# Before we start!

Please run these commands if you want to follow along today:

**On your terminal:**

```
scp -r USERNAME@info.mcmaster.ca:~/../gradstd6/Data_Int_Files ~/Desktop
```

**In Rstudio:**

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("Rbowtie2")
BiocManager::install("Rsamtools")
```

# Analysis and integration of RNA-seq and ChIP-seq data

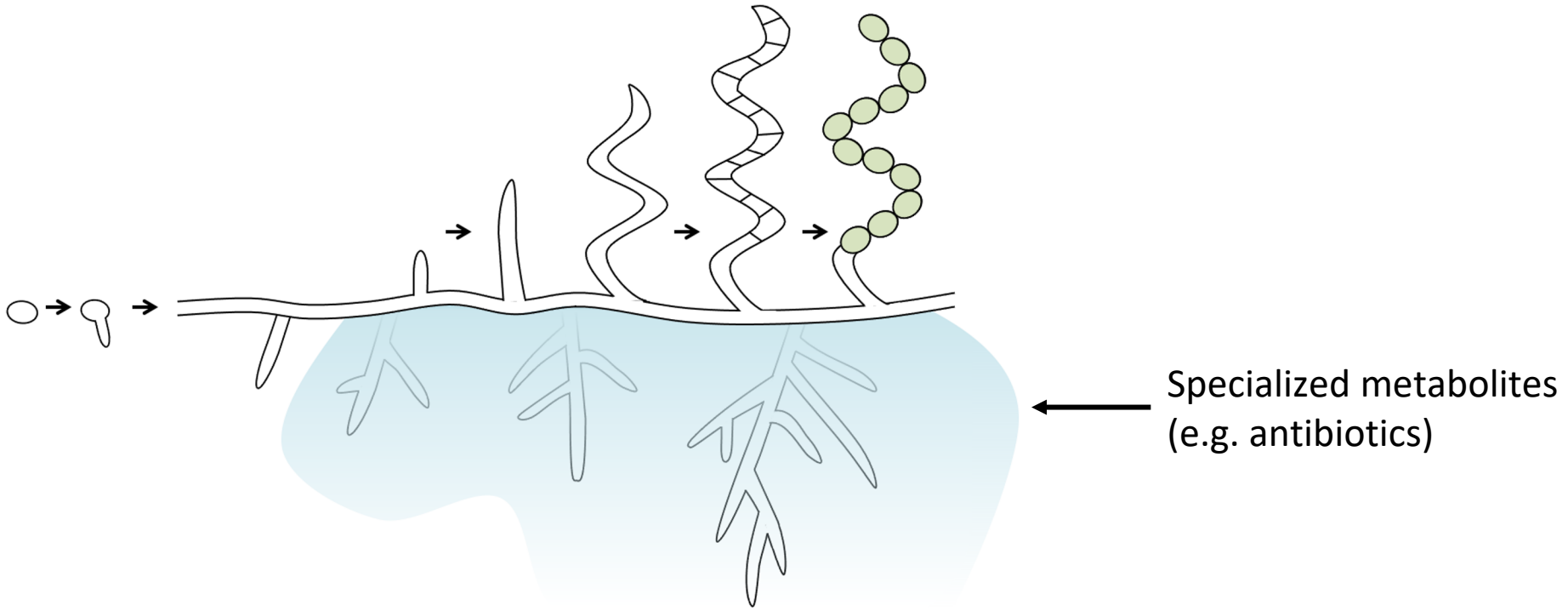
Meagan Archer, Meghan Pepler, Sreedevi Kesavan, & Stephanie Ali Fairbairn

# Outline

- **Introduction**
  - Our model system
  - Packages we use and how they work
- **Workshop #1**
  - Analyzing ChIP sequencing data
- **Quality-check ChIP data**
  - Presentation on ChIPQC
- **Workshop #2**
  - Some data integration
- **Conclusions of our integration**

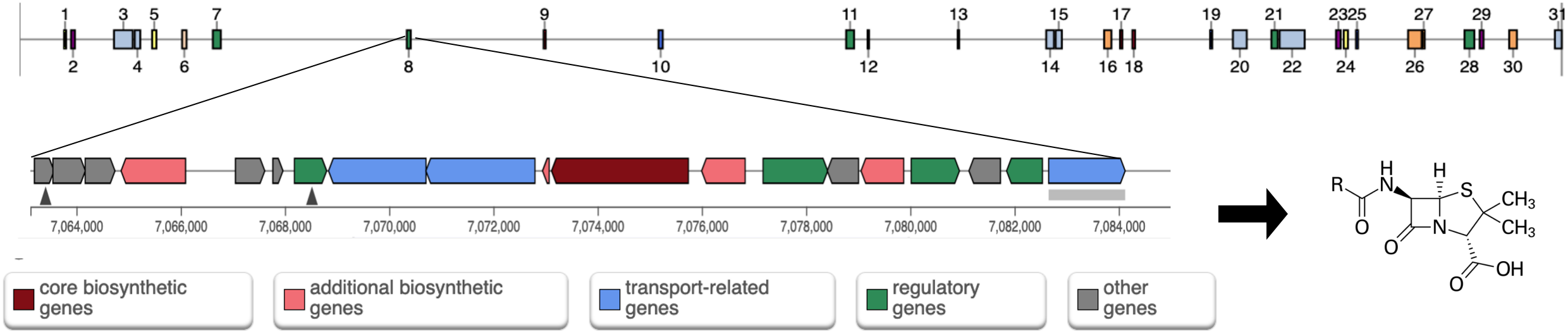
# Introduction to model system: *Streptomyces*

- *Streptomyces* are soil-dwelling Gram-positive bacteria
- Complex multi-stage lifecycle and secondary metabolism



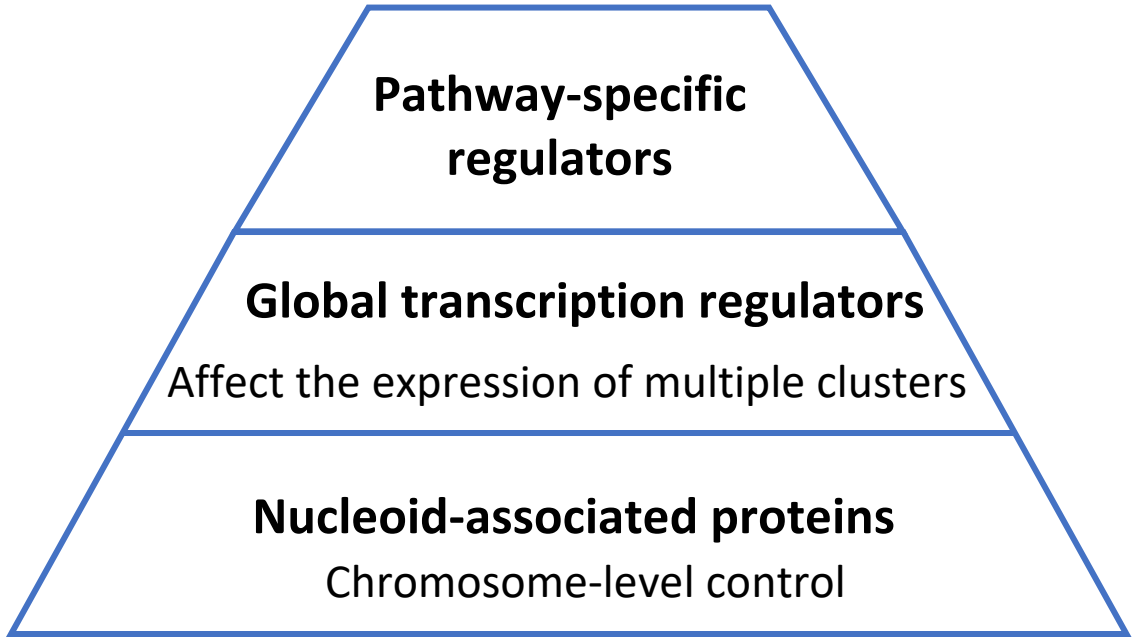
# Biosynthetic gene clusters

- *Streptomyces* have relatively large linear genomes (8-10 Mbp)
- Antibiotic biosynthesis directed by gene clusters
- Most are not expressed under laboratory conditions



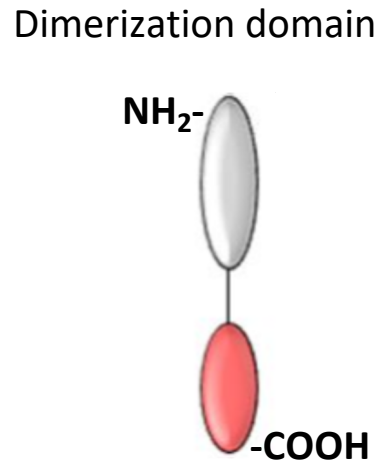
# Regulation of secondary metabolism

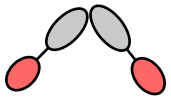
- Secondary metabolism is regulated at many levels
- Nucleoid associated protein: Lsr2



# Lsr2: global regulator?

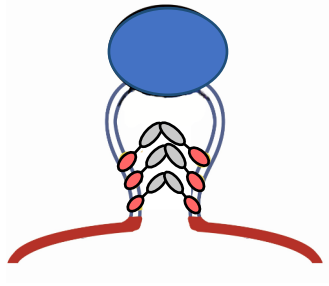
- Absolutely conserved in *Streptomyces*
- DNA-binding and dimerization/polymerization domains
- Binds AT-rich regions of DNA in *Mycobacterium*



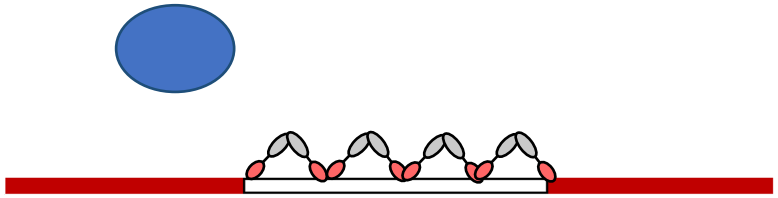



Lsr2

RNA polymerase



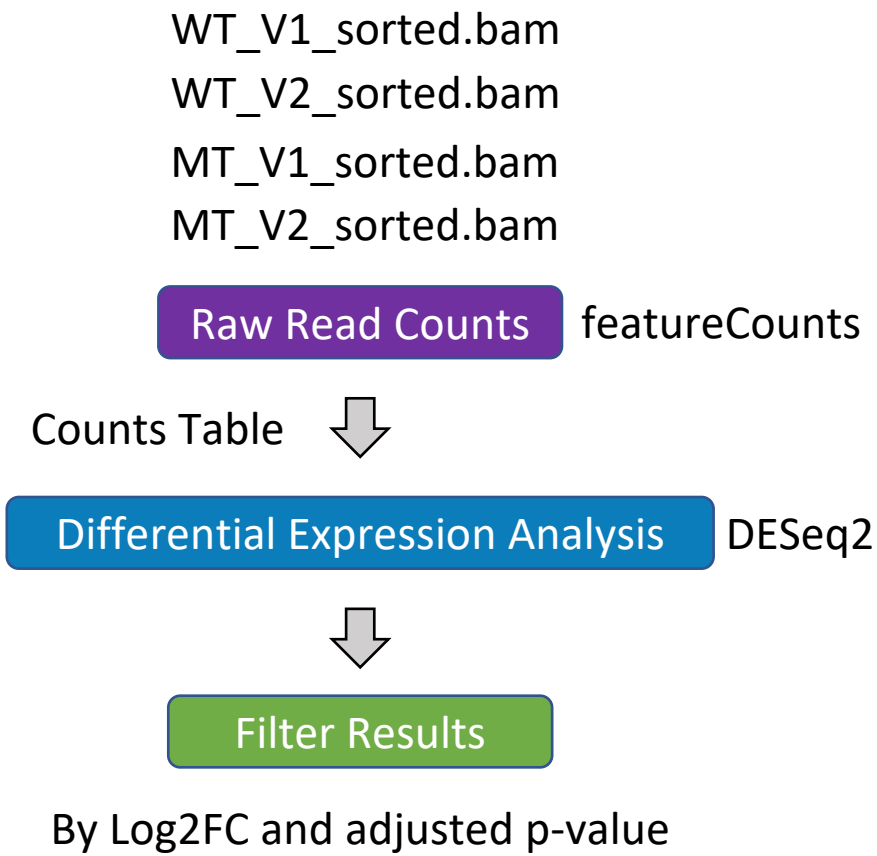
Trapping RNA polymerase



Excluding RNA polymerase at promoter region

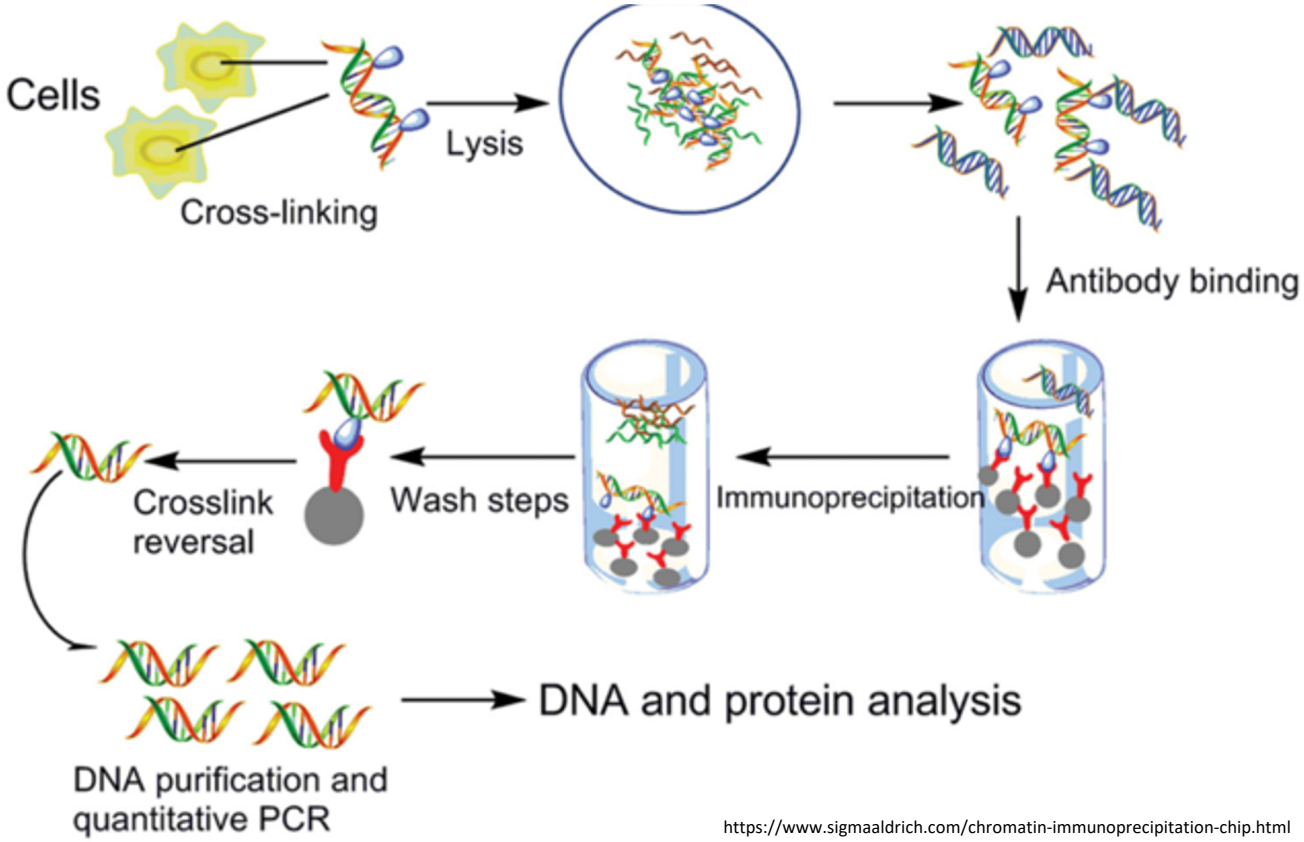
