Statistical analysis of an antibody repertoire

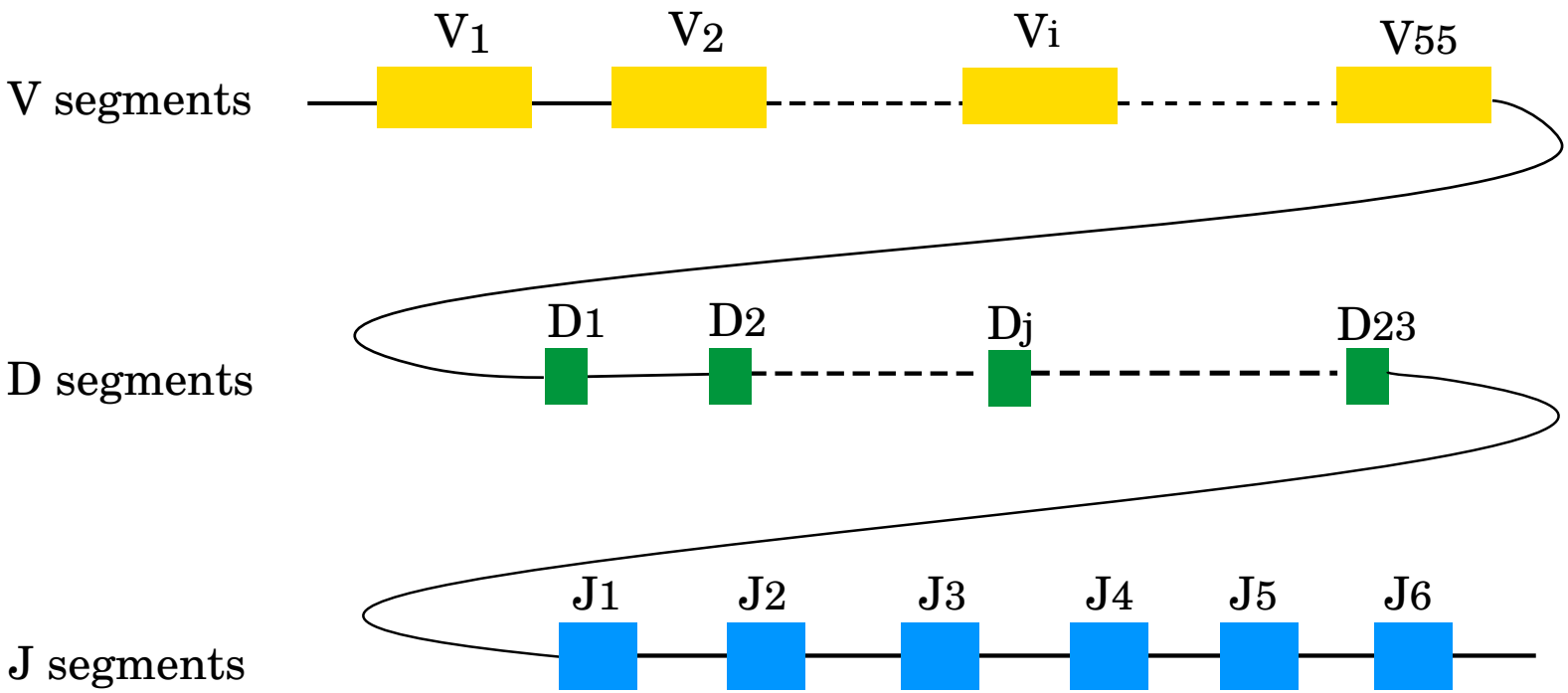
Andrew Bzikadze
seryrzu@gmail.com

September 13, 2015

Outline

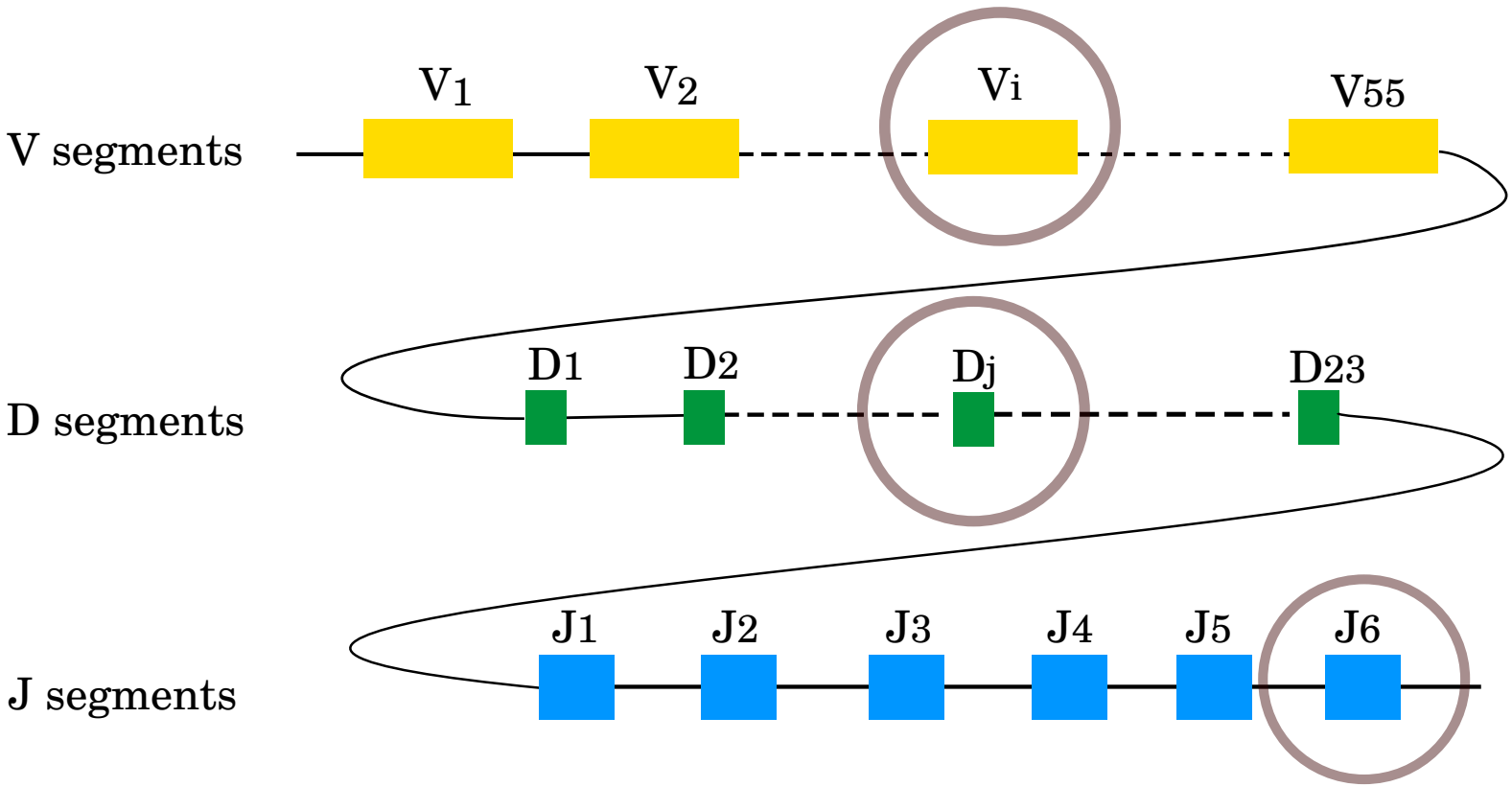
