Statistical analysis of full-length antibody repertoire using immunosequencing data

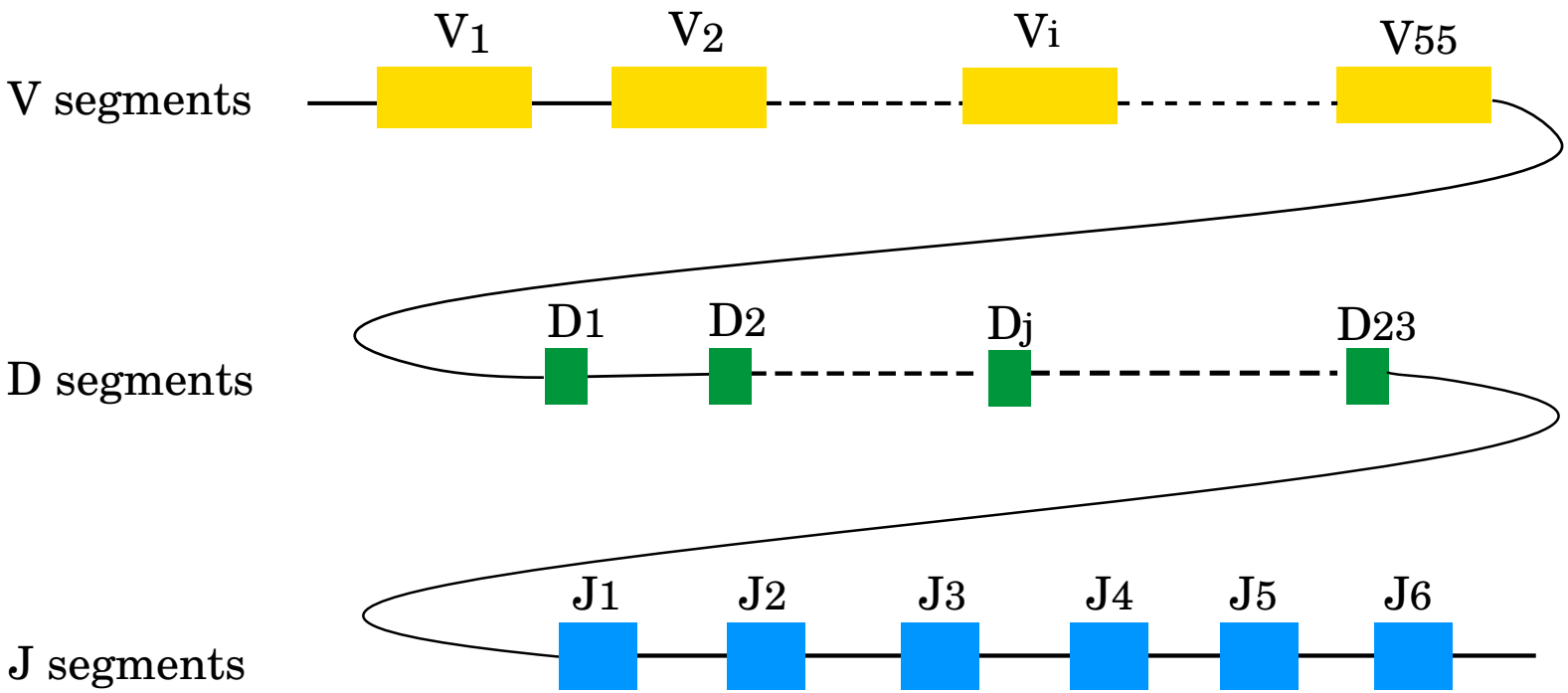
Andrew Bzikadze
seryrzu@gmail.com

September 13, 2015

Outline

