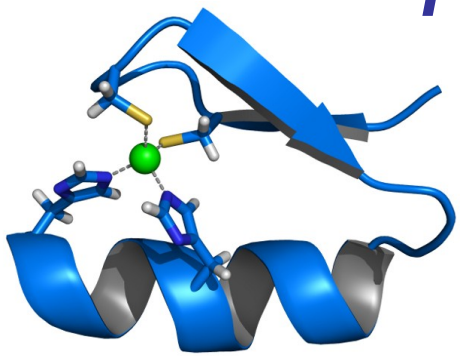
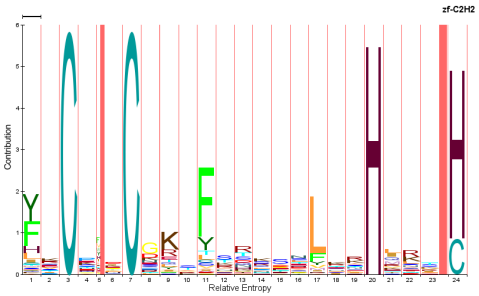
Forget the *E*-value: *q*-values for domains





Protein domains are predicted with statistical models (HMMs), the *E*-value of a score determines significance.

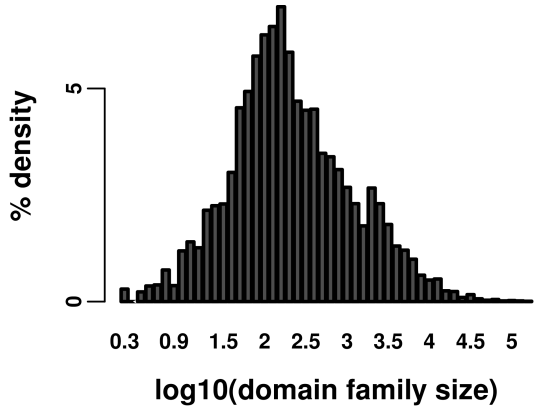
The *E*-value is a bad choice: domain family sizes vary by orders of magnitude!

The False Discovery Rate (*FDR*) is frequently used in microarrays and proteomics (multiple hypothesis testing), connected to posterior error probability.

$$FDR = E/n$$

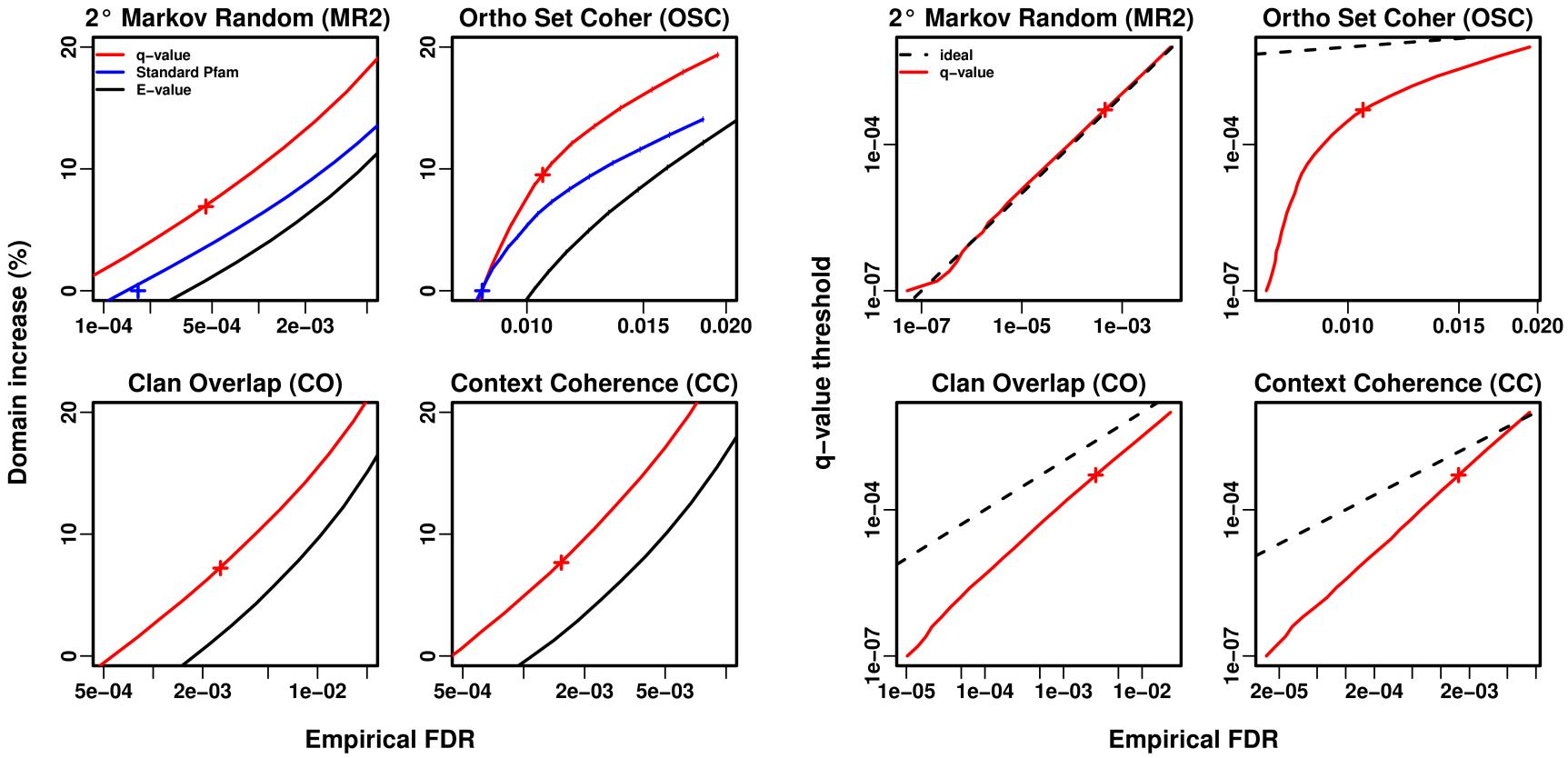
The *q*-value optimally controls the *FDR*. *q*-values are computed from *p*-values.

