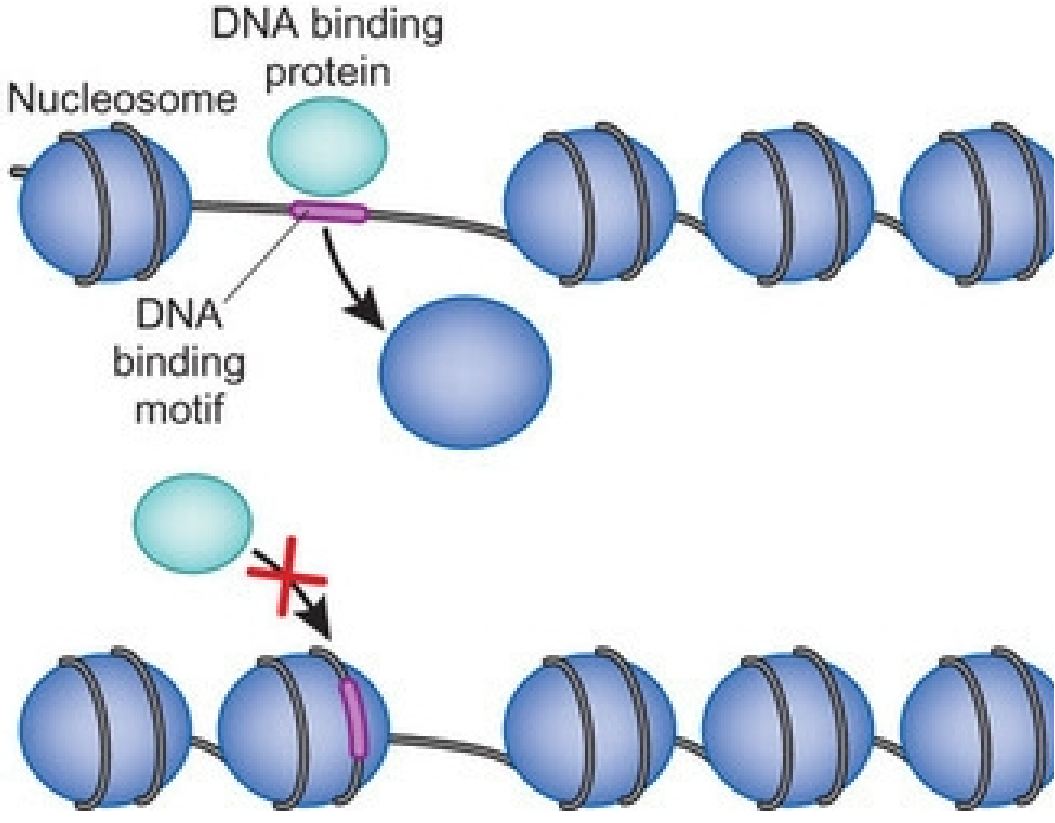
Transcription Factor Binding at Specific and Ubiquitous Open Chromatin Sites Across Multiple Human Cell Types

Martin Hemberg

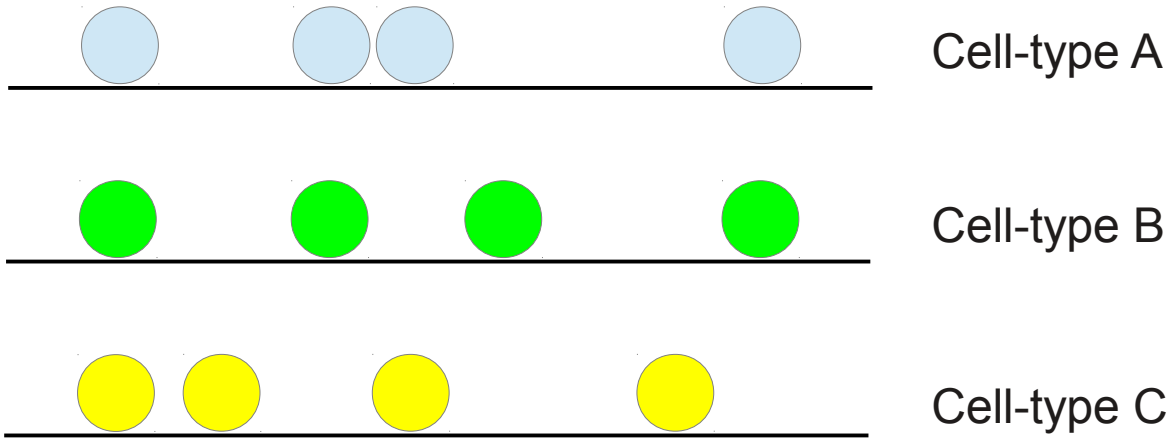
RECOMB Conference
on Systems Biology and Regulatory Genomics
San Francisco
11/13/12

Open chromatin are the accessible locations in the genome

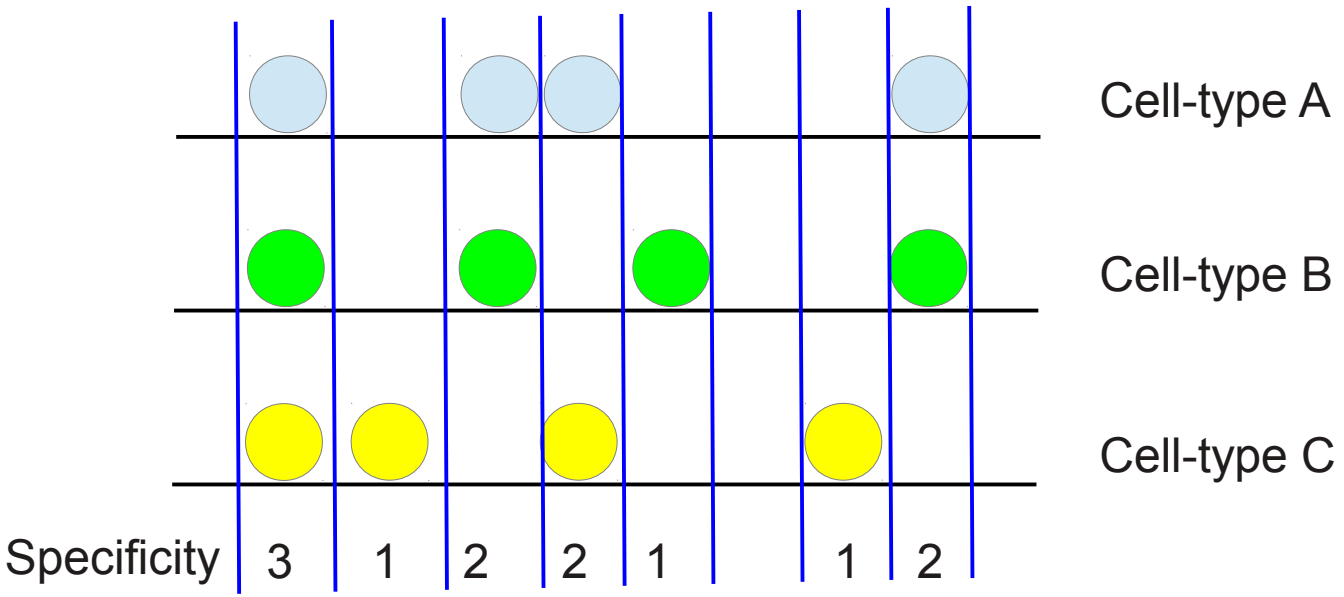
- DnaseI Hypersensitive Sites (DHSs)
- Overlap genes and regulatory elements
- ~100,000/cell-type
  - Cell-type specific



