HHS Public Access

Author manuscript

Mol Biochem Parasitol. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

Mol Biochem Parasitol. 2016 ; 210(1-2): 1–4. doi:10.1016/j.molbiopara.2016.07.005.

Distinct genomic architecture of *Plasmodium falciparum* populations from South Asia

Shiva Kumar[1], **Devaraja G. Mudeppa**[1], **Ambika Sharma**[1,2], **Anjali Mascarenhas**[1,2], **Rashmi Dash**[1,2], **Ligia Pereira**[1,2], **Riaz Basha Shaik**[1,2], **Jennifer N. Maki**[1], **John White III**[1], **Wenyun Zuo**[3], **Shripad Tuljapurkar**[3], **Manoj T. Duraisingh**[4], **Edwin Gomes**[2], **Laura Chery**[1], and **Pradipsinh K. Rathod**[1,*]

[1]Department of Chemistry, University of Washington, Seattle, WA 98195 USA

[2]Department of Medicine, Goa Medical College and Hospital, Bambolim, Goa 403202 India

[3]Department of Biology, Stanford University, Stanford, CA 94305 USA

[4]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA 02115 USA

Abstract

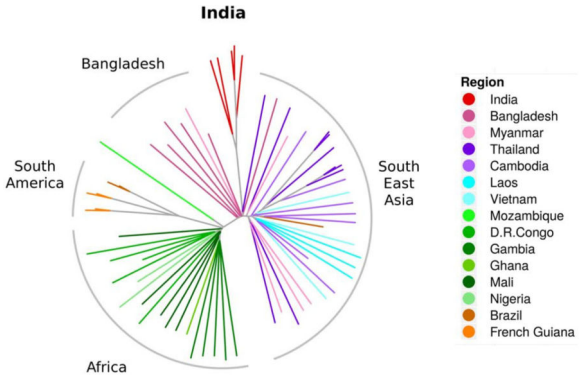
Previous whole genome comparisons of *Plasmodium falciparum* populations have not included collections from the Indian subcontinent, even though two million Indians contract malaria and about 50,000 die from the disease every year. Stratification of global parasites has revealed spatial relatedness of parasite genotypes on different continents. Here, genomic analysis was further improved to obtain country-level resolution by removing *var* genes and intergenic regions from distance calculations. *P. falciparum* genomes from India were found to be most closely related to each other. Their nearest neighbors were from Bangladesh and Myanmar, followed by Thailand. Samples from the rest of Southeast Asia, Africa and South America were increasingly more distant, demonstrating a high-resolution genomic-geographic continuum. Such genome stratification approaches will help monitor variations of malaria parasites within South Asia and future changes in parasite populations that may arise from in-country and cross-border migrations.

Graphical Abstract

Address Correspondence to. Pradipsinh K. Rathod, Department of Chemistry, University of Washington, Seattle, WA, USA 98195, rathod@chem.washington.edu, Fax: +1-206-685-8665 Telephone: +1-206-384-9404.

AUTHOR CONTRIBUTIONS

SK, DGM, JWIII, JNM, MD, LC and PKR conceived and designed the experiments. AS, AM, RD, LP and RBS performed the experiments. SK, DM, JNM, JWIII, WZ, ST, and PKR analyzed the data. SK, LC, and PKR wrote the manuscript. All authors reviewed the final manuscript.



Keywords

Malaria; Populations; Epidemiology; Molecular; Neighbors

The first genome-wide sequencing of pathogenic organisms, including the human malaria parasite *P. falciparum*, proved to be an unprecedented asset for malaria scientists [1]. Investigators, who previously studied individual phenotypes or individual genes and their products, could now study parasites on a genome scale [2–7]. Today, as the world considers global policies to control and eliminate malaria parasites, representative populations of parasites from distinct regions of the world are essential [8–10]. Collective genomes have the potential to capture variations in interactions of malaria parasites with diverse human hosts and mosquito vectors as well as susceptibilities of and resistance to therapeutics.

