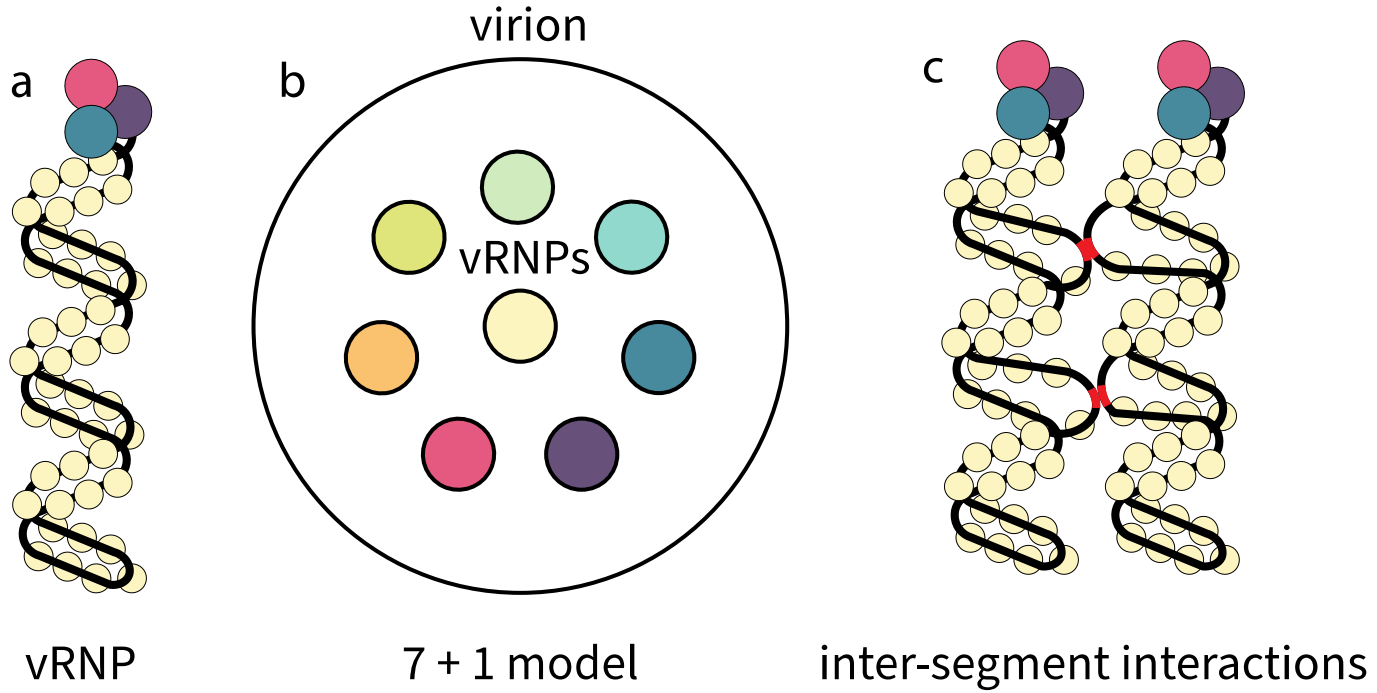
Differential analysis and *de novo* annotation of RNA-RNA interactions with RNAswarm

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

Understanding virus-host interactions

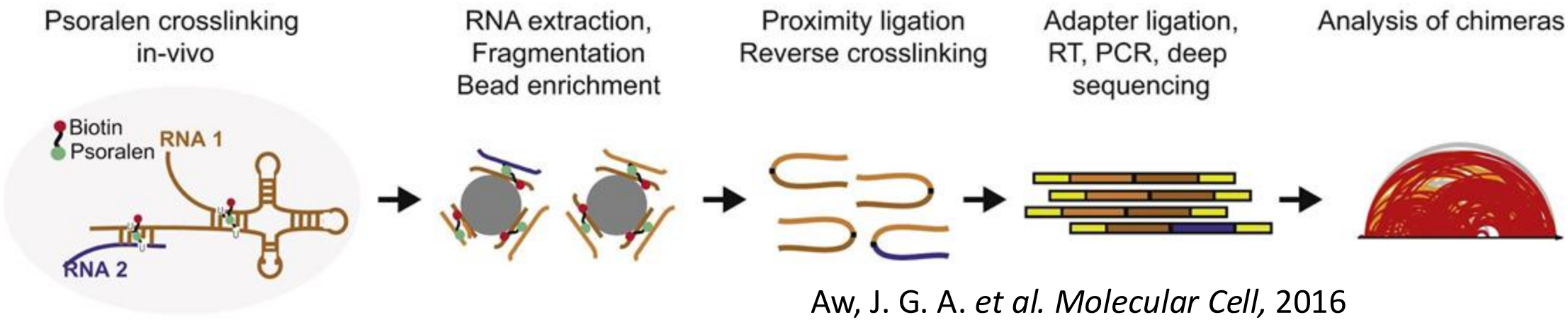
FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

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)



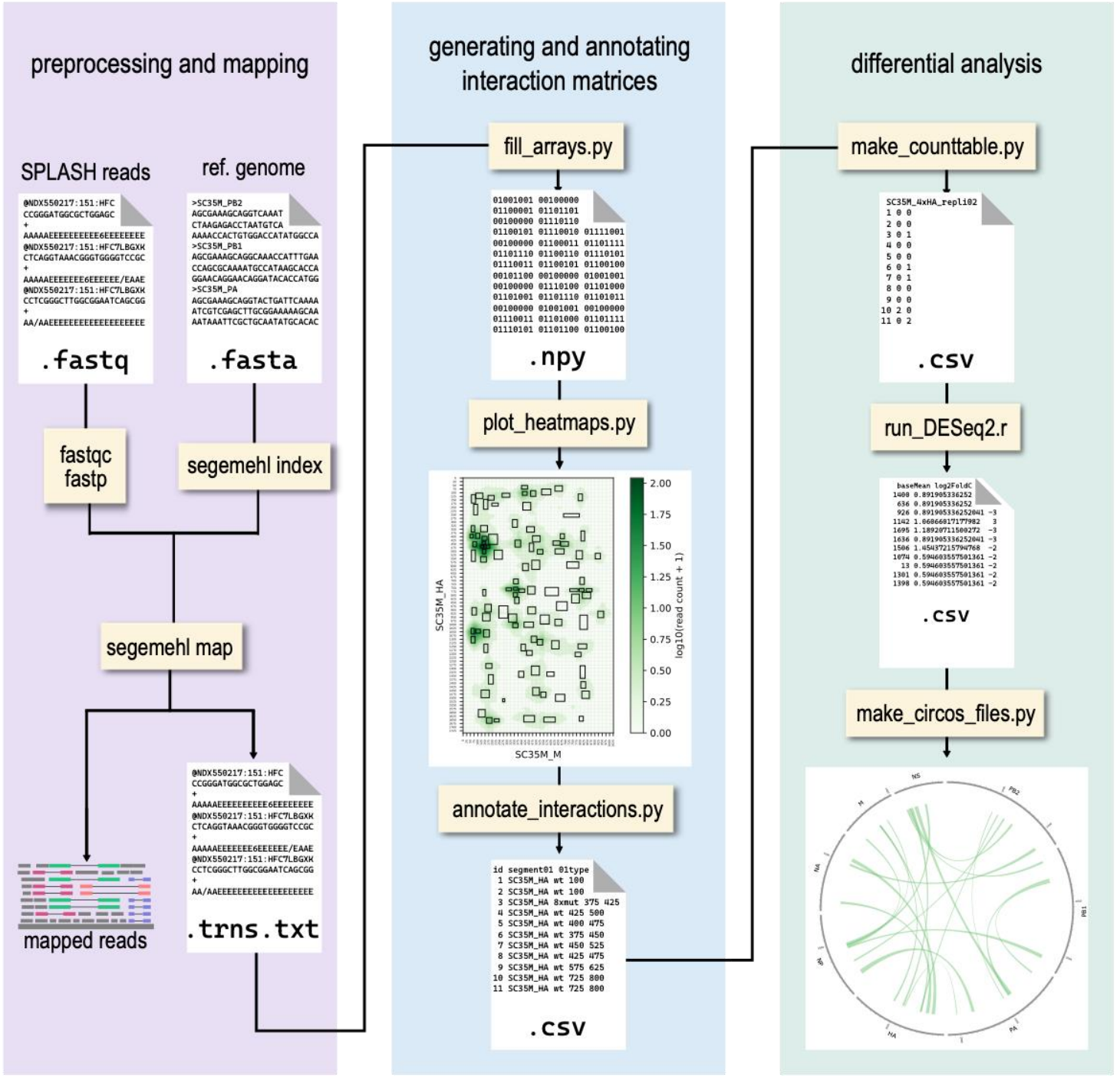
Aw, J. G. A. *et al. Molecular Cell,* 2016

