

RNA-Seq Module 3

Advanced RNA-Seq Analysis Topics and Trouble-Shooting

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

May 24, 2012



UNIVERSITY OF MINNESOTA

Driven to DiscoverSM

RNA-Seq Tutorials

- Tutorial 1: Introductory (Mar. 28 & Apr. 19)
 - RNA-Seq experiment design and analysis
 - Instruction on individual software will be provided in other tutorials
- Tutorial 2: Introductory (Apr. 3 & Apr 24)
 - Analysis RNA-Seq using TopHat and Cufflinks
- **Tutorial 3: Intermediate (May 24)**
 - Advanced RNA-Seq analysis topics and troubleshooting
- Hands-on tutorials (Summer 2012)...



RNA-Seq Module 3

Advanced RNA-Seq Analysis Topics and Trouble-Shooting

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

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

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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Part I

Review and Considerations for Different Goals and Biological Systems

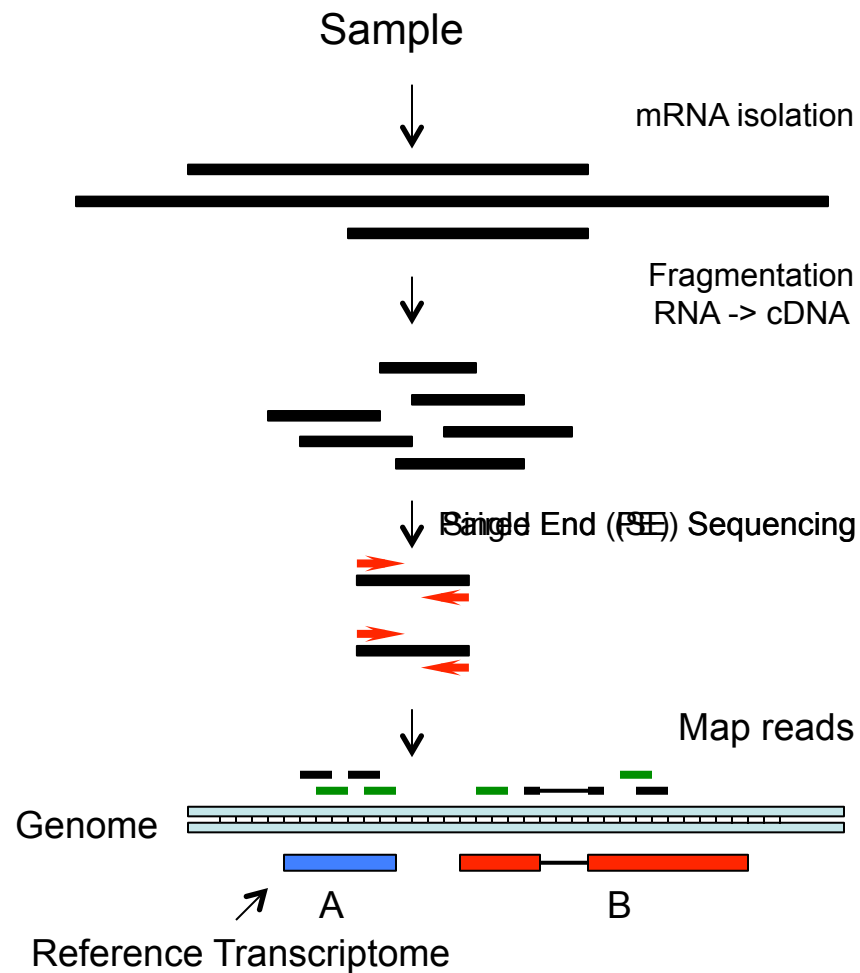
Kevin Silverstein, PhD



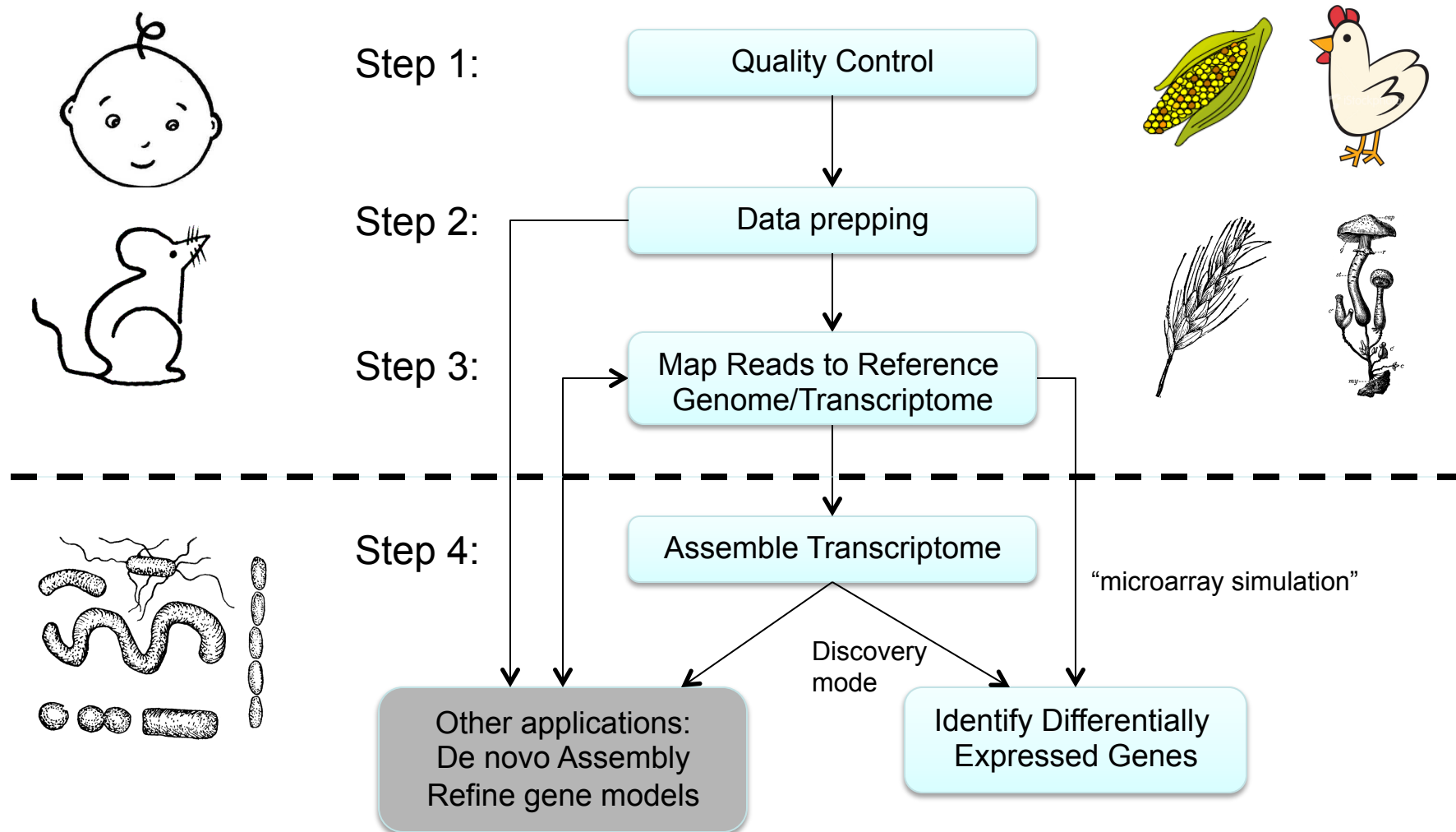
UNIVERSITY OF MINNESOTA

Driven to DiscoverSM

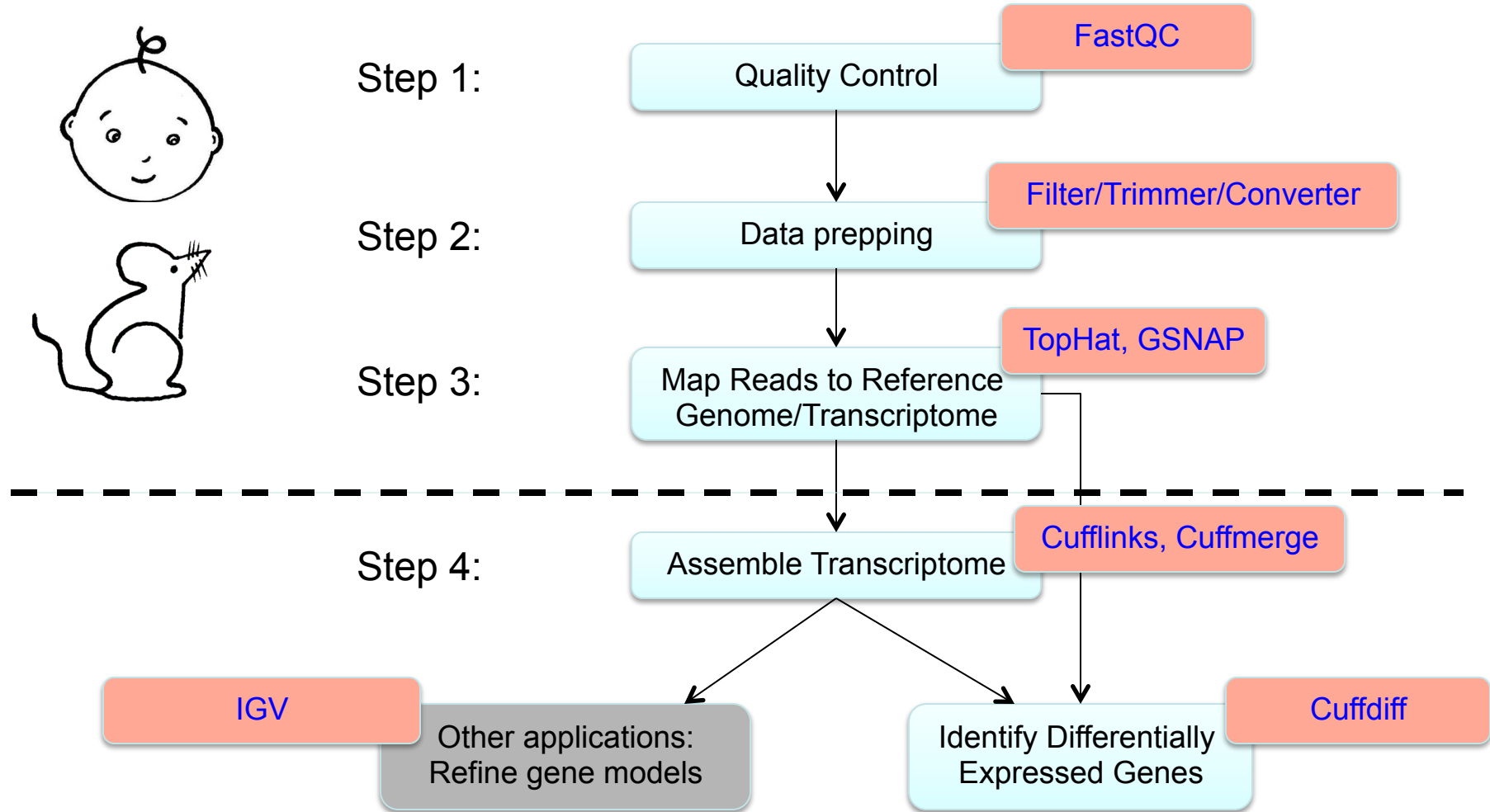
Typical RNA-seq experimental protocol and analysis



