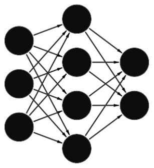
PREDICTION OF PATHOGENIC SNV

Prof. Giorgio Valentini
6 CFU

Luca Cappelletti

Lecture Notes
Year 2017/2018



IT Master Degree
Universiy of Milan
Italy
26 giugno 2018

Indice

Parte I

Dataset

1

Data points

First we begin looking at the dataset, the distributions of the given metrics and the statistical analysis of these data points.

1.1 Retrieving the dataset

The dataset can be downloaded from https://homes.di.unimi.it/valentini/ProgettoBioinformatica1718/data/.

1.2 Data points

In the dataset there are 981389 data points, each one comprised of 26 metrics. The first 356 are pathogenic and all the others are negative.

2

Metrics

2.1 CpGobsExp

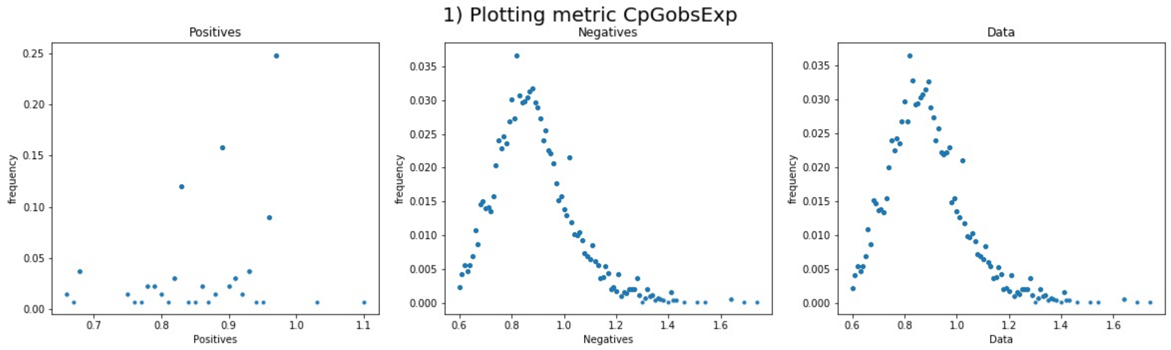


Figura 2.1: Sampling distribution of metric CpGobsExp

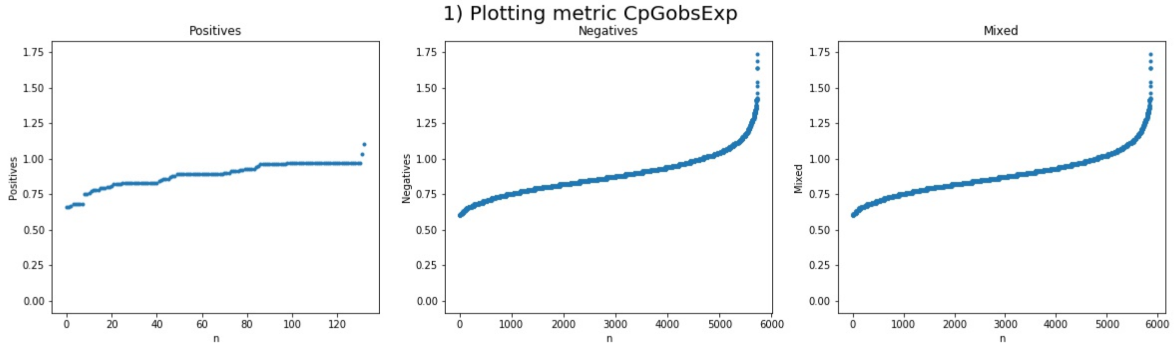


Figura 2.2: Values of metric CpGobsExp

2.2 CpGperCpG

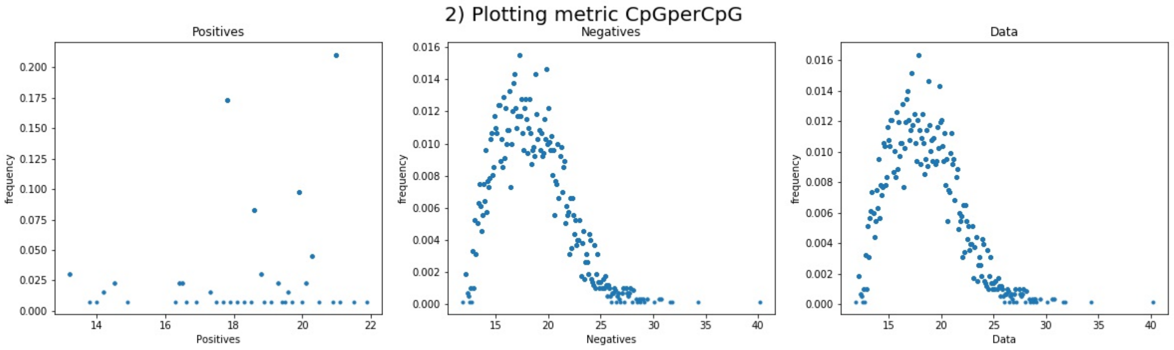


Figura 2.3: Sampling distribution of metric CpGperCpG

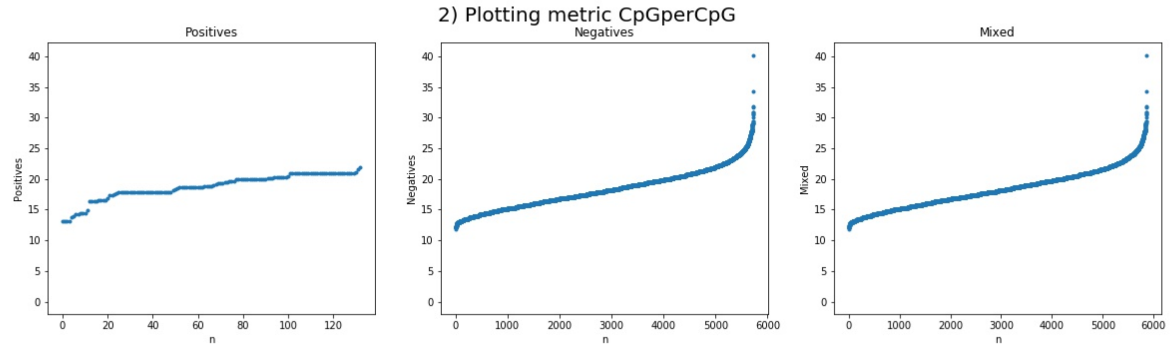


Figura 2.4: Values of metric CpGperCpG

