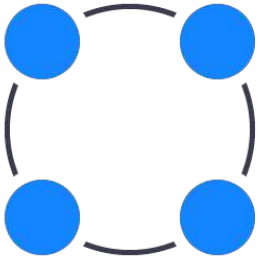


ИНСТИТУТ
БИОИНФОРМАТИКИ

Enhancement of 5-mer based statistical model for SHMs

Project follower and implementer:

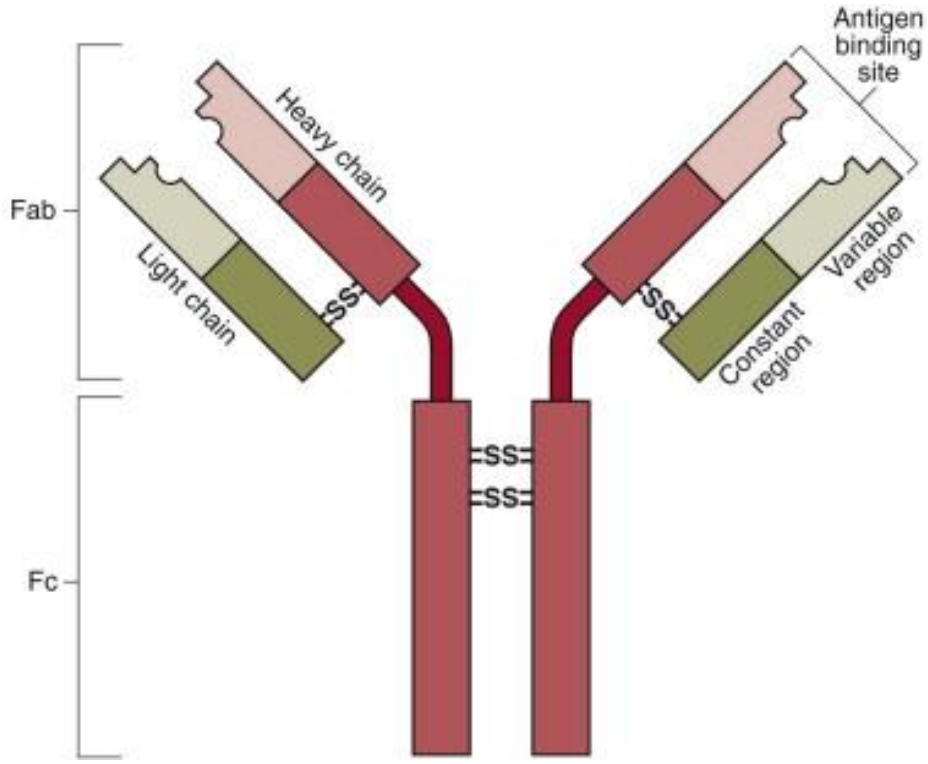
Oksana Ayzsilnieks

Supervisors: Andrey Bzikadze
Yana Safonova

Saint-Petersburg State University Laboratory for Algorithmic Biology

Saint-Petersburg, 2016

Structure of Immunoglobulins



Structure of an immunoglobulin . Fab, antigen binding fragment; Fc, constant fragment.

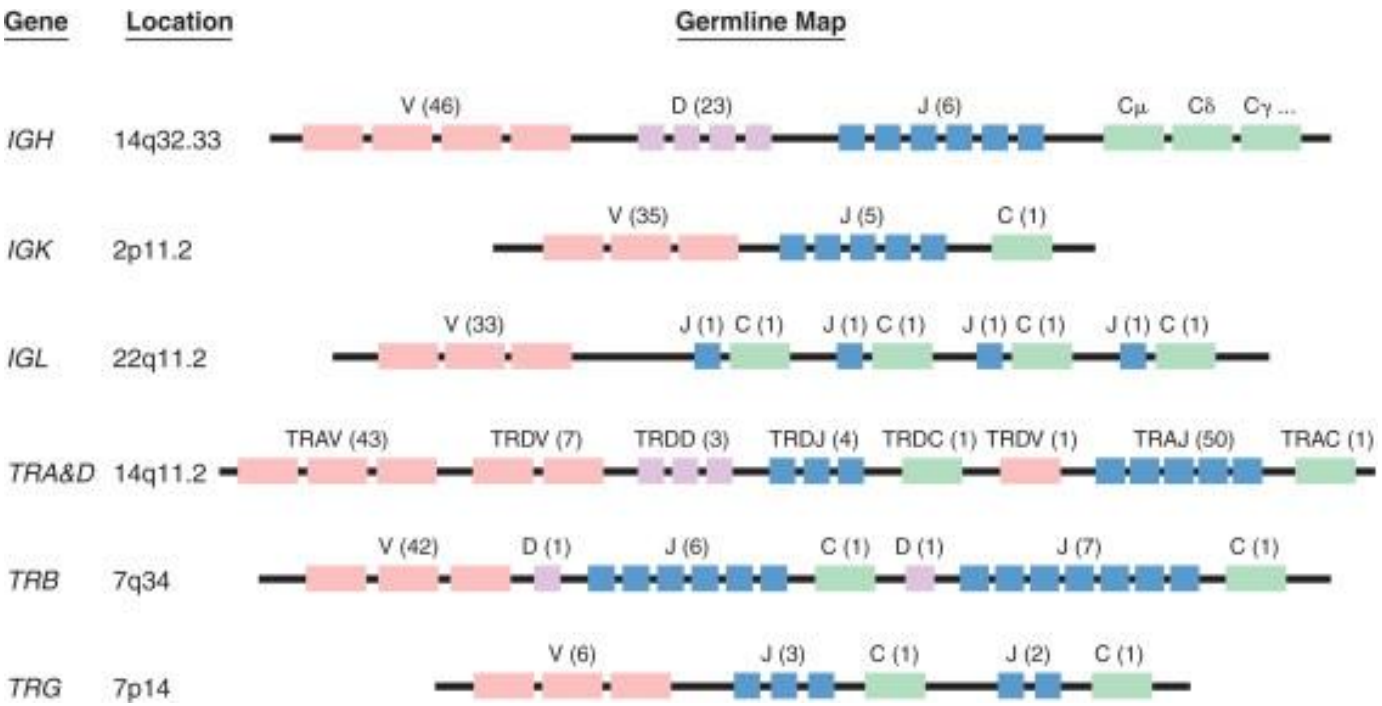
Immunoglobulins
Gaw, Allan, MD PhD FRCPath FFPM PGCertMedEd, Clinical Biochemistry: An Illustrated Colour Text, 26, 52-53

Copyright © 2013 2013, Elsevier Ltd. All rights reserved.

ELSEVIER

ClinicalKey®

Loci of gene of Immunoglobulin



Structural similarities exist between the immunoglobulin (IG) and the T-cell receptor (TR) loci as shown in these maps of the germline configuration. During lymphocyte ontogeny, unique coding sequences are produced through rearrangement of the variable, diversity, joining, and constant regions of each locus. The number of functional alternatives is shown in parentheses for each cluster of elements. Note: T-cell receptor δ (TRD) segments are spliced out during T-cell receptor α (TRA) gene rearrangement. IGH, IG heavy chain gene; IGK, IG kappa light chain gene; IGL, IG lambda light chain gene; TRA, TRB, TRG and TRD, T-cell receptor alpha, beta, gamma and delta genes, respectively; V, variable; D, diversity; C, constant.

