# UQ-Brisbane SISG

## Module 9: Gene Expression and Epigenetic Profiling

SISG Brisbane 2017

Monday February 13, 2017

“Data Normalization”

Greg Gibson

Georgia Institute of Technology

greg.gibson@biology.gatech.com

## Gene Expression Data is analyzed on the log base 2 scale

1. Log transformation makes the data more normally distributed, minimizing biases due to the common feature that a small number of genes account for over half the transcripts

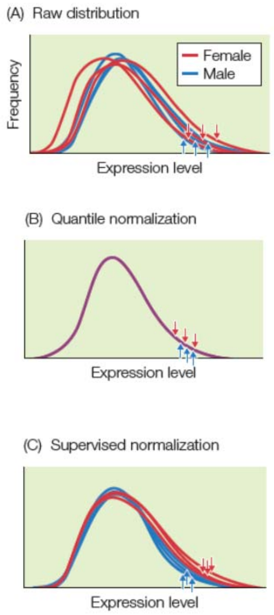
2. Log base 2 is convenient, because in practice most differential expression is in the range of 1.2x to 8x, depending on the contrast of interest and complexity of the sample.

3. It is also intuitively simple to infer fold changes in a symmetrical manner:
   A difference of -1 unit corresponds to half the abundance, and +1 to twice the abundance
   A difference of -2 units corresponds to a quarter the abundance, and +3 to 8-times the abundance

4. The log scale is insensitive to mean centering, so it is simple to just set the mean or median to 0, preserving the relative abundance above or below the sample average

5. It is sometimes useful to add 1 to all values before taking the log, to avoid “0” returning #NUM!

## Rank vs Absolute Expression



Raw data:
no effect

Variance transformed:
no effect

Mean centered:
significant effect

## Hypothesis testing

1. Generally we are interested in asking whether there is a significant difference between two or more treatment group(s) on a gene-by-gene basis

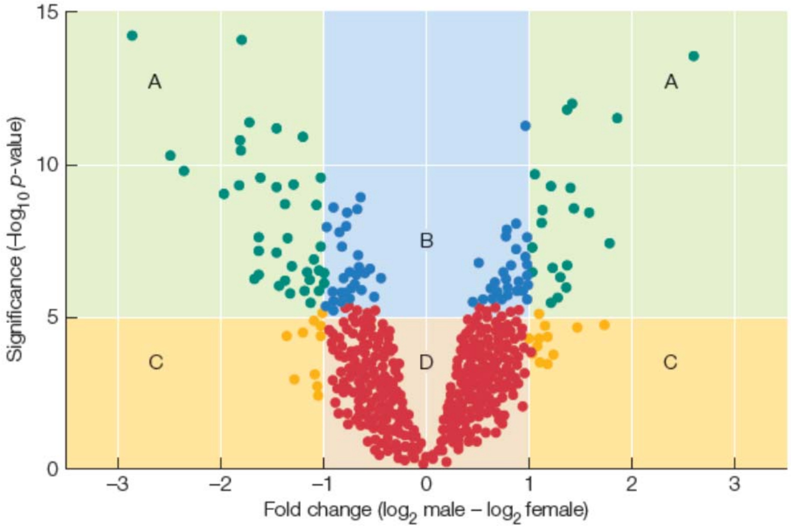
2. For a simple contrast, we can use a t-test to test the hypothesis. Significance is always a function of:
    1. The difference between the two groups: [5,6,4] vs [7,5,6] has a diff of 1
    2. The variance within the groups: [2,5,8] vs [3,6,9] does as well, but is less obvious
    3. The sample size: [5,6,4,4,6,5] and [7,5,6,5,6,7] is better

3. For contrasts involving multiple effects, we usually use General Linear Models in the ANOVA framework (analysis of variance) -
    significance is assessed as the F ratio or between sample to residual sample variance

4. Very robust statistics also allow you to evaluate INTERACTION EFFECTS, namely not just whether two treatments are individually significant, but also whether one depends on the other

