LEIDEN UNIVERSITY MEDICAL CENTER

Usage of k-mer profiles in NGS data

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Calculate distances

We frequently want to know how datasets are related.

Introduction

Calculate distances

We frequently want to know how datasets are related.

Between reference sequences:

- Phylogeny.

Introduction

Calculate distances

We frequently want to know how datasets are related.

Between reference sequences:

- Phylogeny.

Between raw files:

- Phylogeny.
- Quality control.
- Potential measure for the quality of a de novo assembly.

Introduction

Calculate distances

We frequently want to know how datasets are related.

Between reference sequences:

- Phylogeny.

Between raw files:

- Phylogeny.
- Quality control.
- Potential measure for the quality of a de novo assembly.

In combination:

- Metagenomics.
- Quality control.

Introduction

Counting k-mers

We choose a k and count all occurrences of substrings of length k.

Introduction

Counting k-mers

We choose a k and count all occurrences of substrings of length k.

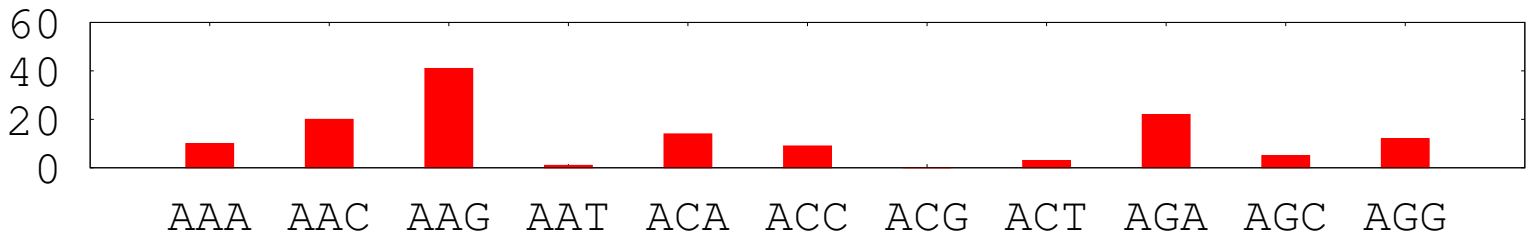


Figure 1: A profile of 3-mer counts.

In Figure 1 we see a part of 3-mer counts; **AAA** occurs 10 times, **AAC** occurs 20 times, etc.

Introduction

Why k-mers?

The usage of k-mers is appealing since:

- They are easy to work with.
- Fingerprinting of samples is possible (even for a small k-mer length (around 10)).
- Can be used to make a database to match against.

Introduction

Why k-mers?

The usage of k-mers is appealing since:

- They are easy to work with.
- Fingerprinting of samples is possible (even for a small k-mer length (around 10)).
- Can be used to make a database to match against.

Some possible extensions:

- How realistic is it to disassemble a "mixed" k-mer profile?

Introduction

Choosing k

k can not be too small:

Introduction

Choosing k

k can not be too small:

- $k = 1$ will result in loss of all subsequence information.
- $k = 2$ will give you information about di-nucleotides.
  - But, pattern growth needs also position information.
- There is only one unique 10-mer in Human (hg18).

Introduction

Choosing k

k can not be too small:

- $k = 1$ will result in loss of all subsequence information.
- $k = 2$ will give you information about di-nucleotides.
  - But, pattern growth needs also position information.
- There is only one unique 10-mer in Human (hg18).

But, k can not be too large either:

Introduction

Choosing k

k can not be too small:

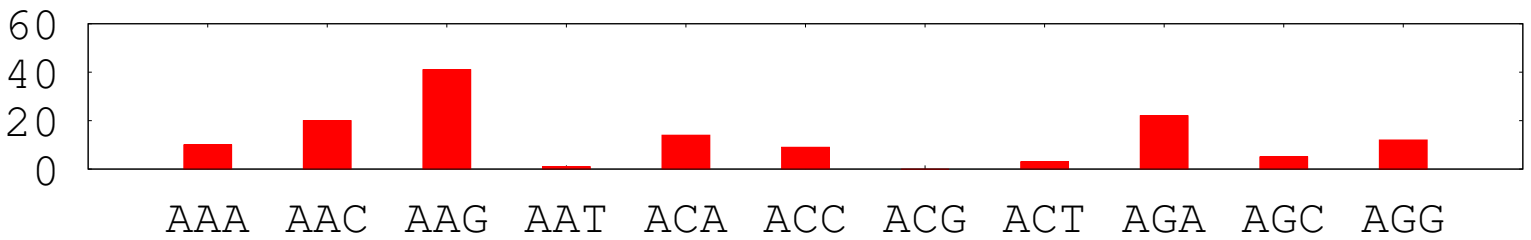
- $k = 1$ will result in loss of all subsequence information.
- $k = 2$ will give you information about di-nucleotides.
  - But, pattern growth needs also position information.
- There is only one unique 10-mer in Human (hg18).

