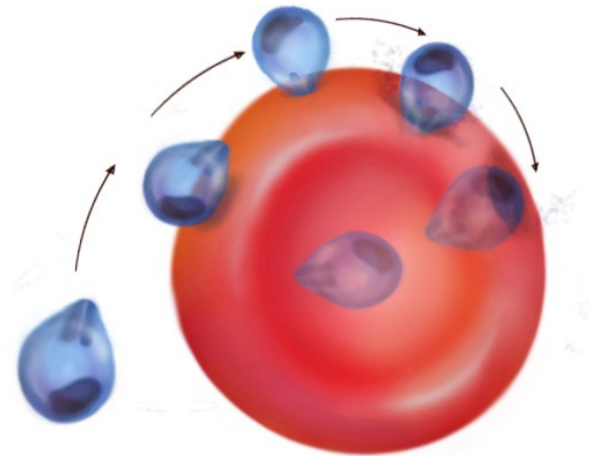
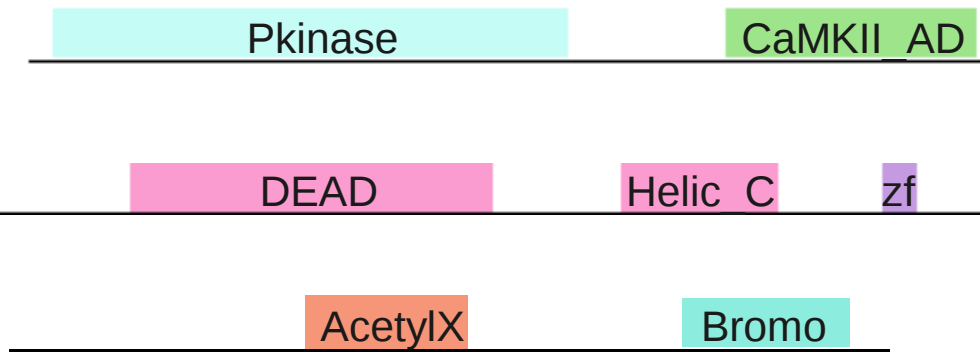


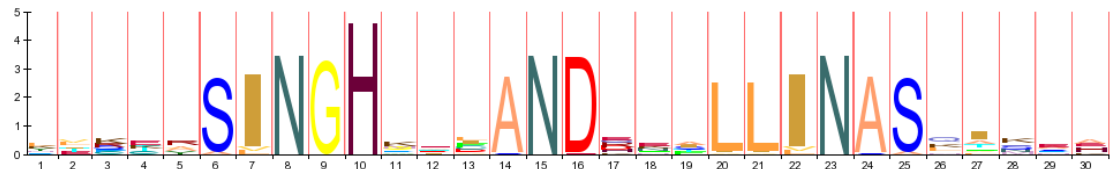
Forget the *E*-value: family-based *q*-values for protein domain prediction, and empirical error detection



Alejandro Ochoa

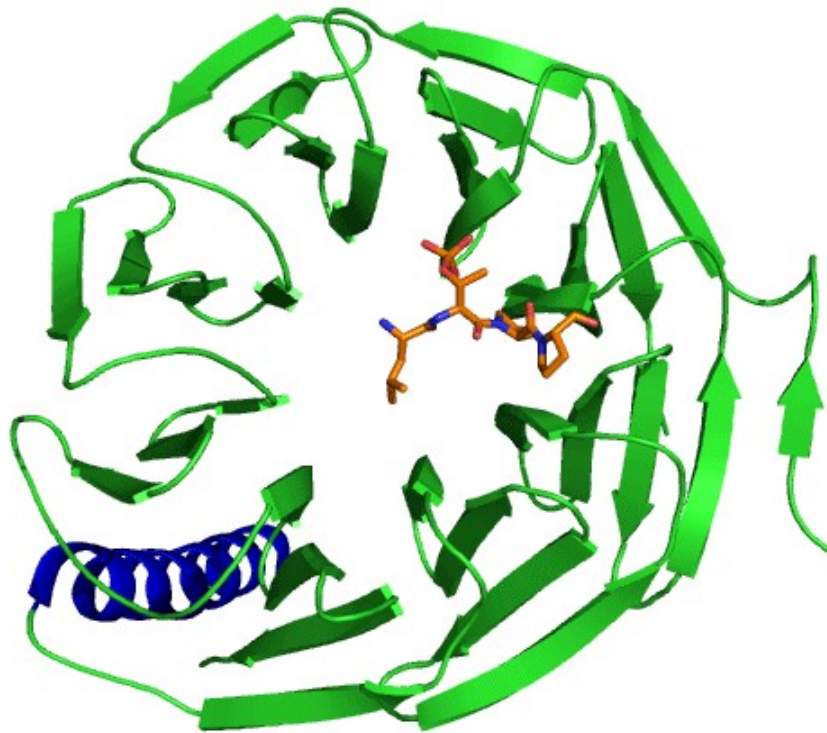
Molecular Biology, Princeton University

Yun Song's group, UC Berkeley, 2013-04-17



Labs

Protein domains

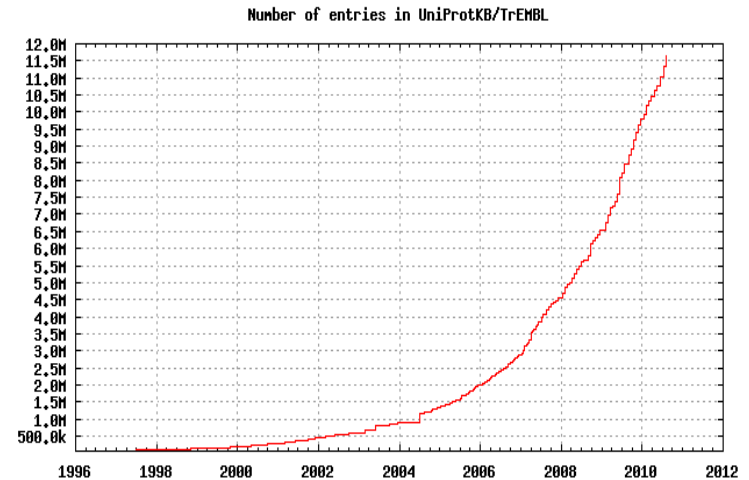
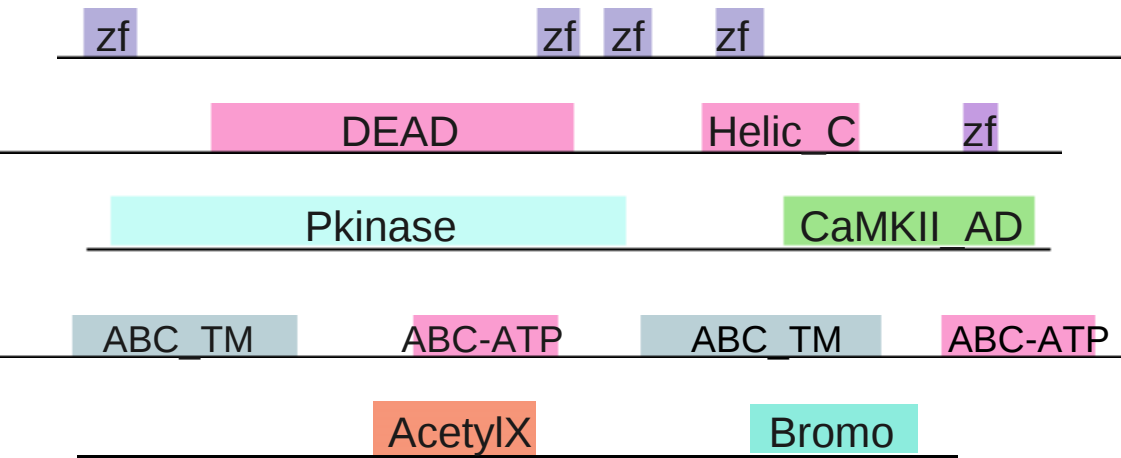


Structure
Evolution
Function

Sequence-based domain prediction:



Why predict domains?



For new sequences, before experiments start...

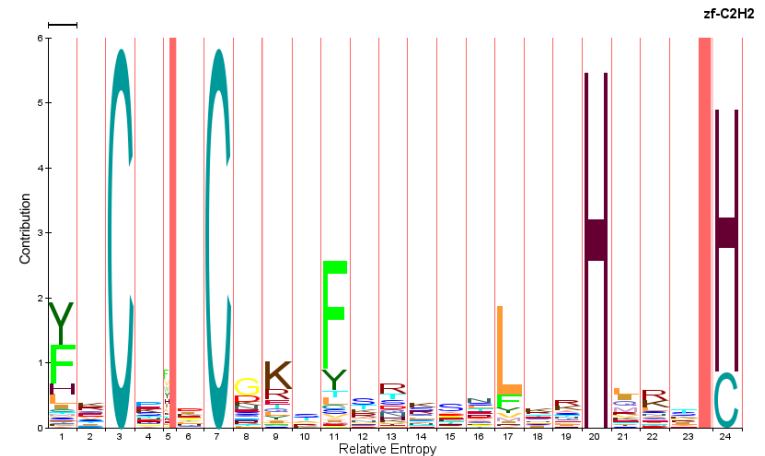
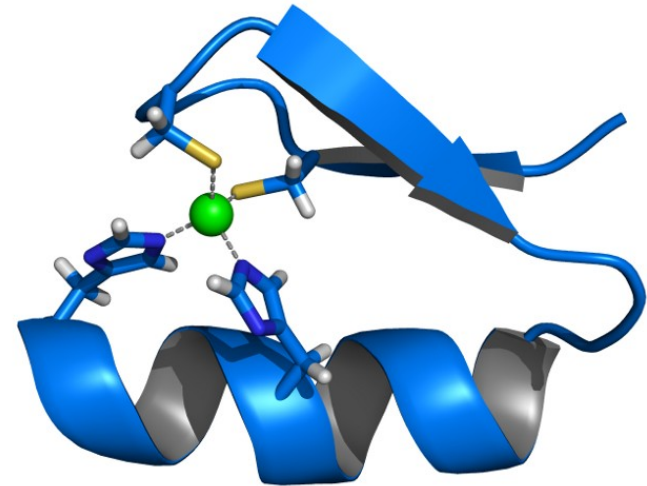
Domains may imply functions

Experimental alternatives are unfeasible as protein databases grow exponentially

Representing Domains

SNAI_DROME/362-385
 SNAI_XENLA/232-255
 SNAI_MOUSE/236-259
 ESCA_DROME/426-449
 SUHW_DROAN/221-243
 TERM_DROME/323-346
 Z020_XENLA/174-196
 EVI1_HUMAN/217-239
 Z02_XENLA/34-59
 EVI1_HUMAN/21-44
 ZNF10_HUMAN/517-539
 ZNF91_HUMAN/238-260
 ZFP58_MOUSE/120-142
 TRAI_CAEEL/306-331
 ZNF76_HUMAN/345-368
 ZN12_MICSA/106-129
 LOLAI_DROME/794-817
 ZNF17_HUMAN/435-457
 ZG32_XENLA/34-56
 TF3A_BUFAM/104-128
 ZG46_XENLA/146-168
 MZFI_HUMAN/412-434
 ZN239_MOUSE/6-28
 ZSC22_HUMAN/352-374
 EGR1_HUMAN/396-418
 SUHW_DROAN/349-373
 CF2_DROME/485-508
 CF2_DROME/401-423
 KRUP_DROME/306-328
 TTY1_HUMAN/383-407
 ZG52_XENLA/61-83
 TTKE_DROME/538-561
 ZNF76_HUMAN/285-309
 SDC1_CAEEL/145-168
 SRYC_DROME/358-380
 SDC1_CAEEL/270-292
 TRAI_CAEEL/276-300
 ESCA_DROME/370-392