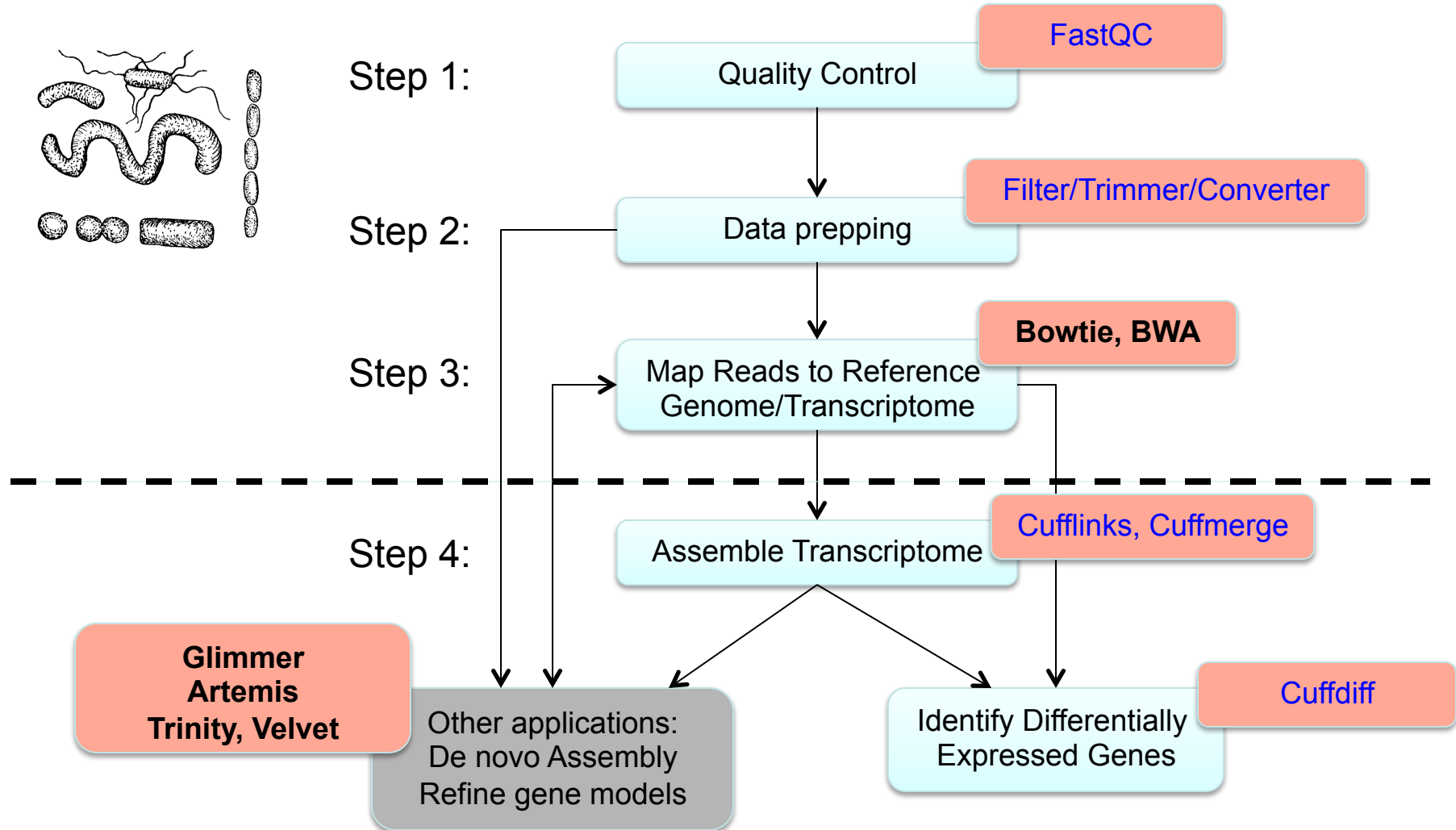
Steps in RNA-Seq data analysis depend on your goals and biological system



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



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



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.



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

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



Step 1:

Quality Control

FastQC

Step 2:

Data prepping

Filter/Trimmer/Converter

Step 3:

Map Reads to Reference
Genome/Transcriptome

TopHat, GSNAP

Step 4:

Assemble Transcriptome

Cufflinks, Cuffmerge

GeneMark, FGeneSH
Trinity, TransABYSS
BLAT

Other applications:
De novo Assembly
Refine gene models

Identify Differentially
Expressed Genes

Cuffdiff



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

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



Step 1:

Quality Control

FastQC

Step 2:

Data prepping

Filter/Trimmer/Converter

Step 3:

Step 4:

Assemble Transcriptome

Cufflinks, Cuffmerge

Trinity, TransABYSS

Other applications:
De novo Assembly
Refine gene models

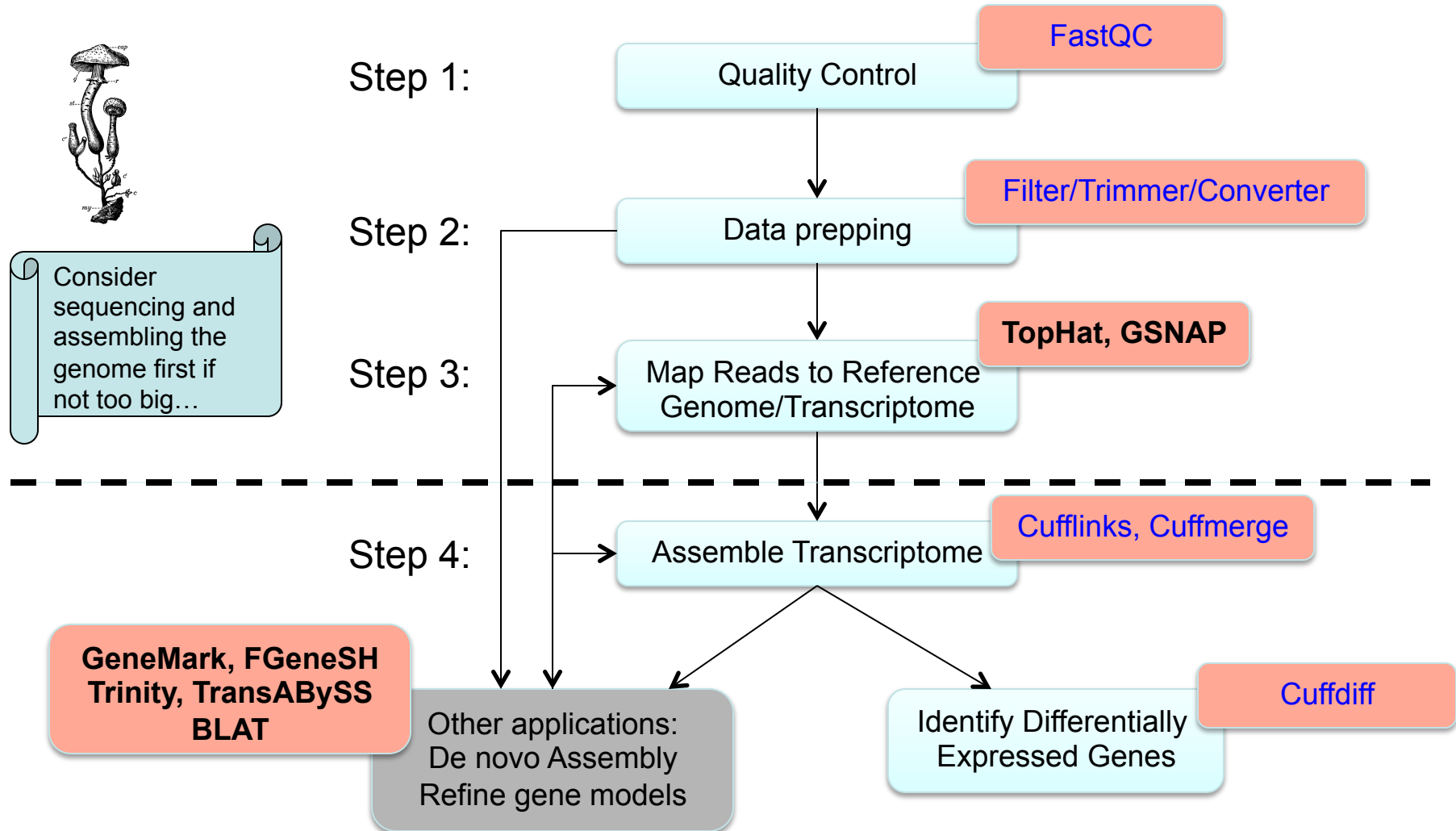
Identify Differentially
Expressed Genes

Cuffdiff



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

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



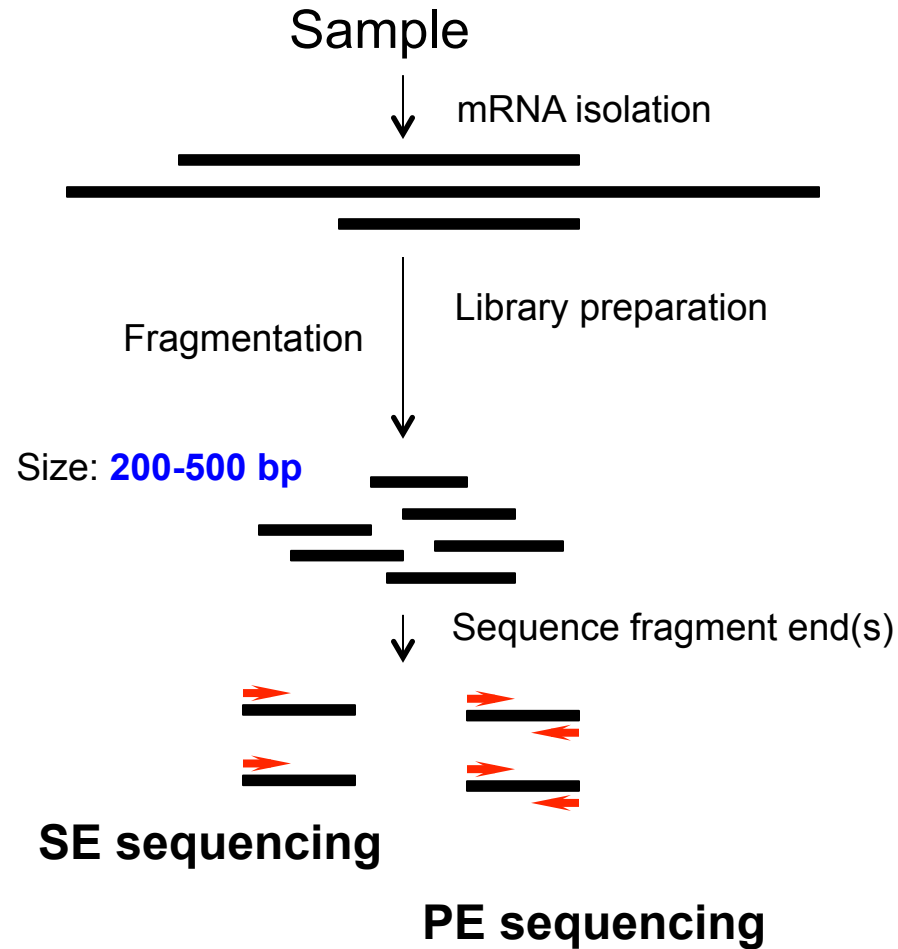
Library construction and sequencing design decisions