Introduction: V(D)J-recombination

B-cells gene locus:



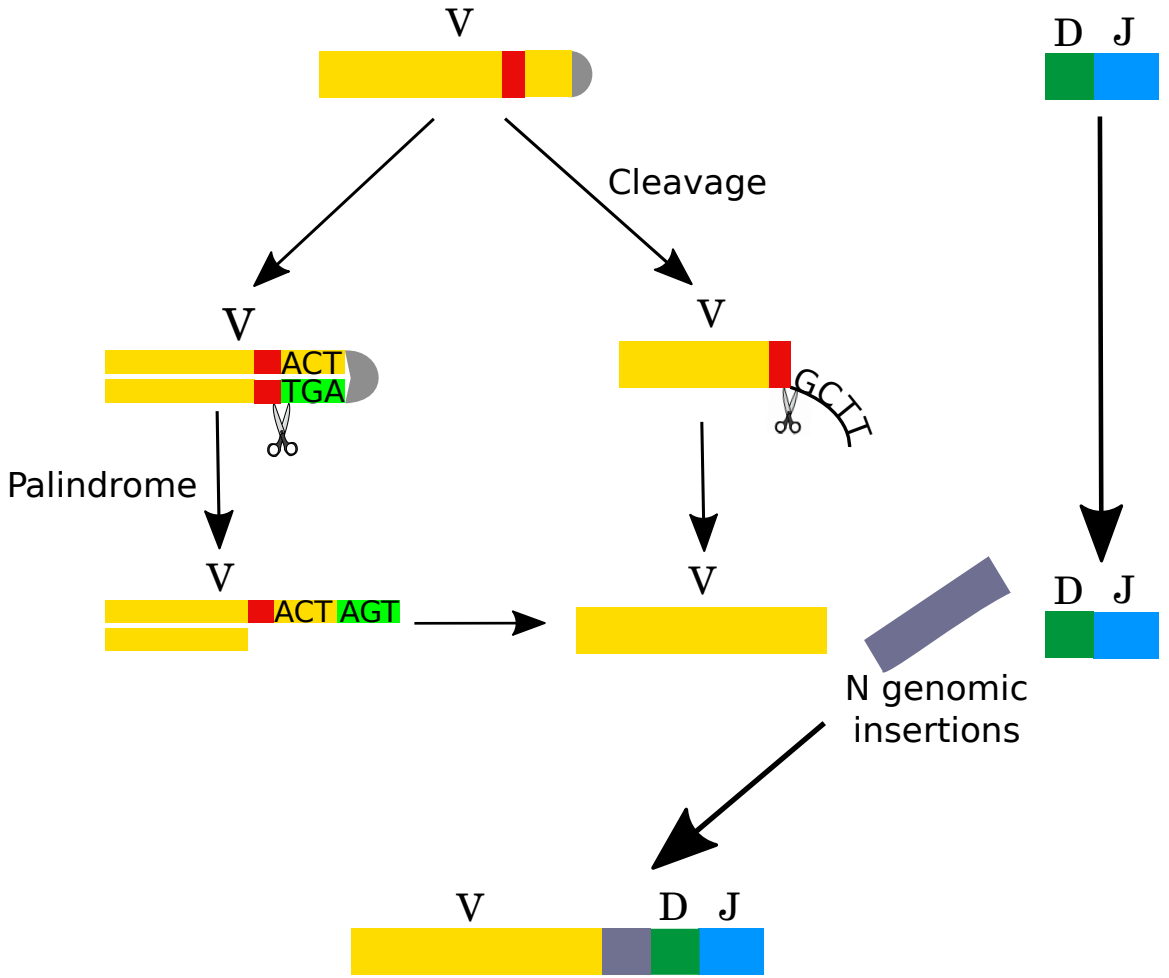
Introduction: V(D)J-recombination

Segment of each type is selected:



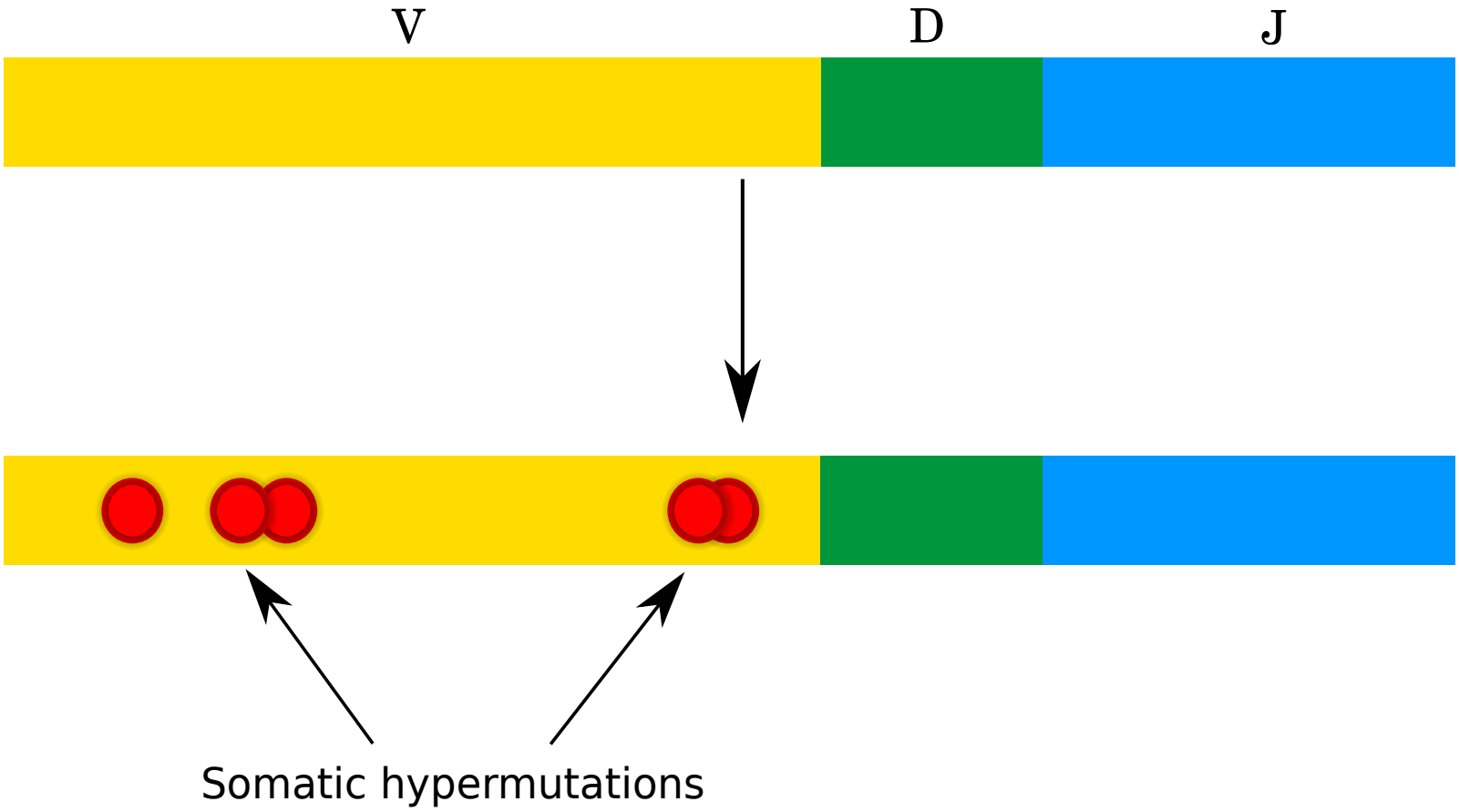
Introduction: V(D)J-recombination

3 types of biochemical events: *palindrome*, *cleavage*, *insertions*.



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 determenistic nature of the V(D)J-recombination, it is reasonable to consider this transform 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:

- Comparison of Antibody Repertoires against Staphylococcus aureus in Healthy Individuals and in Acutely Infected Patients.
- 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.

T-cells:

- Donor Unrestricted T Cells: A Shared Human T Cell Response.
- 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.

Motivation: simulation of a repertoire

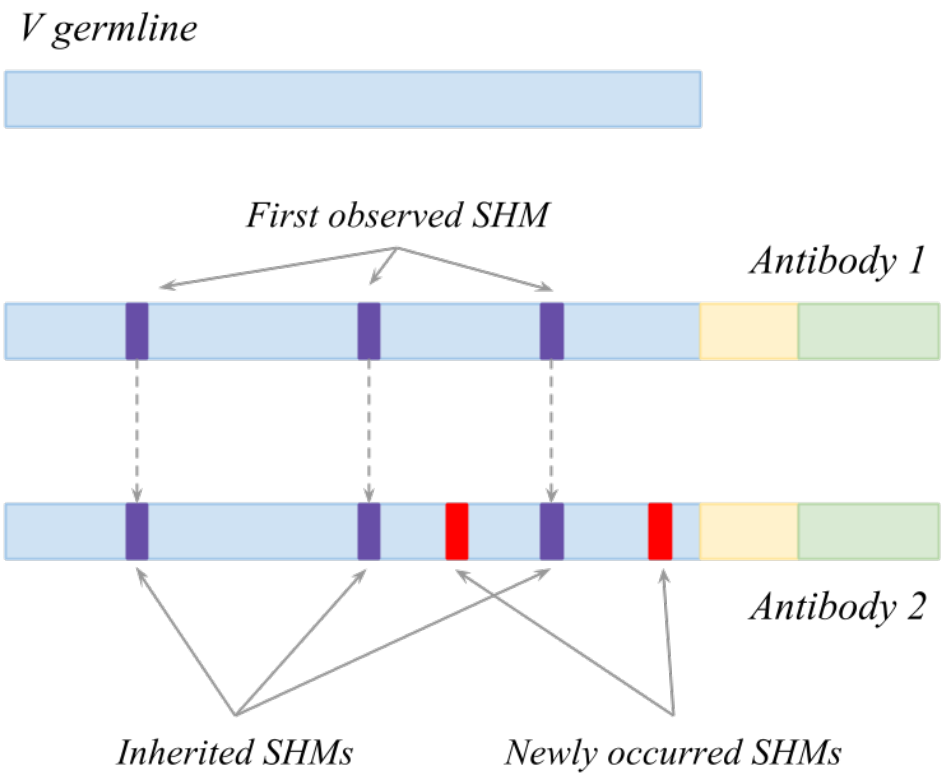
Appropriate statistical model of somatic recombinations potentially improves `IgSimulator`, making it more "realistic".

