# Comparative Genomics of Parasites

INAUGURALDISSERTATION

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

**Philipp Ludin**

aus Ettiswil (LU)

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Pascal Mäser
Prof. Dr. Marcel Tanner
Prof. Dr. Volker Heussler

Basel, den 11. Dezember 2012

Prof. Dr. Jörg Schibler
Dekan

# Table of Contents I

# Acknowledgements

First and foremost, I would like to thank Pascal Mäser for giving me the opportunity to do this PhD thesis, for his supervision, ideas and support. It has been a privilege and a real pleasure to work with such an expert.

I would also like to thank Prof. Marcel Tanner and Prof. Volker Heussler for joining my PhD committee.

I am indebted to Daniel Nilsson for all advice and teaching of the most important bioinformatic skills.

I would like to express my gratitude to Fabrice, Fügi, Nadia, Eva and Chri for the fruitful inputs and the nice time we spent together.

Warmest thanks to all other members of the Parasite Chemotherapy Unit, in particular to Scheuri, Ralf, Sergio, Matze and Reto Brun.

I sincerely acknowledge our collaborators Harry de Koning, Mike Barrett (both University of Glasgow), David Horn (London School of Hygiene and Tropical Medicine), Ben Woodcroft, Stuart Ralph (both University of Melbourne) and people from the Blaxter Lab (University of Edinburgh).

I wish to thank Roche for the 10 Giga-Base Grant which allowed us to sequence several *Trypanosoma brucei* genomes on the 454 platform. I am thankful to the Mathieu-Stiftung, Freiwilligen Akademischen Gesellschaft and Emilia Guggenheim-Schnurr Stiftung for financial support.

A special thank must go to my former working place on the 3rd floor at the ICB in Bern: Mosi, Lucien, Robin, Bernd, Sandro, Patrick, Jan, Baschi, Simon, Gaby, Aline, Marina, Kapila

I am very grateful to Christian Lengeler and the whole SwissTPH football team for spending some nice time on the pitch.

Many thanks must also go to a number of people who provided an unforgettable atmosphere at the SwissTPH: Urs, Beni, Moes, Lucienne, Kate, Thomas, KW, Chrigu, Theresia, Dania, Sophie, Oli, Igor, Chris, Esther, Isabel, Mireille, Cristian, Dirk, Mireia, Araceli, Daniela, Sonia, David, Michael, Raphael, Alex, Susy, Kurt, Christine

Finally, I am deeply grateful to my family and friends for their support and encouragement.

# Summary

Comparative genomics is an emerging field in biology that started in 1995 when the first two genomes of self-replicating organisms had been sequenced. Since then, a plethora of genome sequences from parasites, hosts and vectors has been made available. The comparison of genomes may shed light on the genetic and evolutionary bases of convergent and divergent properties throughout the tree of life. Comparative genomics of parasites may be performed at different levels: (i) between parasites and free-living organisms, (ii) between parasite and host, (iii) within closely related species (family, genus), and (iv) within species.

Studies on all four levels were performed in the framework of this PhD thesis. On the one hand I took advantage of the wealth of genomes available to gain new insights into the molecular nature of host-pathogen interactions, drug target discovery and evolution of parasites in general. On the other hand whole genome sequencing projects were carried out that directly addressed parasite chemotherapy. Three algorithms were invented to study important aspects of parasitology. Automated tools were developed that are widely applicable to parasites and they were included in the *Dirofilaria immits* genome project.

First, whole parasite proteomes were screened for molecular mimicry candidates by comparing parasite sequences to host and control species. Linear epitopes were identified that were present in the host proteome as well as in the parasite but not in free-living control organisms. The designed pipeline returned several interesting hits, most notably a motif in several PfEMP1 variants identical to part of the heparin-binding domain in the cytoadhesive and immunosuppressive serum protein vitronectin. Moreover, a homolog of cytokine suppressor SOCS5 was found in several pathogenic nematodes.

Second, a tool was built that discriminates DNA sequences to the level of species of origin based on palindrome frequency patterns. It relied on the highly specific palindrome occurrence among species for DNA typing. The power of the program was illustrated when the comparison of palindrome frequency patterns provided further evidence for horizontal gene transfer between *D. immitis* and its *Wolbachia* endosymbiont.

Third, a drug target prediction pipeline was designed that is based on the assumption that proteins are likely to be essential if they are highly conserved among related species and if there are no similar proteins in the same proteome. By inclusion of other criteria such as matchlessness in the human proteome, expression in a relevant stage and prediction of druggability, candidates were identified that may serve as starting points for rational drug discovery. When applied to *P. falciparum*, a sizeable list of 40 proteins with proven and new targets was obtained.

Further, whole genome sequencing was conducted of a drug-sensitive *Trypanosoma brucei rhodesiense* STIB900 line and two drug-resistant derivatives STIB900-M and STIB900-P. By comparative genomics, mutations and gene deletions were detected that may confer drug resistance to melarsoprol and pentamidine. Proof-of-principle was the detection of the loss of known determinants of drug susceptibility, the adenosine transporter *TbAT1* and the aquaporin *TbAQP2*. Moreover, a coding mutation occurred in both resistance lines in the gene for the RNA-binding protein UBP1.

In conclusion, comparative genomics is a powerful tool that offers new opportunities in biological research. Comparative genomics can be applied at different levels, from basic research to applied questions such as drug discovery and resistance.

# Table of Abbreviations

| | |
|---|---|
| aa | amino acid |
| AQP | aquaporine |
| AT1 | adenosine transporter 1 |
| ATS | acidic terminal segment |
| BLAST | basic local alignment search tool |
| BLOSUM | blocks substitution matrix |
| bp | basepair(s) |
| CAZymes | carbohydrate-active enzymes |
| COG | clusters of orthologous groups |
| CRIT | complement C2 receptor inhibitory trispanning |
| CSP | circumsporozoite protein |
| D.I. | druggability index |
| DHFR | dihydrofolate reductase |
| DHPS | dihydroopteroate synthetase |
| DNA | deoxyribonucleic acid |
| E/S | excretory/secretory |
| EGF | epidermal growth factor |
| EST | expressed sequence tag |
| e-value | expectancy value |
| GO | gene ontology |
| GPI | glycosylphosphotidylinotisol |
| HAT | human African trypanosomiasis |
| HGT | horizontal gene transfer |
| HMM | hidden Markov model |
| HQ | high quality |
| ICAM | intra-cellular adhesion molecule |
| indel | insertion/deletion |
| kb | kilobase(s) |
| L3 | third-stage larvae |
| LTR | long terminal repeat |

| | |
|---|---|
| MAC | membrane attack complex |
| malERA | Malaria Eradication Research Agenda |
| mapq | mapping quality |
| MASP | mucin-associated surface protein |
| mb | megabase(s) |
| MIF | migration inhibition factor |
| MRPA | multidrug resistance-associated protein |
| NGS | next generation sequencing |
| NTS | N-terminal segment |
| PCR | polymerase chain reaction |
| PEP | phosphoenolpyruvate |
| Perl | practical extraction and report language |
| PEXEL | *Plasmodium* export element |
| PfEMP1 | *Plasmodium falciparum* erythrocyte membrane protein 1 |
| pir | *Plasmodium* interspersed repeat |
| RF | resistance factor |
| rifin | repetitive interspersed family |
| RNA | ribonucleic acid |
| SH2 | src homology 2 |
| SNP | single nucleotide polymorphism |
| SOCS | suppressor of cytokine signalling |
| spp. | species |
| stevor | sub-telomeric variable open reading frame |
| STIB | Swiss Tropical Institute, Basel |
| TGFß | transforming growth factor ß |
| Th | T-helper |
| TM | transmembrane |
| TRAP | thrombospondin-related anonymous protein |
| UBP1 | uridine-rich-binding protein 1 |
| ups | upstream sequence |
| vs. | versus |
| VSG | variant surface glycoprotein |

# *Chapter 1*

## General Introduction

### Comparative Genomics and Applications to Parasites

# General Introduction

## 1 A General Introduction to Comparative Genomics

A landmark in the field of genomics represents undeniably the first sequenced genome of a self-replicating organism, the one of the bacterium *Haemophilus influenzae* [1]. Since then, thousands of sequenced genomes have been made available across the tree of life. The genome of *Saccharomyces cerevisiae* was the first eukaryotic to be published [2]. At this time, species were mainly selected for sequencing projects based on their size, role as a model organism and their relevance to humans [3]. Therefore it was no surprise that *Caenorhabditis elegans* was the first completed metazoan genome [4], as the nematode serves as an important model for multicellular organisms [5]. A further milestone in the era of genomics was certainly the announcement of the human draft genome in 2001 accomplished by the publicly funded Human Genome Project [6] and private company Celera Genomics [7] simultaneously. Researchers now started to systematically compare genomes of model organisms to *Homo sapiens* to gain insights into evolution and human diseases [8–11]. I refer here to comparative genomics as the comparative analysis of the fully sequenced genomes and their predicted encoded proteins between or within species. It can be further complemented by other 'omics' approaches such as proteomics, transcriptomics and metabolomics.

The beginning of comparative genomics marks the completion of the second bacterial genome, *Mycoplasma genitalium* [12], although some researchers date back the starting point to the late 1970s when the first viral genomes were completed [13]. With the completion of the genomes and their analyses, the scale has dramatically changed, from genes to genomes, from kilobases to megabases or even gigabases. Global views on genomes allow the discovery of conserved and divergent regions within and between species at various evolutionary distances [14,15].

The first comparison of complete genomes between organisms revealed a reduced genome of *M. genitalium* that is associated with a strikingly lower number of genes involved in metabolic pathways compared to *H. influenzae* [12]. With only two genomes available, the biological analyses were already put into another dimension by studying fundamental processes such as replication, transcription and translation on a genome-wide scale.

In terms of evolution, the release of the first complete genome of an archaeon, *Methanococcus jannaschii* [16], was a major breakthrough because this provided the first opportunity to explore the three domains of life on the whole genome level, although the genome of *S. cerevisiae* was made public available two months later [2]. The comparison revealed that genes of the cellular information process were more 'eukaryotic-like', whereas genes concerned with energy production, cell division, and metabolism seemed to have their origin in bacteria [16].

With more and more sequences available, the field of comparative genomics has grown rapidly and has become a major part in biological research. A pioneering study was unquestionably the comparative analysis of *Drosophila melanogaster*, *C. elegans* and *S. cerevisiae* [17]. For the first time, the full genomic sequences of three eukaryotic model organisms were available. Rubin et al. uncovered the 'core proteome' of each organism, i.e. the number of distinct proteins where a set of paralogs was taken together as a unit. The researchers pointed out that the 'core proteome' size of the fly is only doubled compared to the single-celled yeast, although Drosophila is a complex metazoan. Moreover, they showed that fly and worm share similar size of distinct protein families, despite the differences related to development and morphology. They concluded that the apparent complexity of an organism is not obtained by the pure number of genes [17].

As technological advances have reduced costs and have enormously accelerated the sequencing process, comparative genomics has got more informative and sensitive by accumulation of genomic data [14]. In the past few years, a large amount of meaningful studies addressing all kinds of biological questions were carried out - among them, I am focusing on eukaryotic parasites.

## 2 Comparative Genomics of Parasites

Parasites frequently have a complex life-cycle, switching between different stages that may include vector, alternate- and definite host. There is a persistent 'arms-race' between host and parasite that is reflected in the genome to some extent. Comparative

| Parasite species | Genome size [Mb] | Protein-coding genes | Year | Reference |
|---|---|---|---|---|
| *Encephalitozoon cuniculi* | 3 | 2000 | 2001 | [18] |
| *Plasmodium falciparum* | 23 | 5300 | 2002 | [19] |
| *Plasmodium yoelii* | 23 | 5900 | 2002 | [20] |
| *Cryptosporidium hominis* | 9 | 4000 | 2004 | [21] |
| *Cryptosporidium parvum* | 9 | 3800 | 2004 | [22] |
| *Entamoeba histolytica* | 24 | 9900 | 2005 | [23] |
| *Leishmania major* | 33 | 8300 | 2005 | [24] |
| *Plasmodium berghei* | 18 | 5900 | 2005 | [25] |
| *Plasmodium chabaudi* | 17 | 5900 | 2005 | [25] |
| *Theileria annulata* | 8 | 3800 | 2005 | [26] |
| *Theileria parva* | 8 | 4000 | 2005 | [27] |
| *Trypanosoma brucei* | 26 | 9100 | 2005 | [28] |
| *Trypanosoma cruzi* | 55 | 12000 | 2005 | [29] |
| *Babesia bovis* | 8 | 3700 | 2007 | [30] |
| *Brugia malayi* | 90 | 11500 | 2007 | [31] |
| *Giardia lamblia* | 12 | 6500 | 2007 | [32] |
| *Leishmania braziliensis* | 32 | 8100 | 2007 | [33] |
| *Leishmania infantum* | 32 | 8200 | 2007 | [33] |
| *Trichomonas vaginalis* | 160 | 60000 | 2007 | [34] |
| *Meloidogyne hapla* | 54 | 14400 | 2008 | [35] |
| *Meloidogyne incognita* | 86 | 19200 | 2008 | [36] |
| *Plasmodium knowlesi* | 24 | 5200 | 2008 | [37] |
| *Plasmodium vivax* | 27 | 5400 | 2008 | [38] |
| *Schistosoma japonicum* | 398 | 13500 | 2009 | [39] |
| *Schistosoma mansoni* | 363 | 11800 | 2009 | [40] |
| *Ascaris suum* | 273 | 18500 | 2011 | [41] |
| *Leishmania mexicana* | 32 | 8300 | 2011 | [42] |
| *Leishmania tarentolae* | 30 | 8200 | 2011 | [43] |
| *Trichinella spiralis* | 64 | 15800 | 2011 | [44] |
| *Dirofilaria immitis* | 84 | 10200 | 2012 | *Chapter 5* |
| *Plasmodium cynomolgi* | 26 | 5700 | 2012 | [45] |

**Table 1**. List of published endoparasite genomes. (Mb: megabases).

genomics can be a powerful discipline to illuminate this host-parasite co-evolution. Indeed, there is a wealth of parasite genomes to be explored (Table 1). With full genomic sequences of host and vector available, the opportunities for the study of each species and their complex interactions make the field of comparative genomics very attractive.

Comparative genome analyses of parasites may be performed at different levels: Comparative genomics (i) between parasites and free-living organisms, (ii) between parasite and host, (iii) within closely related species (family, genus), and (iv) within species. By no means this subdivision is definite; comparative studies may include more than one category as they do for drug target identification or vaccine development (Figure 1).

In the recent years, many studies on parasite genomics were undertaken but it is not the aim of my thesis to cover here the whole diversity of analyses because it is almost impossible to mention them all. Therefore I focus mainly on studies in which endoparasites of our own research are involved, i.e. plasmodia, trypanosomatids and nematodes.

## 2.1 Comparative Genomics between Parasites and Free-living Organisms

Regardless of having an intracelluar or extracelluar lifestyle, the challenges that endoparasites face are manifold: they need to (i) enter the host, (ii) route to the definite location, (iii) develop and/or reproduce, (iv) deal with the host's defense mechanisms, and (v) infect a new host.

Comparison of a non-parasitic genome to a parasitic one may give hints about the specific genomic adaptations during the evolution of a free-living organism into a parasite [46]. Ideally, the compared species are closely related because during evolution genomic divergence unrelated to parasitism may appear.

The first genomic sequenced parasitic eukaryote was *Encephalitozoon cuniculi* in 2001 [18], the smallest in size and the fewest gene number so far (Table 1). The remarkable reduction was manifested by the lack of genes for the tricarboxylic acid cycle and for several biosynthetic pathways, a low diversity of transporters, and gene shortening compared to non-parasitic species [18]. The authors speculated that the shortening is a consequence of reduced protein-protein interactions as a result of gene losses related to

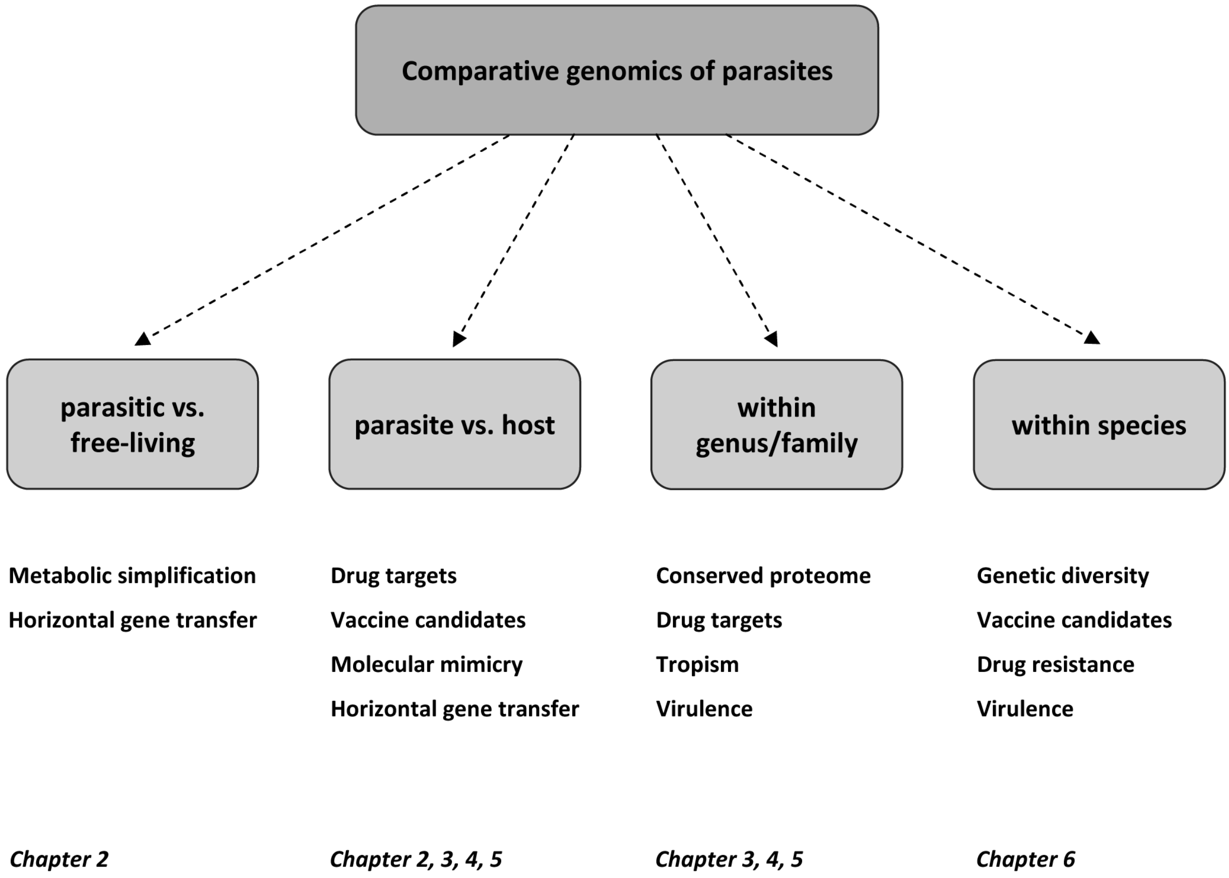


**Figure 1.** Overview on comparative genomics as discussed to parasites herein.

a parasitic adaptation because longer proteins may enable more complex regulation networks [47]. Moreover, *E. cuniculi* lacks mitochondria and peroxysomes [18].

The first sequenced eukaryotic parasite showed already tremendous differences compared to the hitherto sequenced free-living species. A common feature among endoparasites is the loss of metabolic functions during evolution [48]. For example, all obligate endoparasitic protozoa examined to date miss the genes for purine *de novo* synthesis, importing exogenous purines from their hosts [49]. *Brugia malayi* is incapable of purine synthesis as well, and it was suggested that the filarial nematode salvages purines from its *Wolbachia* endosymbionts [31]. Because the *de novo* purine synthesis is an energetically costly pathway, the hypothesis was that parasites primarily lost enzymes of ATP-consuming reactions for economical cause [48]. However, Nerima et al. disproved the proposed hypothesis by comparative analysis of metabolic-networks of free-living and parasitic eukaryotes, where they concluded that ATP-requiring reactions have been preferentially maintained during the course of evolution, whereas NADH- or NADPH-requiring reactions were lost [48].

Beside the convergent trend towards metabolic simplification, more traits among endoparasites were revealed by comparative genomics. For instance, the genome of *Plasmodium falciparum* showed a higher proportion of proteins involved in adhesion and immune evasion when compared to *S. cerevisiae* [19] but one must be careful because the analysis was based on assignment of gene ontology terms [50] which was at that time in its infancy. Another fascinating case of parasitic adaptation comes from plant-parasitic root-knot nematodes *Meloidogyne* spp. Their genomes contained an unexampled set of plant cell wall-degrading, carbohydrate-active enzymes (CAZymes) [35,36,46]. It is increasingly acknowledged that CAZymes were acquired by horizontal gene transfer (HGT) because the most similar proteins were found in bacteria [35,36,46]. It seems that these capture events played a crucial role in the evolution of root-knot nematodes [35,36,46]. Interestingly, diverse cellulases were found in *Pristionchus pacificus*, a necromenetic nematode that lives in association with beetles [51]. Dietrich et al. suggested that the acquisition of cellulases and other genes through HGT played a critical role in the evolutionary transition into a parasite and that *P. pacificus* may display preadaptations to a parasitic lifestyle [46,51].

Our knowledge about nematode parasitsm has benefited hugely from the availability of the genomes of the free-living *C. elegans* [4] and the necromenic *P. pacificus* [51]. This underpins the importance of having a closely related species to compare with to gain new insights into the evolution of parasites. A similar effect could be observed in trypanosomatidae as soon as the genome of the free-living kinetoplastid *Bodo saltans* [52] is fully sequenced or in *Plasmodium* when the genome of *Chromera velia* [53], a photosynthetic alveolate phylogenetically related to apicomplexans, is published.

## 2.2 Comparative Genomics between Host and Parasite

As mentioned earlier, a major accomplishment in the era of comparative genomics was the release of the draft human genome [6,7]. Indeed, research on similarities and differences between host and parasite on a large scale have revolutionized modern biology. In particular, drug target identification has benefited vastly since the release of the host genome because comparison of host and parasite genome may uncover biochemical peculiarities and vulnerable points for chemotherapeutic intervention, such as essential parasite enzymes that are not present in the host [48]. Although not directly linked with the human draft genome release, a striking example is the identification of enzymes targeted to or encoded by the apicoplast, an essential plastid of *Plasmodium* and other apicomplexa [54]. The fact that the apicoplast is derived from the secondary endosymbiosis of a cyanobacterium implies that these enzymes are excellent targets for drug development as many of them are bacterial-like and hence different from the mammalian host [54]. Furthermore, the fact that the present set of apicoplast proteins has been maintained during the reductive evolution of the endosymbiont indicates that they are likely to be essential [54]. Although the apicoplast genome was sequenced already in 1996 [55], the metabolic pathways targeted to the organelle became clearer when the *P. falciparum* genome was obtained as 545 of the 568 proteins predicted to be in the apicoplast are encoded by the nuclear genome [19,54]. In addition to housekeeping process (DNA replication, RNA transcription and Protein synthesis) proteins of prokaryotic origin, the identified anabolic pathways of isoprenoid precursor synthesis, fatty-acid biosynthesis and a partial haem biosynthesis are of particular interest, as they are not found in the vertebrate host of *Plasmodium* [54,56]. This was illustrated by the usage of fosmidomycin, an antibiotic that inhibited the non-mevalonate pathway of isoprenoid synthesis in *P. falciparum* [57]. Intriguingly, Fosmidomycin is currently tested in Phase II as combination-therapy with clindamycin against malaria [58].

As described before (see 2.1), all obligate endoparasitic protozoa are purine auxotrophs, a consequence of the parasitic trend towards metabolic simplification [48]. In contrast, parasites have developed alternative strategies and gained pathways in the course of evolution [56]. Compared to the host genome, parasites possess metabolic pathways absent in humans [56]. For instance, *Leishmania* and *T. cruzi* are able to synthesize cysteine from serine [56]. In many cases, genes involved in these pathways were similar to prokaryotes and therefore were believed to be acquired by HGT [56]. For example, it was suggested that nearly 50 genes were transferred from prokaryotes into the Tritryp lineage [28]. Further, a genome-wide search revealed up to 3% of the proteins comprised on TriTrypDB (http://www.tritrypdb.org) resembled rather prokaryotic than eukaryotic proteins [59].

Aside of drug targets, the development of vaccine candidates and diagnostic tools can benefit from comparative genomics as well by exploiting characteristics that distinguish parasite and host [56]. Nevertheless, genome comparisons that reveal similarities are also of interest. Molecular structures that are shared between host and parasite lead to the concept of molecular mimicry [60]. Molecules that are expressed on the parasites' surface or are secreted may interfere with the host. The benefits of mimicking host molecules are manifold: camouflage, cytoadherence or manipulation of host signalling [61]. The *P. falciparum* genome revealed a putative homolog of human cytokine macrophage inhibitory factor (MIF) that functions as a growth factor and immune-modulator in vertebrates [19]. Similar genes were found in *L. major* [24] (see 2.3). In *Plasmodium knowlesi*, researchers identified amino acid stretches in the extracelluar domain of the *kir* gene family products that resembled the immunoregulatory host protein CD99 [37]. Strikingly, one protein that belongs to the *cyir* gene family in *Plasmodium cynomolgi* had a highly similar region to CD99 as well [45]. Proteins that may interact with the host's immune system were also described in nematodes. Endoparasitc helminths are known as 'masters of immune regulation and host manipulation' [62]. In addition to MIF, predicted proteins similar to the human transforming growth factor β (TGFβ) and interleukin-16 were uncovered by the *B. malayi* genome project [31]. Molecular mimicry candidates were also found in *Ascaris suum*, where Jex et al. specifically turned their interest to the excretory/secretory (E/S) peptides [41]. Amongst others, they identified several C-type lectins similar to low affinity IgE receptors and it is thought that the parasite masks itself with these host-like antigens [41].

In summary, the research community has taken advantage of the availability of the host genome by identifying potential drug targets or molecular mimicry candidates. However, one should not forget to consider the genomes of vectors such as *Anopheles gambiae* [63] or *Aedes aegypti* [64], because they are published as well and their comparisons could significantly augment our understanding about the parasites.

## 2.3 Comparative Genomics within Family/Genus

Comparison of the genomes of closely related species may shed light on the parasite's (i) virulence, (ii) preference for a distinct host or niche and, (iii) lifestyle by characterizing features that are species or genus/family-specific. *Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania* spp. belong to the Trypanosomatidae family that is defined by the presence of a single flagellum and a kinetoplast [65]. Although they have different vectors and life-cycle features, the genomes of these parasites shared about 6200 orthologues gene clusters that were mostly arranged in syntenic order [66]. Genes conserved among closely related species may fulfil important functions and drugs against these potential targets may be effective against more than one species [66]. The causative agents of sleeping sickness, chagas disease and leishmaniasis have peculiar characteristics in common such as polycistronic transcription, trans-splicing, RNA editing and ergosterol biosynthesis [66]. Amino acid alignments of orthologues genes showed an average of 57% identity between *T. brucei* and *T. cruzi*, and 44% identity between *L. major* and both trypanosomes [66], following their expected phylogenetic relationship [67]. Despite the large number of shared genes, the parasites show vast differences in regard to their vector, tissue targeting and their immune evasion mechanisms [66]. These discrepancies may be reflected by species-specific genes. The parasite-specific genes were mostly found on non-syntenic chromosomes at subtelomeric locations and the majority seemed to be members of surface antigens families [65]. Unique gene families that are involved in host-parasite interactions were found towards telomeres in other organisms as well and it was suggested that subtelomeric regions have a higher rate of gene diversification in many organisms [68]. Indeed, most *T. brucei*-specific genes were located near telomeres and were related to the parasite's ability to undergo antigenic variation in the mammalian host [28]. Similar, the largest *T. cruzi*-specific gene families occurred at subtelomeric regions and encoded for the surface proteins mucin and mucin-associated surface proteins (MASPs) [29]. In contrast, most *L. major*-specific genes were randomly dispersed among the chromosomes [24]. Although some of them were identified to be responsible for pivotal metabolic differences between *Leishmania* and the other trypanosomes, 68% have no functional annotation [24]. Interestingly, two closely related genes were found to encode a protein that showed up to 40% identity to MIF homologs from other organisms [24]. It is suggested that *Leishmania* MIF may manipulate the host macrophage response (see 2.2). Further, a comparison of three *Leishmania* species revealed only 200 genes that varied between the investigated species and it seems that they were lineage-specific mainly due to gene loss and pseudogene formation [33]. Moreover, the genomes showed high conservation in terms of synteny and coding sequences [33].

Similar comparisons were carried out with *Plasmodium* spp. [25,69]. To date, approximately 200 species have been described belonging to the genus *Plasmodium* and infect mammals, birds and squamate reptiles (e.g. lizards, snakes) [70]. So far, the fully sequenced genomes of the human malaria parasites *P. falciparum*, *P. vivax*, the monkey parasites *P. knowlesi* and *P. cynomology*, and the rodent parasites *P. berghei*, *P. chabaudi* and *P. yoelii* were published from mammal pathogens (Table 1). The comparison between *P. falciparum* and the sequenced rodent parasites uncovered about 4500 genes conserved among the *Plasmodium* spp. [25]. Moreover, orthologous genes of *P. berghei* and *P. chabaudi* seemed to be under negative selection in general, however, it was suggested that genes likely involved in host-parasite interactions (i. e. genes containing transmembrane domains or signal peptides) are more diverse [25]. As seen for *T. brucei* and *T. cruzi*, comparative genomic analyses revealed that *Plasmodium* species- or species subset-specific genes were preferentially located at dynamically evolving subtelomeric regions [25,68,69]. In *P. falciparum*, the highly variable *var*, repetitive interspersed family (*rifin*) or sub-telomeric variable open reading frame (*stevor*) families were found towards the telomeres [19]. The reference 3D7 *P. falciparum* genome contained 59 *var* genes whose products are exported to the surface of infected red blood cells and permit adhesion to host endothelia through multiple adhesion domains [19,71]. The *var* gene family encodes *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) which is thought to be the predominant virulence factor [69,71]. Only one of the PfEMP1 proteins is expressed at a time and transcriptional switching between *var* genes allows antigenic variation that leads to immune evasion [19,71]. The specific function of *rifin* and *stevor* remains to be solved [19]. Similar families implicated in immune evasion were likewise found in other *Plasmodium* species, namely *vir* in *P. vivax*, *SICAvar* and *kir* in *P. knowlesi*, *cyir* in *P. cynomolgi*, and the *cir/bir/yir* family in rodent-infective parasites [45,72]. They are generally described as *Plasmodium* interspersed repeat (*pir*) multigene families and may have diverse functions such as signaling, trafficking and adhesion [25,72]. In contrast to other *Plasmodium* genomes where only few genes of *pir* families were found outside subtelomeric and telomeric regions, *SICAvar* and *kir* family genes were distributed throughout the *P. knowlesi* genome [37]. Although genes unique to a single *Plasmodium* species were predominantly located near teleomeres, species-specific genes were identified in core regions at synteny breakpoints and intrasyntenic indels [68,69]. Frech et al. compared the genomes of six *Plasmodium* species (*P. falciparum*, *P. vivax*, *P. knowlesi*, *P. berghei*, *P. chabaudi* and *P. yoelii*) to identify possible chromosome-internal species-specific genes to reveal unknown factors related to human diseases including pathogenicity and 'human-mosquito-human' transmissibility [69]. First, they focused on genes that were present in the primate parasites but absent in the rodent parasites to identify candidates linked to the parasite's ability to infect primates. Among 16 identified genes, of particular interest were three key enzymes of thiamine biosynthesis. The authors speculated that the primate host may provide insufficient amount of thiamine to the parasites [69]. Second, genes possibly important for parasite transmission between humans were identified by finding genes conserved between *P. falciparum* and *P. vivax* but absent in *P. knowlesi*. Despite *P. knowlesi* can naturally infect humans, no natural 'human-mosquito-human' transmission has been documented so far [73]. Overall, 13 syntenic genes were identified to be specific to *P. falciparum* and *P. vivax* when compared to *P. knowlesi* [69]. Strikingly, three genes were specifically upregulated in gametocytes and sporozoites in cell cycle expression experiments suggesting a role in parasite development within the vector [69]. Next, the researchers were looking for novel candidate genes that may explain the high virulence of *P. falciparum* compared to *P. vivax* apart from the known *var/rif/stevor* gene families. It is widely believed that *P. vivax* is less virulent because it preferentially infects reticulocytes that comprise only 1-2% of erythrocytes and hence limit hyperparasitaemia [69,74]. Another reason could be that *P. vivax*-infected red blood cells do not need to adhere to the vascular endothelium to avoid splenic clearance because they are more deformable and therefore are not stuck in the spleen [69,74,75]. To narrow down the potential candidate set, only genes were retained that account for features associated with human virulence such as *Plasmodium* export element (PEXEL) motifs [76], signal peptides, transmembrane domains, or coexpression or interaction with known virulence genes [69]. Most of the 15 novel candidate virulence genes had unknown function and their association with high virulence remained speculative. And finally, Frech et al. were looking for genes that were only present in *P. vivax* to gain insights into the parasite's ability to infect reticulocytes and to develop dormant hypnozoite formation [69]. They identified an uncharacterized gene cluster that may be linked to erythrocyte invasion, but they were unsuccessful in identifying genes associated with hypnozoites. However, only recently, the complete genome of *P. cynomology* was published where the authors found nine candidate genes implicated in hypnozoite formation [45].

Although a large amount of new findings were revealed by comparative genomics of closely related species, there is still a long way to understand the unique and common features among parasites of the same genus or family. A huge problem is the large proportion of genes that have unknown function. The inclusion of transcriptome, proteome or metablome data and further experimental work will significantly increase our knowledge about parasite lifestyles and pathogenicity in the future.

## 2.4 Comparative Genomics within Species

The youngest field in comparative genomics of parasitic eukaryotes is the comparison of complete genomes within species. This discipline shares some approaches applied to genus/family comparisons but at a finer level. Differences and similarities within species may give information about evolutionary history, genetic diversity and population structure, other studies may involve evaluation of vaccine candidates, identification of virulence or drug resistance factors [77,78]. So far, only few studies have been published of eukaryotes, most from the malaria field. A genome-wide survey revealed around 47'000 SNPs across the *P. falciparum* lab strains 3D7, HB3, DD2 and 16 geographically distinct isolates [79]. As expected, the study demonstrated that genes associated with antigenic variation and cytoadherence showed the highest nucleotide diversity, whereas housekeeping genes lacked nucleotide variation [79]. Further, recent selective sweeps attributed to chloroquine- and pyrimethamine-resistance were identified by searching for chromosome regions exhibiting low polymorphism in resistant populations compared to sensitives [79]. Another study surveyed genes for potential vaccine candidates where the nucleotide diversity of about 65% of the predicted genes was investigated [80]. Genes annotated as antigens were found to be under positive selection in agreement with the hypothesis that they were exposed to the host immune system [80]. To find further potential immune targets, the authors searched the *P. falciparum* genomes for highly polymorphic genes [80]. Over 50% of the 56 highly polymorphic genes identified had no functional annotation. Intriguingly, 57% contained a signal peptide and/or transmembrane domain suggesting host-parasite interaction [80].

A similar genome-wide variation analysis was undertaken within *P. vivax* strains [81]. It was shown that *P. vivax* exhibits greater genetic diversity than *P. falciparum* [81]. The authors speculated that the lower diversity in *P. falciparum* may be due to drug-induced selective sweeps, creating a bottleneck in recent history which was not the case in *P. vivax* [81]. As seen for *P. falciparum*, most diverse genes and gene families were linked to red blood cell invasion and immune evasion [81]. A similar picture was observed within strains of *P. cynomologi*, the closest relative of *P. vivax*, where genes under positive selection were predicted to have transmembrane domains including annotated antigens and transporters [45].

Comparative genomics within species plays also an increasingly important role in the discovery of new mutations underlying drug resistance. Whole genome sequencing holds the advantage that all mutations can be discovered at the genomic level that occurred between drug-resistant and drug-sensitive lines. A successful study addressed the identification of mutations conferring artemisinin resistance [82]. Comparative genomic analysis between different artemisinin-resistant *P. chabaudi* lines showed that a point mutation in *ubp1*, which encodes a ubiquitin-specific protease, seemed to be the predominant factor leading to resistance as no other shared mutation was found among the resistant lines when compared to the sensitive at the genomic level [82]. The researchers hope to transfer their knowledge to human pathogenic *P. falciparum* and *P. vivax* and use their model not only to detect mutations in response to the current drugs but also for new ones, to obtain genetic markers for resistance prior to the introduction of new drugs, allowing to detect resistance in the field [82,83]. Moreover, the proposed approach can be used to evaluate possible partners for combination therapy to avoid eventual shared resistance mechanisms [82,83].

Although only few publications are available, the power of comparative genomics within species to gain new insights into parasites is indisputable and it will certainly further advance our knowledge in parasitology and other fields.

In summary, comparative genomics is an emerging field that harbors a wide range of applications. Comparative genomics of parasites may address questions at different levels starting with characteristics that distinguish parasites from free-living organisms through to the identification of specific mutations underlying drug resistance. Thus, comparative genomics is a powerful tool that augments our understanding of parasitology and other biological areas. Many genomes from species throughout the tree of life have been sequenced in the past and are freely available to be used and explored. Moreover, new technologies make whole genome sequencing affordable for smaller research projects where researchers can answer their own specific questions beyond the individual gene level. As comparative genomics offers manifold investigations to various fascinating topics, I applied several approaches at different levels in my PhD thesis.

# 3 Objectives

The aim of my PhD thesis was on the one hand to explore and exploit the plethora of genome data available from parasites, on the other hand to produce own sequencing data from species that allow to answer specific questions in key areas of parasite chemotherapy. In this thesis, all four previously described disciplines of comparative genomic analyses (Figure 1) were used to shed more light to various aspects of parasitology. In particular, three generally applicable *in silico* tools were developed and included in the *Dirofilaria immits* genome project. Moreover, drug resistance candidate genes were identified in African trypanosomes by whole genome sequencing. The following specific objectives were accomplished by comparative genomics of parasites:

(i) Invention of an algorithm and development of an automated tool for genome-wide identification of molecular mimicry candidates that can be adopted to any host-pathogen pair (*Chapter 2*)

(ii) Invention of an algorithm and development of an automated tool that allows to discriminate DNA sequences to the level of species based on palindrome frequency patterns (*Chapter 3*)

(iii) Invention of a generally usable *in silico* drug target prediction pipeline and application to *Plasmodium falciparum* (*Chapter 4*)

(iv) Application of the invented algorithms from (i, ii, iii) to the international genome project of *D. immits* (*Chapter 5*)

(v) Whole genome sequencing of *T. b. rhodesiense* STIB900 and its two drug-resistant derivatives STIB900-M and STIB900-P, and identification of candidate resistance mutations (*Chapter 6*)

# References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496–512.

2. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. Science 274: 546, 563–567.

3. Pevsner J (2009) Bioinformatics and Functional Genomics. 2nd ed. Wiley-Blackwell.

4. Genome sequence of the nematode *C. elegans*: a platform for investigating biology (1998) Science 282: 2012–2018.

5. Bürglin TR, Lobos E, Blaxter ML (1998) *Caenorhabditis elegans* as a model for parasitic nematodes. Int. J. Parasitol. 28: 395–411.

6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921. doi:10.1038/35057062

7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351. doi:10.1126/science.1058040

8. Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature 415: 871–880. doi:10.1038/nature724

9. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562. doi:10.1038/nature01262

10. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493–521. doi:10.1038/nature02426

11. Eichinger L, Pachebat JA, Glöckner G, Rajandream M-A, Sucgang R, et al. (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435: 43–57. doi:10.1038/nature03481

12. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270: 397–403.

13. Mushegian AR (2007) 4 - Getting Ready for the Era of Comparative Genomics: The Importance of Viruses. Foundations of Comparative Genomics. Burlington: Academic Press. pp. 33–50. Available: http://www.sciencedirect.com/science/article/pii/B9780120887941500045. Accessed 20 Aug 2012.

14. Hardison RC (2003) Comparative genomics. PLoS Biol. 1: E58. doi:10.1371/journal.pbio.0000058

15. Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. Mol. Ecol 17: 4586–4596. doi:10.1111/j.1365-294X.2008.03954.x

16. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273: 1058–1073.

17. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, et al. (2000) Comparative genomics of the eukaryotes. Science 287: 2204–2215.

18. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, et al. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414: 450–453. doi:10.1038/35106579

19. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498–511. doi:10.1038/nature01097

20. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature 419: 512–519. doi:10.1038/nature01099

21. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, et al. (2004) The genome of *Cryptosporidium hominis*. Nature 431: 1107–1112. doi:10.1038/nature02977

22. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, et al. (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304: 441–445. doi:10.1126/science.1094786

23. Loftus B, Anderson I, Davies R, Alsmark UCM, Samuelson J, et al. (2005) The genome of the protist parasite *Entamoeba histolytica*. Nature 433: 865–868. doi:10.1038/nature03291

24. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. Science 309: 436–442. doi:10.1126/science.1112680

25. Hall N, Karras M, Raine JD, Carlton JM, Kooij TWA, et al. (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science 307: 82–86. doi:10.1126/science.1103717

26. Pain A, Renauld H, Berriman M, Murphy L, Yeats CA, et al. (2005) Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. Science 309: 131–133. doi:10.1126/science.1110418

27. Gardner MJ, Bishop R, Shah T, De Villiers EP, Carlton JM, et al. (2005) Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. Science 309: 134–137. doi:10.1126/science.1110439

28. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The genome of the African trypanosome *Trypanosoma brucei*. Science 309: 416–422. doi:10.1126/science.1112642

29. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, et al. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309: 409–415. doi:10.1126/science.1112631

30. Brayton KA, Lau AOT, Herndon DR, Hannick L, Kappmeyer LS, et al. (2007) Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 3: 1401–1413. doi:10.1371/journal.ppat.0030148

31. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, et al. (2007) Draft genome of the filarial nematode parasite *Brugia malayi*. Science 317: 1756–1760. doi:10.1126/science.1145406

32. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, et al. (2007) Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science 317: 1921–1926. doi:10.1126/science.1143837

33. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, et al. (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat. Genet. 39: 839–847. doi:10.1038/ng2053

34. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, et al. (2007) Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science 315: 207–212. doi:10.1126/science.1132894

35. Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, et al. (2008) Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. Proc. Natl. Acad. Sci. U.S.A. 105: 14802–14807. doi:10.1073/pnas.0805946105

36. Abad P, Gouzy J, Aury J-M, Castagnone-Sereno P, Danchin EGJ, et al. (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nat. Biotechnol. 26: 909–915. doi:10.1038/nbt.1482

37. Pain A, Böhme U, Berry AE, Mungall K, Finn RD, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455: 799–803. doi:10.1038/nature07306

38. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455: 757–763. doi:10.1038/nature07327

39. The *Schistosoma japonicum* genome reveals features of host-parasite interplay (2009) Nature 460: 345–351. doi:10.1038/nature08140

40. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, et al. (2009) The genome of the blood fluke *Schistosoma mansoni*. Nature 460: 352–358. doi:10.1038/nature08160

41. Jex AR, Liu S, Li B, Young ND, Hall RS, et al. (2011) *Ascaris suum* draft genome. Nature 479: 529–533. doi:10.1038/nature10553

42. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, et al. (2011) Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. Genome Res. 21: 2129–2142. doi:10.1101/gr.122945.111

43. Raymond F, Boisvert S, Roy G, Ritt J-F, Légaré D, et al. (2012) Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. Nucleic Acids Res. 40: 1131–1147. doi:10.1093/nar/gkr834

44. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, et al. (2011) The draft genome of the parasitic nematode *Trichinella spiralis*. Nat. Genet. 43: 228–235. doi:10.1038/ng.769

45. Tachibana S-I, Sullivan SA, Kawai S, Nakamura S, Kim HR, et al. (2012) *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat. Genet.. Available: http://www.ncbi.nlm.nih.gov/pubmed/22863735. Accessed 14 Aug 2012.

46. Dieterich C, Sommer RJ (2009) How to become a parasite - lessons from the genomes of nematodes. Trends Genet. 25: 203–209. doi:10.1016/j.tig.2009.03.006

47. Zhang J (2000) Protein-length distributions for the three domains of life. Trends Genet. 16: 107–109.

48. Nerima B, Nilsson D, Mäser P (2010) Comparative genomics of metabolic networks of free-living and parasitic eukaryotes. BMC Genomics 11: 217. doi:10.1186/1471-2164-11-217

49. De Koning HP, Bridges DJ, Burchmore RJS (2005) Purine and pyrimidine transport in pathogenic protozoa: from biology to therapy. FEMS Microbiol. Rev. 29: 987–1020. doi:10.1016/j.femsre.2005.03.004

50. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25: 25–29. doi:10.1038/75556

51. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, et al. (2008) The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat. Genet. 40: 1193–1198. doi:10.1038/ng.227

52. Jackson AP, Quail MA, Berriman M (2008) Insights into the genome sequence of a free-living Kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa). BMC Genomics 9: 594. doi:10.1186/1471-2164-9-594

53. Moore RB, Oborník M, Janouskovec J, Chrudimský T, Vancová M, et al. (2008) A photosynthetic alveolate closely related to apicomplexan parasites. Nature 451: 959–963. doi:10.1038/nature06635

54. Ralph SA, Van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, et al. (2004) Tropical infectious diseases: metabolic maps and functions of the *Plasmodium falciparum* apicoplast. Nat. Rev. Microbiol. 2: 203–216. doi:10.1038/nrmicro843

55. Wilson RJ, Denny PW, Preiser PR, Rangachari K, Roberts K, et al. (1996) Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. J. Mol. Biol. 261: 155–172.

