CUT&Tag

in the Colorado Potato Beetle
or

Elisa Israel,Computational EvoDevo Group, Leipzig University,
Bled 11.02.2025

How UPS f*cked with my PhD

PhD topic

Evolution of Epigenetic Regulation in Beetles

- focus of our study
  - DNA methylation
    - associated with active gene expression
    - very low in insects
    - lost in **some** beetles (not in CPB)
  - histone modifications
    - can alter chromatin structure and affect transcription
    - H3K36me3
    - H3K27ac
    - associated with active gene expression
- both are highly interlinked in vertebrates
- studies in insects are limited

Objective

multi species multi-omics (embryo and adult), combining

- RNAseq (gene transcription)
- EMseq (DNA methylation)
- CUT&Tag (histone modifications)

Objective

multi species multi-omics, combining

- RNAseq (gene transcription)
- EMseq (DNA methylation)
- CUT&Tag (histone modification patterns)
  - H3K36me3: prevalent on gene bodies
  - H3K27ac: associated with regulatory regions

Objective

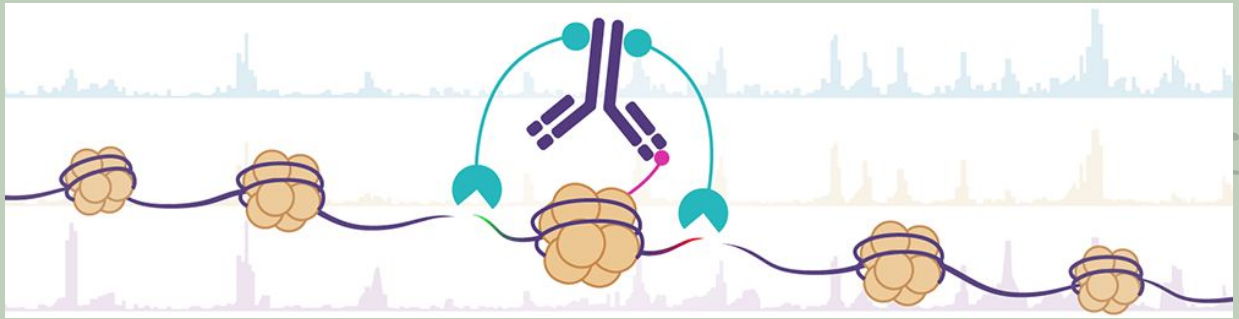
multi species multi-omics, combining

- RNAseq (gene transcription)
- EMseq (DNA methylation)
- **CUT&Tag (histone modifications) in CPB embryos**

Because life would be boring if all would go well …

CUT&Tag

- Cleavage Under Targets and Tagmentation sequencing
- chromatin protein / modification is bound in situ by a specific antibody, which then tethers a protein A-Tn5 transposase fusion protein (pA-Tn5)
- underlying DNA is marked and cleaved
- fragments are of nucleosome length



CUT&Tag

- Cleavage Under Targets and Tagmentation sequencing
- chromatin protein / modification is bound in situ by a specific antibody, which then tethers a protein A-Tn5 transposase fusion protein (pA-Tn5)
- underlying DNA is marked

- improvement to ChIP-seq and CUT&RUN
- high resolution, low background

UPS

Storytime

You always think of all the things that can go wrong …

Storytime

You always think of all the things that can go wrong …

… but some things we just did not anticipate.

Mainz

Münster

Mainz

Münster

Mainz

Münster

Mainz

Münster

Mainz

Münster

Result: Panic.

Panic.

- not enough material as a backup

Panic.

- not enough material as a backup

- package found couple of days later
- samples defrosted
- sequenced anyway, hoping for the best

Panic.

- not enough material as a backup

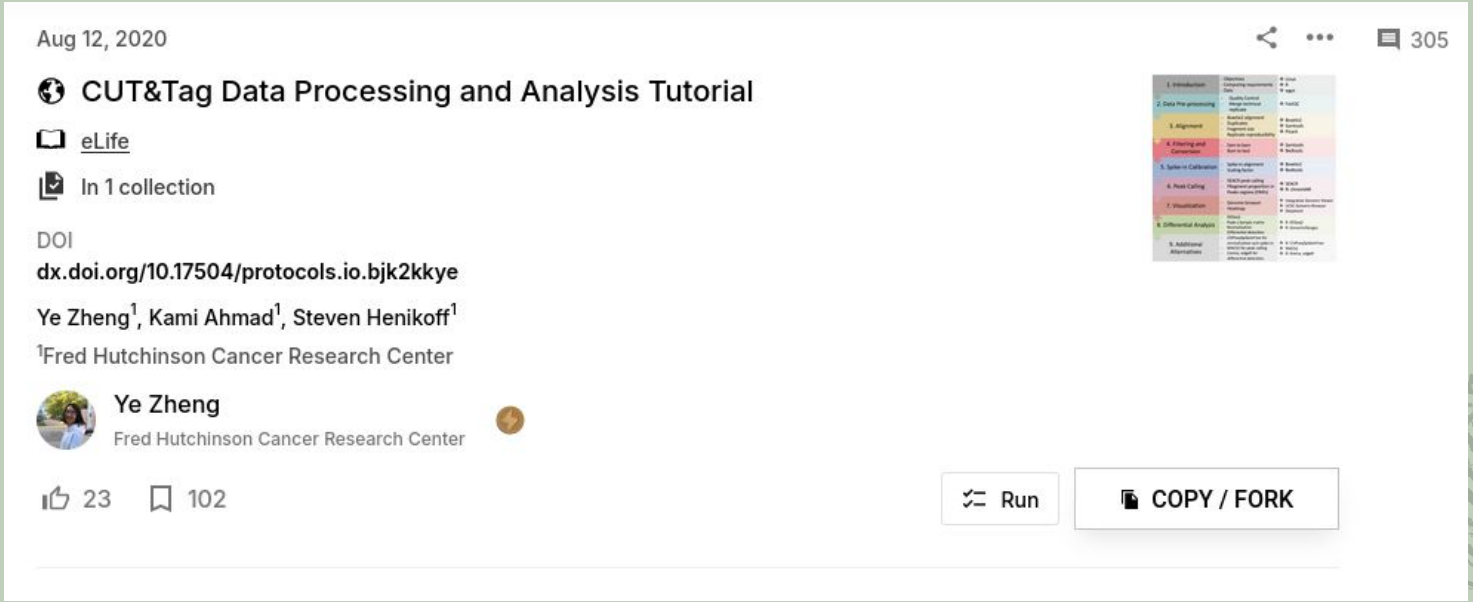
- package found couple of days later
- samples defrosted
- sequenced anyway, hoping for the best

