James Madison University

From the SelectedWorks of Ray Enke Ph.D.

July, 2016

Research & Course-based Applications of the UCSC Genome Browser

Raymond A Enke

Available at: https://works.bepress.com/raymond_enke/58/
Research & Course-based Applications of the UCSC Genome Browser

Ray Enke Ph.D.
James Madison University
Department of Biology
2016 CGEMS Summer Workshop
July 21, 2016
The Impact of “Next Generation” Sequencing

Cost per Genome

Completion of Human Genome Project (2001)

Moore's Law

Source: NHGRI
The Impact of “Next Generation” Sequencing

Cost per Genome

- 10K fold reduction in sequencing cost over 4 yrs (2007-2011)

Source: NHGRI
The Impact of “Next Generation” Sequencing

Roche 454 (2007)  
Illumina (2009)  
Ion Torrent (2010)  
Nanopore MinION (2015)

“Next Gen” Sequencing Platforms

“3rd Generation!”

• Reduced cost of Next gen sequencing platforms have made genome-wide analysis available to the masses

• Undergraduate training in genomics analysis is “critical to the Nation’s health and economy” (NSF IUSE grant solicitation)
Infusing Genomics Analysis into Primarily Undergraduate Institutions (PUIs)...
...and high school education too!

Barriers to integrating genomics analysis into PUI curriculum:
- Cost prohibitive to generate data sets
- Complex bioinformatics analysis & advanced computing capabilities
- Limited instructor resources for training
What is a Genome Browser?

- graphical interface for display of aggregated genomic data
- enable researchers to visualize & browse entire genomes with annotated data
  - gene expression data, mutation data, protein data, etc
Commonly Used Genome Browsers

**Ensembl:**
- European Bioinformatics Institute
- Plants, animals, fungi, protist, prokaryotes

**NCBI:**
- US-funded National Center for Biotechnology Information
- Plants, animals, fungi, protist, prokaryotes

**UCSC Genome Browser:**
- Hosted by University of California, Santa Cruz
- Vertebrates & model organisms only
What we’re not covering today: Ensembl Genome Browsers
Plant, protist, fungi, & bacterial genomes

Ensemble training tutorials: http://uswest.ensembl.org/info/website/tutorials/index.html

- http://plants.ensembl.org/index.html
- http://protists.ensembl.org/index.html
- http://fungi.ensembl.org/index.html
Commonly Used Genome Browsers

**Ensembl:**
- European Bioinformatics Institute
- Plants, animals, fungi, protist, prokaryotes

**NCBI:**
- US-funded National Center for Biotechnology Information
- Plants, animals, fungi, protist, prokaryotes

**UCSC Genome Browser:**
- Hosted by University of California, Santa Cruz
- Vertebrates & model organisms only
- User friendly interface; applications for novices – bioinformaticians
1. **Basic features & beginner activities**
   - basic searches, configuring data, saving & sharing sessions
   - examples of course-embedded application

2. **Creating custom data tracks & intermediate activities**
   - visualizing & integrating your data with browser data
   - examples of research applications

3. **Custom data tracks advanced activity**
   - visualizing & integrating data from the literature with browser data
The UCSC Genome Browser homepage: [http://genome.ucsc.edu/](http://genome.ucsc.edu/)

- Curated by the University of California, Santa Cruz
- Research grade bioinformatics tools
- Intuitive user friendly interface - excellent tutorials available
- Many preloaded data sets
- Custom data builds available
- Link to training modules
Open Helix training modules:
http://www.openhelix.com/ucsc

- 6 training module videos
- ~75 minutes to complete
- excellent overview of basic browser features
Before you begin: create a free account to save & share browser sessions

My Data >>> My Sessions:
- set up free account
- save & return to sessions
- share sessions
- make public sessions
UCSC Genome Browser Homepage
http://genome.ucsc.edu/

- Link to gateway page
- Choose your genome
- Choose genome assembly
- Search for gene
Choose genome, assembly, & genomic region

- 240 eukaryotic genomes
- Vertebrates and model organisms - e.g. C. elegans, yeast, Drosophila
- No proks or plant; try Ensembl browser

Search gene name, chr position, chr#, etc
Different Genomes Have Different Available Tools & Data

- Species may have different data tracks
- Layout, software, & functions are the same
Control what you see in the browser window

**Genome viewer**

**Groups of data (Tracks)**

- Mapping and Sequencing Tracks
- Genes and Gene Prediction Tracks
  - (including sno/miRNA data)
- Phenotype, Disease, Literature Tracks
- mRNA and EST Tracks
- Expression (such as microarray)
- Regulation (including TFBS)
- Comparative Genomics
  - As a group
  - Individual species
- Neandertal, Denisova
- Variation, Repeats
  - (including SNPs, copy number variation)

Many genomes have a ton of available data

Only view what you need (hide all features)
Example browser window: default view

Default view has lots of data
hide all tracks, add back only data you want to view
Example browser window: reconfigured view

