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)

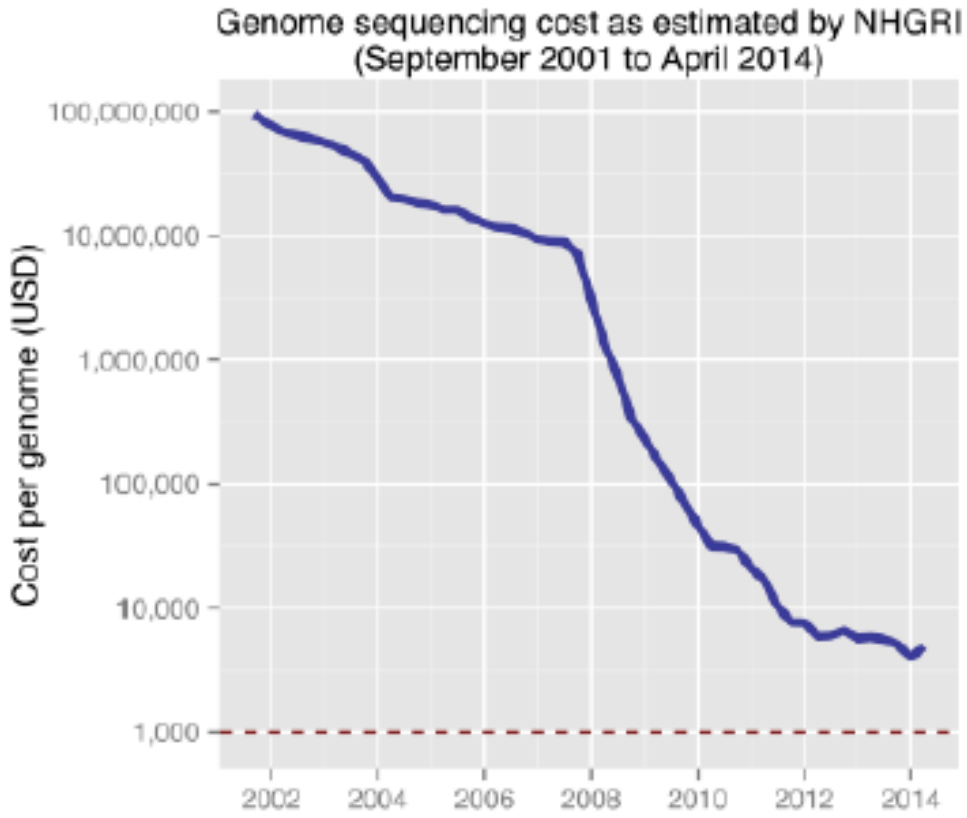
Overview

- **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

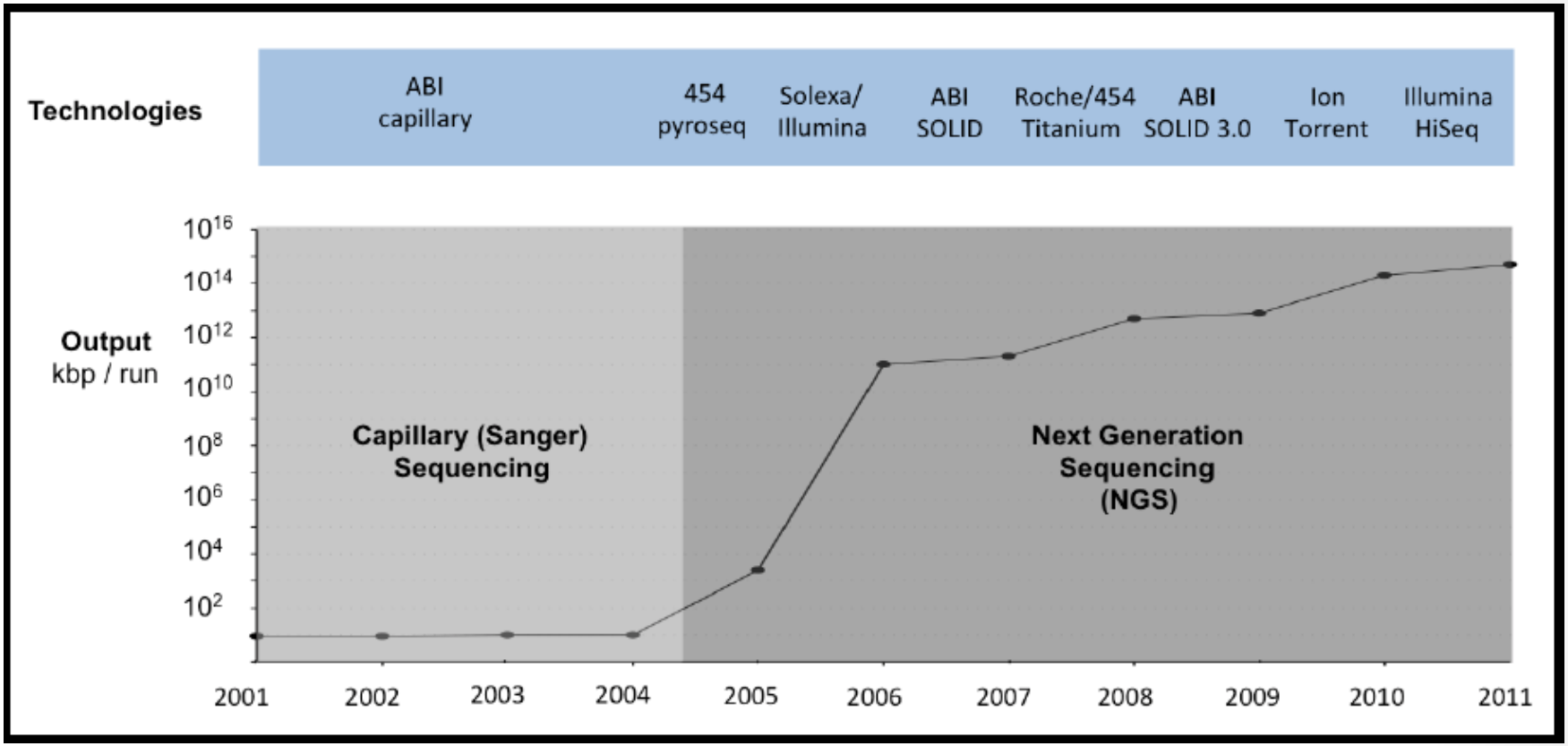
Yoav Gilad, PhD
Scientific Director

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

The Rate of Data Acquisition ....



Broad Applications of NGS

Broad applications of NGS to drug discovery

Applications	Pros of NGS	Cons of NGS	Alternatives	Refs
Mutation detection: personalised medicine	Can sequence large genome regions to identify efficacy markers	Initial setup and running cost for NGS	Large-scale Sanger sequencing technology	[64]
ChIP-Seq: target identification and/or validation and compound profiling for epigenetics	Enables study of epigenetic targets at the whole-genome level	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	Comparative genomic hybridisation	[35,65, 66]
Exome sequencing: target identification and/or drug resistance studies, biomarker discovery	Identify rare variants, using deep sequence coverage	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	Microarray-based technology	[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

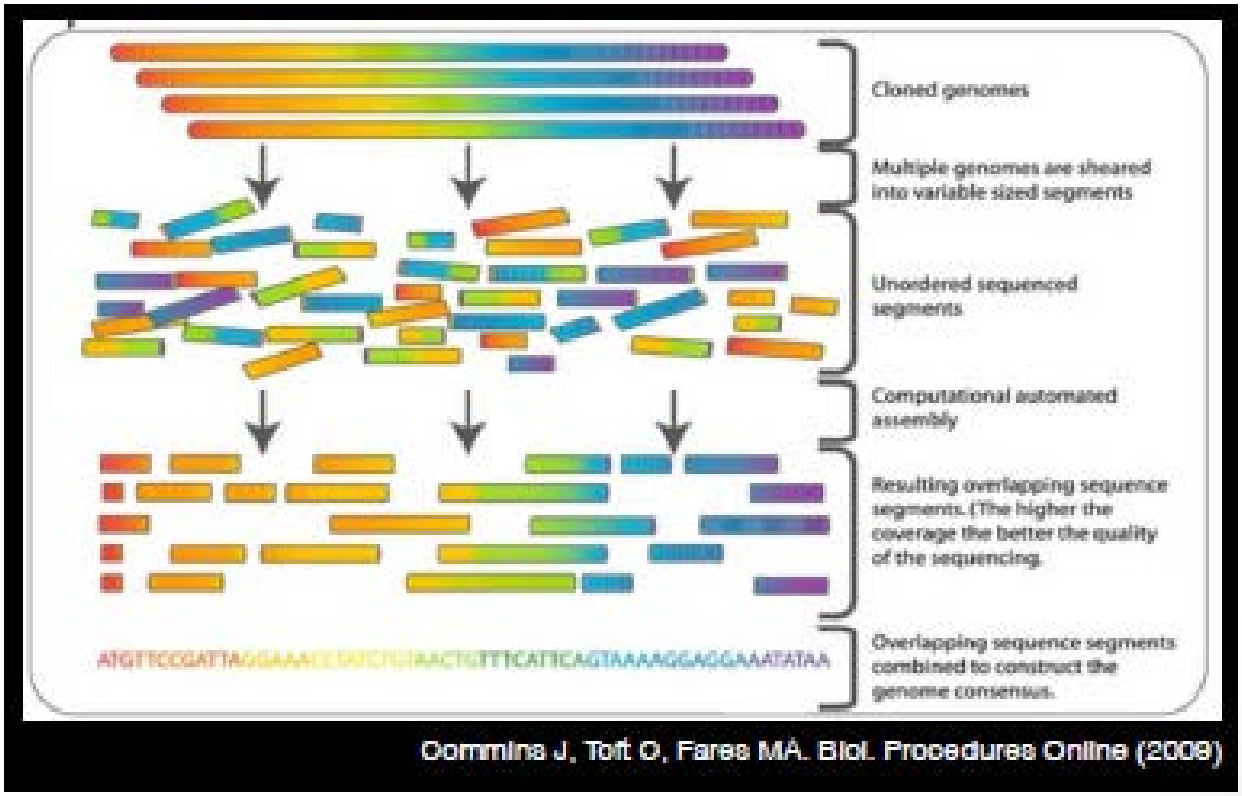
- NGS-based Technology to **qualitatively** and **quantitatively** profile the **full set of transcripts** (i.e., transcriptome), including **mRNAs**, **small RNAs** and other **non-coding RNAs.**

- **Transcriptome profiling** provides a snapshot of gene expression patterns and regulatory elements

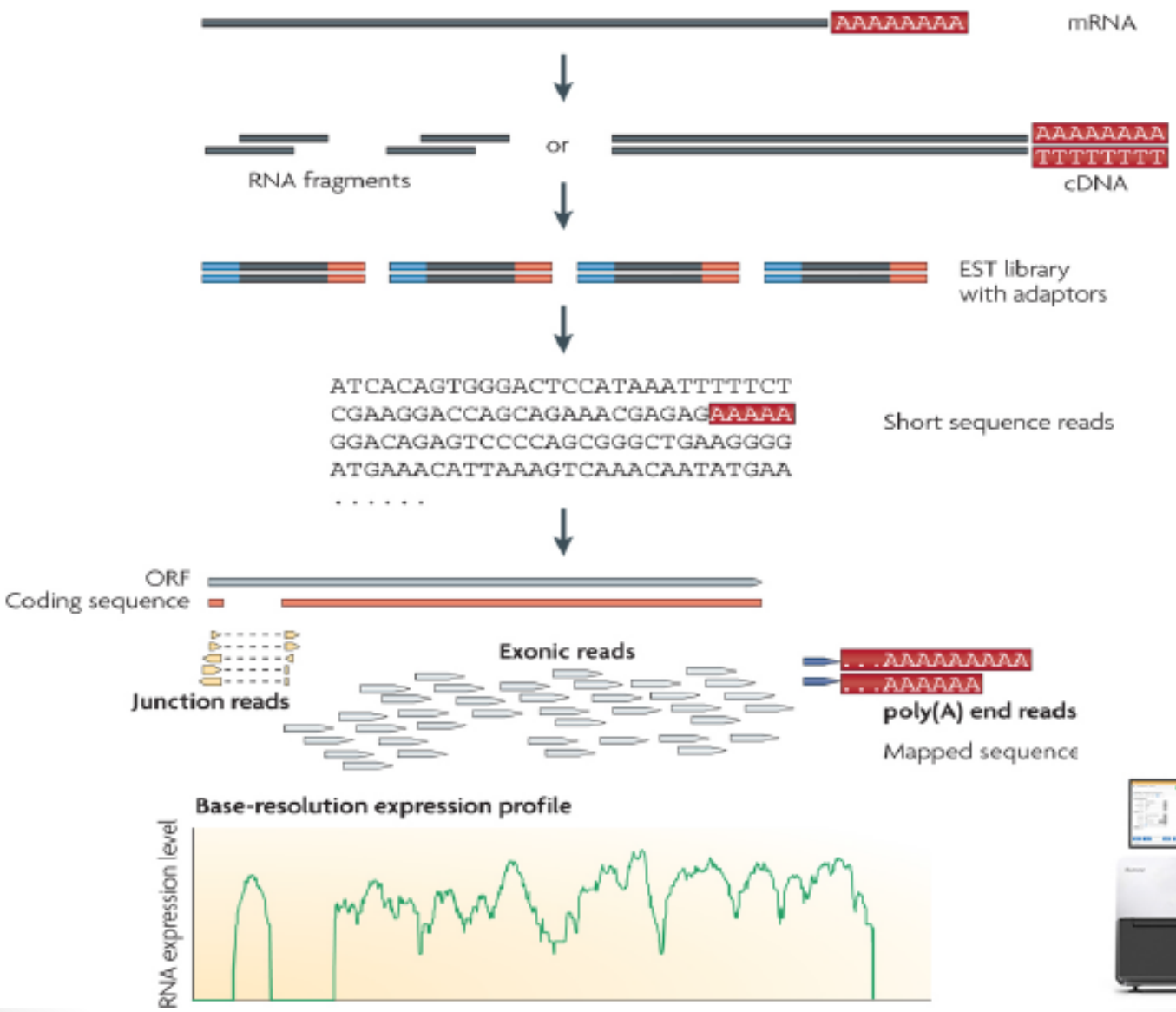
- Although a **transcriptome** only represents a small fraction of the human genome (**<5%**), it is **very complex**, transcripts derived from **alternative splicing**, **gene fusion**, **antisense transcription** and **RNA editing** largely increase the **diversity** of the transcriptome

