

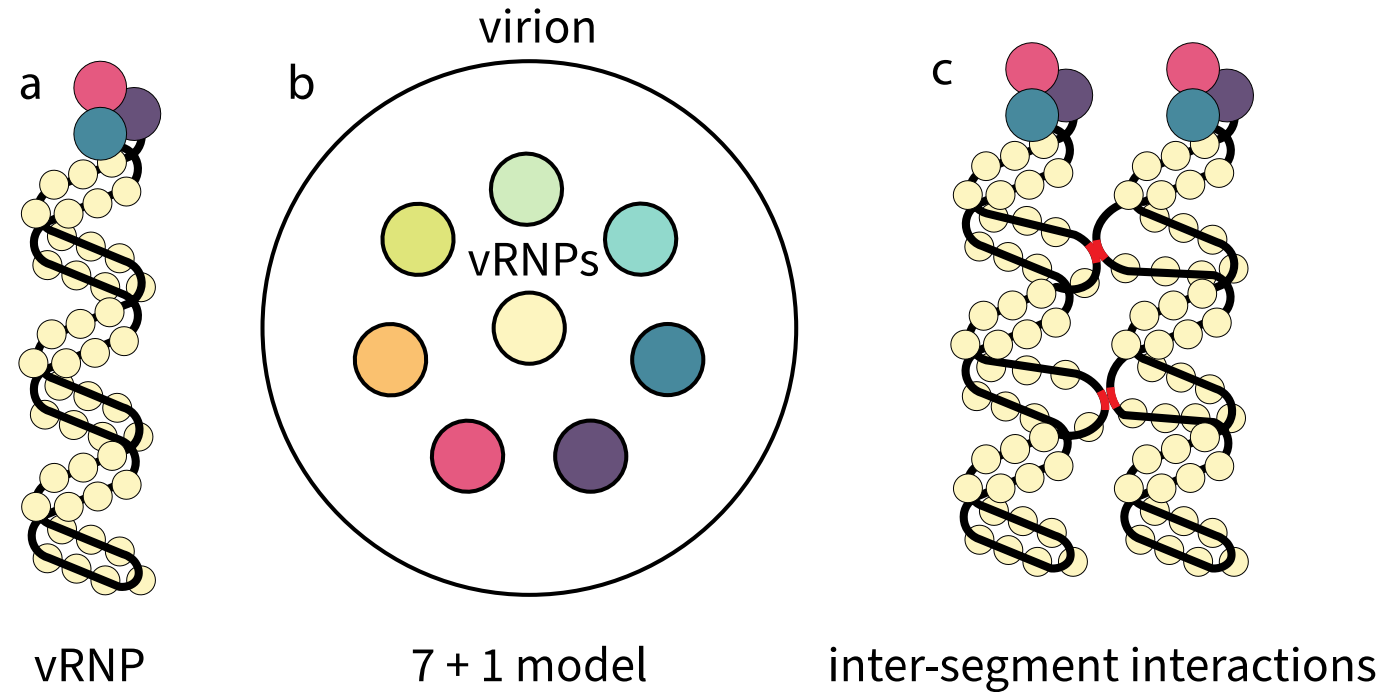
Differential analysis and *de novo* annotation of RNA-RNA interactions with RNAswarm

Gabriel Lencioni Lovate, Celia Jakob, Christian Höner zu Siederdissen, Kevin Lamkiewicz, Hardin Bolte, Martin Schwemmle, Manja Marz



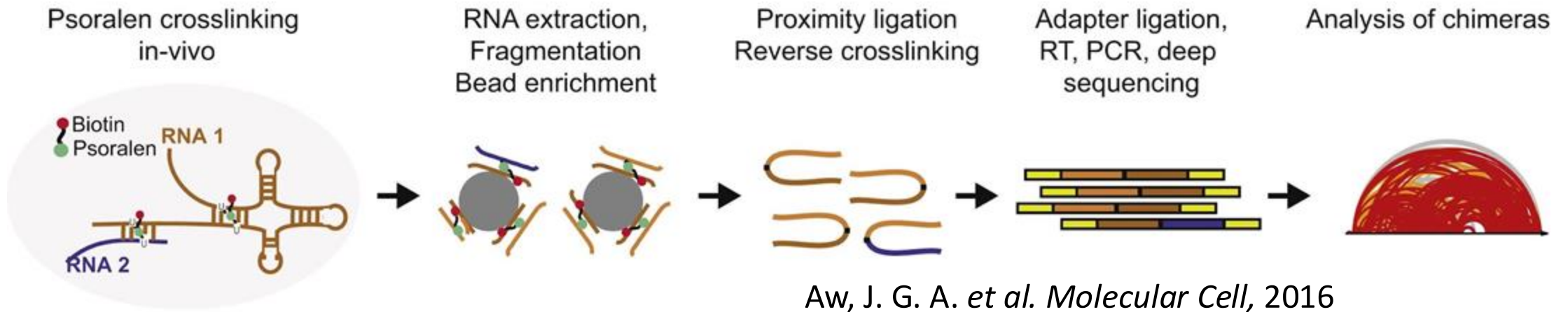
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Our guiding thread: The Influenza A Virus (IAV) genome packaging problem



- a. IAV's genome is organized in viral ribonucleoproteins (vRNPs)
- b. vRNPs are thought to be organized in a 7 + 1 model
- c. Exposed RNA are thought to play a crucial role in IAV genome packaging via RNA-RNA interactions

RNA-crosslinking methodologies can probe RNA-RNA interactions (RRIs)



Implementing a reproducible analysis pipeline

- Automate analysis of chimeras from read trimming to interaction mapping.

Differential analysis

- Evaluating significance of interactions.
- Comparing mutants and wild-type sequences.

preprocessing and mapping

SPLASH reads

```
@NDX550217:151:HFC
CCGGGATGGCGTGGAGC
+
AAAAAEEEEEEEEEEEEEE
@NDX550217:151:HFC7LBGX
CTCAGGTAACGGGTGGGTCGCG
+
AAAAAEEEEEEEEEEEEEE/EAEE
@NDX550217:151:HFC7LBGX
CCTCGGCTTGGCGGAATCAGCGG
+
AA/AAAAEEEEEEEEEEEEEE
```

.fastq

fastqc
fastp

ref. genome

```
>SC35M_PB2
AGCGAAAGCAGGTCAAT
CTAAGAGACCTAATGTCA
AAACCACTGTGGACCATATGGCCA
>SC35M_PB1
AGCGAAAGCAGGCAACCATTTGAA
CCAGCGCAAAATGCCATAAGCACCA
GGAACAGGAACAGGATACACCATGG
>SC35M_PA
AGCGAAAGCAGGTACTGATTCAAAA
ATCGTCGAGCTTGGGAAAAAGCAA
AATAAATTCGCTCAATATGCACAC
```

.fasta

segemehl index

segemehl map

mapped reads

```
@NDX550217:151:HFC
CCGGGATGGCGTGGAGC
+
AAAAAEEEEEEEEEEEEEE
@NDX550217:151:HFC7LBGX
CTCAGGTAACGGGTGGGTCGCG
+
AAAAAEEEEEEEEEEEEEE/EAEE
@NDX550217:151:HFC7LBGX
CCTCGGCTTGGCGGAATCAGCGG
+
AA/AAAAEEEEEEEEEEEEEE
```

.trns.txt

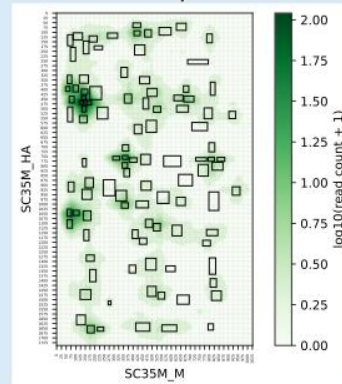
generating and annotating interaction matrices

fill_arrays.py

```
01001001 00100000
01100001 01101101
00100000 01110110
01100101 01110010 01111001
00100000 01100011 01101111
01101110 01100110 01110101
01110011 01100101 01100100
00101100 00100000 01001001
00100000 01110100 01101000
01101001 01101110 01101011
00100000 01001001 00100000
01110011 01101000 01101111
01110101 01101100 01100100
```

.npz

plot_heatmaps.py



annotate_interactions.py

```
id segment01 @type
1 SC35M_HA wt 100
2 SC35M_HA wt 100
3 SC35M_HA 8xmut 375 425
4 SC35M_HA wt 425 500
5 SC35M_HA wt 400 475
6 SC35M_HA wt 375 450
7 SC35M_HA wt 450 525
8 SC35M_HA wt 425 475
9 SC35M_HA wt 575 625
10 SC35M_HA wt 725 800
11 SC35M_HA wt 725 800
```

.csv

differential analysis

make_counttable.py

```
SC35M_4xHA_repL182
1 0 0
2 0 0
3 0 1
4 0 0
5 0 0
6 0 1
7 0 1
8 0 0
9 0 0
10 2 0
11 0 2
```