P. falciparum jumped from gorillas to humans in Central Africa in the distant past, but some of the most interesting differences among human *P. falciparum* parasites probably reflect evolutionary selection pressures resulting from more recent host history [11]. Human settlement into specific geographic regions and ecosystems, with unique climates, available plants and animals for food, and insect vectors, not only affected evolution of protective traits in humans, but also triggered new specialized traits within pathogens. The highly-intertwined genetic relationship between present-day humans, mosquitoes, and malaria parasites is captured in the varying architecture of parasite genomes from different parts of the world. Raw mapping of sequence reads between parasite populations has previously revealed distinct segregation of *P. falciparum* parasites from Africa, Southeast Asia and South America [12, 13]. While some continental separation could have been anticipated based on select markers from parasite genomes [14–16], the extent to which these principles held across the whole genome is remarkable [12, 13]. In the present study we demonstrate that it is possible to achieve country-level resolution of parasite relationships from whole genome sequences with judicious use of bioinformatics tools.

The global collection of more than 500 different assembled whole genome sequences of *P. falciparum* with known geographic history has gaps, particularly related to South Asia. This is a significant deficiency. There are more than 500 million people at risk for malaria in India and the country reports up to two million cases per year [17, 18]. Current estimates of deaths due to malaria in India reach approximately 50,000 per year [17, 18]. It is of great

biological interest and public health importance to know if present-day *P. falciparum* parasites from India are closely related to parasites from Southeast Asia, especially since India shares borders and has growing ties with many countries in Southeast Asia. Thailand, Vietnam, Laos and Cambodia have been under continual antimalarial drug pressure for decades and most drug resistance was first detected in these countries, including the current threat of artemisinin-resistance [19]. To understand future intermingling of parasite genomes, current similarities and differences in the genetic makeup of Indian and Southeast Asian parasites are of particular significance. It is also conceivable that Indian parasites could be closely related to parasites from Africa, given the history of travels over the last eight centuries as well as more recent trade and population exchanges [20].

The Malaria Evolution in South Asia (MESA) program seeks to understand how parasites in the field evolve not just against drugs, but also to gain other evolutionary advantages related to overcoming host immunity and improving transmission. The program runs under the broader US NIH International Centers of Excellence for Malaria Research (ICEMR) initiative [9] and so is a part of a formal government-sanctioned collaboration between India and the US [8, 21].

The clinical protocol guiding sample collections was approved by the institutional review boards of Goa Medical College and Hospital, University of Washington, US NIAID Division of Microbiology and Infectious Diseases, and Government of India Health Ministry Screening Committee. From April 2012 to December 2015 patients at the Goa Medical College (GMC) who were diagnosed with *Plasmodium falciparum* infection (by either rapid diagnostic test or by thin-smear microscopy) were referred to the MESA-ICEMR study team. Non-pregnant individuals between 12 months and 65 years old were given a written and oral description of the study and asked to provide written informed consent. Children between 8 and 18 years old were asked to provide assent in addition to the written consent of a parent or guardian required for all under 18. Study participants provided 4 – 6 mL of venous blood and parasite species was confirmed by RDT (FalciVax, Zephyr Biomedicals, India) and Giemsa-stained slide microscopy. Through December 2015, a total of 1088 *Plasmodium*-positive individuals, 228 of whom had *Plasmodium falciparum* mono-infection, were enrolled by the MESA-ICEMR at GMC. Of the malaria-positive patients enrolled at Goa Medical College and Hospital between 2012 and 2015, 88% were born in 31 Indian states other than Goa. Most of the enrolled patients were male (91%) and many were construction workers (51%) (29).

