PlasmoDraft: a database of Plasmodium falciparum gene function predictions based on postgenomic data

Laurent Brehelin, Jean-François Dufayard, Olivier Gascuel

BMC Bioinformatics

Research article

PlasmoDraft: a database of *Plasmodium falciparum* gene function predictions based on postgenomic data

Laurent Bréhélin*, Jean-François Dufayard and Olivier Gascuel

Address: Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada, 34392 MONTPELLIER, France

Email: Laurent Bréhélin* - brehelin@lirmm.fr; Jean-François Dufayard - Jean-francois.Dufayard@lirmm.fr; Olivier Gascuel - gascuel@lirmm.fr

* Corresponding author

Published: 16 October 2008

BMC Bioinformatics 2008, **9**:440 doi:10.1186/1471-2105-9-440

Received: 13 June 2008
Accepted: 16 October 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/440

Abstract

Background: Of the 5 484 predicted proteins of *Plasmodium falciparum*, the main causative agent of malaria, about 60% do not have sufficient sequence similarity with proteins in other organisms to warrant provision of functional assignments. Non-homology methods are thus needed to obtain functional clues for these uncharacterized genes.

Results: We present PlasmoDraft http://atgc.lirmm.fr/PlasmoDraft/, a database of Gene Ontology (GO) annotation predictions for *P. falciparum* genes based on postgenomic data. Predictions of PlasmoDraft are achieved with a *Guilt By Association* method named Gonna. This involves (1) a predictor that proposes GO annotations for a gene based on the similarity of its profile (measured with transcriptome, proteome or interactome data) with genes already annotated by GeneDB; (2) a procedure that estimates the confidence of the predictions achieved with each data source; (3) a procedure that combines all data sources to provide a global summary and confidence estimate of the predictions. Gonna has been applied to all *P. falciparum* genes using most publicly available transcriptome, proteome and interactome data sources. Gonna provides predictions for numerous genes without any annotations. For example, 2 434 genes without any annotations in the Biological Process ontology are associated 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 are associated with specific GO terms, respectively (740 and 329 with confidence value above 50%).

Conclusion: All predictions along with their confidence values have been compiled in PlasmoDraft, which thus provides an extensive database of GO annotation predictions that can be achieved with these data sources. 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. 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.

Background

Malaria is one of the most prevalent disease in the world, infecting 400 million people every year, and causing 2.7 million deaths, mainly children under 5 years [1]. *Plasmodium falciparum*, the main causative agent of this parasitic disease, develops drug resistance and no effective vaccine

is available. Of the 5 484 coding genes of *P. falciparum* (http://plasmodb.org version 5.4), about 60% do not have sufficient similarity to proteins in other organisms to warrant provision of functional assignments. Thus, almost two-thirds of the proteins appear to be specific to *P. falciparum*, a much higher proportion than observed in other enkaryotes [2]. However, this is likely exacerbated by the high evolutionary distance between *P. falciparum* and other sequenced eukaryotes, so homology detection is a hard task. Because of the extreme AT bias (80%), the high amino acid bias (six amino acids account for more than 50% of the protein composition) and the presence of a large number of low complexity repeat regions that are believed to form non-globular segments [3], standard sequence comparison methods based on BLAST [4] or HMMER [5] could be ineffective [6]. Non-homology methods are thus needed to obtain functional clues for these uncharacterized genes [7].

Methods based on post-genomic data (mainly gene expression and protein interaction) have been proposed. 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 [8,9]. In the same way, protein interaction data are also used since proteins that share common interactors likely share common functions [10-12]. These methods provide functional predictions for the uncharacterized genes, and new clues to be compared with the predictions achieved by homology.

Part of these new post-genomic methods work in a non-supervised way: first a gene clustering algorithm is run on the post-genomic data to cluster the genes into several groups. Then, in each cluster and for each potential function, a statistical test is applied to compare the proportion of genes annotated with this function in the cluster with that in the complete set of genes. Functions that appear to be over-represented in one cluster are used to annotate the uncharacterized genes that belong to this cluster. Several genome-scale studies have been conducted using this principle, *e.g.* [8,13,14].

Some other GBA methods work in a supervised way: first, based on the post-genomic data of already characterized genes, a supervised learning algorithm is run to learn a predictor, *i.e.* a function that takes post-genomic measurements of a given gene as input, and outputs one or several functional predictions for that gene. This predictor is then used to annotate the uncharacterized genes. Typical examples of this approach are, *e.g.* [11,15,16]. Zhou et al. [17] presented OPI, a supervised method that predicts Gene Ontology annotations using gene expression profiles and was applied on *P. falciparum*. Alternative methods work in a semi-supervised way [18]; these use gene clustering as in the non-supervised approach, but clustering is not fully unsupervised as the function of the already characterized genes is used to define the clusters.

In this paper, we present PlasmoDraft http://atgc.lirmm.fr/PlasmoDraft/, a database of Gene Ontology (GO) annotation predictions for *P. falciparum* achieved by applying a GBA predictor named Gonna (for *Gene Ontology Nearest Neighbor Approach*) on several transcriptome (microarray), proteome (mass-spectrometry) and protein-protein interaction data. The Gonna system involves: (1) a supervised k-nearest-neighbor predictor that proposes predictions on the basis of each data source; (2) a cross-validation procedure that estimates the confidence of the predictions achieved with each data source; (3) a procedure that combines the results achieved with the different data sources to estimate a global confidence value of each prediction for each gene. The PlasmoDraft database provides all of these predictions along with their confidence values in a friendly interface that allows easy browsing and querying.

Methods

Gonna proposes annotation predictions in the GO framework. The GO Consortium http://www.geneontology.org has developed a systematic and standardized nomenclature to annotate genes in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF), in a species-independent manner. Each ontology describes generalization relationships between hundreds of terms. The most general term is at the top of the ontology, while the bottom terms are the most specific ones. A gene may be annotated with several GO terms of the same ontology. Moreover, due to the generalization relationship, when a gene is annotated with a term *t*, then it is also annotated with all upper terms that generalize *t* (a principle known as the "*true path rule*" in GO context). In PlasmoDraft, the specificity of a term is evaluated by its prior probability, *i.e.* the proportion of already characterized genes of *P. falciparum* that belong to this term. In this way, the leaves of the ontologies are the most specific terms with low prior probabilities, while the root of the ontology is the most common term with a prior probability of 1. Gonna uses the GO annotations available on PlasmoDB and provided by GeneDB as prior knowledge database to propose new annotations. The GO consortium distinguishes between curator-assigned annotations and automatically-assigned annotations. Curator-

assigned annotations involve annotations that come from experimental data (GO evidence codes IDA, IPI, etc.), or that have been inferred by sequence similarity and curated by an expert (GO evidence code ISS). Automatically-assigned annotations involve all electronically inferred annotations (usually by sequence similarity) that have not been reviewed by an expert (GO evidence code IEA). Here, due to the scarcity of the curator-assigned annotations for *P. falciparum* (~60% annotations possess IEA evidence code only), all available GO annotations are considered, without regard to their evidence code (this choice is further-discussed below). Every gene with an annotation in the considered ontology (whatever its evidence code) is then referred as "characterized".

The predictor

Gonna uses a *k*-nearest neighbor approach [19]. It takes as input two positive integers K and $K' \leq K$ (*e.g.* $K = 6$ and $K' = 4$), one ontology (MF, BP, or CC), and one postgenomic data source D (*e.g.* the microarray data of [14]). With this data source, Gonna computes a function S_D that measures the similarity $S_D(g, h)$ of every gene pair (g, h). For example, if D is a transcriptomic data set then S_D measures the similarity of profiles using the Pearson correlation coefficient. When asked for the GO categories of a gene g, Gonna uses the S_D function to search for the K genes already characterized in the selected ontology by GeneDB, which have the highest level of similarity with g. Then, for each GO term t of the ontology, Gonna looks at these K genes, and if at least K' are associated with t, then g is predicted to be also associated with t; otherwise g is not considered to be in t. Note that when looking at the terms associated with the neighbor genes, Gonna considers all the upper terms generalizing the direct annotations (*i.e.* all terms in the true path rule).

Some choices are critical to insure that Gonna provides relevant and accurate predictions. The first critical choice is related to the similarity measure, which has to capture the "signature" of the gene functions in the data set at hand. When two genes appear to be similar, this should imply that they share common functions. For transcriptomic (microarray) and proteomic (mass-spectrometry) data, we use the Pearson correlation coefficient that gives high similarity to genes with correlated transcriptomic/proteomic profiles. Other similarity measures, as the classical Euclidean metric, could be possible, but the Pearson correlation measure has been shown to perform well to detect functional links in several analyses [20]. For the protein-protein interaction data, we use the Czekanovski-Dice metric [21], which gives high similarity to pairs of genes that share many interactors, and has been shown to perform well to predict biological functions [10].

Another critical choice is related to the K and K' values. K should be neither too large (else some neighbors will not be similar to the studied gene) nor too low (to avoid reduced, non-representative gene samples). With K' the problem is different. If K' is high (close to K), then the proportion of good predictions is likely to be high, but only a few predictions could be achieved on the most specific terms of the ontology, and most of the predictions would involve the most general (and hence less interesting) terms. Conversely, if K' is low, then the proportion of good predictions declines, but more predictions are made on the most specific terms. In PlasmoDraft, we use two pairs of parameters (K, K') for each postgenomic data source: one stringent pair ($K = 6$, $K' = 4$) is used to achieve, for each GO term, a first set of predictions that usually has a high proportion of good predictions (see next section for an estimate of this proportion). Next, a second, non-stringent pair ($K = 6$, $K' = 2$) is used to come up with, for each GO term, another set of predictions that cannot be achieved with the stringent setting, but which usually contains a lower proportion of good predictions.

