Table 3 *Schistosoma japonicum* genome contigs containing the similar signal sequence (n = 34)

*Contig [GenBank Accession No.]	Transcription Strand	Size, kb	Signal Sequence coordinates
SJC_C002611 [CABF01002612]	-	69.7	4848 – 4779
SJC_C002621 [CABF01002622]	-	0.6	276 – 205
SJC_C002622 [CABF01002623]	-	1.6	999 – 928
SJC_C002627 [CABF01002628]	-	3.0	2413 – 2342
SJC_C002629 [CABF01002630]	-	14.5	3856 – 3785
SJC_C013669 [CABF01013761]	-	6.8	3284 – 3217
SJC_C019814 [CABF01020047]	+	29.2	19023 – 19094
SJC_C019817 [CABF01020050]	-	10.1	6502 – 6431
SJC_C019827 [CABF01020060]	-	43.7	42511 – 42440
SJC_C022876 [CABF01022876]	-	12.4	9335 – 9264
SJC_C022884 [CABF01022884]	+	12.9	493 – 564
SJC_C023364 [CABF01023364]	-	12.4	10498 – 10427
SJC_C025268 [CABF01025296]	-	12.1	7860 – 7789
SJC_C027826 [CABF01027854]	+	4.3	433 – 504
SJC_C027833 [CABF01027861]	-	11.9	5768 – 5697
SJC_C027838 [CABF01027866]	+	19.0	12367 – 12438
SJC_C032855 [CABF01032892]	-	22.3	383 – 322
SJC_C032859 [CABF01032896]	-	9.6	4663 – 4602
SJC_C043165 [CABF01043187]	+	4.9	2484 – 2544
SJC_C057153 [CABF01057161]	+	2.8	337 – 408
SJC_C061392 [CABF01061395]	-	7.4	4411 – 4342
SJC_C067189 [CABF01067176]	+	4.9	388 – 459
SJC_C067567 [CABF01067411]	-	3.2	925 – 854
SJC_C070280 [CABF01070230]	-	6.8	271 – 200
SJC_C072631 [CABF01072590]	+	1.9	1646 – 1717
SJC_C072632 [CABF01072591]	+	2.6	446 – 517
SJC_C073741 [CABF01073691]	-	2.4	2389 – 2319
SJC_C075160 [CABF01075030]	-	6.7	5231 – 5160
SJC_C076469 [CABF01076032]	+	3.1	1295 – 1366
SJC_C077101 [CABF01078976]	-	1.1	656 – 585
SJC_C080985 [CABF01080674]	+	2.3	1094 – 1165
SJC_C081391 [CABF01080757]	-	2.1	1918 – 1847
SJC_C081246 [CABF01080893]	-	1.3	1131 – 1060
SJC_C097686 [CABF01092393]	+	5.6	4249 – 4320

(-) are contigs with 'signal sequence' on the negative strand (anti-sense) of the genome while
(+) are contigs with 'signal sequence' on the positive strand (sense) in the genome
*Contigs are representative of disperse duplicated gene loci. We indicated the ranges for the signal sequence motif.

To investigate possible role by repetitive elements (REs) in mediating such dispersed duplication with a clue from previous studies [20,26-29], we performed repeat masking on the putative duplication source locus and the other 33 duplicons and observed a preponderance of flanking REs, especially of non-LTR class prominent of which were the *S. japonicum RTE (*retrotransposable element*)*-like retrotransposon (*SjR2*) and the *Perere* class of retrotransposons *(SjR1)* (Additional file 5). An almost full copy of *SjR2* was found upstream of the coding region of the putative source locus in addition to other six albeit partial copies of *SjR2*. Alignment of the other contigs to the putative duplication source locus revealed that both the dispersed similar signal sequence and the repeat elements are considerably aligned at very similar positions, further showing that they were likely

Table 4 *Schistosoma japonicum* Scaffolds containing the similar signal sequence (n = 18)

Scaffolds [GenBank Accession]	Contigs within the scaffolds
SJC_S000013 [FN330988]	CABF01002611, CABF01002612, CABF01002622, CABF01002623, CABF01002628, CABF01002630
SJC_S000219 [FN331192]	CABF01020047, CABF01020050
SJC_S000220 [FN331193]	CABF01020060
SJC_S000273 [FN331245]	CABF01022876, CABF01022884
SJC_S000284 [FN331256]	CABF01023364
SJC_S000329 [FN331301]	CABF01025296
SJC_S000394 [FN331366]	CABF01027854, CABF01027861, CABF01027866
SJC_S005820 [FN336777]	CABF01067176
SJC_S007785 [FN338731]	CABF01070230
SJC_S008639 [FN339578]	CABF01072590, CABF01072591
SJC_S009177 [FN340103]	CABF01073691
SJC_S010134 [FN341037]	CABF01075030
SJC_S011206 [FN342077]	CABF01076032
SJC_S011724 [FN342573]	CABF01078976
SJC_S014521 [FN345237]	CABF01080674
SJC_S014753 [FN345459]	CABF01080893
SJC_S014868 [FN345568]	CABF01080757
SJC_S026182 [FN354050]	CABF01092393

duplicated from a single source locus. The fact that the duplicons are not absolutely homologous and the degenerative nature of the RE sequences suggests variation within members, typical of evolving genes (Figure 4). Because homology with the other duplicates did not terminate 3` of this putative source locus, we recruited and adjoined two contigs [GenBank:CABF01020061 and GenBank:CABF01020062] downstream of the putative source locus according to the genome assembly information, thereby creating flanking sequences of at least 5 kilobasepairs on each side of the gene duplication source locus. This sequence was then aligned with the genome contigs and scaffolds to identify the exact point at which homology was lost, which could arguably represent the breakpoint of duplication. Further attempt to identify the exact breakpoints was not successful due to unfilled sequencing gaps in the scaffolds but examination of the downstream flanking sequence from the point where homology was terminated showed a prominent retrotransposon of the *Perere* class flanking the duplicated loci 3` of the locus (see short movie in Additional file 5). Taken together, our data show that the duplication source locus was flanked on either side by *RTE*-like and *Perere* class retrotransposons. These two classes of non-LTR retrotransposons have significantly high copy number, making up 12.63 % of the *S. japonicum*

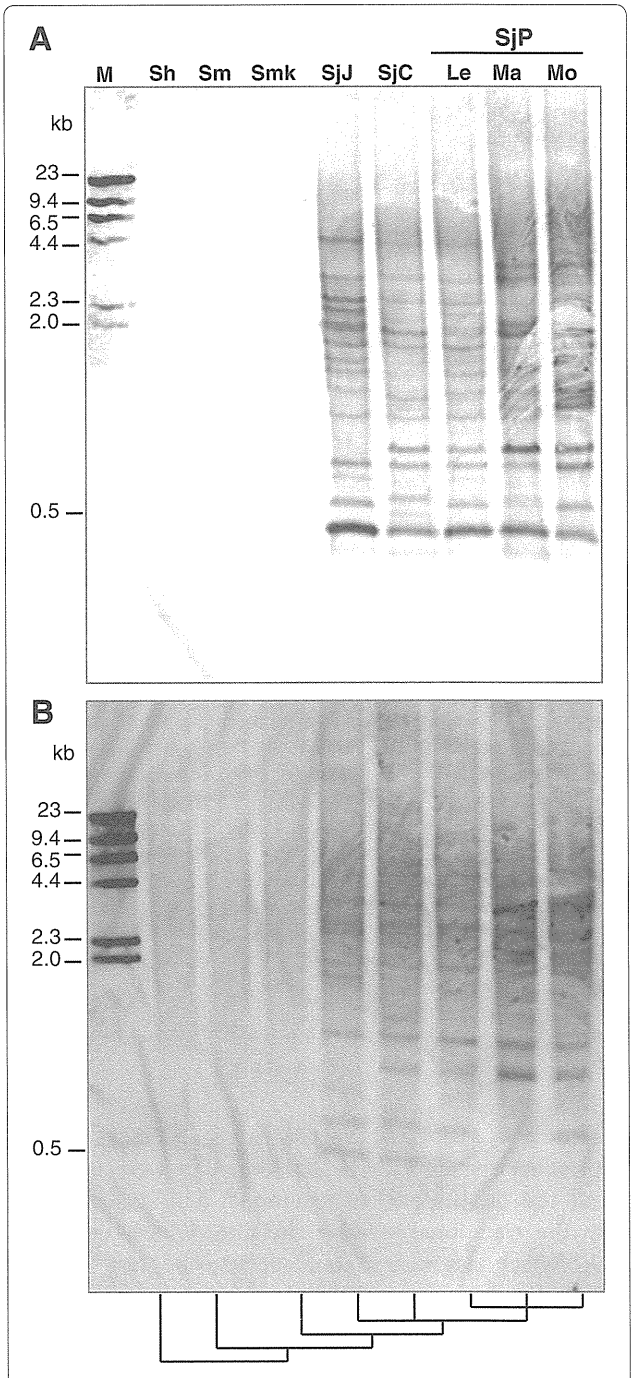

Figure 2 Southern blotting confirms duplicated loci exclusively in *S. japonicum*. (**A**) Southern hybridization with digoxigenin-labeled probes showing the presence of duplicated loci with several bands due to copies of the duplicated source locus. Lanes 2-9 corresponds to *Eco*RI + *Eco*RV double digested genomic DNA of different species and strains of Schistosoma (*S. haematobium*, *S. mansoni*, *S. mekongi*, *S. japonicum* (Japanese, Chinese, Philippines' Leyte, Mindanao and Mindoro isolates). 'M' is Digoxigenin-labeled DNA molecular weight marker. Notice the differential banding pattern among different strains and between isolates of the same strain. (**B**) Same experiment as in (A) was replicated using a different pair of restriction enzymes (EcoRI + *Hind*III). Also inter-strain and intra-strain variation in the banding pattern is apparent.

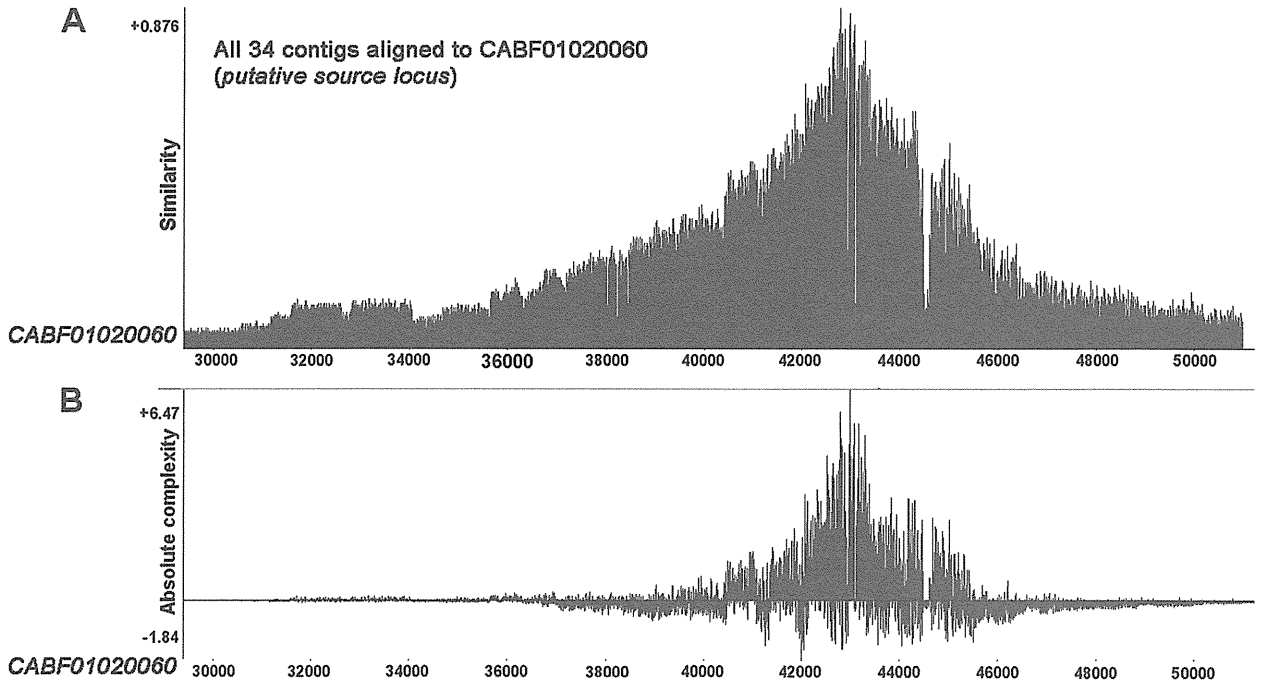

Figure 3 Multi-alignments of 34 *S. japonicum* genome contigs representing duplicated loci to assign putative source locus. Graph shows the sequence similarity (**A**) and absolute complexity (**B**) of the DNA sequence of the 34 contigs in *S. japonicum* genome containing the duplicated loci. This multiple alignment was used to putatively assign the most prominent contig [GenBank:CABF01020060], the longest among the identified dataset (43.7 kb), which significantly covered the length of the other contigs as the putative duplication 'source locus'. The curve shows the probable length of the duplicated locus, terminating with *RTE*-SJ at the 5` end (trimmed out in this figure) and *Perere* at the 3` end. 'Similarity' curve is a measure of the level of similarity of the aligned sequences. The y-axis on the 'similarity' curve will read '1' when the sequences are 100 % similar in each position. The output shows the maximum score on the y-axis. Absolute complexity is a measure of the level of conservation or variability of nucleotides in the aligned sequences. It is a measure of the likelihood that the observed similarity did not occur by chance. The maximum positive score on the y-axis of the 'absolute complexity' curve is expected to be higher than the negative value to exclude any possibility that the observed similarity occurred by chance. The x-axes in both curves represent the nucleotide positions. The numbering of the nucleotides started at the 30000th position in this figure because we trimmed output figure at the 5` end for ease of presentation.

genome [49]. Different degrees of degeneracy of both the coding region and the flanking REs were observed in all the duplicated loci examined. This is consistent with the traditional view of the fate of new duplicons [6,9], which assumes a tendency to be lost because of genetic drift under natural evolution [29,57] while not precluding the possibility for some duplicates to evolve distinct functions either by sub-functionalization or neo-functionalization.