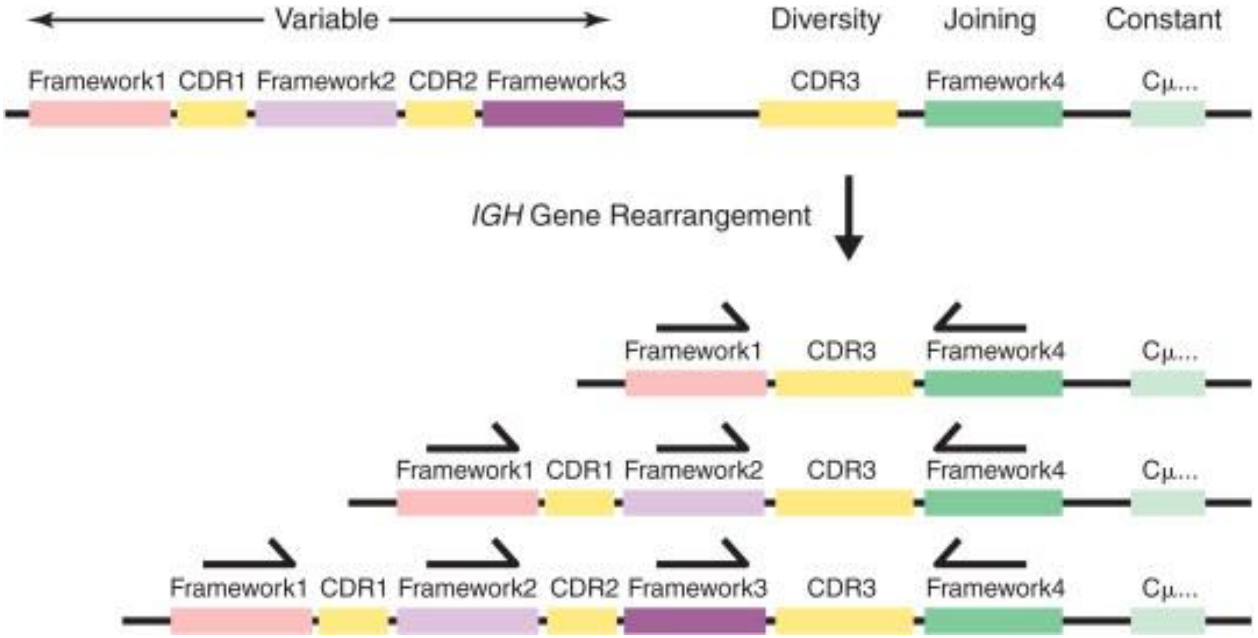
Immunoglobulin and T-Cell Receptor Gene Rearrangement
Gulley, Margaret L., Cell and Tissue Based Molecular Pathology: A Volume in the Foundations in Diagnostic Pathology Series, Chapter 13, 127-134

Copyright © 2009 Copyright © 2009 by Churchill Livingstone, an imprint of Elsevier Inc.

ELSEVIER

ClinicalKey®

Ig gene has FRamework and Complementary Determinig Regions



Polymerase chain reaction (PCR) assays can be designed to amplify across the rearranged immunoglobulin heavy chain gene. In the germline configuration depicted at the top, the variable and joining regions are too far apart (> 1 kilobase) for reliable amplification. However, as depicted below, a rearranged gene juxtaposes the variable and joining segments so that PCR amplification across the spliced variable, diversity, and joining segments is feasible. To maximize detection of the many alternative variable, diversity, and joining variants while minimizing the number of primers required, a cocktail of consensus primers (shown as half arrows) targets each of four framework regions that are relatively well conserved compared to the interspersed complementary determining regions (CDR1, CDR2, and CDR3), which are quite mutation prone.

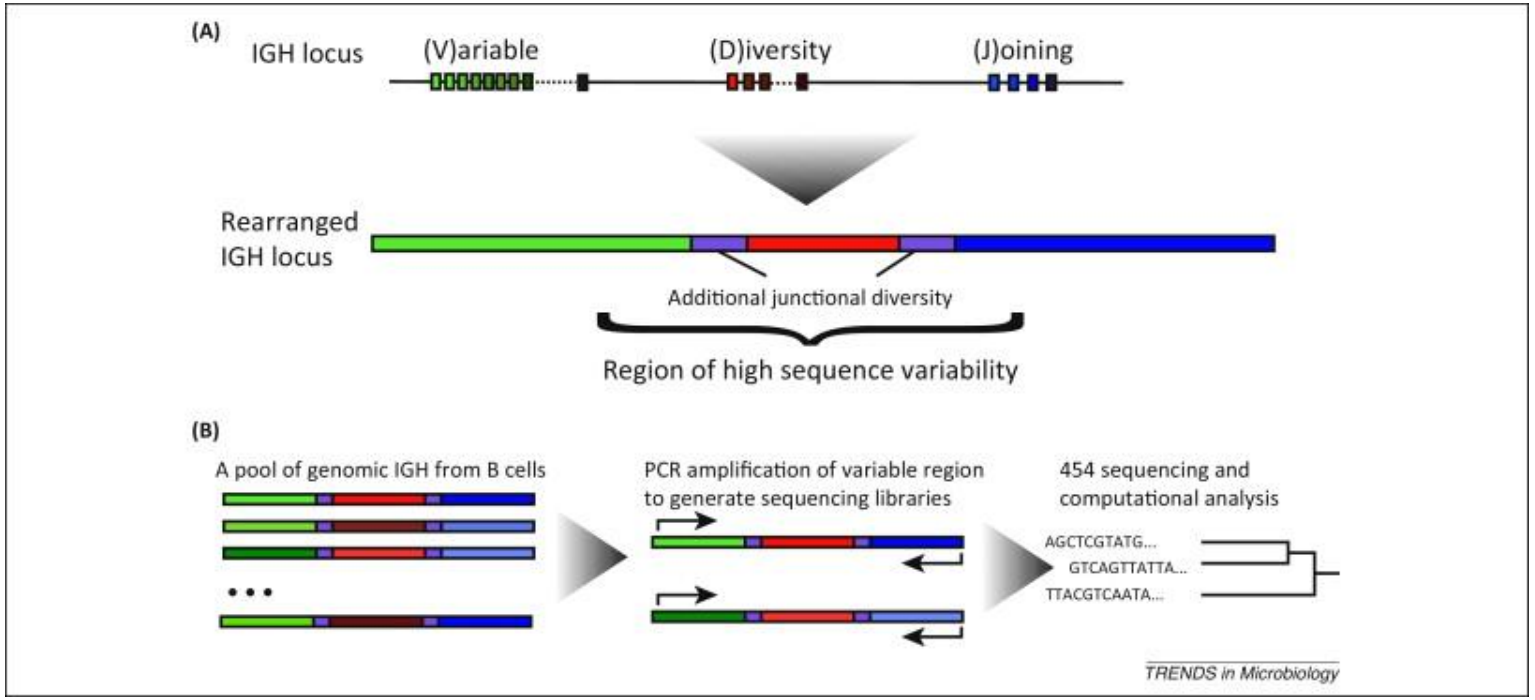
Immunoglobulin and T-Cell Receptor Gene Rearrangement
Gulley, Margaret L., Cell and Tissue Based Molecular Pathology: A Volume in the Foundations in Diagnostic Pathology Series, Chapter 13, 127-134

Copyright © 2009 Copyright © 2009 by Churchill Livingstone, an imprint of Elsevier Inc.

ELSEVIER

ClinicalKey®

IGH locus rearrangment



Probing immunoglobulin diversity and clonal structure using high-throughput sequencing. (A) Combinatorial joining of different V, D, and J gene segments, the removal of and/or addition of non-templated nucleotides at junctions, together with somatic hypermutation at the immunoglobulin heavy-chain (IGH) locus generate tremendous IGH-sequence diversity (and therefore antigen recognition diversity) among different B cells. (B) To analyze the B cell repertoire by sequencing, libraries are generated via PCR amplification of the variable IGH region. These libraries are subjected to 454 pyrosequencing, where a random subset of the amplified IGH fragments is sequenced. As B cell clones have largely unique IGH sequences, clustering analysis of the sequences can reveal the clonal structure of the B cells in the sample.