This *k*-nearest neighbor predictor has several appealing features. It is a direct and simple implementation of the GBA principle, which allows the predictions to be explained by exhibiting the K' genes annotated by GeneDB that support each prediction (see Figure 1). In fact, Gonna uses a basic principle similar to gene expression mining tools as g:profiler [22], which help users to make their own predictions. These tools search for genes with expression profile correlated with that of the studied gene, look for GO terms enriched in the neighboring gene list, and then predict the selected GO terms for the studied gene. Gonna can thus be viewed as a systematic and automatic implementation of this natural principle, combined with confidence estimation and data source aggregation (see below). Moreover, Gonna can be used with any present and future postgenomic data source, as long as there is a relevant similarity measure. Next. Gonna is consistent with the structure of the ontology. This important property means that if any gene is predicted in a GO term t, then it must be predicted in all terms that generalize t. Finally, Gonna has low computing time, which enables intensive use of the cross-validation procedure to assess the confidence of the predictions.

Assessing the predictions

Cross-validation (CV) is a well known procedure to estimate the error rate of supervised classification methods [19]. The leave-one-out version of CV, which we use here, involves: (1) running Gonna on each gene already characterized in GeneDB as if it were an uncharacterized gene, and (2) comparing the predictions to the true annotations. Since no functional information on this gene is supplied to Gonna for the predictions, this procedure provides an unbiased estimate of the method performance [19]. For a given GO term t, the correct predictions in CV involve the genes predicted in t, which are already

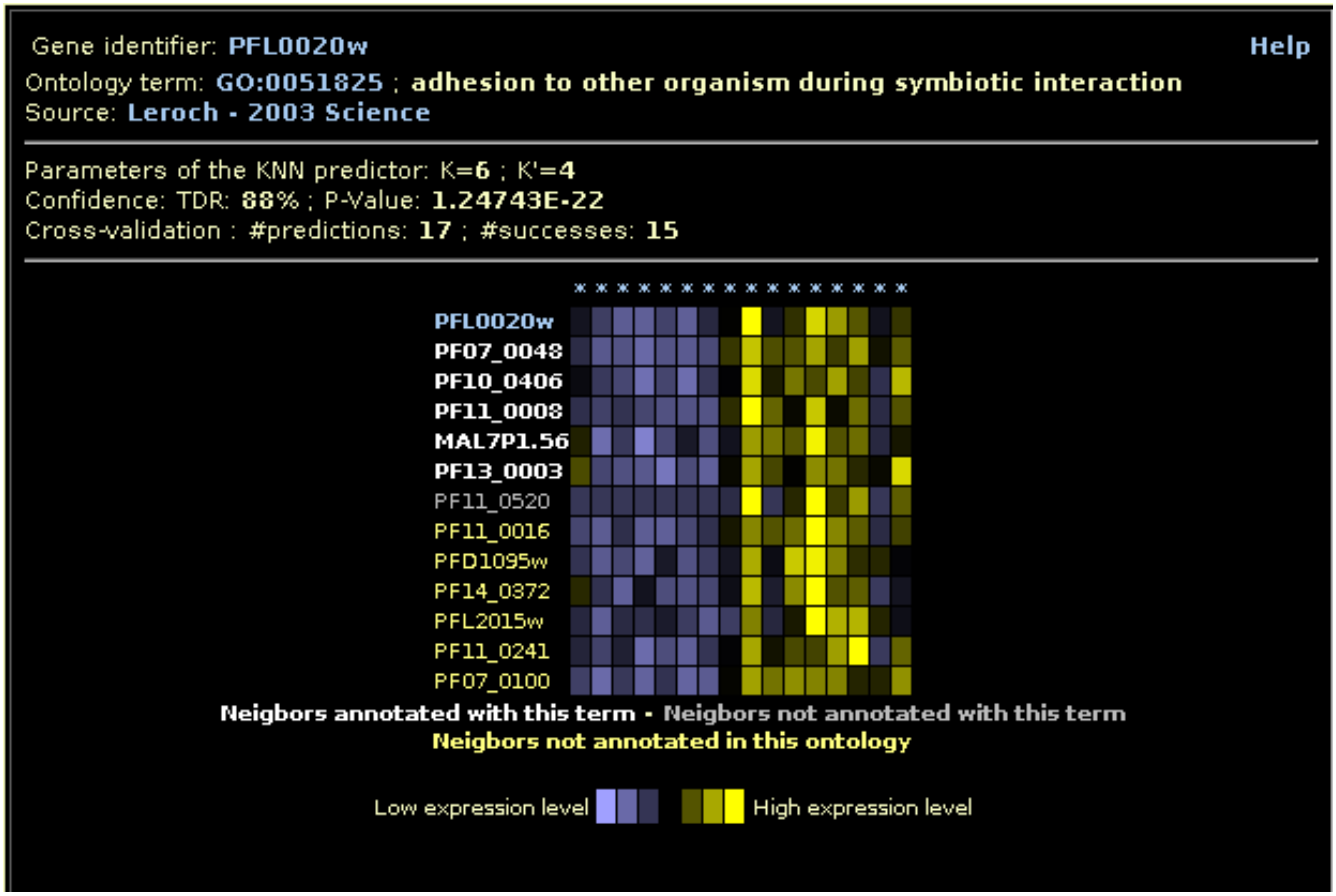


Figure 1

The neighbors view. Profiles of the *K* nearest characterized neighbors that support (white), or does not support (gray), the prediction of gene PFL0020w in term `Adhesion to other organism during symbiotic interaction` (GO:0051825) for the Leroch et al. (2003) data source [14]. For comparison purpose, profiles of the *K* nearest uncharacterized neighbors (yellow) are also reported.

annotated by this term in GeneDB (or by a specialization of *t*, cf. the true path rule); the wrong predictions involve the genes predicted in *t* which are already annotated in the ontology under consideration (MF, BP, or CC) but not with *t* (or one of its specializations). Genes without any annotation in the selected ontology are not taken into account. It is convenient to present all of these quantities in tabular form (see Table 1). For example, p_a denotes the number of genes predicted in the GO term *t* which are annotated with *t* in GeneDB, while n_n denotes the number of genes not annotated with *t* which have not been predicted in *t* in CV. Then, the True Discovery Rate (*TDR*) associated with this GO term (and for a given data source) is estimated by

$$TDR = p_a/(p_a + p_n). \quad (1)$$

For example, a GO term with a *TDR* of 80% means that when Gonna predicts that a gene belongs to this term, this prediction has 80% chance of being correct. Note that due to the incompleteness of the annotations, the above formula may be a conservative estimate of the *TDR*, because some predictions considered as wrong may actually be correct. Moreover, when the sample size ($p_a + p_n$) is low. this *TDR* estimate may not be fully accurate. So, we also compute the p-value of achieving by chance p_a or more correct predictions (among $p_a + p_n$) if the true *TDR* were equal to the prior probability of the term. If this p-value is higher than 5%, then the *TDR* is not considered to be significantly higher than the prior probability. PlasmoDraft reports the *TDR*s of the predictions with a color code that ranges from red (0%) to light green (100%) via yellow (50%), while non-significant *TDR*s appear in gray (cf Figure 2). As described above, two sets of predictions are achieved for each GO term and data source using two parameters *K*'. Therefore, one *TDR* is estimated for each of these sets: the first *TDR* reports the accuracy of the predictions achieved with the stringent predictor, while the sec-

ond *TDR* reports the accuracy of the predictions achieved with the non-stringent predictor but which are not supported by the stringent one. As expected, the first *TDR* is usually higher than the second one. When neither the stringent predictor nor the non-stringent one apply, the gene is said to be "non predicted in *t*".

The advantage of estimating the *TDR* of each GO term rather than estimating a global performance on the whole ontology is that it allows to differentiate GO terms that appear well suited for applying a GBA approach with the considered data source. Indeed, all GO terms cannot be predicted with the same accuracy. First because some terms are more general than others (and thus are *a priori* more likely). But also because some functions (GO terms) have a more apparent signature than others in the considered data source. For example, while they have a similar prior probability (~10%), GO terms `antigenic variation (GO:0020033)` and `post-translational protein modification (GO:0043687)` get 90% and 15% *TDR*s with the microarray data of [14], respectively.

Combining the data sources

When each data source has been used to produce predictions, and *TDR*s have been estimated for each GO term and each source, Gonna combines all of these results to propose a Global Degree of Belief (*GDB*) for each prediction. If gene *g* has been predicted to be associated with GO term *t* by one or several sources, Gonna computes the *GDB* of this prediction in the following way. Let 1,..., *n* and *n* + 1,..., *m* denote the data sources that support, and do no support, the prediction of *g* in *t*, respectively. We use the notation d_i and $\neg d_j$ to indicate that data sources *i* and *j* support and do not support the prediction of *g* in *t*, respectively. We first compute a global confidence score that is a rough estimate of the probability that the prediction is correct, given that it is supported by data sources 1,..., *n* but not by data sources *n* + 1,..., *m*. Using Bayes theorem, this probability can be written as

$$P(t \mid d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m) = \frac{P(d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m \mid t)\times P(t)}{P(d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m)}.$$

$P(t)$ is the prior probability of term *t* (estimated by the proportion of already characterized genes of *P. falciparum* that belong to *t*). $P(d_1,\ldots, d_n, \neg d_{n+1},\ldots, \neg d_m \mid t)$ is the probability that data sources 1,..., *n* and data sources *n* + 1,..., *m* support and do not support the prediction of *g* in *t* when *g* belongs to *t*, respectively. We use the conditional independence assumption [19] to estimate this latter term and the probability $P(d_1,\ldots, d_n, \neg d_{n+1},\ldots, \neg d_m)$:

$$P(d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m \mid t) \simeq P(d_1 \mid t)\times\cdots\times P(d_n \mid t)\times P(\neg d_{n+1} \mid t)\times\cdots\times P(\neg d_m \mid t),$$

$$P(d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m \mid \neg t) \simeq P(d_1 \mid \neg t)\times\cdots\times P(d_n \mid \neg t)\times P(\neg d_{n+1} \mid \neg t)\times\cdots\times P(\neg d_m \mid \neg t),$$

and

$$P(d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m) = P(d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m \mid t)P(t) + P(d_1,\ldots,d_n,\neg d_{n+1},\ldots,\neg d_m \mid \neg t)P(\neg t).$$

