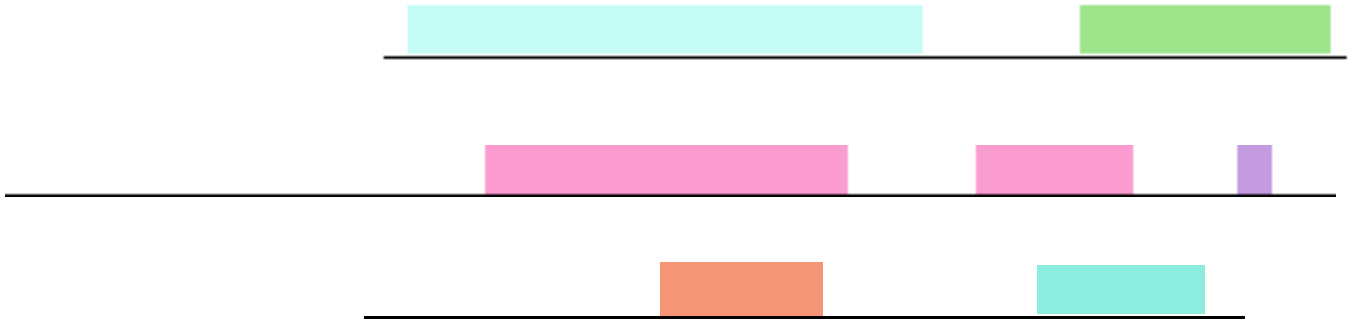
Beyond the *E*-value: stratified statistics for protein domain prediction



Alejandro Ochoa, John D. Storey, Manuel Llinás, Mona Singh

Center for Statistics and Machine Learning, Princeton University

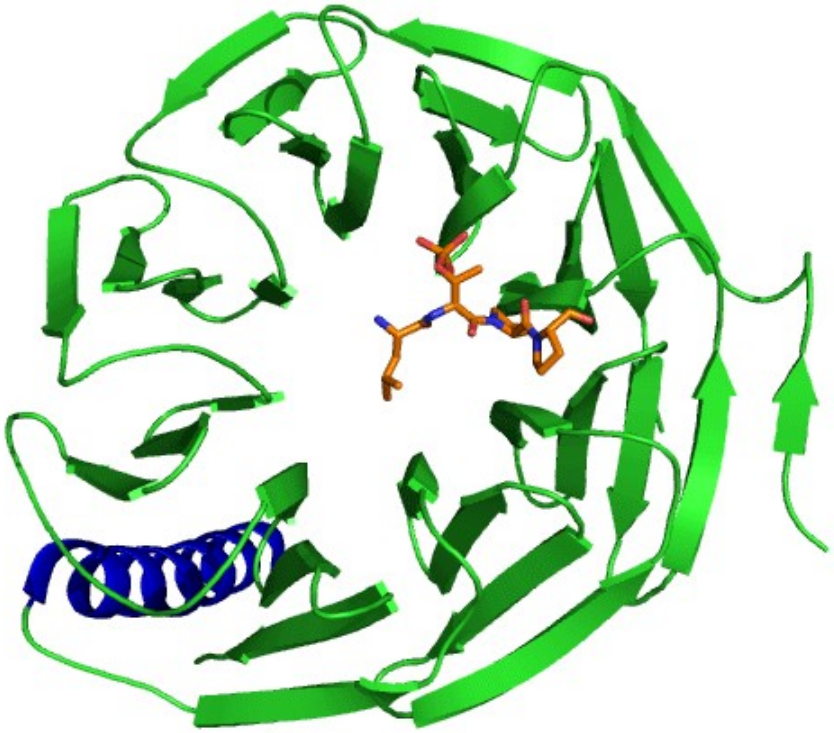
Probabilistic Modeling in Genomics, Cold Spring Harbor Laboratory

2015-10-15

CENTER FOR STATISTICS AND MACHINE LEARNING

PRINCETON UNIVERSITY

Protein domains



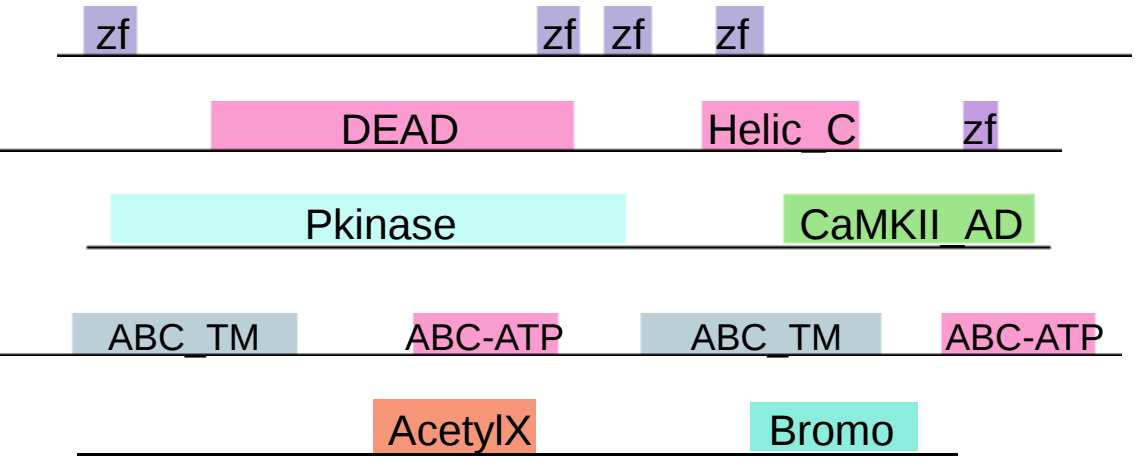
Structure

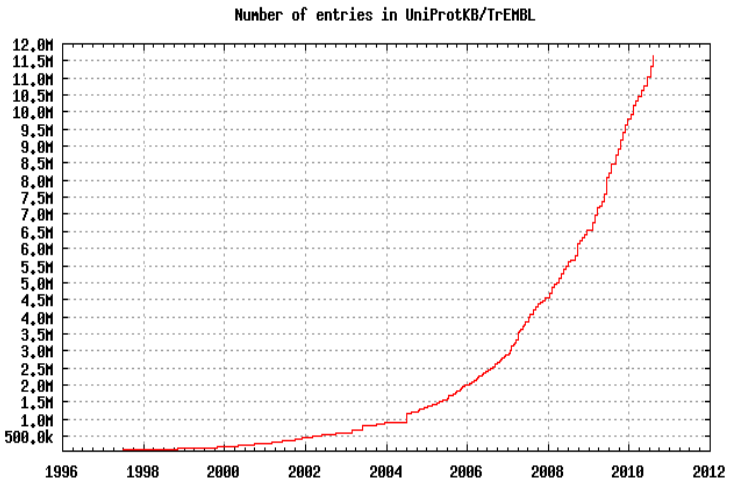
Evolution

Function

Sequence-based domain prediction:

Why predict domains?