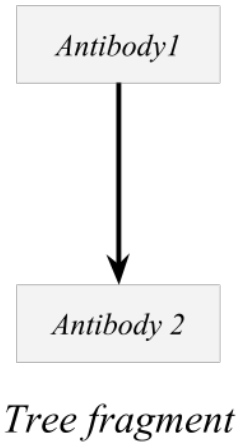
IgSimulator

Statistical Model

Motivation: Clonal trees

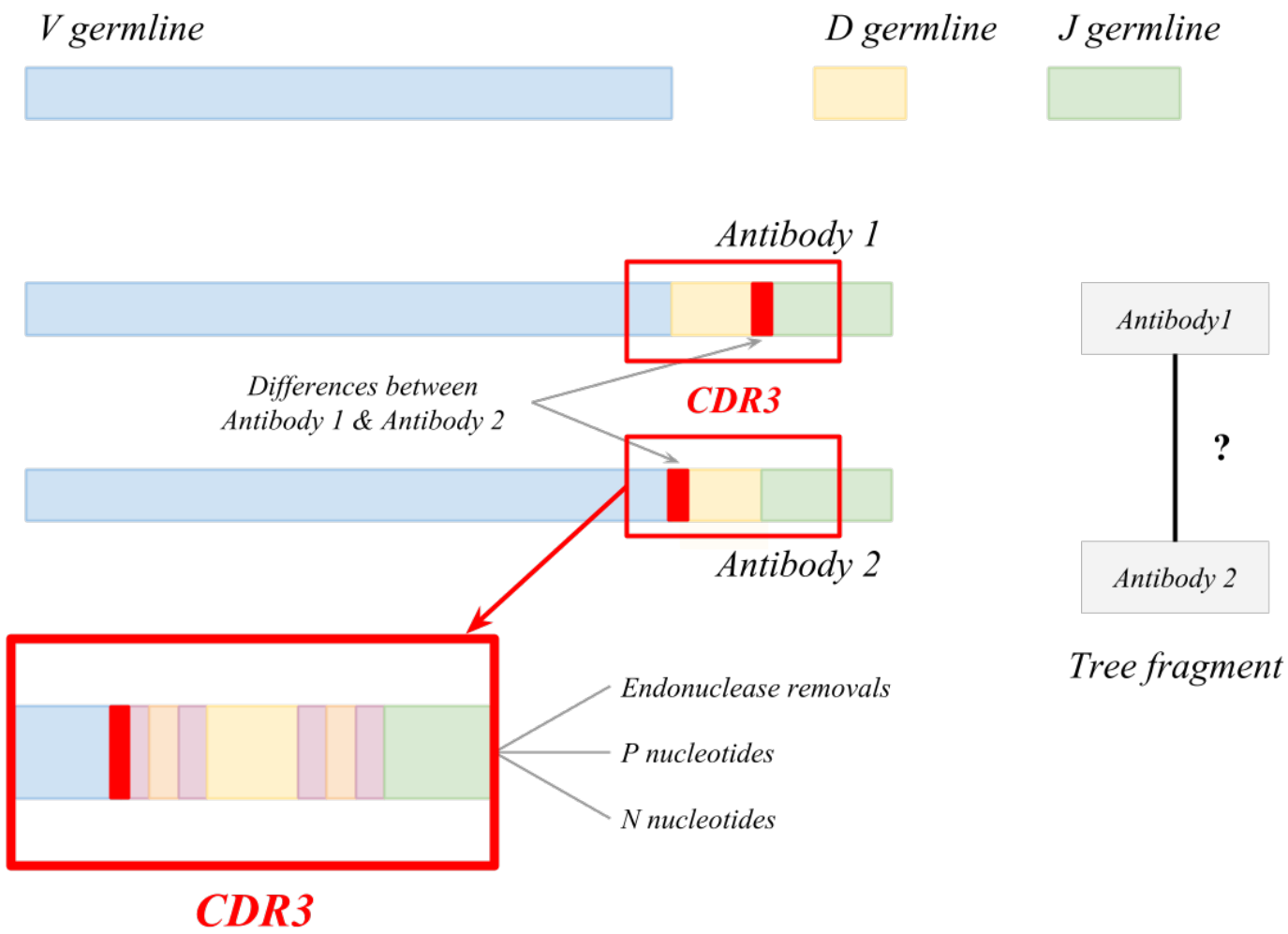
Clear situation:



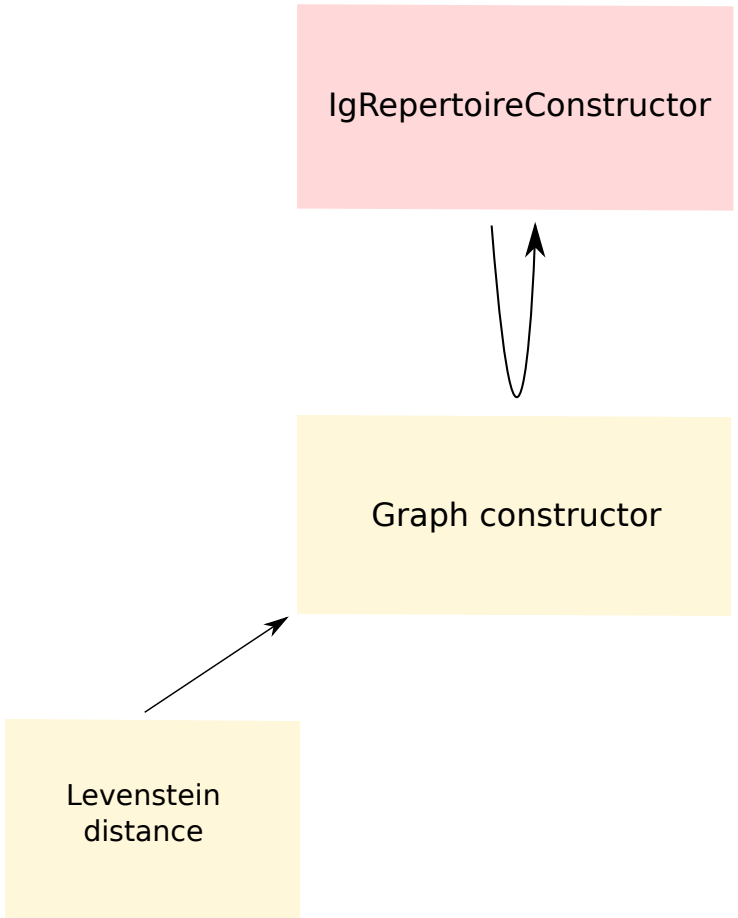


Motivation: Clonal trees

Arguable situation:

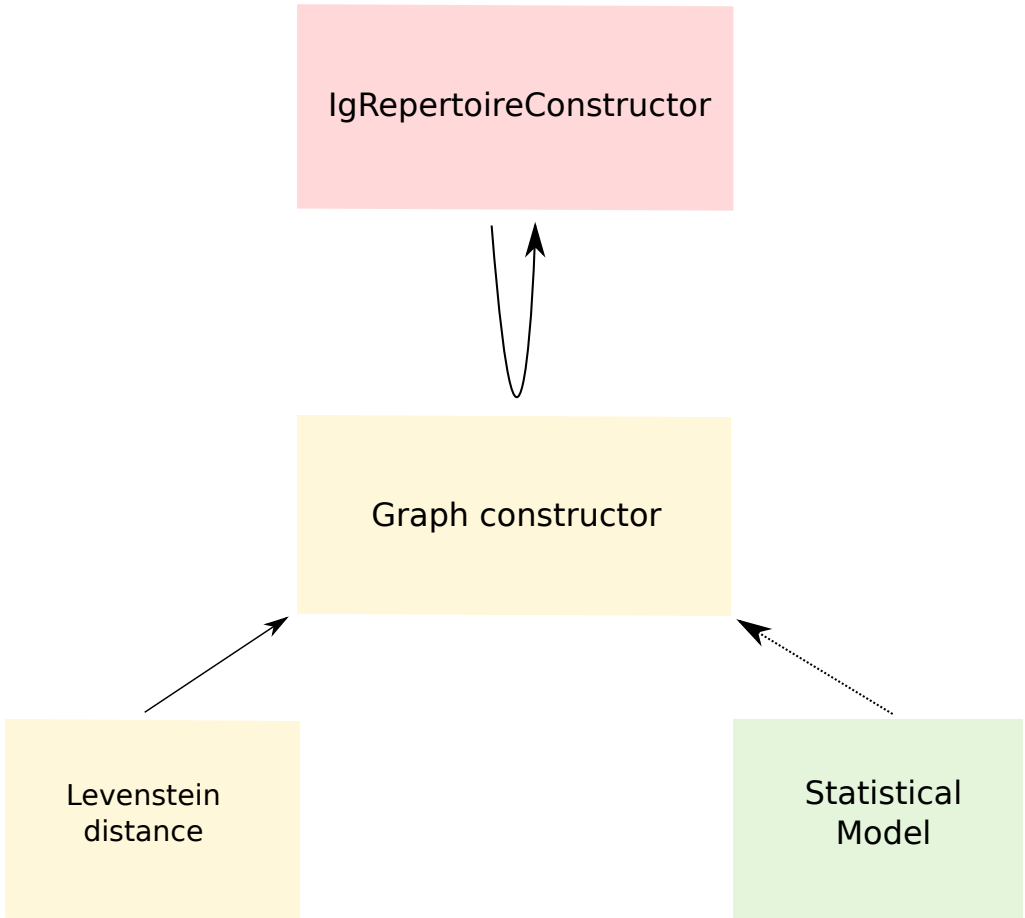


Motivation: IgRepertoireConstructor



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 lots 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 *insertions*?

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 nonproductive 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 *insertions* at the end of a gen.

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 nonproductive 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 *insertions* at the end of a gen.
- 2865(!) parametrs 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 *insertions*

include the task of distinguishing "accidental" and "biological" events and hence require some additional knowledge about the structure of the repertoire.

Considering that and the questions about the article let's firstly concentrate on a simplier problem of seeking **correlation between cleavage and specific gene segments.**

Outline

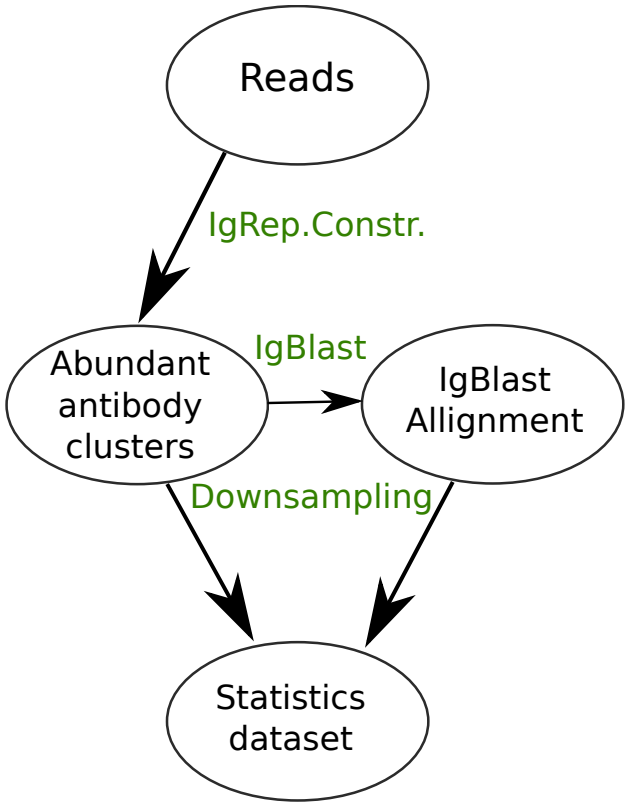
Cleavage and specific gene segments

Problem: In the datasets not only V(D)J-recombinations effects are reflected, but also the result of secondary mutations.

Cleavage and specific gene segments

Problem: In the datasets not only `V(D)J`-recombinations effects are reflected, but also the result of secondary mutations.

- Consider a merged pairs dataset.
- Use `IgRepertoireConstructor` for this dataset and consider highly abundant antibody clusters of the constructed repertoire.
- Apply `IgBlast` to reads from highly abundant clusters.
- Downsampling: consider only such reads, that have alignment score of their segments not less than a threshold according to `IgBlast`.