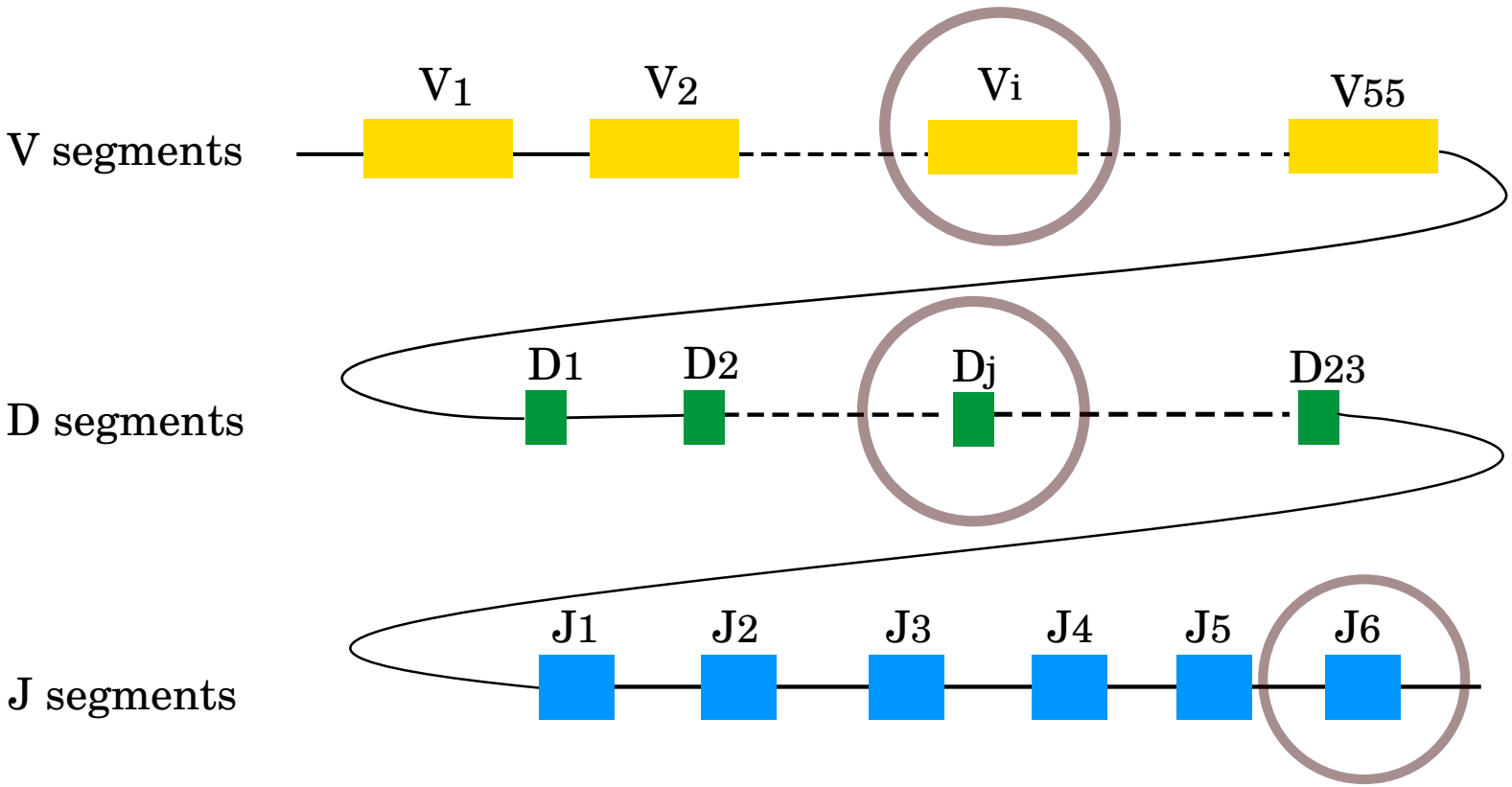
Introduction: V(D)J-recombination

Immunoglobulin heavy chain locus:



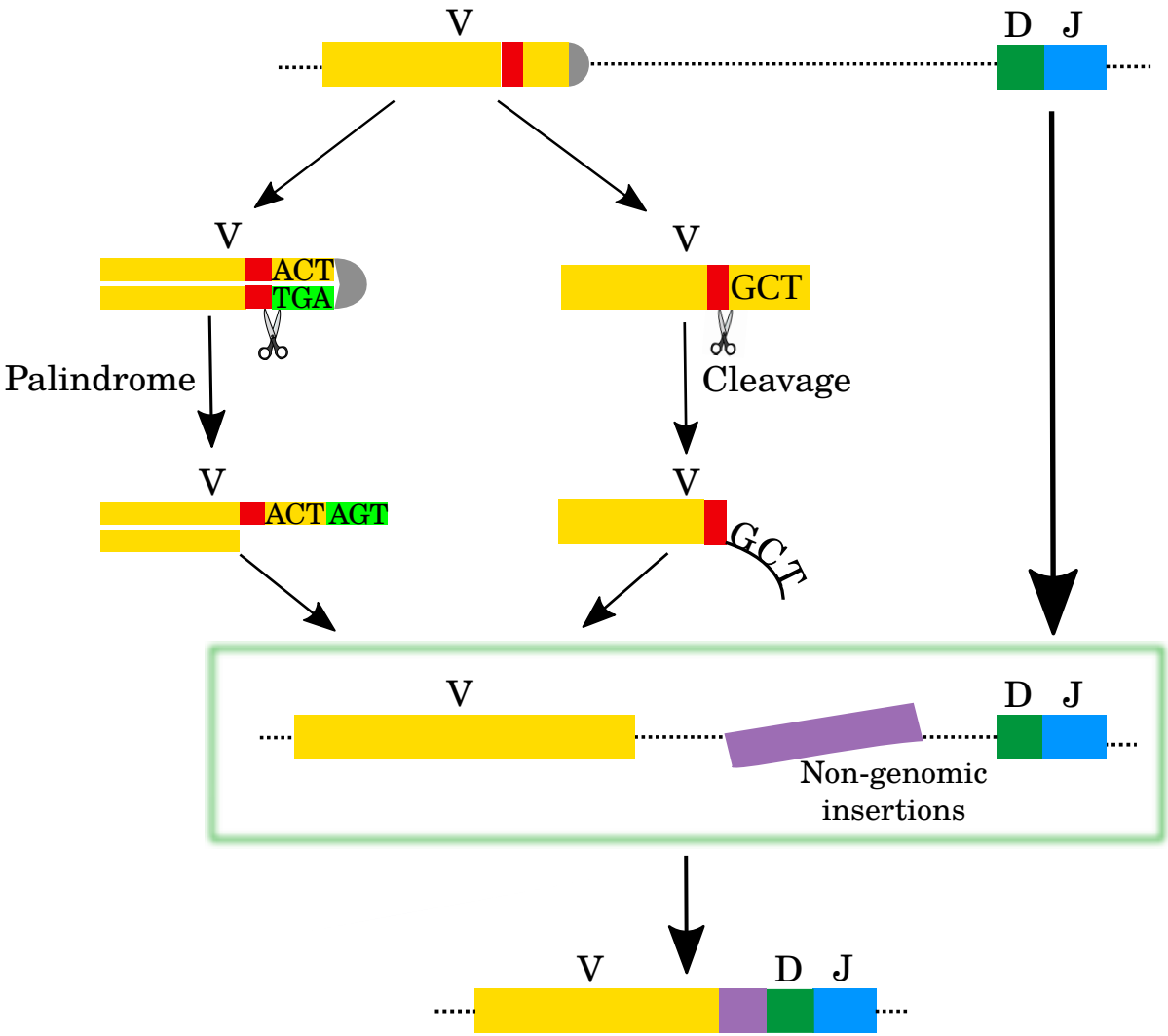
Introduction: V(D)J-recombination

Segment of each type is selected:



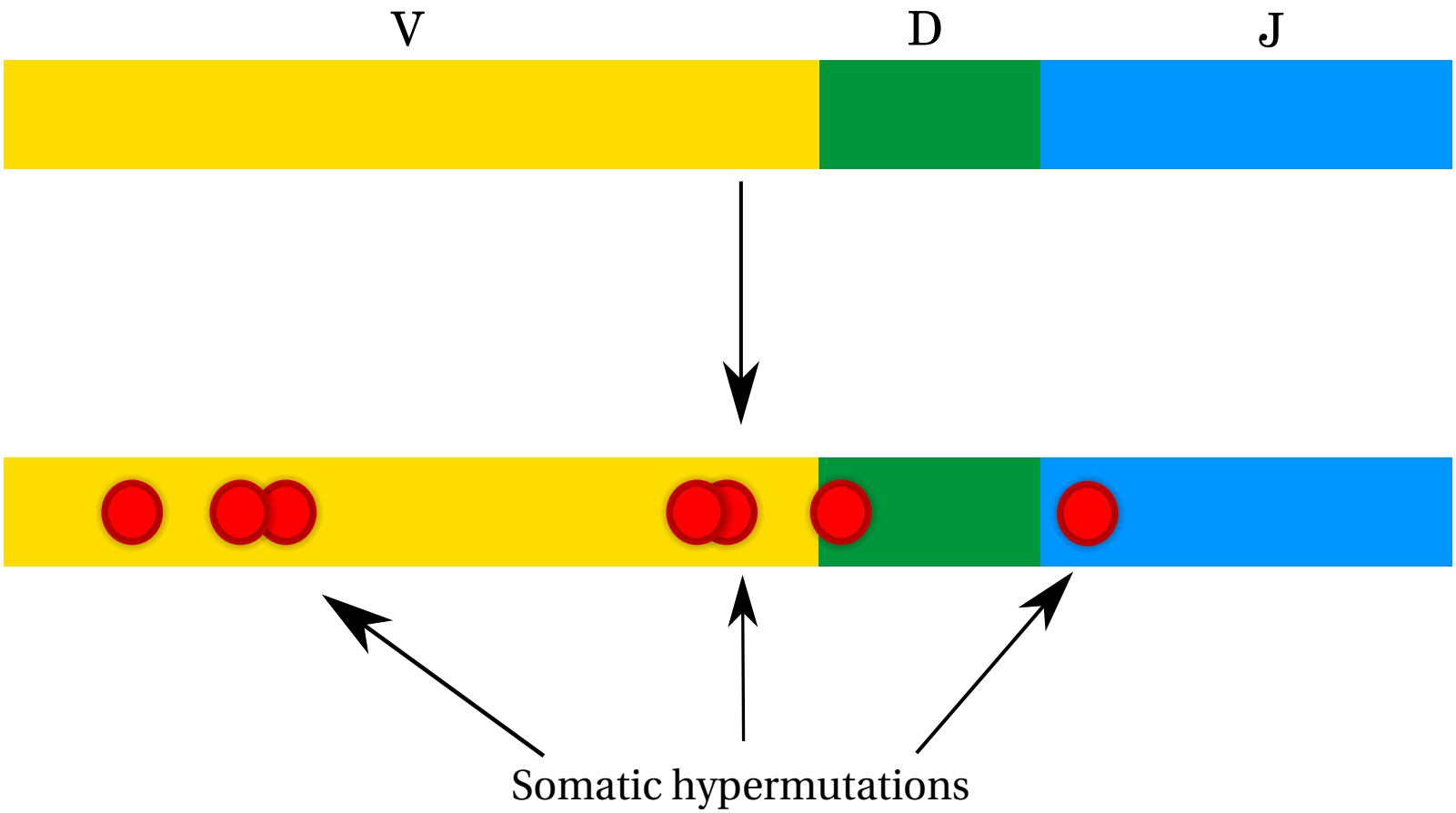
Introduction: V(D)J-recombination

3 types of biochemical events: *palindrome*, *cleavage*, *non-genomic insertion*.



Introduction: V(D)J-recombination

Further optimization of antibody affinity is achieved through extensive mutations referred as *somatic hypermutations*:



Introduction: V(D)J-recombination

Because we do not know the deterministic nature of the V(D)J-recombination, it is reasonable to consider it as a random (stochastic) process.

Hence the analysis of somatic recombination can be done in statistical and simulation terms.

Motivation: comparing different antibody repertoires

B-cells:

- Agnieszka Dryla *et al.*, Comparison of Antibody Repertoires against Staphylococcus aureus in Healthy Individuals and in Acutely Infected Patients — CVI, 2005.
- Beddows S. *et al.*, Comparison of the antibody repertoire generated in healthy volunteers following immunization with a monomeric recombinant gp120 construct derived from a CCR5/CXCR4-using human immunodeficiency virus type 1 isolate with sera from naturally infected individuals — Journal of Virology, 1999.

T-cells:

- Van Rhijn I, Moody DB, Donor Unrestricted T Cells: A Shared Human T Cell Response — Journal of Immunology, 2015.
- Warren RL *et al.*, Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes — Genome Research, 2011.

Motivation: simulation of a repertoire

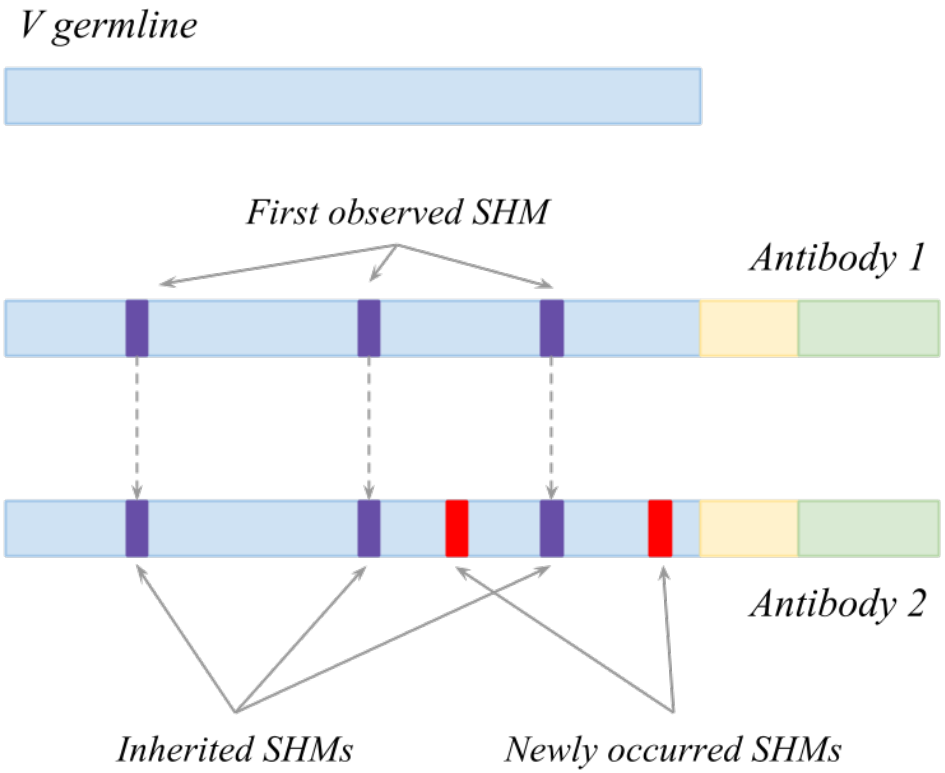
Appropriate statistical model of somatic recombinations potentially improves `IgSimulator`, making it more "realistic".