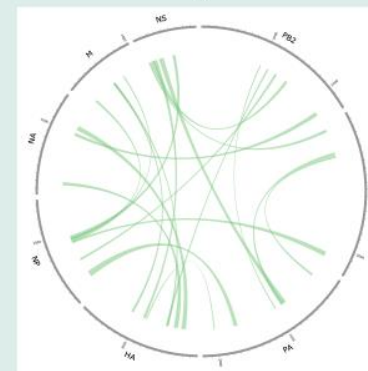
.csv

run_DESeq2.r

```
baseMean log2FoldC
1489 0.891945316252
636 0.891945316252
926 0.891945316252041 -3
1142 1.00866817177982 3
1695 1.18920711508272 -3
1636 0.891945316252041 -3
1506 1.45437215794768 -2
1074 0.594603557501361 -2
13 0.594603557501361 -2
1301 0.594603557501361 -2
1398 0.594603557501361 -2
```

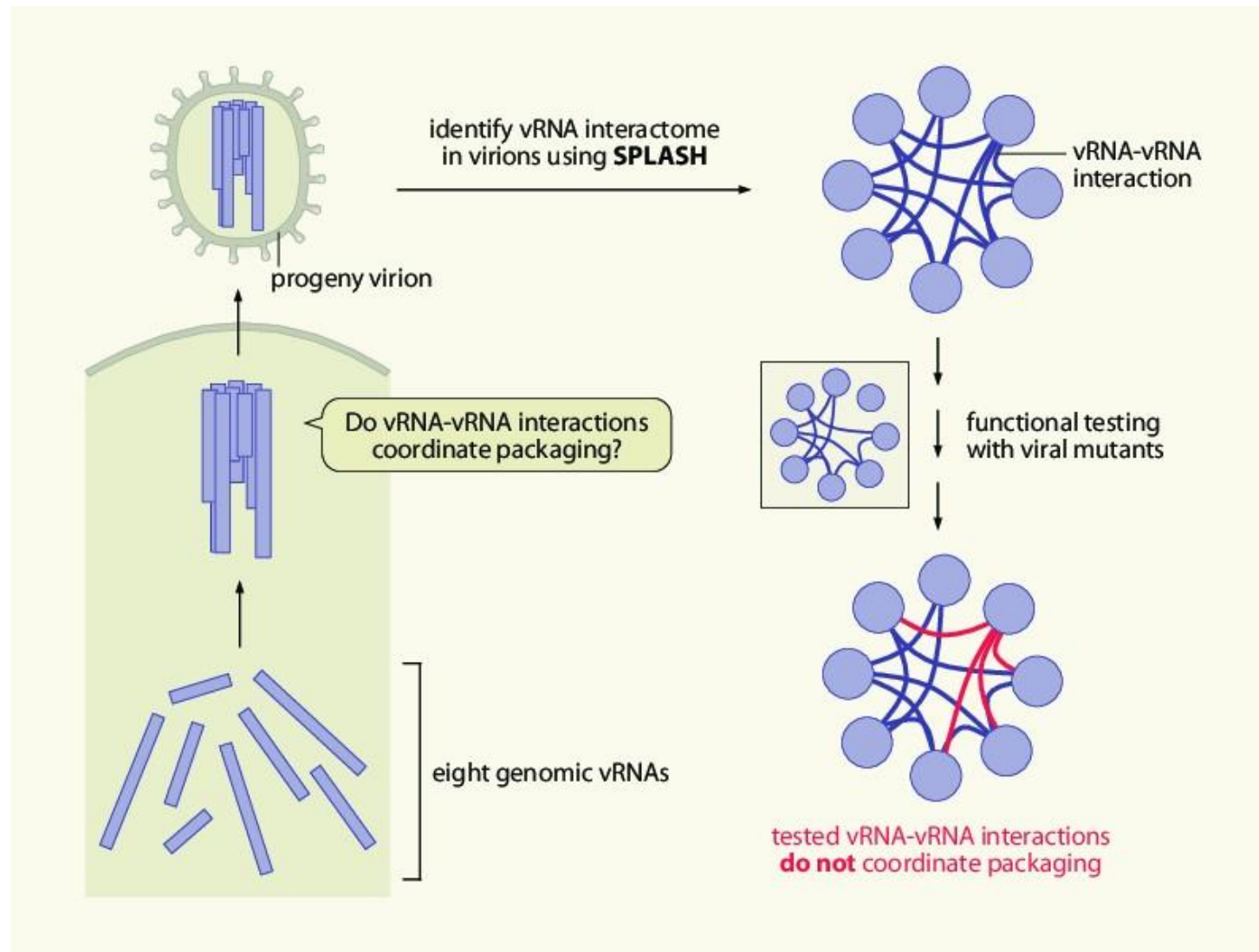
.csv

make_circos_files.py

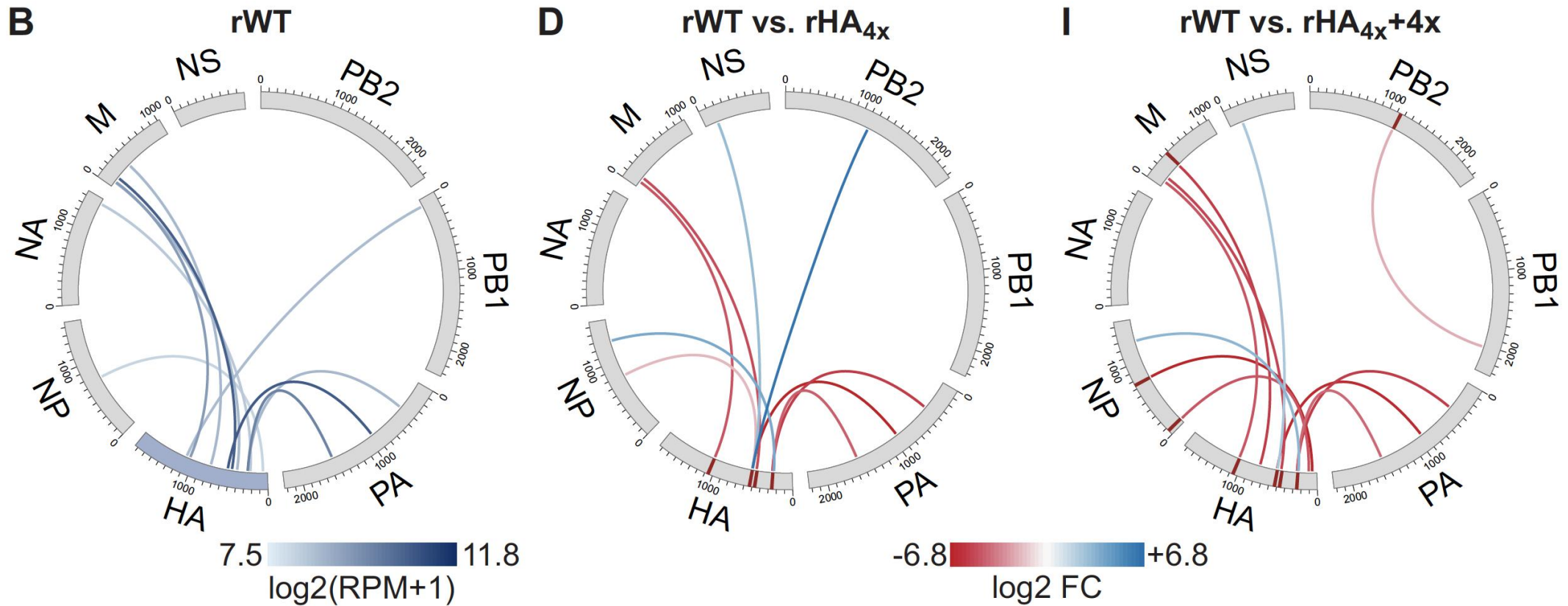


RNAswarm performs differential analyses of SPLASH datasets

Are high-high frequency cross-linked reads packaging signals?