The role of repetitive elements (REs) in dispersed duplication of genomic sequences is fairly documented from previous studies in model organisms [15,20,27,28,30,58,59]. The precise mechanism of this retrotransposon mediated dispersed duplication is not clear but may likely involve RE-mediated DNA level recombination, most likely by non-allelic homologous recombination (NAHR), alternatively called ectopic recombination (see illustration in Additional file 6). Due to their extremely high copy numbers, REs create structural modifications in the genome by providing the requisite highly similar DNA sequences, initiating recombination between non-allelic elements [20,25,60], the result of which could be deletion, shuffling, duplication or transduction of a genomic DNA segment. Structural modifications introduced in the genome by NAHR mechanism can progress between non-homologous chromosomes (inter-chromosomal), between homologous chromosomes (inter-homologous or intra-chromosomal), between sister chromatids (inter-sister chromatid) or within a chromatid (intra-chromatid); giving rise to dispersed duplication of genomic segments, several forms of deletions or may create isodicentric chromosome by forming a mirrored segment in the chromosome by inversion. See detailed cartoon in Additional file 6. [60].

Many studies in other organisms have elucidated the role of REs in mediating sequence duplication, transduction and other structural variations by ectopic recombination mechanism. Notable among these is the human *Alu* element for which several reports suggest a role in

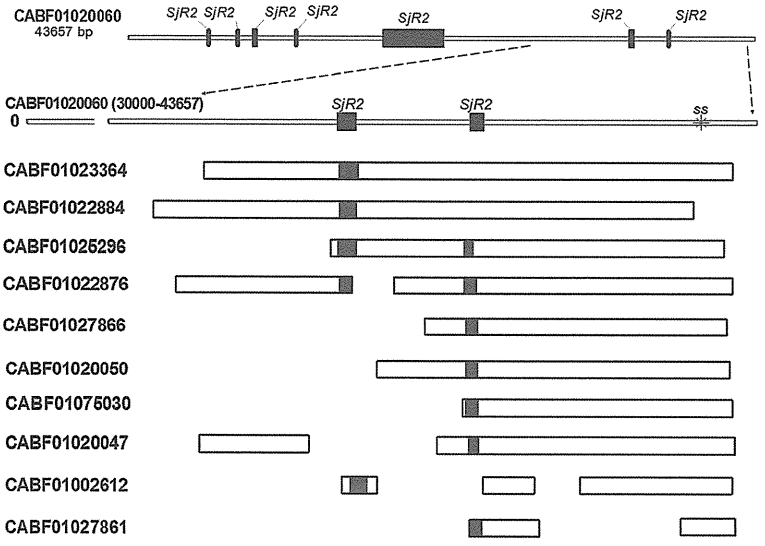

Figure 4 Further evidence that the duplicated genes were duplicons of a single duplication source locus. Apart from the prominent flanking copies of retrotransposons observed around the putative gene duplication source locus [GenBank:CABF01020060], other two short copies of the retrotransposon (*RTE_SJ*) are also found within introns in the coding region. We aligned the source locus with 10 of the duplicons and observed that both the signal sequence and these two partial copies of *RTE_SJ* are relatively aligned at same position, further indicating that the duplicated genes could have originated from a single source locus.

mediating NAHR and other structural modifications in the human genome [7,20,61]. Yang *et al* found an excess of repetitive sequences proximate to the breakpoints of duplicated gene loci in the genome of the fruit fly *Drosophyla melanogaster,* and have suggested that a NAHR mechanism, mediated by REs accounted for the birth of the new duplicons [1,27]. Another study performed on human individuals concluded that NAHR accounted for over 40 % of detected genomic sequence duplications in the human genome [30]. Illegitimate recombination (IR), incomplete crossing over and non-homologous end joining (NHEJ) are other possible mechanisms of gene duplication by DNA-level recombination, but NAHR play a more significant role in producing typical dispersed duplications [1] while the other mechanisms in addition to NAHR are more likely to produce tandem duplicates. Although we could not clearly identify the exact breakpoints of the duplications at both ends still for lack of a reference ancestral homolog and partly due to sequencing gaps, the fact that homology among all the scaffolds examined uniformly terminated at the same point with *Perere* on the 3` end (Figure 3 and Additional file 5), and traces of the observed predominant retrotransposons (*SjR2)* was found at the exact positions as they occur in the putative source locus (Figure 4) confirm that these gene loci could be products of dispersed duplication from a single genomic source locus.

In addition to RE-mediated DNA-level recombination by NAHR, gene duplication events are also attributable to RE-mediated retrotransduction mechanism either on the 5` or 3` directions [27]. Xing *et al* and other groups have demonstrated the role of retrotransposons in the duplication of entire genes and creation of previously un-described genes by analyzing SVA (SINE, VNTR and Alu)-mediated retrotransduction events in the human genome [20,29]. However, we did not specifically identify any chimeric duplicon originating via a retrotransduction mechanism among our datasets. Furthermore, retrotransposons including *SjR2* characteristically encode reverse transcriptase and endonuclease, and can therefore transcribe and 'paste' a gene sequence into new locations in the genome [3,22,62]. However, retrotransposed genes are characteristically intronless since the introns are usually spliced out during the process of retrotransposition. Our duplicons retained their introns, although in some case some portion of the introns may have either degenerated or deleted during duplication and subsequent sequence modifications [3,22,63]. A further evidence that a retrotransposition mechanism is unlikely in our observed cases was that while retrotransposons would not duplicate the promoter regions of duplicated gene based on the process of transcription and insertion of retrocopies [1,57] which leaves the newly retrotransposed sequences to acquire new regulatory sequences from adjacent genes or through mutations in order to be functional [14,19,24]; the protein coding duplicons observed among our duplicated gene loci retained the same or similar core regulatory region and signal sequence as the source locus, suggesting that they may not have been products of retroposition and

may equally explain the parallel assumption of coding potential at their new duplication loci without the need to form chimeric structures with adjacent genes.

Evolution of translatable ORF and evidence of expression of duplicated genes

Some of the duplicons appear degenerative in homology and are relatively shorter than the source locus (Figure 3, also see Additional file 5) thus are consequently redundant and non-coding at the new locations as opined in the canonical view on the fate of new duplicons [6,9]; which assumes a tendency to be lost because of genetic drift under natural evolution [29,57]. However, our data provide evidence that some of the duplicons have evolved into protein coding genes with distinct products at their new loci, the fate of which could tend to either sub-functionalization to the source gene [8,64] or neo-functionalization by acquiring new distinct functions [9,65]. In addition to the two duplicons with alternative splicing variants, which we further explored in the next section, some representatives of the protein coding duplicons were depicted in a supplementary figure (Additional file 7). The nucleotide sequences of these genes are still appreciably similar but accumulation of mutations and other sequence modifications have given rise to novel protein coding ORFs, encoding putatively distinct products. We identified and mapped each cDNA sequence to the genomic contigs using information we generated from GeneMark and GeneQuest gene predictions [66] and confirmed by alignment of the cDNAs to the genomic sequences using NCBI *Spling* program. This approach was necessitated because the fully mapped and annotated genome of *S. japonicum* is not presently available in the public databases. Intriguingly, our results corroborate the available UniGene and GenBank entries. Nevertheless, it is notable that we only assessed the duplicated copies on the basis of possessing the similar signal sequence. There is possibility that some other duplicons from this source locus could be involved in initiating other forms of structural modifications at other loci when incorporated into the coding region of other genes, but this was not investigated here.

To provide evidence of the transcription and expression of the putative source gene locus and some of the duplicons, we performed developmental stage specific RT-PCR using primers that specifically amplify the coding regions of the candidate genes from the cDNA libraries of each stage of *S. japonicum*. RT-PCR results provide evidence of the transcription of some of the duplicons at their new genomic sites in addition to the source locus (Figure 5). The candidate genes analyzed did not show differential developmental stage specific expression, although we did not perform quantitative estimation of expression levels. It is possible that this group represents a potential new family of proteins with similar signal peptides in this zoonotic trematode, which possess other extra distinctive characteristics from other members of the genus *Schistosoma*. We are presently undertaking further research to fully characterize the identified novel protein-coding genes to provide insight into the functional and structural significance of this trend in the genome of *S. japonicum*. The protein products of some of these candidate genes have already been expressed in our laboratory and confirmed by the reactivity of the immune sera with the parasite crude antigen preparations. The data will be reported with the molecular and functional characterization information.

Functional selection by alternative splicing

The precise recognition of exon-intron junction in a precursor mRNA (pre-mRNA) by the splicing machinery is central for the production of functional translatable mRNAs. However, there is often uncertainties in the choice of recognizable splice signals, resulting in a process termed alternative splicing [17], which enables the origination of multiple mRNA transcript variants from a single gene locus [67-69]. Alternative splicing mechanism could result in 'intronization' of an exon or 'exonization' of an intronic sequence. Ideally, the creation of an intron from a previously exonic sequence could lead to the loss of an ORF in coding genes. In evolving genes however, functional selection possibly by mutations may evolve the required splice signals and induce the intronization of an exon in a transcribed but non-coding mRNA gene sequence to create a translatable ORF encoding a functional protein. Conversely, while exonization of an intron could disrupt a translatable ORF in a coding gene, selective pressure may also evolve new splice signals within an intron to yield exons that could create a translatable ORF from a previously non-coding gene locus or a chimeric ORF from a protein-coding gene.

These two mechanisms have been shown from our observations to be capable of creating functional coding-genes from previously non-coding albeit transcribed mRNA sequences. We identified at least two classical evidences of alternative splicing and we propose that in addition to increasing coding potential and genomic diversity [68,69], alternative splicing can also be one of the driving forces of adaptive evolution; producing genetic novelties and functional selection. The most prominent example of alternative splicing was observed in the duplication source locus [GenBank:CABF01020060], which was found to be able to produce a protein-coding mRNA [GenBank:AY570737] in addition to a non-coding mRNA transcript variant [GenBank:FN328299] (Figure 6). An alignment of the DNA sequences of these two transcripts with details of this observation is

