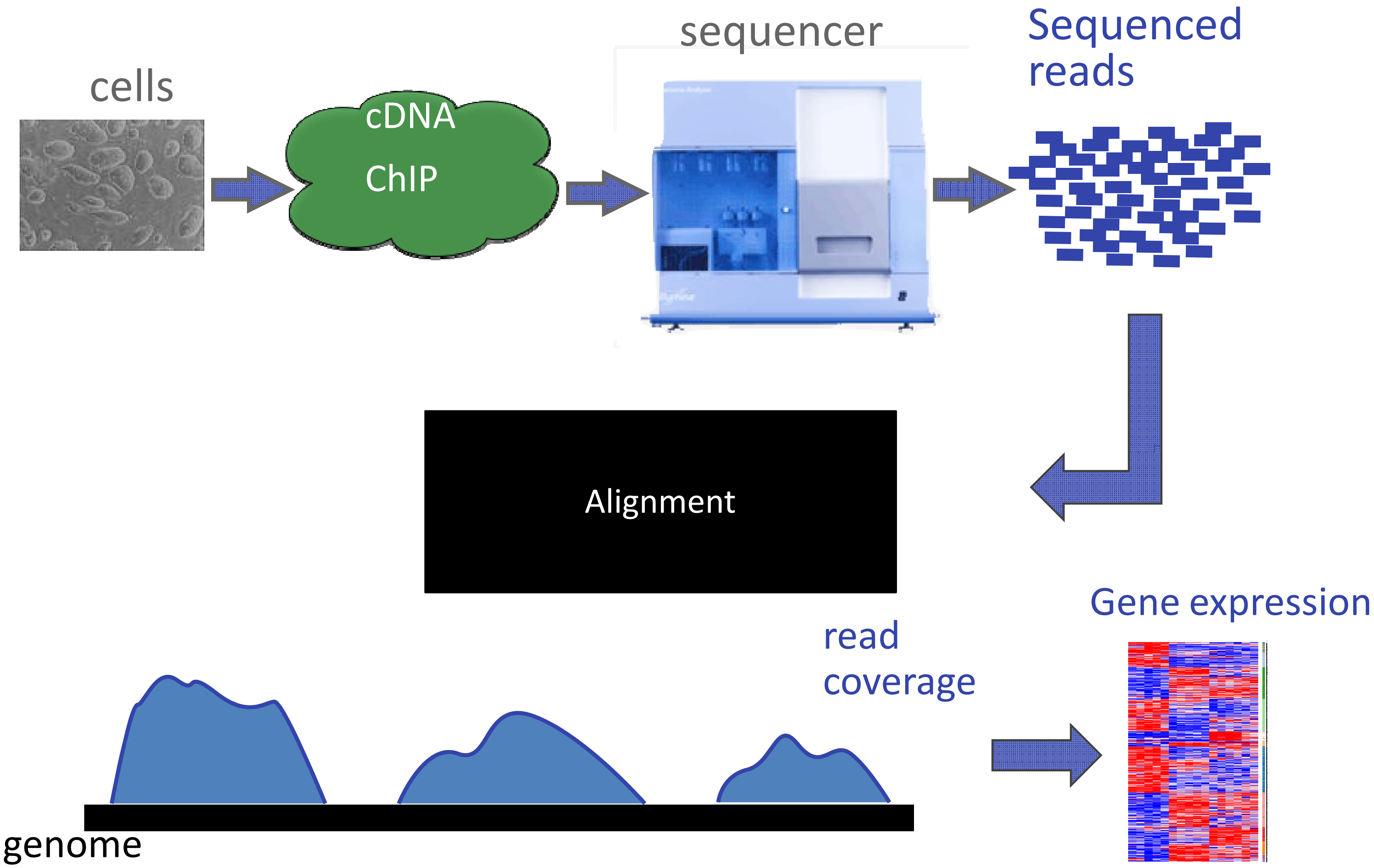
Data: databases, annotations, genomic resources

Week 2

Presenter: Hennady Shulha



cells

cDNA

ChIP

sequencer

Sequenced reads

Alignment

read coverage

Gene expression

genome

Setting up

- Log in into ghpcc06.umassrc.org

```
$ ssh <username>@ghpcc06.umassrc.org
```

- Check that you are in your home directory:

```
$ pwd
$ cd
```

- Create “`biocourse`” directory

```
$ mkdir biocourse
$ cd biocourse
```

Setting up

- Put into current folder file from web

```
$ wget http://biocore.umassmed.edu/biocourse/week2.tar.gz
```

- Unpack compacted files

```
$ tar -xvzf week2.tar.gz
$ cd week2
```

- Unzip this given file

```
$ gunzip reads.fastq.gz
```

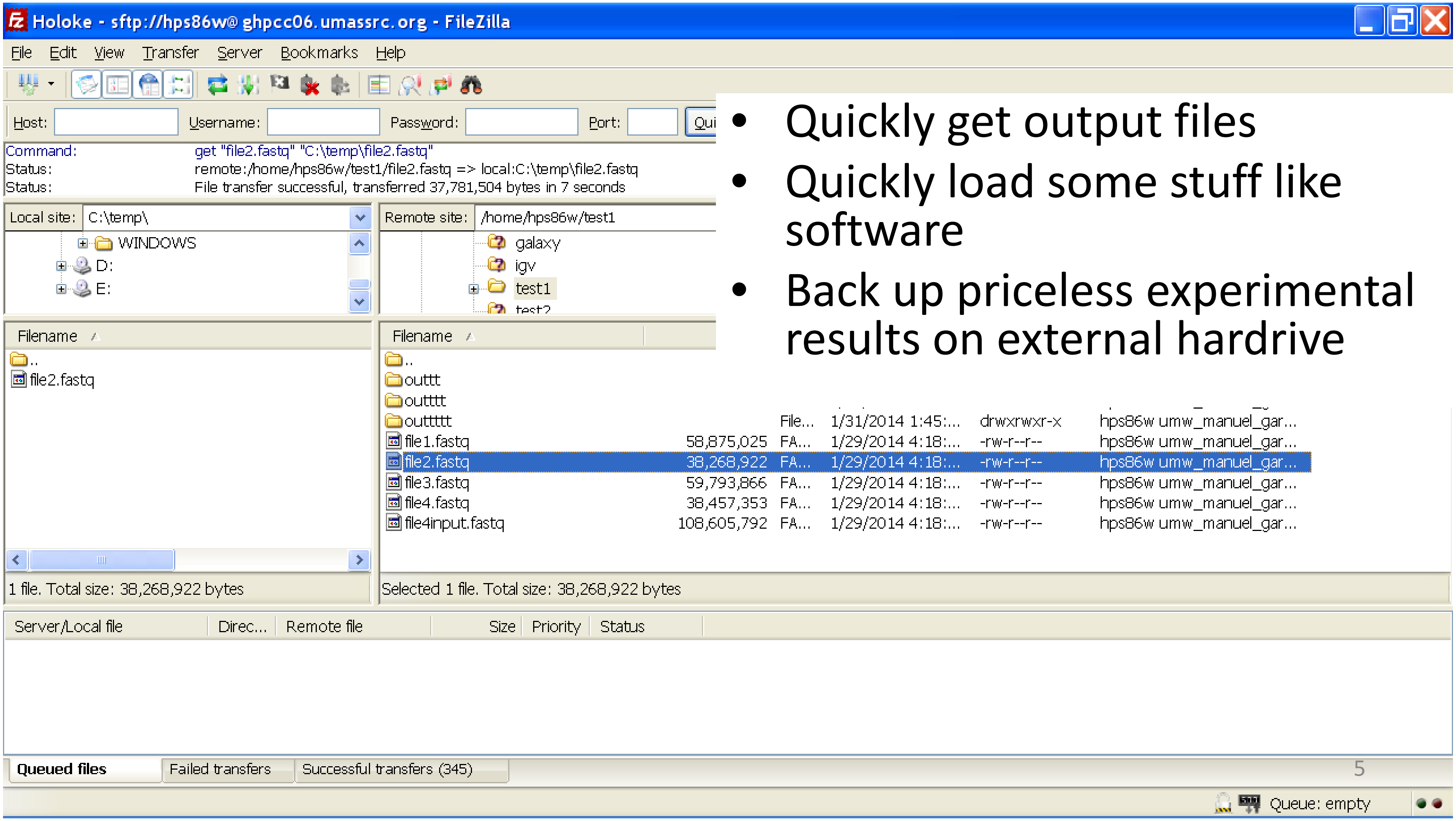
alternative and more common commands for download: “sftp”, “ftp”

sftp/ftp software: Filezilla. https://filezilla-project.org/

File transfer software

Used to connect your PC to any ftp/sftp location

- Quickly get output files
- Quickly load some stuff like software
- Back up priceless experimental results on external hardrive



```
Command:    get "file2.fastq" "C:\temp\file2.fastq"
Status:     remote:/home/hps86w/test1/file2.fastq => local:C:\temp\file2.fastq
Status:     File transfer successful, transferred 37,781,504 bytes in 7 seconds
```

Local site: C:\temp\

```
WINDOWS
D:
E:
```

Remote site: /home/hps86w/test1

```
galaxy
igv
test1
test2
```

```
Filename
..
file2.fastq
```

1 file. Total size: 38,268,922 bytes

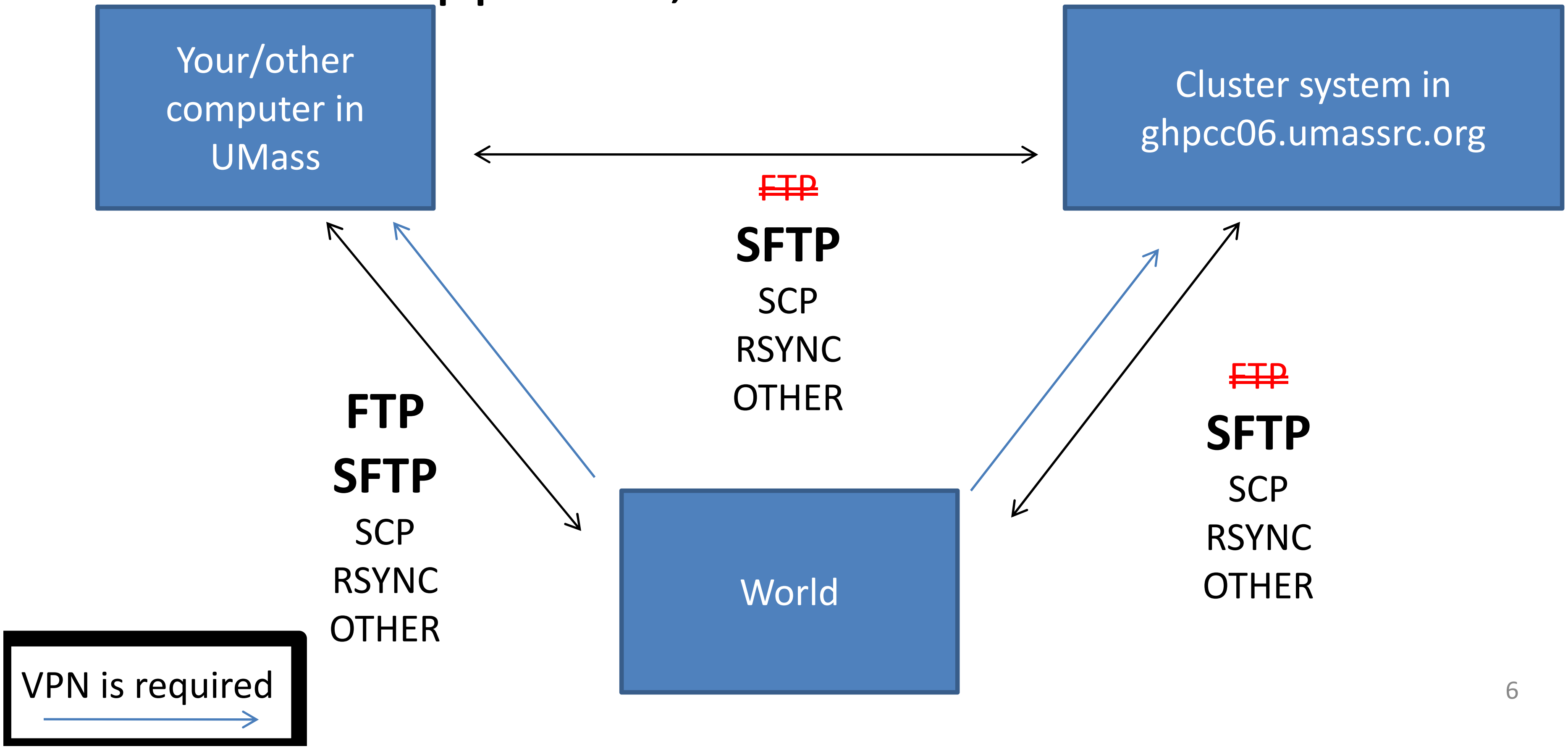
```
Filename
..
outtt
outttt
outtttt                     File...  1/31/2014 1:45:...  drwxrwxr-x   hps86w umw_manuel_gar...
file1.fastq      58,875,025  FA...   1/29/2014 4:18:...  -rw-r--r--   hps86w umw_manuel_gar...
file2.fastq      38,268,922  FA...   1/29/2014 4:18:...  -rw-r--r--   hps86w umw_manuel_gar...
file3.fastq      59,793,866  FA...   1/29/2014 4:18:...  -rw-r--r--   hps86w umw_manuel_gar...
file4.fastq      38,457,353  FA...   1/29/2014 4:18:...  -rw-r--r--   hps86w umw_manuel_gar...
file4input.fastq 108,605,792 FA...   1/29/2014 4:18:...  -rw-r--r--   hps86w umw_manuel_gar...
```

Selected 1 file. Total size: 38,268,922 bytes

Protocols availability

SFTP – supported by Filezilla

SCP – not supported;



Your/other computer in UMass

Cluster system in ghpcc06.umassrc.org

World

~~FTP~~
SFTP
SCP
RSYNC
OTHER

FTP
SFTP
SCP
RSYNC
OTHER

~~FTP~~
SFTP
SCP
RSYNC
OTHER

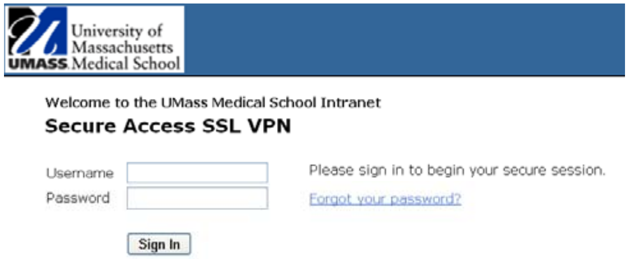
VPN is required

VPN

https://ssl.umassmed.edu

(contact UMass support if you do not have VPN account)

Used to connect your PC through UMass firewall if you are outside of UMass campus.



