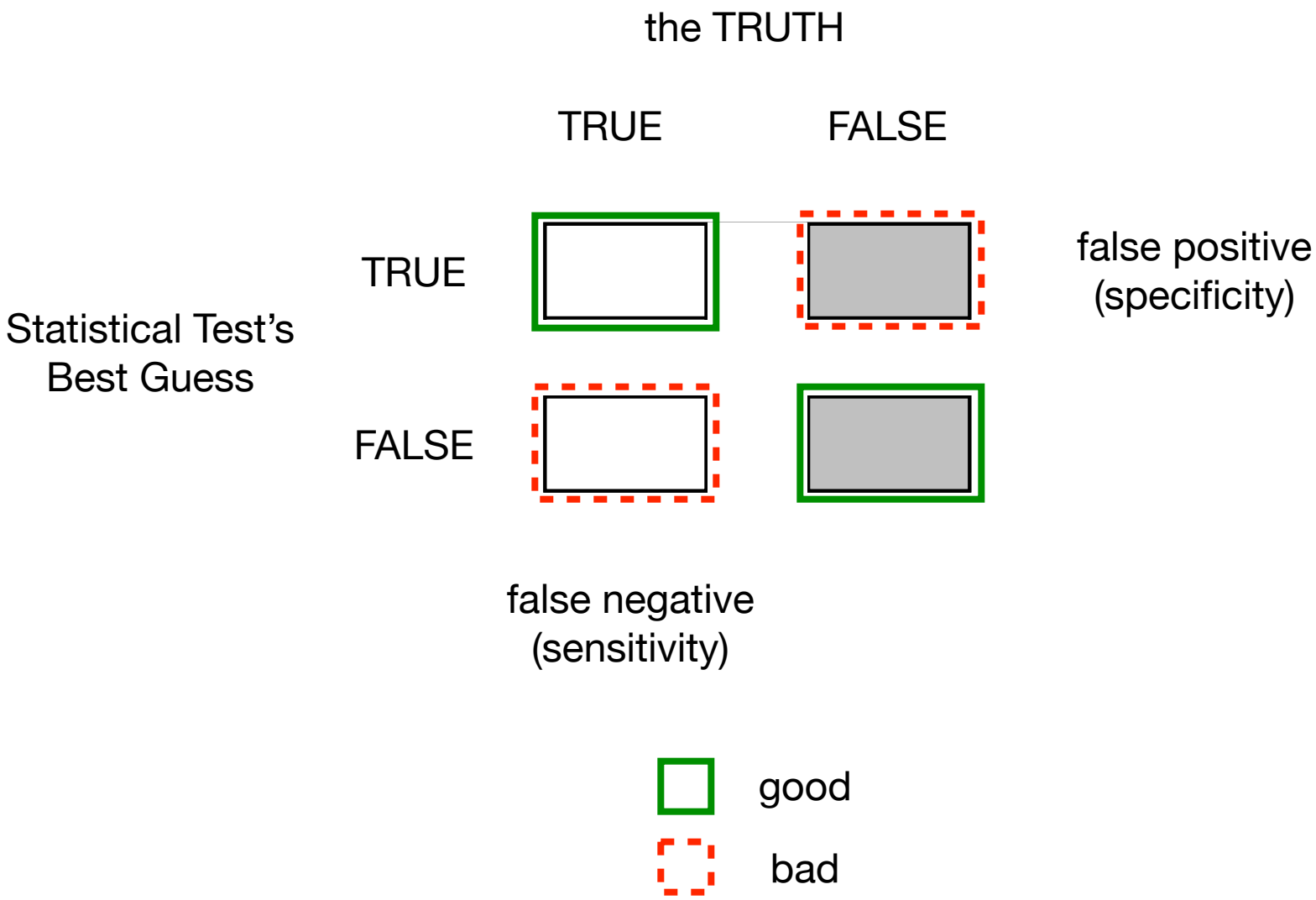
Introduction to RNA-Seq

Part III: Comparing Genes

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology

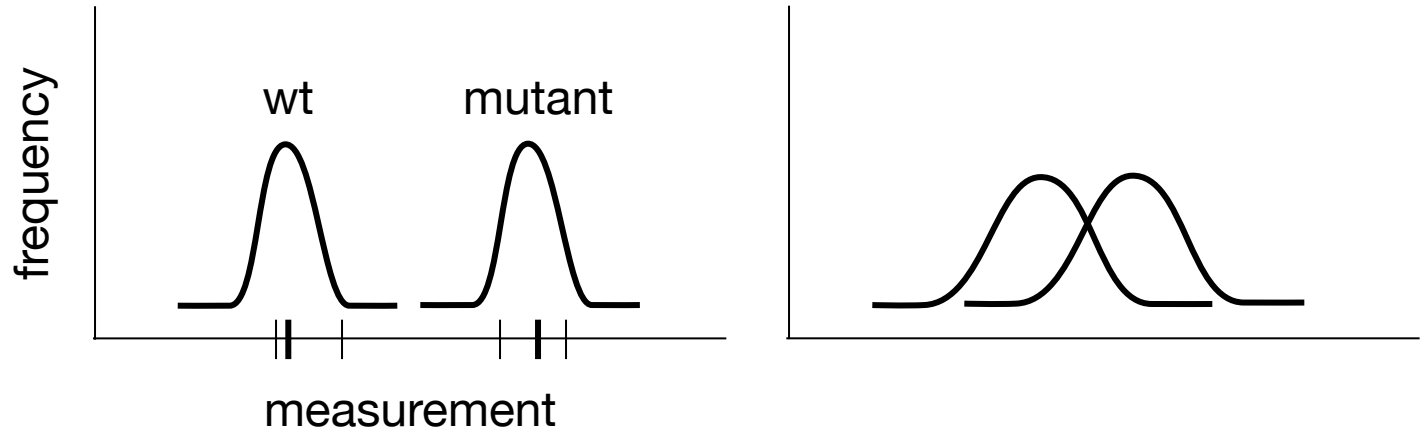
30 July 2013

Goal: More Good, Less Bad



Uncertainty in Sampling Distributions Makes Hypothesis Testing Difficult

H_0: wt = mutant



An Example: Student's t-test

$$t = \frac{\text{difference}}{\text{variation}}$$

wt	100, 105, 110
mutant	120, 115, 125

$t(4) = -3.674$,
$p=0.021$

mean_{wt}	$\frac{100 + 105 + 110}{3} = 105$
mean_{mutant}	$\frac{120 + 115 + 125}{3} = 120$
difference	$\text{mean}_{wt} - \text{mean}_{mutant}$ $120 - 105 = 15$
variation	$s_{x1x2} \times \text{sqrt}(2/n)$ $5 \times \text{sqrt}(2/3) = 4.0825$

Several Complications Arise with Hypothesis Testing of Sequencing Data

Quantitate Genes	Ambiguous Counts
Normalize Counts	Read Depth, Transcriptional Output
Estimate Variance	Count Data, Over Dispersion
Test Significance	
Correct for Multiple Comparisons	10,000+ genes