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



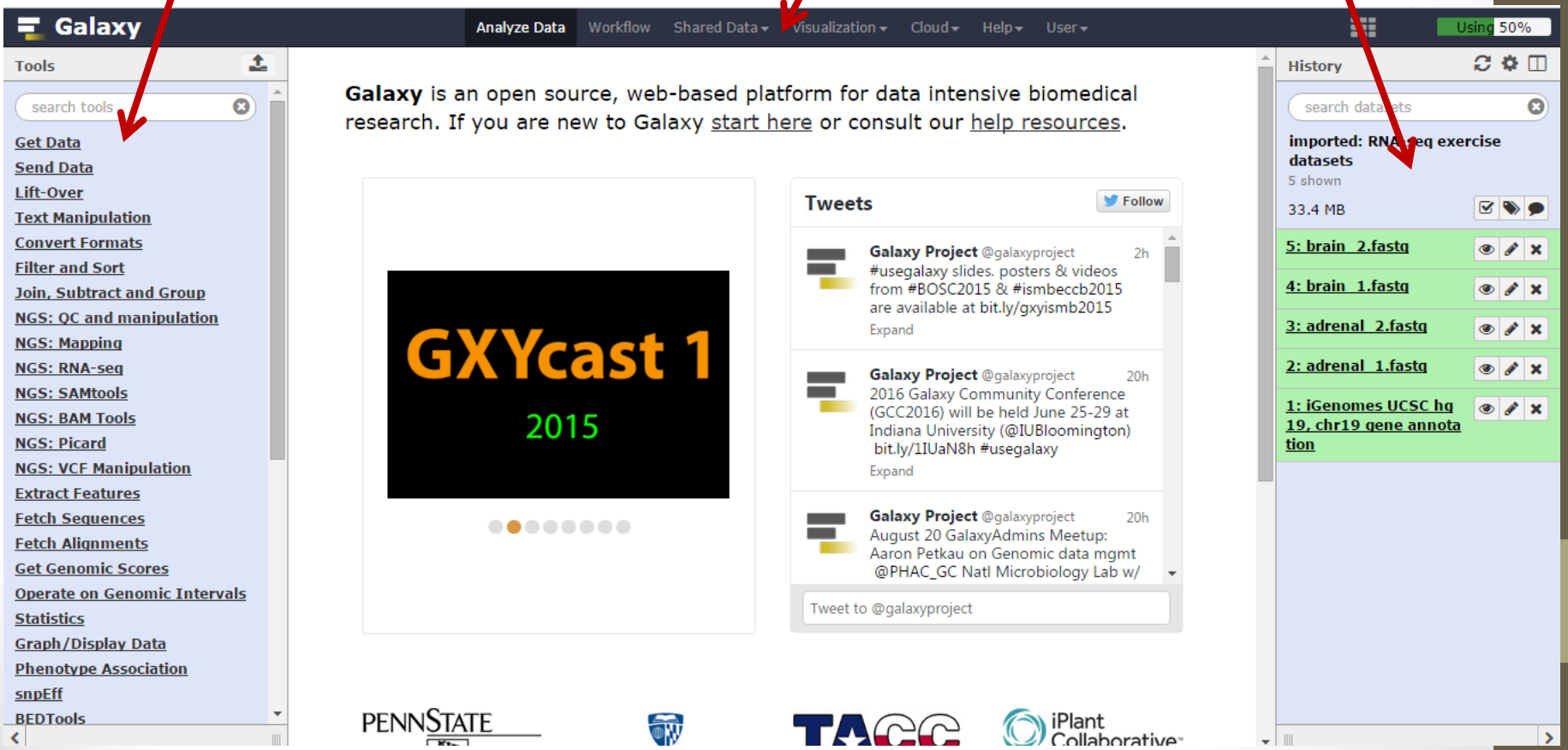
Let's Get Started…

https://usegalaxy.org

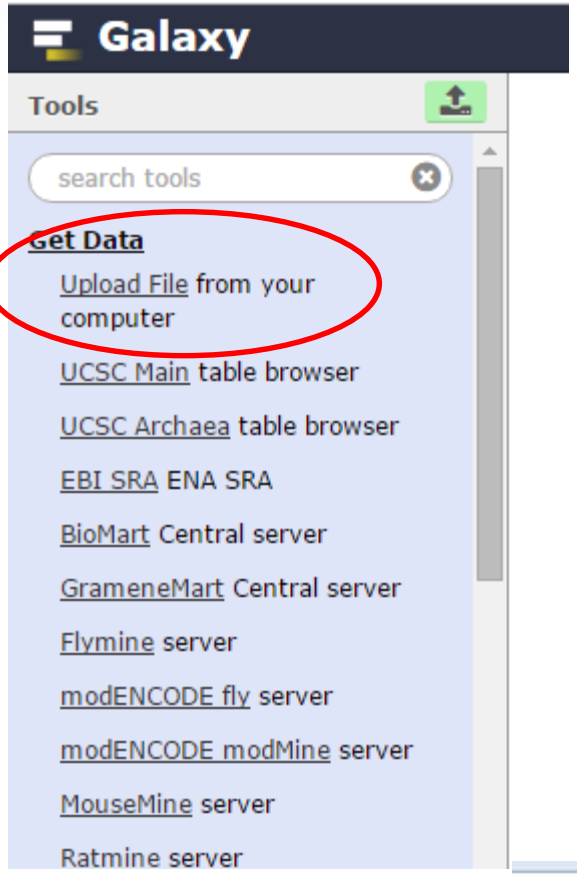
Menu of Analyses

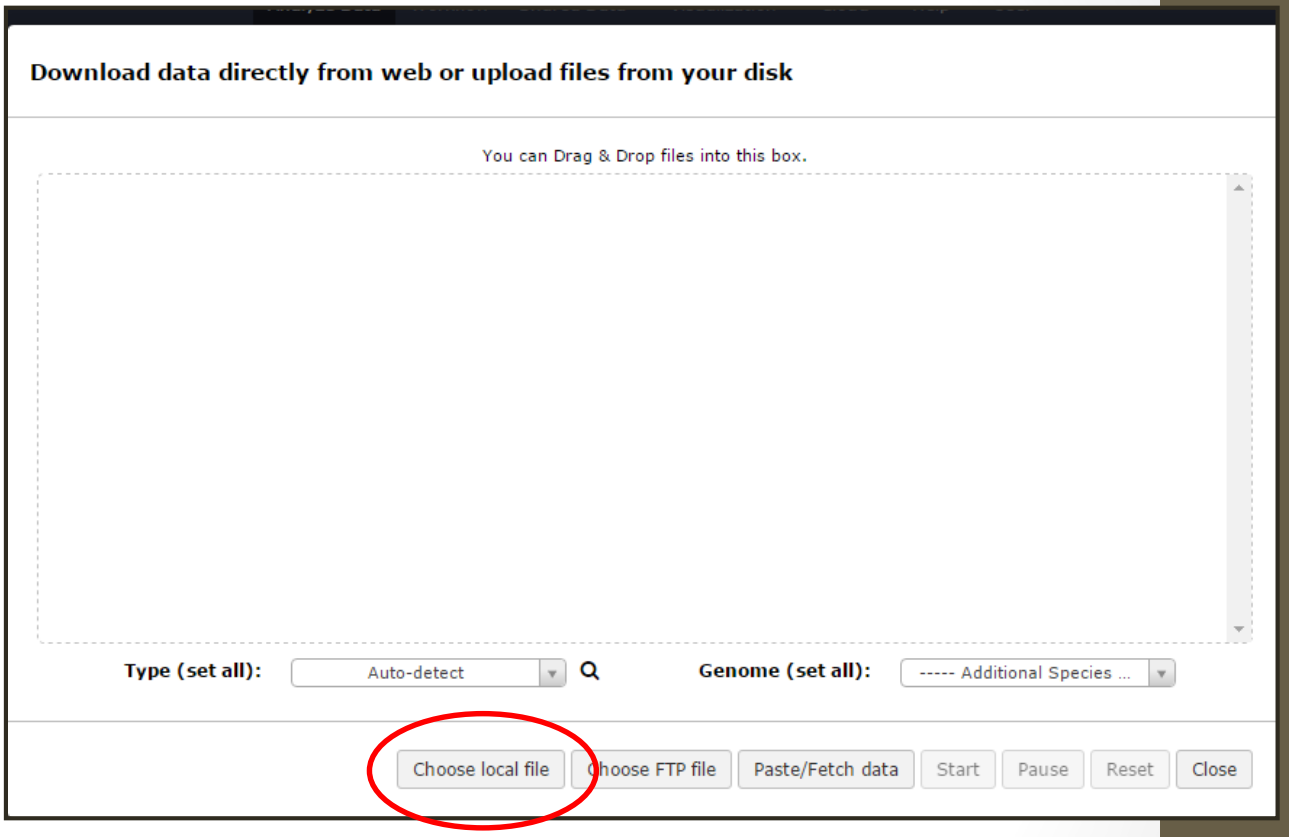
User Interface

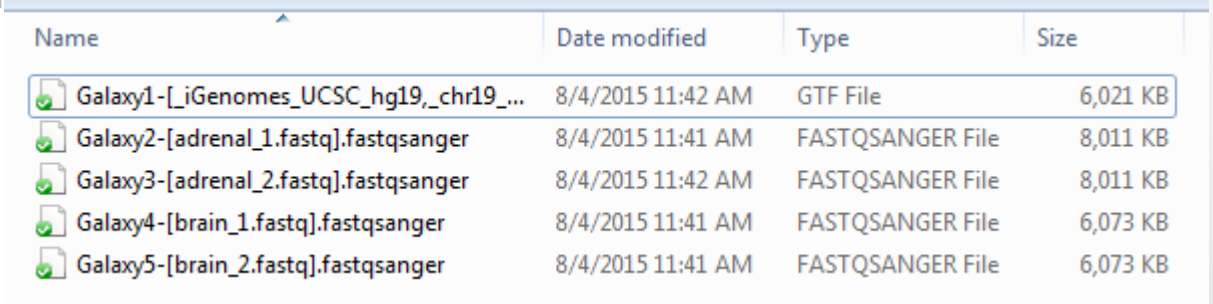
Your Analysis History



Upload Data – 5 Files







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_hg19,_chr19_gene_annotation].gtf	6.2 MB	Auto-det...	----- Additional Sp...		100%
Galaxy2-[adrenal_1.fastq].fastqsanger	8.2 MB	Auto-det...	----- Additional Sp...		100%
Galaxy3-[adrenal_2.fastq].fastqsanger	8.2 MB	Auto-det...	----- Additional Sp...		100%
Galaxy4-[brain_1.fastq].fastqsanger	6.2 MB	Auto-det...	----- Additional Sp...		36%

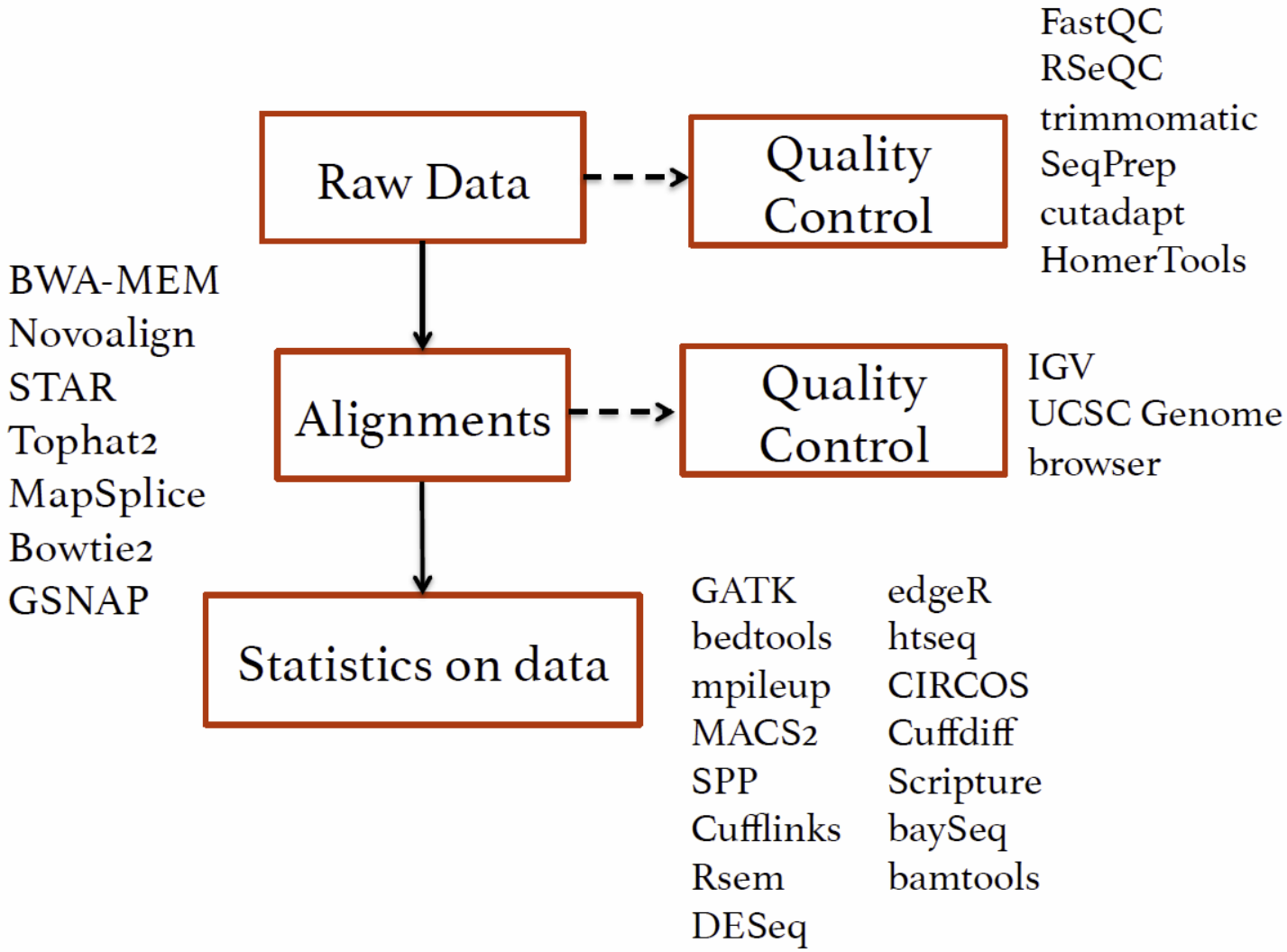
Type (set all): Auto-detect

Genome (set all): ----- Additional Species ...

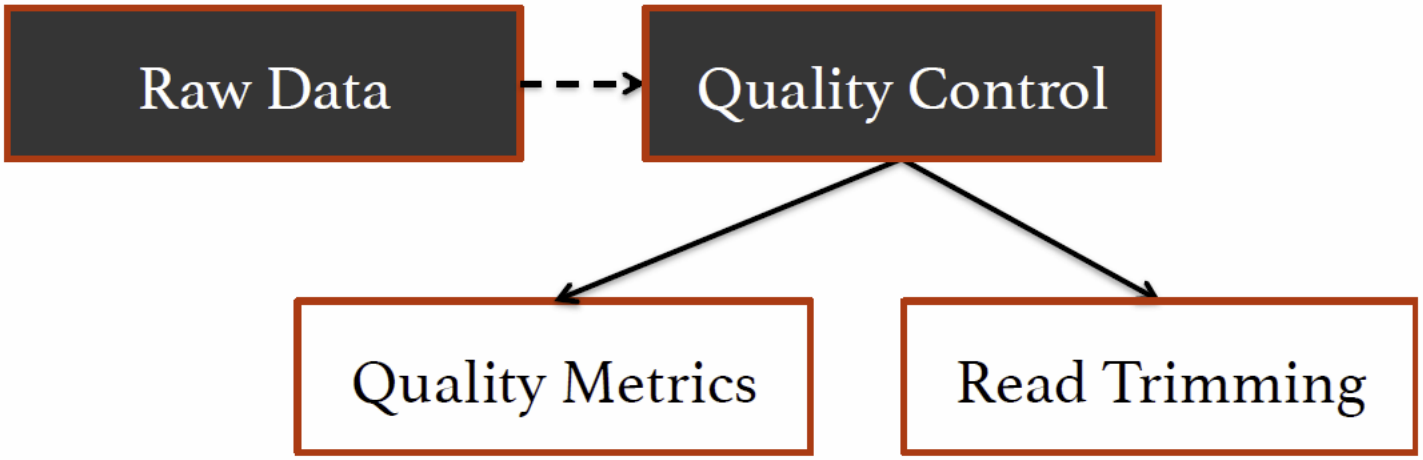
Choose local file | Choose FTP file | Paste/Fetch data | Start | Pause | Reset | Close

Plan B: GenomeSpace Importer

Step 1: Data QC



QC and Trimming off Barcodes



FastQC
RSeQC
PRINSEQ

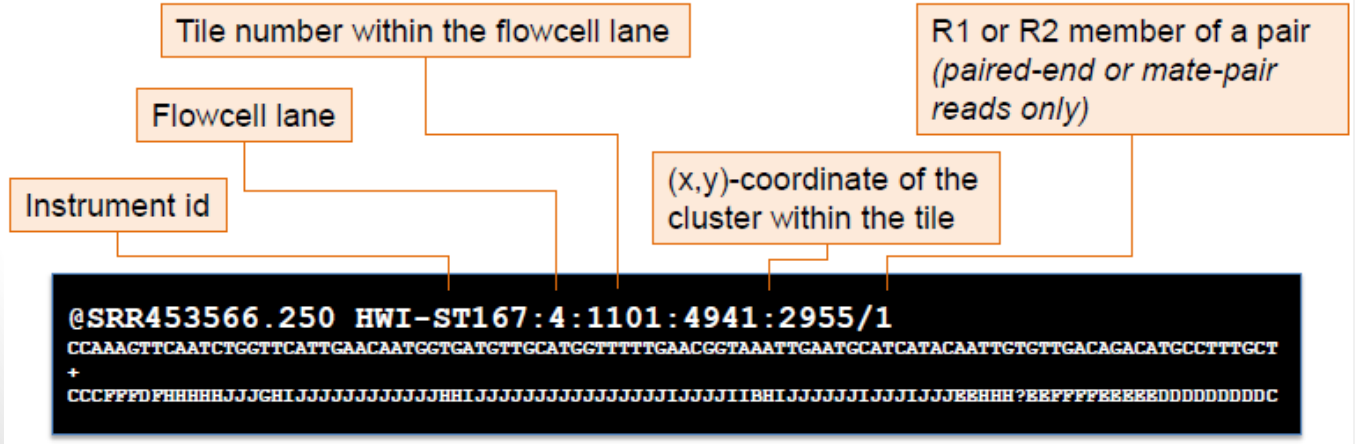
trimmomatic
SeqPrep
cutadapt
HomerTools

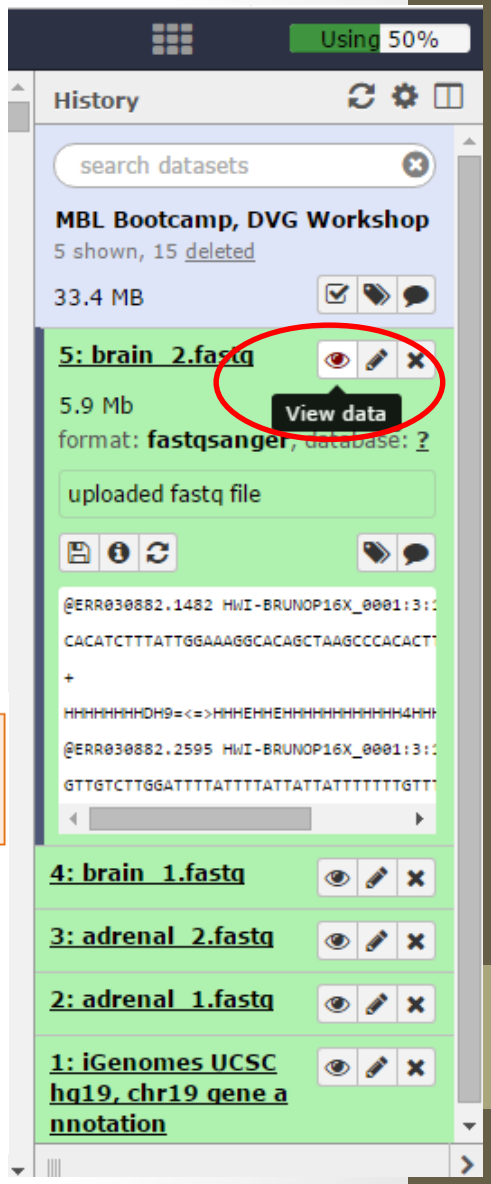
View Raw Fastq Data

Sequence

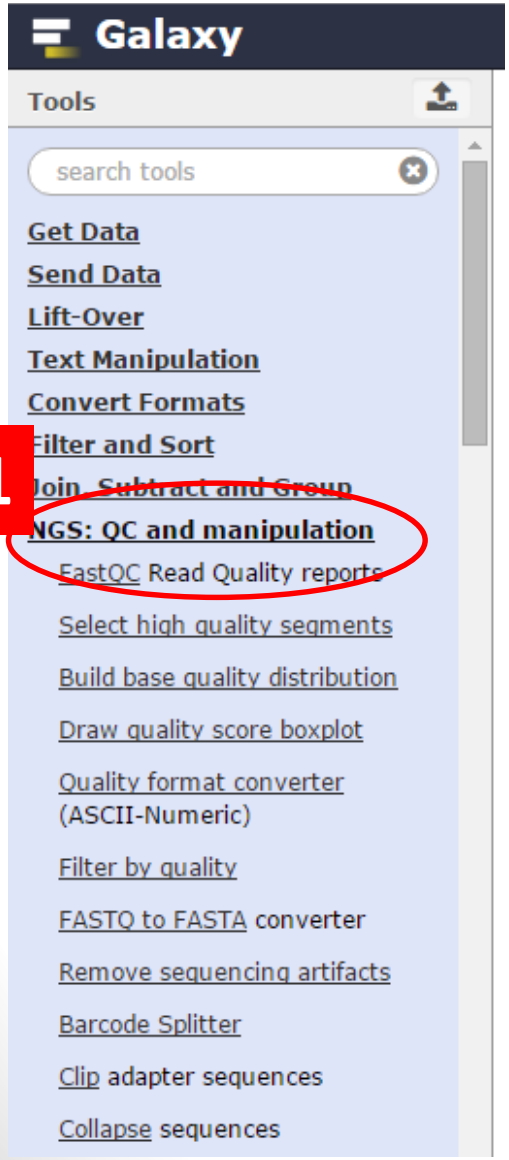
```
@M00956:35:000000000-A8P3G:1:1101:16937:1654 1:N:0:2
GCTAAACTAATGTCAACTTATTTCCTTTTTTCCTTTAGCCACCATCAAAGAA
CAGACGAGAGAGAACAGAACTGAAGCCTGATTTCTTTGATCCAGCTTCAATT
ATGGATGAATCAGTAGGTTTAATTTAAATGTACCCTAGGATTGCAAA
+
9CCCCGGGGGGGGGDGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GFGFGGGGGGCEGGGGGGGGGGGGGGGGGGGGGGGGGCFGGFGGGGGGFFF
FGFFCFEGGGGGGGFGGGGGFFGGGGGGDGGGGGGGGFGFFFCFGFF
```

