Using RNA-seq for gene annotation, quantitation, and functional comparison in non-model organisms

Jeremy Goecks, Anton Nekrutenko, The Galaxy Team, and James Taylor

The Question

What can I do with my Illumina RNA-seq reads from my non-model organism(s) using open source software?

The Question

What can I do with my Illumina RNA-seq reads from my non-model organism(s) using open source software?

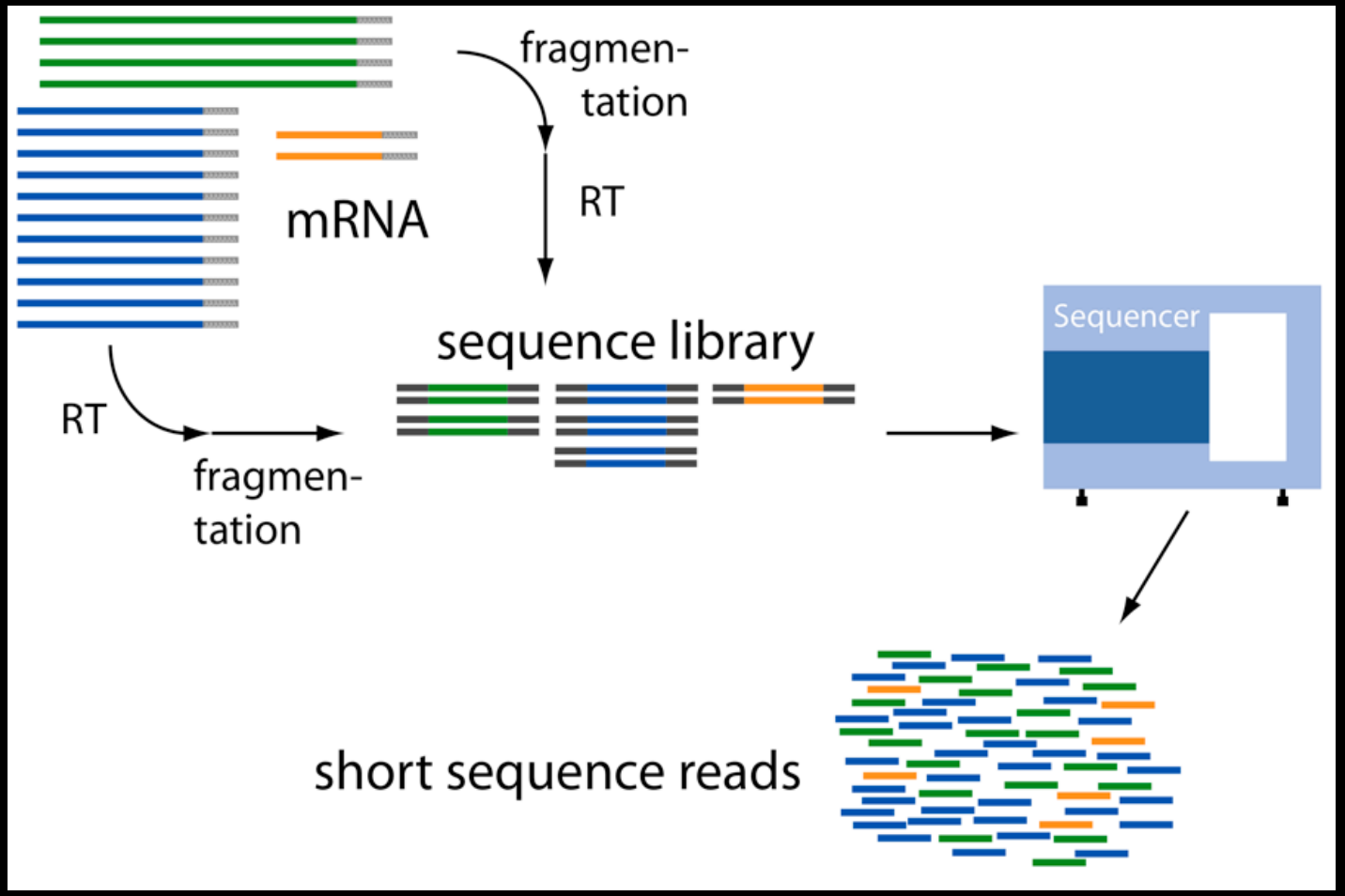
GBs of short, paired-end seq reads

no gene annotation, perhaps no genome

freely available to all, reproducible

RNA-seq

next generation gene expression data



http://www.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics

Current Projects

Study of organism infection

- RNA-seq data from infected, uninfected individuals
- genome but no annotations

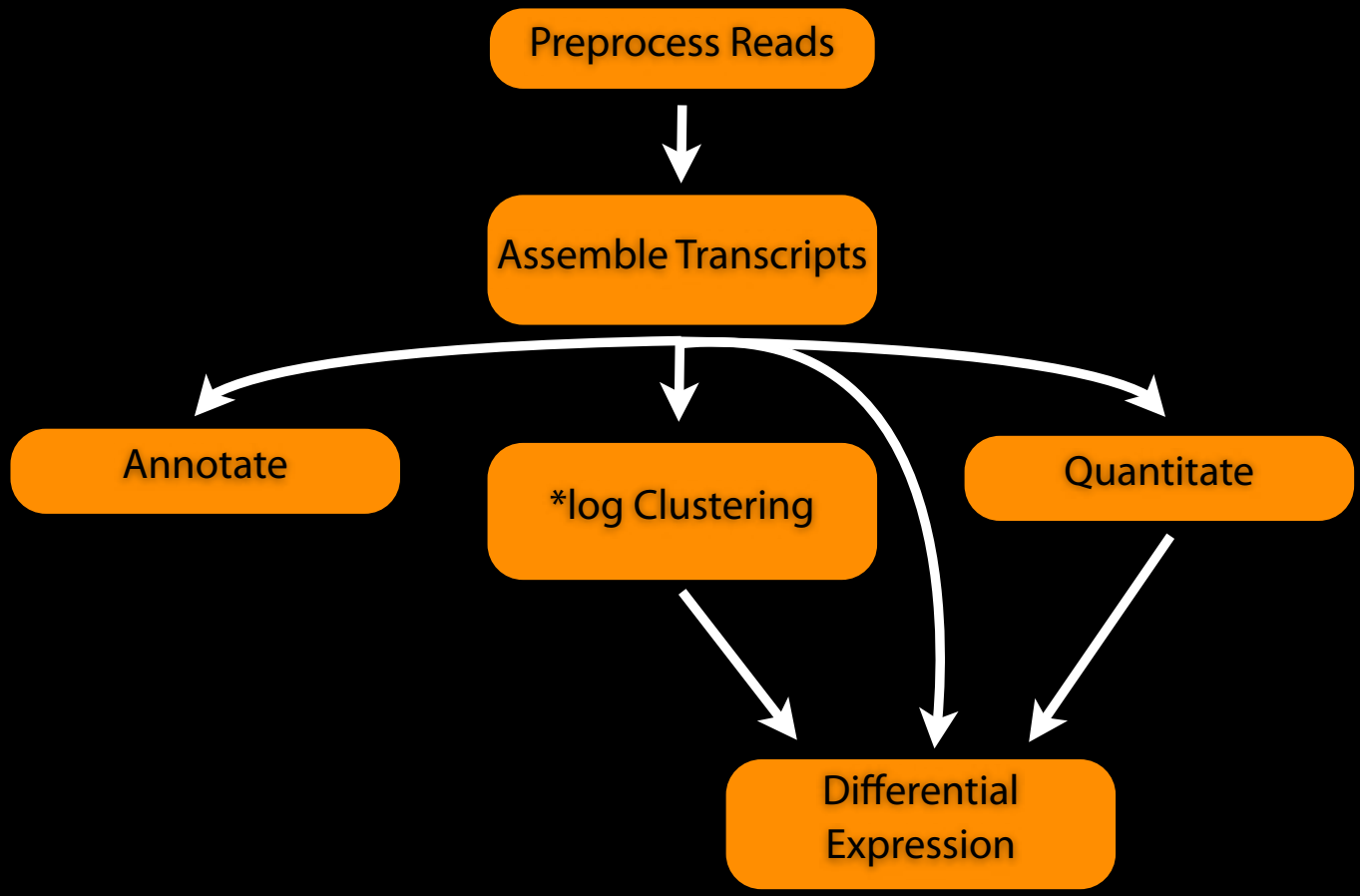
Study of dietary adaptations

- RNA-seq data from 4 related species; single lane of Illumina 76bp paired-end reads per species
- no genome or annotations

Study of parasite venoms

- RNA-seq data from 3 related parasite venoms, two lanes of Illumina 100bp paired-end reads per venom
- no genome or annotations

Overview



*log = homolog and/or ortholog

Desired Outputs

For each organism

- assembled transcripts, quantitated and annotated

For pairs/groups of organisms

- *log clusters
- differential expression amongst genes and/or *logs

What is Galaxy?

Web-based GUI for genomics that requires only a Web browser for **everything**: analysis, workflows, sharing, publication, and visualization

A public web service (http://usegalaxy.org) integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

Open source software that makes integrating your own tools and data and customizing for your own site simple

