Lecture 1

Write a python program to organize your data

copy the raw data, don’t rename (or use symlinks)
if you rename stuff have the code do it (never by hand)
document your code + backup (github)
organize data in a way that it is easy to process

You are likely going to find these module useful!

```
import numpy as np
import scipy.stats as stats
import sys
import shutil
import os
import zipfile
```

```
/bio/share/Bioinformatics_Course.tar
```

numpy and scipy.stats

- see crash course in scipy modules link in class notes
- https://docs.scipy.org/doc/scipy-0.18.1/reference/stats.html
- We will return to this in week 10, when we analyze some of our genomic summaries

import sys

main uses

```
# read from stdin
for line in sys.stdin:
        print line
```

sys.stdout & sys.stderr also useful!

```
# get command line arguments
for arg in sys.argv[1:]:
        print arg
```

```
# useful info
sys.path, sys.platform, sys.version
```

import os

```
# execute a shell command
os.system()
# current working directory
os.getcwd()
# return a list of directories
os.listdir(path)
# change directory
os.chdir(path)
#make a directory
os.mkdir(dir)
#path stuff
os.path.basename()
os.path.abspath()
os.path.exists()
```

```
import shutil

#copy files
shutil.copy(source,destination)
#copy files and metadata
shutil.copy2(source,destination)
#copy entire “tree”
shutil.copytree(source, destination)
#the destination folder must already exist!
```

The data

(check out the “tree” command)

```
Bioinformatics_Course
├── ATACseq
│   ├── README.ATACseq.txt
│   ├── Sample_ACCAGCA-CTCCTTAC_4R009_L1_P050_R1.fq.gz
│   ├── Sample_ACCAGCA-CTCCTTAC_4R009_L1_P050_R2.fq.gz
│   ├── Sample_ACCAGCA-TATGCAGT_4R009_L1_P059_R1.fq.gz
│   ├── Sample_ACCAGCA-TATGCAGT_4R009_L1_P059_R2.fq.gz
│   ├── ...
├── DNAseq
│   ├── ADL06_1_1.fq.gz
│   ├── ADL06_1_2.fq.gz
│   ├── ...
│   ├── ADL09_1_1.fq.gz
│   ├── ...
│   └── README.DNA_samples.txt
└── RNAseq
    ├── RNAseq384plex_flowcell01
    │   ├── Demultiplex_Stats.htm
    │   ├── Project_plex1
    │   │   ├── Sample_1
    │   │   │   ├── 1_CACTTGA_L001_R1_001.fastq.gz
    │   │   │   ├── 1_CACTTGA_L001_R2_001.fastq.gz
    │   │   │   └── SampleSheet.csv
    │   │   ├── Sample_10
    │   │   │   ├── 10_GGAATGT_L001_R1_001.fastq.gz
    │   │   │   ├── 10_GGAATGT_L001_R2_001.fastq.gz
    │   │   │   └── SampleSheet.csv
    │   │   ├── ...
    │   ├── Project_plex2
    │   │   ├── Sample_46
    │   │   │   ├── 46_CACTTGA_L002_R1_001.fastq.gz
    │   │   │   ├── 46_CACTTGA_L002_R2_001.fastq.gz
    │   │   │   └── SampleSheet.csv
    │   │   ├── ...
    │   ├── ...
    ├── RNAseq384_README.txt
    ├── RNAseq384_SampleCoding.txt
    └── RNAseq384_SampleCoding.xlsx
```

- check out the readme for sample mappings
- what might be a better way to organize the data?
  - ATACseq and DNAseq by sample name
  - RNAseq perhaps left as is, and map sample names from within DESeq2

Lecture 2

Illumina data

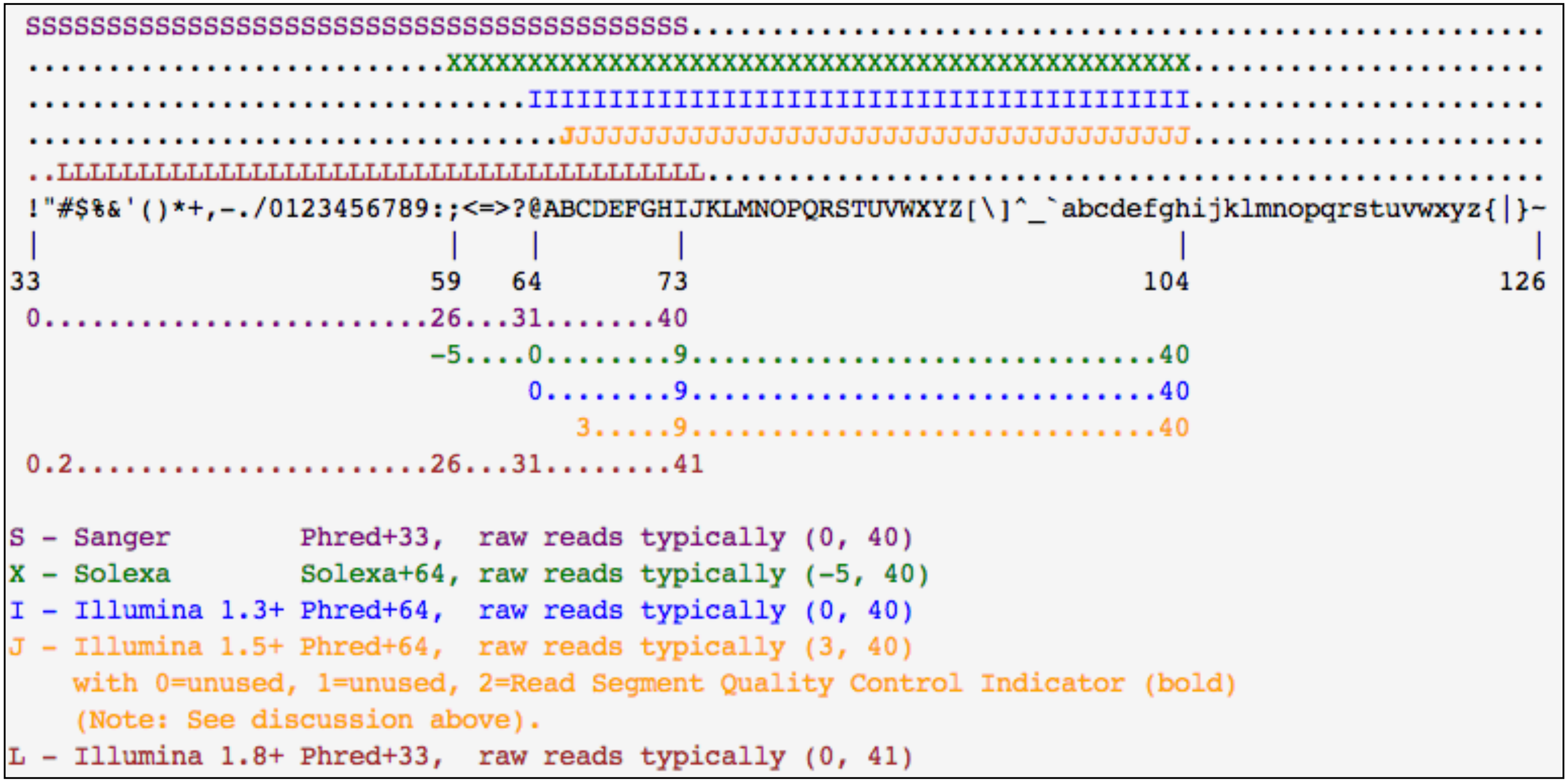
- lots of SE or PE short reads
- current HiSEQ4000 (circa 2016)
  - 400M PE100s = $2.5K
  - 400M PE50s = $1.2K
- “insert size” limited to about 700bp by technology
- for most applications reads mapped to a reference genome (de novo assembly hard because of repeats)
- reads have errors (0.2% per base per read)
- sample being sequenced have SNPs and INDELs relative to reference genome

Error rate

- mismatch relative to reference
- higher at higher cycles (e.g., ends of reads)
- reads have quality scores that tell you confidence in base call
- other problems can occur with library prep
  - contamination
  - adaptory things
  - bad starting template, etc.

raw reads look like ...