The present study provides a first glimpse of *P. falciparum* genomes from India though analysis of five Indian *P. falciparum* whole genome sequences collected between 2012 and 2015 by the MESA-ICEMR. Though the parasites samples included in the present study were collected in Goa, the parasites likely had diverse origins, including the states of Goa, Uttar Pradesh, Bihar, and Assam, based on study participant's reported place of birth and travel history one month prior to sample collection (Table 1). For comparison of relatedness to parasites around the world, whole genome sequencing (WGS) of malarial parasites from global field isolates delivers a comprehensive set of variations present in parasite populations. This, in turn, can provide an accurate approximation of associations between various genome samples [12,13]. In such population studies, with more than 5,000 genes in

each malaria parasite, the relationships between samples are complex. This complexity makes it useful to compare global variation and relatedness between parasites mathematically by imagining the distinct samples as occupying locations in an abstract space (usually in n-1 dimensions, where n is the number of samples under consideration). Principal Component Analysis (PCA) plots identify components, linear combinations of these dimensions, that describe the variation between the samples. In this case and typically, a few principal components capture most of the variation, thus describing the separation of samples using just a few important dimensions. This approach provides a parsimonious view of the stratification present in a set of samples. Technical details of the sequences of the parasites from India, the gathering of non-Indian genome sequences, and the final genomic comparisons are presented in the legend to Figure 1.

Together with genomes from around the world, the relative position of the Indian parasites in the world-map of genetic interconnectedness was identified (Figure 1). The layers of stratification of Indian samples in relation to parasites from other parts of the world was also estimated. Figure 1 shows that, as expected, global isolates fall into different groups based on their geographical origins [12, 13]. In this stratification, Indian isolates (red) clearly segregate into a distinct cluster, different from isolates of Southeast Asia (blue and purple) and even Bangladesh and Myanmar (dark and light pink). A neighbor joining tree, obtained using 'nj' from 'ape' package in R and made from the same distance matrix as that used for PCA, shows that the common ancestor for isolates from Africa and Asia are well-separated (Figure 2). Indian samples (red), lie between those from Bangladesh and Myanmar (pink), slightly removed from samples from Thailand (dark blue) and more removed from those of far Southeast Asian countries of Cambodia, Laos, and Vietnam (lighter blue).

Investigators in South Asia and beyond should benefit from access to the present whole genome sequences of *P. falciparum* from India. The Indian subcontinent has very diverse human genetics, ecosystems, and mosquito populations [21], so future parasite genome collections by the MESA-ICEMR will help capture an even more complete geographic representation of Indian *P. falciparum* parasites. Parasites captured directly from the states of origin of the construction workers in Goa will be of particular interest and will help determine if there is even greater diversity of parasites in India than what we capture from Southwest India. The present collection will also serve as a reference point in time. As the social and economic interactions between the countries of South Asia and Southeast Asia increase, it will be important to track the extent to which parasite genetic structures change, or even merge, over time. Genome-wide analysis of *P. falciparum* SNPs may also help track newly arrived ‘interlopers’ across borders with migrant workers, and possible stable recombinants emerging after mating with local parasite populations in India. Beyond Asia, the parasite genome sequences from India should provide valuable context to interpret population differences amongst global collections of fully-assembled whole genome sequences of *P. falciparum*, especially those from Africa. Such large data should help predict cross-continental efficacy of newly emerging drugs, vaccines, and vector-control measures as well as to track the spread of potential resistance traits across continents.

Acknowledgments

The authors thank all of the study participants and clinical research staff at the Goa Medical College and Hospital who assisted with this work. The authors also gratefully acknowledge Goa Medical College Dean, Dr. Pradeep Naik, and Medical Superintendent, Dr. Sunanda Amonkar, for facilitation and support of the research study. Dr. Anju Verma (The Rotary Blood Bank, New Delhi, India) provided human RBCs and human plasma for parasite culture. This work was part of the 'Malaria Evolution in South Asia' Program Project, an International Center of Excellence for Malaria Research (ICEMR) supported by US NIH/NIAID agreement U19 AI089688 (Program Director, PKR). The authors are most grateful for the administrative and scientific guidance provided by the MESA-ICEMR Scientific Advisory Group, particularly Dr. David Sibley on genomics, and the Government of India representatives Dr. Neena Valecha, Dr. Rashmi Arora, Dr. P.L. Joshi and Dr. Shiv Lal, and US NIH Program Officer Dr. Malla Rao.

