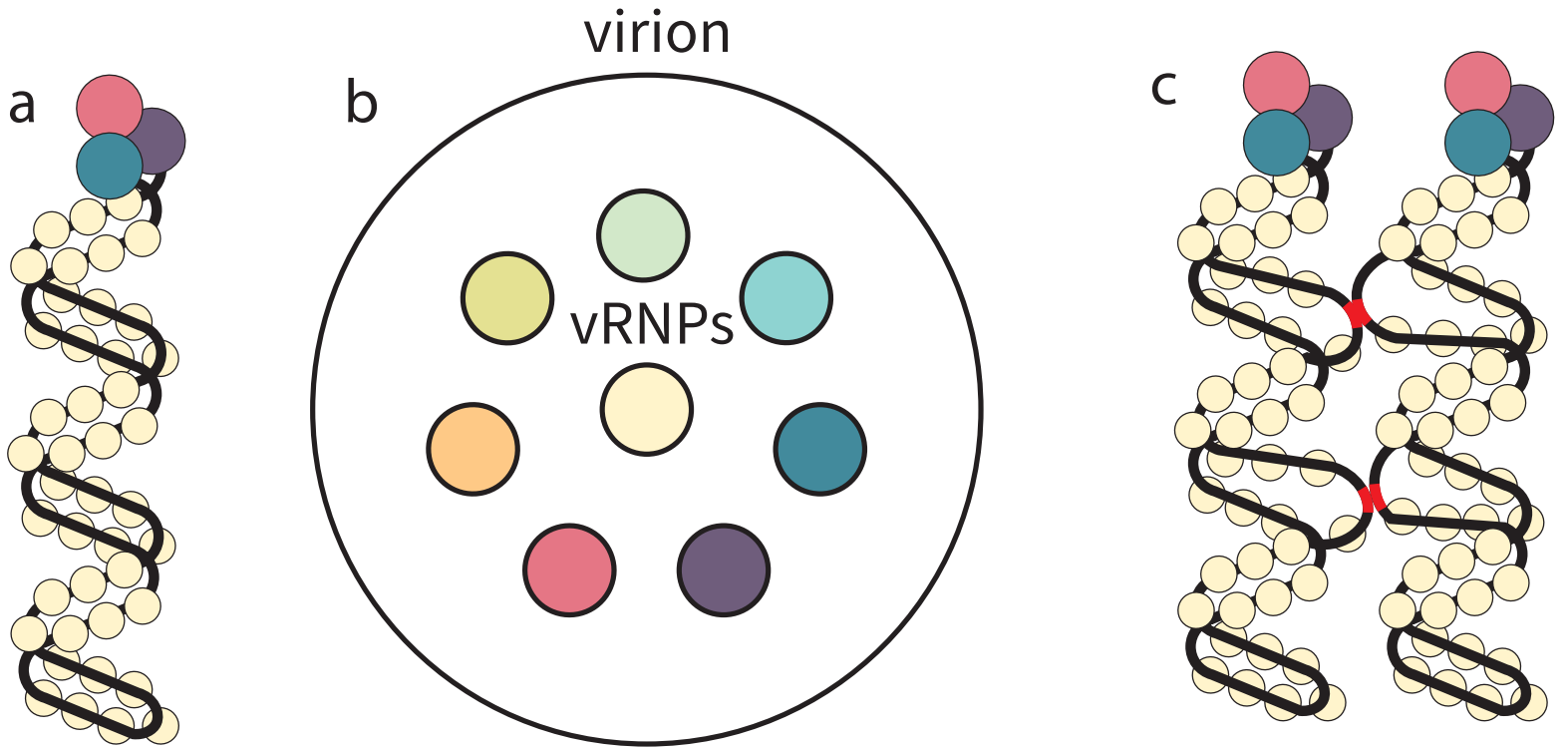
RNAswarm: A pipeline for high-throughput differential RNA-RNA interaction probing

Gabriel Lencioni Lovate

Bled, February 16th 2022

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

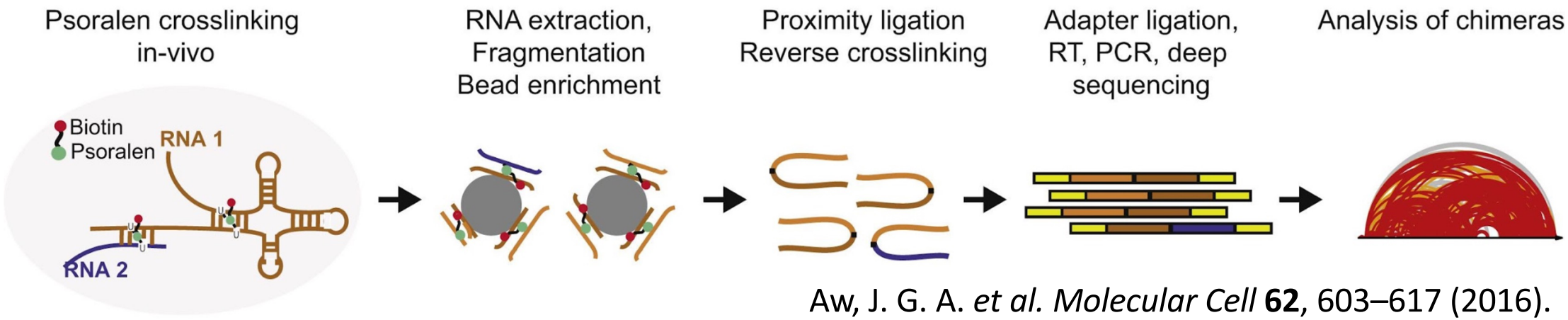
The IAV genome packaging problem



vRNP | 7 + 1 model | inter-segment interactions

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

SPLASH can probe RNA-RNA interactions



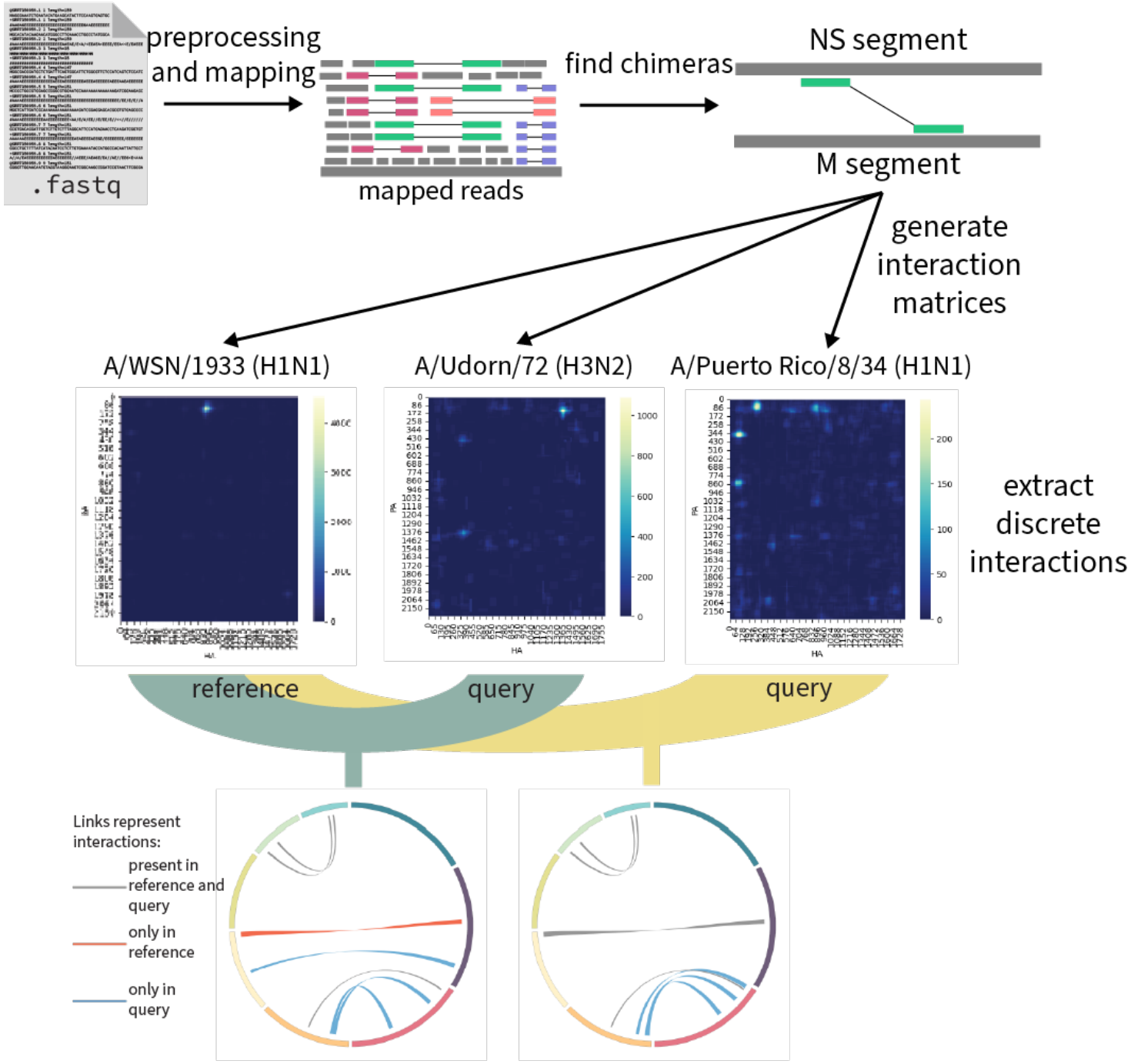
Aw, J. G. A. *et al. Molecular Cell* **62**, 603–617 (2016).

