Medley: Interaction of Virology and Bioinformatics

M. Marz

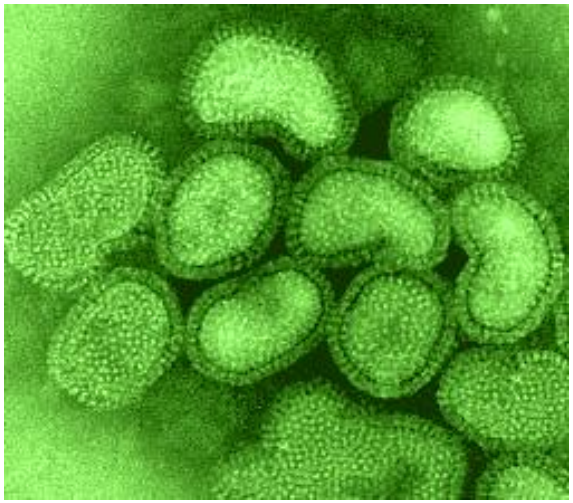
University of Jena, Germany
Chair for HTS Data Analysis

Director of European Virus Bioinformatics Center

Institute of Data Driven Science: MSCJ
Leibniz Institute for Age Research: FLI
Aging Research Center Jena

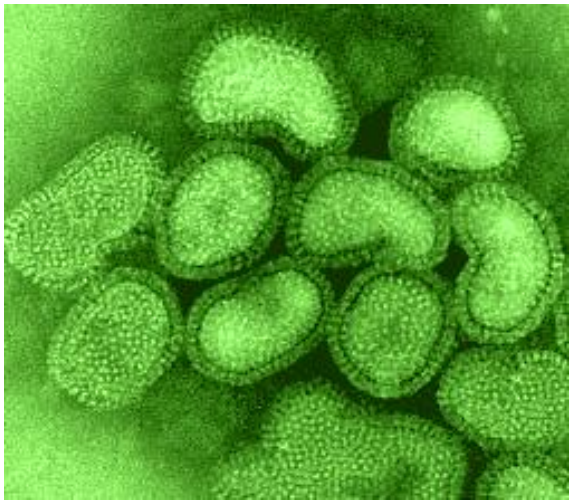
Bled, Slovenia
21.02.2018

Viruses are tiny



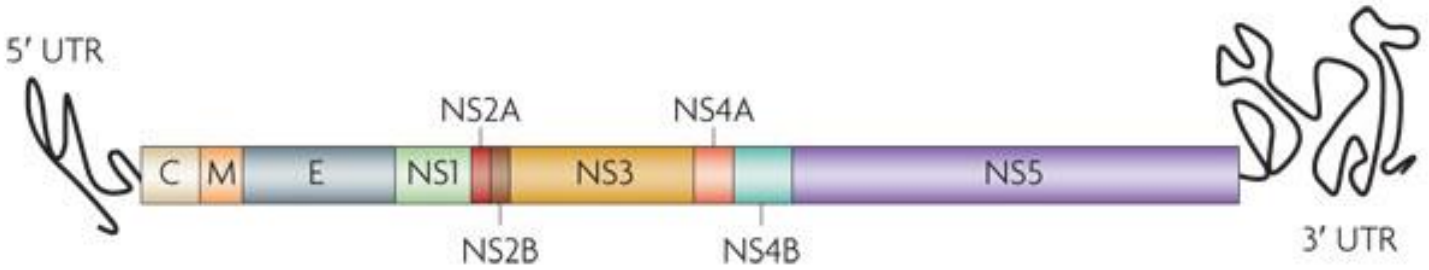
- Size of viruses:

Viruses are tiny



- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm

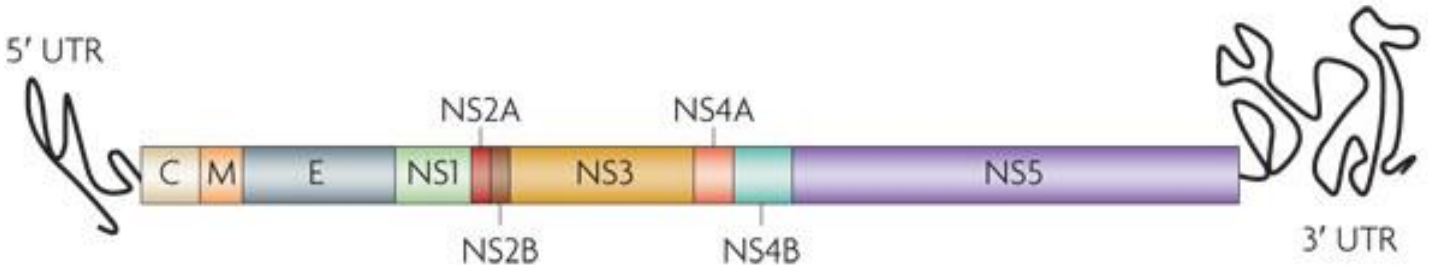
Viruses are tiny



doi:10.1038/nrmicro2460

- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:

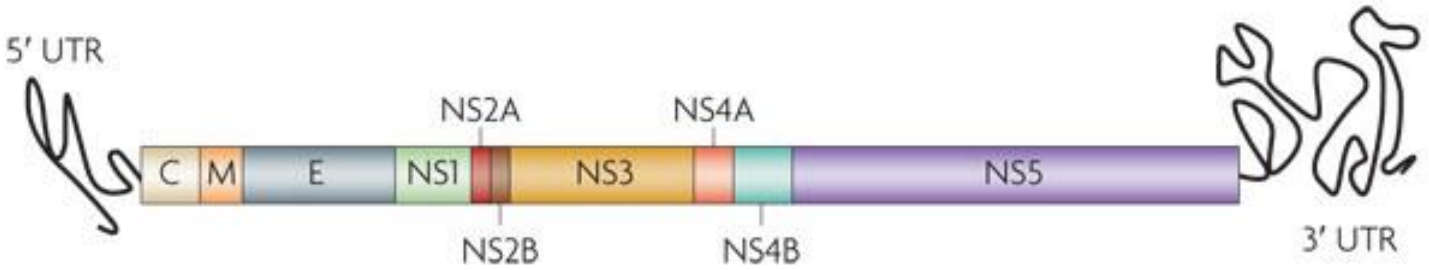
Viruses are tiny



doi:10.1038/nrmicro2460

- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)

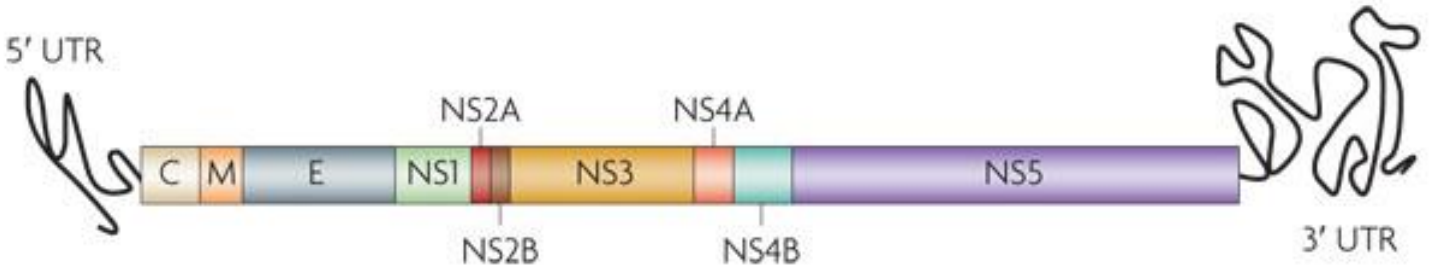
Viruses are tiny



doi:10.1038/nrmicro2460

- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

Viruses are tiny

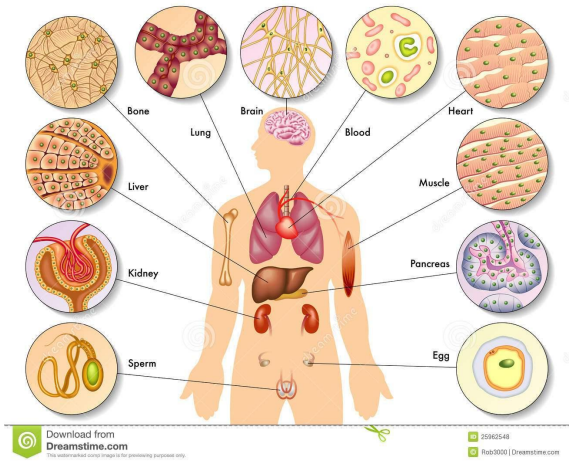


doi:10.1038/nrmicro2460

- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

  Cells@human $\sim 3.72 \cdot 10^{13}$

Viruses are tiny



- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

Cells@human	$\sim 3.72 \cdot 10^{13}$
Bacteria@human	$> 10^{14}$

Viruses are tiny

- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

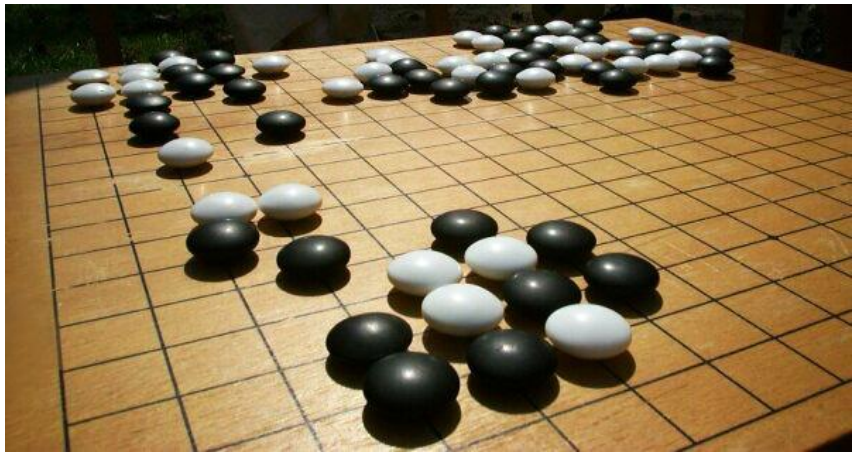
Cells@human	$\sim 3.72 \cdot 10^{13}$
Bacteria@human	$> 10^{14}$
Stars	$\sim 10^{23}$

Viruses are tiny

- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

Cells@human	$\sim 3.72 \cdot 10^{13}$
Bacteria@human	$> 10^{14}$
Stars	$\sim 10^{23}$
Postions on a Go board	$1.7 \cdot 10^{172}$

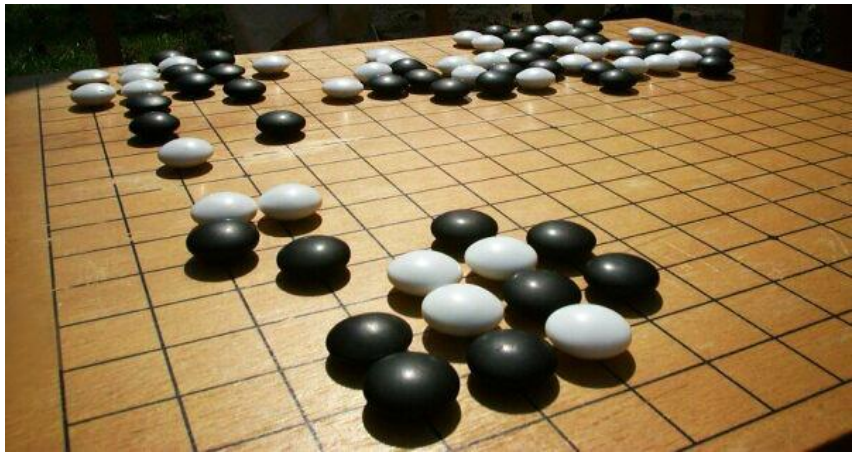
Viruses are tiny



- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

Cells@human	$\sim 3.72 \cdot 10^{13}$
Bacteria@human	$> 10^{14}$
Stars	$\sim 10^{23}$
Postions on a Go board	$1.7 \cdot 10^{172}$
Virus_p@earth	