2.3 CpGperGC

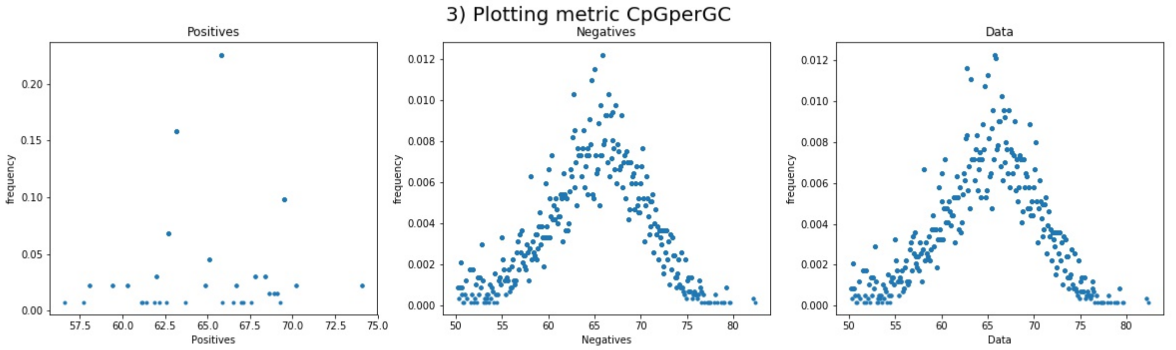


Figura 2.5: Sampling distribution of metric CpGperGC

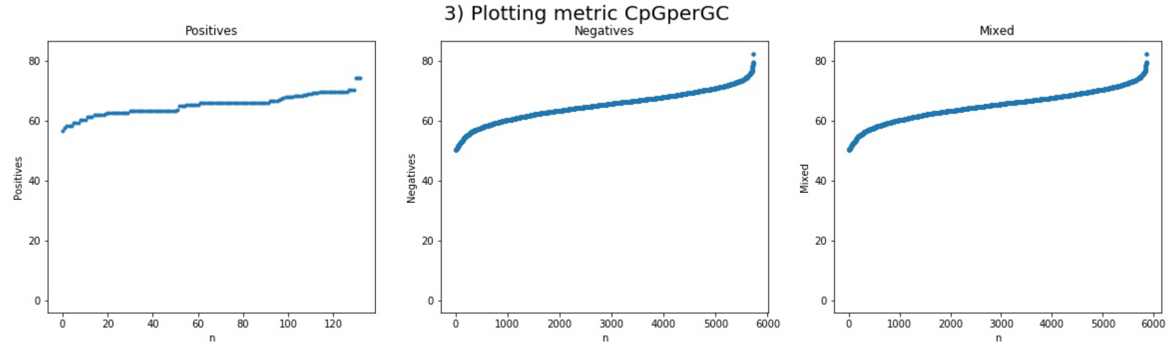


Figura 2.6: Values of metric CpGperGC

2.4 DGVCount

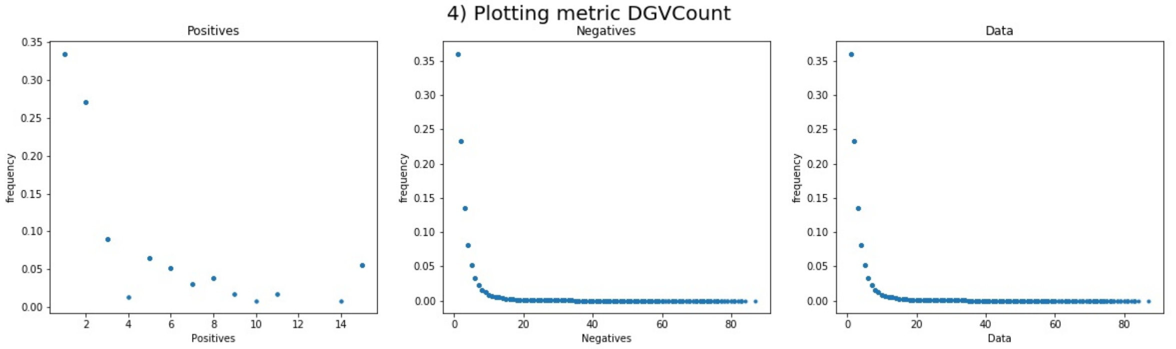


Figura 2.7: Sampling distribution of metric DGVCount

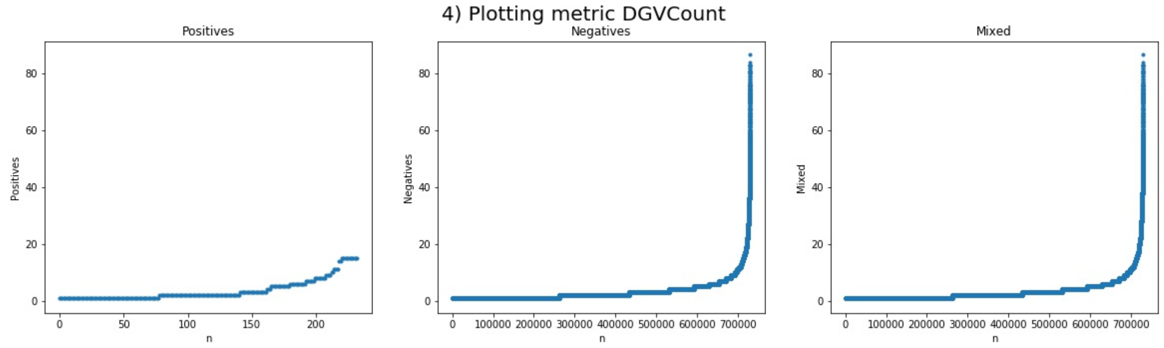


Figura 2.8: Values of metric DGVCount

2.5 DnaseClusteredHyp

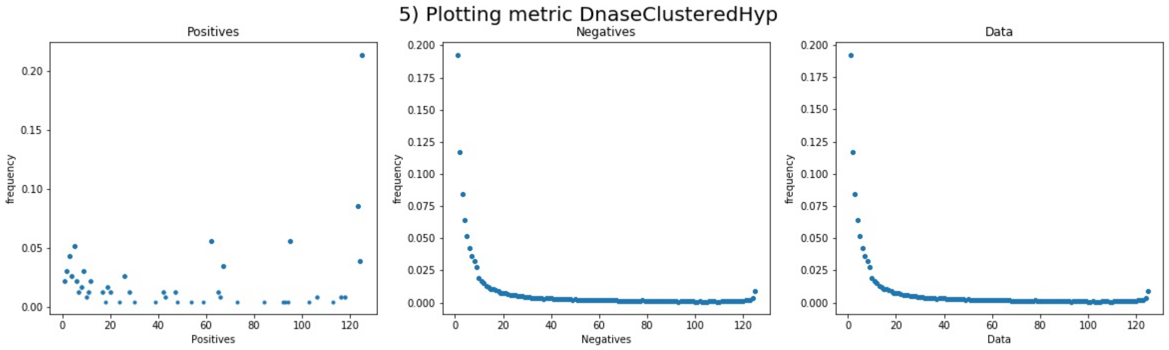


Figura 2.9: Sampling distribution of metric DnaseClusteredHyp

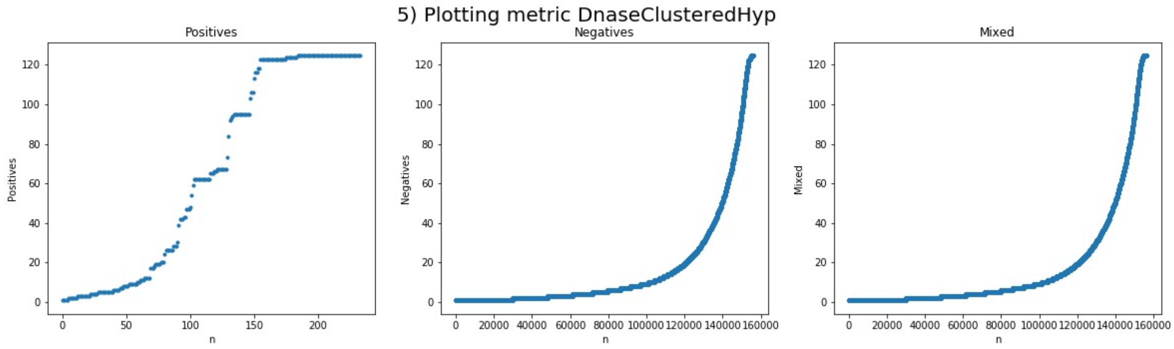


Figura 2.10: Values of metric DnaseClusteredHyp

2.6 DnaseClusteredScore

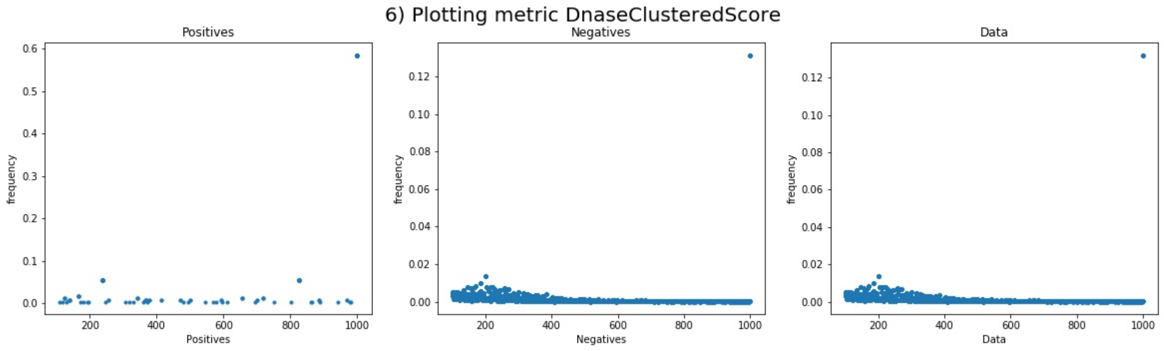


Figura 2.11: Sampling distribution of metric DnaseClusteredScore

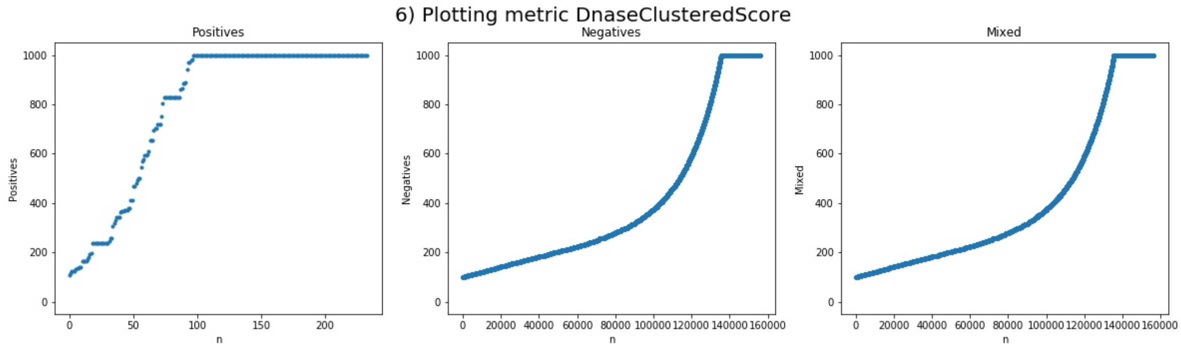


Figura 2.12: Values of metric DnaseClusteredScore

2.7 EncH3K27Ac

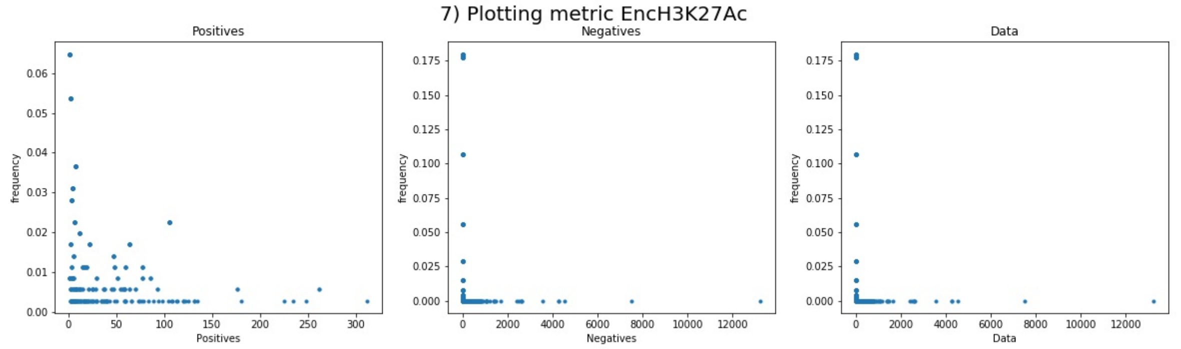


Figura 2.13: Sampling distribution of metric EncH3K27Ac

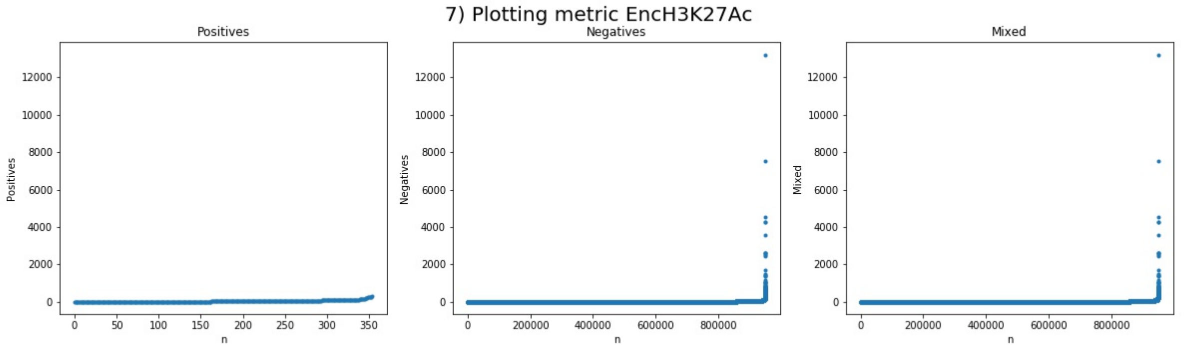


Figura 2.14: Values of metric EncH3K27Ac

2.8 EncH3K4Me1

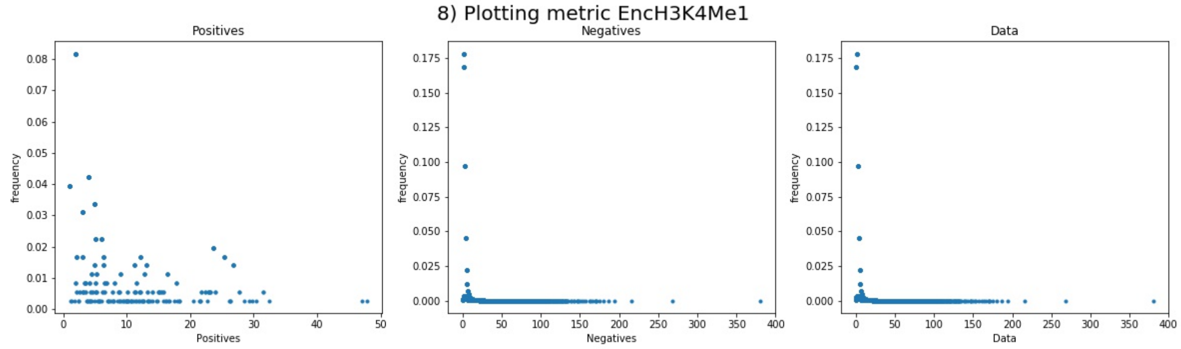