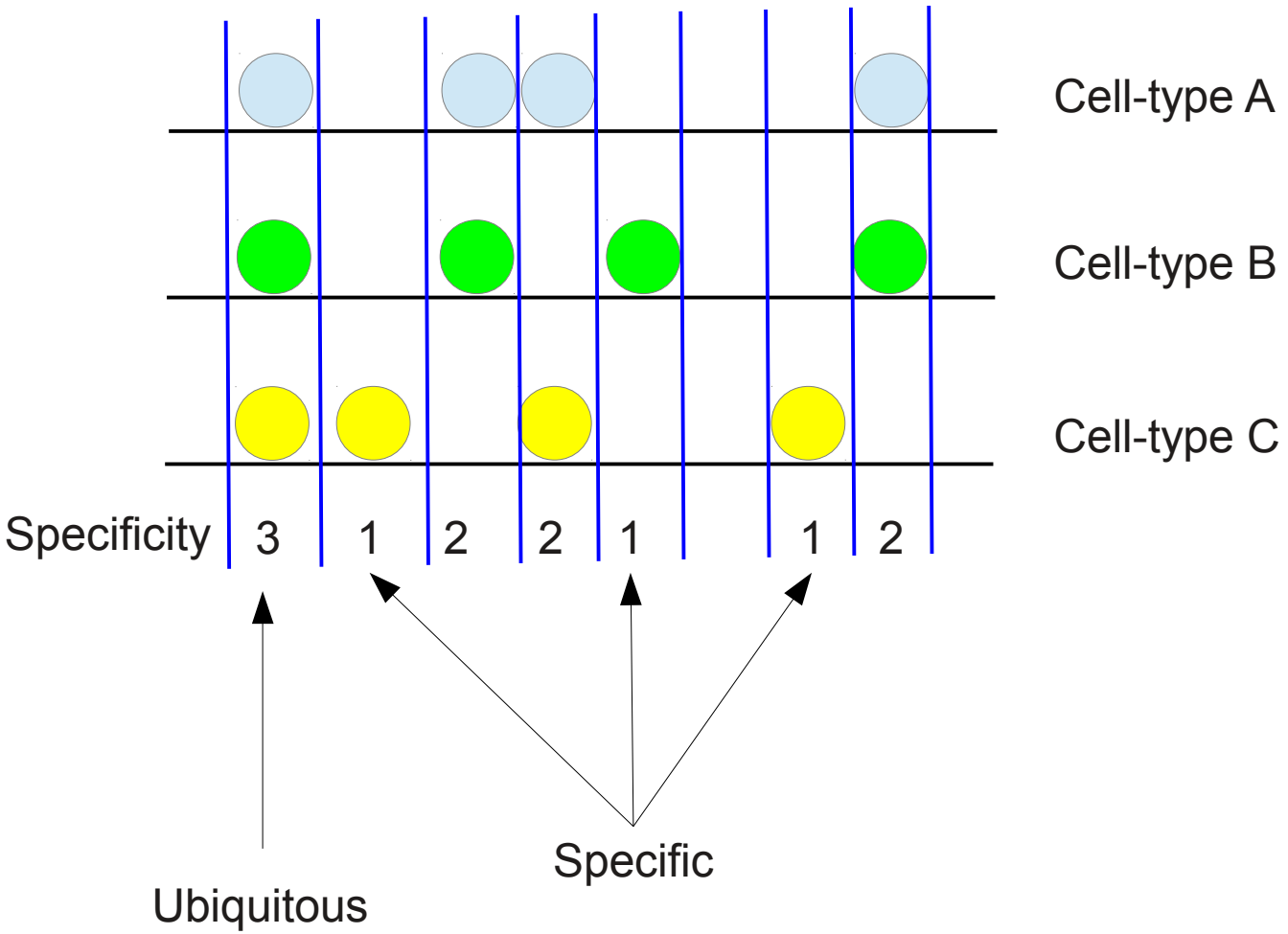
DHS specificity



DHS specificity

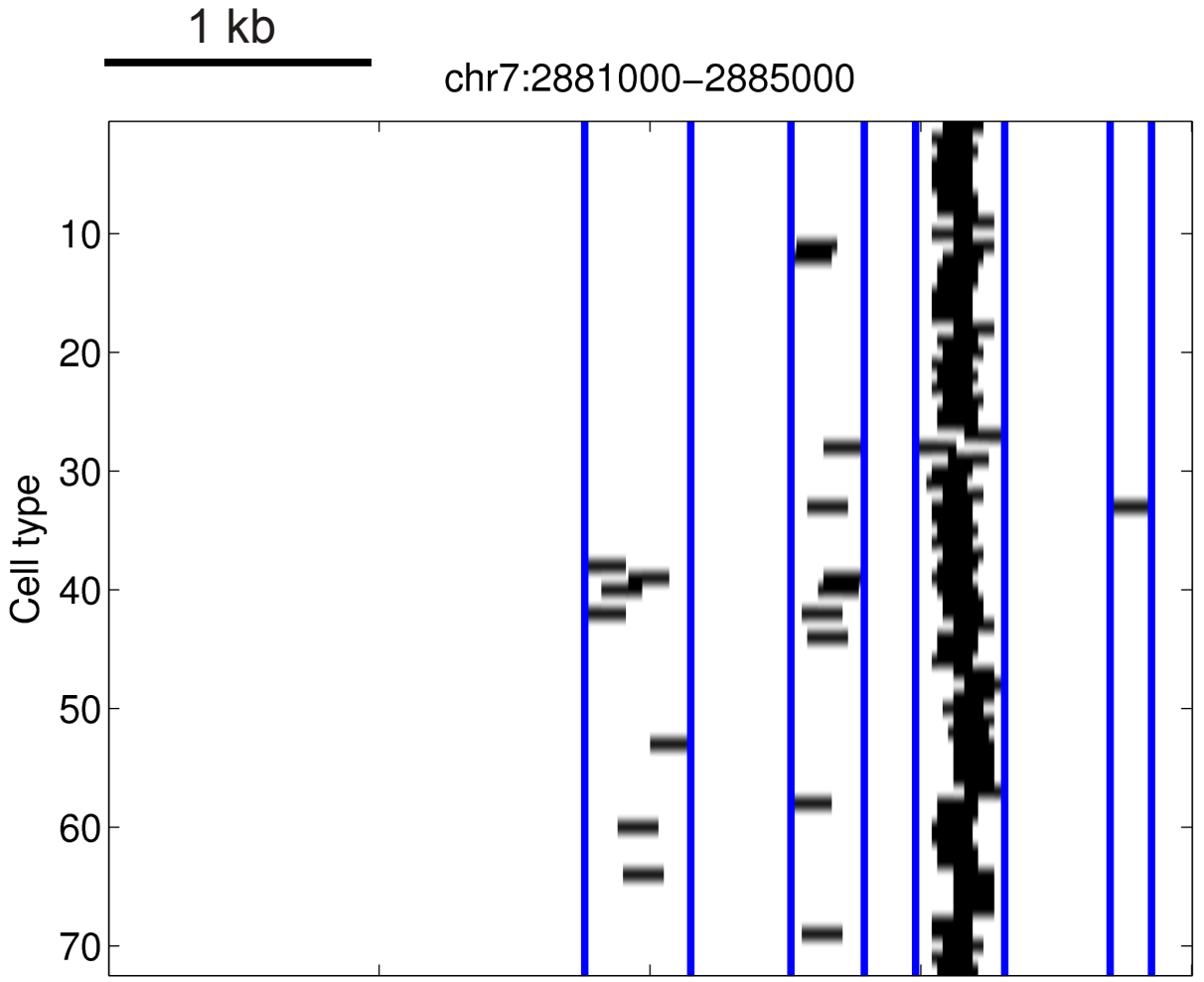


DHS specificity



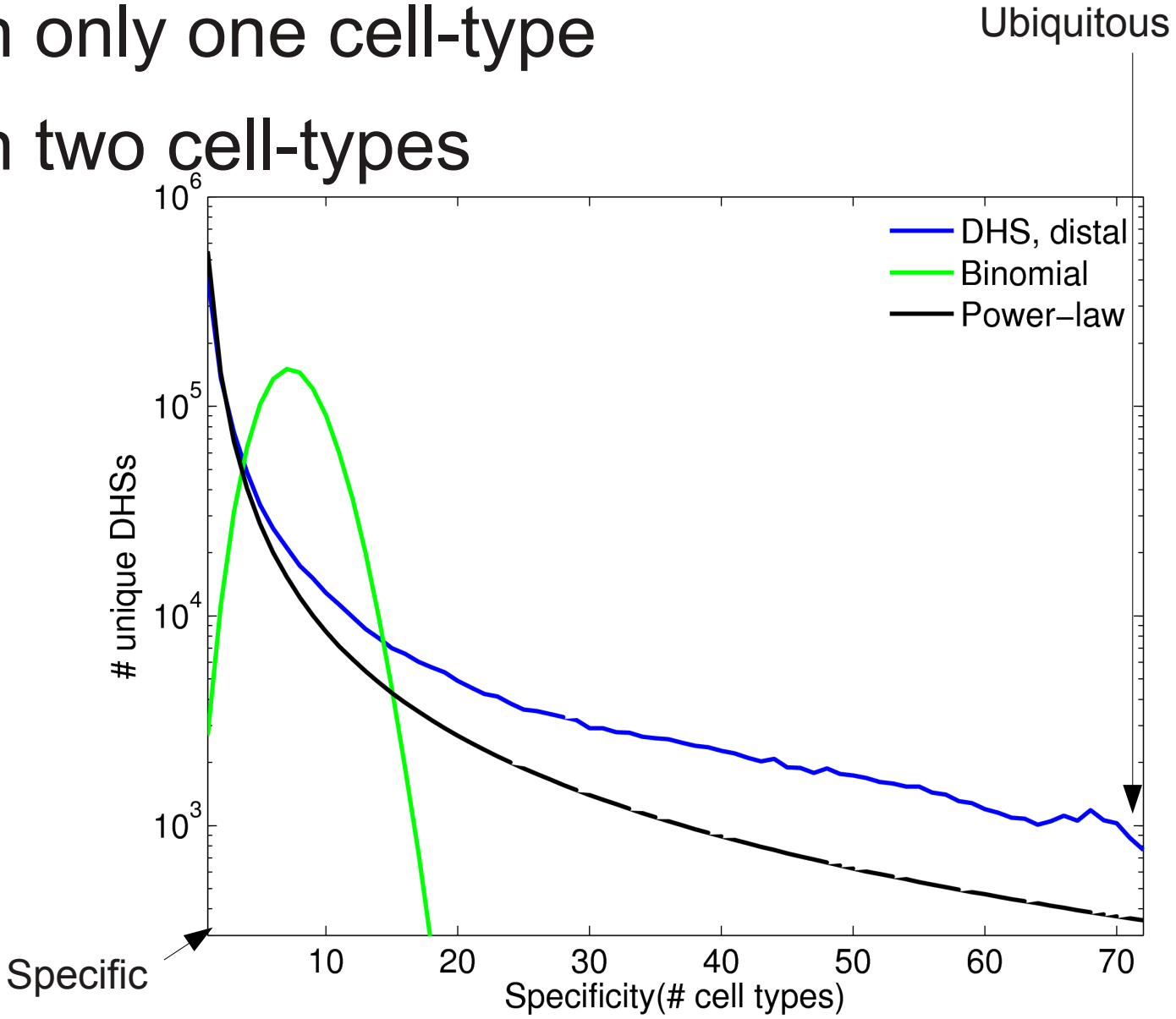
DHS Specificity, 72 cell-types

- ~1 million unique loci, >5 kb from TSS

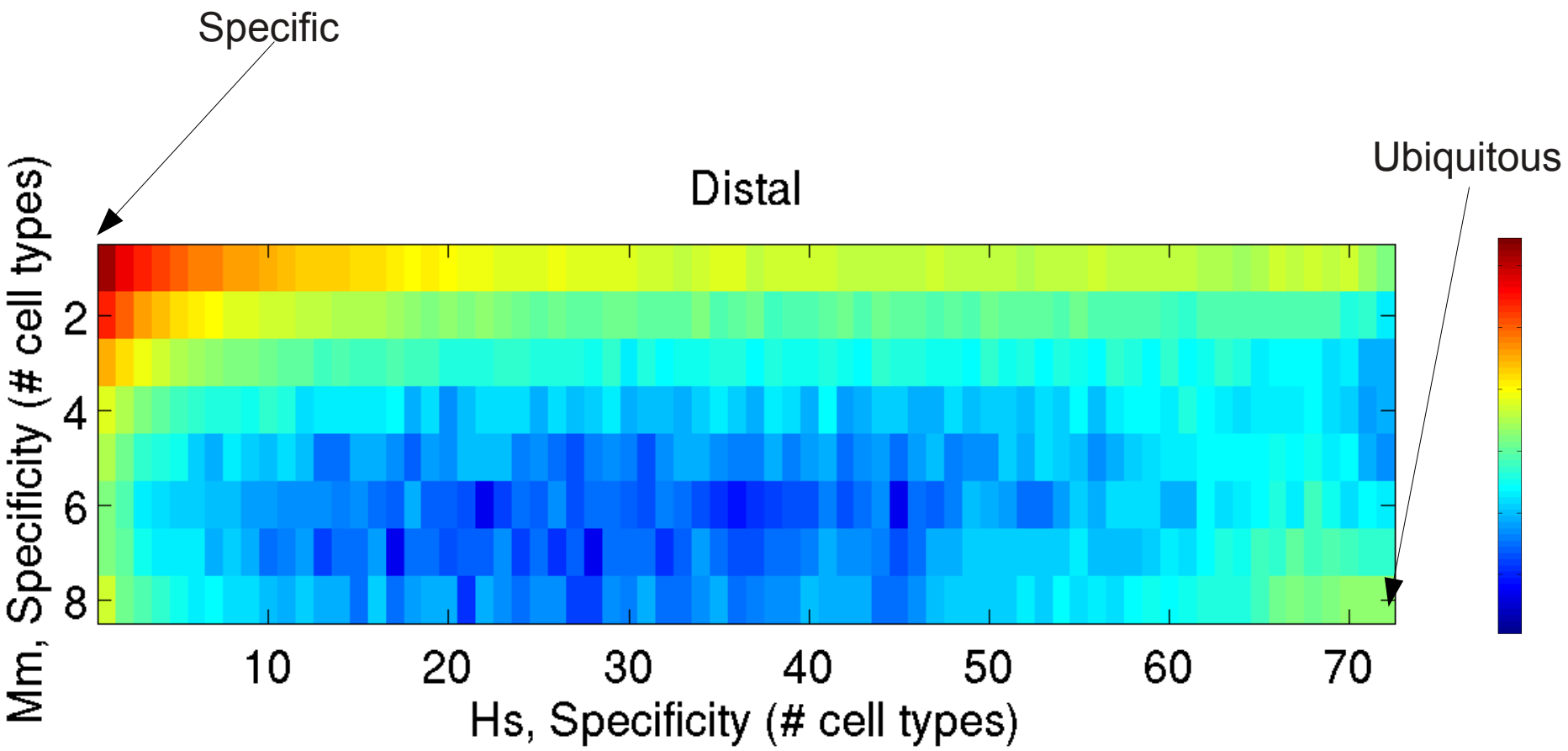


DHS Specificity follows a power-law

- ~42% found in only one cell-type
- ~14% found in two cell-types
- Slope ~ -1.35



Specificity conserved in mouse



Why do we observe the power-law?

- Start with one cell-type with N DHSs

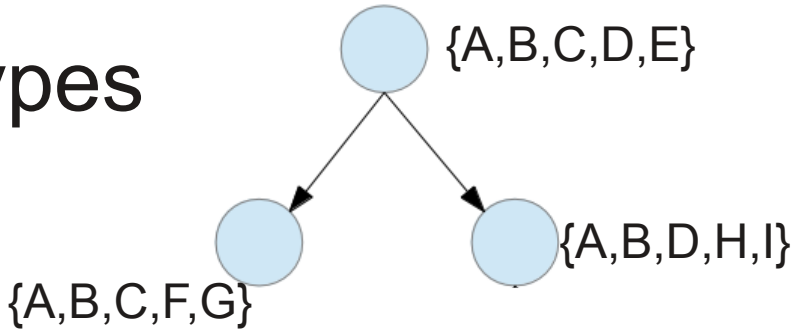
{A,B,C,D,E}

A: 1
B: 1
C: 1
D: 1
E: 1

Why do we observe the power-law?