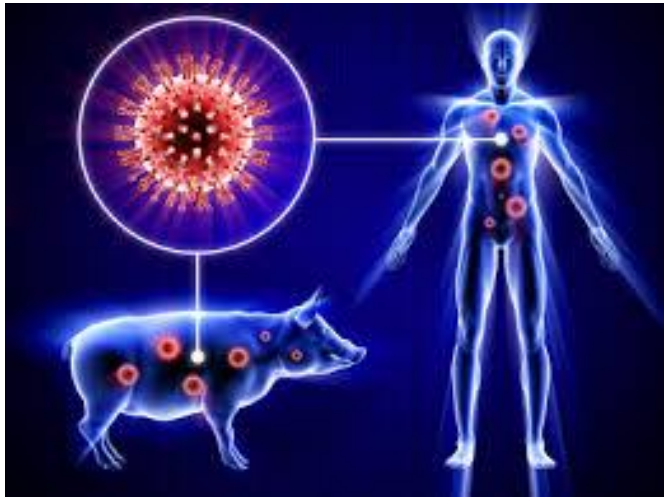
Viruses are tiny



- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

Cells@human	$\sim 3.72 \cdot 10^{13}$
Bacteria@human	$> 10^{14}$
Stars	$\sim 10^{23}$
Postions on a Go board	$1.7 \cdot 10^{172}$
Virus_p@earth	$\sim 10^{31}$

Viruses are tiny

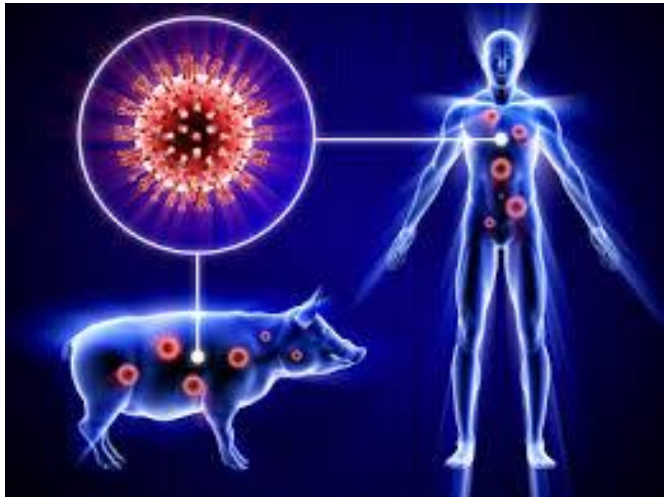


- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

Cells@human	$\sim 3.72 \cdot 10^{13}$
Bacteria@human	$> 10^{14}$
Stars	$\sim 10^{23}$
Postions on a Go board	$1.7 \cdot 10^{172}$
Virus_p@earth	$\sim 10^{31}$

- 3,186 virus species
- 320,000 (unknown) viral species (only @ mammals)

Viruses are tiny



- Size of viruses: 15-440 nm (80 nm); avg. cell size: 10-100 μm
- Size of viral genome:
  3.400-31.000 nt (RNA viruses); $3 \cdot 10^6$ nt (DNA viruses)
- Number of virus particles on earth

Cells@human	$\sim 3.72 \cdot 10^{13}$
Bacteria@human	$> 10^{14}$
Stars	$\sim 10^{23}$
Postions on a Go board	$1.7 \cdot 10^{172}$
Virus_p@earth	$\sim 10^{31}$

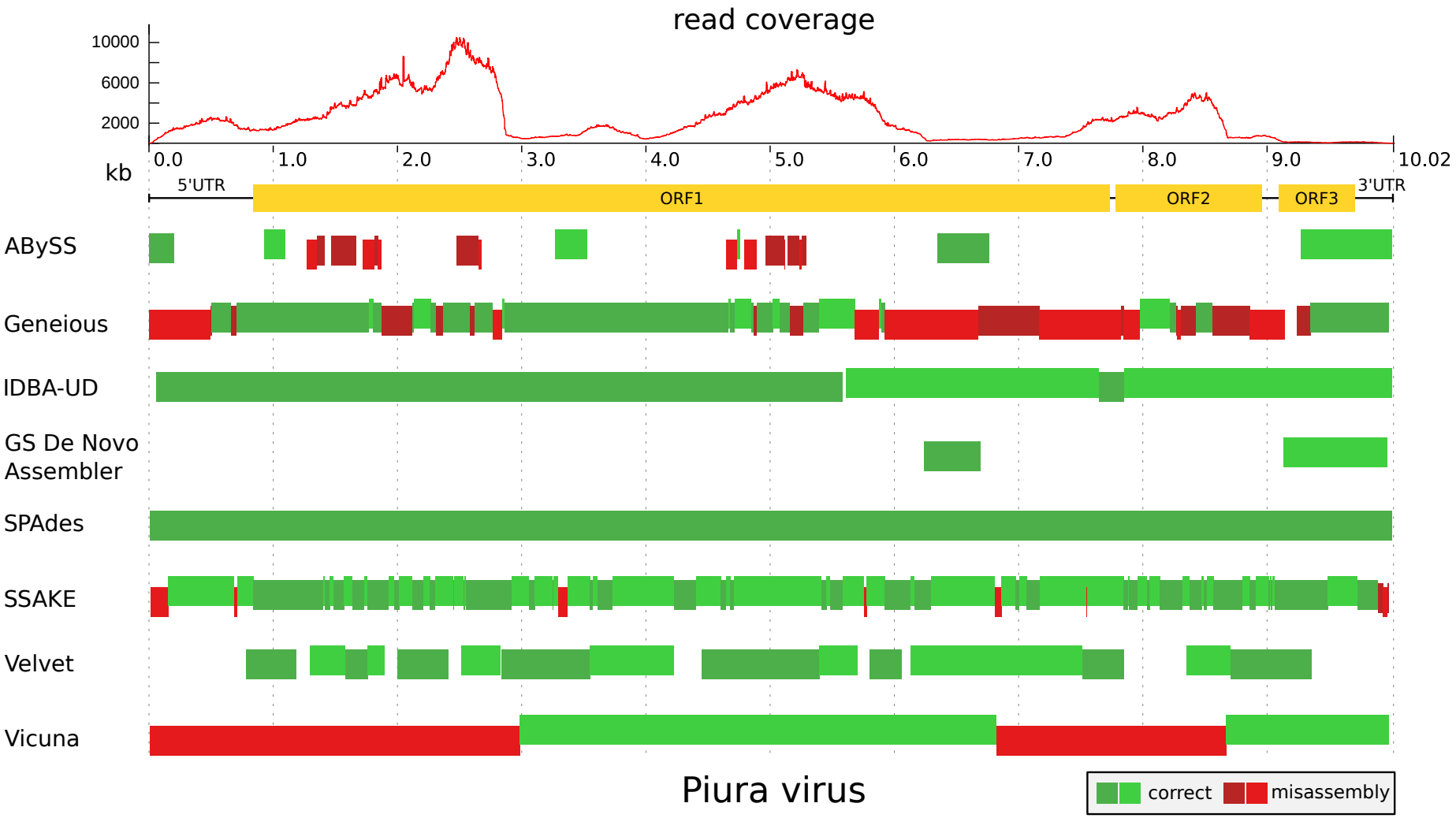
- 3,186 virus species
- 320,000 (unknown) viral species (only @ mammals)
- just 95 years old viruses known (!)
- Human genome (proteins, ncRNAs, viral elements?)
- less than 1% of bioinformaticians deal with viruses

European Virus Bioinformatics Center

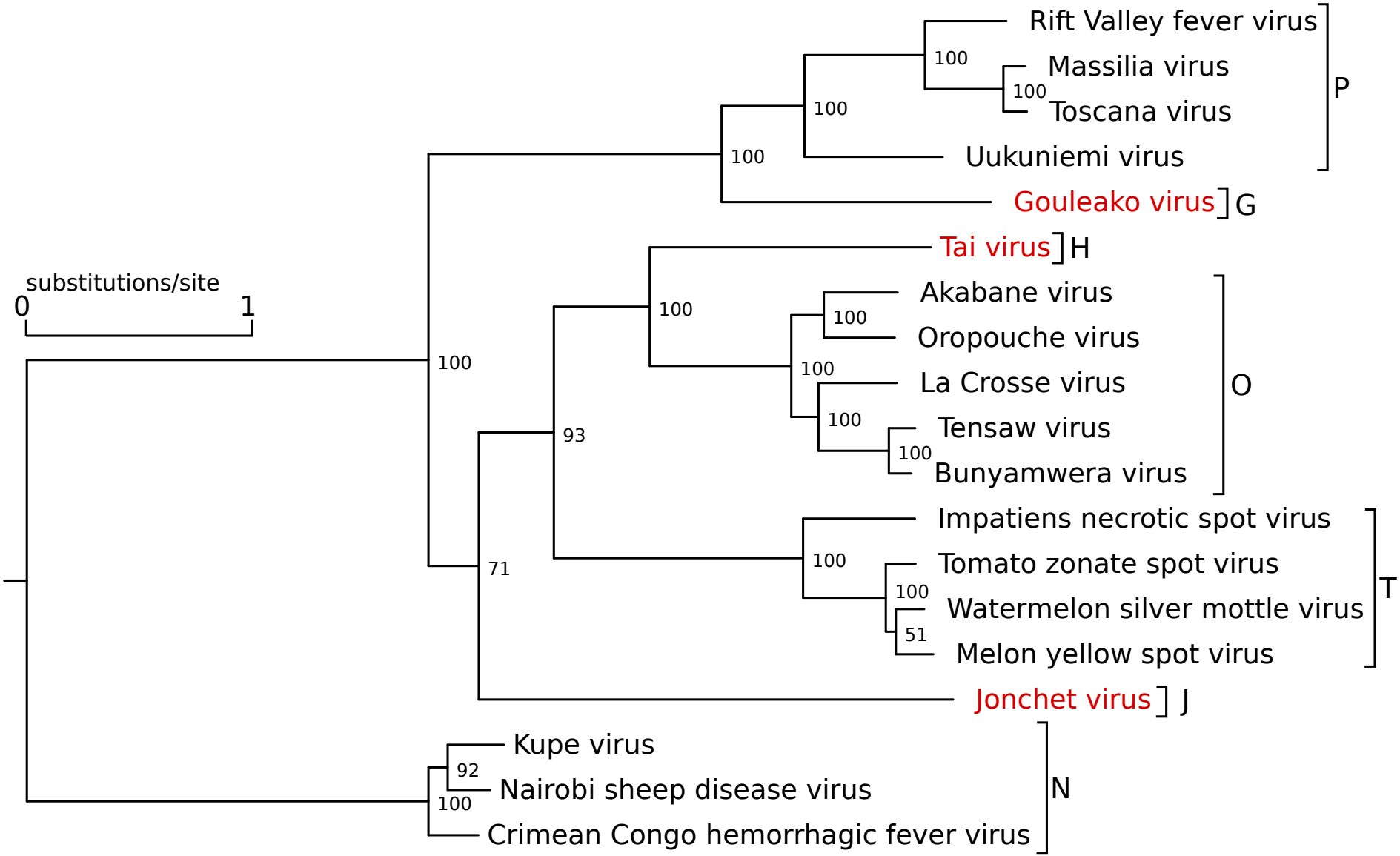
European Virus Bioinformatics Center

www.evbc.uni-jena.de

How to detect novel viruses?



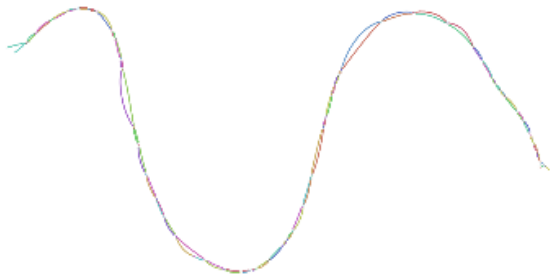
YEAH! Found something!



