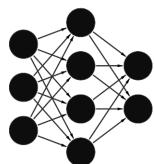


PREDICTION OF PATHOGENIC SNV

Prof. Giorgio Valentini
6 CFU

Luca Cappelletti

Lecture Notes
Year 2017/2018



IT Master Degree
Universiy of Milan
Italy
28 giugno 2018

Indice

I Dataset	4
1 Data points	5
1.1 Retrieving the dataset	5
1.2 Composition	5
1.2.1 Training dataset	5
1.2.2 Testing dataset	5
2 Metrics	6
2.1 How the graphs are realized	6
2.1.1 Metric sample distribution	6
2.1.2 Plot graphs	6
2.1.3 Normalized plot graphs	6
2.2 CpGobsExp	7
2.2.1 Metric sample distribution	7
2.2.2 Metric values	7
2.3 CpGperCpG	8
2.3.1 Metric sample distribution	8
2.3.2 Metric values	8
2.4 CpGperGC	9
2.4.1 Metric sample distribution	9
2.4.2 Metric values	9
2.5 DGVCount	10
2.5.1 Metric sample distribution	10
2.5.2 Metric values	10
2.6 DnaseClusteredHyp	11
2.6.1 Metric sample distribution	11
2.6.2 Metric values	11
2.7 DnaseClusteredScore	12
2.7.1 Metric sample distribution	12
2.7.2 Metric values	12
2.8 EncH3K27Ac	13
2.8.1 Metric sample distribution	13
2.8.2 Metric values	13
2.9 EncH3K4Me1	14
2.9.1 Metric sample distribution	14
2.9.2 Metric values	14
2.10 EncH3K4Me3	15
2.10.1 Metric sample distribution	15
2.10.2 Metric values	15
2.11 GCContent	16
2.11.1 Metric sample distribution	16
2.11.2 Metric values	16
2.12 GerpRS	17
2.12.1 Metric sample distribution	17
2.12.2 Metric values	17
2.13 GerpRSpv	18
2.13.1 Metric sample distribution	18

2.13.2 Metric values	18
2.14 ISCApath	19
2.14.1 Metric sample distribution	19
2.14.2 Metric values	19
2.15 commonVar	20
2.15.1 Metric sample distribution	20
2.15.2 Metric values	20
2.16 dbVARCount	21
2.16.1 Metric sample distribution	21
2.16.2 Metric values	21
2.17 fantom5Perm	22
2.17.1 Metric sample distribution	22
2.17.2 Metric values	22
2.18 fantom5Robust	23
2.18.1 Metric sample distribution	23
2.18.2 Metric values	23
2.19 fracRareCommon	24
2.19.1 Metric sample distribution	24
2.19.2 Metric values	24
2.20 mamPhastCons46way	25
2.20.1 Metric sample distribution	25
2.20.2 Metric values	25
2.21 mamPhyloP46way	26
2.21.1 Metric sample distribution	26
2.21.2 Metric values	26
2.22 numTFBSConserved	27
2.22.1 Metric sample distribution	27
2.22.2 Metric values	27
2.23 priPhastCons46way	28
2.23.1 Metric sample distribution	28
2.23.2 Metric values	28
2.24 priPhyloP46way	29
2.24.1 Metric sample distribution	29
2.24.2 Metric values	29
2.25 rareVar	30
2.25.1 Metric sample distribution	30
2.25.2 Metric values	30
2.26 verPhastCons46way	31
2.26.1 Metric sample distribution	31
2.26.2 Metric values	31
2.27 verPhyloP46way	32
2.27.1 Metric sample distribution	32
2.27.2 Metric values	32
3 Metric distribution summary	33
4 Scatter plot	34
4.1 Scatter plot	34
4.2 Identified data correlations	35
4.2.1 dbVARCount and DGVCount	35
4.2.2 mamPhyloP46way and verPhyloP46way	36
4.2.3 mamPhastCons46way and verPhastCons46way	36
4.2.4 mamPhastCons46way and verPhastCons46way	37
4.2.5 verPhastCons46way and priPhastCons46way	37
4.2.6 priPhastCons46way and mamPhastCons46way	38
II Theory	39
5 Input modelling	40
5.1 Input values	40
5.1.1 Normalized metric	40

5.1.2	Probability	40
5.1.3	Deviation from mean	40
5.2	Feet	40
5.3	Oversampling of positives	41
5.4	Undersampling of negatives	41
5.5	Oversampling and undersampling targets	41
5.6	Absence of information	41
6	Output modelling	42
7	Weight initialization	43
7.1	Weight distribution based on input distribution	43
7.2	Weight distribution based on activation functions and regularization layers	43
8	Locally connected dense layers	44
8.1	Activation function	44
9	Dense layers	45
9.1	Activation function	45
9.2	Regularization	45
9.3	Drop out	45
10	References	46

Parte I

Dataset

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 Composition

1.2.1 Training dataset

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

1.2.2 Testing dataset

In the test dataset there are 190189 data points, still each one comprised of 26 metrics. The first 40 are pathogenic and the following are negative.

2

Metrics

2.1 How the graphs are realized

All the graphs are in triples: positives, negatives and mixed. All the zeros are removed as in most metrics *seemed* to indicate an unknown value.

2.1.1 Metric sample distribution

Are realized by calculating the frequencies and estimating the density distributions parameters via MLE.

2.1.2 Plot graphs

Are realized by sorting the values of the metric.

2.1.3 Normalized plot graphs

Are realized by sorting the values of the metric, with the domain and codomain normalized.

2.2 CpGobsExp

2.2.1 Metric sample distribution

The data points seem to follow a **Beta** distribution with the following parameters:

$$\alpha = 7.6689746880295795 \quad \beta = 6778383.524935903 \\ \text{loc} = -0.09826818916997124 \quad \text{scale} = 306278.3184506849$$

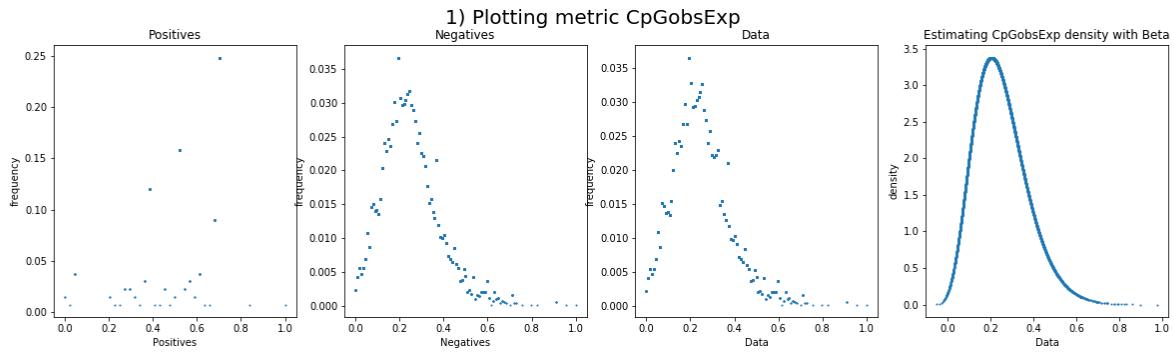


Figura 2.1: Sampling distribution of metric CpGobsExp

2.2.2 Metric values

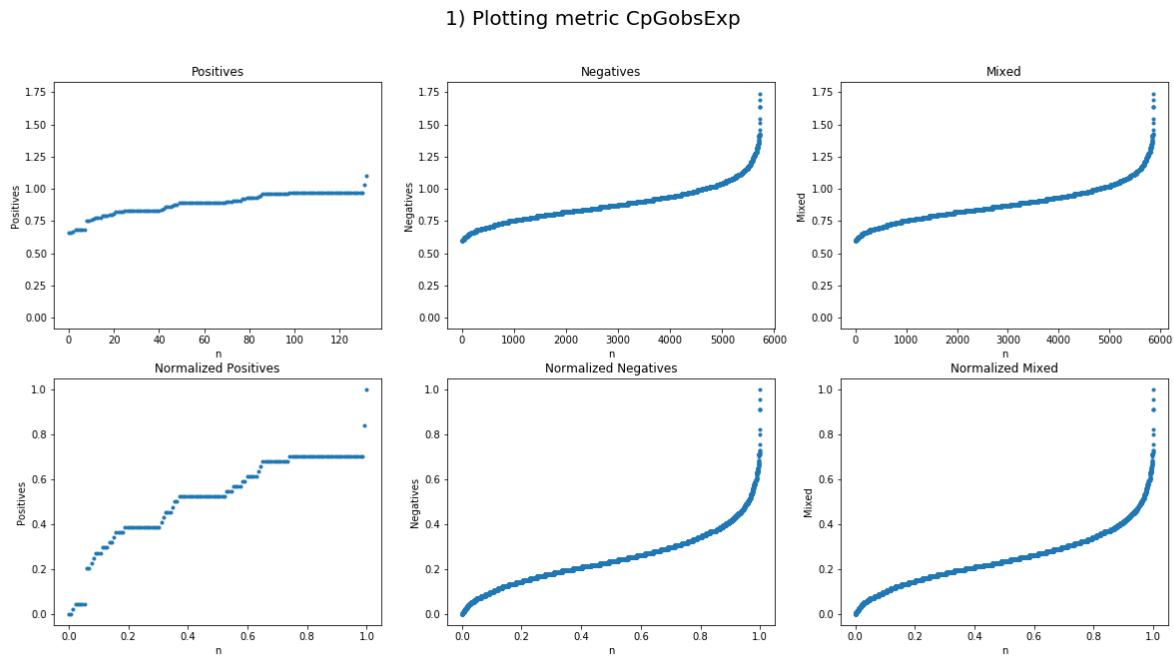


Figura 2.2: Values of metric CpGobsExp

2.3 CpGperCpG

2.3.1 Metric sample distribution

The data points seem to follow a **Beta** distribution with the following parameters:

$$\alpha = 6.402175341881067 \quad \beta = 97129163.31117742 \\ \text{loc} = -0.05698922703576313 \quad \text{scale} = 4337764.42876015$$

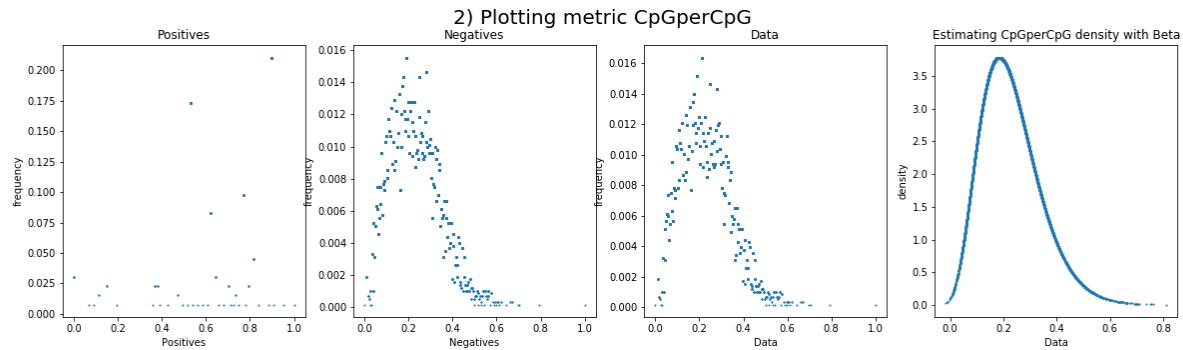


Figura 2.3: Sampling distribution of metric CpGperCpG

2.3.2 Metric values

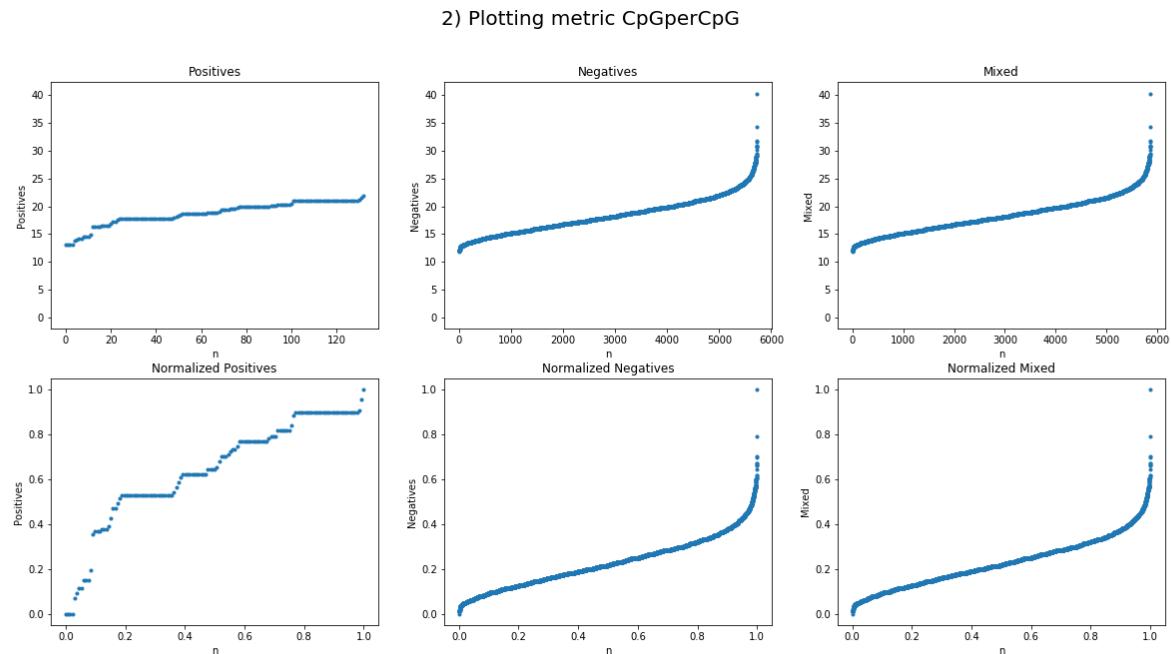


Figura 2.4: Values of metric CpGperCpG

2.4 CpGperGC

2.4.1 Metric sample distribution

The data points seem to follow a **Gaussian** distribution with the following parameters:

$$\mathbb{E}(X) = 0.4602356242601636 \quad \text{Var}(X) = 0.15949294643574352$$

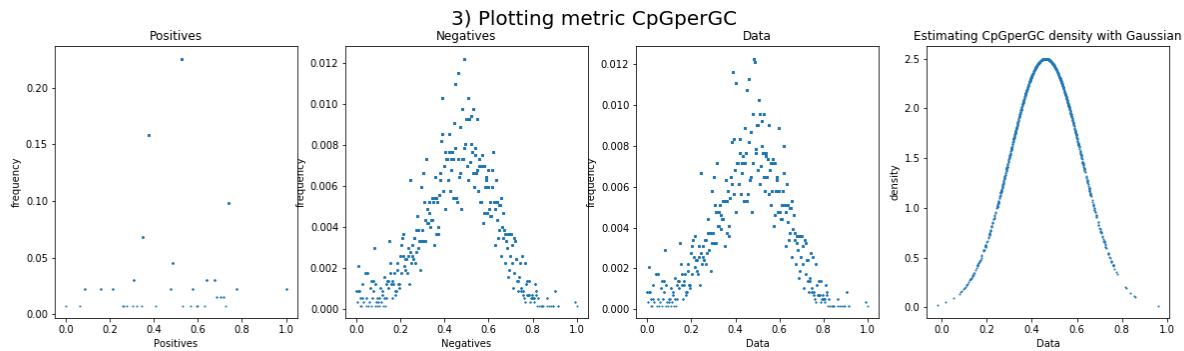


Figura 2.5: Sampling distribution of metric CpGperGC

2.4.2 Metric values

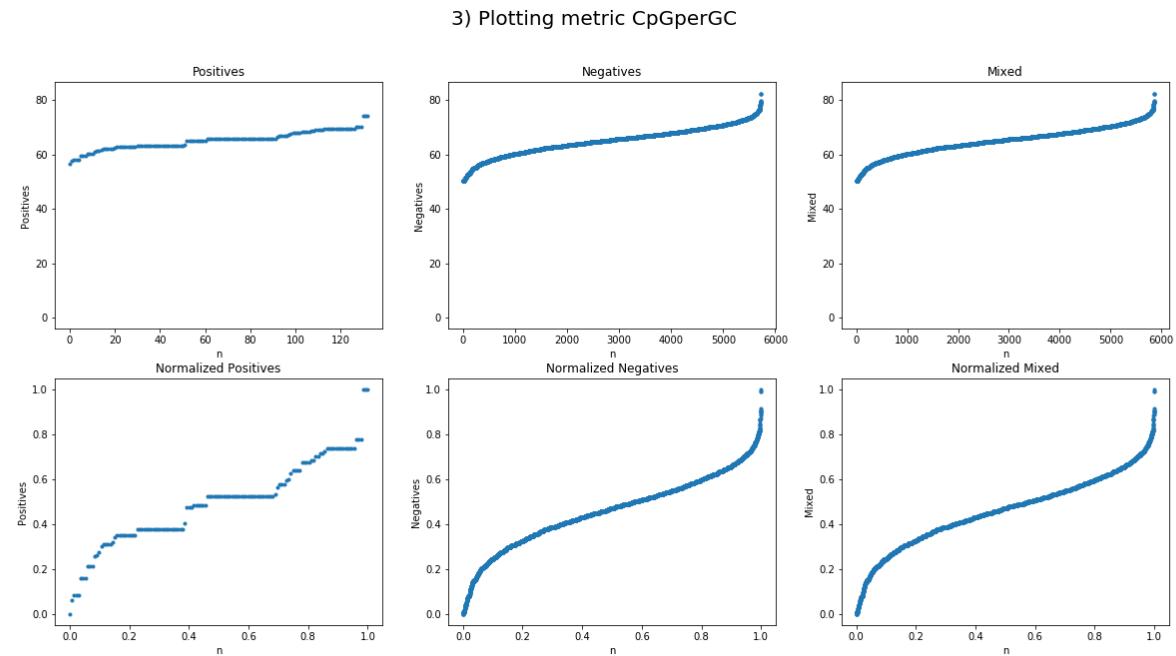


Figura 2.6: Values of metric CpGperGC

2.5 DGVCount

2.5.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.20940038672579409 \quad \text{loc} = -1.1962983066939984e - 30 \quad \text{scale} = 1.2347090894162929$$

