Next Generation Sequencing Software for Data Management, Analysis, and Visualization

Session W14

Tufts University

Tufts University Core Facility

Tools for Next Generation Sequencing Data Analysis

Kip Lord Bodi

Genomics Core Director

Tufts University Core Facility

genomics.med.tufts.edu

Tufts UNIVERSITY

T U C F Tufts University Core Facility

Tufts University Core Facility

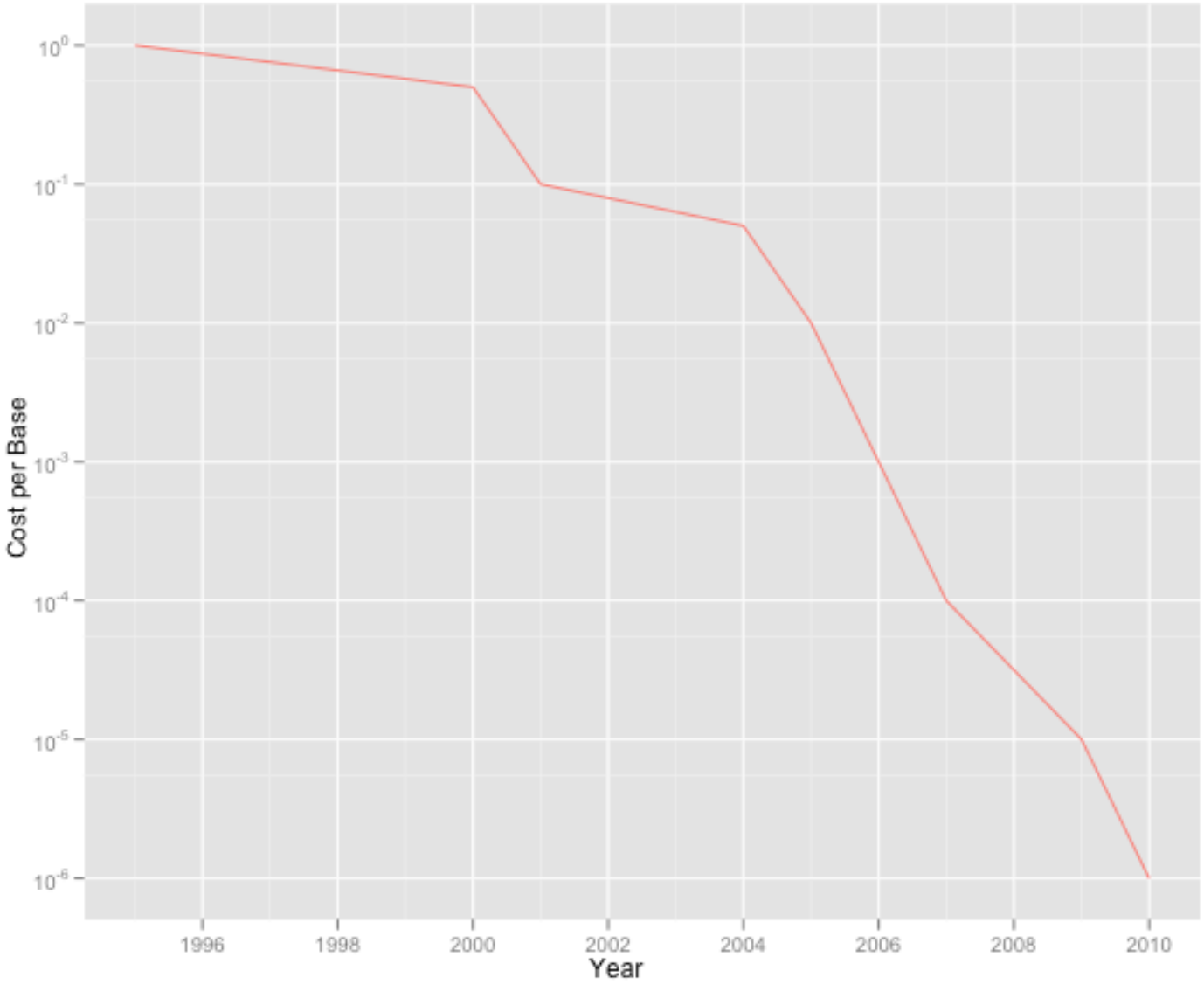
- Established 1987
- Services:
  - DNA Synthesis and Sequencing
  - Peptide Synthesis and Protein Sequencing
  - **Next-Generation Sequencing (2008)**
    - Illumina Genome Analyzer IIx
      - (HiSeq 2000 arriving March 2011)
    - Roche 454 GS FLX
    - http://genomics.med.tufts.edu

NGS: Major Applications at Tufts

- Whole Genome Sequencing
  - Up to 12 microbial genomes per lane on a GAIIx
- Probe genetic networks (**Tn-Seq**)
- Transcriptome Profiling (RNA-Seq)
- Chromatin-IP Sequencing (ChIP-Seq)
  - Including 'pulldown' experiments
- Metagenomics

TUCF Next-Gen Sequencing Services

- Consult and Design Experiments
- Sample Preparation
  - Genomic DNA, RNA
  - Others: Will assist
- Actual Sequencing
  - Single End, Paired End, 40 - 150 Bases
- Data Management
- **Bioinformatics Analysis**



Tufts University Core Facility

Illumina Genome Analyzer IIx

- Short Read Technology
  - 40-150 Bases
  - Paired-End also available
- 240M Reads / Run
  - 30M Reads / Lane

Tufts University Core Facility

Run Statistics

PERFORMANCE PARAMETERS*

READ LENGTH	RUN TIME (DAYS)	CLUSTERS PASSING FILTER	OUTPUT (GB)	THROUGHPUT (GB/DAY)	BASE CALLS WITH Q ≥30	RAW READ ACCURACY	% PERFECT READS
1 × 35 bp	~2	225–250 million	8.0–9.0	~4.0–4.5	75–90%	≥ 99%	≥ 90%
2 × 35 bp	~4	225–250 million	16.0–18.0	~4.0–4.5	75–90%	≥ 99%	≥ 90%
2 × 50 bp	~5	225–250 million	22.5–25.0	~4.5–5.0	75–90%	≥ 99%	≥ 85%
2 × 75 bp	~7.5	225–250 million	34.0–38.0	~4.5–5.0	70–85%	≥ 98.5%	≥ 80%
2 × 100 bp	~9.5	225–250 million	45.0–50.0	~4.75–5.25	≥ 70%	≥ 98%	≥ 70%

SAMPLES

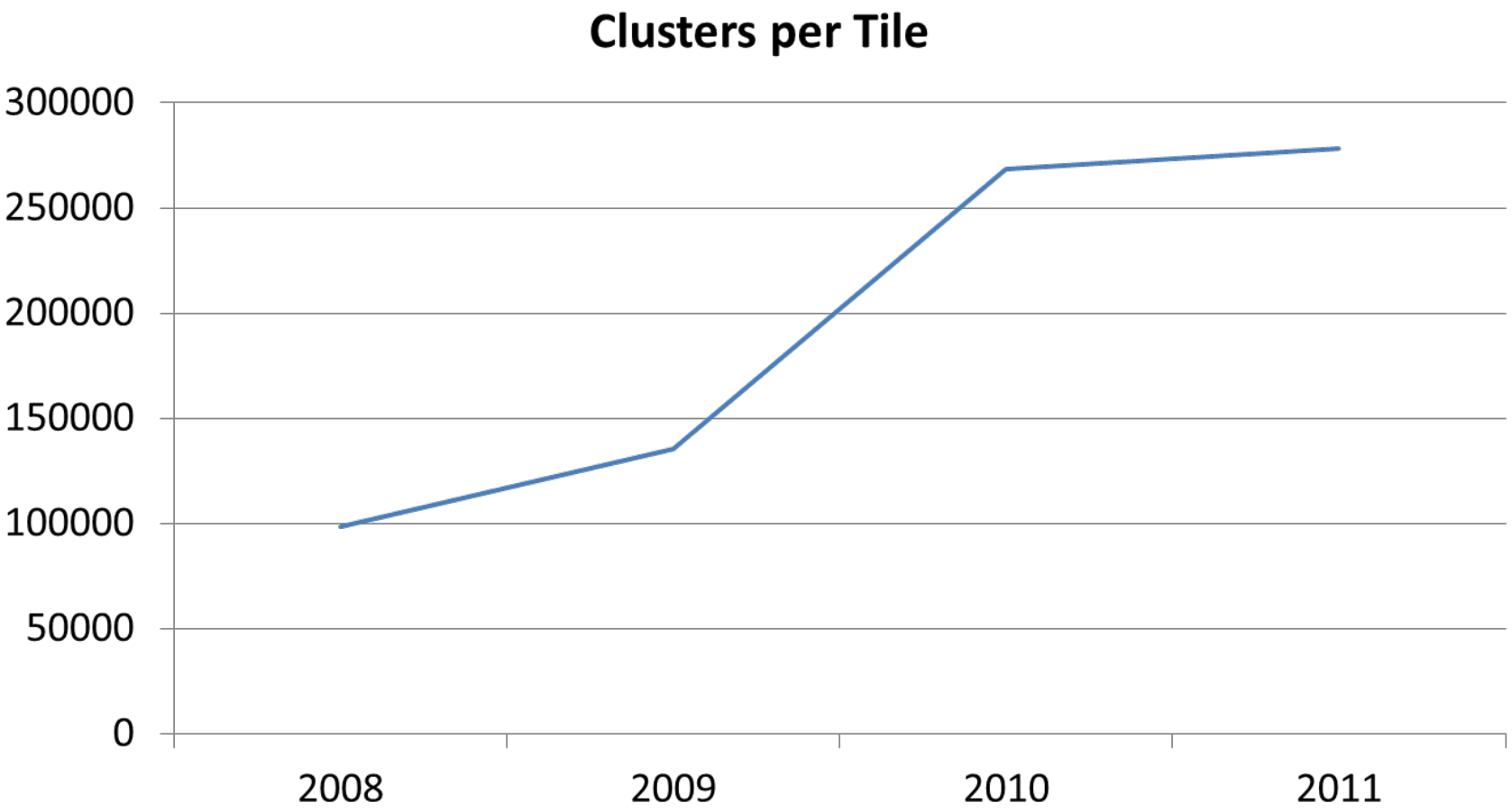
Throughput: eight channels per flow cell, up to 12 samples per channel using Illumina Multiplexing Reagents
Input requirement: 0.1–1.0 µg (single- and paired-end reads), 10 µg (Mate Pair reads)
Genomic DNA sample prep: Three hours hands-on, six hours total for single or paired-end libraries

SERVICE AND SUPPORT

Illumina will ensure that your Genome Analyzer system is properly installed and qualified, and will provide ongoing maintenance and service. This industry-leading support is available in North America, Europe, and Asia.

* gDNA sequencing output generated with cluster densities between 235,000 and 260,000 clusters per tile that pass filtering on a Genome Analyzer IIx using Sequencing Control Software (SCS) 2.6, SBS v4 kits and Cluster Generation v4 kits. Data were analyzed using Pipeline software v1.6.