False Discovery Rate Methods Attempt to Correct for Multiple Comparisons

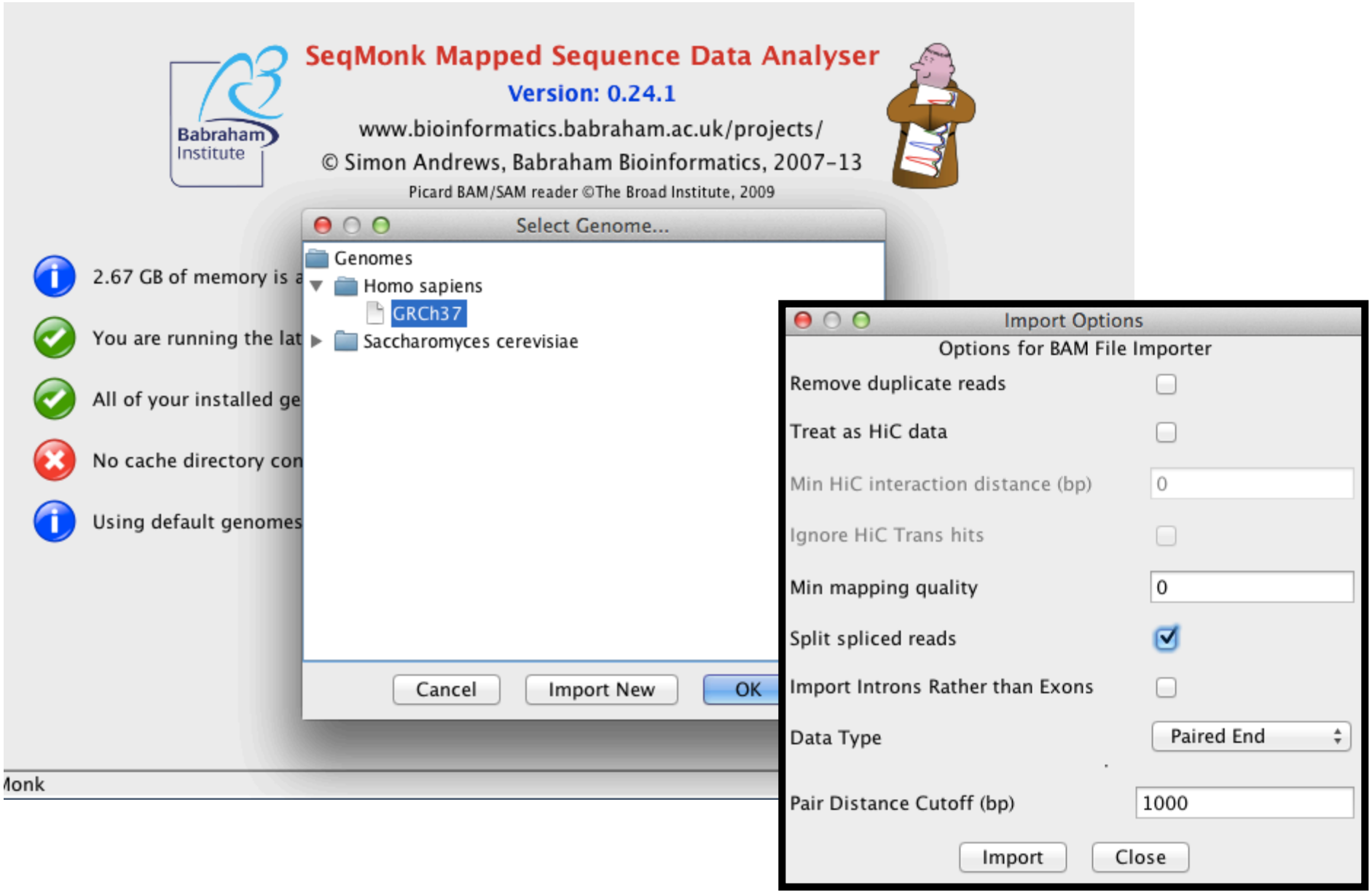
10,000 significance tests

p-value <0.05	FDR corrected <0.05
600 genes	100 genes
expect 500 by chance alone	attempts to limit 5 by chance alone

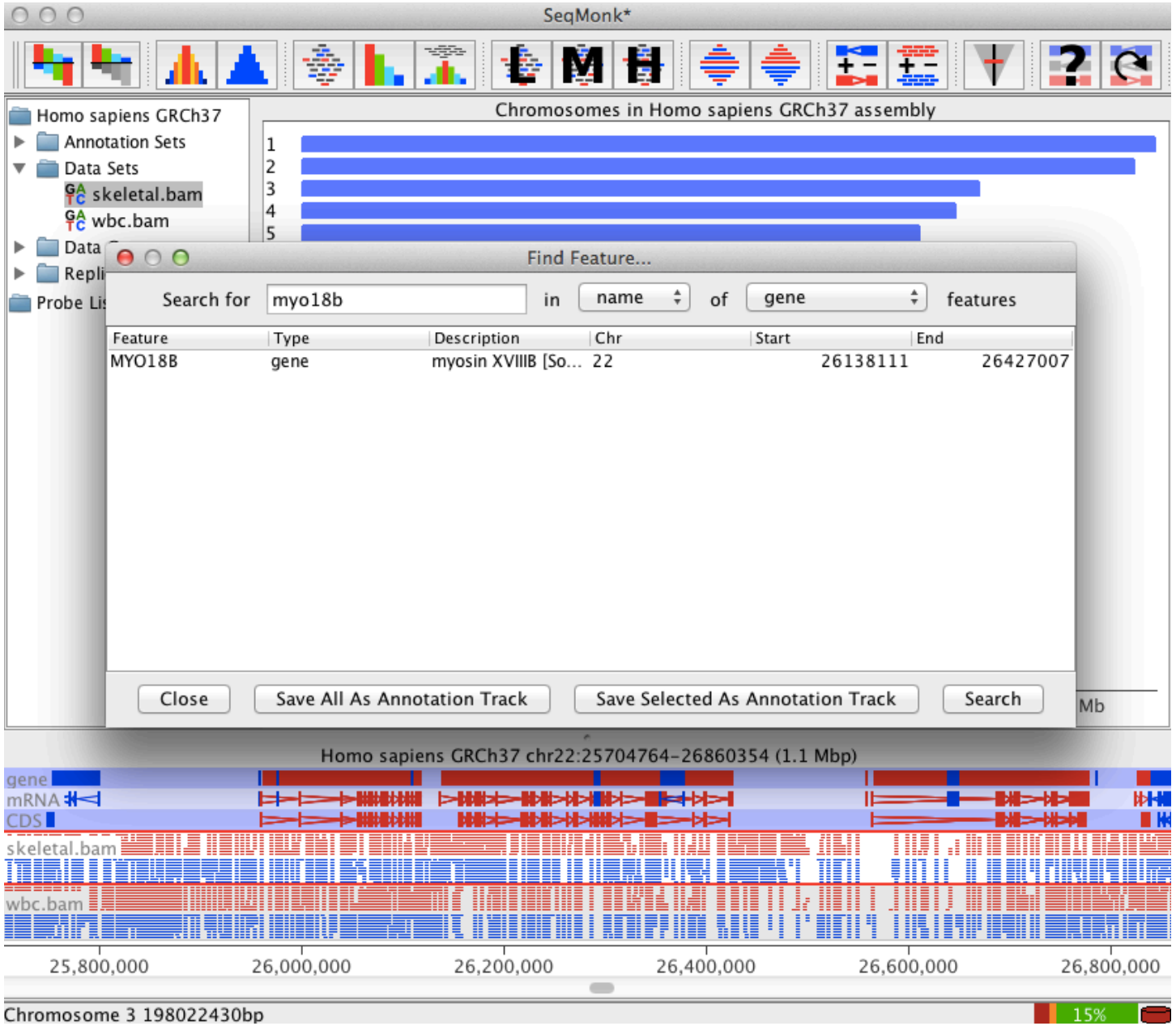
Exploratory Data Analysis

Part I: Mapping Reads	*Part II: Quantitating Abundance*	*Part III: Comparing Genes*
FastQC	Annotations	**SeqMonk**
Bowtie	RNA-SeQC	Cuffdiff
SAMtools	HTSeq	CummeRbund
IGV	Cufflinks	DESeq
TopHat		