References

1. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002; 419:498–511. [PubMed: 12368864]
2. Doolan DL, Apte SH, Proietti C. Genome-based vaccine design: the promise for malaria and other infectious diseases. Int J Parasitol. 2014; 44:901–913. [PubMed: 25196370]
3. Volkman SK, Neafsey DE, Schaffner SF, Park DJ, Wirth DF. Harnessing genomics and genome biology to understand malaria biology. Nat Rev Genet. 2012; 13:315–328. [PubMed: 22495435]
4. Fan E, Baker D, Fields S, Gelb MH, Buckner FS, Van Voorhis WC, et al. Structural genomics of pathogenic protozoa: an overview. Methods Mol Biol. 2008; 426:497–513. [PubMed: 18542886]
5. Kooij TW, Janse CJ, Waters AP. Plasmodium post-genomics: better the bug you know? Nat Rev Microbiol. 2006; 4:344–357. [PubMed: 16582929]
6. Llinas M, DeRisi JL. Pernicious plans revealed: *Plasmodium falciparum* genome wide expression analysis. Curr Opin Microbiol. 2004; 7:382–387. [PubMed: 15358256]
7. Rathod PK, Ganesan K, Hayward RE, Bozdech Z, DeRisi JL. DNA microarrays for malaria. Trends Parasitol. 2002; 18:39–45. [PubMed: 11850013]
8. Narayanasamy K, Chery L, Basu A, Duraisingh MT, Escalante A, Fowble J, et al. Malaria evolution in South Asia: knowledge for control and elimination. Acta Trop. 2012; 121:256–266. [PubMed: 22266213]
9. Rao MR. International Centers of Excellence for Malaria Research. The American Journal of Tropical Medicine and Hygiene. 2015; 93:1–4.
10. Carlton JM, Volkman SK, Uplekar S, Hupalo DN, Pereira Alves JM, Cui L, et al. Population Genetics, Evolutionary Genomics, and Genome-Wide Studies of Malaria: A View Across the International Centers of Excellence for Malaria Research. Am J Trop Med Hyg. 2015; 93:87–98.
11. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014; 15:379–393. [PubMed: 24776769]
12. Miotto O, Almagro-Garcia J, Manske M, Macinnis B, Campino S, Rockett KA, et al. Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. Nat Genet. 2013; 45:648–655. [PubMed: 23624527]
13. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, Amaratunga C, et al. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. Nat Genet. 2015; 47:226–234. [PubMed: 25599401]
14. Ferdig MT, Su XZ. Microsatellite markers and genetic mapping in *Plasmodium falciparum*. Parasitol Today. 2000; 16:307–312. [PubMed: 10858651]
15. Anderson TJ. Mapping drug resistance genes in *Plasmodium falciparum* by genome-wide association. Curr Drug Targets Infect Disord. 2004; 4:65–78. [PubMed: 15032635]
16. Su XZ, Wootton JC. Genetic mapping in the human malaria parasite *Plasmodium falciparum*. Mol Microbiol. 2004; 53:1573–1582. [PubMed: 15341640]
17. Kumar A, Valecha N, Jain T, Dash AP. Burden of malaria in India: retrospective and prospective view. Am J Trop Med Hyg. 2007; 77:69–78. [PubMed: 18165477]

