Bioinformatic strategies to provide functional clues to the unknown genes in Plasmodium falciparum genome

Isabelle Florent, Eric Maréchal, Olivier Gascuel, Laurent Brehelin

BIOINFORMATIC STRATEGIES TO PROVIDE FUNCTIONAL CLUES TO THE UNKNOWN GENES IN *PLASMODIUM FALCIPARUM* GENOME

FLORENT I.*, MARÉCHAL E.**, GASCUEL O.*** & BRÉHÉLIN L.***

Summary:

The fight against *Plasmodium falciparum*, the species responsible for 90 % of the lethal forms of human malaria, took a new direction with the publication of its genome in 2002. However, the hopes that the genome should help bringing to the foreground the expected new "vaccines candidates" or "targets of new medicines" were disappointed by the low number of genes that could be functionally annotated – less than 40 % upon the genome publication, just over 50 % eight years later. This 10 % gain of knowledge was made possible by the efforts of the entire scientific community in many directions which include: the production of transcriptomic and proteomic profiles at various stages of the parasite development and in response to drug or stress treatments; the proteomic study of subcellular compartments; the sequencing of numerous *Plasmodium* related species (allowing whole genome comparisons) and the sequencing of numerous *P. falciparum* strains (allowing investigations of gene polymorphism). In parallel with this production of experimental biological data, the development of original mining tools adapted to the *P. falciparum* specificities quickly appeared as a priority, as the performances of "classical" bioinformatic tools, used successfully for other genomes, had limited efficacy. This was the aim of the PlasmoExplore project launched in 2007. This brief review does not cover all efforts made by the international community to decipher the *P. falciparum* genome but focuses on improvements and novel mining methods investigated by the PlasmoExplore consortium, and some of the lessons we could learn from these efforts.

KEY WORDS: *Plasmodium*, genome, bioinformatics, annotations, novel mining methods.

Résumé : STRATÉGIES BIOINFORMATIQUES POUR PRÉDIRE LA FONCTION DES GÈNES NON-ANNOTÉS DU GÉNOME DE *PLASMODIUM FALCIPARUM*

La lutte contre Plasmodium falciparum, *l'espèce responsable de 90 % des formes mortelles du paludisme chez l'Homme, a pris une nouvelle direction avec la publication de son génome en 2002. Cependant, les espoirs de faire émerger de nouveaux « candidats de vaccins » ou des « cibles de nouveaux médicaments » ont été déçus par le faible nombre de gènes qui ont pu être annotés sur le plan fonctionnel – moins de 40 % au moment de la publication de son génome, juste au-delà de 50 % huit ans plus tard. Ce gain de 10 % de connaissances (parfois imprécises) résulte des efforts menés par toute la communauté scientifique dans plusieurs directions dont : la production de profils transcriptomiques et protéomiques au cours du développement du parasite et en réponse à des traitements médicamenteux ou stressants ; l'étude protéomique de compartiments subcellulaires ; le séquençage d'une diversité d'espèces proches du genre* Plasmodium *et dans le genre* Plasmodium *(permettant des comparaisons au niveau de génome entiers) et le séquençage de nombreuses souches de* P. falciparum *(permettant d'accéder au polymorphisme des gènes). En parallèle de cette production de données biologiques expérimentales, le développement d'outils bioinformatiques originaux adaptés aux spécificités de* P. falciparum *est rapidement apparu comme une priorité, face aux limitations rencontrées avec les outils classiques, utilisés pourtant avec succès sur d'autres génomes. C'est ce défi que le consortium PlasmoExplore, lancé en 2007, s'est proposé de relever. Cette revue ne couvre pas tous les efforts réalisés par la communauté internationale pour décrypter le génome de* P. falciparum, *mais se focalise sur les améliorations aux méthodes classiques et les développements de méthodes originales nouvelles que le consortium PlasmoExplore a mis en œuvre, ainsi que les leçons que l'on peut tirer de tels efforts.*

Mots clé : Plasmodium, *génome, bioinformatique, annotations, nouvelles méthodes de fouille de données.*

* Centre National de la Recherche Scientifique / Muséum National d'Histoire Naturelle, FRE3206, Molécules de Communication et Adaptation des Micro-organismes, Adaptation des Protozoaires à leur Environnent, CP 52, 61, rue Buffon, 75231 Paris Cedex 05, France.
** Commissariat à l'Énergie Atomique, Centre National de la Recherche Scientifique, Université Joseph Fourier Grenoble I, Institut National de la Recherche en Agronomie, Unité Mixte de Recherche 5168 ; Institut de Recherches en Technologies et Sciences pour le Vivant, Grenoble, 17, rue des Martyrs, 38054 Grenoble Cedex 09, France.
*** Méthodes et algorithmes pour la Bioinformatique, LIRMM, Univ. Montpellier 2, CNRS, 161, rue Ada, 34095 Montpellier Cedex 5, France.
Correspondence: Dr Isabelle Florent, FRE3206 CNRS/MNHN, APE, RDDM.
Tel.: +33 (0)1 40 79 35 47 - fax: +33 (0)1 40 79 34 99
E-mail: florent@mnhn.fr

INTRODUCTION

Malaria causative agents include *Plasmodium vivax*, *P. malariae*, *P. ovale*, *P. knowlesi* but *P. falciparum* is the species responsible for 90 % of the lethal forms (Greenwood, 2008). The massive efforts to understand the biology of *P. falciparum* and the etiology, pathology and diagnosis of malaria has led to a unique situation: while our knowledge of this disease is probably one of the most documented amongst infections caused by eukaryotic pathogens,

malaria remains one of the most burdensome (Greenwood, 2009). According to WHO, malaria is one of the most important sources of morbidity and mortality, with 300–500 million clinical cases and 1–2 million deaths annually (Greenwood, 2008). In addition, geographic distribution of endemic regions puts half of the world population at risk. The life cycle of *Plasmodium* species alternates between asexual proliferations in vertebrate hosts (in liver and red blood cells) and a sexual reproduction in an arthropod vector, most notably mosquitoes of the *Anopheles* genus for human parasites (Tuteja, 2007). The high rate of resistance outbreaks and spreading implies a constant need to search for novel prophylaxis and treatments, *i.e.* for the discovery of novel drug targets and antimalarials (Greenwood, 2008). Fight against malaria took a new direction with the publication of the *P. falciparum* and *P. yoelii* genomes in 2002 (Gardner *et al.*, 2002; Carlton *et al.*, 2002). One of the most expected benefits of genome sequencing projects is the capacity to allow the prediction of the complete proteome of a given organism. However, in the case of *P. falciparum*, the capacity to functionally annotate its proteins was profoundly limited by the important genetic distance of *P. falciparum* with the majority of the other species then sequenced (Doolittle, 2002). The extreme wealth of its genome in bases A+T (80 %) leads to a significant composition bias in the proteins coded by these genes (Bastien *et al.*, 2004; Bastien *et al.*, 2005). This and the experimental observation that the majority of its proteins are on average 20 % longer than the reference bodies (Aravind *et al.*, 2003) further complicated the annotation process.

Because a function (often putative, *i.e.* only based on the similarity with a sequence annotated in another organism) could only be proposed to less than 40 % of the genes, the major source of new target candidates for intervention (expected amongst the ~ 60 % genes devoid of annotations) remained unexplored. To face with this problem, several teams launched important post-genomic studies in order to explore the temporal and even spatio-temporal expression of the genes of this and related organisms. A wealth of genomic and post-genomic data has then been produced since the initial publications in 2002 (Carlton *et al.*, 2002; Florens *et al.*, 2002; Gardner *et al.*, 2002), providing a very important framework for data mining (Aurrecoechea *et al.*, 2009b; Winzeler, 2008). Currently, the *Plasmodium* genus is considered one of the best documented in terms of genomic and post-genomic data among eukaryotic pathogens (Birkholtz *et al.*, 2006; Birkholtz *et al.*, 2008).

