Assessing recombination detection methods using viral simulations

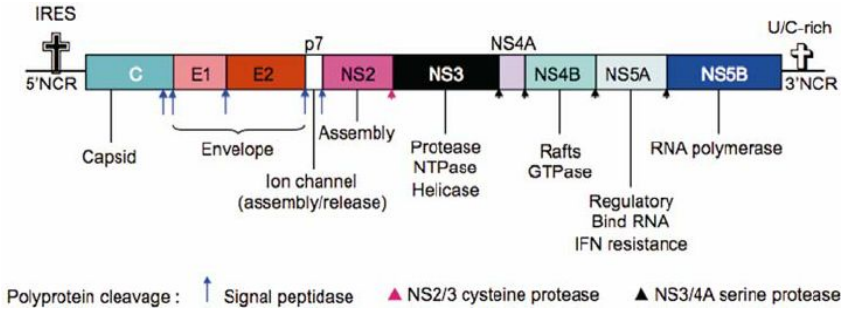
Fred Jaya - Stage 2

28/10/19


@fredjaya1


github.com/fredjaya

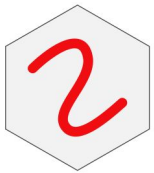
Hepatitis C virus (HCV)



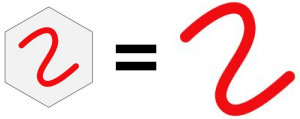
- ~71 million people with chronic infection
- Liver cirrhosis, carcinoma
- Single-stranded positive-sense RNA (+ssRNA)
- ~9000 bp ORF
- 7 genotypes, many subtypes

- Mutation + recombination → diversity
- Host expansion, drug and immune evasion
- **High mutation rates** → RdRp
- **Infrequent recombination**
- NGS has shown circulating recombinants

Genotyping HCV

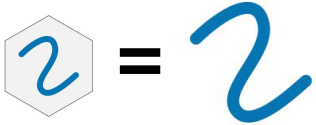


Hybridisation
assay



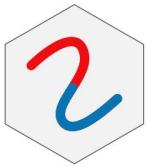
Sequencing

Phylogenetics

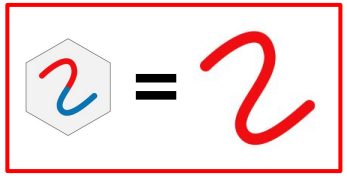


Virus sample → Genotyping → Treatment

Genotyping **Recombinant** HCV



Hybridisation
assay



Sequencing

Phylogenetics

Virus sample → Genotyping → Treatment

Effect of recombination

Hybridisation assays

Commercial kits, VERSANT

Unable to identify recombinants

Phylogenetics

Obscures evolutionary relationships

Treatments

Effectiveness of treatments are type specific

Solution...

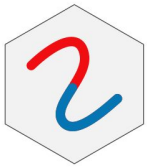
Next-Generation Sequencing (NGS)

Deep sequencing, WGS
Higher-resolution for subtyping
Surveillance of genotypes

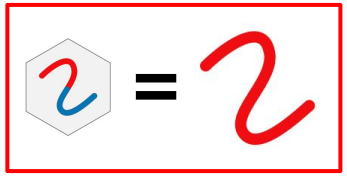
Bioinformatics

Recombination detection methods (RDMs)

Genotyping **Recombinant** HCV



Hybridisation assay



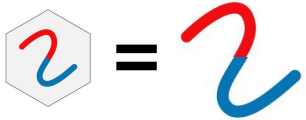
Sequencing

Phylogenetics

Virus sample → Genotyping → Treatment

Assessment of recombination detection methods using viral simulations

Hybridisation assay



Recombination detection

Sequencing

Phylogenetics + bioinformatics

Virus sample → Genotyping → Treatment

RDM benchmarks

INVITED TECHNICAL REVIEW

Analysing recombination in nucleotide sequences

DARREN P. MARTIN,* PHILIPPE LEMEY† and DAVID POSADA‡

Detecting recombination in evolving nucleotide sequences

Cheong Xin Chan, Robert G Beiko and Mark A Ragan*

Evaluation of methods for detecting recombination from DNA sequences: Computer simulations

David Posada* and Keith A. Crandall

A Simulation Study of the Reliability of Recombination Detection Methods

Carsten Wiuf, Thomas Christensen,† and Jotun Hein‡*

Sequence length

Breakpoint accuracy

Sequence diversity

Limitations

- RDMs were not designed to process large (NGS) datasets
- No clear outline which RDMs are most effective for which datasets
  - Simulations are simplified

Research questions

1) Identify methods that are capable of being scaled to process NGS data
2) Explore the effect of evolutionary and sequence properties on RDM behaviour
    a) Can we suggest effective RDMs based on viral properties i.e. mutation, recombination rate?

Next-generation sequencing (NGS)

Sanger ABI 3730xl	PacBio Sequel II HiFi
600 - 1000 bases	~ 15,000 bases (high accuracy)
96 reads	~ 4,000,000 reads (SMRT Cell 8M)

Methodology

1) Viral simulations and parametric sweep
2) Recombination detection method analyses
3) Pipeline design

Simulations

Mutation rate	Recombination rate	Dual infection probability	Sample size
10^{-8}	10^{-8}	0.00	**100**
10^{-7}	**10^{-7}**	**0.05**	**1000**
10^{-6}	**10^{-6}**	0.10	**2500**
10^{-5}	10^{-5}	0.25	**5000**
10^{-4}	10^{-4}	0.50	**10000***
10^{-3}	10^{-3}	1.00	