- Default view has lots of data
- hide all tracks, add back only data you want to view (e.g. annotated genes)
1. **Basic features & beginner activities** (starts on pg 1-2)
   - Pick 1 of 4 activities
   - work in groups of 2-4 (~15-20 min)
   - answer questions activity questions

2. **share your sessions with class when done** (pg 17)
Pedagogical applications of basic browser features:

- Chapter 4 Fate: Discusses the genetic variants in the Huntingtin (HTT) gene associated with Huntingtin’s disease (Bio 481 Genomics)
Pedagogical applications of basic browser features:

CAG repeats in 5' region of human Huntingtin (HTT) gene

Searchable in UCSC Public Sessions: “Ridley Fate”

- Chapter 4 Fate: Discusses the genetic variants in the Huntingtin (HTT) gene associated with Huntingtin’s disease (Bio 481 Genomics)
Pedagogical applications of basic browser features:

- PTC Bitter Taste Perception genotype/phenotype lab (Bio 480 Advanced Molec Bio)
Pedagogical applications of basic browser features:

![Gene visualization and genetic variation data tracks obtained from the UCSC Genome Browser (a). Gene sequence exported from the browser and annotated in the ApE sequence editing software (b).](image)

**FIG 2**

Student computational analysis of the human TAS2R38 gene. Gene visualization and genetic variation data tracks obtained from the UCSC Genome Browser (a). Gene sequence exported from the browser and annotated in the ApE sequence editing software (b).

Berndsen et al., BAMBED, 2016

- PTC Bitter Taste Perception genotype/phenotype lab (Bio 480 Advanced Molec Bio)
3. Creating custom data tracks & high quality figures from browser sessions (intermediate activity; starts on pg 18)

Yeast EGD2 Gene & Upstream region

- EGD2 is a 1 exon reverse strand coded gene encoding the yeast homolog of the human protein NACA. In humans, NACA associates with the transcription factor BTF2 to form the nascent polypeptide-associated complex (NAC). The transcription factor MET31 has previously been demonstrated to bind to the 5′ regulatory region of EDG2. MET31 binding region, consensus binding motif and sequence are displayed. Annealing sites of primers used for EGD2 mutant genotyping are also shown in red.
Click the “add custom tracks” box
Uploading data for custom track

Custom data can be added in 3 different ways:

1. Copy/paste in “tab separated values” data (TSV format)
2. browse & upload a tab separated data (.tsv) file containing formatted data
3. paste a link to a host URL containing formatted data
## Tab Separated Values (TSV) Data

<table>
<thead>
<tr>
<th>Pos</th>
<th>Name</th>
<th>G</th>
<th>PA</th>
<th>AB</th>
<th>R</th>
<th>H</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Caleb_Joseph</td>
<td>100</td>
<td>355</td>
<td>320</td>
<td>38</td>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>1B</td>
<td>Chris_Davis*</td>
<td>160</td>
<td>670</td>
<td>573</td>
<td>100</td>
<td>150</td>
<td>47</td>
</tr>
<tr>
<td>2B</td>
<td>Jonathan_Schoop</td>
<td>86</td>
<td>321</td>
<td>305</td>
<td>34</td>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>SS</td>
<td>J.J._Hardy</td>
<td>114</td>
<td>437</td>
<td>411</td>
<td>45</td>
<td>90</td>
<td>8</td>
</tr>
<tr>
<td>3B</td>
<td>Manny_Machado</td>
<td>162</td>
<td>713</td>
<td>633</td>
<td>102</td>
<td>181</td>
<td>35</td>
</tr>
<tr>
<td>LF</td>
<td>Steve_Pearce</td>
<td>92</td>
<td>325</td>
<td>294</td>
<td>42</td>
<td>64</td>
<td>15</td>
</tr>
<tr>
<td>CF</td>
<td>Adam_Jones</td>
<td>137</td>
<td>581</td>
<td>546</td>
<td>74</td>
<td>147</td>
<td>27</td>
</tr>
</tbody>
</table>

*led team in HRs

- Text file
- Each line is an entry
- Columns denoted by spaces or tabs
- Use underscores as spacers w/in entry
- Can be any type of data (e.g. sports stats)
- example: 8 lines with 8 columns

---

2015 Baltimore Orioles Stats

![Image of Baltimore Orioles players](image-url)
Genome coordinates: TSV data in BED format

*human genome coordinates for the 5 RHO exons

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>stop</th>
<th>exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr3</td>
<td>129247482</td>
<td>129247937</td>
<td>exon1</td>
</tr>
<tr>
<td>chr3</td>
<td>129249719</td>
<td>129249887</td>
<td>exon2</td>
</tr>
<tr>
<td>chr3</td>
<td>129251094</td>
<td>129251259</td>
<td>exon3</td>
</tr>
<tr>
<td>chr3</td>
<td>129251376</td>
<td>129251615</td>
<td>exon4</td>
</tr>
<tr>
<td>chr3</td>
<td>129252451</td>
<td>129254187</td>
<td>exon5</td>
</tr>
</tbody>
</table>