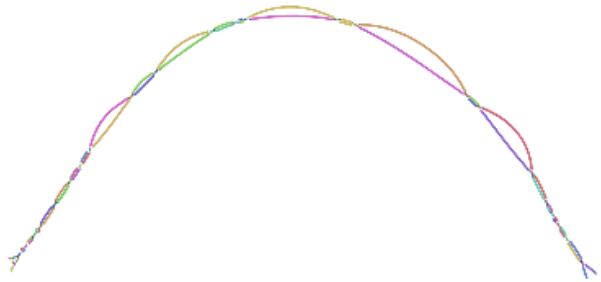
New generation of sequencing methods (2 CoV)

New generation of sequencing methods (2 CoV)

(a) $k = 12$, number of nodes = 535, number of edges = 663

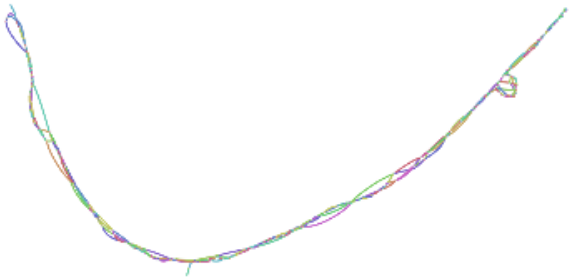


(b) $k = 16$, number of nodes = 155, number of edges = 203

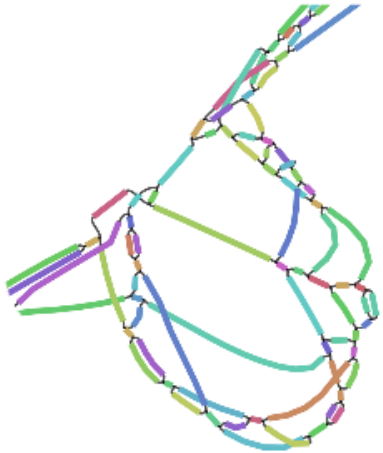


(c) $k = 20$, number of nodes = 75, number of edges = 97

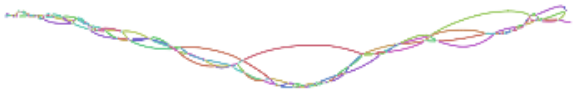
New generation of sequencing methods (3 CoV)



(a) $k = 16$, number of nodes = 609, number of edges = 807

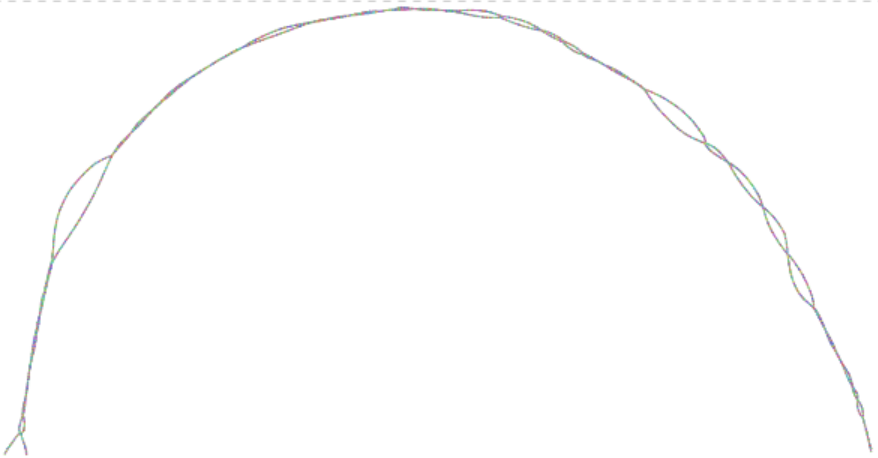


(b) $k = 16$, detail

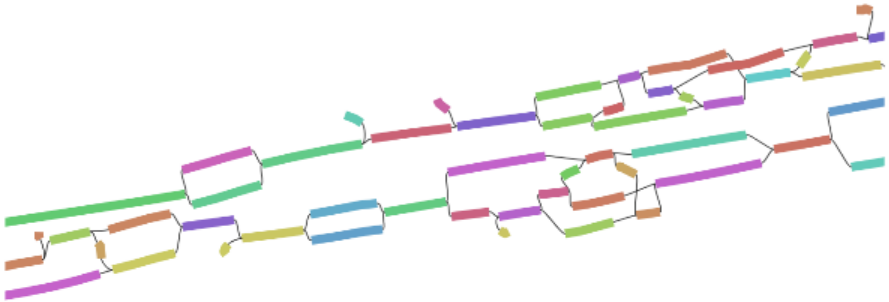


(c) $k = 20$, number of nodes = 289, number of edges = 381

New generation of sequencing methods

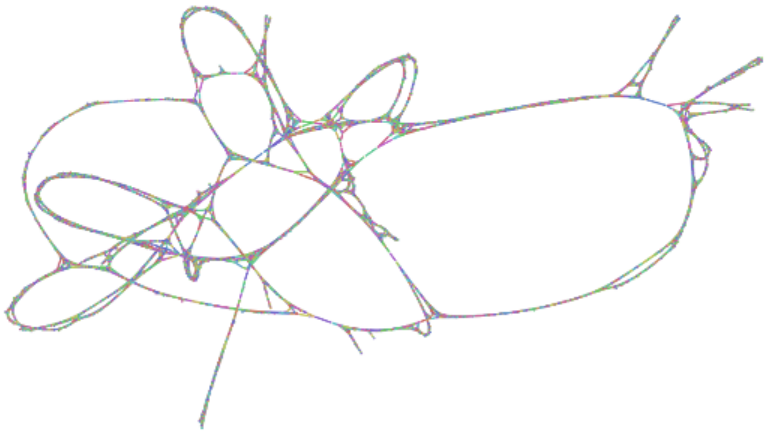


(a) error-probability: 0.5%, $k = 16$, number of nodes = 6170, number of edges = 7840

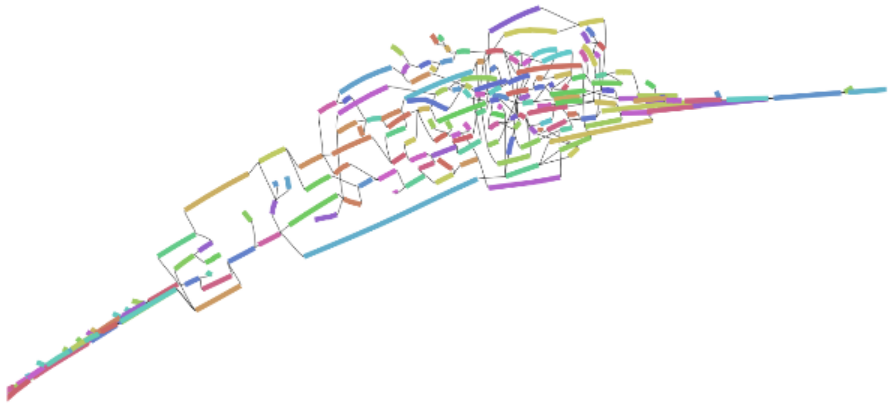


(b) error-probability: 0.5%, $k = 16$, detail

New generation of sequencing methods



(c) error-probability: 5%, $k = 16$, number of nodes = 27017, number of edges = 34084



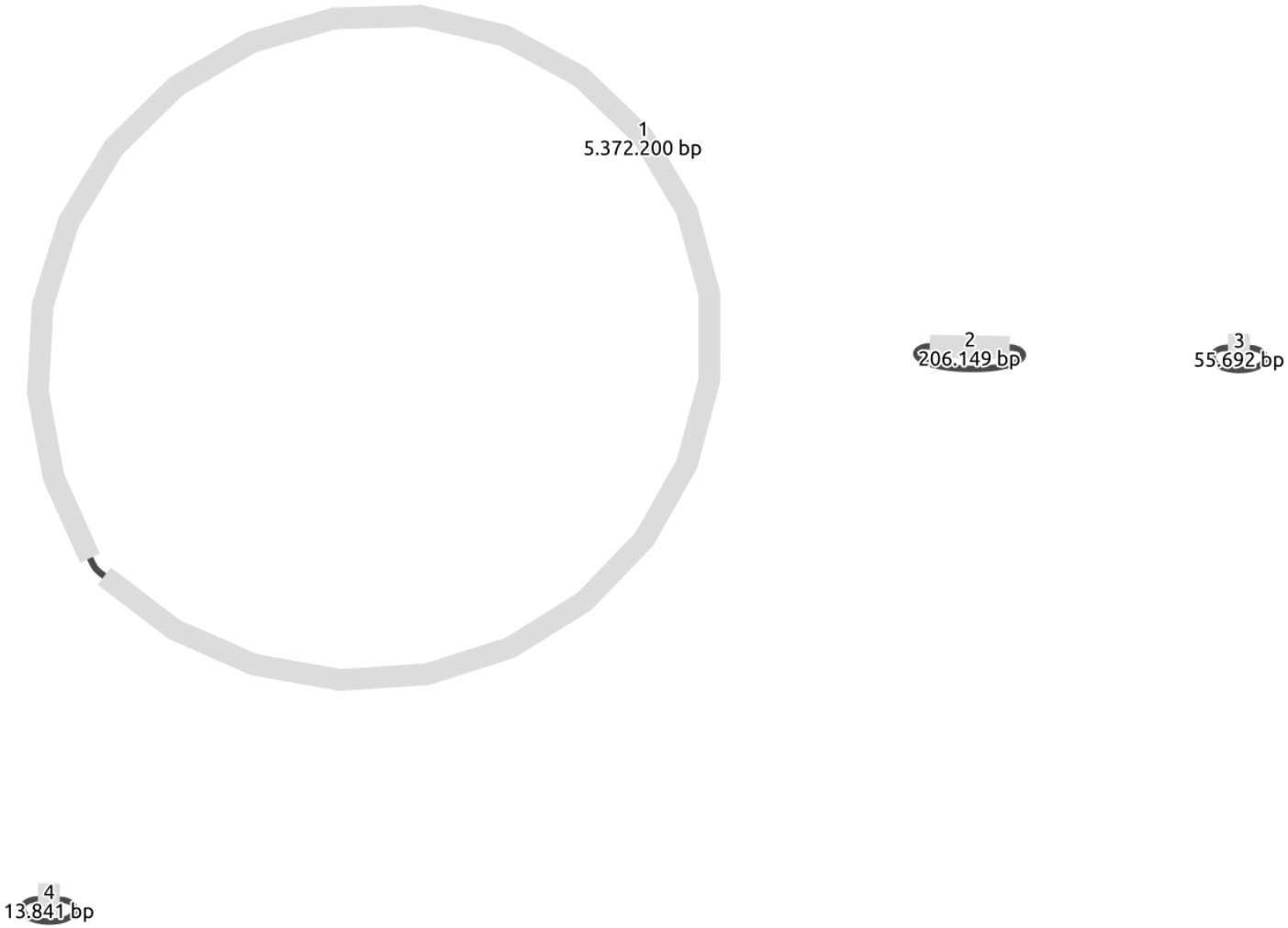
(d) error-probability: 5%, $k = 16$, detail

New generation of sequencing methods

Minion

Klebsiella pneumoniae carbapenem

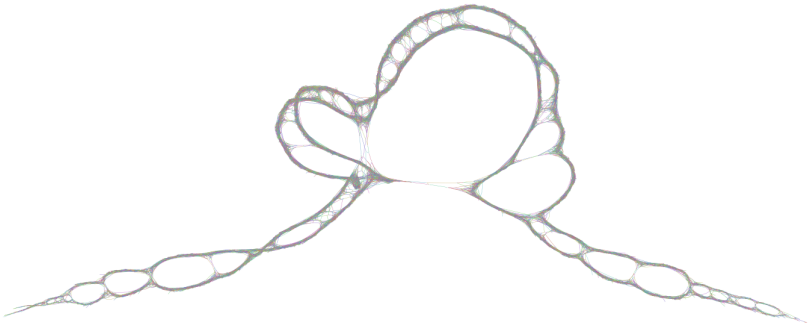
Klebsiella pneumoniae carbapenem



Minion – 2 viruses (BVDV)

