Recombination detection of malaria *var* genes

Qian Feng[1], Kathryn Tiedje[2], Shazia Ruybal[2], Gerry Tonkin-Hill[2], Michael Duffy[2], Karen Day[2], Heejung Shim[1], Yao-ban Chan[1]

[1] Melbourne Integrative Genomics, School of Mathematics and Statistics / School of Bioscience, The University of Melbourne

[2] Department of Microbiology and Immunology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne

November 22, 2019

Malaria Parasite

- Malaria is a serious, sometimes fatal, disease that is caused by a parasitic infection of the red blood cells.

- 2018 World Malaria Report:
  - 219 million malaria cases globally in 2017
  - 435,000 malaria-related deaths in 2017
  - Most cases occur in Africa (93%)

- *Plasmodium falciparum* (the most dangerous parasite) has caused 300,000 deaths each year and 200 million clinical cases.

Malaria Parasite

- Malaria is a serious, sometimes fatal, disease that is caused by a parasitic infection of the red blood cells.

- 2018 World Malaria Report:
  - 219 million malaria cases globally in 2017
  - $435,000$ malaria-related deaths in 2017
  - Most cases occur in Africa (93%)

- *Plasmodium falciparum* (the most dangerous parasite) has caused $300,000$ deaths each year and 200 million clinical cases.

Malaria Parasite

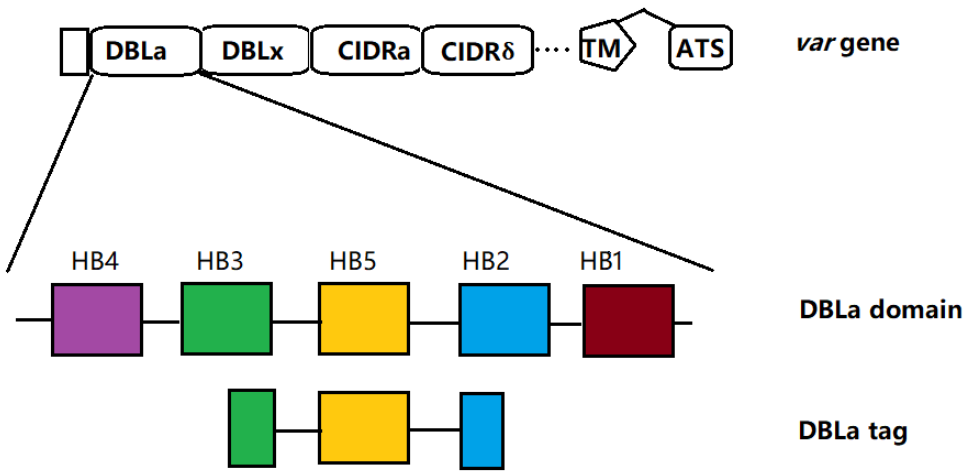
- Malaria is a serious, sometimes fatal, disease that is caused by a parasitic infection of the red blood cells.

- 2018 World Malaria Report:
  - 219 million malaria cases globally in 2017
  - 435,000 malaria-related deaths in 2017
  - Most cases occur in Africa (93%)

- *Plasmodium falciparum* (the most dangerous parasite) has caused 300,000 deaths each year and 200 million clinical cases.

PfEMP1 and *var* architecture

P. falciparum erythrocyte membrane protein 1 (PfEMP1) is the major antigen of malaria parasite *P. falciparum*, encoded by $50 \sim 60$ *var* genes per genome.



These genes are hyper−diverse, principally due to **recombination**.

The study of these *var* genes is thus one core problem in current malaria research, with implications for **future malaria interventions.**

Project aim

The evolution of *var* genes can be studied through the conserved DBLα tags.

We aim to uncover these tags' evolutionary histories by constructing a phylogeny.

- Phylogenetic tree
- **Phylogenetic network**

parent 1: REDTADDKKIHG

parent 2: WALLKNRPNTDP

recombinant: REDTANRPNTDP

What does the phylogenetic tree/network look like?

Project aim

The evolution of *var* genes can be studied through the conserved DBLα tags.

We aim to uncover these tags' evolutionary histories by constructing a phylogeny.

- Phylogenetic tree
- **Phylogenetic network**

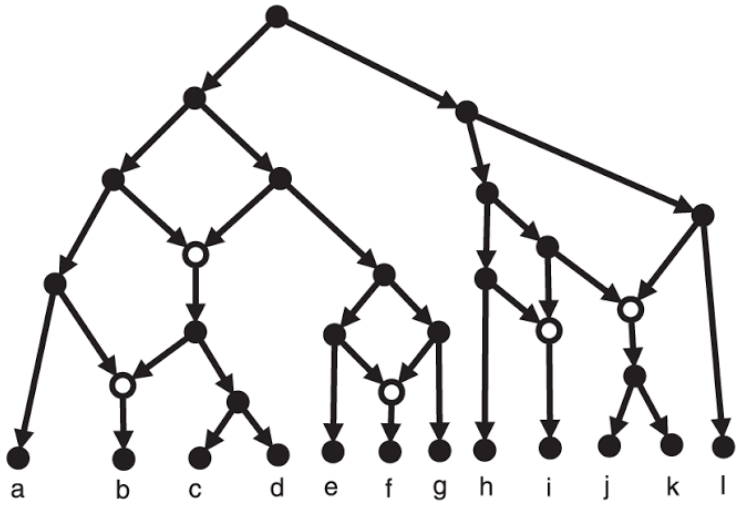
parent 1: REDTADDKKIHG

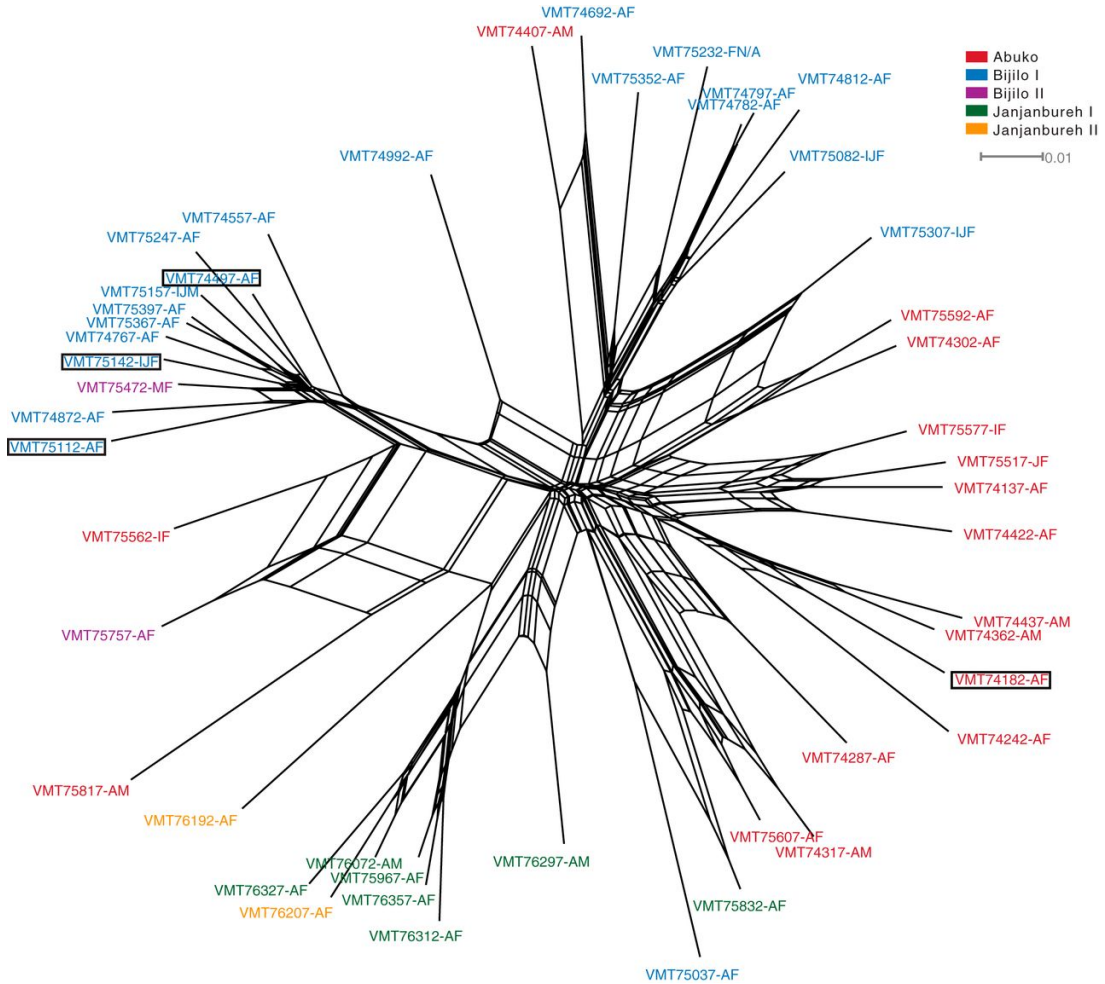
parent 2: WALLKNRPNTDP

recombinant: REDTANRPNTDP

What does the phylogenetic tree/network look like?

Recombination network





Split network from *env* gene sequences generated using the NeighborNet algorithm as implemented in SplitsTree.

Vol 93 Issue 22, Journal of Virology, 31 December 1969

Project aim

The evolution of *var* genes can be studied through the conserved DBLα tags.

We aim to uncover these tags' evolutionary histories by constructing a phylogeny.

- Phylogenetic tree
- **Phylogenetic network**

parent 1: REDTADDKKIHG

parent 2: WALLKNRPNTDP

recombinant: REDTANRPNTDP

What does the phylogenetic tree/network look like?

In order to solve this problem, we should start to finish

Recombination Identification

- ✓ Which sequence is recombined one?
- ✓ Where is the potential breakpoint?

Project aim

The evolution of *var* genes can be studied through the conserved DBLα tags.

We aim to uncover these tags' evolutionary histories by constructing a phylogeny.

- Phylogenetic tree
- **Phylogenetic network**

parent 1: REDTADDKKIHG

parent 2: WALLKNRPNTDP

recombinant: REDTANRPNTDP

What does the phylogenetic tree/network look like?

In order to solve this problem, we should start to finish

Recombination Identification

- ✓ Which sequence is recombined one?
- ✓ Where is the potential breakpoint?

Literature Review

Unfortunately, none of them is appropriate solution for our problem.

We have to solve the following three obstacles:

- large number of sequences 🕒
- no multiple sequence alignment ☹
- no reference genome sequences ✍

☺ Fortunately, we finally work this problem out by a novel algorithm.

Literature Review

Unfortunately, none of them is appropriate solution for our problem.