GAIIx advances: Reagents alone



Tufts University Core Facility

Informatics Crisis

- Many life science researchers may not be able to make use of their data
- Most software is Unix/Linux based
  - Bowtie, MAQ, Samtools
  - Tophat, Cufflinks
  - MACS, ERANGE Peak Caller
- **Many analyses do not fit these tools**
  - Will require custom script or software to be written

Options for Analysis

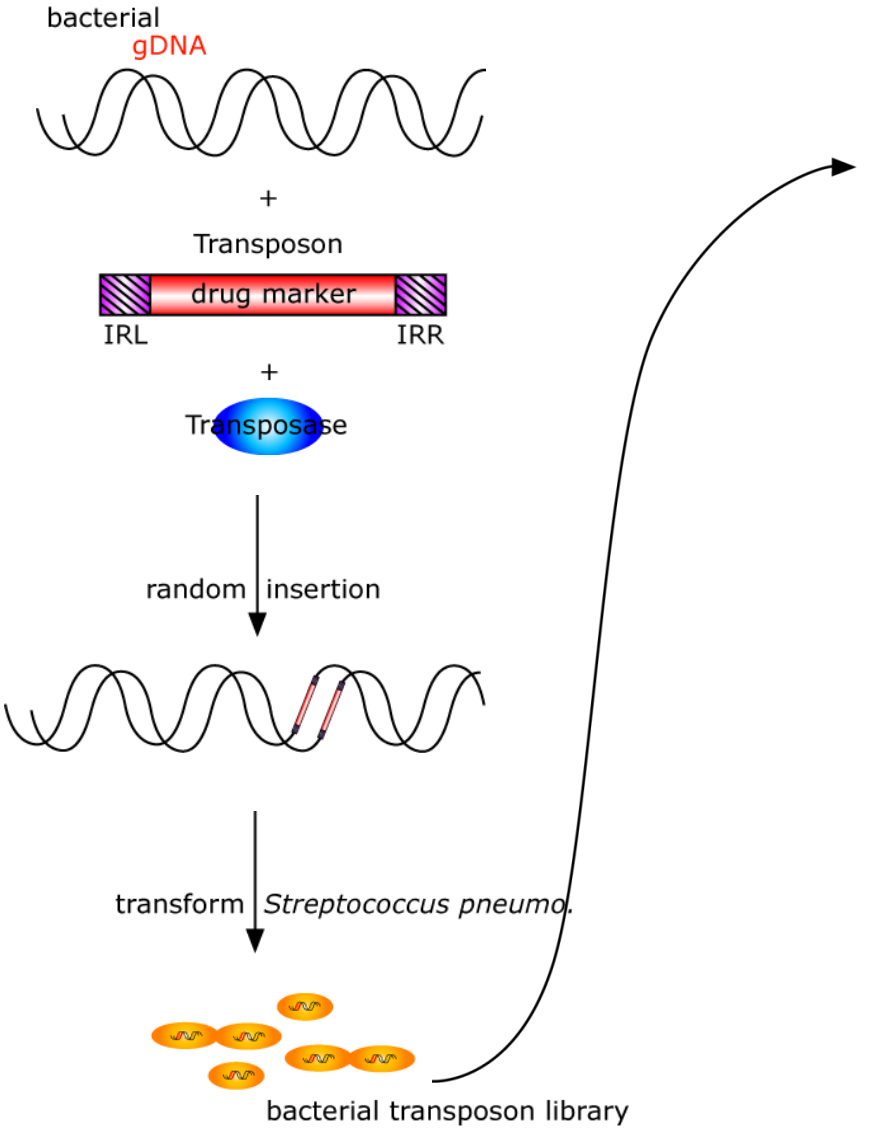
- Dedicated bioinformatician
- Bioinformatics core facility
- Commercial software
  - CLC Genomics Workbench
  - GenomeQuest
- The ambitious postdoc
  - Willing to learn
  - And has the time

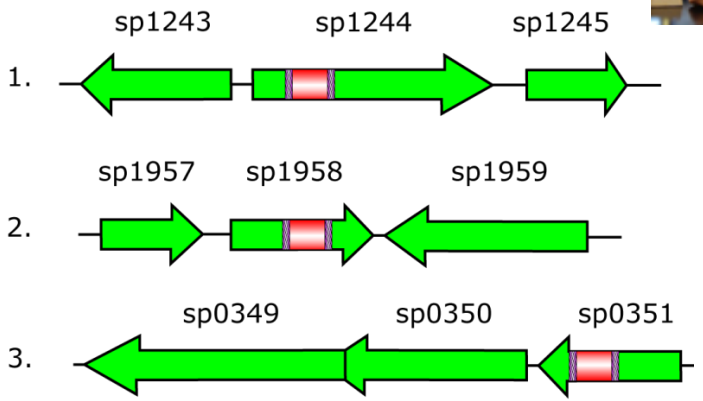
Back to Custom Analyses: Example

- Department of Molecular and Microbiology
- Camilli Laboratory at TUSM
- Studies two microbial pathogens
  - Vibrio cholerae
  - Streptococcus pneumoniae
- Interested in gene networks and function in bacteria
- Many genes are not annotated

quantitative Tn-seq approach

- make a large random insertion library -





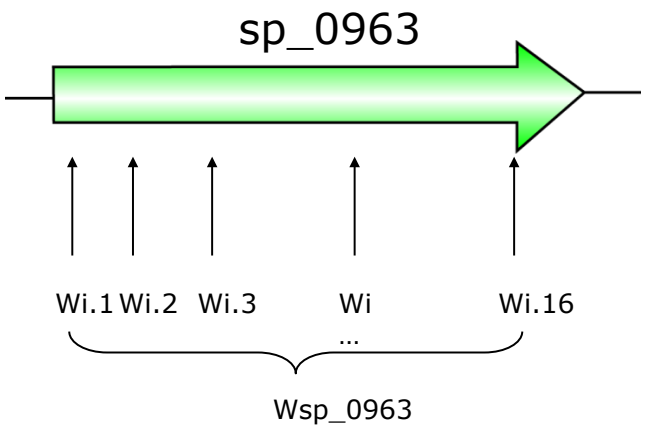
Tn library ready for selection

Tufts University Core Facility

quantitative Tn-seq approach

- composite fitness -

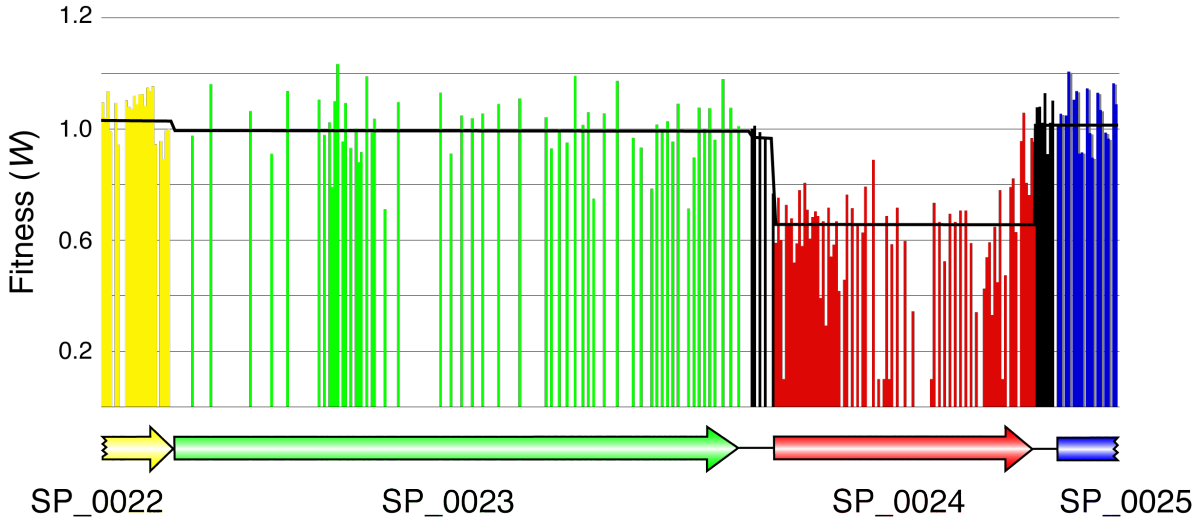
a library consisting of 30,000 insertions on average has 16 insertions per 1kb