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA+
unique_sequence_ID
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4
```

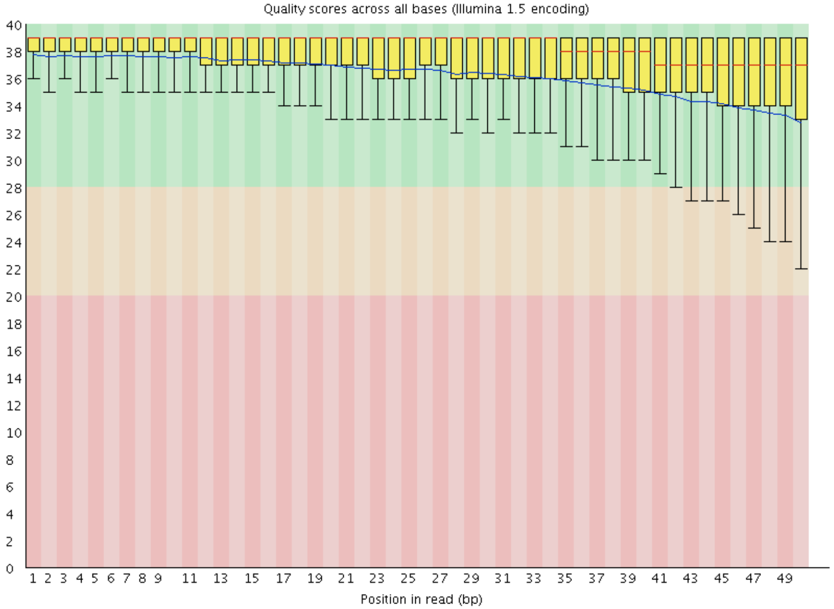


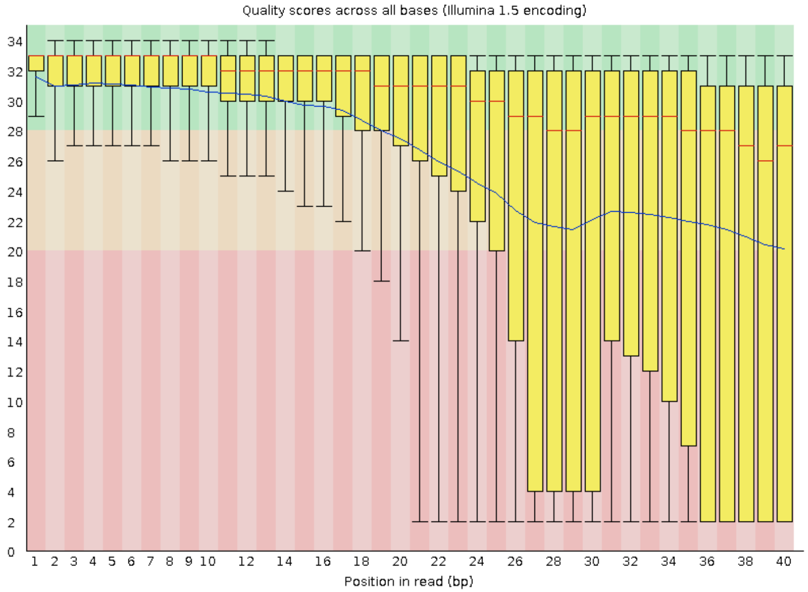
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                       59   64       73                            104                   126
0.........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

http://en.wikipedia.org/wiki/FASTQ_format

fastqc

- difficult to look at and tell if data is good...
- quality scores hard to look at
- fastq file usually gzipped (`zcat blah.fq.gz | head –n 100`)
- fastqc can be helpful for QC

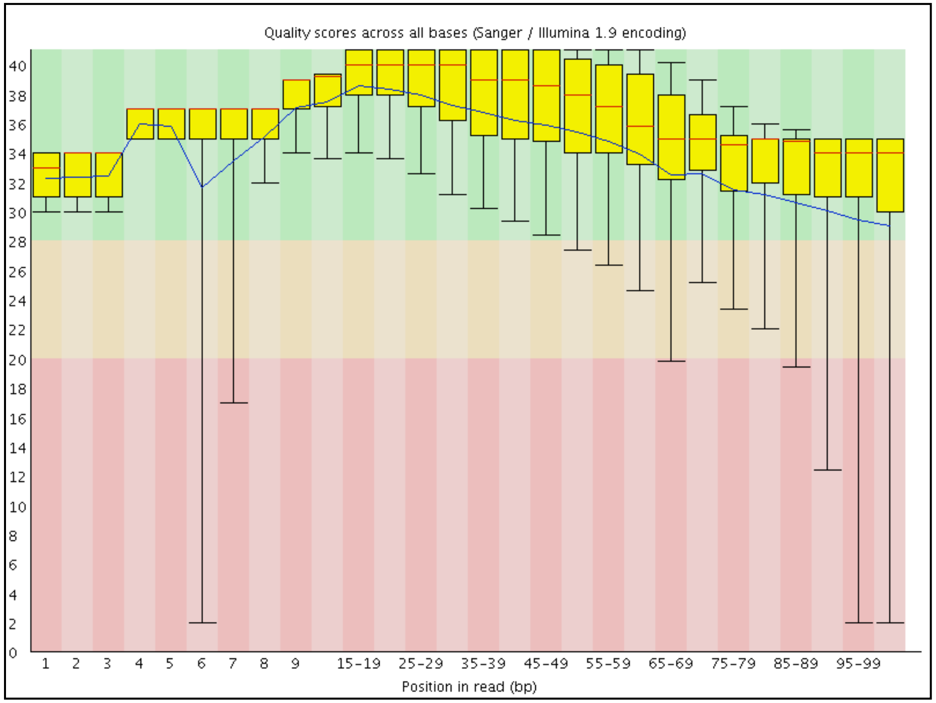




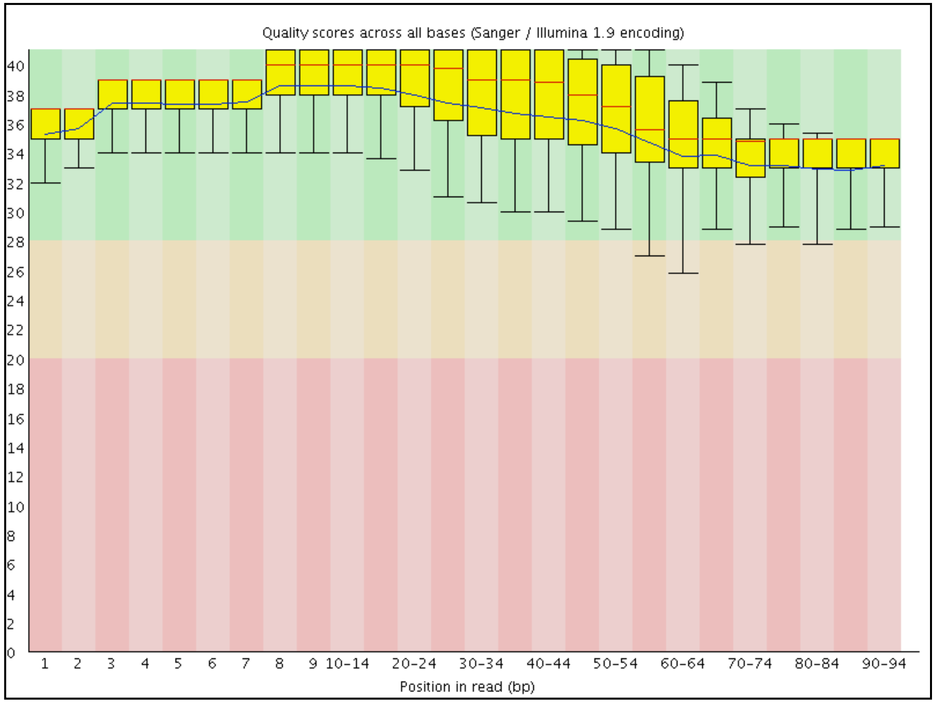
Trimming is an option in some cases

- trimmomatic or fastx-toolkit or write your own in python!
- depending on application you may or may not want to trim
  - assembly

Before quality trimming

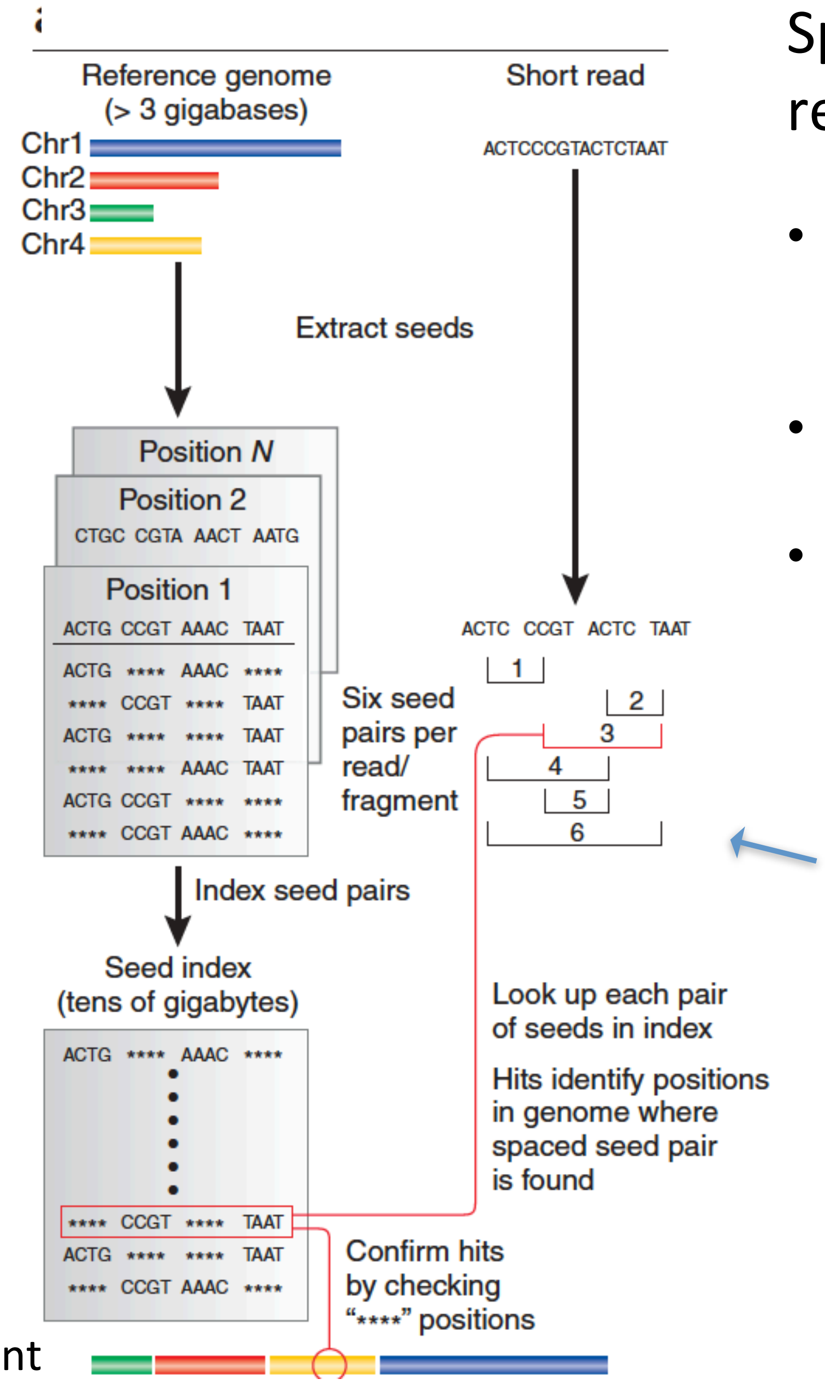


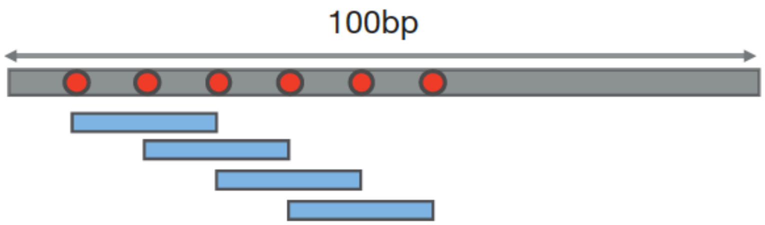
After quality trimming



Spaced-seed indexing of the reference genome

- Need to break up the genome into manageable segments
- Create index of short sequences
- Match seeds against genome index

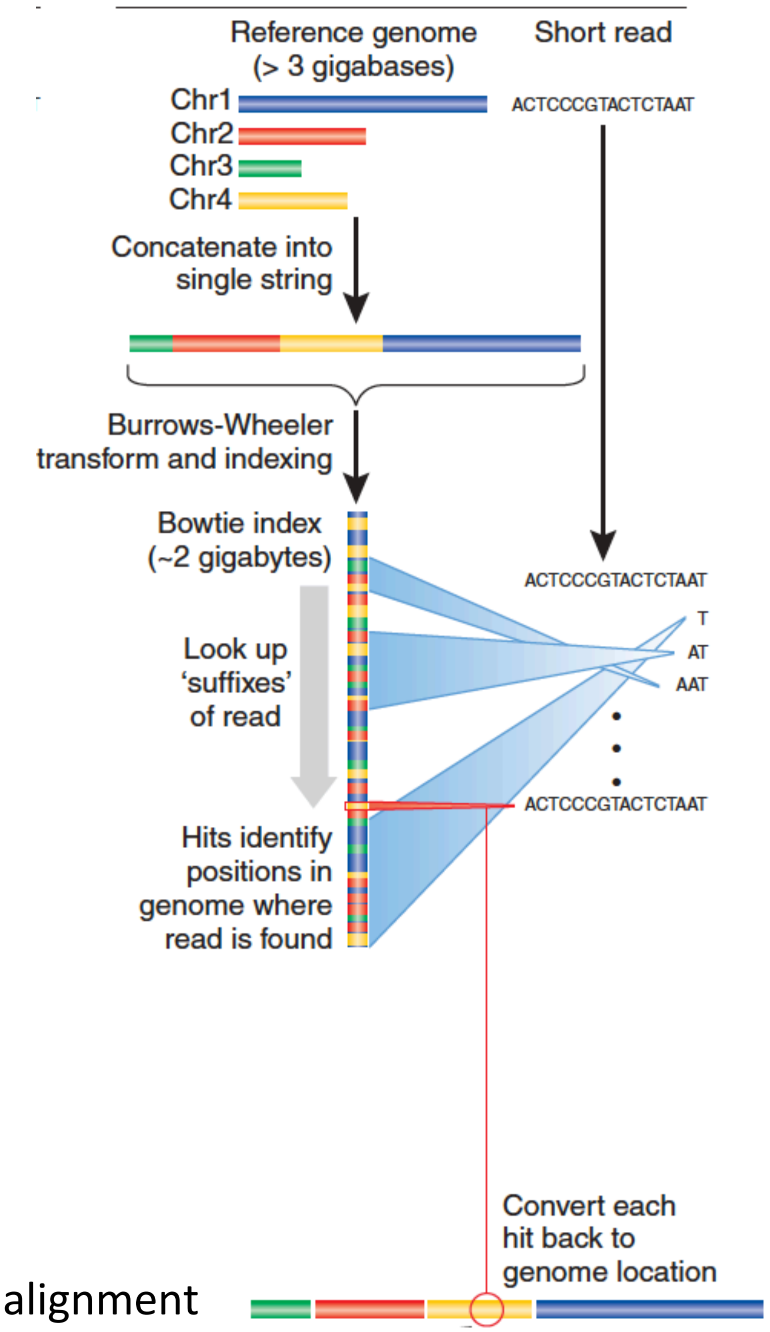




130.132.212.207/mediawiki/images/f/f9/Short_read_2014.pptx

Trapnell and Salzberg, Nat Biotechnology 27:455 (2009)

Reference genome indexing using Burrows-Wheeler transform

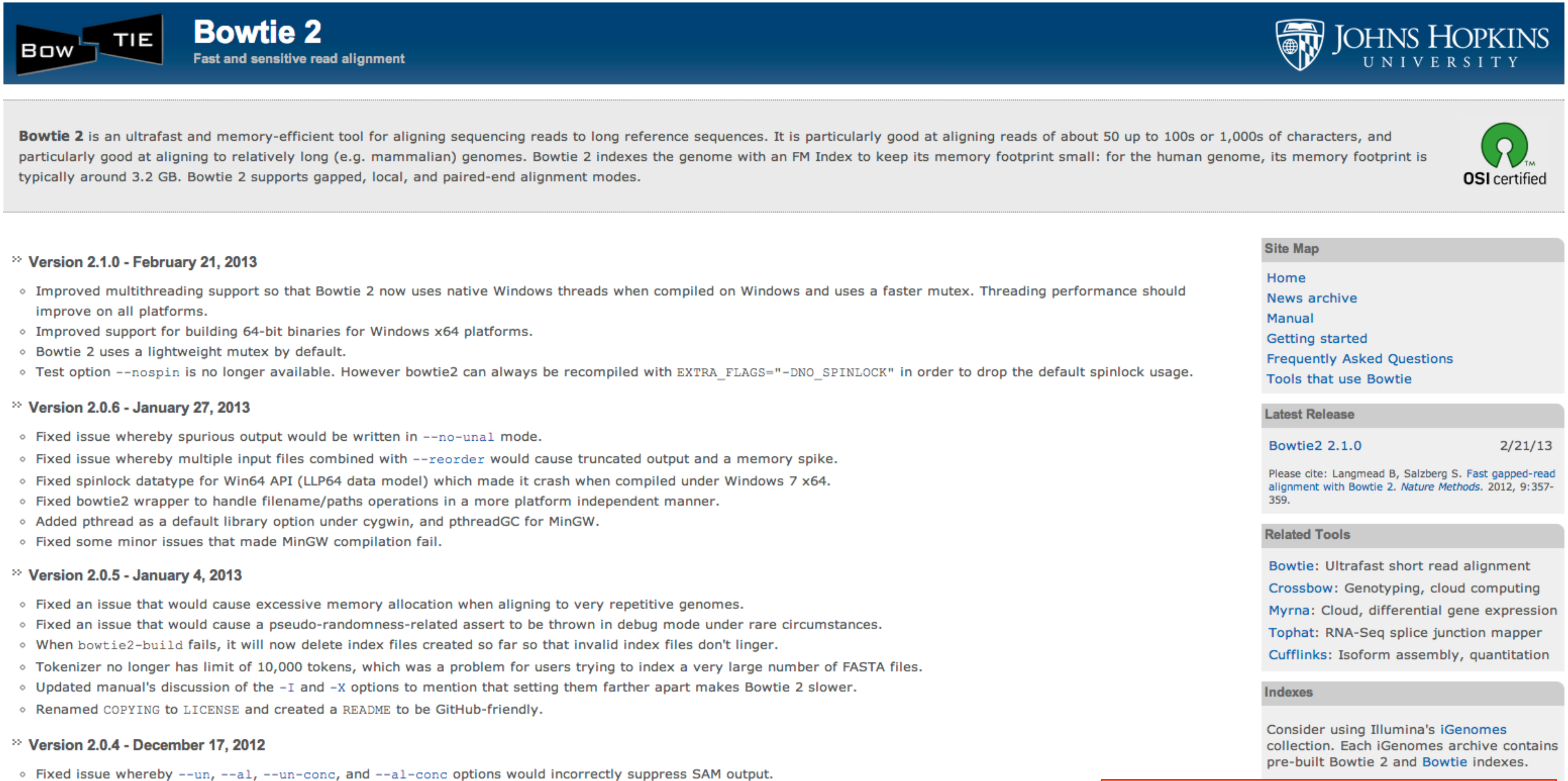

Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Concatenate into single string
Burrows-Wheeler transform and indexing
Bowtie index
(~2 gigabytes)
Look up 'suffixes' of read
ACTCCCGTACTCTAAT
T
AT
AAT
ACTCCCGTACTCTAAT
Hits identify positions in genome where read is found
Convert each hit back to genome location
alignment

- Reversible encoding scheme
- Simplifies genome sequence
- Results in "indexed" genome
- Very rapid alignments

130.132.212.207/mediawiki/images/f/f9/Short_read_2014.pptx

Trapnell and Salzberg, Nat Biotechnology 27:455 (2009)

Bowtie 2



Bowtie 2

Fast and sensitive read alignment

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

Version 2.1.0 - February 21, 2013

- Improved multithreading support so that Bowtie 2 now uses native Windows threads when compiled on Windows and uses a faster mutex. Threading performance should improve on all platforms.
- Improved support for building 64-bit binaries for Windows x64 platforms.
- Bowtie 2 uses a lightweight mutex by default.
- Test option `--nospin` is no longer available. However bowtie2 can always be recompiled with `EXTRA_FLAGS="-DNO_SPINLOCK"` in order to drop the default spinlock usage.

Version 2.0.6 - January 27, 2013

- Fixed issue whereby spurious output would be written in `--no-unal` mode.
- Fixed issue whereby multiple input files combined with `--reorder` would cause truncated output and a memory spike.
- Fixed spinlock datatype for Win64 API (LLP64 data model) which made it crash when compiled under Windows 7 x64.
- Fixed bowtie2 wrapper to handle filename/paths operations in a more platform independent manner.
- Added pthread as a default library option under cygwin, and pthreadGC for MinGW.
- Fixed some minor issues that made MinGW compilation fail.

Version 2.0.5 - January 4, 2013

- Fixed an issue that would cause excessive memory allocation when aligning to very repetitive genomes.
- Fixed an issue that would cause a pseudo-randomness-related assert to be thrown in debug mode under rare circumstances.
- When `bowtie2-build` fails, it will now delete index files created so far so that invalid index files don't linger.
- Tokenizer no longer has limit of 10,000 tokens, which was a problem for users trying to index a very large number of FASTA files.
- Updated manual's discussion of the `-I` and `-X` options to mention that setting them farther apart makes Bowtie 2 slower.
- Renamed `COPYING` to `LICENSE` and created a `README` to be GitHub-friendly.

Version 2.0.4 - December 17, 2012

- Fixed issue whereby `--un`, `--al`, `--un-conc`, and `--al-conc` options would incorrectly suppress SAM output.

Site Map

Home
News archive
Manual
Getting started
Frequently Asked Questions
Tools that use Bowtie

Latest Release

Bowtie2 2.1.0	2/21/13

Please cite: Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

Related Tools

Bowtie: Ultrafast short read alignment
Crossbow: Genotyping, cloud computing
Myrna: Cloud, differential gene expression
Tophat: RNA-Seq splice junction mapper
Cufflinks: Isoform assembly, quantitation

Indexes

Consider using Illumina's iGenomes collection. Each iGenomes archive contains pre-built Bowtie 2 and Bowtie indexes.

Pre-built Indexed genomes

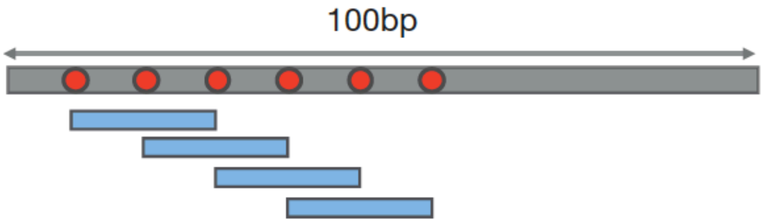
Bowtie 1 and Bowtie 2 indexes are not compatible

130.132.212.207/mediawiki/images/f/f9/Short_read_2014.pptx

Alignments in Bowtie 2

