Education background

- Sept 2010 to July 2014.
  - China Agricultural University (211,985). Top 40 in China
  - Major: Mathematics and Applied Mathematics, School of Science
- Sept 2014 to June 2017.
  - Renmin University of China (211,985). Top 10 in China
  - Major: Epidemiology and Health Statistics, School of Statistics
- Sept 2017 to present
  - The University of Melbourne. Top 3 in Australia
  - School of Mathematics and Statistics
  - Melbourne Integrative Genomics

A scalable method for identifying recombinants from unaligned sequences

Qian Feng[1], Kathryn Tiedje[2,3], Shazia Ruybal[3,4,5,6], Gerry Tonkin-Hill[3,7,8], Michael Duffy[2,3], Karen Day[2,3], Heejung Shim[1], Yao-ban Chan[1,*]

[1] Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia
[2] Department of Microbiology and Immunology, The University of Melbourne, Bio21 Institute, Melbourne, Australia
[3] School of BioSciences, The University of Melbourne, Bio21 Institute, Melbourne, Australia
[4] Population Health and Immunity Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
[5] Department of Medical Biology, The University of Melbourne, Melbourne, Australia
[6] Burnet Institute, Australia
[7] Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Australia
[8] Parasites and Microbes, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom

Malaria

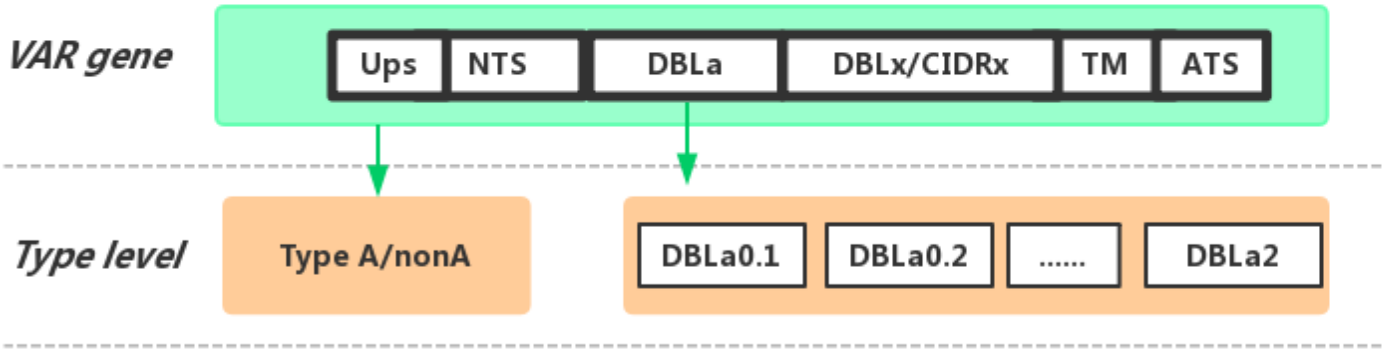
- Malaria is a serious, sometimes fatal, disease that is caused by a parasitic infection of the red blood cells.

- 2019 World Malaria Report
  - 228 million malaria cases globally in 2018, 405,000 malaria-related deaths in 2018.
  - The incidence rate of malaria declined globally between 2010 and 2018, however, the rate of change slowed dramatically, remaining at similar levels from 2014 to 2018.
  - Most cases occur in Africa (93%).

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

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.



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

Project aim

We aim to uncover *var* genes' evolutionary histories by constructing a phylogeny.