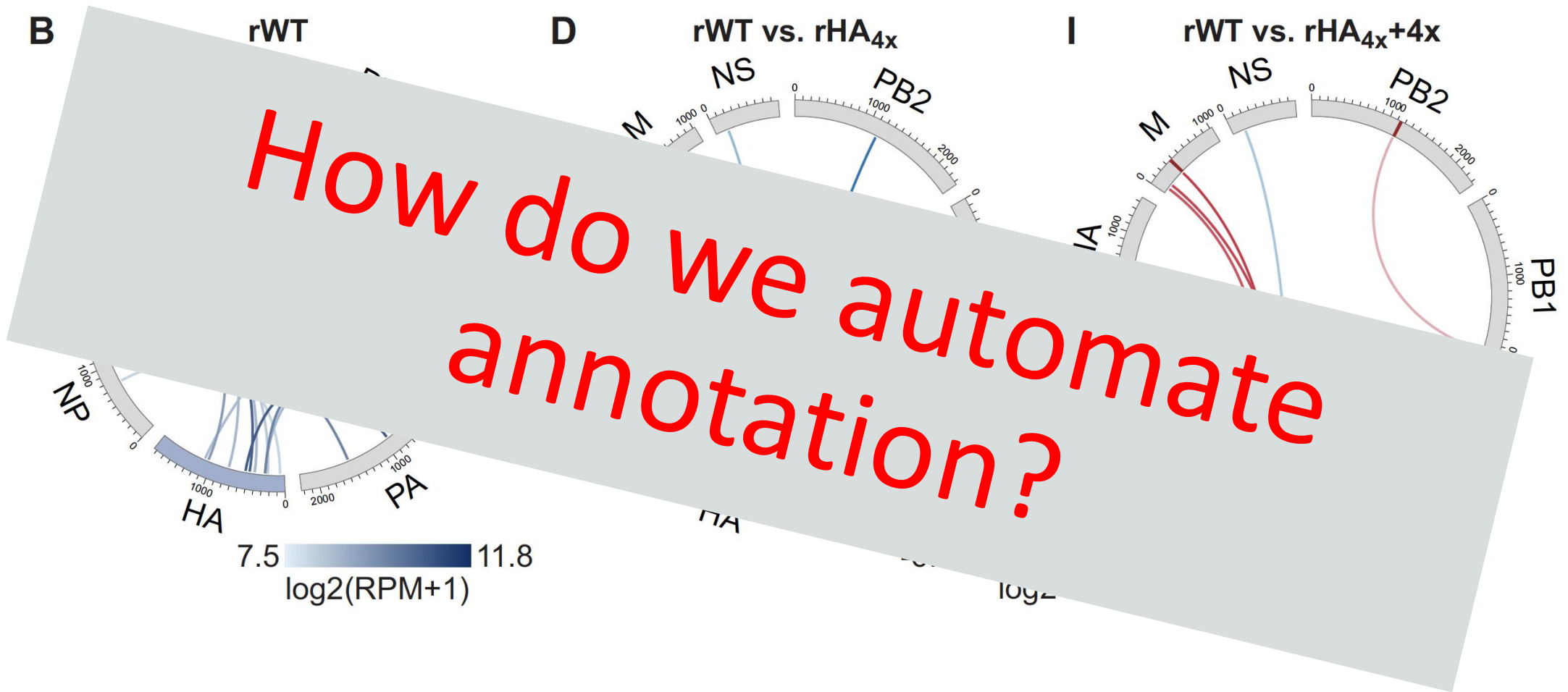


Mutations do not impair packaging of SC35M



Jakob et al. *Nucleic Acids Research*, 2023

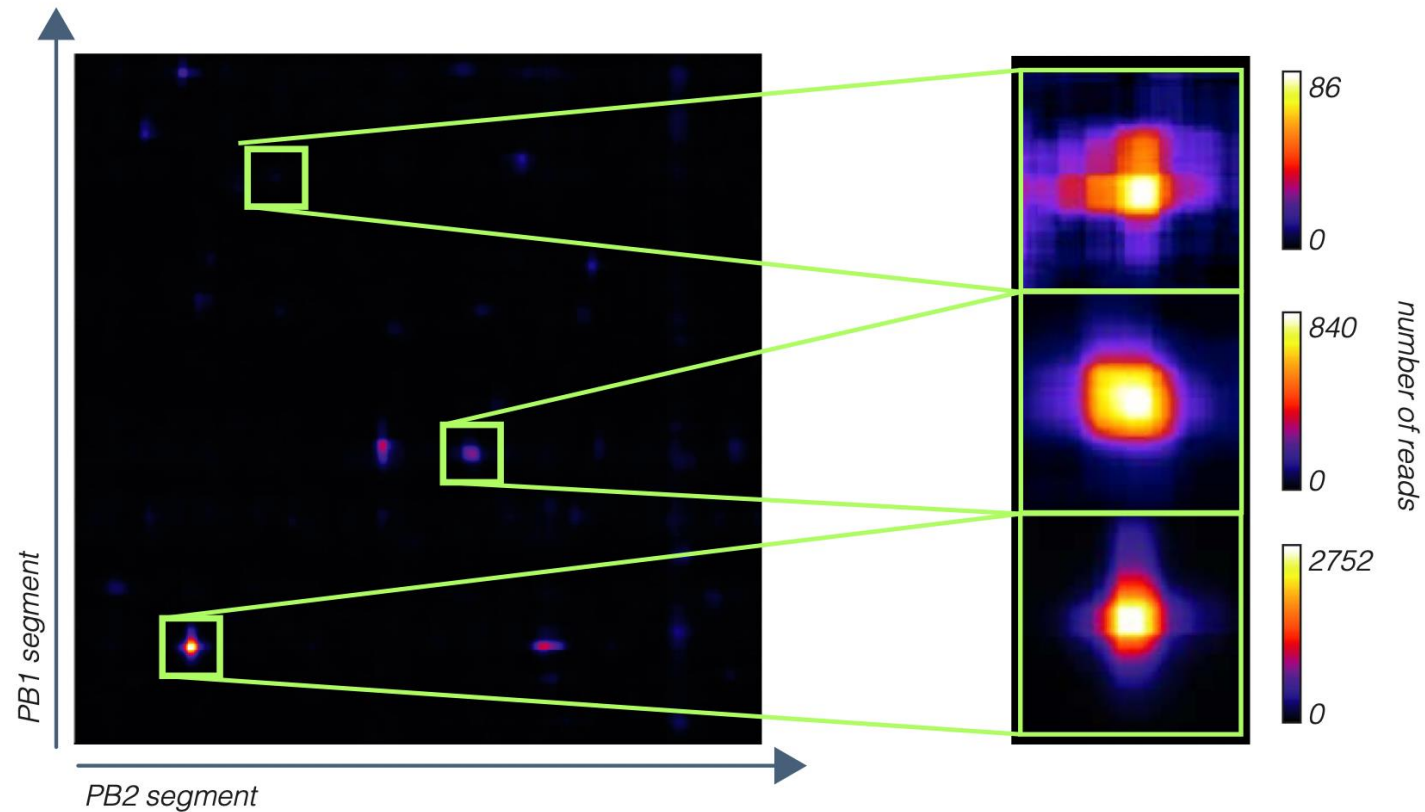
Mutations do not impair packaging of SC35M



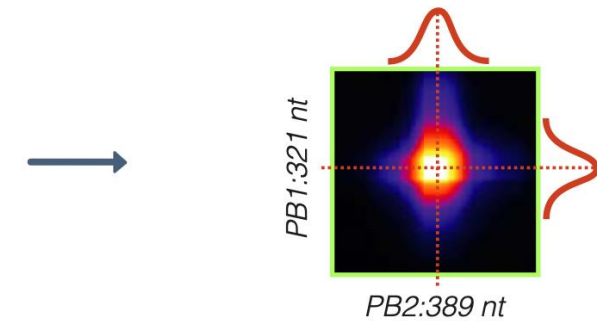
Interaction matrices can be annotated by fitting a 2d gaussian function to hand-picked quadrants

SPLASH identification of interaction loci

1. Visualise interaction matrices between each pair of segments as a heatmap and select loci of interaction



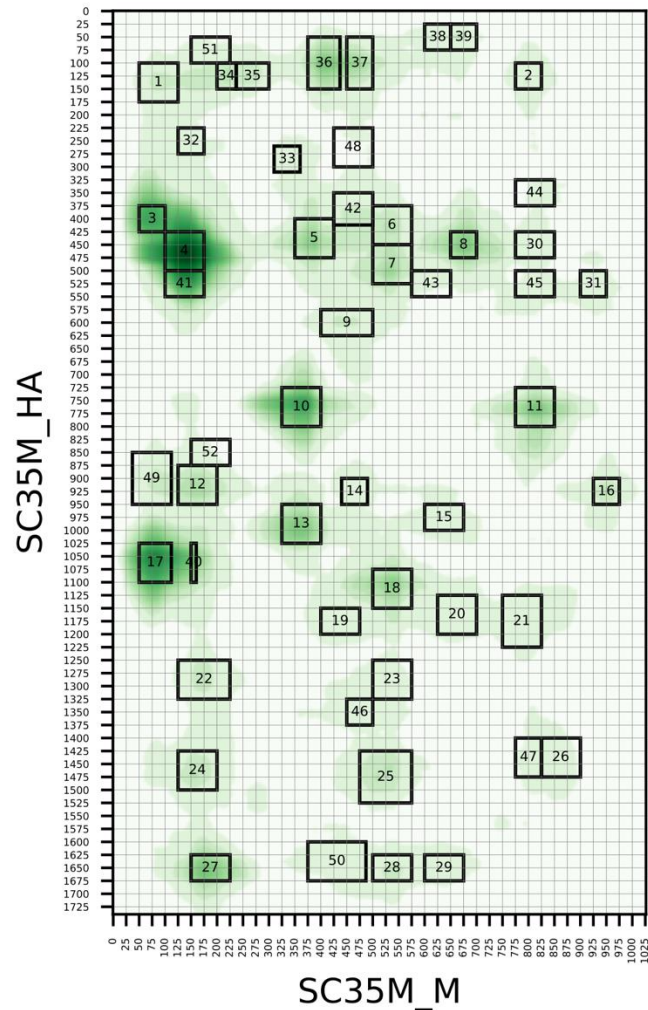
2. Fit a 2D Gaussian function to each locus to extract coordinates of interaction