Implementing a reproducible analysis pipeline

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

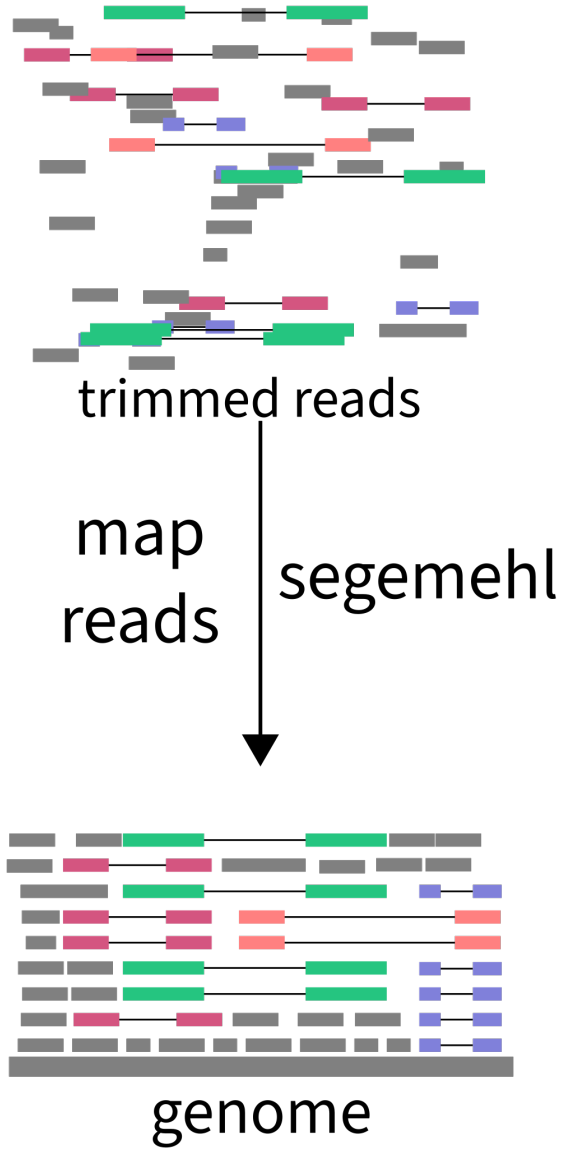
Differential analysis

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



RRI probing with RNAswarm

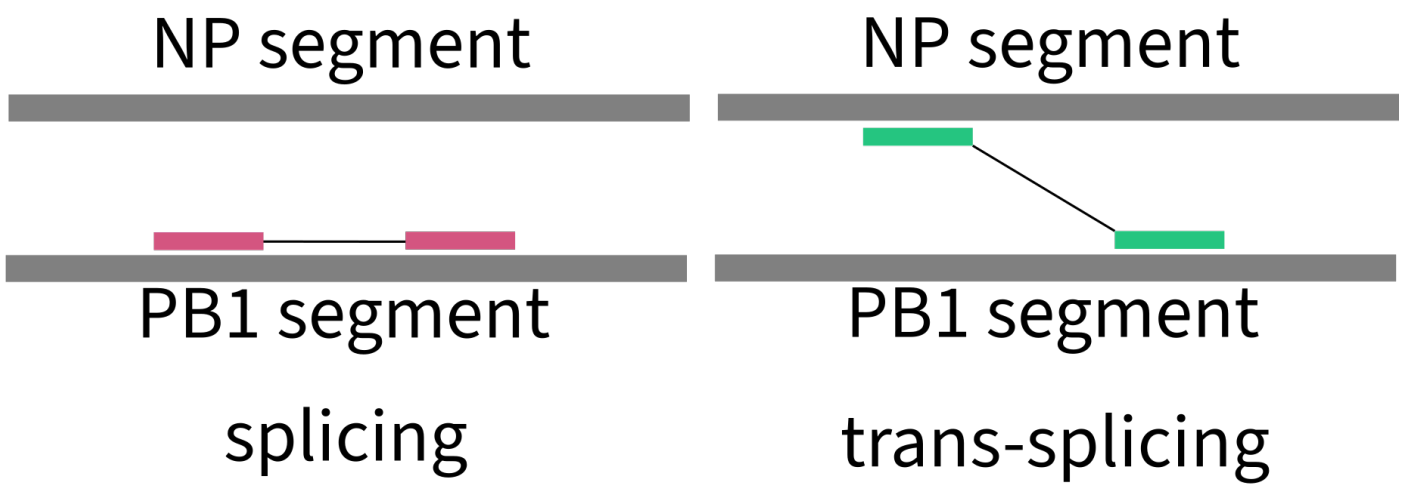
Split-read mapping



- Reads are aligned to the viral genome
- Chimeric reads are splitted during mapping
- Currently available mappers
  - segemehl (default)
  - BWA MEM