YACQ...VCH...KSF SRM...SLLNKHSS...NC
 YQCK...SCS...RTFSRM...SLLHKHEET...GC
 YQCQ...ACA...RTFSRM...SLLHKH QES...GC
 YSCT...SCS...KTFSRM...SLLTKHSEG...GC
 HVCG...KCY...KTFRRL...MSLKKHLEF...C
 LHCR...RCR...TQFSRR...SKLHIHQKL...RC
 FMCA...DCG...RCFSVS...SSLKYHQRI...C
 IKCK...DCG...QMFSST...SSLNKHRRF...C
 YSCA...DCG...KHFSEK...MYLQFHQKNPSEC
 YRCE...DCD...QLFESK...AELADHQKF...PC
 YKCN...QCG...IIFSQM...SPFIVHQIA...H
 YKCE...ECG...KAFKQL...STLTT HKII...C
 IKCE...ECG...KAFSTR...STYYRHQKN...H
 YKCEF.ADCE...KAFSNA...SDRAKHQNR...TH
 YTCS...TCG...KTYRQT...STLAMHKRS...AH
 YRCS...QCG...KAFRRT...SDLSSHRPT...QC
 YECR...HCG...KKYRWK...STLRRHENV...EC
 YECN...KCG...KFFRYC...FTLNRHORV...H
 FVCV...HCG...KGF RDM...YKLSLHLRI...H
 YVCYF.ADCG...QQFRKH...NQLKIHQYI...H
 YVCT...ECG...TSFRVR...PQLRIHLRT...H
 FVCG...DCG...QGFRVS...ARLEEHRV...H
 YKCD...KCG...KGFTRS...SSLVHHSV...H
 YKCG...ECG...KTFSRS...THLTQHQRV...H
 FACD...ICG...RKFARS...DERKRHTKI...H
 YACK...ICG...KDFTRS...YHLKRHQKYS...SC
 YTCP...YCD...KRFTQR...SALTVHTTK...LH
 YTCS...YCG...KSFTQS...NTLKQHTRI...H
 YTCE...ICD...GKFSDS...NQLKSHMLV...H
 YVCPF.DGCN...KKFAQS...TNLKSHILT...H
 YTCT...QCN...KQFSHS...AQLRAHIST...H
 YPCP...FCF...KEFTRK...DMMTAHVKI...IH
 YTCPE.PHCG...RGFTSA...TMYKNHVRI...H
 YMCQ...VCL...TLFSGT...YNLFMHWRT...SC
 YOCD...ICG...QKFVOK...INLTHHARI...H
 YFCH...ICG...TVFIEQ...DMLFKHWRL...H
 NKCEY.PCGG...KEYSRL...ENLKTHRRT...H
 CKCN...LCG...KAFSRP...WLLQGHIRT...H



zf-C2H2

Plasmodium falciparum

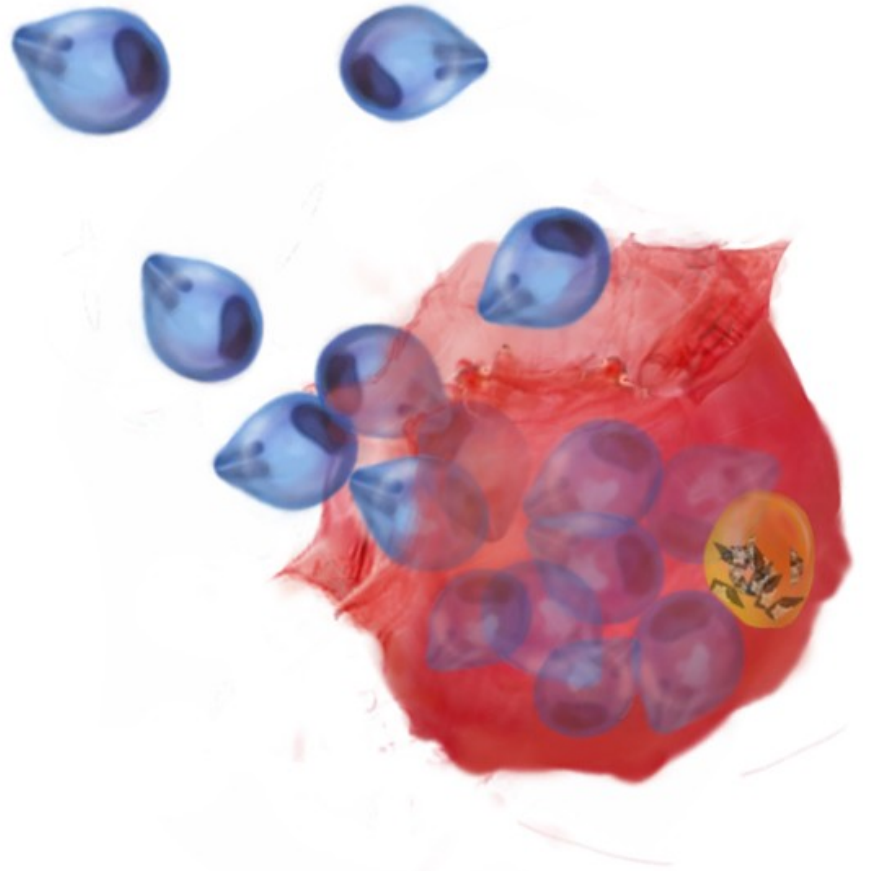
Malaria

Information challenges

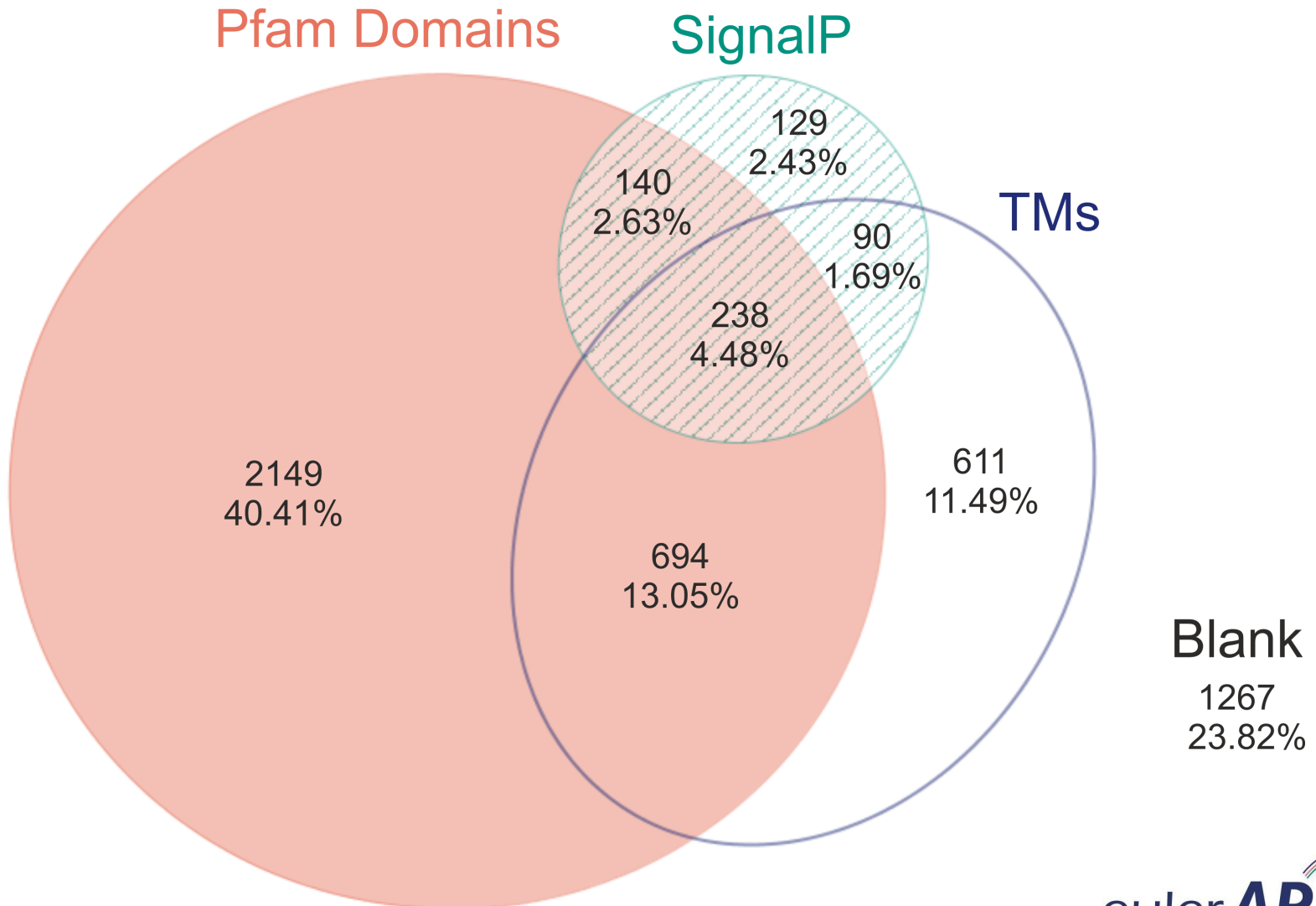
- Diverged eukaryote
- 80% AT-bias
- Low-complexity regions

Annotation

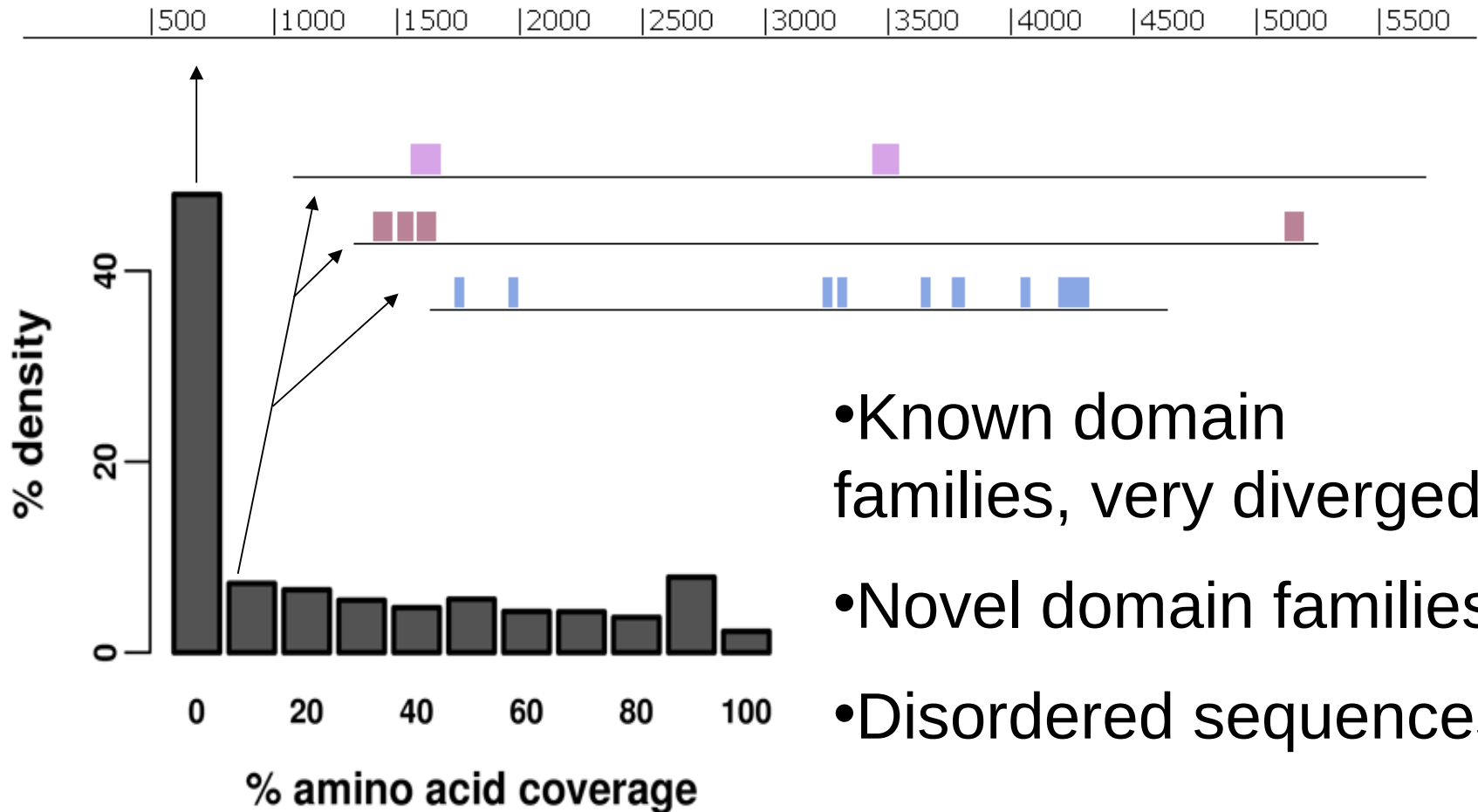
- 5.5K proteins
- 45% unknown function
 - 20% unknown in yeast
- 88% of annotations are bioinformatical



Sequence-based annotation of *Plasmodium falciparum* proteome



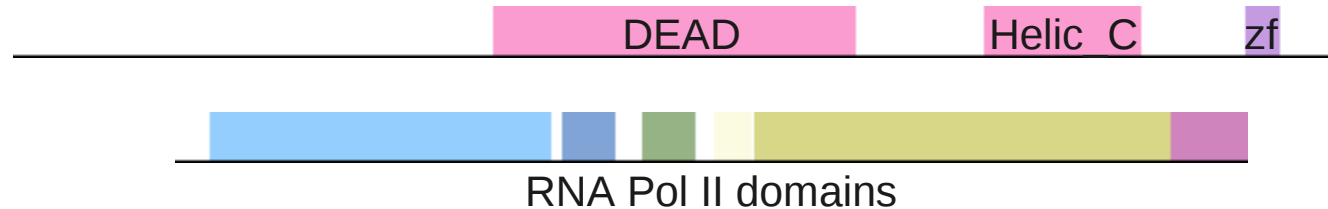
Poor domain coverage of *Plasmodium falciparum*



Outline of results

- Domain prediction using context
 - Application to malaria parasite
- Optimal *FDR* control for domains
 - Family-based *q*-values, adapted to domain problem
 - *Local FDR* optimizes problem, but *q*-values are more robust
 - Large problems with coiled coils, transmembrane domains
 - Implications for sequence models