Datasets were generated between all individual parameters

Simulation 1 - **bold**
Simulation 2 - all; * except n = 10000

SANTA-SIM - viral simulator

Pipeline design

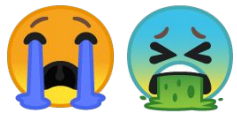
864 unique parameter combinations

→ 3 replicates

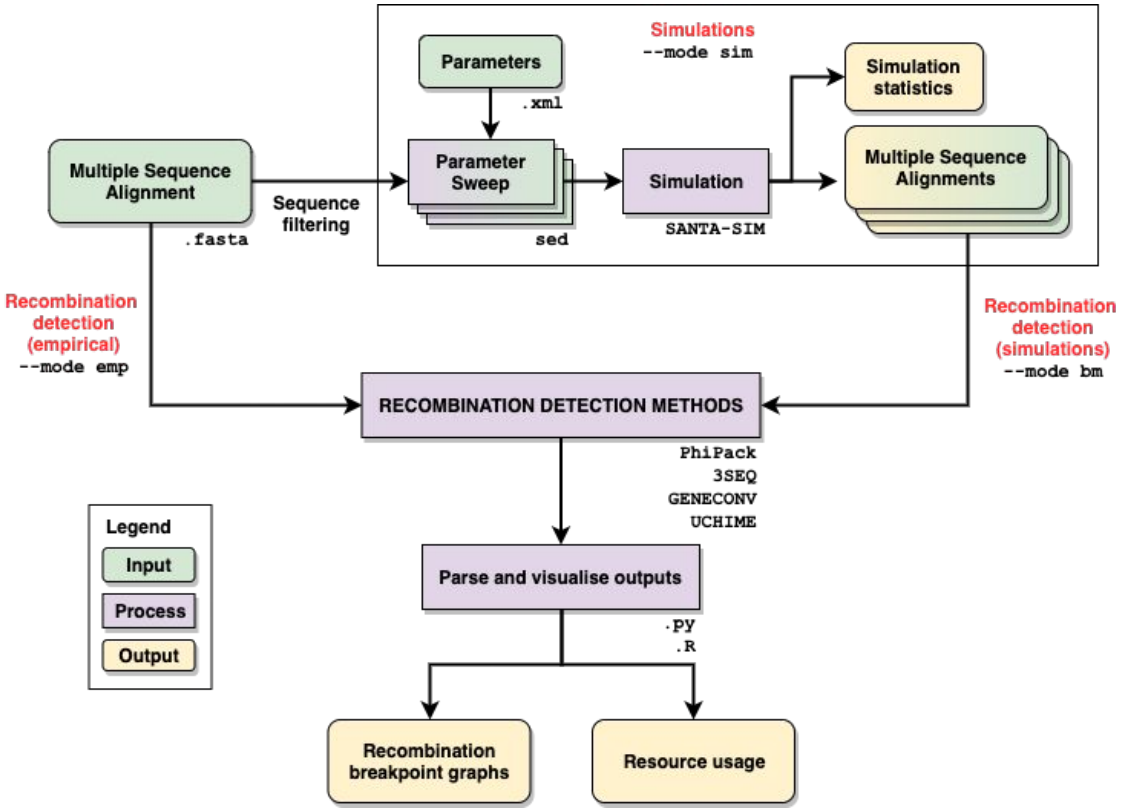
→ 2592 total files to simulate

→ 3 RDMs

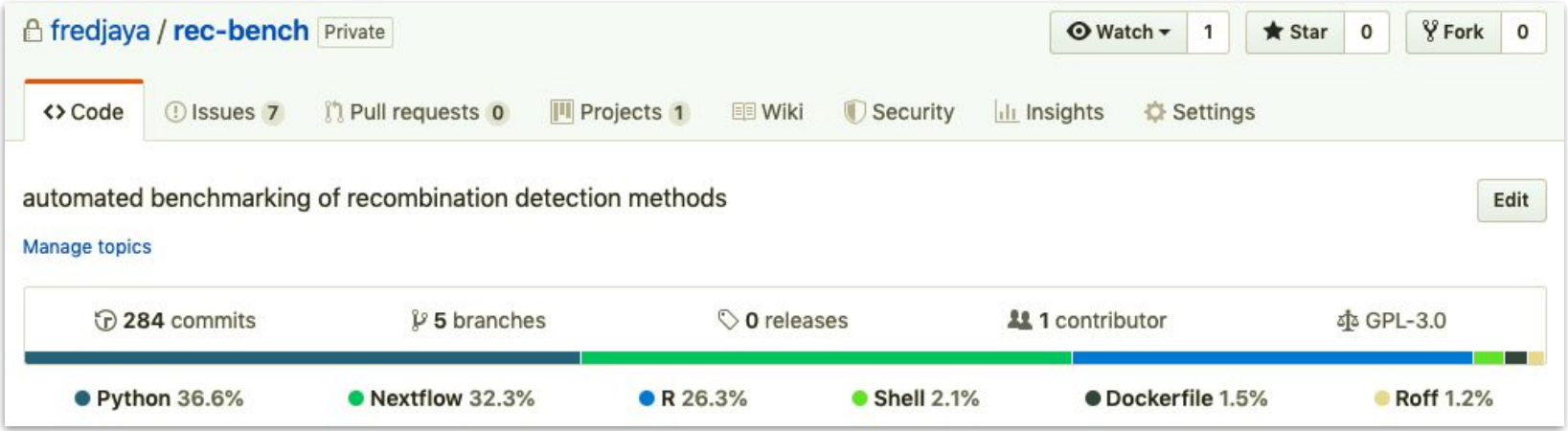
→ 7776 analyses

→ 

Pipeline design



rec-bench



github.com/fredjaya/rec-bench

PhiPack contribution to bioconda

Results (initial)

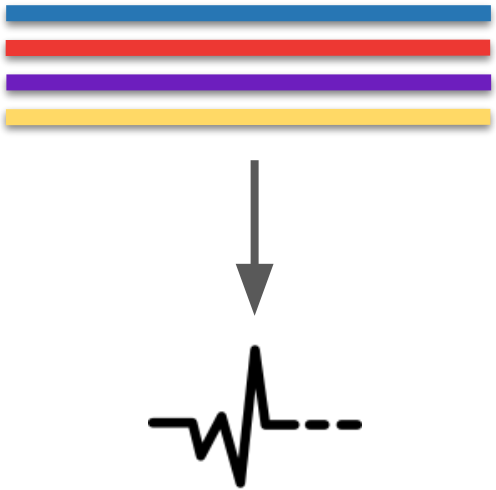
Simulation stats

Breakpoint significance

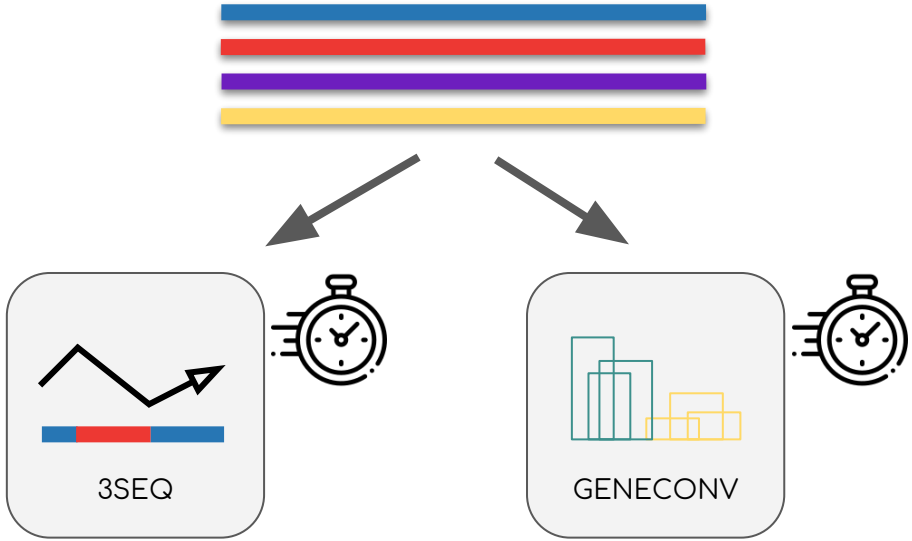
Runtime (wall)