# Experimental design: RNA-seq

- RNA-seq of wild type and Δ*lsr2 Streptomyces venezuelae*
- Overview of data analysis pipeline:



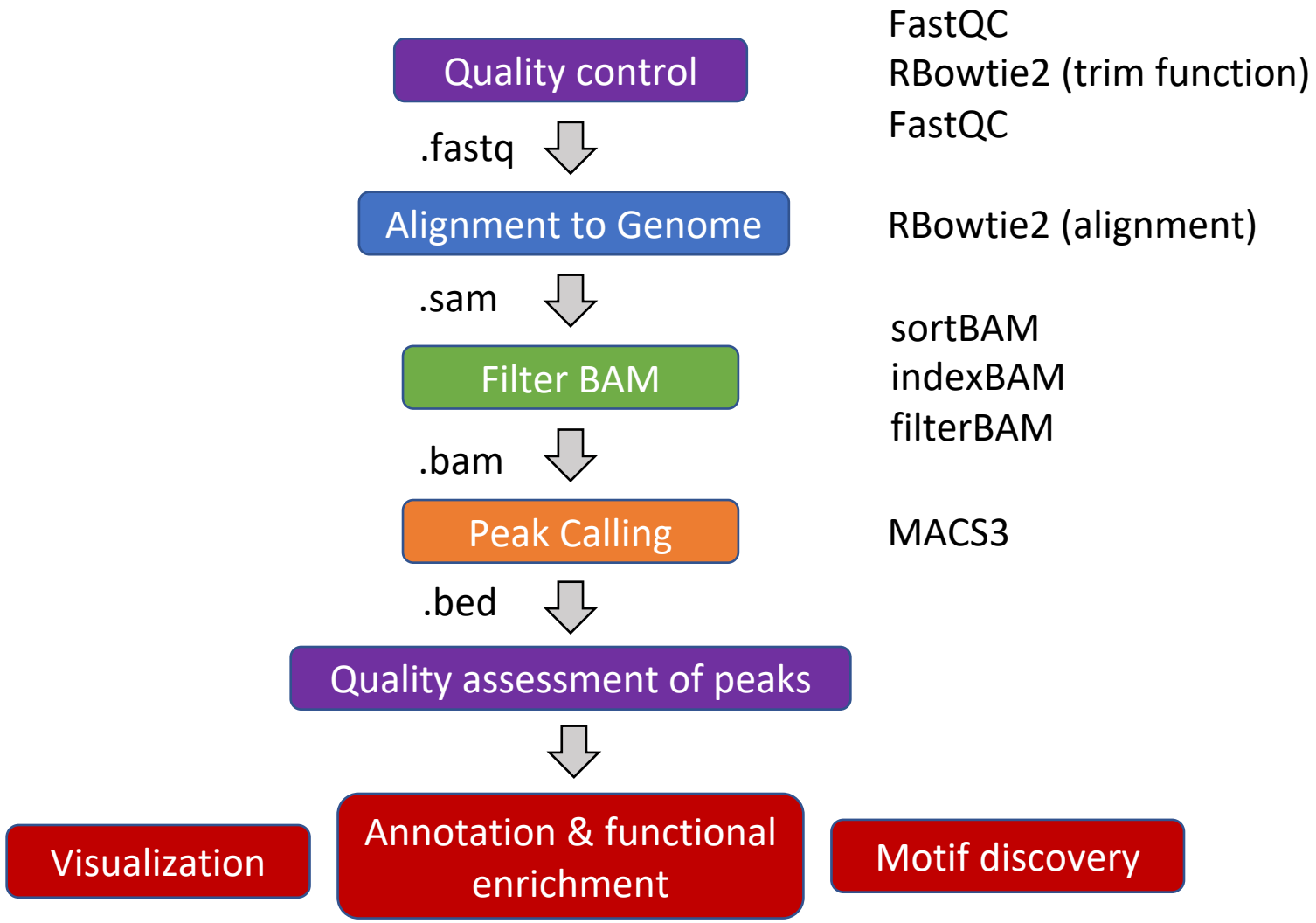
# Experimental design: ChIP-seq

- ChIP-sequencing Lsr2-3XFLAG and Lsr2 unflagged (negative control)




https://www.sigmaaldrich.com/chromatin-immunoprecipitation-chip.html

# ChIP data analysis workflow




# BOWTIE 2: ultrafast, memory-efficient alignment tool

- One of the most popular aligners due to its speed and low memory requirement.
- Accepts FASTQ and FASTA as input files.
- Uses an FM index based on a Burrows–Wheeler transform method
- Good for reads >50 bases
- Supports local, end-to-end and gapped alignment.
- Outputs a set of alignments in SAM format.
- Scores: higher = more similar
- Mapping quality: higher = more unique

# Rsamtools (sort/index/filter)

- Converts SAM output files to BAM files.
- SAM-Sequence Alignment/Map - Human readable text files
- BAM- Binary Alignment/Map
- Sorts BAM files such that the alignments occur in “genome order.”
- Creates a BAM index file (.bai) to allow easy access to overlapping regions.
- Used to filter out unmapped reads in BAM files or create subsets of the original file.
- BAM files can be re-converted to SAM files for easy viewing

# Peak Calling

- Identifies areas in a genome enriched with aligned reads
  - areas where a protein interacts with DNA

- Compares distribution of groups formed against background.

- Two types of enrichment:
  - Broad peak (histone modifications)
  -Narrow peak (transcription factor binding)

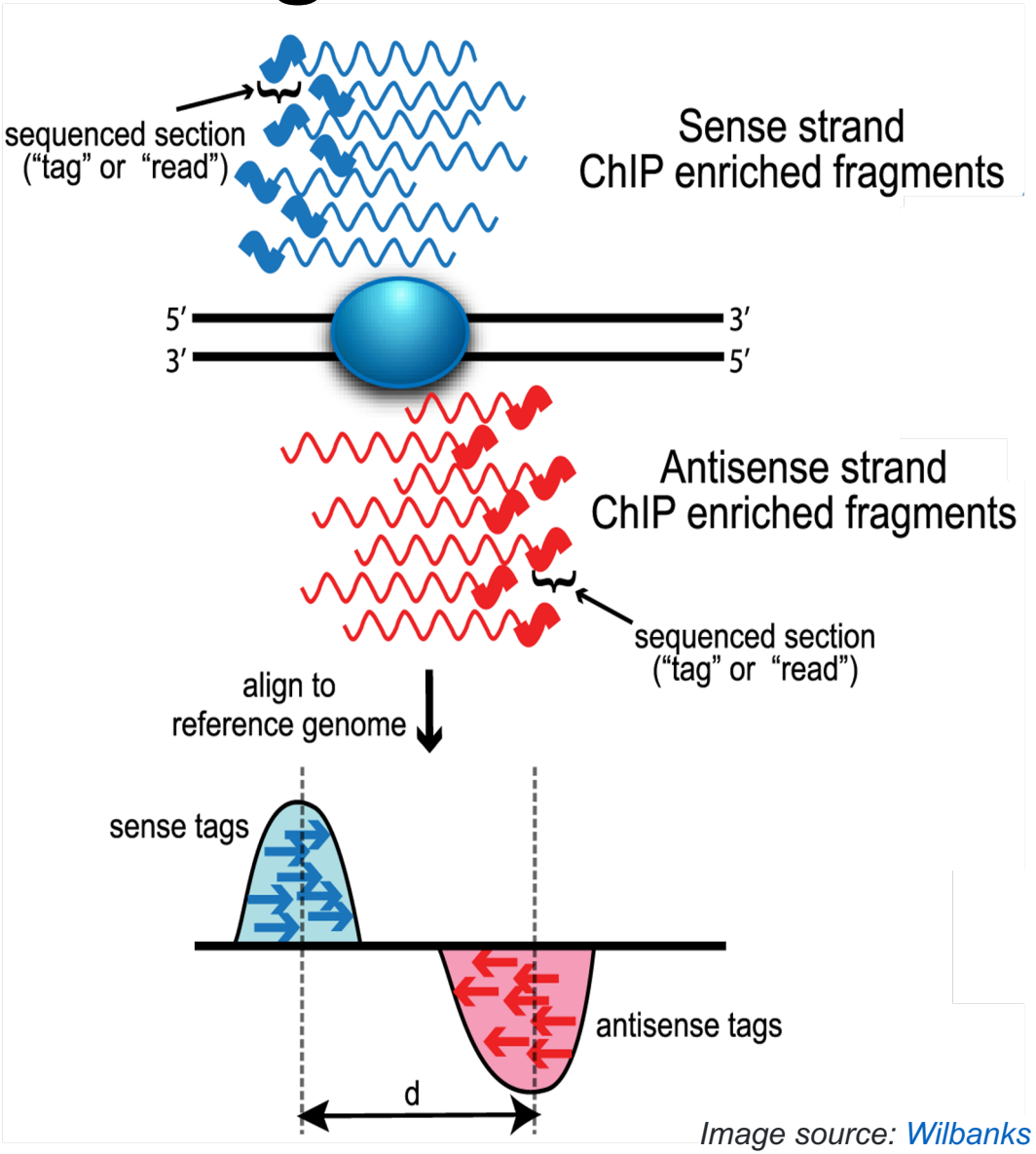



*Image source: Wilbanks and Facciotti, PLoS One 2010*

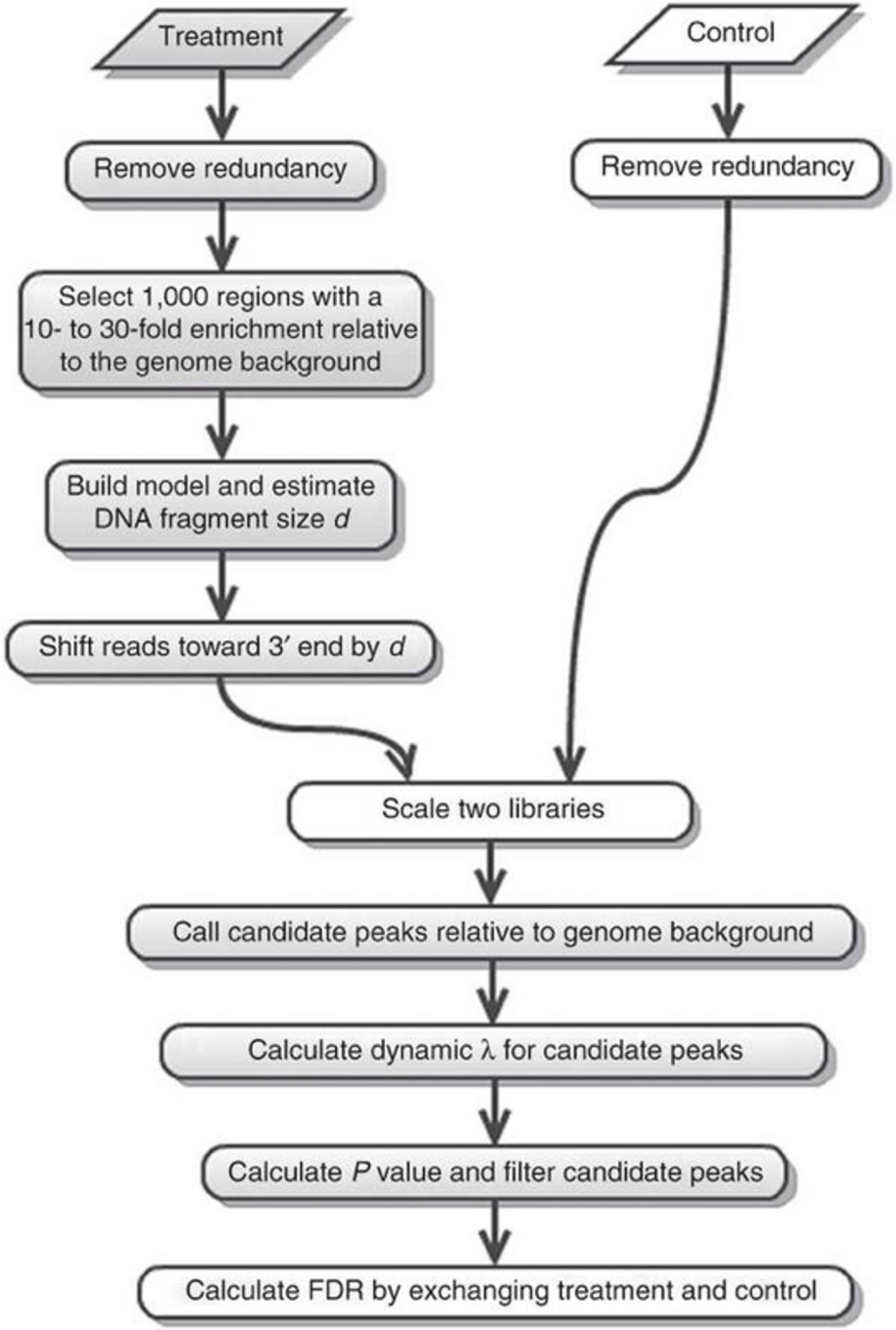
# MACS3

## *Model-based Analysis of ChIP-Seq*

- Developed for the detection of transcription factor binding sites.
