Infernal development

Eric Nawrocki and Claude

Benasque RNA meeting, 16 July 2026

National Library of Medicine
National Institutes of Health

Talk outline

1. Introduction to CMs/Infernal

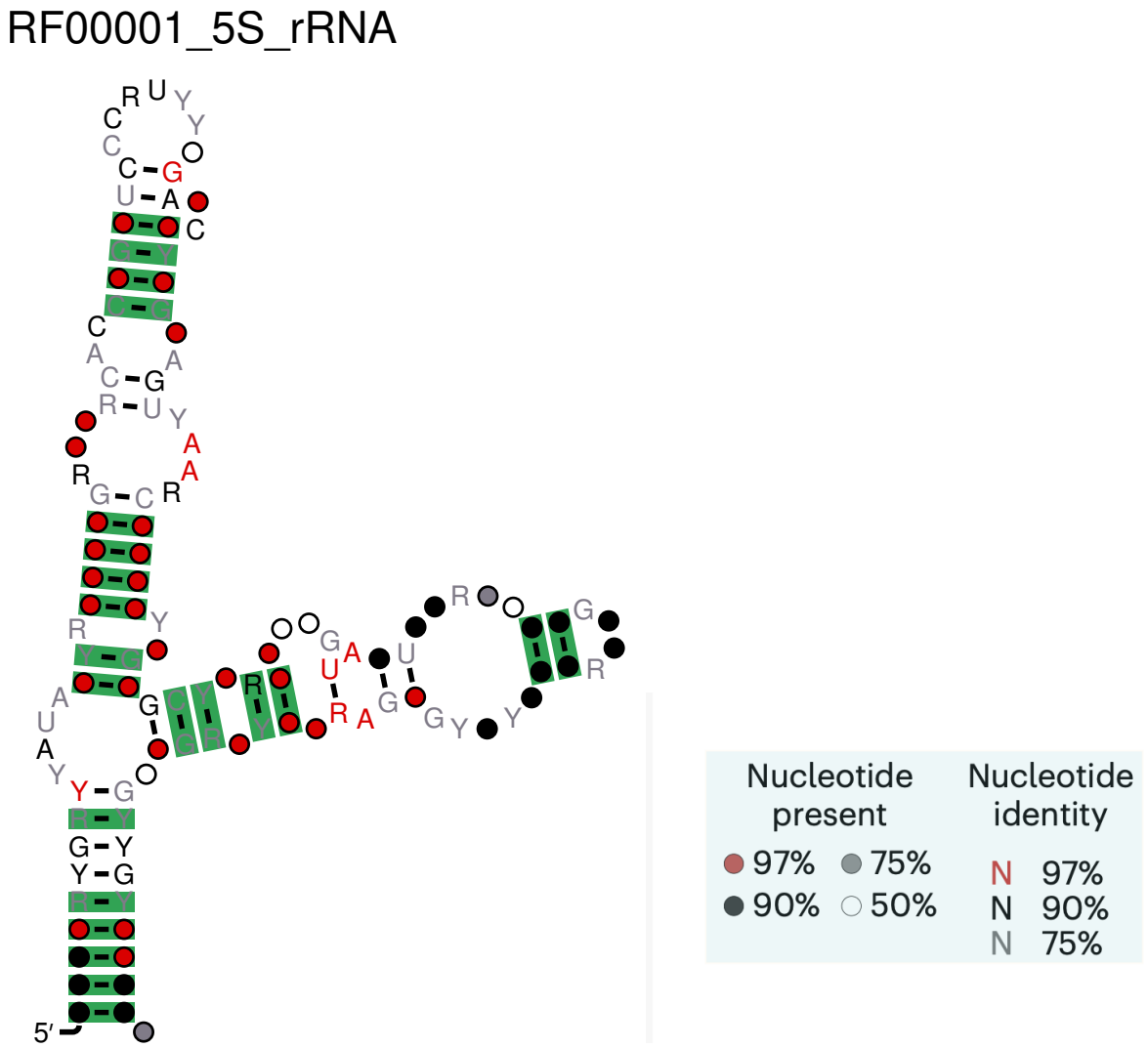
2. Eliminating cmcalibrate

3. Pseudoknot alignment markup

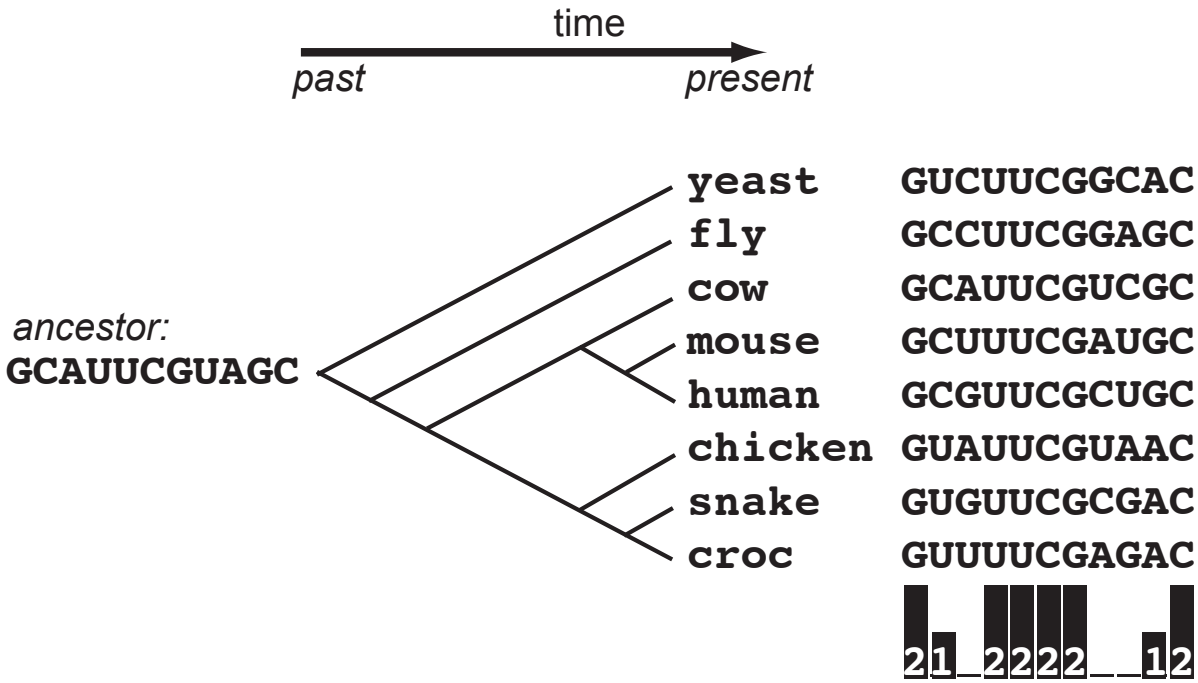
Infernal implements CMs: statistical models of an RNA family

Search sequence databases for (or align) homologs of an RNA family using *both* primary sequence conservation *and* secondary structure conservation

using covariance models (CMs) - profile stochastic context-free grammars



Amount of information in a profile can be measured in bits



Each column is independent

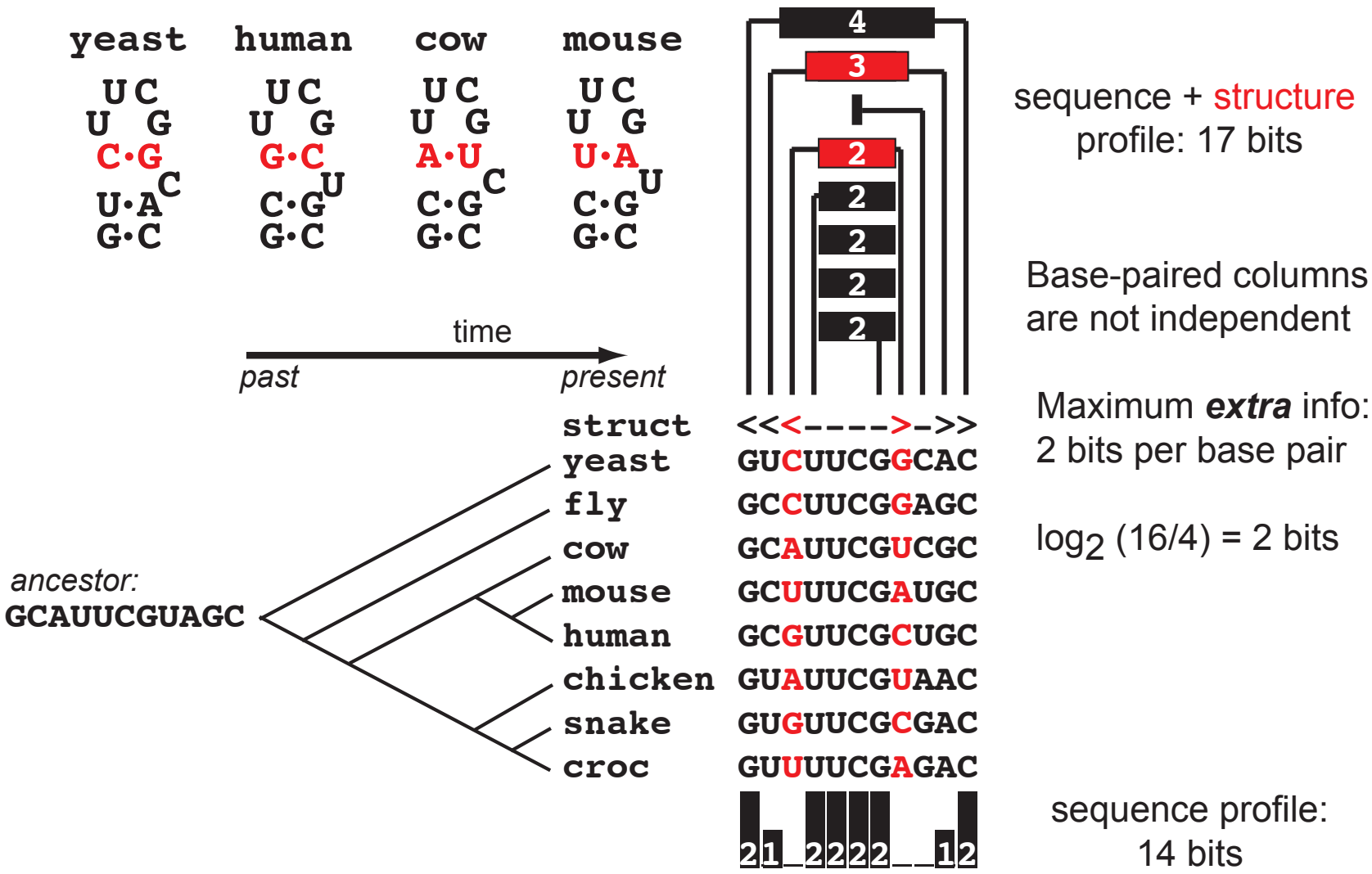
Maximum information is 2 bits per position, if column is 100% conserved.

$\log_2 (4/1) = 2$ bits

sequence profile: 14 bits

expect a match by chance: 1 in 2^{14} nt =~ 16 Kb

Structure contributes additional information from covariation



sequence + structure profile: 17 bits

Base-paired columns are not independent

Maximum ***extra*** info: 2 bits per base pair

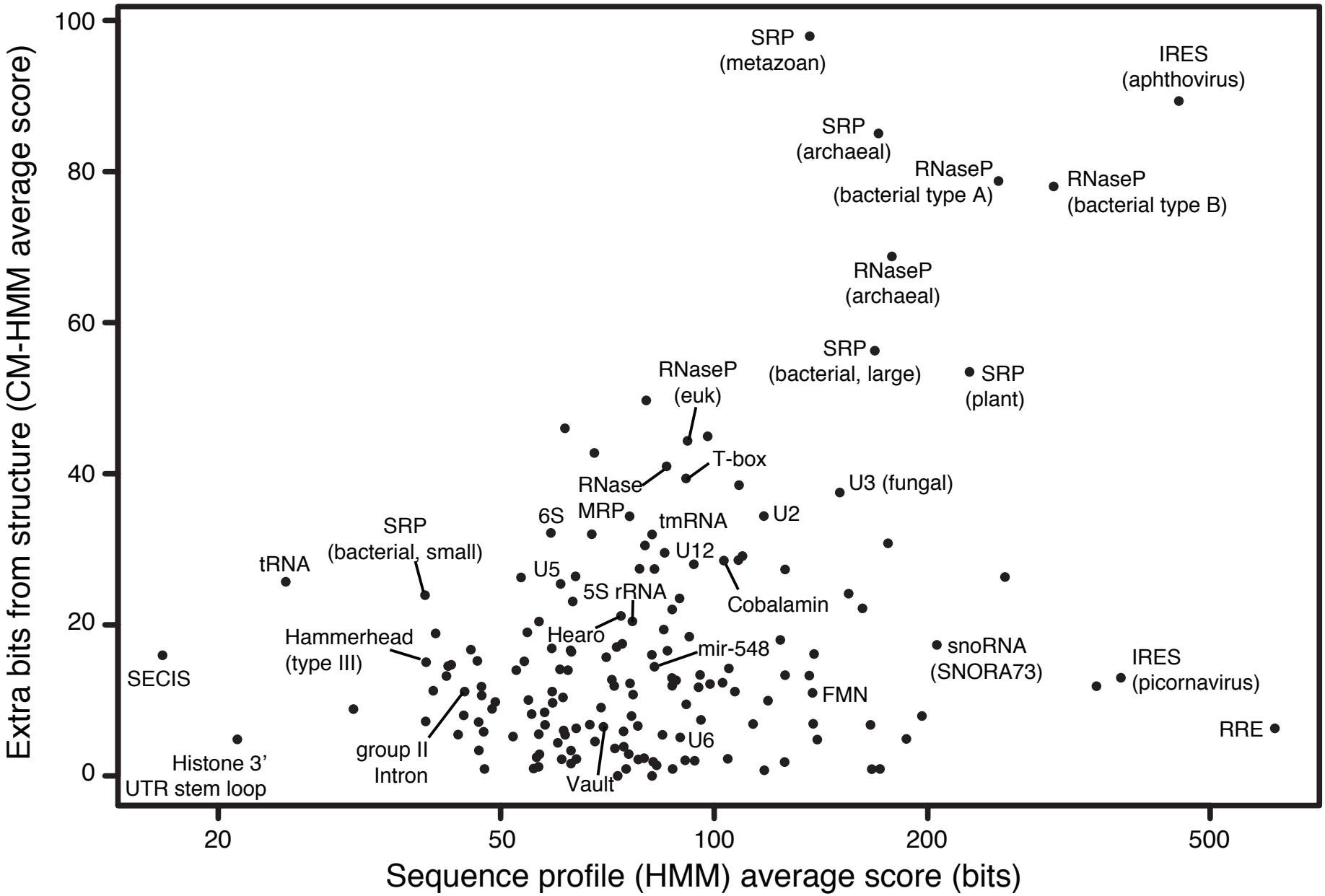
$\log_2 (16/4) = 2$ bits

sequence profile: 14 bits

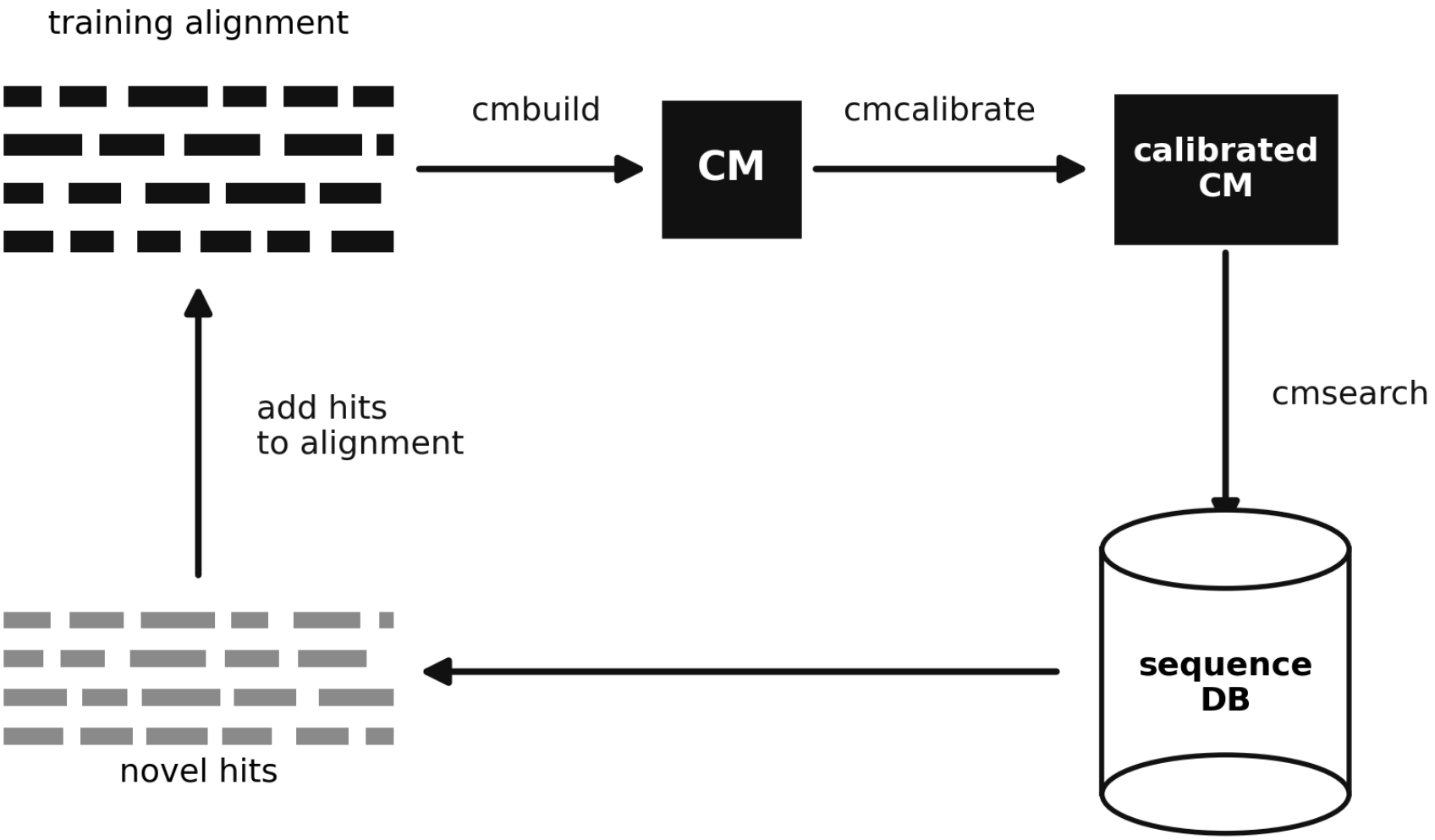
expect a match by chance: 1 in 2^{17} nt =~ 130 Kb