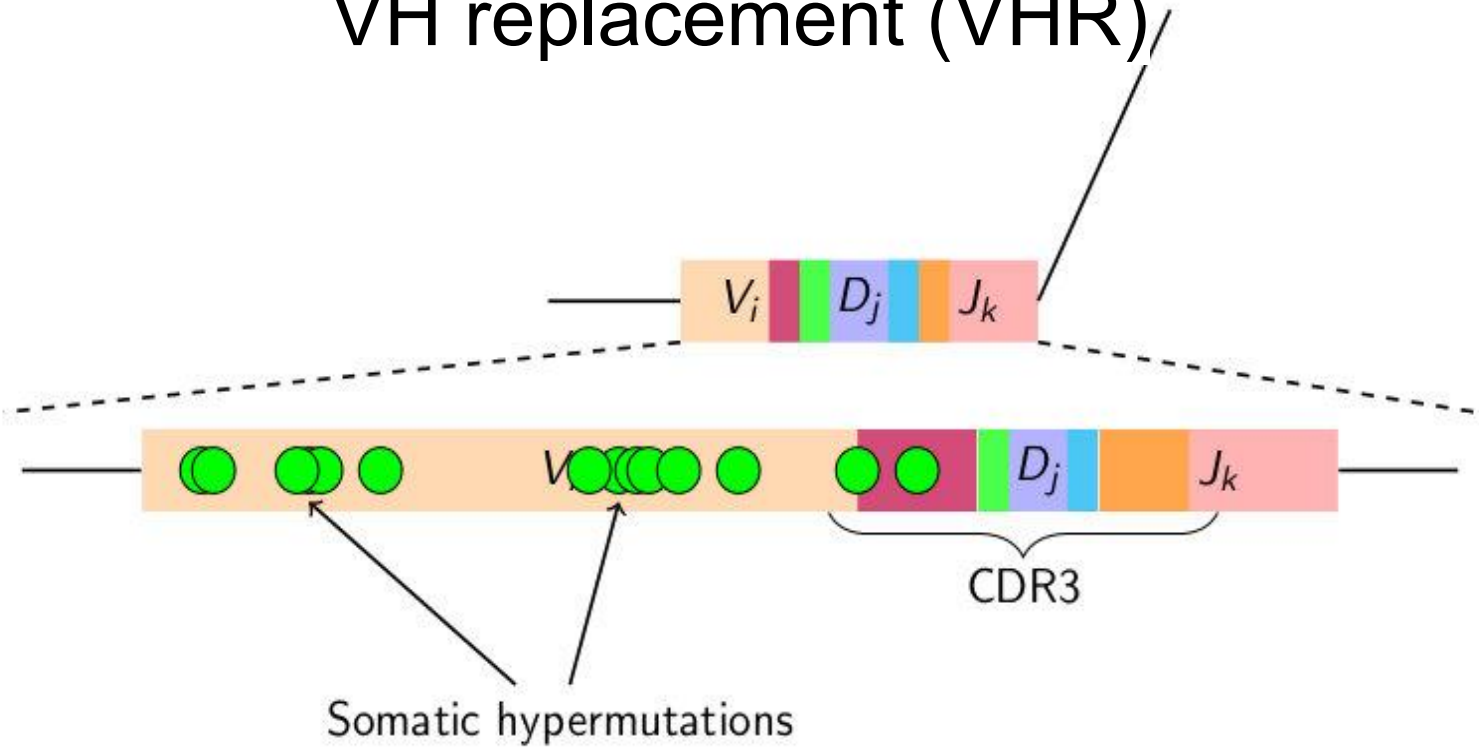
Random yet deterministic: convergent immunoglobulin responses to influenza
Martins, Andrew J., Trends in Microbiology, Volume 22, Issue 9, 488-489

Copyright © 2014

ELSEVIER

ClinicalKey®

Antibodies can be further revised and diversified to drive the affinity maturation through receptor editing of the light chain, **somatic hypermutation**, gene conversion, and VH replacement (VHR)



Yana Safonova' group collaborators from Yale (2013):

Front Immunol. 2013; 4: 358.
Published online 2013 Nov 15. doi: 10.3389/fimmu.2013.00358
PMCID: PMC3828525

Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data

Gur Yaari,[1,2] Jason A. Vander Heiden,[3] Mohamed Uduman,[2] Daniel Gadala-Maria,[3] Namita Gupta,[3] Joel N. H. Stern,[4,5] Kevin C. O'Connor,[4,6] David A. Hafler,[4,7] Uri Laserson,[8] Francois Vigneault,[9] and Steven H. Kleinstein[2,3,*]

Author information ► Article notes ► Copyright and License information ►

Yale's 5-mer SHM models

Targeting model

Expanding **spots**

5-mer	Mutability
...	...
GCCTC	0.12
GCGAC	0.16
ACACT	0.48
AGCTA	3.17
...	...

Substitution model

Expanding AID **locality**

5-mer	A	C	G	T
...	...	...	...	...
ACAAC	0	.24	.48	.29
GGCGT	.22	0	.12	.65
CCGTC	.35	.52	0	.13
TCTAC	.31	.54	.14	0
...	...	...	...	...

5-mers frequency matrices

Computation 5-mers frequency matrix inspired by Yaari et al. basing on **V gene alignment**

5-mer	A	C	G	T
CTATG	16738	689	194	736
CTCAA	37	5092	176	31
GGGGG	221	135	38075	244
ACTCT	54	170	178	19990

Parameters for each 5-mer:

Mut(CTATG) =
(689 + 194 + 736) / 16738

Subst(CTATG) =
(689, 194, 736) / 1619

Novel approach to SHM modeling

5-mer	A	C	G	T
CTATG	16738	689	194	736
CTCAA	37	5092	176	31
GGGGG	221	135	38075	244
ACTCT	54	170	178	19990

Parameters for each 5-mer:

Mut(CTATG) =
(689 + 194 + 736) / 16738

Subst(CTATG) =
(689, 194, 736) / 1619

For each 5-mer the corr. Row:

1. Binomial(Mut) — how many mutations occurred.
2. Multinomial(Subst) — what were the mutations.

Assumption: the parameters are random themselves.
Let Mut / Subst parameters have distribution P / Q.

What are P & Q ?

What have we done on the moment of previous perfomance

1. We tried the code that count the 5-mer frequency matrices
2. We pondered the situation, when mutations occur in FR-CDR sharing 5-mers
3. Studying the literature, we found, that diversification of anti-HIV antibodies occur not only with somatic hypermutation, but with the VH replacement

Further goals of the current project on the moment of previous perfomance

1. The better comprehension of the project's story..**DONE AS POSSIBLE**
2. Enhancement of the model by considering **FR/CDR position** of a 5-mer.. **DONE**
3. Validation of model on repertoires with various **specificity**: **HIV**, autoimmune disorders and, especially, CLL..

Poor quality of HIV data, so VALIDATION ON "HEALTHY(age)" AND "FLU" DATASETS

k-mers' coverage through CDR (left) and FR (right) zones in HIV data

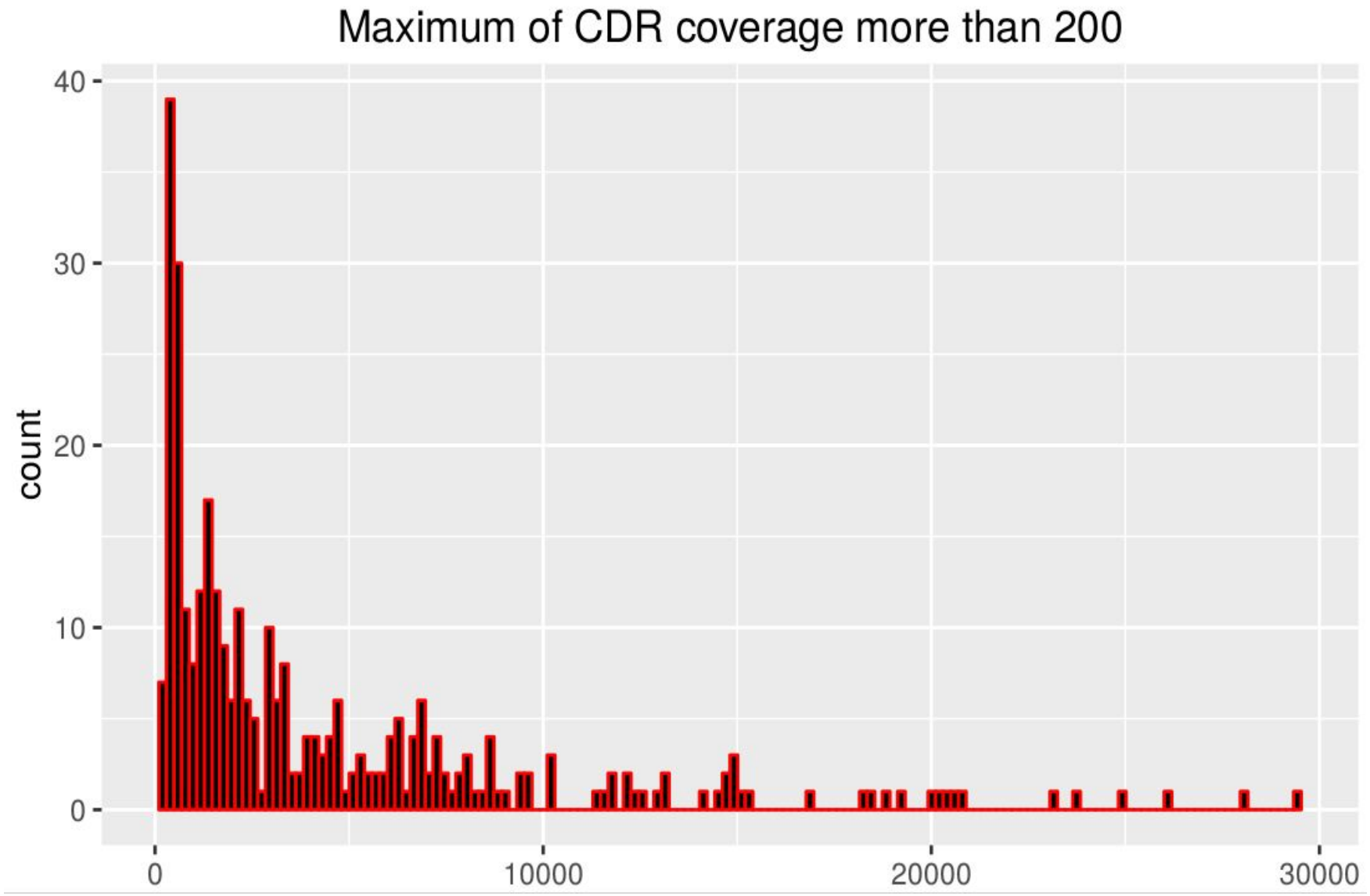
```
;A;C;G;T
AAAAA;0;0;0;0
AAAAC;35;1;0;1
AAAAG;36;0;2;0
AAAAT;0;0;0;0
AAACA;90;1;12;0
AAACC;0;0;0;0
AAACG;0;0;0;0
AAACT;32;1;0;0
AAAGA;0;0;0;0
AAAGC;1435;0;4;0
AAAGG;0;0;0;0
AAAGT;0;0;0;0
AAATA;3822;1;255;1
AAATC;0;0;0;0
AAATG;0;0;0;0
AAATT;0;0;0;0
AACAA;0;274;0;1
AACAC;2;213;53;3
AACAG;0;1270;5;26
AACAT;0;0;0;0
AACCA;0;0;0;0
AACCC;0;508;0;106
AACCG;0;0;0;0
```