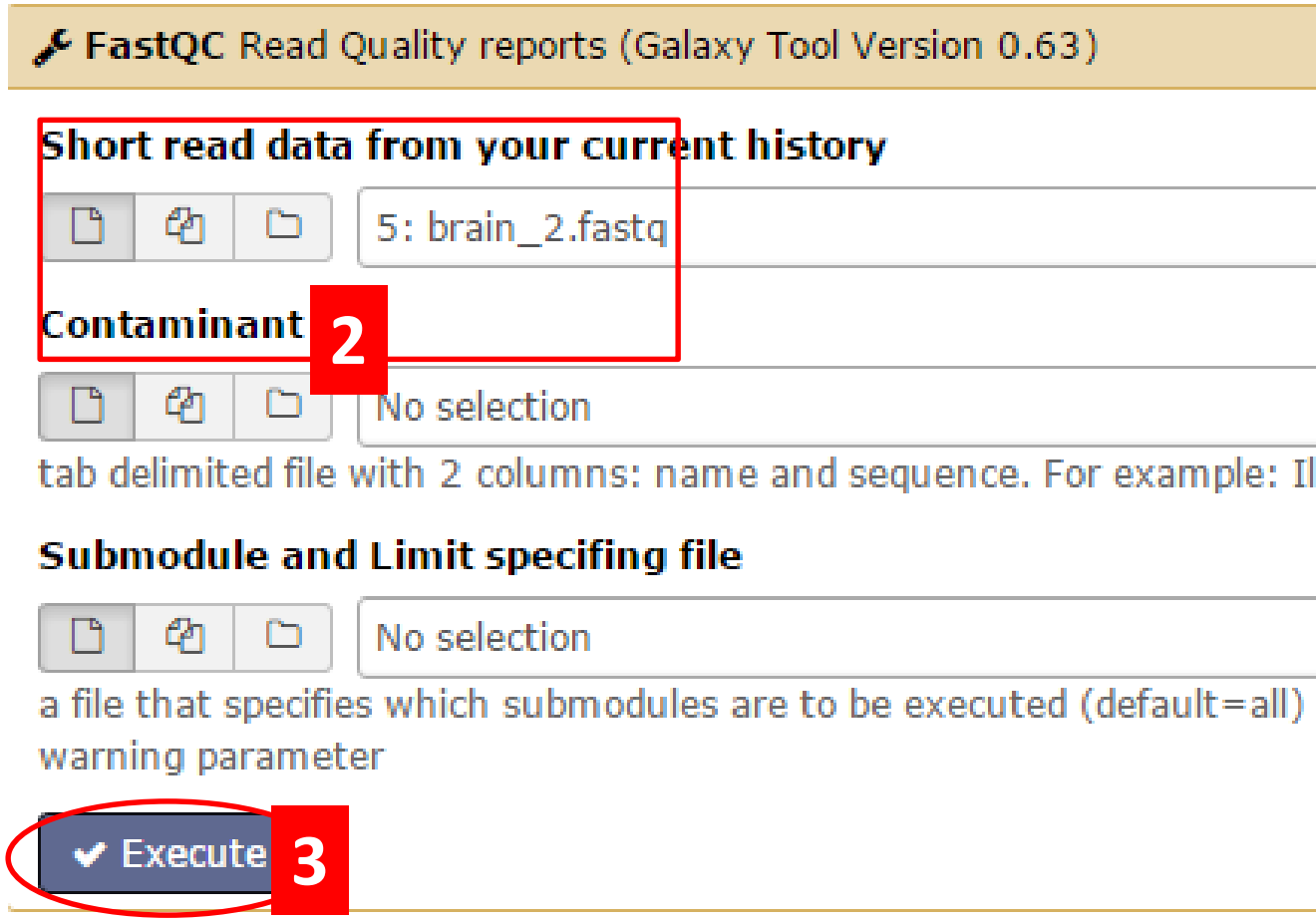
Quality Score





FastQC – Quality Metrics





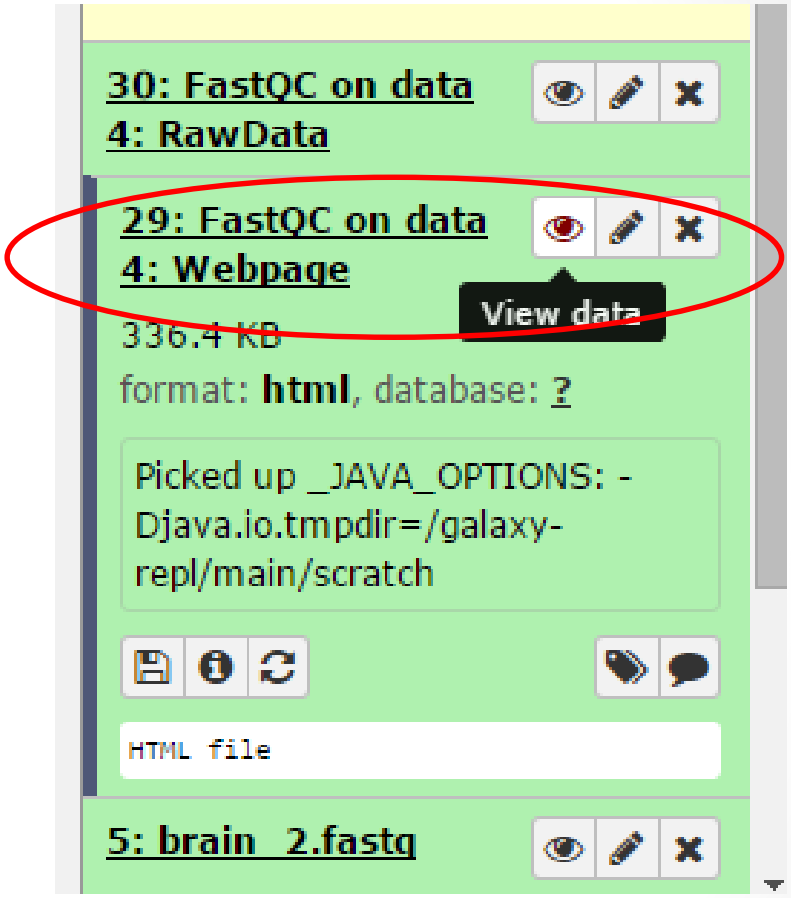
Analyze Data Workflow S

brain_1.fastq FastQC Report

FastQC Report
Mon 10 Aug 2015
brain_1.fastq

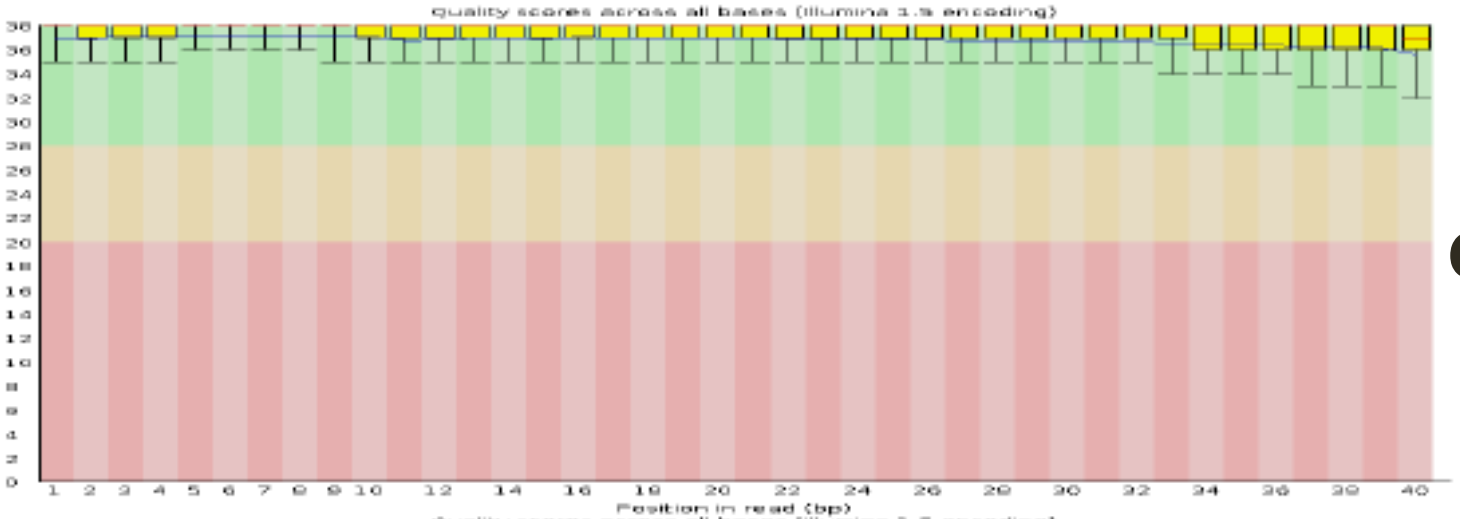
Summary

- 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

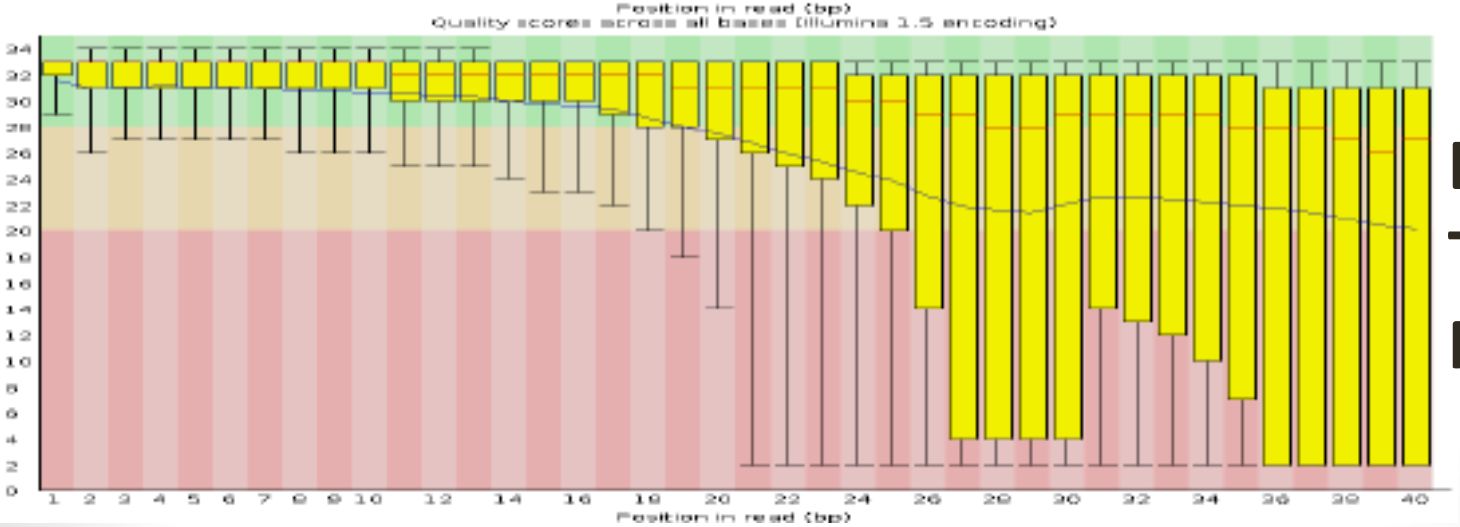


* Note Data X Reference
* Delete “RawData” Files

Per base sequence quality



Good!



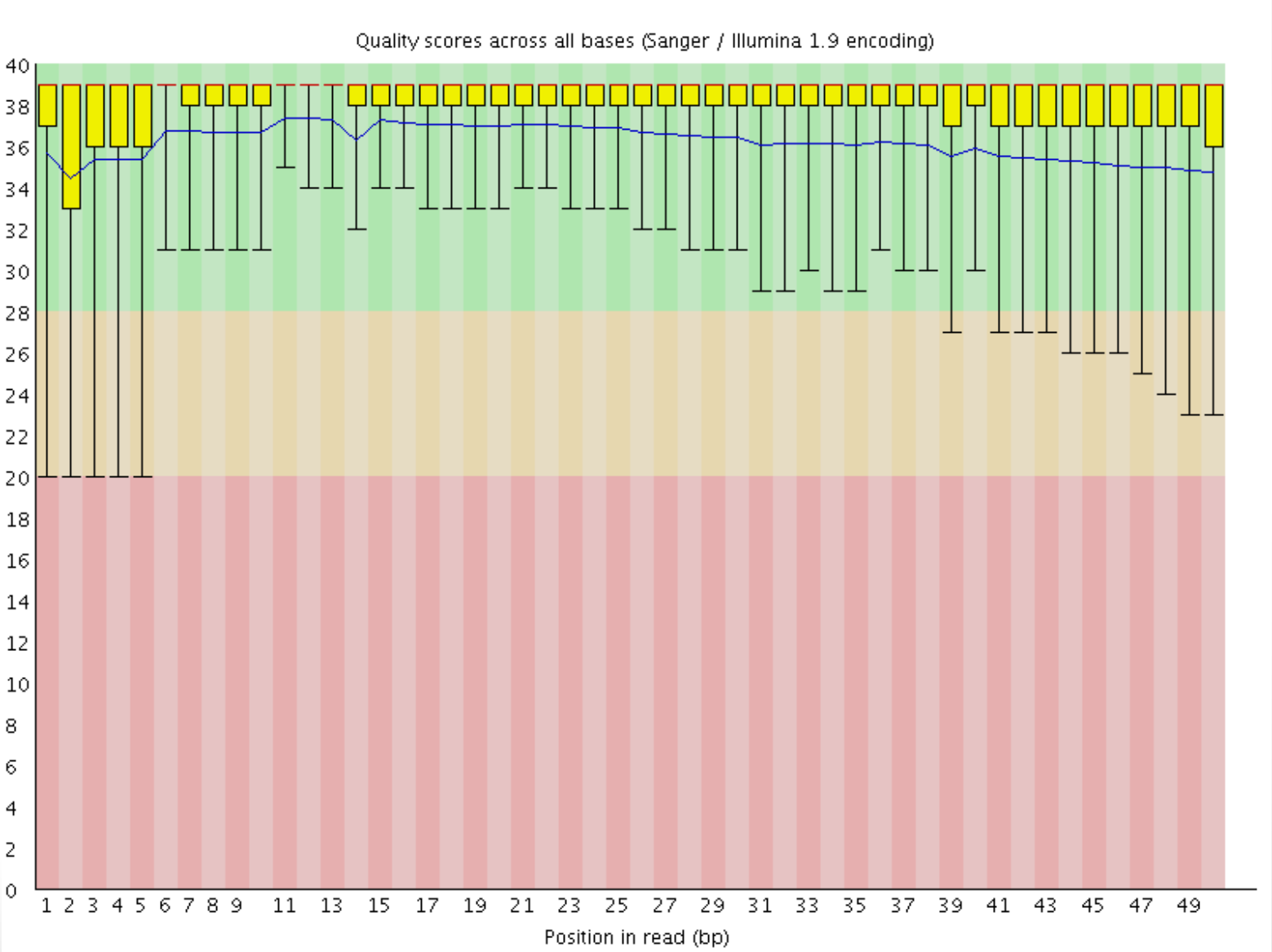
Bad!
Trimming
Needed

FastQC Analyses

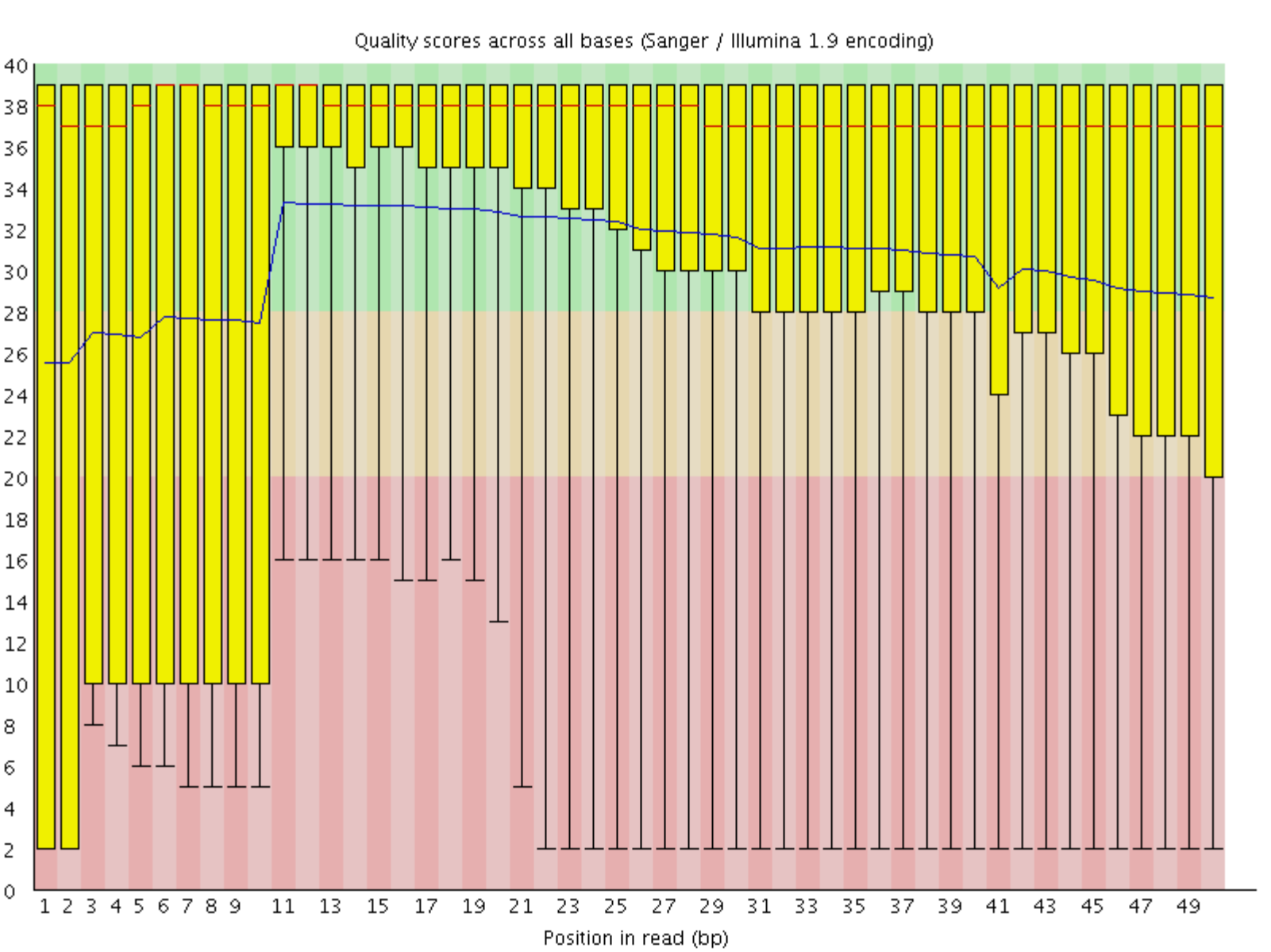
- Assume a median quality score of below 20 to be unusable (yellow bars).
- Given this criterion, is trimming needed for any of the datasets? If so, which base pairs should be trimmed?

- 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

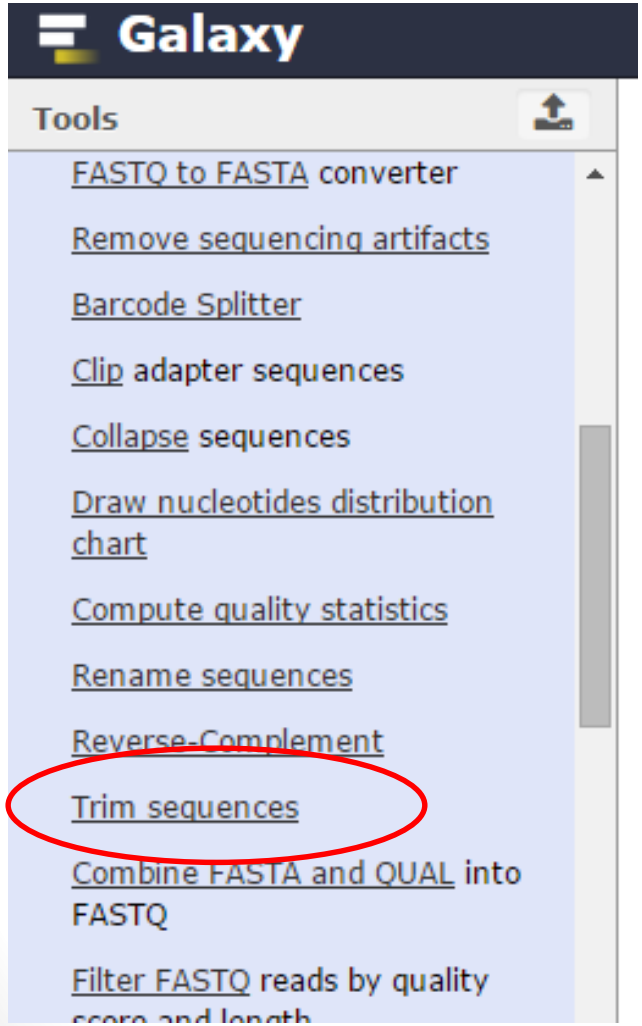

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Position in read (bp)

Brain 1


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Position in read (bp)