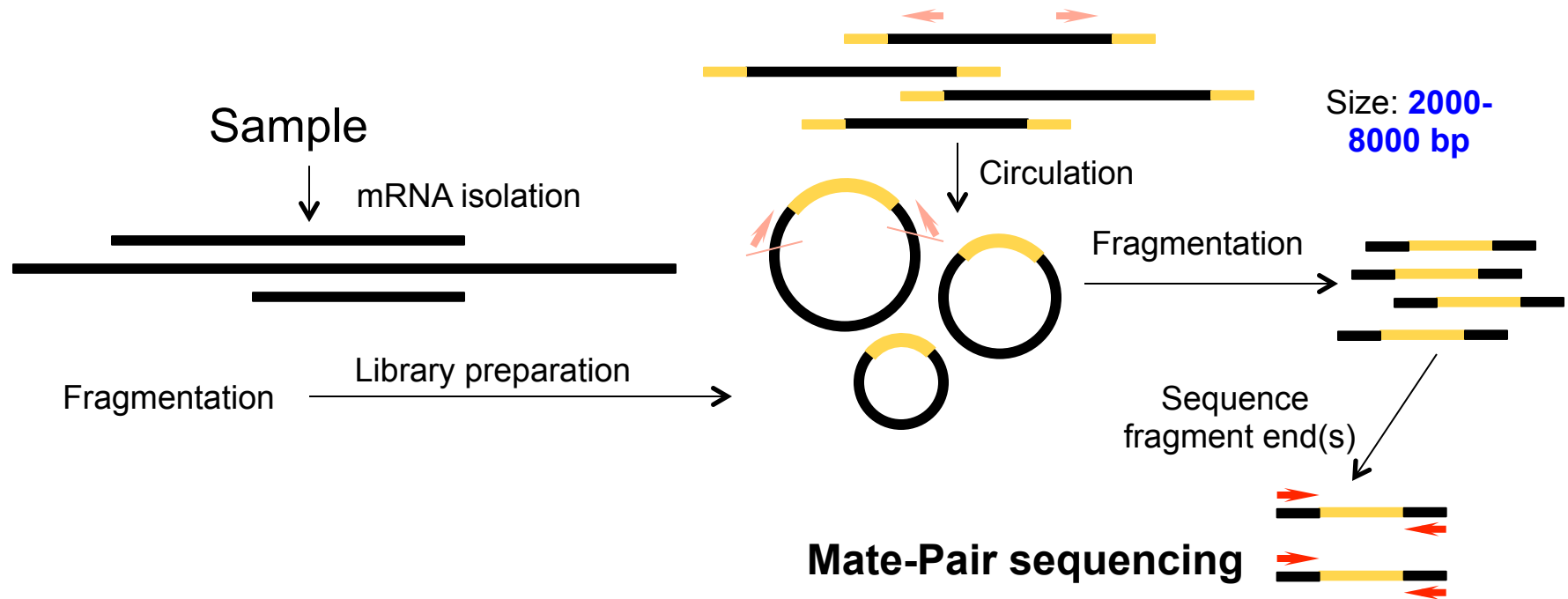
UNIVERSITY OF MINNESOTA

Driven to DiscoverSM

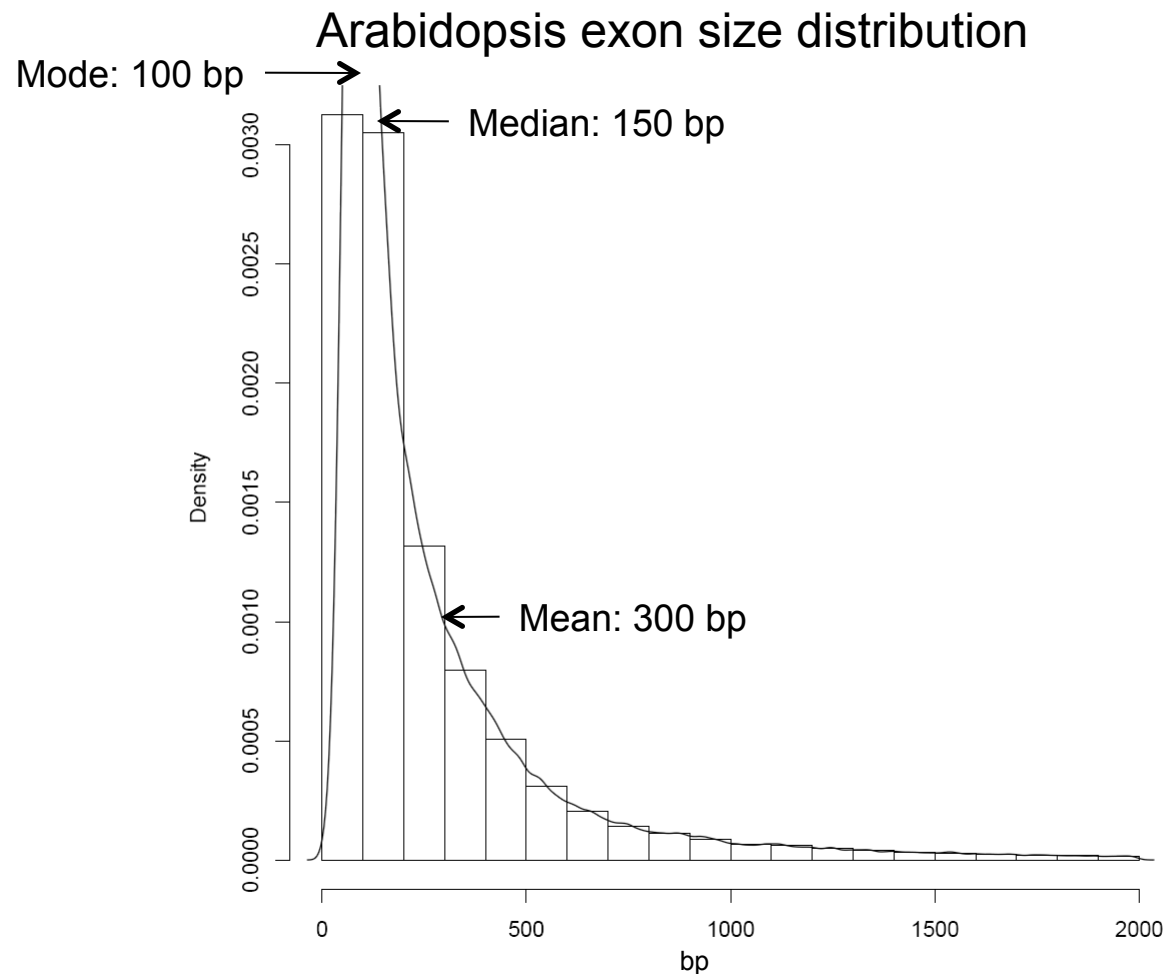
Library type (SE/PE) and insert size



Library type (Mate-pair) and insert size

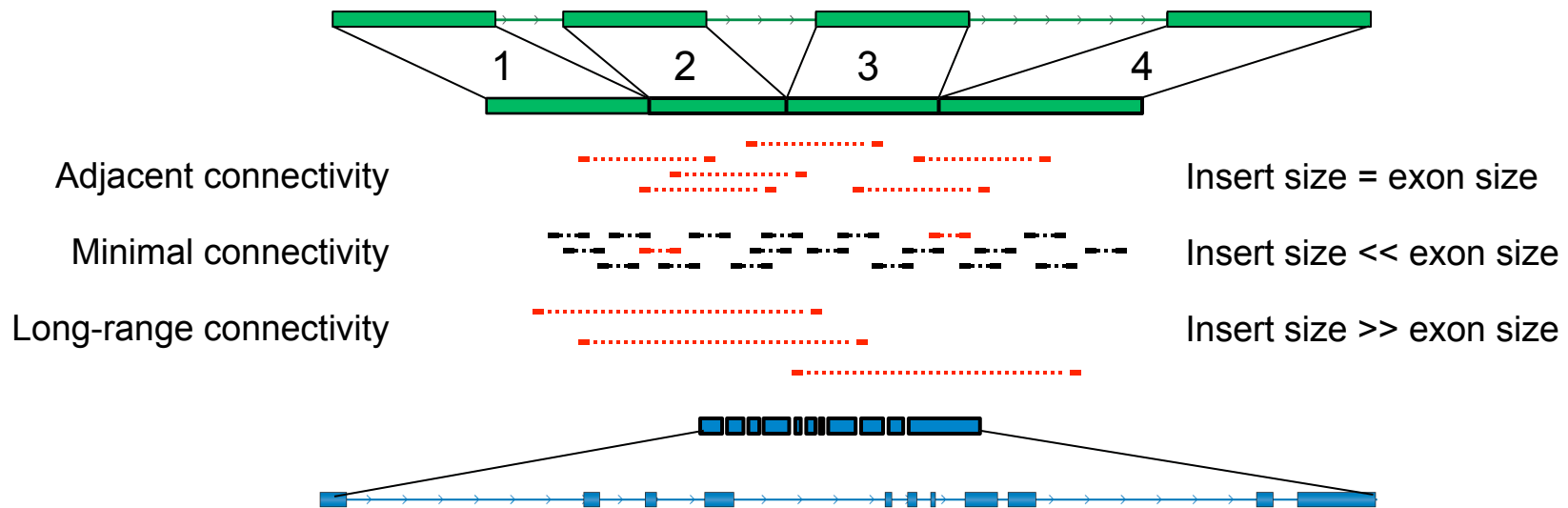


Optimal library size depends on goals and organism: **exon size**



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

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



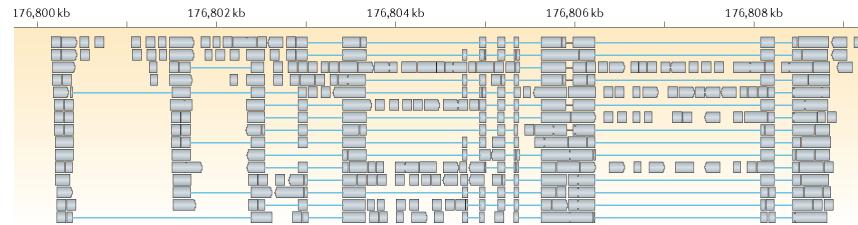
One size doesn't fit all: organisms can differ in exon size distribution



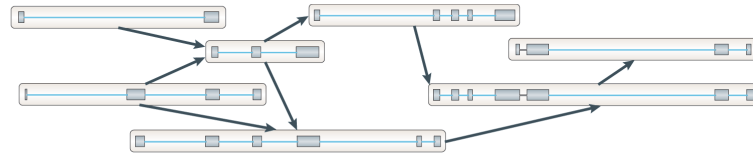
UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

How does connectivity play into the analysis?

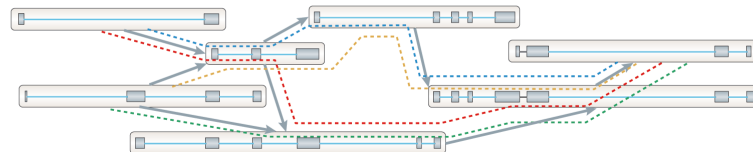
1. splice-align reads to the genome



2. Build a graph representing alternative splicing events



3. Traverse the graph to assemble variants



4. Assemble isoforms



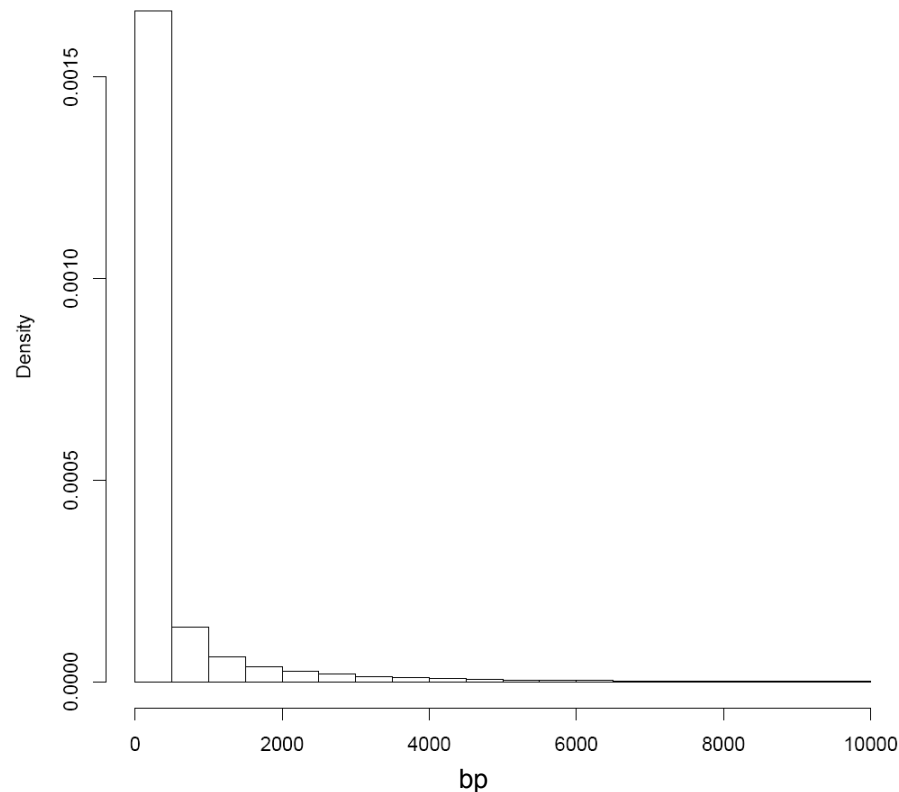
Martin JA and Wang Z. Nat Rev Genet. (2011) 12:671–682.



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

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)

Arabidopsis intron size distribution

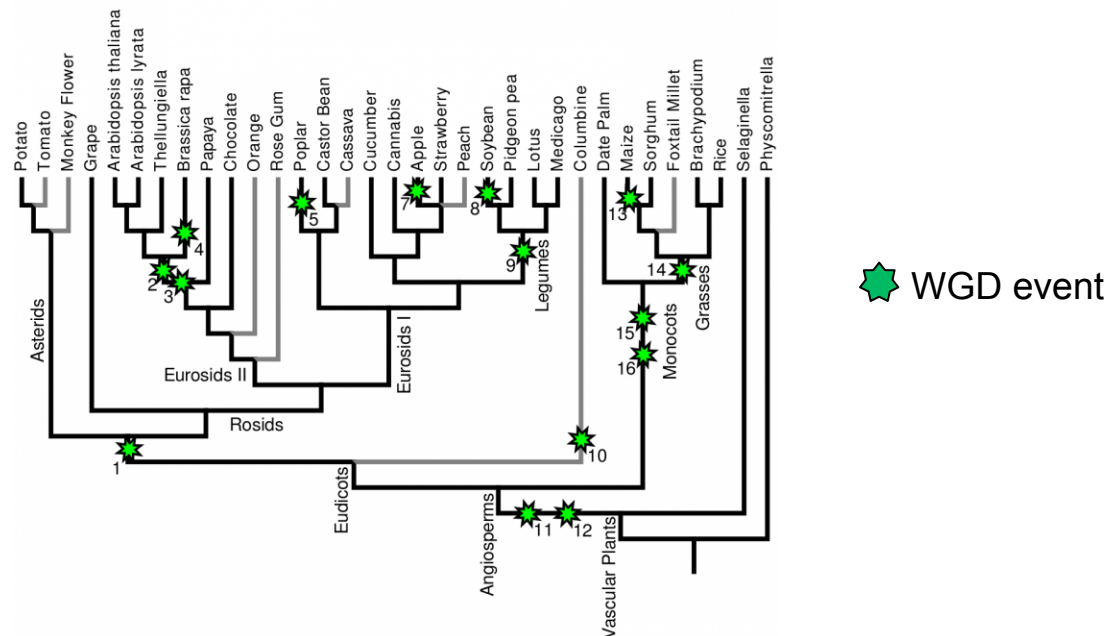


Why not just leave the defaults? (e.g., 70-500,000 bp)

- ~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)