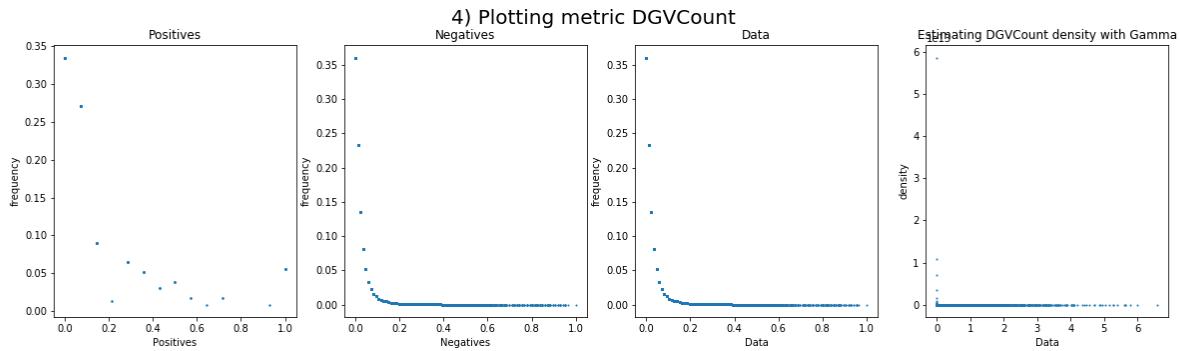


Figura 2.7: Sampling distribution of metric DGVCount

2.5.2 Metric values

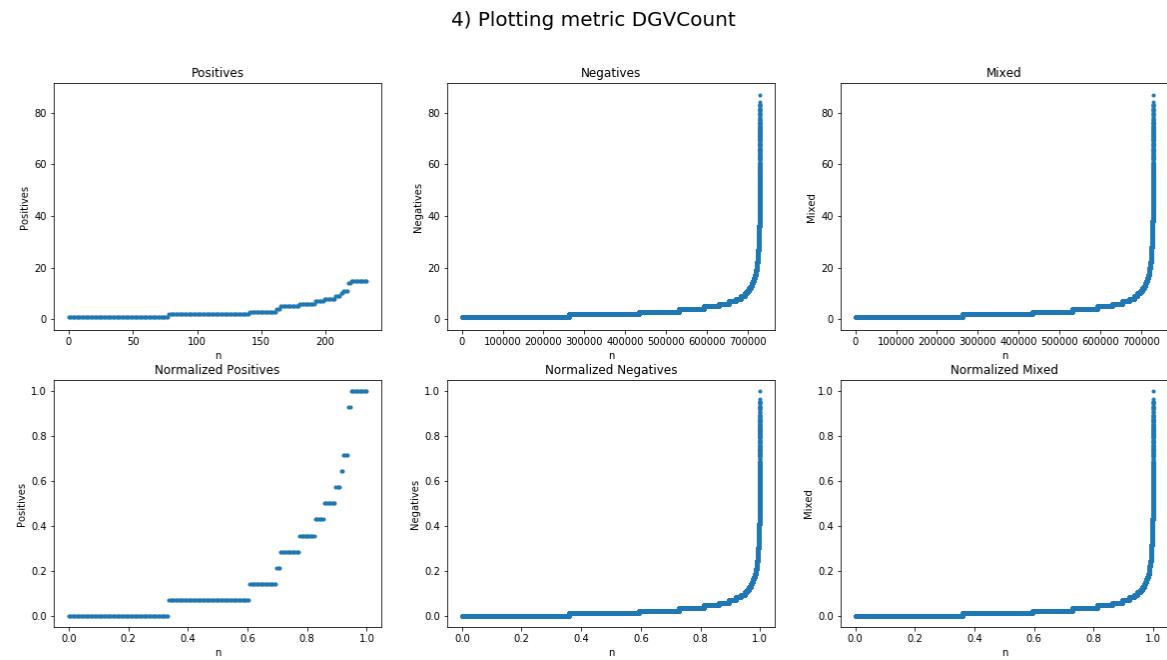


Figura 2.8: Values of metric DGVCount

2.6 DnaseClusteredHyp

2.6.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.4176887081406805 \quad \text{loc} = -3.362626207862299e-29 \quad \text{scale} = 0.3676310948709975$$

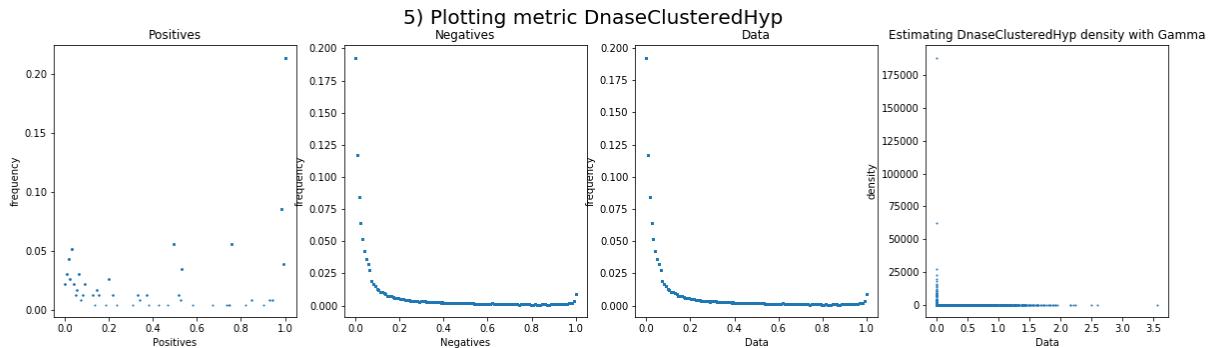


Figura 2.9: Sampling distribution of metric DnaseClusteredHyp

2.6.2 Metric values

5) Plotting metric DnaseClusteredHyp

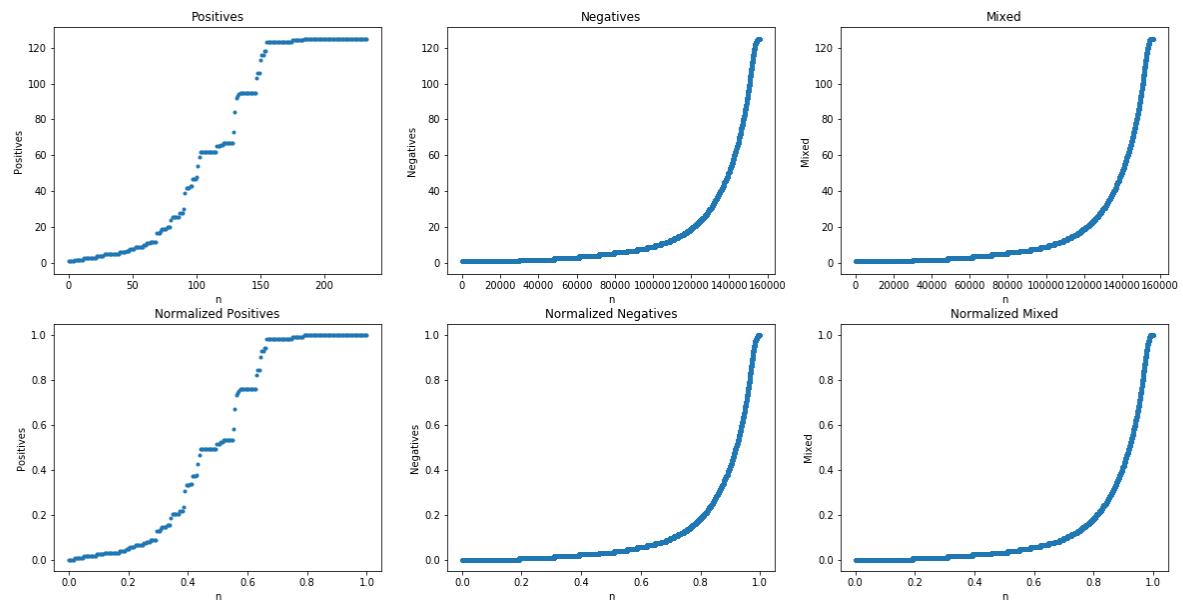


Figura 2.10: Values of metric DnaseClusteredHyp

2.7 DnaseClusteredScore

2.7.1 Metric sample distribution

The data points seem to follow **slightly** a **Beta** distribution with the following parameters:

$$\alpha = 0.2709657632937803 \quad \beta = 0.44530002562349713 \\ \text{loc} = -0.09309893086089688 \quad \text{scale} = 1.0930989308608972$$

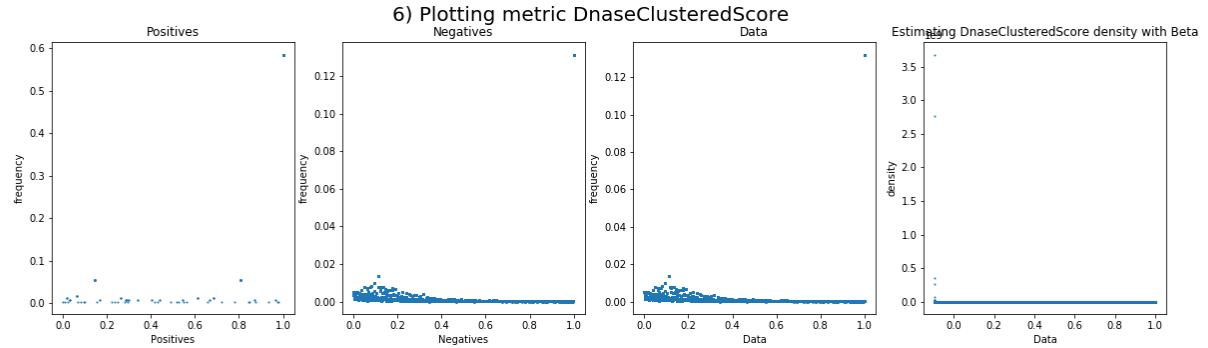


Figura 2.11: Sampling distribution of metric DnaseClusteredScore

2.7.2 Metric values

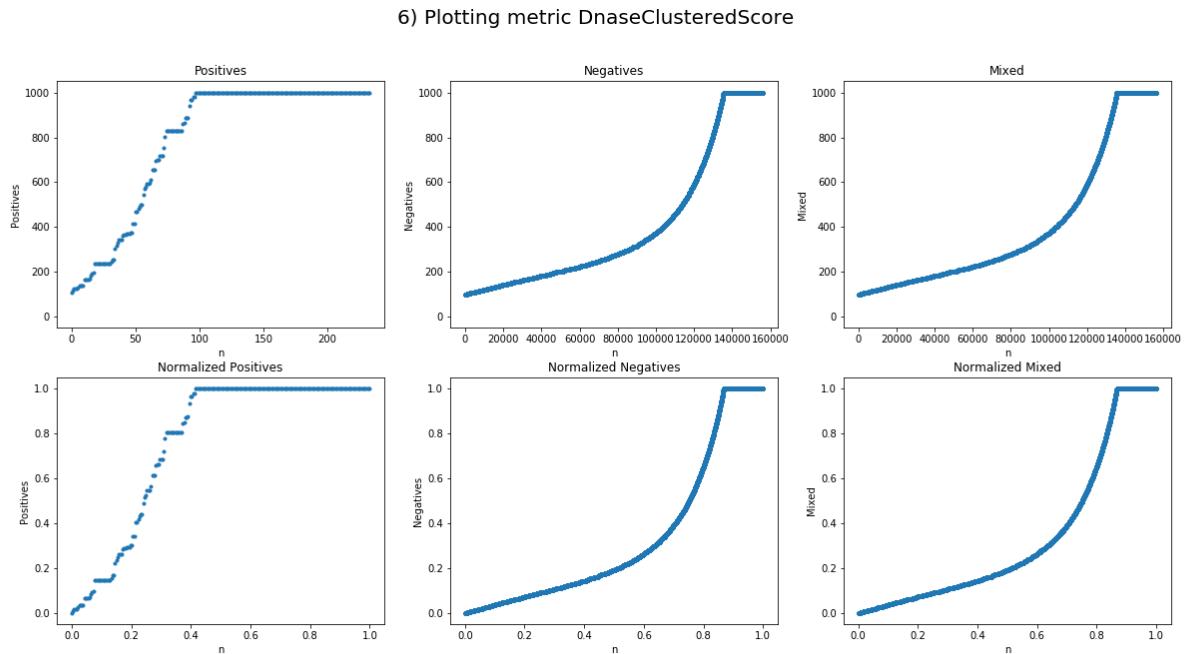


Figura 2.12: Values of metric DnaseClusteredScore

2.8 EncH3K27Ac

2.8.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

$$\alpha = 0.0004042086221537893 \quad \text{loc} = -2.859398162696207e-24 \quad \text{scale} = 0.03076944787133299$$

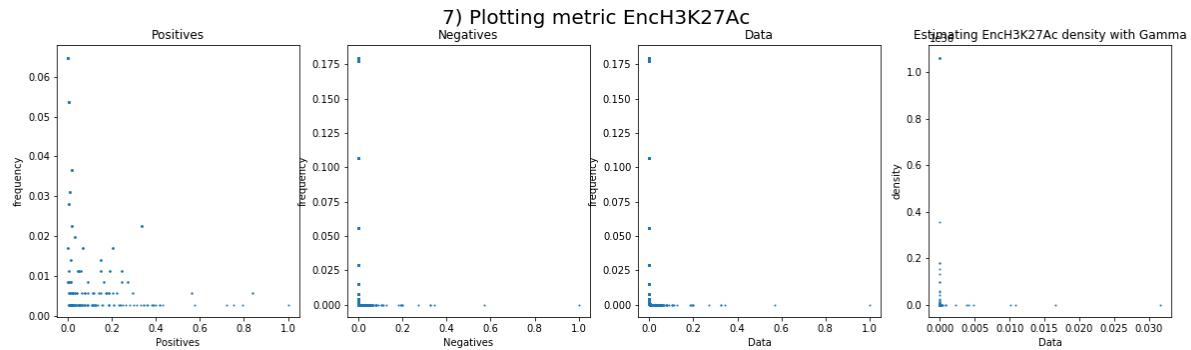


Figura 2.13: Sampling distribution of metric EncH3K27Ac

2.8.2 Metric values

7) Plotting metric EncH3K27Ac

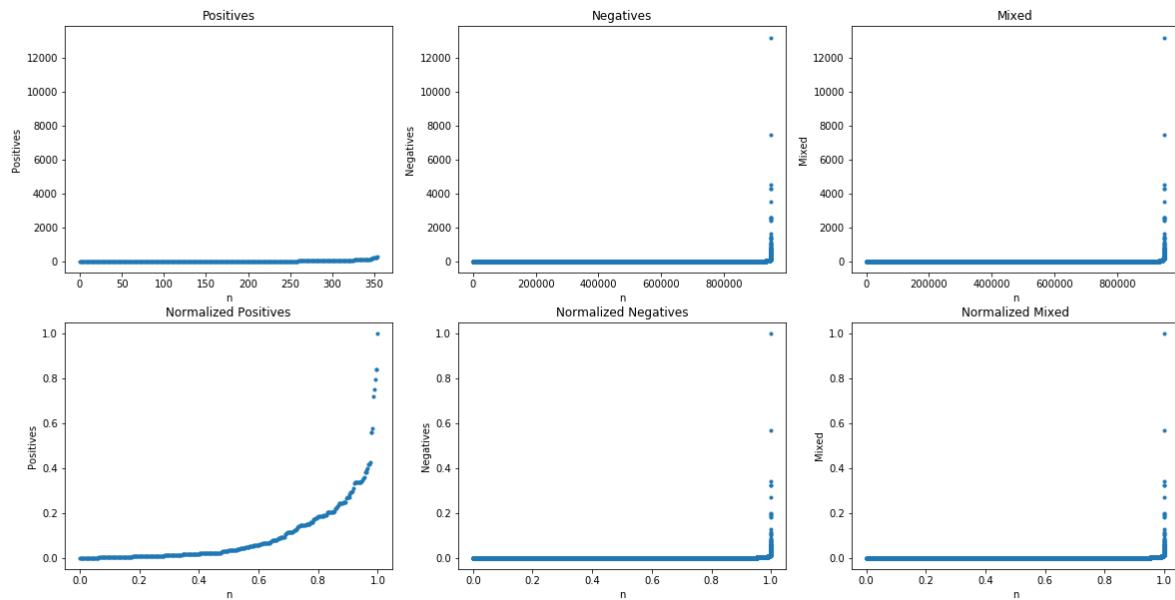


Figura 2.14: Values of metric EncH3K27Ac

2.9 EncH3K4Me1

2.9.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

$$\alpha = 0.22566387737236238 \quad \text{loc} = -6.619765504581537e-27 \quad \text{scale} = 1.396157055181753$$

