INFO
Corresponding Author:
Ritu Gill, Centre for Biotechnology, Maharsh Dayanand University, Rohtak, Haryana, India.
E-mail Id:
rgilcbt@gmail.com
Orcid Id:
https://orcid.org/0000-0002-8959-3089
How to cite this article:
Date of Submission: 2019-12-15
Date of Acceptance: 2019-01-03

ABSTRACT
Histones N-terminal tails are the sites for Post-Translational Modifications (PTMs) that regulate the chromatin structure, thus chromatin associated processes. PTMs include methylation, acetylation, phosphorylation, ubiquitination, sumoylation, and ribosylation. Histone lysine methylation is associated with both transcription activation and repression. The SET domain proteins carry out the histone lysine methylation on the N-terminal tails of histones H3 and H4 and are called Histone Lysine Methyltransferases (HKMTs). A total of ten SET domain genes have been identified in human malarial parasite Plasmodium falciparum. The present study provides detailed computational analysis of PfSET domain family in terms of domain composition, physiochemical properties, subcellular localization, expression profiling and phylogenetic relationships. The work also highlights the conservation of important catalytic residues in PfSETs. The present study provides a detailed insight into the PfSET family, thus opens a platform for further developments.

Keywords: Malaria, Plasmodium falciparum, SET domain, Histone Lysine Methyltransferases

Introduction
Despite significant reduction in malaria incidences since 2010, malaria continues to be a major health concern affecting millions of people in tropical and subtropical regions. Malaria parasite life cycle occurs in two different hosts- a vertebrate and an invertebrate. The pathogen exhibits differential gene expression to cope up with distinct environments it experiences during its life cycle. However, regulation of gene expression is hugely understudied in malaria parasite. Histones covalent modifications (methylation, acetylation, phosphorylation, ubiquitination, sumoylation, ribosylation etc.) carried by chromatin modifying proteins play a significant role in regulation of chromatin structure and gene expression. Chromatin mediated epigenetic regulation plays an important role in malaria parasite. Histone methylation is a widespread covalent modification that occurs on lysine and arginine residues in the N-terminal tails of histones H3 and H4. Histone lysine methylation is involved in a number of processes such as transcriptional regulation, heterochromatin formation, DNA damage response and cancer. Lysine methylation on the histone tails is carried out by a group of proteins containing SET domain. SET domain is approximately 130 amino acids long evolutionary conserved motif present in chromatin associated proteins from yeast to mammals. SET domain proteins act as histone lysine methyl transferases (HKMTs) that transfer methyl group from the cofactor S-adenosyl-L-methione (SAM) to lysine residues of the histone tails. PKMTs regulate
transcription and other cellular functions through site-specific methylation of histones and other substrates. The SET domain is named after its first identification in three Drosophila melanogaster proteins: Su(var)3-9; Suppressor of variegation 3-9, E(z); Enhancer of zeste and Trx; the trithorax-group. The differential gene expression patterns in different stages of Plasmodium life cycle are maintained by number of mechanisms, the most important are covalent histone modifications and three-dimensional chromatin conformation. Methylation marks in H3 and H4 histones are involved in regulation of the parasite erythrocytic cycle especially schizont stage. PfSET vs directed methylation of histone H3K36 results in the repression of all var genes and allows expression of one gene at a time. The mechanism of expression of only one out of the 60 var genes leads to the antigenic variation of the parasite and thus evasion of the host immune system. As SET proteins are involved in various vital cellular and biochemical functions, thus play important role in progression of P. falciparum through different developmental stages. Further, targeting these histone methylation writers PfSETs could be an important strategy in controlling the intractable parasite. Mining of SET genes from P. falciparum and their further bioinformatics analysis will extend our understanding of these proteins in P. falciparum life cycle that can be potential drug targets in future.

