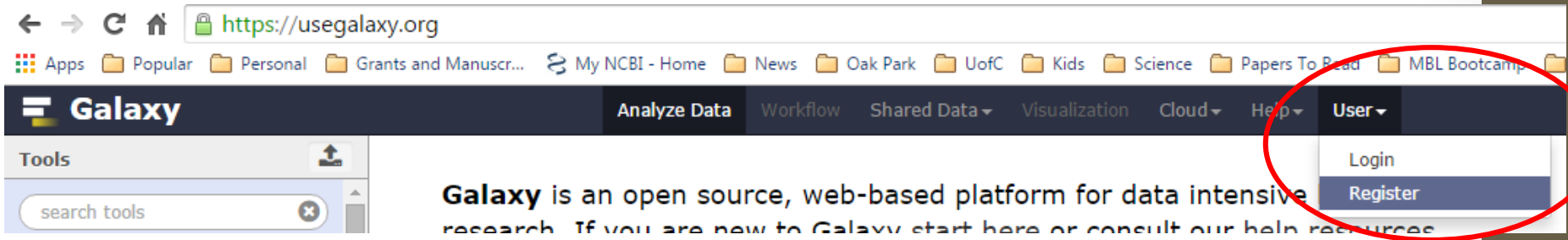


While Setting Up:

1. Go to <https://usegalaxy.org/> and set up account



2. Locate Vander Griend folder on USB drive...

Critical and Quantitative Analyses of Next Generation Sequencing Data

Donald Vander Griend, Ph.D.

Dept. of Surgery, Section of Urology (CCB, DSRB)

Alex Ling (CCB)

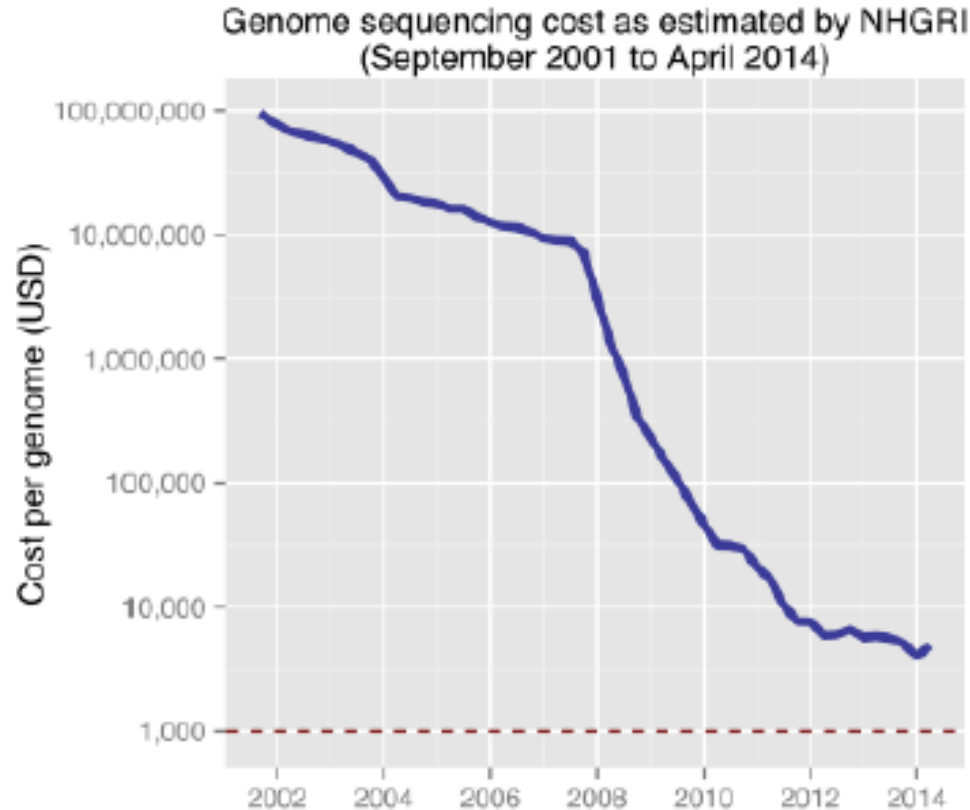
Overview

- **Part 1:** Analyses of FASTQ RNAseq Data
- **Part 2:** Data Visualization
- **Part 3:** Utilizing Online Databases

Next Generation Sequencing (NGS) Today...

- **Impressive advances** in NGS have enabled an immense diversity of **novel applications**.
- The barrier of the **\$1000 genome** has recently been broken.
- Important novel tools for **clinical diagnostics** based on NGS are appearing.
- **Third-generation** technologies may further revolutionize genomics research.
- Significant **challenges** for NGS remain, in particular data storage and processing.

The Price of Sequencing



The price of sequencing a single genome has dropped from the **\$3 billion** spent by the original Human Genome Project 13 years ago to as little as **\$1,000**

Sequencing at UofC

- <https://osrf.uchicago.edu/>
- <https://genomics.uchicago.edu/>

2011 High Throughput Sequencing Services Prices

HIGH THROUGHPUT SEQUENCING SERVICES	CHARGE
Illumina Libraries (standard)	Per library
DNA-SEQ	\$200
RNA-SEQ	\$250
small RNA SEQ	\$250
Illumina Sequencing Runs (HiSeq)	Per lane
50 bp Single-end	\$750 (any number of Lanes)
100 bp Single-end	\$1,100 (8 lanes only)
50 bp Paired-end	\$1,250 (8 lanes only)
100 bp Paired-end	\$1,800 (any number of lanes)
SOLiD Libraries (standard)	Per Library
DNA-SEQ	\$200
RNA-SEQ	\$250
SOLiD Sequencing Runs	Per lane
50 bp Single-end	\$675
75 bp Single-end	\$833
75 bp - 35 bp paired end	\$1,050

CONTACTS

[Pieter W. Faber, PhD](#)
Technical Director
Phone: (773)834-8420

[Yoav Gilad, PhD](#)
Scientific Director

LOCATION

Genomics Facility
University of Chicago -
Knapp Center for Biomedical
Discovery (KCBD)
900 E. 57th Street
Room # 1230C
Chicago, IL 60637

[900 E. 57 Street, Chicago IL
60637](#)

SCHEDULE

The Rate of Data Acquisition

Technologies

ABI
capillary

454
pyroseq

Solexa/
Illumina

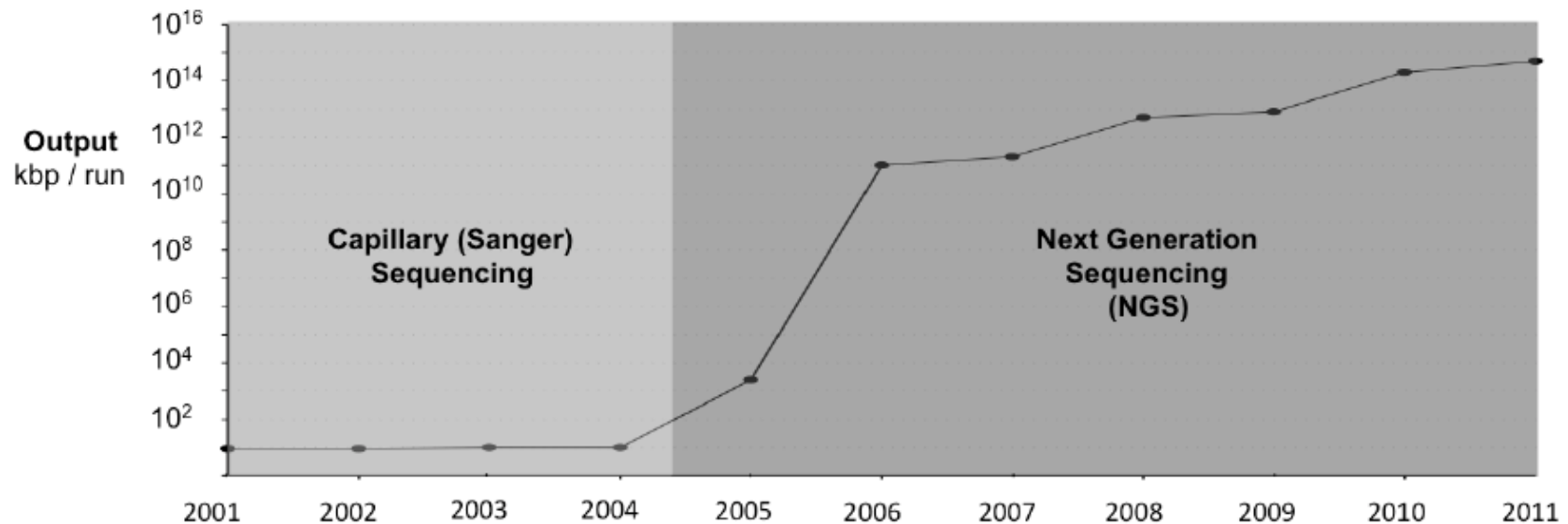
ABI
SOLID

Roche/454
Titanium

ABI
SOLID 3.0

Ion
Torrent

Illumina
HiSeq



Broad Applications of NGS

Broad applications of NGS to drug discovery

<i>Applications</i>	<i>Pros of NGS</i>	<i>Cons of NGS</i>	<i>Alternatives</i>	<i>Refs</i>
Mutation detection: personalised medicine	Can sequence large genome regions to identify efficacy markers	<u>Initial setup and running cost for NGS</u>	Large-scale Sanger sequencing technology	[64]
ChIP-Seq: target identification and/or validation and compound profiling for epigenetics	Enables study of <u>epigenetic targets at the whole-genome level</u>	Many possible algorithms for data analysis and complex data interpretation	ChIP-on-chip assay using microarray-based technology	[44]
CNV: target identification, personalised medicine, for example, cancer	Uncovers all types of CNV; no a priori assumptions about location of CNVs required	Large and complex rearrangements might not be detected	<u>Comparative genomic hybridisation</u>	[35,65,66]
Exome sequencing: target identification and/or drug resistance studies, biomarker discovery	<u>Identify rare variants, using deep sequence coverage</u>	Sequence variation in non-coding regions and introns not detected	Large-scale Sanger sequencing technology	[16]
RNA-Seq: target identification and/or validation by studying differential gene or miRNA expression between normal and diseased tissue	Detects alternative splicing and low expression transcripts; has large dynamic range	Bias during library preparation can result in over-representation of transcript 3' ends	<u>Microarray-based technology</u>	[11,12,67]

The application of next-generation sequencing technologies to drug discovery and development.

[Woollard PM](#), [Mehta NA](#), [Vamathevan JJ](#), [Van Horn S](#), [Bonde BK](#), [Dow DJ](#).

RNA-seq

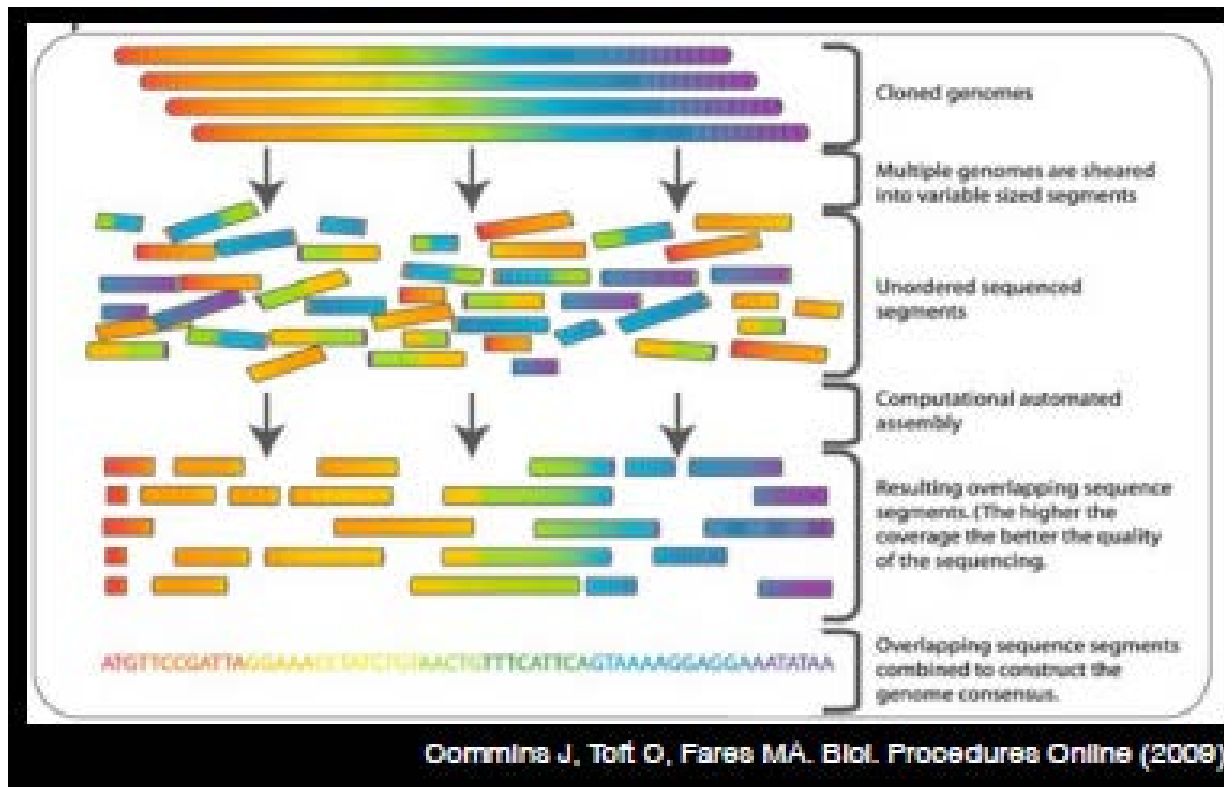
- NGS-based Technology to **qualitatively** and **quantitatively** profile the **full set of transcripts** (i.e., transcriptome), including **mRNAs**, **small RNAs** and other **non-coding RNAs**.
- **Transcriptome profiling** provides a snapshot of gene expression patterns and regulatory elements
- Although a **transcriptome** only represents a small fraction of the human genome (<5%), it is **very complex**, transcripts derived from **alternative splicing**, **gene fusion**, **antisense transcription** and **RNA editing** largely increase the **diversity** of the transcriptome

The Basic NGS Process

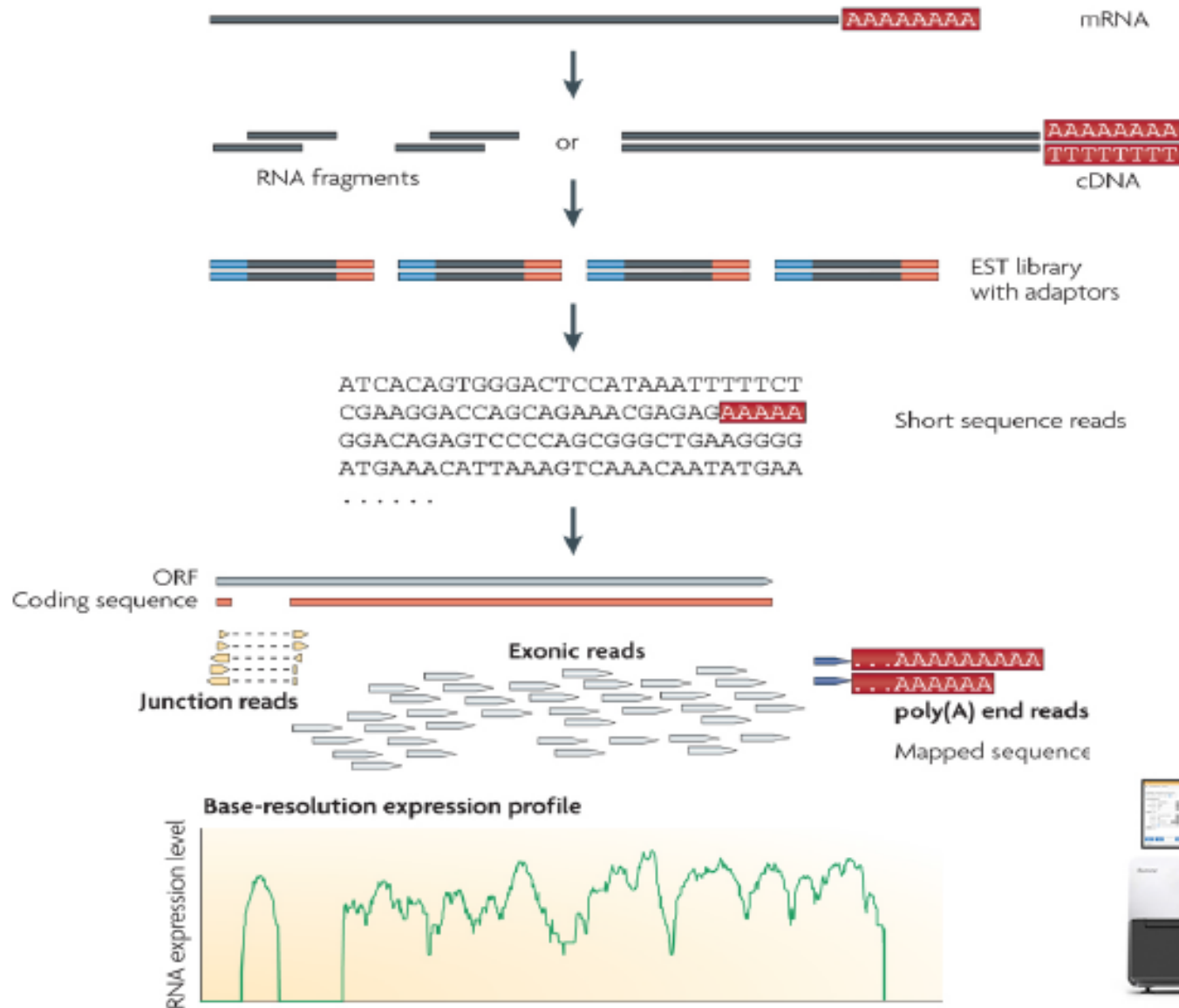
Sample preparation: the target genome is broken into fragments

Physical sequencing: individual bases in each fragment are identified in order

Reconstruction: bioinformatics software aligns overlapping reads from each fragment, allowing the original genome to be constructed



A Typical RNAseq Experiment



Let's Get Started...

