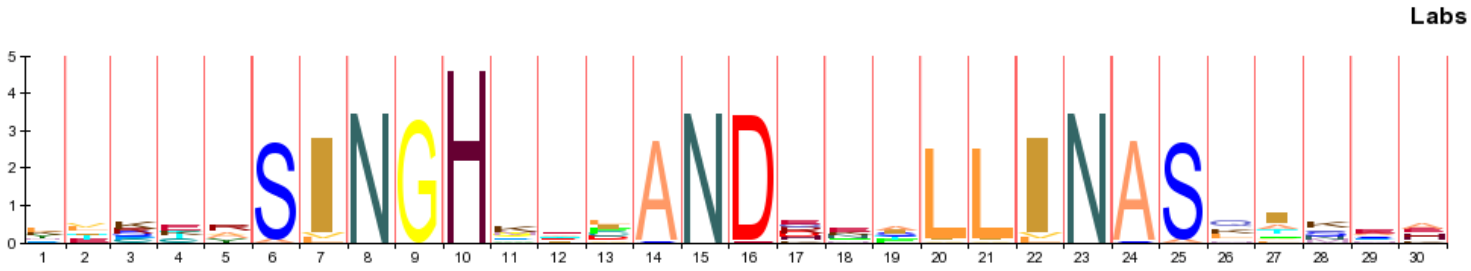
Domain Prediction Using Context (dPUC):

a framework for enhancing protein domain predictions across diverse organisms

Alejandro Ochoa

2010-03-04



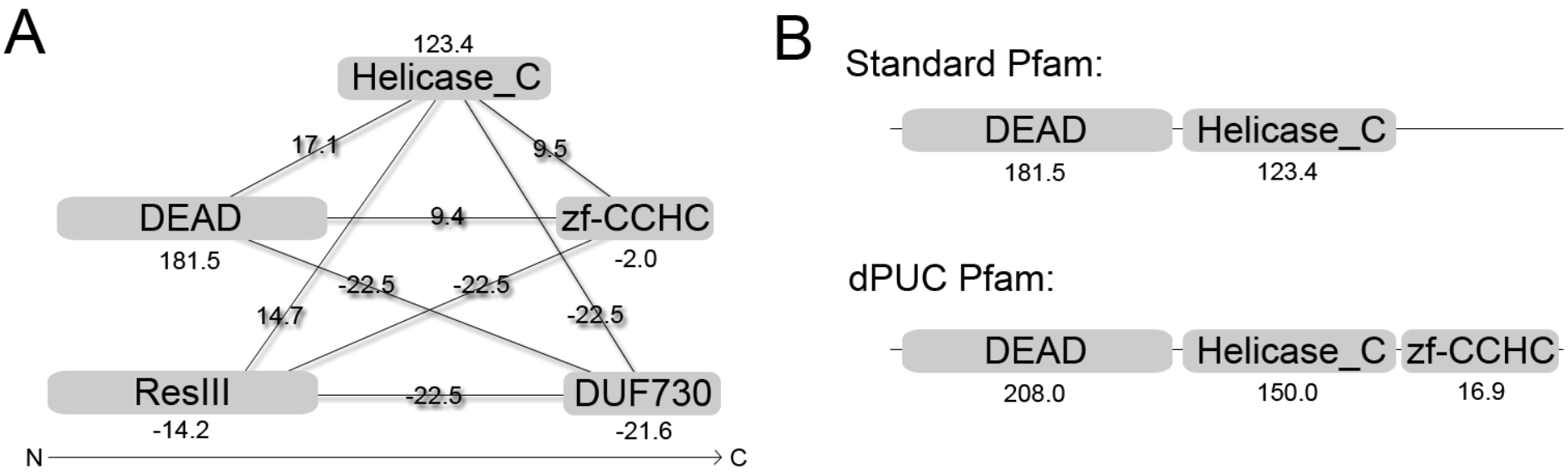
Protein Domains

- Domains
  - Building blocks of proteins
- Domain prediction
  - Pfam
    - ~12K families
  - Best source of function prediction on new genomes
- Domains co-occur in limited combinations
  - **Traditionally scored independently of each other**