thus fitness for each gene is composed of multiple fitness values

quantitative Tn-seq approach

- composite fitness -

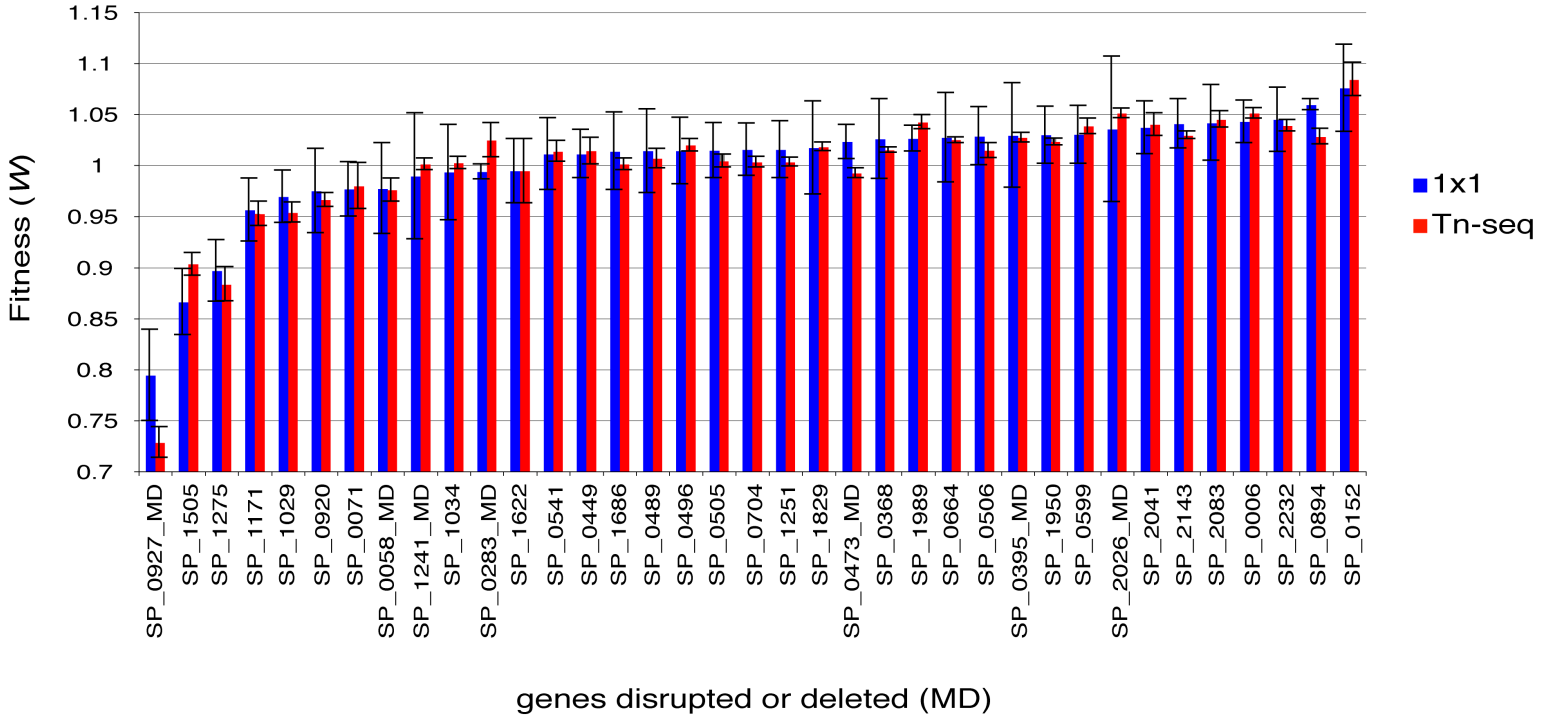


Tufts University Core Facility

quantitative Tn-seq approach

- Tn-seq data vs. 1x1 competition -

7 marked deletions and 30 random Tn insertion colonies were competed against Wt

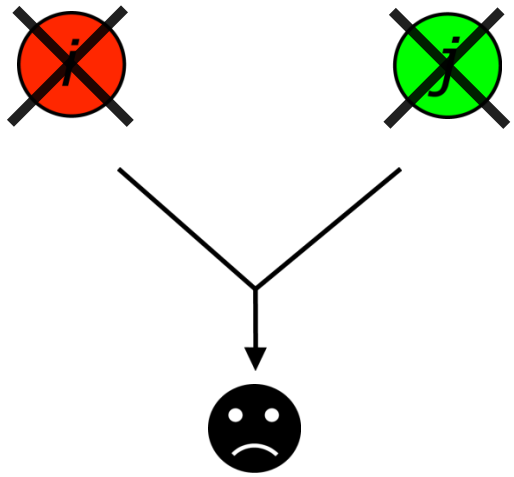


Tufts University Core Facility

genetic interaction

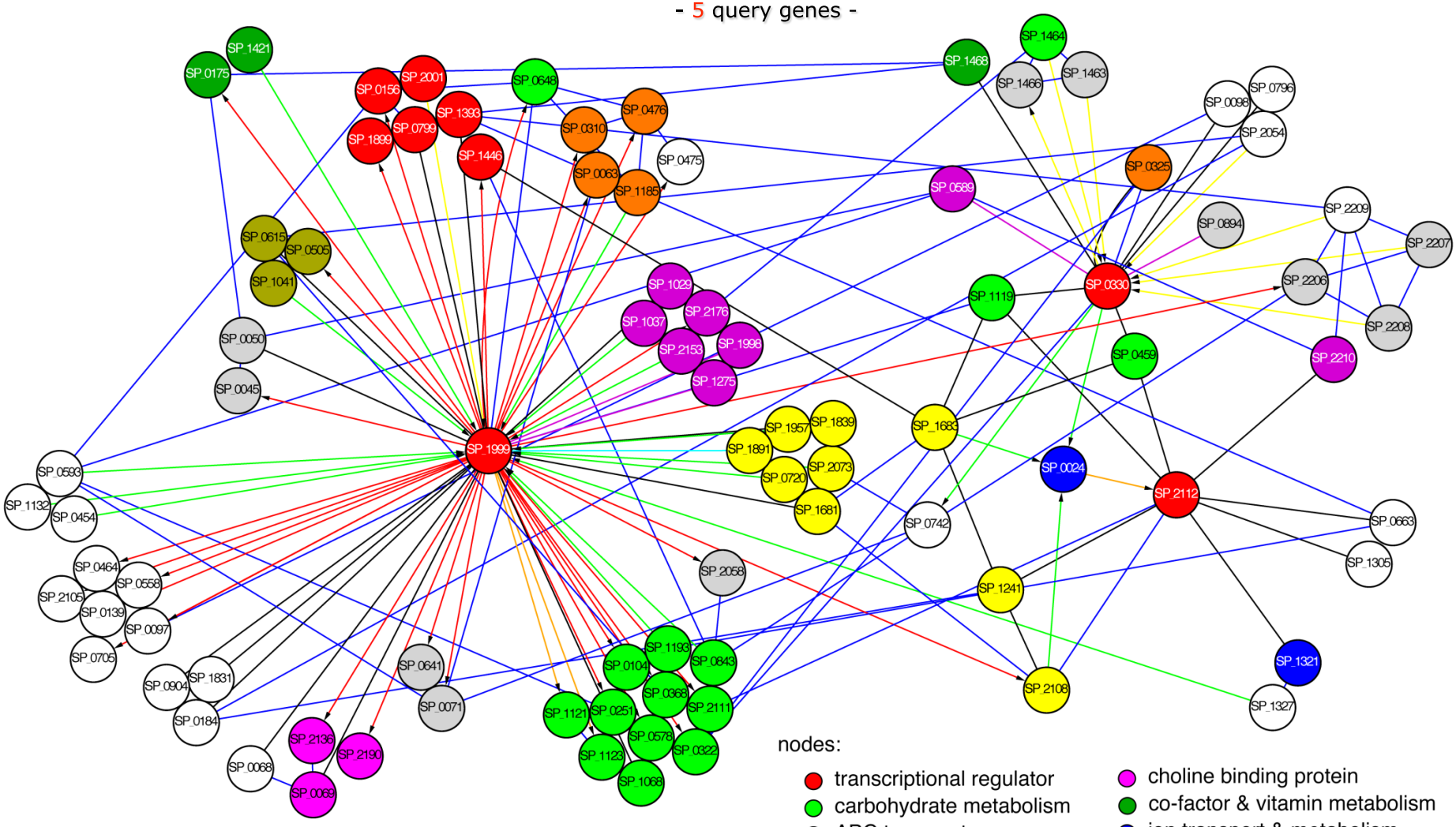
- deviation from a multiplicative model -

genetic interaction: $Wij \neq Wi \times Wj$

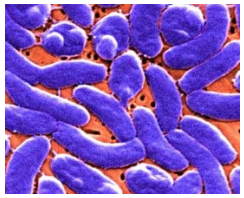


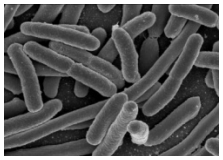
genetic interaction network

- 5 query genes -



Tufts University Core Facility

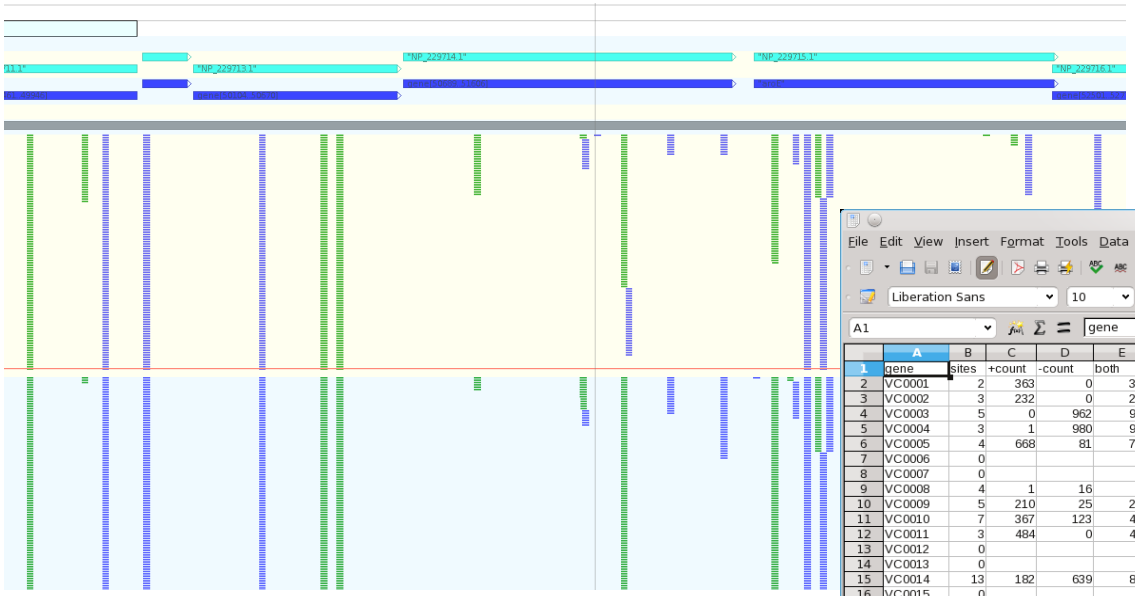




Tufts University Core Facility

Scene Missing

?