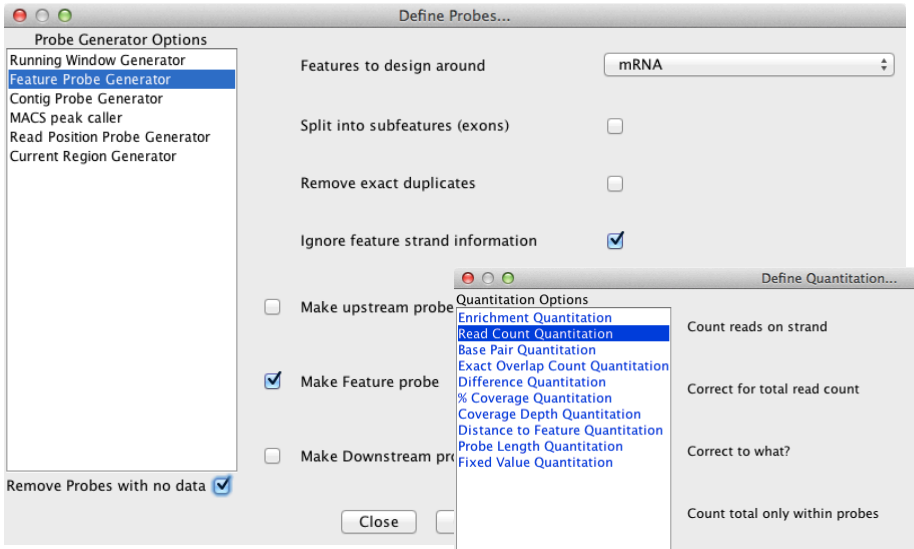
Create Project and Import .BAM Files

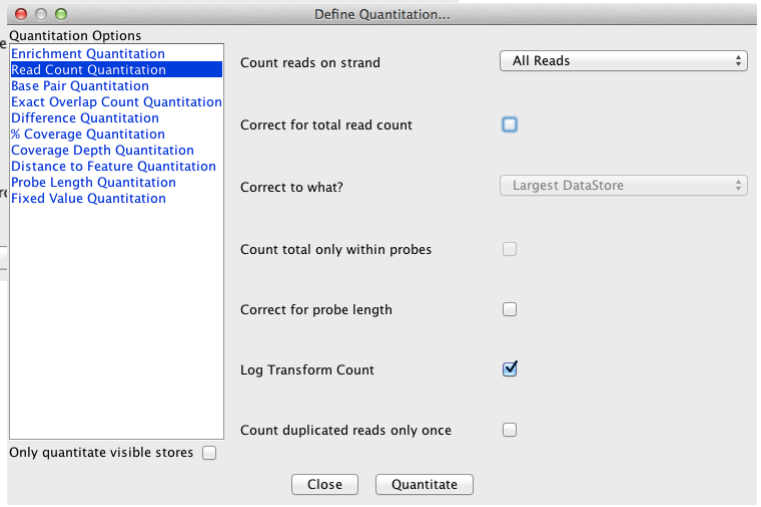


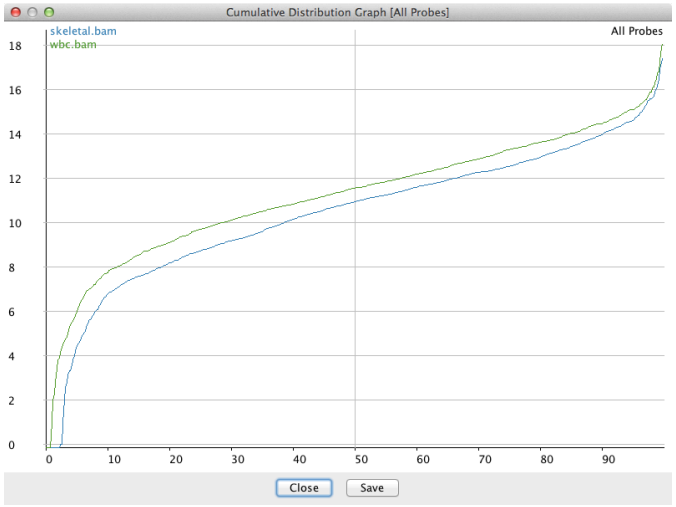
Find Genes and Explore Mapping Data

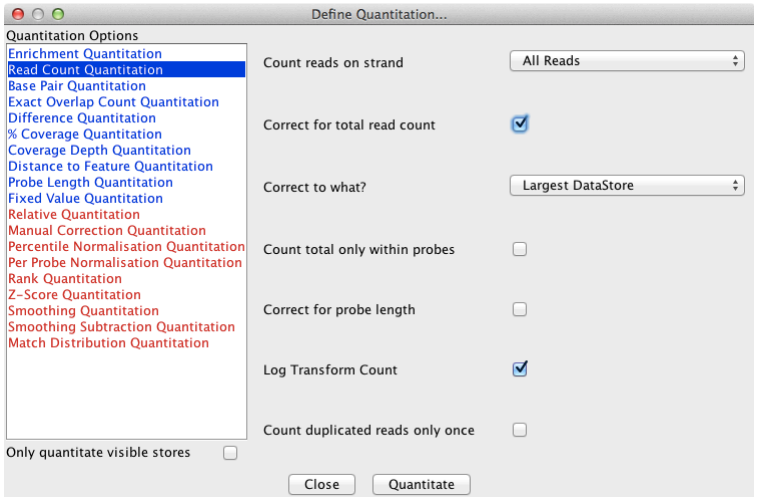


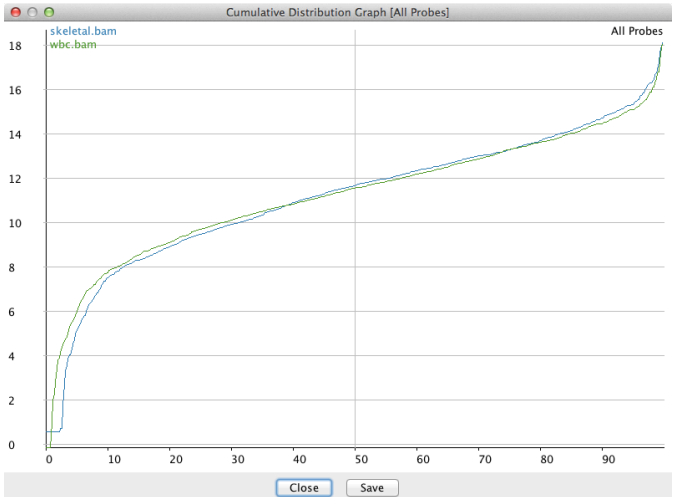
Define Probes, Quantitate, and Normalize