- Start with one cell-type with N DHSs
- Generate two new cell-types

$$P(\text{keep DHS}) = \frac{k^b}{1 + k^b}$$



  - *k* - # cell-types where DHS present
  - *b* – selective advantage of existing DHSs
- Add new DHSs to get N

A: 3
B: 3
C: 2
D: 2
E: 1
F: 1
...

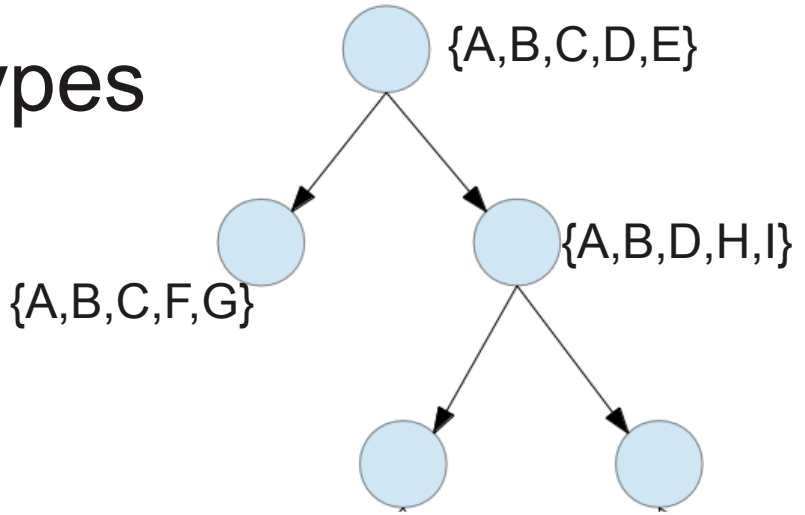
Why do we observe the power-law?

- Start with one cell-type with N DHSs
- Generate two new cell-types

$$P(\text{keep DHS}) = \frac{k^b}{1 + k^b}$$

  - *k* - # cell-types where DHS present
  - *b* – selective advantage of existing DHSs

– Add new DHSs to get N



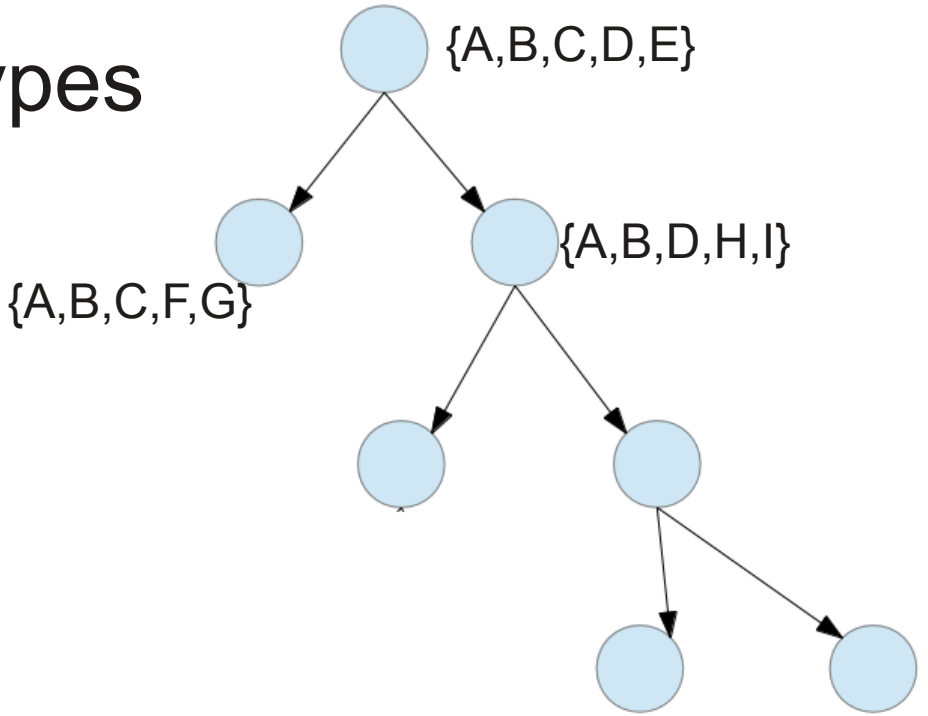
A: 4
B: 3
C: 2
D: 3
E: 1
F: 1
...

Why do we observe the power-law?

- Start with one cell-type with N DHSs
- Generate two new cell-types

$$P(\text{keep DHS}) = \frac{k^b}{1 + k^b}$$

  - *k* - # cell-types where DHS present
  - *b* – selective advantage of existing DHSs
- Add new DHSs to get N
- Expand until C nodes



A: 6
B: 3
C: 2
D: 5
E: 1
F: 1
...

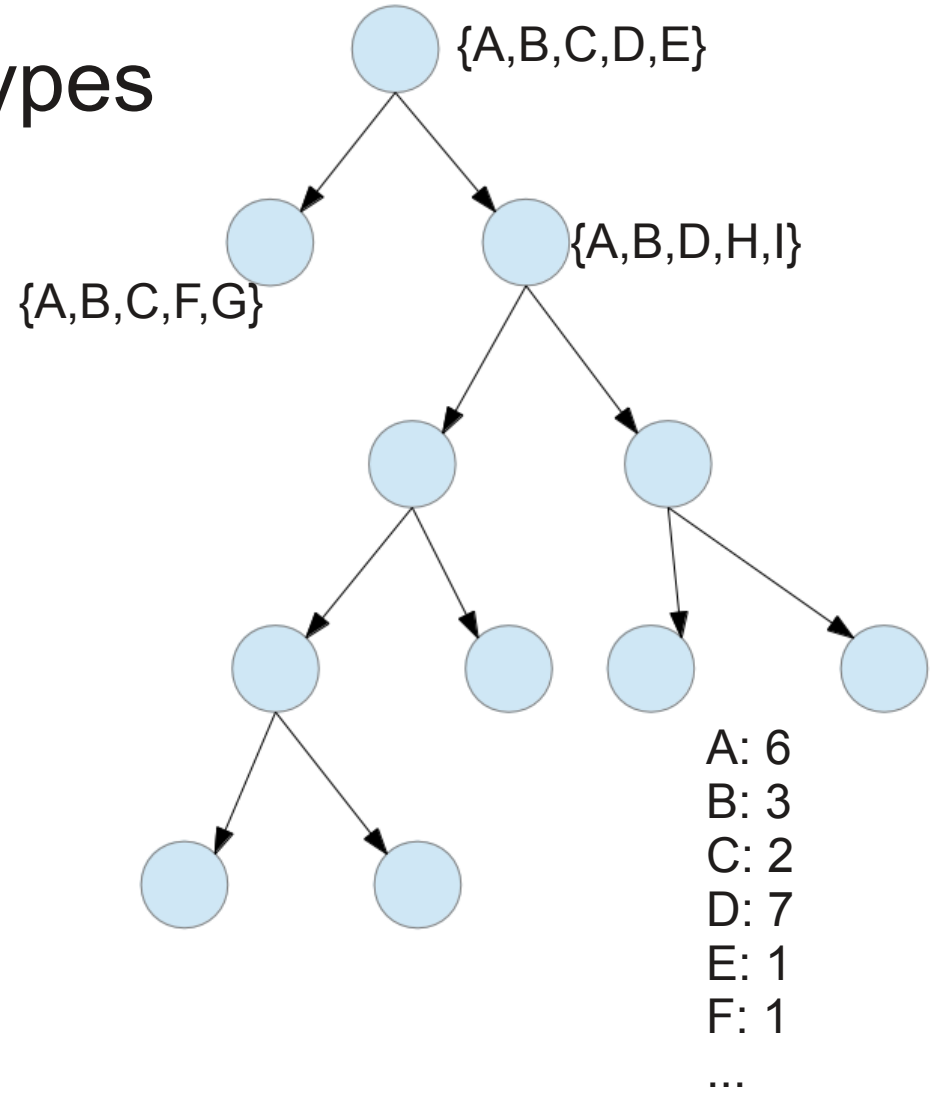
Why do we observe the power-law?

- Start with one cell-type with N DHSs
- Generate two new cell-types

$$P(\text{keep DHS}) = \frac{k^b}{1 + k^b}$$

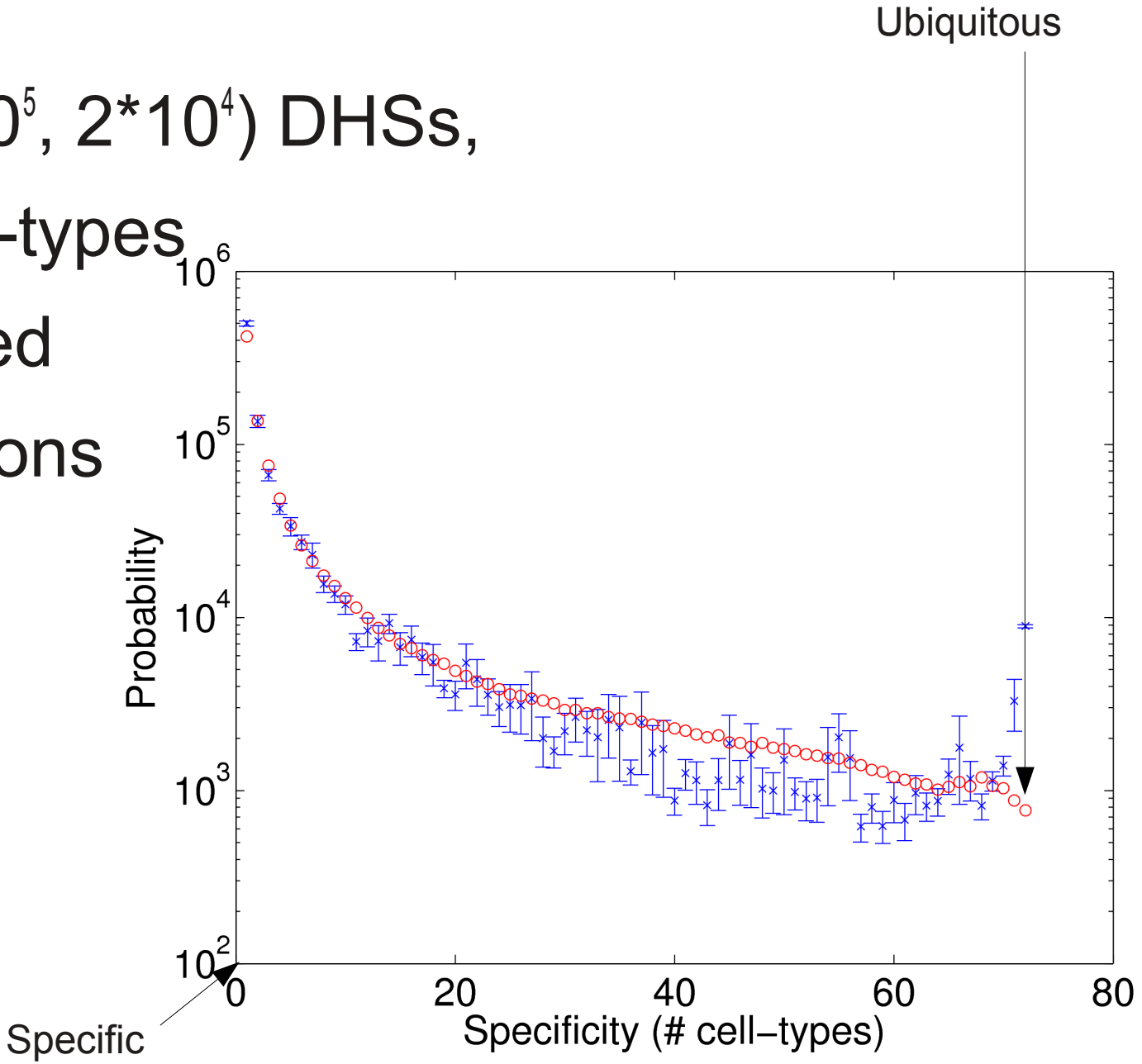
  - *k* - # cell-types where DHS present
  - *b* – selective advantage of existing DHSs