Domain Prediction Using Context: dPUC



Background

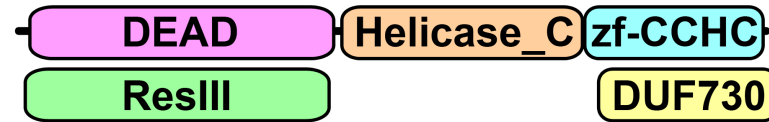
- Domains co-occur in limited combinations
- Domains are scored independently of each other

Idea

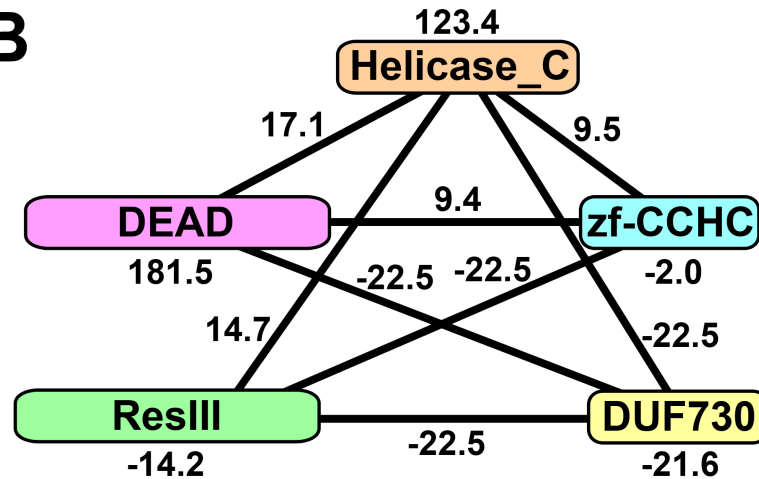
- Score domains in combination
- Context + Sequence evidence

The dPUC method

A



B



C

Standard Pfam:



dPUC Pfam:



<http://compbio.cs.princeton.edu/dpuc/>

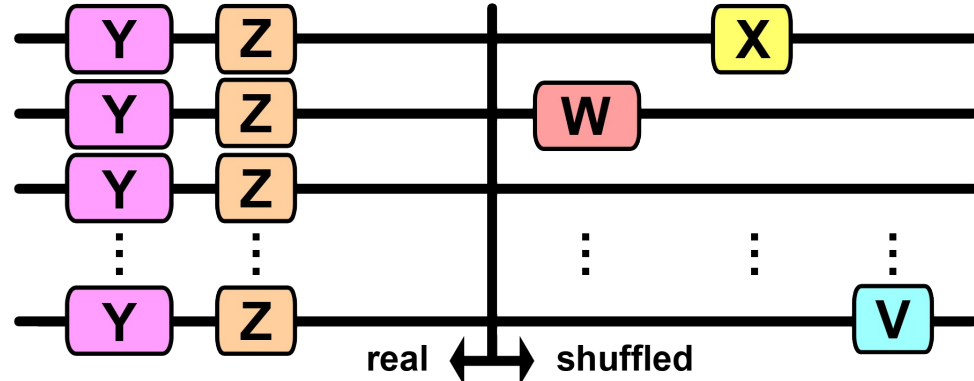
Ochoa, *et al.* BMC Bioinformatics 12, 90 (2011).

Improved signal to noise

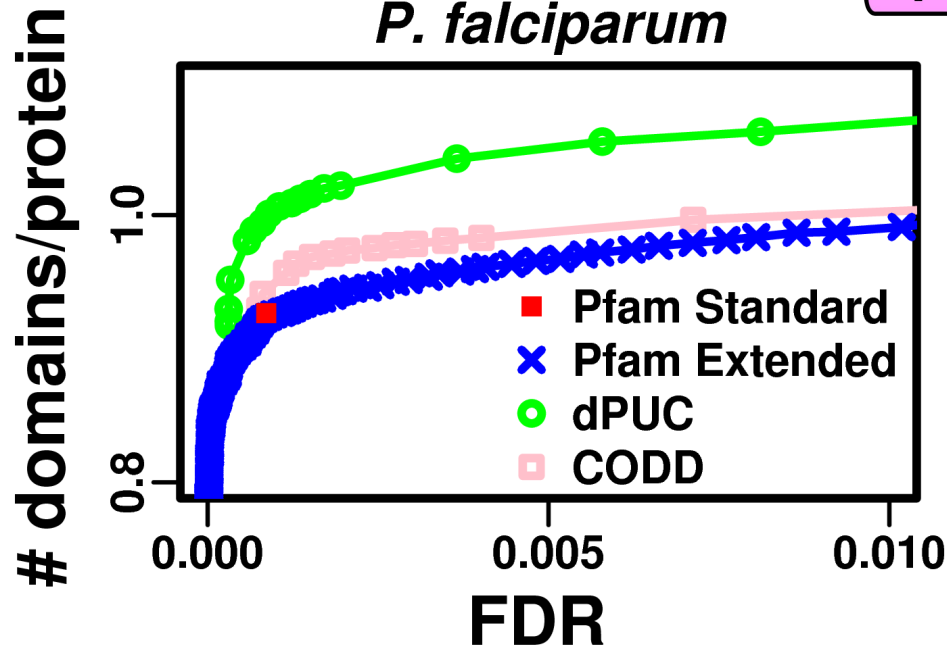
Real protein



Real protein with shuffled sequences



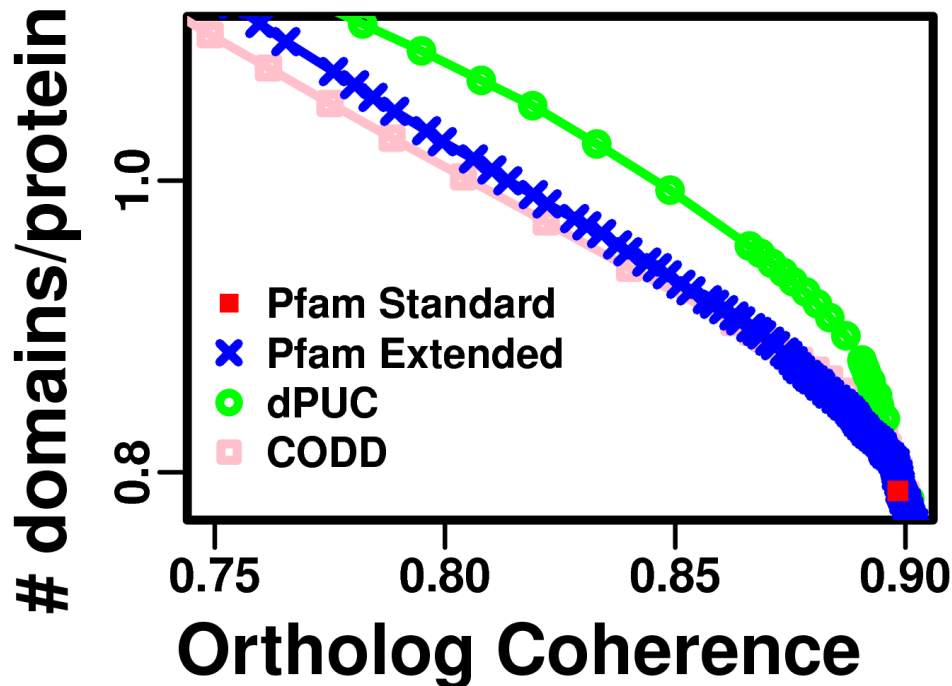
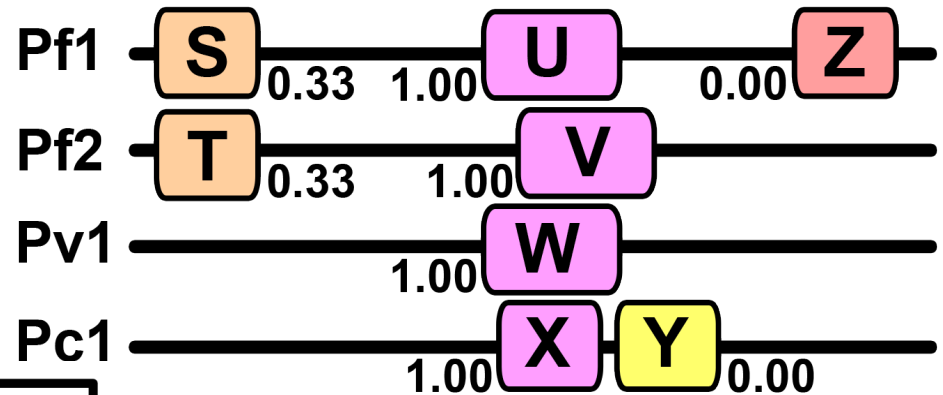
P. falciparum



<http://compbio.cs.princeton.edu/dpuc/>

Ochoa, *et al.* BMC Bioinformatics 12, 90 (2011).

