Identifying Copy Number Variations in Plasmodium falciparum

Brian Bishop
Advisor: Scott Emrich

ABSTRACT

The purpose of this effort is to investigate and develop a new method for identifying copy number variations (CNVs) in the malaria parasite. Malaria is a deadly disease that is extremely difficult to analyze due to a high rate of mutation combined with a high %AT. Because current drugs are becoming less effective in treating *Plasmodium falciparum* infections, specifically in areas of South East Asia, new methods need to be established to better understand emerging resistance in these parasites. Here we propose a new method to find CNVs, which has been previously associated with drug resistance. We use traditional read depth methods (post-alignment) to analyze two drug-selected *P. falciparum* genomes relative to their parent strain. These results are then compared to a well-known CNV finding algorithms, CNVnator. This new approach shows promise to furthering our knowledge of malaria and equips researchers with a new method for analyzing the genome of *P. falciparum*.

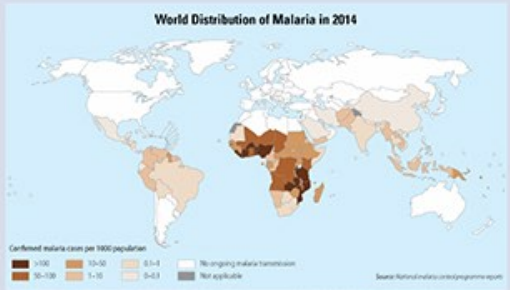

Fig 1.Distribution of Malaria in 2014 (Source: World Malaria Report 2014)

INTRODUCTION

- Malaria is a deadly disease caused by the *P. falciparum* parasite.
- Malaria affects 20 million people in the world every year and kills nearly 1 million each year.
- Drug resistant malaria has become a global concern, with drug resistance developing in areas of Southeast Asia with our latest drug ACT or Artemisinin Combined Therapy.
- Drug resistance has been linked to changes within the parasites genome.
- The genome of *P. falciparum* is difficult to analyze due to the high AT abundance.

OBJECTIVE

- Identify potential sources (genes) for the cause of the ACT resistance developing in Southeast Asia using drug-selected clones to start.
- Develop new methods for analyzing the *P. falciparum* genome.

MATERIALS AND METHODS

Sequencer Data

Sequencer data was collected for 2 ACT drug pressured isolates (2H12 and DD2) and the parent isolate (DD2). The data was gathered by lab based researchers who used drug pressuring techniques to induce in vitro drug resistance. The organisms were sequenced using an Illumina Next-Gen Sequencer. The reads were provided as forward and backward paired end reads in the format of a FastQ format.

Read Depth (Post-alignment)

For each of the isolates, the raw sequencer reads were aligned to the well annotated 3D7 reference genome using the typical SNP calling pipeline. The resulting VCF files was parsed to identify the read depth of each individual nucleotide position (DP field). Skellum values were calculated, by subtracting the 2A2 read depth value from DD2's read depth value (DD2 – 2A2). The resulting Skellum values were used in a Hidden Markov Model that was trained to identify CNVs.

CNVnator (Post-alignment)

CNVnator was run using the final BAM files that were created during the read depth approach. CNVnator was run using a window-size of 250 and the same 3D7 reference genome as in our approach. Since CNVnator finds CNVs based on the reference genome, we compared the results from running the algorithm on DD2 and 2A2. The CNVs that were found in the DD2 strain were compared to those found in 2A2 and were discarded.

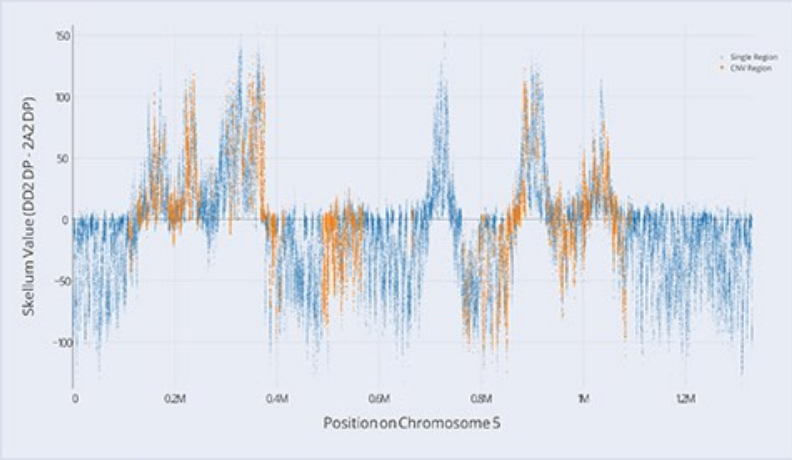

Fig 2. Skellum Values for Chromosome 5 (DD2 - 2A2)

Fig 2. Sample Read Depth Example

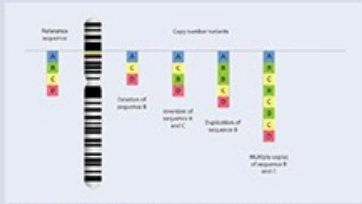

Fig 3. Potential CNVs (Source: EMedMD.com)

RESULTS

- Using the read depth HMM approach and variable windows, our model predicted **4787 potential CNVs** of length 250 or greater.
- Using the well known program CNVnator and fixed (large) windows, **85 potential CNVs** of length 250 or greater were found.

Table 1. Potential CNVs found with a length greater than 250 bases

Chromosome	1	2	3	4	5	6	7
Read Depth HMM	72	111	188	109	243	317	243
CNVNator	1	3	5	10	2	9	11

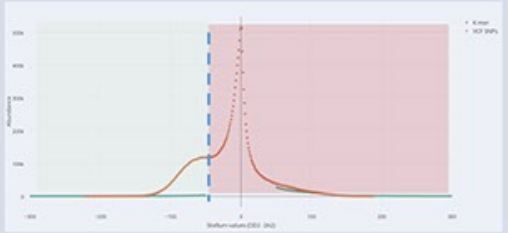

Fig 4. Skellum comparison K-mer vs Read depth

CONCLUSIONS

- Drug resistance with malaria is becoming a global concern and current analysis tools are limited due to the difficult nature of the *P. falciparum* genome.
- The read depth HMM approach shows potential to identify more CNVs at a finer grain (more specific) than many other tools already available.
- New tools need to consistently be developed in order to fully understand the complex nature of this deadly parasite.

ACKNOWLEDGMENTS

We wish to thank Dr. Roland Cooper (Dominican University of California) and Melissa Stephens (Notre Dame Genomics and Bioinformatics Core) for data submission and generation, respectively. This summer research experience was supported by the NSF DISC REU site (PI Thain), Computer Science and Engineering and by members of the Notre Dame Bioinformatics lab.

REFERENCES

1. Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome research, 6, 974–984.