Implementing a reproducible analysis pipeline

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

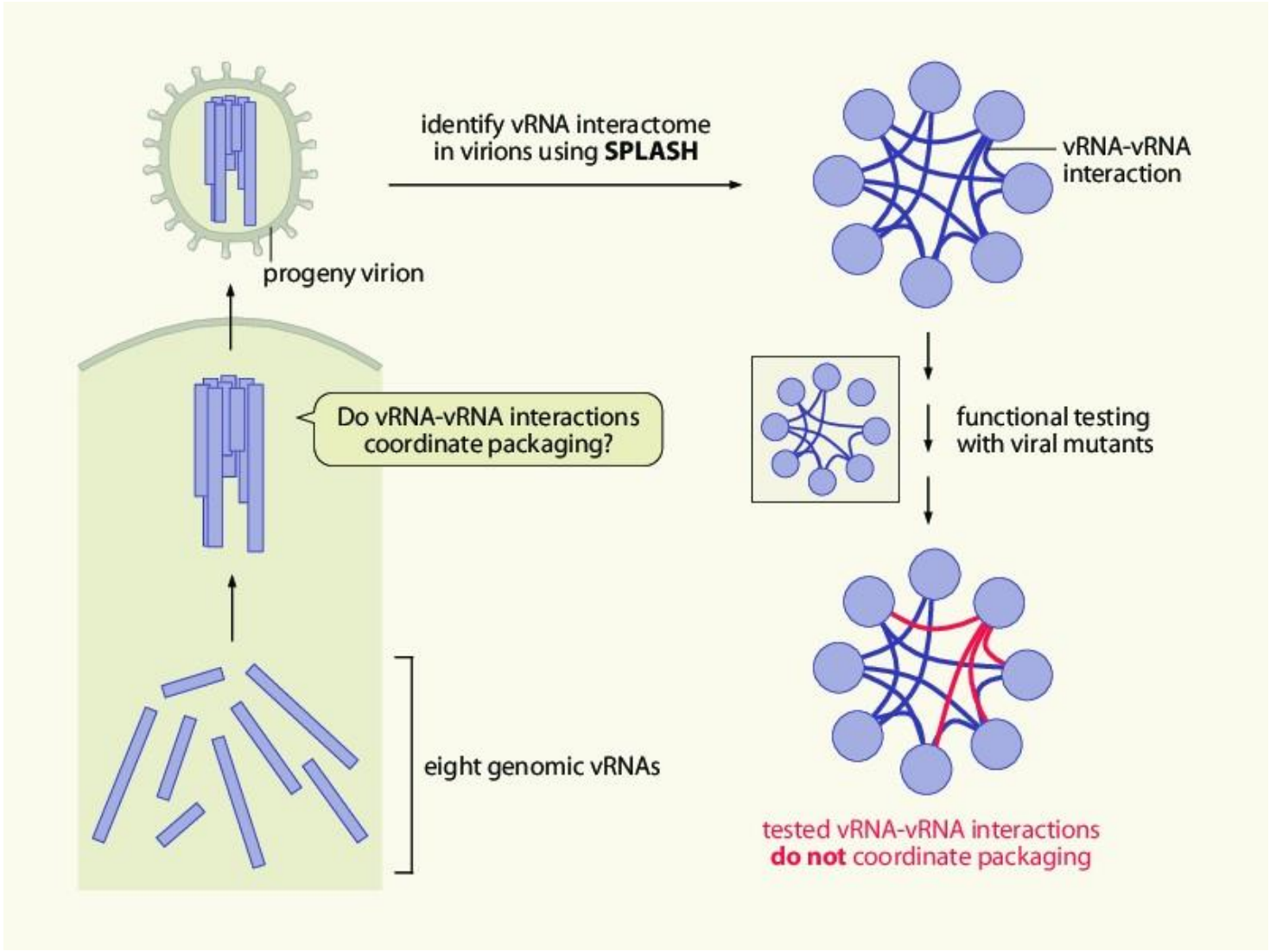
Differential analysis

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



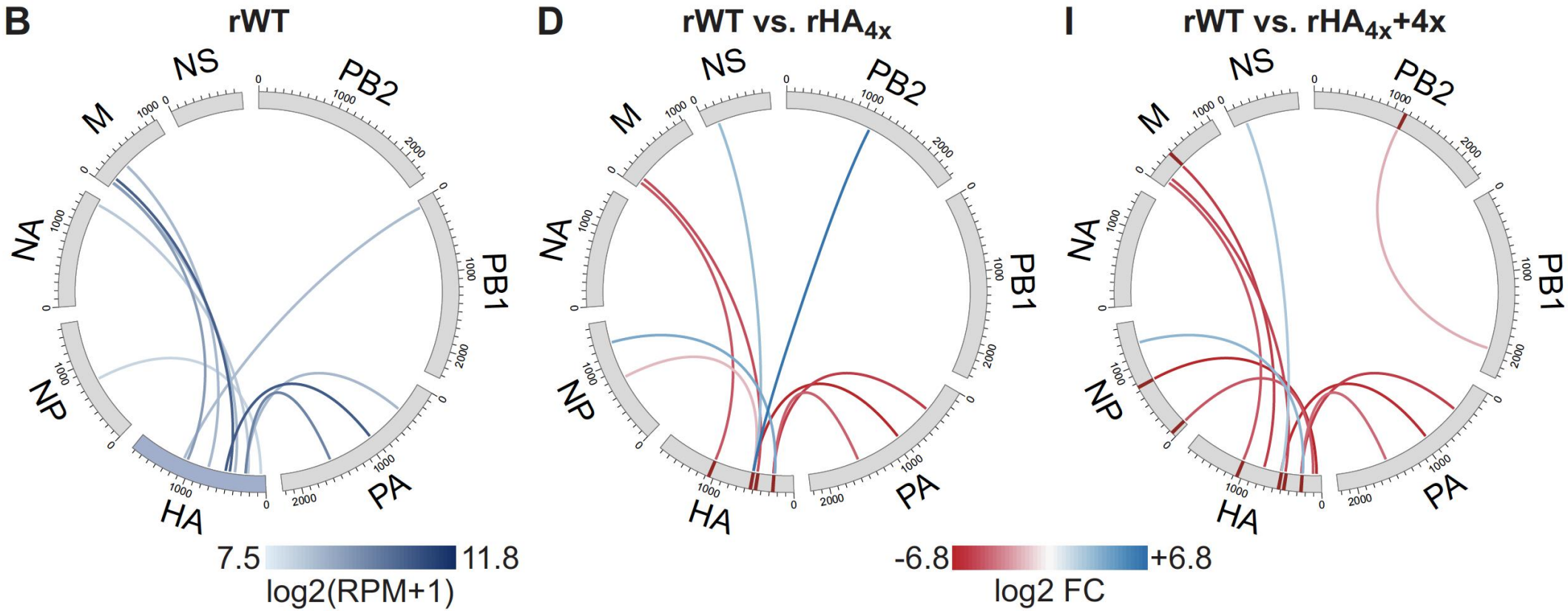
RNAswarm performs differential analyzes of SPLASH datasets

