Using the UCSC genome browser
Credits

Terry Braun

Mary Mangan, Ph.D.

www.openhelix.com
UCSC Genome Browser Credits

Development team: http://genome.ucsc.edu/staff.html

Led by David Haussler and Jim Kent
Dozens of staff and students bring you this software and data

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

Human Genome: Acknowledgements

The UCSC Human Genome Browser Project is conducted in collaboration with the International Human Genome Project. We have worked with many collaborators at other institutions to produce and annotate this reference sequence of the human genome, as described in our recent publications. The UCSC Human Genome Project is made possible by a grant from the National Human Genome Research Institute, directed by Francis Collins, and generous support from the Howard Hughes Medical Institute and the California Institutes for Science and Innovation.
What is a genome, what is a build

• Very critical to get right
• The genome (human) is not done revisions happen
  – If you use a different version you will get different results
• Chr1:123456-123654 seems pretty specific
  – Why are you not happy with that?
UCSC FAQ

• [http://genome.ucsc.edu/FAQ/FAQreleases](http://genome.ucsc.edu/FAQ/FAQreleases)
• Comparison of UCSC and NCBI human assemblies

• Question:
• "How do the human assemblies displayed in the UCSC Genome Browser differ from the NCBI human assemblies?"

• Response:
• Recent human assemblies displayed in the Genome Browser (hg10 and higher) are identical to the NCBI assemblies.
Issues

Question:
- "I noticed that the chromosomal coordinates for a particular gene that I'm looking at have changed since the last time I used your browser. What happened?"

- Response:
- A common source of confusion for users arises from mixing up different assemblies. It is very important to be aware of which assembly you are looking at. Within the Genome Browser display, assemblies are labeled by organism and date. To look up the corresponding UCSC database name or NCBI build number, use the release table."
Question:
"I am confused about the start coordinates for items in the refGene table. It looks like you need to add "1" to the starting point in order to get the same start coordinate as is shown by the Genome Browser. Why is this the case?"

Response:
Our internal database representations of coordinates always have a zero-based start and a one-based end. We add 1 to the start before displaying coordinates in the Genome Browser. Therefore, they appear as one-based start, one-based end in the graphical display. The refGene.txt file is a database file, and consequently is based on the internal representation.

We use this particular internal representation because it simplifies coordinate arithmetic, i.e. it eliminates the need to add or subtract 1 at every step. Unfortunately, it does create some confusion when the internal representation is exposed or when we forget to add 1 before displaying a start coordinate. However, it saves us from much trickier bugs.
Basic use of the genome browser
Customizing your view
Visual Cues on the Genome Browser

Tick marks; a single location (STS, SNP)

3' UTR  exon  exon  exon  ex  5' UTR

Intron and direction of transcription <<< or >>>

Track colors *may* have meaning—for example, UCSC Gene track:
- If there is a corresponding PDB entry = black
- If there is a corresponding reviewed/validated seq = dark blue
- If there is a non-RefSeq seq = lightest blue

Vert. cons.

height of a blue bar is increased likelihood of conservation,
red indicates a likelihood of faster-evolving regions

Alignment indications (Conservation pairs: “chain” or “net” style)
- Alignments = boxes, Gaps = lines
Basic Annotation Track Menus Defined

- **Hide**: removes a track from view
  
  - **Dense**: all items collapsed into a single line
  
  - **Squish**: each item = separate line, but 50% height + packed
  
  - **Pack**: each item separate, but efficiently stacked (full height)

- **Full**: each item on separate line (may need to zoom to fit)
Super Tracks
• Search for data types
• Reset to defaults
• Configure options page
• You control the views with numerous features
• Flip 5` to 3`
Using the Get DNA

- Say you have a region of interest and what to make primers....
- You can use the View->DNA link to get the DNA sequence of the region you are looking at...
Get Sequence from Description Pages

Sequence and Links to Tools and Databases

Genomic Sequence (chr17:7,571,720-7,590,868)  mRNA (may differ from genome)  Protein (354 aa)

<table>
<thead>
<tr>
<th>Schema</th>
<th>BioGPS</th>
<th>CGAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td>Gepis Tissue</td>
<td>H-INV</td>
</tr>
<tr>
<td>Pfam</td>
<td>PubMed</td>
<td>Reactome</td>
</tr>
</tbody>
</table>