Welcome to the UMass Medical School Intranet

Secure Access SSL VPN

Username

Password

Please sign in to begin your secure session.

Forgot your password?

How we represent genomic data?

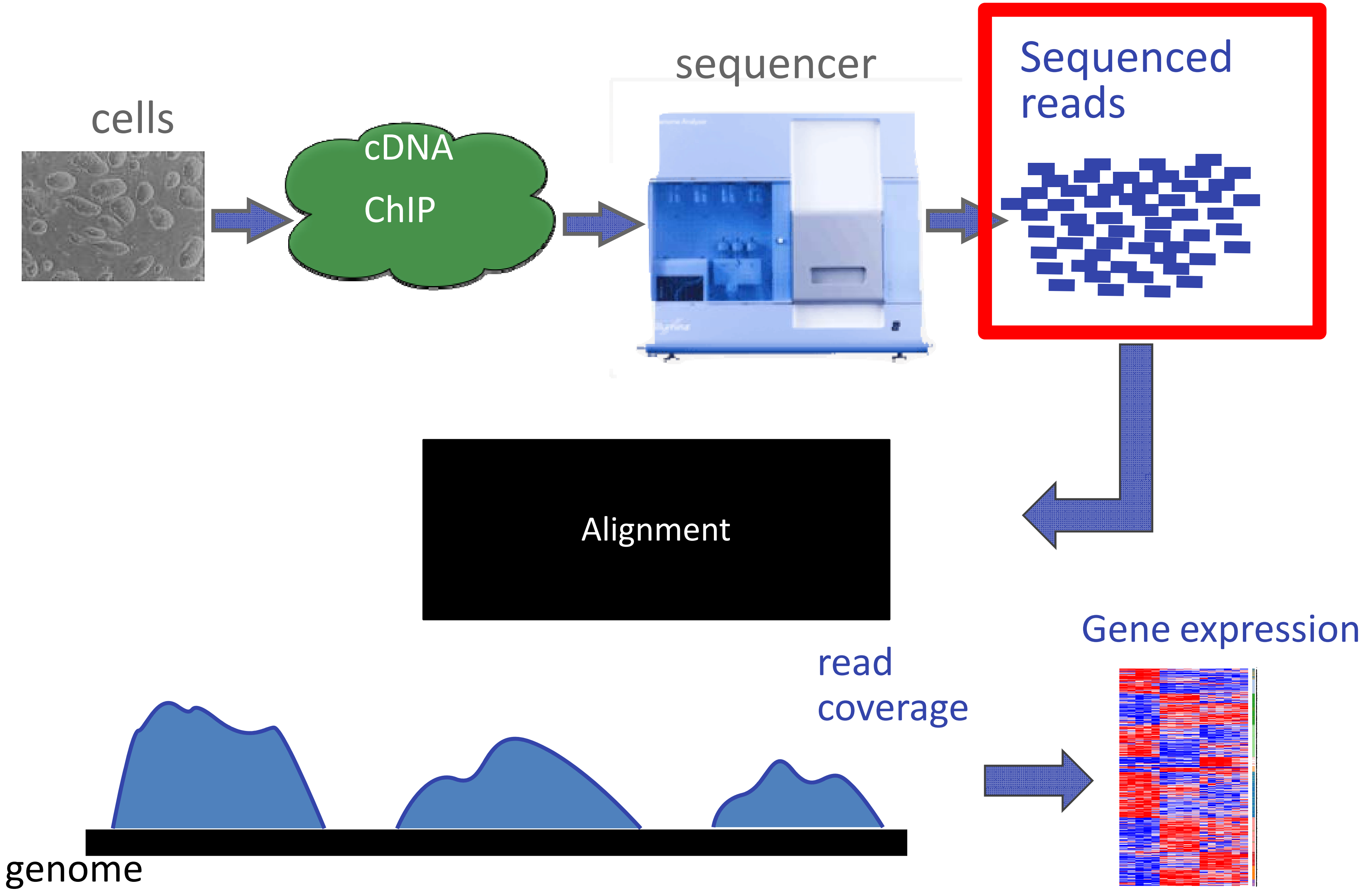
- Genomic sequence
  - Genomes
  - PCR products
- Genomic annotations
  - Genes
  - miRNAs
- Experimental results
  - Sequencing experiment
  - Array hybridization
- Process data for visualization
  - How many reads per base?
  - Probes are on

File formats

- Binary
- Compressed to save space; not directly readable by human; advantage is that only part of file is needed to do visual display
- Text readable
- Can be opened in any text editor; occupy large space (2-3x comparing with binary)

Good command to test what is inside is “head”

http://genome.ucsc.edu/FAQ/FAQformat.html



cells

cDNA

ChIP

sequencer

Sequenced reads

Alignment

read coverage

Gene expression

genome

File formats: Fasta

Used to keep sequences of genomes, proteins etc.

Example:

```
>Sequence_Name Basic description of the sequence
ATCGATCGATGCATGCGAGTCGTAGTCGTAGTCGT
TACGTACGTAGTCGTGTACGTGTAGTCGAGTCGTA
ATCGTACGAGCGAGTCGTTGATGCTGAGTCGTGTC
TACGATGCGAGGCTGTAGTCGTAGTCGTAGTGTCC
TACGACGTGTATGCGTACGGATCGCGATTCGTAGC
```

File formats: Fastq

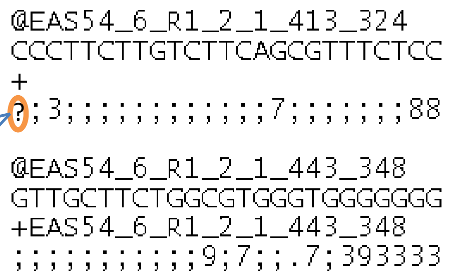
Used mainly to store reads and information about quality

Example:

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88

@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

File formats: Fastq quality

- Quality values



```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
?;3;;;;;;;;;;;;7;;;;;;;88

@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

FASTQ (Phred)

Q = NUMERICS_ID_of_SYMBOL – 33

$Q = -10 \log_{10} P$ where P is probability of incorrect base calling

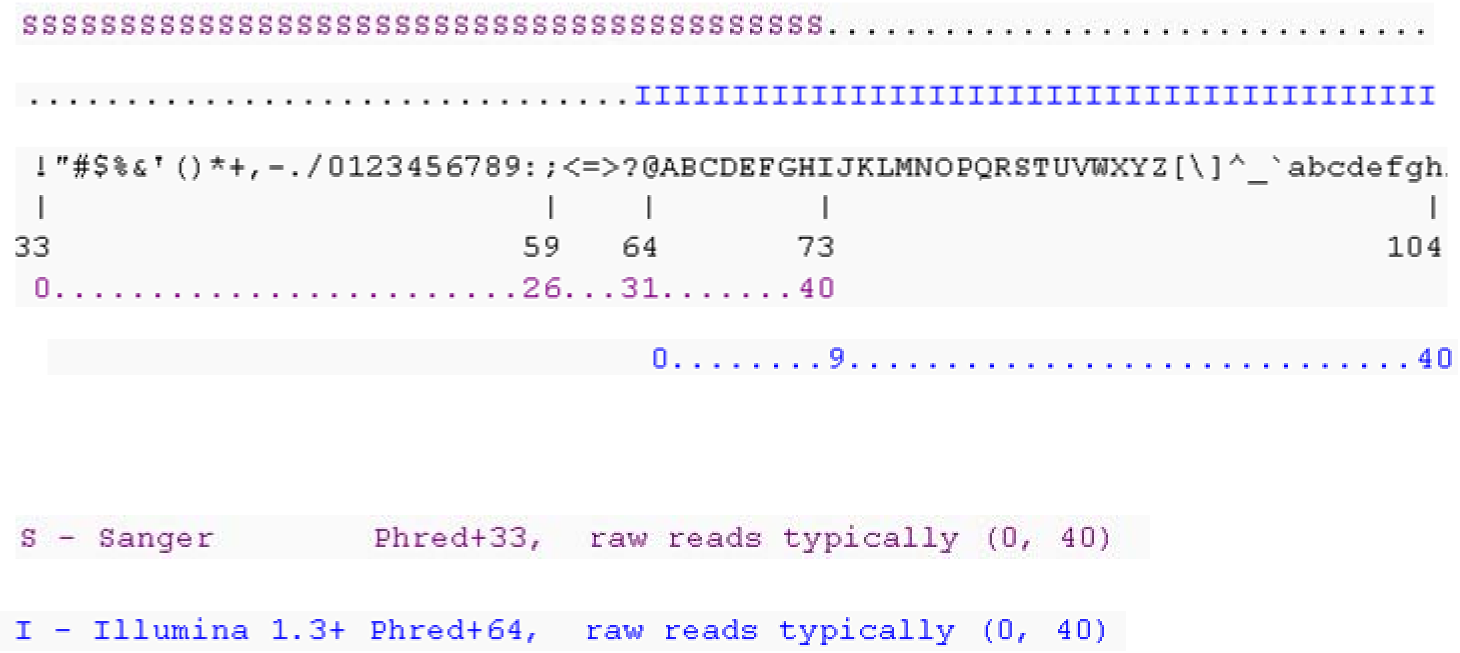
Solexa

Q = 10 * log(1 + 10 ** (NUMERICS_ID_of_SYMBOL - 64) / 10.0)) / log(10)

$Q_{\text{solexa-prior to v.1.3}} = -10 \log_{10} \frac{p}{1-p}$ where P is probability of incorrect base calling

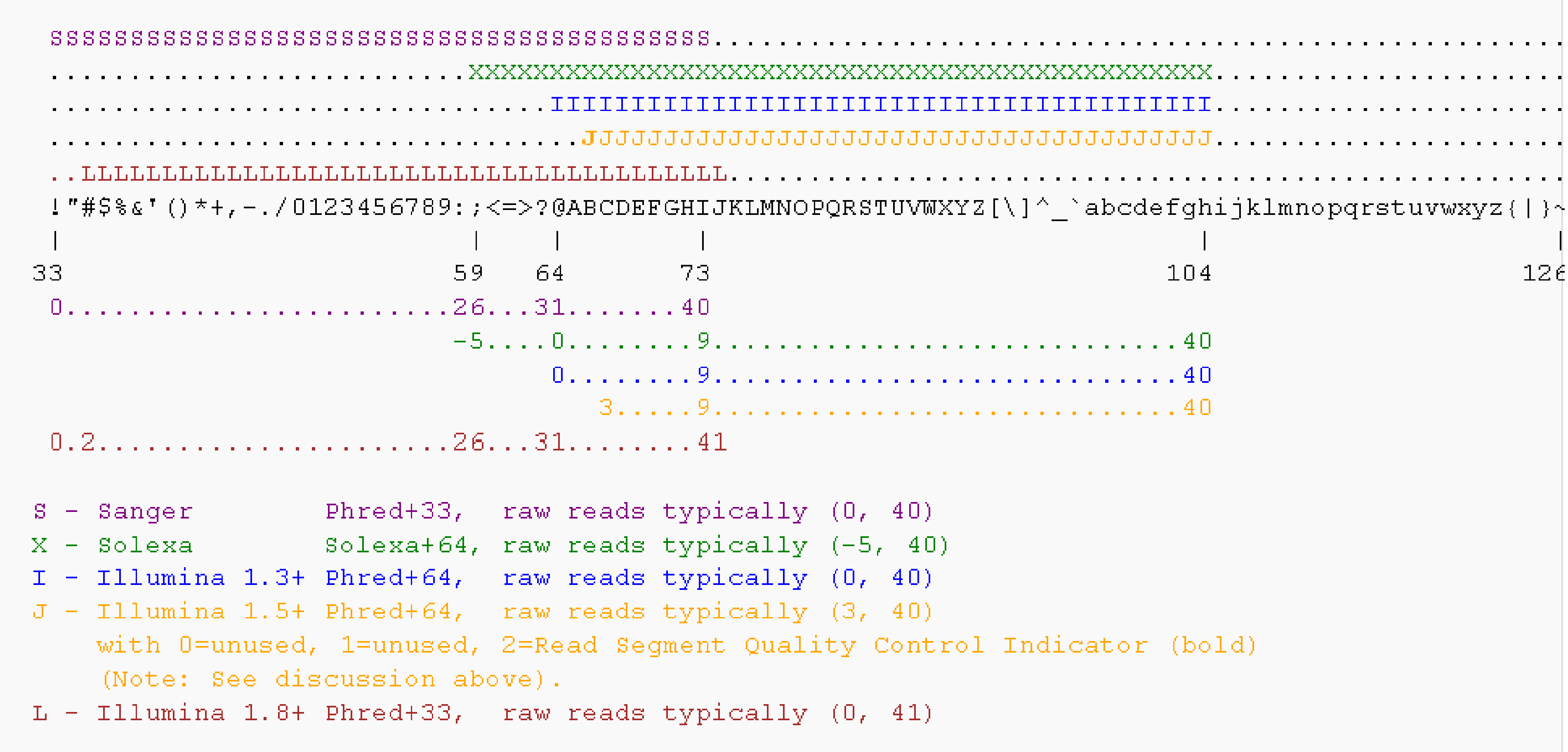
(later they had different Phred related versions (Phred+33; Phred+64))

File formats: Fastq versions



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS...............................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
 |                         |    |        |                              |
33                        59   64       73                            104
 0........................26...31.......40
                                0........9.............................40


S - Sanger        Phred+33,  raw reads typically (0, 40)

I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
```