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



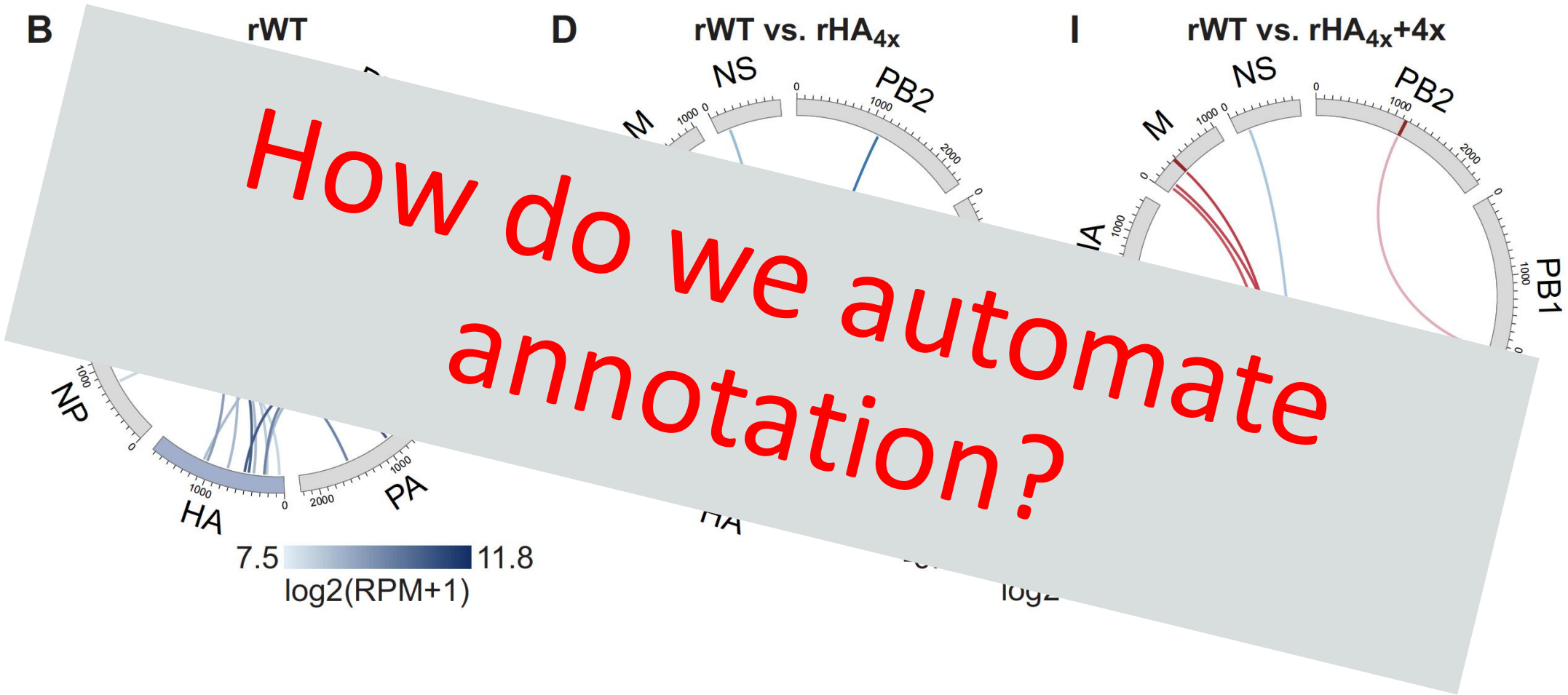
Jakob et al. *Nucleic Acids Research*, 2023

Mutations do not impair packaging of SC35M



Jakob et al. *Nucleic Acids Research*, 2023

Mutations do not impair packaging of SC35M



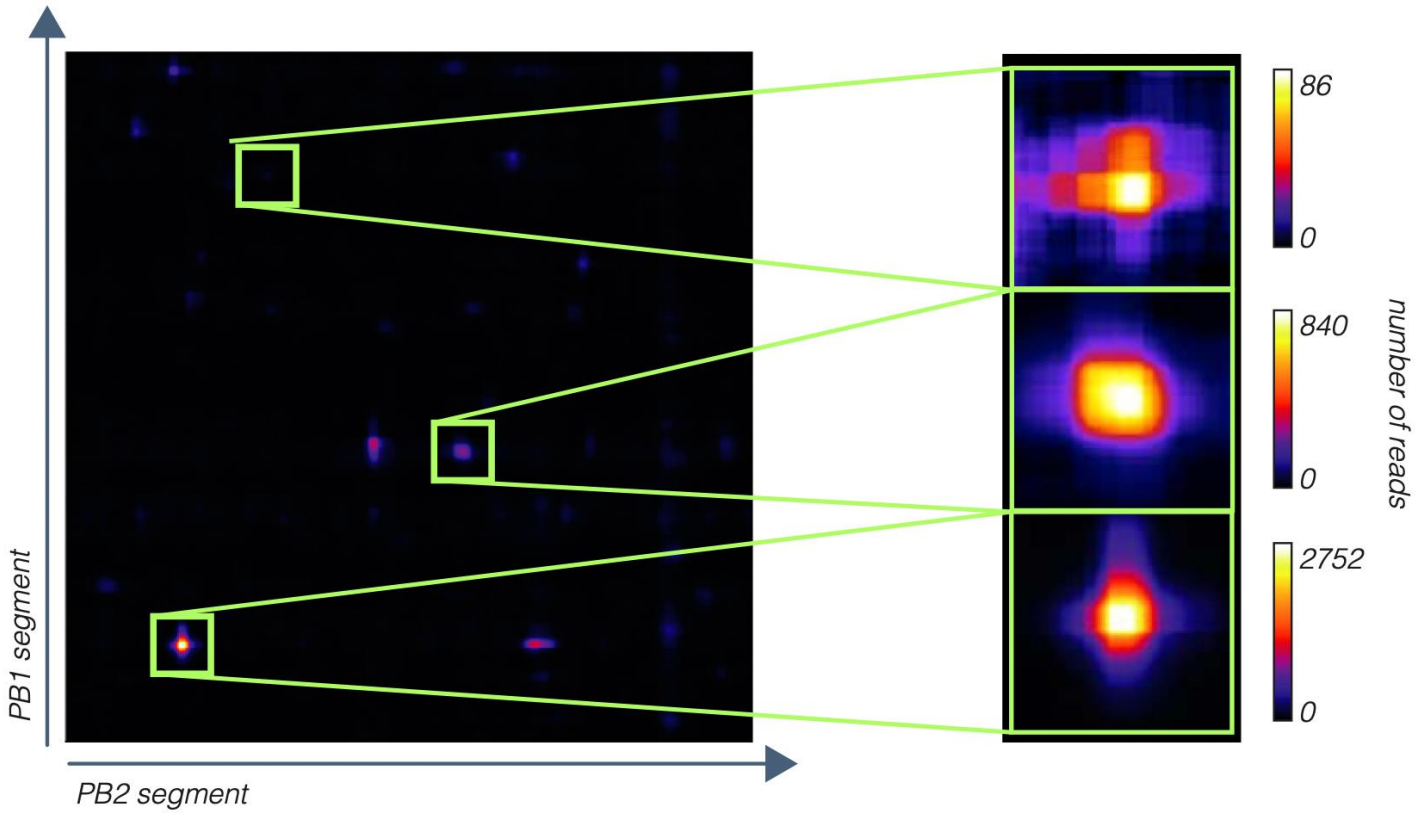
How do we automate annotation?

