www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating 2017: 5.03
TPI 2017; 6(12): 30-34
© 2017 TPI
www.thepharmajournal.com
Received: 08-10-2017
Accepted: 09-11-2017

Comparative genome analysis of *Plasmodium sp*. and identification of unique signature with next generation sequencing technology

Maulik Patel, Harshida Gadhavi, Tanushree Tiwari, Parantap Pandya, Saumya Patel, Rakesh M Rawal and Himanshu A Pandya

Maulik Patel
Department of Botany, Bioinformatics and Climate Change Impacts Management, University School of Sciences, Navrangpura, Ahmedabad, Gujarat, India

Harshida Gadhavi
Department of Botany, Bioinformatics and Climate Change Impacts Management, University School of Sciences, Navrangpura, Ahmedabad, Gujarat, India

Tanushree Tiwari
Department of Biology, York University, Toronto, Ontario-Canada

Parantap Pandya
Department of Botany, Bioinformatics and Climate Change Impacts Management, University School of Sciences, Navrangpura, Ahmedabad, Gujarat, India

Saumya Patel
Department of Botany, Bioinformatics and Climate Change Impacts Management, University School of Sciences, Navrangpura, Ahmedabad, Gujarat, India

Rakesh M Rawal
Department of Life Sciences, University School of Sciences, Gujarat University, Navrangpura, Ahmedabad, Gujarat, India

Himanshu A Pandya
Department of Botany, Bioinformatics and Climate Change Impacts Management, University School of Sciences, Navrangpura, Ahmedabad, Gujarat, India

Correspondence
Maulik Patel
Department of Botany, Bioinformatics and Climate Change Impacts Management, University School of Sciences, Navrangpura, Ahmedabad, Gujarat, India

Abstract

Malaria is a malignant disease which is growing all over the world and its causative agent. *Plasmodium* species easily develops resistant to commonly used antimalarial drugs easily. These empower different strains of *Plasmodium* e.g. *Plasmodium falciparum* and *Plasmodium vivax* to infect humans with malaria.To get the deeper molecular insights, next generation sequencing data were used for further analysis as it has shifted the paradigm of genomics to address biological questions with high confidence and in timely manner. The short reads for above mentioned parasites were retrieved from SRA (Sequence read archive) and de novo assembly was performed. Several novel genes along with known genes were predicted from assembled contigs, Functional annotation followed by gene ontology and pathway analysis. Comparison between species gave structural and functional diversity of the specific genes responsible for disease condition which further can be studied for disease biology.

Keywords: Malaria, Genome Analysis, *de novo* assembly, next generation sequencing (NGS), functional annotation

Introduction

The most common forms of human malaria are caused by *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium knowlesi* and *Plasmodium malariae*. The most important of these are *Plasmodium falciparum* and *Plasmodium vivax* as it can be fatal and lead to death of a person. Malaria was first recognized in 1880 as a disease caused by parasitic infection [1]. It is a malignant blood disease induced by parasites transmitted to humans through the bite of the *Anopheles* mosquito which are better known as malarial vectors. Infected mosquitoes acquit the *Plasmodium* parasite. Infected mosquito bites human and transmits the parasites; than these intruders multiply in the host liver before infecting and destroying red blood cells. *Plasmodium falciparum is* found in tropical and semitropical areas worldwide. Approximately 1 million people are killed by *Plasmodium falciparum* every year [2]. Whereas *Plasmodium vivax* resides in Asia, Latin America and in parts of Africa. It is the most prevalent human malarial parasite [3]. Data and statistics give evidence that malaria is spread all over the world and the malarial parasites develop resistance very soon against the drugs and insecticides. Artemisinin is one of the synthetic derivatives which are a group of drugs used against *Plasmodium falciparum* and *Plasmodium vivax* malaria [34, 35]. In April 2011, the WHO stated that resistance to the most effective antimalarial drug, artemisinin, could unravel national (India) malaria control programs, which have achieved significant progress in the last decade. WHO advocates the rational use of antimalarial drugs and acknowledges the crucial role of community health workers in reducing malaria in the region [36, 37]. According to malaria report 2016 by WHO Malaria growth is decreased by 29% but still *Plasmodium vivax* is being resist with present drugs [31]. So more robust methods are required to control the loss caused by malarial parasites. The NGS genomics approach is useful as it is economically practical and can give effective results to control it.