For new sequences, before experiments start…

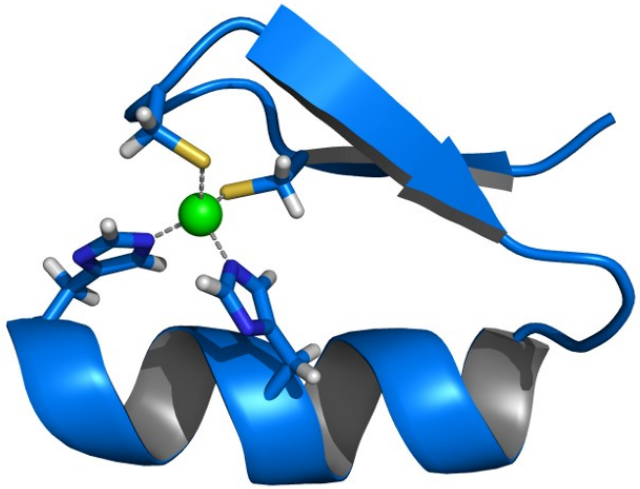
Domains may imply functions

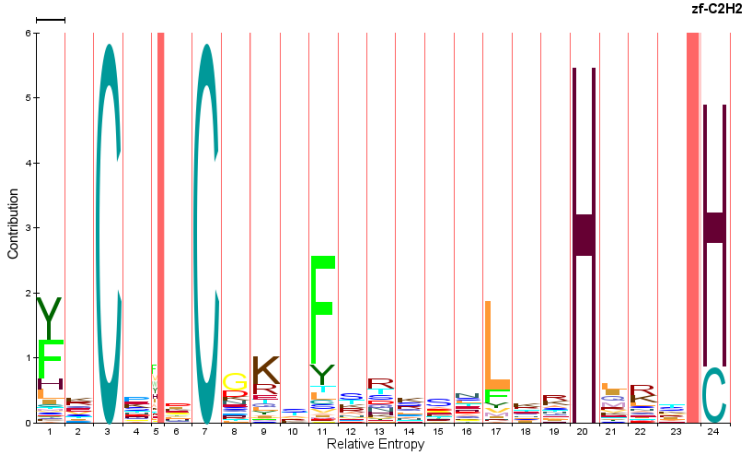
Experimental alternatives are unfeasible since protein databases grow exponentially