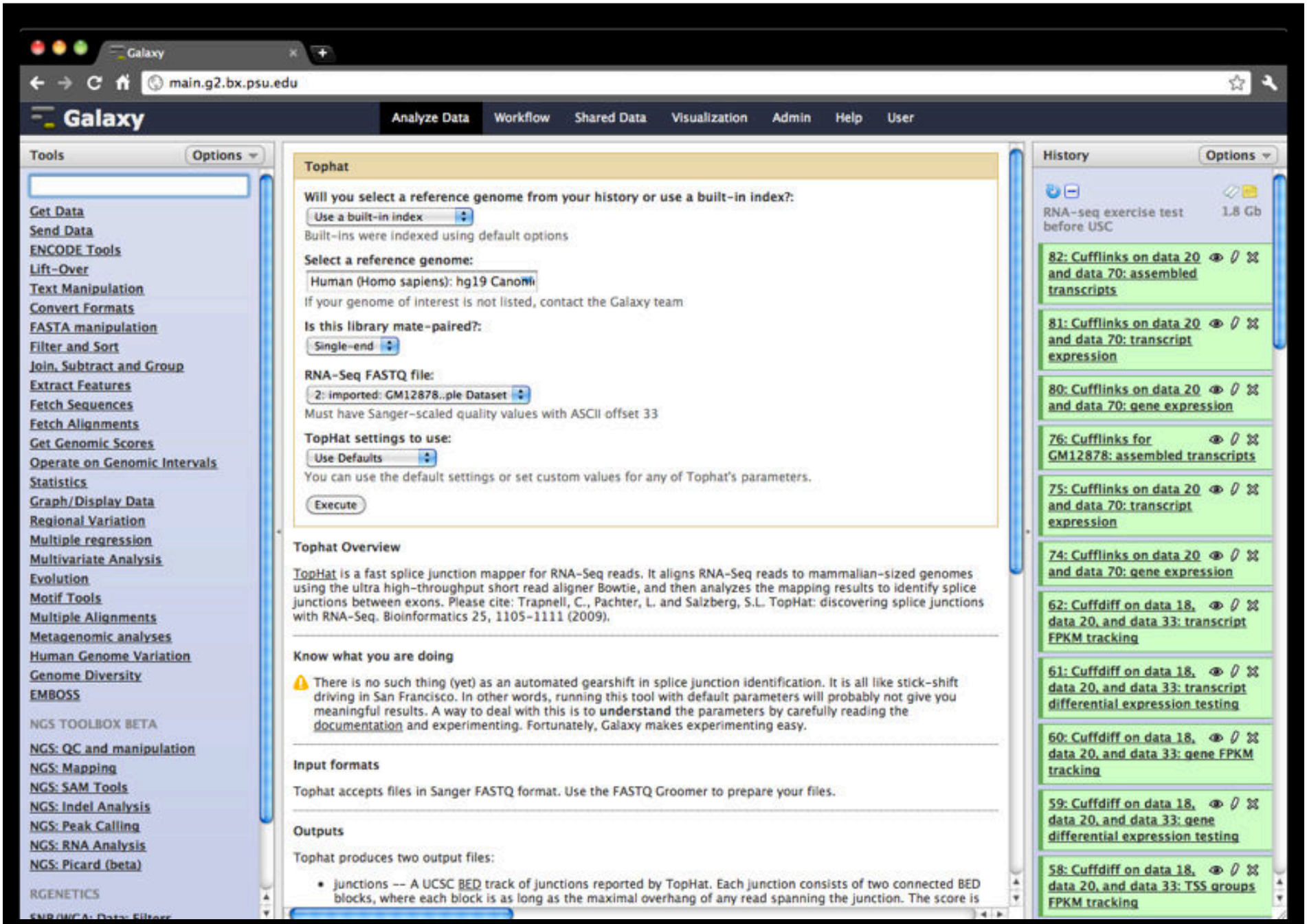
Galaxy in a Nutshell

What you can do in Galaxy

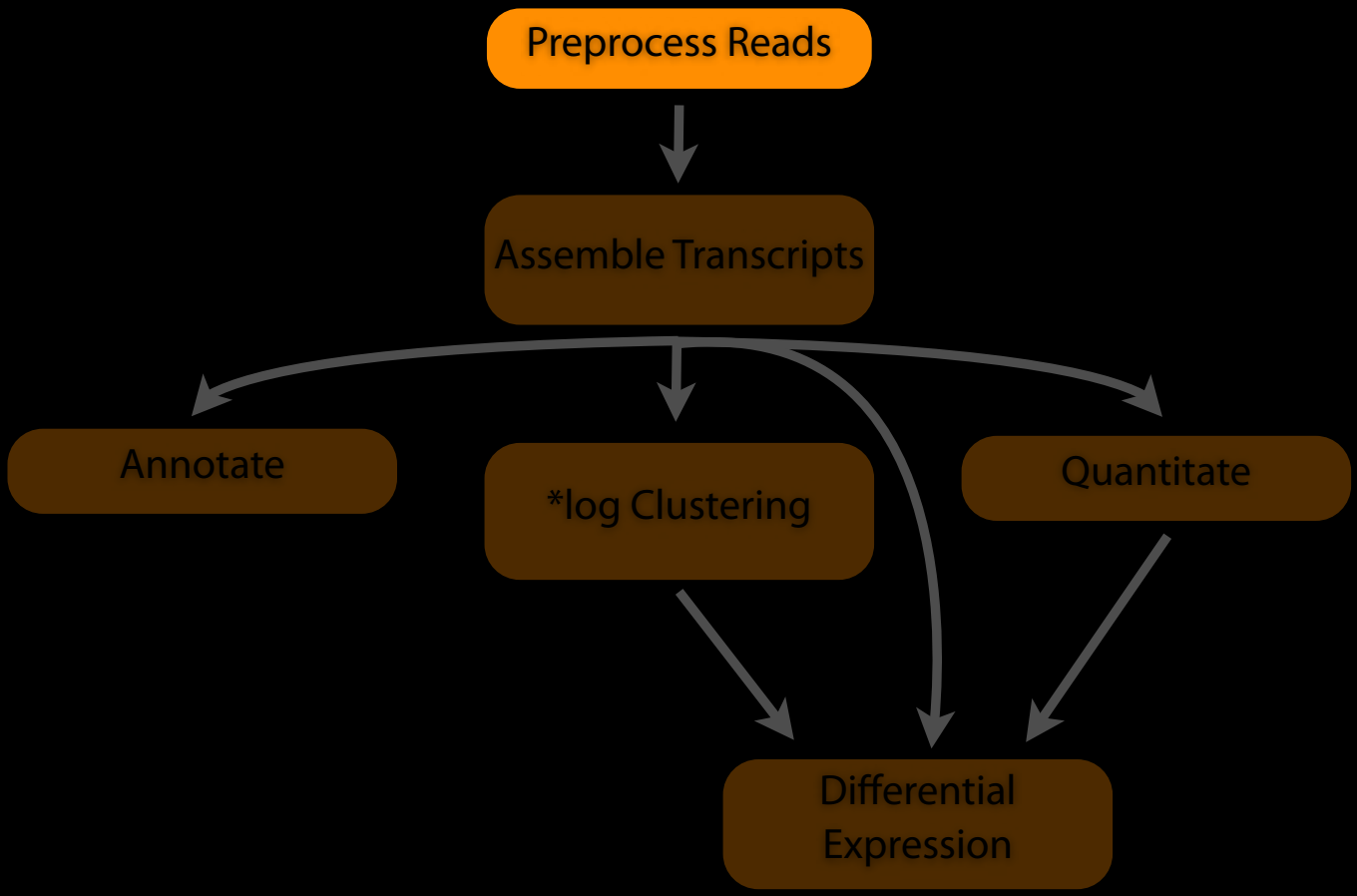
- analysis interface, tools and datasources
- data libraries
- workflows
- visualization
- sharing
- Pages

Where you can use and build Galaxy

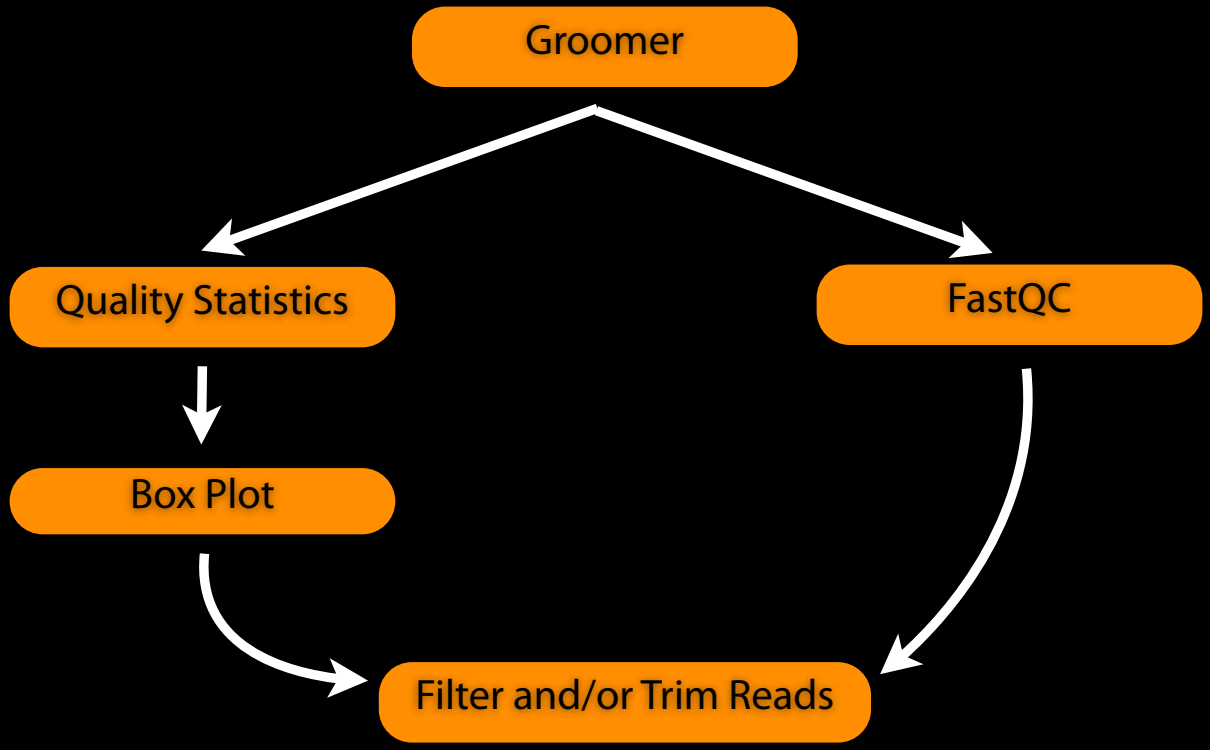
- public website
- local instance
- on the cloud
- tool shed/contributing tools


Galaxy
main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)
RGENETICS
Tophat
Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Built-ins were indexed using default options
Select a reference genome:
Human (Homo sapiens): hg19 Canoni
If your genome of interest is not listed, contact the Galaxy team
Is this library mate-paired?:
Single-end
RNA-Seq FASTQ file:
2: imported: GM12878..ple Dataset
Must have Sanger-scaled quality values with ASCII offset 33
TopHat settings to use:
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.
Execute
Tophat Overview
TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).
Know what you are doing
There is no such thing (yet) as an automated gearshift in splice junction identification. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.
Input formats
Tophat accepts files in Sanger FASTQ format. Use the FASTQ Groomer to prepare your files.
Outputs
Tophat produces two output files:
junctions -- A UCSC BED track of junctions reported by TopHat. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is
History
Options
RNA-seq exercise test before USC
1.8 Gb
82: Cufflinks on data 20 and data 70: assembled transcripts
81: Cufflinks on data 20 and data 70: transcript expression
80: Cufflinks on data 20 and data 70: gene expression
76: Cufflinks for GM12878: assembled transcripts
75: Cufflinks on data 20 and data 70: transcript expression
74: Cufflinks on data 20 and data 70: gene expression
62: Cuffdiff on data 18, data 20, and data 33: transcript FPKM tracking
61: Cuffdiff on data 18, data 20, and data 33: transcript differential expression testing
60: Cuffdiff on data 18, data 20, and data 33: gene FPKM tracking
59: Cuffdiff on data 18, data 20, and data 33: gene differential expression testing
58: Cuffdiff on data 18, data 20, and data 33: TSS groups FPKM tracking

Overview


Preprocess Reads
Assemble Transcripts
Annotate
*log Clustering
Quantitate
Differential Expression

Preprocessing Reads in Galaxy



Blankenberg *et al*. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5.

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Quality Statistics and Box Plot Tool

NGS TOOLBOX BETA

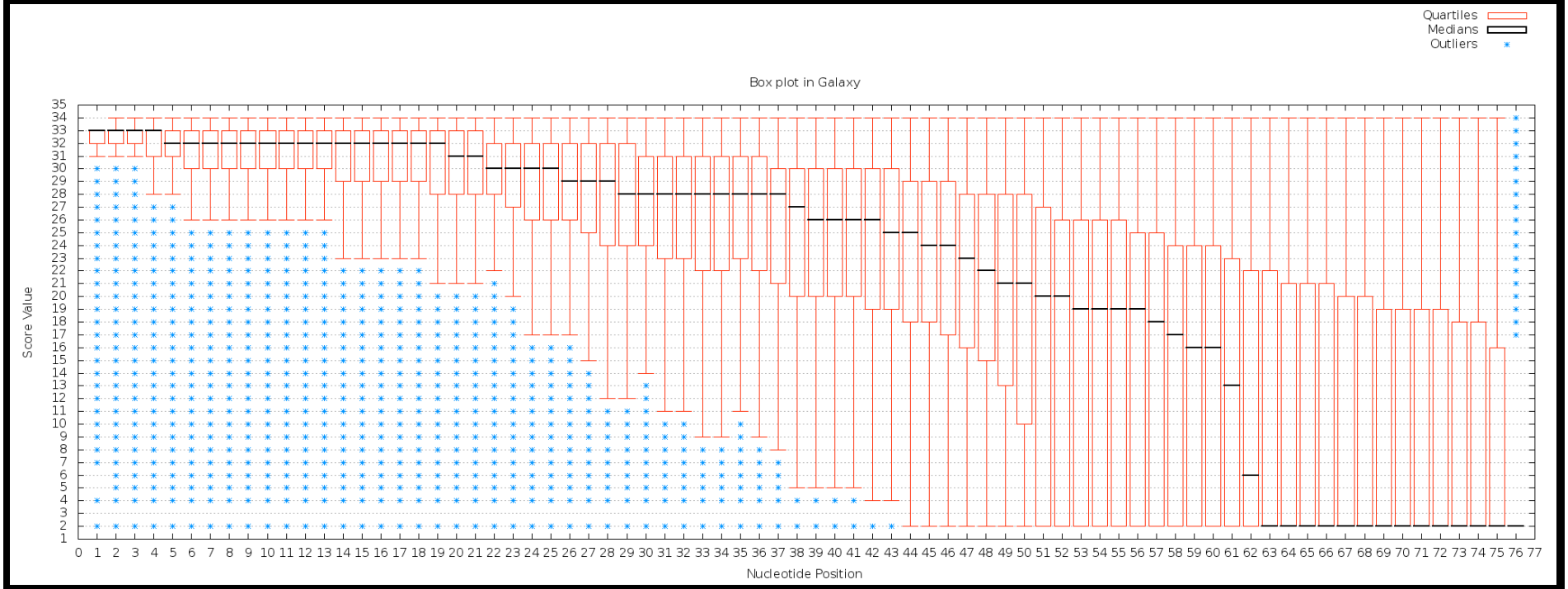
NGS: QC and manipulation

ILLUMINA DATA

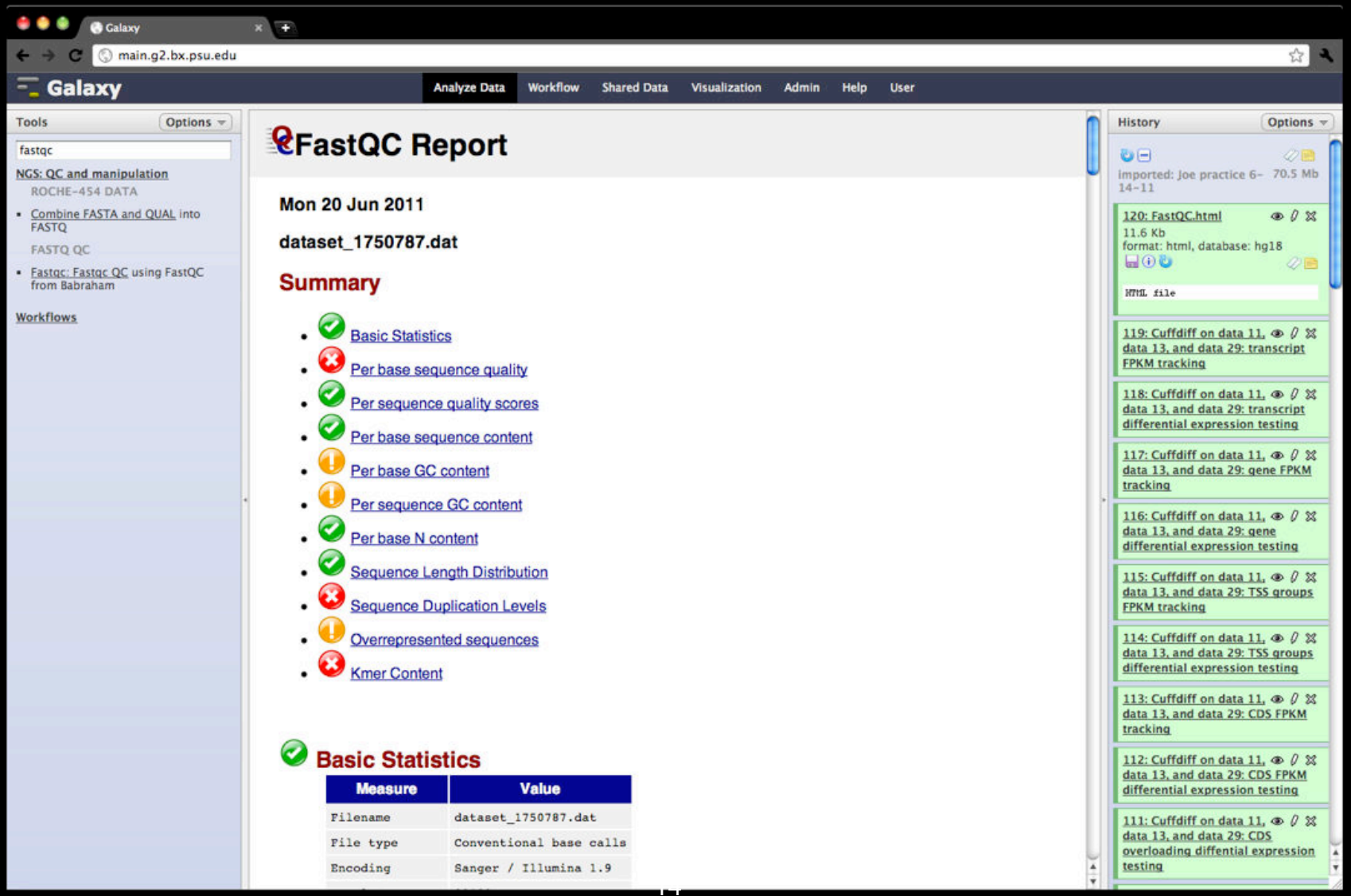
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