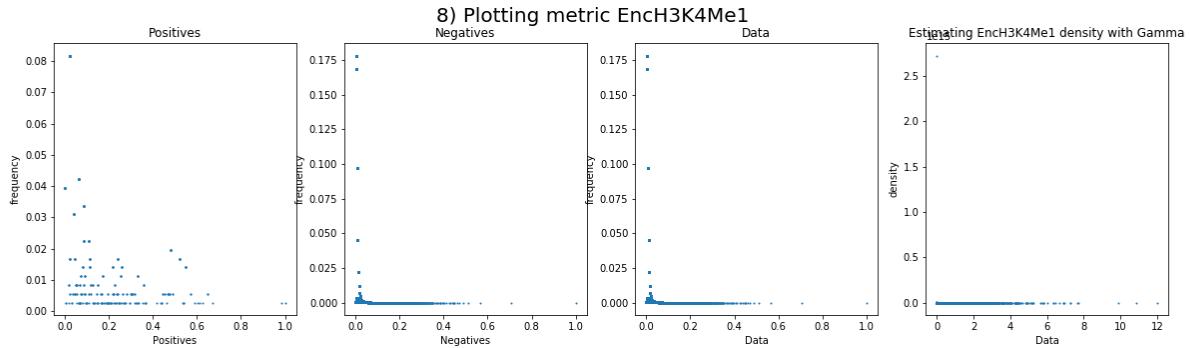


Figura 2.15: Sampling distribution of metric EncH3K4Me1

2.9.2 Metric values

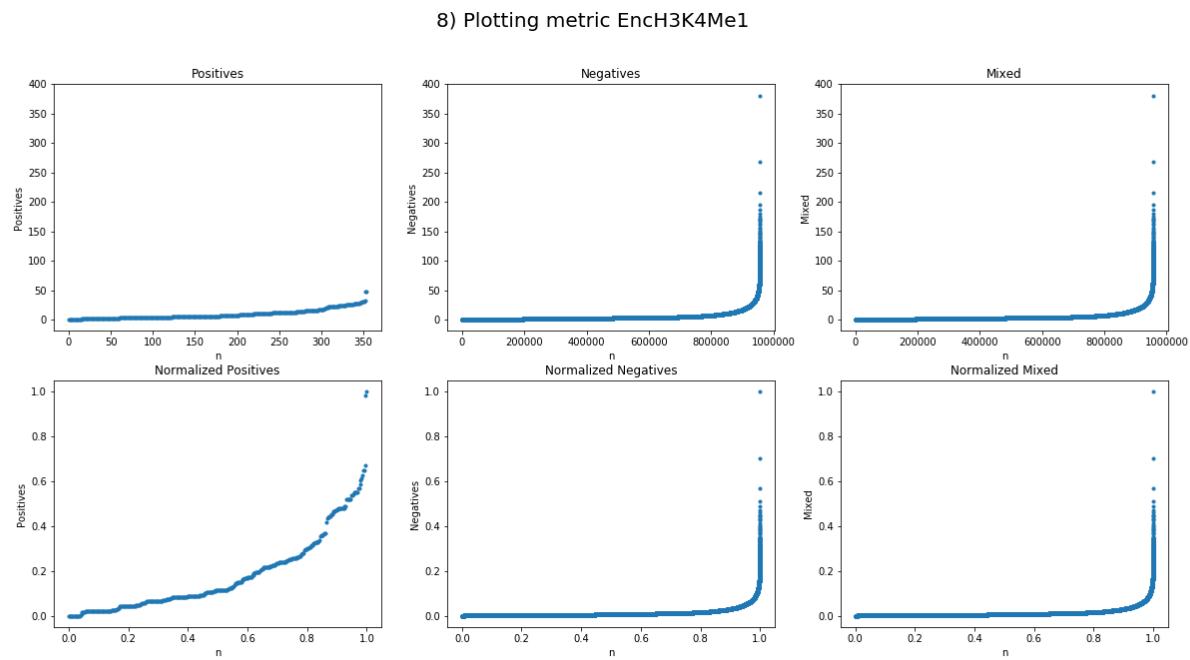


Figura 2.16: Values of metric EncH3K4Me1

2.10 EncH3K4Me3

2.10.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

$$\alpha = 0.007502428717446465 \quad \text{loc} = -3.469650119186857e-25 \quad \text{scale} = 0.04125297431971783$$

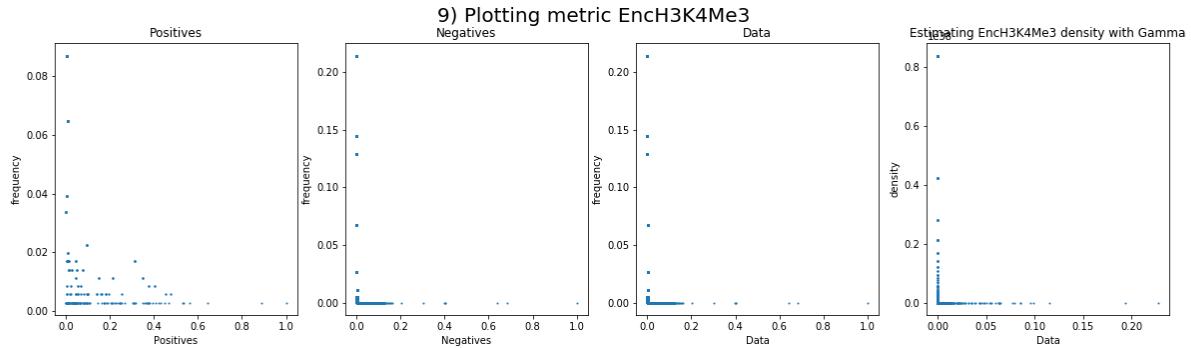


Figura 2.17: Sampling distribution of metric EncH3K4Me3

2.10.2 Metric values

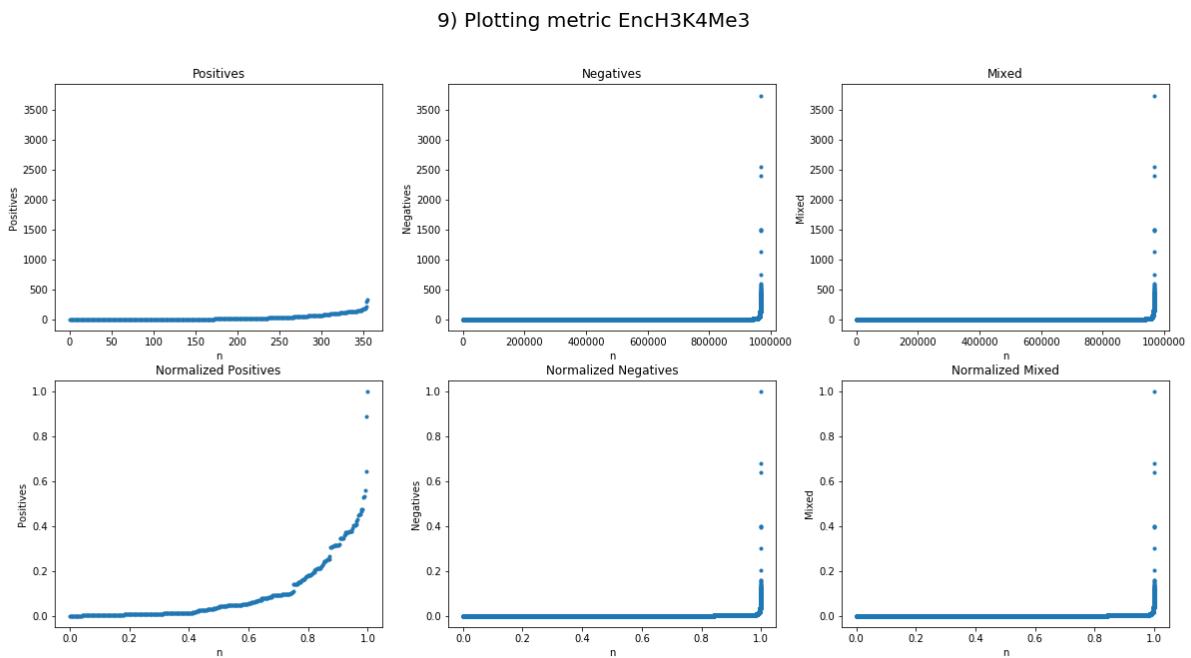


Figura 2.18: Values of metric EncH3K4Me3

2.11 GCContent

2.11.1 Metric sample distribution

The data points seem to be a combination of two **Gaussian** distributions. This will be approximated to one with the following parameters:

$$\mathbb{E}(X) = 0.4482813176478024 \quad \text{Var}(X) = 0.1097424869360011$$

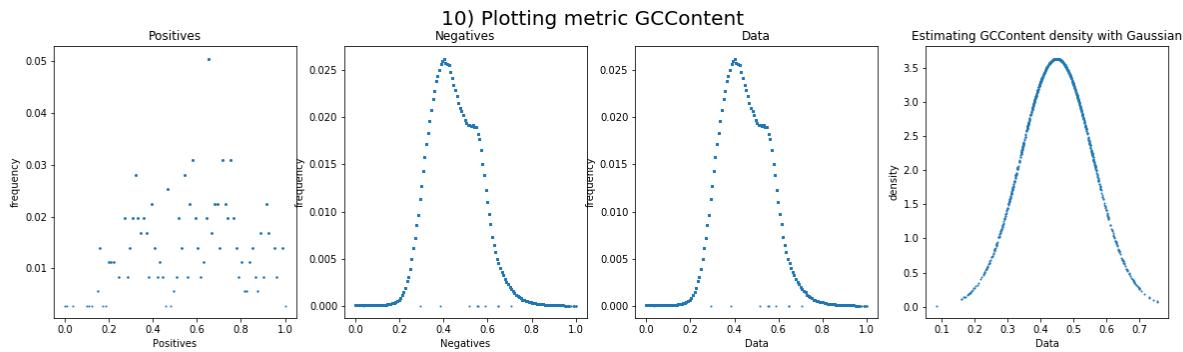


Figura 2.19: Sampling distribution of metric GCContent

2.11.2 Metric values

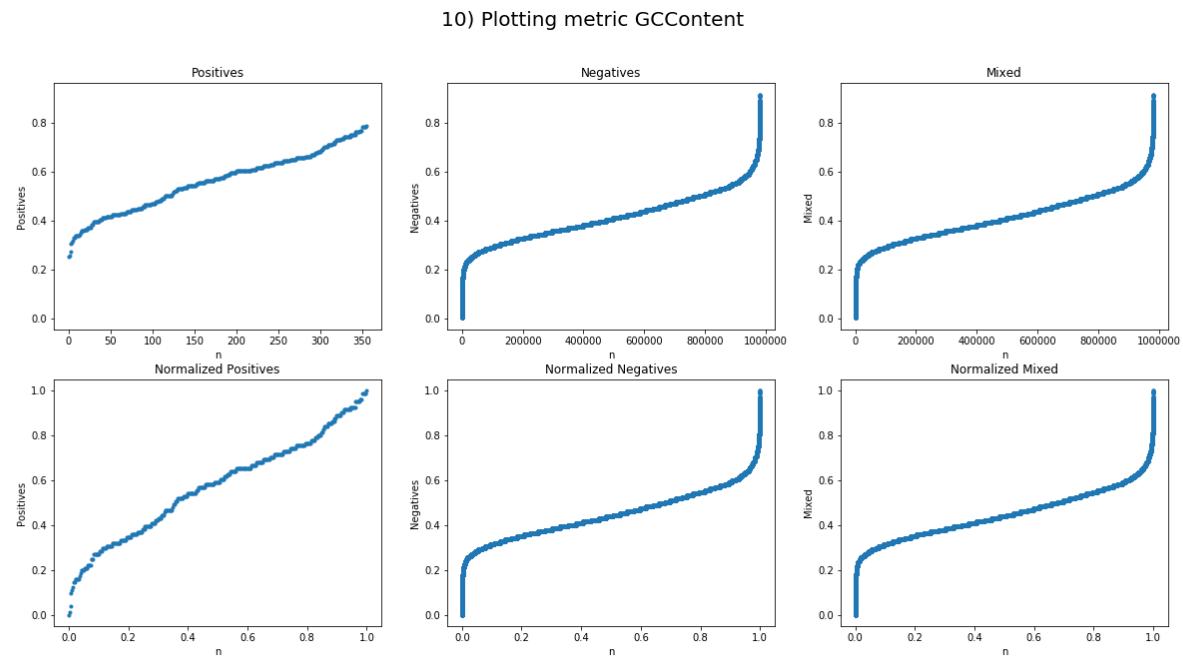


Figura 2.20: Values of metric GCContent

2.12 GerpRS

2.12.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

$$\alpha = 0.8688332877203315 \quad \text{loc} = -1.7081810436826354e-28 \quad \text{scale} = 0.11512094125204281$$

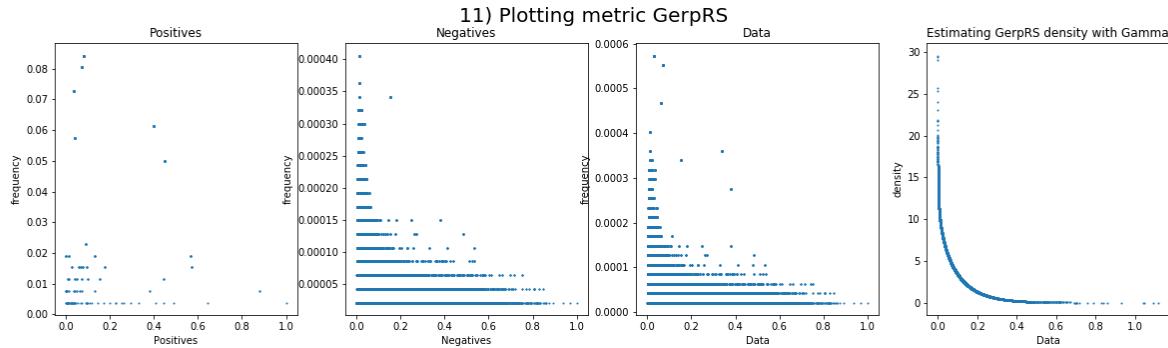


Figura 2.21: Sampling distribution of metric GerpRS

2.12.2 Metric values

11) Plotting metric GerpRS

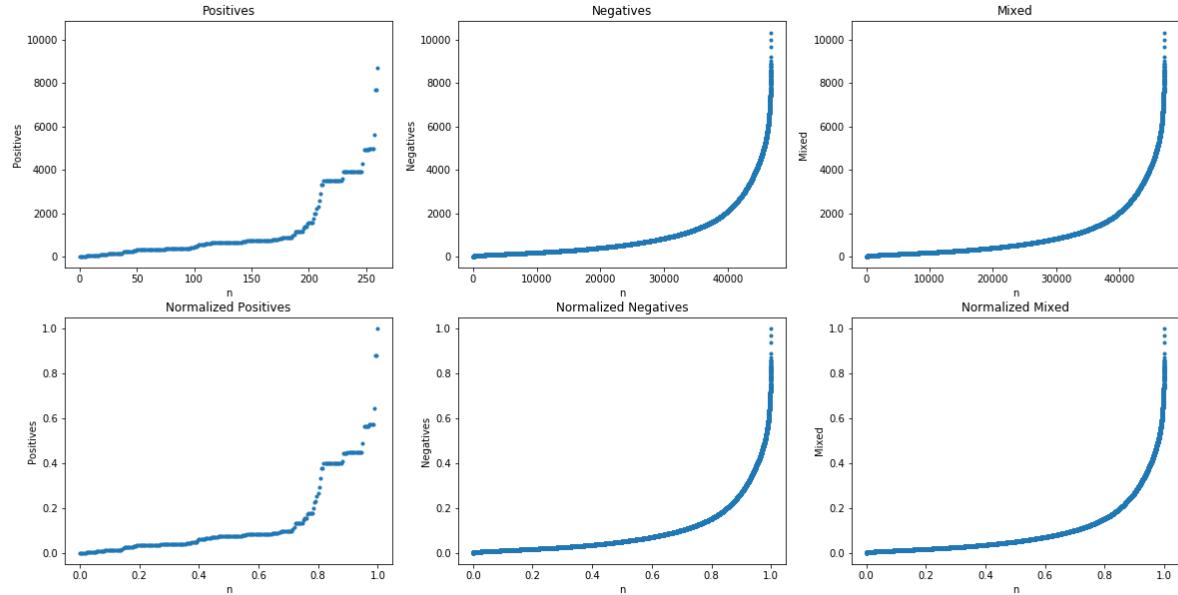


Figura 2.22: Values of metric GerpRS

2.13 GerpRSpv

2.13.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

$$\alpha = 0.5165290213220888 \quad \text{loc} = -6.952792177974854e-30 \quad \text{scale} = 0.2530358950266992$$