Materials and Methods

Identification of SET Domain Containing Genes in P. falciparum (PfSET genes)

Plasmodium genomic resource PlasmoDB version 9.0 (http://PlasmoDB.org) was searched to identify P. falciparum SET domain encoding genes (PfSET genes). Gene text search was primarily used to collect putative PfSET genes from PlasmoDB. Extraction of all PfSET genes was ensured by BLASTp analysis carried out with a threshold expect value of ≤10 using protein sequences of SET domain from various organisms (Drosophila melanogaster, Arabidopsis thaliana, Homo sapiens) as a query and low complexity filter turned off.

Confirmation of SET Domain in Putative Genes

To validate the presence of SET domain in proteins encoded by putative PfSET genes, SMART (http://smart.embl.de/), InterPro (http://www.ebi.ac.uk/interpro/) databases were explored. Protein sequences of each putative PfSET genes were obtained from PlasmoDB followed by confirmation of predicted SET domain by SMART/InterPro.

Analysis of PfSET Genes

All the information about PfSET genes regarding Gene Ids, chromosomal location, genomic position, and number of introns, nucleotide sequence length, molecular weight, amino acid sequence length and Isoelectric Point (IP) was extracted from PlasmoDB. Number of SET domains in a gene was identified by SMART.

Domain Architecture of PfSET Proteins

Domain structures of all PfSET proteins were identified with the SMART database. All the protein sequences of PfSETs were submitted to the SMART database one by one. Domain architecture of all PfSET proteins was drawn manually. Domains were named as identified by SMART.

Prediction of Subcellular Localization of SET Proteins

Various online softwares like Mitoprot (http://ihg.gsf.de/ihg/mitoprot.html), Euk-mPLoc 2.0 server (http://mloc.iro.umontreal.ca/mloc2/) were used to predict subcellular localization of PfSET proteins in P.falciparum. Mitoprot (http://ihg.gsf.de/ihg/mitoprot.html) is a web based computer program for predicting mitochondrion proteins. Any sequence showing mitochondria targeting probability equal to or greater than 0.4 was considered to be a mitochondrial protein. Euk-mLoc 2.0 server (http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/) is used for predicting sub-cellular localization of eukaryotic proteins. PATS (http://modlabcadd.ethz.ch/software/pats/) was used to identify the apicoplast targeted PfSETs.

Expression Profiling

Expression of all PfSET genes at different stages of life cycle of P. falciparum was analysed as per available transcriptome and proteome data at PlasmoDB. Proteomic data was retrieved from Florens et al., Lasonder et al., Le Roch et al., Silvestrini et al., Solyakov et al., Treeck et al., Oehring et al., Lindner et al., Pease et al. Transcriptome data of intraerythrocytic stages as provided by DeRisi group was analysed for existence of mRNA of PfSET genes. The heat maps for PfSETs were constructed using MeV software version 4.9.

Prediction of Human Homologs of PfSET Proteins

Homologs of PfSET proteins in Homo sapiens were predicted by BLASTp analysis of individual PfSET protein sequences against HPRD (Human Protein Reference Database). HPRD is freely available at www.hprd.org. The best hit was taken and further analysed to validate it as a homolog by its domain architecture (SMART and Pfam), functional annotation and protein size. Homologs list was prepared showing NCBI gene ids and product description against the respective PfSET gene.

Multiple Alignment Sequence and Identification of Conserved Features

Protein sequences of only SET domains for all PfSETs were
extracted as identified by SMART database. All the extracted SET domain sequences were aligned with CLUSTAL Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) using default parameters. SET specific conserved residues/signature motifs were identified based on multiple sequence alignment and highlighted manually.

**Generation of Sequence Logo**

To represent level of conservation of specific residues at particular positions in the PfSET domains, sequence logo for PfSET domain was generated using online software WebLogo (http://weblogo.berkeley.edu).

**Formation of Phylogenetic Tree**

The evolutionary connections between organisms are represented graphically through phylogenetic trees. Multiple sequence alignment of protein sequences was performed using the ClustalW program. The resulting file was subjected to phylogenetic analysis using the program PHYlogeny Inference Package (PHYLIP version 3.69). An un-rooted neighbor-joining Phylogenetic tree was constructed by generating 100 random bootstrap replicates using PHYLIP. Resulting tree file was visualized by Mega 5.05 program.