A0.0.agghop.txt - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

Liberation Sans 10

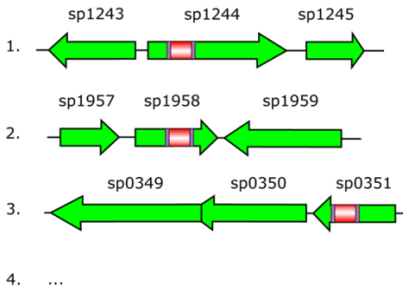
A1 gene

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	gene	sites	+count	-count	both	start	end	strand	length	read_count	expected	dval_genome	dval_genic
2	VC0001	2	363	0	363	235	402	-1	167	363	107.4194493864	3.3792763049	5.5689910298
3	VC0002	3	232	0	232	372	806	-1	434	232	279.1619223576	0.8310588996	1.3695712157
4	VC0003	5	0	962	962	816	2210	-1	1394	962	896.6629487706	1.072866902	1.7680667736
5	VC0004	3	1	980	981	2271	3896	-1	1625	981	1045.2491332512	0.9385322301	1.5466854731
6	VC0005	4	668	81	749	3899	4156	-1	257	749	165.3101706127	4.5308766982	7.4668092847
7	VC0006	0				4123	4479	-1	356				
8	VC0007	0				4492	4629	-1	137				
9	VC0008	4	1	16	17	4871	5608	-1	737	17	474.0606838192	0.0358603879	0.0590973216
10	VC0009	5	210	25	235	5605	6276	-1	671	235	431.6074882533	0.5444761882	0.8972876836
11	VC0010	7	367	123	490	6354	7100	-1	746	490	479.8497559418	1.0211529628	1.6828430637
12	VC0011	3	484	0	484	7224	7352	1	128	484	82.3334701884	5.8785327388	9.6877239789
13	VC0012	0				7397	8815	1	1418				
14	VC0013	0				8873	9973	1	1100				
15	VC0014	13	182	639	821	9991	11082	1	1091	821	701.764187309	1.16990866	1.9279899734
16	VC0015	0				11103	13520	1	2417				
17	VC0016	3	13	0	13	13876	14367	1	491	13	315.8260458008	0.0411618996	0.0678341246
18	VC0017	0				14377	14493	1	116				
19	VC0018	2	94	0	94	14769	15206	1	437	94	281.0916130651	0.3344105467	0.5511030075
20	VC0019	6	238	9	247	15294	16550	-1	1256	247	807.8971762237	0.3057319759	0.5038412008
21	VC0020	0				16869	18935	-1	2066				
22	VC0021	1	116	0	116	18950	19942	-1	992	116	638.0843939601	0.1817941343	0.2995937034
23	VC0022	3	425	76	501	20235	20837	1	602	501	387.2246019798	1.2938227515	2.132198331
24	VC0023	4	446	0	446	20943	21200	1	257	446	165.3101706127	2.6979586213	4.4461908424
25	VC0024	0				21287	21535	-1	248				
26	VC0025	5	141	411	552	21656	21856	1	200	552	128.6460471694	4.2908430702	7.0712378662
27	VC0026	5	545	497	1042	22028	23008	-1	980	1042	630.36563113	1.6530089024	2.7241311212
28	VC0027	6	450	1105	1555	23256	24788	-1	1532	1555	985.4287213175	1.5779933813	2.6005067928
29	VC0028	11	538	428	966	24792	26633	-1	1841	966	1184.1868641942	0.8157496331	1.3443417997
30	VC0029	10	707	56	763	26643	27602	-1	959	763	616.8577961772	1.2369139285	2.0384135391
31	VC0030	3	58	4	62	27593	27877	-1	284	62	182.6773869805	0.33939614	0.5593191821
32	VC0031	8	188	33	221	27889	29535	-1	1646	221	1058.756968204	0.208735344	0.3439923681
33	VC0032	12	1854	223	2077	30069	31592	1	1523	2077	979.6396491948	2.120167351	3.4940004587
34	VC0033	8	763	1092	1855	31682	32566	1	884	1855	568.6155284887	3.2623097806	5.3762321471
35	VC0034	6	1030	121	1151	32644	33246	-1	602	1151	387.2246019798	2.9724351039	4.898523511
36	VC0035	8	10	499	509	33277	34263	-1	986	509	634.225012545	0.8025542827	1.3225960821
37	VC0036	9	467	299	766	34311	35798	-1	1487	766	956.4833607043	0.8008503143	1.3197879705
38	VC0037	4	3	20	23	36152	36430	-1	278	23	178.8180055654	0.1286223942	0.2119675619
39	VC0038	0				36527	36694	-1					
40	VC0039	7	233	1190	1423	36928	37770	1					
41	VC0040	8	2144	6	2150	37848	38495	1					
42	VC0041	4	246	181	427	38534	39190	-1					
43	VC0042	15	1300	1527	2827	39103	40548	-1	1				
44	VC0043	0				40560	41936	-1	1				
45	VC0044	8	48	739	787	41998	43302	-1	1				
46	VC0045	0				43334	44281	-1					
47	VC0046	5	309	24	333	44299	44808	-1					
48	VC0047	13	1034	750	1784	44900	46075	1	1				
49	VC0048	7	292	271	563	46075	47190	1	1				
50	VC0049	3	154	162	316	47193	47669	1	476	316	306.1775922631	1.0320807531	1.7008518799
51	VC0050	8	337	461	798	47679	48251	1	572	798	367.9276949044	2.1689044099	3.5743183194

Sheet1

Sheet 1 / 1 Default STD Sum=0 Average= 100%

$$W_i = \frac{\ln(N_i(t_2) \times d / N_i(t_1))}{\ln((1-N_i(t_2)) \times d / (1-N_i(t_1)))}$$

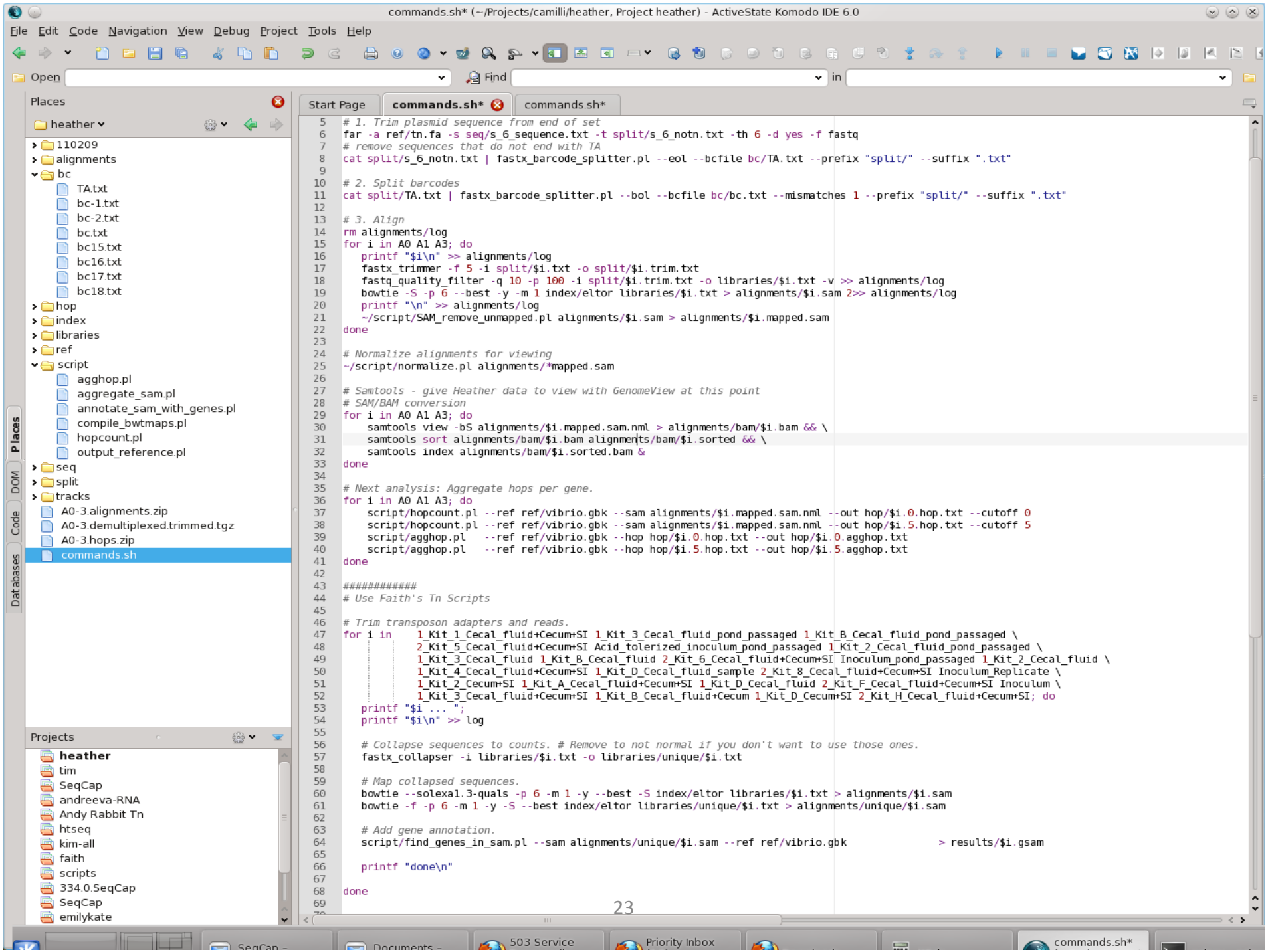


$$d_{gene} = \frac{\frac{Reads\ in\ Gene}{Gene\ Length}}{Reads\ Expected}$$

Tufts University Core Facility

Tn-Seq Bioinformatics Workflow

- **Align** short reads to microbial reference
- Get read **counts** for each insertion site (TA)
- **Compare two time points** to generate fitness score for every insertion
- Aggregate **fitness scores per gene**
- Use **query genes** to look for genetic interactions and create a gene network


commands.sh* (~/Projects/camilli/heather, Project heather) - ActiveState Komodo IDE 6.0
File Edit Code Navigation View Debug Project Tools Help
Open
Find
in
Places
heather
110209
alignments
bc
TA.txt
bc-1.txt
bc-2.txt
bc.txt
bc15.txt
bc16.txt
bc17.txt
bc18.txt
hop
index
libraries
ref
script
agghop.pl
aggregate_sam.pl
annotate_sam_with_genes.pl
compile_bwtmaps.pl
hopcount.pl
output_reference.pl
seq
split
tracks
A0-3.alignments.zip
A0-3.demultiplexed.trimmed.tgz
A0-3.hops.zip
commands.sh
Projects
heather
tim
SeqCap
andreeva-RNA
Andy Rabbit Tn
htseq
kim-all
faith
scripts
334.0.SeqCap
SeqCap
emilykate
Start Page
commands.sh*
commands.sh*
1. Trim plasmid sequence from end of set
far -a ref/tn.fa -s seq/s_6_sequence.txt -t split/s_6_notn.txt -th 6 -d yes -f fastq
remove sequences that do not end with TA
cat split/s_6_notn.txt | fastx_barcode_splitter.pl --eol --bcfile bc/TA.txt --prefix "split/" --suffix ".txt"
2. Split barcodes
cat split/TA.txt | fastx_barcode_splitter.pl --bol --bcfile bc/bc.txt --mismatches 1 --prefix "split/" --suffix ".txt"
3. Align
rm alignments/log
for i in A0 A1 A3; do
printf "$i\n" >> alignments/log
fastx_trimmer -f 5 -i split/$i.txt -o split/$i.trim.txt
fastq_quality_filter -q 10 -p 100 -i split/$i.trim.txt -o libraries/$i.txt -v >> alignments/log
bowtie -S -p 6 --best -y -m 1 index/eltor libraries/$i.txt > alignments/$i.sam 2>> alignments/log
printf "\n" >> alignments/log
~/script/SAM_remove_unmapped.pl alignments/$i.sam > alignments/$i.mapped.sam
done
Normalize alignments for viewing
~/script/normalize.pl alignments/*mapped.sam
# Samtools - give Heather data to view with GenomeView at this point
# SAM/BAM conversion
for i in A0 A1 A3; do
samtools view -bS alignments/$i.mapped.sam.nml > alignments/bam/$i.bam && \
samtools sort alignments/bam/$i.bam alignments/bam/$i.sorted && \
samtools index alignments/bam/$i.sorted.bam &
done
# Next analysis: Aggregate hops per gene.
for i in A0 A1 A3; do
script/hopcount.pl --ref ref/vibrio.gbk --sam alignments/$i.mapped.sam.nml --out hop/$i.0.hop.txt --cutoff 0
script/hopcount.pl --ref ref/vibrio.gbk --sam alignments/$i.mapped.sam.nml --out hop/$i.5.hop.txt --cutoff 5
script/agghop.pl --ref ref/vibrio.gbk --hop hop/$i.0.hop.txt --out hop/$i.0.agghop.txt
script/agghop.pl --ref ref/vibrio.gbk --hop hop/$i.5.hop.txt --out hop/$i.5.agghop.txt
done
############
Use Faith's Tn Scripts
Trim transposon adapters and reads.
for i in 1_Kit_1_Cecal_fluid+Cecum+SI 1_Kit_3_Cecal_fluid_pond_passaged 1_Kit_B_Cecal_fluid_pond_passaged \
2_Kit_5_Cecal_fluid+Cecum+SI Acid_tolerized_inoculum_pond_passaged 1_Kit_2_Cecal_fluid_pond_passaged \
1_Kit_3_Cecal_fluid 1_Kit_B_Cecal_fluid 2_Kit_6_Cecal_fluid+Cecum+SI Inoculum_pond_passaged 1_Kit_2_Cecal_fluid \
1_Kit_4_Cecal_fluid+Cecum+SI 1_Kit_D_Cecal_fluid_sample 2_Kit_8_Cecal_fluid+Cecum+SI Inoculum_Replicate \
1_Kit_2_Cecum+SI 1_Kit_A_Cecal_fluid+Cecum+SI 1_Kit_D_Cecal_fluid 2_Kit_F_Cecal_fluid+Cecum+SI Inoculum \
1_Kit_3_Cecal_fluid+Cecum+SI 1_Kit_B_Cecal_fluid+Cecum 1_Kit_D_Cecum+SI 2_Kit_H_Cecal_fluid+Cecum+SI; do
printf "$i ... ";
printf "$i\n" >> log
Collapse sequences to counts. # Remove to not normal if you don't want to use those ones.
fastx_collapser -i libraries/$i.txt -o libraries/unique/$i.txt
Map collapsed sequences.
bowtie --solexa1.3-quals -p 6 -m 1 -y --best -S index/eltor libraries/$i.txt > alignments/$i.sam
bowtie -f -p 6 -m 1 -y -S --best index/eltor libraries/unique/$i.txt > alignments/unique/$i.sam
Add gene annotation.
script/find_genes_in_sam.pl --sam alignments/unique/$i.sam --ref ref/vibrio.gbk > results/$i.gsam
printf "done\n"
done
Places
DOM
Code
Databases
503 Service
Priority Inbox
commands.sh*