We have to solve the following three obstacles:

- large number of sequences 🕒
- no multiple sequence alignment ☹
- no reference genome sequences ✍

☺ Fortunately, we finally work this problem out by a novel algorithm.

JHMM. Zilversmit et al,2013

T A G T C K D I M M M F

D_1 A G T C

D_2 K D I M

D_3 M - F

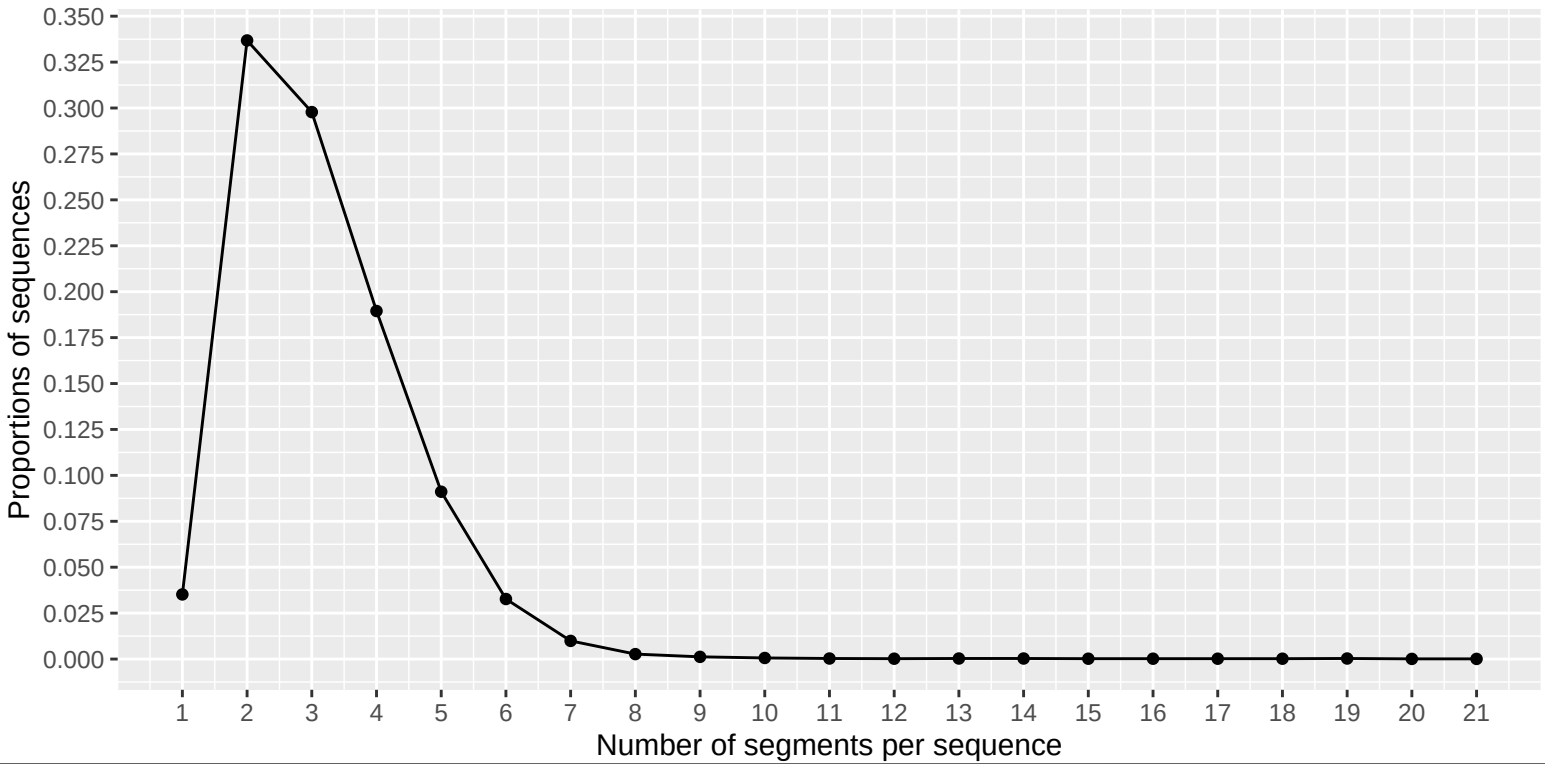
three parents

T A G T C K D I M M

D_1 A G T C

D_2 K D I M M

two parents



JHMM. Zilversmit et al,2013

T A G T C K D I M M M F

D_1 A G T C

D_2 K D I M

D_3 M - F

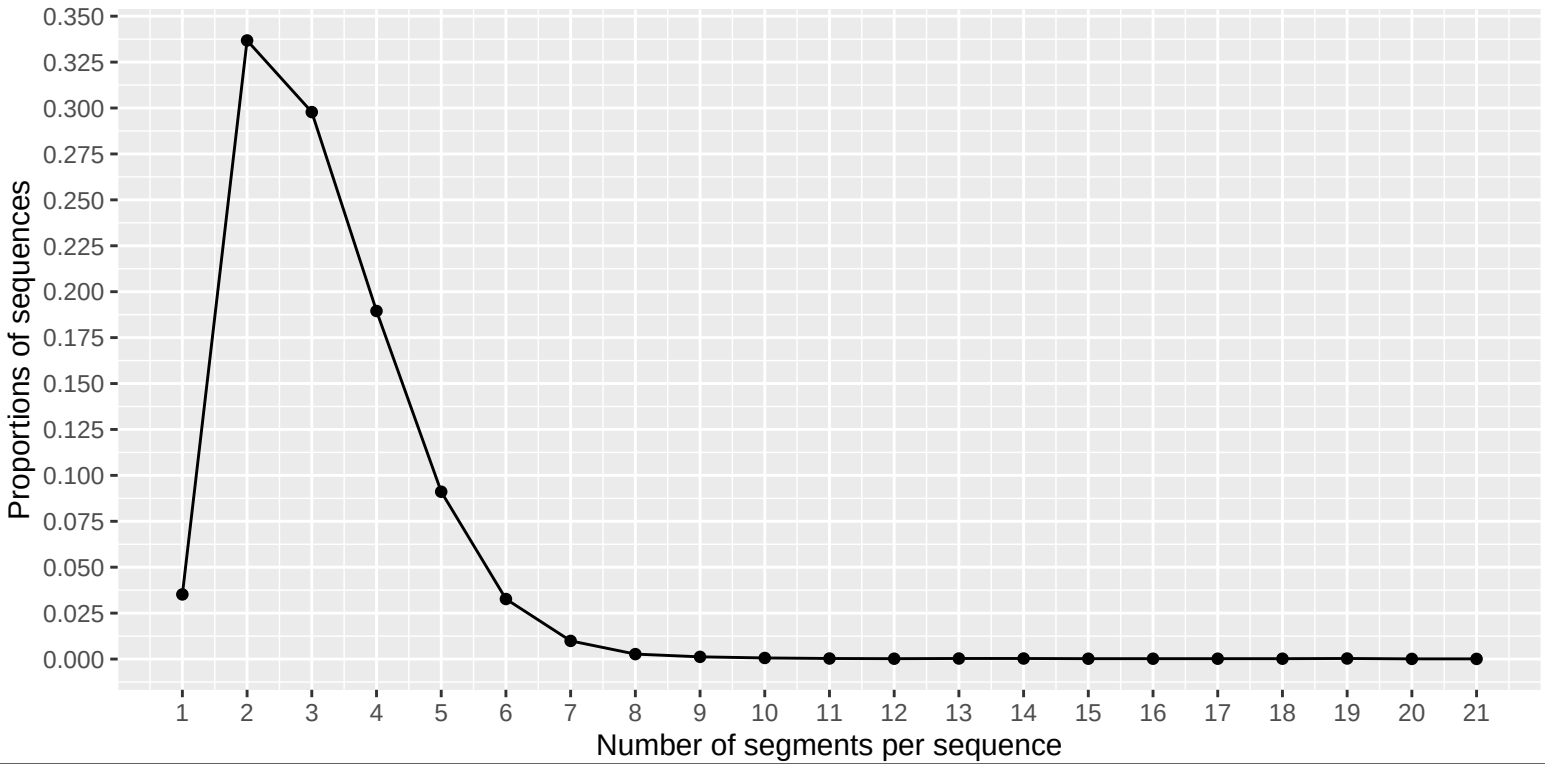
three parents

T A G T C K D I M M

D_1 A G T C

D_2 K D I M M

two parents



JHMM. Zilversmit et al,2013

T A G T C K D I M M M F

D_1 A G T C

D_2 K D I M

D_3 M - F

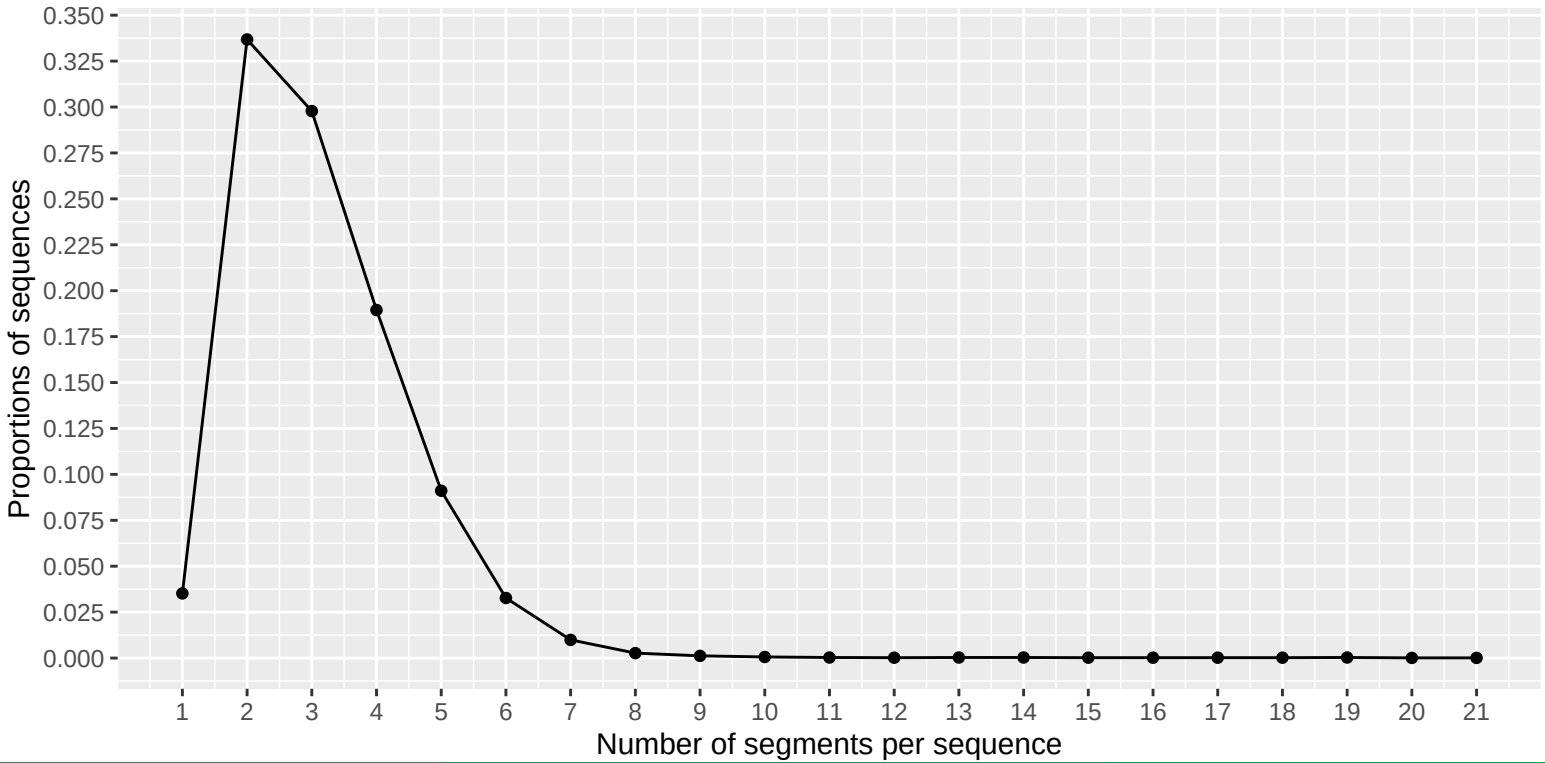
three parents

T A G T C K D I M M

D_1 A G T C

D_2 K D I M M

two parents

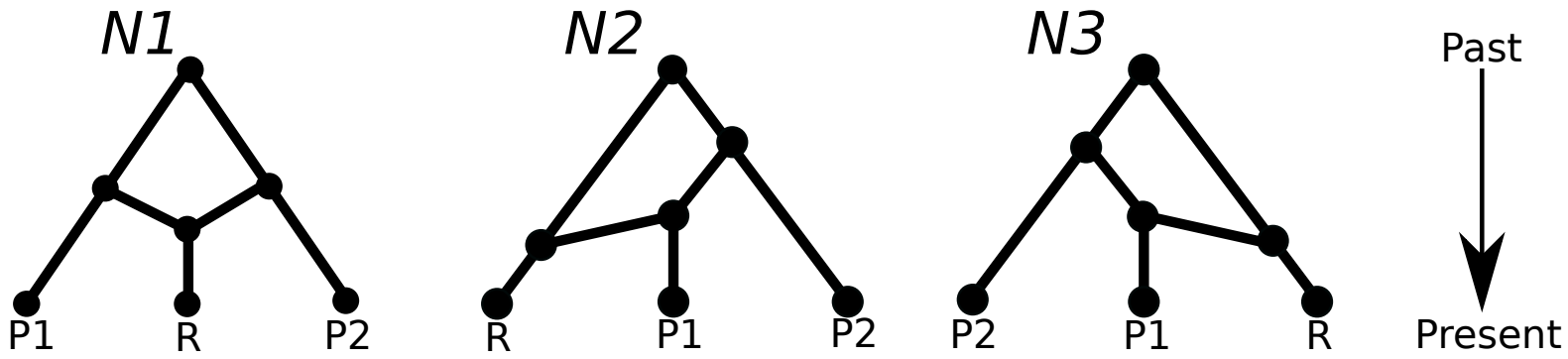


Potential Breakpoints are obtained by JHMM

```
#Input parameters = /vlsci/SG0011/qian-feng/MZmosaic/mosaic -ma -seq ../Protein_translateable_pilot_upper_centroids_run5.fasta -aa -tag ../Prote
target -target target -del 0.00806934718714 -eps 0.2283998284 -rec 0.015
#Created on Fri Aug  3 13:00:25 2018


Target: target_seq150   Length: 132     Llk: -100.954
target_seq150   DIGDIIRGKDLYLSYDKKEKEQRDKLEDNLKGVFAKIHDDVTSGKKKEEAEERYKGDTENYYQLREYWWALNRQDVWKAITCKAPDNAQYFRGTCGGGQNKTQNNCRCDKEKGAKDGDQVPTYFDYVPQ
                |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
db_seq8275      DIGDIIRGKDLYLSYDKKEKEQRDKLEDNLKGVFAKIHDDVTSGKKKEEAEERYKGDTENYYQLREYWWAL
db_seq6677                                                                                     NRQDVWKAITCKAPDNAQYFRGTC
db_seq13430                                                                                                            GGGQNKTQNNC
 db_seq773                                                                                                                        RCDKEKGAKDGDQVPTYFDYVPQ

Target: target_seq139   Length: 123     Llk: -66.833
target_seq139   DIGDIIRGKDLYLGNKKQNEKDREKERLQQNLKEIFKKIHGNLKDAQTHYNDNDENYYKLREDWWNANRQQVWKAMTCSAPDNAKYFRHTCGAGNDRSETKNNCQCISGDPPTYFDYVPQHLR
                |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
db_seq2432      DIGDIIRGKDLYLGNKKQNEKDREKERLQQNLKEIFKKIHGNLKDAQTHYNDNDENYYKLRED
db_seq9149                                                                     WWNANRQQVWKAMTCSAPDNAKYFRHTCGAGNDRSETKNNCQCISGDPPTYFDYVPQHLR

Target: target_seq138   Length: 136     Llk: -81.278