56. Chaudhary K, Roos DS (2005) Protozoan genomics for drug discovery. Nat. Biotechnol. 23: 1089–1091. doi:10.1038/nbt0905-1089

57. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, et al. (1999) Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 285: 1573–1576.

58. Anthony MP, Burrows JN, Duparc S, Jmoehrle J, Wells TN (2012) The global pipeline of new medicines for the control and elimination of malaria. Malar. J. 11: 316. doi:10.1186/1475-2875-11-316

59. Myler PJ (2008) Searching the Tritryp genomes for drug targets. Adv. Exp. Med. Biol. 625: 133–140. doi:10.1007/978-0-387-77570-8_11

60. Damian RT (1964) Molecular mimicry: antigen sharing by parasite and host and its consequences. Am Naturalist 98: 129–149.

61. Ludin, Philipp (2009) Proteome-wide surveys for molecular mimicry in parasites. MSc Thesis.

62. Maizels RM, Balic A, Gomez-Escobar N, Nair M, Taylor MD, et al. (2004) Helminth parasites--masters of regulation. Immunol. Rev 201: 89–116. doi:10.1111/j.0105-2896.2004.00191.x

63. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298: 129–149. doi:10.1126/science.1076181

64. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, et al. (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316: 1718–1723. doi:10.1126/science.1138878

65. Teixeira SM, De Paiva RMC, Kangussu-Marcolino MM, Darocha WD (2012) Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases. Genet. Mol. Biol. 35: 1–17.

66. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science 309: 404–409. doi:10.1126/science.1112181

67. Wright AD, Li S, Feng S, Martin DS, Lynn DH (1999) Phylogenetic position of the kinetoplastids, *Cryptobia bullocki*, *Cryptobia catostomi*, and *Cryptobia salmositica* and monophyly of the genus *Trypanosoma* inferred from small subunit ribosomal RNA sequences. Mol. Biochem. Parasitol. 99: 69–76.

68. Kooij TWA, Carlton JM, Bidwell SL, Hall N, Ramesar J, et al. (2005) A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. PLoS Pathog. 1: e44. doi:10.1371/journal.ppat.0010044

69. Frech C, Chen N (2011) Genome comparison of human and non-human malaria parasites reveals species subset-specific genes potentially linked to human disease. PLoS Comput. Biol. 7: e1002320. doi:10.1371/journal.pcbi.1002320

70. Martinsen ES, Perkins SL, Schall JJ (2008) A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): evolution of life-history traits and host switches. Mol. Phylogenet. Evol. 47: 261–273. doi:10.1016/j.ympev.2007.11.012

71. Pasternak ND, Dzikowski R (2009) PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum*. Int. J. Biochem. Cell Biol. 41: 1463–1466. doi:10.1016/j.biocel.2008.12.012

72. Cunningham D, Lawton J, Jarra W, Preiser P, Langhorne J (2010) The *pir* multigene family of *Plasmodium*: antigenic variation and beyond. Mol. Biochem. Parasitol. 170: 65–73. doi:10.1016/j.molbiopara.2009.12.010

73. Kantele A, Jokiranta TS (2011) Review of cases with the emerging fifth human malaria parasite, *Plasmodium knowlesi*. Clin. Infect. Dis. 52: 1356–1362. doi:10.1093/cid/cir180

74. Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, et al. (2009) Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. Lancet Infect Dis 9: 555–566. doi:10.1016/S1473-3099(09)70177-X

75. Suwanarusk R, Cooke BM, Dondorp AM, Silamut K, Sattabongkot J, et al. (2004) The deformability of red blood cells parasitized by *Plasmodium falciparum* and *P. vivax*. J. Infect. Dis. 189: 190–194. doi:10.1086/380468

76. Marti M, Good RT, Rug M, Knuepfer E, Cowman AF (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 306: 1930–1933. doi:10.1126/science.1102452

77. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461: 1243–1247. doi:10.1038/nature08480

78. Comas I, Borrell S, Roetzer A, Rose G, Malla B, et al. (2012) Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat. Genet. 44: 106–110. doi:10.1038/ng.1038

79. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, et al. (2007) A genome-wide map of diversity in *Plasmodium falciparum*. Nat. Genet. 39: 113–119. doi:10.1038/ng1930

80. Mu J, Awadalla P, Duan J, McGee KM, Keebler J, et al. (2007) Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. Nat. Genet. 39: 126–130. doi:10.1038/ng1924

81. Neafsey DE, Galinsky K, Jiang RHY, Young L, Sykes SM, et al. (2012) The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat. Genet.. Available: http://www.ncbi.nlm.nih.gov/pubmed/22863733. Accessed 14 Aug 2012.

82. Hunt P, Martinelli A, Modrzynska K, Borges S, Creasey A, et al. (2010) Experimental evolution, genetic analysis and genome re-sequencing reveal the mutation conferring artemisinin resistance in an isogenic lineage of malaria parasites. BMC Genomics 11: 499. doi:10.1186/1471-2164-11-499

83. Borges S, Cravo P, Creasey A, Fawcett R, Modrzynska K, et al. (2011) Genomewide scan reveals amplification of mdr1 as a common denominator of resistance to mefloquine, lumefantrine, and artemisinin in *Plasmodium chabaudi* malaria parasites. Antimicrob. Agents Chemother. 55: 4858–4865. doi:10.1128/AAC.01748-10

# *Chapter 2*

## Genome-Wide Identification of Molecular Mimicry Candidates in Parasites

Philipp Ludin[1,2,3], Daniel Nilsson[1], Pascal Mäser[1,2,3]

[1] Institute of Cell Biology, University of Bern, Bern, Switzerland

[2] Swiss Tropical and Public Health Institute, Basel, Switzerland

[3] University of Basel, Basel, Switzerland

*PLoS ONE* (2011)

OPEN ACCESS Freely available online

PLoS one

# Genome-Wide Identification of Molecular Mimicry Candidates in Parasites

**Philipp Ludin**[1,2,3], **Daniel Nilsson**[1¤], **Pascal Mäser**[1,2,3*]

**1** Institute of Cell Biology, University of Bern, Bern, Switzerland, **2** Swiss Tropical and Public Health Institute, Basel, Switzerland, **3** University of Basel, Basel, Switzerland

## Abstract

Among the many strategies employed by parasites for immune evasion and host manipulation, one of the most fascinating is molecular mimicry. With genome sequences available for host and parasite, mimicry of linear amino acid epitopes can be investigated by comparative genomics. Here we developed an *in silico* pipeline for genome-wide identification of molecular mimicry candidate proteins or epitopes. The predicted proteome of a given parasite was broken down into overlapping fragments, each of which was screened for close hits in the human proteome. Control searches were carried out against unrelated, free-living eukaryotes to eliminate the generally conserved proteins, and with randomized versions of the parasite proteins to get an estimate of statistical significance. This simple but computation-intensive approach yielded interesting candidates from human-pathogenic parasites. From *Plasmodium falciparum*, it returned a 14 amino acid motif in several of the PfEMP1 variants identical to part of the heparin-binding domain in the immunosuppressive serum protein vitronectin. And in *Brugia malayi*, fragments were detected that matched to periphilin-1, a protein of cell-cell junctions involved in barrier formation. All the results are publicly available by means of mimicDB, a searchable online database for molecular mimicry candidates from pathogens. To our knowledge, this is the first genome-wide survey for molecular mimicry proteins in parasites. The strategy can be adopted to any pair of host and pathogen, once appropriate negative control organisms are chosen. MimicDB provides a host of new starting points to gain insights into the molecular nature of host-pathogen interactions.

**Citation:** Ludin P, Nilsson D, Mäser P (2011) Genome-Wide Identification of Molecular Mimicry Candidates in Parasites. PLoS ONE 6(3): e17546. doi:10.1371/journal.pone.0017546

**Editor:** Najib El-Sayed, The University of Maryland, United States of America

**Received** October 8, 2010; **Accepted** February 8, 2011; **Published** March 8, 2011

**Funding:** Daniel Nilsson was supported by a fellowship of the Swedish National Science Foundation, and Pacal Mäser was supported by a research professorship grant of the Swiss National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: pascal.maeser@unibas.ch

¤ Current address: Science for Life Laboratory, Karolinska Institutet, Solna, Sweden

## Introduction

Endoparasites are confronted with host defenses at multiple levels: physical barriers, innate immunity, and adaptive immune responses need to be overcome in order to successfully establish an infection and proliferate inside a host. Antigenic variation to escape humoral responses is well documented for the malaria parasites, *Giardia*, African trypanosomes, etc. Further strategies for immune evasion or immune suppression are less well understood. Molecular mimicry as a strategy for immune evasion and host manipulation is well known from viruses [1,2]. While many viruses have a natural propensity to acquire genetic material or proteins from the host cell upon formation of virions, others have by themselves evolved surface proteins for mimicry, e.g. the chemokine receptors of cytomegalovirus [3]. The term molecular mimicry was coined by R. Damian in 1964 and defined as the sharing of antigens between parasite and host [4]. We refer here to molecular mimicry as the display of any structure by the parasite that (i) resembles structures of the host at the molecular level and (ii) confers a benefit to the parasite because of this resemblance. The potential benefits of molecular mimicry include camouflage – as exemplified by the concept of 'eclipsed antigens' which are not recognized as such by the host's immune system due to their similarity to host antigens [5] – and cytoadherence. For intracellular parasites, cytoadherence is a prerequisite to infection. Trypomastigote *T. cruzi* adhere to fibroblasts via the fibronectin receptor, and exogenous peptides with fibronectin RGD motifs inhibited host cell invasion [6,7]. Cytoadherence of *P. falciparum*-infected erythrocytes to microvascular endothelium contributes to cerebral malaria pathology. *P. falciparum* erythrocyte membrane protein 1 (PfEMP1, encoded by the *var* genes) interacts with adhesion molecules such as ICAM-1, CD36, or thrombospondin via different domains [8,9]. Endothelial adherence prevents the infected erythrocytes from passage to the spleen where they would be eliminated. A third reason why parasites might mimic host molecules is signaling. Parasites may mimic hormone receptors to respond to signals from the host, or mimic hormones to send signals to the host. Functional homologues of the mammalian epidermal growth factor (EGF) receptor were described from trypanosomes [10,11] and helminths [12,13]. *Plasmodium* spp. possess at least two surface proteins with EGF motifs, one (Pfs25) expressed in the mosquito [14], the other (MSP1) in the blood-stages where it is critical for erythrocyte invasion [15,16]. Schistosomes send immunosuppressory signals in the form of neuropeptides to both the definite host (man) and the intermediate host (snail) [17]. There are extreme cases of behavioral manipulation of the host by the parasite such as the suicidal diving of grasshoppers infected by hairworms, and there too molecular mimicry is likely to play a role [18].

The first evidence for molecular mimicry between parasite and host came from immunological studies on antisera that cross-reacted with parasite and host. *Ascaris lumbricoides* was found to possess A- and B-like blood group antigens [19]. This was confirmed by more recent studies, which suggested that these antigens had been acquired from host blood [20]. Biosynthesis of human blood group-like antigens was described for *Schistosoma mansoni* [21,22] and *Fasciola hepatica* [23]. However, the function of these antigens produced by the parasite remains to be elucidated. More recently, tools other than antisera were used to address molecular mimicry between parasite and host. Molecular cloning of the involved genes [24,25], elucidation of polysaccharide structures [26], use of monoclonal antibodies [27,28] and synthetic peptides [29] have all contributed to a wealth of evidence that endoparasites take advantage of molecular mimicry to survive in their hosts (see also Table 1). Recurring targets for mimicry by bloodborne pathogens are the components of the complement system, growth hormones and their receptors, and cell adhesion molecules [30]. A parasite's ability to perform molecular mimicry may stem from either having acquired macromolecules from the host (transfer) or from adaptive evolution of the mimicking structures (convergence). Both scenarios are supported by multiple examples from parasites (Table 1). With the rapidly growing number of fully sequenced genomes, direct comparison between host and parasite protein sequences provides a powerful tool to identify molecular mimicry candidates. To our knowledge, however, there has been no systematic approach to study molecular mimicry since parasitology entered the post-genomic era.

Here we develop an *in silico* pipeline to identify molecular mimicry candidates from parasites. In brief, proteome-wide blast surveys were performed with either whole proteins or with overlapping protein fragments to identify similar epitopes in parasite and host. This approach warrants that all linear amino acid epitopes which share significant similarity between parasite and host will be discovered. Searches against control proteomes of free-living eukaryotes served as negative controls to exclude proteins that are generally conserved across phyla, while searches with random sequences allowed to estimate statistical significance. The results are made available by means of an online database for molecular mimicry candidate proteins in pathogens.

## Results and Discussion

### Molecular mimicry surveys with full length protein sequences

In pilot surveys for molecular mimicry candidates we concentrated on endoparasitic helminths since (i) they are known masters of immune evasion and host manipulation, and (ii) a convenient negative control is available in the form of the free-living nematode *C. elegans*. In principal, a mimicry candidate is a parasite protein or motif which bears a high degree of resemblance to a protein of the host but not to those of unrelated control species. Such proteins are readily identified by proteome-wide blast surveys. In a first trial, we ran every predicted protein of *Brugia malayi* with blastp against the proteomes of *H. sapiens* and *C. elegans*. As expected, the *B. malayi* proteins returned significantly ($p<0.0001$, two-tailed Wilcoxon test) higher scores against *C. elegans* than against *H. sapiens*. There were only few *B. malayi* proteins which scored better against the human host (Figure 1, left). The converse picture emerged when the same procedure was carried out with *Schistosoma mansoni* (Figure 1, right) or *S. japonicum* (not shown), where the parasite proteins generally were more similar to human than to *C. elegans* proteins ($p<0.0001$, two-tailed Wilcoxon test). The systemic nature of the phenomenon (Figure 1, right) speaks against molecular mimicry as the underlying selective force since it involves too many housekeeping proteins that do not interact with the host. *C. elegans* and *S. mansoni* are from different metazoan clades, the ecdysozoa and the lophotrochozoa, respectively [31]. While the *S. mansoni* proteins were also more similar to *D. melanogaster* than to *C. elegans* proteins, the overall similarity to human proteins was still the most pronounced (not shown).

The two-dimensional blastp approach allowed to graphically divide the proteome of *B. malayi* into separate quadrants: parasite-specific proteins (lower left in Figure 1, left), generally conserved proteins such as tubulin or ubiquitin (upper right), nematode-specific proteins (upper left), and mimicry candidates (lower right). However, this rough subdivision is prone to false positives caused by the well documented phenomenon of gene loss in *C. elegans* [32]. In order to eliminate proteins which are generally conserved, the negative control was refined to include – in addition to *C. elegans* – a panel of unrelated, free-living eukaryotes whose genomes have been sequenced: *Saccharomyces pombe*, *Arabidopsis thaliana*, *Ciona intestinalis*, and *Trichoplax adhaerens* (Table 2). For the detection of mimicry candidates we focused on human-pathogenic endoparasites known for their mastery in immune evasion, namely *Brugia malayi*, *Schistosoma mansoni*, *Plasmodium falciparum*, *Leishmania major*, *Cryptosporidium parvum*, *Trichomonas vaginalis* and *Trypanosoma cruzi* (Table 2). The predicted proteomes of the parasites were run as blast queries against the control proteomes and against *H. sapiens*. Molecular mimicry candidates were defined as parasite proteins with (i) a blastp score above 100 to the best hit in the human proteome and (ii) a score in *H. sapiens* at least two-fold higher than the best score achieved in the control proteomes. This

**Table 1.** Possible mechanism for molecular mimicry and examples from pathogens.

| Macromolecule | Mimicry by transfer | Mimicry by convergence |
|---|---|---|
| **Nucleic acid** | *Schistosoma mansoni* possesses a CRIT gene which shares 98% identical nucleotides with the human orthologue [25]. | The 3′UTR of the RNA genome of barley yellow dwarf virus mimics the $m^7G$ cap of eukaryotic mRNA to stimulate translation [59]. |
| **Protein** | Pathogenic bacteria, *E. granulosus* and *O. volvulus* decorate themselves with inhibitors of the complement cascade sequestered from the blood [43,60,61,62]. | A 18 aa motif in *P. falciparum* CSP is nearly identical to the cytoadhesive region of mammalian thrombospondin [49] and was shown to bind to hepatocytes [63]. |
| **Sugar** | Trans-sialidases transfer sialic acid from host cells to the surface of the parasite. *T. cruzi* trans-sialidase is a virulence factor in mammals [64]; *T. brucei* trans-sialidase is required for survival in the tsetse fly [65]. | Several pathogenic helminths synthesize the Forssman antigen (globopentosylceramide) [21,66], a glycolipid implicated in cell adhesion and the formation of tight junctions [67]. |

Mimicry by transfer of nucleic acids or convergence of proteins can be identified *in silico* by comparative genomics (CRIT: Complement C2 receptor inhibitor trispanning (CRIT), C4BP: $m^7G$, 7-methyl guanosine; CSP, circumsporozoite protein; Complement-binding protein, CR1: Complement receptor 1, FHL-1: factor-H-like protein-1, fH: factor H, MCP: Membrane cofactor protein, DAF: Decay-accelerating factor).

doi:10.1371/journal.pone.0017546.t001

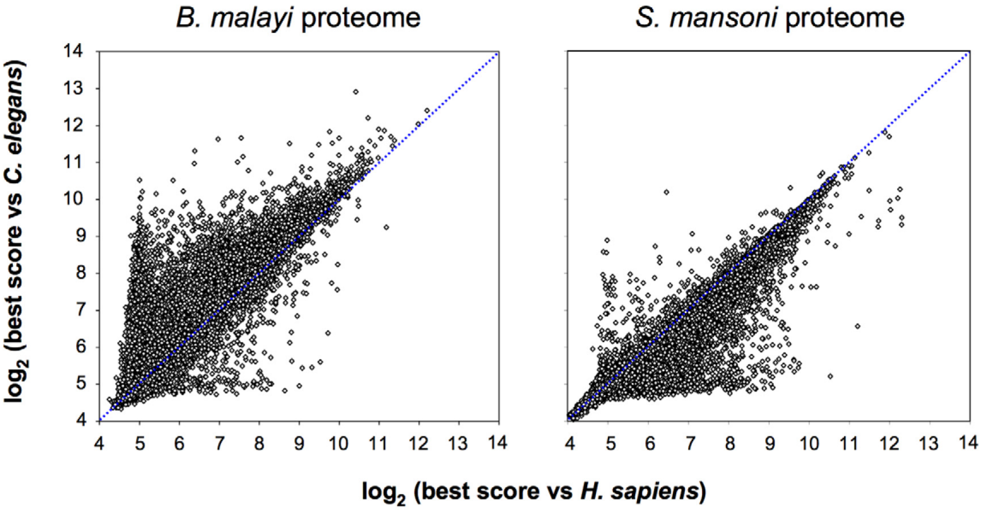


**Figure 1. Scatter plot of the blast scores of all proteins from *B. malayi* (left) and *S. mansoni* (right) vs. the host *H. sapiens* (x-axis) and the control *C. elegans* (y-axis).** Points below the blue dotted line represent parasite proteins with better scores to *H. sapiens* than to *C. elegans*.
doi:10.1371/journal.pone.0017546.g001

search returned 84 hits, most of which from *S. mansoni* (52) and *B. malayi* (15; Table S1). One hit from *B. malayi* was a predicted protein (A8NPN8) with strong similarity to human suppressor of cytokine signaling 5 (SOCS5), in particular to the SH2 domain and the SOCS box (Figure 2). Human SOCS5 was shown to inhibit the IL-4 pathway in T helper cells, promoting T$_H$1 differentiation [33]. The SH2 domain recognizes the target molecule and the SOCS box recruits the ubiquitin complex that mediates proteosomal degradation of the target [34]. SOCS proteins being crucial regulators of both innate and adaptive immunity, the SOCS5-like protein from *B. malayi* is an interesting candidate. However, it does not carry an export signal and it is therefore not clear how it should interact with host proteins. Possibly, it is released when parasites die.

The known mimicry candidate CRIT (complement C2 receptor inhibitory trispanning, Table 1), which is almost identical between *S. mansoni* and *H. sapiens* [35], was not identified here because human CRIT is not included in the reviewed human proteome from Swissprot (Table 2). Searching against the whole human Uniprot dataset readily returned *S. mansoni* CRIT as the top hit. In the classical complement pathway CRIT blocks the formation of C3 convertase by decreasing the association of C2 with C4b; once C2 is attached to the receptor, it cannot be cleaved by C1 to produce C2a and C2b and thus C3 convertase is no longer formed – the classical pathway is disrupted [25]. It is easy to conceive that a parasite gains an advantage in the human body by exhibiting CRIT and diminishing the proinflammatory response. Based on the high level of DNA similarity *S. mansoni* is thought to have acquired the CRIT gene by horizontal transfer [25,35]. However, while CRIT orthologues are present in all of the sequenced *Schistosoma* species and in *T. cruzi*, the only mammals which possess CRIT are man and rat (Figure S1). This enigmatic distribution can only be explained by multiple instances of gene transfer or gene loss in mammals. Postulating a minimal number of horizontal transfers, a parsimonial interpretation would place the origin of the CRIT gene to schistosomes. The gene could have been acquired (exapted) from the parasites by *H. sapiens* and *R. norvegicus* independently, and finally picked up by *T. cruzi* from a mammalian host. In this scenario, only the CRIT of *T. cruzi* would be a case of molecular mimicry.

## Molecular mimicry surveys with fragmented protein sequences

Several known cases of molecular mimicry from parasites (Table 1) involve shorter peptides, e.g. the thrombospondin motif in *P. falciparum* circumsporozoite protein CSP. Such mimicry

**Table 2.** Organisms used in this study.

| Species | Proteins | Source | Ref. |
|---|---|---|---|
| *Brugia malayi* | 11551 | Uniprot | [68] |
| *Cryptosporidium parvum* | 3805 | CryptoDB | [69] |
| *Giardia lamblia* | 5901 | GiardiaDB | [70] |
| *Leishmania major* | 8406 | TritrypDB | [71] |
| *Plasmodium falciparum* | 5479 | PlasmoDB | [72] |
| *Schistosoma mansoni* | 13157 | Sanger | [73] |
| *Trichomonas vaginalis* | 50155 | Uniprot | [74] |
| *Trypanosoma cruzi* | 23031 | TritrypDB | [75] |
| *Homo sapiens* | 20298 | Uniprot | [76] |
| *Aedes aegypti* | 16531 | Vectorbase | [77] |
| *Anopheles gambiae* | 14103 | Vectorbase | [78] |
| *Arabidopsis thaliana* | 36671 | EBI | [79] |
| *Caenorhabditis elegans* | 24143 | Wormbase | [80] |
| *Ciona intestinalis* | 15852 | JGI | [81] |
| *Schizosaccharomyces pombe* | 4977 | EBI | [82] |
| *Trichoplax adhaerens* | 11585 | Uniprot | [83] |

Parasite (top), host (middle), and negative control species (bottom), their predicted number of protein-coding genes, and source of the predicted proteome file (EBI: European Bioinformatics Institute, JGI: Joint Genome Institute).
doi:10.1371/journal.pone.0017546.t002

```
SOCS5     TQIDYIHCLVPDLLQITGNPCYWGVMDRYEAEALLEGKPEGTFLLRDSAQEDYLFSVSFR   420
A8NPN8    TSVHYTNCLVPRLDLIIDSSYYWGIMDRYEAEALLDNKPEGTFLLRDSAQSEYLFSVSFR   144
          *.:.* :**** *  * ... ***:**********:.*************.:********

SOCS5     RYNRSLHARIEQWNHNFSFDAHDPCVFHSSTVTGLLEHYKDPSSCMFFEPLLTISLNRTF   480
A8NPN8    RYKRTLHARIEQKNHRFSFDFSDPSIYSANTITKLISYYKDPTKCLFFEPQLSVPLPRNF   204
          **:*:******* **.****  **.:: :.*:* *:.:****:.*:**** *::.* *.*

SOCS5     PFSLQYICRAVICRCTTYDGIDGLPLPSMLQDFLKEYHYKQKVRVRWLERE---------   531
A8NPN8    VFSLQHLCRARIASLTTYDGVEKLNLPVSLKNFIKEYHYKHPVKTVNYTPDTDLLHAYTL   264
          ****::*** *.  *****:: * **  *::*:******: *:.     :
```

**Figure 2. ClustalW alignment of the candidate mimicry region in A8NPN8 from *B. malayi* to *H. sapiens* SOCS5.** The SH2 domain is shaded in yellow, the SOCS box domain in blue. The N-terminal parts of the two proteins do not share any similarity (not shown).
doi:10.1371/journal.pone.0017546.g002

candidates would not be detected with the above approach using full-length protein sequences. Thus we refined the systematic survey and developed a peptide-based pipeline for detection of mimicry candidates as outlined in Figure 3. In brief, the parasite proteins were converted to a series of overlapping 14-mers, each of which was searched with ungapped blastp against the control proteomes *C. elegans*, *S. pombe*, *A. thaliana*, *C. intestinalis*, or *T. adhaerens*. The 14-mers with high similarity to any sequence of the controls were filtered out using an empirically developed scheme (Figure S2). The remainder of the 14-mers was screened, again with ungapped blastp, against the *H. sapiens* proteome and those exhibiting strong similarity (Figure S2) to a human sequence were identified as molecular mimicry candidates. For this approach, predicted N-terminal protein export signal sequences were removed since they resemble each other and might produce false positive hits. Parasite 14-mers with 100% identity to a human protein were obtained from *B. malayi* (4), *C. parvum* (1), *P. falciparum* (13) and *S. mansoni* (15). 14-mers with 13 identical residues to a human protein were found in all parasites except *G. lamblia*. The number of hits is summarized in Figure 4. As a control, the same approach (Figure 3) was carried out with versions of the pathogen proteomes where every sequence had been scrambled randomly. This yielded not a single 14-mer of 100% identity to a human protein over all the parasites tested, and only 4 with 13 identities in, underscoring the statistical significance of the identified mimicry candidates. The largest differences between real and randomized proteins were observed for the helminths *B. malayi* and *S. mansoni*, and for *P. falciparum*. Selected mimicry candidates from these parasites are listed in Table 3. The selection was based on number of identical residues, Shannon-entropy of the respective 14-mer as a measure of sequence heterogeneity, and GO terms associated with the hit in the human proteome. An overview of all the high-level GO terms of the human proteins which were matched with mimicry candidates from parasites is shown in Table S2. The mimicry candidates of *P. falciparum* enriched for 'Cellular component biogenesis', 'Localization', and 'Growth', while for the helminths *B. malayi* and *S. mansoni* 'Biological adhesion' and 'Rhythmic process' were overrepresented in the human hits (compared to the complete human proteome; Table S2).

Among the most interesting of the identified mimicry candidates was a match of 17 identical amino acids from *B. malayi* to human plasma glutamate carboxypeptidase. The *B. malayi* protein (A8QH34) had been previously detected in excretory-secretory products in abundance [36,37]. Moreover, the identified candidate has 67% identity to ES-62 from the rodent filarial nematode *Acanthocheilonema viteae* (Uniprot ID O76552), a protein with immunomodulatory impact on different host cells depending on the occurrence of phosphorylcholine [38]. The identified candidate stretch shares 14 identical amino acids with ES-62 of *A. viteae*. Other interesting fragments from *B. malayi* matched to human periphilin-1 (Q8NEY8), a protein of cell-cell junctions in differentiated keratinocytes which was proposed to be involved in barrier formation and epidermal integrity [39], and to plasminogen (P00747), the proenzyme of plasmin which dissolves blood clots and acts as a proteolytic factor in various other processes (Table 3).

In *P. falciparum*, the peptide-based approach significantly enriched for exoproteins ($p<0.0001$, two-sided chi square test), i.e. proteins with transmembrane domains or export signal predicted by Phobius [40]. The best hit overall was to human vitronectin. Several of the *var* family gene products turned out to share a stretch of 13 to 16 identical amino acids with vitronectin. The candidate mimicry motif lies in the extracellular part of PfEMP1, close to the predicted transmembrane domain (Figure 5, bottom). The corresponding sequence in vitronectin is in the N-terminal half, in the first of the heparin-binding motifs between the somatomedin and the central hemopexin domains (Figure 5, top). Vitronectin is a multifunctional protein that promotes cell adhesion, stabilizes plasminogen activator inhibitor 1, and inhibits the formation of the pore-forming membrane attack complex (MAC) of the complement system. Vitronectin is abundant in the extracellular matrix and in the serum [41]. Pathogenic bacteria such as *Neisseria meningitides* or *Haemophilus influenzae* decorate themselves with human vitronectin which they acquire form the serum through specific binding partners on their surface [42,43]. Bacteria also exploit human vitronectin for cytoadhesion and host cell invasion [44]. Malaria-infected erythrocytes, however, tested negative for binding to human vitronectin [45]. We identified six PfEMP1 variants possessing the candidate mimicry motif to vitronectin in the *P. falciparum* strain 3D7 and seven in the strain HB3 (Figure 5). The motif is positionally conserved relative to the transmembrane domain of PfEMP1. Searching the non-redundant protein database of GenBank with the corresponding peptide 'NPEQTPVLKPEEEAP' returned significant hits (expectancy $<0.001$) only from *H. sapiens*, Chimpanzee, Orangutan, and *P. falciparum* (not shown). Interestingly, the genome project of the simian and human malaria parasite *P. knowlesi* had uncovered a candidate molecular mimicry motif to the immunoregulatory host protein CD99 in the extracellular domain of the *kir* gene family products [46].

The fragment-based approach for mimicry candidates in *P. falciparum* also returned a triad between host, vector and parasite. Thrombospondin-related anonymous protein (TRAP, PF13_0201)

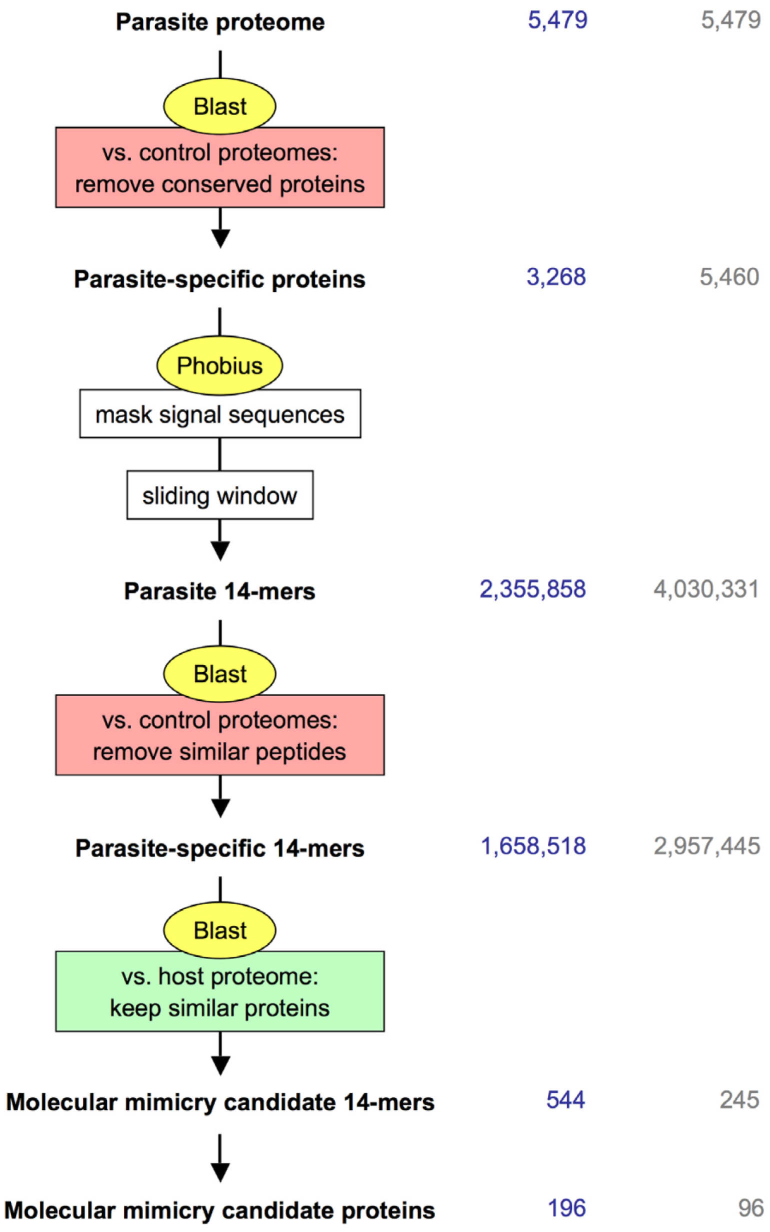


**Figure 3. The *in silico* pipeline for identification of molecular mimicry candidates from parasites.** See Methods for details. The process is illustrated with the actual numbers from the analysis of the *P. falciparum* proteome in blue, respectively a randomized version of it in grey, vs. the host *H. sapiens*.
doi:10.1371/journal.pone.0017546.g003

of *P. falciparum* matched with the human spondin (Q9HCB6) and a hypothetical protein from *A. gambiae* (AGAP012307, not shown). In the human protein, the region lies in the thrombospondin type-I repeat (TSR) domain which binds to heparin sulphate proteoglycans on hepatocytes [47,48]. This mimicked structure was also found on the circumsporozoite protein (CSP) and has been known for a long time [49]. Whereas CSP mediates the binding of the parasites to the human liver, it is suggested that TRAP is crucial for sporozoite locomotion and cell invasion [50,51]. Interestingly, the same part of the TSR domain of TRAP has been matched with the *A. gambiae* proteome and it has been demonstrated with loss-of-function mutations that this region is involved in the sporozoite invasion into mosquito salivary glands [52].

## mimicDB - Database for molecular mimicry candidates from pathogens

All mimicry candidates from parasites to mammalian and insect hosts (Table 2) were stored in a relational database, mimicDB,

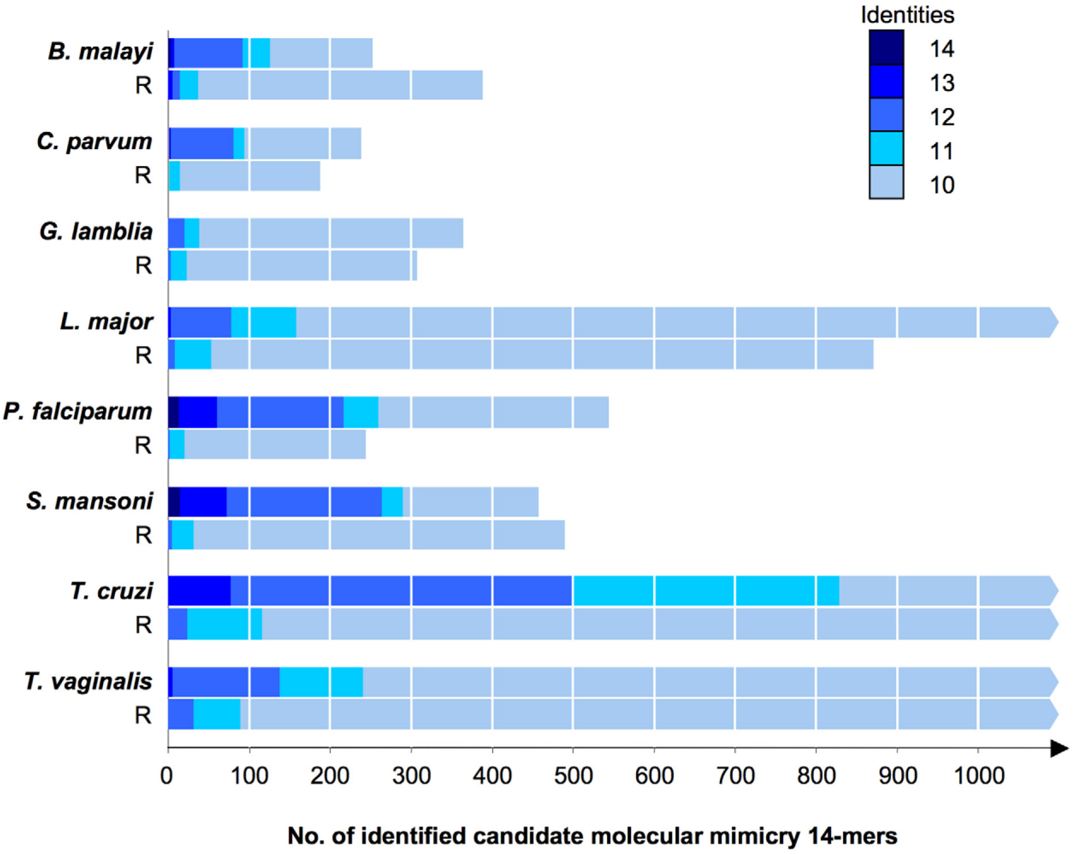


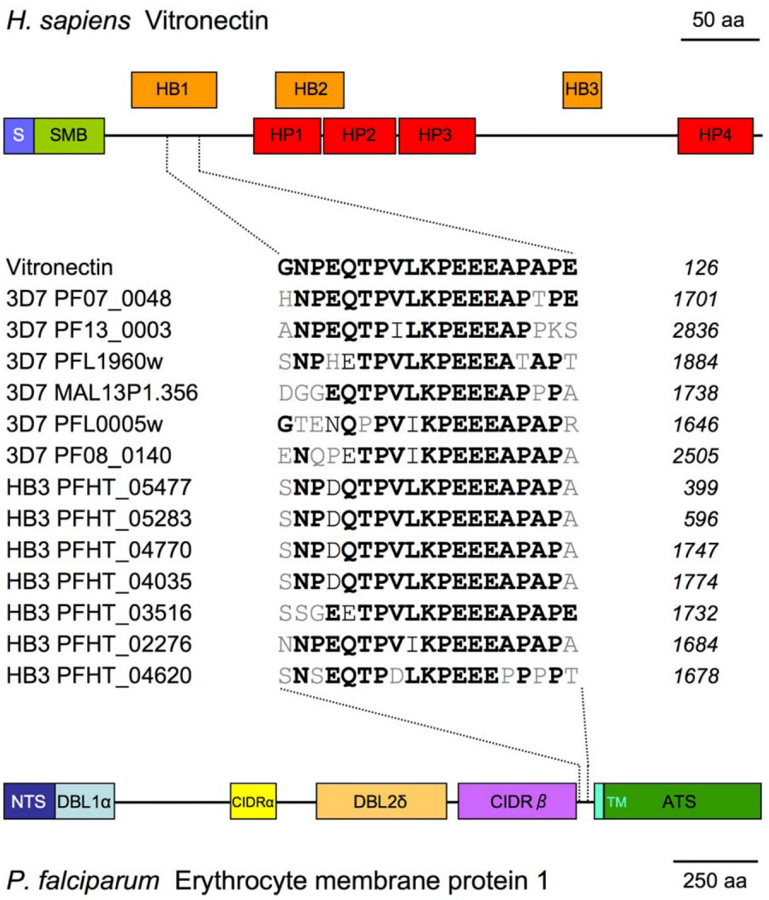
**Figure 4. Numbers of identified candidate molecular mimicry 14-mers from parasite proteomes and randomized versions thereof (R).** Numbers of amino acid identities between the 14-mers and their best hit in the human proteome are color-coded as indicated.
doi:10.1371/journal.pone.0017546.g004

which is publicly accessible via <http://mimicdb.scilifelab.se>. The database was designed for ease of community access to the mimicry data (Figure S3). It can be queried using keywords from gene description, different formats of gene and protein accession numbers and names, and in general on free text on the available data. GO terms are tightly integrated into the database, and queries can be made both on leaf-terms as well as directly onto broader categories higher up in the hierarchy. The queries can be restricted to species using special qualifiers. From the resulting tables, links are provided directly to entries in large public databases (Uniprot,

**Table 3.** Selected mimicry candidates.

| Parasite protein | 14-mer motif | Ent. | Id. | Human protein |
|---|---|---|---|---|
| *Bma* A8PSR3, uncharacterized protein | TFGFVTKMLIEKDP | 3.24 | 12 | Q8NEY8, periphilin-1 |
| *Bma* A8Q9C9, uncharacterized protein | RKSSQKIRMRDVVL | 3.04 | 12 | P00747, plasminogen |
| *Bma* A8QH34, leucyl aminopeptidase | HLDSWDVGQGAMDD | 3.09 | 14 | Q9Y646, plasma glutamate carboxypeptidase |
| *Bma* A8PP49, pregnancy-associated plasma protein E | CYIYEGDGECEPFE | 2.81 | 12 | Q9BXP8, pappalysin-2 |
| *Sma* Smp_111120, insulin receptor kinase substrate | SLEKSQAELKKIRR | 2.90 | 14 | Q9UHR4, insulin receptor tyrosine kinase substrate |
| *Sma* Smp_109770, integrin alpha-4 | APNVSMEIMVPNSF | 3.09 | 13 | P13612, integrin alpha-4 |
| *Pfa* PF07_0048, PfEMP1 | NPEQTPVLKPEEEA | 2.90 | 14 | P04004, vitronectin |
| *Pfa* MAL13P1.34, RED-like protein | SKFMGGDEEHTHLV | 3.38 | 12 | Q13123, IK cytokine |
| *Pfa* PF13_0201, TRAP | WDEWSPCSVTCGKG | 3.24 | 12 | Q9HCB6, spondin-1 |

Hits from *B. malayi* (*Bma*), *S. mansoni* (*Sma*) and *P. falciparum* (*Pfa*) and their human match (Ent, Shannon entropy in bits; Id, number of identities).
doi:10.1371/journal.pone.0017546.t003



| | | |
|---|---|---|
| Vitronectin | GNPEQTPVLKPEEEAPAPE | 126 |
| 3D7 PF07_0048 | HNPEQTPVLKPEEEAPTPE | 1701 |
| 3D7 PF13_0003 | ANPEQTPILKPEEEAPPKS | 2836 |
| 3D7 PFL1960w | SNPHETPVLKPEEEATAPT | 1884 |
| 3D7 MAL13P1.356 | DGGEQTPVLKPEEEAPPPA | 1738 |
| 3D7 PFL0005w | GTENQPPVIKPEEEAPAPR | 1646 |
| 3D7 PF08_0140 | ENQPETPVIKPEEEAPAPA | 2505 |
| HB3 PFHT_05477 | SNPDQTPVLKPEEEAPAPA | 399 |
| HB3 PFHT_05283 | SNPDQTPVLKPEEEAPAPA | 596 |
| HB3 PFHT_04770 | SNPDQTPVLKPEEEAPAPA | 1747 |
| HB3 PFHT_04035 | SNPDQTPVLKPEEEAPAPA | 1774 |
| HB3 PFHT_03516 | SSGEETPVLKPEEEAPAPE | 1732 |
| HB3 PFHT_02276 | NNPEQTPVIKPEEEAPAPA | 1684 |
| HB3 PFHT_04620 | SNSEQTPDLKPEEEPPPPT | 1678 |

**Figure 5. Alignment of human vitronectin (top) and *P. falciparum* PfEMP1 variants (bottom).** Identities to vitronectin are printed in bold black, similarities in black. The known vitronectin domains are the signal sequence (blue), somatomedin-B (green), and hemopexin (red). The known PfEMP1 domains are the N-terminal segment (dark blue), Duffy Binding Like α (light blue), cysteine-rich interdomain region α (yellow), Duffy Binding Like 2d (orange), cysteine-rich interdomain region ß (purple), transmembrane domain (cyan), acidic terminal segment (green).
doi:10.1371/journal.pone.0017546.g005

NCBI) as well as to detailed sequence views. Predicted protein motifs and signal peptides are visualized on the source and target sequences together with the candidate mimicry motifs.

## Conclusion

To our knowledge this is the first *in silico* survey for molecular mimicry candidates in parasites. Its systematic, genome-wide nature warrants that all linear amino acid epitopes involved in molecular mimicry between a given parasite and its host are going to be detected. False positive hits can be tracked by including the appropriate controls: proteomes of free-living species to eliminate the proteins which are generally conserved across phyla, and scrambled versions of the parasite proteomes to estimate for random hits resulting from the sheer number of analyzed sequences. False negatives are more problematic; mimicry by non-linear epitopes composed from amino acids of separate folds (or even separate polypeptides) will not be recognized, and neither are glycosylated epitopes (Table 1). Nevertheless, there are examples of molecular mimicry by linear epitopes which are straightforward to detect by comparative genomics as performed here. Proof of concept was obtained from the fact that the known molecular mimicry motif in TRAP (thrombospondin-related anonymous protein) from *P. falciparum* was detected readily. Many new molecular mimicry candidates were discovered from human parasites, in particular from *B. malayi*, *S. mansoni* and *P. falciparum*, most notably a sequence shared between human vitronectin and several of the *P. falciparum* erythrocyte membrane protein 1 variants. All the identified mimicry candidates are stored in a relational database called mimicDB and searchable on-line. We hope that mimicDB will stimulate research into molecular mimicry of parasites. Given its numerous potential benefits – camouflage, cytoadherence, manipulation of host signaling – molecular mimicry may well be much more common among parasitic microorganisms than currently known.