- Add new DHSs to get N
- Expand until C nodes
- Sample 72 nodes

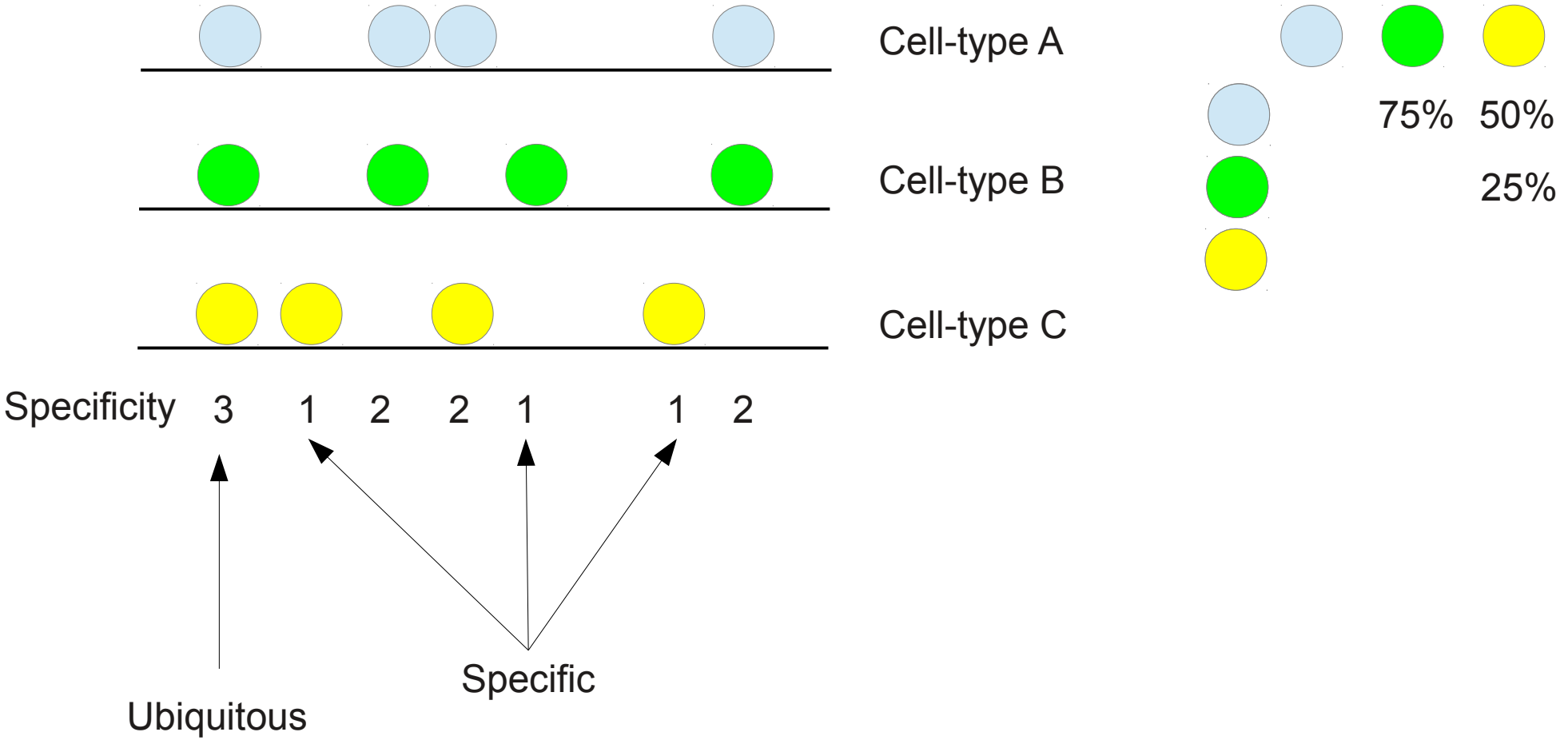


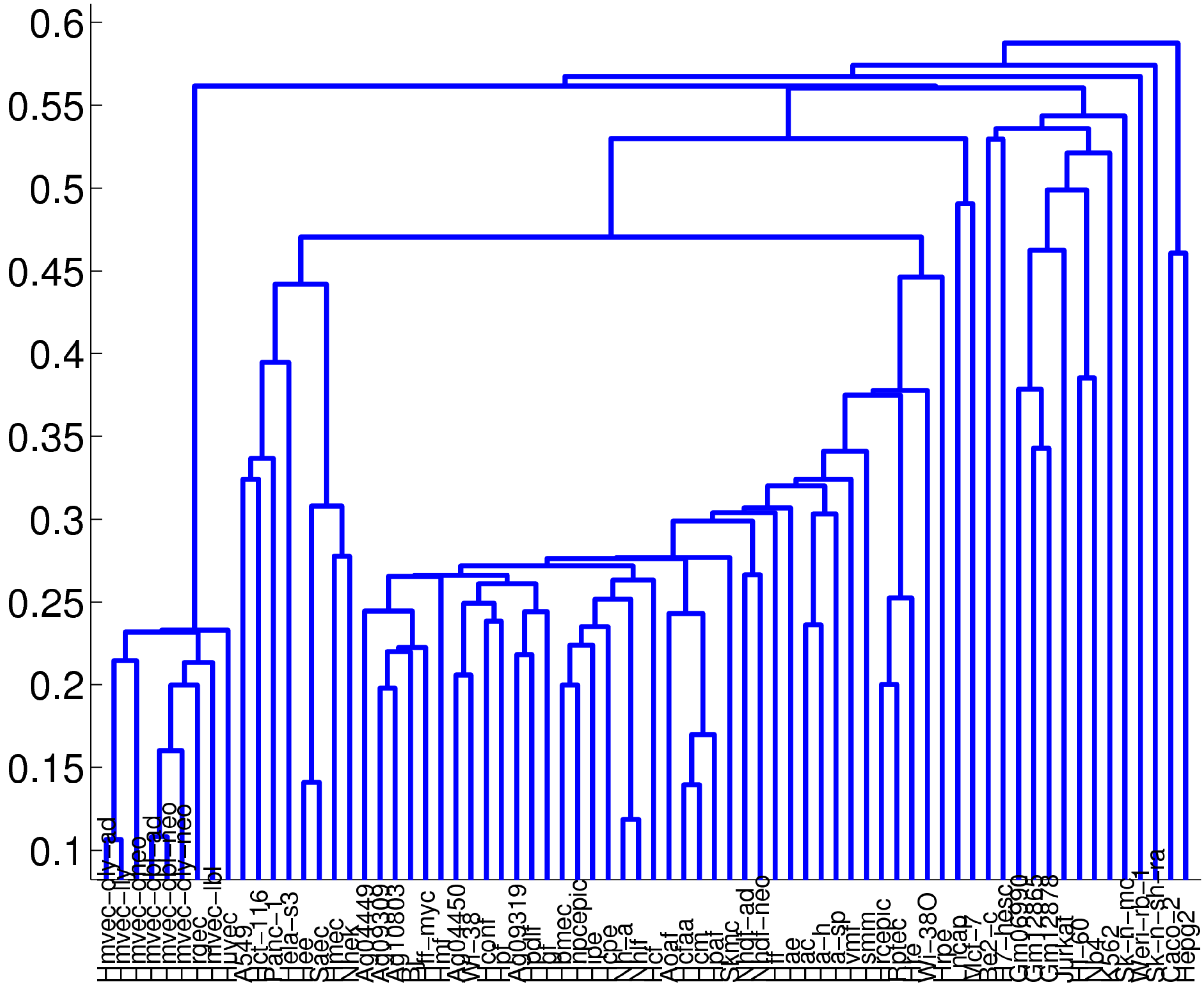
Model yields power-law ~ k^{-b}

- N = Normal(10^5, $2*10^4$) DHSs,
- C = 1,000 cell-types
- S = 72 sampled
- n = 10 repetitions
  - b = 1.35
  - Slope = -1.4