```
@HWI-ST974:58:C059FACXX:2:1201:10589:110434 1:N:0:TGACCA
TGCACACTGAAGGACCTGGAATATGGCGAGAAAACTGAAAATCATGGAAAATGAGAAATACACACTTTAGGACGTG
```

Multiseed alignment (ungapped)



Seed length: 16 nt, every 10 nt

# mismatches: 0

Seeds are extended (gaps allowed) to generate alignment

Match = 2

```
Ref   TGCACACTGAAGGACCTGGAATATGGCGAGAAAACTGAAAATCATGGAAAATGAGAAATACACACTTTAGGACGTG
Read  TGCACACTGAAGGTCCTGGAATATGGCGAGAAAACTGAAAATCATGGAAA--GAGAAATACACACTTTAGGACGTG
```

Mismatch = -6

Gap = -11

-5 to open

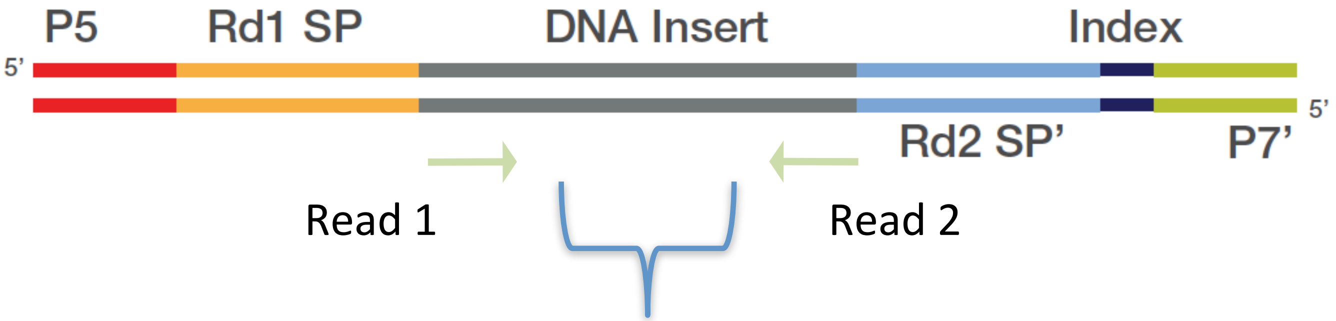
-3 to extend by 1 bp

130.132.212.207/mediawiki/images/f/f9/Short_read_2014.pptx

bwa mem and bowtie2

- most widely used
- most cited
- easiest to use
- free
- “best” by several measures
- recent (and older) version on hpc

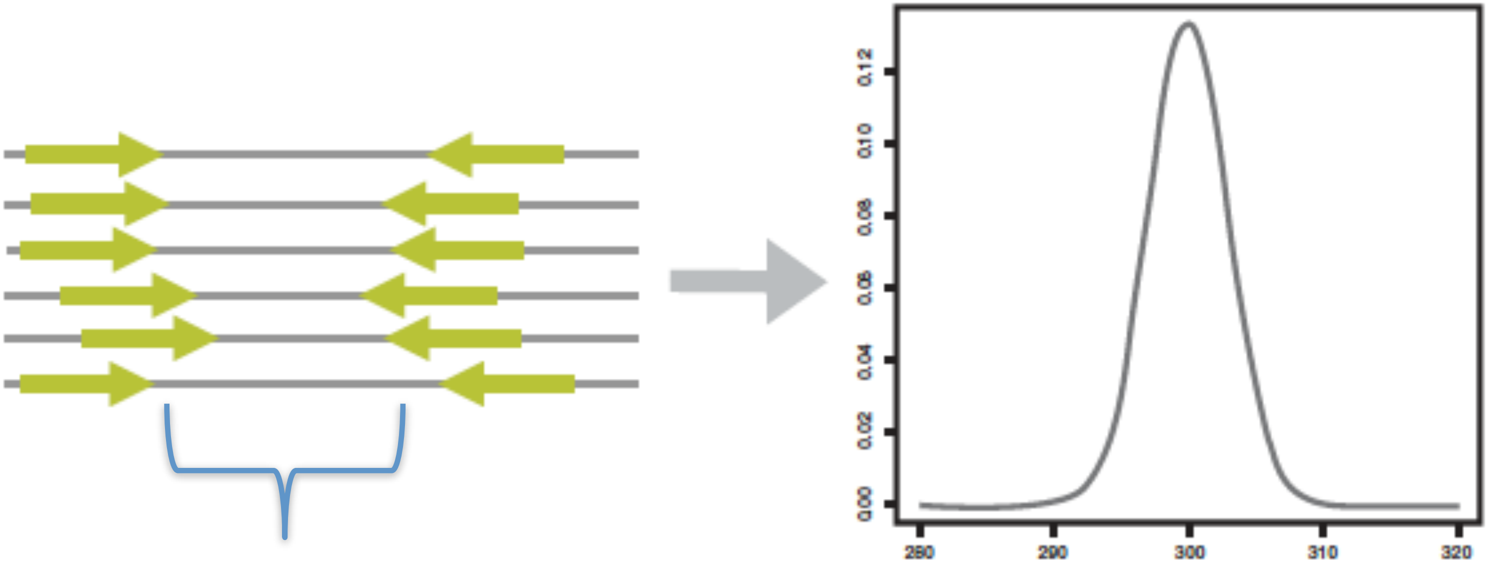
Mapping paired end reads



Read 1

Read 2

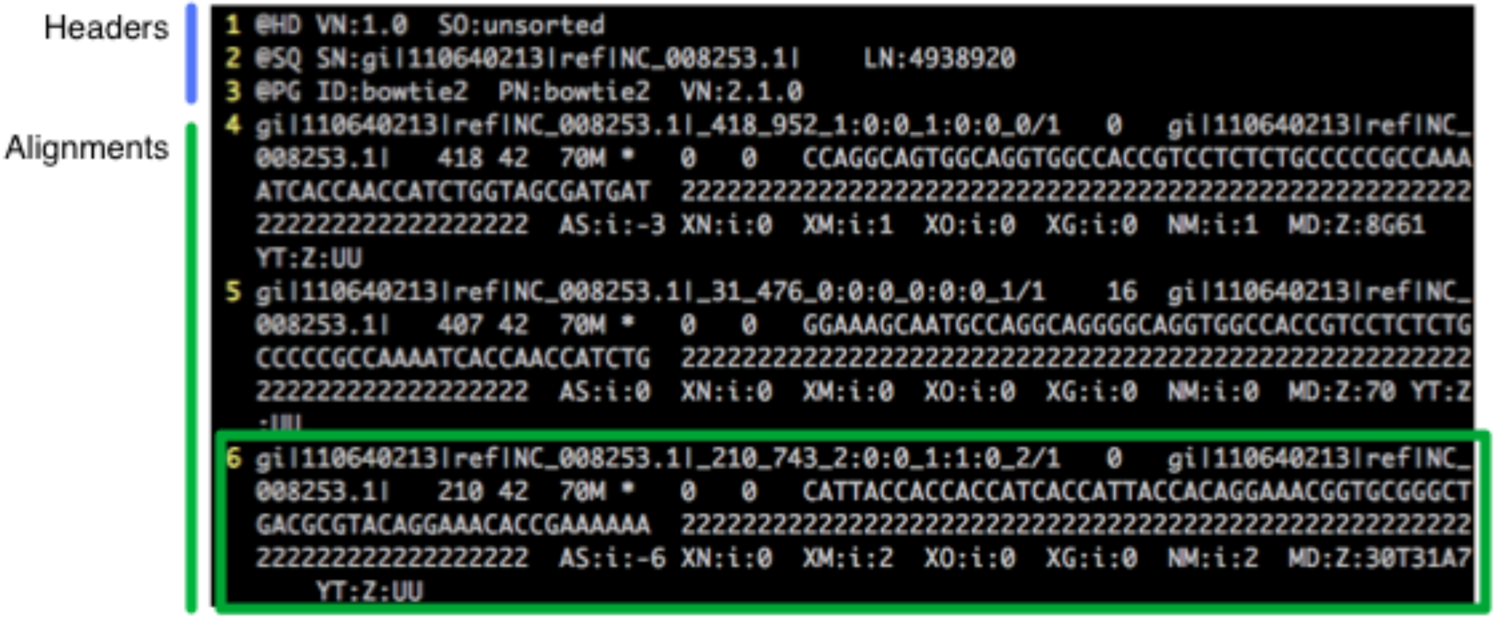
Insert size



Insert size
within specified range

sam/bam file summarizes alignment

https://samtools.github.io/hts-specs/
http://davetang.org/wiki/tiki-index.php?page=SAM
http://genome.sph.umich.edu/wiki/SAM
bam = index-able binary sam

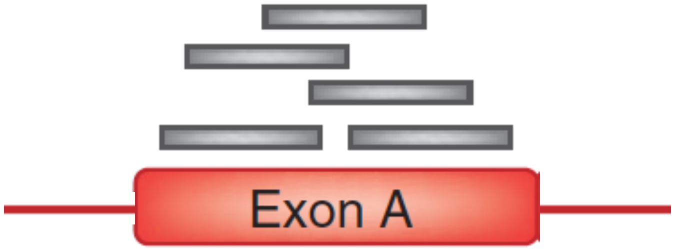