## Materials and Methods

### Proteome files

Predicted proteins from completely sequenced genomes (Table 2) were obtained from ftp.ebi.ac.uk (*Arabidopsis thaliana*, *Schizosaccharomyces pombe*), www.tritrypdb.org (*Leishmania major*, *Trypanosoma cruzi*), www.cryptodb.org (*Cryptosporidium parvum*) www.giardiadb.org (*Giardia lamblia*) www.plasmodb.org (*Plasmodium* *falciparum* 3D7), www.broadinstitute.org (*Plasmodium falciparum* HB3), ftp.vectorbase.org (*Aedes aegypti*, *Anopheles gambiae*), ftp.wormbase.org (*Caenorhabditis elegans*), ftp.sanger.ac.uk (*Schistosoma mansoni*), ftp.jgi-psf.org (*Ciona intestinalis*), and www.uniprot.org (*Brugia malayi*, *Homo sapiens*, *Trichomonas vaginalis*, *Trichoplax adhaerens*).

### Algorithm

BLAST 2.2.17 [53] was obtained from ftp.ncbi.nlm.nih.gov, Phobius 1.01 [40] from <phobius.sbc.su.se>. Automated detection of molecular mimicry candidates as depicted in Figure 3 was performed with Perl scripts, available on request. First, those of the predicted parasite proteins which are generally conserved among eukaryotes were sorted out based on full-length blastp searches against the proteomes of *C. elegans*, *C. intestinalis*, *T. adhaerens*, *S. pombe and A. thaliana*. Sequences which returned an e-value$\leq 10^{-10}$ to any sequence of these control proteomes were filtered out. The remaining parasite proteins were run through Phobius and predicted N-terminal export signal sequences were cut off at the predicted cleavage site. Then, the protein sequences were converted to a series of overlapping 14-mers with a sliding window of increment one. The resulting peptides were screened against the five control proteomes with ungapped blastp, and 14-mers above the empirically determined identity threshold (represented by the red line in Figure S2) were removed. With the remaining, parasite-specific 14-mers, an ungapped blastp search was performed against the host proteome and hits above the empirically determined identity threshold (green line in Figure S2) were considered molecular mimicry candidates. Randomized sequences were generated with 'shuffleseq' of the EMBOSS package [54]. All programs were run on the University of Bern Linux cluster, Ubelix <http://ubelix.unibe.ch>. Multiple sequence alignments were performed using ClustalX [55].

### Database

The mimicDB database (http://mimicDB.scilifelab.se) uses MySQL as its relational database engine. The database was designed as an extension to the GO term [56] database schema for ease of interrogation on the complete GO hierarchy rather than leaf term only (Figure S3). Protein motif predictions were obtained using hmmer 3.0 [57] with the PFAM database v24.0 [58], and signal peptide predictions using Phobius 1.01 [40]. *Ad hoc* Perl scripts were used to import the mimicry pipeline results, predicted motifs and signals as well as calculate Shannon source entropy for peptides. The interface was constructed using Perl and the Titanium extension to CGI.pm. A package to reconstruct the results and database is available from the authors upon request or can be downloaded from the mimicDB web site.

## Supporting Information

**Figure S1 ClustalW dendrogram of CRIT orthologues from *Schistosoma mansoni* (Sma), *S. haematobium* (Sha), *S. japonicum* (Sja), *Trypanosoma cruzi* (Tcr), *H. sapiens* (Hsa), and *R. norvegicus* (Rno).** The scale bar indicates changes per site. Bootstrapping numbers (grey) are given as percent positives of 1,000 rounds.
(TIF)

**Figure S2 The filtering system used in the overlapping fragments approach.** Numbers represent identical amino acid residues. Red line: threshold for negative control species. Green line: threshold for molecular mimicry candidate in mammalian host or insect vector.
(TIF)

**Figure S3 Database schema of mimicDB.** The mimicDB database schema centers around *mimic_sequence*, which represents the individual genes. This table has as attribute tables the actual peptide sequences (*mimic_sequence_seq*) and predicted motifs (*mimic_sequence_motif*). Hits between parts of these genes are collected in *mimic_hit*, which stores the coordinates and properties of the hit. A complexity measure, in the form of Shannon source entropy for each peptide hit is stored in *mimic_hit_entropy*. The database connects to the GO consortium GO term database in that *mimic_sequence* entries that have a GO association are referenced by entries in *mimic_sequence_with_go_association*, where the corresponding GO term db gene_product::id is also a foreign key.
(TIF)

**Table S1 All molecular mimicry candidates identified searching the human proteome with full-length protein sequences from parasites.** Scores are from blastp searches using the BLOSUM62 matrix and default parameters. Ratios are of the score against *H. sapiens* divided by the best score achieved against any of the control species *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Ciona intestinalis*, *Schizosaccharomyces pombe*, or *Trichoplax adhaerens*.
(XLS)

**Table S2 Molecular mimicry candidates identified searching the human proteome with fragmented protein sequences from parasites.** Hits are sorted according to GO (gene ontology) process annotation of the respective human target protein. Enrichment ('Enrich') of GO terms in the identified sets of target proteins is expressed in relation to the abundance of the same GO terms in the complete human proteome (last three columns).
(XLS)

## Acknowledgments

We wish to thank the University of Bern for user time on their Linux cluster Ubelix.

## Author Contributions

Conceived and designed the experiments: PL DN PM. Performed the experiments: PL DN. Analyzed the data: PL DN PM. Contributed reagents/materials/analysis tools: PL DN PM. Wrote the manuscript: PL DN PM. Designed mimicDB: DN.

## References

1. Lambris JD, Ricklin D, Geisbrecht BV (2008) Complement evasion by human pathogens. Nat Rev Microbiol 6: 132–142.
2. Srinivasappa J, Saegusa J, Prabhakar BS, Gentry MK, Buchmeier MJ, et al. (1986) Molecular mimicry: frequency of reactivity of monoclonal antiviral antibodies with normal tissues. J Virol 57: 397–401.
3. Michelson S (2004) Consequences of human cytomegalovirus mimicry. Hum Immunol 65: 465–475.
4. Damian RT (1964) Molecular mimicry: Antigen sharing by parasite and host and its consequences. American Naturalist 98: 129–149.
5. Damian RT (1962) A theory of immunoselection for eclipsed antigens of parasites and its implications for the problem of antigenic polymorphism in man. J Parasitol 48: 16.
6. Ouaissi MA, Afchain D, Capron A, Grimaud JA (1984) Fibronectin receptors on Trypanosoma cruzi trypomastigotes and their biological function. Nature 308: 380–382.
7. Ouaissi MA, Cornette J, Afchain D, Capron A, Gras-Masse H, et al. (1986) Trypanosoma cruzi infection inhibited by peptides modeled from a fibronectin cell attachment domain. Science 234: 603–607.

8. Baruch DI, Gormely JA, Ma C, Howard RJ, Pasloske BL (1996) Plasmodium falciparum erythrocyte membrane protein 1 is a parasitized erythrocyte receptor for adherence to CD36, thrombospondin, and intercellular adhesion molecule 1. Proc Natl Acad Sci U S A 93: 3497–3502.
9. Howell DP, Levin EA, Springer AL, Kraemer SM, Phippard DJ, et al. (2008) Mapping a common interaction site used by Plasmodium falciparum Duffy binding-like domains to bind diverse host receptors. Mol Microbiol 67: 78–87.
10. Hide G, Gray A, Harrison CM, Tait A (1989) Identification of an epidermal growth factor receptor homologue in trypanosomes. Mol Biochem Parasitol 36: 51–59.
11. Ghansah TJ, Ager EC, Freeman-Junior P, Villalta F, Lima MF (2002) Epidermal growth factor binds to a receptor on Trypanosoma cruzi amastigotes inducing signal transduction events and cell proliferation. J Eukaryot Microbiol 49: 383–390.
12. Spiliotis M, Kroner A, Brehm K (2003) Identification, molecular characterization and expression of the gene encoding the epidermal growth factor receptor orthologue from the fox-tapeworm Echinococcus multilocularis. Gene 323: 57–65.
13. Vicogne J, Cailliau K, Tulasne D, Browaeys E, Yan YT, et al. (2004) Conservation of epidermal growth factor receptor function in the human parasitic helminth Schistosoma mansoni. J Biol Chem 279: 37407–37414.
14. Kaslow DC, Quakyi IA, Syin C, Raum MG, Keister DB, et al. (1988) A vaccine candidate from the sexual stage of human malaria that contains EGF-like domains. Nature 333: 74–76.
15. Han HJ, Park SG, Kim SH, Hwang SY, Han J, et al. (2004) Epidermal growth factor-like motifs 1 and 2 of Plasmodium vivax merozoite surface protein 1 are critical domains in erythrocyte invasion. Biochem Biophys Res Commun 320: 563–570.
16. Blackman MJ, Ling IT, Nicholls SC, Holder AA (1991) Proteolytic processing of the Plasmodium falciparum merozoite surface protein-1 produces a membrane-bound fragment containing two epidermal growth factor-like domains. Mol Biochem Parasitol 49: 29–33.
17. Duvaux-Miret O, Stefano GB, Smith EM, Dissous C, Capron A (1992) Immunosuppression in the definitive and intermediate hosts of the human parasite Schistosoma mansoni by release of immunoactive neuropeptides. Proc Natl Acad Sci U S A 89: 778–781.
18. Biron DG, Marche L, Ponton F, Loxdale HD, Galeotti N, et al. (2005) Behavioural manipulation in a grasshopper harbouring hairworm: a proteomics approach. Proc Biol Sci 272: 2117–2126.
19. Oliver-Gonzalez J (1944) Functional antigens in helminths. J Infect Diseases 78: 232–237.
20. Ponce de Leon P, Valverde J (2003) ABO System: molecular mimicry of Ascaris lumbricoides. Rev Inst Med Trop Sao Paulo 45: 107–108.
21. Oliver-Gonzalez J, Torregrosa MV (1944) A substance in animal parasites related to human isoagglutinogens. J Infect Diseases 74: 173–177.
22. Nyame AK, Debose-Boyd R, Long TD, Tsang VC, Cummings RD (1998) Expression of Lex antigen in Schistosoma japonicum and S.haematobium and immune responses to Lex in infected animals: lack of Lex expression in other trematodes and nematodes. Glycobiology 8: 615–624.
23. Ben-Ismail R, Mulet-Clamagirand C, Carme B, Gentilini M (1982) Biosynthesis of A, H, and Lewis blood group determinants in Fasciola hepatica. J Parasitol 68: 402–407.
24. Lu B, PereiraPerrin M (2008) A novel immunoprecipitation strategy identifies a unique functional mimic of the glial cell line-derived neurotrophic factor family ligands in the pathogen Trypanosoma cruzi. Infect Immun 76: 3530–3538.
25. Inal JM, Hui KM, Miot S, Lange S, Ramirez MI, et al. (2005) Complement C2 receptor inhibitor trispanning: a novel human complement inhibitory receptor. J Immunol 174: 356–366.
26. Lehr T, Geyer H, Maass K, Doenhoff MJ, Geyer R (2007) Structural characterization of N-glycans from the freshwater snail Biomphalaria glabrata cross-reacting with Schistosoma mansoni glycoconjugates. Glycobiology 17: 82–103.
27. Holmquist G, Udomsangpetch R, Berzins K, Wigzell H, Perlmann P (1988) Plasmodium chabaudi antigen Pch105, Plasmodium falciparum antigen Pf155, and erythrocyte band 3 share cross-reactive epitopes. Infect Immun 56: 1545–1550.
28. Ponce de Leon P, Foresto P, Valverde J (2005) H antigen presence in an Ascaris lumbricoides extract. Rev Inst Med Trop Sao Paulo 47: 159–160.
29. Ramos M, Alvarez I, Sesma L, Logean A, Rognan D, et al. (2002) Molecular mimicry of an HLA-B27-derived ligand of arthritis-linked subtypes with chlamydial proteins. J Biol Chem 277: 37573–37581.
30. Hall R (1994) Molecular mimicry. Adv Parasitol 34: 81–132.
31. Halanych K (2004) The new view of animal phylogeny. Annu Rev Ecol Evol Syst 35: 229–256.
32. Gamulin V, Muller I, Muller W (2000) Sponge proteins are more similar to those of Homo sapiens than to Caenorhabditis elegans. Biol J Linn Soc 71: 821–828.
33. Seki Y, Hayashi K, Matsumoto A, Seki N, Tsukada J, et al. (2002) Expression of the suppressor of cytokine signaling-5 (SOCS5) negatively regulates IL-4-dependent STAT6 activation and Th2 differentiation. Proc Natl Acad Sci U S A 99: 13003–13008.
34. Yoshimura A, Naka T, Kubo M (2007) SOCS proteins, cytokine signalling and immune regulation. Nat Rev Immunol 7: 454–465.
35. Inal JM (2005) Complement C2 receptor inhibitor trispanning: from man to schistosome. Springer Semin Immunopathol 27: 320–331.
36. Hewitson JP, Harcus YM, Curwen RS, Dowle AA, Atmadja AK, et al. (2008) The secretome of the filarial parasite, Brugia malayi: proteomic profile of adult excretory-secretory products. Mol Biochem Parasitol 160: 8–21.
37. Bennuru S, Semnani R, Meng Z, Ribeiro JM, Veenstra TD, et al. (2009) Brugia malayi excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling. PLoS Negl Trop Dis 3: e410.
38. Goodridge HS, Stepek G, Harnett W, Harnett MM (2005) Signalling mechanisms underlying subversion of the immune response by the filarial nematode secreted product ES-62. Immunology 115: 296–304.
39. Kazerounian S, Aho S (2003) Characterization of periphilin, a widespread, highly insoluble nuclear protein and potential constituent of the keratinocyte cornified envelope. J Biol Chem 278: 36707–36717.
40. Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338: 1027–1036.
41. Schvartz I, Seger D, Shaltiel S (1999) Vitronectin. Int J Biochem Cell Biol 31: 539–544.
42. Blom AM, Hallstrom T, Riesbeck K (2009) Complement evasion strategies of pathogens-acquisition of inhibitors and beyond. Mol Immunol 46: 2808–2817.
43. Singh B, Su YC, Riesbeck K (2010) Vitronectin in bacterial pathogenesis: A host protein used in complement escape and cellular invasion. Mol Microbiol accepted article.
44. Bergmann S, Lang A, Rohde M, Agarwal V, Rennemeier C, et al. (2009) Integrin-linked kinase is required for vitronectin-mediated internalization of Streptococcus pneumoniae by host cells. J Cell Sci 122: 256–267.
45. Sherwood JA, Roberts DD, Marsh K, Harvey EB, Spitalnik SL, et al. (1987) Thrombospondin binding by parasitized erythrocyte isolates in falciparum malaria. Am J Trop Med Hyg 36: 228–233.
46. Pain A, Bohme U, Berry AE, Mungall K, Finn RD, et al. (2008) The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature 455: 799–803.
47. Muller HM, Reckmann I, Hollingdale MR, Bujard H, Robson KJ, et al. (1993) Thrombospondin related anonymous protein (TRAP) of Plasmodium falciparum binds specifically to sulfated glycoconjugates and to HepG2 hepatoma cells suggesting a role for this molecule in sporozoite invasion of hepatocytes. EMBO J 12: 2881–2889.
48. Robson KJ, Frevert U, Reckmann I, Cowan G, Beier J, et al. (1995) Thrombospondin-related adhesive protein (TRAP) of Plasmodium falciparum: expression during sporozoite ontogeny and binding to human hepatocytes. EMBO J 14: 3883–3894.
49. Robson KJ, Hall JR, Jennings MW, Harris TJ, Marsh K, et al. (1988) A highly conserved amino-acid sequence in thrombospondin, properdin and in proteins from sporozoites and blood stages of a human malaria parasite. Nature 335: 79–82.
50. Sultan AA, Thathy V, Frevert U, Robson KJ, Crisanti A, et al. (1997) TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites. Cell 90: 511–522.
51. Menard R (2000) The journey of the malaria sporozoite through its hosts: two parasite proteins lead the way. Microbes Infect 2: 633–642.
52. Matuschewski K, Nunes AC, Nussenzweig V, Menard R (2002) Plasmodium sporozoite invasion into insect and mammalian cells is directed by the same dual binding system. EMBO J 21: 1597–1606.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
54. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.
55. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23: 403–405.
56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
57. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
58. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281–288.
59. Wang S, Browning KS, Miller WA (1997) A viral sequence in the 3′-untranslated region mimics a 5′ cap in facilitating translation of uncapped mRNA. EMBO J 16: 4107–4116.
60. Kraiczy P, Wurzner R (2006) Complement escape of human pathogenic bacteria by acquisition of complement regulators. Mol Immunol 43: 31–44.
61. Diaz A, Ferreira A, Sim RB (1997) Complement evasion by Echinococcus granulosus: sequestration of host factor H in the hydatid cyst wall. J Immunol 158: 3779–3786.
62. Meri T, Jokiranta TS, Hellwage J, Bialonski A, Zipfel PF, et al. (2002) Onchocerca volvulus microfilariae avoid complement attack by direct binding of factor H. J Infect Dis 185: 1786–1793.
63. Cerami C, Frevert U, Sinnis P, Takacs B, Clavijo P, et al. (1992) The basolateral domain of the hepatocyte plasma membrane bears receptors for the circumsporozoite protein of Plasmodium falciparum sporozoites. Cell 70: 1021–1033.
64. Rubin-de-Celis SS, Uemura H, Yoshida N, Schenkman S (2006) Expression of trypomastigote trans-sialidase in metacyclic forms of Trypanosoma cruzi increases parasite escape from its parasitophorous vacuole. Cell Microbiol 8: 1888–1898.

65. Nagamune K, Acosta-Serrano A, Uemura H, Brun R, Kunz-Renggli C, et al. (2004) Surface sialic acids taken from the host allow trypanosome survival in tsetse fly vectors. J Exp Med 199: 1445–1450.
66. Mauss EA (1941) Occurrence of Forssman heterogenic antigen in the nematode, Trichinella spiralis. J Immunol 42: 71–77.
67. Shear HL, Nussenzweig RS, Bianco C (1979) Immune phagocytosis in murine malaria. J Exp Med 149: 1288–1298.
68. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, et al. (2007) Draft genome of the filarial nematode parasite Brugia malayi. Science 317: 1756–1760.
69. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, et al. (2004) Complete genome sequence of the apicomplexan, Cryptosporidium parvum. Science 304: 441–445.
70. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, et al. (2007) Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. Science 317: 1921–1926.
71. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, Leishmania major. Science 309: 436–442.
72. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498–511.
73. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, et al. (2009) The genome of the blood fluke Schistosoma mansoni. Nature 460: 352–358.
74. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, et al. (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. Science 315: 207–212.
75. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, et al. (2005) The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science 309: 409–415.
76. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
77. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, et al. (2007) Genome sequence of Aedes aegypti, a major arbovirus vector. Science 316: 1718–1723.
78. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129–149.
79. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.
80. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012–2018.
81. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, et al. (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science 298: 2157–2167.
82. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, et al. (2002) The genome sequence of Schizosaccharomyces pombe. Nature 415: 871–880.
83. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, et al. (2008) The Trichoplax genome and the nature of placozoans. Nature 454: 955–960.

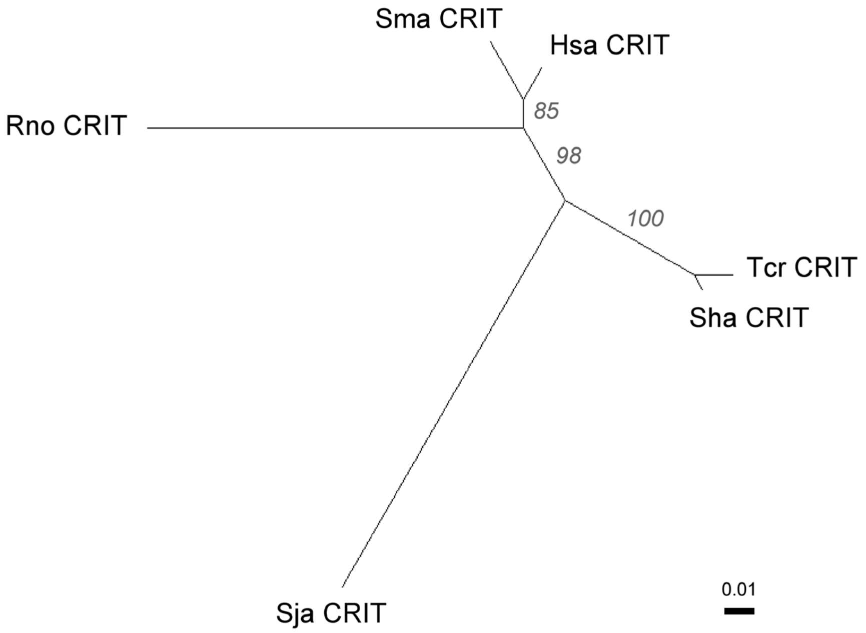
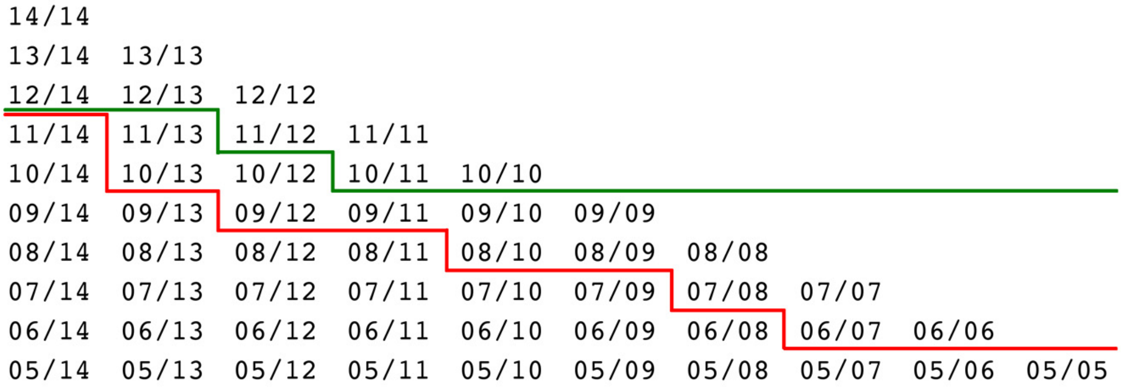


**Figure S1.** ClustalW dendrogram of CRIT orthologues from *Schistosoma mansoni* (Sma), *S. haematobium* (Sha), *S. japonicum* (Sja), *Trypanosoma cruzi* (Tcr), *H. sapiens* (Hsa), and *R. norvegicus* (Rno). The scale bar indicates changes per site. Bootstrapping numbers (grey) are given as percent positives of 1,000 rounds.



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14/14 | | | | | | | | | |
| 13/14 | 13/13 | | | | | | | | |
| 12/14 | 12/13 | 12/12 | | | | | | | |
| 11/14 | 11/13 | 11/12 | 11/11 | | | | | | |
| 10/14 | 10/13 | 10/12 | 10/11 | 10/10 | | | | | |
| 09/14 | 09/13 | 09/12 | 09/11 | 09/10 | 09/09 | | | | |
| 08/14 | 08/13 | 08/12 | 08/11 | 08/10 | 08/09 | 08/08 | | | |
| 07/14 | 07/13 | 07/12 | 07/11 | 07/10 | 07/09 | 07/08 | 07/07 | | |
| 06/14 | 06/13 | 06/12 | 06/11 | 06/10 | 06/09 | 06/08 | 06/07 | 06/06 | |
| 05/14 | 05/13 | 05/12 | 05/11 | 05/10 | 05/09 | 05/08 | 05/07 | 05/06 | 05/05 |

**Figure S2.** The filtering system used in the overlapping fragments approach. Numbers represent identical amino acid residues. Red line: threshold for negative control species. Green line: threshold for molecular mimicry candidate in mammalian host or insect vector.

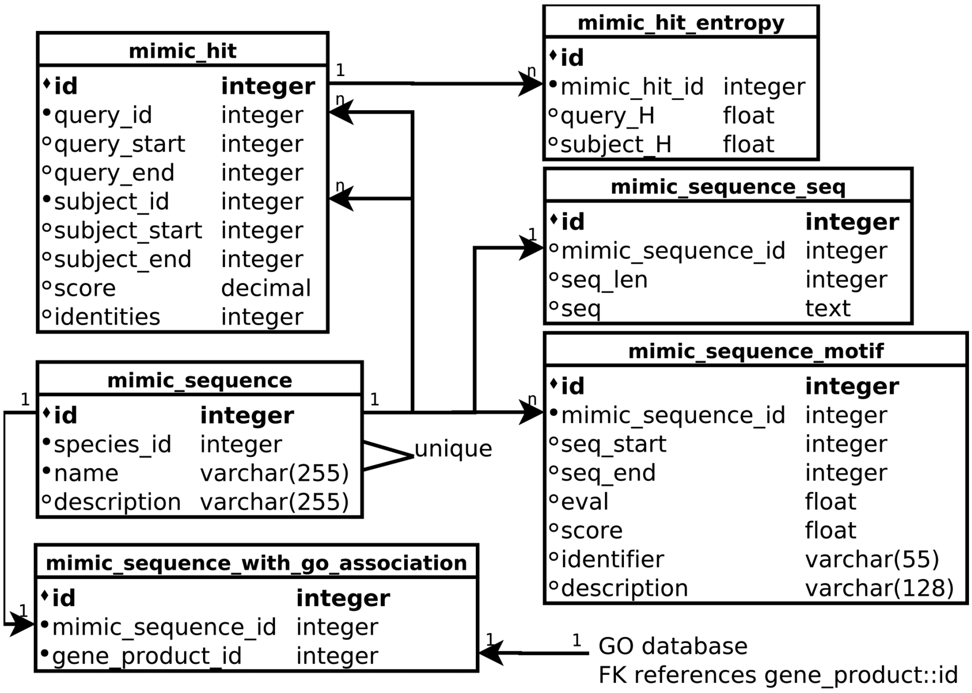


**Figure S3.** Database schema of mimicDB. The mimicDB database schema centers around *mimic_sequence*, which represents the individual genes. This table has as attribute tables the actual peptide sequences (*mimic_sequence_seq*) and predicted motifs (*mimic_sequence_motif*). Hits between parts of these genes are collected in *mimic_hit*, which stores the coordinates and properties of the hit. A complexity measure, in the form of Shannon source entropy for each peptide hit is stored in *mimic_hit_entropy*. The database connects to the GO consortium GO term database in that *mimic_sequence* entries that have a GO association are referenced by entries in *mimic_sequence_with_go_association*, where the corresponding GO term db gene_product::id is also a foreign key.

# *Chapter 3*

## Species-specific Typing of DNA Based on Palindrome Frequency Patterns

Estelle Lamprea-Burgunder[1,†], Philipp Ludin[1,2,3,†], Pascal Mäser[1,2,3]

[1] Institute of Cell Biology, University of Bern, Bern, Switzerland

[2] Swiss Tropical and Public Health Institute, Basel, Switzerland

[3] University of Basel, Basel, Switzerland

[†] These authors contributed equally to this work

*DNA Research* (2011)

DNA Research **18**, 117–124, (2011) doi:10.1093/dnares/dsr004
Advance Access Publication: 23 March 2011

# Species-specific Typing of DNA Based on Palindrome Frequency Patterns

Estelle Lamprea-Burgunder[1,†,‡], Philipp Ludin[1,2,3,†], and Pascal Mäser[1,2,3,*]

*Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland[1]; Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland[2] and University of Basel, 4000 Basel, Switzerland[3]*

*To whom correspondence should be addressed. Tel. +41-61-284-8338. Fax. +41-61-284-8101. E-mail: pascal.maeser@unibas.ch

Edited by Hiroyuki Toh
(Received 20 October 2010; accepted 24 February 2011)

**Abstract**

**DNA in its natural, double-stranded form may contain palindromes, sequences which read the same from either side because they are identical to their reverse complement on the sister strand. Short palindromes are underrepresented in all kinds of genomes. The frequency distribution of short palindromes exhibits more than twice the inter-species variance of non-palindromic sequences, which renders palindromes optimally suited for the typing of DNA. Here, we show that based on palindrome frequency, DNA sequences can be discriminated to the level of species of origin. By plotting the ratios of actual occurrence to expectancy, we generate palindrome frequency patterns that allow to cluster different sequences of the same genome and to assign plasmids, and in some cases even viruses to their respective host genomes. This finding will be of use in the growing field of metagenomics.**

**Key words:** comparative genomics; DNA palindrome; hierarchical clustering

## 1. Introduction

The double helix forms the structural basis of semi-conservative DNA replication.[1,2] Less intuitively, it also has implications on the information content of DNA for double-stranded DNA as such only has about half the storage capacity of single-stranded DNA. This is because a given sequence and its reverse complement, while the same in the double-stranded form, are different entities in single-stranded DNA—except for those sequences which are identical to their reverse complement. Centrally symmetric when double-stranded, such sequences read the same from either 5′ end and are called DNA palindromes (e.g. 5′-GATC-3′). Consider the $4^4 = 256$ different single-stranded sequences of length 4; only the $4^2 = 16$ possible palindromes are unique in the double-stranded form. The remaining 240 form 120 pairs of identical sequences when complemented to the double strands (e.g. 5′-GACT-3′ and 5′-AGTC-3′). Thus, the total number of different double-stranded sequences of length 4 bp is $120 + 16 = 136$. It follows that the generally accepted maximal information content of 2 bit per base pair only holds true if the two sister strands can be distinguished (which requires extra information).

Given that palindromes are the only sequences which are unique in double-stranded DNA, it is not surprising that they are of particular importance in genome biology. Dimeric restriction endonucleases and DNA methyltransferases bind palindromic recognition sites.[3,4] The same applies to transcription factors such as the bacterial trp repressor[5] or the mammalian oestrogen receptor.[6] Palindromes also fulfil an important role as spacers in the prokaryotic CRISPR/Cas system (clustered regularly interspaced short palindromic repeats), which forms the basis of immune memory against bacteriophages and

[†] These authors contributed equally to this work.
[‡] Present address: Roche Diagnostics, 6343 Rotkreuz, Switzerland.

plasmids.[7] Viral and bacterial genomes possess palindromic replication origins.[8,9] Palindromes also contribute to genome instability: as target sites for insertion sequence elements[10] and for homologous strand invasion during recombination,[11] and by inducing double-strand breaks due to hairpin-specific nucleases.[12]

While statistically, palindromes are expected to occur half as often as non-palindromic sequences in double-stranded DNA, they are even rarer than this in natural DNA sequences. Short palindromes were found to be underrepresented in various genomes including bacteriophages,[13] bacteria,[14–16] and fungi.[17,18] In the human genome, palindromes were found to be underrepresented in exons but overrepresented in introns and in upstream regions of genes.[19] In bacteria, restriction endonucleases which cleave palindromic DNA recognition sites were proposed as a selective force against palindromes.[20] In vertebrate genomes, the underrepresentation of palindromes is partly attributable to the drift of CG dinucleotides to TG by deamination of methylated cytosine.[21] Other factors accounting for the scarcity of palindromes in genomic DNA sequences are the potential adverse effects of palindromes on chromatin structure,[17] bias of the mismatch repair system,[22] and selection against palindromes to avoid inappropriate binding of transcription factors.[17] Here, we make use of the large number of available sequenced genomes, scanning them for the occurrence of short palindromes and demonstrating that (i) the underrepresentation of short palindromes is ubiquitous and (ii) the frequency distribution of short palindromes lends itself for species-specific typing of DNA sequences.

**Table 1.** The 16 palindromes of length 4 and an equal number of non-palindromes, their mean ratio *R* of occurrence to expectancy, and variance of *R*, across 200 genomes

| Palindrome | *R* | Var(*R*) | Non-palindrome | *R* | Var(*R*) |
|---|---|---|---|---|---|
| AATT | 0.97 | 0.06 | AAGG/CCTT | **1.23** | 0.09 |
| ATAT | **0.85** | 0.05 | ACAG/CTGT | 1.04 | 0.08 |
| TATA | **0.68** | 0.07 | ACTG/CAGT | 0.94 | 0.06 |
| TTAA | **0.84** | 0.10 | AGAC/GTCT | **0.87** | 0.05 |
| ACGT | **0.63** | 0.12 | CAAC/GTTG | **1.10** | 0.06 |
| AGCT | 1.13 | 0.13 | CAGA/TCTG | **1.16** | 0.12 |
| CATG | 0.99 | 0.09 | CCAA/TTGG | **1.24** | 0.12 |
| CTAG | **0.67** | 0.14 | CTGA/TCAG | 1.06 | 0.08 |
| GATC | **0.86** | 0.16 | GAGA/TCTC | **1.14** | 0.09 |
| GTAC | **0.71** | 0.05 | GGAA/TTCC | **1.30** | 0.11 |
| TCGA | **0.84** | 0.39 | GTCA/TGAC | **0.87** | 0.04 |
| TGCA | **1.15** | 0.15 | GTGA/TCAC | 0.94 | 0.04 |
| CCGG | 0.89 | 0.24 | TCCT/AGGA | **1.29** | 0.12 |
| CGCG | **0.62** | 0.27 | TGAG/CTCA | 1.07 | 0.07 |
| GCGC | **0.80** | 0.28 | TGGT/ACCA | **1.13** | 0.08 |
| GGCC | **1.15** | 0.32 | TGTC/GACA | 0.92 | 0.04 |
| Overall | **0.86** | 0.19 | Overall | **1.08** | 0.10 |

Values significantly deviating from 1 are given in bold ($P <$ 0.01, one-way ANOVA followed by Dunnett's multiple comparison test).

## 2. Materials and methods

### 2.1. *Genome sequences*

Genomic DNA sequences were analysed of 200 species from 10 different phylogenetic groups, 20 species per group: vertebrates, invertebrates, fungi, plants, protozoa, bacteria, archaea, mitochondria, dsDNA viruses, and retroviruses. Complete genomes or chromosomes were analysed if available, otherwise large contigs of at least 100 kb. Twenty was the maximal number of available genome sequences for groups like invertebrates or plants; to be able to compare the variances between the different groups, the number of species per group was therefore fixed to 20 (randomly selected). See Supplementary Table S1 for a complete list of species including accession numbers.

### 2.2. *Calculation of palindrome expectancy*

The number $N$ of different DNA palindromes of length $l$ is given by:

$$N(l) = 4^{l/2} \tag{1}$$

since palindromes are centrally symmetric. The expectancy ($E$) of a palindrome (pal) of length $l_{\mathrm{pal}}$ and GC ratio $\mathrm{gc}_{\mathrm{pal}}$ in a DNA sequence (seq) of length $l_{\mathrm{seq}}$ and GC ratio $\mathrm{gc}_{\mathrm{seq}}$ is:

$$E(\mathrm{pal}, \mathrm{seq}) = \left(\frac{\mathrm{gc}_{\mathrm{seq}}}{2}\right)^{\mathrm{gc}_{\mathrm{pal}} \times l_{\mathrm{pal}}} \times \left(\frac{1 - \mathrm{gc}_{\mathrm{seq}}}{2}\right)^{(1-\mathrm{gc}_{\mathrm{pal}}) \times l_{\mathrm{pal}}} \times l_{\mathrm{seq}} \tag{2}$$

The ratio $R$ of palindrome frequency was defined as:

$$R(\mathrm{pal}, \mathrm{seq}) = \frac{n}{E(\mathrm{pal}, \mathrm{seq})} \tag{3}$$

where $n$ is the actual occurrence of the palindrome (pal) in the sequence (seq).

### 2.3. *Counting of palindromes*

The counting of palindromes in DNA sequences (Fasta format) was performed with a Perl script, available on request under the GNU public licence. Input

DNA sequences were first rid of all perfect repeats (word size four or longer, repeated in tandem for at least five times) to avoid a possible bias from telomeric or centromeric repeats. Then, the occurrence was counted of each of the 16 different palindromes of length 4 (Table 1). To allow for comparison of variance, the same number of control sequences were included that did not contain any palindromic duplets nor compatible ends (Table 1). Each of these controls was counted alongside its reverse complement to render the result independent of the DNA strand searched. The $\log_2$ of the ratios $R$ of occurrence to expectancy for each 4-mer palindrome was plotted as vectors of 16 components, which were clustered by average linkage based on the city-block distance (i.e. the sum of absolute differences in the components of a given pair of vectors). Clustering was performed with the programs Cluster and TreeView[23] from the Eisen lab (http://rana.lbl.gov/eisen/).

### 2.4. *Random controls*

Random sequences of variable length were generated with the program *makenucseq* of the EMBOSS package.[24] Randomly selected, non-overlapping 10 kb fragments of bacterial genomes were generated with a self-made Perl script using *srand*(time) as the random number seed.

## 3. Results and discussion

### 3.1. *Palindrome occurrence across the tree of life*

We counted the occurrence of the 16 palindromic words of length 4 (Table 1), along with an equal number of non-palindromic words of length 4 (Table 1), in DNA sequences of selected genomes. Twenty different species were analysed for each of 10 different phylogenetic groups, i.e. the vertebrates, invertebrates, fungi, plants, protozoa, mitochondria, bacteria, archaea, double-stranded DNA viruses, and retroviruses. Perfect repeats were removed from the input sequences to avoid introduction of a trivial bias from regions of extremely low complexity such as telomeric or centromeric repeats. For each input DNA sequence and each 4-mer word, we then calculated the ratio $R$ of actual occurrence of the word divided by the expected number of occurrences, given its GC content and that of the input DNA sequence. Most of the palindromes were underrepresented ($R < 1$) across all genomes analysed. Overall, the palindromes exhibited a mean $R$ of 0.86, in contrast to a mean $R$ of 1.08 for the non-palindromic controls (Table 1). The underrepresentation of palindromes was most pronounced in vertebrate genomes, plants, double-stranded DNA viruses, and retroviruses (Fig. 1). Contrary to previous reports,[20] palindromes were underrepresented even in mitochondrial genomes, demonstrating that the infrequence of palindromes in prokaryote genomes cannot solely be explained by the selective pressure exerted by restriction enzymes. Additional selective forces against palindromes might comprise their impact on DNA structure or their role as transcription factor-binding sites.[17] Whatever the underlying forces, short palindromes are underrepresented in all kinds of genomes (Fig. 1). Exactly which palindromes and how strongly depend on the source of the DNA. Interestingly, the inter-genome frequencies of short palindromes exhibit more than twice the variance of the non-palindromic control sequences (22 versus 9%; Table 1), whereas the intra-genome frequencies, e.g. between different chromosomes of the same organism, are uniform (Figs 2–4). This renders short palindromes optimally suited for the typing of DNA.

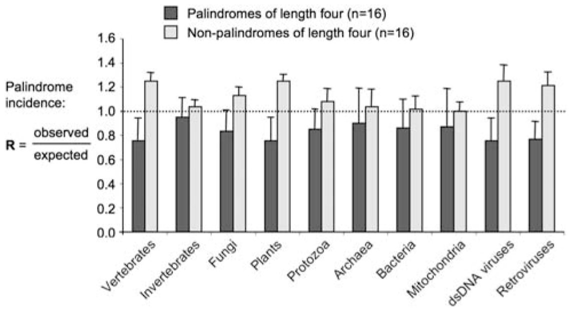


**Figure 1.** Frequency of palindromes throughout a diverse selection of genomes. Palindrome frequency is expressed as the ratio ($R$) of occurrence to expectancy. Palindromes are underrepresented ($R < 1$, dotted line) in all kinds of genomes, most strongly in vertebrates, plants, and viruses, and they exhibit about twice the inter-species variance in frequency (error bars) than non-palindromes. Twenty different genomes were analysed per group (see Section 2).

### 3.2. *Clustering of DNA based on palindrome frequency*

Here, we represent a given DNA sequence by a vector of 16 numbers: for each of the 16 palindromes of length 4, the $\log_2$ of the ratio $R$ of actual to expected frequency (given the GC content of the analysed DNA and that of the palindrome). When such vectors, generated from a diverse selection of DNA sequences, were aligned and hierarchically clustered based on the city-block distance, different DNA sequences of the same species readily grouped together (see Fig. 2 for a representative set of diverse genomes). The clustering worked for all kinds of genome sequences tested—eukaryote, prokaryote, plastid, or virus—but the topology of the resulting tree was not phylogenetically meaningful (Fig. 2). The lack of a large-scale phylogenetic signal

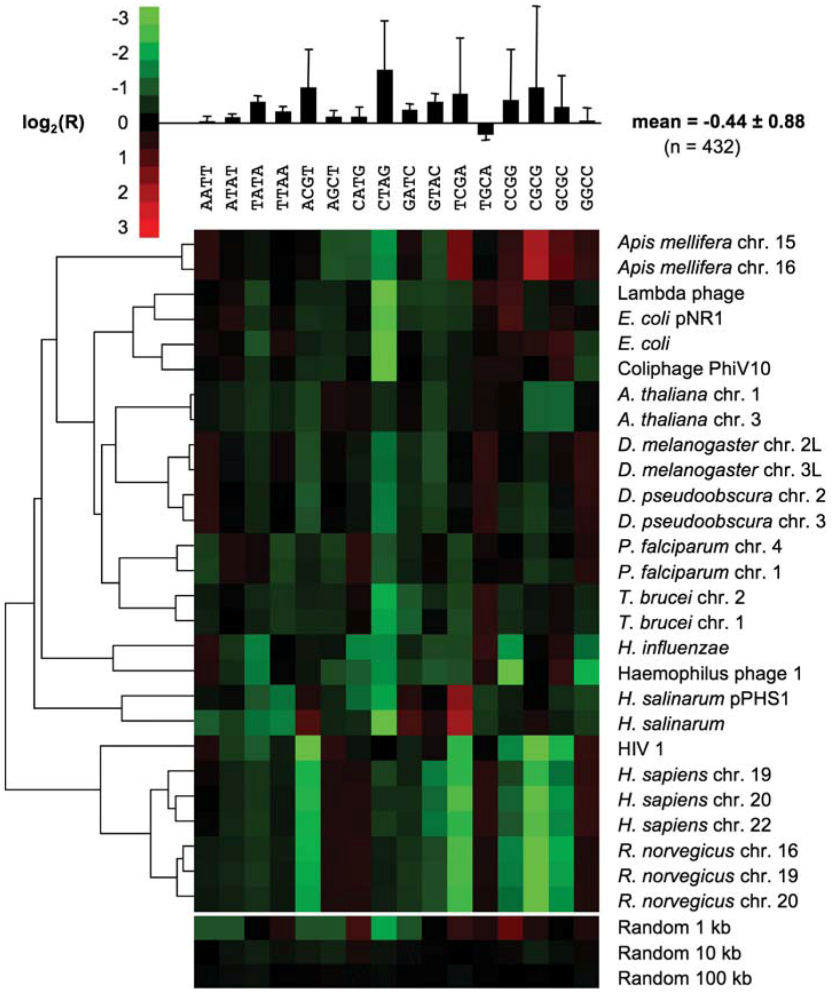


**Figure 2.** Examples of palindrome frequency patterns. Frequency of the 16 palindromes of length 4 in selected genomes, expressed as $\log_2$ of ratio ($R$) of actual to expected occurrence. Hierarchical clustering was performed based on the city-block distance.[23] (Top) Mean and variance by palindrome. (Bottom) The signals from three random sequences are shown for comparison.