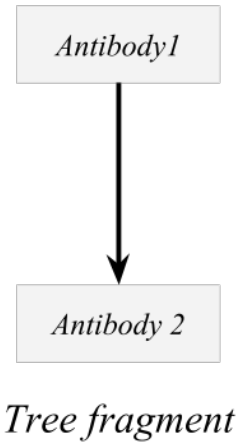
IgSimulator

Statistical Model

Motivation: Clonal trees

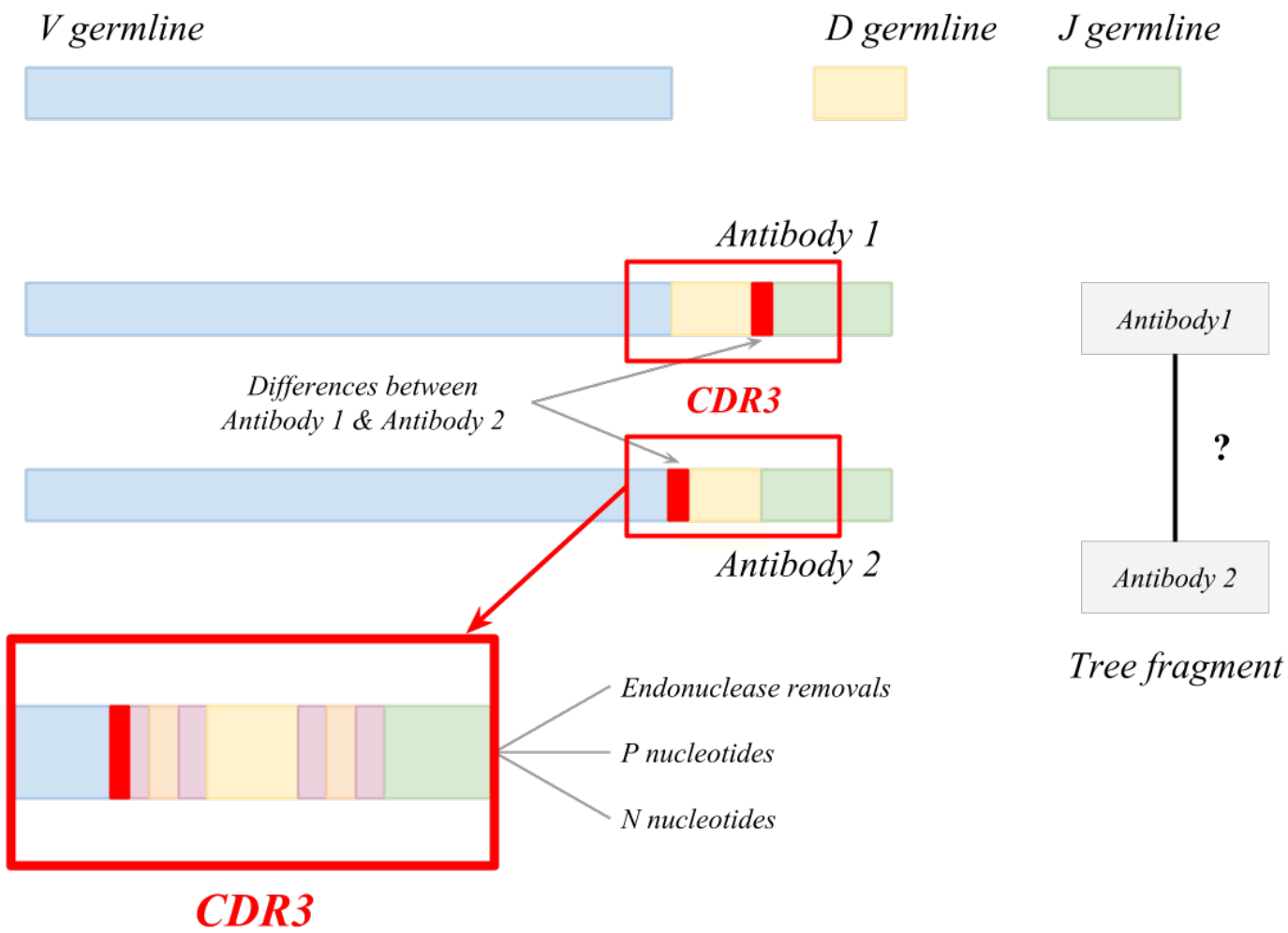
Clear situation:



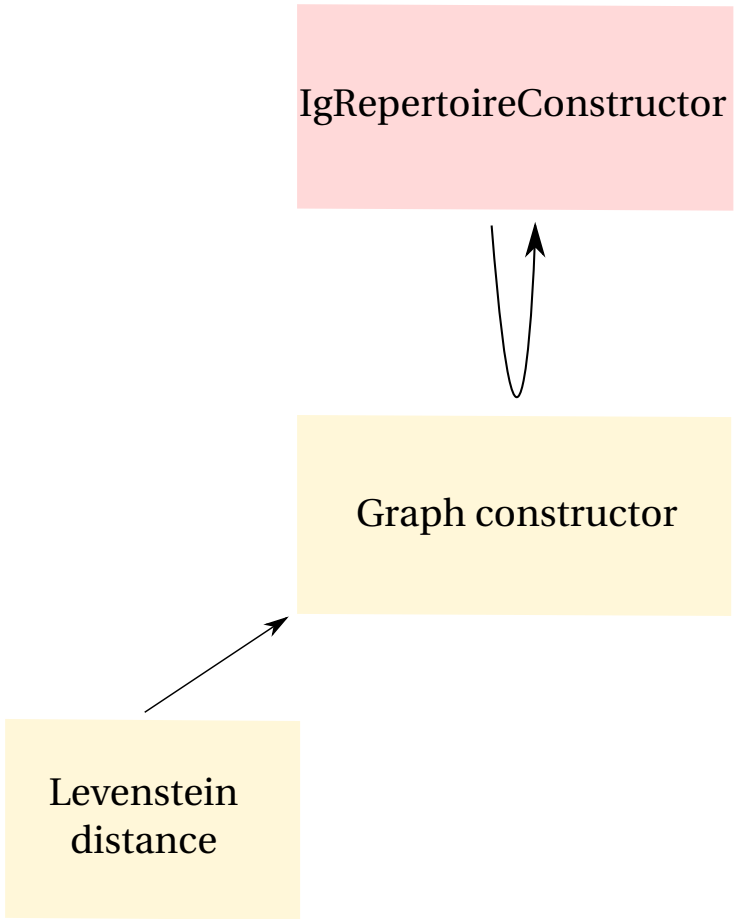


Motivation: Clonal trees

Arguable situation:

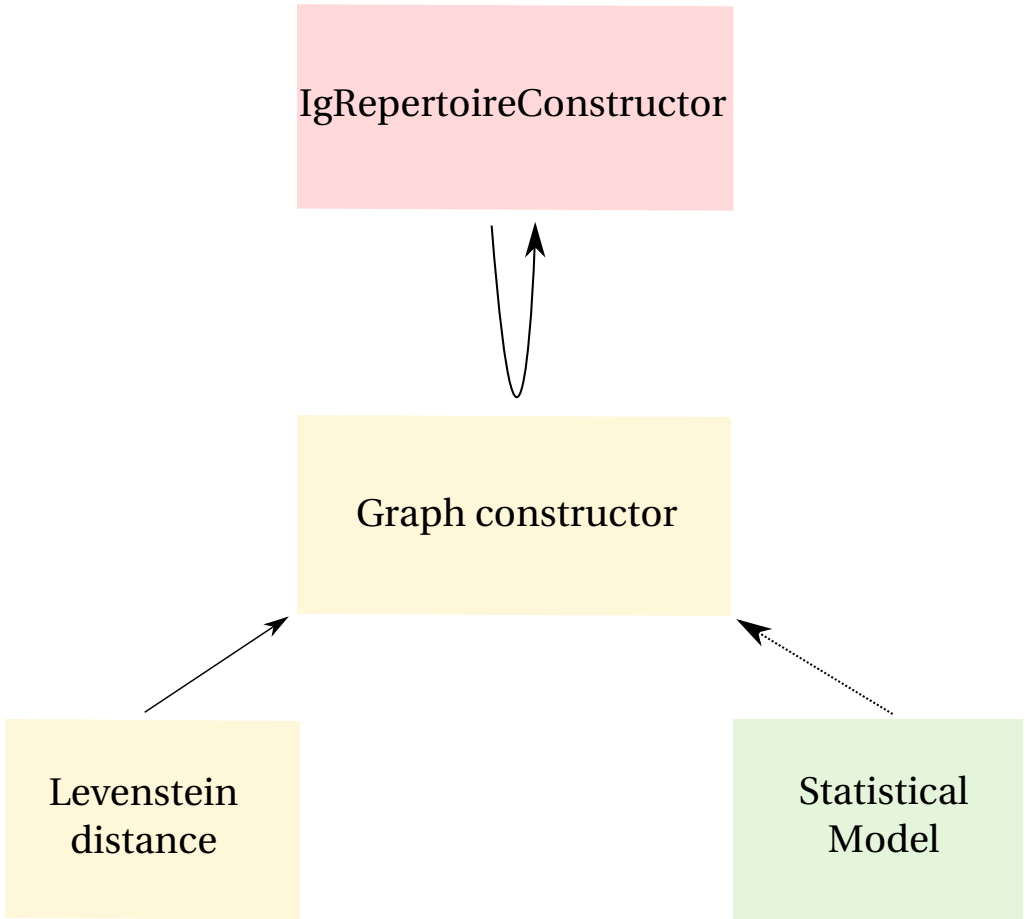


Motivation: `IgRepertoireConstructor`



The current release of the `IgRepertoireConstructor` uses Levenstein distance to construct edges in the graph.

Motivation: IgRepertoireConstructor



The statistical model could suggest a more delicate approach.

Tasks

There are plenty of tasks. To name a few:

- What is the correlation between `D-J` and `V-DJ` joining?
- Is there any correlation between the *cleavage* / *palindromes* and specific gene segments?
- What are the properties of the *non-genomic insertion*?

Background

An article about the distribution law of CDR3 generating recombinations for T-cells:
Anand Murugana, Thierry Morab, Aleksandra M. Walczakc and Curtis G. Callan — 2012:

- Analysis is focused on non-productive CDR3s.
- Suggested model sets joint distribution over the set of discrete variables: *identities* of V-, D-, J- genes, number of *deletions* from the end of a segment, *palindromic* nucleotides and *non-genomic insertion* at the end of a gene.

Background

An article about the distribution law of CDR3 generating recombinations for T-cells:
Anand Murugana, Thierry Morab, Aleksandra M. Walczakc and Curtis G. Callan — 2012:

- Analysis is focused on non-productive CDR3s.
- Suggested model sets joint distribution over the set of discrete variables: *identities* of V-, D-, J- genes, number of *deletions* from the end of a segment, *palindromic* nucleotides and *non-genomic insertion* at the end of a gene.
- 2865(!) parameters to estimate.

Questions about the paper:

- Is the suggested model really adequate (including the problem of potential overfitting)?
- Are the results statistically significant?
- Are similar results true for B-cells?

Two types of events

The goals

- correlations between *palindromes* and specific gene segments,
- properties of the *non-genomic insertion*

include the task of distinguishing "biological" (meaningful) events from "accidental" (random) observations and hence require additional knowledge about repertoire structure.

Considering that first we decided to concentrate on a problem of computation of correlation between **cleavage and specific gene segments**.

Two types of events

The goals

- correlations between *palindromes* and specific gene segments,
- properties of the *non-genomic insertion*

include the task of distinguishing "biological" (meaningful) events from "accidental" (random) observations and hence require additional knowledge about repertoire structure.

Considering that first we decided to concentrate on a problem of computation of correlation between **cleavage and specific gene segments**.

Unlike two problems above *cleavage* can be detected precisely according to the alignment.

Outline

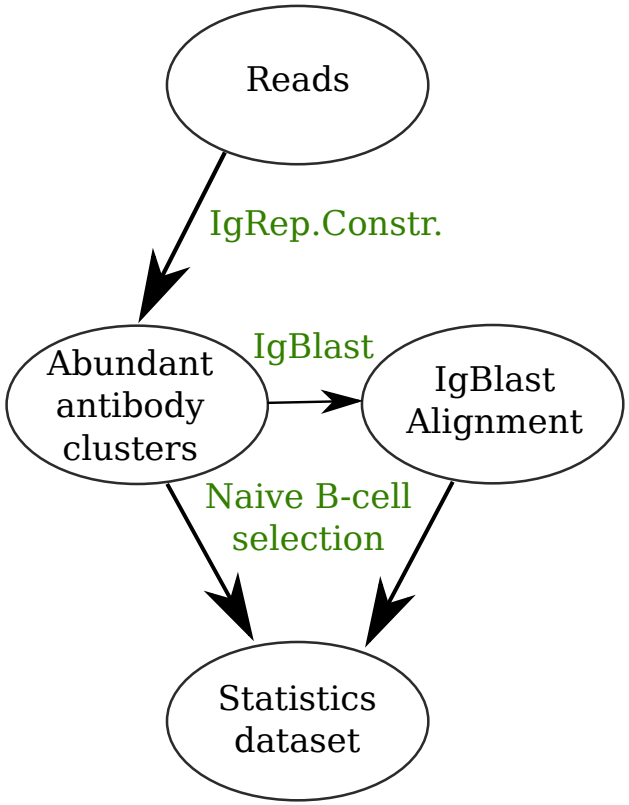
Cleavage and specific gene segments

Antibody repertoire is a result of V(D)J recombination and secondary diversification. To analyze properties of V(D)J recombination only we perform the following steps:

Cleavage and specific gene segments

Antibody repertoire is a result of V(D)J recombination and secondary diversification. To analyze properties of V(D)J recombination only we perform the following steps:

- Constuct a full-length antibody repertoire using `IgRepertoireConstructor`.
- Consider highly abundant antibody clusters of the constructed repertoire.
- Apply `IgBlast` to reads corresponding to selected clusters.
- To avoid effects of the secondary diversification filter out reads with *low* alignment score.



Cleavage and specific gene segments: V-genes

Age datasets — 9 heavy chain datasets sequenced from human serum samples without cell sorting.

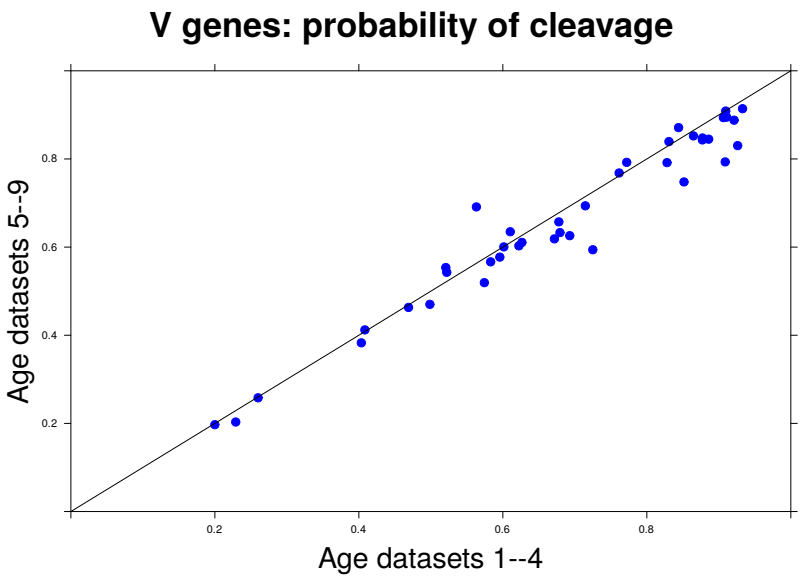


Figure : Age datasets. The point is the V gene segments. Pearson correlation is 0.98.

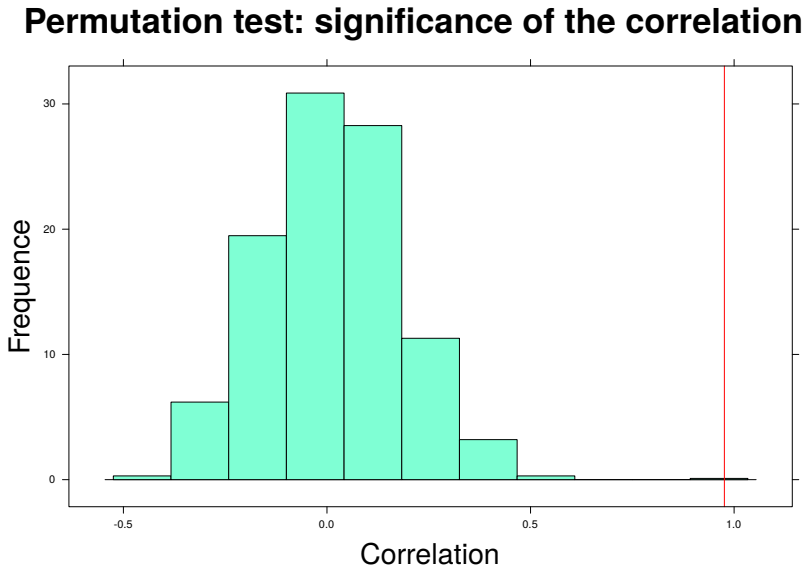


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

Cleavage and specific gene segments: V-genes

Age datasets — 9 heavy chain datasets sequenced from human serum samples without cell sorting.

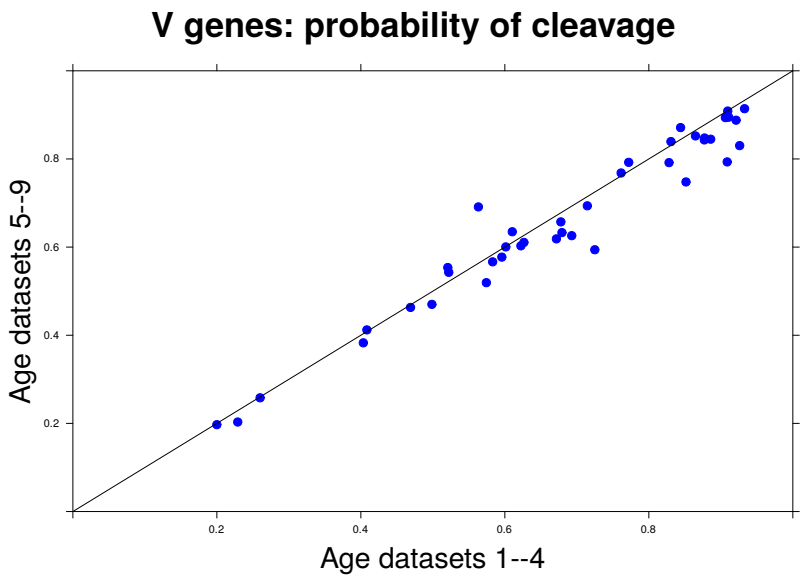


Figure : Age datasets. The point is the V gene segments. Pearson correlation is 0.98.