target_seq138   DIGDIIRGRDLYLGNRKKKRNGKETERDKLEQLKDIFKKIHEGLDHRIKSKYNGDTPYYYQLREDWWTANRHTVWKAITCSDDLKDNRYFRQTCSDTHGSSVAIHYCRCNDDKPDDDKPNTDPPTYFD
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| | ||||||||||
 db_seq778      DIGDIIRGRDLYLGNRKKKRNGKETERDKLEQLKDIFKKIHEGLDHRIKSKYNGDTPYYYQLR
db_seq15930                                                                    EDWWTANRHTVWKAITCSDDLKDNRYFRQTCSDTHGSSVAIHYCRCNDDKPDADNPNTDPPTYFD

Target: target_seq135   Length: 121     Llk: -76.681
target_seq135   DIGDIVRGRDLFHGNDEEKKQRKQLDDKLKDIFKNIKKENRDVNKLTNEKVREYWWYANRATIWKALTCDVKDNTYFRPTCNGKERTKGYCRCDDKTSGGKPGSNADQVPTYFDYVPQYLR
                |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||| |||||||||||||||||||||||||||||||||||||
db_seq3424      DIGDIVRGRDLFHGNDEEKKQRKQLDDKLKDIFKNIKKENRDVNKLTNEKVREYWWYANRATIWKALTCDVKDNTYFRATCNG
db_seq15159                                                                                        EERTKGYCRCDDKTSGGKPGSNADQVPTYFDYVPQYLR

Target: target_seq134   Length: 127     Llk: -91.465
target_seq134   DIGDIIRGKDLFIGYDERDRKEKQKIQDNLKDIFAKIHEELNGEAENHYNGDKQNNFYQLREDWWYANRETVWKAMTCSDDLNNSSYFRATCSDSADEKGPSVAKNKCTCNNGDVPTYFDYVPQFLR
                ||||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
db_seq1637      DIGDIIRGKDLFIGYDERDRKEKQKIQDNLKDIFAKIHEELNGEAQNHYNGDKQNNFYQLREDWWYANRETVWK
db_seq10996                                                                               AMTCSDDLNNSSYFRATCS
db_seq7467                                                                                                    DSADEKGPSVAKNKCTCNNGDVPTYFDYVPQFLR
```

Which one is true recombinant for two parents case?

Consider triple sequences each time and find the most probable recombinant sequence.



Our target is to find right one as accurately as possible and try to use the least time.

There is one key common in these three networks, P1 and P2 have very similar distance along sequences.

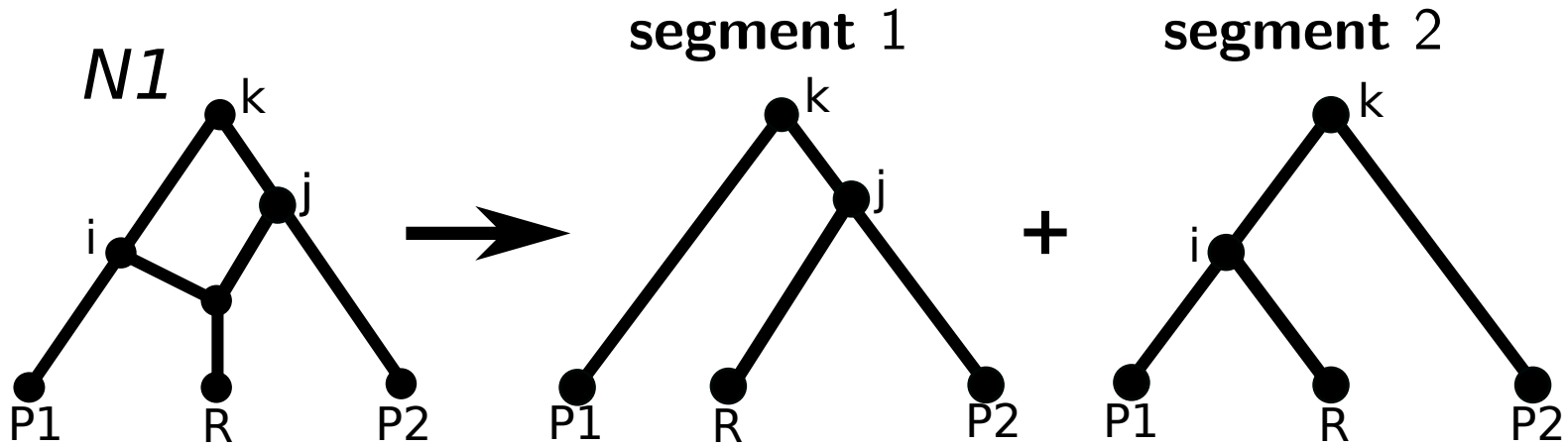
Key step in proposed algorithm

By computing the absolute value of segment distance differences, the smallest difference indicates two non-recombinant sequence.

$|D_1(P_1, P_2) - D_2(P_1, P_2)| = k - k = 0$; indicating R is recombinant.

$|D_1(R, P_2) - D_2(R, P_2)| = k - j$;

$|D_1(R, P_1) - D_2(R, P_1)| = k - i$;



Algorithm

Step 1: **Partial alignment results** are obtained using the jumping hidden Markov model (Zilversmit *et al.*)

Step 2: for triple in triple list:

if (segment length < 10): remove its closest triple(s).

else: **MAFFT** alignment is used to complement, forming one equal-length triple, go to step 3.