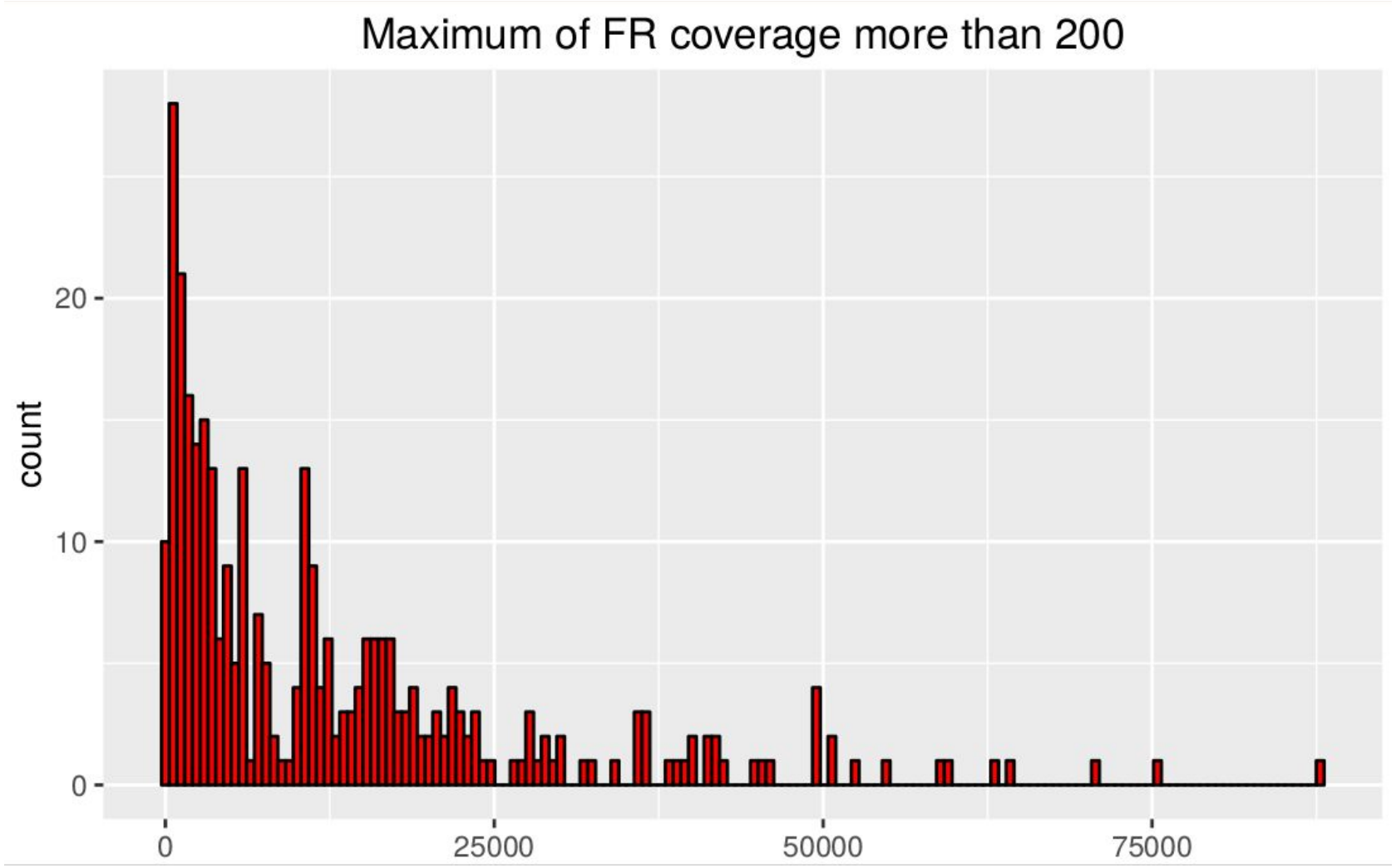
```
;A;C;G;T
AAAAA;97;1;5;0
AAAAC;861;0;0;0
AAAAG;687;2;9;0
AAAAT;13;0;0;0
AAACA;38;0;2;0
AAACC;770;0;2;0
AAACG;0;0;0;0
AAACT;1722;3;10;95
AAAGA;2472;0;1;2
AAAGC;1014;0;15;0
AAAGG;1015;0;18;0
AAAGT;2;0;0;0
AAATA;1;0;0;0
AAATC;754;1;3;0
AAATG;7711;296;563;83
AAATT;0;0;0;0
AACAA;0;6;0;0
AACAC;4;6609;59;200
AACAG;905;4900;1199;177
AACAT;1;454;0;4
AACCA;0;93;8;13
AACCC;0;105;1;4
AACCG;0;430;12;145
AACCT;1;1;0;0
AACGA;0;0;0;0
AACGC;1;1162;1;20
AACGG;0;0;0;0
```

