Statistical Analysis of Illumina merged pair reads

Andrew Bzikadze
seryrzu@gmail.com

September 13, 2015

Outline

Introduction

Research direction: analysis of input merged pair reads using statistical and simulation methods.

Introduction

Research direction: analysis of input merged pair reads using statistical and simulation methods.

Questions:

- What is the distribution law of nucleotide subsequences of merged pair reads?
- Is there any correlation between biological events (for instance, between *cleavage* / *palindromes* and specific gene segments)?
- What model describes biological events the best (for example, *insertions* at the `VD-`, `DJ-` junction)?

Motivation

Motivation: knowledge of the distribution of nucleotide sequences potentially helps with

- Simulation of pair reads: improvements of `IgSimulator`.
- Dealing with Clonal Trees. ???
- Comparison of different antibody repertoires.
- String metrics for measuring the difference between two sequences: improvements of `IgRepertoireConstructor`.

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 CDR3.
- Suggested model sets joint distribution over the set of discrete variables: *identities* of V-, D-, J- gens, 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 CDR3.
- Suggested model sets joint distribution over the set of discrete variables: *identities* of V-, D-, J- gens, 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?

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?

Difficult to answer, should start with something simpler.

Two types of events

It is reasonable to classify events (palindromes, cleavage, etc.) into "biological" and "accidental".
Example: if we detect a palindrome at the end of a segment, while clavage has also happened, then we defenitely detect an *accidental* (noise) palindrome.

The goals

- correlations between palindromes and specific gene segments;
- distribution law of nucleotide subsequences

include the task of classification the events into two types 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 ????.

Cleavage and specific gene segments

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

Scheme:

- Consider a merged pair dataset.

Cleavage and specific gene segments

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

Scheme:

- Consider a merged pair dataset.
- Use `IgRepertoireConstructor` for this dataset and consider highly abundant antibody clusters of the constructed repertoire.

Cleavage and specific gene segments

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

Scheme:

- Consider a merged pair dataset.
- Use `IgRepertoireConstructor` for this dataset and consider highly abundant antibody clusters of the constructed repertoire.
- Apply `IgBlast` to reads from highly abundant clusters.

Cleavage and specific gene segments

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

Scheme:

- Consider a merged pair 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`.

Scheme is invariant of biological event we are interested in.

Cleavage and specific gene segments: V-gens

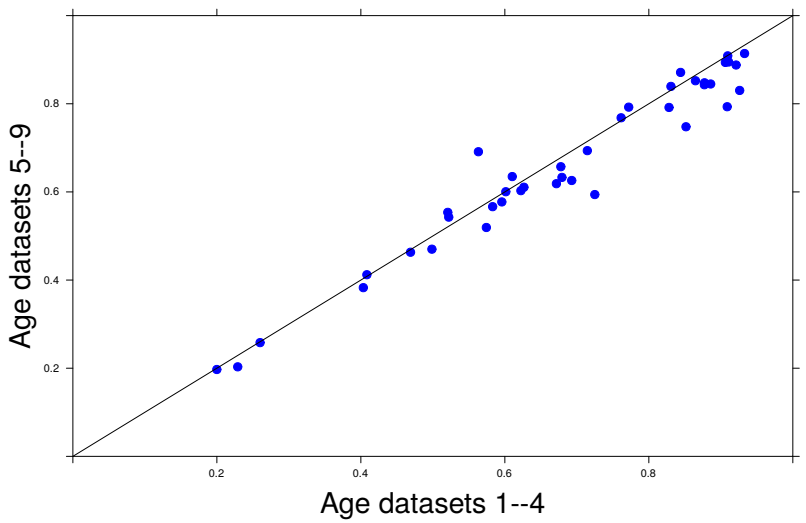


Figure : Age datasets. The point — is the gen. Pearson correlation is 0.98.

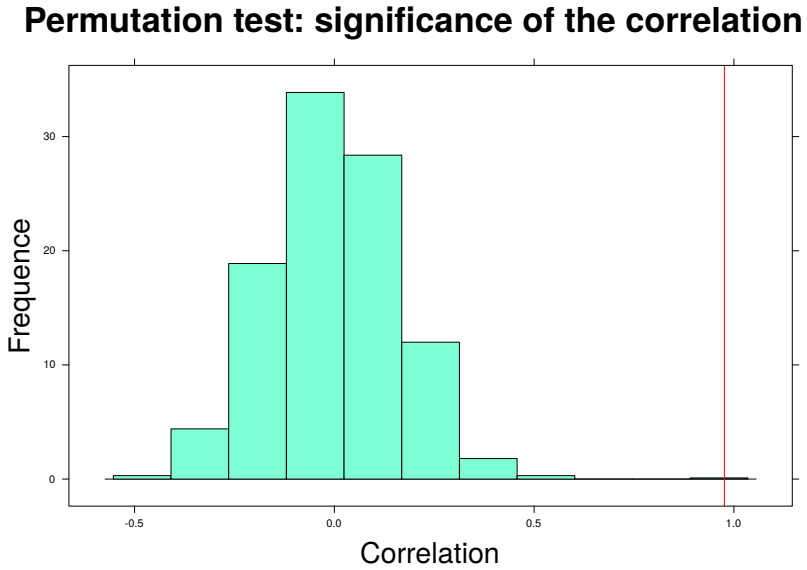


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

Cleavage and specific gene segments: V-gens

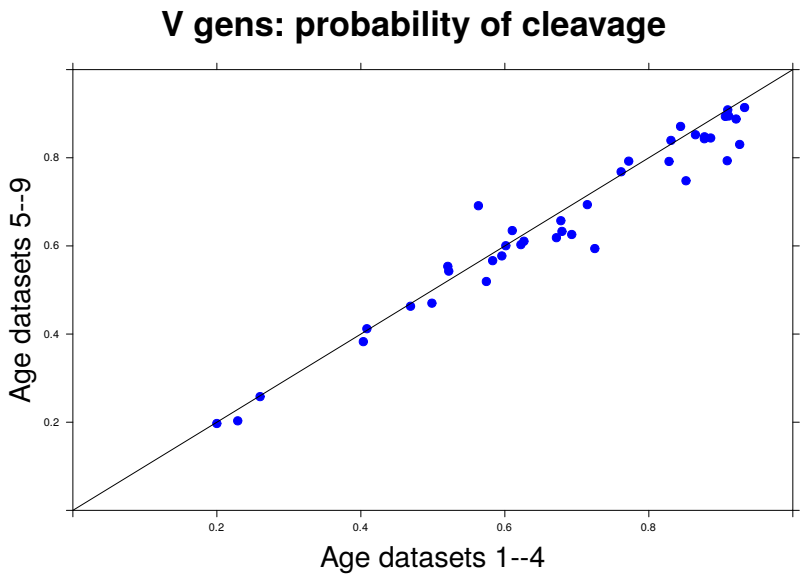


Figure : Age datasets. The point — is the gen. Pearson correlation is 0.98.

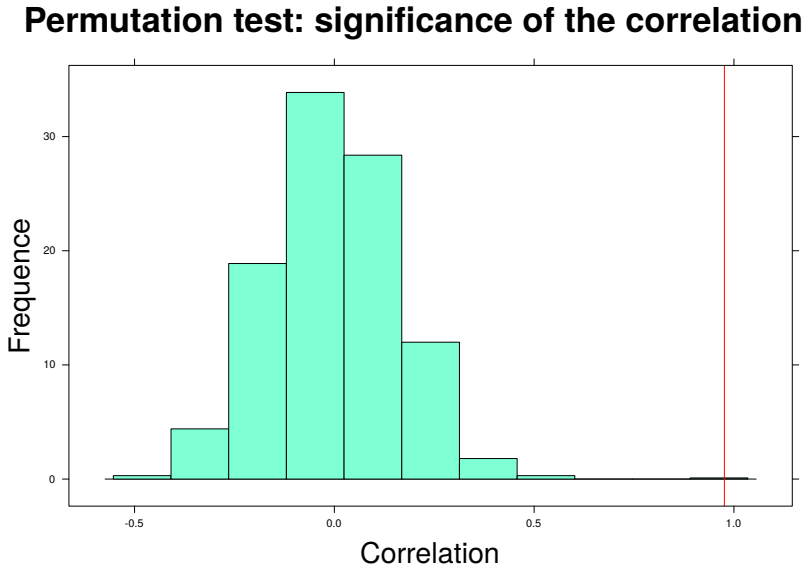


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

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

Cleavage and specific gene segments: V-gens

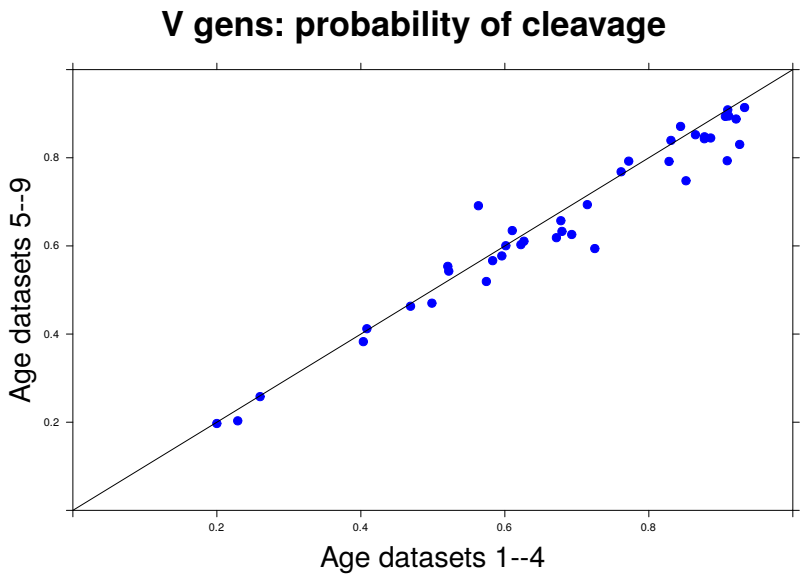


Figure : Age datasets. The point — is the gen. Pearson correlation is 0.98.

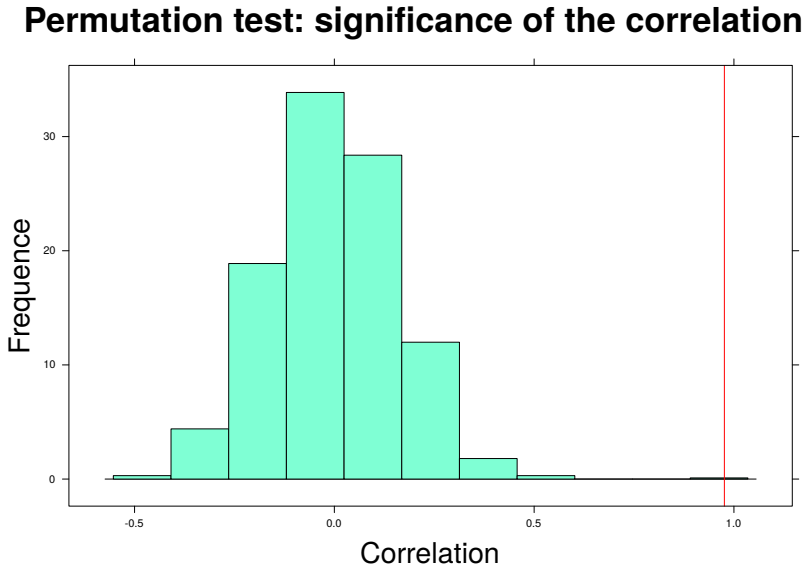


Figure : Histogram of statistics of permutation test that shows the significance of Pearson correlation.

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

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

Further goals to clusterize gens

As long as the probabilities of cleavage for gens don't introduce an obvious way to clusterize V-gens (and hence to reduce the number of parameters in the model for distribution) the next steps will be correlations between the gen and

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

Outline

Two types of palindromes

It is known that if cleavage took place, then no palindrome can happen. This is only partially true. "Accidental" palindrome can still happen, but it won't have "biological" nature.

- The simplest model is that nucletides are distributed **uniformly**.
- In that model the length of "accidental" palindrome has $\mathrm{Geom}(3/4)$ distribution.

Two types of palindromes

Emperical (Age-datasets) and theoretical distribution in log scale:

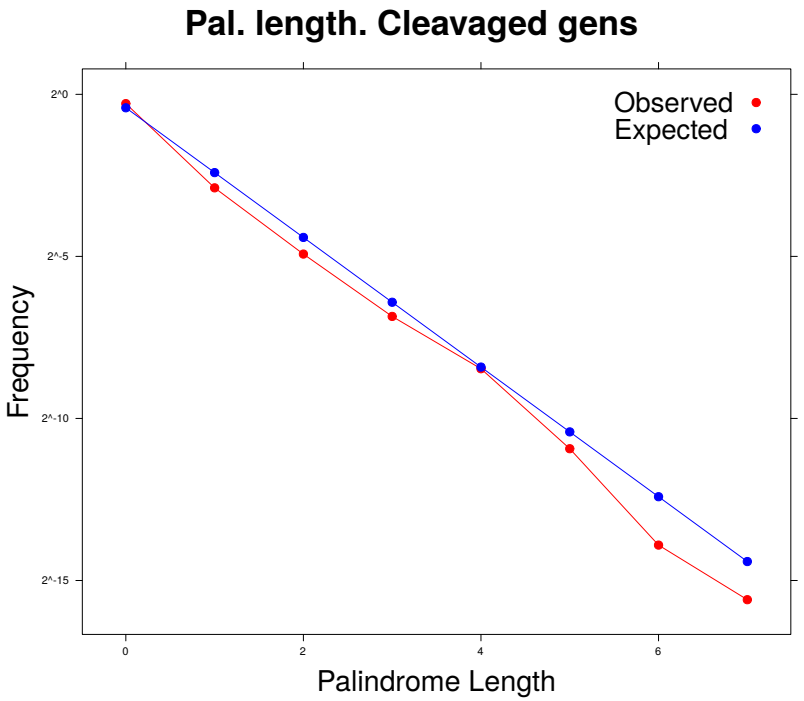


Figure : The mean of length is 0.33.

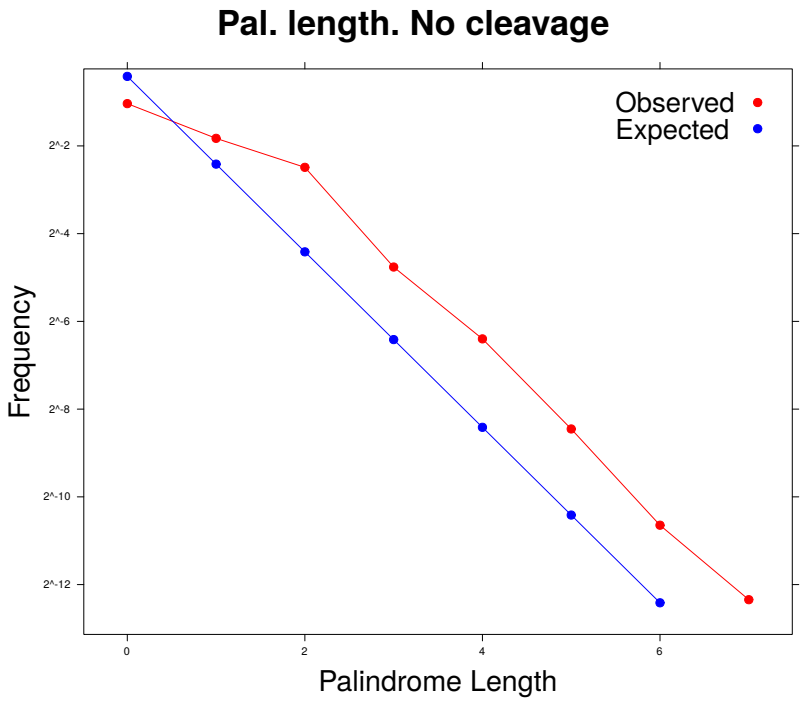
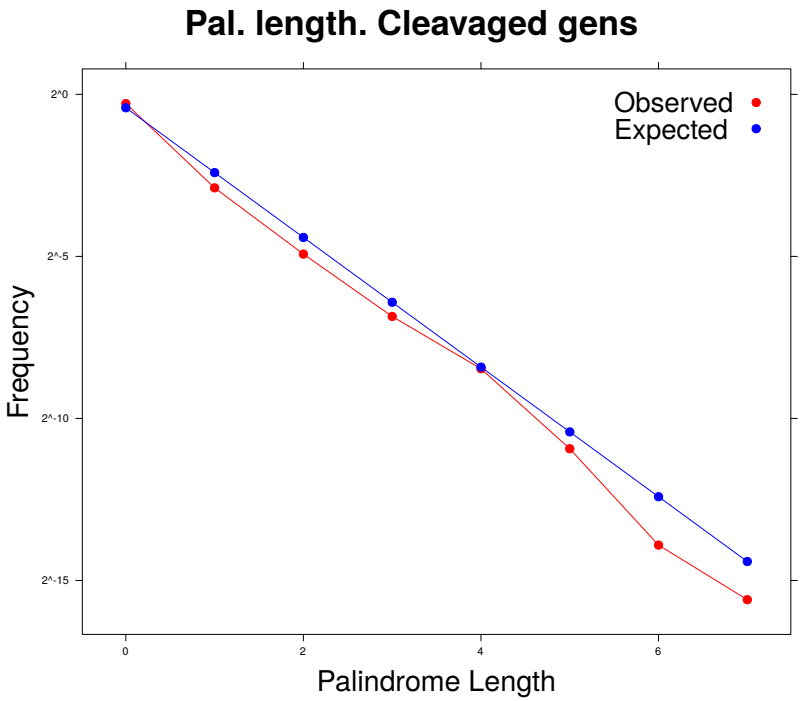


Figure : The mean of length is 0.82.

Two types of palindromes

Emperical (Age-datasets) and theoretical distribution in log scale:



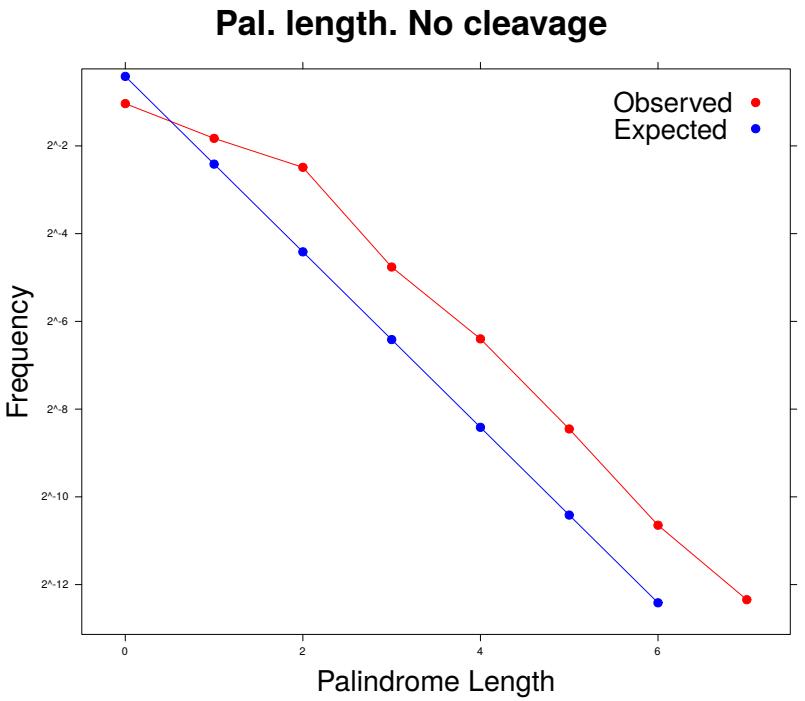


Figure : The mean of length is 0.33.

Figure : The mean of length is 0.82.

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

What else about the palindromes

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

Next steps

To sum up let's revisit main important questions:

- What is the distribution of the nucleotides?
- What events are correlated strongly / weakly with each other?
- How to distinguish "biological" and "accidental" events?

Also checking whether the model advised in the article is adequate for B-cells (and trying to reduce the number of the parametrs) seems to be reasonable.

Further analysis will be concentrated on this problems.