Step 3: Calculate all the pairwise segment distances in the left and right partitions.

Step 4: Compute the absolute value of segment distance differences, **the smallest difference infers two non-recombinant sequences**.

$$Rec := \{R, P_1, P_2\} \backslash arg \min_{P_1P_2, RP_1, RP_2} \{|d^{s_1}_{P_1P_2} - d^{s_2}_{P_1P_2}|, |d^{s_1}_{RP_1} - d^{s_2}_{RP_1}|, |d^{s_1}_{RP_2} - d^{s_2}_{RP_2}|\}$$

Step 5: **Bootstrap** the characters in each partition with replacement, repeat above two steps 100 times to get a statistical support value for inferred recombinant.

Simulation workflow

- Each specific setting replicates 100 times.
- Simulate one arbitrary tree without any recombination.
- Simulate the recombining sequences with randomly-distributed breakpoints.

snakemake

- Biologically realistic scenarios:
  - ☞ varying recombinant proportion
  - ☞ varying point mutation scale
  - ☞ varying indel events (indel rate and fragment size distribution)
  - ☞ changing protein sequence length
  - ☞ changing empirical amino-acid models

Simulation workflow

- Each specific setting replicates 100 times.
- Simulate one arbitrary tree without any recombination.
- Simulate the recombining sequences with randomly-distributed breakpoints.

snakemake

- Biologically realistic scenarios:
  - ☞ varying recombinant proportion
  - ☞ varying point mutation scale
  - ☞ varying indel events (indel rate and fragment size distribution)
  - ☞ changing protein sequence length
  - ☞ changing empirical amino-acid models

Accuracy measures

- In each replicate, we get the classification results.
- In each setting, summarize each evaluation metric result (mean and 95% confidence).

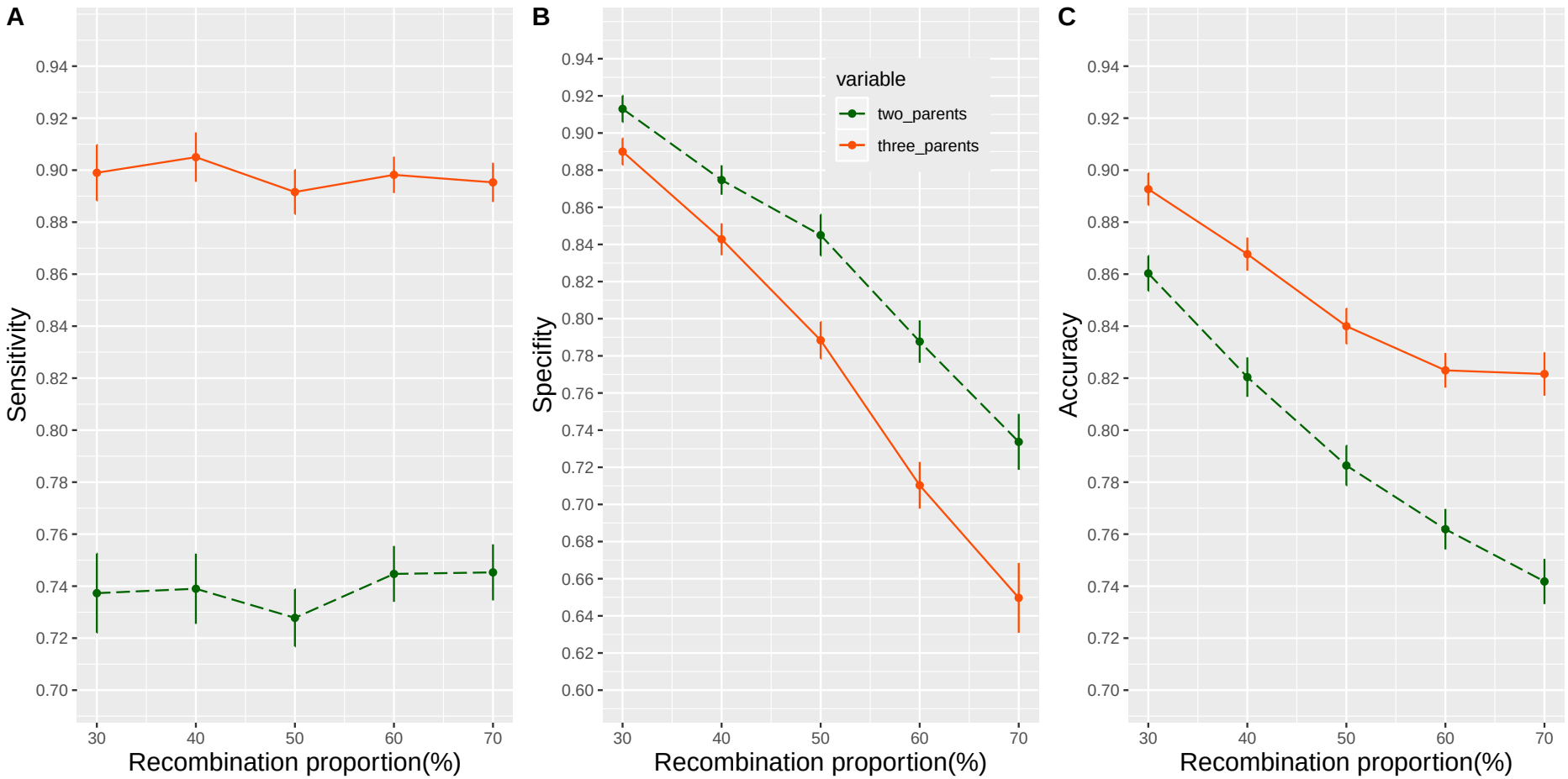
		Actual Condition		
	Total Samples	Actual Positive	Actual Negative	
Output of Classifier	Classify Positive	TP	FP	PPV (Precision)
	Classify Negative	FN	TN	
		TPR (Recall)	TNR (Specificity)	ACC F-measure MCC

Two cases are considered:

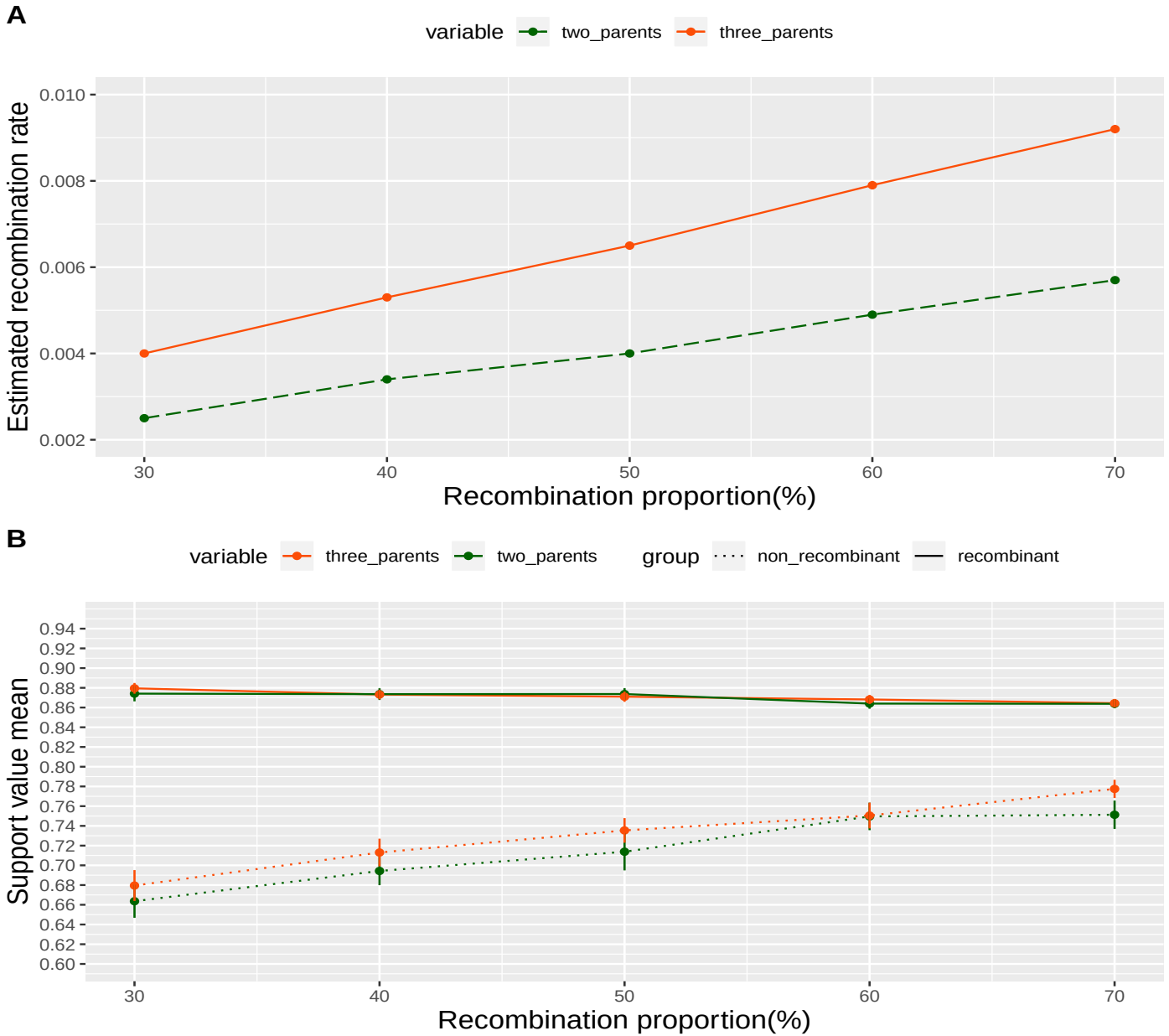
(1) Each recombinant has only two parents.

(2) Each recombinant has three parents.

