

## Supplementary tables and figures for the manuscript “Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*”

Nicolas Terrapon<sup>1,2</sup>, Olivier Gascuel<sup>1</sup>, Éric Maréchal<sup>2</sup> and  
Laurent Bréhélin<sup>1</sup>,

<sup>1</sup>LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada 34392 Montpellier Cedex 5 France

<sup>2</sup>CEA Grenoble iRTSV/LPCV, 17 rue des Martyrs, 38054 Grenoble cedex 9 France

**Supp. Table 1.** Number of distinct Pfam domains and protein coverage in eukaryote proteomes (Pfam v23.0)

Organism	Proteome size	Pfam domains	Coverage
<i>A. Gambiae</i>	12 347	2 991	74%
<i>A. thaliana</i>	34 517	3 125	74%
<i>C. elegans</i>	22 637	2 953	65%
<i>D. melanogaster</i>	16 224	3 129	72%
<i>D. rerio</i>	31 844	3 384	84%
<i>H. sapiens</i>	40 252	3 914	68%
<i>S. cerevisiae</i>	5 862	2 369	76%
<i>P. falciparum</i>	5 460	1 429	53%
<i>P. vivax</i>	5 432	1 415	50%
<i>P. yoelii</i>	7 724	1 313	42%

This table reports the proteome size, the total number of distinct Pfam domains identified in the proteins, and the proportion of proteins where at least one Pfam domain is known.

**Supp. Table 2.** Proportion of domains in low complexity regions

	Domains in low comp. region	Residues in low comp. region
Known domains	3811/5782 (66%)	30
New domains <i>FDR</i> < 10%	292/452 (64%)	30
New domains <i>FDR</i> < 20%	454/818 (56%)	26

This table reports the proportion of known and new Pfam domains in *P. falciparum* proteins that overlap with a long low complexity region (length > 10 residues) (PlasmoDB 5.5), and the average number of residues in the low complexity regions of the overlapping domains. For example, 3811 out of the 5782 already known Pfam domains overlap with a long low complexity region. Moreover, in average, the low complexity region involves 30 residues in these 3811 domains.

**Supp. Table 3.** Newly certified domains in *P. vivax* and *P. yoelii* proteins

	<i>FDR</i>	$\leq 10\%$				$\leq 20\%$			
		Valid. dom.	Pfam	Interp.	Pot.	All	Pfam	Interp.	Pot.
<i>P. vivax</i>	Certif. Dom.	279	76	65	348	343	253	101	517
	New Interp.	227	46	56	274	290	274	89	417
	New Dom. Types	77	26	22	94	106	101	32	150
<i>P. yoelii</i>	Certif. Dom.	233	140	42	298	289	329	123	485
	New Interp.	195	98	32	236	249	267	106	406
	New Dom. Types	66	35	10	77	87	103	35	144

“Valid. dom.” indicates the type of validating domains used for certifications: “Pfam”, known Pfam domains from InterProScan; “Interp.”, known InterPro (non-Pfam) domains from InterProScan; “Pot.”, potential domains themselves; “All” indicates the results achieved when combining the 3 types of validating domains. “Certif. dom.” denotes the number of new certified domains, “New Interp.” indicates the number of certifications allowing us to identify a new InterPro Entry for the protein, and “New Dom. Types” indicates the number of domain types that were never previously detected in any *P. vivax* or *P. yoelii* proteins.

**Supp. Table 4.** New GO annotations of *P. vivax* and *P. yoelii* proteins

	<i>FDR</i>	Single	Combin. with	Unannot.
		Domains	Certified Dom.	prot.
<i>P. vivax</i>	$\leq 10\%$	144	119	37
	$\leq 20\%$	230	142	55
<i>P. yoelii</i>	$\leq 10\%$	122	99	28
	$\leq 20\%$	248	111	44

“Single Domains” is the number of new GO annotations brought by a single domain certified by our approach; “Combin. Known Dom.” is the number of GO annotations that can be deduced from combinations of already known domains thanks to inferred associations between domain combinations and GO annotations; “Combin. with Certified Dom.” is the number of supplementary GO annotations (different from the 2 previous columns) that can be deduced from combinations involving a newly certified domain. “Total Prot.” is the total number of proteins involved, and “Unannot. Prot.” is the number of proteins without any annotation for which an annotation has been proposed.

