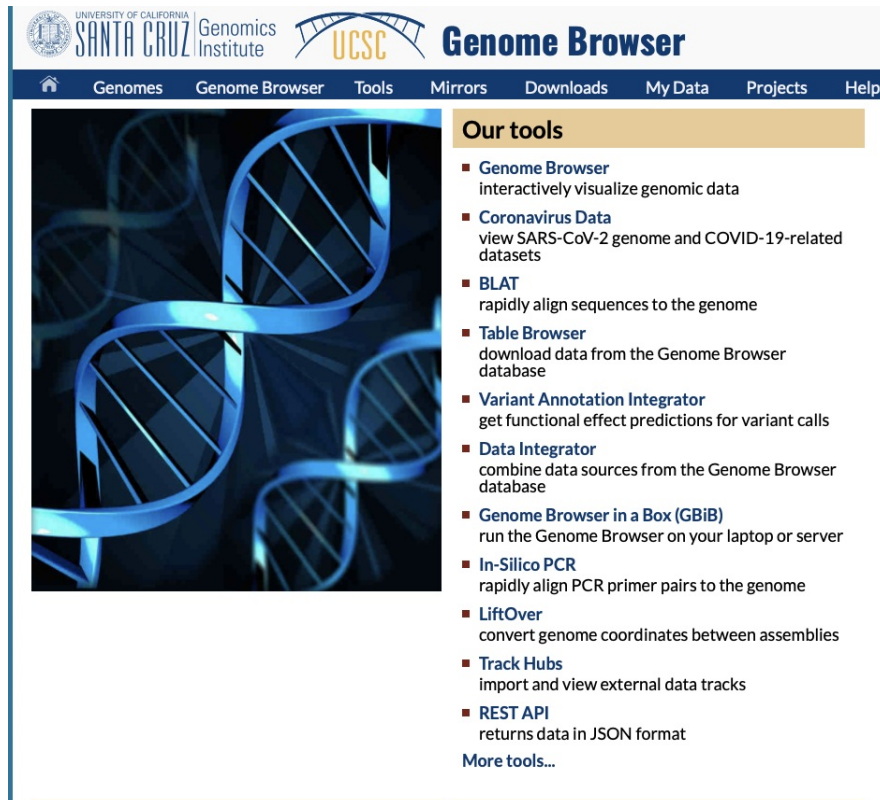
### 1 Genome Browser

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](https://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 use the human Gene **ACE2** as our starting point, enter that and hit **GO**

### Browse/Select Species

**POPULAR SPECIES**

Human
 Mouse
 Rat
 Zebrafish
 Fruitfly

Worm
 Yeast

**REPRESENTED SPECIES**

### Find Position

**Human Assembly**  
Dec. 2013 (GRCh38/hg38)

**Position/Search Term**  
Enter position, gene symbol or search terms  
Current position: chrX:15,560,138-15,602,945

**Human Genome Browser - hg38 assembly**

UCSC Genome Browser assembly ID: hg38  
Sequencing/Assembly provider ID: Genome Reference Consortium  
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 1:  
Assembly accession: [GCA\\_000001405.27](#)  
NCBI Genome ID: 51 (Homo sapiens (human))  
NCBI Assembly ID: 5800238 (GRCh38.p12, GCA\_000001405.27)  
BioProject ID: [PRJNA31257](#)

**Search the assembly:**

- By position or search term: Use the "position or search term" box
- By gene name: Type a gene name into the "search term" box, click
- By track type: Click the "track search" button to find Genome Browser tracks

**Download sequence and annotation data:**

- Using rsync (recommended)
- Using FTP
- Using HTTP
- Data use conditions and restrictions
- Acknowledgments

**Assembly Details**

The GRCh38 assembly is the first major revision of the human genome

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.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