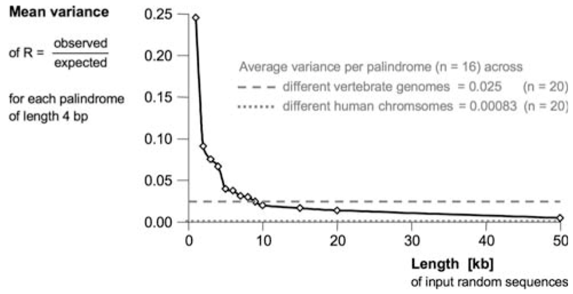


**Figure 3.** Variance of palindrome frequencies in random DNA sequence of different lengths ($n = 20$ for each). The mean variance for each palindrome of length 4 across the 20 different sequences is compared with those across the first 20 human chromosomes (dotted grey line) and across the 20 different vertebrate chromosomes analysed in Fig. 1 (see Supplementary Table S1).

was equally apparent from the analysis of the complete set of 200 genomes (Supplementary Fig. S1). The resolution of palindrome frequency clustering would increase further by using the 64 different palindromes of length 6. However, this would also require the input sequences to be longer. On the basis of the random sequences included in Fig. 2, the present approach appeared to work for sequences longer than about 10 kb. To obtain a better estimate on the minimally required size of input DNA, we analysed randomly generated sequences of increasing length (Fig. 3). Above 9 kb, the average variance of $R$ per palindrome dropped below the value obtained for different vertebrate chromosomes (0.025, dashed grey line in Fig. 3). For comparison, the average variance of $R$ per palindrome across human

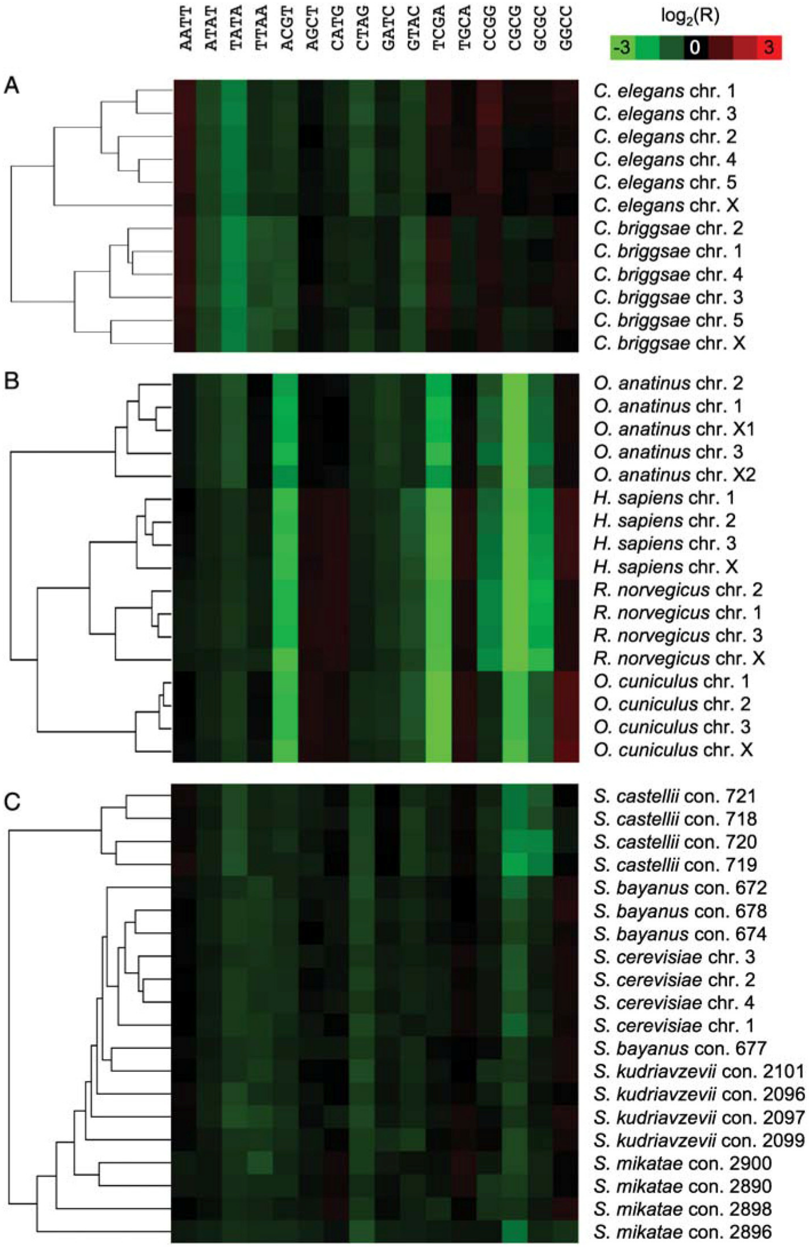


**Figure 4.** Case studies on *Caenorhabditis* spp. (A), mammalian chromosomes (B), and *sensu stricto* yeasts (C). Most of the chromosomes are correctly resolved by clustering based on palindrome frequency. Perfect tandem repeats were removed prior to analysis to avoid trivial differences from repetitive regions. Note the striking difference between vertebrate and invertebrate DNA.

chromosomes was 0.0008 (dotted grey line in Fig. 3), demonstrating again that the variance of palindrome frequency is much lower intra- than inter-genome.

Invertebrates exhibiting the smallest inter-genome variance of palindrome frequency (Fig. 1), we chose *Caenorhabditis* species to challenge its power of discrimination. The complete nuclear genomes of *C. briggsae* and *C. elegans* were compared as described above and all the chromosomes were correctly resolved in spite of the weak patterns (Fig. 4A). Clustering based on palindrome frequency also segregated different mammalian chromosomes which, in contrast to invertebrate DNA, showed the characteristic pattern caused by strong

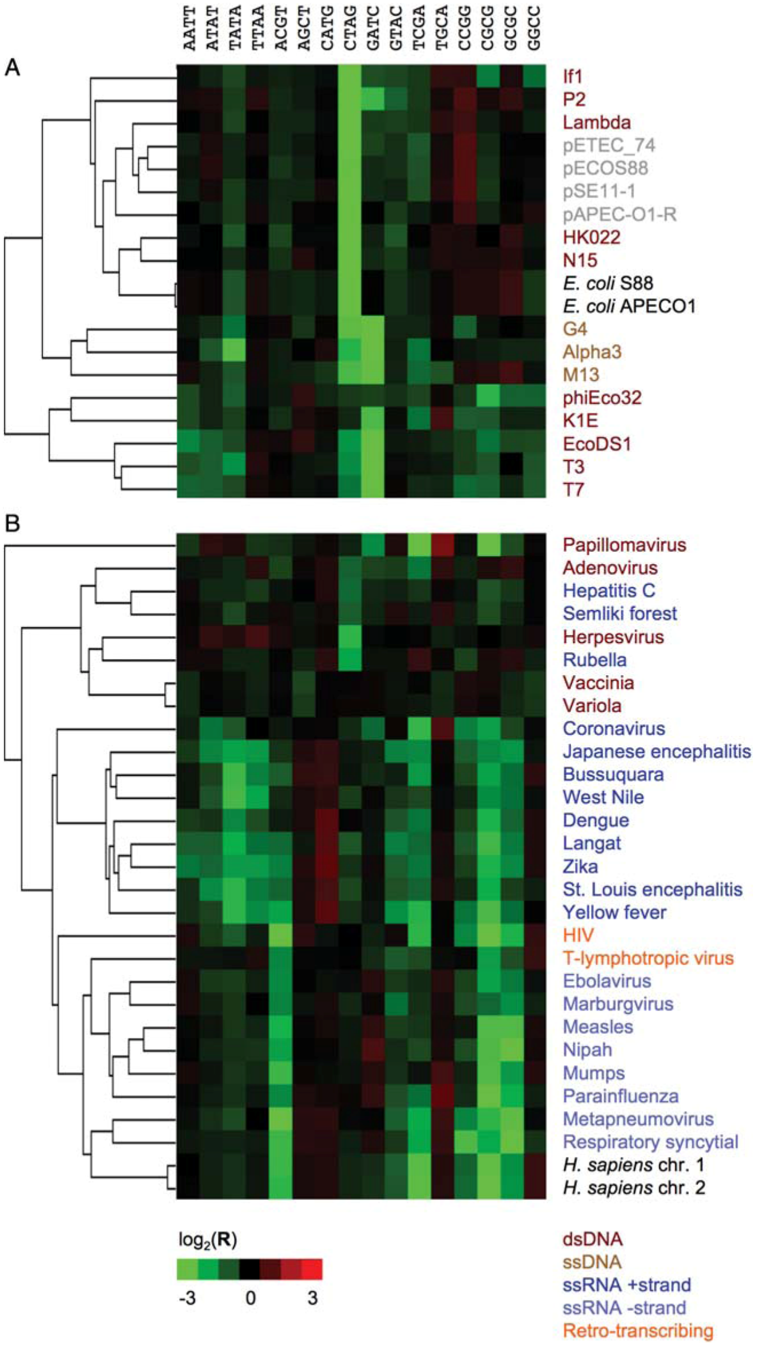


**Figure 5.** Palindrome frequency patterns of host genomic DNA (A, *E. coli*; B, *Homo sapiens*; labelled in black) and associated viruses (colour-coded according to nucleic acid type of the genome) or plasmids (grey).

underrepresentation of palindromes containing a CG dinucleotide (ACGT, TCGA, CCGG, GCGC, and CGCG; Fig. 4B). This is in agreement with the model that in vertebrates, DNA methylation is restricted to cytosines followed by guanine (CpG), whereas in invertebrates, cytosines are methylated in a wider context.[25] Spontaneous mutation of the palindromic CG to the non-palindromic TG by deamination of methylated cytosine thus eliminates short palindromes from vertebrate DNA. The limit of resolution of palindrome frequency clustering was reached with a data set of highly similar *sensu stricto* yeasts.[26] The different chromosomes of the closely related species *Saccharomyces cerevisiae*, *S. bayanus*, *S. mikatae*, and *S. kudriavzevii* did not segregate perfectly; those of the more distantly related *S. castellii* did (Fig. 4C).

Clustering based on palindrome frequency also worked for prokaryotes, generating species-specific patterns for archaea as well as bacteria. Prokaryote genomes exhibited highly diverse patterns (Supplementary Fig. S1). Natural plasmids of *Escherichia coli* clearly clustered with the host DNA (Fig. 5A). The same applied to certain dsDNA bacteriophages such as Lambda or P2. However, other dsDNA phages such as T3, as well as all analysed ssDNA phages, did not exhibit the same palindrome frequency patterns as *E. coli* (Fig. 5A). An interesting picture emerged when comparing human viruses: while all ssRNA minus-strand viruses and the retro-transcribing HIV clustered with human DNA, dsDNA viruses and ssRNA plus-strand viruses did not (Fig. 5B).

### 3.3. *Potential application to metagenomics*

The quickly developing field of environmental shotgun sequencing allows metagenomic analyses of communities of microorganisms, the majority of which cannot be cultured in the lab and have therefore remained undetected until recently.[27] A key challenge in interpreting environmental shotgun sequencing data is the binning of non-overlapping DNA scaffolds into groups which, ideally, correspond to the different species of microorganisms present.[28] Standard methods such as similarity searches to known genomes or phylogenetic analysis of marker genes are of limited use when dealing with DNA fragments sampled from previously undescribed species.[28] Di-, tri- and tetra-nucleotide frequencies have been proposed to provide DNA signatures.[29–31] Palindrome frequencies carrying a species-specific signal (Figs 2 and 4), the ratios of occurrence to expectancy as applied here may also be useful to bin environmental shotgun sequencing data, provided that the contigs to be analysed are longer than 9 kb (Fig. 3). From the 2007 *Sorcerer II* Global Ocean Sampling Expedition, which at that time predominantly produced novel sequences,[32] the hundred largest contigs, sized between 11 and 59 kb, were analysed as described above. This revealed a diverse picture of palindrome frequency patterns with several major clusters (Supplementary Fig. S2). However, the analysed sequences still did not return high-quality hits when searched with *blastn*[33] against the NCBI non-redundant nucleotide collection, with only one exception of 99% identity to *Prochlorococcus* phage P-SSM4 (GenBank accession no. AY940168). Thus, it was not possible to assess the benefit of palindrome frequency clustering with this data set. To nevertheless test the potential of the method, we randomly selected 10 non-overlapping fragments of length 10 kb from each of the 20 different bacterial genomes analysed in Fig. 1 (Supplementary Table S1). When these 200 sequences were clustered according to palindrome frequency patterns, over 90% of them correctly assembled according to species of origin.

## 4. Conclusion

Accustomed to reading DNA as linear sequences, we tend to forget that it is double-stranded in nature. In double-stranded DNA, the only sequences which are unique are palindromes. Here, we confirm the notion that short palindromes are underrepresented across all different kinds of genomes. The frequency distribution of short palindromes exhibits highest inter-species but low intra-species variance. We take advantage of this to type DNA based on palindrome frequency, generating highly specific patterns which discriminate the DNA from different species, clustering together sequences from the same species. The method allows for the assignment of plasmids and certain viruses to their respective host genomes. Although the underlying selective forces are not fully understood, these patterns are highly useful for analysis of DNA sequences of unknown origin, such as those generated by the gigabase in metagenomic high-throughput sequencing surveys. Palindrome frequency ratios as presented here could be incorporated into more sophisticated classifiers such as self-organizing maps,[34] Bayesian classifiers,[35] or support vector machines.[36] Concentrating on palindromes may help to estimate the diversity of microbial communities and to bin different, non-overlapping sequences originating from the same genome, to classify sequences by comparison to reference patterns, and to assign plasmids and bacteriophages to their respective host genomes.

**Supplementary data:** Supplementary data are available at www.dnaresearch.oxfordjournals.org.

## Funding

We wish to thank the Swiss National Science Foundation for financial support.

## References

1. Watson, J. and Crick, F. 1953, Genetical implications of the structure of deoxyribonucleic acid, *Nature*, **171**, 964–7.

2. Chagin, V.O., Stear, J.H. and Cardoso, M.C. 2010, Organization of DNA replication, *Cold Spring Harb. Perspect. Biol.*, **2**, a000737.
3. Pingoud, A. and Jeltsch, A. 2001, Structure and function of type II restriction endonucleases, *Nucleic Acids Res.*, **29**, 3705–27.
4. Zinoviev, V.V., Yakishchik, S.I., Evdokimov, A.A., Malygin, E.G. and Hattman, S. 2004, Symmetry elements in DNA structure important for recognition/methylation by DNA [amino]-methyltransferases, *Nucleic Acids Res.*, **32**, 3930–4.
5. Czernik, P.J., Shin, D.S. and Hurlburt, B.K. 1994, Functional selection and characterization of DNA binding sites for trp repressor of *Escherichia coli*, *J. Biol. Chem.*, **269**, 27869–75.
6. Welboren, W.J., Stunnenberg, H.G., Sweep, F.C. and Span, P.N. 2007, Identifying estrogen receptor target genes, *Mol. Oncol.*, **1**, 138–43.
7. Horvath, P. and Barrangou, R. 2010, CRISPR/Cas, the immune system of bacteria and archaea, *Science*, **327**, 167–70.
8. Leung, M.Y., Choi, K.P., Xia, A. and Chen, L.H. 2005, Nonrandom clusters of palindromes in herpesvirus genomes, *J. Comput. Biol.*, **12**, 331–54.
9. Zakrzewska-Czerwinska, J., Jakimowicz, D., Zawilak-Pawlik, A. and Messer, W. 2007, Regulation of the initiation of chromosomal replication in bacteria, *FEMS Microbiol. Rev.*, **31**, 378–87.
10. Schoner, B. and Kahn, M. 1981, The nucleotide sequence of IS5 from *Escherichia coli*, *Gene*, **14**, 165–74.
11. Zhou, Z.H., Akgun, E. and Jasin, M. 2001, Repeat expansion by homologous recombination in the mouse germ line at palindromic sequences, *Proc. Natl Acad. Sci. USA*, **98**, 8326–33.
12. Nasar, F., Jankowski, C. and Nag, D.K. 2000, Long palindromic sequences induce double-strand breaks during meiosis in yeast, *Mol. Cell. Biol.*, **20**, 3449–58.
13. Duggleby, R.G. 1981, A paucity of palindromes in phi X174, *J. Theor. Biol.*, **93**, 143–55.
14. Elhai, J. 2001, Determination of bias in the relative abundance of oligonucleotides in DNA sequences, *J. Comput. Biol.*, **8**, 151–75.
15. Karlin, S., Mrazek, J. and Campbell, A.M. 1997, Compositional biases of bacterial genomes and evolutionary implications, *J. Bacteriol.*, **179**, 3899–913.
16. Fuglsang, A. 2003, Distribution of potential type II restriction sites (palindromes) in prokaryotes, *Biochem. Biophys. Res. Commun.*, **310**, 280–5.
17. Burge, C., Campbell, A.M. and Karlin, S. 1992, Over- and under-representation of short oligonucleotides in DNA sequences, *Proc. Natl Acad. Sci. USA*, **89**, 1358–62.
18. Lisnic, B., Svetec, I.K., Saric, H., Nikolic, I. and Zgaga, Z. 2005, Palindrome content of the yeast *Saccharomyces cerevisiae* genome, *Curr. Genet.*, **47**, 289–97.
19. Lu, L., Jia, H., Droge, P. and Li, J. 2007, The human genome-wide distribution of DNA palindromes, *Funct. Integr. Genomics*, **7**, 221–7.
20. Gelfand, M. and Koonin, E.V. 1997, Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes, *Nucleic Acids Res.*, **25**, 2430–9.
21. Bird, A. 1986, CpG-rich islands and the function of DNA methylation, *Nature*, **321**, 209–13.
22. Bhagwat, A.S. and McClelland, M. 1992, DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome, *Nucleic Acids Res.*, **20**, 1663–8.
23. Eisen, M., Spellman, P., Brown, P. and Botstein, D. 1998, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl Acad. Sci. USA*, **95**, 14863–8.
24. Rice, P., Longden, I. and Bleasby, A. 2000, EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet.*, **16**, 276–7.
25. Mandrioli, M. 2007, A new synthesis in epigenetics: towards a unified function of DNA methylation from invertebrates to vertebrates, *Cell. Mol. Life Sci.*, **64**, 2522–4.
26. Cliften, P., Sudarsanam, P., Desikan, A., et al.. 2003, Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting, *Science*, **301**, 71–6.
27. Tringe, S.G. and Rubin, E.M. 2005, Metagenomics: DNA sequencing of environmental samples, *Nat. Rev. Genet.*, **6**, 805–14.
28. Eisen, J.A. 2007, Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes, *PLoS Biol.*, **5**, e82.
29. Karlin, S. and Ladunga, I. 1994, Comparisons of eukaryotic genomic sequences, *Proc. Natl Acad. Sci. USA*, **91**, 12832–6.
30. Karlin, S., Ladunga, I. and Blaisdell, B.E. 1994, Heterogeneity of genomes: measures and values, *Proc. Natl Acad. Sci. USA*, **91**, 12837–41.
31. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. and Glockner, F.O. 2004, Application of tetranucleotide frequencies for the assignment of genomic fragments, *Environ. Microbiol.*, **6**, 938–47.
32. Rusch, D.B., Halpern, A.L., Sutton, G., et al. 2007, The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific, *PLoS Biol.*, **5**, e77.
33. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
34. Dick, G.J., Andersson, A.F., Baker, B.J., et al., 2009, Community-wide analysis of microbial genome sequence signatures, *Genome Biol.*, **10**, R85.
35. Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I. and Coster, J. 2001, Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier, *Genome Res.*, **11**, 1404–9.
36. McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P. and Rigoutsos, I. 2007, Accurate phylogenetic classification of variable-length DNA fragments, *Nat. Methods*, **4**, 63–72.

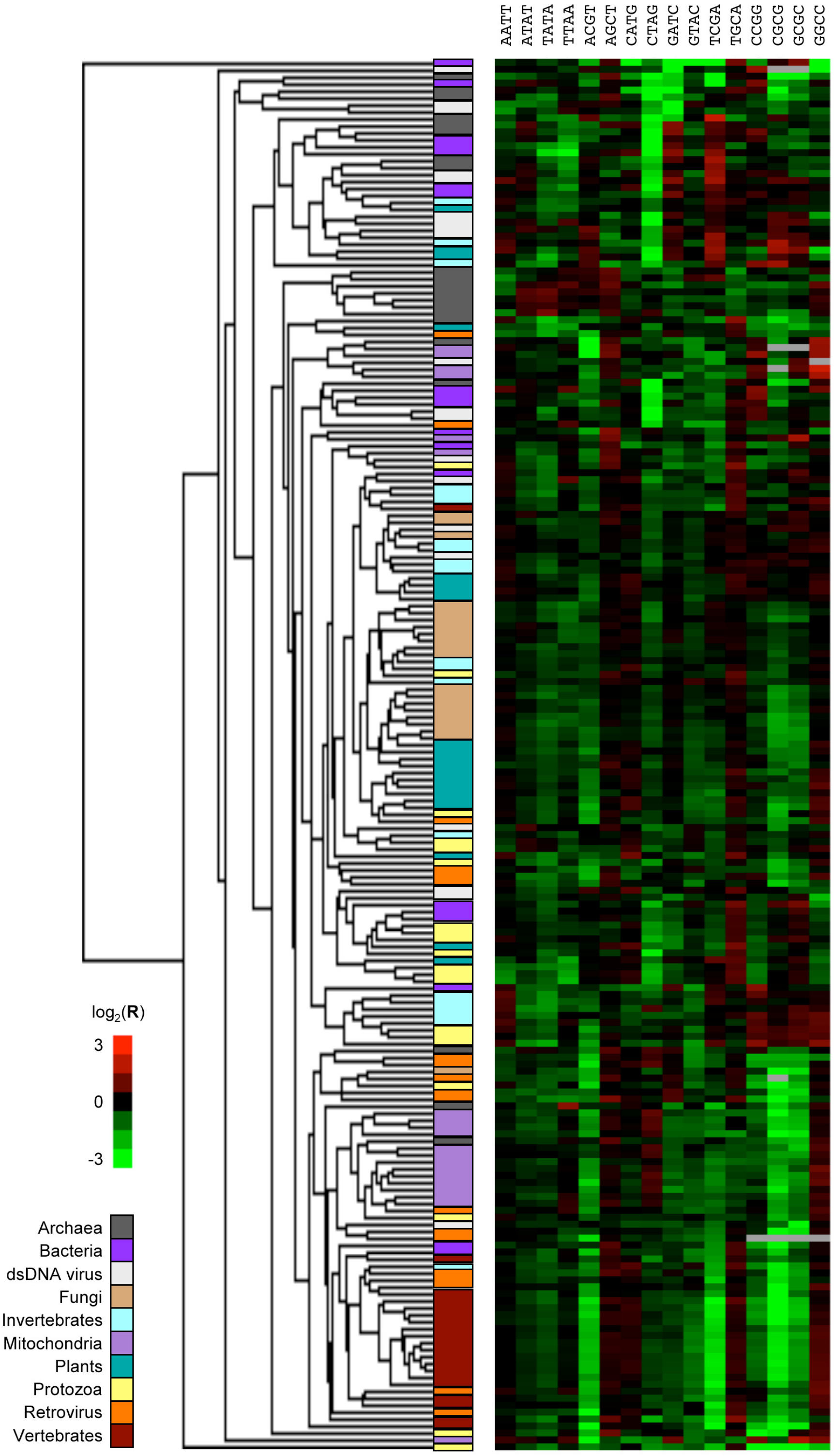


**Figure S1.** Hierarchical clustering of different genomes, 20 each of different phylogenetic groups (color codes), according to palindrome frequencies. R denotes the ratio of occurrence to expectancy of each of the 16 different palindromes of length four. Undefined R (grey) means that a given palindrome did not occur at all. Note that while there are local clusters of related genomes, the palindrome frequencies do not carry a large scale phylogenetic signal.

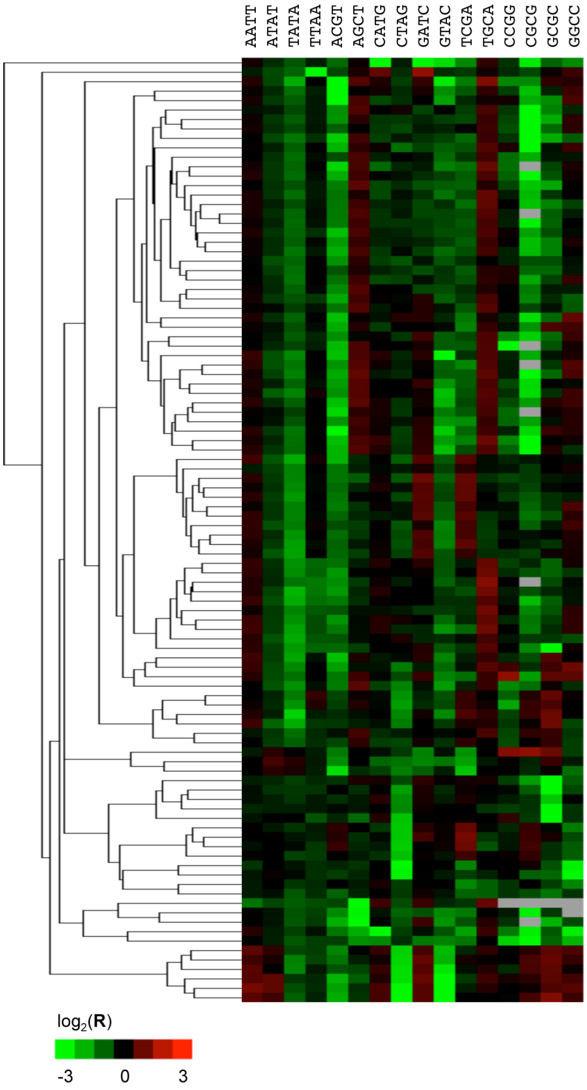


**Figure S2.** Hierarchical clustering of the 100 largest contigs of the *Sorcerer II* Global Ocean Sampling Expedition according to palindrome frequencies. R denotes the ratio of occurrence to expectancy of each of the 16 different palindromes of length four. Undefined R (grey) means that a given palindrome did not occur at all.

# *Chapter 4*

## *In Silico* Prediction of Antimalarial Drug Target Candidates

Philipp Ludin[a,b], Ben Woodcroft[c], Stuart A. Ralph[c], Pascal Mäser[a,b]

[a] Swiss Tropical and Public Health Institute, Basel, Switzerland

[b] University of Basel, Basel, Switzerland

[c] Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Victoria 3010, Australia

*International Journal for Parasitology: Drugs and Drug Resistance* (2012)

International Journal for Parasitology: Drugs and Drug Resistance 2 (2012) 191–199

Contents lists available at SciVerse ScienceDirect

International Journal for Parasitology:
Drugs and Drug Resistance

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

# In silico prediction of antimalarial drug target candidates

Philipp Ludin [a,b], Ben Woodcroft [c,1], Stuart A. Ralph [c], Pascal Mäser [a,b,*]

[a] Swiss Tropical and Public Health Institute, 4002 Basel, Switzerland
[b] University of Basel, 4000 Basel, Switzerland
[c] Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Victoria 3010, Australia

ARTICLE INFO

*Article history:*
Received 19 April 2012
Received in revised form 28 June 2012
Accepted 3 July 2012
Available online 17 July 2012

*Keywords:*
*Plasmodium falciparum*
Phylogenomics
Gene essentiality
Drug target

ABSTRACT

The need for new antimalarials is persistent due to the emergence of drug resistant parasites. Here we aim to identify new drug targets in *Plasmodium falciparum* by phylogenomics among the *Plasmodium* spp. and comparative genomics to *Homo sapiens*. The proposed target discovery pipeline is largely independent of experimental data and based on the assumption that *P. falciparum* proteins are likely to be essential if (i) there are no similar proteins in the same proteome and (ii) they are highly conserved across the malaria parasites of mammals. This hypothesis was tested using sequenced Saccharomycetaceae species as a touchstone. Consecutive filters narrowed down the potential target space of *P. falciparum* to proteins that are likely to be essential, matchless in the human proteome, expressed in the blood stages of the parasite, and amenable to small molecule inhibition. The final set of 40 candidate drug targets was significantly enriched in essential proteins and comprised proven targets (e.g. dihydropteroate synthetase or enzymes of the non-mevalonate pathway), targets currently under investigation (e.g. calcium-dependent protein kinases), and new candidates of potential interest such as phosphomannose isomerase, phosphoenolpyruvate carboxylase, signaling components, and transporters. The targets were prioritized based on druggability indices and on the availability of in vitro assays. Potential inhibitors were inferred from similarity to known targets of other disease systems. The identified candidates from *P. falciparum* provide insight into biochemical peculiarities and vulnerable points of the malaria parasite and might serve as starting points for rational drug discovery.

## 1. Introduction

Drug discovery programs launched by the Medicines for Malaria Venture and other product-development partnerships have culminated in the development of promising new antimalarial compounds such as the synthetic peroxide OZ439 (Charman et al., 2011) and the spiroindolone NITD 609 (Rottmann et al., 2010), which are currently undergoing clinical trials. In spite of these recent successes, it is pivotal to maintain early phase drug discovery to prevent the antimalarial drug development pipeline from draining. Due to the propensity of the parasite to become drug-resistant (Muller and Hyde, 2010; Sa et al., 2011), the need for new antimalarial chemotypes will persist until the human-pathogenic *Plasmodium* spp. are eventually eradicated.

Rational post-genomic drug discovery is based on the screening of large chemical libraries – either virtually or in high-throughput format – against a given target enzyme of the parasite. A persistent bottleneck for target-based approaches is the identification of a suitable drug target in the first place. This enzyme should be essential for survival of the parasite and sufficiently different from its closest counterpart in the human host to be inhibited selectively. Experimental tools to validate candidate drug targets are limited for the malaria parasites. Gene silencing by RNAi does not seem to be feasible (Baum et al., 2009). Gene replacement with selectable markers is (Triglia et al., 1998), but it is inherently problematic to call a gene essential from failing to knock it out. Inducible degradation of *Plasmodium falciparum* proteins that have been fused to a FKBP-destabilization domain (Armstrong and Goldberg, 2007) is currently the most conclusive method for antimalarial target validation. However, none of the reverse genetic methods is practicable at the genome-wide scale. Adding up to the challenges with *Plasmodium* molecular biology is the lack of a phylogenetically close model organism that could serve as a point of reference – as is the case with parasitic nematodes, where essentiality of genes may be estimated based on the RNAi phenotypes (Schindelman et al., 2011) of orthologues in *Caenorhabditis elegans*.

* Corresponding author at: Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland. Tel.: +41 61 284 8338; fax: +41 61 284 8101.

*E-mail addresses:* philipp.ludin@unibas.ch (P. Ludin), b.woodcroft@uq.edu.au (B. Woodcroft), saralph@unimelb.edu.au (S.A. Ralph), pascal.maeser@unibas.ch (P. Mäser).

[1] Current address: School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, The University of Queensland, St. Lucia, Queensland 4072, Australia.

2211-3207/$ - see front matter © 2012 Australian Society for Parasitology Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijpddr.2012.07.002

Several bioinformatic approaches have previously been employed to help identify or prioritize drug targets for *Plasmodium* parasites. These include techniques based on automated identification of important steps in metabolic pathways (Yeh et al., 2004; Fatumo et al., 2009; Huthmacher et al., 2010; Plata et al., 2010), techniques that combine chemical starting points and protein-based queries (Joubert et al., 2009), as well as the use of the TDR-targets web-resource (http://www.tdrtargets.org) (Magarinos et al., 2012) to prioritize drug targets through the combination of multiple data types relevant to drug development (Crowther et al., 2010).

Here we try to predict antimalarial drug targets in silico, building on previous approaches by other labs for predicting essentiality of proteins based on phylogeny (Doyle et al., 2010; Waterhouse et al., 2010). We define a protein as a candidate antimalarial drug target if it (i) has conserved orthologues in all of the mammalian-pathogenic *Plasmodium* spp.; (ii) has no other match in *P. falciparum*; (iii) does not have a good match in the human proteome either; (iv) is expressed in the trophozoite and gametocyte stages; and (v) is predicted to function as an enzyme, receptor, or transporter. The rationale is that conserved genes often fulfill essential functions – or rather that genes fulfilling essential functions are under negative selection. A conserved single copy gene that lacks close matches in the same genome is likely to be indispensable. Starting from the complete predicted proteome of *P. falciparum* (Gardner et al., 2002), we applied consecutive filters to extract all candidate drug targets that meet the above criteria.

## 2. Material and methods

### 2.1. Datasets

The predicted *Plasmodium* spp. proteomes were downloaded from PlasmoDB (http://www.plasmodb.org/common/downloads) (Aurrecoechea et al., 2009), the *Saccharomyces cerevisiae* proteome from SGD (Saccharomyces genome database; http://www.downloads.yeastgenome.org/) (Engel et al., 2010), the *Homo sapiens* proteome from EBI (ftp://www.ftp.ebi.ac.uk/pub/databases/integr8/fasta/proteomes) (Mulder et al., 2008), and all others from UniProt (http://www.uniprot.org/taxonomy) (Magrane and Consortium, 2011). *P. falciparum* 3D7 cell cycle expression data (Le Roch et al., 2003) were obtained from PlasmoDB, using as a threshold for expression $E > 10$ and probe signal distribution $\log P < -0.5$ as proposed by the authors (Le Roch et al., 2003). *S. cerevisiae* deletion phenotype data were from SGD (http://www.downloads.yeastgenome.org/curation/literature/phenotype_data.tab). Proteins were termed essential if the phenotype of the knock-out (mutant type = 'null') of the corresponding gene was 'inviable'. The TDRtargets web resource (http://www.tdrtargets.org) (Magarinos et al., 2012), as well as the BRENDA database (http://www.brenda-enzymes.org) (Scheer et al., 2011) was used to identify proteins with precedence for interaction with small molecule chemical inhibitors. Essentiality of *Plasmodium berghei* orthologues was found through RMgmDB (Rodent Malaria genetically modified Parasites; http://www.pberghei.eu/) (Janse et al., 2011).

### 2.2. Programs

InParanoid was obtained from http://www.inparanoid.sbc.su.se and QuickParanoid from http://www.pl.postech.ac.kr/QuickParanoid. BLAST 2.2.17 (Altschul et al., 1990) was downloaded from NCBI (ftp.ncbi.nlm.nih.gov), Needle (from the EMBOSS suite (Rice et al., 2000)) from http://www.emboss.sourceforge.net. All programs were run on the Basel Biocomputing Linux cluster (http://www.bc2.unibas.ch).

### 2.3. Algorithms

The pipeline for comparative genomics and target prediction was run in a semi-automated way, combining the existing programs (see Section 2.2) with self-made Perl scripts for parsing, re-formatting, and analysis of outputs. Druggability scores were extracted from the TDRtargets website (Magarinos et al., 2012). These druggability predictions are numerical values between 0 and 1, with 1 corresponding to more druggable targets. Scores reflect a combination of factors including degree of similarity to a large library of targets within the ChEMBL database (Hopkins et al., 2011; Gaulton et al., 2012), with empirically determined interactions with drug like compounds, possession of physiochemical features enriched in known drug targets, and the drug-like quality of the inhibitors known to interact with those homologues. These scores are outputs of a combination of multiple discriminators based on empirically-characterized drug targets, but are untested in *Plasmodium*, other than through concordance of high-scoring proteins with the very few known antimalarial drug targets. Proteins with higher druggability scores are likely to be more amenable to inhibition with drug-like molecules than proteins with low druggability scores, and hence are used here as one aspect of the ranking system, but there is lack of parasite-specific direct data for these algorithms with which to assign a cutoff for designating an acceptable score.

## 3. Results and discussion

### 3.1. Using yeast as a touchstone for essentiality prediction

The central hypothesis implemented in the present target discovery pipeline is, that a protein is likely to be essential if there are no other similar proteins in the same proteome but conserved orthologues in all the related species. Before applying it to malaria parasites, we tested the hypothesis on the yeast *S. cerevisiae* where essentiality of proteins has been tested by systematic knock-out of the respective genes (Winzeler et al., 1999). The current null mutant dataset in the *Saccharomyces* Genome Database comprises 5477 genes, of which 1109 were found to be essential. To test whether a given *S. cerevisiae* protein is unique, its best match (excluding itself) in the same proteome was identified with Blastp (Altschul et al., 1990), then and a Needleman-Wunsch global alignment (Needleman and Wunsch, 1970) was performed between the two proteins. The median global identity between a *S. cerevisiae* protein and its best match was 15.7% (red bar in Fig. 1). 14% of the proteins with a match above the median were essential, compared to 24% of those below the median (Fig. 1). The difference was

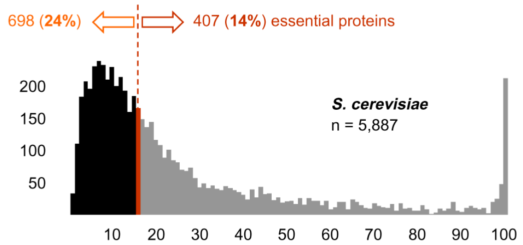


**Fig. 1.** *S. cerevisiae* non-redundant proteins. Frequency distribution of the similarities for all predicted *S. cerevisiae* proteins to their next best match in the same proteome. Of the 1109 essential proteins, 698 were below the median identity of 15.7% (red bar) and 407 above (4 proteins had exactly 15.7% identity to their next best match). The black area represents the proteins below the threshold of ⩽15% global identity. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

significant ($p < 0.0001$, two-sided Chi-square test) and supported the hypothesis that 'unmatched' proteins are more likely to be essential (Hannay et al., 2008). Setting the cut-off at ⩽15% global identity 2648 *S. cerevisiae* proteins passed (black area in Fig. 1), 26% of which were essential (Fig. 2).

To determine which of the *S. cerevisiae* proteins are conserved across the Saccharomycetaceae, a set of related (Fitzpatrick et al., 2006) and fully sequenced yeasts served as the reference: *Candida glabrata*, *Ashbya gossypii*, *Kluyveromyces lactis*, *Candida albicans*, *Debaryomyces hansenii*, and *Yarrowia lipolytica*. The predicted proteomes (Table 1) were fed into InParanoid and QuickParanoid (Ostlund et al., 2010), programs that utilize reciprocal, proteome-wide Blastp searches to build groups of orthologous proteins and rank them. A total of 3546 *S. cerevisiae* proteins possessed orthologues in all the included yeasts. The highly conserved orthology clusters, i.e. those above median similarity, were significantly enriched for essential proteins (Fig. 2): 29% of the *S. cerevisiae* proteins that belonged to a conserved cluster were essential ($p < 0.0001$, two-sided Chi-square test). Finally the intersect of the two sets, unmatched and conserved, contained 632 proteins of which 40% were essential (Fig. 2). The yeast results supporting our working hypothesis, we went on to apply it to malaria parasites.

### 3.2. Essentiality prediction in P. falciparum

The predicted *P. falciparum* proteome exhibits a strikingly low degree of redundancy, with the large majority of proteins exhibiting but a low level of similarity to their next best match in the same proteome (Fig. 3). As described above (Section 3.1) for *S. cerevisiae*, the closest hit to any *P. falciparum* protein was determined with Blastp. Similarity was measured by global alignment and expressed as percent identical amino acids. The median global identity between a *P. falciparum* protein and its next best match was only 13% (red bar in Fig. 3), and 3078 of the 5410 proteins passed the threshold of ⩽15% global identity to their next best match (black area in Fig. 3).

The conserved *Plasmodium* proteins were identified by comparing *P. falciparum* to all other available genome sequences of mammalian malaria parasites, namely *P. berghei*, *Plasmodium chabaudi* and *Plasmodium yoelii* (mouse), *Plasmodium knowlesi* (primate), and *Plasmodium vivax* (human tertian malaria). The six genomes encode a total of 33,704 predicted proteins (Table 1). The common intersect of proteins from orthology clusters represented in all the six species was 26,377, as determined with InParanoid and QuickParanoid (Ostlund et al., 2010). This amounts to 78% of all proteins, reflecting the close relationship of the malaria parasites. The number of proteins in the common clusters was conserved in all the

**Table 1**
Proteomes used in this study, their sizes and sources.

| Species | Proteins | Source | Version |
|---|---|---|---|
| *P. falciparum* | 5410 | PlasmoDB | 8.1 |
| *P. vivax* | 5396 | PlasmoDB | 8.1 |
| *P. berghei* | 4861 | PlasmoDB | 8.1 |
| *P. chabaudi* | 5119 | PlasmoDB | 8.1 |
| *P. knowlesi* | 5194 | PlasmoDB | 8.1 |
| *P. yoelii* | 7724 | PlasmoDB | 8.1 |
| *H. sapiens* | 67,172 | Integr8 | 08/2011 |
| *S. cerevisiae* | 5887 | SGD | R64-1-1 |
| *A. gossypii* | 4761 | UniProt | 08/2011 |
| *C. albicans* | 5727 | UniProt | 08/2011 |
| *C. glabrata* | 5197 | UniProt | 08/2011 |
| *D. hansenii* | 6242 | UniProt | 08/2011 |
| *K. lactis* | 5074 | UniProt | 08/2011 |
| *Y. lipolytica* | 6414 | UniProt | 08/2011 |

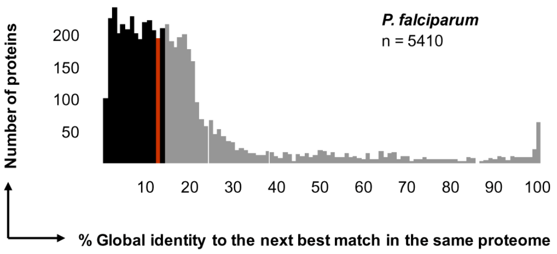
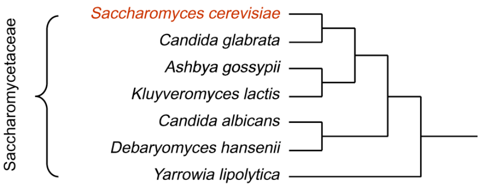


**Fig. 3.** *P. falciparum* non-redundant proteins. Frequency distribution of the similarities for all predicted *P. falciparum* proteins to their next best match in the same proteome. The median is shown in red, the black area represents the proteins which, passing the threshold of ⩽15% global identity, are assumed to be enriched for essential ones.



| | Total | Essential | |
|---|---|---|---|
| *S. cerevisiae* proteome | 5,477 | 1,109 | 20% |
| Unmatched >15% global identity | 2,648 | 677 | 26% |
| Conserved in Saccharomycetaceae | 1,773 | 514 | 29% |
| **Unmatched & conserved** | **632** | **253** | **40%** |

**Fig. 2.** Testing for essentiality with yeasts. The cladogram depicts the phylogeny of the fully sequenced Saccharomycetaceae (Fitzpatrick et al., 2006). Eliminating the *S. cerevisiae* proteins which possess matches in the same proteome (gray area in Fig. 1), and those which are not conserved across all the analyzed yeasts, enriches for essential proteins.

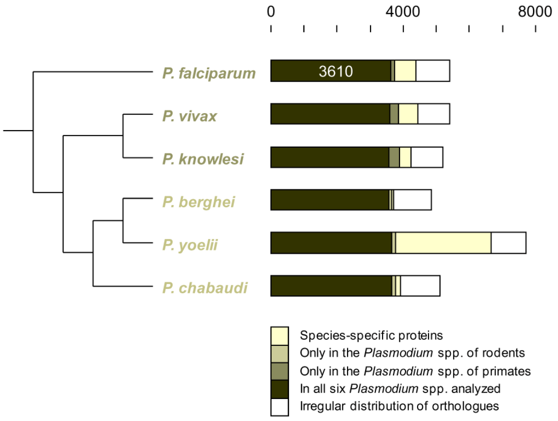


**Fig. 4.** The conserved malaria proteome. The cladogram depicts the phylogeny of the fully sequenced malaria parasites of mammals (Martinsen et al., 2008). The bars divide each predicted proteome into proteins present only in the species itself, proteins with orthologues only in the *Plasmodium* spp. sharing intermediate hosts, and proteins with orthologues in all the analyzed species (3610 in *P. falciparum*). The white sections represent proteins with an irregular distribution insofar as the presence of orthologues did not correlate with host specificity (nor with phylogeny).

analyzed *Plasmodium* spp., and the size of their proteomes correlated with the number of species-specific proteins (Fig. 4). Of the 5410 predicted *P. falciparum* proteins, 3610 (67%) possessed at least one orthologue in all of the other *Plasmodium* species (Fig. 4). 630 *P. falciparum* proteins did not have an orthologue in any other species and 139 had an orthologue in *P. vivax* and *P. knowlesi* but not in the malaria parasites of rodents (Fig. 4). The different protein subsets outlined in Fig. 4 for each parasite are available as a Supplementary excel file (supplement1.xls). The intersect of the two subsets, *P. falciparum* unique (*n* = 3078) and *P. falciparum* highly conserved (i.e. above median, *n* = 1805), contained 944 proteins which we posit – based on the results with *S. cerevisiae* – to be enriched for essential ones.

### 3.3. Comparative genomics between parasite and host

For each predicted protein of *P. falciparum* the best match in the human proteome was identified with Blastp, ranking the obtained hits by ascending expectancy (*E*-values). The similarity between a *P. falciparum* protein and this closest match from *H. sapiens* was then quantified with a Needleman-Wunsch global alignment (since the local alignment provided by Blast may be misleading if the similarity between two otherwise unrelated proteins is confined to a smaller common domain). The frequency distribution of global identity to human proteins in the *P. falciparum* proteome is shown in Fig. 5. The most conserved proteins between parasite and host were ubiquitins, histones, calmodulin, and tubulin. The median global identity between a *P. falciparum* protein and its best match in *H. sapiens* was 10% (red bar in Fig. 5). Proteins exhibiting >10% identity to the host (gray area in Fig. 5) were removed from the potential target space, which thereby was trimmed down to 426 proteins (Fig. 6). Some of these still had highly significant Blastp *E*-values to a human protein, e.g. the *P. falciparum* kinases listed in Table 2, since a short region of high similarity in two otherwise disparate proteins can return a high local alignment score (in the case of the *P. falciparum* kinases, this region is the catalytic domain). Nevertheless, selective targeting might be possible against other domains.