Galaxy http://usegalaxy.org

Making Computational Analyses Accessible

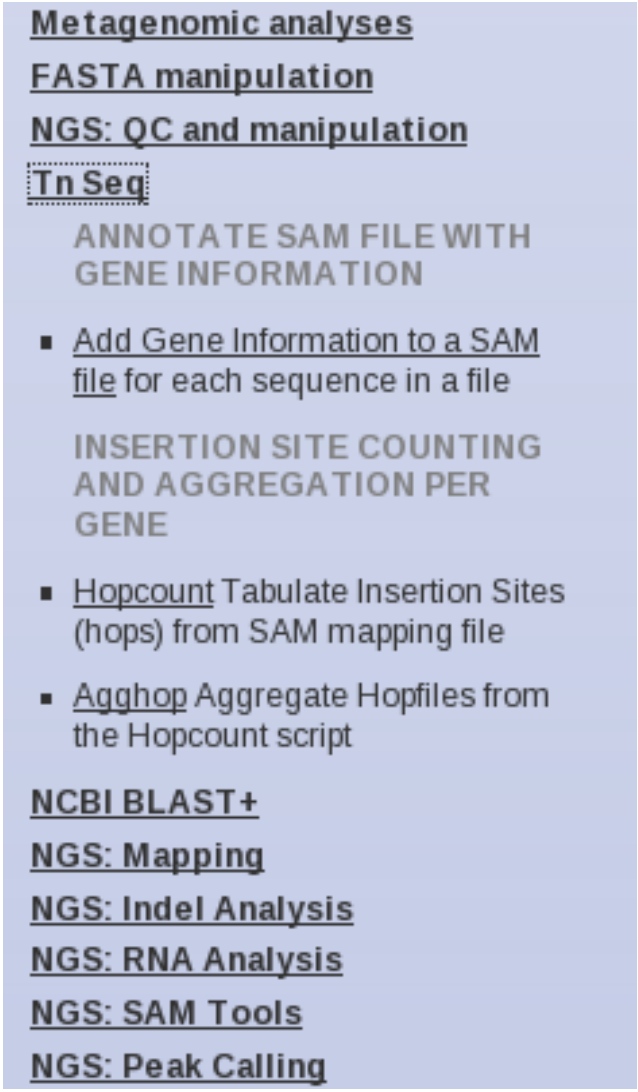
- Web-based, Visual Platform Analysis tool
  - No need to download or install software
  - Workflows for Genomics Research
- NGS Toolset
  - QC, manipulation, mapping, peak calling, RNA-Seq
- Can be installed on your own computer (or on a local server)
  - Integrate Perl, Python, Shell scripts

Tn-Seq Analyses

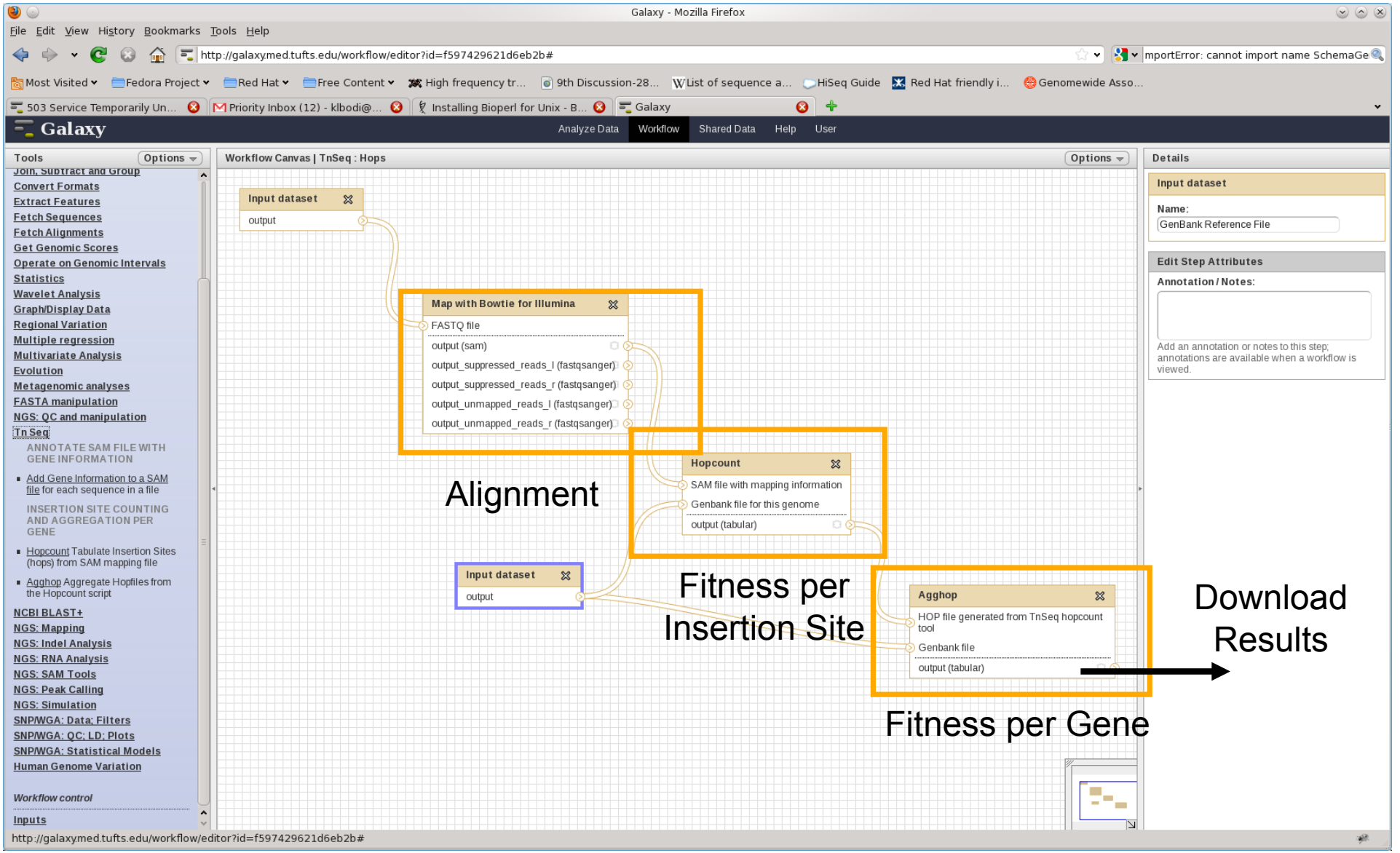
- Issue
  - Every time researcher needs to change one parameter in his workflow, it requires our bioinformatics team to re-run the entire analysis
- Solution
  - Integrate workflow into Galaxy
  - Now the researcher can run it himself!
  - Also, others can start using Tn-Seq by using a shared Tn-Seq workflow