CURRENT AND PERSPECTIVE DATA

Current status of genomic and post-genomic data for *Plasmodium* and related species

As of 2010, six *Plasmodium* species (*P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi* and *P. knowlesi*) have been sequenced and genomic data is also available for *P. gallinaceum* and *P. reichenowi*, opening an avenue for fruitfull comparative genomic analyses (Carlton *et al.*, 2008; Carlton *et al.*, 2002; Dechamps *et al.*, 2010; Gardner *et al.*, 2002; Hall *et al.*, 2005; Kooij *et al.*, 2005; Liew *et al.*, 2010; Pain *et al.*, 2008). In the case of *P. falciparum*, sequences of several strains along with sequences of massive collections of field samples available in the near future allow researchers to investigate the polymorphism and evolution of the parasite (Achidi *et al.*, 2008; Ahouidi *et al.*, 2010; Florent *et al.*, 2009; Jeffares *et al.*, 2007; Volkman *et al.*, 2007). Transcriptomic data, providing a global view of gene expression during the parasite's development (in human, in insects, *in vitro*) or following drug or stress treatments are now widely available for *P. falciparum* and related species (Birkholtz *et al.*, 2008; Bozdech *et al.*, 2003a; Clark *et al.*, 2008; Dahl *et al.*, 2006; Hall *et al.*, 2005; Hu *et al.*, 2010; Le Roch *et al.*, 2008; Le Roch *et al.*, 2003; Llinas *et al.*, 2006; Natalang *et al.*, 2008; Shock *et al.*, 2007; Tarun *et al.*, 2008; Winzeler, 2008; Young *et al.*, 2005). Proteomic data are also available. The first studies, dedicated to monitoring protein expression profiles in key developmental stages in *P. falciparum* and rodent models of malaria (Florens *et al.*, 2004; Florens *et al.*, 2002; Hall *et al.*, 2005; Khan *et al.*, 2005; Lasonder *et al.*, 2002; Lasonder *et al.*, 2008; Mair *et al.*, 2006; Tarun *et al.*, 2008) were completed by various proteomic analyses of diverse sub-cellular compartments of the parasite or the infected host cells, such as the parasitophorous vacuole (Nyalwidhe & Lingelbach, 2006), food vacuole (Lamarque *et al.*, 2008), invasive organelles (Lal *et al.*, 2009; Sam-Yellowe *et al.*, 2008; Sam-Yellowe *et al.*, 2004) or even rafts (Sanders *et al.*, 2007). Proteomic data from drug treated parasites also start to emerge (Radfar *et al.*, 2008). Finally, interactome data have been either produced from experimental investigations (LaCount *et al.*, 2005) or predicted from theoretical models (Date & Stoeckert, 2006).

This important source of genomic and post-genomic data and information thus provides virtually infinite ways for data mining toward the identification of biologically relevant targets. Most data have been collected and integrated into the PlasmoDB database (Aurrecoechea *et al.*, 2009b; Stoeckert *et al.*, 2006). The web interface of PlasmoDB provides several tools to query the genomic and post-genomic data by keywords, gene IDs, sequence alignments, functional annotations, etc.,

as well as to formulate complex queries by Boolean combinations of simple ones (Aurrecoechea *et al.*, 2009b). In addition, a similar web portal called EuPathDB allows the combined access to genomic and post-genomic data from major eukaryotic pathogens, evolutionarily related (*Cryptosporidium*, *Neospora*, *Toxoplasma*) or not (*Encephalitozoon*, *Entamoeba*, *Giardia*, *Leishmania*, *Trichomonas* and *Trypanosoma*) to *Plasmodium* (Aurrecoechea *et al.*, 2009a). Hence, upon rationally designed queries on specific biological or metabolic processes, lists of genes that are important for pathology, invasion, replication or any relevant question may easily be produced. Lists of genes that encode proteins with no homologue in the human genome (an important feature to lower toxicity risks) and with little allelic variation (to lower risks of resistance spreading) may even be produced. The recent identification of the proteins involved in the export machinery in *P. falciparum* beautifully illustrates how appropriate data mining, combined with experimental validation, allows major breakthrough in the understanding of basic mechanisms of the parasite biology (de Koning-Ward *et al.*, 2009). Nevertheless, such lists highly depend on the accuracy of the scientific question, the quality of the experimental data and also, the quality of the annotations associated with these data (Saidani *et al.*, 2009).

MOVING THE CURSOR?

Despite the very large amount of genomic and post-genomic data available for the *Plasmodium* genus (Aurrecoechea *et al.*, 2009b) and related pathogens (Aurrecoechea *et al.*, 2009a), a major limitation in this field still remains the relative paucity of functional annotations attached to the ~ 5,400 *P. falciparum* genes and their orthologs in other *Plasmodium* species. Since the 60 % hypothetical genes of the genome publication in 2002 (Gardner *et al.*, 2002), the cursor separating unannotated and annotated genes has shifted slowly. The proportion remained very stable for a couple of years (Birkholtz *et al.*, 2006), falling to 57 % only in 2008 and 47 % in 2009 (see PlasmoDB, www.plasmodb.org). Importantly, the amount of functional information of the annotated genes is often very low, limited to the detection of conserved protein domains that provide little clues to the possible function. Thus, beside the production of new data, a part of the scientific community has engaged into the search of novel *in silico* methods to improve the annotation status of the *P. falciparum* genome.

Launched in 2007, at a date where 60 % genes still remained hypothetical, the PlasmoExplore consortium (supported by the French National Agency for Research, ANR) has combined the expertise of computer scientists in algorithmics and bioinformatics, experts in evolutionary sciences and biologists to investigate how provisional functional annotation could be proposed, and possibly validated, for all these genes. Various strategies were proposed by this consortium so as to help ascribing novel potential functions to previously "unknown" proteins, thus providing important insights for drug and vaccine target discovery. These methods are described and discussed hereafter. We first present sequence-based methods, including their limited use in the field of malaria and how searching for domain co-occurrence in plasmodial proteins helped uncovering 585 new domains and 387 new Gene Ontology annotations in the *P. falciparum* proteins (Terrapon *et al.*, 2009). We then present our investigations in the field of non-homology methods based on the Guilt By Association principle which, combined with *P. falciparum* post-genomic data and Gene Ontology allowed to compute functional predictions for all *P. falciparum* proteins within PlasmoDraft (Brehelin *et al.*, 2008). Finally, we present a method combining homology and non-homology based approaches to exploit conserved co-expressed genes between *P. falciparum*, *Saccharomyces cerevisiae* and *Drosophila melanogaster*, in order to propose annotation transfers for several dozens of genes for which homology scores were not sufficient (Bréhélin *et al.*, 2010).

EXPLORING SEQUENCE-BASED METHODS: THE NECESSITY TO ADAPT CLASSICAL APPROACHES TO PECULIARITIES OF THE *P. FALCIPARUM* GENOME

Annotation of genomes is primarily based on homology-based transfers. A working hypothesis was that classical sequence alignment and domain detection methods failed to accurately detect homologies in the *Plasmodium* genome. This working hypothesis had been previously investigated by (Bastien *et al.*, 2005), based on the observation that the *P. falciparum* genome was compositionally biased (*i.e.* AT rich) and that this bias was a problem for alignments with unbiased reference sequences. The bias is however heterogeneous along the genome:

- intergenic regions are more AT-rich (> 90 %);
- intraprotein low complexity segments, supposed not important for the function, are more AT-rich and more biased at amino acid level;
- several proteins are not compositionally biased and very well detected by alignment methods.

Although some proteins are believed to be really specific to the *P. falciparum* species, the *Plasmodium* genus or the Apicomplexan phylum, other proteins have likely diverged from their ancestor, following the compositional bias constraint in such a manner that they are not detected any more by using classical methods. How could we correct these classical methods to detect similarity?

On one hand, sequence similarity may be detected based on *sequence alignments*, using methods like BLAST/FASTA (Altschul *et al.*, 1990; Lipman & Pearson, 1985). On the other hand, *models of sequence families,* based on multiple alignments capturing subtle features of the homology thanks to the diversity of its members, can also be computed. Hidden Markov Models (HMMs) are widely utilized for this purpose and allow modelling entire proteins, or specific protein domains (Durbin *et al.*, 1998).

Concerning sequence alignment, the possibility to correct the substitution matrices used for alignments was envisaged, with adjustments leading to non-symmetrical matrices (Bastien *et al.*, 2005). The statistics for biased *vs* unbiased alignments have also been examined (Bastien & Marechal, 2008).