- resequencing would mean to redo RNAseq and EMseq as well

Data Analysis

Based on

Protocol of Zheng et al. (2020)



Analysis Pipeline

1. Pre-Processing, Quality Control
2. Alignment
3. Check mapping efficiency, fragment size and replicate reproducibility
4. Filtering
5. Spike-In Calibration
6. Peak Calling
7. Visualization
8. Combining Results

Analysis Pipeline

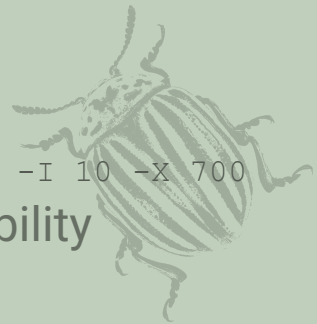
1. Pre-Processing, **Quality Control**
2. Alignment
3. Check **mapping efficiency,** fragment size and **replicate reproducibility**
4. Filtering
5. **Spike-In Calibration**
6. **Peak Calling**
7. Visualization
8. Combining Results

Analysis Pipeline

1. Pre-Processing
   - ➢ quality control: fastqc (version 0.12.1)
   - ➢ adapter removal: cutadapt (version 4.8)

   - ➢ higher GC content than expected
   - ➢ 35.5% expected vs. 38-43% observed
   - ➢ might be due to defrosting?
2. Alignment
   - ➢ trimmed reads to reference genome (CPB atlas)
   - ➢ bowtie2 (version 2.5.3), parameters:

     ```
     --end-to-end --very-sensitive --no-mixed --no-discordant --phred33 -I 10 -X 700
     ```
3. Check mapping efficiency, fragment size and replicate reproducibility

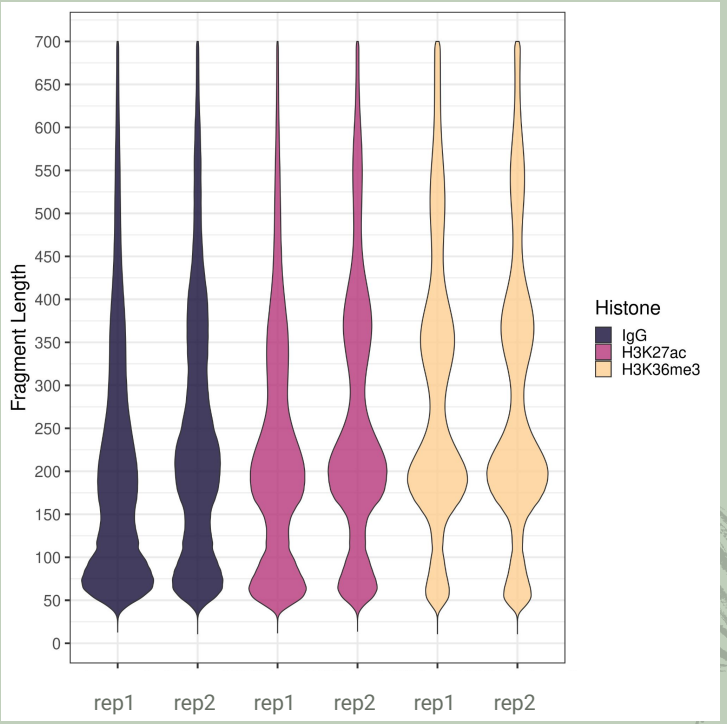


Mapping Efficiency

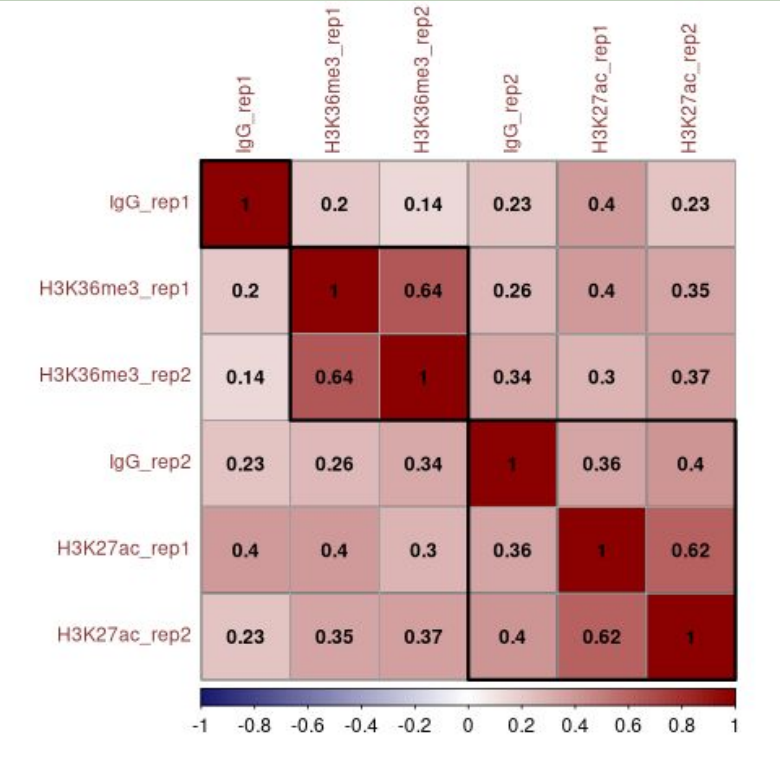
	IgG (rep1)	IgG (rep2)	H3K27ac (rep1)	H3K27ac (rep2)	H3K36me3 (rep1)	H3K36me3 (rep2)
reads	16,393,724	4,948,671	20,568,774	14,155,082	20,187,579	19,151,354
aligned 0 times (in %)	11.25	37.15	8.75	10.94	11.95	19.72
aligned 1 time (in %)	40.08	26.85	59.9	52.77	56.13	50.96
aligned >1 times (in %)	48.67	35.99	31.35	36.29	31.92	29.32
overall alignment rate (in %)	**88.75**	**62.85**	**91.25**	**89.06**	**88.05**	**80.28**

