Evidence of Massive Horizontal Gene Transfer Between Humans and *Plasmodium vivax*

Daniel Z. Bar

Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem,
Jerusalem 91904, Israel. Tel: +972-2-6585981 E-mail: daniel.bar@mail.huji.ac.il

Nature Precedings : doi:10.1038/npre.2011.5690.1 : Posted 16 Feb 2011

Abstract

The horizontal transfer of DNA between different organisms is a major force shaping the genomes of prokaryotes, but is considered to have a minor role in eukaryotes, with only a handful of known examples, mostly of limited size. The nucleotide databases of *Plasmodium* genomes were divided into small fragments and compared to human, as well as to other *Plasmodium* genomes. This computational approach revealed that the *Plasmodium vivax* genome is interlaced with multiple DNA fragments that were likely acquired via horizontal transfer from humans. Contamination is a major concern in such studies; moreover, it must be determined if the identified homologies might be due to chance. These reservations are supported by the fact that the identified homologous sequences were found to be predominantly within short contigs. Re-sequencing of candidate sites using distinct isolations of *P. vivax* genomic DNA showed deletions not found in the human genome, and with much greater similarity to the *P. vivax* than human genome. Moreover, the identified fragments were enriched for mRNA coding sequences and genes that are known to be functionally important for *P. vivax*, including nitric oxide synthase 1 (neuronal) adaptor and Interleukin-1 family, suggesting a functional role. These results are important for two reasons. First, a directional massive horizontal transfer of genetic material from humans to another eukaryote is shown for the first time. This sheds light on parasite evolution, co-adaptation and immune evasion. Second, the DNA found is enriched for Interleukin-1 family, which is known to be essential for malaria protection. This indicates a functional role and might serve to better understand how *Plasmodium vivax* and the immune system interact.

Introduction

The transfer of DNA between different organisms, horizontal gene transfer (HGT), has long been recognized as a major force shaping the genomes of prokaryotes. In eukaryotes however, HGT have been reported only in a few incidences and on a limited scale. These include transfer to Rotifera from multiple sources, bacteria to fungal events, passage of transposons between parasitic bugs and there vertebrate host, the acquisition of carotenoid biosynthesis enzymes by aphids from fungus (Gilbert et al.; Moran et al.; Hall et al. 2005a; Gladyshev et al. 2008), and some controversial human to *Plasmodium* HGT events (Deitsch et al. 2001a; Deitsch et al. 2001b; Striepen et al. 2002; Striepen et al. 2004; Templeton et al. 2004). In the latter, disagreement arose either due to suspected host DNA contamination or lack of clarity as to the mechanism causing parasite-host sequence similarity (Pain et al. 2008). The recent high coverage sequencing of multiple *Plasmodium* genomes, alongside with that of humans and other mammals, enables the systematic exploration of putative HGT events (PHE) at a whole genome scale (Institute; Lander et al. 2001; Venter et al. 2001; Carlton et al. 2002; Waterston et al. 2002; Hall et al. 2005b; Carlton et al. 2008; Aurrecoechea et al. 2009).

The *Plasmodium* genus is a mosquito borne parasitic protozoa and the causative agent of human malaria. Over two hundred *Plasmodium* species are known, infecting a variety of hosts, from reptiles to mammals, usually in a species specific manner. Only a small fraction of all *Plasmodium* species can infect humans, among these *Plasmodium falciparum,* the cause of most infections and deaths; *Plasmodium vivax*, the second most common cause of infection, that can lay latent in the human liver and is responsible for most causes of recurring (quartan) malaria; and *Plasmodium knowlesi***,** that infects mainly macaques but has been shown to infect humans both naturally and artificially (Fig. 1A). It is herein shown that cross-kingdom HGT from humans is likely to have had a significant, and probably functional, contribution to the *P. vivax* genome.