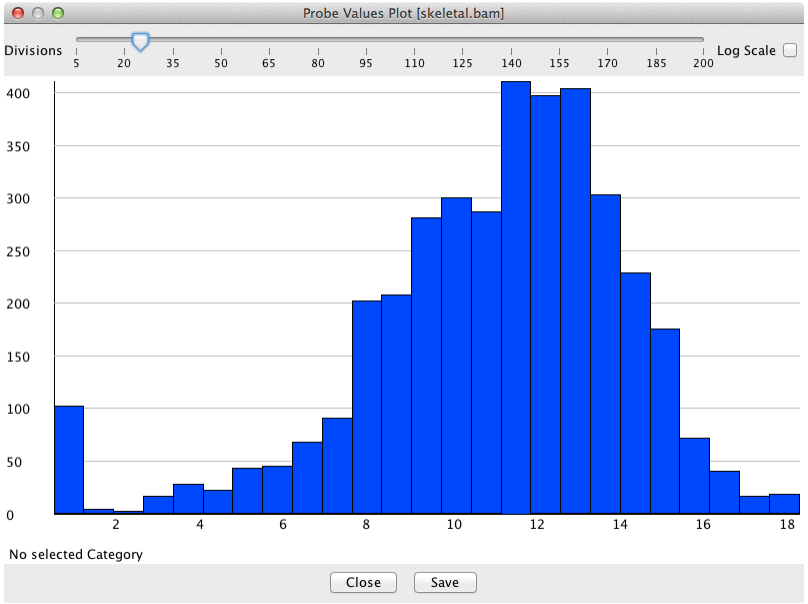


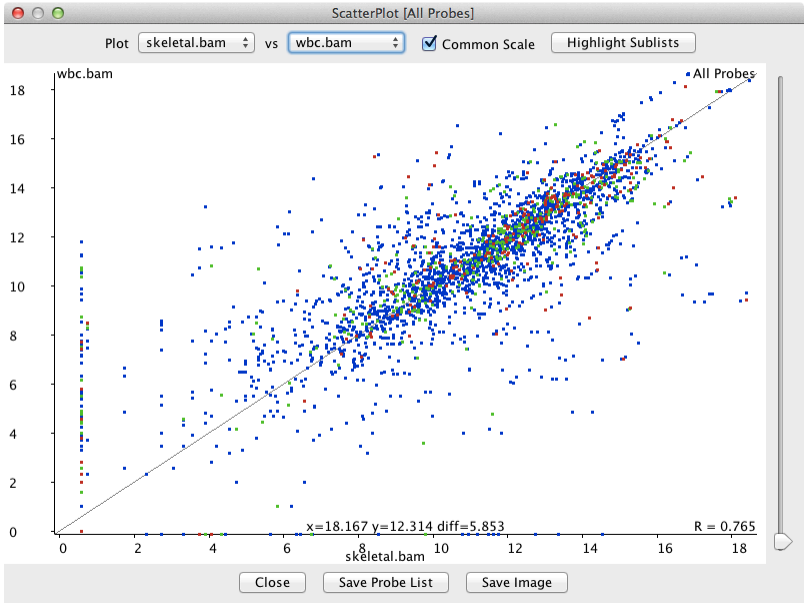


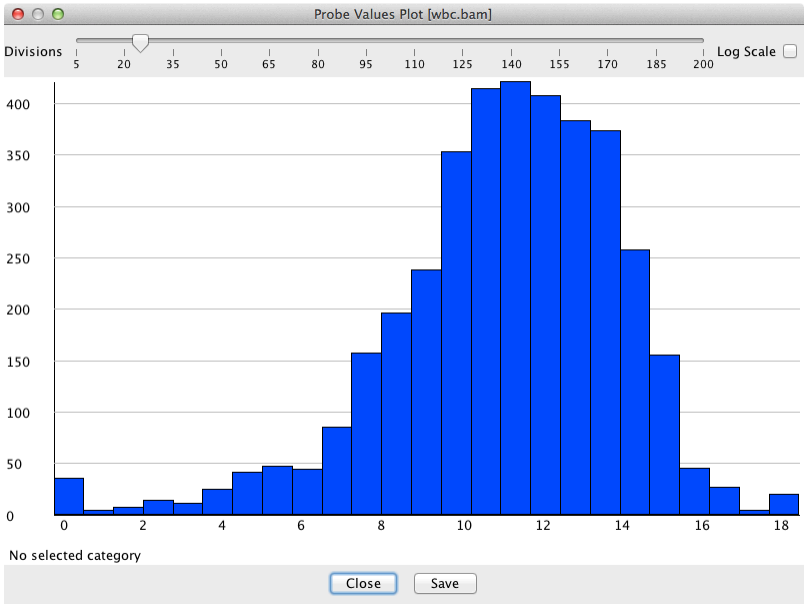


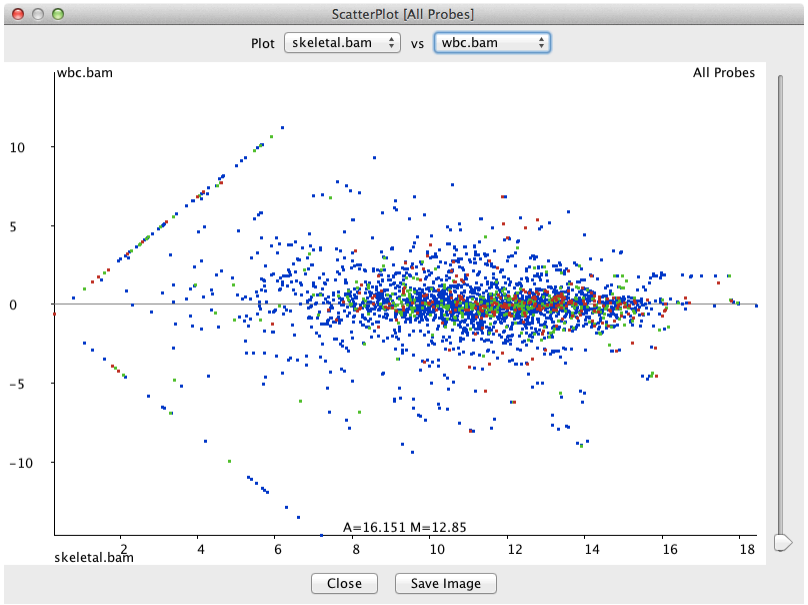


Visualize Data Multiple Ways

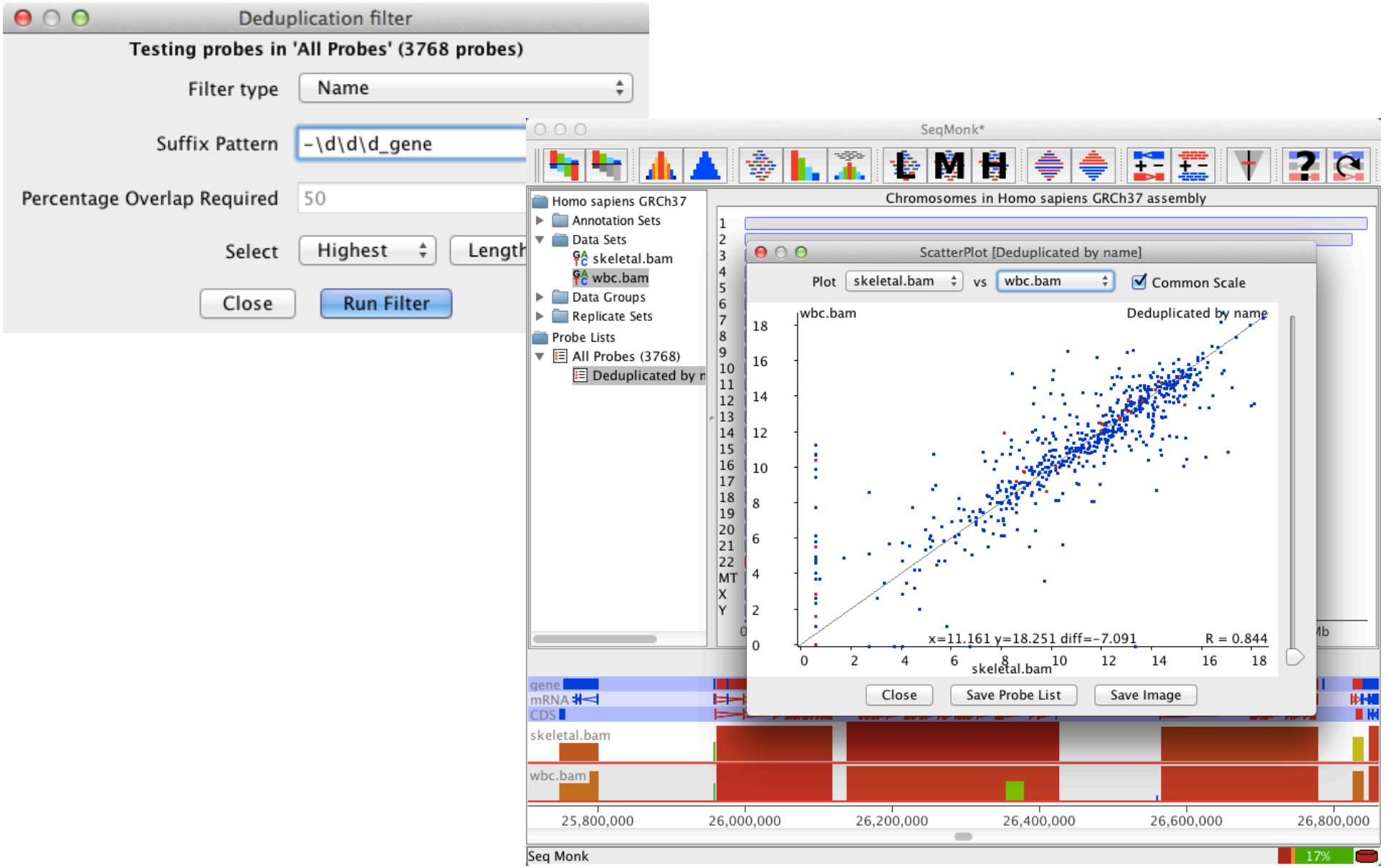




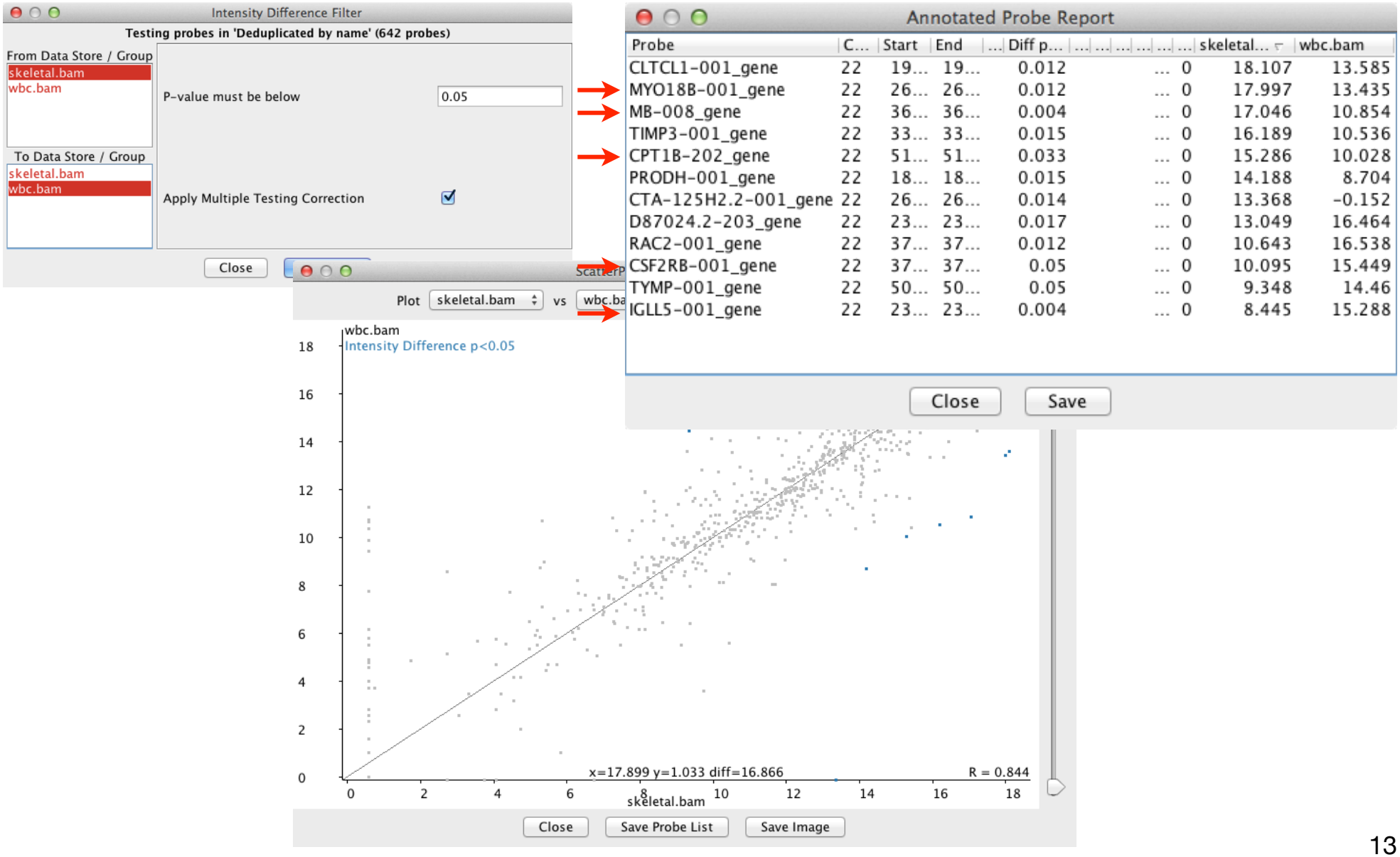




Deduplicate Multiple Isoforms



Identify Differentially Abundant Genes



Comparing Genes - Strategy A

Part I: Mapping Reads	*Part II: Quantitating Abundance*	*Part III: Comparing Genes*
FastQC	Annotations	SeqMonk
Bowtie	RNA-SeQC	**Cuffdiff**
SAMtools	HTSeq	CummeRbund
IGV	Cufflinks	DESeq
TopHat		

Choose Your Own Adventure: Library Normalization Methods

classic-fpkm v0.7

Library size factor is set to 1 - no scaling applied to FPKM values or fragment counts.

default for Cufflinks

upper-quartile-norm v1.0

With this option, Cufflinks normalizes by the upper quartile of the number of fragments mapping to individual loci instead of the total number of sequenced fragments. This can improve robustness of differential expression calls for less abundant genes and transcripts.

geometric v2.0

FPKMs and fragment counts are scaled via the median of the geometric means of fragment counts across all libraries, as described in Anders and Huber (Genome Biology, 2010). This policy identical to the one used by DESeq. default for Cuffdiff

quartile v2.1