3. Fragment Length Distribution

- fragments should have nucleosome length
- shorter fragments due to
  - tagmentation of DNA at nucleosome surface (typically 50-100 bp)
  - background noise



Replicate Reproducibility (Pearson)



5. Spike-In Calibration

- *E. coli* DNA is carried along with pA-Tn5 protein and gets tagmented non-specifically during reaction
- assumption: two experiments start with same amount of pA-Tn5
  - fixed amount of *E. coli* DNA
  - ratio of fragments mapped to *E. coli* genome is the same for a series of samples
  - *E. coli* reads can be used to normalize epitope abundance in a set of experiments

Of course, this idea makes perfect sense, but

5. Spike-In Calibration

- *E. coli* DNA is carried along with p… …ts tagmented non-specifically during reacti…
- assumption: two experi… …A-Tn5
  - fixed amount of *E. c*…
  - ratio of fragme… of samples
  - *E. coli* reads … …t of experiments

Oh boy did that not work

Of course, this idea ma…

The same culprits

	IgG (rep1)	IgG (rep2)	H3K27ac (rep1)	H3K27ac (rep2)	H3K36me3 (rep1)	H3K36me3 (rep2)
reads	16,393,724	4,948,671	20,568,774	14,155,082	20,187,579	19,151,354
aligned 0 times	99.7%	98.06%	100%	100%	100%	99.49%
aligned 1 time	1,657 0.01%	57,279 1.16%	23 0%	21 0%	103 0%	75,578 0.39%
aligned >1 times	2,625 0.02%	38,838 0.78%	18 0%	9 0%	64 0%	21,580 0.11%
reads aligned to *E. coli*	**4,282 0.03%**	**96,117 1.94%**	**41 0%**	**30 0%**	**167 0%**	**97,158 0.51%**

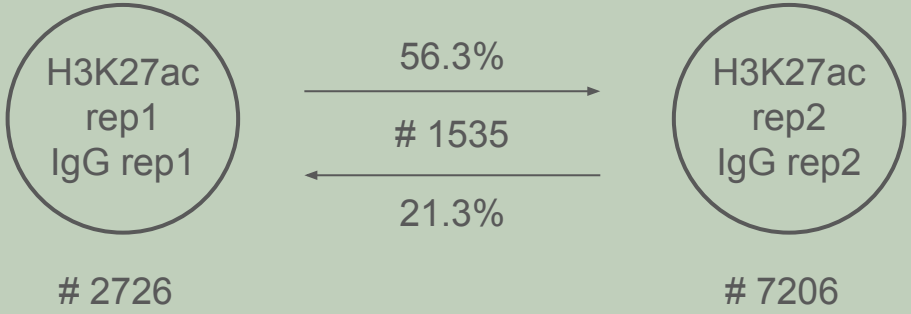
5. Spike-In Calibration

- ➢ almost no signal left if data is normalized using this approach
- ➢ instead use build-in normalization of peak caller `SEACR`

6. Peak Calling

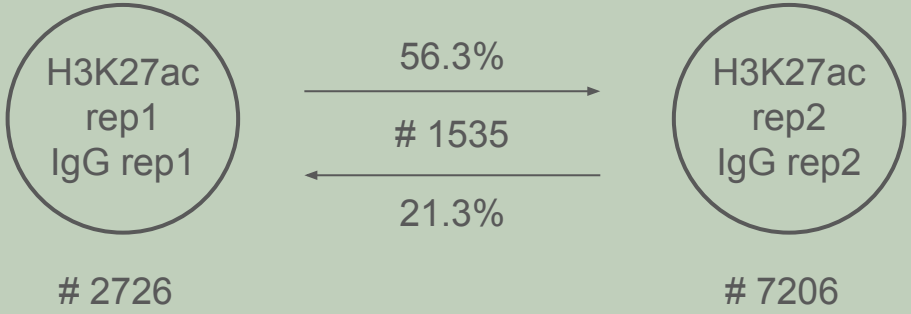
- feature coverage: `bedtools genomecov` (version 2.31.1)
- using `SEACR` (version 1.3), parameters: `'norm', 'stringent'`
- calls peaks and enriched regions from chromatin profiling data with low background
- with or without IgG control

Peak Reproducibility by overlap

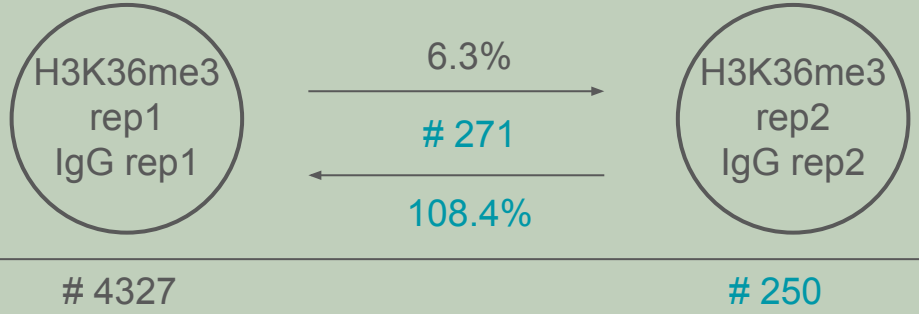


Calculate the number of peaks that appear in both replicates.