Representing Domains

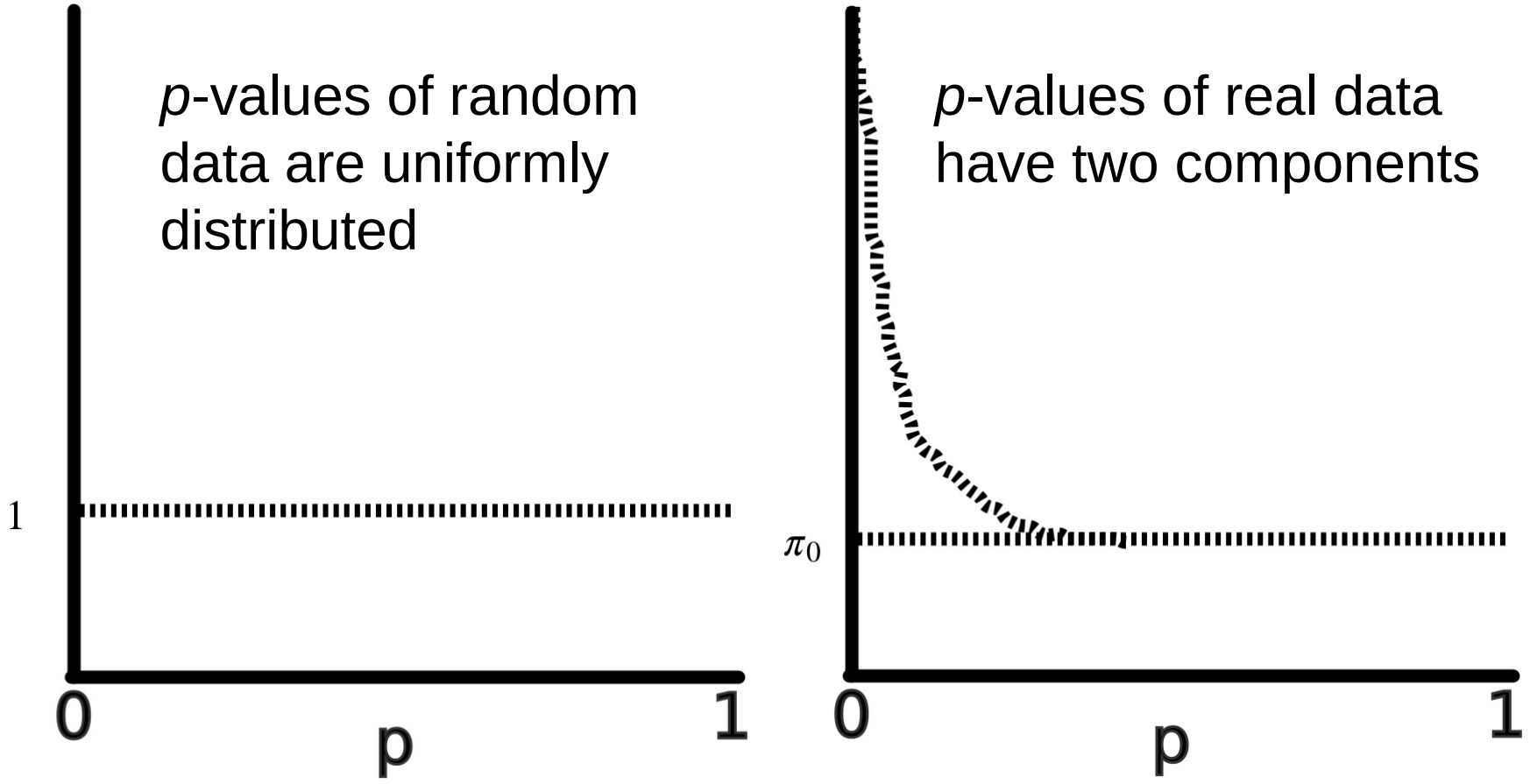
```
SNAI_DROME/362-385      YACQ...VCH...KSFSRM...SLLNKHSSS..NC
SNAI_XENLA/232-255      YQCK...SCS...RTFSRM...SLLHKHEET..GC
SNAI_MOUSE/236-259      YQCQ...ACA...RTFSRM...SLLHKHQES..GC
ESCA_DROME/426-449      YSCT...SCS...KTFSRM...SLLTKHSEG..GC
SUHW_DROAN/221-243      HVCG...KCY...KTFRRL...MSLKKHLEF...C
TERM_DROME/323-346      LHCR...RCR...TQFSRR...SKLHIHQKL..RC
ZO20_XENLA/174-196      FMCA...DCG...RCFSVS...SSLKYHQRI...C
EVI1_HUMAN/217-239      IKCK...DCG...QMFSTT...SSLNKHRRF...C
ZO2_XENLA/34-59         YSCA...DCG...KHFSEK...MYLQFHQKNPSEC
EVI1_HUMAN/21-44        YRCE...DCD...QLFESK...AELADHQKF..PC
ZNF10_HUMAN/517-539     YKCN...QCG...IIFSQN...SPFIVHQIA...H
ZNF91_HUMAN/238-260     YKCE...ECG...KAFKQL...STLTTHKII...C
ZFP58_MOUSE/120-142     IKCE...ECG...KAFSTR...STYYRHQKN...H
TRA1_CAEEL/306-331      YKCEF.ADCE...KAFSNA...SDRAKHQNR..TH
ZNF76_HUMAN/345-368     YTCS...TCG...KTYRQT...STLAMHKRS..AH
ZN12_MICSA/106-129      YRCS...QCG...KAFRRT...SDLSSHRRT..QC
LOLA1_DROME/794-817     YECR...HCG...KKYRWK...STLRRHENV..EC
ZNF17_HUMAN/435-457     YECN...KCG...KFFRYC...FTLNRHQRV...H
ZG32_XENLA/34-56        FVCV...HCG...KGFRDN...YKLSLHLRI...H
TF3A_BUFAM/104-128      YVCYF.ADCG...QQFRKH...NQLKIHQYI...H
ZG46_XENLA/146-168      YVCT...ECG...TSFRVR...PQLRIHLRT...H
MZF1_HUMAN/412-434      FVCG...DCG...QGFVRS...ARLEEHRRV...H
ZN239_MOUSE/6-28        YKCD...KCG...KGFTRS...SSLLVHHSV...H
ZSC22_HUMAN/352-374     YKCG...ECG...KTFSRS...THLTQHQRV...H
EGR1_HUMAN/396-418      FACD...ICG...RKFARS...DERKRHTKI...H
SUHW_DROAN/349-373      YACK...ICG...KDFTRS...YHLKRHQKYS.SC
CF2_DROME/485-508       YTCP...YCD...KRFTQR...SALTVHTTK..LH
CF2_DROME/401-423       YTCS...YCG...KSFTQS...NTLKQHTRI...H
KRUP_DROME/306-328      YTCE...ICD...GKFSDS...NQLKSHMLV...H
TYY1_HUMAN/383-407      YVCPF.DGCN...KKFAQS...TNLKSHILT...H
ZG52_XENLA/61-83        YTCT...QCN...KQFSHS...AQLRAHIST...H
TTKB_DROME/538-561      YPCP...FCF...KEFTRK...DNMTAHVKI..IH
ZNF76_HUMAN/285-309     YTCPE.PHCG...RGFTSA...TNYKNHVRI...H
SDC1_CAEEL/145-168      YMCQ...VCL...TLFGHT...YNLFMHWRT..SC
SRYC_DROME/358-380      YQCD...ICG...QKFVQK...INLTHHARI...H
SDC1_CAEEL/270-292      YFCH...ICG...TVFIEQ...DNLFKHWRL...H
TRA1_CAEEL/276-300      NKCEY.PGCG...KEYSRL...ENLKTHRRT...H
ESCA_DROME/370-392      CKCN...LCG...KAFSRP...WLLQGHIRT...H
```

. . . . . .