**Results and Discussion**

**Identification and Listing of PfSET Genes in *P. falciparum***

By gene text searching and BLASTp analysis carried out with a threshold e-value of ≤10 using protein sequences of SET domain from various organisms (*D. melanogaster*, *A. thaliana*, *H. sapiens*) as a query, we were able to extract total 10 putative PfSET genes from PlasmoDB. However, out of 10 putative PfSET genes, SET domain was confirmed in 8 PfSET genes by Pfam and SMART databases (Table 1). Remaining 2 putative PfSET genes did not encode any SET domain according to Pfam and SMART. The HKMTs without SET domain are reported. A previous study although reported nine SET genes in *P. falciparum*.

**Analysis of PfSET Genes**

Total 10 putative PfSET genes were found to be widely distributed over chromosomes 4-6, 8, 9, 11-13 as shown in Figure 1. A maximum of 2 genes were located on chromosome 12 (PF3D7_1214200 & PF3D7_1221000) and 13 (PF3D7_1322100 & PF3D7_1355300). Out of 10 PfSET putative genes, only 5 genes were found to contain introns. Number of introns varied between a range of minimum 1 (PF3D7_1322100 & PF3D7_0910000) to maximum 6 (PF3D7_0508100) in these genes. Length of PfSET proteins varied from a minimum of 178 amino acids (PF3D7_1214200) to a maximum of 6753 amino acids (PF3D7_0629700) as shown in Figure 2. There were large variations in isoelectric point (pI) values; ranging from acidic 4.2 (PF3D7_1115200) to basic 9.51 (PF3D7_1214200) and molecular weight from 20.7 kDa (PF3D7_1214200) to 796.1 kDa (PF3D7_0629700).

**Table 1. Description of PfSET genes**

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Gene ID</th>
<th>Product description</th>
<th>Chromosome No. [Gene location]</th>
<th>Nucleotide length (bp)</th>
<th>Protein length (aa)</th>
<th>Molecular weight (Da)</th>
<th>No. of SET (SMART &amp; Pfam)</th>
<th>Isoelectric point (pI)</th>
<th>No. of Introns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>PF3D7_0629700</td>
<td>SET domain protein, putative (SET 1)</td>
<td>6 [PF3D7_06_v3: 1,221,941 - 1,242,922 (+)]</td>
<td>20259</td>
<td>6753</td>
<td>796053</td>
<td>1</td>
<td>8.77</td>
<td>0</td>
</tr>
<tr>
<td>2.</td>
<td>PF3D7_1322100</td>
<td>SET domain protein, putative (SET 2)</td>
<td>13 [PF3D7_13_v3: 920,971 - 929,449 (+)]</td>
<td>7644</td>
<td>2548</td>
<td>300717</td>
<td>1</td>
<td>7.43</td>
<td>1</td>
</tr>
<tr>
<td>3.</td>
<td>PF3D7_0827800</td>
<td>SET domain protein, putative (SET 3)</td>
<td>8 [PF3D7_08_v3: 1,198,858 - 1,206,057 (+)]</td>
<td>7197</td>
<td>2399</td>
<td>283552</td>
<td>1</td>
<td>8.3</td>
<td>5</td>
</tr>
<tr>
<td>4.</td>
<td>PF3D7_0910000</td>
<td>SET domain protein, putative (SET 4)</td>
<td>9 [PF3D7_09_v3: 452,399 - 455,743 (-)]</td>
<td>3342</td>
<td>1114</td>
<td>131892</td>
<td>1</td>
<td>8.6</td>
<td>1</td>
</tr>
<tr>
<td>5.</td>
<td>PF3D7_1214200</td>
<td>histone-lysine N-methyltransferase, putative (SET 5)</td>
<td>12 [PF3D7_12_v3: 607,379 - 608,006 (-)]</td>
<td>534</td>
<td>178</td>
<td>20663</td>
<td>1</td>
<td>9.51</td>
<td>4</td>
</tr>
<tr>
<td>No.</td>
<td>Gene ID</td>
<td>Product Description</td>
<td>Location</td>
<td>No. of Introns</td>
<td>Isoelectric Point</td>
<td>Molecular Weight (kDa)</td>
<td>Amino Acid Sequence Length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----------------</td>
<td>----------------------------------------------------------</td>
<td>------------------------</td>
<td>----------------</td>
<td>-------------------</td>
<td>-------------------------</td>
<td>---------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>PF3D7_1355300</td>
<td>histone-lysine N-methyltransferase, putative (SET 6)</td>
<td>13 [PF3D7_13_v3: 2,195,449 - 2,196,978 (+)]</td>
<td>1</td>
<td>8.82</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PF3D7_1115200</td>
<td>histone-lysine N-methyltransferase (SET 7)</td>
<td>11 [PF3D7_11_v3: 576,773 - 580,276 (+)]</td>
<td>1</td>
<td>4.2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>PF3D7_0403900</td>
<td>SET domain protein, putative (SET 8)</td>
<td>4 [PF3D7_04_v3: 219,087 - 222,838 (+)]</td>
<td>1</td>
<td>8.56</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>PF3D7_0508100</td>
<td>SET domain protein, putative (SET 9)</td>
<td>5 [PF3D7_05_v3: 331,680 - 336,704 (+)]</td>
<td>0</td>
<td>8.91</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>PF3D7_1221000</td>
<td>histone-lysine N-methyltransferase, putative (SET 10)</td>
<td>12 [PF3D7_12_v3: 836,912 - 843,901 (-)]</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Information regarding product description, gene location, nucleotide sequence length, no. of introns, isoelectric point, molecular weight and amino acid sequence length is extracted from Plasmo DB. Confirmation of SET domain in each gene is based on SMART and Pfam database.