Graph/Display Data

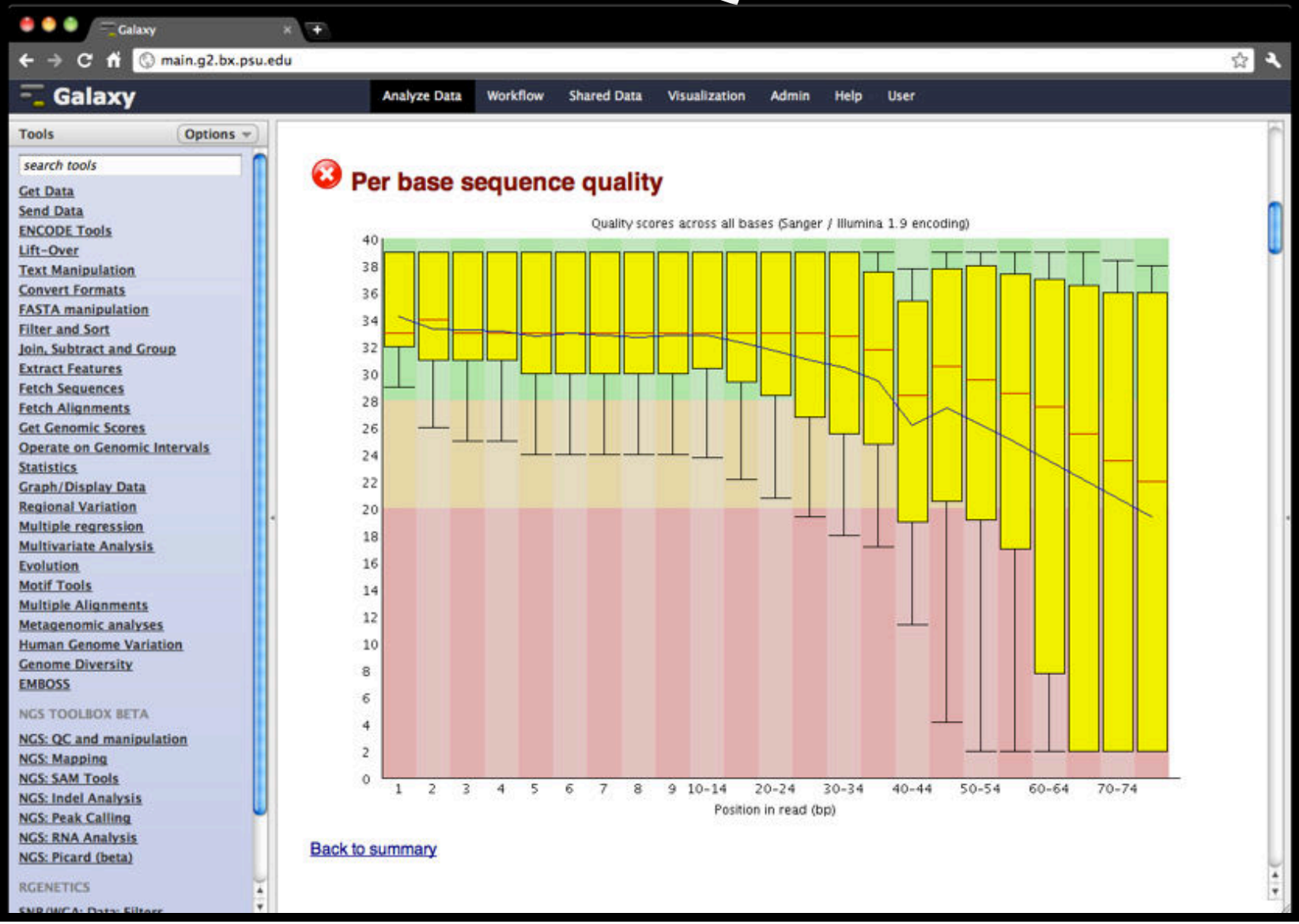
- Histogram of a numeric column
- Scatterplot of two numeric columns
- Plotting tool for multiple series and graph types
- Boxplot of quality statistics



FastQC


Galaxy
main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Tools
Options
fastqc
NGS: QC and manipulation
ROCHE-454 DATA
Combine FASTA and QUAL into FASTQ
FASTQ QC
Fastqc: Fastqc QC using FastQC from Babraham
Workflows
FastQC Report
Mon 20 Jun 2011
dataset_1750787.dat
Summary
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Basic Statistics
Measure
Value
Filename
dataset_1750787.dat
File type
Conventional base calls
Encoding
Sanger / Illumina 1.9
History
Options
imported: Joe practice 6-14-11
70.5 Mb
120: FastQC.html
11.6 Kb
format: html, database: hg18
HTML file
119: Cuffdiff on data 11, data 13, and data 29: transcript FPKM tracking
118: Cuffdiff on data 11, data 13, and data 29: transcript differential expression testing
117: Cuffdiff on data 11, data 13, and data 29: gene FPKM tracking
116: Cuffdiff on data 11, data 13, and data 29: gene differential expression testing
115: Cuffdiff on data 11, data 13, and data 29: TSS groups FPKM tracking
114: Cuffdiff on data 11, data 13, and data 29: TSS groups differential expression testing
113: Cuffdiff on data 11, data 13, and data 29: CDS FPKM tracking
112: Cuffdiff on data 11, data 13, and data 29: CDS FPKM differential expression testing
111: Cuffdiff on data 11, data 13, and data 29: CDS overloading diffential expression testing

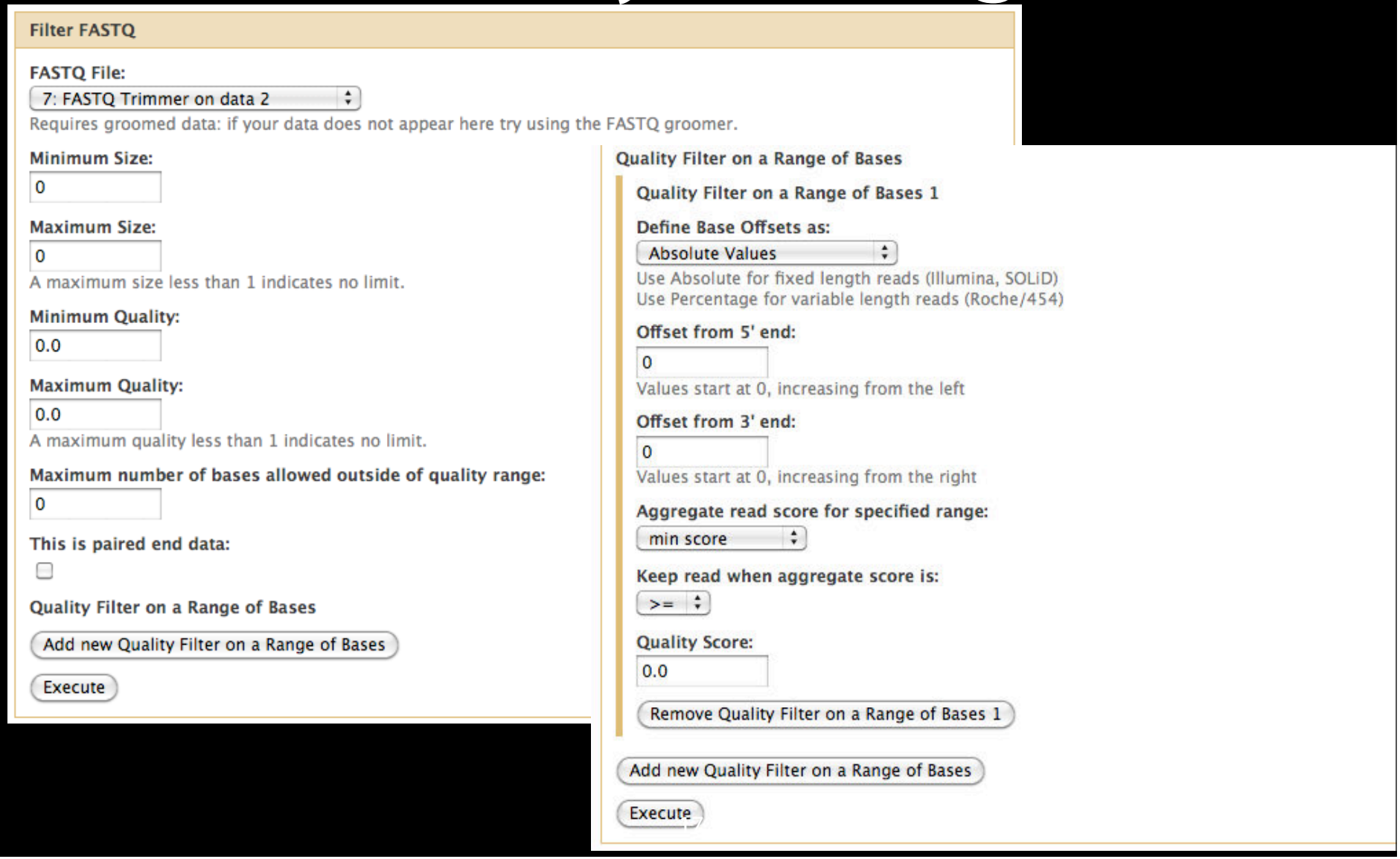
FastQC


Galaxy
main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Tools
Options
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)
RGENETICS
Per base sequence quality
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
10-14
20-24
30-34
40-44
50-54
60-64
70-74
Position in read (bp)
Back to summary

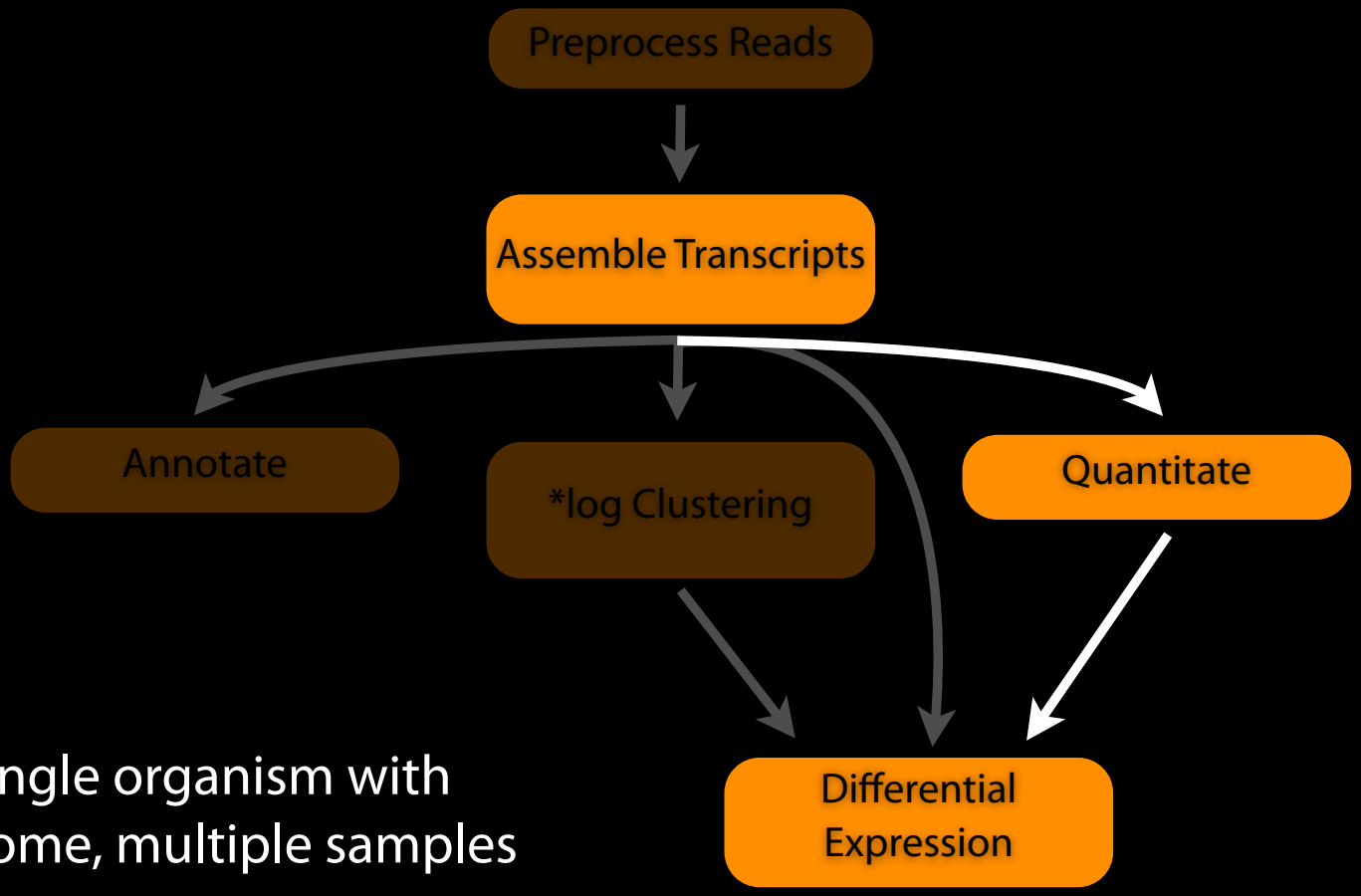
FastQC


Galaxy
main.g2.bx.psu.edu
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Tools
Options
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)
RGENETICS
Per sequence quality scores
Quality score distribution over all sequences
Average Quality per read
10000
8000
6000
4000
2000
0
2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
Mean Sequence Quality (Phred Score)
Back to summary

Quality Filtering

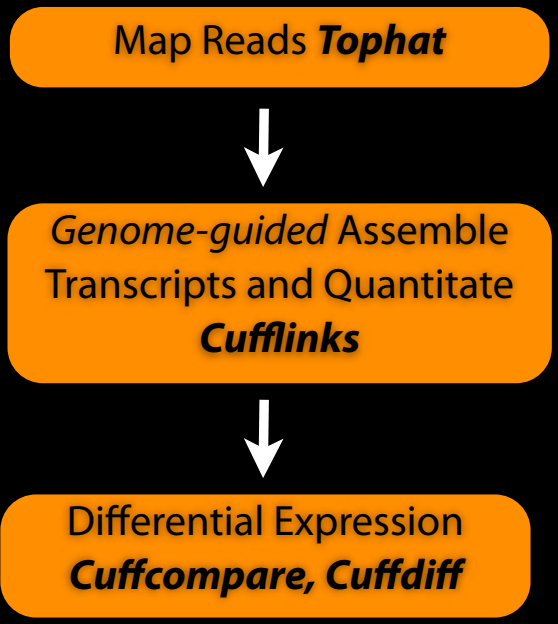


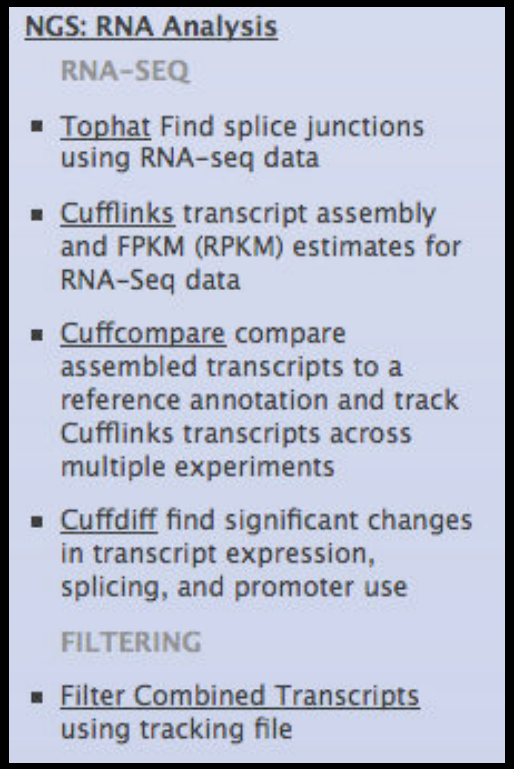
Overview: Single Organism, Genome but no Annotations