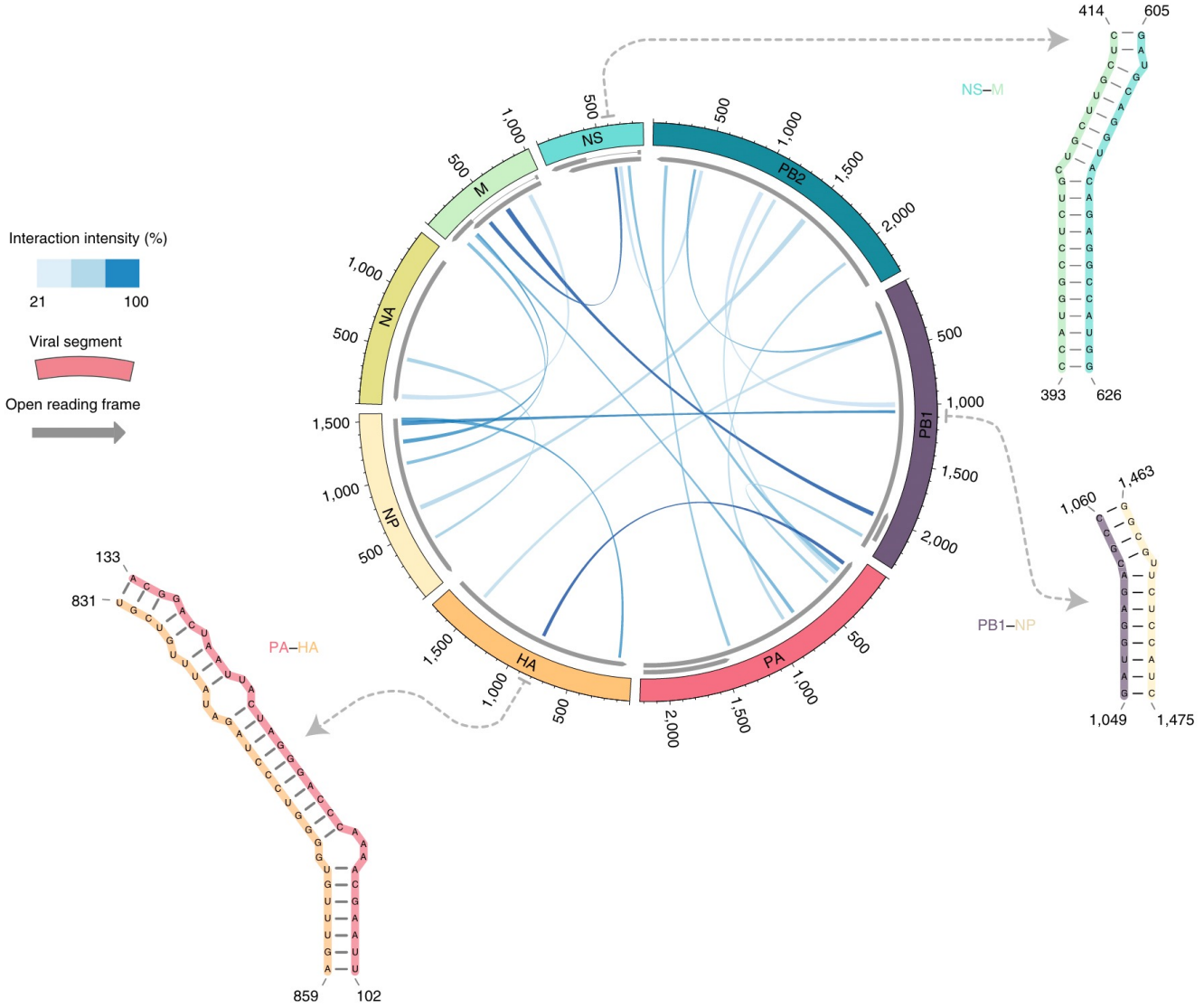
Finding chimeric reads

- 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



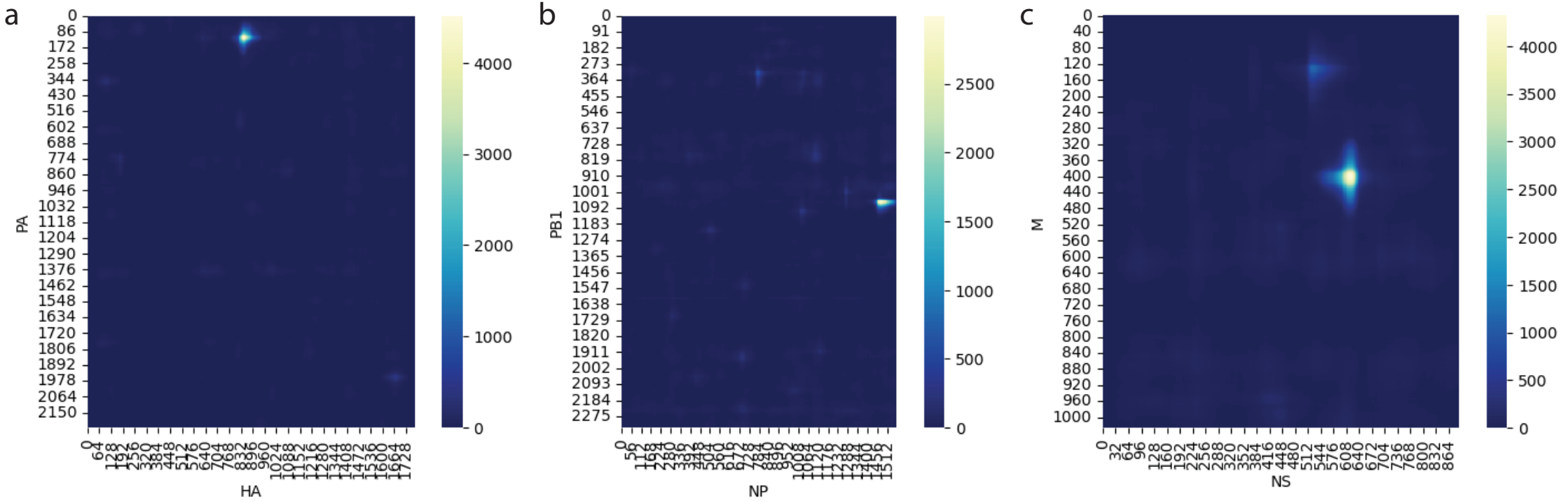
Finding chimeric reads

- intersegment RNA interactions in the IAV genome of the WSN (H1N1) IAV strain.
- Validating the pipeline with published datasets



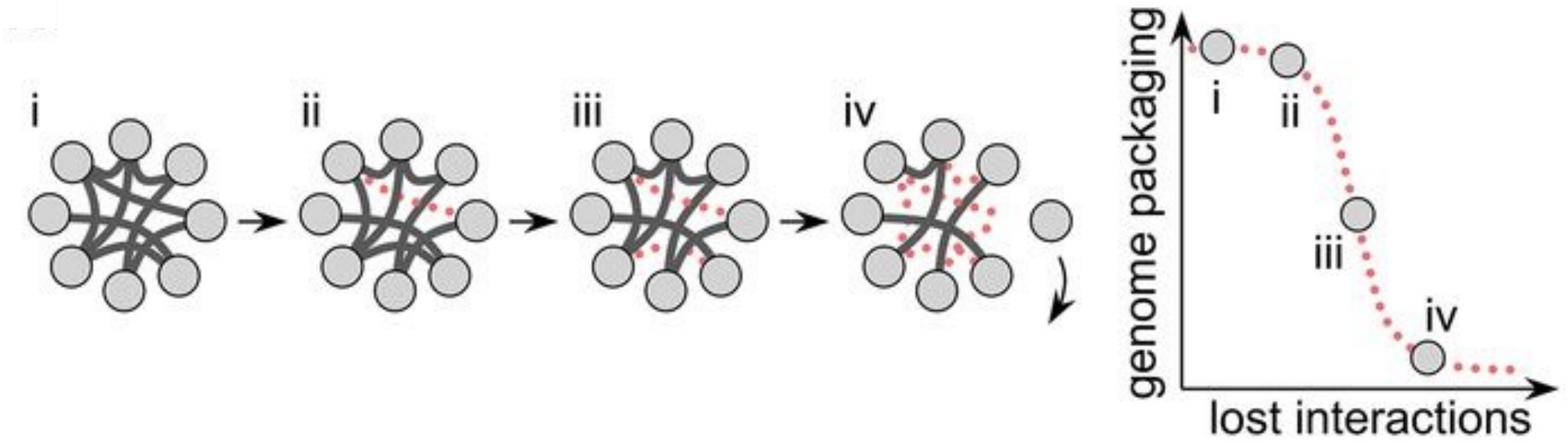
Dadonaite, B. et al. Nat Microbiol 2019, 4 (11), 1781–1789

Validating known RRIs



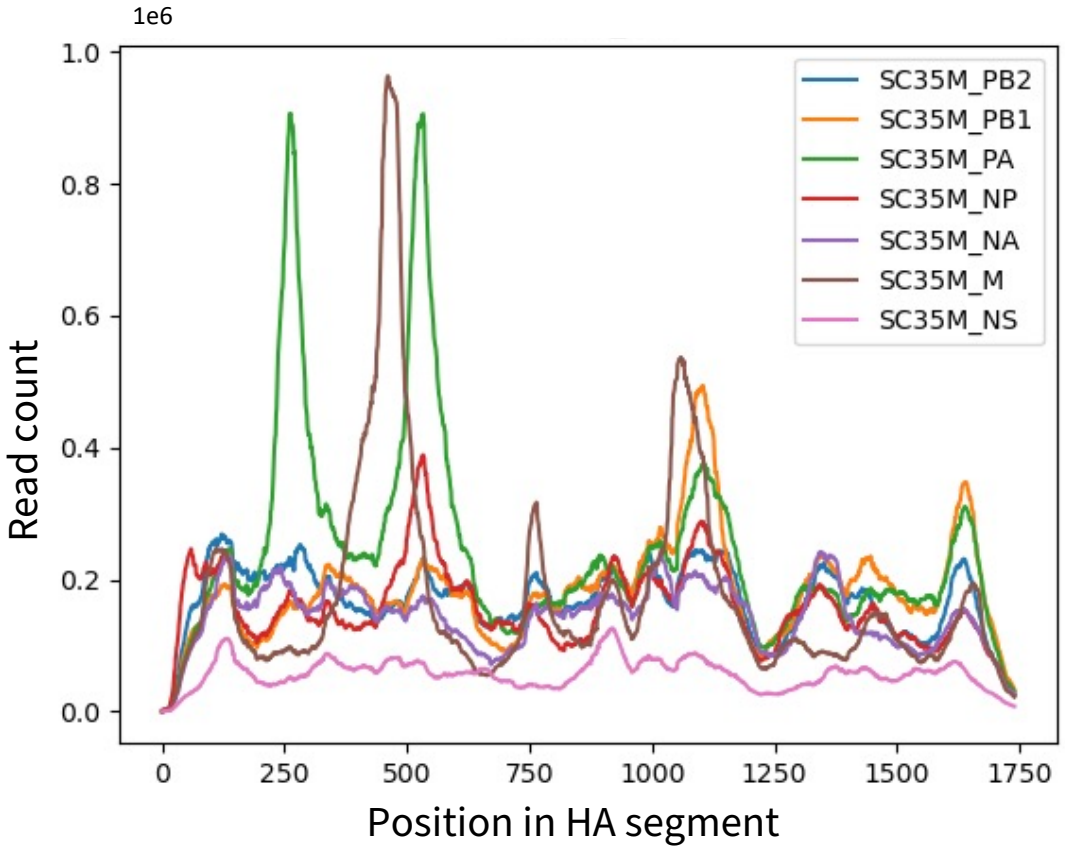
a. Interactions between PA (positions 133 to 102) and HA (positions 831 to 859) segments.
b. Interactions between PB1 (positions 1049 to 1060) and NP (positions 1463 to 1475) segments.
c. Interactions between M (positions 393 to 414) and NS (positions 605 to 626) segments.