- Main function is to call peaks.
- Evaluates the significance of enriched ChIP regions.
- Can be used with ChIP sample alone or with a control.
- Considers both sequencing tag position and orientation.

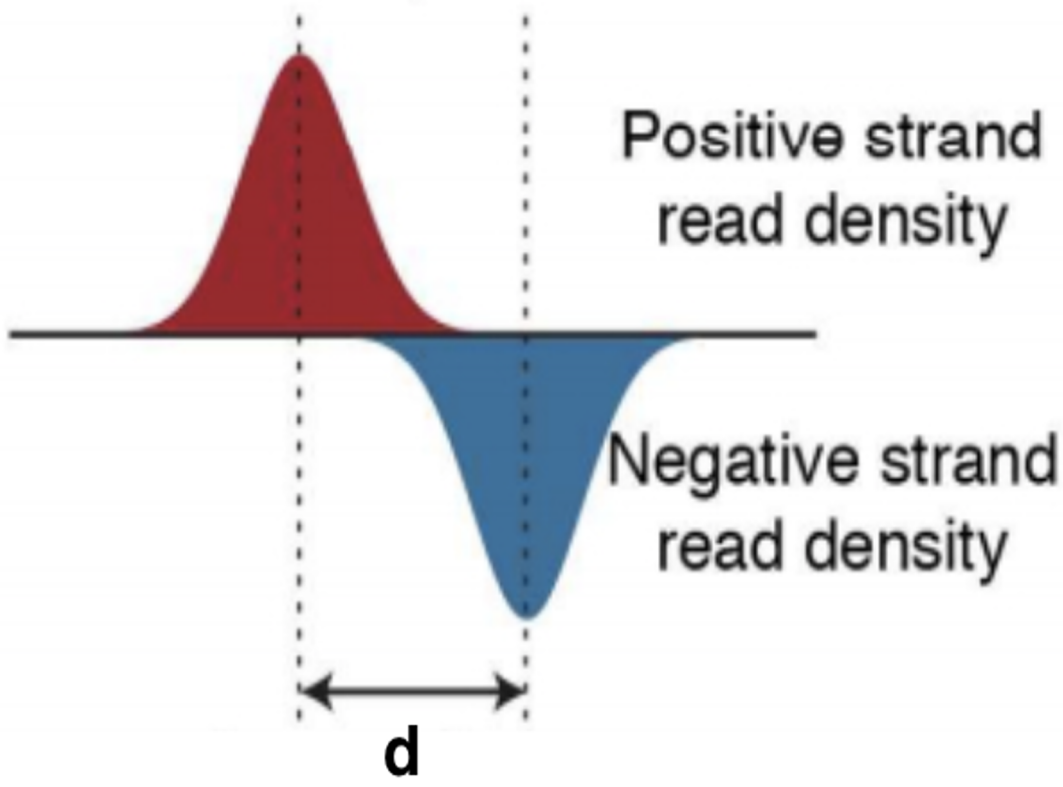
# MACS3- Workflow

- Reads with the same start position considered duplicates
- Good or bad duplicates?
- Bona fide peaks will have multiple overlapping reads with offsets
- True binding sites show bimodal enrichment pattern (paired peaks)
- Paired peaks from ChIP sample used to build a model



# MACS3-Modeling the shift size

- Random selection of 1000 high-quality peaks
- Distance between the modes of the two peaks is defined as 'd'
- All tags shifted by d/2 to get most likely interaction site
- Peaks are detected using a window size of 2d
- Poisson distribution used to calculate significant enrichment
- Fold enrichment calculated using tag ratio and λlocal



# MACS3- Output Files

- R script showing peak model- _model.R
- Excel file- _peaks.xls
- Bed files- _summits.bed
- Narrow peak files- _.peaks.narrowPeak

# Workshop time!

Analysis and quality check ChIP-seq data

# Quality check: ChIPQC

- Bioconductor package that takes the .bam files and peak calls to automatically compute quality metrics
- To run the analysis you need to make a sample sheet with specific headers:
  - **SampleID**: Identifier string for sample
  - **Tissue, Factor, Condition**: Identifier strings for up to three different factors (You will need to have all columns listed. If you don't have infomation, then set values to NA)
  - **Replicate**: Replicate number of sample
  - **bamReads**: file path for BAM file containing aligned reads for ChIP sample
  - **ControlID**: an identifier string for the control sample
  - **bamControl**: file path for bam file containing aligned reads for control sample
  - **Peaks**: path for file containing peaks for sample
  - **PeakCaller**: Identifier string for peak caller used. Possible values include “raw”, “bed”, “narrow”, “macs”

# ChIPQC report

## Mapping, Filtering and Duplication rate

This section presents the mapping quality, duplication rate and distribution of reads in known genomic features.

**Table 2.** Number and percantage of mapped,duplicated and MapQ filter passing reads

| ID | Tissue | Factor | Condition | Replicate | Unmapped | Mapped | Pass MapQ Filter and Dup | Total Dup% | Pass MapQ Filter% | Pass MapQ Filter and Dup% |
|---|---|---|---|---|---|---|---|---|---|---|
| FLAG1 | | | | 1 | 0 | 2493808 | 0 | 0 | 78 | 0 |
| FLAG2 | | | | 2 | 0 | 2294689 | 0 | 0 | 78 | 0 |
| UNFLAG1 | NA | NA | NA | NA | 0 | 2223632 | 0 | 0 | 95 | 0 |
| UNFLAG2 | NA | NA | NA | NA | 0 | 3516903 | 0 | 0 | 95 | 0 |

**Table 2** shows the absolute number of total, mapped, passing MapQ filter and duplicated reads. The percent of mapped reads passing quality filter and marked as duplicates (Non-Redundant Fraction?) are also included.

Description of read filtering and flag metrics:

- **Total Dup%**-Percentage of all **mapped** reads which are marked as **duplicates.**
- **Pass MapQ Filter%**-Percentage of all **mapped** reads which**pass MapQ quality** filter
- **Pass MapQ Filter and Dup%**-Percentage of all reads which pass **MapQ filter** and are marked as**duplicates.**

# ChIPQC report: RiP (reads in peaks)

- Represents the percentage of reads that overlap 'called peaks'

**Figure 5.** Density plot of the number of reads in peaks

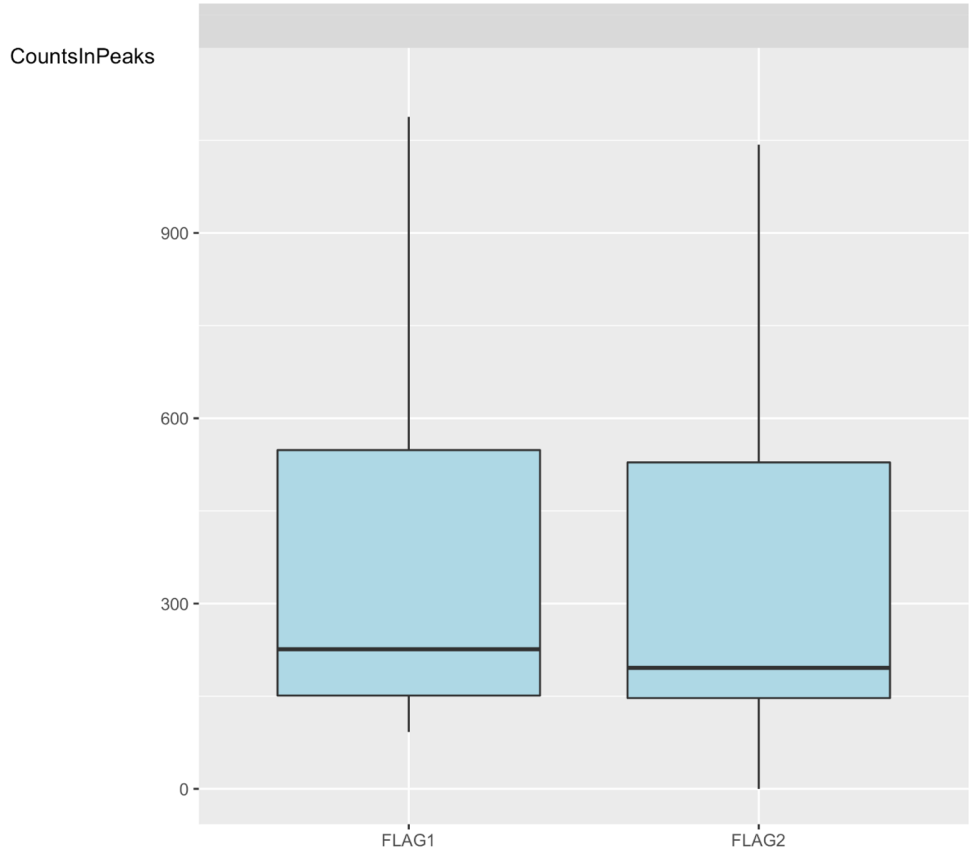
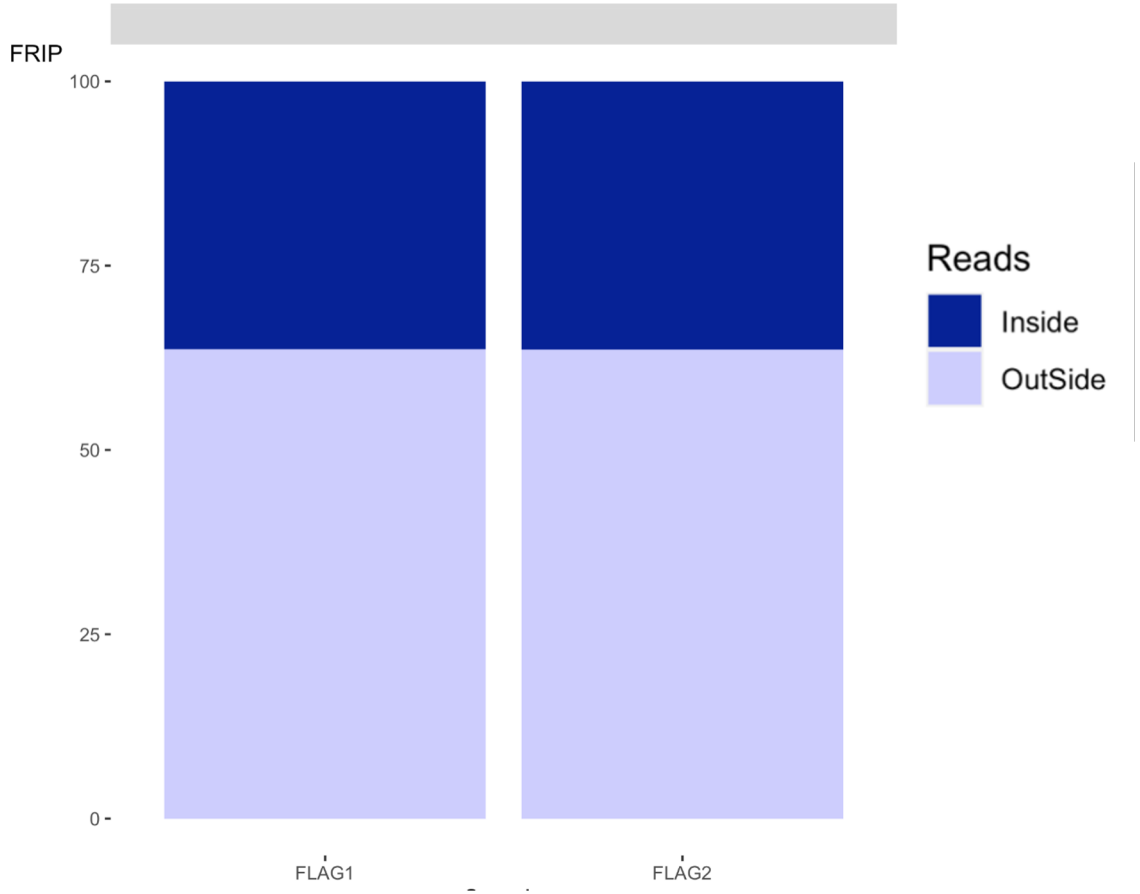



**Figure 4.** Barplot of the percentage number of reads in peaks

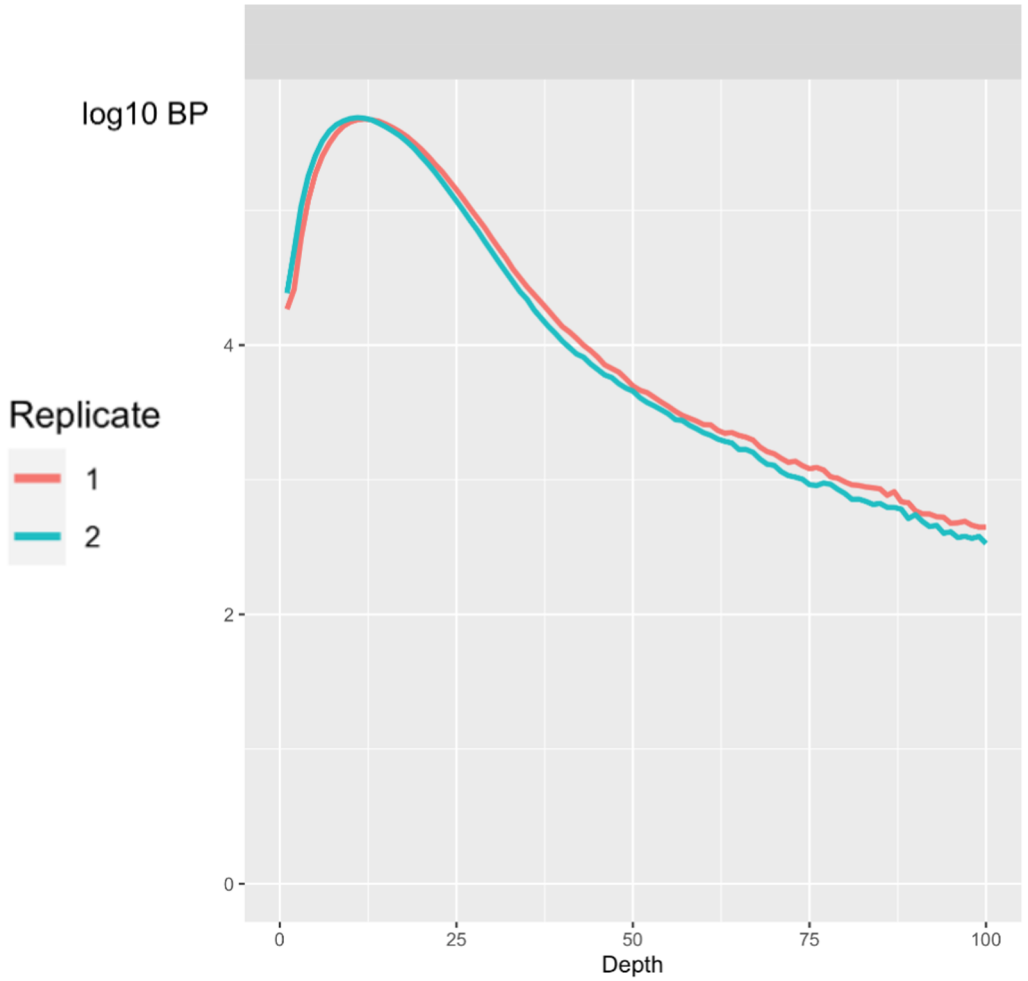


# ChIPQC report: SSD

- Represents the **uniformity of coverage of reads** across the genome

- In samples with low enrichment, the reads will have most positions (high values on y-axis) in the genome with low pile-up (low x-axis values)

**Figure 1.** Plot of the log2 base pairs of genome at differing read depths



# ChIPQC report: Peak signal strength