<https://usegalaxy.org>

Menu of Analyses

User Interface

Your Analysis History

The screenshot displays the Galaxy web-based platform interface. The top navigation bar includes links for **Analyze Data**, **Workflow**, **Shared Data**, **Visualization**, **Cloud**, **Help**, and **User**. A status bar on the right indicates "Using 50%".

Menu of Analyses: A red arrow points to the left sidebar, which contains a search bar and a list of tool categories: **Get Data**, **Send Data**, **Lift-Over**, **Text Manipulation**, **Convert Formats**, **Filter and Sort**, **Join, Subtract and Group**, **NGS: QC and manipulation**, **NGS: Mapping**, **NGS: RNA-seq**, **NGS: SAMtools**, **NGS: BAM Tools**, **NGS: Picard**, **NGS: VCF Manipulation**, **Extract Features**, **Fetch Sequences**, **Fetch Alignments**, **Get Genomic Scores**, **Operate on Genomic Intervals**, **Statistics**, **Graph/Display Data**, **Phenotype Association**, **snpEff**, and **BEDTools**.


User Interface: A red arrow points to the main content area. It features a header stating: "Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#)." Below this is a large banner for "GXYcast 1 2015". To the right of the banner is a "Tweets" section displaying three tweets from the Galaxy Project (@galaxyproject), including announcements about the 2016 Galaxy Community Conference and a meetup.


Your Analysis History: A red arrow points to the right sidebar, titled "History". It shows a search bar and a list of datasets under the heading "imported: RNA-seq exercise datasets". The list includes five datasets: "5: brain_2.fastq", "4: brain_1.fastq", "3: adrenal_2.fastq", "2: adrenal_1.fastq", and "1: iGenomes UCSC hg19, chr19 gene annotation". Each dataset entry has icons for viewing, editing, and deleting.

The footer of the page features logos for Penn State, TACC, and iPlant Collaborative.

Upload Data – 5 Files

Galaxy

Tools 


search tools 

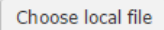
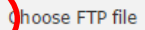





Get Data






- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- MouseMine server
- Ratmine server

Download data directly from web or upload files from your disk

You can Drag & Drop files into this box.




















Type (set all):  Genome (set all):



      

Name	Date modified	Type	Size
 Galaxy1-[_iGenomes_UCSC_hg19_chr19_...	8/4/2015 11:42 AM	GTF File	6,021 KB
 Galaxy2-[adrenal_1.fastq].fastqsanger	8/4/2015 11:41 AM	FASTQSANGER File	8,011 KB
 Galaxy3-[adrenal_2.fastq].fastqsanger	8/4/2015 11:42 AM	FASTQSANGER File	8,011 KB
 Galaxy4-[brain_1.fastq].fastqsanger	8/4/2015 11:41 AM	FASTQSANGER File	6,073 KB
 Galaxy5-[brain_2.fastq].fastqsanger	8/4/2015 11:41 AM	FASTQSANGER File	6,073 KB

Download data directly from web or upload files from your disk

Please wait...2 out of 5 remaining.

Name	Size	Type	Genome	Settings	Status
 Galaxy1- [_iGenomes_UCSC_hg 19_chr19_gene_anno tation].gtf	6.2 MB	Auto-det...  	----- Additional Sp... 		100% 
 Galaxy2- [adrenal_1.fastq].fast qsanger	8.2 MB	Auto-det...  	----- Additional Sp... 		100% 
 Galaxy3- [adrenal_2.fastq].fast qsanger	8.2 MB	Auto-det...  	----- Additional Sp... 		100% 
 Galaxy4- [brain_1.fastq].fastqsa nger	6.2 MB	Auto-det...  	----- Additional Sp... 		 36% 

Type (set all): Auto-detect   Genome (set all): ----- Additional Species ... 

Choose local file

Choose FTP file

Paste/Fetch data

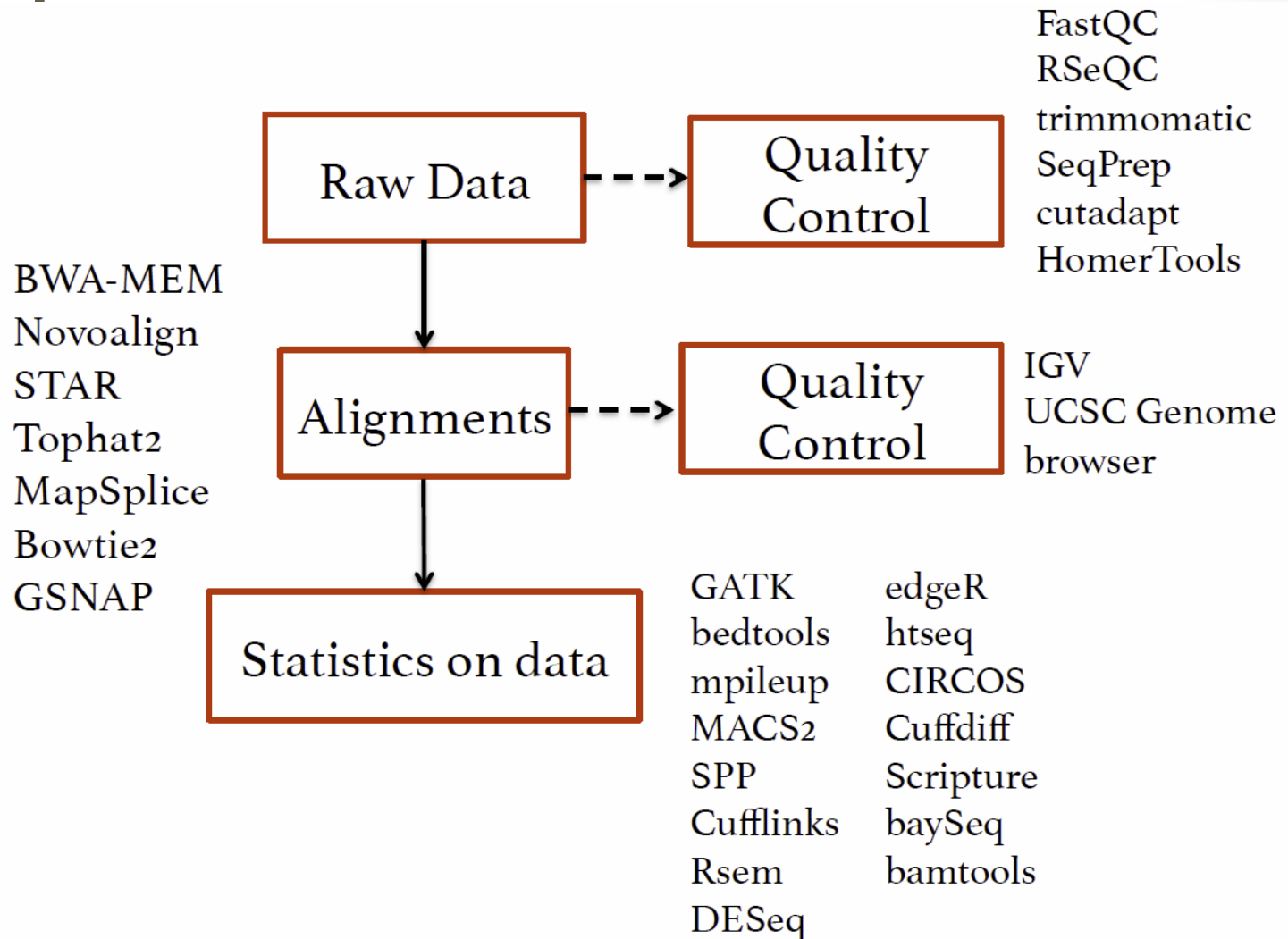
Start

Pause

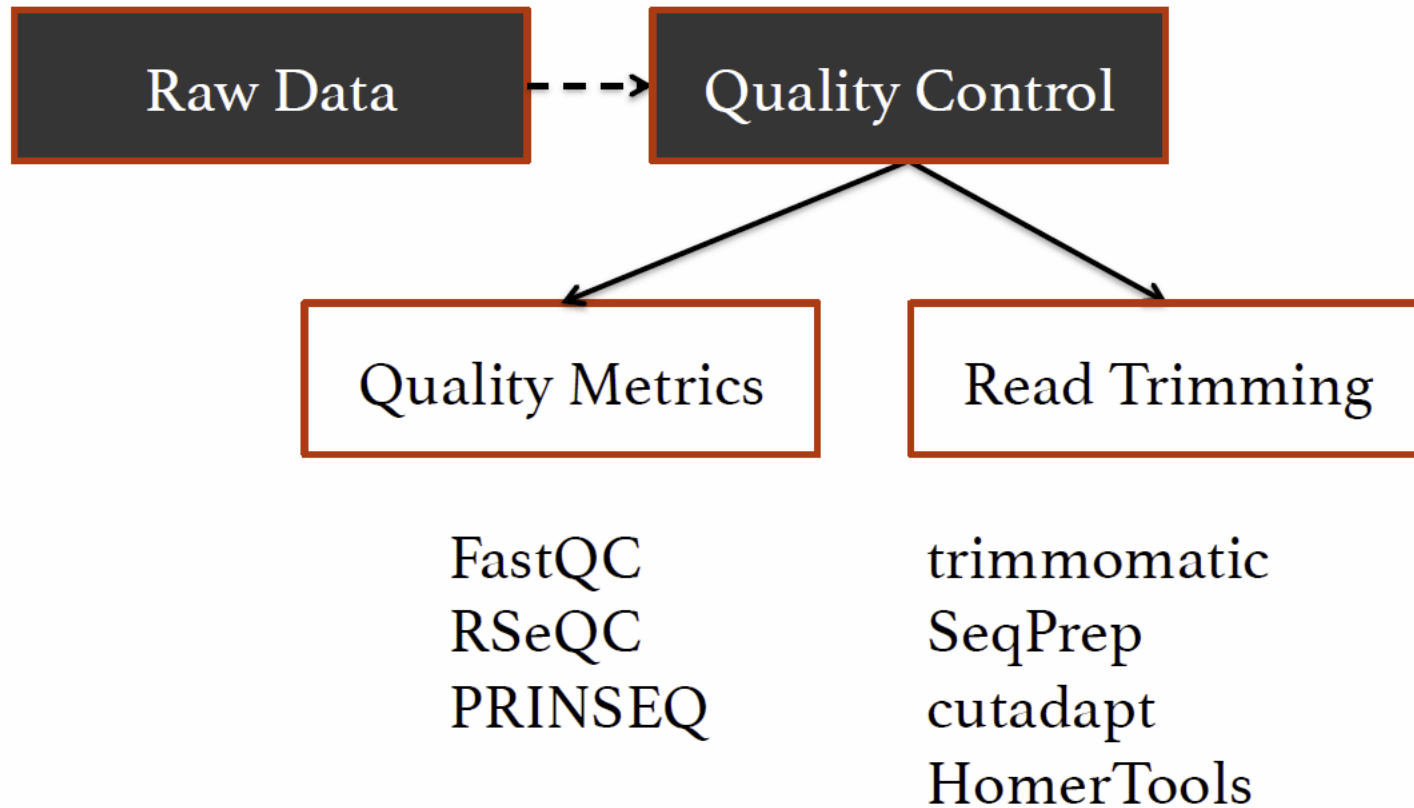
Reset

Close

Step 1: Data QC



QC and Trimming off Barcodes



History

search datasets

MBL Bootcamp, DVG Workshop
5 shown, 15 [deleted](#)

33.4 MB

5: brain_2.fastq 5.9 Mb
format: fastqsanger, database: ?

uploaded fastq file

View data

4: brain_1.fastq

3: adrenal_2.fastq

2: adrenal_1.fastq

1: iGenomes UCSC hg19, chr19 gene annotation

FastQC – Quality Metrics

Galaxy

Tools

search tools

[Get Data](#)

[Send Data](#)

[Lift-Over](#)

[Text Manipulation](#)

[Convert Formats](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[NGS: QC and manipulation](#)

[FastQC Read Quality reports](#)

[Select high quality segments](#)

[Build base quality distribution](#)

[Draw quality score boxplot](#)

[Quality format converter \(ASCII-Numeric\)](#)

[Filter by quality](#)

[FASTQ to FASTA converter](#)

[Remove sequencing artifacts](#)

[Barcode Splitter](#)

[Clip adapter sequences](#)

[Collapse sequences](#)

FastQC Read Quality reports (Galaxy Tool Version 0.63)

Short read data from your current history

5: brain_2.fastq

Contaminant

No selection

tab delimited file with 2 columns: name and sequence. For example: IL

Submodule and Limit specifying file


No selection

a file that specifies which submodules are to be executed (default=all)
warning parameter

✓ Execute

Analyze Data Workflow S

brain_1.fastq FastQC Report

 FastQC Report
Mon 10 Aug 2015
brain_1.fastq

Summary

- ✓ [Basic Statistics](#)
- ✗ [Per base sequence quality](#)
- ✗ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✗ [Per base sequence content](#)
- ✗ [Per sequence GC content](#)
- ✗ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ! [Sequence Duplication Levels](#)
- ! [Overrepresented sequences](#)
- ✓ [Adapter Content](#)
- ! [Kmer Content](#)