Recombination detection methods

Key findings

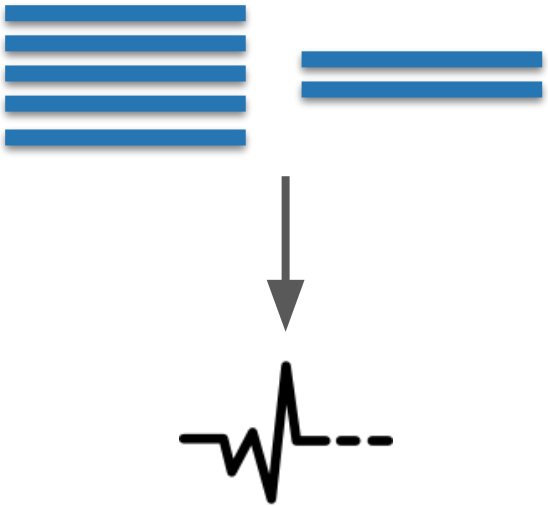


1. Diversity determines recombination detection

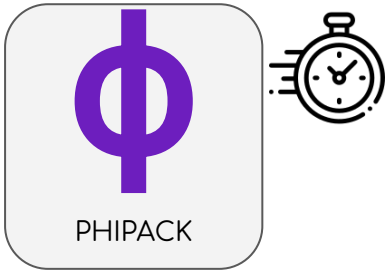


2. Diversity affects runtime in 3SEQ and GENECONV

Key findings



3. Sample size affects recombination detection

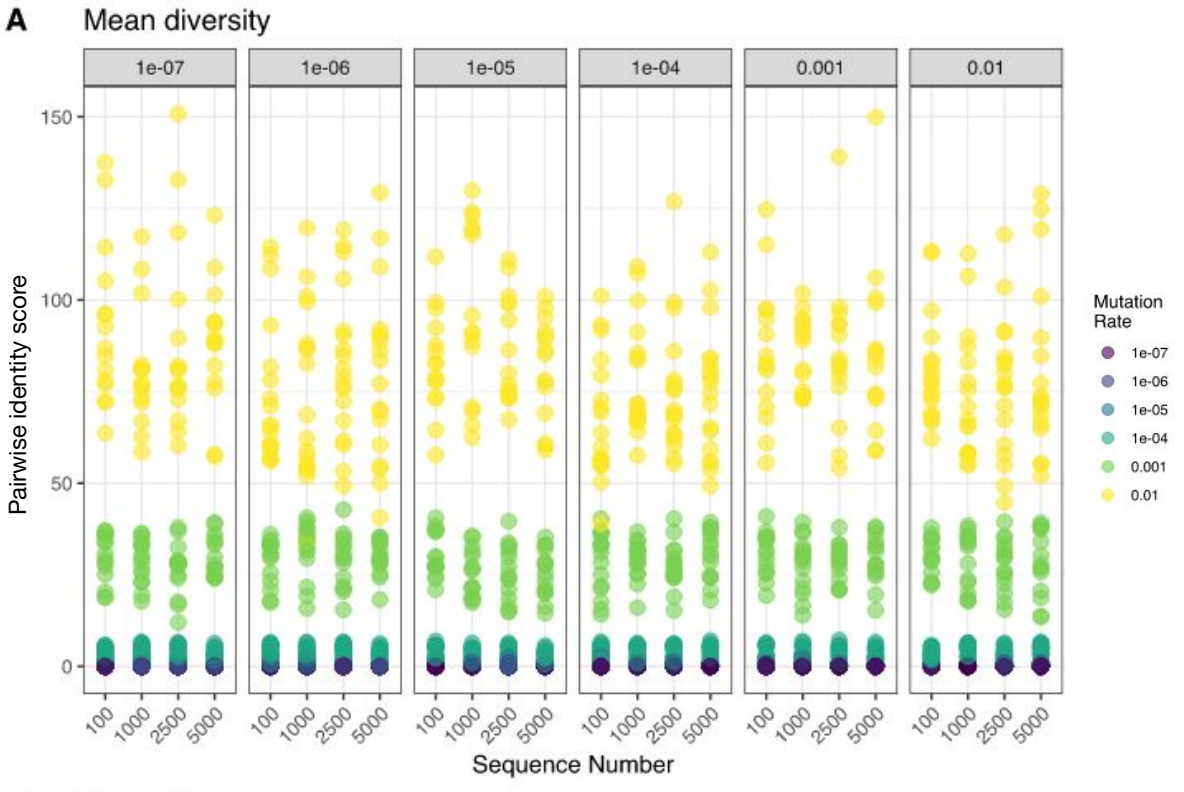


4. PhiPack is scalable

Results

Simulations

Mutation rate is correlated with output sequence diversity



Recombination detection

Significance of detected signal is determined by algorithm/method

Pairwise Homoplasy Index

Observed number of states between two sites

No sites removed

Hypergeometric random walk

Triplets compared

No sites removed

‘BLAST-like’

Pairwise comparison

Monomorphic sites removed

Recombination detection

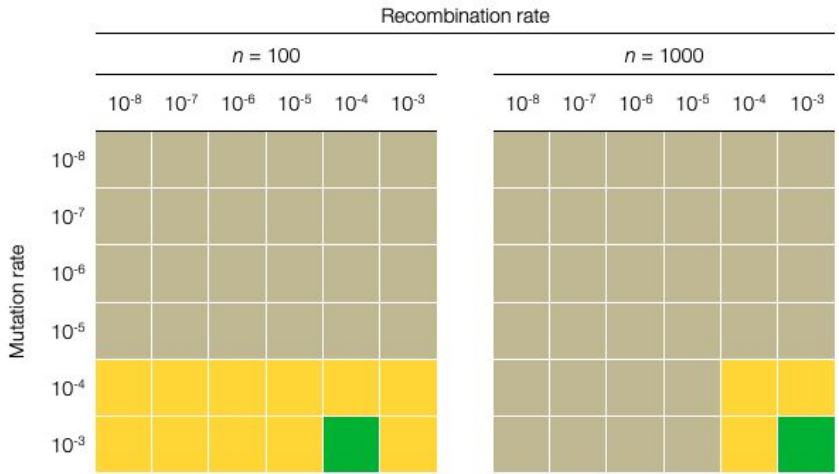
No recombination detected
Some recombination detected
Frequent recombination detected

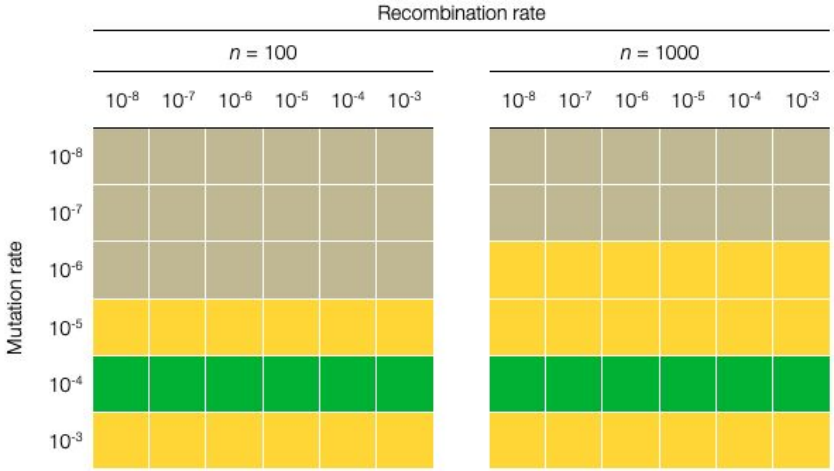
3SEQ

- Monomorphic sites = uninformative
- Sensitive to high divergence
- Sensitive to recombination with *n*?

GENECONV

- Monomorphic sites = uninformative
- Very sensitive to a specific range
- *n* = affects diversity

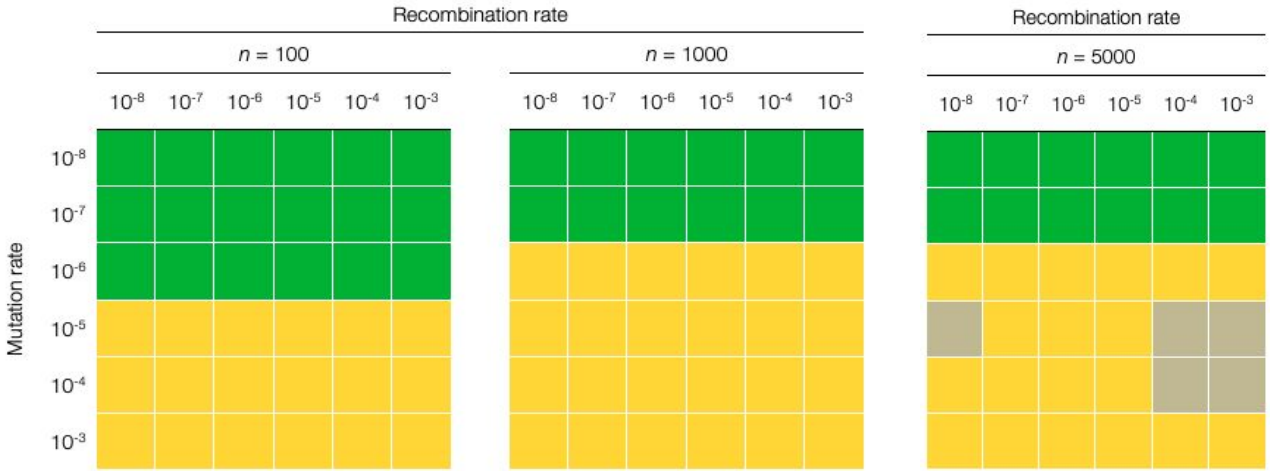




Recombination detection

PHIPACK

- Very sensitive
- Powerful at low divergence
- Increasing *n* reduces false positives?