File formats: Fastq versions



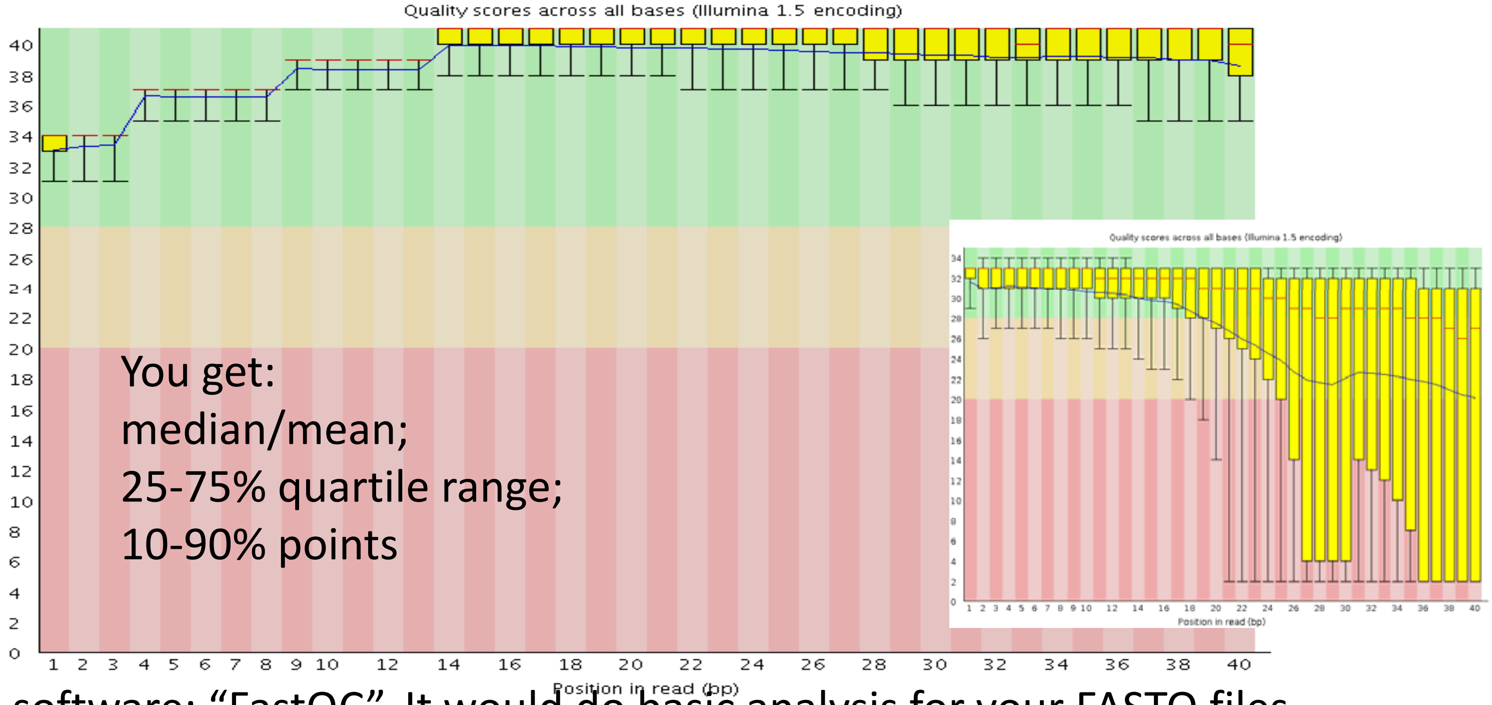
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  ..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)
```

I.e. if you are going to use quality scores (by aligners) – would be better to know a source that generated it!!!

Fastq quality

- Alternative: check if general picture is fine and not use it.

h
d
`
\
X
T
P
L
H
D
@

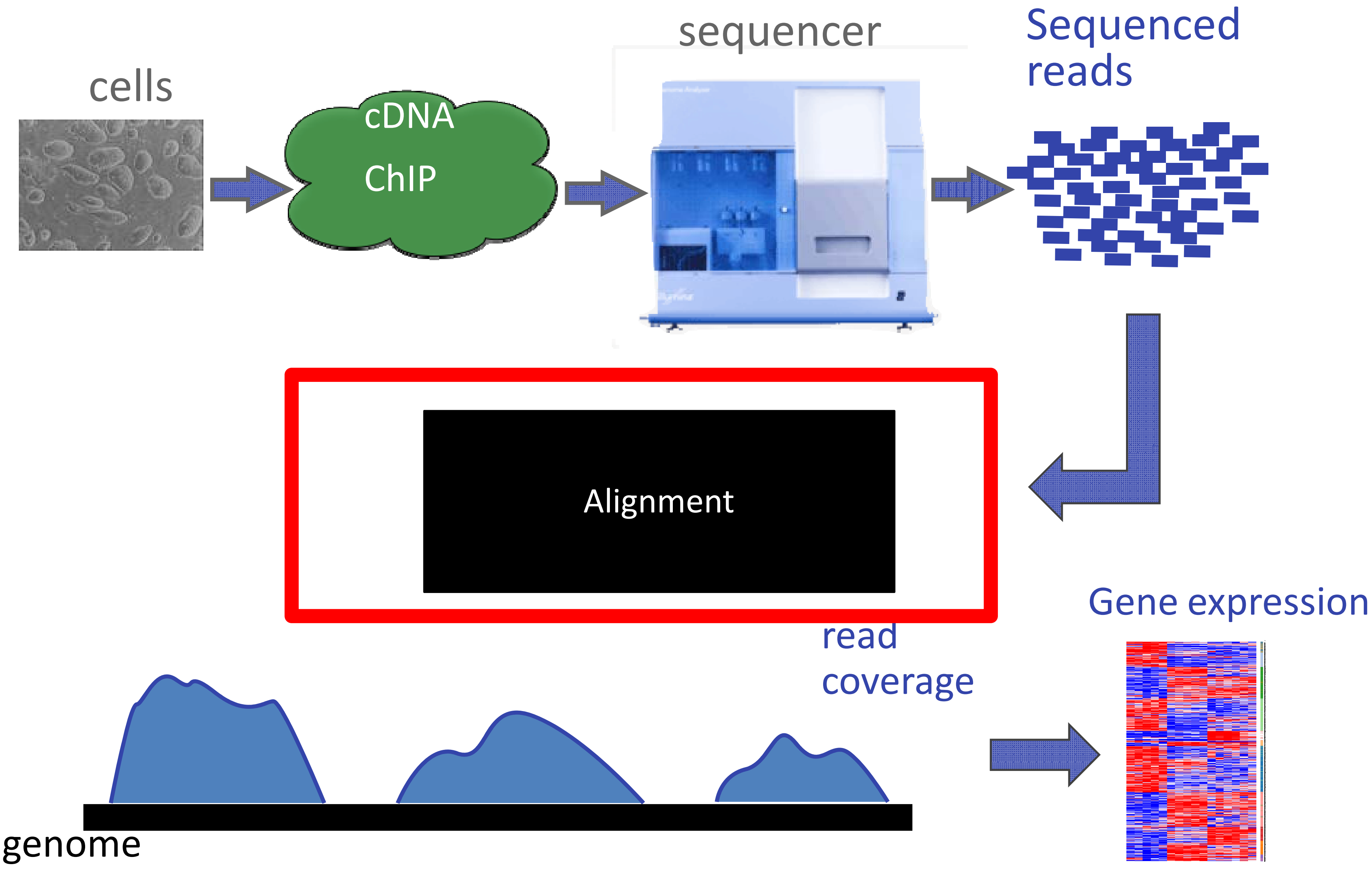


You get:
median/mean;
25-75% quartile range;
10-90% points

Very useful software: “FastQC”. It would do basic analysis for your FASTQ files.

```
$ module load fastqc/0.10.1
$ fastqc -help
$ fastqc  reads.fastq
```

```
$  module avail
```



cells

cDNA

ChIP

sequencer

Sequenced reads

Alignment

read coverage

genome

Gene expression

BAM/SAM: Read alignment format

- Source of a bam file
  - Fastq → aligner → BAM: Alignment results
- Aligners:
  - Bowtie: Aligns contiguous reads
  - Tophat: Aligns spliced reads

To load bowtie:

```
$  module load bowtie/1.0.0
```

File formats: SAM/BAM

Information about genomic locations, plus other useful things like base qualities, number of mismatches etc.

```
@HD VN:1.0
@SQ SN:1 LN:249250621
Sequence1 113 chr1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
0;==-==9;>>>>>>=>>>>>>>>>>>=>>>>>>>>>
XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0
```

BAM contains the same info but in compressed, binary format.

http://samtools.sourceforge.net/SAMv1.pdf

File formats: SAM/BAM

```
@HD VN:1.0
@SQ SN:1 LN:249250621
```

Header: basic description; sizes of chromosomes; other

http://samtools.sourceforge.net/SAMv1.pdf

File formats: SAM/BAM

```
@HD VN:1.0
@SQ SN:1 LN:249250621
Sequence1
```

Sequence name: whatever the read was named in fastq file

http://samtools.sourceforge.net/SAMv1.pdf

File formats: SAM/BAM

```
@HD VN:1.0
@SQ SN:1 LN:249250621
Sequence1 113 chr1 497
```

Genomic position: position how it was mapped by a mapper

http://samtools.sourceforge.net/SAMv1.pdf

File formats: SAM/BAM

```
@HD VN:1.0
@SQ SN:1 LN:249250621
Sequence1 113 chr1 497 37 37M
```

CIGAR string: contains information about mapping like gaps.

http://samtools.sourceforge.net/SAMv1.pdf

File formats: SAM/BAM

```
@HD VN:1.0
@SQ SN:1 LN:249250621
Sequence1 113 chr1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
```

Sequence by itself: mapped sequence

http://samtools.sourceforge.net/SAMv1.pdf

File formats: SAM/BAM

```
@HD VN:1.0
@SQ SN:1 LN:249250621
Sequence1 113 chr1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>
```

Quality: quality from FASTQ file

http://samtools.sourceforge.net/SAMv1.pdf

File formats: SAM/BAM

```
@HD VN:1.0
@SQ SN:1 LN:249250621
Sequence1 113 chr1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>
XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0
```

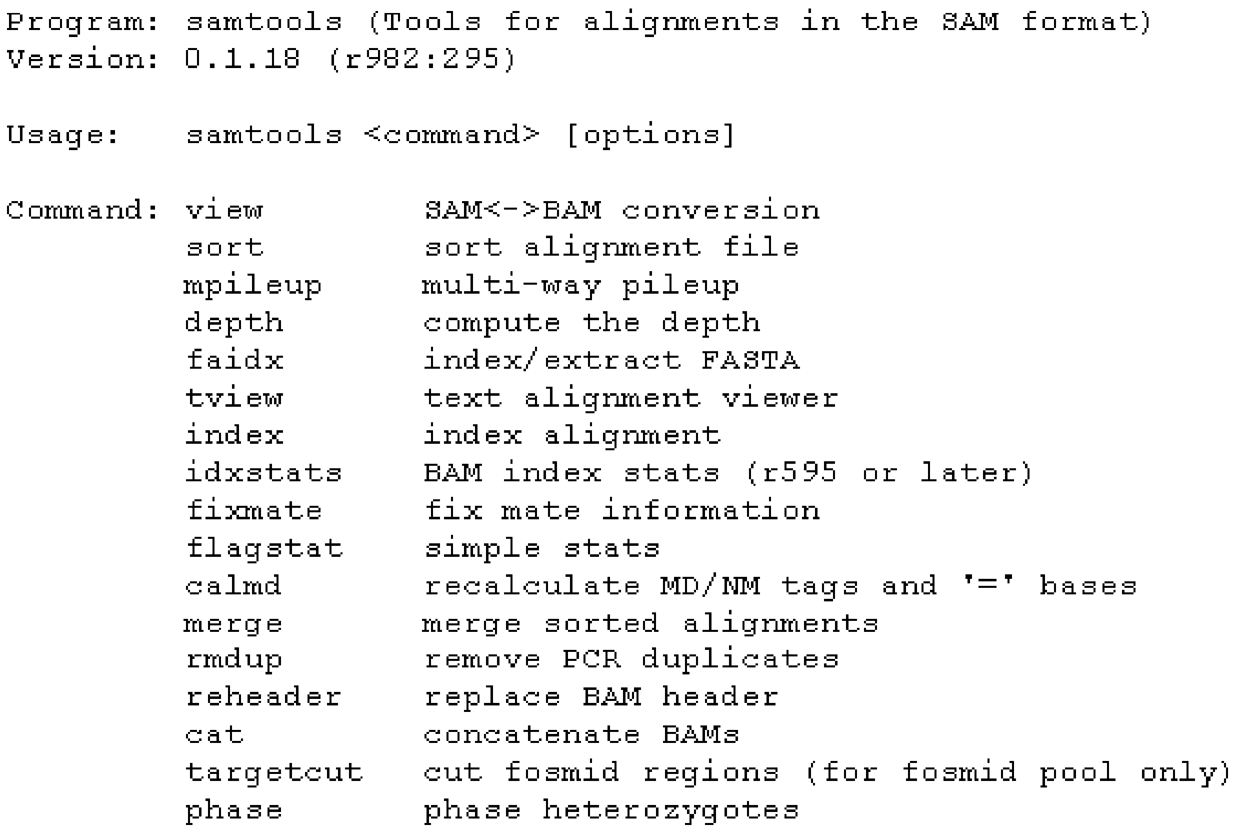
Other info: for example, NM:i:0 means 0 mismatches

http://samtools.sourceforge.net/SAMv1.pdf

File conversion

- SAMTOOLS