Improved ortholog coherence on *Plasmodium* species



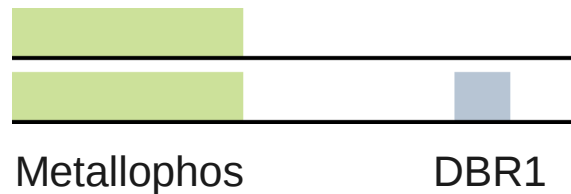
New predictions

Phosphatase -> RNA lariat debranching enzyme

P. falciparum

Std Pfam

dPUC



S. cerevisiae

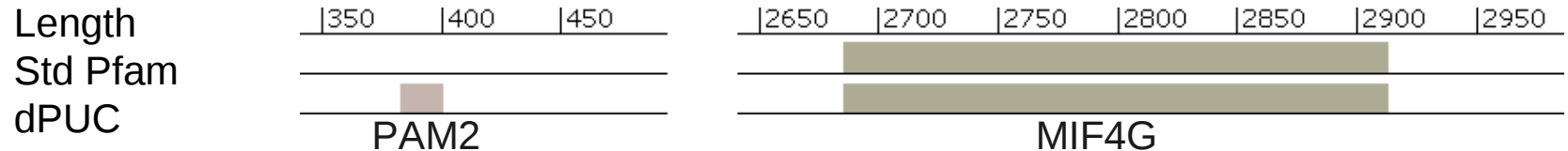
Std Pfam & dPUC



New predictions

MIF4G domain-containing protein ->
Poly-A binding protein-interacting protein 1

P. falciparum



H. sapiens

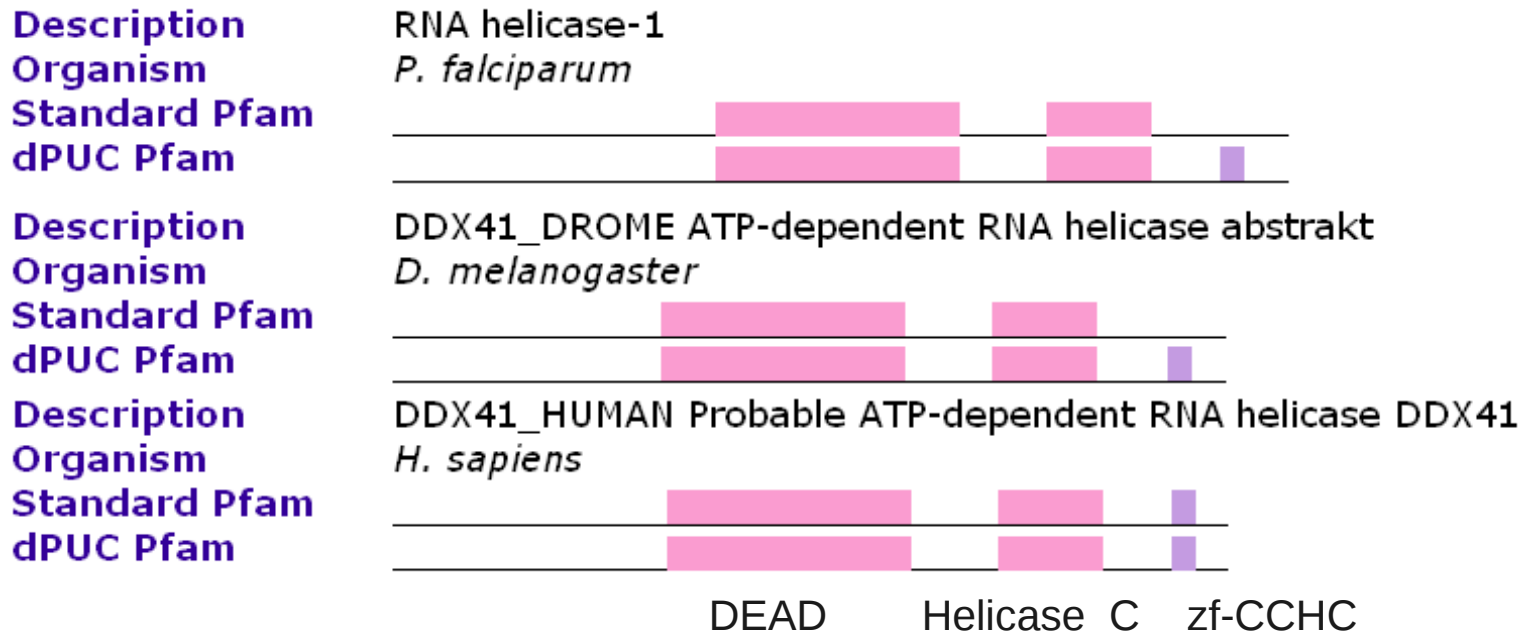


Interested malaria curator:
Hagai Ginsburg,
Hebrew U of Jerusalem

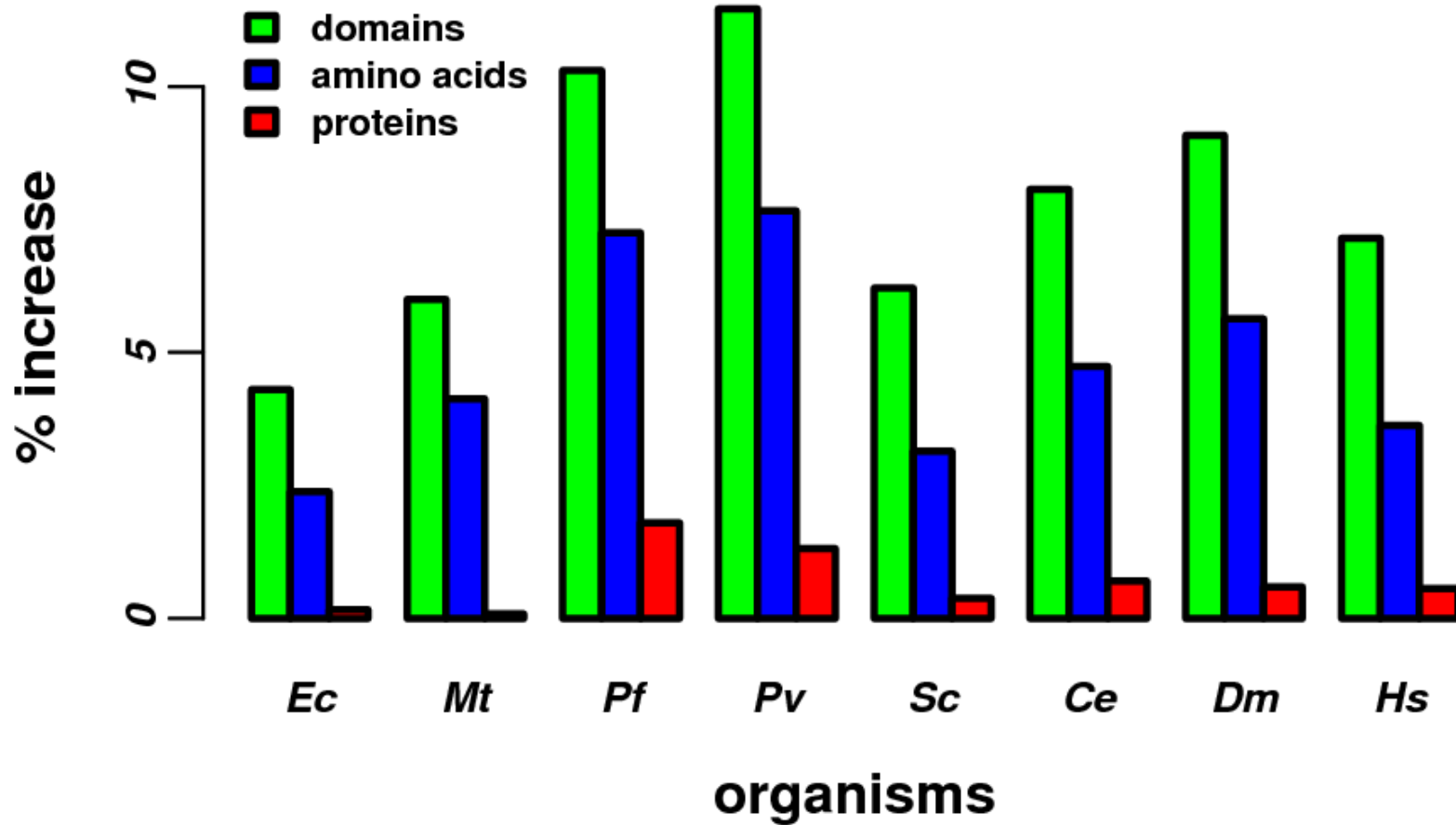
<http://compbio.cs.princeton.edu/dpuc/>
Ochoa, *et al.* BMC Bioinformatics 12, 90 (2011).

New predictions

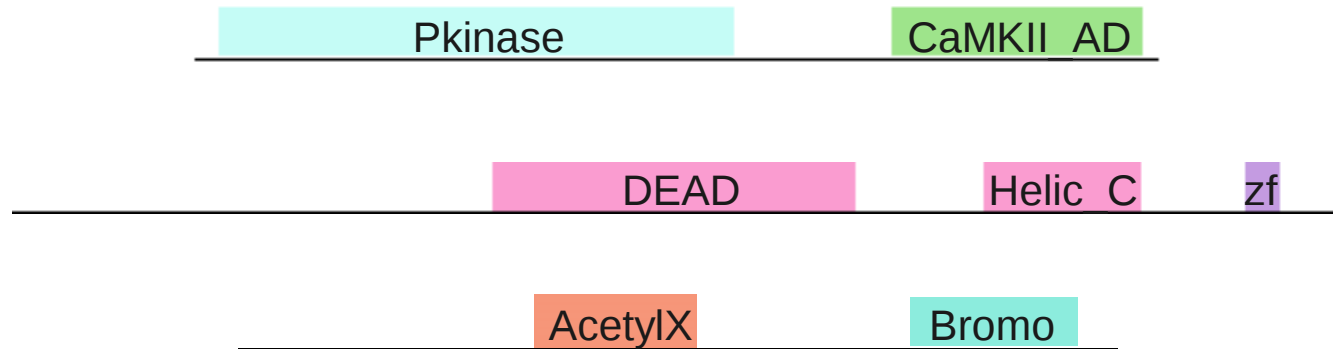
RNA helicase -> mRNA sequestration



dPUC increases coverage

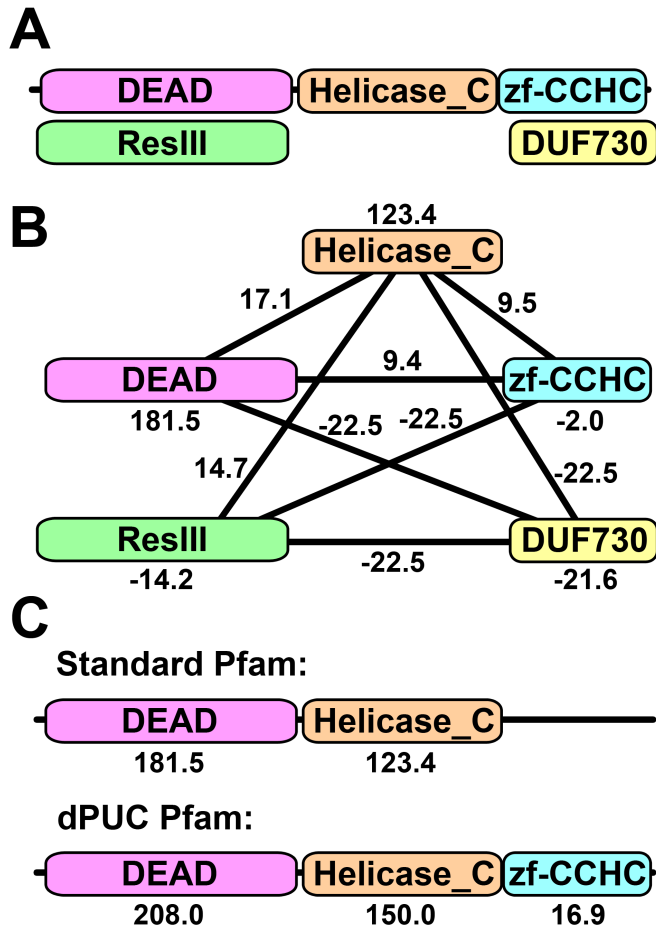


Domain context



Complements sequence evidence
Improves domain predictions
Works best on diverged organisms

The dPUC method: problems



Domain scores are normalized by curated thresholds

Why do *E*-values perform worse?

Outline of results

- Domain prediction using context
 - Application to malaria parasite
- **Optimal *FDR* control for domains**
 - Family-based q -values, adapted to domain problem
 - *Local FDR* optimizes problem, but q -values are more robust
 - Large problems with coiled coils, transmembrane domains
 - Implications for sequence models

The *E*-value: the statistic of sequence analysis

First genomes
 1976: virus
 1995: bacteria
 1996: yeast
 1998: worm
 2000: human

Query Type

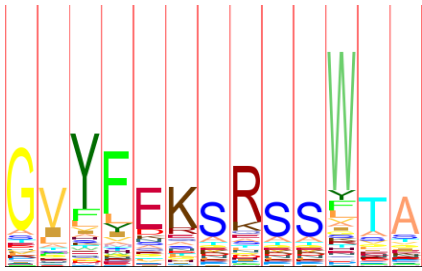
Database Type

Software

GVYFEKSRSSWTA

GVRQRKNSNKWVS

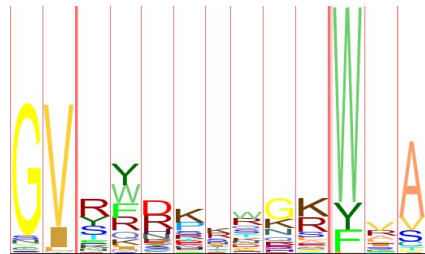
**FASTA (Z 1985)
 BLAST (*E* 1990)**



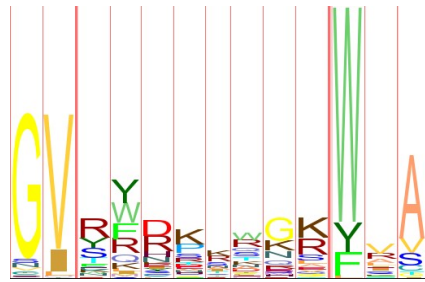
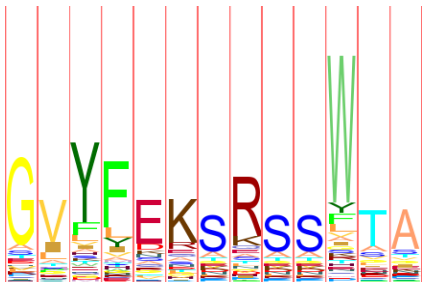
GVRQRKNSNKWVS

PSI-BLAST (*E* 1997)

GVYFEKSRSSWTA



**SAM (Z 1994, ~*E* 2005)
 HMMer (bit 1995, ~*E* 1998, *E* 2008)
 Pfam DB (curated 1997)
 IMPALA, RPS-BLAST (*E* 1999)**

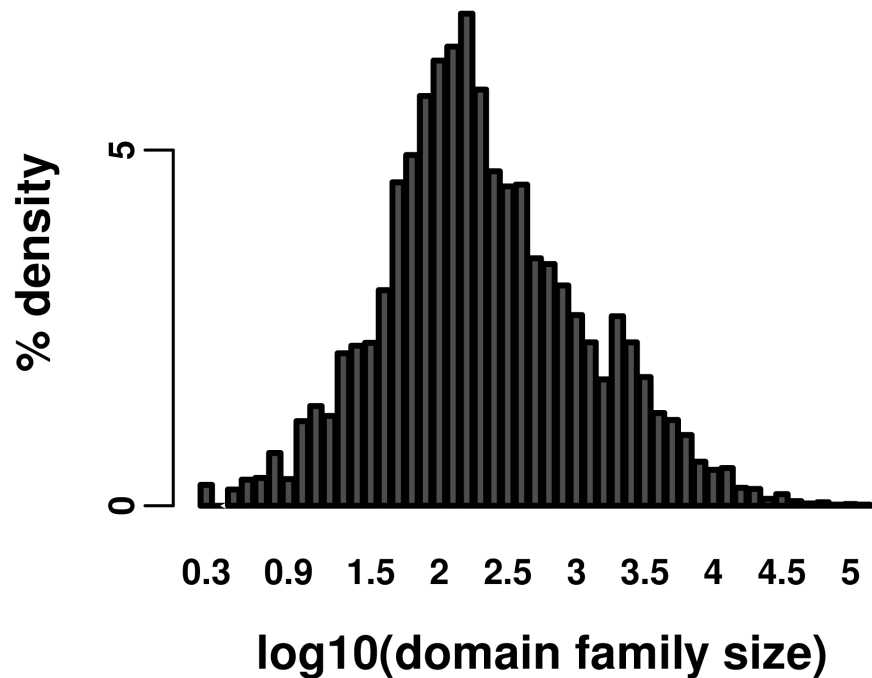


**PRC (bit 2002, ~*E* 2004)
 HHsearch (~*E*, *P* 2004)
 CORAL (*E* 2009)**

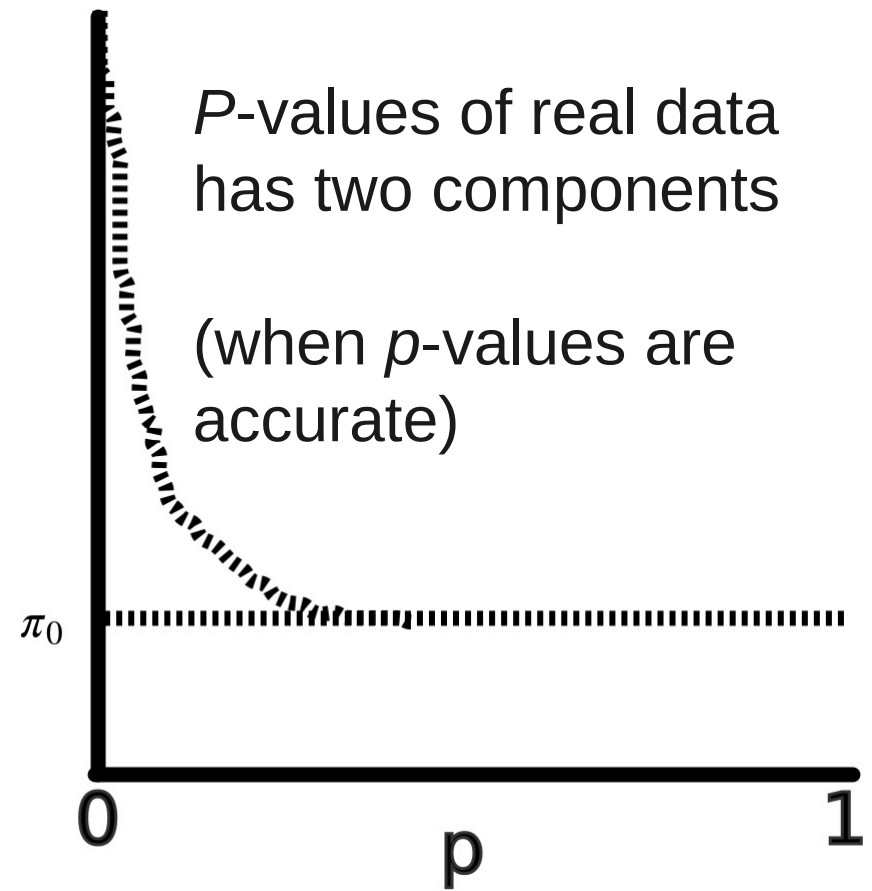
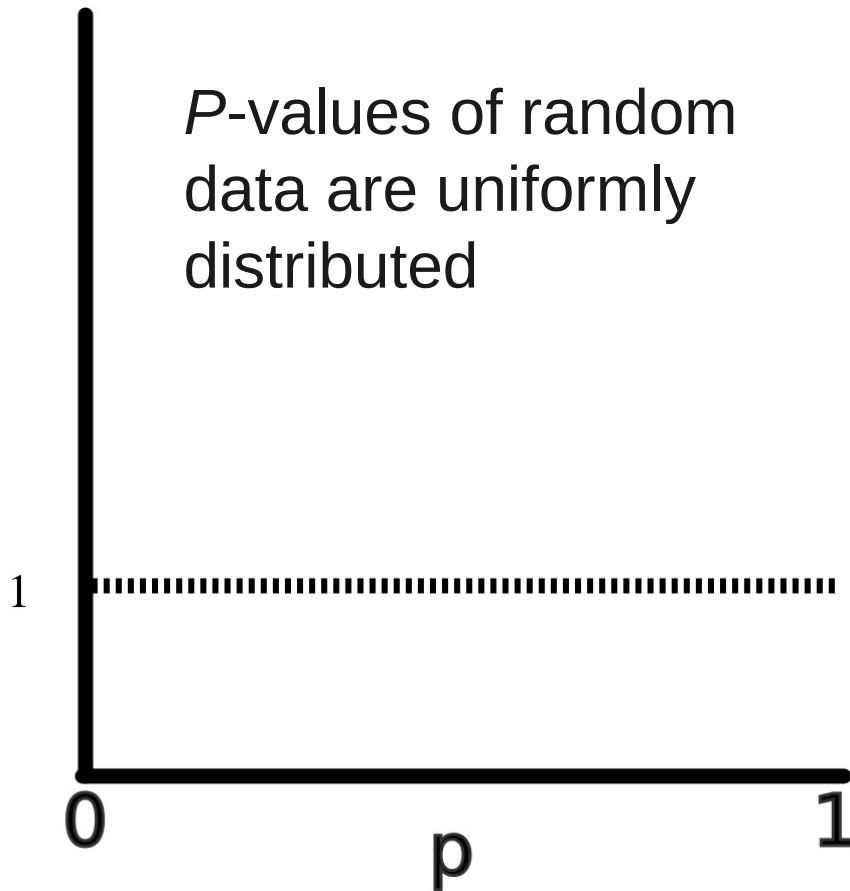
***FDR*: BH 1995, *q* 2002**

$$FDR = E/n$$

FDR = average posterior error probability

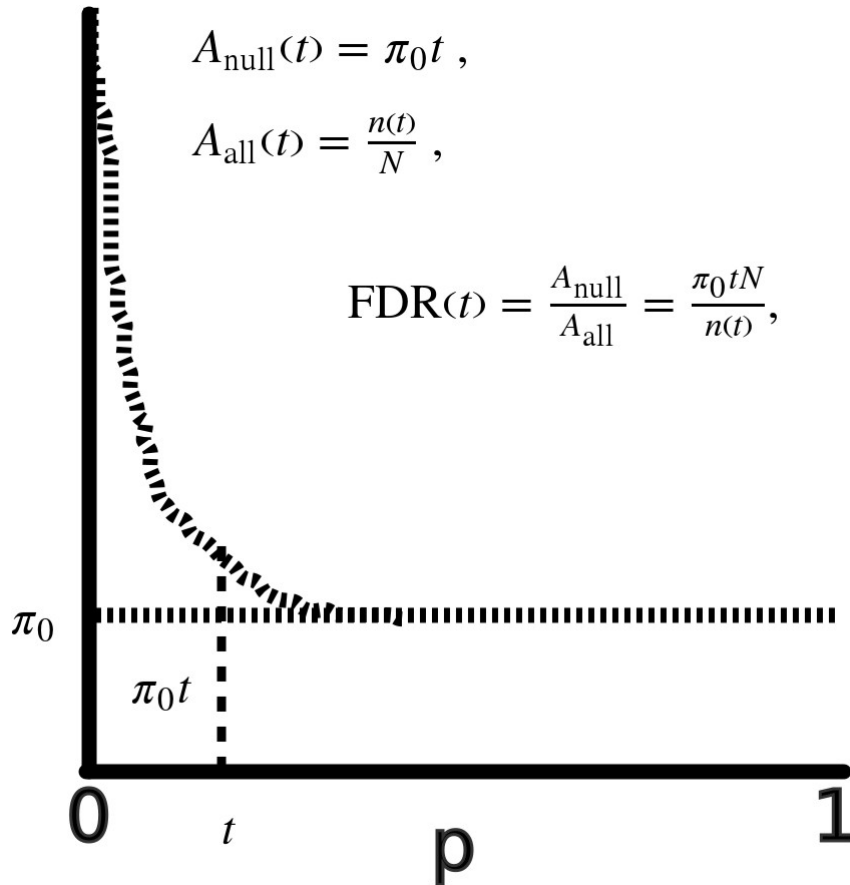


Computing q -values



Step 1: estimate π_0 (proportion of data that is false)

Computing q -values



Step 2: Directly estimate $FDR(t)$ for all thresholds t

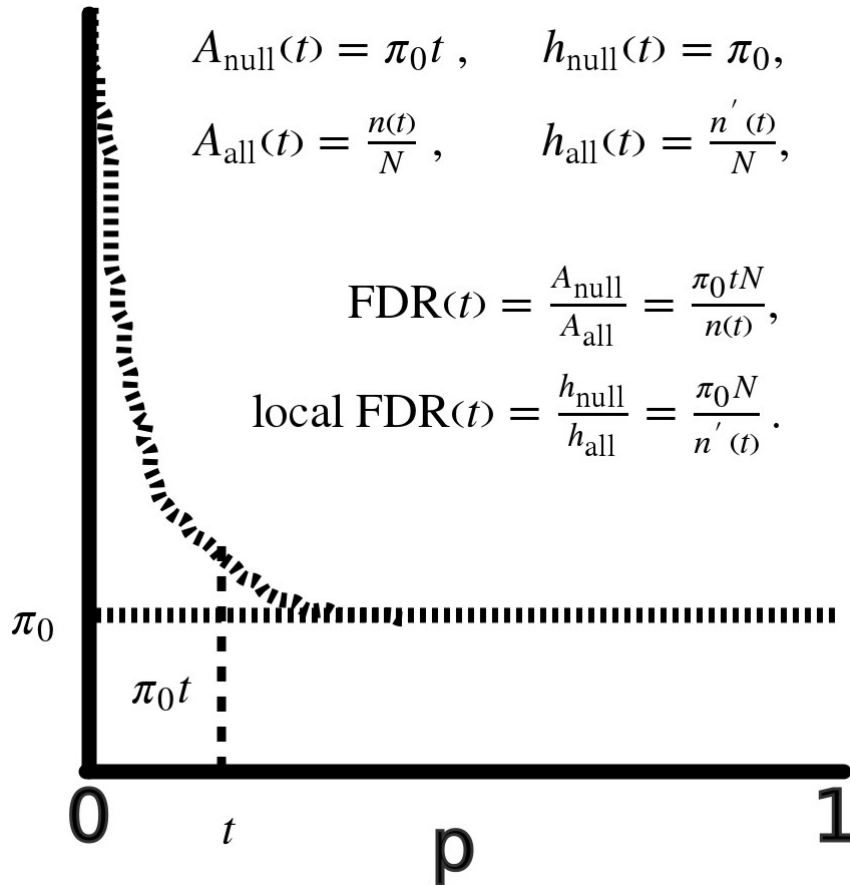
– $N = \#$ tests

– $n(t) = \#$ sig tests

Step 3: Ensure monotonicity

$$q(p) = \min_{t; p \leq t} FDR(t)$$

FDR and local FDR



Local FDR =
Posterior Error
Probability (*PEP*)

FDR = average *PEP*
of significant
predictions

Local FDRs optimize domain prediction

Find domain family thresholds t_i (for each family i) to maximize predictions

$$M = \sum n_i(t_i),$$

while constraining the combined *FDR* of all families to Q

$$Q \geq \frac{\sum \pi_{0,i} \cdot t_i \cdot N_i}{\sum n_i(t_i)} = \frac{\sum FDR_i(t_i) \cdot n_i(t_i)}{\sum n_i(t_i)}.$$

Equal family *local FDRs* solve this optimization!
Also holds when constraining combined *E-value*!

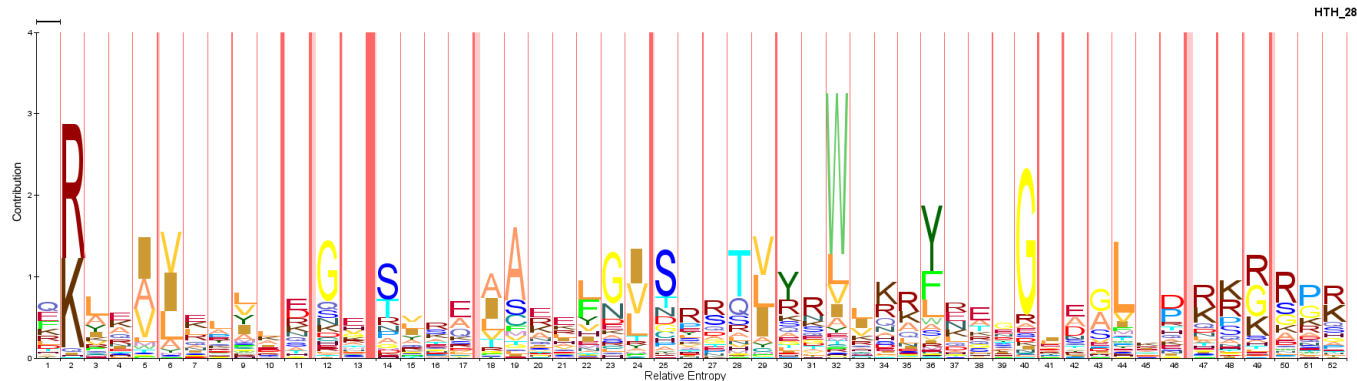
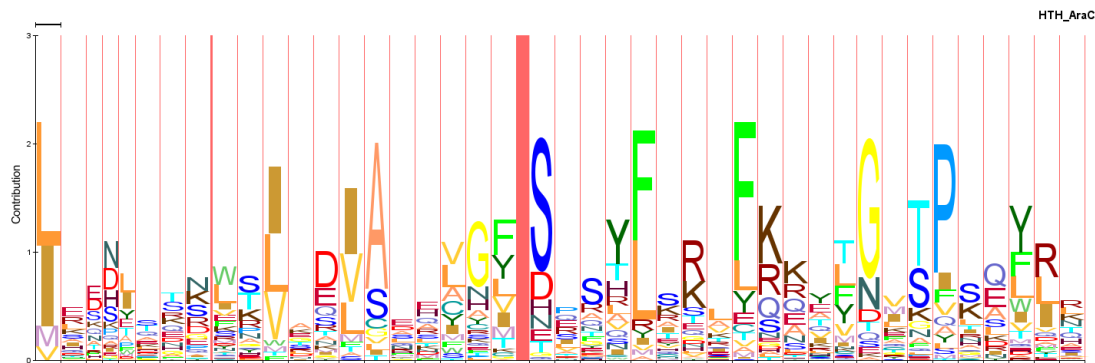
Q-values for domains