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

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

- Phylogenetic tree
- **Phylogenetic network**

parent 1: REDTADDKKIHG
parent 2: WALLKNRPNTDP
recombinant: REDTANRPNTDP

What does the phylogenetic tree/network look like?

phylogenetic tree

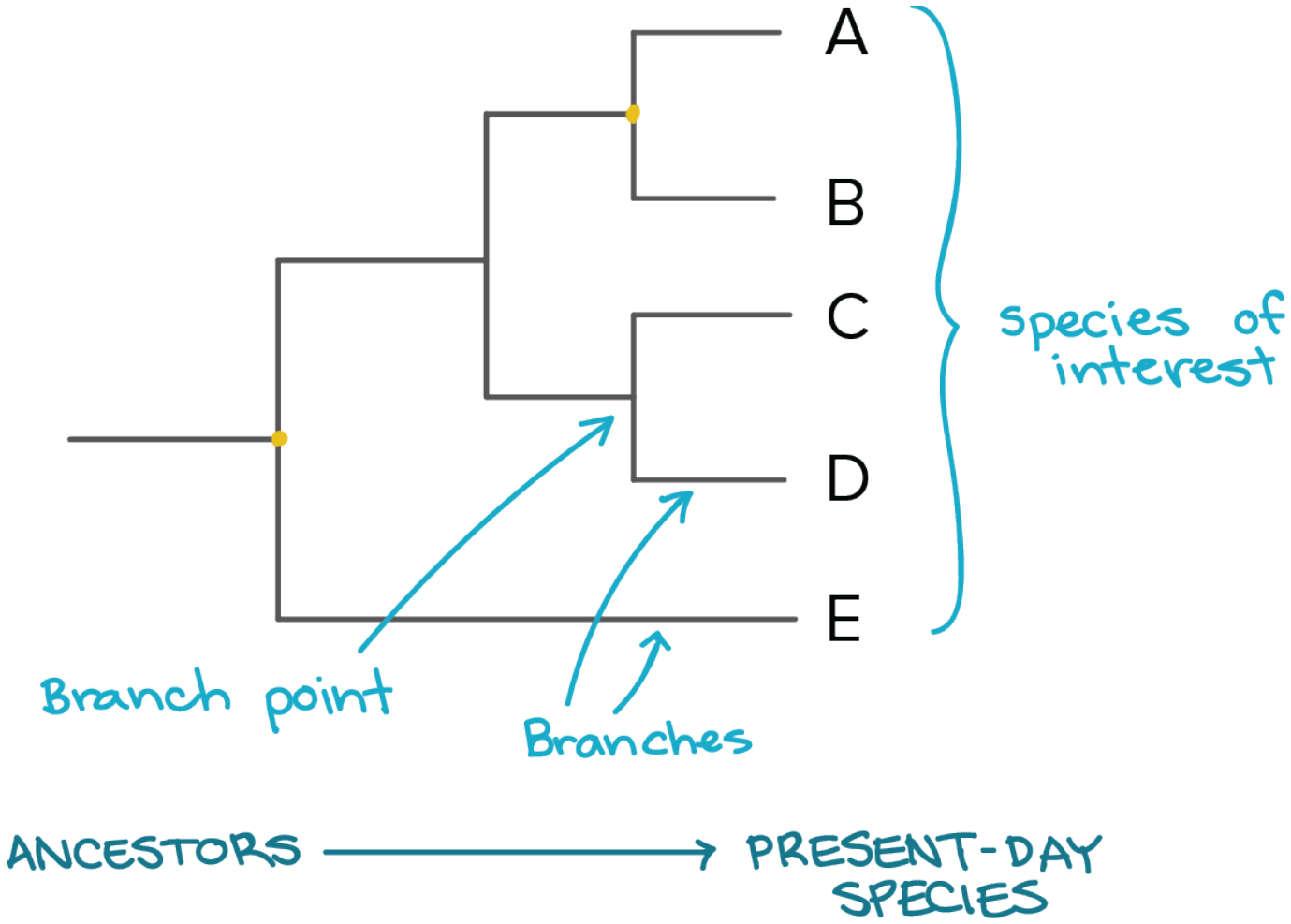
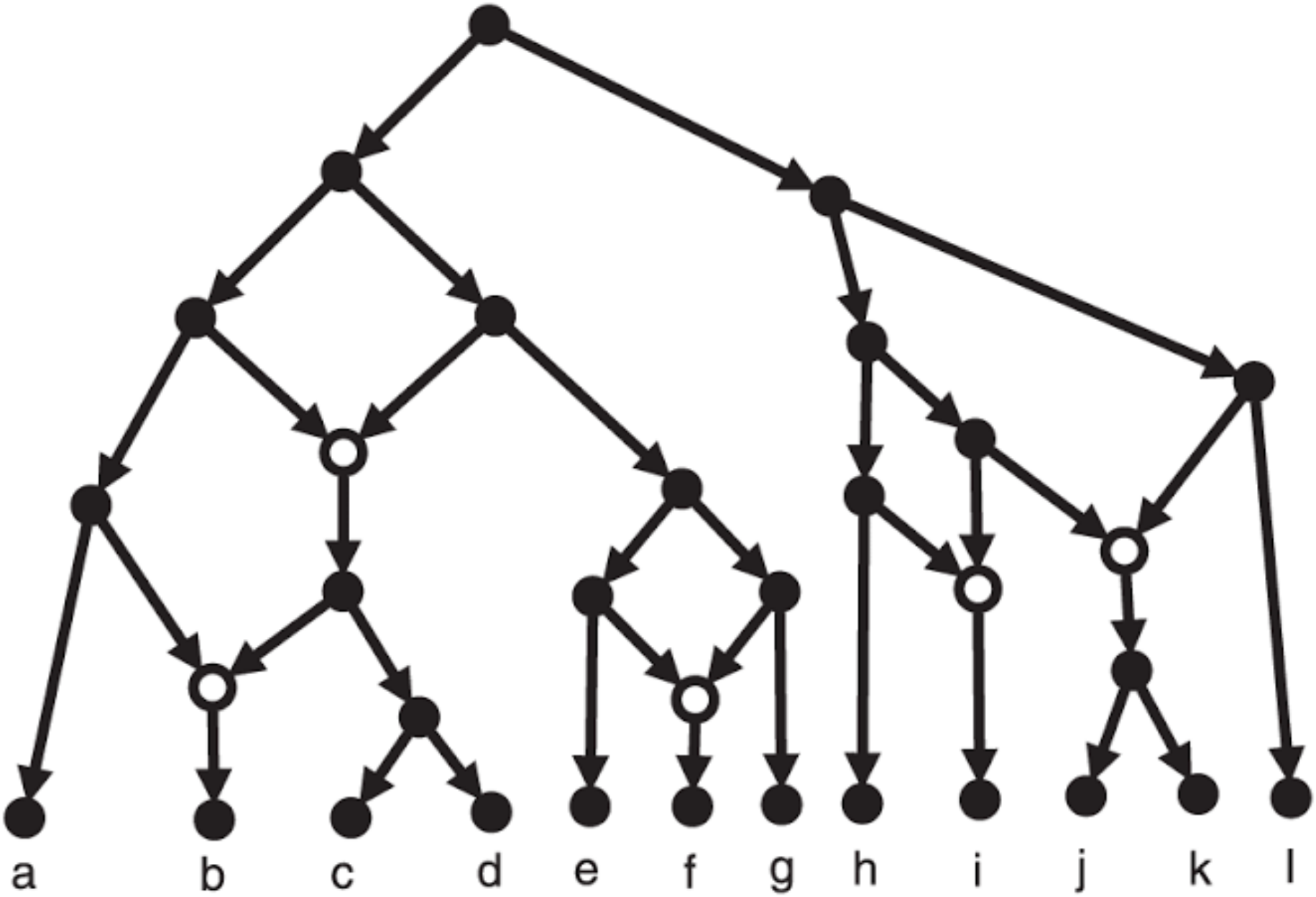