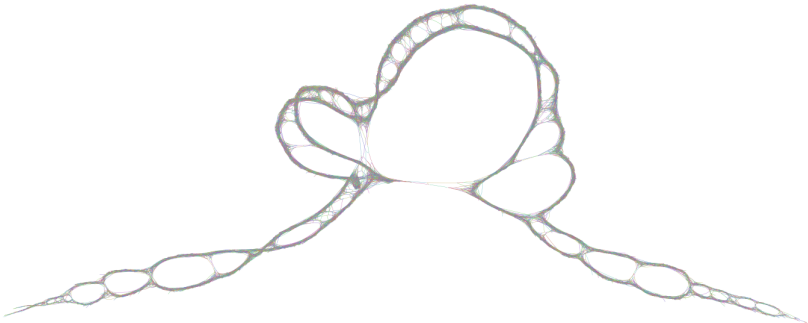
Minion – 2 viruses (BVDV)

e=0.10 k=20

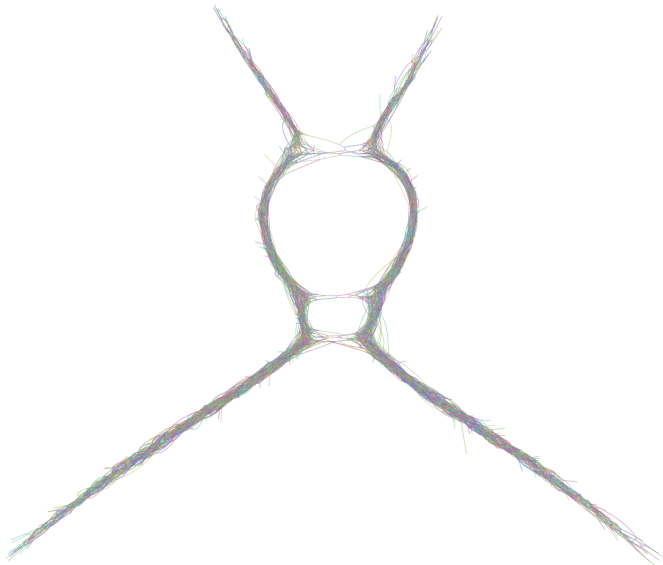


Minion – 2 viruses (BVDV)

e=0.10 k=20

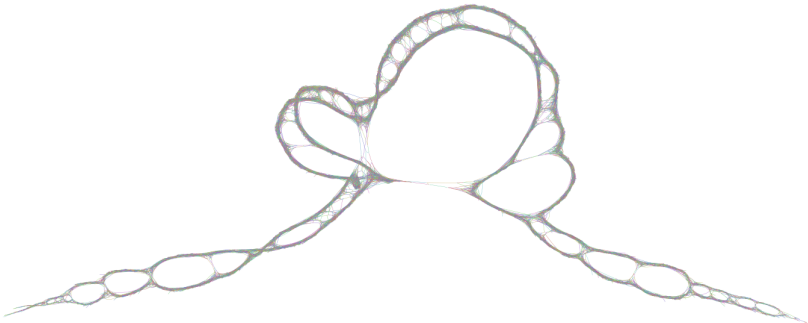


e=0.10 k=30

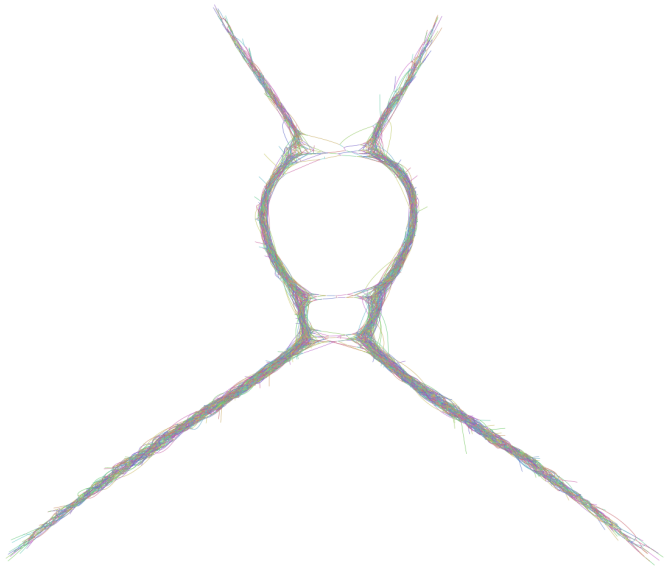


Minion – 2 viruses (BVDV)

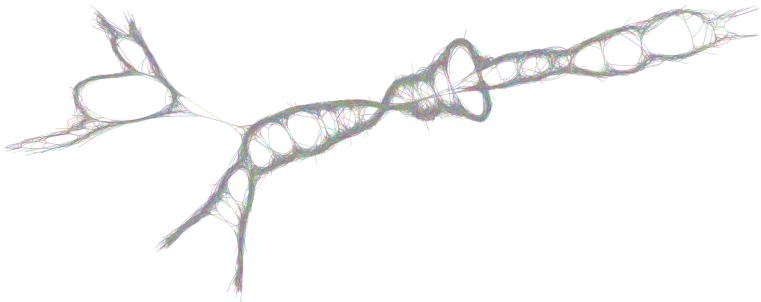
e=0.10 k=20



e=0.10 k=30

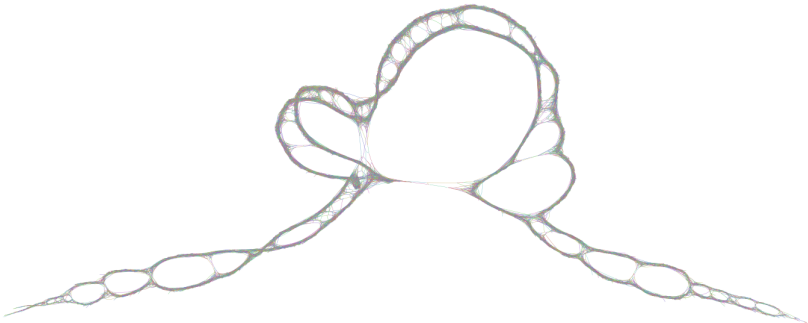


e=0.15 k=20

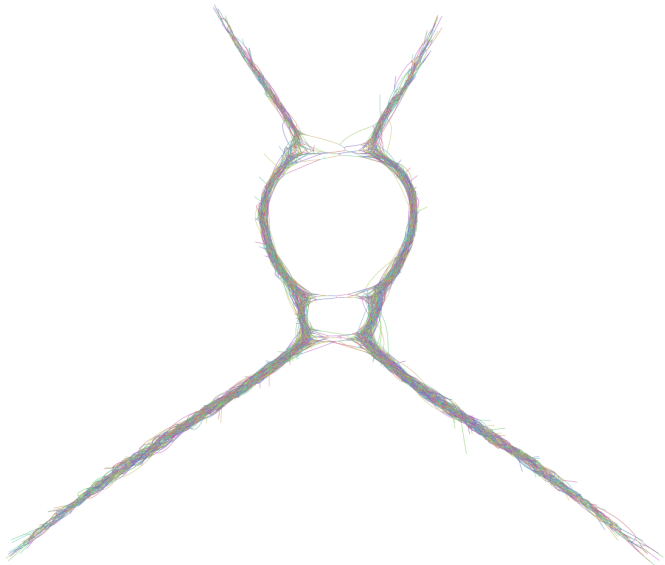


Minion – 2 viruses (BVDV)

