Statistical modeling of gene expression using epigenomic signals

Part 1 - Data preparation

Santa Kirezi, Nadia Ponts, Gaël Le Trionnaire, David Causeur

2023-10-17

Contents

1 Introduction

2 Importing data

Epigenomic and transcriptomic signals are stored into three different files, one for each replicate. Data are imported into three `R` objects, of type `list`, named `l_express1`, `l_express2` and `l_express3` for the transcriptomic signals and `l_epigeno1`, `l_epigeno2` and `l_epigeno3` for the epigenomic signals.

The number of signals in all lists is the same, namely the number of genes:

Components of all lists are named using the gene names provided in the raw data files. The 5 first gene names are extracted from list `l_express1` and shown below for illustration:

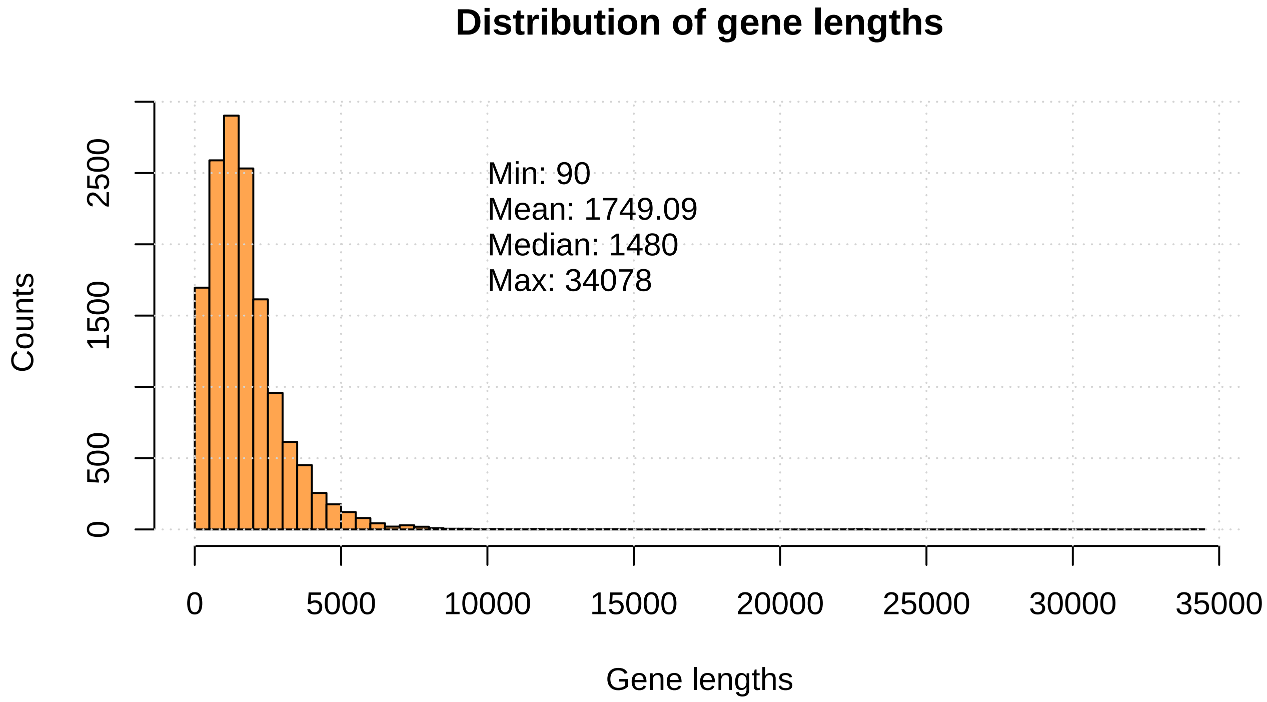
Table 1: Number of signals in all data tables

Express1	Express2	Express3	Epigeno1	Epigeno2	Epigeno3
14145	14145	14145	14145	14145	14145

```r
gene_names <- names(l_express1)
gene_names[1:5]
```

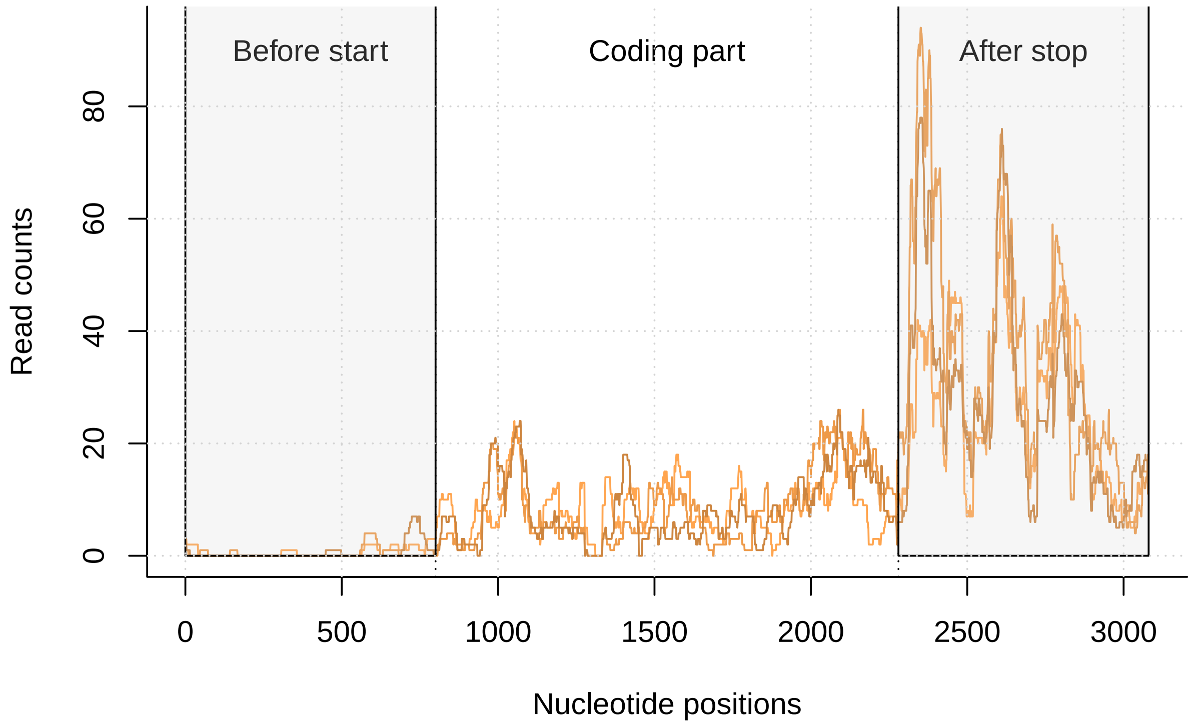
```
[1] "FGRAMPH1_01T00001" "FGRAMPH1_01T00003" "FGRAMPH1_01T00005"
[4] "FGRAMPH1_01T00007" "FGRAMPH1_01T00009"
```

All signals are pointwise measurements, either transcriptomic or epigenomic, at nucleotide positions in-between the start and stop codons of a gene but also at the 800 positions before the start and 800 positions after the stop codon. Therefore, gene lengths can be deduced from the signal lengths by subtracting 1600. The plot below displays the distribution of gene lengths across the genome:



For an illustrative purpose, both the epigenomic and transcriptomic signals of a gene, arbitrarily chosen with length close to the median gene length, are plotted below:

Expression signals for gene FGRAMPH1_01T01949



Epigenomic signals for gene FGRAMPH1_01T01949

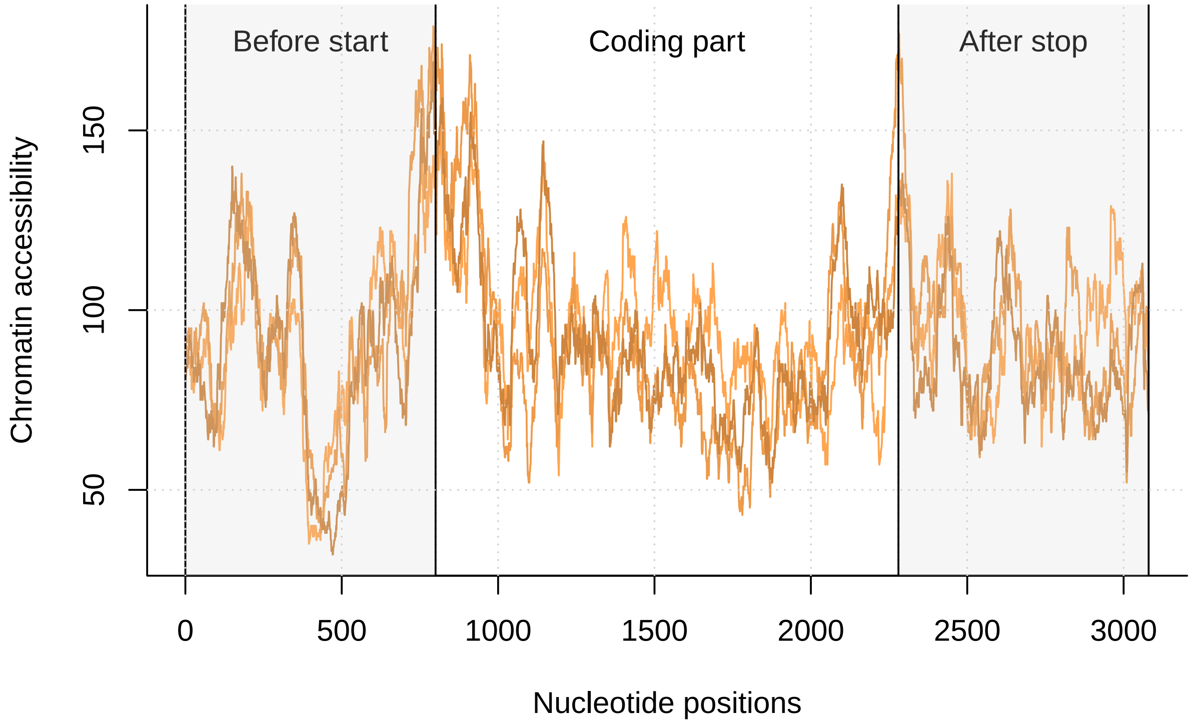


Table 2: Number of genes in each chromosome

chrom	Freq
HG970332	4390
HG970333	3648
HG970334	3085
HG970335	3022

3 Data quality control

In this Section, transcriptomic and epigenomic data are explored to identify genes for which the signals should be suspected of poor quality and therefore removed from the subsequent analysis.

3.1 Overlapping genes

In the raw data, the sequences of nucleotides before the start codon and after the stop codon of each gene has been arbitrarily chosen to be 800 positions long. However, due to the density of the genome, the sequence after one stop codon for a gene may overlap with the sequence before the start codon for another one. This would affect the epigenomic signals measured for genes too close on the genome.

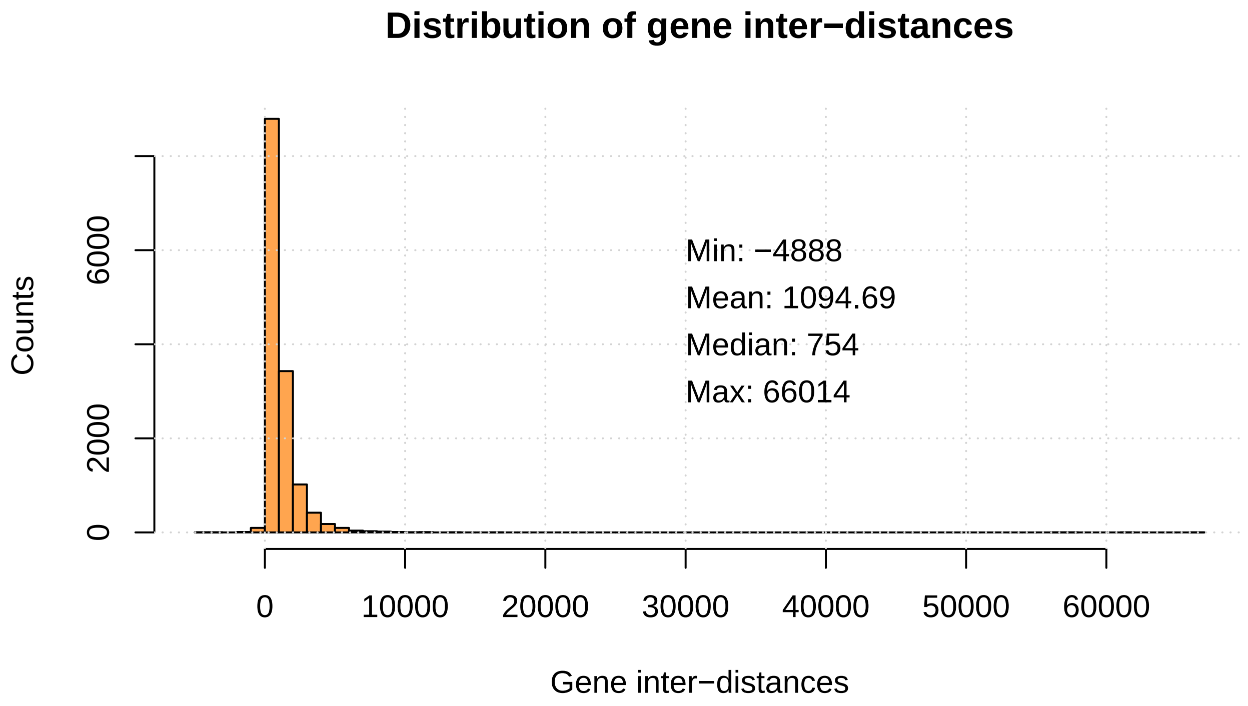
In order to identify genes with a large overlap, we use information about the genome structure available in the file `FungiDB-58_FgraminearumPH-1.UpDown.800bp.sort.bed`. It is deduced from this file that the genome is made of four chromosomes with the following distribution of the numbers of genes:

The distance between two genes will be calculated as the number of nucleotide positions between the stop codon of a gene and the start codon of the following gene.

The following plot displays an histogram of the gene inter-distances:

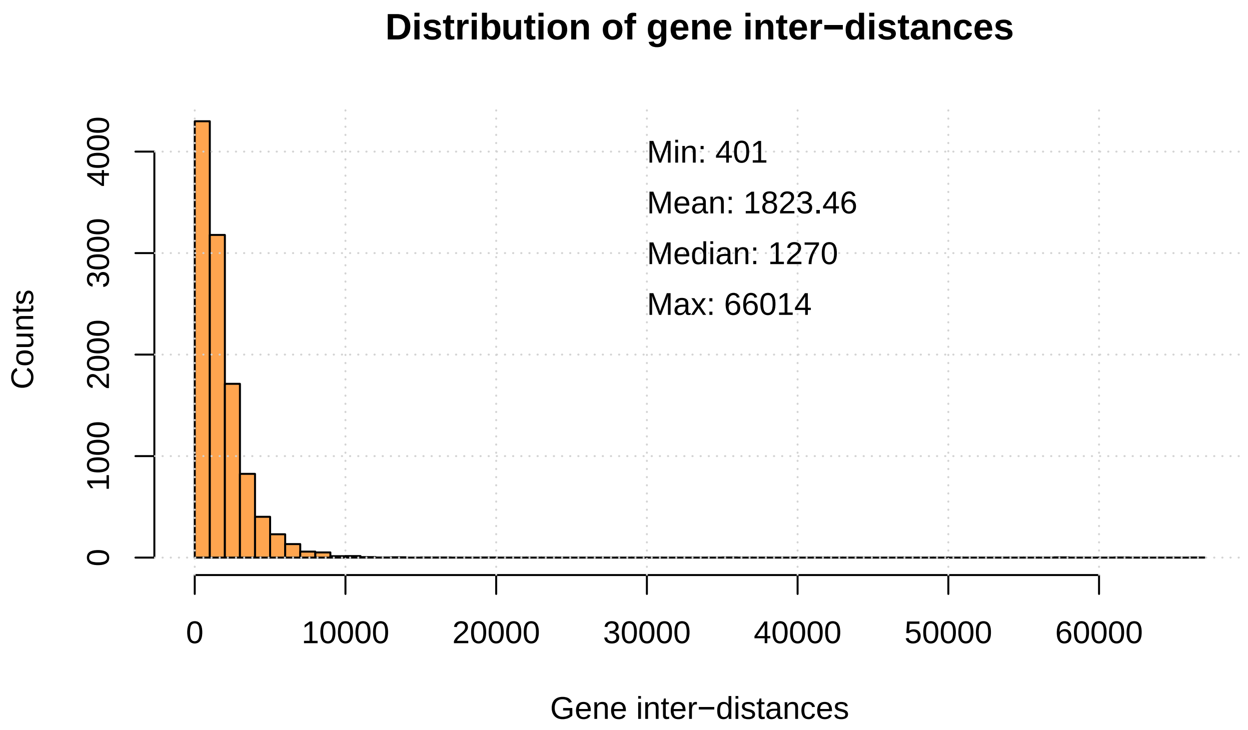
Table 3: Start and end positions of two overlapping genes

	Start	End
FGRAMPH1_01T00045	70886	71196
FGRAMPH1_01T00043	70921	71208



Note that, in some few cases, inter-gene distance can be negative. For example, the two following genes are overlapping:

Since the sequence before the start codon is suspected to be of major importance to explain the transcription, genes whose start codon is closer than 400 nucleotide positions from the stop codon of the preceding gene are removed from the data.



