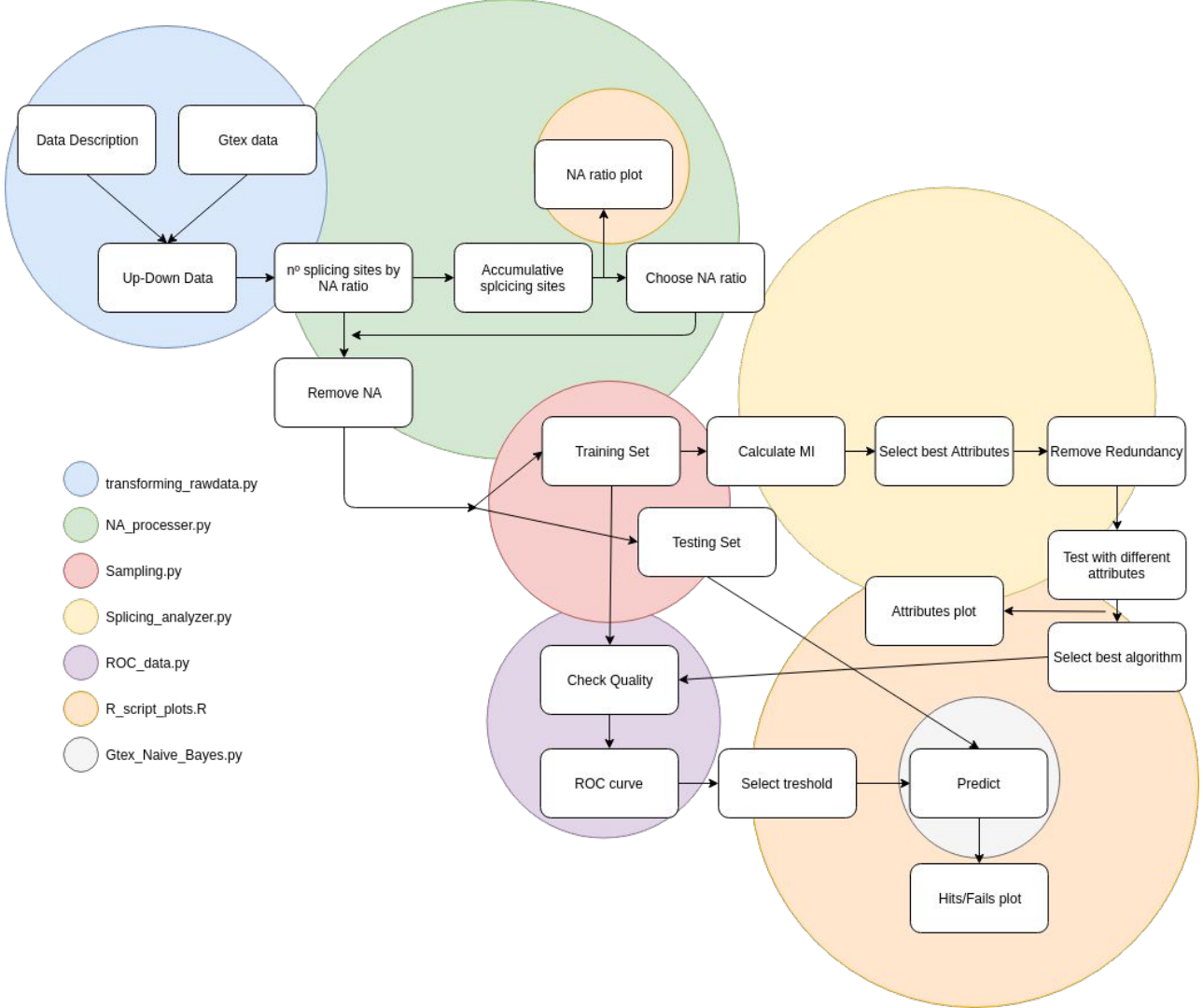
# Naive Bayes model to classify tissue types and brain regions using splicing patterns