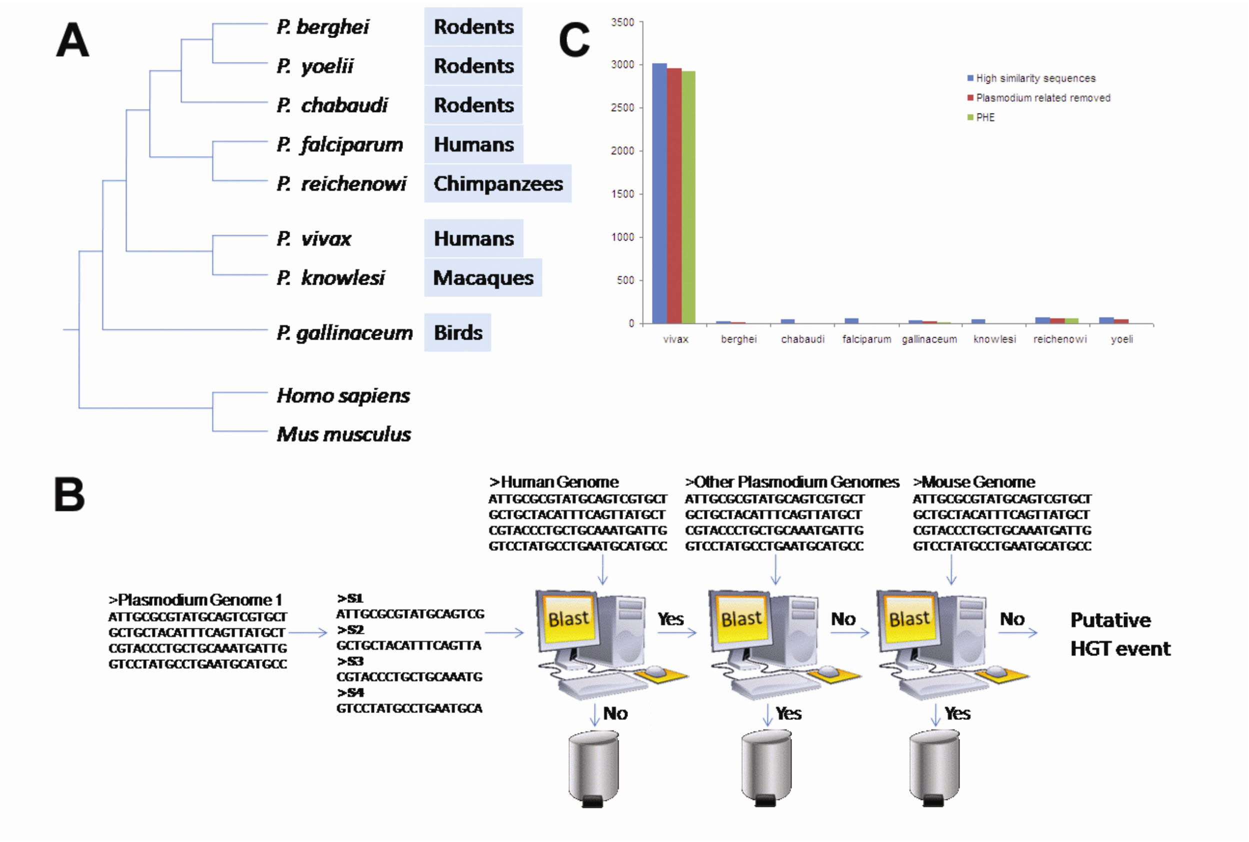

Fig. 1. (**A**) Phylogenetic tree of species discussed. Background colored text indicates typical host. (**B**) Scheme of the comparison model. The genomic sequence of a *Plasmodium* is divided into multiple 60bp sequences. At first these were compared to the human genome, and sequences showing high similarity were kept. Next, sequences that show equal or higher similarity to any of 7 *plasmodium* species were removed. Finally, sequences showing equal or greater similarity to the mouse genome were removed. (**C**) Model results. Blue bars count *Plasmodium* sequences showing e-values < 10^{-7} when searched in the human genome. Red bars count previously found sequences with no closer homolog in any of the other 7 *Plasmodium* species. Green bars count sequences that additionally lack a closer homolog in the mouse genome (PHE).

Results

To determine whether HGT from humans played a role in the recent evolution of some *Plasmodium* species, the genomes of eight *Plasmodium* genome nucleotide databases were divided into 60 letters long fragments and compared to the human genome using BLASTN (Zhang et al. 2000; Morgulis et al. 2008). These genomes were the common human infectious *P. falciparum* and *P.vivax*, the rarely human infectious *P. knowlesi*, the primate infectious *P. reichenowi*, as wells as *P. gallinaceum*, *P. yoelii*, *P. berghei*, and *P. chabaudi*. Fragments showing similarity with expected values (e-value) larger then 10^{-7} were dismissed as coincidental and not pursued further. The *Plasmodium* fragment queries were further compared to the seven additional *Plasmodium* genomes used in this study. By doing so, highly conserved sequences, or sequences that can be accounted for by Mendelian inheritance, were removed. Finally, a comparison to the mouse genome was performed, serving as an aid in delimiting the origin of the HGT sequences (Fig. 1B). The remaining sequences, originating from a *Plasmodium* genome and having close homologs in the human genome, but not in any other *Plasmodium* genome tested, were considered PHE events. All but *P. vivax* resulted in little (N≤60) or no PHE, the former likely to be false positives (See Methods). *P. vivax* resulted in 3009 sequences with close homologs in the human genome, of these 2966 (98.6%) showed greater similarity to the human genome then to any other *Plasmodium* examined. Of the latter, 2921 (98.5%) were closer to the human genome then to the mouse genome (Fig. 1C).