**Table 1.** Summary of ChIP-seq filtering and quality metrics.

| ID | Tissue | Factor | Condition | Replicate | Reads | Dup% | ReadL | FragL | RelCC | SSD | RiP% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FLAG1 | | | | 1 | 2493808 | 0 | 76 | 154 | 0.99 | 110 | 36 |
| FLAG2 | | | | 2 | 2294689 | 0 | 76 | 187 | 0.91 | 100 | 36 |
| UNFLAG1 | NA | NA | NA | NA | 2223632 | 0 | 76 | 153 | 0.86 | 19 | NA |
| UNFLAG2 | NA | NA | NA | NA | 3516903 | 0 | 76 | 153 | 0.87 | 25 | NA |

- At least SSD is larger in FLAG than control
- A "good" or enriched sample typically has regions of significant read pile-up (larger differences in coverage) so a higher SSD is more indicative of better enrichment

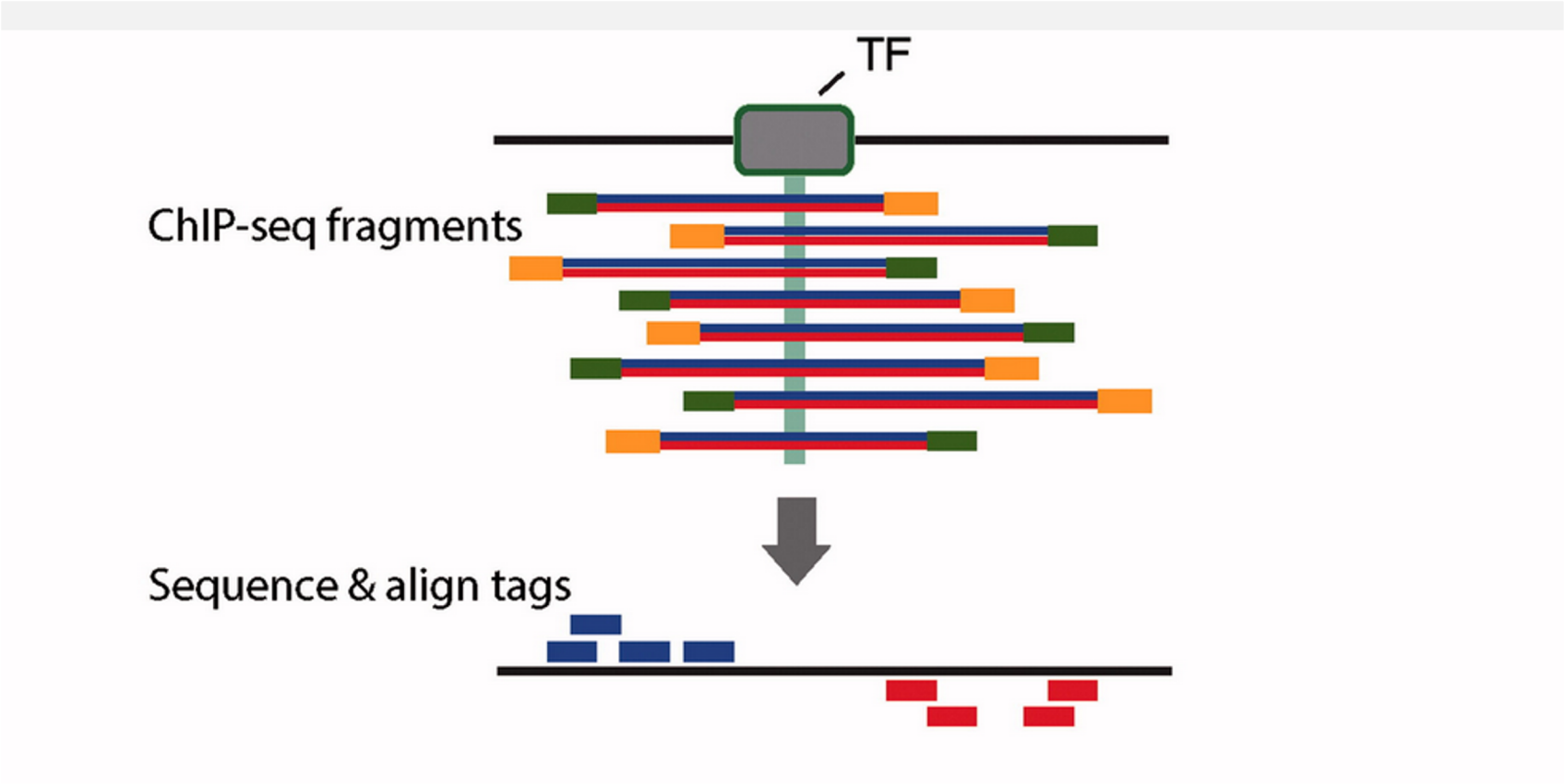
# ChIPQC report: Peak signal strength

**Table 1.** Summary of ChIP-seq filtering and quality metrics.

| ID | Tissue | Factor | Condition | Replicate | Reads | Dup% | ReadL | FragL | RelCC | SSD | RiP% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FLAG1 | | | | 1 | 2493808 | 0 | 76 | 154 | 0.99 | 110 | 36 |
| FLAG2 | | | | 2 | 2294689 | 0 | 76 | 187 | 0.91 | 100 | 36 |
| UNFLAG1 | NA | NA | NA | NA | 2223632 | 0 | 76 | 153 | 0.86 | 19 | NA |
| UNFLAG2 | NA | NA | NA | NA | 3516903 | 0 | 76 | 153 | 0.87 | 25 | NA |

- FragL values should be similar between samples
- RelCC (relative strand cross-correlation coefficient) values larger than 1 would suggest good signal-to-noise
  - Our data is good but not great in that sense

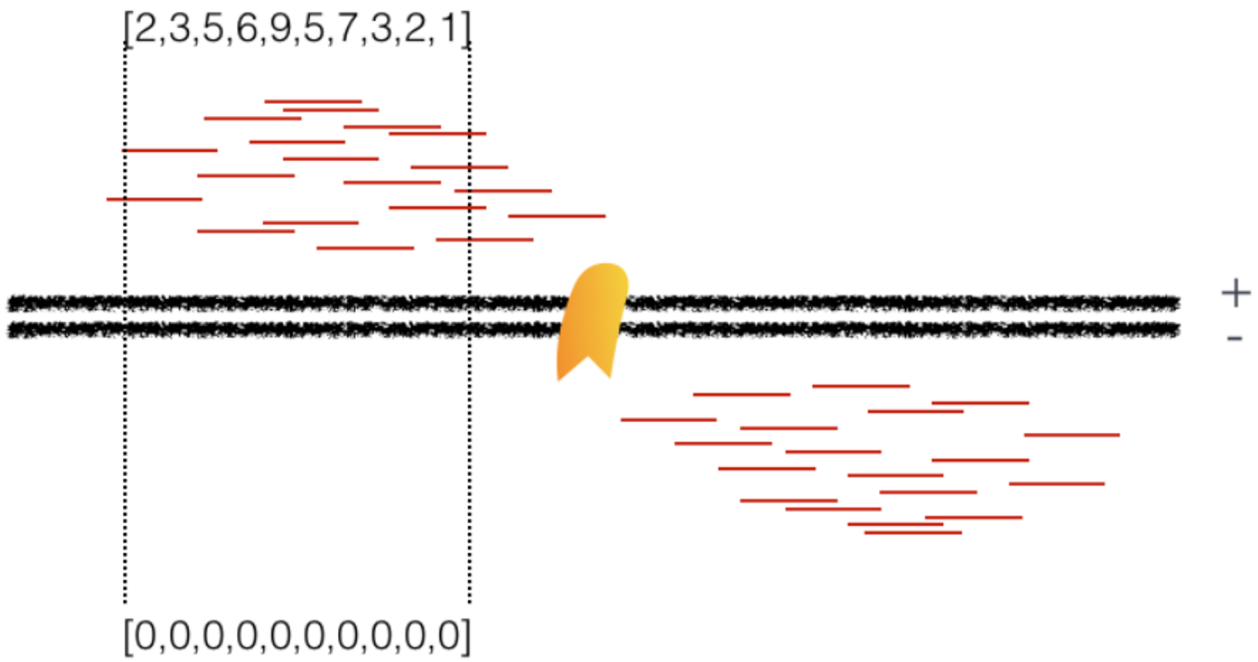
# ChIPQC report: Strand cross-correlation



# ChIPQC report: Strand cross-correlation

How are they generated?