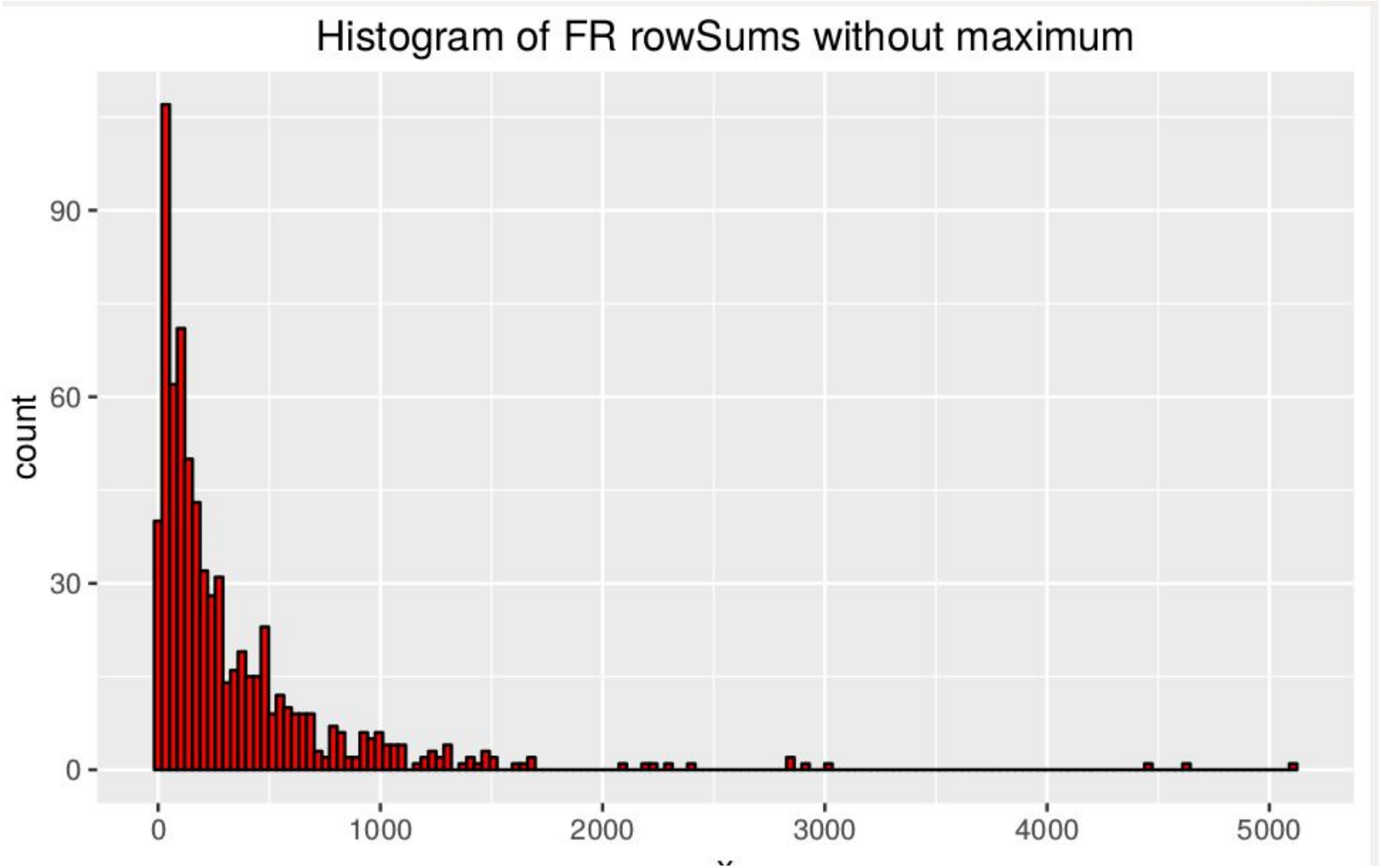
k-mers' coverage through CDR (left) and FR (right) zones in age data

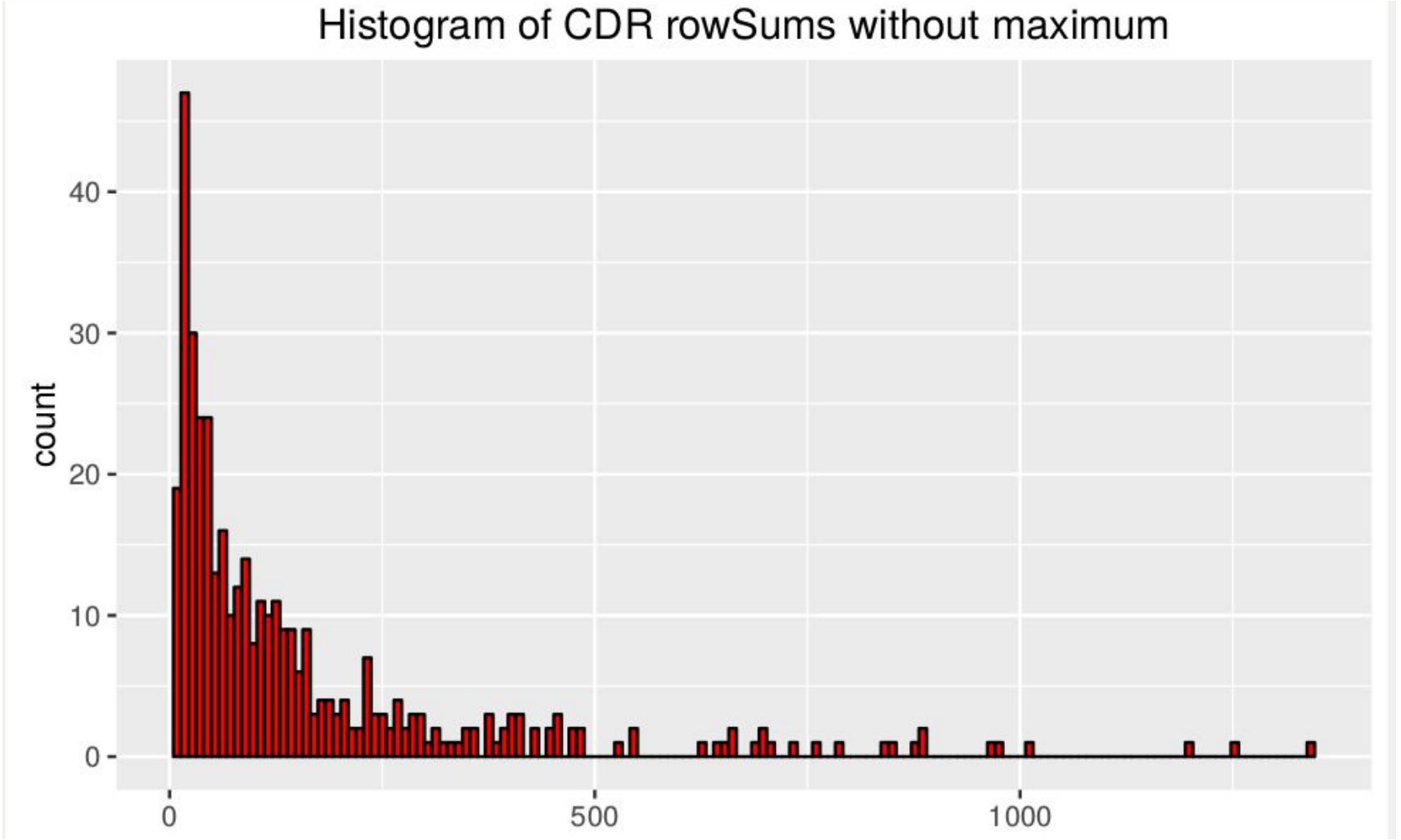
```
AGTTC;0;0;0;0
AGTTG;3;19;11;1314
AGTTT;0;0;0;4
ATAAA;5784;88;128;36
ATAAC;1;0;0;0
ATAAG;484;4;14;0
ATAAT;441;8;3;7
ATACA;6001;51;15;57
ATACC;2056;18;49;15
ATACG;23;0;0;0
ATACT;539;2;4;6
ATAGA;0;0;0;0
ATAGC;3917;7;121;8
ATAGG;1126;15;12;15
ATAGT;8424;84;259;14
ATATA;2658;59;30;36
ATATC;3005;132;76;206
ATATG;4267;107;16;146
ATATT;0;0;0;0
ATCAA;45;5112;44;38
ATCAC;0;9;0;0
ATCAG;72;8569;55;103
ATCAT;89;11437;59;55
```

```
AAAAC;2378;7;61;8
AAAAG;3999;38;62;14
AAAAT;549;4;18;1
AAACA;560;8;12;0
AAACC;2341;20;63;18
AAACG;5;0;0;0
AAACT;8071;109;95;20
AAAGA;2772;4;14;1
AAAGC;2669;24;23;15
AAAGG;4498;22;141;21
AAAGT;401;0;49;1
AAATA;412;1;19;1
AAATC;3643;11;103;17
AAATG;17030;131;154;132
AAATT;9;0;0;0
AACAA;0;163;1;0
AACAC;72;11325;71;291
AACAG;66;16804;79;319
AACAT;14;1970;9;68
AACCA;86;11199;199;506
AACCC;57;10885;92;502
AACCG;13;950;33;57
AACCT;0;2;33;0
AACGA;0;4;0;0
AACGC;34;5659;17;92
AACGG;0;0;0;0
AACGT;5;140;7;2
```


Maximum of CDR coverage more than 200
count
40
30
20
10
0
0
10000
20000
30000


Maximum of FR coverage more than 200
count
20
10
0
0
25000
50000
75000


Histogram of FR rowSums without maximum
count
90
60
30
0
0
1000
2000
3000
4000
5000


Histogram of CDR rowSums without maximum
count
40
30
20
10
0
0
500
1000

Length FR_list

Length_CDR-list

Length(FR_list & CDR_list)

869	854
523	515
448	433

After filter input with absent k-mer

Compare 228 hypothesis among 10 patients (age data)

p_adjusted with method “Holm”
get 32% p_value < 0.05

Compare 228 hypothesis among 40 patients (age & flu data)

Better?