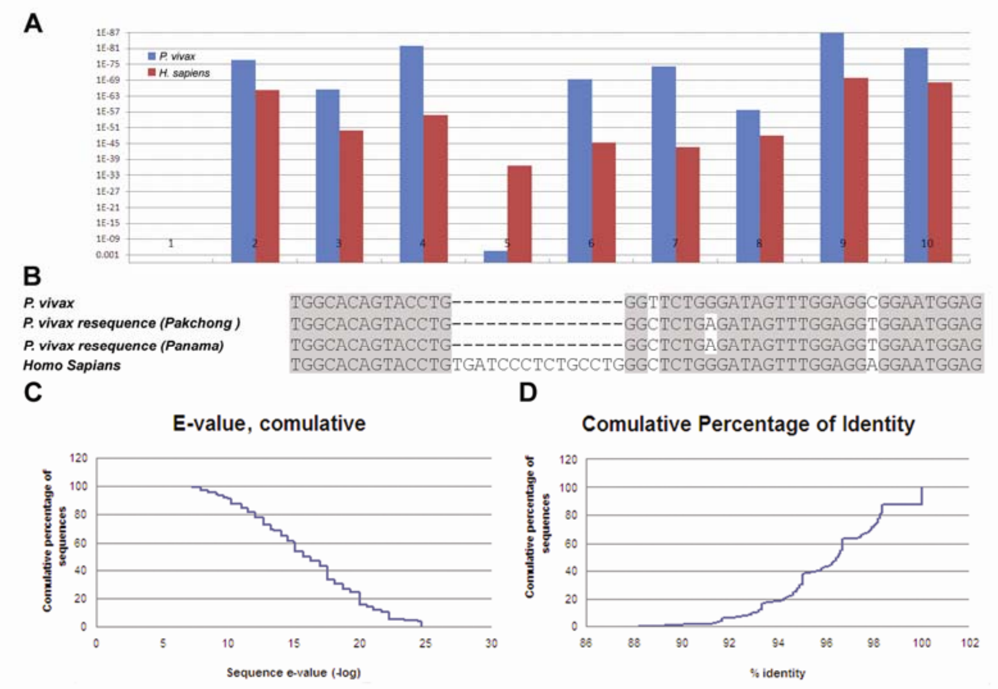

Fig 2. PHE found are unlikely to be a false positive. (**A**) e-values of 10 re-sequenced canditade regions, compared to the *P. vivax* genome (red) and the human genome (blue). The first primer set did not result in a single band; the fifth resulted in a sequence only at the fourth attempt. (**B**) Alignment of the human genome of the known *P. vivax* genome and re-sequencing validation (primer set 2, Supplementary Table 4). (**C**) e-values of 2921 PHE sequences from *P. vivax,* logarithmic scale. (D) Cumulative percent identity between the 2921 PHE and human genome homologs.

A major concern when searching for HGT in parasites is the contamination of genome nucleotide databases with sequence information derived from genetic material from the host. This has led in the past to false classifications of human to *Plasmodium* HGT events. Carlton et al (Carlton et al. 2008) performed a rigorous quality control process that accompanied the sequencing and publication of the *P. vivax* genome, showing very low contamination rates. However, almost all PHE found fall into 790 short (median length: 879) contigs, with only 8 contigs longer then 8kb (data not shown). Moreover, most of these contigs show high similarity to the human genome outside the found PHE. To verify this is not due to contamination, 10 primer sets (Supplementary Table 4) were selected to identify features combining sequences similar to human DNA with characteristics not found in the human genome. For example, primer set 2 was selected for showing similarity to the human genome, but with a deletion in the middle of the region of similarity (Fig. 2B). In case of a contamination, it is unlikely that the same deletion will be seen in an independently derived genomic DNA sample. Of the 10 primer sets, one resulted in multiple bands and 9 showed bands at the expected size. These were verified by sequencing, with 8 of them showing great (e-value $< 10^{-43}$) similarity to the human genome, and much greater similarity to the *P. vivax* genome (e-value $< 10^{-57}$, Figure 2A, Supplementary Table 4). One sequence (Set 5, Supplementary Table 4) did not yield a clear sequence in three out of four attempts. The fourth attempt resulted in a sequence showing e-value of 10^{-36} compared to the human genome and only mild similarity (e-value $< 3\times10^{-5}$) compared to the *P. vivax* genome. Set 2 was verified using a second distinct genomic DNA extraction (Methods) with similar results (Figure 2B). As it is improbable that two separate isolations will have a contamination with the same sequence variations, not found in the human genome, these results are unlikely to be due to contamination. Any new contamination is expected to show greater similarity to the contaminating genome, while here 8 out of 10 show much greater similarity to the *P. vivax* genome, including deletions not found in the contaminating genome.