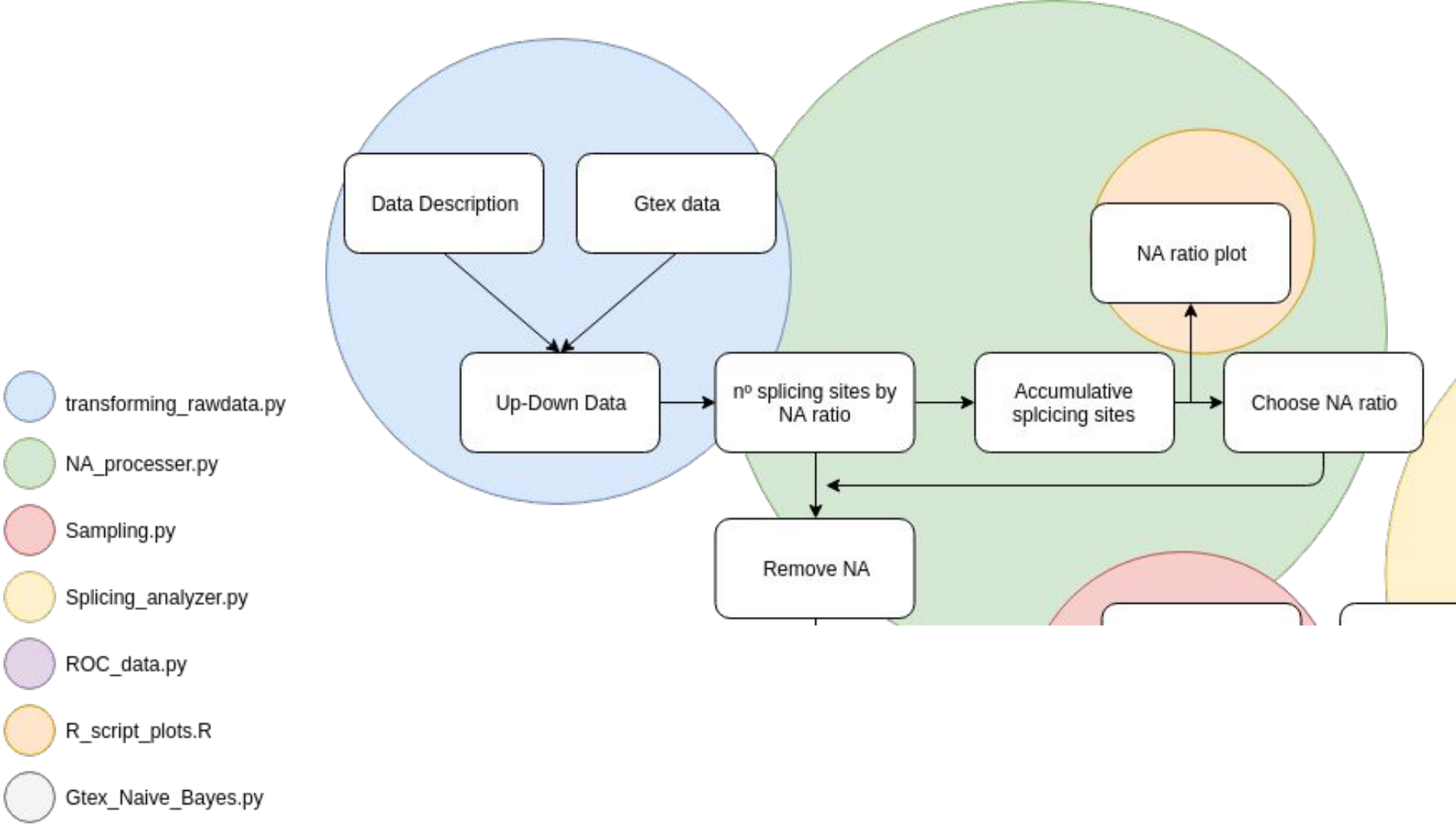
Oriol Gracia, Júlia Mir, Helena Rodríguez