But, k can not be too large either:

- Almost all 18-mers are unique in the Human genome.
- Since this is not error tolerant:
  - Read errors.
  - Assembly errors.

Introduction

Comparing k-mer profiles



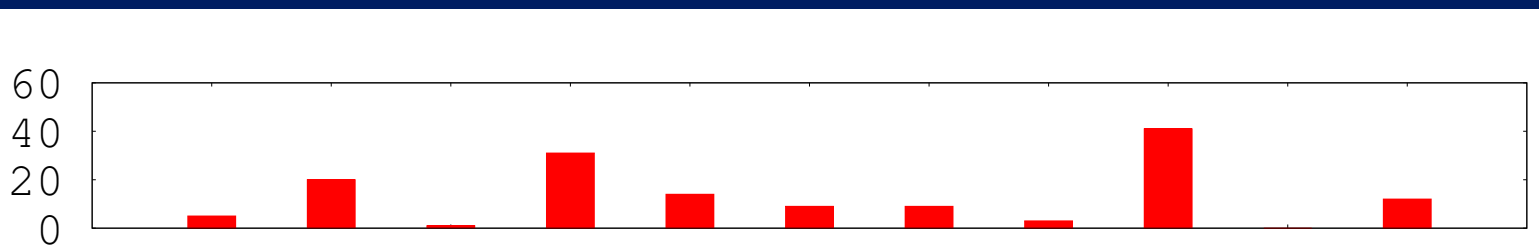
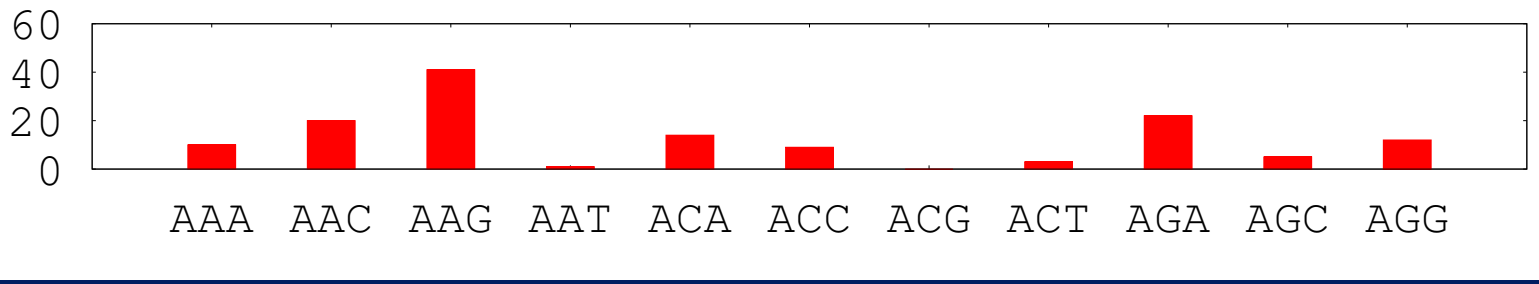


Figure 2: Two profiles of k-mer counts.