Trim Sequences

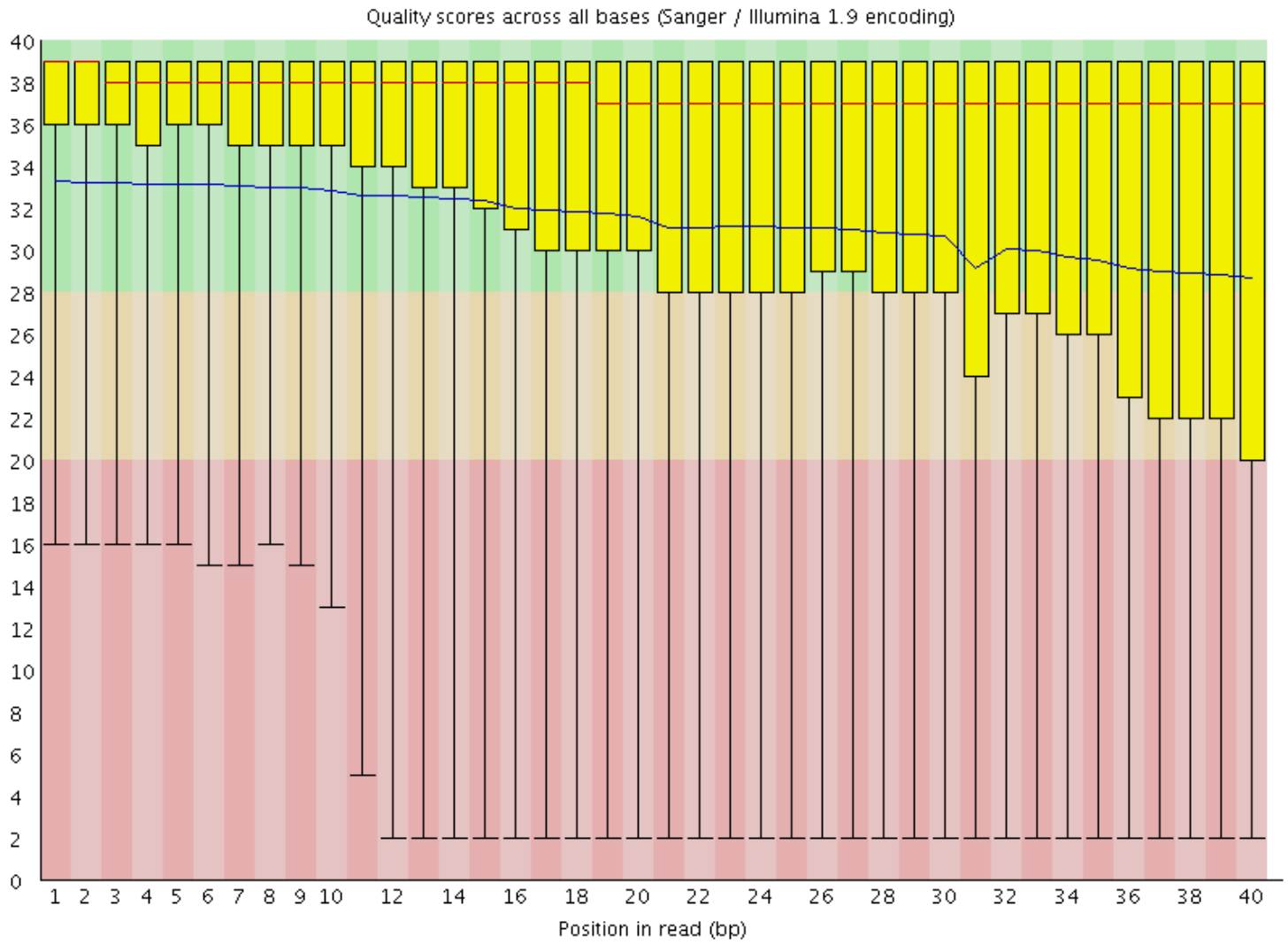
Brain 2



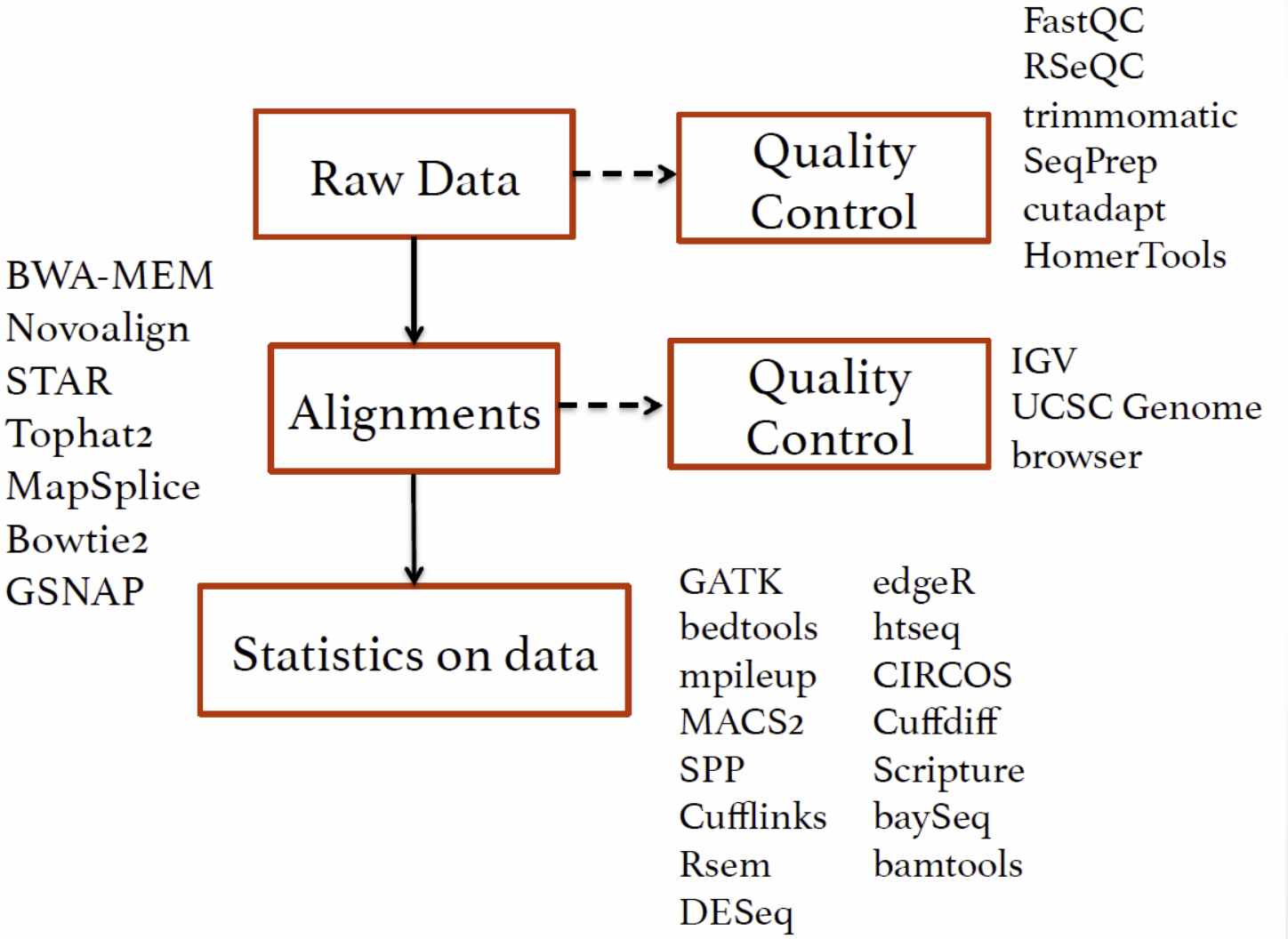
Now run FastQC again on the Trimmed Sequence and see what it looks like…

How does "Per base sequence quality" graph look now?

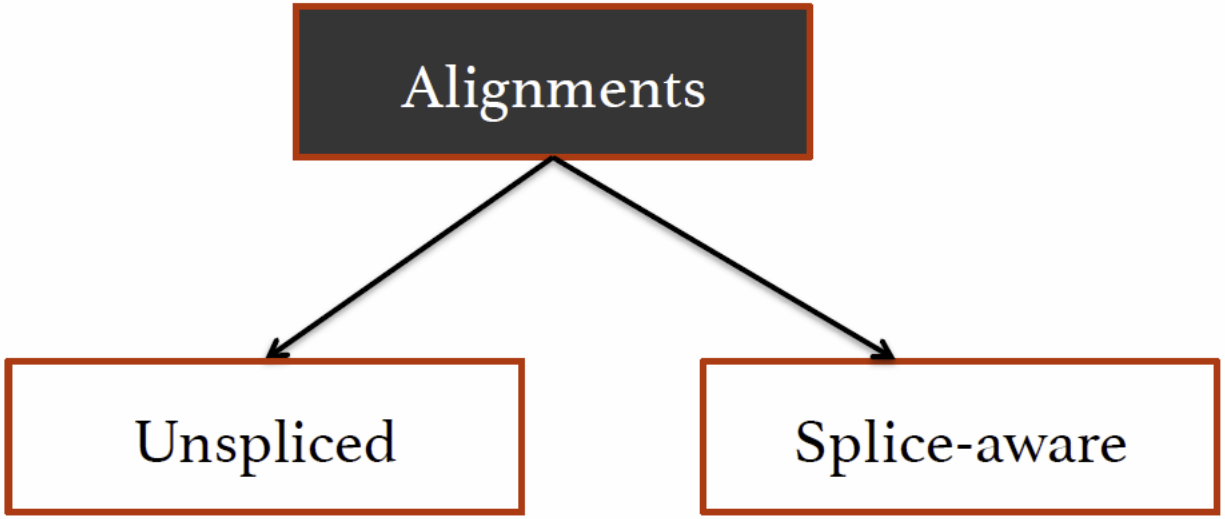
QC of Brain 1 Trimmed



Step 2: Aligning to Reference Genome



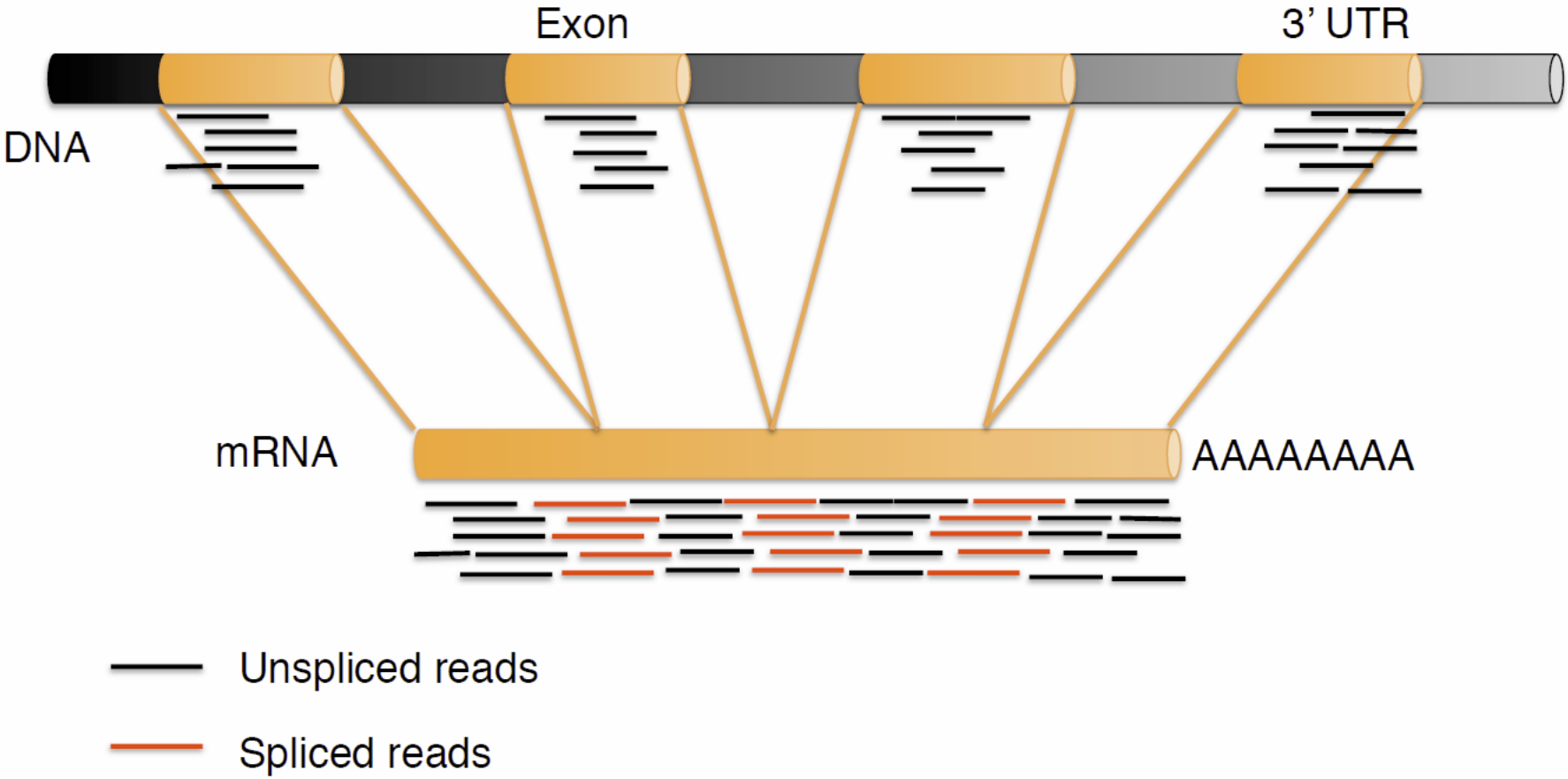
Mapping: Aligning Reads to Genes



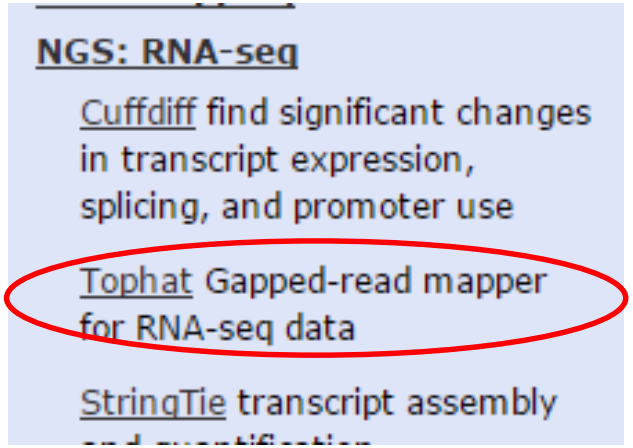
BWA-MEM
Novoalign
Bowtie2

STAR
Tophat2
MapSplice
GSNAP

Aligning Reads to a Reference Genome (Hg19)

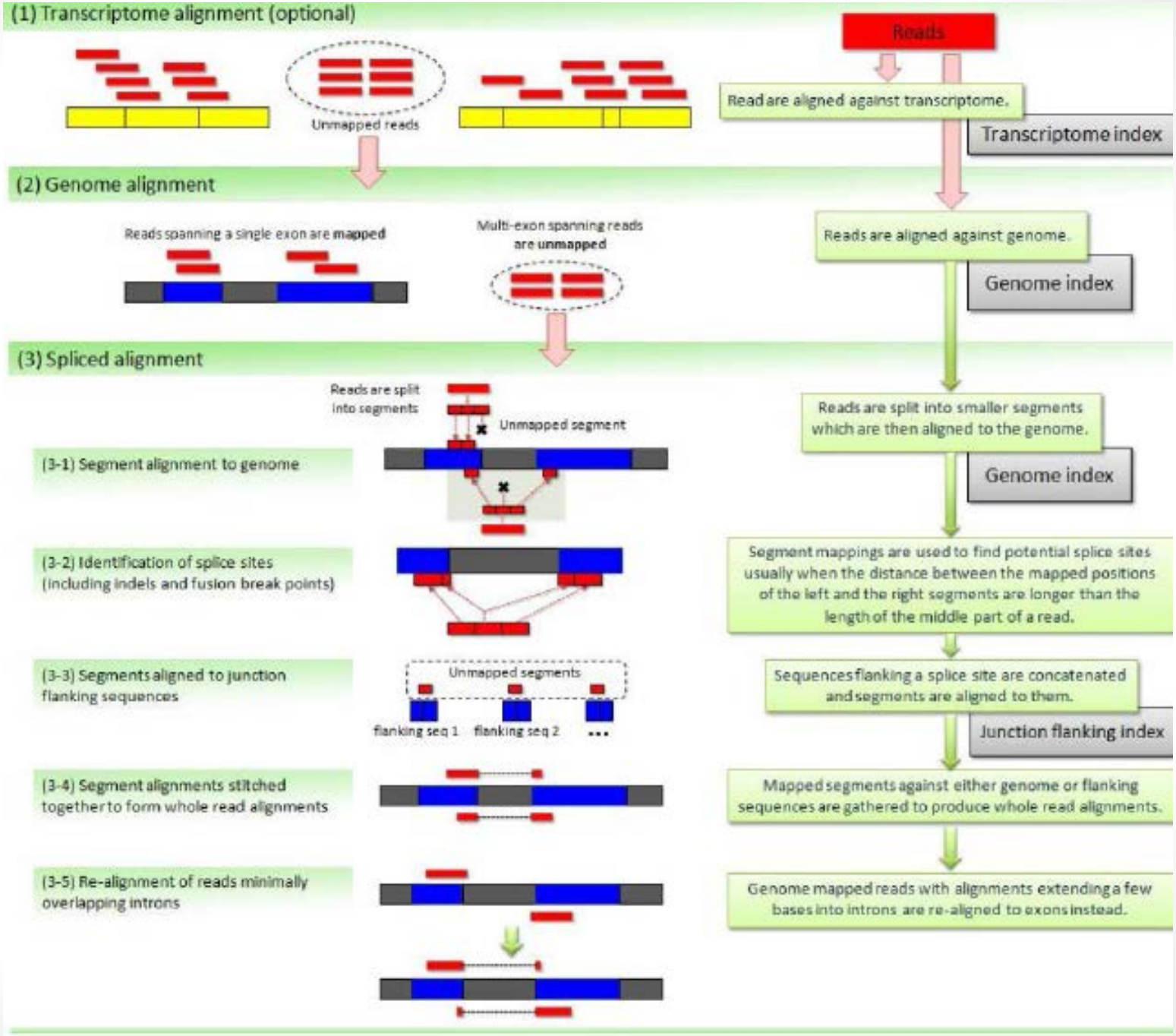


NGS: RNA-seq - TopHat



- 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)
Unmapped reads
Reads
Read are aligned against transcriptome.
Transcriptome index
(2) Genome alignment
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
Reads are aligned against genome.
Genome index
(3) Spliced alignment
Reads are split into segments
Unmapped segment
(3-1) Segment alignment to genome
Reads are split into smaller segments which are then aligned to the genome.
Genome index
(3-2) Identification of splice sites (including indels and fusion break points)
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.
(3-3) Segments aligned to junction flanking sequences
Unmapped segments
flanking seq 1
flanking seq 2
Sequences flanking a splice site are concatenated and segments are aligned to them.
Junction flanking index
(3-4) Segment alignments stitched together to form whole read alignments
Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.
(3-5) Re-alignment of reads minimally overlapping introns
Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.