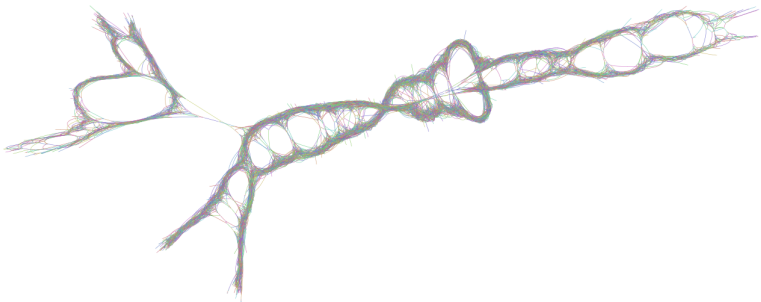
e=0.10 k=20



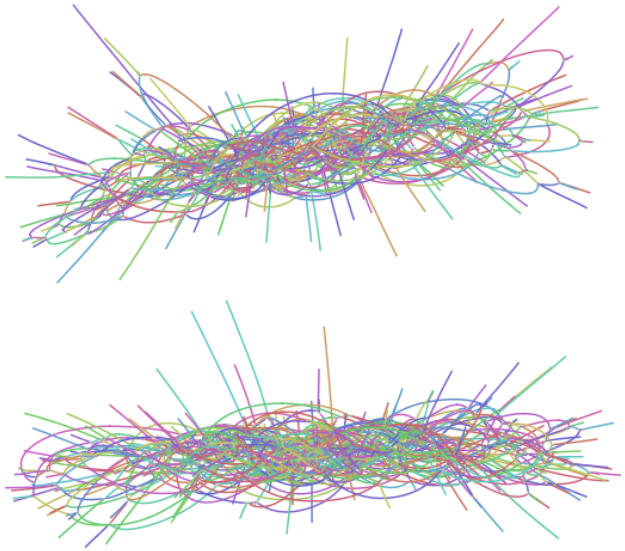
e=0.10 k=30



e=0.15 k=20

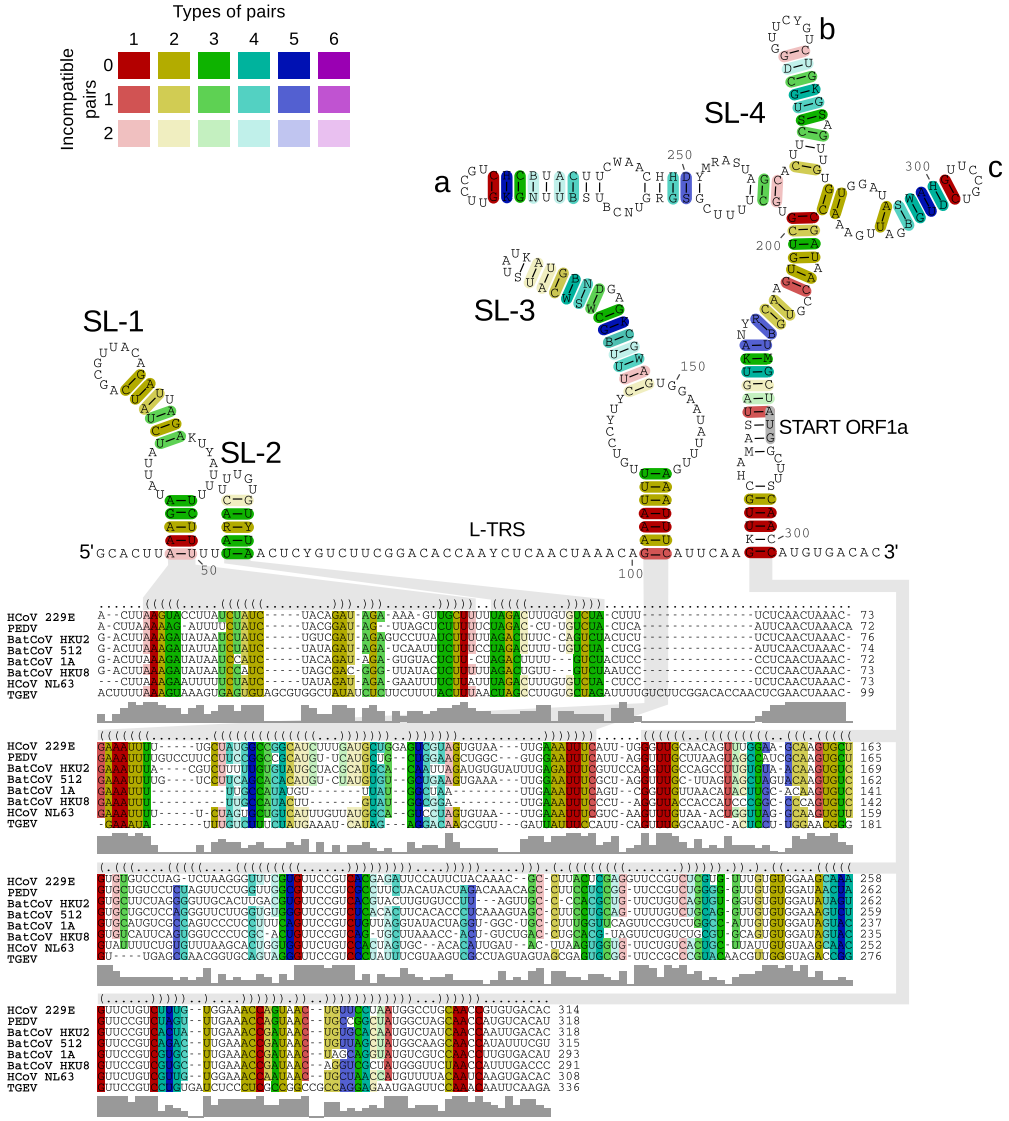


e=0.15 k=30



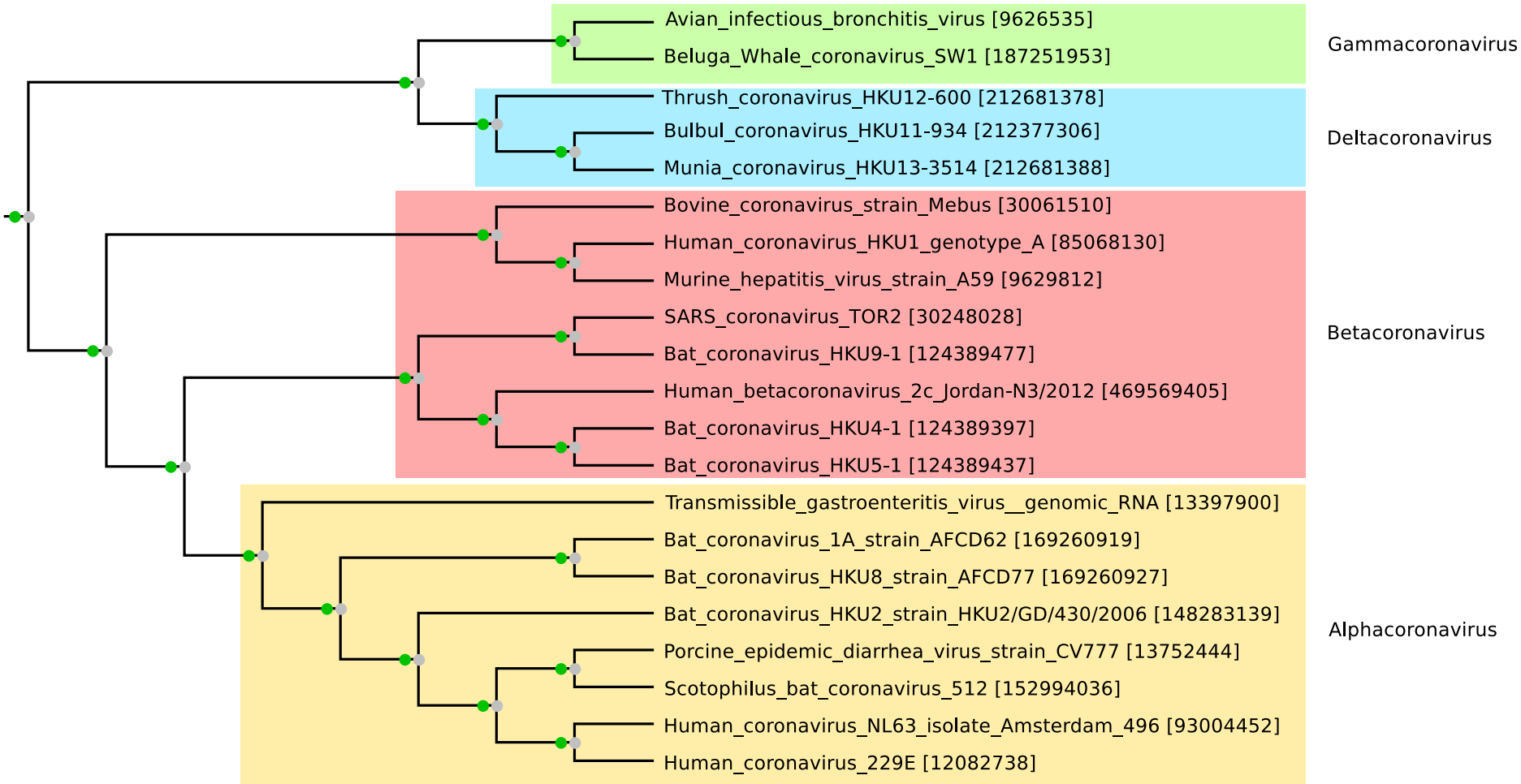
Secondary structures in RNA viruses: Coronaviruses

Secondary structures in RNA viruses: Coronaviruses

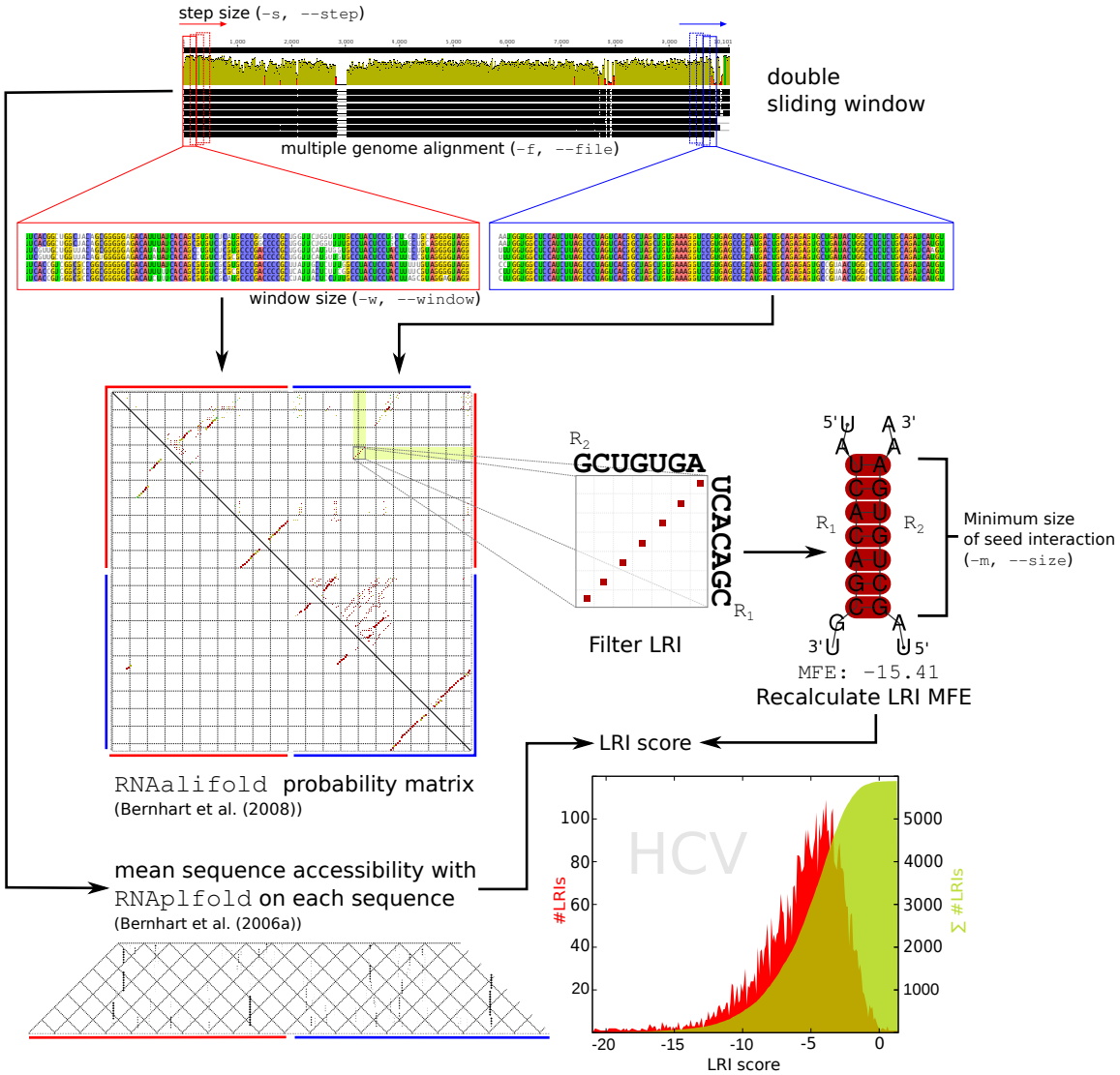


Clustering of secondary structures

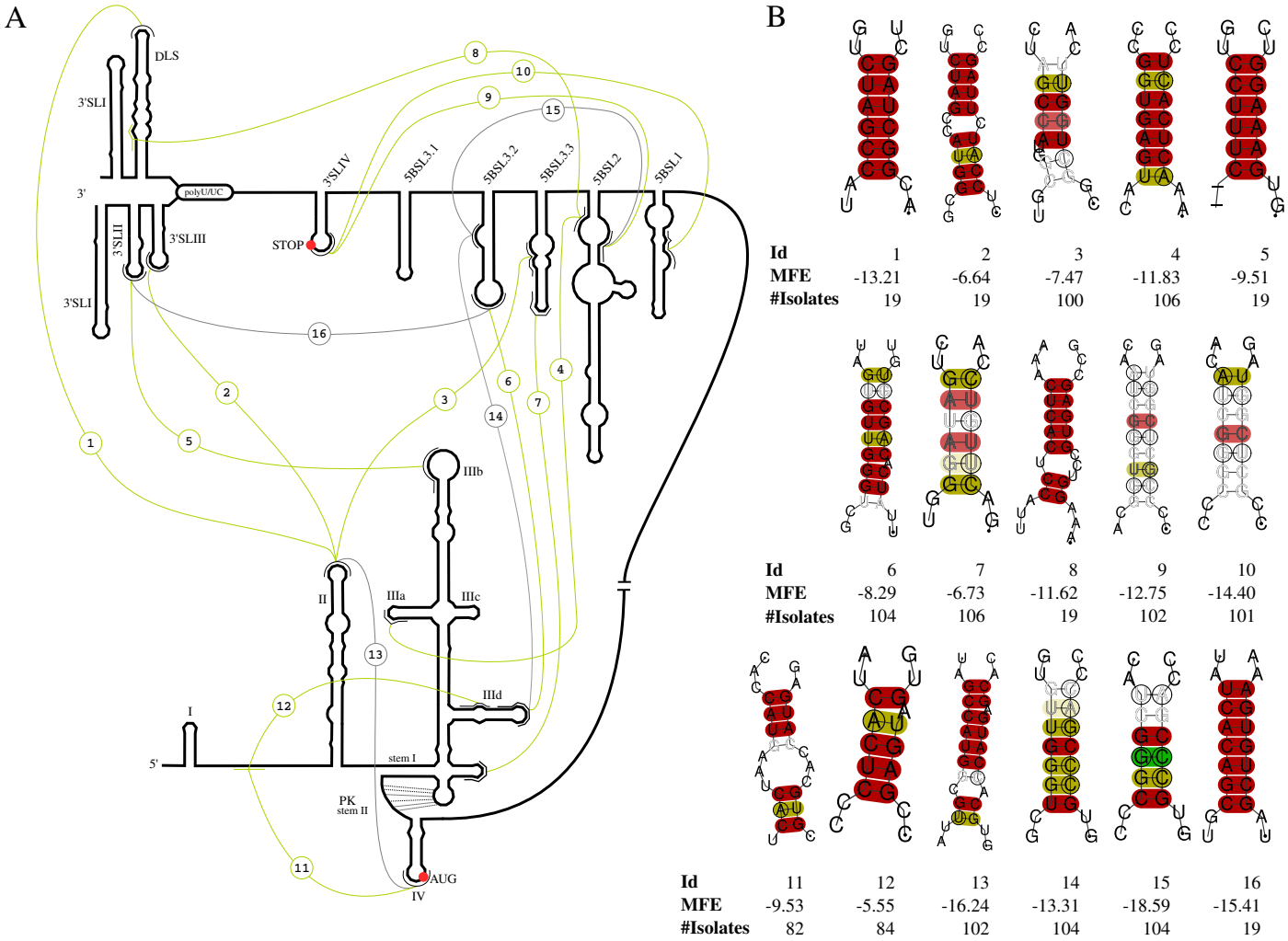
Clustering of secondary structures



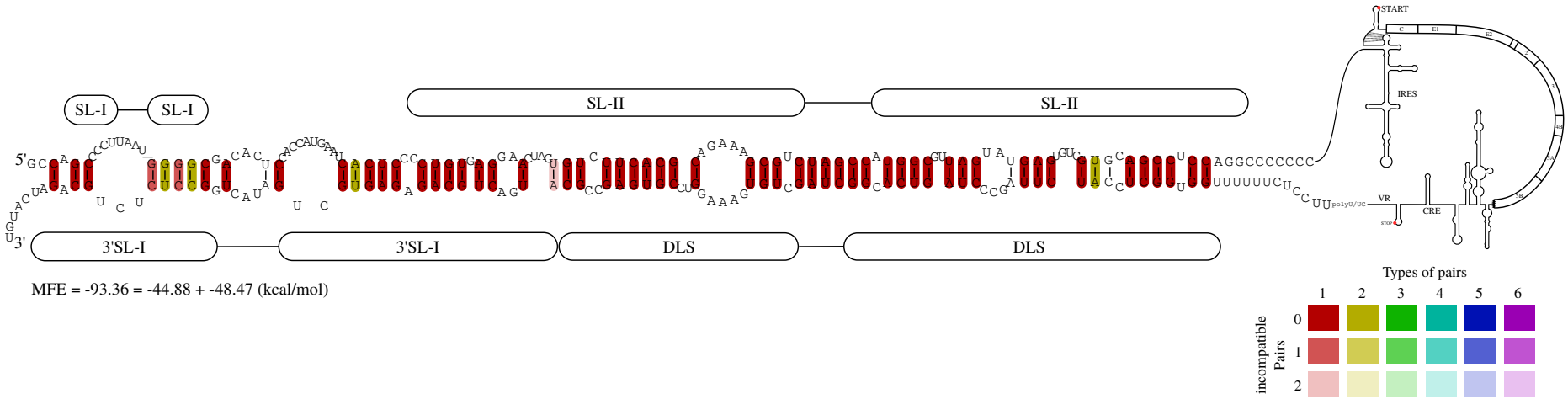
Secondary structures in RNA viruses: Long-range interactions