A novel track explored by the consortium concerned the methods for the detection of protein domains (Terrapon *et al.*, 2009). Protein domains are sequential and structural motifs that are found independently in different proteins and in different combinations and that can be viewed as functional sub-units of proteins. The objective was to improve the sensitivity of HMM domain detection by exploiting the tendency of the domains to appear preferentially with a few other favourite domains in multi-domain proteins. When sequence similarity alone is not sufficient to warrant the presence of a particular domain in a multi-domain protein, this approach can enable its detection on the basis of the co-occurrence of another domain in the same protein. Figure 1 presents one example of such a new domain discovery in the *P. falciparum* protein PF08_0070. While classical methods do detect the RAP domain in this protein and do not have enough signal to detect the FAST-1 domain, the co-occurrence algorithm validates the presence of this FAST-1 domain in PF08_0070 based on its computed co-occurrence with the RAP domain in proteins. Performance of detection of novel domains was effectively improved and several hundreds of new domains (585) were discovered, allowing proposing new functional annotations in the corresponding proteins (Terrapon *et al.*, 2009). A database of these newly discovered domains was made accessible for *P. falciparum*, *P. vivax* and *P. yoelii*, and was next extended to several other eukariotic pathogenes (EuPathDomains: http:/www.atgc-montpellier.fr/EuPathDomains/) (Ghouila *et al.*, 2010).

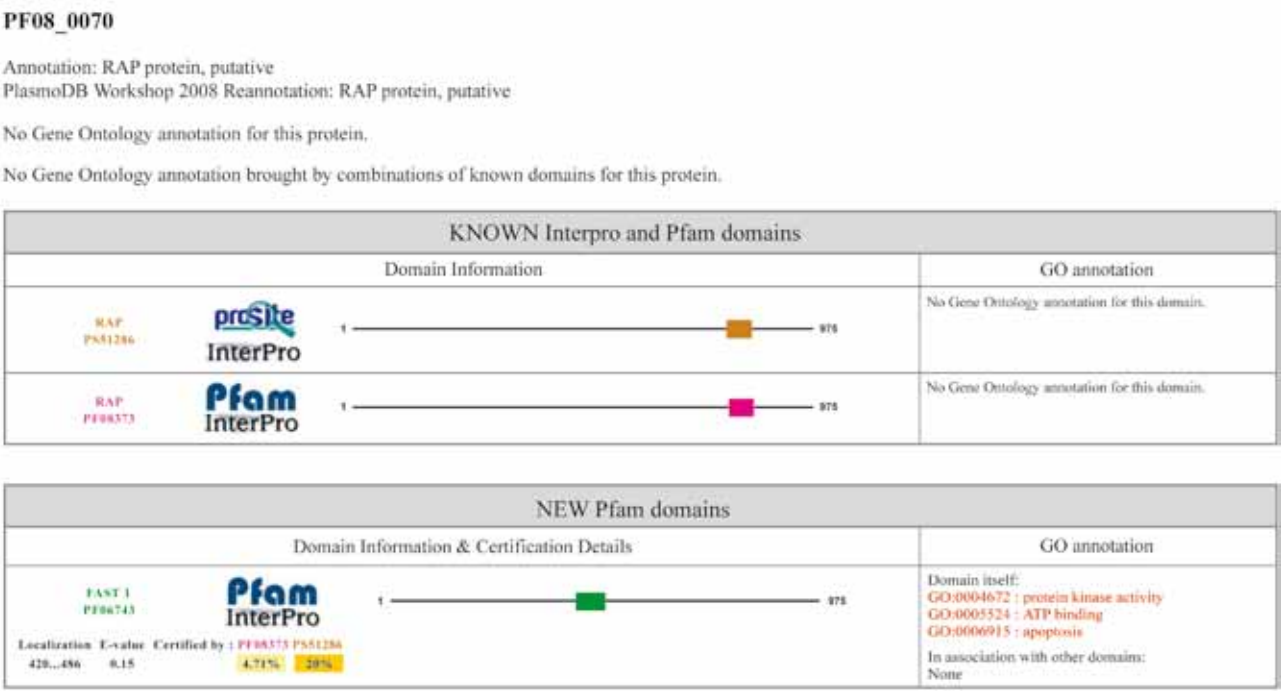


Fig. 1. – Illustration of domain detection by using Co-occurrent Domain Discovery (http:/www.atgc-montpellier.fr/EuPathDomains/). This example illustrates the discovery of a new domain in the RAP protein (PF08_0070)**.** The upper panel indicates known domains detected in the protein by classical Interpro scan (http://www.ebi.ac.uk/interpro/), and already annotated for the query protein. The lower panel indicates the new domains discovered in this protein by using the co-occurrence algorithm. In this example, the RAP domain (PS51286 and PF08373 respectively in Prosite/Interpro and Pfam/Interpro databases) is a known domain of PF08_0070 while the FAST-1 domain has not been identified as true domain of PF08_0070 by the Interpro scan because the E-value of 0.15 is above the confidence threshold. However, the co-occurrence algorithm, by knowing that RAP and FAST-1 domains are often found together in proteins, strongly suggests that, in spite of individual E-value of 0.15 above confidence threshold, the FAST-1 domain has a high probability (estimated error < 4.71 %) to be present in the PF08_0070 protein. In addition to previously known and new protein domains in query *P. falciparum* proteins, the co-occurrence interface displays the localizations of the different domains and the GO terms associated with the newly discovered domain (GO:0004672, GO:0005524 and GO:0006915 in this example).

Exploring non-homology based methods: gene annotation transfers based on Guilt By Association principle and Gene Ontology

Another possibility to cope with the difficult problem of homology detection is to use methods based on post-genomic data to obtain functional clues for the uncharacterized genes. These are commonly called Guilt By Association (GBA) methods. Contrary to sequence homology which involves inter-species annotation transfers, *i.e.* genes characterized in other species are used to annotate genes of the newly sequenced genome, GBA approaches involve intra-species annotation transfers: the genes already characterized in the genome, *e.g.* by wet experiments or using sequence homology, are used for the annotation of the other genes (Guilt By Association principle). Gene expression data are often used, since genes with similar transcriptomic profiles likely share common functional roles (Eisen *et al.*, 1998). This approach was first proposed for *P. falciparum* by (Le Roch *et al.*, 2003), and next carried out in an extensive way by the PlasmoExplore consortium, with the creation of the PlasmoDraft database (Bréhélin *et al.*, 2008).