Image modified from Taxonomy and phylogeny: Figure 2 by Robert Bear et al., CC BY 4.0

phylogenetic network



Project aim

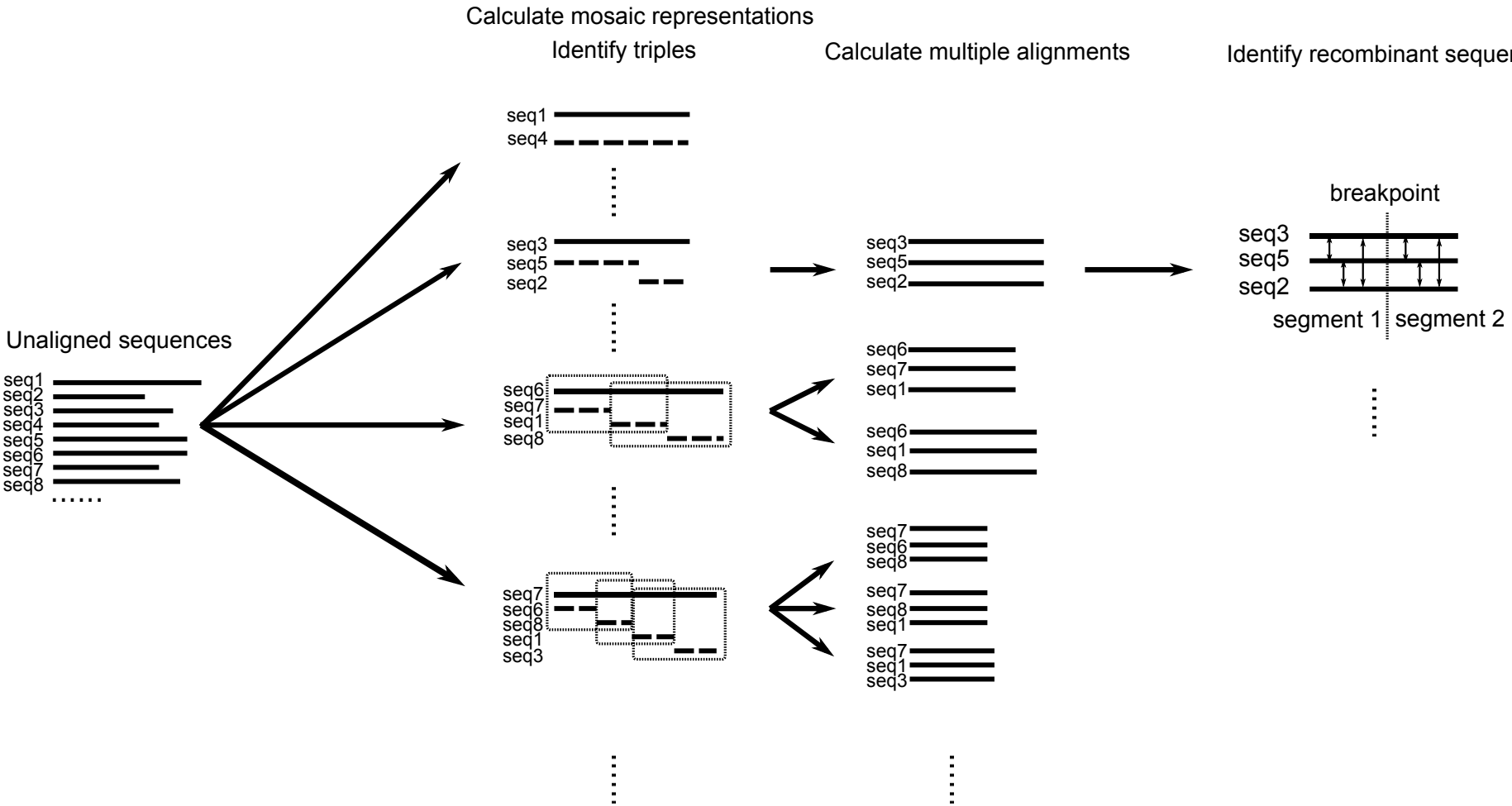
We aim to uncover DBLα sequences' evolutionary histories by constructing a phylogenetic network.