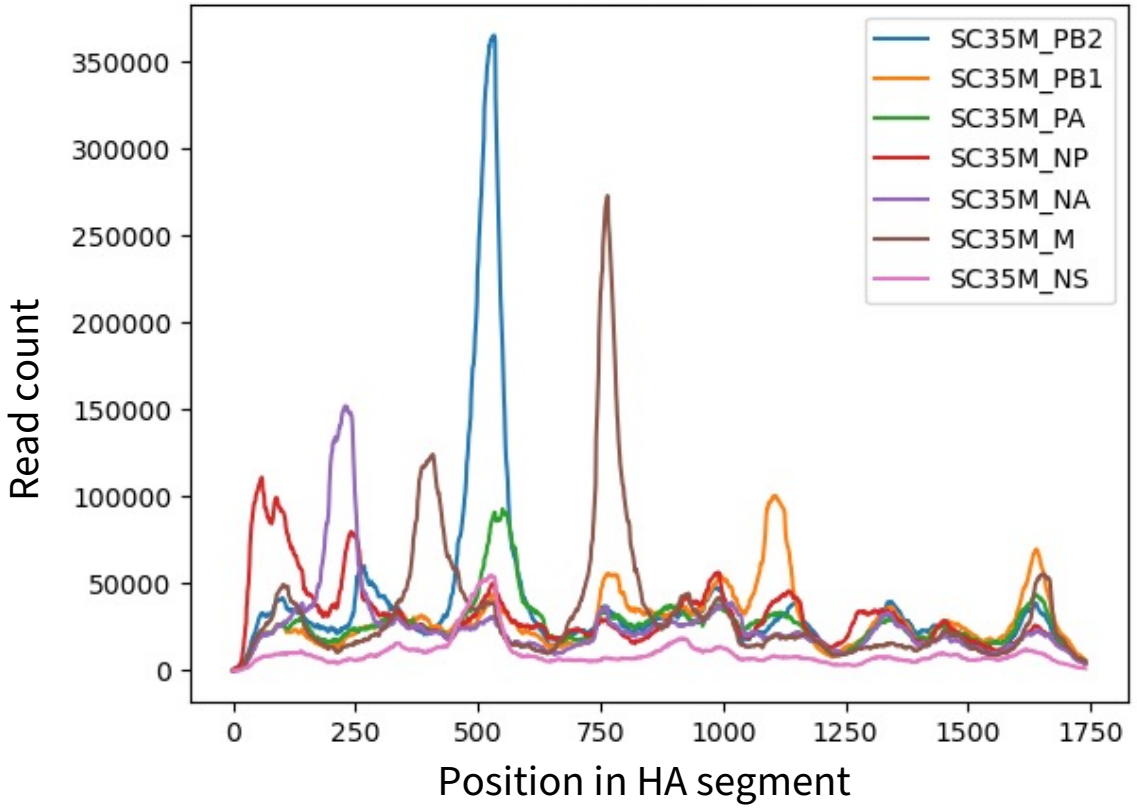
Mapping the interactome of SC35M (H7N7) strain with SPLASH



- Successive disruption of inter-segment interactions
- Expected reduction in fitness

Mapping the interactome of SC35M (H7N7) strain with SPLASH





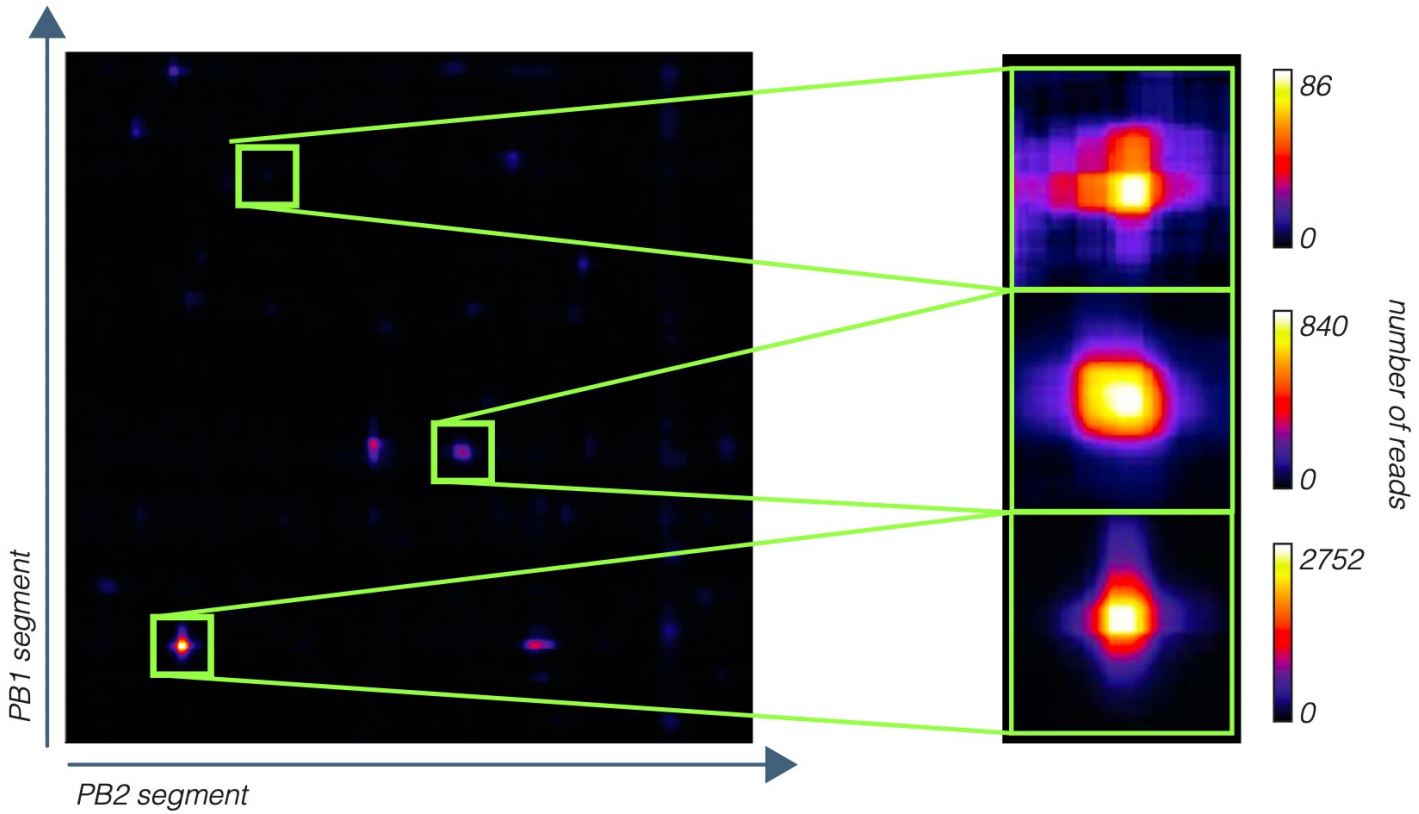
Wild-type

4xHA mutant
No fitness reduction!

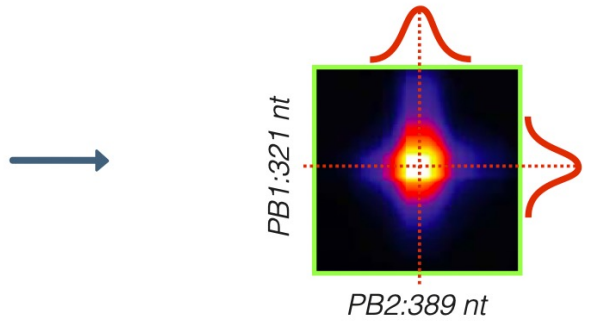
Defining discrete interactions

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, B. et al. Nat Microbiol 2019, 4 (11), 1781–1789

Using Gaussian Mixture Models (GMMs) to identify discrete interactions

- Independent from user intervention
- Can be fit to the data using an expectation-maximization algorithm:
  1. Assume random components.
  2. Compute for each point a probability of being generated by each component of the model.
  3. Tweak the parameters to maximize the likelihood of the data given those assignments.
  4. Repeat 2 and 3 until convergence.