```
@HD VN:1.0  SO:unsorted
@SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
@PG ID:bowtie2  PN:bowtie2  VN:2.1.0
gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
ATCACCAACCATCTGGTAGCGATGAT  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
YT:Z:UU
gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
CCCCCGCCAAAATCACCAACCATCTG  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
:UU
gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
GACGCGTACAGGAAACACCGAAAAAA  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
    YT:Z:UU
```

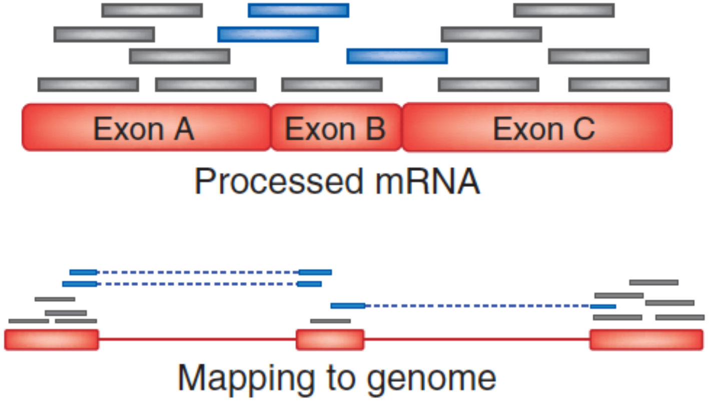
Each row describes a single alignment of a raw read against the reference genome.
Each alignment has 11 mandatory fields, followed by any number of optional fields.

RNA is special

Exome or Genome



Transcriptome



- aligning RNAseq reads to a genome must allow for large “gaps” (= introns) not just SNPs and small INDELs
- a GFF/GTF describing KNOWN gene structures can aid this process
- special tool for this called tophat
- tophat on hpc

Lecture 3

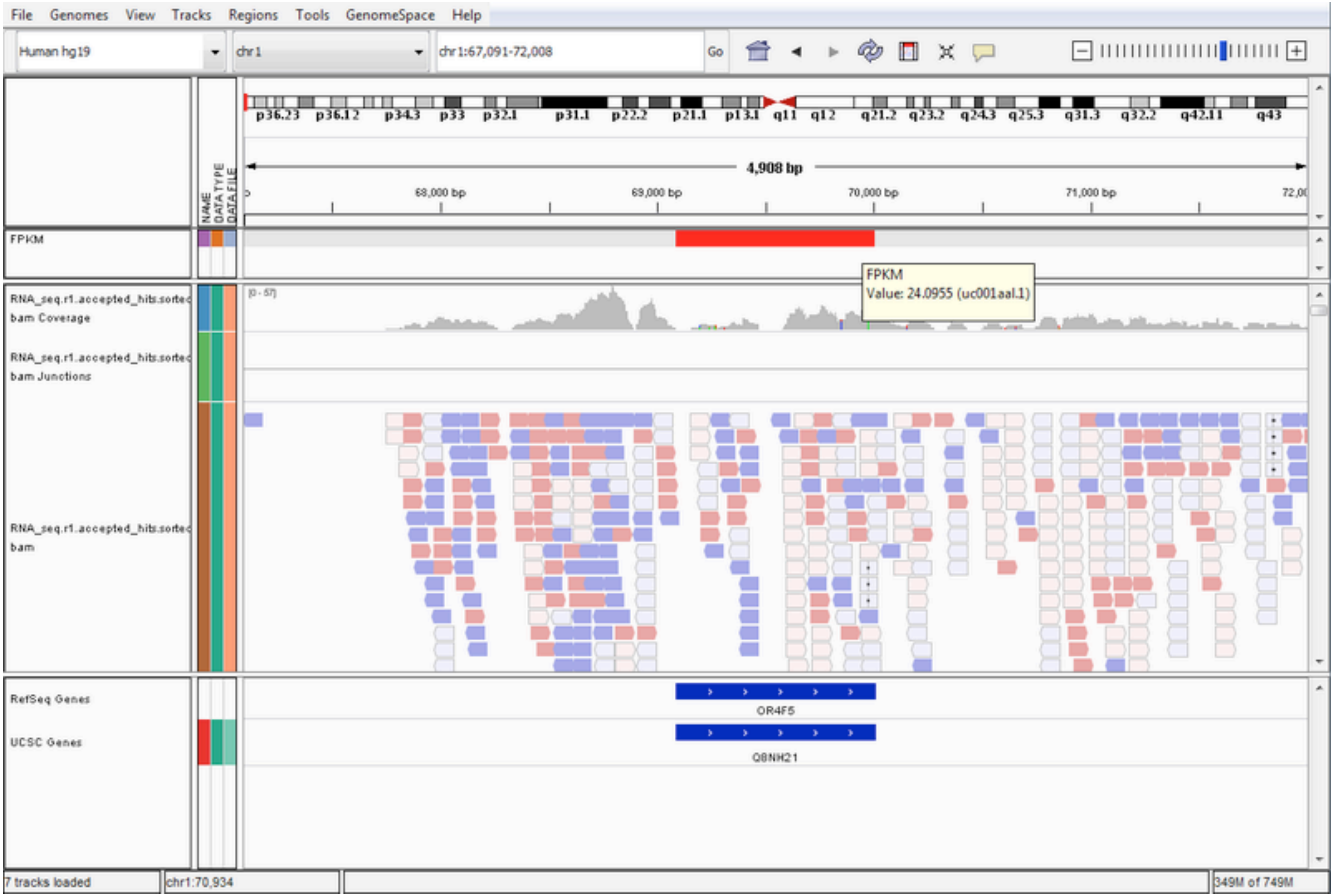
…do something with alignments…

Sometimes it is helpful to look at alignments

igv viewer seems most widely used

http://software.broadinstitute.org/software/igv/

import bam files
import GFF
runs on your desktop



Human hg19

chr1

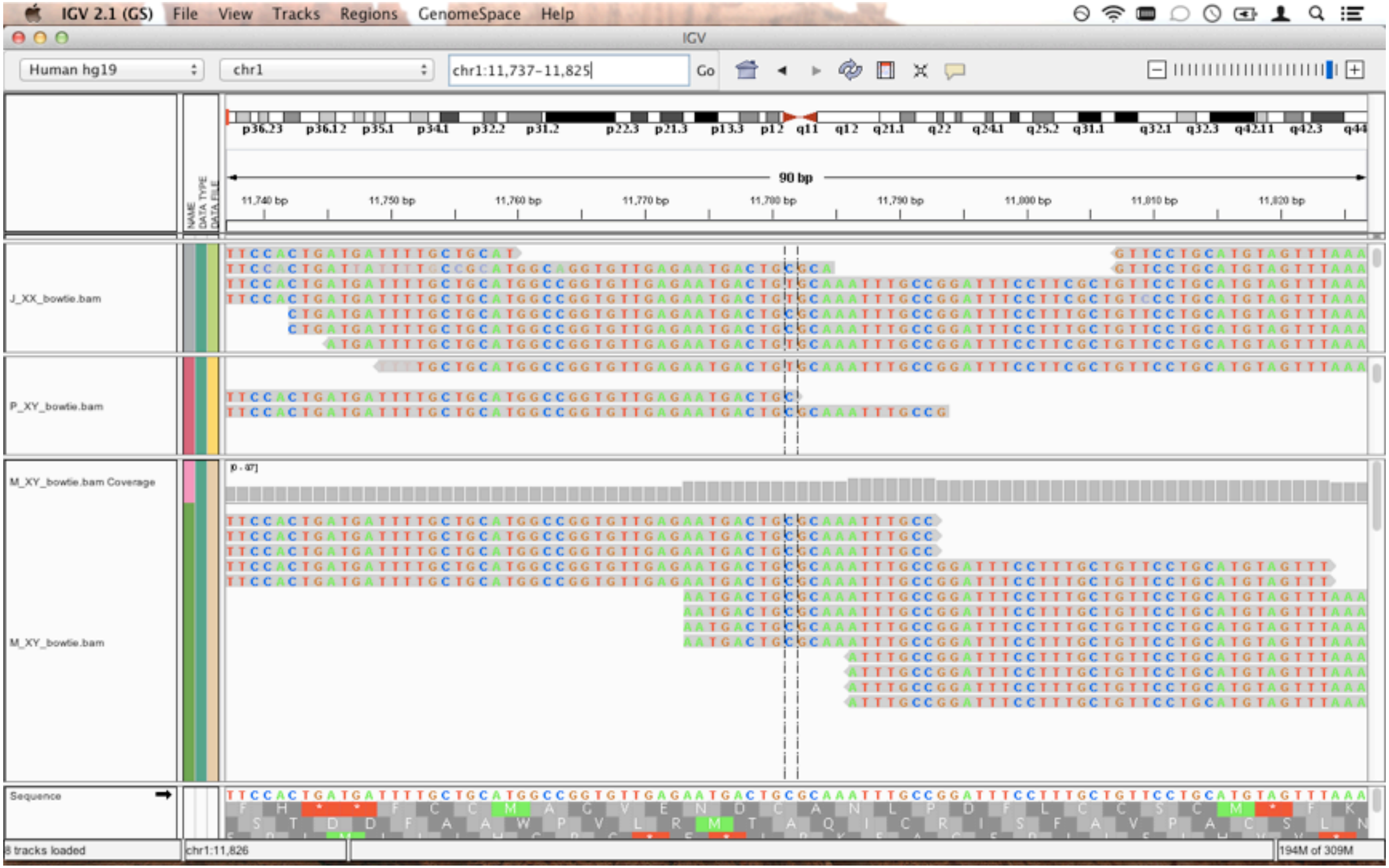
chr1:67,091-72,008

FPKM
Value: 24.0955 (uc001aal.1)

7 tracks loaded

chr1:70,934

349M of 749M



Human hg19

chr1

chr1:11,737-11,825

Call SNPs from DNAseq

- GATK pipeline pretty “industry standard”
- on hpc
- yuk – intermediate files, lots of switches
- java...
- SNP calls in VCF file

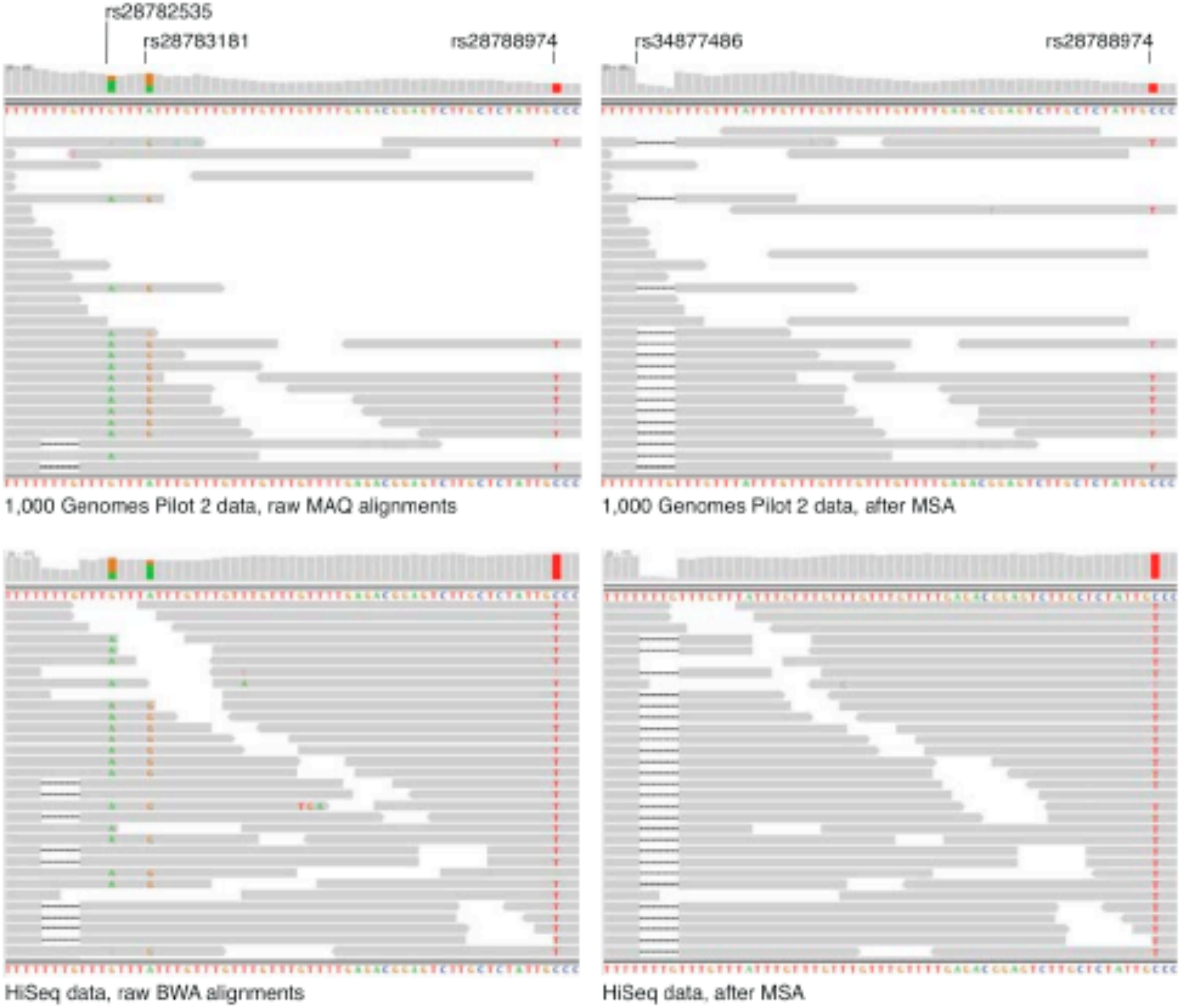
GATK "pipeline"

http://barcwiki.wi.mit.edu/wiki/SOPs/variant_calling_GATK

- align reads using bwa mem & index
- mark duplicates (poolseq, deep seq, SEseq...)
- add read groups (GATK needs these)
- merge bam files across samples
- Indel realignment with "RealignerTargetCreator" and "IndelRealigner"
- Base Recalibration (need Gold standard SNPs)
- Call variants (diploids = HaplotypeCaller) or UnifiedGenotyper
- Annotate (need GFF file & external programs)
- Filter variants for HQ calls (strand bias etc)

IndelRealigner

Before



1,000 Genomes Pilot 2 data, raw MAQ alignments

1,000 Genomes Pilot 2 data, after MSA

HiSeq data, raw BWA alignments

HiSeq data, after MSA

After

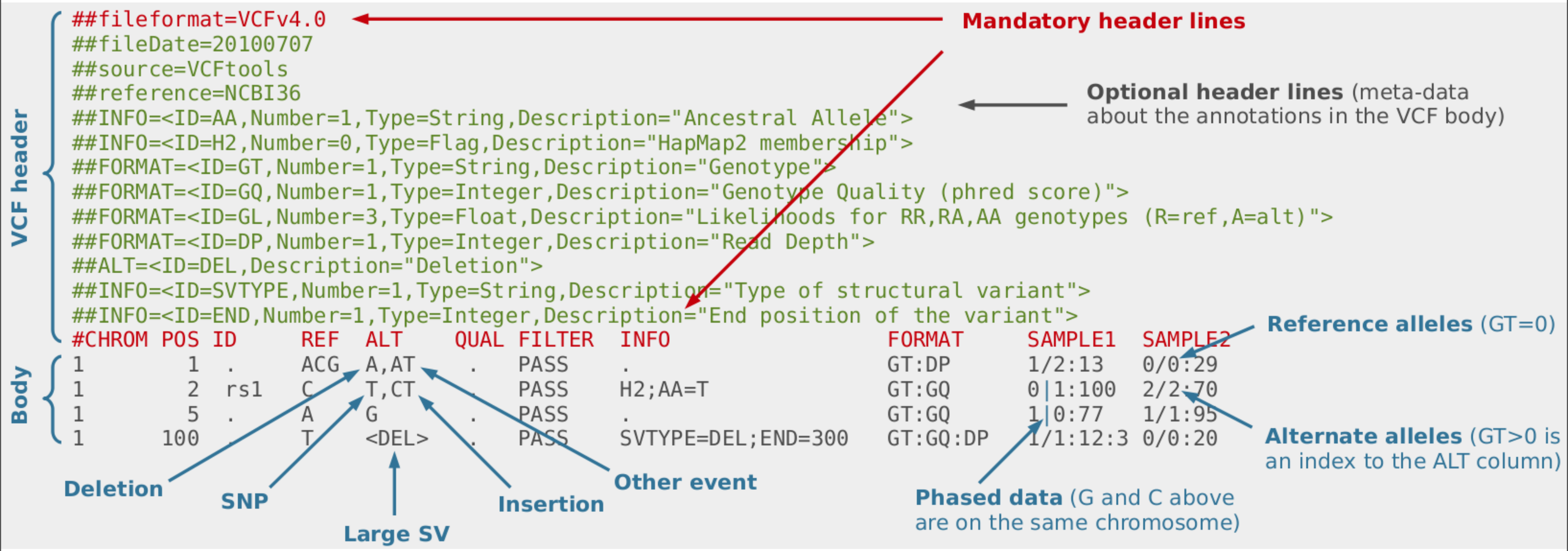
DePristo, M., Banks, E., Poplin, R. et. al, A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Gen.

FilterVariants

the potential for lots of bad karma here, depending on downstream goals. What these filters do it label (and eventually throw out) SNPs that do not satisfy certain rules. So this could impact various downstream analyses for sure (e.g., number of segregating sites, site frequency spectrum, etc).

