Research Proposal

Project Title: Phy1ogenetic modeling of malaria *var* genes

Proposed Course: PhD (Mathematics and Statistics) at the University of Melbourne, 2017-2021

Student Name: Qian Feng

Telephone (+86)15201421685

1 Introduction

The malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*, which are transmitted by infected mosquitoes, account for more than 95% of all human malaria infections [5]. P. falciparum causes an estimated 200 million clinical cases and 300,000 malaria-related deaths annually, pre-dominantly in sub-Saharan Africa. P. vivax causes an estimated 16 million clinica1 cases of malaria (nearly half of all malaria cases outside Africa), predominantly in Asia, Oceania and Central and South America, To control and eliminate these parasites, it is important to understand their origins, evolutionary history and evolutionary mechanisms.

A key component of the virulence of *Plasmodium falciparum* is the ability of the parasites to escape elimination by causing infected erythrocytes (red blood cells) to adhere to the vascular membrane. This ability is mediated by members of the P. *falciparum* erythrocyte membrane protein 1 (PfEMP1) family, which are expressed by a gene family known as the *var* genes [1]. The *var* genes exhibit extreme levels of diversity, allowing the parasites to repeatedly evade detection from the human immune system and remain in the bloodstream over a period of many years.

The *var* genes are a very large multi-gene family (family of duplicate genes), comprising 40 to 60 genes per genome. They evolve extremely rapidly, have a very high leve1 of sequence diversity, and do not have stable locations in the genome. The primary mechanism of the evolution of *var* genes is genetic recombination, combined with single-base mutation.

Recombination is an evolutionary mechanism where two segments of different copies of a gene are “spliced” together, creating a novel gene sequence. This may occur between two alleles in a population1 or between two duplicate copies of a gene in a multi-gene family. This latter type of recombination, known as *paralog exchange* or non-allelic homologous recombination, is the driving force behind *var* gene evolution

and diversity [2].

To fully understand the mechanisms behind *var* gene evolution, we must first reconstruct its evolutionary history. Traditionally, evolutionary history is depicted by means of a phylogenetic tree or phylogeny, a mathematical tree where leaves represent current-day species and interna1 branches represent extinct ancestors. There are a wide varirty of methods for reconstructing such trees, generally from protein or DNA genome sequences.

Unfortunately, due to the complexity and rapidity of *var* gene evolution, traditional phy1ogenetic methods do not work for this gene family. Most importantly, paralog exchange breaks the fundamental assumption of phylogenetics, that evolution can be expressed in the form of a tree. Thus specialized mathematical methods are needed to analyze this complex gene family.

2 Project aims and methodology

This project aims to deve1op mathematical methods to reconstruct the evolutionary history of *var* genes. This will proceed in three parts:

1. We will develop methods to construct phylogenetic trees for *var* genes. As noted above, the extreme diversity, rapid evolution and lack of a stable location in the genome a1l combine to make traditional phylogenetic methods work poorly or simply fail altogether for this gene family. In particular, the first step of constructing a phy1ogeny usually involves performing a multiple sequence alignment (MSA) of the available sequences; this enables only homologous characters to be compared to each other. Due to the diversity of *var* genes, even the alignment problem is difficult to solve; the method of Zilversmit et a1[6], based on hidden Markov models, was deve1oped specifically for this problem. This work will be a first step towards the construction of phy1ogenetic trees for *var* genes. A specialized stochastic model for *var* gene evolution will then be constructed, and a phylogeny built through maximum-likelihood or Bayesian methods.

   An alternative approach may simply be to skip the alignment step altogether by using an alignment-free (AF) method. AF methods are usually based on frequency counts of short subsequences and have been used in phylogenetics for phylogeny construction and detection of lateral genetic transfer [3].

2. We will develop methods for the inference of paralog exchange. Most phylogeny construction methods do not allow for any form of recombination, but reconstruct evolutionary history only in the form of a tree. In order to properly understand the

evolution of *var* genes, we must be able to infer recombination events. We will model evolutionary history with a phylogenetic network, rather than a tree, to include recombination. There are a variety of methods for inferring these networks, with many coming from the related field of population genetics and studied under the term Ancestral Recombination Graph (ARG). However, these methods are often slow and do not scale to large datasets. Moreover, they are intended for the inference of allelic recombination, rather than paralog exchange.

We will devise method specifically for the detection of paralog exchange. Comparison of gene phylogenies to the species tree, by way of mathematical mappings called *reconciliations* [4], will be useful here, as they allow inference of which genes are duplicates or not, reducing the search space and allowing for the construction of efficient algorithms.

3. We will apply the developed methods to real-life *var* gene datasets. There are world-leading experts in malaria at the University of Melbourne (Karen Day and Kathryn Tiedje) who have expressed an interest in this project and have access to up-to-date *var* gene databases. By applying our methods to real-life datasets, we can understand the mechanisms behind the evolution of *var* genes, and in particular deduce the frequency of recombination and the role it plays in the diversification of the gene family. This will provide a better understanding of the way the *Plasmodium* parasite evades human immune responses, and will give information which may eventually lead to advancements in the treatment of malaria.

Aim	Time
Phylogenetic tree inference for war genes	2017 - 2018
Inference of paralog exchange	2018 - 2019
Application to real-life datasets	2019 - 2020
Write and submit thesis	2020- 2021

Table1: Timeline for the project

3 Outcomes and timeline

This project will produce at least 3 papers in leading journals, at least 1 presentation at a significant international conference in computational biology, and a completed PhD thesis. The timeline for this project is given in Table1.

References

[1] D. S. Chen, A. E. Barry, A. Leliwa-Sytek, T.-A. Smith, 1. Peterson, S. M. Brown, F. Migot-Nabias, P. Deloron, M. M. Kortok, K. Marsh, et al. A molecular epidemiological study of var gene diversity to characterize the reservoir of plasmodium falciparum in humans in africa. PloS One, 6(2); e16629, 2011.

[2] A. Claessens, W. L. Hamilton, M. Kekre, T. D. Otto, A, Faizullabhoy, J. C. Rayner, and D. Kwiatkowski. Generation of antigenic diversity in *Piasmodium faiciparum* by structured rearrangement off Var genes during mitosis, PLoS Genet., 10(12):e1004812, 2014.

[3] Y. Cong, Y. Chan, and M. A. Ragan. A novel alignment-free method for detection of latera1 genetic transfer based on TF-IDF. Sci. Rep, 6, 2016.

[4] J.-P. Doyon, V. Ranwez, V. Daubin, and V. Berry. Models, algorithms and programs for phy1ogeny reconciliation. Brief. Bioinform, 12(5):392-400, 2011.

[5] D. E. Loy, W. Liu, Y. Li, G. H. Learn, L. J, Plenderleith, S. A. Sundararaman, P. M. Sharp, and B. H. Hahn. Out of africa: origins and evolution of the human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax.* Int. J, Parasito1, 2016.

[6] M. M. Zilversmit, E. K. Chase, D. S. Chen, P. Awada1la, K. P. Day, and G. McVean. Hyper-variable antigen genes in malaria have ancient roots. BMC Evol. Biol, 13(1):110, 2013.