Recombination detection

Significance of detected signal is determined by algorithm/method

Pairwise Homoplasy Index

Observed number of states between two sites

No sites removed

Hypergeometric random walk

Triplets compared

No sites removed

‘BLAST-like’

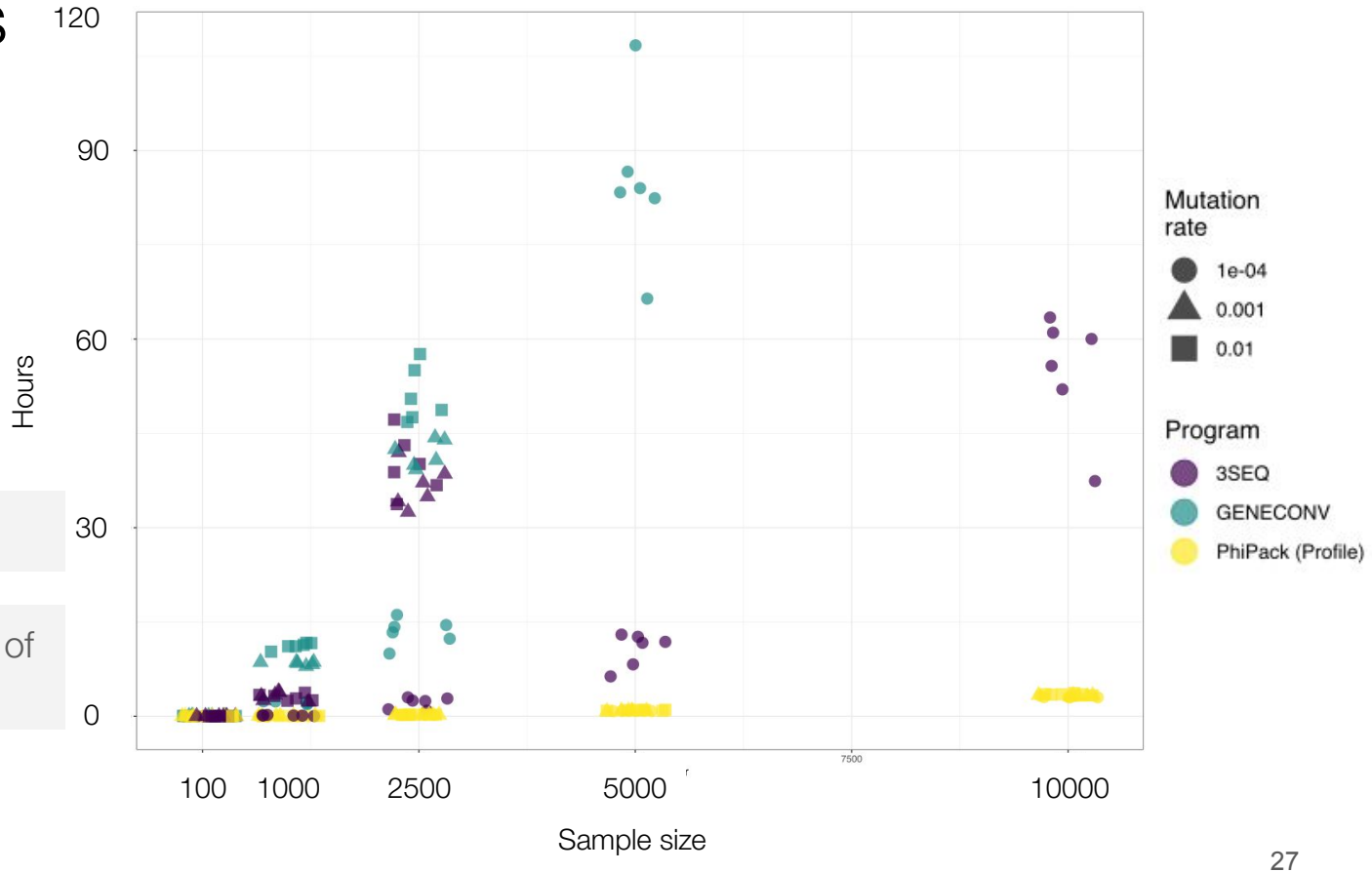
Pairwise comparison

Monomorphic sites removed

Runtimes

PhiPack is scalable

Mutation rate affects speed of 3SEQ and GENECONV



Runtimes

Wall time is determined by detection algorithm/method

Pairwise Homoplasy Index

Observed number of states between two sites

No sites removed

Hypergeometric random walk

Triplets compared

No sites removed

'BLAST-like'

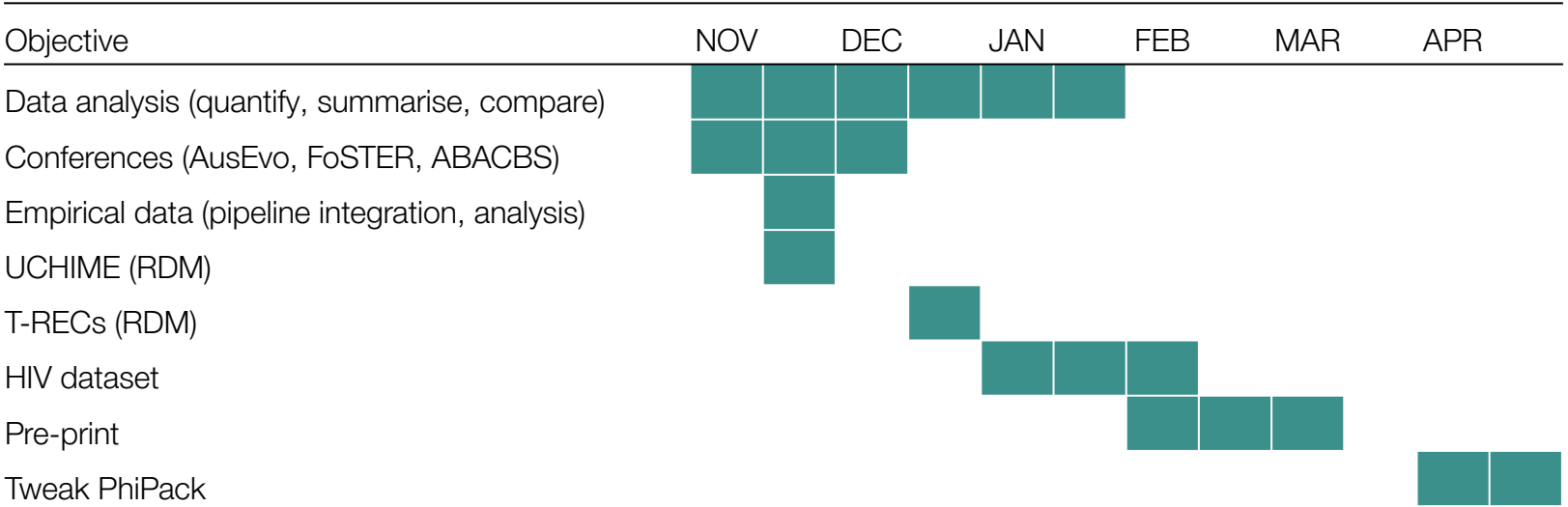
Pairwise comparison

Monomorphic sites removed

Limitations

- Lack of detail
  - Presence/absence or positive/negative
  - Summarise data for comparison
- Unreliable reporting of simulated breakpoints
  - SANTA-SIM mod: cumulative breakpoints across generations
  - Doesn’t account for false positives
  - No measure of accuracy
- Doesn’t reflect empirical data
  - Generation of unfit sequences/populations
  - Mutation = diversity, what’s the effect of recombination on sims?
  - Sequence diversity underrepresented
- Low replicates