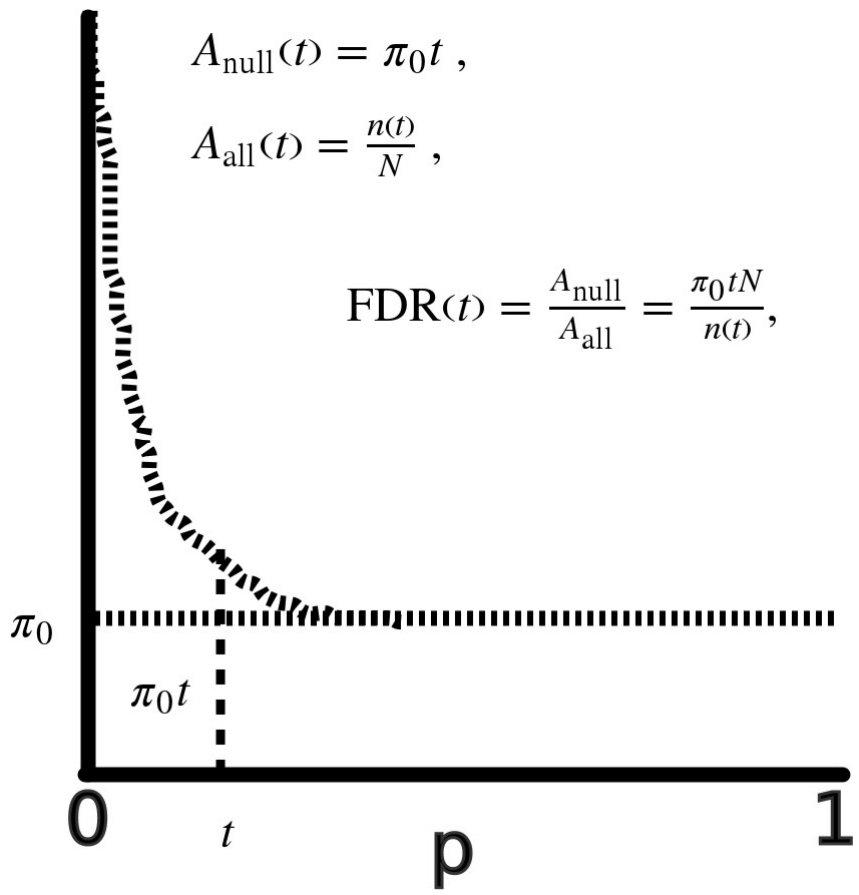


Computing *q*-values



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

Computing *q*-values



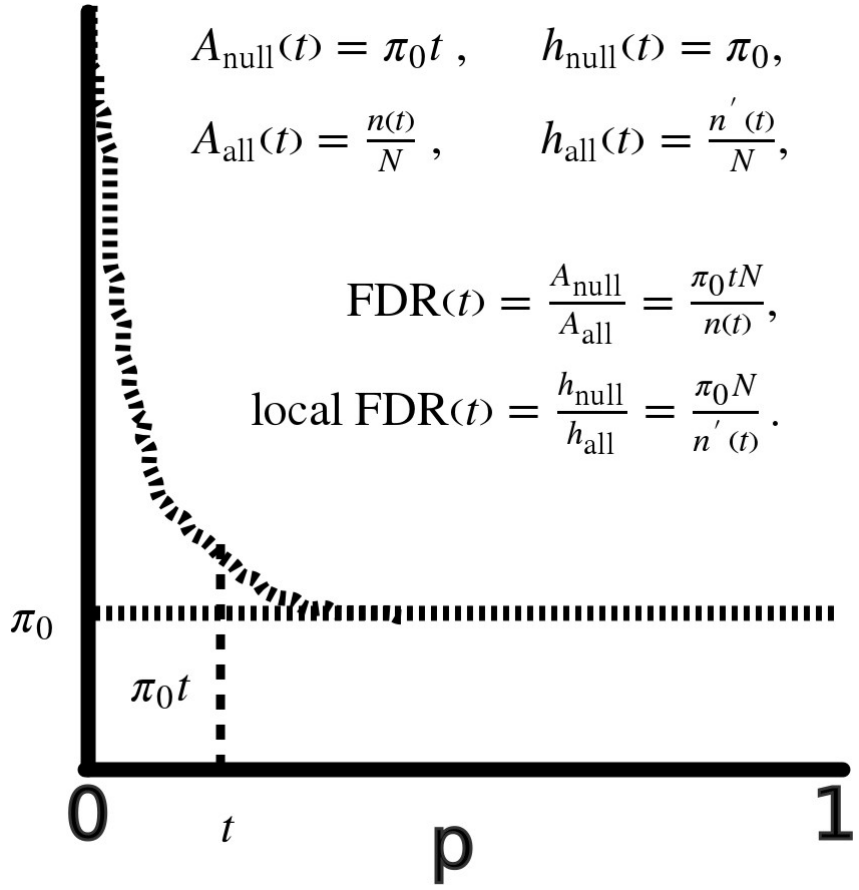
Step 2: Directly estimate FDR(*t*) for all thresholds *t*

- *N* = # tests
- *n(t)* = # sig tests

Step 3: Ensure monotonicity

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

The local FDR



Local FDR = Posterior Error Probability

FDR = average local FDR of significant predictions

Local FDRs are optimal for stratified problems

Goal: Find thresholds t_i for each stratum (family) *i* that maximize predictions

$$\sum_i n_i\left(t_i\right),$$

constraining the combined *FDR* across strata to *Q:*

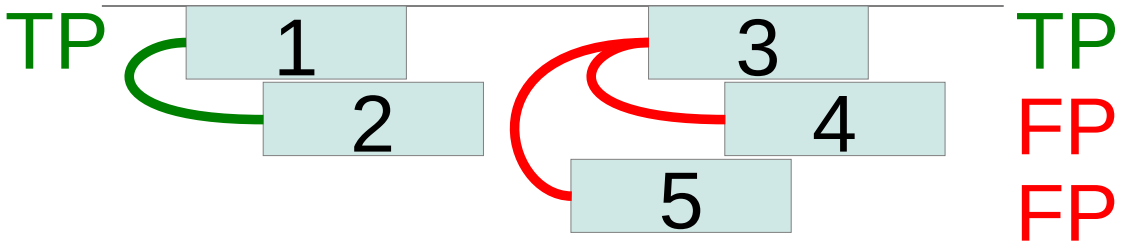
$$\frac{\sum_i \pi_{0,i} t_i N_i}{\sum_i n_i\left(t_i\right)} \leq Q\,.$$

Necessary condition of solution:
Equal per-stratum *local FDR* thresholds!