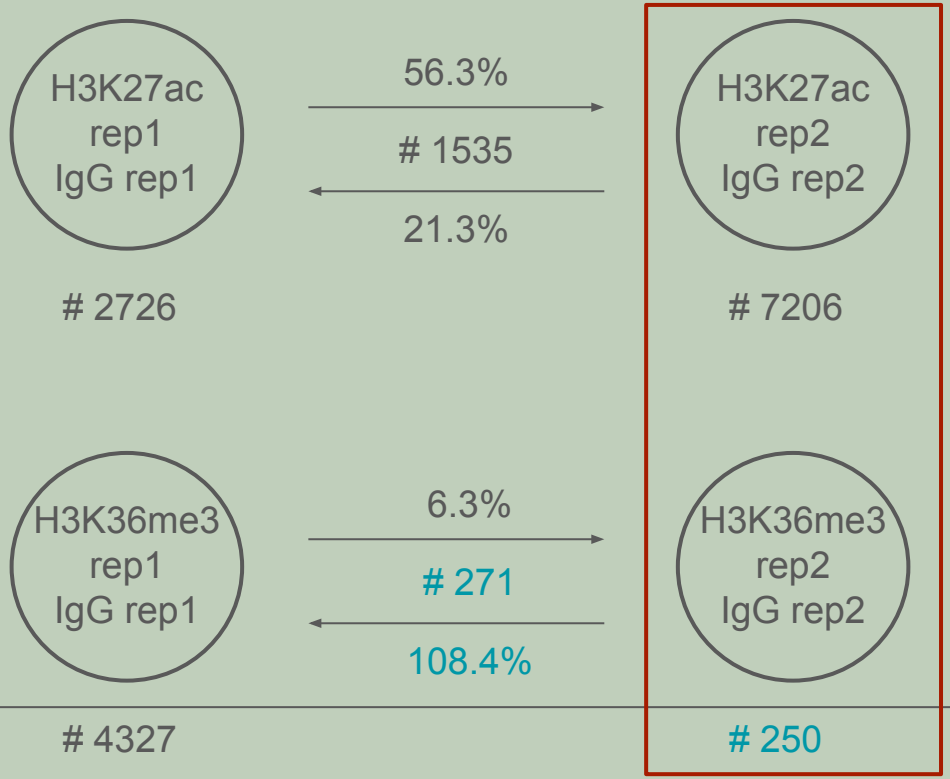
Peak Reproducibility by overlap



some peaks from one sample might overlap with multiple peaks from the other



Peak Reproducibility by overlap



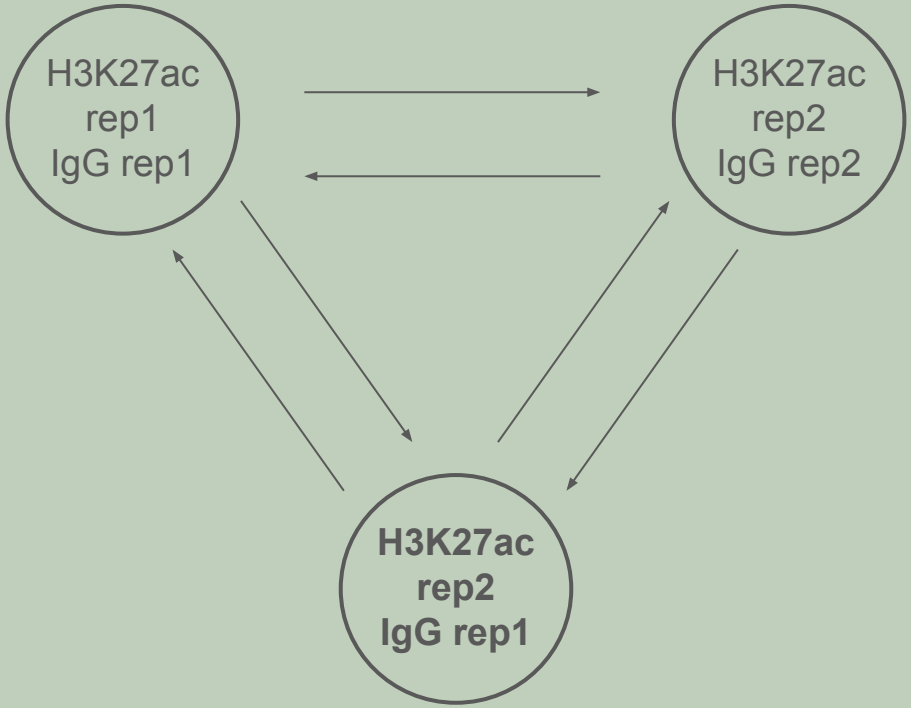
This doesn't look right.

Peak Reproducibility

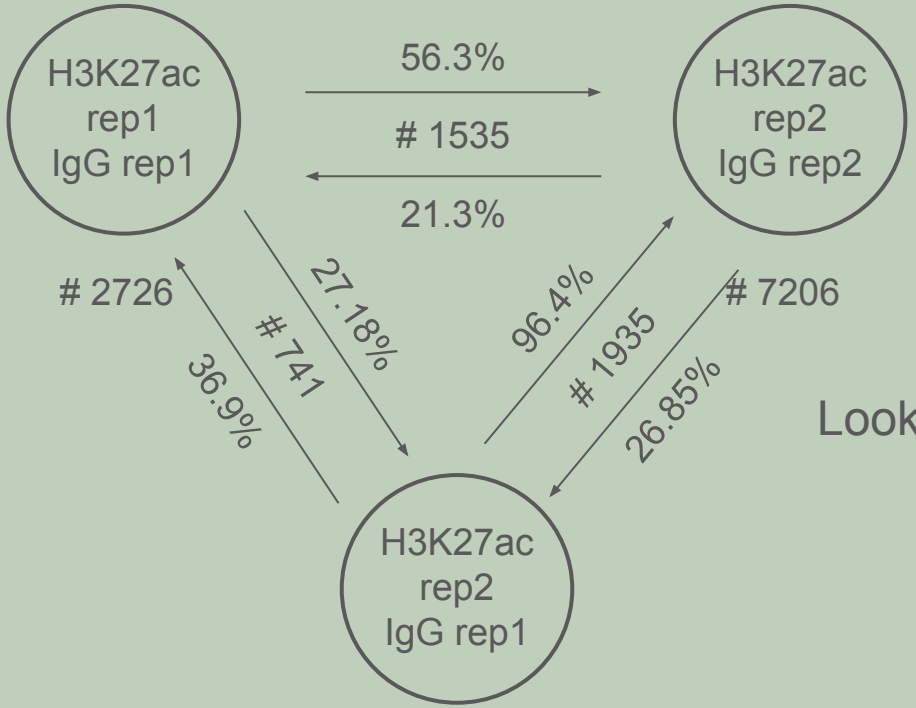
Some labs only use one IgG control for all replicates.