```
# SNPs to start with
-V rawSNPS-Q30.vcf
# SNPs must be >5bp from INDEL
--mask inDels-Q30.vcf --maskExtension 5 --maskName InDel
# SNPs within 10bp of one another are masks (alignments can be poor)
--clusterWindowSize 10
#  these SNPs are poor at cross-validation
--filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)”
       --filterName "BadValidation"
# low quality
--filterExpression "QUAL < 30.0" --filterName "LowQual"
# low depth of ALT allele
--filterExpression "QD < 5.0" --filterName "LowVQCBD"
# SNP not consistent on Watson and Crick strands
--filterExpression "FS > 60" --filterName "FisherStrand"
# final SNPs
-o Q30-SNPs.vcf
```

VCF file format



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT     QUAL FILTER  INFO                 FORMAT      SAMPLE1    SAMPLE2
1        1  .     ACG  A,AT     .   PASS    .                    GT:DP       1/2:13     0/0:29
1        2  rs1   C    T,CT     .   PASS    H2;AA=T              GT:GQ       0|1:100    2/2:70
1        5  .     A    G        .   PASS    .                    GT:GQ       1|0:77     1/1:95
1      100  .     T    <DEL>    .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP    1/1:12:3   0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Phased data (G and C above are on the same chromosome)
Deletion
SNP
Large SV
Insertion
Other event

VCF file format

There are tools for working with VCF files

- -vcftools

It might be fun to look at the sensitivity of downstream stuff to the filters...

You can also roll your own

- -estimate ALT frequency from VCF (poolseq) - week 5

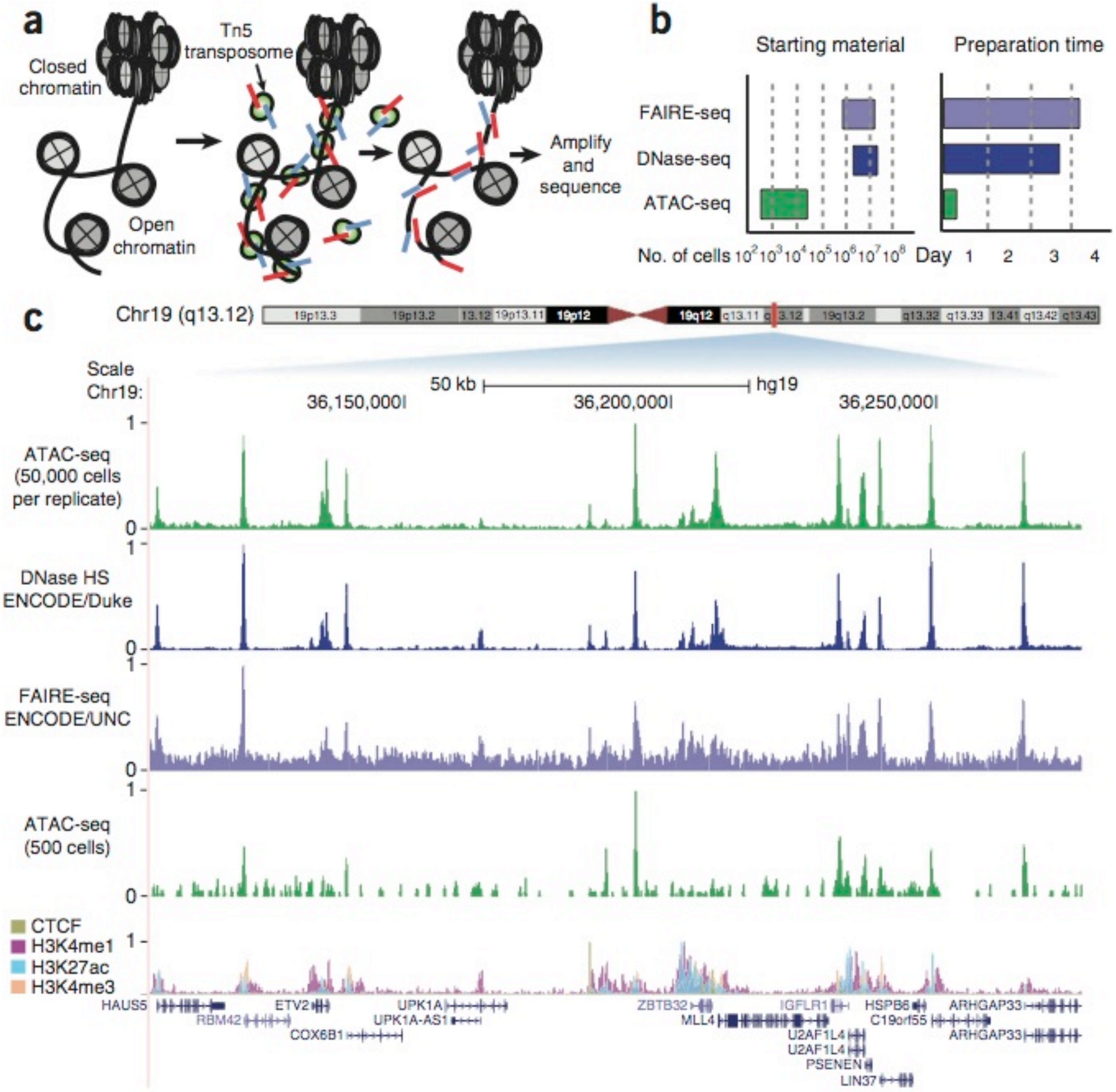
ATACseq

usually short PE reads, the goal is to eventually map the “cut sites”

macs seems to be an important part of “peak-caller” pipelines

but this is sort of hard, so initially we will just look at coverage

But eventually we could try and get to here
https://github.com/kundajelab/atac_dnase_pipelines
https://docs.google.com/document/d/1f0Cm4vRyDQDu0bMehHD7P7KOMxTOP-HiNoIvL1VcBt8/edit#



http://www.nature.com/nmeth/journal/v10/n12/full/nmeth.2688.html

DEseq2

RNAseq is a field in and of itself

in theory you can find new splice variants, quantify isoforms, etc.

in practice the first thing we do is look for differential expression at the level of each gene

expression highly variable, so experiments often employ biological replicates

DEseq2 is an R package that relies on other packages that are part of "Bioconductor". So I have created an Rstudio instance on tprout with the libraries in place

https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
https://www.bioconductor.org/packages/release/bioc/manuals/DESeq2/man/DESeq2.pdf
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/
https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf

What does DEseq do?

reads in raw counts (# reads aligning) per gene per sample

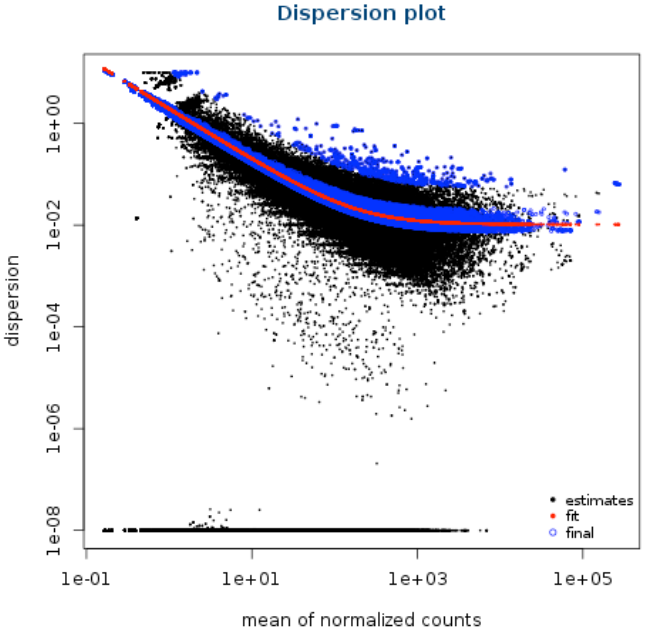
so you sort of need to define genes with a gtf file

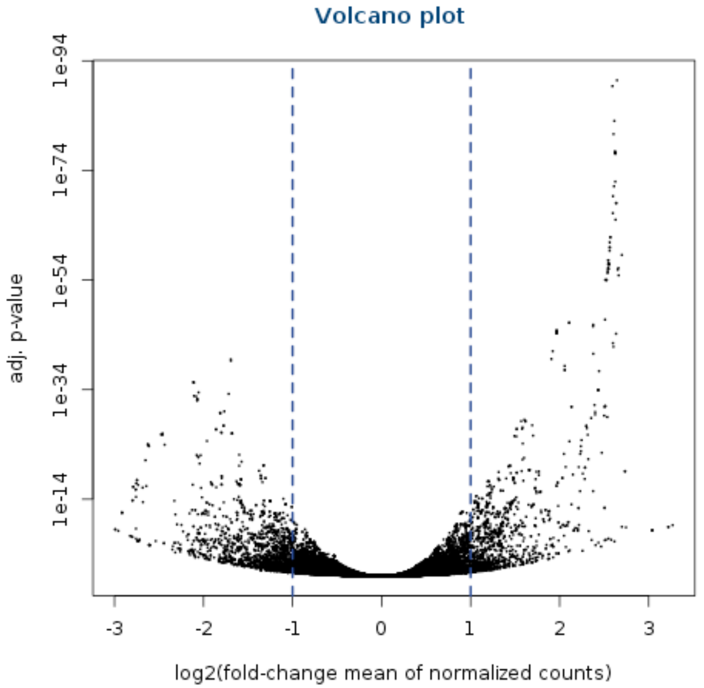
normalize counts across samples

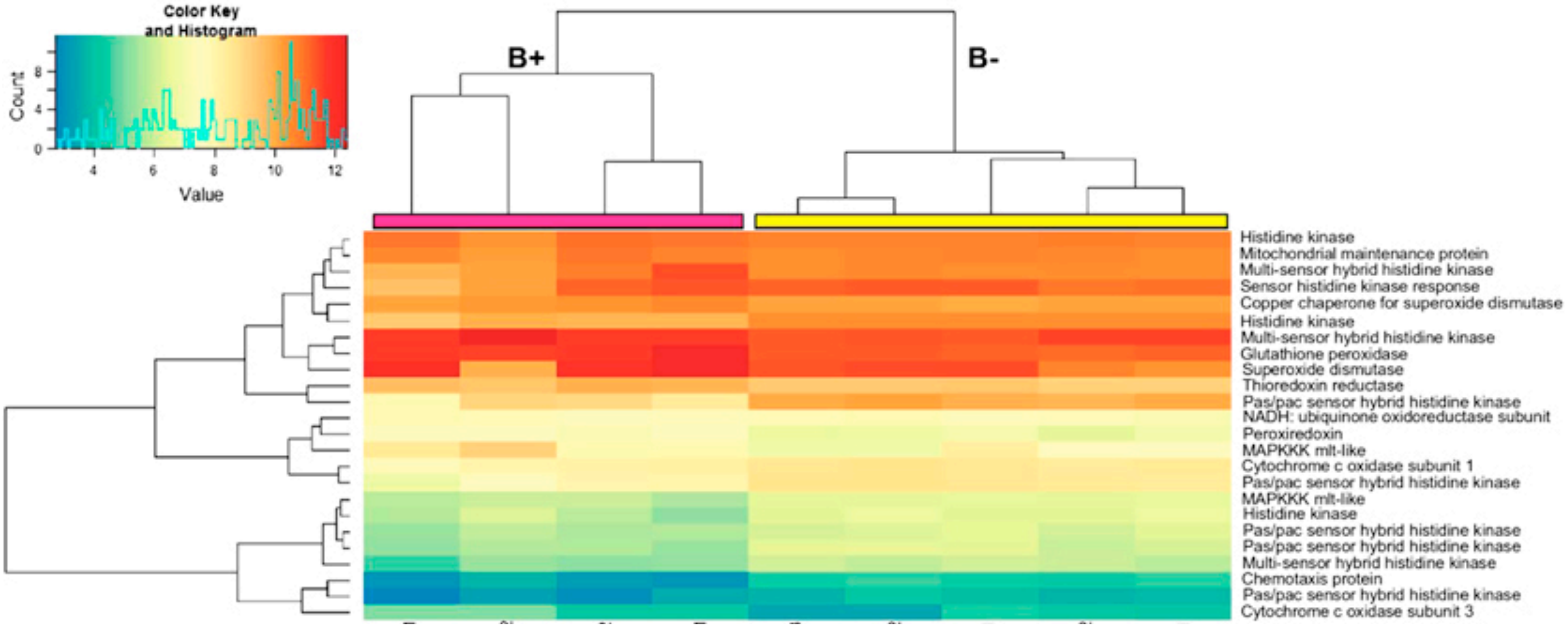
“variance shrinkage” to do statistical testing

statistical testing

lots of plots to make sure the wheels have not come off the bus


Dispersion plot
dispersion
1e+00
1e-02
1e-04
1e-06
1e-08
1e-01
1e+01
1e+03
1e+05
mean of normalized counts
estimates
fit
final


Volcano plot
adj. p-value
1e-94
1e-74
1e-54
1e-34
1e-14
-3
-2
-1
0
1
2
3
log2(fold-change mean of normalized counts)

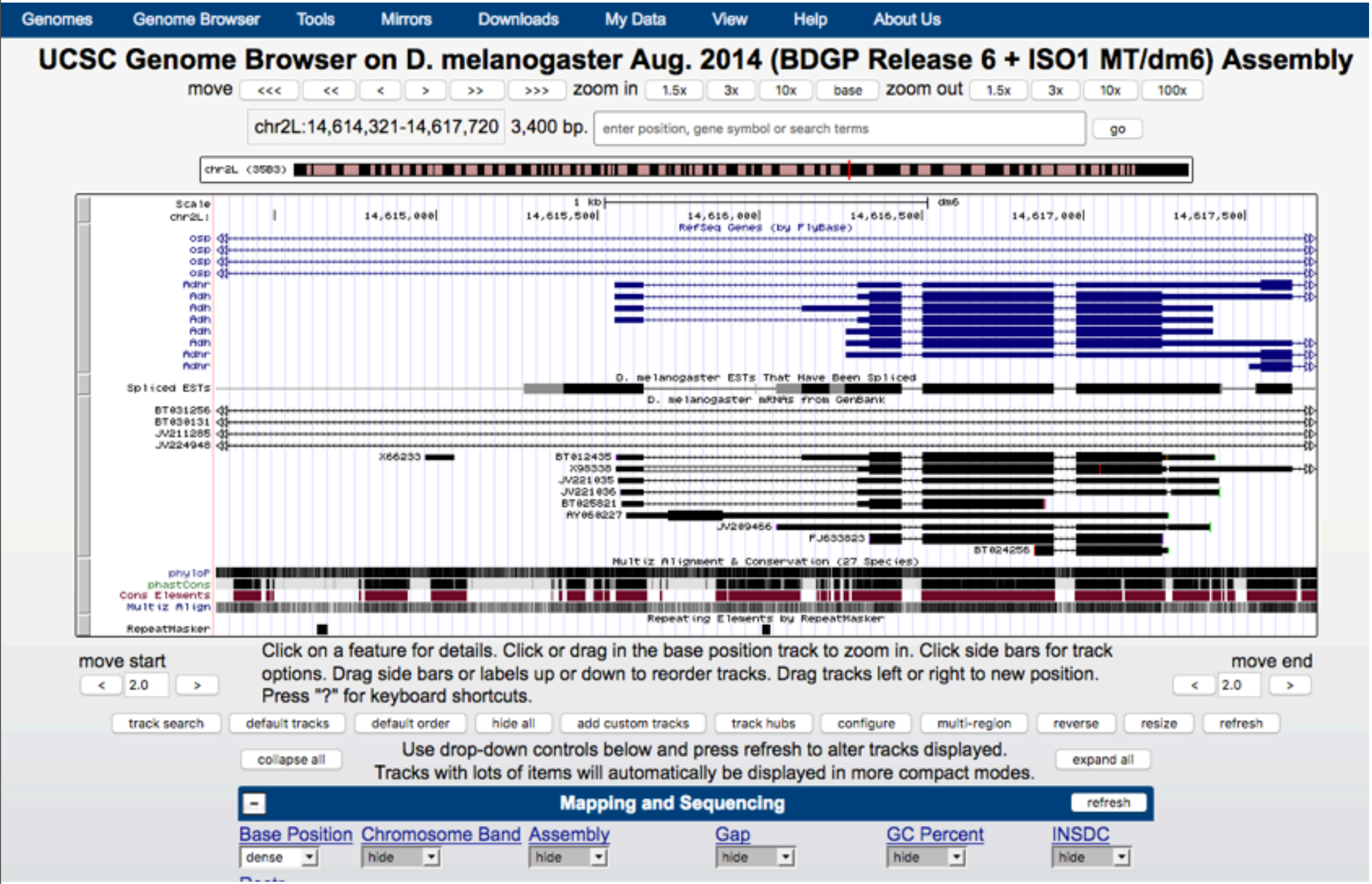

Color Key and Histogram
Count
0
4
8
4
6
8
10
12
Value
B+
B-
Histidine kinase
Mitochondrial maintenance protein
Multi-sensor hybrid histidine kinase
Sensor histidine kinase response
Copper chaperone for superoxide dismutase
Histidine kinase
Multi-sensor hybrid histidine kinase
Glutathione peroxidase
Superoxide dismutase
Thioredoxin reductase
Pas/pac sensor hybrid histidine kinase
NADH: ubiquinone oxidoreductase subunit
Peroxiredoxin
MAPKKK mlt-like
Cytochrome c oxidase subunit 1
Pas/pac sensor hybrid histidine kinase
MAPKKK mlt-like
Histidine kinase
Pas/pac sensor hybrid histidine kinase
Pas/pac sensor hybrid histidine kinase
Multi-sensor hybrid histidine kinase
Chemotaxis protein
Pas/pac sensor hybrid histidine kinase
Cytochrome c oxidase subunit 3

Lecture 4

Santa Cruz Genome Browser

a way of representing genomes in browser
great for models
(but can do custom genomes)
“tracks summarize stuff known”
you can upload your own tracks
you can “host” your own tracks

A tutorial: http://www.sciencedirect.com/science/article/pii/S0888754308000451



UCSC Genome Browser on D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) Assembly

chr2L:14,614,321-14,617,720 3,400 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

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.

Base Position: dense

Chromosome Band: hide

Assembly: hide

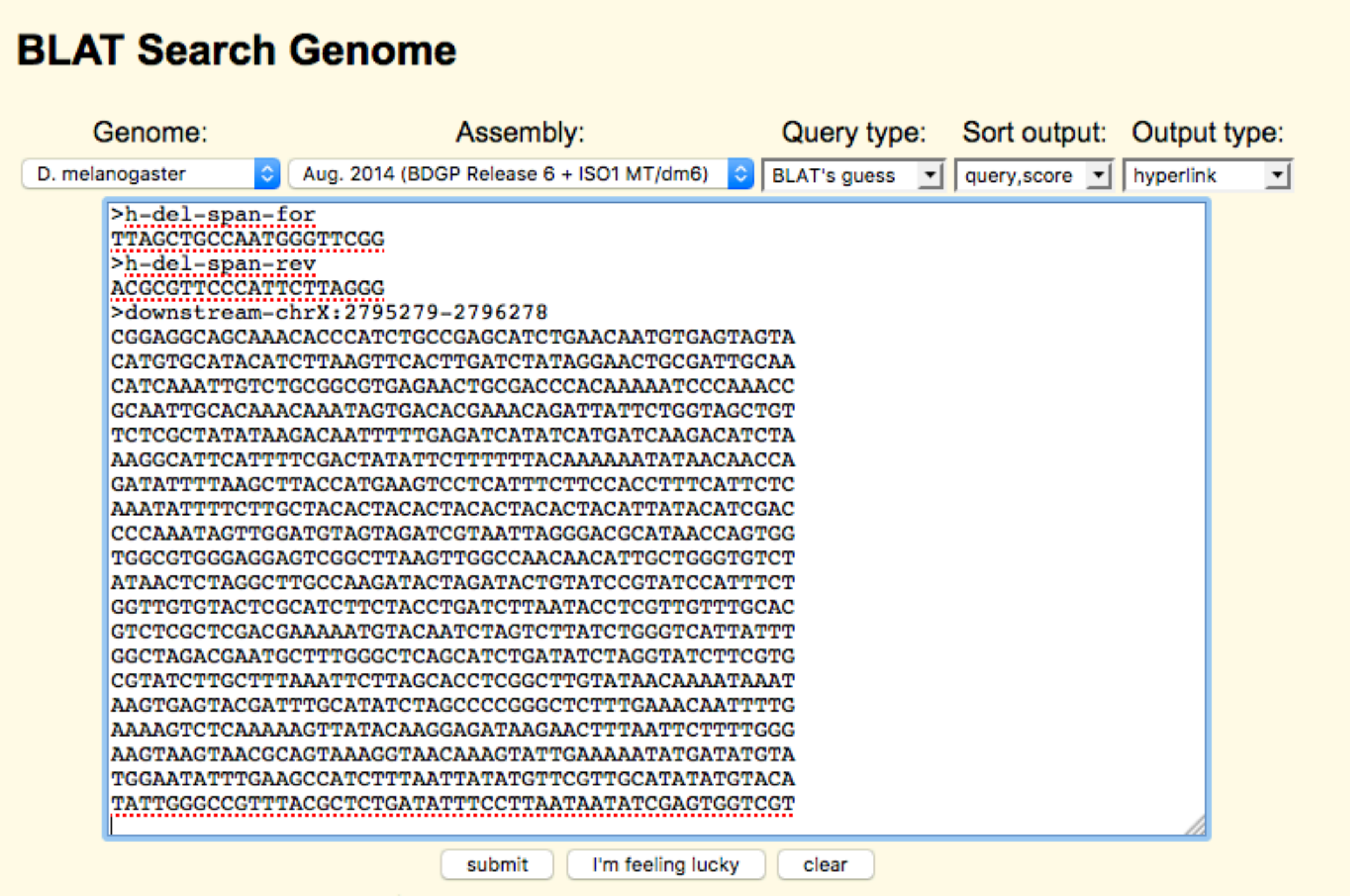
Gap: hide

GC Percent: hide

INSDC: hide

BLAT

tools -> BLAT
quickly see the context of some sequence fragment



BLAT Search Genome

Genome: D. melanogaster

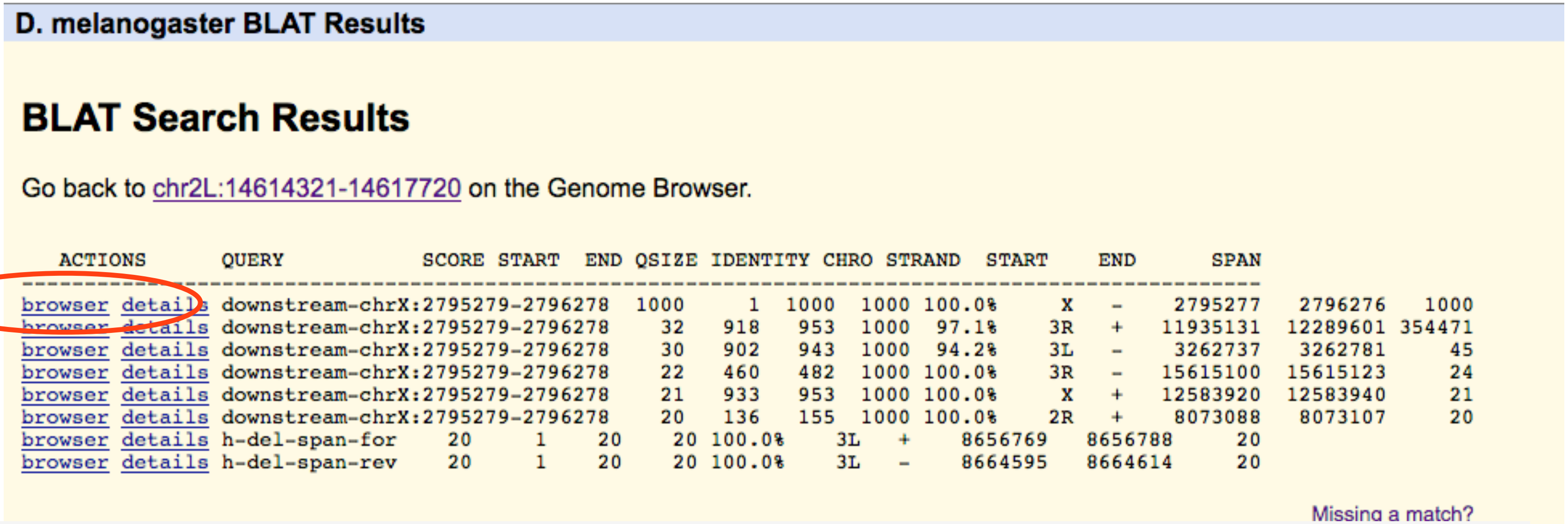
Assembly: Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)

Query type: BLAT's guess

Sort output: query,score

Output type: hyperlink