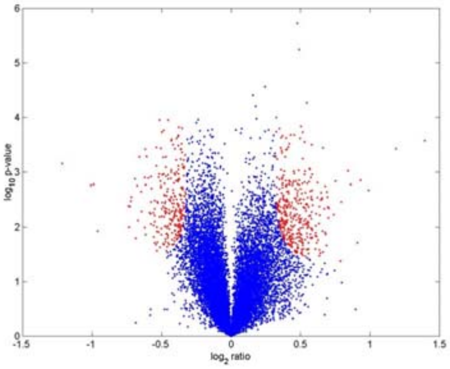
5. Given a list of p-values and DE estimates, we need to evaluate a significance threshold, which is usually
    done using False Discovery Rate (FDR) criteria, either B-H or a qvalue

## Transcriptome Volcano plots



## Variance estimators

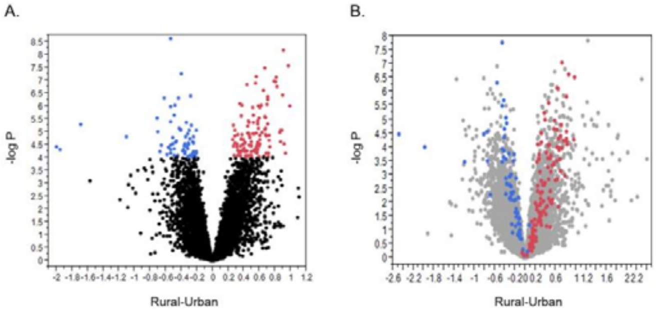
- Gene-specific approach means that the power for each gene varies, but shrinkage can equilibrate the variance
- Permutation approach may be more appropriate where you have many treatments with low replication



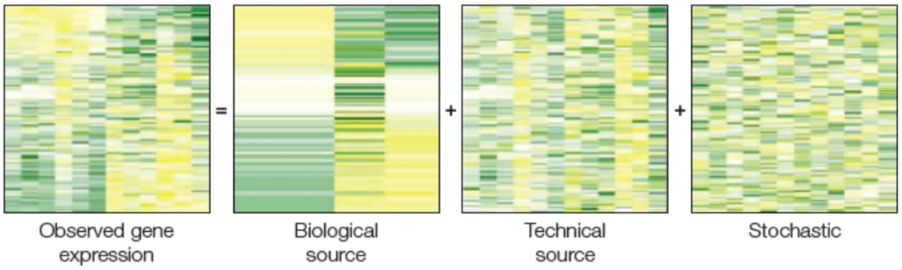
Gary Churchill, Katie Kerr

## Beware false negatives and pathway annotations

1. Although powerful, DE analysis is also intrinsically under-powered, so there is a high false negative rate
2. Consequently, when you see a gene set annotated as “perturbed by drug x in cell-type y of females with disease z”, beware! Most likely a replicate of the experiment would give a completely different list.
3. Conversely, some annotations, eg “Lupus-associated genes” have multiple completely different lists.



## The normalization challenge - visual



John Storey

## The normalization challenge – in math

General model:

$$\underbrace{g_{ij}}_{\text{gene expression}} = \underbrace{b_{i0}}_{\text{baseline expression}} + \underbrace{b_{i1}y_j}_{\text{phenotype effect}} + \underbrace{c_i a_j}_{\text{known batch}} + \underbrace{d_i u_j}_{\text{unknown artifact}} + \underbrace{e_{ij}}_{\text{meas. error}}$$

Control probe model:

$$\underbrace{g_{ij}}_{\text{gene expression}} = \underbrace{b_{i0}}_{\text{baseline expression}} + \underbrace{d_i u_j}_{\text{unknown artifact}} + \underbrace{e_{ij}}_{\text{meas. error}}$$

Normalization strategy (for SVA):

1. Identify the genes that are only affected by unknown artifacts
2. Perform a decomposition of the data for just these genes to identify estimates of the artifacts.
3. Include the artifact estimates in subsequent analyses as if they were known.