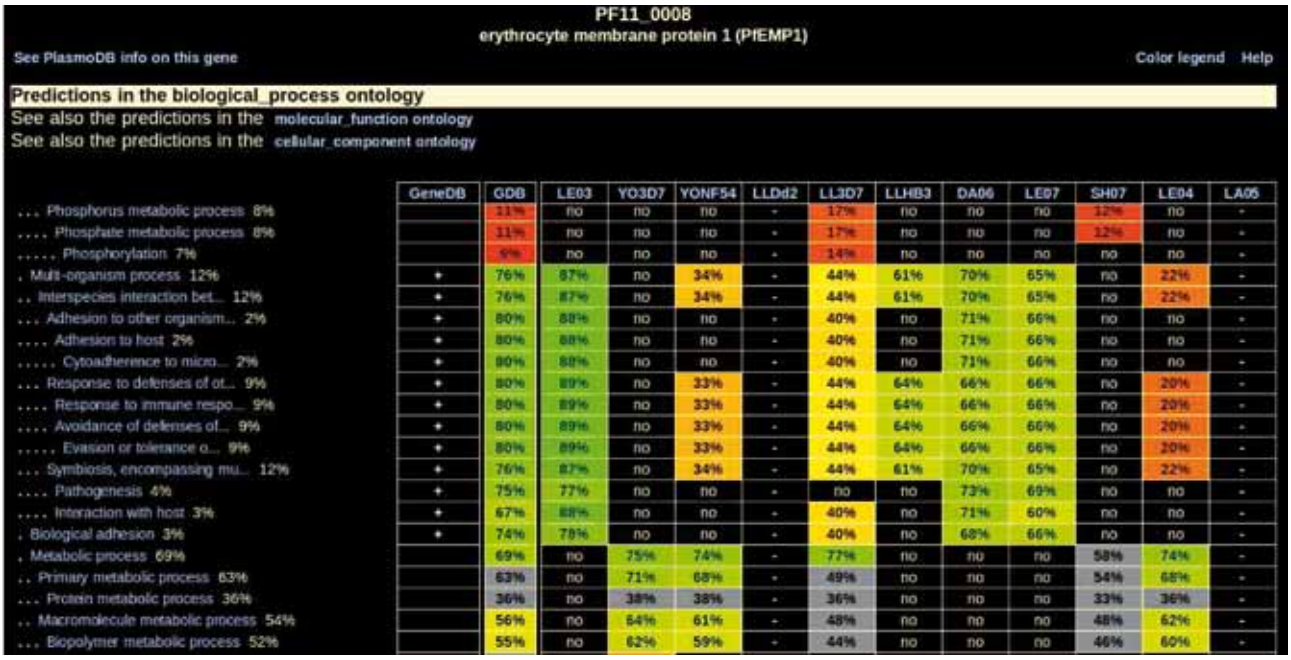
PlasmoDraft (http://www.atgc-montpellier.fr/PlasmoDraft/) is a database of Gene Ontology (GO) annotation predictions for the genes of *P. falciparum*. It involves both the ~ 60 % of genes that lack GO annotation, and the genes that are already annotated in GeneDB (http://www.genedb.org, Sanger Institute, GB). Predictions of PlasmoDraft are based on transcriptome, proteome and interactome data, and are thus complementary to predictions achieved by sequence homology. In Plasmodraft, transcriptomic data cover the life and the intraerytrocytic cycles (Bozdech *et al.*, 2003a; Le Roch *et al.*, 2008; Llinas *et al.*, 2006), the sexual stage (Young *et al.*, 2005) and several drug treated parasites (Dahl *et al.*, 2006; Le Roch *et al.*, 2008; Shock *et al.*, 2007). Data from one proteome (Florens *et al.*, 2002) and one interactome (LaCount *et al.*, 2005) studies were also taken into account.

PlasmoDraft is based on the predictions of a GBA predictor named Gonna that proposes GO annotations for a gene, according to the similarity of its profile (measured on transcriptome, proteome, or interactome data) with those of genes already annotated in GO by GeneDB (Bréhélin *et al.*, 2008). Moreover, Gonna provides an estimate of the confidence of its prediction. Gonna has been applied to most post-genomic data sources publicly available for *P. falciparum*, and all predictions along with their confidences have been compiled into PlasmoDraft (see Fig. 2 for an example). Moreover, for each prediction, a global degree of belief computed by combining the different data sources is also provided (column GDB on Fig. 2). The database can be accessed in different ways. A global view allows for a quick inspection of the GO terms that are predicted with high confidence, depending on the various data sources. Moreover, a gene view and a GO term view allow for the search of potential GO terms attached to a given gene, and genes that potentially belong to a given GO term (Bréhélin *et al.*, 2008). PlasmoDraft therefore presents the predictions achieved by all data sources in a friendly way, accompanied with error estimates. For example, 2,434 genes without any annotations in the Biological Process ontology were associated, through PlasmoDraft, with specific GO terms (*e.g.* Rosetting, Antigenic variation), and among these, 841 have confidence values above 50 %. In the Cellular Component and Molecular Function ontologies, 1,905 and 1,540 uncharacterized genes were associated with specific GO terms, respectively (740 and 329 with confidence value above 50 %).

This effort of the PlasmoExplore consortium completes that of the group of E. Winzeler at the Scripps Research Institute, who used a specifically developed method – Ontology-based Pattern Identification, OPI – (Zhou *et al.*, 2008)) to propose GO functional predictions on the basis of a single new transcriptomic dataset covering all life cycle stages of the parasite and combining *P. yoelii* and *P. falciparum* gene expression data (http://www.scripps.edu/cb/winzeler/resources.htm). Importantly, both PlasmoDraft and the OPI predictions are now indexed in PlasmoDB (release 6.0).

Combining homology and non-homology methods: gene annotation transfers based on inter-species conserved coexpression

Another strategy to address the homology issue entailed by the divergence of *P. falciparum* proteins was to use gene expression data to increase the sensitivity of homology methods. This way, the gene expression data are not used to propose intra-species annotation transfers as in GBA methods, but to assess the inter-species transfers when the homology is doubtful. The approach starts from a set of likely orthologs between a reference species (for example *S. cerevisiae*) and *P. falciparum,* and two microarray data sets that monitor the level of expression of the genes of the two species. The *coexpression context* of a gene in one of the species is defined as the set of genes that appear to be coexpressed with this gene in the microarray data of the species. Then, if two homologs with potentially weak sequence similarity have similar coexpression contexts – *I.e.*, if their coexpression contexts share a sufficiently high number of orthologs – they likely have the same function (Fig. 3, Table I). The approach was used with Yeast and Drosophila and enabled us to propose with high confidence new/refined annotations for several dozens hypothetical/putative *P. falciparum* genes. These annotations were integrated into the PlasmoDB



	GeneDB	GDB	LE03	YO3D7	YONF54	LLDd2	LL3D7	LLHB3	DA06	LE07	SH07	LE04	LA05
... Phosphorus metabolic process 8%		11%	no	no	no	-	17%	no	no	no	12%	no	-
.... Phosphate metabolic process 8%		11%	no	no	no	-	17%	no	no	no	12%	no	-
..... Phosphorylation 7%		9%	no	no	no	-	14%	no	no	no	no	no	-
. Multi-organism process 12%	+	76%	87%	no	34%	-	44%	61%	70%	65%	no	22%	-
.. Interspecies interaction bet... 12%	+	76%	87%	no	34%	-	44%	61%	70%	65%	no	22%	-
... Adhesion to other organism... 2%	+	80%	88%	no	no	-	40%	no	71%	66%	no	no	-
.... Adhesion to host 2%	+	80%	88%	no	no	-	40%	no	71%	66%	no	no	-
..... Cytoadherence to micro... 2%	+	80%	88%	no	no	-	40%	no	71%	66%	no	no	-
... Response to defenses of ot... 9%	+	80%	89%	no	33%	-	44%	64%	66%	66%	no	20%	-
.... Response to immune respo... 9%	+	80%	89%	no	33%	-	44%	64%	66%	66%	no	20%	-
.... Avoidance of defenses of... 9%	+	80%	89%	no	33%	-	44%	64%	66%	66%	no	20%	-
..... Evasion or tolerance o... 9%	+	80%	89%	no	33%	-	44%	64%	66%	66%	no	20%	-
... Symbiosis, encompassing mu... 12%	+	76%	87%	no	34%	-	44%	61%	70%	65%	no	22%	-
.... Pathogenesis 4%	+	75%	77%	no	no	-	no	no	73%	69%	no	no	-
.... Interaction with host 3%	+	67%	88%	no	no	-	40%	no	71%	60%	no	no	-
. Biological adhesion 3%	+	74%	78%	no	no	-	40%	no	68%	66%	no	no	-
. Metabolic process 69%		69%	no	75%	74%	-	77%	no	no	no	58%	74%	-
.. Primary metabolic process 63%		63%	no	71%	68%	-	69%	no	no	no	54%	68%	-
... Protein metabolic process 36%		36%	no	38%	38%	-	36%	no	no	no	33%	36%	-
.. Macromolecule metabolic process 54%		56%	no	64%	61%	-	48%	no	no	no	48%	62%	-
... Biopolymer metabolic process 52%		55%	no	62%	59%	-	44%	no	no	no	46%	60%	-