# Introduction

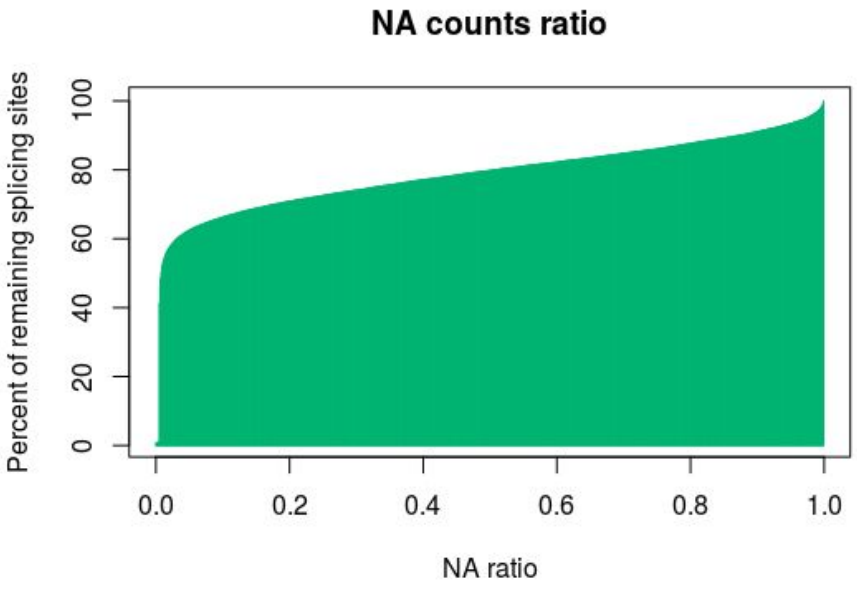
- Exon-skipping events.
- Proportion of RNA molecules that include the exon from all RNA molecules.
- Per tissue - per splicing site
- Brain samples
- Tissue samples
- Gtex format:
  - columns → tissues
  - rows → splicing sites

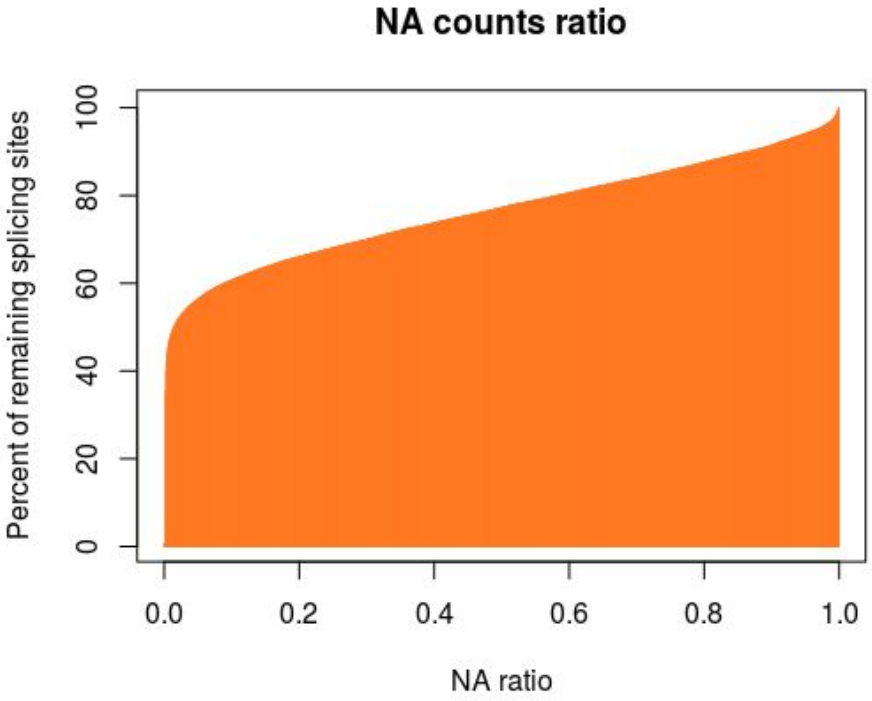
# Pipeline


Data Description
Gtex data
Up-Down Data
nº splicing sites by NA ratio
Accumulative splcicing sites
NA ratio plot
Choose NA ratio
Remove NA
Training Set
Testing Set
Calculate MI
Select best Attributes
Remove Redundancy
Test with different attributes
Attributes plot
Select best algorithm
Check Quality
ROC curve
Select treshold
Predict
Hits/Fails plot
transforming_rawdata.py
NA_processer.py
Sampling.py
Splicing_analyzer.py
ROC_data.py
R_script_plots.R
Gtex_Naive_Bayes.py


Data Description
Gtex data
Up-Down Data
nº splicing sites by NA ratio
Accumulative splcicing sites
Choose NA ratio
NA ratio plot
Remove NA
transforming_rawdata.py
NA_processer.py
Sampling.py
Splicing_analyzer.py
ROC_data.py
R_script_plots.R
Gtex_Naive_Bayes.py