Antibodies are not always the good guys

Skull X-ray showing osteolytic lesions of myeloma.

Immunoglobulins
Gaw, Allan, MD PhD FRCPath FFPM PGCertMedEd, Clinical Biochemistry: An Illustrated Colour Text, 26, 52-53

Copyright © 2013 2013, Elsevier Ltd. All rights reserved.

ELSEVIER

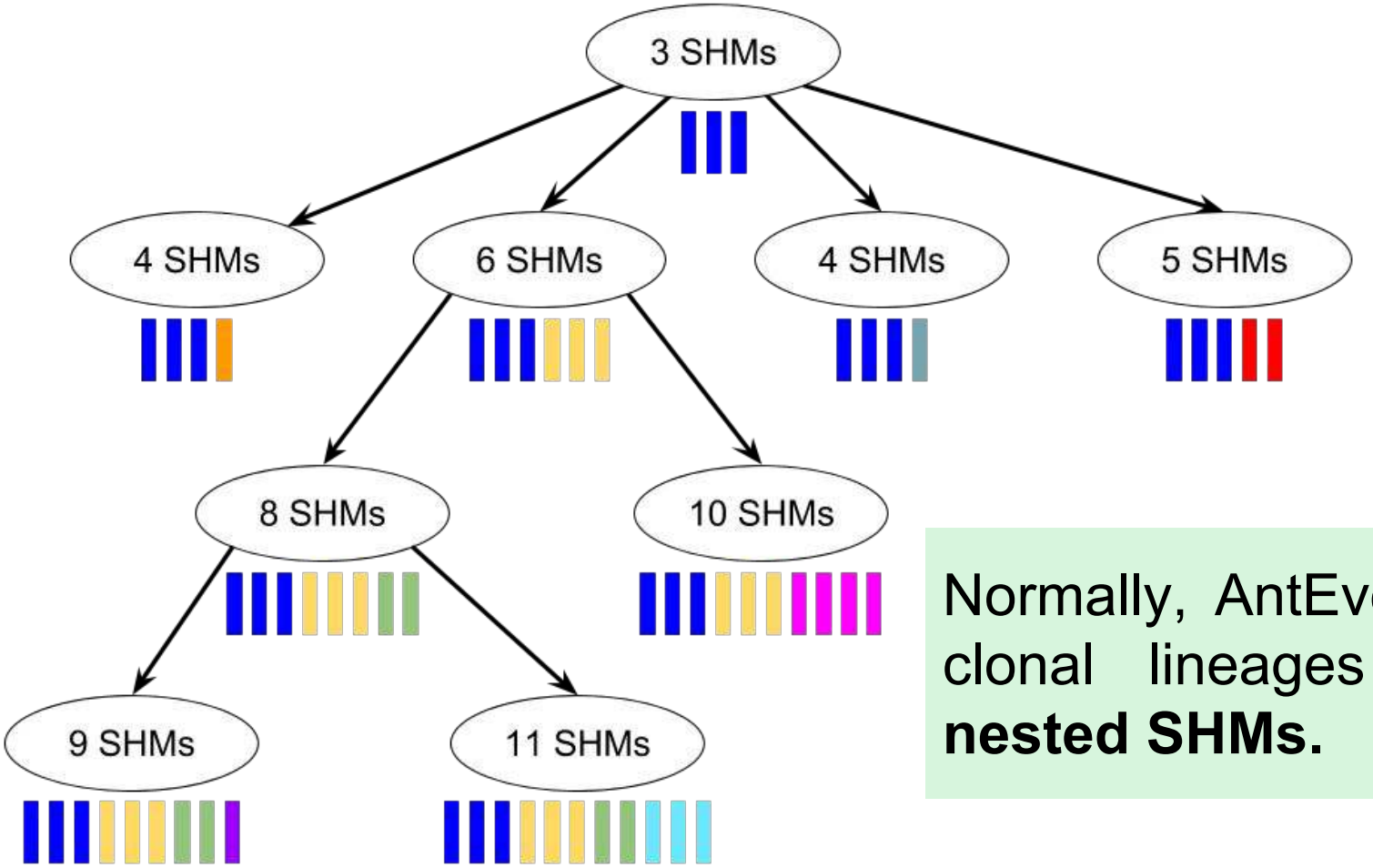
ClinicalKey®

Thank you for your attention

Additional information

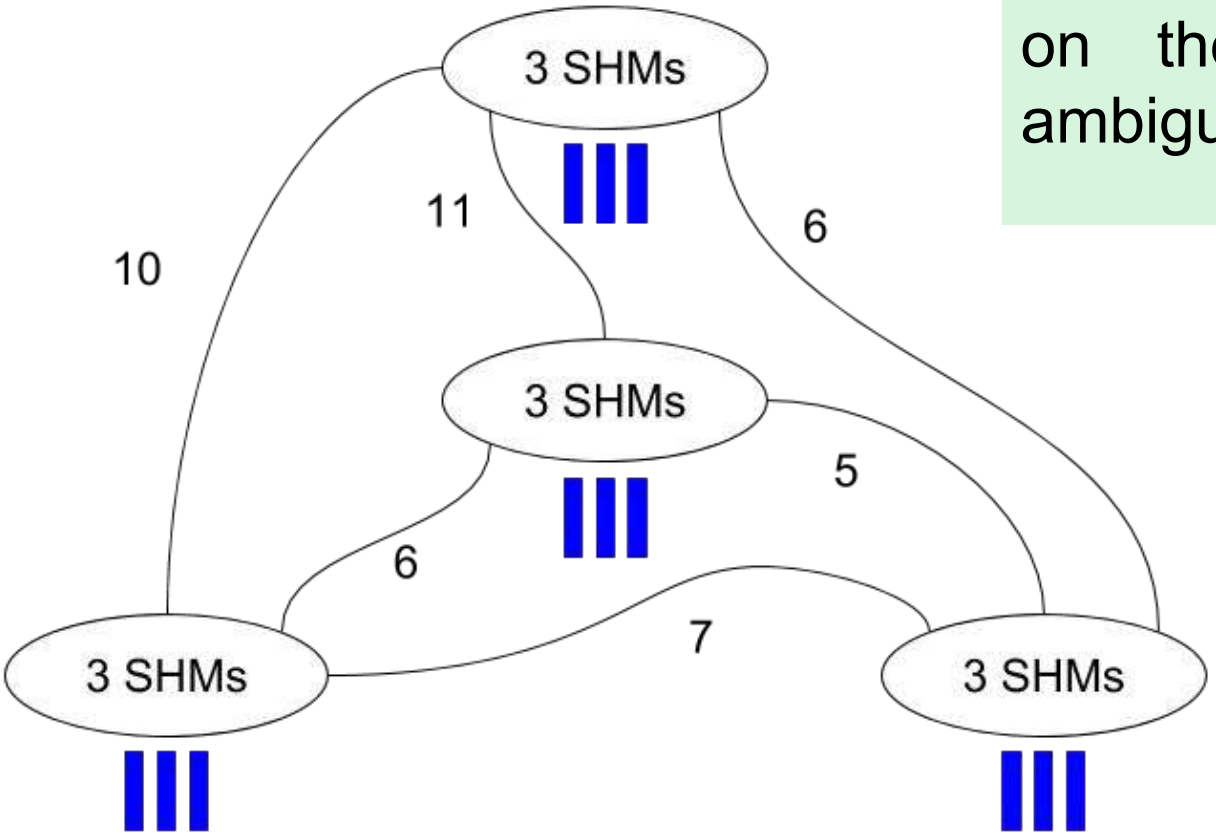
Why probabilistic model for SHMs?

Construction of the evolutionary trees.