Empirical FDR tests

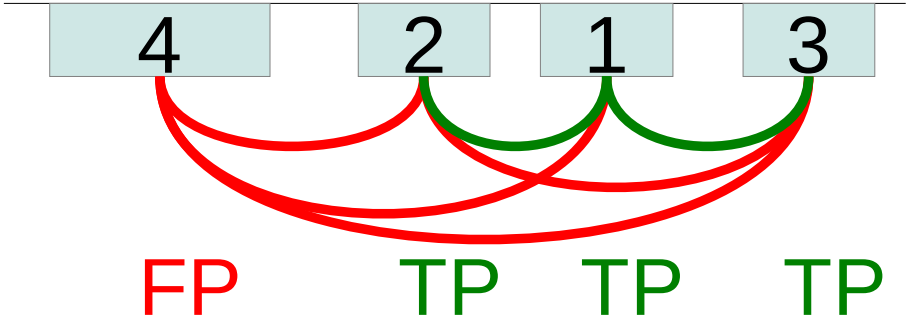
A

Clan Overlap



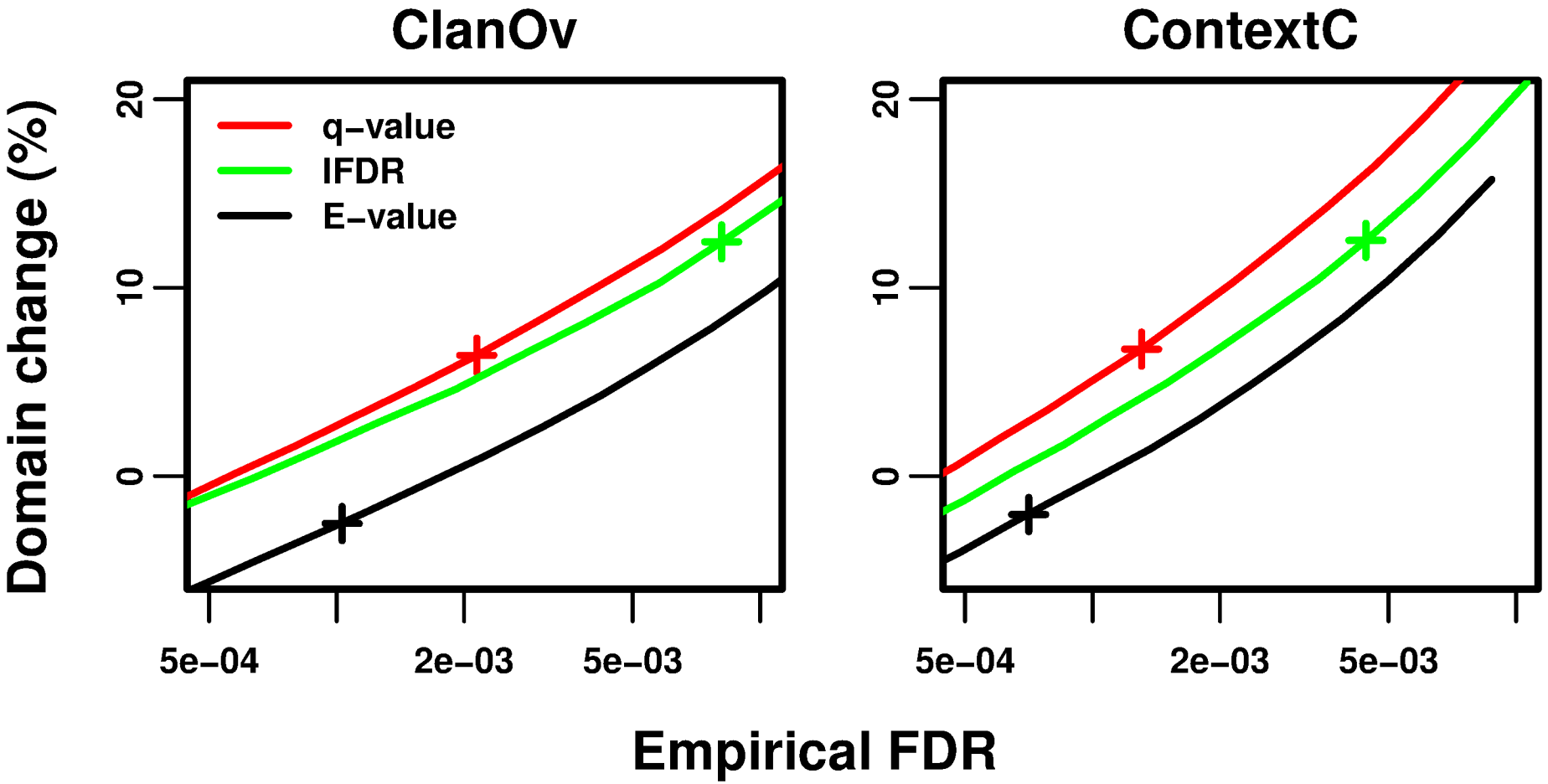
B

Context Coherence

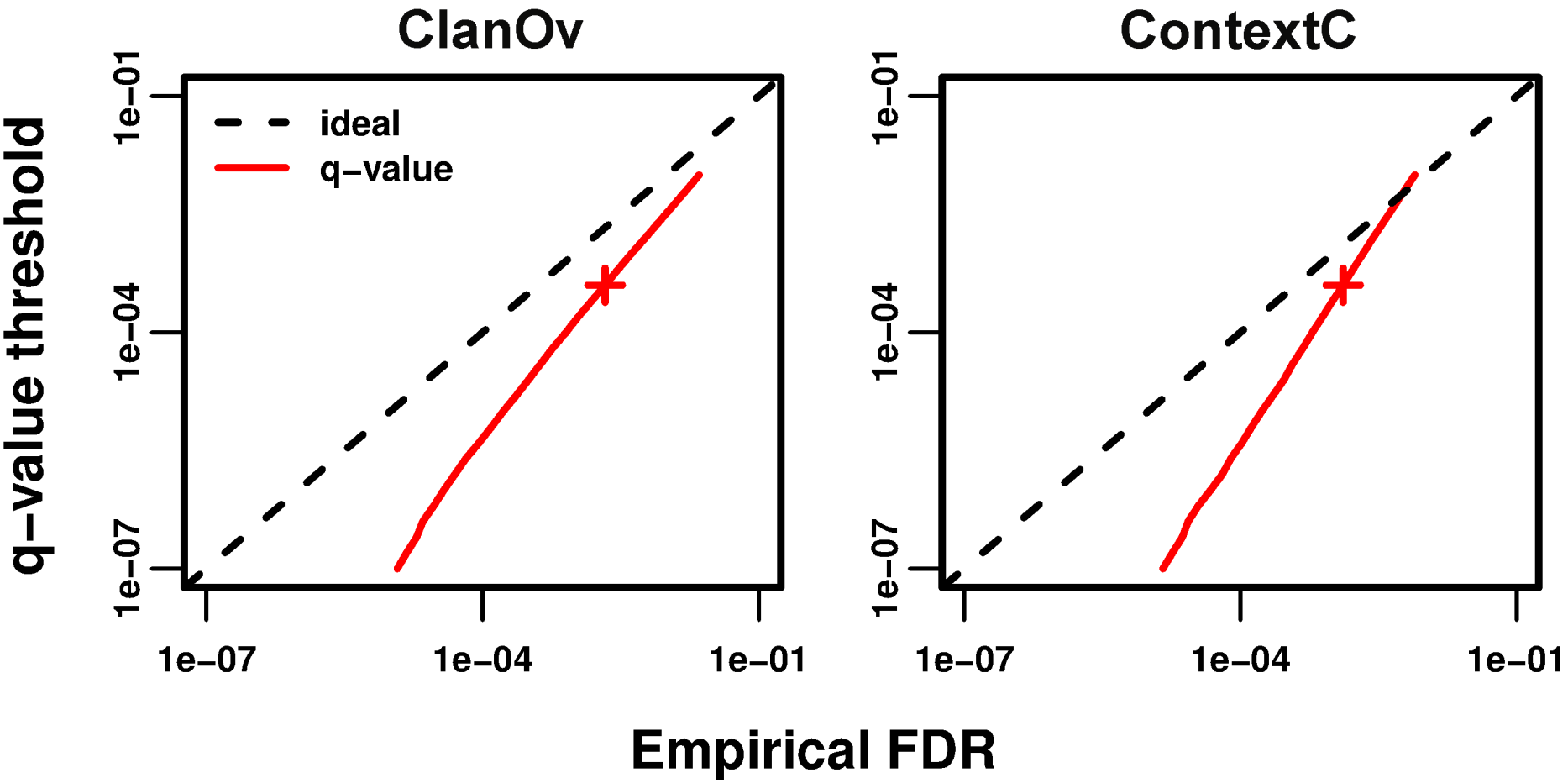


FDR Tests adapted from:
Mistry, *et al*. Nucl Acids Res. 2013;41: e121–e121.
Terrapon, *et al*. BMC Bioinformatics. 2012;13: 67.
Ochoa, *et al*.. BMC Bioinformatics. 2011;12: 90.

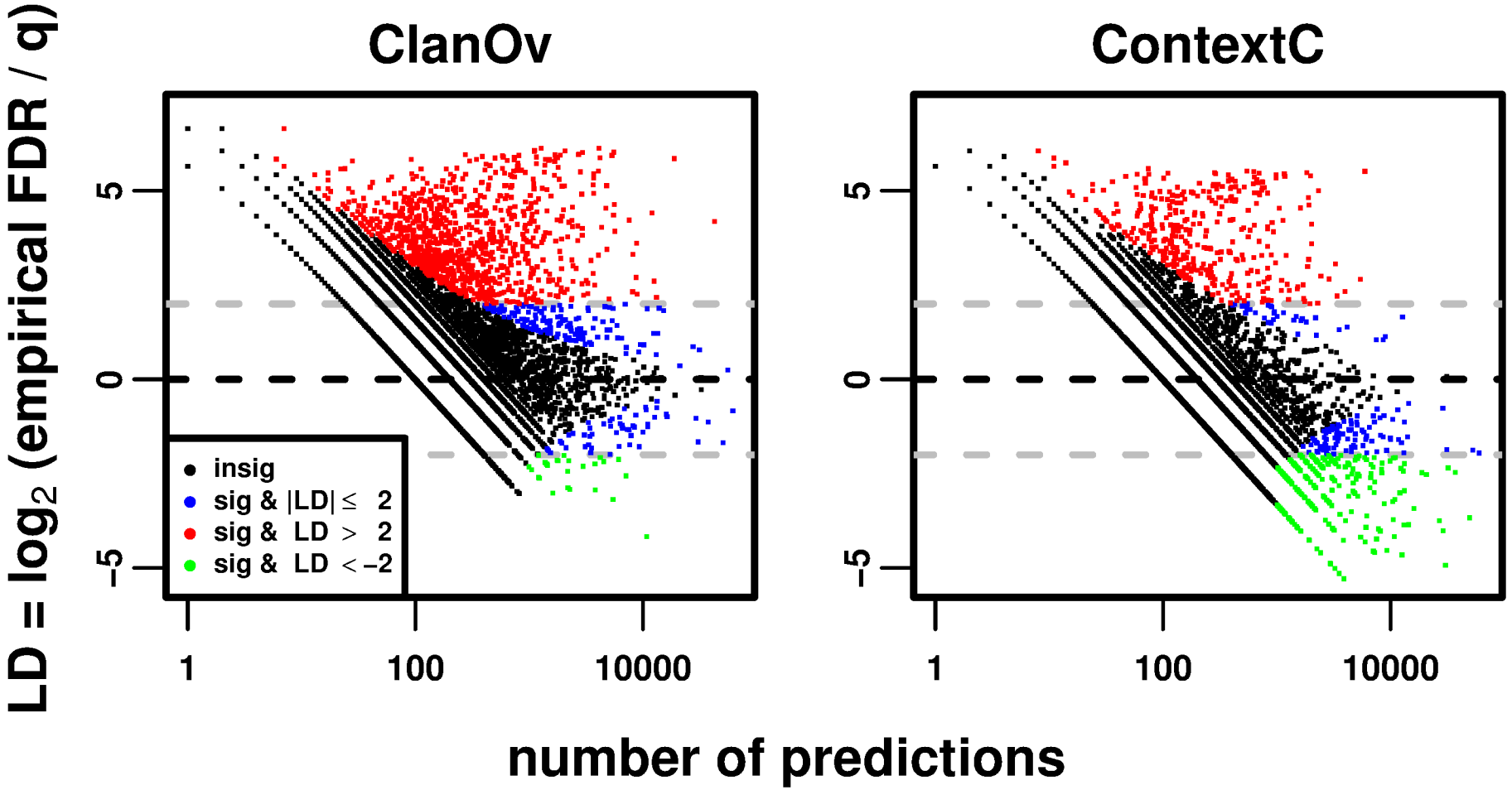
Stratified statistics improve upon *E*-values