Note that ssva (supervised sva) estimates the control probes from external data

Jeff Leek, 2014. *Nucl Acids Res.* **42**: e161 “svaseq: removing batch effects and other unwanted noise from sequencing data”

## Common Strategies for Data Normalization

1. Linear Centering

   log(fluorescence) = μ + Array + Residual

   may also include covariates in the model (eg RIN, cell abundance)

2. Fractional Centering (counts per million)

   RNA-Seq data is usually transformed to the cpm scale to adjust for library size,

   and edgeR makes an additional TMM adjustment for high abundance biases

3. Unsupervised Variance transformations

   (a) Sample (and/or transcript) standardization to z-score

   (b) Inverse Normal Rank Transformation

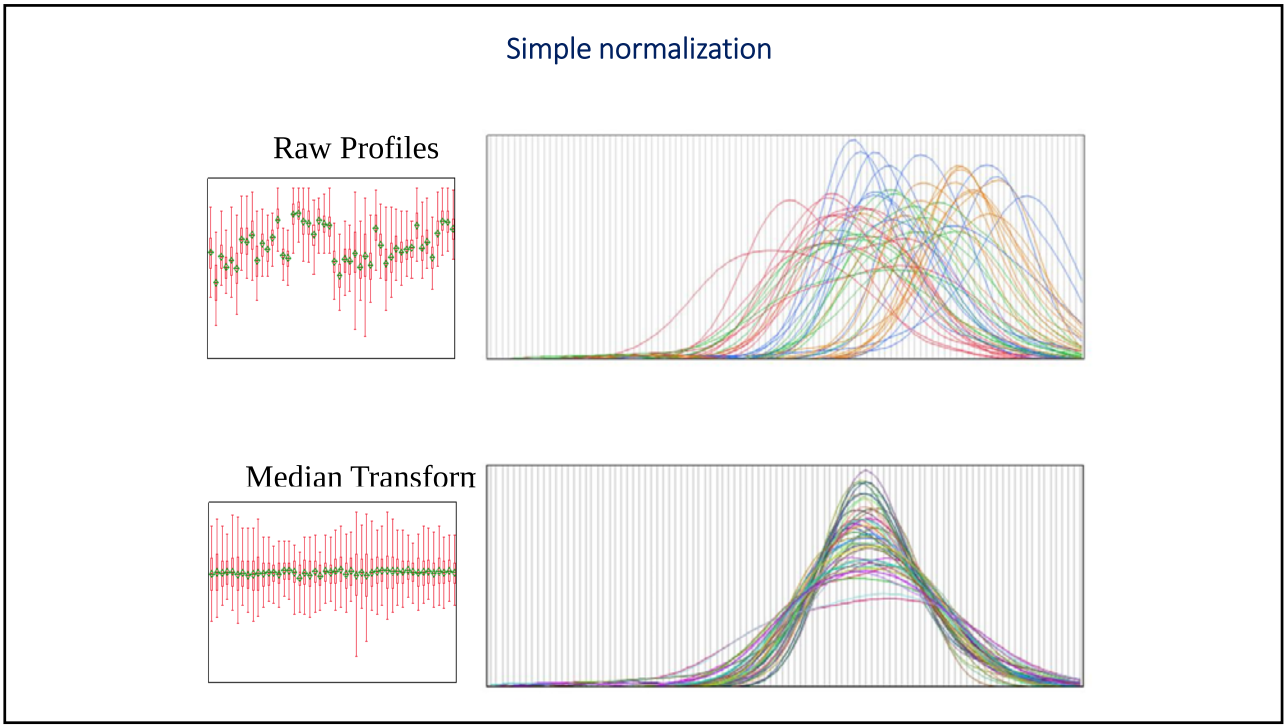
   (c) Quantile normalization

4. Supervised normalization

   (a) PEER factors (a Bayesian approach)