reducing expected false positives by $2^3 = 8$-fold

Levels of sequence and structure conservation in RNA families



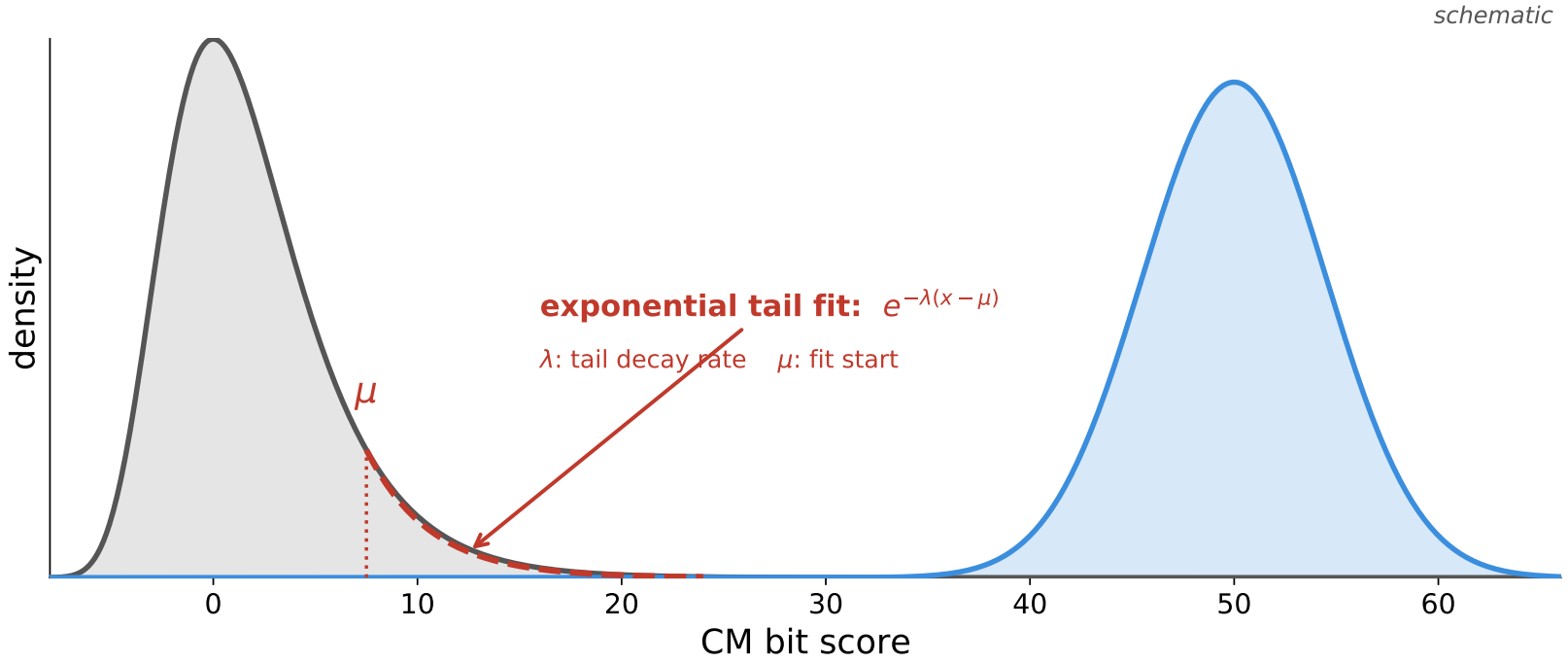
CM workflow


training alignment
cmbuild
CM
cmcalibrate
calibrated CM
cmsearch
add hits to alignment
sequence DB
novel hits

Calibration enables E-value reporting

- to assign *E-values* to search hits

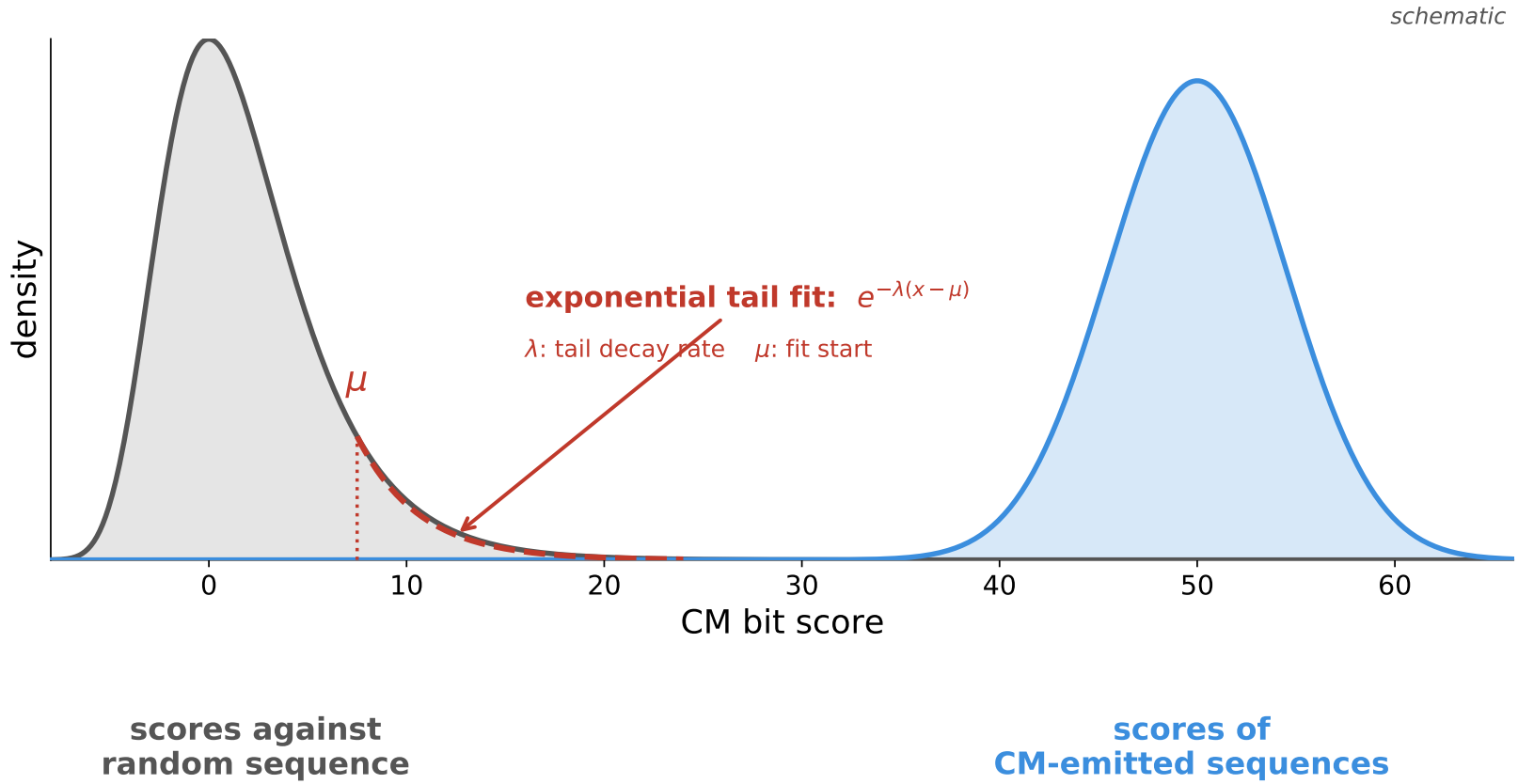
- E-value = expected number of hits at a given score in random, nonhomologous sequence



n

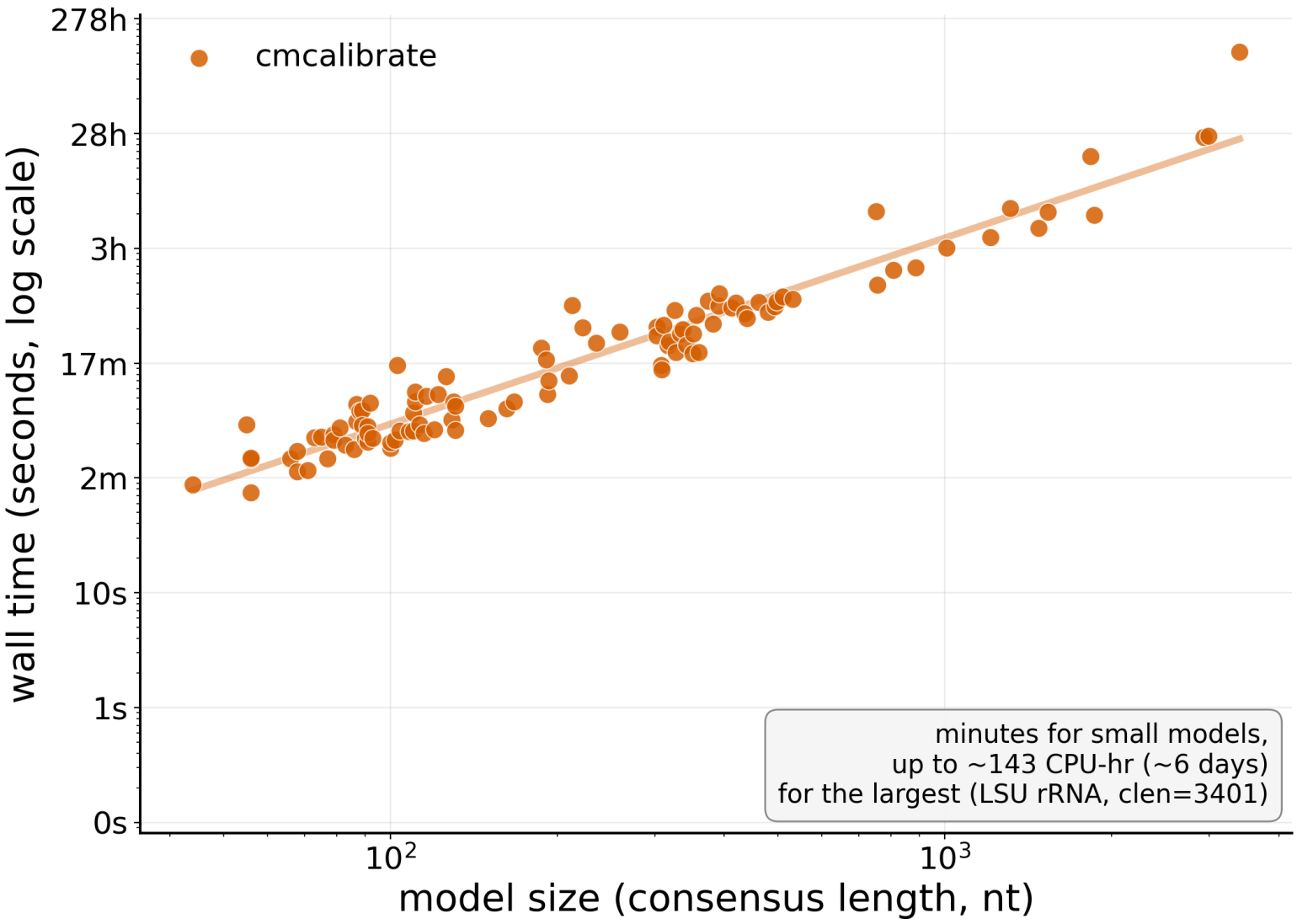
Calibration is slow

- to fit far tail's slope (λ) we need search a lot of random sequence
- CM search DP (CYK / Inside) scales as $O(LN^2 \log N)$,
  vs $O(LN)$ for a profile HMM (N = model length, L = target length)
- we speed up *cmsearch* with HMM filters and HMM bands* - but calibration can't use them: the filters remove nearly all random sequence

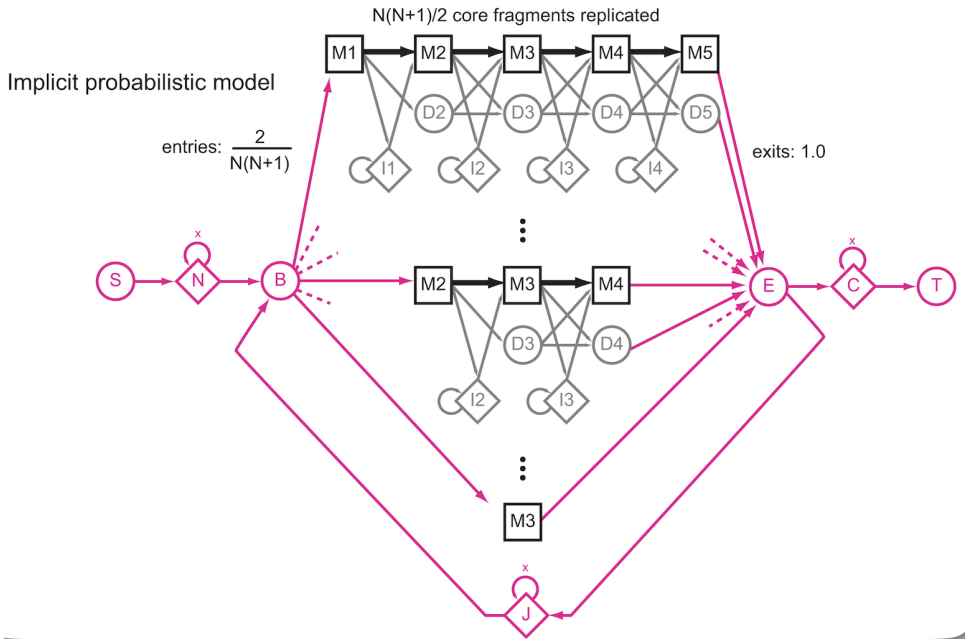


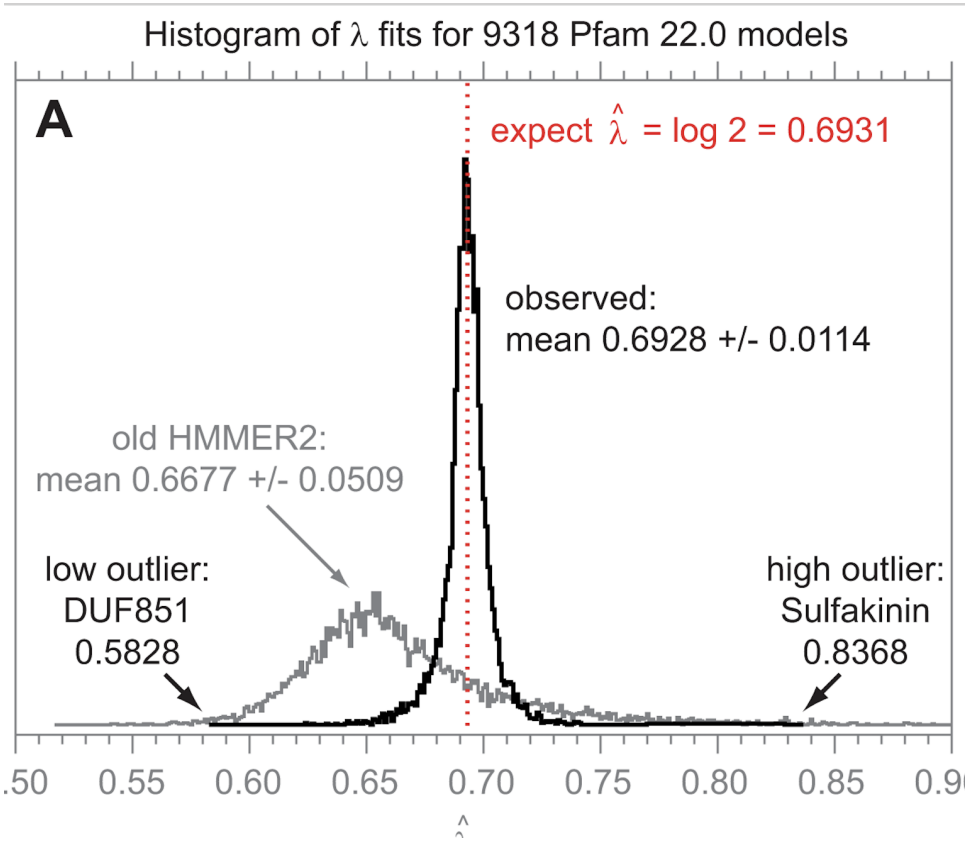
*E.P. Nawrocki & S.R. Eddy, "Infernal 1.1: 100-fold faster RNA homology searches," *Bioinformatics* **29**(22):2933–2935 (2013).