18. Murray CJ, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, et al. Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet. 2012; 379:413–431. [PubMed: 22305225]
19. Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, Tarning J, et al. Artemisinin resistance in *Plasmodium falciparum* malaria. N Engl J Med. 2009; 361:455–467. [PubMed: 19641202]
20. Faulde MK, Rueda LM, Khaireh BA. First record of the Asian malaria vector *Anopheles stephensi* and its possible role in the resurgence of malaria in Djibouti, Horn of Africa. Acta Trop. 2014; 139:39–43. [PubMed: 25004439]
21. Kumar A, Chery L, Biswas C, Dubhashi N, Dutta P, Dua VK, et al. Malaria in South Asia: prevalence and control. Acta Trop. 2012; 121:246–255. [PubMed: 22248528]
22. Llinas M, Deitsch KW, Voss TS. Plasmodium gene regulation: far more to factor in. Trends Parasitol. 2008; 24:551–556. [PubMed: 18929512]
23. Guizetti J, Scherf A. Silence, activate, poise and switch! Mechanisms of antigenic variation in *Plasmodium falciparum*. Cell Microbiol. 2013; 15:718–726. [PubMed: 23351305]
24. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011; 2011:17.
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]
27. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27:2987–2993. [PubMed: 21903627]
28. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2009.
29. Chery, et al. Demographic and clinical profiles of *Plasmodium falciparum* and *Plasmodium vivax* patients at a government tertiary care centre in southwestern India. Submitted.

Highlights

- *Plasmodium falciparum* genomes from India are underrepresented in global collections
- Five distinct *Plasmodium falciparum* genomes from India have been sequenced
- Compared to global parasites, *Plasmodium falciparum* from India are distinct
- Indian parasites are most closely related to those from Bangladesh and Thailand