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

Recombinants Identification

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

A schematic of the algorithm



Advantages of proposed algorithm

- applicable to large number of sequences
- no need of multiple sequence alignment

- no need of reference genome sequences
- applicable not only in malaria, it holds great promise for many general applications to diverse gene families
- allow to analyze the properties of recombinants after application
  - How the proportion of recombinants change with time and space, for instance, comparison between wet and dry season?
  - How the breakpoint positions in recombinant sequences distribute (with time)?
  - Comparison between recombinants and non-recombinants
  - .....

Conclusion

limitation

Given the algorithm complexity, it would be more and more time-consuming if number of sequences increases.

Future work

- **Modify the JHMM in proposed algorithm so as to accommodate more input sequences and execute efficiently.**
- **Further application to real datasets.**
  - Explore the temporal and spatial features for the identified recombinants in bigger Ghana dataset, or even in global dataset.
- **Construct phylogenetic networks for these DBLα sequences.**
- **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. Qixin He, Dr. Zitong Li,
- Bobbie Shaban, Andrew Siebel and MIG students♡

Back up

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

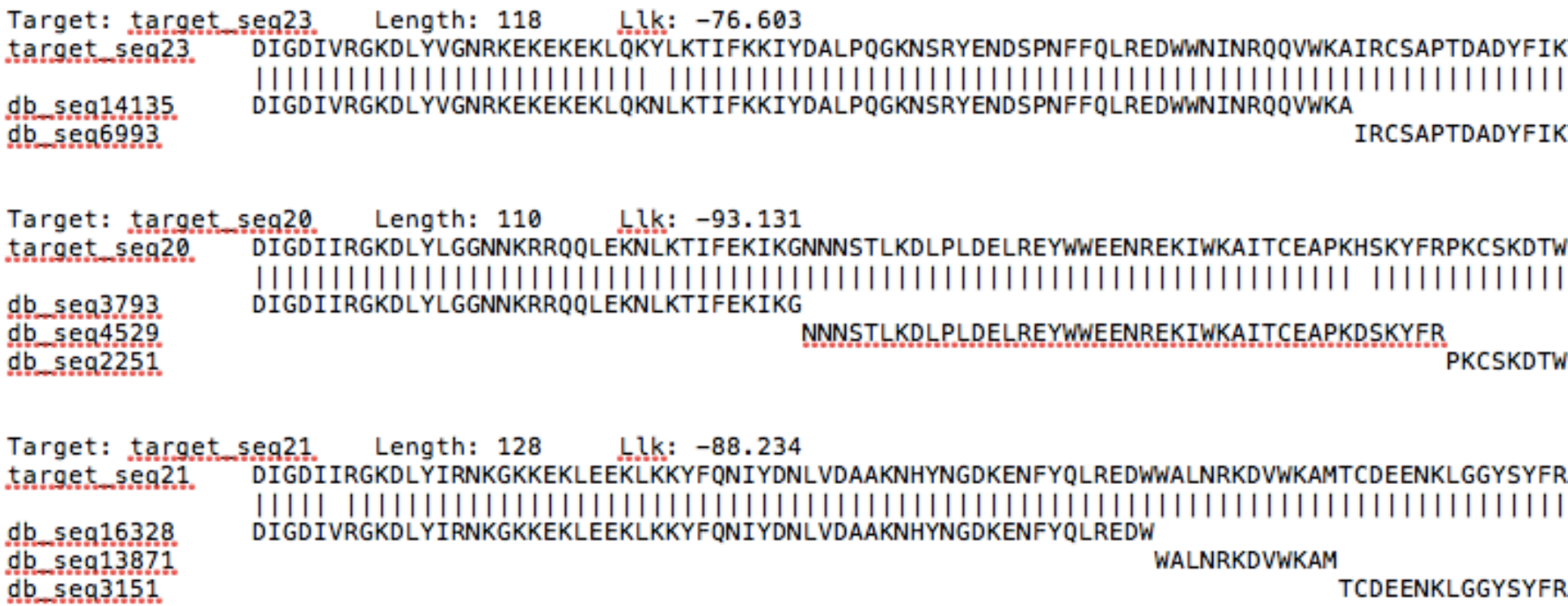