Calibration is slow

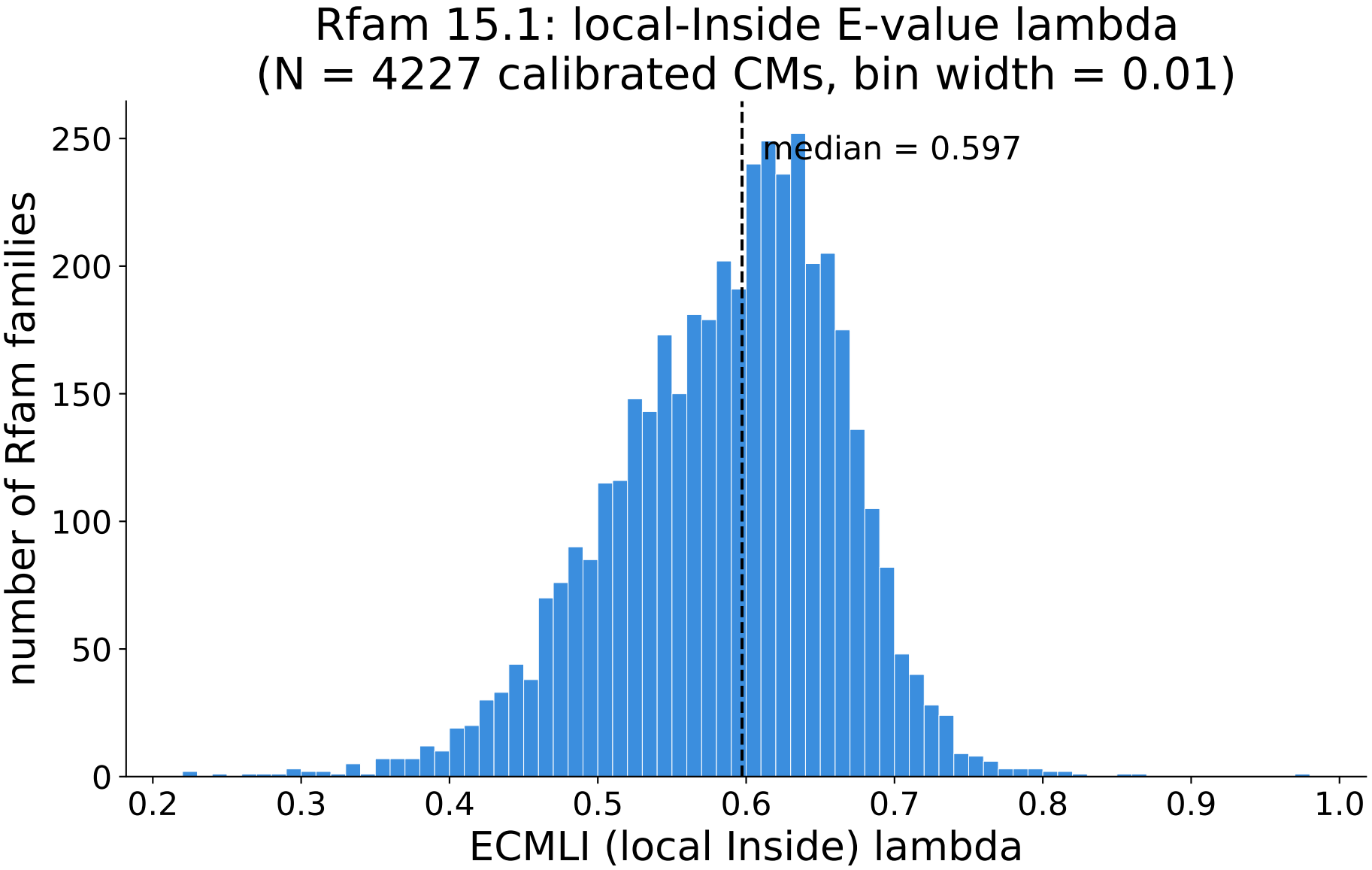


HMMER3 gets E-value statistics analytically



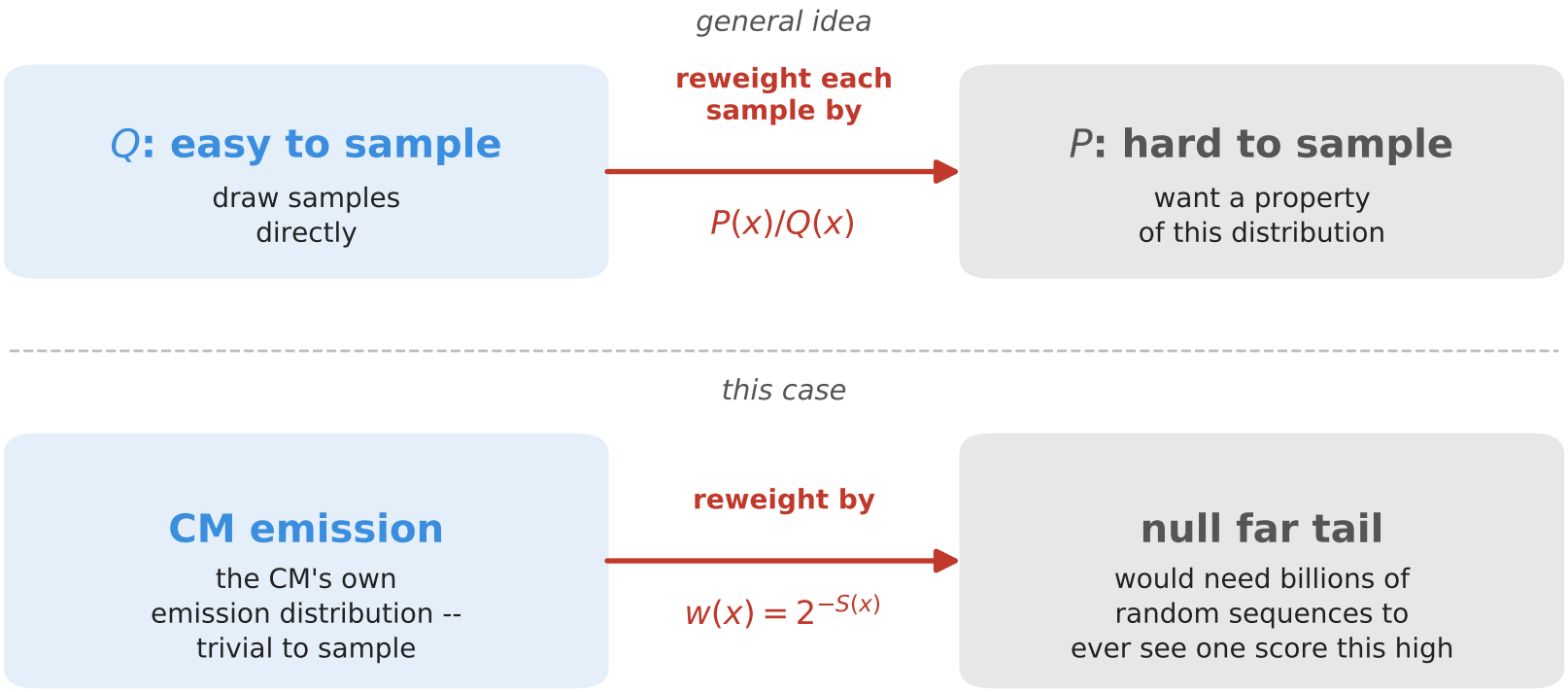


Eddy (2008), PLoS Comput. Biol. 4:e1000069


Rfam 15.1: local-Inside E-value lambda
(N = 4227 calibrated CMs, bin width = 0.01)
median = 0.597
number of Rfam families
0
50
100
150
200
250
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ECMLI (local Inside) lambda

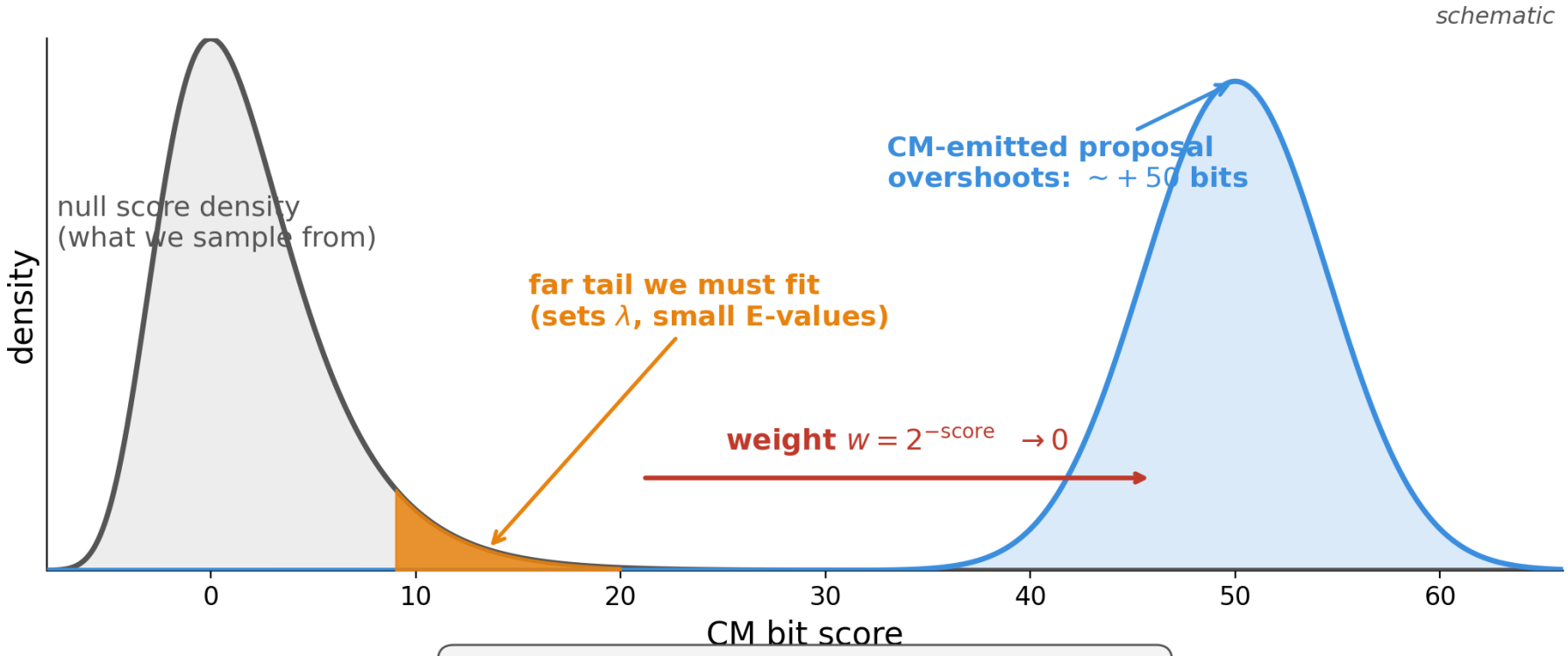
Importance sampling: sample the easy distribution, reweight to the hard one

- want a property of a distribution P that's hard to sample directly
- instead sample an easy Q, and reweight by $P(x)/Q(x)$



$S(x) = \log_2 \frac{P_{\text{CM}}(x)}{P_{\text{null}}(x)}$ (the CM bit score is itself the log-odds of emission vs. null)

Importance sampling fails: distributions too far apart



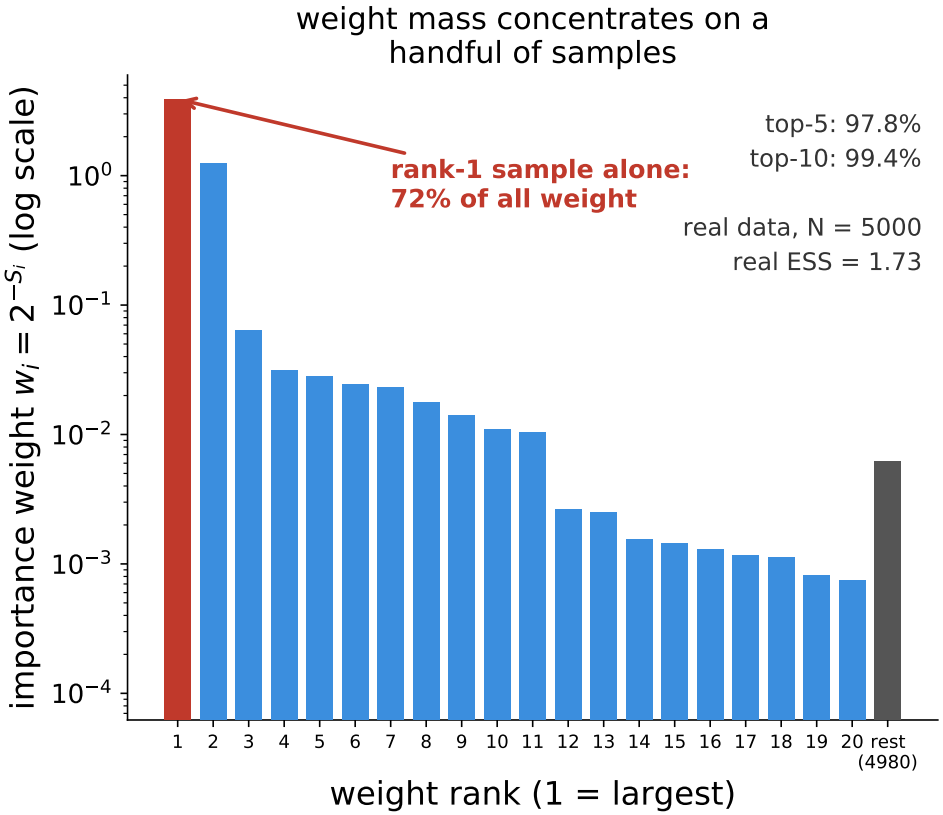
$$w(x) = \frac{P_{\text{null}}(x)}{q(x)} = 2^{-\text{score}} \qquad \text{ESS} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2} \approx 1$$

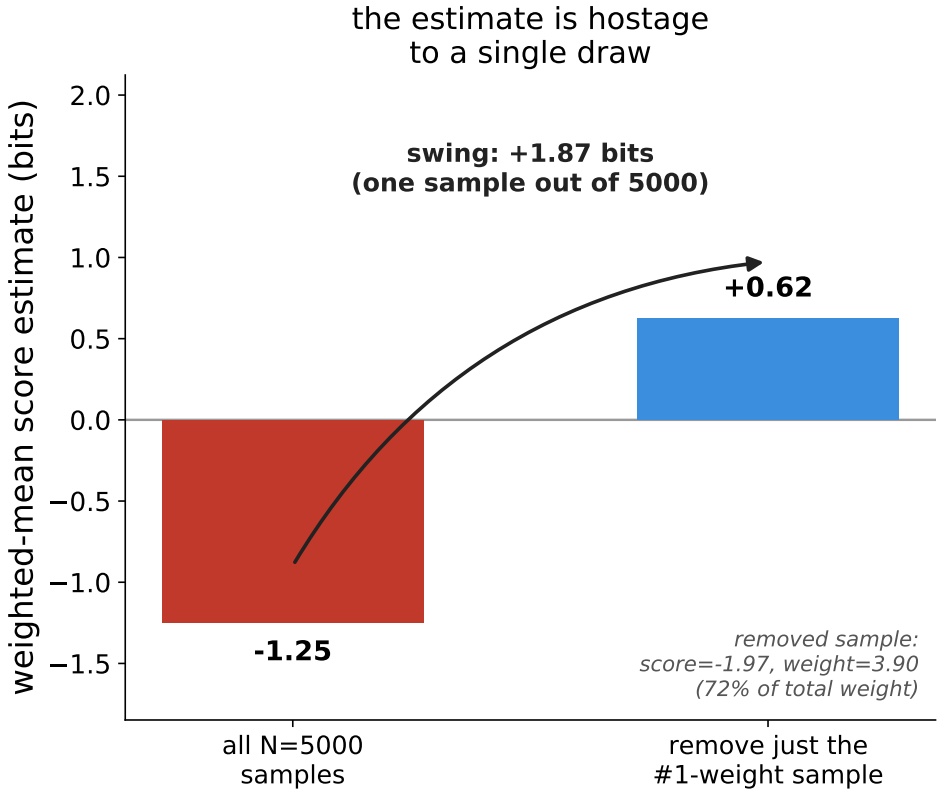
a few sequences carry all the weight ⇒ effective sample size ≈ 1 — worst for large / glocal models

Few samples hold the whole estimate hostage

- real N=5000 sequences emitted from the 5S rRNA CM (RF00001); weight $w_i = 2^{-S_i}$ per sample; real ESS = 1.73 out of N=5000

- the single largest-weight sample carries 72% of all weight – removing just that one sample swings the weighted-mean estimate by +1.87 bits





I tried to fix this various ways — none worked

attempted fix	**result**
flatten the CM's probabilities, to sample closer to random sequence	still collapses: ESS stays ≤ 2
reject any emitted sequence outside a target score window	biased the fitted slope 3$\times$ (0.14 vs. true 0.42)
design/mutate sequences to force them into the tail	fitted slope always converges to the same fixed value, regardless of the true one
sample directly from the tail with MCMC (no reweighting)	accurate, but cost grows quadratically with model size – slower than what it replaces for exactly the largest models

Skip sampling and predict the E-value parameters (λ, μ) directly

- λ, μ aren't random, they're a function of the CM (its composition, structure, and length). So they should be predictable from model properties.

Skip sampling and predict the E-value parameters (λ, μ) directly

- λ, μ aren't random, they're a function of the CM (its composition, structure, and length). So they should be predictable from model properties.

- Thousands of already-calibrated Rfam CMs give us a free training/testing set: (CM $\rightarrow$ λ, μ) data points.

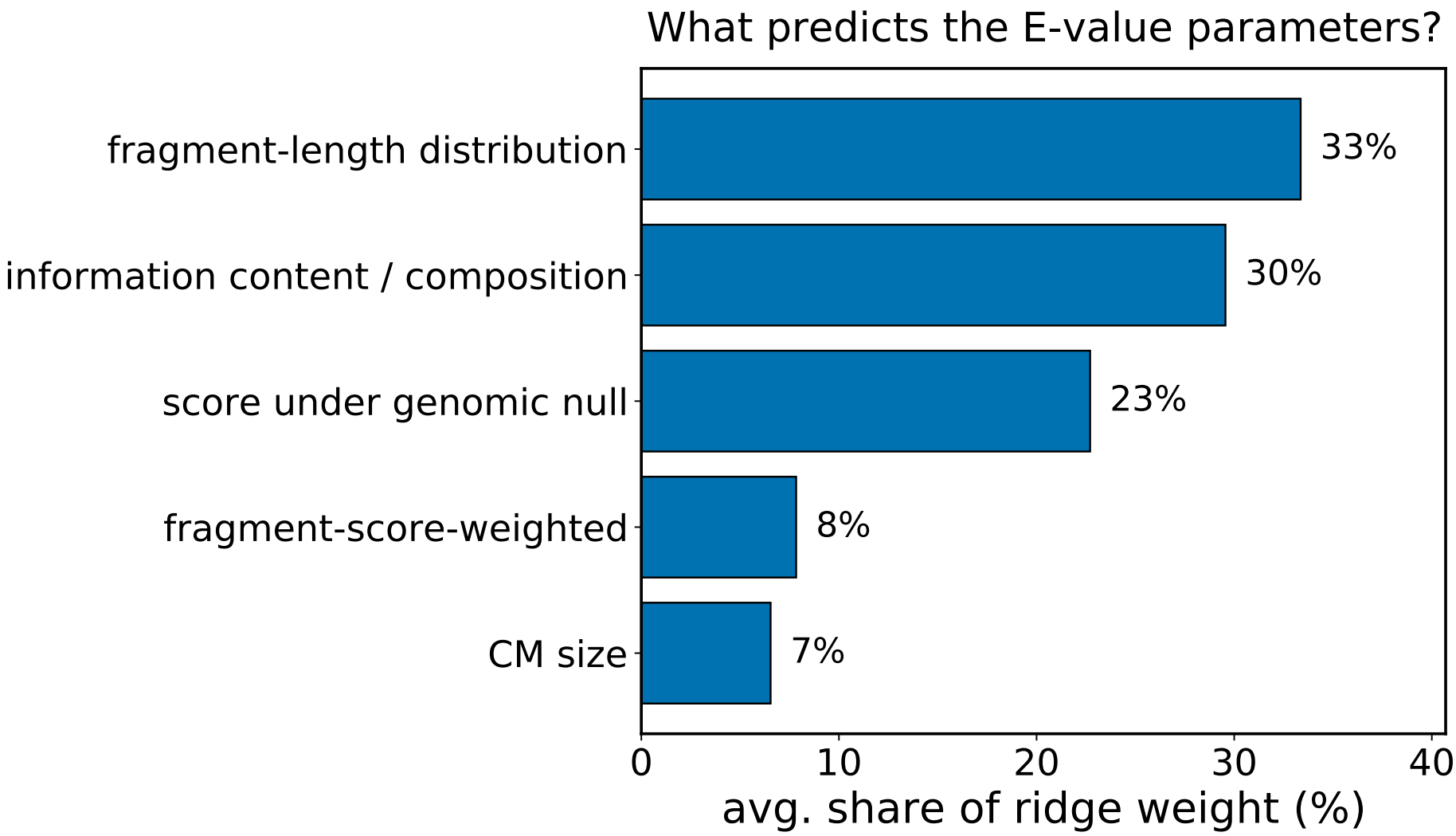
Skip sampling and predict the E-value parameters (λ, μ) directly

- λ, μ aren't random, they're a function of the CM (its composition, structure, and length). So they should be predictable from model properties.