Introduced domain-specific features

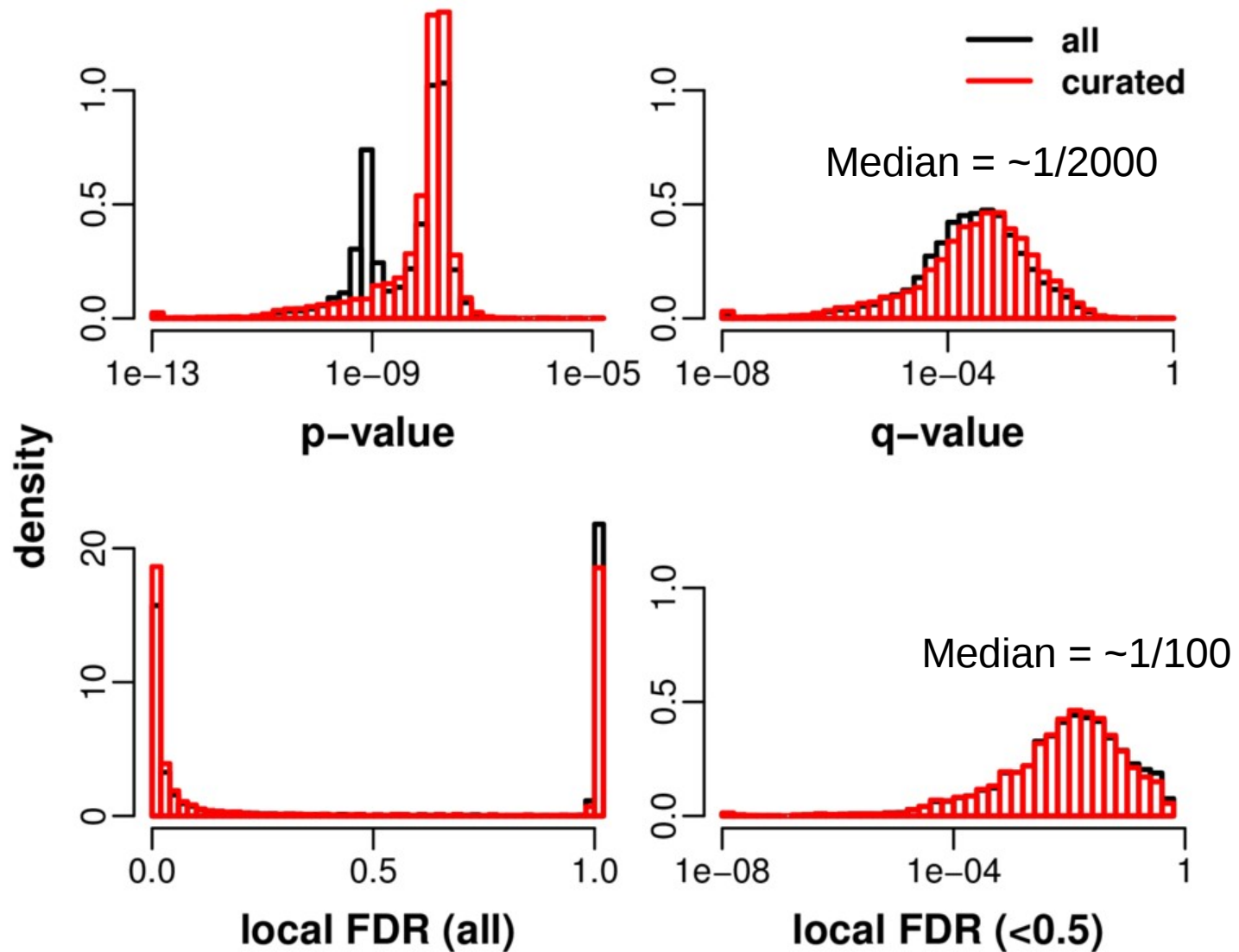
- Incomplete list of p -values (HMMER3, BLAST and descendants)
- Correction for domain overlaps (heuristic)

Two members of Pfam “Clan” HTH

$$FDR = E/n$$



Pfam curated thresholds



Empirical null models

$$FDR = \#FP / (\#FP + \#TP)$$

2nd order Markov Random Sequences

- FP if domain came from random sequence

Improved from Ochoa, *et al.* BMC Bioinformatics 12, 90 (2011).

Ortholog Set Coherence

- FP if orthologs don't predict any homologous domains

Improved from Ochoa, *et al.* BMC Bioinformatics 12, 90 (2011).

Clan Overlap

- FP if domain overlaps stronger non-homologous domain

Inspired/adapted from S. Eddy (p.c., 2012).

Context Coherence

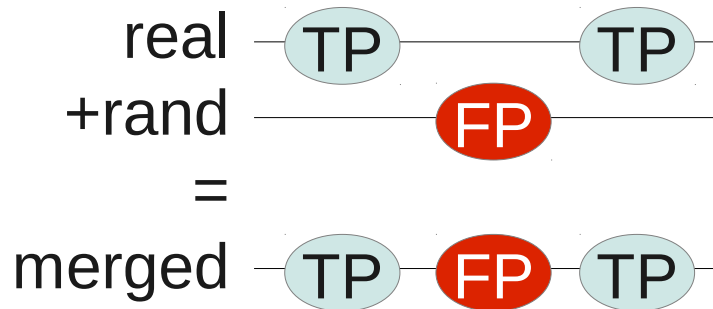
- FP if domain doesn't co-occur with any stronger domains

Inspired/adapted from Terrapon, *et al.* BMC Bioinformatics 13, 67 (2012).