TopHat Output: BAM Format

SAM/BAM Format

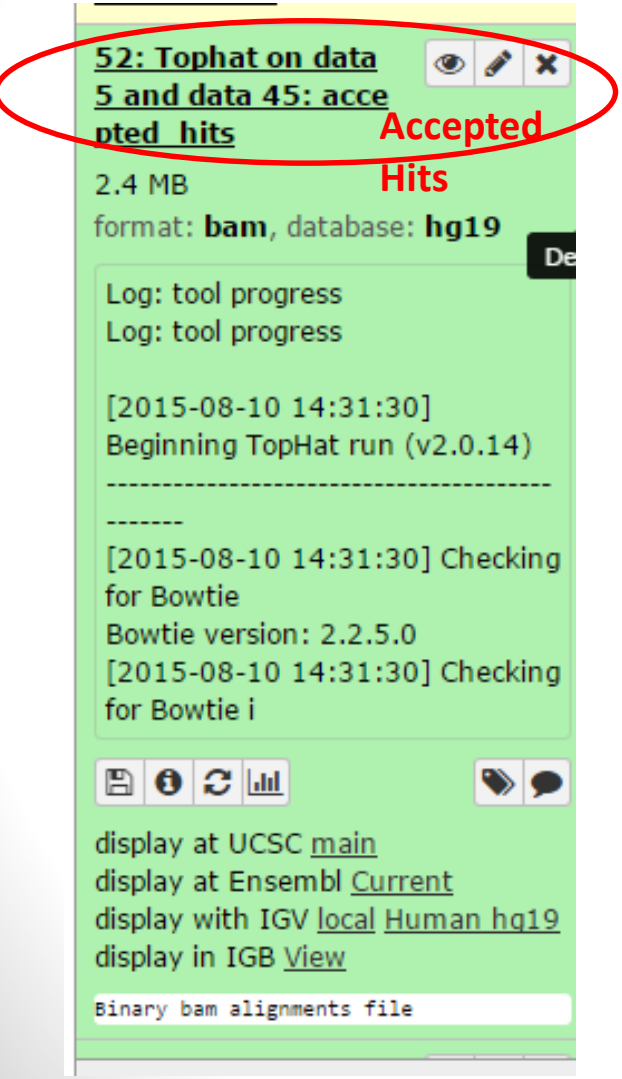
1	2	3	4	5	6	7	8	9	10	11	12
R001	83	ref	37	30	9M	=	7	-39	CAGCGCAT	CAGCGCAT	TAG

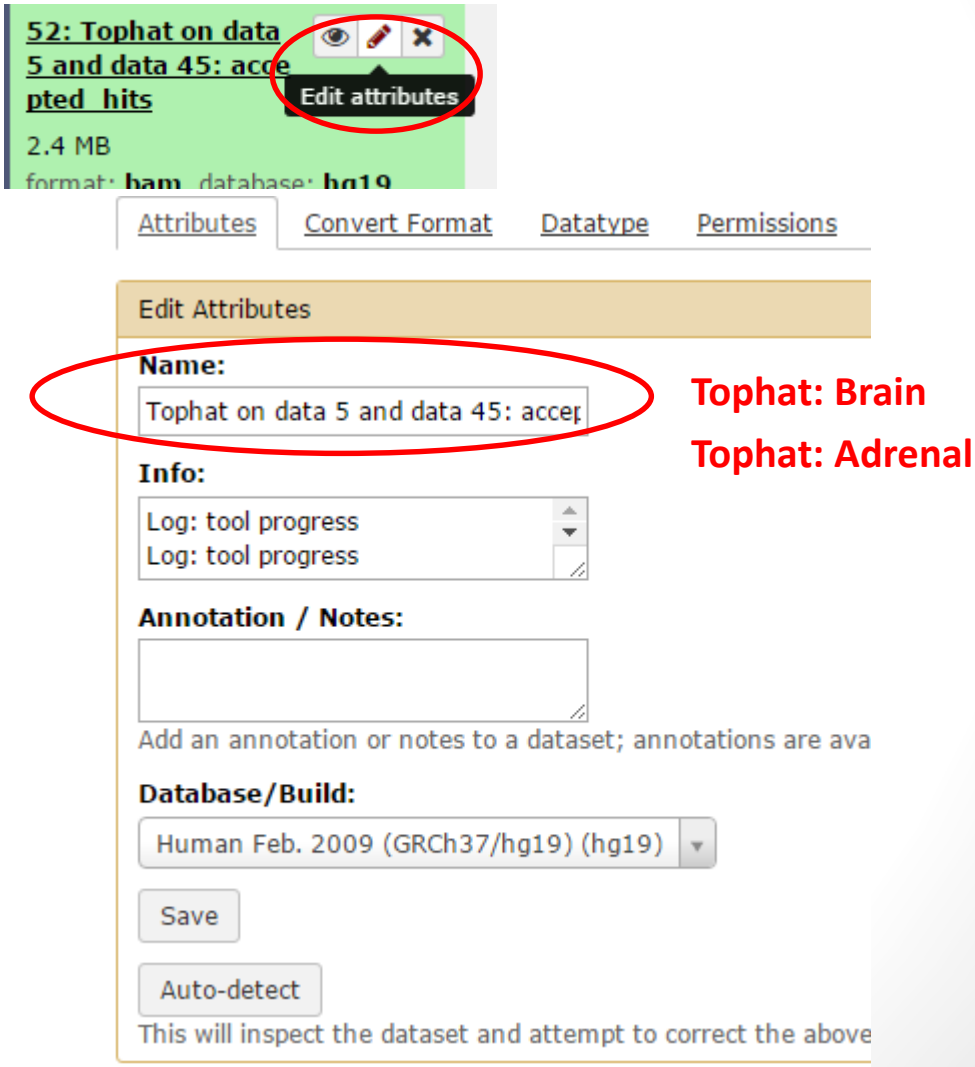
COLUMNS:

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

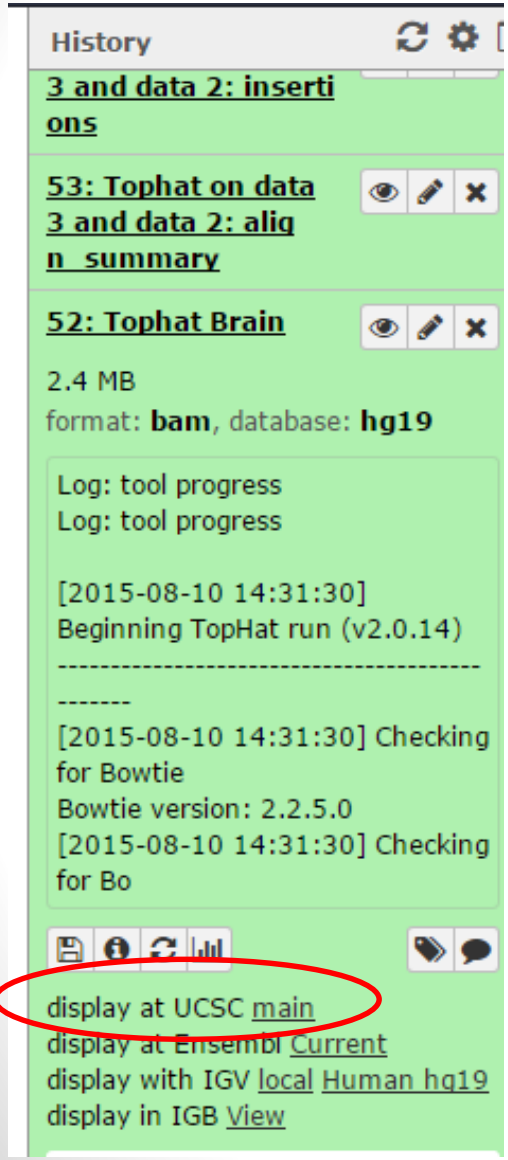
Keeping Organized

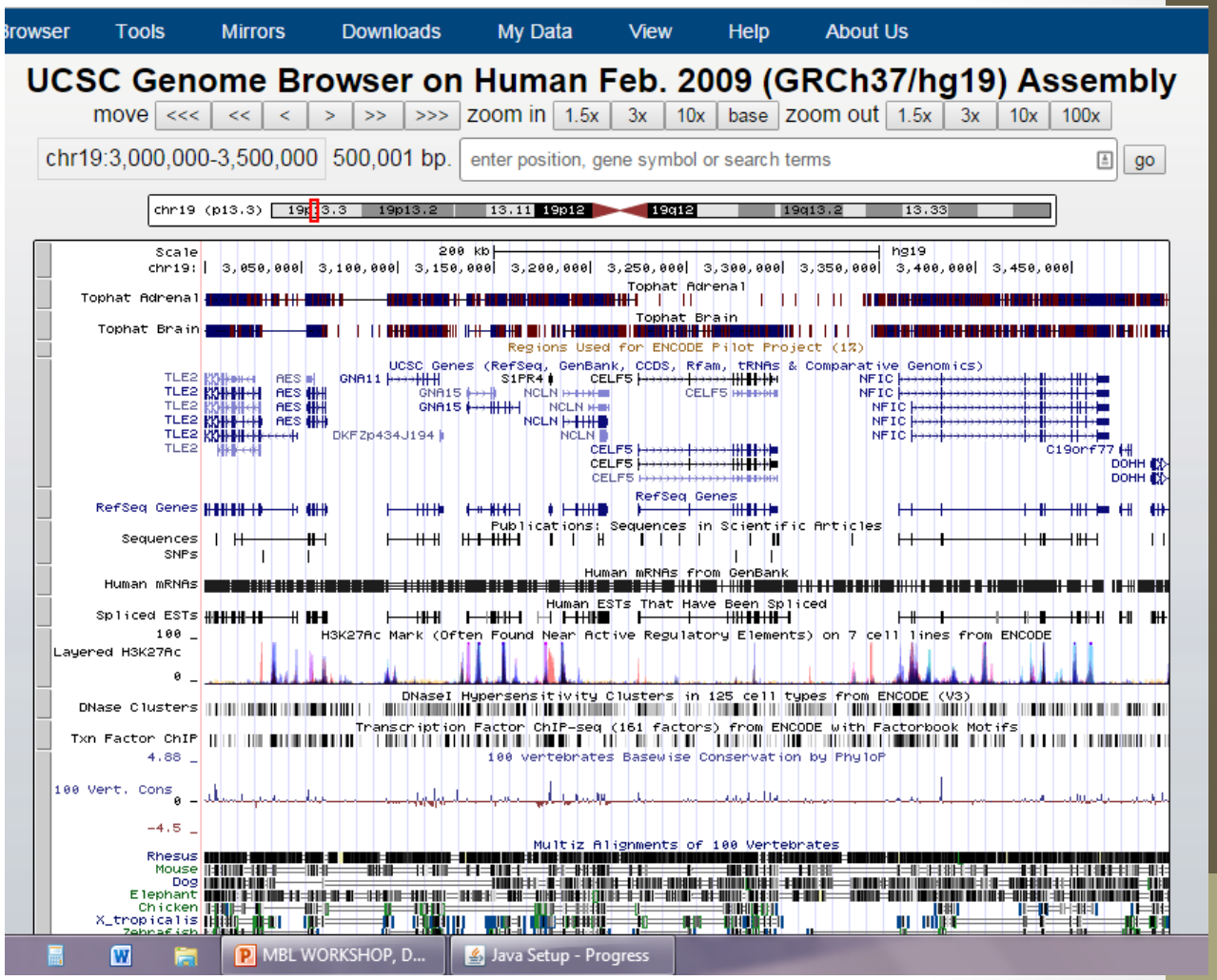




World's Most Accurate Pie Chart

Visualizing Data

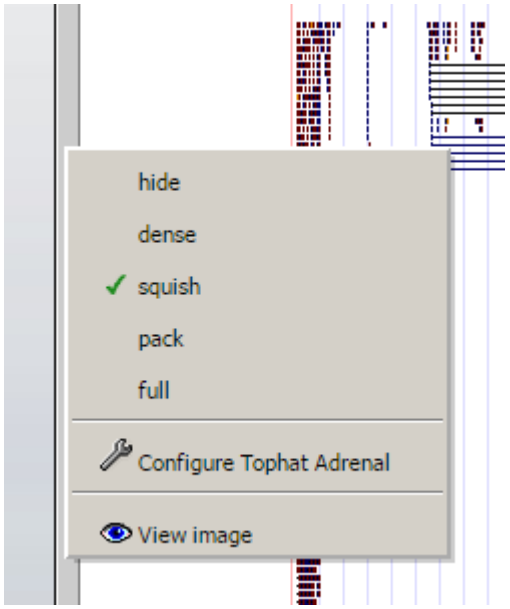




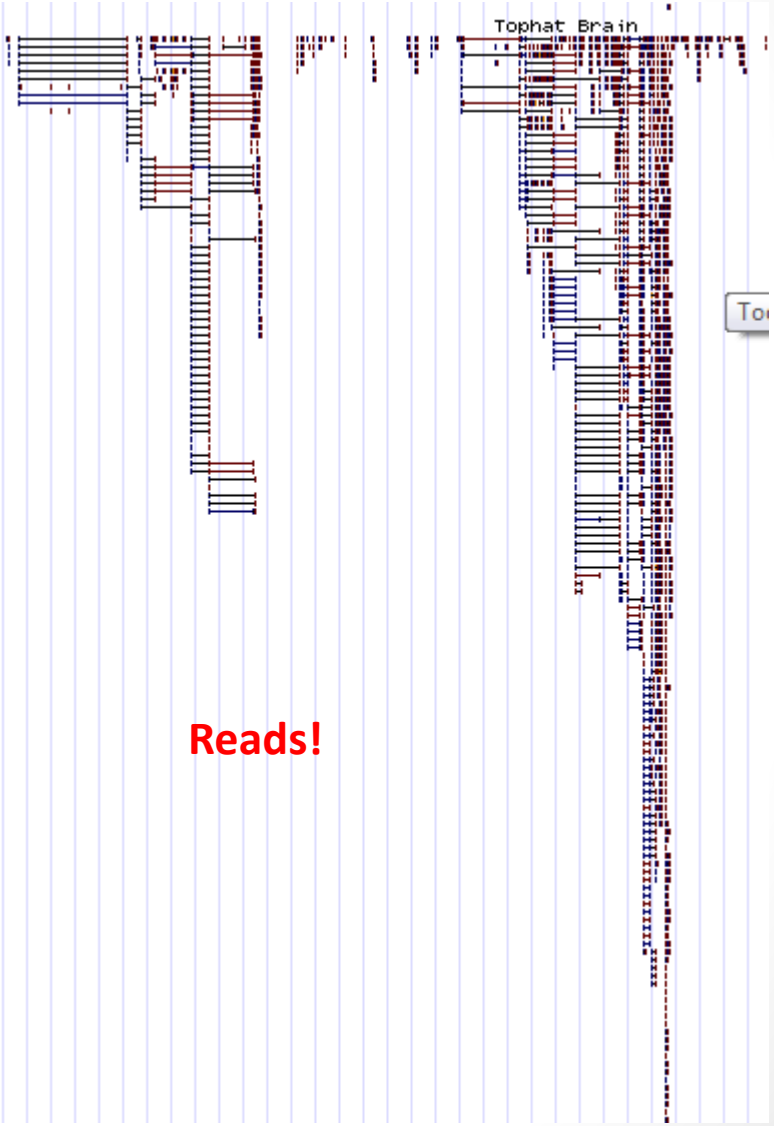
Search: chr19:3,000,000-3,500,000

UCSC Genome Browser

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

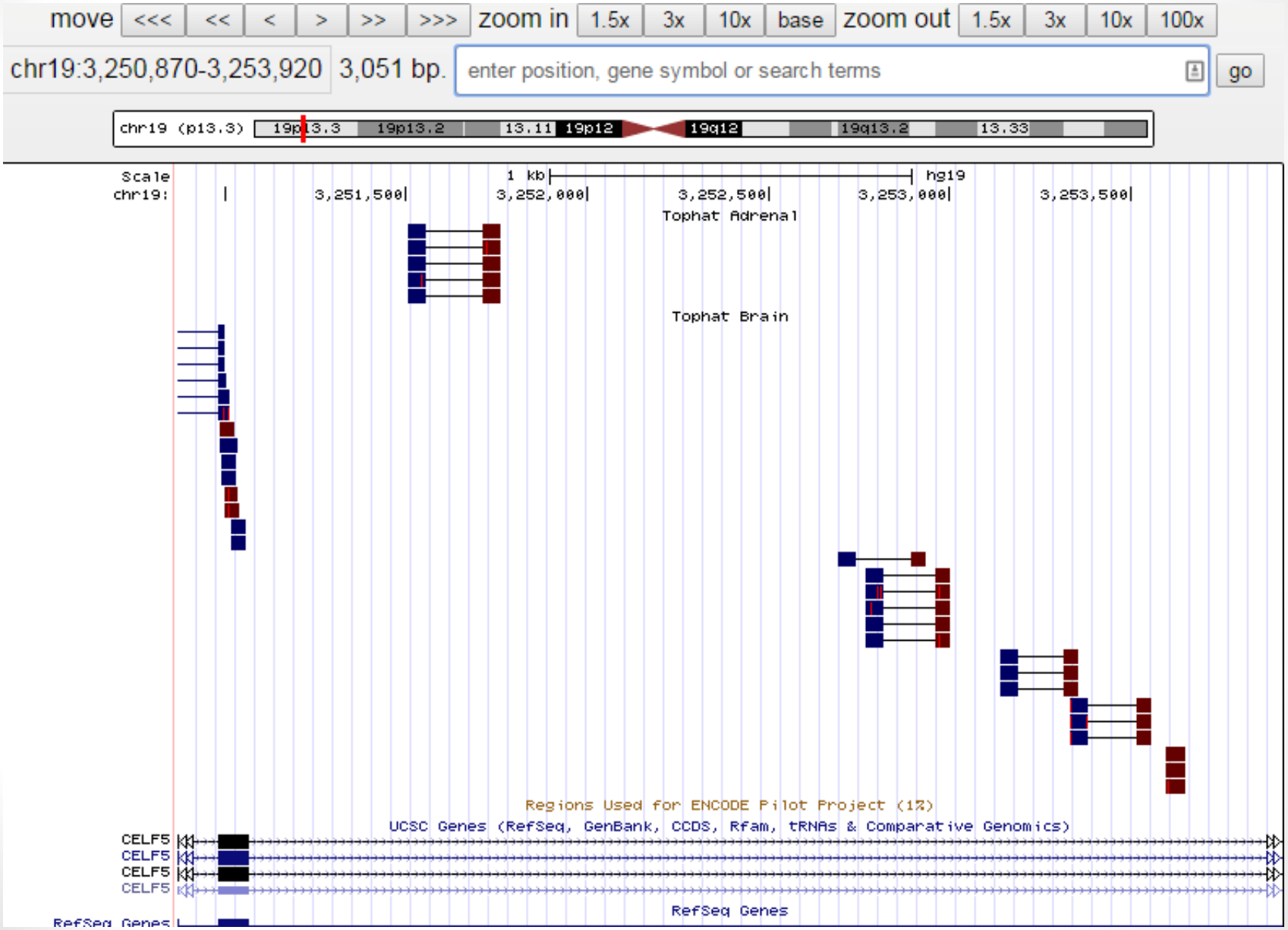


Right Click, Select “squish” view

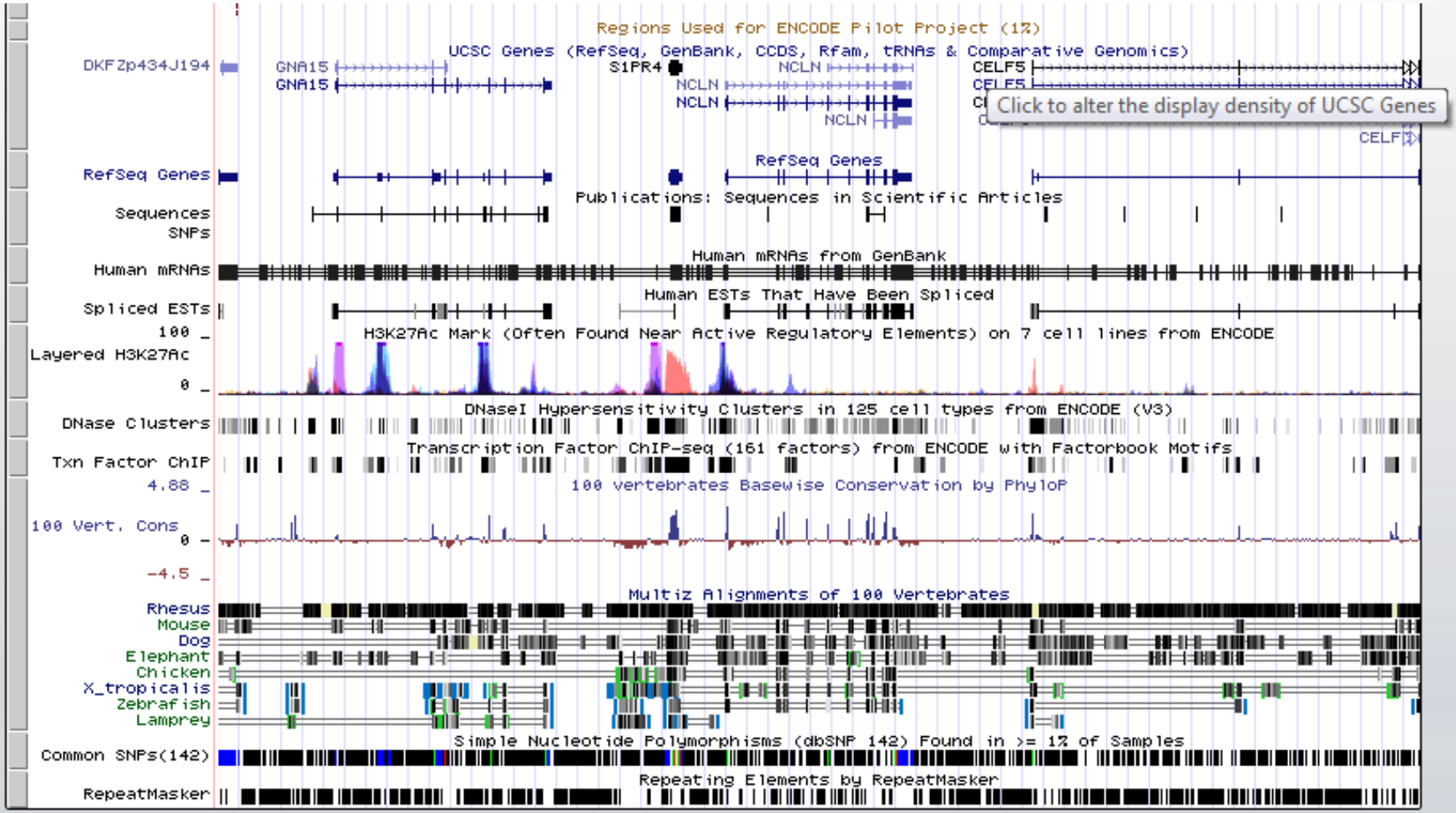


Reads!