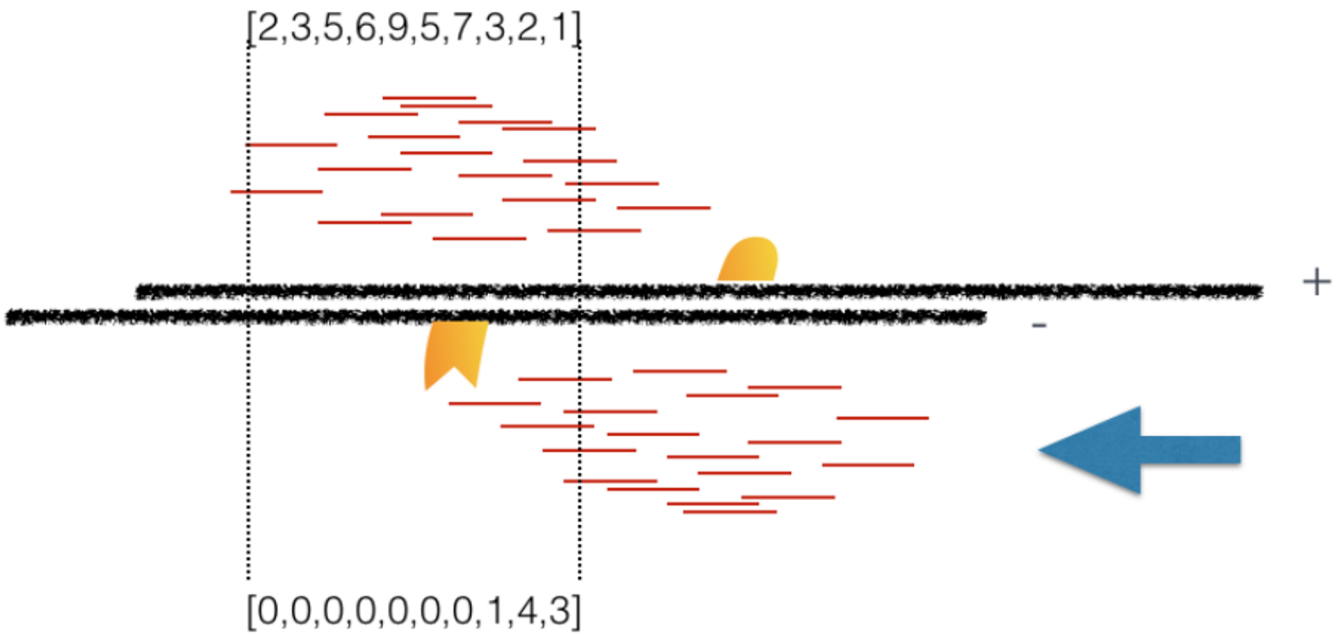
***Plot 1:*** *At strand shift of zero, the Pearson correlation between the two vectors is 0.*



# ChIPQC report: Strand cross-correlation

How are they generated?

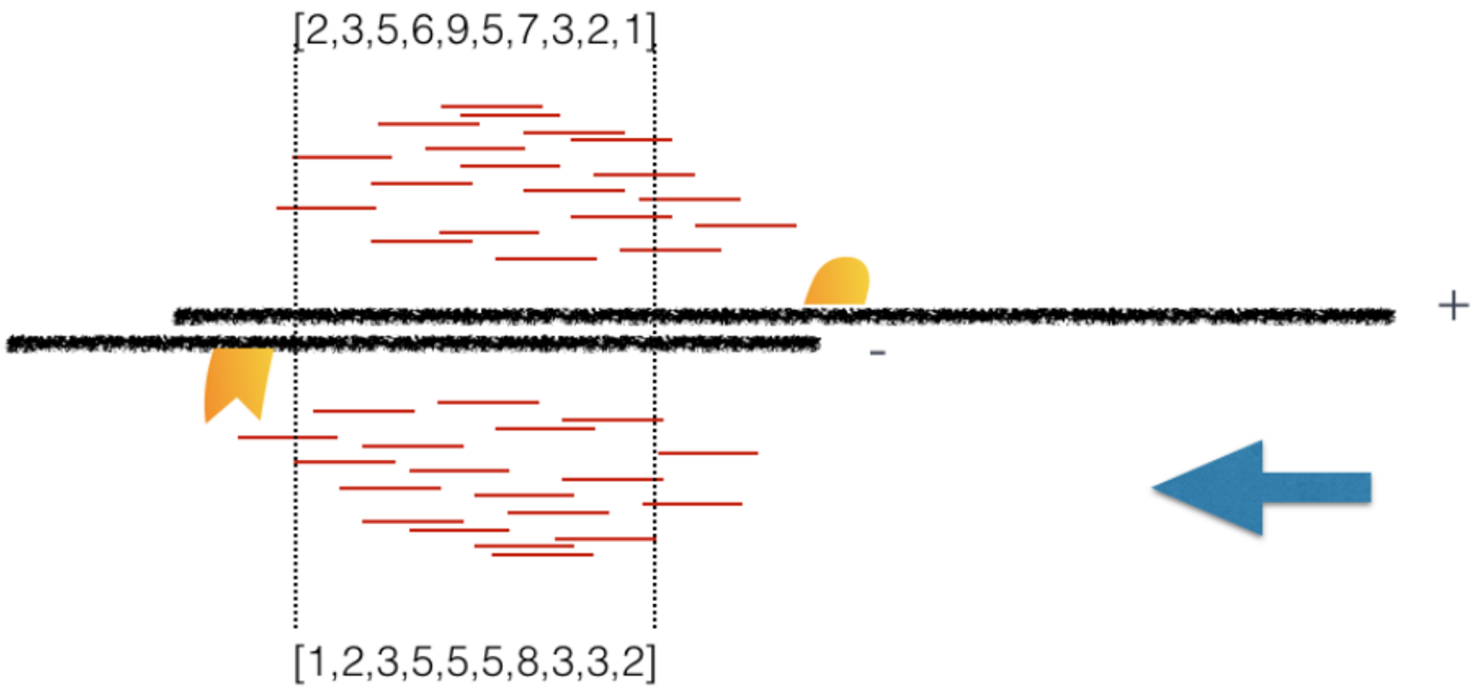
***Plot 2:*** *At strand shift of 100bp, the Pearson correlation between the two vectors is 0.389.*



# ChIPQC report: Strand cross-correlation

How are they generated?

***Plot 3:*** *At strand shift of 175bp, the Pearson correlation between the two vectors is 0.831.*



# ChIPQC report: Strand cross-correlation

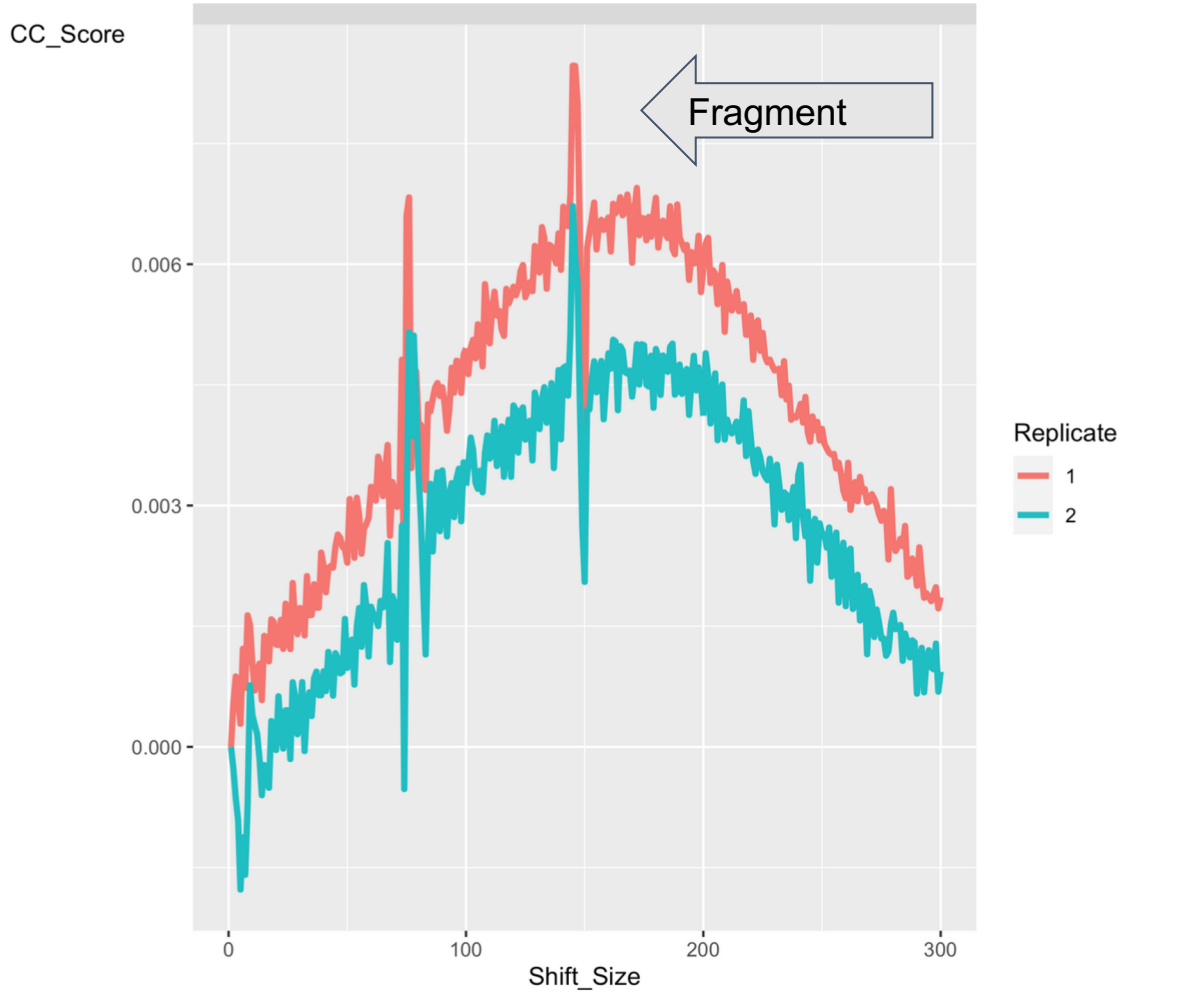
- When this process is completed we will have a table of values mapping each base pair shift to a Pearson correlation value
- These values are computed for every peak for each chromosome
  - values are multiplied by a scaling factor and then summed across all chromosomes
- Once the final cross-correlation values have been calculated, they can be plotted (Y-axis) against the shift value (X-axis) to generate a cross-correlation plot!

# ChIPQC report: Strand cross-correlation

The cross-correlation plot typically produces two peaks:

- a peak of enrichment corresponding to the predominant **fragment length** (highest correlation value)
- a peak corresponding to the **read length** ("phantom" peak)

**Figure 2.** Plot of CrossCoverage score after successive strand shifts



# Integrating CHIP-seq and RNA-seq

## Why?

- To further characterize gene regulation, ie. the interactions between protein-binding and gene expression