**Mapping and Sequencing**

Base Position (den) hide

P12 Fix Patches hide

P12 Alt Haplotypes hide

P12 Assembly hide

Centromeres hide

P12 Chromosome Band hide

Clone Ends hide

FISH Clones hide

P12 Gap hide

P12 GC Percent hide

GRC Contigs hide

GRC Incident hide

Hg19 Diff hide

Hg19 Mapping hide

P12 INSDC hide

LRG Regions hide

Mappability... hide

P12 RefSeq Acc hide

Restr Enzymes hide

Scaffolds hide

Short Match hide

STS Markers hide

**Genes and Gene Predictions**

P12 GENCODE v32 (full) hide

Updated NCBI RefSeq hide

P12 Other RefSeq hide

P12 Updated All GENCODE... hide

P12 AUGUSTUS hide

CCDS hide

CRISPR Targets (squ) hide

Geneid Genes hide

P12 Genscan Genes hide

IKMC Genes Mapped hide

LRG Transcripts hide

MANE select v0.9 hide

P12 MGC Genes hide

Non-coding RNA... hide

Old UCSC Genes hide

P12 ORFeome Clones hide

P12 Pfam in UCSC Gene hide

RetroGenes V9 hide

SGP Genes hide

SIB Genes hide

TransMap V5... hide

P12 UCSC Alt Events hide

UniProt hide

**Phenotype and Literature**

mRNA and EST

Expression

Regulation

**Comparative Genomics**

Conservation Cons 7 Verts (full) hide

Cons 20 Mammals hide

Cons 30 Primates hide

Primate Chain/Net hide

Placental Chain/Net hide

Vertebrate Chain/Net hide

**Variation**

**Repeats**

- 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:

**Human Gene ACE2 (ENST00000427411.1) Description and Page Index**

**Description:** Homo sapiens angiotensin I converting enzyme 2 (ACE2), transcript variant 2, mRNA. (from RefSeq NM\_021804)

**RefSeq Summary (NM\_021804):** The protein encoded by this gene belongs to the angiotensin-converting enzyme family of dipepti gene suggests that it may play a role in the regulation of cardiovascular and renal function, as well as fertility. In addition, the encod

**Gencode Transcript:** ENST00000427411.1

**Gencode Gene:** ENSG00000130234.10

**Transcript (Including UTRs)**

**Position:** hg38 chrX:15,561,033-15,602,069 **Size:** 41,037 **Total Exon Count:** 19 **Strand:** -

**Coding Region**

**Position:** hg38 chrX:15,561,905-15,600,911 **Size:** 39,007 **Coding Exon Count:** 18

Page Index	Sequence and Links	UniProtKB Comments	MalaCards	CTD	RNA-Seq Expression
Microarray Expression	RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions
Pathways	Other Names	Methods			

Data last updated: 2019-09-04

**Sequence and Links to Tools and Databases**

Genomic Sequence (chrX:15,561,033-15,602,069)	mRNA (may differ from genome)	Protein (805 aa)
Gene Sorter	Genome Browser	Other Species FASTA
CGAP	Ensembl	Entrez Gene
HPRD	Lynx	MGI
Reactome	Stanford SOURCE	UniProtKB
		Wikipedia
	Gene interactions	Table Schema
	ExonPrimer	GeneCards
	neXtProt	OMIM
		PubMed
		BioGPS
		HGNC

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.

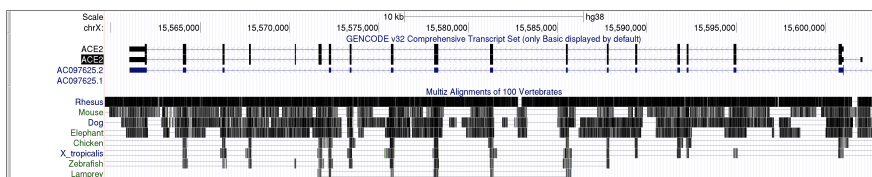
1. look at the genomic sequence of ACE2 **excluding** UTRs and introns
2. break the genomic sequence into exonic subsets, take a look

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:

**Conservation Track Settings**

**Vertebrate Multiz Alignment & Conservation (100 Species)** (+All Comparative Genomics tracks)

Maximum display mode:

Select views (Help):

Multiz Alignments

**Multiz Alignments Configuration**

Species selection:

**Primate**

chimp gorilla orangutan gibbon rhesus  
crab-eating macaque baboon green monkey marmoset squirrel monkey  
bushbaby

**Euarchontoglires**

chinese tree shrew squirrel lesser Egyptian jerboa prairie vole chinese hamster  
golden hamster mouse rat naked mole-rat guinea pig  
chinchilla brush-tailed rat rabbit pika

**Laurasiatheria**

pig alpaca bactrian camel dolphin killer whale  
tibetan antelope cow sheep domestic goat horse  
white rhinoceros cat dog ferret panda  
pacific walrus weddell seal black flying-fox megabat david's myotis (bat)  
microbat big brown bat hedgehog shrew star-nosed mole

**Afrotheria**

elephant cape elephant shrew manatee cape golden mole tenrec  
aardvark

**Mammal**

armadillo opossum tasmanian devil wallaby platypus

**Birds**

saker falcon peregrine falcon collared flycatcher white-throated sparrow medium ground finch  
zebra finch tibetan ground jay budgerigar parrot scarlet macaw  
rock pigeon mallard duck chicken turkey

**Sarcopterygii**

american alligator green sea turtle painted turtle chinese softshell turtle spiny softshell turtle  
lizard x. tropicalis coelacanth