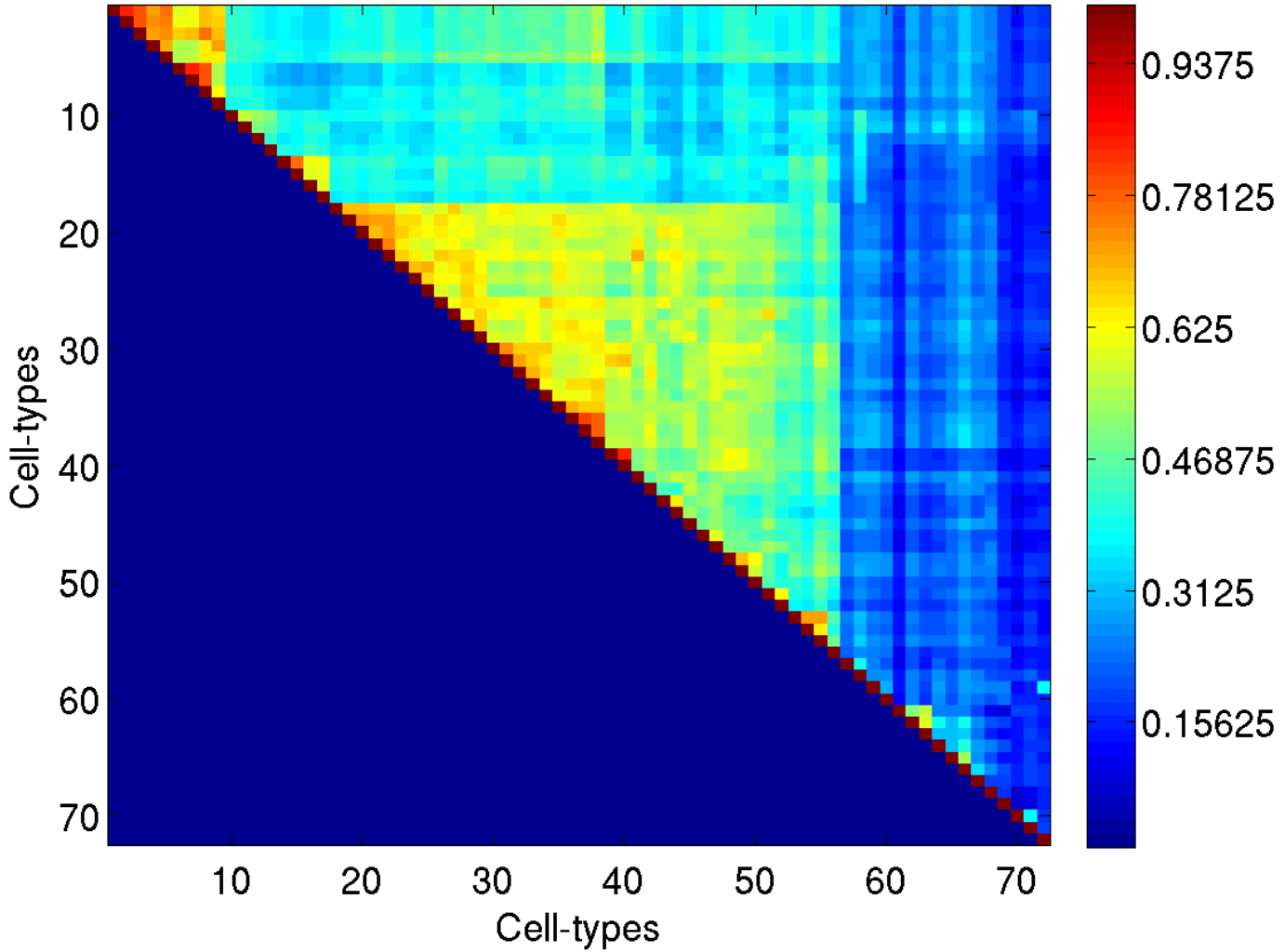


Cell-type distance can be inferred from DHS overlap

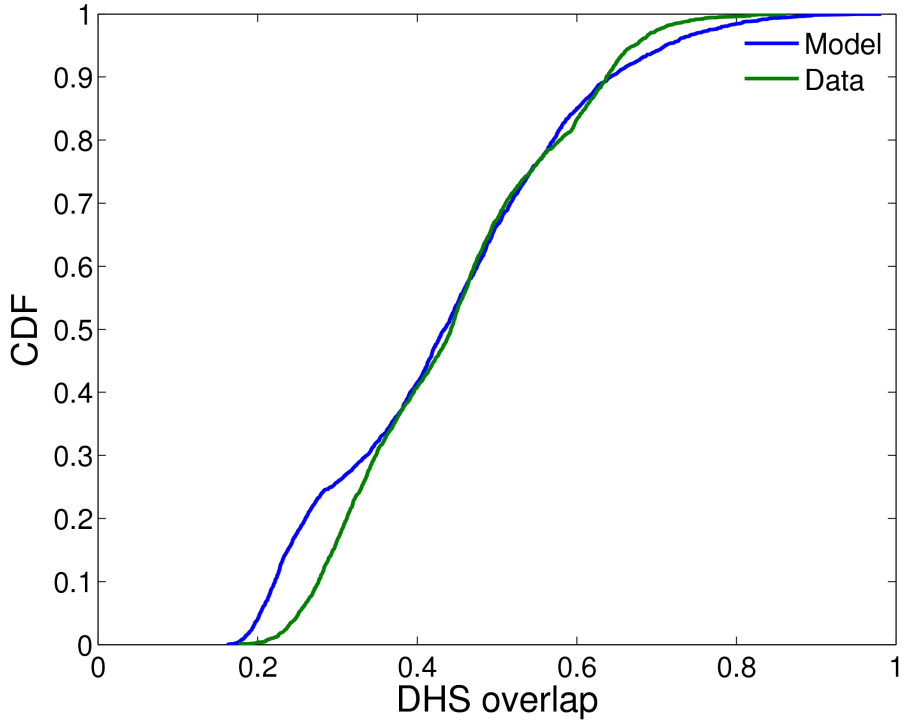



0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
Hmvec-dly-ad
Hmvec-ly
Hmvec-dneo
Hmvec-dbl-ad
Hmvec-dbl-neo
Hmvec-dly-neo
Hrgec
Hmvec-lbl
Huvec
A549
Hct-116
Panc-1
Hela-s3
Hee
Saec
Hmec
Nhek
Ag04449
Ag09309
Ag10803
Bj
Hff-myc
Hmf
Ag04450
Wi-38
Hconf
Hpf
Ag09319
Hpdlf
Hbf
Hbmec
Hnpcepic
Hipe
Hcpe
Nh-a
Nhlf
Hcf
Aoaf
Hcfaa
Hcm
Hpaf
Skmc
Nhdf-ad
Nhdf-neo
Hff
Hae
Hac
Ha-h
Ha-sp
Hvmf
Hsmm
Hrcepic
Rptec
Hre
Wi-38O
Hrpe
Lncap
Mcf-7
Be2_c
H7-hesc
Gm06990
Gm12865
Gm12878
Jurkat
Hl-60
Nb4
K562
Sk-n-mc
Weri-rb-1
Sk-n-sh_ra
Caco-2
Hepg2

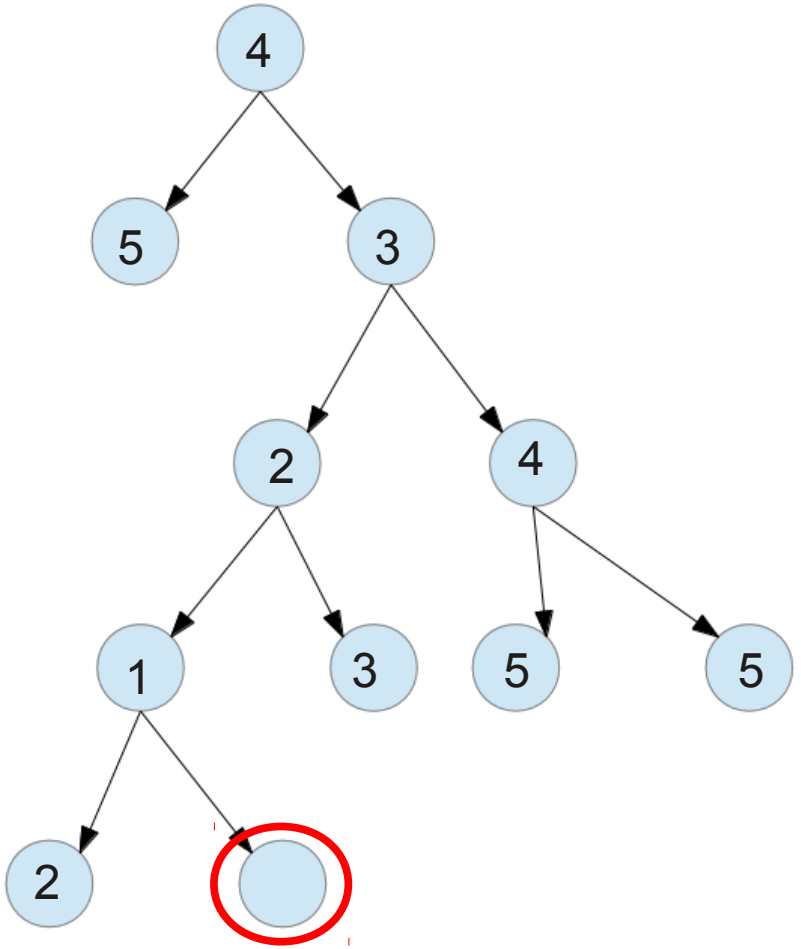
Distribution of pair-wise overlaps



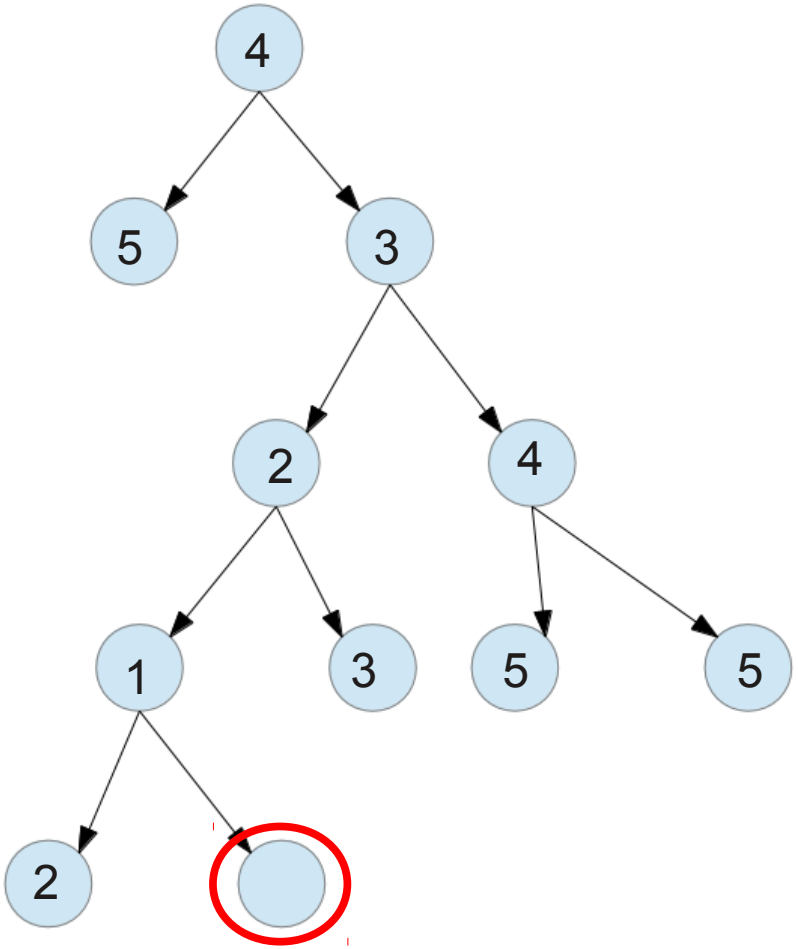
Model predicts distribution of pairwise overlaps

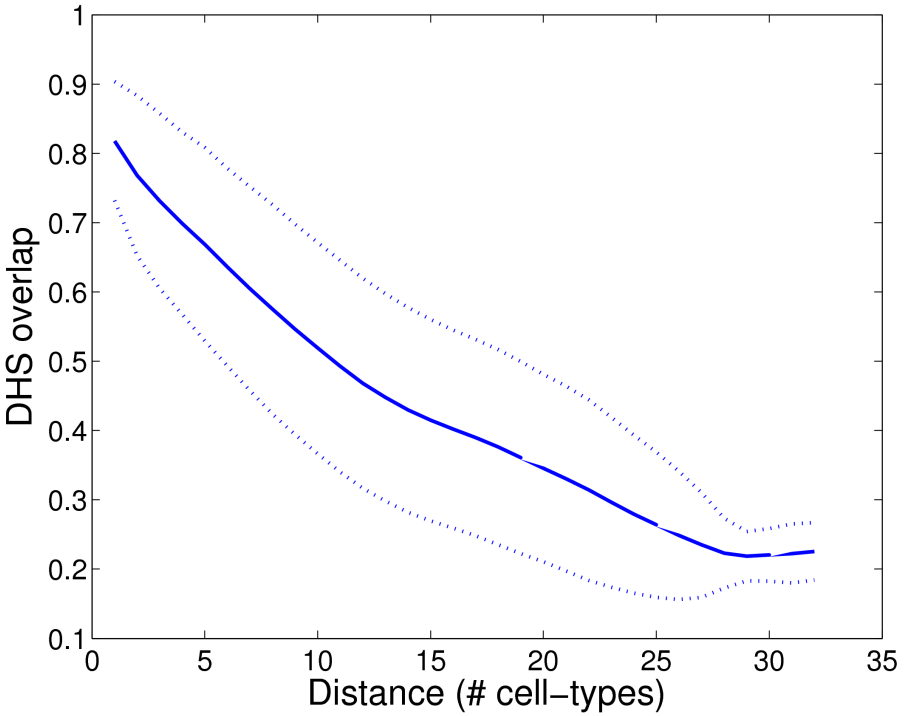


The overlap is inversely proportional to the distance

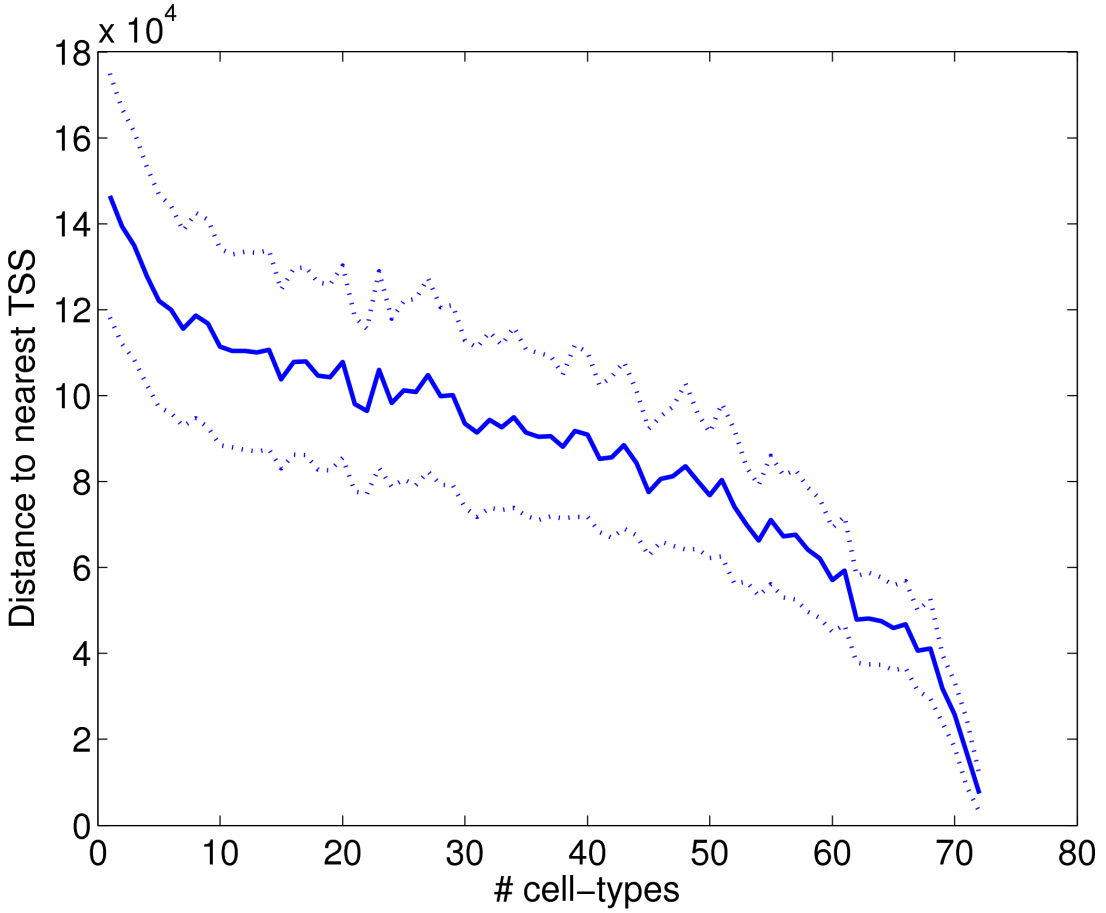


The overlap is inversely proportional to the distance



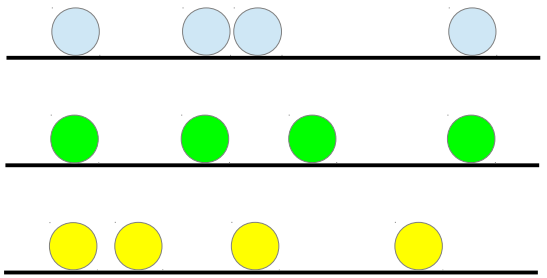


Specific DHSs are located further from known TSSs



WTF! Where are those TFs?