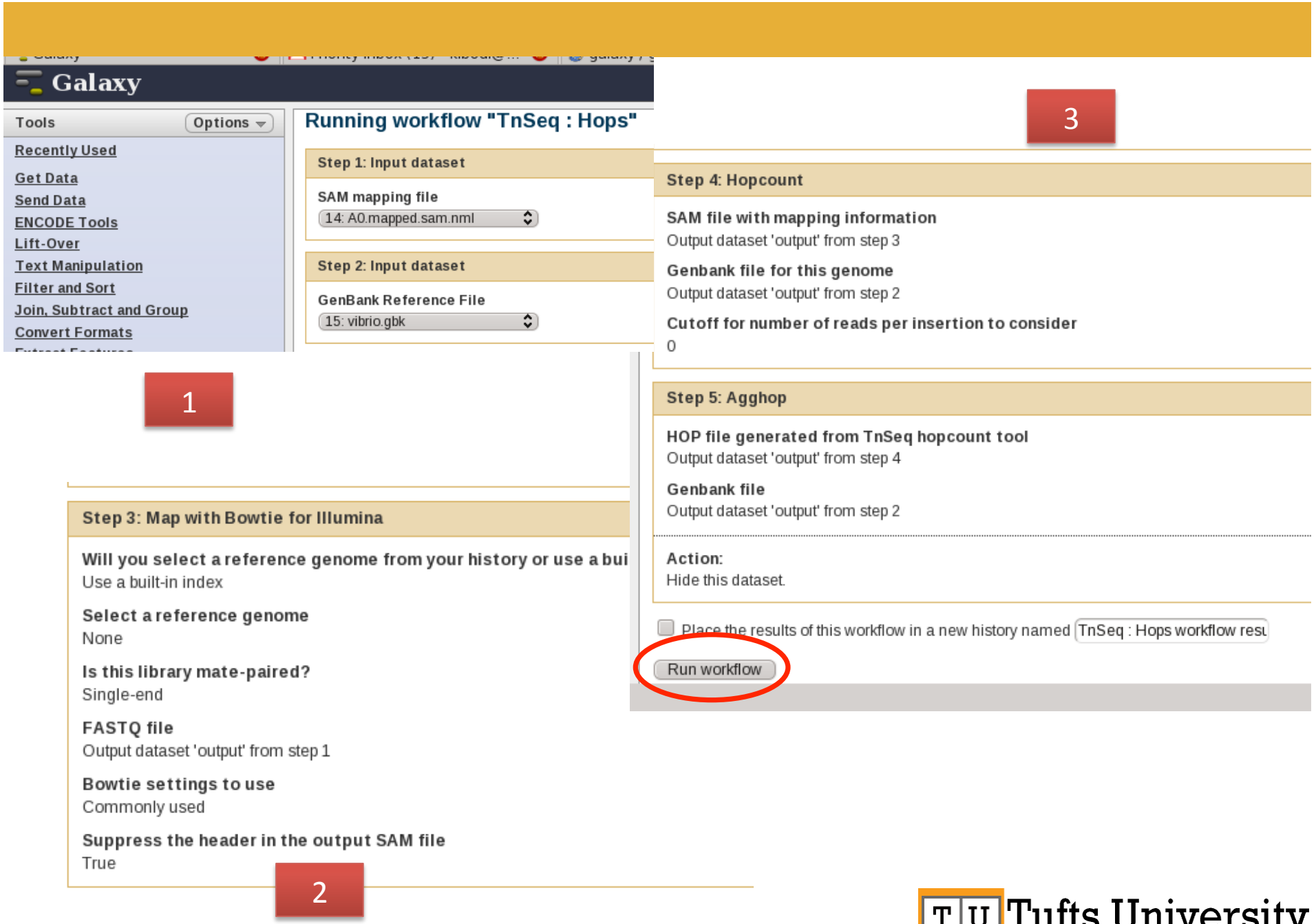
Tool Integration

- Add TnSeq scripts as custom tools in Galaxy
- Sharing
  - Can publish tools, histories, workflows to all or selected users

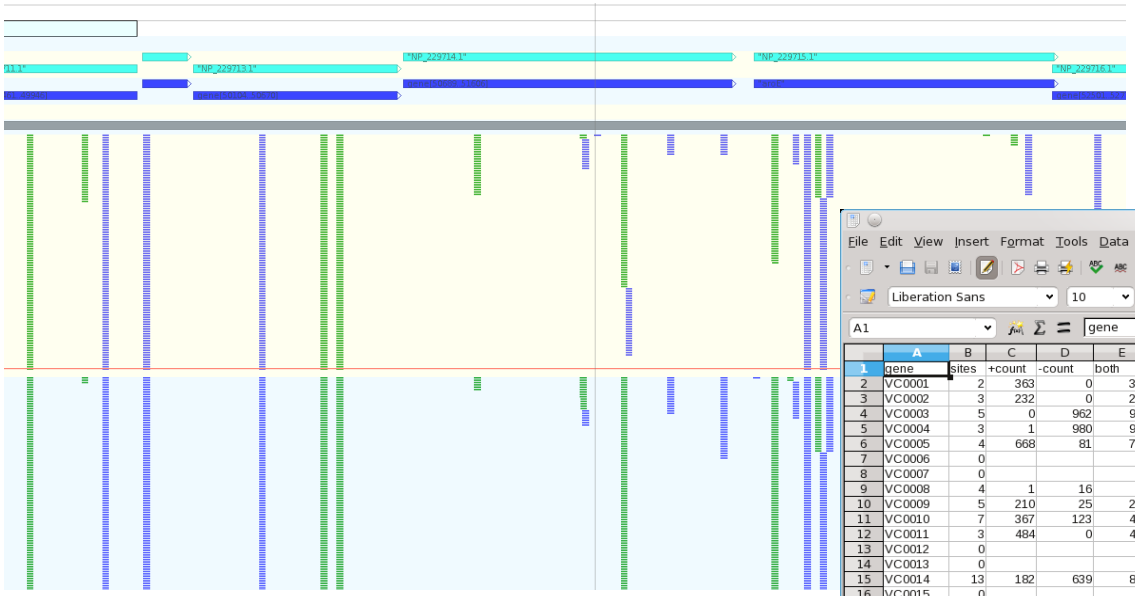


Tufts University Core Facility


Galaxy - Mozilla Firefox
Galaxy
Analyze Data
Workflow
Shared Data
Help
User
Tools
Options
Workflow Canvas | TnSeq : Hops
Details
Input dataset
output
Map with Bowtie for Illumina
FASTQ file
output (sam)
output_suppressed_reads_l (fastqsanger)
output_suppressed_reads_r (fastqsanger)
output_unmapped_reads_l (fastqsanger)
output_unmapped_reads_r (fastqsanger)
Alignment
Hopcount
SAM file with mapping information
Genbank file for this genome
output (tabular)
Input dataset
output
Fitness per Insertion Site
Agghop
HOP file generated from TnSeq hopcount tool
Genbank file
output (tabular)
Download Results
Fitness per Gene
Name:
GenBank Reference File
Edit Step Attributes
Annotation / Notes:
Add an annotation or notes to this step; annotations are available when a workflow is viewed.
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
Tn Seq
ANNOTATE SAM FILE WITH GENE INFORMATION
Add Gene Information to a SAM file for each sequence in a file
INSERTION SITE COUNTING AND AGGREGATION PER GENE
Hopcount Tabulate Insertion Sites (hops) from SAM mapping file
Agghop Aggregate Hopfiles from the Hopcount script
NCBI BLAST+
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: Peak Calling
NGS: Simulation
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Human Genome Variation
Workflow control
Inputs
http://galaxymed.tufts.edu/workflow/editor?id=f597429621d6eb2b#



Tufts University Core Facility



A0.0.agghop.txt - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

Liberation Sans 10

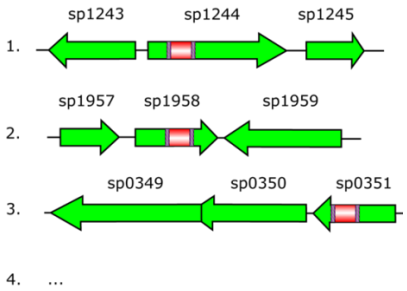
A1 gene

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	gene	sites	+count	-count	both	start	end	strand	length	read_count	expected	dval_genome	dval_genic
2	VC0001	2	363	0	363	235	402	-1	167	363	107.4194493864	3.3792763049	5.5689910298
3	VC0002	3	232	0	232	372	806	-1	434	232	279.1619223576	0.8310588996	1.3695712157
4	VC0003	5	0	962	962	816	2210	-1	1394	962	896.6629487706	1.072866902	1.7680667736
5	VC0004	3	1	980	981	2271	3896	-1	1625	981	1045.2491332512	0.9385322301	1.5466854731
6	VC0005	4	668	81	749	3899	4156	-1	257	749	165.3101706127	4.5308766982	7.4668092847
7	VC0006	0				4123	4479	-1	356				
8	VC0007	0				4492	4629	-1	137				
9	VC0008	4	1	16	17	4871	5608	-1	737	17	474.0606838192	0.0358603879	0.0590973216
10	VC0009	5	210	25	235	5605	6276	-1	671	235	431.6074882533	0.5444761882	0.8972876836
11	VC0010	7	367	123	490	6354	7100	-1	746	490	479.8497559418	1.0211529628	1.6828430637
12	VC0011	3	484	0	484	7224	7352	1	128	484	82.3334701884	5.8785327388	9.6877239789
13	VC0012	0				7397	8815	1	1418				
14	VC0013	0				8873	9973	1	1100				
15	VC0014	13	182	639	821	9991	11082	1	1091	821	701.764187309	1.16990866	1.9279899734
16	VC0015	0				11103	13520	1	2417				
17	VC0016	3	13	0	13	13876	14367	1	491	13	315.8260458008	0.0411618996	0.0678341246
18	VC0017	0				14377	14493	1	116				
19	VC0018	2	94	0	94	14769	15206	1	437	94	281.0916130651	0.3344105467	0.5511030075
20	VC0019	6	238	9	247	15294	16550	-1	1256	247	807.8971762237	0.3057319759	0.5038412008
21	VC0020	0				16869	18935	-1	2066				
22	VC0021	1	116	0	116	18950	19942	-1	992	116	638.0843939601	0.1817941343	0.2995937034
23	VC0022	3	425	76	501	20235	20837	1	602	501	387.2246019798	1.2938227515	2.132198331
24	VC0023	4	446	0	446	20943	21200	1	257	446	165.3101706127	2.6979586213	4.4461908424
25	VC0024	0				21287	21535	-1	248				
26	VC0025	5	141	411	552	21656	21856	1	200	552	128.6460471694	4.2908430702	7.0712378662
27	VC0026	5	545	497	1042	22028	23008	-1	980	1042	630.36563113	1.6530089024	2.7241311212
28	VC0027	6	450	1105	1555	23256	24788	-1	1532	1555	985.4287213175	1.5779933813	2.6005067928
29	VC0028	11	538	428	966	24792	26633	-1	1841	966	1184.1868641942	0.8157496331	1.3443417997
30	VC0029	10	707	56	763	26643	27602	-1	959	763	616.8577961772	1.2369139285	2.0384135391
31	VC0030	3	58	4	62	27593	27877	-1	284	62	182.6773869805	0.33939614	0.5593191821
32	VC0031	8	188	33	221	27889	29535	-1	1646	221	1058.756968204	0.208735344	0.3439923681
33	VC0032	12	1854	223	2077	30069	31592	1	1523	2077	979.6396491948	2.120167351	3.4940004587
34	VC0033	8	763	1092	1855	31682	32566	1	884	1855	568.6155284887	3.2623097806	5.3762321471
35	VC0034	6	1030	121	1151	32644	33246	-1	602	1151	387.2246019798	2.9724351039	4.898523511
36	VC0035	8	10	499	509	33277	34263	-1	986	509	634.225012545	0.8025542827	1.3225960821
37	VC0036	9	467	299	766	34311	35798	-1	1487	766	956.4833607043	0.8008503143	1.3197879705
38	VC0037	4	3	20	23	36152	36430	-1	278	23	178.8180055654	0.1286223942	0.2119675619
39	VC0038	0				36527	36694	-1					
40	VC0039	7	233	1190	1423	36928	37770	1					
41	VC0040	8	2144	6	2150	37848	38495	1					
42	VC0041	4	246	181	427	38534	39190	-1					
43	VC0042	15	1300	1527	2827	39103	40548	-1	1				
44	VC0043	0				40560	41936	-1	1				
45	VC0044	8	48	739	787	41998	43302	-1	1				
46	VC0045	0				43334	44281	-1					
47	VC0046	5	309	24	333	44299	44808	-1					
48	VC0047	13	1034	750	1784	44900	46075	1	1				
49	VC0048	7	292	271	563	46075	47190	1	1				
50	VC0049	3	154	162	316	47193	47669	1	476	316	306.1775922631	1.0320807531	1.7008518799
51	VC0050	8	337	461	798	47679	48251	1	572	798	367.9276949044	2.1689044099	3.5743183194