three parents

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

```
Target: target_seq23    Length: 118     Llk: -76.603
target_seq23     DIGDIVRGKDLYVGNRKEKEKEKLQKYLKTIFKKIYDALPQGKNSRYENDSPNFFQLREDWWNINRQQVWKAIRCSAPTDADYFIK
                 |||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
db_seq14135      DIGDIVRGKDLYVGNRKEKEKEKLQKNLKTIFKKIYDALPQGKNSRYENDSPNFFQLREDWWNINRQQVWKA
db_seq6993                                                                                  IRCSAPTDADYFIK


Target: target_seq20    Length: 110     Llk: -93.131
target_seq20     DIGDIIRGKDLYLGGNNKRRQQLEKNLKTIFEKIKGNNNSTLKDLPLDELREYWWEENREKIWKAITCEAPKHSKYFRPKCSKDTW
                 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
db_seq3793       DIGDIIRGKDLYLGGNNKRRQQLEKNLKTIFEKIKG
db_seq4529                                           NNNSTLKDLPLDELREYWWEENREKIWKAITCEAPKDSKYFR
db_seq2251                                                                                    PKCSKDTW


Target: target_seq21    Length: 128     Llk: -88.234
target_seq21     DIGDIIRGKDLYIRNKGKKEKLEEKLKKYFQNIYDNLVDAAKNHYNGDKENFYQLREDWWALNRKDVWKAMTCDEENKLGGYSYFR
                 ||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
db_seq16328      DIGDIVRGKDLYIRNKGKKEKLEEKLKKYFQNIYDNLVDAAKNHYNGDKENFYQLREDW
db_seq13871                                                                  WALNRKDVWKAM
db_seq3151                                                                               TCDEENKLGGYSYFR
```