```
$ module load samtools/0.0.19
```



```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.18 (r982:295)

Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         idxstats    BAM index stats (r595 or later)
         fixmate     fix mate information
         flagstat    simple stats
         calmd       recalculate MD/NM tags and '=' bases
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
         reheader    replace BAM header
         cat         concatenate BAMs
         targetcut   cut fosmid regions (for fosmid pool only)
         phase       phase heterozygotes
```

File conversion

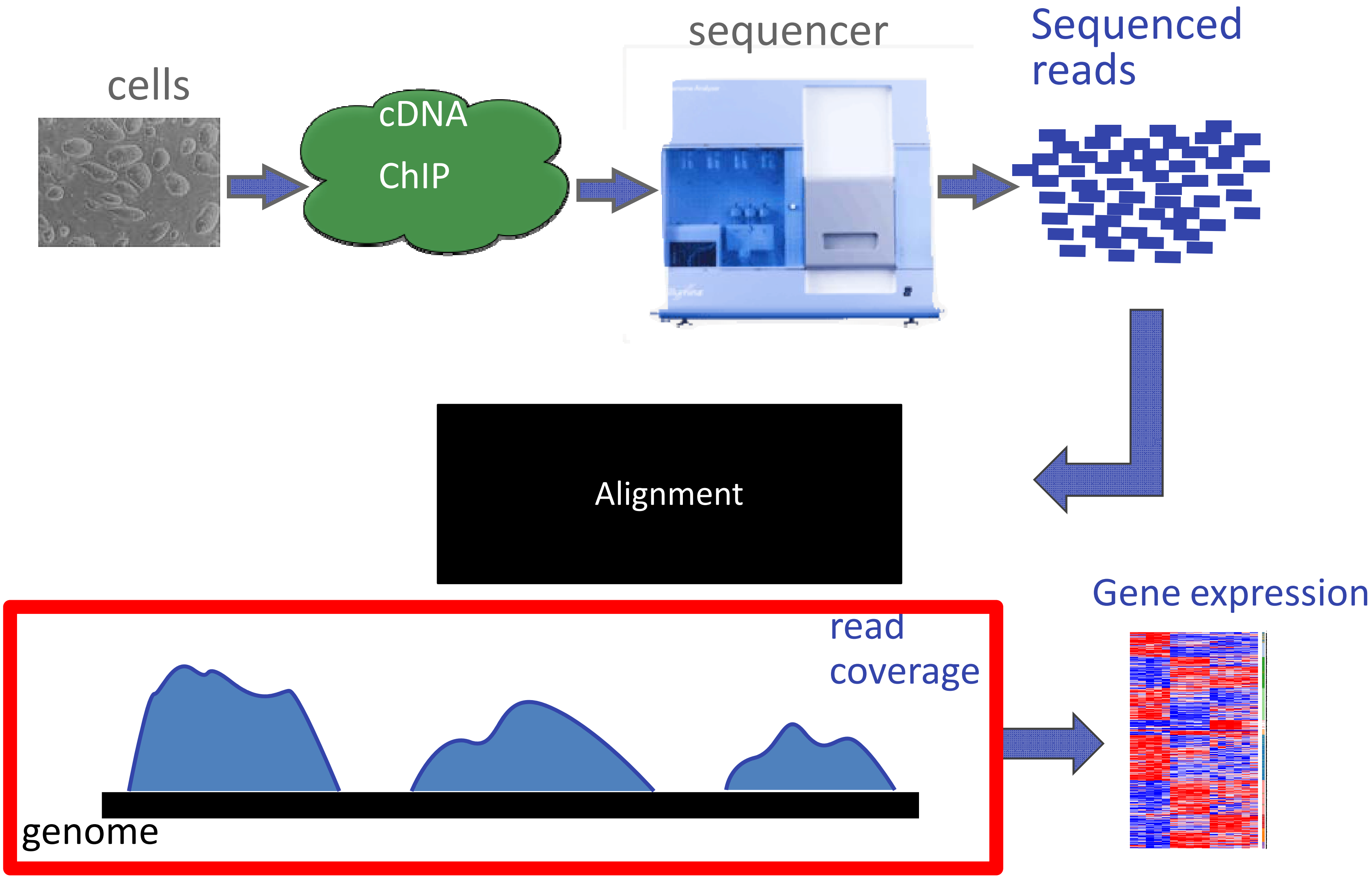
- SAM<->BAM conversion

```
$ samtools view -b -S reads.sam > reads.bam
[samopen] SAM header is present: 25 sequences.
```

```
$ samtools view -h reads.bam | less
```

```
$ samtools sort
Usage: samtools sort [-on] [-m <maxMem>] <in.bam> <out.prefix>
```

```
$ samtools sort reads.bam sorted
```



cells

cDNA

ChIP

sequencer

Sequenced reads

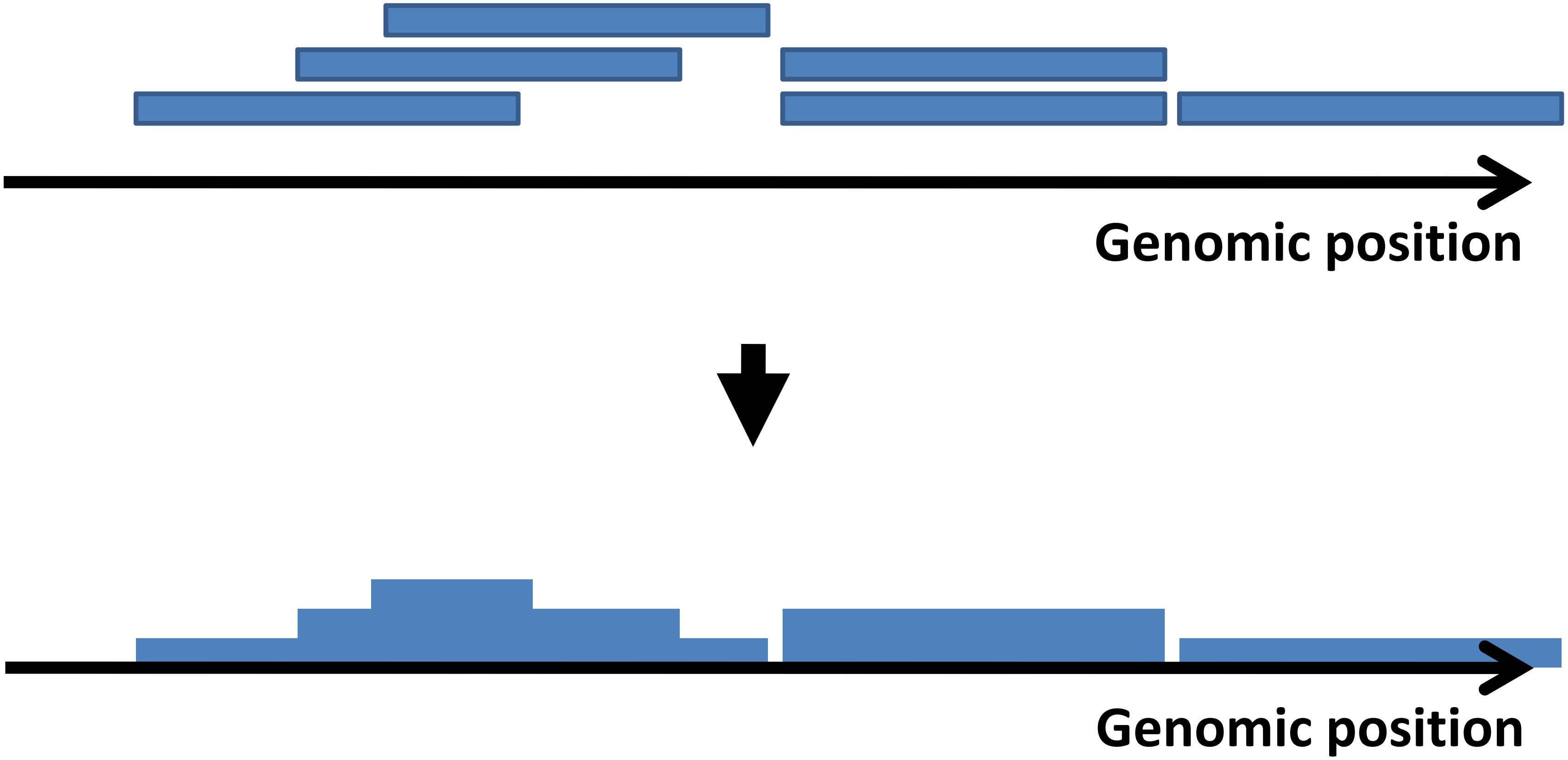
Alignment

read coverage

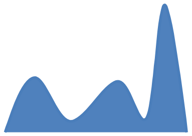
genome

Gene expression

Aggregation



File formats: WIG/bigWIG



Aggregated information about genomic locations.

VARIABLE Step

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

is equivalent to:

```
variableStep chrom=chr2 span=5
300701 12.5
```

bigWIG contains the same info but in compressed, binary format.

File formats: WIG/bigWIG

Aggregated information about genomic locations.

FIXED Step

```
fixedStep chrom=chr3 start=400601 step=100
11
22
33
```

displays the values 11, 22, and 33 as single-base regions on chromosome 3 at positions 400601, 400701, and 400801, respectively. Adding span=5 to the declaration line:

```
fixedStep chrom=chr3 start=400601 step=100 span=5
11
22
33
```

Tools

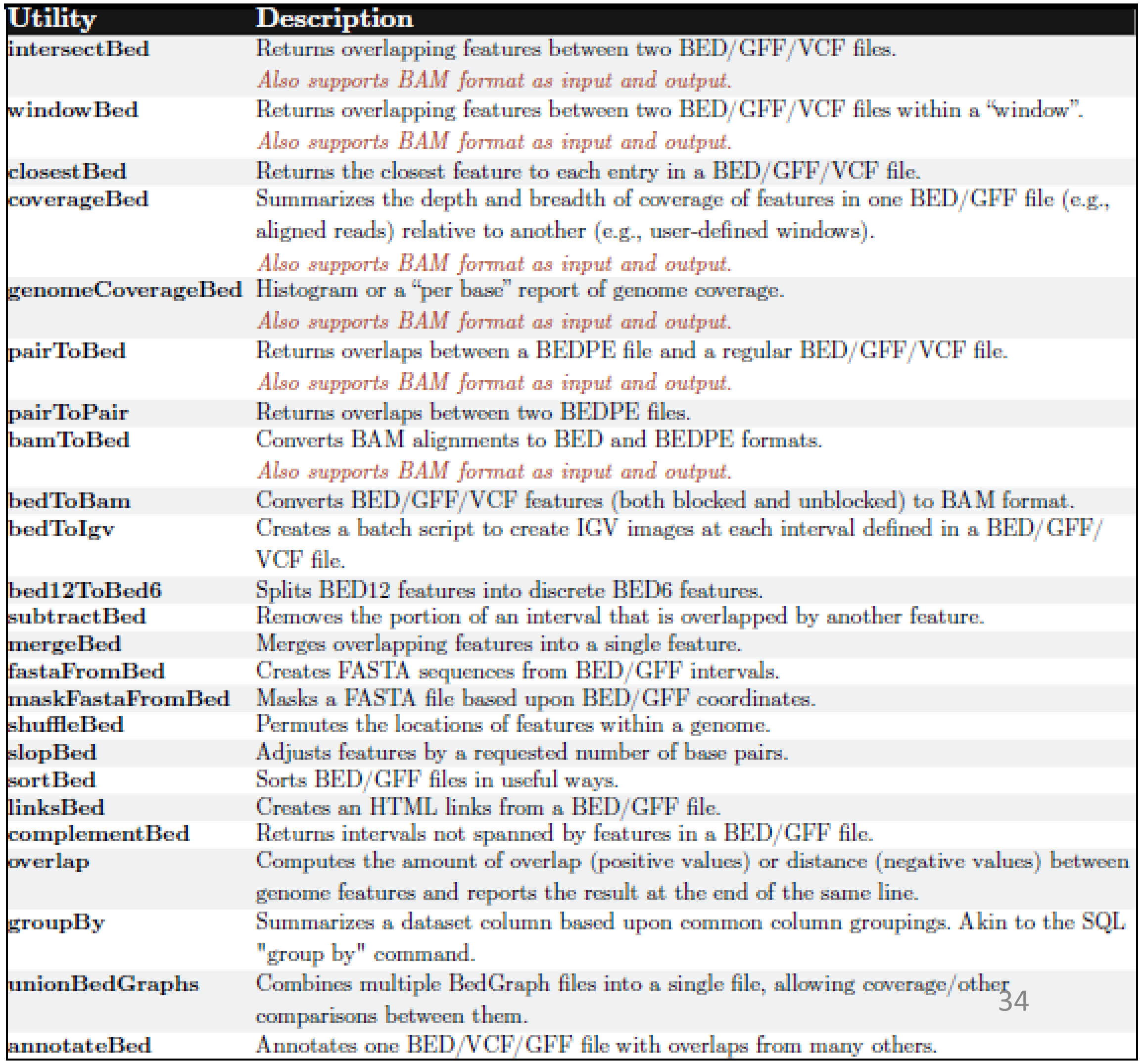
- Bedtools (BED manipulation but BAM support is available)

https://bedtools.googlecode.com/files/BEDTools-User-Manual.v4.pdf