FPKMs and fragment counts are scaled via the ratio of the 75 quartile fragment counts to the average 75 quartile value across all libraries.

Identify Differences at Gene-level, Isoform-level, and More Using Cuffdiff

```
$ cd bodymap
$ cuffdiff -o skeletal-wbc_diff_out -L skeletal,wbc
  ~/genomes/human/Homo_sapiens.GRCh37.72-chr22.gtf
  skeletal_thout/accepted_hits.bam wbc_thout/accepted_hits.bam
```

```
[12:31:47] Loading reference annotation.
[12:31:47] Inspecting maps and determining fragment length distributions.
> Map Properties:
>   Total Map Mass: 1505904.96
>   Fragment Length Distribution: Empirical (learned)
>                 Estimated Mean: 195.77
>              Estimated Std Dev: 78.52
> Map Properties:
>   Total Map Mass: 2627229.15
>   Fragment Length Distribution: Empirical (learned)
>                 Estimated Mean: 179.15
>              Estimated Std Dev: 34.55
[12:33:09] Modeling fragment count overdispersion.
[12:33:09] Testing for differential expression and regulation in locus.
> Processed 745 loci.                          [*************************] 100%
Performed 1842 isoform-level transcription difference tests
```

```
$ cuffdiff -o skeletal-wbc_diff_out -L skeletal,wbc
  -u -b ~/genomes/human/hg19-chr22.fa
  ~/genomes/human/Homo_sapiens.GRCh37.72-chr22.gtf
  skeletal_thout/accepted_hits.bam wbc_thout/accepted_hits.bam
```