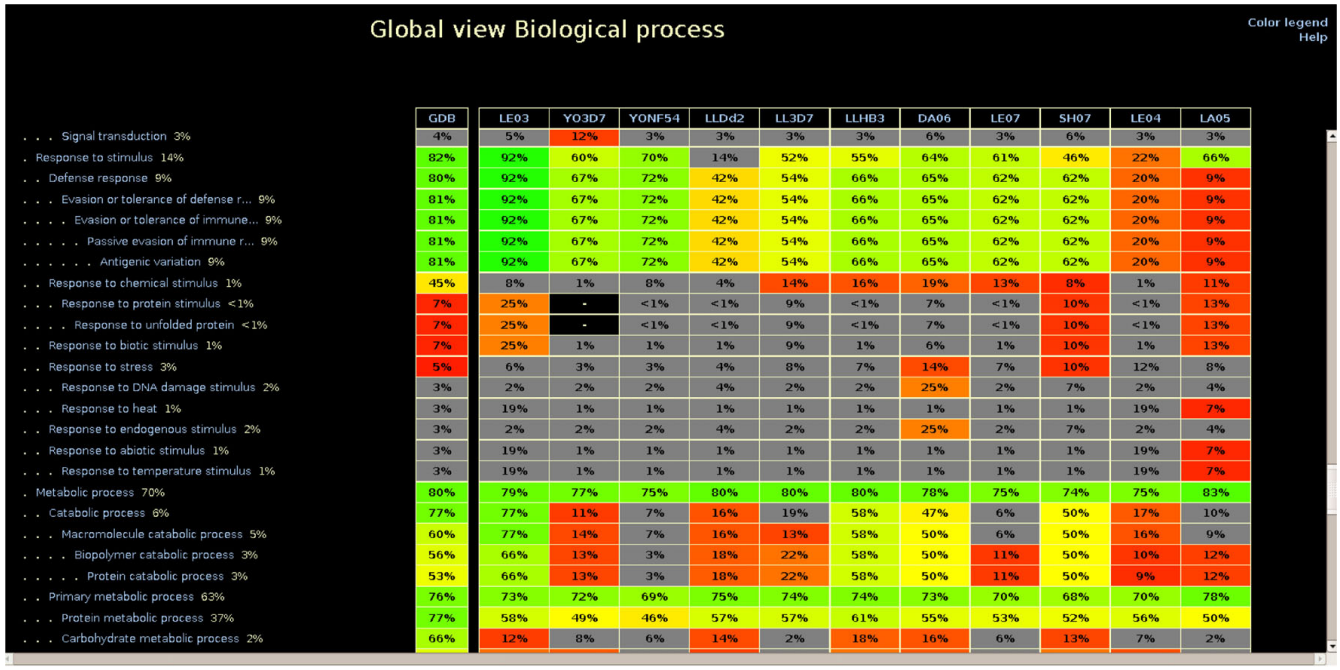
Terms $P(d_i|t)$, $P(\neg d_i|t)$, $P(d_i|\neg t)$, and $P(\neg d_i|\neg t)$ are estimated with the quantities computed in the CV and displayed in Table 1. For example, $P(d_i|t)$ is the probability that data source *i* supports *t* when the gene belongs to *t*; it is estimated by the ratio $p_a/(p_a + n_a)$. $P(\neg d_i|\neg t)$ is the probability that data source *i* does not support *t* when the gene does not belongs to *t*; it is estimated with $n_n/(p_n + n_n)$.

Thus, from the three above equations, the conditional probability of *t* can be roughly estimated and it constitutes our global confidence score. This score reflects the likelihood of the predictions: genes with high (near 1) confidence scores are more likely to be associated with *t* than genes with low (near 0) confidence scores. However, due to the independence assumption, this score cannot be interpreted as the probability of *t*. Hence, it is discretized in 4 score categories (very low [0.0, 0.25], low]0.25, 0.5], high]0.5, 0.75], and very high]0.75,1.0]). The true discovery rate associated with each category is estimated by way of a last cross-validation procedure: this is done by computing the proportion of successes among already characterized genes that have been predicted in the considered GO term with a confidence score in this category. These cross-validated true discovery rates then represent our *GDB*. For example, a prediction associated with a *GDB* of 80% means that 80% of the predictions belonging to the same score category in this GO term are correct in the CV procedure. As for the *TDR*s, we also compute the p-

Table 1: Correct and wrong predictions associated with a given GO term *t* in the CV procedure

	predicted in	not predicted in
annotated with	p_a	n_a
not annotated with	p_n	n_n

p_a and p_n denote the number of genes predicted in the GO term *t* which are, and are not, annotated with *t* in GeneDB, respectively. n_a and n_n denote the number of genes not predicted in *t* which are. and are not, annotated with *t*, respectively.



	GDB	LE03	YO3D7	YONF54	LLDd2	LL3D7	LLHB3	DA06	LE07	SH07	LE04	LA05
. . . Signal transduction 3%	4%	5%	12%	3%	3%	3%	3%	6%	3%	6%	3%	3%
. Response to stimulus 14%	82%	92%	60%	70%	14%	52%	55%	64%	61%	46%	22%	66%
. . Defense response 9%	80%	92%	67%	72%	42%	54%	66%	65%	62%	62%	20%	9%
. . . Evasion or tolerance of defense r... 9%	81%	92%	67%	72%	42%	54%	66%	65%	62%	62%	20%	9%
. . . . Evasion or tolerance of immune... 9%	81%	92%	67%	72%	42%	54%	66%	65%	62%	62%	20%	9%
. . . . . Passive evasion of immune r... 9%	81%	92%	67%	72%	42%	54%	66%	65%	62%	62%	20%	9%
. . . . . Antigenic variation 9%	81%	92%	67%	72%	42%	54%	66%	65%	62%	62%	20%	9%
. . Response to chemical stimulus 1%	45%	8%	1%	8%	4%	14%	16%	19%	13%	8%	1%	11%
. . . Response to protein stimulus <1%	7%	25%	-	<1%	<1%	9%	<1%	7%	<1%	10%	<1%	13%
. . . . Response to unfolded protein <1%	7%	25%	-	<1%	<1%	9%	<1%	7%	<1%	10%	<1%	13%
. . Response to biotic stimulus 1%	7%	25%	1%	1%	1%	9%	1%	6%	1%	10%	1%	13%
. . Response to stress 3%	5%	6%	3%	3%	4%	8%	7%	14%	7%	10%	12%	8%
. . . Response to DNA damage stimulus 2%	3%	2%	2%	2%	4%	2%	2%	25%	2%	7%	2%	4%
. . . Response to heat 1%	3%	19%	1%	1%	1%	1%	1%	1%	1%	1%	19%	7%
. . Response to endogenous stimulus 2%	3%	2%	2%	2%	4%	2%	2%	25%	2%	7%	2%	4%
. . Response to abiotic stimulus 1%	3%	19%	1%	1%	1%	1%	1%	1%	1%	1%	19%	7%
. . . Response to temperature stimulus 1%	3%	19%	1%	1%	1%	1%	1%	1%	1%	1%	19%	7%
. Metabolic process 70%	80%	79%	77%	75%	80%	80%	80%	78%	75%	74%	75%	83%
. . Catabolic process 6%	77%	77%	11%	7%	16%	19%	58%	47%	6%	50%	17%	10%
. . . Macromolecule catabolic process 5%	60%	77%	14%	7%	16%	13%	58%	50%	6%	50%	16%	9%
. . . . Biopolymer catabolic process 3%	56%	66%	13%	3%	18%	22%	58%	50%	11%	50%	10%	12%
. . . . . Protein catabolic process 3%	53%	66%	13%	3%	18%	22%	58%	50%	11%	50%	9%	12%
. . Primary metabolic process 63%	76%	73%	72%	69%	75%	74%	74%	73%	70%	68%	70%	78%
. . . Protein metabolic process 37%	77%	58%	49%	46%	57%	57%	61%	55%	53%	52%	56%	50%
. . . Carbohydrate metabolic process 2%	66%	12%	8%	6%	14%	2%	18%	16%	6%	13%	7%	2%

Figure 2
An extract of the Biological Process global view. This view presents a summary of all of the best *GDB*s and *TDR*s that are associated with each GO term and data source. Clicking on any term opens the corresponding GO term view.

value of the *GDB*s. If this p-value is higher than 5%, then the *GDB* is not considered to be significantly higher than the prior probability of the term, and it appears in gray in PlasmoDraft.

The discretization procedure we use, sometimes known as the *equal interval width* method, could be replaced by other methods, such as the *equal frequency interval* method or more sophisticated methods based on entropy minimization [23]. However, it is a classical and simple method that has shown to give good performance on numerous data sets [24].

The independence assumption is often used in statistical machine learning, and forms the basis of the "naive Bayes" predictor, which was shown to be fairly accurate in a number of applications [19]. One interesting feature of this predictor (and hence of the *GDB*) is that it is not much affected by irrelevant or poor quality data sources [25]. Indeed, when a source i is not relevant for a specific GO term t, either because it has not been designed for screening this type of information or because of the poor quality of the data, terms $P(d_i|t)$ and $P(d_i|\neg t)$ tend to be equal. Therefore, the numerical quantities related to this data source tend to cancel in the numerator and denominator pairs of the confidence score. This prevents the *GDB* from pollution by irrelevant or too noisy data sources.

Results

Data

To produce the PlasmoDraft database, Gonna has been applied to most publicly available postgenomic data sources we were aware. 9 transcriptomic (microarray), 1 proteomic (mass-spectrometry), and 1 protein-protein interaction data sets were used. Below is a short description of each data set, indexed by the name used in PlasmoDraft.

• LE03: Le Roch et al. (2003) data set [14]. A transcriptomic data set that covers 9 stages of the entire cycle of strain 3D7: 6 asexual intraerythrocytic stages, plus the merozoite, gametocyte, and salivary gland sporozoite stages. Measurements for ~5 100 genes.

• YO3D7: Young et al. (2005) data set [26]. A transcriptomic data set that covers the sexual developmental cycle (gametocytes) of strain 3D7. Measurements for ~5 100 genes.

• YONF54: Young et al. (2005) data set [26]. Same data set as Y03D7, for strain NF54.

• LLHB3: Llinas et al. (2006) data set [27,28]. A transcriptomic data set that covers 48 h of the intraerythrocytic developmental cycle of strain HB3. Measurements for ~4 200 genes.