Figura 2.15: Sampling distribution of metric EncH3K4Me1

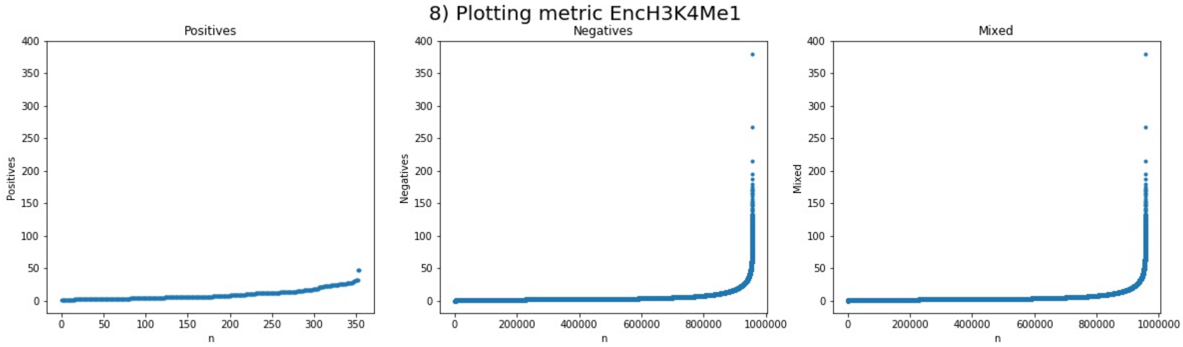


Figura 2.16: Values of metric EncH3K4Me1

2.9 EncH3K4Me3

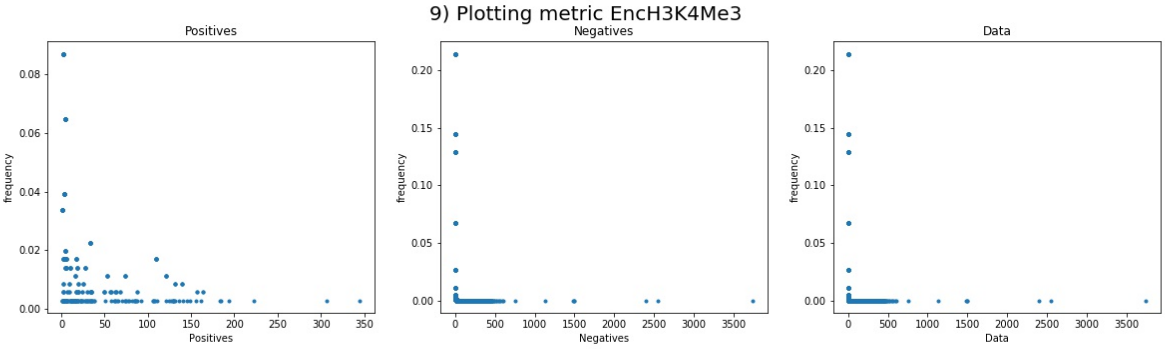


Figura 2.17: Sampling distribution of metric EncH3K4Me3

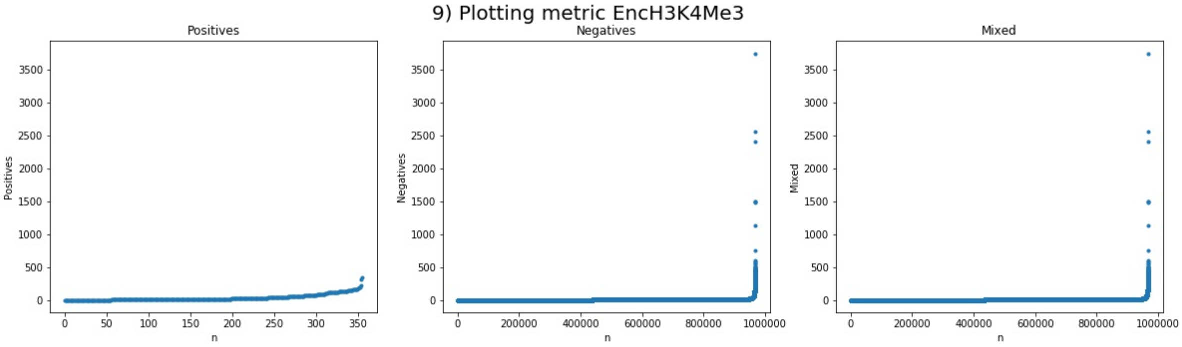


Figura 2.18: Values of metric EncH3K4Me3

2.10 GCContent

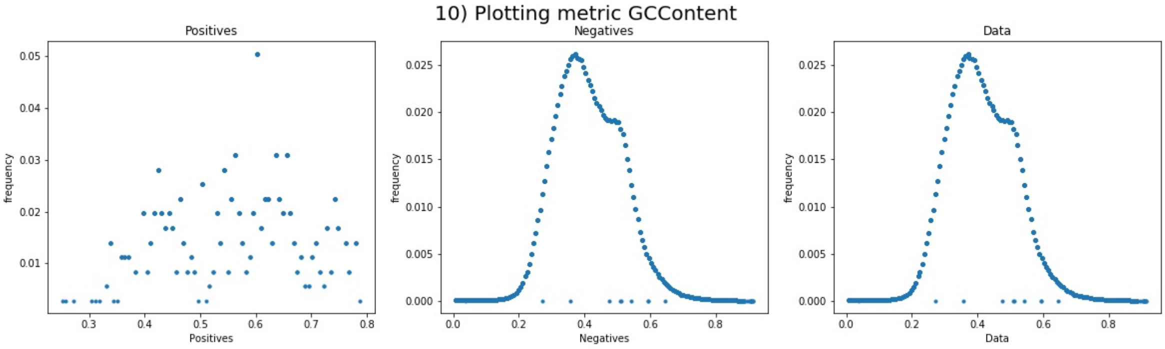


Figura 2.19: Sampling distribution of metric GCContent

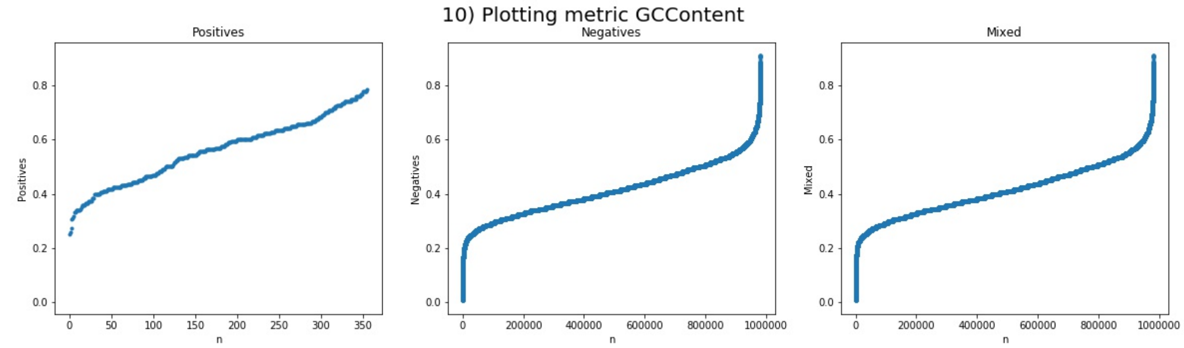


Figura 2.20: Values of metric GCContent

2.11 GerpRS

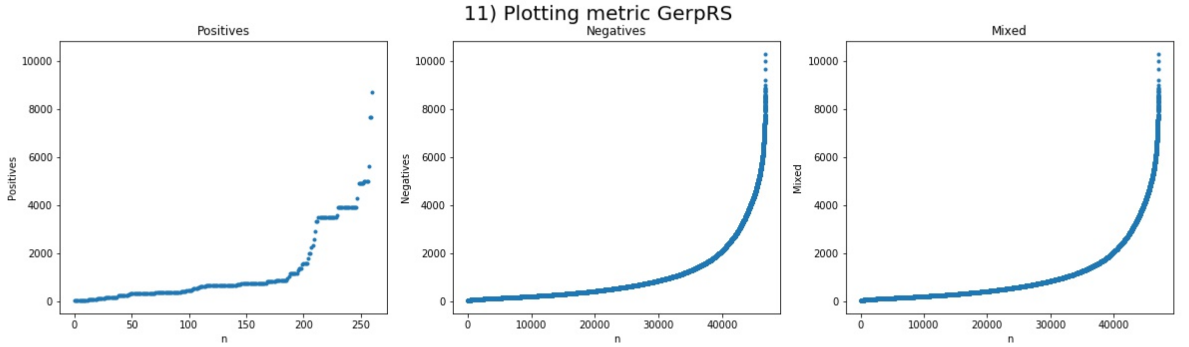


Figura 2.21: Values of metric GerpRS

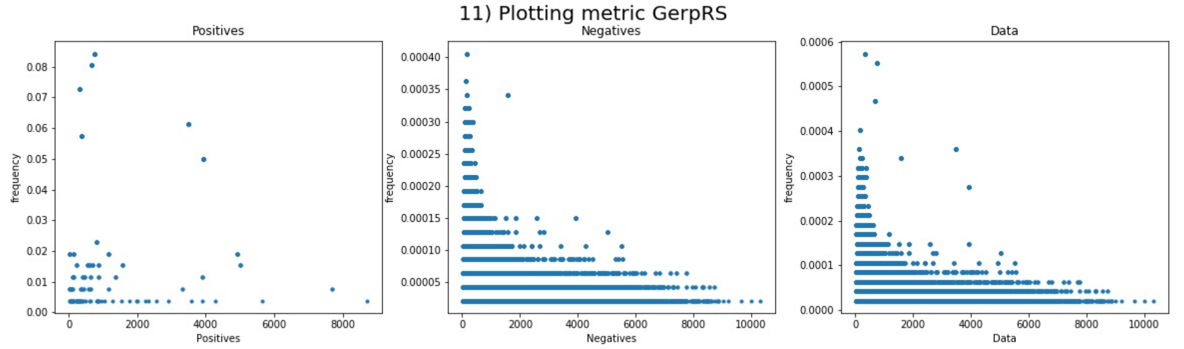


Figura 2.22: Sampling distribution of metric GerpRS

2.12 GerpRSpv

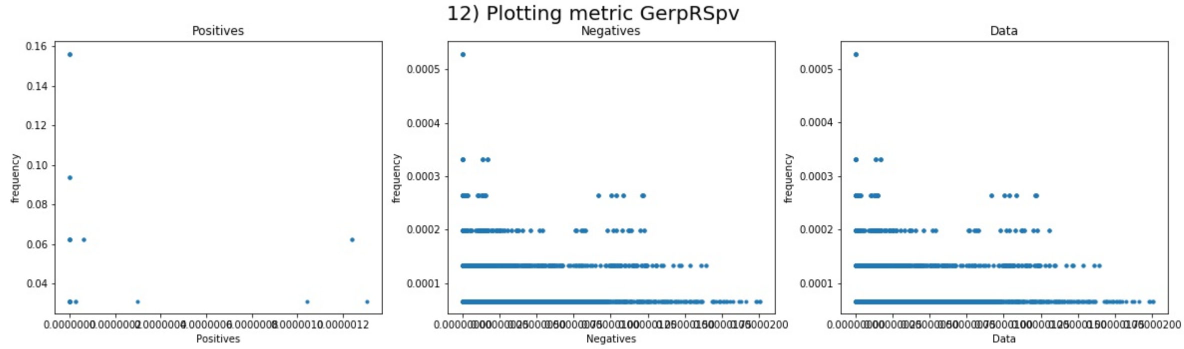


Figura 2.23: Sampling distribution of metric GerpRSpv

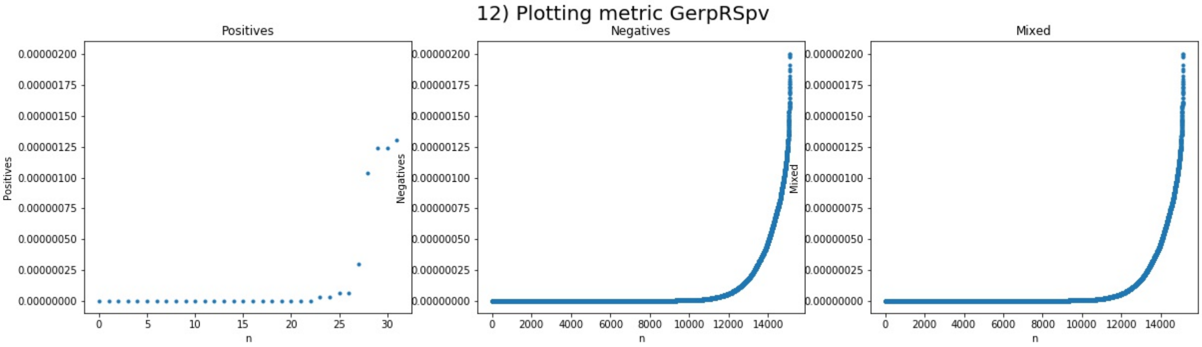


Figura 2.24: Values of metric GerpRSpv

2.13 ISCApath

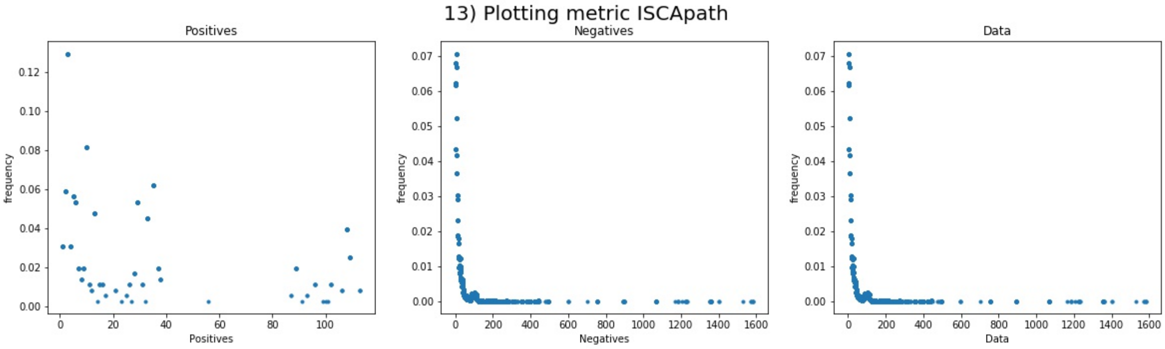


Figura 2.25: Sampling distribution of metric ISCApath

