Intermediate RNA-Seq

Tips, Tricks and Non-Human Organisms

Kevin Silverstein PhD, John Garbe PhD and
Ying Zhang PhD,
Research Informatics Support System (RISS)
MSI

September 25, 2014

Slides available at www.msi.umn.edu/tutorial-materials

RNA-Seq Tutorials

- Tutorial 1
  - RNA-Seq experiment design and analysis
  - Instruction on individual software will be provided in other tutorials
- Tutorial 2 – Thursday Sept. 25
  - Advanced RNA-Seq Analysis topics
- Hands-on tutorials -
  - Analyzing human and potato RNA-Seq data using Tophat and Cufflinks in Galaxy
  - Human: Thursday Oct. 2
  - Potato: Tuesday Oct. 14

RNA-seq Tutorial 2

Tips, Tricks and Non-Human Organisms

Part I: Review and Considerations for Different Goals and Biological Systems (Kevin Silverstein)

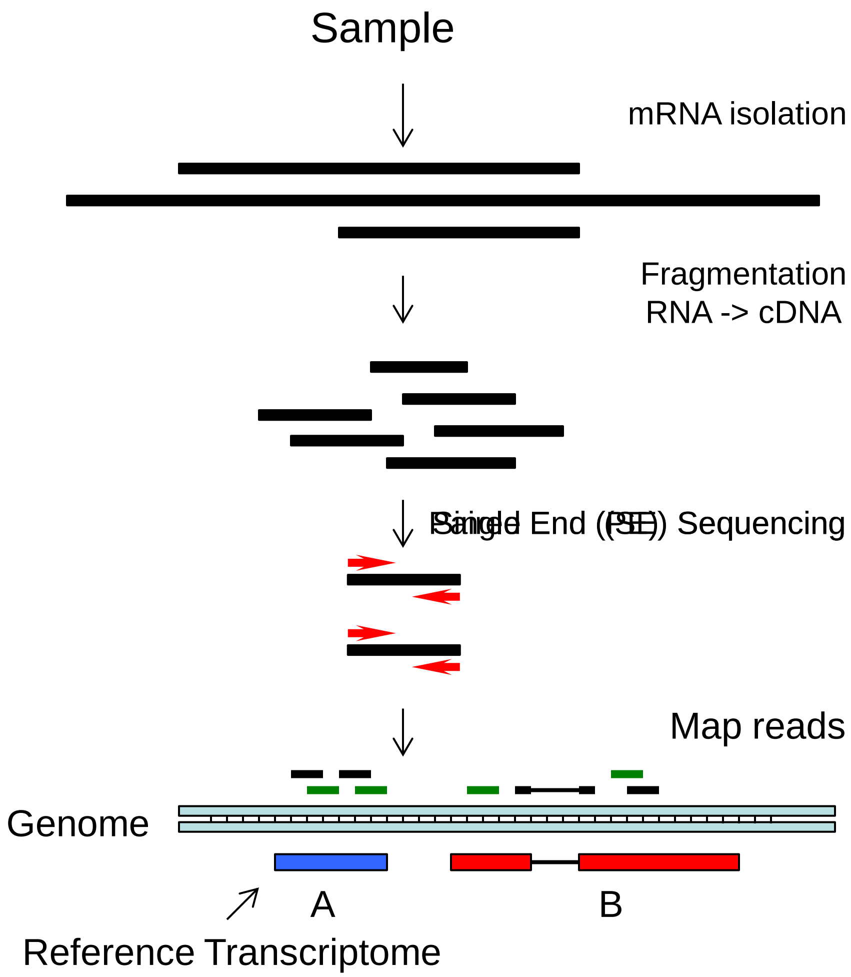
Part II: Read Mapping Statistics and Visualization (John Garbe)

Part III: Post-Analysis Processing – Exploring the Data and Results (Ying Zhang)

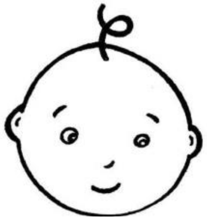
Part I

Review and Considerations for Different Goals and Biological Systems

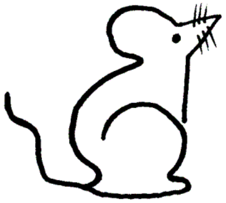
Typical RNA-seq experimental protocol and analysis



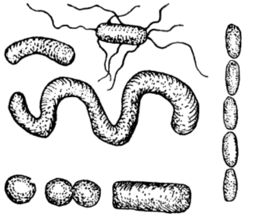
Steps in RNA-Seq data analysis depend on your goals and biological system

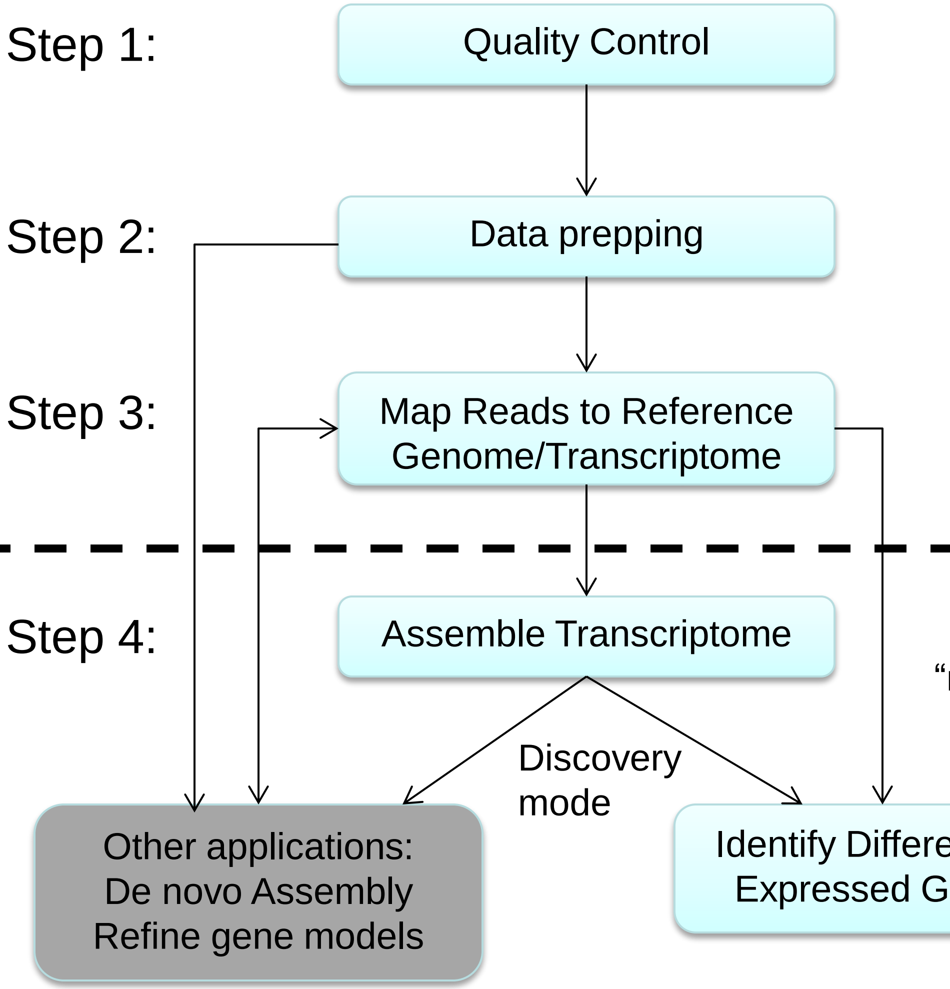


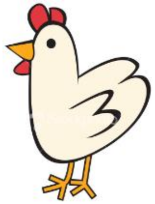
KIR
HLA

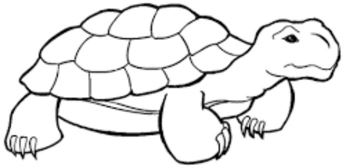


PACIFIC BIOSCIENCES™



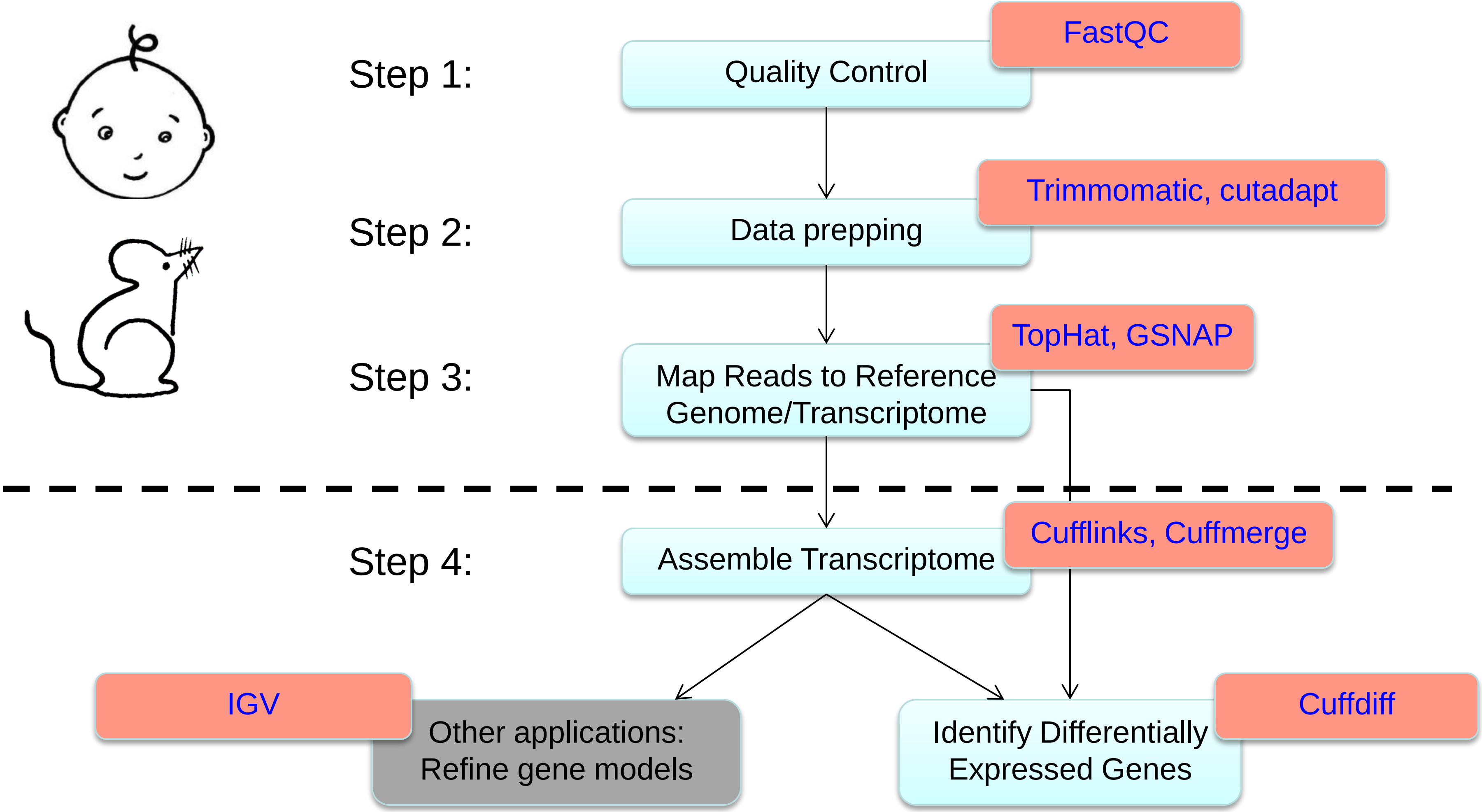






“microarray simulation”

Programs used in RNA-Seq data analysis depend on your goals and biological system



Step 1: Quality Control — FastQC

Step 2: Data prepping — Trimmomatic, cutadapt

Step 3: Map Reads to Reference Genome/Transcriptome — TopHat, GSNAP

Step 4: Assemble Transcriptome — Cufflinks, Cuffmerge

Other applications: Refine gene models — IGV

Identify Differentially Expressed Genes — Cuffdiff

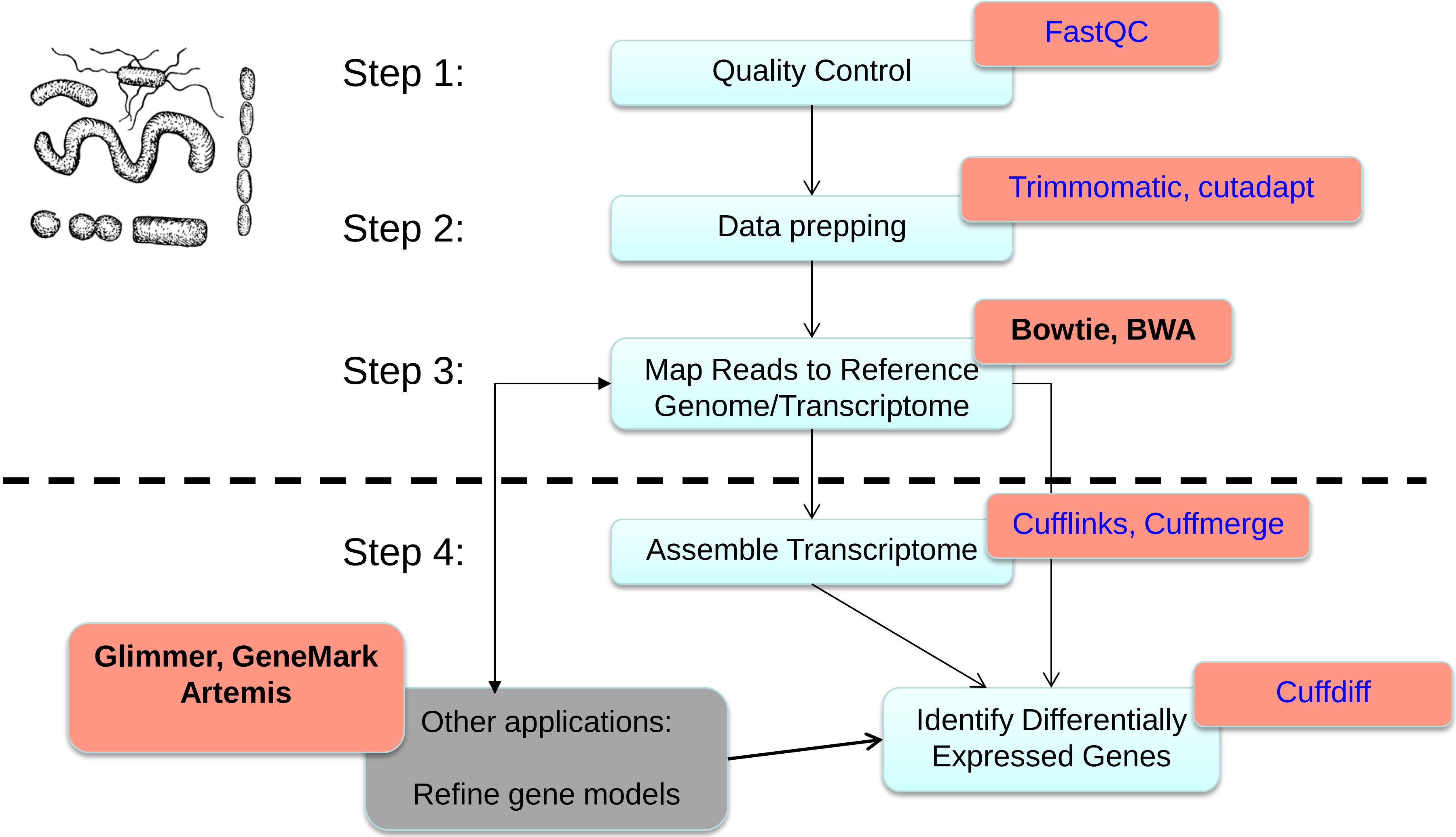
Specific Note for Prokaryotes

- Cufflinks developer:

"We don't recommend assembling bacteria transcripts using Cufflinks at first. If you are working on a new bacteria genome, consider a computational gene finding application such as Glimmer."

- For bacteria transcriptomes:
  - Genome available: do genome annotation first then reconstruct the transcriptome.
  - No genome: try *de novo* assembly of the transcriptome, followed by gene annotation.