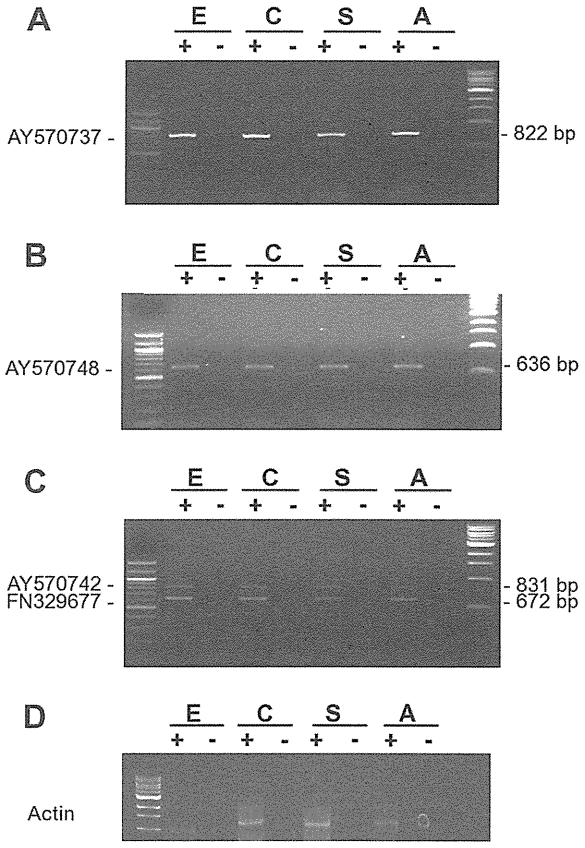

Figure 5 RT-PCR showing expression patterns of some of the duplicons in the developmental stages of the parasite. RT-PCR using cDNA libraries of the parasite egg (E), cercaria (C), schistosomula (S) and mixed sex adult (A) as template provide evidence of the transcription and expression of some of the duplicons. The pairs of primers used were designed to amplify the entire coding sequences of the mRNAs. No differential expression pattern was observed but quantitative expression levels were not investigated. (**A**) Evidence of expression of SjCP1084 protein coding mRNA [GenBank:AY570737], transcribed from the putative source locus [GenBank:CABF01020060]. See Table 2 for list of other similar transcripts. Notably, a non-coding transcript variant [GenBank: FN328299] can also be transcribed from the same locus. See Figure 6 and Additional file 8 for more details. (**B**) Expression of SjCP3842 protein coding mRNA [GenBank:AY570748] predictably transcribed from [GenBank:CABF01002612]. See Table 2 for list of other similar transcripts in the database. (**C**) Expression of SjCP1531 [GenBank:AY570742] predictably transcribed from [GenBank: CABF01023364]. See Table 2 for list of other similar transcripts. Notably, a non-coding transcript variant [GenBank:FN329677] can also be transcribed from the same locus (second band). See Figure 6 and Additional file 9 for details. (**D**) S. japonicum Actin gene was used as internal control to qualify the samples.

presented in the on-line published supporting information (Additional file 8). UniGene entries also suggest that the two transcripts are from the same locus (Table 2). An extra intron donor and acceptor sites were found within the first exon of the non-coding mRNA transcript [GenBank:FN328299]. While the transcription model of the non-coding variant did not recognize the extra splice signals and thus retained the intron of about 1 kb, the coding mRNA variant [GenBank:AY570737] recognized the splice sites and created an ORF from the gene by splicing out an intron thereby giving rise to the 5`untranslated region (5`UTR) and the first exon of a protein-coding gene encoding a protein product of 271 residues (SjCP1084). Additionally, another pair of splice acceptor and donor sites evolving at exon 5 of the non-coding variant resulted in the splicing out of a portion of the exon, all contributing in creating a translatable ORF in the protein coding variant (See Additional file 8 for details).

On the other hand, exons 5 and 6 of a coding mRNA variant [GenBank:AY570742] predictably transcribed from one of the progeny loci [GenBank:CABF01023364] were skipped in a non-coding shorter variant [GenBank: FN329677] without a translatable ORF (Figure 6 and Additional file 9). We observed that the sequences of exons 5 and 6 were similar and was repeated five times *in tandem* within this locus, but only two copies of the tandemly duplicated potential exons were incorporated into the coding sequence of the mRNA to create exons 5 and 6 of a protein-coding ORF of 274 codons (SjCP1531). These results represent typical models of alternative splicing by intronization and exonization respectively.

Although in evolutionary perspective, intron retention that creates a translatable ORF is considered more plausible than the reverse process; our data show that both mechanisms are potentially possible. Other groups have also identified intron gains recently in mammalian and rodent retrogenes [68,69]. The identification of non-coding mRNA variant alternatively transcribed from a single gene locus with a protein coding mRNA (Figure 6) is evidence that a novel protein-coding gene can originate from previously transcribed regions that contain the necessary transcription elements and provide RNA material for a protein translation machine [2,39,68]. Exon repetition has also been observed from our data to exist in this organism and could be instrumental in expanding the organism`s coding potential. The 'parallel' expression of the non-coding variant alongside the protein-coding transcripts is of significance and could suggest further that the gene may have been recently evolved. Non-coding RNAs have also been shown to perform some regulatory roles at various levels during gene expression [2,68,70]. This could be further explored with our data set. In the two described cases in our analyses, we have treated the non-coding isoforms as evolutionally preceding the coding variants; nevertheless, the reverse could also be the case. In addition to these two cases, we also identified a two-nucleotide insertion into a non-coding mRNA sequence [GenBank: FN330540] that yielded

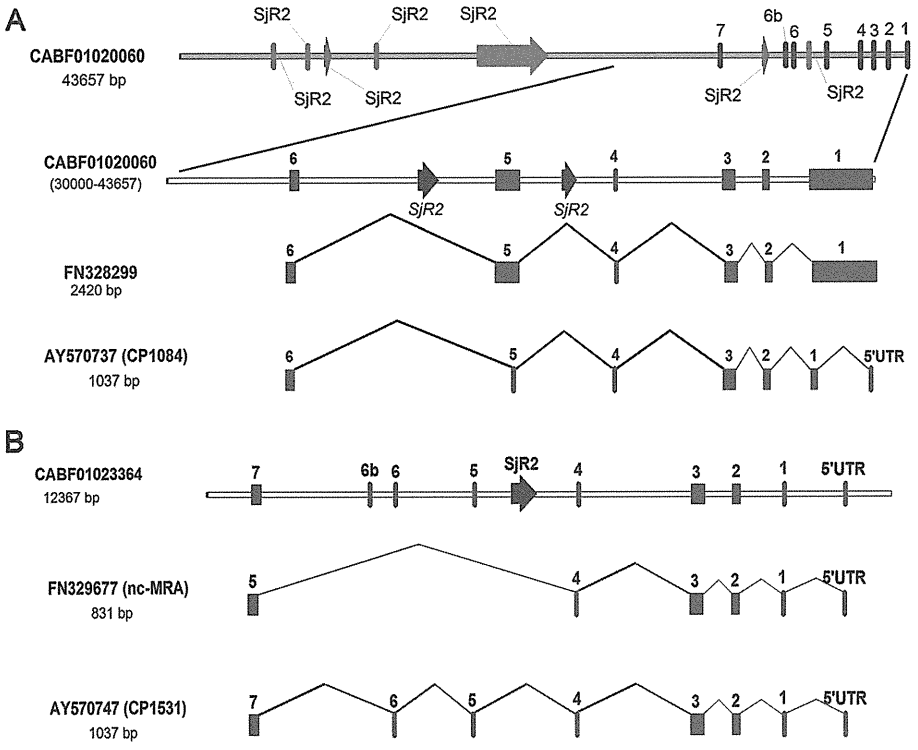

Figure 6 Splice models of some duplicons with evidence of alternatively splicing. (**A**) SjCP1084 protein coding mRNA [GenBank:AY570737] and a non-coding transcript [GenBank:FN328299] are products of alternative splicing. Based on gene prediction from the contigs using GeneQuest and GeneMark, and alignment of cDNAs to genome sequences using *Spling* program, we observed that two mRNA transcript variants were produced from [GenBank:CABF01020060]. An extra splice site was evolved in the first exon of the non-coding transcript [GenBank: FN328299]. When the splice site is recognized, an ORF encoding a protein coding mRNA [GenBank:AY570737] variant is created. The images were created from computer simulation of real DNA sequences using Vector NTI program. Also see a supplementary figure in Additional file 8 for more details. (**B**) SjCP1531 protein coding mRNA [GenBank:AY570742] and a non-coding transcript [GenBank:FN329677] are products of alternative splicing. Two mRNA transcript variants can be produced from a contig representing on of the duplicated loci [GenBank: CABF01023364]. Two extra splice sites were not utilized in the transcription of the non-coding transcript [GenBank:FN329677]. When the splice sites were recognized, exons 5 and 6 of a translatable ORF were created to produce a protein coding mRNA [GenBank:AY570742] variant. Refer to RT-PCR result in Figure 5 (C) where two bands of exact size and sequence as the two variants described above are apparent on the agarose gel electrophoresis image. Also see Additional file 9 for more details.

the coding mRNA of schistosomula protein with the similar signal peptide, with many similar transcripts in the database. However, this last observation could be an artifact from sequencing error since the existence of the non-coding transcript was not traceable to the genomic sequence.

Conclusions

We have passably delineated the possible mechanism that led to the identification of several protein coding genes with similar signal sequence, following lead from our work that isolated secreted proteins candidate genes using SST. A trend was described in the genome of *S. japonicum* whereby a 'newly evolved' gene served as a source locus for dispersed duplication events leading to the formation of several expressed genes with similar transcription core promoter region and signal sequence. We further found that the duplicated gene locus was flanked by non-long terminal repetitive elements (REs), especially of the *RTE*-like and *Perere* class. We therefore inferred that REs may have played an important role in this dispersed gene duplication by creating the requisite homologous DNA sequence that mediate a DNA-level recombination, most probably by a non-allelic homologous recombination (NAHR) mechanism. Our findings also provide evidence of logical sequential process of novel gene origination by evolution of transcription core elements followed by translatable ORF. While similar RE mediated phenomena had been observed in other organisms, unlike our dataset, most analyses have centered on the model organisms. Our data contribute to the accumulating evidence that REs mediate diverse recombination events leading to novel gene origination and other evolutionary novelties.

Methods

BLAST search

We had earlier identified a particular 81 nucleotides (27 amino acids) sequence, which was commonly utilized as signal sequence by several of our signal sequence trap (SST) isolated *S. japonicum* cDNAs (Table 1) [47]. The sequence of this signal sequence was employed as query to search for matches in the GenBank non-redundant nucleotide sequence database and expressed sequence tags (ESTs) database for all organisms using BLASTN program in National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) [71]. A search on the NCBI UniGene database [72] that provides information on sets of transcript sequences that appear to come from the same transcription locus was performed to ascertain redundancy and group identified transcripts. For genome-wide searches, the same query sequence and program were used to search the WGS reads from all organisms with sequenced genomes deposited in the NCBI genome databases. In a similar search in the protein database, the amino acid and nucleotide sequence of the same signal sequence was used as query for BLASTP and BLASTX searches respectively. Conserved domain architecture searches on all translation products of the SST identified candidate genes were performed using the conserved domain architecture retrieval tool on NCBI website [73] and compared with same analyses on the ProDom database of protein domain families available online at [74].

Multiple alignments

All multiple sequence alignments of DNA and protein sequences were performed in parallel with ClustalW on MegAlign program in Lasergene 7 DNASTAR software, NCBI bl2seq, COBALT multiple alignment programs, and *Multialin* interface software [75]. cDNA-to-genome sequence alignments were computed using the free NCBI *Spling* program [75]. The latest update of the *S. japonicum* genome map is accessible at [52]. Phylogenetic and molecular evolutionary analyses were conducted using *MEGA* version 5 [76].

Gene prediction

Gene predictions were performed using the GeneQuest program (Lasergene 7 DNASTAR) to predict potential coding regions, starts, stops, acceptors and donor sites using Borodovsky matrix files for *Caenorhabditis elegans*; and the results compared with that of the Eukaryotic GeneMark.hmm [66] gene prediction server provided freely on the website of Georgia Institute of Technology, Atlanta, USA.

Repeat masking

The whole sequences of all the genome contigs bearing the similar signal sequence were screened against a reference collection of repetitive DNA elements in the RepBase database available at the Genomic Information Research Institute website, using the CENSOR repeat masking software [77]. Sequence analysis figures were generated using real DNA sequences on Vector NTI Advanced 11.0 (Invitrogen).

Designation of putative duplication source locus and probable breakpoint

Reference to a parent gene is required for accurate determination of duplication breakpoint. However, in absence of a reference homolog, we putatively selected the most prominent contig [GenBank:CABF01020060], the longest among the identified dataset (43.7 kb), which significantly covered the length of the other contigs (Figure 3, also see Additional file 5) as the putative duplication source locus and utilized it as such for most of the analyses performed in this study. When the contigs were aligned with the putative source locus, homology was not lost till the 3` end of the aligned sequences. We therefore recruited two contigs [GenBank:CABF01020061 and GenBank:CABF01020062] downstream of the source locus based on genome assembly information, thereby generating at least 5 kb flanking sequences on either side of the duplication source locus. This sequence was then aligned with the genome contigs and scaffolds to identify the exact point where sequence identity disappeared. This point was arguably chosen as the possible duplication breakpoint and utilized as such in our discussions. We further attempted to identify a recurrent consensus sequence at the breakpoints but this was hampered by several sequencing gaps in the partially assembled scaffolds.