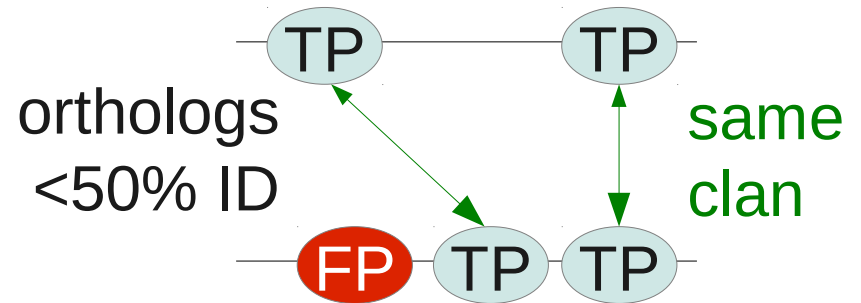
Empirical null models

$$FDR = \#FP / (\#FP + \#TP)$$

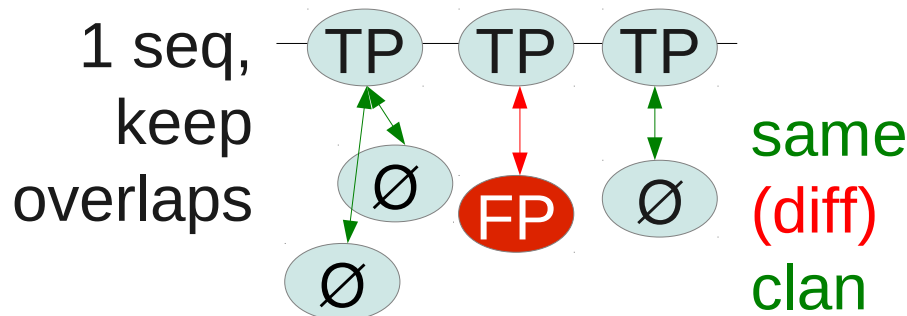
2nd order Markov Random



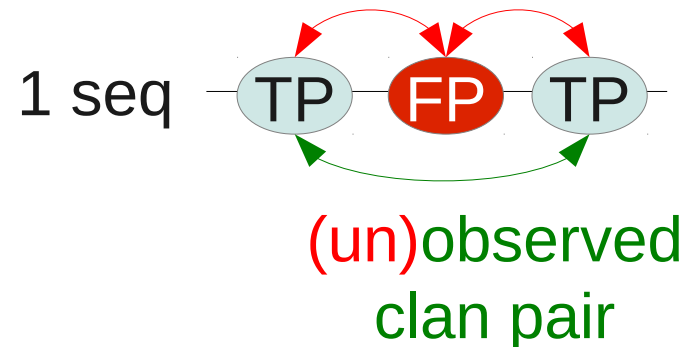
Ortholog Set Coherence



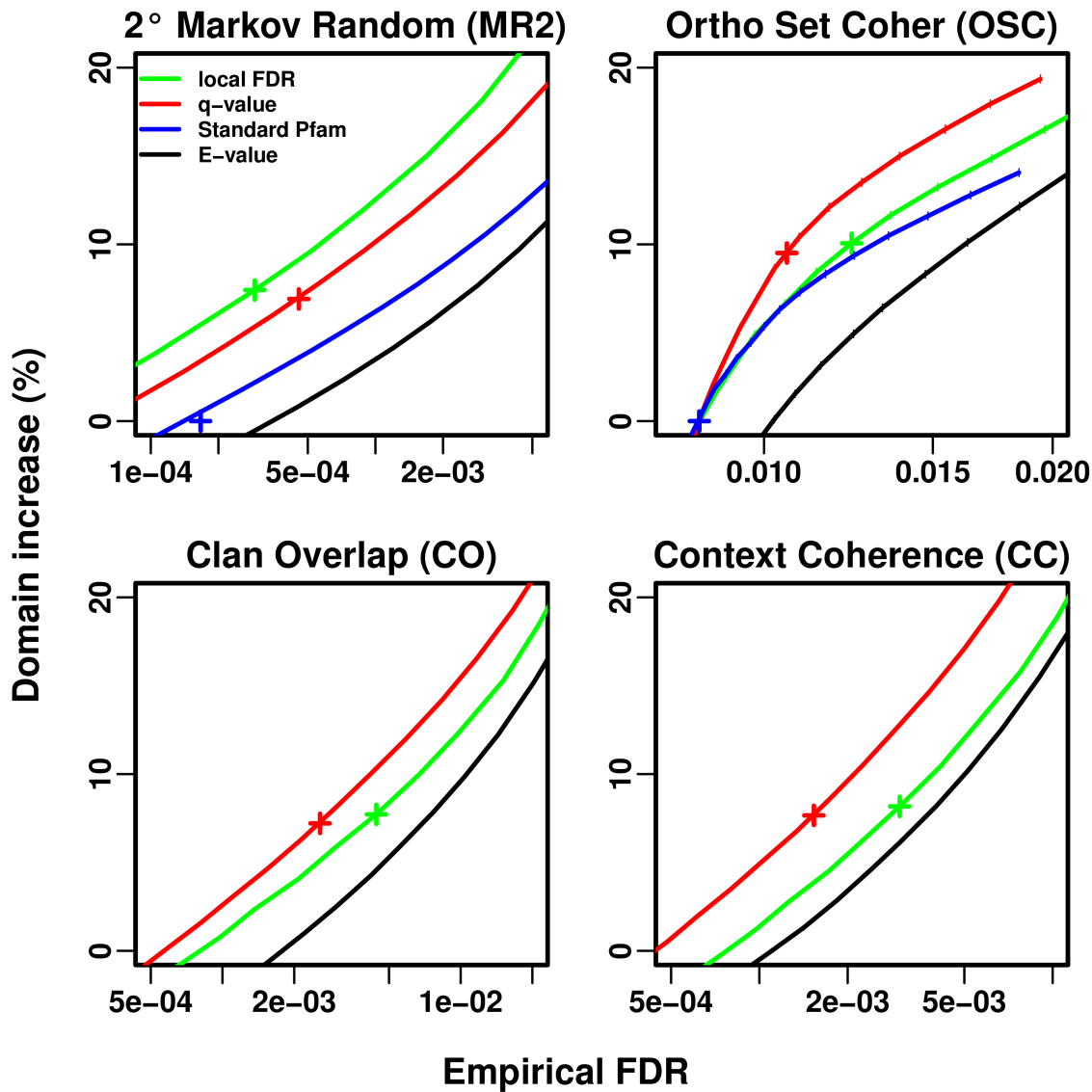
Clan Overlap



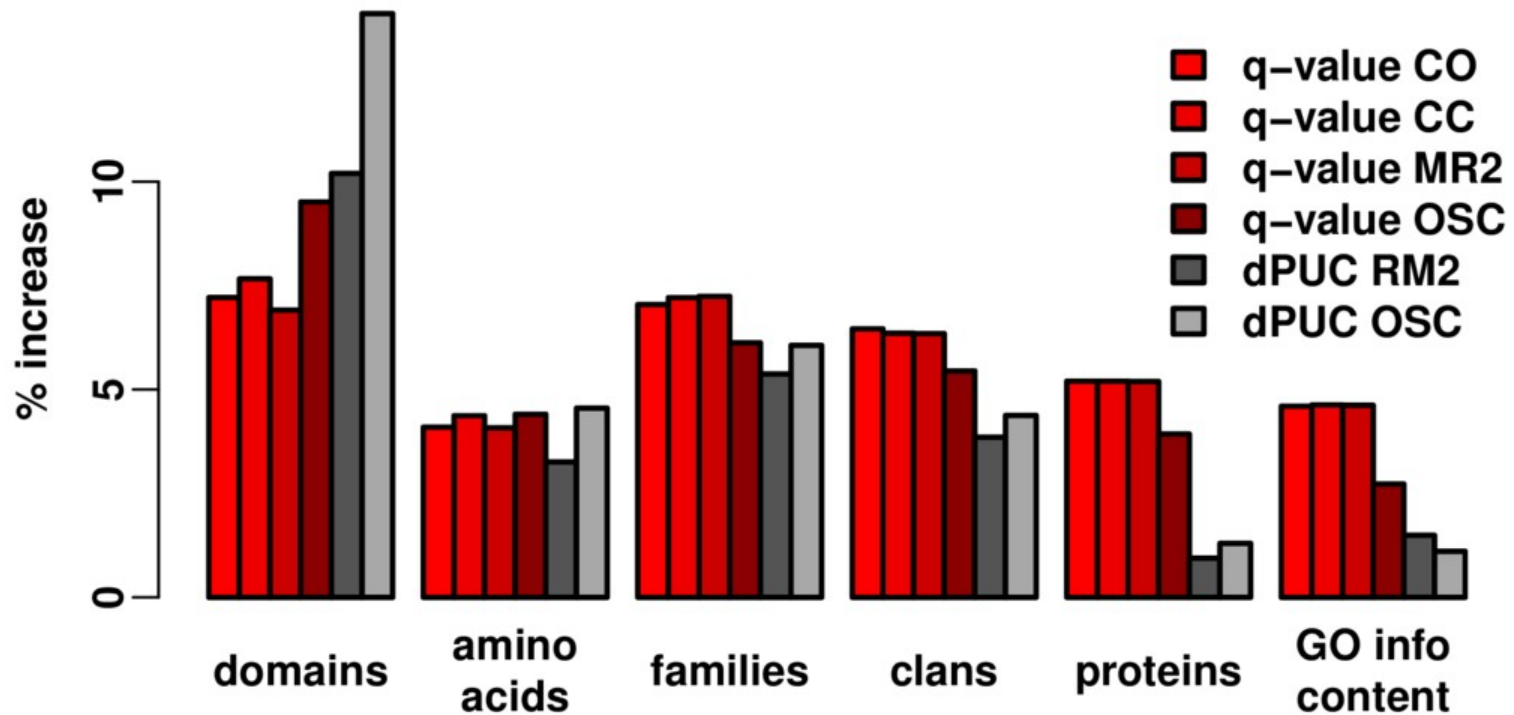
Context Coherence



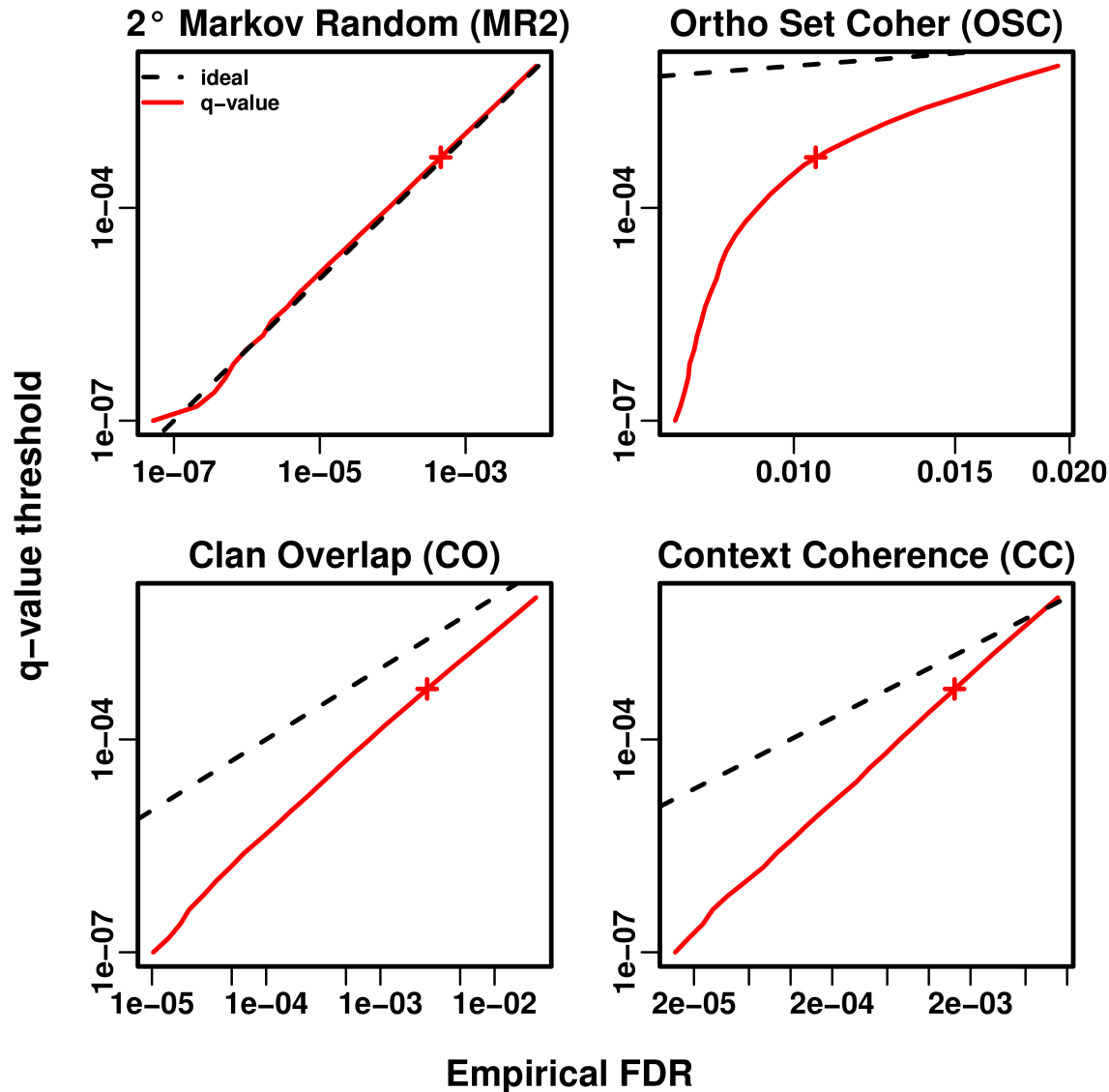
$q > local\ FDR$



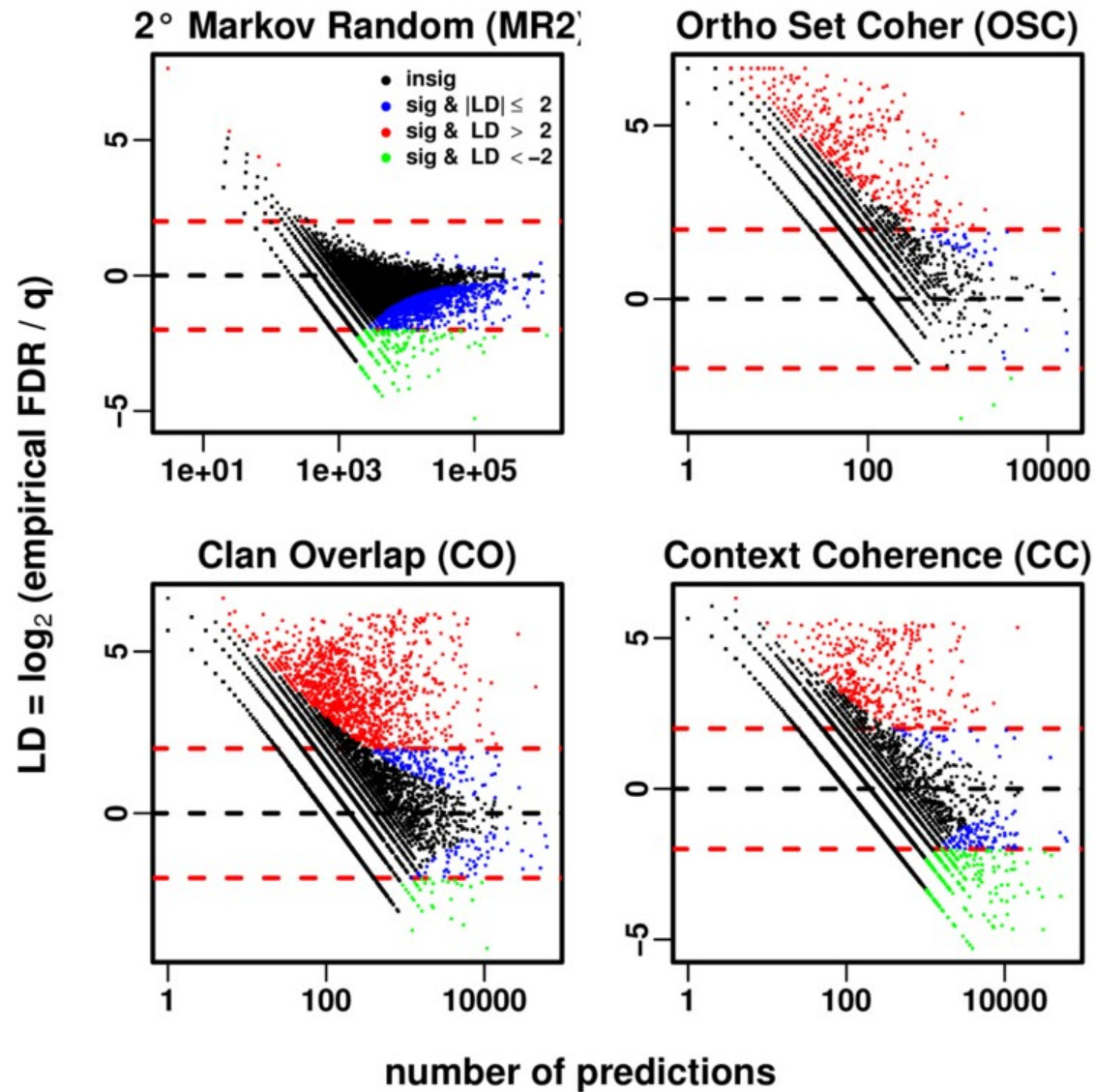
Improving more than just domains



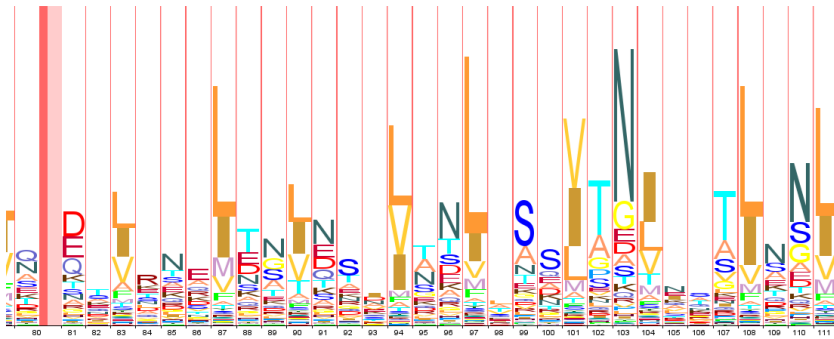
The q -values underestimate empirical $FDRs$



Per family analysis of noise



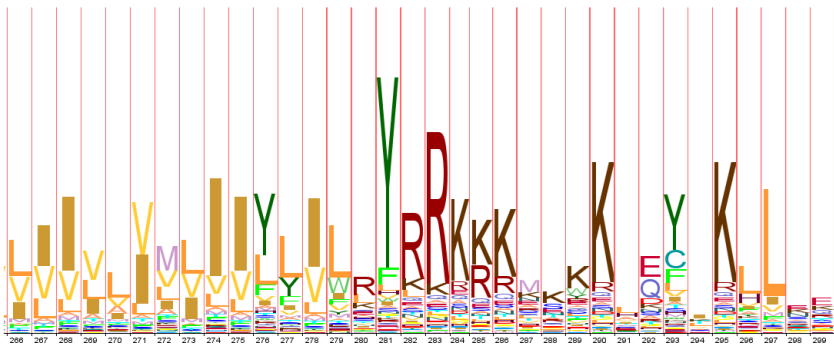
Classes of noisy domains



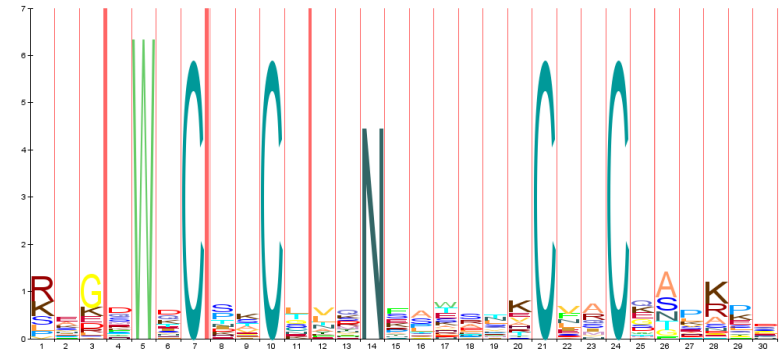
Coiled coils: 403 (3%)