**Fish**

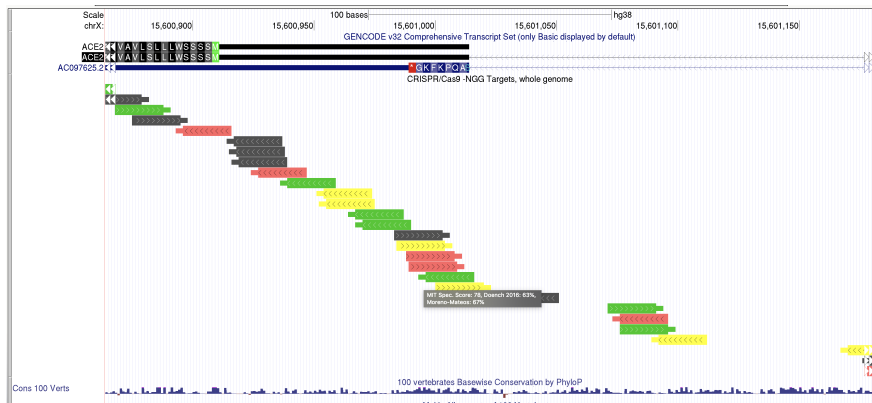
tetraodon fugu yellowbelly pufferfish nile tilapia princess of Burundi  
burton's mouthbreeder zebra mbuna pundamilia nyererei medaka southern platyfish  
stickleback atlantic cod zebrafish mexican tetra (cavefish) spotted gar  
lamprey

**Phylogenetic Tree**

Human  
Chimp  
Gorilla  
Orangutan  
Gibbon  
Rhesus  
Crab-eating macaque  
Baboon  
Green monkey  
Marmoset  
Squirrel monkey  
Bushbaby  
Chinese tree shrew  
Squirrel  
Lesser Egyptian jerboa  
Prairie vole  
Chinese hamster  
Golden hamster  
Mouse  
Rat  
Naked mole rat  
Guinea pig  
Chinchilla  
Brush-tailed rat  
Rabbit  
Pika  
Pig  
Alpaca  
Bactrian camel  
Dolphin  
Killer whale  
Tibetan antelope  
Cow  
Sheep  
Domestic goat  
Horse  
White rhinoceros  
Cat  
Dog  
Ferret  
Panda  
Pacific walrus  
Weddell seal  
Black flying fox  
Megabat  
Big brown bat  
David's myotis bat  
Microbat  
Hedgehog  
Shrew  
Star-nosed mole  
Elephant  
Cape elephant shrew  
Manatee  
Cape golden mole  
Tenrec  
Aardvark  
Armadillo  
Opossum  
Tasmanian devil

- 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.

[illegible]