# NA ratio




Brain

Other tissues

# Training and Testing sets

BRAIN

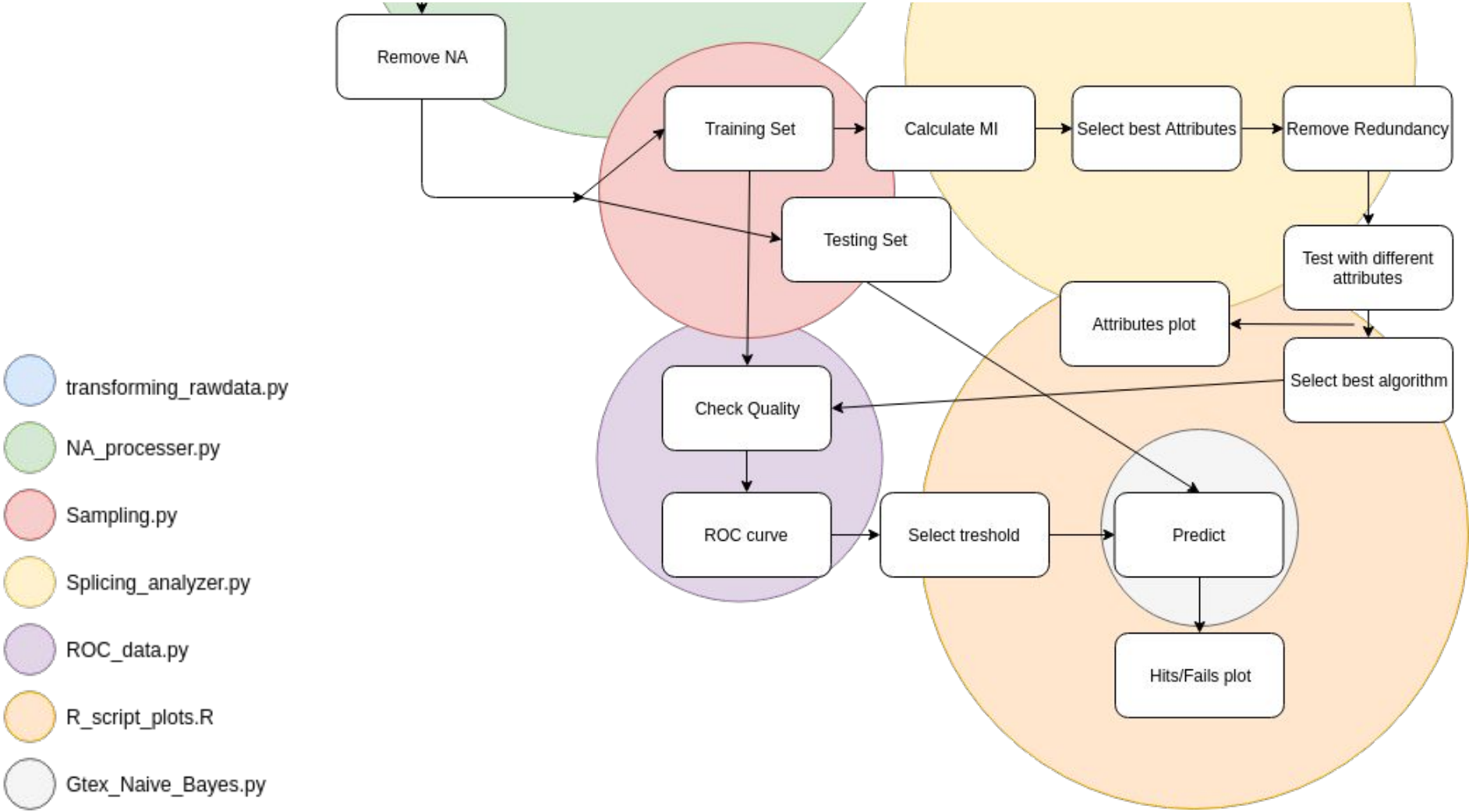
69 Amygdala
83 Anterior_Cingulate_Cortex_(Ba24)
109 Caudate_(Basal_Ganglia)
97 Cerebellar_Hemisphere
119 Cerebellum
105 Cortex
102 Frontal_Cortex_(Ba9)
84 Hippocampus
82 Hypothalamus
104 Nucleus_Accumbens_(Basal_Ganglia)
81 Putamen_(Basal_Ganglia)
60 Spinal_Cord_(Cervical_C-1)
57 Substantia_Nigra