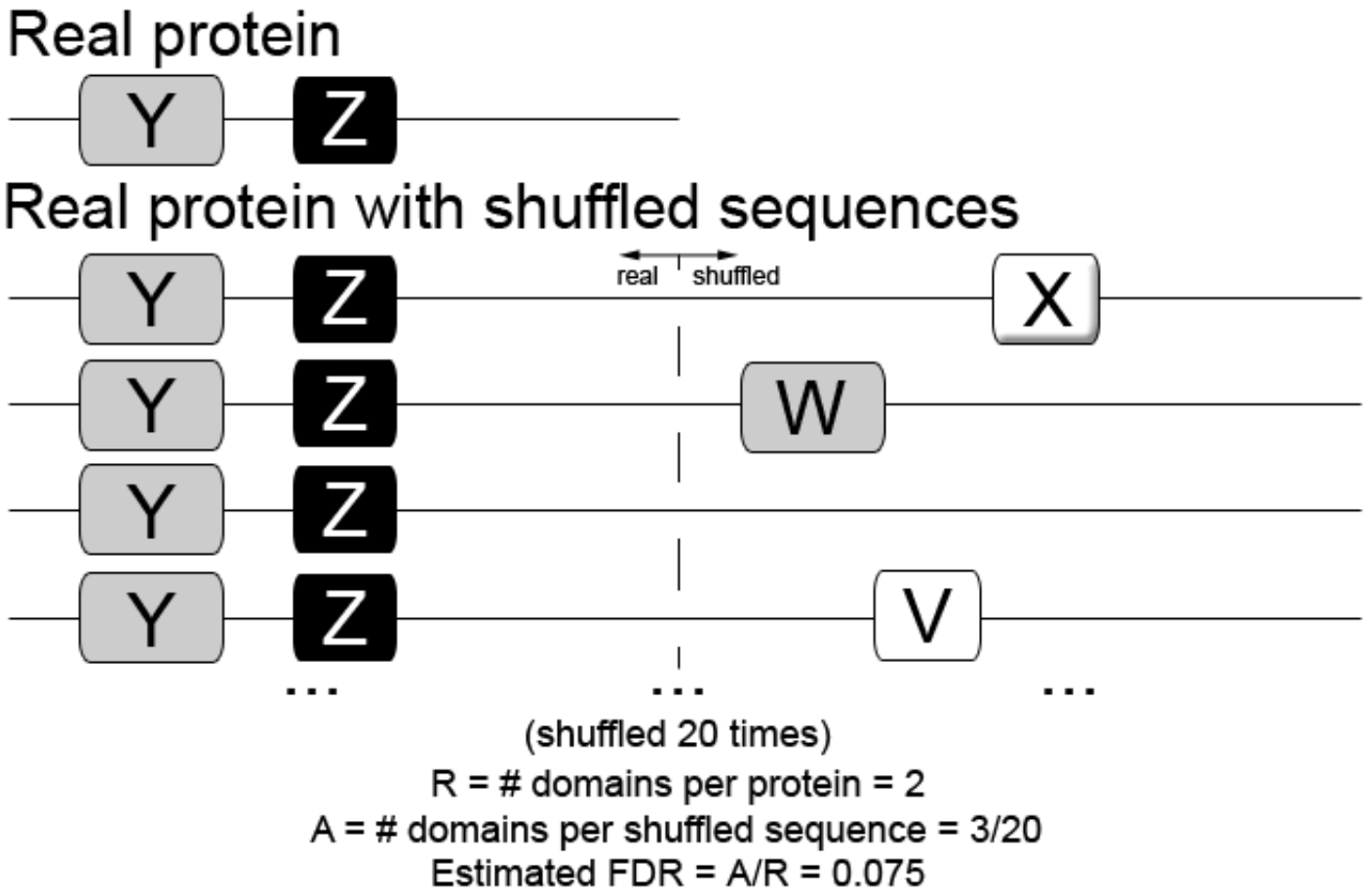
Domain context

- Background
  - Domains co-occur in limited combinations
  - Domains are scored independently of each other
- Idea
  - Score domains together
- Goals
  - High quality predictions
  - Fast and practical
    - For whole genomes and on web