>uc002gij.3 (TP53) length=354

MDDMLSPDDLEQQWFTDPSEPPEAFMPEAAPPVAPAAPAPFPAPAPSWPLESSSV
QKTYQQSYGFRLGFQLGTAQTCYSPALMNLCQALTCPDQLWVSTPFPGRTVR
MAIKQKSMQMTETVVRAPCRHCRSDDGALAPQHLIRVEGNLIREYLDNRNTFHRHSVY
YEPEVGSDDTTYHHYMYCNSSSCMHGAINRPIILTIITTEDSSGNLLGRNFEVRCAC
RDRRTENRJKGEPHHELPPGSTKRALPNNTESSPQFEEKFLDGEYFTLQIRGRE
MFRELNEALELKDQAGKEPGSSRKHSSHLKSKKQGSTSRHKLMKGTEOYPSD
Using the E-PCR

• So you designed the primers.
• Will they work?
• Will they amplify what you want?
• Tools -> Insilico PCR
Blat: Tools-> Blat

• An alignment tool similar to BLAST from FAQ:

  • “Blat of DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments”.

  • “On proteins, Blat uses 4-mers rather than 11-mers, finding protein sequences of 80% and greater similarity to the query of length 20+ amino acids. However, BLAST and psi-BLAST at NCBI can find much more remote matches.”
POP quiz

• What gene is this peptide from? (hint: human)
• MDEPPFSEAAALEQALGEPCDLDAALLTDIEDMLQLINNQDSDFPGLFDPPPYAGSGAGGTD
• What gene is 5` of this gene?
• What gene is 3` of this gene?
  – What tissue is this gene (the gene 3` to the peptide) highly expressed in?
Converting from one build to another:

• Tools->Liftover
  – BED files?
  – Command line?

• OK how about
  – View -> In other Genomes
  – chr17:38496978-38530994 (hg18)
    • Find it in hg38
Using the gene sorter

- From Homepage click GeneSorter
- Similarity types
- Filtering criteria
- Ordering
- Exporting results
VisiGene
Using the table browser

- Powerful way to pull out specific sets of data.
- Can be used for further visualization
- Pulling out subset of sequences
- Pulling out tabular data about the genome
Underlying Database (MySQL)

visualize

search & download
Table example

• Find all coding SNPs in the region: chr17:41243452-41277468
• Filter based on the codon you want to work with.
Creating your own tracks

• Open notepad

• [https://genome.ucsc.edu/FAQ/FAQformat.html](https://genome.ucsc.edu/FAQ/FAQformat.html)

• Enter:

```bash
browser position chr7:127471196-127495720 browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration"
visibility=2 itemRgb="On"
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
```
Capturing output

- Screen capture
- View->PDF/PS
- Export sequences to another program
UCSC Exercise #1

• Pull up SAA2 in the genome browser
• Make sure you have at least "UCSC Genes" and "RefSeq Genes" turned on – I like "Pack"
• Zoom out until you can see SAA1, SAA2, and SAA4 (hint: they are close by)
• Knowing what we know about the similarity between SAA1 and SAA2 – what inference could we make about the origins of SAA1 and SAA2?
• Turn on "Alt Events" (under "Gene and Gene Prediction Tracks" – after reading the description)
• What does "retained Intron" annotation from the "Alt Events" track mean?
• You can also turn on "SIB Alt-Splicing" under mRNA and EST tracks – to see alternatively spliced ESTs that have been observed
• Finally – see if you can find SAA3 – and view it in the same window as SAA1, SAA2 and SAA4
• What do you think about the origins of SAA3? What is SAA3?
UCSC Exercise #2

- The primers Kischner reported in the paper to amplify this exon were:
  - CTGTGATCCTAACGCAAGAC
  - TAGAAATCACATCATAGCAC
  - Can you find this exon? How?
  - Be sure to turn on the "Alt Events" track in "Gene and Gene Prediction Tracks"
UCSC Exercise #3

• Get all regions that have a simple repeated sequence
  – *GCGGC*
  – That overlap a gene
  – View your track in the genome browsers
  – Download as a bed file