### 3.4. Blood-stage expression of putative targets

The majority of genes in *P. falciparum* are expressed in a stage-dependent manner (Llinas and DeRisi, 2004), and it is in the asexual and gametocyte stages that we attempt to find targets given their roles in pathology and transmission. We relied on a published microarray dataset (Le Roch et al., 2003) of *P. falciparum* cultures synchronized by sorbitol treatment to identify the potential targets that are present in the blood stages. Retaining only the products of genes which are expressed in the ring stages, trophozoites, or schizonts, reduced the target space to 288 proteins (Fig. 6). The complete list is available as a Supplementary excel file (supplement2.xls). Over half of these (*n* = 178) could not be assigned a putative function. Although of potential interest since conserved in the malaria parasites and absent in humans, these proteins were (for the time being) removed from the candidate target set. Other proteins were removed manually because they were not likely to have a function that is susceptible to inhibition, e.g. structural proteins or surface antigens. Focusing on proteins with a clear, though sometimes putative, functional annotation in PlasmoDB as enzyme, receptor or transporter resulted in a final set of 40 *P. falciparum* candidate drug targets (Fig. 6).

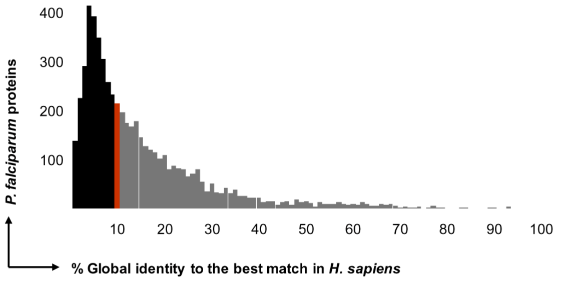


**Fig. 5.** Parasite vs. host. Frequency distribution of percent global identity to *H. sapiens* in the *P. falciparum* proteome. The red bar marks the median, the gray areas the *P. falciparum* proteins that were excluded from the potential target space.

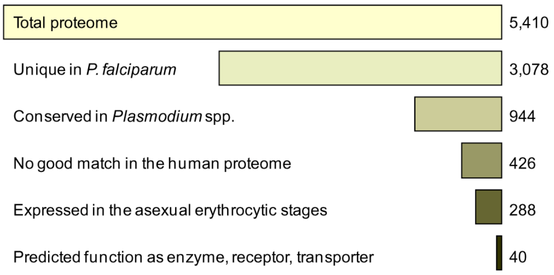


**Fig. 6.** From genome to target. Overview on the in silico pipeline to narrow down the potential target space of *P. falciparum*. The 40 identified candidates are listed in Table 2.

### 3.5. Assessing essentiality of the 40 identified candidate targets

For four of the predicted essential proteins attempts have been made to disrupt the corresponding gene in *P. falciparum* (Table 3). In three cases disruption was attempted through deletion by recombination but no gene deletion could be obtained (Santiago et al., 2004; Wang et al., 2004; Abdi et al., 2010). For the fourth gene ispH (HMBPP reductase; PFA0225w), introduction of a hammerhead ribozyme targeting the transcript was lethal (Vinayak and Sharma, 2007). Thus, each of these four genes experimentally examined appeared to be essential. Similarly, a small number of these proteins' orthologues in *P. berghei* or *Toxoplasma gondii* have been the subject of directed gene knock-out or mutagenesis experiments aimed at assaying their essentiality. The orthologues of the kinases PF13_0166 and CDPK7 (PF11_0242) have been suggested to be essential in *P. berghei* (Tewari et al., 2010). In *T. gondii*, the orthologue of the apurinic/apyrimidinic endonuclease (PF13_0176) has been shown to essential by conditional knockout (Onyango et al., 2011) and GAP40 (PFE0785c) is part of the essential glideosome complex, though the protein itself has not been verified as essential (Frenal et al., 2010). Only one gene was found to be non-essential. The *P. berghei* CDPK6 (PF11_0239) orthologue has been successfully deleted by double crossover (Coppi et al., 2007). These experimental assays for essentiality thus indicate eight of the predicted proteins to be essential, and only one to be non-essential, supporting our method to enrich for essentiality.

### 3.6. Prediction of druggability

In order to recognize the important role of chemical accessibility in ranking of putative drug targets we looked for druggability information on the predicted essential targets. Druggability describes the properties of a protein that make them able to interact with a drug-like molecule. Some proteins have physiochemical and structural properties that make them unlikely to ever bind a compound with the necessary characteristic to be a drug, and whole genome analyses have estimated that only 10% of a genome may represent druggable proteins (Hopkins and Groom, 2002). As part

**Table 2**
Forty candidate drug targets from *P. falciparum* identified as outlined in Fig. 6, their putative function, Blastp *E*-value to the best hit in the human proteome, and druggability index (D.I.). The categories (*italics*) are arbitrary and mainly meant to illustrate the functional diversity of the identified targets.

| Accession | Putative function | %Id *Hsa* | *E Hsa* | D.I. |
|---|---|---|---|---|
| *Metabolism* | | | | |
| PF08_0095 | Dihydropteroate synthetase (2.5.1.15) | 0.8 | 8.5 | 0.8 |
| MAL8P1.156 | Mannose-6-P isomerase (5.3.1.8) | 6.9 | 1E-8 | 0.6 |
| PF08_0068 | FAD-dependent monooxygenase | 6.0 | 0.07 | 0.5 |
| PF14_0334 | Ornithine aminotransferase (2.6.1.13) | 4.3 | 1E-14 | 0.3 |
| PF13_0234 | PEP carboxykinase (4.1.1.49) | 4.0 | 3.6 | 0.2 |
| PF14_0246 | PEP carboxylase (4.1.1.31) | 3.8 | 4.5 | 0.1 |
| PFI1180w | Patatin-like phospholipase | 6.3 | 7.4 | 0.1 |
| PF10_0221 | HMB-PP synthase (GcpE, 1.17.7.1) | 6.5 | 2.2 | |
| PFA0225w | HMB-PP reductase (1.17.1.2) | 7.3 | 5.3 | |
| MAL7P1.88 | Thioredoxin-like protein | 3.1 | 4.9 | |
| PFI1340w | Fumarate hydratase (Fumarase, 4.2.1.2) | 10.0 | 2.7 | |
| PF08_0132 | Glutamate dehydrogenase C (1.4.1.2) | 8.1 | 7.2 | |
| PFD0285c | Lysine decarboxylase (4.1.1.18) | 5.6 | 7.4 | |
| PFL0620c | Glycerol-3-P acyltransferase (2.3.1.15) | 9.8 | 1.0 | |
| PF14_0250 | Lipase | 3.3 | 8.1 | |
| *Nucleic acids* | | | | |
| PF14_0273 | RNA methyltransferase | 9.8 | 7E-30 | 0.7 |
| PF13_0080 | Telomerase reverse transcriptase (2.7.7.49) | 0.3 | 0.04 | 0.5 |
| PF13_0176 | Apur./apyr. endonuclease (4.2.99.18) | 2.6 | 2.6 | 0.2 |
| MAL13P1.311 | Exonuclease | 5.9 | 7E-4 | 0.1 |
| MAL13P1.328 | DNA topoisomerase VI, b subunit | 4.9 | 7.1 | |
| PFE0675c | DNA photolyase (4.1.99.3) | 3.7 | 0.06 | |
| PF13_0048 | NUDIX hydrolase (3.6.1.17) | 8.4 | 8E-35 | |
| PF08_0111 | RNA helicase | 7.8 | 8E-15 | |
| PFC0980c | RNA triphosphatase | 8.5 | 8.6 | |
| *Signaling* | | | | |
| PF11_0239 | $Ca^{2+}$-dependent protein kinase 6 | 7.8 | 9E-34 | 0.5 |
| PF11_0242 | $Ca^{2+}$-dependent protein kinase 7 | 9.0 | 6E-41 | 0.5 |
| PF13_0258 | Serine/threonine protein kinase TKL3 | 4.8 | 3E-21 | 0.5 |
| PFA0515w | Phosphatidylinositol-4-P-5-kinase (2.7.1.68) | 7.0 | 1E-24 | 0.5 |
| PF13_0166 | Protein kinase | 8.9 | 7E-4 | 0.3 |
| PF07_0024 | Inositol phosphatase | 7.3 | 1E-31 | 0.3 |
| MAL7P1.18 | Serine/threonine protein kinase | 4.7 | 5E-13 | 0.1 |
| MAL7P1.64 | G protein-coupled receptor | 2.5 | 93 | 0.1 |
| PF14_0143 | Atypical protein kinase, ABC-1 family | 3.8 | 4E-26 | |
| PF14_0614 | Phosphatase | 6.1 | 4E-5 | |
| *Protease/chaperone* | | | | |
| PF14_0382 | Stromal-processing peptidase | 4.4 | 0.001 | 0.5 |
| PF10_0032 | DnaJ protein | 2.4 | 7E-4 | |
| *Transporter* | | | | |
| PF11_0092 | Mechanosensitive ion channel | 5.0 | 11 | 0.1 |
| PF13_0019 | $Na^+$:$H^+$ antiporter | 6.6 | 0.01 | 0.1 |
| PFI0785c | Sugar transporter | 2.8 | 0.09 | 0.1 |
| PFE0785c | Metabolite transporter/glideosome component | 0.7 | 10 | |

of the TDRtargets project a druggability score was assigned to enzymes from the *P. falciparum* proteome (Aguero et al., 2008; Magarinos et al., 2012). This score predicts the suitability of the protein for binding a drug like molecule, and is an aggregate of several factors including physiochemical and structural features of the protein and sequence similarity to validated druggable proteins from other organisms (Al-Lazikani et al., 2007; Gaulton et al., 2012). Of the 40 targets predicted as essential, 23 have positive druggability scores (Table 2). Those with the highest scores are predicted to have more druggable structures and might therefore be prioritized over potential targets with low or zero druggability scores. The higher druggability scores are more likely to be indicative of promising drug target-like proteins, but any positive score indicates structural or physiochemical similarity to proteins with some precedence for interaction with a drug-like molecule.

To search for proteins with specific compounds as starting points for inhibitor development we also searched for proteins that had been identified as the targets of inhibitors, or that had orthologues with identified inhibitors. Searches were conducted of the TDRTargets database, the ChEMBL database and the BRENDA database to identify such targets. Of the 40 predicted essential *P. falciparum* genes, 2 have been directly validated as the targets of inhibitors. One of these, dihydropteroate synthetase (DHPS; PF08_0095) is one of the very few validated targets of clinically used antimalarials, consistent with the value of our approach for identifying essential targets. An additional 13 proteins have orthologues with identified small molecule inhibitors. Identification of these inhibitors not only provides useful starting points for identifying small molecules with which to interrogate the *Plasmodium* targets, but also indicates that these are likely to be potentially druggable targets. Compounds were curated to exclude facile inhibitors such as natural amino acids, ATP and NaCl, however no systematic attempt was made to rate the drug-like quality of the predicted inhibitors. For the kinases in the list a slightly modified approach was taken, as many kinase inhibitors are well known to be cross reactive against multiple kinases. An empirical cut-off of 30% local sequence identity in the kinase domains was therefore used to identify homologues for potential inhibitors of from the TDR/ChEMBL data (Table 3).

### 3.7. Experimental amenability of the predicted targets

Tractability of a target for developing an assay in which inhibition is a major factor in prioritizing promising drug targets in

**Table 3**
Identified targets, evidence for essentiality in apicomplexan parasites (*Pfa, P. falciparum*; *Pbe, P. berghei*; *Tgo, T. gondii*), known and inferred inhibitors, and published in vitro assays.

| Accession | Enzyme | Essentiality | Inhibitors | Assay | References |
|---|---|---|---|---|---|
| PF08_0095 | DHPS | *Pfa* | Sulfonamides | Yes | Wang et al. (2010) |
| PF14_0273 | RNA methyltransferase | | 8 inferred | Yes | Riguet et al. (2005), Leber et al. (2009) |
| MAL8P1.156 | M6P isomerase | | 8 | | Wells et al. (1995), Salvati et al. (2001) and Foret et al. (2009) |
| PF11_0242 | CDPK7 | *Pbe* | 72 inferred | | 12 Publications |
| PF11_0239 | CDPK6 | Not in *Pbe* | 2488 inferred | | 189 Publications |
| PFA0515w | PI-4-P-5-kinase | | 1 | | Kobayashi et al. (2005) and Santiago et al. (2004) |
| PF13_0080 | Telomerase RT | | Berberine | | Sriwilaijareon et al. (2002) |
| PF13_0258 | TKL3 | *Pfa* | | Yes | Abdi et al. (2010) |
| PF13_0166 | Protein kinase | *Pbe* | | | |
| PF14_0334 | OAT | | >10 | Yes | Strecker (1965), Levillain et al. (2000), Wang et al. (2007), Stranska et al. (2008) and Gafan et al. (2001) |
| PF13_0176 | AP endonuclease | *Tgo* | >20 | Yes | Onyango et al. (2011), Zawahir et al. (2009), Bapat et al. (2010), Adhikari et al. (2012) and Haltiwanger et al. (2000) |
| PF13_0234 | PEP carboxykinase | | | Yes | Hayward (2000) |
| PF14_0246 | PEP carboxylase | | 9 | Yes | Pairoba et al. (1996), McDaniel and Siu (1972) and Jenkins (1989) |
| PFD0285c | Lysine decarboxylase | | 4 | | Yamamoto et al. (1991) and Takatsuka et al. (1999) |
| PFI1340w | Fumarase | | 2 | Yes | Flint (1994), Beeckmans and Van Driessche (1998) and Bulusu et al. (2011) |
| PFL0620c | G3P acyltransferase | *Pfa* | >40 | Yes | Santiago et al. (2004), Yamashita and Numa (1981), Coleman (1988) and Wydysh et al. (2009) |
| PF13_0048 | NUDIX hydrolase | | >20 | | Guranowski et al. (2003, 2009) and Branson et al. (2009) |
| PF08_0132 | GDH C | | >20 | Yes | Li et al. (2007, 2009), Rodriguez-Acosta et al. (1999) |
| PFA0225w | HMB-PP reductase | *Pfa* | | Yes | Vinayak and Sharma (2007) and Rohrich et al. (2005) |

infectious diseases because the smaller research communities, compared to those working on human proteins, means fewer protein reagents are available. This is exacerbated in malaria research because the *P. falciparum* genome is extremely A + T rich, and many proteins have long low-complexity insertions, making recombinant protein production challenging. High throughput attempts to generate recombinant *Plasmodium* enzymes have relatively high failure rates (Abdi et al., 2010), so the presence of an existing assay is a particularly relevant criteria for target prioritization. We therefore prioritized proteins where researchers had already developed a direct in vitro assay for the *P. falciparum* enzyme. Eleven of the 40 predicted essential enzymes have had in vitro assays described in *P. falciparum*, although in some cases these may not be readily amenable to high throughput inhibitor screening.

### 3.8. Candidate antimalarial drug targets

The approach presented here, although starting from the complete proteome, is by no means all-embracing in the sense that it would discover every potential drug target. Many essential proteins and known drug targets were missed, such as *P. falciparum* enzymes which possess human orthologues but exhibit exploitable differences nevertheless. These could be differences in sequence, as is the case with glyceraldehyde-3-phosphate dehydrogenase (Satchell et al., 2005) or in regulation, as demonstrated for dihydrofolate reductase (Zhang and Rathod, 2002). Other proven targets were eliminated because they possess several paralogues, e.g. the falcipains (Teixeira et al., 2011). Using rigorous filters, we tried to maximize the specificity of target prediction at the cost of sensitivity, taking into account a large number of false negatives. Our aim was not to identify all potential antimalarial drug targets but rather to come up with a sizeable list of promising targets that (i) may serve as starting points for rational drug discovery and (ii) provide new insights into peculiarities and vulnerable points of the malaria parasites. The applied selection criteria were rigorous (Fig. 6), so the list of potential targets could be expanded by using less stringent filters. The present approach identified 40 candidate drug targets from *P. falciparum* which are shown in Table 2. We are confident that this set contains promising leads because it includes known targets such as dihydropteroate synthetase (PF08_0095), the target of sulfonamides (dihydrofolate synthetase is not included because it is 20% identical to human mitochondrial tetrahydrofolylpolyglutamate synthase, EC 6.3.2.17), and enzymes of the non-mevalonate pathway for isoprenoid synthesis. Another potential target in the apicoplast could be stromal processing peptidase (PF14_0382), which has already been characterized in *P. falciparum* (van Dooren et al., 2002). Investigational targets identified include glutamate dehydrogenase C (PF08_0132) (Aparicio et al., 2010; Storm et al., 2011) and Ca$^{2+}$-dependent protein kinases (PF11_0239, PF11_0242) (Kato et al., 2008; Ojo et al., 2010; Lim et al., 2012). In addition, there are enzymes which are being targeted in other systems, e.g. apurinic/apyrimidinic endonuclease (PF13_0176) (Tell et al., 2010) or telomerase reverse transcriptase (PF13_0080) (Chen et al., 2009) by antitumor agents, or phosphomannose isomerase (MAL8P1.156) by antifungals. Phosphomannose is a key enzyme in yeast cell wall synthesis (Mastrolorenzo et al., 2000; Wills et al., 2001); to our knowledge, the *P. falciparum* orthologue has so far not been studied. A surprising find in the list of candidate targets was phosphoenolpyruvate carboxylase (PF14_0246), the enzyme responsible for carbon fixation in C4 plants. The *P. falciparum* orthologue was characterized and proposed to fix atmospheric carbon as well (McDaniel and Siu, 1972). If this process is essential, PEP carboxylase could be an attractive target since inhibitors have been identified as C4 plant herbicides (Madhusudana et al., 1980; Jenkins, 1989; Pairoba et al., 1996). Further candidate targets of interest could be the identified signaling enzymes and nutrient transporters (Table 2).

## 4. Conclusion

Translating a pathogen's genome sequence into new drugs is a key challenge to post-genomic biology. Several strategies for the prediction of drug targets that are essential and unique to the parasite have been proposed (Payne et al., 2001; Joyce and Palsson, 2008; Holman et al., 2009; Crowther et al., 2010; Deng et al., 2010; Doyle et al., 2010; Magarinos et al., 2012). Here we implement phylogenomic comparison within a clade of related parasites to infer on protein essentiality, rather than referring to a model organism which may be understood in detail but phylogenetically distant to the parasites. In addition, we estimate redundancy across the parasite's proteome based on global alignments to quantify protein similarity. Thus the proposed target identification pipeline does not rely on experimental data to predict essentiality, and therefore is generally applicable. The identified candidate targets from *P. falciparum* are, by definition, present in all the *Plasmodium* spp. analyzed. Conservation of targets between *P. falciparum* and *P. vivax* is important since eventual new drugs need to be active against both parasites. Conservation of the candidate targets between human- and rodent-pathogenic *Plasmodium* spp. is desirable as well, since mouse models are indispensable in the drug development process. The presented target discovery pipeline is corroborated by the presence of known antimalarial targets in the output list (Table 2). We hope that the newly identified candidate targets will support the development of novel antimalarials.

## Acknowledgments

We wish to thank the Basel Computational Biology Center for runtime on their Linux cluster. SAR is funded by an Australian Research Council Future Fellowship (FT0990350). BJW was supported by a University of Melbourne Research scholarship.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijpddr.2012.07.002.

## References

Abdi, A., Eschenlauer, S., Reininger, L., Doerig, C., 2010. SAM domain-dependent activity of PfTKL3, an essential tyrosine kinase-like kinase of the human malaria parasite *Plasmodium falciparum*. Cell. Mol. Life Sci. 67, 3355–3369.

Adhikari, S., Manthena, P.V., Kota, K.K., Karmahapatra, S.K., Roy, G., Saxena, R., Uren, A., Roy, R., 2012. A comparative study of recombinant mouse and human apurinic/apyrimidinic endonuclease. Mol. Cell. Biochem. 362, 195–201.

Aguero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F.S., Campbell, R.K., Carmona, S., Carruthers, I.M., Chan, A.W., Chen, F., Crowther, G.J., Doyle, M.A., Hertz-Fowler, C., Hopkins, A.L., McAllister, G., Nwaka, S., Overington, J.P., Pain, A., Paolini, G.V., Pieper, U., Ralph, S.A., Riechers, A., Roos, D.S., Sali, A., Shanmugam, D., Suzuki, T., Van Voorhis, W.C., Verlinde, C.L., 2008. Genomic-scale prioritization of drug targets: the TDR Targets database. Nat. Rev. Drug Discov. 7, 900–907.

Al-Lazikani, B., Gaulton, A., Paolini, G., Lanfear, J., Overington, J., Hopkins, A.L. (2007). The molecular basis of predicting druggability. In: Bioinformatics – From Genomes to Therapies. In: The Holy Grail: Molecular Function, vol. 3, Wiley, T.L. Weinheim, pp. 1315–1334.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Aparicio, I.M., Marin-Menendez, A., Bell, A., Engel, P.C., 2010. Susceptibility of *Plasmodium falciparum* to glutamate dehydrogenase inhibitors–a possible new antimalarial target. Mol. Biochem. Parasitol. 172, 152–155.

Armstrong, C.M., Goldberg, D.E., 2007. An FKBP destabilization domain modulates protein levels in *Plasmodium falciparum*. Nat. Methods 4, 1007–1009.

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert Jr., C.J., Treatman, C., Wang, H., 2009. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 37, D539–543.

Bapat, A., Glass, L.S., Luo, M., Fishel, M.L., Long, E.C., Georgiadis, M.M., Kelley, M.R., 2010. Novel small-molecule inhibitor of apurinic/apyrimidinic endonuclease 1 blocks proliferation and reduces viability of glioblastoma cells. J. Pharmacol. Exp. Ther. 334, 988–998.

Baum, J., Papenfuss, A.T., Mair, G.R., Janse, C.J., Vlachou, D., Waters, A.P., Cowman, A.F., Crabb, B.S., de Koning-Ward, T.F., 2009. Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. Nucleic Acids Res. 37, 3788–3798.

Beeckmans, S., Van Driessche, E., 1998. Pig heart fumarase contains two distinct substrate-binding sites differing in affinity. J. Biol. Chem. 273, 31661–31669.

Branson, K.M., Mertens, H.D., Swarbrick, J.D., Fletcher, J.I., Kedzierski, L., Gayler, K.R., Gooley, P.R., Smith, B.J., 2009. Discovery of inhibitors of lupin diadenosine 5′,5‴-P(1), P(4)-tetraphosphate hydrolase by virtual screening. Biochemistry 48, 7614–7620.

Bulusu, V., Jayaraman, V., Balaram, H., 2011. Metabolic fate of fumarate, a side product of the purine salvage pathway in the intraerythrocytic stages of *Plasmodium falciparum*. J. Biol. Chem. 286, 9236–9245.

Charman, S.A., Arbe-Barnes, S., Bathurst, I.C., Brun, R., Campbell, M., Charman, W.N., Chiu, F.C., Chollet, J., Craft, J.C., Creek, D.J., Dong, Y., Matile, H., Maurer, M., Morizzi, J., Nguyen, T., Papastogiannidis, P., Scheurer, C., Shackleford, D.M., Sriraghavan, K., Stingelin, L., Tang, Y., Urwyler, H., Wang, X., White, K.L., Wittlin, S., Zhou, L., Vennerstrom, J.L., 2011. Synthetic ozonide drug candidate OZ439 offers new hope for a single-dose cure of uncomplicated malaria. Proc. Natl. Acad. Sci. USA 108, 4400–4405.

Chen, H., Li, Y., Tollefsbol, T.O., 2009. Strategies targeting telomerase inhibition. Mol. Biotechnol. 41, 194–199.

Coleman, R.A., 1988. Hepatic sn-glycerol-3-phosphate acyltransferases: effect of monoacylglycerol analogs on mitochondrial and microsomal activities. Biochim. Biophys. Acta 963, 367–374.

Coppi, A., Tewari, R., Bishop, J.R., Bennett, B.L., Lawrence, R., Esko, J.D., Billker, O., Sinnis, P., 2007. Heparan sulfate proteoglycans provide a signal to Plasmodium sporozoites to stop migrating and productively invade host cells. Cell Host Microbe 2, 316–327.

Crowther, G.J., Shanmugam, D., Carmona, S.J., Doyle, M.A., Hertz-Fowler, C., Berriman, M., Nwaka, S., Ralph, S.A., Roos, D.S., Van Voorhis, W.C., Aguero, F., 2010. Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. PLoS Negl. Trop. Dis. 4, e804.

Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A.A., Hassett, D.J., Lu, L.J., 2010. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. Nucleic Acids Res. 39, 795–807.

Doyle, M.A., Gasser, R.B., Woodcroft, B.J., Hall, R.S., Ralph, S.A., 2010. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. BMC Genomics 11, 222.

Engel, S.R., Balakrishnan, R., Binkley, G., Christie, K.R., Costanzo, M.C., Dwight, S.S., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Hong, E.L., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R., Oughtred, R., Park, J., Skrzypek, M.S., Weng, S., Wong, E.D., Dolinski, K., Botstein, D., Cherry, J.M., 2010. Saccharomyces Genome Database provides mutant phenotype data. Nucleic Acids Res. 38, D433–436.

Fatumo, S., Plaimas, K., Mallm, J.P., Schramm, G., Adebiyi, E., Oswald, M., Eils, R., Konig, R., 2009. Estimating novel potential drug targets of *Plasmodium falciparum* by analysing the metabolic network of knock-out strains in silico. Infect. Genet. Evol. 9, 351–358.

Fitzpatrick, D.A., Logue, M.E., Stajich, J.E., Butler, G., 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. 6, 99.

Flint, D.H., 1994. Initial kinetic and mechanistic characterization of Escherichia coli fumarase A. Arch. Biochem. Biophys. 311, 509–516.

Foret, J., de Courcy, B., Gresh, N., Piquemal, J.P., Salmon, L., 2009. Synthesis and evaluation of non-hydrolyzable D-mannose 6-phosphate surrogates reveal 6-deoxy-6-dicarboxymethyl-D-mannose as a new strong inhibitor of phosphomannose isomerases. Bioorg. Med. Chem. 17, 7100–7107.

Frenal, K., Polonais, V., Marq, J.B., Stratmann, R., Limenitakis, J., Soldati-Favre, D., 2010. Functional dissection of the apicomplexan glideosome molecular architecture. Cell Host Microbe 8, 343–357.

Gafan, C., Wilson, J., Berger, L.C., Berger, B.J., 2001. Characterization of the ornithine aminotransferase from *Plasmodium falciparum*. Mol. Biochem. Parasitol. 118, 1–10.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100–1107.

Guranowski, A., Starzynska, E., McLennan, A.G., Baraniak, J., Stec, W.J., 2003. Adenosine-5′-O-phosphorylated and adenosine-5′-O-phosphorothioylated polyols as strong inhibitors of (symmetrical) and (asymmetrical) dinucleoside tetraphosphatases. Biochem. J. 373, 635–640.

Guranowski, A., Starzynska, E., Pietrowska-Borek, M., Rejman, D., Blackburn, G.M., 2009. Novel diadenosine polyphosphate analogs with oxymethylene bridges replacing oxygen in the polyphosphate chain: potential substrates and/or inhibitors of Ap4A hydrolases. FEBS J. 276, 1546–1553.

Haltiwanger, B.M., Karpinich, N.O., Taraschi, T.F., 2000. Characterization of class II apurinic/apyrimidinic endonuclease activities in the human malaria parasite, *Plasmodium falciparum*. Biochem. J. 345 (Pt 1), 85–89.

Hannay, K., Marcotte, E.M., Vogel, C., 2008. Buffering by gene duplicates: an analysis of molecular correlates and evolutionary conservation. BMC Genomics 9, 609.

Hayward, R.E., 2000. *Plasmodium falciparum* phosphoenolpyruvate carboxykinase is developmentally regulated in gametocytes. Mol. Biochem. Parasitol. 107, 227–240.

Holman, A.G., Davis, P.J., Foster, J.M., Carlow, C.K., Kumar, S., 2009. Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of Brugia malayi. BMC Microbiol. 9, 243.

Hopkins, A.L., Groom, C.R., 2002. The druggable genome. Nat. Rev. Drug Discov. 1, 727–730.

Hopkins, A.L., Bickerton, G.R., Carruthers, I.M., Boyer, S.K., Rubin, H., Overington, J.P., 2011. Rapid analysis of pharmacology for infectious diseases. Curr. Top. Med. Chem. 11, 1292–1300.

Huthmacher, C., Hoppe, A., Bulik, S., Holzhutter, H.G., 2010. Antimalarial drug targets in *Plasmodium falciparum* predicted by stage-specific metabolic network analysis. BMC Syst. Biol. 4, 120.

Janse, C.J., Kroeze, H., van Wigcheren, A., Mededovic, S., Fonager, J., Franke-Fayard, B., Waters, A.P., Khan, S.M., 2011. A genotype and phenotype database of genetically modified malaria-parasites. Trends Parasitol. 27, 31–39.

Jenkins, C.L., 1989. Effects of the phosphoenolpyruvate carboxylase inhibitor 3,3-dichloro-2-(dihydroxyphosphinoylmethyl)propenoate on photosynthesis: C(4) selectivity and studies on C(4) photosynthesis. Plant Physiol. 89, 1231–1237.

Joubert, F., Harrison, C.M., Koegelenberg, R.J., Odendaal, C.J., de Beer, T.A., 2009. Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. Malar. J. 8, 178.

Joyce, A.R., Palsson, B.O., 2008. Predicting gene essentiality using genome-scale in silico models. Methods Mol. Biol. 416, 433–457.

Kato, N., Sakata, T., Breton, G., Le Roch, K.G., Nagle, A., Andersen, C., Bursulaya, B., Henson, K., Johnson, J., Kumar, K.A., Marr, F., Mason, D., McNamara, C., Plouffe, D., Ramachandran, V., Spooner, M., Tuntland, T., Zhou, Y., Peters, E.C., Chatterjee, A., Schultz, P.G., Ward, G.E., Gray, N., Harper, J., Winzeler, E.A., 2008. Gene expression signatures and small-molecule compounds link a protein kinase to *Plasmodium falciparum* motility. Nat. Chem. Biol. 4, 347–356.

Kobayashi, T., Takematsu, H., Yamaji, T., Hiramoto, S., Kozutsumi, Y., 2005. Disturbance of sphingolipid biosynthesis abrogates the signaling of Mss4, phosphatidylinositol-4-phosphate 5-kinase, in yeast. J. Biol. Chem. 280, 18087–18094.

Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., Winzeler, E.A., 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301, 1503–1508.

Leber, W., Skippen, A., Fivelman, Q.L., Bowyer, P.W., Cockcroft, S., Baker, D.A., 2009. A unique phosphatidylinositol 4-phosphate 5-kinase is activated by ADP-ribosylation factor in *Plasmodium falciparum*. Int. J. Parasitol. 39, 645–653.

Levillain, O., Diaz, J.J., Reymond, I., Soulet, D., 2000. Ornithine metabolism along the female mouse nephron: localization of ornithine decarboxylase and ornithine aminotransferase. Pflugers Arch. 440, 761–769.

Li, M., Allen, A., Smith, T.J., 2007. High throughput screening reveals several new classes of glutamate dehydrogenase inhibitors. Biochemistry 46, 15089–15102.

Li, M., Smith, C.J., Walker, M.T., Smith, T.J., 2009. Novel inhibitors complexed with glutamate dehydrogenase: allosteric regulation by control of protein dynamics. J. Biol. Chem. 284, 22988–23000.

Lim, D.C., Cooke, B.M., Doerig, C., Saeij, J.P., 2012. Toxoplasma and Plasmodium protein kinases: roles in invasion and host cell remodelling. Int. J. Parasitol. 42, 21–32.

Llinas, M., DeRisi, J.L., 2004. Pernicious plans revealed: *Plasmodium falciparum* genome wide expression analysis. Curr. Opin. Microbiol. 7, 382–387.

Madhusudana, Rao.I., Swamy, P.M., Das, V.S.R., 1980. Herbricide (sic) inhibited phosphoenolpyruvate carboxylase in leaves of six nonsucculent scrub species. Z. Pflanzenphysiol 99, 69–74.

Magarinos, M.P., Carmona, S.J., Crowther, G.J., Ralph, S.A., Roos, D.S., Shanmugam, D., Van Voorhis, W.C., Aguero, F., 2012. TDR Targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res. 40, D1118–1127.

Magrane, M., Consortium, U., 2011. UniProt Knowledgebase: A Hub of Integrated Protein Data. Database, Oxford, bar009.

Martinsen, E.S., Perkins, S.L., Schall, J.J., 2008. A three-genome phylogeny of malaria parasites (Plasmodium and closely related genera): evolution of life-history traits and host switches. Mol. Phylogenet. Evol. 47, 261–273.

Mastrolorenzo, A., Scozzafava, A., Supuran, C.T., 2000. Antifungal activity of Ag(I) and Zn(II) complexes of aminobenzolamide (5-sulfanilylamido-1,3,4-thiadiazole-2-sulfonamide) derivatives. J. Enzym. Inhib. 15, 517–531.

McDaniel, H.G., Siu, P.M., 1972. Purification and characterization of phosphoenolpyruvate carboxylase from Plasmodium berghei. J. Bacteriol. 109, 385–390.

Mulder, N.J., Kersey, P., Pruess, M., Apweiler, R., 2008. In silico characterization of proteins: UniProt, InterPro and Integr8. Mol. Biotechnol. 38, 165–177.

Muller, I.B., Hyde, J.E., 2010. Antimalarial drugs: modes of action and mechanisms of parasite resistance. Future Microbiol. 5, 1857–1873.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Ojo, K.K., Larson, E.T., Keyloun, K.R., Castaneda, L.J., Derocher, A.E., Inampudi, K.K., Kim, J.E., Arakaki, T.L., Murphy, R.C., Zhang, L., Napuli, A.J., Maly, D.J., Verlinde, C.L., Buckner, F.S., Parsons, M., Hol, W.G., Merritt, E.A., Van Voorhis, W.C., 2010. *Toxoplasma gondii* calcium-dependent protein kinase 1 is a target for selective kinase inhibitors. Nat. Struct. Mol. Biol. 17, 602–607.

Onyango, D.O., Naguleswaran, A., Delaplane, S., Reed, A., Kelley, M.R., Georgiadis, M.M., Sullivan Jr., W.J., 2011. Base excision repair apurinic/apyrimidinic endonucleases in apicomplexan parasite *Toxoplasma gondii*. DNA Repair (Amst.) 10, 466–475.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L., 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38, D196–203.

Pairoba, C.F., Colombo, S.L., Andreo, C.S., 1996. Flavonoids as inhibitors of NADP-malic enzyme and PEP carboxylase from C4 plants. Biosci. Biotechnol. Biochem. 60, 779–783.

Payne, D.J., Holmes, D.J., Rosenberg, M., 2001. Delivering novel targets and antibiotics from genomics. Curr. Opin. Investig. Drugs 2, 1028–1034.

Plata, G., Hsiao, T.L., Olszewski, K.L., Llinas, M., Vitkup, D., 2010. Reconstruction and flux-balance analysis of the *Plasmodium falciparum* metabolic network. Mol. Syst. Biol. 6, 408.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.

Riguet, E., Desire, J., Boden, O., Ludwig, V., Gobel, M., Bailly, C., Decout, J.L., 2005. Neamine dimers targeting the HIV-1 TAR RNA. Bioorg. Med. Chem. Lett. 15, 4651–4655.

Rodriguez-Acosta, A., de Dominguez, N., Aguilar, I., Giron, M.E., 1999. Detection of glutamate dehydrogenase enzyme activity in *Plasmodium falciparum* infection. Indian J. Med. Res. 109, 152–156.

Rohrich, R.C., Englert, N., Troschke, K., Reichenberg, A., Hintz, M., Seeber, F., Balconi, E., Aliverti, A., Zanetti, G., Kohler, U., Pfeiffer, M., Beck, E., Jomaa, H., Wiesner, J., 2005. Reconstitution of an apicoplast-localised electron transfer pathway involved in the isoprenoid biosynthesis of *Plasmodium falciparum*. FEBS Lett. 579, 6433–6438.

Rottmann, M., McNamara, C., Yeung, B.K., Lee, M.C., Zou, B., Russell, B., Seitz, P., Plouffe, D.M., Dharia, N.V., Tan, J., Cohen, S.B., Spencer, K.R., Gonzalez-Paez, G.E., Lakshminarayana, S.B., Goh, A., Suwanarusk, R., Jegla, T., Schmitt, E.K., Beck, H.P., Brun, R., Nosten, F., Renia, L., Dartois, V., Keller, T.H., Fidock, D.A., Winzeler, E.A., Diagana, T.T., 2010. Spiroindolones, a potent compound class for the treatment of malaria. Science 329, 1175–1180.

Sa, J.M., Chong, J.L., Wellems, T.E., 2011. Malaria drug resistance: new observations and developments. Essays Biochem. 51, 137–160.

Salvati, L., Mattu, M., Tiberi, F., Polticelli, F., Ascenzi, P., 2001. Inhibition of *Saccharomyces cerevisiae* phosphomannose isomerase by the NO-donor S-nitroso-acetyl-penicillamine. J. Enzym. Inhib. 16, 287–292.

Santiago, T.C., Zufferey, R., Mehra, R.S., Coleman, R.A., Mamoun, C.B., 2004. The *Plasmodium falciparum* PfGatp is an endoplasmic reticulum membrane protein important for the initial step of malarial glycerolipid synthesis. J. Biol. Chem. 279, 9222–9232.

Satchell, J.F., Malby, R.L., Luo, C.S., Adisa, A., Alpyurek, A.E., Klonis, N., Smith, B.J., Tilley, L., Colman, P.M., 2005. Structure of glyceraldehyde-3-phosphate dehydrogenase from *Plasmodium falciparum*. Acta Crystallogr. D: Biol. Crystallogr. 61, 1213–1221.

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Sohngen, C., Stelzer, M., Thiele, J., Schomburg, D., 2011. BRENDA, the enzyme information system in 2011. Nucleic Acids Res. 39, D670–676.

Schindelman, G., Fernandes, J.S., Bastiani, C.A., Yook, K., Sternberg, P.W., 2011. Worm Phenotype Ontology: integrating phenotype data within and beyond the *C. elegans* community. BMC Bioinform. 12, 32.

Sriwilaijareon, N., Petmitr, S., Mutirangura, A., Ponglikitmongkol, M., Wilairat, P., 2002. Stage specificity of *Plasmodium falciparum* telomerase and its inhibition by berberine. Parasitol. Int. 51, 99–103.

Storm, J., Perner, J., Aparicio, I., Patzewitz, E.M., Olszewski, K., Llinas, M., Engel, P.C., Muller, S., 2011. *Plasmodium falciparum* glutamate dehydrogenase a is dispensable and not a drug target during erythrocytic development. Malar. J. 10, 193.

Stranska, J., Kopecny, D., Tylichova, M., Snegaroff, J., Sebela, M., 2008. Ornithine delta-aminotransferase: an enzyme implicated in salt tolerance in higher plants. Plant Signal. Behav. 3, 929–935.

Strecker, H.J., 1965. Purification and properties of rat liver ornithine delta-transaminase. J. Biol. Chem. 240, 1225–1230.

Takatsuka, Y., Onoda, M., Sugiyama, T., Muramoto, K., Tomita, T., Kamio, Y., 1999. Novel characteristics of Selenomonas ruminantium lysine decarboxylase capable of decarboxylating both L-lysine and L-ornithine. Biosci. Biotechnol. Biochem. 63, 1063–1069.

Teixeira, C., Gomes, J.R., Gomes, P., 2011. Falcipains, *Plasmodium falciparum* cysteine proteases as key drug targets against malaria. Curr. Med. Chem. 18, 1555–1572.

Tell, G., Fantini, D., Quadrifoglio, F., 2010. Understanding different functions of mammalian AP endonuclease (APE1) as a promising tool for cancer treatment. Cell. Mol. Life Sci. 67, 3589–3608.

Tewari, R., Straschil, U., Bateman, A., Bohme, U., Cherevach, I., Gong, P., Pain, A., Billker, O., 2010. The systematic functional analysis of Plasmodium protein kinases identifies essential regulators of mosquito transmission. Cell Host Microbe 8, 377–387.

Triglia, T., Wang, P., Sims, P.F.G., Hyde, J., Cowman, A., 1998. Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. EMBO J. 17, 3807–3815.

van Dooren, G.G., Su, V., D'Ombrain, M.C., McFadden, G.I., 2002. Processing of an apicoplast leader sequence in *Plasmodium falciparum* and the identification of a putative leader cleavage enzyme. J. Biol. Chem. 277, 23612–23619.

Vinayak, S., Sharma, Y.D., 2007. Inhibition of *Plasmodium falciparum* ispH (lytB) gene expression by hammerhead ribozyme. Oligonucleotides 17, 189–200.

Wang, G., Shang, L., Burgett, A.W., Harran, P.G., Wang, X., 2007. Diazonamide toxins reveal an unexpected function for ornithine delta-amino transferase in mitotic cell division. Proc. Natl. Acad. Sci. USA 104, 2068–2073.

Wang, P., Wang, Q., Aspinall, T.V., Sims, P.F., Hyde, J.E., 2004. Transfection studies to explore essential folate metabolism and antifolate drug synergy in the human malaria parasite *Plasmodium falciparum*. Mol. Microbiol. 51, 1425–1438.

Wang, P., Wang, Q., Yang, Y., Coward, J.K., Nzila, A., Sims, P.F., Hyde, J.E., 2010. Characterisation of the bifunctional dihydrofolate synthase-folylpolyglutamate synthase from *Plasmodium falciparum*; a potential novel target for antimalarial antifolate inhibition. Mol. Biochem. Parasitol. 172, 41–51.

Waterhouse, R.M., Zdobnov, E.M., Kriventseva, E.V., 2010. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. Genome Biol. Evol. 3, 75–86.

Wells, T.N., Scully, P., Paravicini, G., Proudfoot, A.E., Payton, M.A., 1995. Mechanism of irreversible inactivation of phosphomannose isomerases by silver ions and flamazine. Biochemistry 34, 7896–7903.

Wills, E.A., Roberts, I.S., Del Poeta, M., Rivera, J., Casadevall, A., Cox, G.M., Perfect, J.R., 2001. Identification and characterization of the *Cryptococcus neoformans* phosphomannose isomerase-encoding gene, MAN1, and its impact on pathogenicity. Mol. Microbiol. 40, 610–620.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D.J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J.L., Riles, L., Roberts, C.J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R.K., Veronneau, S., Voet, M., Volckaert, G., Ward, T.R., Wysocki, R., Yen, G.S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., Davis, R.W., 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 285, 901–906.

Wydysh, E.A., Medghalchi, S.M., Vadlamudi, A., Townsend, C.A., 2009. Design and synthesis of small molecule glycerol 3-phosphate acyltransferase inhibitors. J. Med. Chem. 52, 3317–3327.

Yamamoto, S., Imamura, T., Kusaba, K., Shinoda, S., 1991. Purification and some properties of inducible lysine decarboxylase from *Vibrio parahaemolyticus*. Chem. Pharm. Bull. (Tokyo) 39, 3067–3070.

Yamashita, S., Numa, S., 1981. Glycerophosphate acyltransferase from rat liver. Methods Enzymol. 71 (Pt C), pp. 550–554.

Yeh, I., Hanekamp, T., Tsoka, S., Karp, P.D., Altman, R.B., 2004. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. Genome Res. 14, 917–924.

Zawahir, Z., Dayam, R., Deng, J., Pereira, C., Neamati, N., 2009. Pharmacophore guided discovery of small-molecule human apurinic/apyrimidinic endonuclease 1 inhibitors. J. Med. Chem. 52, 20–32.

Zhang, K., Rathod, P.K., 2002. Divergent regulation of dihydrofolate reductase between malaria parasite and human host. Science 296, 545–547.