Figure 1. Distribution of PfSET genes on the chromosomes of *P. falciparum*

Figure 2. Graphical representation of protein sequence length of PfSETs with their respective gene IDs at Y-axis
Domain Architecture of PfSET Proteins

The domain architecture of proteins encoded by PfSET genes is based on the presence of SET domain and other additional domains as identified by SMART database. Out of total ten PfSETs, six PfSETs (PF3D7_0827800, PF3D7_0910000, PF3D7_1214200, PF3D7_1355300, PF3D7_1115200 and PF3D7_0403900) were found to encode only SET domain. Remaining four PfSETs were found to have other domains besides SET domain as shown in Figure 3. Two PfSETs (PF3D7_0629700, PF3D7_132200) were found to have 4 PHD domains additional to SET domain. PF3D7_0629700 was identified to encode one Bromo domain and one post SET domain besides PHD and SET domains. PF3D7_0508100 and PF3D7_1221000 were found to have no SET domain. But these were annotated as SET genes at PlasmoDB. PfSET9 contains Ankyrin repeats (AKN domain) whereas PfSET10 (PF3D7_1221000) harbors one PHD, one RING and three PbH1 domains.

Prediction of Subcellular Localization of PfSET Proteins

PfSET sequences were analyzed for putative signal sequences by MITOPROT, Euk-mpLoc 2.0 server and PATS and depicted in Figure 4. Out of eight PfSETs bearing nuclear localization signals, six (55%) were found to be exclusive to this organelle as shown in Figure 4 (a). PfSET9 (PF3D7_0508100) was found to be present in nucleus and cytoplasm both, whereas PfSET5 (PF3D7_1214200) was found to be targeted to nucleus and mitochondria. PfSET6 (PF3D7_1355300) was found to be cytoplasmic exclusively. No PfSET was found to be exclusively mitochondrial or apicoplast targeted. However, PfSET7 (PF3D7_1115200) was predicted to be shuttling between cytoplasm and mitochondria. Cellular localization of PfSET proteins is represented in Figure 4 (b). The predicted nuclear localization for most of PfSETs confirms their role in chromatin related processes.

The domain architecture is based on the presence of SET and other additional domains as identified by SMART and Pfam. Protein sizes are not to scale.