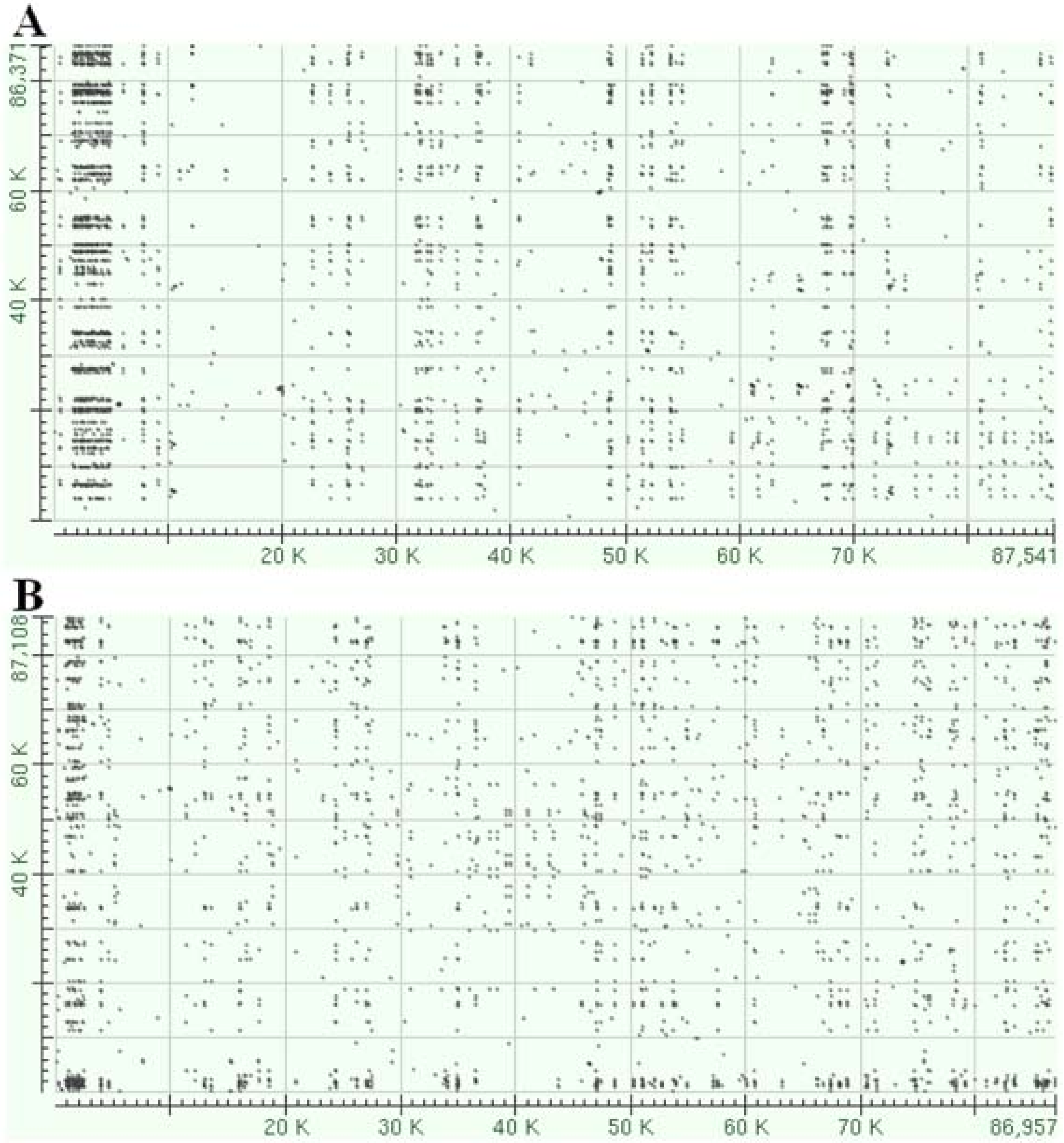

Fig 3. Internal comparison of *P. vivax* PHE. (**A**) A dot matrix comparing *P. vivax* PHE, with sequences divided at the middle. (**B**) A dot matrix comparing *P. vivax* PHE, with sequences divided alternately. Dots indicate alignment between query and subject sequences.

Five independent lines of evidence suggest that the results shown here do not represent errors in the sequencing and assemble of the *P. vivax* genome, but genuine HGT events. (i) *P. vivax* was isolated for sequencing from squirrel monkeys (*Saimiri boliviensis boliviensis*). Of the PHE found, only 18% have highly similar sequences (e-value < 10^{-7}) in the squirrel monkey genome, making host contamination unlikely. This does not rule out a lab contamination. (ii) PHE sequences from the *P. vivax* genome are 96% similar to human, (PHE with a squirrel monkey homolog show 93.8% similarity), unlike results for the control genomes, that overall show >99% similarity between *Plasmodium* and either human or host sequences (See Methods). This percent of identity falls far below the expected similarity for a contamination, considering the accuracy of current generation sequencing machines (>99%) and the genomic coverage (X10; ref. (Carlton et al. 2008)). (iii) Both cumulative PHE e-values, plotted on a logarithmic scale, and percent identity, appear almost linear. This is consistent with multiple HGT events, where integrated fragments accumulate mutations over time, but not with a contamination (Figure 2C, D) (iv) The sequences found are enriched for mRNA (9.5% compared to 1.1% expected(Venter et al. 2001); $P < 3\times10^{-14}$, cumulative binomial distribution). (v) Results are enrichement ($P < 6\times10^{-12}$; benjamini correction $< 4\times10^{-8}$) for IL1 familliy (See Methods). Overall these results suggest the sequences found are probably genuine HGT events incorporated next to repetitive elements or low complexity sequences. This can be validated by Fluorescent in situ hybridization in a specialized lab.