The number of remaining genes in the data is now 10936.

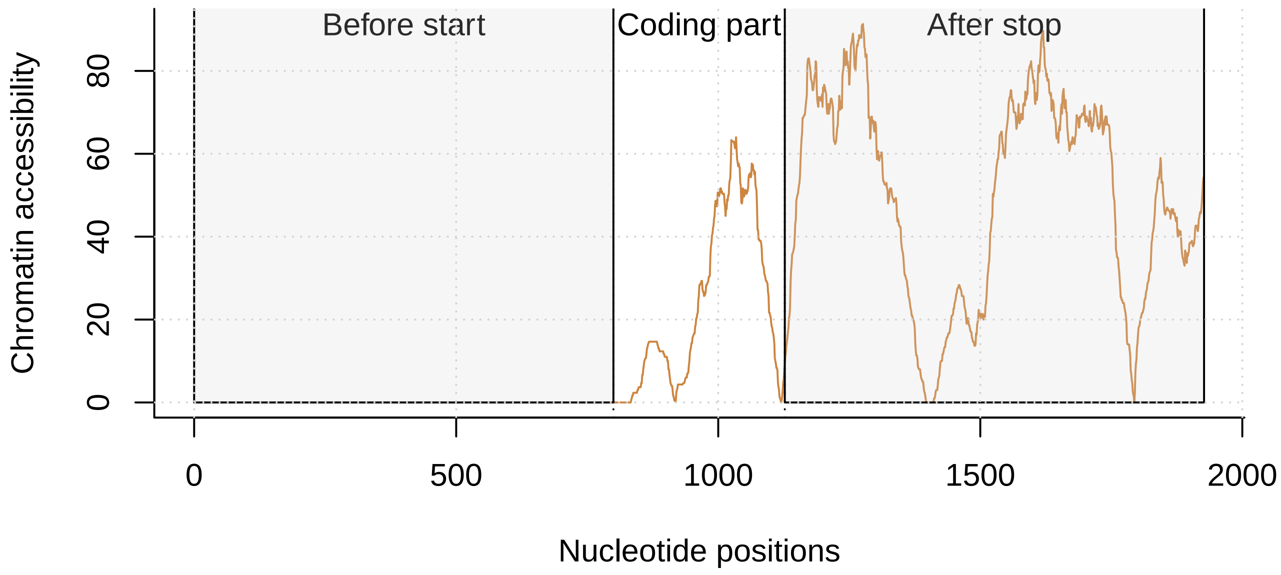
3.2 Intervals of zeroes in epigenomic signals

Genes whose mean epigenomic signal is constantly zero before the start codon or after the stop codon, or show at least one sequence of more than hundred consecutive zeroes within the coding part will be considered as problematic in the following.

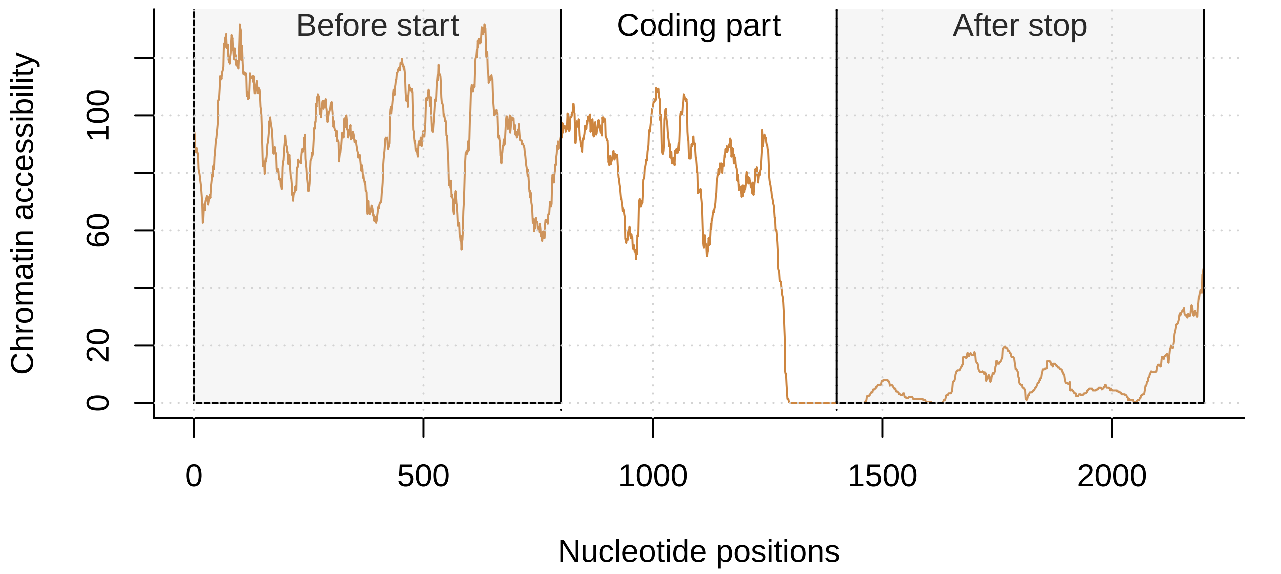
First, transcriptomic and epigenomic signals are averaged over replicates, the mean signals being stored as lists in objects `l_mean_express` and `l_mean_epigeno` respectively.

Genes whose mean epigenomic signal is constantly zero before the start codon or after the stop codon, or show at least one sequence of more than hundred consecutive zeroes within the coding part are identified:

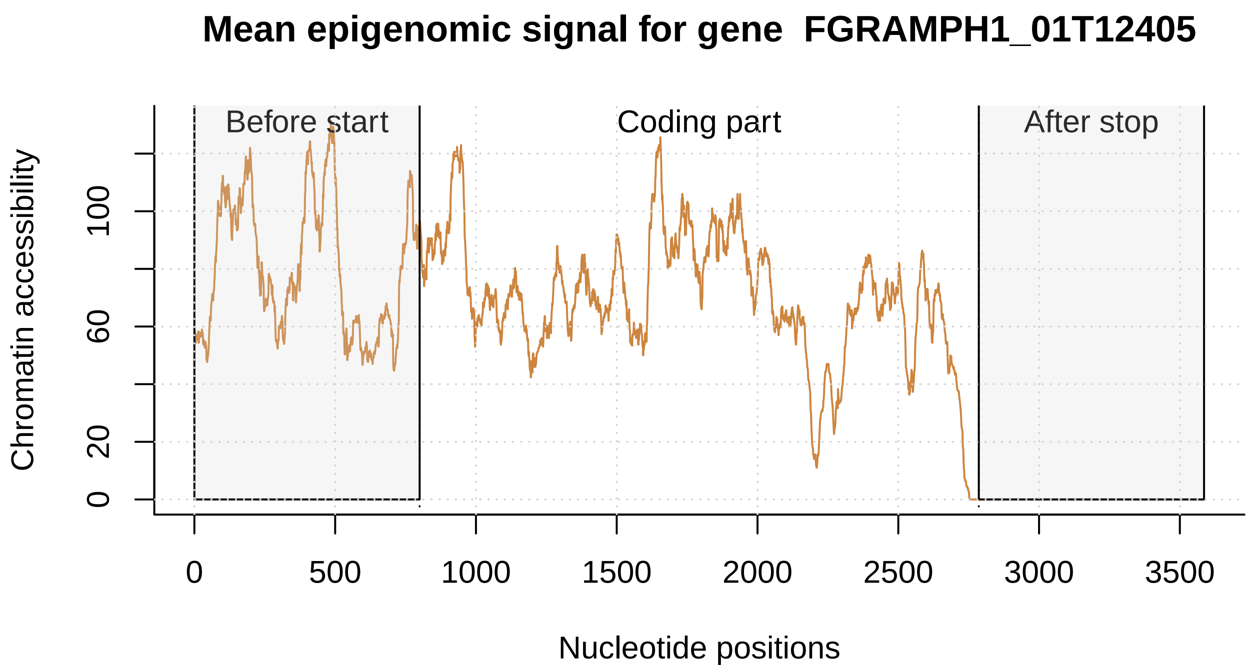
The following plot shows an example of a gene whose epigenomic signal is constantly zero before the stop codon:


Mean epigenomic signal for gene FGRAMPH1_01T08185

Similarly, the plot below shows an example of a gene whose epigenomic signal has one or more long sequence of consecutive zeroes within the coding part:


Mean epigenomic signal for gene FGRAMPH1_01T00005

Finally, the plot below shows an example of a gene whose epigenomic signal is constantly zero after the stop codon:



The number of genes with anomalies such as illustrated by the three plots above is 141. Those genes are removed from the subsequent analysis.