Jakob et al. *Nucleic Acids Research*, 2023

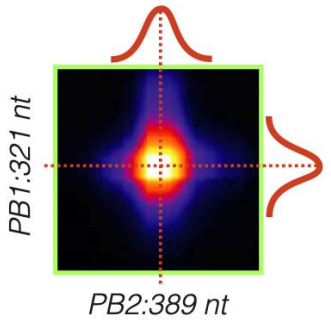
Interaction matrices can be annotated by fitting a 2d gaussian function to hand-picked quadrants

SPLASH identification of interaction loci

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



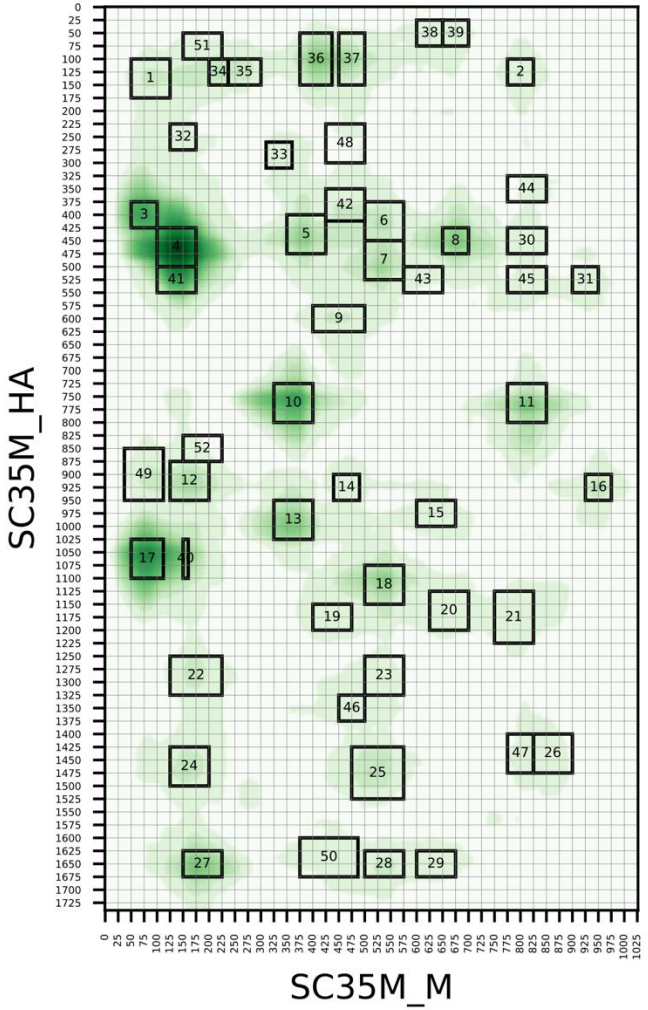
2. Fit a 2D Gaussian function to each locus to extract coordinates of interaction



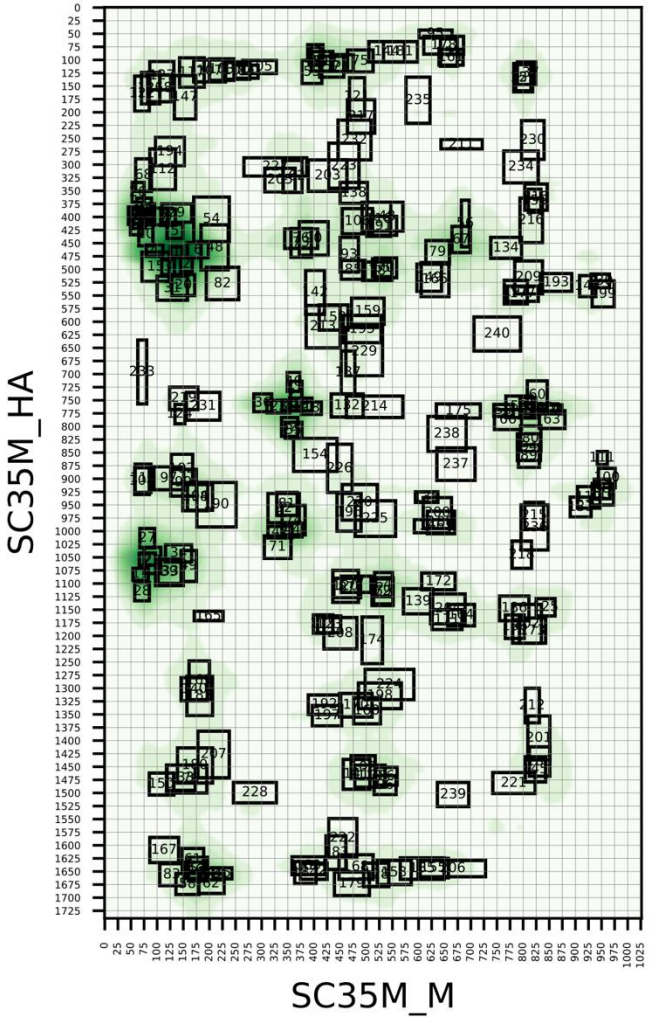
Dadonaite et al. *Nat. Microbiol* (2019)

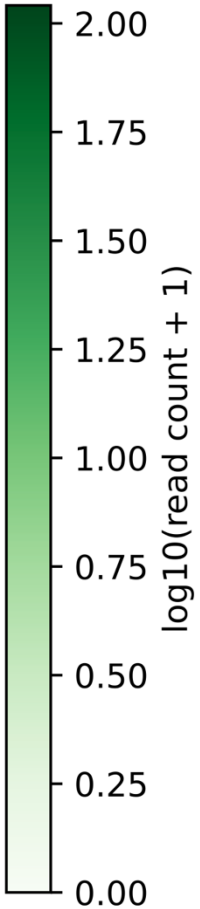
We can use heuristics to deduplicate the annotations