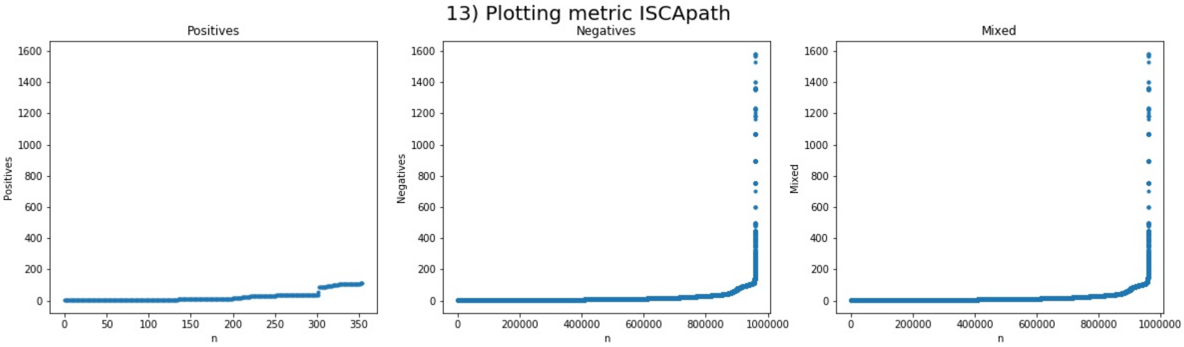


Figura 2.26: Values of metric ISCApath

2.14 commonVar

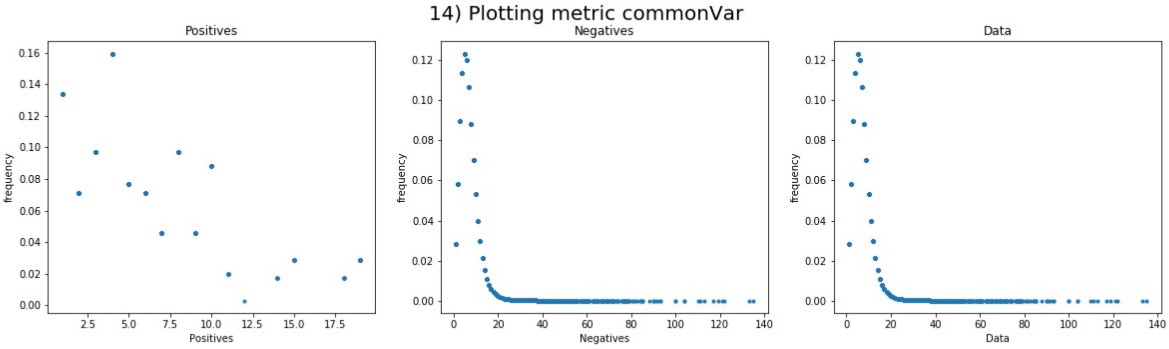


Figura 2.27: Sampling distribution of metric commonVar

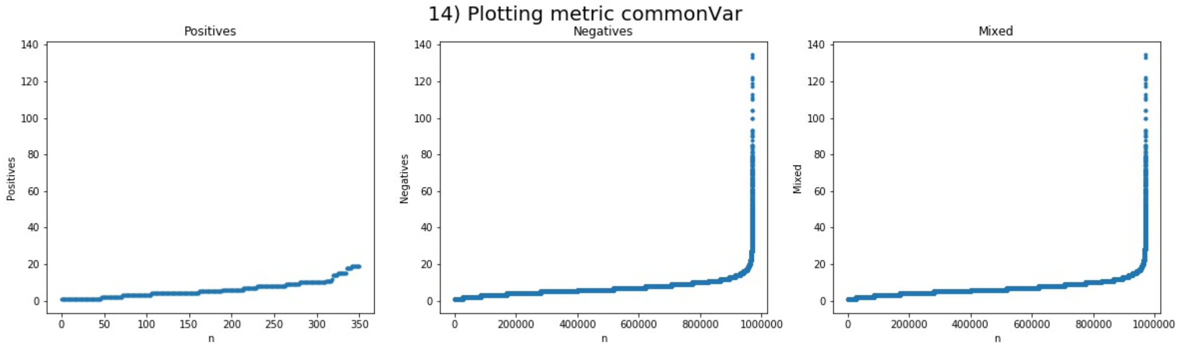


Figura 2.28: Values of metric commonVar

2.15 dbVARCount

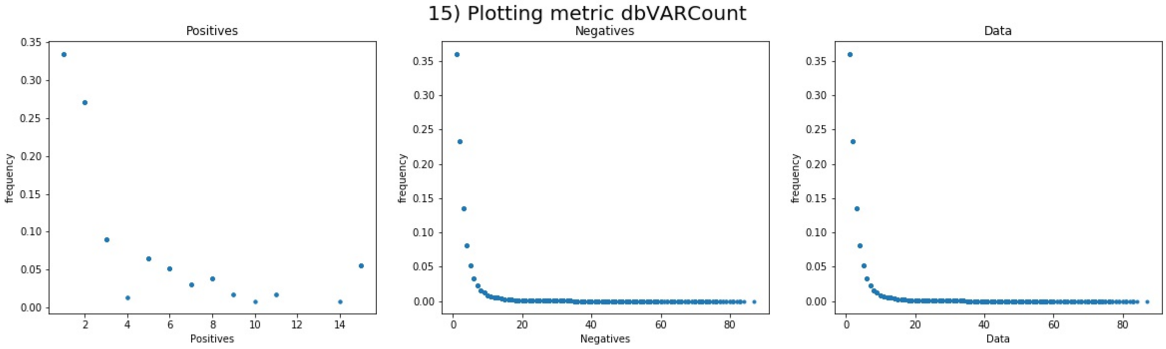


Figura 2.29: Sampling distribution of metric dbVARCount

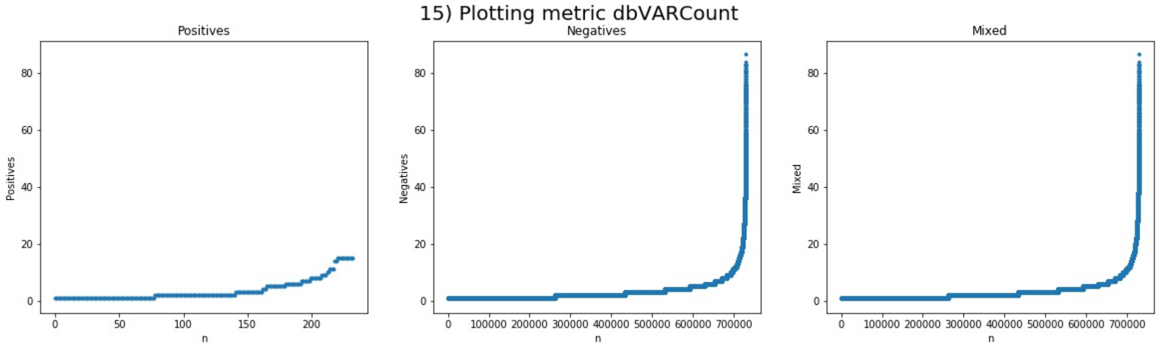


Figura 2.30: Values of metric dbVARCount

2.16 fantom5Perm

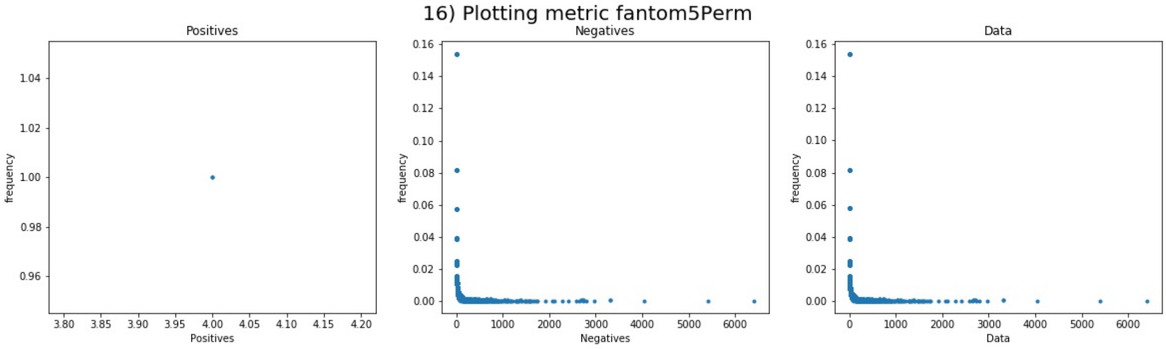


Figura 2.31: Sampling distribution of metric fantom5Perm

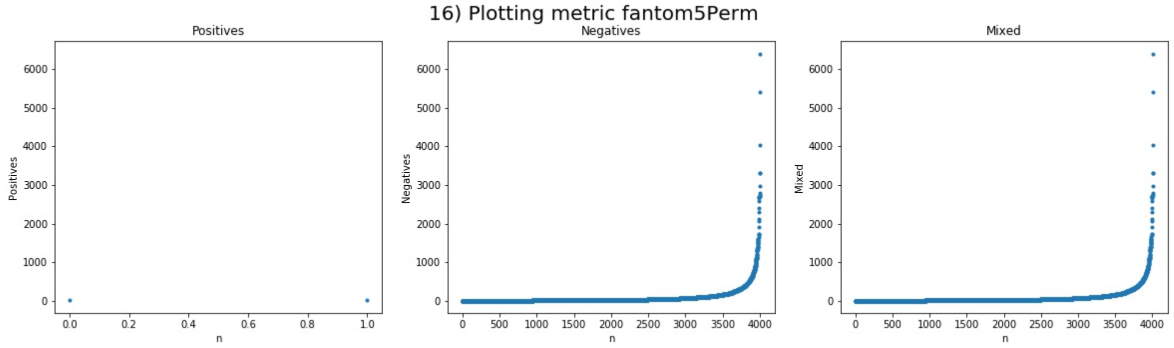


Figura 2.32: Values of metric fantom5Perm

2.17 fantom5Robust

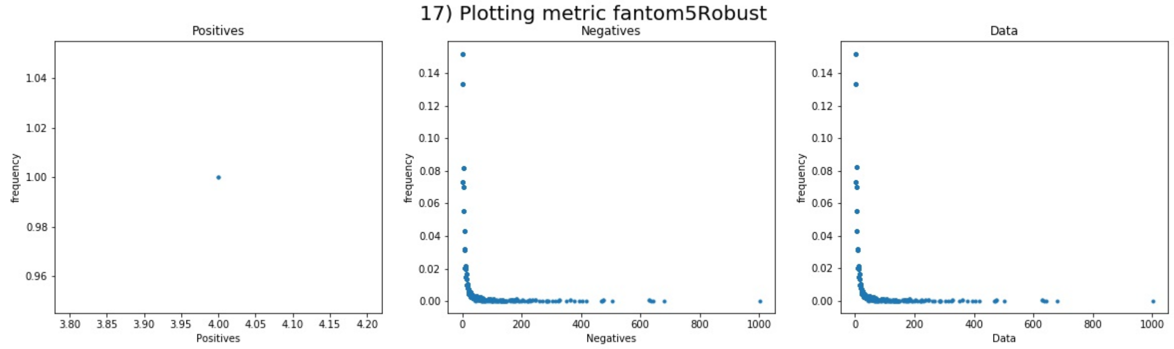


Figura 2.33: Sampling distribution of metric fantom5Robust

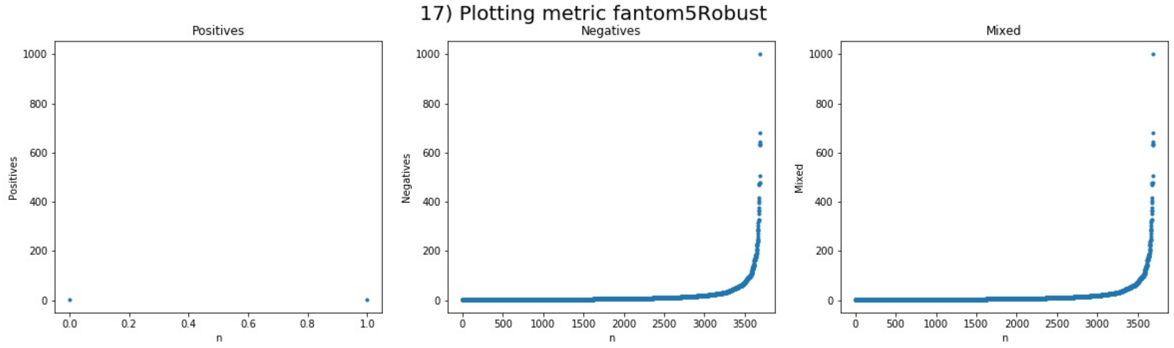


Figura 2.34: Values of metric fantom5Robust

2.18 fracRareCommon

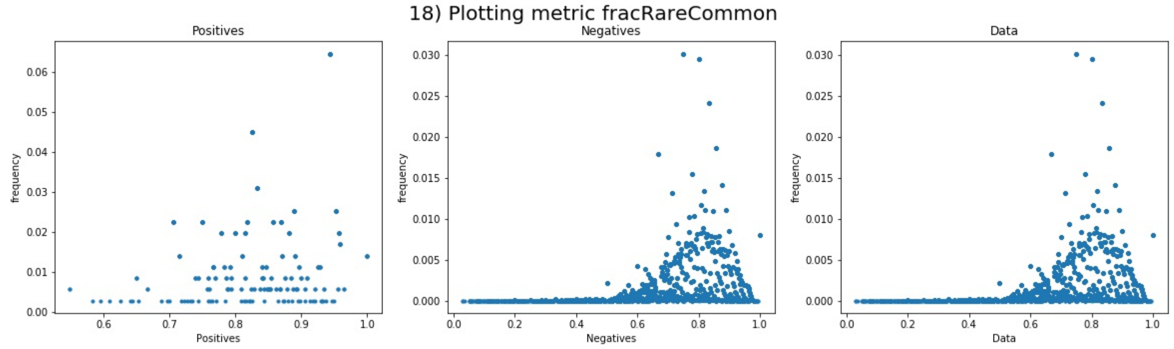


Figura 2.35: Sampling distribution of metric fracRareCommon

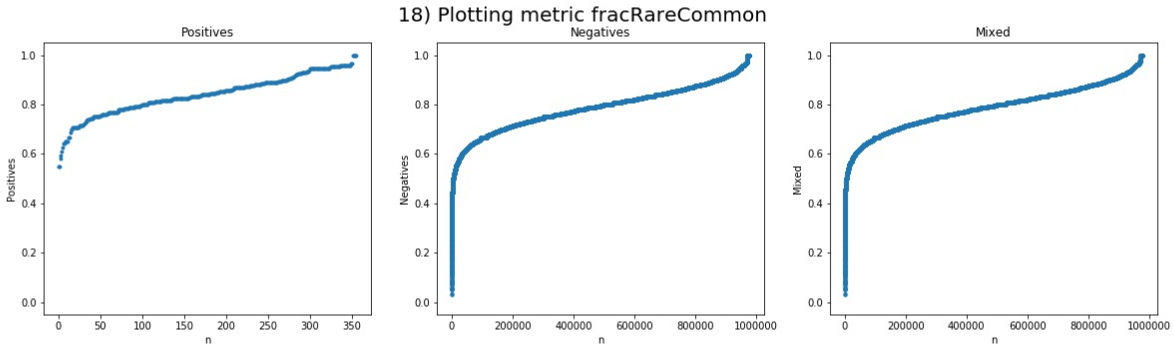


Figura 2.36: Values of metric fracRareCommon