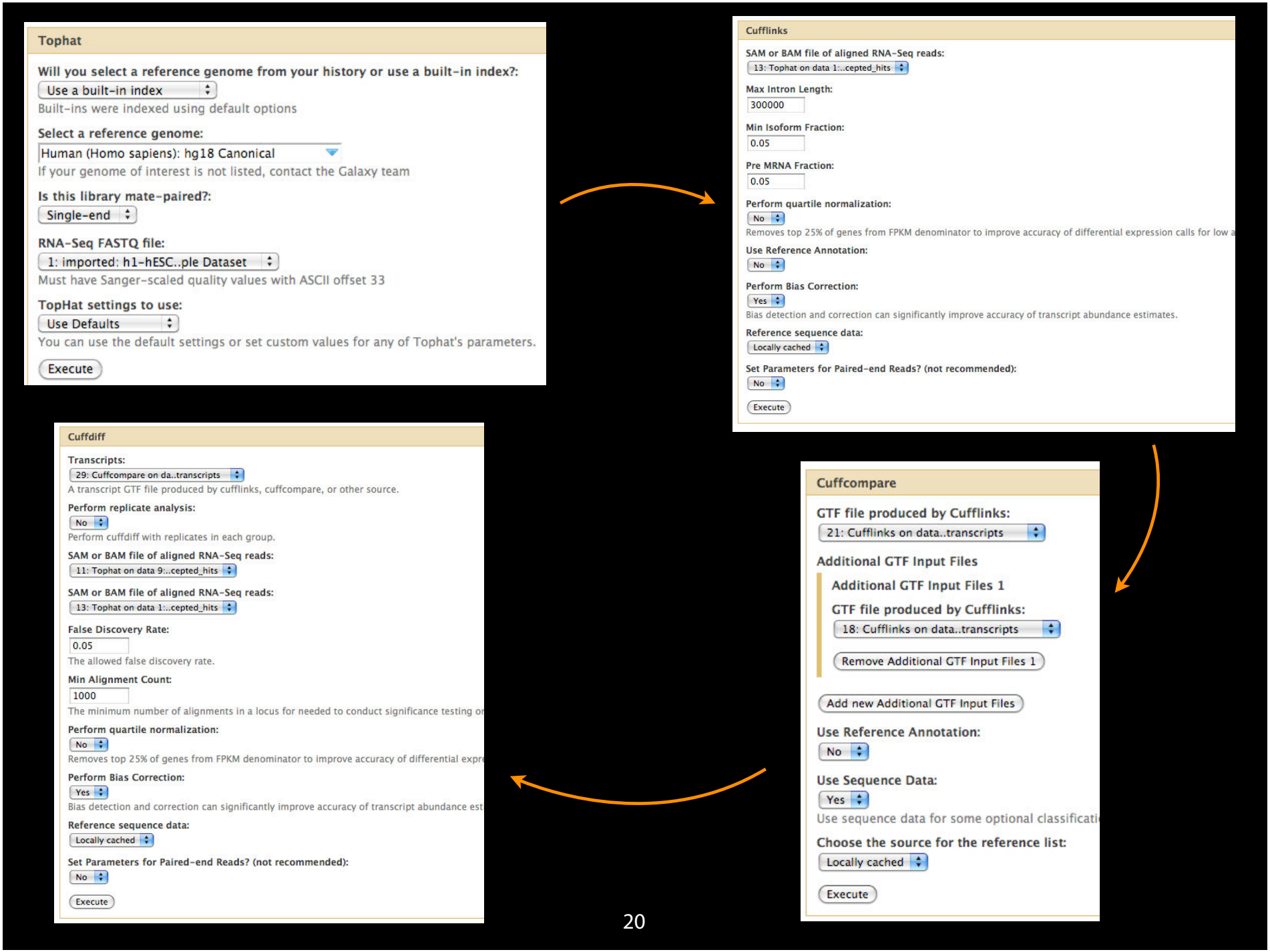
single organism with genome, multiple samples

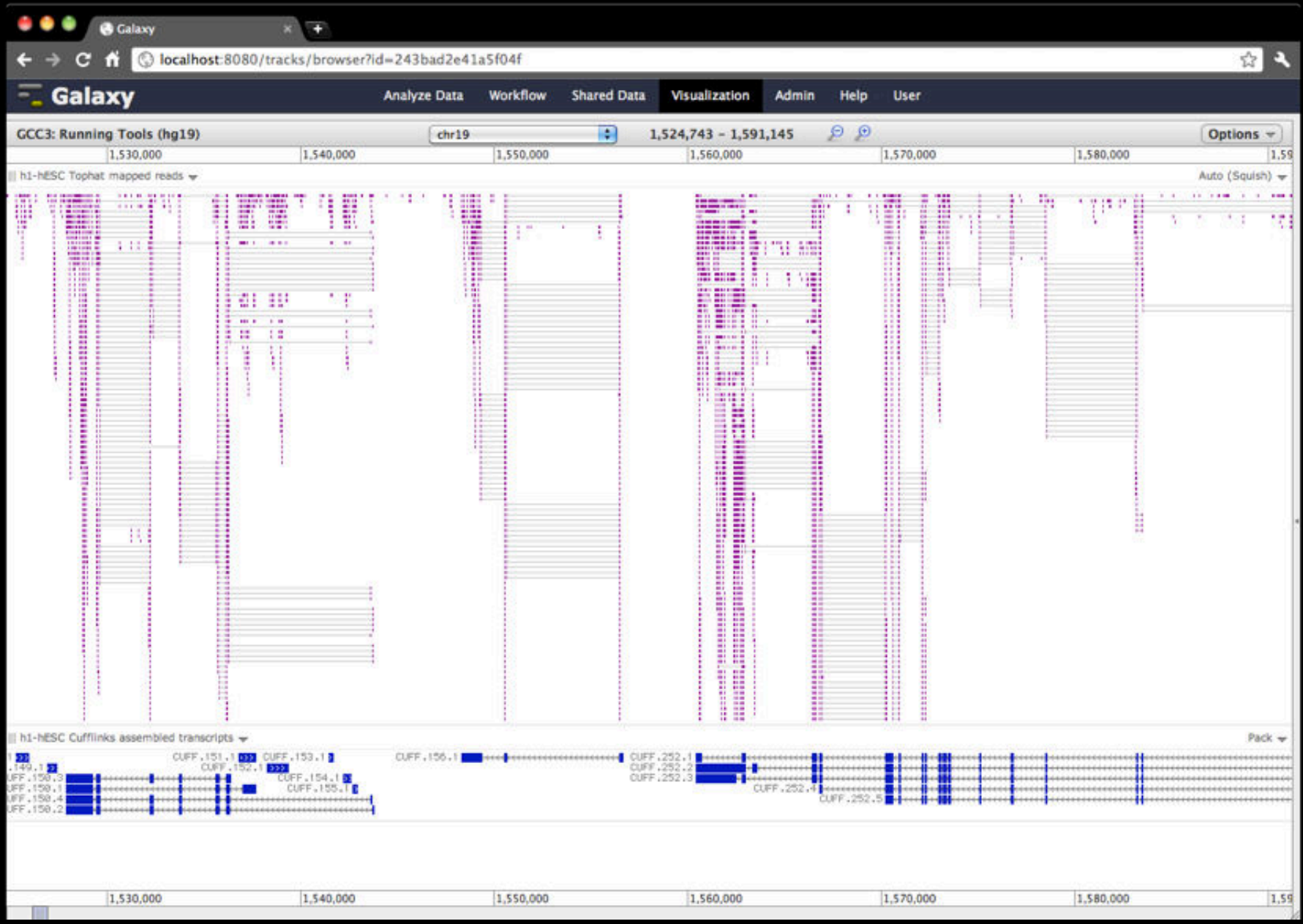
Genome but no Annotations (using Galaxy)

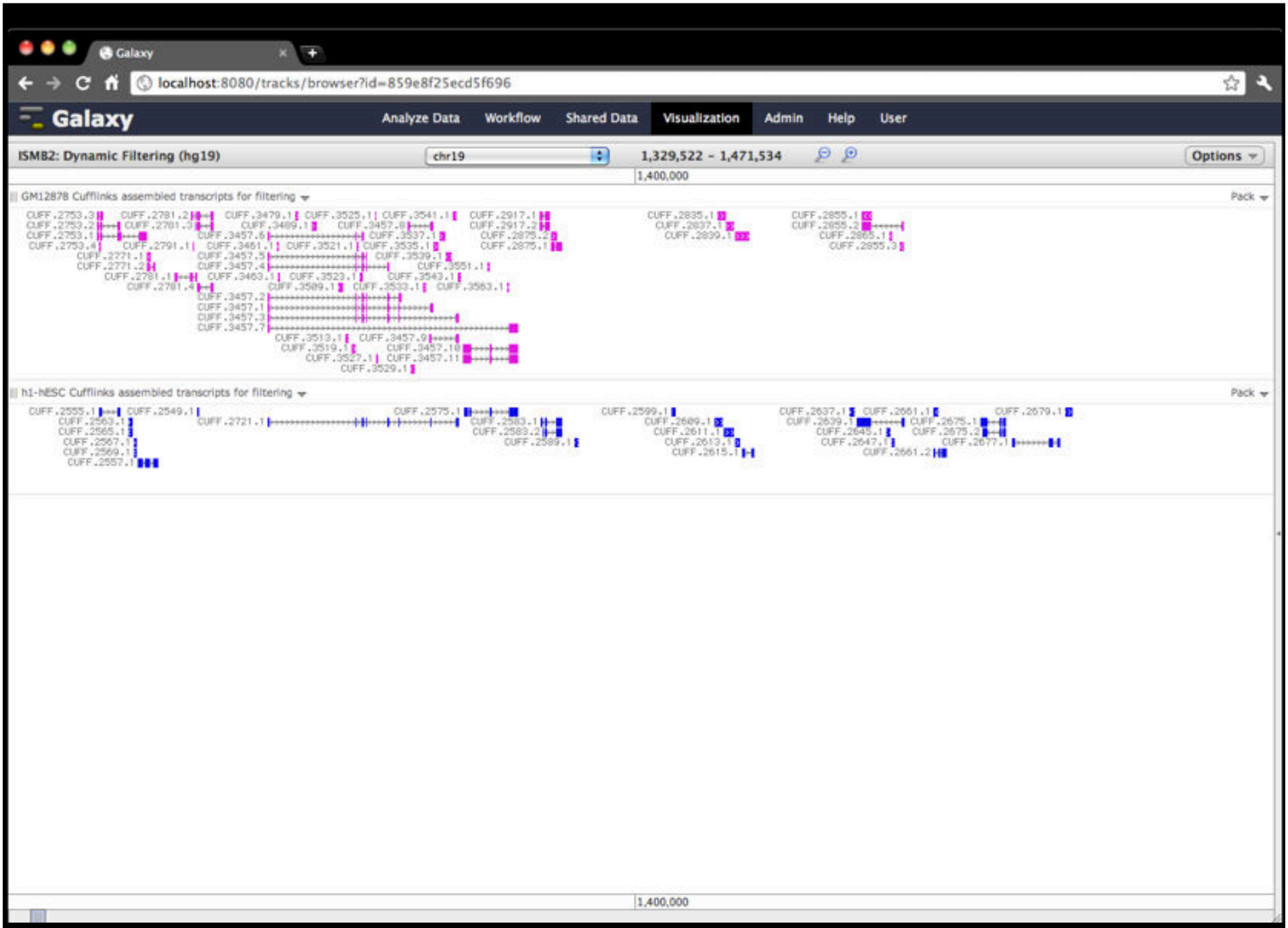




1. Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).
2. Trapnell et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621


Tophat
Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Built-ins were indexed using default options
Select a reference genome:
Human (Homo sapiens): hg18 Canonical
If your genome of interest is not listed, contact the Galaxy team
Is this library mate-paired?:
Single-end
RNA-Seq FASTQ file:
1: imported: h1-hESC..ple Dataset
Must have Sanger-scaled quality values with ASCII offset 33
TopHat settings to use:
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.
Execute
Cufflinks
SAM or BAM file of aligned RNA-Seq reads:
13: Tophat on data 1:..cepted_hits
Max Intron Length:
300000
Min Isoform Fraction:
0.05
Pre MRNA Fraction:
0.05
Perform quartile normalization:
No
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low a
Use Reference Annotation:
No
Perform Bias Correction:
Yes
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.
Reference sequence data:
Locally cached
Set Parameters for Paired-end Reads? (not recommended):
No
Execute
Cuffcompare
GTF file produced by Cufflinks:
21: Cufflinks on data..transcripts
Additional GTF Input Files
Additional GTF Input Files 1
GTF file produced by Cufflinks:
18: Cufflinks on data..transcripts
Remove Additional GTF Input Files 1
Add new Additional GTF Input Files
Use Reference Annotation:
No
Use Sequence Data:
Yes
Use sequence data for some optional classificati
Choose the source for the reference list:
Locally cached
Execute
Cuffdiff
Transcripts:
29: Cuffcompare on da..transcripts
A transcript GTF file produced by cufflinks, cuffcompare, or other source.
Perform replicate analysis:
No
Perform cuffdiff with replicates in each group.
SAM or BAM file of aligned RNA-Seq reads:
11: Tophat on data 9:..cepted_hits
SAM or BAM file of aligned RNA-Seq reads:
13: Tophat on data 1:..cepted_hits
False Discovery Rate:
0.05
The allowed false discovery rate.
Min Alignment Count:
1000
The minimum number of alignments in a locus for needed to conduct significance testing or
Perform quartile normalization:
No
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expre
Perform Bias Correction:
Yes
Bias detection and correction can significantly improve accuracy of transcript abundance est
Reference sequence data:
Locally cached
Set Parameters for Paired-end Reads? (not recommended):
No
Execute