## What can you do?

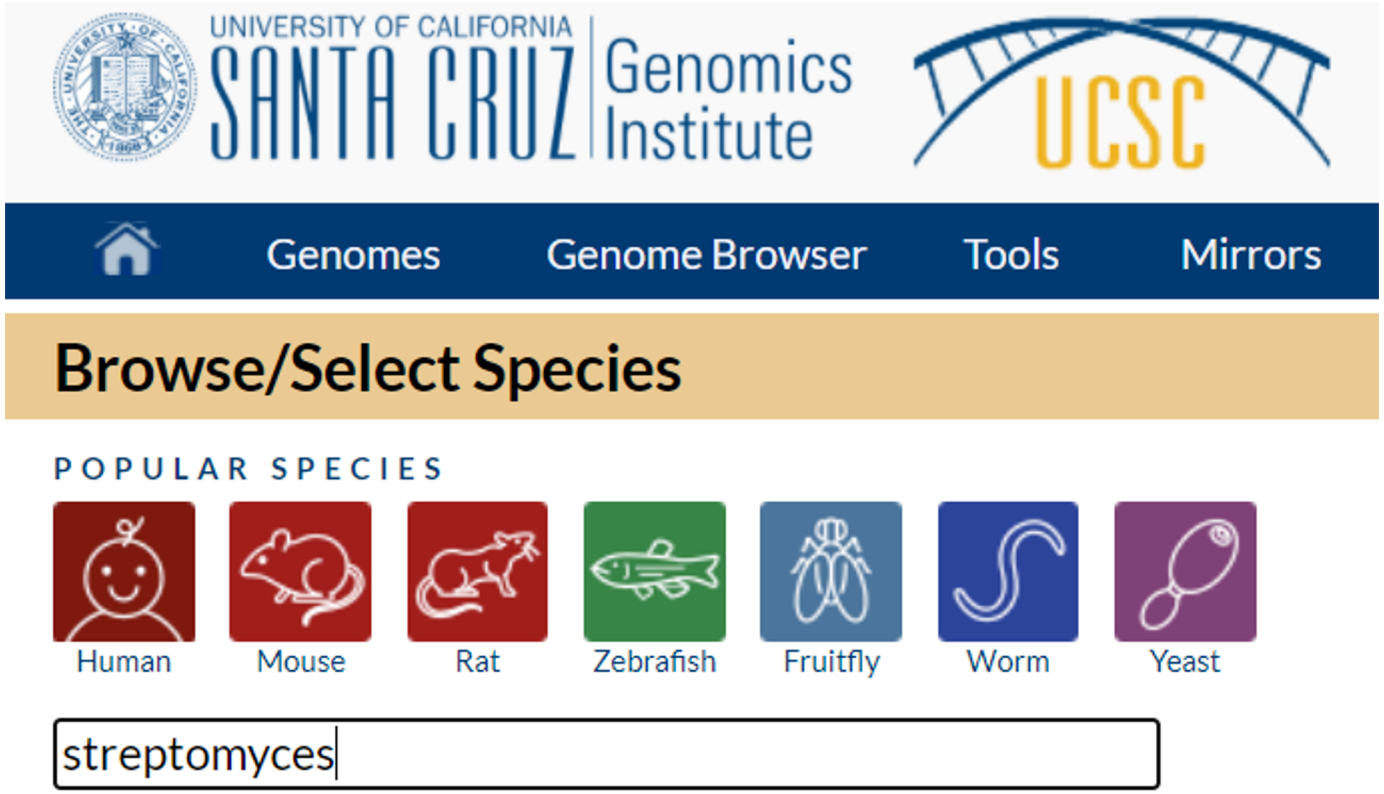
- Find significantly differentially expressed genes that overlap in CHIP-seq and RNA-seq outputs
- What you do next depends on what your research question is focussed on:
  - Identifying target genes, co-regulators or epigenetic cofactors
  - Predicting transcription factor binding or gene expression

(Höllbacher *et al*., 2020)

# A Cautionary Tale - *Streptomyces*

- There are limitations with working with less common organisms (and prokaryotes)
- PubMed query: rna-sequencing AND chip-sequencing AND "bacteria"
  - Returned **19** results relevant to our work (Choudhary *et al*., 2020; Mahmud *et al*., 2020; Bogue *et al*., 2020; DuPai *et al*., 2020; Lee *et al*., 2020; Hurst-Hess *et al*., 2019; Rioualen *et al*., 2019; Kroner *et al*., 2019; Jaskólska *et al*., 2018; Shao *et al*., 2018; Fishman *et al*., 2018, Pan *et al.,* 2018; Park *et al*., 2017; Markel *et al.,* 2016; Schulz *et al*., 2015, Lam *et al*., 2014; Balasubramanian *et al.,* 2014; Uplekar *et al*., 2013 Jutras *et al*., 2012)
- Tools, such as BETA (Binding and Expression Target Analysis), are ill-adapted to work with prokaryotic data

# A Cautionary Tale - *Streptomyces*



# (1) Identifying Target Genes

**(1)** Contingency tables and Fisher's test

**(2)** Correlation between log2fold change in expression and peak intensity

**(3)** BETA software package

(Höllbacher *et al*., 2020)

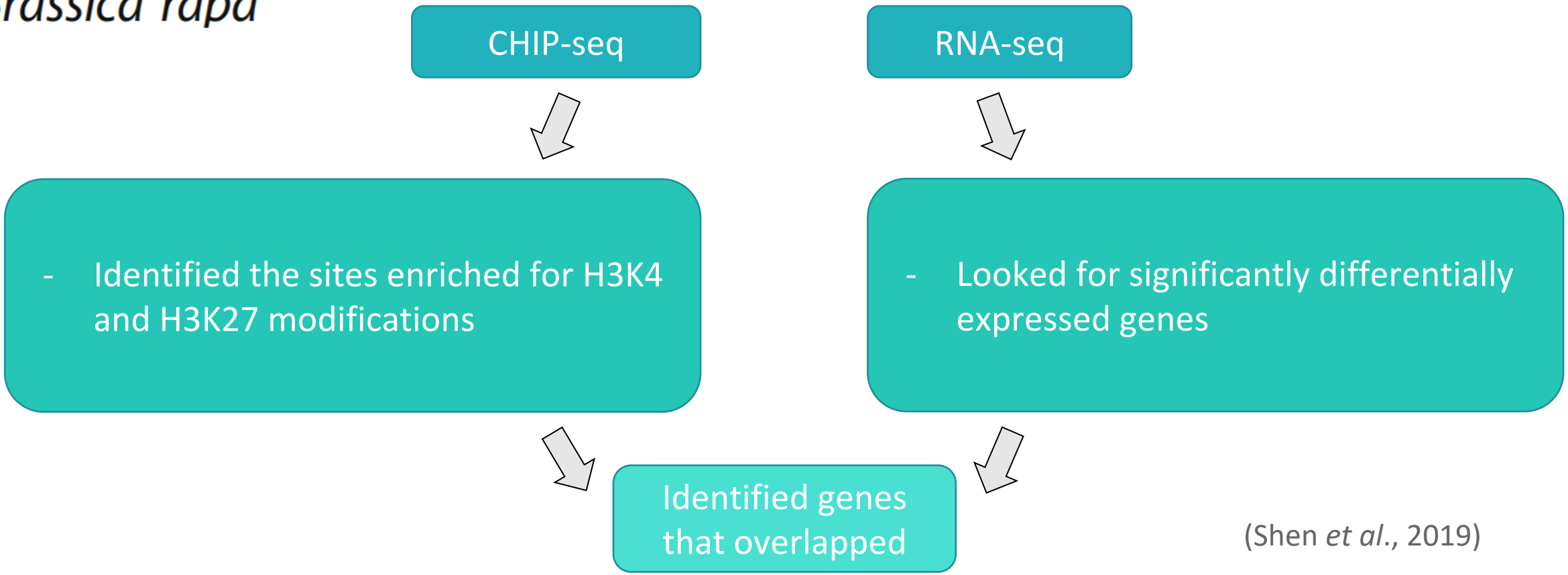
# (2) Predicting Gene Expression

- Chromatin and **histone modifications** are accepted predictors of gene expression
- Active area of debate: Can transcription factor binding strength be used to predict gene expression levels?

(Höllbacher *et al*., 2020)

RESEARCH ARTICLE Open Access

# Comparative transcriptome analysis and ChIP-sequencing reveals stage-specific gene expression and regulation profiles associated with pollen wall formation in *Brassica rapa*

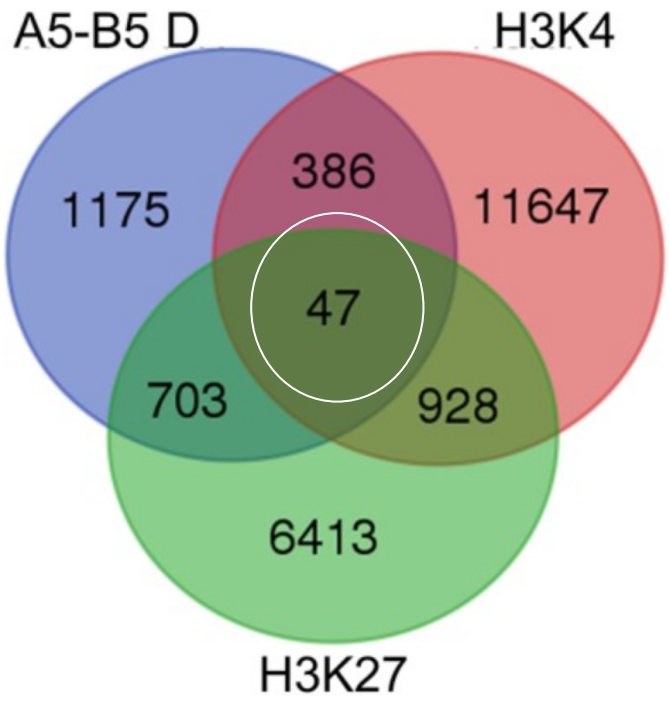



(Shen *et al.*, 2019)

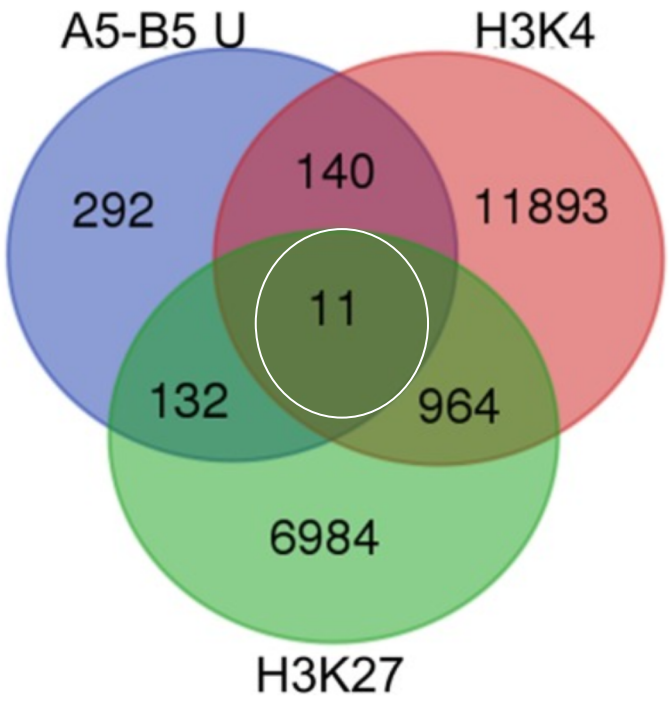
RESEARCH ARTICLE Open Access

# Comparative transcriptome analysis and ChIP-sequencing reveals stage-specific gene expression and regulation profiles associated with pollen wall formation in *Brassica rapa*

## Downregulated




## Upregulated




A5-B5 = mature pollen stage

(Shen *et al.*, 2019)

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Host-Microbe Biology

# Genome-Scale Mapping Reveals Complex Regulatory Activities of RpoN in *Yersinia pseudotuberculosis*

CHIP-seq

- Identified binding sites of RpoN

RNA-seq

- Looked for significantly differentially expressed genes

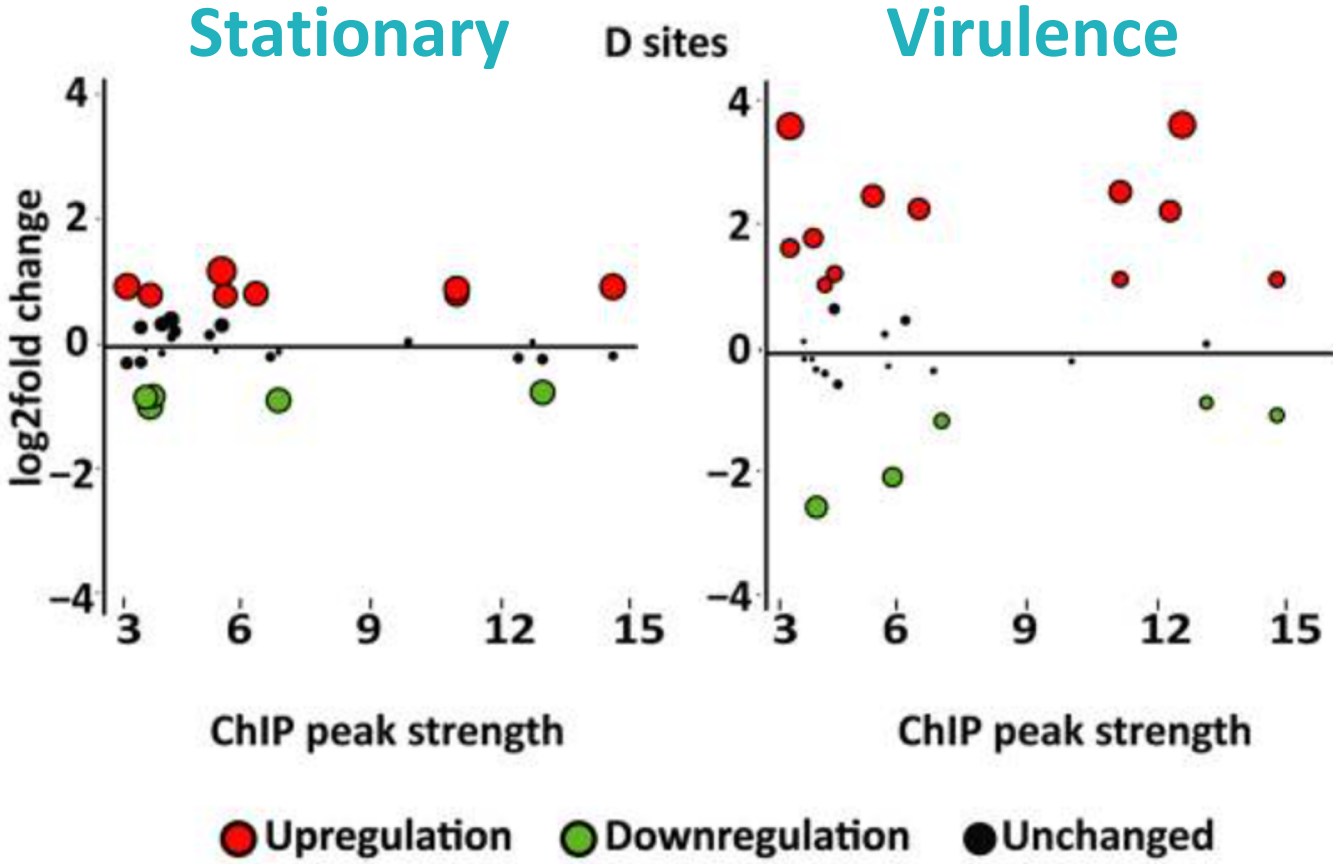
Identified binding sites that mediate positive and negative regulation

(Choudhary *et al.*, 2020)

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Host-Microbe Biology

# Genome-Scale Mapping Reveals Complex Regulatory Activities of RpoN in *Yersinia pseudotuberculosis*




(Choudhary *et al.*, 2020)

# Visualizing Combined CHIP-seq and RNA-seq

## Our Questions