Figure 3. Domain architecture PfSET proteins
Expression Profiling of PfSET Genes

In order to study expression of PfSET genes during intraerythrocytic developmental cycle (IDC) of the malaria parasite life cycle, we analyzed transcriptome data by DeRisi group (Figure 5a). DeRisi group represented the transcriptome of IDC for P. falciparum 3D7 at one hour time interval with 53 time points. The Proteome data of various studies was compiled in Figure 5b.

PF3D7_0629700 was found to have constant expression profile during IDC according to transcriptome data. The protein of PF3D7_0629700 was expressed in T, S, M G and Sp. PF3D7_1322100 was identified to be upregulated in ring and having a constant expression profile in trophozoite and schizont stages at mRNA level and its protein was detected in R,T,S,Gt and SGS. PF3D7_0827800 gene was predicted to have constant expression throughout all three erythrocytic stages at transcript level. However, a real-time quantitative expression analysis of PfSETs highlighted discrepancies in transcript abundance of PF3D7_0827800 as compared to microarray datasets with a peak expression in schizont stage. The Proteomic datasets suggested its protein existence at R,T,S,G and Sp. PF3D7_0910000 protein was found to be SGS specific. PF3D7_1355300 was found to be upregulated in R at mRNA whereas its protein was restricted to S, EG, LG and Sp stages. Transcripts of PF3D7_1115200, PF3D7_0403900 and PF3D7_0508100 were found to be upregulated in S. PF3D7_1115200 protein expressed in R, T, S, and Sp. PF3D7_0403900 and PF3D7_0508100 were found to be restricted to S and M.

Expression profiling at transcript and protein level of PfSET genes suggested a mixture of relationships between transcriptome and proteome datasets. Further, present analysis also revealed disparities between different proteome datasets.
a) Heat map of Derisi microarray data of Llinas et al.\textsuperscript{31} b) Heat map of proteome and phosphoproteome data retrieved from Florens et al.\textsuperscript{2}, Lasonder et al.\textsuperscript{3}, Le Roch et al.\textsuperscript{3}, Silvestrini et al.\textsuperscript{23}, Solyakov et al.\textsuperscript{25}, Treeck et al.\textsuperscript{27}, Oehring et al.\textsuperscript{28}, Lindner et al.\textsuperscript{29}, Pease et al.\textsuperscript{30}. Different stages of *P. falciparum* are represented as: R, rings; T, trophozoites; S, schizonts; M, merozoites; G, gametocytes; Sp, sporozoites; Gt, gamete; EG, early gametocyte; MG, mature gametocyte; OOC, oocyst; ODS, oocyst derived sporozoites; SGS, salivary gland sporozoites; phospho E, phospho-enriched; phosphoD, phospho-depleted; N, nuclear; and C, cytoplasm.

**Figure 5.** Expression Profiling of *PfSETs*

**Table 2.** List of human homologs of *PfSET* proteins

<table>
<thead>
<tr>
<th>S. No.</th>
<th><em>P. falciparum</em> Gene ID</th>
<th>Product Description</th>
<th>Protein sequence length (aa)</th>
<th>Homo sapiens Gene ID</th>
<th>Product Description</th>
<th>Protein sequence length (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>PF3D7_0629700</td>
<td>SET domain protein, putative (SET 1)</td>
<td>6753</td>
<td>NP_005924.2</td>
<td>Histone-lysine N-methyltransferase MLL1</td>
<td>3969</td>
</tr>
<tr>
<td>2.</td>
<td>PF3D7_1322100</td>
<td>SET domain protein, putative (SET 2)</td>
<td>2548</td>
<td>NP_075447.1</td>
<td>Histone-lysine N-methyltransferase NSD3</td>
<td>1437</td>
</tr>
<tr>
<td>3.</td>
<td>PF3D7_1355300</td>
<td>histone-lysine N-methyltransferase, putative (SET 6)</td>
<td>509</td>
<td>NP_073580.1</td>
<td>SET and MYND domain-containing protein 3</td>
<td>428</td>
</tr>
<tr>
<td>4.</td>
<td>PF3D7_1115200</td>
<td>SET domain protein, putative (SET 7)</td>
<td>793</td>
<td>NP_006053.1</td>
<td>SET and MYND domain-containing protein 5</td>
<td>418</td>
</tr>
<tr>
<td>5.</td>
<td>PF3D7_0403900</td>
<td>SET domain protein, putative (SET 8)</td>
<td>1186</td>
<td>NP_065115.3</td>
<td>N-lysine methyltransferase SETD8</td>
<td>393</td>
</tr>
</tbody>
</table>
Prediction of Human Homologs of PfSETs