# *Chapter 5*

## The Genome of the Heartworm, *Dirofilaria immitis*, Reveals Drug and Vaccine Targets

The *Dirofilaria immitis* Genome Consortium

*The FASEB Journal* (2012)

*The FASEB Journal* • Research Communication

# The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets

**Christelle Godel,*,†,‡,1 Sujai Kumar,§,1 Georgios Koutsovoulos,§,2 Philipp Ludin,*,†,2 Daniel Nilsson,¶ Francesco Comandatore,# Nicola Wrobel,‖ Marian Thompson,‖ Christoph D. Schmid,*,† Susumu Goto,** Frédéric Bringaud,†† Adrian Wolstenholme,‡‡ Claudio Bandi,# Christian Epe,‡ Ronald Kaminsky,‡ Mark Blaxter,§,‖ and Pascal Mäser*,†,3**

*Swiss Tropical and Public Health Institute, Basel, Switzerland; †University of Basel, Basel, Switzerland; ‡Novartis Animal Health, Centre de Recherche Santé Animale, St. Aubin, Switzerland; §Institute of Evolutionary Biology and ‖The GenePool Genomics Facility, School of Biological Sciences, University of Edinburgh, Edinburgh, UK; ¶Department of Molecular Medicine and Surgery, Science for Life Laboratory, Karolinska Institutet, Solna, Sweden; #Dipartimento di Scienze Veterinarie e Sanità Pubblica, Università degli studi di Milano, Milan, Italy; **Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, Japan; ††Centre de Résonance Magnétique des Systèmes Biologiques, Unité Mixte de Recherche 5536, University Bordeaux Segalen, Centre National de la Recherche Scientifique, Bordeaux, France; and ‡‡Department of Infectious Diseases and Center for Tropical and Emerging Global Disease, University of Georgia, Athens, Georgia, USA

ABSTRACT **The heartworm *Dirofilaria immitis* is an important parasite of dogs. Transmitted by mosquitoes in warmer climatic zones, it is spreading across southern Europe and the Americas at an alarming pace. There is no vaccine, and chemotherapy is prone to complications. To learn more about this parasite, we have sequenced the genomes of *D. immitis* and its endosymbiont *Wolbachia*. We predict 10,179 protein coding genes in the 84.2 Mb of the nuclear genome, and 823 genes in the 0.9-Mb *Wolbachia* genome. The *D. immitis* genome harbors neither DNA transposons nor active retrotransposons, and there is very little genetic variation between two sequenced isolates from Europe and the United States. The differential presence of anabolic pathways such as heme and nucleotide biosynthesis hints at the intricate metabolic interrelationship between the heartworm and *Wolbachia*. Comparing the proteome of *D. immitis* with other nematodes and with mammalian hosts, we identify families of potential drug targets, immune modulators, and vaccine candidates. This genome sequence will support the development of new tools against dirofilariasis and aid efforts to combat related human pathogens, the causative agents of lymphatic filariasis and river blindness.—Godel, C., Kumar, S., Koutsovoulos, G., Ludin, P., Nilsson, D., Comandatore, F., Wrobel, N., Thompson, M., Schmid, C. D., Goto, S., Bringaud, F., Wolstenholme, A., Bandi, C., Epe, C., Kaminsky, R., Blaxter, M., Mäser, P. The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets. *FASEB J.* 26, 4650–4661 (2012). www.fasebj.org**

*Key Words: comparative genomics · filaria · transposon* · Wolbachia

THE HEARTWORM *DIROFILARIA IMMITIS* (Leidy, 1856) is a parasitic nematode of mammals. The definitive host is the dog; however, it also infects cats, foxes, coyotes, and, very rarely, humans (1). Dirofilariasis of dogs is a severe and potentially fatal disease. Adult nematodes of 20 to 30 cm reside in the pulmonary arteries, and the initial damage is to the lung. The spectrum of subsequent pathologies related to chronic heartworm infection is broad, the most serious manifestation being heart failure. Recent rapid spread of *D. immitis* through the United States and southern Europe (2, 3) is being favored by multiple factors. Global warming is expand-

Abbreviations: EST, expressed sequence tag; HMM, hidden Markov model; HSP, high-scoring pair; LTR, long terminal repeat; RNA-Seq, transcriptome shotgun sequencing; SOCS, suppressor of cytokine signaling; wBm *Wolbachia* endosymbiont of *Brugia malayi;* wDi, *Wolbachia* endosymbiont of *Dirofilaria immitis*

[1] These authors contributed equally to this work.

[2] These authors contributed equally to this work.

[3] Correspondence: Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland. E-mail: pascal.maeser@unibas.ch

doi: 10.1096/fj.12-205096

This article includes supplemental data. Please visit *http://www.fasebj.org* to obtain this information.

ing the activity season of vector mosquitoes, increasing their abundance and the likelihood of transmission of the parasite, and there are growing numbers of pets, reservoir animals, and "traveling" dogs (2, 3).

*D. immitis* is an onchocercid filarial nematode, related to important parasites of humans, such as *Onchocerca volvulus*, the agent of river blindness. The *D. immitis* lifecycle is typical for Onchocercidae. Microfilariae, shed into the bloodstream by adult females, are ingested by a mosquito (various species, including *Aedes*, *Anopheles*, and *Culex* spp.) where they develop into third-stage larvae (L3) and migrate to the labium. Feeding by an infected mosquito introduces L3 into the skin. The prepatent period in the newly bitten dog is 6–9 mo, during which the injected larvae undergo two further molts and migrate *via* muscle fibers to the pulmonary vasculature, where the adult nematodes develop. At present, diagnosis is effective only for patent infections, because it is based on detection of circulating microfilariae or antigens from mature females. Treatment of dirofilariasis is also problematic, because the arsenical melarsomine dihydrochloride, the only adulticide approved by the U.S. Food and Drug Administration, can cause adverse neurological reactions. Treatment carries a significant risk of lethality due to blockage of the pulmonary artery by dead nematodes. No vaccine is available. These issues, together with the alarming increasing spread of *D. immitis*, prompted the American Heartworm Society to recommend year-round chemoprophylactic treatment of dogs (4) to kill the larval stages before they develop into adults. This requires monthly administration of anthelmintics, predominantly macrocyclic lactones, such as ivermectin, milbemycin, or moxidectin.

Human-infective parasites related to *D. immitis* cause subcutaneous filariasis and river blindness and are endemic in tropical and subtropical regions around the globe, with an estimated 380 million people affected (5). Improved diagnostics, new drugs, and, ultimately, effective vaccines are sorely needed. The sequencing of the *Brugia malayi* genome provides a platform for rational drug design, but by itself this single sequence cannot distinguish between idiosyncratic and shared targets that could be exploited for control (6).

Most of the filarial nematodes that cause diseases in humans and animals, including *D. immitis*, *O. volvulus*, *Wuchereria bancrofti*, and *B. malayi*, have been shown to harbor intracellular symbiotic bacteria of the genus *Wolbachia* (*e.g.*, refs. 7–10). These bacteria are vertically transmitted to the nematode progeny, *via* transovarial transmission. In most of the infected nematode species, all individuals are infected (reviewed in ref. 11). Even though the exact role of *Wolbachia* in filarial biology has not yet been determined, these bacteria are thought to be beneficial to the nematode host. Indeed, antibiotics that target *Wolbachia* have been shown to have deleterious effects on filarial nematodes, blocking reproduction, inducing developmental arrest, and killing adult nematodes (*e.g.*, refs. 7–9). This has led to development of research projects with the aim of developing anti-*Wolbachia* chemotherapy as a novel strategy for the control of filarial diseases. *Wolbachia* has also been implicated in the immunopathogenesis of filarial diseases, with a role in the development of pathological outcomes, such as inflammation and clouding of the cornea that is typical of river blindness (12). The genome of *Wolbachia* is thus an additional source of potential drug targets (7–10), but a single genome cannot reveal shared *vs.* unique biochemical weaknesses.

The human pathogenic Onchocercidae do not represent an attractive market for the pharmacological industry, because projected incomes from impoverished communities in developing endemic nations would be unlikely to cover the costs of drug development. The heartworm may hold a possible solution to this problem, because the market potential for novel canine anthelmintics is big, given the costs for heartworm prevention of $75–100/dog/yr and the estimated number of 80 million dogs in the United States (13). Choosing drug targets that are likely to be conserved in related, human pathogenic species may benefit both canine and human medicine. Here we present the draft genome sequences of *D. immitis* and its *Wolbachia* endosymbiont (wDi) and use these data to investigate the relationship between nematode and endosymbiont and identify new drug and vaccine targets.

## MATERIALS AND METHODS

### *D. immitis* isolates and DNA sequencing

We sequenced two canine *D. immitis* isolates, one from Pavia, Italy, and the other from Athens, Georgia, USA. The Pavia isolate was established in a laboratory lifecycle after primary isolation from an infected dog. Adult Pavia nematodes used for DNA extraction were recovered after necropsy of dogs infected as a control group in ongoing investigations (permit FR401e/08 from the Veterinary Office Canton de Fribourg, Switzerland). The Athens nematodes used for DNA and RNA extraction were from a naturally infected dog necropsied as part of routine clinical surveillance and were not from an established strain. Genomic DNA was extracted (QIAamp DNA extraction kit; Qiagen, Valencia, CA, USA) from individual adult female nematodes from Pavia and Athens isolates, and RNA was extracted (RNeasy kit; Qiagen) from individual female and male nematodes from Athens. Whole-genome shotgun sequences were generated at The GenePool Genomics Facility (University of Edinburgh, Edinburgh, UK) and at Fasteris SA (Geneva, Switzerland) using Illumina GAIIx and HiSeq2000 instruments (Illumina, Inc., San Diego, CA, USA). Several short insert (100- to 400-bp) paired-end amplicon libraries and long insert (3- to 4-kb) mate-pair amplicon libraries were made, and data from four of these were used in the final assembly (details are given on the Web site http://www.dirofilaria.org). These yielded a raw data total of 28 Gb in 295 million reads [European Bioinformatics Institute (EBI) Short Read Archive, accession number ERA032353; http://www.ebi.ac.uk/ena/]. After trimming low-quality bases (Phred score <20) and filtering out reads with uncalled bases or length <35 b, 271 million reads were used for assembly (Supplemental Table S1).

### Nuclear, mitochondrial, and *Wolbachia* genome assemblies

The short-read data were assembled using ABySS 1.2.3 (14). A number of test assemblies were performed using other assemblers, and a range of parameters was tested within ABySS, and the final, optimal assembly was performed using a k-mer length of 35 and scaffolding with the paired-end data only. Assembly qualities were assessed using summary statistics including maximizing the N50 (the contig length at which 50% of the assembly span was in contigs of that length or greater), maximum contig length, and total number of bases in contigs (see Supplemental Table S1) and using biological optimality assessment, such as maximizing the coverage of published *D. immitis* expressed sequence tag (EST) sequences and maximizing the number of *B. malayi* genes matched and the completeness of representation of core eukaryotic genes (using CEGMA; ref. 15). Redundancy due to allelic polymorphism was reduced with CD-HIT-EST (16), merging contigs that were ≥97% identical over the full length of the shorter contig. The mitochondrial genome was assembled by mapping the reads to the published *D. immitis* mitochondrial genome (17) and predicting a consensus sequence of the mitochondrial genomes of the Athens and Pavia nematodes separately. The wDi genome was assembled by first identifying likely wDi contigs in the whole assembly with BLASTn (18) using all *Wolbachia* genomes from EMBL-Bank, and then collecting all raw reads (and their pairs; $n$=6,912,659) that mapped to these putative wDi genome fragments. The reduced set of likely wDi reads was then assembled using an independently optimized ABySS parameter set, using mate-pair information where available. Mitochondrial and wDi contigs were removed from the full assembly to leave the final nuclear assembly.

### Transcriptome shotgun sequencing (RNA-Seq) assembly

The preparation of amplicon libraries and RNA-Seq analysis were performed following standard Illumina TruSeq protocols. A total of 11,019,886 (male) and 21,643,293 (female) read pairs of length 54 b were produced on the Illumina GAIIx platform (ArrayExpress accession number E-MTAB-714; ENA study accession number ERP000758). After quality filtering, the remaining 31,396,183 pairs were assembled with Trans-ABySS using k-mer values from 23 to 47 in steps of 4 (Supplemental Table S1).

### *D. immitis* nuclear genome protein-coding gene prediction and analysis

Repeats in the *D. immitis* genomic assembly were identified and masked using RepeatMasker 3.2.9 (19), including all "Nematoda" repeats in the RepBase libraries (20). The MAKER 2.08 annotation pipeline (21) was used to identify protein-coding genes based on evidence from the RNA-Seq assembly, alignments to the *B. malayi* proteome (WormBase release WS220; http://wormbase.sanger.ac.uk/), predictions made by the *ab initio* gene finder SNAP (22), and predictions from the *ab initio* gene finder Augustus (23) based on the Augustus hidden Markov model (HMM) profiles for *B. malayi*. MAKER predicted 11,895 gene models, and, with alternative splicing, a total of 12,872 transcripts and peptides. We compared the nuclear proteome of *D. immitis* with those of four other species for which complete genome data are available and which span the phylogenetic diversity of the phylum Nematoda (*B. malayi, Ascaris suum, Caenorhabditis elegans,* and *Trichinella spiralis*). The complete proteomes were compared using all-against-all BLAST, and then clustered using OrthoMCL (24). OrthoMCL clusters were postprocessed to classify clusters by their species content and analyzed with reference to the robust molecular phylogeny of the Nematoda (25). The prediction of *D. immitis* orthologs from *B. malayi*, *C. elegans*, *Homo sapiens,* and *Canis lupus* to identify drug targets was performed with InParanoid (26).

### Analyses of orthology and divergence in filarial *Wolbachia*

The wDi genome was annotated with the RAST server (27), an online resource that uses best-practice algorithms to perform both gene finding and gene functional annotation. Selected metabolic pathways were annotated based on enzyme lists from the KEGG Pathway database (28), after an HMM profile was generated for each enzyme (29) from a ClustalW (30) multiple alignment of a redundancy-reduced set of all the manually curated entries in UniProt (31). Analysis of orthology was performed using the BLAST reciprocal best-hits algorithm (32), with the following cutoff values: $E$ value 0.1 and ID percentage 60%. Protein distance for each pair of orthologs was calculated using Protdist in Phylip 3.69 (33) with the Dayhoff PAM matrix option. Proteins were allocated to functional categories using BLAST against the COG database. Protein distances were then analyzed based on COG categories: within each category we calculated the average distances of protein pairs. To evaluate whether some categories were significantly more variable than others, we performed the Kruskal-Wallis test on COG categories containing more than one ortholog pair. The pairwise Mann-Whitney test was then performed to detect pairs of COG categories that displayed significant differences in their average variation.

### Identification of *Wolbachia* insertions in nematode genomes

To identify potential lateral genetic transfers from *Wolbachia* to the host nuclear genome, the nuclear genome was queried against the 921-kbp wDi genome using the dc-megablast option in BLASTN (NCBI-blast+2.2.25) with default settings. All high-scoring pairs (HSPs) longer than 100 bp with >80% identity were kept. Overlapping HSP coordinates on the nuclear genome were merged, and sequences from these coordinates were extracted to obtain putative nuclear *Wolbachia* DNA elements. The *B. malayi* nuclear genome was screened with the *B. malayi Wolbachia* (wBm) genome in the same way. The small numbers of *Wolbachia* insertions identified in the nuclear genomes of *Acanthocheilonema viteae* and *Onchocerca flexuosa* (34) were surveyed for matches to the wDi and wBm genomes and cross-compared with the insertion sets from the complete *D. immitis* and *B. malayi* genomes using reciprocal best BLAST searches and filtering alignments shorter than 100 bp. Reciprocal best BLAST matches were isolated and single-linkage clustered.

## RESULTS

### Genome assembly of *D. immitis* and its *Wolbachia* symbiont

Genome sequence was generated from single individuals of *D. immitis* isolated from naturally infected dogs, one from Athens, Georgia (USA) and the other from Pavia (Italy). A total of 16 Gb of raw data was retained after rigorous quality checks, corresponding to ~170-fold coverage of the *D. immitis* nuclear genome (likely to be ~95 Mb, similar to related Onchocercidae). The ABySS (35) assembler performed best based on statistical and biological measures (Supplemental Table S1). The mitochondrial and *Wolbachia* wDi genomes were assembled independently. The final nuclear assembly contained 84.2 Mb of sequence in 31,291 scaffolds with an N50 of 10,584 bases (**Table 1**). The draft genome of wDi consists of 2 scaffolds spanning 0.92 Mb. We identified 99% of previously deposited genome survey sequences putatively from wDi (GenBank accession numbers ET041559 to ET041665) within our wDi assembly. The wDi genome was 16% smaller than that of wBm (1.08 Mb; GenBank accession number NC_006833), and there was significant breakage of synteny between the two genomes, as has been observed between other *Wolbachia*.

The *D. immitis* and wDi genomes, the annotations we have made on these, and additional technical details and analyses are available through a dedicated genome browser (http://www.dirofilaria.org).

### Lack of genetic diversity between the sequenced *D. immitis* isolates

Even though the two sequenced *D. immitis* came from independent isolates from different continents, they showed low genetic differentiation, allowing the raw sequencing data from both nematodes to be coassembled. We mapped the reads from each nematode back to the draft assembly and identified only 32,729 high-quality single-nucleotide variations, a very low per-nucleotide diversity rate of 0.04%. We identified the sequences corresponding to 11 polymorphic microsatellite loci used previously to analyze the *D. immitis* population structure in North America (36) and genotyped our two isolates *in silico* by counting the predicted numbers of microsatellite repeats at each locus. Both our nematodes could be classified within the diversity of the eastern United States population. The mitochondrial genomes of the two isolates differed at only 6 sites (and were thus >99.9% identical). Surprisingly, compared with the published, Australian *D. immitis* mitochondrion (17), both had many shared differences (each was only 99.5% identical to the published *D. immitis* mitochondrion). Because the ~70 differences were often clumped and were unique in the published *D. immitis* mitochondrial genome compared both with our two genomes and with the genomes of five other filarial nematodes, we suggest that many of these are sequencing errors in the published genome.

TABLE 1. *Comparison of the genome assemblies of* D. immitis*,* B. malayi*, and* C. elegans

| Characteristic | *D. immitis* | *B. malayi* | *C. elegans* |
|---|---|---|---|
| Assembly size (Mb) | 84.2 | 93.6[a] | 100.3 |
| Protein-coding gene models | 11,375 | 11,434 | 20,517 |
| Genes per megabase | 135 | 122 | 205 |
| Predicted proteins | 12,344 | 11,460 | 31,249 |
| Protein-coding sequence (%) | 18.0 | 13.8 | 25.4 |
| Median exons per gene | 5 | 5 | 6 |
| Median exon size (b) | 142 | 139 | 147 |
| Median intron size (b) | 226 | 213 | 73 |
| Overall GC content (%) | 28.3 | 30.2 | 35.4 |
| Exon GC content (%) | 37.4 | 39.4 | 43.4 |
| Intron GC content (%) | 26.6 | 27.2 | 32.5 |

*B. malayi* data are from the GenBank RefSeq dataset; *C. elegans* data from the WS230 dataset. [a]70.8 Mb scaffolds + 17.5 Mb short contigs.

### A metazoan genome without active transposable elements

The *D. immitis* genome was surveyed for the three main classes of transposable elements [DNA transposons, long terminal repeat (LTR) retrotransposons, and non-LTR retrotransposons] with tBLASTn (18) using the transposon-encoded proteins as queries. No traces of active or pseudogenized DNA transposons or non-LTR retrotransposons were found, but 376 fragments of LTR retrotransposons of the BEL/Pao family (37) were identified. None of these fragments were predicted to be functional, because all contained frame shifts and stop codons in the likely coding sequence. The *D. immitis* Pao pseudogenes were most similar to Pao family retrotransposons from *B. malayi* (6). In *B. malayi*, several of the Pao retrotransposons are likely to be active, because they have complete open reading frames and LTRs. Overall, however, *B. malayi* has a lower density of Pao elements and fragments (3.4 Pao/Mb, 8.3% of which are predicted to be functionally intact) compared with *D. immitis* (4.6 Pao/Mb, none of which were intact).

### *D. immitis* nuclear proteome

Protein coding genes were predicted in the nuclear assembly using the MAKER pipeline (21), integrating evidence-based (RNA-Seq and known protein mapping) and *ab initio* methods. Of the 11,375 gene models, 897 were predicted to generate alternate transcripts (Table 1). The total number of predicted proteins of length ≥100 aa was 10,179, similar to the 9807 predicted in *B. malayi.* Based on matches to *D. immitis* ESTs and core eukaryotic genes (15), the *D. immitis* proteome was likely to be near-complete. Protein-coding exons occupy ~18% of the genome of *D. immitis* and 14% of the genome of *B. malayi* (Table 1), but in *C. elegans* there are nearly twice as many genes, and exons cover ~30% of the genome. The median global identity between a *D. immitis* protein and its best match (as determined by BLASTp) in *B. malayi* was 75%.

*D. immitis* proteins were clustered with the complete proteomes of four other nematode species. These clusters were classified and mapped onto the phylogenetic tree of the five species based on the placement of the deepest node that linked the species that contributed members (**Fig. 1**). The *D. immitis* proteome included 3199 proteins (31% of the total proteome) that were unique to this species, a proportion similar to that found in *B. malayi* (27%), but many fewer (and a lower proportion) compared with those for the other species (for example, *C. elegans* had 63% of its proteome in species-unique

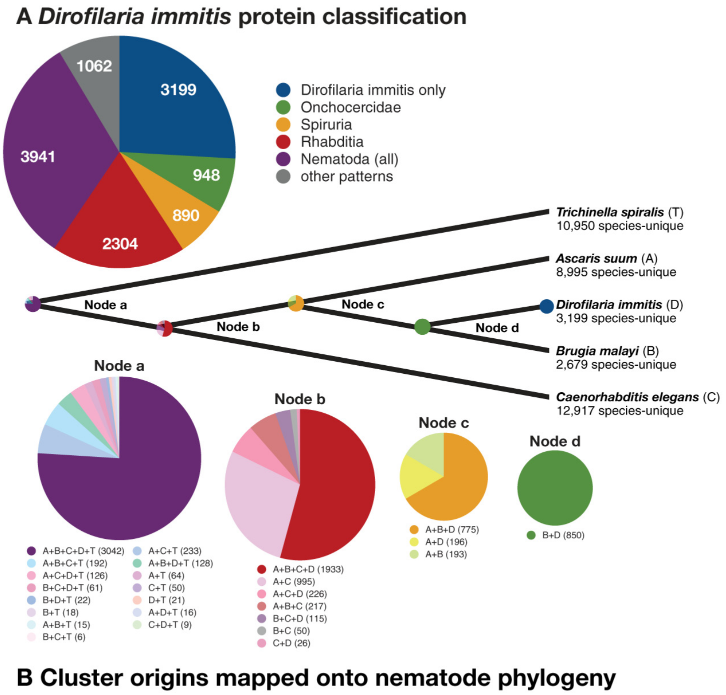


**Figure 1.** Conserved and novel genes in *D. immitis*. The *D. immitis* proteome was clustered with those of *B. malayi, A. suum, C. elegans,* and *T. spiralis*. Clusters were then classified based on the membership from the five species according to the current phylogeny of the phylum Nematoda. *A*) Pie chart showing the distribution of classification of *D. immitis* proteins: *D. immitis* only, singletons and clusters only found in *D. immitis*; Onchocercidae, clusters with members only from *D. immitis* and *B. malayi*; Spiruria, clusters with members only from Onchocercidae and *A. suum*; Rhabditia clusters with members only from Spiruria and *C. elegans*; Nematoda, clusters with members from all five species (*i.e.*, Rhabditia and *T. spiralis*); and other patterns, clusters with members not fitting simply into the phylogenetic schema (probably arising from gene loss, lack of predictions, or failure to cluster in one or more species). *B*) Cluster numbers and patterns of conservation mapped onto the phylogeny of the five species.

clusters). This difference may be partly due to the 850 proteins in clusters uniquely shared by the relatively closely related *D. immitis* and *B. malayi*, but these clusters only raise the proportion of proteins in phylogenetically local clusters to 47%.

### *D. immitis* genes homologous to known antinematode drug targets

An array of drugs are effective against nematode parasites (**Table 2**). Of these, flubendazole (38), mebendazole

TABLE 2. *Candidate drug targets, top-down search: current anthelmintics and their known targets in* C. elegans *and orthologs in* D. immitis

| Chemical class | Drug | Target | *C. elegans* | *D. immitis* |
|---|---|---|---|---|
| Benzimidazole | Albendazole<br>Flubendazole<br>Mebendazole | β-Tubulin | BEN-1 | DIMM36740 |
| Imidazothiazole | Levamisole | nACh receptor | LEV-1 | |
| | | | LEV-8 | |
| | | | UNC-29 | DIMM30000 |
| | | | UNC-38 | DIMM45965 |
| | | | UNC-63 | DIMM08405 |
| Macrocyclic lactone | Ivermectin<br>Milbemycin<br>Moxidectin<br>Selamectin | Glutamate receptor | AVR-14 | DIMM16610 |
| | | | AVR-15 | |
| | | | GLC-1 | |
| | | | GLC-2 | DIMM25280, DIMM21120 |
| | | | GLC-3 | |
| | | | GLC-4 | DIMM22030 |
| | | GABA receptor | EXP-1 | DIMM57890 |
| | | | GAB-1 | |
| | | | UNC-49 | DIMM33210 |
| Cyclodepsipeptide | Emodepside | $K^+$ channel | SLO-1 | DIMM33710 |
| | | Latrophilin GPCR | LAT-1 | DIMM37270, DIMM37275 |
| | | | LAT-2 | DIMM17690 |
| Aminoacetonitrile derivative | Monepantel | nACh receptor | ACR-23 | |
| | | | DES-2 | |

nAChR, nicotinic acetylcholine; GPCR, G protein-coupled receptor.

(39), levamisole (40), ivermectin, milbemycin, moxidectin, and selamectin (41, 42) have been demonstrated to be active against *D. immitis*. Many drug targets have been identified, particularly through forward genetics in the model nematode *C. elegans* (43) (Table 2). Prominent among these targets are neuronal membrane proteins, highlighting the importance of the neuromuscular junction as a hotspot of anthelmintic drug action. *D. immitis* appears to lack some known targets, notably members of the DEG-3 subfamily of acetylcholine receptors, which contains the presumed targets of monepantel (44). This contrasts with *B. malayi*, which possesses orthologs of DEG-3 and DES-2 (45). In *C. elegans*, the target space of levamisole and ivermectin comprises a large number of ligand-gated ion channels. Although these drugs are effective against heartworm, some of these ion channels do not have an ortholog in *D. immitis* (Table 2), indicating that those present are sufficient to confer drug susceptibility. The identified *D. immitis* orthologs of the known anthelmintic targets can now be monitored in suspected cases of drug resistance.

### New drug target candidates in *D. immitis*

New potential drug targets were identified *in silico* through an exclusion-inclusion strategy (46, 47). Starting from the complete set of predicted *D. immitis* proteins, we excluded proteins that had an ortholog in the dog or human proteome or had multiple paralogs in *D. immitis*. We included proteins that had a *C. elegans* ortholog essential for survival or development (based on RNAi phenotypes) and had predicted function as an enzyme or receptor. Among the 20 candidates identified (**Table 3**) were several proven drug targets, such as RNA-dependent RNA polymerase (antiviral), apurinic/apyrimidinic endonuclease and hedgehog proteins (anticancer; ref. 48), UDP-galactopyranose mutase (against mycobacteria, ref. 49; and kinetoplastids, ref. 50), sterol-C24-methyltransferase (antifungal; ref. 51), and the insecticide target chitin synthase (52). The *D. immitis* orthologs of these enzymes may serve as starting points for the development of new anthelmintics.

### Immune modulators and vaccine candidates

Filarial nematodes modulate the immune systems of their mammalian hosts to promote their own survival and fecundity, but the exact mechanisms used remain enigmatic. Proteases such as leucyl aminopeptidase and protease inhibitors such as serpins and cystatins have been implicated in disruption of immune signal processing (53), and we identified *D. immitis* leucyl amino-

TABLE 3. *Candidate drug targets, bottom-up search*

| *D. immitis* protein | Predicted function | *B. malayi* ortholog | *H. sapiens* log$_{10}$ (*E*) | *C. lupus* log$_{10}$ (*E*) | *C. elegans* RNAi |
|---|---|---|---|---|---|
| Nucleic acid synthesis and repair | | | | | |
| DIMM09370 | RNA-dependent RNA polymerase | BM06623 | 0.28 | 0.11 | Lethal |
| DIMM23395 | Apurinic/apyrimidinic endonuclease | BM17151 | >1 | −0.12 | Lethal |
| Glycosylation and sugar metabolism | | | | | |
| DIMM15580 | dTDP-4-dehydrorhamnose 3,5-epimerase | BM18305 | 0.23 | 0.04 | Lethal |
| DIMM03355 | β-1,4-Mannosyltransferase | BM20353 | 0.95 | 0.94 | Lethal |
| DIMM44525 | UDP-galactopyranose mutase | BM01820 | 0.36 | 0.08 | Molt defective |
| DIMM36945 | Chitin synthase | BM18745, BM02779 | −3.52 | −4.00 | Lethal |
| Lipid metabolism | | | | | |
| DIMM52545 | Lipase | BM01258, BM03783 | 0.08 | −1.60 | Lethal |
| DIMM13730 | Sterol-C24-methyltransferase (Erg11) | BM20515 | −3.10 | −4.00 | Lethal |
| DIMM28375 | Methyltransferase | BM18889 | −0.03 | −0.15 | Lethal |
| Transport | | | | | |
| DIMM21065 | Aquaporin | BM04673 | −2.05 | −0.52 | Lethal |
| Signal transduction | | | | | |
| DIMM13570 | Nuclear hormone receptor | | −2.40 | −4.52 | Lethal |
| DIMM11130 | G protein-coupled receptor | BM19106 | −1.06 | −0.59 | Lethal |
| DIMM32415 | G protein-coupled receptor | | −1.26 | −1.57 | Lethal |
| DIMM39455 | G protein-coupled receptor | | −2.70 | −1.96 | Lethal |
| DIMM13630 | Groundhog protein | | >1 | 0.04 | Lethal |
| DIMM47150 | Warthog protein | BM01098 | >1 | 0.78 | Lethal |
| DIMM03220 | Warthog protein | BM01043, BM17326, BM08657 | >1 | 0.32 | Lethal |
| DIMM11410 | Haloacid dehalogenase-like hydrolase | BM19541 | −3.22 | −4.00 | Lethal |
| DIMM13420 | Apoptosis regulator CED-9 | BM01838 | 0.77 | −1.02 | Lethal |

Potential drug targets were filtered from the predicted *D. immitis* proteome using the following criteria: *1*) presence of an ortholog in *C. elegans* that has as an RNAi phenotype lethal, L3_arrest, or molt_defective; *2*) absence of a significant BLAST match ($E > 10^{-5}$) in the predicted proteomes of *H. sapiens* and *C. lupus familiaris*; and *3*) predicted function as an enzyme or receptor.

peptidase, as well as 3 cystatins, and many serpins (**Table 4**). Another route to modulation is through recruitment of nematode homologs of ancient system molecules that have been redeployed in the mammalian immune system, such as TGF-β and macrophage migration inhibition factor (MIF). In *D. immitis*, we identified 2 MIF genes, orthologs of the MIF-1 and MIF-2 genes of *B. malayi* and *O. volvulus* and 4 TGF-β homologs (Table 4). Another proposed route to modulation is by mimicry of immune system signals. We identified a homolog of suppressor of cytokine signaling 5 (SOCS5), a negative regulator of the JAK/STAT pathway and inhibitor of the IL-4 pathway in T-helper cells, promoting TH1 differentiation (54). Several viruses induce host SOCS protein expression for immune evasion and survival (55). Interestingly, SOCS5 homologs were also identified in the animal-parasitic nematodes *B. malayi*, *D. immitis*, *Loa loa*, *A. suum*, and *T. spiralis*, but were absent from the free-living *C. elegans*, the necromenic *Pristionchus pacificus*, and the plant parasitic *Meloidogyne* spp. *D. immitis* and other filarial nematodes (56) may use SOCS5 homologs to mimic host SOCS5. We also identified a homolog of IL-16, a PDZ domain-containing, pleiotropic cytokine (57). In mammals, IL-16 acts *via* the CD4 receptor to modulate the activity of a wide range of immune effector cells, including T cells and dendritic cells (58). Again, this molecule was only present in parasitic nematodes (including *A. suum*; ref. 59) and was absent from genomes of free-living and plant parasitic species. We suggest that these molecules and perhaps other mimics of cytokines and modulators belong to the effector toolkit used by filarial nematodes to build an immunologically compromised niche.

We surveyed the *D. immitis* genome for molecules currently proposed as vaccine candidates in other onchocercids (60, 61) and identified homologs for all 14 classes of molecules (Table 4).

### Analysis of the wDi genome: the *D. immitis-Wolbachia* symbiosis

wDi genes were predicted using the RAST online server. We performed an orthology analysis comparing wBm and wDi and found 538 shared proteins. There were 259 (with 8 duplicated) and 329 (with 4 duplicated) unique genes, respectively, for wBm and wDi. COG analysis showed that the total number of genes in each COG category was similar in the two organisms. Analysis of pairwise protein distances between wDi and wBm in different COG categories indicated that there was significant variation (Kruskal-Wallis $P=0.00077$) and pairwise Mann-Whitney tests identified 2 of the 14 high-level COG categories as having elevated divergence between the two *Wolbachia*. The COG categories showing elevated divergence were M (cell wall, mem-

TABLE 4. D. immitis *potential immune modulators and orthologs of onchocercid vaccine candidates*

| *D. immitis* protein | *B. malayi* ortholog | Description | Potential |
|---|---|---|---|
| DIMM39040, DIMM39045 | BM18548 | Pi-class glutathione *S*-transferase (GSTP) | VC |
| DIMM29150 | BM02625 | Tropomyosin (TMY) | VC |
| DIMM29270 | BM00759, BM19824 | Fatty acid and retinoic acid binding protein (FAR) | VC |
| DIMM47055 | BM03010 | Fructose bisphosphate aldolase (FBA) | VC |
| DIMM59360 | | Astacin metalloprotease MP1 | VC |
| DIMM37935, DIMM46475 | BM01859, BM09541, BM14520 | Chitinase (CHI) | VC |
| DIMM48695 | BM21967, BM08119 | Abundant larval transcript 1 (ALT); unknown function (also known as SLAP) | VC |
| DIMM48700 | BM20051 | "RAL-2," unknown function; DUF148 superfamily (also known as SXP-1) | VC |
| DIMM62215, DIMM45570, DIMM58880 | BM03177, BM05783, BM16294 | Activation associated proteins [ASP, also known as venom allergen homologs (VAH)] | VC |
| DIMM58690 | BM02480 | "OV103" *Onchocerca* vaccine candidate of unknown function | VC |
| DIMM12355 | BM07484, BM22082 | "B8" *Onchocerca* vaccine candidate of unknown function | VC |
| DIMM55190, DIMM50565, DIMM48395 | BM00175, BM14240, BM04930, BM07956 | "B20" *Onchocerca* vaccine candidate of unknown function | VC |
| DIMM56580 | BM05118 | Cysteine proteinase inhibitor 2 (CPI-2) | VC/IM |
| DIMM18905 | BM04900 | Cysteine proteinase inhibitor 3 (CPI-3) | VC/IM |
| DIMM11425 | BM21284 | Interleukin-16-like (IL16) | IM |
| DIMM57180 | BM00325 | Leucyl aminopeptidase (LAP) | IM |
| DIMM28945 | BM06847 | Suppressor of cytokine signaling 5 (SOCS5) | IM |
| DIMM42430 | BM07480 | Macrophage migration inhibitory factor (MIF-1) | IM |
| DIMM40455 | BM16561 | Macrophage migration inhibitory factor 2 (MIF-2) | IM |
| DIMM23225 | BM17713 | Transforming growth factor β (TGF) homolog of *C. elegans* TIG-2 | IM |
| DIMM37585 | BM20852 | TGF homologue of *C. elegans* DAF-7 | IM |
| DIMM29335 | BM21753 | TGF homologue of *C. elegans* DBL-1/CET-1 | IM |
| DIMM61250 | BM18112 | TGF homologue of *C. elegans* UNC-129 | IM |

*B. malayi* orthologs are referred to by their designation in WormBase WS230. IM, immune modulator; VC, vaccine candidate.

brane, and envelope biogenesis) and S (function unknown).

The relationship between filarial nematodes and their *Wolbachia* endosymbionts is thought to be a mutualistic symbiosis (62), because extended treatment of infected mammals with tetracycline and other antibiotics results in clearance of the nematodes. The bases of this symbiosis remain unclear. It has been proposed that wBm provides *B. malayi* with additional sources of critical metabolites such as heme and riboflavin (63). We interrogated the wDi genome to examine the symbiont's biochemical capabilities. *C. elegans* and other nematodes (including *B. malayi*, and, on the basis of the genome sequence presented here, *D. immitis*) are deficient in heme synthesis but wBm has an intact heme pathway (**Fig. 2**) and a CcmB heme exporter, suggesting that it may support its host by providing heme. wBm has a complete pathway from succinyl-CoA to heme (one apparently missing component, HemG, may be substituted by a functional HemY). wDi lacks both HemY and HemG (and the recently described HemJ that can perform the same transformation). This step in the heme pathway is apparently absent in other bacteria, and so this may not indicate a nonfunctional heme synthesis pathway. Further anabolic pathways absent in *D. immitis* but present in wDi are purine and pyrimidine *de novo* synthesis (Fig. 2).

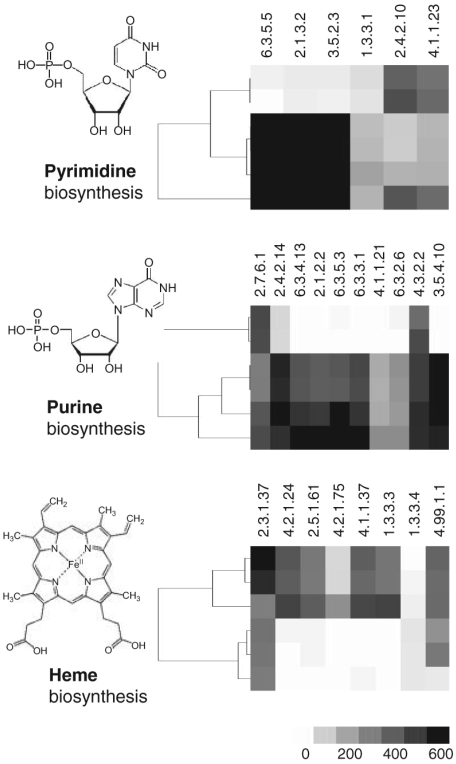


**Figure 2.** Anabolic pathways in *Wolbachia* and *Dirofilaria*. Selected pathways were identified by screening the predicted proteomes with HMM profiles representing each enzyme in the pathway using HMMer (29). The proteomes were hierarchically clustered (77) based on city block distance between the vectors consisting of the best scores (represented as a heat plot) obtained against each profile. A complete prediction of *D. immitis* metabolic pathways is available online at the Draft Genomes page of the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/catalog/org_list1.html).

wBm is deficient in folate synthesis because it lacks dihydrofolate reductase and dihydroneopterin aldolase. wDi has both these genes, suggesting that it can use dihydroneopterin as an input to folate metabolism. *Wolbachia* wMel from *Drosophila melanogaster* has both these enzymes, and they are variably present in other alphaproteobacteria. Whether this pathway contributes to the nematode symbiosis is unclear, but it does highlight another component of *Wolbachia* metabolism that may be accessible to drug development. Further wDi gene products that might be exploited as drug targets include nucleic acid synthesis and cell division proteins, such as FtsZ and DnaB, the fatty acid synthesis enzymes FabZ and AcpS, components of the Sec protein secretion system, and, possibly, the peptidoglycan synthesis enzymes of the *Mur* operon. All these are unique proteins in wDi, do not have counterparts in mammals, and are being developed as antibiotic drug targets for bacterial infections (64–68).

Horizontal gene transfer from *Wolbachia* to host nuclear genomes is common in animals harboring this endosymbiont (63), and it has been proposed that these transfers may confer new functionality to the nuclear genome (34, 69, 70), although this is unlikely (71). We identified 868 elements, spanning 219 kb, of >100 b with ≥80% identity to wDi. The *Wolbachia* origin of these elements was supported by clustering based on the frequency distribution patterns (Supplemental Fig. S1) of tetramer palindromes (72). We did not identify the putative complex *Wolbachia* insertion discussed by Dunning-Hotopp *et al.* (70) involving the antigen *Dg2* gene. We found a version of the *Dg2* gene in our predicted transcriptome that contained standard nematode introns, but no evidence of the construct previously described that had the introns of *Dg2* largely replaced with sequences that match 100% to the wDi genome. It is likely that this sequence is a laboratory or computational artifact, especially because the construct includes a cloning vector sequence in addition to *Wolbachia*.

Only 9 of our identified elements matched >80% of the length of a wDi open reading frame and were not interrupted by frame-shifting insertions or deletions or stop codons. Only one of these putative lateral gene transfers had a match to a *Wolbachia* protein of known function (transcription termination factor, NusB). We found no evidence of transcription of this gene in the male and female RNA-Seq data. We applied the same procedure for finding *Wolbachia* insertions to the *B. malayi* genome and identified 654 insertions spanning 327 kb. Only 31 pairs of insertions that were probably derived from homologous *Wolbachia* genes were found (in both of the two genomes). None of these shared insertions had complete open reading frames. Comparison with the *Wolbachia* insertions in the partial genomes of *A. viteae* and *O. flexuosa*, onchocercid nematodes that have lost their symbionts (34), revealed no insertions shared by all four species. Only 48 insertions were shared by 2 species and 5 were shared by 3. The number of shared fragments was as would be expected from homoplasious, random insertion of *Wolbachia* fragments independently into their host genomes. If ~25% of the genome was randomly transferred in all species, the number of shared fragments expected by chance would be ~45 (0.25×0.25×750 fragments). We thus tentatively conclude that, although elements from wDi have transferred to the nuclear genome, there is no evidence of their functional integration into nematode biology.

## DISCUSSION

The *D. immitis* genome sequence described here is only the second to be determined for an onchocercid nematode, despite the social and economic importance of these parasites. Three genomes were cosequenced: the mitochondrial (at ~4000-fold read coverage of the 13.6-kb genome; this had been determined previously; ref. 17); the genome of the *Wolbachia* symbiont wDi (at ~1000-fold coverage of the 0.9-Mb genome); and the nuclear genome (at ~150-fold coverage of the estimated 95-Mb genome). We used high-throughput, short-read Illumina technology, stringent quality filtering and optimized assembly methods to derive genomes of good draft quality (73). After redundancy reduction, the span of the nuclear assembly was 84.2 Mb, slightly smaller than the 88.3 Mb assembled for *B. malayi* (6). Overall, although the number of scaffolds was approximately equivalent, the contiguity of the *D. immitis* genome assembly was lower than that of *B. malayi*, because of the availability of long-range scaffolding information for the latter species. The predicted nuclear gene set was much smaller than that of *C. elegans*, but of a size similar to that of *B. malayi*. The two onchocercid nematodes also have a lower proportion of species-unique proteins. These two differences may be a feature of the Onchocercidae, because the unpublished *L. loa* genome has only 15,444 predicted proteins (Filarial Worms Sequencing Project, Broad Institute of Harvard and MIT; http://www.broadinstitute.org/). Another possibility is that the richer analytic environment for *C. elegans* in particular has permitted the identification of many unique genes using biological evidence (such as transcript information). We will continue to develop and improve the assembly and annotation of *D. immitis* and wDi as additional tools and biological resources become available.

Two peculiarities of the assembled *D. immitis* genome are striking: the lack of genetic diversity and the lack of active transposable elements. The lack of diversity was convenient, in that it allowed us to pool data obtained from two different *D. immitis* isolates, one from Pavia, Italy, and the other from Athens, Georgia, USA. Polymorphisms called from the independent sequencing of the two isolates yielded a per-nucleotide diversity of 0.04%. Both sequenced isolates fall within the single eastern United States population defined by microsatellite analyses (36). The hypovariability may be a result of the recent admixture of European and American heartworm populations through movement of domestic animals or arise from the very recent introduction of heartworm into the New World by Europeans (74). The first report of dirofilariasis in the United States dates from only 1847, as opposed to a 1626 observation from Italy. The lack of genetic diversity in the nuclear genome will make identification of mutations conferring drug resistance much easier. The lack of DNA transposons and active retrotransposons in *D. immitis* is a strong negative result, because active elements are easy to identify (they are present in multiple, highly similar copies). We identified only fragmented and functionally inactivated segments of Pao-type retrotransposons, similar to those found in and probably still active in *B. malayi*. To our knowledge, this is the first metazoan genome devoid of active transposable elements. The presence of putatively active Pao elements in *B. malayi* suggests that their loss was an evolutionary recent event in *D. immitis*.

The *Wolbachia* wDi genome, with 823 predicted proteins, complements the *D. immitis* nuclear genome in that it encodes enzymes for anabolic pathways that are missing in the latter, *e.g.*, biosynthesis of heme, purine, or pyrimidines (Fig. 2). In contrast to wBm, wDi also carries the genes for folate synthesis, suggesting that folate too might be supplied by the endosymbiont. However, essential metabolites could also be taken up from the mammalian or insect host, and so it remains to be shown whether such metabolites are actually delivered from wDi to *D. immitis*. Analysis of orthology between wBm and wDi revealed that both organisms possess many unique genes (approximately one-third of the total gene complement of each genome). The representation of genes in the different COG categories was similar for wBm and wDi, suggesting that most gene losses occurred before the split of the two lineages or that there have been no biases in gene losses/acquisition after the evolutionary separation. Analysis of protein distances revealed that proteins involved in cell wall/membrane biogenesis (COG category M) displayed more variation between the two organisms compared with the other functional categories. It is reasonable to conclude that the interface between the symbiotic bacterium and the host environment is a place where evolutionary rates are elevated, either as part of an arms race underpinning conflict between the two genomes or as a feature of the dynamic exploitation of the interface in adaptation of the symbiosis. In any case, the endosymbiont, being essential for proliferation of *D. immitis*, represents a target for control of the heartworm. Screening the predicted wDi proteome returned expected antibiotic drug targets such as Fts and Sec proteins, but also the products of the *Mur* operon required for peptidoglycan synthesis.