Let’s introduce a third set …

Peak Reproducibility by overlap

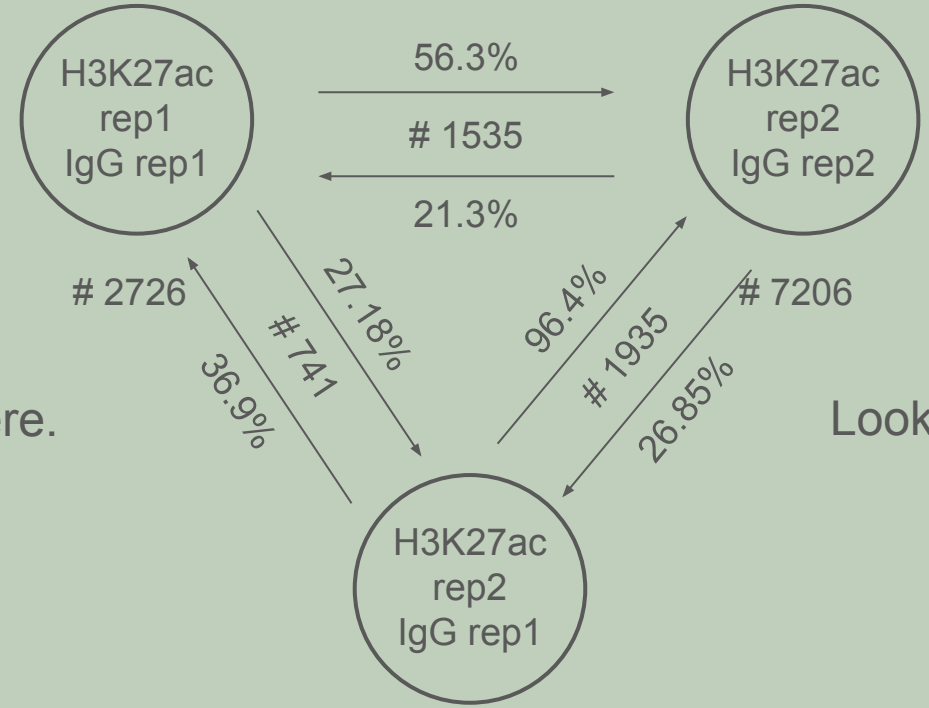


Peak Reproducibility by overlap



Looks good here …

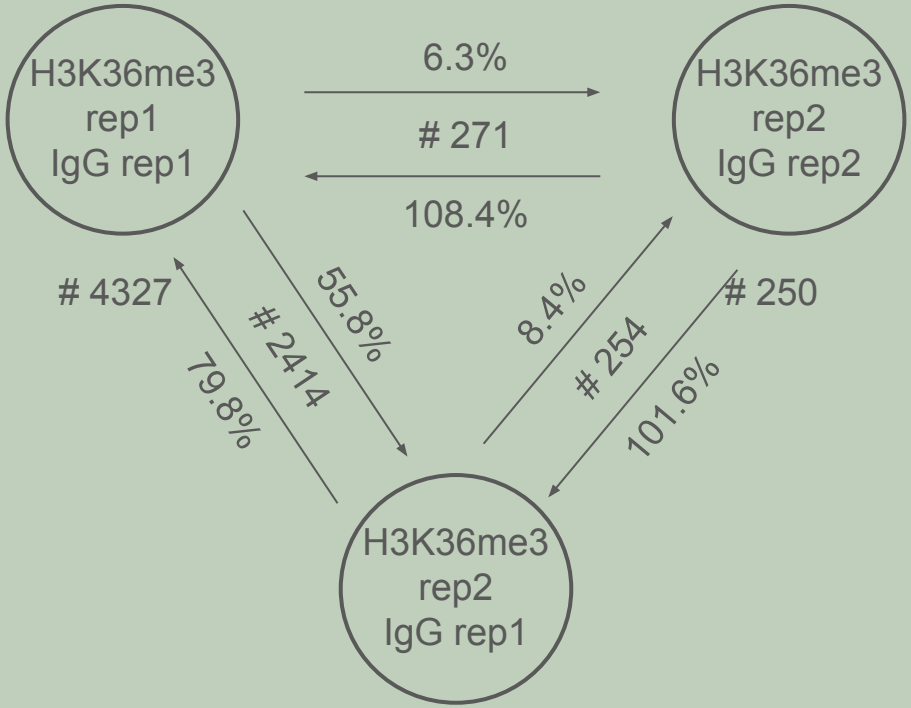
Peak Reproducibility by overlap



… but not here.

Looks good here …

Peak Reproducibility by overlap



Peak Reproducibility

- IgG rep2
  - does not reduce background noise for H3K27ac
  - almost completely eliminates the signal of H3K36me3
- using IgG rep1 for both sample replicates
  - improves H3K36me3 results
  - but impairs results for H3K27ac

Well … What now?

New Nemesis Unlocked: IgG replicate 2

SEND HELP. PLEASE.

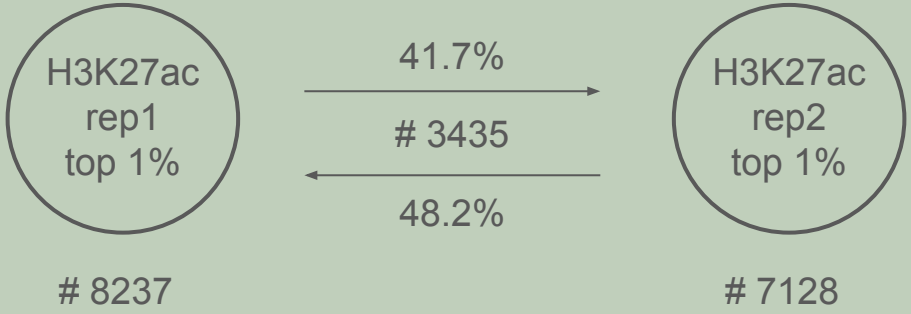
Me, trying to make sense of it all.
Symbolic image.