Normally, AntEvolo infers clonal lineages through **nested SHMs.**

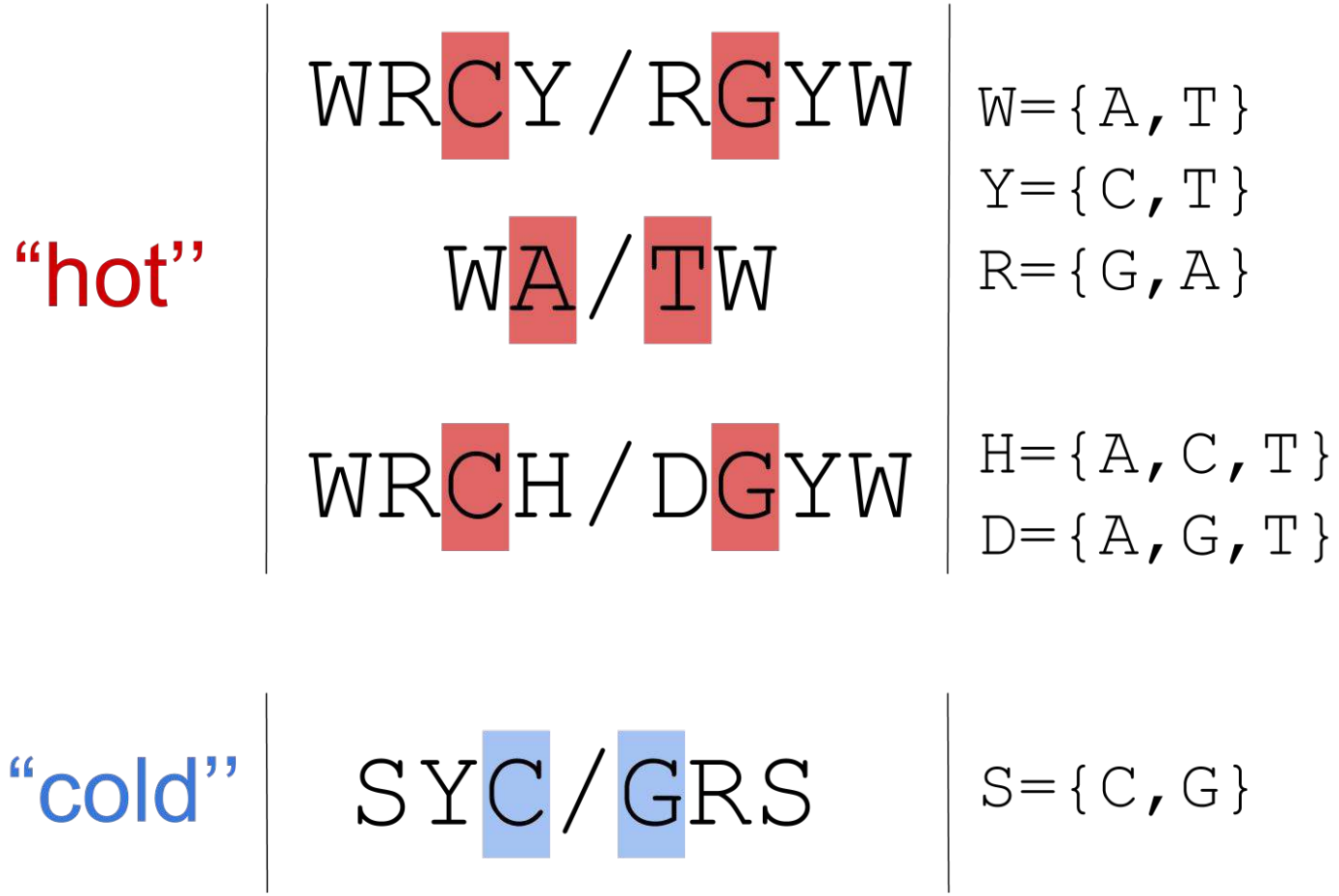
Unclear situation



The model could elaborate on the **orientation** of ambiguous edges.

Hot-/cold-spots

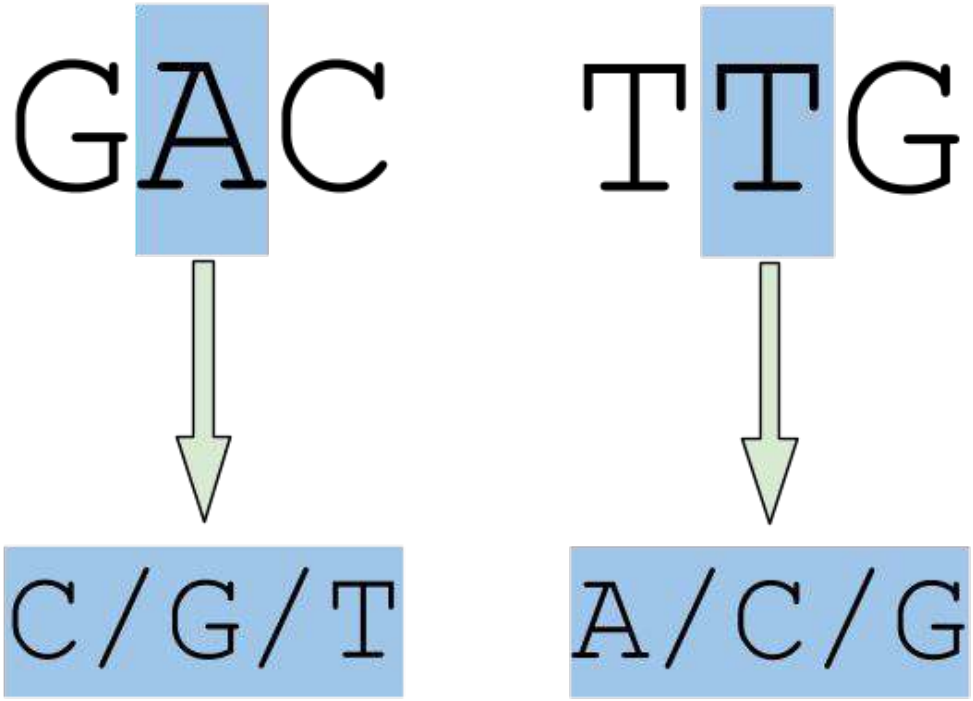
Mutations are not uniformly distributed



Rogozin and Kolchanov, *Biochimica et Biophysica Acta*, 1992

Locality of AID enzyme activity

Surrounding bases heavily influence SHMs.



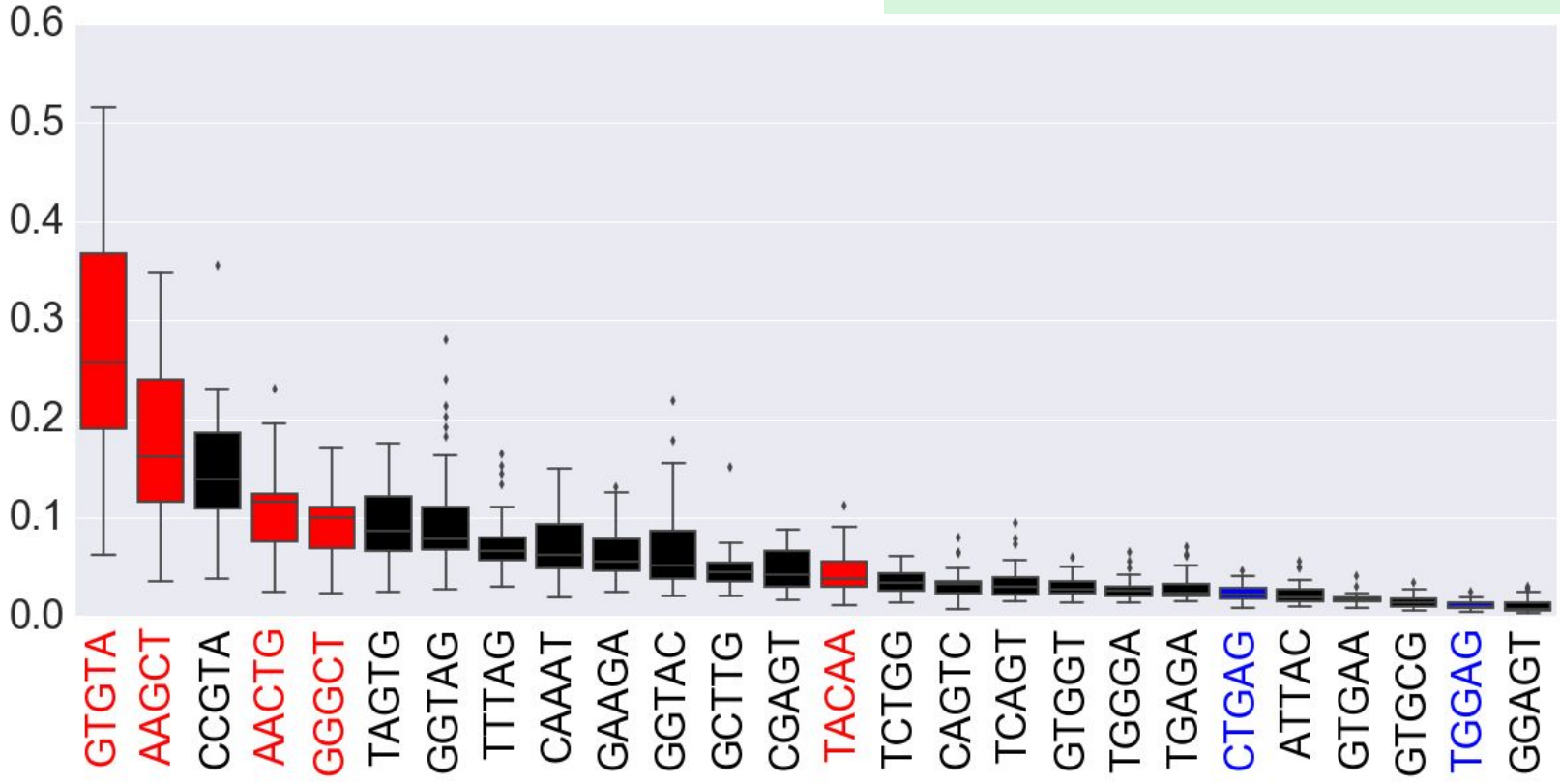
Kleinshtein, *Molecular Immunology*, 2011

Natural variations of parameters

We noticed natural **diversity** in parameters' estimations between individuals.