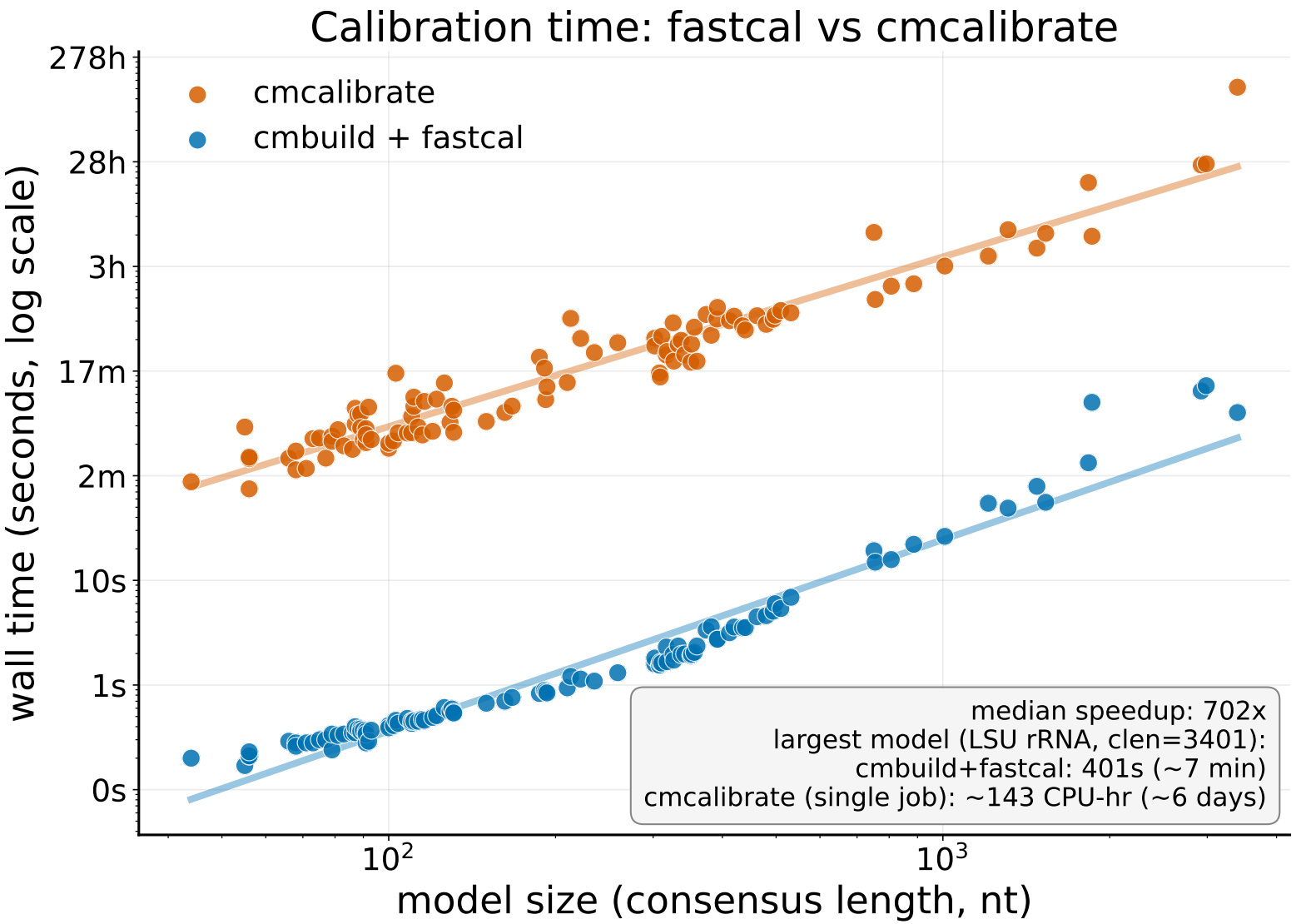
- Thousands of already-calibrated Rfam CMs give us a free training/testing set: (CM $\rightarrow$ λ, μ) data points.

- Ridge regression: linear, L2-regularized (robust with many correlated features, hard to overfit), and interpretable (you can read off what predicts what)

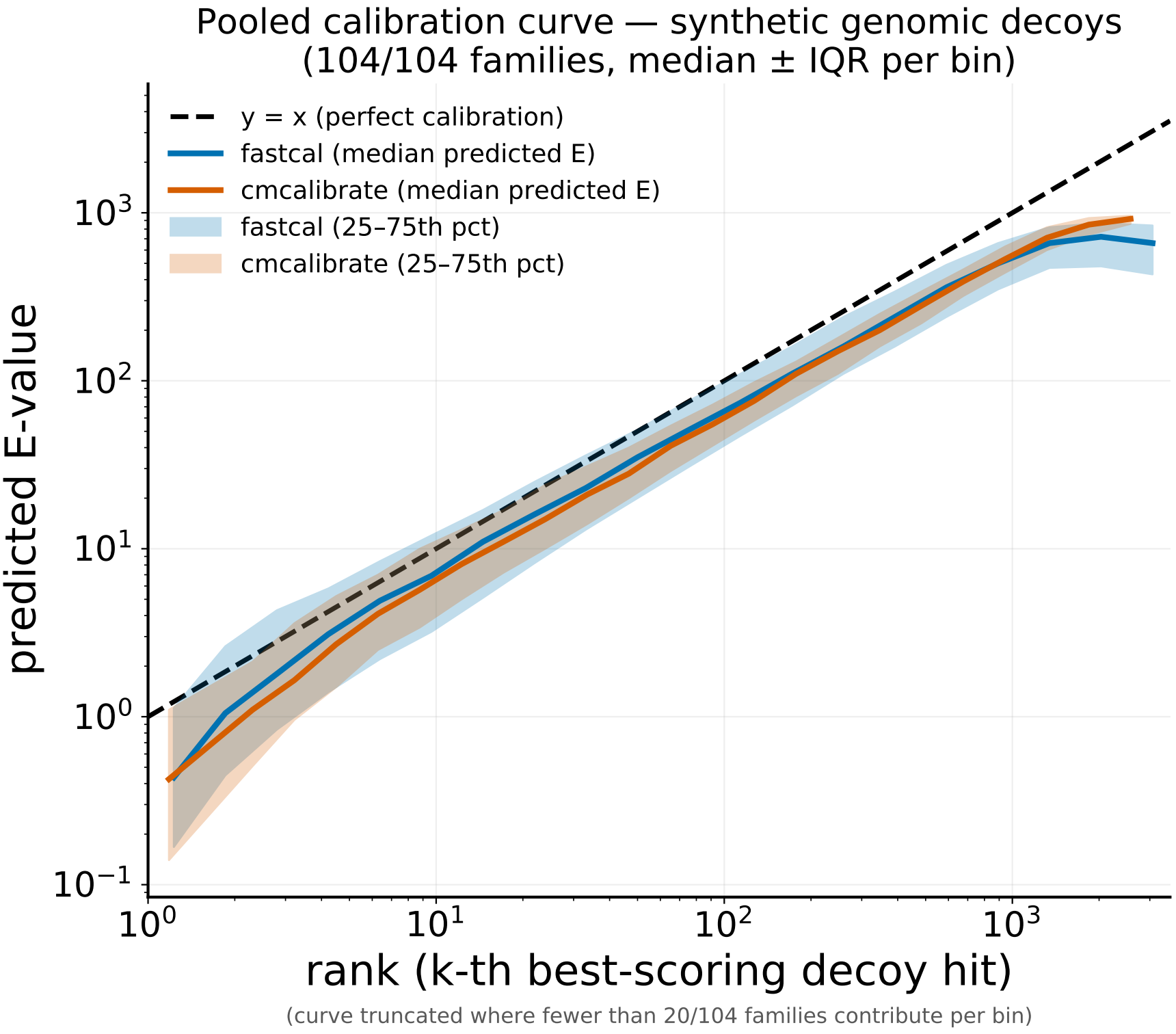
The most important CM features

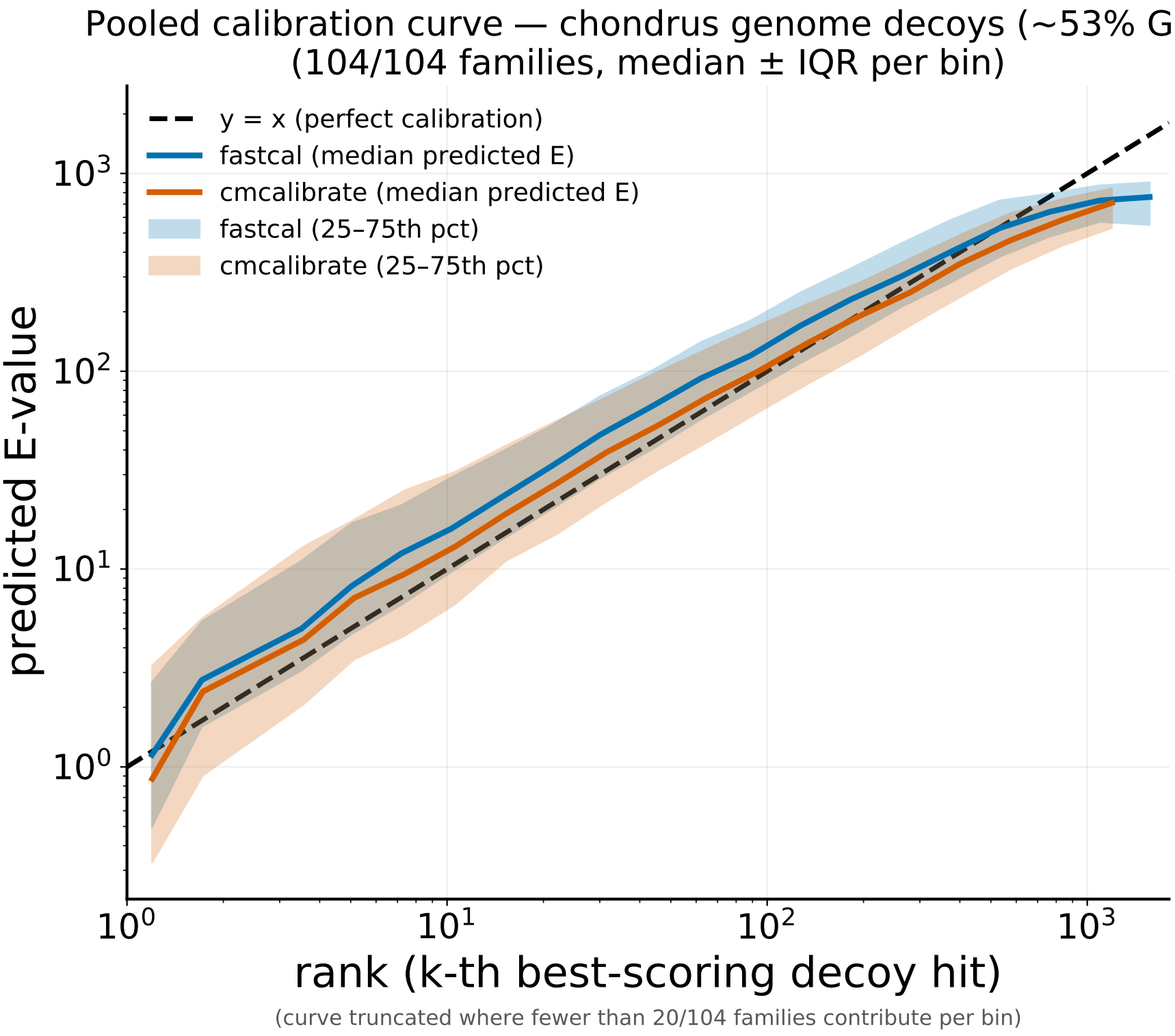


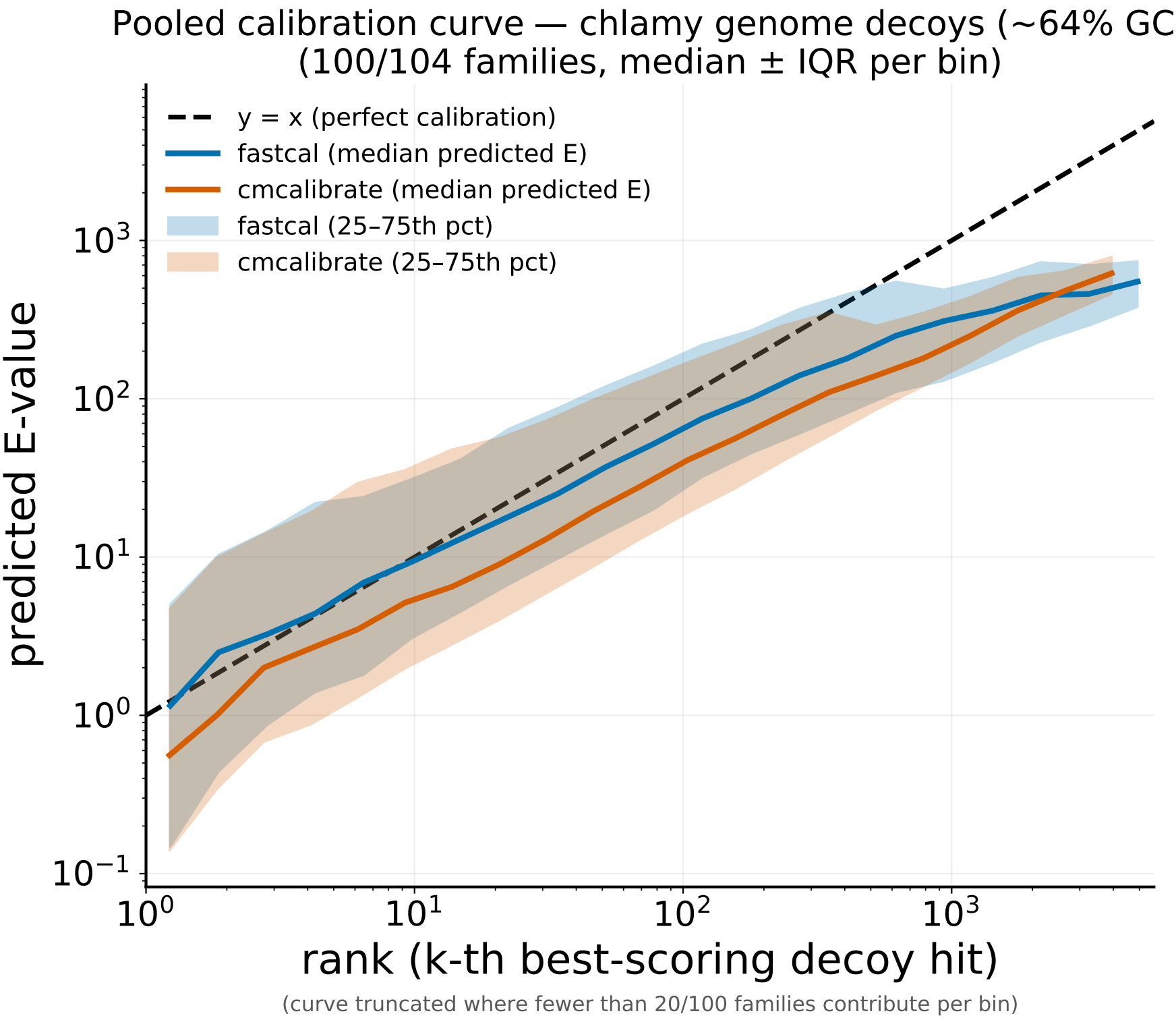
Fast

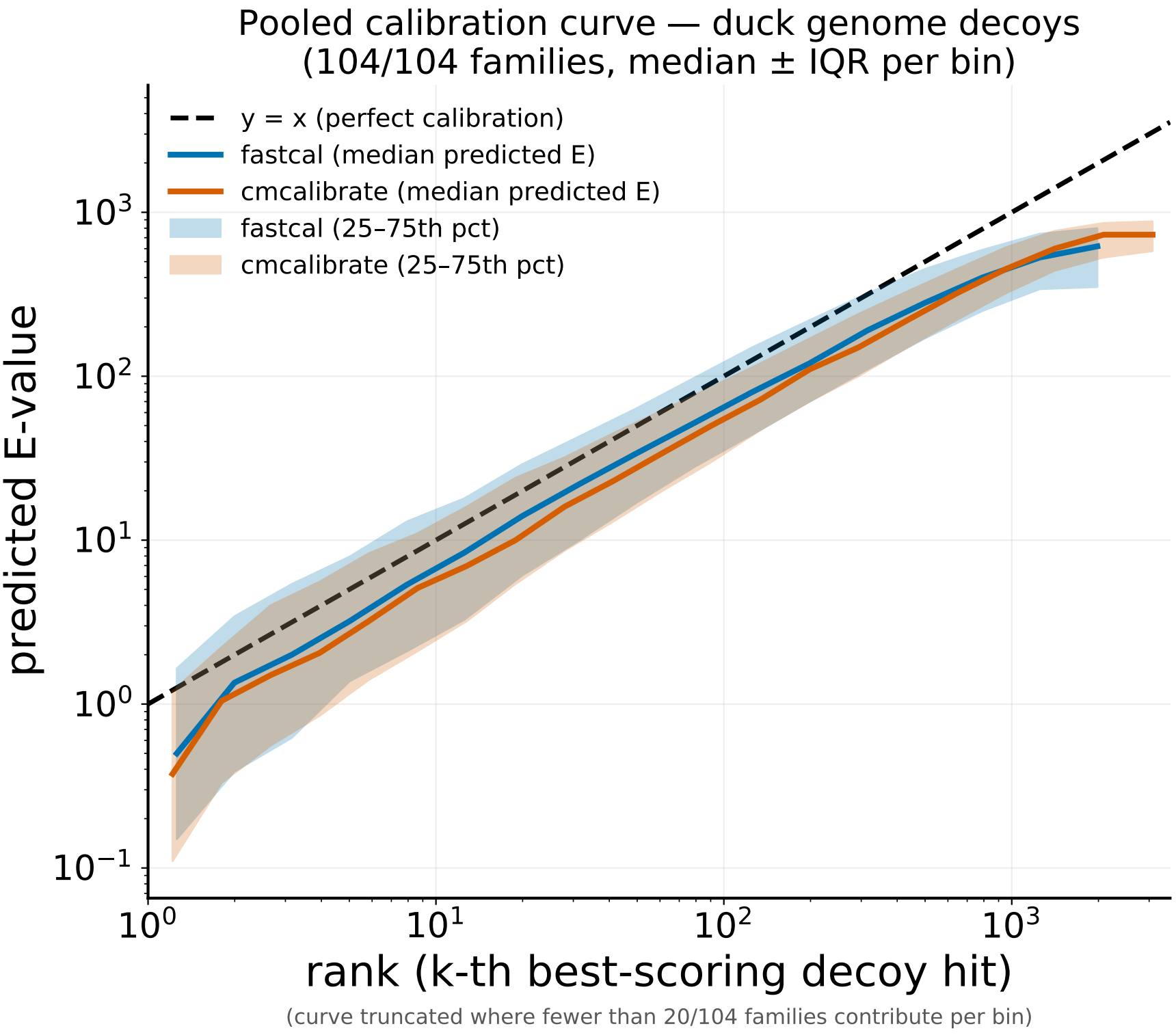


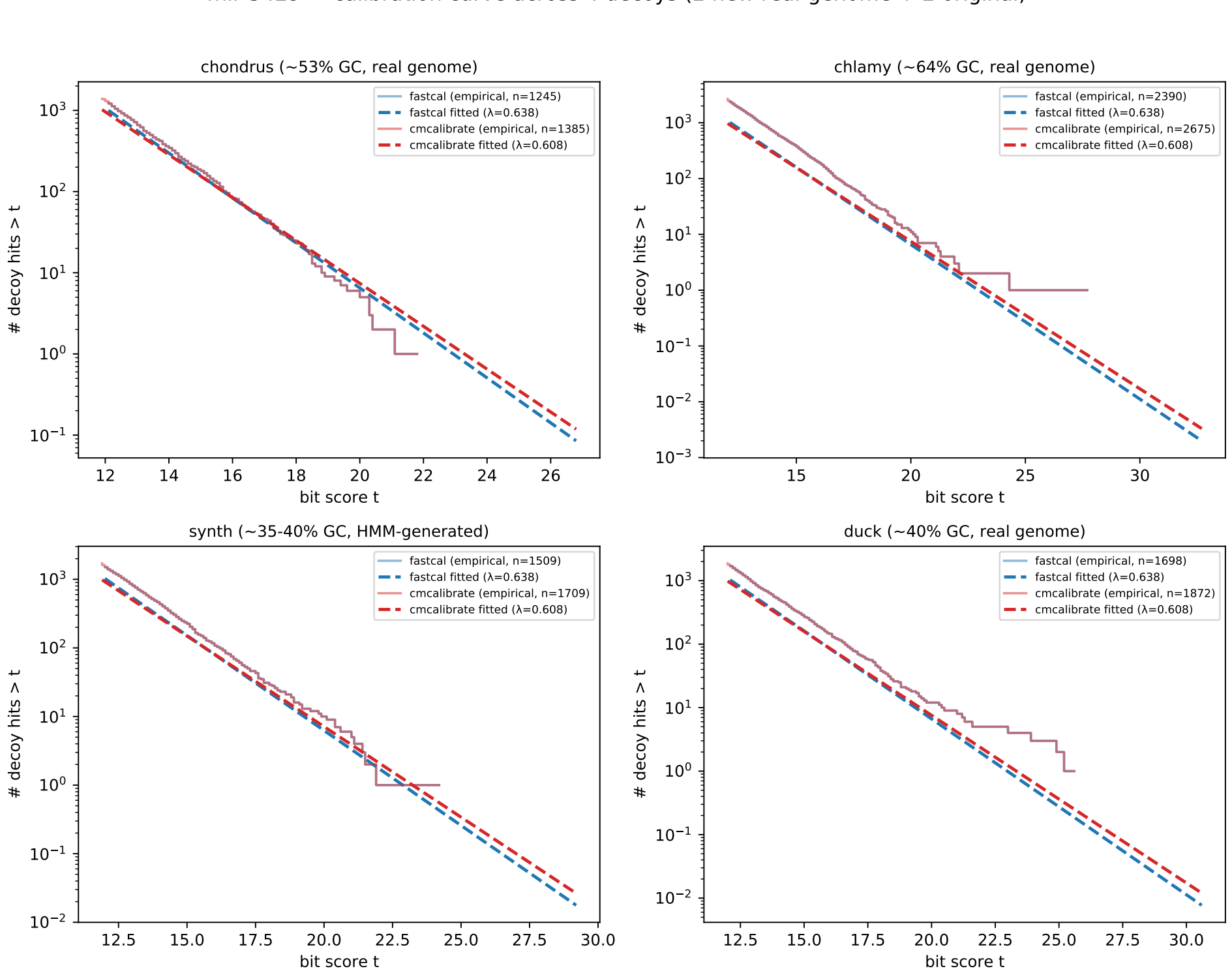
104 models, clen 44–3401 — **median speedup** ~**702**× (range 39–2634×), conservative in fastcal's favor (cmcalibrate ran `--cpu 4`, fastcal ran single-threaded)