• LLDd2: Llinas et al. (2006) data set [28]. Same data set as LLHB3, for strain Dd2.

• LL3D7: Llinas et al. (2006) data set [28]. Same data set as LLHB3, for strain 3D7.

• DA06: Dahl et al. (2006) data set [29]. A transcriptomic data set that covers two 48 h life cycles of doxycyclin treated parasites. Measurements for ~5 300 genes.

• SH07: Shock et al. (2007) data set [30]. A transcriptomic data set analysing mRNA decay during the intraerythrocytic developmental cycle. Measurements for ~5 300 genes.

• LE07: A transcriptomic data set analysing the parasite response to choline analog T4 during the intraerythrocytic life cycle. See series GSE4582 in the NCBI Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo/. Measurements for ~5 100 genes.

• LE04: Le Roch et al. (2004) data set [31,32]. A proteomic data set that covers 7 stages of the entire cycle of strain 3D7: the ring, trophozoite, schizont, merozoite, gametocyte, gamete, and salivary gland sporozoite stages. Measurements for ~2 900 genes.

• LA05: LaCount et al. (2005) data set [33]. A protein-protein interaction data set. Measurements for ~1 300 genes.

The Gene Ontology file (revision 5.754) and the gene annotations file (revision 1.54) were downloaded from the GO website.

Accessing the database

Users can access the predictions by browsing the database or querying for a specific gene, GO term, or keyword. Results are displayed using three types of views: a global view, a gene view, and a GO term view. In each view, *TDR*s and *GDB*s are represented with a color code that ranges from red (0%) to light green (100%) via yellow (50%); non-significant *TDR*s or *GDB*s (see Method) are in gray.

The global views

There is one global view for each gene ontology (Molecular Function, Biological Process, and Cellular Component). A global view (see Figure 2) shows all GO terms of the selected ontology where predictions are made. These are represented in a hierarchical way which respects the ontology structure. Each term is followed by its prior probability, the best *GDB* found for a gene predicted in this term, and the best *TDR* associated with each data source for this term.

The GO term view

The GO term views show all genes that are predicted in any given term by Gonna (see Figure 3). Two views are available for each GO term: one for uncharacterized genes that have no annotation in GeneDB for the ontology at hand, and the other one for genes that are already annotated in this ontology in GeneDB (but not obligatory with this term). For the latter, a '+' symbol after the gene name indicates that the gene is already annotated by the term. Additional information about predictions is provided by clicking on a specific *TDR*. This opens a new window presenting the *K* genes that support, or do not support, the prediction for the corresponding data source, along with their associated profiles (for the transcriptomic and proteomic sources, see Figure 1). A link towards the AmiGO website http://amigo.geneontology.org allows the user to quickly retrieve additional information on this term.

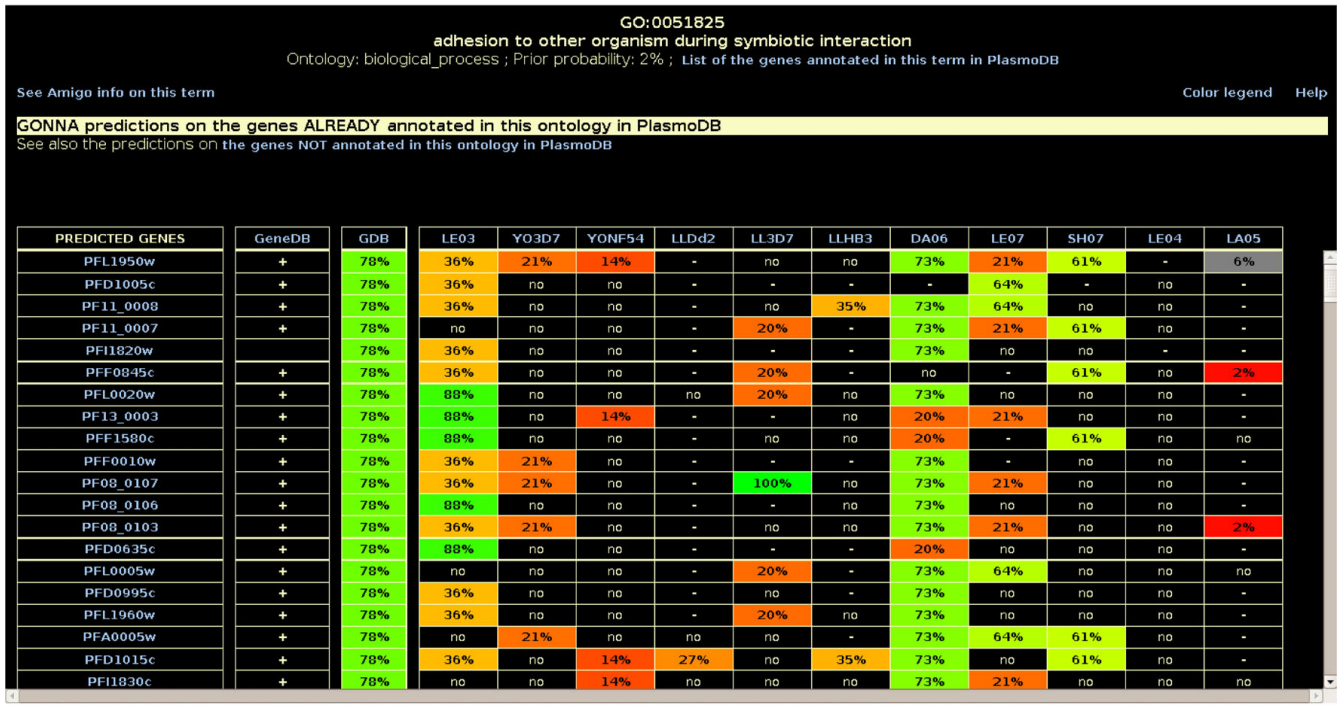
The gene view

The gene view displays the different GO terms that are predicted for each gene by Gonna. These terms are shown in a hierarchical way which follows the ontology structure (see Figure 4). There are three gene views for each gene, which correspond to the three GO ontologies. Each term is followed by its prior probability, the *GDB* of the prediction, and the *TDR*s associated with all data sources that support it. Moreover, for genes that already possess GeneDB annotations in the selected ontology, a ' + ' symbol after the term name indicates that this term already annotates this gene in GeneDB. As for the term view, clicking on a specific *TDR* opens a new window that provides additional information about the corresponding prediction. A link to PlasmoDB allows the user to quickly retrieve additional information on this gene. Note that *TDR*s and *GDB*s associated with the terms usually increase when scrolling toward the top of the ontology, because the prior probabilities of the terms increase. However, they may also decrease sometimes: If a GO term t is a generalization of one term t' with a good postgenomic signature (high *TDR*) and one term t'' with a poor signature (low *TDR*), genes predicted in t'' may have an unfavorable impact on the *TDR* estimation of t which may be lower than that of t'.

Discussion

Annotation quality

Quantity and quality of the available annotations used in the prior knowledge database to generate the predictions is a key point of any GBA approach. For *P. falciparum*, both quantity and quality are questionable. For example, in the BP ontology, of the 1799 genes (35%) possessing annotations, only 228 (13%) have annotations with experimental evidence; annotations of the 1571 remaining genes come from sequence similarity with proteins of other organisms (ISS and IEA evidence codes), and for 1067



GO:0051825
adhesion to other organism during symbiotic interaction
Ontology: biological_process ; Prior probability: 2% ; List of the genes annotated in this term in PlasmoDB

See Amigo info on this term — Color legend — Help

GONNA predictions on the genes ALREADY annotated in this ontology in PlasmoDB
See also the predictions on the genes NOT annotated in this ontology in PlasmoDB