Programs used in RNA-Seq data analysis depend on your goals and biological system



Step 1: Quality Control — FastQC

Step 2: Data prepping — Trimmomatic, cutadapt

Step 3: Map Reads to Reference Genome/Transcriptome — **Bowtie, BWA**

Step 4: Assemble Transcriptome — Cufflinks, Cuffmerge

Identify Differentially Expressed Genes — Cuffdiff

Other applications:

Refine gene models

Glimmer, GeneMark Artemis

Visualizing microbial data in Artemis

All mapped reads

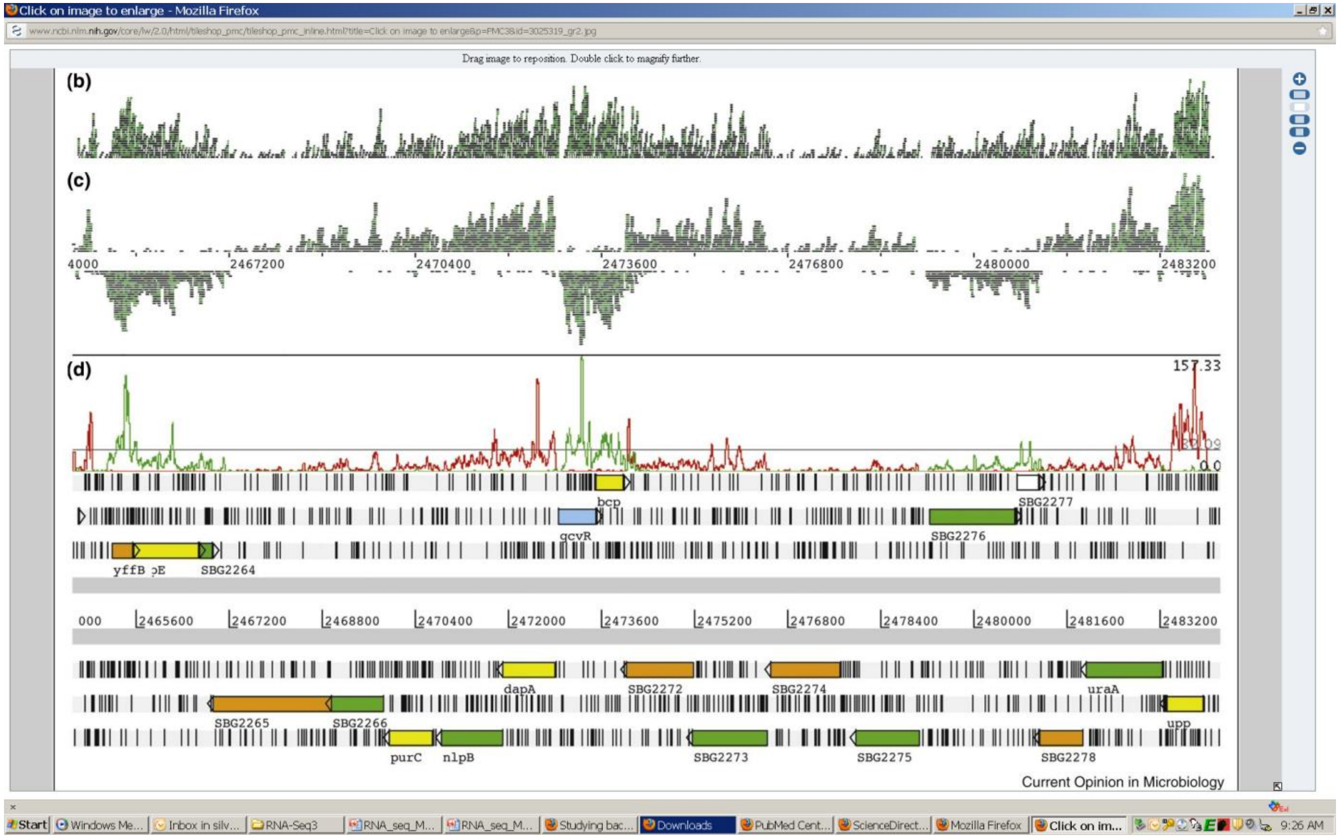
Reverse reads

Forward reads

Strand-specific coverage

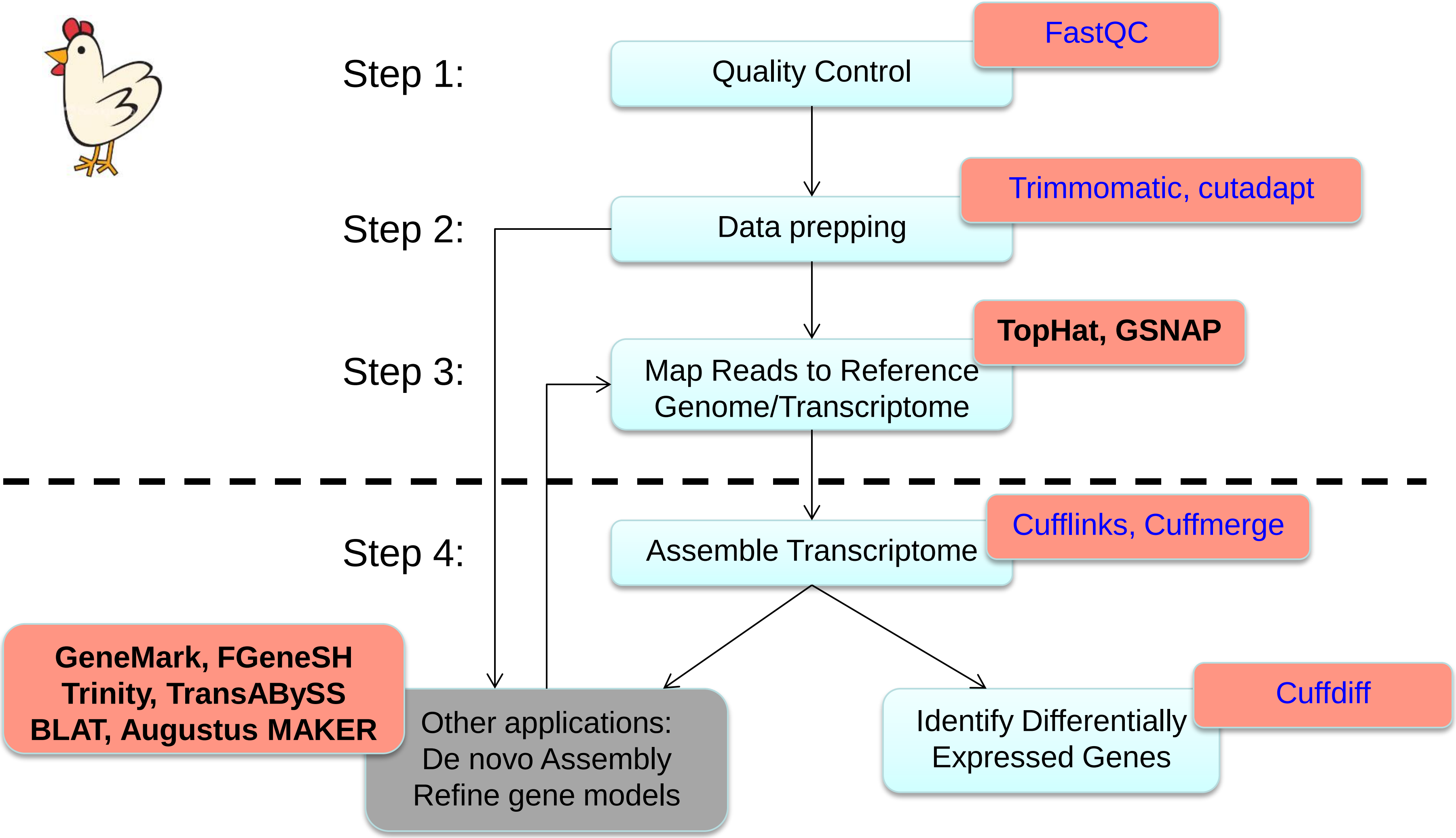
Forward genes

Reverse genes



Croucher NJ and Thomson NR. Curr Opin Microbiol. (2010) 13:619–624.

Programs used in RNA-Seq data analysis depend on your goals and biological system



Step 1:

Quality Control

FastQC

Step 2:

Data prepping

Trimmomatic, cutadapt

Step 3:

Map Reads to Reference Genome/Transcriptome

TopHat, GSNAP

Step 4:

Assemble Transcriptome

Cufflinks, Cuffmerge

GeneMark, FGeneSH
Trinity, TransABySS
BLAT, Augustus MAKER

Other applications:
De novo Assembly
Refine gene models

Identify Differentially Expressed Genes

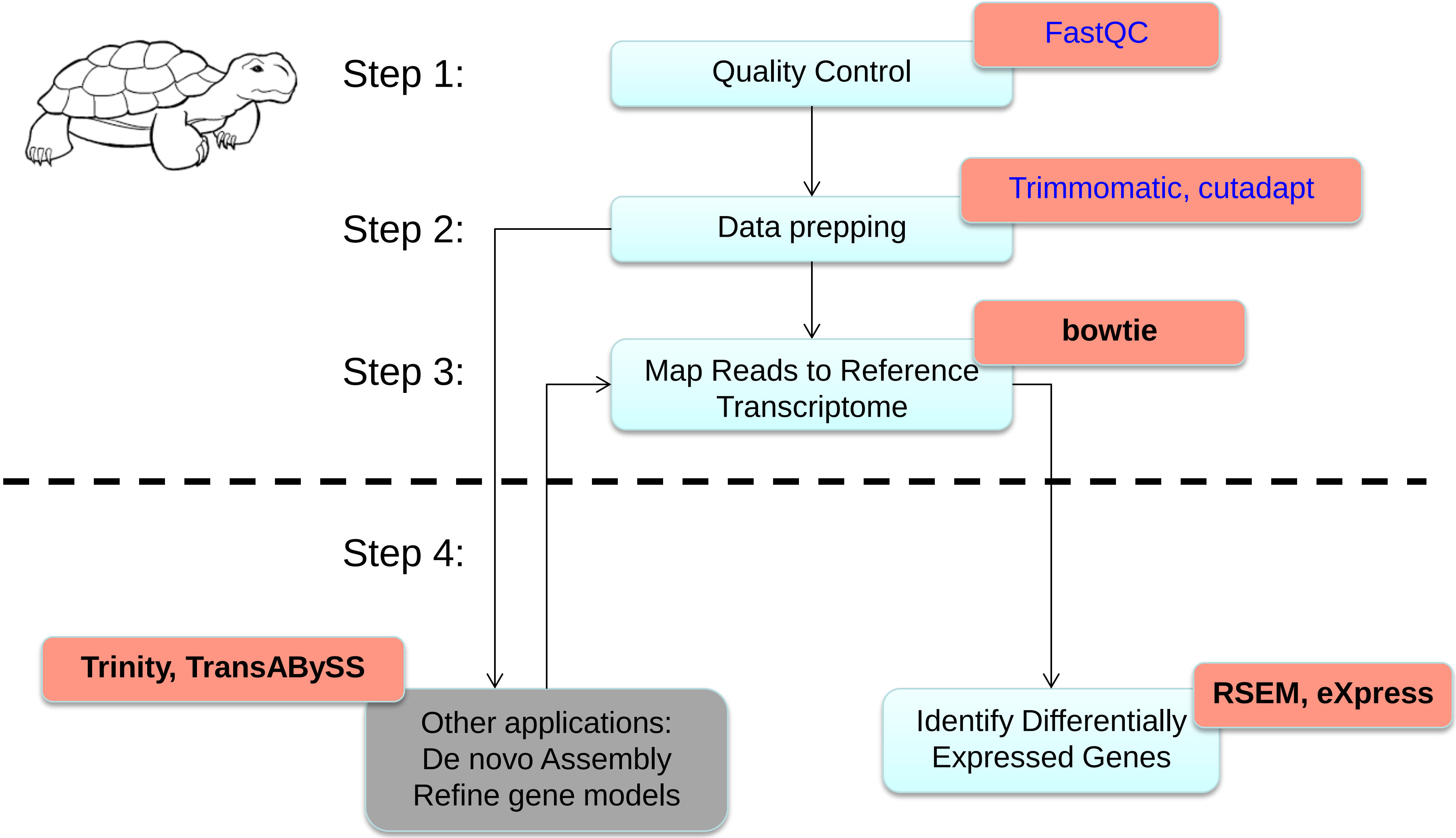
Cuffdiff

Augustus creates superior gene models using RNA-seq data

http://augustus.gobics.de/binaries/readme.rnaseq.html

- Ideal for organisms with draft genome sequence and poor (or no) gene models
- Utilizes intron/exon boundaries to provide “hints” to the *de novo* gene prediction
  - Bonus for predictions that match boundaries
  - Penalties for predictions that conflict

Programs used in RNA-Seq data analysis depend on your goals and biological system



Step 1:

Quality Control

FastQC

Step 2:

Data prepping

Trimmomatic, cutadapt

Step 3:

Map Reads to Reference Transcriptome

bowtie

Step 4:

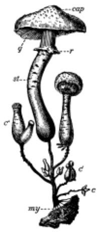
Trinity, TransABySS

Other applications:
De novo Assembly
Refine gene models

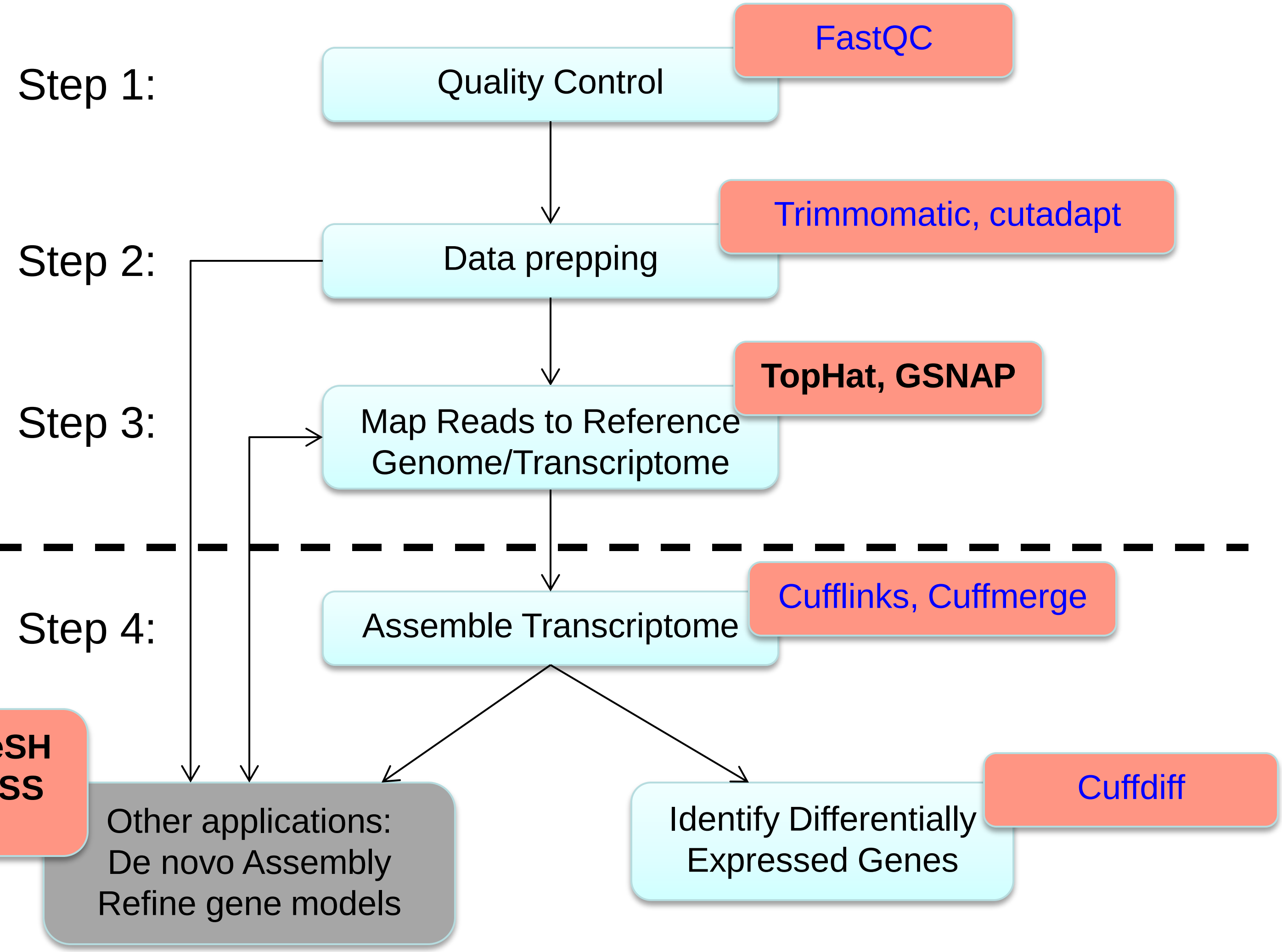
Identify Differentially Expressed Genes

RSEM, eXpress

Programs used in RNA-Seq data analysis depend on your goals and biological system



Consider sequencing and assembling the genome first via **PacBio** if < 50 MB…



Step 1: Quality Control — FastQC

Step 2: Data prepping — Trimmomatic, cutadapt

Step 3: Map Reads to Reference Genome/Transcriptome — **TopHat, GSNAP**

Step 4: Assemble Transcriptome — Cufflinks, Cuffmerge

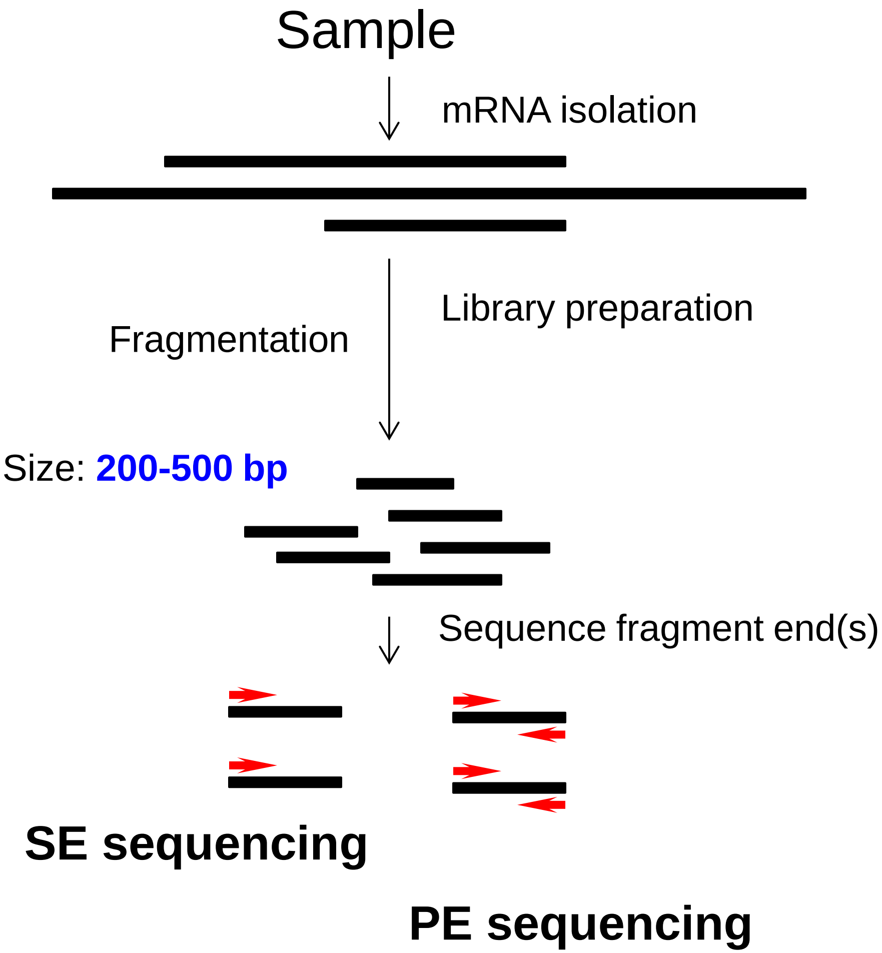
Identify Differentially Expressed Genes — Cuffdiff

Other applications: De novo Assembly Refine gene models

GeneMark, FGeneSH
Trinity, TransABySS
BLAT, MAKER

Library construction and sequencing design decisions

Library type (SE/PE) and insert size



Sample

mRNA isolation

Fragmentation

Library preparation

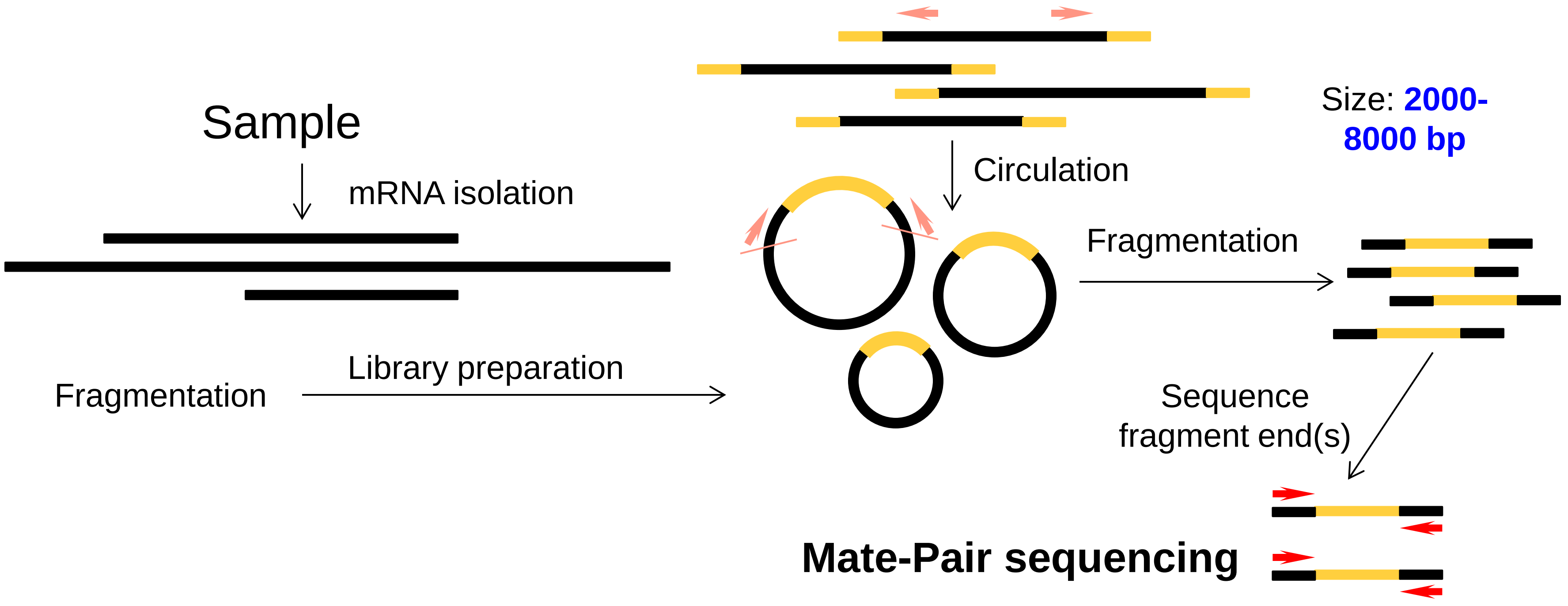
Size: **200-500 bp**

Sequence fragment end(s)

SE sequencing

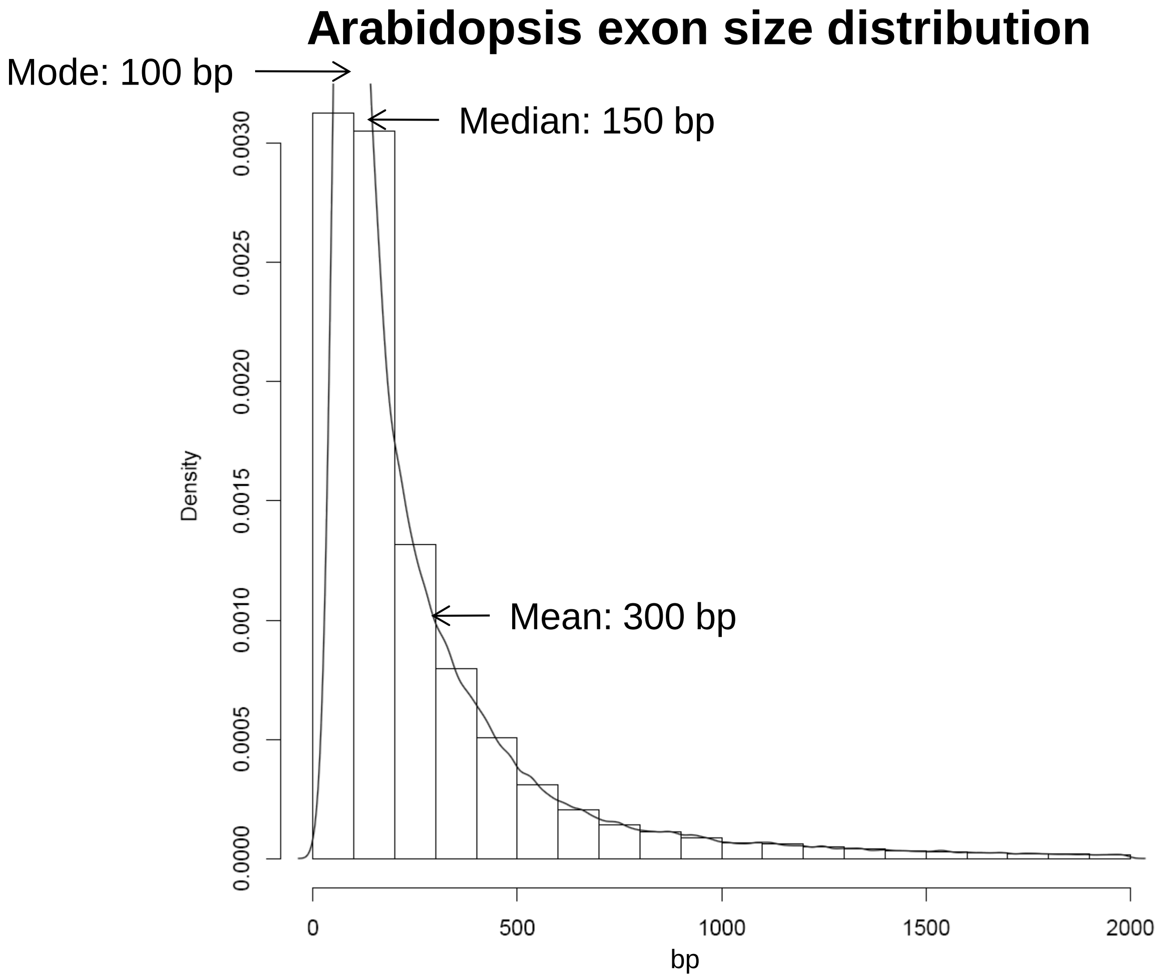
PE sequencing

Library type (Mate-pair) and insert size

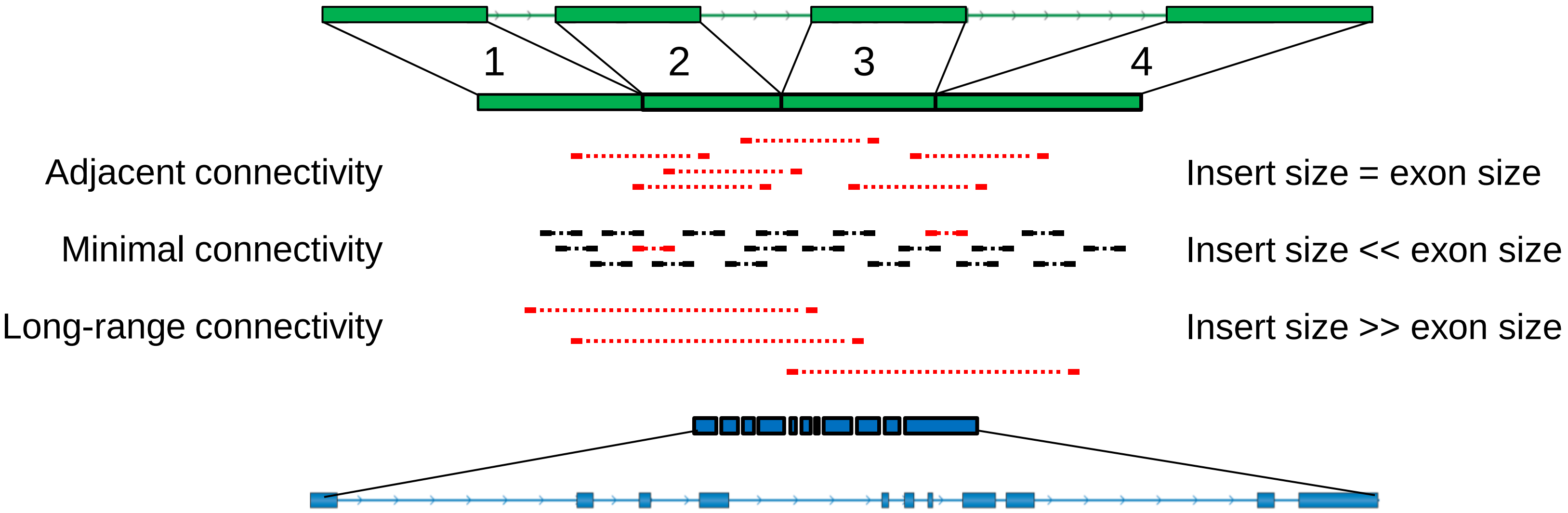


Mate-Pair sequencing

Optimal library size depends on goals and organism: ***exon size***



Optimal library size depends on goals and organism: ***exon size***



Adjacent connectivity

Insert size = exon size

Minimal connectivity

Insert size << exon size

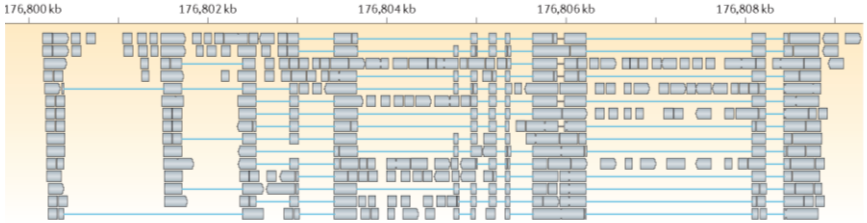
Long-range connectivity

Insert size >> exon size

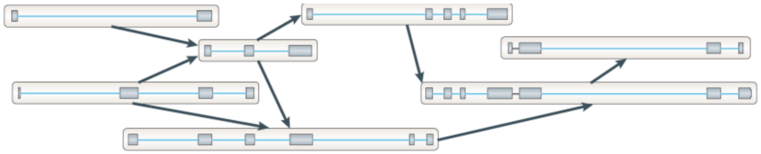
One size doesn’t fit all: organisms can differ in exon size distribution

How does connectivity play into the analysis?

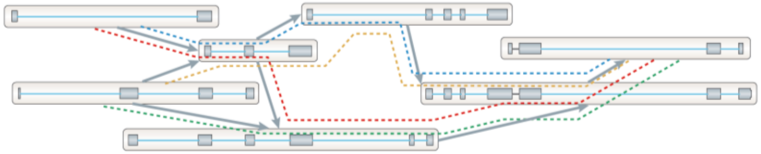
1. splice-align reads to the genome



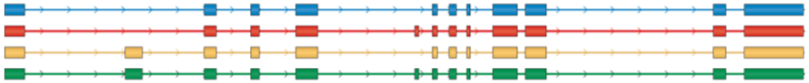
2. Build a graph representing alternative splicing events



3. Traverse the graph to assemble variants

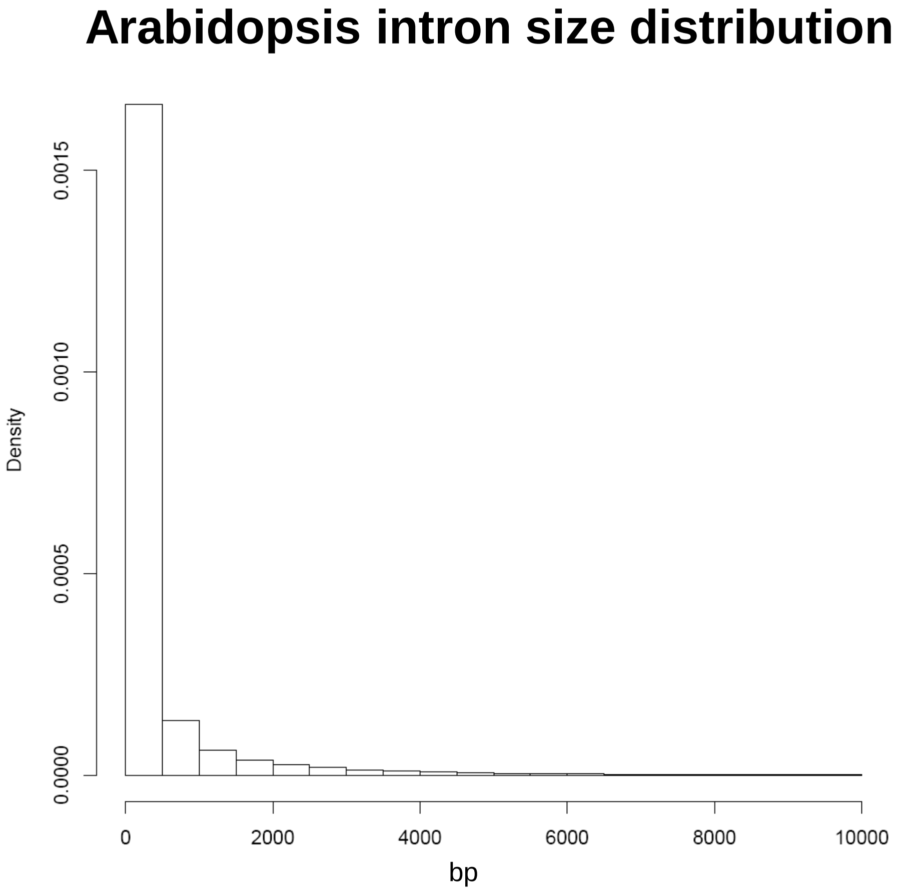


4. Assemble isoforms



Martin JA and Wang Z. Nat Rev Genet. (2011) 12:671–682.

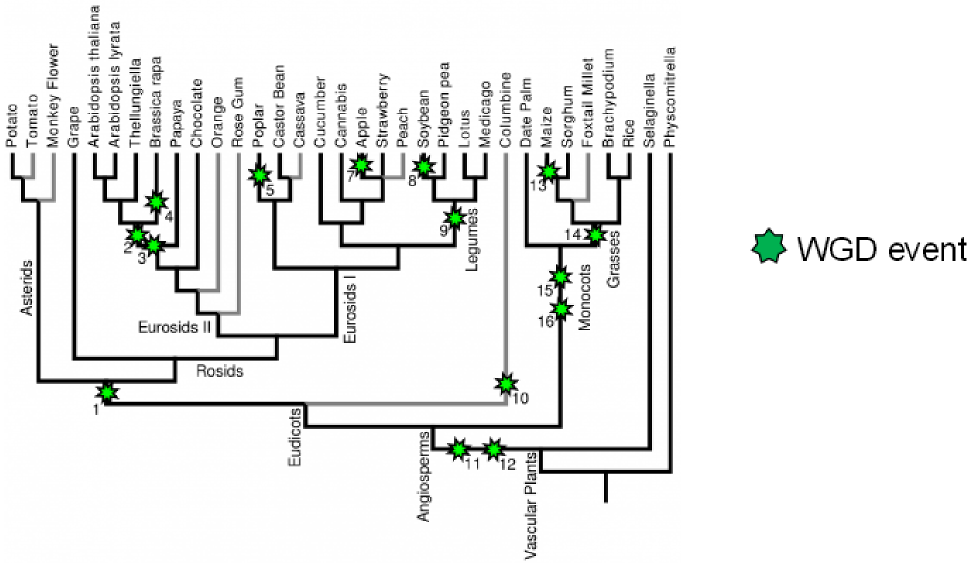
Some algorithms (e.g., tophat) exhaustively look for candidate splices in a specified distance pegged to the expected intron size distribution (default 70-500,000)



Why not just leave the defaults? (e.g., 70-500,000 bp)

- ~3500 Arabidopsis introns < 70 bp
- Huge increase in computation time
- Will accumulate spurious long-range splice junctions

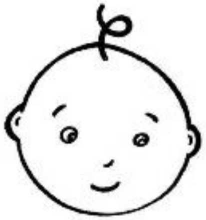
Many plant genomes have undergone ancient Whole Genome Duplications (WGDs)

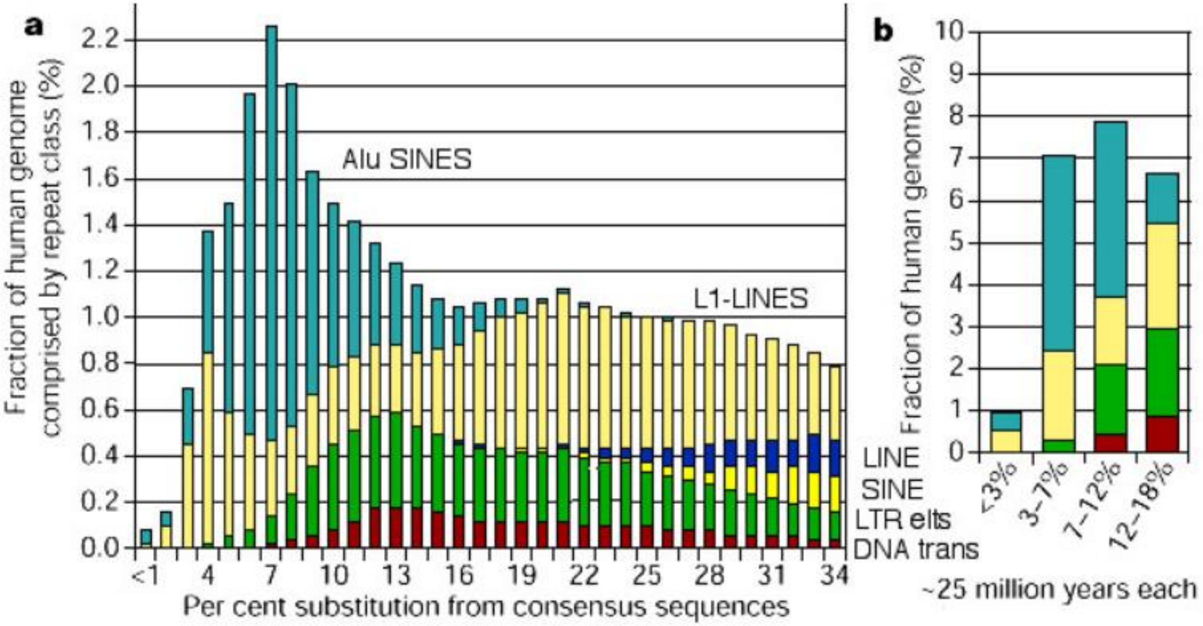


http://genomevloution.org

- Difficulty mapping uniquely to related gene family members
- Abundance levels (e.g., FPKMs) can become skewed for members of large gene families
- Both PE strategies and longer reads help to distinguish paralogs

Some genomes are rife with repetitive elements

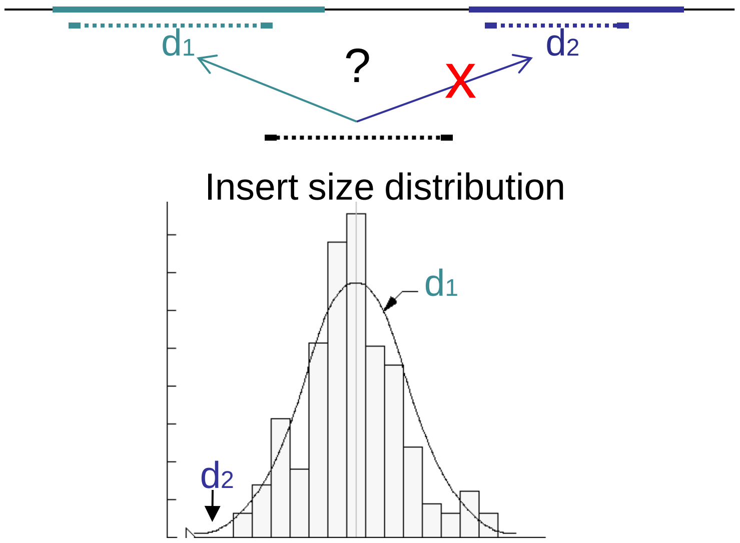




http://genomevoloution.org

- 50%, 65% of the human and maize genome are repeat elements, respectively (repbase, Kronmiller et al., Plant Phys 2008;)
- PE, mate-pair strategies and multiple insert sizes help to uniquely map repeats
- Long reads can help for small-scale or simple repeats

Why is PE is crucial for repetitive genomes and those with paralogous gene families?



2 x 50 bp is better than 1 X 100 bp for most applications and systems.

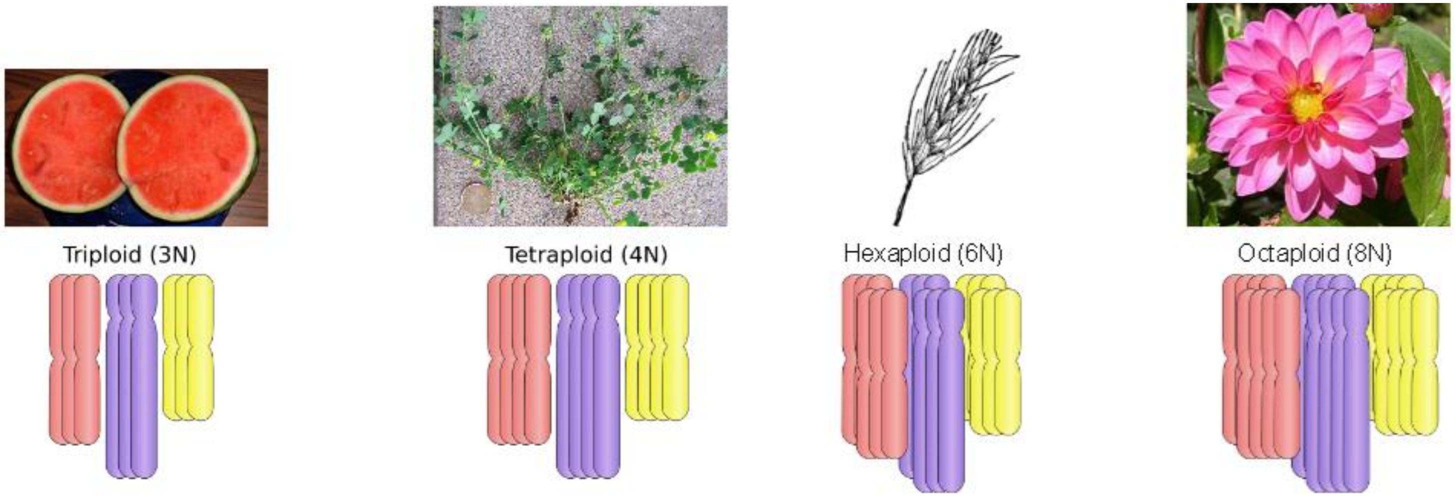
Sequencing depth needed depends on transcriptome size and the project goals

- **Sequencing Depth** is the average read coverage of target sequences
  - Sequencing depth = total number of reads X read length / estimated target sequence length
  - Example, for a 5MB transcriptome, if 1Million 50 bp reads are produced, the depth is 1 M X 50 bp / 5M ~ 10 X

- Average coverage may be misleading, since expression levels can vary more than 5 orders of magnitude!

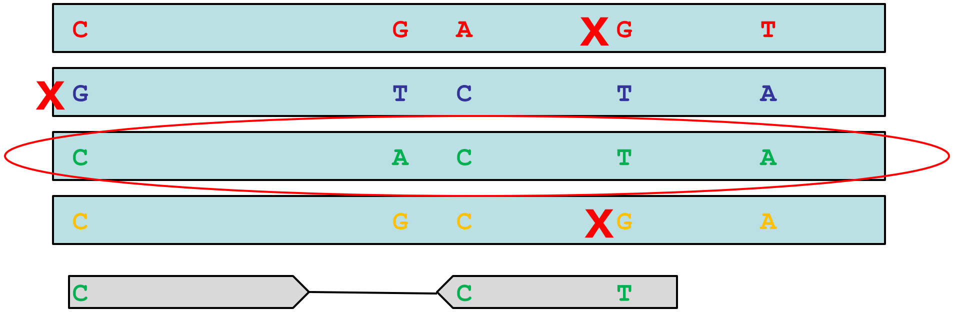
- Differential expression requires less depth than assembly, gene model refinement and structural variant discovery.

Polyploidy is particularly problematic



- Difficult to distinguish alleles from paralogs
- Genome assembly often intractable
- Need care in design of transcriptome experiment

Certain applications and biological systems will require special design considerations for maximal resolution



- **Polyploid genomes may require long reads, multiple insert sizes and custom software to distinguish among highly similar alleles at each locus.**
- **Ditto for those who wish to interrogate allele-specific differential expression (e.g., maternal or paternal impriting).**

Genome size characteristics (iGenomes)

Species	Number of genes	Transcriptome size (Mbp)	Mode| Avg **exon** size	**Intron** size range (1%|99%)	% genome repetitive	% genes in families*
Homo sapiens	29230	70.1	100|300	77|107000	47	20
Mus musculus	24080	61.4	100|300	78|100000	44	NA
*Gallus gallus***	4906	11.1	100|230	73|120000	10	NA
Drosophila melanogaster	18436	30.1	150|450	30|25000	32	7
Caenorhabditis elegans	23933	28.0	110|220	43|8000	4	24
Arabidopsis thaliana	27278	51.1	70|300	46|4900	9	35
Saccharomyces cerevisiae	6692	8.9	75|1200	20|2600	1	36
*Escherichia coli****	4290	0.6	NA	NA	3	52

* % genes with at least one paralog in the COG database (unicellular) or included in the COG lineage specific expansion (LSE) list. (These percentages are likely systematic underestimates)

** Poor annotation is suspected for iGenomes UCSC-based Gallus gallus (galGal3)

*** http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/E/Esch.coli.html; ecocyc; Gur-Arie, Genome Res 2000;.

Summary of Library Construction and Sequencing Decisions

	1	2	3	4
Project Goals:	*De novo* Assembly of transcriptome	Refine gene model	Differential Gene Expression	Identification of structural variants
Library Type:	PE, Mated PE	PE, SE	PE	PE, Mated PE
Sequencing Depth:	Extensive (> 50 X)	Extensive	Moderate (10 X ~ 30 X)	Extensive

- SE may be OK for (3) DGE if you have a good annotation and a simple genome.
- Strand-specific library creation may be necessary for organisms with a large percentage of genes that overlap on opposite strands (e.g., yeast, bacteria), or if you're interested in antisense regulation.
- Consider PacBio sequencing for goals #1, #2 and #4 above!

ENCODE RNA-Seq guidelines

Sample Replicates and Pooling Decisions

	1	2	3	4
Project Goals	*De novo* Assembly of transcriptome	Refine gene model	Differential Gene Expression	Identification of structural variants
Pooling OK?	No	Yes	No	Yes, for discovery
Biological Replicates?	Yes	Yes, if not pooling	Yes	Yes, if not pooling

- Pooling may be advisable if RNA is limited or if not interested in biological variability.

As a general rule, the following biological replicates are advisable for DGE:

  - 3+ for cell lines and pooled samples
  - 5+ for inbred lines (e.g., BL6 mice, NILs, RILs)
  - 20+ for human samples

istockphoto.com

Part II

Read Mapping Statistics and Visualization

John Garbe, PhD

Mapping Statistics

How well did my sequence library align to my reference?

Mapping Statistics

- Mapping Output
  - SAM (text) / BAM (binary) alignment files
  - Summary statistics (per read library)
    - % reads with unique alignment
    - % reads with multiple alignments
    - % reads with no alignment
    - % reads properly paired (for paired-end libraries)
    - Mean and standard deviation of insert size

SAM specification: http://samtools.sourceforge.net/SAM1.pdf

Mapping Statistics

- SAM Tools
- Tophat

Mapping Statistics – SAMtools

- Galaxy
  - – NGS: SAM Tools -> flagstat
- MSI Command line
  - – Module load samtools
  - – samtools flagstat accepted_hits.bam

Mapping Statistics – SAMtools

- SAMtools output

```
% samtools flagstat accepted_hits.bam
31443374 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
31443374 + 0 mapped (100.00%:-nan%)
31443374 + 0 paired in sequencing
15771038 + 0 read1
15672336 + 0 read2
15312224 + 0 properly paired (48.70%:-nan%)
29452830 + 0 with itself and mate mapped
1990544 + 0 singletons (6.33%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Mapping Statistics – tophat

- Galaxy
  - MSI -> tophat
- Command line
  - module load tophat
  - tophat_out/align_summary.txt

Mapping Statistics – tophat

- align_summary.txt output (paired-end reads)

```
Left reads:
               Input:  12000000
              Mapped:  11392868 (94.9% of input)
            of these:   4329227 (38.0%) have multiple alignments (111 have >20)
Right reads:
               Input:  12000000
              Mapped:  11211546 (93.4% of input)
            of these:   4231651 (37.7%) have multiple alignments (105 have >20)
94.2% overall read alignment rate.

Aligned pairs:  10982574
     of these:   3246926 (29.6%) have multiple alignments
          and:    313704 ( 2.9%) are discordant alignments
88.9% concordant pair alignment rate.
```

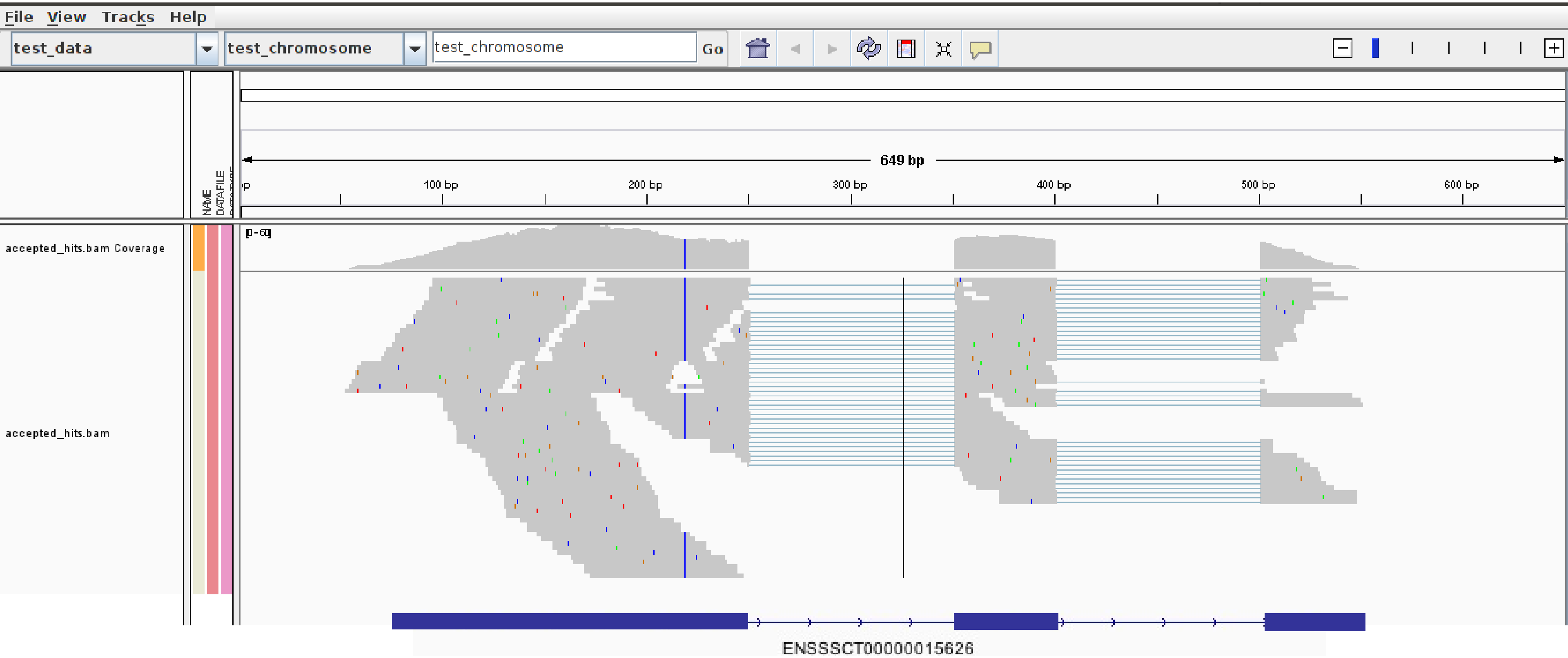
Mapping Visualization

- Integrative Genomics Viewer (IGV)
  - Fast genome browser
  - Supports array-based and next-generation sequence data, and genomic annotations
  - Free Java program

igv Integrative Genomics Viewer

http://www.broadinstitute.org/igv/home

Mapping Visualization

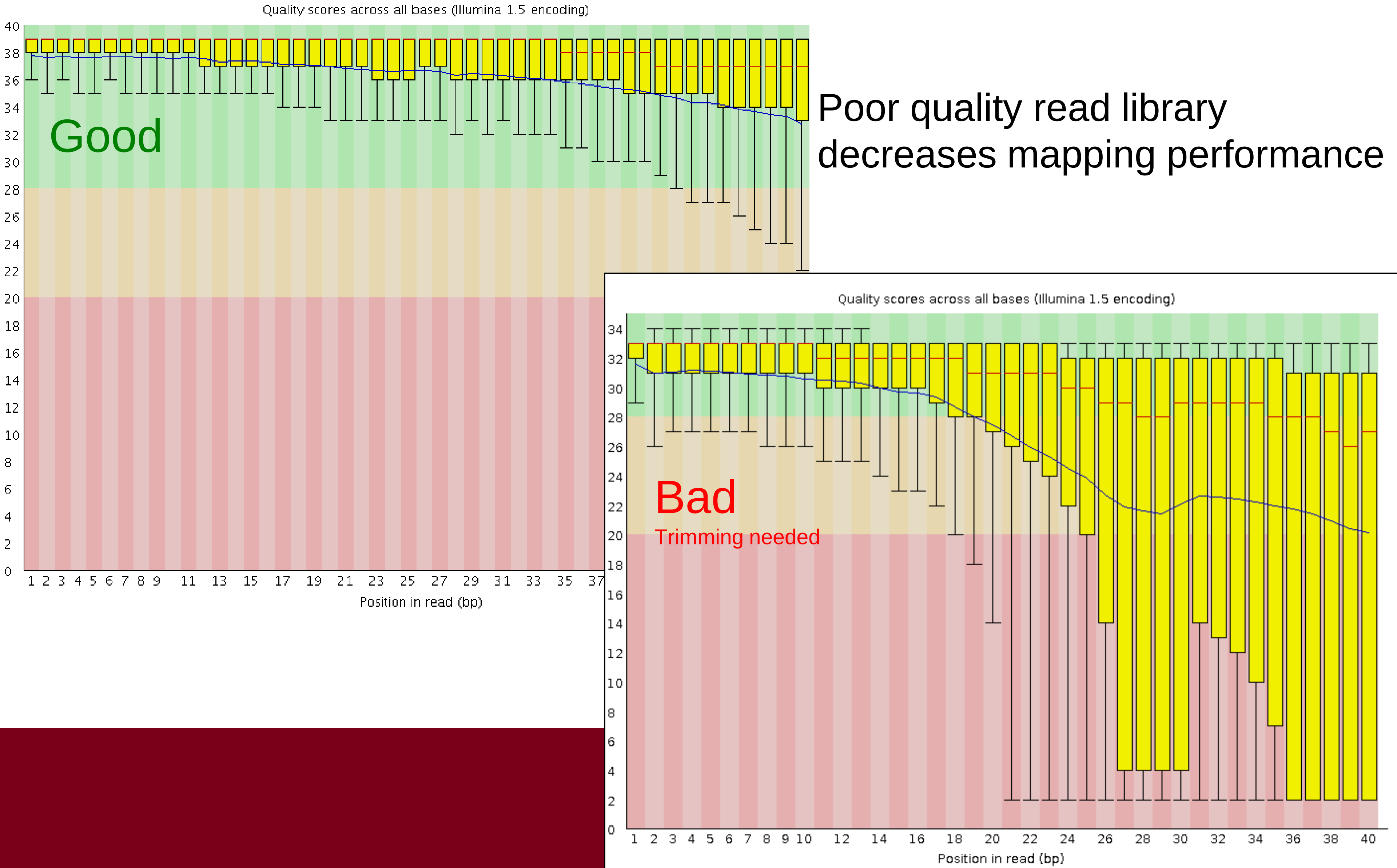


Bam file viewed with IGV

Causes of poor mapping

- Poor quality sequence library
- Contaminated sequence library
- Poor quality reference
- Divergence between sequenced population and reference
- Corrupted files
- Poor choice of mapping software
- Bug in mapping software
- Improper alignment parameters
- Repetitive genome
- Mislabeled samples
- Short read length (< 50bp)
- …

Poor Quality Library



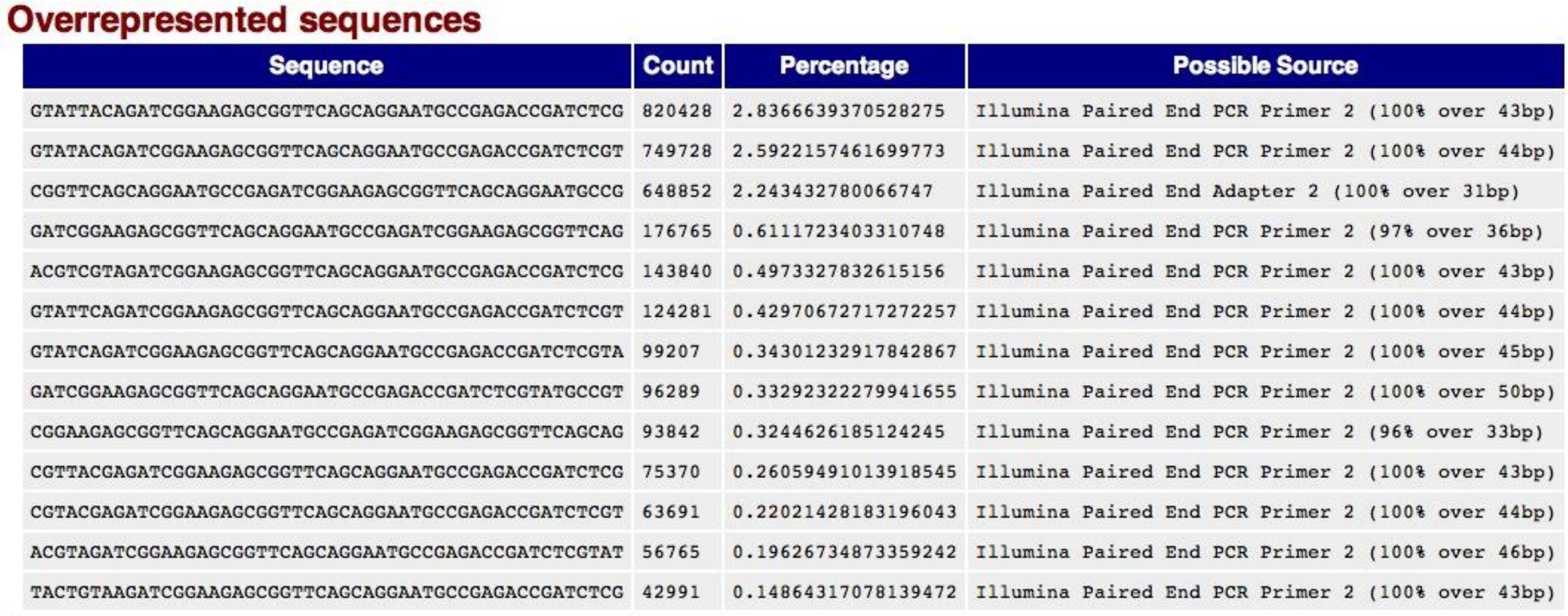
Good

Bad

Trimming needed

Poor quality read library decreases mapping performance

Contaminated sequence library



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GTATTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	820428	2.8366639370528275	Illumina Paired End PCR Primer 2 (100% over 43bp)
GTATACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	749728	2.5922157461699773	Illumina Paired End PCR Primer 2 (100% over 44bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCG	648852	2.243432780066747	Illumina Paired End Adapter 2 (100% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAG	176765	0.6111723403310748	Illumina Paired End PCR Primer 2 (97% over 36bp)
ACGTCGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	143840	0.4973327832615156	Illumina Paired End PCR Primer 2 (100% over 43bp)
GTATTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	124281	0.42970672717272257	Illumina Paired End PCR Primer 2 (100% over 44bp)
GTATCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTA	99207	0.34301232917842867	Illumina Paired End PCR Primer 2 (100% over 45bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGT	96289	0.33292322279941655	Illumina Paired End PCR Primer 2 (100% over 50bp)
CGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAG	93842	0.3244626185124245	Illumina Paired End PCR Primer 2 (96% over 33bp)
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	75370	0.26059491013918545	Illumina Paired End PCR Primer 2 (100% over 43bp)
CGTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT	63691	0.22021428183196043	Illumina Paired End PCR Primer 2 (100% over 44bp)
ACGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	56765	0.19626734873359242	Illumina Paired End PCR Primer 2 (100% over 46bp)
TACTGTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG	42991	0.14864317078139472	Illumina Paired End PCR Primer 2 (100% over 43bp)

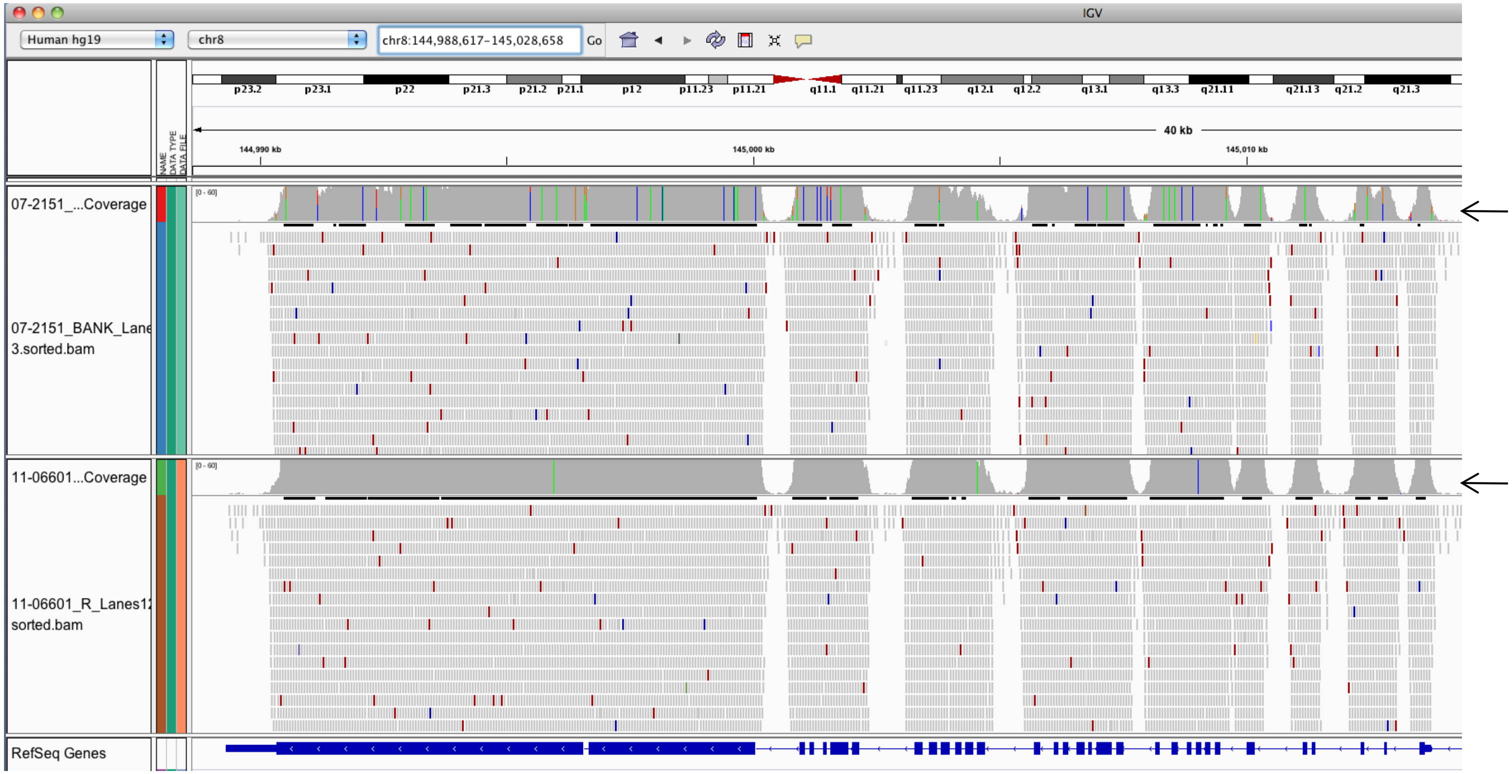
FastQC output showing ~10% adapter contamination

Poor Quality Reference

Sus scrofa 9.2	Sus scrofa 10.2	
46%	48%	mapped, properly paired
17%	20%	mapped, wrong insert size
9%	9%	singleton
26%	22%	no mapping

Mapping performance improves due to improvement in Pig genome build

Divergence between sequenced population and reference



Human hg19

chr8

chr8:144,988,617-145,028,658

← Many SNPs

← Few SNPs

Large and small sequence divergence between two human samples and the human reference genome

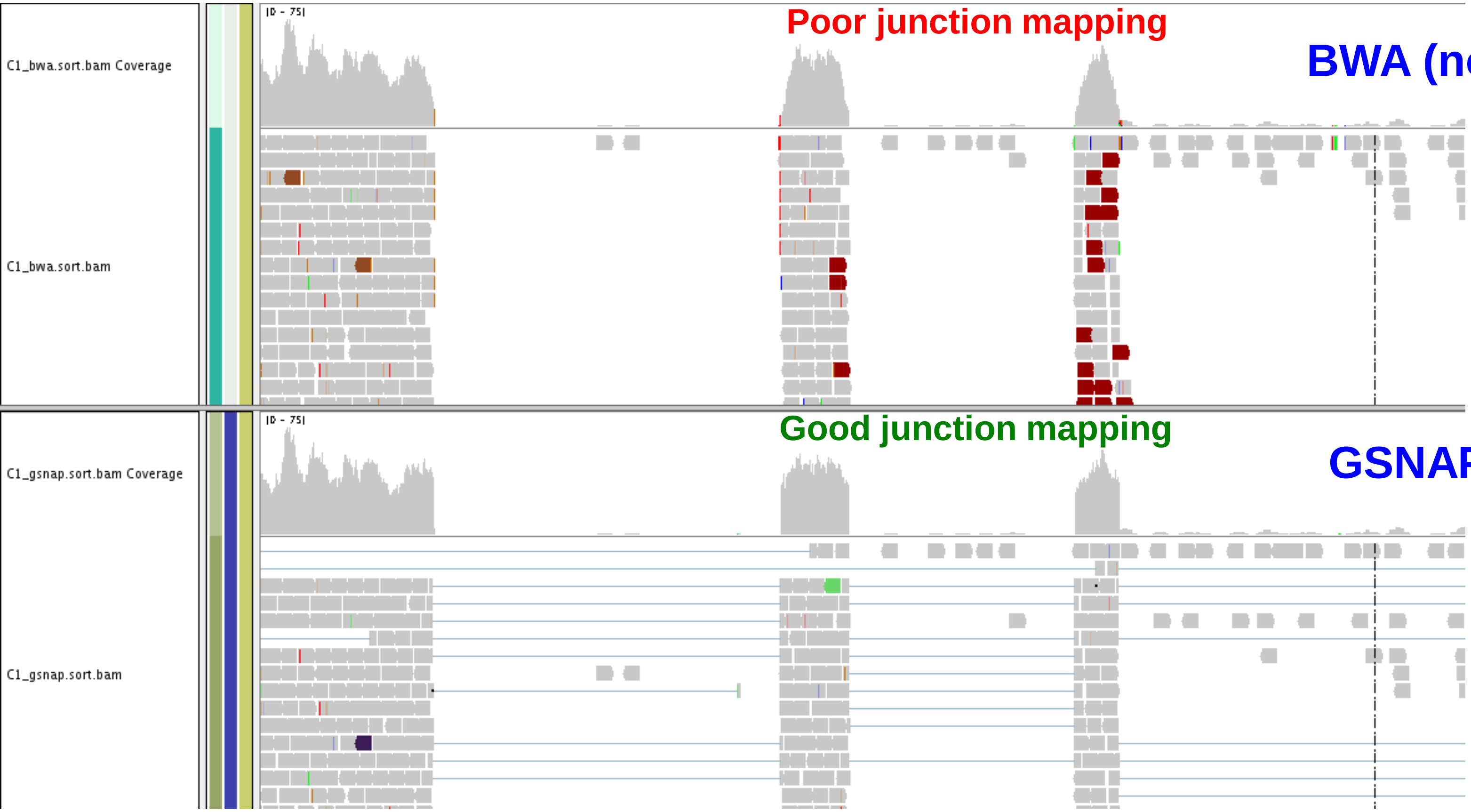
Corrupted files

R1.fastq	R2.fastq
Read1	Read 1
Read 2	Read 2
Read 3	Read 4
Read 4	Read 5

Correct fastq file	Corrupted fastq file	
48%	22%	mapped, properly paired
20%	46%	mapped, wrong insert size
9%	10%	singleton
22%	22%	no mapping

Unsynchronized paired-end fastq file decreases percentage of properly-paired reads

Poor choice of mapping software



Poor junction mapping

BWA (not splice aware)

Good junction mapping

GSNAP (splice aware)

Bug in software

Tophat 2.0.0	Tophat 2.0.1	
35%	48%	mapped, properly paired
33%	20%	mapped, wrong insert size
10%	9%	singleton
22%	22%	no mapping

New “bugfix” release of Tophat improves mapping performance

Improper alignment parameters

Correct inner distance (60)	Incorrect inner distance (220)	
48%	43%	mapped, properly paired
20%	25%	mapped, wrong insert size
9%	10%	singleton
22%	22%	no mapping

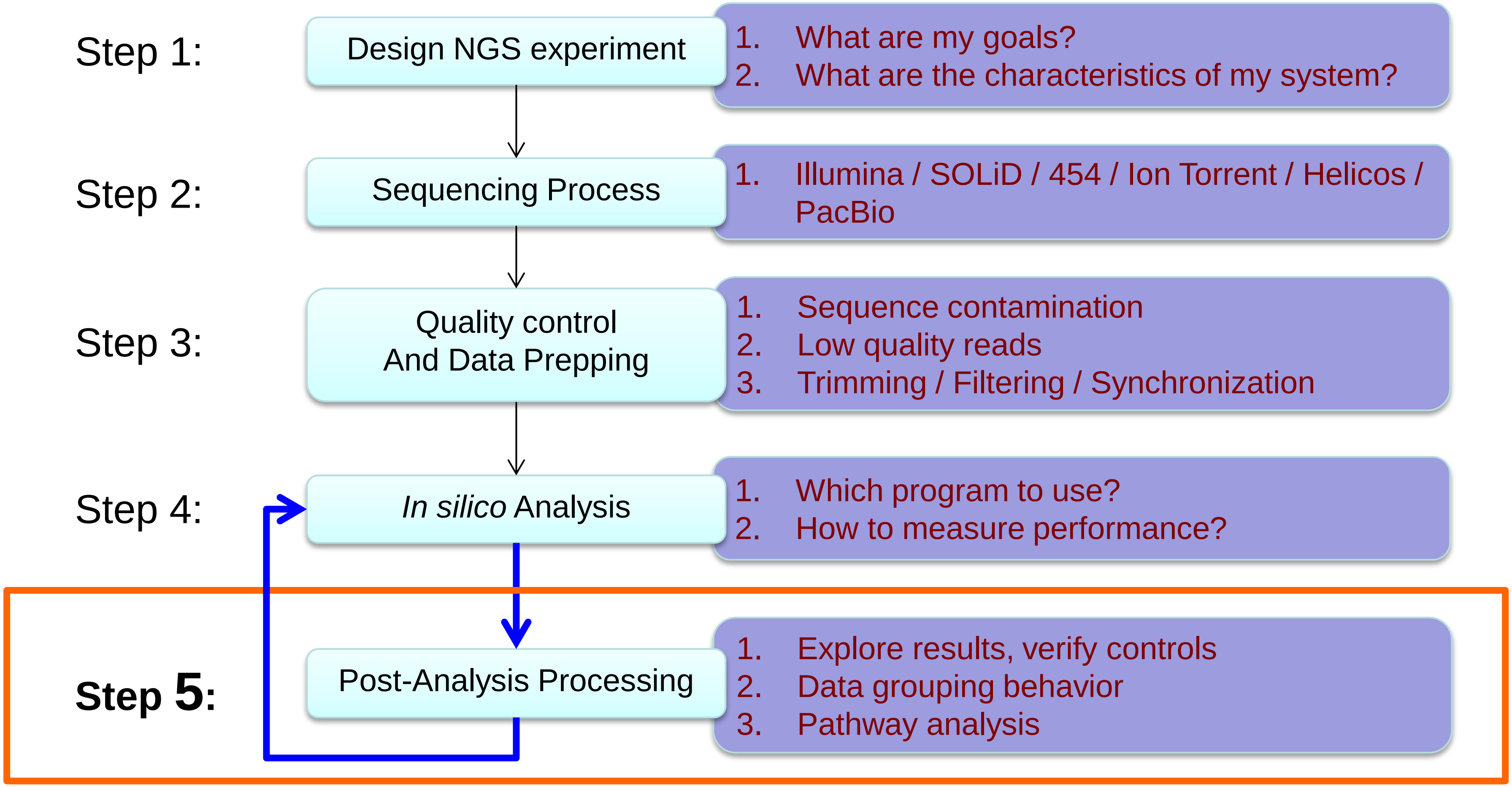
Incorrect “inner mate pair distance” parameter decreases mapping performance

Part III

Post-Analysis Processing - Exploring the Data and Results

Ying Zhang, PhD

Workflow of a typical NGS project



Step 1: Design NGS experiment

1. What are my goals?
2. What are the characteristics of my system?

Step 2: Sequencing Process

1. Illumina / SOLiD / 454 / Ion Torrent / Helicos / PacBio

Step 3: Quality control And Data Prepping

1. Sequence contamination
2. Low quality reads
3. Trimming / Filtering / Synchronization

Step 4: *In silico* Analysis

1. Which program to use?
2. How to measure performance?

Step 5: Post-Analysis Processing

1. Explore results, verify controls
2. Data grouping behavior
3. Pathway analysis

Widely-used Tools for Data Exploration

- Direct visualization of “positive controls”:
  - IGV viewer
  - UCSC Genome Browser
- Statistical checks of data structure:
  - PCA: principle component analysis
  - MDS: multi-dimension scaling
  - Unsupervised clustering and Heatmap
- System-level Analysis:
  - IPA: ingenuity pathway analysis

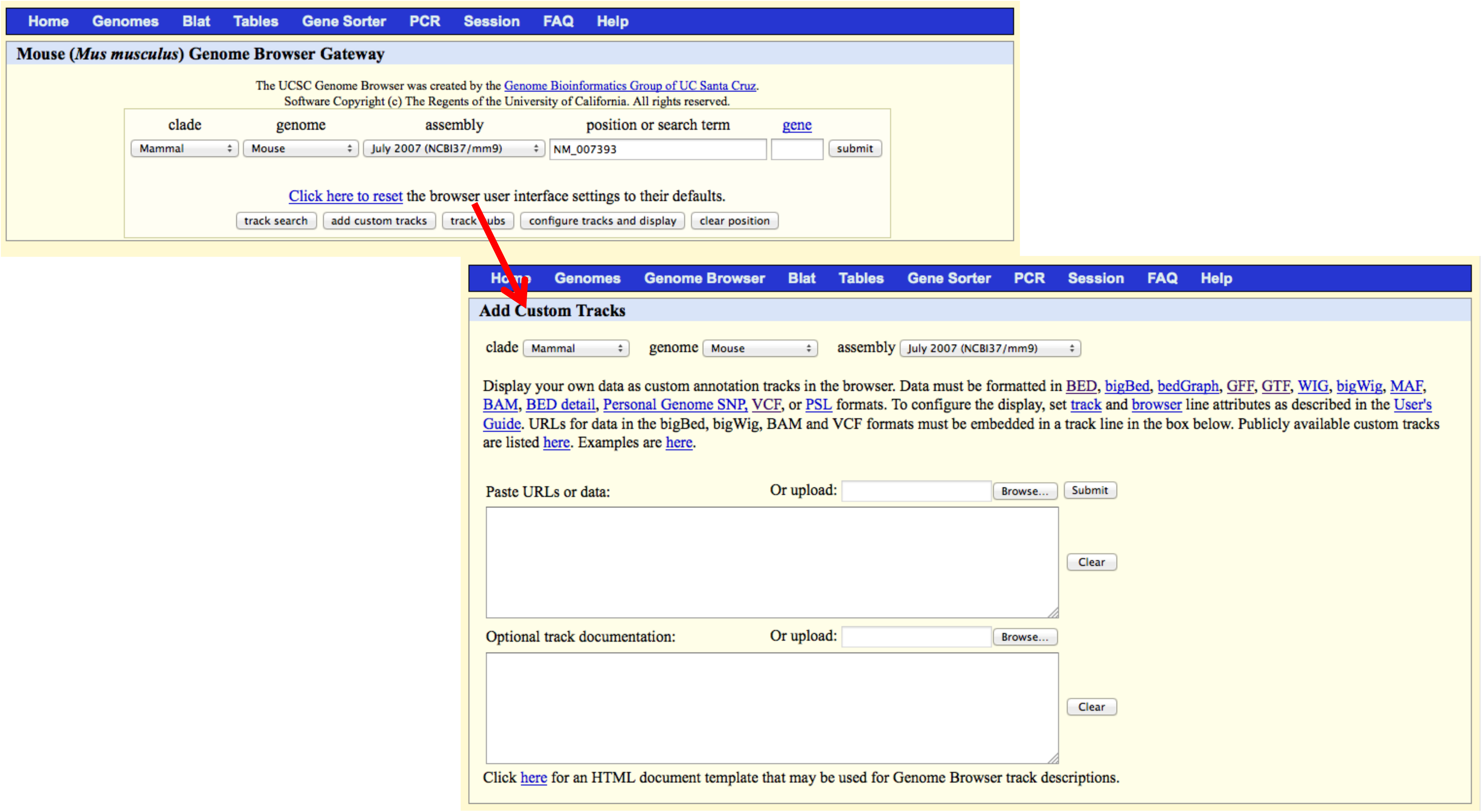
Integrative Genomics Viewer (IGV)

- Fast genome browser
- Supports array-based and next-generation sequence data, and genomic annotations
- Free Java program
- Launch:
  - From Galaxy
  - From Desktop: allocate enough memory 💣

igv Integrative Genomics Viewer

http://www.broadinstitute.org/igv/home

UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway)



Mouse (*Mus musculus*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	gene	
Mammal	Mouse	July 2007 (NCBI37/mm9)	NM_007393		submit

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

Add Custom Tracks

clade Mammal genome Mouse assembly July 2007 (NCBI37/mm9)

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

Paste URLs or data: Or upload:

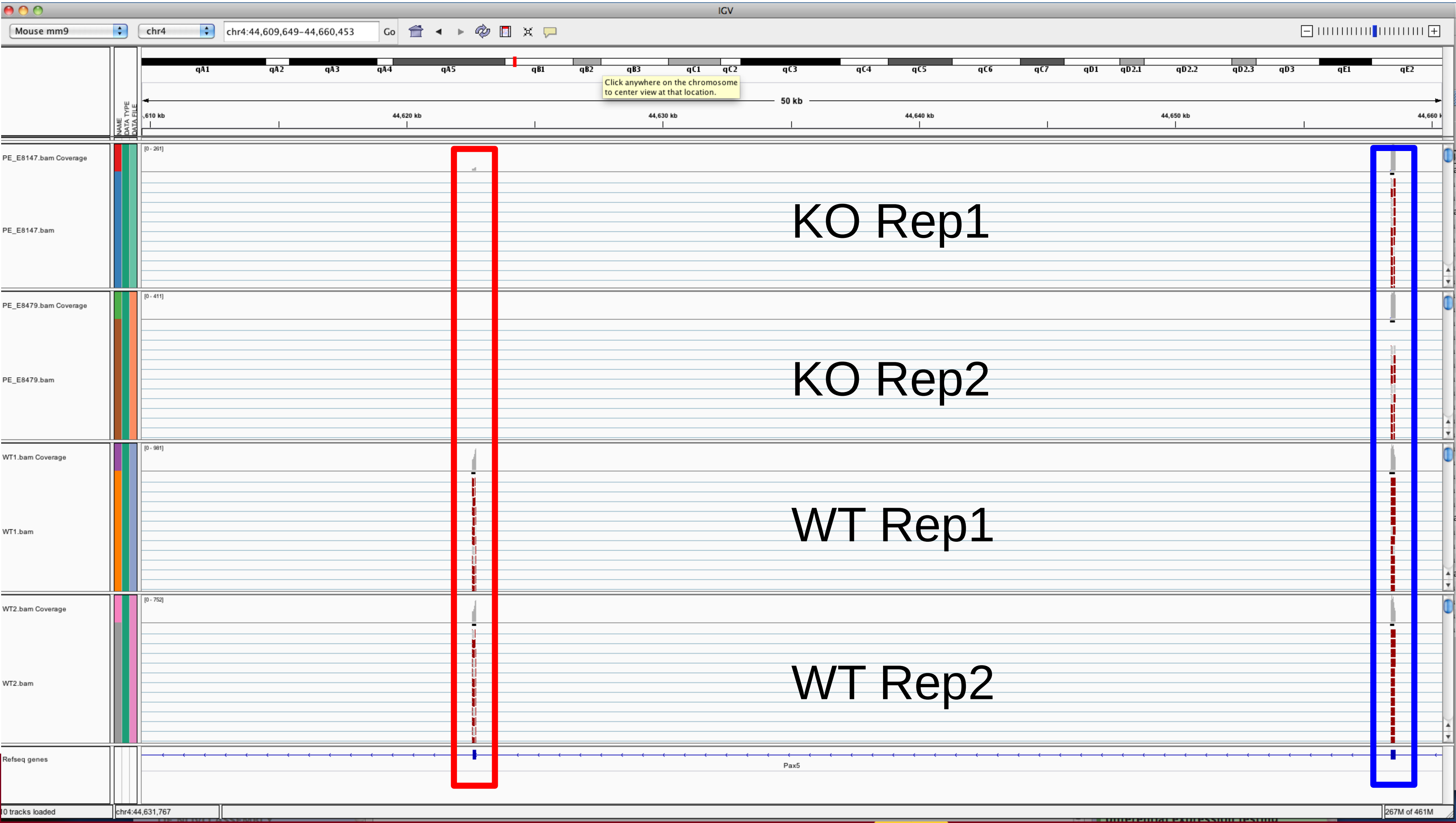
Optional track documentation: Or upload:

Click here for an HTML document template that may be used for Genome Browser track descriptions.

No. 1 in your Check-List

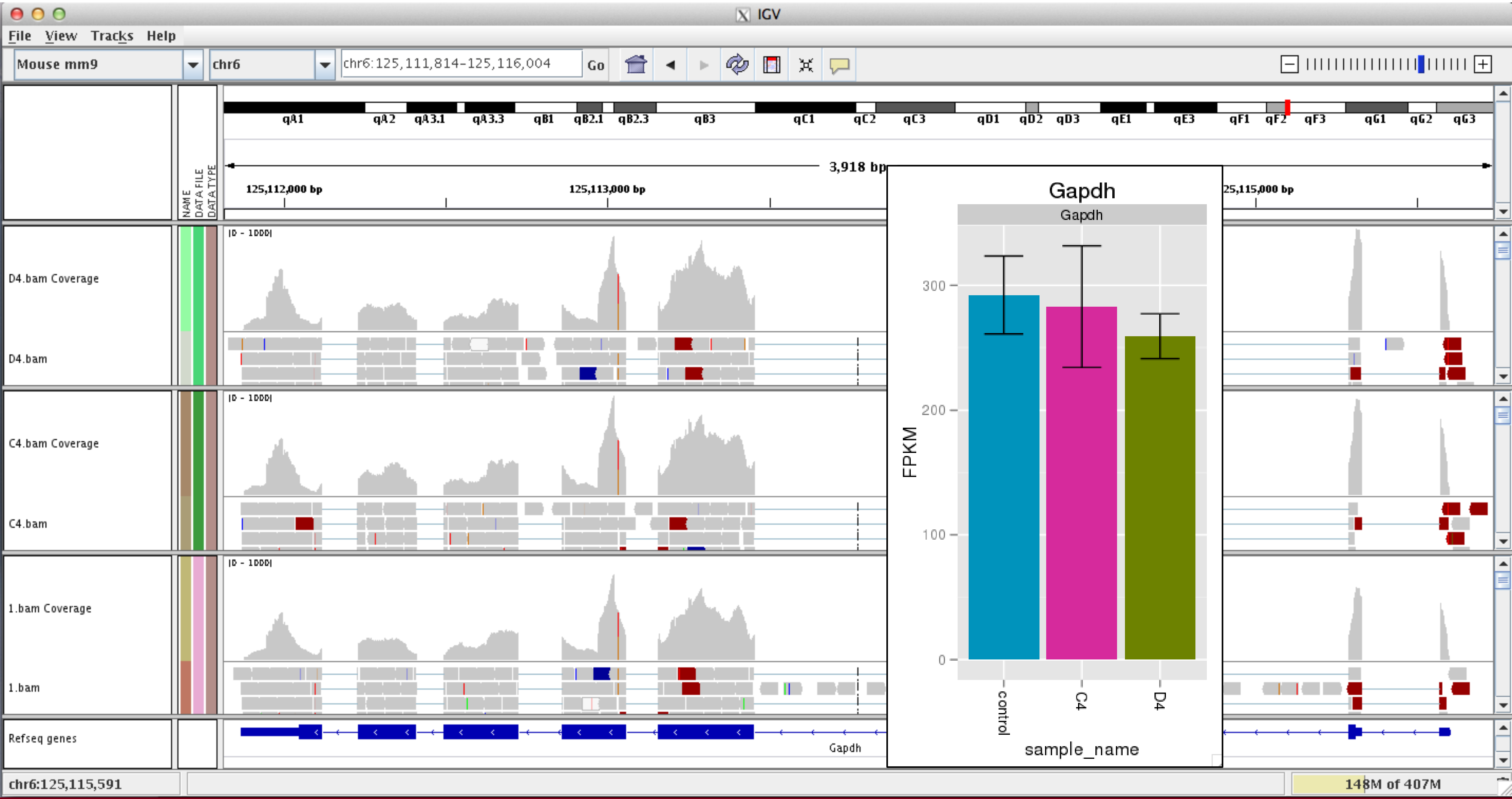
“Does my data behave as expected?”

Visualizing results–
Example I: no reads mapped at knock-out site



Data Courtesy of Dr. Mike Farrar and Dr. Lynn Harris (unpublished data)

Example II: Housekeeping genes should behave similarity across multiple samples



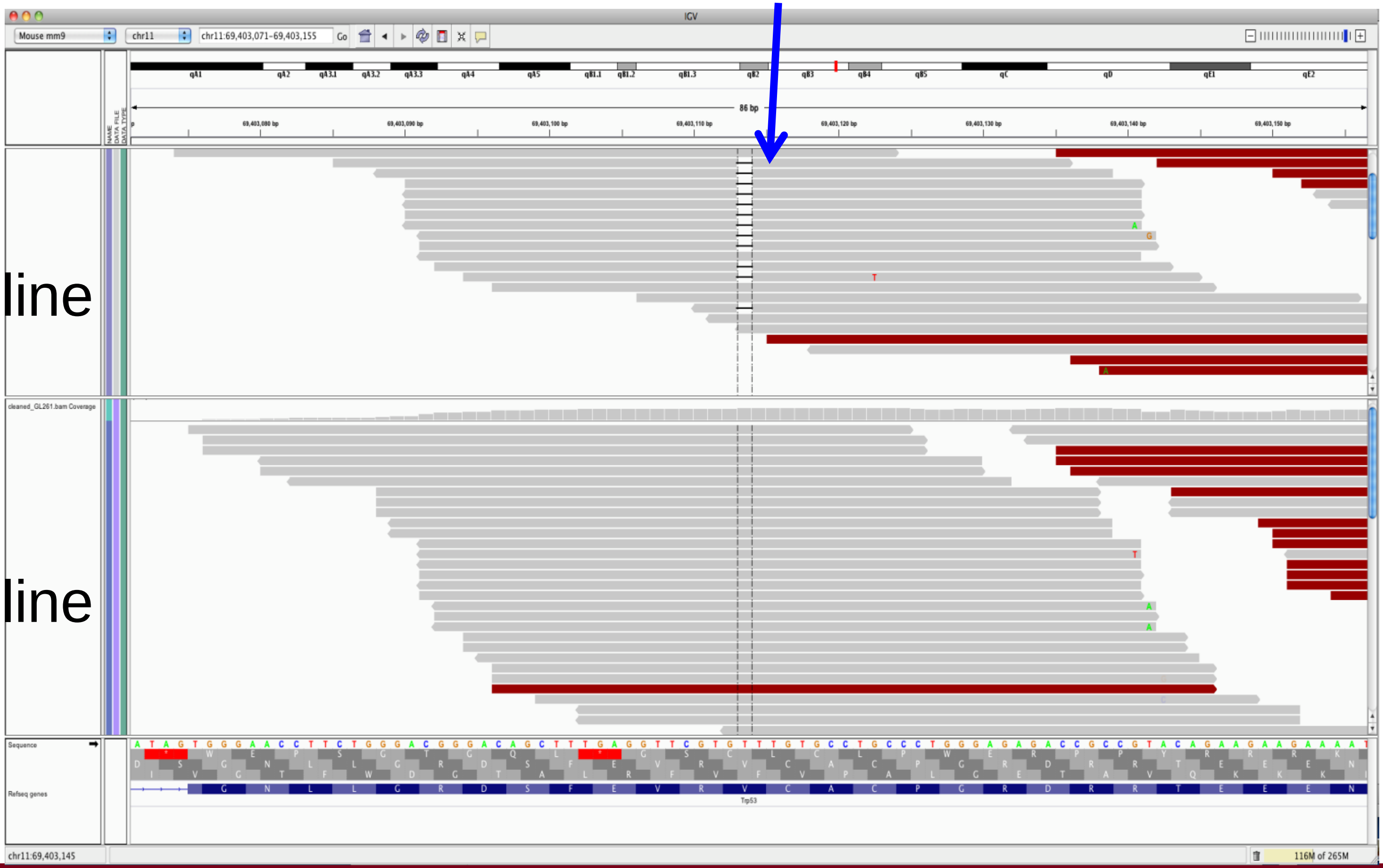
Data Courtesy of Dr. David Bernlohr and Dr. Ann Hertzel (unpublished data)

Example III: review of known biomarkers, for example, known SNPs and indels

Heterozygous deletion of 'T' with 46% penetrance

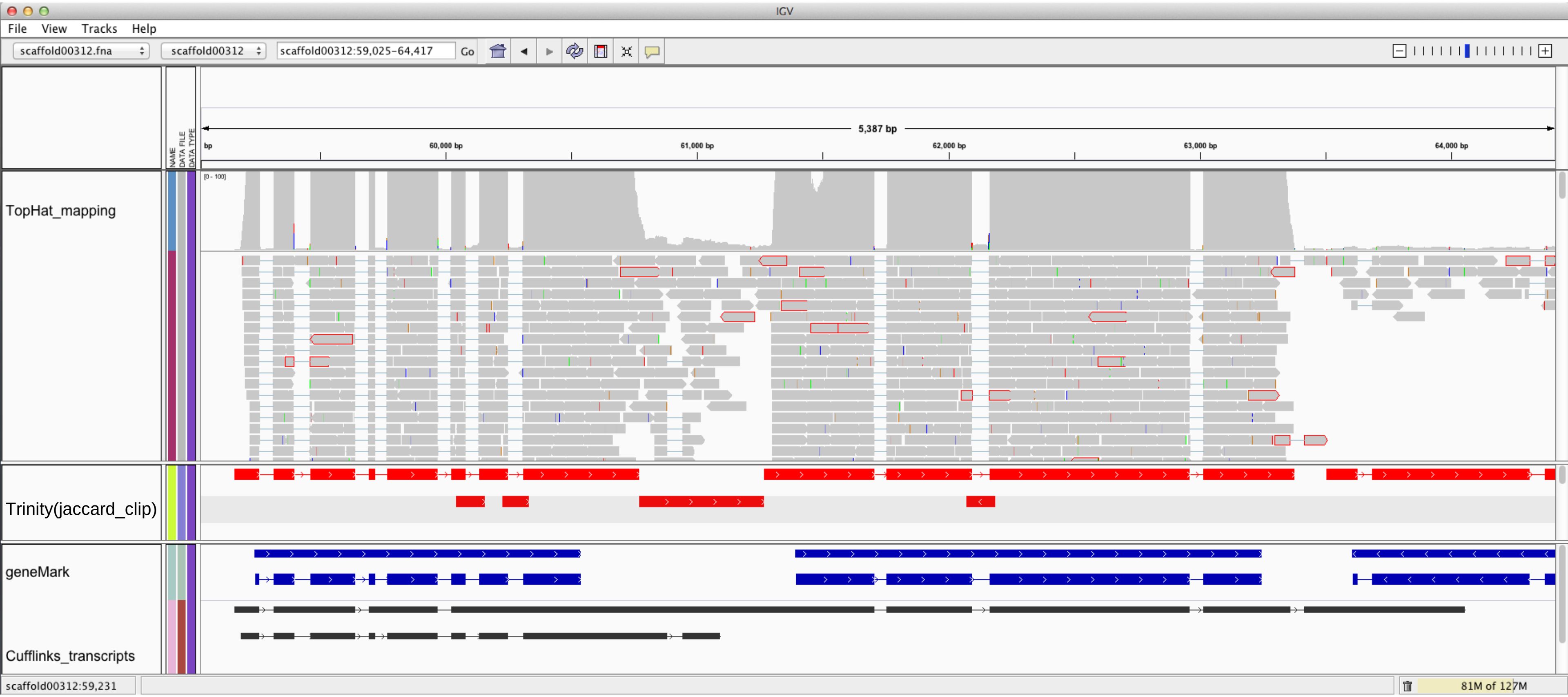
Cancer cell line

Control cell line



Data Courtesy of Dr. John Ohlfest and Dr. Flavia Popescu (unpublished data)

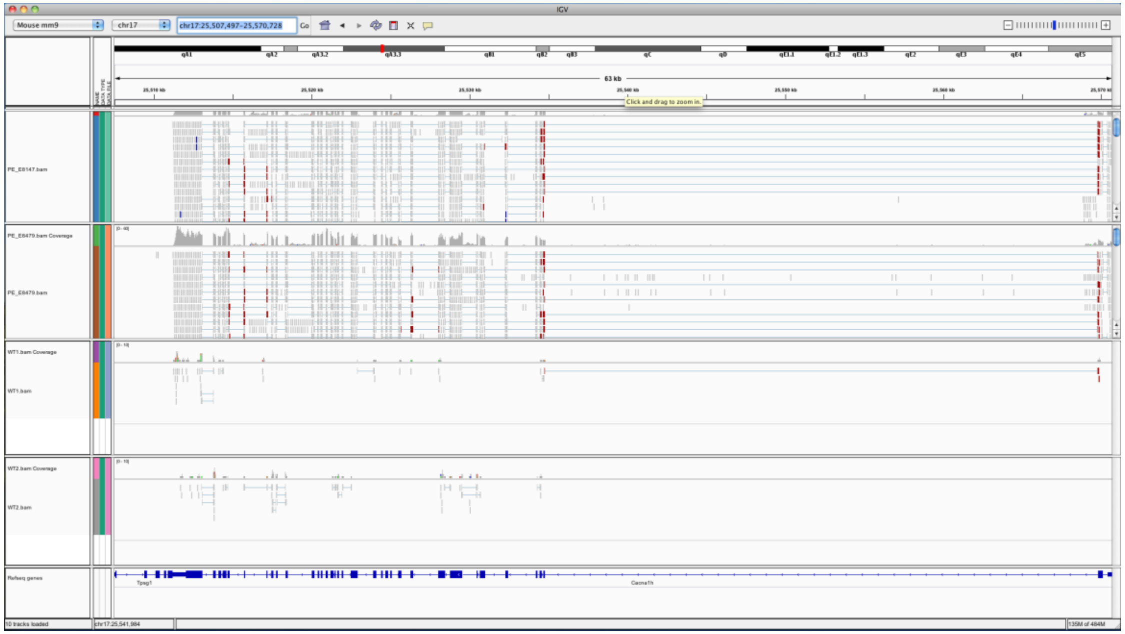
Example IV: Try different tools/parameters to identify limitations of software



Data courtesy of Dr. Steve Gantt and Dr. Karen Tang (unpublished data)

Warning: don’t throw the baby out with the bathwater…

Cuffdiff: “Min Alignment Count” must be satisfied in **all** samples – too high a value will remove genes not expressed in one condition but strongly expressed in another!



Mut Rep 1

Mut Rep 2

Wt Rep 1

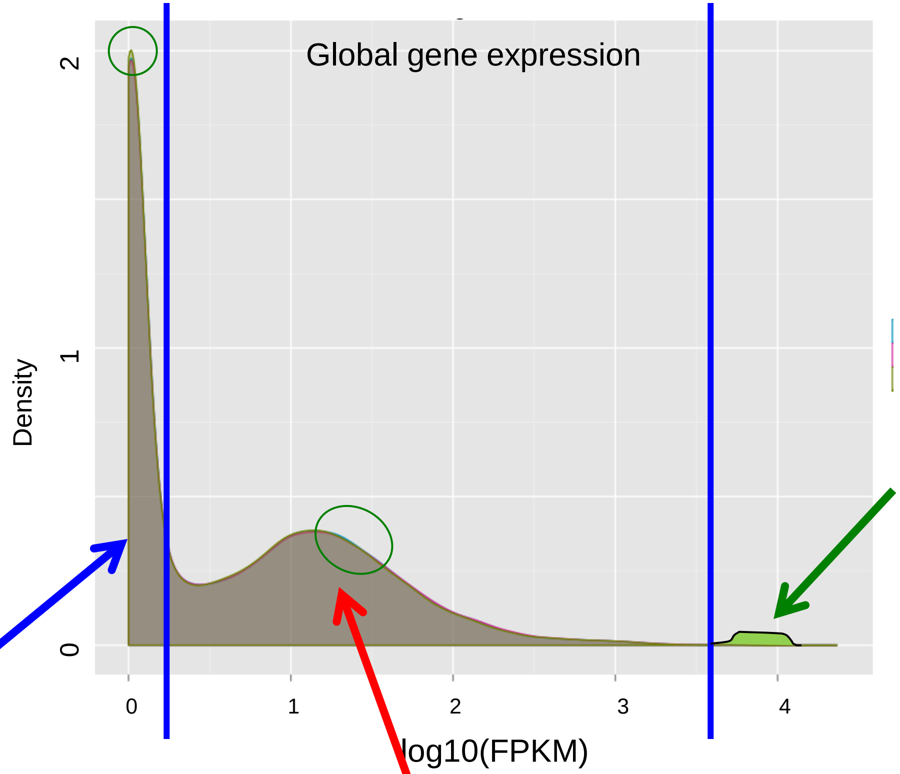
Wt Rep 2

This gene was reported as DE with “Min Alignment Count” = 10, but not with 100.

No. 2 in your Check-List

"What is the global behavior of my data?"

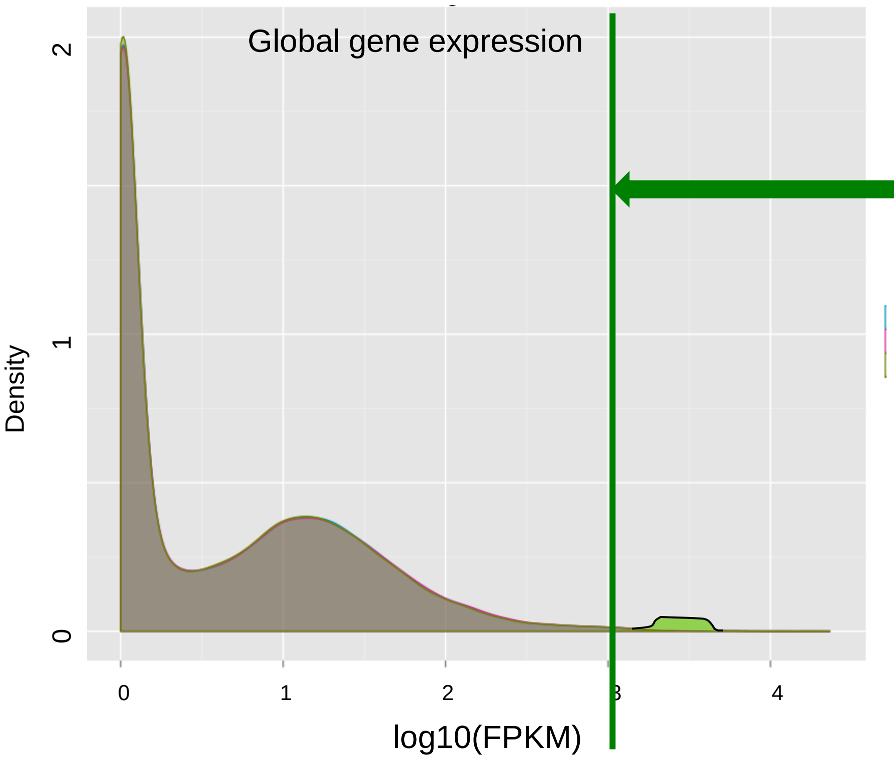
Explore the global distribution of data



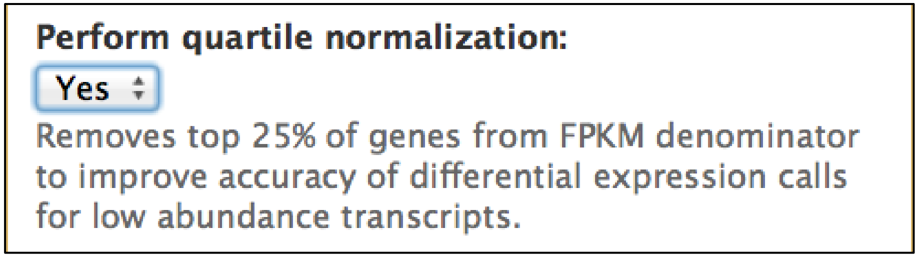
Very few genes have an usually high expression.

Many genes will have little or no expression.

A set of genes have a high expression.



Exclude the highly-expressed genes for highly-unbalanced expression between conditions.
Set “yes” to “Perform quartile normalization”.



Perform quartile normalization:

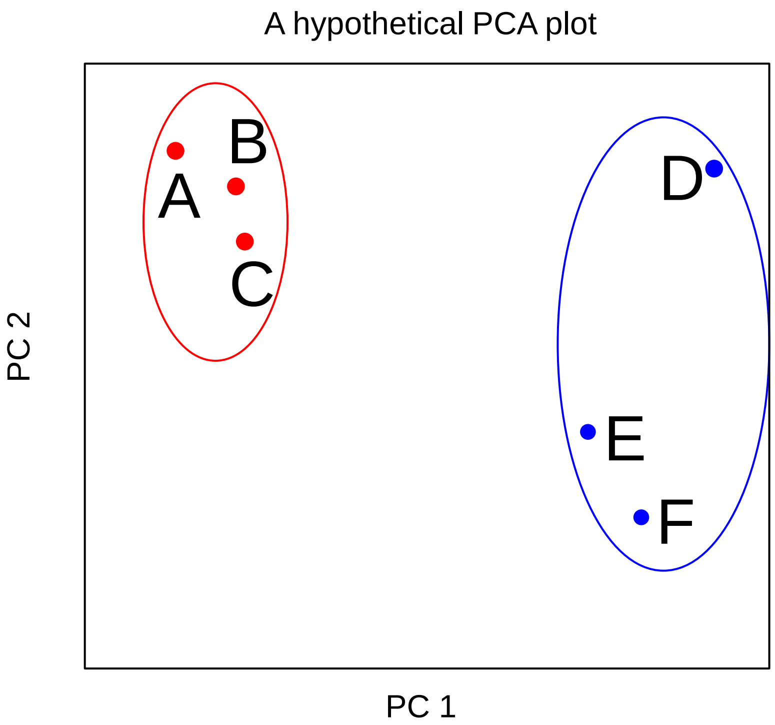
Yes

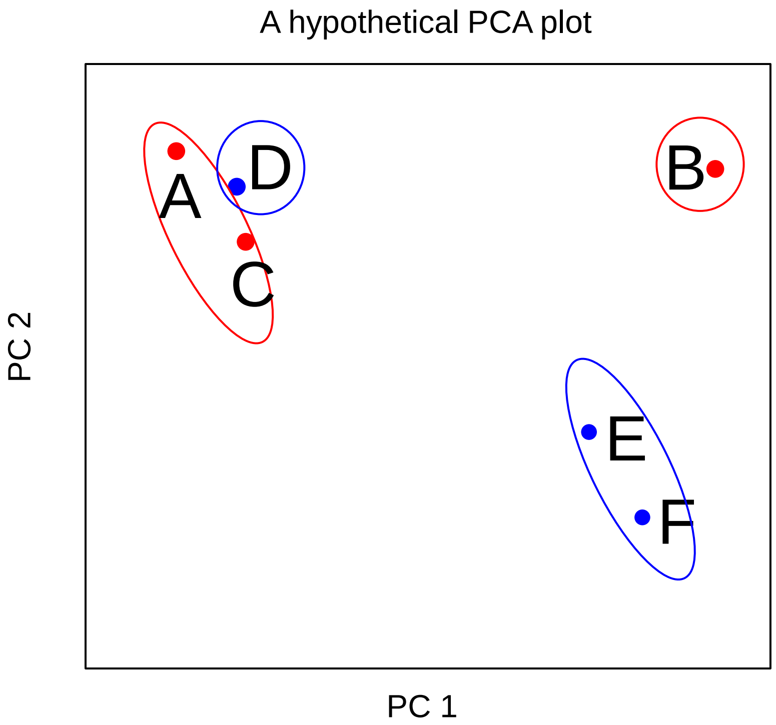
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Example: red cell blood compared to other tissue

Statistical Checks of data structure – Multi-Variable Analysis

- Biological replicates should show grouping behavior in multi-variable analysis:
  - innate consistency between samples


A hypothetical PCA plot


A hypothetical PCA plot

Within-group variation: non-biological variations

- Source of non-biological variation:
  - Batch effect
    - How were the samples collected and processed? Were the samples processed as groups, and if so what was the grouping?
  - Non-synchronized cell cultures
    - Were all the cells from the same genetic backgrounds and growth phase?
  - Use of the technical replicates (not recommended!) rather than biological replicates

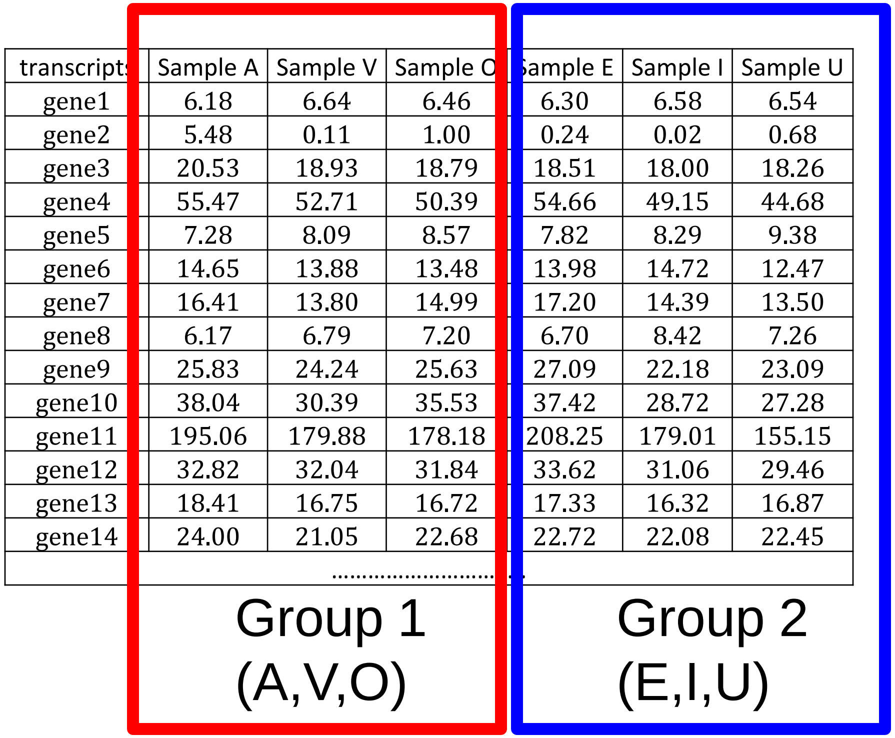
How to check for data variation?

- Principle Component Analysis (PCA)
  - Uses an orthogonal transformation
  - The first principle component has the largest possible variance
- Multi-Dimensional Scaling (MDS)
  - Computes Euclidean distances among all pairs of samples
- Unsupervised Clustering / heatmap
  - Identify the hidden structure in “unlabeled” data
- Tools:
  - Galaxy
  - Statistical Package: R, SPSS, MatLab
  - Partek and Genedata Expressionist (Analyst)

Steps in PCA analysis

1. Construct the multiple variable matrix → 2. Run PCA analysis and explore the result

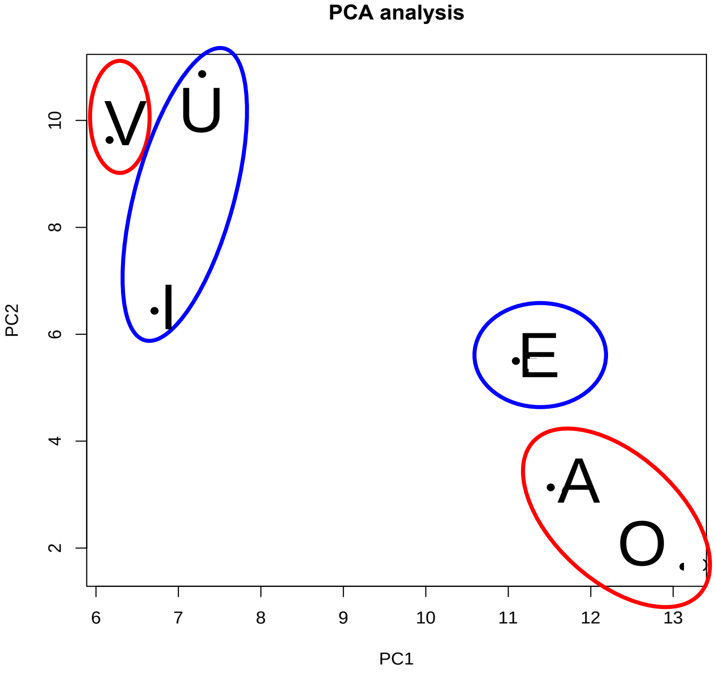
e.g. tables of FPKM values



transcripts	Sample A	Sample V	Sample O	Sample E	Sample I	Sample U
gene1	6.18	6.64	6.46	6.30	6.58	6.54
gene2	5.48	0.11	1.00	0.24	0.02	0.68
gene3	20.53	18.93	18.79	18.51	18.00	18.26
gene4	55.47	52.71	50.39	54.66	49.15	44.68
gene5	7.28	8.09	8.57	7.82	8.29	9.38
gene6	14.65	13.88	13.48	13.98	14.72	12.47
gene7	16.41	13.80	14.99	17.20	14.39	13.50
gene8	6.17	6.79	7.20	6.70	8.42	7.26
gene9	25.83	24.24	25.63	27.09	22.18	23.09
gene10	38.04	30.39	35.53	37.42	28.72	27.28
gene11	195.06	179.88	178.18	208.25	179.01	155.15
gene12	32.82	32.04	31.84	33.62	31.06	29.46
gene13	18.41	16.75	16.72	17.33	16.32	16.87
gene14	24.00	21.05	22.68	22.72	22.08	22.45
			....			

Group 1
(A,V,O)

Group 2
(E,I,U)



Heatmap: Unsupervised clustering

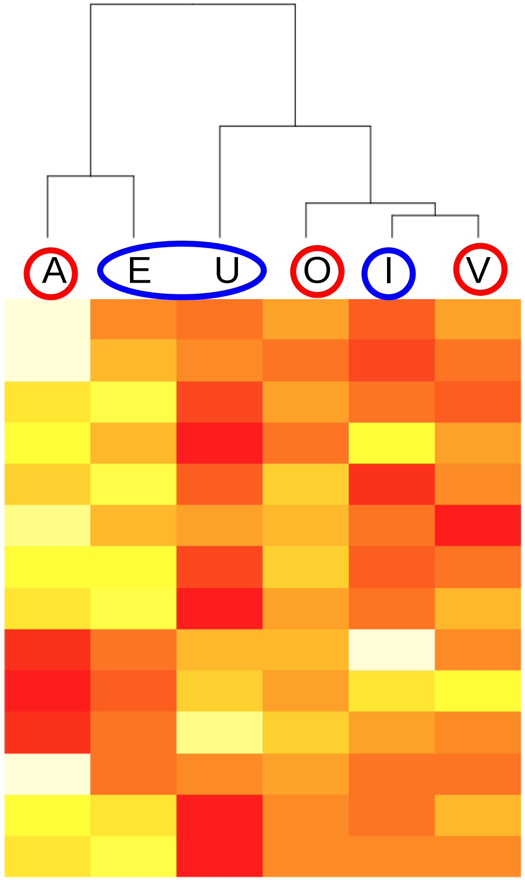
1. Construct the multiple variable matrix → 2. Run Unsupervised Clustering and generate Heatmap

e.g. tables of FPKM values

transcripts	Sample A	Sample V	Sample O	Sample E	Sample I	Sample U
gene1	6.18	6.64	6.46	6.30	6.58	6.54
gene2	5.48	0.11	1.00	0.24	0.02	0.68
gene3	20.53	18.93	18.79	18.51	18.00	18.26
gene4	55.47	52.71	50.39	54.66	49.15	44.68
gene5	7.28	8.09	8.57	7.82	8.29	9.38
gene6	14.65	13.88	13.48	13.98	14.72	12.47
gene7	16.41	13.80	14.99	17.20	14.39	13.50
gene8	6.17	6.79	7.20	6.70	8.42	7.26
gene9	25.83	24.24	25.63	27.09	22.18	23.09
gene10	38.04	30.39	35.53	37.42	28.72	27.28
gene11	195.06	179.88	178.18	208.25	179.01	155.15
gene12	32.82	32.04	31.84	33.62	31.06	29.46
gene13	18.41	16.75	16.72	17.33	16.32	16.87
gene14	24.00	21.05	22.68	22.72	22.08	22.45
			…………………………			

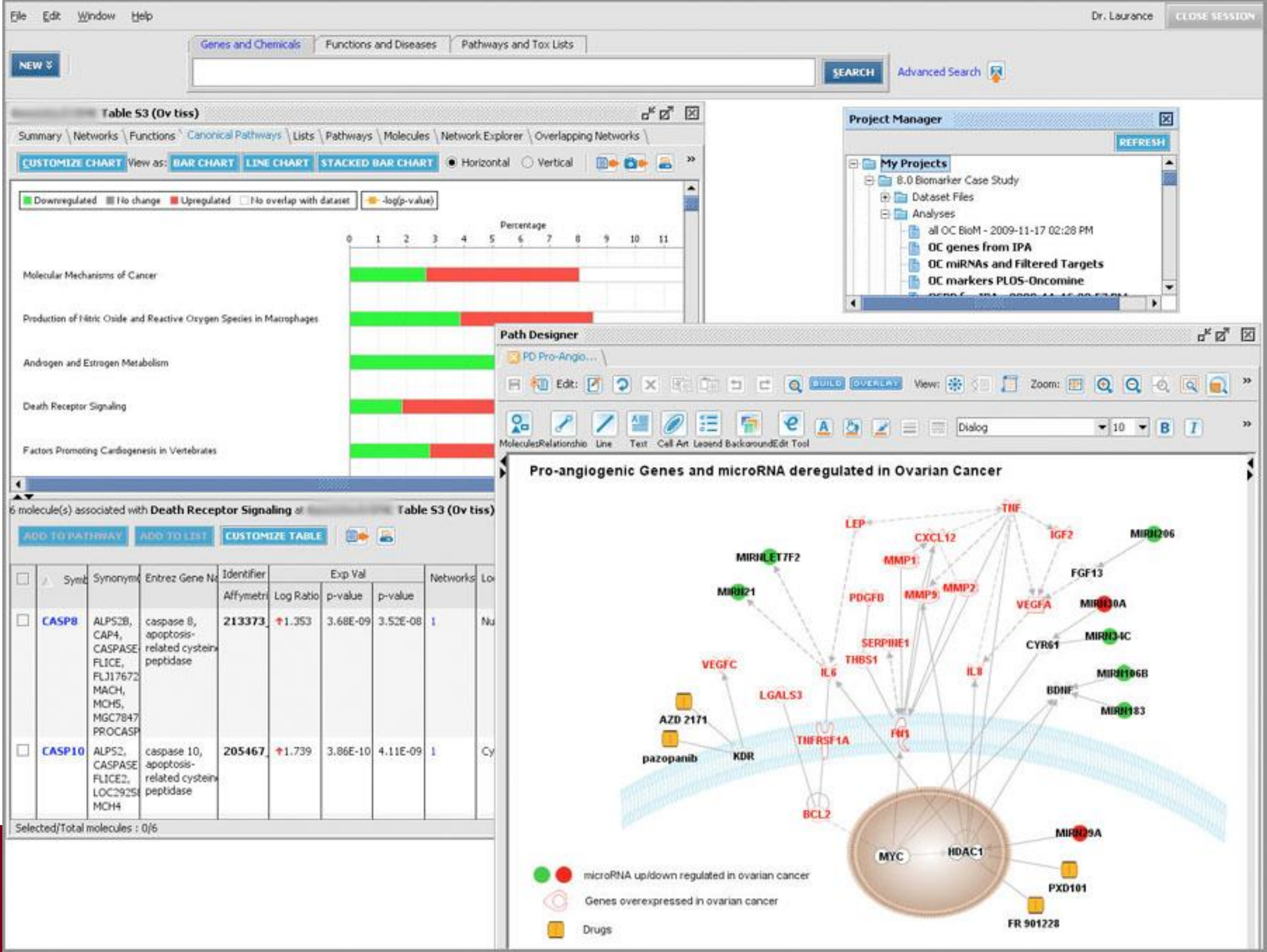
Group 1
(A,V,O)

Group 2
(E,I,U)



Exploring data at system-level: Ingenuity Pathway analysis

- Using the differentially expressed genes
- Connecting the genes with known knowledge
- Testing for the significance of the identified network
- Check the details at:
  - http://ingenuity.com/products/pathways_analysis.html
- Primarily for mammalian systems
- Consider MapMan for plants
  - http://mapman.gabipd.org/web/guest/mapman



[redacted] Table S3 (Ov tiss)

Project Manager

- My Projects
  - 8.0 Biomarker Case Study
    - Dataset Files
    - Analyses
      - all OC BioM - 2009-11-17 02:28 PM
      - **OC genes from IPA**
      - **OC miRNAs and Filtered Targets**
      - **OC markers PLOS-Oncomine**

Pro-angiogenic Genes and microRNA deregulated in Ovarian Cancer

●● microRNA up/down regulated in ovarian cancer

Genes overexpressed in ovarian cancer

Drugs

6 molecule(s) associated with **Death Receptor Signaling** at [redacted] Table S3 (Ov tiss)

□	Symb	Synonym	Entrez Gene N	Identifier	Exp Val			Networks	Lo
				Affymetri	Log Ratio	p-value	p-value		
□	**CASP8**	ALPS2B, CAP4, CASPASE FLICE, FLJ17672 MACH, MCH5, MGC7847 PROCASP	caspase 8, apoptosis- related cystein peptidase	**213373_**	↑1.353	3.68E-09	3.52E-08	1	Nu
□	**CASP10**	ALPS2, CASPASE FLICE2, LOC2925 MCH4	caspase 10, apoptosis- related cystein peptidase	**205467_**	↑1.739	3.86E-10	4.11E-09	1	Cy

Selected/Total molecules : 0/6

Discussion and Questions?

- Get Support at MSI:
  - Email: help@msi.umn.edu
  - General Questions:
    - Subject line: “RISS:…”
  - Galaxy Questions:
    - Subject line: “Galaxy:…”