Fig. 2. – Example of functional predictions achieved by PlasmoDraft.
A view of the predictions achieved for gene PF11_0008 (PfEMP1), in the biological_process ontology, using PlasmoDraft. GO terms are shown in a hierarchical way following the ontology structure (left side). Each term is followed by a number corresponding to its prior probability, which is the proportion of all *P. falciparum* genes annotated by the term in GeneDB (http://www.geneDB.org/). For example, 2 % of *the P. falciparum* genes are annotated "Adhesion to host". The first column of the table indicates whether the query gene (PF11_0008) is currently annotated (+) or not (no sign) with the corresponding GO term in the GeneDB. The 12 next columns correspond each to a given post-genomic data set: e.g. LE03 refers to the data published in (Le Roch *et al.*, 2003), LLDd2, LL3D7, LLHB3 refer to the data published in (Llinas *et al.*, 2006) (see (Bréhélin *et al.*, 2008) or http://www.atgc-montpellier.fr/PlasmoDraft/ for the definition of the other datasets). GDB corresponds to the global degree of belief computed by considering all datasets. Percentages in the table reflect the probability of success of the predictions, estimated by cross-validation for each data source, and a colour code is used to allow quick visualization: green for high probability, orange/yellow for medium probability and red for low or very low probability. Coloured entries therefore support the prediction with various probabilities while black cells do not support (« no » entries) the prediction. The « - » entries mean that no data are available in the source for this gene. This example illustrates that the gene PF11_008 is predicted by PlasmoDraft to be annotated with all terms between "Multi-organism process" and "Biological adhesion" by several series of experimental data (LE03, LLHB3 in parts, DA06, LE07) and with very high probability.

database and allowed, in particular, the most complete inventory of ribosomal proteins (GO:003735), an inventory of *Plasmodium* proteins involved in ribosomal biogenesis and assembly (GO:0042254), rRNA metabolic pathway (GO:0016072) and tRNA processing (GO:0008033) (http://www.lirmm.fr/~brehelin/co-coexpression/) (Bréhélin *et al.,* 2010).

FUTURE CHALLENGES

Lessons from these efforts can be summarized briefly: although it is difficult to move the cursor between unknown and functionally annotated genes in *Plasmodium*, it is possible, and the contribution of novel bioinformatic strategies to this gain of knowledge (~ 10 % these last years) has proven essential.

The PlasmoExplore consortium, in parallel with other groups worldwide, has thus put a substantial effort to improve available methods to compare *Plasmodium* genomic and post-genomic data with those of other better characterized organisms, and to exploit these comparisons to propagate annotations from genes of known function to *P. falciparum* genes of unknown functions. In parallel, the consortium had the concern to propose the provisional annotations in specific databases, like the PlasmoDraft and EuPathDomains, and to transfer those that have been curated by biologist experts to the PlasmoDB database (see Table II for an inventory of relevant web sites for these databases). A future challenge will be of course to validate experimentally as many of the predictions as possible.

The benefits of *in silico* approaches has however the potency to go far beyond functional predictions of hypothetical genes, although this is a critical problem in the case of malaria. *In silico* experiments can also provide additional information to help selecting in annotated genes, those which could be the most potent target candidates for therapeutic interventions. For example,

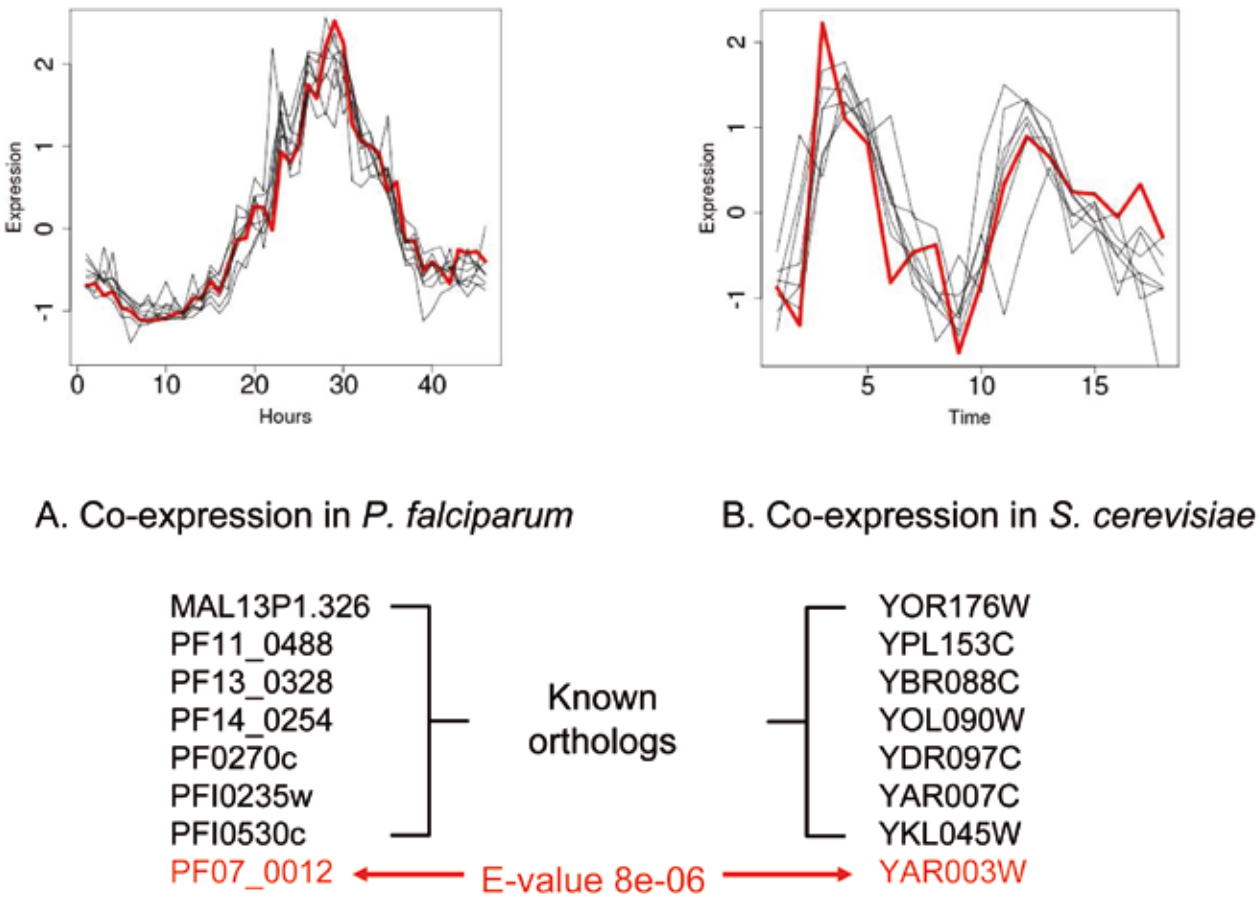


Fig. 3. - Rationale for annotation transfer using conserved co-expression between unrelated species.
In this example, a cluster of seven *P. falciparum* genes (black lines) with known orthologs in *S. cerevisiae* are co-expressed both in the intraerythrocytic cycle of *P.falciparum* ((Bozdech *et al.*, 2003b): **A**) and in the yeast cell cycle ((Spellman *et al.*, 1998): **B**). Two genes with uncertain homology also show similar expression profiles (red curves). These two genes are PF07_0012 and YAR003W. According to BLAST, PF07_0012 has two potential homologs in yeast: YAR003W (e-value 8. 10-6) and YOR195W (e-value 4.10-4). YAR003W is a subunit of the COMPASS (Set1C) complex, which methylates histone H3 on lysine 4 and is required in transcriptional silencing near telomeres, while YOR195W is a Kinetochore-associated protein required for normal segregation of chromosomes in meiosis and mitosis. Hence, the apparent conservation of coexpression between PF07_0012 and YAR003W, allows us to transfer the annotation of YAR003W to PF07_0012. See Table I for PlasmoDB and SGD (Saccharomyces Genome database) current annotations for these 16 genes.

in silico screening (docking) can provide useful information for identifying drug candidates (de Beer *et al.*, 2009) The Discovery database (http://malport.bi.up.ac.za:8150/) was developed to help mining the malaria genome using drug candidate structures (Joubert *et al.*, 2009). Another important information lies in the prediction of subcellular localization of proteins, since this provides important clues regarding pathogenesis in *Plasmodium*. The GBA principle can also be used for this purpose, because expression profiles of genes that are specific to a given organelle also display similar expression patterns. The PlasmoDraft predictions relative to the Cellular Component part of the Gene Ontology constitutes a first step in this direction.