30: FastQC on data
4: RawData

29: FastQC on data
4: Webpage

336.4 KB
format: **html**, database: ?

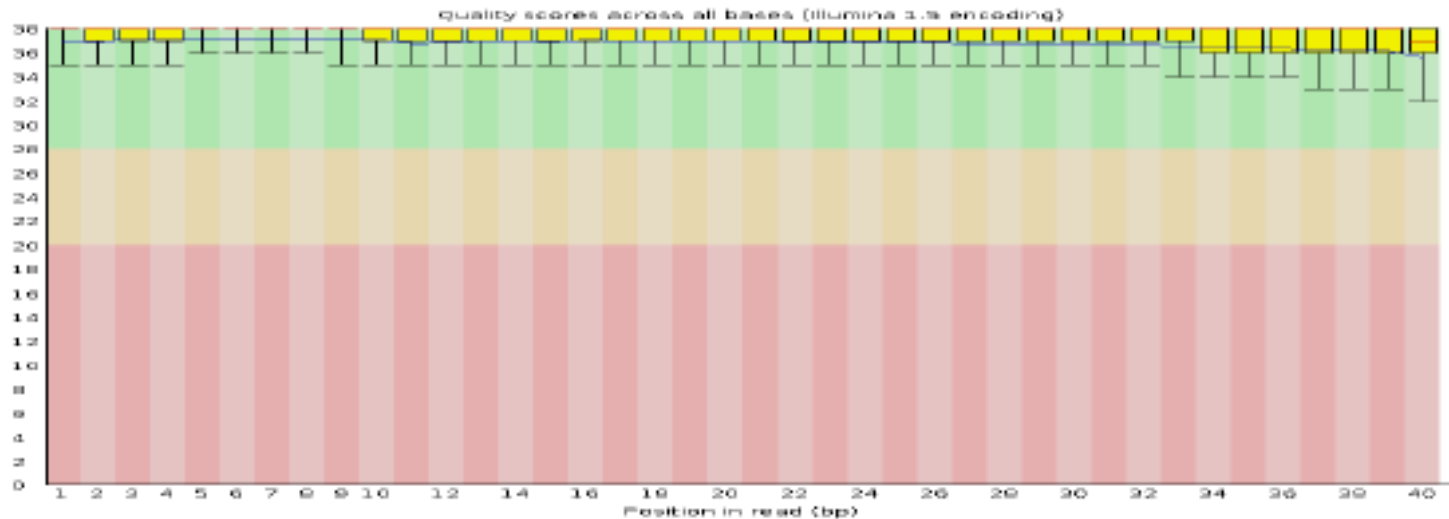
Picked up _JAVA_OPTIONS: -
Djava.io.tmpdir=/galaxy-
repl/main/scratch

HTML file

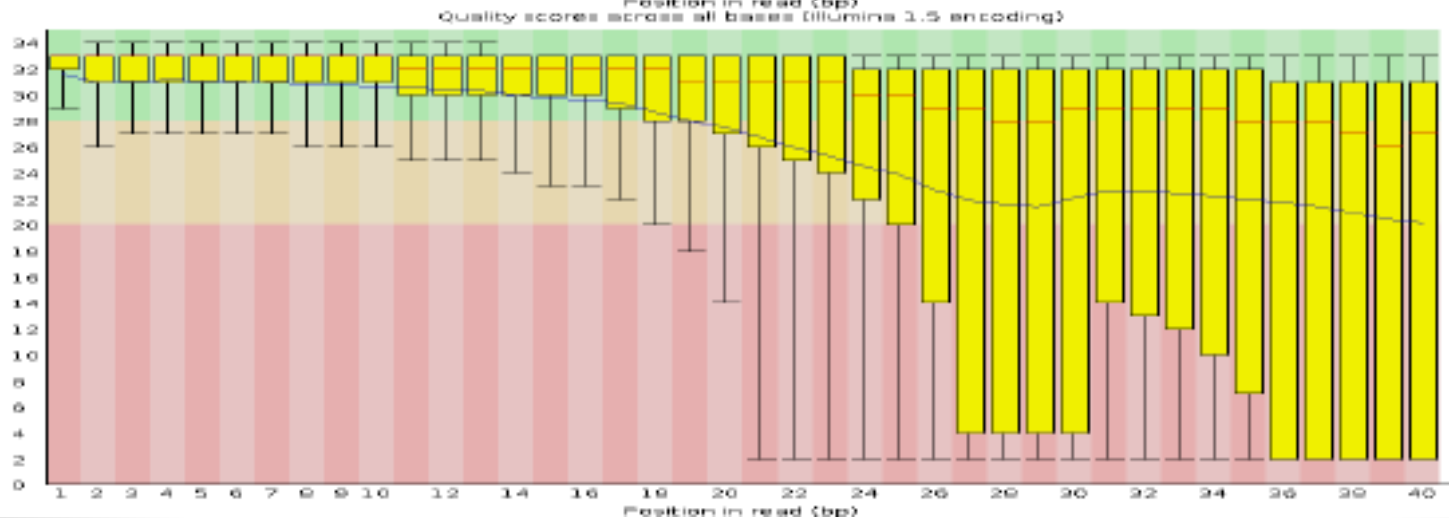
5: brain_2.fastq

- * Note Data X Reference
- * Delete "RawData" Files

Per base sequence quality



Good!

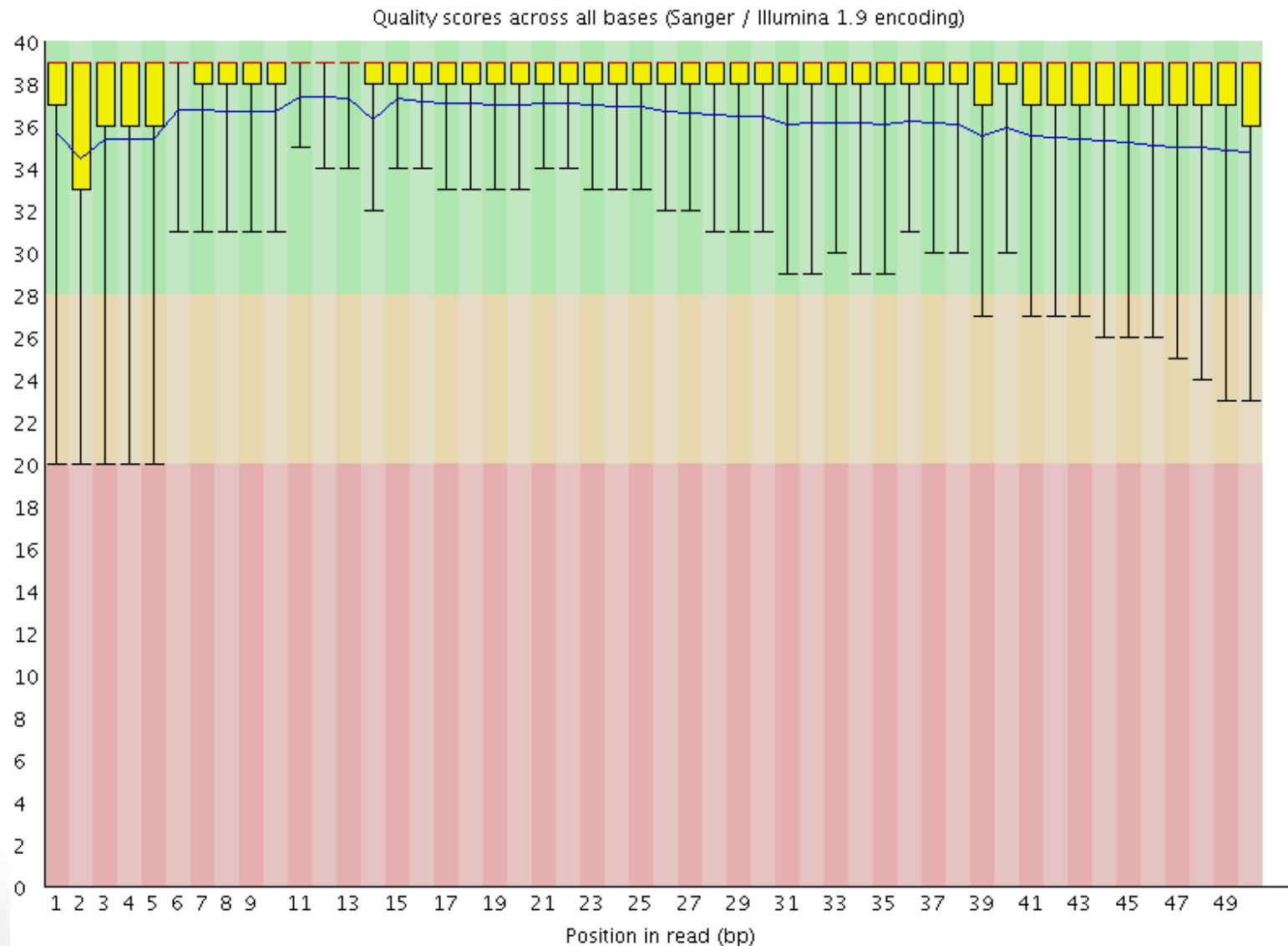


Bad!
Trimming
Needed

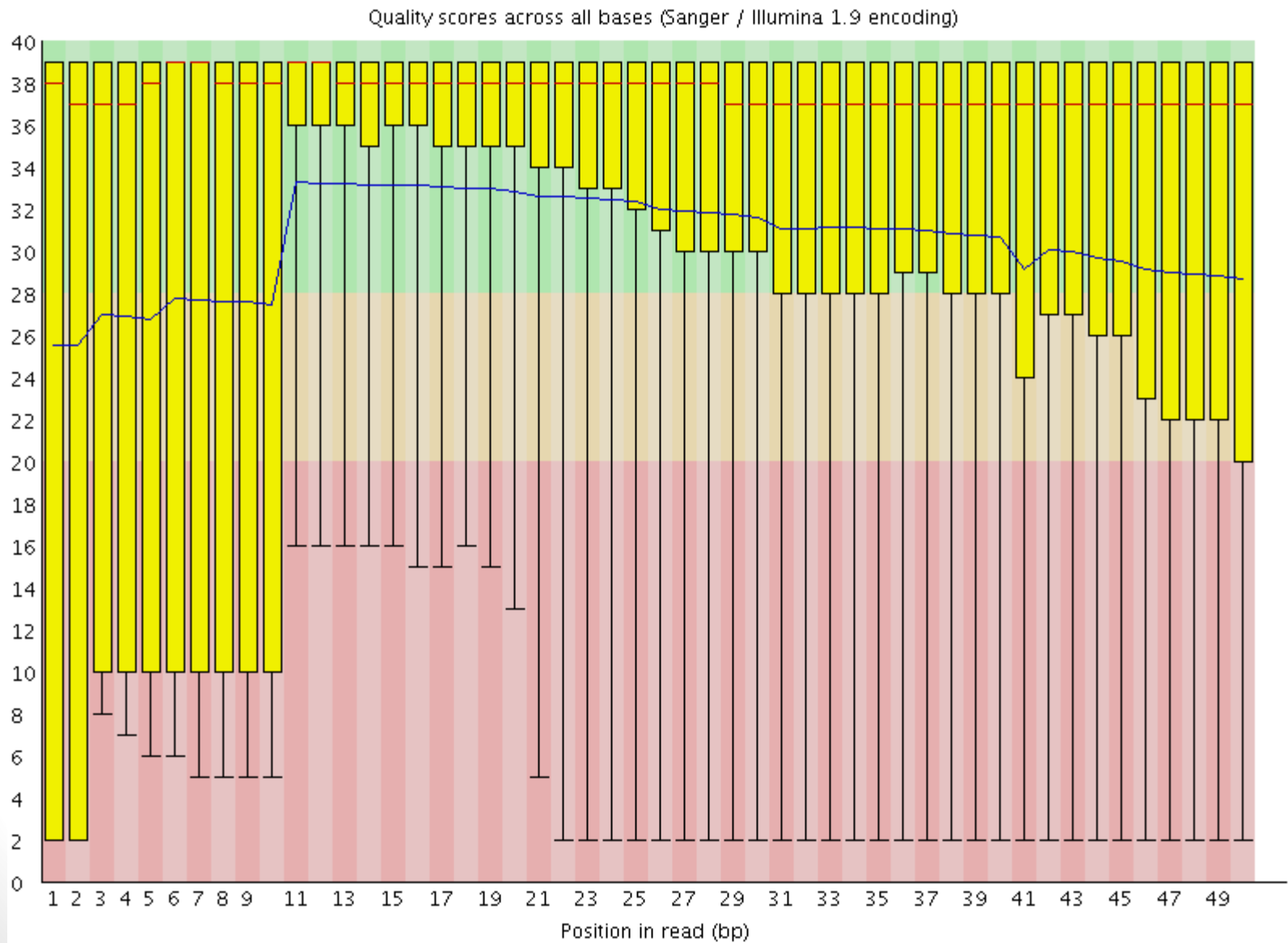
FastQC Analyses

- Assume a median quality score of below 20 to be unusable (yellow bars).
- Given this criterion, is trimming needed for any of the datasets? If so, which base pairs should be trimmed?
- RNAseq datasets do not pass per sequence GC content and sequence duplication levels. This is expected because FastQC is designed for DNA sequencing data, not RNA sequencing.
- What unique characteristics of RNAseq data will cause these two QC checks to fail?