```
$ module load bedtools/2.17.0

$ head a.bed
chr1  100   200
chr1  1000   2000

$ head b.bed
chr1  150   250

$ intersectBed -a a.bed -b b.bed
chr1  150   200
```

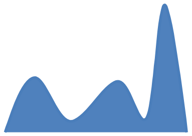
Tools

- Bedtools



Utility	Description
intersectBed	Returns overlapping features between two BED/GFF/VCF files. *Also supports BAM format as input and output.*
windowBed	Returns overlapping features between two BED/GFF/VCF files within a "window". *Also supports BAM format as input and output.*
closestBed	Returns the closest feature to each entry in a BED/GFF/VCF file.
coverageBed	Summarizes the depth and breadth of coverage of features in one BED/GFF file (e.g., aligned reads) relative to another (e.g., user-defined windows). *Also supports BAM format as input and output.*
genomeCoverageBed	Histogram or a "per base" report of genome coverage. *Also supports BAM format as input and output.*
pairToBed	Returns overlaps between a BEDPE file and a regular BED/GFF/VCF file. *Also supports BAM format as input and output.*
pairToPair	Returns overlaps between two BEDPE files.
bamToBed	Converts BAM alignments to BED and BEDPE formats. *Also supports BAM format as input and output.*
bedToBam	Converts BED/GFF/VCF features (both blocked and unblocked) to BAM format.
bedToIgv	Creates a batch script to create IGV images at each interval defined in a BED/GFF/VCF file.
bed12ToBed6	Splits BED12 features into discrete BED6 features.
subtractBed	Removes the portion of an interval that is overlapped by another feature.
mergeBed	Merges overlapping features into a single feature.
fastaFromBed	Creates FASTA sequences from BED/GFF intervals.
maskFastaFromBed	Masks a FASTA file based upon BED/GFF coordinates.
shuffleBed	Permutes the locations of features within a genome.
slopBed	Adjusts features by a requested number of base pairs.
sortBed	Sorts BED/GFF files in useful ways.
linksBed	Creates an HTML links from a BED/GFF file.
complementBed	Returns intervals not spanned by features in a BED/GFF file.
overlap	Computes the amount of overlap (positive values) or distance (negative values) between genome features and reports the result at the end of the same line.
groupBy	Summarizes a dataset column based upon common column groupings. Akin to the SQL "group by" command.
unionBedGraphs	Combines multiple BedGraph files into a single file, allowing coverage/other comparisons between them.
annotateBed	Annotates one BED/VCF/GFF file with overlaps from many others.

File formats: bedGraph



CHR	Start	End	Value
chr1	1	100	100
chr1	200	300	20

```
$ genomeCoverageBed -bg -ibam sorted.bam -g mm10.chrom.sizes >
out.bg
```

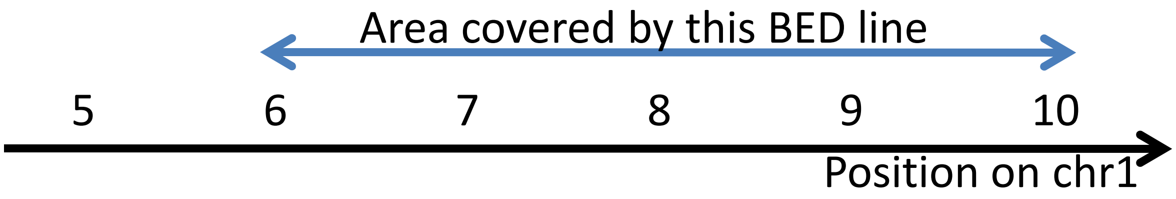
track type=bedGraph name=test_track

File formats: BED

Information about genomic locations. TAB delimited.

- Simple case:

chr1 5 10



Area covered by this BED line

5 6 7 8 9 10

Position on chr1

- Advanced (12 columns, details in the link on slide9):

chr22 100 500 cloneA 96 + 100 500 0 2 4,5, 0,35
chr22 200 600 cloneB 90 - 200 600 0 2 4,5, 0,36

File formats: GTF

Gene information (usually from some external databases)

TAB separated information

<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

