Visualize mapped data at UCSC genome browser

UCSC Genome Browser tracks

The UCSC Genome Browser is an invaluable resource both for obtaining public sequencing data and for visualizing it.

Tip Sometimes the UCSC Genome Browser at http://genome.ucsc.edu/ is pretty slow -- after all, it's a resource shared among the Eukaryotic genomics community. But there's also a second "Beta test" version of the browser at http://hgwdev.cse.ucsc.edu/. It has slightly newer (and possibly less stable) code, but fewer people use it.

- http://genome.ucsc.edu/ **Genome Browser**, submit
- navigaion
    - type GAPDH in gene box, **jump**
    - note zoom out/zoom in buttons; click on position or click/drag
- track detail
    - click "Simply Nucleotide Polymorphisms (dbSnp build 130)" to expand track detail
    - click on one of the SNP to expand track detail
        - then click on the snp name to see details
- selecting/hiding tracks
    - under "Regulation" section, change "ENCODE Regulation" track from "show" to "hide", **refresh**
    - right click "Multiz Alignments", **hide**
    - under "Phenotype and Disease Association" change GWAS Catalog from "hide" to "squish", **refresh**
- type PRNP in gene box, **jump**.
    - click on "NHGRI Catalog..." track description to expand detail
    - note correspondence between SNPs (SNP 132) and disease SNPs (GWAS)
    - click on one of the disease SNPs for detail

Configuring custom tracks

The UCSC Genome Browser has a "Custom Tracks" feature that lets you visualize your data using the Genome Browser web application. This data is visible only to you, not publically (unless you choose to share a link to it with others).

There are two approaches to visualizing your data in the UCSC Genome Browser:

1. Directly upload a data file, in one of the supported formats.
    - Your data is copied over the Internet to UCSC, where it is stored in tables and displayed as you browse.
    - Appropriate for small to medium size files (up to a few MB).
2. Host your data locally, and configure the UCSC Genome Browser with its URL.
    - Your data resides in a location accessible via an HTTP or FTP public URL (e.g., our **/corral-repl/utexas/BioITeam/web** directory). No data is copied to UCSC. You only tell the browser where to find the data when it is needed.
    - Appropriate for large data sets (e.g. BAM files) that can be indexed for fast retrieval.

BED data

BED format is a simple 3 to 9 column format for location-oriented data.

See supported data formats for custom tracks for more information and examples.

VCF data

VCF data can only be configured as a URL, not uploaded directly. Directions are found at http://genome.ucsc.edu/goldenPath/help/vcf.html.

- The VCF file must be sorted by chromosome and position (most tools produce VCFs like this).
- The VCF file must be compressed using **bgzip**:

```
module load tabix  # also loads bgzip
cd $BI/web
bgzip progeria_ctcf.vcf
```

- The VCF file must be indexed using **tabix**:

```
tabix -p vcf progeria_ctcf.vcf.gz
```

This has already been done, and the resulting files are at this URL: http://loving.corral.tacc.utexas.edu/bioiteam/ucsc_custom_tracks/, filename **progeria_ctcf.vcf.gz**. These are hg18 SNP calls from published Iyer Lab CTCF ChIP-seq data in Progeria cells. The VCF file was produce using Broad's GATK.

- Add custom tracks (be sure to pick assembly March 2006, NCBI36/hg18)
- Here is the track configuration line

```
track type=vcfTabix name="progeria_ctcf_snp_calls" bigDataUrl="http://loving.corral.tacc.utexas.edu
/bioiteam/ucsc_custom_tracks/progeria_ctcf.vcf.gz"
```

BAM data

BAM data can only be configured as a URL, not uploaded directly. Directions are found at http://genome.ucsc.edu/goldenPath/help/bam.html.

- The BAM file must be sorted and indexed using **samtools**. The .bam and .bai index file must reside in the same directory.

This has already been done, and the resulting files are at this URL: http://loving.corral.tacc.utexas.edu/bioiteam/ucsc_custom_tracks/, filename **hela_totrna.sorted.bam**. This is SE RNAseq data mapped directly to the human genome, hg19.

- Add custom tracks (be sure to pick assembly Feb 2009, NCBI37/hg19)
- Here is the track configuration line

```
track type=bam name="hela_rnaseq" bigDataUrl="http://loving.corral.tacc.utexas.edu/bioiteam
/ucsc_custom_tracks/hela_totrna.sorted.bam"
```

Here is another example, using paired end RNAseq data as processed using a **tophat/cufflinks** pipeline:

```
track type=bam name="rnaseq_bam" pairEndsByName=Y bigDataUrl="http://loving.corral.tacc.utexas.edu/bioiteam
/ucsc_custom_tracks/accepted_hits.sorted.bam"
```

Downloading annotation data

For RNAseq you often need a GTF file, but how do you find them? One way is to download annotations from the UCSC Table browser in GTF format:

- http://genome.ucsc.edu/cgi-bin/hgTables
  - clade: Mammal, genome: Human, assembly: **hg19**
  - group: Genes and Gene Prediction tracks, track: RefSeq genes
  - output format: GTF - gene transfer format
  - optional: enter filename in typein box
  - **get output**

A couple of exercises

Exercise: Altzheimer's disease SNP

Using the UCSC Genome Browser, determine whether Craig Venter or James Watson has a higher risk of Altzheimer's disease.

Hints

APOE gene.

Variation & Repeats, **Genome Variants**

Phenotype & Disease Assocations, **GWAS Catalog**
A solution

Exercise 2

Using the UCSC Genome Browser, find and download a list of high-sequencing-depth regions in BED format.

Hints

group: Mapping and Sequencing tracks

track: Hi Seq Depth

A solution