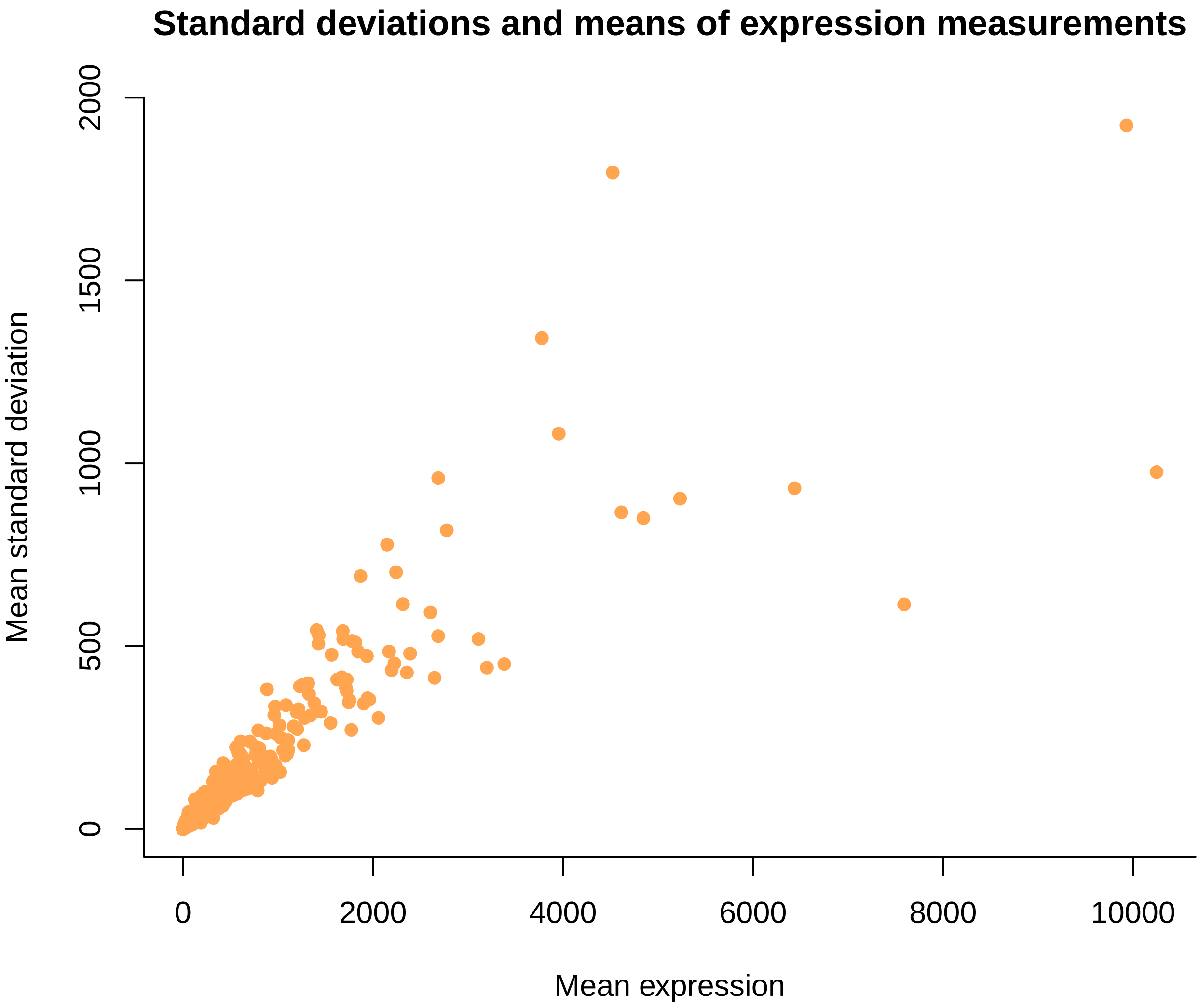
3.3 Signals with large variability across replicates

There is no straightforward correspondence between replicates of transcriptomic and epigenomic data. Since the goal is to study the association between both types of signals, replicates will be used in a preprocessing step to identify genes for which the signals would not be reproducible.

3.3.1 Expression signals

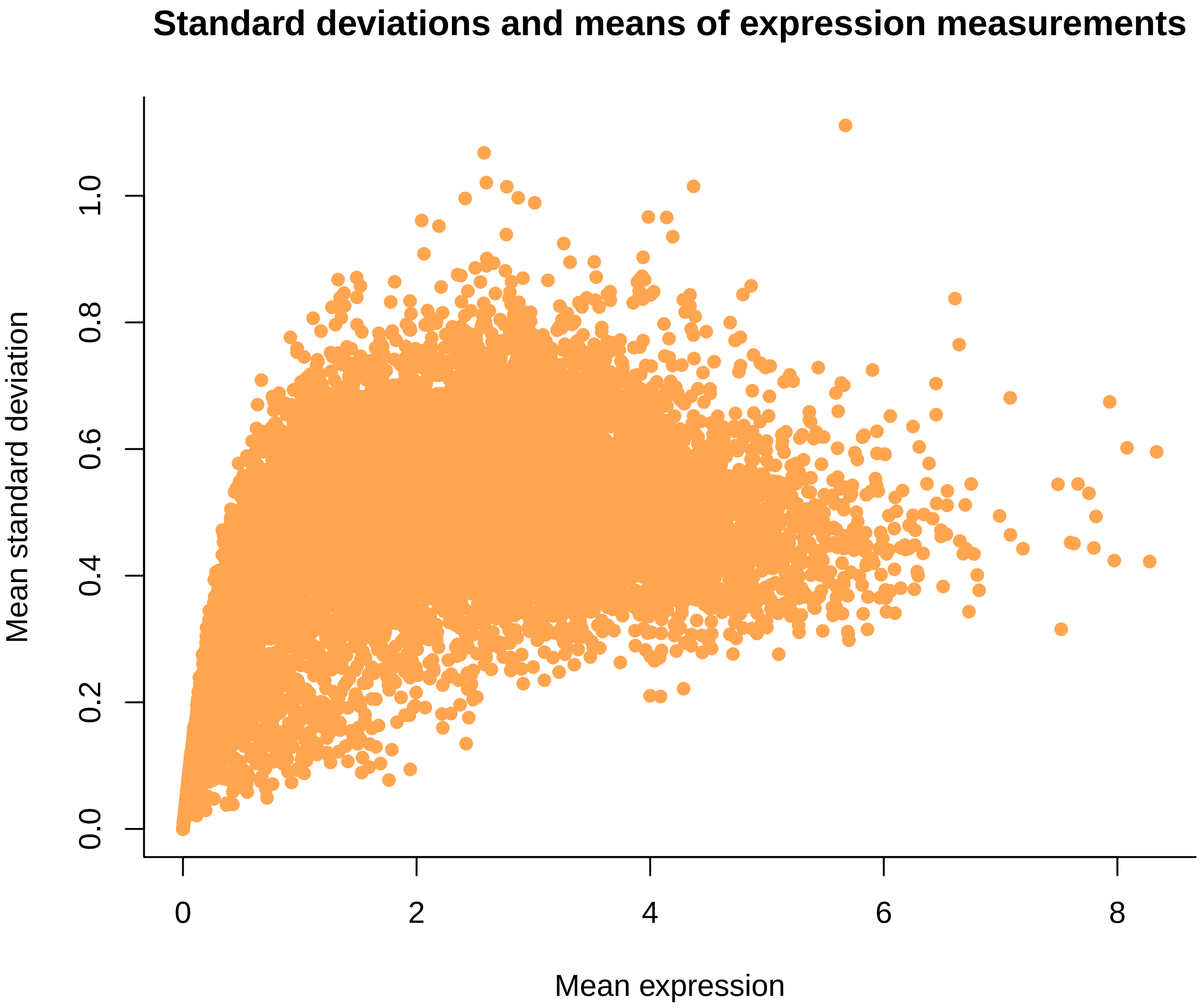
First, at each nucleotide position, the standard deviation over the three replicates of the expression measurements is derived for each gene.