Cleavage and specific gene segments: V-genes

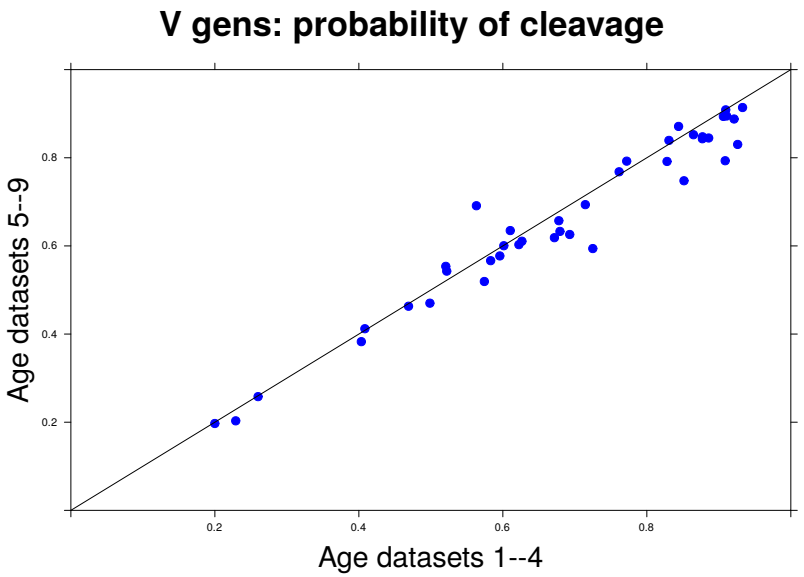


Figure : Age datasets. The point — is the gen. Pearson correlation is 0.98.

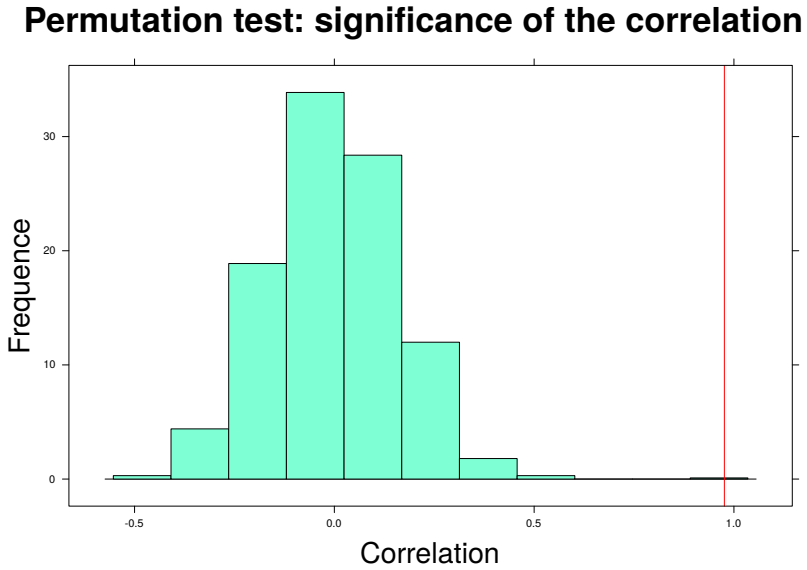


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

Cleavage and specific gene segments: V-genes

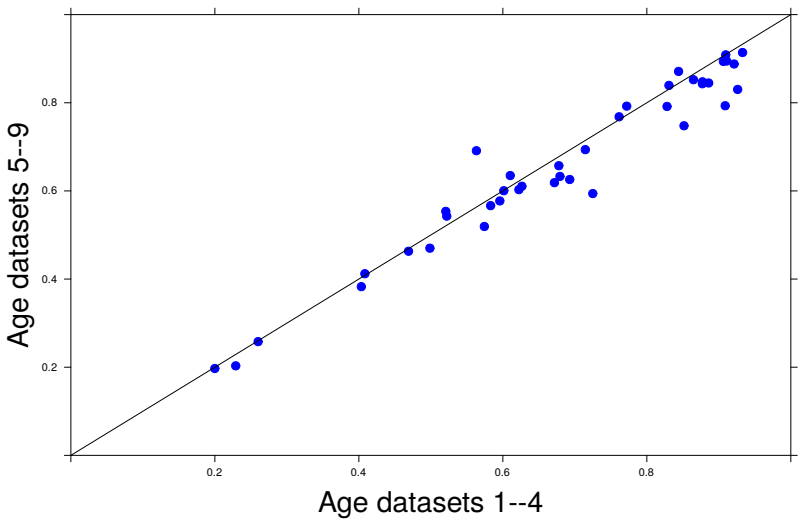


Figure : Age datasets. The point — is the gen. Pearson correlation is 0.98.

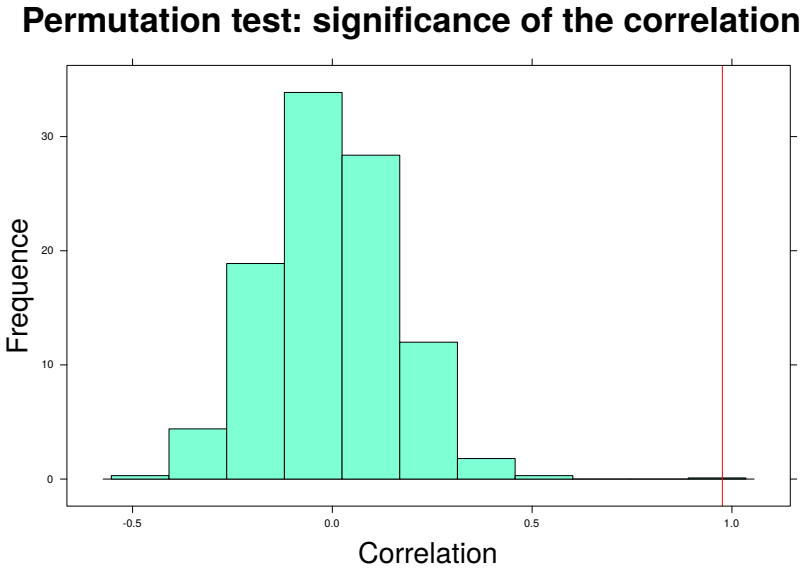


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

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

Cleavage and specific gene segments: V-genes

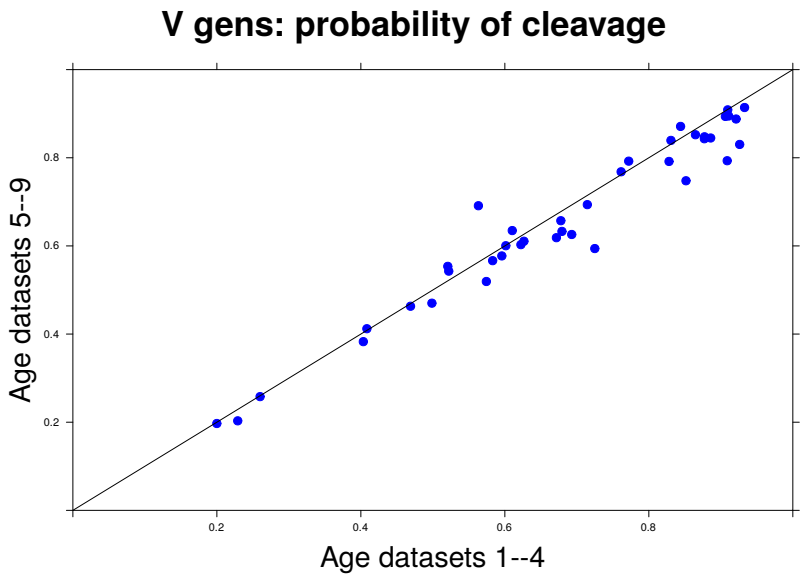


Figure : Age datasets. The point — is the gen. Pearson correlation is 0.98.