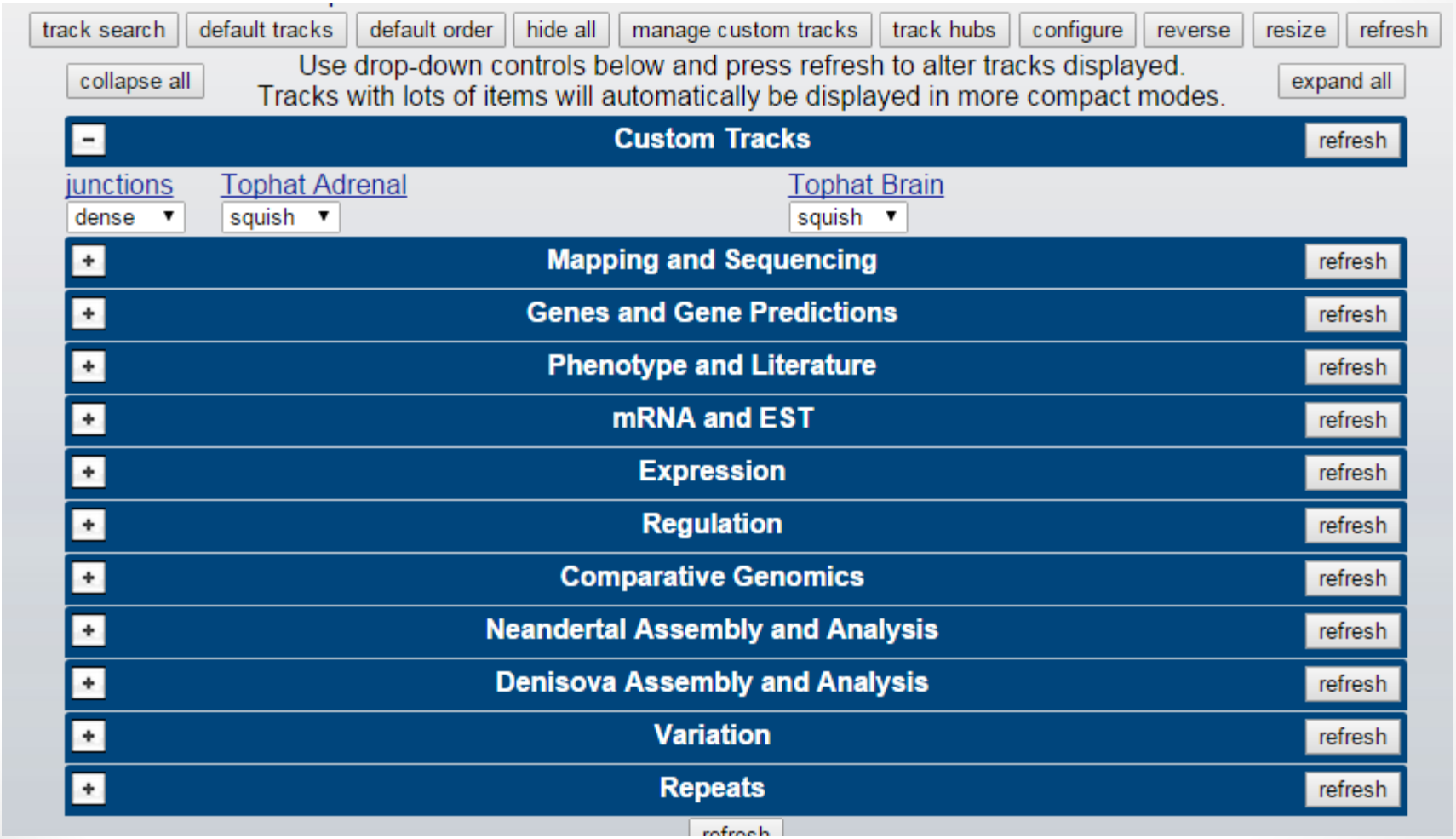
Narrow Down - Search for: chr19:3,250,772-3,253,822



UCSC Browser Comparisons



Lots and Lots of Comparisons…

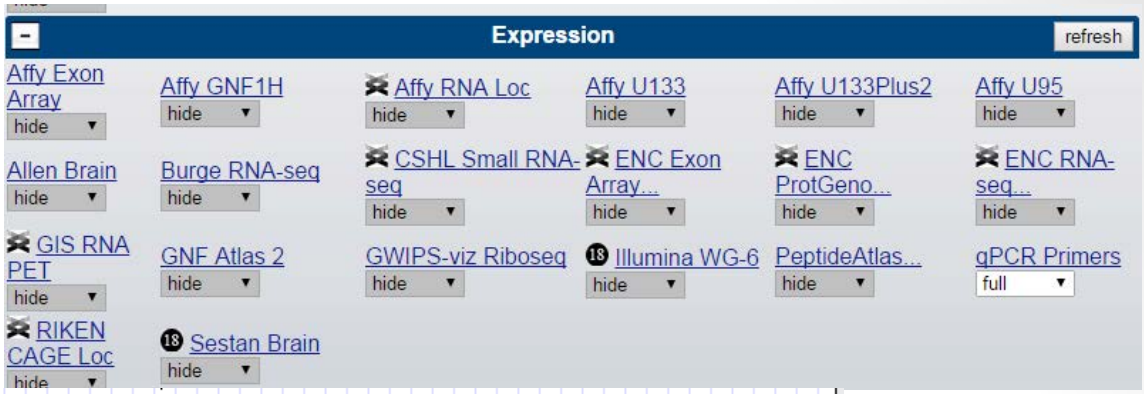


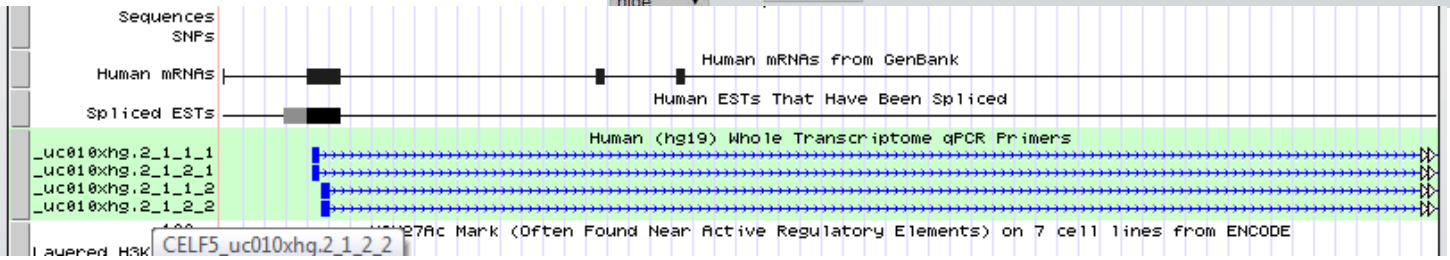
Challenges

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?

qPCR Primers





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

PRIMER PAIR NAME: CELF5_uc010xhg.2_1_1_1

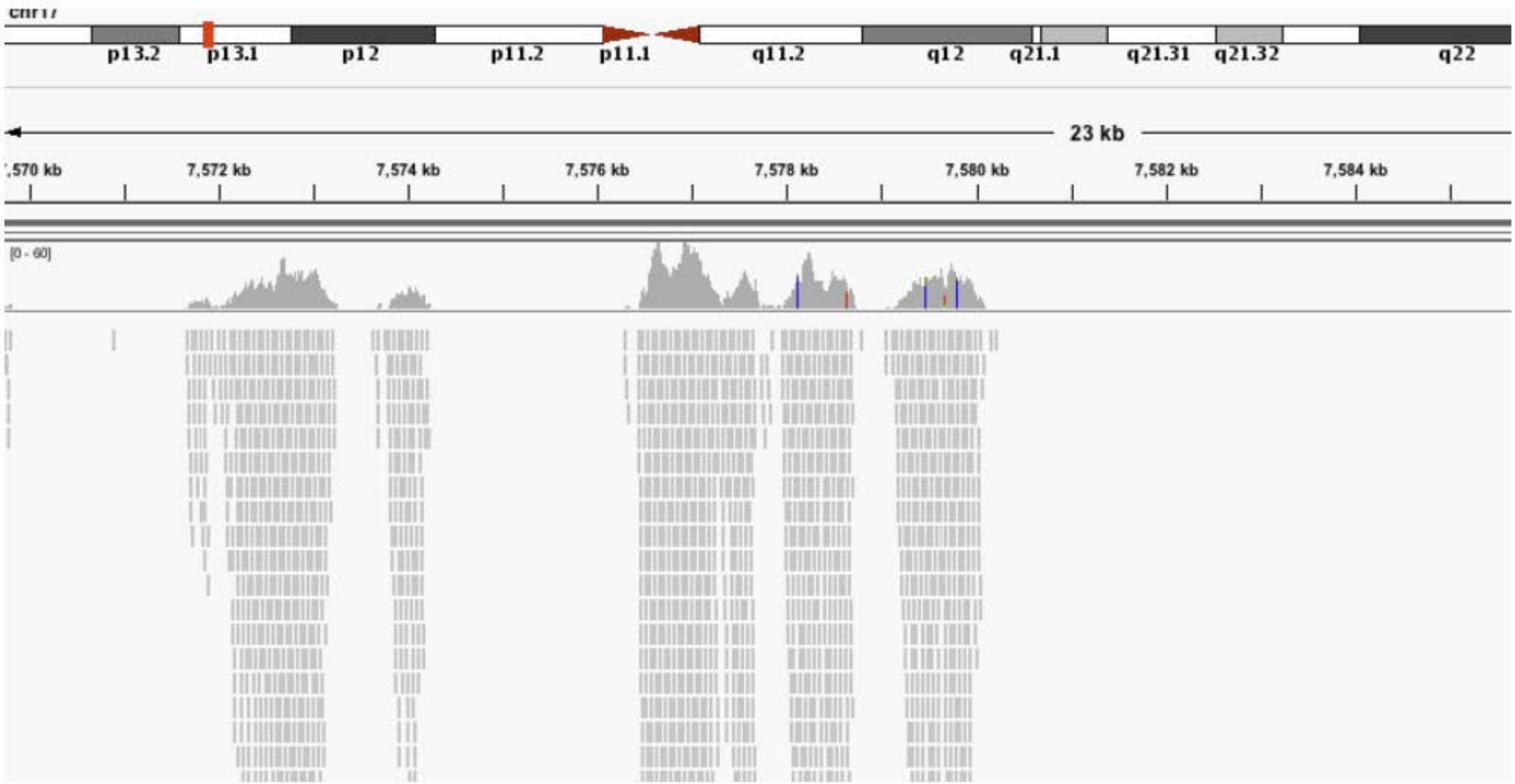
OLIGO	len	tm	gc%	any	3'	seq
FORWARD	20	60.533	55.000	4.00	2.00	TCACCTACTGTGCCAGGGAT
REVERSE	18	60.557	61.111	3.00	2.00	GCTTTCACTGTCCGCAGG

PAIR ANY COMPL:	PAIR 3' COMPL:	SNP_flag	RNA_edit_flag	number_specific_products	PRODUCT SIZE
4.00	3.00	0	0	1	113

Link to UCSC in-silico PCR

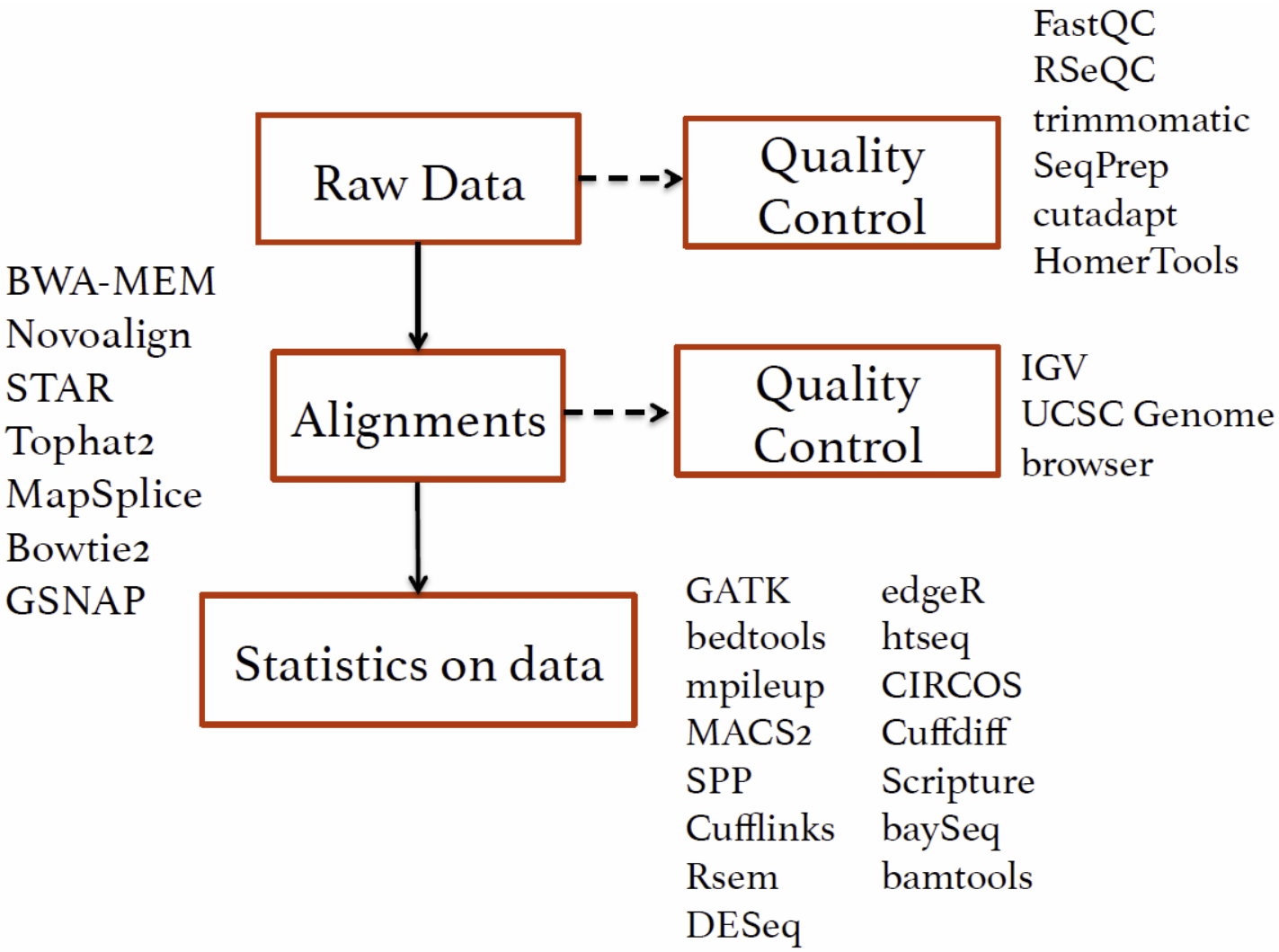
Data Visualization – IGV

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

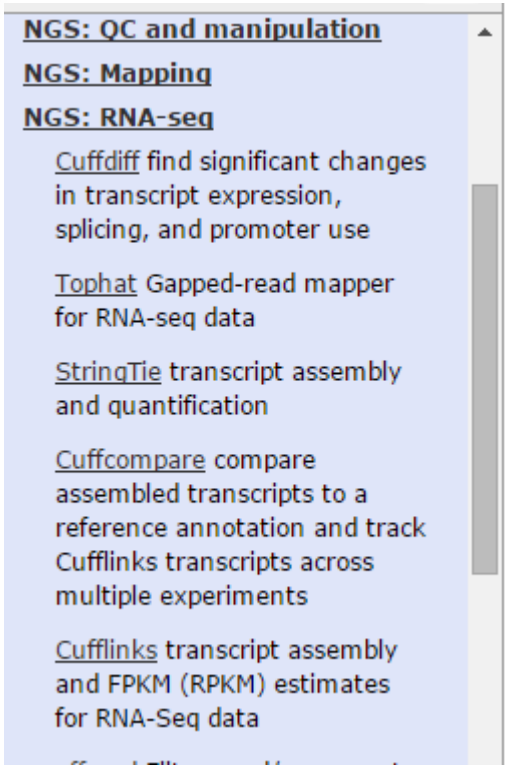


The 2008 World Submarine Racing Championships

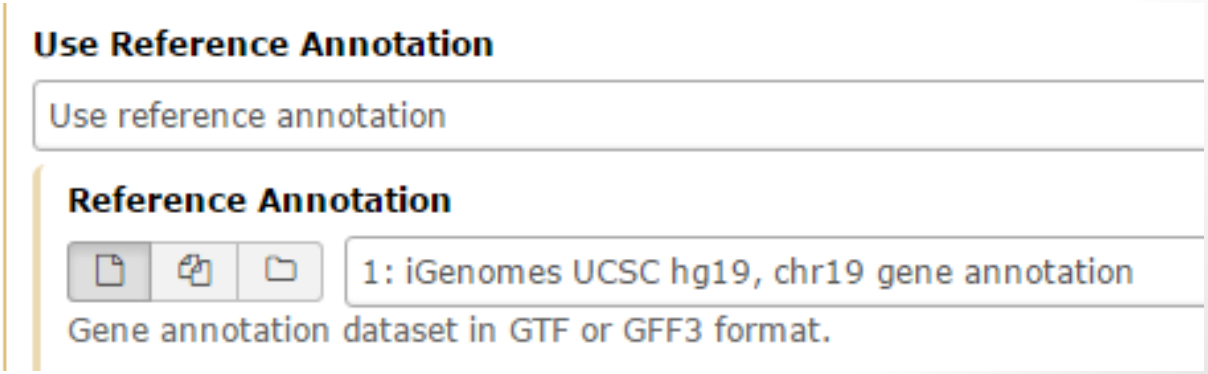
Step 3: Assembling Reads into Full-Length Transcripts



Cufflinks: Read Assembly



- 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”



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:

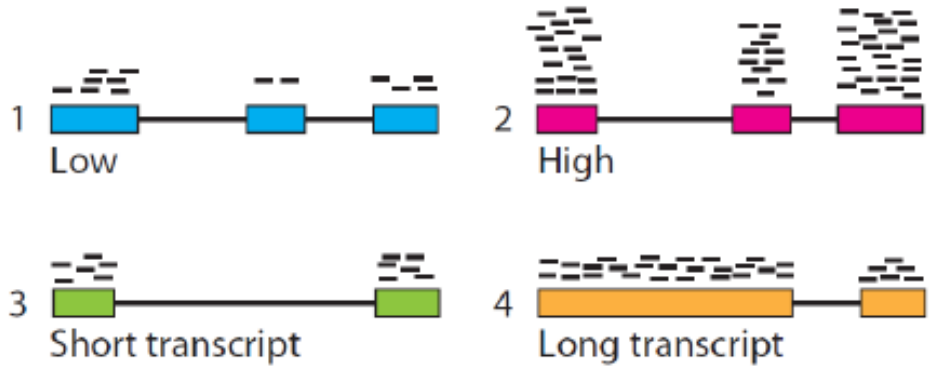
- num-threads
- GTF
- GTF-guide (assemble novel isoforms)
- library-type
- max-intron-length
- min-intron-length
- Many more!