The expression measurements being read counts, standard deviation highly depends on the mean expression. In order to illustrate this point, both the curves of standard deviations and mean expressions are summarized by their mean value over the nucleotide positions. The following plot shows the relationship between those average standard deviations and mean expressions:



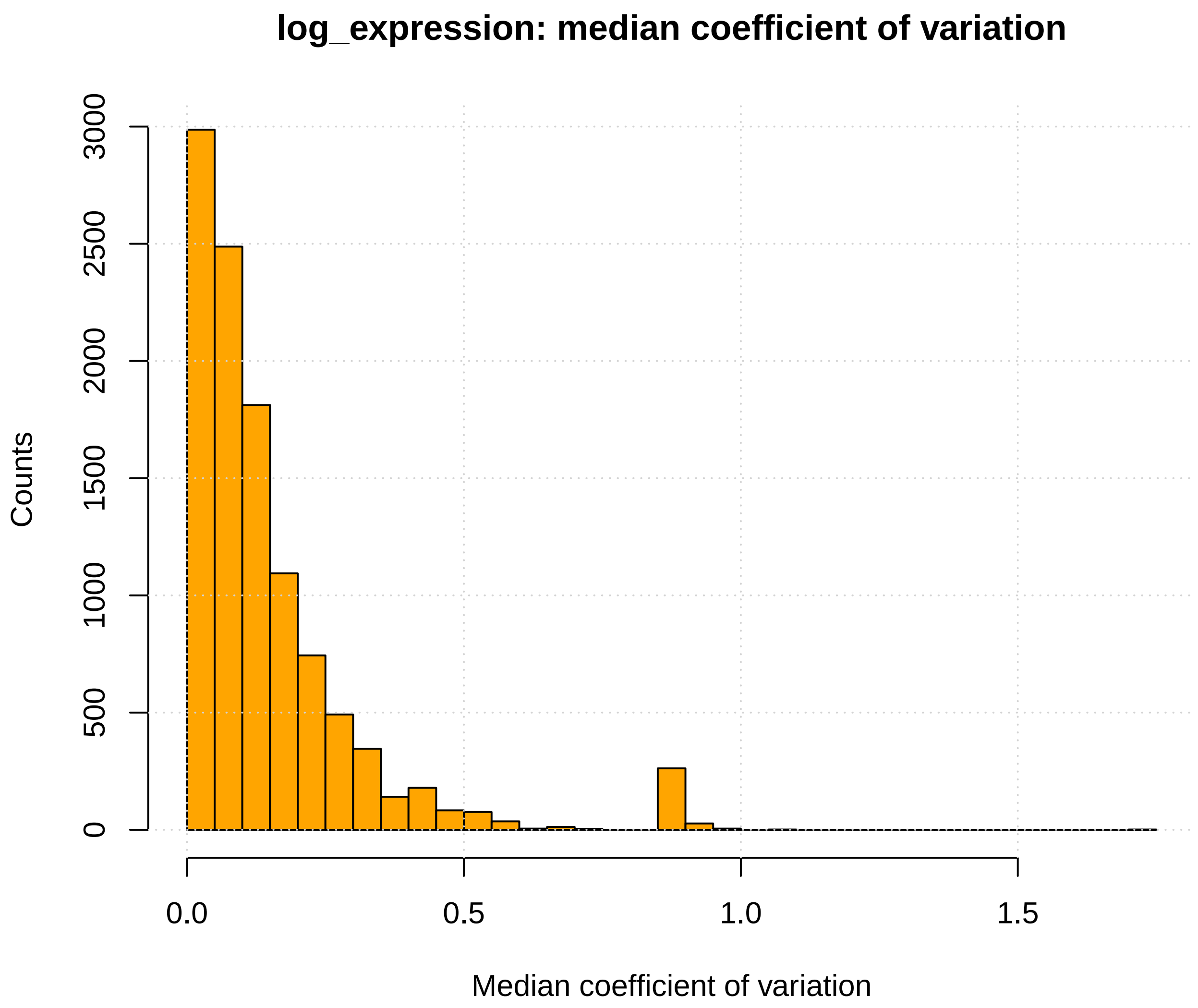
Therefore, a decision rule for excluding genes with poorly reproducible signals that would be based on the standard deviation solely would favor the exclusion of genes with large mean expression.

This issue is addressed hereafter by a $\log_2$-transformation of read counts. The following plot of the relationship between average standard deviations and mean expressions after $\log_2$-transformation confirms that variability of expression is less straightforwardly dependent of the mean level, except for the lowest mean expressions:



The ratio of the standard deviation to the mean, namely the coefficient of variation, is used hereafter, to evaluate the variability of signals over the replicates and identify outliers. For each nucleotide position and each gene, a coefficient of variation is derived to measure the variability among replicates and, for each gene, the global variability of the whole expression signal is evaluated by the median coefficient of variation.

The distribution of the median coefficient of variation over the genome is shown below:

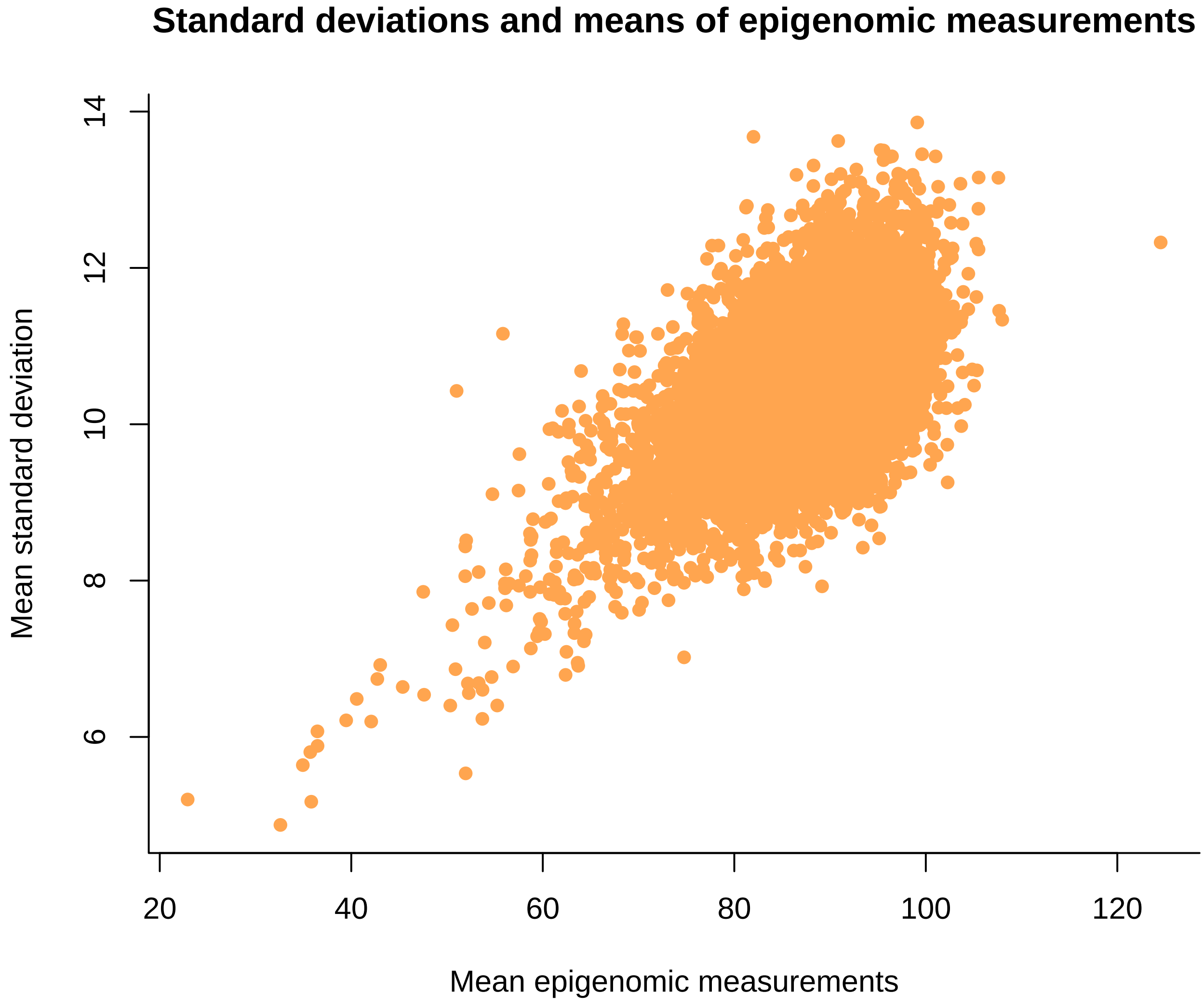


The plot above reveals some few genes for which the median coefficient of variations is abnormally large, close to one but smaller. The number of those genes, with median coefficient of variation larger than 0.8, is 296. However, since all coefficients of variation are smaller than 1, it is decided not to remove any genes based on the former inspection of the variability of expression signals across replicates.

3.3.2 Epigenomic signals

As above for expression signals, at each nucleotide position, the standard deviation over the three replicates of the epigenomic measurements is derived for each gene.