Illustration of dPUC method



Measuring signal to noise



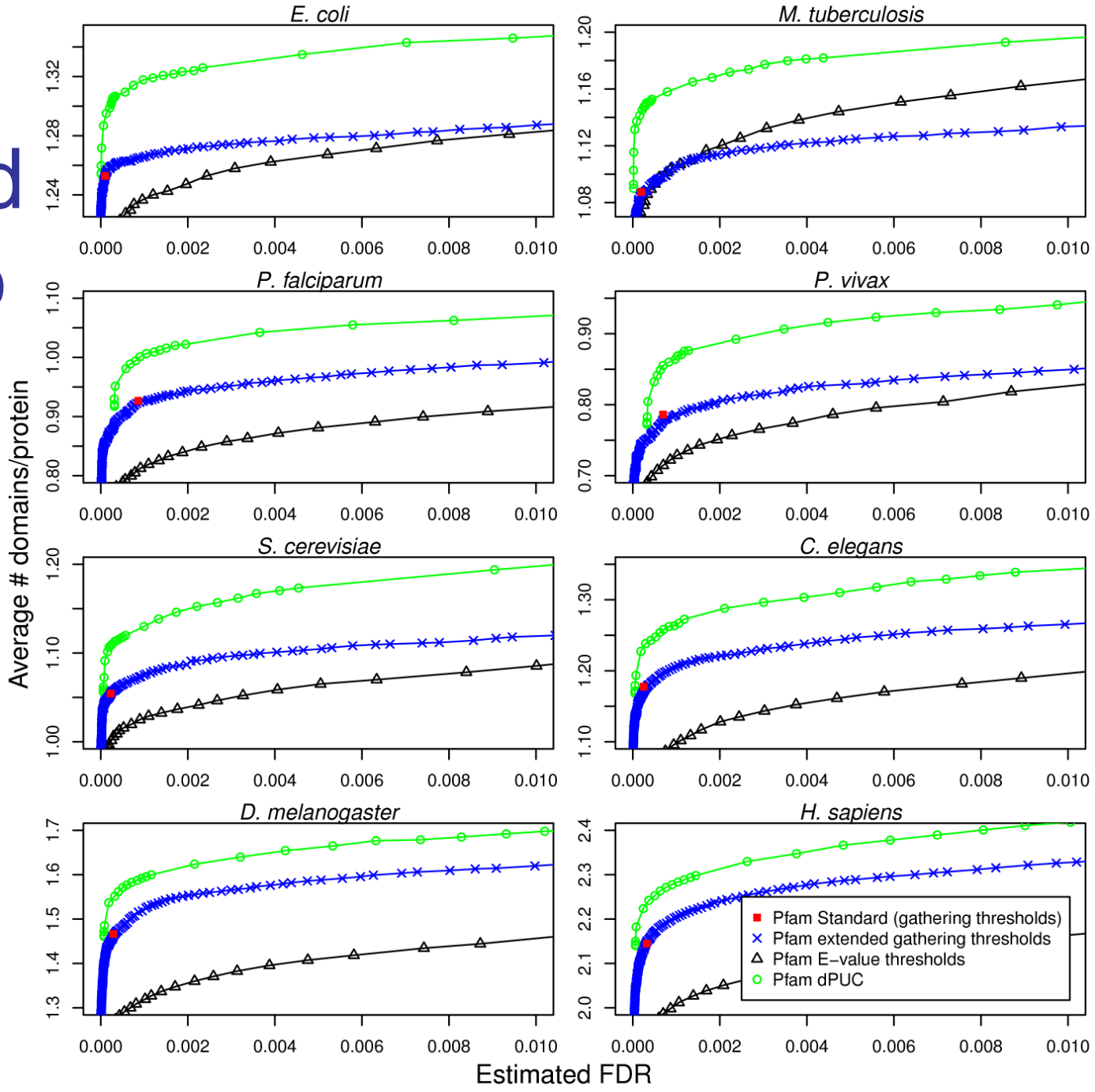
(shuffled 20 times)