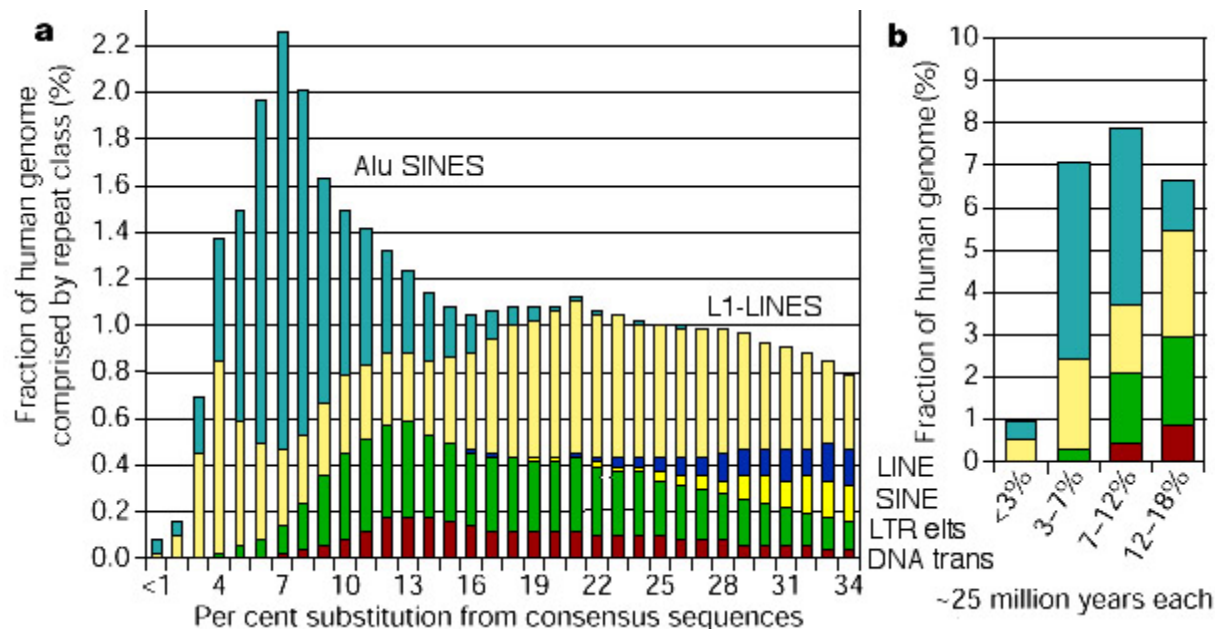
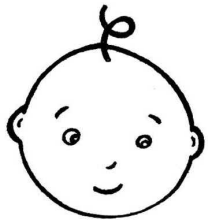
<http://genomeevolution.org>

- 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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Some genomes are rife with repetitive elements



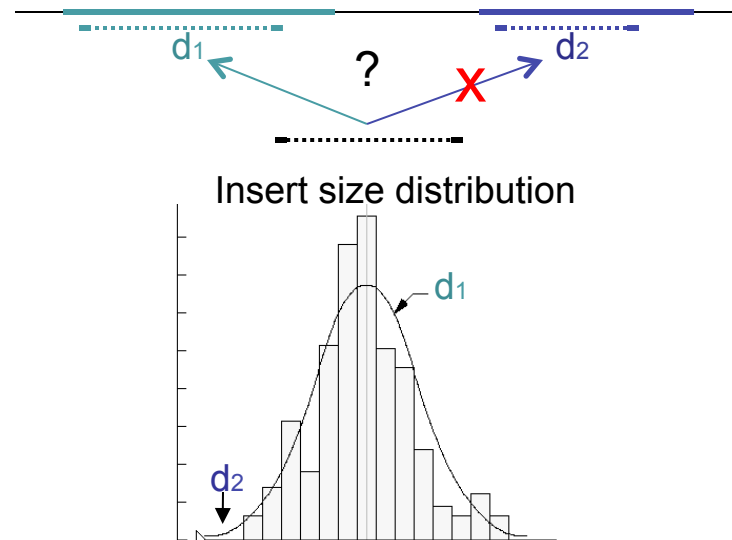
<http://genomeevolution.org>

- 50%, 65% of the human and maize genome are repeat elements, respectively (rebase, 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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Why is PE 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.



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Sequencing depth needed depends on transcriptome size and the project goals

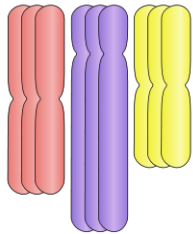
- **Sequencing Depth** is the average read coverage of target sequences
 - Sequencing depth = total number of reads X read length / estimated target sequence length
 - Example, for a 5MB transcriptome, if 1 Million 50 bp reads are produced, the depth is $1\text{ M} \times 50\text{ bp} / 5\text{M} \sim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



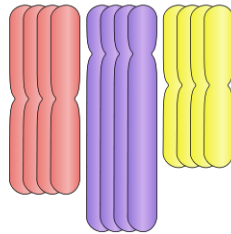
Polyploidy is particularly problematic



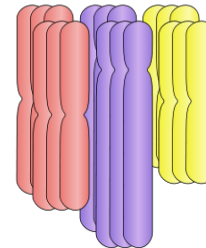
Triploid (3N)



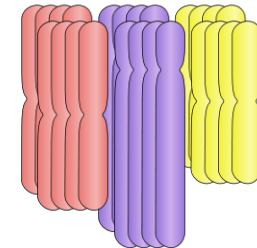
Tetraploid (4N)



Hexaploid (6N)



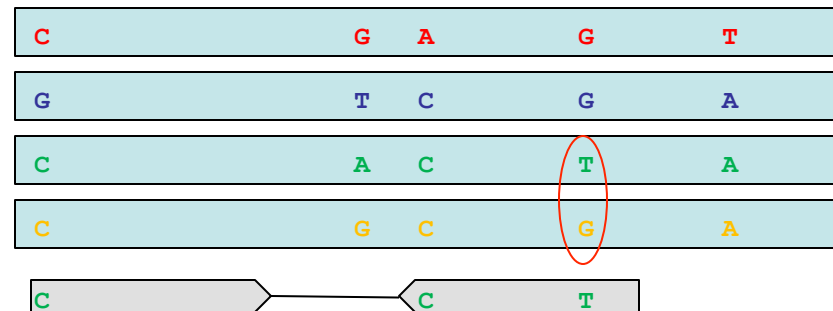
Octaploid (8N)



- 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 imprinting).



Genome size characteristics (iGenomes)

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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Summary of Library Construction and Sequencing Decisions

| | 1 | 2 | 3 | 4 |
|-------------------|--|-------------------|------------------------------|---------------------------------------|
| Project Goals: | <i>De novo</i> 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. bacteria, yeast), or if you're interested in antisense regulation.

Sample Replicates and Pooling Decisions

| | 1 | 2 | 3 | 4 |
|------------------------|--|---------------------|------------------------------|---------------------------------------|
| Project Goals | <i>De novo</i> 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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Part II

Read Mapping Statistics and Visualization

John Garbe, PhD



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Mapping Statistics

How well did my sequence library align to my reference?



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

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>



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Mapping Statistics

- SAM Tools
- Picard
- Tophatstats



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 – Picard

- Galaxy
 - NGS: Picard (beta) -> SAM/BAM Alignment Summary Metrics
- Command line:
 - module load picard-tools
 - java -Xmx2g -jar
CollectAlignmentSummaryMetrics.jar
INPUT=accepted_hits.bam OUTPUT=stats.txt



Mapping Statistics – Picard

- Picard output

| CATEGORY | TOTAL_READS |
|----------------|-------------|
| FIRST_OF_PAIR | 14739626 |
| SECOND_OF_PAIR | 14653925 |
| PAIR | 29393551 |



Mapping Statistics – tophatstats

- Galaxy
 - MSI -> tophatstats
- Command line
 - module load tophatstats



Mapping Statistics – tophatstats

- Tophatstats output (paired-end reads)

```
% tophatstats.pl accepted_hits.bam L1_R1_sample1.fastq
Input files: accepted_hits.bam      L1_R1_sample1.fastq
250000 total read pairs in fastq file
120004 (48.00%) read pairs mapped with correct insert size
           (116869 with unique alignments)
50536 (20.21%) read pairs mapped with wrong insert size
           (49351 with unique alignments)
24368 (9.75%) read pairs with only one read in the pair mapped
           (23544 with unique alignments)
55092 (22.04%) read pairs with no mapping
60.13bp average inner distance between read pairs
```



Mapping Visualization

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

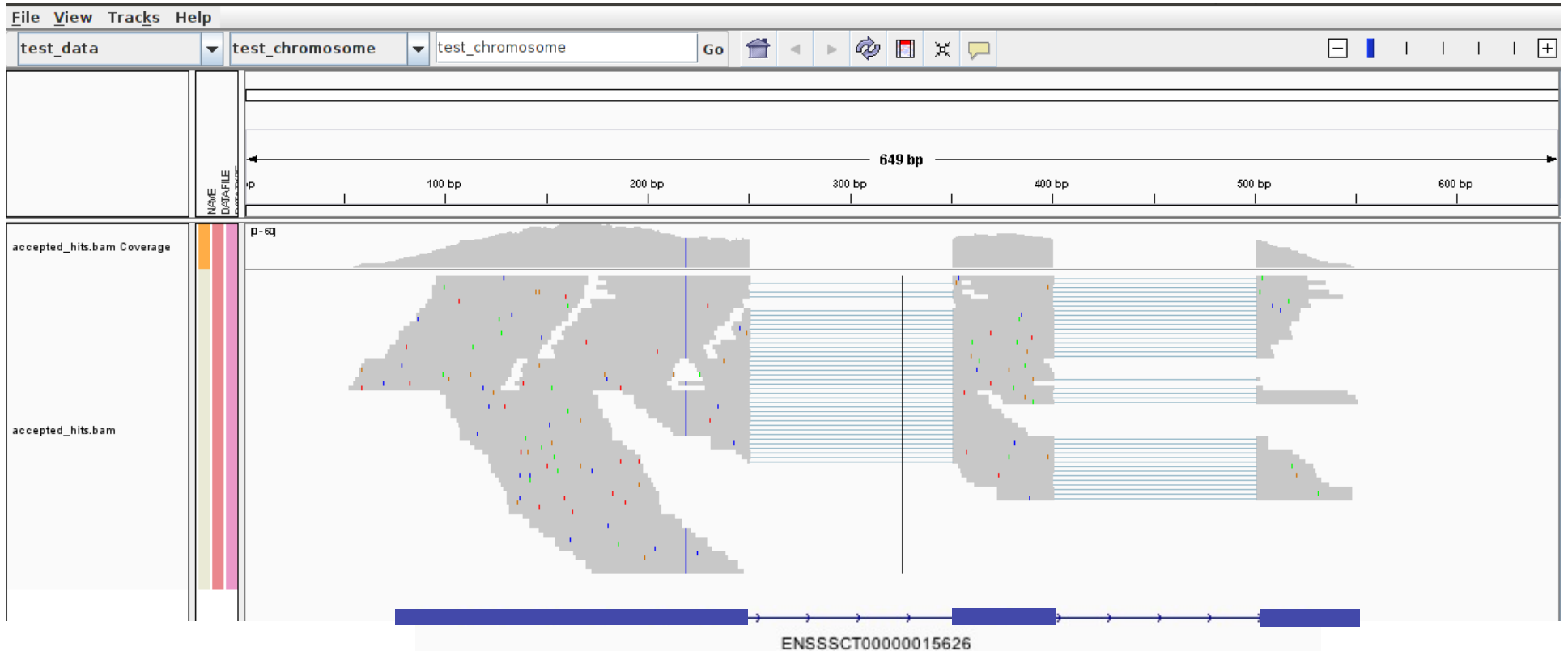


<http://www.broadinstitute.org/igv/home>



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Mapping Visualization



Bam file viewed with IGV



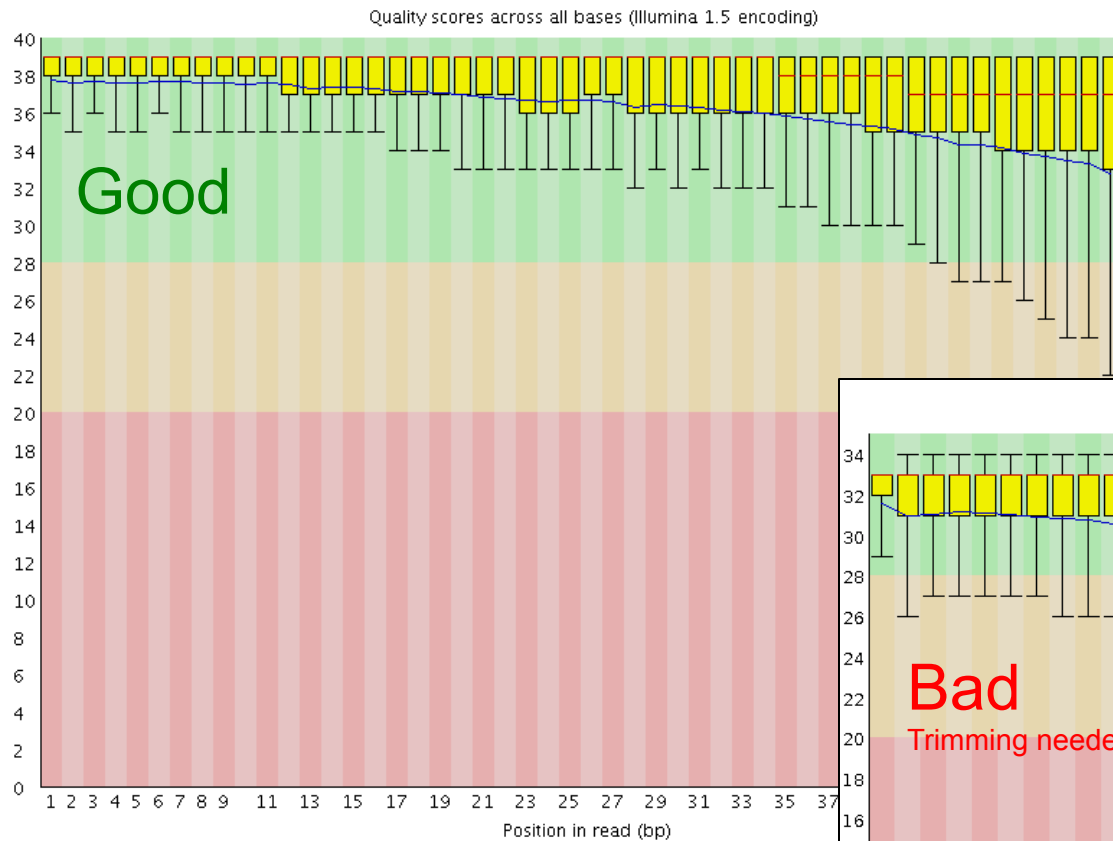
UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Causes of poor mapping