The GBA principle has been widely applied to *P. falciparum* (by the PlasmoExplore consortium and others) and provided functional clues for many uncharacterized genes. However, it is worth noting that, depending on the type of function (or protein localisation), the accuracy of the approach can be low. Llinás and del Portillo warned against using only GBA methods, showing that many genes that were involved in absolutely unrelated functions, showed close expression profiles during the asexual development of *P. falciparum* (Llinas & del Portillo, 2005). Hence, a fundamental understanding of how parasite genes are regulated is critical to go one step forward, and to develop novel therapeutic strategies against this organism. Apart from a few specific genes, we remain largely ignorant of the mechanisms underlying gene expression control, and, more specifically, on the relative role of transcriptional and post-transcriptional regulation in the parasite. New insights

PlasmoDB code	Current PlasmoDB annotation	SGD code	Current SGD annotation
MAL13P1.326	Ferrochelatase, putative	YOR176W	Ferrochelatase, a mitochondrial inner membrane protein, catalyzes the insertion of ferrous iron into protoporphyrin IX, the eighth and final step in the heme biosynthetic pathway
PF11_0488	Ser/Thr protein kinase	YPL153C	Protein kinase, required for cell-cycle arrest in response to DNA damage; activated by trans autophosphorylation when interacting with hyperphosphorylated Rad9p; also interacts with ARS1 and plays a role in initiation of DNA replication
PF13_0328	Proliferating cell nuclear antigen	YBR088C	Proliferating cell nuclear antigen (PCNA), functions as the sliding clamp for DNA polymerase delta; may function as a docking site for other proteins required for mitotic and meiotic chromosomal DNA replication and for DNA repair
PF14_0254	DNA missmatch repair protein MSH2p, putative	YOL090W	MutS Homolog, Protein that forms heterodimers with Msh3p and Msh6p that bind to DNA mismatches to initiate the mismatch repair process; contains a Walker ATP-binding motif required for repair activity; Msh2p-Msh6p binds to and hydrolyzes ATP
PF0270c	DNA repair protein, putative	YDR097C	MutS Homolog, Protein required for mismatch repair in mitosis and meiosis, forms a complex with Msh2p to repair both single-base & insertion-deletion mispairs; potentially phosphorylated by Cdc28p
PFI0235w	Replication factor A-related protein, putative	YAR007C	Replication factor A, Subunit of heterotrimeric Replication Protein A (RPA), which is a highly conserved single-stranded DNA binding protein involved in DNA replication, repair, and recombination
PFI0530c	DNA primase large subunit, putative	YKL045W	DNA primase, Subunit of DNA primase, which is required for DNA synthesis and double-strand break repair
PF07_0012	Conserved unknown function	YAR003W	Set1c, WD40 repeat protein, Subunit of the COMPASS (Set1C) complex, which methylates histone H3 on lysine 4 and is required in transcriptional silencing near telomeres; WD40 beta propeller superfamily member with similarity to mammalian Rbbp7

Table I. – PlasmoDB and SGD codes and annotations for 16 genes presented on Fig. 3.
SGD stands for *Saccharomyces* Genomic Database (http://www.yeastgenome.org/)

General repositories of genomic and post-genomic data for malaria	
EupathDB	http://eupathdb.org/eupathdb/
PlasmoDB	http://plasmodb.org/plasmo/
GeneDB	http://www.genedb.org/
Specific websites for novel predictions	
PlasmoExplore	http://www.lirmm.fr/~brehelin/PlasmoExplore/
EuPathDomains	http:/www.atgc-montpellier.fr/EuPathDomains/
PlasmoDraft	http://www.atgc-montpellier.fr/PlasmoDraft/
Coexpression	http://www.lirmm.fr/~brehelin/co-coexpression/
OPI DB	http://chemlims.com/OPI21/MServlet.ChemInfo
Specific database combining malaria protein and putative antimalarial drug queries	
Discovery DB	http://malport.bi.up.ac.za:8150/

Table II. – List of web sites mentioned in this article (See text for details).

into the complex processes that regulate parasite proliferation, development and differentiation in its different host tissues are on the malaria research agenda. Elucidating the genes involved in key functions such as transmission success, immune evasion, drug resistance, as well as the global understanding of how gene expression is altered during the host-parasite interaction, are expected to provide promising new drug targets. In addition, this would shed light on the mode of action of known drugs, and help understanding the *P. falciparum* resistance mechanisms. *In silico* analysis of gene expression mechanisms represents therefore one of the next important challenges in malaria bioinformatics.

ACKNOWLEDGEMENTS

This work was made possible by financial support of the Agence Nationale de la Recherche, PlasmoExplore project ANR-06-CIS6-MDCA-14-01, initiated and headed by OG and bringing together the authors of this review and their research teams.

REFERENCES

ACHIDI E.A. *ET AL.* A global network for investigating the genomic epidemiology of malaria. *Nature*, 2008, *456*, 732-737.

AHOUIDI A.D., BEI A.K., NEAFSEY D.E., SARR O., VOLKMAN S., MILNER D., COX-SINGH J., FERREIRA M.U., NDIR O., PREMJI Z., MBOUP S. & DURAISINGH M.T. Population genetic analysis of large sequence polymorphisms in *Plasmodium falciparum* blood-stage antigens. *Infect. Genet. Evol.*, 2010, *10*, 200-206.

ALTSCHUL S.F., GISH W., MILLER W., MYERS E.W. & LIPMAN D.J. Basic local alignment search tool. *J. Mol. Biol.*, 1990, *215*, 403-410.

ARAVIND L., IYER L.M., WELLEMS T.E. & MILLER L.H. *Plasmodium* biology: genomic gleanings. *Cell*, 2003, *115*, 771-785.

AURRECOECHEA C., BRESTELLI J., BRUNK B.P., CARLTON J.M., 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., MORRISON H.G., NAYAK V., PENNINGTON C., PINNEY D.F., ROOS D.S., ROSS C., STOECKERT C.J., JR., SULLIVAN S., TREATMAN C. & WANG H. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res.*, 2009a, *37*, D526-530.

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 C.J., JR., TREATMAN C. & WANG H. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.*, 2009b, *37*, D539-543.

BASTIEN O., LESPINATS S., ROY S., METAYER K., FERTIL B., CODANI J.J. & MARECHAL E. Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using *Arabidopsis thaliana* as a reference. *Gene*, 2004, *336*, 163-173.

BASTIEN O. & MARECHAL E. Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores. *BMC Bioinformatics*, 2008, *9*, 332.

BASTIEN O., ROY S. & MARECHAL E. Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. *C. R. Biol.*, 2005, *328*, 445-453.

BIRKHOLTZ L., VAN BRUMMELEN A.C., CLARK K., NIEMAND J., MARECHAL E., LLINAS M. & LOUW A.I. Exploring functional genomics for drug target and therapeutics discovery in Plasmodia. *Acta Trop.*, 2008, *105*, 113-123.

BIRKHOLTZ L.M., BASTIEN O., WELLS G., GRANDO D., JOUBERT F., KASAM V., ZIMMERMANN M., ORTET P., JACQ N., SAIDANI N., ROY S., HOFMANN-APITIUS M., BRETON V., LOUW A.I. & MARECHAL E. Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space? *Malar. J.*, 2006, *5*, 110.

BOZDECH Z., LLINAS M., PULLIAM B.L., WONG E.D., ZHU J. & DERISI J.L. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.*, 2003a, *1*, E5.

BOZDECH Z., ZHU J., JOACHIMIAK M.P., COHEN F.E., PULLIAM B. & DERISI J.L. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol.*, 2003b, 4.

BREHELIN L., DUFAYARD J.F. & GASCUEL O. PlasmoDraft: a database of *Plasmodium falciparum* gene function predictions based on postgenomic data. *BMC Bioinformatics*, 2008, *9*, 440.

BREHELIN L., FLORENT I., GASCUEL O. & MARECHAL E. Assessing functional annotation transfers with inter-species conserved coexpression: application to *Plasmodium falciparum*. *BMC Genomics*, 2010, *11*, 35.