manual curation



automated



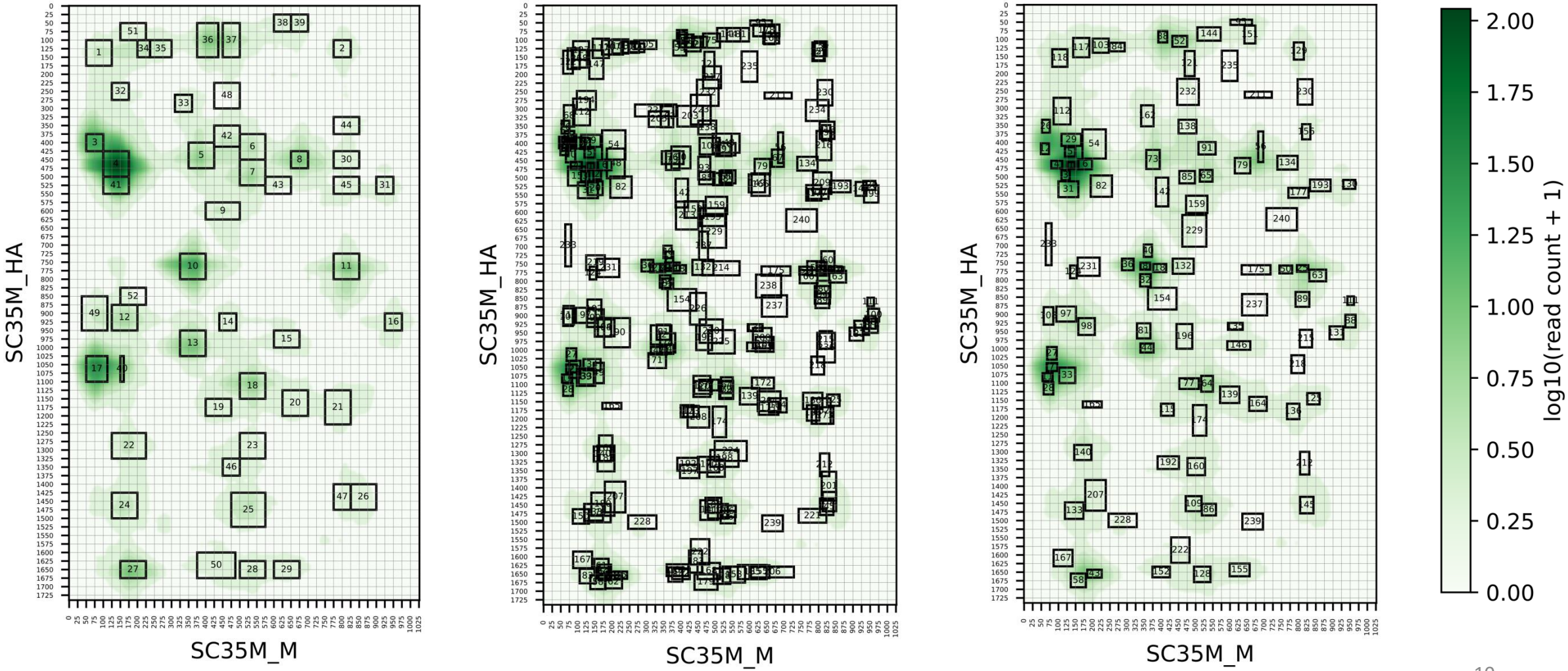


We can use heuristics to deduplicate the annotations

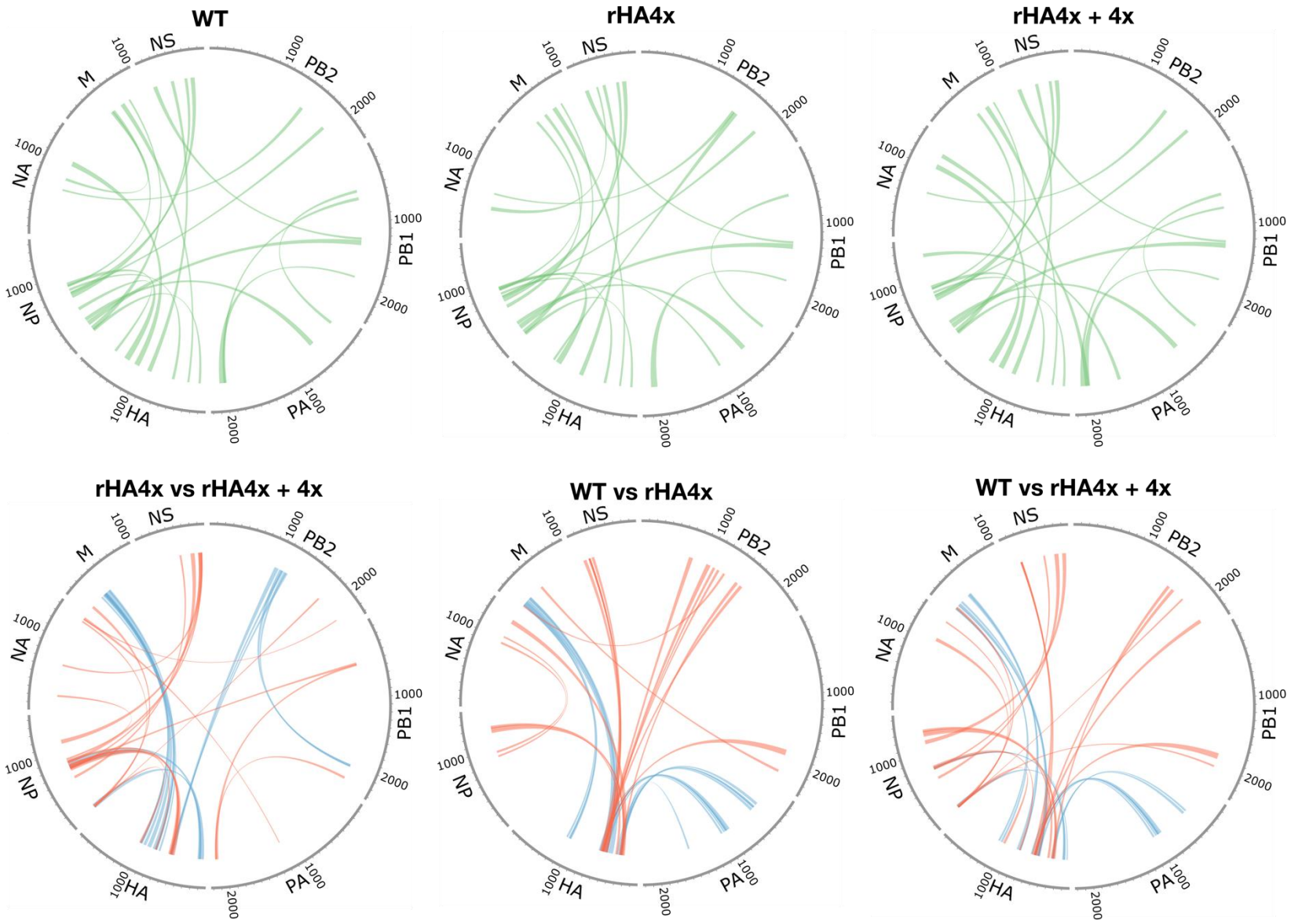
manual curation

automated

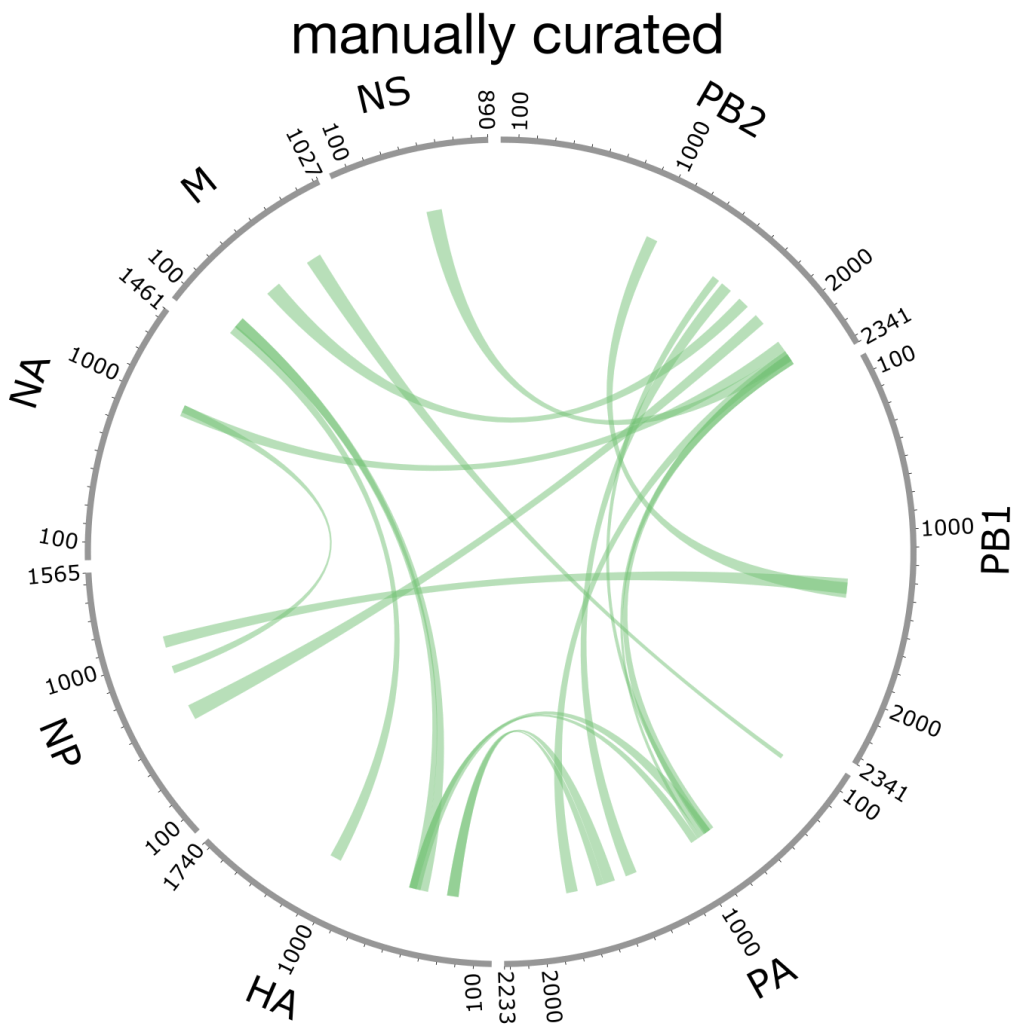
automated w/ deduplication

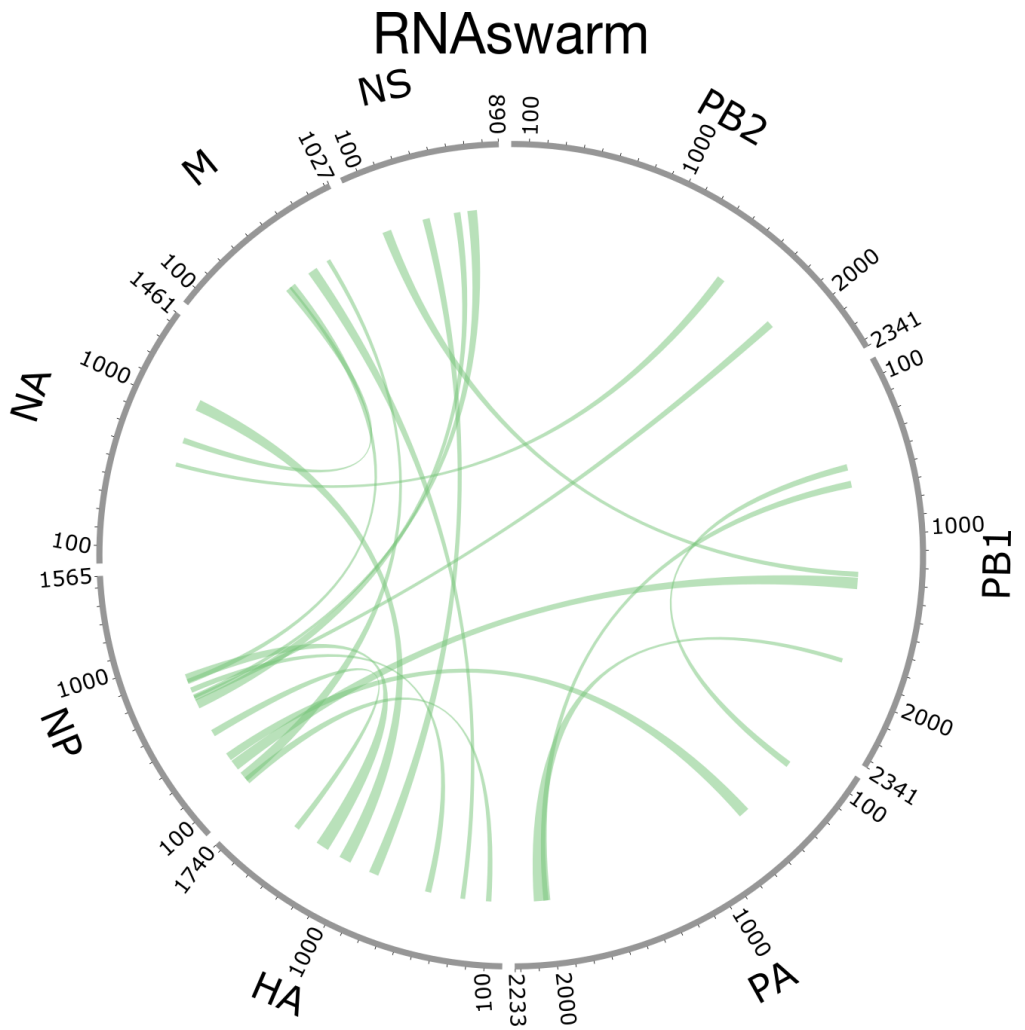


RNAswarm *de novo annotations*

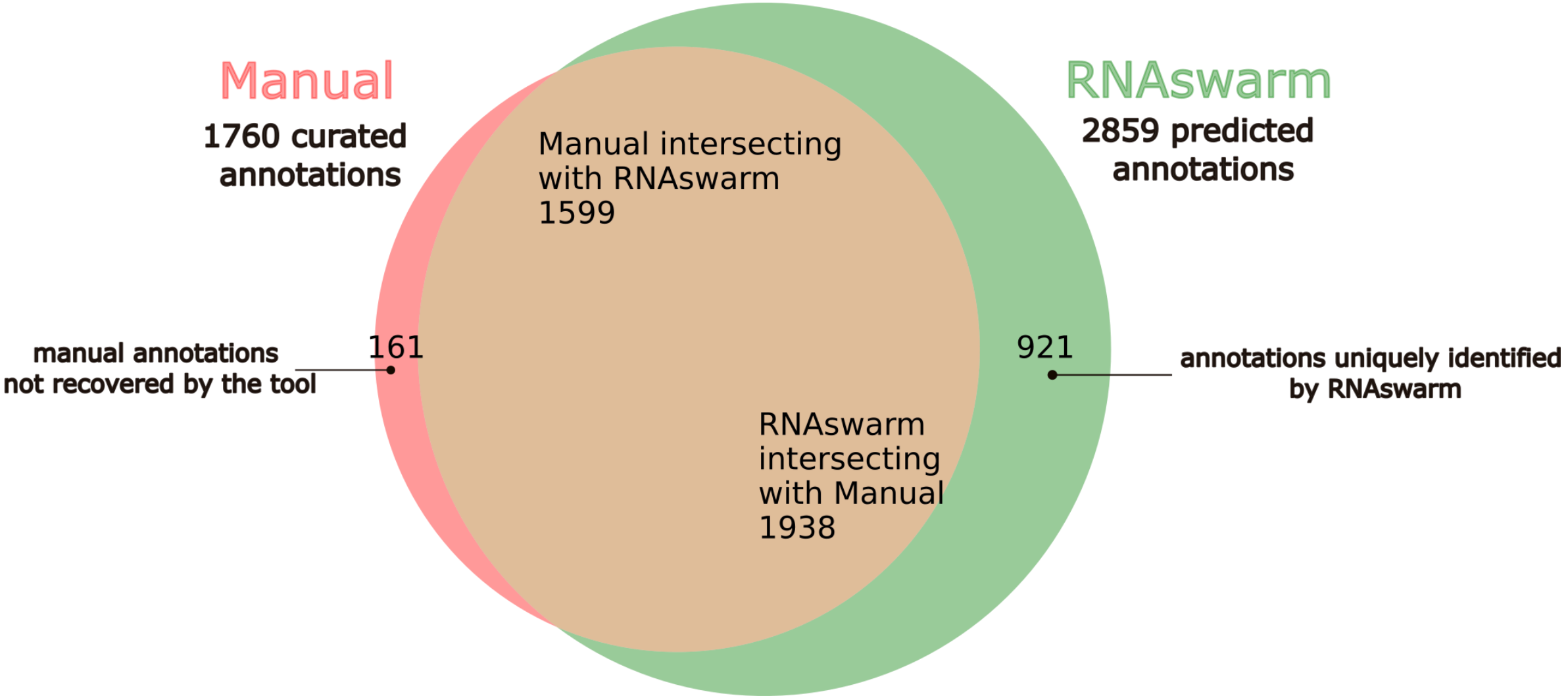