2.19 mamPhastCons46way

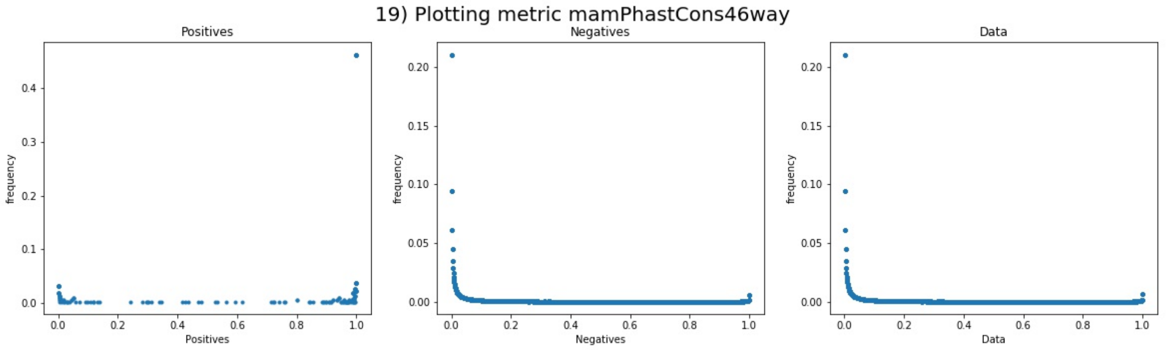


Figura 2.37: Sampling distribution of metric mamPhastCons46way

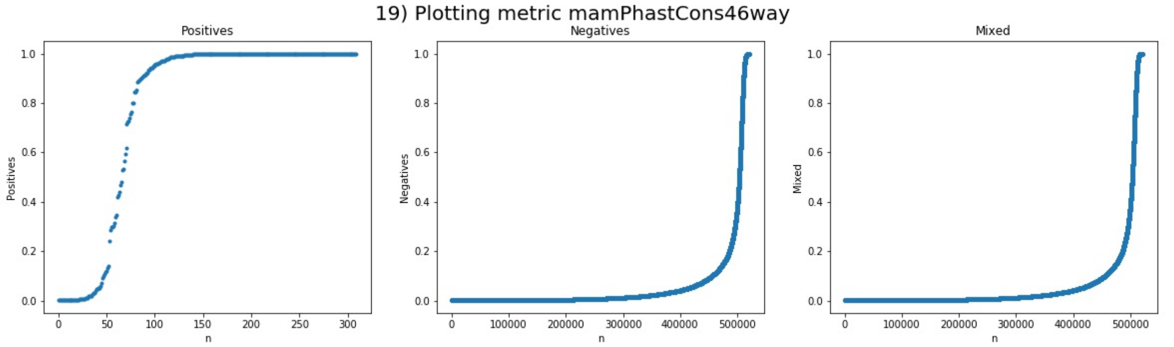


Figura 2.38: Values of metric mamPhastCons46way

2.20 mamPhyloP46way

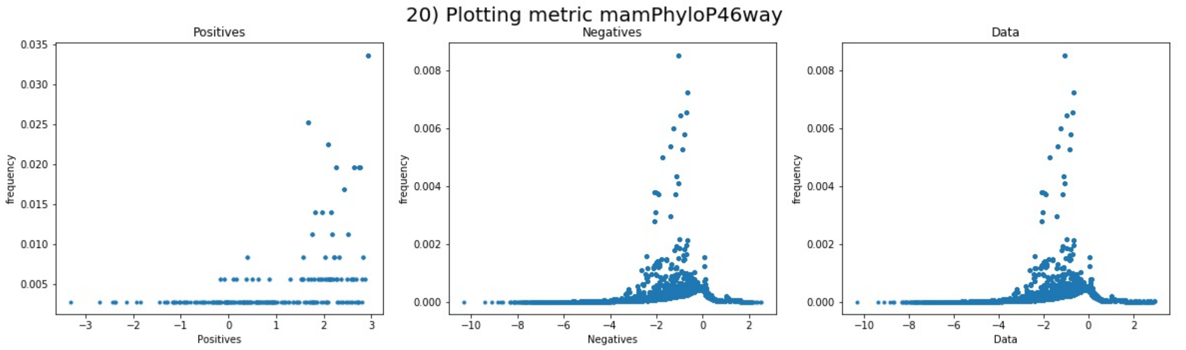


Figura 2.39: Sampling distribution of metric mamPhyloP46way

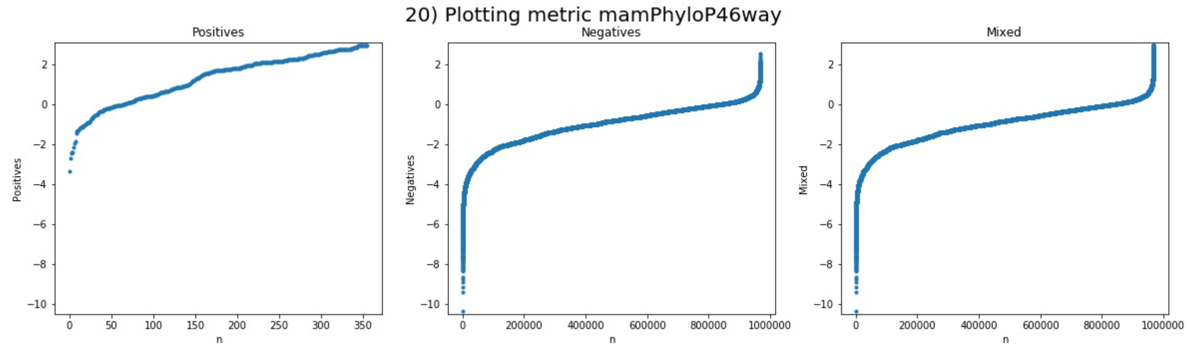


Figura 2.40: Values of metric mamPhyloP46way

2.21 numTFBSConserved

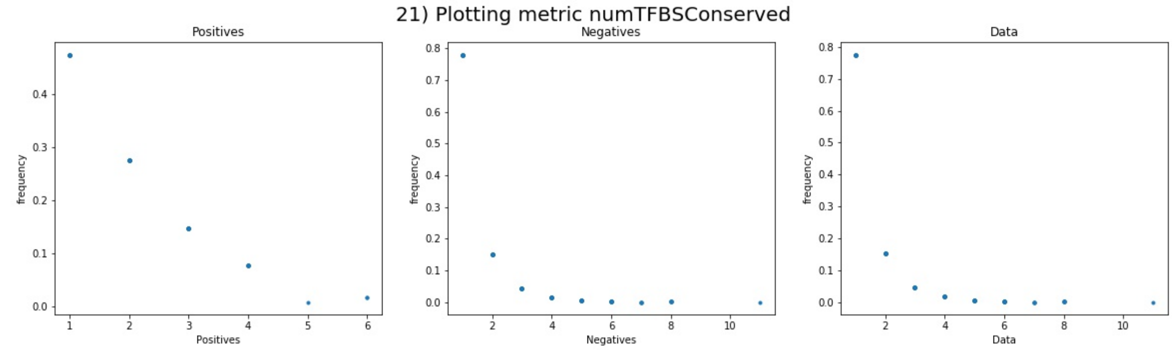


Figura 2.41: Sampling distribution of metric numTFBSConserved

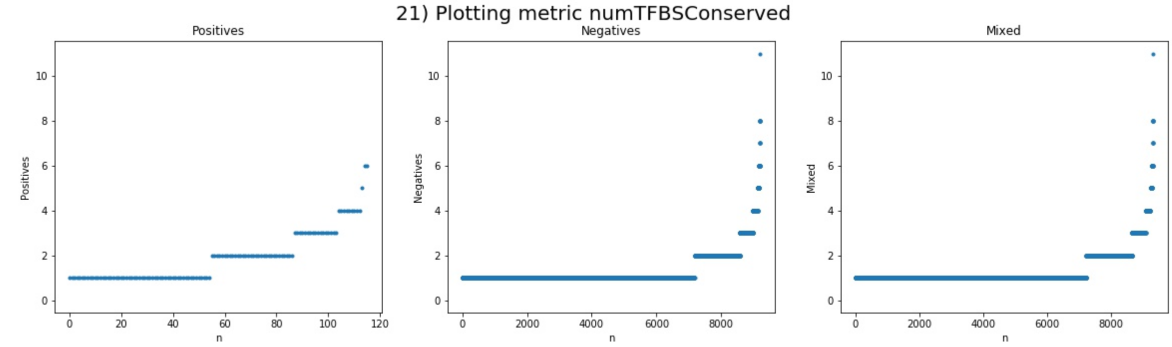


Figura 2.42: Values of metric numTFBSConserved

2.22 priPhastCons46way

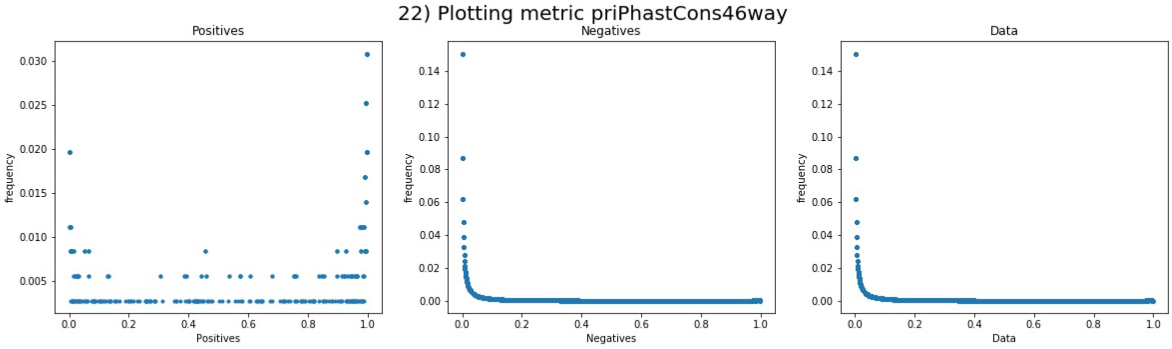


Figura 2.43: Sampling distribution of metric priPhastCons46way

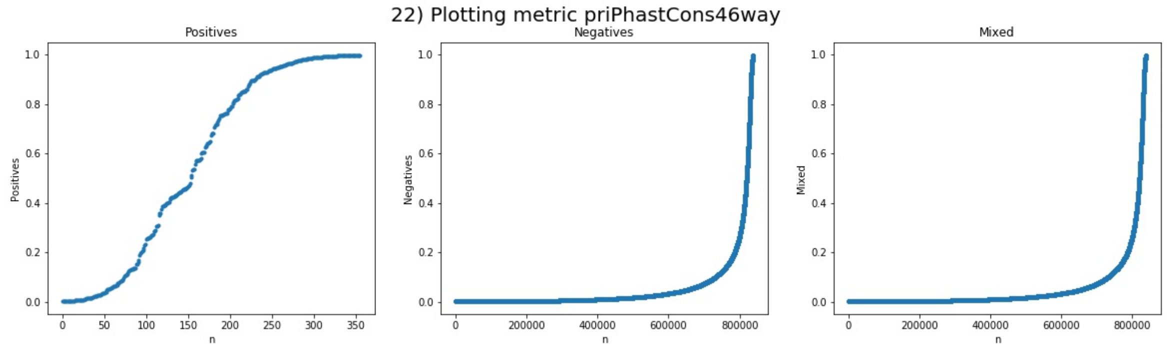


Figura 2.44: Values of metric priPhastCons46way

2.23 priPhyloP46way

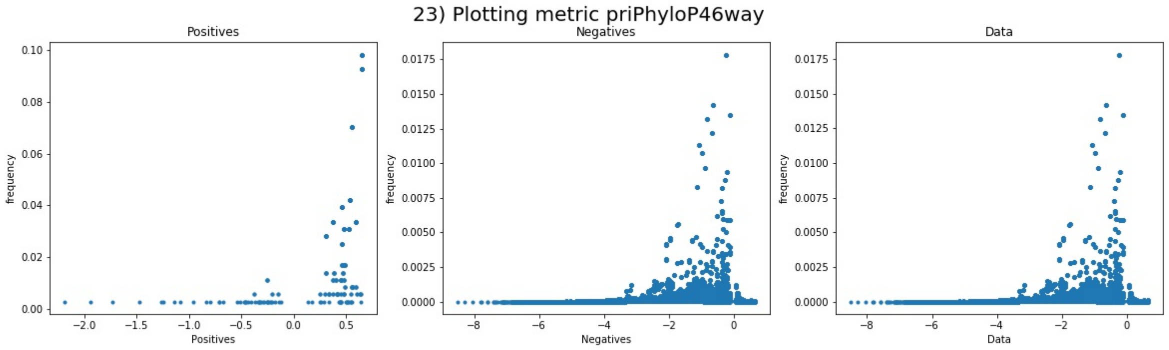


Figura 2.45: Sampling distribution of metric priPhyloP46way

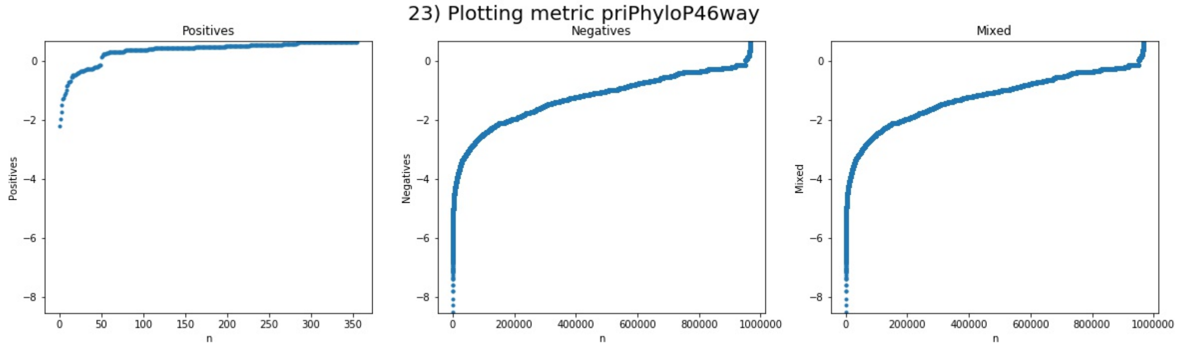