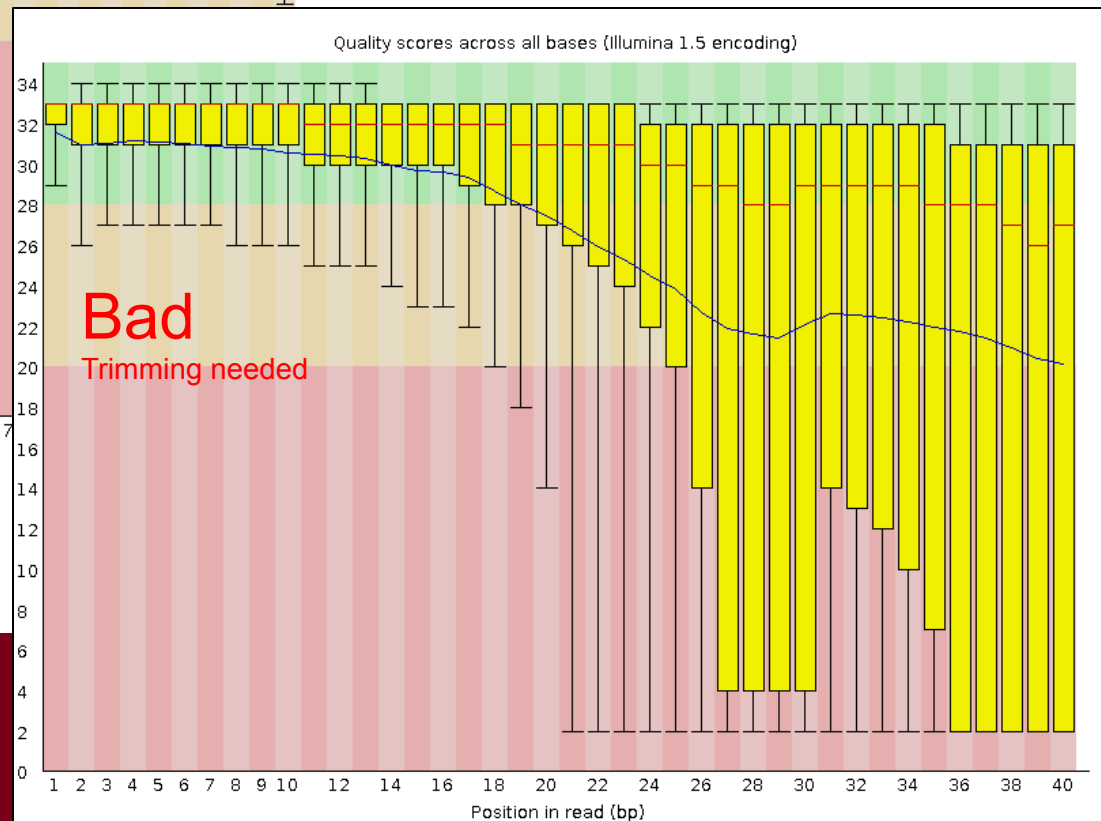
- Improper alignment parameters
- Poor quality sequence library
- Contaminated sequence library
- Poor quality reference
- Repetitive genome
- Divergence between sequenced population and reference
- Mislabeled samples
- Corrupted files
- Short read length
- Poor choice of mapping software
- Bug in mapping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



Poor Quality Library



Poor quality read library
decreases mapping performance



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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Poor Quality Reference

Sus scrofa 9.2

46%

17%

9%

26%

Sus scrofa 10.2

48%

20%

9%

22%

mapped, properly paired

mapped, wrong insert size

singleton

no mapping

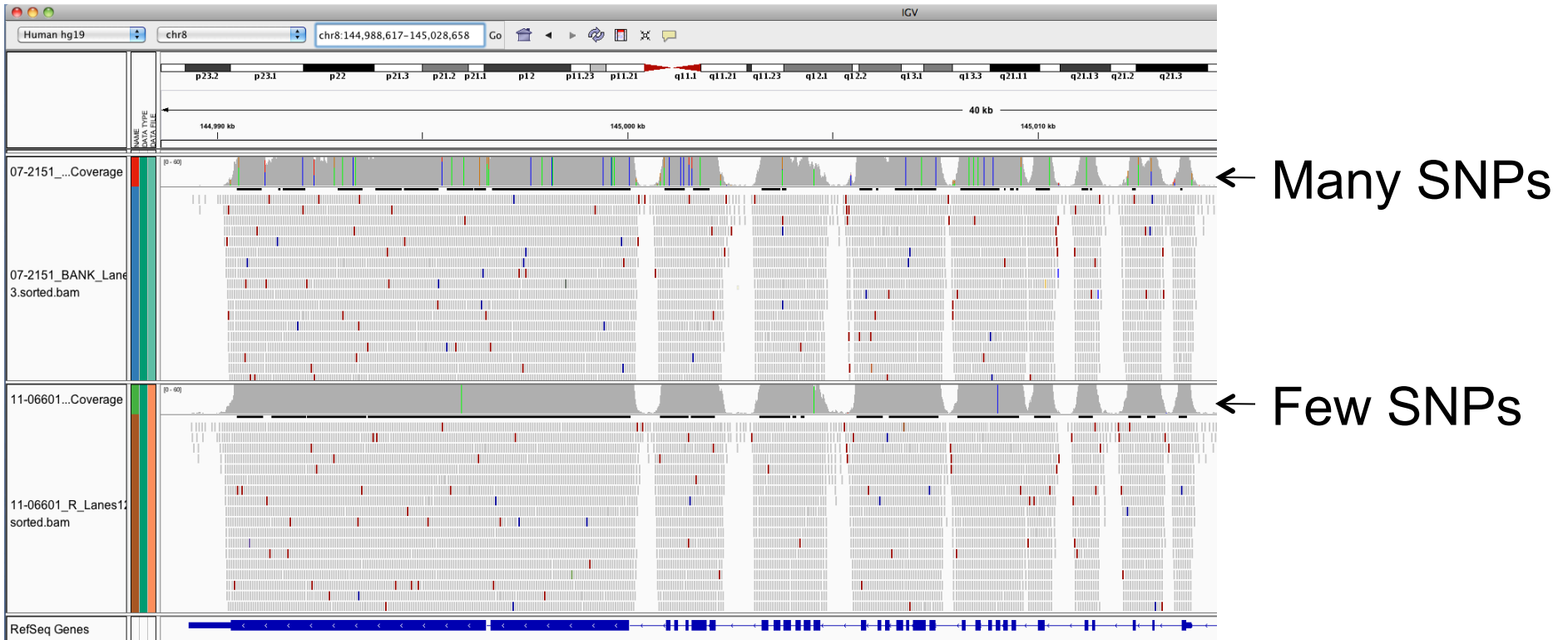
Mapping performance improves due
to improvement in Pig genome build



UNIVERSITY OF MINNESOTA

Driven to DiscoverSM

Divergence between sequenced population and reference



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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Contaminated sequence library

Overrepresented sequences

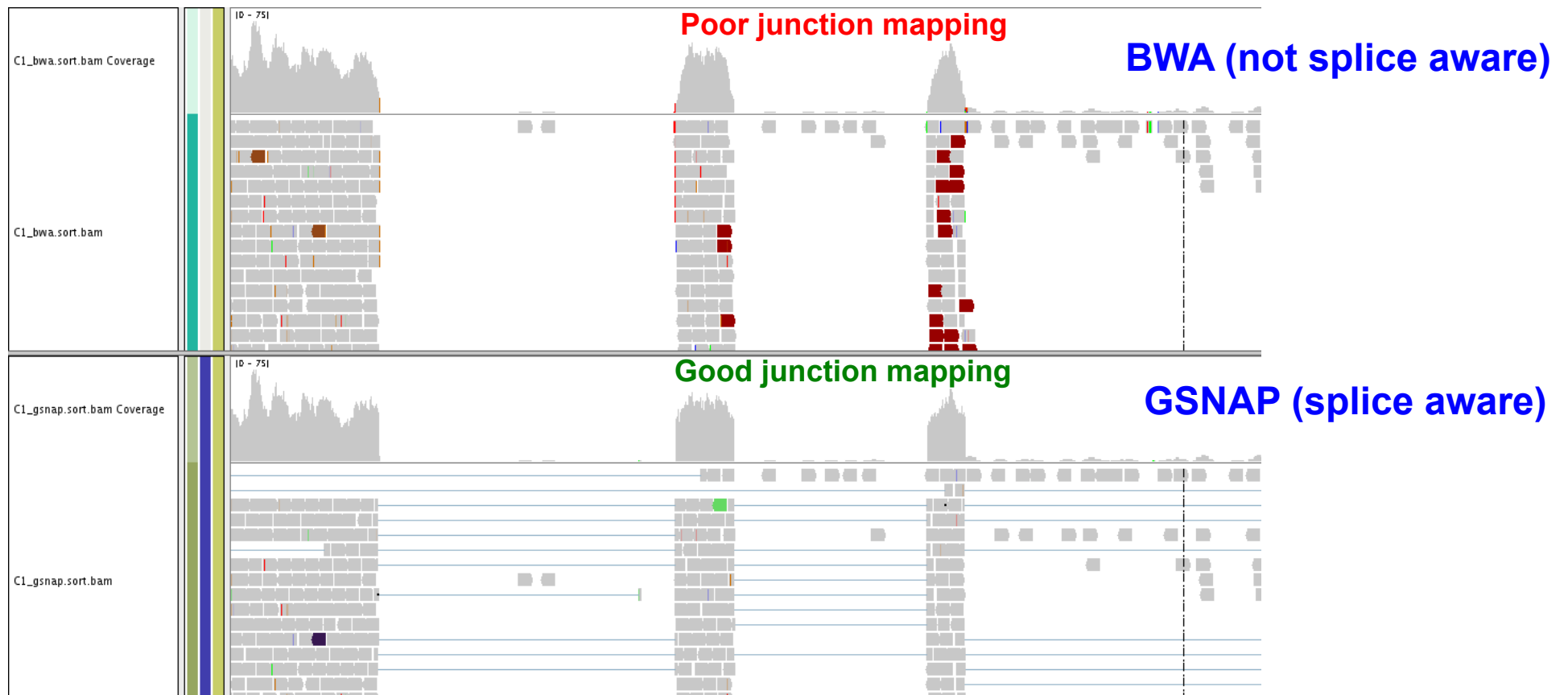
| Sequence | Count | Percentage | Possible Source |
|---|--------|---------------------|---|
| GTATTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCG | 820428 | 2.8366639370528275 | Illumina Paired End PCR Primer 2 (100% over 43bp) |
| GTATACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT | 749728 | 2.5922157461699773 | 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) |
| GTATTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGT | 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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Poor choice of mapping software



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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Corrupted files

| 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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Part III