Next-generation sequencing describe a number of different modern sequencing technologies including: Illumina (Solexa) sequencing, Roche 454 sequencing, Ion torrent: Proton / PGM sequencing [21, 22, 28]. These advanced technologies allow us to sequence DNA and RNA much faster and cheaper than the previously used Sanger sequencing and as such have revolutionized the study of molecular biology and genomics [10, 26, 27].

Whole genome analysis with NGS involves recombinant DNA, DNA sequencing, and bioinformatics [4, 5]. It optionally includes studies of intra genomic phenomena such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome [6, 23]. In contrast, the investigation of the functions and roles of single genes is a primary focus of molecular biology or genetics and is a common topic of modern biological and medical research. Research of single genes does not fall into the definition of genomics unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genomes networks [7, 8]. Genome assembly is carried out with sequenced reads, which are the fragments of the actual genome [20, 29]. The assembled genome depends on heuristics and the data available [9, 30]. The analysis of assembled scaffolds or contigs functional annotation is must [24, 25]. This involves classifying various regions of the assembled genome in coding and non-coding regions [11, 12]. The annotated regions information helps in doing the downstream analysis for curing and extrapolated target information.

Methods

Raw data was retrieved from NCBI-SRA database with the following Accession numbers: SRR767807 [32] (*Plasmodium falciparum*), SRR1568115 [33] (*Plasmodium vivax*) on the basis of clinical isolates and literature study. [13] SRA data was extracted into fastq format using fastq-dump from SRA Toolkit 2.5.7 tool [18].

De novo Assembly and Functional Annotation

The raw data was quality checked and trimmed to remove the adapters as well as other contamination using Trimmomatic 0.35. The clean reads were further used for de-novo genome assembly using SOAPdenovo v2.04 [17] with the de Bruijn graph data structure and the wide range of odd k-mer values. The contigs generated by kmer 62 were selected based on N50, closest to genome size, minimum gaps and contig number. These contigs were further processed to minimize the number of N's using gap closer which results in scaffolds.

Assembled scaffolds larger than 500 bp were used for downstream analysis. The genes were predicted using GeneMark.hmm eukaryotes 0.64 [38] from scaffolds of *Plasmodium falciparum and Plasmodium vivax.* The predicted genes were functionally annotated against NR database using blastX program of Blast suite. The functionally annotated genes were categorized in three different classes of Gene Ontology using Blast2GOsoftware (version 2.3.5) [22]. The predicted genes were mapped on pathways using KEGG Automatic Annotation Server (KAAS). While to study and correlate two parasites orthologous nature between the annotated genes was studied using OrthoMCL [39].

Results

The SRA entries from NCBI had 10,386,609 paired end raw reads for *Plasmodium falciparum and* 26,145,770 for *Plasmodium vivax*. The raw data was filtered for low quality, duplication, adapter content and high quality reads were further used (Table 1). The quality of reads was analysed using FastQC. Approximately there were 85% of the *Plasmodium falciparum* and 60% of the *Plasmodium vivax* high quality reads remaining after filtration. These reads were assembled and used for downstream analysis.

Table 1: *P. falciparum* and *P. vivax* Data Statistics

Plasmodium Species	No. of Raw reads	No. of bases	No. of HQ reads	No. of bases of HQ reads
P. falciparum	20773218	1246393080	14499220	867401652
P. vivax	52291540	5281440000	44152412	4310767982

The high quality reads were assembled in 8,161 and 7,797 scaffolds significantly high N50 of 3,109 and 7,811 for *Plasmodium falciparum and Plasmodium vivax* respectively with minimum scaffold of 500 bp (Table 2). The gene prediction was carried out on scaffolds with accurate hmm model.

Table 2: Whole genome assembly features and functional annotation of *P. falciparum* and *P. vivax.*

Assembly Features		
Description	***P. falciparum***	***P. vivax***
Total genome size(bp)	18039487	24399378
Number of scaffolds	8161	7797
Maximum scaffold length(bp)	44139	59371
Minimum scaffold length(bp)	500	500
Average scaffold length(bp)	2210	3129
Scaffold N50	3109	7811
GC (%)	42.42 %	43.38 %
Gene features and functional annotation		
No. of genes	6269	6400
Total gene length(bp)	4387404	6436785
Average gene length (bp)	699	1005
Genes with blast hits	5764	5149
Genes with non-blast hits	505	1251
GC (%)	44.78 %	45.42 %

The predicted genes were aligned against NCBI NR database using BlastX, functionally annotating genes from both the species. The best hit was selected from the blast results based on e-value of 1e-03 and highest alignment score. The best hits of maximum genes shared highest homology with *Plasmodium falciparum* Dd2.