- BED = Browser Extensible Data
- TSV formatted data, but with 3 defined & 1 optional heading
  1. Chr
  2. Start
  3. Stop
  4. Can be any data (usually a line descriptor)
Custom tracks: 1) copy/paste BED formatted data

- Paste BED data into custom tracks data window
- do not include headers
- Hit submit > hit go
Custom tracks: 1) copy/paste BED formatted data

- Paste BED data into custom tracks data window (do not include headers)
- Hit submit > hit go
Custom tracks: 1. copy/paste BED formatted data

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>stop</th>
<th>exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr3</td>
<td>129247482</td>
<td>129247937</td>
<td>exon1</td>
</tr>
<tr>
<td>chr3</td>
<td>129249719</td>
<td>129249887</td>
<td>exon2</td>
</tr>
<tr>
<td>chr3</td>
<td>129251094</td>
<td>129251259</td>
<td>exon3</td>
</tr>
<tr>
<td>chr3</td>
<td>129251376</td>
<td>129251615</td>
<td>exon4</td>
</tr>
<tr>
<td>chr3</td>
<td>129252451</td>
<td>129254187</td>
<td>exon5</td>
</tr>
</tbody>
</table>

- Paste BED data into custom tracks data window (do not include headers)
- Hit submit > hit go > zoom out to view full gene
4th column in BED format is whatever you want

- Change descriptor field to whatever you want to see associated with each element in your custom track (e.g. q-value, ID#, rank#, etc)
Edit custom **track name** and **track description**

- Select “manage custom tracks” under display window
- Click into “user track” to edit your custom track
Edit custom **track name** and **track description**

- Change the track name and track description to your liking
- Must have `=` and `'` characters
Edit custom **track name** and **track description**

- Hit submit > hit go > zoom out to view full gene
- Track name & description are now customized
- Customize colors?
Edit custom track color

Update Custom Track: protein coding exonic sequence [hg19]

Update your custom track configuration, data, and/or documentation. Data must be formatted in MAF, narrowPeak, Personal Genome SNP, PSL, or WIG formats. To configure the display, see bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a track.

Edit configuration:

track name='exonic_seq' description='protein coding exonic sequence' color=0,0,255

- Navigate back track editor window
- Add text “color=0,0,255” (this is the RGB value for blue)
  - pure white =255,255,255; pure black =0,0,0
  - pure red=255,0,0; pure green=0,255,0; mix and match to get your fav shade
- Hit submit > hit go > zoom out to view full gene
Edit custom track color

- Navigate back track editor window
- Add text “color=0,0,255” (this is the RGB value for blue)
  - pure white =255,255,255; pure black =0,0,0
  - pure red=255,0,0; pure green=0,255,0; mix and match to get your fav shade
- Hit submit > hit go > zoom out to view full gene
3. Creating custom data tracks (intermediate activity; starts on pg 18)
   - Pick 1 of 4 activities
   - work in groups of 2-4 (~15-20 min)
   - Make a high resolution image in MS PowerPoint of your session (pg 22)

4. Save & share your sessions with class when done (pg 17)
EGD2 is a 1 exon reverse strand coded gene encoding the yeast homolog of the human protein NACA. In humans, NACA associates with the transcription factor BTF2 to form the nascent polypeptide-associated complex (NAC). The transcription factor MET31 has previously been demonstrated to bind to the 5’ regulatory region of EDG2. MET31 binding region, consensus binding motif and sequence are displayed. Annealing sites of primers used for EGD2 mutant genotyping are also shown in red.
ICD-1 is a 5 exon reverse strand coded gene encoding the *C. elegans* homolog of the human transcription factor BTF3. Arsenovic *et al.*, 2012 demonstrated that icd-1 mutations result in misfolded protein stress in the ER triggering the unfolded protein response (UPR). Annealing sites of primers used for ICD-1 mutant genotyping are shown in royal blue.
CFTR is a 26 exon sense strand coded gene encoding the Cystic Fibrosis Transmembrane Regulator which is a trans membrane protein chloride channel. Mutations in the CFTR gene are responsible for the syndromic disease cystic fibrosis (CF). The Arg117His point mutation resulting in a single amino acid change in the and ΔF508 deletion mutation
Research applications of custom tracks features:

- Custom ChIP-seq & RNA-seq data, PCR sites integrated with UCSC data
5. Advanced user activity (pg 23)
- visualizing & integrating data from the literature with browser data

CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors
Joseph C. Corbo, Karen A. Lawrence, Marcus Karlstetter, Connie A. Myers, Musa Abdelaziz, William Dirkes, Karin Weigelt, Martin Seifert, Vladimir Benes, Lars G. Fritsche, Bernhard H.F. Weber and Thomas Langmann

* These authors contributed equally to this work.
5. Advanced user activity (pg 23)
- visualizing & integrating data from the literature with browser data
Thanks!

contact: enkerajmu.edu
Twitter: @Enke_Lab

2016 CGEMS Summer Workshop

2016 CGEMS Summer Workshop

2016 CGEMS Summer Workshop

2016 CGEMS Summer Workshop