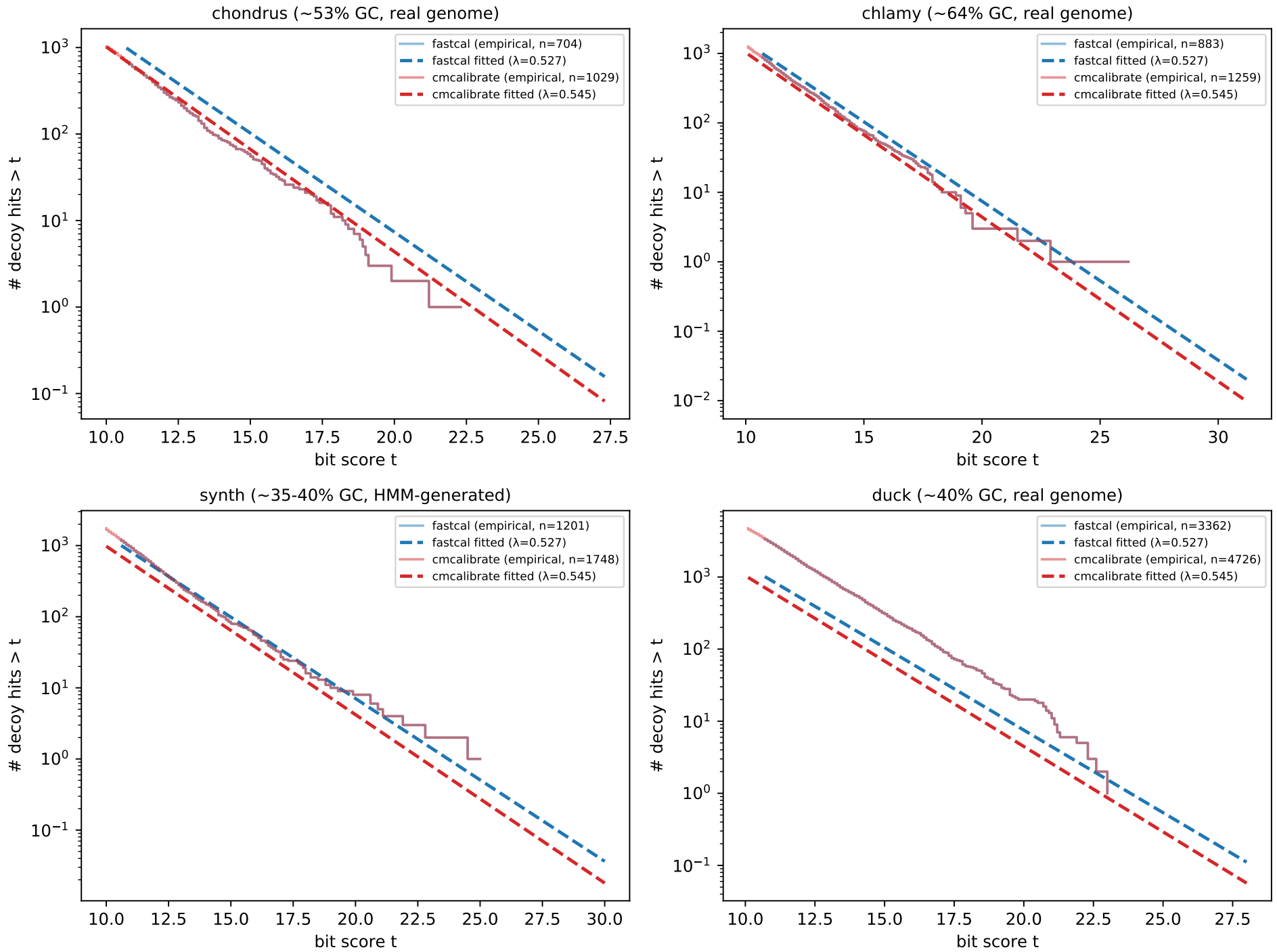

Pooled calibration curve — synthetic genomic decoys
(104/104 families, median ± IQR per bin)
y = x (perfect calibration)
fastcal (median predicted E)
cmcalibrate (median predicted E)
fastcal (25–75th pct)
cmcalibrate (25–75th pct)
predicted E-value
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
rank (k-th best-scoring decoy hit)
(curve truncated where fewer than 20/104 families contribute per bin)

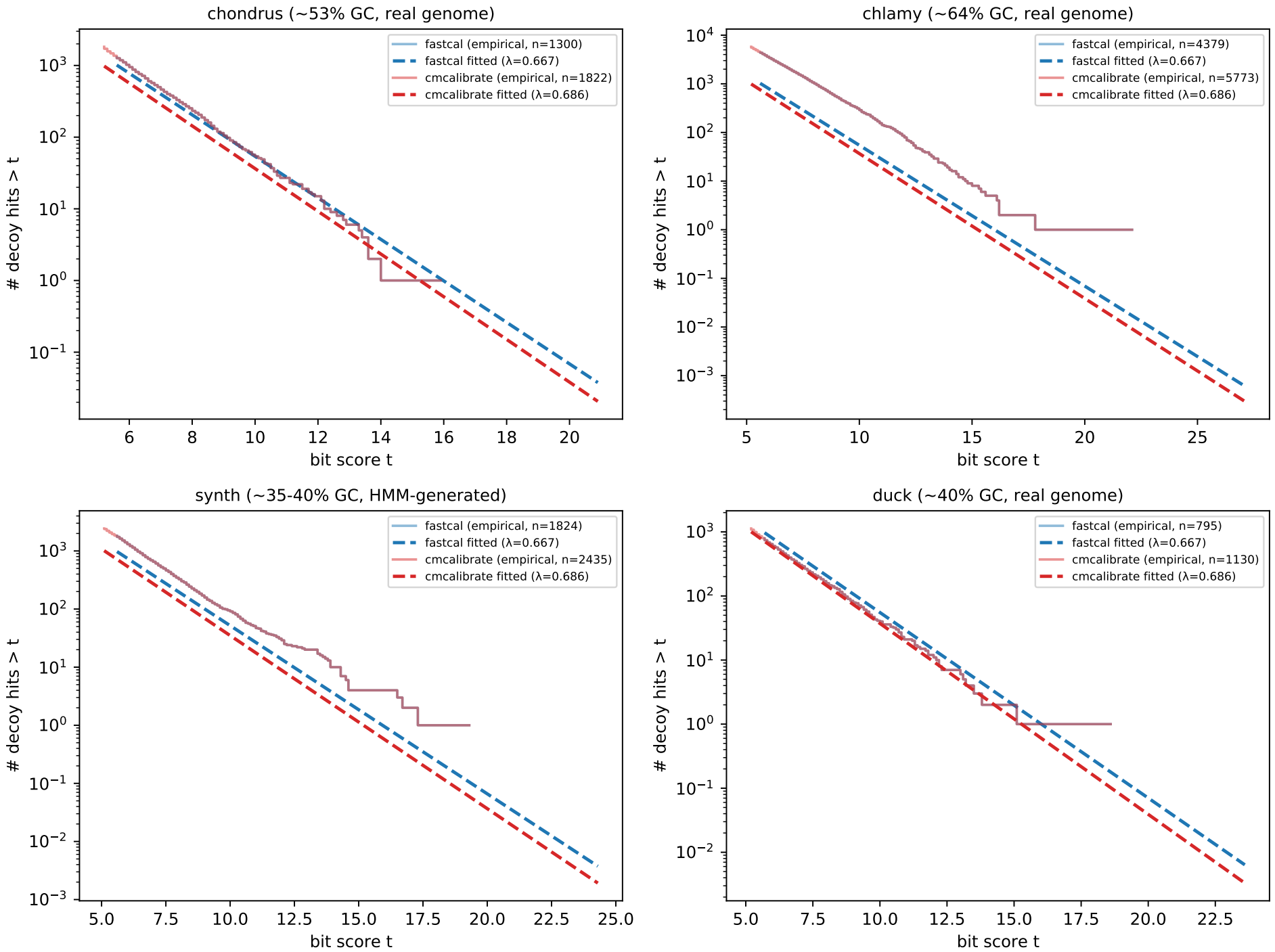

Pooled calibration curve — chondrus genome decoys (~53% GC)
(104/104 families, median ± IQR per bin)
y = x (perfect calibration)
fastcal (median predicted E)
cmcalibrate (median predicted E)
fastcal (25–75th pct)
cmcalibrate (25–75th pct)
predicted E-value
10^0
10^1
10^2
10^3
rank (k-th best-scoring decoy hit)
(curve truncated where fewer than 20/104 families contribute per bin)


Pooled calibration curve — chlamy genome decoys (~64% GC)
(100/104 families, median ± IQR per bin)
y = x (perfect calibration)
fastcal (median predicted E)
cmcalibrate (median predicted E)
fastcal (25–75th pct)
cmcalibrate (25–75th pct)
predicted E-value
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
rank (k-th best-scoring decoy hit)
(curve truncated where fewer than 20/100 families contribute per bin)

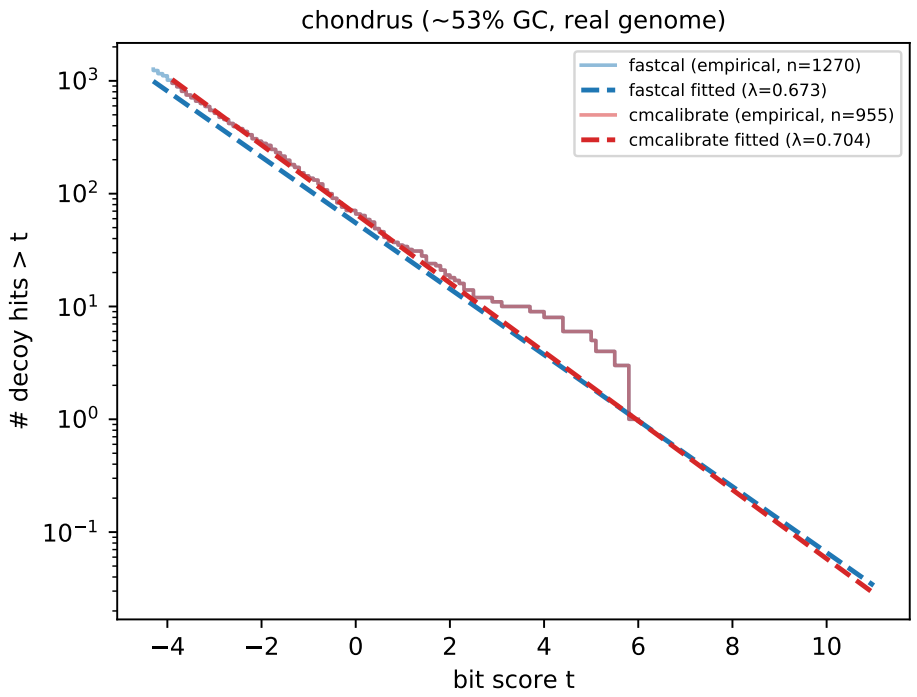

Pooled calibration curve — duck genome decoys
(104/104 families, median ± IQR per bin)
y = x (perfect calibration)
fastcal (median predicted E)
cmcalibrate (median predicted E)
fastcal (25–75th pct)
cmcalibrate (25–75th pct)
predicted E-value
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
rank (k-th best-scoring decoy hit)
(curve truncated where fewer than 20/104 families contribute per bin)

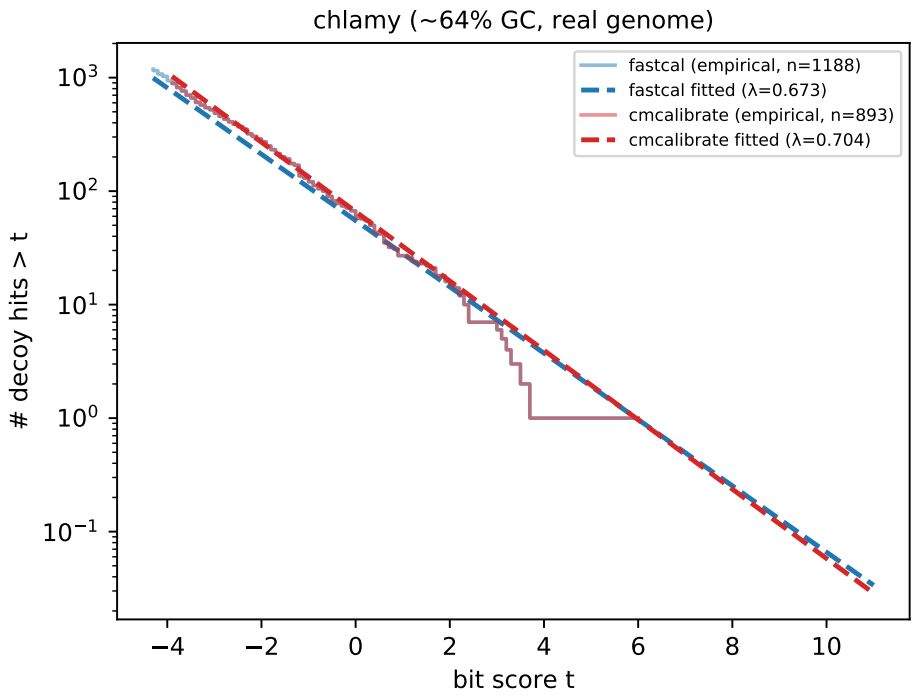

mir-5429 — calibration curve across 4 decoys (2 new real-genome + 2 original)
chondrus (~53% GC, real genome)
fastcal (empirical, n=1245)
fastcal fitted (λ=0.638)
cmcalibrate (empirical, n=1385)
cmcalibrate fitted (λ=0.608)
decoy hits > t
bit score t
chlamy (~64% GC, real genome)
fastcal (empirical, n=2390)
fastcal fitted (λ=0.638)
cmcalibrate (empirical, n=2675)
cmcalibrate fitted (λ=0.608)
decoy hits > t
bit score t
synth (~35-40% GC, HMM-generated)
fastcal (empirical, n=1509)
fastcal fitted (λ=0.638)
cmcalibrate (empirical, n=1709)
cmcalibrate fitted (λ=0.608)
decoy hits > t
bit score t
duck (~40% GC, real genome)
fastcal (empirical, n=1698)
fastcal fitted (λ=0.638)
cmcalibrate (empirical, n=1872)
cmcalibrate fitted (λ=0.608)
decoy hits > t
bit score t

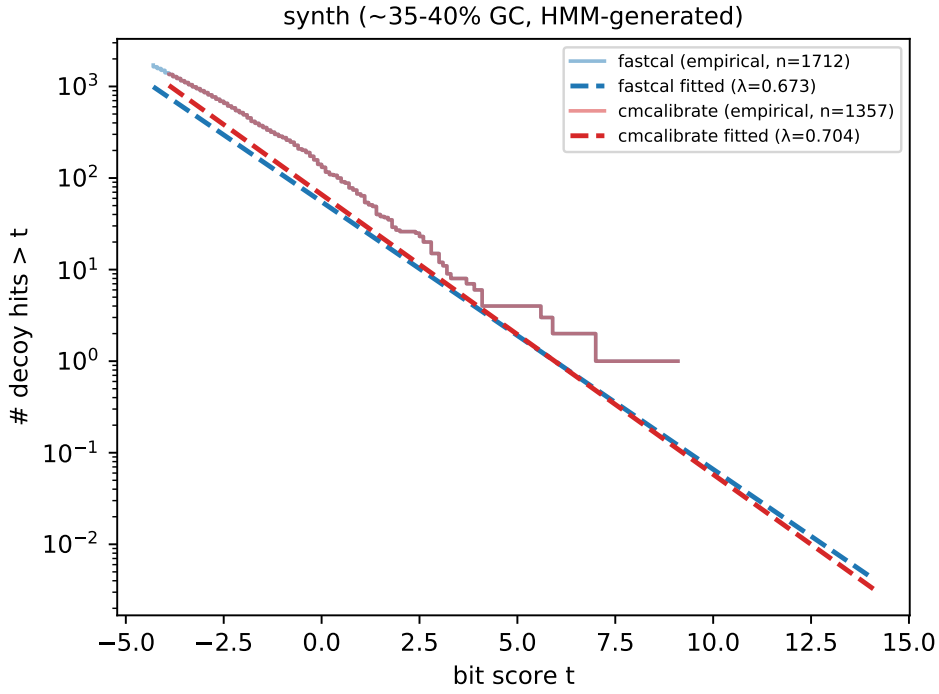

SNORA30 — calibration curve across 4 decoys (2 new real-genome + 2 original)
chondrus (~53% GC, real genome)
fastcal (empirical, n=704)
fastcal fitted (λ=0.527)
cmcalibrate (empirical, n=1029)
cmcalibrate fitted (λ=0.545)
chlamy (~64% GC, real genome)
fastcal (empirical, n=883)
fastcal fitted (λ=0.527)
cmcalibrate (empirical, n=1259)
cmcalibrate fitted (λ=0.545)
synth (~35-40% GC, HMM-generated)
fastcal (empirical, n=1201)
fastcal fitted (λ=0.527)
cmcalibrate (empirical, n=1748)
cmcalibrate fitted (λ=0.545)
duck (~40% GC, real genome)
fastcal (empirical, n=3362)
fastcal fitted (λ=0.527)
cmcalibrate (empirical, n=4726)
cmcalibrate fitted (λ=0.545)
decoy hits > t
bit score t