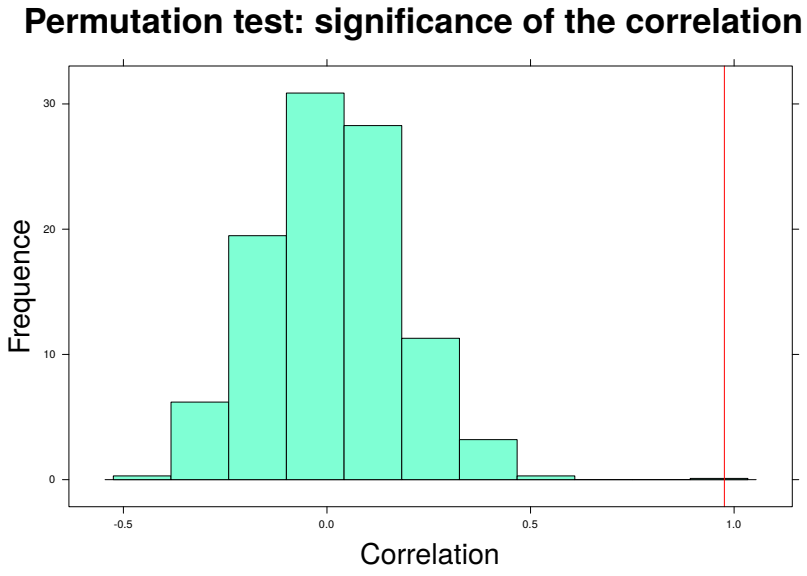


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

Hence reads in the dataset are dependent standard pooled Z-test for equal propotions is not applicable.

Further steps

Find out factors to clusterize V segments.

Probability of *cleavage* is not appropriate. It is reasonable to seek a way to clusterize V-genes by

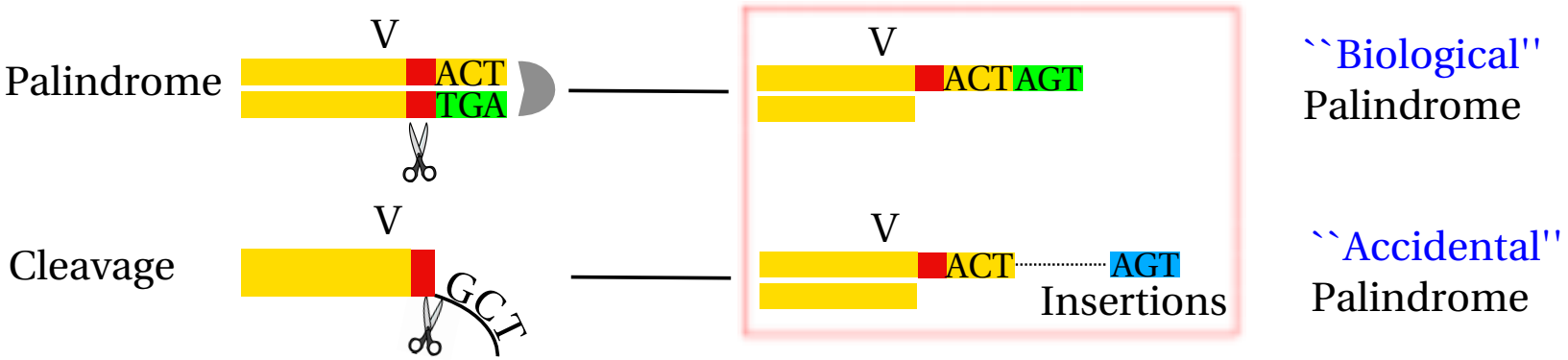
- *palindromes* length;
- GC-content;
- different type of genes (V vs J *etc...*).

Outline

Two types of palindromes

If a *cleavage* took place, then no "biological" *palindrome* could happen.

An "accidental" *palindrome* can still happen due to the *non-genomic insertions*.



Two types of palindromes

- The simplest model assumes that every nucleotide ξ in the sequence is distributed **uniformly**:

$$\mathbb{P}(\xi = x) = \frac{1}{4} \text{ where } x \in \{'A', 'C', 'G', 'T'\}.$$

- In that model the length η of an "accidental" palindrome has $\mathrm{Geom}(3/4)$ distribution, so

$$\mathbb{P}(\eta = n) = \frac{3}{4^{n+1}} \text{ for all } n \in \mathbb{N}_0.$$

Two types of palindromes

Emperical (Age-datasets) and $\mathrm{Geom}(3/4)$ distributions in log scale:

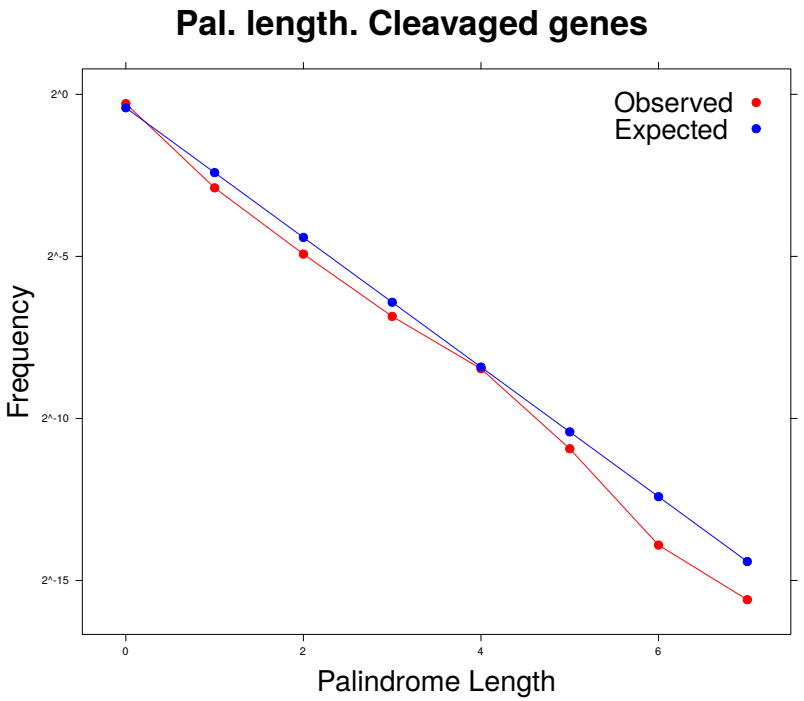


Figure : The mean of length is 0.33.