```
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1"
```

Visualization

- UCSC browser

  http://genome.ucsc.edu

  http://biocore.umassmed.edu/ucsc.html

- ENSEMBL

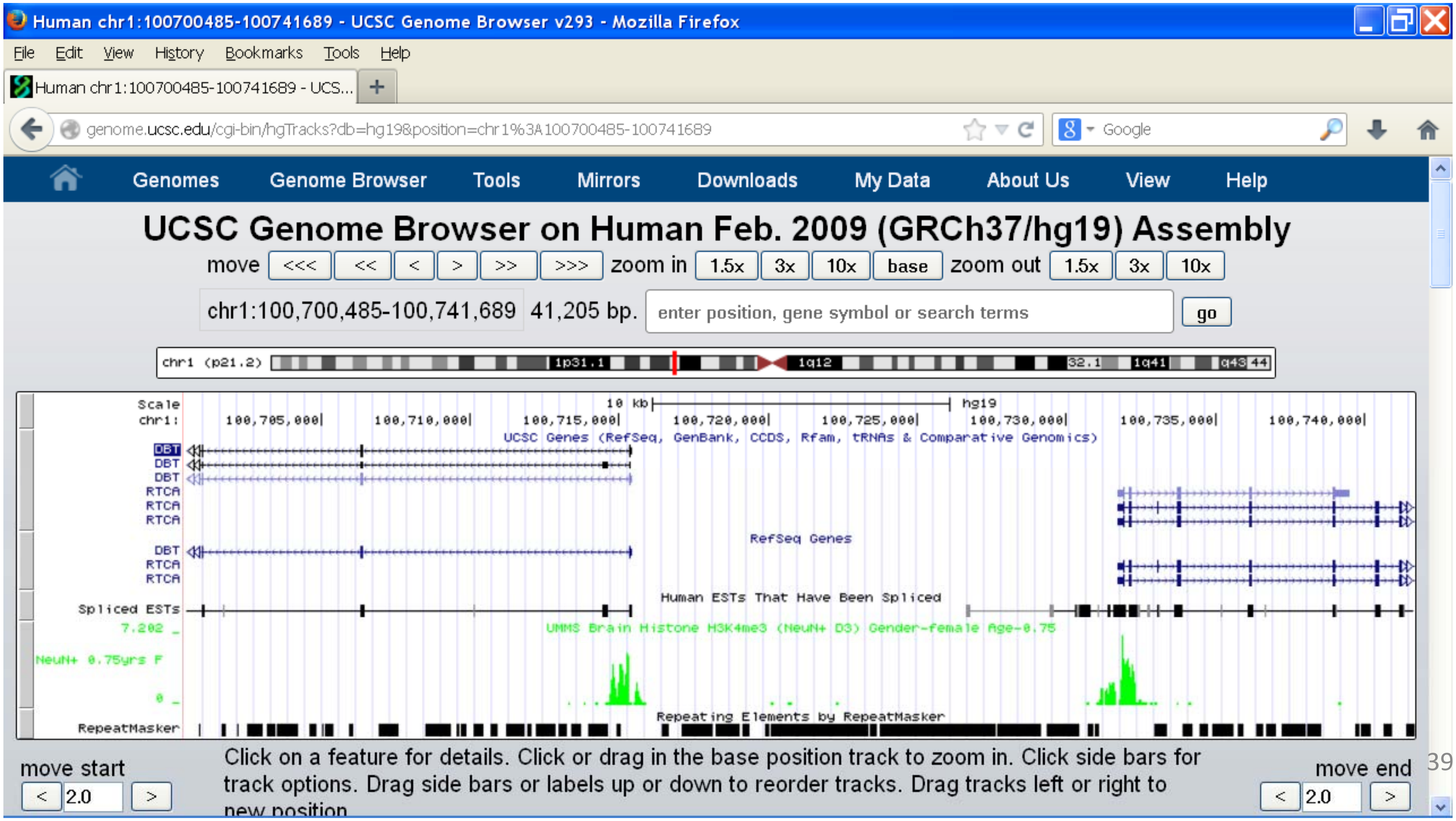
  http://www.ensembl.org/index.html

- IGV browser

  http://www.broadinstitute.org/igv

Visualization

- UCSC – handles bunch of different things



Human chr1:100700485-100741689 - UCSC Genome Browser v293 - Mozilla Firefox

genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr1%3A100700485-100741689

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr1:100,700,485-100,741,689 41,205 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.

Visualization

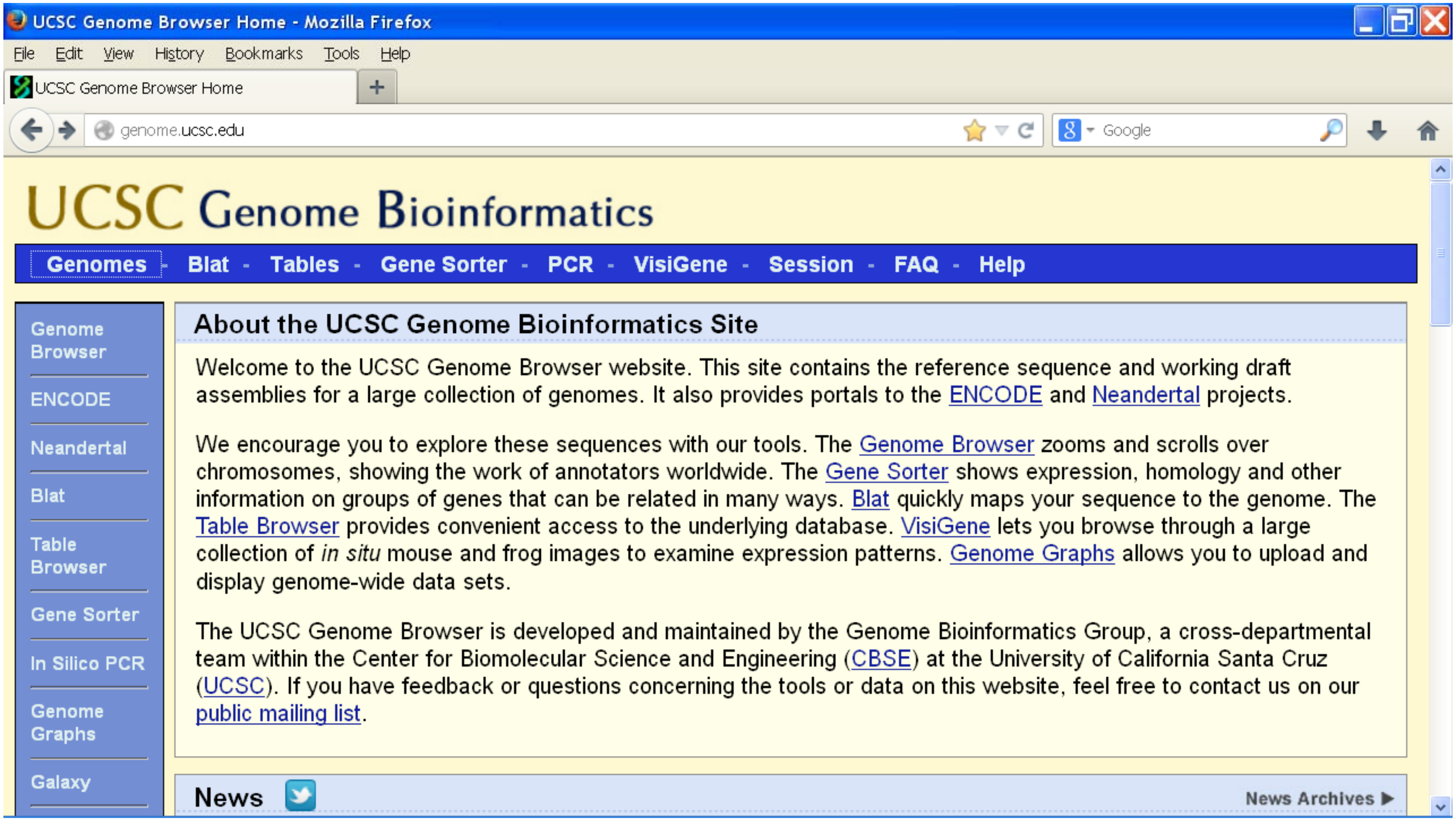
http://biocore.umassmed.edu/ucsc.html

to use for visualization, any other activities where you would expect heavy competition on ucsc.edu.

http://genome.ucsc.edu/index.html

to use for tools that are not under heavy competition like BLAT, datatables download.

Visualization


UCSC Genome Browser Home - Mozilla Firefox
File Edit View History Bookmarks Tools Help
UCSC Genome Browser Home
genome.ucsc.edu
Google
UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help
Genome Browser
ENCODE
Neandertal
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
News
News Archives ▶

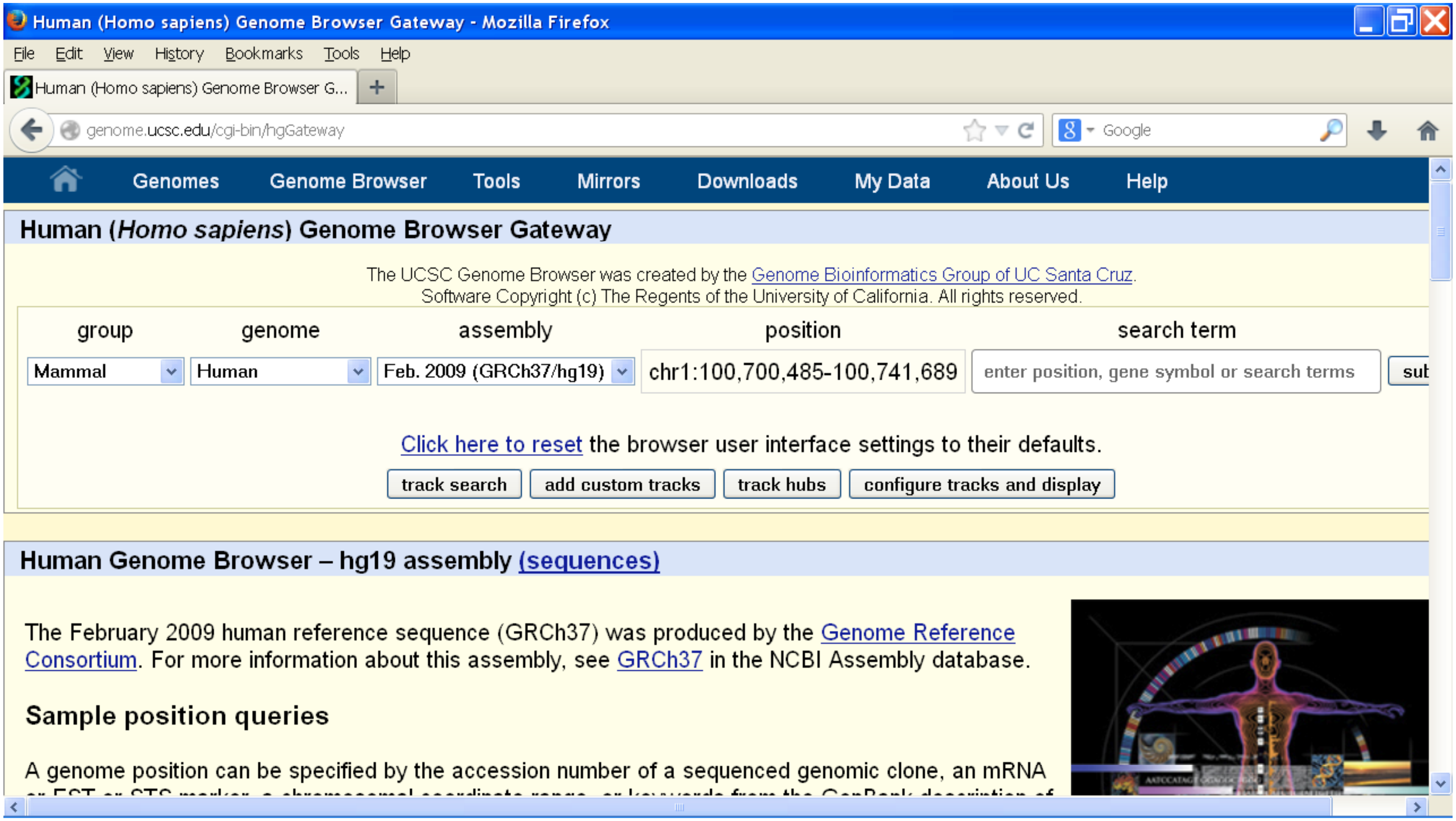
About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

Visualization



Human (Homo sapiens) Genome Browser Gateway - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Human (Homo sapiens) Genome Browser G...
genome.ucsc.edu/cgi-bin/hgGateway
Google
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
Help
track search
add custom tracks
track hubs
configure tracks and display
sub

Human (*Homo sapiens*) Genome Browser Gateway

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.

group	genome	assembly	position	search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr1:100,700,485-100,741,689	enter position, gene symbol or search terms

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

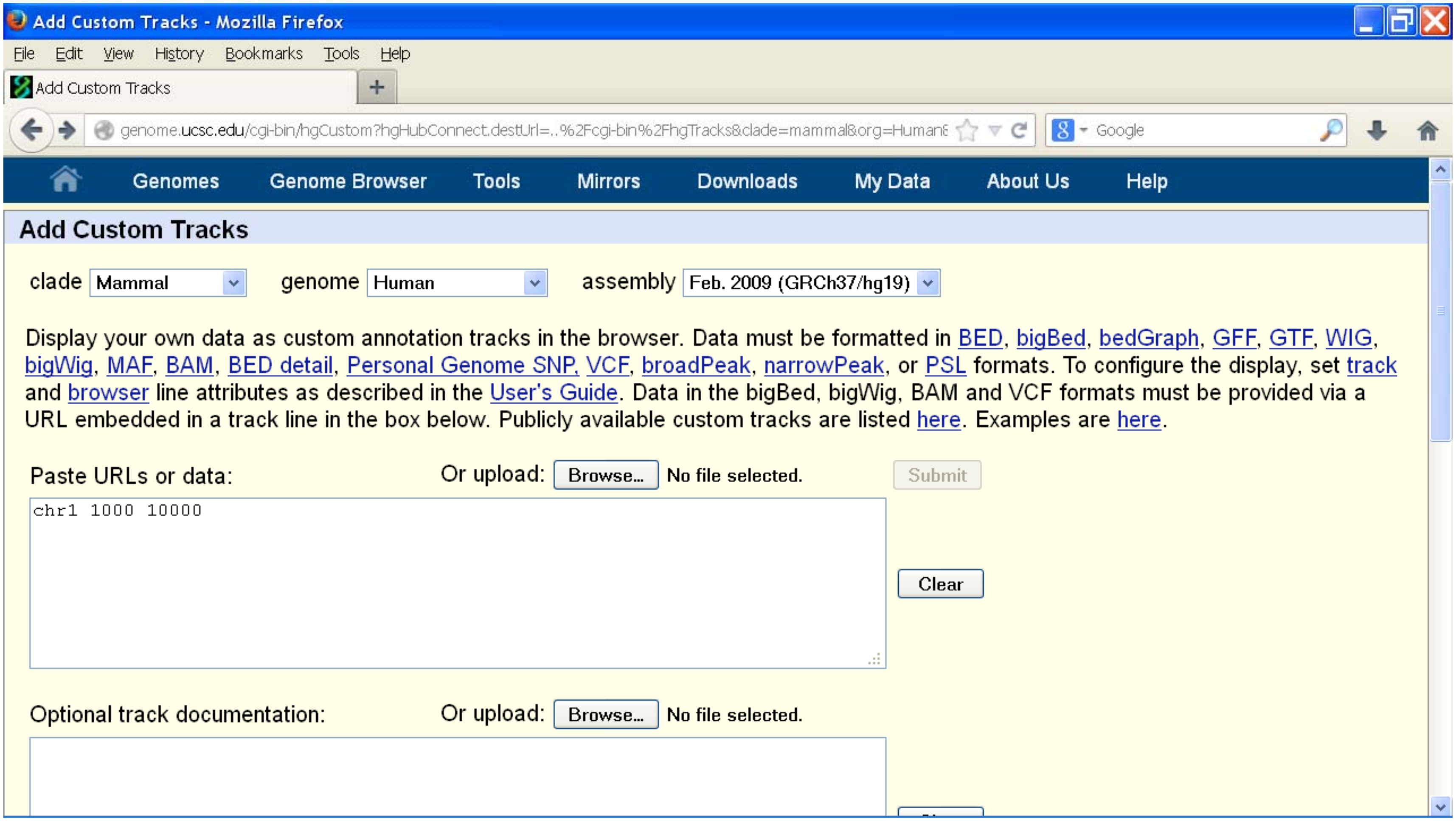
Human Genome Browser – hg19 assembly (sequences)

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of

Visualization


Add Custom Tracks - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Add Custom Tracks
genome.ucsc.edu/cgi-bin/hgCustom?hgHubConnect.destUrl=..%2Fcgi-bin%2FhgTracks&clade=mammal&org=Human&
Google
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
Help

Add Custom Tracks

clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

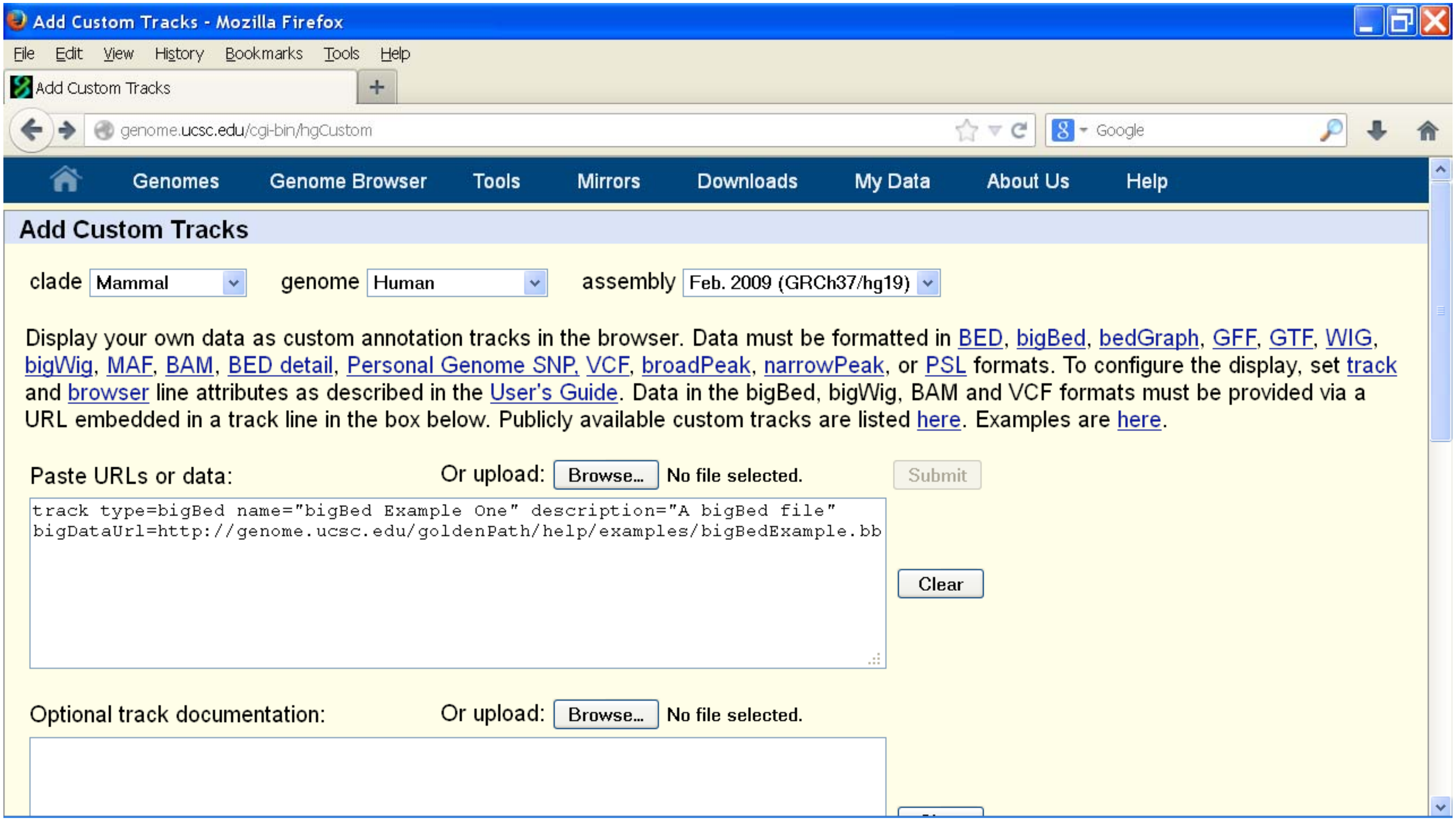
Display your own data as custom annotation tracks in the browser. Data must be formatted in BED, bigBed, bedGraph, GFF, GTF, WIG, bigWig, MAF, BAM, BED detail, Personal Genome SNP, VCF, broadPeak, narrowPeak, or PSL formats. To configure the display, set track and browser line attributes as described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

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