Further work



Acknowledgements

Darling Lab

- Prof. Aaron Darling
- Dr. Barbara Brito
- Dr. Matt DeMaere
- Dr. Kay Anantanawat
- Dr. Mathieu Fourment
- Dr. Leigh Monahan
- Dr. Mike Imelfort
- Dr. Liza Kretzschmar
- Kevin Ying
- Daniela Gaio
- Nehleh Kargarfard
- Brent Bevear
- Sidaswar Krishnan
- Christian Cabato

the ithree institute

UTS eResearch

SHDRSA

Binfies

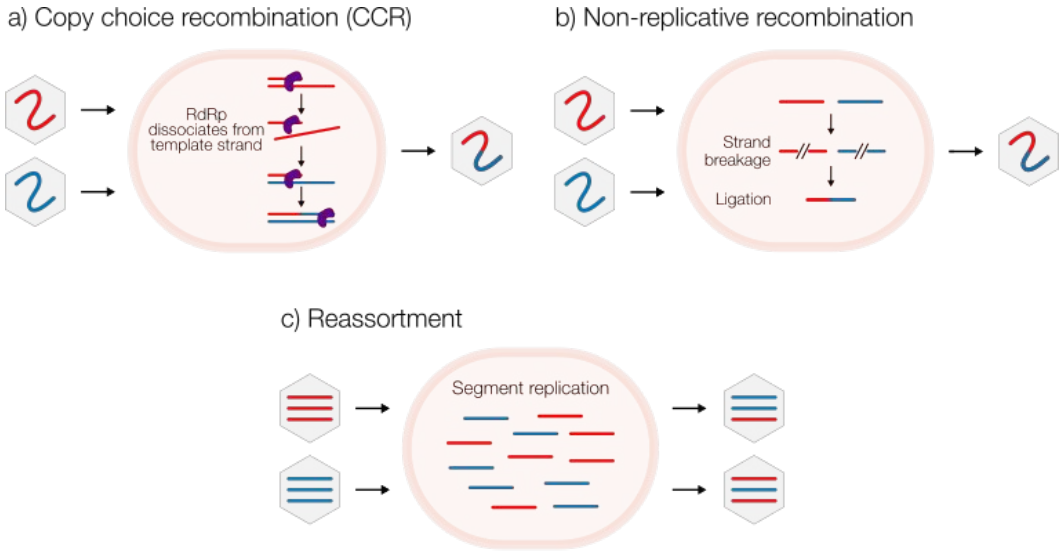
Level 7 HDRs

@fredjaya1

github.com/fredjaya

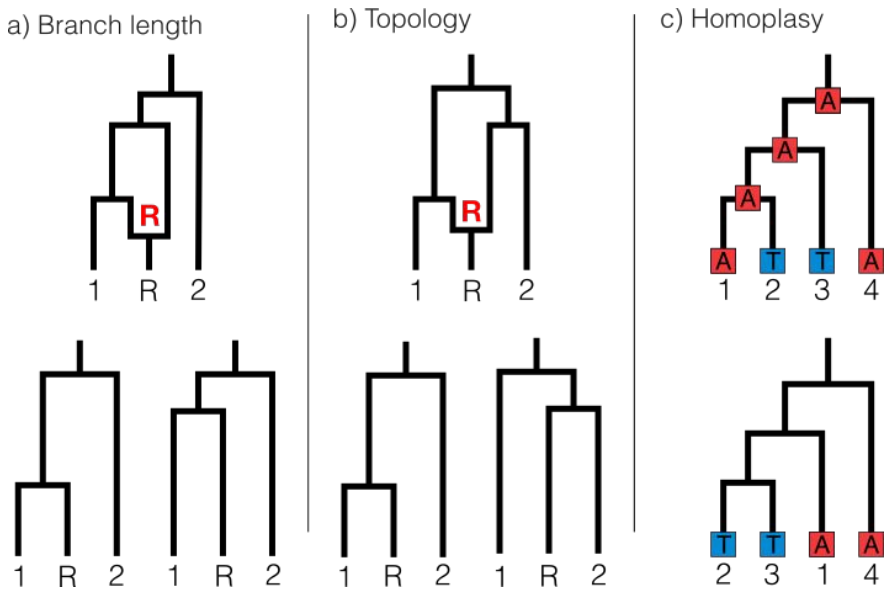
Recombination (viral)

- Generation of a child genome with two parental genomes
- Co-infection of a host cell
- Affects evolution - changes viral phenotype and fitness



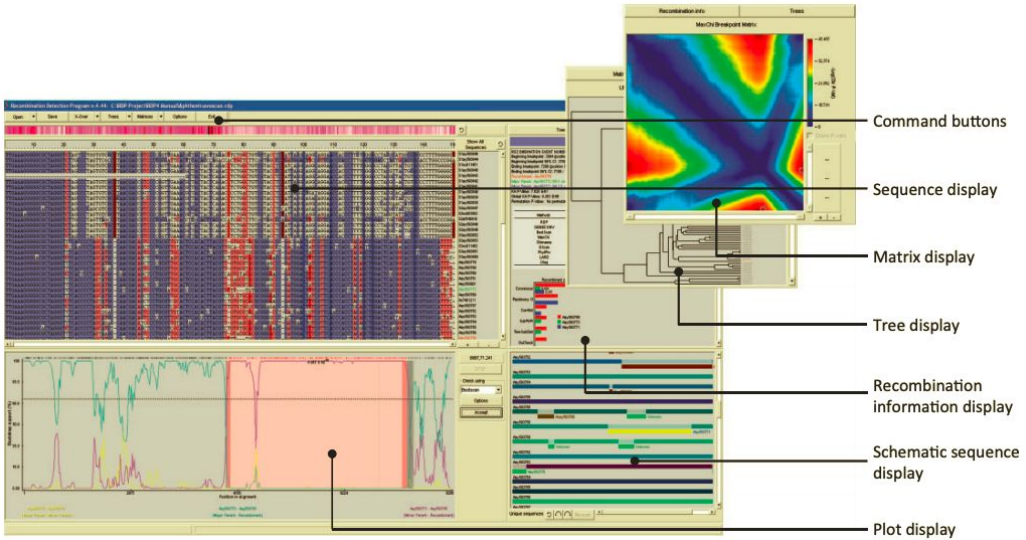
Confounds phylogeny

- Recombinants have two ancestral histories
- Violates assumptions → inaccurate estimations



Recombination detection methods

- Test and account for recombination prior to analyses (sequences)
- Different statistical tests and algorithms
- Performance differs based on sequence properties
  - Recombination rate
  - Sequence diversity
  - Sample size