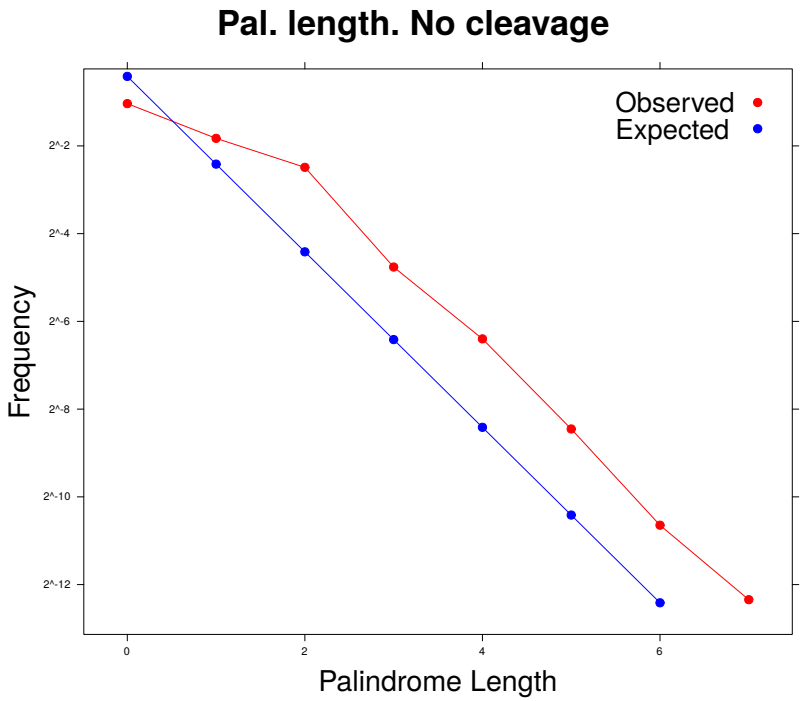
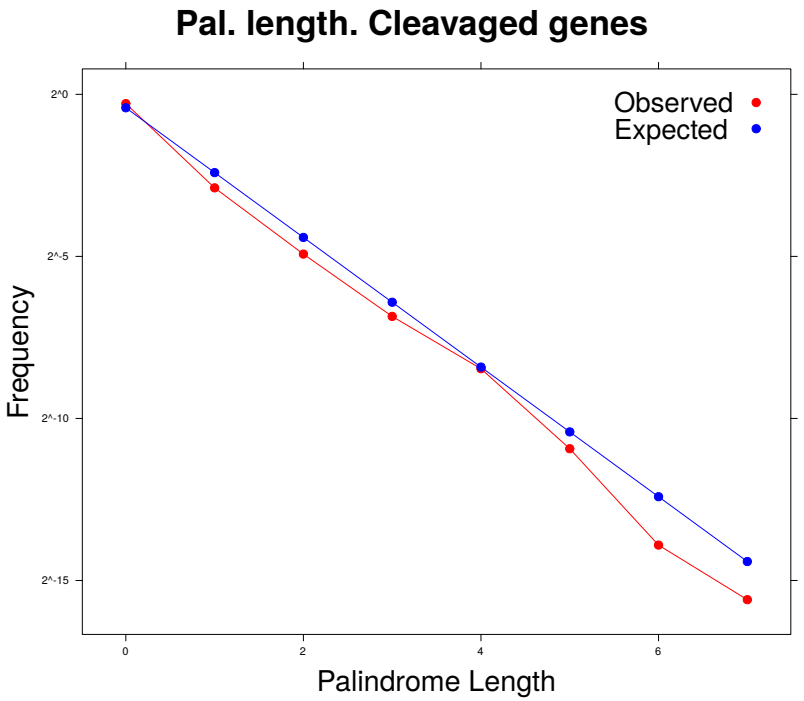


Figure : The mean of length is 0.82.

Two types of palindromes

Emperical (Age-datasets) and $\mathrm{Geom}(3/4)$ distributions in log scale:



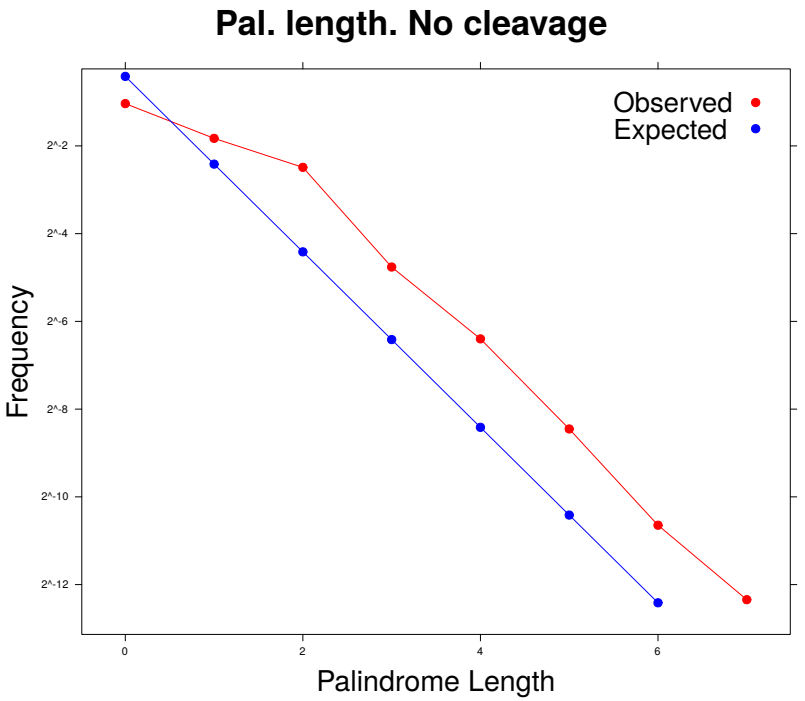


Figure : The mean of length is 0.33.

Figure : The mean of length is 0.82.

Hence reads in the dataset are dependent goodness-of-fit χ^2-test is not applicable.

Next steps

- To find out the distribution of “biological” palindrome length.
- To construct more adequate model for nucleotide distribution.

Summary

Statistical analysis of an antibody repeitore is very promising research with lots of applications. Some of them:

- Comparing antibodies repertoires.
- Simulation of a repertoire.
- Edges in clonal trees.
- Improvement of `IgRepertoireConstructor`.

Our goals:

- Construct an adequate statistical model of the `V(D)J`-recombination.
- Find out correlations between segments of genes.
- Find out correlations between various biological events, including *cleavage*, *palindromes* and *non-genomic-insertions*.