```
chr1 1000 10000
```

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

Visualization



Add Custom Tracks - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Add Custom Tracks
genome.ucsc.edu/cgi-bin/hgCustom
Google
Genomes Genome Browser Tools Mirrors Downloads My Data About Us Help

Add Custom Tracks

clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

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

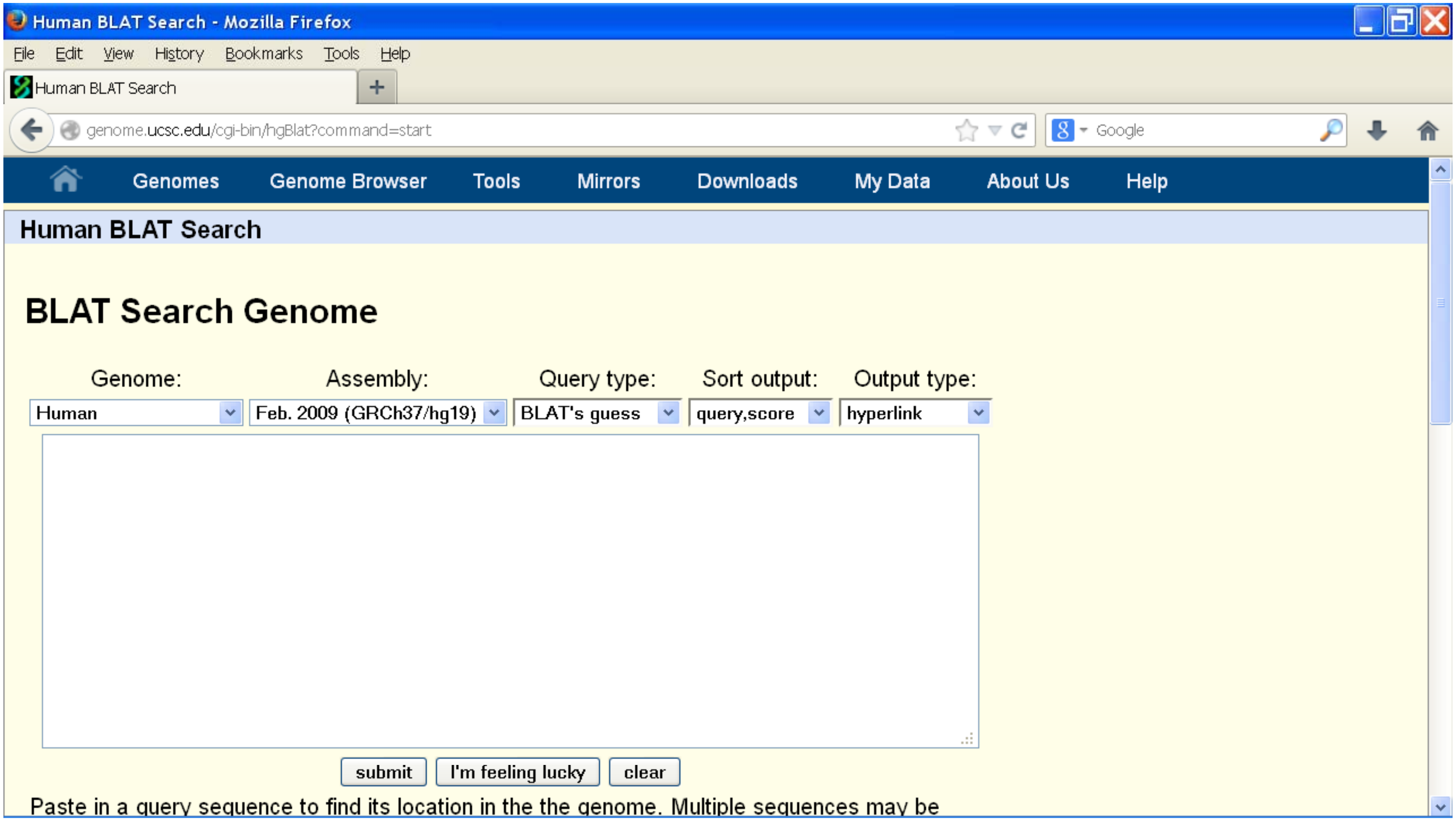
Paste URLs or data: Or upload: Browse... No file selected.