Figura 2.46: Values of metric priPhyloP46way

2.24 rareVar

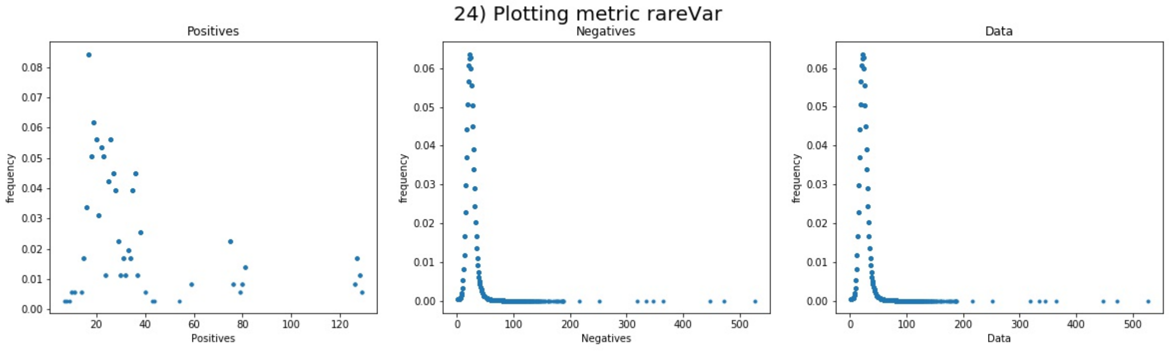


Figura 2.47: Sampling distribution of metric rareVar

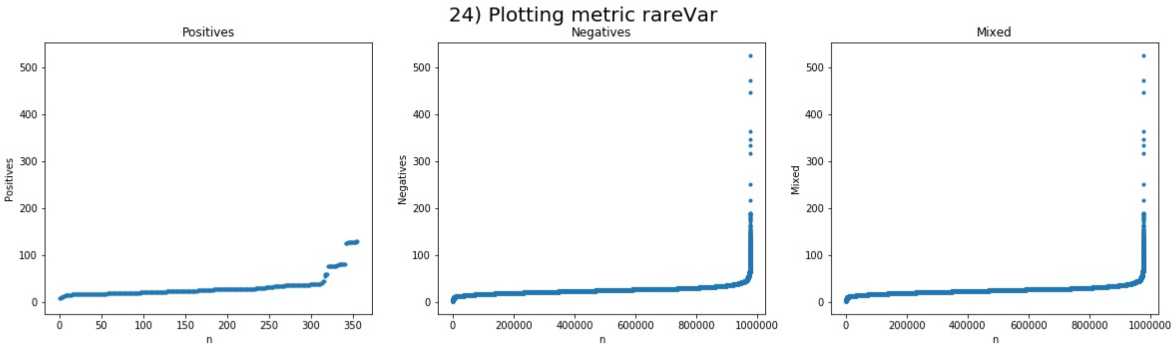


Figura 2.48: Values of metric rareVar

2.25 verPhastCons46way

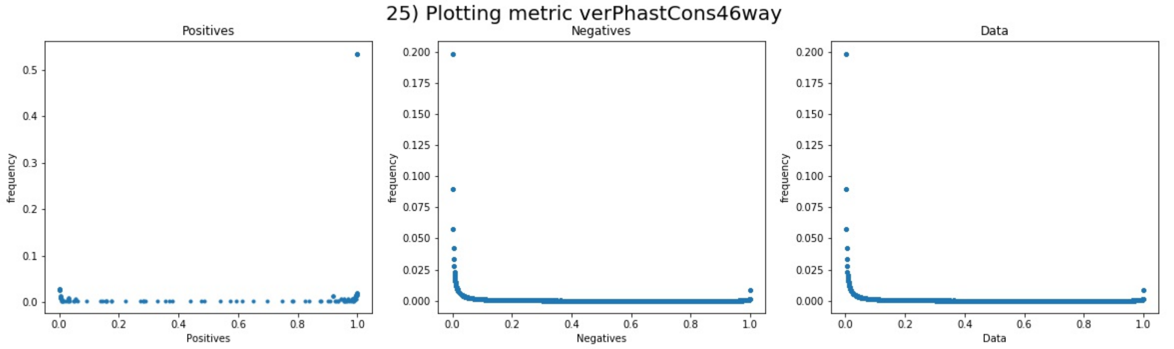


Figura 2.49: Sampling distribution of metric verPhastCons46way

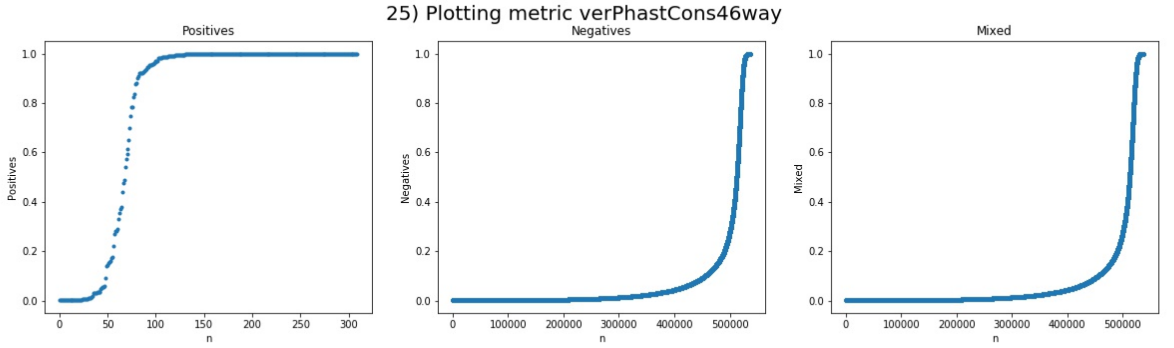


Figura 2.50: Values of metric verPhastCons46way

2.26 verPhyloP46way

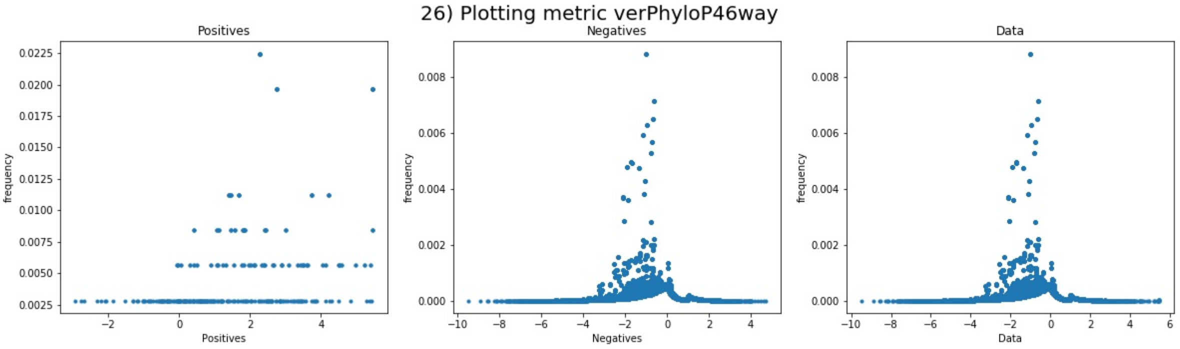


Figura 2.51: Sampling distribution of metric verPhyloP46way

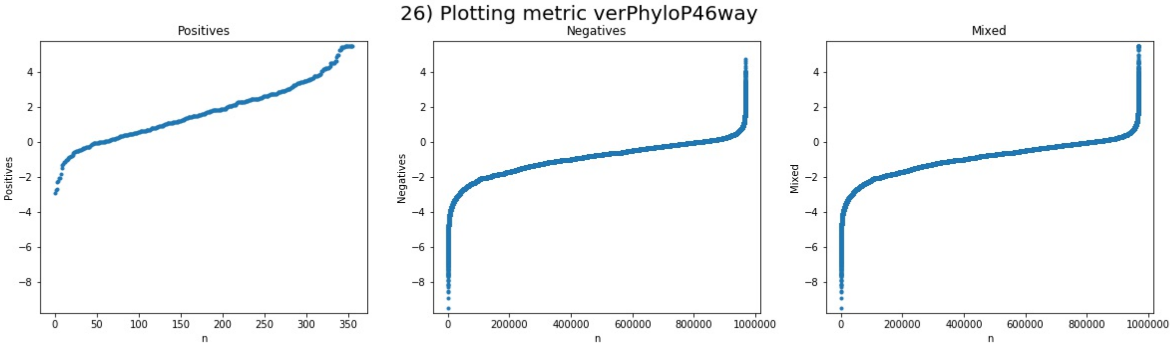


Figura 2.52: Values of metric verPhyloP46way

Parte II

Theory

3 Input modelling

3.1 Input values

The values used for each metric are the 3 following:

3.1.1 Normalized metric

Clearly one of the important metrics is the metric itself, that will be normalized to allow for input in $[0,1]$ range, zero mean and unary variance:

$$\text{metric}' = \frac{\text{metric} - \min\{\text{metric values}\}}{\max\{\text{metric values}\} - \min\{\text{metric values}\}}$$