Multiple mechanisms can theoretically cause HGT; for example, DNA fragments can penetrate at random through small membrane ruptures; elements of viral origin can transfer specific DNA and RNA molecules across membranes; and selfish element can leave the host genome and incorporate into foreign DNA, if adjacent. The lengths of the PHE containing contigs are generally too short for viral elements, although it cannot be excluded that viruses facilitate the transport or were purged from the *Plasmodium* genome. This is supported by the fact that only a handful of PHE sequences map to known viral genomes (data not shown). To further deduce if these sequences are derived from a common origin and reveal repetitions, as the later might indicate a functional role, an internal comparison was made. Not surprisingly, it was found that PHE sequences tend to be similar to each other (Fig. 3), indicating either a common origin or a functional role. For example, S60429 (Supplementary Table 1) is closely related to 13 other PHE (e-value < 10^{-18}). As it does not map to any known virus or transposon, a functional role was investigated. Most of S60429, along with its other closely related PHE, are mapped to nitric oxide synthase 1 (neuronal) adaptor (NOS1AP; e-value = 5×10^{-11}). To a different location in this gene are also mapped S60439 and its 12 homologs. NOS1AP functions as an adapter protein, linking neuronal nitric oxide synthase (nNOS) to targets such as Dexras1. NO is known to have a major role in host-parasite relation, but the nature of this role remains controversial. It has been claimed that high NO has a protective effect against malaria(Anstey et al. 1996) but also hypothesized that parasite-induced NO production may be the cause of parasite-related anemia(Anstey et al. 1999). Comparison of all PHE sequences with the NOS1AP gene revealed that 22% of PHE sequences cover 13% of NOS1AP in a discontinuous manner (Fig 4, Supplementary Table 2). A global view of the human homologs of PHE revealed enrichment for Interleukin-1 (IL1) family, of which 7 members had mapped PHE (Figure 5). IL1 family is known to be essential for malaria protection and polymorphism in IL1, that has a receptor antagonist is mapped to a PHE, is involved in susceptibility to malaria. While these results do not confer information as to how *P. vivax* adjusts to its host, they do imply that HGT events were fixed in the population via functional selection. Still, much needs to be done in order to understand how incorporated DNA can serve the parasite. The fragments found are generally too short to mimic fully functional human proteins, but might serve to imitate protein domains or serve as non coding RNA expression regulators.

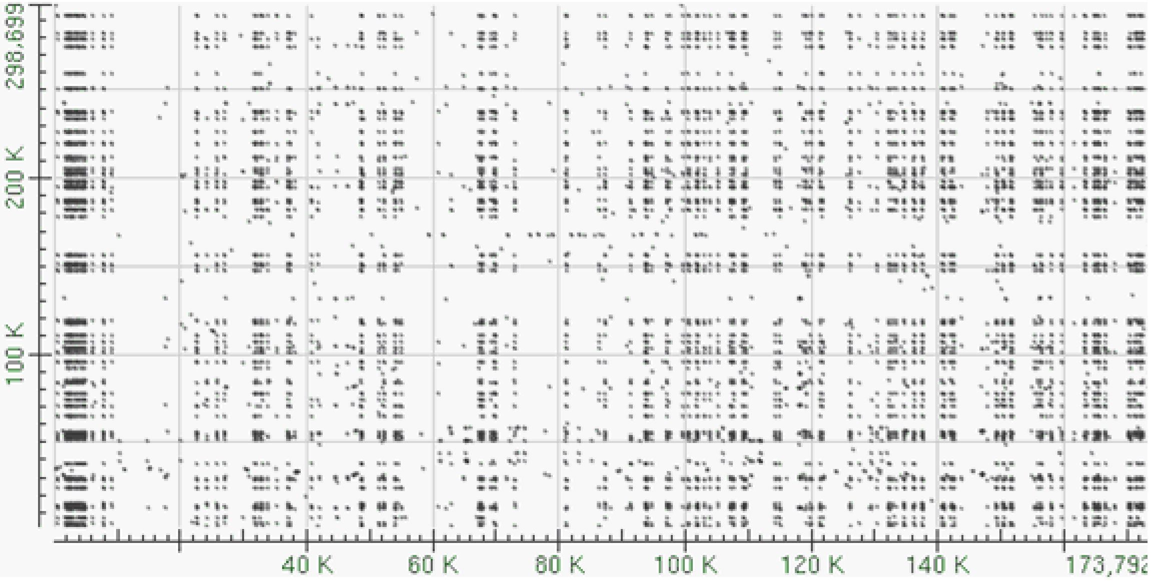

Fig 4. A dot matrix comparing NOS1AP (vertical) with all *P. vivax* PHE sequences (horizontal). Dots indicate alignment between query and subject sequences.