Parasites, genomic DNA and developmental stage mRNA samples

Chinese strain of *S. japonicum* (hereafter abbreviated as *Sj*) was obtained from Jiangsu Provincial Institute of Parasitic Diseases Wuxi, Jiangsu Province, PR China, while the Philippine and Japanese strains of *S. japonicum* in addition to *S. mekongi (Smk)* samples, were maintained in the Laboratory of Tropical Medicine and Parasitology, Dokkyo Medical University, Tochigi, Japan. *S. mansoni* (*Sm*) adult worms were maintained by, and kindly provided by the Department of Parasitology, Institute of Tropical Medicine, Nagasaki University, Japan. *S. haematobium* (*Sh*) sample was from Department of Immunology and Parasitology, University of Occupational and Environmental Health, Kitakyushu, Japan. Total genomic DNA was purified from cut tissues of mixed sex adult worms from different species of

Schistosoma using QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's instructions. Qualification and quantification of genomic DNA extract was assessed by gel electrophoresis and ND-1000 spectrophotometer (NanoDrop, USA). To obtain sufficient amount of genomic DNA for southern hybridization experiments, the whole genome of each sample was amplified using the GenomePhi DNA Amplification Kit (GE Healthcare) according to the manufacturer's instructions. Equally, total RNA was extracted from parasite eggs, cercariae, 24 h cultured schistosomulae and adult worms of *S. japonicum* according to the instruction manual of PureLink Micro-to-midi total RNA Purification System Kit (Invitrogen).

Reverse transcription polymerase chain reaction (RT-PCR)

mRNA from eggs, cercariae, 24 h culture schistosomulae and adult worms of the Chinese strain of *S. japonicum* was used for RT-PCR. The first strand cDNA was synthesized from the total RNA of each developmental stage by using oligo (dT) primer according to the instruction manual of High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) and the resulting cDNA was used as template for RT-PCR. The *S. japonicum* actin gene was used for internal quality assurance. The cDNA sequences of some selected SST identified secreted candidate genes were amplified using pairs of sequence specific primers designed according to the *S. japonicum* transcriptome data [49] in the NCBI public database. All RT-PCR amplicons were analyzed using gel electrophoresis and confirmed by sequencing using the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystem).

Southern hybridization

Southern hybridization was performed following standard procedures [78] using the DIG nonradioactive labeling and detection system (Roche, Germany). Briefly, the hybridization probe labeled with DIG-dUTP was synthesized using PCR DIG synthesis kit (Roche, Germany) according to the manufacturer's instructions, and labeling was confirmed by size disparity with unlabeled amplicon as a result of slower migration in agarose gel due to digoxigenin labeling. Genomic DNA from different species of *Schistosoma* (*Sh, Sm, Smk, Sj* Japanese (Yamanashi), *Sj* Chinese (Jiangsu) and *Sj* Philippines (Leyte, Mindanao and Mondoro isolates) were double digested with three different pairs of restriction enzymes (*Eco*RI + *Eco*RV, *Eco*RI + *Hind*III and *Bam*HI + *Hind*III) to achieve the best possible fragmentation of the genomic DNA. The digested genomic DNA fragments were electrophoresed through 1 % (w/v) agarose gel, depurinated in 250 mM HCl, and denatured by incubating in two changes of denaturing solution for 15 min each (0.5 M NaOH, 1.5 M NaCl). The gels were then neutralized by incubation in two changes of neutralizing solution (0.5 M Tris-HCl at pH7.5, 1.5 M NaCl) for 15 min each, and DNA was transferred to a positively charged nylon membrane (Roche, Germany) by capillary action overnight using 20x SSC solution (3 M NaCl, 300 mM sodium citrate at pH 7.0). The transferred DNA was fixed to the membrane by baking in an oven at 80 °C for 2 hours after rinsing briefly in 2x SSC. After prehybridizing the membrane in 10 ml hybridization buffer (5x SSC, 0.1 % N-lauroylsarcosine (w/v), 0.02 % SDS (w/v), 1 % blocking solution (Roche, Germany)) for 30mins in a hybridization bag, 5μl of the PCR generated hybridization probe was mixed in 50μl of double deionized water, denatured by boiling for 5mins and introduced into the hybridization bag and incubated overnight with shaking at 50 °C. The membrane was washed in two changes of low stringent wash buffer (2x SSC, 0.1 % SDS) for 5mins each at RT, and twice in high stringent wash buffers (0.5x SSC, 0.1 % SDS) for 15 min each at 65 °C. The hybridized probe was then detected using anti-Digoxigenin antibody (Roche, Germany) using CSPD as the chemiluminiscent substrate according to the manufacturer's instructions. The blot was then visualized by exposing to chemiluminiscence for 10 min in a LAS-4000 mini image reader (Fujifilm).

Additional files

Additional file 1: Schematics of some of the mechanisms of novel gene origination. Apart from the pioneering idea of gene duplication [6], there are other mechanisms by which new genes are born. These include but not limited to exon shuffling or exon "scrambling" (a) [4,14-18]; fission or fusion of genes (b) [1,3,22], horizontal gene transfer between organisms (c) [31-33], *de novo* origination of protein coding genes from previously non-coding sequences (d) [2,3,34-40], retrotransposition by retrotransposons yielding intronless chimeric genes (e) [18-25], transduction of adjacent DNA by transposable elements (f) or may involve a repetitive element mediated DNA level recombination by a non-allelic homologous recombination (NAHR) mechanism (g) [7,20,26-30]. The figure was adapted from [3].

Additional file 2: Multiple alignments of signal sequence trap (SST) isolated cDNAs showing similar signal sequence. The similar promoter region including the signal sequence is boxed. The two arrows indicate the 'ATG' start positions utilized in the transcript ORF of the candidate mRNA sequences.

Additional file 3: Phylogenetic tree of the genus *Schistosoma* showing the possible origination point of new duplicated genes. The species phylogeny was adapted from [53] as inferred from DNA sequencing, comparative molecular genomics and karyotyping. This phylogenetic tree was manually simulated and thus the length of the branches does not estimate dates or time scale. The tree shows the *S. japonicum* clade and a representative each of the other clades in the genus including the species that reinvaded Asia from Africa. See review in [53]. Based on the result of the southern hybridization in Figure 2, the species and strains that contain the duplicated genes encoding products with similar signal sequence are colored green and we inferred that the most probable time point estimate (black dot) of the gene's emergence could be after the other species in the *S. japonicum* group (in parenthesis) have diverged.

Additional file 4: Expected fragments on restriction map of genome scaffolds correspond to bands on southern blots. To confirm dispersed duplication hypothesis and to exclude the possibilty of overlapping among the loci, the restriction map of six of the genome scaffolds bearing duplicated loci were generated (A). Using same restriction endonuclease enzymes as in the generated maps, we performed southern hybridization using restriction digested genomic DNA from *S. japonicum* species and strains, and were able to match the expected fragment sizes with the observed bands on the hybridization blots. The contigs and the expected probe binding sites were labeled followed by their sequence ranges. We denoted the respective restriction digested fragments with probe binding site using alphabets with their expected restriction digestion product sizes in parenthesis (*E* + *E*: *Eco*RI + *Eco*RV; *E* + *H*: *Eco*RI + *Hind*III; *B* + *H* = *Bam*HI + *Hind*III). As shown in (B), we were able to match the expected fragment with the southern blot bands, labeled using their corresponding alphabetic codes. Probe binding site on the positive strand were colored 'green' while the antisense sites were colored 'red'. The tiny vertical lines on the graphics represent the cutting sites of the selected restriction enzymes. The restriction map and the image were generated using DNADynamo sequence analysis software.

Additional file 5: Simulations using our raw data to show DNA-Level recombination mediated by REs by NAHR mechanism. The movie created from a Powerpoint presentation (Additional file 10) represents the basic approach we utilized in our analysis to show evidence of DNA level recombination by a non-allelic homologous recombination mechanism. The raw data obtained from BLAST searches and RepBase repetitive element prediction report was used to present a simulation that demonstrates that the duplicated locus is flanked on 5` and 3` ends by retrotransposons of the classes *RTE_SJ* and *Perere* respectively. We proposed that these repetitive elements could have provided the requisite homologous stretch of DNA that is required for such DNA level recombination. NAHR can be inter-chromosomal, intra-chromosomal, inter-sister chromatid, or intra-chromatid to give rise to disperse duplicates of the intervening genomic locus. This movie was created from an original Powerpoint presentation (Additional file 10)

Additional file 6: A simplified illustration of repetitive element mediated DNA level non-allelic homologous recombination (NAHR). Repetitive elements provide the requisite homologous DNA sequence for DNA level recombination between non-allelic pairs by a NAHR mechanism. NAHR can occur within a chromosome (intra-homologous chromosomal), between chromosomes (inter-chromosomal), between sister-chromatids or within a chromatid to give rise to disperse duplicates of the intervening genomic locus. The figure was adapted from [60]. Also see Additional file 1 for a cartoon of NAHR and other mechanisms of new gene origination, and [20,26-29,60] for review.

Additional file 7: Splicing models of some protein-coding representatives of the young duplicons. Based on gene prediction from the contigs using GeneQuest and GeneMark and alignment of cDNAs to genome sequences using *Spling* program, we married the predicted products to the transcriptome database of this parasite and found that some of the duplicons are able to code for distinct gene products. Some of the transcription loci can encode two mRNA transcript variants. The significance of this was further explored in Figure 6.

Additional file 8: SjCP1084 protein coding mRNA [GenBank: AY570737] and a non-coding transcript [GenBank:FN328299] are products of alternative splicing Based on gene prediction from the contigs using GeneQuest and GeneMark, and alignment of cDNAs to genome sequences using *Spling* program, we observed that two mRNA transcript variants were produced from [GenBank:CABF01020060]. This figure is same as Figure 6 (A) but we have in addition presented the aligned sequence of the two transcripts showing details of alternative splicing. An extra splice site was evolved in the first exon of the non-coding transcript [GenBank:FN328299]. When the splice site is recognized, an ORF encoding SjCP1084 protein coding mRNA [GenBank:AY570737] variant is created.

Additional file 9: SjCP1531 protein coding mRNA [GenBank: AY570742] and a non-coding transcript [GenBank:FN329677] are products of alternative splicing. Based on gene prediction from the contigs using GeneQuest and GeneMark, and alignment of cDNAs to genome sequences using *Spling* program, we observed that two mRNA transcript variants were produced from [GenBank:CABF01023364]. This figure is same as Figure 6 (B) but we have in addition provided the aligned sequence of the two transcripts showing details of alternative splicing. Two extra splice sites were not utilized in the transcription of the non-coding transcript [GenBank:FN329677]. When the splice sites were recognized, exons 5 and 6 of a translatable ORF were created to produce SjCP1531 protein coding mRNA [GenBank:AY570742] variant. Refer to RT-PCR result in Figure 5 (C) where two bands (exact size and sequence as the two variants described above) are seen on the agarose gel electrophoresis image.

Additional file 10: Simulations using our raw data to show DNA-Level recombination mediated by REs by NAHR mechanism. This Powerpoint presentation was used to create the movie in Additional file 5.

Abbreviations

SST: Signal sequence trap; NAHR: Non-allelic homologous recombination; DLR: DNA level recombination; NHEJ: Non-homologous end joining; IR: Illegitimate recombination; RE: Repetitive element; ORF: Open reading frame; WGS: Whole genome shotgun.

Competing interests

The authors declare that they have no competing interests.

Author contributions

ECM participated in the conception and design of the study, in-silico analyses, molecular experiments, data analysis and interpretation and drafted the manuscript. YC carried out the signal sequence trap (SST) and participated in in-silico analyses. MK participated in the design of the study, SST, in-silico analyses, molecular experiments and data interpretation. MNS participated in in-silico analyses, molecular experiments, data interpretation and revised the manuscript. DB participated in molecular experiments and data analyses. MK_2, NH, YC_2 and YO maintained parasite life cycle and participated in molecular experiments. SH participated in data interpretation, supervision and revised manuscript for intellectual content. KH participated in the conception and design of the study, SST, in-silico analyses, data interpretation, revised the manuscript and general coordination. All authors approved final version of the manuscript.

Acknowledgements

We would like to thank Prof. Kenji Hirayama's lab members (Department of Immunogenetics, Institute of Tropical Medicine, Nagasaki University) for insightful discussions. ECM is a recipient of the Japanese Government Ministry of Education, Culture, Sports, Science and Technology (MEXT) PhD fellowship. This study was supported in part by the Global Center of Excellence (GCOE) Program (2008-2011); Grant-in-Aid for 21st century COE program (2003-2008), Nagasaki University; and Grant-in-Aid for Scientific Research B (22406009) and C (23590489) from the Japanese Government Ministry of Education, Culture, Sports, Science and Technology (MEXT). The funding agency played no role in conducting the study, drafting the manuscript and the decision to publish.