(a) Input normalization to $[0,1]$ range

$$\text{metric}'' = \frac{\text{metric'} - \mathbb{E}\left(\{\}|\text{ metric' values}\right)}{\sqrt{\text{Var}\left(\text{metric' values}\right)}}$$

(b) Input normalization to zero mean and unary variance

3.1.2 Rarity

Another value we will be using in the input layer of the network is the rarity of the metric value, modelled as the surprise of extracting the given value from the estimated metric distribution extrapolated out of the sampling distribution.

If M is the estimated metric distribution and m is the value assumed by the metric in the given data point, we can model **rarity** as follows:

$$\text{rarity}(m) = 1 - M(m)$$

Figura 3.2: Rarity

3.1.3 Entropy

The third and final value used will be **entropy**, obtained using the estimated metric probability:

$$H(x) = -\mathbb{P}\left(x\right)\log\mathbb{P}\left(x\right)$$

Figura 3.3: Entropy

3.2 Feet

The input layer is comprised of 26 (number of metrics) *feet*, meaning tiny networks that are used to limit the initial linear combination of the metric input values to themselves.

Each feet is modelled as a locally connected dense layer.

3.3 Oversampling of positives

Since the positive values are just the 0.036% of the dataset we'll oversample these to weight more these values. Since the variance of positive data points is too high to extrapolate a distribution to generate significant new fuzzy data points, simple duplication will be used.

3.4 Undersampling of negatives

Since the negative values are more than the 99.96% of the dataset we'll undersample these to weight less these values.

3.5 Absence of information

Absence of information about a given metric will be modelled as **zeros**, meaning all values relative to the given absent metric for that data point will be treated as zero.

4 Output modelling

The output layer of the neural network is modelled by two neurons, one representing the positive class and one the negative class.

5 Weight initialization

5.1 Gaussian noise initialization

6
Locally connected dense layers

6.1 Leaky RELU

7 Dense layers

7.1 SELU

7.2 Drop out

Parte III

Code