Bayesian Information Criterion (BIC)

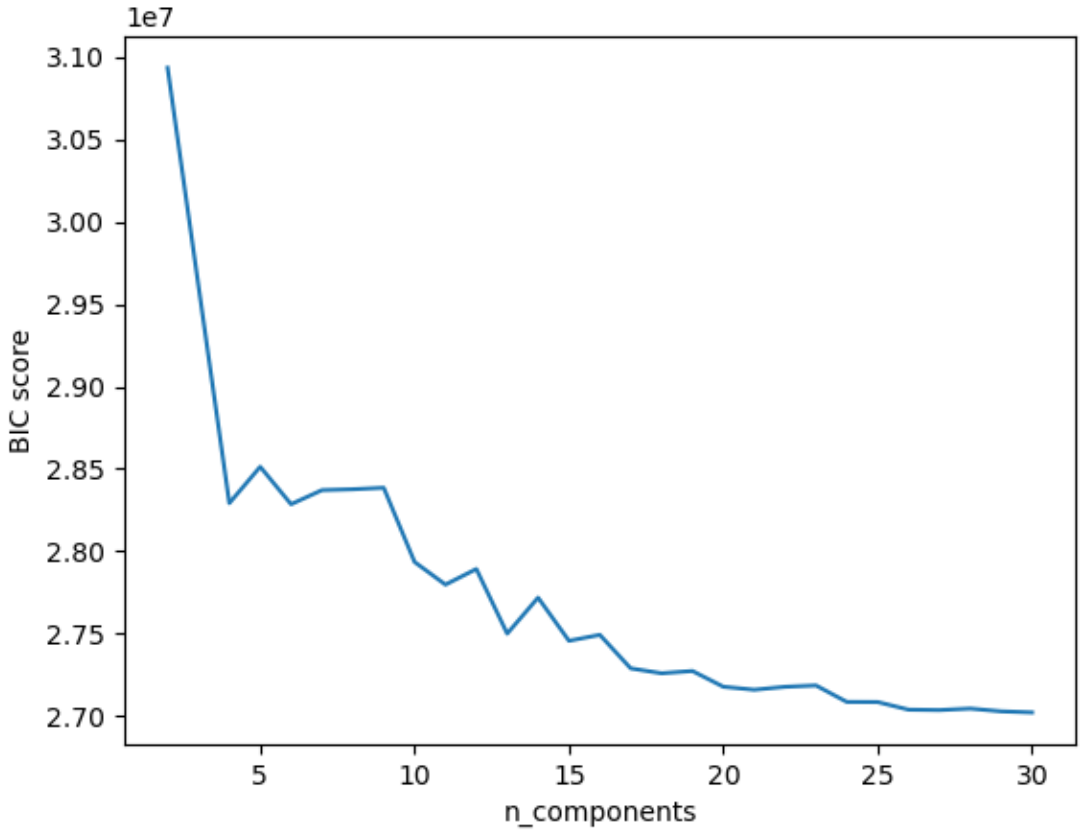
$$BIC(\theta) = -2\, logL(\theta) + k\, log\, i$$

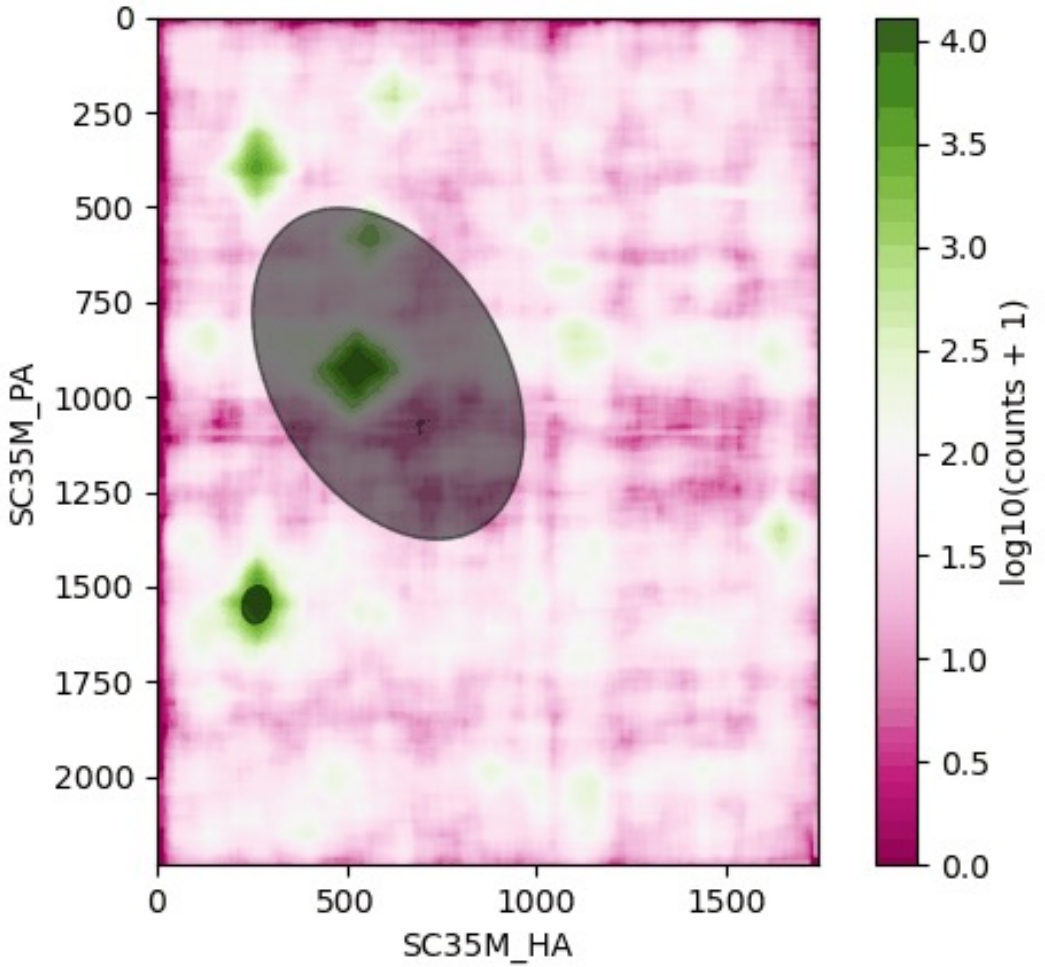
$L(\theta)$ being the achieved likelihood for a model θ with c components

k parameters

i data points

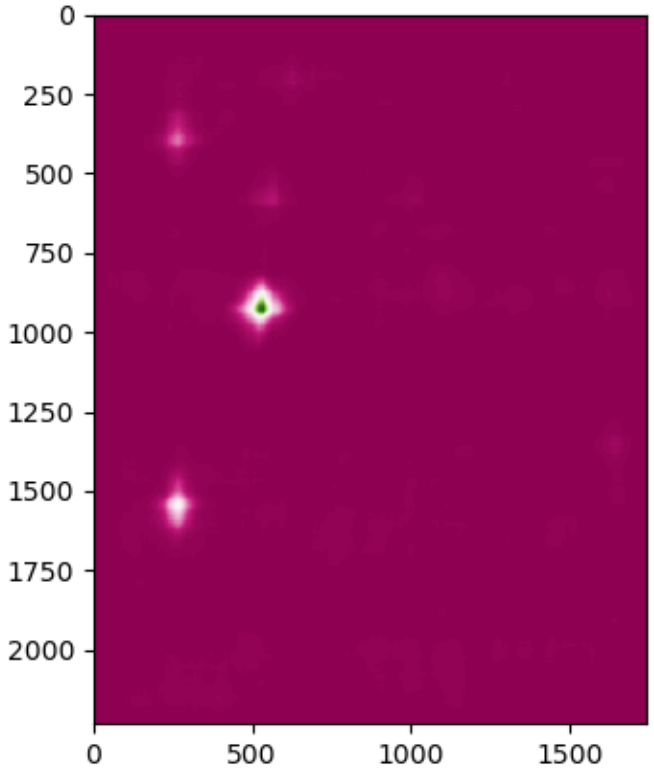
Defining number of features using the BIC