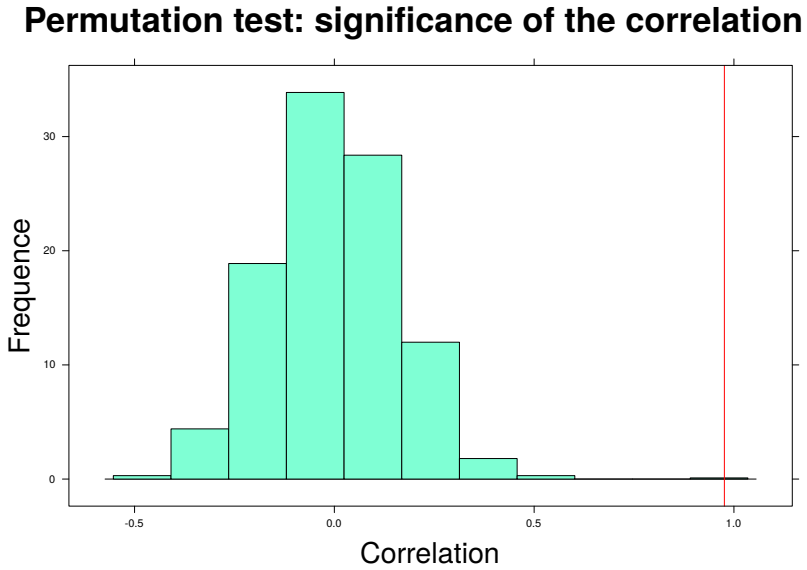


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

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

Remark: No obvious way to clusterize V-genes effectively.

Further goals to clusterize genes

It is reasonable to seek a way to clusterize V-genes by

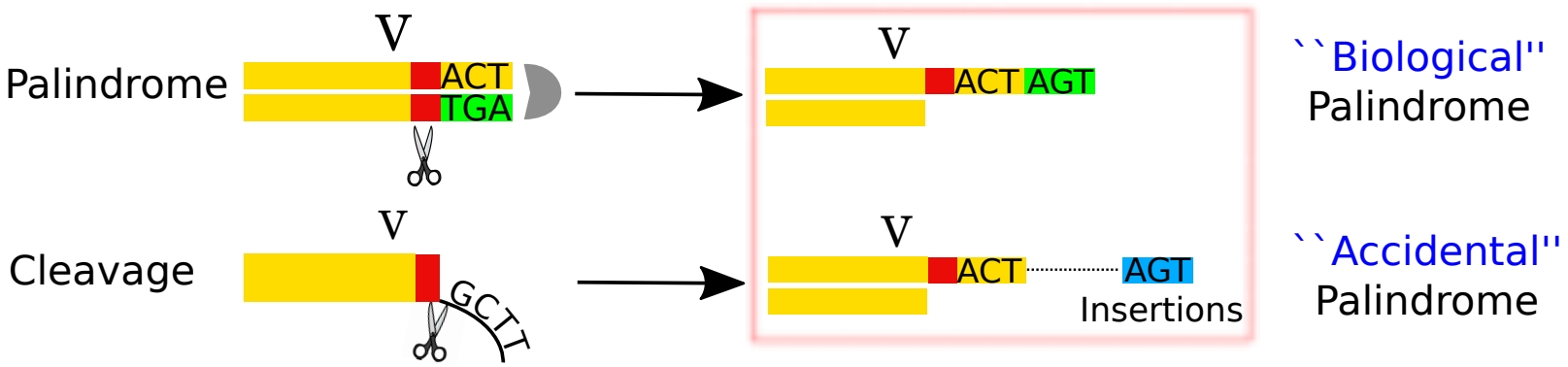
- *palindromes* length;
- GC-content;
- different type of genes (V vs J *etc. . .*).

Outline

Two types of palindromes

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

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



Two types of palindromes

- The simplest model is that any nucletide 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 $\mathrm{Geom}(3/4)$ distribution in log scale:

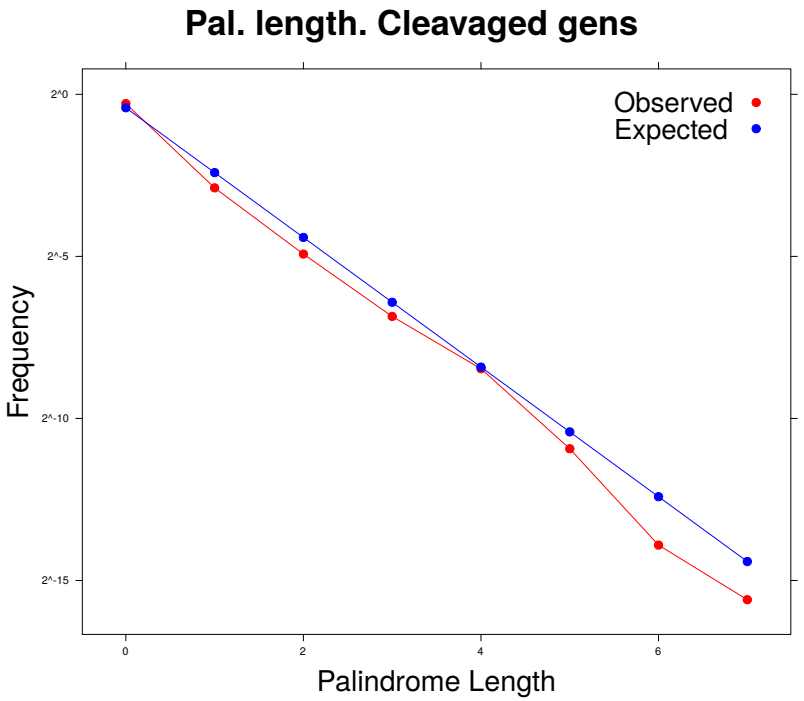


Figure : The mean of length is 0.33.

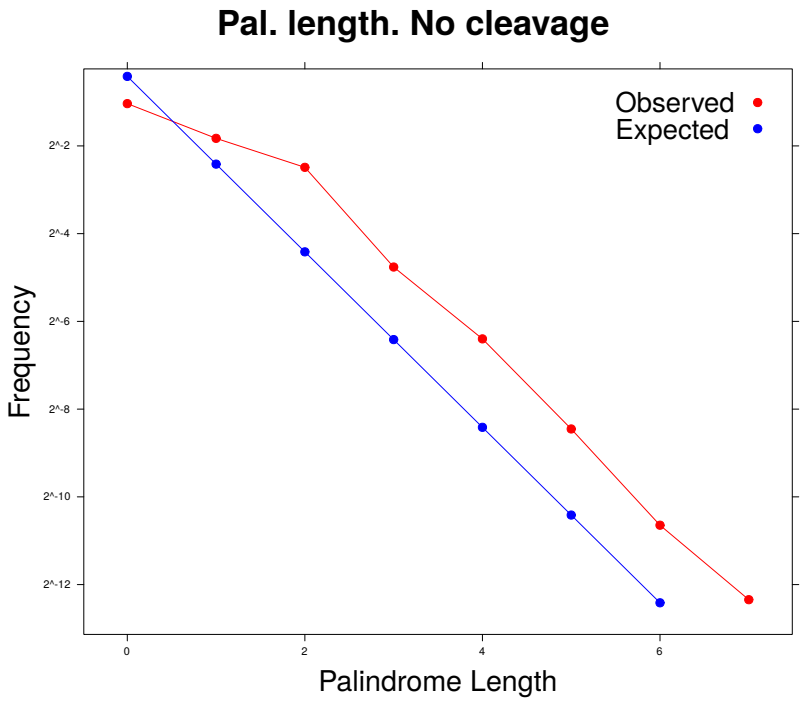
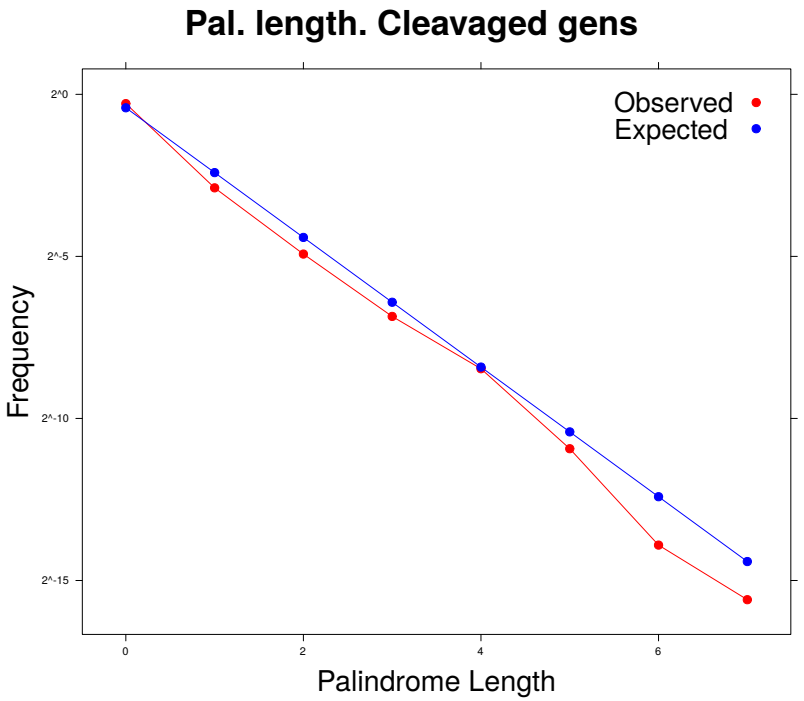


Figure : The mean of length is 0.82.

Two types of palindromes

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



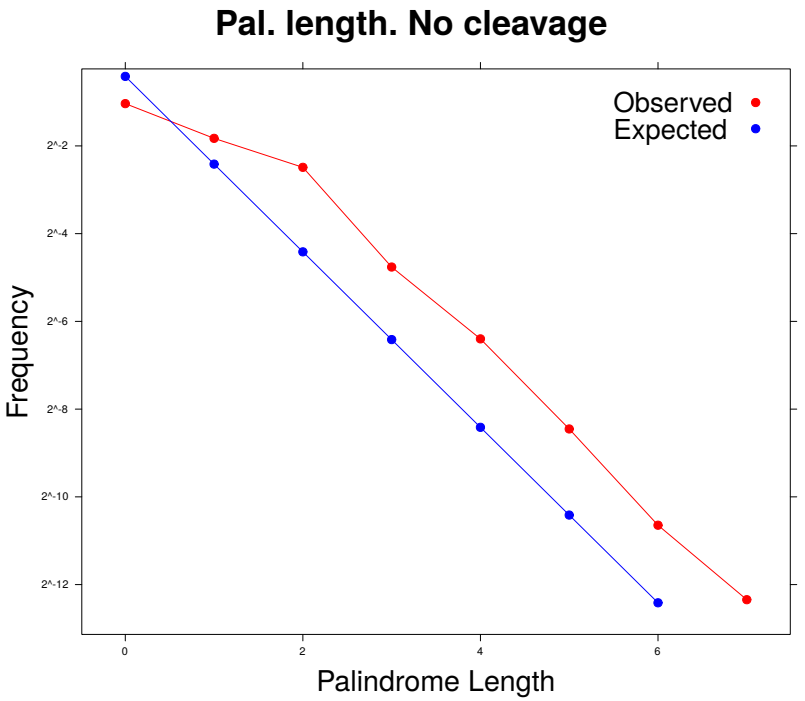


Figure : The mean of length is 0.33.

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.
- To construct more adequate model for nucleotide distribution.