Adrenal 1

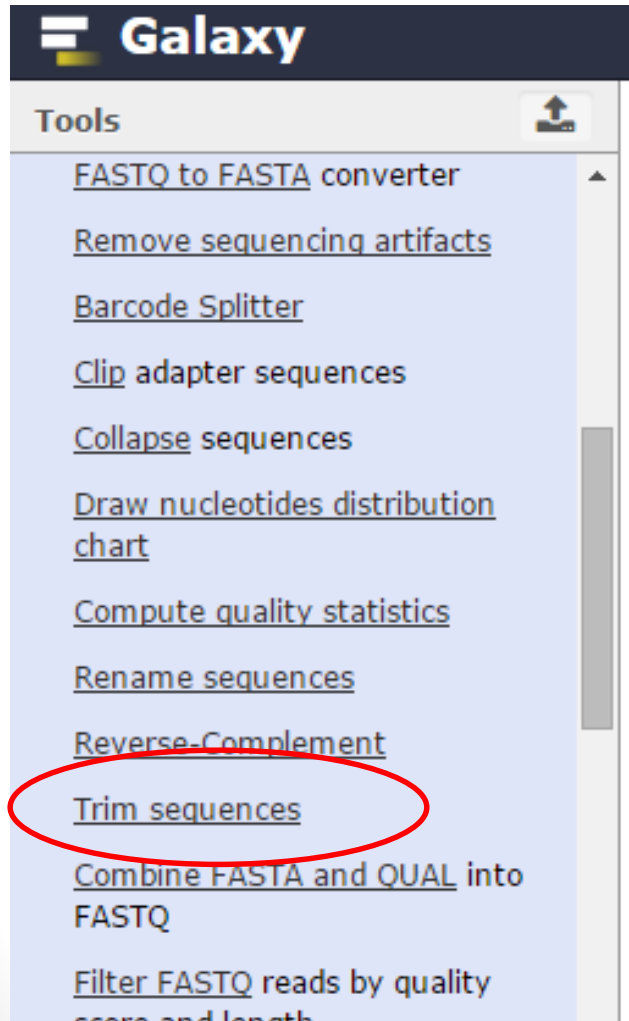


Brain 1



Trim Sequences

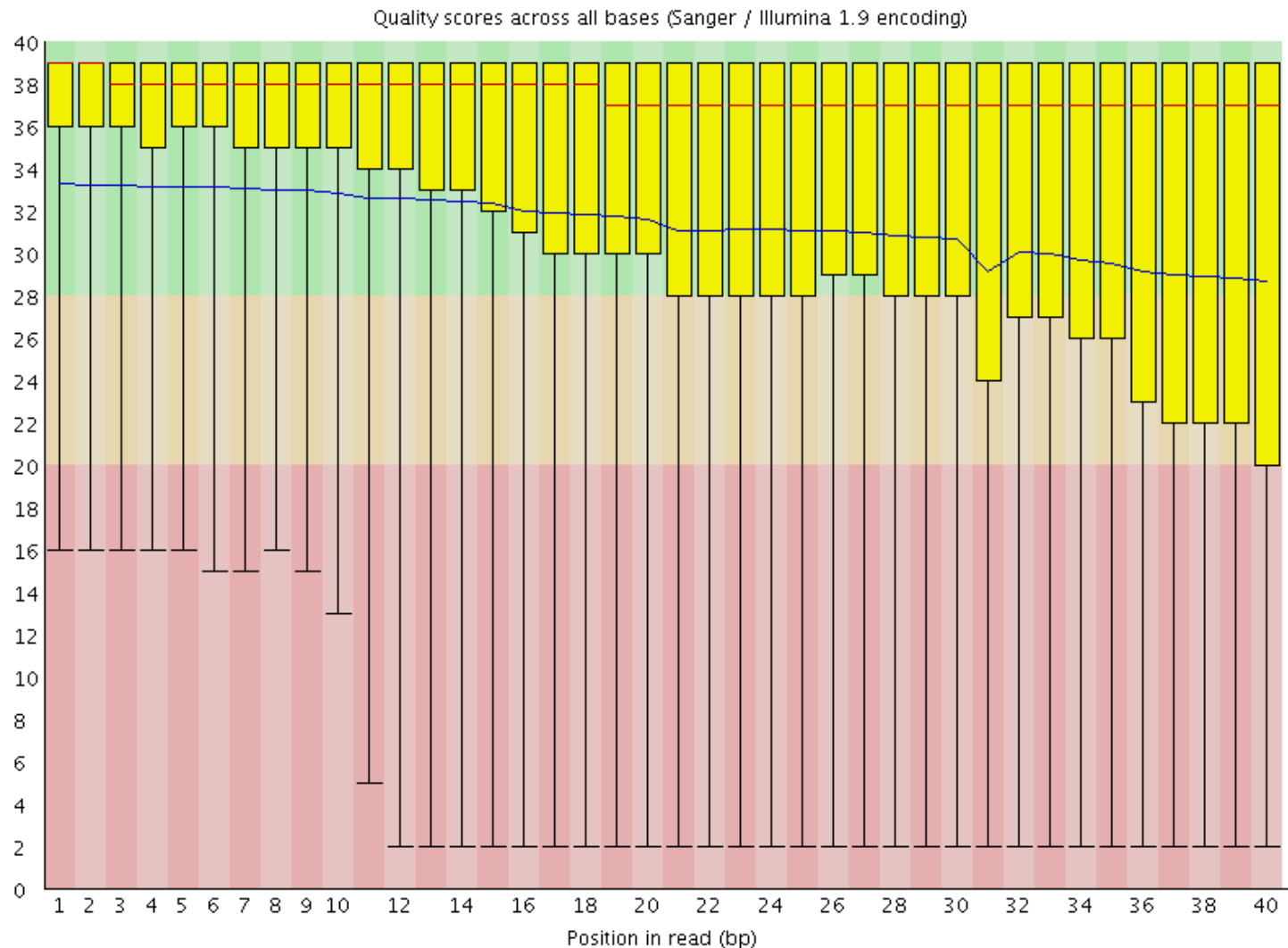
Brain 2



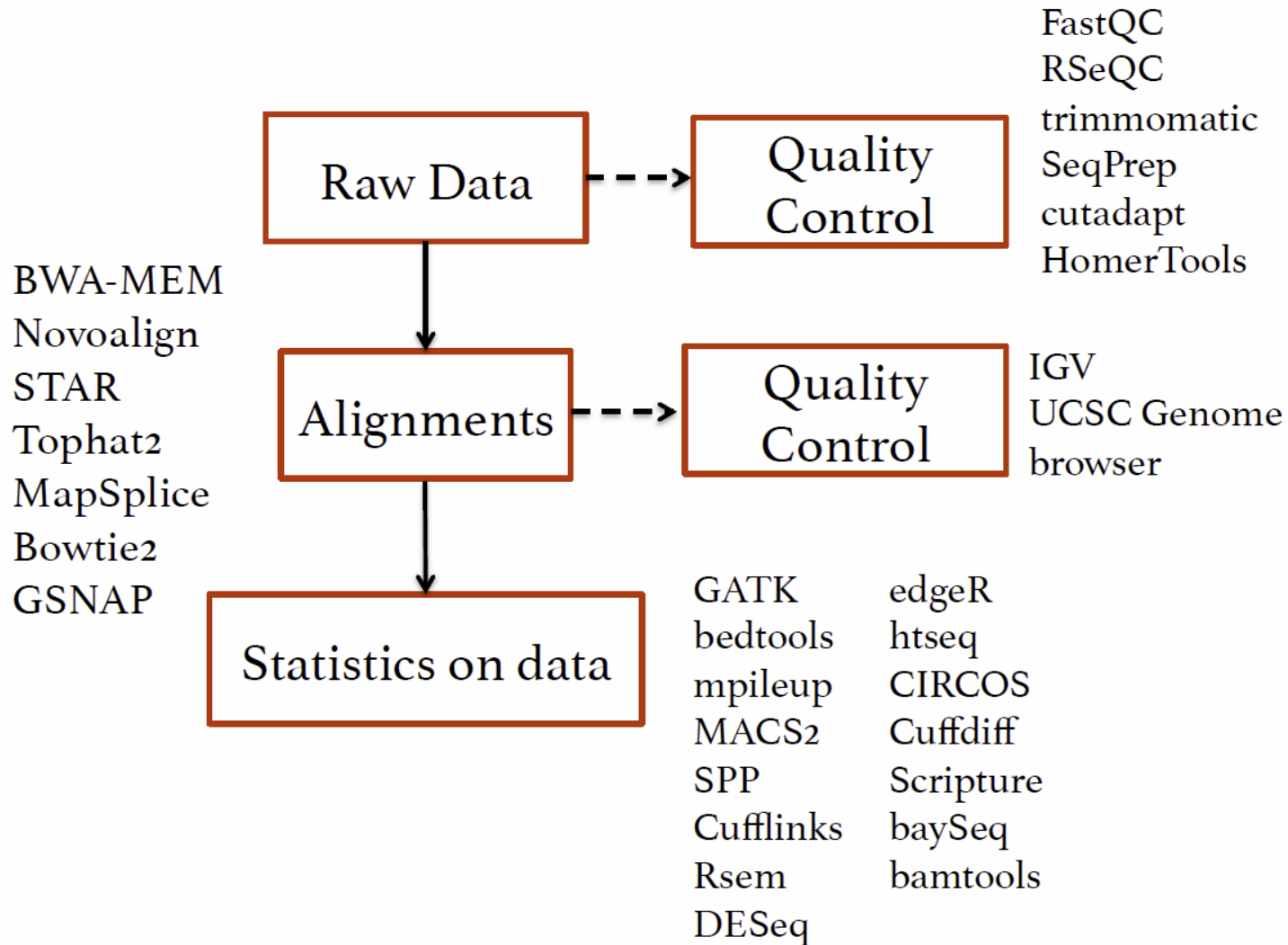
Now run FastQC again on the Trimmed Sequence and see what it looks like...

How does “Per base sequence quality” graph look now?

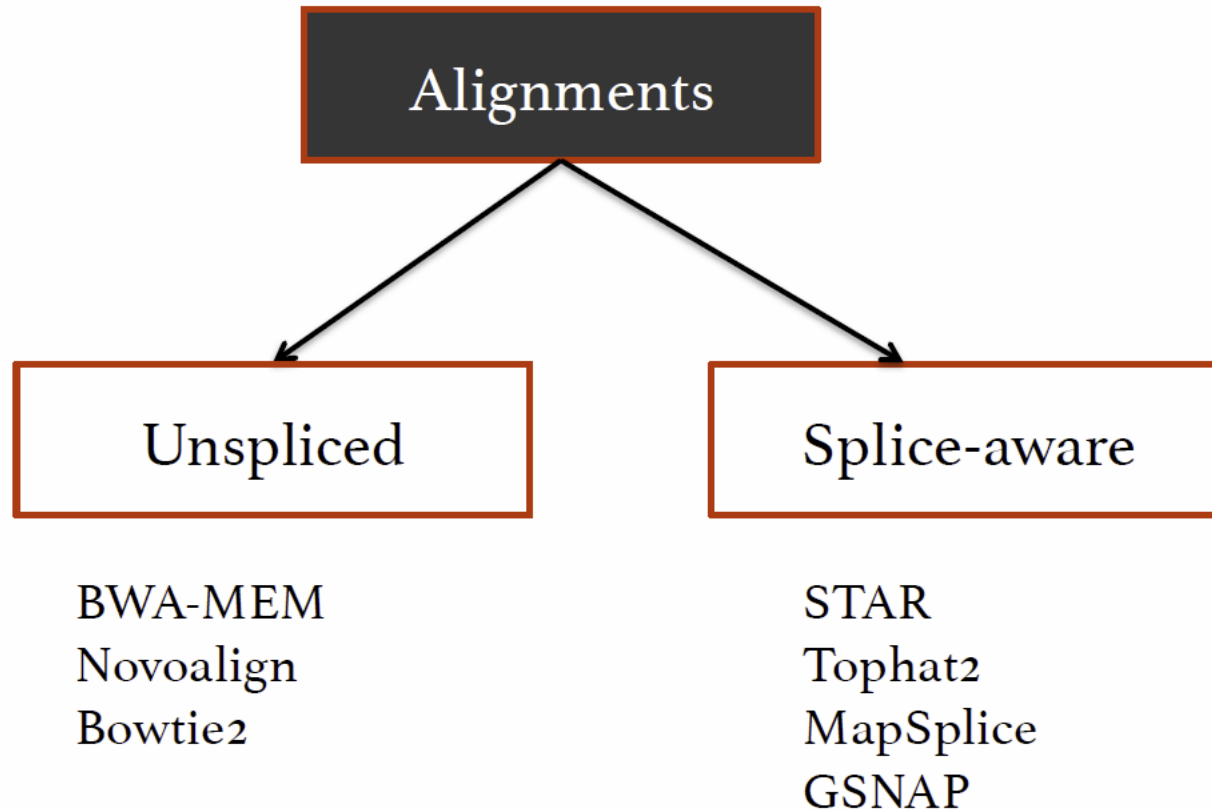
QC of Brain 1 Trimmed



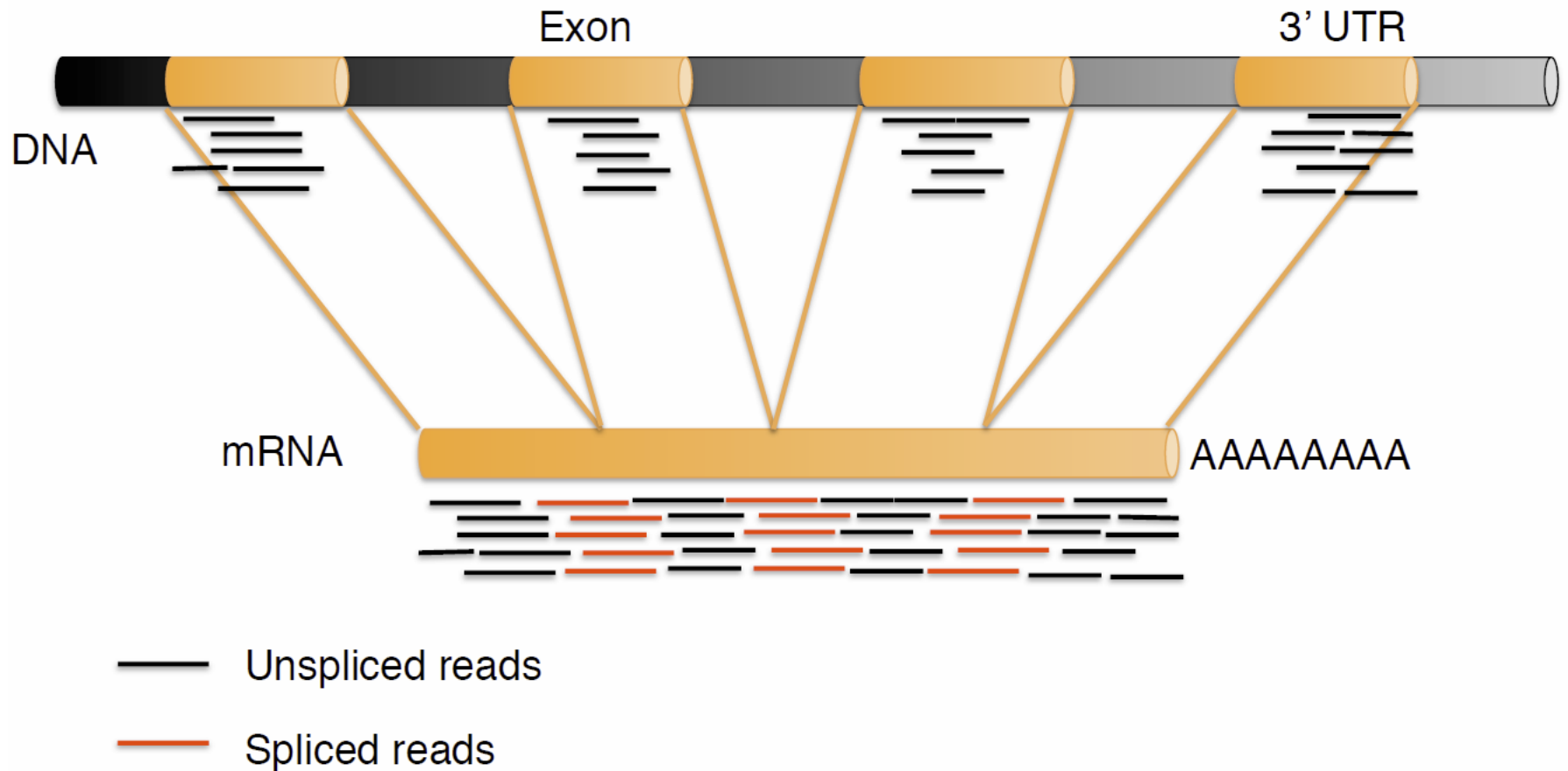
Step 2: Aligning to Reference Genome



Mapping: Aligning Reads to Genes



Aligning Reads to a Reference Genome (Hg19)



NGS: RNA-seq - TopHat

NGS: RNA-seq

Cuffdiff find significant changes in transcript expression, splicing, and promoter use

Tophat Gapped-read mapper for RNA-seq data

StringTie transcript assembly and quantification

- Paired vs. UnPaired Seq Runs
 - Tophat tool to map RNAseq reads to the hg19 build.
 - Because the reads are paired, you'll need to set mean inner distance between pairs; this is the average distance in basepairs between reads.
 - Use a mean inner distance of 110 for BodyMap data.
-
- Is this single-end or paired-end data? Paired end, individual datasets
 - Mean Inner Distance between Mate Pairs: 110
 - Select a reference genome: Human (Homo sapiens) (b37):hg19
 - Leave all defaults same...
 - Execute
 - Do again on your own for other organ (Adrenal/Brain)

(1) Transcriptome alignment (optional)

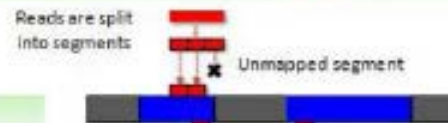


(2) Genome alignment



(3) Spliced alignment

(3-1) Segment alignment to genome



(3-2) Identification of splice sites (including indels and fusion break points)



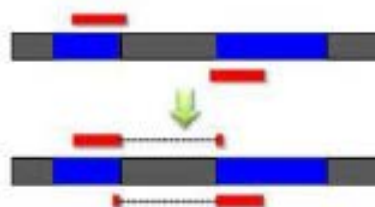
(3-3) Segments aligned to junction flanking sequences



(3-4) Segment alignments stitched together to form whole read alignments



(3-5) Re-alignment of reads minimally overlapping introns



Reads

Reads are aligned against transcriptome.

Transcriptome index

Reads are aligned against genome.

Genome index

Reads are split into smaller segments which are then aligned to the genome.

Genome index

Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.

Sequences flanking a splice site are concatenated and segments are aligned to them.

Junction flanking index

Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.

Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.