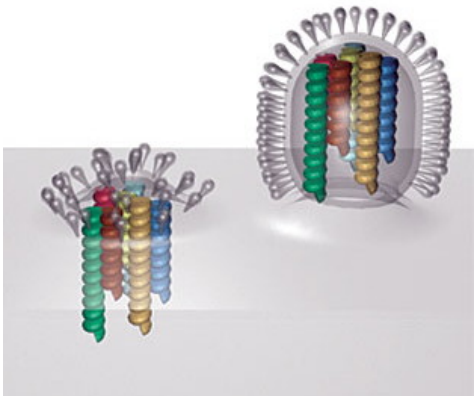
Impact of secondary structures in RNA viruses: Long-range interactions in HCV



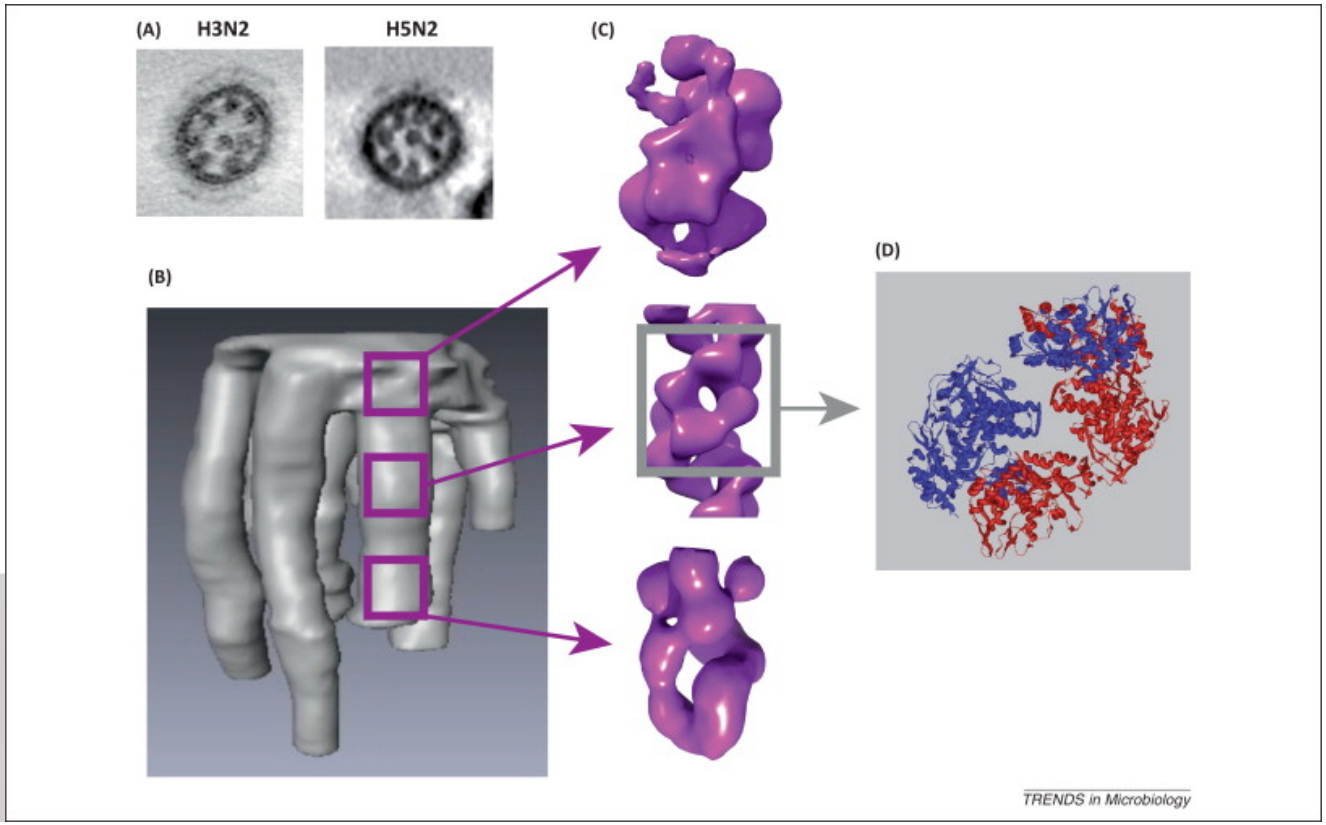
Secondary structures in RNA viruses: Circularization of HCV