Introduction

Comparing k-mer profiles



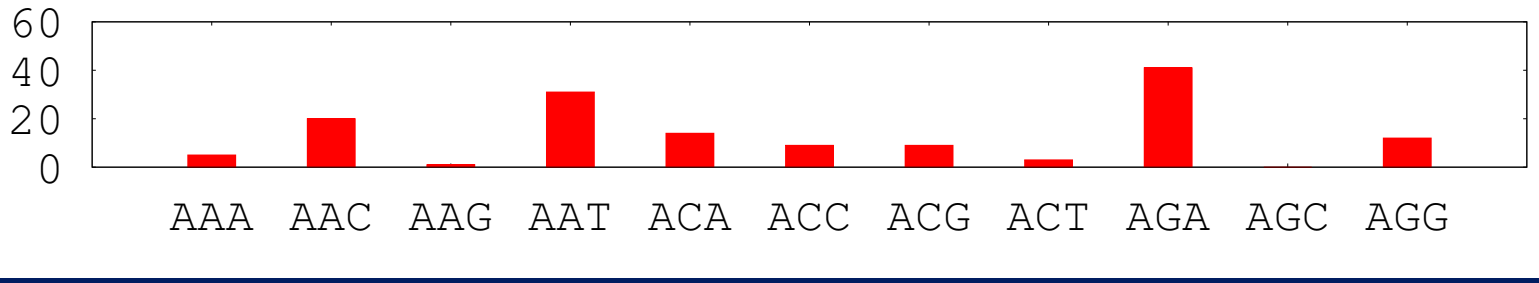


Figure 2: Two profiles of k-mer counts.

How to express this difference with one value.

Introduction

Pairwise distance function

We use the following function:

$$f(x,y) = \frac{|x-y|}{(x+1)(y+1)}$$

Introduction

Pairwise distance function

We use the following function:

$$f(x,y) = \frac{|x-y|}{(x+1)(y+1)}$$

Properties:

- $f(0,1) = \frac{1}{2}$
- $f(0,1) > f(7,8)$

Introduction

Pairwise distance function

We use the following function:

$$f(x,y) = \frac{|x-y|}{(x+1)(y+1)}$$

Properties:

- $f(0,1) = \frac{1}{2}$
- $f(0,1) > f(7,8)$

This is desirable:

- The fact that a k-mer is not present is more important than the number of times it is present.
- Differences in the low end of the spectrum are more important than ones at the high end.

Introduction

Multiset distance function

Let f be a function $f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with finite supremum M and the following properties:

$$
\begin{array}{lcll}
f(x,y) & = & f(y,x) & \text{for all } x,y \in \mathbb{R}_{\geq 0} \\
f(x,x) & = & 0 & \text{for all } x \in \mathbb{R}_{\geq 0} \\
f(x,0) & \geq & M/2 & \text{for all } x \in \mathbb{R}_{>0} \\
f(x,y) & \leq & f(x,z) + f(z,y) & \text{for all } x,y,z \in \mathbb{R}_{\geq 0}
\end{array}
$$

Introduction

Multiset distance function

Let f be a function $f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with finite supremum M and the following properties:

$$\begin{array}{lcll} f(x,y) & = & f(y,x) & \text{for all } x,y \in \mathbb{R}_{\geq 0} \\ f(x,x) & = & 0 & \text{for all } x \in \mathbb{R}_{\geq 0} \\ f(x,0) & \geq & M/2 & \text{for all } x \in \mathbb{R}_{>0} \\ f(x,y) & \leq & f(x,z)+f(z,y) & \text{for all } x,y,z \in \mathbb{R}_{\geq 0} \end{array}$$

For a multiset X, let $S(X)$ denote its underlying set. For multisets X, Y with $S(X), S(Y) \subseteq \{1, 2, \ldots, n\}$ we define

$$d_f(X,Y) = \frac{\sum_{i=1}^{n} f(x_i, y_i)}{|S(X) \cup S(Y)| + 1}$$

Reference sequences

Results for reference sequences

Comparisons from a previous study.

		Human			
		0	1	2	≥ 3
Chimp	0	150,783,349	4,486,933	1,216,093	498,090
	1	3,212,656	7,352,318	3,737,739	2,333,341
	2	602,927	2,621,970	4,011,169	4,907,515
	≥ 3	145,530	950,955	2,697,230	78,877,641

Table 1: Differences between Human and Chimpanzee ($k = 14$)

Reference sequences

Results for reference sequences

Comparisons from a previous study.

		Human			
		0	1	2	≥ 3
Chimp	0	150,783,349	4,486,933	1,216,093	498,090
	1	3,212,656	7,352,318	3,737,739	2,333,341
	2	602,927	2,621,970	4,011,169	4,907,515
	≥ 3	145,530	950,955	2,697,230	78,877,641

Table 1: Differences between Human and Chimpanzee ($k = 14$)

Observations:

- There are over six million 14-mers that can identify Human DNA in a mix of Human and Chimpanzee DNA.
- Nearly half a million 14-mers are abundant in Human, but are not present in the Chimpanzee.