55 → training
2 → testing

TISSUES

377 Heart
28 Kidney
110 Liver
288 Lung
396 Muscle
278 Nerve

80% → training
20% → testing


Remove NA
Training Set
Calculate MI
Select best Attributes
Remove Redundancy
Testing Set
Test with different attributes
Attributes plot
Select best algorithm
Check Quality
ROC curve
Select treshold
Predict
Hits/Fails plot
transforming_rawdata.py
NA_processer.py
Sampling.py
Splicing_analyzer.py
ROC_data.py
R_script_plots.R
Gtex_Naive_Bayes.py

# Calculate MI

**Information gain** $IG(S,A) = MI(S,A) = H(S) - H(S \mid A)$

$$H(S) = -\sum_{s=\{classes\}} P(s)\log_2 P(s)$$

$$H(S \mid A) = -\sum_{a=\{values\}} \sum_{s=\{classes\}} P(s,a)\log_2 \frac{P(s,a)}{P(a)}$$

$$H(S \mid A) = -\sum_{a=\{values\}} P(a) \sum_{s=\{classes\}} P(s \mid a)\log_2 P(s \mid a)$$

# Avoiding Redundancy

| | Tissue A | Tissue A | Tissue B | MI |
|---|---|---|---|---|
| splicing A | up | down | up | 0.02 |
| splicing B | up | up | up | 0.01 |