Author details

[1]Department of Immunogenetics, Institute of Tropical Medicine (NEKKEN), and Global COE Program, Nagasaki University, 1-12-4 Sakamoto, 852-8523, Nagasaki, Japan. [2]Laboratory on Technology for Parasitic Disease Prevention and Control, Jiangsu Institute of Parasitic Diseases, 117 Yangxiang, Meiyuan, Wuxi 214064, People's Republic of China. [3]Laboratory of Tropical Medicine and Parasitology, Dokkyo Medical University, Tochigi, Japan. [4]Department of Immunology and Parasitology, The University of Occupational and Environmental Health, Kitakyushu, Japan. [5]Department of Parasitology, Institute of Tropical Medicine (NEKKEN), and Global COE Program, Nagasaki University, 1-12-4 Sakamoto, 852-8523, Nagasaki, Japan. [6]Department of Parasitology and Entomology, Faculty of Bioscience, Nnamdi Azikiwe University, P.M.B. 5025, Awka, Nigeria.

Received: 2 February 2012 Accepted: 11 June 2012
Published: 20 June 2012

References

1. Zhou Q, Wang W: **On the origin and evolution of new genes-a genomic and experimental perspective.** *Journal of Genetics and Genomics* 2008, **35**(11):639–648.
2. Cai J, Zhao R, Jiang H, Wang W: **De novo origination of a new protein-coding gene in Saccharomyces cerevisiae.** *Genetics* 2008, **179**(1):487–496.
3. Long M, Betran E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old.** *Nat Rev Genet* 2003, **4**(11):865–875.
4. Long M, Deutsch M, Wang W, Betrán E, Brunet FG, Zhang J: **Origin of New Genes: Evidence from Experimental and Computational Analyses.** *Genetica* 2003, **118**(2):171–182.
5. Kaessmann H: **Origins, evolution, and phenotypic impact of new genes.** *Genome Res* 2010, **20**(10):1313–1326.
6. Ohno S: *Evolution by gene duplication*. New York: Springer-Verlag; 1970.
7. Bailey JA, Liu G, Eichler EE: **An Alu transposition model for the origin and expansion of human segmental duplications.** *Am J Hum Genet* 2003, **73**(4):823–834.
8. Gao X, Lynch M: **Ubiquitous internal gene duplication and intron creation in eukaryotes.** *Proc Natl Acad Sci* 2009, **106**(49):20818–20823.
9. Katju V, Lynch M: **On the formation of novel genes by duplication in the Caenorhabditis elegans genome.** *Mol Biol Evol* 2006, **23**(5):1056–1067.
10. Gu Z, Rifkin SA, White KP, Li W-H: **Duplicate genes increase gene expression diversity within and between species.** *Nat Genet* 2004, **36**(6):577–579.
11. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models.** *Nat Rev Genet* 2010, **11**(2):97–108.
12. Osada N, Innan H: **Duplication and gene conversion in the Drosophila melanogaster genome.** *PLoS genetics* 2008, **4**(12):e1000305.
13. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W: **On the origin of new genes in Drosophila.** *Genome Res* 2008, **18**(9):1446–1455.
14. Long M, Langley CH: **Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila.** *Science* 1993, **260**(5104):91–95.
15. Moran JV, DeBerardinis RJ, Kazazian HH Jr: **Exon shuffling by L1 retrotransposition.** *Science* 1999, **283**(5407):1530–1534.
16. Patthy L: **Genome evolution and the evolution of exon-shuffling–a review.** *Gene* 1999, **238**(1):103–114.
17. Shao X, Shepelev V, Fedorov A: **Bioinformatic analysis of exon repetition, exon scrambling and trans-splicing in humans.** *Bioinformatics* 2006, **22**(6):692–698.
18. Elrouby N, Bureau TE: **Bs1, a New Chimeric Gene Formed by Retrotransposon-Mediated Exon Shuffling in Maize.** *Plant Physiol* 2010, **153**(3):1413–1424.
19. Wang W, Brunet FG, Nevo E, Long M: **Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster.** *Proc Natl Acad Sci U S A* 2002, **99**(7):4448–4453.
20. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** *Nat Rev Genet* 2009, **10**(10):691–703.
21. Courseaux A, Nahon JL: **Birth of two chimeric genes in the Hominidae lineage.** *Science* 2001, **291**(5507):1293–1297.
22. Kaessmann H, Vinckenbosch N, Long M: **RNA-based gene duplication: mechanistic and evolutionary insights.** *Nat Rev Genet* 2009, **10**(1):19–31.
23. Shapiro J: **Transposable elements as the key to a 21st century view of evolution.** *Genetica* 1999, **107**(1):171–179.
24. Okamura K, Nakai K: **Retrotransposition as a Source of New Promoters.** *Mol Biol Evol* 2008, **25**(6):1231–1238.
25. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH Jr: **Retrotransposition of marked SVA elements by human L1s in cultured cells.** *Hum Mol Genet* 2011, **20**(17):3386–3400.
26. Johnson ME, Cheng Z, Morrison VA, Scherer S, Ventura M, Gibbs RA, Green ED, Eichler EE: **Recurrent duplication-driven transposition of DNA during hominoid evolution.** *Proc Natl Acad Sci U S A* 2006, **103**(47):17626–17631.
27. Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, Zhang Y, Zhao R, Brunet F, Peng L, *et al*: **Repetitive Element-Mediated Recombination as a Mechanism for New Gene Origination in *Drosophila*.** *PLoS genetics* 2008, **4**(1):e3.
28. Shapiro JA: **A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering.** *Gene* 2005, **345**(1):91–100.
29. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA: **Emergence of primate genes by retrotransposon-mediated sequence transduction.** *Proc Natl Acad Sci* 2006, **103**(47):17608–17613.
30. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, *et al*: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**(7191):56–64.
31. Eugene V, Koonin KSM, Aravind L: **Horizontal Gene Transfer in Prokaryotes: Quantification and Classification.** *Annu Rev Microbiol* 2001, **55**:709–742.
32. Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S, *et al*: **Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes.** *Science* 2007, **317**(5845):1753–1756.
33. Keeling PJ, Palmer JD: **Horizontal gene transfer in eukaryotic evolution.** *Nat Rev Genet* 2008, **9**(8):605–618.
34. Yang Z, Huang J: **De novo origin of new genes with introns in Plasmodium vivax.** *FEBS Lett* 2011, **585**(4):641–644.
35. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ: **Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression.** *Proc Natl Acad Sci* 2006, **103**(26):9935–9939.
36. Volff J-N: **Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes.** *BioEssays* 2006, **28**(9):913–922.
37. Begun DJ, Lindfors HA, Kern AD, Jones CD: **Evidence for de Novo Evolution of Testis-Expressed Genes in the Drosophila yakuba/Drosophila erecta Clade.** *Genetics* 2007, **176**(2):1131–1137.
38. Knowles DG, McLysaght A: **Recent de novo origin of human protein-coding genes.** *Genome Res* 2009, **19**(10):1752–1759.
39. Wu DD, Irwin DM, Zhang YP: **De novo origin of human protein-coding genes.** *PLoS genetics* 2011, **7**(11):e1002379.
40. Guerzoni D, McLysaght A: **De novo origins of human genes.** *PLoS genetics* 2011, **7**(11):e1002381.
41. Snel B, Bork P, Huynen M: **Genome evolution. Gene fusion versus gene fission.** *Trends Genet* 2000, **16**(1):9–11.
42. Wilson MS, Mentink-Kane MM, Pesce JT, Ramalingam TR, Thompson R, Wynn TA: **Immunopathology of schistosomiasis.** *Immunol Cell Biol* 2006, **85**(2):148–154.
43. Cass CL, Johnson JR, Califf LL, Xu T, Hernandez HJ, Stadecker MJ, Yates JR III, Williams DL: **Proteomic analysis of Schistosoma mansoni egg secretions.** *Mol Biochem Parasitol* 2007, **155**(2):84–93.
44. Edungbola LD, Cha YN, Bueding E, Schiller EL: **Granuloma formation around exogenous eggs of Schistosoma mansoni and Schistosoma japonicum in mice.** *Afr J Med Med Sci* 1982, **11**(2):75–79.
45. Burke ML, Jones MK, Gobert GN, Li YS, Ellis MK, McManus DP: **Immunopathogenesis of human schistosomiasis.** *Parasite Immunology* 2009, **31**(4):163–176.
46. McManus DP, Loukas A: **Current Status of Vaccines for Schistosomiasis.** *Clin Microbiol Rev* 2008, **21**(1):225–242.
47. Yu C, Zhang F, Yin X, Kikuchi M, Hirayama K: **Isolation of the cDNAs encoding secreted and membrane binding proteins from egg of *Schistosoma japonicum* (Chinese strain).** *Acta Parasitologica* 2008, **53**(1):110–114.
48. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Loren van Themaat EV, Brown JKM, Butcher SA, Gurr SJ, *et al*: **Genome Expansion and Gene Loss in Powdery Mildew Fungi Reveal Tradeoffs in Extreme Parasitism.** *Science* 2010, **330**(6010):1543–1546.
49. Feng Liu YZ, Wang Zhi-Qin, Lu Gang, Zheng Huajun, Brindley Paul J, McManus Donald P, Blair David, Zhang Qin-hua, Zhong Yang, Wang Shengyue, Han Ze-Guang, Chen Zhu: **The Schistosoma japonicum genome reveals features of host–parasite interplay.** *Nature* 2009, **460**:345–351 (16 July 2009).
50. Hu W, Brindley PJ, McManus DP, Feng Z, Han Z-G: **Schistosome transcriptomes: new insights into the parasite and schistosomiasis.** *Trends in Molecular Medicine* 2004, **10**(5):217–225.
51. Hu W, Yan Q, Shen DK, Liu F, Zhu ZD, Song HD, Xu XR, Wang ZJ, Rong YP, Zeng LC, *et al*: **Evolutionary and biomedical implications of a Schistosoma japonicum complementary DNA resource.** *Nat Genet* 2003, **35**(2):139–147.
52. Gene DB; [http://www.genedb.org/Homepage/Sjaponicum].
53. Lawton SP, Hirai H, Ironside JE, Johnston DA, Rollinson D: **Genomes and geography: genomic insights into the evolution and phylogeography of the genus Schistosoma.** *Parasit Vectors* 2011, **4**:131.
54. Lockyer AE, Olson PD, Ostergaard P, Rollinson D, Johnston DA, Attwood SW, Southgate VR, Horak P, Snyder SD, Le TH, *et al*: **The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of Schistosoma Weinland, 1858.** *Parasitology* 2003, **126**(Pt 3):203–224.

55. Le TH, Humair PF, Blair D, Agatsuma T, Littlewood DT, McManus DP: **Mitochondrial gene content, arrangement and composition compared in African and Asian schistosomes.** *Mol Biochem Parasitol* 2001, **117**(1):61–71.
56. Chen ST, Cheng HC, Barbash DA, Yang HP: **Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster.** *PLoS genetics* 2007, **3**(7):e107.
57. Szak ST, Pickeral OK, Landsman D, Boeke JD: **Identifying related L1 retrotransposons by analyzing 3' transduced sequences.** *Genome Biol* 2003, **4**(5):R30.
58. Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, *et al*: **High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes.** *The Plant Cell Online* 2006, **18**(8):1791–1802.
59. Thomas JH, Schneider S: **Coevolution of retroelements and tandem zinc finger genes.** *Genome Res* 2011, **21**(11):1800–1812.
60. Sasaki M, Lange J, Keeney S: **Genome destabilization by homologous recombination in the germ line.** *Nat Rev Mol Cell Biol* 2010, **11**(3):182–195.
61. Cordaux R: **The human genome in the LINE of fire.** *Proc Natl Acad Sci U S A* 2008, **105**(49):19033–19034.
62. Laha T, Brindley PJ, Smout MJ, Verity CK, McManus DP, Loukas A: **Reverse transcriptase activity and untranslated region sharing of a new RTE-like, non-long terminal repeat retrotransposon from the human blood fluke, Schistosoma japonicum.** *Int J Parasitol* 2002, **32**(9):1163–1174.
63. Zhang YE, Vibranovski MD, Krinsky BH, Long M: **A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes.** *Bioinformatics* 2011, **27**(13):1749–1753.
64. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154**(1):459–473.
65. Conant GC, Wolfe KH: **Turning a hobby into a job: How duplicated genes find new functions.** *Nat Rev Genet* 2008, **9**(12):938–950.
66. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucleic Acids Res* 2005, **33**(20):6494–6506.
67. Modrek B, Lee C: **A genomic view of alternative splicing.** *Nat Genet* 2002, **30**(1):13–19.
68. Keren H, Lev-Maor G, Ast G: **Alternative splicing and evolution: diversification, exon definition and function.** *Nat Rev Genet* 2010, **11**(5):345–355.
69. Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I: **Primate and rodent specific intron gains and the origin of retrogenes with splice variants.** *Mol Biol Evol* 2011, **28**(1):33–37.
70. Mattick JS, Makunin IV: **Non-coding RNA.** *Hum Mol Genet* 2006, **15**(Spec No 1):R17–R29.
71. NCBI: *Basic Local Alignment Search Tool (BLAST)*: [http://blast.ncbi.nlm.nih.gov/Blast.cgi].
72. NCBI: *UniGene Database*; [http://www.ncbi.nlm.nih.gov/unigene?term=txid6182[organism]].
73. NCBI: *Conserved Domain Architecture Retrieval Tool*: [http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml].
74. *ProDom Database*: [http://prodom.prabi.fr/prodom/current/html/home.php].
75. Corpet F: **Multiple sequence alignment with hierarchical clustering.** *Nucleic Acids Res* 1988, **16**(22):10881–10890.
76. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** *Mol Biol Evol* 2011, **28**(10):2731–2739.
77. Kohany O, Gentles AJ, Hankus L, Jurka J: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.** *BMC Bioinforma* 2006, **7**:474.
78. Southern E: **Southern blotting.** *Nat Protoc* 2006, **1**(2):518–525.

doi:10.1186/1471-2164-13-260
Cite this article as: Mbanefo *et al.*: Origin of a novel protein-coding gene family with similar signal sequence in *Schistosoma japonicum*. *BMC Genomics* 2012 **13**:260.