Gene ontology study was carried out using Blast2GO software. In *Plasmodium falciparum* 2,896, 2,437 and 2406 genes were involved in cellular process, molecular function, and biological process respectively. Whereas for *Plasmodium vivax* 1,918, 2,061 and 1486 genes were involved in cellular process, molecular function and biological process respectively. The distribution of GO terms is represented in (figure 1) using WEGO.

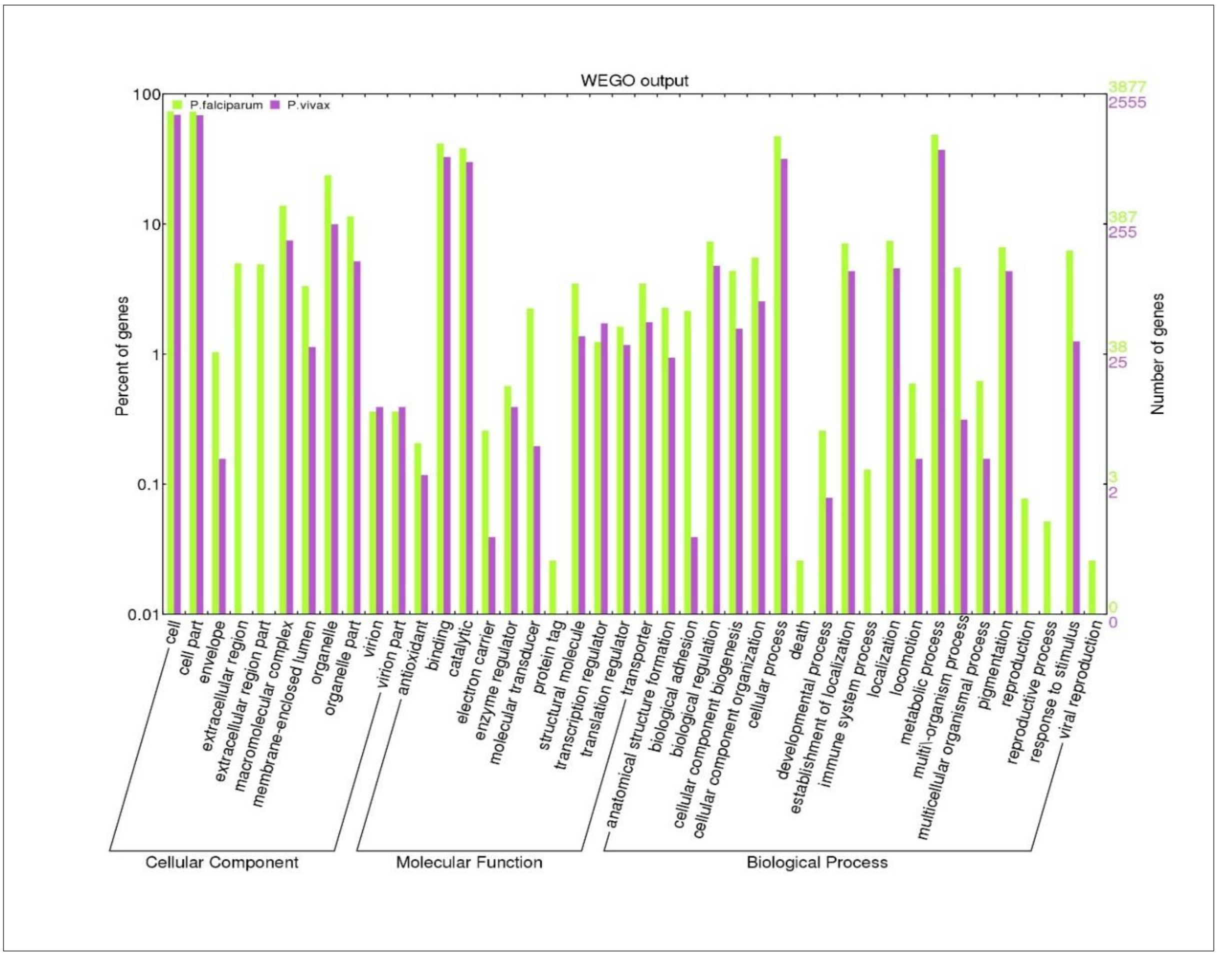

Fig 1: Comparison of GO classification for *Plasmodium falciparum* (green) and *Plasmodium vivax* (purple) with NR database. It is divided into three subcategories as cellular component, molecular function, and biological process.