Many of the anthelmintics used in human medicine were originally developed for the veterinary sector. We pursued two approaches to identify potential drug targets in *D. immitis*: top-down, starting from the known anthelmintic targets of *C. elegans* (Table 2), and bottom-up, narrowing down the predicted *D. immitis* proteome to a list of essential, unique, and druggable targets (Table 3). Although the majority of the current anthelmintics activate their target (thereby interfering with synaptic signal transduction), the aim of the second approach was to identify inhibitable targets. The criteria applied—presence of an essential ortholog in *C. elegans*, absence of any significantly similar protein in human or dog, and absence of paralogs in *D. immitis*—admittedly missed many of the known anthelmintic targets, *e.g.*, proteins that are not conserved in *C. elegans* or that possess a mammalian ortholog. The aim of the approach was to maximize the specificity of *in silico* target prediction at the cost of low sensitivity. Our goal was to end up with a manageable, rather than complete, list of unique *D. immitis* proteins that are likely to be essential and druggable. Some of the candidates identified are worth further investigation, based on their presumed role in signal transduction, *e.g.*, the nematode-specific G protein-coupled receptors or hedgehog proteins (Table 3). Others have already been validated as drug targets in other systems: sterol-C-24-methyltransferase (EC 2.1.1.41) is a target of sinefungin, chitin synthase (EC 2.4.1.16) is the target of the insecticide lufenuron, and the mannosyltransferase bre-3 is required for interaction of *Bacillus thuringiensis* toxin with intestinal cells (52). The discovery of new *D. immitis* drug targets would be timely because resistance to macrocyclic lactones has recently been reported from the southern United States (75).

Filarial nematodes modulate the immune systems of their hosts in complex ways that result in an apparently intact immune system that ignores a large parasite residing, sometimes for decades, in tissues or the bloodstream. They may also require intact immune systems to develop properly (76). Often immune responses result in a pathologic condition for the host in addition to parasite clearance, and *Wolbachia* may exacerbate these responses (12). We identified a wide range of putative immunomodulatory molecules and, in addition, highlight two *D. immitis* products that may deflect or distract the host immune response: one similar to SOCS5 and the other similar to IL-18. The host-encoded versions of both of these molecules have been implicated in antifilarial immune responses. Strategies for development of a vaccine against filariases depend on delivering the correct antigens to the right arm of the immune system, avoiding induction of dangerous responses, and deflecting or stopping immune suppression by the parasite. We identified homologs of all the current roster of filarial vaccine candidates in our genome, and these can now be moved rapidly into testing in the dog heartworm model. In addition, we defined a large number of potentially secreted *D. immitis* proteins that may contribute to the host-parasite interaction and also be accessible to the host immune system.

Onchocercid parasites share not only a fascinating biology involving immune evasion, arthropod vectors, and *Wolbachia* endosymbionts but also a pressing need for new drugs, improved diagnostic methods, and, ideally, vaccines. We hope that the genome sequence of the heartworm presented here will contribute to an increased understanding of its biology and to new leads for control.

The authors thank Claudio and Marco Genchi (University of Milan, Milan, Italy) for providing *D. immitis* from Pavia, Italy, and Andrew Moorhead (University of Georgia, Athens, GA, USA) for the *D. immitis* from Athens, GA, USA. *D. immitis* genomic and RNA sequencing was performed by Fasteris SA (Geneva, Switzerland), and The GenePool Genomics Facility (Edinburgh, UK). The authors are grateful for financial support from the Novartis Fellowship Program (C.G.), the School of Biological Sciences, University of Edinburgh (S.K. and G.K.), the UK Biotechnology and Biological Sciences Research Council (G.K.), the Swedish Research Council (D.N.), the UK Medical Research Council (G0900740; M.B., N.W., and M.T.), the UK Natural Environment Research Council (R8/H10/561; M.B.), the Centre National de Recherche Scientifique (F.B.), the European collaboration Enhanced Protective Immunity Against Filariasis (EPIAF), a focused research project (Specific International Cooperative Action) of the EU (award 242131; M.B.) and the Swiss National Science Foundation (P.M.).

## REFERENCES

1. Lee, A. C., Montgomery, S. P., Theis, J. H., Blagburn, B. L., and Eberhard, M. L. (2010) Public health issues concerning the widespread distribution of canine heartworm disease. *Trends Parasitol.* **26,** 168–173
2. Genchi, C., Rinaldi, L., Mortarino, M., Genchi, M., and Cringoli, G. (2009) Climate and *Dirofilaria* infection in Europe. *Vet. Parasitol.* **163,** 286–292
3. Traversa, D., Di Cesare, A., and Conboy, G. (2010) Canine and feline cardiopulmonary parasitic nematodes in Europe: emerging and underestimated. *Parasit. Vectors* **3,** 62
4. American Heartworm Society (2010) *Diagnosis, Prevention, and Management of Heartworm* (Dirofilaria immitis) *Infection in Dogs*, American Heartworm Society, Wilmington, DE, USA
5. World Health Organization (2008) Global programme to eliminate lymphatic filariasis. Progress report and conclusions of the meeting of the technical advisory group on the global elimination of lymphatic filariasis. *Wkly. Epidemiol. Rec.* **83,** 333–348
6. Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J. E., Delcher, A. L., Guiliano, D. B., Miranda-Saavedra, D., Angiuoli, S. V., Creasy, T., Amedeo, P., Haas, B., El-Sayed, N. M., Wortman, J. R., Feldblyum, T., Tallon, L., Schatz, M., Shumway, M., Koo, H., Salzberg, S. L., Schobel, S., Pertea, M., Pop, M., White, O., Barton, G. J., Carlow, C. K., Crawford, M. J., Daub, J., Dimmic, M. W., Estes, C. F., Foster, J. M., Ganatra, M., Gregory, W. F., Johnson, N. M., Jin, J., Komuniecki, R., Korf, I., Kumar, S., Laney, S., Li, B. W., Li, W., Lindblom, T. H., Lustigman, S., Ma, D., Maina, C. V., Martin, D. M., McCarter, J. P., McReynolds, L., Mitreva, M., Nutman, T. B., Parkinson, J., Peregrin-Alvarez, J. M., Poole, C., Ren, Q., Saunders, L., Sluder, A. E., Smith, K., Stanke, M., Unnasch, T. R., Ware, J., Wei, A. D., Weil, G., Williams, D. J., Zhang, Y., Williams, S. A., Fraser-Liggett, C., Slatko, B., Blaxter, M. L., and Scott, A. L. (2007) Draft genome of the filarial nematode parasite *Brugia malayi. Science* **317,** 1756–1760
7. Slatko, B. E., Taylor, M. J., and Foster, J. M. (2010) The *Wolbachia* endosymbiont as an anti-filarial nematode target. *Symbiosis* **51,** 55–65

8. Hoerauf, A., Nissen-Pahle, K., Schmetz, C., Henkle-Duhrsen, K., Blaxter, M. L., Buttner, D. W., Gallin, M. Y., Al-Qaoud, K. M., Lucius, R., and Fleischer, B. (1999) Tetracycline therapy targets intracellular bacteria in the filarial nematode *Litomosoides sigmodontis* and results in filarial infertility. *J. Clin. Invest.* **103,** 11–18
9. Bandi, C., McCall, J. W., Genchi, C., Corona, S., Venco, L., and Sacchi, L. (1999) Effects of tetracycline on the filarial worms *Brugia pahangi* and *Dirofilaria immitis* and their bacterial endosymbionts *Wolbachia. Int. J. Parasitol.* **29,** 357–364
10. Bazzocchi, C., Mortarino, M., Grandi, G., Kramer, L. H., Genchi, C., Bandi, C., Genchi, M., Sacchi, L., and McCall, J. W. (2008) Combined ivermectin and doxycycline treatment has microfilaricidal and adulticidal activity against *Dirofilaria immitis* in experimentally infected dogs. *Int. J. Parasitol.* **38,** 1401–1410
11. Taylor, M. J., Bandi, C., and Hoerauf, A. (2005) *Wolbachia* bacterial endosymbionts of filarial nematodes. *Adv. Parasitol.* **60,** 245–284
12. Saint Andre, A., Blackwell, N. M., Hall, L. R., Hoerauf, A., Brattig, N. W., Volkmann, L., Taylor, M. J., Ford, L., Hise, A. G., Lass, J. H., Diaconu, E., and Pearlman, E. (2002) The role of endosymbiotic *Wolbachia* bacteria in the pathogenesis of river blindness. *Science* **295,** 1892–1895
13. American Pet Products Association (2012) *APPA National Pet Owners Survey*, American Pet Products Association, Greenwich, CT, USA
14. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A. L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J., Hoodless, P. A., and Birol, I. (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7,** 909–912
15. Parra, G., Bradnam, K., and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067
16. Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26,** 680–682
17. Hu, M., Gasser, R. B., Abs El-Osta, Y. G., and Chilton, N. B. (2003) Structure and organization of the mitochondrial genome of the canine heartworm, *Dirofilaria immitis. Parasitology* **127,** 37–51
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410
19. Chen, N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformat.* Chapter 4, Unit 4.10
20. Kapitonov, V. V., and Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genetics* **9,** 411–412; author reply 414
21. Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18,** 188–196
22. Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformat.* **5,** 59
23. Stanke, M., and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33,** W465–W467
24. Li, L., Stoeckert, C. J., Jr., and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13,** 2178–2189
25. Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., Vida, J. T., and Thomas, W. K. (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature* **392,** 71–75
26. Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38,** D196–203
27. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75
28. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27,** 29–34
29. Eddy, S. R. (1995) Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3,** 114–120
30. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680
31. Mulder, N. J., Kersey, P., Pruess, M., and Apweiler, R. (2008) In silico characterization of proteins: UniProt, InterPro and Integr8. *Mol. Biotechnol.* **38,** 165–177
32. Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314,** 1041–1052
33. Felsenstein, J. (1989) PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* **5,** 164–166
34. McNulty, S. N., Foster, J. M., Mitreva, M., Dunning Hotopp, J. C., Martin, J., Fischer, K., Wu, B., Davis, P. J., Kumar, S., Brattig, N. W., Slatko, B. E., Weil, G. J., and Fischer, P. U. (2010) Endosymbiont DNA in endobacteria-free filarial nematodes indicates ancient horizontal genetic transfer. *PLoS One* **5,** e11029
35. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123
36. Belanger, D. H., Perkins, S. L., and Rockwell, R. F. (2010) Inference of population structure and patterns of gene flow in canine heartworm (*Dirofilaria immitis*). *J. Parasitol.* **97,** 602–609
37. Eickbush, T. H., and Malik, H. S. (2002) Origins and evolution of retrotransposons. In *Mobile DNA II* (Craig, A. G., Craigie, R., Gellert, M., and Lambowitz, A. M., eds), ASM Press, Washington, DC
38. Guerrero, J., Campbell Seibert, B. P., Newcomb, K. M., Michael, B. F., and McCall, J. W. (1983) Activity of flubendazole against developing stages of *Dirofilaria immitis* in dogs. *Am. J. Vet. Res.* **44,** 2405–2406
39. McCall, J. W., and Crouthamel, H. H. (1976) Prophylactic activity of mebendazole against *Dirofilaria immitis* in dogs. *J. Parasitol.* **62,** 844–845
40. Carlisle, C. H., Atwell, R. B., and Robinson, S. (1984) The effectiveness of levamisole hydrochloride against the microfilaria of *Dirofilaria immitis. Aust. Vet. J.* **61,** 282–284
41. Campbell, W. C. (1982) Efficacy of the avermectins against filarial parasites: a short review. *Vet. Res. Commun.* **5,** 251–262
42. McCall, J. W. (2005) The safety-net story about macrocyclic lactone heartworm preventives: a review, an update, and recommendations. *Vet. Parasitol.* **133,** 197–206
43. Sangster, N. C., Song, J., and Demeler, J. (2005) Resistance as a tool for discovering and understanding targets in parasite neuromusculature. *Parasitology* **131**(Suppl.), S179–S190
44. Kaminsky, R., Ducray, P., Jung, M., Clover, R., Rufener, L., Bouvier, J., Weber, S. S., Wenger, A., Wieland-Berghausen, S., Goebel, T., Gauvry, N., Pautrat, F., Skripsky, T., Froelich, O., Komoin-Oka, C., Westlund, B., Sluder, A., and Mäser, P. (2008) A new class of anthelmintics effective against drug-resistant nematodes. *Nature* **452,** 176–180
45. Williamson, S. M., Walsh, T. K., and Wolstenholme, A. J. (2007) The cys-loop ligand-gated ion channel gene family of *Brugia malayi* and *Trichinella spiralis*: a comparison with *Caenorhabditis elegans. Invert. Neurosci* **7,** 219–226
46. Doyle, M. A., Gasser, R. B., Woodcroft, B. J., Hall, R. S., and Ralph, S. A. (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* **11,** 222
47. Holman, A. G., Davis, P. J., Foster, J. M., Carlow, C. K., and Kumar, S. (2009) Computational prediction of essential genes in an unculturable endosymbiotic bacterium, *Wolbachia of Brugia malayi. BMC Microbiol.* **9,** 243
48. Abbotts, R., and Madhusudan, S. (2010) Human AP endonuclease 1 (APE1): from mechanistic insights to druggable target in cancer. *Cancer Treat. Rev.* **36,** 425–435
49. Borrelli, S., Zandberg, W. F., Mohan, S., Ko, M., Martinez-Gutierrez, F., Partha, S. K., Sanders, D. A., Av-Gay, Y., and Pinto,

B. M. (2010) Antimycobacterial activity of UDP-galactopyranose mutase inhibitors. *Int. J. Antimicrob. Agents* **36,** 364–368

50. Oppenheimer, M., Valenciano, A. L., and Sobrado, P (2011) Biosynthesis of galactofuranose in kinetoplastids: novel therapeutic targets for treating leishmaniasis and Chagas' disease. *Enzyme Res.* 2011.415976
51. Ganapathy, K., Kanagasabai, R., Nguyen, T. T., and Nes, W. D. (2011) Purification, characterization and inhibition of sterol C24-methyltransferase from *Candida albicans. Arch. Biochem. Biophys.* **505,** 194–201
52. Griffitts, J. S., Huffman, D. L., Whitacre, J. L., Barrows, B. D., Marroquin, L. D., Muller, R., Brown, J. R., Hennet, T., Esko, J. D., and Aroian, R. V. (2003) Resistance to a bacterial toxin is mediated by removal of a conserved glycosylation pathway required for toxin-host interactions. *J. Biol. Chem.* **278,** 45594–45602
53. Maizels, R. M., Gomez-Escobar, N., Gregory, W. F., Murray, J., and Zang, X. (2001) Immune evasion genes from filarial nematodes. *Int. J. Parasitol.* **31,** 889–898
54. Yoshimura, A., Naka, T., and Kubo, M. (2007) SOCS proteins, cytokine signalling and immune regulation. *Nat. Rev. Immunol.* **7,** 454–465
55. Akhtar, L. N., and Benveniste, E. N. (2011) Viral exploitation of host SOCS protein functions. *J. Virol.* **85,** 1912–1921
56. Ludin, P., Nilsson, D., and Mäser, P. (2011) Genome-wide identification of molecular mimicry candidates in parasites. *PLoS One* **6,** e17546
57. Baier, M., Bannert, N., Werner, A., Lang, K., and Kurth, R. (1997) Molecular cloning, sequence, expression, and processing of the interleukin 16 precursor. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 5273–5277
58. Cruikshank, W. W., Kornfeld, H., and Center, D. M. (2000) Interleukin-16. *J. Leukoc. Biol.* **67,** 757–766
59. Wang, J., Czech, B., Crunk, A., Wallace, A., Mitreva, M., Hannon, G. J., and Davis, R. E. (2011) Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles. *Genome Res.* **21,** 1462–1477
60. Lustigman, S., James, E. R., Tawe, W., and Abraham, D. (2002) Towards a recombinant antigen vaccine against *Onchocerca volvulus. Trends Parasitol.* **18,** 135–141
61. Makepeace, B. L., Jensen, S. A., Laney, S. J., Nfon, C. K., Njongmeta, L. M., Tanya, V. N., Williams, S. A., Bianco, A. E., and Trees, A. J. (2009) Immunisation with a multivalent, subunit vaccine reduces patent infection in a natural bovine model of onchocerciasis during intense field exposure. *PLoS Negl. Trop. Dis.* **3,** e544
62. Fenn, K., and Blaxter, M. (2004) Are filarial nematode *Wolbachia* obligate mutualist symbionts? *Trends Ecol. Evol.* **19,** 163–166
63. Fenn, K., and Blaxter, M. (2006) *Wolbachia* genomes: revealing the biology of parasitism and mutualism. *Trends Parasitol.* **22,** 60–65
64. Li, Z., Garner, A. L., Gloeckner, C., Janda, K. D., and Carlow, C. K. (2011) Targeting the *Wolbachia* cell division protein FtsZ as a new approach for antifilarial therapy. *PLoS Negl. Trop. Dis.* **5,** e1411
65. Ma, S. (2012) The development of FtsZ inhibitors as potential antibacterial agents. *ChemMedChem* **7,** 1161–1172
66. Chan, D. I., and Vogel, H. J. (2010) Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem. J.* **430,** 1–19
67. Segers, K., and Anne, J. (2011) Traffic jam at the bacterial sec translocase: targeting the SecA nanomotor by small-molecule inhibitors. *Chem. Biol.* **18,** 685–698
68. Katz, A. H., and Caufield, C. E. (2003) Structure-based design approaches to cell wall biosynthesis inhibitors. *Curr. Pharm. Des.* **9,** 857–866
69. Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N., Bhattacharyya, A., Kapatral, V., Kumar, S., Posfai, J., Vincze, T., Ingram, J., Moran, L., Lapidus, A., Omelchenko, M., Kyrpides, N., Ghedin, E., Wang, S., Goltsman, E., Joukov, V., Ostrovskaya, O., Tsukerman, K., Mazur, M., Comb, D., Koonin, E., and Slatko, B. (2005) The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* **3,** e121
70. Dunning-Hotopp, J. C., Clark, M. E., Oliveira, D. C., Foster, J. M., Fischer, P., Munoz Torres, M. C., Giebel, J. D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R. V., Shepard, J., Tomkins, J., Richards, S., Spiro, D. J., Ghedin, E., Slatko, B. E., Tettelin, H., and Werren, J. H. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317,** 1753–1756
71. Blaxter, M. (2007) Symbiont genes in host genomes: fragments with a future? *Cell Host Microbe* **2,** 211–213
72. Lamprea-Burgunder, E., Ludin, P., and Maser, P. (2010) Species-specific typing of DNA based on palindrome frequency patterns. *DNA Res.* **18,** 117–124
73. Chain, P. S., Grafham, D. V., Fulton, R. S., Fitzgerald, M. G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D. C., Buhay, C., Cole, J. R., Ding, Y., Dugan, S., Field, D., Garrity, G. M., Gibbs, R., Graves, T., Han, C. S., Harrison, S. H., Highlander, S., Hugenholtz, P., Khouri, H. M., Kodira, C. D., Kolker, E., Kyrpides, N. C., Lang, D., Lapidus, A., Malfatti, S. A., Markowitz, V., Metha, T., Nelson, K. E., Parkhill, J., Pitluck, S., Qin, X., Read, T. D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R. L., Sutton, G., Thomson, N. R., Tiedje, J. M., Weinstock, G., Wollam, A., and Detter, J. C. (2009) Genomics. Genome project standards in a new era of sequencing. *Science* **326,** 236–237
74. Bowman, D. D., and Atkins, C. E. (2009) Heartworm biology, treatment, and control. *Vet. Clin. North Am. Small Anim. Pract.* **39,** 1127–1158, vii
75. Bourguinat, C., Keller, K., Bhan, A., Peregrine, A., Geary, T., and Prichard, R. (2011) Macrocyclic lactone resistance in *Dirofilaria immitis. Vet. Parasitol.* **181,** 388–392
76. Babayan, S. A., Read, A. F., Lawrence, R. A., Bain, O., and Allen, J. E. (2010) Filarial parasites develop faster and reproduce earlier in response to host immune effectors that determine filarial life expectancy. *PLoS Biol.* **8,** e1000525
77. Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14863–14868

*Received for publication February 27, 2012.*
*Accepted for publication July 30, 2012.*

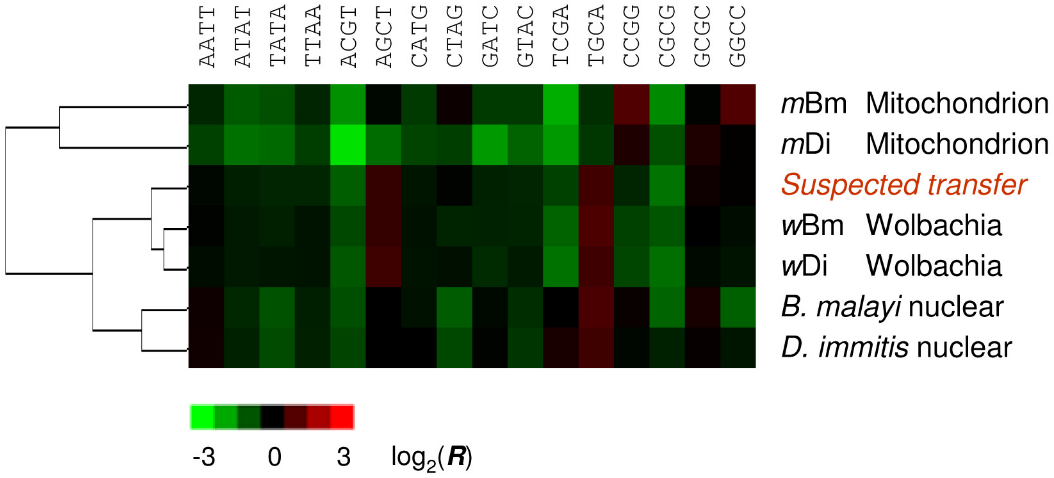


**Figure S1. Clustering of DNA based on palindrome frequency.** Nuclear genomic DNA sequences of *D. immitis* (this work) and *B. malayi* (GenBank accession DS239371) were compared to *Wolbachia w*Di (this work) and *w*Bm (NC_006833), mitochondria *m*Di (NC_005305) and *m*Bm (NC_004298), and to joined DNA fragments (*'Suspected transfer'*) assumed to have been horizontally transferred from *w*Di to the *D. immitis* nucleus (see main text). For each input sequence, the actual occurrence of the 16 possible palindromes of length 4 (top labels) was divided by the expected occurrence given its GC-content. These ratios ***R*** were transformed to $\log_2$ (numbers represented as heat map where green indicates underrepresentation of the palindrome) and the input sequences were clustered based on city-block distance of their 16-element ***R*** vectors. The *'Suspected transfer'* DNA fragments of *D. immitis* indeed clustered with the *Wolbachia* genomes, providing further evidence that they had been acquired by the *D. immitis* nucleus through horizontal transfer.

Lamprea-Burgunder, Ludin, Mäser (2011) Species-specific typing of DNA based on palindrome frequency patterns. DNA Research 18: 117

# *Chapter 6*

## Comparative Genomics of Drug Resistance in *Trypanosoma brucei rhodesiense*

Philipp Ludin[1,2] , Fabrice Graf[1,2] , Christina Kunz-Renggli[1,2] , Pascal Mäser[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Basel, Switzerland

*Working manuscript*

# Introduction

Human African trypanosomiasis (HAT) or sleeping sickness is a fatal disease caused by *Trypanosoma brucei rhodesiense* and *T. b. gambiense*, protozoan parasites that are transmitted by the tsetse fly. *Trypanosoma brucei* spp. proliferate extracellular in the bloodstream of their mammalian hosts and evade the immune system by antigenic variation. The periodic exchange of surface proteins makes the development of a vaccine highly unlikely and therefore chemotherapy is the only way to control the disease [1]. The treatment of sleeping sickness relies on just five drugs, selected based on the causative subspecies and the stage of the disease: In the first, haemolymphatic stage, patients are treated with suramin (*T. b. rhodesiense*, introduced in 1916) or pentamidine (*T. b. gambiense*, 1937). In the second stage, when the trypanosomes have invaded the central nervous system, melarsoprol (both forms, 1949), eflornithine (*T. b. gambiense* only, 1977) and nifurtimox-eflornithine combination therapy NECT (*T. b. gambiense* only, 2009) are used [2]. All of the current drugs are difficult to administer and suffer from severe adverse effects. Melarsoprol, in particular, holds unacceptable toxicity, 5% of the treated patients die from reactive encephalopathy [3]. Furthermore, treatment failure rates of up to 30% are reported throughout sub-Saharan Africa [4–7] possibly indicating the spread of drug resistant trypanosomes.

New and expectantly safer drugs are under development. Fexinidazole [8] is currently in clinical Phase II/III (www.dndi.org). However, until new drugs are available, it is essential to maximally extend the life-spans of the current ones and therefore it is crucial to understand the molecular mechanisms of drug resistance.

The known mechanisms of drug resistance were recently discussed by Barrett et al. [9]. Most investigations have addressed melarsoprol resistance. Two observations were repeatedly made, namely (i) a declined drug uptake by drug resistant trypanosomes and (ii) cross-resistance between melarsoprol and pentamidine. Cellular uptake of melarsoprol was linked to the adenosine transporter P2 encoded by *TbAT1*, the high affinity pentamidine transporter HAPT1 (gene unknown) [10] and the aquagylcerolporin *TbAQP2* [11]. Loss of function of any of these transporters – AT1, HAPT1 or AQP2 – is thought to give rise to melarsoprol-pentamidine cross-resistance. Mutations in *TbAT1* were shown to cause drug resistance in *T. brucei* lab strain [12]. The same mutations also occur in the field [13]. Whether they contribute to melarsoprol treatment failures is not resolved [14,15]. Furthermore, overexpression of the putative melarsoprol-trypanothione export pump encoded by *TbMRPA* was shown to cause melarsoprol resistance [16,17].

The following MelB resistance factors were observed for null mutant bloodstream-form *T. brucei*: 2.5 for *TbAT1* [10] and 2 for *TbAQP2* [11]. Trypanosomes with MRPA ovexpression were 6-fold resistant [18], respectively 10-fold in a *tbat1*$^{-/-}$ null background [18]. Given these results, the molecular mechanisms of high-level resistance of pentamidine and merlarsoprol have not been fully resolved. In order to gain more insights into the mechanisms of drug resistance, the drug-sensitive *T. b. rhodesiense* STIB900 strain had been selected independently *in vitro* for melarsoprol and pentamidine resistance, giving rise to the two lines STIB900-M and STIB900-P, respectively [19]. Both lines showed a high and stable cross-resistance between melarsoprol and pentamidine. Here, we perform whole genome sequencing of the parental STIB900 and its resistant derivatives STIB900-M and STIB900-P, aiming to identify new drug resistance genes at the molecular level by comparative genomics.

## Material and Methods

### *T. b. rhodesiense* lines and maintenance

*Trypanosoma brucei rhodesiense* STIB900 is a derivative of STIB704 which was isolated from a male patient at St. Francis Hospital in Ifakara, Tanzania. After several passages in rodents and a cyclic passage in *Glossina morsitans morsitans*, a cloned population was adapted to axenic growth. *T. b. rhodesiense* STIB900 was selected independently *in vitro* for melarsoprol (STIB900-M) and pentamidine (STIB900-P) as described [19]. Female Naval Medical Research Institute (NMRI) mice from Harlan (Netherlands) were inoculated with $10^6$ trypanosomes of STIB900, STIB900-M and STIB900-P, respectively. Tail blood was checked every second day for trypanosomes. When the parasitaemia was high, trypanosomes were harvested. The trypanosomes were separated from the blood cells on DEAE-cellulose columns as described [20].

### DNA isolation

Genomic DNA of the 3 *T. b. rhodesiense* lines was isolated from bloodstream-form trypanosomes propagated in mice by phenol/chloroform extraction. The purity of the obtained DNA was assessed by PCR, using primers for mouse Glyceraldehyde-3-phosphate-dehydrogenase (GAPDH). Mouse cDNA was kindly provided by Dr. Hansjörg Keller (Novartis Pharma AG, Basel) as a positive control. All *T. b. rhodesiense* DNA samples tested negative for mouse GAPDH.

### Whole genome sequencing analysis

Whole genome sequencing was carried out by 454 Life Sciences (Branford, US) on the Genome Sequencer FLX Titanium. Two shotgun runs per line were performed. FASTQ format was extracted from .sff files using 'SFF converter' from Galaxy (main.g2.bx.psu.edu) [21]. High quality (HQ) reads were mapped to the reference genome *T. b. brucei* 927 from EBI-EMBL (www.ebi.ac.uk, October 2011) using SMALT (ftp.sanger.ac.uk/pub4/resources/software/smalt). The reference genome was indexed with wordlength 13 and skipstep 1. Consensus sequence and variants relative to the reference genome were identified by SAMtools [22] using 'mpileup' command. *Ad hoc* perl scripts were used to compare nucleotide variants between the mapped reads of STIB900, STIB900-M, STIB900-P and *T. b. brucei 927* genome. SNPs were called if they had a read depth of at least 5 high quality bases (DP4 ≥ 5) and a read mapping quality of minimum 20 (mapq ≥ 20). Variants that had reads at the same position had to differ to the reference base to avoid false negatives as the reference genome is a haploid mosaic of the diploid chromosomes [23]. A SNP of STIB900-M or STIB900-P was only called if the coverage was at least 5 DP4 in STIB900 to eliminate false positives due to low coverage. The remaining SNPs, indels and gene deletions were inspected manually using Artemis (ftp.sanger.ac.uk/pub4/resources/software/artemis) [24]. The software used .embl annotation files (www.ebi.ac.uk, October 2011) to distinguish of intergenic from intragenic, and non-synonymous from synonymous mutations.

### Restriction digest of *TbAT1*

*TbAT1* was PCR-amplified with primers TbAT1_F: GAAATCCCCGTCTTTTCTCAC / TbAT1_R: ATGTGCTGAGCCTTTTTCCTT flanking of the ORF ($T_{annealing}$ = 56°C). The PCR product was purified on a silica membrane column (Nucleospin gel and PCR clean up (Macherey Nagel)) according to the supplier's protocol. The product was digested with NruI (New England Biolabs) and run on a 1.5% agarose gel.

### Status of *AQP2/3* locus

A 3.2 kb fragment encompassing the complete *AQP2/3* locus was PCR-amplified with primers AQP2/3_F: AAGAAGGCTGAAACTCCACTTG / AQP2/3_R: TGCACTCAAAAACAGGAAAAGA ($T_{annealing}$ = 58°C). The products were run on a 0.8% agarose gel.

# Results

### Whole genome sequencing analysis

The genomes of *T. b. rhodesiense* STIB900, STIB900-M and STIB900-P were sequenced on the 454 platform [25] with a calculated coverage of 18-fold (STIB900) and 21-fold (STIB900-M, STIB900-P) (Table 1). There is evidence that *T. b. brucei* and *T. b. rhodesiense* are not reproductively isolated taxa and they may differ in just one gene, the *SRA* gene [26]. Given this information, HQ reads were mapped to the reference genome *T. b. brucei* 927, 'minichromosomes' were not included [23]. *T. b. brucei* 927 is currently the best annotated *Trypanosoma brucei* genome. Reads with low mapping quality (mapq < 20) were removed, i.e. sequences that have a significant number of mismatches compared to the reference genome or reads that mapped equally well to at least one other location (non-unique/repetitive sequences or regions containing low complexity). There were only minor differences between the lines regarding reference genome overall coverage and number of genes covered with high quality bases (Table 1).

Compared to the *T. b. brucei* 927, 112'565 common SNPs were found among STIB900, STIB900-M and STIB900-P (Table 1). Only 2 were shared between both resistant lines, 31 were specific to STIB900-M and 11 were only detected in STIB900-P. However, the number of variants is an underestimate, as a mutation was only counted if all mapped reads showed differences to the reference base. The reason for the preclusion of heterozygous alleles is that the available *T. b. brucei* 927 genome is a haploid mosaic of the diploid chromosomes [23], thus variation between *T. b. rhodesiense* reads and the reference genome may occur due to heterozygosity within the *T. b. brucei* 927 sequences itself. Although we can not exclude all false-positives, the likelihood that a detected SNP is a true-positive is very high as one would expect to find at least a few reads that show the same base as the reference if there is a heterozygous allele. With this filter, we tried to maximize the specificity of a SNP at the cost of sensitivity, taking into account missing true-positives. In addition, false negative calls could occur in regions with insufficient high quality read coverage.

### Gene deletion

In terms of drug resistance, genes that are lost in STIB900-M and/or STIB900-P compared to the parental line are of high interest. It turned out that STIB900-M had not only lost *TbAT1* (Tb927.5.286b) as previously published, but a whole region of over 25 kb encompassing *TbAT1* and the adjacent genes (Tb927.5.288b, Tb927.5.289b, Tb927.5.291b, Tb927.5.292b) (Figure 1). This probably explains the failure of amplification a PCR product with primers in the 5' and 3' untranslated regions of *TbAT1* [19]. In addition to this large deletion at the *TbAT1* locus, a deletion of 1.8 kb was detected at the *AQP2/3* locus (Figure 2). Strikingly, the same deletion occurred also in STIB900-P (Figure 2). The gene deletion was confirmed by Sanger sequencing (data not shown) and by PCR (Figure 3). Further, suspected regions of deletion or small indels were inspected manually using the Artemis software, but no further distinct case was located.

**Genome-wide SNP analysis**

Compared to the drug-sensitive STIB900, one coding mutation was found in both resistant lines, three specifically in STIB900-M and one mutation only in STIB900-P (Table 2), this mutation was in *TbAT1*. In STIB900-P, a non-synonymous substitution was identified at position 430, leading from a neutral glycine to a positively charged arginine. We confirmed the point mutation with Sanger sequencing and a restriction digest of the *TbAT1* PCR product with the endonuclease NruI that specifically cuts at the position where the mutation occurs (Figure 4). The PCR product of STIB900-P was not completely digested, although both 454 and Sanger sequencing showed a homozygous mutation. Furthermore, the gene is transcribed as seen in RNAseq studies (Graf et al., unpublished). To date, the variant has not been detected in published genotyping data from laboratory strains [12] or from field isolates [13,14]. Surprisingly, the mutation was not previously reported in STIB900-P although the gene had been sequenced [19], suggesting a sequencing error or misinterpretation in the previous study.

In addition to a coding mutation in an atypical VSG (Tb927.5.4950), two non-synonymous substitutions in a hypothetical protein (Tb927.3.5720) were specific to STIB900-M (Table 2). A blastp search against the non-redundant protein database from NCBI (blast.ncbi.nlm.nih.gov) showed no siginificant similarity to known proteins, but there are paralogs within the genome. At the genomic level, Tb927.3.5720 exhibited 91% identity to the adjacent Tb927.3.5700, Tb927.3.5710 and Tb927.3.5730. No reads were mapped to these genes because they are almost identical to each other.

With regard to cross-resistance, mutations specific to both resistant lines are of particular interest since *T. b. rhodesiense* STIB900-M and STIB900-P have the same pattern of cross-resistance. The same coding variant was found in the gene of uridine-rich-binding protein 1 (UBP1) (Tb11.03.0620) in STIB900-M and STIB900-P. Compared to the drug-sensitive STIB900, STIB900-M and STIB900-P carry a leucine instead of an arginine in the RNA-binding motif of the protein at position 131 (Table 2; Figure 5).

The approach presented here detected by no means all mutations that may be involved in drug resistance. As mentioned above, the number of variants is an underestimate, and in addition, so far we only looked at coding regions and at the genomic level. For instance, a mutation in a regulatory region could lead to an over/underexpression of a gene that may lead to a reduced drug uptake. To complete our catalogue of drug resistance candidates, we are currently performing RNAseq studies to find differences at the expression levels (Graf et al., unpublished).

# Discussion

Drug resistance in African trypanosomes has been investigated since the pioneering studies of Paul Ehrlich [27]. Until recently, genes involved in resistance were identified based on hypotheses related to possible mechanisms. In the last few years, functional genomics approaches have emerged that dramatically changed the scale. New techniques allow high-throughput studies beyond the individual gene level, investigating whole genome, transcriptome, proteome and metabolome to gain insights into the mechanisms of drug resistance [9]. Here, we took advantage of whole genome sequencing. The 454 reads obtained from the three *T. b. rhodesiense* lines were mapped to the reference genome *T. b. brucei* 927 [23]. The idea was to use the published *T. b. brucei* genome as a triangulation point (and not to compare *T. b. rhodesiense* to *T. b. brucei*). Compared to the drug-sensitive STIB900 line, we found that both resistant lines have lost *TbAQP2* (Figure 3). A new coding mutation was detected in *TbAT1* of STIB900-P, a large deletion encompassing *TbAT1* and neighbouring genes in STIB900-M (Figure 1). Furthermore, we detected a non-synonymous substitution in *TbUBP1* that has not been linked to resistance so far. Strikingly, the same mutation was independently fixed in both resistant lines. The variant led to an amino acid change from a positively charged arginine to a neutral leucine. Intriguingly, the substitution occurred in the highly conserved RNA recognition motif (RRM). The conservation among kinetoplastids is illustrated in Figure 5. Interestingly, it was shown that the corresponding arginine residue in the *T. cruzi* homolog plays a crucial role in RNA binding [28]. Moreover, TcUBP1 was implicated in regulation of small mucin gene (SMUG) mRNA levels [29]. However, a specific role has not been attributed to *TbUBP1* so far. It was suggested that *TbUBP1* is essential for parasite growth and that it may be involved in mRNA regulation [30]. In general, cross-resistance may occur if the drugs share a common uptake/efflux mechanism and/or a common target. It is assumed that the mode of action of pentamidine and melarsoprol are different [31]. One explanation could be that the amino acid substitution changes the binding properties of UBP1 to the negatively charged RNA. This may affect mRNA levels of other genes. Up- or downregulation of a specific gene (i. e. transporter) should be detectable with RNAseq studies (Graf et al., unpublished). However, the mutation could also have a more systemic effect, i.e. it could stabilize mRNAs of several genes that lead to an overproduction of proteins that were affected by the drug itself or by oxidative stress.

A crucial experiment will be to test, by double homozygous deletion, whether simultaneous loss of *TbAT1* and *TbAQP2* fully explains the observed multidrug resistant phenotype of *T. b. rhodesiense* STIB900-M or whether additional mutations, such as that identified in *TbUBP1*, must be involved.

An important point will be to evaluate whether the variations are relevant to the field. Mutations conferring resistance are often associated with fitness costs. For instance, it was shown in *Mycobacterium tuberculosis* that rifampin-resistant mutants had a reduced fitness when compared with the sensitive ancestor in competition assays and that mutations bearing the highest fitness cost never appeared in clinical isolates [32].

Once relevant determinants of drug resistance are selected, a diagnostic tool may be eventually developed that allows to monitor drug resistance in the field. By designing specific primers that can discriminate between drug-resistant and drug-sensitive lines, a loop-mediated isothermal amplification (LAMP) test [33] may have great potential to be applied in the field.

# References

1. Stuart K, Brun R, Croft S, Fairlamb A, Gürtler RE, et al. (2008) Kinetoplastids: related protozoan pathogens, different diseases. J Clin Invest 118: 1301–1310. doi:10.1172/JCI33945.

2. Brun R, Blum J, Chappuis F, Burri C (2010) Human African trypanosomiasis. Lancet 375: 148–159. doi:10.1016/S0140-6736(09)60829-1.

3. Kennedy PGE (2008) The continuing problem of human African trypanosomiasis (sleeping sickness). Ann Neurol 64: 116–126. doi:10.1002/ana.21429.

4. Kibona SN, Matemba L, Kaboya JS, Lubega GW (2006) Drug-resistance of *Trypanosoma b. rhodesiense* isolates from Tanzania. Trop Med Int Health 11: 144–155. doi:10.1111/j.1365-3156.2005.01545.x.

5. Stewart ML, Krishna S, Burchmore RJS, Brun R, De Koning HP, et al. (2005) Detection of arsenical drug resistance in *Trypanosoma brucei* with a simple fluorescence test. Lancet 366: 486–487. doi:10.1016/S0140-6736(05)66793-1.

6. Matovu E, Enyaru JC, Legros D, Schmid C, Seebeck T, et al. (2001) Melarsoprol refractory *T. b. gambiense* from Omugo, north-western Uganda. Trop Med Int Health 6: 407–411.

7. Robays J, Nyamowala G, Sese C, Betu Ku Mesu Kande V, Lutumba P, et al. (2008) High failure rates of melarsoprol for sleeping sickness, Democratic Republic of Congo. Emerging Infect Dis 14: 966–967. doi:10.3201/eid1406.071266.

8. Mäser P, Wittlin S, Rottmann M, Wenzler T, Kaiser M, et al. (2012) Antiparasitic agents: new drugs on the horizon. Current opinion in pharmacology. Available: http://www.ncbi.nlm.nih.gov/pubmed/22652215. Accessed 19 September 2012.

9. Barrett MP, Vincent IM, Burchmore RJS, Kazibwe AJN, Matovu E (2011) Drug resistance in human African trypanosomiasis. Future Microbiol 6: 1037–1047. doi:10.2217/fmb.11.88.

10. Matovu E, Stewart ML, Geiser F, Brun R, Mäser P, et al. (2003) Mechanisms of arsenical and diamidine uptake and resistance in *Trypanosoma brucei*. Eukaryotic Cell 2: 1003–1008.

11. Baker N, Glover L, Munday JC, Aguinaga Andrés D, Barrett MP, et al. (2012) Aquaglyceroporin 2 controls susceptibility to melarsoprol and pentamidine in African trypanosomes. Proc Natl Acad Sci USA 109: 10996–11001. doi:10.1073/pnas.1202885109.

12. Mäser P, Sütterlin C, Kralli A, Kaminsky R (1999) A nucleoside transporter from *Trypanosoma brucei* involved in drug resistance. Science 285: 242–244.

13. Matovu E, Geiser F, Schneider V, Mäser P, Enyaru JC, et al. (2001) Genetic variants of the *TbAT1* adenosine transporter from African trypanosomes in relapse infections following melarsoprol therapy. Mol Biochem Parasitol 117: 73–81.

14. Maina N, Maina KJ, Mäser P, Brun R (2007) Genotypic and phenotypic characterization of *Trypanosoma brucei gambiense* isolates from Ibba, South Sudan, an area of high melarsoprol treatment failure rate. Acta Trop 104: 84–90. doi:10.1016/j.actatropica.2007.07.007.

15. Kazibwe AJN, Nerima B, De Koning HP, Mäser P, Barrett MP, et al. (2009) Genotypic status of the *TbAT1/P2* adenosine transporter of *Trypanosoma brucei gambiense* isolates from Northwestern Uganda following melarsoprol withdrawal. PLoS Negl Trop Dis 3: e523. doi:10.1371/journal.pntd.0000523.

16. Mäser P, Kaminsky R (1998) Identification of three ABC transporter genes in *Trypanosoma brucei* spp. Parasitol Res 84: 106–111.

17. Shahi SK, Krauth-Siegel RL, Clayton CE (2002) Overexpression of the putative thiol conjugate transporter *TbMRPA* causes melarsoprol resistance in *Trypanosoma brucei*. Mol Microbiol 43: 1129–1138.

18. Lüscher A, Nerima B, Mäser P (2006) Combined contribution of *TbAT1* and *TbMRPA* to drug resistance in *Trypanosoma brucei*. Mol Biochem Parasitol 150: 364–366. doi:10.1016/j.molbiopara.2006.07.010.

19. Bernhard SC, Nerima B, Mäser P, Brun R (2007) Melarsoprol- and pentamidine-resistant *Trypanosoma brucei rhodesiense* populations and their cross-resistance. Int J Parasitol 37: 1443–1448. doi:10.1016/j.ijpara.2007.05.007.

20. Lanham SM, Godfrey DG (1970) Isolation of salivarian trypanosomes from man and other mammals using DEAE-cellulose. Exp Parasitol 28: 521–534.

21. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15: 1451–1455. doi:10.1101/gr.4086505.

22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi:10.1093/bioinformatics/btp352.

23. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The genome of the African trypanosome *Trypanosoma brucei*. Science 309: 416–422. doi:10.1126/science.1112642.

24. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.

25. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380. doi:10.1038/nature03959.

26. Balmer O, Beadell JS, Gibson W, Caccone A (2011) Phylogeography and taxonomy of *Trypanosoma brucei*. PLoS Negl Trop Dis 5: e961. doi:10.1371/journal.pntd.0000961.

27. Ehrlich P (1907) Chemotherapeutische Trypanosomen-Studien. Berliner Klinische Wochenschrift 9.

28. Volpon L, D’Orso I, Young CR, Frasch AC, Gehring K (2005) NMR structural study of TcUBP1, a single RRM domain protein from *Trypanosoma cruzi*: contribution of a beta hairpin to RNA binding. Biochemistry 44: 3708–3717. doi:10.1021/bi047450e.

29. D’Orso I, Frasch AC (2001) TcUBP-1, a developmentally regulated U-rich RNA-binding protein involved in selective mRNA destabilization in trypanosomes. J Biol Chem 276: 34801–34809. doi:10.1074/jbc.M102120200.

30. Hartmann C, Benz C, Brems S, Ellis L, Luu V-D, et al. (2007) Small trypanosome RNA-binding proteins TbUBP1 and TbUBP2 influence expression of F-box protein mRNAs in bloodstream trypanosomes. Eukaryotic Cell 6: 1964–1978. doi:10.1128/EC.00279-07.