Discussion

Malaria and the evolution of resistance to it have played a key role in shaping the human genome. Multiple genetic traits, diseases and mutations originated and prospered as *Plasmodium* defense or evasion mechanisms. Among this can be found glucose-6-phosphate dehydrogenase (G6PD) deficiency, the most common human enzyme defect (Mehta et al. 2000). It is now shown that the human genome probably had a direct influence on that of *P. vivax,* and it is hypothesized that in the evolutionary arms race, HGT is used by parasites to improve its fitness in the host. In total, 2921 PHE were found, spanning 174kb, or 0.7% of the *Plasmodium* genome. As all sequences were compared to the *P. knowlesi* genome, the time frame for these changes is no more than 4.7 million years. Taking into account the mutation rate, HGT emerges as a major source for genetic diversity, with a contribution on the same order of magnitude as random mutations (Cornejo et al. 2006). Such a quick rate of foreign DNA adoption, along with initial results indicating enrichment in process known to be critical for parasite fitness, point to the possibility that *P. vivax* systematically incorporates DNA from its host to manipulate immune response pathways deleterious to parasite persistence. By doing so, it might manage to cope with new genetic adaptations to become the most widespread human-infecting species of *Plasmodium*.

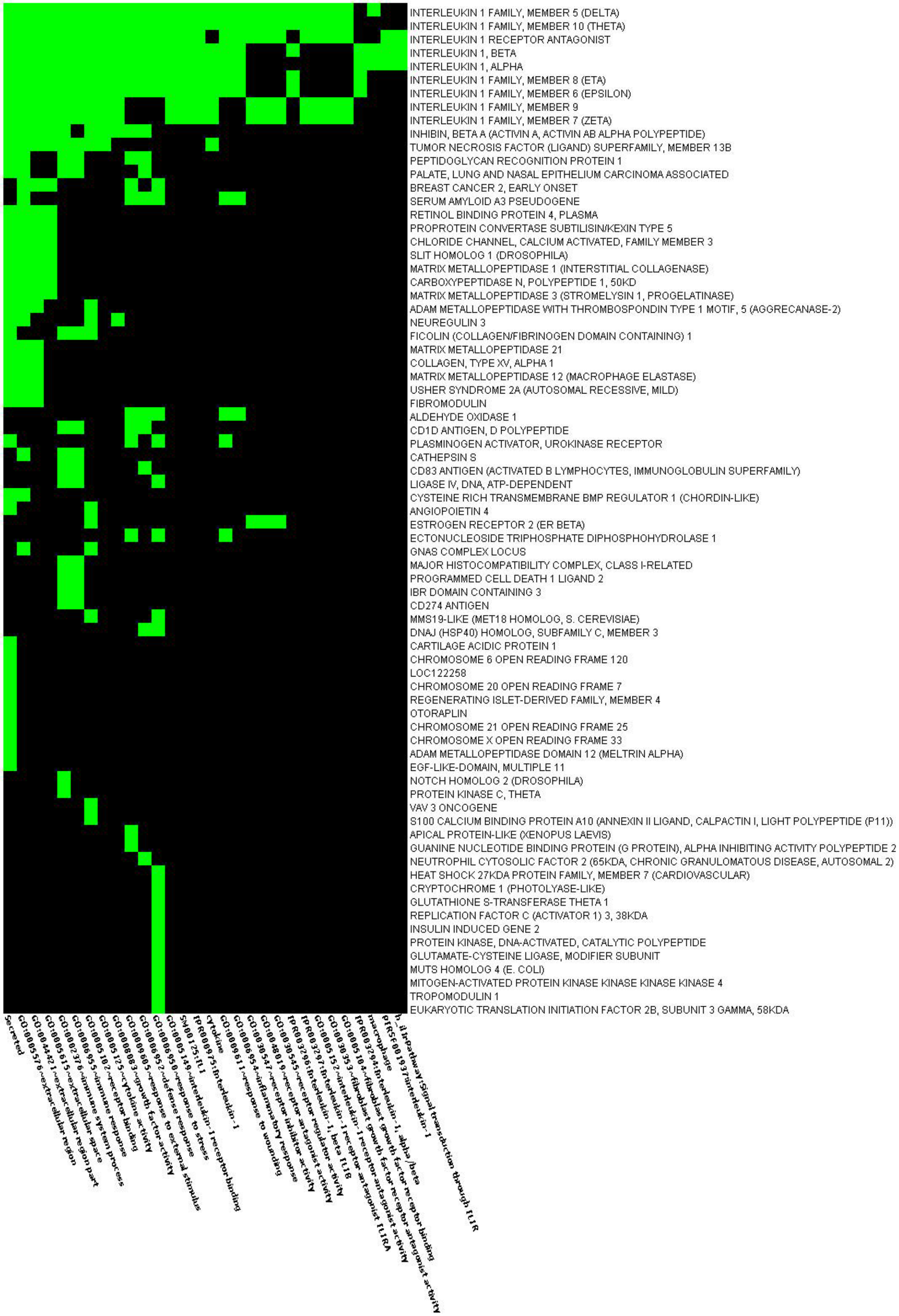

Fig 5. IL1 family is enriched in *P. vivax* PHE. A 2D view of Functional annotation clustering using DAVID.