TopHat Output: BAM Format

SAM/BAM Format




1	2	3	4	5	6	7	8	9	10	11	12
R001	83	ref	37	30	9M	=	7	-39	CAGCGCAT	CAGCGCAT	TAG

COLUMNS:

1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	Reference sequence NAME
4	POS	Int	1-based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next fragment
8	PNEXT	Int	Position of the mate/next fragment
9	TLEN	Int	observed Template LENgth
10	SEQ	String	fragment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33≈

Renaming Files

Keeping Organized

52: Tophat on data 5 and data 45: accepted hits   

Accepted Hits







2.4 MB

format: **bam**, database: **hg19**

Log: tool progress
Log: tool progress




[2015-08-10 14:31:30]
Beginning TopHat run (v2.0.14)

[2015-08-10 14:31:30] Checking for Bowtie
Bowtie version: 2.2.5.0
[2015-08-10 14:31:30] Checking for Bowtie i

display at UCSC [main](#)
display at Ensembl [Current](#)
display with IGV [local](#) [Human hg19](#)
display in IGB [View](#)

Binary bam alignments file

52: Tophat on data 5 and data 45: accepted hits    **Edit attributes**

2.4 MB

format: **bam**, database: **hg19**

[Attributes](#) [Convert Format](#) [Datatype](#) [Permissions](#)

Edit Attributes

Name:

Tophat on data 5 and data 45: accepted hits

Info:

Log: tool progress
Log: tool progress

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available

Database/Build:

Human Feb. 2009 (GRCh37/hg19) (hg19)

Save

Auto-detect

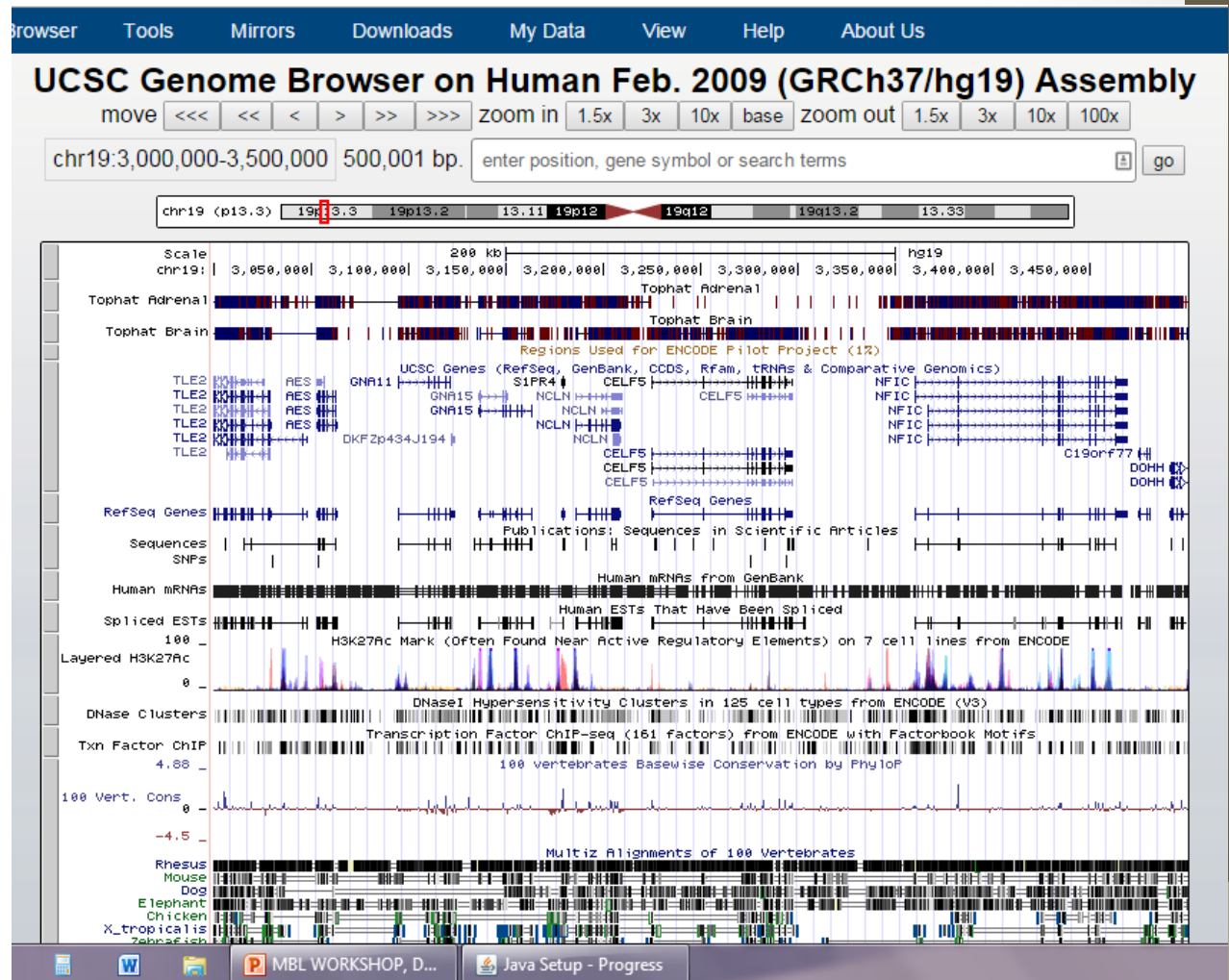
This will inspect the dataset and attempt to correct the above

Tophat: Brain
Tophat: Adrenal

World's Most Accurate Pie Chart



Visualizing Data



Search: chr19:3,000,000-3,500,000

History

3 and data 2: insertions

53: Tophat on data 3 and data 2: alignment summary

52: Tophat Brain

2.4 MB

format: bam, database: hg19

Log: tool progress

Log: tool progress

[2015-08-10 14:31:30]
Beginning TopHat run (v2.0.14)

[2015-08-10 14:31:30] Checking
for Bowtie
Bowtie version: 2.2.5.0
[2015-08-10 14:31:30] Checking
for Bo

display at UCSC main

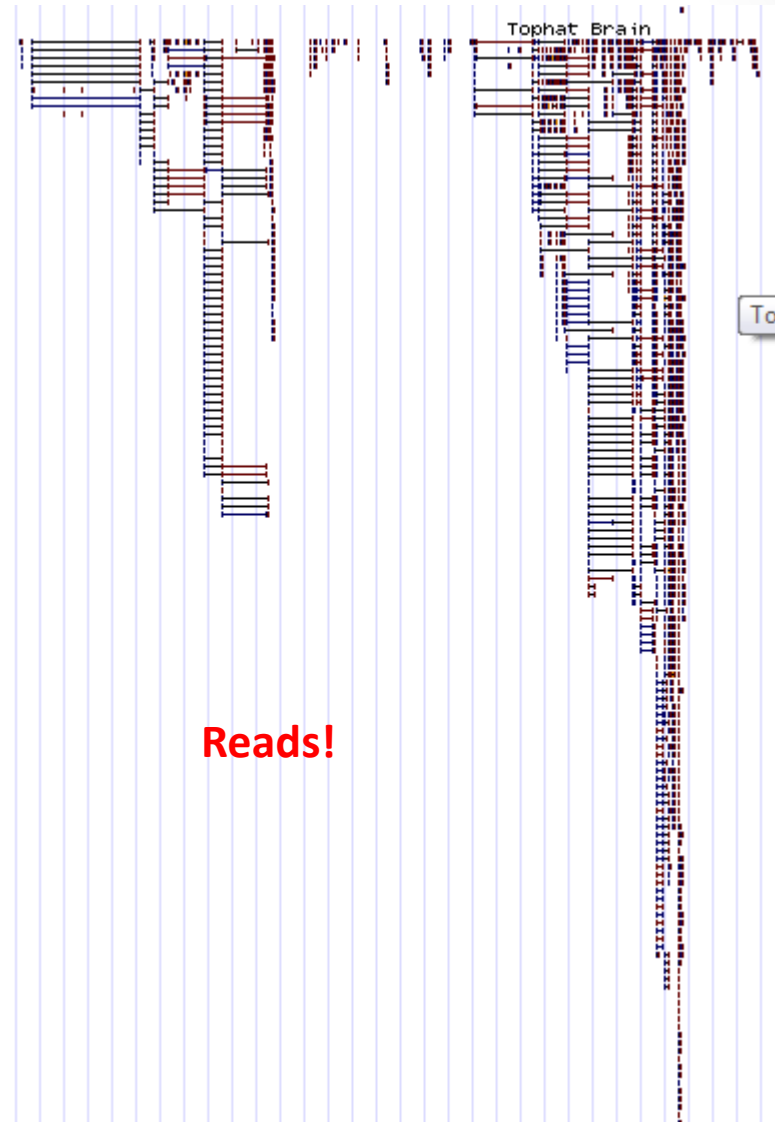
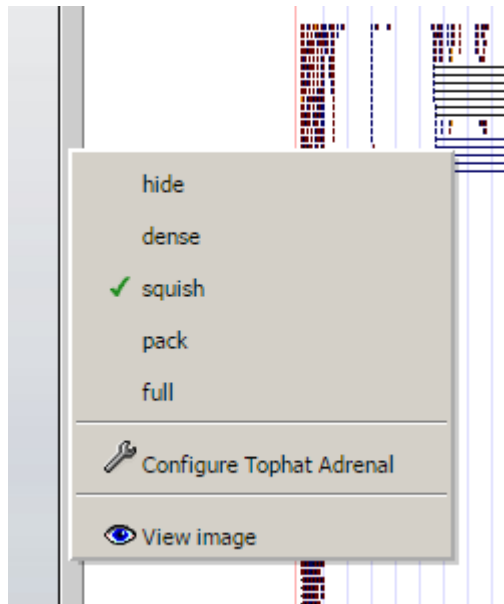
display at Ensembl Current

display with IGV local Human hg19

display in IGB View

UCSC Genome Browser

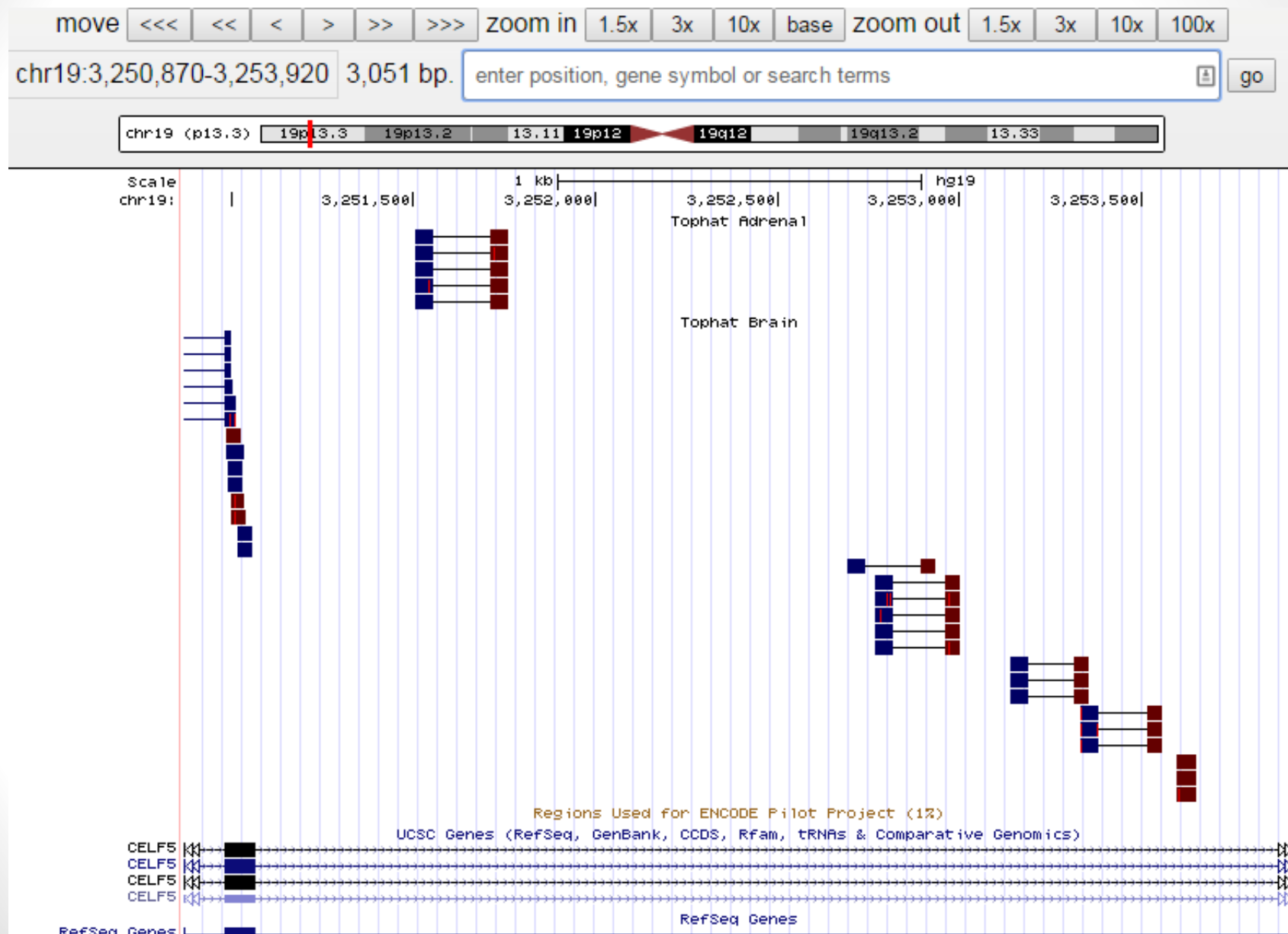
chr19:3,121,510-3,274,076



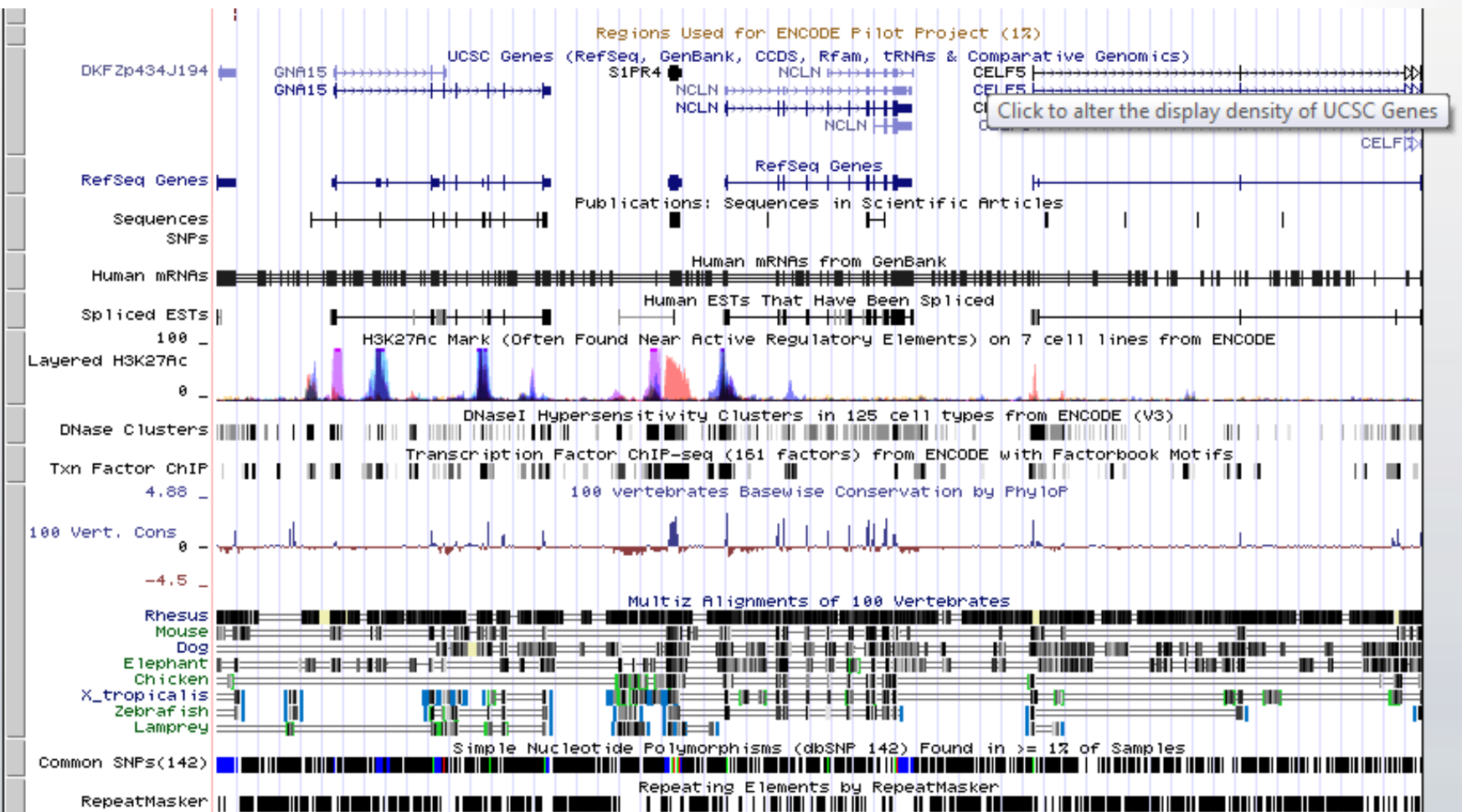
Reads!