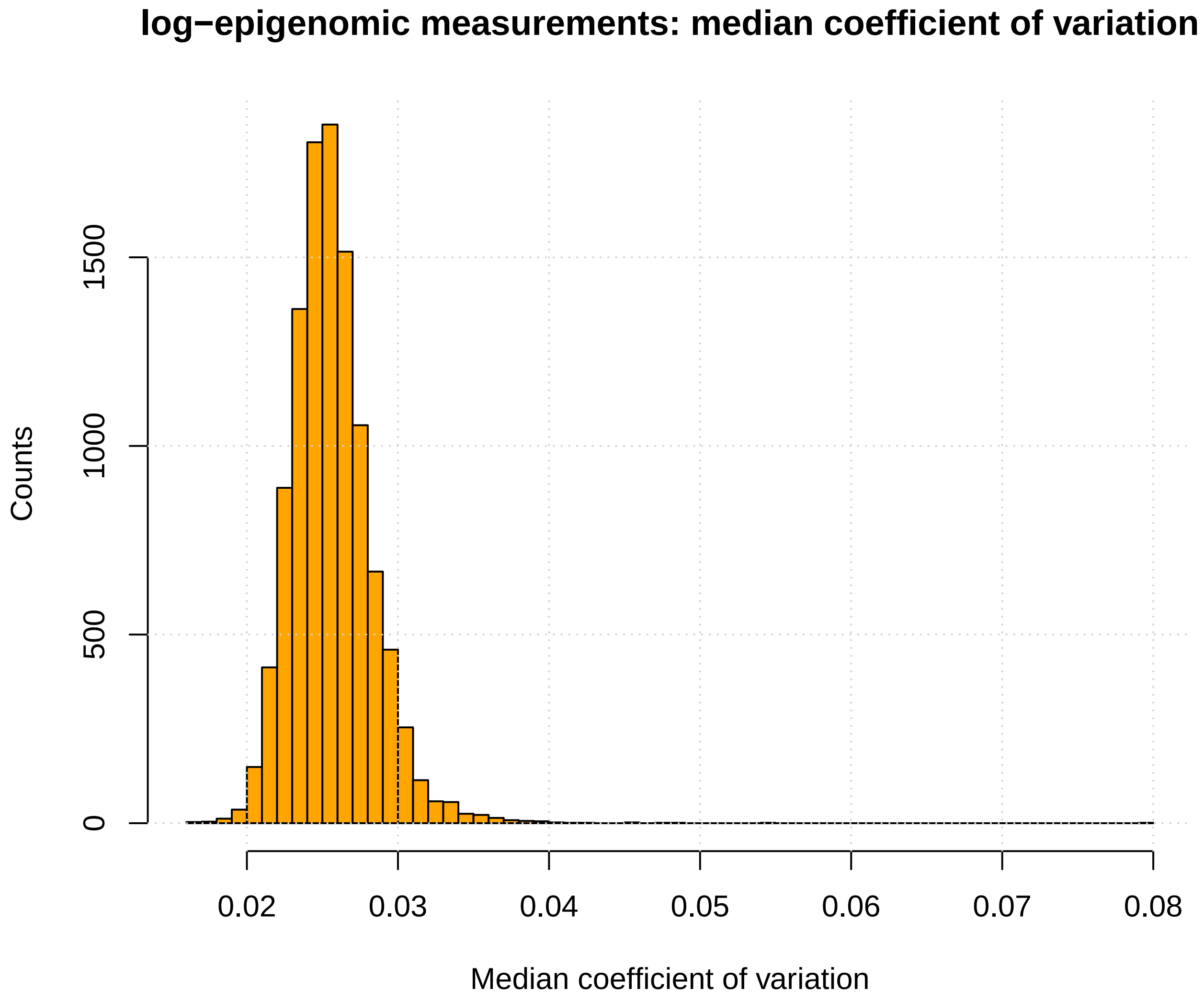
For the epigenomic measurements also, standard deviation highly depends on the mean values. After summarizing the curves of standard deviations and mean expressions by their mean value over the nucleotide positions, the following plot shows the relationship between those average standard deviations and mean expressions:



As above for expression signals, epigenomic signals are $\log_2$-transformed.

For each nucleotide position and each gene, a coefficient of variation is derived to measure the variability among replicates and, for each gene, the global variability of the whole epigenomic signal is evaluated by the median coefficient of variation.

The distribution of the median coefficient of variation over the genome is shown below:



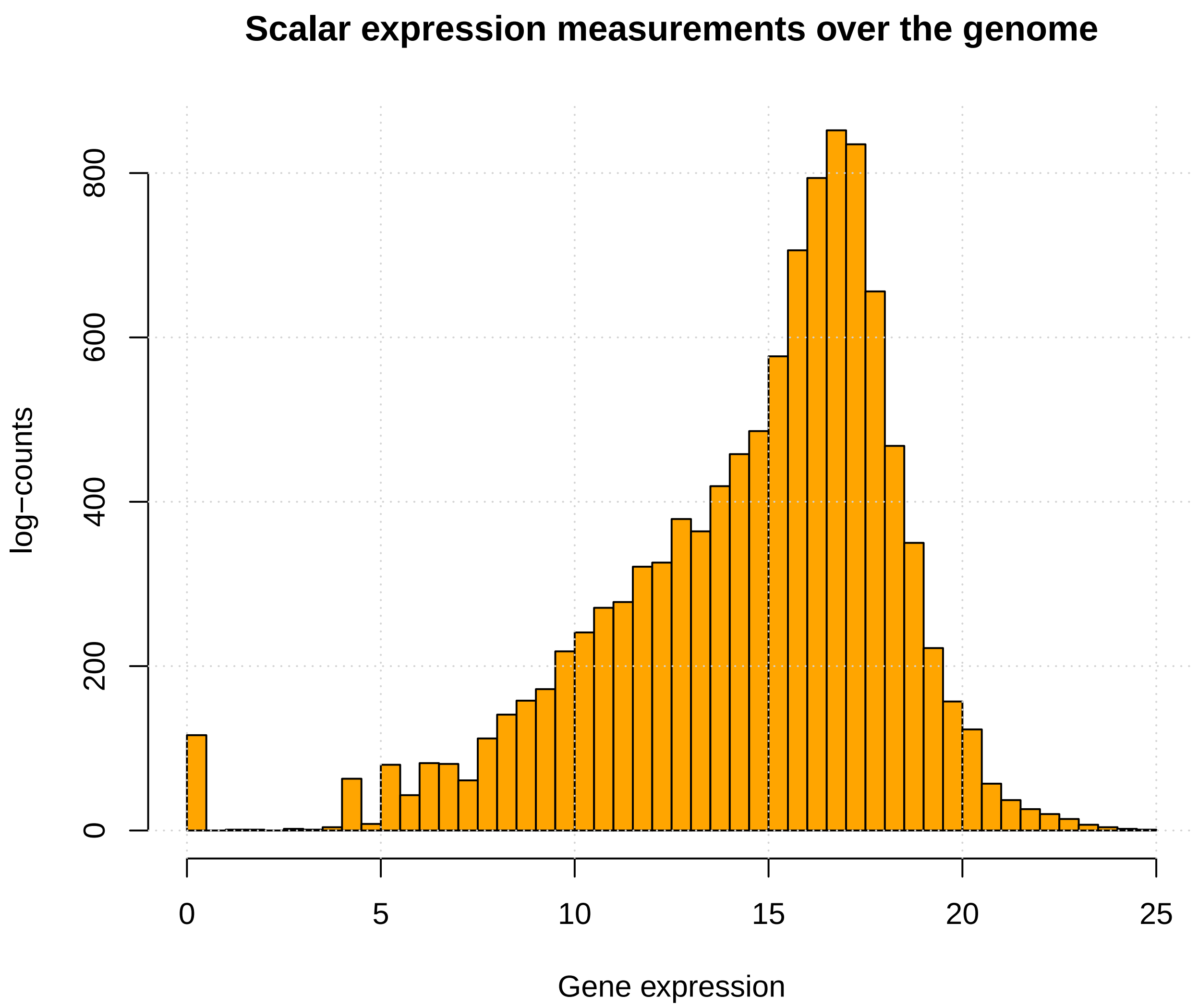
It is deduced from the plot above that the epigenomic signals are obviously much more reproducible than the expression signals, with most median coefficients of variation smaller than 0.04. As for expression signals, all genes are considered as having satisfactorily reproducible epigenomic signal to be kept within the subsequent analysis.

4 Genome-wide coding of signals

Modeling genome-wide association between epigenomic ans transcriptomic signals requires that both type of signals are coded by variables that are identically defined for each gene.