Prediction of homologs of PfSET proteins in Homo sapiens was carried out using BLASTp of PfSET protein sequences against HPRD (Human Protein Reference Database). Out of total 10 PfSET genes, we could predict human homologs for 5 PfSET genes (Table 2). All PfSETs are longer in length than their human homologs highlighting parasite specific insertions. Further, it will be interesting to compare domain structure and sequences of these PfSET proteins with their human homologs to identify parasite specific features.

Multiple Sequence Alignment of PfSET Genes

SET domain harbors four signature motifs- motif 1 (G-X-G), motif II (YXG), motif III (RFINHXCXPN) and motif IV (ELXFDY). By multiple sequence alignment, we were able to recognize all four motifs I, II, III and IV of SET domain as shown in Figure 6. Motif 1 and II were found to be less conserved as compared to motifs III and IV. Motifs III and IV form a pseudo knot fold which brings the two most-conserved sequence motifs III (RFINHXCPN) and IV (ELXFDY) of the SET domain together to form an active site. The conserved residues (G-X-G) of motif 1 and NH residues of the motif III are involved in hydrogen bonding and van der Waals interactions with the cofactor AdoMet. Further, Y residue of motif IV was found to be conserved in the lysine binding cleft which form hydrogen-bond with amino group of K residue and align it for a methyltransfer with AdoMet.
Generation of Sequence Logo of *PfSET*

A sequence logo for SET domain of *P. falciparum* was generated by Web logo based on multiple sequence alignment of all 8 SET domain sequences identified in *P. falciparum*. Generated *PfSET* domain sequence logo is shown in Figure 7. The overall height of each stack is proportional to the residue conservation at that position and the height of each letter corresponds to frequency of particular residue at that position. The logo shows the level of conservation of four signature motif consensus sequences as predicted by multiple sequence alignment. In addition to these motifs conservation of some more residues like K (lysine), Y (tyrosine), A (alanine), R (arginine) and D (aspartic acid) is also revealed.

Phylogenetic Analysis of *PfSETs*

An un-rooted neighbor-joining Phylogenetic tree for 10 *PfSET* protein sequences was constructed using Phylip 3.69 and was visualized with MEGA 5.05 program (Figure 8). Phylogenetic analysis revealed that PfSET1, PfSET9, PfSET4, PfSET5 and PfSET10 are grouped together in one calde. PfSET7 and PfSET8 are closely related to each other and grouped together. However, PfSET3 and PfSET6 revealed distant relationships.
Conclusion
The present study provides detailed bioinformatics analysis of PFSET family. In this study, we identified 10 PFSET genes in P. falciparum. Further, we carried out detailed analysis of their domain composition, physicochemical properties, subcellular localization, expression profiling and phylogenetic relationships. The work highlighted the unique domain composition and conservation of important catalytic residues in PFSETs. Transcriptome and proteome profiling of PFSETs exhibited a mixture of linear and nonlinear relationships and revealed stage specific PFSETs. In essence, genome-wide analysis of PFSET family in P. falciparum will provide a platform for further experimental studies.

Acknowledgement
MK, PC thankfully acknowledge the Maharshi Dayanand University, Rohtak, Haryana for University Research Fellowship. RG acknowledges the UGC and DST funding.

Conflict of Interest: None

References

ISSN: 0019-5138
DOI: https://doi.org/10.24321/0019.5138.201934