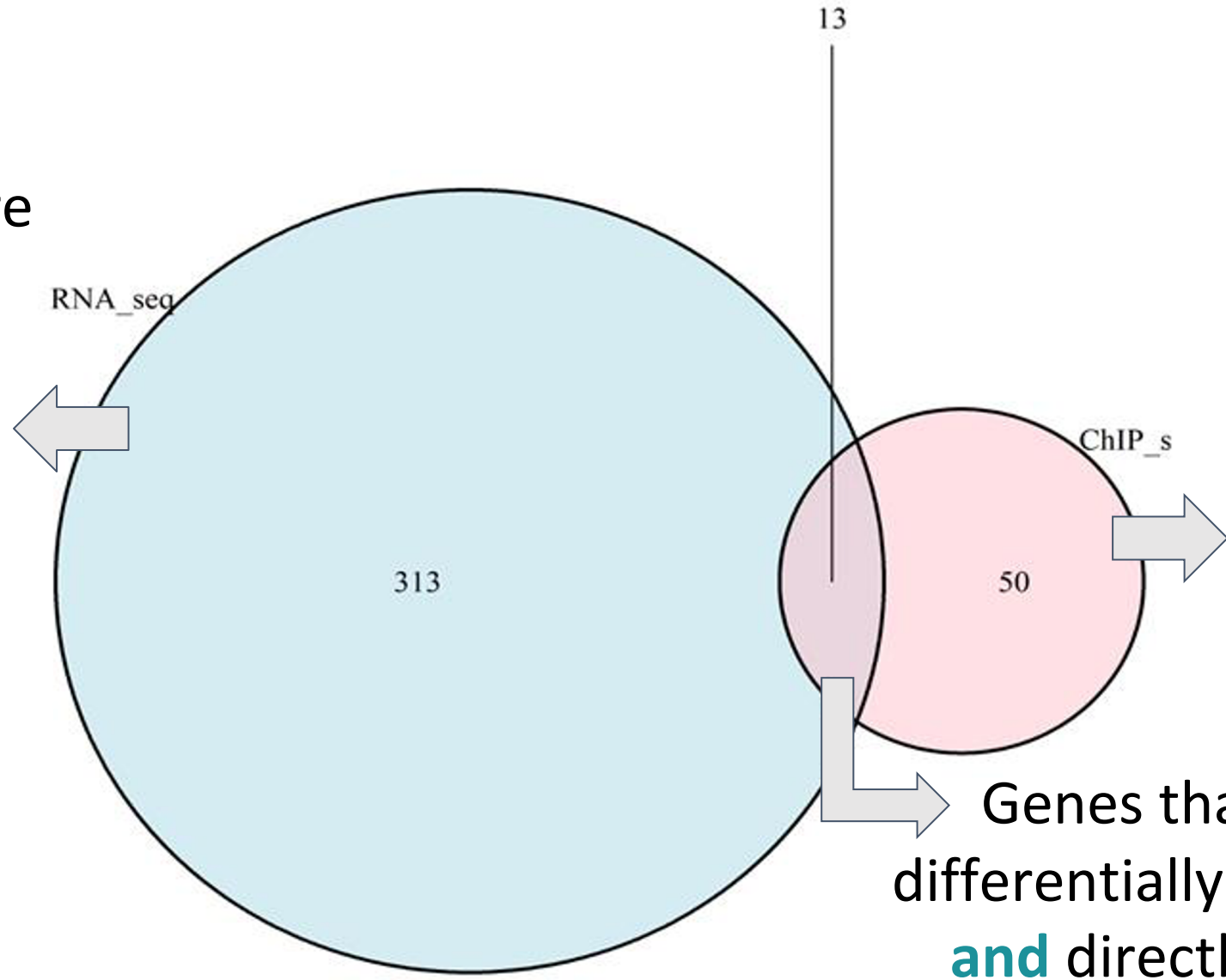
1. When you delete *lsr2*, is there is a change in expression of genes that are bound by/near Lsr2?
2. Does Lsr2 act as a repressor?
3. Are biosynthetic gene clusters particularly regulated by Lsr2 (*i.e.* does deleting *lsr2* stimulate antibiotic production)?

# Workshop time!

Integrating and visualizing CHIP-seq and RNA-seq data

1. When you delete *lsr2*, is there is a change in expression of genes that are bound by/near Lsr2?

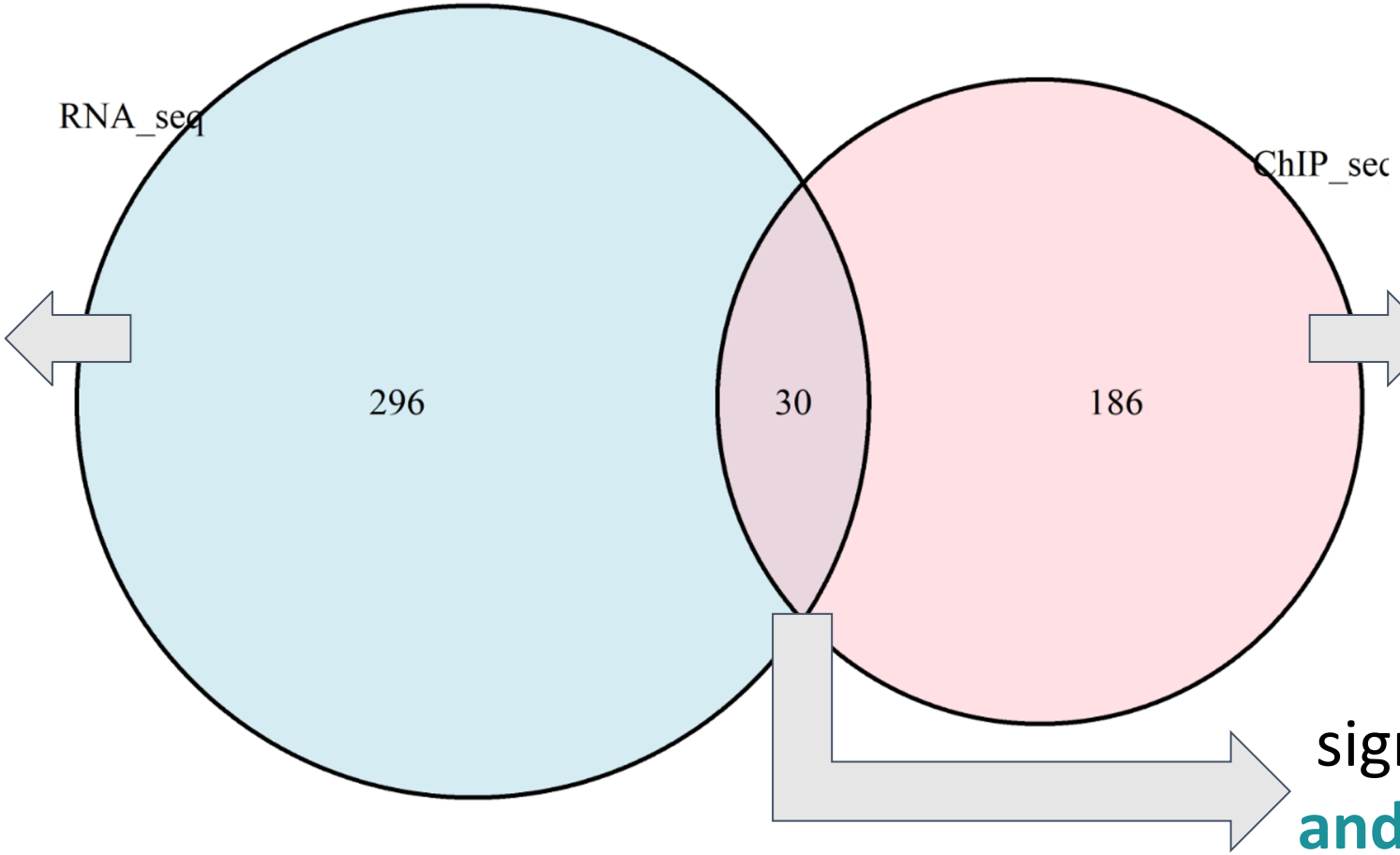
Genes that were significantly differentially expressed




Genes that were directly bound by Lsr2

Genes that were differentially expressed **and** directly bound

Genes that were significantly differentially expressed

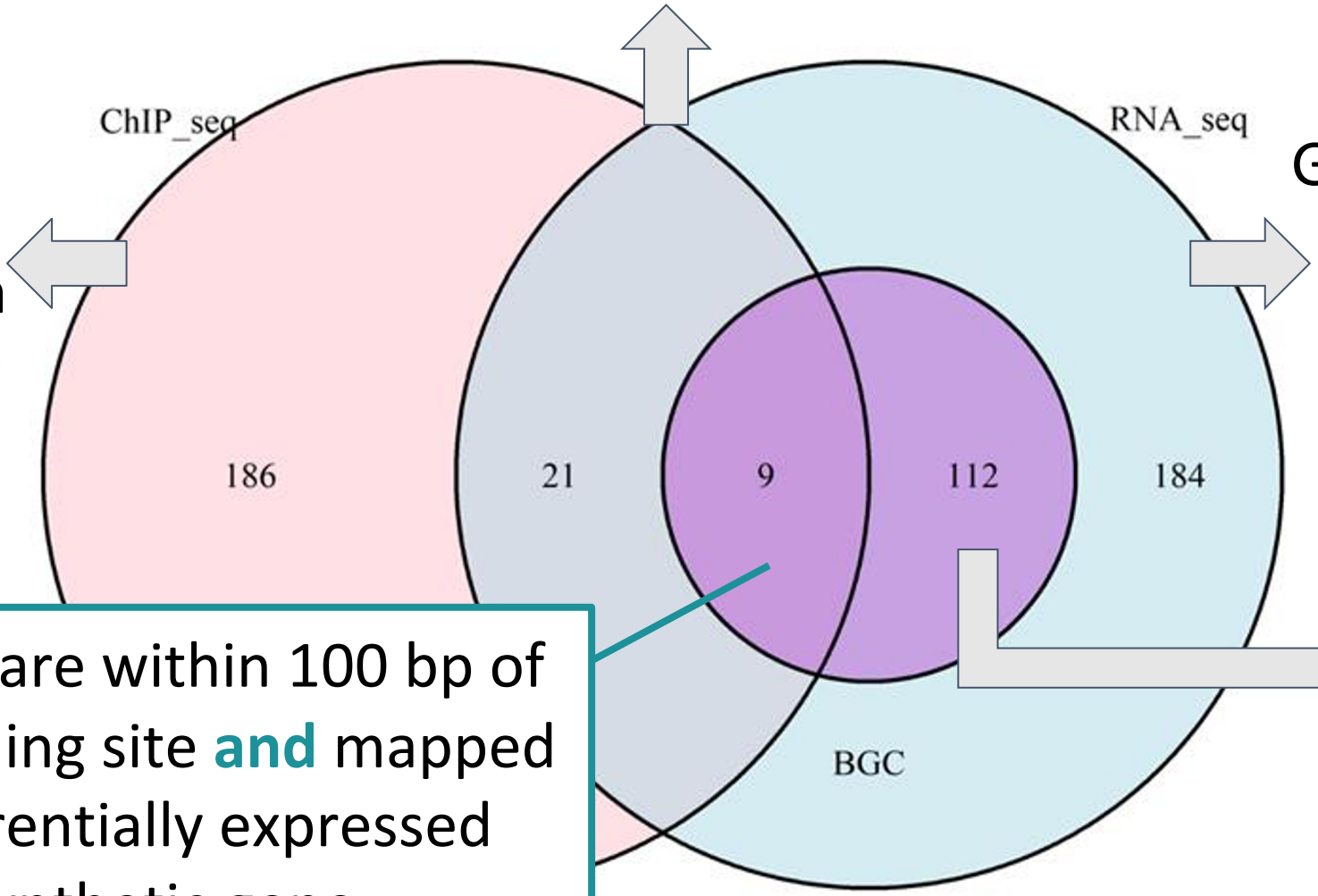



Genes that were within 100 bp of an Lsr2 binding site

Genes that were significantly expressed **and** were within 100 bp of an Lsr2 binding site

Genes that were differentially expressed **and** were within 100 bp of an Lsr2 binding site

Genes that were within 100 bp of an Lsr2 binding site




Genes that were significantly differentially expressed

Genes that are within 100 bp of an Lsr2 binding site **and** mapped to a differentially expressed biosynthetic gene

Differentially expressed biosynthetic genes

2. Does Lsr2 act as a repressor?

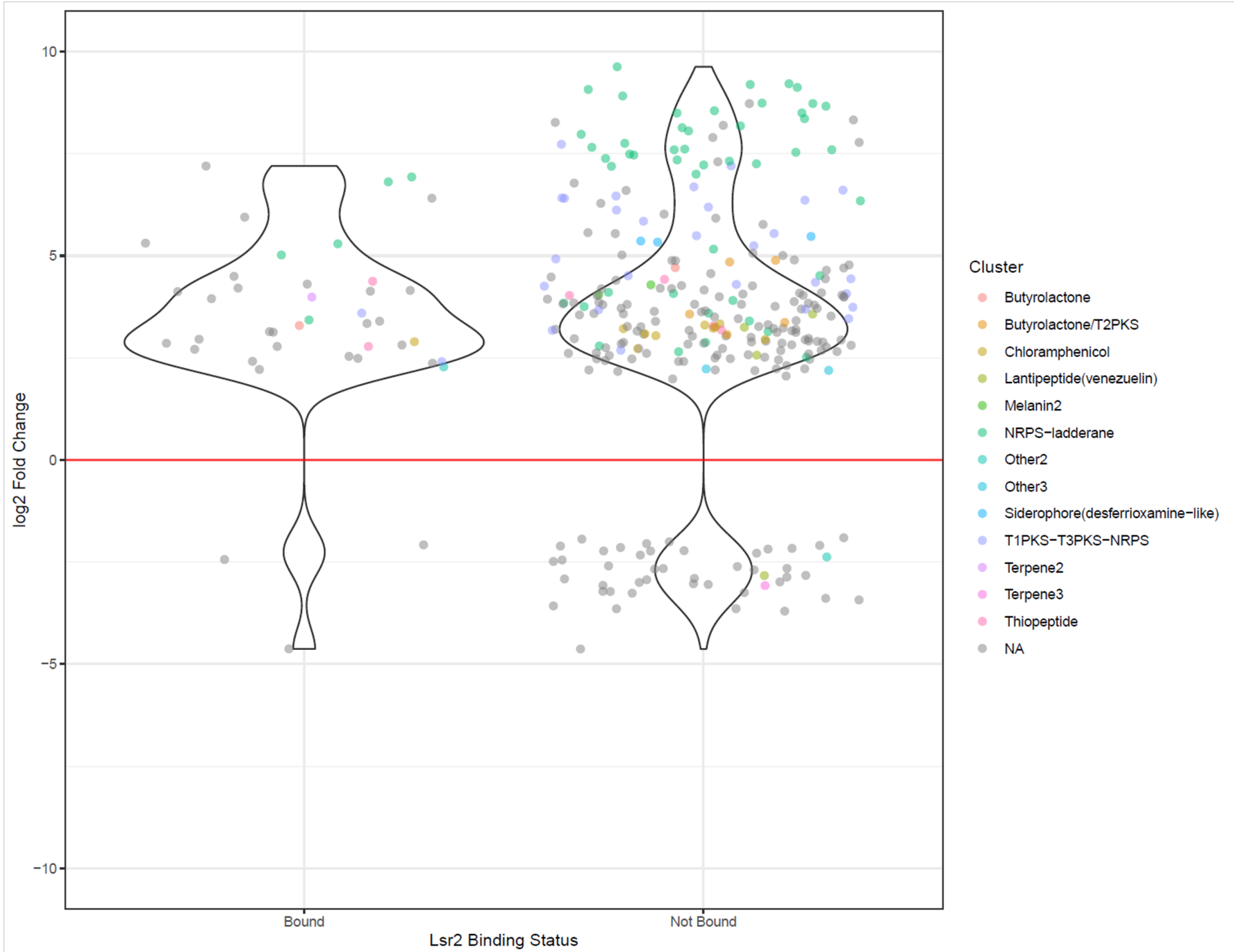
AND

3. Are biosynthetic gene clusters particularly regulated by Lsr2 (i.e. does deleting lsr2 stimulate antibiotic production)?