nº up
nº down

nº up
nº down

if nºup splicingA == nº up splicingB → +1 hit

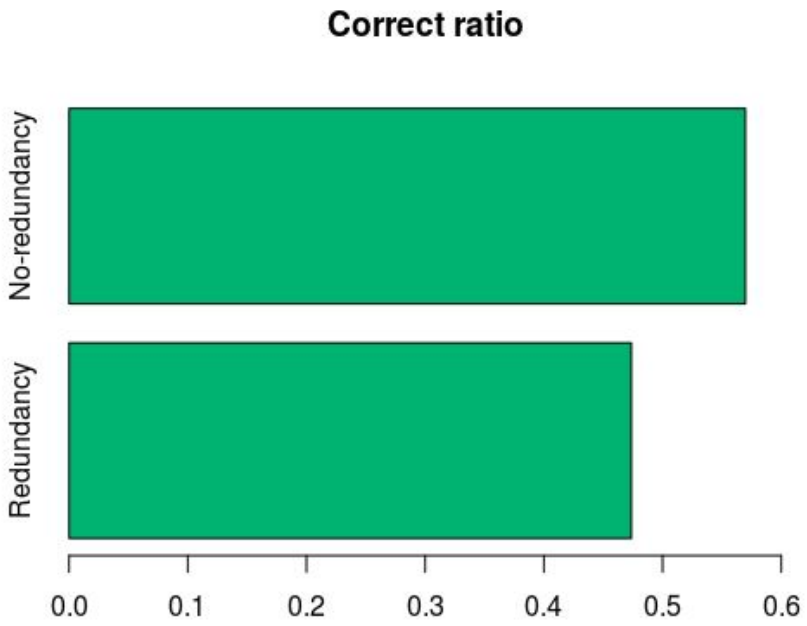
# Avoiding Redundancy

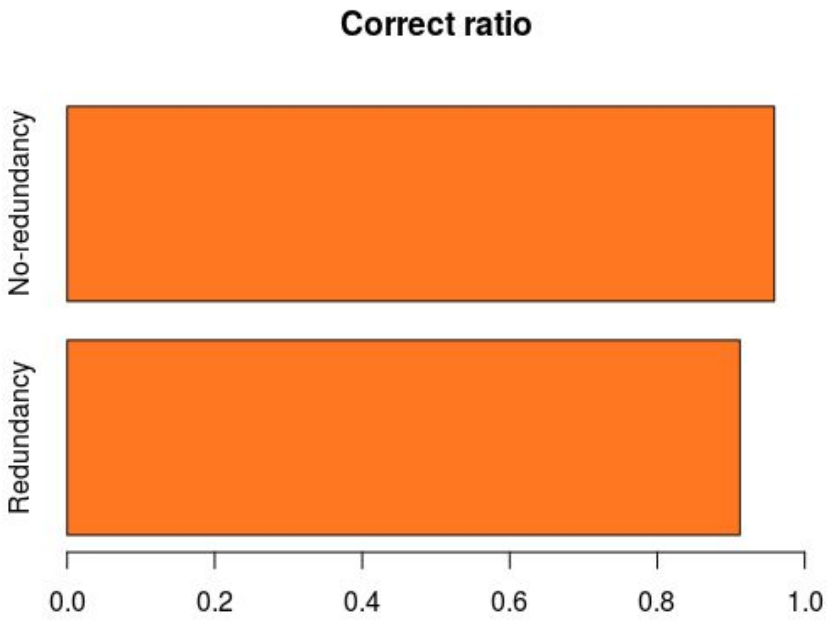
| | splicing A | splicing B | splicing C | splicing D |
|---|---|---|---|---|
| splicing A | | | | |
| splicing B | 3 | | | |
| splicing C | 1 | 1 | | |
| splicing D | 2 | 2 | 2 | |

MI A > MI B

| | splicing A | splicing C | splicing D |
|---|---|---|---|
| splicing A | | | |
| splicing C | 2 | | |
| splicing D | 1 | 2 | |