Peak Reproducibility

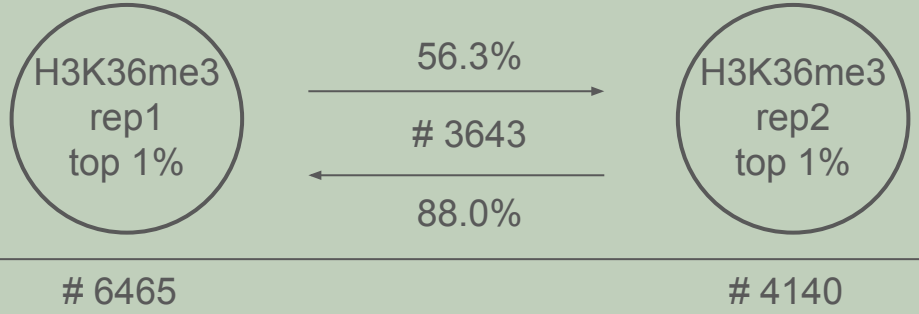
How about getting rid of IgG controls entirely?

Look at the set of 1% top peaks.

Peak Reproducibility

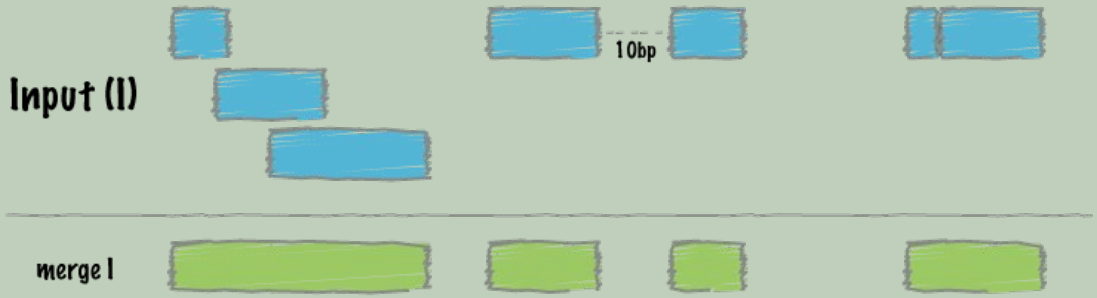


Looks better



High-Confidence Peaks

- choose top 2.5% of peaks
- only keep peaks present in both replicates
- merge the corresponding peaks
- result: high-confidence set of reproducible peaks



7. Visualization of Enrichment Patterns

- `deepTools` (version 3.5.5)
  - `computeMatrix: --scale-regions`
  - `plotHeatmap`
- visualization of whole gene:
  - gene length normalized to a length of 5 kb
  - 3 kb upstream and downstream of the gene body
- visualization of TSS and TES:
  - 0.5 kb upstream and downstream of feature

Multi-Omics

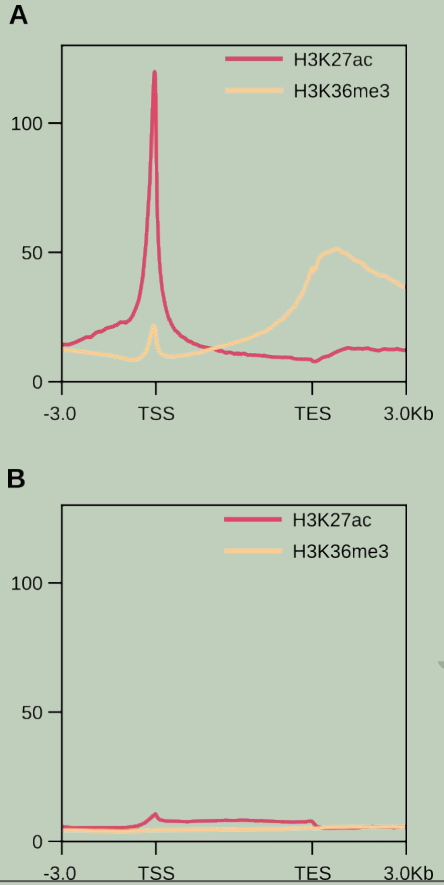
Multi-Omics: CUT&Tag and RNAseq

histone modifications and gene expression

- majority of peaks are covering genes
  - 79% for H3K27ac → 21% in intergenic regions
  - 83% for H3K36me3 → 17% in intergenic regions
- expressed genes
  - 66% have H3K27ac or H3K36me3 peaks
  - 58%have peaks for **both** modifications
- only 10% of not expressed genes show an overlap with either modification

Enrichment Patterns

- H3K27ac - expressed genes
  - prominent, narrow peak at TSS
  - small dip around the TES
- H3K36me3 - expressed genes
  - small peak at TSS
  - steep incline towards TES
  - gradual decline far into downstream flanking region
- almost no signal in not expressed genes