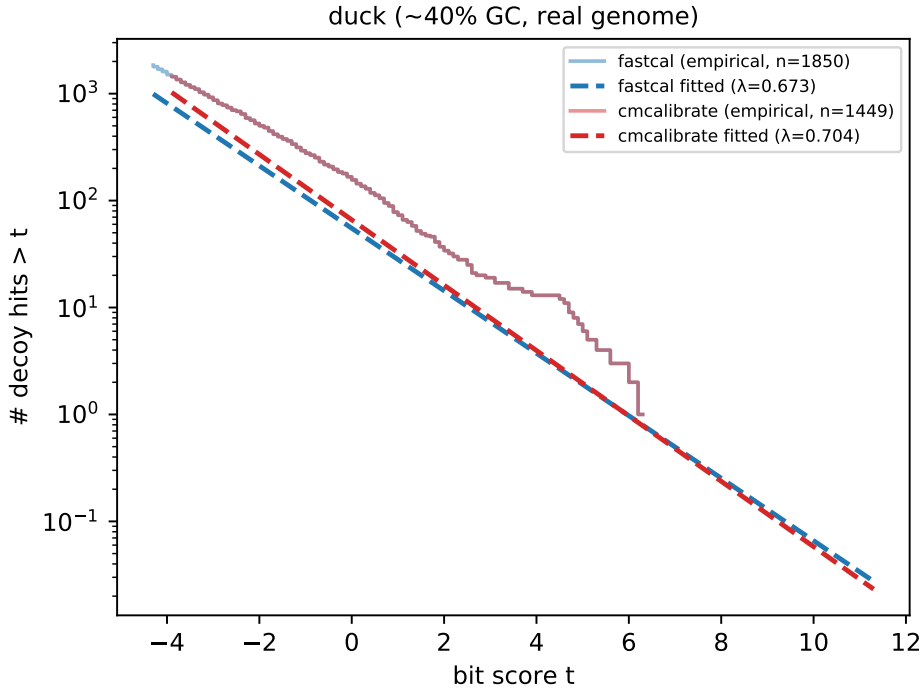

IRES_c-myc — calibration curve across 4 decoys (2 new real-genome + 2 original)
chondrus (~53% GC, real genome)
fastcal (empirical, n=1300)
fastcal fitted (λ=0.667)
cmcalibrate (empirical, n=1822)
cmcalibrate fitted (λ=0.686)
decoy hits > t
bit score t
chlamy (~64% GC, real genome)
fastcal (empirical, n=4379)
fastcal fitted (λ=0.667)
cmcalibrate (empirical, n=5773)
cmcalibrate fitted (λ=0.686)
decoy hits > t
bit score t
synth (~35-40% GC, HMM-generated)
fastcal (empirical, n=1824)
fastcal fitted (λ=0.667)
cmcalibrate (empirical, n=2435)
cmcalibrate fitted (λ=0.686)
decoy hits > t
bit score t
duck (~40% GC, real genome)
fastcal (empirical, n=795)
fastcal fitted (λ=0.667)
cmcalibrate (empirical, n=1130)
cmcalibrate fitted (λ=0.686)
decoy hits > t
bit score t

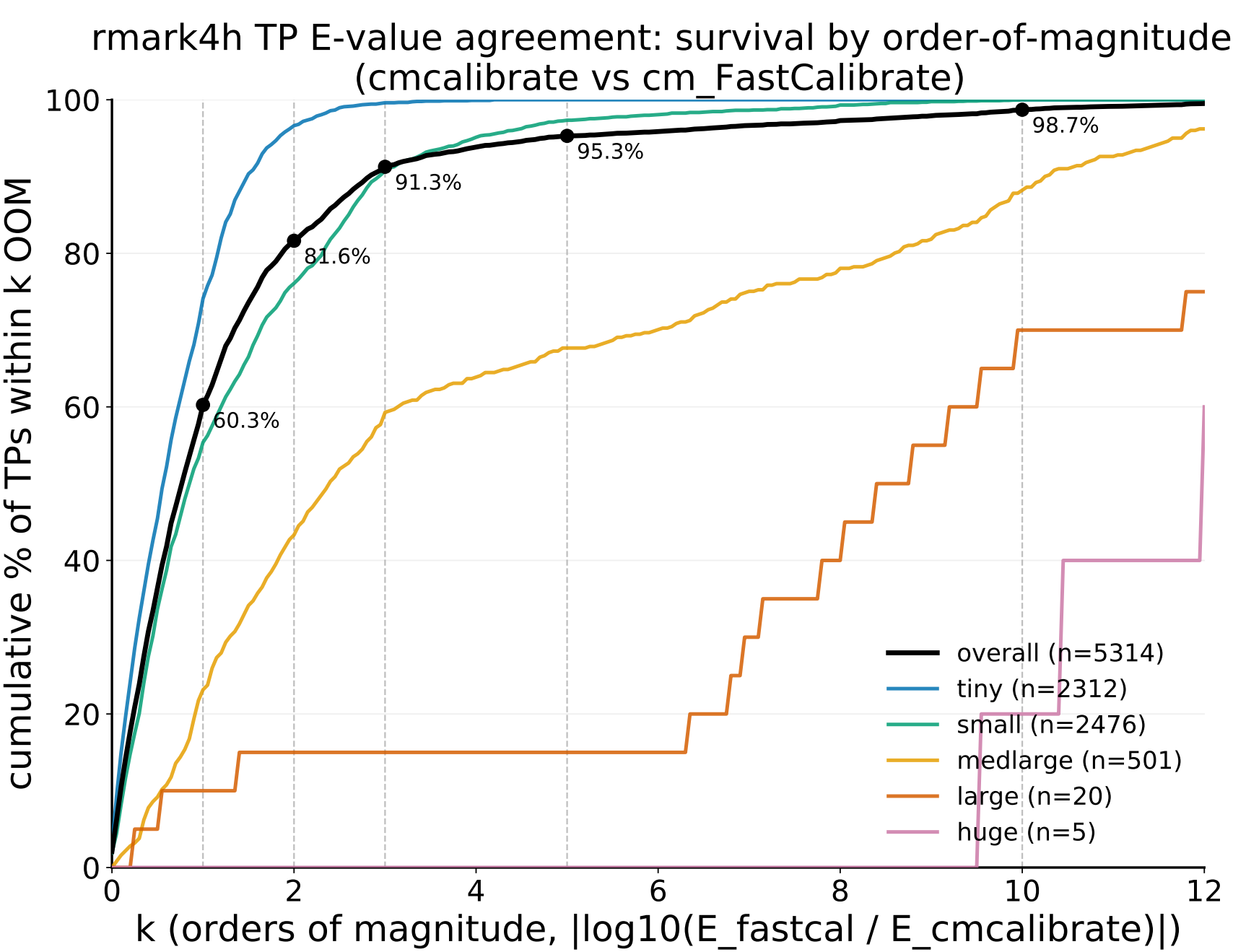
SSU_rRNA_eukarya — calibration curve across 4 decoys (2 new real-genome + 2 original)