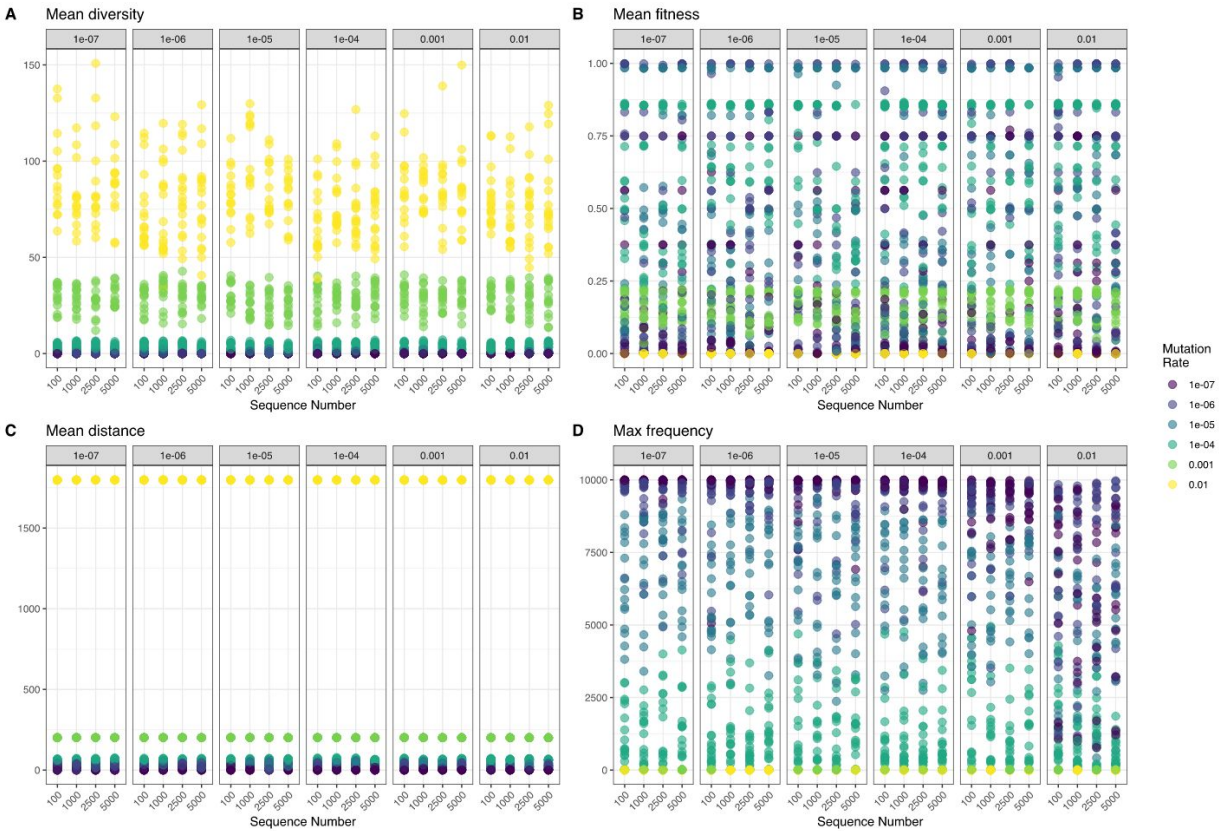


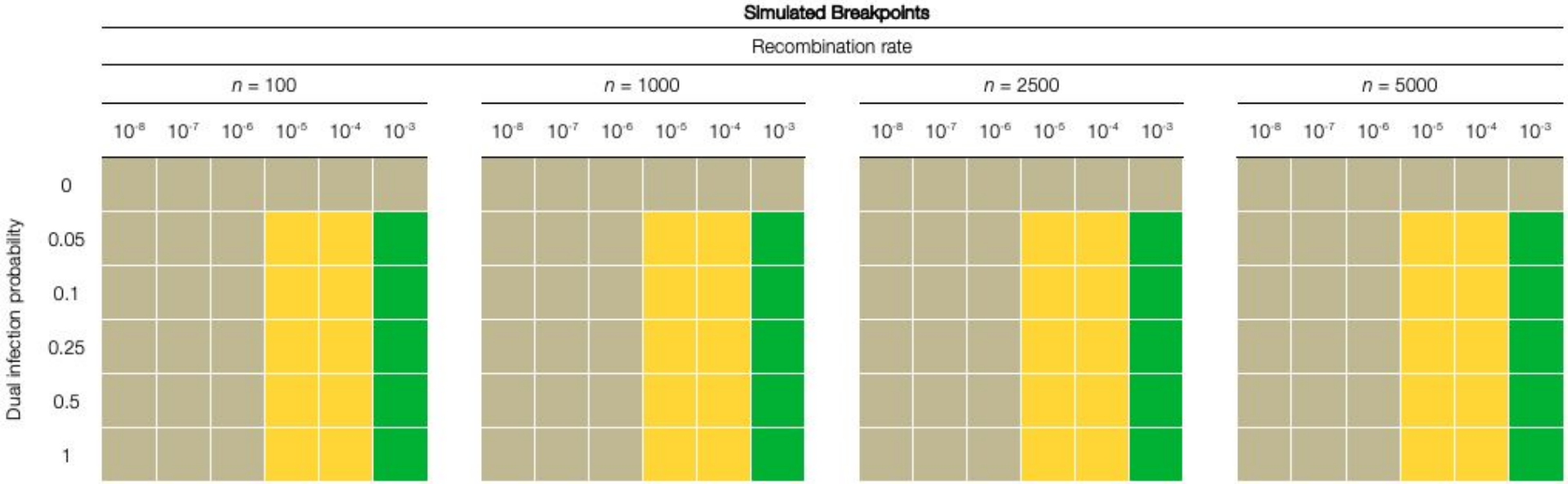
Simulations

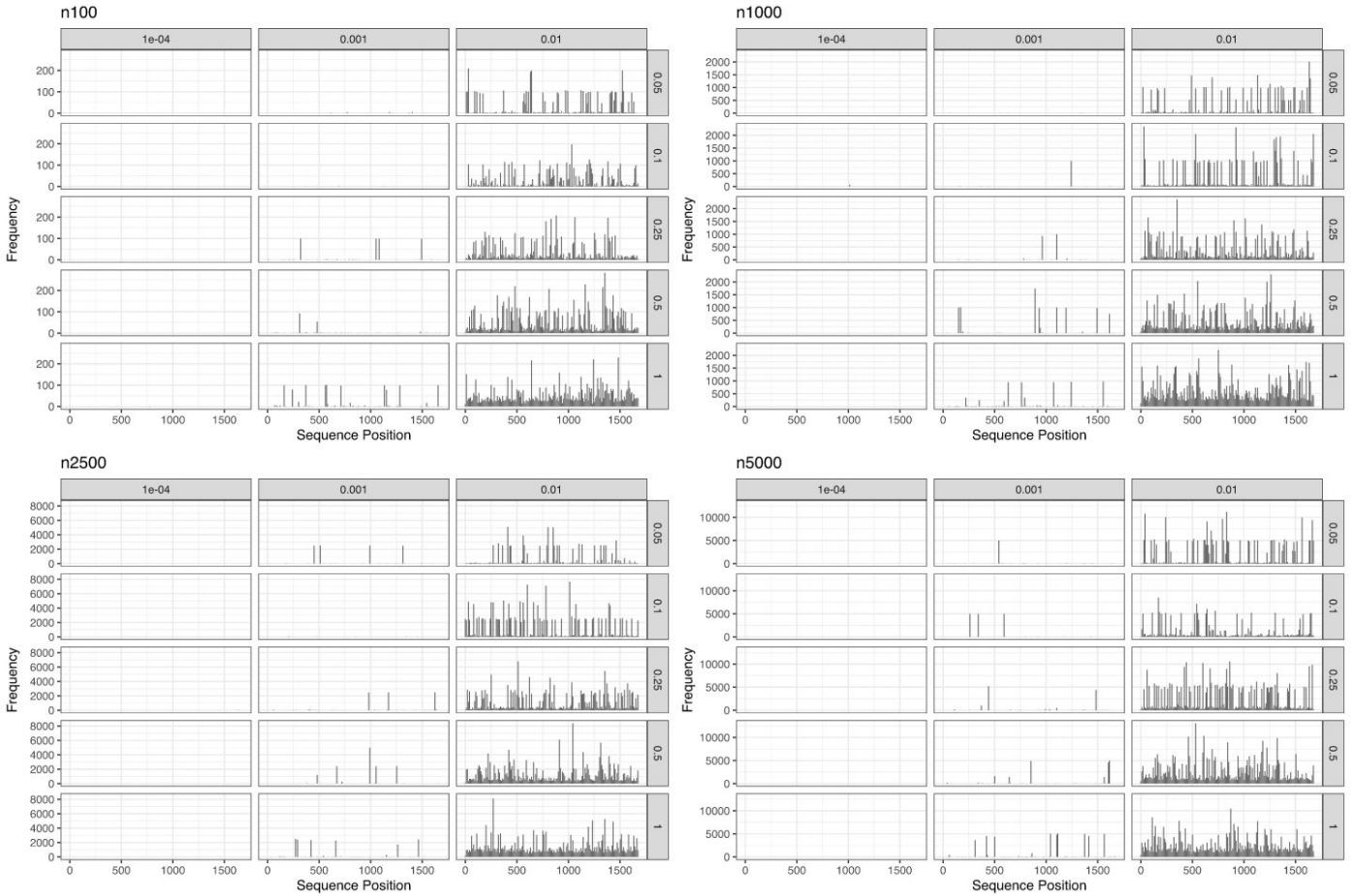
	sim_1			
n	PhiPack (Profile)	3SEQ	GENECONV	UCHIME
100	**Y**	**Y**	**Y**	N
1000	**Y**	**Y**	**Y**	N
2500	**Y**	**Y**	**Y**	N
5000	**Y**	**Y**	**Y**	N
10000	**Y**	**Y**	**Y**	N

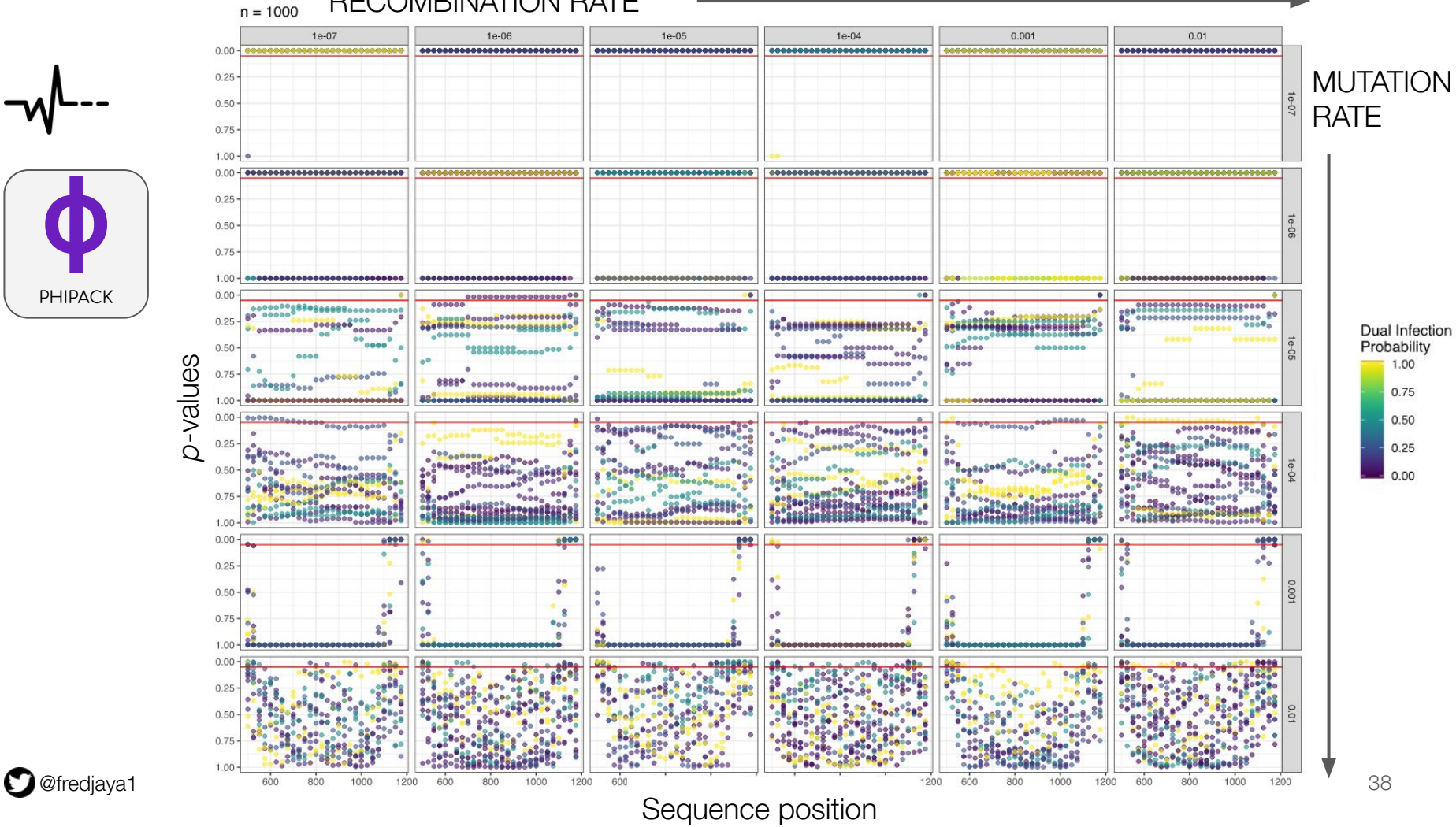
	sim_2			
n	PhiPack (Profile)	3SEQ	GENECONV	UCHIME
100	**Y**	**Y**	**Y**	**Y**
1000	**Y**	**Y**	**Y**	**Y**
2500	N	N	N	N
5000	**Y**	N	N	**Y**
10000	N	N	N	N

Y = Analysed
N = Not analysed




Simulated Breakpoints
Recombination rate
n = 100
n = 1000
n = 2500
n = 5000
10-8 10-7 10-6 10-5 10-4 10-3
Dual infection probability
0
0.05
0.1
0.25
0.5
1


n100
n1000
n2500
n5000
1e-04
0.001
0.01
0.05
0.1
0.25
0.5
1
Frequency
Sequence Position