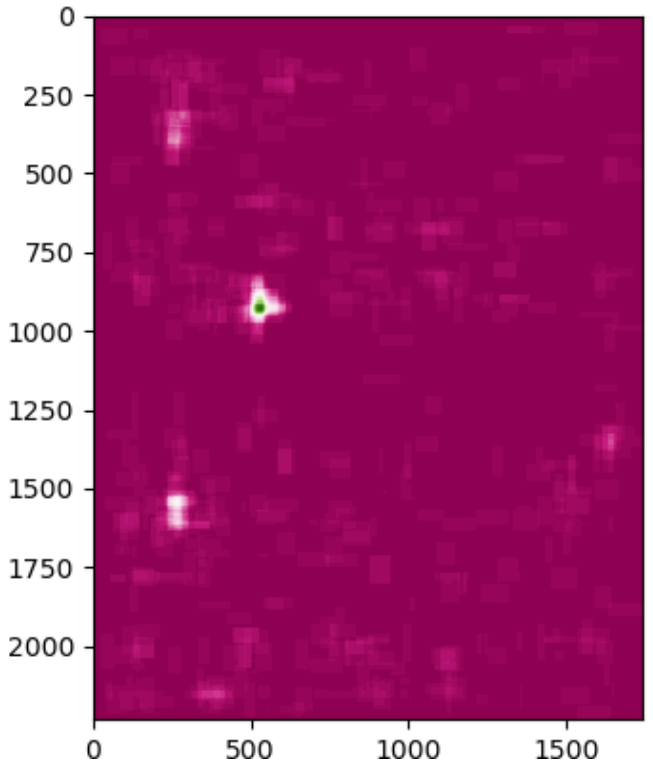
Merging matrices from multiple replicates

replicate 01



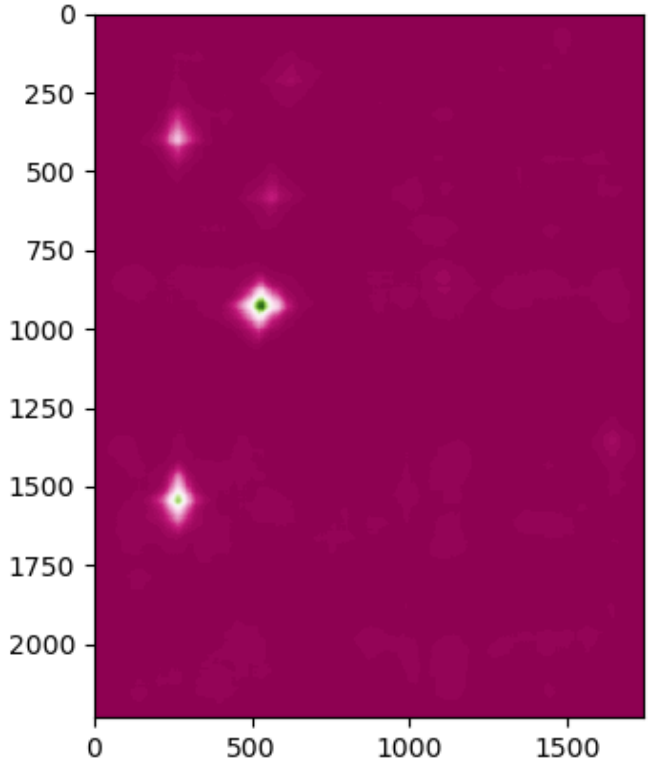
+

replicate 02



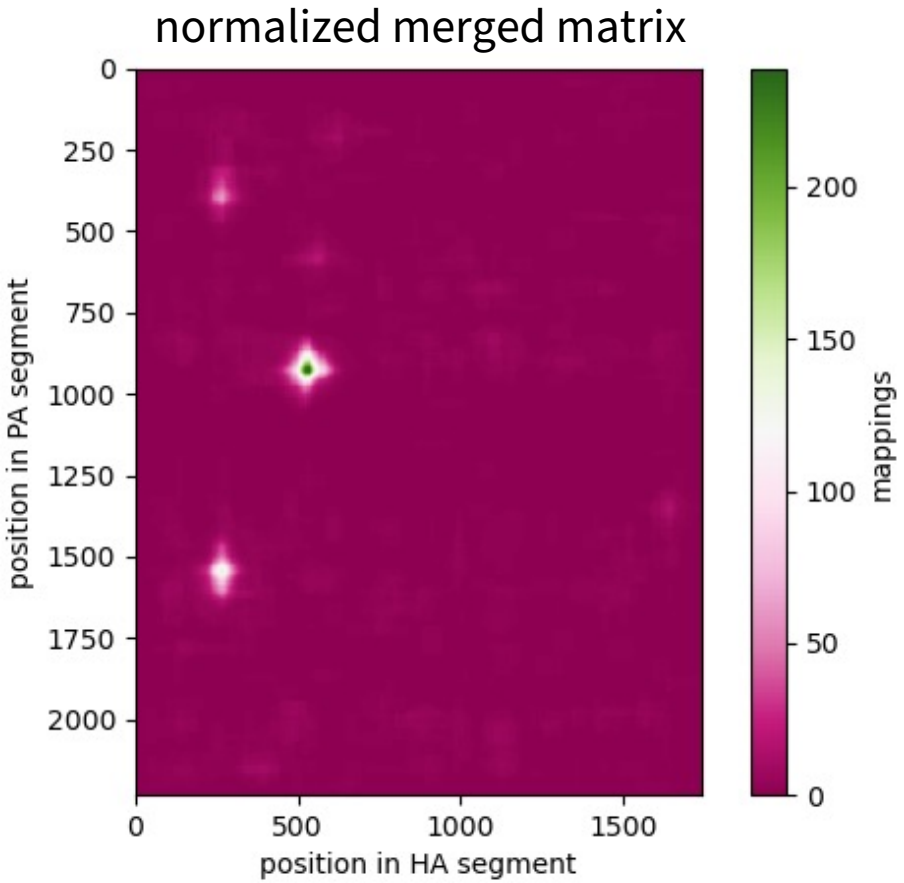
+

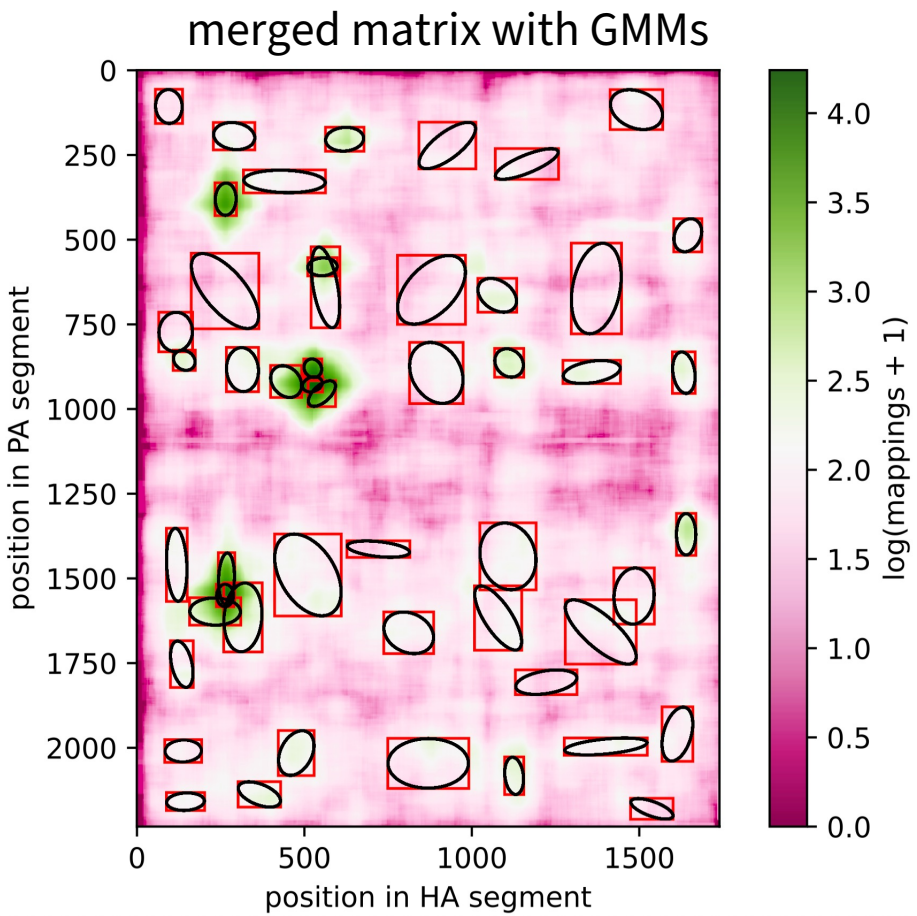
replicate 03



Sum up matrices from multiple replicates

Merging annotations





Fit GMM to merged matrix – 50 components

Generating annotation tables

	segment	start	end	segment	start	end
1	HA	100	175	M	50	125
7	HA	100	150	M	775	825
8	HA	375	425	M	50	100
9	HA	425	500	M	75	175
10	HA	400	475	M	350	425
11	HA	375	450	M	500	575
12	HA	450	525	M	500	575
13	HA	425	475	M	650	700
14	HA	575	625	M	400	500
15	HA	725	800	M	325	400
16	HA	725	800	M	775	850
17	HA	875	950	M	125	200
18	HA	950	1025	M	325	400
19	HA	900	950	M	440	490
20	HA	950	1000	M	600	675