# Violin Plots Using ggplot2

- Combination of a boxplot and a kernel density plot
  - **Boxplot** - displays the distribution of numerical data as quartiles, with minimum and maximum scores and the median
  - **Kernel Density Plot** - estimates the probability density function of a continuous variable (log2 fold change of significantly differentially expressed genes) and add a smooth curve

- Compares the distribution of the numerical samples (log2 fold change of significantly differentially expressed genes) across categories (Lsr2 bound or not bound)


10
5
0
−5
−10
log2 Fold Change
Bound
Not Bound
Lsr2 Binding Status
Cluster
Butyrolactone
Butyrolactone/T2PKS
Chloramphenicol
Lantipeptide(venezuelin)
Melanin2
NRPS−ladderane
Other2
Other3
Siderophore(desferrioxamine−like)
T1PKS−T3PKS−NRPS
Terpene2
Terpene3
Thiopeptide
NA

# Thank you!

Questions?

# References


1. Höllbacher, B., Balázs, K., Heinig, M., & Uhlenhaut, N. H. (2020). Seq-ing answers: current data integration approaches to uncover mechanisms of transcriptional regulation. Computational and structural biotechnology journal.

2. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 2018 Jul 18. doi: 10.1093/bioinformatics/bty648.

3. Shen, X., Xu, L., Liu, Y., Dong, H., Zhou, D., Zhang, Y., Lin, S., Cao, J., & Huang, L. (2019). Comparative transcriptome analysis and ChIP-sequencing reveals stage-specific gene expression and regulation profiles associated with pollen wall formation in Brassica rapa. BMC genomics, 20(1), 264. https://doi.org/10.1186/s12864-019-5637-x

4. Choudhary, K. S., Kleinmanns, J. A., Decker, K., Sastry, A. V., Gao, Y., Szubin, R., Seif, Y., & Palsson, B. O. (2020). Elucidation of Regulatory Modes for Five Two-Component Systems in Escherichia coli Reveals Novel Relationships. mSystems, 5(6), e00980-20. https://doi.org/10.1128/mSystems.00980-20

5. Mahmud, A., Nilsson, K., Fahlgren, A., Navais, R., Choudhury, R., Avican, K., & Fällman, M. (2020). Genome-Scale Mapping Reveals Complex Regulatory Activities of RpoN in Yersinia pseudotuberculosis. mSystems, 5(6), e01006-20. https://doi.org/10.1128/mSystems.01006-20

6. Bogue, M. M., Mogre, A., Beckett, M. C., Thomson, N. R., & Dorman, C. J. (2020). Network Rewiring: Physiological Consequences of Reciprocally Exchanging the Physical Locations and Growth-Phase-Dependent Expression Patterns of the Salmonella fis and dps Genes. mBio, 11(5), e02128-20. https://doi.org/10.1128/mBio.02128-20

# References


7. DuPai, C. D., Wilke, C. O., & Davies, B. W. (2020). A Comprehensive Coexpression Network Analysis in Vibrio cholerae. mSystems, 5(4), e00550-20. https://doi.org/10.1128/mSystems.00550-20

8. Lee, J. H., Yoo, J. S., Kim, Y., Kim, J. S., Lee, E. J., & Roe, J. H. (2020). The WblC/WhiB7 Transcription Factor Controls Intrinsic Resistance to Translation-Targeting Antibiotics by Altering Ribosome Composition. mBio, 11(2), e00625-20. https://doi.org/10.1128/mBio.00625-20

9. Hurst-Hess, K., Biswas, R., Yang, Y., Rudra, P., Lasek-Nesselquist, E., & Ghosh, P. (2019). Mycobacterial SigA and SigB Cotranscribe Essential Housekeeping Genes during Exponential Growth. mBio, 10(3), e00273-19. https://doi.org/10.1128/mBio.00273-19

10. Rioualen, C., Charbonnier-Khamvongsa, L., Collado-Vides, J., & van Helden, J. (2019). Integrating Bacterial ChIP-seq and RNA-seq Data With SnakeChunks. Current protocols in bioinformatics, 66(1), e72. https://doi.org/10.1002/cpbi.72

11. Kroner, G. M., Wolfe, M. B., & Freddolino, P. L. (2019). Escherichia coli Lrp Regulates One-Third of the Genome via Direct, Cooperative, and Indirect Routes. Journal of bacteriology, 201(3), e00411-18. https://doi.org/10.1128/JB.00411-18

12. Jaskólska, M., Stutzmann, S., Stoudmann, C., & Blokesch, M. (2018). QstR-dependent regulation of natural competence and type VI secretion in Vibrio cholerae. Nucleic acids research, 46(20), 10619–10634. https://doi.org/10.1093/nar/gky717

13. Shao, X., Zhang, X., Zhang, Y., Zhu, M., Yang, P., Yuan, J., Xie, Y., Zhou, T., Wang, W., Chen, S., Liang, H., & Deng, X. (2018). RpoN-Dependent Direct Regulation of Quorum Sensing and the Type VI Secretion System in Pseudomonas aeruginosa PAO1. Journal of bacteriology, 200(16), e00205-18. https://doi.org/10.1128/JB.00205-18

# References

14. Fishman, M. R., Zhang, J., Bronstein, P. A., Stodghill, P., & Filiatrault, M. J. (2018). Ca2+-Induced Two-Component System CvsSR Regulates the Type III Secretion System and the Extracytoplasmic Function Sigma Factor AlgU in Pseudomonas syringae pv. tomato DC3000. Journal of bacteriology, 200(5), e00538-17. https://doi.org/10.1128/JB.00538-17

15. Pan Y, Liang F, Li RJ, Qian W. MarR-Family Transcription Factor HpaR Controls Expression of the vgrR-vgrS Operon of Xanthomonas campestris pv. campestris. Mol Plant Microbe Interact. 2018 Mar;31(3):299-310. doi: 10.1094/MPMI-07-17-0187-R. Epub 2018 Jan 3. PMID: 29077520.

16. Park DM, Overton KW, Liou MJ, Jiao Y. Identification of a U/Zn/Cu responsive global regulatory two-component system in Caulobacter crescentus. Mol Microbiol. 2017 Apr;104(1):46-64. doi: 10.1111/mmi.13615. Epub 2017 Jan 23. PMID: 28035693.

17. Markel, E., Stodghill, P., Bao, Z., Myers, C. R., & Swingle, B. (2016). AlgU Controls Expression of Virulence Genes in Pseudomonas syringae pv. tomato DC3000. Journal of bacteriology, 198(17), 2330–2344. https://doi.org/10.1128/JB.00276-16

18. Schulz, S., Eckweiler, D., Bielecka, A., Nicolai, T., Franke, R., Dötsch, A., Hornischer, K., Bruchmann, S., Düvel, J., & Häussler, S. (2015). Elucidation of sigma factor-associated networks in Pseudomonas aeruginosa reveals a modular architecture with limited and function-specific crosstalk. PLoS pathogens, 11(3), e1004744. https://doi.org/10.1371/journal.ppat.1004744

19. Lam, H. N., Chakravarthy, S., Wei, H. L., BuiNguyen, H., Stodghill, P. V., Collmer, A., Swingle, B. M., & Cartinhour, S. W. (2014). Global analysis of the HrpL regulon in the plant pathogen Pseudomonas syringae pv.

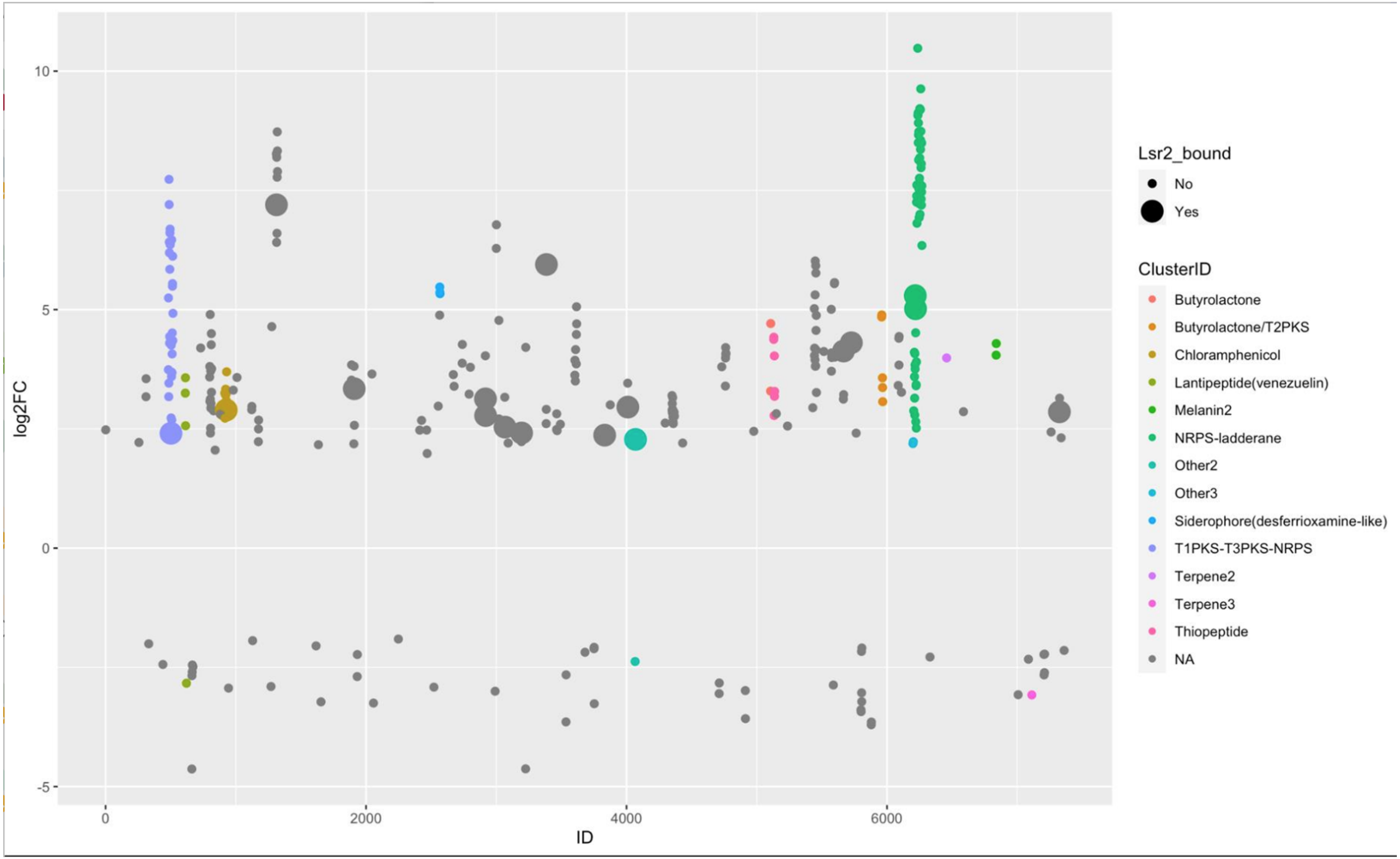
# References

20. Balasubramanian, D., Kumari, H., Jaric, M., Fernandez, M., Turner, K. H., Dove, S. L., Narasimhan, G., Lory, S., & Mathee, K. (2014). Deep sequencing analyses expands the Pseudomonas aeruginosa AmpR regulon to include small RNA-mediated regulation of iron acquisition, heat shock and oxidative stress response. Nucleic acids research, 42(2), 979–998. https://doi.org/10.1093/nar/gkt94

21. Uplekar, S., Rougemont, J., Cole, S. T., & Sala, C. (2013). High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in Mycobacterium tuberculosis. Nucleic acids research, 41(2), 961–977. https://doi.org/10.1093/nar/gks1260

22. Jutras, B. L., Bowman, A., Brissette, C. A., Adams, C. A., Verma, A., Chenail, A. M., & Stevenson, B. (2012). EbfC (YbaB) is a new type of bacterial nucleoid-associated protein and a global regulator of gene expression in the Lyme disease spirochete. Journal of bacteriology, 194(13), 3395–3406. https://doi.org/10.1128/JB.00252-12

# Supplementary Slides


10
5
0
-5
log2FC
0
2000
4000
6000
ID
Lsr2_bound
No
Yes
ClusterID
Butyrolactone
Butyrolactone/T2PKS
Chloramphenicol
Lantipeptide(venezuelin)
Melanin2
NRPS-ladderane
Other2
Other3
Siderophore(desferrioxamine-like)
T1PKS-T3PKS-NRPS
Terpene2
Terpene3
Thiopeptide
NA