CARLTON J.M., ADAMS J.H., SILVA J.C., BIDWELL S.L., LORENZI H., CALER E., CRABTREE J., ANGIUOLI S.V., MERINO E.F., AMEDEO P., CHENG Q., COULSON R.M., CRABB B.S., DEL PORTILLO H.A., ESSIEN K., FELDBLYUM T.V., FERNANDEZ-BECERRA C., GILSON P.R., GUEYE A.H., GUO X., KANG'A S., KOOIJ T.W., KORSINCZKY M., MEYER E.V., NENE V., PAULSEN I., WHITE O., RALPH S.A., REN Q., SARGEANT T.J., SALZBERG S.L., STOECKERT C.J., SULLIVAN S.A., YAMAMOTO M.M., HOFFMAN S.L., WORTMAN J.R., GARDNER M.J., GALINSKI M.R., BARNWELL J.W. AND FRASER-LIGGETT C.M. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, 2008, *455*, 757-763.

CARLTON J.M., ANGIUOLI S.V., SUH B.B., KOOIJ T.W., PERTEA M., SILVA J.C., ERMOLAEVA M.D., ALLEN J.E., SELENGUT J.D., KOO H.L., PETERSON J.D., POP M., KOSACK D.S., SHUMWAY M.F., BIDWELL S.L., SHALLOM S.J., VAN AKEN S.E., RIEDMULLER S.B., FELDBLYUM T.V., CHO J.K., QUACKENBUSH J., SEDEGAH M., SHOAIBI A., CUMMINGS L.M., FLORENS L., YATES J.R., RAINE J.D., SINDEN R.E., HARRIS M.A., CUNNINGHAM D.A., PREISER P.R., BERGMAN L.W., VAIDYA A.B., VAN LIN L.H., JANSE C.J., WATERS A.P., SMITH H.O., WHITE O.R., SALZBERG S.L., VENTER J.C., FRASER C.M., HOFFMAN S.L., GARDNER M.J. & CARUCCI D.J. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, 2002, *419*, 512-519.

CLARK K., DHOOGRA M., LOUW A.I. & BIRKHOLTZ L.M. Transcriptional responses of *Plasmodium falciparum* to alpha-difluoromethylornithine-induced polyamine depletion. *Biol. Chem.*, 2008, *389*, 111-125.

DAHL E.L., SHOCK J.L., SHENAI B.R., GUT J., DERISI J.L. & ROSENTHAL P.J. Tetracyclines specifically target the apicoplast of the malaria parasite *Plasmodium falciparum*. *Antimicrob. Agents Chemother.*, 2006, *50*, 3124-3131.

DATE S.V. & STOECKERT C.J., JR. Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. *Genome Res.*, 2006, *16*, 542-549.

DE BEER T.A., WELLS G.A., BURGER P.B., JOUBERT F., MARECHAL E., BIRKHOLTZ L. & LOUW A.I. Antimalarial drug discovery: in silico structural biology and rational drug design. *Infect. Disord. Drug Targets*, 2009, *9*, 304-318.

DE KONING-WARD T.F., GILSON P.R., BODDEY J.A., RUG M., SMITH B.J., PAPENFUSS A.T., SANDERS P.R., LUNDIE R.J., MAIER A.G., COWMAN A.F. & CRABB B.S. A newly discovered protein export machine in malaria parasites. *Nature*, 2009, *459*, 945-949.

DECHAMPS S., MAYNADIER M., WEIN S., GANNOUN-ZAKI L., MARECHAL E. & VIAL H.J. Rodent and nonrodent malaria

parasites differ in their phospholipid metabolic pathways. *J. Lipid Res.*, 2010, *51*, 81-96.

Doolittle R.F. Biodiversity: microbial genomes multiply. *Nature*, 2002, *416*, 697-700.

Durbin R., Eddy S., Krogh A. & Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. *Cambridge University Press*, 1998, *46*-132.

Eisen M.B., Spellman P.T., Brown P.O. & Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, 1998, *95*, 14863-14868.

Florens L., Liu X., Wang Y., Yang S., Schwartz O., Peglar M., Carucci D.J., Yates J.R., 3rd & Wub Y. Proteomics approach reveals novel proteins on the surface of malaria-infected erythrocytes. *Mol. Biochem. Parasitol.*, 2004, *135*, 1-11.

Florens L., Washburn M.P., Raine J.D., Anthony R.M., Grainger M., Haynes J.D., Moch J.K., Muster N., Sacci J.B., Tabb D.L., Witney A.A., Wolters D., Wu Y., Gardner M.J., Holder A.A., Sinden R.E., Yates J.R. & Carucci D.J. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, 2002, *419*, 520-526.

Florent I., Porcel B.M., Guillaume E., Da Silva C., Artiguenave F., Marechal E., Brehelin L., Gascuel O., Charneau S., Wincker P. & Grellier P. A *Plasmodium falciparum* FcB1-schizont-EST collection providing clues to schizont specific gene structure and polymorphism. *BMC Genomics*, 2009, *10*, 235.

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. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 2002, *419*, 498-511.

Ghouila A., Terrapon N., Gascuel O., Guerfali F.Z., Laouini D., Maréchal É & Bréhélin L. EuPathDomains : The divergent domain database for eukaryotic pathogenes. *Infection, Genetics & Evolution*, 2010 (in press).

Greenwood B. Progress in malaria control in endemic areas. *Travel Med. Infect. Dis.*, 2008, *6*, 173-176.

Greenwood B. Can malaria be eliminated? *Trans. R. Soc. Trop. Med. Hyg.*, 2009, *103* (Suppl. 1), S2-S5.

Hall N., Karras M., Raine J.D., Carlton J.M., Kooij T.W., Berriman M., Florens L., Janssen C.S., Pain A., Christophides G.K., James K., Rutherford K., Harris B., Harris D., Churcher C., Quail M.A., Ormond D., Doggett J., Trueman H.E., Mendoza J., Bidwell S.L., Rajandream M.A., Carucci D.J., Yates J.R., 3rd, Kafatos F.C., Janse C.J., Barrell B., Turner C.M., Waters A.P. & Sinden R.E. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, 2005, *307*, 82-86.

Hu G., Cabrera A., Kono M., Mok S., Chaal B.K., Haase S., Engelberg K., Cheemadan S., Spielmann T., Preiser P.R., Gilberger T.W. & Bozdech Z. Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*. *Nat. Biotechnol.*, 2010, *28*, 91-98.

Jeffares D.C., Pain A., Berry A., Cox A.V., Stalker J., Ingle C.E., Thomas A., Quail M.A., Siebenthall K., Uhlemann A.C., Kyes S., Krishna S., Newbold C., Dermitzakis E.T. & Berriman M. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat. Genet.*, 2007, *39*, 120-125.

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

Khan S.M., Franke-Fayard B., Mair G.R., Lasonder E., Janse C.J., Mann M. & Waters A.P. Proteome analysis of separated male and female gametocytes reveals novel sex-specific *Plasmodium* biology. *Cell*, 2005, *121*, 675-687.

Kooij T.W., Carlton J.M., Bidwell S.L., Hall N., Ramesar J., Janse C.J. & Waters A.P. A *Plasmodium* Whole-Genome Synteny Map: Indels and Synteny Breakpoints as Foci for Species-Specific Genes. *PLoS Pathog*, 2005, *1*, e44.

LaCount D.J., Vignali M., Chettier R., Phansalkar A., Bell R., Hesselberth J.R., Schoenfeld L.W., Ota I., Sahasrabudhe S., Kurschner C., Fields S. & Hughes R.E. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, 2005, *438*, 103-107.

Lal K., Prieto J.H., Bromley E., Sanderson S.J., Yates J.R., 3rd, Wastling J.M., Tomley F.M. & Sinden R.E. Characterisation of *Plasmodium* invasive organelles; an ookinete microneme proteome. *Proteomics*, 2009, *9*, 1142-1151.

Lamarque M., Tastet C., Poncet J., Demettre E., Jouin P., Vial H.J. & Dubremetz J.F. Food vacuole proteome of the malarial parasite *Plasmodium falciparum*. *Proteomics*, 2008, *2*, 1361-1374.

Lasonder E., Ishihama Y., Andersen J.S., Vermunt A.M., Pain A., Sauerwein R.W., Eling W.M., Hall N., Waters A.P., Stunnenberg H.G. & Mann M. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature*, 2002, *419*, 537-542.