MALARIA JOURNAL

RESEARCH Open Access

Toll-like receptor 9 (TLR9) polymorphism associated with symptomatic malaria: a cohort study

Ahmeddin H Omar[1], Michio Yasunami[1], Akiko Yamazaki[1], Hiroki Shibata[1], Michael F Ofori[2], Bartholomew D Akanmori[2], Mohammed Nasir Shuaibu[1], Mihoko Kikuchi[1,3] and Kenji Hirayama[1*]

Abstract

Background: In areas mesoendemic for malaria transmission, symptomatic individuals play a significant role as reservoirs for malaria infection. Understanding the pathogenesis of symptomatic malaria is important in devising tools for augmenting malaria control. In this study, the effect of TLR9 polymorphisms on susceptibility to symptomatic malaria was investigated among Ghanaian children.

Methods: Four hundred and twenty nine (429) healthy Ghanaian children, aged three to eleven years (3–11 years), were enrolled into a cohort study and actively followed up for symptomatic malaria for one year. Four TLR9 single nucleotide polymorphisms (SNPs) namely: rs187084 (C-1486 T), rs5743836(C-1237 T), rs352139 (G + 1174A) and rs352140 (G + 2848A) were genotyped by direct sequencing, and their attributable and relative risks for symptomatic malaria determined. TLR9 haplotypes were inferred using the PHASE software and analysed for the risk of symptomatic malaria. A luciferase assay was performed to investigate whether the TLR9 haplotypes influence TLR9 promoter activity.

Results: The rs352139 GG genotype showed a significantly increased relative risk of 4.8 for symptomatic malaria ($P = 0.0024$) and a higher mean parasitaemia ($P = 0.04$). Conversely, the rs352140 GG genotype showed a significantly reduced relative risk of 0.34 ($P = 0.048$). TLR9 haplotypes analyses showed that TTAG haplotype was significantly associated with reduced relative risk of 0.2 for symptomatic malaria ($P = 4\times10^{-6}$) and a lower mean parasitaemia (0.007), while CTGA haplotype had an increased relative risk of 3.3 ($P = 0.005$). Functional luciferase reporter gene expression assay revealed that the TTA haplotype had a significantly higher promoter activity than the CCG, CTG and TCG haplotypes.

Conclusions: Taken together, these findings indicate a significant association of TLR9 gene polymorphisms with symptomatic malaria among Ghanaian children in Dangme-West district.

Keywords: Cohort study, TLR9, Symptomatic malaria, Genetic susceptibility, Genetic polymorphism, Haplotype, Luciferase promoter assay

Background

Despite the tremendous achievements in malaria control over the past few years, malaria still remains a major public health problem in the endemic countries. According to the WHO world malaria report, there were an estimated malaria cases and deaths of 225 million and 781,000, respectively, in the year 2009 [1]. Malaria has a wide spectrum of infections that ranges from asymptomatic malaria, symptomatic (mild) malaria, to severe (complicated) malaria infections. Although severe malaria, which constitutes 1-2% of all malaria cases, is responsible for most of the malaria deaths, symptomatic malaria accounts for the major burden of the disease's morbidity [2]. Moreover, while asymptomatic gametocyte carriers are the major transmission reservoir in areas hyperendemic for malaria [3], symptomatic individuals play an increased role as reservoir for malaria

* Correspondence: hiraken@nagasaki-u.ac.jp
[1]Department of Immunogenetics, Institute of Tropical Medicine (NEKKEN) and Global COE Program, Nagasaki University, 1-12-4 Sakamoto, Nagasaki, 852-8523, Japan
Full list of author information is available at the end of the article

infections in areas with low intensity of malaria transmission [4].

Among the factors that have been postulated to determine the clinical outcome of malaria infections are host genetic make-up and immunity [5]. The contribution of genetics in susceptibility to malaria has been well documented, with the sickle cell trait, haemoglobin C variants and glucose-6-phosphate dehydrogenase deficiency conferring protection against severe malaria [6]. On the other hand, α-thalassaemia shows a more complex effect of increasing susceptibility to mild malaria while conferring protection against severe malaria [7]. Thus, it is possible that the mechanism responsible for protection against mild and severe malaria may be different from each other. This hypothesis has been reinforced by the fact that in malaria endemic areas, immunity to severe malaria is achieved at a time when susceptibility to mild malaria is still high [8].

Toll like receptors (TLR) are innate immune receptors that recognize pathogen associated molecular patterns and trigger activation of the intracellular signal cascade that induces transcription of inflammatory cytokines, type 1 interferons and chemokines [9]. In addition, stimulation of TLRs also leads to dendritic cell maturation and induction of adaptive immune response [10]. TLR2 and TLR4 have been reported to recognize *Plasmodium falciparum* glycosylphosphatidylinositol (GPI) [11], while TLR9 has been reported to recognize malaria haemozoin [12] and/or *Plasmodium* DNA-bound haemozoin [13].

TLR9-encoding gene is located on chromosome 3p21.3 and spans approximately 5 kb. It consists of 2 exons and encodes 1032 amino acids [14]. The TLR9 gene has been implicated in pathogenesis of severe malaria both in murine model and in humans. In animal studies, it has been demonstrated that mice deficient in TLR9 survived cerebral malaria better than the wild type mice [15] and inhibition of TLR9 activations by TLR9 agonist conferred protection against cerebral malaria in mice [16]. A number of single nucleotide polymorphisms (SNPs) for TLR9 have been identified, and four of these SNPs; rs187084, rs5743836, rs352139 and rs352140, have been reported to show high heterozygocity among three major US ethnic groups [17]. Several studies have focused on the relationship of TLR9 polymorphisms and severe malaria phenotypes, with some reporting positive association. For example, in a study in Ghana on malaria in pregnancy (n = 304), the TLR9 promoter polymorphism rs187084 C allele was associated with increased risk of low birth weight among term infants [18], while a Brazilian study among adults (n = 304) with mild malaria associated the promoter polymorphisms rs187084 C allele and rs5743836 C allele with high parasitaemia (>10,000 parasite/μl) [19]. However, these effects of TLR9 promoter polymorphisms on severe malaria phenotypes were not replicated in a large family and population-based association study from Malawi and Gambia (n > 6000) that found no convincing association of the four common TLR9 SNPs with severe malaria [20]. On the other hand, few studies have focussed on the effect of TLR9 polymorphisms on susceptibility to mild malaria. Two studies, carried out in Brazil [19] and Iran [21], reported no influence of TLR9 promoter polymorphism on susceptibility to mild malaria in their respective populations. However, the effect of TLR9 polymorphisms on mild malaria among the African population is not well understood.

In the present study, healthy Ghanaian children living in an area mesoendemic for malaria transmission were recruited into a prospective cohort study for one year, and the effect of the TLR9 SNPs on susceptibility to symptomatic malaria investigated.

Methods

Study area and subjects

Details about the study area, population, set-up of the cohort and the follow-up protocol has previously been described [22]. In brief, the cohort study was conducted at Asutuare, a sub-district of Dangme-West District in the Great Accra region of Ghana. The district has two rainy seasons in a year, April to July and October to December, and consequently malaria transmission is seasonal, with the peak transmission coinciding with the period of the major rainy seasons while the dry seasons having low malaria transmission [23]. It is estimated that individuals in Dodowa, (the district headquarter for Dangme-West district), are exposed to about 20 infective bites per year, and 98% of the infections are due to *P. falciparum* [23]. Four hundred and twenty nine (429) healthy Ghanaian children, aged 3–11 years, were enrolled into a one-year prospective cohort study, from June 2007 to July 2008, which spanned across three (3) rainy seasons. Only one child per household was recruited in order to avoid inclusion of closely related individuals. The study participants were actively followed up for clinical malaria symptoms with regular home visits of two-week intervals. During the visits, data on the health status of the participants for the previous two weeks were collected using a standard questionnaire and their body temperatures measured. Individuals suspected of clinical malaria were referred to the community health centre where medical examinations and blood smear for malaria were carried out and treatment given according to the recommendation of the Ghanaian Ministry of Health.

Phenotype definition

In this study, a participant was considered to be suffering from symptomatic mild malaria, if he or she had a

temperature that is greater than or equal to 38.0°C, with a parasite load that is greater than or equal to 5,000 parasite per microliter (μl) of blood. Informed consent was obtained from each participant's parents/guardian after a detailed explanation of all procedures in the study. The study was approved by the Ghanaian Ministry of Health, Institutional Review Board (IRB) of Noguchi Memorial Institute of Medical Research, University of Ghana and IRB of Institute of Tropical Medicine, Nagasaki University.

DNA extraction and polymerase chain reaction (PCR)

Sample collection and storage has been described in detail elsewhere [22]. Genomic DNA was extracted from blood sample using QIAamp DNA blood mini kit (Qiagen, Tokyo, Japan).

The PCR amplification for the rs187084, rs5743836, and rs352139 were carried out using the forward primer 5′-CTGTGGACATCGATATCGGTGT-3′ (PF3) and reverse primer 5′-AAGCTTCGCTGCGGCAGAA ACCCTGT-3′ (i1R), while for rs352140; forward primer 5′-TCTAGACATCATGCTGGCCATGACC-3′ (2 F8) and reverse primer 5′-CAGAGCCACTCAACAGTG GACT-3′ (2R5) were used. Each PCR reaction contained 5.0 μl of 2xKOD FX buffer, 2.0 μl of 2 mM dNTP, 0.4 μl of forward primer, 0.4 μl of reverse primer, 0.2 μl of KOD Polymerase, 1.5 μl of water, and 0.5 μl of DNA and the PCR conditions used were as follows: One cycle at 95°C for 5 min, 35 cycles at 95°C for 1 min, 35 cycles at 64°C for 1 min, and one cycle at 72°C for 4 min.

SNP genotyping

The TLR9 SNPs studied were identified by analysis of direct nucleotide sequencing of PCR products. The sequencing was carried out using the 3730 DNA Analyzer Applied Biosystems sequencing machine with the sequencing primers 5′-GGGTGTACA TAATTCAGCAG-3′, 5′-GGCAAAGGAGCTCAGGA GTG-3′, 5′-GGAAGAACTTCTGCAGGTAG-3′, and 5′-GGAGAAGGTCTGGCTGCAG-3′ for rs187084, rs5743836, rs352139 and rs352140, respectively. The sequencing data was analysed using sequence scanner version 1.0, Applied Biosystems.

Plasmid construction for luciferase gene reporter assay

Using the PCR primers PF3 and i1R (named above) a DNA fragment, nucleotide positions −2268 to +1233 (NM_17442), which encompassed the promoter, exon 1 and intron 1 regions of the TLR9 gene was amplified from genomic DNA of four selected study subject. The four subjects were homozygous carriers of TTA, CTG, TCG, and CCG for SNPs at positions −1486, -1237 and +1174, respectively. Topo cloning of the PCR products were carried out using Zero blunt® Topo® cloning kit and transformed into One Shot® TOP10 Competent cells. Positive clones were selected, plasmid extracted and presence of the expected fragment confirmed by restriction digestion analysis and DNA sequencing. The cloned fragment was then treated with BamHI, HindIII and BclI restriction enzymes to obtain a 2754-bp DNA fragment, nucleotide position (−1521 to +1233). This fragment was then cloned into BglII and HindIII restriction sites of the firefly luciferase reporter vector pGL4.10 (Promega, Madison, WI) resulting in construction of four (4) promoter-intron plasmids (Figure 1a). The presence of the insertions and their orientation were confirmed by restriction digestion analysis and DNA sequencing.