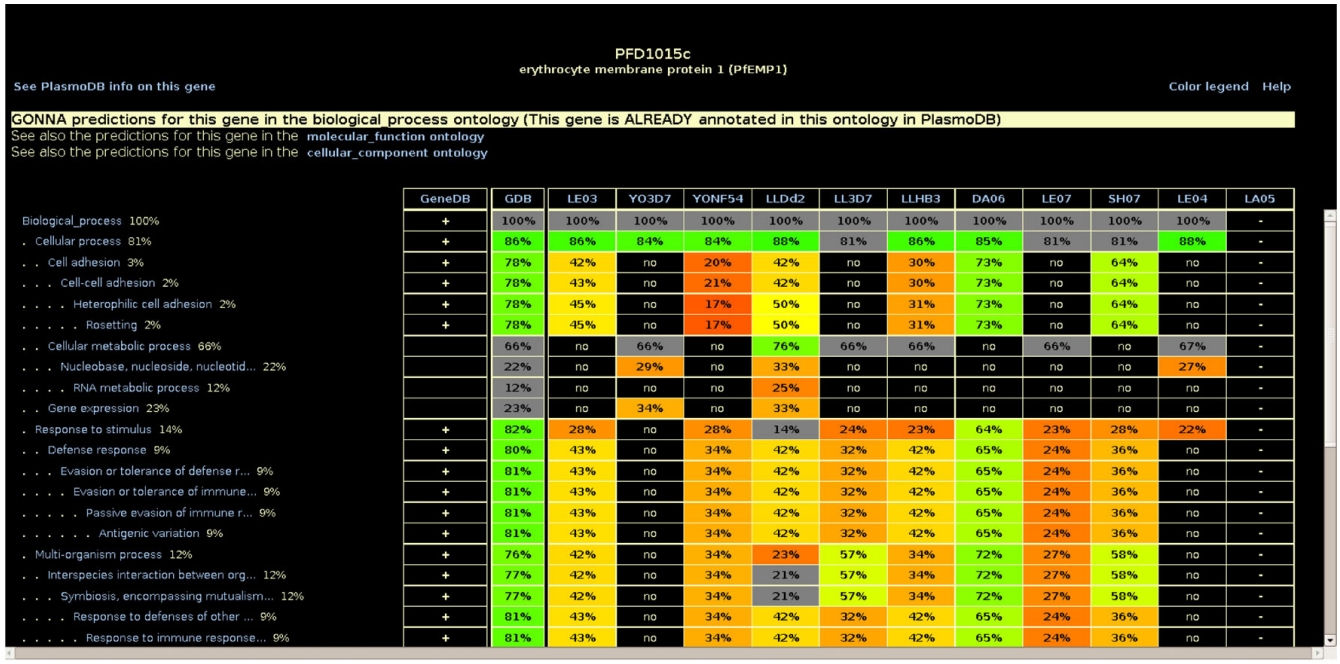
PREDICTED GENES	GeneDB	GDB	LE03	YO3D7	YONF54	LLDd2	LL3D7	LLHB3	DA06	LE07	SH07	LE04	LA05
PFL1950w	+	78%	36%	21%	14%	-	no	no	73%	21%	61%	-	6%
PFD1005c	+	78%	36%	no	no	-	-	-	-	64%	-	no	-
PF11_0008	+	78%	36%	no	no	-	no	35%	73%	64%	no	no	-
PF11_0007	+	78%	no	no	no	-	20%	-	73%	21%	61%	no	-
PFI1820w		78%	36%	no	no	-	-	-	73%	no	no	-	-
PFF0845c	+	78%	36%	no	no	-	20%	-	no	-	61%	no	2%
PFL0020w	+	78%	88%	no	no	no	20%	no	73%	no	no	no	-
PF13_0003	+	78%	88%	no	14%	-	-	no	20%	21%	no	no	-
PFF1580c	+	78%	88%	no	no	-	no	no	20%	-	61%	no	no
PFF0010w	+	78%	36%	21%	no	-	-	-	73%	-	no	no	-
PF08_0107	+	78%	36%	21%	no	-	100%	no	73%	21%	no	no	-
PF08_0106	+	78%	88%	no	no	-	-	no	73%	no	no	no	-
PF08_0103	+	78%	36%	21%	no	-	no	no	73%	21%	no	no	2%
PFD0635c	+	78%	88%	no	no	-	-	-	20%	no	no	no	-
PFL0005w	+	78%	no	no	no	-	20%	-	73%	64%	no	no	no
PFD0995c	+	78%	36%	no	no	-	no	-	73%	no	no	no	-
PFL1960w	+	78%	36%	no	no	-	20%	no	73%	no	no	no	-
PFA0005w	+	78%	no	21%	no	no	no	-	73%	64%	61%	no	-
PFD1015c	+	78%	36%	no	14%	27%	no	35%	73%	no	61%	no	-
PFI1830c	+	78%	no	no	14%	no	no	no	73%	21%	no	no	no

Figure 3
An extract of the predictions achieved in term "adhesion to other organism during symbiotic interaction" (GO:0051825). The "no" entry indicates that the data source does not support the prediction, while "-" means that no data are available in the source for this gene. By clicking on a *TDR*, the *K* characterized nearest neighbors that support/do not support this prediction are shown (see Figure 1). Clicking on any gene opens the corresponding gene view.

genes (68%) these annotations have IEA code, indicating that they have not been reviewed by a curator. Moreover, of the 431 different BP GO terms associated with the *P. falciparum* genes when considering all annotations, 172 (40%) are associated with IEA annotations only. For example, all annotations involving BP GO terms `ATP biosynthetic process (GO:0006754)`, `immune response (GO:0006955)` or `methylation (GO:0032259)`, as well as their descendants terms, possess IEA code only. Hence, we decided to consider all available GO annotations when generating the PlasmoDraft database. Removing all non-curated annotations from the prior knowledge database would eliminate not only numerous characterized genes, but also numerous GO terms, which would render impossible any new prediction in these parts of the ontology.

Experiments on a well annotated organism

In these conditions, it was relevant to check the method on a well annotated organism, using only experimental evidence code annotations as input for the predictions and for estimating the *TDR*s. To this end, we applied Gonna on the transcriptomic data set published by Spellman et al. (1998) [34], which monitors the expression level of yeast genes along the cell cycle. The same parameters as for *P. falciparum* were used, *i.e.* neighbor genes were selected using the Pearson correlation coefficient and we used two sets of parameters (K, K'): ($K = 6$, $K' = 4$) and ($K = 6$, $K' = 2$). All annotations different from IEA, ISS and RCA were used (gene annotation file revision 1.1323, downloaded from the GO website), which involves 4 165 genes characterized in the BP ontology, and a total of 1 220 different GO terms. The *TDR*s were estimated for each GO term by cross-validation. Figure 5 represents the *TDR*s associated with all BP GO terms where predictions are proposed by Gonna, as a function of the prior probability of the terms. We see that for numerous terms, predictions are made with a *TDR* significantly higher than the prior probability of the term, which shows the potential of the approach to decipher biological functions from gene expression data. For comparison purpose, the same experiment was achieved on *P. falciparum* with the time series of Bozdech et al. (2003) [27] using all available BP annotations (see Figure 6). While, as expected, the number of GO terms where predictions are made is lower than for yeast, numerous GO terms are also predicted with high *TDR*s. Though the reliability of these predictions could depends on the prior (IEA) annotations, the similarity of Figures 5 and 6 is quite encouraging and shows that *P. fal-*



PFD1015c
erythrocyte membrane protein 1 (PfEMP1)

See PlasmoDB info on this gene — Color legend Help

GONNA predictions for this gene in the biological process ontology (This gene is ALREADY annotated in this ontology in PlasmoDB)
See also the predictions for this gene in the molecular_function ontology
See also the predictions for this gene in the cellular_component ontology

	GeneDB	GDB	LE03	YO3D7	YONF54	LLDd2	LL3D7	LLHB3	DA06	LE07	SH07	LE04	LA05
Biological_process 100%	+	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	-
. Cellular process 81%	+	86%	86%	84%	84%	88%	81%	86%	85%	81%	81%	88%	-
. . Cell adhesion 3%	+	78%	42%	no	20%	42%	no	30%	73%	no	64%	no	-
. . . Cell-cell adhesion 2%	+	78%	43%	no	21%	42%	no	30%	73%	no	64%	no	-
. . . . Heterophilic cell adhesion 2%	+	78%	45%	no	17%	50%	no	31%	73%	no	64%	no	-
. . . . . Rosetting 2%	+	78%	45%	no	17%	50%	no	31%	73%	no	64%	no	-
. . Cellular metabolic process 66%		66%	no	66%	no	76%	66%	66%	no	66%	no	67%	-
. . . Nucleobase, nucleoside, nucleotid... 22%		22%	no	29%	no	33%	no	no	no	no	no	27%	-
. . . . RNA metabolic process 12%		12%	no	no	no	25%	no	no	no	no	no	no	-
. . Gene expression 23%		23%	no	34%	no	33%	no	no	no	no	no	no	-
. Response to stimulus 14%	+	82%	28%	no	28%	14%	24%	23%	64%	23%	28%	22%	-
. . Defense response 9%	+	80%	43%	no	34%	42%	32%	42%	65%	24%	36%	no	-
. . . Evasion or tolerance of defense r... 9%	+	81%	43%	no	34%	42%	32%	42%	65%	24%	36%	no	-
. . . . Evasion or tolerance of immune... 9%	+	81%	43%	no	34%	42%	32%	42%	65%	24%	36%	no	-
. . . . . Passive evasion of immune r... 9%	+	81%	43%	no	34%	42%	32%	42%	65%	24%	36%	no	-
. . . . . Antigenic variation 9%	+	81%	43%	no	34%	42%	32%	42%	65%	24%	36%	no	-
. Multi-organism process 12%	+	76%	42%	no	34%	23%	57%	34%	72%	27%	58%	no	-
. . Interspecies interaction between org... 12%	+	77%	42%	no	34%	21%	57%	34%	72%	27%	58%	no	-
. . . Symbiosis, encompassing mutualism... 12%	+	77%	42%	no	34%	21%	57%	34%	72%	27%	58%	no	-
. . . . Response to defenses of other ... 9%	+	81%	43%	no	34%	42%	32%	42%	65%	24%	36%	no	-
. . . . . Response to immune response... 9%	+	81%	43%	no	34%	42%	32%	42%	65%	24%	36%	no	-

Figure 4

An extract of predictions achieved for gene PFD1015c in the BP ontology. The "no" entry indicates that the data source does not support the prediction, while "-" means that no data are available in the source for this gene. By clicking on a *TDR*, the *K* characterized nearest neighbors that support/do not support this prediction are shown (Figure 1). Clicking on any term opens the corresponding GO term view.

ciparum annotations are globally consistent, as they are mostly recovered using a transcriptomic data set.

Contents of the database

By browsing the PlasmoDraft database, several predictions clearly involve rare GO terms (*i.e.* with low prior probability) with high *TDR*s or *GDB*s. For example, in the BP ontology, 16 uncharacterized genes are predicted in `establishment of localization (GO:0051234)` (prior probability 15%, *GDB* 78%), 25 uncharacterized genes are predicted in `Rosetting (GO:0020013)` (prior probability 2%, *GDB* 78%), and 50 uncharacterized genes are in `Pathogenesis (GO:0009405)` (prior probability 4%, *GDB* 75%). Similarly (but with lower *GDB*s), 13 uncharacterized genes are predicted in `Ubiquitin-dependent protein catabolic proces(GO:0006511)` (prior probability 2%, *GDB* 50%), and 12 uncharacterized genes are in `Biopolymecata bolic process (GO:0043285)` (prior probability 3%, *GDB* 56%). Moreover the best *TDR*s are not always achieved with the same data source. For example, for the `Antigenic variation (GO:0020033)` term, the LE03 data [14] provides more accurate predictions than the LLHB3/LLDd2/L13D7 series [27,28], may be because this function has a more apparent expression signature when considering the entire life cycle of the organism. For functions such as `DNA packaging (GO:0006323)` however, the highest *TDR* is achieved with the LLHB3 data set [27] because the function is better monitored at the cell cycle level.