4.1 Summarizing signals over replicates

First, for each gene, the epigenomic and transcriptomic signals are obtained by taking the mean over the replicates and applying a $\log_2$ transformation. Consistently with the usual way of measuring the expression of a gene, the $\log_2$ transformed sum over the nucleotide positions of the mean transcriptomic signals is also calculated. A histogram of those scalar expression values is shown below:



4.2 Signal alignment

Within the two sequences of nucleotides before the start codon and after the stop codon, both epigenomic and transcriptomic signals are measured for each gene on the same grid of nucleotide positions. However, since the genes have different lengths, the signals are not aligned on the same grid in-between the start and stop codons. Since the goal is to infer on an asso-

Table 4: Length of the first two genes in the data

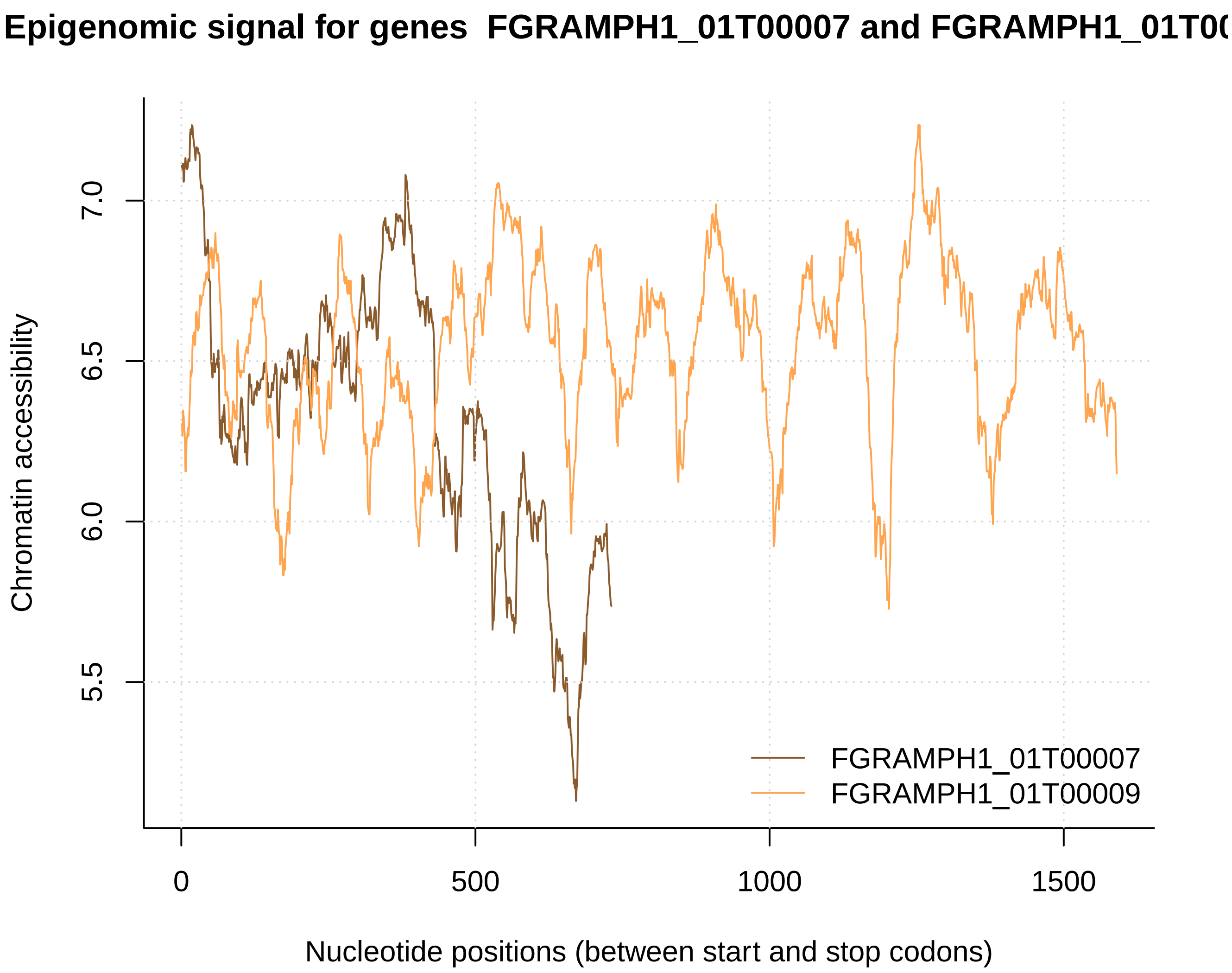
	Length
FGRAMPH1_01T00007	731
FGRAMPH1_01T00009	1590

ciation model between epigenome and transcriptome for all genes, signals need to be coded identically for each gene.

This could be done by using standard non-functional summary statistics such as mean levels over the nucleotide positions, standard deviations, number of positive peaks, and so on. The search for such summary statistics that do not explicitly account for the spatial support of signals over the nucleotide positions exposes to making choices that are not guided by a prior biological knowledge and therefore could lead to irrelevant or incomplete summaries.

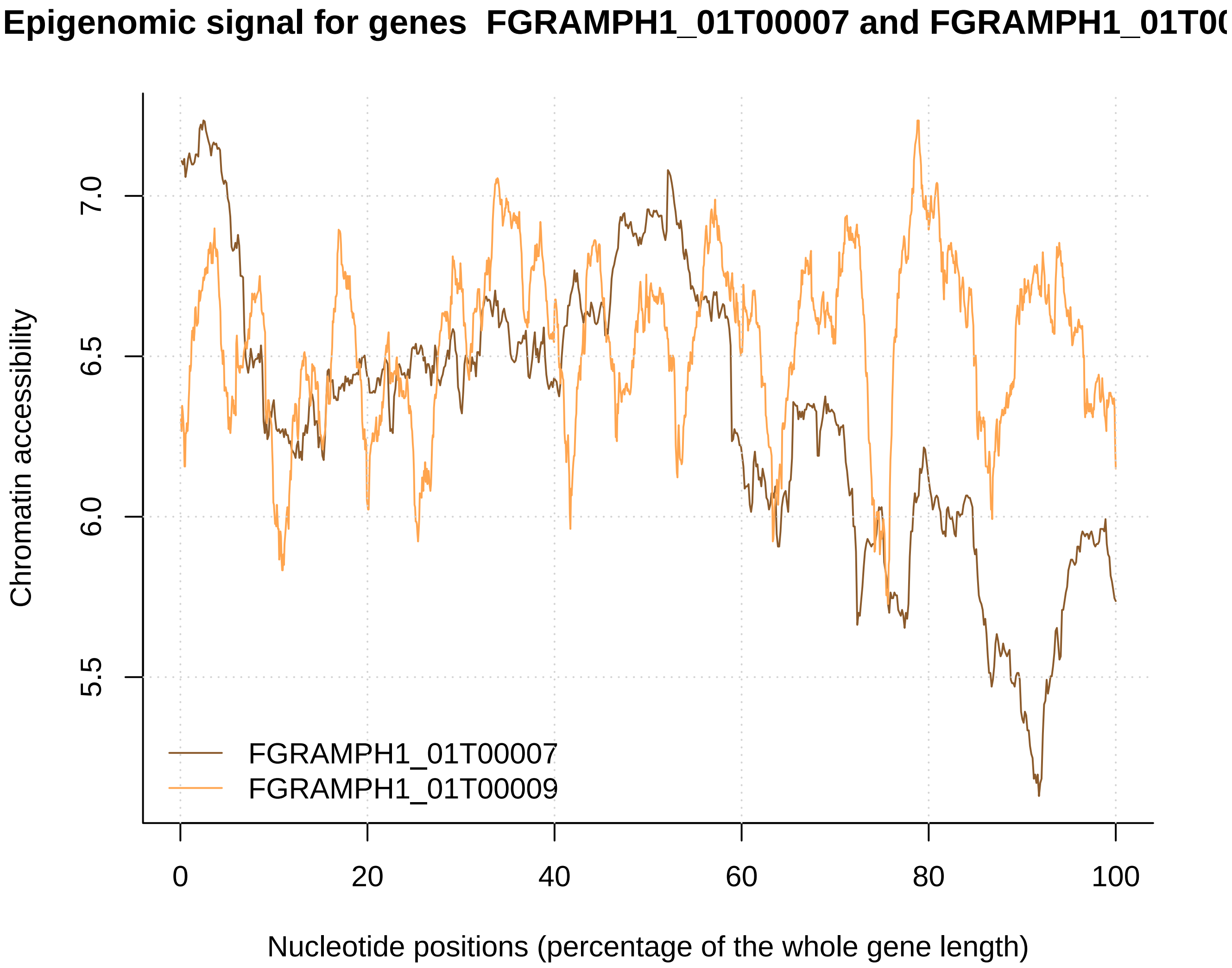
Our approach will aim to avoid such choices and keep as most as possible the information contained in the patterns of variations over the sequence of nucleotide. Let us consider the first two genes in the data for illustration. Their name and length is given below:

The following plot shows the epigenomic signals of those two genes between the start and stop



codons:

Since those two genes have different lengths, the two signals are not aligned in the sense that a given nucleotide position do not correspond to the same relative distance from the start codon, expressed as a percentage of the whole gene length. The plot below shows the epigenomic signals of the two genes between the start and stop codons, when considering the nucleotide positions as their relative distance with respect to the start codon:



Although the two epigenomic signals are now aligned on the same interval of relative nucleotide positions, the pointwise measurements of those signals do not yet correspond to the same relative positions, as illustrated by the table below, giving the first five relative positions for each gene:

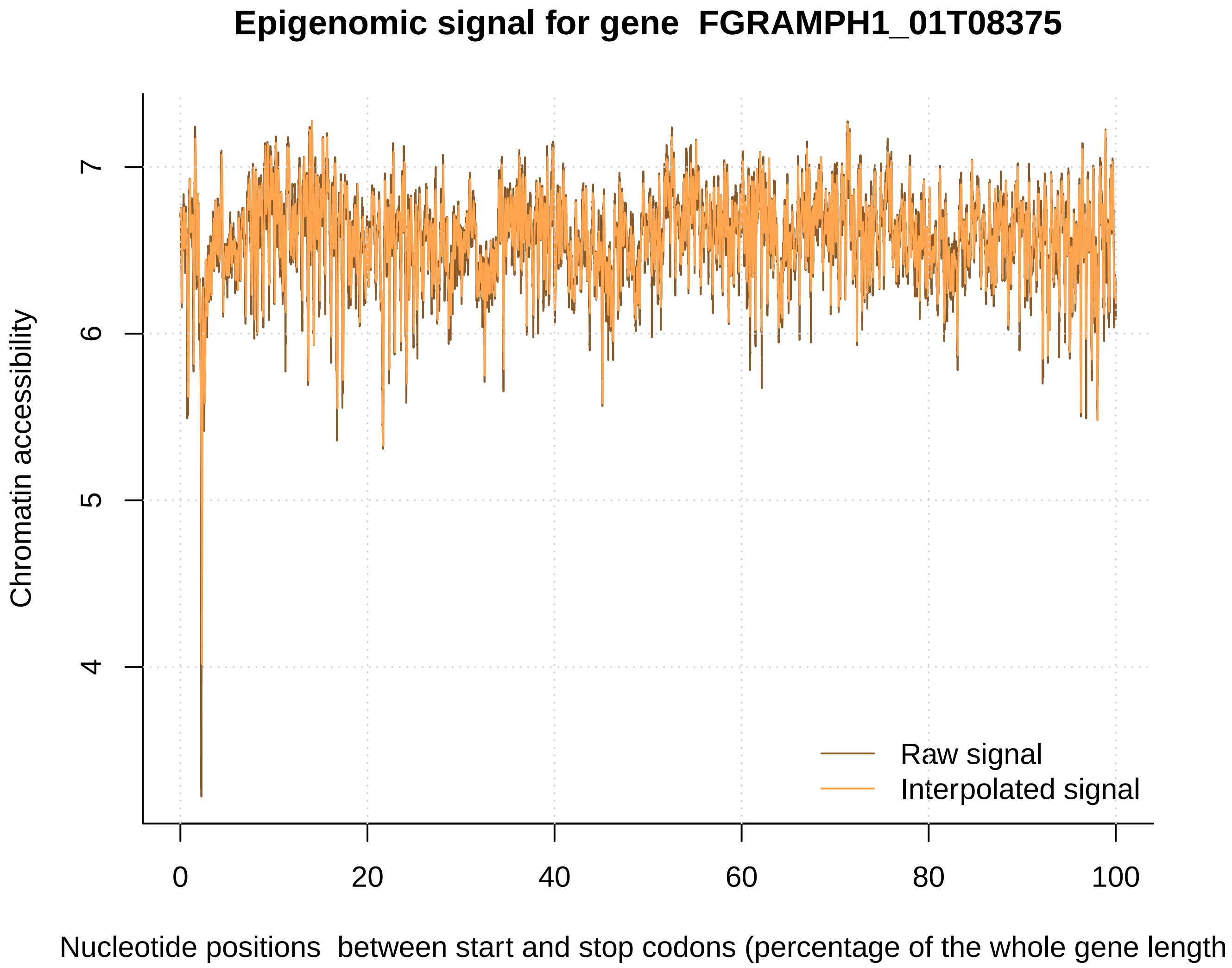
Table 5: First five relative nucleotide positions

	1	2	3	4	5
FGRAMPH1_01T00007	0.14	0.27	0.41	0.55	0.68
FGRAMPH1_01T00009	0.06	0.13	0.19	0.25	0.31

Therefore, a common grid of relative nucleotide positions is now chosen, and the measurements of the signals at each point on this grid are obtained by linear interpolations between the two closest measurements. The choice of this common grid is guided by a consensus between a small number of points, leading to smoothing potentially interesting local variations in long signals and a large number of points, which induces both a large oversampling for small genes and a large number of variables in the final dataset.

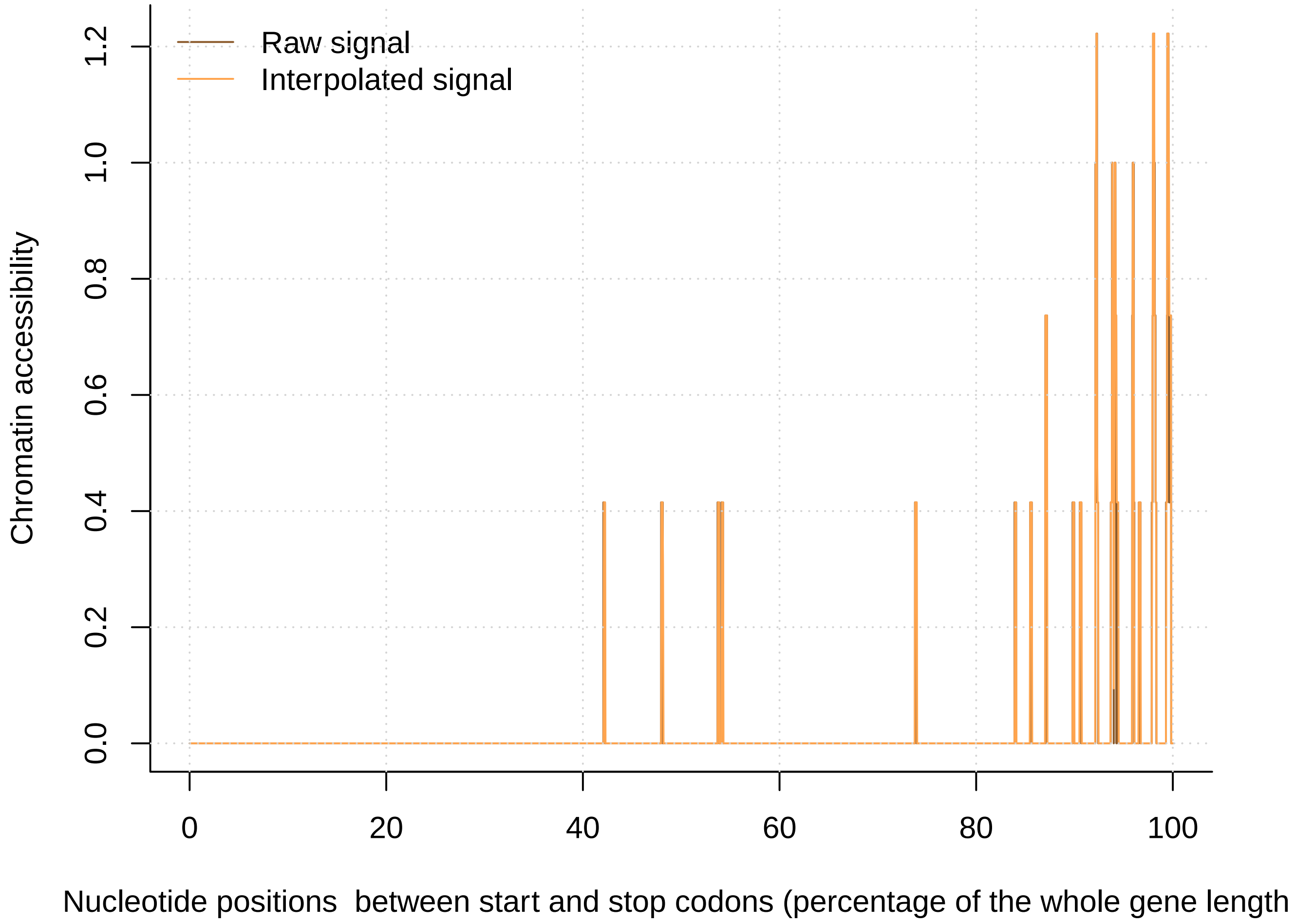
Since two thirds of the genes have less than 2000 measurements, it is decided conservatively to consider a grid of 2000 regularly spaced relative positions, which means that the distance between two points in this grid is 0.05%.

The largest interpolation error is supposed to be made for the longest gene, with length 2.9066×10^4. The following plot superimposes the raw and interpolated curves between the start and stop codons for this gene, which confirms the close approximation of the raw signal with the present choice of a grid of 2000 relative nucleotide positions.



A similar plot is produced for the same gene with the raw and interpolated transcriptomic signals:

Epigenomic signal for gene FGRAMPH1_01T08375


Raw signal
Interpolated signal
Chromatin accessibility
Nucleotide positions between start and stop codons (percentage of the whole gene length