**Supp. Table 5.** Proportion of known and new *P. falciparum* domains present in *P. vivax* homologous proteins

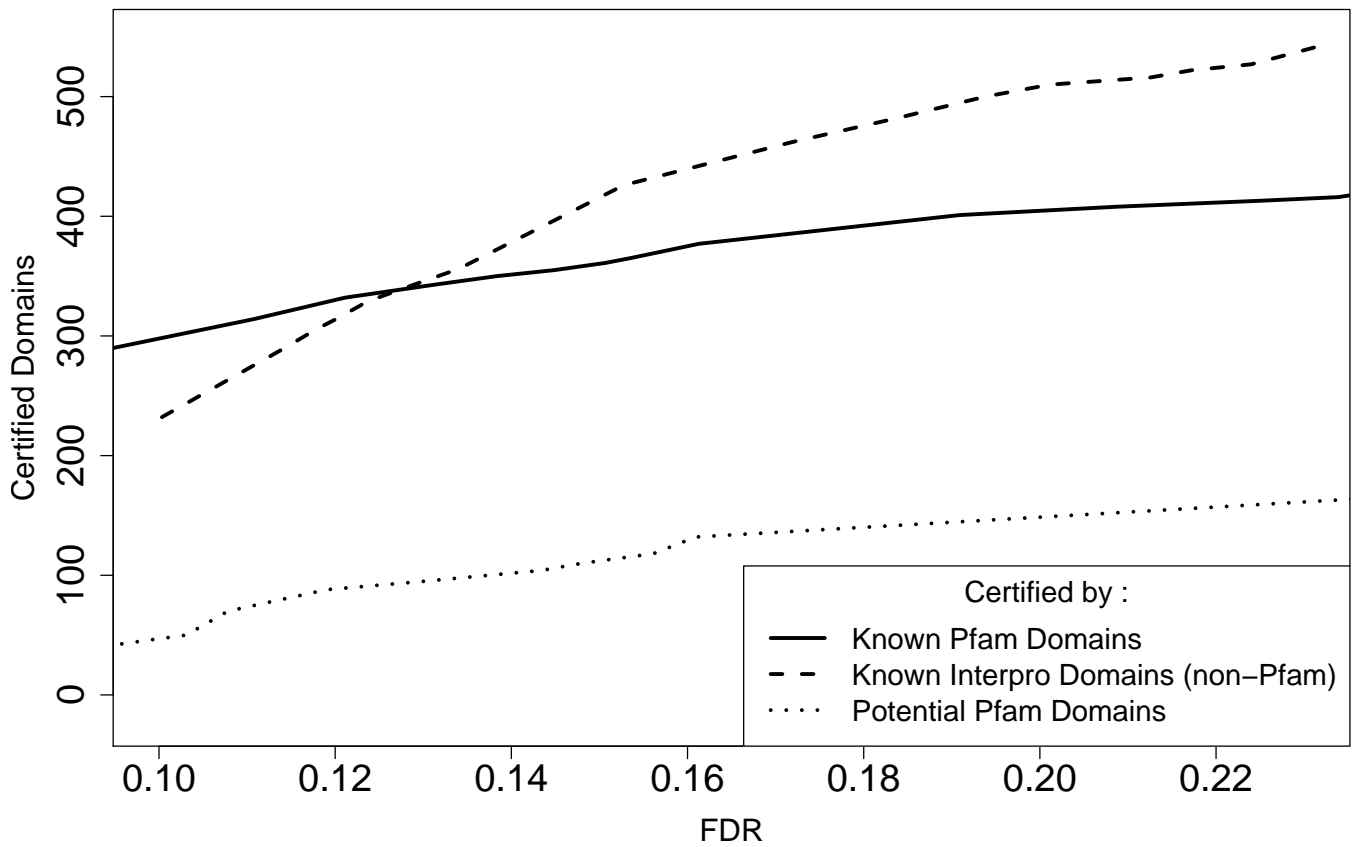
		Known dom.	<i>P. vivax</i>	
			New dom. <i>FDR</i> < 10%	New dom. <i>FDR</i> < 20%
<i>P. falciparum</i>	Known dom.	2985/3143 (95%)	34/3143 (1%)	55/3143 (2%)
	New dom. <i>FDR</i> < 10%	46/323 (14%)	205/323 (63%)	222/323 (69%)
	New dom. <i>FDR</i> < 20%	54/548 (10%)	233/548 (43%)	277/548 (51%)