Sheet1

Sheet 1 / 1 Default STD Sum=0 Average= 100%

$$W_i = \frac{\ln(N_i(t_2) \times d / N_i(t_1))}{\ln((1-N_i(t_2)) \times d / (1-N_i(t_1)))}$$



$$d_{gene} = \frac{\frac{Reads\ in\ Gene}{Gene\ Length}}{Reads\ Expected}$$

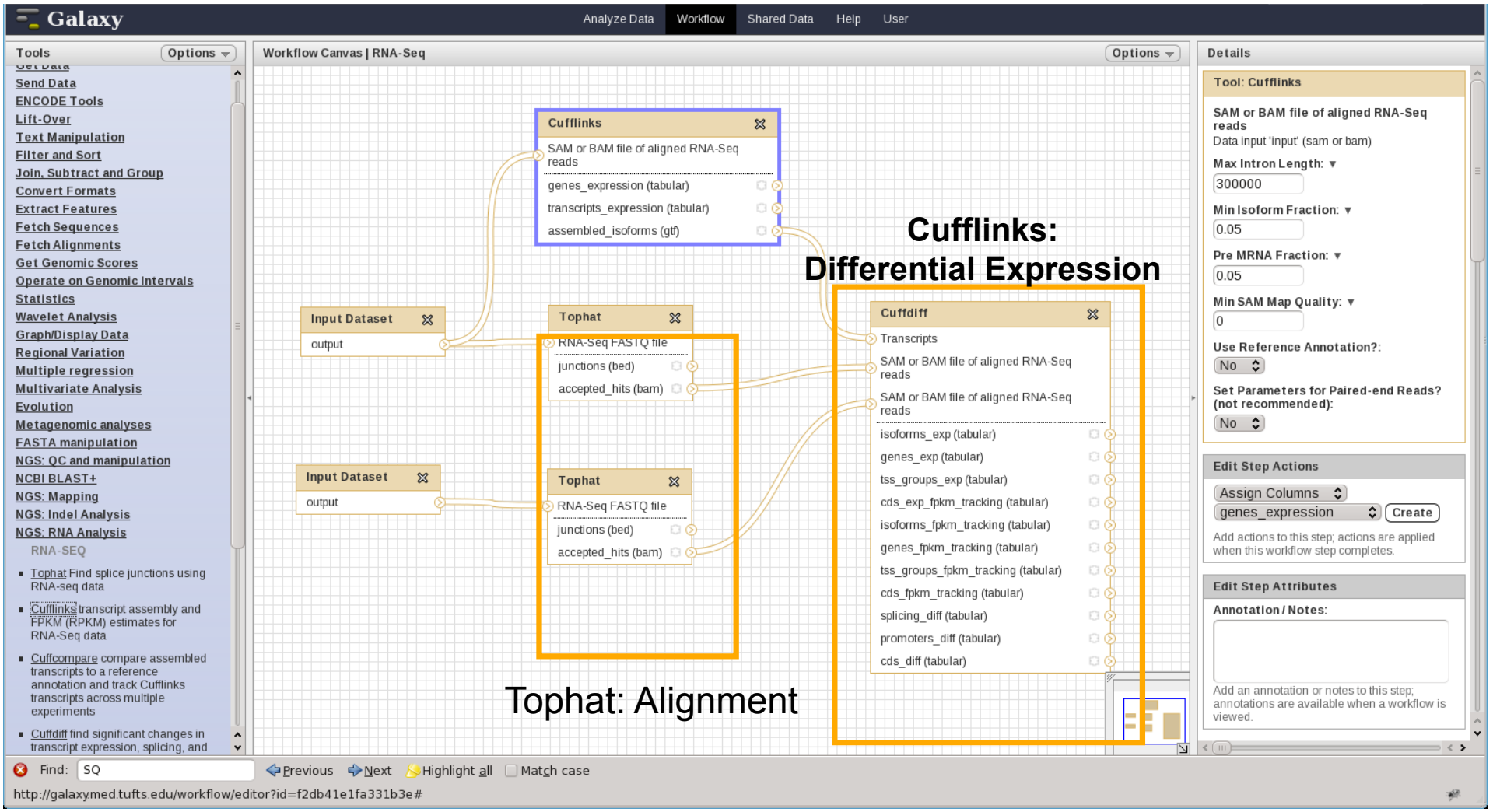
Tufts University Core Facility

Basic Workflow for NGS Analyses

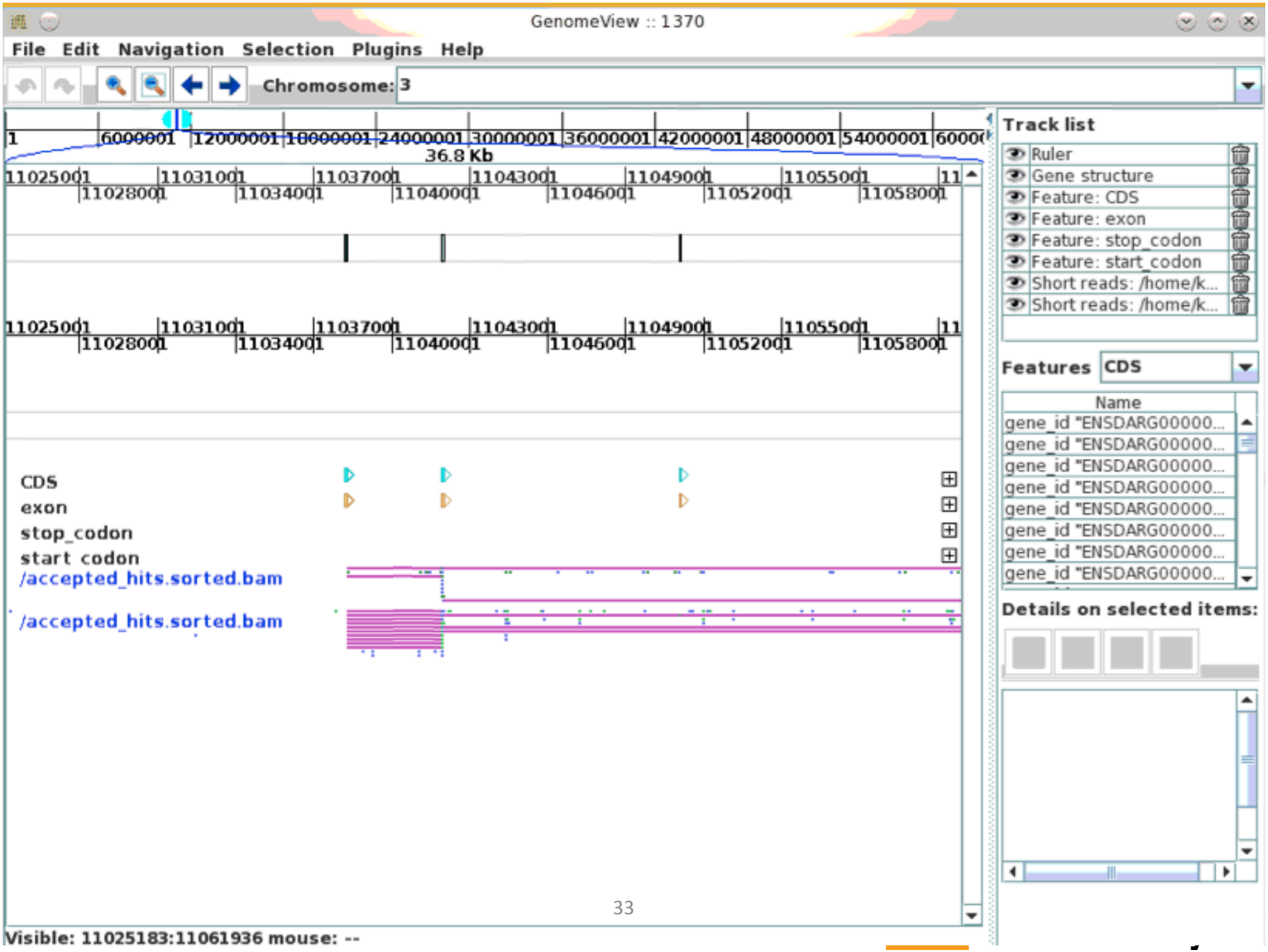
- Sequence Data (FASTQ format)
- Read mapping (Bowtie, MAQ, etc.)
- Visualization
  - GenomeView, IGV, CLC Genomics Workbench, UCSC Genome Browser
- Analysis
  - SNP / Indel Detection
  - Peak calling
  - **Differential Expression**