   (b) Surrogate Variable Analaysis (SVA) with COMBAT

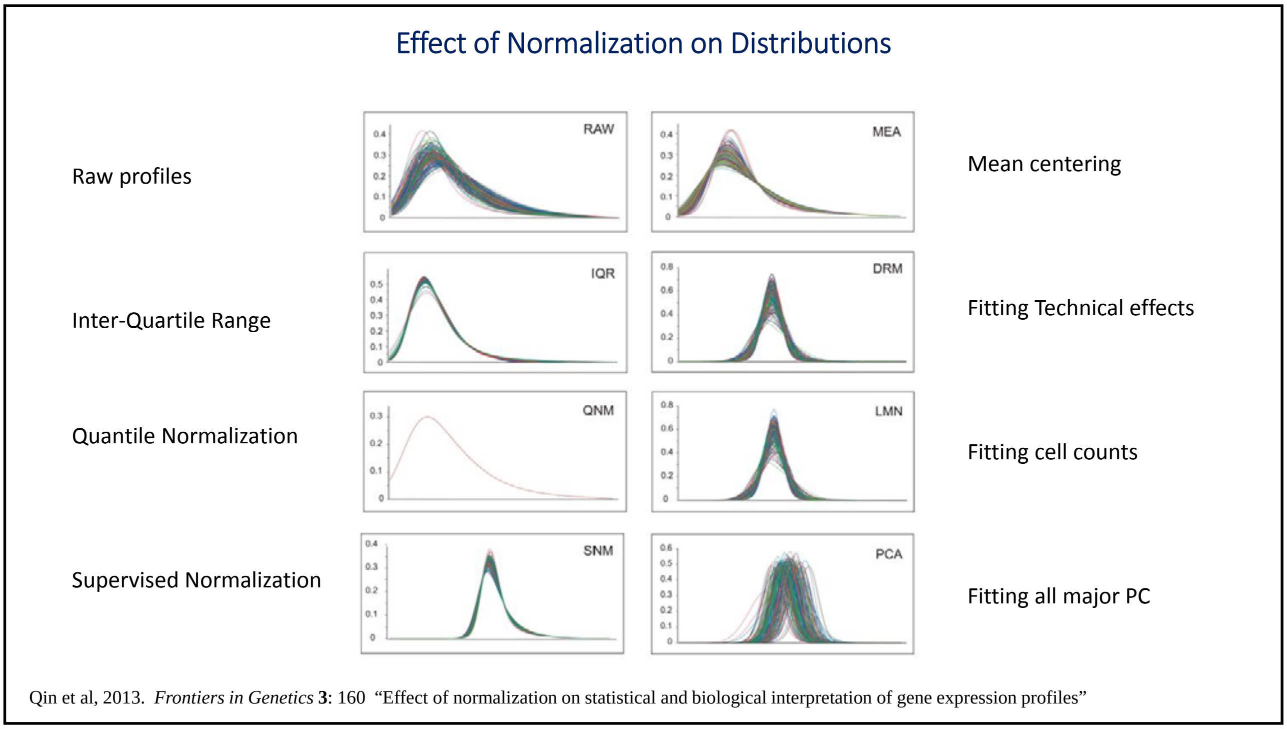
   (c) Supervised Normalization of Microarrays (SNM)



## Simple normalization

Raw Profiles

Median Transforn



## Effect of Normalization on Distributions

Raw profiles

Inter-Quartile Range

Quantile Normalization

Supervised Normalization

Mean centering

Fitting Technical effects

Fitting cell counts

Fitting all major PC

Qin et al, 2013. *Frontiers in Genetics* **3**: 160 “Effect of normalization on statistical and biological interpretation of gene expression profiles”

## Effect of Normalization on Covariance

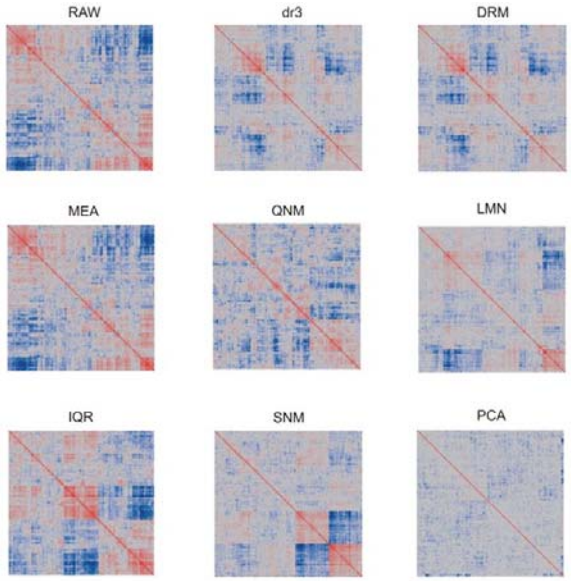
Each heat map shows the pairwise correlation between samples of the first 10 PC of the dataset.

Red implies highly similar profiles, Blue dissimilar.

