Identifying recombinant DBLα sequences in *Plasmodium falciparum*

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

Conclusion

- We propose a distance-based approach to detect recombinants in a large set of unaligned sequences.
- Algorithm enjoys high accuracy over a wide range of biologically realistic scenarios by simulations.
- Recombinant happens more frequently in the same ups (upstream promoter sequences) type group in a large dataset of DBLα tags from a high-transmission area of Ghana.

Background and Aim

- The *var* genes of the malaria parasite *Plasmodium falciparum* encode the PfEMP1 antigen, which controls the ability of the parasite to evade the human immune response system.
- These genes are hyper-diverse, principally due to recombination between the many (~ 60) genes per genome. The study of these genes is thus one core problem in current malaria research, with implications for future malaria interventions.
- The evolution of *var* genes can be studied through a conserved part of one of their domains, called DBLα tags.

- But which tags are recent recombinants, and how to locate breakpoint(s) for each recombinant? The evolutionary process of these recombinants has three different possibilities.

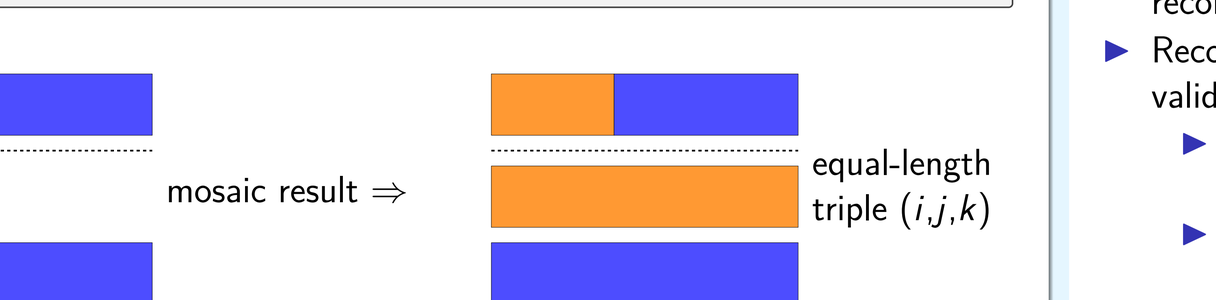

Figure: At network N1, R' s two parents are P1 and P2. At network N2, R's parents are P1 and the ancestor of P1 and P2. N3 is the mirror of N2.

There is one key common in these three networks, **P1 and P2 have very similar distance among sequences**, which are both non-recombinants, providing idea for algorithm step 4.

Simulation Results

We simulated 100 replicates for each setting, breakpoints are uniformly distributed along the recombined sequences, segments in recombinants are from independent datasets.

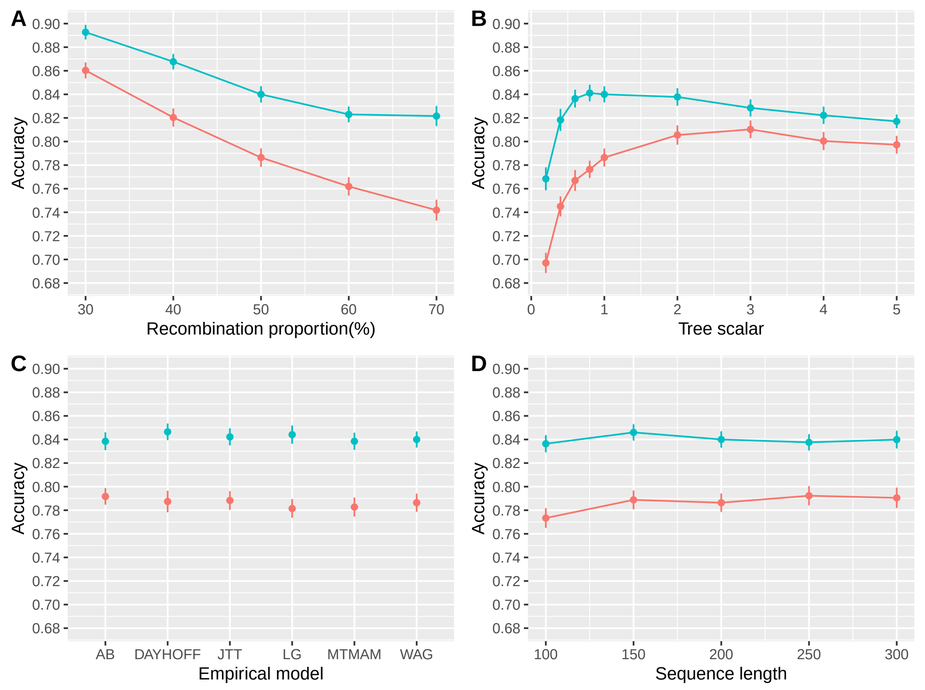

Figure: Red means each recombinant has two parents, green means three.

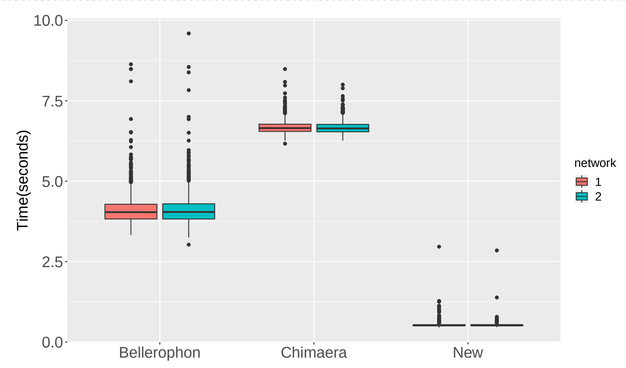

Figure: New method enjoys higher efficiency than Chimaera and Bellerophon with known breakpoints.

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.

$$Rec := \{i,j,k\} \backslash arg \min_{ij,ik,jk} \{|d_{ij}^{s_1} - d_{ij}^{s_2}|, |d_{ik}^{s_1} - d_{ik}^{s_2}|, |d_{jk}^{s_1} - d_{jk}^{s_2}|\}$$

i,j,k refers to each index of three sequences in one specific triple.

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.

Ghana Data Application Results

- 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. Before that a pilot study was conducted as well. Currently, we are analysing the pilot data involving 161 isolates, but we plan to analyse the whole survey data in the near future.
- 14801 out of 17335 representative protein sequences are identified recombinants.
- Recombinant happens more frequently in the same ups group statistically, validating previously reported theory.
  - HMMER and protein-protein BLAST methods are employed separately to get two ups classification systems.
  - At domain level, most domains also show same conclusion.

Table: P values for statistically proportional tests

	same ups parents	same ups family
A and non-A	$< 2.2e-16$	$< 2.2e-16$
A, B and C	$< 2.2e-16$	$< 2.2e-16$

Extensions

1. This novel algorithm is applicable not only in malaria, but also in RNA sequencing in cancer bioinformatics, in the context of detecting gene fusions.
2. We aim to construct phylogenetic networks for these DBLα tags, inferring their evolutionary histories.
3. Classification of semi-conserved upstream promoter sequences and its relationship with DBLα sequences.

Zilversmit, Martine M., et al., Hypervariable antigen genes in malaria have ancient roots. *BMC evolutionary biology* 13.1 (2013), p.110.
Tiedje, Kathryn E., et al. "Seasonal variation in the epidemiology of asymptomatic Plasmodium falciparum infections across two catchment areas in Bongo District, Ghana." *The Amer. J. Tro* 97.1 (2017): 199-212.

Website: https://qianfeng2.github.io Funding body: China Scholarship Council - University of Melbourne PhD Scholarship fengq2@student.unimelb.edu.au