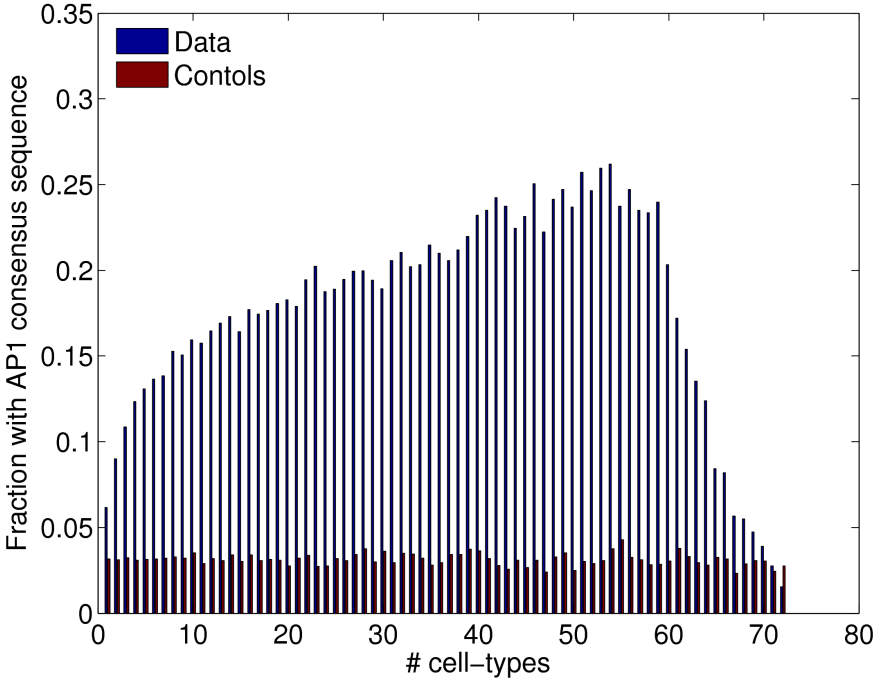
Summary

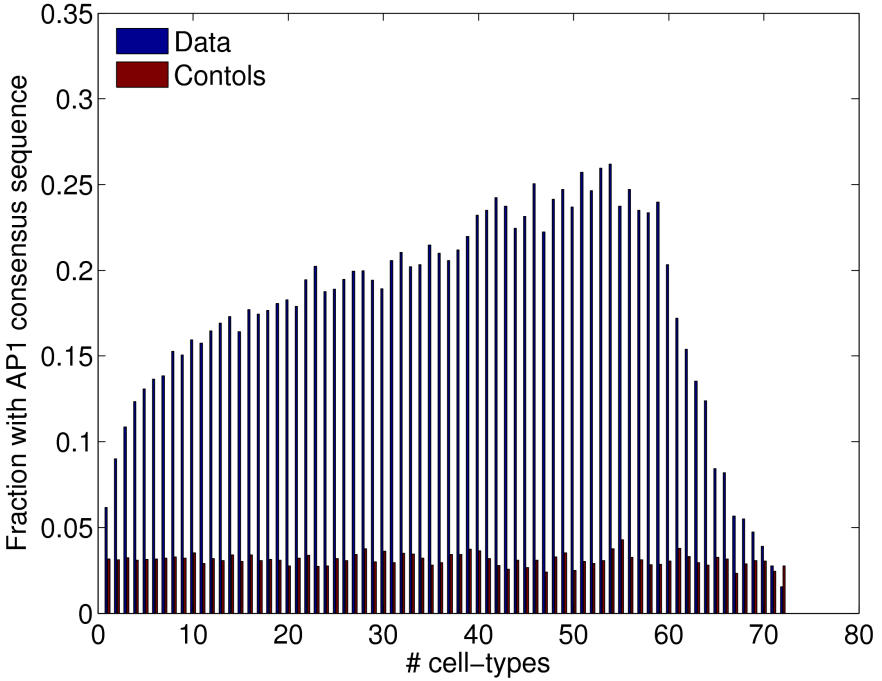


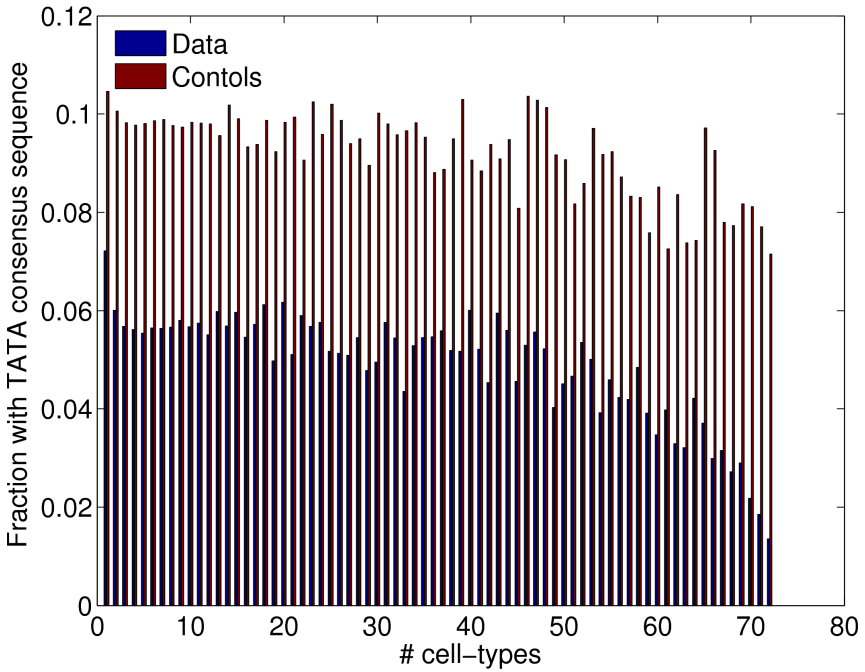
- DHS specificity follows a power-law
- Generative model
  - Parent DHS used as template
  - Selection rule: $P(\text{keep DHS}) = \dfrac{k^b}{1+k^b}$
  - Selective advantage *b*
- Specificity related to function

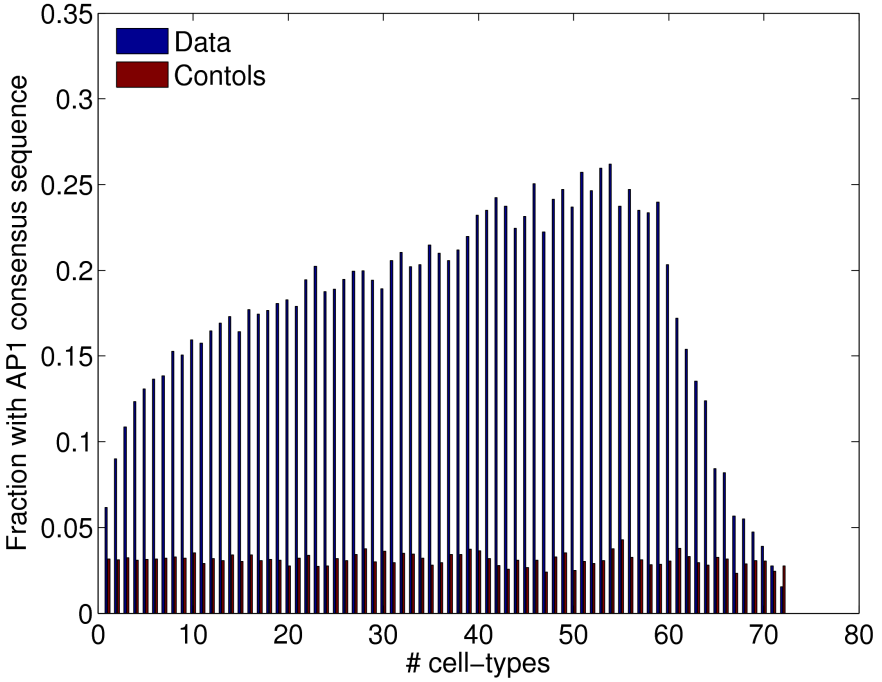
Acknowledgements

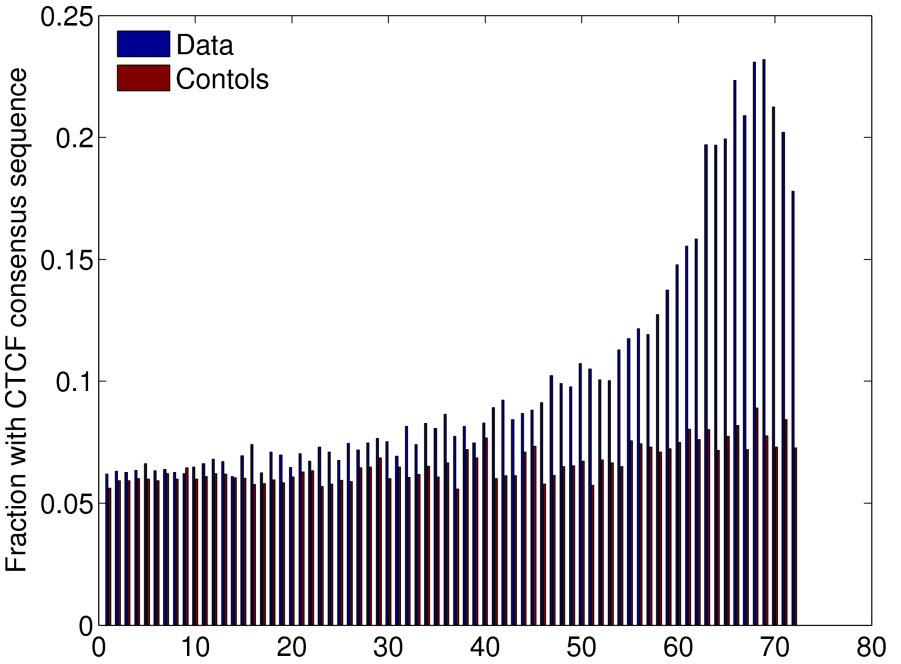
- Gabriel Kreiman
- Wui Ip
- Ben Tsuda
- Enrique Tobis