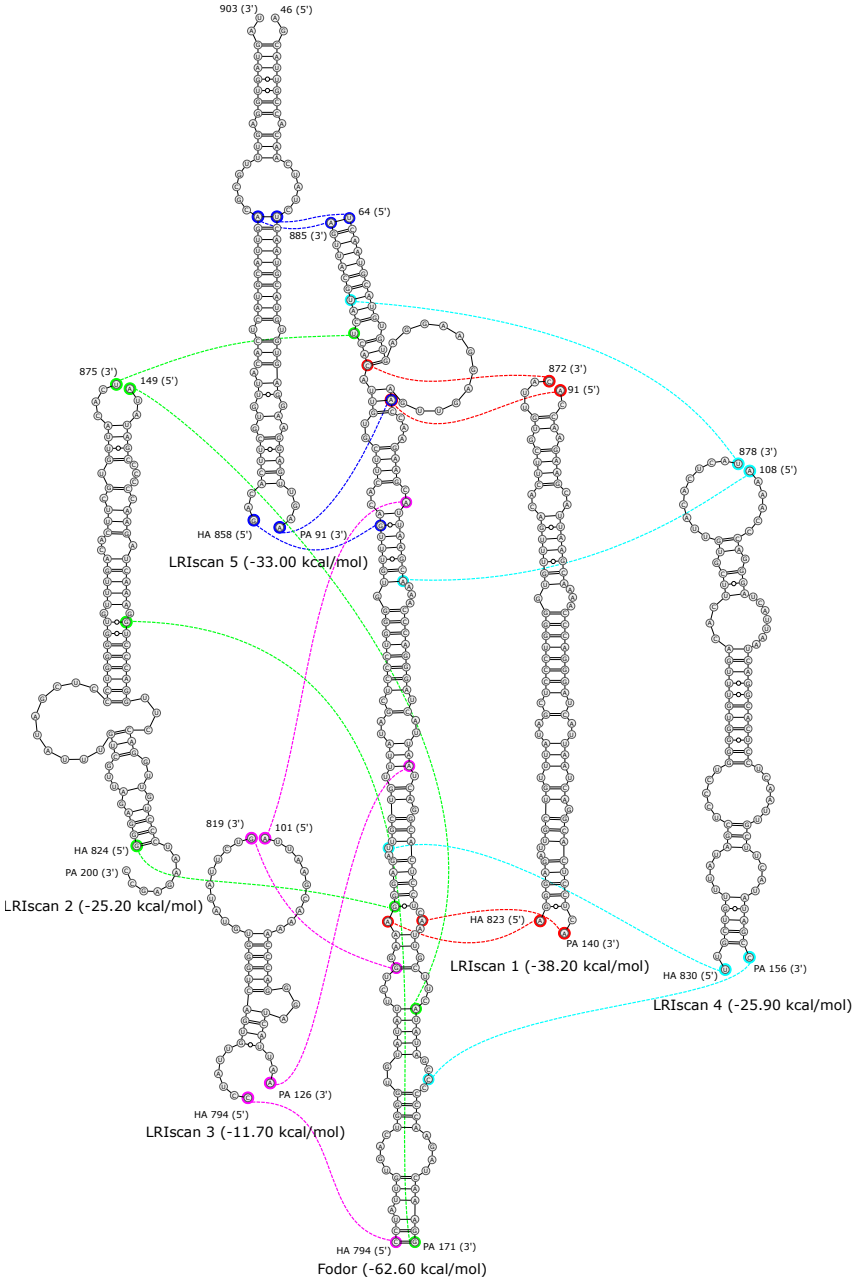
Packaging in Influenza



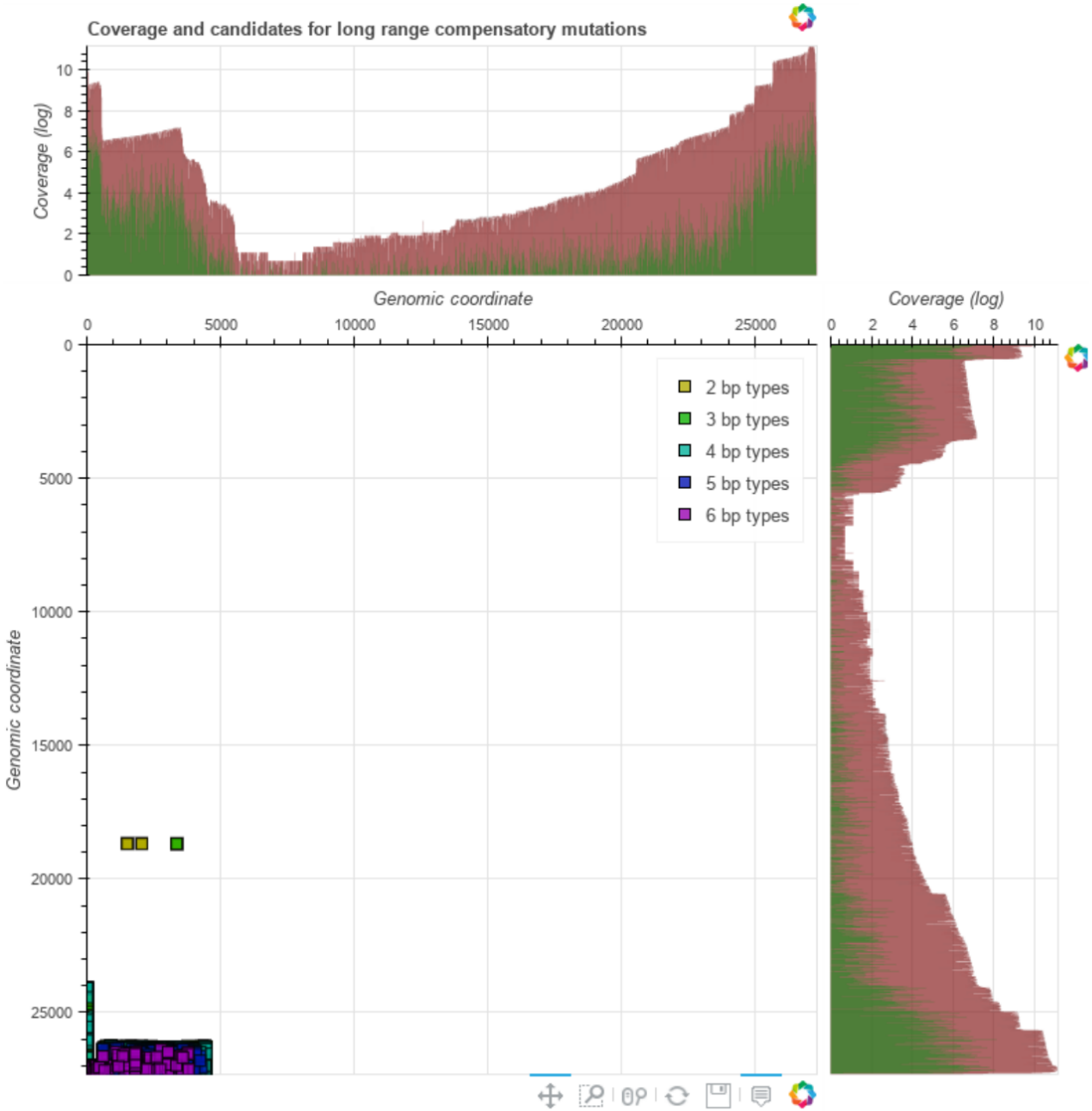
Packaging in Influenza



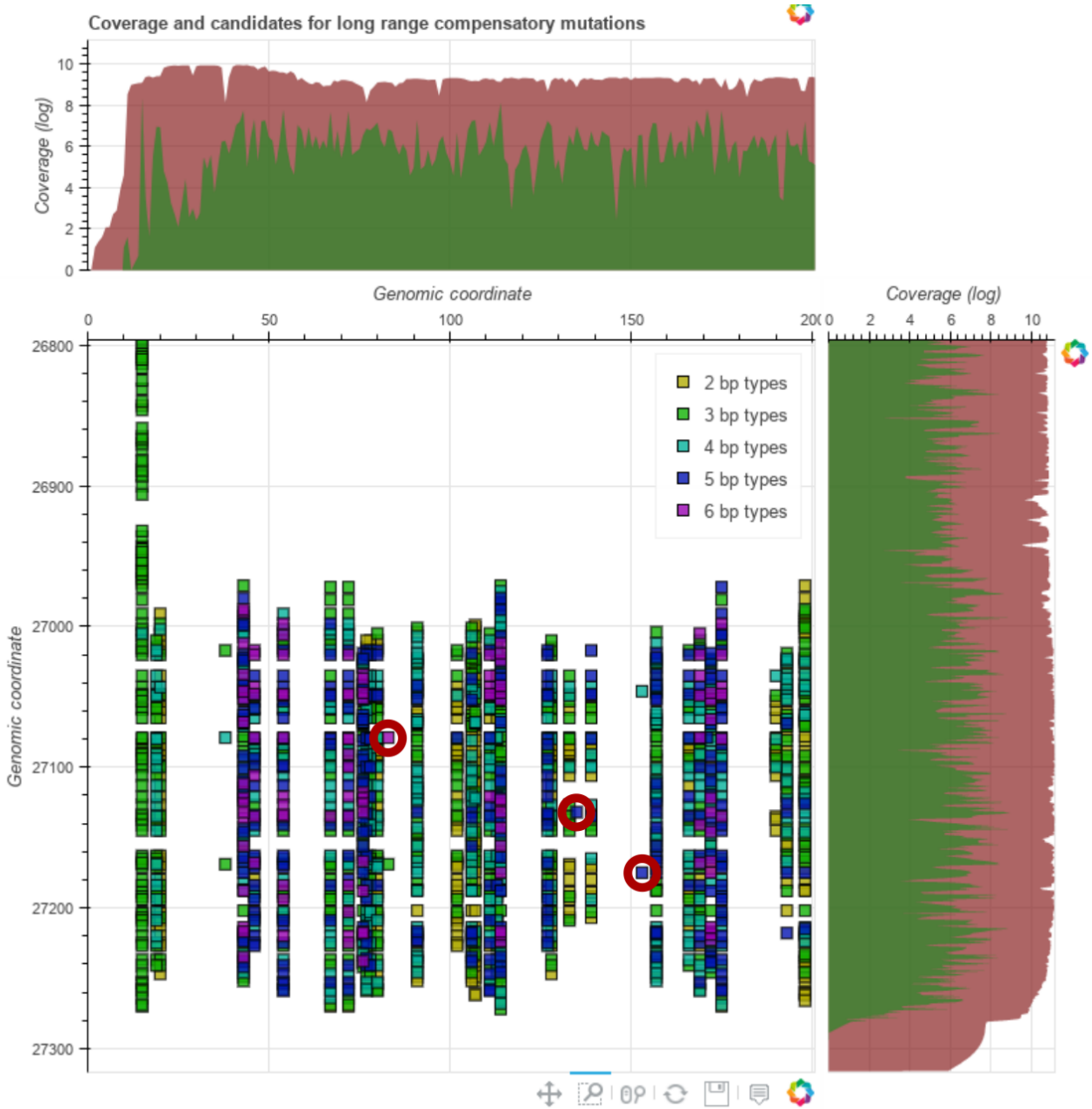
Packaging in Influenza



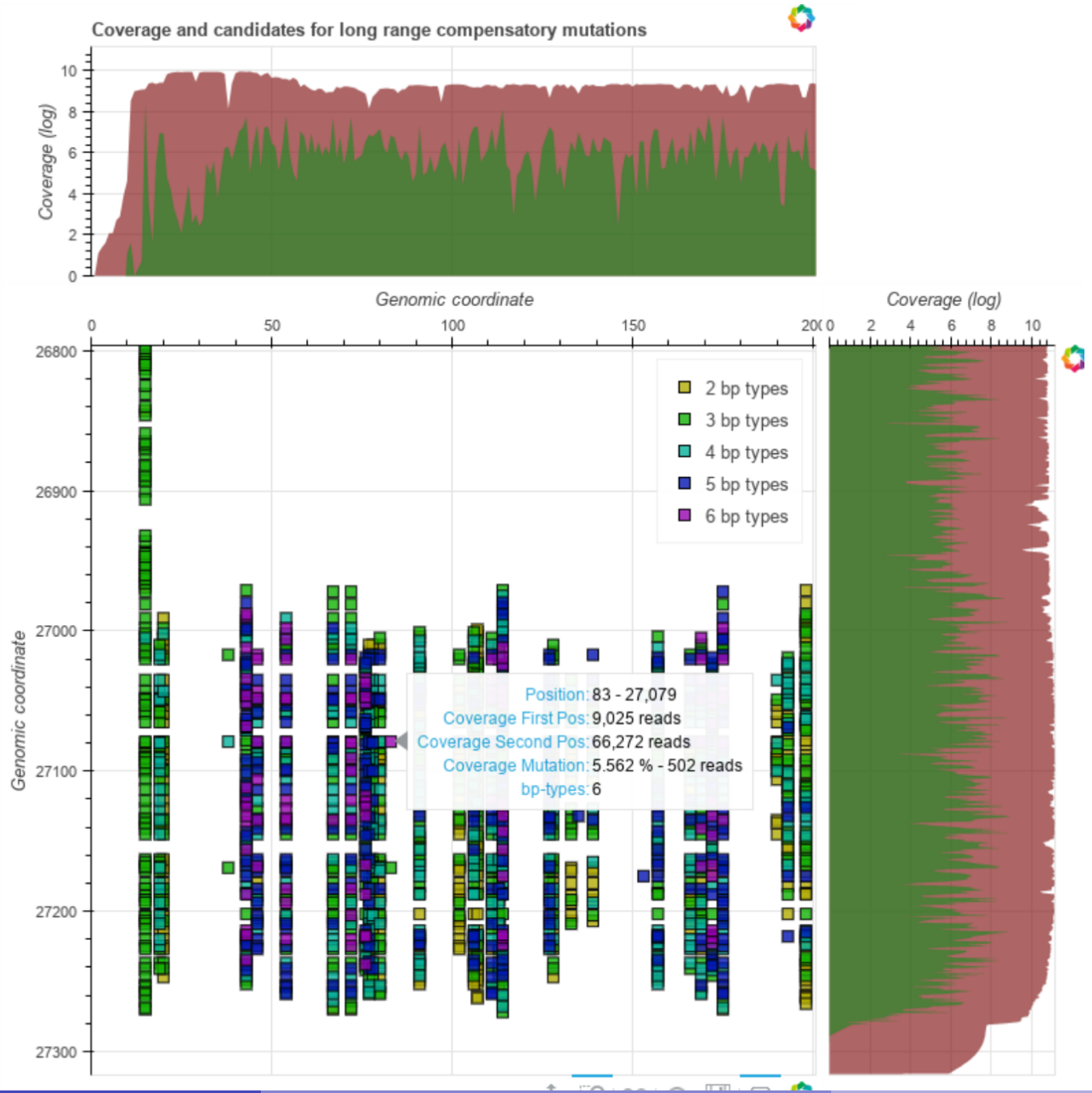
Use Minion: Location



Use Minion: Location



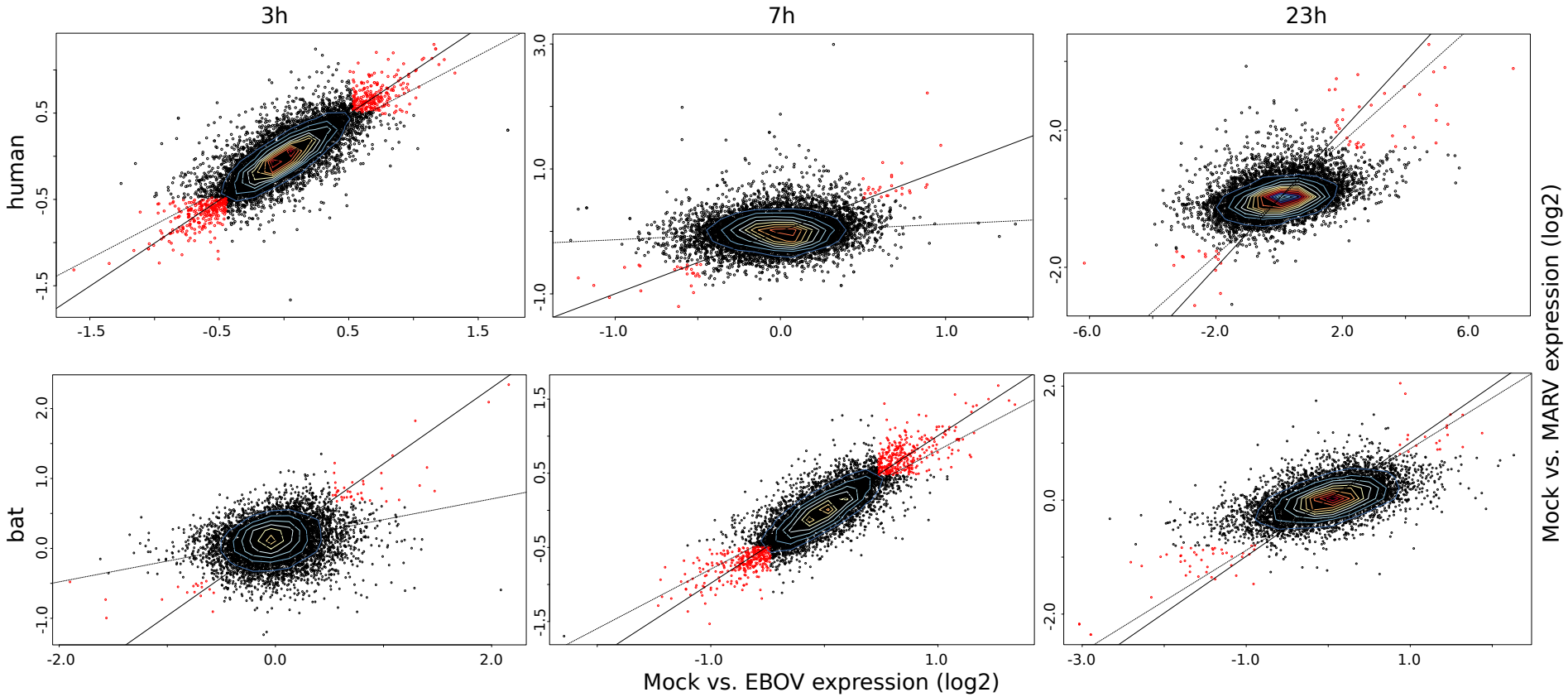
Use Minion: Location



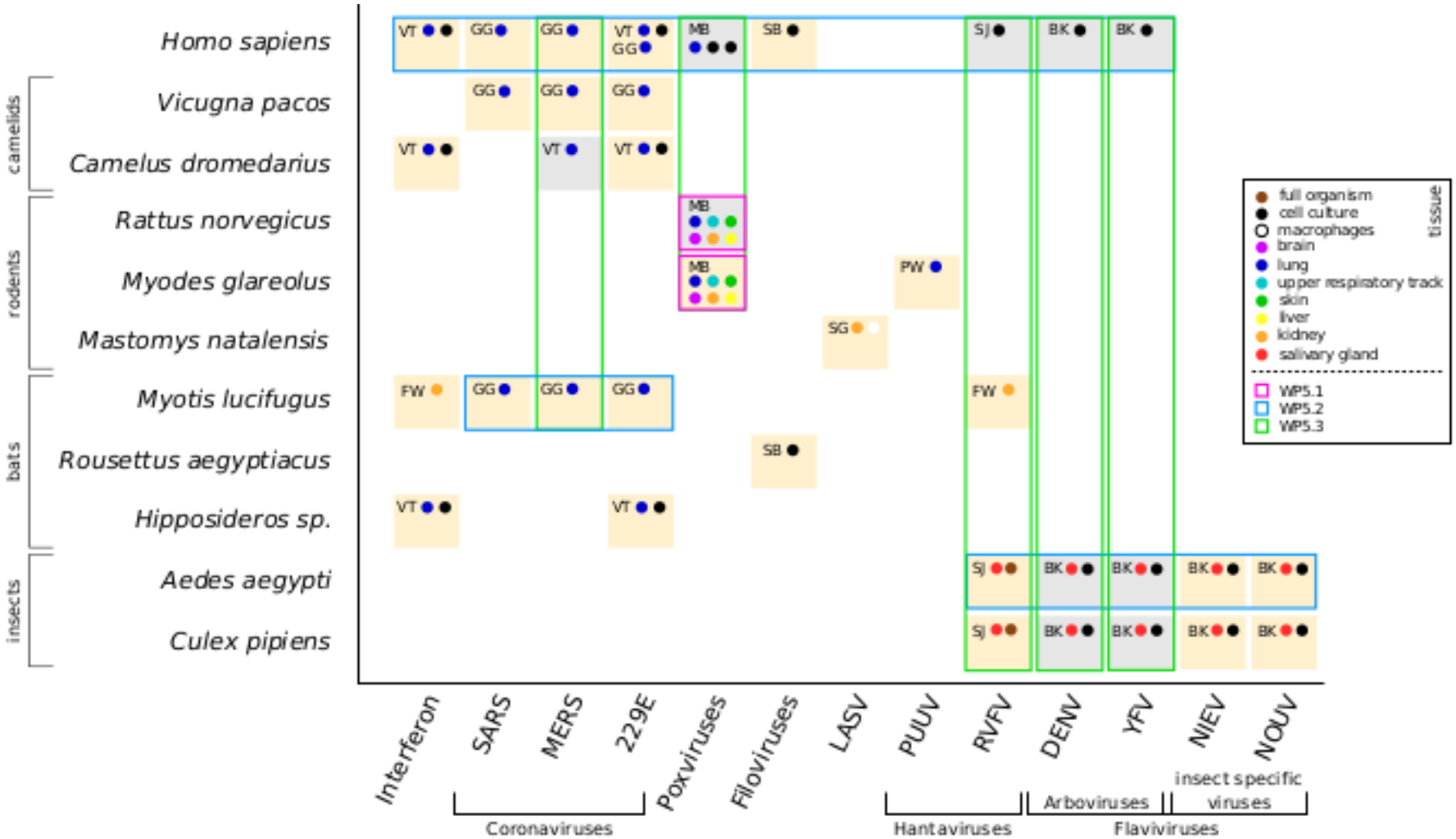
Hackathon – Virus-Host Interaction – Ebola

Hackathon – Virus-Host Interaction – Ebola

$\sim$ 50% of reads map onto viral genome

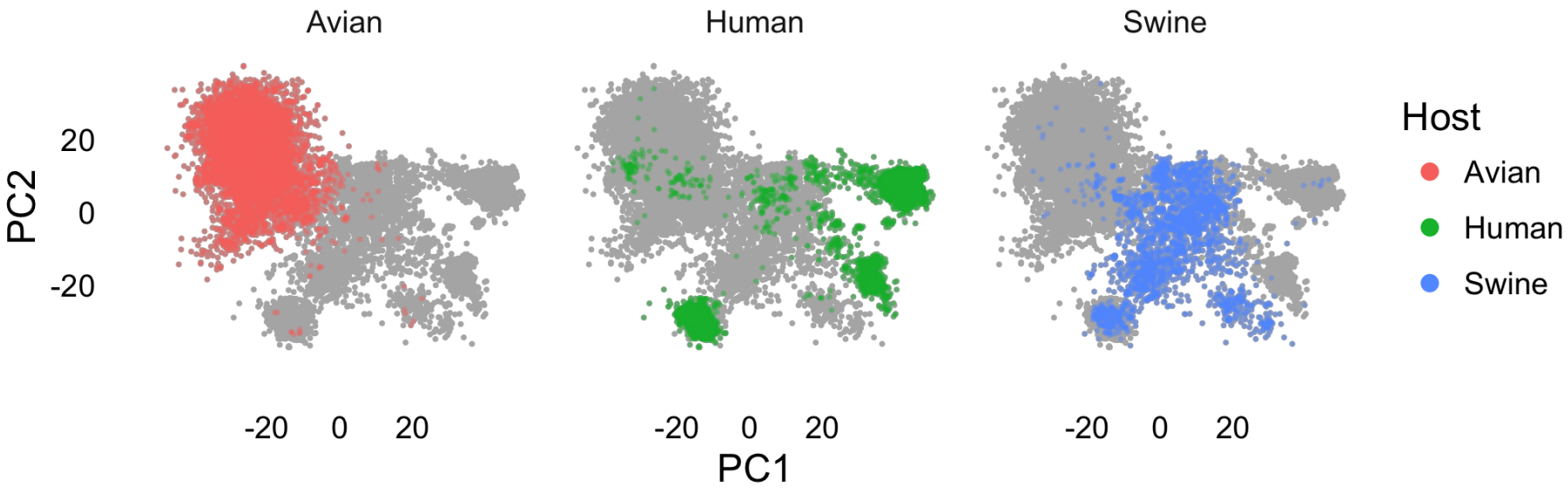


A systematic approach in understand host reactions



Give me your virus and I tell you the host

Give me your virus and I tell you the host



Give me your virus and I tell you the host

$y \backslash \hat{y}$	Avian	Human	Swine	All
Avian	3207	49	13	3269
Human	6	4470	82	4558
Swine	9	10	849	868
All	3222	4529	944	8695

metric \ host	Avian	Human	Swine	All
accuracy				0.95
recall	0.99	0.95	0.94	
precision	0.88	1.00	0.89	
F1-score	0.94	0.97	0.93	

Codon usage

$y \backslash \hat{y}$	Avian	Human	Swine	All
Avian	5524	227	576	6327
Human	168	11314	1251	12733
Swine	25	800	995	1820
All	5717	12341	2822	20880

Dinucleotides

$y \backslash \hat{y}$	Avian	Human	Swine	All
Avian	4294	1048	985	6327
Human	173	8376	4184	12733
Swine	17	630	6342	1820
All	4484	10054	6342	20880

A database of viruses?

A database of viruses?

Zitat: “NEIN NEIN NEIN”

- data structure
- various viruses
- 'pan genomics'
- photos
- alignments
- links
- data policy
- partial update

A database of viruses?

A database of viruses?

```
# Create a JSON schema from zoo's templates to validate any data cell
insertions.
zoo schema (core(metadata(influenza),annotation)) \
> schema.json
# Or use your own.
zoo schema --fp path/to/file (a(b,c)) > schema.json

# Stream GenBank records to data cell, and validate schema.
zoo load --source ncbi --fmt json \
--ids accessions.txt --stdout - | \
zoo init --db mockA --cell foo --validate schema.json -
# ... Initializing data cell.
# ... 42 entries inserted into cell "original".
# ... Primary key assigned to field "_id".
# ... inspect cell and commit

zoo status --db mockA --cell foo --example

# Make and commit changes (like you would with Git).
zoo commit --db mockA --cell foo original
# ... Dumping data cell.
# ... Minhash signature computed for molecule type: DNA
```

A database of viruses?

```
# share
mkdir send
cp original.json send/
dat share send/
# ... Syncing Dat Archive: .../send
# ... Link:
# dat://73401e1b931164763eccsomelonglinkcefc718ebf49f6b4fe4dbad7

# In a faraway place, our collaborator (B) clones a copy of our
cell and adds it to her "zoo" of other data cells.
mkdir receive
dat clone <link> receive/
zoo add --db mockB --cell foo --primkey genbank.accession \ receive/original.
json
# ... Loading data cell.
# ... Index created on field "genbank.accession".
# ... 39 documents inserted in cell "foo".
# ... 3 duplicates skipped.

# Meanwhile, original.json was modified. B want his zoo to reflect
# the changes:
dat pull receive/
```

A database of viruses?

```
# diff it
zoo diff --db mockA --cell foo bar.json > diff.json