IGH **mutabilities** used for **targeting** model:

The same for both **light** chains.

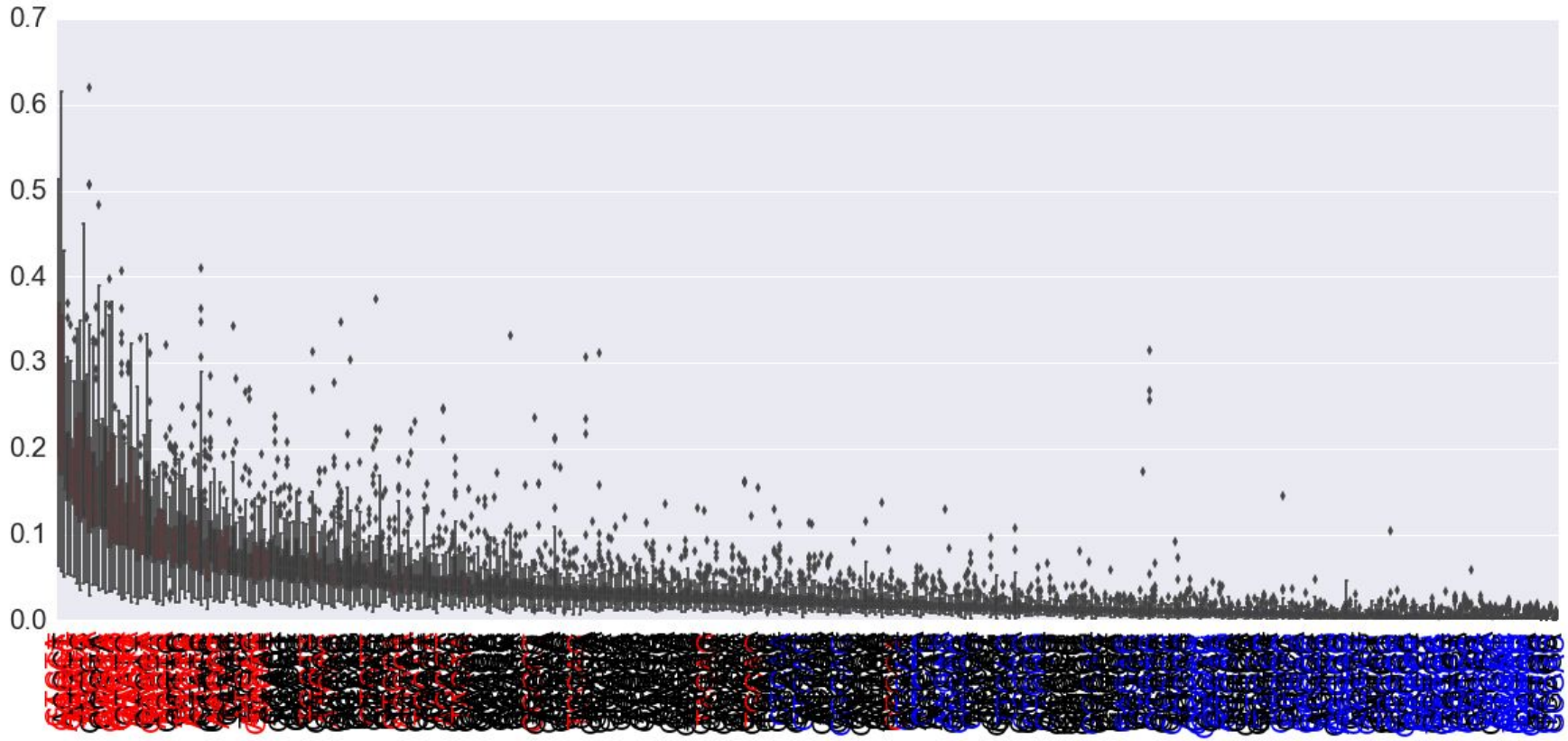


We checked **significance** of the variance between datasets by comparing it to variance inside the dataset.

Natural variations of parameters

IGH **mutabilities** used for **targeting** model:

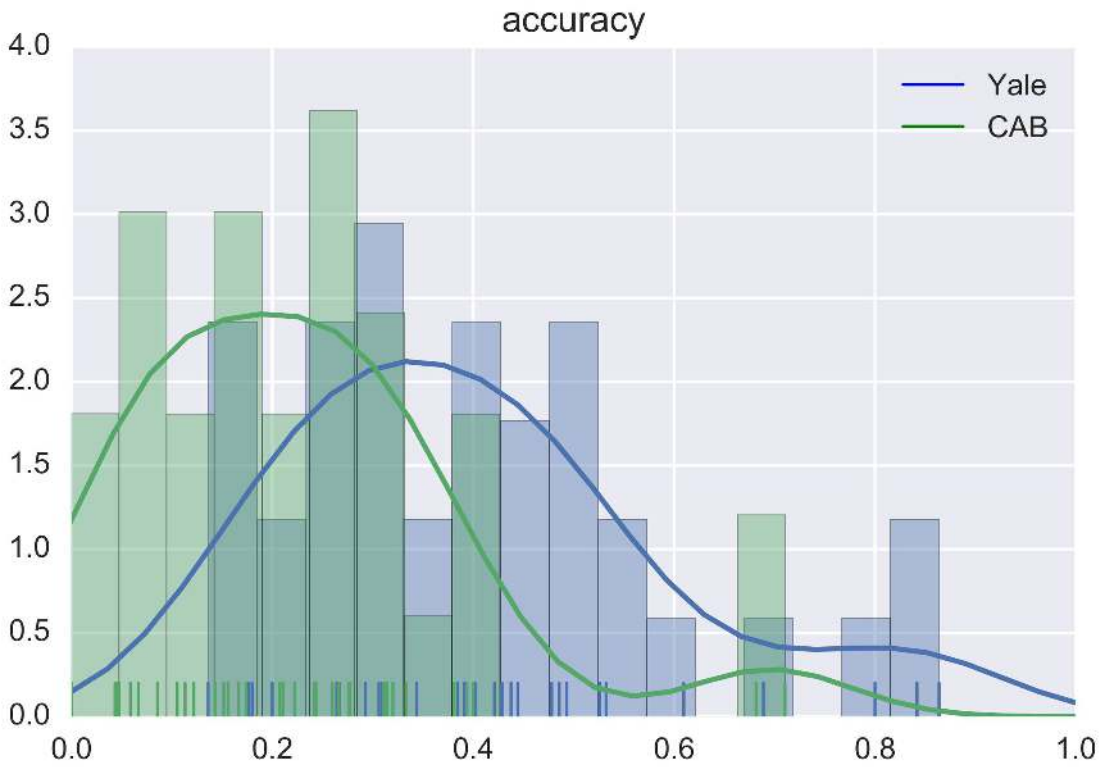
The same for both **light** chains.



We also checked that **hot-/cold-spots** indeed have **high/low mutability**.

CDR3 test, IGH

For each tree in AbVitro datasets we calculate % **wrong orientation** suggested by the model (*accuracy*).



Histogram median:

Yale: 0.39

CAB: 0.21

Equal mean hyp.:

Pvalue ~ 10^{-5}

Full accuracy:

Yale: 0.4

CAB: 0.23