Reference sequences

Results for reference sequences

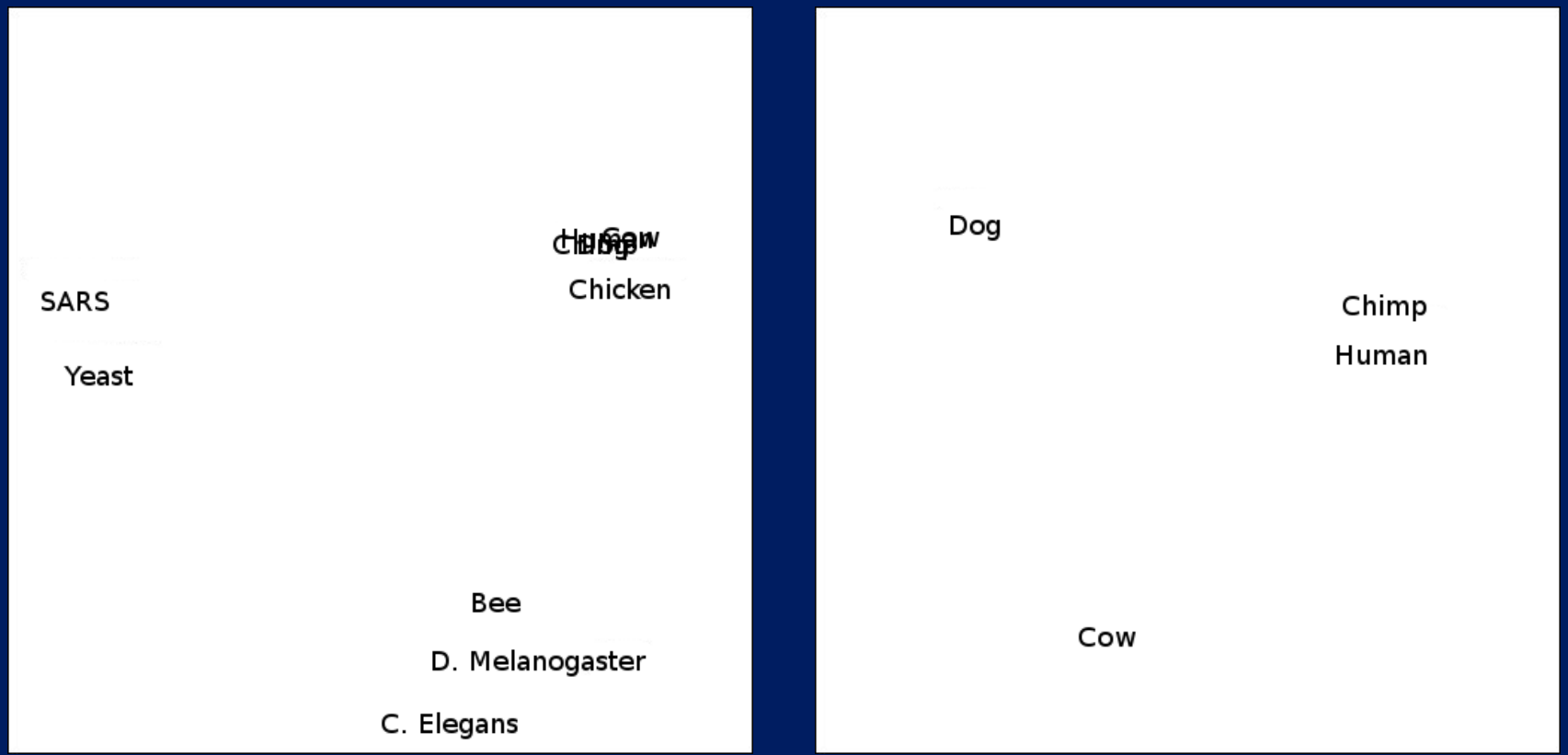


Figure 3: Left: ten species. Right: four mammals.

Application on raw data

Fastq files

We want to extend this to *raw* data.

Application on raw data

Fastq files

We want to extend this to *raw* data.

- Often we do not have a reference sequence.
- Sometimes we do not know which species we are looking at.

Application on raw data

Fastq files

We want to extend this to *raw* data.

- Often we do not have a reference sequence.
- Sometimes we do not know which species we are looking at.

Characteristics:

- Short reads.
- Unique k-mers can be present more than once.
- k-mers that are present in the sample can be missing in this file.