```
>h-del-span-for
TTAGCTGCCAATGGGTTCGG
>h-del-span-rev
ACGCGTTCCCATTCTTAGGG
>downstream-chrX:2795279-2796278
CGGAGGCAGCAAACACCCATCTGCCGAGCATCTGAACAATGTGAGTAGTA
CATGTGCATACATCTTAAGTTCACTTGATCTATAGGAACTGCGATTGCAA
CATCAAATTGTCTGCGGCGTGAGAACTGCGACCCACAAAAATCCCAAACC
GCAATTGCACAAACAAATAGTGACACGAAACAGATTATTCTGGTAGCTGT
TCTCGCTATATAAGACAATTTTTGAGATCATATCATGATCAAGACATCTA
AAGGCATTCATTTTCGACTATATTCTTTTTTACAAAAAATATAACAACCA
GATATTTTAAGCTTACCATGAAGTCCTCATTTCTTCCACCTTTCATTCTC
AAATATTTTCTTGCTACACTACACTACACTACACTACATTATACATCGAC
CCCAAATAGTTGGATGTAGTAGATCGTAATTAGGGACGCATAACCAGTGG
TGGCGTGGGAGGAGTCGGCTTAAGTTGGCCAACAACATTGCTGGGTGTCT
ATAACTCTAGGCTTGCCAAGATACTAGATACTGTATCCGTATCCATTTCT
GGTTGTGTACTCGCATCTTCTACCTGATCTTAATACCTCGTTGTTTGCAC
GTCTCGCTCGACGAAAAATGTACAATCTAGTCTTATCTGGGTCATTATTT
GGCTAGACGAATGCTTTGGGCTCAGCATCTGATATCTAGGTATCTTCGTG
CGTATCTTGCTTTAAATTCTTAGCACCTCGGCTTGTATAACAAAATAAAT
AAGTGAGTACGATTTGCATATCTAGCCCCGGGCTCTTTGAAACAATTTTG
AAAAGTCTCAAAAAGTTATACAAGGAGATAAGAACTTTAATTCTTTTGGG
AAGTAAGTAACGCAGTAAAGGTAACAAAGTATTGAAAAATATGATATGTA
TGGAATATTTGAAGCCATCTTTAATTATATGTTCGTTGCATATATGTACA
TATTGGGCCGTTTACGCTCTGATATTTCCTTAATAATATCGAGTGGTCGT
```



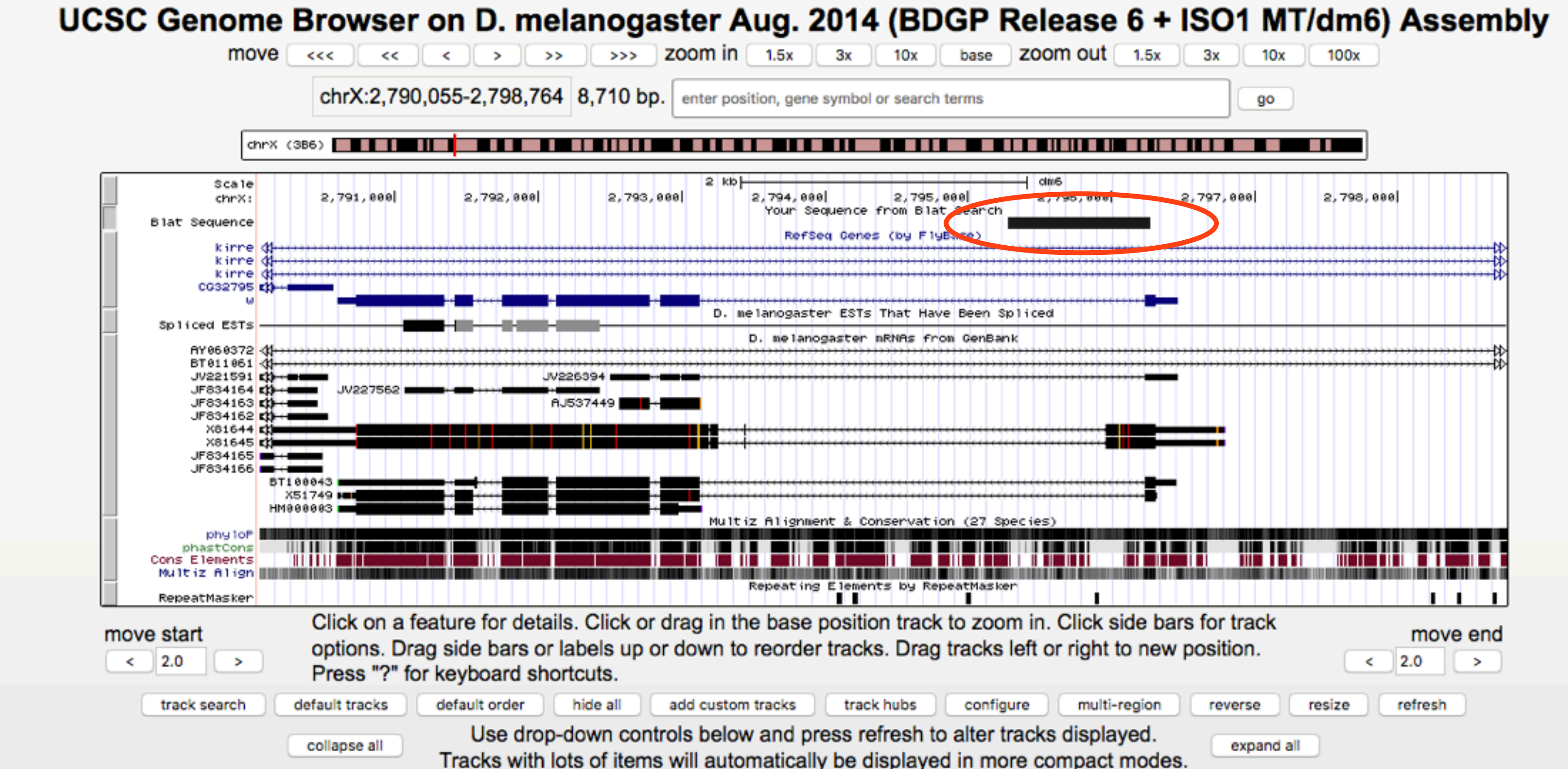
D. melanogaster BLAT Results

BLAT Search Results

Go back to chr2L:14614321-14617720 on the Genome Browser.