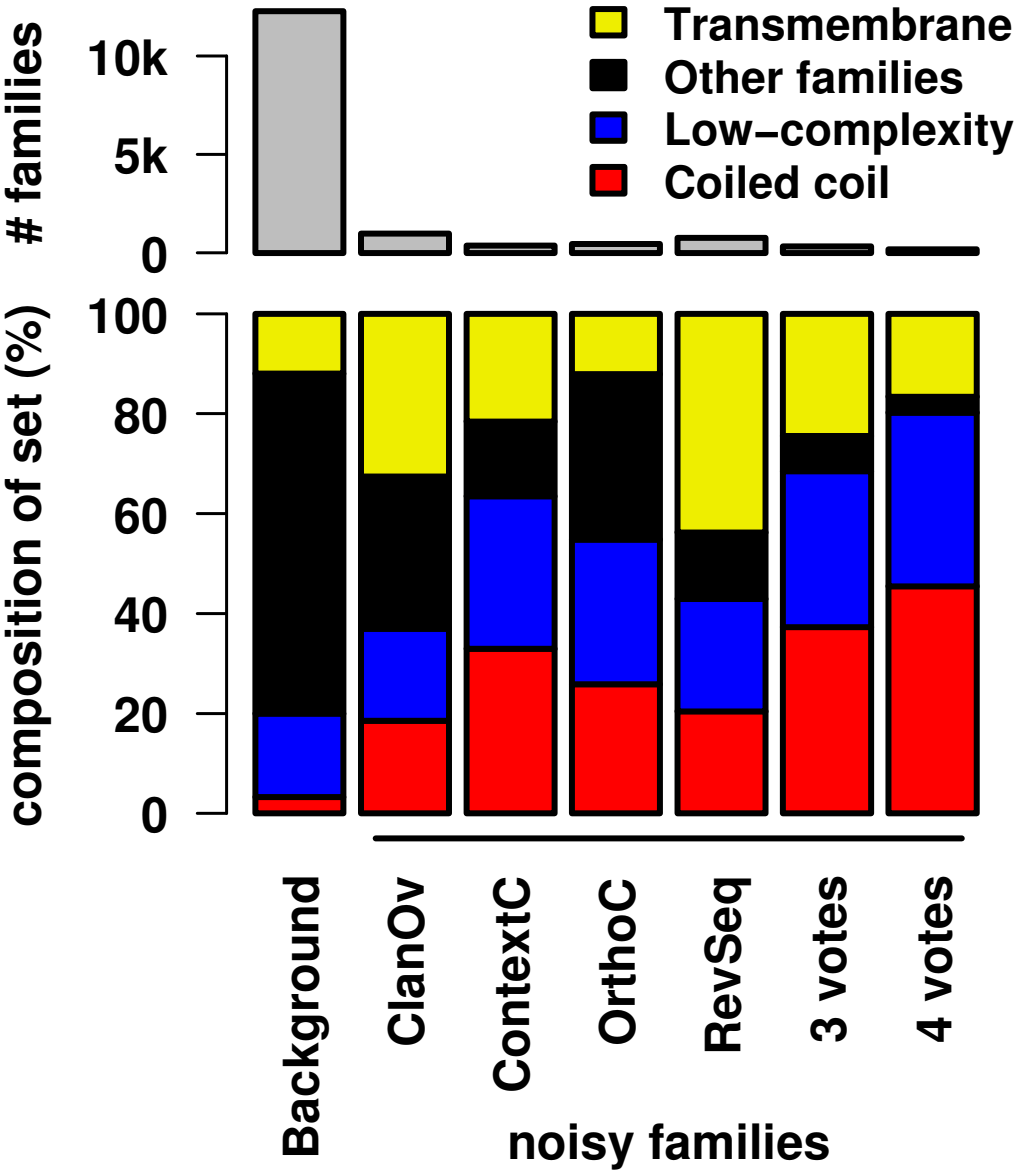
Disagreement between *q*-values and empirical FDRs