Right Click, Select “squish” view

Narrow Down - Search for: chr19:3,250,772-3,253,822



UCSC Browser Comparisons



Lots and Lots of Comparisons...

track search default tracks default order hide all manage custom tracks track hubs configure reverse resize refresh

collapse all

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.

expand all

-

Custom Tracks

refresh

[junctions](#)

[Tophat Adrenal](#)

[Tophat Brain](#)

dense ▾

squish ▾

squish ▾

+ Mapping and Sequencing

refresh

+ Genes and Gene Predictions

refresh

+ Phenotype and Literature

refresh

+ mRNA and EST

refresh

+ Expression

refresh

+ Regulation

refresh

+ Comparative Genomics

refresh

+ Neandertal Assembly and Analysis

refresh

+ Denisova Assembly and Analysis

refresh

+ Variation

refresh

+ Repeats

refresh

refresh

Challenges

chr19:3,250,772-3,253,822

- Compare your reads to the Neandertal Genome.
- What transcription factors bind within this region?
- What SNPs are in this region?
- Can you find qPCR primers to amplify this region?

Sequences SNPs

Human mRNAs from GenBank

Spliced ESTs

Human (hg19) Whole Transcriptome qPCR Primers

27Aic Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

CEL5 uc010xhg.2_1_2_2

Human (hg19) Whole Transcriptome qPCR Primers (CELF5_uc010xhg.2_1_1_1)

Click here for primer details: [95629](#)

Item: CELF5_uc010xhg.2_1_1_1
Position: [chr19:3250995-3273911](#)
Band: 19p13.3
Genomic Size: 22917
Strand: +
[View DNA for this feature](#) (hg19/Human)

[View table schema](#)

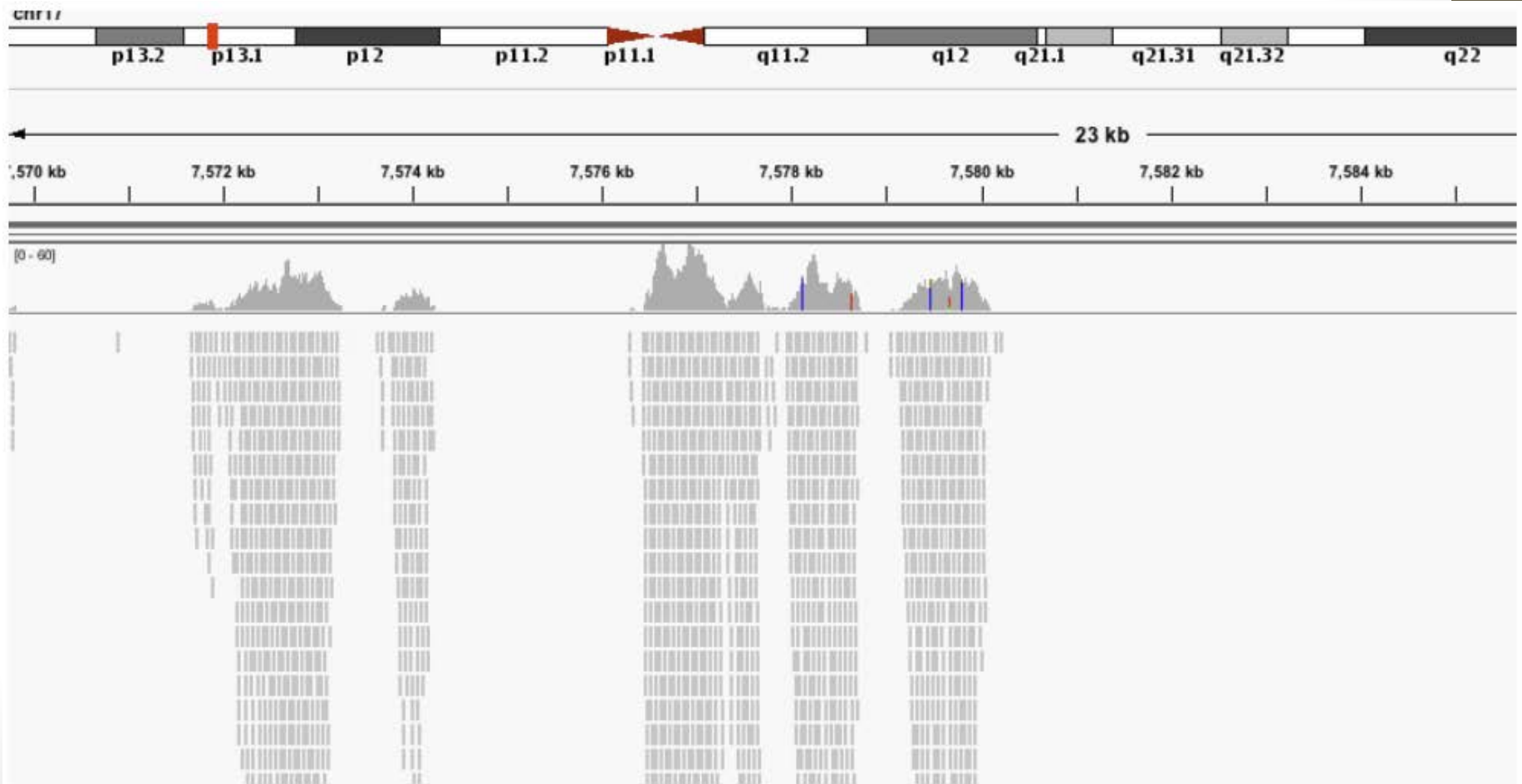
[Go to qPCR Primers track controls](#)

OLIGO	len	tm	gc%	any	3'	seq
FORWARD	20	60.533	55.000	4.00	2.00	TCACCTACTGTGCCAGGGAT
REVERSE	18	60.557	61.111	3.00	2.00	GCTTTCACTGTCCGCAGG

[Link to UCSC in-silico PCR](#)

Data Visualization – IGV

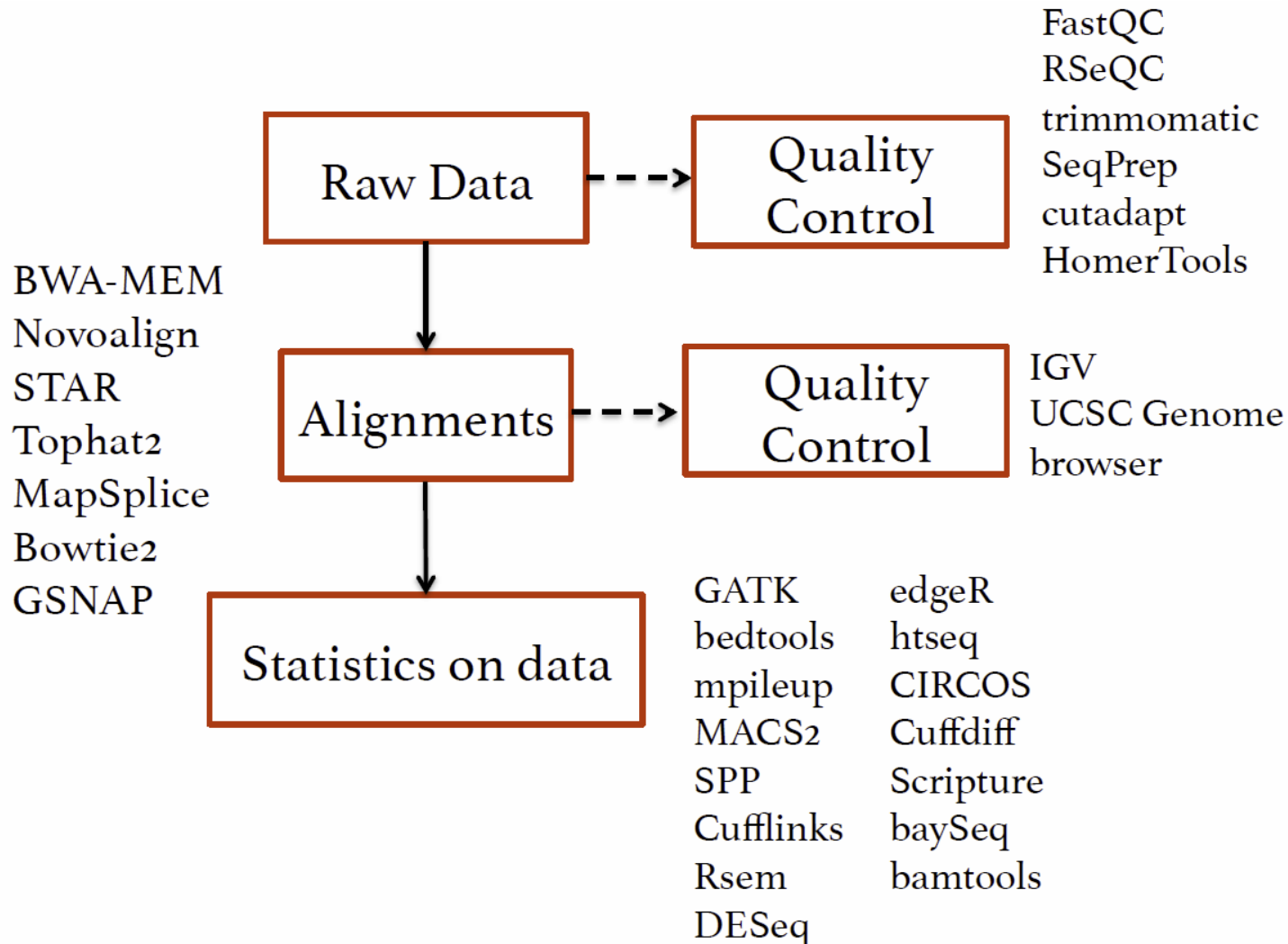
<https://www.broadinstitute.org/igv/>



The 2008 World Submarine Racing Championships



Step 3: Assembling Reads into Full-Length Transcripts



Cufflinks: Read Assembly

NGS: QC and manipulation

NGS: Mapping

NGS: RNA-seq

Cuffdiff find significant changes in transcript expression, splicing, and promoter use

Tophat Gapped-read mapper for RNA-seq data

StringTie transcript assembly and quantification

Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

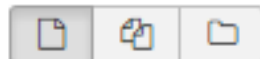
Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

- Takes reads from a BAM file and assembles the reads into mRNA transcripts.
- Select your BAM File and Run with Default Settings
 - Tophat Brain
 - Tophat Adrenal
 - **“Use Reference Annotation”**
- Rename “Assembled Transcripts” to “Cufflinks Brain” or “Cufflinks Adrenal”

Use Reference Annotation

Use reference annotation

Reference Annotation



1: iGenomes UCSC hg19, chr19 gene annotation

Gene annotation dataset in GTF or GFF3 format.

Cufflinks

Inputs and Outputs

Input:

- Alignment files (BAM)
- Indexed reference genome (FASTA)
- Known gene annotations (GTF) *Optional*

Output:

- Transcript assembly (GTF)
- Transcript abundance estimation (FPKM)

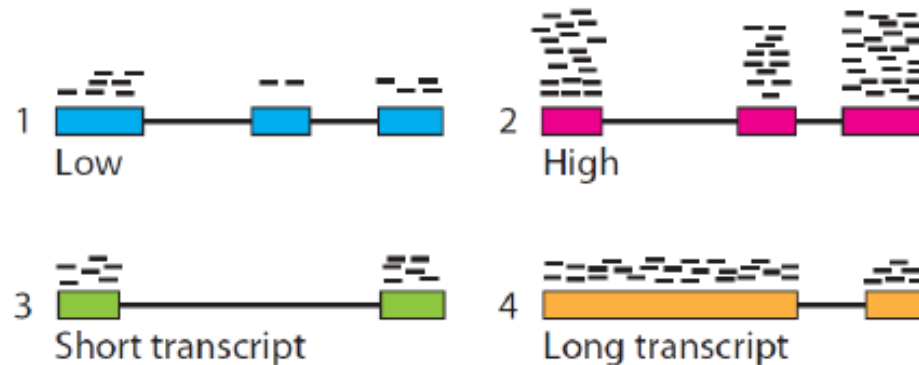
Parameters:

- num-threads
- GTF
- GTF-guide (assemble novel isoforms)
- library-type
- max-intron-length
- min-intron-length
- Many more!

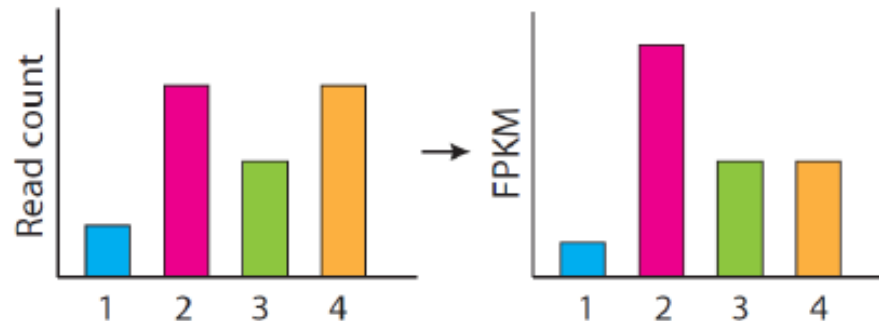
FPKM

$$\text{FPKM} = \frac{\text{total fragments}}{\text{mapped reads (millions)} * \text{exon length (KB)}}$$

FPKM: **F**ragments **P**er
Kilobase per **M**illion
mapped reads
(total fragments)



RPKM: **R**eads **P**er
Kilobase per **M**illion
mapped reads
(total exons reads)



<http://www.broadinstitute.org>

Number of fragments normalized by

- ✓ Transcript length (Kb)
- ✓ Total number of mapped reads (Million)

Cuffmerge

Produces a merged transcripts dataset that includes all transcripts in both datasets. Datasets info are still maintained, but used for differential analyses.

Galaxy Analyze Data Workflow Shared Data Visualize

Tools

NGS: QC and manipulation

NGS: Mapping

NGS: RNA-seq

- [Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use
- [Tophat](#) Gapped-read mapper for RNA-seq data
- [StringTie](#) transcript assembly and quantification
- [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- [gffread](#) Filters and/or converts GFF3/GTF2 records
- [Cuffmerge](#) merge together several Cufflinks assemblies**

Cuffmerge merge together several Cufflinks assemblies (Galaxy Tool Version)

GTF file(s) produced by Cufflinks

- 94: Cufflinks on data 1 and data 52: Skipped Transcripts
- 92: Cufflinks Brain**
- 82: Cufflinks Adrenal
- 1: iGenomes UCSC hg19, chr19 gene annotation

Additional GTF Inputs (Lists)

+ Insert Additional GTF Inputs (Lists)

Use Reference Annotation

Yes

Reference Annotation

- 1: iGenomes UCSC hg19, chr19 gene annotation**

Requires an annotation file in GFF3 or GTF format.