Transient transfection assay

THP-1 (human acute monocytic leukaemia) cells were obtained from the American type culture collection (ATCC®) and cultured as per their instructions in RPMI-1640 medium supplemented with 10% foetal calf serum (FCS) and Antibiotic-Antimycotic, 100X (AAS) (Gibco, Life technologies). After more than 90% confluence was achieved in a cell culture flask, 1×10^6 cells/well were then seeded into a 24-well plates; and transiently transfected with 2 μg of each of the four TLR9 plasmid constructs (designated as TTA, CCG, CTG and TCG promoter-intron plasmids) and 2 μg of an empty pGL4.10 vector (promega) as a negative control. All the transfections were performed using Fugene HD® (Promega) according to the manufacturer's recommendations. After 48 h of incubation, the cells were transferred from the 24-well plate to 1.5 ml micro-centrifuge tubes, centrifuged at 200 rpm for 5 min and washed with 1 ml of phosphate buffered saline (PBS). The cells were lysed by incubation with passive lysis buffer (promega) for 15 min at room tempreture, harvested and cleared by centrifugation at 15,000 rpm for 3 min. Luciferase activity of three independent transfections were determined in wallac 1420 multilabel counter® (Perkin Elmer) using luciferase assay systems (promega). All transfections were performed in triplicate.

Statistical analysis

Data were analysed with GraphPad prism (Software version 5.00, Inc; San Diego California USA). TLR9 SNPs genotype and allele frequencies were calculated by direct counting. Consistency of genotype frequency with Hardy-Weinberg Equilibrium (HWE) was determined by comparing the observed number of different genotypes with those expected under the HWE for the estimated allele frequency [24].

The incidence, attributable risk, relative risk and 95% confidence intervals (CIs) were calculated and their significances tested using t-statistics. The p value of <0.05 was considered to be significant after Bonferroni's corrections were made for multiple testing. Briefly, the incidence of

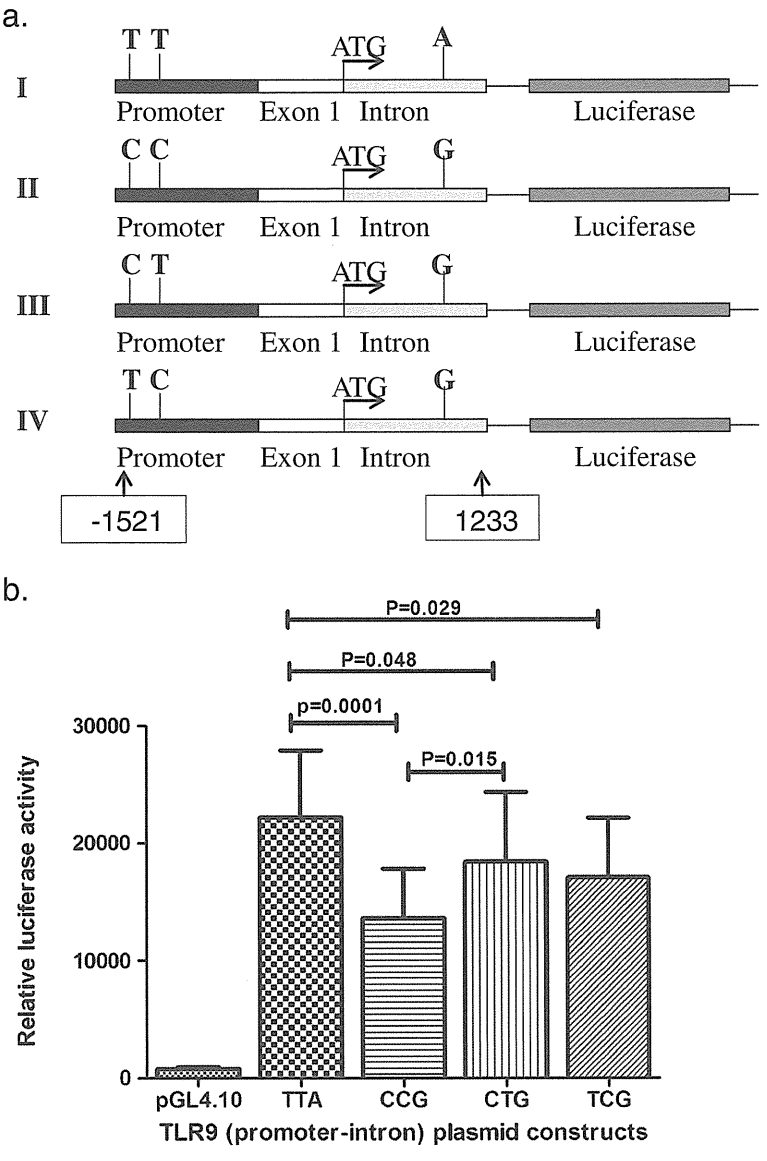

Figure 1 Luciferase reporter activity of TLR9 (promoter-intron) plasmid constructs. a), schematic of four reporter gene constructs that contains TLR9 promoter- intron region with TTA, CCG, CTG and TCG at position −1486, -1237, and +1174 polymorphic sites, respectively. **b**), luciferase expression of the four constructs in THP-1 cells and pGL4.10 as negative control. The luciferase activity levels are mean values of three (3) independent experiments that were all carried out in triplicate. Anova test and t-test were used to test for statistical significance. P value of <0.05 was considered to be statistically significant.

symptomatic malaria was calculated by counting the number of events in the adjusted observation period (person per year) for each genotype or haplotype. The attributable risk was calculated by subtracting the incidence of symptomatic malaria in non-carriers of a genotype/haplotype from carriers of the genotype/haplotype, while the relative risk was calculated by dividing the incidence of symptomatic malaria for a genotype/haplotype with that of non-carriers of the genotype or haplotype as previously described [25]. Statistical significance of attributable risk was evaluated by assuming null hypothesis, that attributable risk follows normal distribution N(0, SE), where SE was estimated by the formula described in reference [25].

Haplotypes for TLR9 SNPs were estimated from non-phased genotype data with maximization likelihood algorithm by the use of PHASE version 2.1.1 [26,27]. Also, pairwise linkage disequilibrium between each pair of

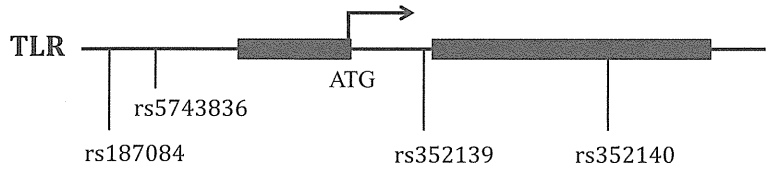

Figure 2 Schematic drawing of the Toll-Like Receptor 9 gene structure and positions of the four TLR9 SNPs. The four SNPs studied are located at position −1486, -1237, +1174 and +2848, respectively. The positions were calculated by taking the A of TLR9 ATG start codon as position 1 based on Genbank accession no. NM_17442.

SNP was determined using the software Haploview [28]. The heterogeneity of mean parasite density for different genotype carriers was determined by Analysis of variance (ANOVA), then the genotype effect assessed with student's t-test comparing mean parasite density between carriers and non-carriers of each genotype.

Results

TLR9 genotype frequencies

In this study, four common TLR9 SNPs, (two located in the promoter region, one intronic and one in the coding region of the second exon (Figure 2)), were genotyped by direct sequencing for the 429 Ghanaian cohort children who were successfully followed up for one year. The minor allele frequencies for rs187084 (C-1486 T), rs5743836(C-1237 T), rs352139 (G + 1174A) and rs352140 (G + 2848A) were found to be 0.35, 0.38, 0.41 and 0.25, respectively, in our study population. The genotype distributions of all the SNPs examined did not show significant deviation from HWE.

Linkage disequilibrium

Pairwise linkage disequilibrium (LD) for the four SNPs was analysed using haploview software and a weak LD was observed between the SNPs rs352139 and rs352140 ($D^1 = 1$, $r^2 = 0.237$); rs352139 and rs5743836 ($D^1 = 0.857$, $r^2 = 0.319$); and rs352139 and rs187084 ($D^1 = 0.846$, $r^2 = 0.277$) (Table 1).

Table 1 The strength of pairwise linkage disequilibrium for the four TLR9 SNPs analysed in the study

r^2 D^1	rs352140	rs352139	rs5743836	rs187084
rs352140		**0.237**	0.032	0.175
rs352139	**1.0**		**0.319**	**0.277**
rs5743836	0.4	**0.857**		0.014
rs187084	0.535	**0.846**	0.124	

D^1 is the measure for deviation of the observed haplotype frequency from the expected frequency, with D^1 =1 showing that the examined loci are completely dependent on one another. r^2 is the correlation coefficient between pair of two loci and $r^2 > 0.4$ showing a meaningful association while $r^2 > 0.2$ being suggestive.

Effect on incidence of symptomatic malaria

Assessment of the effect of TLR9 genotypes on the risk of symptomatic mild malaria was carried out by calculating the incidence of malaria, the relative risk and attributable risk for each TLR9 genotype studied (Table 2). The assessment revealed that the GG genotype for the rs352139 (G + 1174A) intronic SNP increased the attributable risk for malaria by 0.116 events per person per year, with an incidence of 0.146 vs. 0.03 for rs352139 GG vs. non-GG carriers, respectively, and a significant relative risk of 4.8 even after Bonferroni's correction for multiple testing ($p_{corrected} = 0.0024$) (Table 2). In contrast, the rs352140 (G + 2848A) GG genotype decreased the attributable risk by 0.076 events per person per year, with an incidence of 0.04 vs. 0.116 for rs352140 GG vs. non-GG carriers, respectively, and a significant relative risk of 0.34 ($p_{corrected} = 0.048$) (Table 2). The promoter polymorphisms rs187084 (C-1486 T) and rs5743836(C-1237 T) did not show a significant influence on the incidence of symptomatic malaria in the study population (Table 2).

Influence on parasitaemia

The influence of TLR9 polymorphisms on parasitaemia was evaluated by comparing the mean parasite densities at the point of febrile episodes among carriers of each genotype. It was observed that individuals with rs352139 (G + 1174A) GG genotype had a significantly higher mean parasitaemia level compared with non-GG (GA + AA) genotype (19,376 vs. 6,086 parasite/μl, respectively, $p = 0.01$) (Figure 3a), while individuals with the rs5743836 (C-1237 T) TT genotype had a significantly lower mean parasite load compared with the non-TT (TC + CC) genotype (5,501 vs. 16,973 parasite/μl, respectively, $p = 0.03$) (Figure 3)b. No effect on mean parasitaemia was detected for the TLR9 SNPs rs187084 (C-1486 T) and rs352140 (G + 2848A) (Figure 3c and 3d).

TLR9 haplotype analysis

Eleven (11) TLR9 haplotypes were inferred using the Phase software and their estimated haplotype frequencies

Table 2 Incidence, relative risk and attributable risk of mild malaria among carriers of TLR9 SNP's genotypes

[a]SNPs	Incidence	Relative risk	Attributable risk	[b]CI	*p* value	[c]p_c
rs187084(C-1486 T)						
TT vs. CC + TC	0.04	0.37	−0.061	(−0.11)-(−0.01)	**0.01**	ns
CC vs. TT + TC	0.20	3.7	0.144	0.03-0.26	**0.01**	ns
rs5743836(C-1237 T)						
TT vs. CC + TC	0.04	0.38	−0.060	(−0.11)-(−0.01)	**0.01**	ns
CC vs. TT + TC	0.09	1.29	0.020	(−0.05)-(0.09)	0.6	ns
rs352139(G + 1174A)						
GG vs. AA + GA	**0.15**	**4.8**	**0.116**	**0.05-0.18**	**0.0003**	**0.0024**
AA vs. GG + GA	0.03	0.3	−0.055	(−0.10)-(−0.01)	0.03	ns
rs352140 (G + 2848A)						
GG vs. AA + GA	**0.04**	**0.3**	**−0.076**	(−0.13)-(−0.02)	**0.006**	**0.048**
AA vs. GG + GA	0.11	1.63	0.044	(−0.08)-(0.17)	0.5	ns

[a]SNPs: Single nucleotide polymorphisms; [b]CI: confidence intervals; [c]p_c = corrected *p* value by bonferroni method by factor of 8; ns: not statistically significant. The incidence shown is for the genotype highlighted in bold, the relative and attributable risk shown are for the same genotype when compared with the non-carriers of the genotype. Corrected p value <0.05 was considered to be statistically significant.

in the study population are as follows; TTAG: 39.1%, CTGA: 17.3%, TCGG: 15.1%, CCGG: 14.6%, TCGG: 5.9%, TTGG: 2.6%, CCAG: 2.0%, TTGA: 1.8%, CTGG: 1.1%, TCAG: 0.2% and CTAG: 0.2% (for rs187084 (C-1486 T), rs5743836(C-1237 T), rs352139 (G + 1174A) and rs352140 (G + 2848A), respectively). The major haplotypes with frequencies greater than 5% were considered for further analysis on their possible effect on the risks for symptomatic malaria. The result revealed that the TTAG haplotype significantly reduced the attributable risk by 0.115 events per person per year, with an incidence of 0.03 vs. 0.147 for carriers of TTAG vs. non-TTAG haplotypes, respectively; and a significant relative risk of 0.22 ($p_{corrected} = 4\times10^{-6}$) (Table 3). Conversely, the CTGA haplotype significantly increased the attributable risk by 0.103 events per person per year, with an