There was significant number of genes participating in pathways. The maximum genes from both the species were involved in purine metabolism pathway followed by thiamine metabolic pathway. Effective response was found in a relation with T-cell receptor signalling pathway which involves in gene expression and regulation (figure 2).

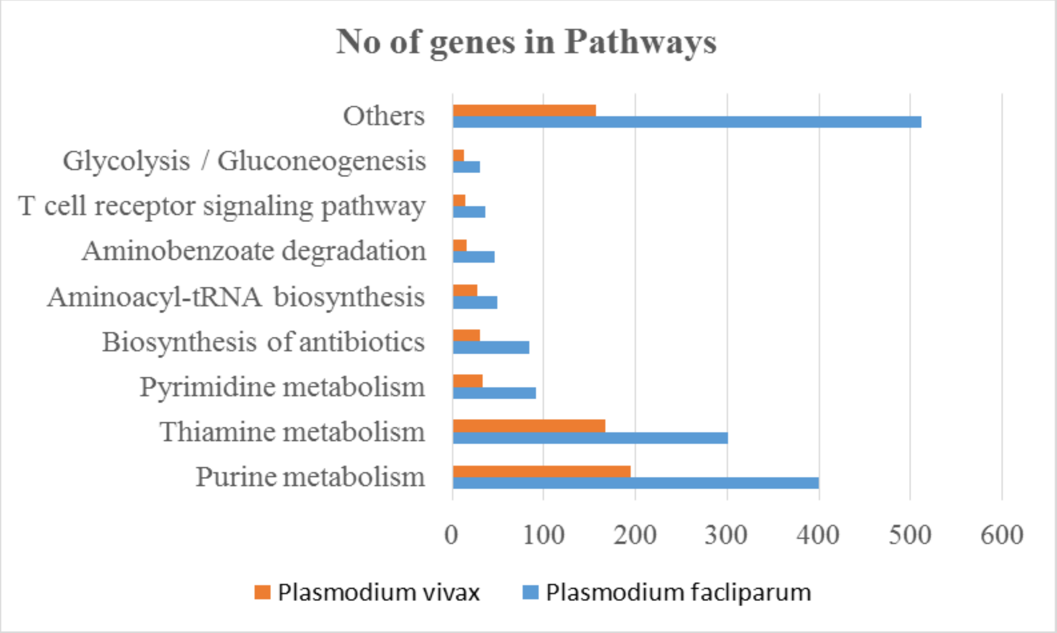

Fig 2: No of genes in pathways for both sps.

Orthologus genes were found using OrthoMCL.In *Plasmodium falciparum* 126 uniq groups and *Plasmodium vivax* 181 groups were found.The results are shown in (Table 3 and Figure 3).

Table 3: OrthoMCL result statistics

Parameters	No. of genes
Total Genes in *Plasmodium falciparum*	6269
Total Genes in *Plasmodium vivax*	6400
Total Genes in Groups	1721
Total Unique Groups in *Plasmodium falciparum*	186
Total Unique Groups in *Plasmodium vivax*	121

Ortho MCL study includes number of genes involved in total groups that is 1721 of P.falciparum and P.vivax and Unique groups in both the species.

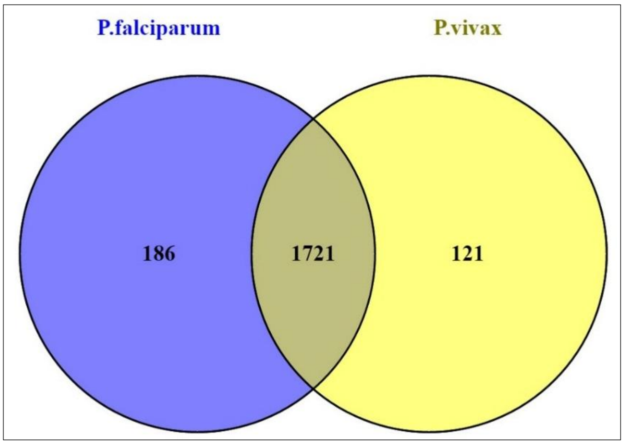

Fig 3: Ortho MCL result shows the unique groups between *P. falciparum* and *P. vivax* that is 186 and 121 respectively and number of common groups in both the species are 1721.