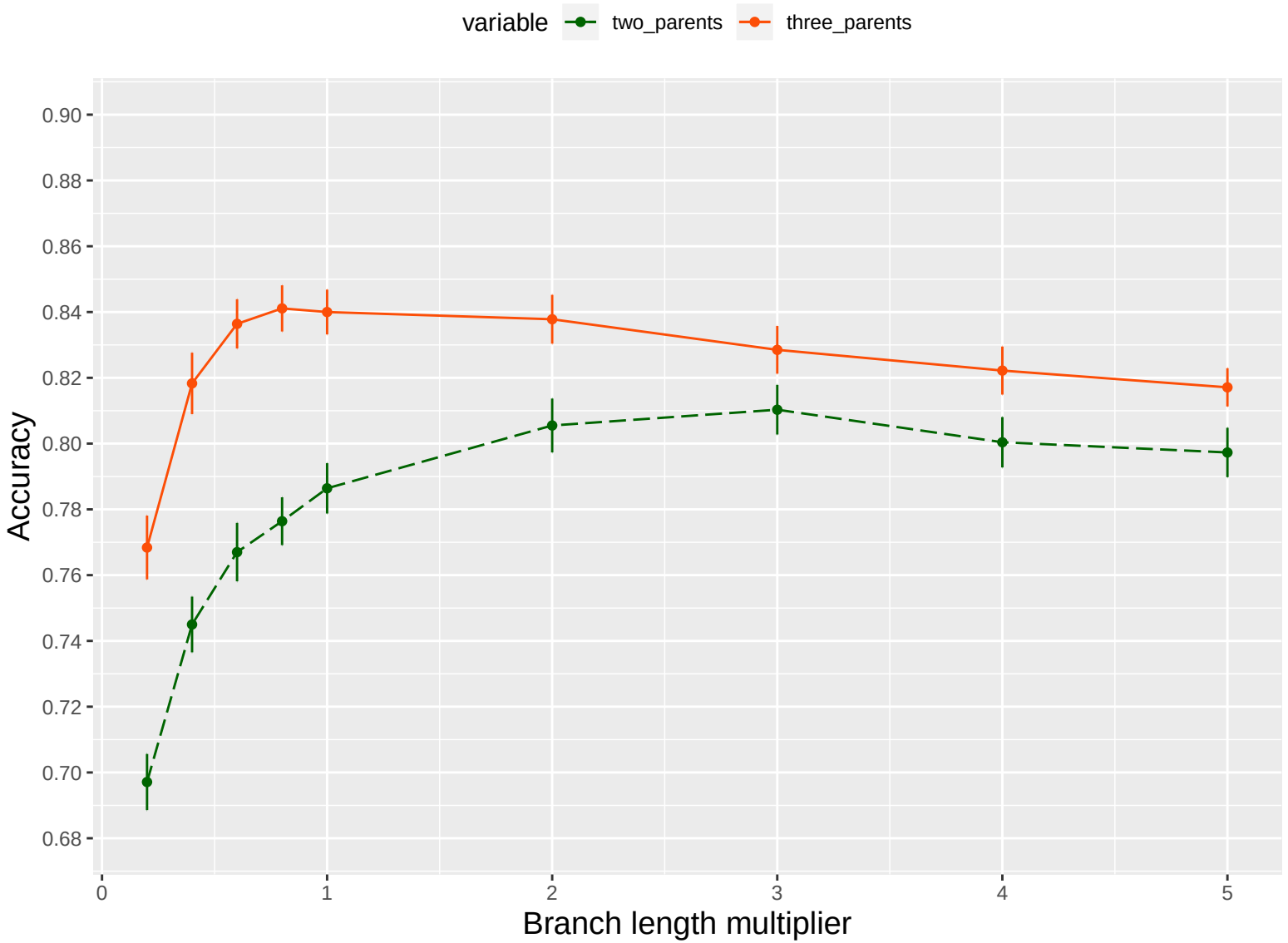
Simulation results when changing recombination proportion



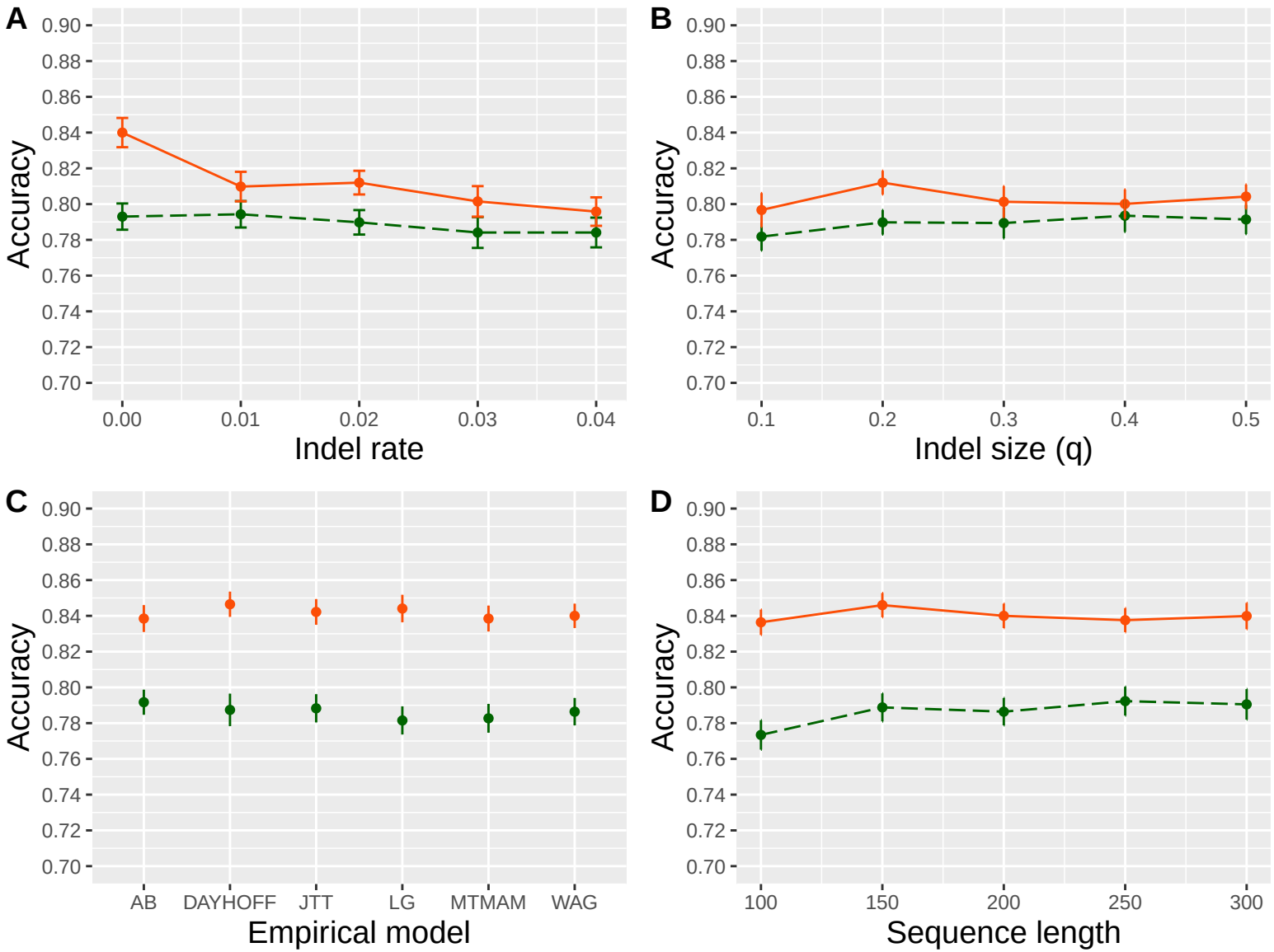
Simulation results when changing recombination proportion



Accuracy for mutation rate

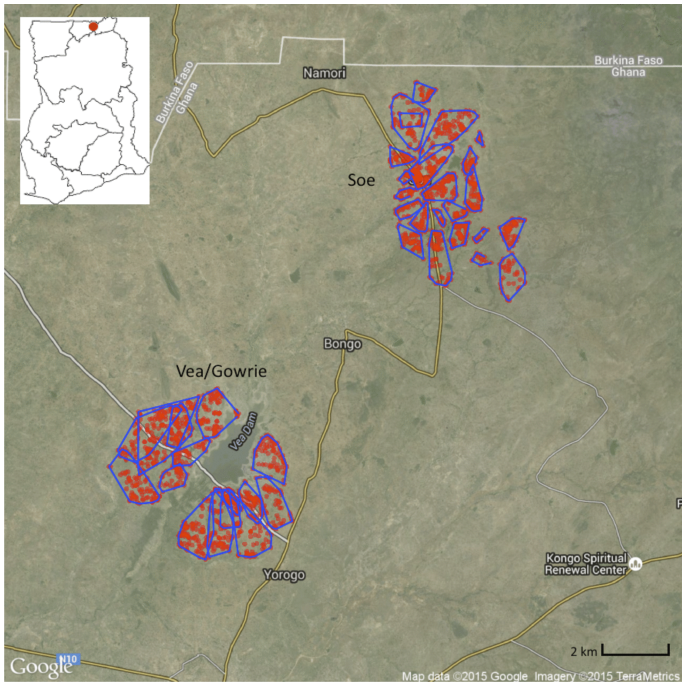


Accuracy for different settings



Application to a pilot study involving 161 isolates

- Two surveys were investigated in two catchment areas (Vea/Gowrie, Soe) in the Bongo District of north east Ghana (Tiedje *et al*, 2017).
- In this district, malaria was ranked as the **most threatening public disease**.



- 14801 out of 17335 (85.38%) representative protein sequences are identified recombinants.

Most positive results in real data application

- Recombinant happens more frequently not only in the same ups type group, but also in the same DBLα sub domains statistically!

	Same ups parents	Same ups family
A and non-A	0.989($0.850^{\star}$)	0.985($0.776^{\star}$)
A, B and C	0.655($0.509^{\star}$)	0.510($0.304^{\star}$)

Same domain parents	Same domain family
0.310($0.079^{\star}$)	0.206($0.010^{\star}$)

$\star$ refers to P value less than $2.2e-16$

- Non-recombinant DBLα types are significantly more likely to be observed in 10 or more isolates than recombinant DBLα types.

Most positive results in real data application

- Recombinant happens more frequently not only in the same ups type group, but also in the same DBLα sub domains statistically!