Multi-omics: EMseq and RNAseq

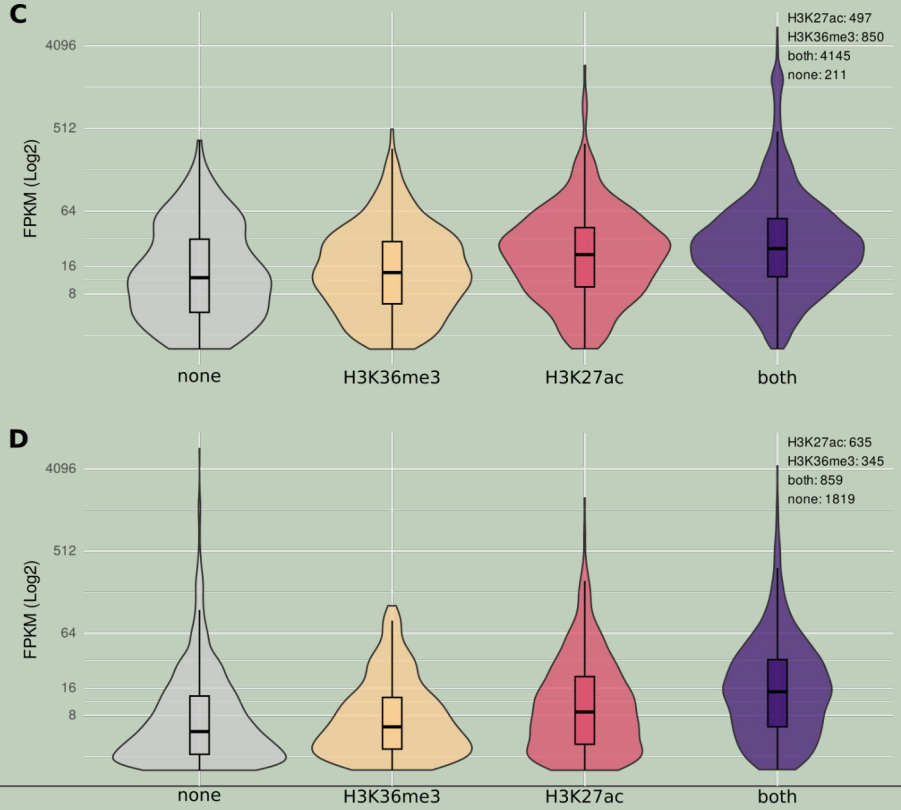
- adding DNA methylation status (mCpGs)
- dividing genes into four subsets

not methylated / expressed	methylated / expressed
not methylated / not expressed	methylated / not expressed

Histone Modifications, Methylation and Gene Expression

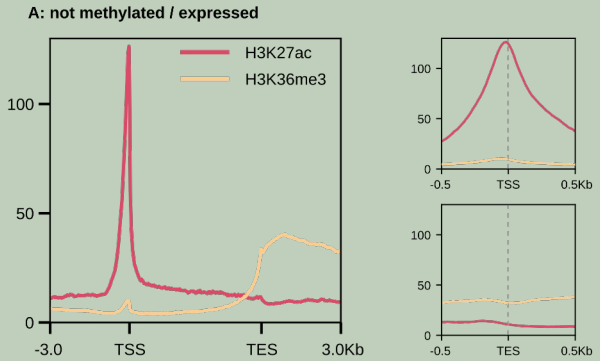
methylated / expressed

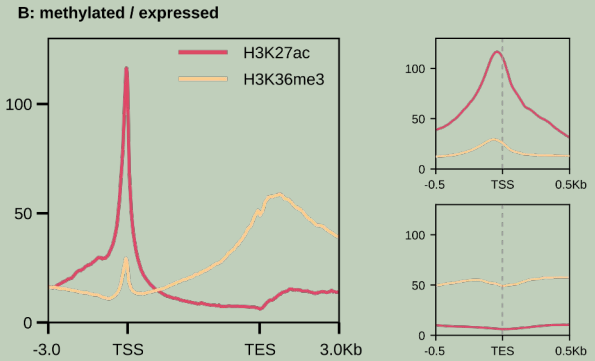
not methylated / expressed



Enrichment Patterns

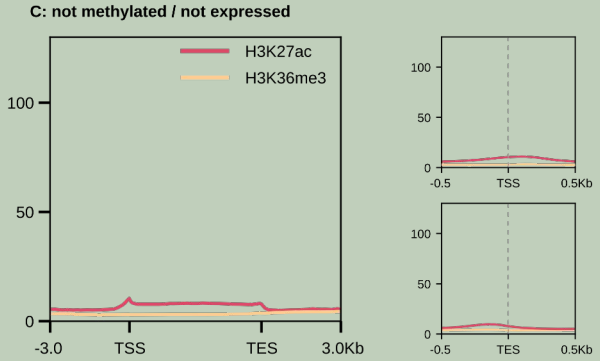
not methylated / expressed

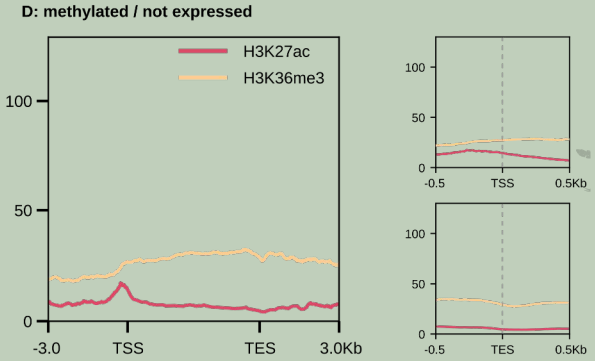




methylated / expressed

not methylated / not expressed





methylated / not expressed

Outlook

- results for *L. decemlineata* submitted
  - preprint can be found on bioRxiv 10.1101/2025.01.09.632173
- ongoing: analysis of *T. castaneum*
  - data looks much better
- planned for (currently in the lab)
  - *L. decemlineata* (adult)
  - *Nicrophorus vespilloides* (embryo, adult)
  - *Onthophagus taurus* (adult)
  - *Tenebrio molitor* (embryo, adult)
- fine-tuning of our analysis pipeline

Thank you!