Discussion

Malaria is a lethal disease and its causative parasites known as *plasmodium* which is being very resistant to antibiotics, so there is need to control this disease. Researchers are working hard on methods for prevention of malarial infection, advance diagnosis and treatment, with just one malaria vaccine close to being licensed so far. Fast progress in DNA sequencing technology has made for a substantial decrease in cost and a consequential increase in throughput and accuracy. With more and more organisms being sequenced, a wave of genetic bioinformatics tools basic introduction of microbial tools data is drowning the world every day. Progress in genomics has been moving towards transformation in sequencing technology. It is now within reach for individual research groups in the eco-evolutionary and conservation community to generate *de novo* draft genome sequences for any organism of choice. Because of the cost and considerable effort involved in such an endeavour. Researchers developed a novel method known as *de novo* genome assembly by analysing sequencing data from high-throughput short read sequencing technology. At a fraction of the traditional cost and without using reference sequence assembly of large genomes can be done. The combination method of WGS (whole-genome sequencing) analysis and functional annotation for the comparison between two plasmodium species *Plasmodium falciparum* and *Plasmodium vivax* is done. Functional annotation plays an important role with Whole genome analysis for the study of gene function and various pathway related information. Total 36 genes from *Plasmodium falciparum* and 13 genes from *Plasmodium vivax* showed relation with T-cell receptor signalling pathway. In pathway analysis genes involved in T cell receptor signalling pathway plays an important role in T cell receptor (TCR) signal transduction is initiated by the recognition of cognate peptide–MHC molecules. The LAT signalosome propagates signal branching to three major signalling pathways, the Ca2+, the mitogen-activated protein kinase (MAPK) kinase and the nuclear factor-κB (NF-κB) signalling pathways, leading to the mobilization of transcription factors that are critical for gene expression and essential for T cell growth and differentiation. As a result of comparative analysis between *Plasmodium falciparum* and *Plasmodium vivax* orthologous unique groups were found for both the species. Moreover 186 and 121 unique group of orthologous gene signatures were found in *Plasmodium falciparum* and *Plasmodium vivax*. This functional genomics studies for *Plasmodium* spices can be further extended in future towards target based drug discovery for these human pathogen.

Conclusion

Comparative genome analysis helped us in getting deeper insights about the genetic composition of two parasites with the help of functional annotation at different level. Total 36 genes from *Plasmodium falciparum* and 13 genes from *Plasmodium vivax* showed relation with T-cell receptor signalling pathway which indicates a relation with response of the immune system. These data will be helpful for more studies on the structural and functional pathology of these pathogens. The advancement and unique gene signatures of all plasmodium species may be used to identify novel antigens and development of vaccines.

Conflict of interest

The authors have declared that no competing interests exist.

References

1. CDC. Laveran and the Discovery of the Malaria Parasite, accessed 23 March, 2016.
2. Weimin Liu, Yingying Li, Gerald Learn H, Rebecca Rudicell S, Joel Robertson D, Brandon Keele F *et al*. (23 September). Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. Nature. 2010; 467(7314):420-425.
3. Collins WE, Sullivan JS, Nace D, Williams T, Williams A, Barnwell JW. Observations on the sporozoite transmission of *Plasmodium vivax* to monkeys". J. Parasitol. 2008; 94(1):287-8. bioinformatics tools basic introduction of microbial tools
4. National Human Genome Research Institute (2010-11-08). A Brief Guide to Genomics. Genome.gov. Retrieved, 2011-12-03.
5. Concepts of genetics (10th Ed.). San Francisco: Pearson Education. 2012.ISBN 9780321724120
6. Pevsner J. Bioinformatics and functional genomics (2nd Ed.). Hoboken, N.J: Wiley-Blackwell, 2009. ISBN 9780470085851.
7. National Human Genome Research Institute (2010-11-08). FAQ about Genetic and Genomic Science. Genome.gov. Retrieved 2011-12-03.
8. Culver KW, Labow MA. (2002-11-08). Genomics. In Robinson R. Genetics. Macmillan Science Library. Macmillan Reference USA. ISBN 0028656067.
9. Mardis ER. A decade's perspective on DNA sequencing

technology. Nature, 2011; 470:198-203.