Galaxy
localhost:8080/tracks/browser?id=243bad2e41a5f04f
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
GCC3: Running Tools (hg19)
chr19
1,524,743 - 1,591,145
Options
1,530,000
1,540,000
1,550,000
1,560,000
1,570,000
1,580,000
h1-hESC Tophat mapped reads
Auto (Squish)
h1-hESC Cufflinks assembled transcripts
Pack
CUFF.151.1
CUFF.153.1
CUFF.152.1
CUFF.154.1
CUFF.155.1
CUFF.156.1
CUFF.252.1
CUFF.252.2
CUFF.252.3
CUFF.252.4
CUFF.252.5


Galaxy
localhost:8080/tracks/browser?id=859e8f25ecd5f696
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
ISMB2: Dynamic Filtering (hg19)
chr19
1,329,522 - 1,471,534
Options
1,400,000
GM12878 Cufflinks assembled transcripts for filtering
Pack
h1-hESC Cufflinks assembled transcripts for filtering
Pack
1,400,000

Quantitation

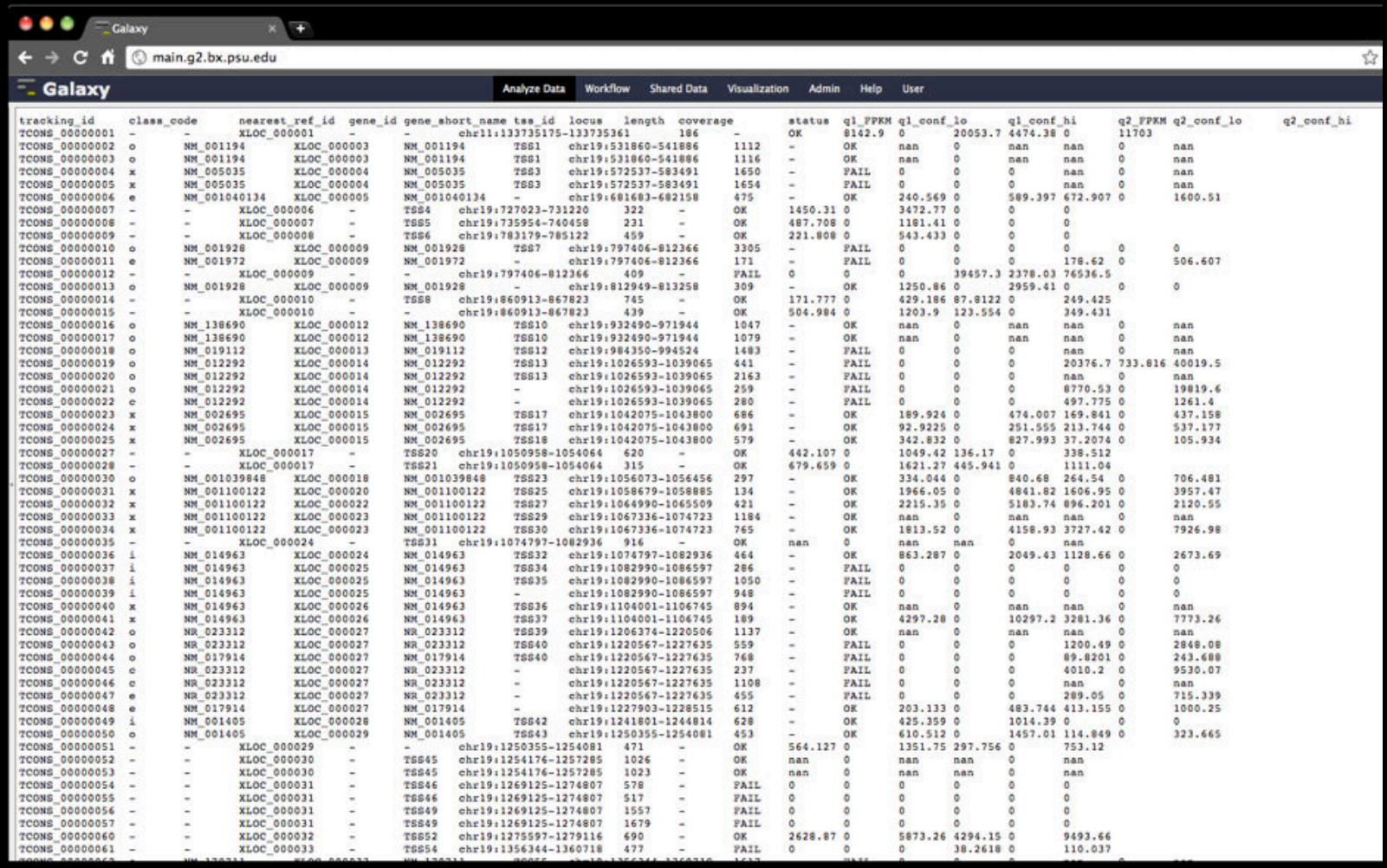
Read to transcript mapping difficult because read can map to multiple transcripts

FPKM = fragments per kilobase of exon per millions of reads

- normalize by exon length and sample size

1 FPKM ~ 1 transcript per cell in mouse

Quantitation and Differential Expression



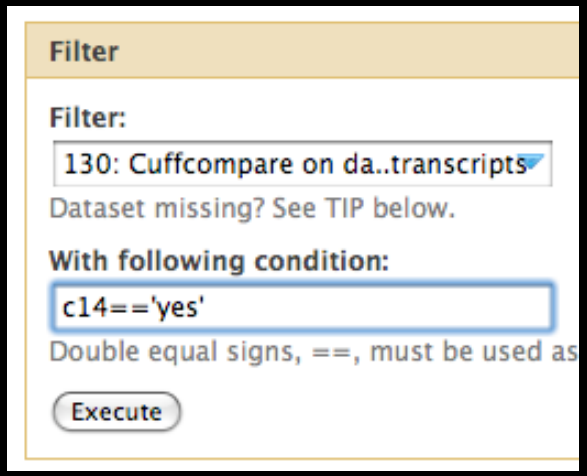
tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	status	q1_FPKM	q1_conf_lo	q1_conf_hi	q2_FPKM	q2_conf_lo	q2_conf_hi
TCONS_00000001	-	-	XLOC_000001	-	-	chr11:133735175-133735361	186	-	OK	8142.9	0	20053.7	4474.38	0	11703
TCONS_00000002	o	NM_001194	XLOC_000003	NM_001194	TSS1	chr19:531860-541886	1112	-	OK	nan	0	nan	nan	0	nan
TCONS_00000003	o	NM_001194	XLOC_000003	NM_001194	TSS1	chr19:531860-541886	1116	-	OK	nan	0	nan	nan	0	nan
TCONS_00000004	x	NM_005035	XLOC_000004	NM_005035	TSS3	chr19:572537-583491	1650	-	FAIL	0	0	0	nan	0	nan
TCONS_00000005	x	NM_005035	XLOC_000004	NM_005035	TSS3	chr19:572537-583491	1654	-	FAIL	0	0	0	nan	0	nan
TCONS_00000006	e	NM_001040134	XLOC_000005	NM_001040134	-	chr19:681683-682158	475	-	OK	240.569	0	589.397	672.907	0	1600.51
TCONS_00000007	-	-	XLOC_000006	-	TSS4	chr19:727023-731220	322	-	OK	1450.31	0	3472.77	0	0	
TCONS_00000008	-	-	XLOC_000007	-	TSS5	chr19:735954-740458	231	-	OK	487.708	0	1181.41	0	0	
TCONS_00000009	-	-	XLOC_000008	-	TSS6	chr19:783179-785122	459	-	OK	221.808	0	543.433	0	0	
TCONS_00000010	o	NM_001928	XLOC_000009	NM_001928	TSS7	chr19:797406-812366	3305	-	FAIL	0	0	0	0	0	0
TCONS_00000011	e	NM_001972	XLOC_000009	NM_001972	-	chr19:797406-812366	171	-	FAIL	0	0	0	178.62	0	506.607
TCONS_00000012	-	-	XLOC_000009	-	-	chr19:797406-812366	409	-	FAIL	0	0	0	39457.3	2378.03	76536.5
TCONS_00000013	o	NM_001928	XLOC_000009	NM_001928	-	chr19:812949-813258	309	-	OK	1250.86	0	2959.41	0	0	0
TCONS_00000014	-	-	XLOC_000010	-	TSS8	chr19:860913-867823	745	-	OK	171.777	0	429.186	87.8122	0	249.425
TCONS_00000015	-	-	XLOC_000010	-	-	chr19:860913-867823	439	-	OK	504.984	0	1203.9	123.554	0	349.431
TCONS_00000016	o	NM_138690	XLOC_000012	NM_138690	TSS10	chr19:932490-971944	1047	-	OK	nan	0	nan	nan	0	nan
TCONS_00000017	o	NM_138690	XLOC_000012	NM_138690	TSS10	chr19:932490-971944	1079	-	OK	nan	0	nan	nan	0	nan
TCONS_00000018	o	NM_019112	XLOC_000013	NM_019112	TSS12	chr19:984350-994524	1483	-	FAIL	0	0	0	nan	0	nan
TCONS_00000019	o	NM_012292	XLOC_000014	NM_012292	TSS13	chr19:1026593-1039065	441	-	FAIL	0	0	0	20376.7	733.816	40019.5
TCONS_00000020	o	NM_012292	XLOC_000014	NM_012292	TSS13	chr19:1026593-1039065	2163	-	FAIL	0	0	0	nan	0	nan
TCONS_00000021	o	NM_012292	XLOC_000014	NM_012292	-	chr19:1026593-1039065	259	-	FAIL	0	0	0	8770.53	0	19819.6
TCONS_00000022	c	NM_012292	XLOC_000014	NM_012292	-	chr19:1026593-1039065	280	-	FAIL	0	0	0	497.775	0	1261.4
TCONS_00000023	x	NM_002695	XLOC_000015	NM_002695	TSS17	chr19:1042075-1043800	686	-	OK	189.924	0	474.007	169.841	0	437.158
TCONS_00000024	x	NM_002695	XLOC_000015	NM_002695	TSS17	chr19:1042075-1043800	691	-	OK	92.9225	0	251.555	213.744	0	537.177
TCONS_00000025	x	NM_002695	XLOC_000015	NM_002695	TSS18	chr19:1042075-1043800	579	-	OK	342.832	0	827.993	37.2074	0	105.934
TCONS_00000027	-	-	XLOC_000017	-	TSS20	chr19:1050958-1054064	620	-	OK	442.107	0	1049.42	136.17	0	338.512
TCONS_00000028	-	-	XLOC_000017	-	TSS21	chr19:1050958-1054064	315	-	OK	679.659	0	1621.27	445.941	0	1111.04
TCONS_00000030	o	NM_001039848	XLOC_000018	NM_001039848	TSS23	chr19:1056073-1056456	297	-	OK	334.044	0	840.68	264.54	0	706.481
TCONS_00000031	x	NM_001100122	XLOC_000020	NM_001100122	TSS25	chr19:1058679-1058885	134	-	OK	1966.05	0	4841.82	1606.95	0	3957.47
TCONS_00000032	x	NM_001100122	XLOC_000022	NM_001100122	TSS27	chr19:1064990-1065509	421	-	OK	2215.35	0	5183.74	896.201	0	2120.55
TCONS_00000033	x	NM_001100122	XLOC_000023	NM_001100122	TSS29	chr19:1067336-1074723	1184	-	OK	nan	0	nan	nan	0	nan
TCONS_00000034	x	NM_001100122	XLOC_000023	NM_001100122	TSS30	chr19:1067336-1074723	765	-	OK	1813.52	0	4158.93	3727.42	0	7926.98
TCONS_00000035	-	-	XLOC_000024	-	TSS31	chr19:1074797-1082936	916	-	OK	nan	0	nan	nan	0	nan
TCONS_00000036	i	NM_014963	XLOC_000024	NM_014963	TSS32	chr19:1074797-1082936	464	-	OK	863.287	0	2049.43	1128.66	0	2673.69
TCONS_00000037	i	NM_014963	XLOC_000025	NM_014963	TSS34	chr19:1082990-1086597	286	-	FAIL	0	0	0	0	0	0
TCONS_00000038	i	NM_014963	XLOC_000025	NM_014963	TSS35	chr19:1082990-1086597	1050	-	FAIL	0	0	0	0	0	0
TCONS_00000039	i	NM_014963	XLOC_000025	NM_014963	-	chr19:1082990-1086597	948	-	FAIL	0	0	0	0	0	0
TCONS_00000040	x	NM_014963	XLOC_000026	NM_014963	TSS36	chr19:1104001-1106745	894	-	OK	nan	0	nan	nan	0	nan
TCONS_00000041	x	NM_014963	XLOC_000026	NM_014963	TSS37	chr19:1104001-1106745	189	-	OK	4297.28	0	10297.2	3281.36	0	7773.26
TCONS_00000042	o	NR_023312	XLOC_000027	NR_023312	TSS39	chr19:1206374-1220506	1137	-	OK	nan	0	nan	nan	0	nan
TCONS_00000043	o	NR_023312	XLOC_000027	NR_023312	TSS40	chr19:1220567-1227635	559	-	FAIL	0	0	0	1200.49	0	2848.08
TCONS_00000044	o	NM_017914	XLOC_000027	NM_017914	TSS40	chr19:1220567-1227635	768	-	FAIL	0	0	0	89.8201	0	243.688
TCONS_00000045	c	NR_023312	XLOC_000027	NR_023312	-	chr19:1220567-1227635	237	-	FAIL	0	0	0	4010.2	0	9530.07
TCONS_00000046	c	NR_023312	XLOC_000027	NR_023312	-	chr19:1220567-1227635	1108	-	FAIL	0	0	0	nan	0	nan
TCONS_00000047	e	NR_023312	XLOC_000027	NR_023312	-	chr19:1220567-1227635	455	-	FAIL	0	0	0	289.05	0	715.339
TCONS_00000048	e	NM_017914	XLOC_000027	NM_017914	-	chr19:1227903-1228515	612	-	OK	203.133	0	483.744	413.155	0	1000.25
TCONS_00000049	i	NM_001405	XLOC_000028	NM_001405	TSS42	chr19:1241801-1244814	628	-	OK	425.359	0	1014.39	0	0	0
TCONS_00000050	o	NM_001405	XLOC_000029	NM_001405	TSS43	chr19:1250355-1254081	453	-	OK	610.512	0	1457.01	114.849	0	323.665
TCONS_00000051	-	-	XLOC_000029	-	-	chr19:1250355-1254081	471	-	OK	564.127	0	1351.75	297.756	0	753.12
TCONS_00000052	-	-	XLOC_000030	-	TSS45	chr19:1254176-1257285	1026	-	OK	nan	0	nan	nan	0	nan
TCONS_00000053	-	-	XLOC_000030	-	TSS45	chr19:1254176-1257285	1023	-	OK	nan	0	nan	nan	0	nan
TCONS_00000054	-	-	XLOC_000031	-	TSS46	chr19:1269125-1274807	578	-	FAIL	0	0	0	0	0	0
TCONS_00000055	-	-	XLOC_000031	-	TSS46	chr19:1269125-1274807	517	-	FAIL	0	0	0	0	0	0
TCONS_00000056	-	-	XLOC_000031	-	TSS49	chr19:1269125-1274807	1557	-	FAIL	0	0	0	0	0	0
TCONS_00000057	-	-	XLOC_000031	-	TSS49	chr19:1269125-1274807	1679	-	FAIL	0	0	0	0	0	0
TCONS_00000060	-	-	XLOC_000032	-	TSS52	chr19:1275597-1279116	690	-	OK	2628.87	0	5873.26	4294.15	0	9493.66
TCONS_00000061	-	-	XLOC_000033	-	TSS54	chr19:1356344-1360718	477	-	FAIL	0	0	0	38.2618	0	110.037