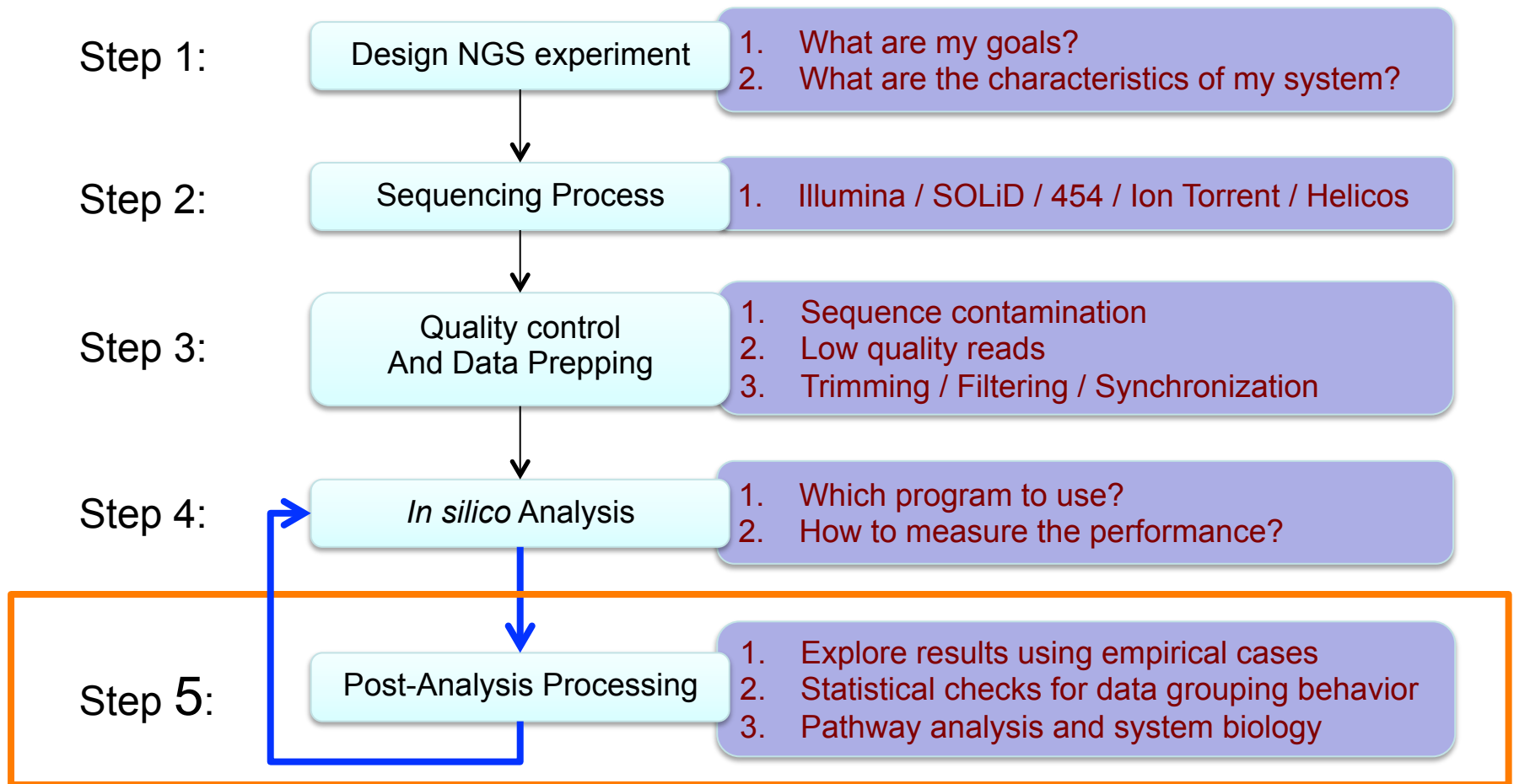
Post-Analysis Processing - Exploring the Data and Results

Ying Zhang, PhD



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Workflow of a typical NGS project




Widely-used Tools in Data Exploring

- 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 



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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

UCSC Genome Browser

(<http://genome.ucsc.edu/cgi-bin/hgGateway>)

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

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.

track search add custom tracks track hubs configure tracks and display clear position

Home Genomes **Genome Browser** Blat Tables Gene Sorter PCR Session FAQ Help

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: Browse... Submit

Clear

Optional track documentation: Or upload: Browse...

Clear

Click [here](#) for an HTML document template that may be used for Genome Browser track descriptions.

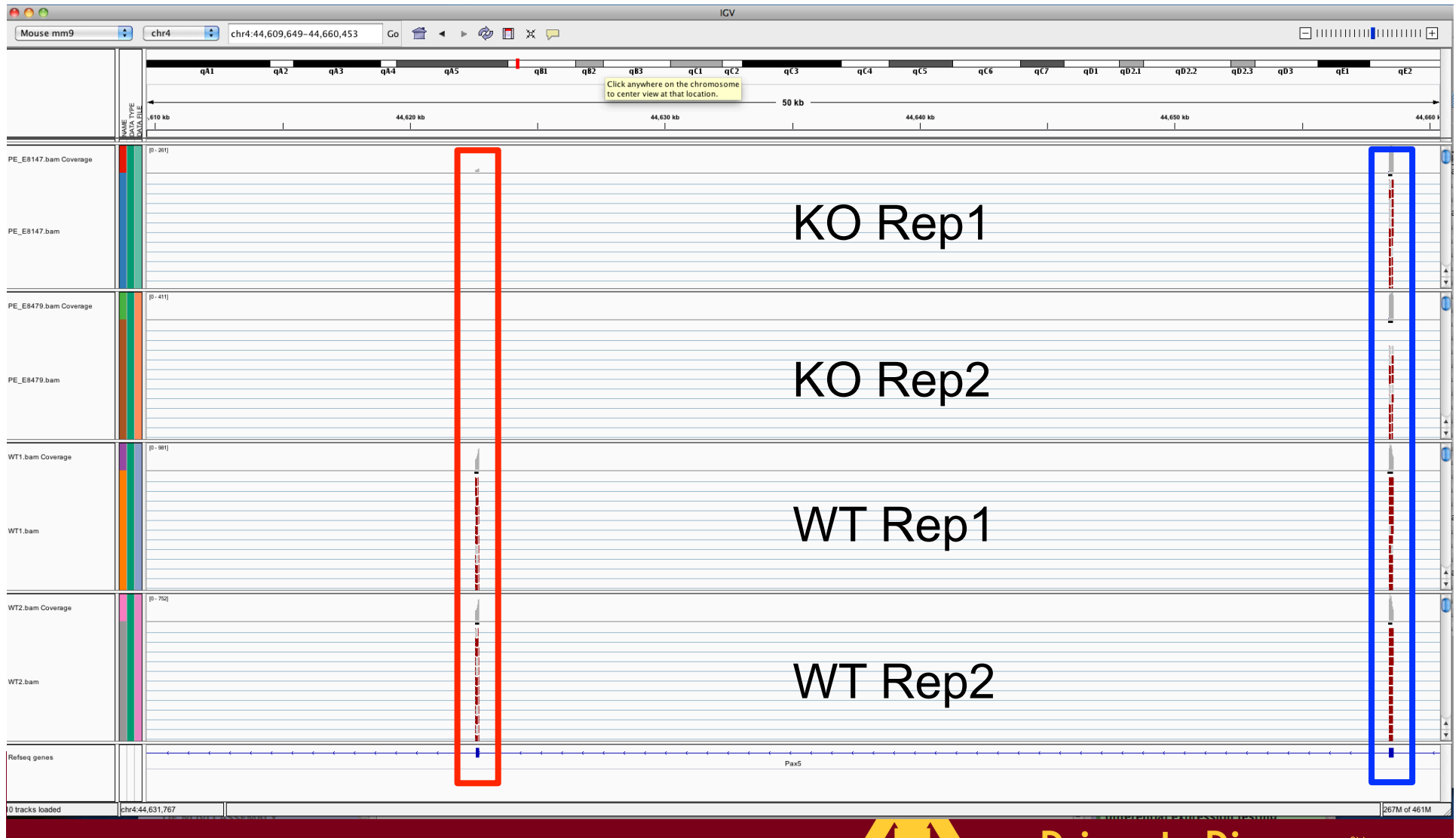
No. 1 in your Check-List

“Are my data behaving as expected?”



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Exploring results using Empirical Cases – Example I: no reads mapped at knock-out site

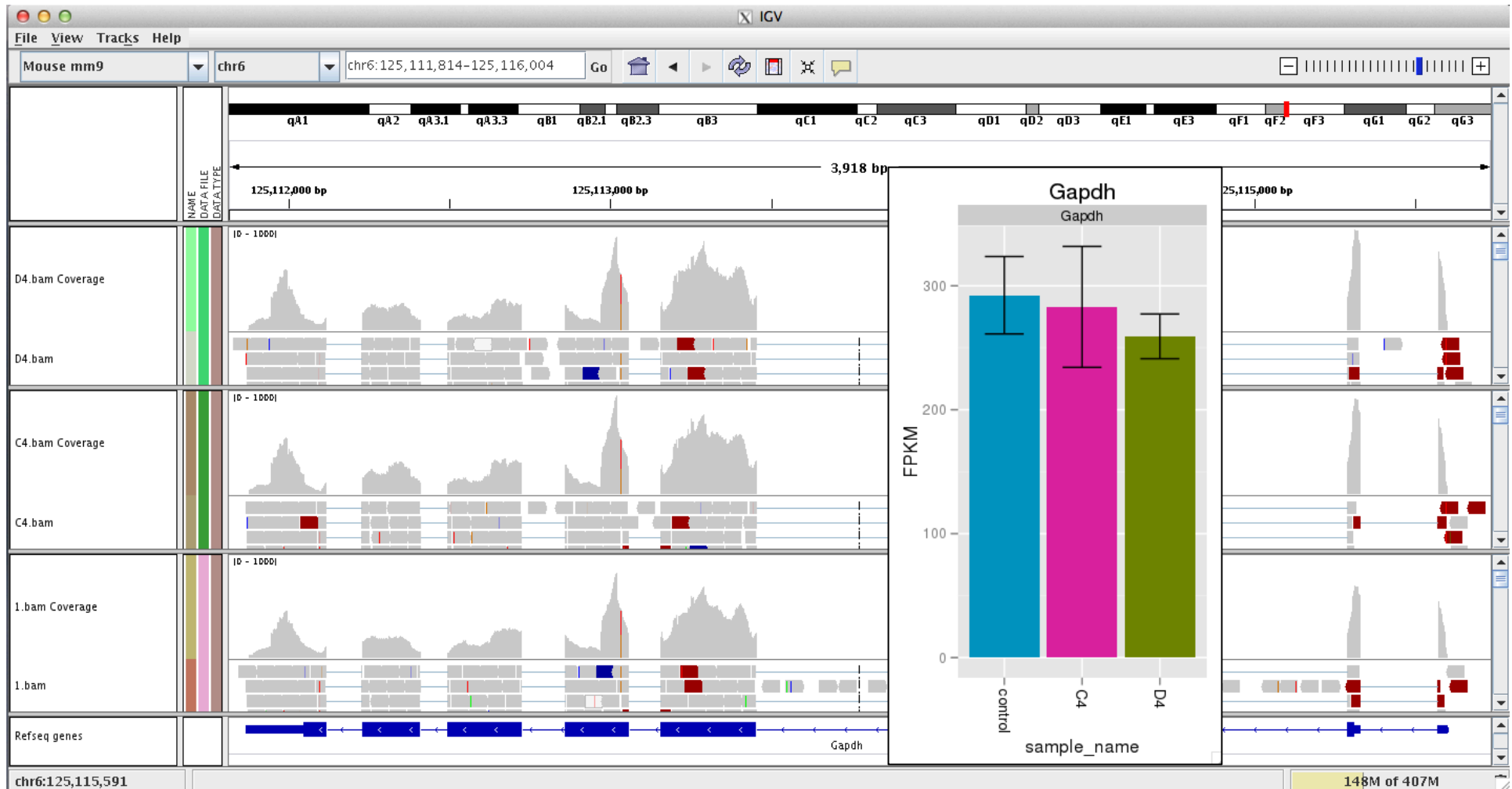


Data Courtesy of Dr. Mike Farrar and Dr. Lynn Harris (unpublished data)