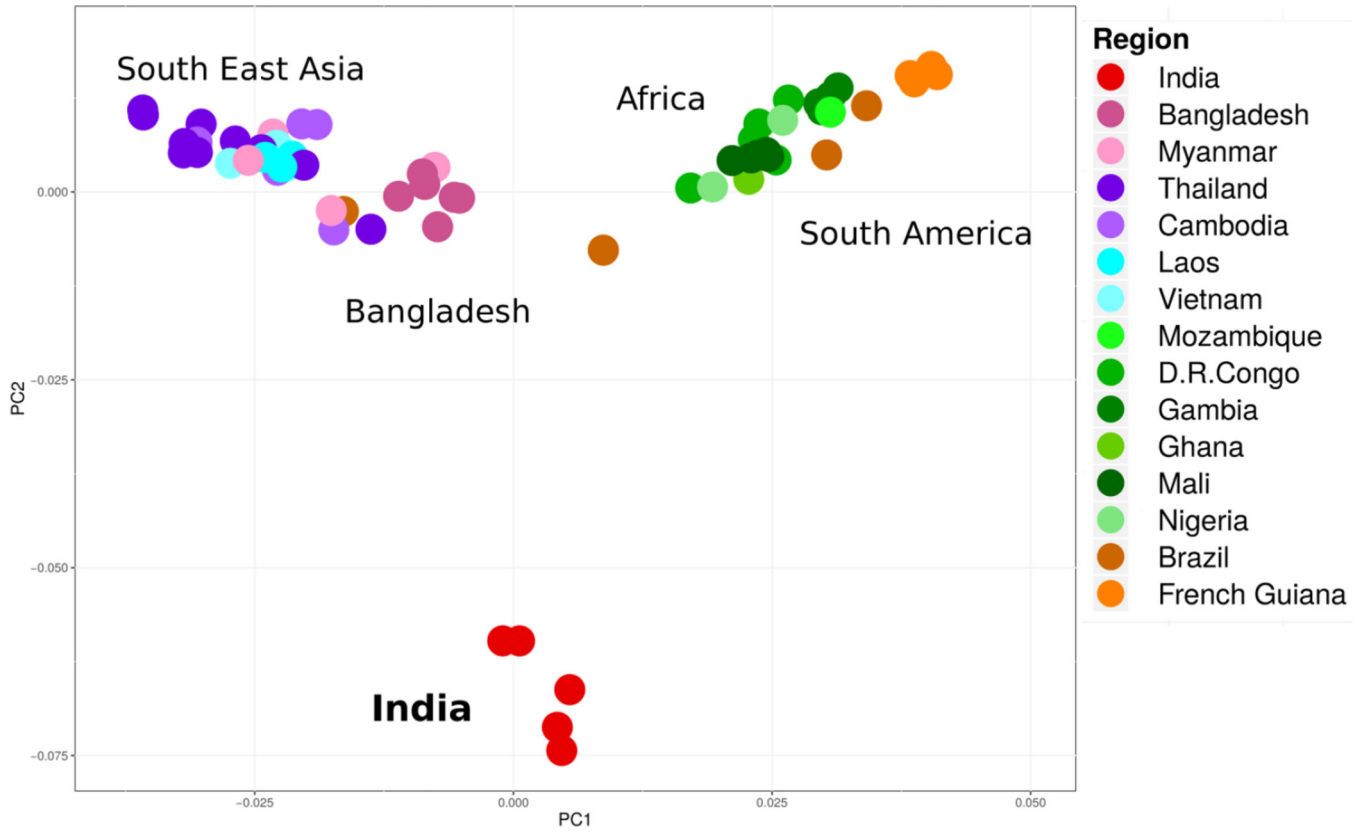


Figure 1.

Positioning of *P. falciparum* genomes from India amongst global isolates of the human malaria parasites using Principal Component Analysis (PCA). Whole-genome sequences of non-Indian isolates from different parts of the world were downloaded from European Nucleotide Archive (ENA; accession numbers given below). Reads from each WGS sample were filtered for quality using Trim Galore [24]. Cutoffs of 70bp for minimal read length and quality score of 28 were used to obtain high quality reads. Reads were then prepared according to 'pre-processing' step of best-practices guidelines from Broad Institute. Specifically, BWA mem algorithm was used to align reads to reference sequence obtained from PlasmoDB (Pf v9) [25]. Reads were then sorted and duplicates were marked using Picard tools 'SortSam' and 'MarkDuplicates', respectively. Indels were then realigned and base quality scores recalibrated using GATK tools [26]. Variants were called using 'Mpileup' from Samtools for all the samples taken together [27]. All analysis was done on an SGE cluster of EC2 instances from Amazon web services (AWS) using starcluster (http://star.mit.edu/cluster/). A single vcf (variant call format) file of all variants was used to calculate pair-wise nucleotide distance. Distances between sample i and sample j were calculated as $(((DV_i/DP_i) - (DV_j/DP_j))^2)^{1/2}$, where DV_i and DV_j are number of variant reads at a given position in the genome of sample i and j, respectively. DP_i and DP_j are the depth of coverage at the corresponding position in sample i and j, respectively. The values were summed for all positions and normalized for varying depths in samples. This was achieved by dividing the raw distance score by the number of positions that passed the criterion for a distance calculation. Criterion for distance calculation was given as a position where depth

was 5 or greater in both samples i and j. Pairwise distance matrix obtained in this manner was used for classical multi-dimensional scaling (MDS) in R using 'cmdscale'. MDS was applied to distances (a quantitative measure of dissimilarity, often arising from our biological understanding) between pairs of samples and is equivalent to PCA under the condition that these distances are Euclidian. Spatial matrix obtained from cmdscale was used to make plots using 'ggplot2' [28]. Regions and countries represented in this plot (and reference numbers for the genomes) are as follows: **South America**: Brazil (ERR012881, ERR012888, ERR023690, SRR530164 and SRR629057) and French Guiana (SRR834923, SRR834924, SRR834925 and SRR834926). **Africa**: Gambia (ERR015439, ERR018904, ERR019543, ERR019548 and ERR020108), Mali (SRR1011071, SRR1011167, SRR1011205, SRR1011289 and SRR1210129), Ghana (ERR015396), Nigeria (ERR426048 and ERR426142), D.R. Congo (ERR404207, ERR404242, ERR426005, ERR426127 and ERR426136) and Mozambique (SRR1029755). **India** (2NRN, 5TBX, JZT6, TCFT and XGK2). **South East Asia**: Bangladesh (ERR404184, ERR404203, ERR404229, ERR404257, ERR426068 and ERR426133), Central Myanmar (ERR175486, ERR175510, ERR175521, ERR404208 and ERR426076), Thailand (ERR164693, ERR164704, ERR164710, ERR216522, ERR221495, ERR337557, ERR337564, ERR404221, ERR404248 and ERR426140), Cambodia (ERR039180, ERR042214, ERR063632, ERR067573, ERR171601, ERR171634 and ERR216553), Laos (ERR216499, ERR221479 and ERR388787) and Vietnam (ERR180075, ERR180092 and ERR426013). The raw sequence reads for the five parasites from five South Asian patients are now in the NCBI-SRA database. SRA accession for the study is: SRP075579 and the run accession numbers for the 5 samples are 1) xgk2: SRR3575059, 2) hxjy: SRR3575060, 3) durf: SRR3575061, 4) 5tbx: SRR3575062, and 5) gabv: SRR3575063. SNP variants emerging from the present parasite genome sequences will be available on the PlasmoDB web-site (http://plasmodb.org/plasmo/).

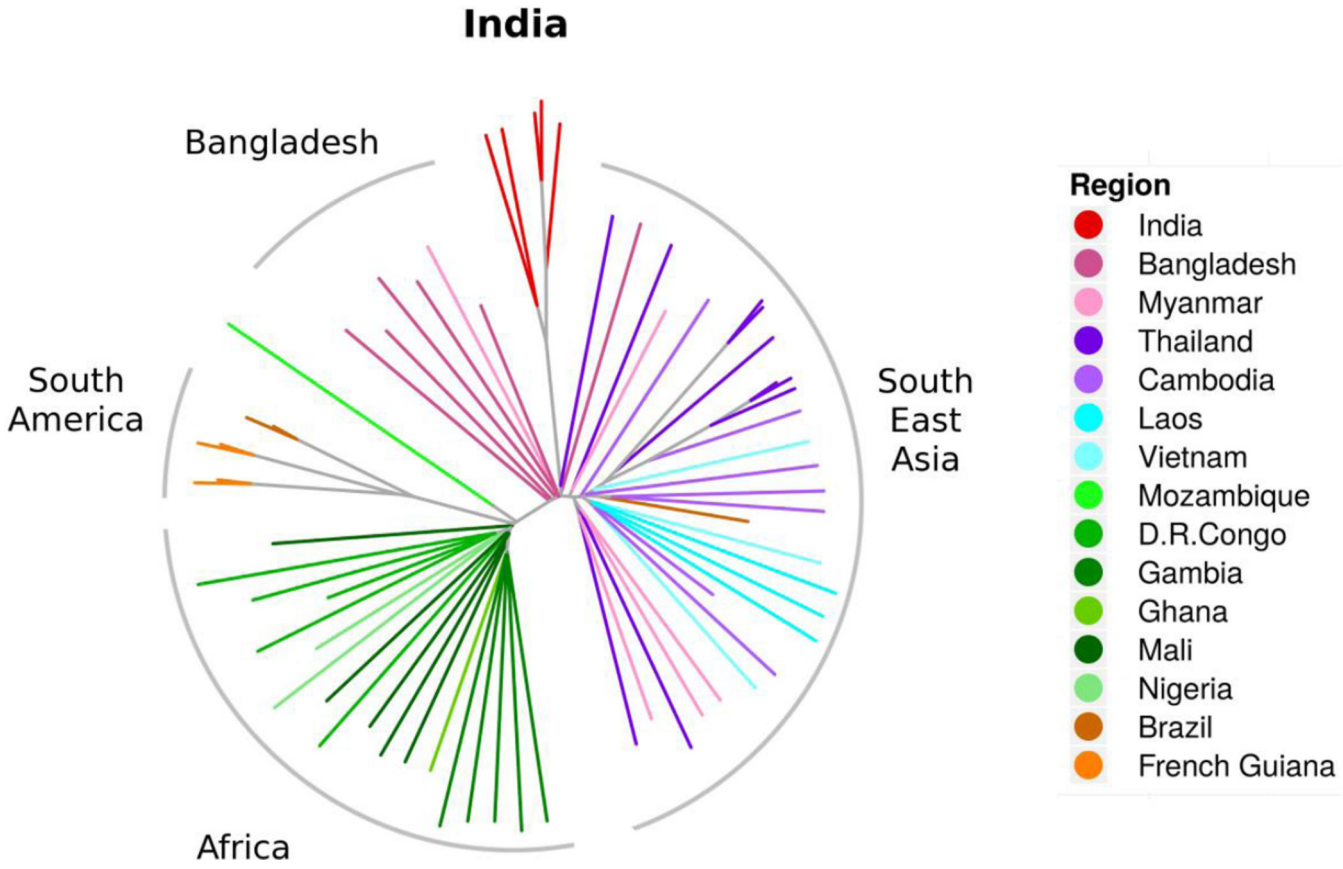


Figure 2.

Positioning of *P. falciparum* genomes from India amongst global isolates of the human malaria parasites using un-rooted neighbor joining tree, *excluding var* genes and non-coding sequences. Non-coding sequences and variable surface protein genes were removed and the distance analysis was based on coding sequences where basic cellular metabolism and physiology functions must reside. This grouping reflected true geographic origins of the parasite samples. African samples showed preferential joining of parasites from Congo and from Gambia. Parasites from West and Central Africa were well-separated from the Mozambique parasite. In Southeast Asia, there was a continual separation of parasites from Bangladesh to Myanmar/Thailand, to Cambodia, and to Laos and Vietnam. Parasites in Thailand were less tightly concentrated, but may reflect the movement of patients and parasites in and out of this country. In this context, the Indian *P. falciparum* isolates remained cleanly segregated from Southeast Asian, African, South American and Bangladesh lines. The continental clusters were well-defined with minimum ambiguity. In an earlier effort, we performed distance analysis using whole genomes (including var genes and intergenic regions) superimposed on the *P. falciparum* 3D7 cell line. This generated some uncertain relationships between isolates and their geographic origins (data not shown). We hypothesize that the ancestral genomic relationships of the parasites may be influenced by subsets of highly-variable parasite genes formed based on interactions with patient immune responses. Genes coding for surface proteins on infected erythrocytes are under heavy selective pressure and reshuffle amongst themselves at very high rates and

acquire point mutations [22, 23]. In addition, non-coding DNA spread across the chromosome forms a substantial part of the parasite genome and could also drift more compared to internal protein coding genes. As shown above, removal of var genes and intergenic regions provided a reliable picture of relationships between genomes and origin of samples.

Table 1

Malaria Patient data associated with sequenced genomes

Patient ID	**XGK2**	**5TBX**	**durf**	**hxjy**	**gabv**
Enrollment Date	Aug-12	Aug-12	Aug-12	Apr-13	Jun-15
State of Birth	Uttar Pradesh (UP)	Outside India	UP	Bihar	Assam
Travels	Maharashtra, UP	No/Goa	No/Goa	Tamil Nadu, Bihar	No/Goa
Temperature	97.8	99.7	104.8	103.8	100.6
% Parasitemia	3.6	0.2	1.9	2.64	2.04
Inpatient	Y	N	N	Y	N
Severe	Y	NA	NA	Y	NA