Mean centering does not alter the covariance

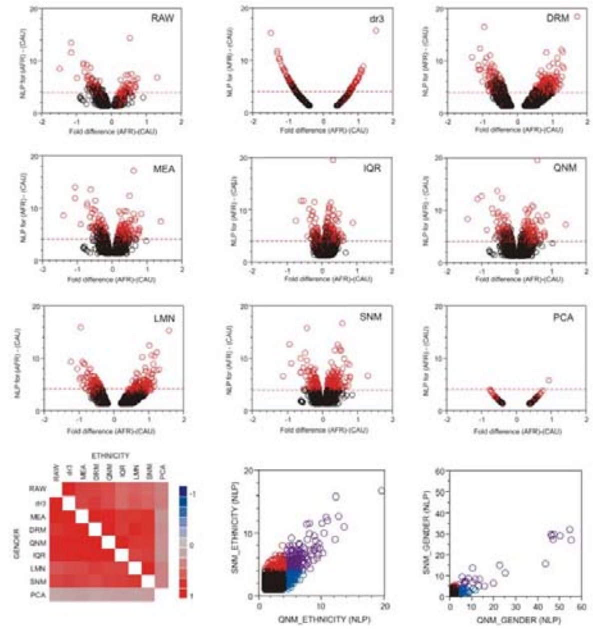
Fitting 16 PC (SV) removes most of the covariance (bottom right)

The SNM model has not accounted for a large source of covariance



Qin et al, 2013. *Frontiers in Genetics* **3**: 160 "Effect of normalization on statistical and biological interpretation of gene expression profiles"

## Effect of Normalization on Significance of Differential Expression



## The special need to adjust for CPM in RNASeq data

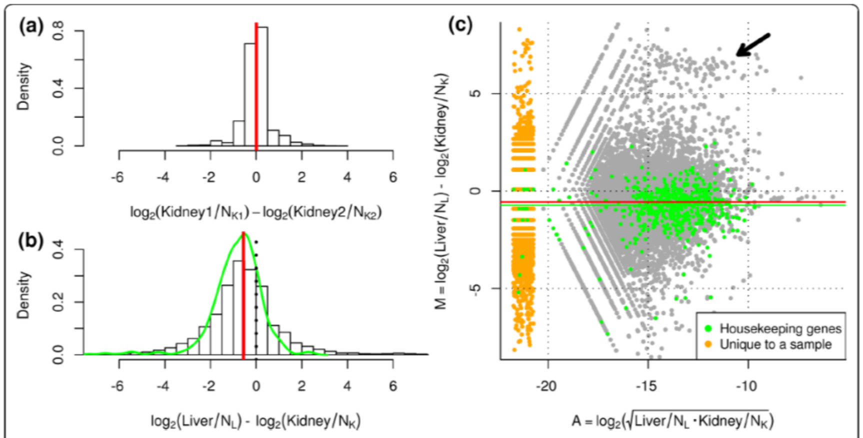


**Figure 1 Normalization is required for RNA-seq data**. Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping genes. **(c)** An M versus A plot comparing liver and kidney shows a clear offset from zero. Green points indicate 545 housekeeping genes, while the green line signifies the median log-ratio of the housekeeping genes. The red line shows the estimated TMM normalization factor. The smear of orange points highlights the genes that were observed in only one of the liver or kidney tissues. The black arrow highlights the set of prominent genes that are largely attributable for the overall bias in log-fold-changes.

TMM =

Trimmed Mean
of the M-values

Robinson and Oshlack, 2010. *Genome Biol* **11**: R25 "A scaling normalization method for differential expression analysis of RNA-seq data"

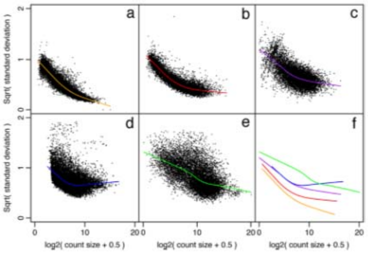
## How to perform DE analysis on the normalized data

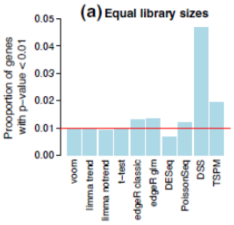
1. Treat it the same as microarray data – use limma or GLM to fit gene specific models
   assuming common variance (not advised for RNASeq)

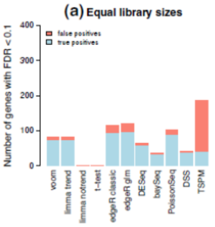
2. Include the identified SV as terms in the limma or GLM models –
   loses the ability to control variance drawing info across probes

3. Convert the normalized values back to cpm scale and analyze in EdgeR or DEseq2
   but I have not seen this done in the literature