	Same ups parents	Same ups family
A and non-A	0.989($0.850^{\star}$)	0.985($0.776^{\star}$)
A, B and C	0.655($0.509^{\star}$)	0.510($0.304^{\star}$)

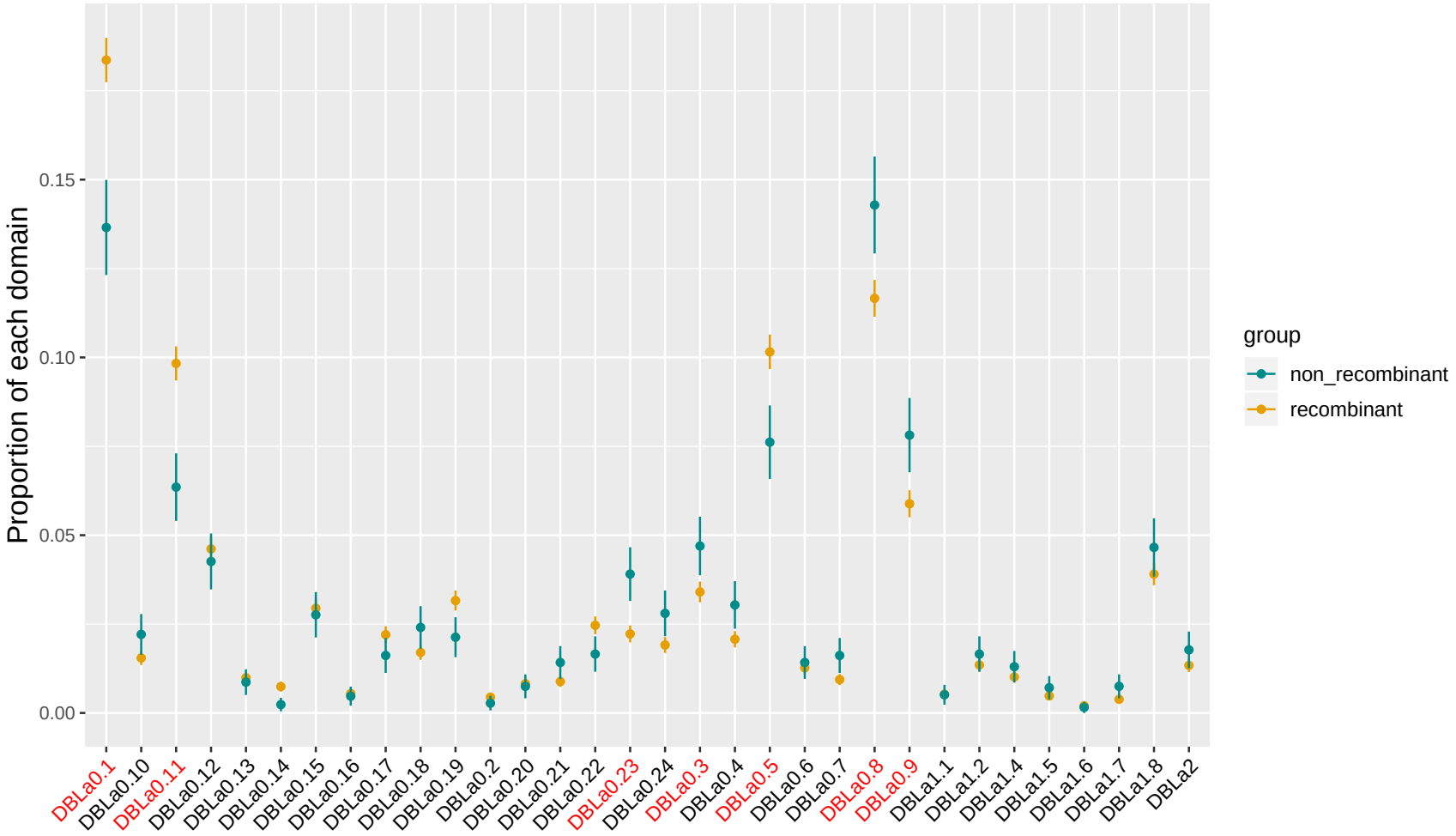
Same domain parents	Same domain family
0.310($0.079^{\star}$)	0.206($0.010^{\star}$)

$\star$ refers to P value less than $2.2e-16$

- Non-recombinant DBLα types are significantly more likely to be observed in 10 or more isolates than recombinant DBLα types.

Difference at domain level

- Some special domains are found to be different in terms of proportions between recombinant and non-recombinant groups.



Conclusion

Extension:

This novel algorithm is applicable not only in malaria, but also in RNA sequencing in cancer bioinformatics, in the context of detecting gene fusions.

Future work:

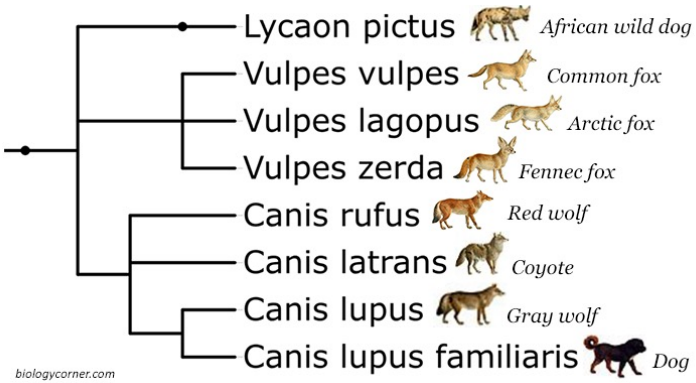
- **Construct phylogenetic networks for these DBLα sequences.**
- **Further application to real datasets.**
  - Explore the spatial and geographical features for the identified recombinants in bigger Ghana dataset, or even in global dataset.
- **Soft classification of semi-conserved upstream promoter sequences and explore its relationship with DBLα sequences.**

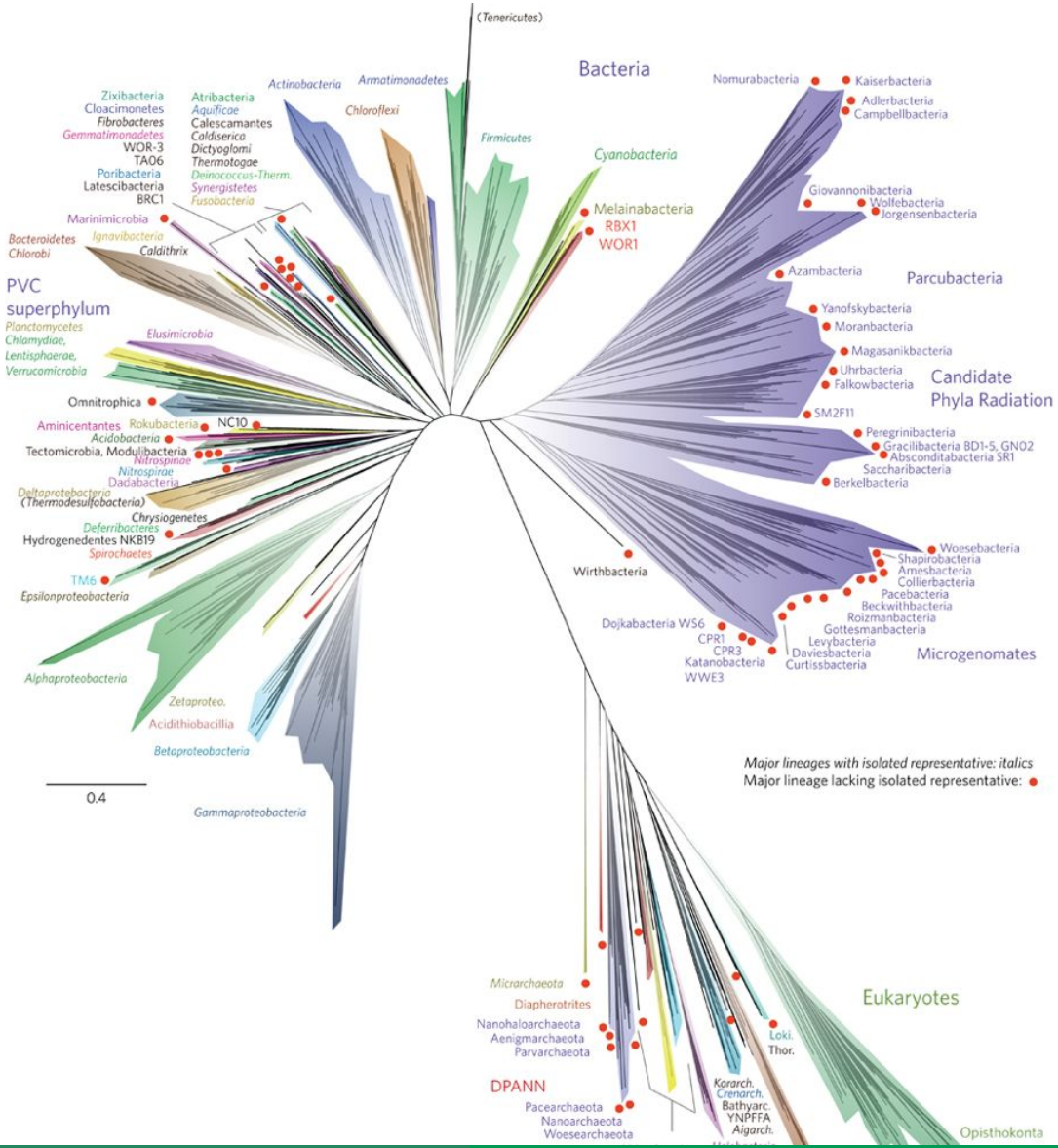
Acknowledgement

- **Dr. Yao-ban Chan**
- **Dr. Heejung Shim**
- Collaborators: Gerry Tonkin-Hill, Dr. Kathryn Tiedje, Prof. Karen Day
- Dr. Zitong Li, Dr. Qixin He
- Bobbie Shaban, Andrew Siebel and MIG students♡

Back up

phylogenetic tree

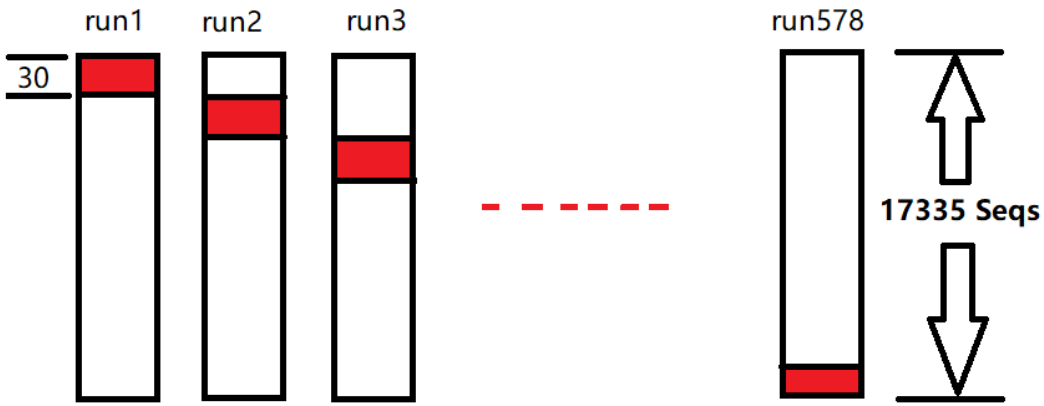




from understanding Evolution. 2019. University of California Museum of Paleontology. 4th November 2019

http://evolution.berkeley.edu.

Actual implementation



- 17335 protein sequences;
- 578 tasks for aligning each sequence per iteration;
- Each iteration runs on a high performing computing cluster (Helix);
- Each task which is assigned one core and 25 Gb memory runs approximately one hour.

Underlying assumptions

- Only one sequence is recombinant.
- Sequence lengths are all equal.
- The trees forming network are all ultrametric.
- Assume the breakpoint is given.