PHIPACK

@fredjaya1

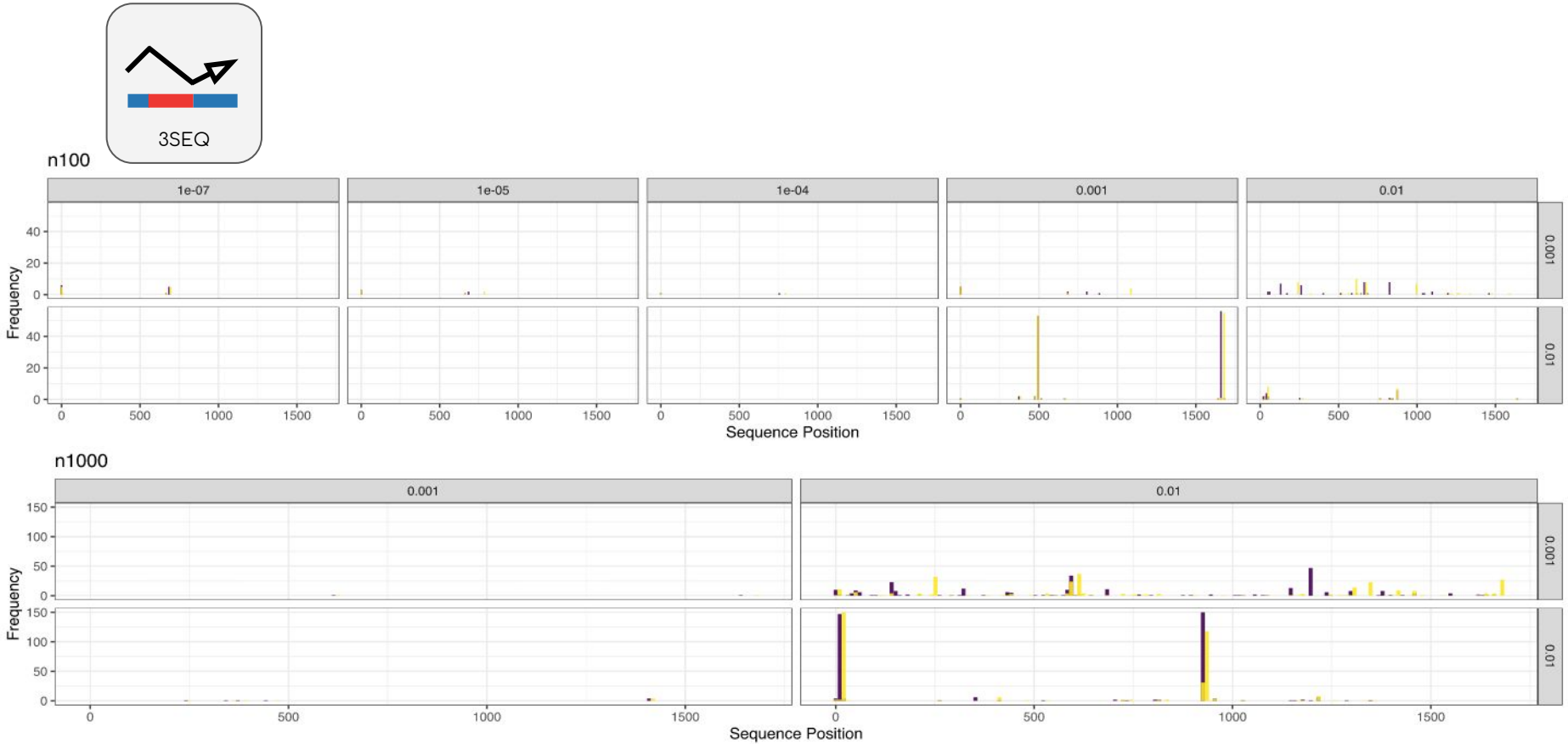


Figure 7. Frequency distribution of significant, paired breakpoint locations (start - purple, end - yellow) detected by 3SEQ from *sim*$_2$. Horizontal facets denote recombination rate and vertical facets denote mutation rate.

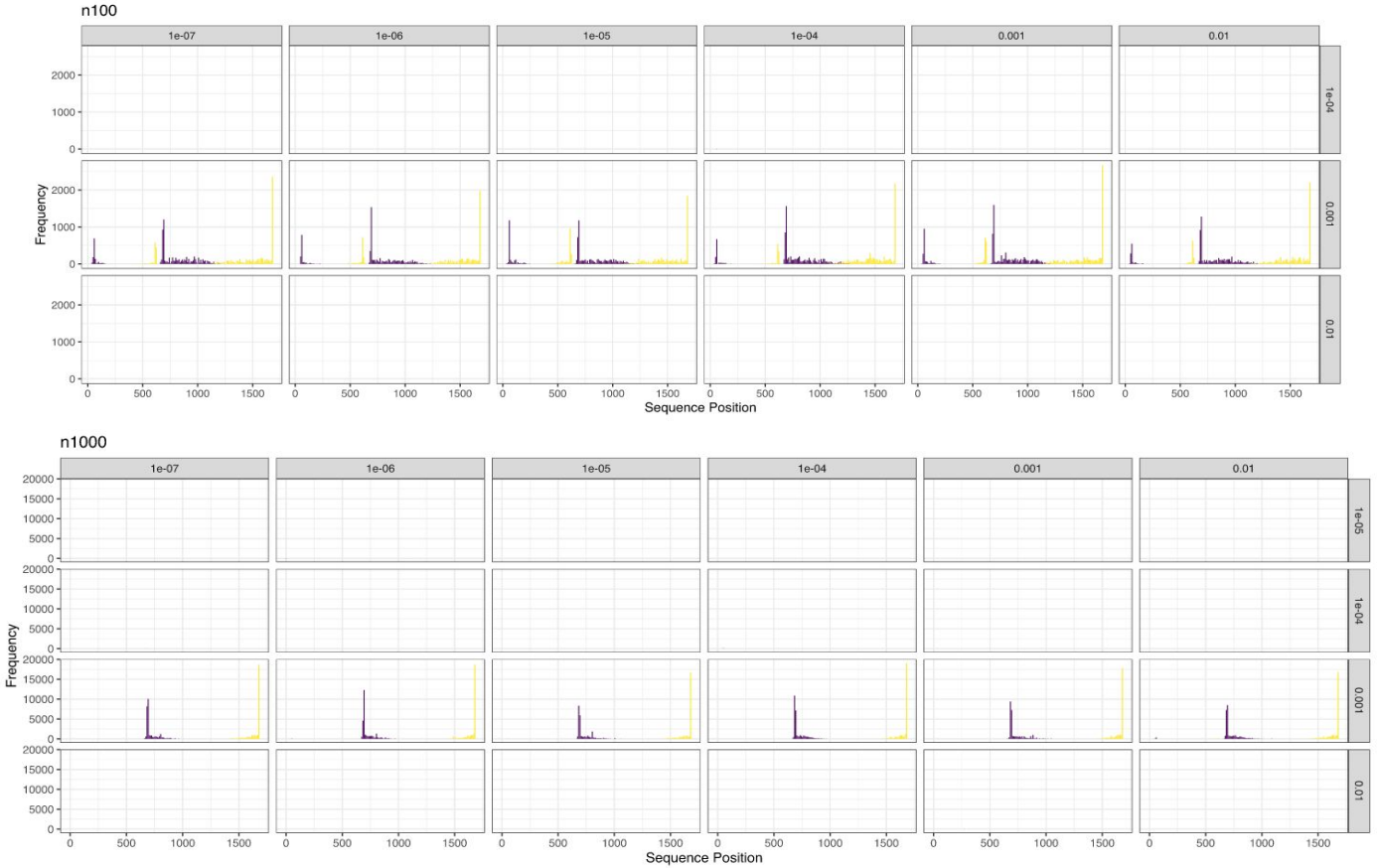


Figure 6. Frequency distribution of significant, paired breakpoint locations (start - purple, end - yellow) detected by GENECONV from *sim*$_2$. Horizontal facets denote recombination rate and vertical facets denote mutation rate.