Our novel *q*-value adaptation for domains: incomplete *p*-values (due to heuristic filters) and domain overlap removal.



Alejandro Ochoa – ochoa@princeton.edu

q-values are better!

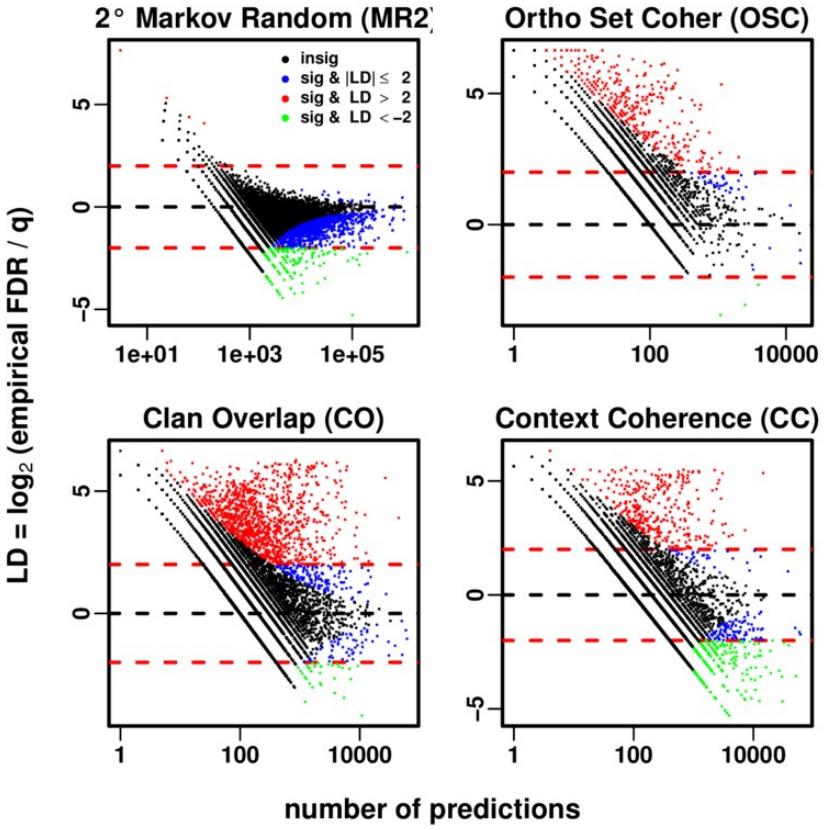


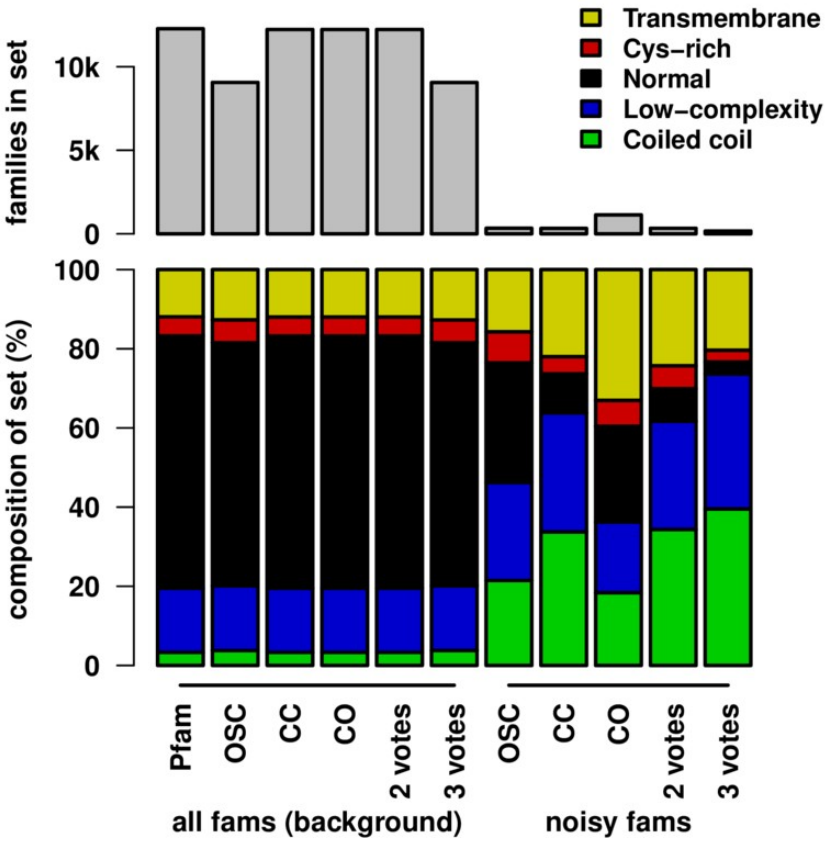
Novel empirical null models to test theory

Empirical *FDR*s and *q*-values don't agree: *p*-values imperfect

Alejandro Ochoa – ochoa@princeton.edu

Repetitive patterns enriched in noisy families





Noisy families pose a problem for automatic threshold selection!

Alejandro Ochoa – ochoa@princeton.edu