Sanity Check Cuffdiff Results

```
$ ls skeletal-wbc_diff_out
```

```
cds.diff                genes.fpkm_tracking      splicing.diff
cds_exp.diff            isoform_exp.diff         tss_group_exp.diff
cds.fpkm_tracking       isoforms.fpkm_tracking   tss_groups.fpkm_tracking
gene_exp.diff           promoters.diff
```

```
$ less -S skeletal-wbc_diff_out/gene_exp.diff
```

```
... gene     locus     sample_1 sample_2 value_1 value_2 log2(FC) p_value     q_value
... MAPK8IP2 chr22:51  skeletal   wbc     6.34865 5.52243 -0.20114 0.830636    0.888678
... BID      chr22:18  skeletal   wbc     421.018 1639.14  1.96098 4.33436e-07 4.66621e-06
... TYMP     chr22:50  skeletal   wbc     69.3598 2429.55  5.13044 3.39897e-17 1.67278e-15
```

```
$ sort -gk13 skeletal-wbc_diff_out/gene_exp.diff | less -S
```

```
... gene     locus     sample_1 sample_2 value_1 value_2 log2(FC) p_value     q_value
... MN1      chr22:28  skeletal   wbc     1472.48 6.18269 -7.8958  0           0
... SMTN     chr22:31  skeletal   wbc     859.493 20.2536 -5.40723 0           0
... MAPK12   chr22:50  skeletal   wbc     3381.69 3.68006 -9.8438  0           0
... MB       chr22:36  skeletal   wbc     66696.7 39.5082 -10.7212 0           0
... CPT1B    chr22:51  skeletal   wbc     2558.69 34.0215 -6.23282 0           0
... APOBEC3G chr22:39  skeletal   wbc     46.1024 1589.57  5.10765 0           0
... PRODH    chr22:18  skeletal   wbc     1462.35 1.69778 -9.75043 5.42101e-20 4.66885e-18
... MYO18B   chr22:26  skeletal   wbc     8649.96 0.06154 -17.1008 5.42101e-20 4.66885e-18
```

Manipulating Cuffdiff Output

Part I: Mapping Reads

- FastQC
- Bowtie
- SAMtools
- IGV
- TopHat

Part II: Quantitating Abundance

- Annotations
- RNA-SeQC
- HTSeq
- Cufflinks

Part III: Comparing Genes

- SeqMonk
- Cuffdiff
- **CummeRbund**
- DESeq

Load Cuffdiff Output with CummeRbund

"Allows for persistent storage, access, exploration, and manipulation of Cufflinks high-throughput sequencing data. In addition, provides numerous plotting functions for commonly used visualizations."

```
$ cd
$ R
```

```
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
        [...]
Type 'q()' to quit R.
```

```
> getwd()
> setwd("bodymap")
> getwd()
> setwd("skeletal-wbc_diff_out")
> getwd()
> library(cummeRbund)
> cuff_data <- readCufflinks()
```

Trapnell, *et al.*, 2012 Nature Protocols

Search Documentation for Information About CuffSet

```
> cuff_data
```

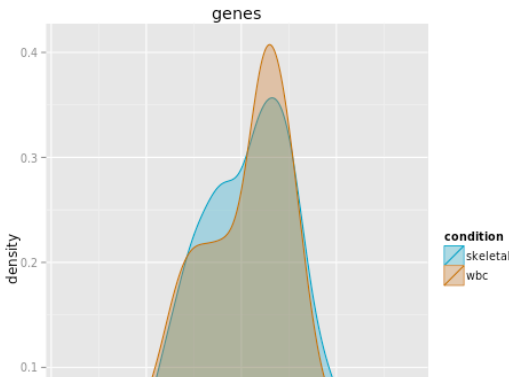
```
CuffSet instance with:
     2 samples
     1207 genes
     4335 isoforms
     0 TSS
     0 CDS
     0 promoters
     0 splicing
     0 relCDS
```

```
> ?CuffSet
> ??CuffSet
> ?CuffSet-class
> ?"CuffSet-class"
```

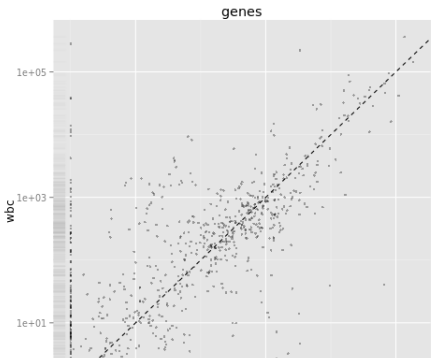
```
CuffSet-class       package:cummeRbund       R Documentation

Class "CuffSet"

Description:
```

Plot Summary Information

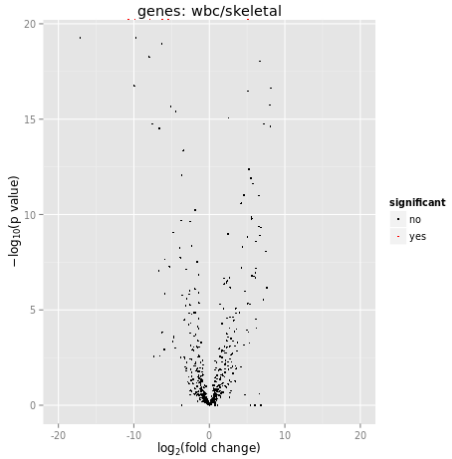
```
> png("density.png")
> csDensity(genes(cuff_data))
> dev.off()
```




```
> png("scatter.png")
> csScatter(genes(cuff_data), 'skeletal', 'wbc')
> dev.off()
```



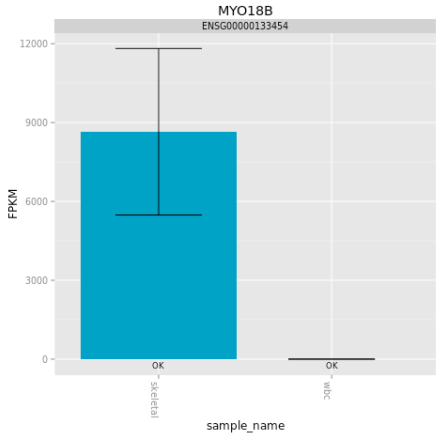

```
> png("volcano.png")
> csVolcano(genes(cuff_data), 'skeletal', 'wbc')
> dev.off()
```



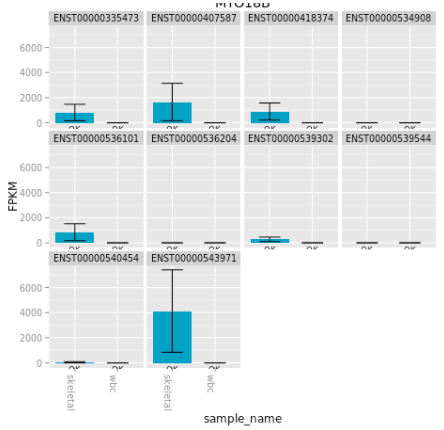
Plot Gene-Specific Information

```
> mygene <- getGene(cuff_data, 'MYO18B')
> mygene
> mygene@annotation
> mygene@fpkm
```

```
> png("myo18b.png")
> expressionBarplot(mygene)
> dev.off()
```




```
> png("myo18b-splice.png")
> expressionBarplot(isoforms(mygene))
> dev.off()
```



Subset Cuffdiff Results

```
> genes(cuff_data)
> diffData(genes(cuff_data))
> nrow(diffData(genes(cuff_data)))
> head(diffData(genes(cuff_data)))
```

gene_id	sample_1	sample_2	value_1	value_2	log2(FC)	p_value	q_value
ENSG35	skeletal	wbc	6.34865	5.52243	-0.201146	8.30636e-01	8.88678e-01
ENSG75	skeletal	wbc	421.01800	1639.14000	1.960980	4.33436e-07	4.66621e-06
ENSG08	skeletal	wbc	69.35980	2429.55000	5.130440	3.39897e-17	1.67278e-15
ENSG70	skeletal	wbc	556.77000	305.11900	-0.867710	1.17985e-02	3.69207e-02
ENSG08	skeletal	wbc	2.23837	24.41550	3.447280	8.59235e-03	2.83260e-02
ENSG11	skeletal	wbc	549.95200	631.16500	0.198712	5.53161e-01	6.87378e-01

```
> gene_diff_data <- diffData(genes(cuff_data))
> str(gene_diff_data)
> gene_diff_data$significant
> sig_gene_data <- subset(gene_diff_data, (significant == 'yes'))
> sig_gene_data
> nrow(sig_gene_data)
> head(sig_gene_data)
```