Application on raw data

Fastq files

We want to extend this to *raw* data.

- Often we do not have a reference sequence.
- Sometimes we do not know which species we are looking at.

Characteristics:

- Short reads.
- Unique k-mers can be present more than once.
- k-mers that are present in the sample can be missing in this file.

Things to think about:

- Normalisation.

Application on raw data

A dataset of owls

Figure 4: Western brown fish owl.

Four samples, paired end.

Application on raw data

First results

		1		2		3		4	
		1	2	1	2	1	2	1	2
1	1	0.000							
	2	0.257	0.000						
2	1	0.649	0.655	0.000					
	2	0.663	0.670	0.206	0.000				
3	1	0.427	0.428	0.754	0.784	0.000			
	2	0.430	0.427	0.763	0.786	0.364	0.000		
4	1	0.552	0.556	0.254	0.262	0.774	0.783	0.000	
	2	0.584	0.589	0.260	0.254	0.814	0.822	0.146	0.000

Table 2: k-mer difference between 8 fastq files.

Application on raw data

First results

		1		2		3		4	
		1	2	1	2	1	2	1	2
1	1	0.000							
	2	0.257	0.000						
2	1	0.649	0.655	0.000					
	2	0.663	0.670	0.206	0.000				
3	1	0.427	0.428	0.754	0.784	0.000			
	2	0.430	0.427	0.763	0.786	0.364	0.000		
4	1	0.552	0.556	0.254	0.262	0.774	0.783	0.000	
	2	0.584	0.589	0.260	0.254	0.814	0.822	0.146	0.000

Table 2: k-mer difference between 8 fastq files.

Observations:

- The distance between two files of one sample is low.

Application on raw data

First results

		1		2		3		4	
		1	2	1	2	1	2	1	2
1	1	0.000							
	2	0.257	0.000						
2	1	0.649	0.655	0.000					
	2	0.663	0.670	0.206	0.000				
3	1	0.427	0.428	0.754	0.784	0.000			
	2	0.430	0.427	0.763	0.786	0.364	0.000		
4	1	0.552	0.556	0.254	0.262	0.774	0.783	0.000	
	2	0.584	0.589	0.260	0.254	0.814	0.822	0.146	0.000

Table 2: k-mer difference between 8 fastq files.

Observations:

- The distance between two files of one sample is low.
- There is an apparent low distance between sample 2 and 4.

Application on raw data

Sizes of the datasets

The sizes of the datasets differ quite a lot.

sample	size
1	6.1
2	31.6
3	3.9
4	30.7

Table 3: Sizes of the samples in gigabytes.

Application on raw data

Sizes of the datasets

The sizes of the datasets differ quite a lot.

sample	size
1	6.1
2	31.6
3	3.9
4	30.7

Table 3: Sizes of the samples in gigabytes.

The chance that a k-mer is present in a small dataset is low.

- Sample 2 and 4 are more alike than sample 3 internally.
- This may give an indication on how deep to sequence.
- Apparently, global normalisation is not effective.

Variations on the theme

Possible ways to deal with this data

Throwing out data:

Variations on the theme

Possible ways to deal with this data

Throwing out data:

- Random.
  - Not in favour, probably the “random” selection of sequences by the sequencer is not random.

Variations on the theme

Possible ways to deal with this data

Throwing out data:

- Random.
  - Not in favour, probably the “random” selection of sequences by the sequencer is not random.
- Low abundant sequences.
  - Worth a try.

Variations on the theme

Possible ways to deal with this data

Throwing out data:

- Random.
  - Not in favour, probably the “random” selection of sequences by the sequencer is not random.
- Low abundant sequences.
  - Worth a try.

Different distance measures:

- Euclidean distance
- Positive multiset
- Relative multiset

Variations on the theme

Possible ways to deal with this data

Throwing out data:

- Random.
  - Not in favour, probably the “random” selection of sequences by the sequencer is not random.
- Low abundant sequences.
  - Worth a try.

Different distance measures:

- Euclidean distance No good results
- Positive multiset
- Relative multiset

Variations on the theme

Possible ways to deal with this data

Throwing out data:

- Random.
  - Not in favour, probably the “random” selection of sequences by the sequencer is not random.
- Low abundant sequences.
  - Worth a try.