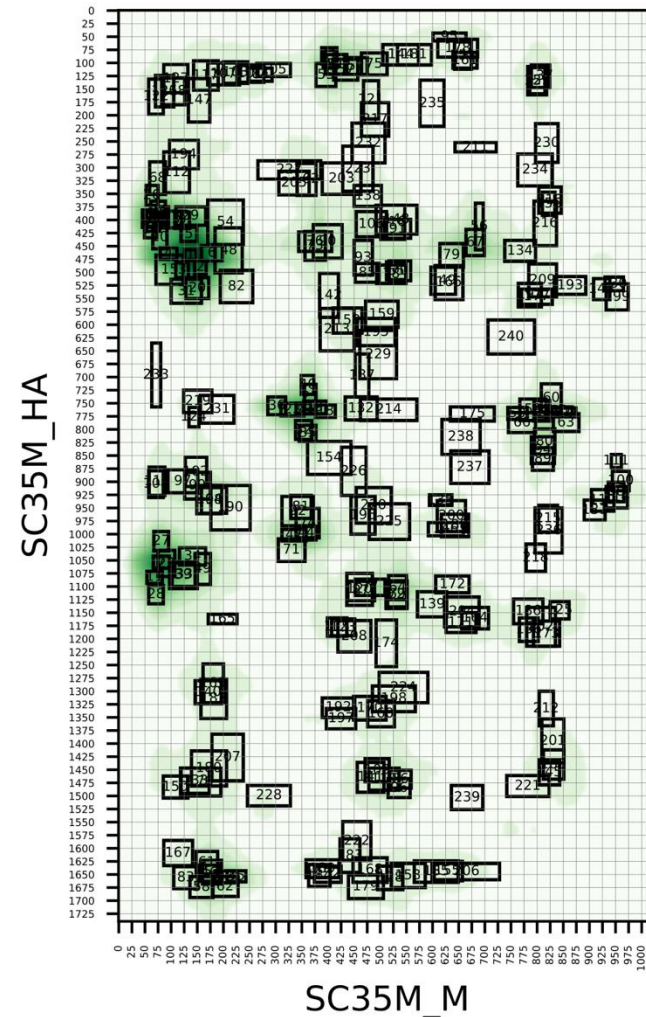
Dadonaite et al. *Nat. Microbiol* (2019)