Joachim Kurtz

Sonja Prohaska

Zoe Marie Länger

DFG Deutsche Forschungsgemeinschaft

GENOMIC BASIS OF EVOLUTIONARY INNOVATIONS

Evol

Backup - Methods

Alignment

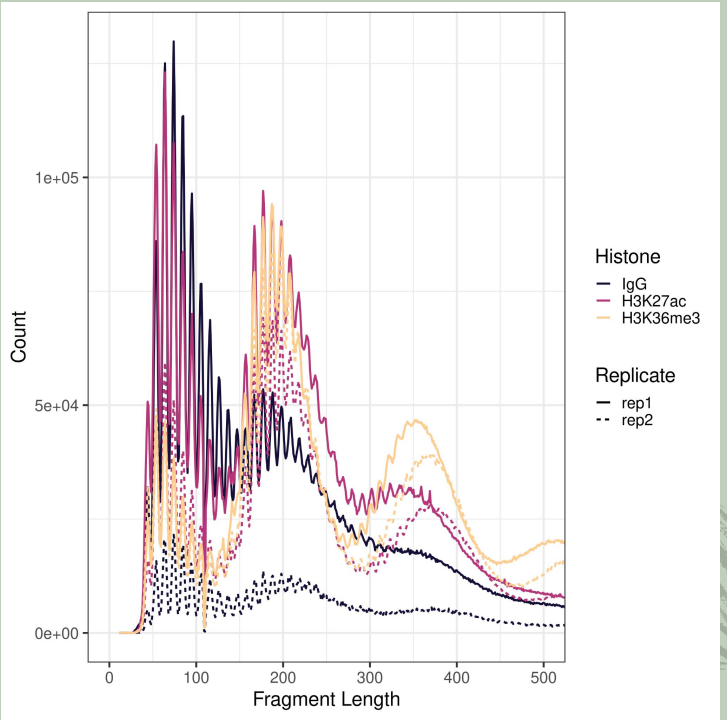
- trimmed reads to reference genome
  - GCF_031307605.1 (*T. castaneum*)
  - Leptinotarsa_decemlineata_01 (CPB atlas)
- `bowtie2` (version 2.5.3), parameters:

```
--end-to-end --very-sensitive --no-mixed --no-discordant --phred33 -I 10 -X 700
```

- alternative: `--local` instead of `--end-to-end`
  - allows mismatches at start and end of read
  - higher overall alignment rate, but with way more multi-aligned reads

3. Fragment Length Distribution

- tagmentation of DNA at nucleosome surface leads to a 10 bp sawtooth periodicity
- typical for successful CUT&Tag



Alignment Results - *L. decemlineata* (embryo)

	IgG (rep1)	**IgG (rep2)**	**H3K27ac (rep1)**	**H3K27ac (rep2)**	**H3K36me3 (rep1)**	**H3K36me3 (rep2)**
reads	16,393,724	4,948,671	20,568,774	14,155,082	20,187,579	19,151,354
aligned 0 times (in %)	11.25 *4.43*	37.15 *32.97*	8.75 *3.08*	10.94 *4.30*	11.95 *5.59*	19.72 *14.35*
aligned 1 time (in %)	40.08 *27.88*	26.85 *19.71*	59.9 *50.80*	52.77 *40.95*	56.13 *46.11*	50.96 *42.40*
aligned >1 times (in %)	48.67 *67.69*	35.99 *47.33*	31.35 *46.12*	36.29 *54.75*	31.92 *48.30*	29.32 *43.25*
overall alignment rate (in %)	**88.75** ***95.57***	**62.85** ***67.03***	**91.25** ***96.92***	**89.06** ***95.70***	**88.05** ***94.41***	**80.28** ***85.65***

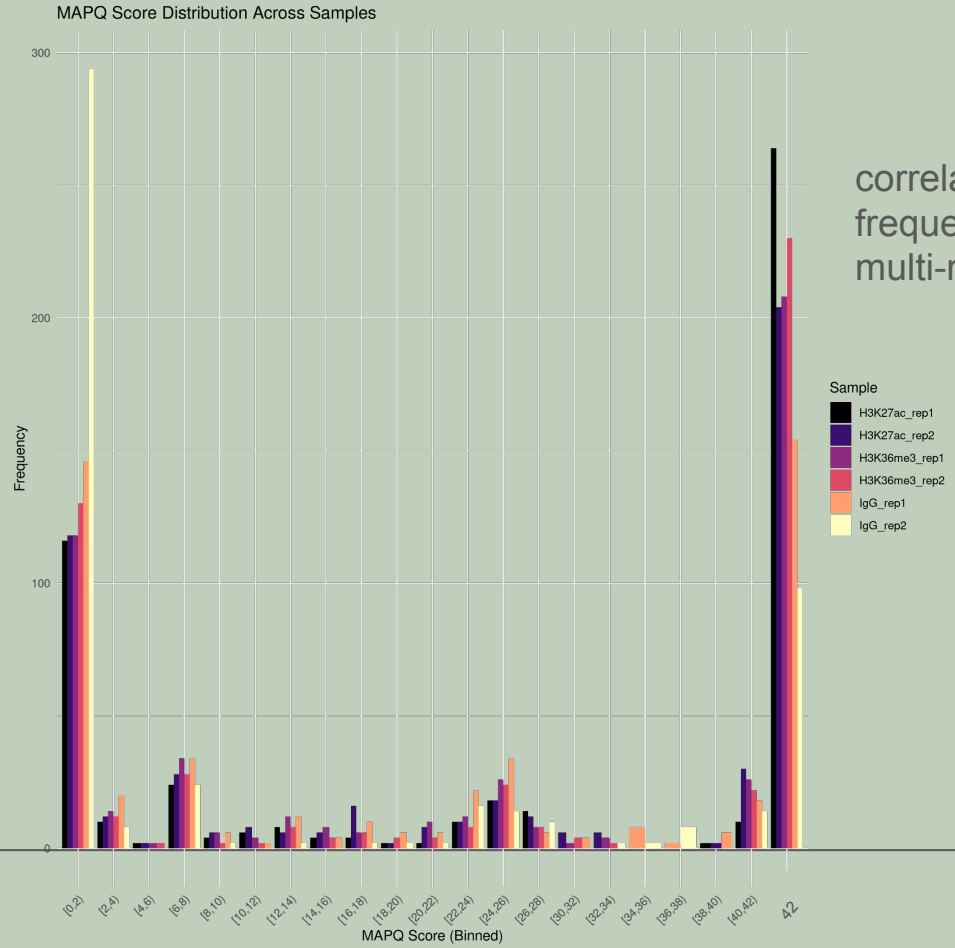
100% --end-to-end
100% --local

Filtering

- `bowtie2` assigns quality score to each mapped read
- `MAPQ(x)` scores are between 0-42
- unique fragments reach scores up to 42, but
- value will be automatically set to 1 for reads that can be aligned multiple times

- reads are often filtered with `MAPQ(x) = 2`
  → only uniquely mapped reads are kept

MAPQ(x)



correlates well with the frequency of multi-mappable reads

Replicate Reproducibility

- genome is split into 500 bp bins
- Pearson correlation of the log2-transformed values of read counts is calculated between replicate data sets
- midpoint of each fragment used to infer which bin it belongs to