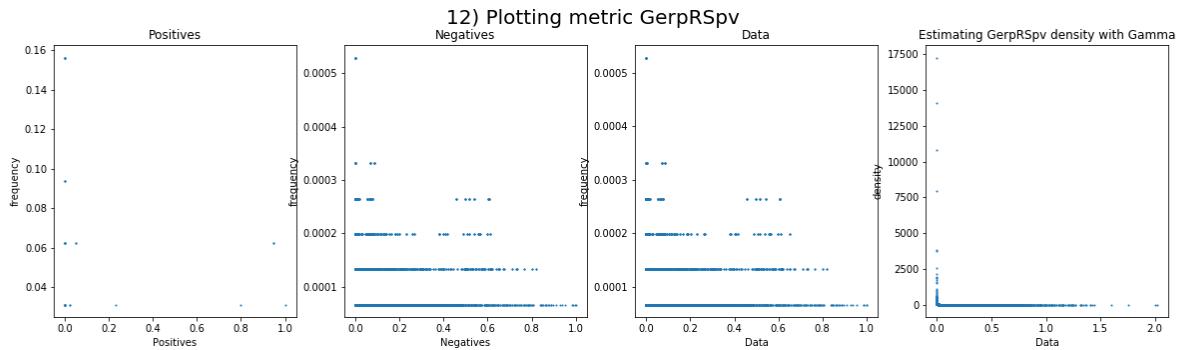


Figura 2.23: Sampling distribution of metric GerpRSpv

2.13.2 Metric values

12) Plotting metric GerpRSpv

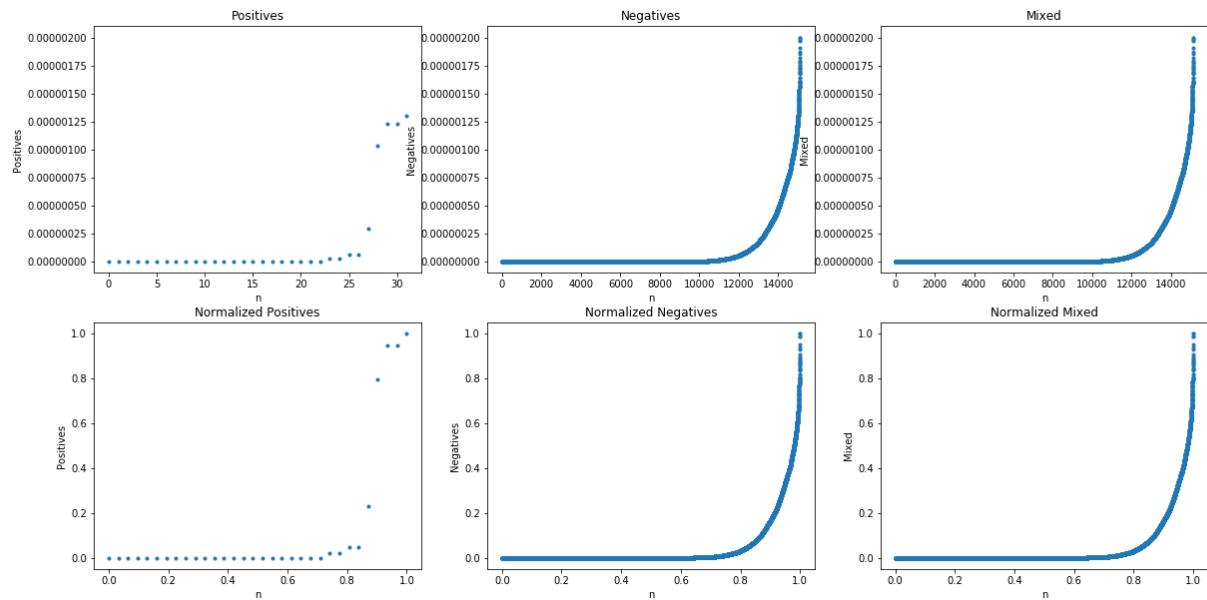


Figura 2.24: Values of metric GerpRSpv

2.14 ISCApath

2.14.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.08318618903703257 \quad \text{loc} = -1.9358902729364646e - 30 \quad \text{scale} = 1.2606790181148981$$

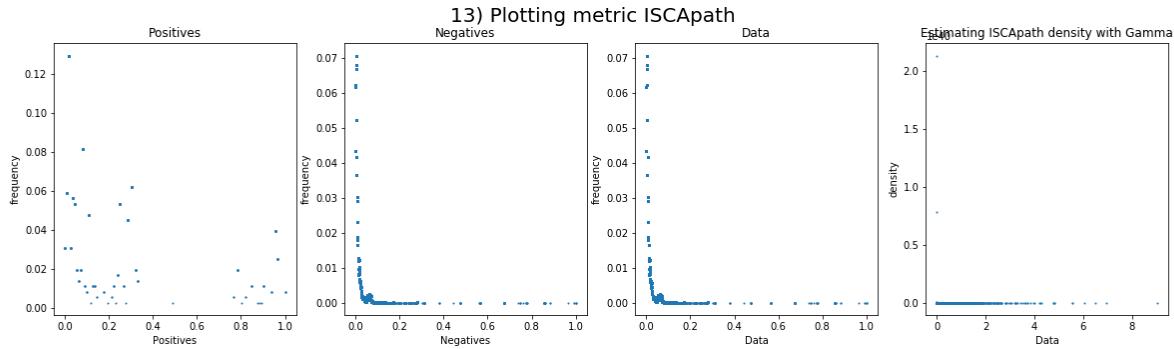


Figura 2.25: Sampling distribution of metric ISCApath

2.14.2 Metric values

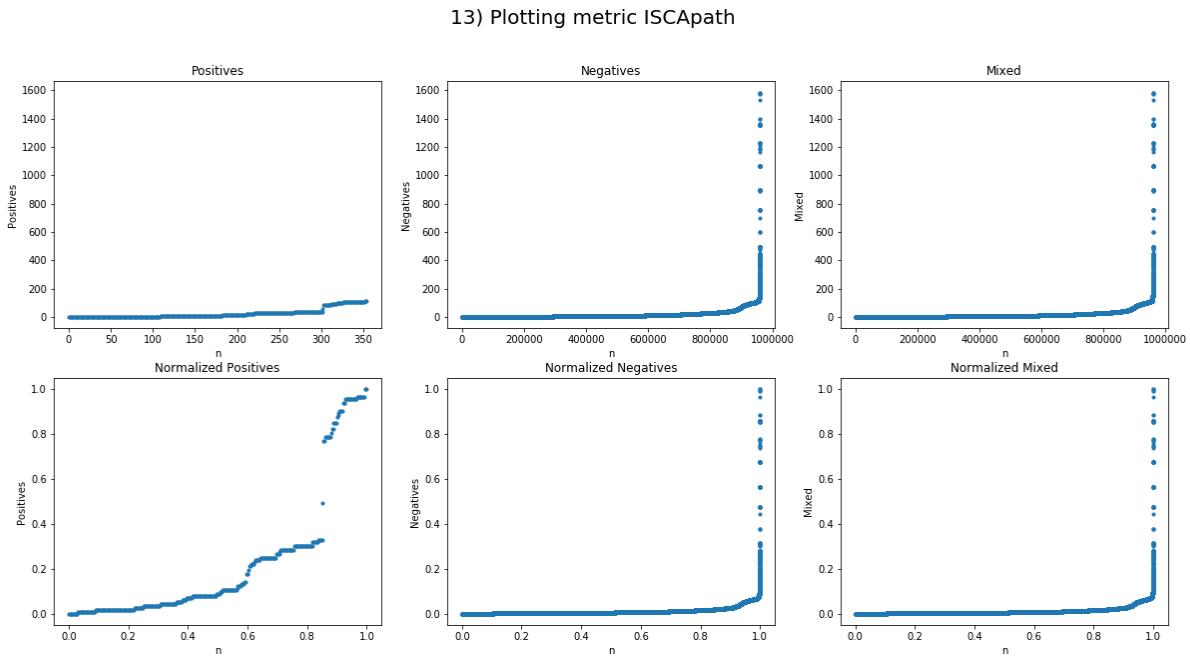


Figura 2.26: Values of metric ISCApath

2.15 commonVar

2.15.1 Metric sample distribution

The data points seem to follow an **Exponential Weibull** distribution with the following parameters:

$$\alpha = 5.038707296051438 \quad \beta = 1.0160276119461702 \\ \text{loc} = -0.012528678364149837 \quad \text{scale} = 0.025052745155722922$$

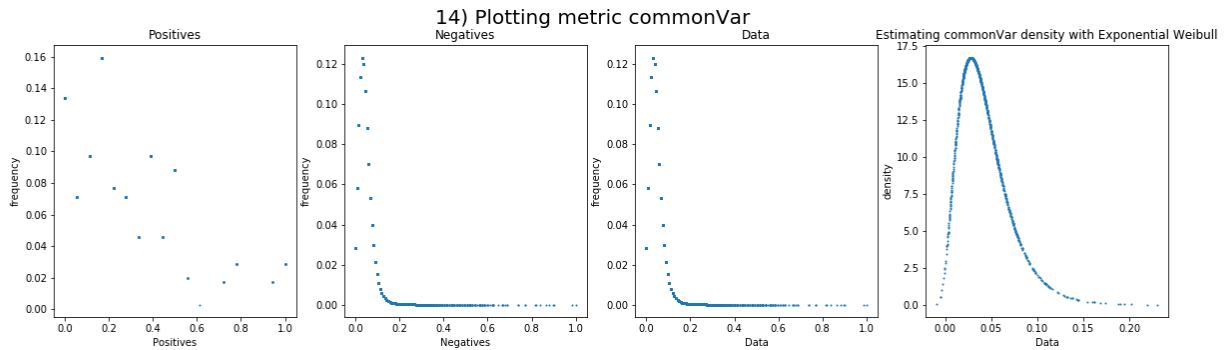


Figura 2.27: Sampling distribution of metric commonVar

2.15.2 Metric values

14) Plotting metric commonVar

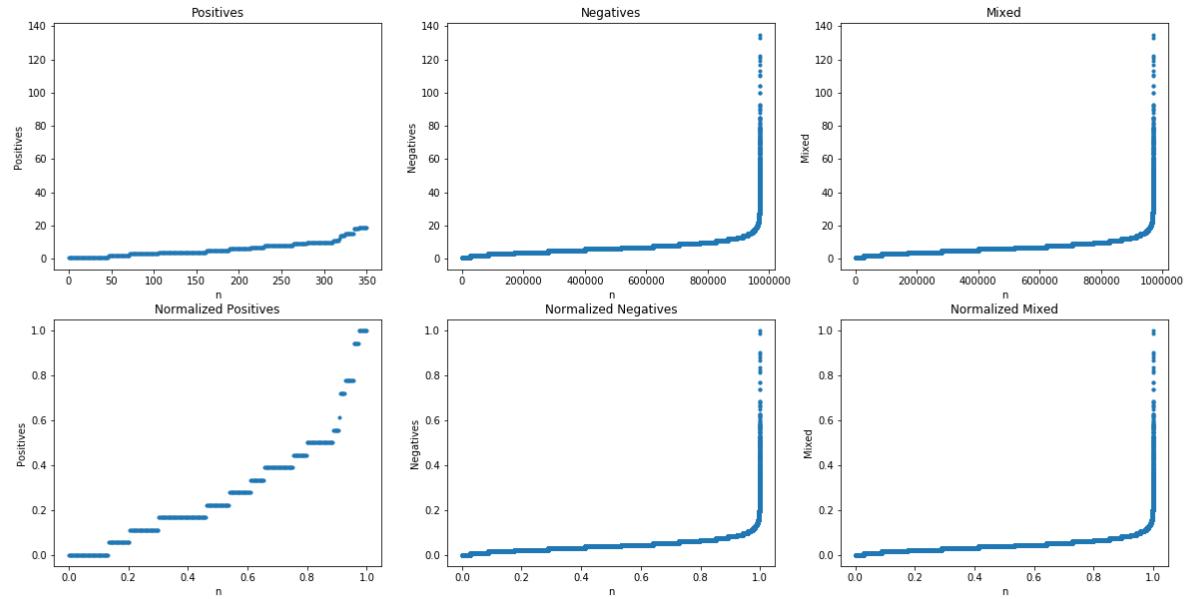


Figura 2.28: Values of metric commonVar

2.16 dbVARCount

2.16.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.20940038672579409 \quad \text{loc} = -1.1962983066939984e-30 \quad \text{scale} = 1.2347090894162929$$

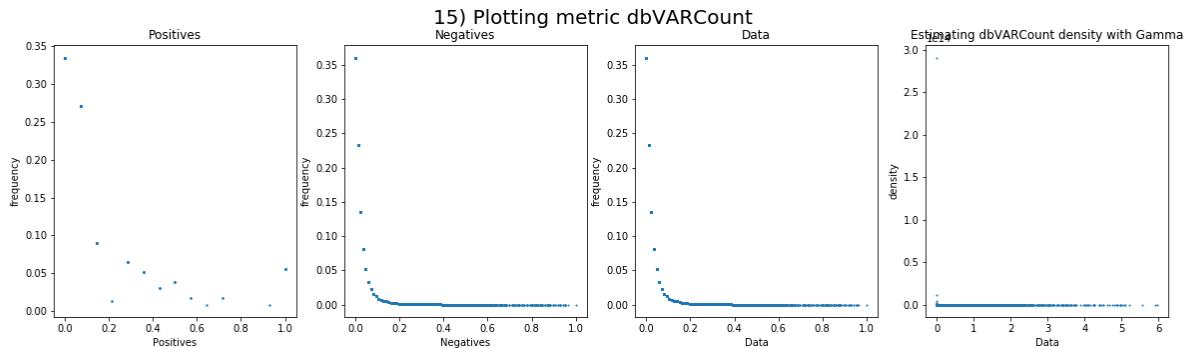


Figura 2.29: Sampling distribution of metric dbVARCount

2.16.2 Metric values

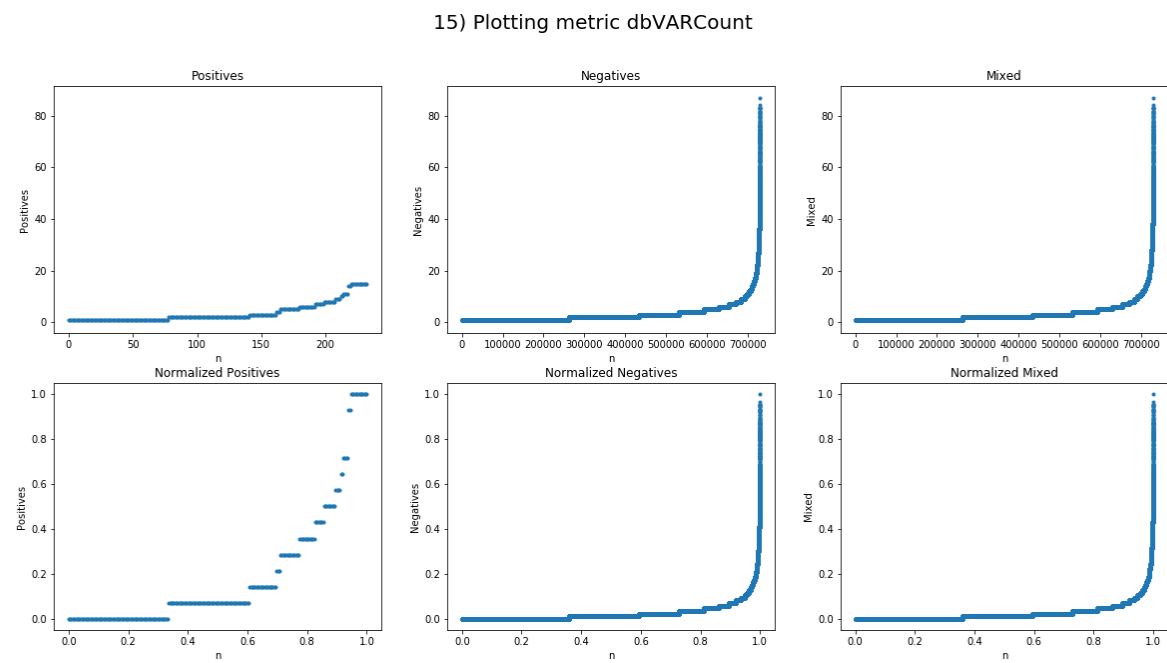


Figura 2.30: Values of metric dbVARCount

2.17 fantom5Perm

2.17.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.06895533706017208 \quad \text{loc} = -3.220296247423778e-30 \quad \text{scale} = 1.2605014923175824$$