```
   ACTIONS      QUERY                 SCORE START  END QSIZE IDENTITY CHRO STRAND  START    END       SPAN
---------------------------------------------------------------------------------------------------------------
browser details downstream-chrX:2795279-2796278  1000     1  1000  1000 100.0%    X    -     2795277  2796276   1000
browser details downstream-chrX:2795279-2796278    32   918   953  1000  97.1%   3R    +    11935131 12289601 354471
browser details downstream-chrX:2795279-2796278    30   902   943  1000  94.2%   3L    -     3262737  3262781     45
browser details downstream-chrX:2795279-2796278    22   460   482  1000 100.0%   3R    -    15615100 15615123     24
browser details downstream-chrX:2795279-2796278    21   933   953  1000 100.0%    X    +    12583920 12583940     21
browser details downstream-chrX:2795279-2796278    20   136   155  1000 100.0%   2R    +     8073088  8073107     20
browser details h-del-span-for    20     1    20    20 100.0%    3L    +    8656769  8656788     20
browser details h-del-span-rev    20     1    20    20 100.0%    3L    -    8664595  8664614     20
```

Missing a match?



UCSC Genome Browser on D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) Assembly

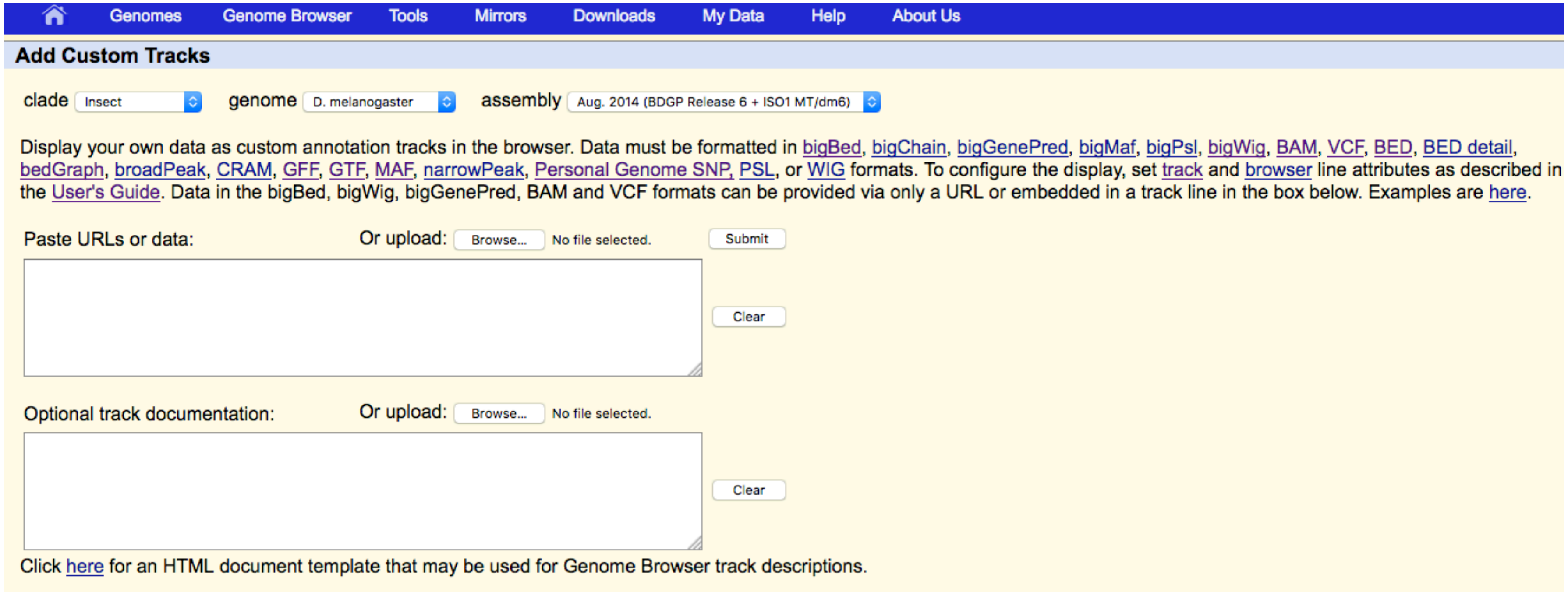
chrX:2,790,055-2,798,764 8,710 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

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.

BED

quickly add “annotations” (under add custom tracks)



Add Custom Tracks

clade Insect genome D. melanogaster assembly Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)

Display your own data as custom annotation tracks in the browser. Data must be formatted in bigBed, bigChain, bigGenePred, bigMaf, bigPsl, bigWig, BAM, VCF, BED, BED detail, bedGraph, broadPeak, CRAM, GFF, GTF, MAF, narrowPeak, Personal Genome SNP, PSL, or WIG formats. To configure the display, set track and browser line attributes as described in the User's Guide. Data in the bigBed, bigWig, bigGenePred, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Examples are here.

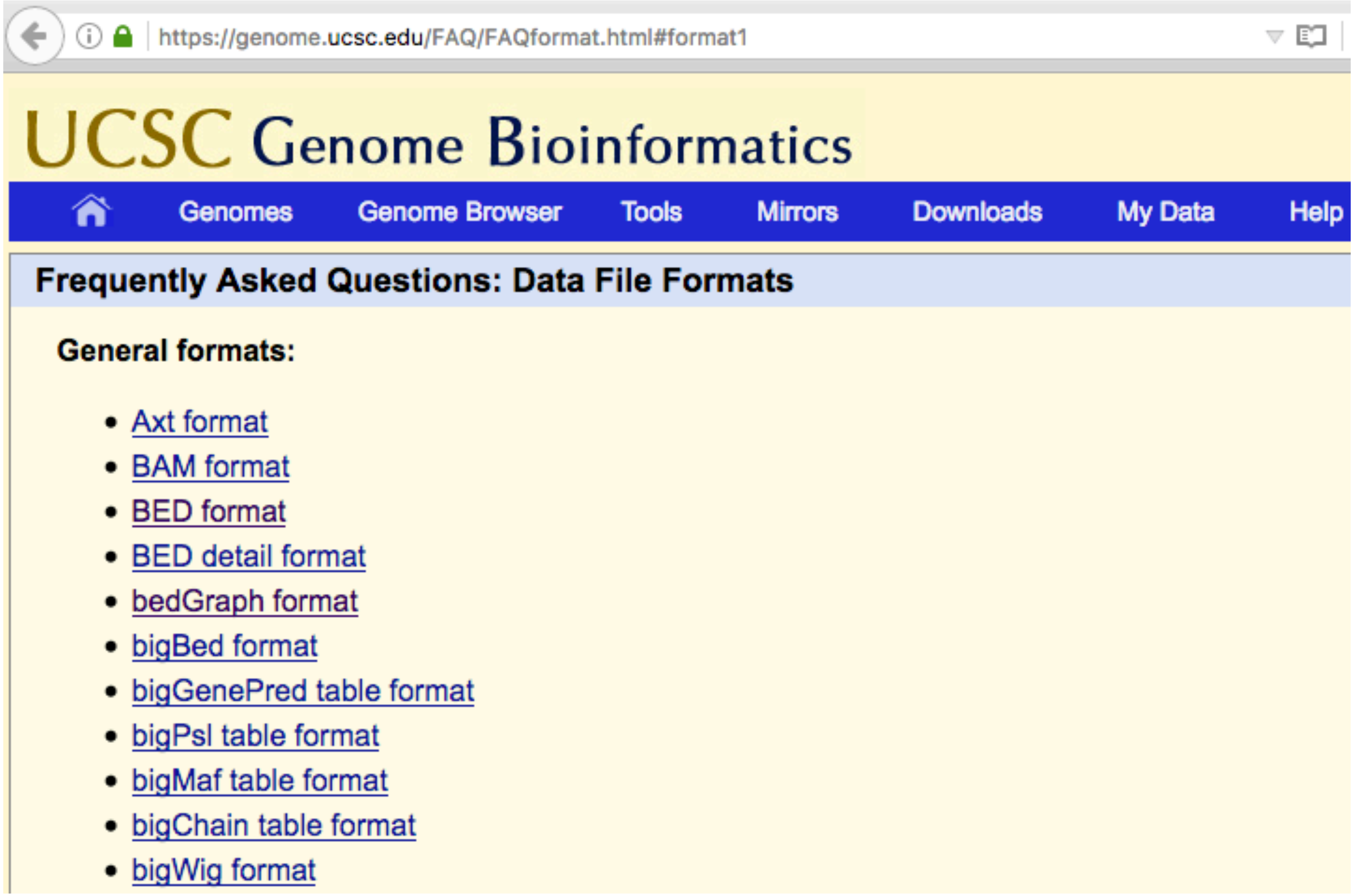
Paste URLs or data: Or upload: Browse... No file selected. Submit

Clear

Optional track documentation: Or upload: Browse... No file selected.

Clear

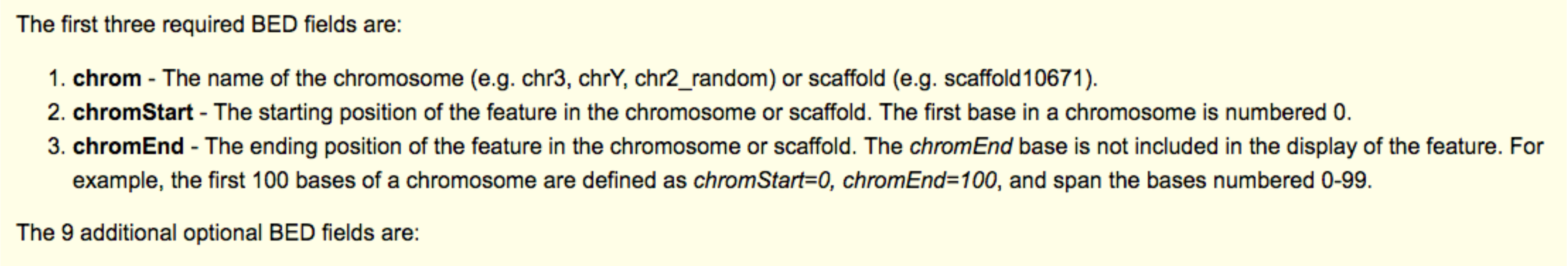
Click here for an HTML document template that may be used for Genome Browser track descriptions.


https://genome.ucsc.edu/FAQ/FAQformat.html#format1
UCSC Genome Bioinformatics
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Help

Frequently Asked Questions: Data File Formats

General formats:

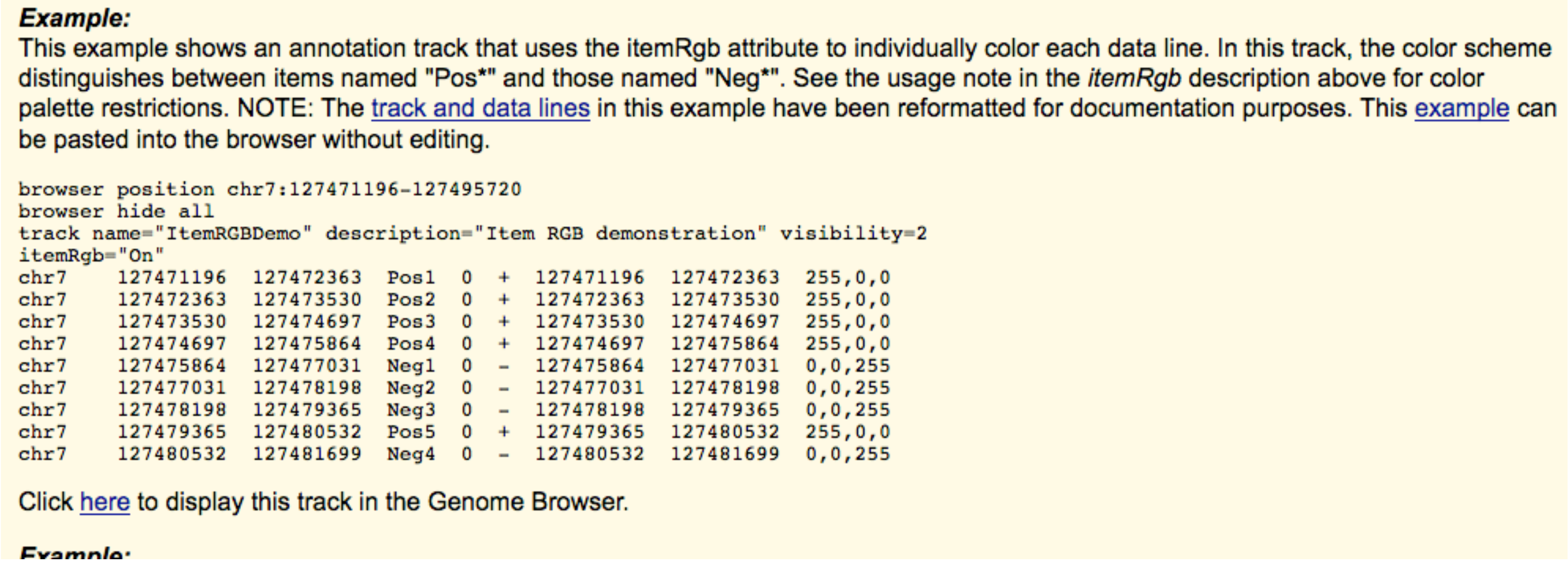
- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigWig format



The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:



Example:
This example shows an annotation track that uses the itemRgb attribute to individually color each data line. In this track, the color scheme distinguishes between items named "Pos*" and those named "Neg*". See the usage note in the *itemRgb* description above for color palette restrictions. NOTE: The track and data lines in this example have been reformatted for documentation purposes. This example can be pasted into the browser without editing.

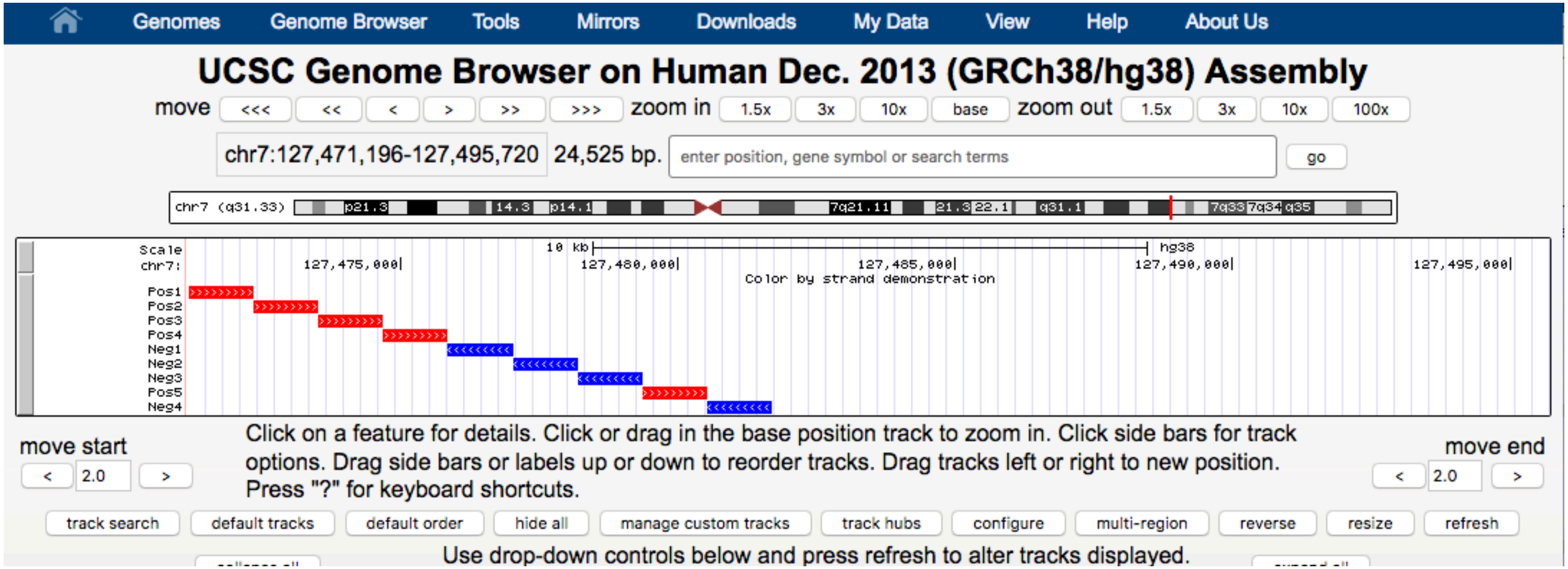
```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2
itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    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
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