Manual vs RNAswarm





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

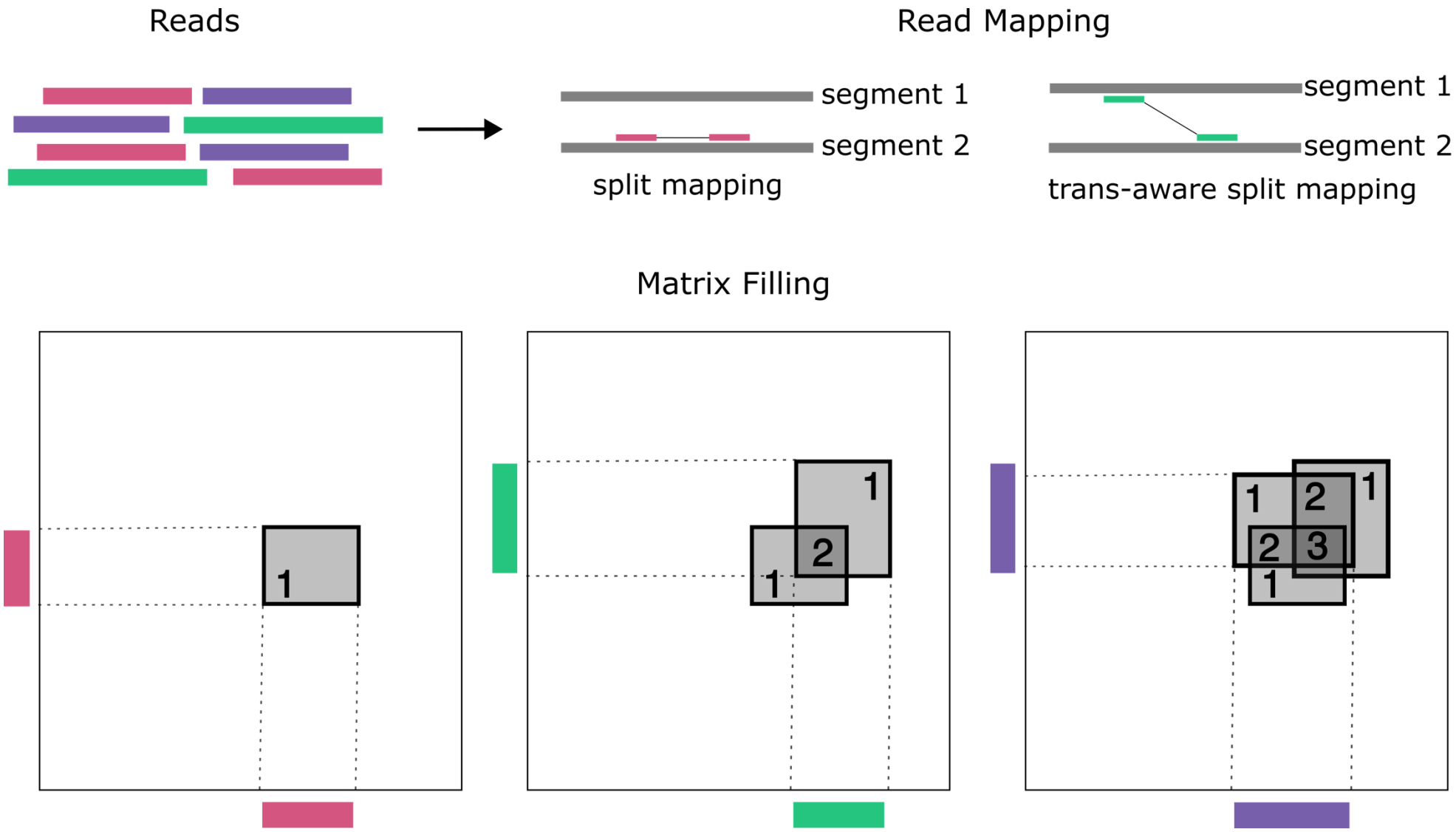
Manja Marz's Group
Christian Höner zu Siederdissen
Kevin Lamkiewicz

Martin Schwemmle's group
Hardin Bolte
Celia Jakob

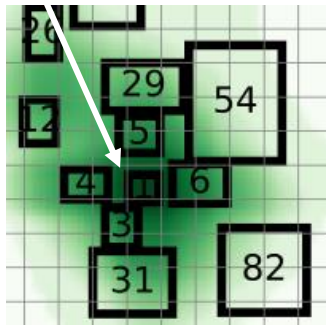
Research funded by EU's Horizon 2020 program, under the MSCA ITN grant agreement no. 955974