R = # domains per protein = 2

A = # domains per shuffled sequence = 3/20

Estimated FDR = A/R = 0.075

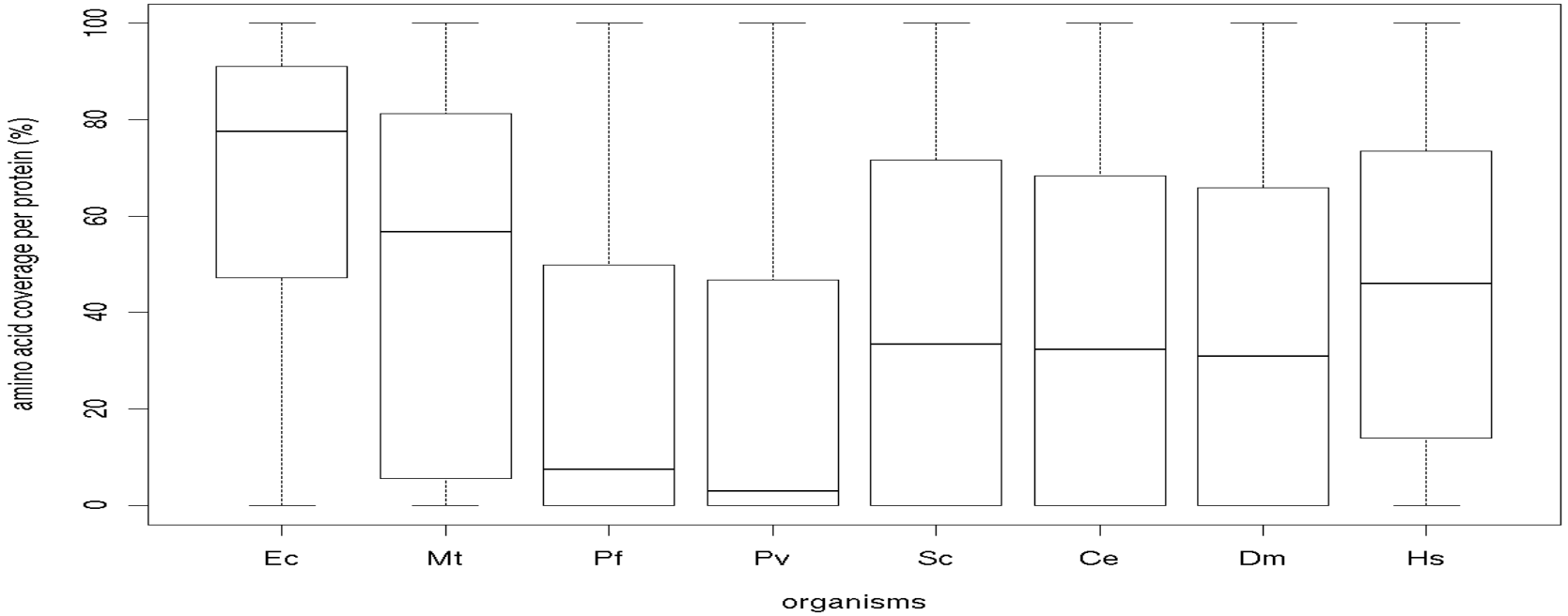
Improved signal to noise



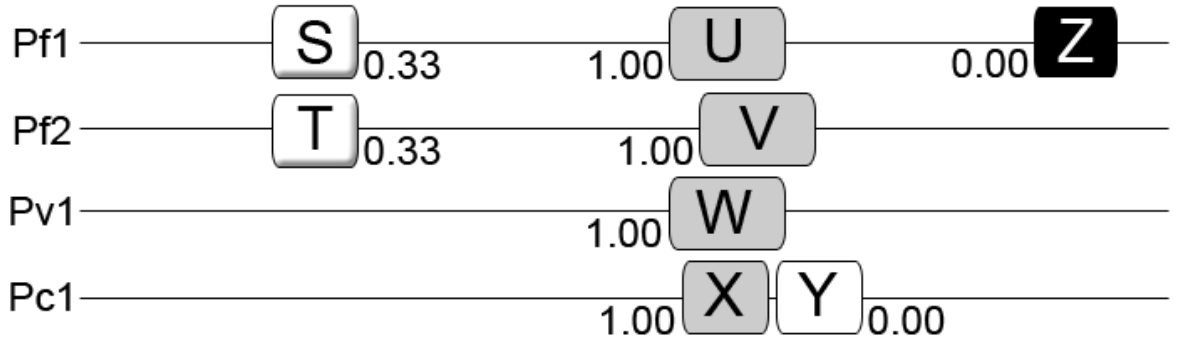
dPUC % improvement compared to Standard Pfam