Lasonder E., Janse C.J., van Gemert G.J., Mair G.R., Vermunt A.M., Douradinha B.G., van Noort V., Huynen M.A., Luty A.J., Kroeze H., Khan S.M., Sauerwein R.W., Waters A.P., Mann M. & Stunnenberg H.G. Proteomic profiling of *Plasmodium* sporozoite maturation identifies new proteins essential for parasite development and infectivity. *PLoS Pathog*, 2008, *4*, e1000195.

Le Roch K.G., Johnson J.R., Ahiboh H., Chung D.W., Prudhomme J., Plouffe D., Henson K., Zhou Y., Witola W., Yates J.R., Mamoun C.B., Winzeler E.A. & Vial H. A systematic approach to understand the mechanism of action of the bisthiazolium compound T4 on the human malaria parasite, *Plasmodium falciparum*. *BMC Genomics*, 2008, *9*, 513.

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. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 2003, *301*, 1503-1508.

Liew K.J., Hu G., Bozdech Z. & Peter P.R. Defining species specific genome differences in malaria parasites. *BMC Genomics*, 2010, *11*, 128.

LIPMAN D.J. & PEARSON W.R. Rapid and sensitive protein similarity searches. *Science*, 1985, *227*, 1435-1441.

LLINAS M., BOZDECH Z., WONG E.D., ADAI A.T. & DERISI J.L. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res.*, 2006, *34*, 1166-1173.

LLINAS M. & DEL PORTILLO H.A. Mining the malaria transcriptome. *Trends Parasitol.*, 2005, *21*, 350-352.

MAIR G.R., BRAKS J.A., GARVER L.S., WIEGANT J.C., HALL N., DIRKS R.W., KHAN S.M., DIMOPOULOS G., JANSE C.J. & WATERS A.P. Regulation of sexual development of *Plasmodium* by translational repression. *Science*, 2006, *313*, 667-669.

NATALANG O., BISCHOFF E., DEPLAINE G., PROUX C., DILLIES M.A., SISMEIRO O., GUIGON G., BONNEFOY S., PATARAPOTIKUL J., MERCEREAU-PUIJALON O., COPPEE J.Y. & DAVID P.H. Dynamic RNA profiling in *Plasmodium falciparum* synchronized blood stages exposed to lethal doses of artesunate. *BMC Genomics*, 2008, *9*, 388.

NYALWIDHE J. & LINGELBACH K. Proteases and chaperones are the most abundant proteins in the parasitophorous vacuole of *Plasmodium falciparum*-infected erythrocytes. *Proteomics*, 2006, *6*, 1563-1573.

PAIN A., BOHME U., BERRY A.E., MUNGALL K., FINN R.D., JACKSON A.P., MOURIER T., MISTRY J., PASINI E.M., ASLETT M.A., BALASUBRAMMANIAM S., BORGWARDT K., BROOKS K., CARRET C., CARVER T.J., CHEREVACH I., CHILLINGWORTH T., CLARK T.G., GALINSKI M.R., HALL N., HARPER D., HARRIS D., HAUSER H., IVENS A., JANSSEN C.S., KEANE T., LARKE N., LAPP S., MARTI M., MOULE S., MEYER I.M., ORMOND D., PETERS N., SANDERS M., SANDERS S., SARGEANT T.J., SIMMONDS M., SMITH F., SQUARES R., THURSTON S., TIVEY A.R., WALKER D., WHITE B., ZUIDERWIJK E., CHURCHER C., QUAIL M.A., COWMAN A.F., TURNER C.M., RAJANDREAM M.A., KOCKEN C.H., THOMAS A.W., NEWBOLD C.I., BARRELL B.G. & BERRIMAN M. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, 2008, *455*, 799-803.

RADFAR A., DIEZ A. & BAUTISTA J.M. Chloroquine mediates specific proteome oxidative damage across the erythrocytic cycle of resistant *Plasmodium falciparum*. *Free Radic. Biol. Med.*, 2008, *44*, 2034-2042.

SAIDANI N., GRANDO D., VALADIE H., BASTIEN O. & MARECHAL E. Potential and limits of in silico target discovery - Case study of the search for new antimalarial chemotherapeutic targets. *Infect. Genet. Evol.*, 2009, *9*, 359-367.

SAM-YELLOWE T.Y., BANKS T.L., FUJIOKA H., DRAZBA J.A. & YADAV S.P. *Plasmodium yoelii*: novel rhoptry proteins identified within the body of merozoite rhoptries in rodent *Plasmodium* malaria. *Exp. Parasitol.*, 2008, *120*, 113-117.

SAM-YELLOWE T.Y., FLORENS L., WANG T., RAINE J.D., CARUCCI D.J., SINDEN R. & YATES J.R., 3RD Proteome analysis of rhoptry-enriched fractions isolated from *Plasmodium* merozoites. *J. Proteome Res.*, 2004, *3*, 995-1001.

SANDERS P.R., CANTIN G.T., GREENBAUM D.C., GILSON P.R., NEBL T., MORITZ R.L., YATES J.R., 3RD, HODDER A.N. & CRABB B.S. Identification of protein complexes in detergent-resistant membranes of *Plasmodium falciparum* schizonts. *Mol. Biochem. Parasitol.*, 2007, *154*, 148-157.

SHOCK J.L., FISCHER K.F. & DERISI J.L. Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.*, 2007, *8*, R134.

SPELLMAN P.T., SHERLOCK G., ZHANG M.Q., IYER V.R., ANDERS K., EISEN M.B., BROWN P.O., BOTSTEIN D. & FUTCHER B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 1998, *9*, 3273-3297.

STOECKERT C.J., JR., FISCHER S., KISSINGER J.C., HEIGES M., AURRECOECHEA C., GAJRIA B. & ROOS D.S. PlasmoDB v5: new looks, new genomes. *Trends Parasitol.*, 2006, *22*, 543-546.

TARUN A.S., PENG X., DUMPIT R.F., OGATA Y., SILVA-RIVERA H., CAMARGO N., DALY T.M., BERGMAN L.W. & KAPPE S.H. A combined transcriptome and proteome survey of malaria parasite liver stages. *Proc. Natl Acad. Sci. USA*, 2008, *105*, 305-310.

TERRAPON N., GASCUEL O., MARECHAL E. & BREHELIN L. Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*. *Bioinformatics*, 2009, *25*, 3077-3083.

TUTEJA R. Malaria - an overview. *FEBS J.*, 2007, *274*, 4670-4679.

VOLKMAN S.K., SABETI P.C., DECAPRIO D., NEAFSEY D.E., SCHAFFNER S.F., MILNER D.A., JR., DAILY J.P., SARR O., NDIAYE D., NDIR O., MBOUP S., DURAISINGH M.T., LUKENS A., DERR A., STANGE-THOMANN N., WAGGONER S., ONOFRIO R., ZIAUGRA L., MAUCELI E., GNERRE S., JAFFE D.B., ZAINOUN J., WIEGAND R.C., BIRREN B.W., HARTL D.L., GALAGAN J.E., LANDER E.S. & WIRTH D.F. A genome-wide map of diversity in *Plasmodium falciparum*. *Nat. Genet.*, 2007, *39*, 113-119.

WINZELER E.A. Malaria research in the post-genomic era. *Nature*, 2008, *455*, 751-756.

YOUNG J.A., FIVELMAN Q.L., BLAIR P.L., DE LA VEGA P., LE ROCH K.G., ZHOU Y., CARUCCI D.J., BAKER D.A. & WINZELER E.A. The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.*, 2005, *143*, 67-79.

ZHOU Y., RAMACHANDRAN V., KUMAR K.A., WESTENBERGER S., REFOUR P., ZHOU B., LI F., YOUNG J.A., CHEN K., PLOUFFE D., HENSON K., NUSSENZWEIG V., CARLTON J., VINETZ J.M., DURAISINGH M.T. & WINZELER E.A. Evidence-based annotation of the malaria parasite's genome using comparative expression profiling. *PLoS One*, 2008, *3*, e1570.

Reçu le 11 juin 2010
Accepté le 28 juillet 2010