The lost slides

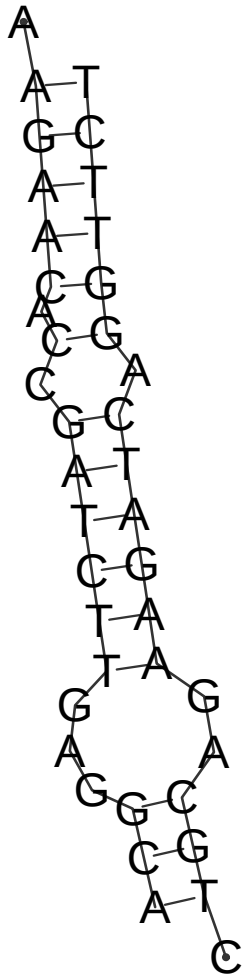
Matrix filling



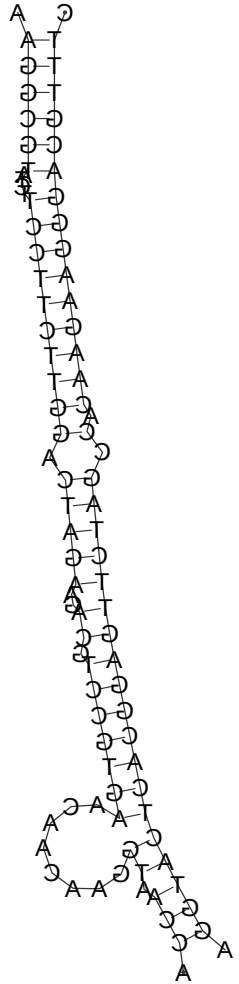
Structure prediction using RNAduplex (or RNAcofold)



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

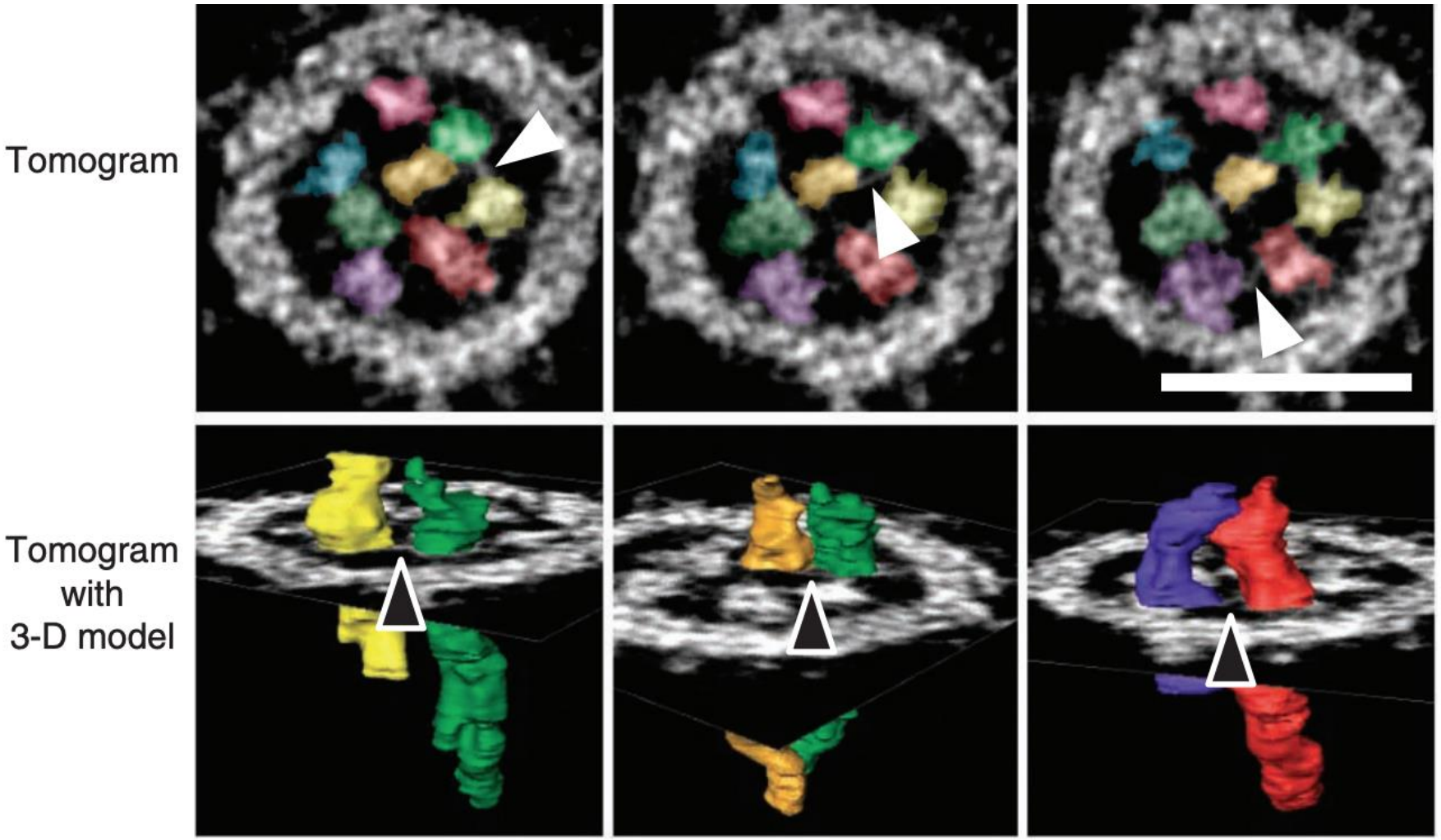
Can be fit to the data using an expectation-maximization algorithm:

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

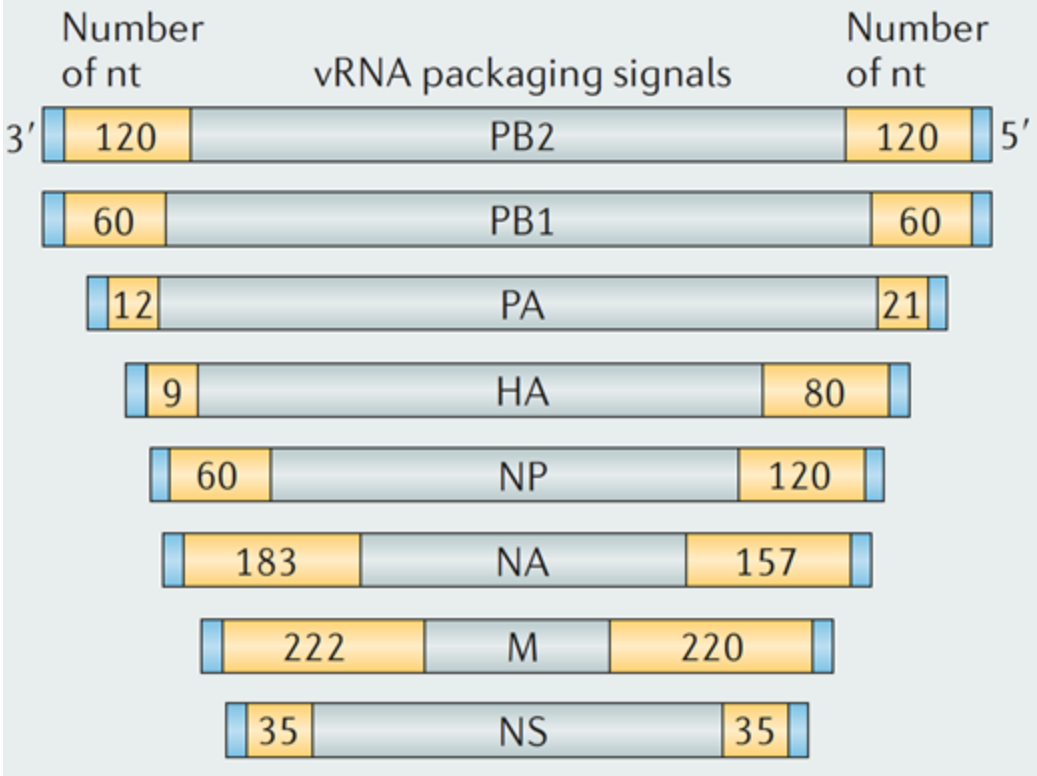


- Genome topology of IAV supports the 7 + 1 model
- vRNPs connected by a string-like structure

Noda, T. et al. Nat Commun 2012, 3 (1), 639.

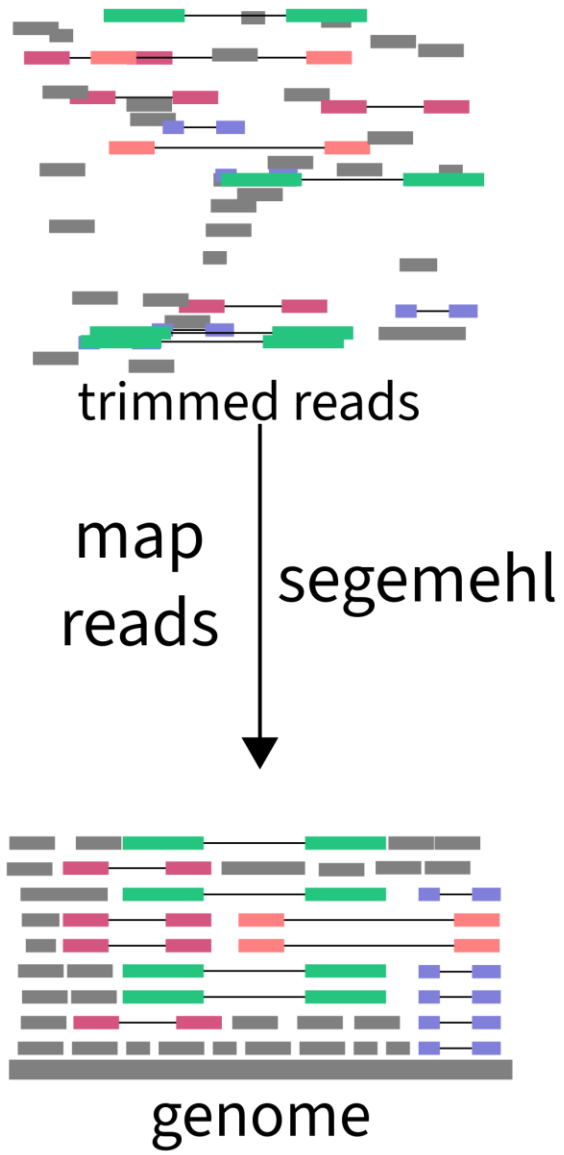
Known vRNA packaging signals

- 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.

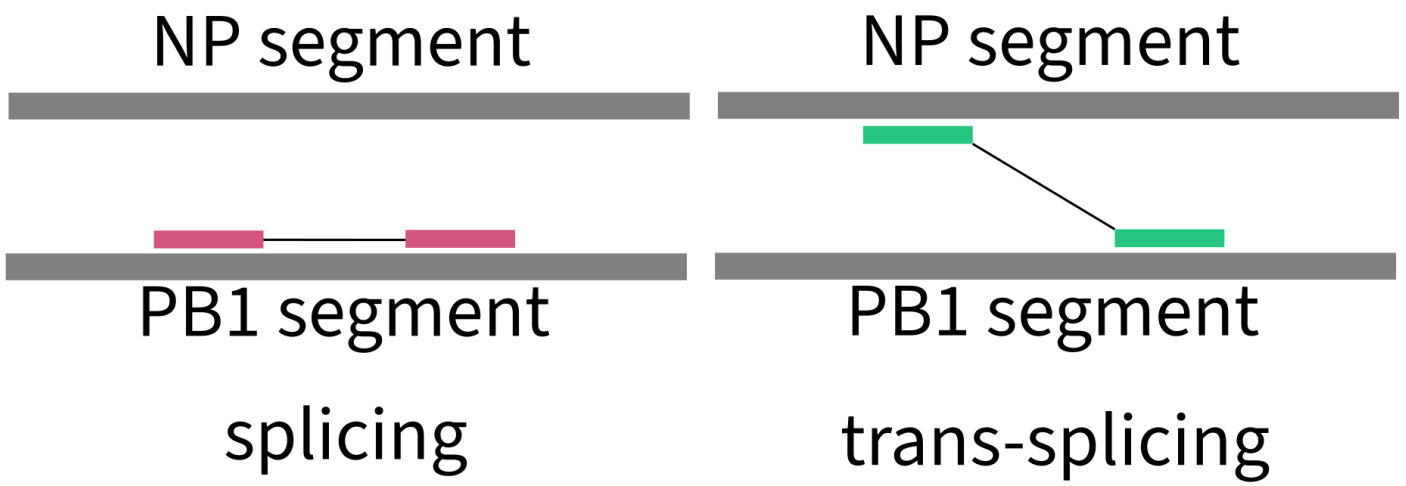
Split-read mapping



- 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.