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

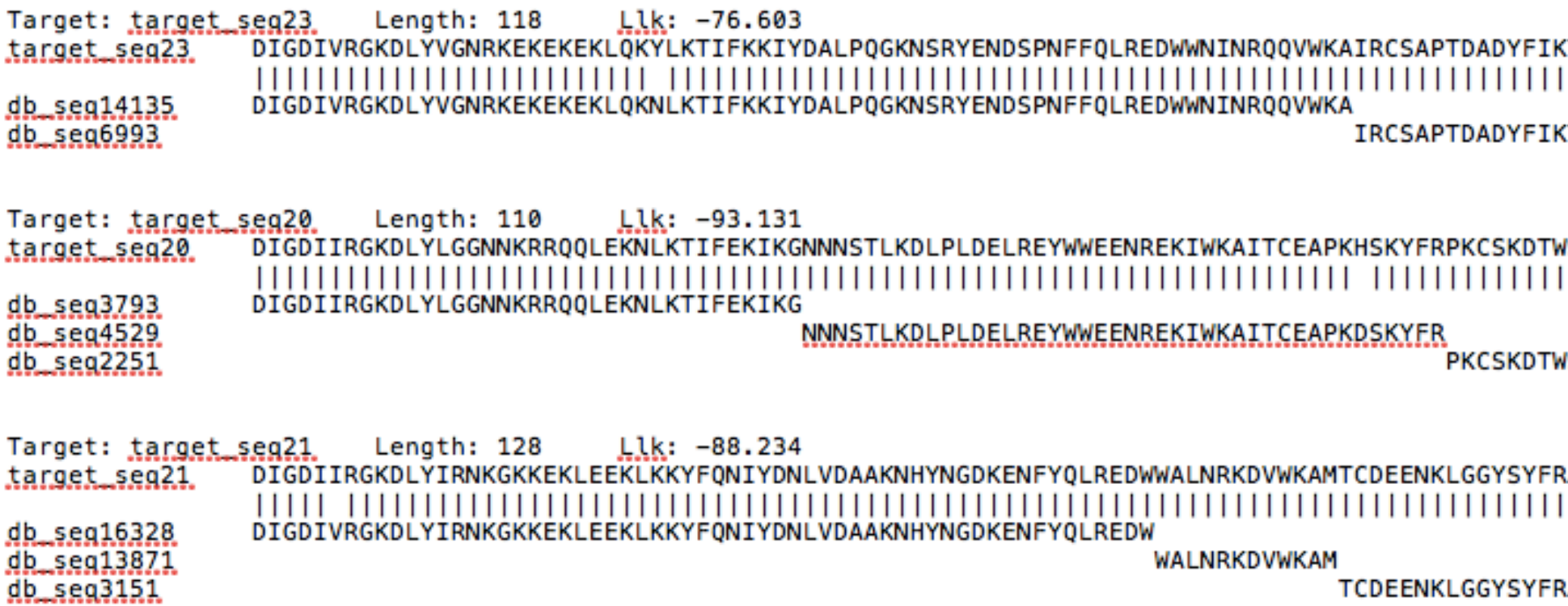
three parents

T A G T C K D I M M

D_1 A G T C

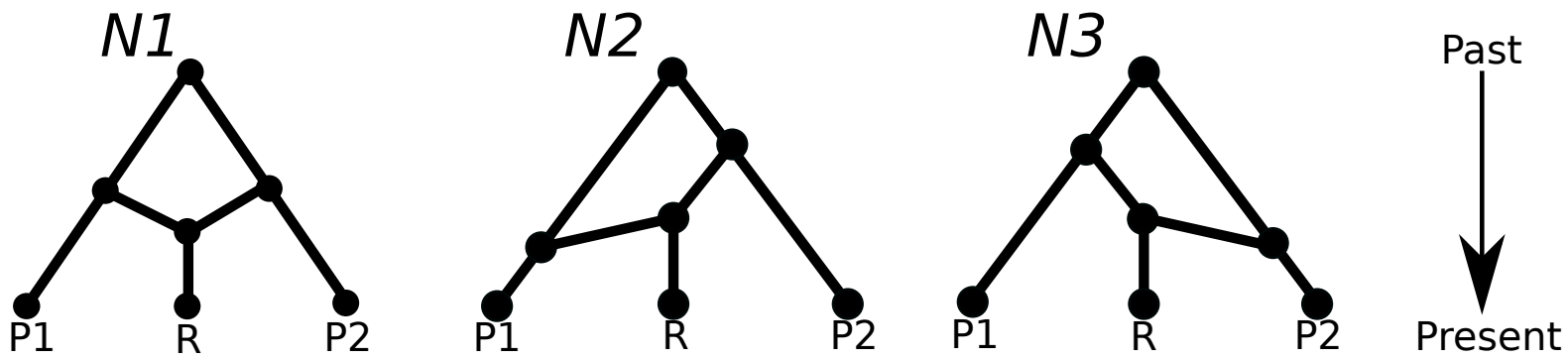
D_2 K D I M M

two parents



Which one is true recombinant for two parents case?