More result

Given the idea from Shazia in her Uganda paper, one of her comments is 'upsA DBLα types were significantly more likely to be observed in 10 or more isolates (ie. more conserved in the population) than upsB/C types.'
Here are two problems:
(1) Is this comment still true in Ghana pilot data?

Yes, it is ($P < 0.001$).

(2) Are recombinant DBLα types significantly more likely to be observed in 10 or more isolates than nonrecombinant DBLα types?

No, they aren't. It's the opposite. Non-recombinant DBLα types are significantly more likely to be observed in 10 or more isolates than recombinant DBLα types ($P = 0.047$).

Third difference between JHMM and PHMM: parameters

In PHMM, gap open and gap extension probability are δ and ϵ;

In JHMM, gap open and gap extension probability are $\delta + \frac{\pi_I}{L}\rho; \epsilon + \frac{\pi_I}{L}\rho$
Consequently, recombination parameter ρ is introduced in transition matrix.

Parameter estimation in JHMM:

Step 1: δ, ϵ by Viterbi training algorithm;

Step 2: ρ by calculating composite likelihood.

Step 3: Calculate the Viterbi path with above parameters.

Third difference between JHMM and PHMM: parameters

Target seq: A G T C I F K K M F - - K D D

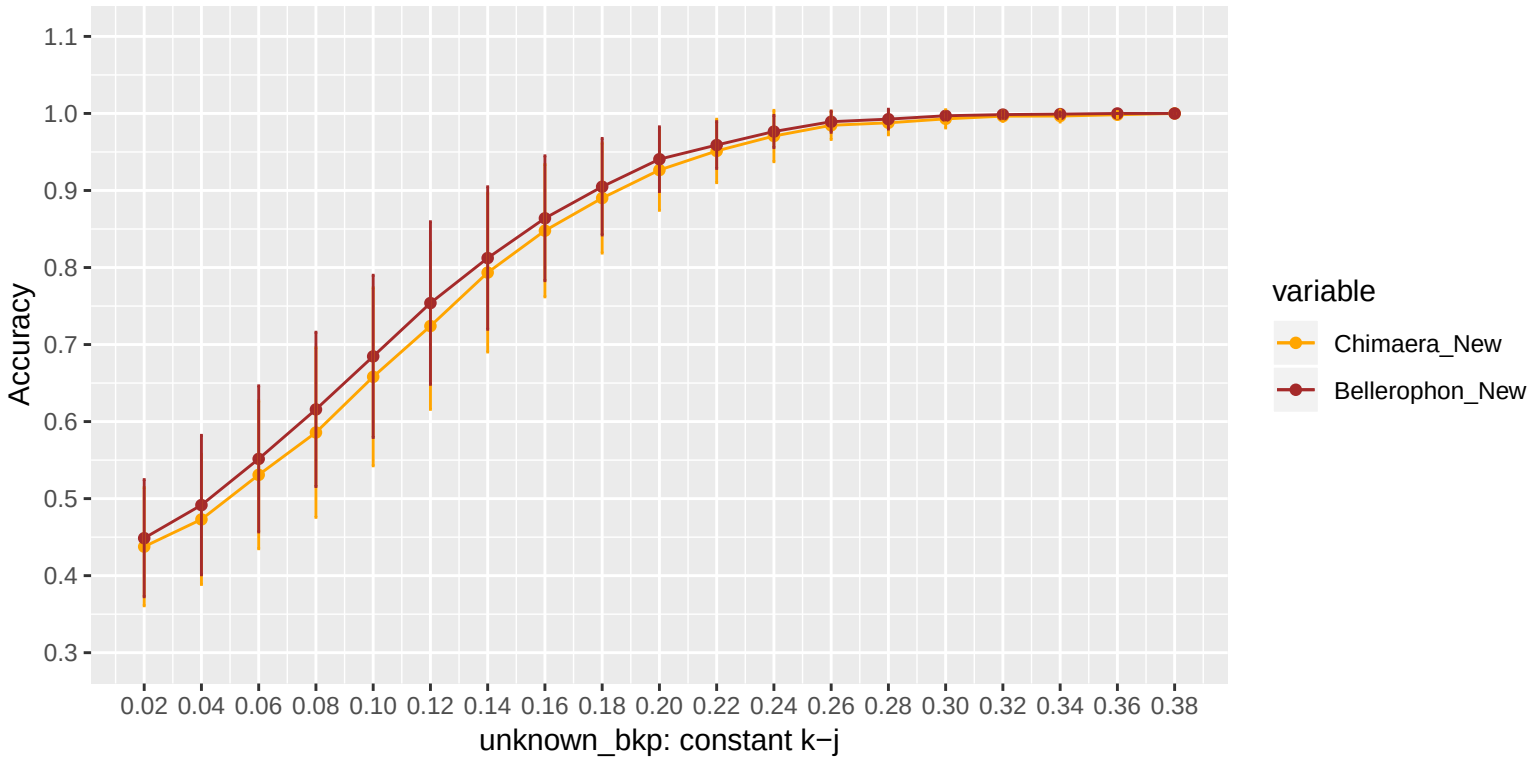
Source seq1: A G T - - F

Source seq2 K K M F Y Y K D D

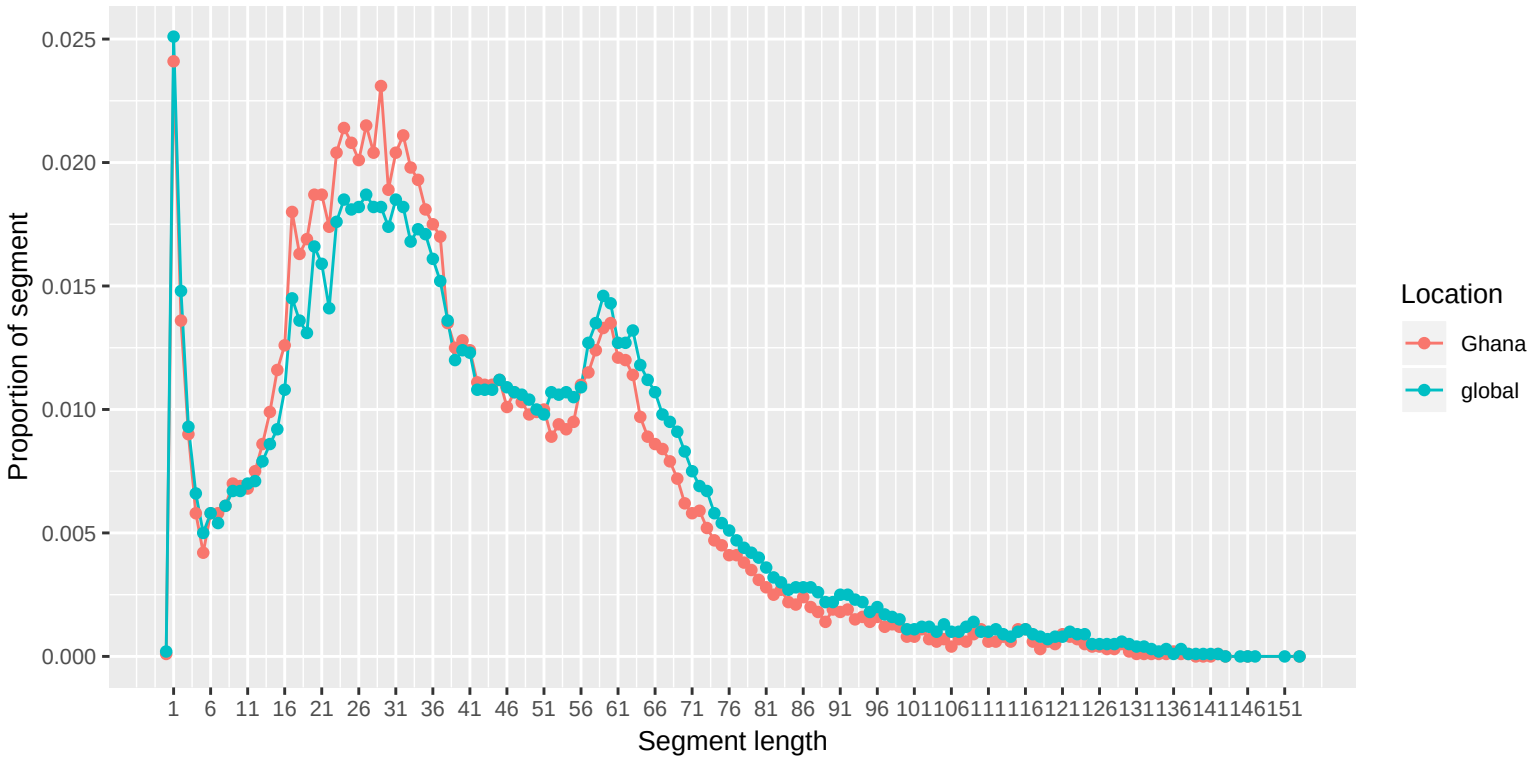
Hidden states: $M_{11}, M_{12}, M_{13}, I_{14}\ I_{15}, M_{16},$

$M_{25}, M_{26}, M_{27}, M_{28}\ D_{29}, D_{210}\ M_{211}, M_{212}$

More result


Accuracy
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.38
unknown_bkp: constant k−j
variable
Chimaera_New
Bellerophon_New

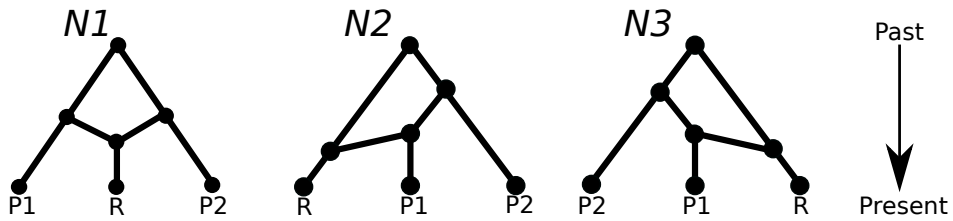
More result


0.025
0.020
0.015
0.010
0.005
0.000
Proportion of segment
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101106111116121126131136141146151
Segment length
Location
Ghana
global