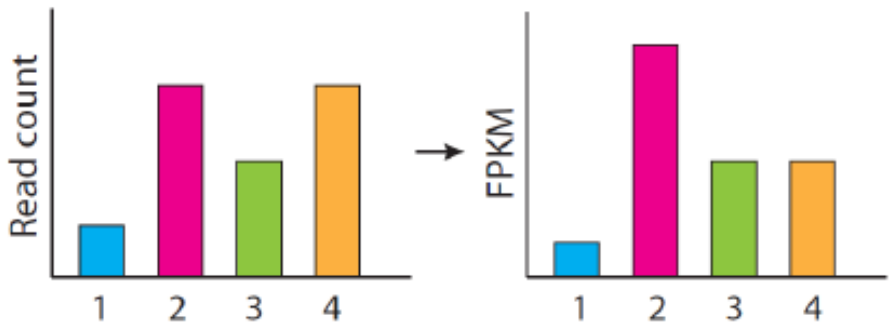
FPKM

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

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

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





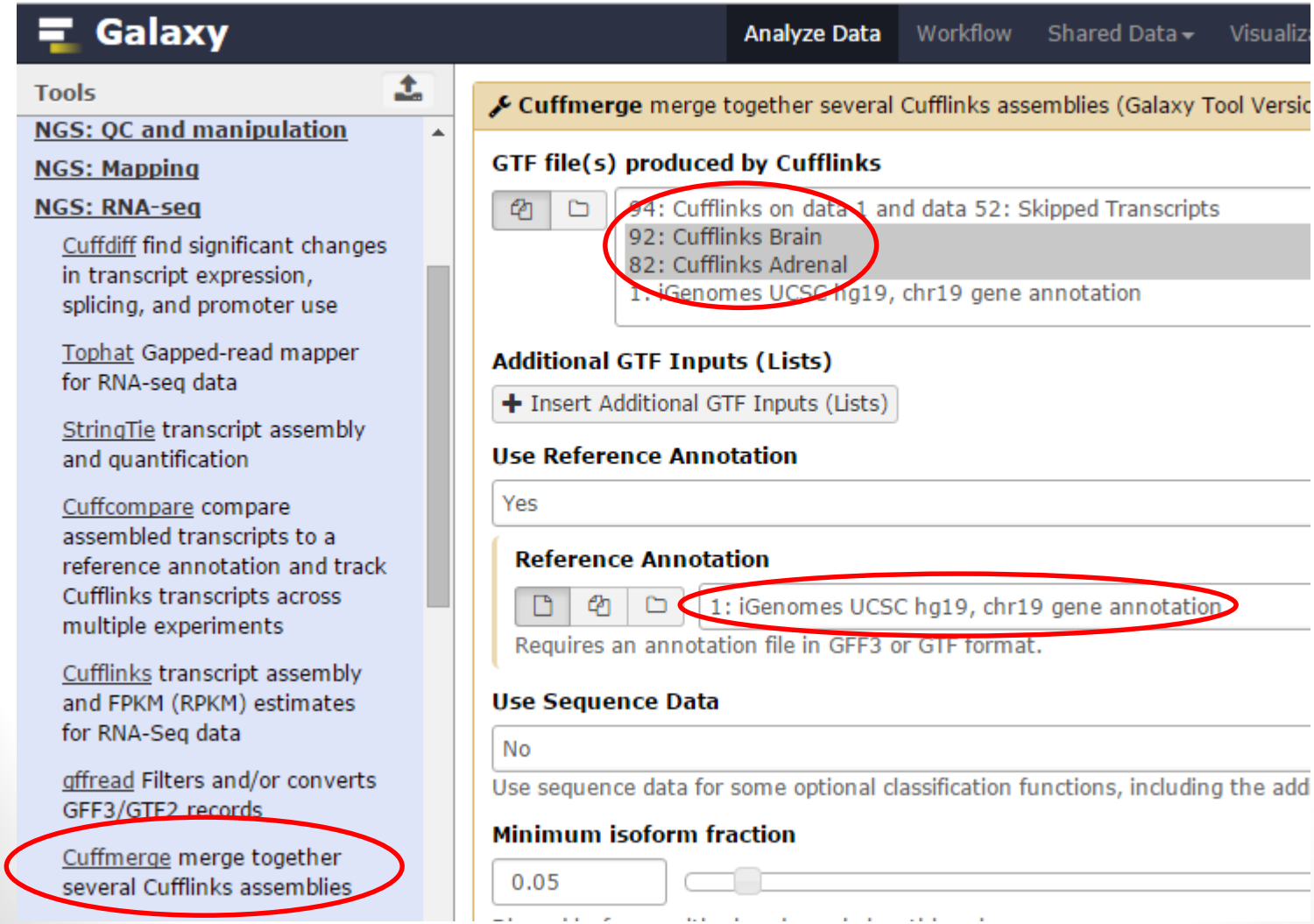
http://www.broadinstitute.org

Number of fragments normalized by

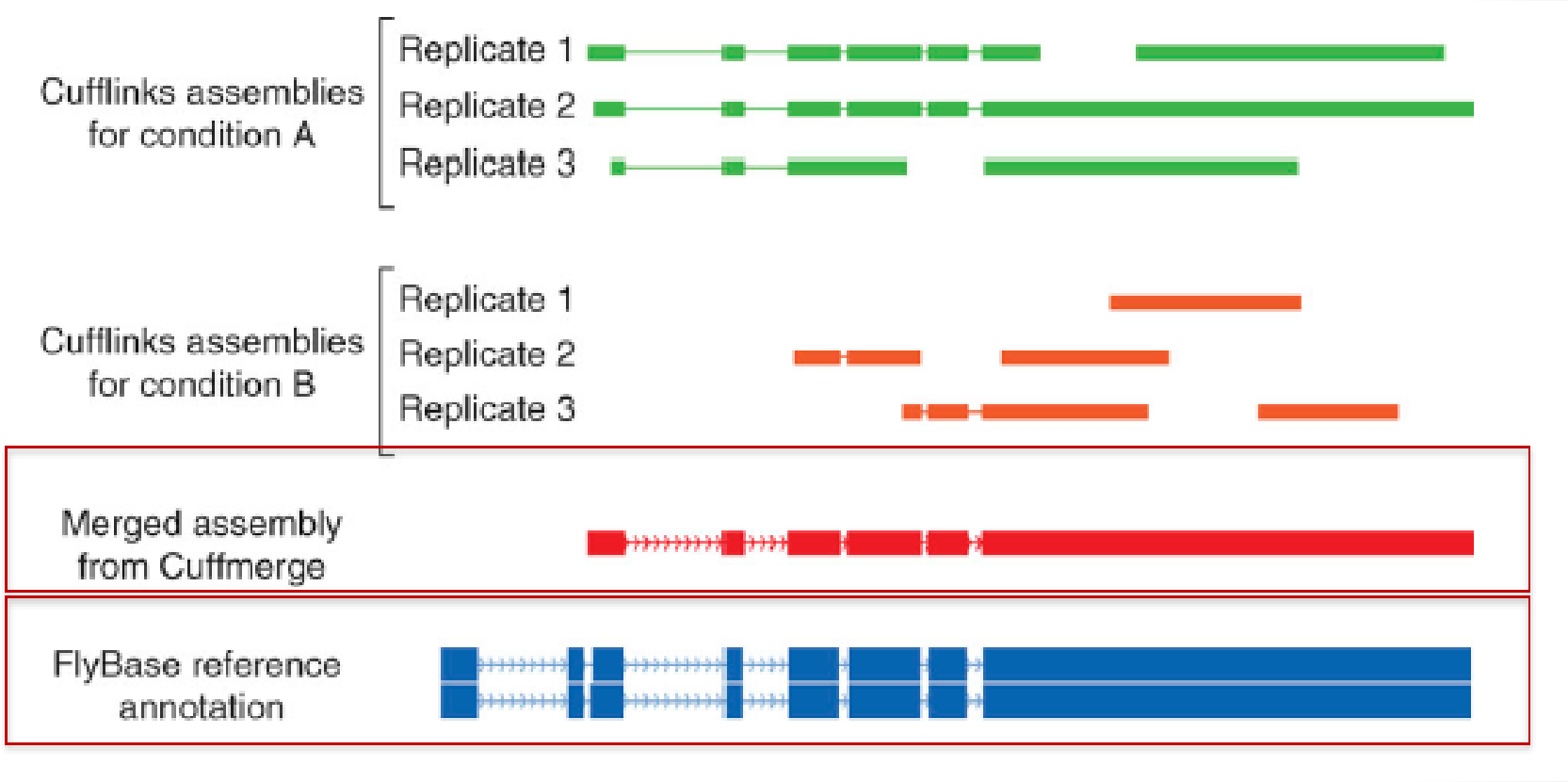
- ✓ Transcript length (Kb)
- ✓ Total number of mapped reads (Million)

Cuffmerge

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



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.

Trapnell Nature 2012

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.

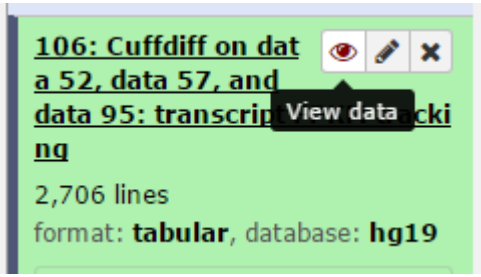
There are multiple FPKM file types:

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

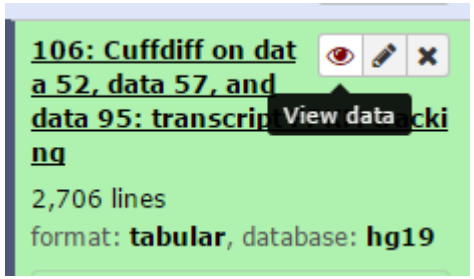
http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html#cuffdiff-output-files

Cuffdiff: Output Styles

Transcript FPKM Tracking



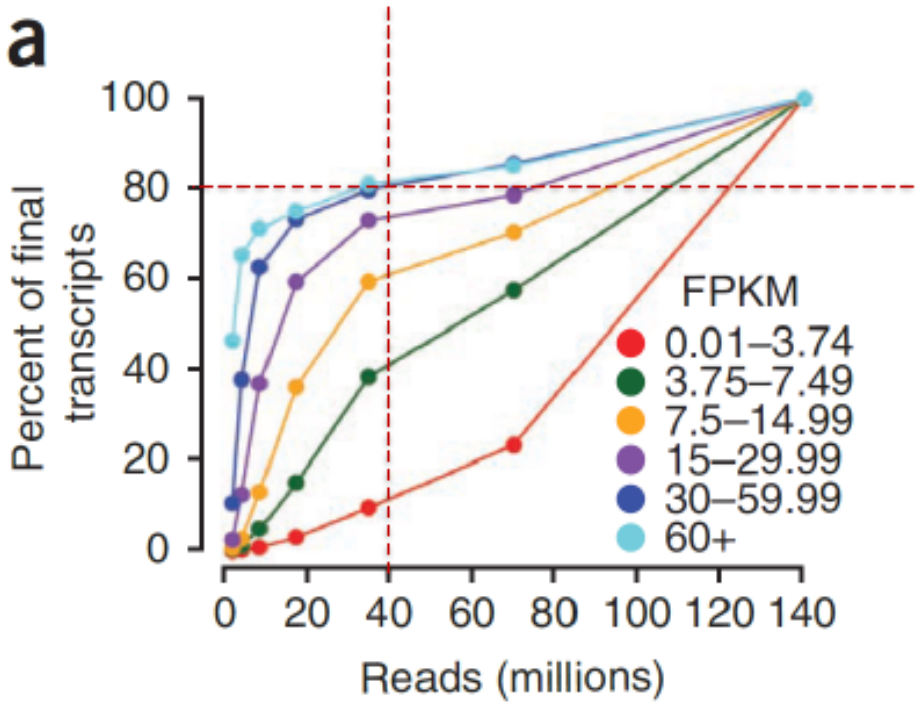
Transcript differential expression testing



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….



Trapnell et al. Nature Protocol 2012

ENCODE saturation analysis

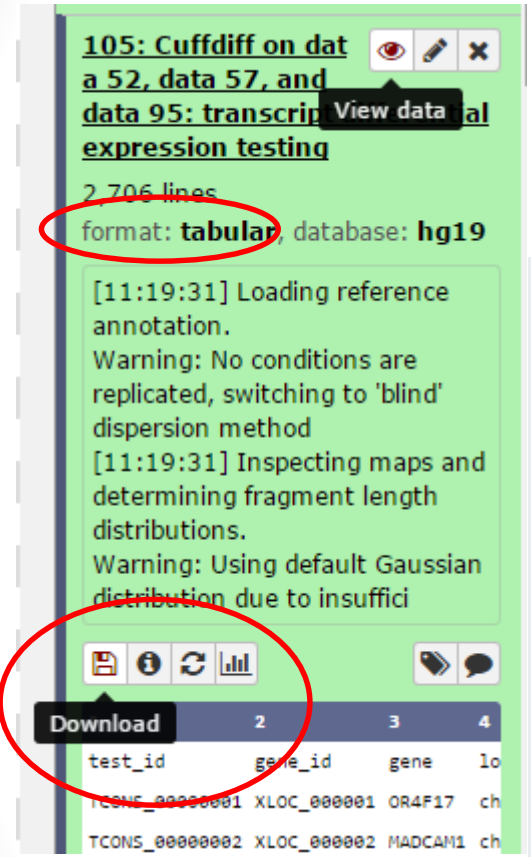
- 214 million 2x100bp PE reads
- H1 human embryonic stem cells
- 80% of the genes with FPKM≥10 are detected by ~36 million mapped reads per sample
- Genes with FPKM<10: ~80 million mapped reads per sample

Sims et al., 2014 Nature Reviews

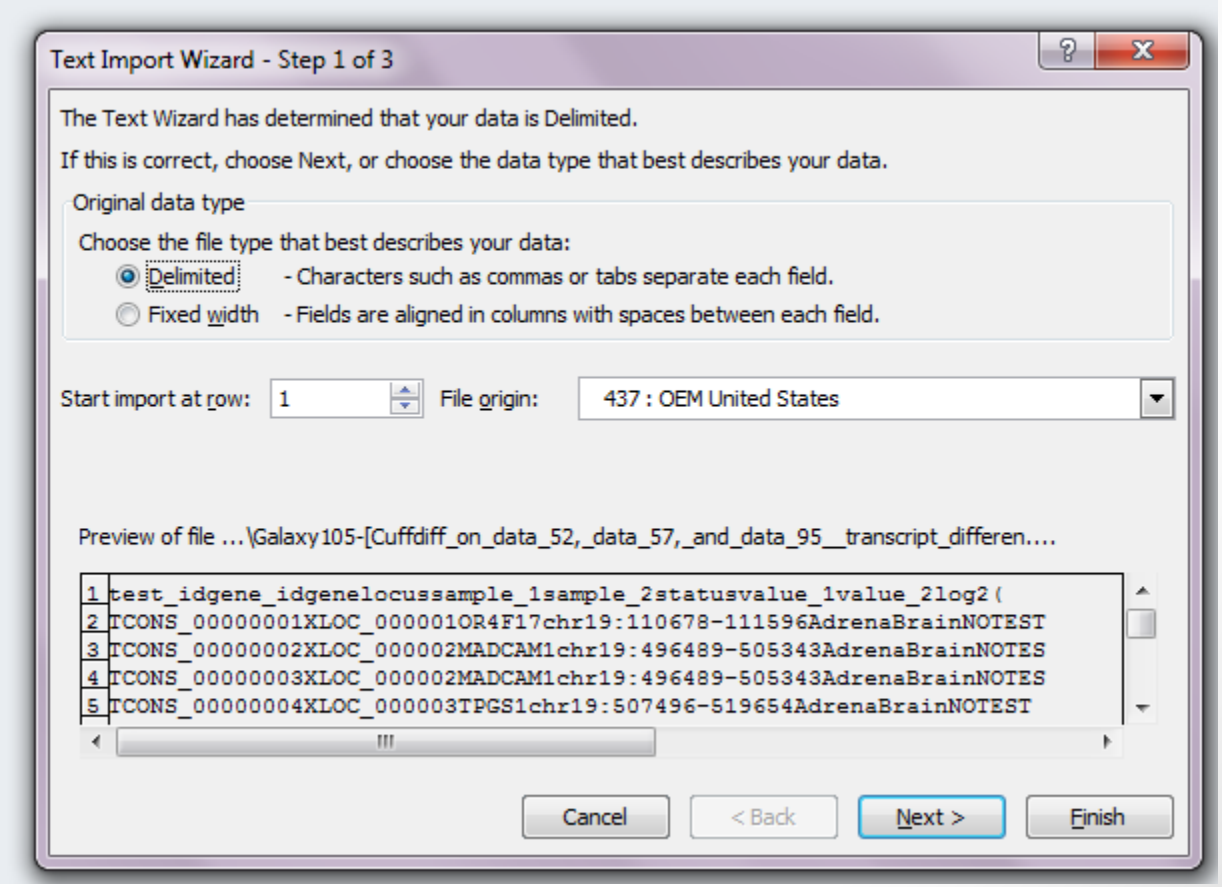
Discussion Points:

- **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.



Understand Your Data…

B2 XLOC_000001

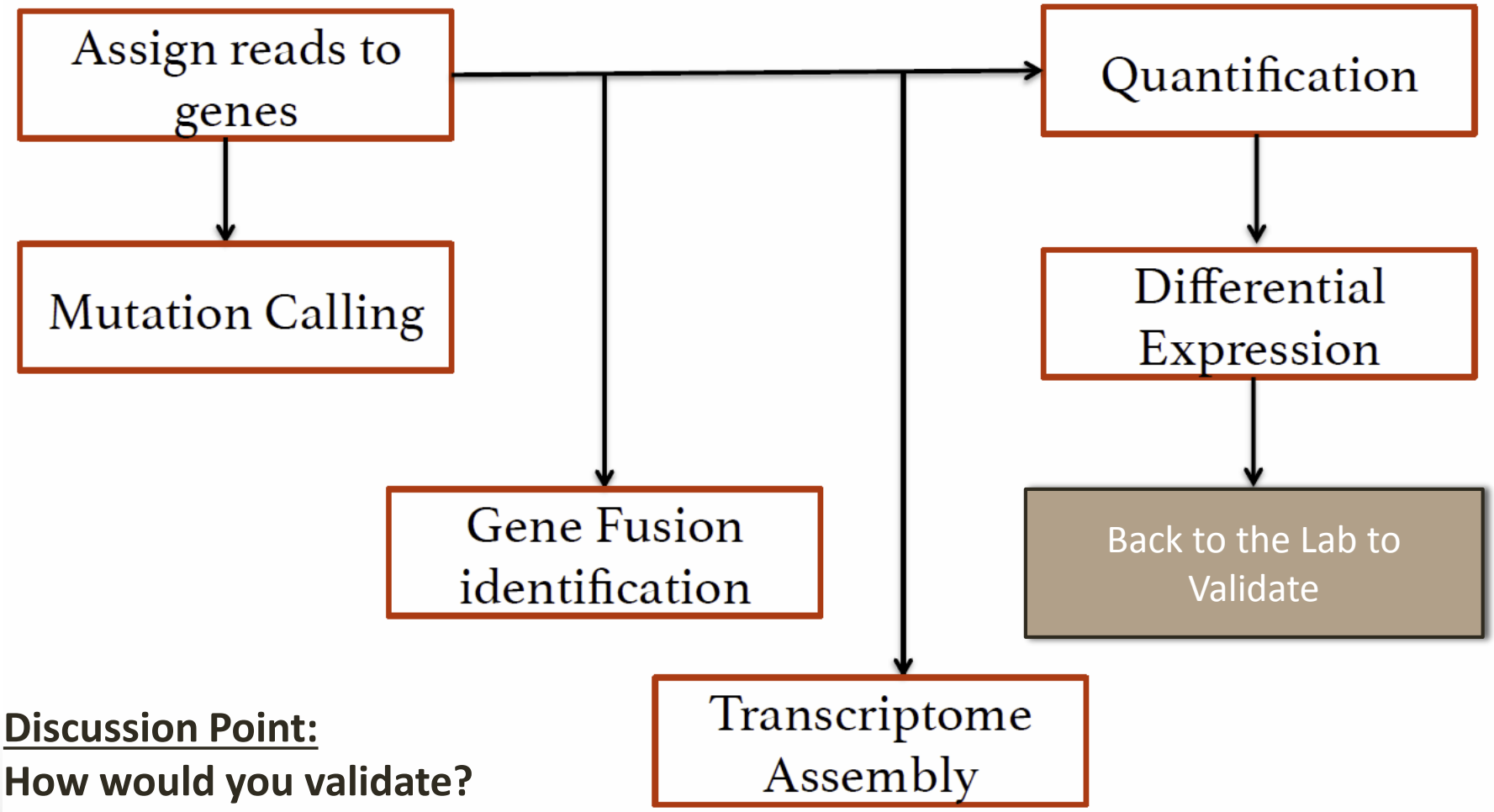
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
2	TCONS_00000001	XLOC_000001	OR4F17	chr19:110678-111596	Adrena	Brain	NOTEST	0	0	0	0	1	1	no
3	TCONS_00000002	XLOC_000002	MADCAM1	chr19:496489-505343	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?

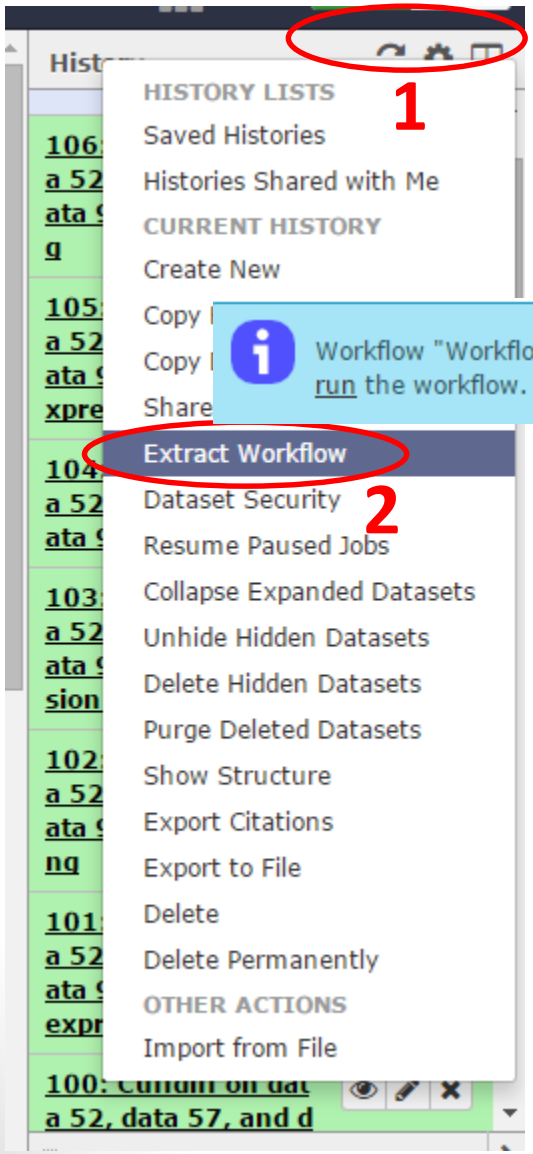
Data Obtained, now back to the Lab to Validate!

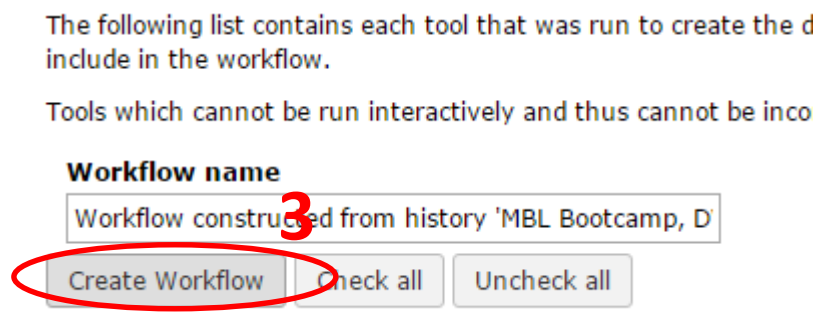


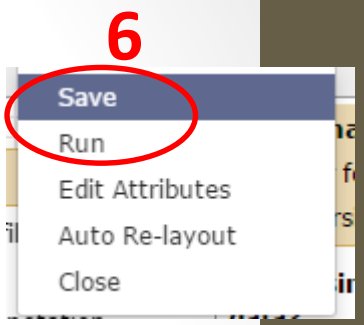
<u>Discussion Point:</u>
How would you validate?

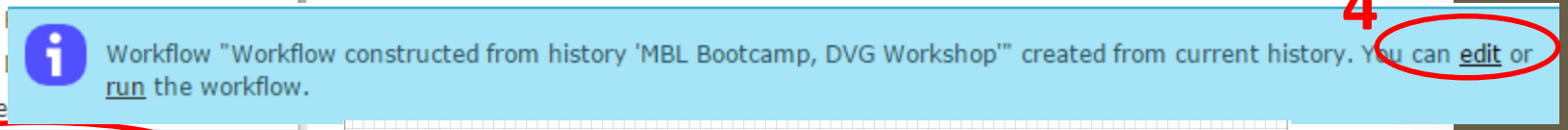
Create a Workflow

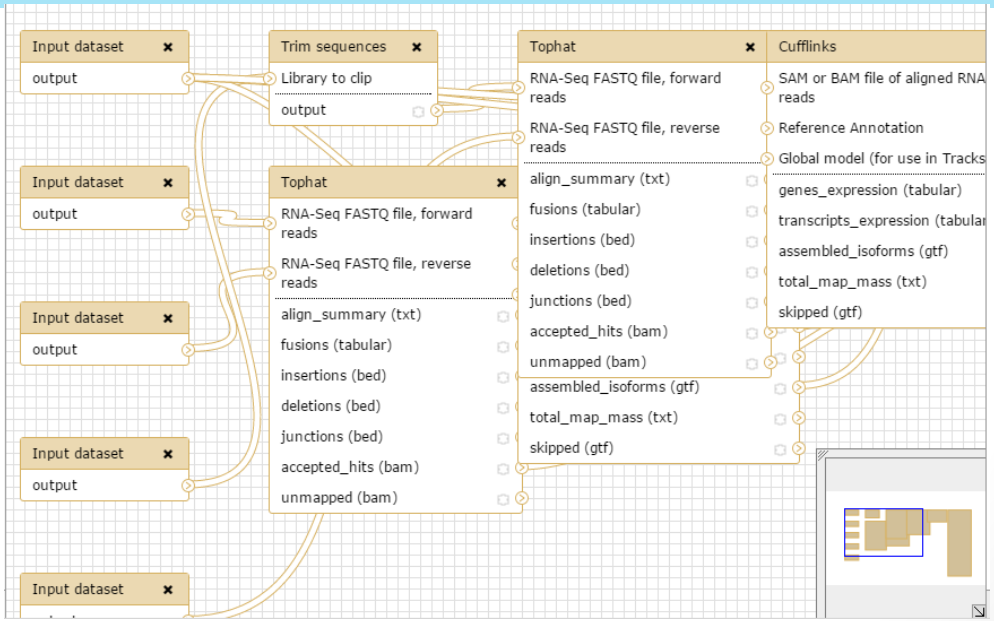
To use, share, and publish

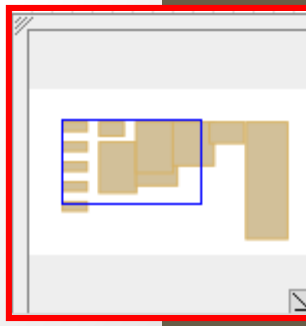












Many Options for Analyses Tools…

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[8]	Seed methods	Short-read mapping package (SHRiMP)[43]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[19]	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[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[30]			
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 expectaion 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[77]			
		Differential Expression analysis of count data (DESeq)[78]			
		Myrna[75]	Cloud-based permutation method		

http://www.ncbi.nlm.nih.gov/pubmed/21623353

Finding Publically Available Data

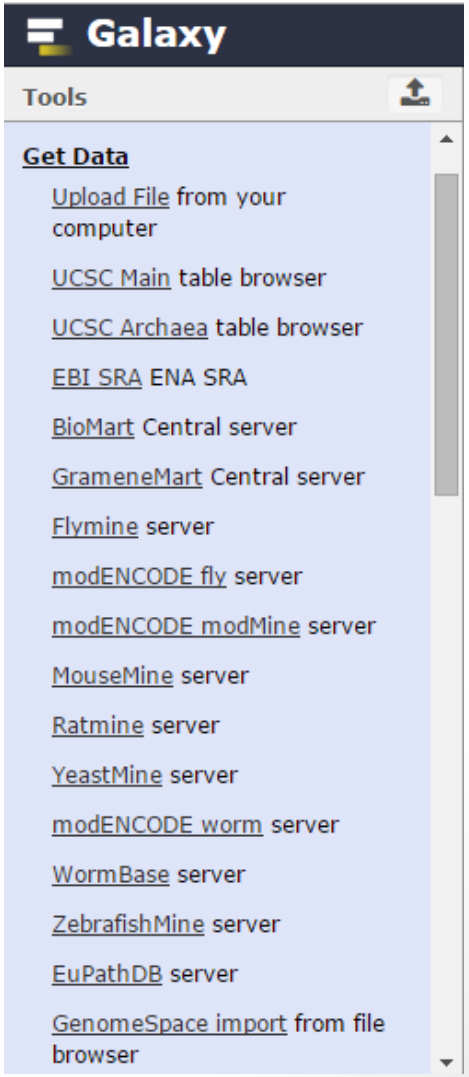
Federal-Funded Studies Must Make Datasets Publically Available

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:

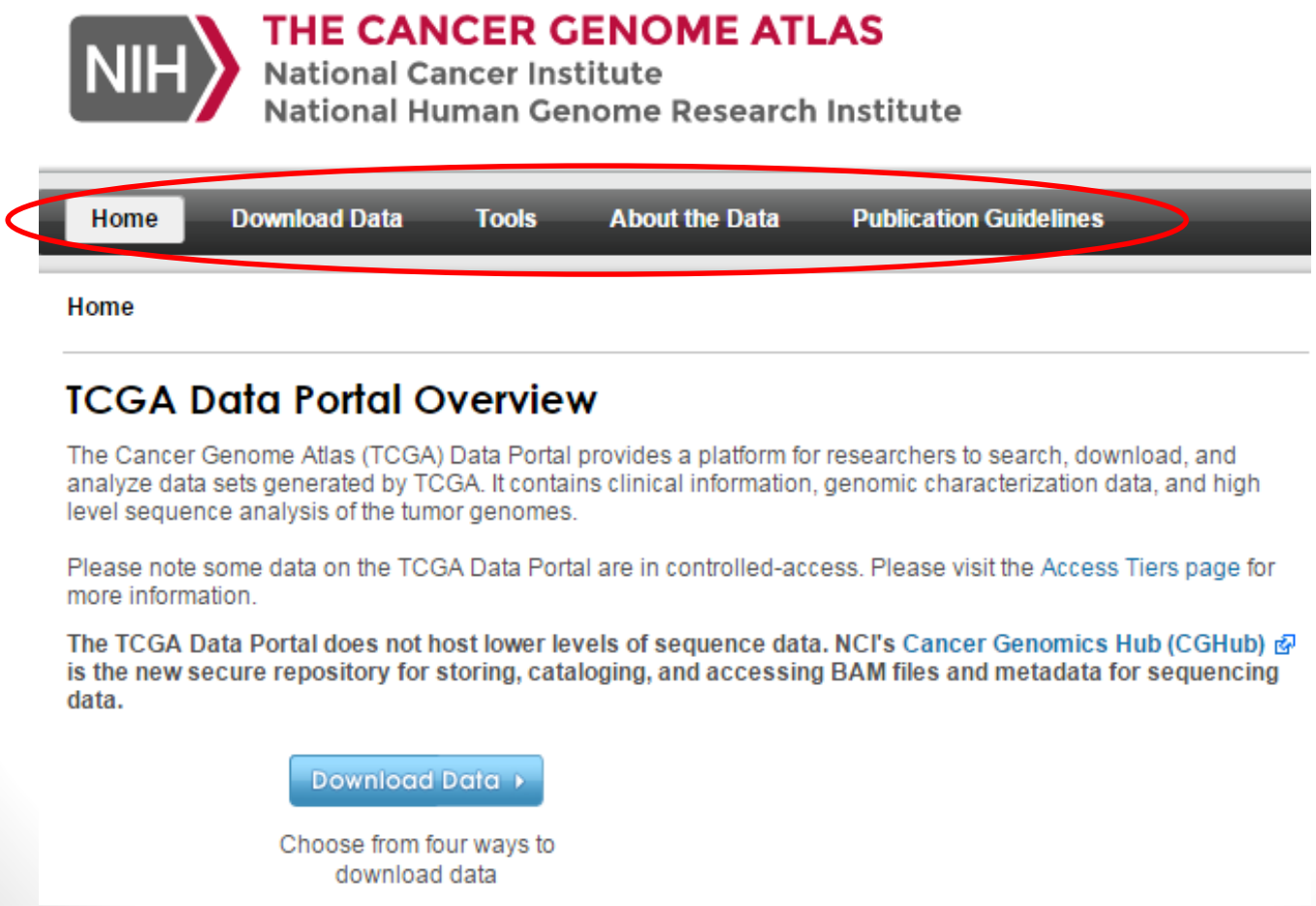
by name | by name and cell line | by ATC code

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



Let's Play

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

Query | Download Data

Select Cancer Study:

Search... | 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)**

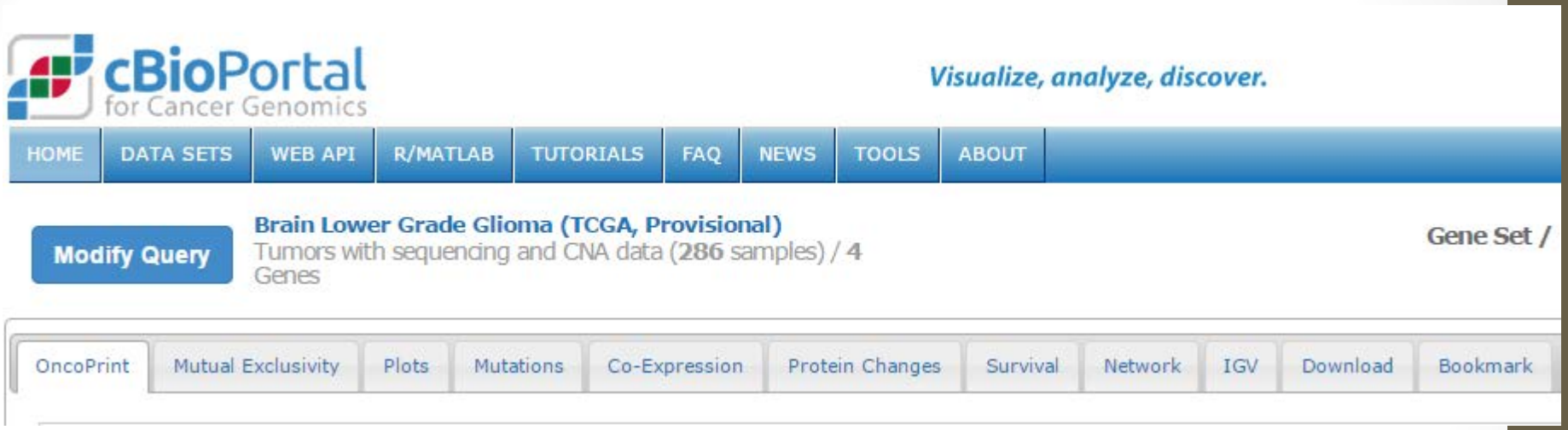
Select Data Type Priority: Mutation and CNA | Only Mutation | Only CNA

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

User-defined List

Enter HUGO Gene Symbols or Gene Aliases

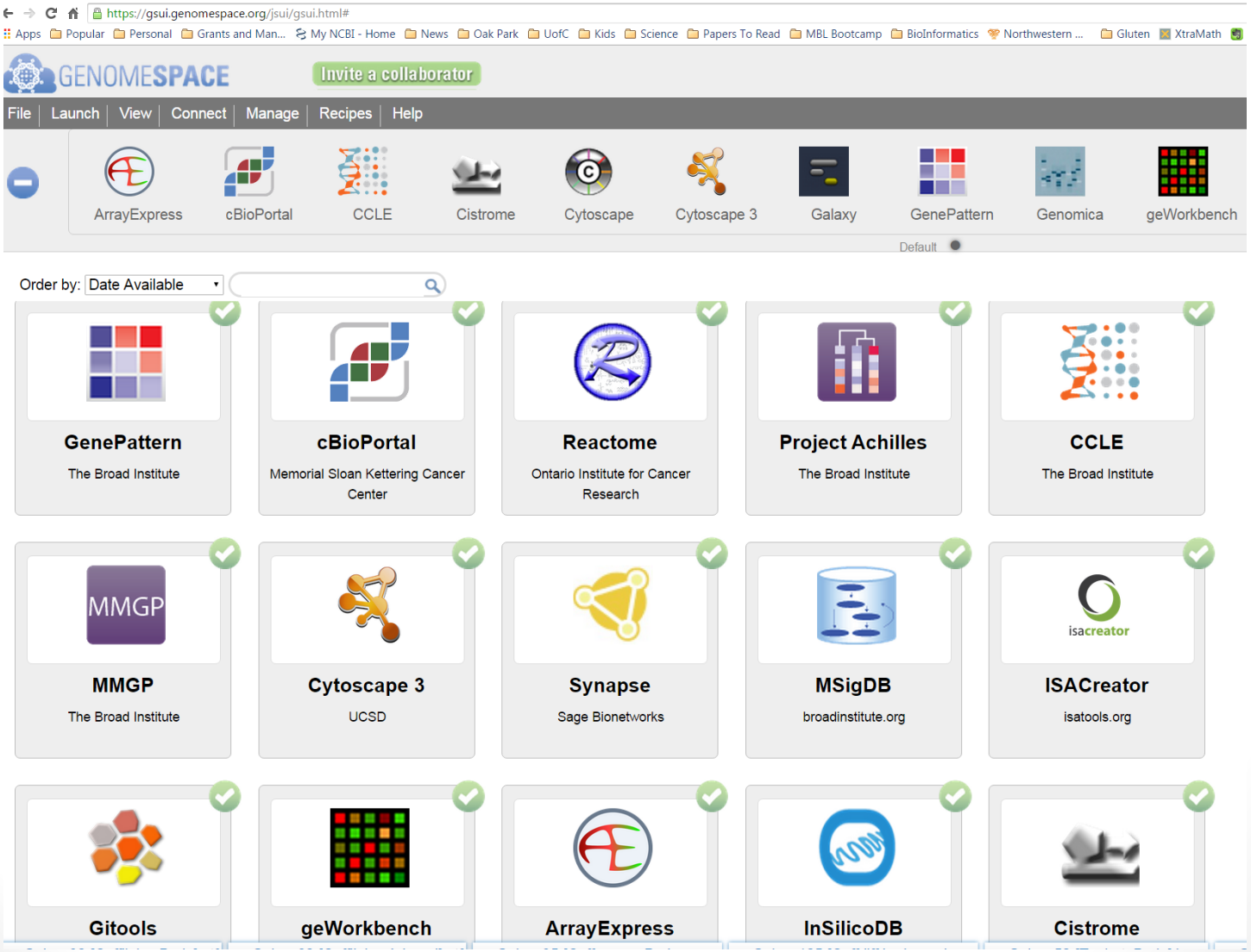
cBioPortal: Outputs



- **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/



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

GENOMESPACE | Log out carcinoman

> / > Home > Public > carcinoman > MBL Files

- .. (To Parent)
- [x] Galaxy1-[_iGenomes_UCSC_hg19,_chr19_gene_annotation].gtf
- [x] Galaxy2-[adrenal_1.fastq].fastqsanger
- [x] Galaxy3-[adrenal_2.fastq].fastqsanger
- [x] Galaxy4-[brain_1.fastq].fastqsanger
- [x] Galaxy5-[brain_2.fastq].fastqsanger
- Cufflinks GTF Files
- TopHat BAM Files

5 files selected | Send to Galaxy