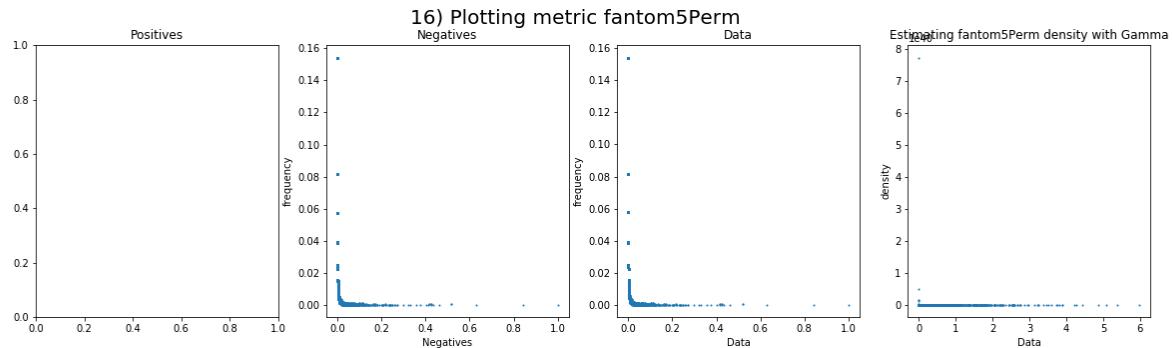


Figura 2.31: Sampling distribution of metric fantom5Perm

2.17.2 Metric values

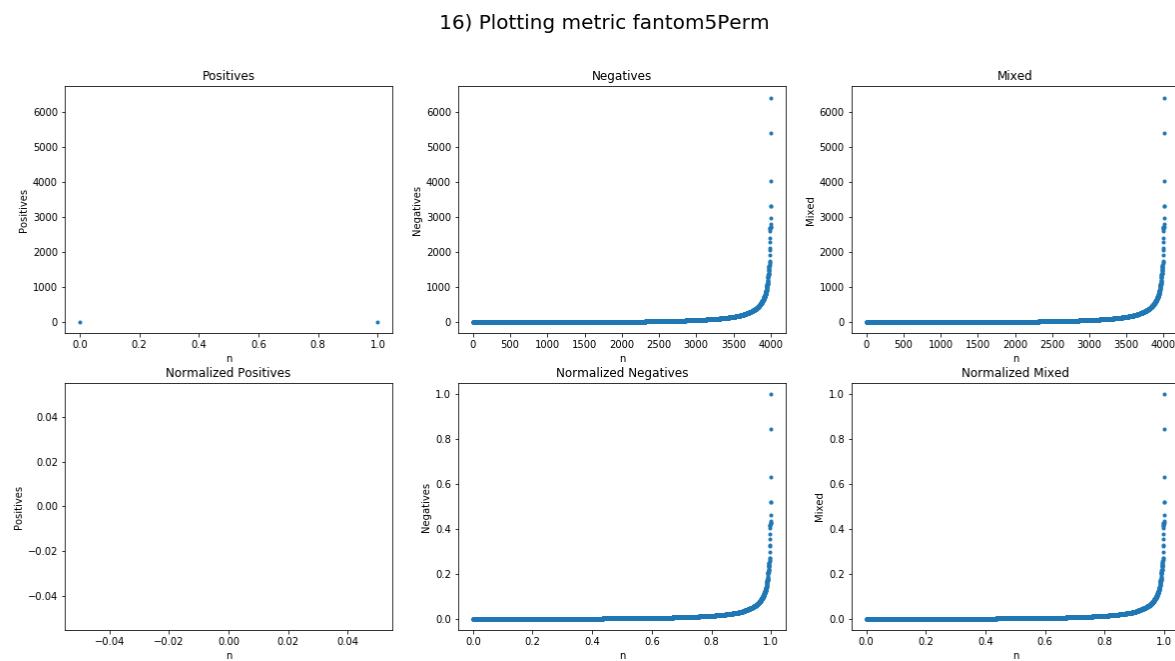


Figura 2.32: Values of metric fantom5Perm

2.18 fantom5Robust

2.18.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.08983952110680529 \quad \text{loc} = -3.220296247423778e-30 \quad \text{scale} = 1.2605014923175824$$

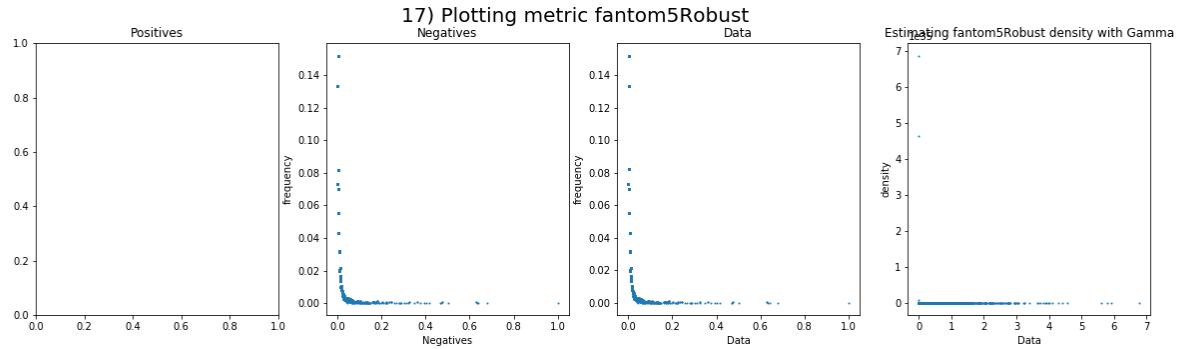


Figura 2.33: Sampling distribution of metric fantom5Robust

2.18.2 Metric values

17) Plotting metric fantom5Robust

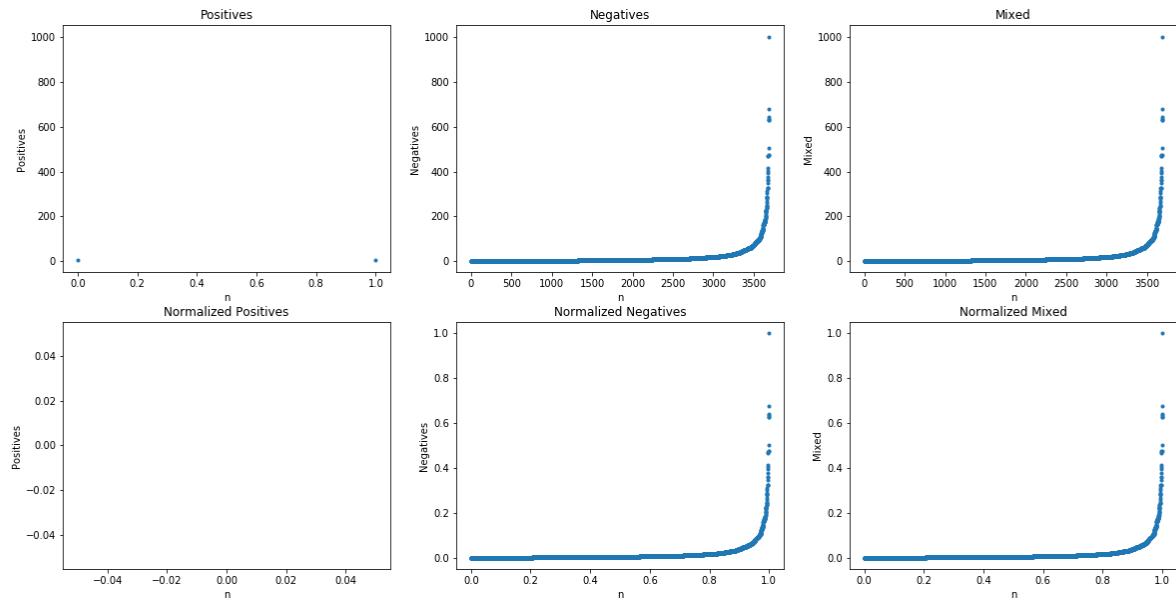


Figura 2.34: Values of metric fantom5Robust

2.19 fracRareCommon

2.19.1 Metric sample distribution

The data points seem to follow an **Beta** distribution with the following parameters:

$$\alpha = 2772.739504773501 \quad \beta = 14.986077009876375 \\ \text{loc} = -69.93503912437342 \quad \text{scale} = 71.09741090721741$$

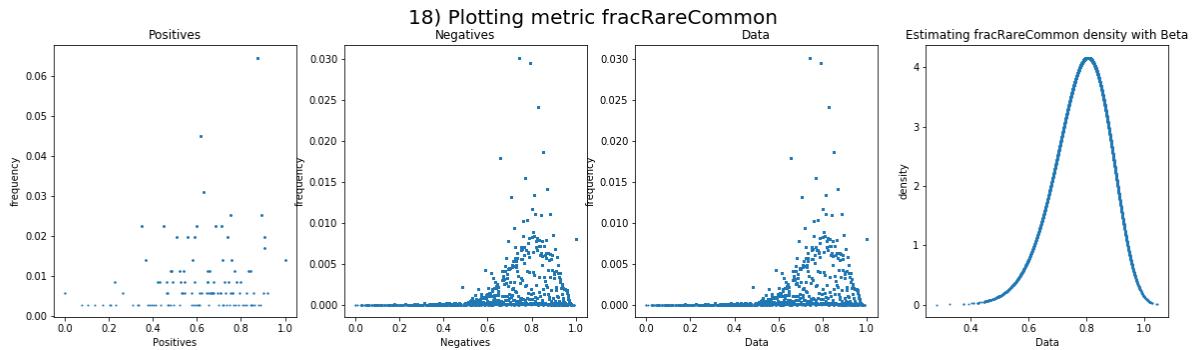


Figura 2.35: Sampling distribution of metric fracRareCommon

2.19.2 Metric values

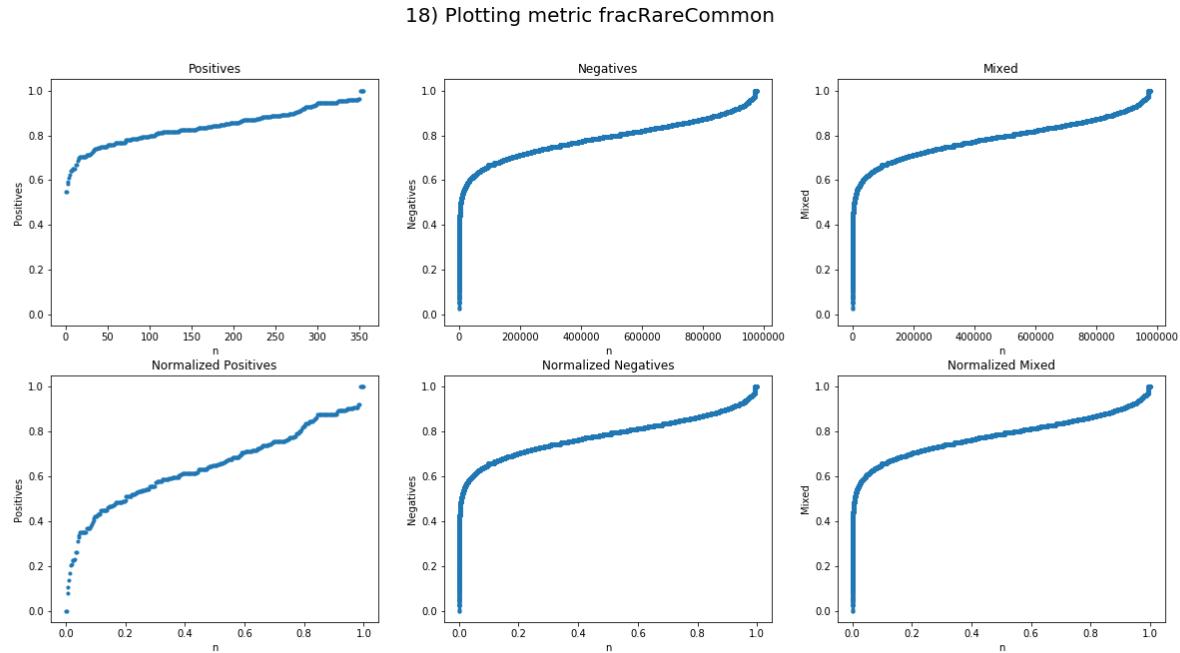


Figura 2.36: Values of metric fracRareCommon

2.20 mamPhastCons46way

2.20.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.3215099801387991 \quad \text{loc} = -6.260887365023215e - 31 \quad \text{scale} = 0.45230902834164866$$

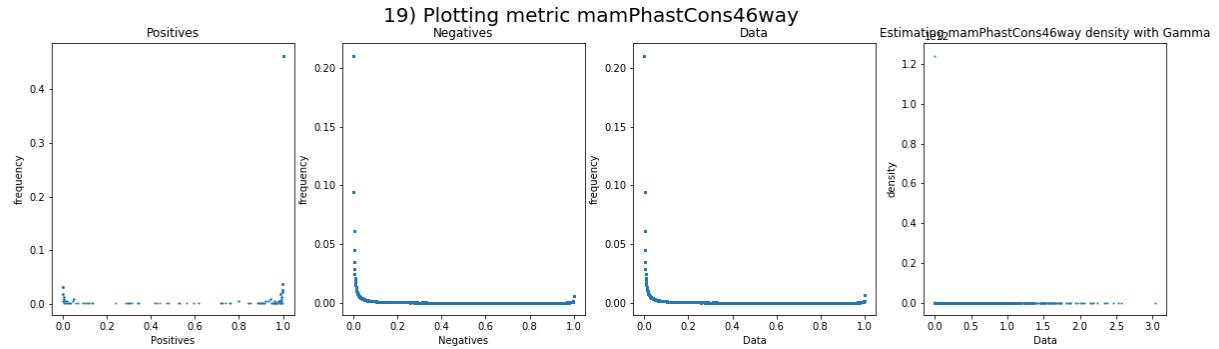


Figura 2.37: Sampling distribution of metric mamPhastCons46way

2.20.2 Metric values

19) Plotting metric mamPhastCons46way

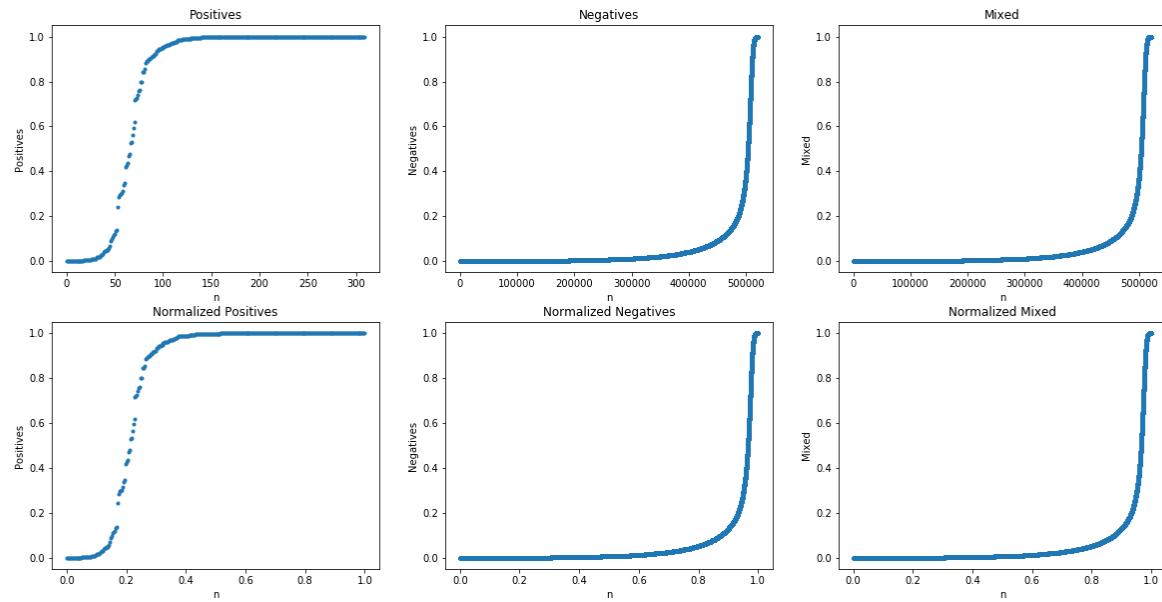


Figura 2.38: Values of metric mamPhastCons46way

2.21 mamPhyloP46way

2.21.1 Metric sample distribution

The data points seem to follow a **Gaussian** distribution with the following parameters:

$$\mathbb{E}(X) = 0.7032457913828309 \quad \text{Var}(X) = 0.07627203289198752$$

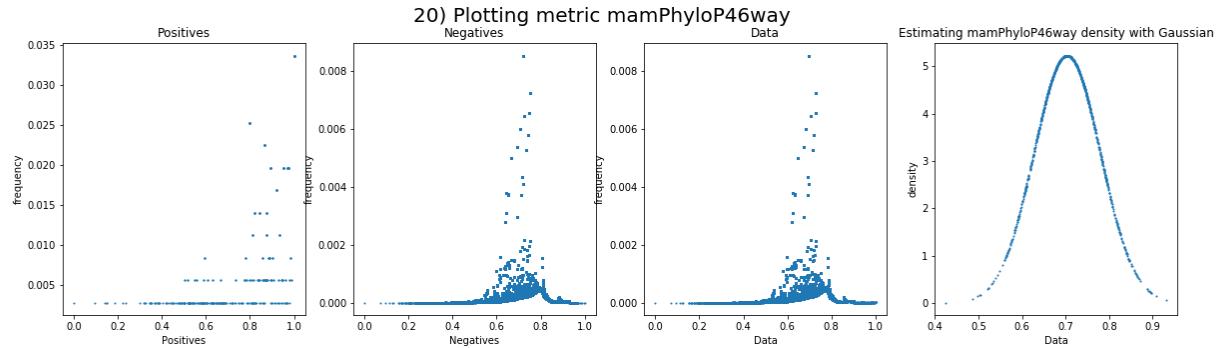


Figura 2.39: Sampling distribution of metric mamPhyloP46way

2.21.2 Metric values

20) Plotting metric mamPhyloP46way

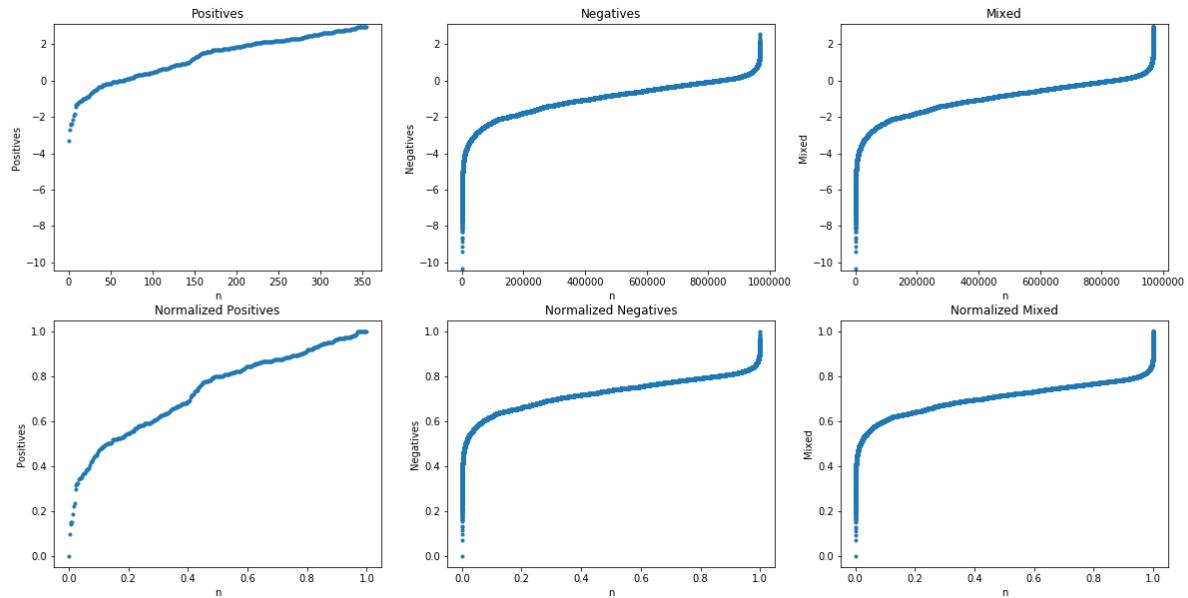


Figura 2.40: Values of metric mamPhyloP46way

2.22 numTFBSConserved

2.22.1 Metric sample distribution

The data points seem to follow a **exponential** distribution with the following parameters:

$$\mathbb{E}(X) = -4.600037873301623e-12 \quad \text{Var}(X) = 0.033419421646804975$$

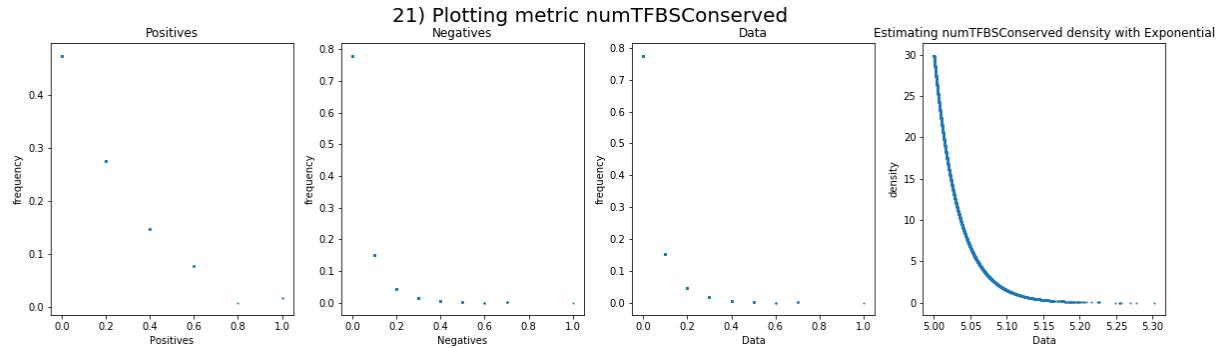


Figura 2.41: Sampling distribution of metric numTFBSConserved

2.22.2 Metric values

21) Plotting metric numTFBSConserved

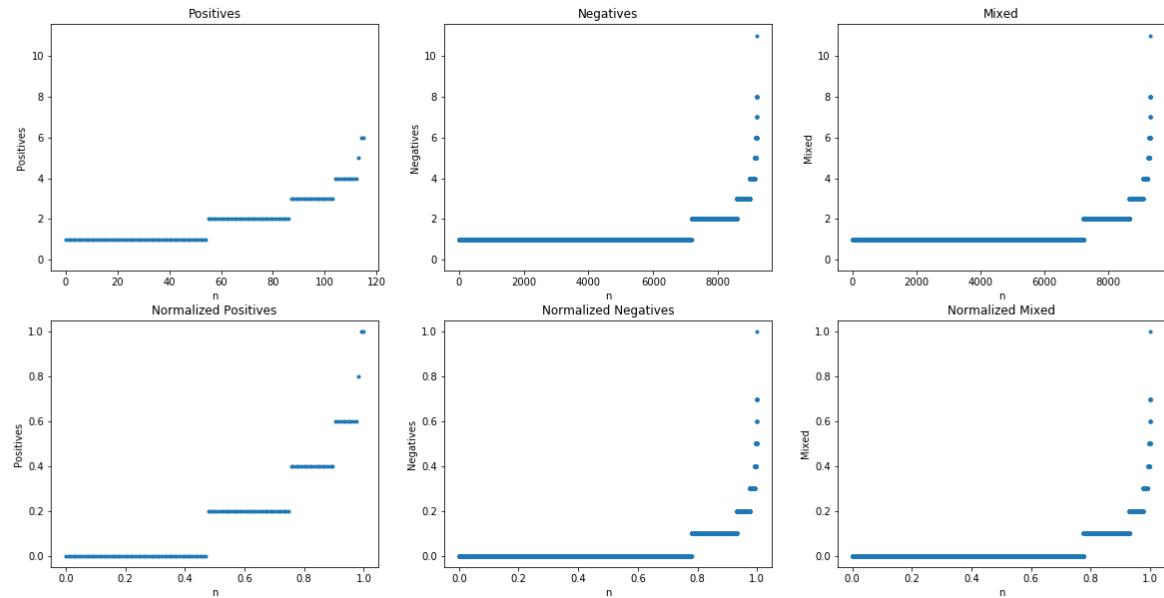


Figura 2.42: Values of metric numTFBSConserved

2.23 priPhastCons46way

2.23.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.2836383862597563 \quad \text{loc} = -1.8643137904859329e - 31 \quad \text{scale} = 0.37399746075497264$$

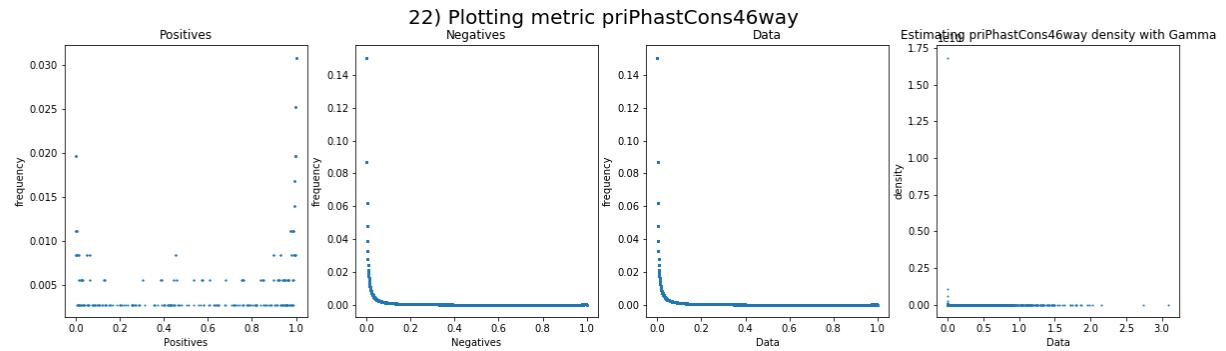


Figura 2.43: Sampling distribution of metric priPhastCons46way

2.23.2 Metric values

22) Plotting metric priPhastCons46way

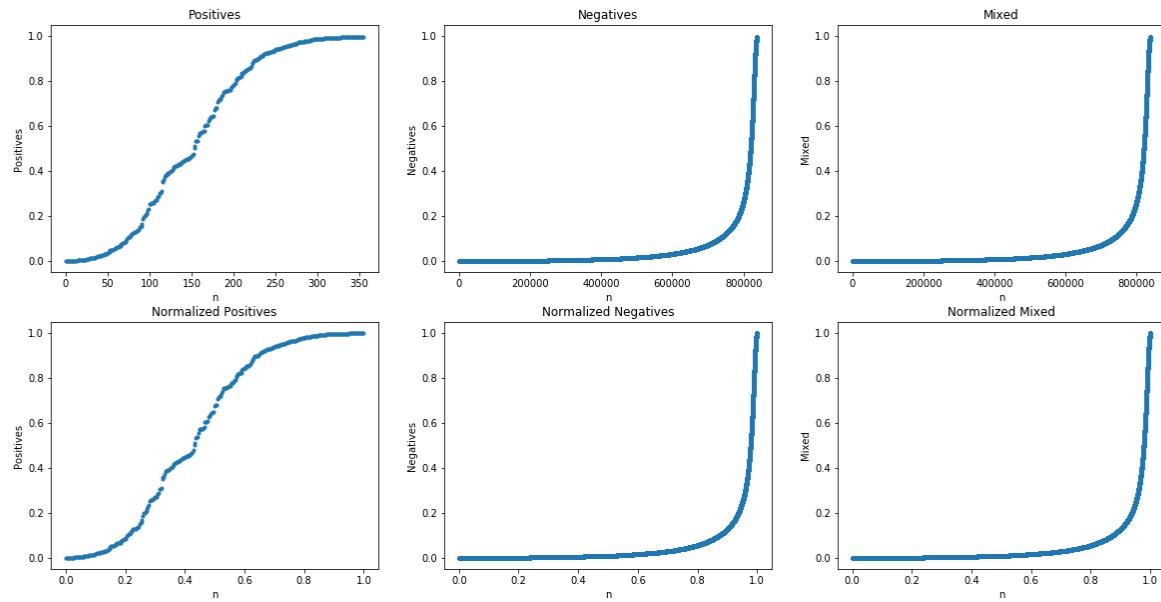


Figura 2.44: Values of metric priPhastCons46way

2.24 priPhyloP46way

2.24.1 Metric sample distribution

The data points seem to follow an **Beta** distribution with the following parameters:

$$\alpha = 2095270.7440875275 \quad \beta = 4.199025269606416 \\ \text{loc} = -103376.03746996864 \quad \text{scale} = 103377.03863437689$$

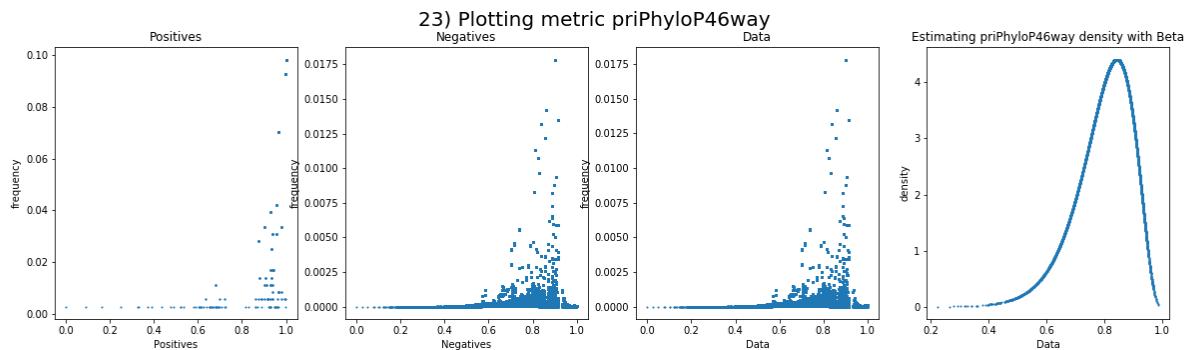


Figura 2.45: Sampling distribution of metric priPhyloP46way

2.24.2 Metric values

23) Plotting metric priPhyloP46way

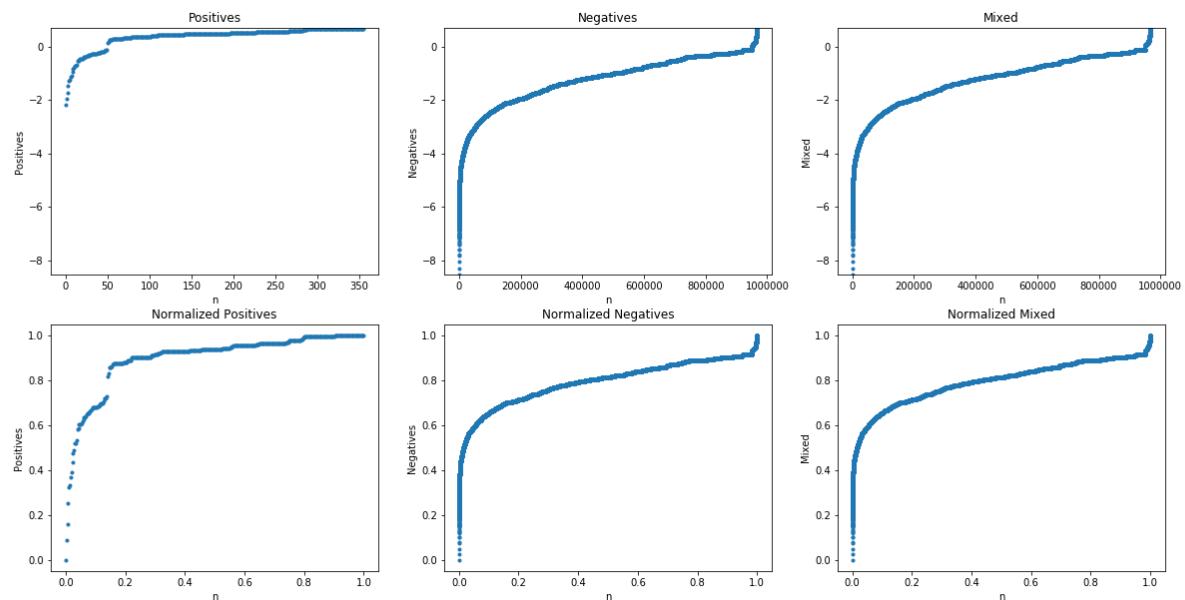


Figura 2.46: Values of metric priPhyloP46way

2.25 rareVar

2.25.1 Metric sample distribution

The data points seem to follow an **Beta** distribution with the following parameters:

$$\alpha = 14.148202647100376 \quad \beta = 7669045.025220526 \\ \text{loc} = -0.008523116473417407 \quad \text{scale} = 28973.953544984728$$