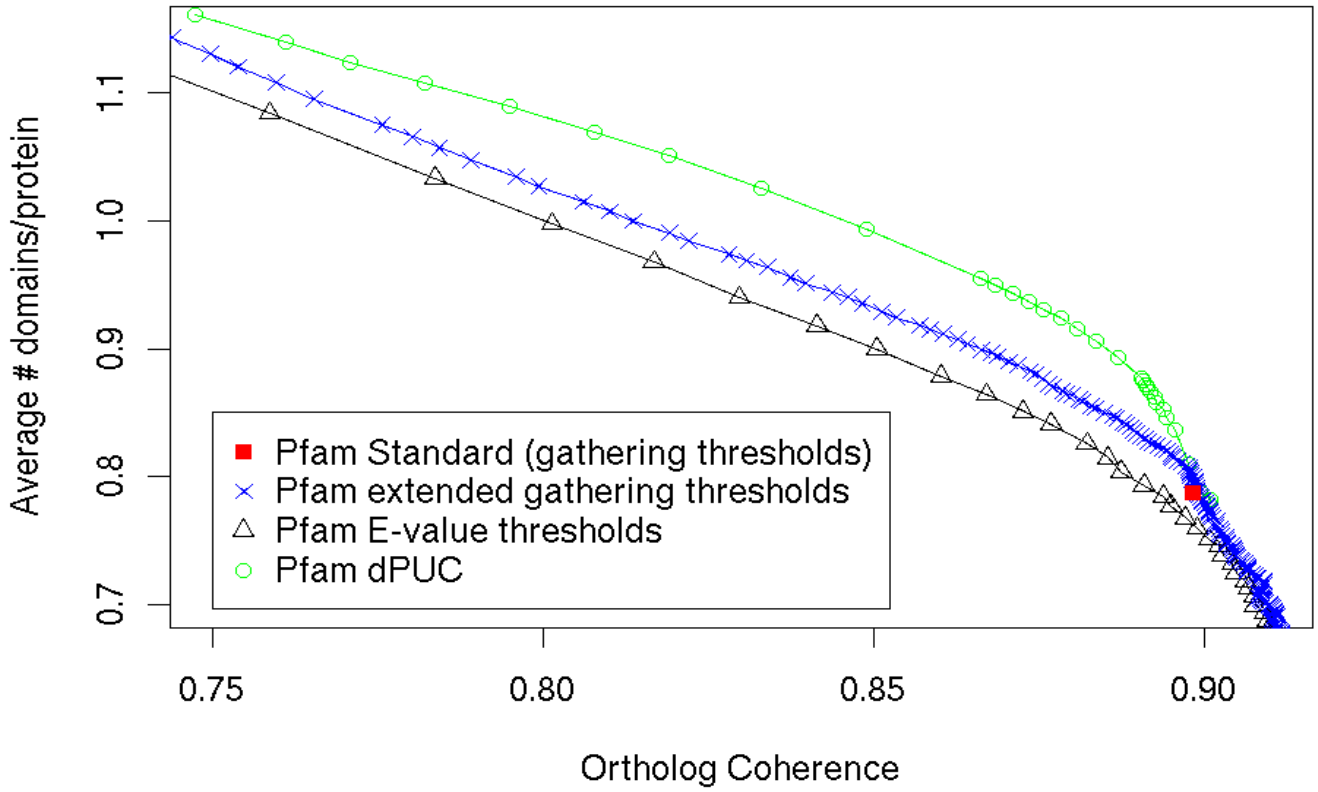
	E. c.	*M. t.*	*P. f.*	*P. v.*	*S. c.*	*C. e.*	*D. m.*	*H. s.*
Domains	4.30	6.00	10.30	11.46	6.21	8.07	9.08	7.15
Amino acids	2.38	4.13	7.25	7.66	3.14	4.74	5.63	3.63
Proteins	0.16	0.08	1.80	1.31	0.38	0.70	0.59	0.56