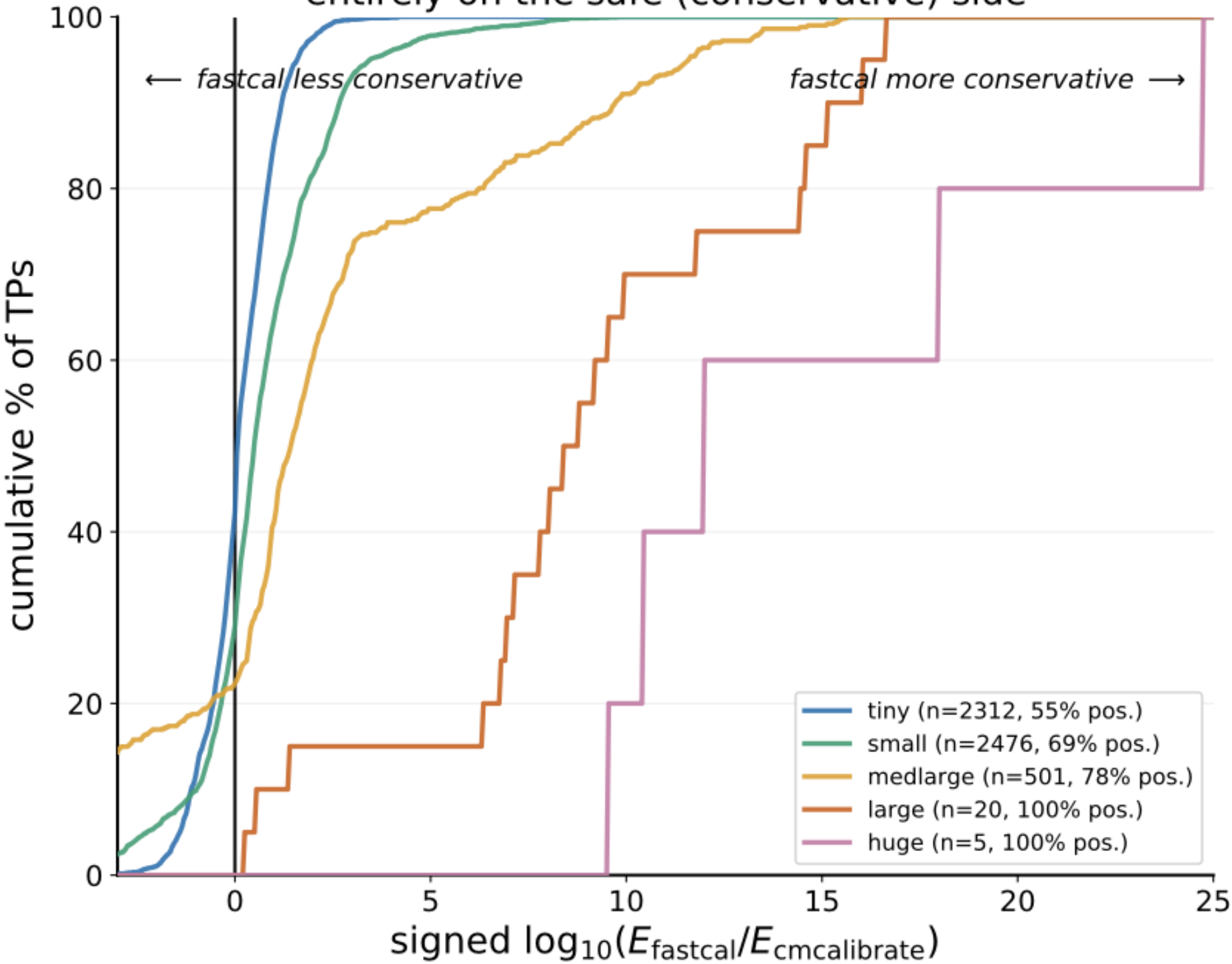


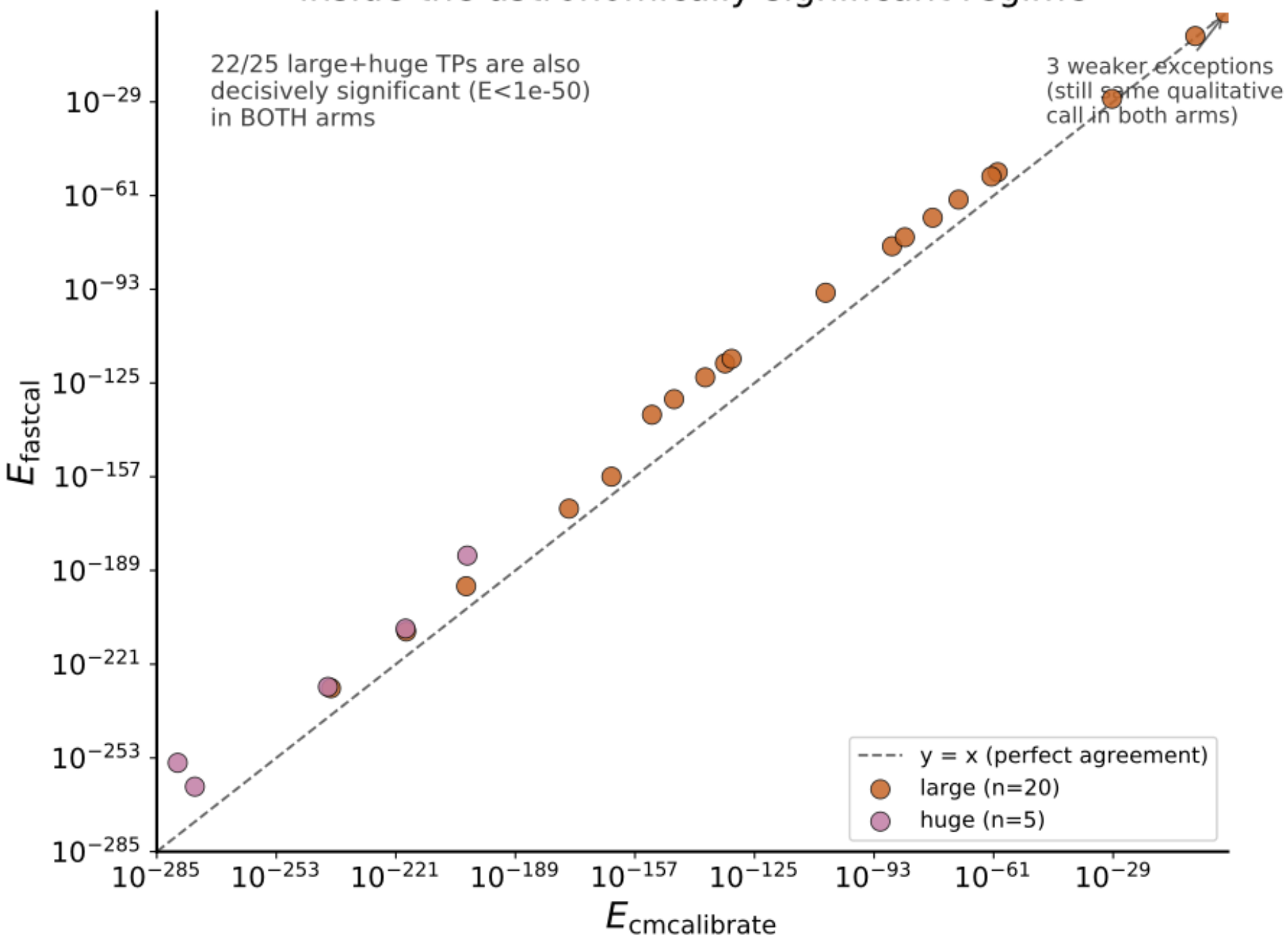


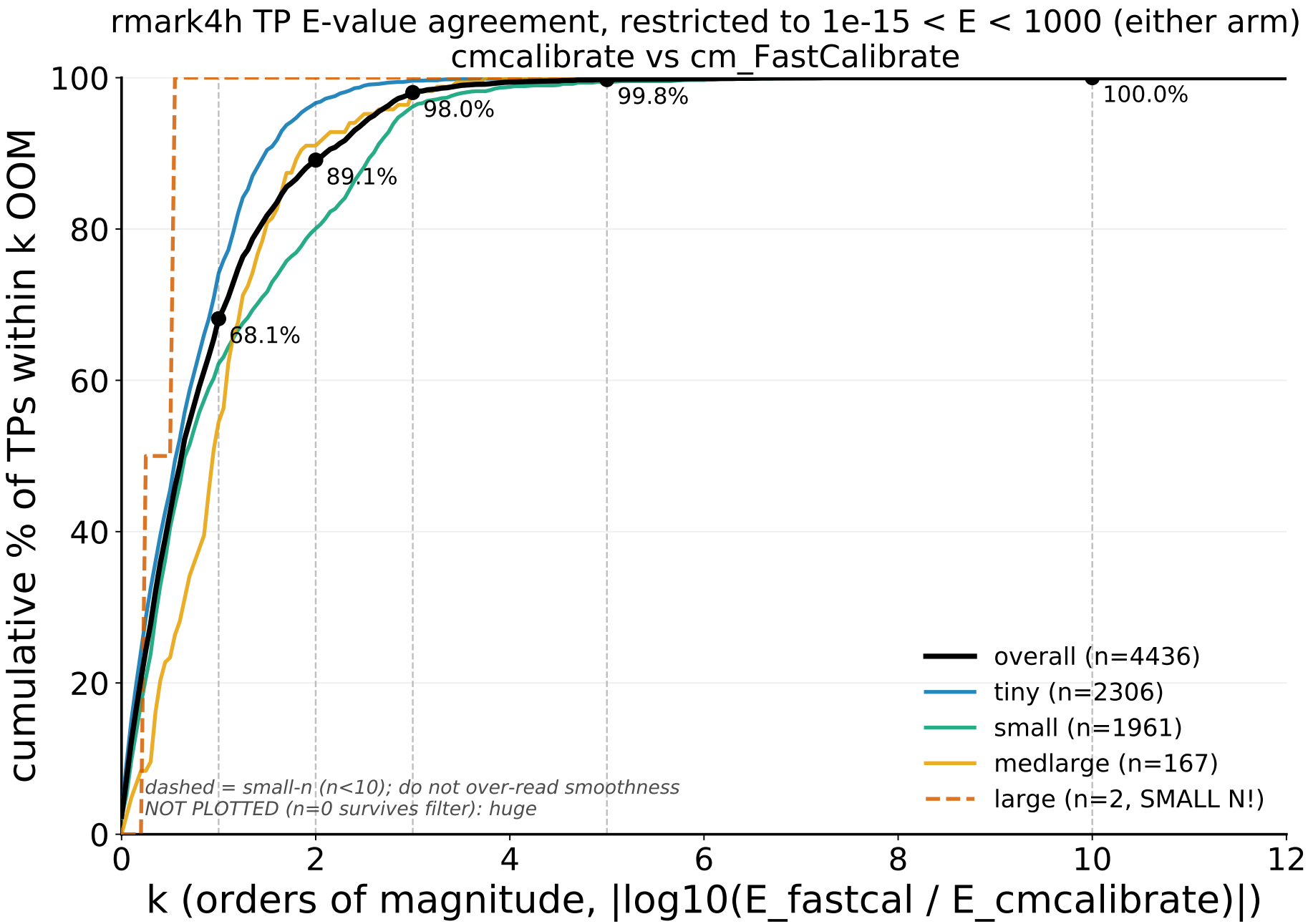


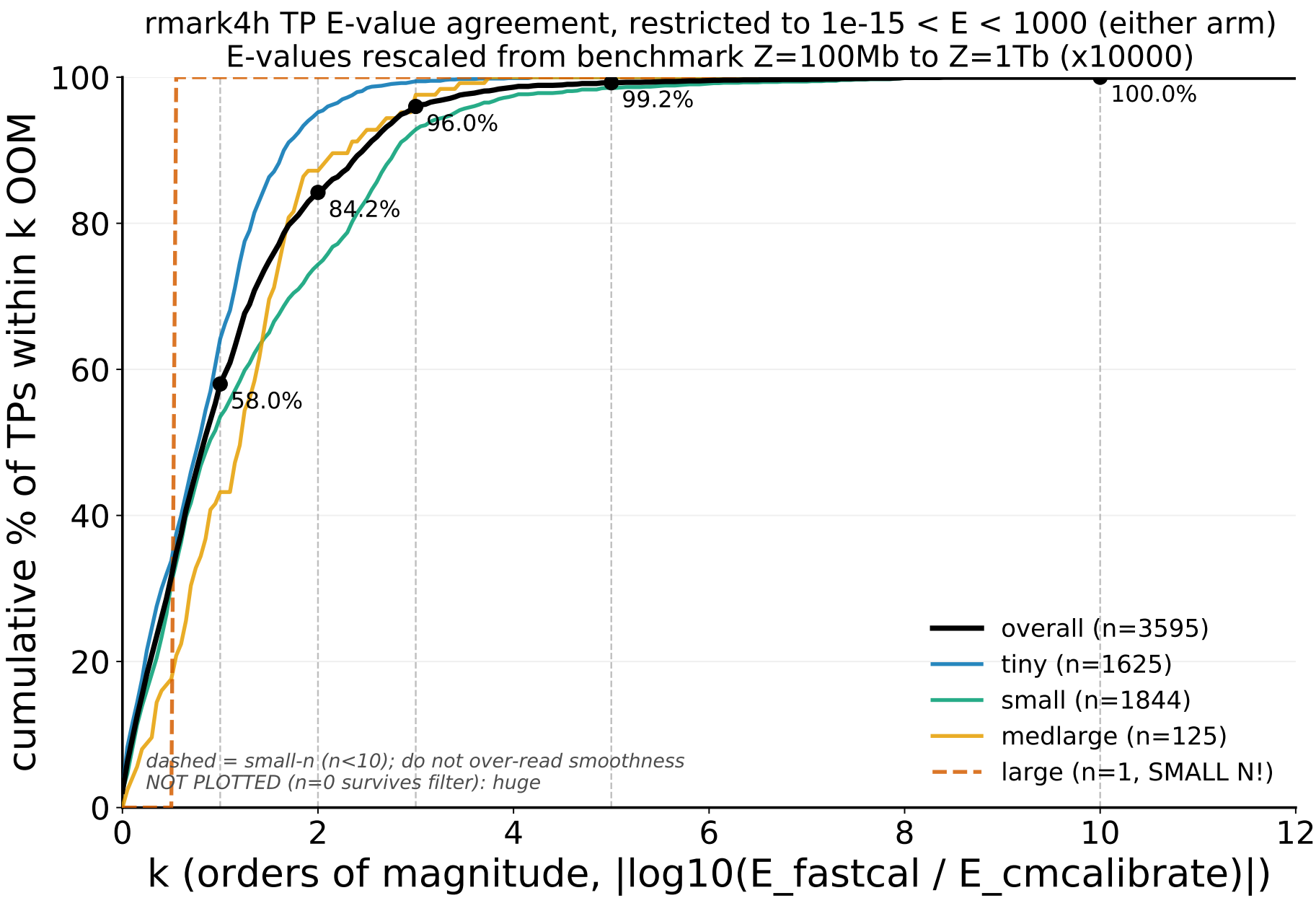


n=5314 matched true positives, rmark4h — Pearson r=0.991 on $\log_{10}$(E);
91.3% agree within 3 orders of magnitude

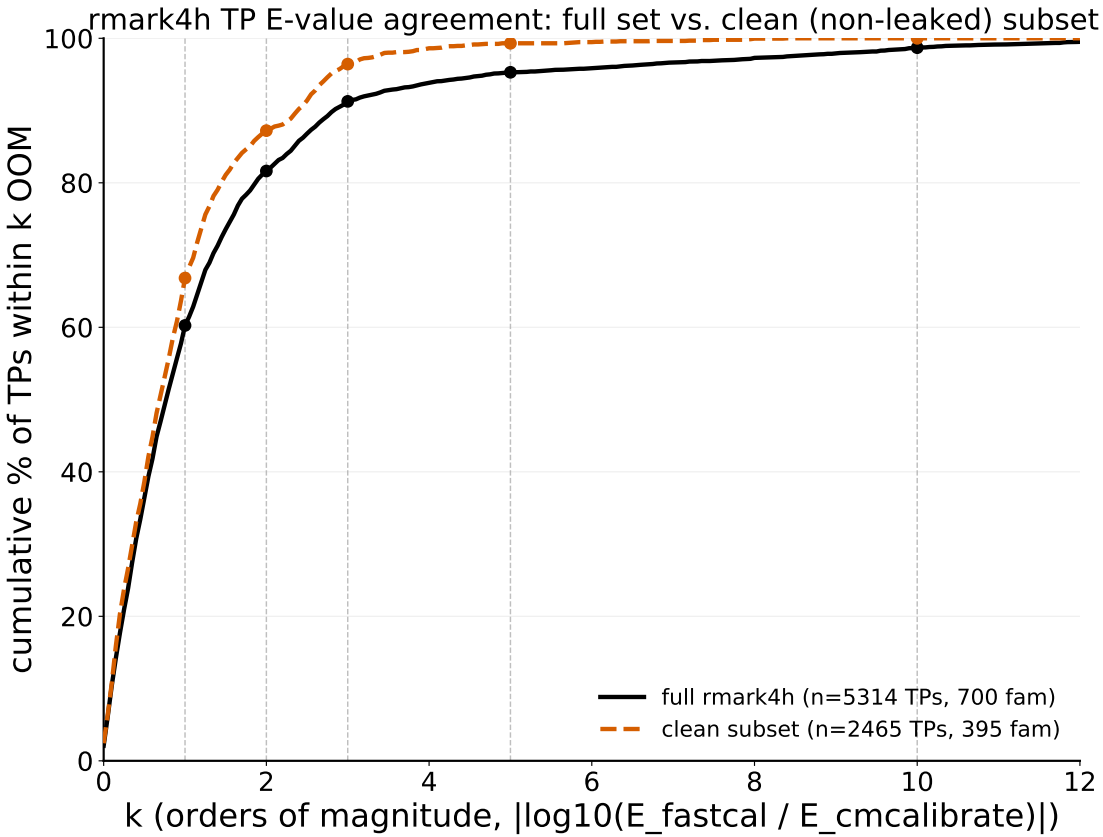

(A) Direction is visible: large/huge sit entirely on the safe (conservative) side
← fastcal less conservative
fastcal more conservative →
cumulative % of TPs
signed log10(Efastcal/Ecmcalibrate)
0
20
40
60
80
100
0
5
10
15
20
25
tiny (n=2312, 55% pos.)
small (n=2476, 69% pos.)
medlarge (n=501, 78% pos.)
large (n=20, 100% pos.)
huge (n=5, 100% pos.)


(B) The large/huge OOM gap lives entirely inside the astronomically-significant regime
22/25 large+huge TPs are also decisively significant (E<1e-50) in BOTH arms
3 weaker exceptions (still same qualitative call in both arms)
E_{fastcal}
$E_{\text{cmcalibrate}}$
10^{-29}
10^{-61}
10^{-93}
10^{-125}
10^{-157}
10^{-189}
10^{-221}
10^{-253}
10^{-285}
y = x (perfect agreement)
large (n=20)
huge (n=5)


rmark4h TP E-value agreement, restricted to 1e-15 < E < 1000 (either arm)
cmcalibrate vs cm_FastCalibrate
cumulative % of TPs within k OOM
k (orders of magnitude, |log10(E_fastcal / E_cmcalibrate)|)
68.1%
89.1%
98.0%
99.8%
100.0%
overall (n=4436)
tiny (n=2306)
small (n=1961)
medlarge (n=167)
large (n=2, SMALL N!)
dashed = small-n (n<10); do not over-read smoothness
NOT PLOTTED (n=0 survives filter): huge


rmark4h TP E-value agreement, restricted to 1e-15 < E < 1000 (either arm)
E-values rescaled from benchmark Z=100Mb to Z=1Tb (x10000)
cumulative % of TPs within k OOM
k (orders of magnitude, |log10(E_fastcal / E_cmcalibrate)|)
0
20
40
60
80
100
0
2
4
6
8
10
12
58.0%
84.2%
96.0%
99.2%
100.0%
dashed = small-n (n<10); do not over-read smoothness
NOT PLOTTED (n=0 survives filter): huge
overall (n=3595)
tiny (n=1625)
small (n=1844)
medlarge (n=125)
large (n=1, SMALL N!)

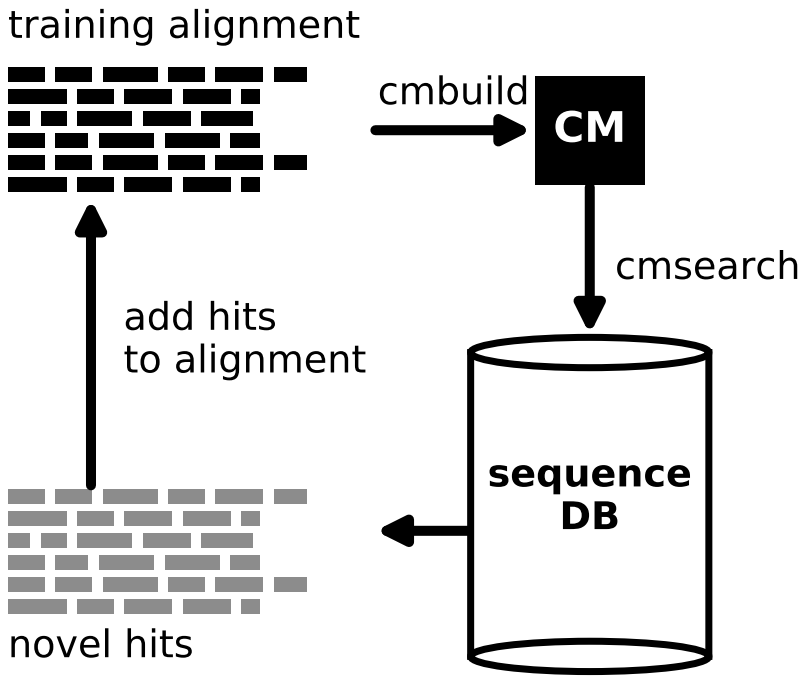
Does this hold up on non-training families?



395/700 rmark4h families with zero training overlap (2465/5314 TPs) - agreement equal-or-better than the full set: $\leq$1 OOM 60.3%$\rightarrow$66.8%, $\leq$3 OOM 91.3%$\rightarrow$96.4%, $\leq$5 OOM 95.3%$\rightarrow$99.3% (full$\rightarrow$clean)

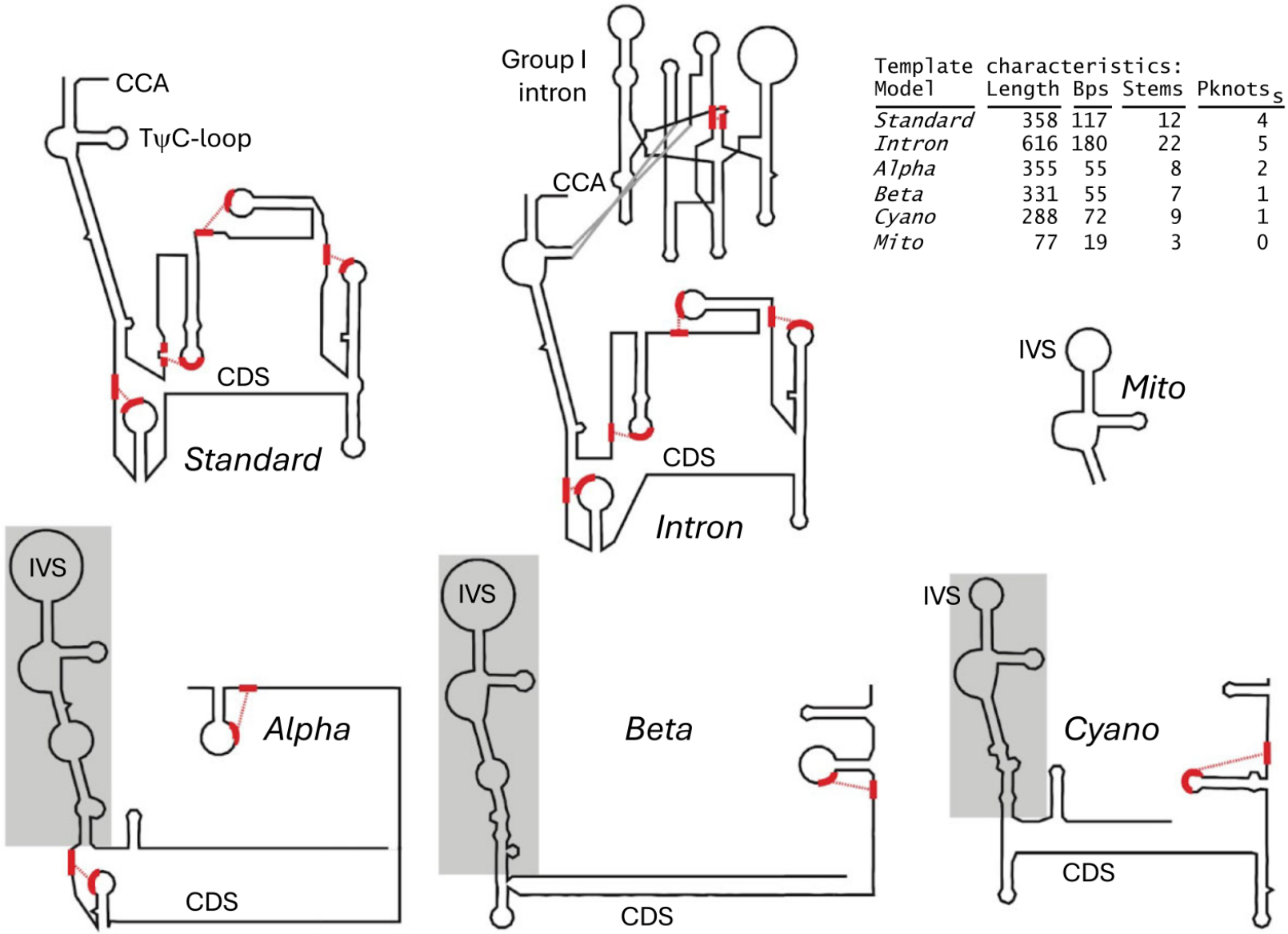
caveat: this clean subset has **zero** *large/huge-bucket coverage (every real large/huge Rfam family was already in training) - it validates tiny/small, not the biggest models; the LOOCV-over-synthetic-CMs result covers large/huge separately*

cmcalibrate eliminated!?



I'm currently trying to decide if `cmcalibrate` should be left in Infernal as a fallback for very highly specialized use cases like AT-rich models (and for Claude-skeptics).

Pseudoknot alignment markup



Template characteristics:

Model	Length	Bps	Stems	Pknots
Standard	358	117	12	4
Intron	616	180	22	5
Alpha	355	55	8	2
Beta	331	55	7	1
Cyano	288	72	9	1
Mito	77	19	3	0

Figure 3. The six tmRNA secondary structure templates, generated by R2DT in "thumbnail" format using the consensus sequence for each template's CM. Pseudoknot stems not modeled by the CM are highlighted, with each half connected by dashed lines. The boxed regions in the alpha, beta, and cyano templates are the tRNA domains, which include all base pairs modeled by the new, more general tmRNA_permuted Rfam model. Counts of base pairs (Bps) and stems in the table include pseudoknots.