Quantitation and Differential Expression

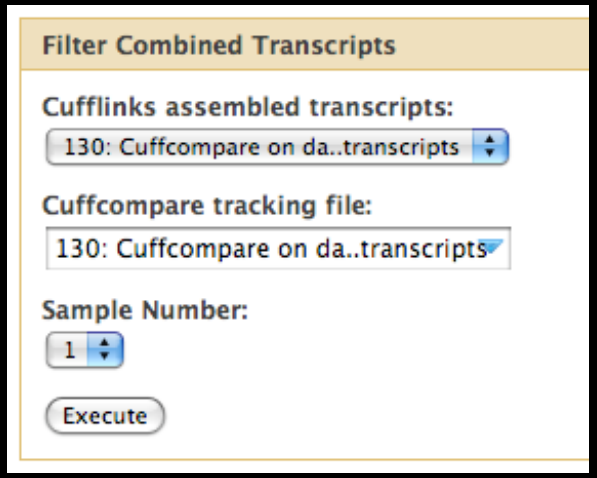
tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage		status	q1_FPKM	q1_conf_lo	q1_conf_hi	q2_FPKM	q2_conf_lo	q2_conf_hi
TCONS_00000001	-	-	XLOC_000001	-	-	chr11:133735175-133735361	186	-	OK	8142.9	0	20053.7	4474.38	0	11703	
TCONS_00000002	o	NM_001194	XLOC_000003	NM_001194	TSS1	chr19:531860-541886	1112	-	OK	nan	0	nan	nan	0	nan	
TCONS_00000003	o	NM_001194	XLOC_000003	NM_001194	TSS1	chr19:531860-541886	1116	-	OK	nan	0	nan	nan	0	nan	
TCONS_00000004	x	NM_005035	XLOC_000004	NM_005035	TSS3	chr19:572537-583491	1650	-	FAIL	0	0	0	nan	0	nan	
TCONS_00000005	x	NM_005035	XLOC_000004	NM_005035	TSS3	chr19:572537-583491	1654	-	FAIL	0	0	0	nan	0	nan	
TCONS_00000006	e	NM_001040134	XLOC_000005	NM_001040134	-	chr19:681683-682158	475	-	OK	240.569	0	589.397	672.907	0	1600.51	
TCONS_00000007	-	-	XLOC_000006	-	TSS4	chr19:727023-731220	322	-	OK	1450.31	0	3472.77	0	0	0	
TCONS_00000008	-	-	XLOC_000007	-	TSS5	chr19:735954-740458	231	-	OK	487.708	0	1181.41	0	0	0	
TCONS_00000009	-	-	XLOC_000008	-	TSS6	chr19:783179-785122	459	-	OK	221.808	0	543.433	0	0	0	
TCONS_00000010	o	NM_001928	XLOC_000009	NM_001928	TSS7	chr19:797406-812366	3305	-	FAIL	0	0	0	0	0	0	0
TCONS_00000011	e	NM_001972	XLOC_000009	NM_001972	-	chr19:797406-812366	171	-	FAIL	0	0	0	178.62	0	506.607	
TCONS_00000012	-	-	XLOC_000009	-	-	chr19:797406-812366	409	-	FAIL	0	0	0	39457.3	2378.03	76536.5	
TCONS_00000013	o	NM_001928	XLOC_000009	NM_001928	-	chr19:812949-813258	309	-	OK	1250.86	0	2959.41	0	0	0	
TCONS_00000014	-	-	XLOC_000010	-	TSS8	chr19:860812-867831	745	-	OK	171.777	0	428.186	87.8122	0	249.425	

Filtering

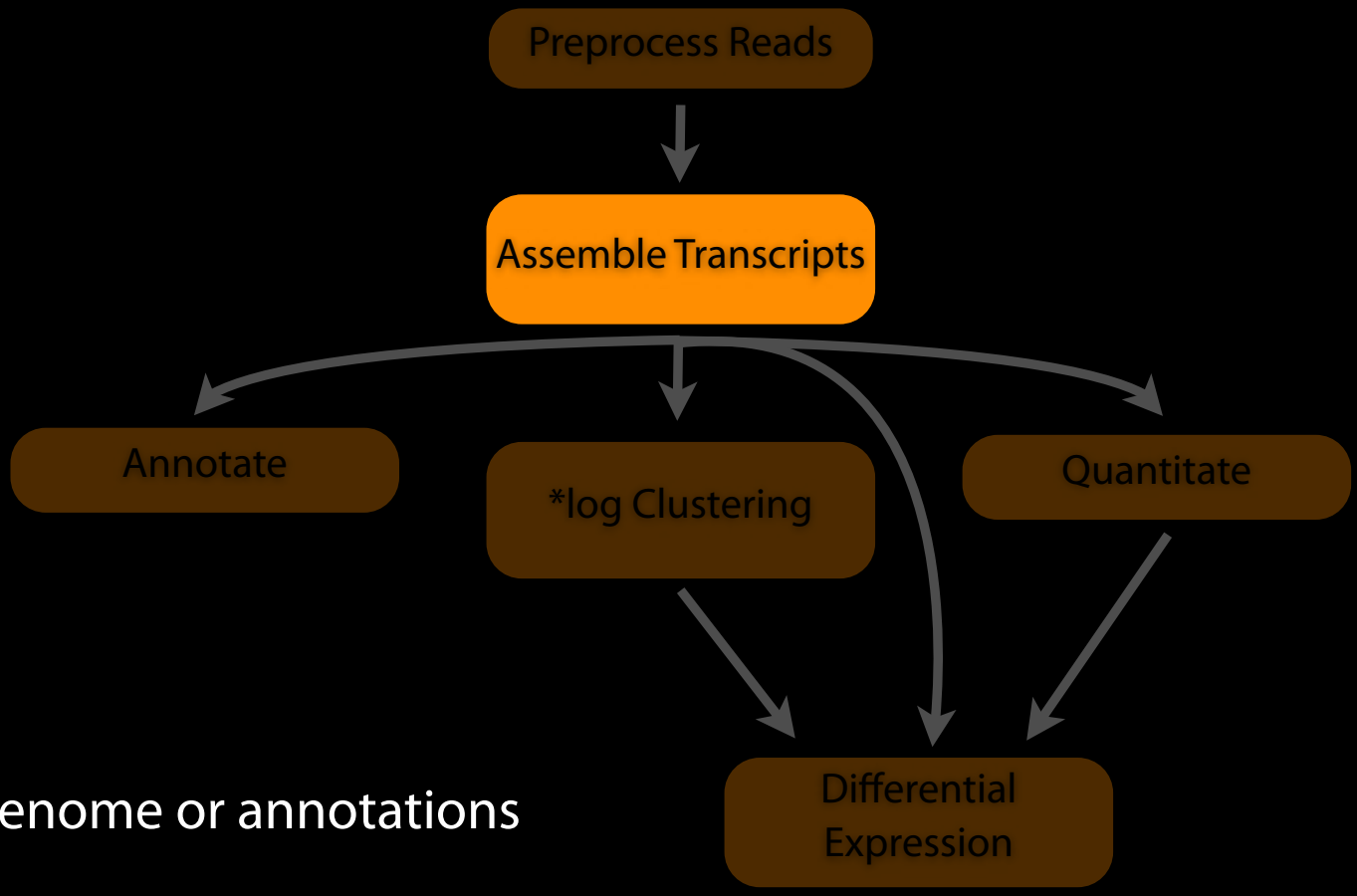
Filter for differentially expressed elements



Filter combined transcripts for those that are differentially expressed



Overview: *de novo*



no genome or annotations

De Novo Assembly

Assemble reads by looking for overlap amongst reads and contigs

- k = starting overlap between reads

Highly resource intensive

- lots of memory and multiple cores required
- difficult to do with desktop computer

Wrappers coming to Galaxy soon

De Novo RNA-seq Assemblers

	Trans-ABySS	Trinity
basic idea	combine multiple DNA assemblies	single assembly oriented toward RNA
runtime (24 cores)	7-9 days	1-2 days
disk space required	huge	small
scaffolding via PE reads	yes	no
strand-specific	no	yes

Robertson et al. (2010). "De novo assembly and analysis of RNA-seq data." Nature Methods 7(11): 909-912.
Grabherr et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol.

De Novo RNA-seq Assemblers

	Trans-ABySS	Trinity
basic idea	combine multiple DNA assemblies	single assembly oriented toward RNA
runtime (24 cores)	7-9 days	1-2 days
disk space required	huge	small
scaffolding via PE reads	yes	no
strand-specific	no	yes

Robertson et al. (2010). "De novo assembly and analysis of RNA-seq data." Nature Methods 7(11): 909-912.
Grabherr et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol.

De Novo RNA-seq Assemblers

	Trans-ABySS	Trinity
basic idea	combine multiple DNA assemblies	single assembly oriented toward RNA
runtime (24 cores)	7-9 days	1-2 days
disk space required	huge	small
scaffolding via PE reads	yes	no
strand-specific	no	yes

Robertson et al. (2010). "De novo assembly and analysis of RNA-seq data." Nature Methods 7(11): 909-912.
Grabherr et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol.

>comp17428_c0_seq1_FPKM_all:10.443_FPKM_rel:10.443_len:198_path:[0]
AGAAAACTTTTTTTTGGTAAAACAAAACAATTTTCTATTTTTGCAGTAAAATTTACGAAT
CATAATTTTGGGGGAAAATTTTTGATATTAATGTAAACTTGTCATAAGAGGGGAAAATTG
TGATACATTTTCCGCCATCTTGTGGGACAAGCAGAAAGTATTATTTGACATTTCGTAAAT
TTTACTGTGCGGCTTGTT

>comp16008_c0_seq1_FPKM_all:15.937_FPKM_rel:15.937_len:460_path:[0]
TTCGACAGAGACCAGCTCCAGCGAATTCGGCAAGATTTAAGTCGCAGAAACCGAGTTTTT
GGTAGGACCTGCCTCCCTTCAATTCCTTTCTAACTGATATCCTTAGATTACAGGGGTCTA
AGACTCCAGTAGATGCATTGGCACTCATTTTACAAAGGAAGTCAAATTTTGCATTCCACC
TAACAGTGTGGTCCTGCACTTCCATTCTTGTGGAATTGTCGGTGAAGGAGCCATCCAGTA
GTCGAAGTTTGGCAAAGAGTACAGCACTGACGAAAGGCACAGCAGTCAGCTCCTCGAGAA
CAATTTCCACGGAAAATTTATATTTCTTCTTTTTCATCATGAATGCCATTTCACTCAACC
CCGGACACTTCACTTTGATCGCGCTTTCGGAGCTCTTGAGCTCTCAGGCACTTCACTTTC
CAGAAAATAATAATTCTTCGTTTATCAAAAAAAAGAAGAG

>comp17416_c0_seq1_FPKM_all:13.273_FPKM_rel:13.273_len:642_path:[0]
TCCTGACAAACAACGCCCCCTGCTTCGTTTGGCGAACGTTTTGGTACAACTTTAACTCCA
CAATCCGCAGATCTTGATAAACAAAATCCAAGAAAAATAGGATCAGCAATTTTAACAAGA
ATATTGTCTACTGAAAATACATGTACACTATTTATTCCGCGCTTCTTCATATTATCCAAT
ACGCCTTGGATTTTTAATGCACGATAAAGTCCTCCATTTCCATCTGGTGCTTTTGATATG
TGACATTTTTCATCCAATATTATTTTACCATCAAAAGTAAAACAGGGTAGCATCTCTTGT
TTAAATGCTTTTATATTTTCTTTCTTTAATCCAAAATAATCATTTTTTTCAAAATATAAT
AAAGTTGTTTCGTGAGTTGCTTCACTTGTCATAATATACCATGTAATATGACCATGTTTT
TTGAATTTTTCTTCAGCAAGTTCTTGTAGTTTACGAATTCTTAAAGCTTGCAAACGAAAT
AGAGTGCTGTGTGAAGGAAGTCCAACATCATACATTCCCTTGGGATAATTTACACCAAGT
CGAGTACCTTGACCTCCAGCCAATAAAAGTATAGCGACTCGACCATCGGAAATTTCCTTT