10. Metzker ML. Sequencing technologies - the next generation. Nature Review Genetics, 2010; 11:31-46.
11. Van El, Cornel CG, Borry MC, Hastings P, Fellmann RJ, Hodgson F *et al*. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. European Journal of Human Genetics. 2013; 21:1:S1-5.
12. Van El, Cornel CG, Borry MC, Hastings P, Fellmann RJ, Hodgson F *et al*. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. European Journal of Human Genetics. 2013; 21(1):S1-5.
13. Alberts Bruce, Johnson Alexander, Lewis Julian, Raff Martin, Roberts Keith, Walter Peter. "8". Molecular biology of the cell (5th Ed.). New York: Garland Science. 2008, 550. ISBN 0-8153-4106-7.
14. Genome data containing website: http://www.ncbi.nlm.nih.gov/sra Retrived 11-2-2016
15. SRA toolkit referred website: http://www.ncbi.nlm.nih.gov/books/NBK158900
16. FastQC : http://www.bioinformatics.babraham.ac.uk/projects/fastqc
17. Bolger M, Marc Lohse, Bjoern Usadel, Max Planck. Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany. 2 Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen, Germany. 3 Institut of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, and 52425 Jülich, Germany Trimmomatic: A flexible trimmer for Illumina Sequence Data Anthony.
18. Ana Conesa, Stefan Götz. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics PMCID: PMC2375974
19. Mark Borodovsky, Alex Lomsadze. Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES Unit–4.610 PMCID: PMC3204378
20. Howison M, Zapata F, Dunn CW. Toward a statistically explicit understanding of de novo sequence assembly. Bioinformatics. PubMed PMID: 24021385.
21. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. PubMed PMID: 22942022.
22. Song K, Ren J, Reinert G, Deng M, Waterman MS, Sun F. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. PubMed Central PMCID: PMC4017329.
23. Finotello F, Lavezzo E, Fontana P, Peruzzo D, Albiero A, Barzon L *et al*. Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. PubMed PMID: 22021898.
24. Ekblom R, Wolf JB. A field guide to whole-genome sequencing, assembly and annotation. PubMed PMID: 25553065; PubMed Central PMCID: PMC4231593.
25. Norgren RB Jr. Improving genome assemblies and annotations for nonhuman primates. PubMed PMID: 24174438; PubMed Central PMCID: PMC3814395.
26. Park ST, Kim J. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. PubMed PMID: 27915479; PubMed Central PMCID: PMC5169091.
27. Yalcin B, Adams DJ, Flint J, Keane TM. Next-generation sequencing of experimental mouse strains. PubMed PMID: 22772437; PubMed Central PMCID: PMC3463794.
28. Matochko WL, Derda R. Next-generation sequencing of phage-displayed peptide libraries. PubMed PMID: 25616338.
29. Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. PubMed PMID: 22676195; PubMed Central PMCID: PMC3960634.
30. El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: four stages of data processing and computational challenges. PubMed PMID: 24348224; PubMed Central PMCID: PMC3861042.
31. Malaria Report 2016: http://www.who.int/malaria/publications/world-malaria-report-2016/report/en/
32. SRA Data for *Plasmodium falciparum* : https://www.ncbi.nlm.nih.gov/sra/?term=SRR767807
33. SRA Data for *Plasmodium vivax* : https://www.ncbi.nlm.nih.gov/sra/?term=SRR1568115
34. White NJ. Assessment of the pharmacodynamic properties of antimalarial drugs *in vivo*. Antimicrob. Agents Chemother. 1997; 41(7):1413-22. PMC 163932. PMID 9210658
35. Douglas NM, Anstey NM, Angus BJ, Nosten F, Price RN. Artemisinin combination therapy for vivax malaria, 2010. PMC 3350863. PMID 20510281.
36. Winzeler EA, Manary MJ. Drug resistance genomics of the antimalarial drug artemisinin. Genome Biology, 2014. PMC 4283579. PMID 25470531.
37. Drugs immunity ‘may’ fail malaria fight. The Jakarta Post, April 23, 2011.
38. Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury Chernoff1, Mark Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm Nucleic Acids Research. 2005; 33:6494-6506.
39. Li Li, Christian J. Stoeckert Jr, David Roos S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 2003; 13:2178-2189.