This work suggests that multiple HGT events among eukaryotes are a major evolutionary force in at least one species. As previously proposed (Anderson 1970), transfer of DNA between organisms can speed up the rate of evolution and be most beneficial to organisms able to exploit such mechanisms. A question remains why this path was not taken by *P. falciparum.* It is hypothesized that the long duration of *P. vivax* in a latent state both gives an opportunity to acquire and experiment with foreign DNA and requires a better adaptation for a more subtle long-term presence. The approach used here can be duplicated to search for widespread HGT between many of the sequenced genomes, finding underling complexity between interacting organisms.

Materials and Methods

Data sources

Plasmodium genome sequences were obtained from the PlasmoDB, release 6.1(Aurrecoechea et al. 2009). The human genome was downloaded from NCBI BLAST database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/human_genomic).

Data processing

Unique identifiers were inserted between file lines in all *Plasmodium* genomes, resulting in a 60 bp fragmentation of the genome. To enable effecent comparison, plasmodium sequences were compared to the assembled human genome using locally installed BLAST (version 2.2.21;ref. (Altschul et al. 1997)) with the megablast option and the following parameters "-b 1 -v 1 -e

0.0000001 -D 0 -a 2 -M 100000 -d human_genomic". This was followed by netblast (version 2.2.21) for passing results without the megablast option using the nr database '-p blastn -a 2 -m 9 -e 0.0000001 -I T -v 1 -b 1 -u "txid5820[ORGN] NOT specific_plasmodium[ORGN]"' and similarly for mouse, squirrel monkey and human. For mRNA detection, a comparison was made to the refseq rna database with bit score equal or higher then obtained from human DNA (nr database) included. All other comparisons were made with the web BLAST interface using the blastn option. Data manipulation was performed with MS excel and perl scripts.

Quality control

Table S3 shows the number of sequences for each of the *Plasmodium* genome at different processing stages. Precent of identity was calculated only for sequences found above the treshold. All passing results combined, excluding *P. vivax,* showed >99% identity to either to the human or to the host genome, indicating a possible contamination. This was also true for each genome seperatly, except a single sequence of *P. Yoelii*. *P. vivax* showed an average 96.2% identity, with a range of 88.1-100% and a standard deviation of 2.5%.

Verification

Ten primer sets (Table S4) were selected for containing features not found in the human genome. These included single nucleotide changes, indels and primer sequences not found in the human genome. Two separete isolations of Genomic DNA, from Panama and Pakchong (MRA-342G and MRA-343G) were obtained from the Malaria Research and Reference Reagent Resource Center (MR4; http://www.mr4.org/). PCR was used to amplify all ten sets using MRA-342G, set 2 was also a amplified using MRA-343G. PCR was conducted as follows: 5 min at 95°C followed by 40 cycles of 30 sec denaturation at 95°C, 45 sec annealing between 50-65°C and 1 min elongation (72°C). Finally 10 min of elongation was allowed. PCR products were ran on a 1.5% agarose gel. All but primer set 1 showed bands at the expected size, primer set 1 showed multiple bands at various sizes. These bands were cut, DNA extraced with HiYield Gel/PCR DNA Fragment Extraction Kit (RBC bioscience) and send for sequencing.

Internal comparison

Sequences obtained were self compared in two manners. First by splitting at the middle of the list, thus keeping contig structure and second by comparing odd versus even line numbers (Fig 3). By doing so, each sequence was compared to 75% of all other sequences with only two runs. Comparison was made using two sequence BLAST with default parameters.

Global comparisons

A few sequences showing multiple homologs in the internal comparison were found to be closely related to parts of NOS1AP. Thus, a two sequence BLAST comparison between NOS1AP and all *P. vivax* PHE was made using default parameters. This resulted in a 22% of PHE sequences covering 13% of NOS1AP in a clustered but discontinues manner (Fig. 4).

To search for other genes and process involved, gene identifiers were subbmitted to DAVID bioinformatic tools(Dennis et al. 2003; Huang da et al. 2009). A strong enrichement ($P < 6\times10^{-12}$; benjamini correction $< 4\times10^{-8}$) for IL1 familliy was noticed. Multiple found members of this familly are enriched for secrition and extacelular regions (Fig. 1).

HGT aquesition rate

The seperation of *P. vivax* from *P. knowlesi* followed the seperation from *P. fragile*, and thus is upper-bound to 4.7 million years ago. PHE account for 0.7% of the *P. vivax* genome, thus the minimal horizontal DNA acquisition rate needed is ~ 1.5×10^{-9} base pair insertions per base pair per year, compared to estimates mutation rate ranging from 4.31×10^{-9} to 3.21×10^{-9} mutations per base pair per year(Cornejo et al. 2006).