Subset of interactions between HA and PA segments

Generating count tables

	WT 1	WT 2	WT 3	WT 4	WT 5	WT 6	WT 7	8xmut 1	8xmut 2	8xmut 3
1	829	30	402	67	720	939	158	194	57	67
7	200	21	173	31	305	364	90	57	28	31
8	1687	21	2976	506	4210	3188	322	309	108	316
9	10780	446	18616	4024	35023	29023	3375	641	191	513
10	1215	40	1130	164	2715	1548	320	146	55	102
11	1227	52	840	167	1339	1310	408	178	71	74
12	1465	50	961	207	1599	1799	545	232	103	144
13	844	32	617	136	1501	1131	217	31	11	12
14	588	24	451	98	527	668	248	73	46	86
15	2156	115	2444	1067	4715	7047	357	17	19	11
16	1002	34	1054	456	2610	2714	510	341	250	616
17	1119	41	675	201	1524	968	334	118	73	95
18	905	27	694	158	2666	2056	358	28	27	32
19	249	11	163	60	238	170	77	14	29	30
20	410	11	212	51	311	410	124	56	26	51

Subset of interactions between HA and PA segments

DESeq2 analysis

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
19_SC35M_PA_875_975_SC35M_HA_475_575	4025.24702844095	5.44992388413147	0.412178339933248	13.2222471588732	6.52839589810325E-40	1.11570285898584E-36
15_SC35M_HA_725_800_SC35M_M_325_400	778.985831810769	4.85712665285643	0.508802449218027	9.54619353802501	1.34549250518297E-21	4.59889338271539E-19
7_SC35M_HA_25_125_SC35M_NP_700_800	567.893675396786	4.38247740250625	0.477201751669411	9.18369932041298	4.16530520926159E-20	1.01692951466115E-17
22_SC35M_HA_1025_1100_SC35M_M_50_125	1222.07109733955	4.35537239806334	0.425830858563133	10.2279398274694	1.48647902590521E-24	6.35098163817999E-22
1_SC35M_PA_350_450_SC35M_HA_225_300	1028.51894139388	4.3256368895604	0.420790386651096	10.2797901919443	8.69165997649876E-25	4.95134896661212E-22
1_SC35M_HA_50_150_SC35M_NP_25_100	372.473095581194	4.28886550049346	0.492876988261682	8.70169556022441	3.26961748808314E-18	5.58777628713408E-16
63_SC35M_PA_800_875_SC35M_HA_475_550	1473.50519238682	4.074986883656	0.430166159271509	9.47305313499564	2.71781336342697E-21	7.74123839682781E-19
47_SC35M_HA_500_550_SC35M_M_100_175	1566.85974057496	4.00783192747044	0.446692112725806	8.97224690853357	2.90526275374386E-19	6.20636755768531E-17
25_SC35M_PA_1500_1625_SC35M_HA_225_300	2463.73510135188	3.21986846664587	0.362543951904923	8.88131894002822	6.60722125324147E-19	1.25463790242107E-16
13_SC35M_HA_425_475_SC35M_M_650_700	187.105200255042	2.69980868056625	0.324404208119329	8.32236023144666	8.62287741565874E-17	1.33968159121462E-14
9_SC35M_HA_425_500_SC35M_M_75_175	4204.08293463218	2.53944878879569	0.383988927945846	6.61333857301696	3.75747711688346E-11	4.58680599482417E-09
62_SC35M_PA_275_350_SC35M_HA_250_350	652.147938245227	2.52482220880395	0.370004042561463	6.82376925215496	8.86822879605578E-12	1.26298358437161E-09
18_SC35M_HA_950_1025_SC35M_M_325_400	259.282450816198	2.30850194063679	0.366462407872042	6.29942359993129	2.98754509766311E-10	3.40380971460417E-08
3_SC35M_HA_225_300_SC35M_NP_275_350	53.7945521206507	2.29941618892772	0.446209203623051	5.15322447465746	2.56045164124669E-07	2.30305887099505E-05
39_SC35M_PB2_1100_1200_SC35M_PB1_2000_2075	569.990736179057	2.01426376710235	0.373611087881521	5.39133829920247	6.99348731526449E-08	6.63992767877057E-06

Comparison between SC35M (H7N7) and SC35M 8x (H7N7)

Outlook: Better extraction of interactions

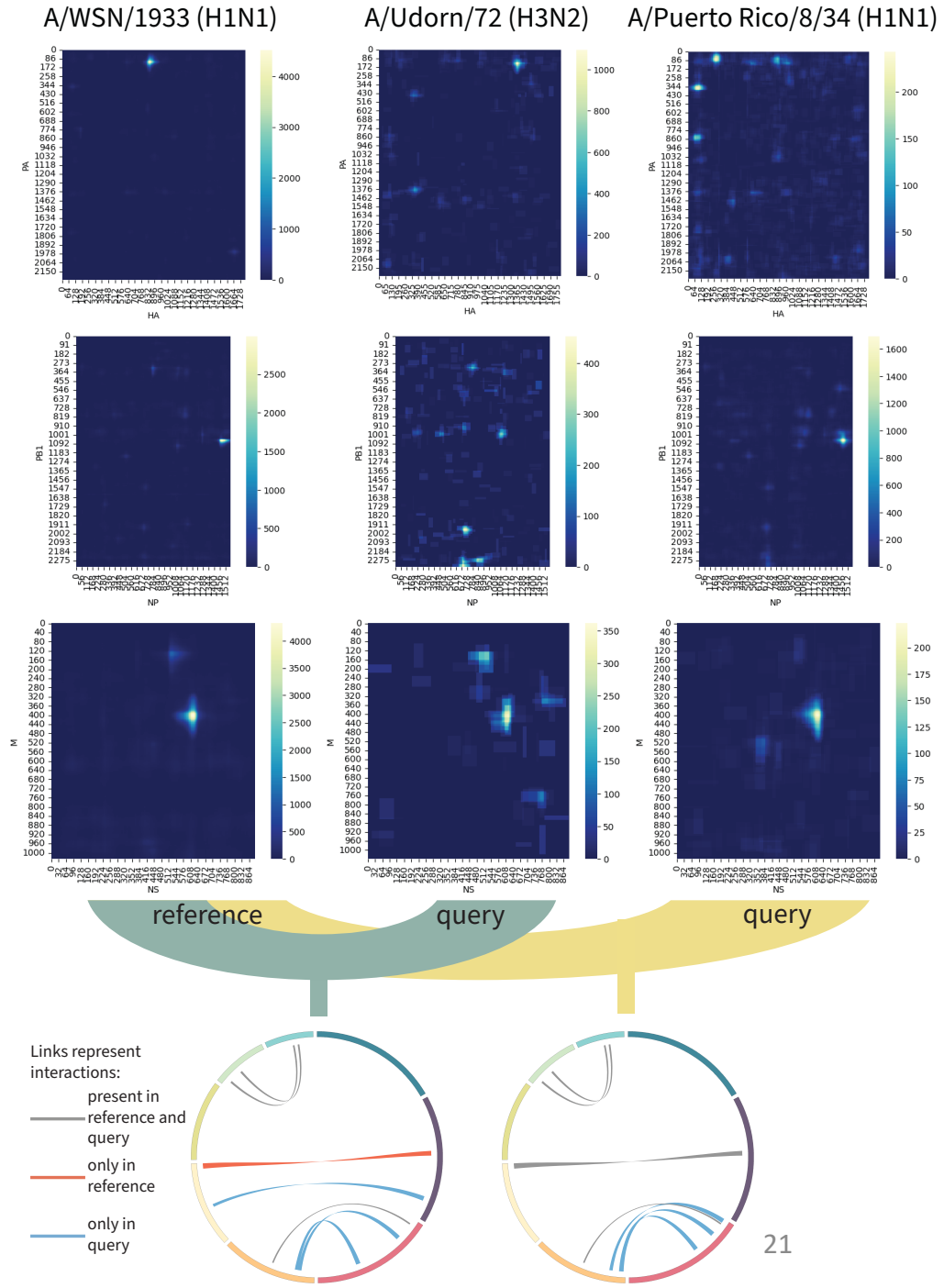
- Merged matrices can be “noisy”
- Convoluted gaussians when too many components are present
- Refine interaction region

Outlook: Validating GMM-identified annotations

- Experts are able to draw boundaries around interactions
- We are comparing GMM-identified annotations with manually curated annotations from experts

Outlook: Differential RRI

- Interaction matrices are processed, in order to identify differentially structured regions in the genome
- In the example, two IAV strains (A/Udorn/72 and A/Puerto Rico/8/34) are compared against WSN.