Amino acid coverage of Standard Pfam



Improved ortholog coherence on PlasmoDB





Predictions on *P. falciparum* novel to PlasmoDB

Protein ID	Original Domains	Additional New Context Domains	Current annotation (PlasmoDB 6.0)	Suggested reannotation (this study)
PFL0980w	CwfJ_C_1	CwfJ_C_2	Cons. *P* prot., unk. func.	Pre-mRNA splicing factor (*CwfJ* hom.)
PF13_0222	Metallophos	DBR1	Phosphatase	RNA lariat debranching enzyme (*DBR1* hom.)
PF11_0086	MIF4G	PAM2	MIF4G dom. cont. prot.	Poly(A)-binding prot.-interacting prot. 1 (*PAIP1*) hom.
PFE1390w	DEAD, Helicase_C	zf-CCHC	RNA helicase-1	Post-translational mRNA regulation (*Abstrakt* hom.)
PF08_0130	WD40	Utp13	WD-repeat prot.	**U3 ribonucleoprot. comp. (*PWP2* hom.)**
PF14_0456	WD40	Utp12	Cons. *P* prot., unk. func.	**U3 ribonucleoprot. comp. (*DIP2* hom.)**
PF10_0128	WD40	Utp13	WD-repeat prot.	**U3 ribonucleoprot. comp. (*UTP13* hom.)**
PFI1025w	RRM_1	Lsm_interact	RNA binding prot.	**U4/U6 snRNA-associated-splicing factor (*PRP24* hom.)**
PFL0985c	DUF367	RLI	Cons. prot., unk. func.	**Ribosome biogenesis regulator (*TSR3* hom.)**
MAL8P1.19	DEAD, Helicase_C	DBP10CT	RNA helicase	**Ribosomal biogenesis RNA helicase prot. (*DBP10* hom.)**
PFE0560c	MORN	Avl9	MORN repeat prot.	Atypical *AVL9* trans. prot. hom. w/ MORN doms.
PFL1455w	DUF202, SPX	VTC	Cons. *P* prot., unk. func.	Vacuolar transporter chaperone (*VTC2/3/4* hom.)
PFL2255w	TPR_2	F-box	Cons. *P* prot., unk. func.	DNA replication origin binding prot. (*DIA2* hom.)
PFF1070c	UPF0004, Radical_SAM	TRAM	Radical SAM prot.	tRNA modification enzyme (*MiaB* hom.) or CDK5 regulatory subunit-associated prot. 1
PFL1045w	DUF814	FbpA	Cons. prot., unk. func.	FbpA dom. prot.
MAL13P1.182	RanBPM_CRA	LisH	Cons. *P* prot., unk. func.	GID8 hom.
MAL13P1.79		zf-CCCH, WD40	Cons. *P* prot., unk. func.	CCCH zinc finger prot.
MAL13P1.37		zf-B_box	Cons. *P* prot., unk. func.	Tripartite motif prot.