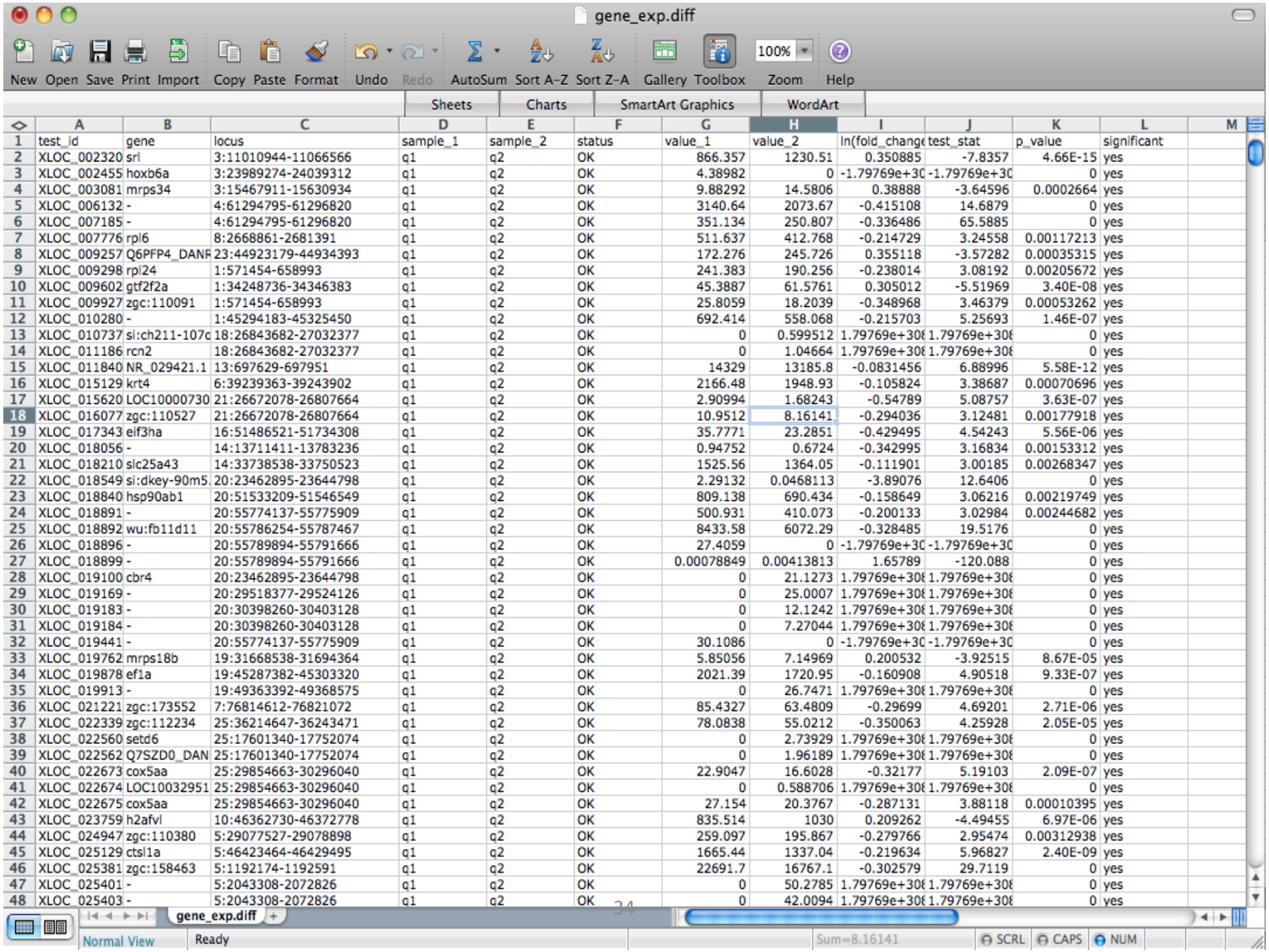
RNA-Seq with Tophat & Cufflinks

- Tophat: spliced read mapper for RNA-Seq
  - Take unmapped reads, look for splice junctions
  - Output in SAM format
    - SAMtools
      - Call variations and filter by depth and quality
- Cufflinks
  - Differential Expression Analysis Tool
  - Output per Gene, Transcript FPKM
  - Report significant changes in expression



Tufts University Core Facility


GenomeView :: 1370
File Edit Navigation Selection Plugins Help
Chromosome: 3
1
6000001
12000001
18000001
24000001
30000001
36000001
42000001
48000001
54000001
36.8 Kb
11025001
11028001
11031001
11034001
11037001
11040001
11043001
11046001
11049001
11052001
11055001
11058001
CDS
exon
stop_codon
start codon
/accepted_hits.sorted.bam
/accepted_hits.sorted.bam
Track list
Ruler
Gene structure
Feature: CDS
Feature: exon
Feature: stop_codon
Feature: start_codon
Short reads: /home/k...
Short reads: /home/k...
Features
CDS
Name
gene_id "ENSDARG00000...
Details on selected items:
Visible: 11025183:11061936 mouse: --



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	test_id	gene	locus	sample_1	sample_2	status	value_1	value_2	ln(fold_change	test_stat	p_value	significant	
2	XLOC_002320	srl	3:11010944-11066566	q1	q2	OK	866.357	1230.51	0.350885	-7.8357	4.66E-15	yes	
3	XLOC_002455	hoxb6a	3:23989274-24039312	q1	q2	OK	4.38982	0	-1.79769e+30	-1.79769e+30	0	yes	
4	XLOC_003081	mrps34	3:15467911-15630934	q1	q2	OK	9.88292	14.5806	0.38888	-3.64596	0.0002664	yes	
5	XLOC_006132	-	4:61294795-61296820	q1	q2	OK	3140.64	2073.67	-0.415108	14.6879	0	yes	
6	XLOC_007185	-	4:61294795-61296820	q1	q2	OK	351.134	250.807	-0.336486	65.5885	0	yes	
7	XLOC_007776	rpl6	8:2668861-2681391	q1	q2	OK	511.637	412.768	-0.214729	3.24558	0.00117213	yes	
8	XLOC_009257	Q6PFP4_DANR	23:44923179-44934393	q1	q2	OK	172.276	245.726	0.355118	-3.57282	0.00035315	yes	
9	XLOC_009298	rpl24	1:571454-658993	q1	q2	OK	241.383	190.256	-0.238014	3.08192	0.00205672	yes	
10	XLOC_009602	gtf2f2a	1:34248736-34346383	q1	q2	OK	45.3887	61.5761	0.305012	-5.51969	3.40E-08	yes	
11	XLOC_009927	zgc:110091	1:571454-658993	q1	q2	OK	25.8059	18.2039	-0.348968	3.46379	0.00053262	yes	
12	XLOC_010280	-	1:45294183-45325450	q1	q2	OK	692.414	558.068	-0.215703	5.25693	1.46E-07	yes	
13	XLOC_010737	si:ch211-107c	18:26843682-27032377	q1	q2	OK	0	0.599512	1.79769e+308	1.79769e+308	0	yes	
14	XLOC_011186	rcn2	18:26843682-27032377	q1	q2	OK	0	1.04664	1.79769e+308	1.79769e+308	0	yes	
15	XLOC_011840	NR_029421.1	13:697629-697951	q1	q2	OK	14329	13185.8	-0.0831456	6.88996	5.58E-12	yes	
16	XLOC_015129	krt4	6:39239363-39243902	q1	q2	OK	2166.48	1948.93	-0.105824	3.38687	0.00070696	yes	
17	XLOC_015620	LOC10000730	21:26672078-26807664	q1	q2	OK	2.90994	1.68243	-0.54789	5.08757	3.63E-07	yes	
18	XLOC_016077	zgc:110527	21:26672078-26807664	q1	q2	OK	10.9512	8.16141	-0.294036	3.12481	0.00177918	yes	
19	XLOC_017343	eif3ha	16:51486521-51734308	q1	q2	OK	35.7771	23.2851	-0.429495	4.54243	5.56E-06	yes	
20	XLOC_018056	-	14:13711411-13783236	q1	q2	OK	0.94752	0.6724	-0.342995	3.16834	0.00153312	yes	
21	XLOC_018210	slc25a43	14:33738538-33750523	q1	q2	OK	1525.56	1364.05	-0.111901	3.00185	0.00268347	yes	
22	XLOC_018549	si:dkey-90m5.	20:23462895-23644798	q1	q2	OK	2.29132	0.0468113	-3.89076	12.6406	0	yes	
23	XLOC_018840	hsp90ab1	20:51533209-51546549	q1	q2	OK	809.138	690.434	-0.158649	3.06216	0.00219749	yes	
24	XLOC_018891	-	20:55774137-55775909	q1	q2	OK	500.931	410.073	-0.200133	3.02984	0.00244682	yes	
25	XLOC_018892	wu:fb11d11	20:55786254-55787467	q1	q2	OK	8433.58	6072.29	-0.328485	19.5176	0	yes	
26	XLOC_018896	-	20:55789894-55791666	q1	q2	OK	27.4059	0	-1.79769e+30	-1.79769e+30	0	yes	
27	XLOC_018899	-	20:55789894-55791666	q1	q2	OK	0.00078849	0.00413813	1.65789	-120.088	0	yes	
28	XLOC_019100	cbr4	20:23462895-23644798	q1	q2	OK	0	21.1273	1.79769e+308	1.79769e+308	0	yes	
29	XLOC_019169	-	20:29518377-29524126	q1	q2	OK	0	25.0007	1.79769e+308	1.79769e+308	0	yes	
30	XLOC_019183	-	20:30398260-30403128	q1	q2	OK	0	12.1242	1.79769e+308	1.79769e+308	0	yes	
31	XLOC_019184	-	20:30398260-30403128	q1	q2	OK	0	7.27044	1.79769e+308	1.79769e+308	0	yes	
32	XLOC_019441	-	20:55774137-55775909	q1	q2	OK	30.1086	0	-1.79769e+30	-1.79769e+30	0	yes	
33	XLOC_019762	mrps18b	19:31668538-31694364	q1	q2	OK	5.85056	7.14969	0.200532	-3.92515	8.67E-05	yes	
34	XLOC_019878	ef1a	19:45287382-45303320	q1	q2	OK	2021.39	1720.95	-0.160908	4.90518	9.33E-07	yes	
35	XLOC_019913	-	19:49363392-49368575	q1	q2	OK	0	26.7471	1.79769e+308	1.79769e+308	0	yes	
36	XLOC_021221	zgc:173552	7:76814612-76821072	q1	q2	OK	85.4327	63.4809	-0.29699	4.69201	2.71E-06	yes	
37	XLOC_022339	zgc:112234	25:36214647-36243471	q1	q2	OK	78.0838	55.0212	-0.350063	4.25928	2.05E-05	yes	
38	XLOC_022560	setd6	25:17601340-17752074	q1	q2	OK	0	2.73929	1.79769e+308	1.79769e+308	0	yes	
39	XLOC_022562	Q7SZD0_DAN	25:17601340-17752074	q1	q2	OK	0	1.96189	1.79769e+308	1.79769e+308	0	yes	
40	XLOC_022673	cox5aa	25:29854663-30296040	q1	q2	OK	22.9047	16.6028	-0.32177	5.19103	2.09E-07	yes	
41	XLOC_022674	LOC10032951	25:29854663-30296040	q1	q2	OK	0	0.588706	1.79769e+308	1.79769e+308	0	yes	
42	XLOC_022675	cox5aa	25:29854663-30296040	q1	q2	OK	27.154	20.3767	-0.287131	3.88118	0.00010395	yes	
43	XLOC_023759	h2afvl	10:46362730-46372778	q1	q2	OK	835.514	1030	0.209262	-4.49455	6.97E-06	yes	
44	XLOC_024947	zgc:110380	5:29077527-29078898	q1	q2	OK	259.097	195.867	-0.279766	2.95474	0.00312938	yes	
45	XLOC_025129	ctsl1a	5:46423464-46429495	q1	q2	OK	1665.44	1337.04	-0.219634	5.96827	2.40E-09	yes	
46	XLOC_025381	zgc:158463	5:1192174-1192591	q1	q2	OK	22691.7	16767.1	-0.302579	29.7119	0	yes	
47	XLOC_025401	-	5:2043308-2072826	q1	q2	OK	0	50.2785	1.79769e+308	1.79769e+308	0	yes	
48	XLOC_025403	-	5:2043308-2072826	q1	q2	OK	0	42.0094	1.79769e+308	1.79769e+308	0	yes	

Summary

- A small NGS core can still assist end users with bioinformatics analysis
- Tools that make analyses accessible to the end user help make this possible
  - Galaxy
  - GenomeQuest
  - CLC Genomics Workbench
  - Genomatix
  - Etc.

Acknowledgments

- Tufts Core Facility
  - Michael Berne
  - Tom Phimmasen
  - James Schiemer
  - Anna Wong
  - Cho Low
  - Paula Nguyen
  - Christy Le
- Galaxy Team
  - Anton Nekrutenko
  - James Taylor
  - Greg von Kuster
- Broad Institute
  - Thomas Abeel
- Camilli Laboratory
  - Tim van Opijnen
  - David Lazinski
  - Heather Kamp
- Yelick Laboratory
  - Viktoria Andreeva
- CTRC Core Facility
  - Albert Tai
  - Jennifer Curcuru
  - Fang Liu

Tufts University Core Facility