We estimated the amount of new information provided by PlasmoDraft in a systematic way. For the BP ontology, PlasmoDraft proposes significant annotations on GO terms of low prior probability (below 25%) for 3 900 genes, among which 2 434 have no BP annotations in GeneDB. With CC and MF ontologies, 1 905 and 1 540 uncharacterized genes are annotated by PlasmoDraft on low prior probability GO terms, respectively. The interest of these annotations of course depends on the associated *GDB*. Thus, given a *GDB* threshold (*e.g.* 75%) and an ontology, for each uncharacterized gene in this ontology we searched the GO term with the lowest prior probability wherein the gene is predicted with a statistically significant *GDB* above the threshold. Figure 7 summarizes these results on the three ontologies. From this figure we see, for example, that for the BP ontology 290 uncharacterized genes in GeneDB are predicted with a *GDB* above 75% (red curve) on a GO term with a prior probability below 0.10. In the same manner, 1 025 uncharacterized genes are predicted with a *GDB* above 50% (blue curve) on a GO term with a prior probability below 0.25. For the CC

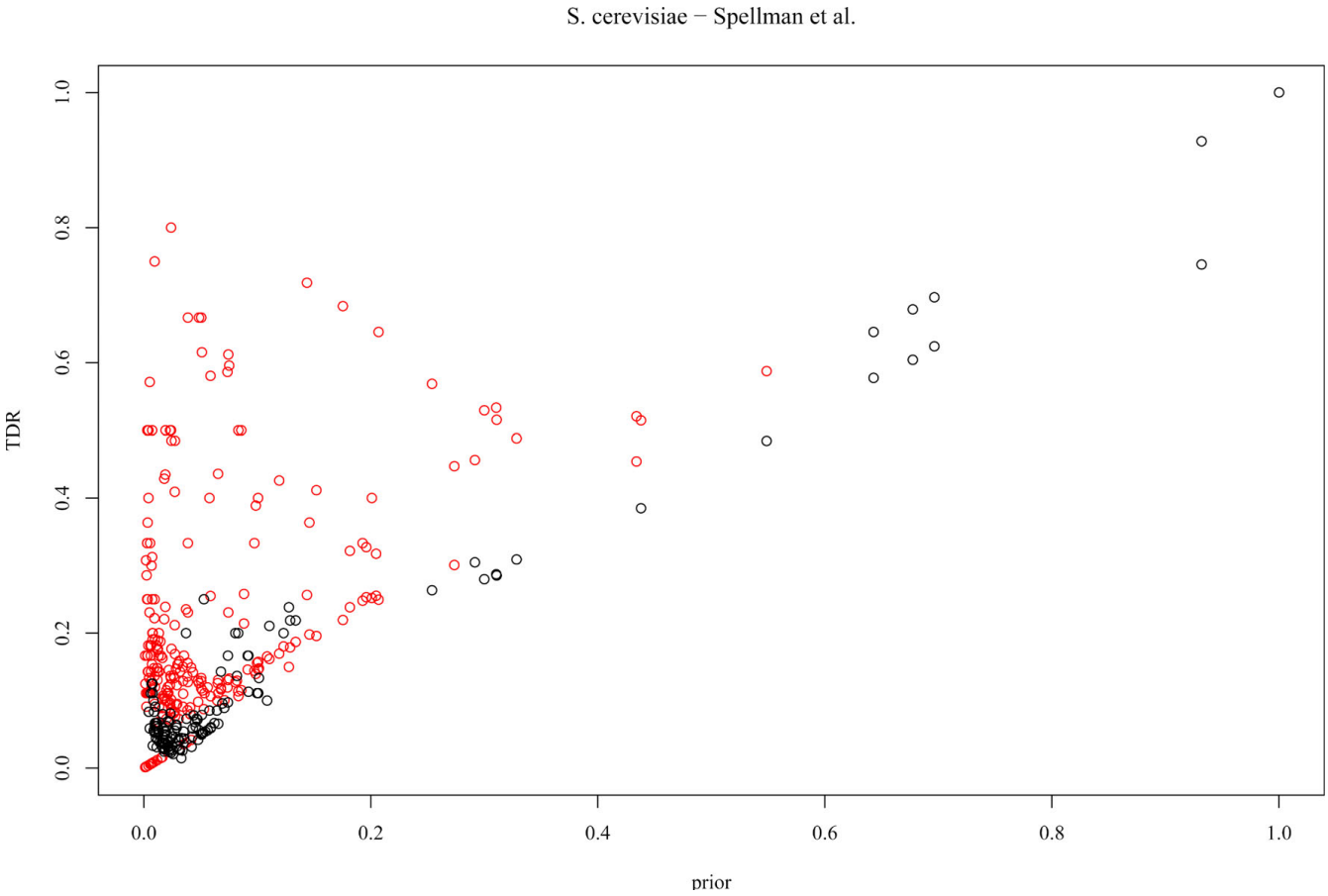


Figure 5
Gonna performance on yeast. Gonna was applied to the transcriptomic data set published by Spellman et al. (1998) [34] using experimental evidence code annotations only as prior knowledge database. *TDR*s of all BP GO terms where predictions are proposed by Gonna are plotted as a function of the prior probability of the terms. Red and black points indicate significant and non-significant *TDR*s, respectively.

and MF ontologies, 740 and 329 genes are predicted with a *GDB* above 50% on a GO term with a prior probability below 0.25, respectively. Note that only genes without any annotation in GeneDB in the selected ontology are considered in this measure, while the PlasmoDraft database also provides additional annotations for many genes that are already annotated in this ontology. By comparing the results achieved on the different ontologies, we see that the BP ontology provides the best results. This is not surprising, as the signature detected in the postgenomic data by GBA methods are mostly characteristic of biological processes [8]. However, by an information propagation phenomenon, the BP signatures may sometimes help for predicting annotations in the two other ontologies. This happens, for example, when many genes with a given molecular function (or exported in a particular cellular component) are involved in a biological process with a strong signature. For example. GO term `host cell plasma membrane (GO:0020002)` in the CC ontology is associated with high *GDB* (72%), because most genes belonging to this term are also associated with the biological process `Defense response (GO:0006952)` which is well recognized.

A similar approach can be used to estimate the amount of new information provided by each data source independently. For example, Figure 8 reports the number of uncharacterized genes in the BP ontology that can be annotated with a *TDR* above 75%, 50% and 25% by the transcriptomic data of Bozdech et al. (2003) [27], and by the interactomic data of LaCount et al. (2005) [33]. We can see that more than 73 genes are associated with a GO term of prior probability below 10% with a TDR above 50% using the transcriptomic data, while 10 genes only are predicted with the same thresholds using interactomic data. This indicates that the interactome tends to provide less functional signal than the transcriptome, partly because less genes are monitored.

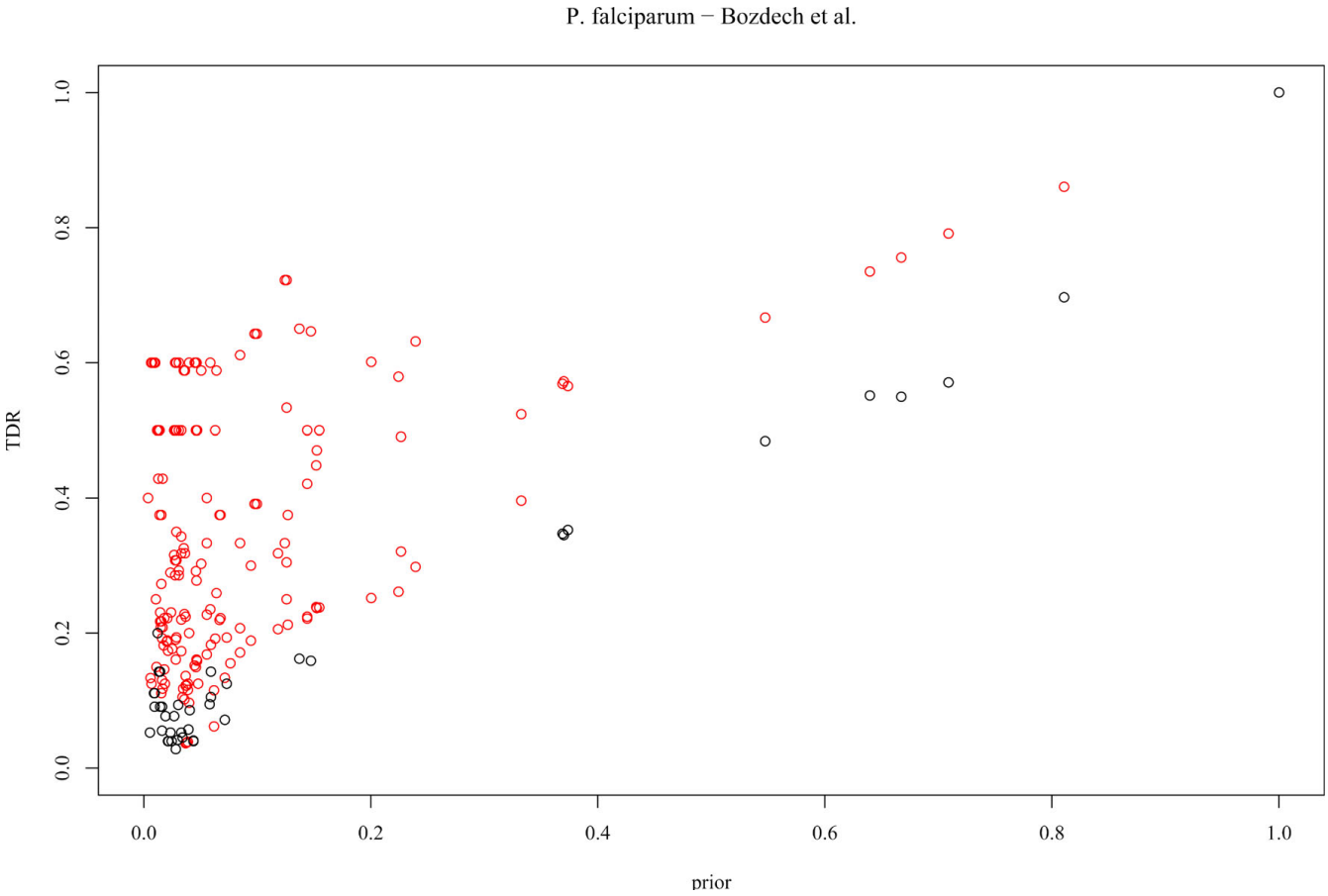


Figure 6
Gonna performances on the transcriptomic data set published in Bozdech et al. (2003) [27]. *TDR*s of all BP GO terms where predictions are proposed by Gonna are plotted as a function of the prior-probability of the terms. Red and black points indicate significant and non-significant *TDR*s, respectively.