We can use heuristics to deduplicate the annotations

manual curation

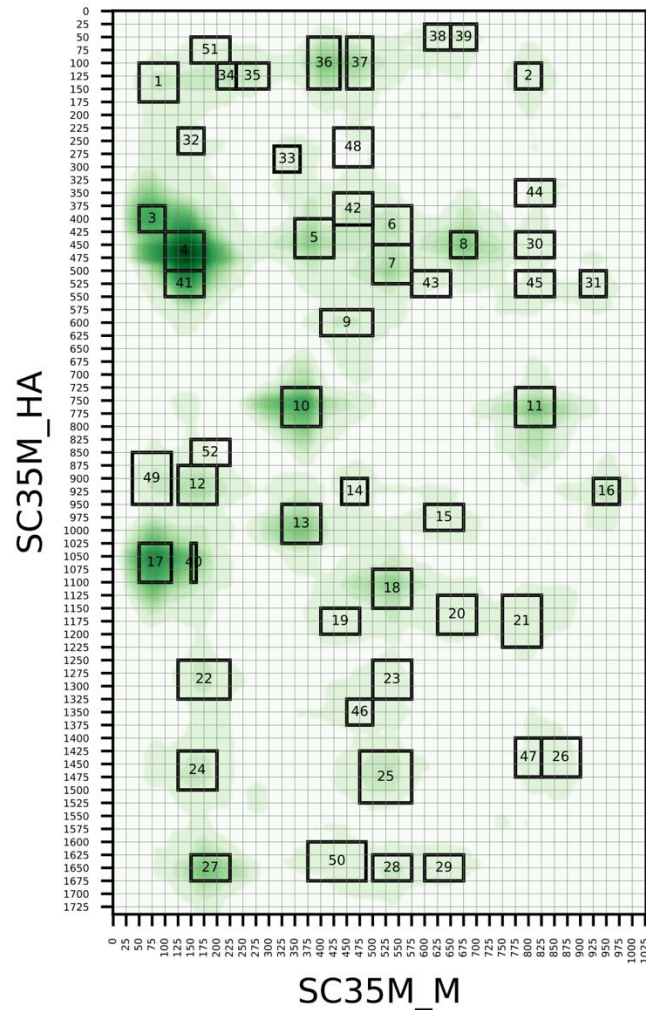


automated

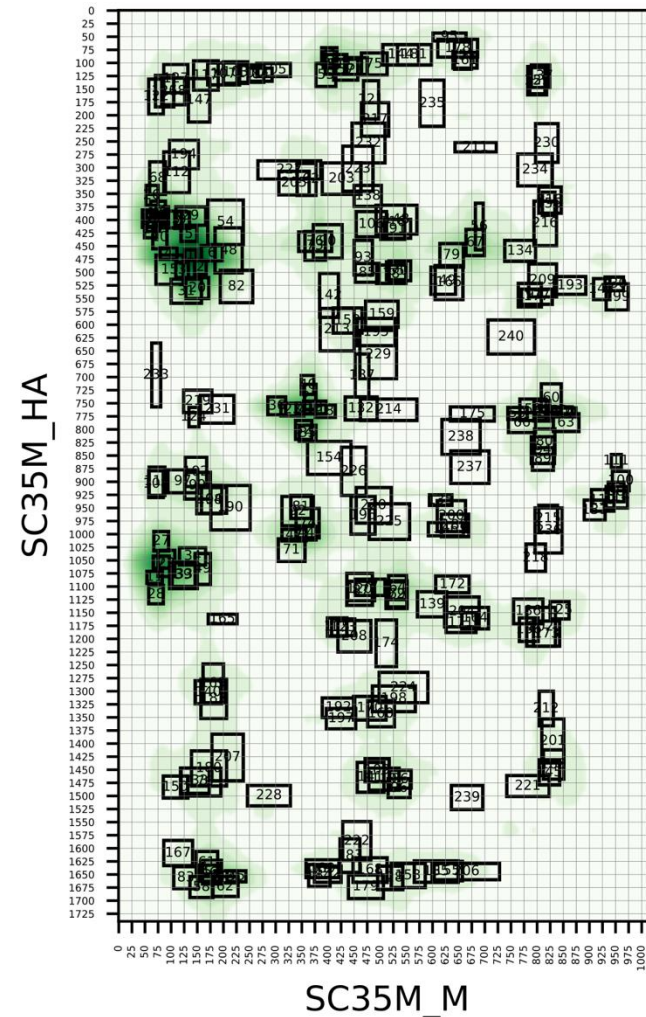


We can use heuristics to deduplicate the annotations

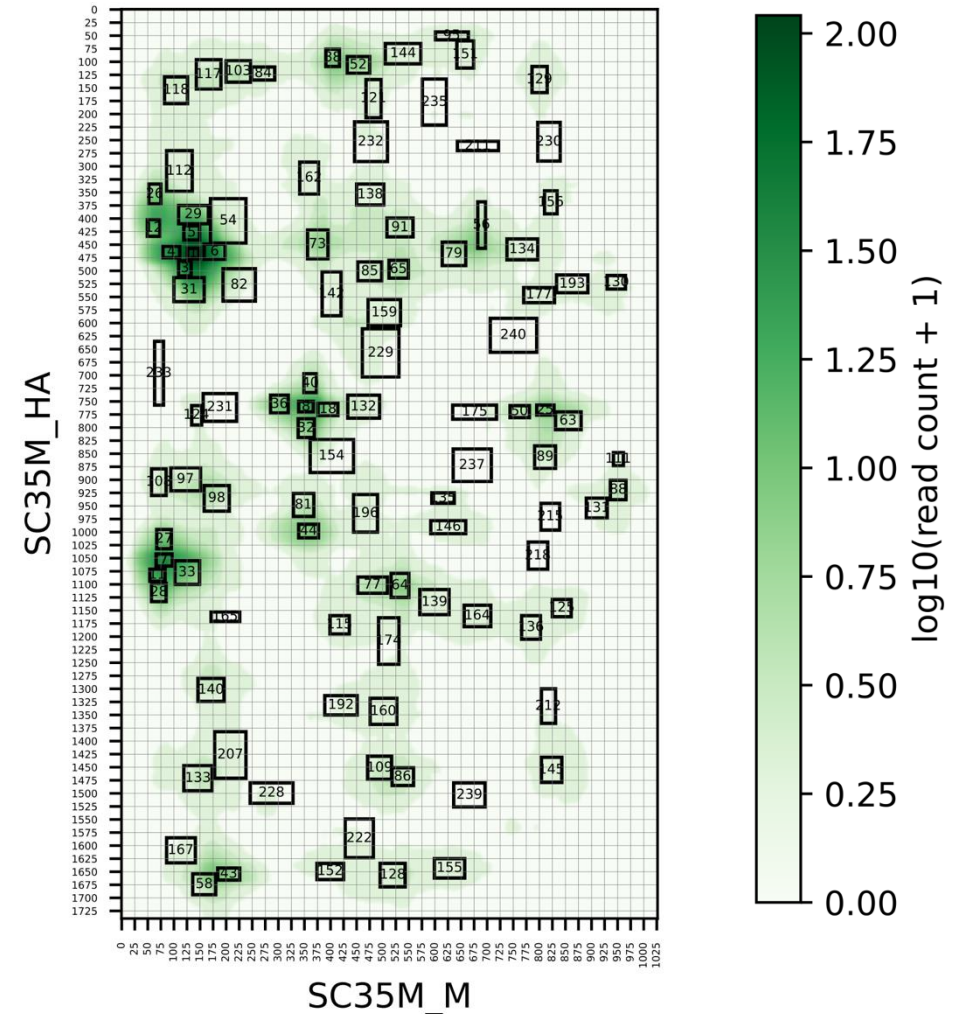
manual curation



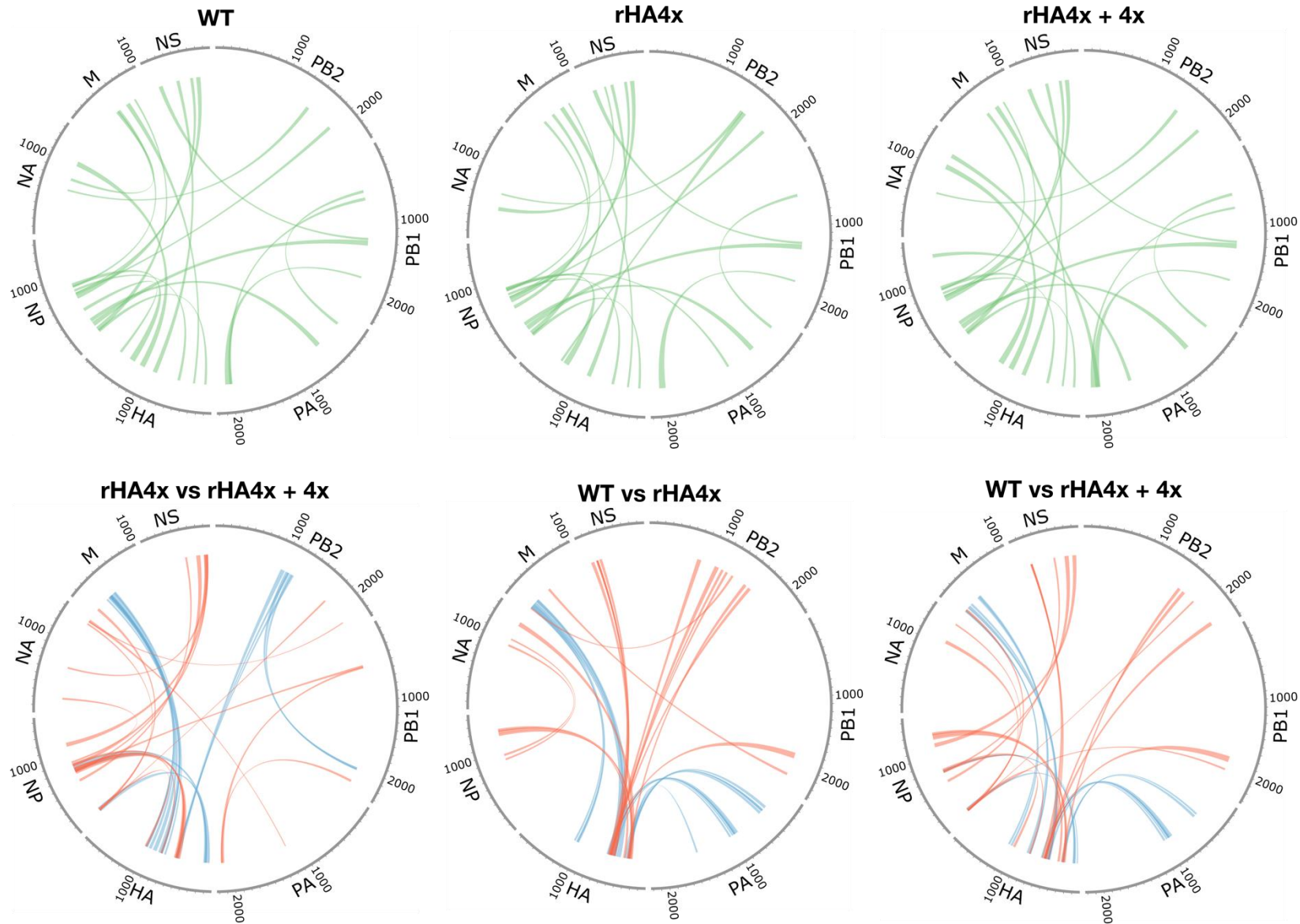
automated



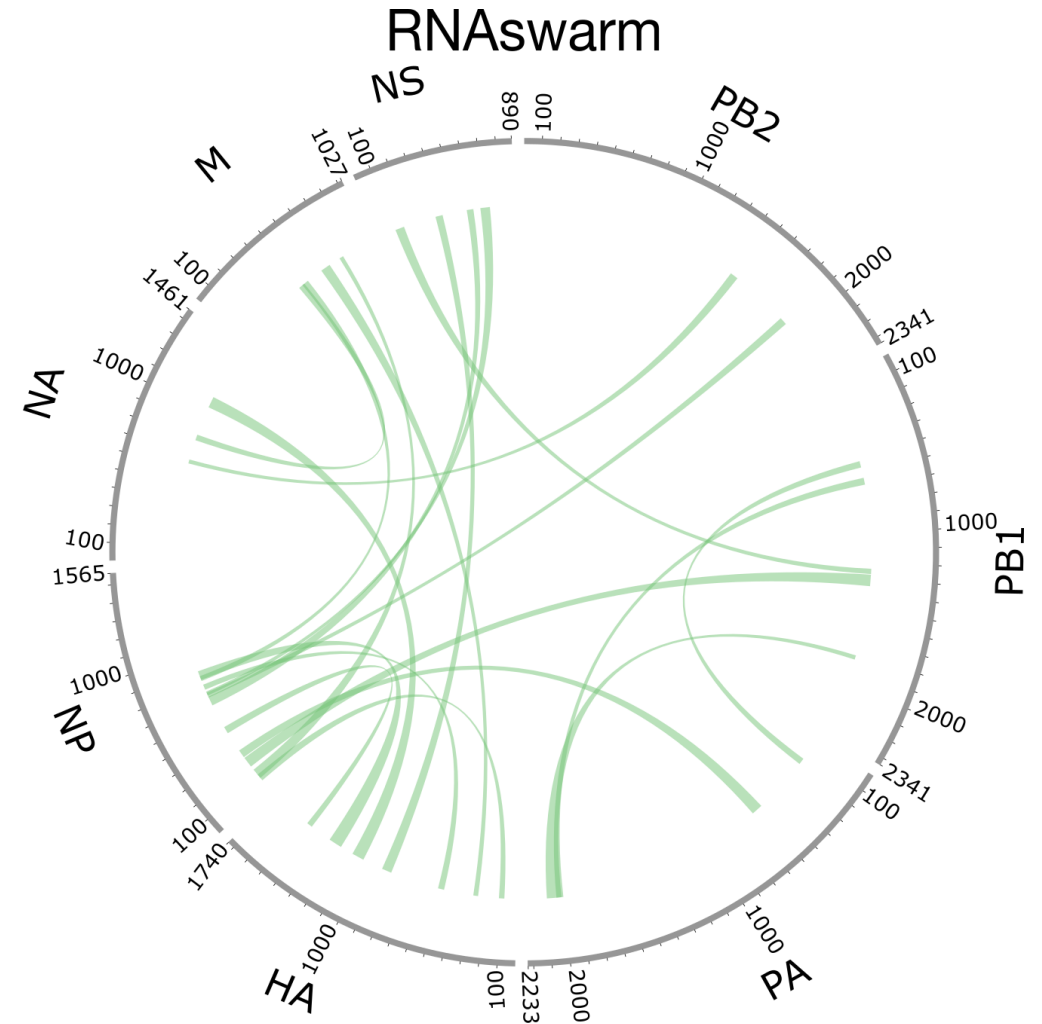
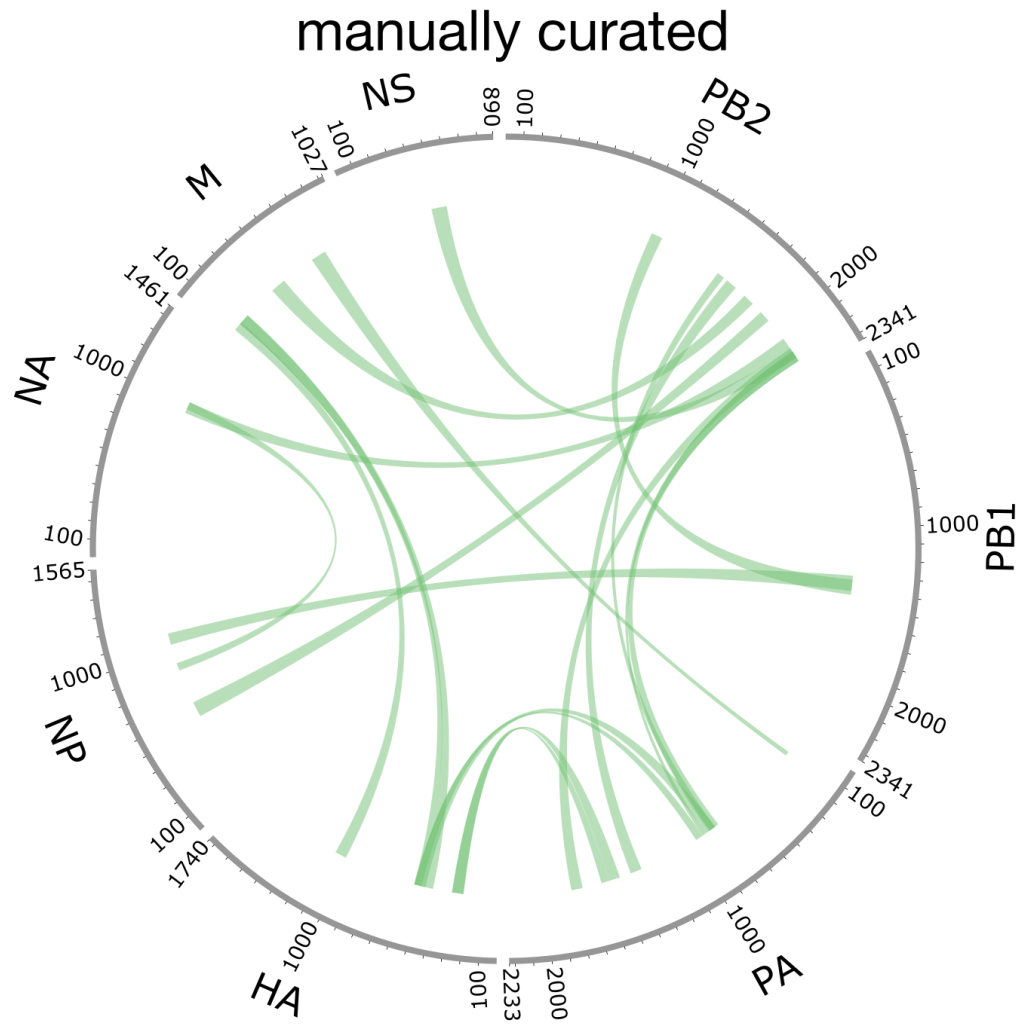
automated w/
deduplication