```
track type=bigBed name="bigBed Example One" description="A bigBed file"
bigDataUrl=http://genome.ucsc.edu/goldenPath/help/examples/bigBedExample.bb
```

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

Visualization

http://genome.ucsc.edu/index.html


Human BLAT Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Human BLAT Search
genome.ucsc.edu/cgi-bin/hgBlat?command=start
Google
Genomes Genome Browser Tools Mirrors Downloads My Data About Us Help
submit
I'm feeling lucky
clear

Human BLAT Search

BLAT Search Genome

Genome:	Assembly:	Query type:	Sort output:	Output type:
Human	Feb. 2009 (GRCh37/hg19)	BLAT's guess	query,score	hyperlink

Paste in a query sequence to find its location in the the genome. Multiple sequences may be

AGCGAATTGGAATGACCTAACATTTCTGTGACATCT

Visualization/Databases

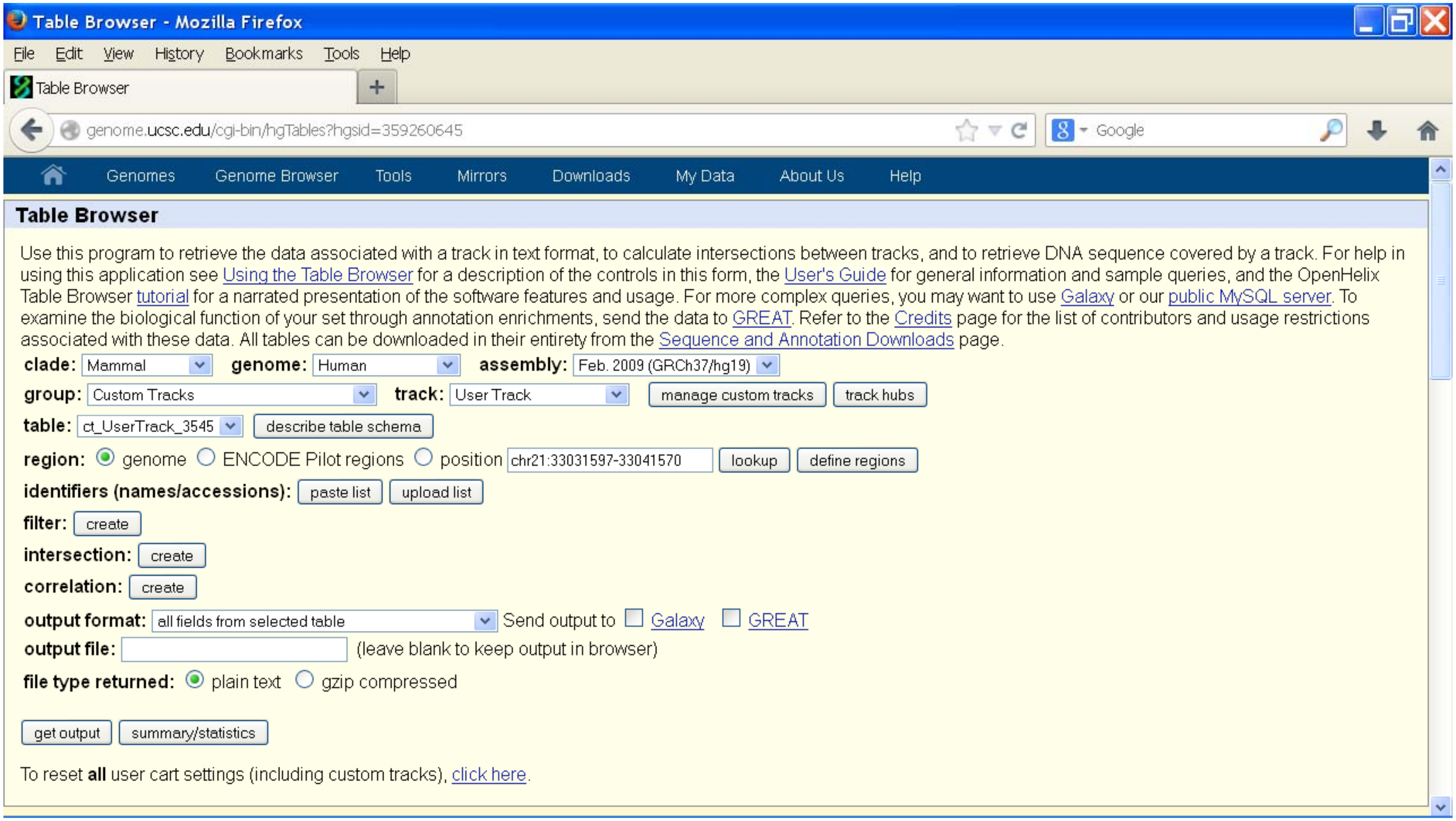


Table Browser - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Table Browser
genome.ucsc.edu/cgi-bin/hgTables?hgsid=359260645
Google
Genomes Genome Browser Tools Mirrors Downloads My Data About Us Help

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Custom Tracks **track:** User Track manage custom tracks track hubs

table: ct_UserTrack_3545 describe table schema

region: ● genome ○ ENCODE Pilot regions ○ position chr21:33031597-33041570 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: all fields from selected table Send output to □ Galaxy □ GREAT

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

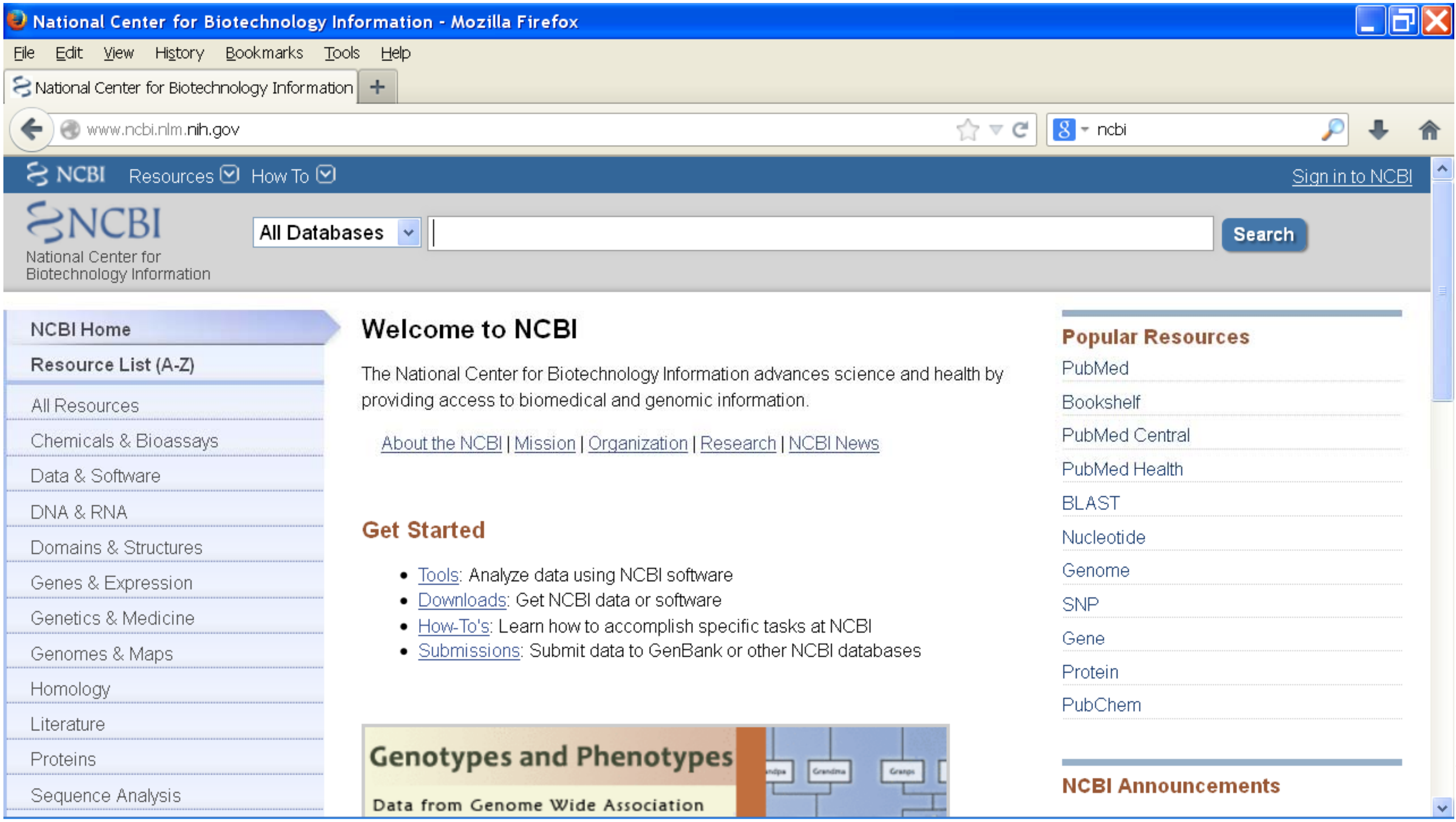
get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Try to download Human hg19 RefSeq genes with “fields selection”.

Databases

- NCBI (http://www.ncbi.nlm.nih.gov)


National Center for Biotechnology Information - Mozilla Firefox
File Edit View History Bookmarks Tools Help
National Center for Biotechnology Information
www.ncbi.nlm.nih.gov
ncbi
NCBI Resources How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | NCBI News

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Popular Resources

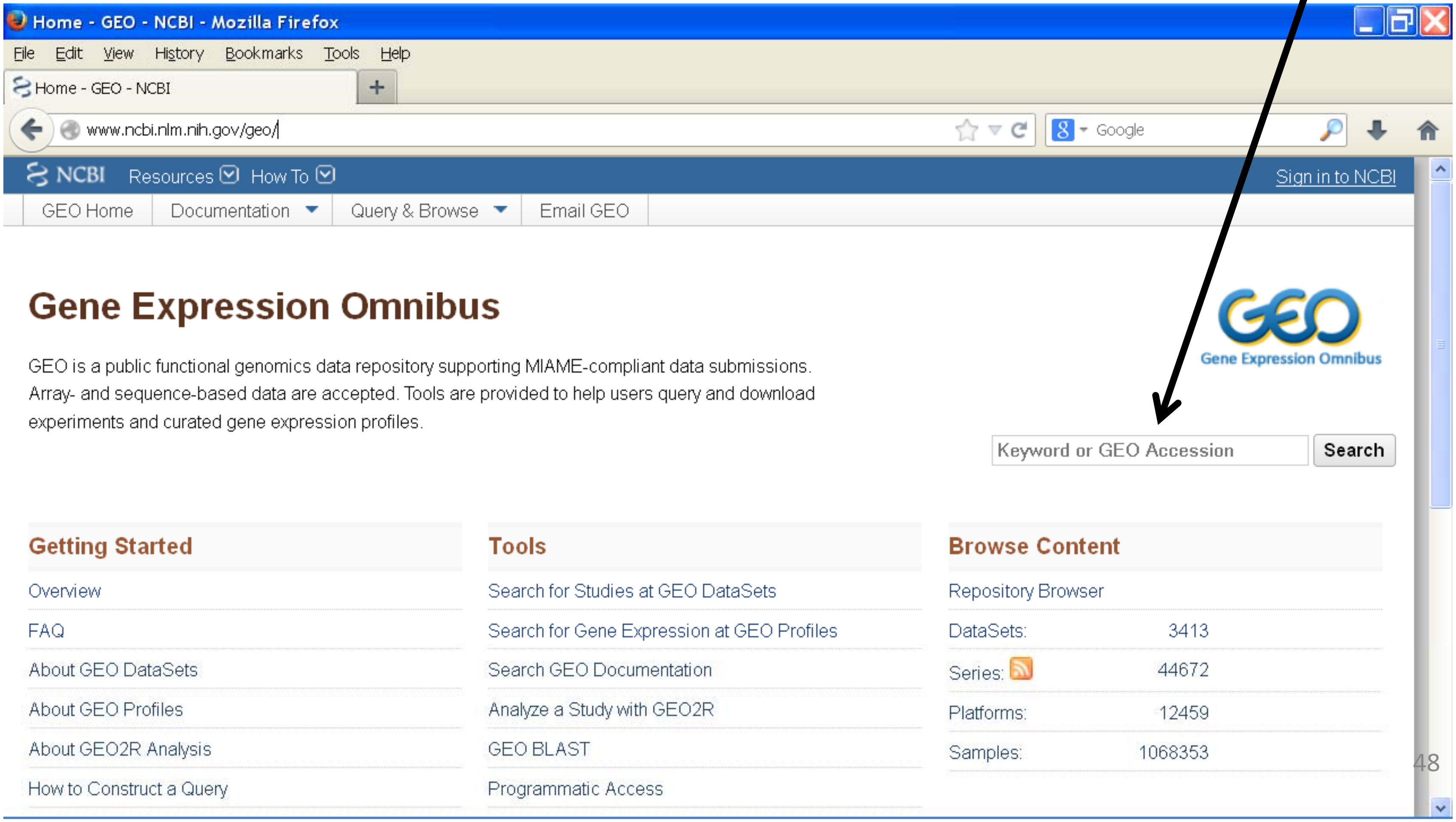
- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

Databases

- NCBI download

The Gene Expression Omnibus (GEO): GSE44690


Home - GEO - NCBI - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Home - GEO - NCBI
www.ncbi.nlm.nih.gov/geo/
Google
NCBI Resources How To
Sign in to NCBI
GEO Home Documentation Query & Browse Email GEO
GEO
Gene Expression Omnibus
Keyword or GEO Accession
Search

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access

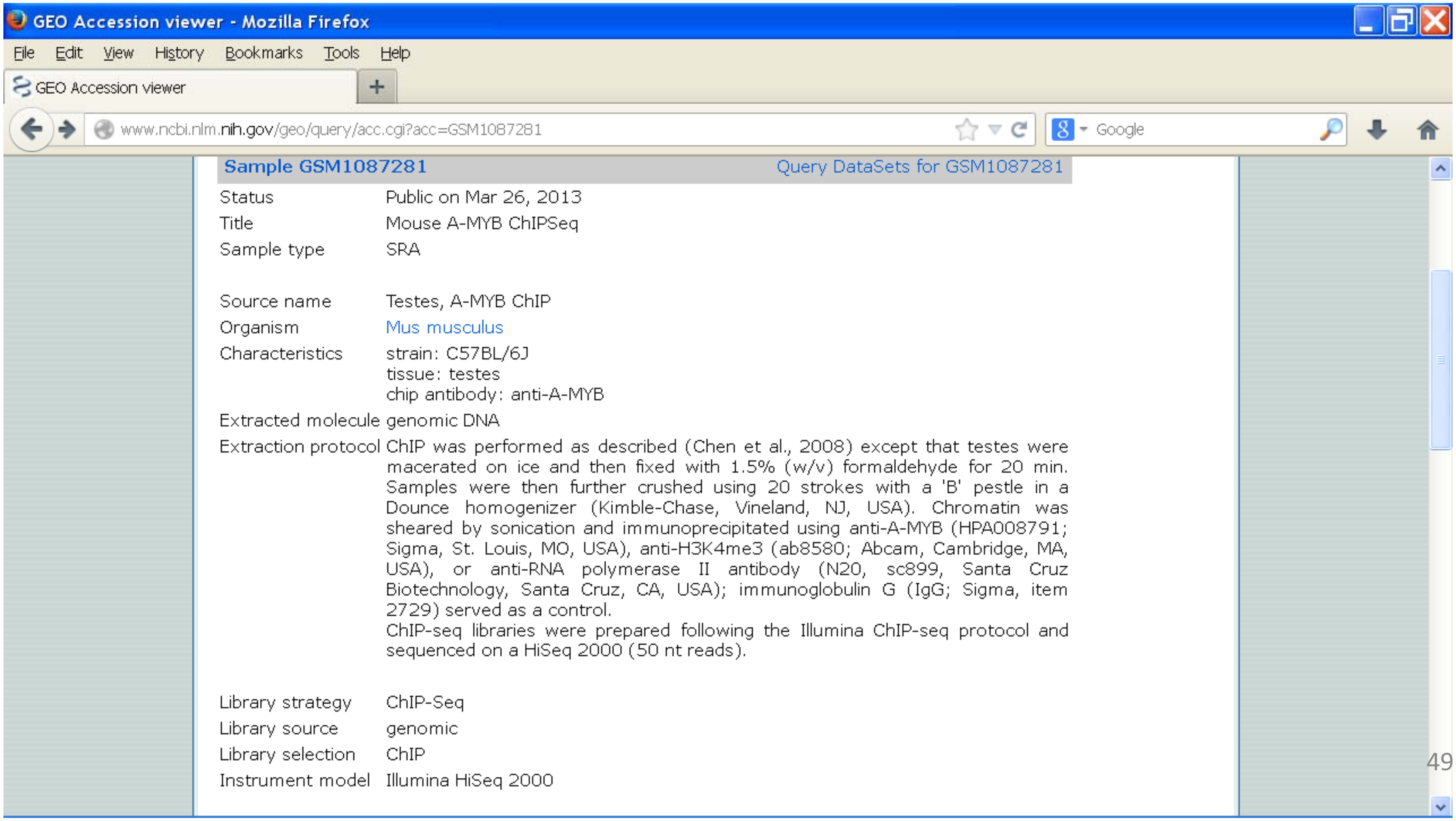
Browse Content

Repository Browser

DataSets:	3413
Series:	44672
Platforms:	12459
Samples:	1068353

Databases

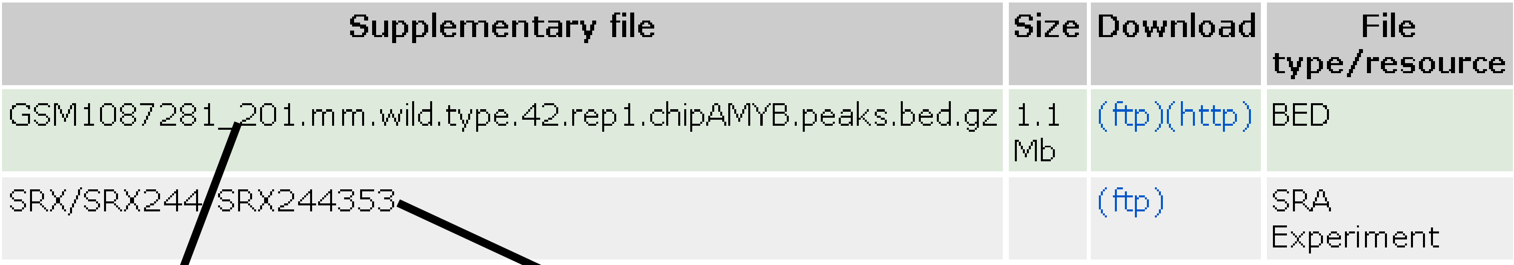
- NCBI download


GEO Accession viewer - Mozilla Firefox
File Edit View History Bookmarks Tools Help
GEO Accession viewer
+
www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1087281
Google

Sample GSM1087281	Query DataSets for GSM1087281
Status	Public on Mar 26, 2013
Title	Mouse A-MYB ChIPSeq
Sample type	SRA
Source name	Testes, A-MYB ChIP
Organism	Mus musculus
Characteristics	strain: C57BL/6J tissue: testes chip antibody: anti-A-MYB
Extracted molecule	genomic DNA
Extraction protocol	ChIP was performed as described (Chen et al., 2008) except that testes were macerated on ice and then fixed with 1.5% (w/v) formaldehyde for 20 min. Samples were then further crushed using 20 strokes with a 'B' pestle in a Dounce homogenizer (Kimble-Chase, Vineland, NJ, USA). Chromatin was sheared by sonication and immunoprecipitated using anti-A-MYB (HPA008791; Sigma, St. Louis, MO, USA), anti-H3K4me3 (ab8580; Abcam, Cambridge, MA, USA), or anti-RNA polymerase II antibody (N20, sc899, Santa Cruz Biotechnology, Santa Cruz, CA, USA); immunoglobulin G (IgG; Sigma, item 2729) served as a control. ChIP-seq libraries were prepared following the Illumina ChIP-seq protocol and sequenced on a HiSeq 2000 (50 nt reads).
Library strategy	ChIP-Seq
Library source	genomic
Library selection	ChIP
Instrument model	Illumina HiSeq 2000

Databases

- NCBI download

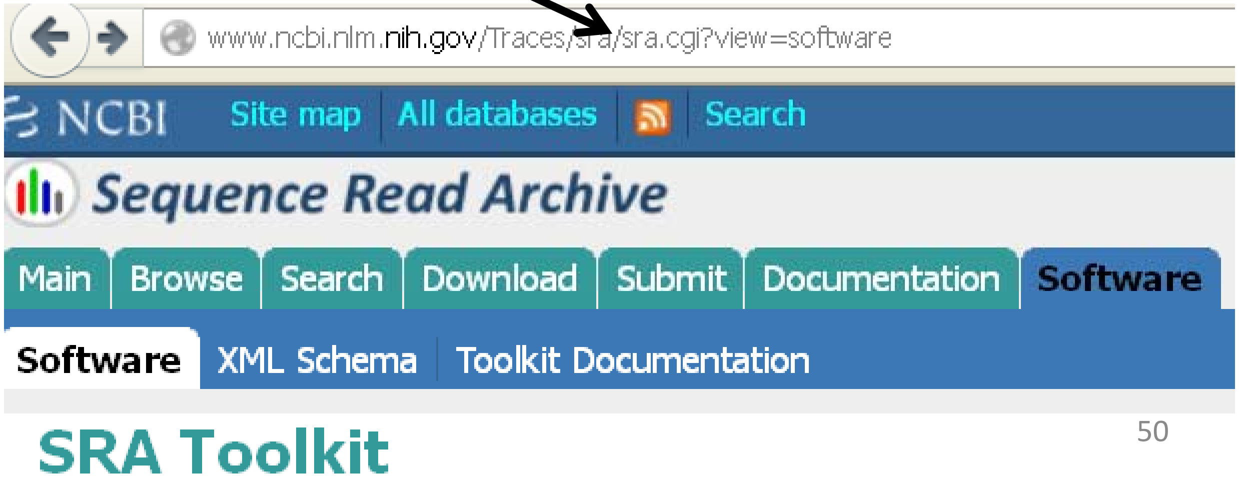


Supplementary file	Size	Download	File type/resource
GSM1087281_201.mm.wild.type.42.rep1.chipAMYB.peaks.bed.gz	1.1 Mb	(ftp)(http)	BED
SRX/SRX244/SRX244353		(ftp)	SRA Experiment

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

```
$ module load sratoolkit/2.3.4-2
$ illumina-dump data.sra
```

BED format is already described



Questions?

Homework

Use information from today’s seminar and get instructions from this location:

Server:	ghpcc06.umassrc.org
Folder/file:	/project/umw_biocore/seminar/Step2.docx