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
```



```

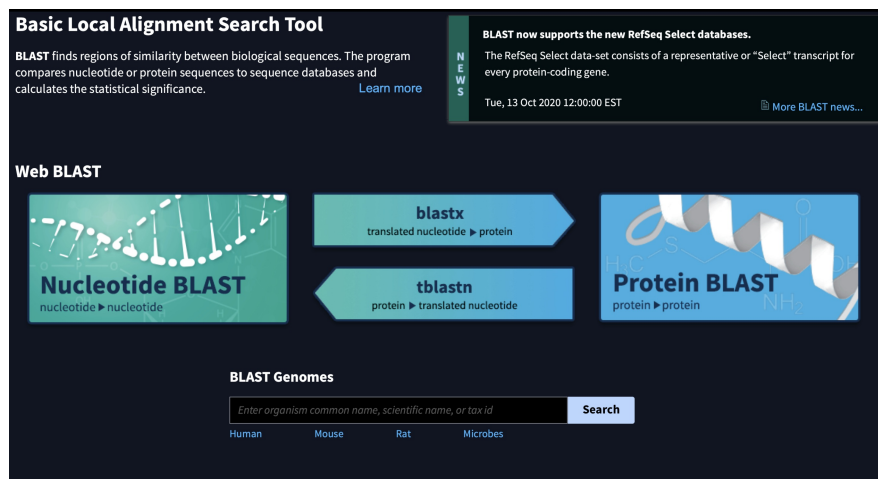
< --- ACT --- >

# score each local base
  ACTAGTGACTGATCG
  ACT**T*****
TGTACTGTACTGATCACTGATC
  ACT****T**T*A**G
  ACTAGTGACTGATCG

```

**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](https://www.ncbi.nlm.nih.gov/) 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
ATGTCCAGCTCCTCTGGCTCCTTCTCAGCCTTGTTGCTGTTACTACTGC
TCAGTCCCTCACCAGGAAAATGCCAAGACATTTTTAAACAACCTTTAATC
AGGAAGCTGAAGACCTGTCTTATCAAAGTTCACCTGCTTCTTGAATTAT AATACTAACATTACTGAAGAAAATGCCAAAAGATG

>seq_2
ATGTCTGATAATGGACCCAAAATCAGCGAAATGCACCCCGCATTACGTT
TGGTGGACCTCAGATTCAGTGGCAGTAACGAGAAATGGAGAACGCAGTG
GGGCGGATCAAAACAACGTCGGCCCCAAGTTTACCAATAATACTGCG
TCTTGGTTCACCGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTCCC
```

```
TCGAGGACAAGGCGTTCCAATTACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAA
```

```
>seq_3
ATGTCAAGCTCTTCTGGCTCCTTCTCAGCCTTGTGCTGTAAGTCTGCTGC
TCAGTCCACCATTGAGGAACAGGCCAAGACATTTTGGACAAGTTTAACC
ACGAAGCCGAAGACCTGTTCTATCAAAGTTCACCTGCTTCTTGAATTAT 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.

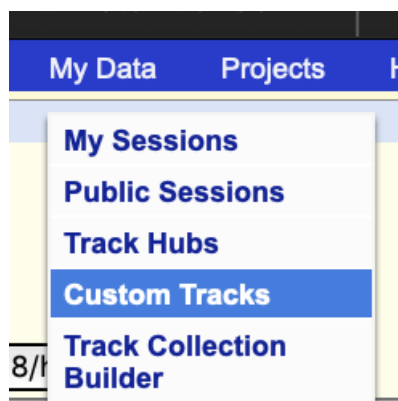
One of my favorite ways to view the alignments is with the matrix alignment view provided under the **Alignments** tab and selected with **Flat query-anchored with dots for identities**. It gives a nice overview of how your query matches up with the database sequences and where variance starts to show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

- 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)

**Add Custom Tracks**

clade  genome  assembly

Display your own data as custom annotation tracks in the browser. Data must be formatted in [bigBed](#), [bigBedChart](#), [bigWig](#), [bigWigChart](#), [BAM](#) or [VCF](#). See the [Data Format Guide](#).

Data in the bigBed, bigWig, bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a [track](#).

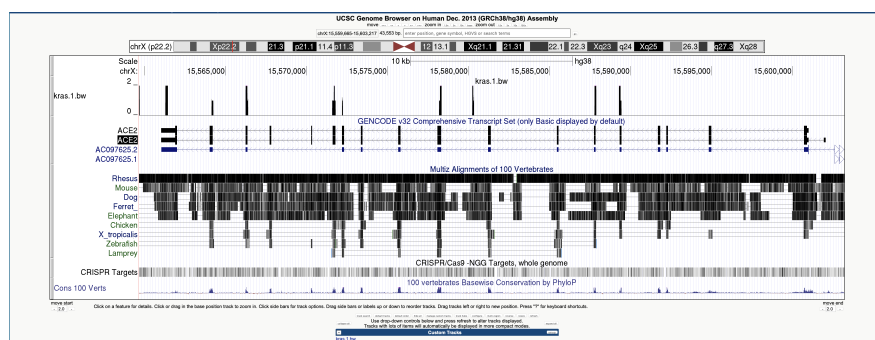
Please note a much more efficient way to load data is to use [Track Hubs](#), which are loaded from the [Track Hubs Portal](#).

Paste URLs or data:  Or upload:  no file selected

Optional track documentation:  Or upload:  no file selected

Click [here](#) for an HTML document template that may be used for Genome Browser track descriptions.

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)

**My Sessions**

Show  entries Search:

session name (click to load)	created on	assembly	view/edit details	delete this session	share with others?	post in public listing?	send to mail
<a href="#">aale_hg38reps</a>	2020-04-12	hub_2090001_hg38reps	<a href="#">details</a>	<a href="#">delete</a>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<a href="#">Email</a>
<a href="#">ipsc.exrna.tracks</a>	2020-02-24	hg38	<a href="#">details</a>	<a href="#">delete</a>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<a href="#">Email</a>
<a href="#">kras.and.ctrl.atac</a>	2020-02-26	hg38	<a href="#">details</a>	<a href="#">delete</a>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<a href="#">Email</a>
<a href="#">panc_plasma</a>	2020-05-27	hg38	<a href="#">details</a>	<a href="#">delete</a>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<a href="#">Email</a>

Showing 1 to 4 of 4 entries Previous  Next

**Save Settings**

Save current settings as named session:

name:  ☐ allow this session to be loaded by others [submit](#)

**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.