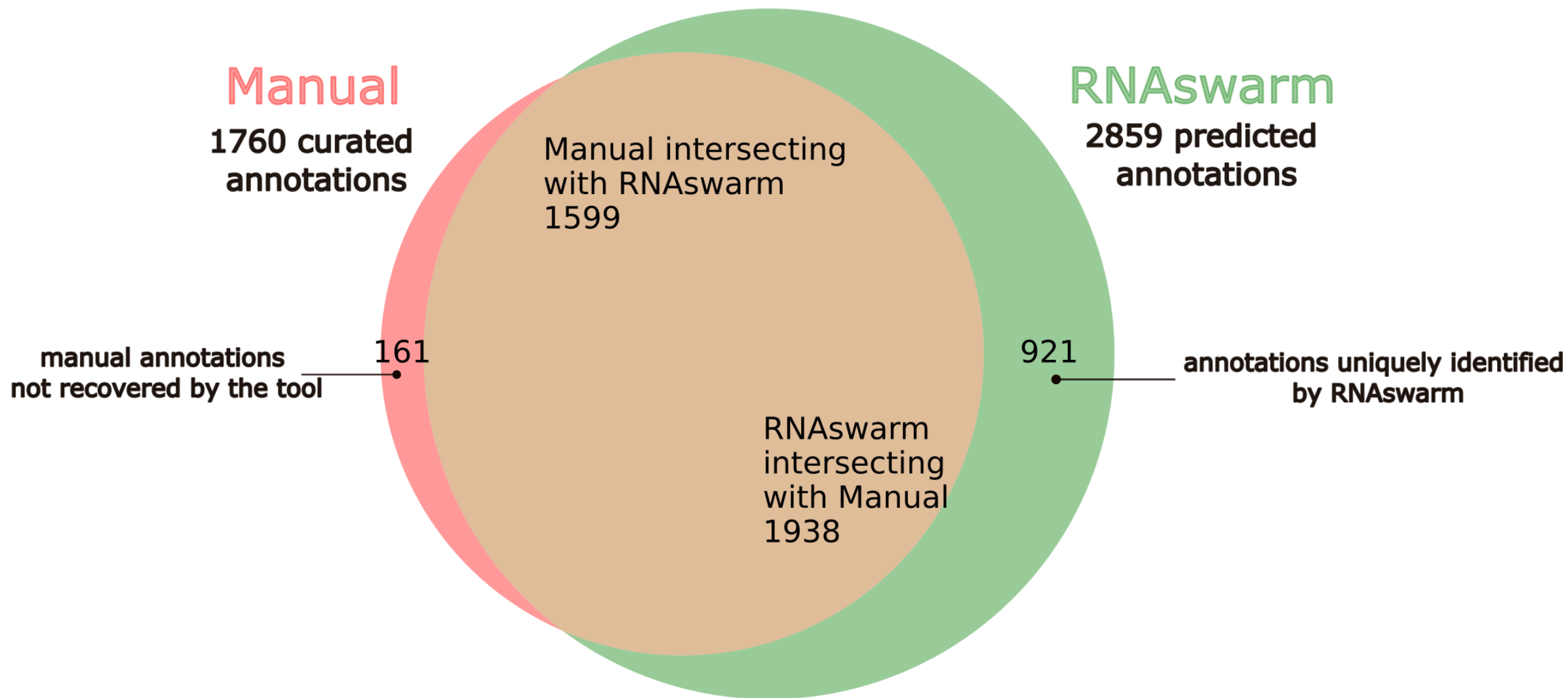
RNAswarm *de novo* annotations



Manual vs RNAswarm



Manual vs RNAswarm



So what do I need to use RNAswarm?

- Input: fasta and fastq files
- Output: Interaction matrices, heatmaps, differential interaction table, and circos plots

And what can I do with RNAswarm?

- RNAswarm can deal with data resulting from a broad range of proximity ligation methods (not only SPLASH)
- We can identify discrete interactions in an unsupervised fashion using GMMs
- We can compare the prevalence of discrete interactions across conditions using DESeq2

With a Little Help from My Friends



Manja Marz's Group

Christian Höner zu Siederdisen
Kevin Lamkiewicz

Martin Schwemmle's group

Hardin Bolte
Celia Jakob

Research funded by EU's Horizon 2020 program, under the MSCA ITN grant agreement no. 955974

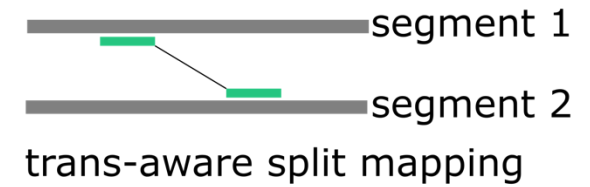
The lost slides

Matrix filling

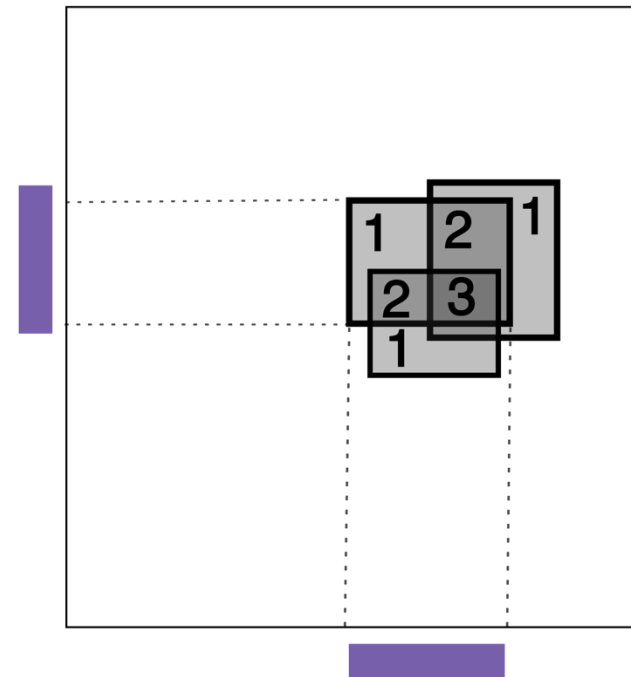
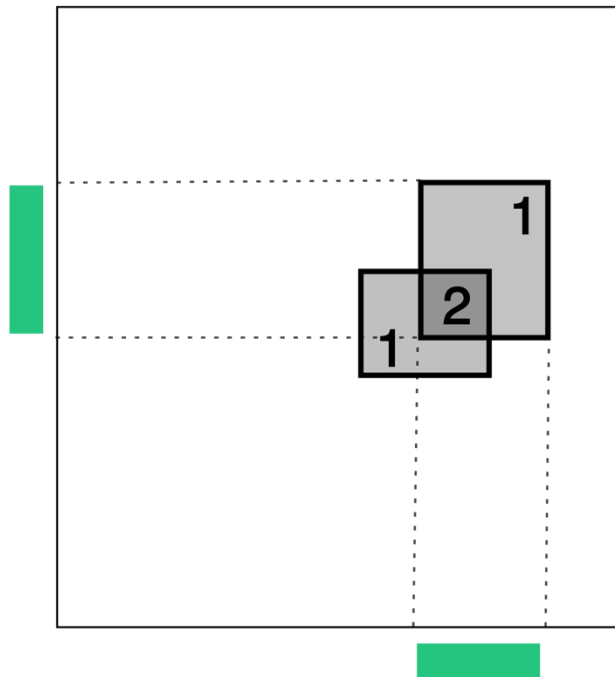
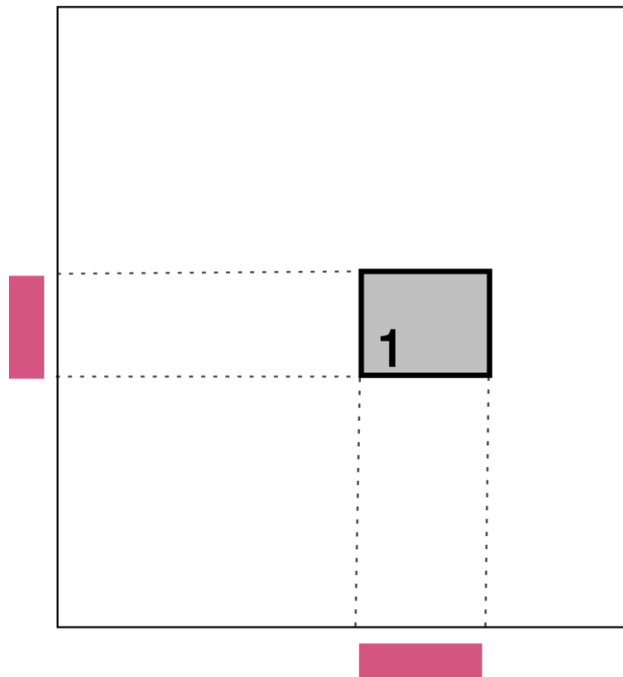
Reads