This table reports the proportion of known and new Pfam domains in *P. falciparum* proteins with a known *P. vivax* homologue, which are also present in the known/new Pfam domains of their *P. vivax* homologue. For example, 3143 known Pfam domains are in a *P. falciparum* protein with a known *P. vivax* homologue. Among these, 2985 are also known in their *P. vivax* homologue, and 34 are among the newly certified domains with *FDR* below 10% of the *P. vivax* proteome. Similarly, 323 newly certified *P. falciparum* domains with a *FDR* < 10% are in a protein with a known *P. vivax* homologue. Among these, 46 are already known in their *P. vivax* homologue, and 205 are newly certified with a *FDR* < 10% in their *P. vivax* homologue.

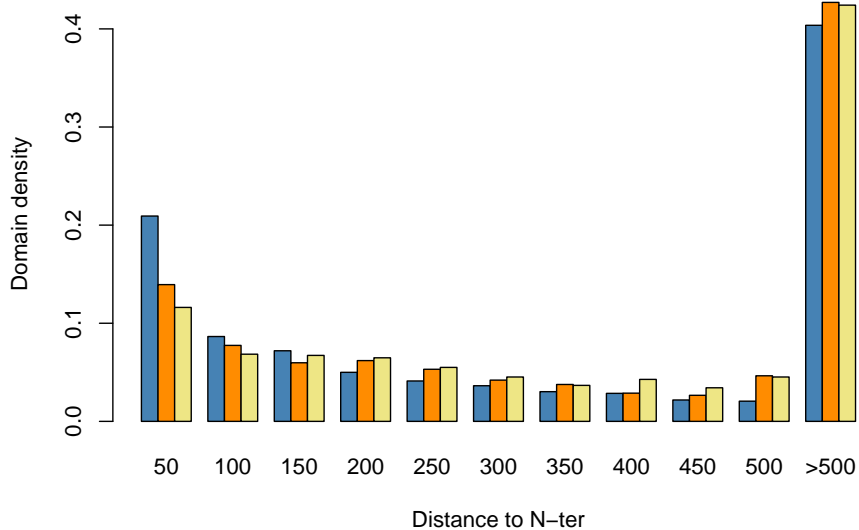
**Supp. Table 6.** Proportion of known and new *P. falciparum* domains present in *P. yoelii* homologous proteins

		Known dom.	<i>P. yoelii</i>	
			New dom. <i>FDR</i> < 10%	New dom. <i>FDR</i> < 20%
<i>P. falciparum</i>	Known dom.	2700/2998 (90%)	42/2998 (1%)	50/2998 (2%)
	New dom. <i>FDR</i> < 10%	41/314 (13%)	158/314 (50%)	169/314 (54%)
	New dom. <i>FDR</i> < 20%	47/538 (9%)	185/538 (34%)	228/538 (42%)

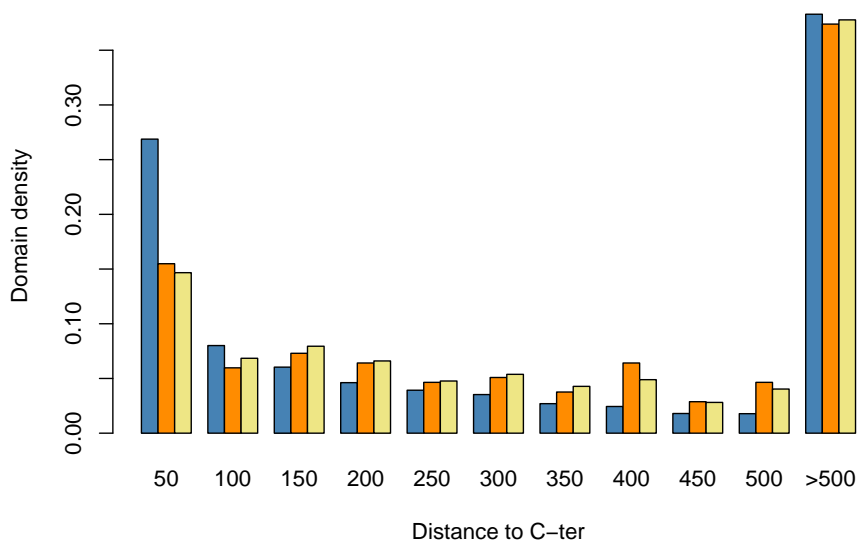
This table reports the proportion of known and new Pfam domains in *P. falciparum* proteins with a known *P. yoelii* homologue, which are also present in the known/new Pfam domains of their *P. yoelii* homologue. See Supp. Table 5 legend.