Evaluating Assemblies

Expectations

- 80% reassembly = success

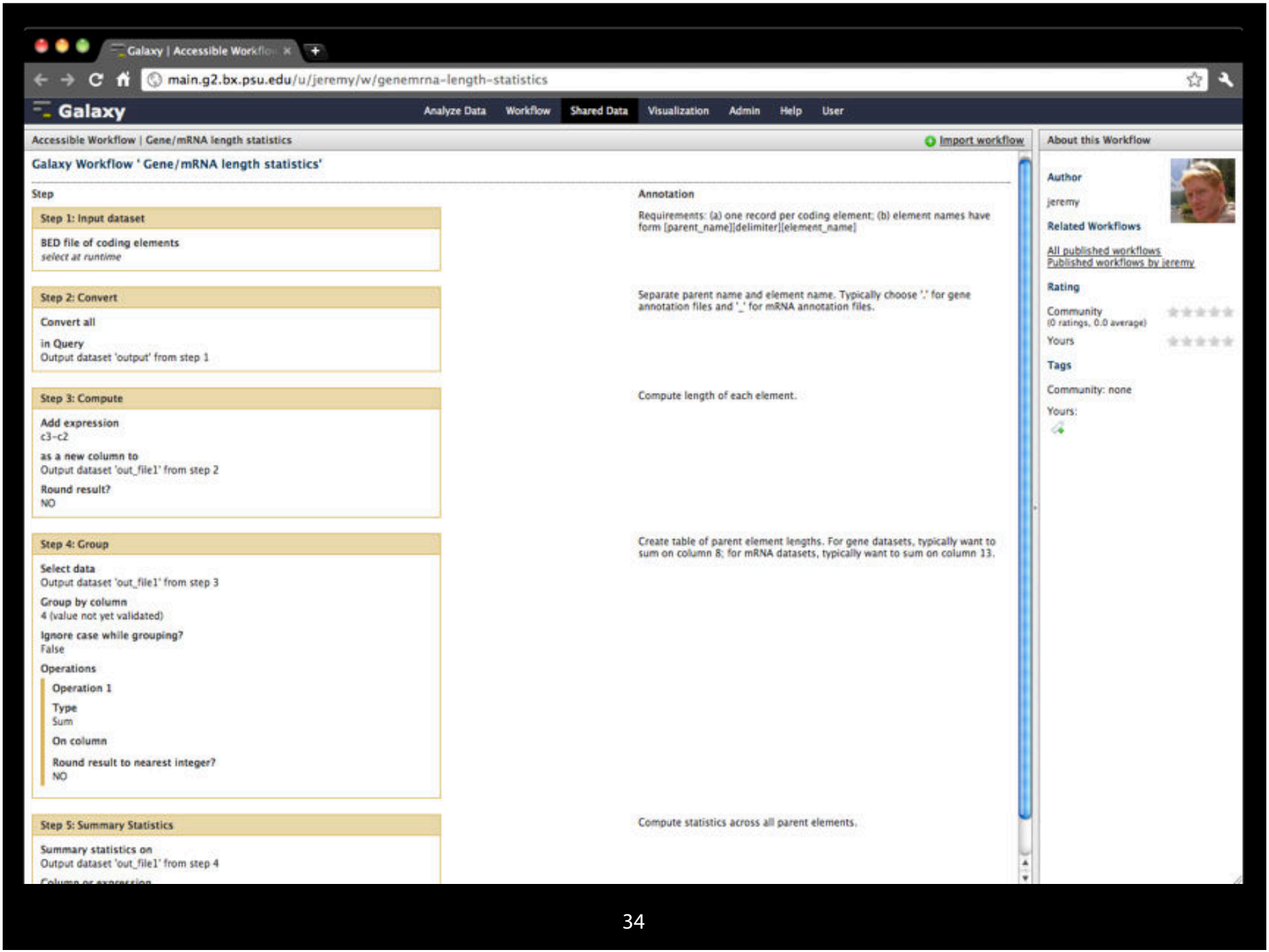
Statistics

n	n:100	n:N50	min	median	mean	N50	max	sum
650976	64671	6276	100	198	618	1744	17063	4.00E+07

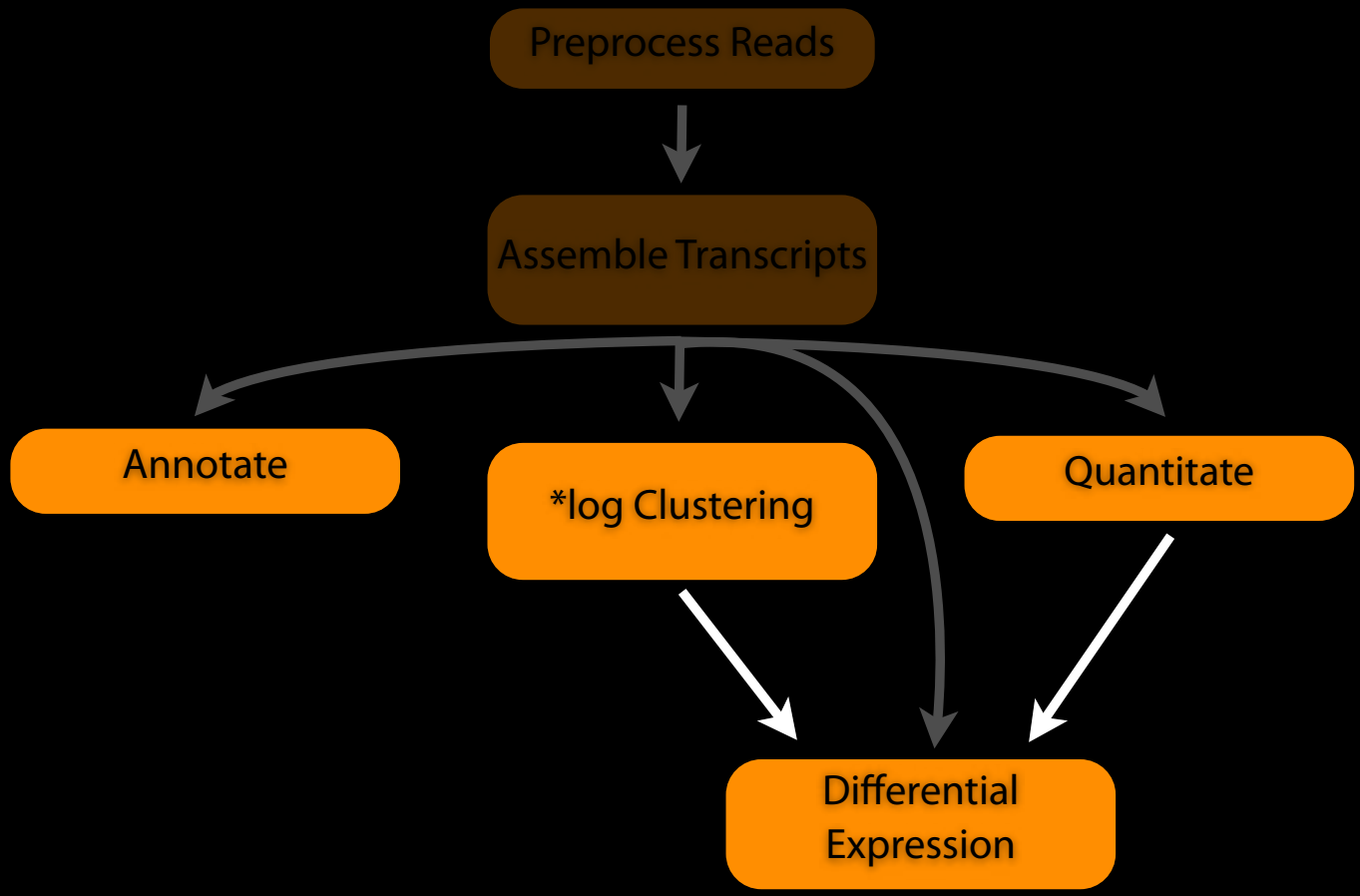
Transcript length distribution

hg19

mean	stdev	0%	25%	50%	75%	100%
1722	1352	30	832	1497	2208	81928

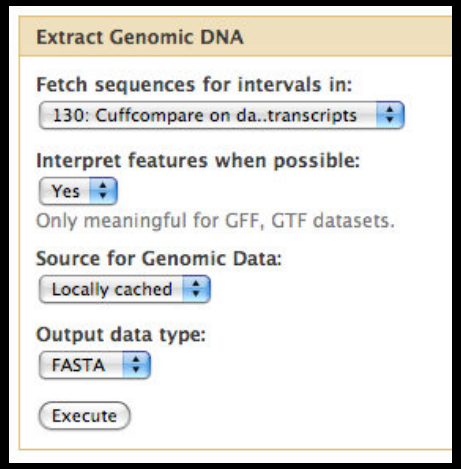

Galaxy | Accessible Workflow
main.g2.bx.psu.edu/u/jeremy/w/genemrna-length-statistics
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Accessible Workflow | Gene/mRNA length statistics
Import workflow
Galaxy Workflow ' Gene/mRNA length statistics'
Step
Annotation
Step 1: Input dataset
BED file of coding elements
select at runtime
Requirements: (a) one record per coding element; (b) element names have form [parent_name][delimiter][element_name]
Step 2: Convert
Convert all
in Query
Output dataset 'output' from step 1
Separate parent name and element name. Typically choose '.' for gene annotation files and '_' for mRNA annotation files.
Step 3: Compute
Add expression
c3-c2
as a new column to
Output dataset 'out_file1' from step 2
Round result?
NO
Compute length of each element.
Step 4: Group
Select data
Output dataset 'out_file1' from step 3
Group by column
4 (value not yet validated)
Ignore case while grouping?
False
Operations
Operation 1
Type
Sum
On column
Round result to nearest integer?
NO
Create table of parent element lengths. For gene datasets, typically want to sum on column 8; for mRNA datasets, typically want to sum on column 13.
Step 5: Summary Statistics
Summary statistics on
Output dataset 'out_file1' from step 4
Compute statistics across all parent elements.
About this Workflow
Author
jeremy
Related Workflows
All published workflows
Published workflows by jeremy
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community: none
Yours:

Overview

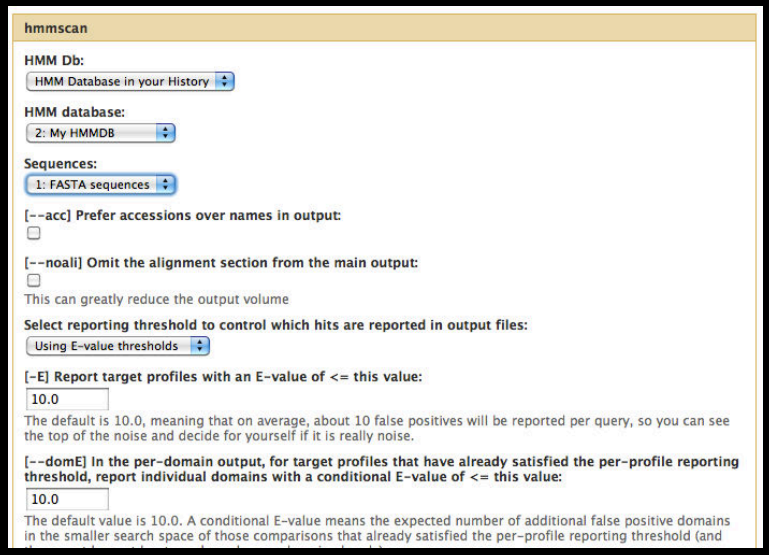

Preprocess Reads
Assemble Transcripts
Annotate
*log Clustering
Quantitate
Differential Expression