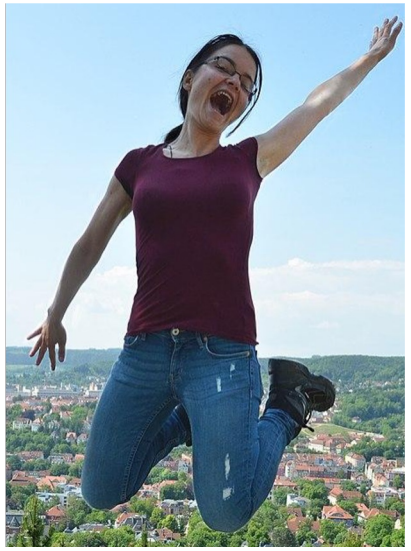
Summary

- Input: **fasta** & **fastq** file
- Output: Interaction matrices, differential interaction table & heatmaps
- Modular and reproducible nextflow pipeline
- In principle RNAswarm can deal with data resulting from a broad range of proximity ligation methods
- We can identify discrete interactions in an unsupervised fashion using GMMs
- We can compare the prevalence of discrete interactions across conditions using DESeq2
- Source code will be available in **spring**

Acknowledgments

Kevin Lamkiewicz

Manja Marz



Theoretical Bioinformatics Lab

Christian Höner zu Siederdissen

Universitätsklinikum Freiburg

Celia Jakob

Hardin Bolte

Martin Schwemmle

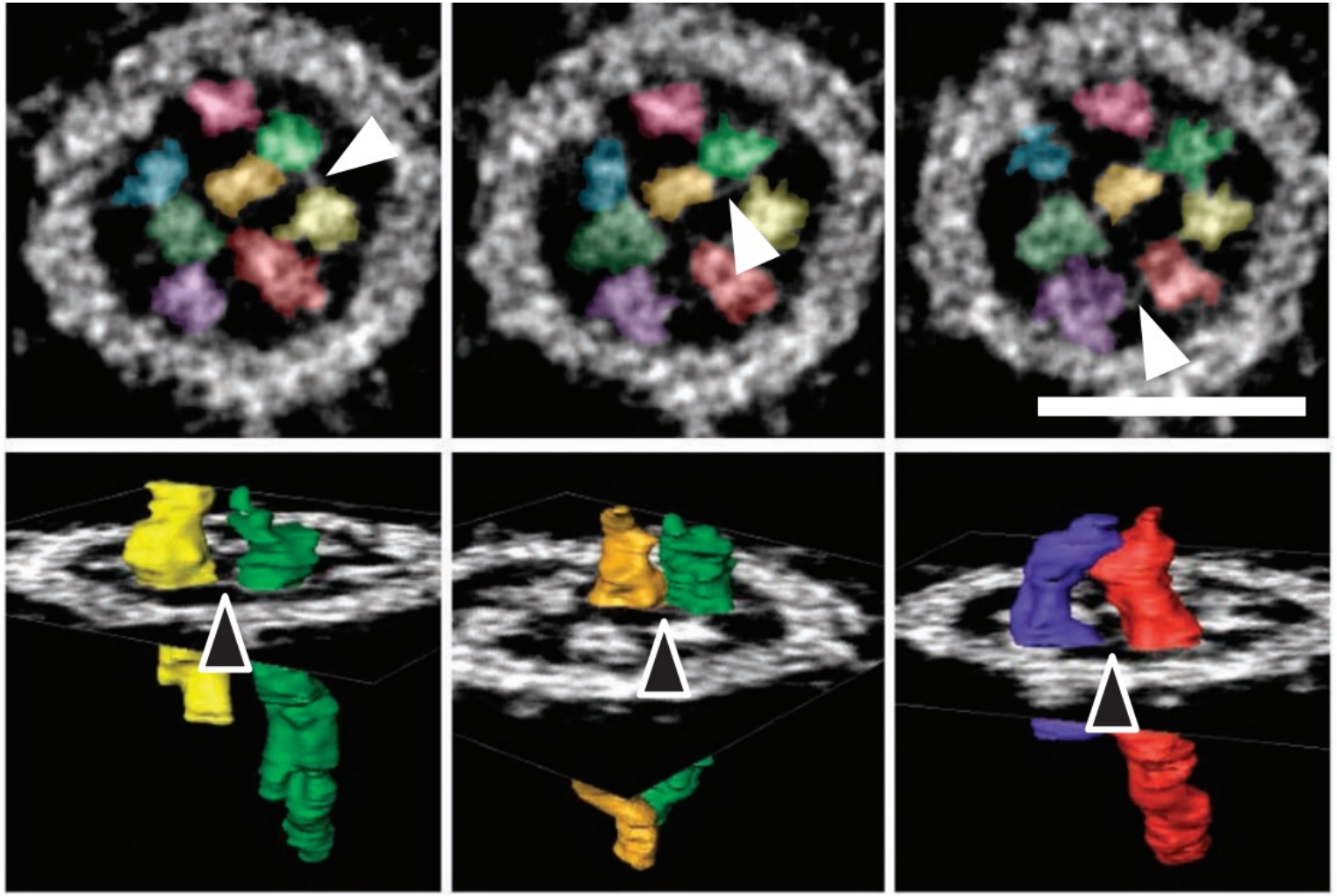
Understanding virus-host interactions

This research was funded by European Union's Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF)

Acknowledgments

RNA Bioinformatics and High-throughput Analysis Group

The IAV genome packaging problem

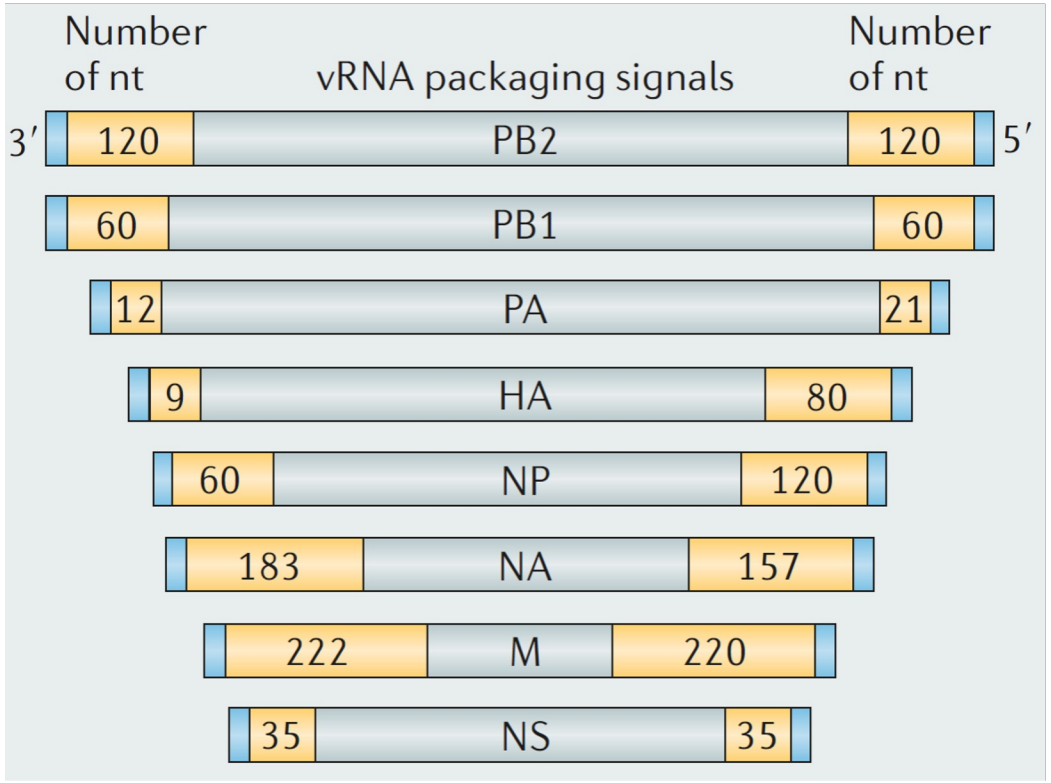


- Genome topology of IAV supports the 7 + 1 model
- vRNPs connected by a string-like structure

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.