4. Output the normalized dataset to VOOM, which estimates the mean-variance relationship from the data
   rather than assuming a negative binomial, uses this in linear models







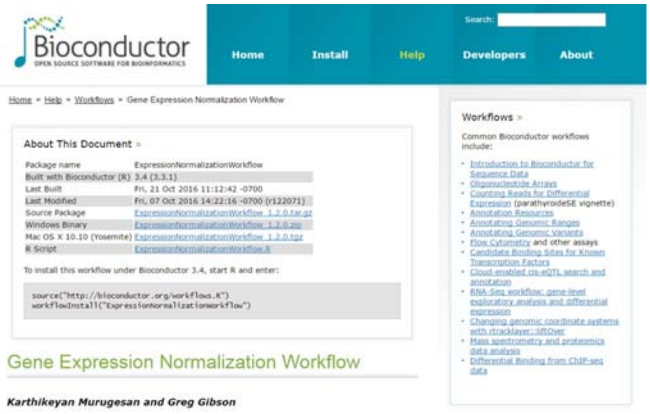
Law et al, 2014. *Genome Biol* **15**: R29 "voom: precision weights unlock linear model analysis tools for RNA-seq read counts"

## Our Standard Analytical strategy

1. Normalize the samples
2. Extract the Principal components of gene expression
3. Ask whether the major PC are correlated with technical covariates such as Batch or RNA quality
4. If they are, renormalize to remove those effects
5. As much as possible, analyze the dataset in several different ways to (i) confirm that the findings are not sensitive to your analytical choice, and (ii) gain insight into what may cause differences, eg find confounding factors

## An Expression Workflow in Bioconductor

https://www.bioconductor.org/help/workflows/ExpressionNormalizationWorkflow/



Home » Help » Workflows » Gene Expression Normalization Workflow

About This Document »

| | |
|---|---|
| Package name | ExpressionNormalizationWorkflow |
| Built with Bioconductor (R) | 3.4 (3.3.1) |
| Last Built | Fri, 21 Oct 2016 11:12:42 -0700 |
| Last Modified | Fri, 07 Oct 2016 14:22:16 -0700 (r122071) |
| Source Package | ExpressionNormalizationWorkflow_1.2.0.tar.gz |
| Windows Binary | ExpressionNormalizationWorkflow_1.2.0.zip |
| Mac OS X 10.10 (Yosemite) | ExpressionNormalizationWorkflow_1.2.0.tgz |
| R Script | ExpressionNormalizationWorkflow.R |

To install this workflow under Bioconductor 3.4, start R and enter:

```
source("http://bioconductor.org/workflows.R")
workflowInstall("ExpressionNormalizationWorkflow")
```

Gene Expression Normalization Workflow

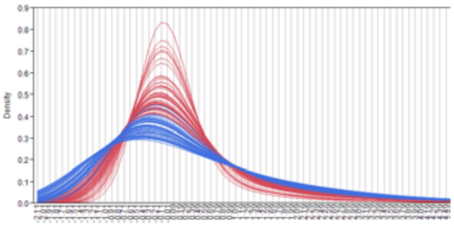
***Karthikeyan Murugesan and Greg Gibson***

Workflows »

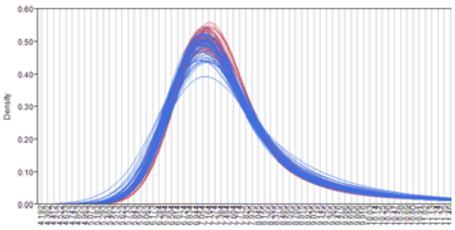
Common Bioconductor workflows include:

- Introduction to Bioconductor for Sequence Data
- Oligonucleotide Arrays
- Counting Reads for Differential Expression (parathyroideSE vignette)
- Annotation Resources
- Annotating Genomic Ranges
- Annotating Genomic Variants
- Flow Cytometry and other assays
- Candidate Binding Sites for Known Transcription Factors
- Cloud-enabled cis-eQTL search and annotation
- RNA-Seq workflow: gene-level exploratory analysis and differential expression
- Changing genomic coordinate systems with rtracklayer::liftOver
- Mass spectrometry and proteomics data analysis
- Differential Binding from ChIP-seq data

BEFORE SNM



AFTER SNM


Density