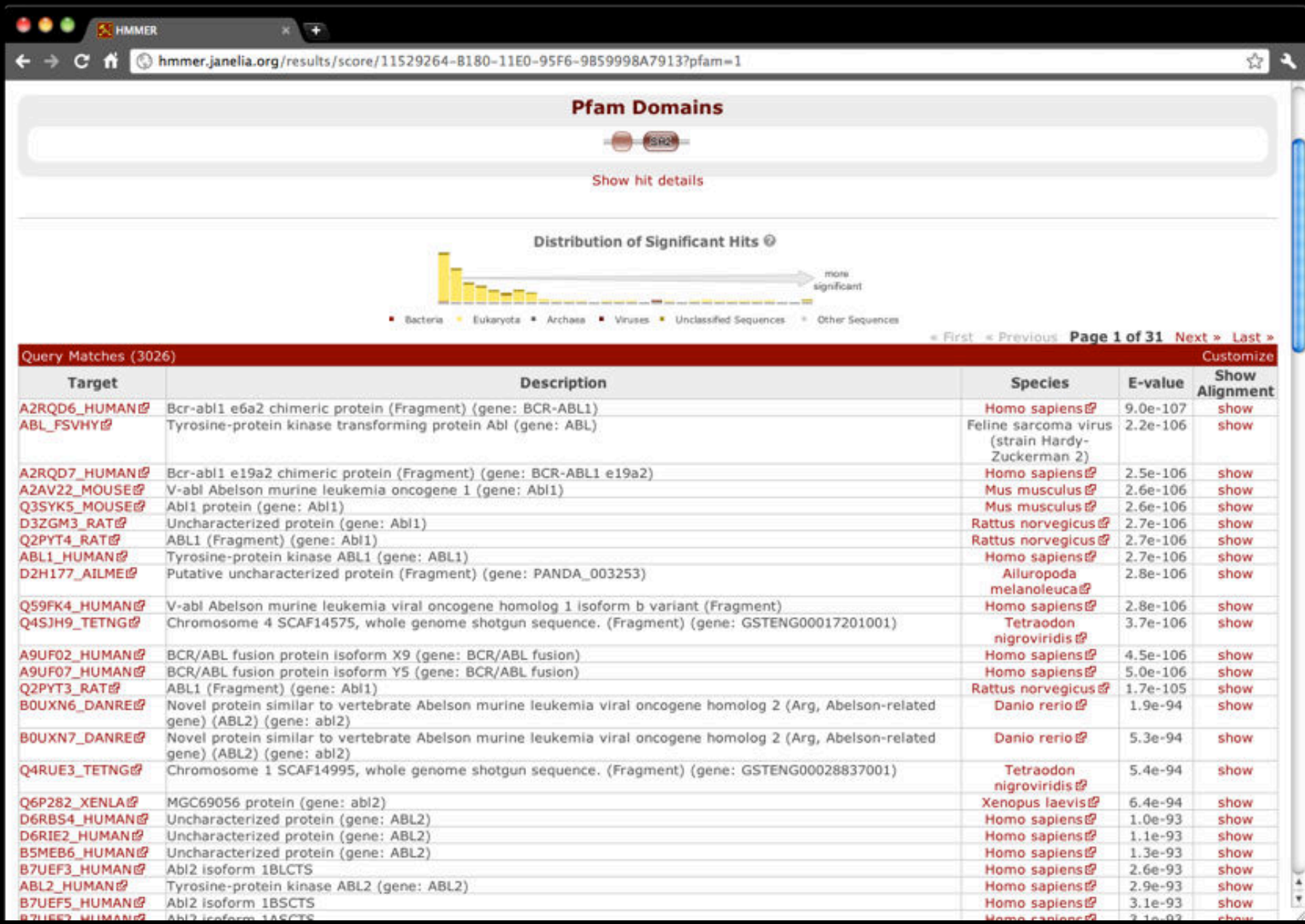
Sequence Annotation *de novo*

1. If necessary, extract genomic DNA
   - using GFF/features!



2. Annotation via HMMScan
   - PFAM database
   - http://hmmer.janelia.org/





Target	Description	Species	E-value	Show Alignment
A2RQD6_HUMAN	Bcr-abl1 e6a2 chimeric protein (Fragment) (gene: BCR-ABL1)	Homo sapiens	9.0e-107	show
ABL_FSVHY	Tyrosine-protein kinase transforming protein Abl (gene: ABL)	Feline sarcoma virus (strain Hardy-Zuckerman 2)	2.2e-106	show
A2RQD7_HUMAN	Bcr-abl1 e19a2 chimeric protein (Fragment) (gene: BCR-ABL1 e19a2)	Homo sapiens	2.5e-106	show
A2AV22_MOUSE	V-abl Abelson murine leukemia oncogene 1 (gene: Abl1)	Mus musculus	2.6e-106	show
Q3SYK5_MOUSE	Abl1 protein (gene: Abl1)	Mus musculus	2.6e-106	show
D3ZGM3_RAT	Uncharacterized protein (gene: Abl1)	Rattus norvegicus	2.7e-106	show
Q2PYT4_RAT	ABL1 (Fragment) (gene: Abl1)	Rattus norvegicus	2.7e-106	show
ABL1_HUMAN	Tyrosine-protein kinase ABL1 (gene: ABL1)	Homo sapiens	2.7e-106	show
D2H177_AILME	Putative uncharacterized protein (Fragment) (gene: PANDA_003253)	Ailuropoda melanoleuca	2.8e-106	show
Q59FK4_HUMAN	V-abl Abelson murine leukemia viral oncogene homolog 1 isoform b variant (Fragment)	Homo sapiens	2.8e-106	show
Q4SJH9_TETNG	Chromosome 4 SCAF14575, whole genome shotgun sequence. (Fragment) (gene: GSTENG00017201001)	Tetraodon nigroviridis	3.7e-106	show
A9UF02_HUMAN	BCR/ABL fusion protein isoform X9 (gene: BCR/ABL fusion)	Homo sapiens	4.5e-106	show
A9UF07_HUMAN	BCR/ABL fusion protein isoform Y5 (gene: BCR/ABL fusion)	Homo sapiens	5.0e-106	show
Q2PYT3_RAT	ABL1 (Fragment) (gene: Abl1)	Rattus norvegicus	1.7e-105	show
B0UXN6_DANRE	Novel protein similar to vertebrate Abelson murine leukemia viral oncogene homolog 2 (Arg, Abelson-related gene) (ABL2) (gene: abl2)	Danio rerio	1.9e-94	show
B0UXN7_DANRE	Novel protein similar to vertebrate Abelson murine leukemia viral oncogene homolog 2 (Arg, Abelson-related gene) (ABL2) (gene: abl2)	Danio rerio	5.3e-94	show
Q4RUE3_TETNG	Chromosome 1 SCAF14995, whole genome shotgun sequence. (Fragment) (gene: GSTENG00028837001)	Tetraodon nigroviridis	5.4e-94	show
Q6P282_XENLA	MGC69056 protein (gene: abl2)	Xenopus laevis	6.4e-94	show
D6RBS4_HUMAN	Uncharacterized protein (gene: ABL2)	Homo sapiens	1.0e-93	show
D6RIE2_HUMAN	Uncharacterized protein (gene: ABL2)	Homo sapiens	1.1e-93	show
B5MEB6_HUMAN	Uncharacterized protein (gene: ABL2)	Homo sapiens	1.3e-93	show
B7UEF3_HUMAN	Abl2 isoform 1BLCTS	Homo sapiens	2.6e-93	show
ABL2_HUMAN	Tyrosine-protein kinase ABL2 (gene: ABL2)	Homo sapiens	2.9e-93	show
B7UEF5_HUMAN	Abl2 isoform 1BSCTS	Homo sapiens	3.1e-93	show

*log Clustering *de novo*

(Quasi-)Reciprocal BLA(s)T

- find best matches between two sets of transcripts
- repeat for more species

LastZ

- designed for long reads/contigs
- high-quality mapping amongst *log clusters

Quantitation *de novo*

Read to transcript mapping difficult because read can map to multiple transcripts

FPKM = fragments per kilobase of exon per millions of reads

- normalize by exon length and sample size

1 FPKM ~ 1 transcript per cell in mouse

RSEM *de novo*

Expectation-maximization model

Uses TPM, not FPKM

- when mean transcript length is 1k, 1 TPM=1 FPKM

Be consistent

Li, B., V. Ruotti, et al. (2010). "RNA-Seq gene expression estimation with read mapping uncertainty." Bioinformatics 26(4): 493-500.

```
k71:94571     34285.52    0.00129541301452086    k71:94571
k70:110756u   34714.33    0.00147532810537204    k70:110756u
k72:79128u    35163.91    0.00139980345826473    k72:79128u
k71:95729u    36039.45    0.00164266197886051    k71:95729u
k73:62819u    36888.41    0.0018459751145641     k73:62819u
k73:62364u    37274.56    0.00180992945323558    k73:62364u
k73:63082u    37919.99    0.00136337550194011    k73:63082u
k72:78714     38233.64    0.00237886076778222    k72:78714
k73:62636     38362.56    0.0003752718439721 68  k73:62636
k71:95841     39041.42    0.00189572242891473    k71:95841
k73:63028u    39759.97    0.0015827631212604     k73:63028u
k72:78108     40155.68    0.00184053788631075    k72:78108
k71:93994     40967.55    0.00177806349073446    k71:93994
k74:38786u    44067.87    0.00220525090976878    k74:38786u
k72:78714     38233.64    0.00237886076778222    k72:78714
k73:62636     38362.56    0.0003752718439721 68  k73:62636
k71:95841     39041.42    0.00189572242891473    k71:95841
k73:63028u    39759.97    0.0015827631212604     k73:63028u
k72:78108     40155.68    0.00184053788631075    k72:78108
k71:93994     40967.55    0.00177806349073446    k71:93994
```

Differential Expression *de novo*

Simple: combine results from *log clustering and quantitation

Summary

For each organism

- assembled transcripts, quantitated and annotated

For pairs/groups of organisms

- *log clusters
- differential expression amongst genes and/or *logs

EMORY

Enis Afgan

Dannon Baker

Dave Clements

Jeremy Goecks

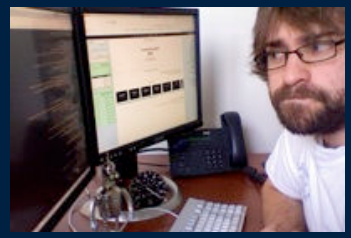
Kanwei Li

James Taylor

PENNSTATE

1855

Dan Blankenberg



Nate Coraor

Jennifer Jackson

Greg von Kuster

Kelly Vincent

Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

Thanks! Questions?

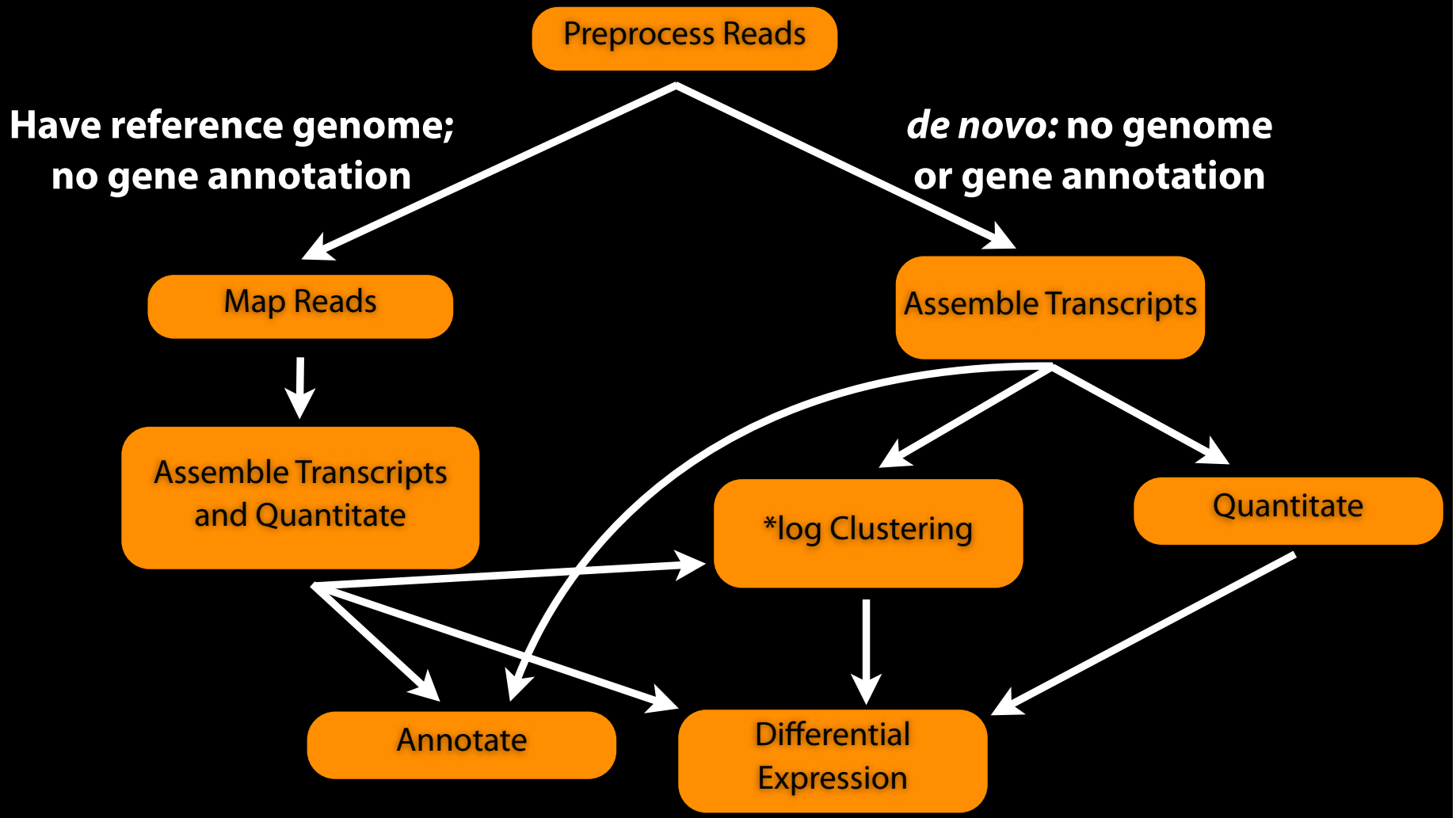
Slides: http://dl.dropbox.com/u/4745528/goecks_rnaseq_workshop_2011.pdf

Galaxy

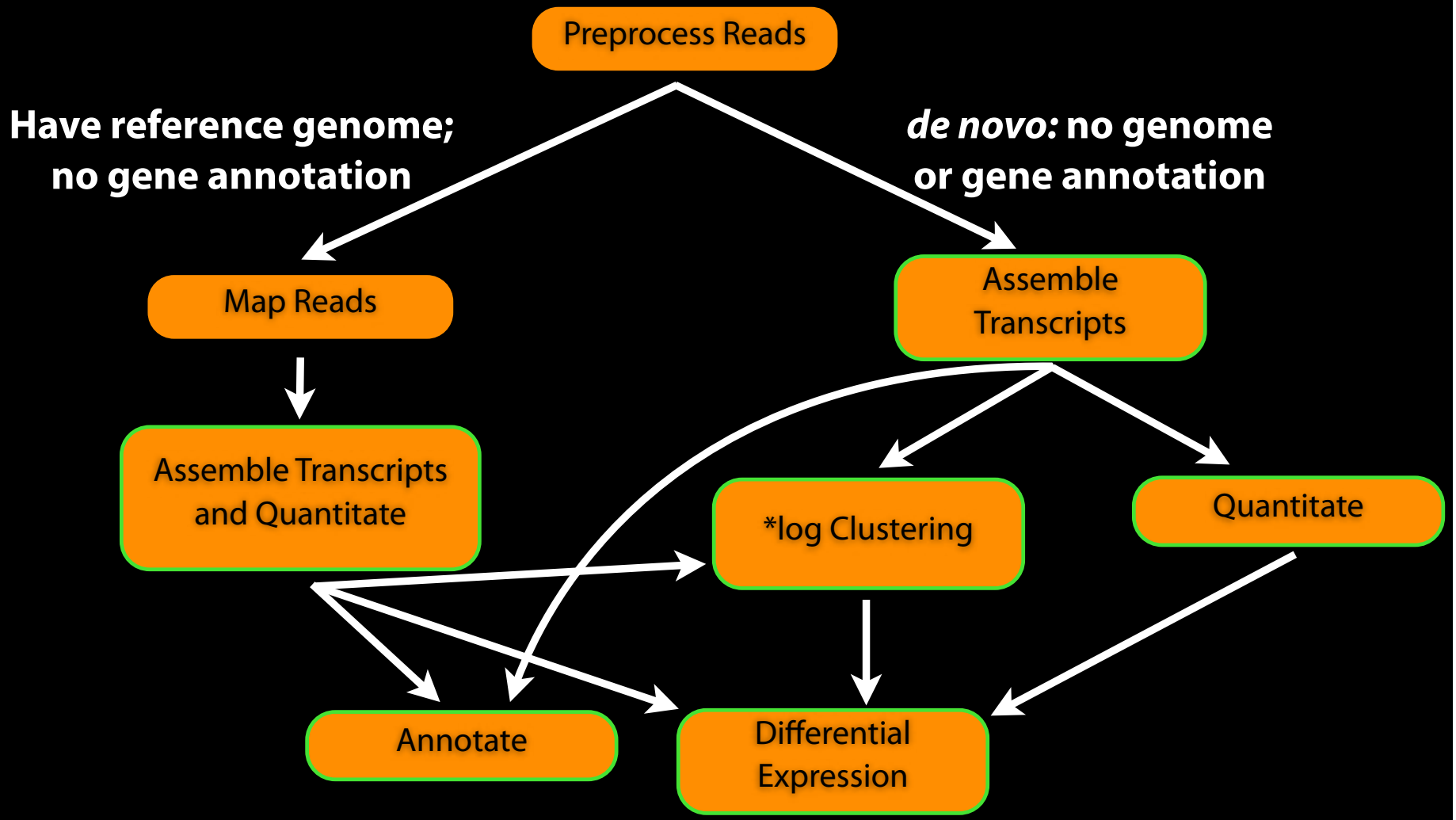
- public server: http://usegalaxy.org
- download: http://getgalaxy.org

jeremy.goecks@emory.edu

Overview


Preprocess Reads
Have reference genome; no gene annotation
de novo: no genome or gene annotation
Map Reads
Assemble Transcripts
Assemble Transcripts and Quantitate
*log Clustering
Quantitate
Annotate
Differential Expression

Overview


Preprocess Reads
Have reference genome; no gene annotation
de novo: no genome or gene annotation
Map Reads
Assemble Transcripts
Assemble Transcripts and Quantitate
*log Clustering
Quantitate
Annotate
Differential Expression