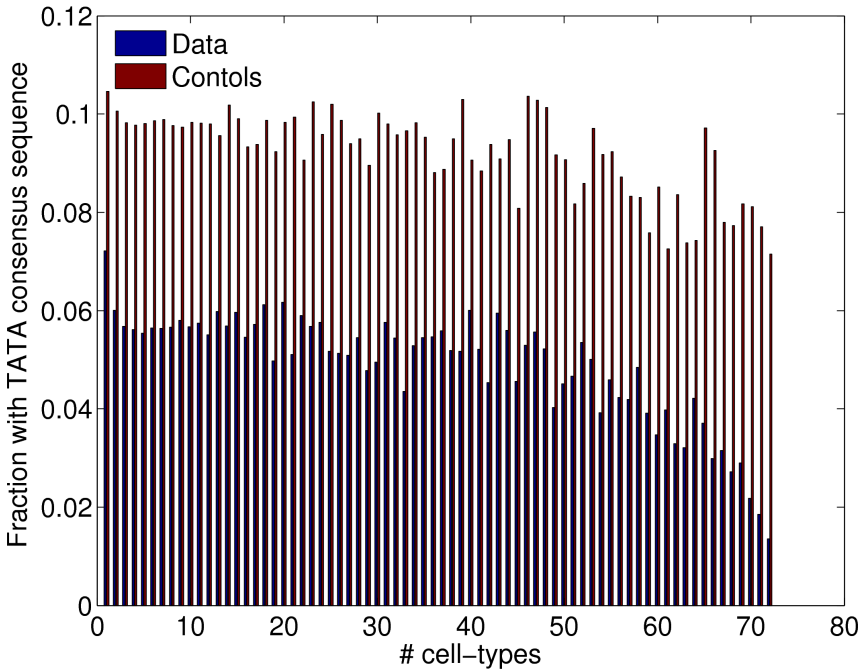

Data
Contols
Fraction with AP1 consensus sequence
cell−types
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80

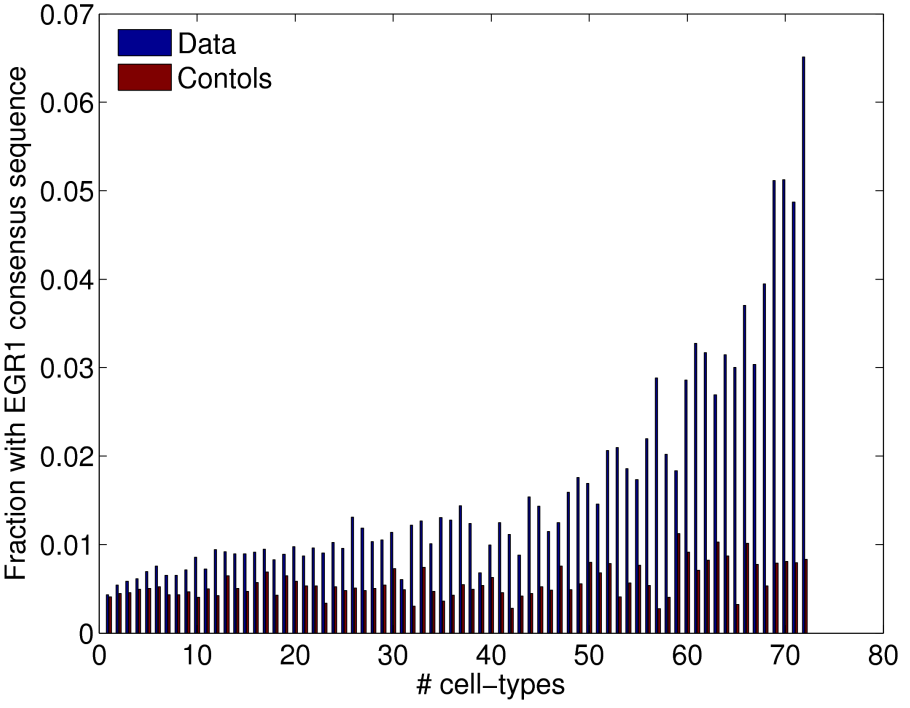

Data
Contols
Fraction with AP1 consensus sequence
cell−types
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70
80


Data
Contols
Fraction with TATA consensus sequence
cell−types
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80

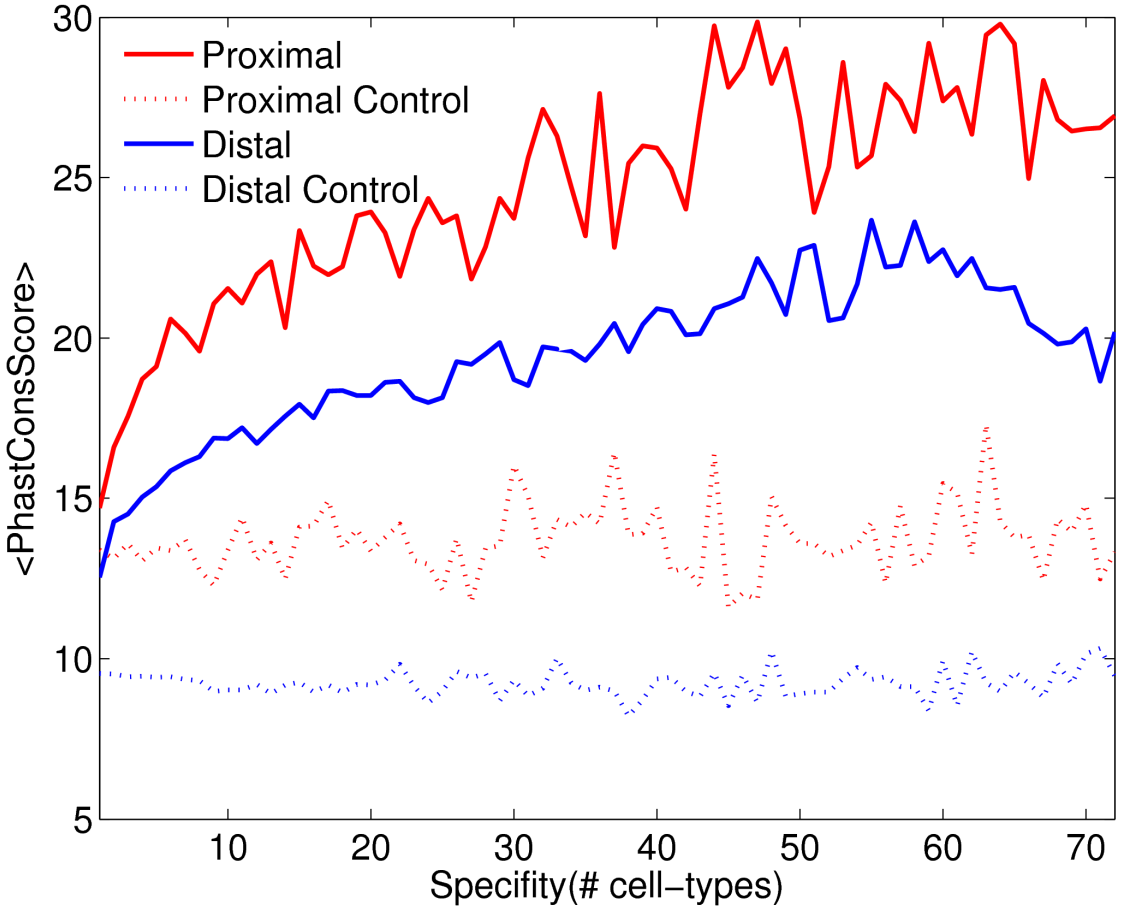

Data
Contols
Fraction with AP1 consensus sequence
cell−types
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0
10
20
30
40
50
60
70
80


Data
Contols
Fraction with CTCF consensus sequence
0
0.05
0.1
0.15
0.2
0.25
0
10
20
30
40
50
60
70
80

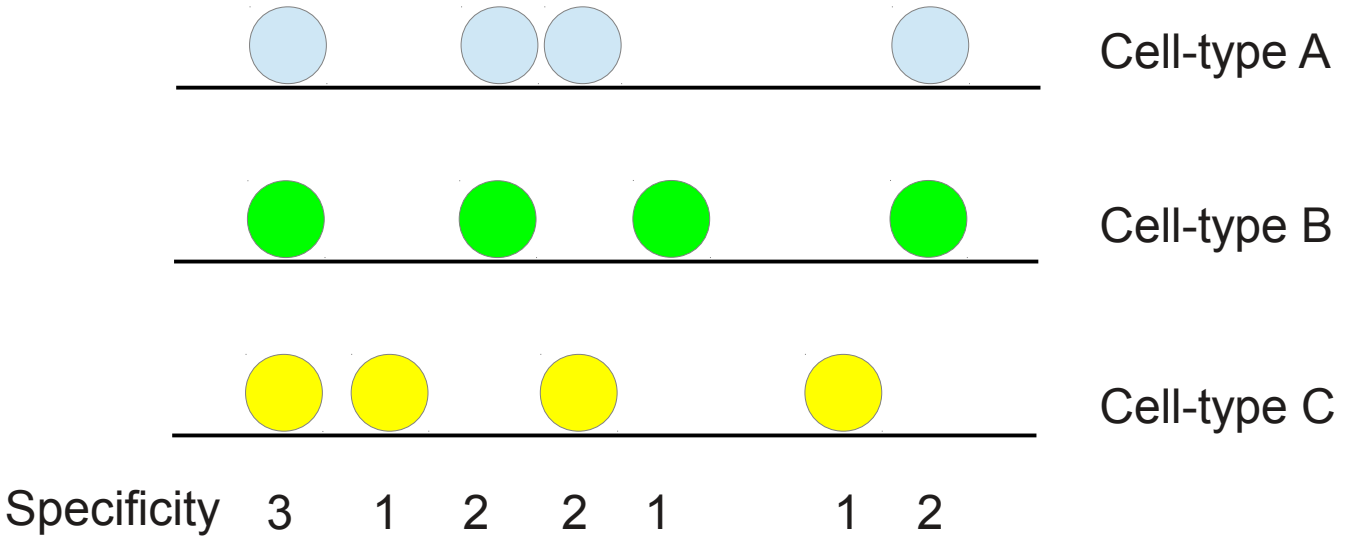

Data
Contols
Fraction with TATA consensus sequence
cell−types
0
0.02
0.04
0.06
0.08
0.1
0.12
0
10
20
30
40
50
60
70
80


Data
Contols
Fraction with EGR1 consensus sequence
cell−types
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
10
20
30
40
50
60
70
80