Predictions on *P. falciparum* agreeing with other domain databases, with reannotations

Protein ID	Original Domains	Additional New Domains	Suggested reannotation (this study)
PFE1240w	Radical_SAM, Wyosine_form	Flavodoxin_1	tRNA modification enzyme (*TYW1* hom.)
PFF1490w	THF_DHG_CYH_C	THF_DHG_CYH	Tetrahydrofolate dehydrogenase/cyclohydrolase (*MTD1* hom., *MIS1/ADE3* hom. w/o FTHFS dom.)
MAL8P1.139	DDA1*	WD40	Regulator of (H+)-ATPase in Vacuolar membrane (*RAV1* hom.)
PF08_0124	CactinC_cactus	Cactin_mid	*Cactin* hom.
PF10_0152		NTP_transf_2, PAP_assoc	Non-canonical cytoplasmic specific poly(A) RNA polymerase prot. (*cid13* hom.)
MAL13P1.170	NTP_transf_2	PAP_assoc	Non-canonical poly(A) RNA polymerase prot. (*PAP2/TRF5* hom.)
PFI1560c	DUF21	CBS, cNMP_binding	Required for mitochondrial morphology (*MAM3* hom.)
PF10_0126		WD40	Phosphoinositide binding prot. (*HSV2* hom.)
PFI0510c	BRCT	IMS	DNA repair prot. (*REV1* hom.)
MAL13P1.54	WD40	LisH	Alternative splicing regulator (*Smu-1* hom.)
PF14_0052	cobW	CobW_C	Cobalamin (vitamin B12) synthesis prot.
PF08_0012	SET, Pre-SET	YDG_SRA	Histone lysine N-methyltransferase
PFE1445c		FG-GAP	T-cell immunomodulatory prot. (human TIP hom.)
PFL0975w	IQ	RCC1	Unconventional myosin fused to IQ and RCC1 domains
PF11_0276	Abhydro_lipase	Abhydrolase_1	Steryl ester hydrolase (*TGL1/YEH1/YEH2* hom.)
PF13_0190	Aha1_N	TPR_2, TPR_1	Chaperone binding prot.
PF11_0287	CRAL_TRIO	CRAL_TRIO_N	CRAL/TRIO prot.
PF11_0197	Ank	ACBP	Acyl-CoA-binding prot.
PF14_0647	TLD	TBC	*Rab* GTPase activator
PFL0575w	Amino_oxidase, Thi4*	PHD	PHD finger and flavin containing amine oxidoreductase
MAL13P1.246	E1-E2_ATPase	Cation_ATPase_C	E1-E2 ATPase
PF11_0116		Nol1_Nop2_Fmu	Nol1/Nop2/Fmu-like prot.
MAL7P1.127		Pkinase	*Rab* GTPase activator and prot. kinase
PFC0425w		zf-C3HC4, PHD	PHD finger prot.
PFI0975c		RCC1	Regulator of chromosome condensation
PFD0900w		RCC1	Regulator of chromosome condensation
MAL7P1.132		Pkinase	Prot. kinase
PFF0810c		Ras	*Ras* GTPase
PFL1990c		zf-CCHC, RRM_1	RNA binding prot.
PF07_0066		RRM_1	RNA binding prot.
PF13_0147		RRM_1	RNA binding prot.
PFF1120c		EGF	EGF-like membrane prot.
PF14_0262	WD40	TPR_1	WD40 and TPR repeats prot.
PFI0275w		WD40	WD40 repeat and EF hand prot.
PF10_0285		WD40	WD40 repeat prot.
PF11_0195		WD40	WD40 repeat prot.
PF14_0640		WD40	WD40 repeat prot.
MAL13P1.308		Arm	ARM repeat prot.

- Mona Singh & Lab
  - Jesse Farnham
  - Peng Jiang
  - Zia Khan
  - Anton Persikov
  - Jimin Song
  - Tao Yue
- Thesis Committee
  - Leonid Krugliak
  - Saeed Tavazoie

- Manuel Llinás & Lab
  - Tracey Campbell
  - Erandi De Silva
  - Björn Kafsack
  - Elyse Kozlowski
  - Yael Marshall
  - Jessica O'Hara
  - Kellen Olszewski
  - Heather Painter
  - Louis Sarry
  - Irene Ying
- NSF GRFP