Different distance measures:

- Euclidean distance — No good results
- Positive multiset — Relatively good results
- Relative multiset

Variations on the theme

Possible ways to deal with this data

Throwing out data:

- Random.
  - Not in favour, probably the “random” selection of sequences by the sequencer is not random.
- Low abundant sequences.
  - Worth a try.

Different distance measures:

- Euclidean distance — No good results
- Positive multiset — Relatively good results
- Relative multiset — Not done yet

Variations on the theme

Possible ways to deal with this data

Throwing out data:

- Random.
  - Not in favour, probably the “random” selection of sequences by the sequencer is not random.
- Low abundant sequences.
  - Worth a try.

Different distance measures:

- Euclidean distance — No good results
- Positive multiset — Relatively good results
- Relative multiset — Not done yet

The last algorithm is rather expensive.

Variations on the theme

Look at relative k-mer occurrences

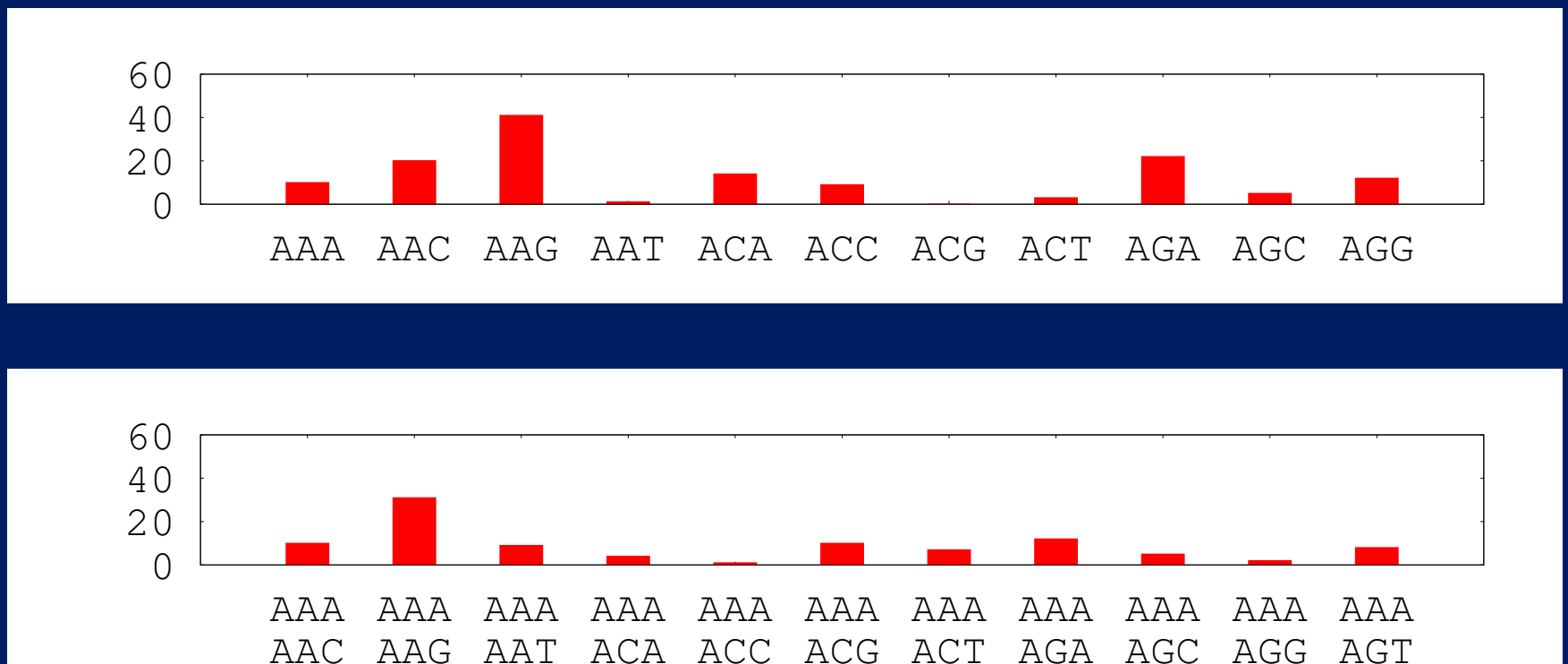


Figure 5: Conversion to differences in counts.

Questions?

Acknowledgements:

Walter Kosters
Hendrik Jan Hoogeboom
Ken Kraaijeveld
Johan den Dunnen

https://humgenprojects.lumc.nl/svn/k-mer