Measuring noise per domain family

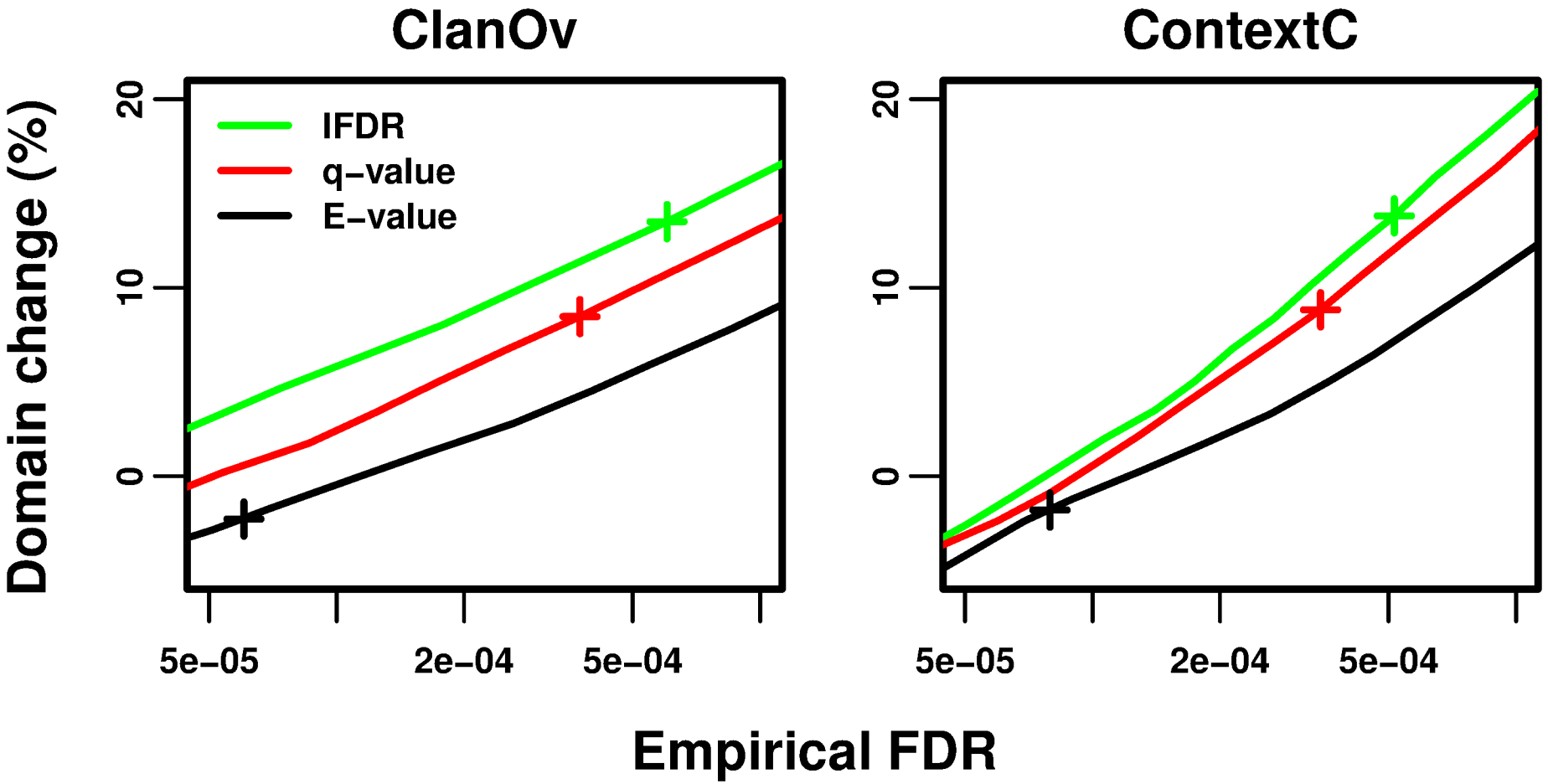


Structural classes enriched in noise

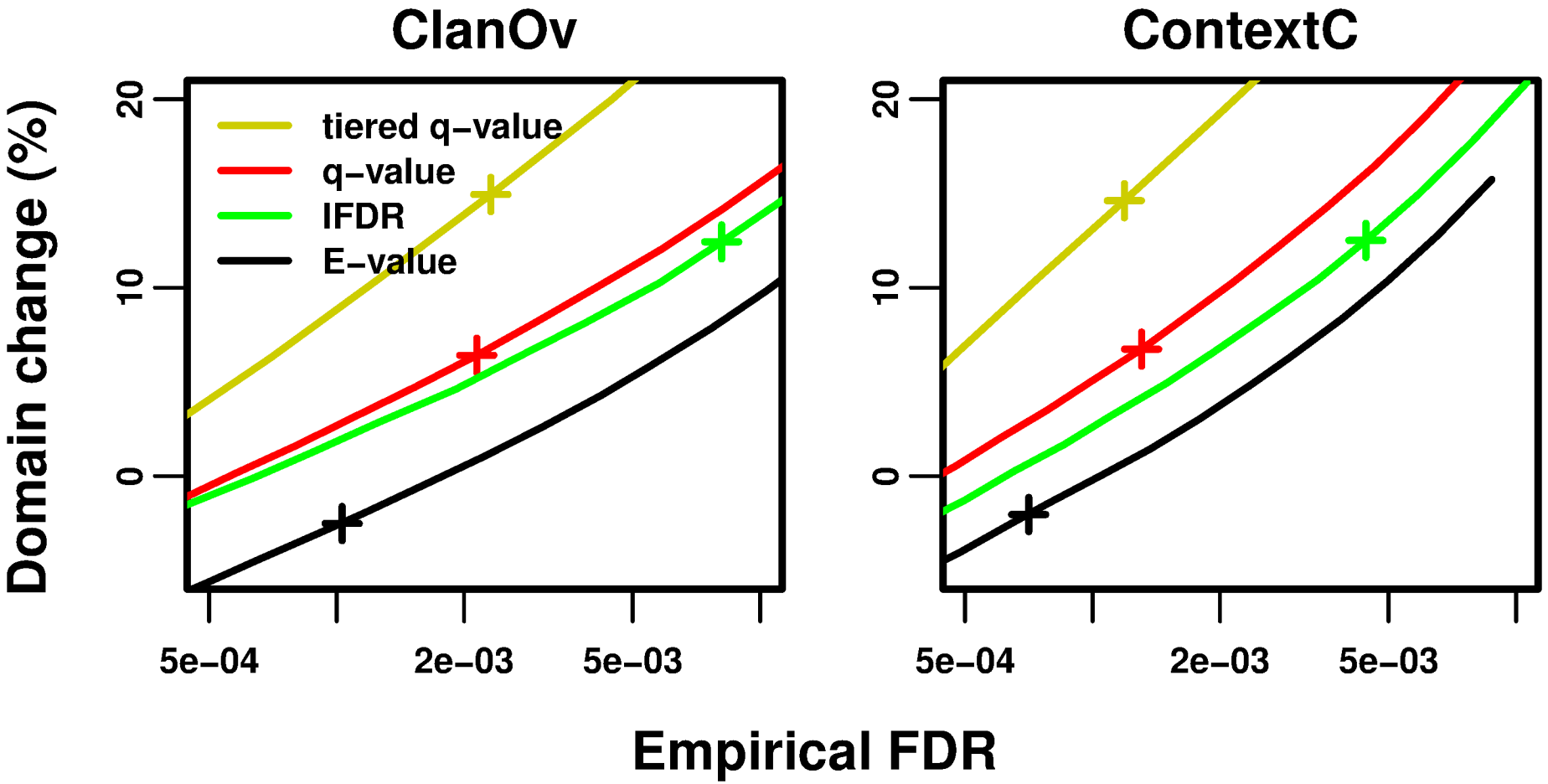


Similar findings:
- Rackham *et al*. JMB 403, 480–493 (2010).
- Wong, *et al.* PLoS Comput Biol 6, e1000867 (2010).
- Mistry, *et al*. Nucl Acids Res. 41: e121–e121 (2013).

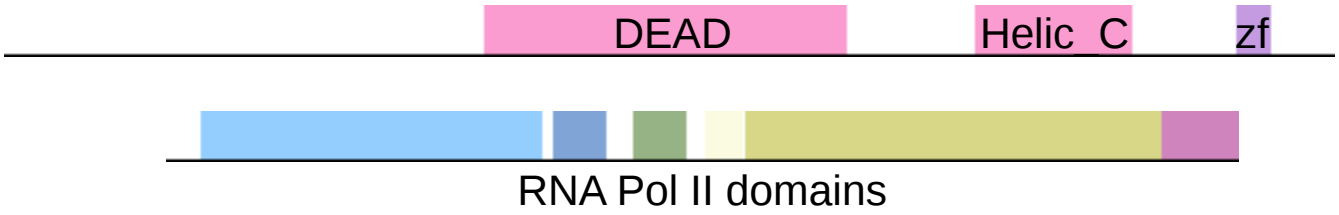
Local FDR outperforms *q*-values in families with correct stats



Tiered *q*-values borrows strength from repeating domains



Domain Prediction Using Context: dPUC



Background

- Domains co-occur in limited combinations

Idea

- Score domains in combination

Future work

- Use *q*-values or local FDRs to improve dPUC

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

Conclusions

Local FDRs are optimal for stratified problems

But *q*-values are more robust to imperfect *p*-values

For domains, repetitive families have inacurate *p*-values

Our FDR-based easy-to-calculate statistics improve domain prediction compared to using *E*-values

Future work:

- Developing methods that combine these statistics with domain context
- Improve *p*-values by improving the standard null model for protein sequences

PRINCETON UNIVERSITY

Mona Singh,
Computer Science, Princeton

Manuel Llinás,
Biochemistry and Molecular Biology, Penn State.

John D. Storey,
Center for Statistics and Machine Learning, Princeton

Funding: NIH, NSF

Software:

DOM STRAT <STATS>

https://github.com/alexviiia/DomStratStats

dPUC2

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