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



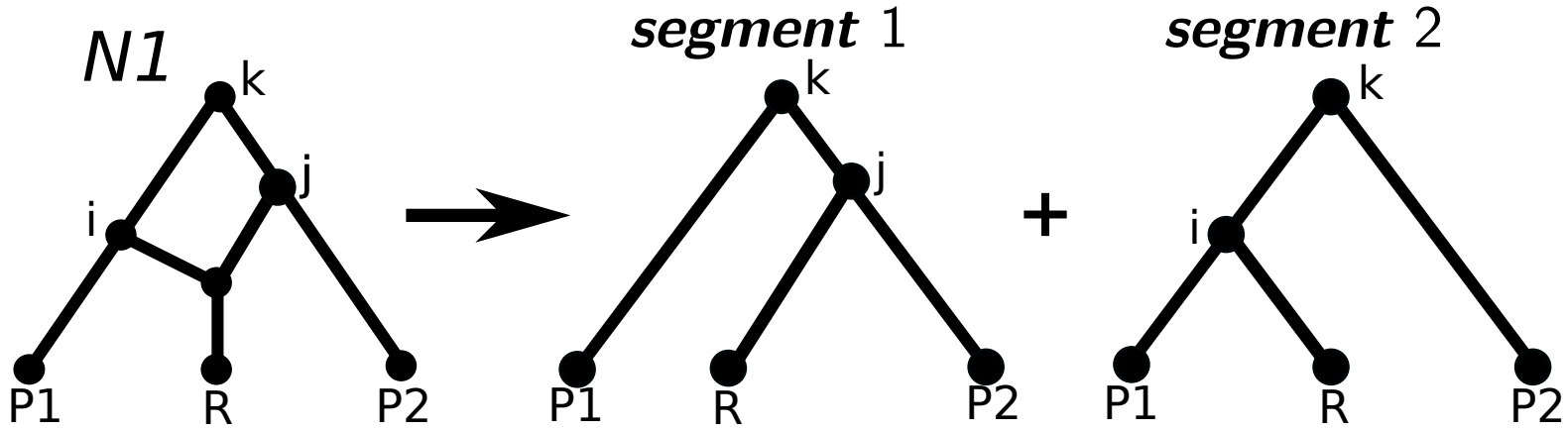
Our target is to find right one as accurately as possible and try to use the least time.
There is one key common in these three networks, two non-recombinants have very similar distance along sequences.

Key step in proposed algorithm

Core: non-recombinants have similar evolutionary distance in each triple.

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

$|D_1(P_1, P_2) - D_2(P_1, P_2)| = k - k = 0$; 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.

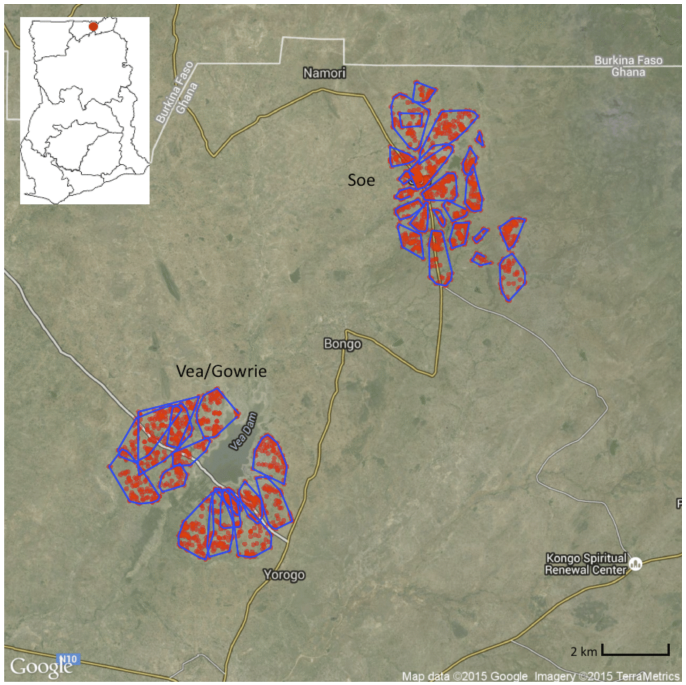
Step 4: Compute the absolute value of segment distance differences, **the smallest difference infers two non-recombinant sequences**.

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

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.

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.

Tiedje et.al,2017

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*)	0.985(0.776*)
A, B and C	0.655(0.509*)	0.510(0.304*)
	Same domain parents	Same domain family
	0.310(0.079*)	0.206(0.010*)

$\star$ 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*)	0.985(0.776*)
A, B and C	0.655(0.509*)	0.510(0.304*)
	Same domain parents	**Same domain family**
	0.310(0.079*)	0.206(0.010*)

$\star$ 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.