Write Cuffdiff Subsets to New Files

```
> write.table(sig_gene_data, 'diff_genes.txt', sep='\t',
  row.names = F, col.names = T, quote = F)
```

```
> sig_gene_data2 <- subset(gene_diff_data, (q_value < 0.1))
> nrow(sig_gene_data2)
> write.table(sig_gene_data2, 'diff_genes2.txt', sep='\t',
    row.names = F, col.names = T, quote = F)
```

```
> sig_gene_data3 <- subset(gene_diff_data, (log2_fold_change > 2))
> nrow(sig_gene_data3)
> write.table(sig_gene_data3, 'diff_genes3.txt', sep='\t',
    row.names = F, col.names = T, quote = F)
```

Comparing Genes - Strategy B

Part I: Mapping Reads	*Part II: Quantitating Abundance*	*Part III: Comparing Genes*
FastQC	Annotations	SeqMonk
Bowtie	RNA-SeQC	Cuffdiff
SAMtools	HTSeq	CummeRbund
IGV	Cufflinks	**DESeq**
TopHat		

Declare Study Design for DESeq

```
> setwd(“~/bodymap”)
> library(“DESeq”)
> bodymapDesign <- data.frame(
      sample = c("1","2"),
      filename = c("skeletal.counts", "heart.counts"),
      condition = c("skeletal", "heart")
   )
> ls()
```

```
[1] "bodymapDesign"  "cuff_data"      "gene_diff_data" "mygene"
[5] "sig_gene_data"  "sig_gene_data2" "sig_gene_data3"
```

Anders and Huber, 2010 Genome Biology

Access Information in a Data Frame Multiple Ways

```
> bodymapDesign
```

```
  sample         filename condition
1      1 skeletal.counts  skeletal
2      2    heart.counts     heart
```

```
> str(bodymapDesign)
> dim(bodymapDesign)
> bodymapDesign[ 1,2 ]
```

```
[1] skeletal.counts
Levels: heart.counts skeletal.counts
```

```
> bodymapDesign[ 1, ]
> bodymapDesign[ ,2 ]
> str(bodymapDesign)
> bodymapDesign$sample
```

Load HTSeq Counts into DESeq

```
> cds <- newCountDataSetFromHTSeqCount(bodymapDesign,
                                       directory = "~/bodymap")
> cds
> ?"CountDataSet-class"
> conditions(cds)
> counts(cds)
> str(counts(cds))
> dim(counts(cds))
```

```
[1] 1209    2
```

```
> counts(cds)[ ,1 ]
> counts(cds)[ ,2 ]
> counts(cds)[ order(counts(cds)[,1]), ]
> counts(cds)[ order(counts(cds)[,2]), ]
```

Estimate Size Factors to Normalize Read Depth

```
> head(counts(cds, normalized = FALSE))
```

```
ENSG00000008735   24   38
ENSG00000015475 1251 1254
ENSG00000025708  100  382
ENSG00000025770 1248  670
ENSG00000040608    4   19
ENSG00000054611 1558  848
```

```
> cds <- estimateSizeFactors(cds)
> sizeFactors(cds)
```

```
0.9808511 1.0195227
```

```
> head(counts(cds, normalized = TRUE))
```

```
ENSG00000008735   24.468545   37.27234
ENSG00000015475 1275.422890 1229.98733
ENSG00000025708  101.952269  374.68514
ENSG00000025770 1272.364322  657.17027
ENSG00000040608    4.078091   18.63617
ENSG00000054611 1588.416357  831.76177
```

Choose Your Own Adventure: Estimate Dispersions Methods

poisson *cufflinks only*

The Poisson model is used, where the variance in fragment count is predicted to equal the mean across replicates.
Not recommended.

pooled *default*

Use the samples from all conditions with replicates to estimate a single pooled empirical dispersion value and assign it to all samples.

per-condition

For each condition with replicates, compute a gene's empirical dispersion value by considering the data from samples for this condition. For samples of unreplicated conditions, the maximum of empirical dispersion values from the other conditions is used. (Note: This method was called "normal" in previous versions.)

blind

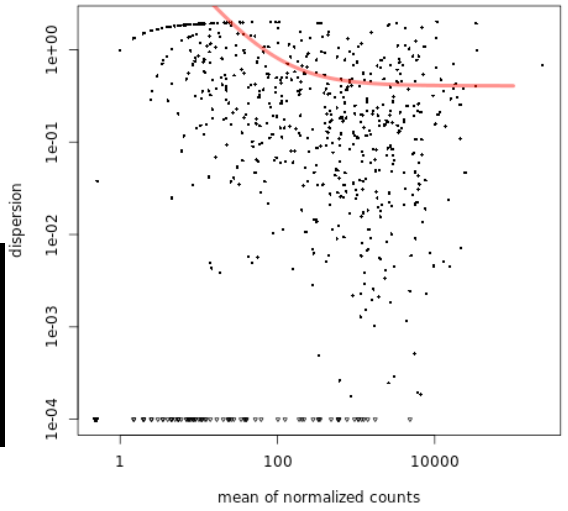
Ignore the sample labels and compute a gene's empirical dispersion value as if all samples were replicates of a single condition. This can be done even if there are no biological replicates. This method can lead to loss of power; see the vignette for details.

Estimate Dispersions (for Data with No Replicates)

```
> cds <- estimateDispersions(cds, method = "blind",
                             sharingMode = "fit-only")
> head(fData(cds))
```

```
ENSG00000008735  1.7137051
ENSG00000015475  0.4378672
ENSG00000025708  0.5750726
ENSG00000025770  0.4474878
ENSG00000040608  3.9611769
ENSG00000054611  0.4390024
```

```
> png("dispersion.png")
> plotDispEsts(cds)
> dev.off()
```




```
> fData(cds)[ order(counts(cds)[,1]), ]
```

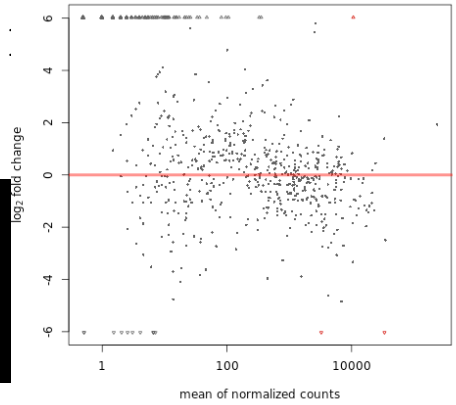
```
   [1]        Inf 20.9901960 41.5747597  4.1482803 82.7438871  8.6394578
   [7]        Inf  2.9787028        Inf        Inf  0.4094231 27.8517173
                                    ...
[1201]  0.4075418  0.4078288  0.4077216  0.4079949  0.4075410  0.4073861
[1207]  0.4067915  0.4068305  0.4058031
```