Use Sequence Data

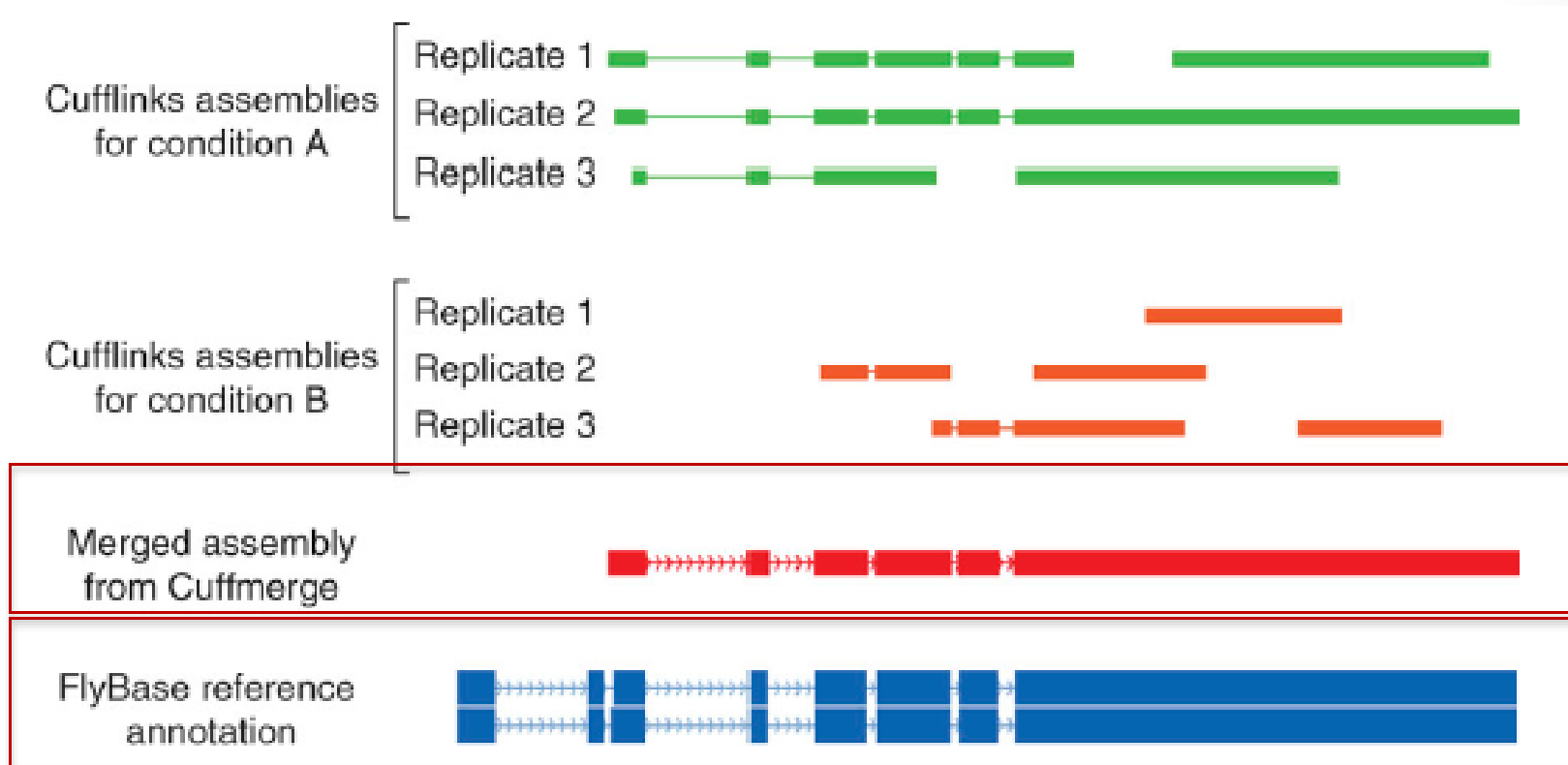
No

Use sequence data for some optional classification functions, including the add

Minimum isoform fraction

0.05


How Cufflinks/Cuffmerge Works...



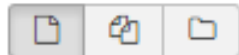
Cuffmerge will merge sequences if they overlap, and agree on splicing, and are in the same orientation. Differential transcripts are not merged.

Cuffdiff

Isolate List of Differentially-Expressed Genes

 **Cuffdiff** find significant changes in transcript expression, splicing, and promoter use (Galaxy Tool Version 2.2.1.2)

Transcripts



95: Cuffmerge on data 1, data 92, and data 82: merged transcripts

A transcript GFF3 or GTF file produced by cufflinks, cuffcompare, or other source.

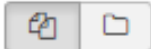
Condition

1: Condition

Name

Adrena

Replicates



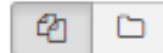
57: Tophat Adrenal
52: Tophat Brain

2: Condition

Name

Brain

Replicates



57: Tophat Adrenal
52: Tophat Brain

Cuffdiff Outputs

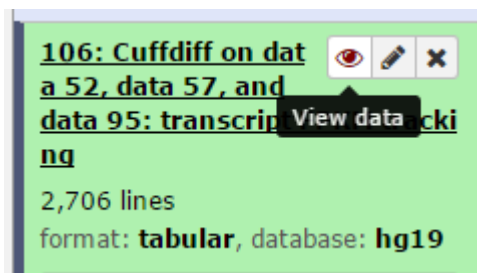
- Cuffdiff calculates the FPKM of each transcript, primary transcript, and gene in each sample.
- Primary transcript and gene FPKMs are computed by summing the FPKMs of transcripts in each primary transcript group or gene group.

There are multiple FPKM file types:

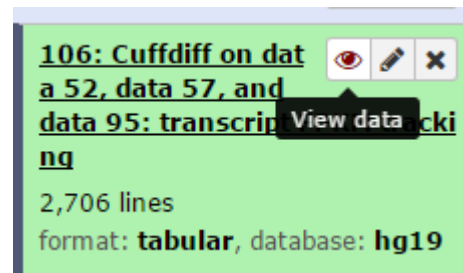
- Isoforms
- Genes
- CDS (coding sequence)
- TSS (transcription start sites)
- Promoter

Cuffdiff: Output Styles

Transcript FPKM Tracking



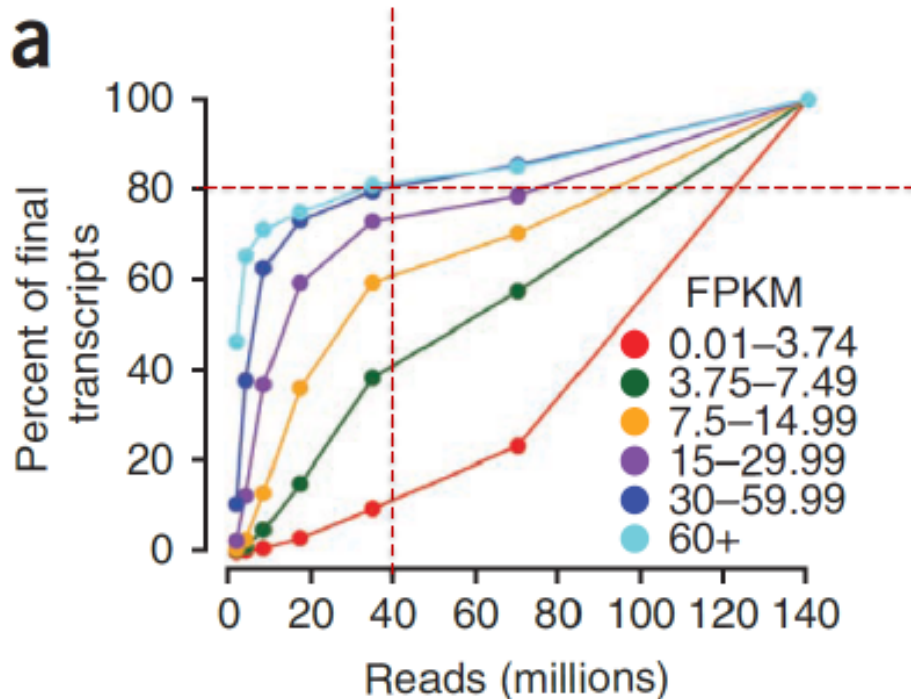
Transcript differential expression testing



7	8	9	10	11	12	13	14
status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
NOTEST	0	0	0	0	1	1	no
NOTEST	0	0	0	0	1	1	no
NOTEST	0	0	0	0	1	1	no
NOTEST	0	0	0	0	1	1	no
NOTEST	0	0	0	0	1	1	no
NOTEST	0	0	0	0	1	1	no
NOTEST	34.4514	0	-inf	0	1	1	no
NOTEST	0	0	0	0	1	1	no
NOTEST	0.0797287	0	-inf	0	1	1	no
NOTEST	0	0	0	0	1	1	no

OK; NOTEST (not enough alignments); LOWDATA; HIDATA; FAIL
Why so many marked “NOTEST”?

Coverage Affects Expression Estimation....



ENCODE saturation analysis

- 214 million 2x100bp PE reads
- H1 human embryonic stem cells
- 80% of the genes with FPKM ≥ 10 are detected by ~36 million mapped reads per sample
- Genes with FPKM < 10 : ~80 million mapped reads per sample

Trapnell et al. Nature Protocol 2012

Sims et al., 2014 Nature Reviews

Discussion Points:

- Always consider low-abundance transcripts and rare events, such as splice variants. Can you ever over-sequence?
- What other information can you use to cross-check the accuracy of your data and gene detection?

Let's Get this Data into Excel...

These are standard tab-separated values; default input settings usually work into any spreadsheet-style program.

105: Cuffdiff on data 52, data 57, and data 95: transcript expression testing

2,706 lines

format: **tabular**, database: hg19

[11:19:31] Loading reference annotation.

Warning: No conditions are replicated, switching to 'blind' dispersion method

[11:19:31] Inspecting maps and determining fragment length distributions.

Warning: Using default Gaussian distribution due to insuffici



Download

test_id	gene_id	gene	loc
TCONS_00000001	XLOC_000001	OR4F17	chr19:110678-111596
TCONS_00000002	XLOC_000002	MADCAM1	chr19:496489-505343

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

- ☒ Delimited - Characters such as commas or tabs separate each field.
- ☐ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row:

1

File origin:

437 : OEM United States

Preview of file ...Galaxy105-[Cuffdiff_on_data_52_data_57_and_data_95_transcript_differen...

```
1 test_idgene_idgenelocussample_1sample_2statusvalue_1value_2log2(  
2 TCONS_00000001XLOC_000001OR4F17chr19:110678-111596AdrenaBrainNOTEST  
3 TCONS_00000002XLOC_000002MADCAM1chr19:496489-505343AdrenaBrainNOTES  
4 TCONS_00000003XLOC_000002MADCAM1chr19:496489-505343AdrenaBrainNOTES  
5 TCONS_00000004XLOC_000003TPGS1chr19:507496-519654AdrenaBrainNOTEST
```

Cancel

< Back

Next >

Finish

Understand Your Data...

B2		X	✓	f _x	XLOC_000001									
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
2	TCONS_00000001	XLOC_000001	OR4F17	chr19:110678-111596	Adrena	Brain	NOTEST	0	0	0	0	1	1	no
3	TCONS_00000002	XLOC_000002	MADCAM1	chr19:496489-505343	Adrena	Brain	NOTEST	0	0	0	0	1	1	no
4	TCONS_00000003	XLOC_000002	MADCAM1	chr19:496489-505343	Adrena	Brain	NOTEST	0	0	0	0	1	1	no
5	TCONS_00000004	XLOC_000003	TPGS1	chr19:507496-519654	Adrena	Brain	NOTEST	0	0	0	0	1	1	no
6	TCONS_00000005	XLOC_000004	CDC34	chr19:531732-542087	Adrena	Brain	NOTEST	0	0	0	0	1	1	no

Gene

Locus

OK vs.
NOTEST

FPKM

Adrenal

FPKM

Brain

Log2(Fold_Change): The (base 2) log of the fold change y/x

Test_stat: The value of the test statistic used to compute significance of the observed change in FPKM

P_value: The uncorrected p-value of the test statistic

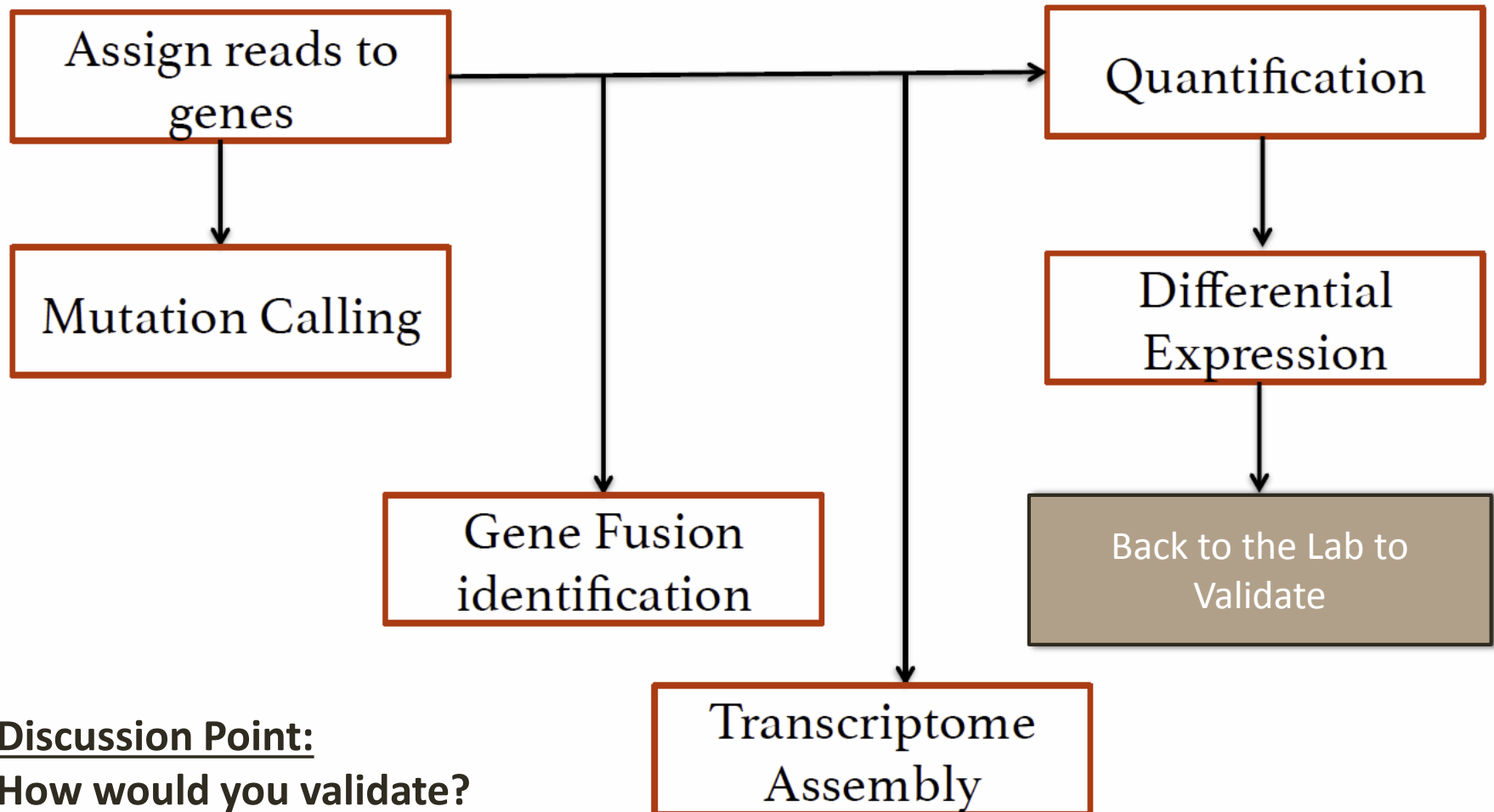
Q_value: The FDR-adjusted p-value of the test statistic

Q1: How many genes are significantly changed?

Q2: Why so few?

Q3: Why are there duplicate genes listed?

Data Obtained, now back to the Lab to Validate!



Discussion Point:
How would you validate?

To use, share, and publish

Tools which cannot be run interactively and thus cannot be inco

Workflow constructed from history 'MBL Bootcamp, D'

Check all

Uncheck all

Workflow "Workflow constructed from history 'MBL Bootcamp, DVG Workshop'" created from current history. You can edit or run the workflow.

Dataset Security

Resume Paused Jobs

Collapse Expanded Datasets

Unhide Hidden Datasets

Delete Hidden Datasets

Purge Deleted Datasets

[Show Structure](#)

Export Citations

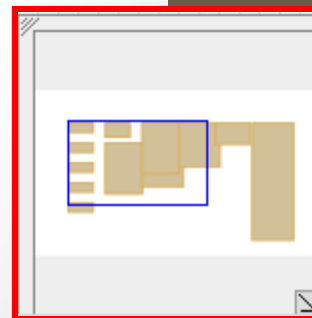
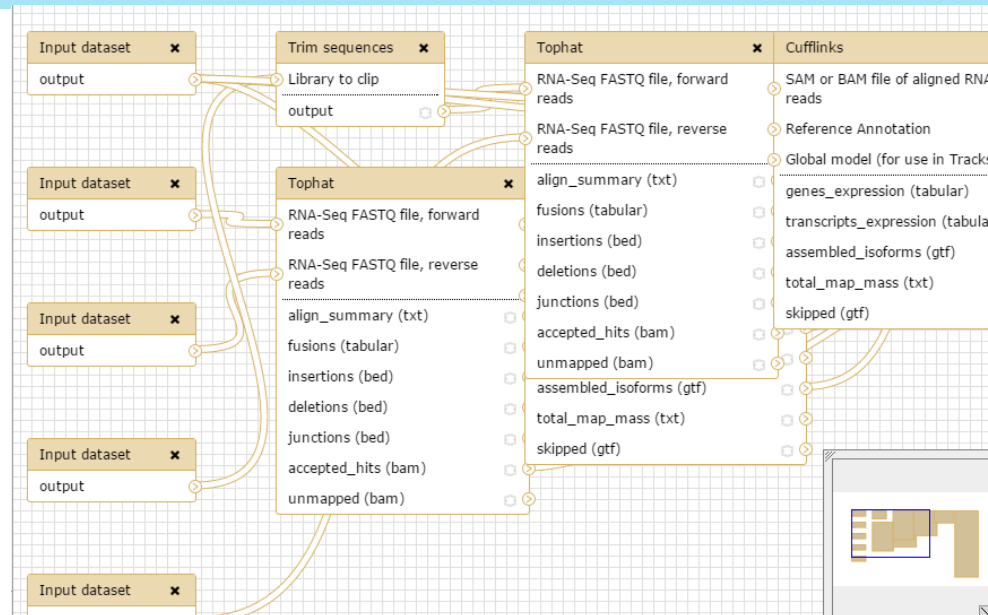
Export to File

Delete

Delete Permanently

OTHER ACTIONS

Import from File



Many Options for Analyses Tools...