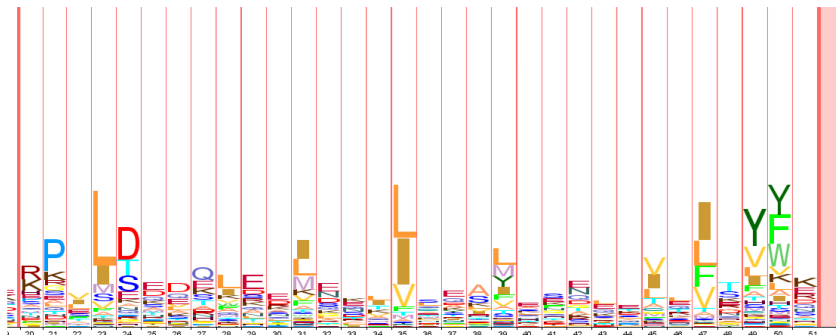
Transmembranes: 1466 (12%)



Low complexity: 1987 (16%)



Cysteine-rich: 587 (5%)

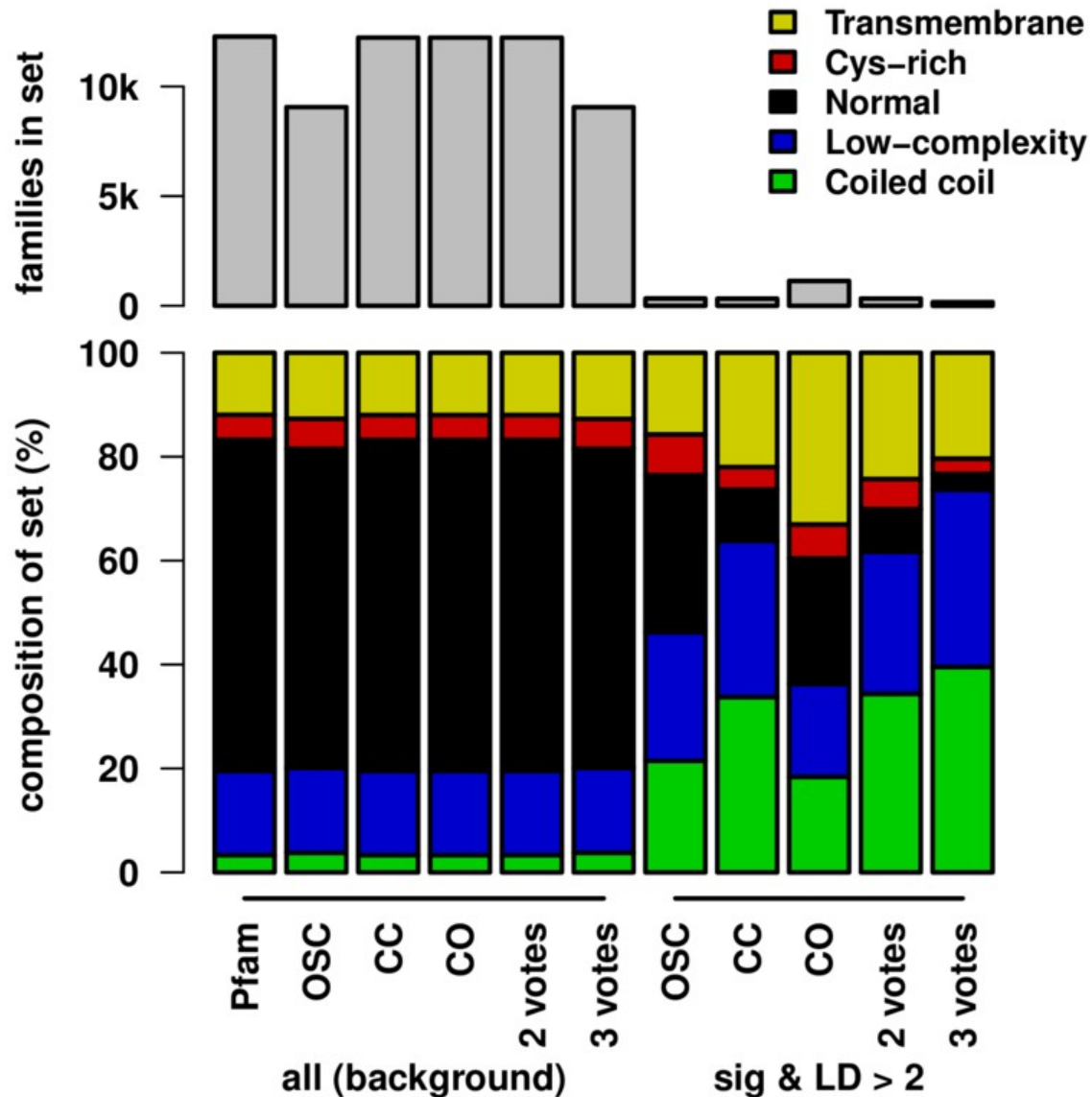


**Normal:
7830 (64%)**

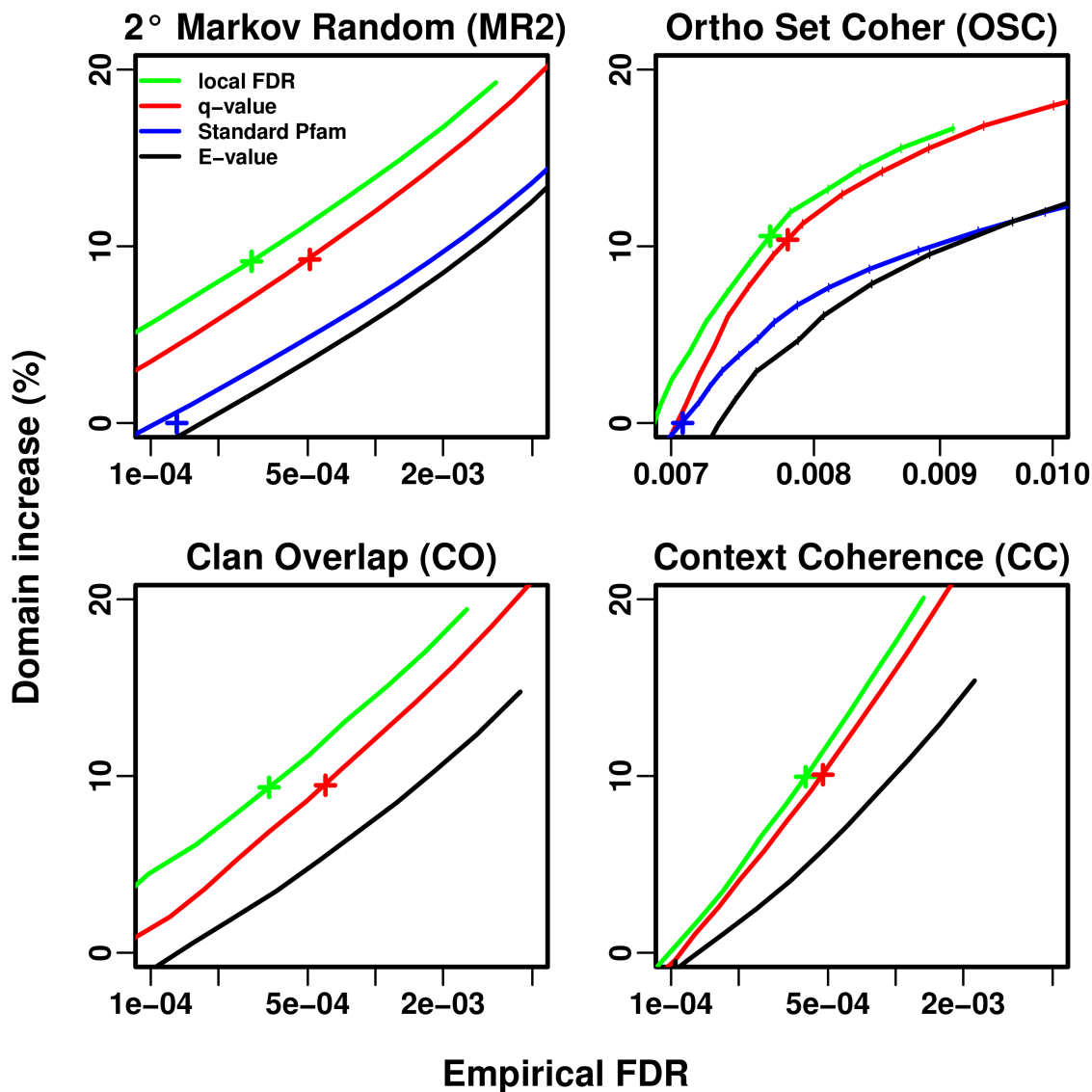
Similar but non-homologous families
(different clans)

Rackham *et al.* JMB 403, 480–493 (2010).
Wong, *et al.* PLoS Comput Biol 6, e1000867 (2010).

Classes enriched in noisy domains



Local FDR > *q* in families with correct stats



Conclusions

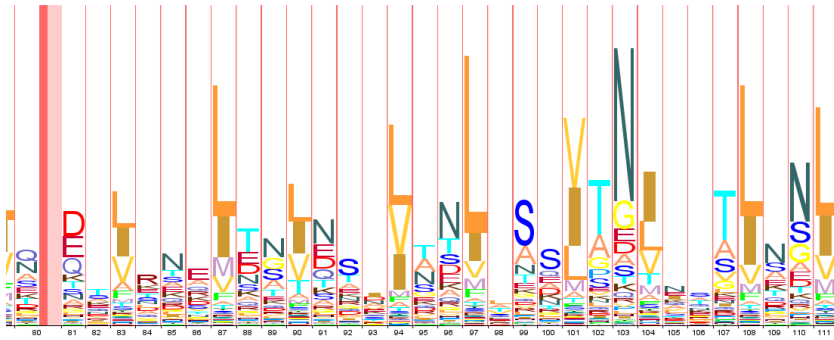
The q -value and *local FDR* are better for domains

- And likely better for regular sequence database searches, iterated searches, orthology prediction
- E -values do not control posterior error probability

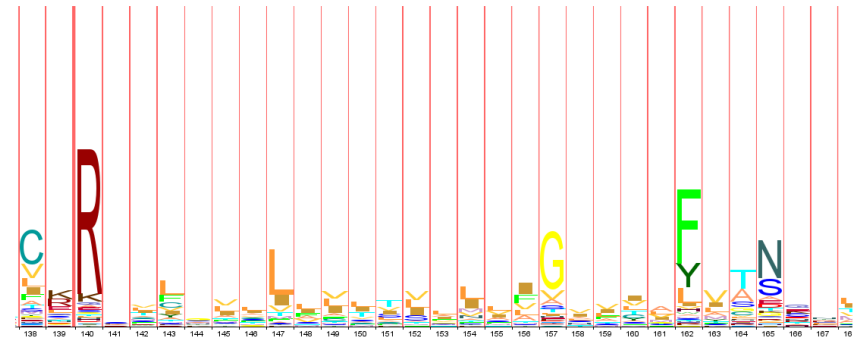
Presented novel empirical null models

- Needed to verify theory is correct
- Uses common-sense biological information and real, full protein sequences
- Structural benchmarks (i.e. SCOP) are limited to well-studied, single domains from model organisms, and exclude coiled coils and transmembrane domains

Noisy domains



Coiled coils



Transmembranes

These remain large problems in all sequence analysis.

Solutions?

- Cannot prevent or ignore these queries
- Masking removes too much information
- Benchmarks not powerful enough to give better thresholds
- Can we properly handle these common, correlated patterns that do not imply homology?



PRINCETON
UNIVERSITY



- **Mona Singh, Computer Science**

- Jesse Farnham
- Dario Ghersi
- Peng Jiang
- Shilpa Nadimpalli
- Anton Persikov
- Yuri Pritykin
- Pawel Przytycki
- Josh Wetzel

- **Thesis Committee**

- John Storey
- Leonid Krugliak
- Saeed Tavazoie

- **Manuel Llinás, Molecular Biology**

- Lindsey Altenhofen
- Katie Baska
- Simon Cobbold
- Valerie Crowley
- Björn Kafsack
- Ian Lewis
- Jessica O'Hara
- Heather Painter
- Joana Santos
- Ariel Schieler
- April Williams

- **NSF GRFP**

<http://compbio.cs.princeton.edu/dpuc/>