Identify Differentially Abundant Genes

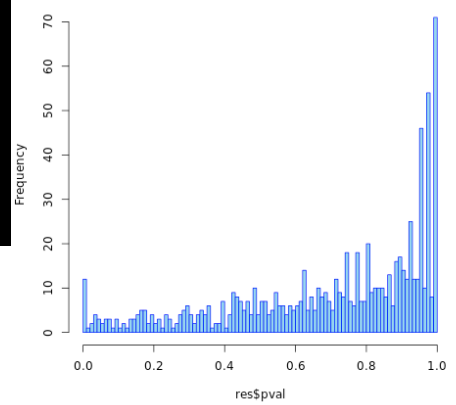
```
> res <- nbinomTest( cds, "skeletal", "heart" )
> res
> head(res)
> str(res)
```

```
'data.frame':   1209 obs. of  8 variables:
 $ id          : chr  "ENSG35" "ENSG75" "ENSG08
 $ baseMean    : num  30.9 1252.7 238.3 964.8 1
                                    ...
```

```
> png("maplot.png")
> plotMA(res)
> dev.off()
```




```
> png("pval.png")
> hist(res$pval, breaks = 100, col = "skyblue",
       border="blue", main = "")
> dev.off()
```



Subset DESeq Results

```
> head(res[ order(res$log2FoldChange, -res$baseMeanA), ])
```

	id	baseMeanA	baseMeanB	log2(FC)	pval	padj
ENSG78		14.273318	0	-Inf	0.6208583	1
ENSG60		13.253795	0	-Inf	0.6466301	1
ENSG17		13.253795	0	-Inf	0.6466301	1
ENSG94		13.253795	0	-Inf	0.6466301	1
ENSG03		8.156182	0	-Inf	0.7897517	1
ENSG86		6.117136	0	-Inf	0.8524392	1

```
> resSig <- res[ res$padj < 0.1, ]
> head(resSig[ order(resSig$padj), ])
```

	id	baseMeanA	baseMeanB	log2(FC)	pval	padj
ENSG70		0.000	21305.06769	Inf	8.8888e-10	7.1910e-07
ENSG58		6484.164	43.15745	-7.231167	6.3338e-05	2.5620e-02
ENSG71		66761.405	642.45750	-6.699269	1.1109e-04	2.9959e-02
<NA>		NA	NA	NA	NA	NA
<NA>		NA	NA	NA	NA	NA
<NA>		NA	NA	NA	NA	NA

Paired Samples and Multi-Factor Designs

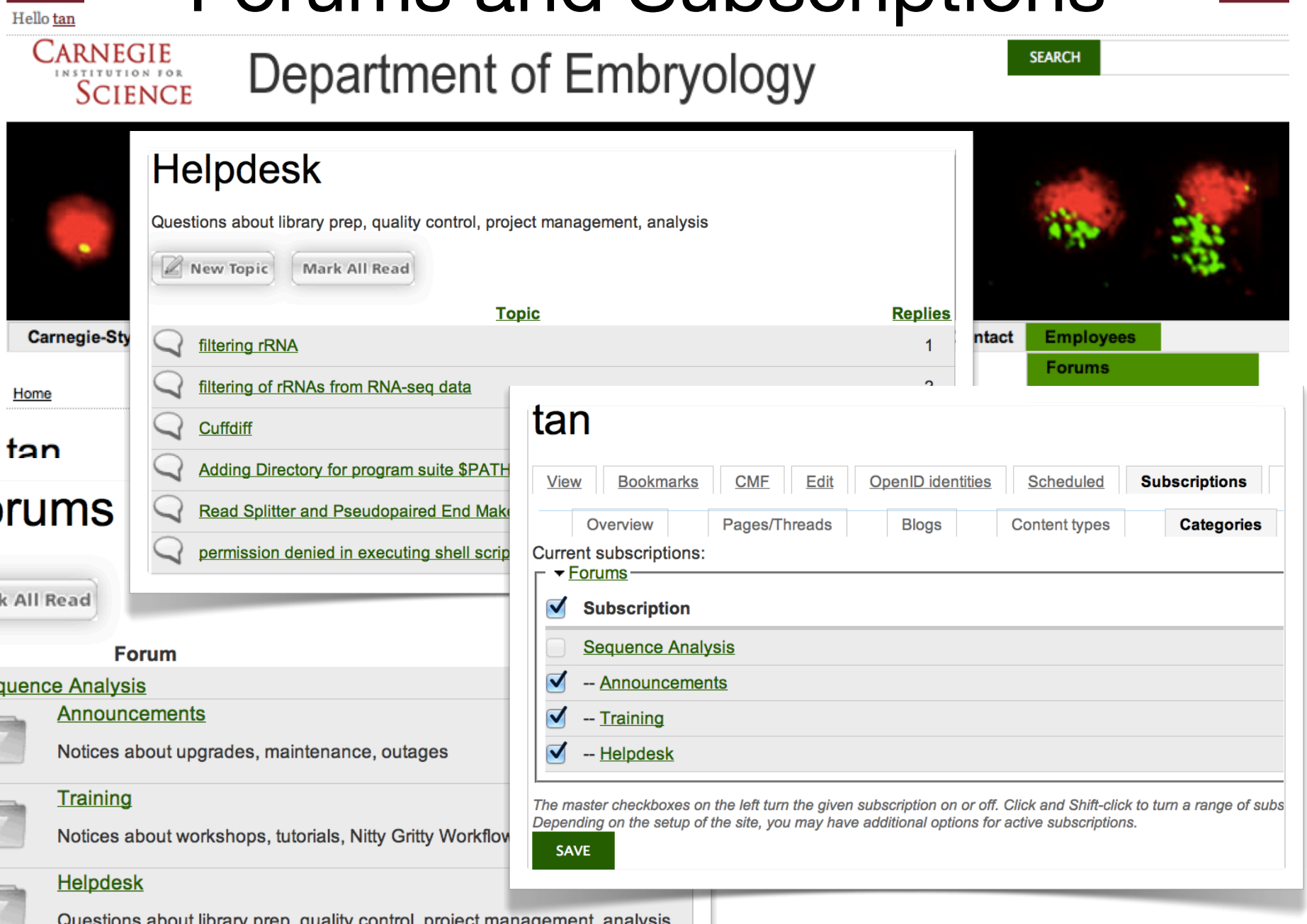
Patient	Condition A	Condition B
1	normal	cancer
2	normal	cancer
3	normal	cancer

wild-type	mutant
wild-type	mutant
+drug	+drug

Wrap-Up

Part I: Mapping Reads	*Part II: Quantitating Abundance*	*Part III: Comparing Genes*
FastQC	Annotations	SeqMonk
Bowtie	RNA-SeQC	Cuffdiff
SAMtools	HTSeq	CummeRbund
IGV	Cufflinks	DESeq
TopHat		

Forums and Subscriptions



Workshop Slides and More

Carnegie-Style Science | Labs | Research Programs | Faculty | Resources | Seminars | Outreach | Contact

Home

Tan Lab

COURSE MATERIALS

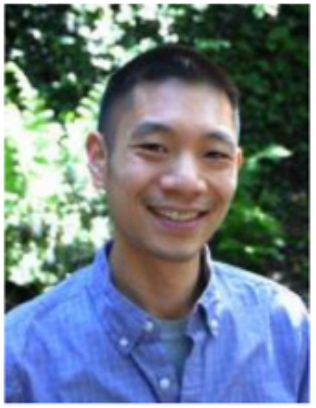
Workshops

- Intro to RNA-Seq - Summer 2013 - Day 1 / Day 2 / Day 3
- Intro to Unix - Spring 2013 - CCG Short Course - Day 1 slides / Day 2 slides

Perl Thursdays

- July 18, 2013 - refining session
- June 27, 2013 - create average coverage map for a set of genes (top200.cov groupCoverage.pl)
- May 30, 2013 - refining session (gt1kbexon-v2.pl)
- May 2, 2013 - "Hello, world!", filter for exons on + strand >1 kb (mouse.gtf gt1kbexon.pl)

Nitty Gritty Workflows



Frederick Tan
Bioinformatics
Office Telephone: (410) 246-3084
Lab Telephone: (410) 246-3083
Department Fax: (410) 243-6311
Email

Lab Members »

Rosa Alcazar, Visiting Scientist
- Research Summary
- Course Material
- Sequence Analysis