Supplementary data is available at https://sites.google.com/site/hgtsuppdata/

Acknowledgements

I thank Yosef Gruenbaum for making this research possible; Ron Dzikowski and Thomas Templeton for helpful discussion and help in the preparation of the manuscript, verification design and *P.vivax* genomic DNA.

Author Disclosure Statement

No competing financial interests exist

References and Notes

. ftp://ftp.ncbi.nlm.nih.gov/blast/db/human_genomic.

Altschul, S. F., T. L. Madden, A. A. Schaffer, et al. 1997. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Anderson, N. G. 1970. "Evolutionary significance of virus infection." Nature **227**(5265): 1346-7.

Anstey, N. M., D. L. Granger, M. Y. Hassanali, et al. 1999. "Nitric oxide, malaria, and anemia: inverse relationship between nitric oxide production and hemoglobin concentration in asymptomatic, malaria-exposed children." Am J Trop Med Hyg **61**(2): 249-52.

Anstey, N. M., J. B. Weinberg, M. Y. Hassanali, et al. 1996. "Nitric oxide in Tanzanian children with malaria: inverse relationship between malaria severity and nitric oxide production/nitric oxide synthase type 2 expression." J Exp Med **184**(2): 557-67.

Aurrecoechea, C., J. Brestelli, B. P. Brunk, et al. 2009. "PlasmoDB: a functional genomic database for malaria parasites." Nucleic Acids Res **37**(Database issue): D539-43.

Carlton, J. M., J. H. Adams, J. C. Silva, et al. 2008. "Comparative genomics of the neglected human malaria parasite Plasmodium vivax." Nature **455**(7214): 757-63.

Carlton, J. M., S. V. Angiuoli, B. B. Suh, et al. 2002. "Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii." Nature **419**(6906): 512-9.

Cornejo, O. E. and A. A. Escalante. 2006. "The origin and age of Plasmodium vivax." Trends Parasitol **22**(12): 558-63.

Deitsch, K., C. Driskill and T. Wellems. 2001a. "Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes." Nucleic Acids Res **29**(3): 850-3.

Deitsch, K. W., J. M. Carlton, J. C. Wootton, et al. 2001b. "Host sequences in Plasmodium falciparum and Plasmodium vivax genomic DNA: horizontal transfer or contamination artifact?" FEBS Lett **491**(1-2): 164-5.

Dennis, G., Jr., B. T. Sherman, D. A. Hosack, et al. 2003. "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Gilbert, C., S. Schaack, J. K. Pace, 2nd, et al. "A role for host-parasite interactions in the horizontal transfer of transposons across phyla." Nature **464**(7293): 1347-50.

Gladyshev, E. A., M. Meselson and I. R. Arkhipova. 2008. "Massive horizontal gene transfer in bdelloid rotifers." Science **320**(5880): 1210-3.

Hall, C., S. Brachat and F. S. Dietrich. 2005a. "Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae." Eukaryot Cell **4**(6): 1102-15.

Hall, N., M. Karras, J. D. Raine, et al. 2005b. "A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses." Science **307**(5706): 82-6.

Huang da, W., B. T. Sherman and R. A. Lempicki. 2009. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.

Institute, P. S. U. a. t. W. T. S. "Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute."

Lander, E. S., L. M. Linton, B. Birren, et al. 2001. "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Mehta, A., P. J. Mason and T. J. Vulliamy. 2000. "Glucose-6-phosphate dehydrogenase deficiency." Baillieres Best Pract Res Clin Haematol **13**(1): 21-38.

Moran, N. A. and T. Jarvik. "Lateral transfer of genes from fungi underlies carotenoid production in aphids." Science **328**(5978): 624-7.

Morgulis, A., G. Coulouris, Y. Raytselis, et al. 2008. "Database indexing for production MegaBLAST searches." Bioinformatics **24**(16): 1757-64.

Pain, A., U. Bohme, A. E. Berry, et al. 2008. "The genome of the simian and human malaria parasite Plasmodium knowlesi." Nature **455**(7214): 799-803.

Striepen, B., A. J. Pruijssers, J. Huang, et al. 2004. "Gene transfer in the evolution of parasite nucleotide biosynthesis." Proc Natl Acad Sci U S A **101**(9): 3154-9.

Striepen, B., M. W. White, C. Li, et al. 2002. "Genetic complementation in apicomplexan parasites." Proc Natl Acad Sci U S A **99**(9): 6304-9.

Templeton, T. J., L. M. Iyer, V. Anantharaman, et al. 2004. "Comparative analysis of apicomplexa and genomic diversity in eukaryotes." Genome Res **14**(9): 1686-95.

Venter, J. C., M. D. Adams, E. W. Myers, et al. 2001. "The sequence of the human genome." Science **291**(5507): 1304-51.

Waterston, R. H., K. Lindblad-Toh, E. Birney, et al. 2002. "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Zhang, Z., S. Schwartz, L. Wagner, et al. 2000. "A greedy algorithm for aligning DNA sequences." J Comput Biol **7**(1-2): 203-14.