**Supp. Fig. 1. Number of certifications achieved by the 3 types of validating domains.** Number of certifications as a function of the *FDR* achieved by known Pfam domains, known InterPro domains, and potential Pfam domains.

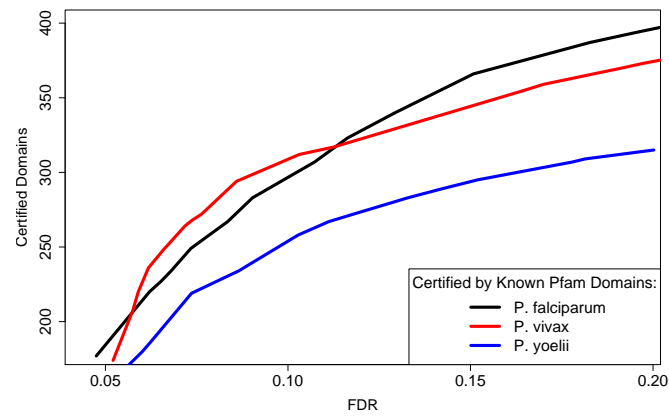


(a)

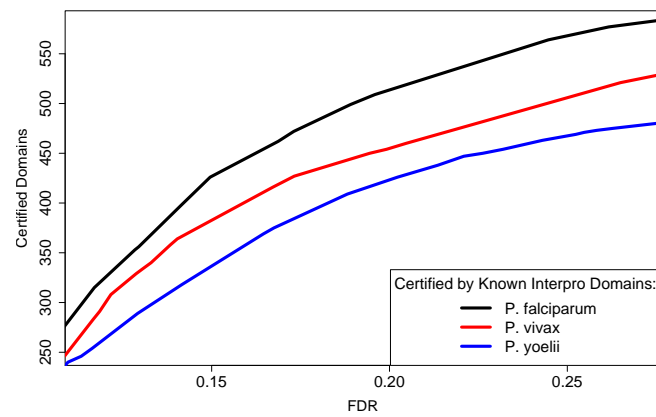


(b)

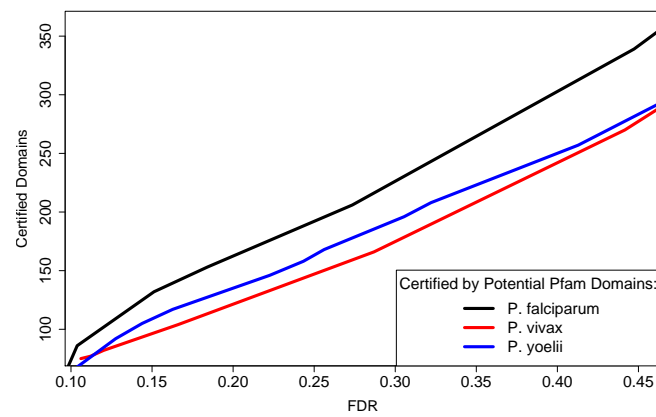
**Supp. Fig. 2. Distribution of distances between domains and protein ends.** Distances are expressed in number of residues. Blue: known domains; orange: newly certified domains with  $FDR < 10\%$ ; yellow: newly certified domains with  $FDR < 20\%$ . (a) Distance to protein N-terminus. (b) Distance to protein C-terminus. As we can see, newly certified domains are not closer to protein ends than the already known domains. As domain evolution events (especially domain loss due to loss of functionality) occurs primarily at protein ends (Weiner *et al.*, 2006), this tends to prove that the new domains are not more affected by these events than the already known domains are.



(a)



(b)



(c)

**Supp. Fig. 3. Number of certifications achieved by the 3 types of validating domains for *P. falciparum*, *P. vivax* and *P. yoelii*.** Number of certifications as a function of the *FDR* achieved by known Pfam domains (a), known InterPro domains (b), and potential Pfam domains (c).