# ... Searching for changes (delta).
# ... Done.
# We can pipe this, too.
zoo diff --db mockA --cell foo bar.json | head -n2
# Apply changes to data cell.
zoo diff --patch --db mockA --cell foo diff.json
# ... Loading and applying delta.
# ... Done.


# pull
zoo pull --db mockB --cell foo receive/modified.json
# ... Updating cell's md5 hashes.
# ... / 0 Elapsed Time: 0:00:00
# ...
# ... 38 entries unchanged.
# ... 4 entries replaced.
```

A database of viruses?

```
# Now put data cells into your favourite analysis workflow,
# then use zoo's API to import/ export the results, like
# multiple sequence or reference-based alignments, phylogenetic
# trees, secondary structure ... happy exploratory
# data analysis. Also, set global vars to reduce typing.
ZOODB=mockB
ZOOCELL=foo
zoo digest --encode tree.nexus
zoo digest --decode msa.mafft.fa

# Not yet implemented: Send metadata about cell to a registry,
# so others can discover it.
zoo push ...

# Create a sequence Bloom tree (SBT) from the minhash
# signatures of a given cell.
zoo sbt_index --db ref --cell virus --ksize 16 --nsketch \
1000 virusref
# ... Initialize SBT.
# ... Compute minhash signatures for selected documents.
# ... k-mer size: 16, sketch size: 1000
# ... \ 9158 Elapsed Time: 0:01:45
# ... Save SBT.
```

A database of viruses?

```
# Export, e.g. to fasta, JSON or stdout.
zoo dump --query q.json --selection \
_id,meta.date,meta.geo.cou,seq \
--delim "|" --fmt fasta dump.fa

zoo dump --query q.json --selection _id,seq \
--fmt fasta - | head

# Pipe into sourmash.
zoo dump --query q.json --selection _id,seq --fmt fasta - | \
sourmash compute -k 16 -n 100 --singleton --out q.sig -

# Done, lets get some coffee.
zoo drop --db mockB --cell foo --force
zoo destroy --db mockB --force
```

Acknowledgements

Acknowledgements

- *RNA Group Jena:* Emanuel Barth, **Nelly Mostajo Berrospi**, **Markus Fricke**, Martin Hölzer, Franziska Hufsky, **Konrad Sachse**, Monika Keilich, Diana Morales, Wittaya Chaiwangyen, Akash Srivastava, **Florian Mock**, **Daniel Desiro**, Maximillian Collatz, Lisa Barf, **Adrian Viehweger**, **Sebastian Krautwurst**, Marie Lataretu, Bashar Ibrahim, **Kevin Lamkievicz**, Javed Iqbal, **Celia Diezel**
- *Marburg:* **Stephan Becker**, Marcus Lechner
- *Giessen:* **John Ziebuhr**, **Michael Niepmann**, **Friedemann Weber**
- *Berlin:* **Christian Drosten**
- *München:* **Dimitrij Frishman**
- *Freiburg:* Rolf Backofen, **Georg Kochs**, Annegret Wilde, Wolfgang Hess
- *Leipzig:* Peter Stadler, Sonja Prohaska, Steve Hoffmann
- *Lübeck:* Christine Klein, **Oliver Tautz**
- *FLI Riems:* Volkmar Liebscher, **Martin Beer**
- *Wien:* Ivo Hofacker, Christoph Flamm, Josef Leydold, **Peter Schuster**
- *Paris:* Oliver Bensaude, Anne-Catherine Dock-Breggeron
- *Glasgow:* **Massimo Palmarini**
- *Arizona:* Julian Chen
- *Christchurch:* Paul Gardner
- *Zürich:* **Niko Beerenwinkel**

DFG SPP-1569: "Ecology and species barriers in emerging viral diseases"; **ZAJ – RegenerAging**; **CRC-1076**: "AquaDiva"; **CRC-TR124**: "FungiNet"; **DAAD**; **iDiv**; **InfectControl 2020**; **DFG MA 5082/9-1**; **MGH, CCXDP**