Method comparison

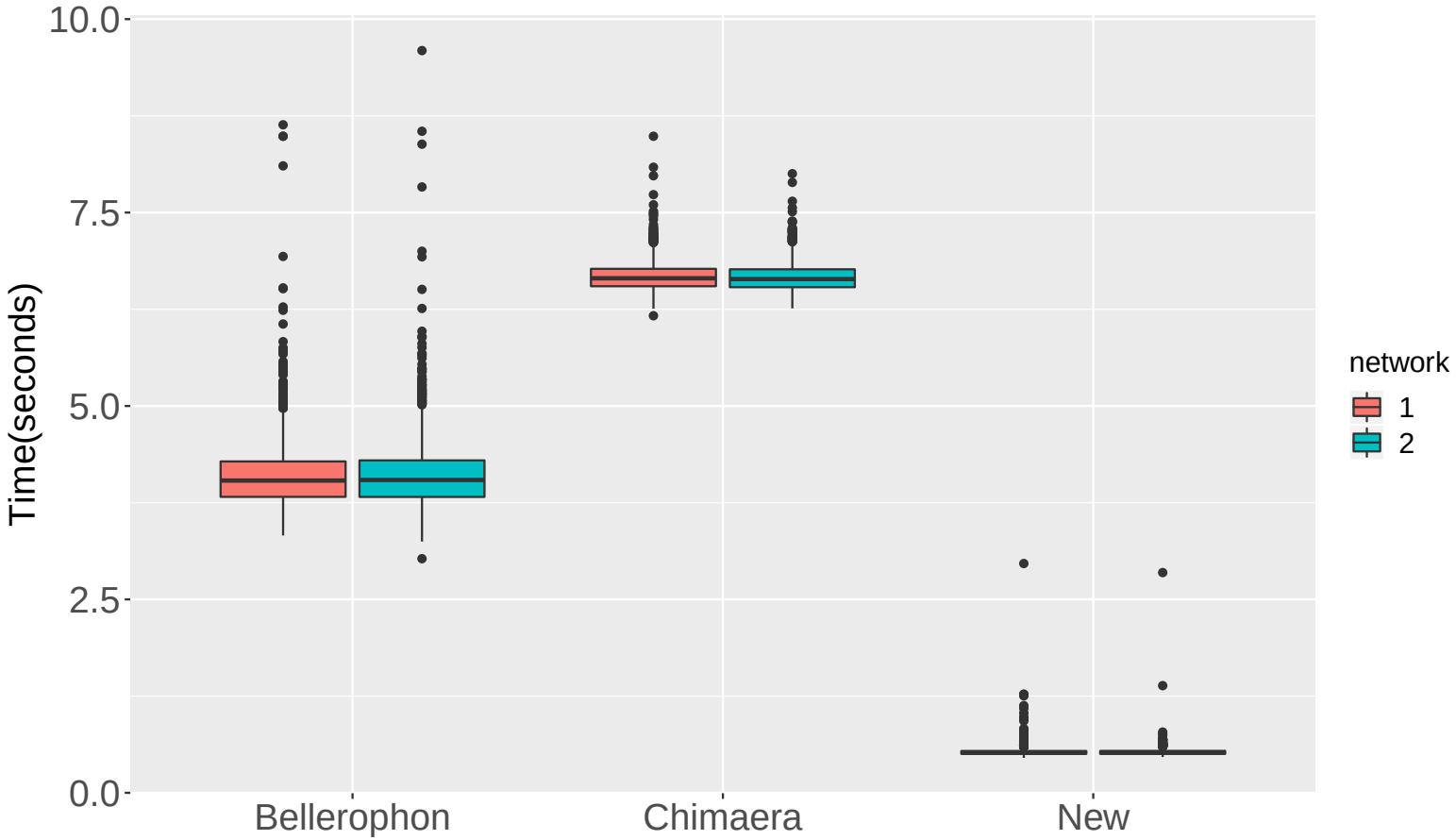
We compare Chimaera and Bellerophon at known and unknown breakpoints cases.

And we focus on the simplest case in which there are three sequences only



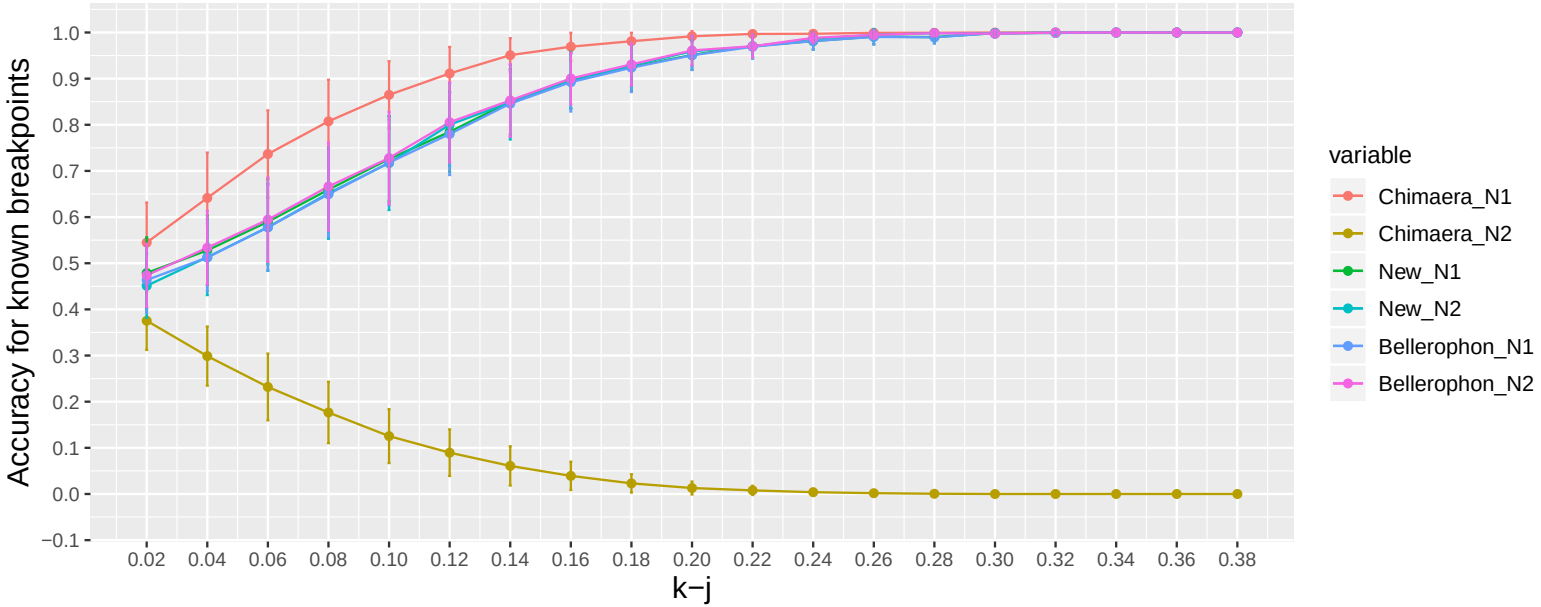
- N1 and N2 involve three branch lengths: i, j and k, i<j<k, sequential number from $[0.1, 0.5]$, interval is 0.02.
- For each network and each specific branch length setting, we simulate 100 groups of 400bp DNA sequences (setting for evolving sequences is the same with Posada et.al (2001))

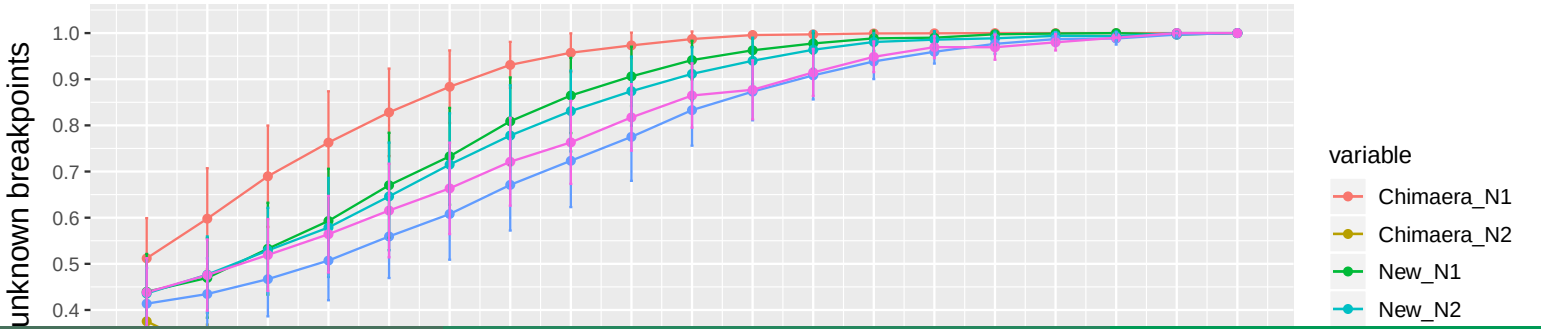
Simulation result about efficiency



Simulation result about accuracy

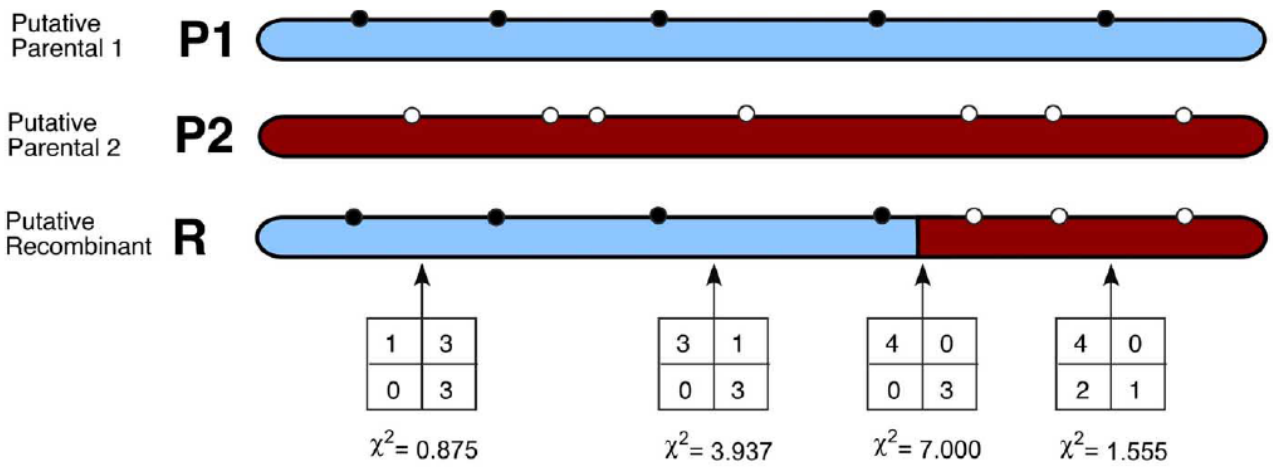
When breakpoint is unknown, we employ the identified breakpoint from Chimaera (very similar result with bellerophon).





Literature Review

- Substitution Distribution (Chimaera; Posada et.al 2001)



- Distance Methods (Bellerophon; Thomas et.al 2004)
  - Normally fast
  - Phylogeny does not need to be known
- Phylogenetic Methods, Compatibility Methods.