Driven to DiscoverSM

Example II: Housekeeping genes should behave similarity across multiple samples



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



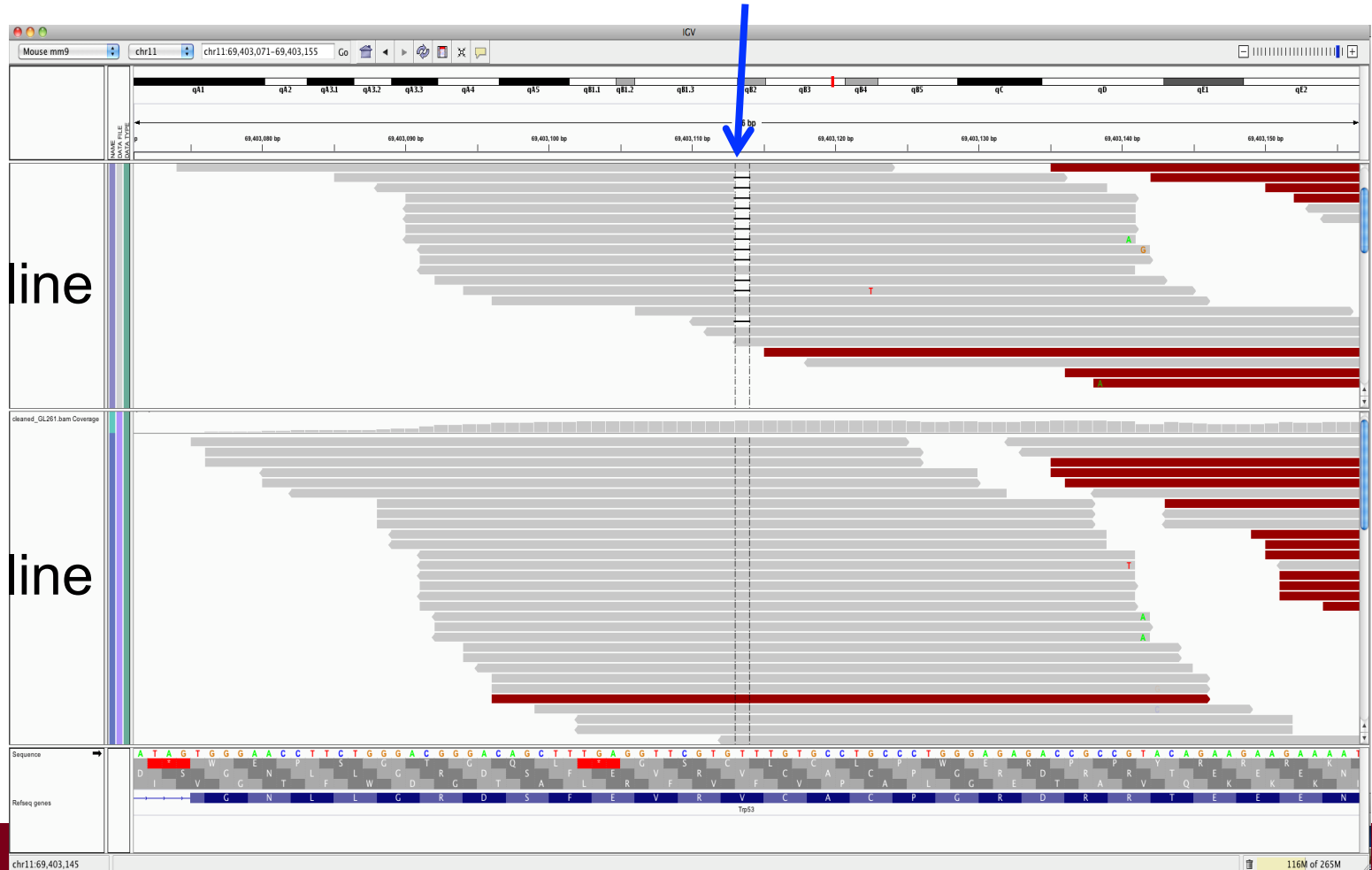
UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Example III: review of known biomarkers, for example, known SNP and indel

Heterozygous deletion of 'T' with 46% penetrance

Cancer cell line

Control cell line



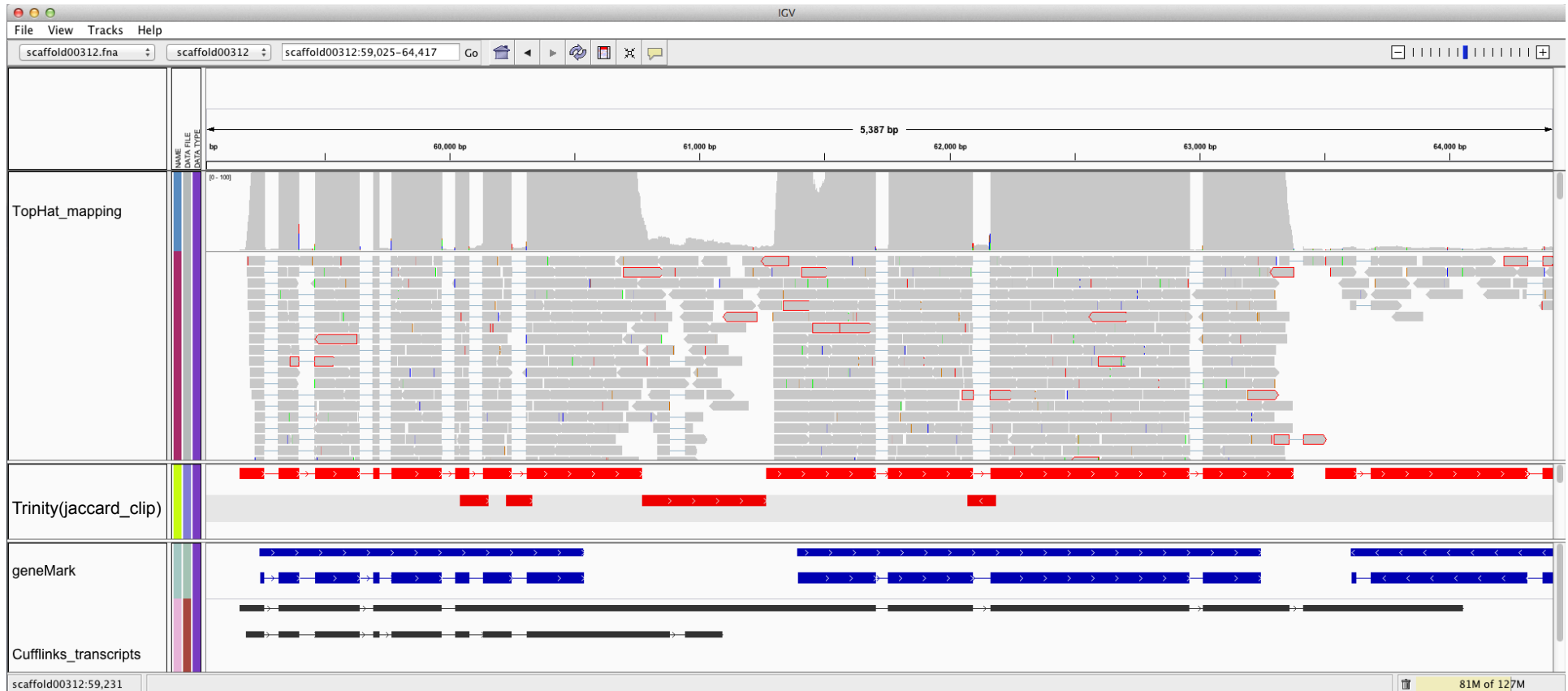
Data Courtesy of Dr. John Ohlfest and Dr. Flavia Popescu (unpublished data)



UNIVERSITY OF MINNESOTA

Driven to DiscoverSM

Example IV: detect the caveat of programs




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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

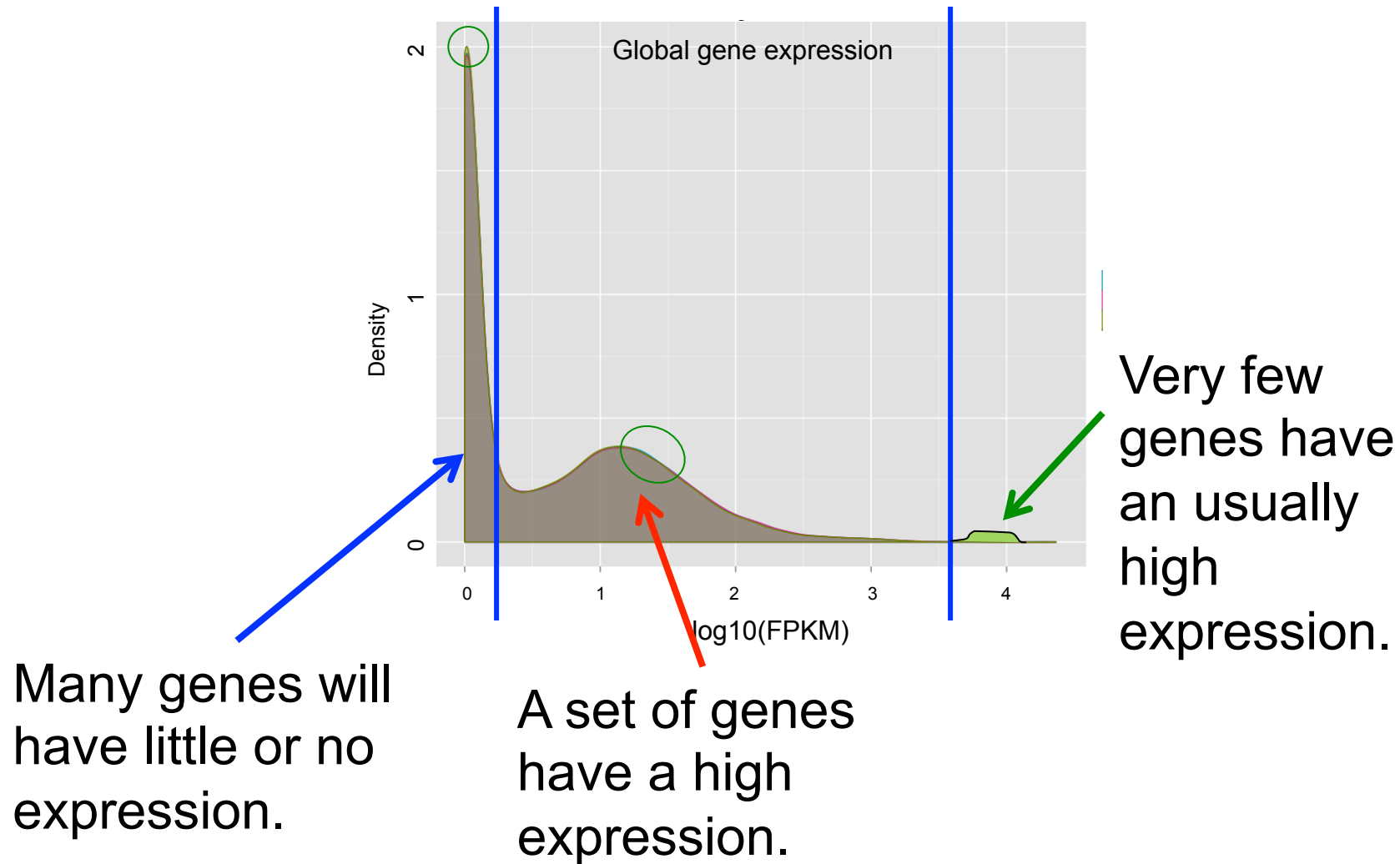
Specific Notes for Prokaryotes' samples

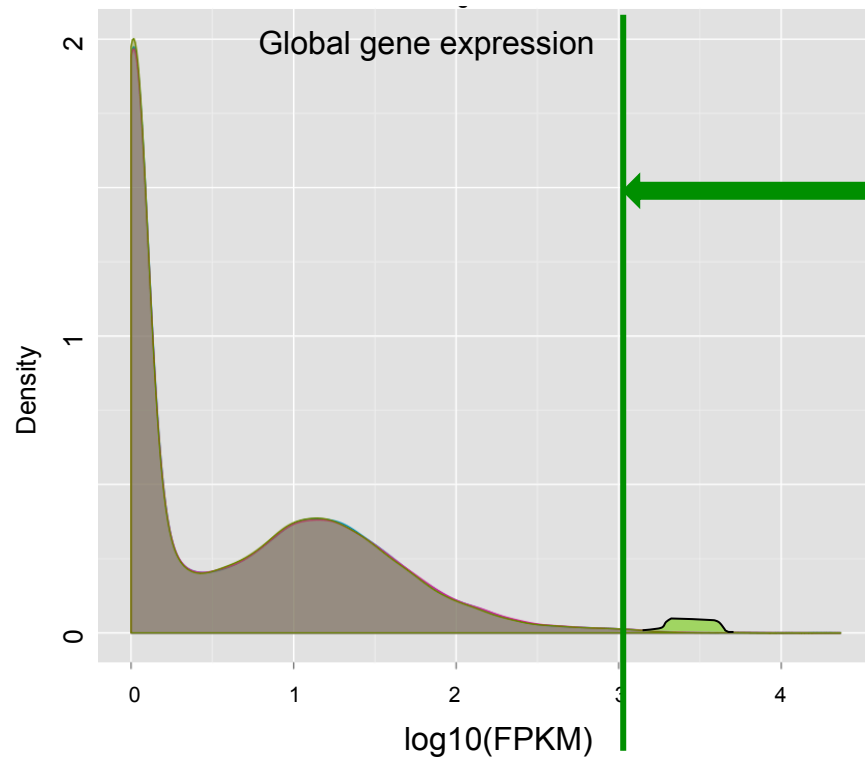
- Cufflinks developer: 

“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.”
- So for bacteria transcriptome:
 - If the genome is available, do genome annotation first then reconstruct the transcriptome.
 - If the genome is not available, try *de novo* assembly of the transcriptome, followed by gene annotation.



Explore the global distribution of data





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



Perform quartile normalization:

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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Warning: don't throw the baby with the bathwater...