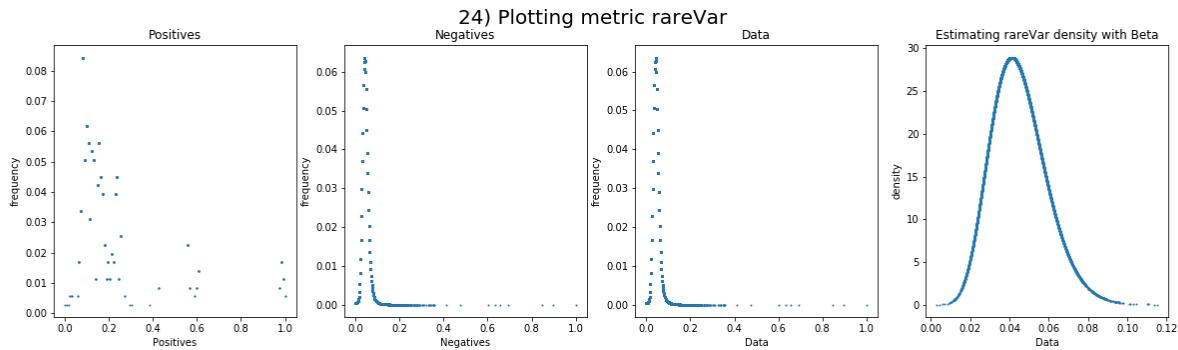


Figura 2.47: Sampling distribution of metric rareVar

2.25.2 Metric values

24) Plotting metric rareVar

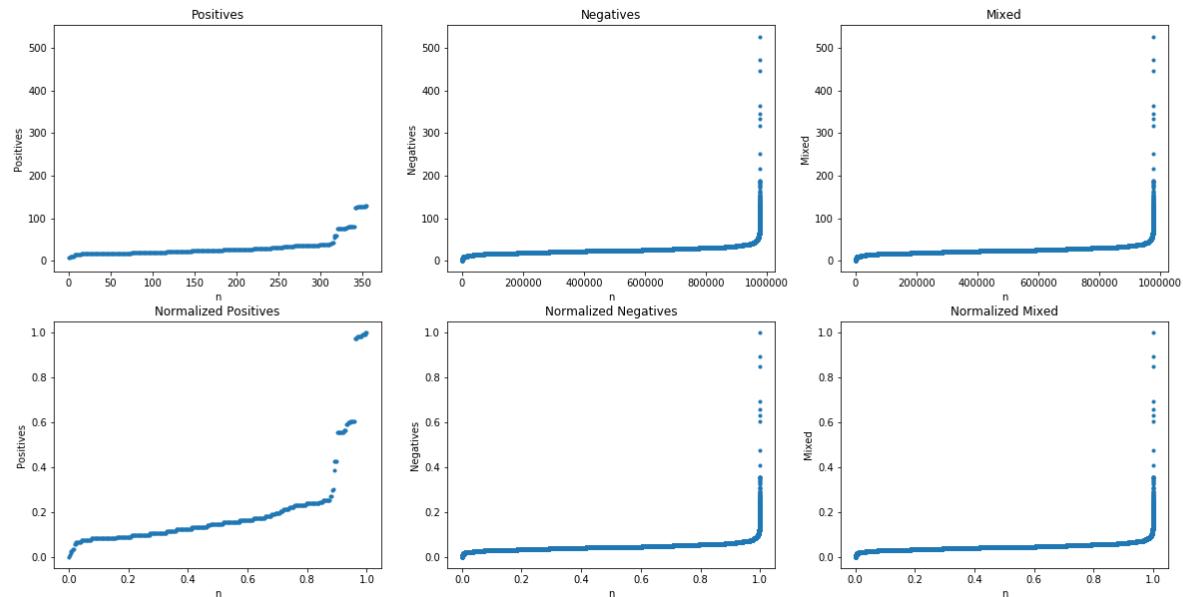


Figura 2.48: Values of metric rareVar

2.26 verPhastCons46way

2.26.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

$$\alpha = 0.4378982063415524 \quad \text{loc} = -2.5307968883256733e - 31 \quad \text{scale} = 0.43138079305533483$$

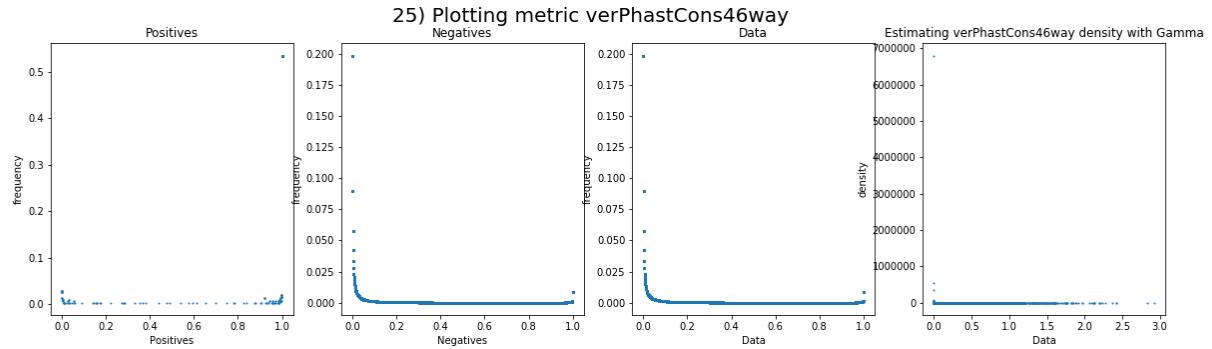


Figura 2.49: Sampling distribution of metric verPhastCons46way

2.26.2 Metric values

25) Plotting metric verPhastCons46way

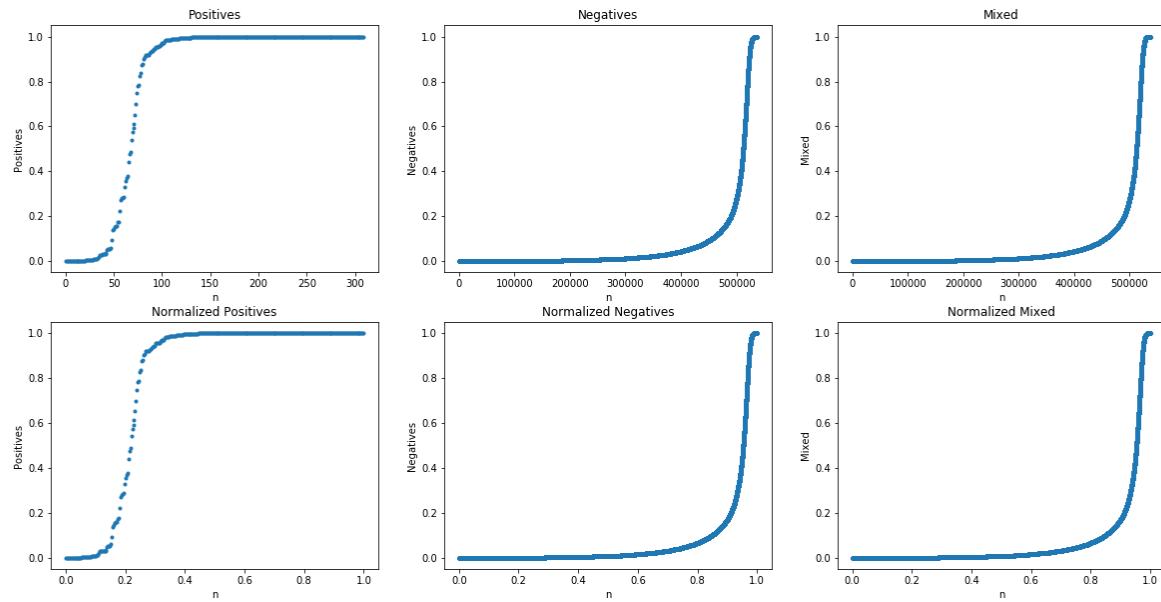


Figura 2.50: Values of metric verPhastCons46way

2.27 verPhyloP46way

2.27.1 Metric sample distribution

The data points seem to follow a **Gaussian** distribution with the following parameters:

$$\mathbb{E}(X) = 0.5723779382558164 \quad \text{Var}(X) = 0.0662715947139185$$

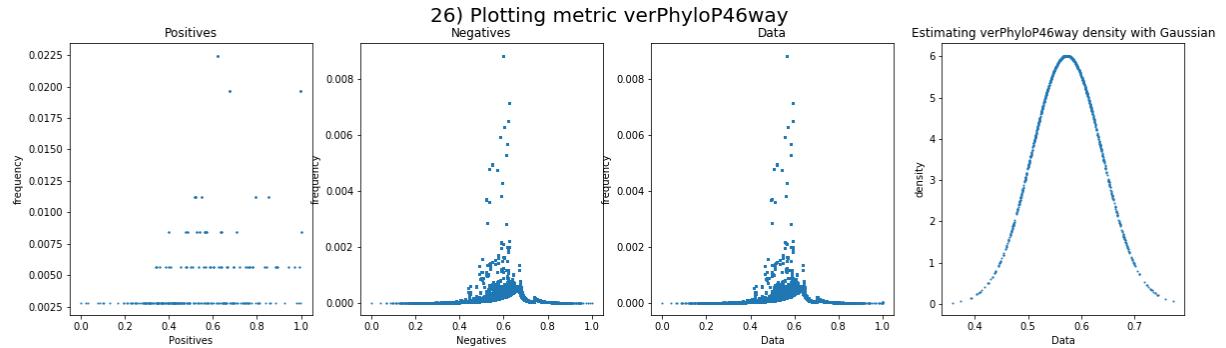


Figura 2.51: Sampling distribution of metric verPhyloP46way

2.27.2 Metric values

26) Plotting metric verPhyloP46way

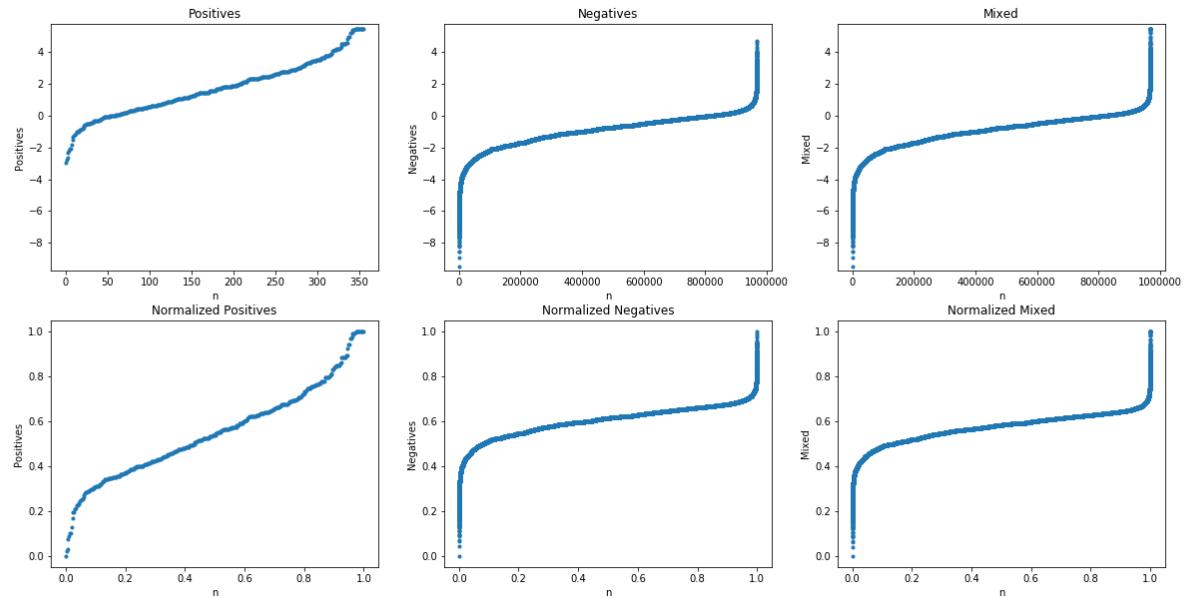


Figura 2.52: Values of metric verPhyloP46way

3

Metric distribution summary

The metrics seem to follow these sample distributions:

Metric	Distribution
CpGobsExp	Beta
CpGperCpG	Beta
CpGperGC	Gaussian
DGVCount	Gamma
DnaseClusteredHyp	Gamma
EncH3K27Ac	Gamma
GCContent	Gaussian
EncH3K4Me3	Gamma
ISCApath	Gamma
DnaseClusteredScore	Beta
EncH3K4Me1	Gamma
GerpRS	Gamma
GerpRSpv	Gamma
commonVar	Exponential Weibull
dbVARCount	Gamma
fantom5Perm	Gamma
fantom5Robust	Gamma
mamPhastCons46way	Gamma
priPhastCons46way	Gamma
rareVar	Beta
verPhastCons46way	Gamma
numTFBSConserved	Exponential
fracRareCommon	Beta
priPhyloP46way	Beta
verPhyloP46way	Gaussian
mamPhyloP46way	Gaussian

Tabella 3.1: Metrics and their distribution

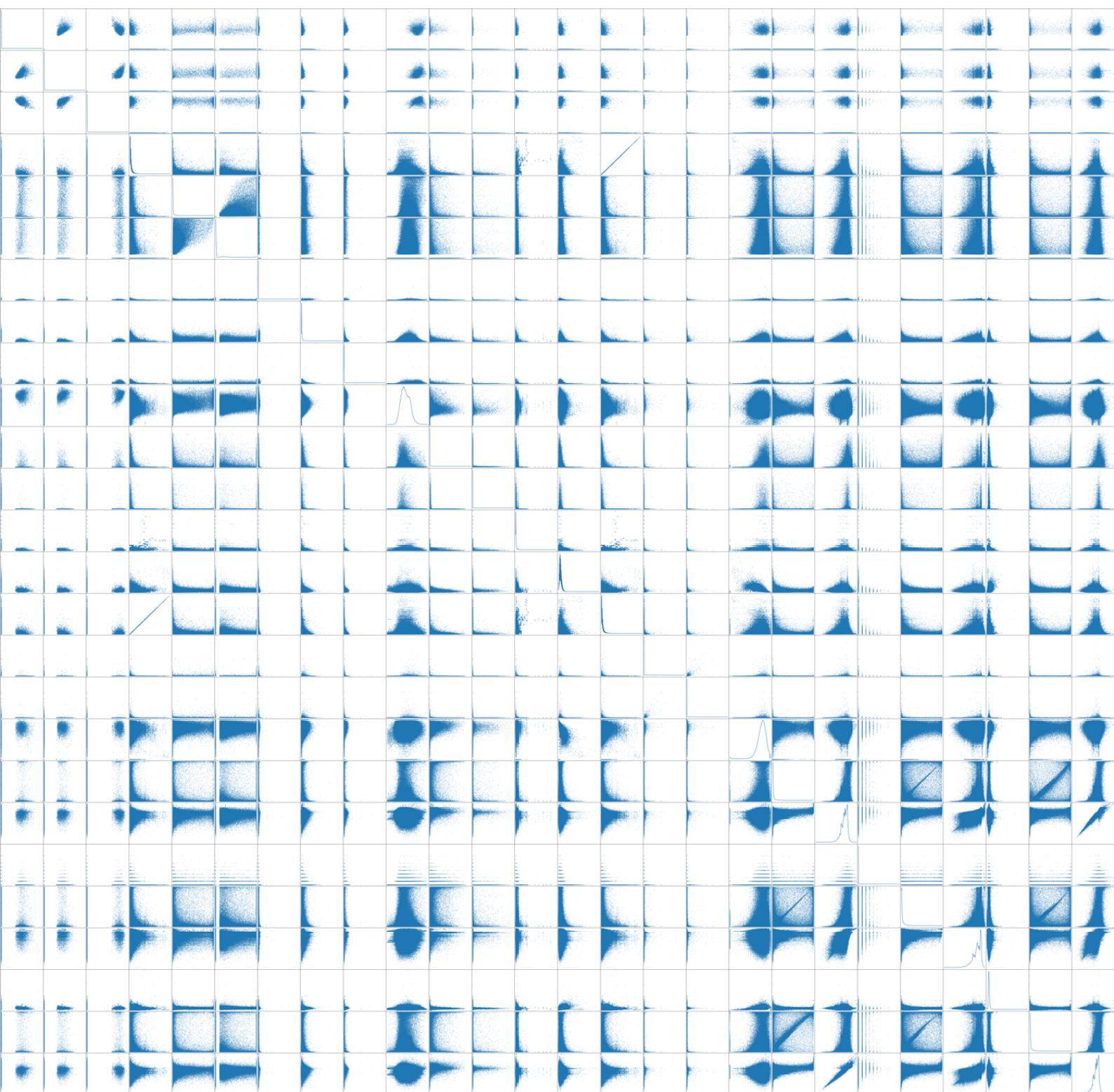
4

Scatter plot

We now proceed to draw a scatter plot trying to identify eventual data correlations.