Assessment of the GDB*s*

*TDR*s and *GDB*s are estimated by cross-validation by applying Gonna on the already characterized genes. This procedure produces unbiased estimates of the method accuracy, provided that the uncharacterized genes share approximately the same distribution as the characterized ones [19]. However, since *TDR*s and *GDB*s are sometimes estimated on small numbers of predictions, users should be aware that for some specific GO terms, the accuracy on the uncharacterized genes may differ from the reported *TDR*s and *GDB*s. Nonetheless, these measures provide valuable indications on the potential functions of genes by pointing out the most likely GO terms. To assess this point, we compared the PlasmoDraft predictions proposed for the uncharacterized genes to the annotations of their homologous genes in yeast when these are known. We looked on the 986 genes without BP annotations that have been predicted with high *GDB* (above 50%) on specific BP terms (prior probability below 25%). As expected, few genes among these 986 can be associated with a characterized orthologous gene in *S. cerevisiae*. However, a reciprocal best hit procedure using BLASTP with an e-value cutoff of 10^{-5} allows to find *S. cerevisiae* orthologues for 141 genes. Among these 141 orthologous pairs, 63 (45%) have "concordant" annotations with the high *GDB* predictions. Here we consider that annotations are concordant if at least half of the terms with prior probability below 25% are shared by the *S. cerevisiae* orthologue. As expected, this proportion decreases when using PlasmoDraft predictions with lower *GDB*s. For example, 2 271 genes without BP annotations are predicted with a *GDB* between 25% and 50% on a GO term with prior probability below 25%. Among these, 245 can be associated with *S. cerevisiae* orthologues by reciprocal best hit, and 71 (29%) have concordant annotations.

Comparison with the predictions of Zhou et al. (2008) [35]

During the writing of this article, another database [35] of gene function predictions based on the OPI method described in reference [17] was published. Briefly, OPI is

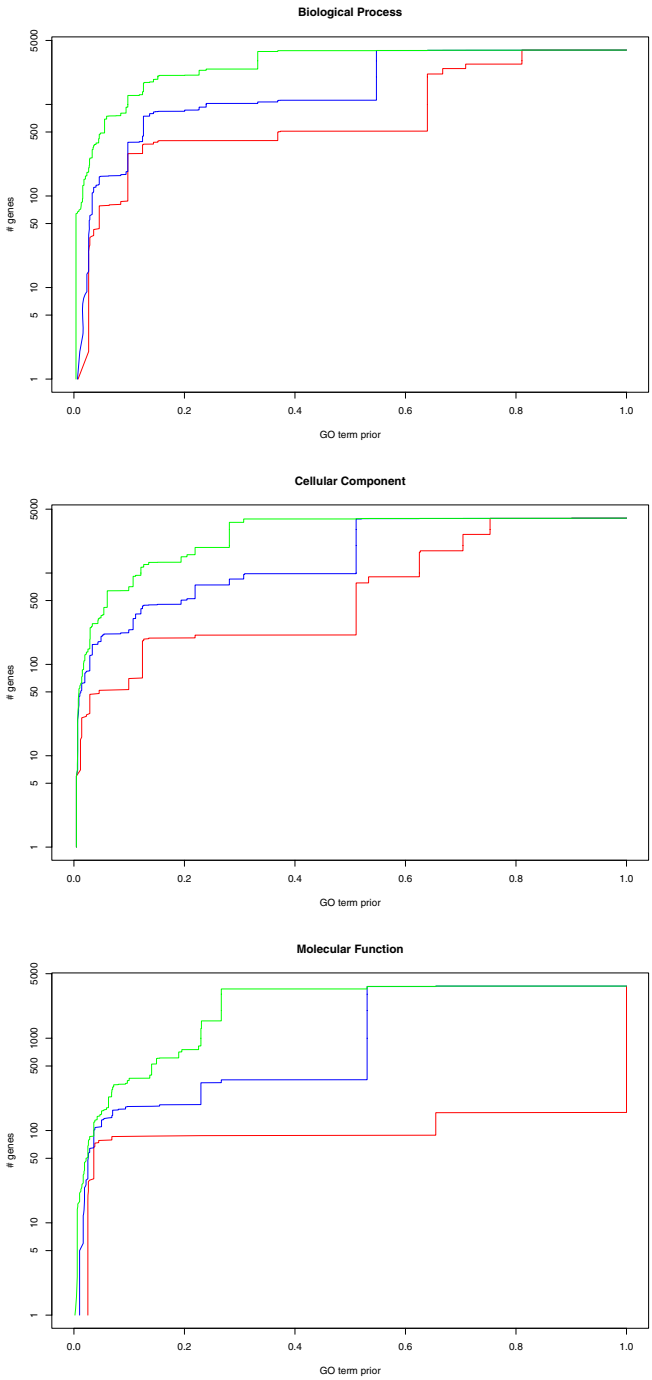


Figure 7
Estimate of the amount of new information supplied in PlasmoDraft. Estimates for the BP (up) CC (middle) and MF (down) ontologies. Red, blue and green lines represent the results achieved with *GDB* thresholds of 75%, 50% and 25%, respectively. The *x*-axis gives the prior probabilities of the terms, while the *y*-axis (in log scale) reports the number of uncharacterized genes in the ontology that have been predicted with a *GDB* above the threshold, on a GO term with prior probability below *x*.

a supervised method that works as follows. For each GO term, OPI uses a set of "seed" genes already annotated with this term to construct an average expression profile. Next, all genes (annotated or not) are ranked according to their similarity to this average profile and a statistical test is used to identify the rank cutoff that includes the largest number of seed genes within the smallest cluster size. All genes before this cutoff are then considered as potentially related to the GO term under consideration. The database [35] exploits a single new transcriptomic data set covering all life cycle stages of the parasite and combining gene expressions from both *P. yoelii* and *P. falciparum*. As both the methods and data sources are different, this database and PlasmoDraft provide different and complementary information. OPI provides BP annotations for 1 902 different genes, among which 1 036 have no BP annotations in GeneDB. When looking at the PlasmoDraft predictions with *GDB* above 50% (which involves 1 111 uncharacterized genes in BP), only 230 also have BP predictions in OPI. However, when both methods propose BP predictions for a gene, the predictions are often similar. Indeed, of the 230 common genes, 94 have concordant predictions – *i.e.* at least half of the predictions of one of the methods involving terms with a prior probability below 25% are also predicted by the other method. Differences in specificity of OPI and PlasmoDraft can also be observed by comparing the *GDB* and *FDR* (false discovery rate) estimates associated with a given GO term by PlasmoDraft and OPI, respectively. Recall that the *GDB* is actually the *TDR* associated with the predictor that combines all data sources. Moreover, by definition, the *FDR* equals 1 minus the *TDR* on this term. While *FDR*s of OPI are not estimated by cross-validation, we can nevertheless get a rough idea of which method provides the best results for a given GO term. For example, OPI obtains the highest *TDR*s on terms like `Entry into host (GO:0044409)`, or `Mitochodrion organization and biogenesis (GO:0007005)` (~90% and ~30% vs. 36% and ~5%), while for terms like `Interaction between organisms (GO:0044419)` or `Rosetting (GO:0020013)`, PlasmoDraft obtains the best results (77% and 78% vs. 25% and no statistically significant *TDR*). On the whole, it thus appears that the two databases use quite different data sources and provide interesting information on different types of functions and different genes, so the community will likely benefit from both.

Conclusion

We presented PlasmoDraft, an extensive database of GO annotation predictions that are achieved by Guilt By Association using most postgenomic data available to date for *P. falciparum*. All predictions come with a confidence estimate computed by cross-validation. The database is presented in a friendly interface that allows easy browsing