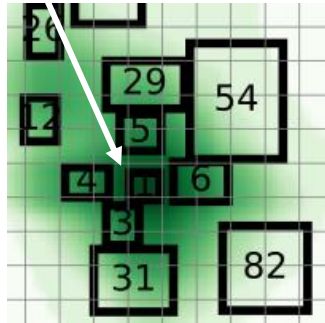
Read Mapping



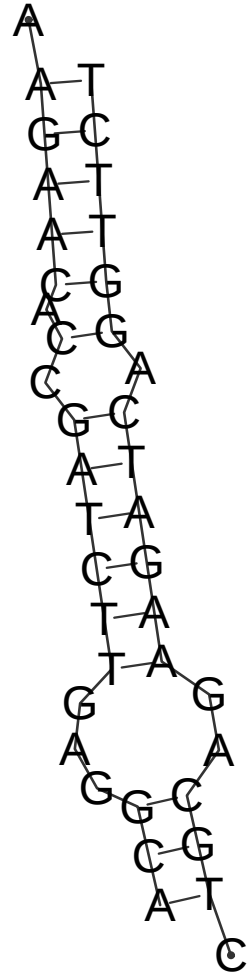
Matrix Filling



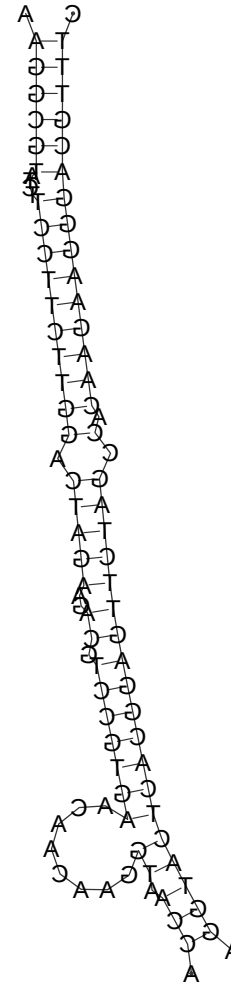
Structure prediction using RNAduplex (or RNAcofold)



Identified region



25-nt extended



Refining annotations with structure predictions

Raw	.(((((.((((((...(((	&	.)))..))))))..))))))
Extend 5 nt	.(((((((((((.(((((((((((((((	&	.)))))))).))..))))))..))))))
Extend 10 nt	.(((((((((((.(((((((((((((((	&	.)))))))).))..))))))..))))))
Extend 15 nt	.(((((((((((.(((((((((((((((	&	.)))))))).))..))))))..))))))
Extend 20 nt	.(((((((((((.(((((((((((((((	&	.)))..)).....)))))))).))..))))))..))))))
Sequence	CTTTGCAGGGAAGAACACCGATCTTGAGGCACTCATGGA	&	ACCAATGGAACAACAAGTGCCTGCAGAAGATCAGGTTCTTCCTTCTATGCGGAA

Gaussian Mixture Models (GMMs) can be used to identify discrete interactions

Independent from user intervention

Can be fit to the data using an expectation-maximization algorithm:

1. Initialization: Assume random components.
2. Expectation Step (E-step): Compute for each point a probability of being generated by each component of the model.
3. Maximization Step (M-step): Update the parameters to maximize the likelihood of the data given those assignments.
4. Repeat 2 and 3 until convergence.

Pseudocode: deduplicate the annotations

Initialize bestRegions as an empty map

For each targetRegion in regions:

 Set highestRatio to 0

 Set bestRegion to None

 For each otherRegion in regions:

 If targetRegion intersects with otherRegion:

 Calculate ratio as readCount of otherRegion divided by area of otherRegion

 If ratio is greater than highestRatio:

 Update highestRatio to this new ratio

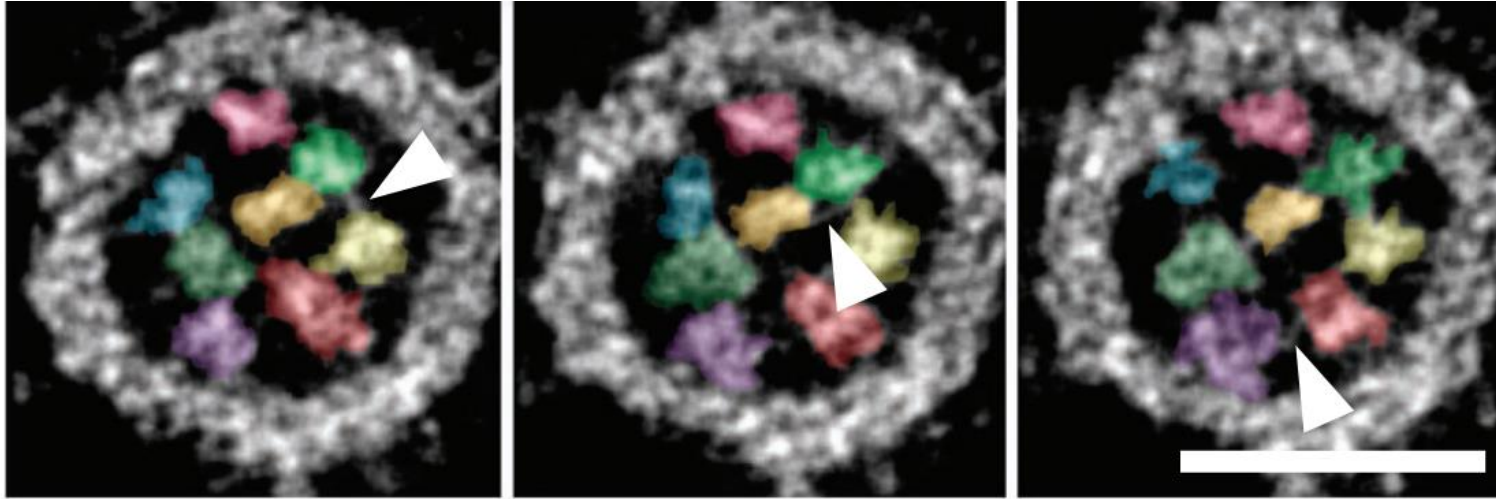
 Update bestRegion to otherRegion

 Assign bestRegion as the optimal choice for targetRegion in bestRegions map

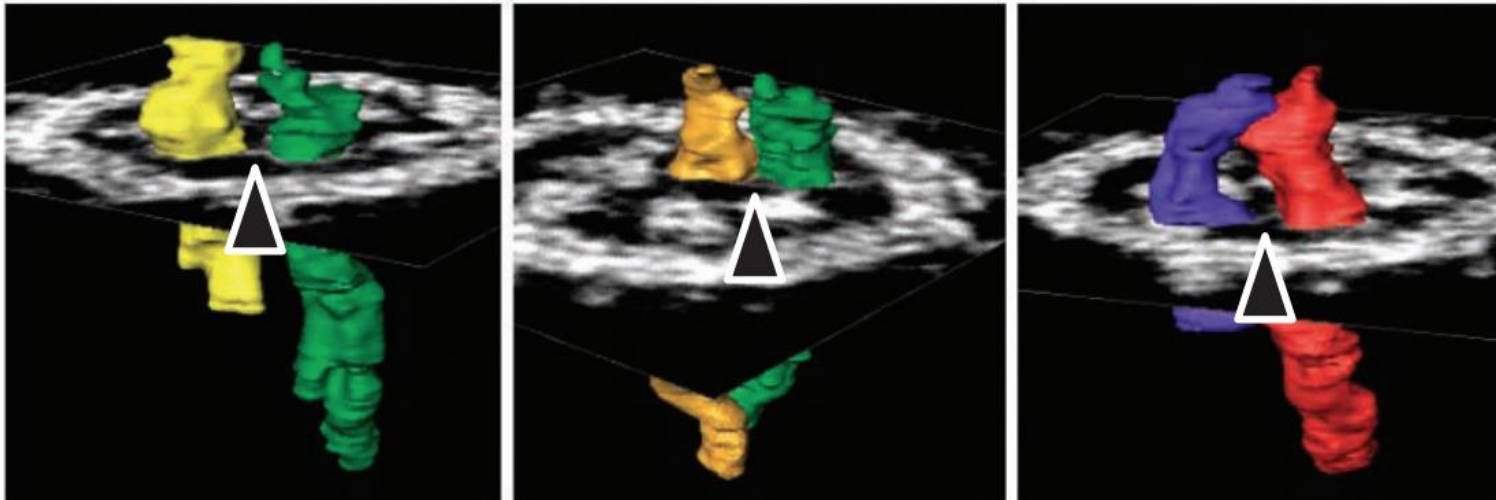
Return bestRegions

The IAV genome packaging problem

Tomogram



Tomogram
with
3-D model

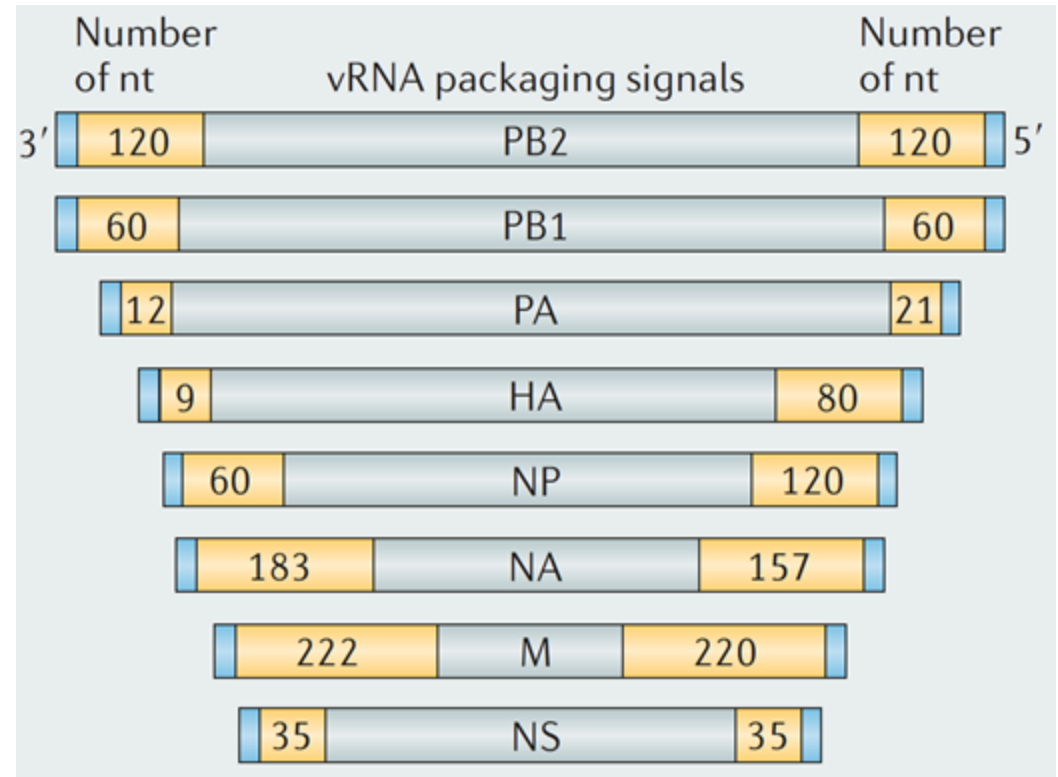


- Genome topology of IAV supports the 7 + 1 model
- vRNPs connected by a string-like structure

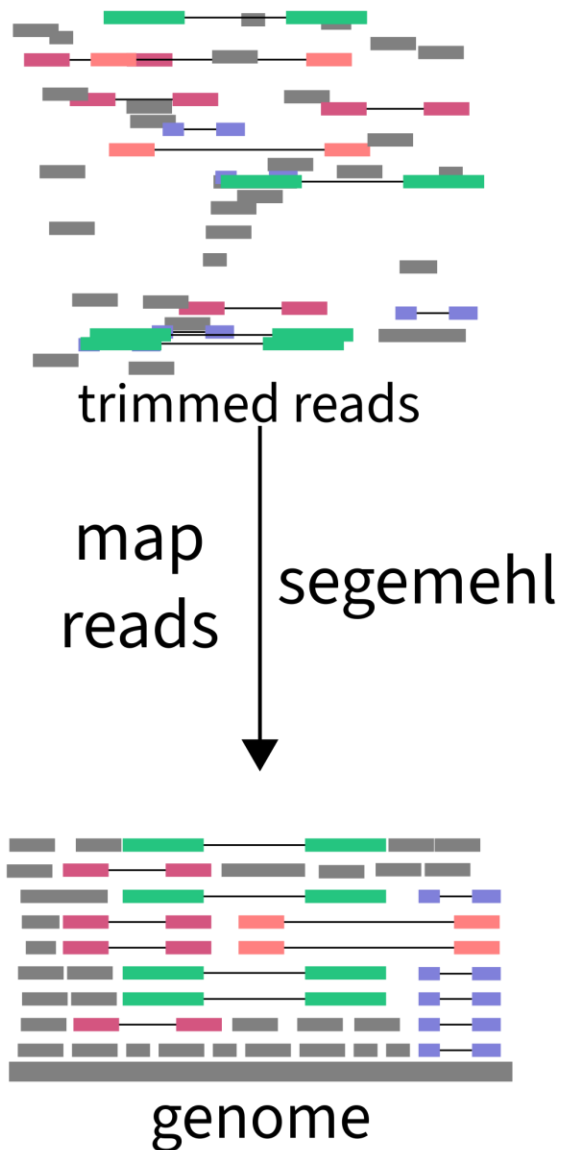
Known vRNA packaging signals

- Concentrated on 3' and 5'
- Internal signals are not the same in all strains

Eisfeld, Amie J. et al. Nature Reviews Microbiology 13, no. 1 (January 2015): 28–41.



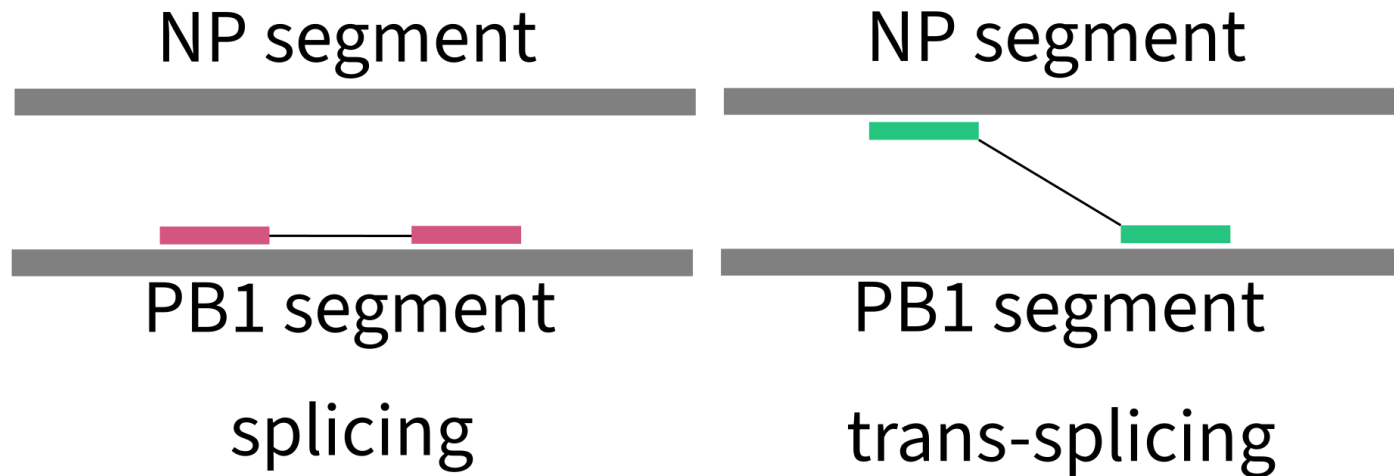
Split-read mapping



- Reads are aligned to the viral genome
- Chimeric reads are splitted during mapping
- segemehl as default mapper

Finding chimeric reads

- We rely on tools built for mapping splicing-product RNAs
- Segemehl classify spliced reads in 3 categories:
 - BED files for single and multi splicing
 - Custom file for trans-spliced reads



Interaction matrices are processed to identify differentially structured regions in the genome.