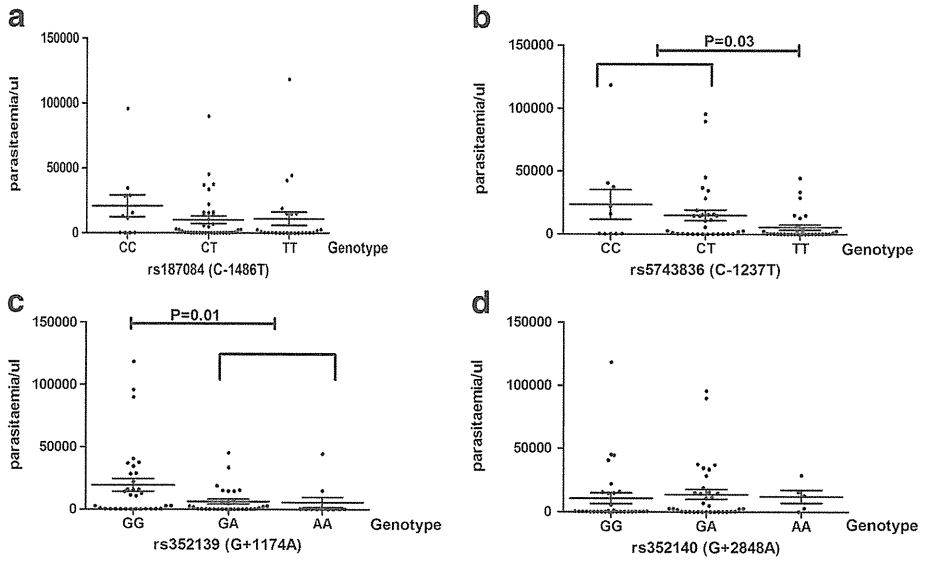

Figure 3 The distribution of mean parasitaemia for carriers of the TLR9 SNP genotypes 3a): The mean parasite load for the rs352139 (G + 1174A) GG, GA and AA genotypes are 19,376; 6,268; 5,589 parasite/μl of blood. The *GG genotype was significantly associated with higher parasitaemia compared with the non-GG genotype (19,376 vs. 6,086 parasite/μl of blood, p = 0.01). **3b):** The mean parasite load for the rs5743836 (C-1237 T) CC, CT and TT genotypes are 23,532; 14,924; 5,501 parasite/μl of blood, respectively. The *TT genotype was significantly associated with lower parasitaemia compared with the non-TT genotype (5,501 vs. 16,973 parasite/μl of blood, p = 0.03). **3c):** The mean parasite load for the rs187084 (C-1486 T) CC, CT and TT genotypes are 20,691; 10,089; 11,117 parasite/μl of blood, respectively. **3d).** The mean parasite load for the rs352140 (G + 2848A) GG, GA and AA genotypes are 10,738; 13776; 12088 parasite/μl of blood. Statistical analysis was performed using ANOVA and student's t-test. P value of <0.05 was considered to be statistically significant.

Table 3 Incidence, relative risk and attributable risk of mild malaria among carriers of TLR9 haplotypes

[a]Haplotype	Incidence	Relative risk	Attributable risk	[b]CI	*p* value	[c]p_c
TTAG vs. non-TTAG	**0.03**	**0.2**	**−0.115**	**(−0.16)-(−0.07)**	**$8x10^{-7}$**	**$4x10^{-6}$**
CTGA vs. non-CTGA	**0.15**	**3.3**	**0.103**	**0.04-0.16**	**0.001**	**0.005**
TCGG vs. non TCGG	0.10	1.5	0.032	(−0.03)-(0.09)	0.3	ns
CCGG vs. non-CCGG	0.12	2.1	0.063	0.0-0.1	0.05	ns
TCGA vs. non-TCGA	0.09	1.2	0.012	(−0.07)-(0.09)	0.8	ns

[a]Haplotype for rs187084 (C-1486 T), rs5743836 (C-1237 T), rs352139 (G + 1174A) and rs352140 (G + 2848A), respectively; [b]CI: confidence intervals; [c]p_c = corrected *p* value by Bonferroni's method by a factor of 5; ns: not statistically significant. The incidence shown is for the haplotype highlighted in bold, the relative and attributable risk shown are for the same haplotype when compared with the non-carriers of the haplotype. Corrected p value <0.05 was considered to be statistically significant.

incidence of 0.147 vs. 0.044 for carriers of CTGA vs. non-CTGA haplotypes, respectively; and a significant relative risk of 3.3 ($p_{corrected} = 0.005$) (Table 3). The other three common haplotypes (TCGG, CCGG and TCGA) were not observed to significantly influence the risk for symptomatic malaria in our cohort study.

On analysing the effect of TLR9 haplotypes on parasitaemia, it was observed that the TTAG haplotype was significantly associated with lower mean parasitaemia, with carriers of TTAG having a mean parasite load of 6,178 parasite/μl compared with 20,114 parasite/μl of the non-TTAG carriers (p = 0.007) (Figure 4). No significant effect on parasitaemia was observed for the haplotypes CTGA, TCGG, CCGG and TCGA in our study population (Figure 4).

Luciferase reporter gene assay

To further examine the functional relevance of the haplotypes in terms of the TLR9 gene transcription, four promoter plasmids containing the TLR9 promoter, exon 1 and intron 1 regions were prepared (Figure 1a) and transfected into THP-1 cells. The four constructs, represents the five major TLR9 haplotype in our population, with TTA, CTG, TCG, and CCG SNPs at positions −1486, -1237 and +1174, respectively (Figure 1a). All the four constructs consistently showed a significant higher relative luciferase activity than the negative control (empty pGL4.10 vector). It was observed that the TTA promoter plasmid has a significantly higher luciferase activity compared with all the other three promoter plasmids, i.e. the CCG, CTG and TCG (P = 0.0001, P = 0.048 and P = 0.029, respectively) (Figure 1b). It was also observed that the CTG promoter plasmid had a significantly higher luciferase activity than the CCG promoter plasmid (P = 0.015) (Figure 1b).

Discussion

The TLR9 polymorphisms have been postulated to have a cis-regulatory effect on TLR9 expression [20] and also shown to alter cytokine levels during severe malaria infections [29]. In this study, the effect of TLR9 gene polymorphisms on symptomatic malaria was investigated and TLR9 polymorphisms and haplotypes were significantly associated with susceptibility to symptomatic malaria among Ghanaian children.

In this cohort study, the intronic polymorphism rs352139 (G + 1174A) GG genotype was significantly associated with increased risk for symptomatic malaria and high parasitaemia. Previous functional study has reported that this intronic SNP has a regulatory effect on TLR9 expression, with the rs352139 G allele in combination with the promoter rs187084 (C-1486 T) C allele having a down-regulatory effect, while the rs352139 A allele in combination with rs187084 T allele having an up-regulatory effect on TLR9 expression among patients with systemic lupus erythematosus [30]. However, it is currently not clear how this intronic SNP induces such a phenotype change. It is possible that it influences signalling by creating an alternative splicing site and thus, affecting the mRNA transcription and the final protein product. Equally, the rs352139 SNP could be a likely marker in LD with a polymorphic regulatory region that controls TLR9 expression or a functional coding region SNP. In contrast to our finding, Campino *et al* [20] reported an association of the rs352139 A allele with severe malaria among Malawian population but not in the Gambian population [20]. This discrepancy may in part be explained by the difference in phenotypes and populations examined, and also by the different role pro-inflammatory cytokines play in the pathogenesis of symptomatic and severe malaria. For example, an early high IFN-γ response has been reported to confer protection against symptomatic malaria episodes [31,32], while an over-production of IFN-γ has been associated with susceptibility to cerebral malaria [33].

The synonymous coding SNP, rs352140 (G + 2848A) GG genotype was associated with protection from symptomatic malaria but no such protective effect was observed for parasitaemia among our cohort population. This rs352140 SNP is linked to the rs352139 (G + 1174A) ($D^1 = 1$, $r^2 = 0.237$) in the study population, with the rs352140 G allele linked to rs352139 A allele.

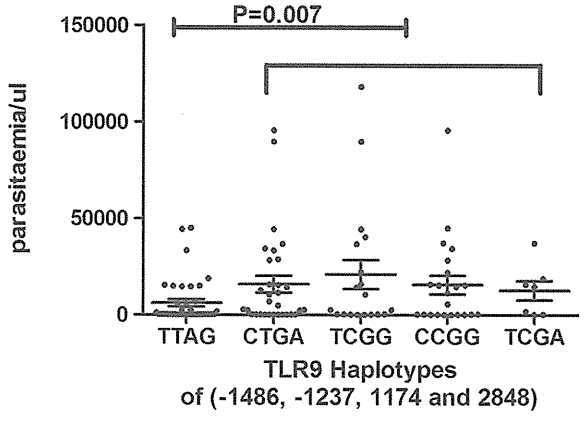

Figure 4 The distribution of mean parasitaemia for carriers of the TLR9 haplotypes. The mean parasite load for the haplotypes are TTAG, CTGA, TCGG, CCGG and TCGA are 6,178; 15,992; 21,034; 15,631; 12,714 parasite/μl of blood. The *TTAG haplotype was significantly associated with lower parasitaemia compared with the non-TTAG haplotype (6,178 vs. 20,114 parasite/μl of blood, p = 0.007). Statistical analysis was performed using ANOVA and student's t-test. P value of <0.05 was considered to be statistically significant.

Thus, the observed protective effect could partly be attributed to the effect of the rs352139 A allele which has been shown to have an up-regulatory effect on TLR9 expression [30].

The promoter polymorphism rs5743836 (C-1237 T) TT genotype was associated with low parasitaemia but no effect on susceptibility to symptomatic malaria was observed in this study. Our finding is consistent with that of Leorrati *et al* [19] who reported an association of rs5743836 TT genotype with low parasitaemia (<10000 parasite/μl) among adults with mild malaria in the Brazilian Amazon [19]. The rs5743836 TT variant has been shown to have a higher promoter activity than the CC genotype [34], and thus, could result in increased pro-inflammatory cytokine production during malaria infection leading to successful control and elimination of malaria parasites.

On haplotype analysis, the TLR9 TTAG (−1486 T, -1237 T, +1174A, +2848 G) haplotype, the most frequent haplotype (39.1%) in the study population, was significantly associated with protection against symptomatic malaria and high parasitaemia. These findings were strongly supported by the result of luciferase reporter gene expression assay which showed a significantly higher promoter activity for the TTA (−1486 T, -1237 T, +1174A) haplotype compared with the CCG, CTG and TCG haplotypes (Figure 1b). It was also observed that the CTG haplotype had a higher promoter activity than the CCG haplotype (Figure 1b). Taken together, the haplotype with A allele for rs352139 (TT**A**) consistently had a significantly higher promoter effect than all the haplotypes with G allele for the same SNP (i.e. CC**G**, CT**G** and TC**G**). It was also observed that the haplotypes with T allele for rs5743836 (i.e. TT**A**, CT**G**) had a higher promoter effect than those with C allele for the same SNP (i.e. C**C**G, T**C**G) and that carriers of the T**TA** haplotype had the highest promoter activity, the C**TG** haplotype had an intermediate activity, and the C**CG** having the lowest promoter activity. Thus, it is possible that these two SNPs (the rs5743836 (C-1237 T) and rs352139 (G + 1174A)) have a greater role of influencing TLR9 gene transcription and expression. This luciferase assay result is in agreement with the findings of Tao *et al* 2007, who reported that the rs187084 (C-1486 T) T allele in combination with rs352139 (G + 1174A) A allele had a higher TLR9 expression compared with the rs187084 C allele in combination with rs352139 G allele among Japanese population who are not polymorphic for rs5743836 (C-1237 T) (i.e. TTA had higher TLR9 expression than CTG) (31). Therefore, it can be inferred that the TTAG haplotype may exert an increased TLR9 transcriptional activity resulting in higher TLR9 expression and subsequently higher pro-inflammatory cytokine production, thus conferring protection against symptomatic malaria phenotype. The TTAG haplotype has also been associated with increased risk for meningococcal meningitis among Dutch children [35] and also associated with protection from ulcerative colitis among the Japanese population [36]. The different effects of TLR9 haplotypes in different disease phenotypes could be explained by the different roles inflammatory cytokines play in pathogenesis of these diseases. While robust pro-inflammatory cytokines has been reported to provide protection against clinical mild malaria [31] and against experimental colitis