Specific DHSs are less conserved

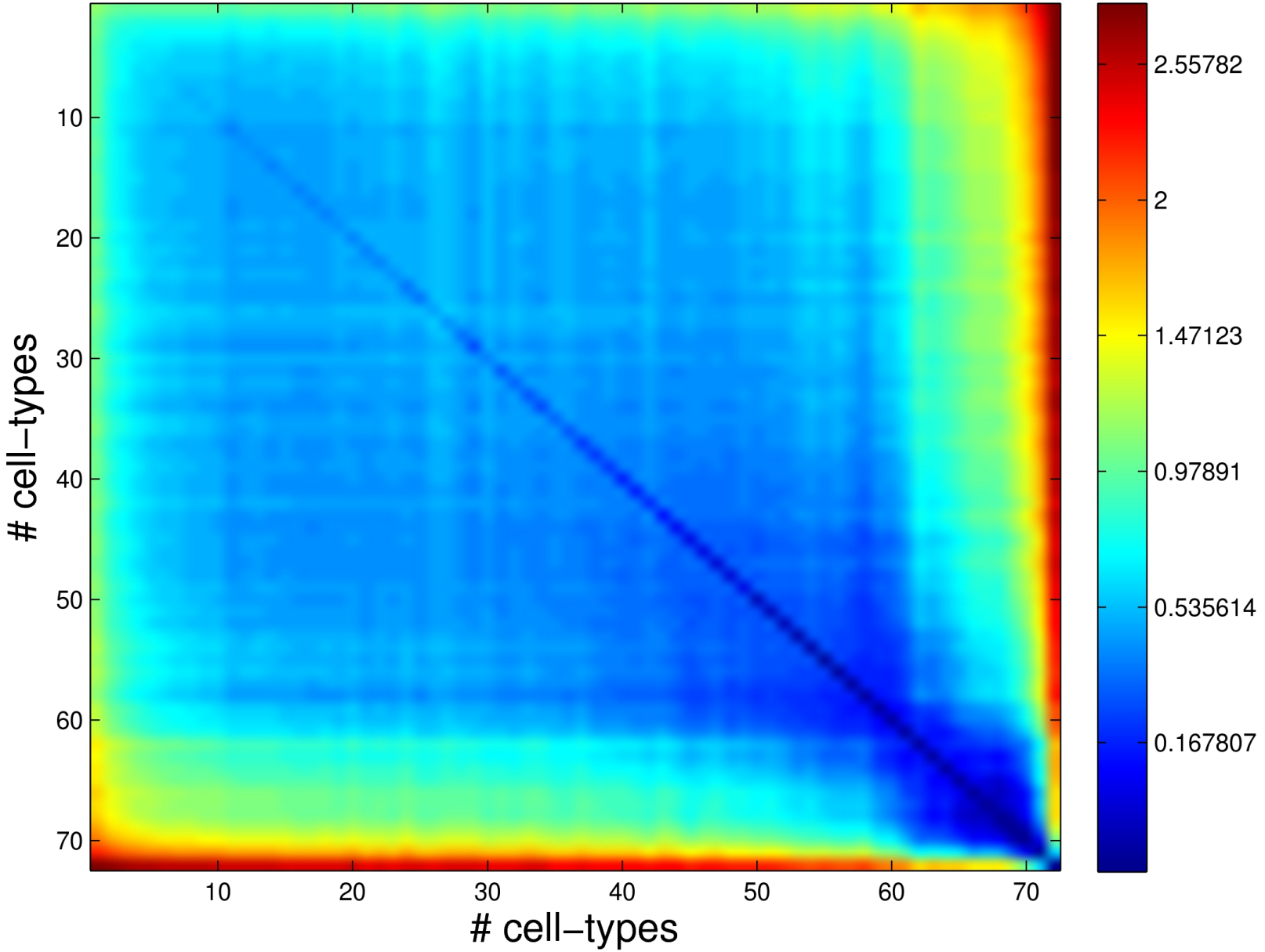


Sequence similarities of DHSs



		Specificity	
Word	1	2	3
AAAAAA	.001	.0002	.0005
AAAAAC	.002	.001	.0043
AAAAAG	…….		
AAAAAT			
AAAACA			
…..			

Difference between sequence distributions

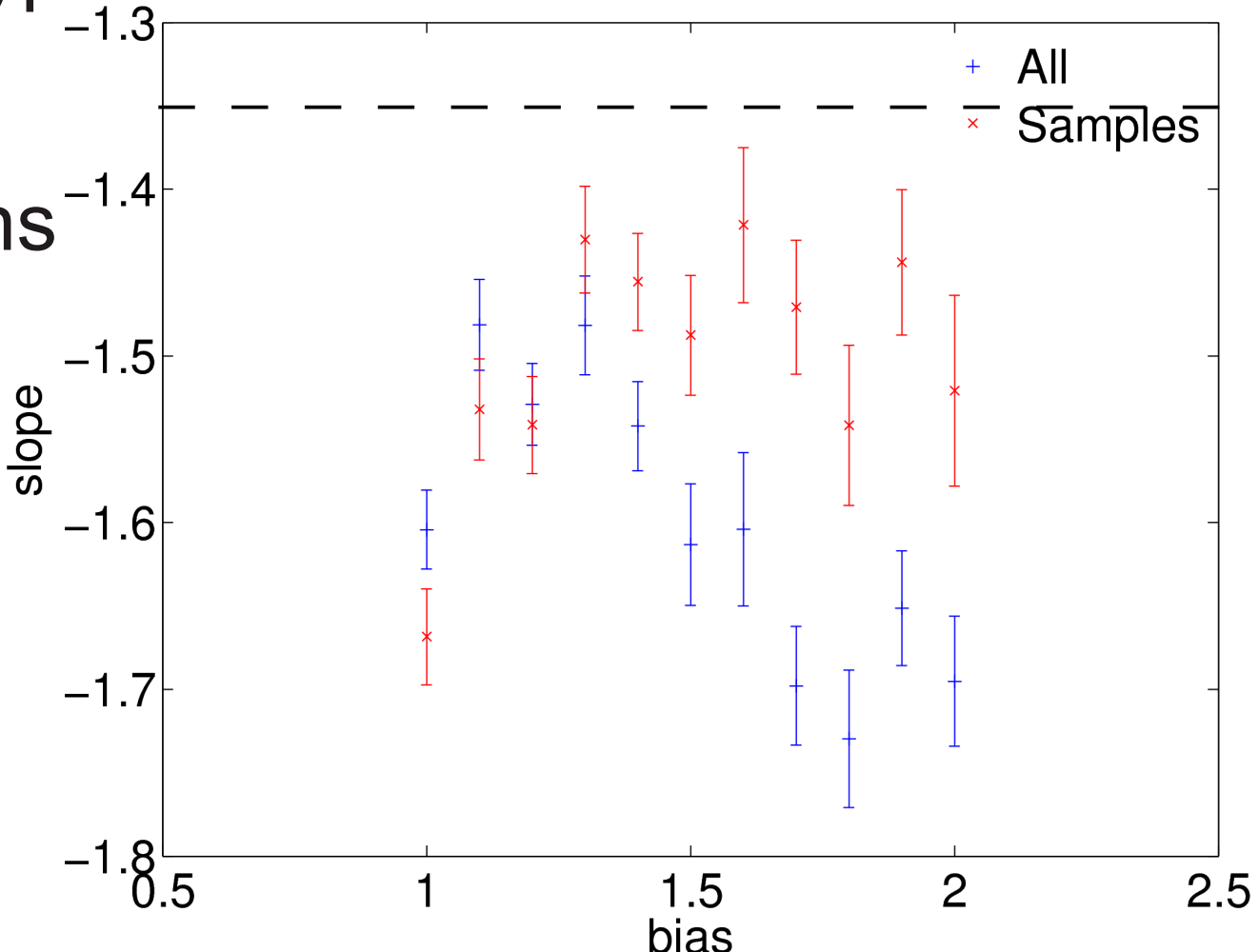


Functional role can be predicted by histone modifications and TFs

Ubiquitous DHSs have higher GC

Monte Carlo Simulations

- N = 100,000 DHSs
- C = 1,000 cell-types
- S = 72 sampled
- n = 10 repetitions



Random branching process

$$P(Specificity = k) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{Q(k,i)}_{\text{Existing DHS having specificity k}} + \underbrace{R(k,i)}_{\text{New DHS having specificity k}}$$

Random branching process

$$P(Specificity = k) \approx \frac{1}{n} \sum_{i=1}^{n} S(k) = S(k)$$

- Assume that tree is very large

Random branching process

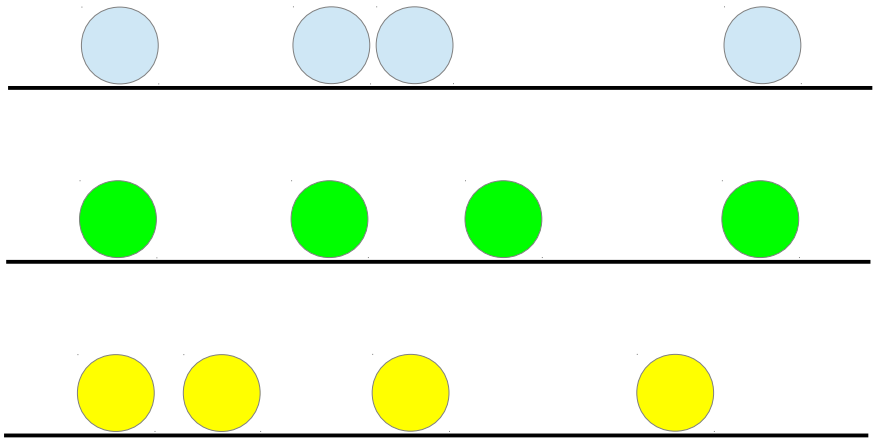
$$P(Specificity = k) \approx \frac{1}{n} \sum_{i=1}^{n} S(k) = S(k)$$

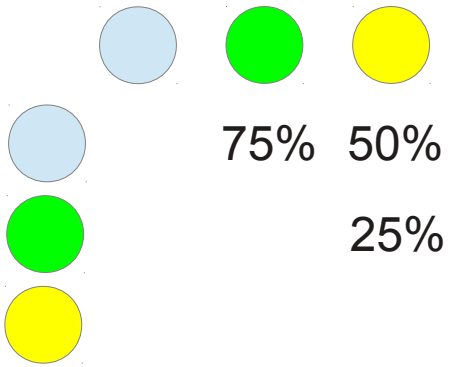
$$S(k) = \frac{1}{1+k^b} \prod_{j=1}^{k} \frac{j^b}{1+j^b} = \frac{k^b (k-1)^b \dots 2^b}{(1+k^b)^2 (1+(k-1)^b) \dots (1+2^b) 2} \geq \frac{1}{1+k^b} \geq k^{-b}$$

Signal strength

Many near-ubiquitous DHSs are insulators

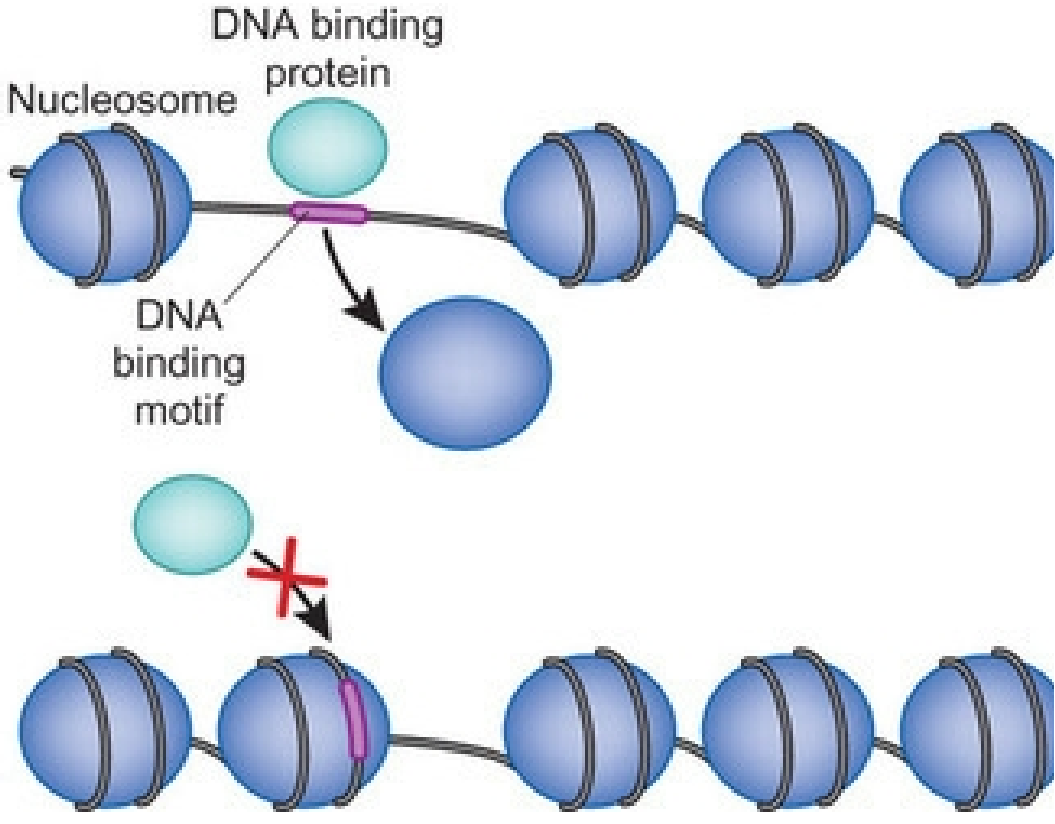
DHS overlap





What determines the location of DHSs?

- Identify distribution of DHSs
  - Functional characteristics
  - Generative model



Uniform distribution of specificities

- ~100,000 DHSs per cell-type
  - ~1,000 of each specificity