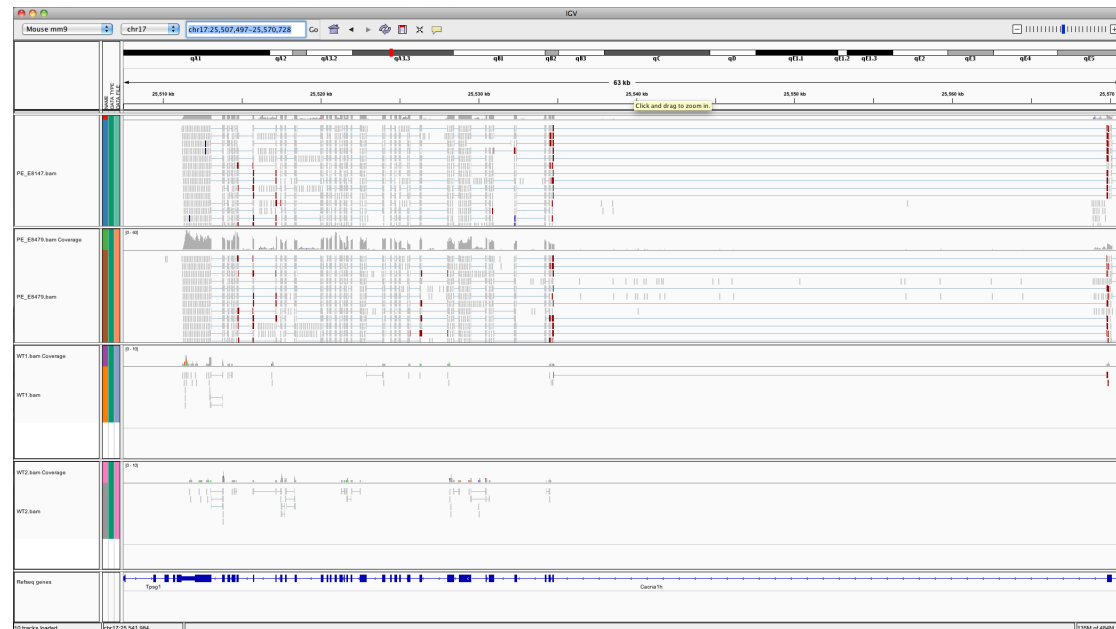
Cuffdiff: “Min Alignment Count” must be satisfied in **all** samples – too high a value will remove genes not expressed in one condition but strongly expressed in another!

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.

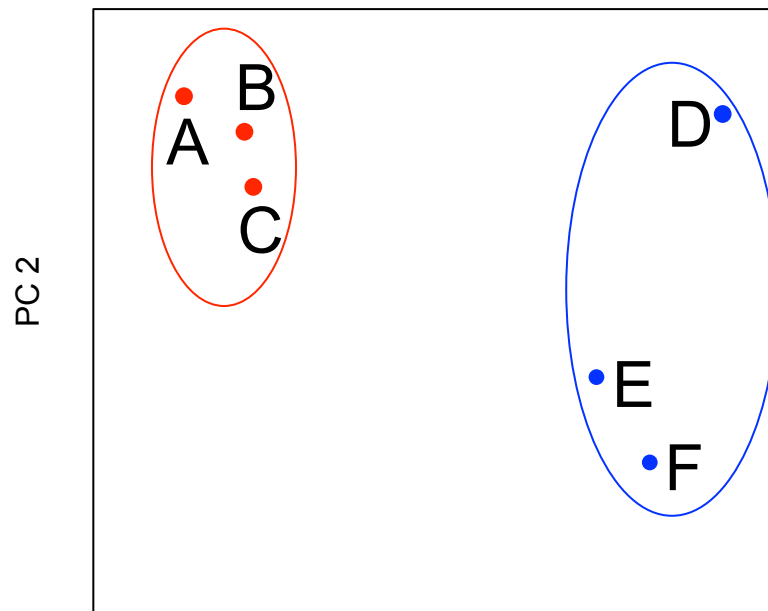


UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Statistical Checks of data structure – Multi-Variable Analysis

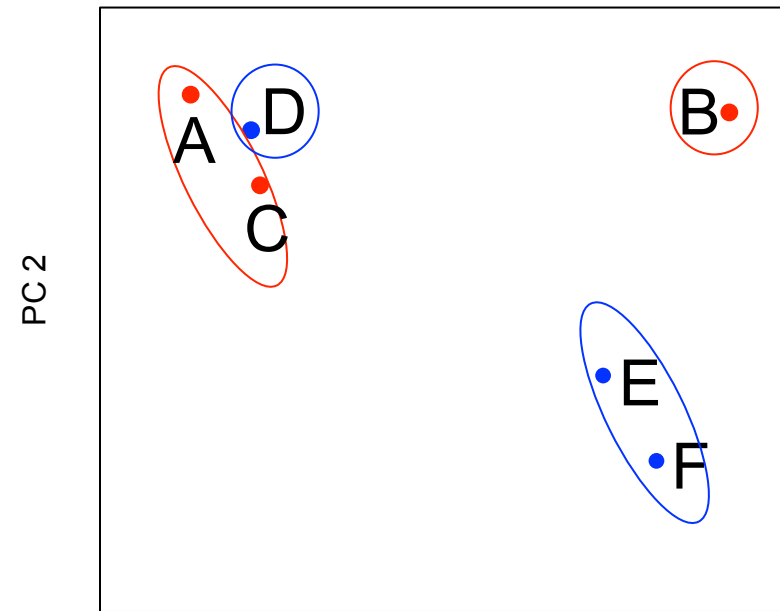
- Biological replicates should show grouping behavior in multi-variable analysis:
 - innate consistence between samples

A hypothetical PCA plot



PC 1

A hypothetical PCA plot



PC 1



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

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 technical replicates 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



Steps in PCA analysis

1. Construct the multiple variable matrix



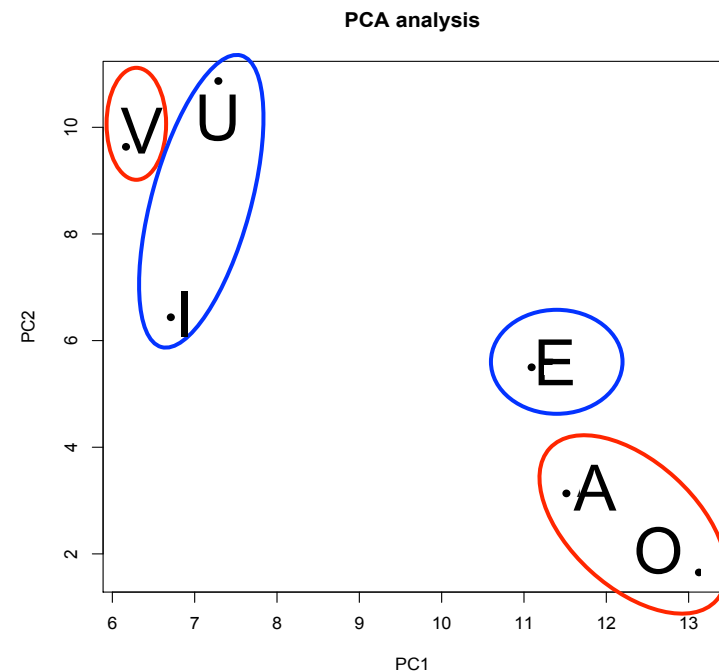
2. Run PCA analysis and explore the result

e.g. tables of FPKM values

| transcript | 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)



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM

Heatmap: Unsupervised clustering

1. Construct the multiple variable matrix

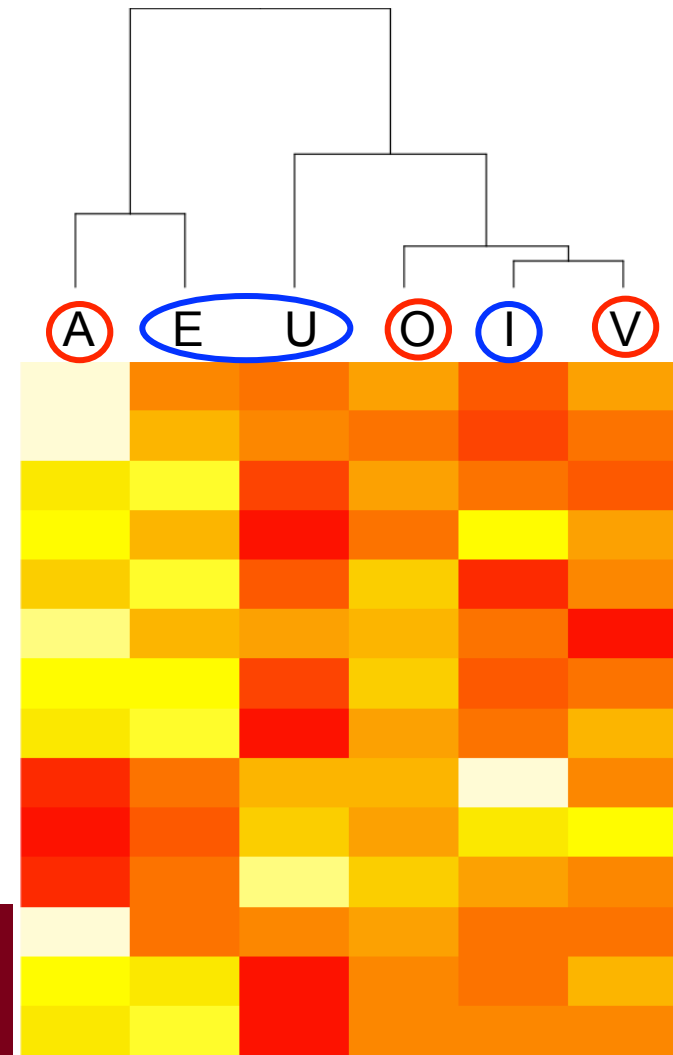
2. Run Unsupervised Clustering and generate Heatmap

e.g. tables of FPKM values

| transcript | 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)

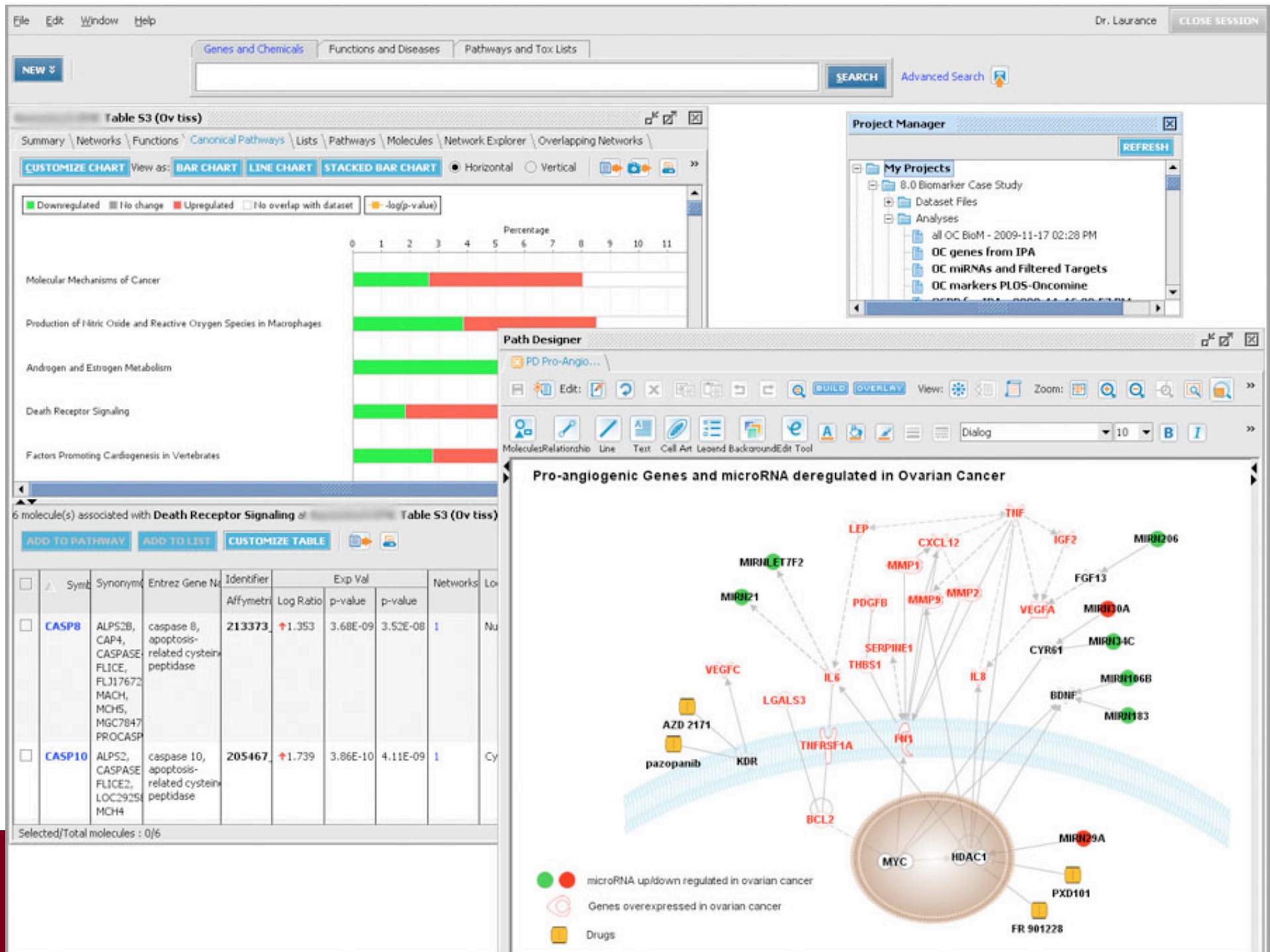


Exploring data at system-level: Ingenuity Pathway analysis

- 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



UNIVERSITY OF MINNESOTA
Driven to DiscoverSM



Discussion and Questions?

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

