Analysis of the Sub-Telomeric Region of the Malaria Parasite *P. falciparum* Using aCGH Data

Diego Diez\(^1\) Ulf Ribacke\(^2\) Craig Wheelock\(^2\) 
diez@kuicr.kyoto-u.ac.jp ulf.ribacke@smi.ki.se craig@para-docs.org 
Mats Wahlgren\(^2\) Susumu Goto\(^1\) 
mats.wahlgren@smi.ki.se goto@kuicr.kyoto-u.ac.jp 

\(^1\) Bioinformatics Center, Institute for Chemical Research, University of Kyoto, Uji, Kyoto 611-0011, Japan 
\(^2\) Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280 SE-171 77, Stockholm, Sweden

**Keywords:** malaria, copy number polymorphism

1 Introduction

Malaria is an infectious disease caused by the protozoan parasite of the genus Plasmodium of which *P. falciparum* is the most deadly species. Over 500 million people suffer from malaria each year, resulted in more than one million deaths, almost 90% of which occur in Sub-Sahara Africa. *P. falciparum* uses a mechanism called antigenic variation to evade the immune system of the human host, by expressing different highly variable surface proteins in the membrane of the infected erythrocytes. Several gene families are involved in this process including *var*, *rifin* and *stevor* [1]. Most of these gene families are known to be located in the sub-telomeric region of the chromosome, where they are subject to different mutational and rearrangement processes.

Recently, a study on copy number polymorphism (CNP) in different strains of *P. falciparum* has been conducted, using the malaria cDNA array from the Karolinska Institutet, but only non sub-telomeric regions have been studied so far. The extent of CNP events in the sub-telomeric regions and how this could be influencing the pathogenicity of the protozoan is still not known in detail. In this work we describe the development of a tool for the analysis of genome CNP using the R platform of statistical analysis, and the application to the study of the sub-telomeric region of the *P. falciparum* genome.

2 Method and Results

We have used the data from an array comparative genomic hybridization (aCGH) study comparing the sequenced *P. falciparum* strain against 9 different laboratory and field strains. The data set is available through the ArrayExpress database as E-MEXP-494. The sequences for the genes and chromosomes and the gene description file were obtained from the PlasmoDB website. The sequences of the array probes were mapped to the chromosomes using the Blat alignment software in the “fastMap” mode (i.e. 100% sequence identity and no introns).

The analysis was carried out using the R platform for statistical analysis [2] and several packages of the Bioconductor project [3]. We created several tools to integrate the gene description and Blat output files for further analysis within R. Also, a tool for visualization and analysis of the genomic and microarray data was developed (Figure 1). An R package with all this tools was prepared and is available for download.
The microarray data was loaded into R, following background correction and normalization using the limma package [4] from Bioconductor. An automatic method based on the distribution of background spots was applied to detect and select high background spots that can introduce some bias to the analysis. These spots were down-weighted in the normalization and later in the linear model computation.

After the pre-processing a linear model was fitted to each gene using limma, including a Dye effect to account the extent of Dye bias in the experiments. Finally, genes were selected based on an absolute fold change. We found that, as expected, most of the genes in the sub-telomeric regions seems to show CNP events, indicating that these regions are subject of major chromosomal rearrangements. Surprisingly, only deletion events were found and almost no duplication (Figure 1).

3 Discussion

The sub-telomeric regions are composed mainly of genes of the var and rifin families. Most of the probes against these gene families show a decrease in intensity compare to the reference strain (3D7), suggesting a deletion in these genes. This can have two explanations: i) The sample strains lack most of the var and rifin genes. This hypothesis is unlikely as then, the mechanisms of evading the immune system could not be functional. ii) The strains do have var and rifin genes, indicating that even in the same species but different strains, variability in the gene family is so high that the probes designed against the 3D7 strain are not able to detect the signal from the other strains.

Acknowledgments

This work was supported by a Strategic Japanese-Swedish Cooperative Program on “Multidisciplinary BIO” grant.

References