Nawrocki, Petrov, Williams (2025), NARGAB, https://doi.org/10.1093/nargab/lqaf019

```
>> AP009178.1  Nitratiruptor sp. SB155-2 genomic DNA, complete genome
 rank     E-value  score  bias mdl mdl from   mdl to       seq from      seq to       acc trunc   gc
 ----   --------- ------ ----- --- -------- --------    ----------- -----------      ---- ----- ----
  (1) !   1.1e-65  225.6   0.0  cm        1      358 []           1         375 + [] 0.94    no 0.54

                                             v                                                 NC
               [[[[[[[,,,,,,,,,,,,,(((((((((-(((((((((((-(((((,,,,,,,<<<<<<________________>>> CS
tmRNA.i1ra   1 GGGGgUGaacuGGuUUCGACgggaauguugauuccaaaagcuGCAuGcCGAGguugccggugacCUCGUaAAAcaaccg 79
               GGGGGUGA +UGG UUCGACG:GA :: :  :::CA:A:GC:GC:UGCCG    UG::G:  AC  CGUAAAA+  :C:
AP009178.1   1 GGGGGUGACAUGGCUUCGACGGGAGCAGUC-UGUCACAGGCUGCGUGCCGGCU-UGAGGA--ACGCCGUAAAAAU-UCC 74
               *****************************7.*******************999.9****9..9**********65.9** PP

                                                                                               NC
               >>>,,,,,,,,,,,,,,,,,,,,,,,,,,,,,.................,,,,,<<<<<<<_______>>>>>>>.,,< CS
tmRNA.i1ra  80 gcaAAAaaAUAAaUGCcAAcaAuaAuuacuau.................cGCucucGCuGcuUAAuaagCaGCga.gcC 140
               :CAAAAAAAUA A GC+AACAAUA + A+UAU                  GC+ :: ::G:UUAA AA:C:: ::  CC
AP009178.1  75 UCAAAAAAAUAGACGCAAACAAUACAGAUUAUagaccggcuuacgcaguAGCAGCGUAAGUUUAACAAACUUACGgUCC 153
               *******************************************888887888888888888888888888888799** PP

                                                                                               NC
               <<<<<.<<<<<<<<_________>>>>>>>>-...->>>>>>,,,..,,,,,,,,,,,,.,<<<<<<<<<_________ CS
tmRNA.i1ra 141 guccc.cccagcccuuGCCuguugggcuggga...aggggcGUCA..uaaaagcgGGaU.agccuuuagguuuGcCugg 212
               :UC:C CC:::C::U GCC+ +U::G:::GGA   +G:GG:GUCA   + +AG GGG+U  G:::::GGU UGCCU
AP009178.1 154 CUCGCaCCUCUCUUU-GCCCUCUAAGAGAGGAuuuUGCGGGGUCAucCUCUAGAGGGCUuGGAUGAGCGGU-UGCCUA- 229
               *****8888899988.*******9999888889**************************888888888887.899999. PP

                                                                                               NC
               __>>>>>>>>>,,,,,,,,,,,,,,,,,<<<<<<<<<<.-<<<________>>>.->>>>>>>>>>,,,,,,.,,,,,, CS
tmRNA.i1ra 213 ggccuaaaggcGAAaauuaaagaGGCUagccuuaagga.agccUGUcugUuggc.auccuuaaggcGAAAau.aAAaaa 288
                GCC::::::CGA   +  ++ AGGCU G::U:A:::A  ::: G   G U:::  U:::U:A::CGA A+U A  A
AP009178.1 230 -GCCGCUCAUCGAC--ACCUUUAGGCUGGGUUGAUACAgGUUUCGGGCGCUGAAcCUGUAUCAACCGAGAUUgAUCAGC 305
               .6888888889998..**********************8888888888888888************************* PP

                                           v                                          NC
               ,,,,,,,,)))))----))))))))))))))))))))),,,,,<<<<<_______>>>>>]]]]]]]:::: CS
tmRNA.i1ra 289 caGACUAAgCaUGUAGAagcuuuuggaauaacauuuucGGACGcGGGUUCGAuUCCCgCCAcCUCCACCA 358
                  ACUA+GCA:GUAGA:GC:U:UG::: : :: UU:CGGAC CGGGUUCGAUUCCCG CACCUCCACCA
AP009178.1 306 GCUACUACGCACGUAGAGGCCUGUGGUGGGUUGUUUCCGGACUCGGGUUCGAUUCCCGACACCUCCACCA 375
               ********************************************************************** PP


NC line: 'v' is negative scoring, noncanonical basepair
```

```
>> AP009178.1  Nitratiruptor sp. SB155-2 genomic DNA, complete genome
 rank     E-value  score  bias mdl mdl from   mdl to       seq from      seq to       acc trunc   gc
 ----   --------- ------ ----- --- -------- --------    ----------- -----------      ---- ----- ----
  (1) !   1.1e-65  225.6   0.0  cm        1      358 []           1         375 + [] 0.94    no 0.54

                                              v                ==$$x          x$$==            PS
               [[[[[[[,,,,,,,,,,,,((((((((((-((((((((((((-(((((AAAAA,<<<<<<___aaaaa________>>> CS
tmRNA.i1ra   1 GGGGgUGaacuGGuUUCGACgggaauguugauuccaaaagcuGCAuGcCGAGguugccggugacCUCGUaAAAcaaccg 79
               GGGGGUGA +UGG UUCGACG:GA :: :  :::CA:A:GC:GC:UGCCG    UG::G:  AC  CGUAAAA+  :C:
AP009178.1   1 GGGGGUGACAUGGCUUCGACGGGAGCAGUC-UGUCACAGGCUGCGUGCCGGCU-UGAGGA--ACGCCGUAAAAAU-UCC 74
               *****************************7.*******************999.9****9..9**********65.9** PP

                                                                                               PS
               >>>,,,,,,,,,,,,,,,,,,,,,,,,,,,,,.................,,,,,<<<<<<<_______>>>>>>>.,,< CS
tmRNA.i1ra  80 gcaAAAaaAUAAaUGCcAAcaAuaAuuacuau.................cGCucucGCuGcuUAAuaagCaGCga.gcC 140
               :CAAAAAAAUA A GC+AACAAUA + A+UAU                  GC+ :: ::G:UUAA AA:C:: ::  CC
AP009178.1  75 UCAAAAAAAUAGACGCAAACAAUACAGAUUAUagaccggcuuacgcaguAGCAGCGUAAGUUUAACAAACUUACGgUCC 153
               ********************************************888888788888888888888888888888799** PP

                               $==$$$=                            =$$$==$               =x==xx PS
               <<<<<.<<<<<<<<__BBBBBBB>>>>>>>>-...->>>>>>,,,..,,,,bbbbbbb,.,<<<<<<<<<___CCCCCC CS
tmRNA.i1ra 141 guccc.cccagcccuuGCCuguugggcuggga...aggggcGUCA..uaaaagcgGGaU.agccuuuagguuuGcCugg 212
               :UC:C CC:::C::U GCC+ +U::G:::GGA   +G:GG:GUCA   + +AG GGG+U  G::::::GGU UGCCU
AP009178.1 154 CUCGCaCCUCUCUUU-GCCCUCUAAGAGAGGAuuuUGCGGGGUCAucCUCUAGAGGGCUuGGAUGAGCGGU-UGCCUA- 229
               *****8888899988.*******9999888889**************************888888888887.899999. PP

                                    xx==x=                 =x$xxxx                        xxx  PS
               __>>>>>>>>>,,,,,,,,,,cccccc,<<<<<<<<<<.-<<<_DDDDDDD>>>.->>>>>>>>>>,,,,,,.,,ddd, CS
tmRNA.i1ra 213 ggccuaaaggcGAAaauuaaagaGGCUagccuuaagga.agccUGUcugUuggc.auccuuaaggcGAAAau.aAAaaa 288
                GCC::::::CGA   +  ++ AGGCU G::U:A:::A  ::: G   G U:::  U:::U:A::CGA A+U A  A
AP009178.1 230 -GCCGCUCAUCGAC--ACCUUUAGGCUGGGUUGAUACAgGUUUCGGGCGCUGAAcCUGUAUCAACCGAGAUUgAUCAGC 305
               .6888888889998..**********************88888888888888888************************ PP

                x$x=                       v                                         PS
               ,dddd,,,)))))----)))))))))))))))))))))),,,<<<<<_______>>>>>]]]]]]]:::: CS
tmRNA.i1ra 289 caGACUAAgCaUGUAGAagcuuuuggaauaacauuuucGGACGcGGGUUCGAuUCCCgCCAcCUCCACCA 358
                  ACUA+GCA:GUAGA:GC:U:UG::: : :: UU:CGGAC CGGGUUCGAUUCCCG CACCUCCACCA
AP009178.1 306 GCUACUACGCACGUAGAGGCCUGUGGUGGGUUGUUUCCGGACUCGGGUUCGAUUCCCGACACCUCCACCA 375
               ********************************************************************** PP


PS line: 'v' is negative scoring, noncanonical nested basepair
         '=' is unchanged pseudoknotted basepair
         'x' changed pseudoknotted basepair, noncanonical
         '$' changed pseudoknotted basepair, canonical WC or GU
```

```
>> FR891194.1  Firmicutes bacterium CAG:272, WGS project CBFW01000000 data, scaffold: scf321, whole geno
 rank     E-value  score  bias mdl mdl from   mdl to       seq from      seq to       acc trunc   gc
 ----   --------- ------ ----- --- -------- --------    ---------- ----------      ---- ----- ----
  (3) !   2.4e-54  186.5   0.0  cm        1      358 []          1         365 + [] 0.91    no 0.58

                                        v  v    v           $===x          x===$                 PS
                 [[[[[[[,,,,,,,,,,,,(((((((((-((((((((((-(((((AAAAA,<<<<<<___aaaaa_______>>>>>>. CS
tmRNA.i1ra     1 GGGGgUGaacuGGuUUCGACgggaauguugauuccaaaagcuGCAuGcCGAGguugccggugacCUCGUaAAAcaaccggca. 82
                 GGGG::G+A+ GGUUUCGACGGG:A:: :G :::C :::::::GC::GC GAG  :G: ::+ A CUC U+AAAC  :: :C:
FR891194.1     1 GGGGACGUAAAGGUUUCGACGGGGACUCUG-AAGCAGGGUAAGCGGGCAGAGC-CGGGACA-AUCUCUUUAAACU-GUUCCAa 79
                 ****************************.**********************.8888887.***********76.8888889 PP

                                                 v                 v                             PS
                 ..,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,<<<<<<<_______>>>>>>>.....................,,<<<<< CS
tmRNA.i1ra    83 ..AAAaaAUAAaUGCcAAcaAuaAuuacuaucGCucucGCuGcuUAAuaagCaGCga.....................gcCgucc 144
                    AAA+ UAAA GC AACA U A+U++    GC   +:CUGC+U A   GCAG:                        CCGU:C
FR891194.1    80 ccUAAAUUUAAACGCUAACACUGAAUUA----GC---UGCUGCCUGAACGGCAGU--gacgcggacuuaagcuccgCCCGUUC 153
                 ****************************....66...57888888865566665..559999999999999999*****  PP

                       v  $xx=xxxv                        xxx=xx$           v     =x=$$$  v          PS
                 <<<<<<<<<__BBBBBBB>>>>>>>-->>>>>>,,,,,,,bbbbbbb,,<<<.<<<<<_...__CCCCCC__>>>>>>>>>, CS
tmRNA.i1ra   145 ccccagcccuuGCCuguugggcugggaaggggcGUCAuaaaagcgGGaUagcc.uuuaggu..uuGcCuggggccuaaaggcG 224
                 ::CC:G::     CUG++ ::C:GG: A:G:GCGUCAU AA   G GA +G:: ::::     U+G+C + +   ::::::CG
FR891194.1   154 AUCCGGAGAG-UACUGCGGCUCCGGA-AUGAGCGUCAUCAAGCAGUGAACGUGcGCGGUCAguUCGUCCUUAUAGCCGCCACG 234
                 *9**9*****.9999999****9**8.9*********6666666666677775666666698688888866666666666 PP

                          $$$=x=        vvv     x$=$xxx    vvv               xxx  $=$x              PS
                 ,,,,,,,,,cccccc,<<<<<<<<<<-<<<_DDDDDDD>>>->>>>>>>>>>,,,,,,,,ddd,,dddd,,,)))))----)) CS
tmRNA.i1ra   225 AAaauuaaagaGGCUagccuuaaggaagccUGUcugUuggcauccuuaaggcGAAAauaAAaaacaGACUAAgCaUGUAGAag 307
                     +U AA +GGCU ::C:::A    : : G+C G U: :    U:::G::GA A++A  AA C G+ UA+GC::G AGA::
FR891194.1   235 ---GAUGAAAGGGCUGCUCCGGAAAAGCUUCGCCGGCUUGGCGACUCCGGGGGAGAUCAGUAAGCCGGAUACGCCCGAAGAAA 314
                 ...568999**************************************************************************  PP

                    v    v v                                    PS
                 ))))))))))))))))))),,,<<<<<_______>>>>>]]]]]]]:::: CS
tmRNA.i1ra   308 cuuuuggaauaacauuuucGGACGcGGGUUCGAuCCCgCCAcCUCCACCA 358
                 :::: G:::+: ::U:UUCGGACGCGGGUUCGAUUCCCGCC::CUCCACCA
FR891194.1   315 GCCCCGCGGAUCGGUUUUCGGACGCGGGUUCGAUUCCCGCCGUCUCCACCA 365
                 ************************************************  PP



PS line: 'v' is negative scoring, noncanonical nested basepair
         '=' is unchanged pseudoknotted basepair
         'x' changed pseudoknotted basepair, noncanonical
         '$' changed pseudoknotted basepair, canonical WC or GU
```

`cmalign` Stockholm output: pseudoknot A/a (well-conserved)

```
AP009178.1:366342-366716/1-375             .-UGUCACAGGCUGCGUGCCGGCU.-UGAGGA-.-ACGCCGUAAAAAU-UCCUCA...AA
#=GR AP009178.1:366342-366716/1-375   PP ..*******************999..9****9...9**********65.9*****...**
#=GR AP009178.1:366342-366716/1-375   PS .v:::__:_:__:__:___==$$x.-__::_:-.-_x$$==_____+_-:_::__...__
CP001085.1:311523-311892/370-1             .AGGUUUGAAGUAGCAUGUCGAGC.-AGA----.-UUCUCGUUAAA--A---UCU...-A
#=GR CP001085.1:311523-311892/370-1   PP .888888889**************..555......8**********..4...555....*
#=GR CP001085.1:311523-311892/370-1   PS .v::::::vv_:v_____:====x.-:_:----.-_x====_+___--_---:_:...-_
CP002116.1:2157433-2157788/356-1           aCACUCACCGGUUGCAAGCGGAGG.-UGCCCUU.CUCCUCCUUAAUCC-AGGGCA..aAC
#=GR CP002116.1:2157433-2157788/356-1 PP 989**********************..*******.************77.******..***
#=GR CP002116.1:2157433-2157788/356-1 PS .v:::__:v:_::___:__$====.-____::_.__====$_+__+__-::____...__
CP001744.1:2441287-2441647/1-361           aUACGGAAUGGUGGCGUGUCGUGGuUGAUCAGAcGGCCACGUAAAAAU-CUGAUC...AA
#=GR CP001744.1:2441287-2441647/1-361 PP 6**********************8777777779************66.777777...8*
#=GR CP001744.1:2441287-2441647/1-361 PS .v:v::__::_:v__:__:==$==._:::_:_+._+==$==_____+_-_:_:::...__
FR891194.1:2245-2609/1-365                 .-AAGCAGGGUAAGCGGGCAGAGC.-CGGGACA.-AUCUCUUUAAACU-GUUCCAaccUA
#=GR FR891194.1:2245-2609/1-365       PP ..**********************..8888887..***********76.8888889****
#=GR FR891194.1:2245-2609/1-365       PS .v:::_v::::::__::__$===x.-:_:_::+.-_x===$_+_____-::_:_:...__
#=GC SS_cons                               .((((((((((((-(((((AAAAA.,<<<<<<_.__aaaaa________>>>>>>...,,
#=GC RF                                    .auuccaaaagcuGCAuGcCGAGg.uugccggu.gacCUCGUaAAAcaaccggca...AA
#=GC bp_cons                               .066**8*68*82.*********4..8*****....4****........*****8.....
#=GC bp_cov                                .735695**5534.0690690956..95*7**....65909........**7*59.....
```

`cmalign` Stockholm output: pseudoknot A/a (well-conserved)

```
AP009178.1:366342-366716/1-375           .-UGUCACAGGCUGCGUGCCGGCU.-UGAGGA-.-ACGCCGUAAAAAU-UCCUCA...AA
#=GR AP009178.1:366342-366716/1-375   PP ..********************999..9****9...9**********65.9*****...**
#=GR AP009178.1:366342-366716/1-375   PS .v:::__:_:__:__:___==$$x.-__::_:-.-_x$$==_____+_-:_::__...__
CP001085.1:311523-311892/370-1           .AGGUUUGAAGUAGCAUGUCGAGC.-AGA----.-UUCUCGUUAAA--A---UCU...-A
#=GR CP001085.1:311523-311892/370-1   PP .888888889**************..555......8**********..4...555....*
#=GR CP001085.1:311523-311892/370-1   PS .v::::::vv_:v_____:====x.-:_:----.-_x====_+___--_---:_:...-_
CP002116.1:2157433-2157788/356-1         aCACUCACCGGUUGCAAGCGGAGG.-UGCCCUU.CUCCUCCUUAAUCC-AGGGCA..aAC
#=GR CP002116.1:2157433-2157788/356-1 PP 989**********************..*******.************77.******..***
#=GR CP002116.1:2157433-2157788/356-1 PS .v:::__:v:_::___:__$====.-____::_.__====$_+__+__-::____...__
CP001744.1:2441287-2441647/1-361         aUACGGAAUGGUGGCGUGUCGUGGuUGAUCAGAcGGCCACGUAAAAAU-CUGAUC...AA
#=GR CP001744.1:2441287-2441647/1-361 PP 6**********************8777777779************66.777777...8*
#=GR CP001744.1:2441287-2441647/1-361 PS .v:v::__::_:v__:__:==$==._:::_:_+._+==$==_____+_-_:_:::...__
FR891194.1:2245-2609/1-365               .-AAGCAGGGUAAGCGGGCAGAGC.-CGGGACA.-AUCUCUUUAAACU-GUUCCAaccUA
#=GR FR891194.1:2245-2609/1-365       PP ..**********************..8888887..***********76.8888889****
#=GR FR891194.1:2245-2609/1-365       PS .v:::_v::::::__::__$===x.-:_:_::+.-_x===$_+_____-::_:_:...__
#=GC SS_cons                             .((((((((((((-(((((AAAAA.,<<<<<<_.__aaaaa________>>>>>>...,,
#=GC RF                                  .auuccaaaagcuGCAuGcCGAGg.uugccggu.gacCUCGUaAAAcaaccggca...AA
#=GC bp_cons                             .066**8*68*82.*********4..8*****....4****........*****8.....
#=GC bp_cov                              .735695**5534.0690690956..95*7**....65909........**7*59.....
```

--

- 'bp_cons' (base-pair conservation): per consensus base-pair column, how often that pair is actually formed as a canonical Watson-Crick/G-U pair across the alignment's sequences (0-9,*);

- 'bp_cov' (base-pair covariation): a separate, complementary measure, whether substitutions on the two sides of a pair are statistically coupled (compensatory), i.e. mutual information between the paired columns (0-9,*);

In this alignment 'bp_cons' is * across nearly the whole A/a pair (this pair is structurally almost always formed as WC/GU), and 'bp_cov' ranges 0-9, showing some of the pairs covary across the 5 sequences.

cmalign Stockholm output: pseudoknot C/c (poorly-conserved)

```
AP009178.1:366342-366716/1-375           UuGGAU.GAGCGGU..-UGCCUA.--GCCGCUCAUCGAC.--ACCUUUAGGCUG.GGUUG
#=GR AP009178.1:366342-366716/1-375   PP *88888.8888887...899999...6888888889998...***********.*****
#=GR AP009178.1:366342-366716/1-375   PS _.__::.::::___..-_=x==x.x-___::::::____.--+__++xx==x=_._::_:
CP001085.1:311523-311892/370-1           A.CACG.UUCAGUU..UUUCCUAuCAUUCUGAACGUUGA.--AAAAAUAGGAUA.GUUGA
#=GR CP001085.1:311523-311892/370-1   PP 6.6666.6666666..56666667555666666666544...79**********.*****
#=GR CP001085.1:311523-311892/370-1   PS _.+:_:.__:__v_..__xx==x.x+_v__:__:_:___.--++___xx==xx_._::::
CP002116.1:2157433-2157788/356-1         UuACCC.UUUUUGU..UCGGCUG.UGGCCUAAAGGGGAC.AUUUUAUCAGUCUUaGGUCG
#=GR CP002116.1:2157433-2157788/356-1 PP 8*****.*******..8999999.8999999********.99999999999999899999
#=GR CP002116.1:2157433-2157788/356-1 PS _._:__.___vv__..._+=$x=x.x___vv_____:___._+__+_+xx=x$=_._::v:
CP001744.1:2441287-2441647/1-361         U.GGCU.GCGGGUCaa-UGCGACcUACGCCCGCAGCAAGaUUUGUGGGACGCUG.GCACU
#=GR CP001744.1:2441287-2441647/1-361 PP *.9***.******995.55554455589********999888888889999999.***99
#=GR CP001744.1:2441287-2441647/1-361 PS _.___:.::::_:_..-_=xxxx.$+_:_:::::_____.++__+__$xxxx=_.:_:::
FR891194.1:2245-2609/1-365               A.CGUGcGCGGUCAguUCGUCCU.UAUAGCCGCCACG--.-GAUGAAAGGGCUG.CUCCG
#=GR FR891194.1:2245-2609/1-365       PP 6.77775666666986888888.6666666666666....568999*******.*****
#=GR FR891194.1:2245-2609/1-365       PS _.+_:::::::_v_..._+=x=$$.$+_v_::::::__--.-_+____$$$=x=_.::_::
#=GC SS_cons                             ,.,<<<.<<<<<<_..__CCCCC.C__>>>>>>>>>,,,.,,,,,,,,cccccc,.<<<<<
#=GC RF                                  U.agcc.uuuaggu..uuGcCug.gggccuaaaggcGAA.aauuaaagaGGCUa.gccuu
#=GC bp_cons                             ...***.***886.....82682.5..688******..........528628..***8*
#=GC bp_cov                              ...99*.6*9*2*.....01055.5..*2*9*6*99..........555010..57699

//
```

In this alignment 'bp_cons' for the pseudoknot is lower and patchier (mostly 2-8, never reaching *) and 'bp_cov' is weaker too (0-5): less consistent evidence for these pairs than for A/a in previous slide.

Infernal 1.2 development is ongoing

- We always take feature requests, but now is an especially good time
- Longstanding goal: continue to make Infernal more efficient
  - target of using Infernal for 10-35Kb viral genomes to drive development

Acknowledgements

- David Landsman
- Sean Eddy
- Tom Jones

Thanks, Eric and Elena!

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine