Genomics Data 9 (2016) 118–125

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

A cross strain *Plasmodium falciparum* microarray optimized for the transcriptome analysis of *Plasmodium falciparum* patient derived isolates

Amit Kumar Subudhi [a,1,2], P.A. Boopathi [a,1], Sheetal Middha [b], Jyoti Acharya [b], Sudha Narayana Rao [c], Raja C. Mugasimangalam [c], Paramendra Sirohi [b], Sanjay K. Kochar [b], Dhanpat K. Kochar [d], Ashis Das [a,*]

[a] Department of Biological Sciences, Birla Institute of Technology and Science (BITS), Pilani, Rajasthan, India
[b] Department of Medicine, S.P. Medical College, Bikaner, Rajasthan, India
[c] Genotypic Technology Pvt. Ltd., Bangalore, India
[d] Rajasthan University of Health Sciences, Jaipur, Rajasthan, India

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form 10 May 2016
Accepted 6 July 2016
Available online 20 July 2016

Keywords:
Plasmodium falciparum
Microarray
Uncomplicated malaria
Complicated malaria
PfEMP1
RIFIN

ABSTRACT

Malarial parasite *P. falciparum*, an apicomplexan protozoan has a 23.3 MB nuclear genome and encodes ~5600 transcripts. The genetic diversity of the parasite within and across geographical zones is a challenge to gene expression studies which are essential for understanding of disease process, outcome and developing markers for diagnostics and prognostics. Here, we describe the strategy involved in designing a custom *P. falciparum* 15K array using the Agilent platform and Genotypic's Right Design methodology to study the transcriptome of Indian field isolates for which genome sequence information is limited. The array contains probes representing genome sequences of two distinct geographical isolates (i.e. 3D7 and HB3) and sub-telomeric *var* gene sequences of a third isolate (IT4) known to adhere in culture condition. Probes in the array have been selected based on their efficiency to detect transcripts through a 244K array experimentation. Array performance for the 15K array, was evaluated and validated using RNA materials from *P. falciparum* clinical isolates. A large percentage (91%) of the represented transcripts was detected from Indian *P. falciparum* patient isolates. Replicated probes and multiple probes representing the same gene showed perfect correlation between them suggesting good probe performance. Additional transcripts could be detected due to inclusion of unique probes representing HB3 strain transcripts. Variant surface antigen (VSA) transcripts were detected by optimized probes representing the VSA genes of three geographically distinct strains. The 15K cross strain *P. falciparum* array has shown good efficiency in detecting transcripts from *P. falciparum* parasite samples isolated from patients. The low parasite loads and presence of host RNA makes arrays a preferred platform for gene expression studies over RNA-Seq.

1. Introduction

Malaria imposes a significant burden on world health. There were an estimated 149–303 million cases of infections along with 0.48 million deaths reported worldwide in the year 2015 [39]. In the year 2014, 1.6 million confirmed cases were reported from the 10 malaria endemic South-East Asian countries, of which, India registered the highest number of confirmed cases (70%) [39]. *Plasmodium falciparum* is often considered the most virulent among the 5 *Plasmodium* species known to infect humans. Malaria elimination programs remain challenged due to the emergence and reemergence of drug resistant parasites and inability to develop an effective vaccine [11,14].

The failure to develop a successful vaccine and emergence of drug resistant parasites can partly be attributed to the complex life cycle of the parasite. Interestingly, many basic aspects of its biology, while in the host, still need to be explored. This has been partially achieved in recent years due to the advancement in techniques like microarray (DNA microarrays, Protein binding microarrays, SNP microarrays), array-comparative genomic hybridization (aCGH) [19], RNA-seq [26], ChIP-chip [9], Chip-seq [3], mass spectrometry techniques [25], transfection techniques [5], mutagenesis methods [2], gene knock out techniques [8,20] and imaging techniques. All these techniques have helped in exploring some fundamental questions, which was not possible earlier.

Abbreviations: PfENO, enolase (PF10_0155); PfCP, conserved hypothetical protein (PF14_0683); PfSec14, Sec 14 domain containing protein (PF1280w); PfROM3, rhomboid protease 3 (MAL8P1.16); PfGK, glycerol kinase (PF13_0269).

* Corresponding author.
E-mail addresses: amit4help@gmail.com (A.K. Subudhi), boopathiarunachalama@gmail.com (P.A. Boopathi), sheetumig21@gmail.com (S. Middha), jyotiacharya2@gmail.com (J. Acharya), Sudha.rao@genotupic.co.in (S.N. Rao), raja@genotypic.co.in (R.C. Mugasimangalam), drpsirohi@gmail.com (P. Sirohi), drskkochar@rediffmail.com (S.K. Kochar), drdkkochar@yahoo.com (D.K. Kochar), ashisd28@gmail.com, adas@pilani.bits-pilani.ac.in (A. Das).
[1] These authors contributed equally to this work.
[2] Current address: Pathogen Genomics Group, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

http://dx.doi.org/10.1016/j.gdata.2016.07.006

The evolution of *P. falciparum* DNA microarrays has started few years before the release of its first sequence draft using PCR amplified inserts from a *P. falciparum* DNA library (shotgun microarray) or cDNAs [12,21]. Following the release of the first draft of *P. falciparum* 3D7 genome sequence from Malaria Genome Consortium (The Sanger center *Plasmodium falciparum* Genome Project, Stanford Genome Technology Malaria Genome Project, TIGR *Plasmodium falciparum* Genome Database), many oligonucleotide based whole genome microarrays has been designed, Importantly high density short oligonucleotide (25mer) microarray on Affymetrix platform came from Winzeler lab [18] and long oligonucleotide (70mer) spotted microarray came from DeRisi lab [4]. Upon release of subsequent updates of *P. falciparum* genome database annotation, many new versions of microarrays were evolved and utilized various platforms (To see the comprehensive list of *P. falciparum* microarrays designed till date, please refer to NCBI's Gene Expression Omnibus website http://www.ncbi.nlm.nih.gov/geo/.

Use of microarrays has provided new avenues for understanding expression and regulation of genes at a genome-wide level. DNA microarrays have been utilized to show the stage specific temporal expression pattern of the parasite during its intra-erythrocytic developmental cycle (IDC) [4,18,21]. Other microarray based studies have also helped to explore the transcriptional status of genes in other life cycle stages like gametocyte and sporozoite stages, differences and commonality of transcriptional programming at a cross strain level [19], transcripts possibly involved in gametocytogenesis [32,41], in vivo transcriptional states of the parasite in infected patients [6], soft-wired transcriptional reprogramming in response to drug perturbations [23,36] or environmental cues [24] and expression pattern of variant surface antigens particularly in clinical conditions [37]. Although the microarray technique has helped in addressing many fundamental questions, its use has been limited due to its cost and comparatively large amount of starting biological material required which is a limitation in the case of clinical isolates. Most of the arrays for *P. falciparum* have been designed based on a single reference strain genome (3D7 laboratory strain) [4,13,18]. These arrays have not considered the possibilities of genomic variations among strains that could limit transcript detection.

A cross strain *P. falciparum* custom 15K microarray has been designed in a platform, which provides in many ways the flexibility to design a custom array. This extends from ordering the number of slides, allowing updation of the microarray design as and when required. Further, due to the cRNA amplification strategy adopted, the amount of starting material required makes studies with field derived samples more feasible. The use of ink-jet based in situ synthesized 60mer long oligonucleotide probes in the selected platform enhances its sensitivity and due to outstanding spot uniformity, it reduces the inter-array variability. It has a broad dynamic range of signal intensity and shows low and uniform background signal [16]. Probes in our 15K cross-strain array represent the genomes of two geographically distinct strains i.e. 3D7 and HB3 and the *var* gene sequences of a third strain (IT4). Probes have been validated using RNA material from Indian *P. falciparum* clinical isolates in a 244K array format [33] and best probes were chosen for the 15K array design. The pre-probe selection experiment helped in identification of probes for efficient detection of transcript sequences of field isolates. Here, we report, the further validation of our cross strain *P. falciparum* custom 15K arrays using 13 clinical isolates from patients with uncomplicated/complicated symptoms due to *P. falciparum* infection.

2. Results and discussion

This study intends to provide validation of data from the custom cross strain microarray. The overall examination of the microarray data is presented as a analysis to validate the strength and accuracy of the array design and procedures.

2.1. Probe selection for 15K Plasmodium falciparum cross strain whole genome GXP microarray

To screen best probes for transcriptome studies of *Plasmodium falciparum* Indian field isolates for which genome sequence information is limited, a 244K custom whole genome array was designed on an Agilent platform with 60mer oligonucleotide probes (Agilent Microarray Design Identity number (AMADID) 024956). Probes represent *P. falciparum* 3D7 strain transcript sequences (PlasmoDB v5.3) [1] and NCBI EST sequences (NCBI 2007). Details about the 244K array can be found in our previous report [33].

Probe selection for 15K array was carried out by considering only sense probes representing PlasmoDB transcripts. GE gMedianSignal (green channel) and GE_rMedianSignal intensity (red channel) for each probe was collected. Maximum value of the GE green and red channel intensity of each probe representing a transcript was calculated, ranked and probe with maximum GE_Median signal intensity selected for each transcript. Probes with signal intensity >96 (~twice the average background intensity) were taken for consideration (3025 probes). Transcripts for which representing probes did not qualify the criteria of showing intensity >96, probes were selected based on their proximity to the 3′ end of the representing sequences. In this way, 5629 probes representing 3D7 transcript sequences were collected from the 244K array.

2.2. Plasmodium falciparum 15K cross strain whole genome GXP microarray design

We also designed and included probes against sequences of two other geographically distinct sequenced *P. falciparum* strains i.e. HB3 and IT4 so as to represent diverse sequences of *P. falciparum* in the array. To design a cross strain *P. falciparum* array, transcript sequences were collected from PlasmoDB v6.3 for 3D7 strain (5595 transcript sequences), from Broad Institute for HB3 strain (5623 transcript sequences) and from NCBI for IT4 strain (80 *var* gene transcript sequences) and a unique database was made. Total number of sequences thus collected were 11,298 out of which 9842 were unique sequences (Although sequence similarity was observed between them) and 1427 transcripts sequences of HB3 were identical to 3D7 transcripts sequences (determined from BLAST results). Probes collected from the 244K array experiment were BLASTed against the database and probes having only single significant hit with any of the strain were selected. New probes were designed for the transcripts for which no probes were found, using Agilent eArray tool. In the case of transcripts for which specific probes could not be designed, probes having minimum number of hits were selected (probes with a potential to cross hybridize) and these were flagged. Criterion for significant hit was alignment of 30 bp or more with >84% identity with the transcript sequence. Although, the criteria for significant hit was as mentioned above, 90% of the probes showed 60 bp alignment with 100% identity to the transcript sequences. All probes, collected and designed were BLASTed against the human transcriptome to check for their cross hybridizing potential with the human transcripts. Potential cross hybridizing probes were removed from the array design. A total of 6362 user defined long oligonucleotide (60mer) probes were designed against the transcript sequences present in the database. Based on the BLAST results, probes were annotated as specific to any one strain (Single significant hit within that strain-3D7 specific/HB3 specific/IT4 specific), common in between any two or more strain (Single significant hit within each strain-3D7 and HB3 or 3D7 and IT4 or HB3 and IT4 or 3D7, HB3 and IT4) and cross hybridizing probes (multiple significant hits within a strain). Array was re-annotated recently using the PlasmoDB v8.2. After the re-annotation, the final design of the *P. falciparum* cross strain whole genome GXP array (AMADID: 024956) is summarized in Supplementary Table S1 and S2. Out of 6362 probes present in the array, 6120 probes can be used against the transcripts of any one of the strain, 194