Click here to display this track in the Genome Browser.

Example:

Really useful for things like:

- locations of exons
- locations of other “features” -- likes a PCR product or a “peak”



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr7:127,471,196-127,495,720 24,525 bp.

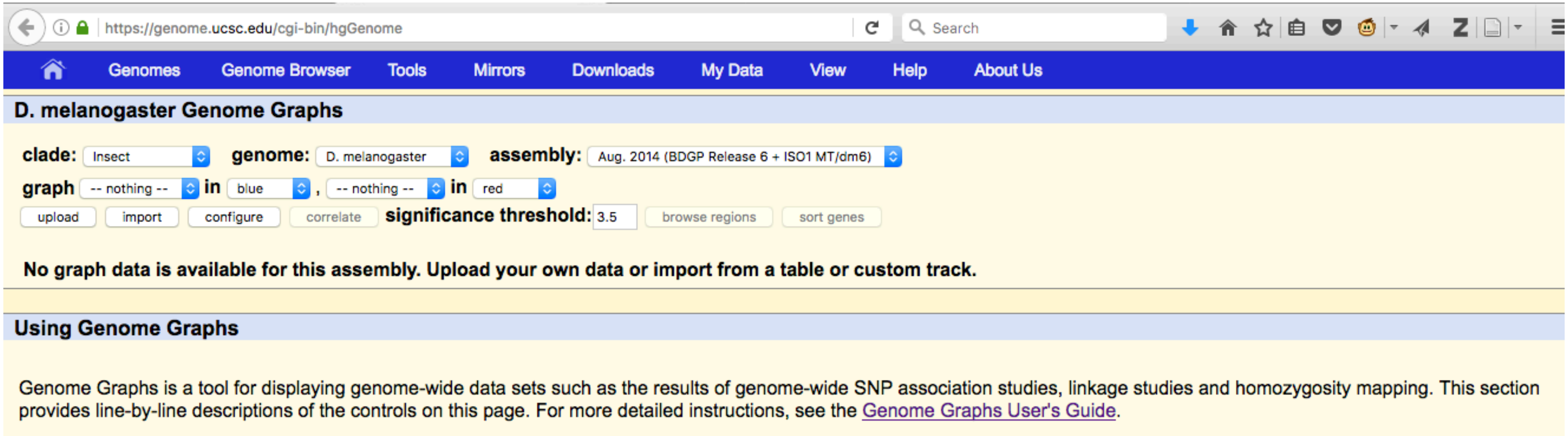
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

Use drop-down controls below and press refresh to alter tracks displayed.

when you paste it in ... make sure you switch to humans!

Genome Graph

quickly add “graph-like object” (under tools)



https://genome.ucsc.edu/cgi-bin/hgGenome

D. melanogaster Genome Graphs

clade: Insect **genome:** D. melanogaster **assembly:** Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)

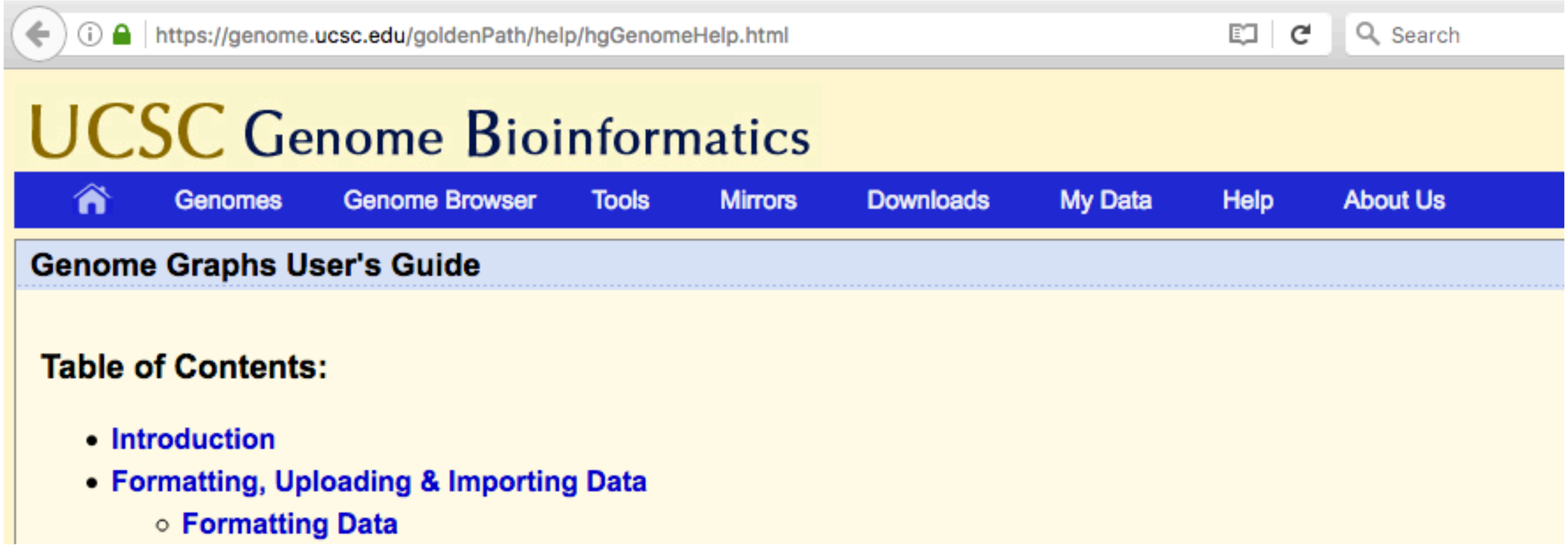
graph -- nothing -- **in** blue , -- nothing -- **in** red

significance threshold: 3.5

No graph data is available for this assembly. Upload your own data or import from a table or custom track.

Using Genome Graphs

Genome Graphs is a tool for displaying genome-wide data sets such as the results of genome-wide SNP association studies, linkage studies and homozygosity mapping. This section provides line-by-line descriptions of the controls on this page. For more detailed instructions, see the Genome Graphs User's Guide.



https://genome.ucsc.edu/goldenPath/help/hgGenomeHelp.html

UCSC Genome Bioinformatics

Genome Graphs User's Guide

Table of Contents:

- Introduction
- Formatting, Uploading & Importing Data
  - Formatting Data

Genome Graph format

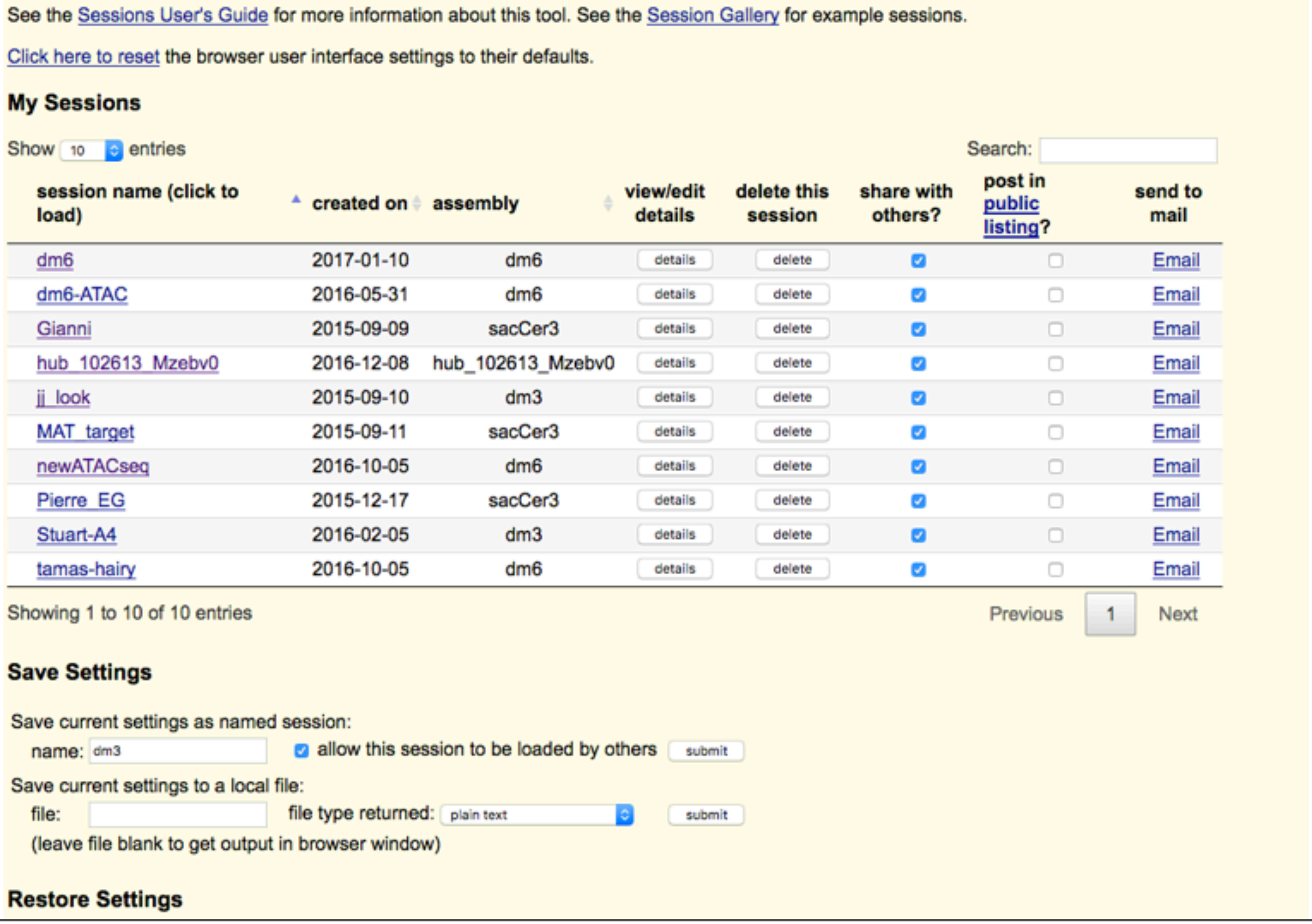
chromosome [tab] basepair [tab] score [return]
chr2L\t1456765\t13.2\n

Really useful for things like:

- LOD scores at markers
- coverage at markers
- HMM states at markers
- etc.

Its fun to share

myData -> Sessions



See the Sessions User's Guide for more information about this tool. See the Session Gallery for example sessions.

Click here to reset the browser user interface settings to their defaults.

My Sessions

Show 10 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
dm6	2017-01-10	dm6	details	delete	☑	☐	Email
dm6-ATAC	2016-05-31	dm6	details	delete	☑	☐	Email
Gianni	2015-09-09	sacCer3	details	delete	☑	☐	Email
hub_102613_Mzebv0	2016-12-08	hub_102613_Mzebv0	details	delete	☑	☐	Email
jj_look	2015-09-10	dm3	details	delete	☑	☐	Email
MAT_target	2015-09-11	sacCer3	details	delete	☑	☐	Email
newATACseq	2016-10-05	dm6	details	delete	☑	☐	Email
Pierre_EG	2015-12-17	sacCer3	details	delete	☑	☐	Email
Stuart-A4	2016-02-05	dm3	details	delete	☑	☐	Email
tamas-hairy	2016-10-05	dm6	details	delete	☑	☐	Email

Showing 1 to 10 of 10 entries

Previous 1 Next

Save Settings

Save current settings as named session:

name: dm3 ☑ allow this session to be loaded by others submit

Save current settings to a local file:

file: file type returned: plain text submit

(leave file blank to get output in browser window)

Restore Settings

..but...

- shared session don’t last forever
- if you use them occasionally they don’t die
- or make your own “track hub”
- https://genome.ucsc.edu/goldenpath/help/hgTrackHubHelp.html
-

...too big to upload...

- limit on size of file you can upload to SCGB
- track hubs are an answer ... but are not quick
- you can host one of the compressed binary index formats supported by the Genome Browser -- these are not uploaded
  - bigBed, bigGenePred, bigPsl, bigChain, bigMaf, bigWig, BAM, CRAM, HAL or VCF (most useful)
  - this is sort of poorly documented
  - https://www.ncbi.nlm.nih.gov/pmc/articles/