4.1 Scatter plot

A scatter plot with higher resolution is available in the project repository.



4.2 Identified data correlations

Data correlations seem to exist between:

4.2.1 dbVARCount and DGVCount

There is a strong correlation between these two metrics:

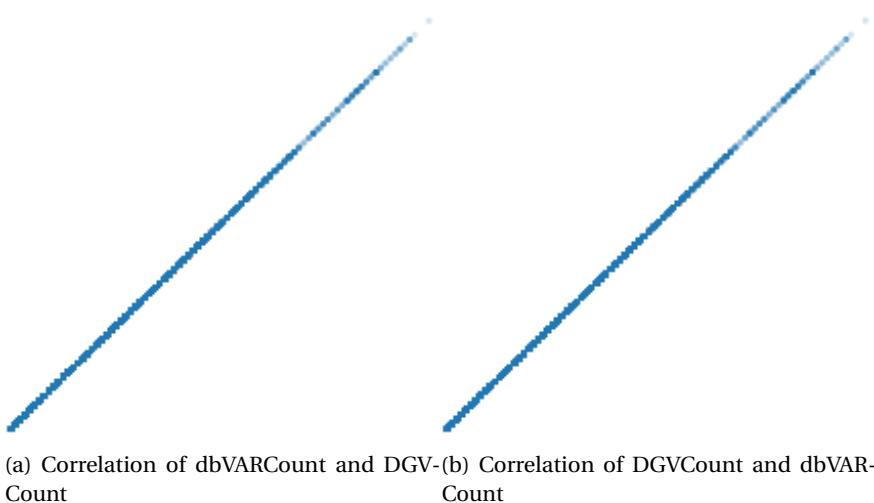


Figura 4.1: The two metrics DGVCount and dbVARCount are strongly correlated

The two metrics seem **highly** correlated, if not the **same metric**. This means that one of the two could be removed from the dataset, as it does not add any useful information.

4.2.2 mamPhyloP46way and verPhyloP46way

There is some correlation between these two metrics:

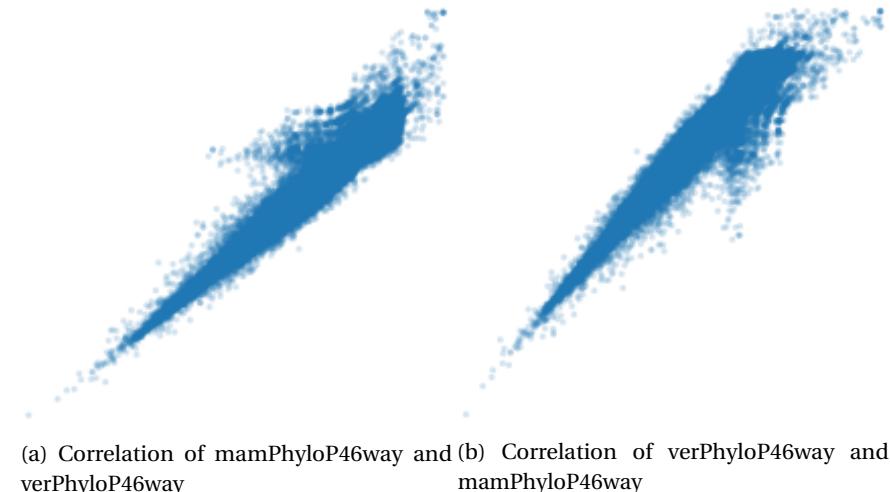
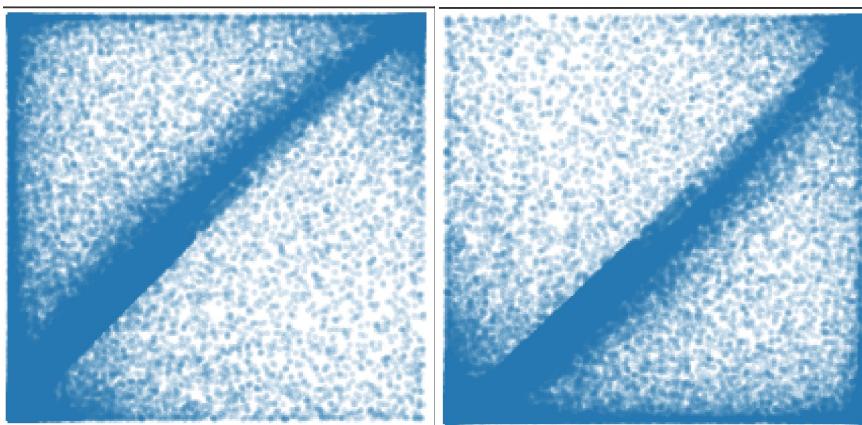


Figura 4.2: The two metrics verPhyloP46way and mamPhyloP46way are softly correlated

The two metrics seem **slightly** correlated, but not enough to consider removing one of the two.

4.2.3 mamPhastCons46way and verPhastCons46way

There is some correlation between these two metrics:



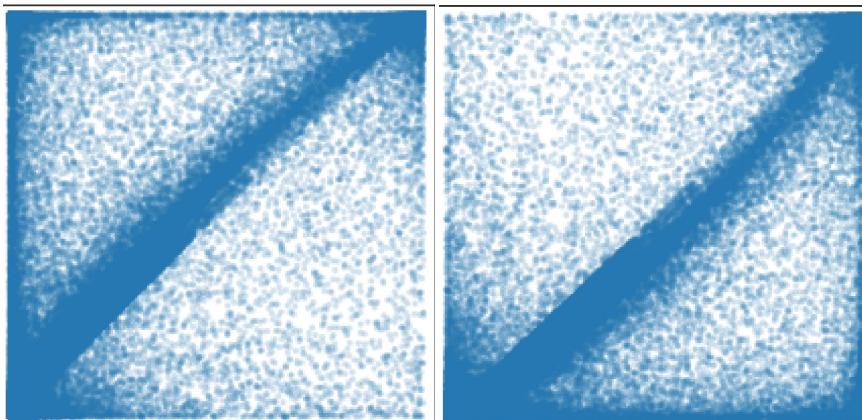
(a) Correlation of mamPhyloP46way and verPhyloP46way (b) Correlation of mamPhastCons46way and verPhastCons46way

Figura 4.3: The two metrics mamPhastCons46way and verPhastCons46way are softly correlated

The two metrics seem **slightly** correlated, but not enough to consider removing one of the two.

4.2.4 mamPhastCons46way and verPhastCons46way

There is a some correlation between this two metrics:



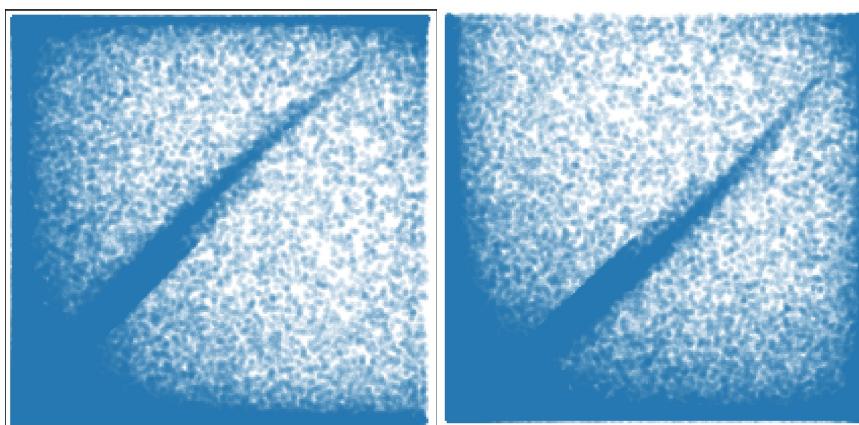
(a) Correlation of mamPhyloP46way and verPhyloP46way (b) Correlation of mamPhastCons46way and verPhastCons46way

Figura 4.4: The two metrics mamPhastCons46way and verPhastCons46way are softly correlated

The two metrics seem **slightly** correlated.

4.2.5 verPhastCons46way and priPhastCons46way

There is a some correlation between this two metrics:



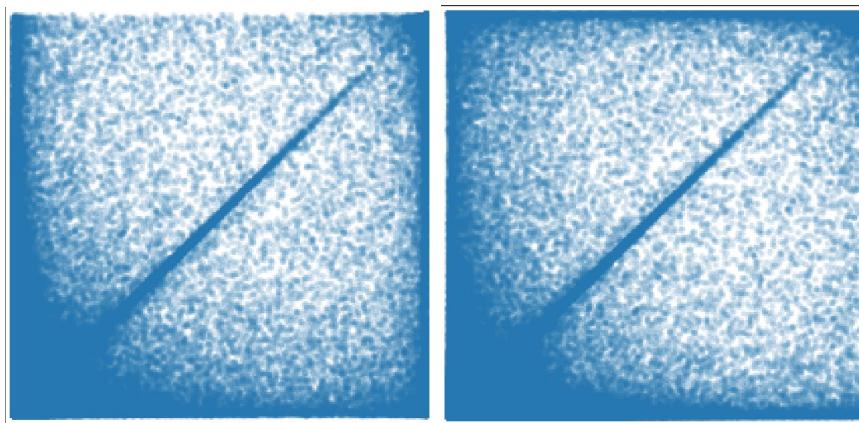
(a) Correlation of priPhastCons46way and
(b) Correlation of verPhastCons46way and
verPhastCons46way priPhastCons46way

Figura 4.5: The two metrics verPhastCons46way and priPhastCons46way are softly correlated

The two metrics seem **slightly** correlated.

4.2.6 priPhastCons46way and mamPhastCons46way

There is some correlation between these two metrics:



(a) Correlation of mamPhastCons46way and
(b) Correlation of priPhastCons46way and
priPhastCons46way mamPhastCons46way

Figura 4.6: The two metrics priPhastCons46way and mamPhastCons46way are softly correlated

The two metrics seem **slightly** correlated.

Parte II

Theory

5

Input modelling

5.1 Input values

The values used for each metric are the 3 following:

5.1.1 Normalized metric

Clearly one of the important metrics is the metric itself, that will be normalized to allow for input in $[-1, 1]$ range:

$$m' = 2 \cdot \frac{\text{metric} - \min \{\text{metric values}\}}{\max \{\text{metric values}\} - \min \{\text{metric values}\}} - 1$$

Figura 5.1: Input normalization to $[-1, 1]$ range

5.1.2 Probability

Another value we will be using in the input layer of the network is the probability of the metric value, modelled from the estimated sampling distribution of the normalized metric:

If M is the estimated metric distribution cumulative distribution function (CDF), m is the value assumed by the metric in the given data point and ϵ is a range, we can model the **probability** in the given value as follows:

$$\mathbb{P}(m' - \epsilon \leq X \leq m' + \epsilon) = M(m' + \epsilon) - M(m' - \epsilon)$$

Figura 5.2: Probability

5.1.3 Deviation from mean

The third and last input value is the deviation from the normalized value and the mean:

$$\text{deviation}(m) = m' - \mathbb{E}(X)$$

Figura 5.3: Deviation from mean

5.2 Feet

The input layer is comprised of 25 (number of metrics, excluded the one recognized to be in strong correlation to another) *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, with a window of 3 neurons.

5.3 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.

5.4 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.

5.5 Oversampling and undersampling targets

Oversampling and undersampling target will be to have a training dataset with 1% of positives and 99% of negatives.

5.6 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.

6

Output modelling

The output layer of the neural network is modelled by **two** neurons, one representing the positive class and one the negative class, with a **sigmoid** as activation function.

7

Weight initialization

7.1 Weight distribution based on input distribution

Since input values are not from any particular distribution or hold properties such as $\mathbb{E}(X) = 0$ or $\text{Var}(X) = 1$ (in some metrics mean and variance are far from these values) they do not suggest to use any specific distribution.

7.2 Weight distribution based on activation functions and regularization layers

The codomain values from the activation functions, being SELU for most of the network, tend to hold the properties of $\mathbb{E}(X) = 0$ and $\text{Var}(X) = 1$ (<http://arxiv.org/abs/1706.02515>). These values are then regularized to penalize extreme weights that may appear when variance starts with a value significantly away from 1.

For these properties weight will be initialized by extracting them from a Gaussian with $\mathbb{E}(X) = 0$ and $\text{Var}(X) = 1$.

8

Locally connected dense layers

The first two layers will be locally connected dense layers, to exploit the positional information of the input values.

Other than the group of triples, input will be sorted by distribution kind so that the initial interpolations happen mostly with data from the same distribution family.

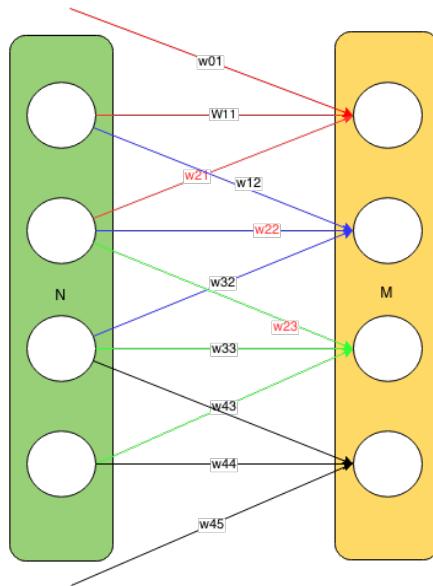


Figura 8.1: Locally connected layer

8.1 Activation function

We'll be using a **leaky relu** for the locally connected dense layers with $\alpha = 0.3$.

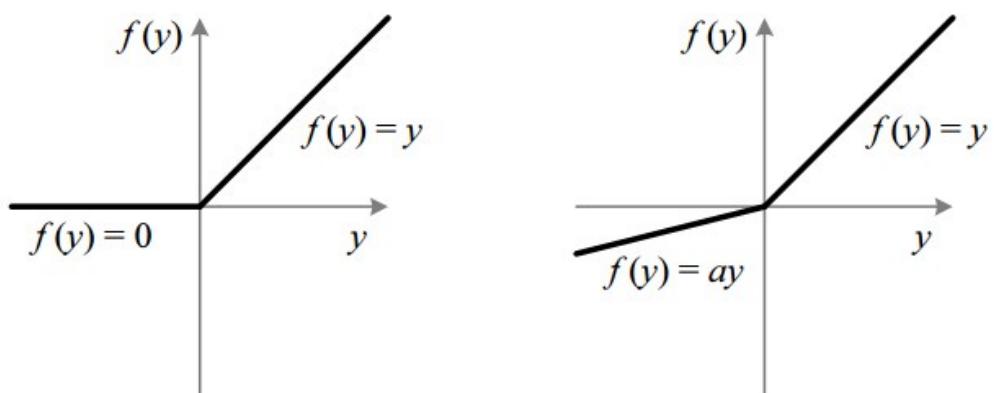


Figura 8.2: RELU and Leaky RELU

9

Dense layers

For the following hidden layers we will be using dense connected layers, with a piramidal structure (reducing the number of the neurons from 26 to 2).

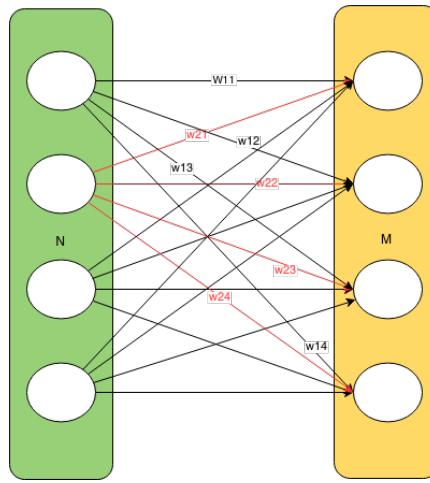


Figura 9.1: Dense connected layer

9.1 Activation function

We'll be using again **leaky relu** and experimenting with **SELU** for the dense layers:

$$\text{selu}(x) = \lambda \begin{cases} x & x > 0 \\ \alpha e^x - \alpha & x \leq 0 \end{cases}$$

Figura 9.2: SELU

9.2 Regularization

Regularization layers will be alternated to the dense layers to penalize weight extreme growth.

9.3 Drop out

In addition to regularization, also **drop out** of 10% of neurons per hidden layer will be applied.

10

References

LatexTools does not compile references at this time.