31. Barrett MP, Fairlamb AH (1999) The biochemical basis of arsenical-diamidine crossresistance in African trypanosomes. Parasitol Today (Regul Ed) 15: 136–140.

32. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, et al. (2006) The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. Science 312: 1944–1946. doi:10.1126/science.1124410.

33. Njiru ZK, Mikosza ASJ, Matovu E, Enyaru JCK, Ouma JO, et al. (2008) African trypanosomiasis: sensitive and rapid detection of the sub-genus *Trypanozoon* by loop-mediated isothermal amplification (LAMP) of parasite DNA. Int J Parasitol 38: 589–599. doi:10.1016/j.ijpara.2007.09.006.

# Tables

| | STIB900 | STIB900-M | STIB900-P |
|---|---|---|---|
| HQ reads total | 1'418'416 | 1'513'957 | 1'482'415 |
| Average read length | 330 | 362 | 371 |
| Calculated coverage | 18 | 21 | 21 |
| **Mapping statistics** | | | |
| Mapped reads (all mapping quality) | 1'212'557 | 1'205'716 | 1'201'165 |
| Mapped reads (mapq ≥ 20) | 897'213 | 959'151 | 959'663 |
| % coverage of RG with HQ bases (DP4 ≥ 5) | 82% | 82% | 82% |
| % gene coverage (gene length ≥ 95%) | 84% | 81% | 83% |
| **SNP statistics** | | | |
| Overall | 112'565 | 112'598 | 112'578 |
| CDS | 46'453 | 46'458 | 46'455 |
| Non-synonymous | 19'575 | 19'579 | 19'577 |

**Table 1**. Whole genome sequencing and mapping statistics when compared to *T. b. brucei* 927 reference genome. (RG: reference genome, HQ: high quality).

| Chr | Gene ID | Annotation | STIB900 | STIB900-M | STIB900-P | Substitution |
|---|---|---|---|---|---|---|
| 3 | Tb927.3.5720 | hypothetical protein | A | C | A | K188N |
| 3 | Tb927.3.5720 | hypothetical protein | T | C | T | F189L |
| 5 | Tb927.5.4950 | VSG, atypical | G | T | G | T347K |
| 5* | Tb927.5.286b | adenosine transporter 1 | G | G | C | G430R |
| 11 | Tb11.03.0620 | RNA-binding protein | G | T | T | R131L |

**Table 2**. Coding point mutations occurred in drug-resistant lines. * *TbAT1* is currently not included in the dataset of chromosome 5, it is included in BAC26D11.

# Figures

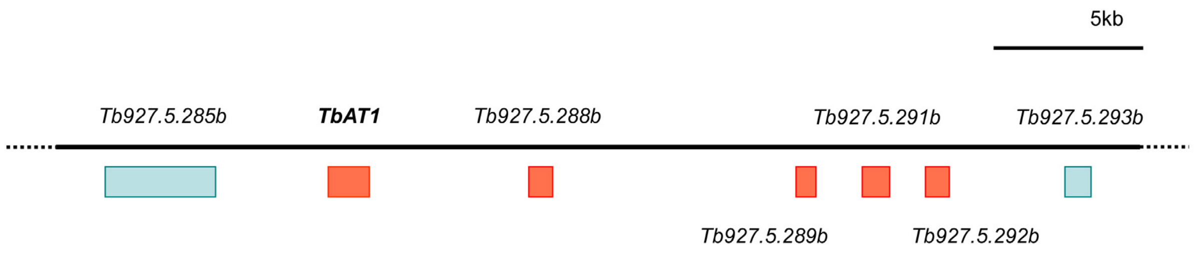


**Figure 1.** Deletion encompassing *TbAT1* and neighbouring genes. Genes depicted in red were lost in STIB900-M.

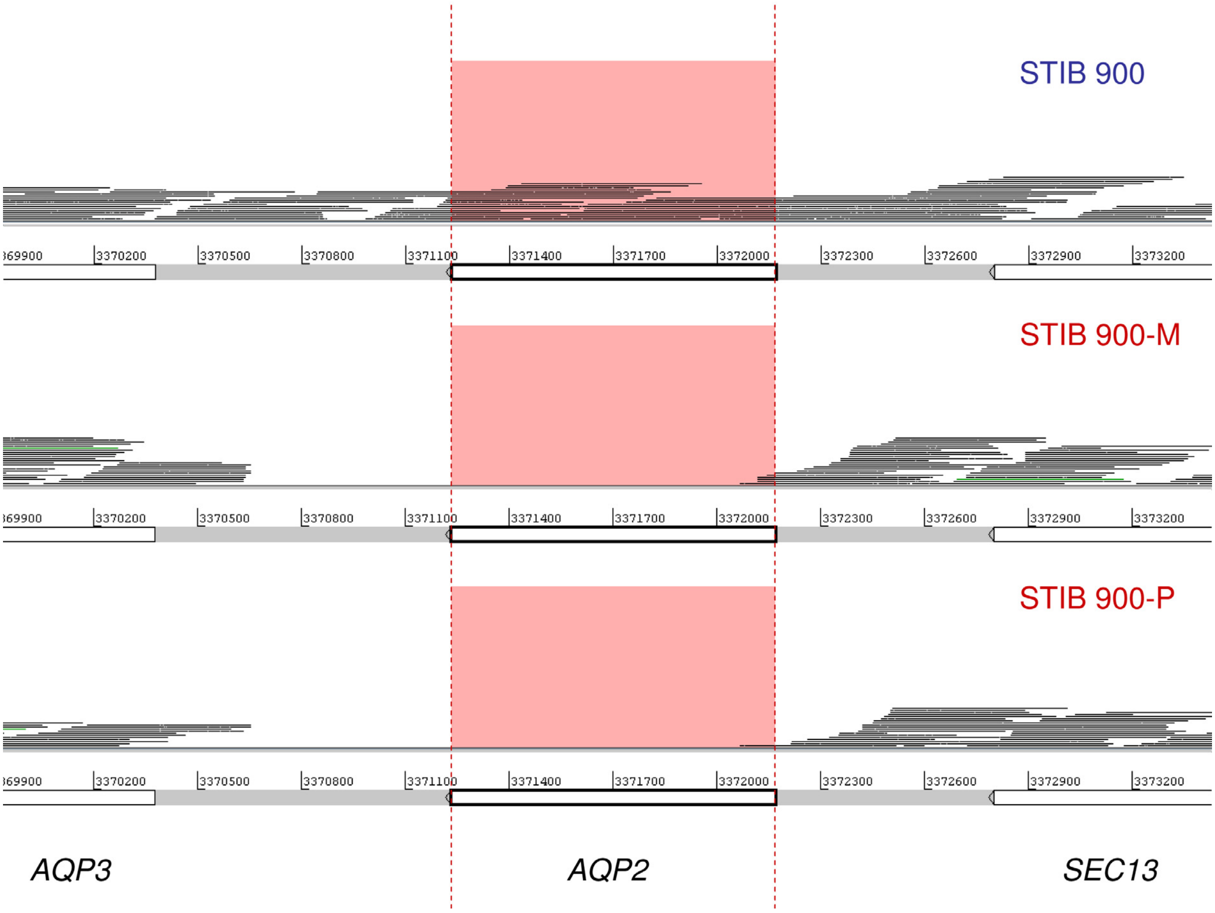


**Figure 2.** Loss of *AQP2* in the drug-resistant lines. Reads from STIB900, STIB900-M and STIB900-P were mapped to the reference genome *T. b. brucei* 927 using SMALT. Numbers indicate the genomic position on chromosome 10. Arrows represent the length of the genes and their orientation. Artemis was used to visualize the data.

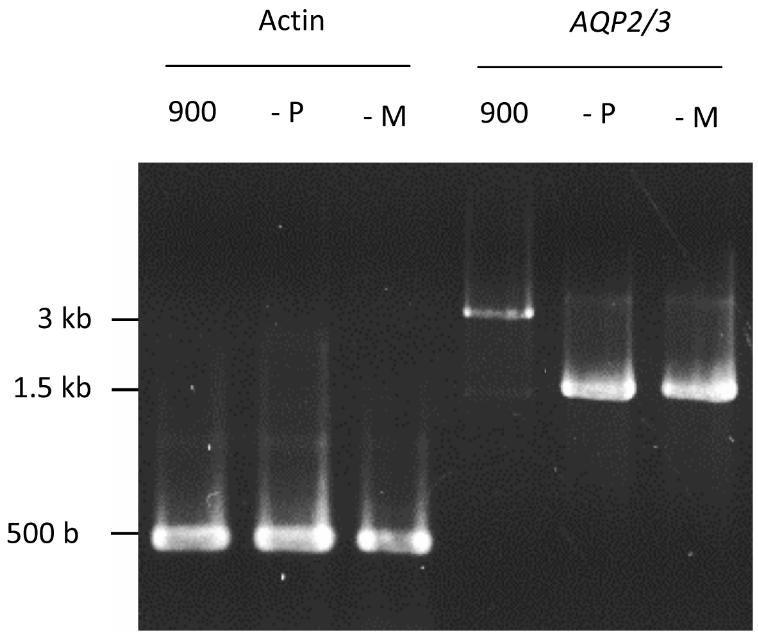


**Figure 3.** Probing for *AQP2/3* locus by PCR. Smaller PCR products were obtained from STIB900-P and STIB900-M when compared to STIB900. Actin served as a positive control.

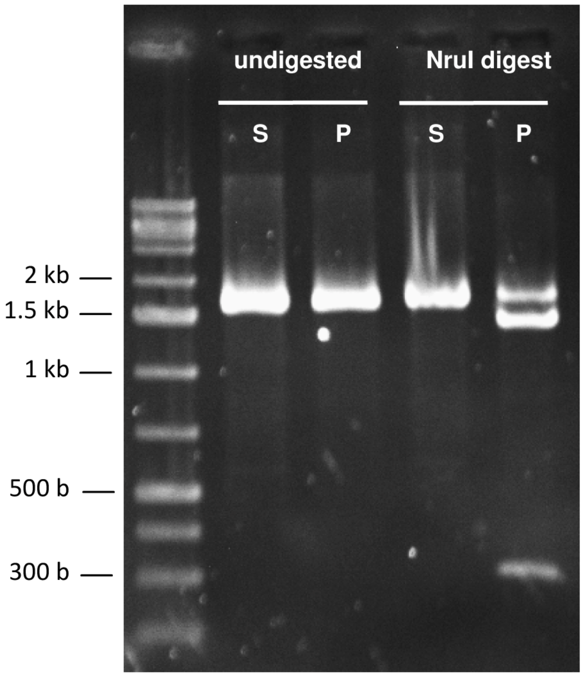


**Figure 4.** Restriction digest of the *TbAT1* PCR products from sensitive parental (S) and pentamidine-selected (P) trypanosomes with the endonuclease NruI.

```
                                                                                  ↓
STIB900-M   RNLMVNYIPTTVDEVQLRQLFERFGAIESVKIVCDRETRQSRGYGFVKFQSASSAQQAIASLNGFVILNKLLKVALAASGHQR
STIB900-P   RNLMVNYIPTTVDEVQLRQLFERFGAIESVKIVCDRETRQSRGYGFVKFQSASSAQQAIASLNGFVILNKLLKVALAASGHQR
STIB900     RNLMVNYIPTTVDEVQLRQLFERFGAIESVKIVCDRETRQSRGYGFVKFQSASSAQQAIASLNGFVILNKRLKVALAASGHQR
TbUBP1      RNLMVNYIPTTVDEVQLRQLFERFGAIESVKIVCDRETRQSRGYGFVKFQSASSAQQAIASLNGFVILNKRLKVALAASGHQR
TcoUBP1     RNLMVNYIPTTVDEVQLRQLFERFGPIESVKIVCDRETRQSRGYGFVKFQSAASAQQAIASLNGFVILNKRLKVALAASGHQR
TcrUBP1     RNLMVNYIPTTVDEVQLRQLFERYGPIESVKIVCDRETRQSRGYGFVKFQSGSSAQQAIAGLNGFNILNKRLKVALAASGHQR
LmUBP1      RNLMVNYIPTTVDEVQLRQLFERFGPIEGVKIVCDRETRQSRGYGFVKYHSAASAQQAVNELNGFNILNKRLKVALAASGNQR
            ***********************:*.**.*******************::*.:*****:  **** **** *********:**
```

**Figure 5.** ClustalW alignment of the RRM motif from UBP1 homologs of kinetoplastids. Arrow indicates position where the amino acid substitution occurred in the drug-resistant STIB900-M and STIB900-P. (Tb: *Trypanosoma brucei brucei*, Tco: *T. congolense*, Tcr: *T. cruzi*, Lm: *Leishmania major*).

# *Chapter 7*

## General Discussion and Conclusion

# General Discussion

## 1 Rationale, Objectives and Aim of the Present Thesis

The genome contains the evolutionary history as well as the functions of an organism. However, the DNA sequence of a single species alone may reveal only a few secrets. Instead, comparative genomics to other species will yield information on the genetic and evolutionary bases of shared and distinct properties throughout the tree of life. Indeed, our knowledge about genomes and their functions have been dramatically enhanced in the post-genomic era. Comparative genomics is an emerging field in biology that started when the first two sequenced genomes of self-replicating organisms were available, *Haemophilus influenzae* and *Mycoplasma genitalium* [1]. Since then, an enormous number of genomes from parasites, hosts and vectors were sequenced. Sequencing and assembly technologies have been further developed which led to remarkably reduced costs and time to complete a genome sequence.

Comparative genomics of parasites may be studied in different ways: (i) between parasites and free-living organisms, (ii) between parasite and host, (iii) within closely related species (family, genus), and (iv) within the same species. In my PhD thesis, all four disciplines were touched in various projects. The aim was to explore and benefit from the wealth of available genomes on the one side and being an active part in whole genome sequencing projects on the other side. In particular, three algorithms (*Chapter 2, 3, 4*) were invented that are of general interest to the parasitology community, and they were applied in the genome project of *Dirofilaria immitis* (*Chapter 5*). Further, the genomes of *Trypanosoma brucei rhodesiense* STIB900 and its two drug-resistant derivatives STIB900-M and STIB900-P were sequenced to identify the mutations underlying drug resistance (*Chapter 6*). For the further discussion, *Chapter 2-5* will be taken together and discussed in a general manner separately from *Chapter 6*.

## 2 Inventions of Algorithms and their Applications in Parasitology

With the rapidly growing number of sequenced genomes, the collected data need to be stored in customized databases for ease of community access. Nowadays, well-curated databases such as NCBI, EMBL and Uniprot exist, which arrange the wealth of data in a user-friendly way either by storing their data directly on their servers or by linking them to other databases. For eukaryotic parasites, EuPathDB (eupathdb.org) and WormBase (wormbase.org) are currently the best portals for accessing genomic-scale datasets. Besides the data collection, there is a plethora of open-source programs available for similarity searches, orthology detection, transmembrane domain or signal prediction, etc.  They can be accessed online on a web interface or downloaded as standalones. However, if complex biological questions are addressed on a large scale, the output of these programs needs to be parsed and fed into other programs to obtain the anticipated results. In bioinformatics, automated pipelines are developed that combine existing programs with e. g. Perl scripts for parsing, reformatting and analysis for output. Perl is a simple but effective programming language. In my thesis, algorithms were invented that were partly based on existing programs, and Perl scripts were created to build tools that are automated and widely applicable.

Although some genome projects documented molecular mimicry candidates in their publications [2–7], no systematic approach has been conducted to identify molecular mimicry in the post-genomic era. The first developed algorithm was therefore to perform a genome-wide survey for molecular mimicry candidates (*Chapter 2*) which was built on my M. Sc. thesis [8]. In brief, I screened parasite proteomes independently with either full-length proteins or overlapping peptide fragments for linear epitopes that were present in the host proteome but were not found in free-living control organisms. Compared to my master thesis, several improvements were made, namely (i) a more accurate filtering system, (ii) the use of randomized sequences to estimate significance of the identified candidates, (iii) the setting up of an online database (mimicdb.scilifelab.se) designed for community access to the mimicry data, and (iv) the development of a fully automated tool with exchangeable parasite, host and control proteomes.

Amongst the identified mimicry candidates, my personal highlights were the SOCS5 homolog of *B. malayi* (*Chapter 2*, Figure 2) and the motif shared between human vitronectin and the extracellular domain of several PfEMP1 variants of *P. falciparum* (*Chapter 2*, Figure 5). SOCS proteins are important regulators of the immune system [9] and we proposed that *B. malayi* may mimic SOCS5 to manipulate the host's immune system (*Chapter 2*). Intriguingly, SOCS5 homologs were also present in the genomes of the animal-parasitic nematodes *D. immitis*, *A. suum* and *T. spiralis* but were not identified in the plant-parasitic *Meloidogyne* spp., the free-living *C. elegans* and necromenic *P. pacificus* (*Chapter 5*). Strikingly, various viruses have developed mechanisms to induce host SOCS protein expression for immune evasion and survival [10]. However, SOCS5-like proteins from animal-parasitic nematodes did not carry an N-terminal signal peptide [11], but they were predicted to be targeted to the non-classical secretory pathway using the SecretomeP method [12]. The SOCS5 homolog was not detected in E/S products of *B. malayi* proteomic analysis [13–15] but was found in the somatic fraction of L3 larvae [15]. Transcriptomic analysis showed expression in all the profiled lifecycle stages (microfilariae, adults, L3 and L4) [16]. Therefore, it seems that the gene is expressed but more experimental data are needed to clarify the function of SOCS5 homologs in parasitic nematodes and their roles in host-parasite interaction.

In the genome project of *D. immitis*, we expanded our survey for host manipulating candidates with blast searches of known immune modulators from parasitic nematodes. Most of these cannot be detected with the presented mimicry approach because they share similarities to *C. elegans* proteins (and *C. elegans* was used as a negative filter; see *Chapter 2*, p. 28). Amongst others, a *D. immitis* MIF-1 homolog was found that was previously described in other nematodes as well as in *Leishmania* and *Plasmodium* (*Chapter 1*). In general, the identified repertoire of immune modulators seemed to be rich, in agreement with the view of parasitic helminths role as 'masters of immune regulation' [17].

In *P. falciparum*, we found several fragments from different PfEMP1 variants in all investigated strains (3D7, HB3, DD2) that shared amino acid identity with human vitronectin (*Chapter 2*, Figure 5). This was the best hit over all parasites analysed and very striking as only significant hits were returned from vitronectin and PfEMP1 when searched with the shared amino acid stretch on non-redundant protein databases. On the host side, the motif was detected in vitronectin, a protein that is abundant in plasma and present in the extracellular matrix of various human tissues [18]. The function of vitronectin is manifold, most notably it is implicated in negative immunomodulation and cytoadherence [19]. Moreover, *Moraxella catarrhalis*, *Haemophilus influenzae* and *Neisseria gonorrhoeae* recruit host vitronectin to their surface to evade the complement system, *Streptococcus pneumoniae* and *Streptococcus pyogenes* utilize host vitronectin for adherence and internalisation [20]. The mimicked region lay in the first heparin-binding domain; however, a specific function has not been attributed so far. On the parasite side, the corresponding amino acid stretch was predicted in several PfEMP1 variants extracellularly close to the transmembrane domain (*Chapter 2*, Figure 5). PfEMP1 is expressed on the surface of infected erythrocytes and thus involved in host-parasite interactions. Therefore it is easy to conceive that the proposed mimicry motif would hold advantage for the parasite. In the 3D7 genome, the *var* gene family was subdivided into 16 types based on their domain type and order, or into three major classes (upsA, upsB, upsC) according to their upstream regions [2]. So far, a minimum of 12 human receptors have been documented to mediate adhesion to infected erythrocytes and it is thought that different subsets of PfEMP1s bind to distinct receptors which may influence disease severity [21,22]. PfEMP1 variants containing the amino acid stretch similar to vitronectin were not grouped to a specific type or class. However, the motif was identified as a homology block (IGHvar29_1726_1743) when 399 PfEMP1 sequences were aligned [23]. Overall, 43 PfEMP1s across eight genomes (*P. falciparum* clone 3D7, HB3, DD2, IT4/FCR3, PFCLIN, RAJ226, IGH and *P. reichenowi* clone PREICH) were identified when searching on the VarDom server (www.cbs.dtu.dk/services/VarDom). In general, it is difficult to demonstrate molecular mimicry experimentally because by definition *in vivo* models are needed to obtain conclusive results. Unfortunately, this is not possible for *P. falciparum* and in the case of PfEMP1, *P. berghei* or *P. chabaudi* are not an option as they do not have *var* genes. Moreover, PfEMP1 and Vitronectin are multifunctional proteins with several domains which render the experiments even more difficult.

By comparative genomics, parasite linear epitopes were detected that resembled host sequences. BLAST was implemented to conduct direct sequence comparisons between parasite and host. However, neither mimicry by glycosylated peptides nor by non-linear epitopes were identified. Although the number of 3D or carbohydrate structures from molecules stored in databases is growing steadily, it is unrealistic to perform genome-wide surveys with conformational epitopes or carbohydrate compositions at the moment. Moreover, parasites may mimic lipids [24] or microRNAs [25] to manipulate their host, mechanisms we did not cover in our study. Further, the pipeline is dependent on the appropriate choice of the free-living negative controls. It is crucial to have a convenient mixture of unrelated free-living organisms to eliminate generally conserved proteins and motifs. Nevertheless, the survey vastly benefits if a closely related species is available. For instance, *C. elegans* serves as a convenient control for parasitic nematodes as it excludes genus-specific sequences unassociated to mimicry. A reduction of artefacts would certainly be observed in kinetoplastids with the inclusion of *Bodo saltans,* or with *Chromera velia* in apicomplexan parasites.

To make our algorithm accessible to the research community, a fully-automated tool was developed that can be adopted to any parasite-host pair and free-living control species. Moreover, we made our results publicly available by setting up mimicDB (mimicdb.scilifelab.se), a database where our mimicry candidates are stored. Thus I believe that our work will contribute to a better understanding of host-parasite relationships at the molecular level.

As advanced sequencing technologies open up new opportunities so do we face new challenges. High-throughput methods allow whole genome sequencing not only for a single species but for whole communities as well. A major challenge in exploiting these data is the binning of non-overlapping DNA contigs into groups that correspond to the different species present in the community [26]. In my PhD thesis, an algorithm was invented that allows to discriminate DNA sequences to the level of species based on palindrome frequency patterns (*Chapter 3*). In summary, we made use of the general underrepresentation of short palindromes in all kinds of genomes, and the observation that palindromes exhibit the highest inter-species but a low intra-species variance. The degree of underrepresentation of a given palindrome depends on the organism; the pattern of palindrome frequency distribution is highly species-specific. In our analysis, we included the 16 palindromic words of length 4. Therefore, every investigated DNA sequence converts to a vector of 16 numbers which allows to group sequences by clustering based on distance. With this method, we were able to cluster together DNA from the same species, and to assign plasmids and certain viruses to their corresponding host genomes. Nevertheless, it is important to mention that no phylogenetic tree can be inferred from palindrome typing. Moreover, a critical point is the minimal sequence length needed as input. At the moment, most commercial available sequencing machines produce reads shorter than 1 kb [27], however, new technologies promise significantly longer reads [28]. But as long as the reads are shorter than 9 kb, they need to be concatenated to long enough contigs prior to their usage in our program. In general, the longer the input sequence, the more accurate the pattern. A higher discriminative power could also be achieved by the usage of the 64 palindromic words of length 6, but then the input sequences would need to be even longer.

To test the presented method, the developed tool was applied to the sequences obtained from the 2007 'Sorcerer II Global Ocean Sampling Expedition' [29], where the hundred largest contigs were analysed (*Chapter 3*, Figure S2). In fact, we were capable of producing major clusters which appeared to be species-specific. However, we could not assess our results because the genomes of the investigated species were not stored in the NCBI nucleotide collection. Nonetheless, the potential of our application was underpinned when we assembled randomly picked DNA pieces correctly to their respective species of origin in over 90% (*Chapter 3*, p. 49).

In the *D. immits* genome project, we took advantage of the palindrome frequency patterns to provide further evidence for the insertion of *Wolbachia* DNA into the nuclear genome of *D. immits* through horizontal gene transfer. We demonstrated that the sequences of suspected horizontal gene transfer clearly clustered together with the *Wolbachia* genome (*Chapter 5*, Figure S1). *Wolbachia*-host transfer has been reported for several species [30], but the evidences supplied from whole genome data are rarely with any doubt [31] and hence our method sheds more light on the proposed transfer sequences, provided that the sequences are long enough.

Taken together, we developed a tool that clusters DNA sequences to their species of origin by comparative palindromics. The power of the program lies in its simple application and high resolution. Even closely related species such as *C. elegans* and *C. briggsae* were correctly resolved. The method can in principal be adopted to any DNA sequence which needs to be assigned to its species of origin.

The increasing availability of genomic sequences from parasites is also supporting drug discovery. As many parasites are genetically intractable, the prediction and prioritization of promising drug targets are pivotal steps in rational drug design [32,33]. In the frame of my PhD thesis, an algorithm was developed to predict drug target candidates based on phylogenomics among closely related species and comparative genomics to the host (*Chapter 4*). Even though not all targets are encoded by essential genes [33], a desirable drug target should be essential and specific for the parasite. The prediction of essentiality plays a critical role in target-based approaches. At the moment, reverse genetic techniques such as homologous recombination or gene silencing are in many parasites impractical at the single gene level, not to mention on a genome-wide scale [33]. Moreover, related model organisms from which essential genes may be inferred are not available for a majority of parasites. Therefore, an essentiality prediction was developed that is largely independent of experimental data and hence widely applicable. The hypothesis was that a protein is less likely to be dispensable if there are no other similar proteins in the same proteome, but conserved orthologues in all the related species. The hypothesis was tested with the *S. cerevisiae* deletion phenotype data set, where the applied algorithm significantly enriched for essential genes. Further, the drug target prediction pipeline was refined by the inclusion of more criteria: To be considered as potential drug target, the protein needs to be matchless in the host proteome, expressed in a relevant stage, and predicted to be druggable.

The presented approach for target prediction mainly differs from published ones [32–34] in that it relies on evolutionary evidence and comparative genomics between closely related species to predict essentiality, and not on model organisms. In addition, amino acid comparisons were based on Needleman-Wunsch global alignments instead of Smith-Waterman local alignments to avoid misleading similarity interpretation due to small common domains.

As functional genomics approaches are limited in *Plasmodium* spp., and no related model organism is available and there is an urgent need for new antimalarial drugs, *P. falciparum* was an ideal pathogen for the designed drug target pipeline. Importantly, the genomes from different *Plasmodium* spp. were available that permit a reliable orthology analysis. Starting with the complete set of 5400 proteins, we came up with a sizeable list of 40 candidates which contained proven and new targets. Strikingly, of the nine proteins investigated by reverse genetics in *Plasmodium* spp. that were in our list, eight appeared to be essential, backing our fundamental hypothesis to enrich for essentiality. A further improvement could be obtained by inclusion of network connectivity information as it is widely believed that highly connected genes are more likely to be essential [33]. However, available *P. falciparum* network data were incomplete and therefore not incorporated. As the idea was to compile a sizeable set of highly promising targets, we missed many plausible candidates by applying stringent filters. Although not of first priority, of potential interest are the 178 hypothetical proteins. Among them, there might be very exciting targets as they would be most likely novel and provide new opportunities to tackle *Plasmodium* species. Therefore, the high number of proteins with unknown function is a real bottleneck in malaria research. Nevertheless, to be considered as a promising protein in a target-based approach, a functional annotation needs to be at hand. From my point of view, the most prospective candidates to begin with are the predicted phosphomannose isomerase and phosphenolpyruvate carboxylase. Since both enzymes were successfully inhibited in other systems by antifungals or herbicides respectively, the drug development process would not need to start from scratch. The potential of agrochemicals against protozoan parasites was recently demonstrated when over 600 compounds were tested against Malaria, Sleeping Sickness, Leishmaniasis and Chagas disease [35].

Another point I would like to discuss is the large discrepancy between the number of conserved proteins among *Plasmodium* spp. detected in *Chapter 4* and the analysis from Hall et al. [36], where they mentioned an 'universal *Plasmodium* set' of about 4500 genes. In our approach, we identified about 3600 *P. falciparum* proteins that contained at least one ortholog in all of the other included *Plasmodium* species. In comparison to the analysis conducted by Hall et al., where the complete genomes of *P. falciparum*, *P. berghei*, *P. chabaudi* and *P. yoelii* were considered, we included the genomes of the primate-infective *P. knowlesi* and *P. vivax* as well. As the method to detect orthologs was similar, the differences need to have appeared as a result of incorporation of more genomic data. Hence this is a good example of how comparative genomics got more informative and precise by incorporation of additional reliable sequences.

The fact that the identified targets from our pipeline are by definition highly conserved among *Plasmodium* species holds the advantage that an eventual drug might be active against *P. vivax* as well, a highly desirable property in line with the Malaria Eradication Research Agenda (malERA) [37]. Moreover, as mouse models are crucial in drug development process, the conservation among rodent parasites is an important plus. Apart from efficacy and toxicity tests, rodent malaria parasites can also be used to obtain resistance markers before the introduction of new drugs to the market as well as for the evaluation of potential combination therapies [38].

For the filarial genome project of *D. immitis*, the prediction pipeline was modified due to the availability of a related model organism. The main difference was that essentiality was inferred from *C. elegans* homologs predicted to be essential for survival and development from RNAi analyses. As for *P. falciparum*, we identified proven targets as well as novel candidates. Although we have enriched for essential genes in the *S.*

*cerevisiae* test set and there is evidence for the *P. falciparum* predicted drug targets, in my opinion, it is more convenient to infer essentiality from functional genomics studies from closely related organisms, if data are complete and trustworthy. However, as mentioned above, related model organisms are frequently not available and thus I believe that our pipeline will advance the field of rational drug discovery against parasites in general.

In summary, I made use of the wealth of genomes and open-source programs available to learn more about the evolution of parasites and to identify vulnerable points. By comparative genomics, algorithms addressing specific biological questions were invented and automated tools based on existing programs and self-developed Perl scripts were built that are available to the scientific community.

## 3 Drug Resistance in African Trypanosomes

With the arriving of next generation sequencing technologies (NGS) [27], the opportunities for genomic research have been markedly increased. Whereas in the past the large sequencing centers appointed which species to sequence, nowadays even small research labs are able to sequence their organism of interest. In the frame of my PhD thesis, whole genome sequencing was performed to reveal mutations underlying drug resistance in African trypanosomes (*Chapter 6*). The power of this method lies in its high resolution as all genomic mutations can be rapidly identified at a time, if the produced sequences are of high quality. The study was built on a previous work at our institute, where the drug-sensitive *T. b. rhodesiense* STIB900 line was selected independently *in vitro* for melarsoprol (STIB900-M) and pentamidine (STIB900-P) resistance [39]. Due to the '10 Giga Grant' from Roche (Basel, CH), the genomes were sequenced on the 454 platform [40] by 454 Life Science (Branford, US). The company provided high quality reads for all three lines and a *de novo* assembly of STIB900. In a first step, the reads of each line were mapped to the reference genome *T. b. brucei* 927 [41]. Available programs were combined with Perl scripts to parse the outputs and automate the variant detection. About 80% of the reference genome was covered with at least 5 high quality bases by STIB900, STIB900-M or STIB900-P, respectively (*Chapter 6*, Table 1).

No reads mapped to the remaining parts of the genome mainly because they contained non-unique regions or they were of low complexity. The threshold coverage of 5 high quality bases was low, especially for a diploid organism. In my opinion, a good value would be of at least 15 as it would allow more precise differentiation between sequencing errors and true variants. A further issue was that the reference genome was a haploid mosaic of the diploid chromosomes. However, the *T. b. brucei* 927 is currently the best annotated genome that is publicly available. Instead of mapping reads to a reference genome, the usage of the *de novo* assembly of STIB900 would have been an option. With blastn, gene annotation transfer from *T. b. brucei* 927 to STIB900 would have been straightforward as *Trypanosoma brucei* genes typically have no introns [42]. However, the $N_{50}$ of the provided assembly was only 6 kb. $N_{50}$ represents 'the contig length at which 50% of the assembly span was in contigs of that length or greater' (*Chapter 5*, p. 69) and is often used as measurement for assembly quality [43]. In comparison, the $N_{50}$ of the 2.2 release of the *D. immitis* genome (nematodes.org/genomes/dirofilaria_immitis) was 38 kb. Moreover, the produced sequences contained errors in homopolymer stretches which meant that almost every single gene prediction would have needed to be manually corrected; hence we decided to map all the reads to the *T. b. brucei* reference genome.

By comparative genomics between the drug-sensitive STIB900 and its drug-resistant derivates STIB900-M and STIB900-P, new as well as proven variants were identified. Overall, more mutations might have been detected with a higher coverage, e.g. in combination with other NGS platforms such as Illumnia. Nevertheless, proof-of-principle was the detection of the loss of *TbAT1* and *TbAQP2*, two validated resistance factors [44,45]. Whether the loss of both genes is sufficient to explain the high-level of resistance in STIB900-M and STIB900-P, needs to be further investigated, e.g. by constructing a *tbat1*$^{-/-}$*aqp2*$^{-/-}$ line. If a double homozygous deletion cannot explain the high resistance factors, the coding mutation in the conserved RNA-binding motif of *TbUBP1* will be a very interesting candidate to follow up as the same mutation was independently fixed in both lines. Moreover, it needs to be resolved whether *TbAT1* in STIB900-P is still partly functional as it may explain the lower resistance factors of STIB900-P compared to STIB900-M.

Our approach focused on coding regions, although mutations in regulatory elements may have an effect on resistance as well. Therefore, a study of RNA expression levels was conducted with the spliced leader trapping protocol [47] and the results are currently analysed (Graf et al., unpublished). In addition, the status of the *AQP* locus is currently tested in *T. b. gambiense* field isolates of relapse patients (Graf et al., unpublished).

In summary, major efforts have been made recently to reveal new drug resistance candidates with new high-throughput methods. These approaches have brought the research community a step closer to the development of molecular genetic markers for monitoring drug resistance in the field. However, candidate genes need to be validated and tested in isolates of relapse patients. Moreover, it was shown that the handling of whole genome sequencing data is feasible and open-source programs can be implemented to detect mutations or deletions at the genomic level.

# Conclusion

Comparative genomics is a powerful discipline that has striking impacts on genomic research and the entire field of biology. A plethora of genomes from parasites, hosts and vectors is already available and the number is steadily increasing as new technologies reduces cost and time to sequence complete genomes. In the framework of my PhD thesis, algorithms were invented and automated tools were built that are widely applicable to parasites. First, a pipeline was developed that takes a whole parasite proteome and by comparing to the host and control species, it returns molecular mimicry candidates. The pipeline revealed promising epitopes from various parasites that may play crucial roles in host-parasite interactions. Second, an algorithm was developed that takes advantage of the species-specific nature of palindrome frequency patterns. The built tool is easy to use and able to discriminate DNA sequences to the level of species. Third, an *in silico* prediction was developed that translates a whole parasite proteome into a sizeable list of promising drug targets. 40 candidates were obtained from *P. falciparum* that may serve as a starting point for rational drug discovery. Further, as a part of an international genome project, the previously described tools were applied to the heartworm *D. immitis*. Moreover, whole genome sequencing of drug-resistant and drug-sensitive *T. b. rhodesiense* lines was performed. By comparative genomics, candidate genes were revealed that are involved in resistance to melarsoprol and pentamidine.

Taken together, it was shown that the wealth of available genomes offers new opportunities to study parasites at different levels. Comparative genomics significantly enhances our understanding of parasites and it accelerates the discovery of new drugs and vaccine candidates, that eventually lead to the control of infectious diseases. Further, it can be concluded that whole genome sequencing is an effective method to reveal new drug resistance candidates and the necessary data processing is feasible with open-source programs. Moreover, this thesis illustrated also limitations of genomic studies: experimental data are still indispensable to validate new findings and the inclusion of other 'omics' approaches may further augment our understanding. Nevertheless, comparative genomics is an emerging biological discipline which will certainly continue to play a pivotal role in parasitology.

## References

1. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270: 397–403.

2. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498–511. doi:10.1038/nature01097.

3. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. Science 309: 436–442. doi:10.1126/science.1112680.

4. Pain A, Böhme U, Berry AE, Mungall K, Finn RD, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455: 799–803. doi:10.1038/nature07306.

5. Tachibana S-I, Sullivan SA, Kawai S, Nakamura S, Kim HR, et al. (2012) *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat Genet. Available: http://www.ncbi.nlm.nih.gov/pubmed/22863735. Accessed 14 August 2012.

6. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, et al. (2007) Draft genome of the filarial nematode parasite *Brugia malayi*. Science 317: 1756–1760. doi:10.1126/science.1145406.

7. Jex AR, Liu S, Li B, Young ND, Hall RS, et al. (2011) *Ascaris suum* draft genome. Nature 479: 529–533. doi:10.1038/nature10553.

8. Ludin, Philipp (2009) Proteome-wide surveys for molecular mimicry in parasites. MSc Thesis.

9. Yoshimura A, Naka T, Kubo M (2007) SOCS proteins, cytokine signalling and immune regulation. Nat Rev Immunol 7: 454–465. doi:10.1038/nri2093.

10. Akhtar LN, Benveniste EN (2011) Viral exploitation of host SOCS protein functions. J Virol 85: 1912–1921. doi:10.1128/JVI.01857-10.

11. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785–786. doi:10.1038/nmeth.1701.

12. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S (2004) Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 17: 349–356. doi:10.1093/protein/gzh037.

13. Hewitson JP, Harcus YM, Curwen RS, Dowle AA, Atmadja AK, et al. (2008) The secretome of the filarial parasite, *Brugia malayi*: proteomic profile of adult excretory-secretory products. Mol Biochem Parasitol 160: 8–21. doi:10.1016/j.molbiopara.2008.02.007.

14. Bennuru S, Semnani R, Meng Z, Ribeiro JMC, Veenstra TD, et al. (2009) *Brugia malayi* excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling. PLoS Negl Trop Dis 3: e410. doi:10.1371/journal.pntd.0000410.

15. Bennuru S, Meng Z, Ribeiro JMC, Semnani RT, Ghedin E, et al. (2011) Stage-specific proteomic expression patterns of the human filarial parasite *Brugia malayi* and its endosymbiont *Wolbachia*. Proc Natl Acad Sci USA 108: 9649–9654. doi:10.1073/pnas.1011481108.

16. Choi Y-J, Ghedin E, Berriman M, McQuillan J, Holroyd N, et al. (2011) A deep sequencing approach to comparatively analyze the transcriptome of lifecycle stages of the filarial worm, *Brugia malayi*. PLoS Negl Trop Dis 5: e1409. doi:10.1371/journal.pntd.0001409.

17. Maizels RM, Balic A, Gomez-Escobar N, Nair M, Taylor MD, et al. (2004) Helminth parasites--masters of regulation. Immunol Rev 201: 89–116. doi:10.1111/j.0105-2896.2004.00191.x.

18. Preissner KT, Seiffert D (1998) Role of vitronectin and its receptors in haemostasis and vascular remodeling. Thromb Res 89: 1–21.

19. Schvartz I, Seger D, Shaltiel S (1999) Vitronectin. Int J Biochem Cell Biol 31: 539–544.

20. Singh B, Su Y-C, Riesbeck K (2010) Vitronectin in bacterial pathogenesis: a host protein used in complement escape and cellular invasion. Mol Microbiol 78: 545–560. doi:10.1111/j.1365-2958.2010.07373.x.

21. Kraemer SM, Smith JD (2006) A family affair: *var* genes, PfEMP1 binding, and malaria disease. Curr Opin Microbiol 9: 374–380. doi:10.1016/j.mib.2006.06.006.

22. Janes JH, Wang CP, Levin-Edens E, Vigan-Womas I, Guillotte M, et al. (2011) Investigating the host binding signature on the *Plasmodium falciparum* PfEMP1 protein family. PLoS Pathog 7: e1002032. doi:10.1371/journal.ppat.1002032.

23. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T (2010) *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. PLoS Comput Biol 6. doi:10.1371/journal.pcbi.1000933.

24. Wanderley JLM, Moreira MEC, Benjamin A, Bonomo AC, Barcinski MA (2006) Mimicry of apoptotic cells by exposing phosphatidylserine participates in the establishment of amastigotes of *Leishmania (L) amazonensis* in mammalian hosts. J Immunol 176: 1834–1839.

25. Manzano-Román R, Siles-Lucas M (2012) MicroRNAs in parasitic diseases: Potential for diagnosis and targeting. Mol Biochem Parasitol. doi:10.1016/j.molbiopara.2012.10.001.

26. Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol 5: e82. doi:10.1371/journal.pbio.0050082.

27. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31–46. doi:10.1038/nrg2626.

28. Bashir A, Klammer AA, Robins WP, Chin C-S, Webster D, et al. (2012) A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol 30: 701–707. doi:10.1038/nbt.2288.

29. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77. doi:10.1371/journal.pbio.0050077.

30. Dunning Hotopp JC (2011) Horizontal gene transfer between bacteria and animals. Trends Genet 27: 157–163. doi:10.1016/j.tig.2011.01.005.

31. Boto L (2010) Horizontal gene transfer in evolution: facts and challenges. Proc Biol Sci 277: 819–827. doi:10.1098/rspb.2009.1679.

32. Kumar S, Chaudhary K, Foster JM, Novelli JF, Zhang Y, et al. (2007) Mining predicted essential genes of *Brugia malayi* for nematode drug targets. PLoS ONE 2: e1189. doi:10.1371/journal.pone.0001189.

33. Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. BMC Genomics 11: 222. doi:10.1186/1471-2164-11-222.

34. Crowther GJ, Shanmugam D, Carmona SJ, Doyle MA, Hertz-Fowler C, et al. (2010) Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. PLoS Negl Trop Dis 4: e804. doi:10.1371/journal.pntd.0000804.

35. Witschel M, Rottmann M, Kaiser M, Brun R (2012) Agrochemicals against Malaria, Sleeping Sickness, Leishmaniasis and Chagas Disease. PLoS Negl Trop Dis 6: e1805. doi:10.1371/journal.pntd.0001805.

36. Hall N, Karras M, Raine JD, Carlton JM, Kooij TWA, et al. (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science 307: 82–86. doi:10.1126/science.1103717.

37. A research agenda for malaria eradication: drugs (2011). PLoS Med 8: e1000402. doi:10.1371/journal.pmed.1000402.

38. Hunt P, Martinelli A, Modrzynska K, Borges S, Creasey A, et al. (2010) Experimental evolution, genetic analysis and genome re-sequencing reveal the mutation conferring artemisinin resistance in an isogenic lineage of malaria parasites. BMC Genomics 11: 499. doi:10.1186/1471-2164-11-499.

39. Bernhard SC, Nerima B, Mäser P, Brun R (2007) Melarsoprol- and pentamidine-resistant *Trypanosoma brucei rhodesiense* populations and their cross-resistance. Int J Parasitol 37: 1443–1448. doi:10.1016/j.ijpara.2007.05.007.

40. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380. doi:10.1038/nature03959.

41. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The genome of the African trypanosome *Trypanosoma brucei*. Science 309: 416–422. doi:10.1126/science.1112642.

42. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GAM (2010) Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res 38: 4946–4957. doi:10.1093/nar/gkq237.

43. Zhang W, Chen J, Yang Y, Tang Y, Shang J, et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS ONE 6: e17915. doi:10.1371/journal.pone.0017915.

44. Mäser P, Sütterlin C, Kralli A, Kaminsky R (1999) A nucleoside transporter from *Trypanosoma brucei* involved in drug resistance. Science 285: 242–244.

45. Baker N, Glover L, Munday JC, Aguinaga Andrés D, Barrett MP, et al. (2012) Aquaglyceroporin 2 controls susceptibility to melarsoprol and pentamidine in African trypanosomes. Proc Natl Acad Sci USA 109: 10996–11001. doi:10.1073/pnas.1202885109.

46. Hartmann C, Benz C, Brems S, Ellis L, Luu V-D, et al. (2007) Small trypanosome RNA-binding proteins TbUBP1 and TbUBP2 influence expression of F-box protein mRNAs in bloodstream trypanosomes. Eukaryotic Cell 6: 1964–1978. doi:10.1128/EC.00279-07.

47. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, et al. (2010) Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. PLoS Pathog 6: e1001037. doi:10.1371/journal.ppat.1001037.

48. Alsford S, Eckert S, Baker N, Glover L, Sanchez-Flores A, et al. (2012) High-throughput decoding of antitrypanosomal drug efficacy and resistance. Nature 482: 232–236. doi:10.1038/nature10771.

# Curriculum Vitae

## Lectures

During my studies I attended lectures and courses given by

M. Altmann, K. Ammann, U. Baumann, R. Brun, S. Christen, B. Engelhardt, B. Erni, U. Feller, J. Frey, B. Gottstein, R. Häner, M. Hediger, J. Hulliger, H. Imboden, J. Kohli, C. Kuhlemeier, B. Lanzrein, P. Mäser, O. Mühlemann, W. Nentwig, D. Rentsch, B. Richner, I. Roditi, W. Salzburger, A. Schaller, A. Schneider, D. Schümperli, T. Seebeck, B. Stadler, B. Suter, M. Täuber