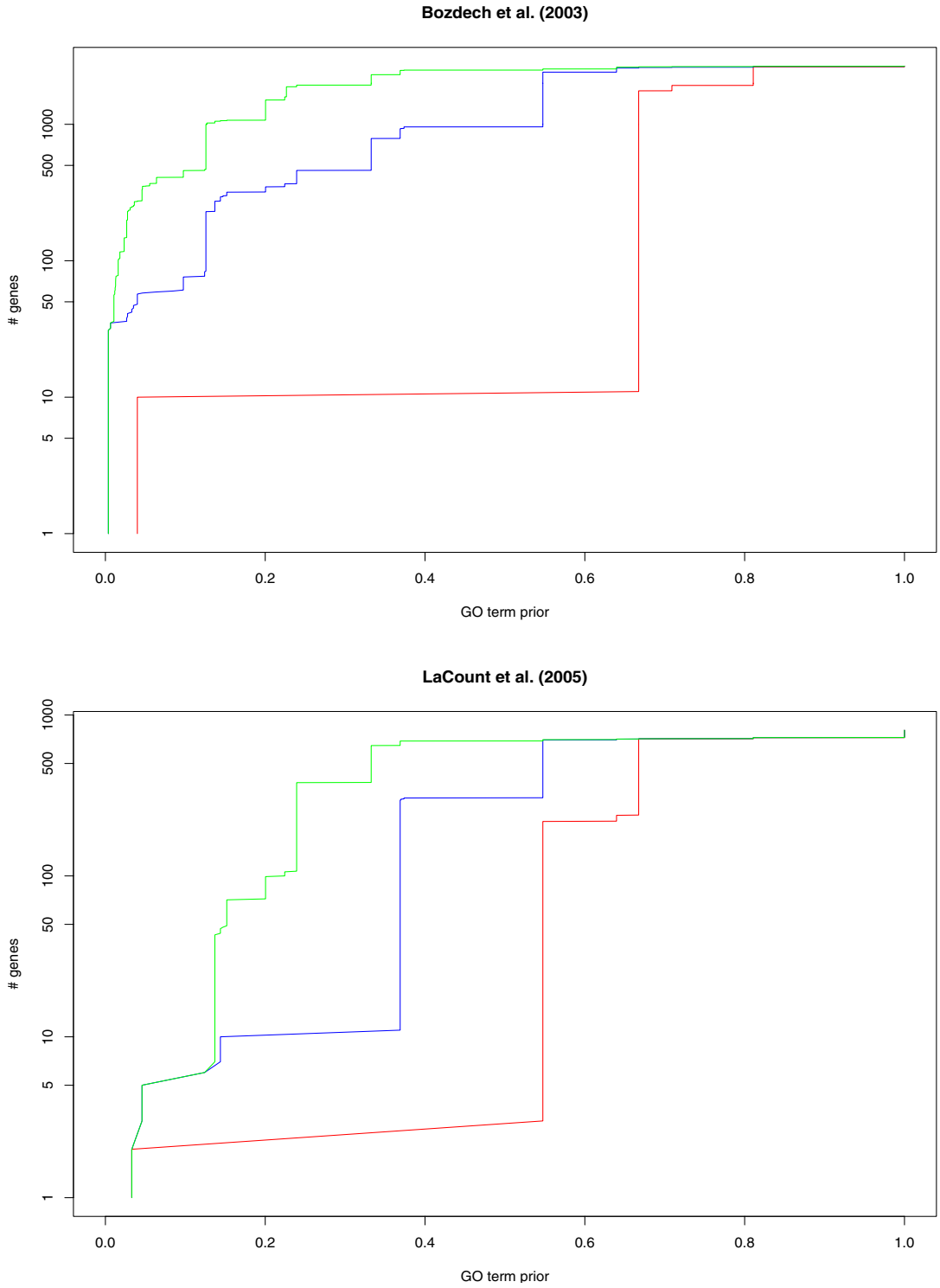


Figure 8

Estimate of the amount of new information supplied by the transcriptomic data source of Bozdech et al. (2003) [27] **and the interactomic data source of LaCount et al. (2005)** [33]. Red, blue and green lines represent the results achieved with *TDR* thresholds of 75%, 50% and 25%, respectively. The *x*-axis gives the prior probabilities of the terms, while the *y*-axis (in log scale) reports the number of uncharacterized genes in the ontology that have been predicted with a *TDR* above the threshold, on a GO term with prior probability below *x*.

and querying, and proposes high confidence annotations for several hundreds of genes without any annotations, as well as additional annotations for many already characterized genes. One prospect is the integration of *compendiums* of gene expression data sets as new data sources in PlasmoDraft. These data, obtained by concatenation of several data sets of diversified biological conditions, have shown to often provide strong biological function signatures [36]. However, predictions based on these data may be difficult to interpret for biologists, and their integration opens new issues in data selection and combination.

As mentioned in the Methods, one advantage of Gonna concerns its genericness that allows its use on any new data, as long as a relevant similarity measure can be computed; a set of scripts then enables regeneration of the database to integrate the new data set in a fully automated way. This also holds for the GO annotations used as prior knowledge, and the new annotations provided by the community in the future will be easily integrated. Most notably, we are aware that a collegiate effort for re-annotating *P. falciparum* proteins should provide new/curated functional annotations in the near future. This should improve both the quantity and the quality of the PlasmoDraft predictions. In the same way, while in the current version of PlasmoDraft all GO annotations are considered (*i.e.* including automatically-assigned annotations) due to the scarcity of curated annotations, it is possible that the re-annotation effort will enable the use of only curator-assigned annotations in the subsequent versions of PlasmoDraft. Thanks to these new advances, PlasmoDraft should become more and more accurate and useful to the community.

Availability and requirements

PlasmoDraft is freely available at http://atgc.lirmm.fr/PlasmoDraft/

Authors' contributions

LB conceived, designed and implemented the method, carried out the analyses, designed the database and drafted the manuscript. JFD designed and developed the database. OG initiated the project, designed the method, participated in the analyses, designed the database and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by the PlasmoExplore project of the French National Research Agency (ANR-Q6-CIS6-MDCA-14). We thank all members of this project for useful discussions.

References

1. Sachs J, Malancy P: **The economic and social burden of malaria.** *Nature* 2002, **415(6872):**680-685.
2. Gardner M, Hall N, Fung E, White O, Berriman M, Hyman R, Carlton J, Pain A, Nelson K, Bowman S, Paulson I, James K, Eisen J, Rutherford K, Salzberg S, Craig A, Kyes S, Chan M, None V, Shallom S, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather M, Vaidya A, Martin D, Fairlamb A, Fraunholz M, Roos D, Ralph S, McFadden G, Cummings L, Subramanian G, Mungall C, Venter J, Carucci D, Hoffman S, Newbold C, Davis R, Fraser C, Barrell B: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419(6906):**498-511.
3. Pizzi E, Frontali C: **Low-complexity regions in Plasmodium falciparum proteins.** *Genome Res* 2001, **11(2):**218-229.
4. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3):**403-410.
5. Sonnhammer E, Eddy S, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains.** *Nucleic Acids Res* 1998, **26:**320-322.
6. Bastion O, Roy S, Marechal E: **Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions.** *C R Biol* 2005, **328(5):**445-453.
7. Marcotte E: **Computational genetics: finding protein function by nonhomology methods.** *Curr Opin Struct Biol* 2000, **10(3):**359-365.
8. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95(25):**14863-14868.
9. Lockhart D, Winzeler E: **Genomics, gene expression and DNA arrays.** *Nature* 2000, **405(6788):**827-836.
10. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B: **Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.** *Genome Biology* 2003, **5:**.
11. Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein-protein interaction networks.** *Nat Biotechnol* 2003, **21(6):**697-700.
12. Chen Y, Xu D: **Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae.** *Nucleic Acids Research* 2004, **32(21):**6414-6424.
13. Wu L, Hughes T, Davierwala A, Robinson M, Stoughton R, Altschuler S: **Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31(3):**255-265.
14. Le Roch K, Zhou Y, Blair P, Grainger M, Moch J, Haynes J, Vega PDL, Holder A, Batalov S, Carucci D, Winzeler E: **Discovery of gene function by expression profiling of the malaria parasite life cycle.** *Science* 2003, **301(5639):**1503-1508.
15. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C, Furey TS, Manuel Ares J, Haussler D: **Knowledge-based Analysis of Microarray Gene Expression Data Using Support Vector Machines.** *Proc Natl Acad Sci USA* 2000, **97(1):**262-267.
16. Mateos A, Dopazo J, Janson R, Tu Y, Gerstein M, Stolovitzky G: **Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons.** *Genome Research* 2002, **12(11):**1703-1715.
17. Zhou Y, Young J, Santrosyan A, Chen K, Yan S, Winzeler E, Young J, Fivelman Q, Blair P, do la Vega P, KG LR, Zhou Y, Carucci D, Baker D, Winzeler E: **In silico gene function prediction using ontology-based pattern identification.** *Bioinformatics* 2005, **21(7):**1237-1245.
18. Toronen P: **Selection of informative clusters from hierarchical cluster tree with gene classes.** *BMC Bioinformatics* 2004, **5:**.
19. Hastie T, Tibshirani R, Friedman J: *The elements of statistical learning* Springer; 2001.
20. Yona G, Dirks W, Rahman S, Lin D: **Effective similarity measures for expression profiles.** *Bioinformatics* 2006, **22(13):**1616-1622.
21. Dice L: **Measure of the amount of ecologic association between species.** *Ecology* 1945, **26:**297-302.
22. Reimand J, Kull M, Peterson H, Hansen J, Vilo J: **g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments.** *Nucleic Acids Res* 2007:W193-W200.
23. Chmielewski MR, Grzynala-Busse JW: **Global discretization of continuous attributes as preprocessing for machine learning.** *International Journal of Approximate Reasoning* 1996, **15:**319-331.
24. Dougherty J, Kohavi R, Sahami M: **Supervised and Unsupervised Discretization of Continuous Features.** *International Conference on Machine Learning* 1995:194-202.
25. Langley P, Sage S: **Induction of selective Bayesian classifiers.** In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* Morgan Kaufmann; 1994:399-406.

26. Young J, Fivelman Q, Blair P, do la Vega P, Le Roch K, Zhou Y, Carncci D, Baker D, Winzeler E: **The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification.** *Mol Biochem Parasitol* 2005, **143:**67-79.
27. Bozdech Z, Llinas M, Pulliam B, Wong E, Zhu J, DeRisi J: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1:**.
28. Llinas M, Bozdech Z, Wong E, Adai A, DeRisi J: **Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains.** *Nucleic Acids Res* 2006, **34(4):**1166-1173.
29. Dahl E, Shock J, Shenai B, Gut J, DeRisi J, Rosenthal P: **Tetracyclines specifically target the apicoplast of the malaria parasite Plasmodium falciparum.** *Antimicrob Agents Chemother* 2006, **50(9):**3124-3131.
30. Shock J, Fischer K, Derisi J: **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(7):**.
31. Florens L, Washburn M, Raine J, Anthony R, Grainger M, Haynes J, Moch J, Muster N, Sacci J, Tabb D, Witney A, Wolters D, Wu Y, Gardner M, Holder A, Sinden R, Yates J, Carucci D: **A proteomic view of the Plasmodium falciparum life cycle.** *Nature* 2002, **419(6906):**520-526.
32. Le Roch K, Johnson J, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan S, Williamson K, Holder A, Carucci D 3rd, Yates J, Winzeler E: **Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle.** *Genome Res* 2004, **14(11):**2308-2318.
33. LaCount D, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth J, Schoenfeld L, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes R: **A protein interaction network of the malaria parasite Plasmodium falciparum.** *Nature* 2005, **438(7064):**103-107.
34. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PC), Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9(12):**3273-3297.
35. Zhou Y, Ramachandran V, Kumar KA, Westenberger S, Refour P, Zhou B, Li F, Young JA, Chen K, Plouffe D, Henson K, Nussenzweig V, Carlton J, Vinetz JM, Duraisingh MT, Winzeler EA: **Evidence-Based Annotation of the Malaria Parasite's Genome Using Comparative Expression Profiling.** *PLoS ONE* 2008, **3(2):**e1570.
36. Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, Storz G, Botstein D, Brown P: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11(12):**4241-4257.