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners ⁸	Seed methods	Short-read mapping package (SHRiMP) ⁴¹	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy ³⁹	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie ⁴³			
		BWA ⁴⁴	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice ⁵²	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap ⁵⁰			
		TopHat ⁵¹	Uses Bowtie alignments		
	Seed-extend methods	GSNAP ⁵³	Can use SNP databases		
		QPALMA ⁵⁴	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture ²⁸	Reports all isoforms		
		Cufflinks ²⁹	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet ⁵¹	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq ⁴⁷	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE) ²⁰	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA) ⁸²	Quantifies using unique reads		
	Isoform quantification	Cufflinks ²⁹	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO ³³			
		RNA-seq by expectation maximization (RSEM) ⁶⁹			
Differential expression		Cuffdiff ²⁹	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq ⁷⁹	Uses a normal distribution		
		EdgeR ²⁷			
		Differential Expression analysis of count data (DESeq) ⁷⁸			
		Myrna ⁷⁵	Cloud-based permutation method		

Finding Publically Available Data



Federal-Funded Studies Must Make Datasets Publically Available

Gene Expression Omnibus (GEO)

<http://www.ncbi.nlm.nih.gov/geo/>

Both sequencing and array data

Sequence Read Archive (SRA)

<http://www.ncbi.nlm.nih.gov/sra/>

Sequencing data

European Nucleotide Archive (ENA)

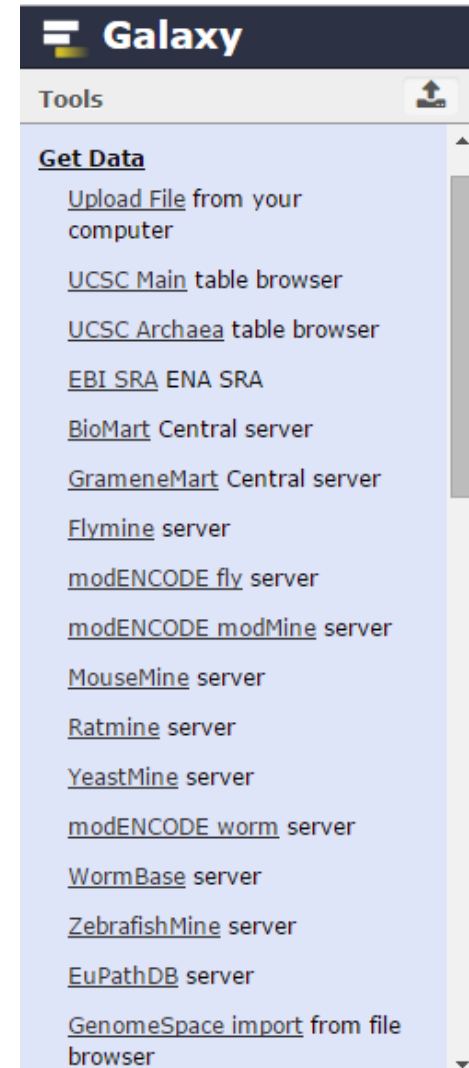
<http://www.ebi.ac.uk/ena>

Sequencing data

UCSC Genome Browser

<http://genome.ucsc.edu/>

Can import directly into Galaxy



Other Big Sequencing Projects

- ENCODE: Encyclopedia of DNA Elements. The **ENCODE** Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI).
- Illumina Body Map
- 1000 Genomes
- TCGA: The Cancer Genome Atlas
- Cancer Cell Line Encyclopedia (CCLE)
- cMAP: The Connectivity Map (or CMap) is a catalog of gene-expression data collected from human cells treated with chemical compounds and genetic reagents.

cMAP Drug Targets

<http://www.broadinstitute.org/cmap>





















total instances: 6100 , signature: HDACi (Glaser), export: Excel

search:

by name

by name and cell line

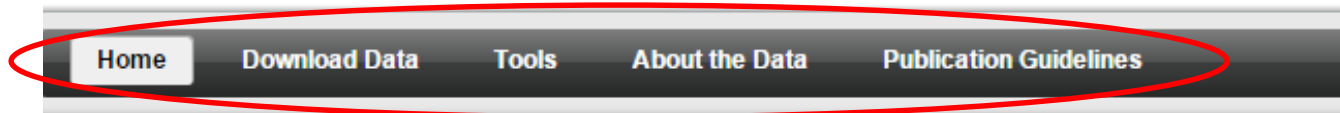
by ATC code

rank	cmap name	mean	n	enrichment	p	specificity	% non-null	
1	vorinostat	0.865	12	0.973	0.00000	0.0201	100	
2	trichostatin A	0.786	182	0.895	0.00000	0.0095	97	
3	geldanamycin	0.484	15	0.705	0.00000	0.0163	100	
4	fluphenazine	0.388	18	0.629	0.00000	0.0155	88	
5	trifluoperazine	0.392	16	0.625	0.00000	0.0625	87	
6	thioridazine	0.440	20	0.599	0.00000	0.1278	85	
7	tanespimycin	0.431	62	0.574	0.00000	0.0259	87	
8	sirolimus	0.337	44	0.491	0.00000	0.0542	77	
9	LY-294002	0.324	61	0.486	0.00000	0.0738	68	
10	valproic acid	0.304	57	0.359	0.00000	0.0263	61	
11	CP-690334-01	0.507	8	0.735	0.00002	0.0121	87	
12	rifabutin	0.735	3	0.971	0.00004	0.0052	100	
13	5707885	0.549	4	0.913	0.00004	0.0000	100	
14	pioglitazone	-0.337	11	-0.646	0.00004	0.0061	72	
15	6-bromoindirubin-3'-oxime	-0.532	7	-0.770	0.00008	0.0047	85	
16	withaferin A	0.542	4	0.896	0.00010	0.0632	100	
17	wortmannin	0.382	18	0.501	0.00010	0.1355	77	
18	ivermectin	0.461	5	0.858	0.00012	0.0215	100	
19	prochlorperazine	0.362	16	0.524	0.00014	0.1262	68	
20	suloctidil	0.553	4	0.888	0.00016	0.0182	100	

<< < **1** 2 3 4 5 > >>

TCGA: The Cancer Genome Atlas

- <https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp>



Home

TCGA Data Portal Overview

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes.

Please note some data on the TCGA Data Portal are in controlled-access. Please visit the [Access Tiers](#) page for more information.

The TCGA Data Portal does not host lower levels of sequence data. NCI's [Cancer Genomics Hub \(CGHub\)](#) is the new secure repository for storing, cataloging, and accessing BAM files and metadata for sequencing data.

[Download Data](#) ▶

Choose from four ways to
download data

Let's Play

<http://www.cbioportal.org/index.do>

Query **Download Data**

Select Cancer Study:
 No studies selected.

☐ Biliary Tract (3)

☐ Cholangiocarcinoma (3)

☐ Intrahepatic Cholangiocarcin... (Johns Hopkins University, Nature Genetics 2013)☐ Cholangiocarcinoma (National Cancer Centre of Singapore, Nature Genetics 2013)☐ Cholangiocarcinoma (National University of Singapore, Nature Genetics 2012)

☐ Bladder Urinary Tract (5)

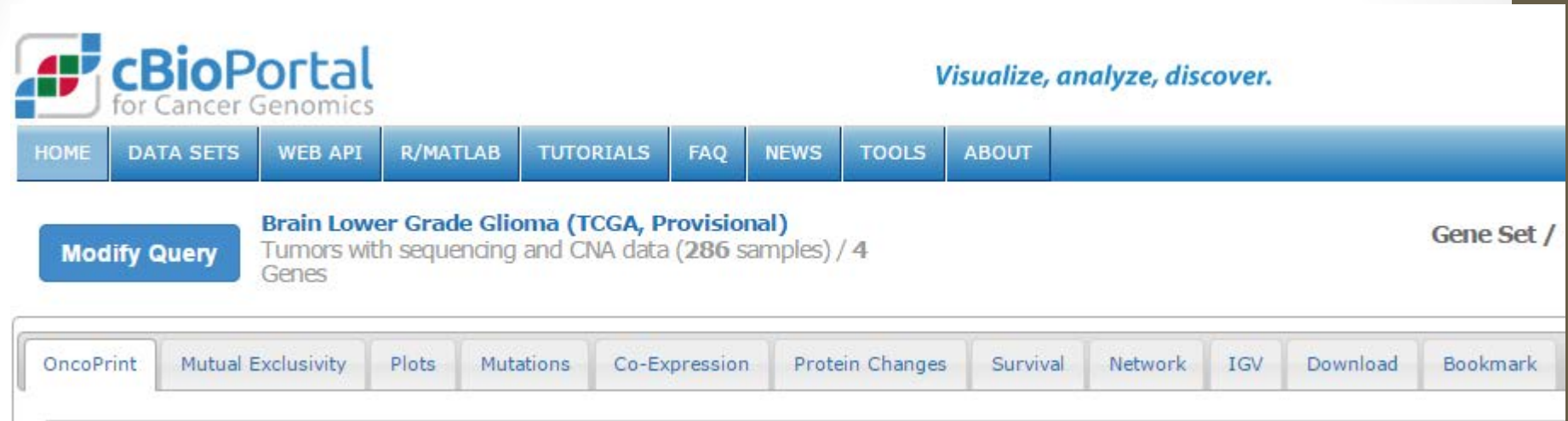
☐ Bladder Urothelial Carcinoma (5)

☐ Bladder Urothelial Carcinoma (BGI, Nature Genetics 2013)☐ Bladder Cancer (MSKCC, JCO 2013)

Select Data Type Priority: ☒ Mutation and CNA ☐ Only Mutation ☐ Only CNA

Enter Gene Set: Advanced: Onco Query Language (OQL)

cBioPortal: Outputs



The screenshot shows the cBioPortal website. At the top left is the cBioPortal logo with the tagline "for Cancer Genomics". To the right is the slogan "Visualize, analyze, discover." Below this is a blue navigation bar with links: HOME, DATA SETS, WEB API, R/MATLAB, TUTORIALS, FAQ, NEWS, TOOLS, and ABOUT. Below the navigation bar, there is a "Modify Query" button on the left. To its right, the query is identified as "Brain Lower Grade Glioma (TCGA, Provisional)" with details "Tumors with sequencing and CNA data (286 samples) / 4 Genes". On the far right of this section is a link "Gene Set /". Below this is a horizontal toolbar with buttons for various analysis tools: OncoPrint, Mutual Exclusivity, Plots, Mutations, Co-Expression, Protein Changes, Survival, Network, IGV, Download, and Bookmark.

- Are genes mutated in cancer cancers?
- What about changes in expression?
- What pathways are they connected to?
- Do changes in expression impact patient survival?
- What other proteins are co-expressed with our gene set?
- Are changes mutually exclusive?

One Stop Bioinformatics Shopping...

<https://gsui.genomespace.org/>

← → ↻ 🌐 <https://gsui.genomespace.org/jsui/jsui.html#>
















📁 Apps 📁 Popular 📁 Personal 📁 Grants and Man... 📁 My NCBI - Home 📁 News 📁 Oak Park 📁 UofC 📁 Kids 📁 Science 📁 Papers To Read 📁 MBL Bootcamp 📁 Bioinformatics 📁 Northwestern ... 📁 Gluten 📁 XtraMath 📁

GENOMESPACE [Invite a collaborator](#)

File | Launch | View | Connect | Manage | Recipes | Help

ArrayExpress cBioPortal CCLE Cistrome Cytoscape Cytoscape 3 Galaxy GenePattern Genomica geWorkbench

Order by: 🔍

 GenePattern The Broad Institute	 cBioPortal Memorial Sloan Kettering Cancer Center	 Reactome Ontario Institute for Cancer Research	 Project Achilles The Broad Institute	 CCLE The Broad Institute
 MMGP The Broad Institute	 Cytoscape 3 UCSD	 Synapse Sage Bionetworks	 MSigDB broadinstitute.org	 ISACreator isatools.org
 Gitools	 geWorkbench	 ArrayExpress	 InSilicoDB	 Cistrome

Overview

- **Part 1:** Analyses of FASTQ RNAseq Data
- **Part 2:** Data Visualization
- **Part 3:** Utilizing Online Databases

Wrap Up...

- Good experimental design and quality of samples are still critical.
 - Garbage in, Garbage out.
 - RNA quality
- Consult with a statistician b/4 diving in. Biological replicates vs. technical replicates vs. read depth.
- UChicago Center for Research Informatics
 - <http://cri.uchicago.edu/>
- Before doing the sequencing yourself, search to see if someone else has already done it.
- Good way to generate hypothesis, still require validation in lab.

Acknowledgements

- UChicago CRI faculty and staff
- Alex Ling, CCB
- MBL Staff
- Vicky Prince, Stephanie Palmer, and Stefano Allesina

Contact me:

prostate@uchicago.edu



GenomeSpace

1. GenomeSpace Import
2. Public Folder
3. Carcinoman
4. MBL Files

Get Data

[Upload File](#) from your computer

[UCSC Main](#) table browser

[UCSC Archaea](#) table browser

[EBI SRA](#) ENA SRA

[BioMart](#) Central server

[GrameneMart](#) Central server

[Flymine](#) server

[modENCODE fly](#) server

[modENCODE modMine](#) server

[MouseMine](#) server

[Ratmine](#) server

[YeastMine](#) server

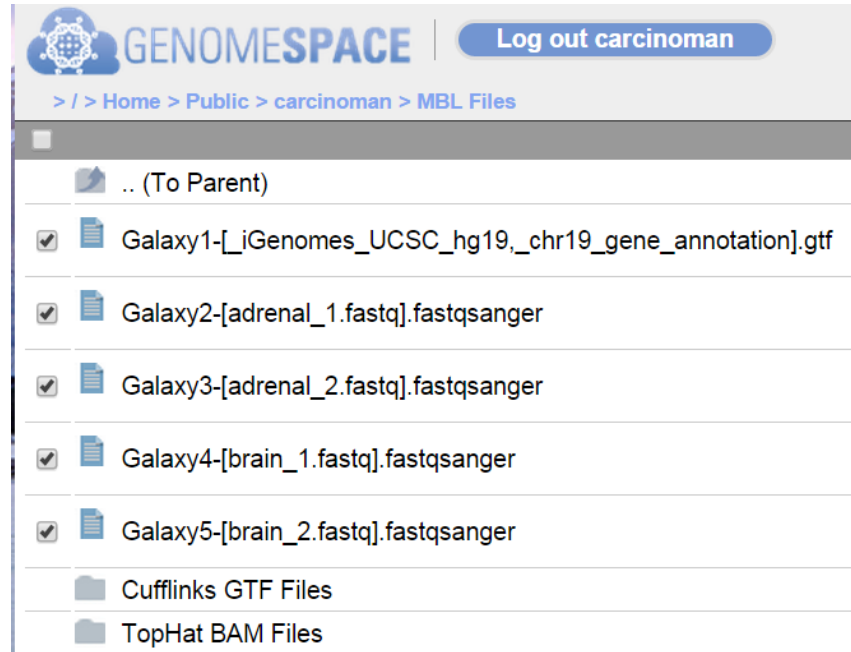
[modENCODE worm](#) server

[WormBase](#) server

[ZebrafishMine](#) server

[EuPathDB](#) server

[GenomeSpace import](#) from file browser



5 files selected

[Send to Galaxy](#)