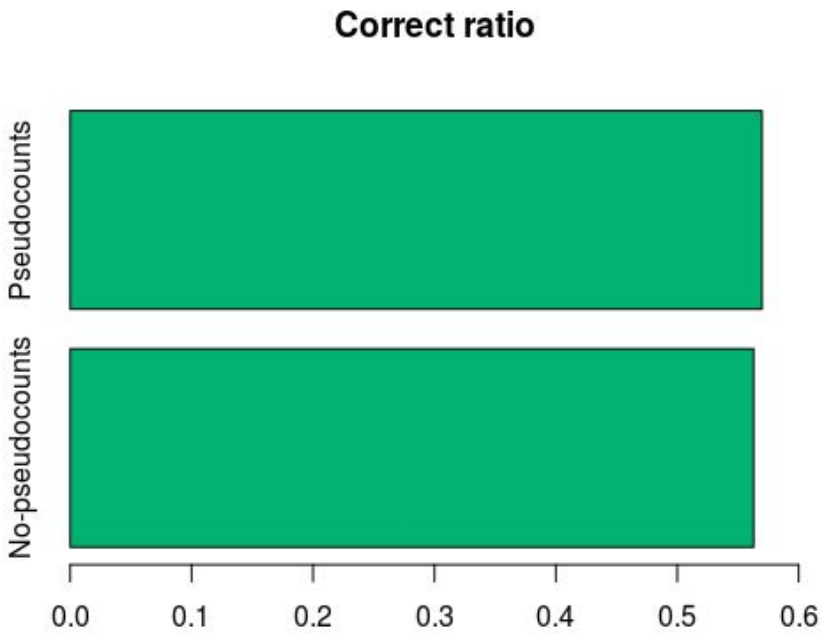
# Redundancy

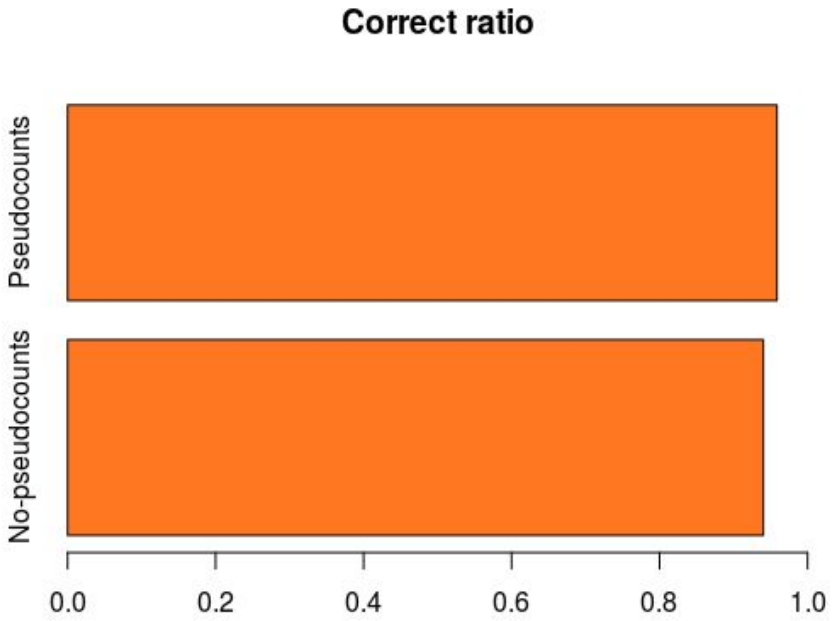



Brain Other tissues

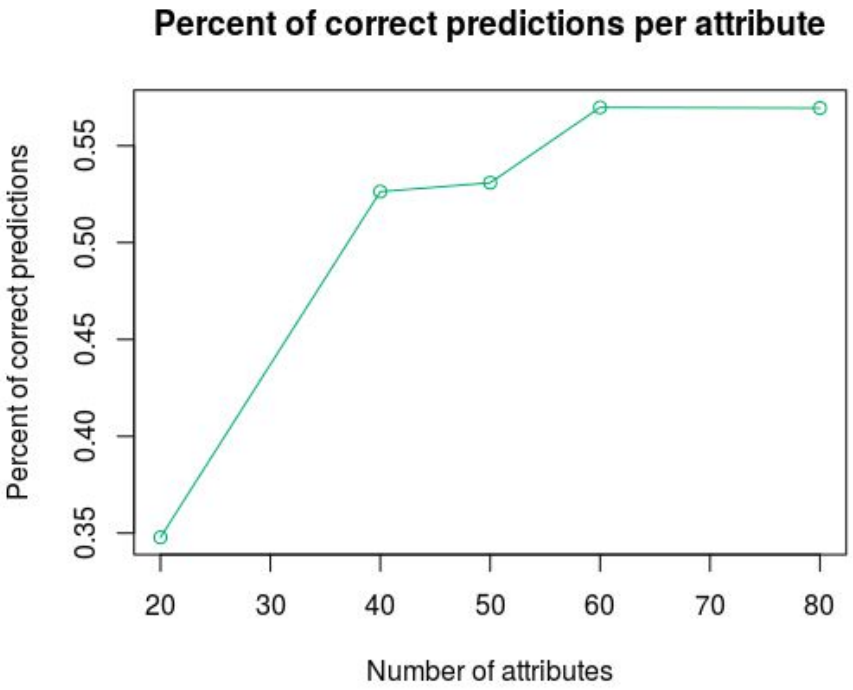
# Pseudocounts

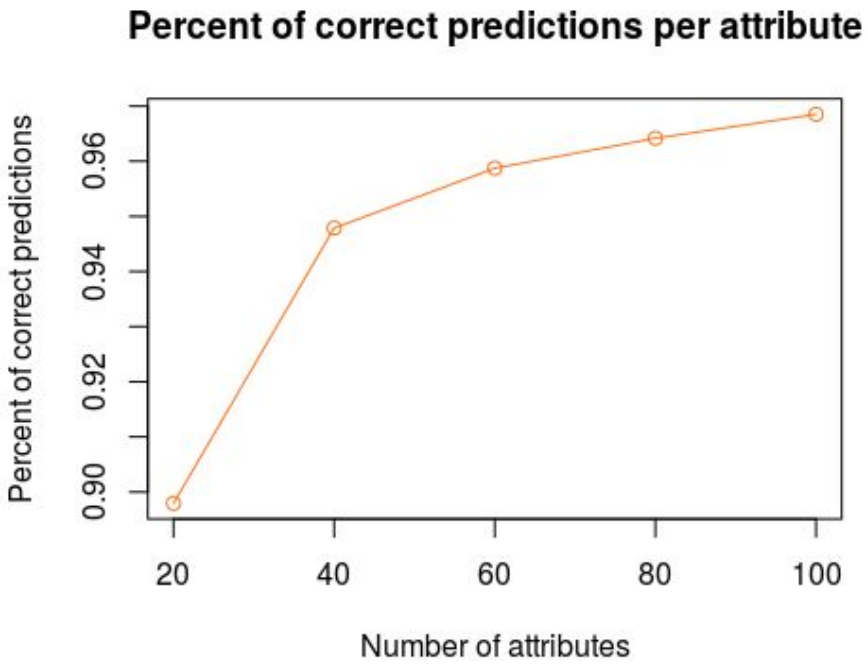


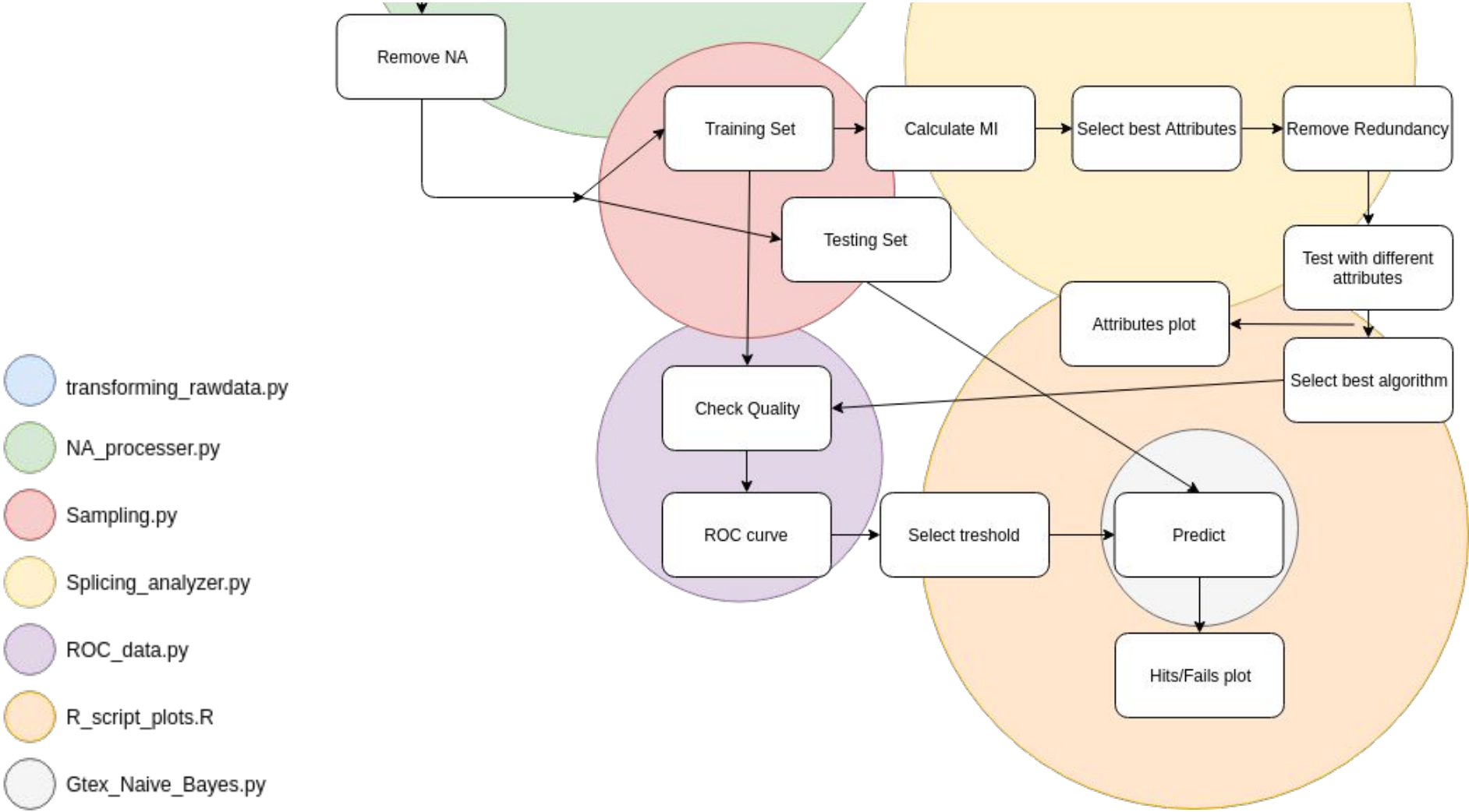

Brain Other tissues

# Correct predictions per attribute




Brain Other tissues


Remove NA
Training Set
Calculate MI
Select best Attributes
Remove Redundancy
Testing Set
Test with different attributes
Attributes plot
Select best algorithm
Check Quality
ROC curve
Select treshold
Predict
Hits/Fails plot
transforming_rawdata.py
NA_processer.py
Sampling.py
Splicing_analyzer.py
ROC_data.py
R_script_plots.R
Gtex_Naive_Bayes.py

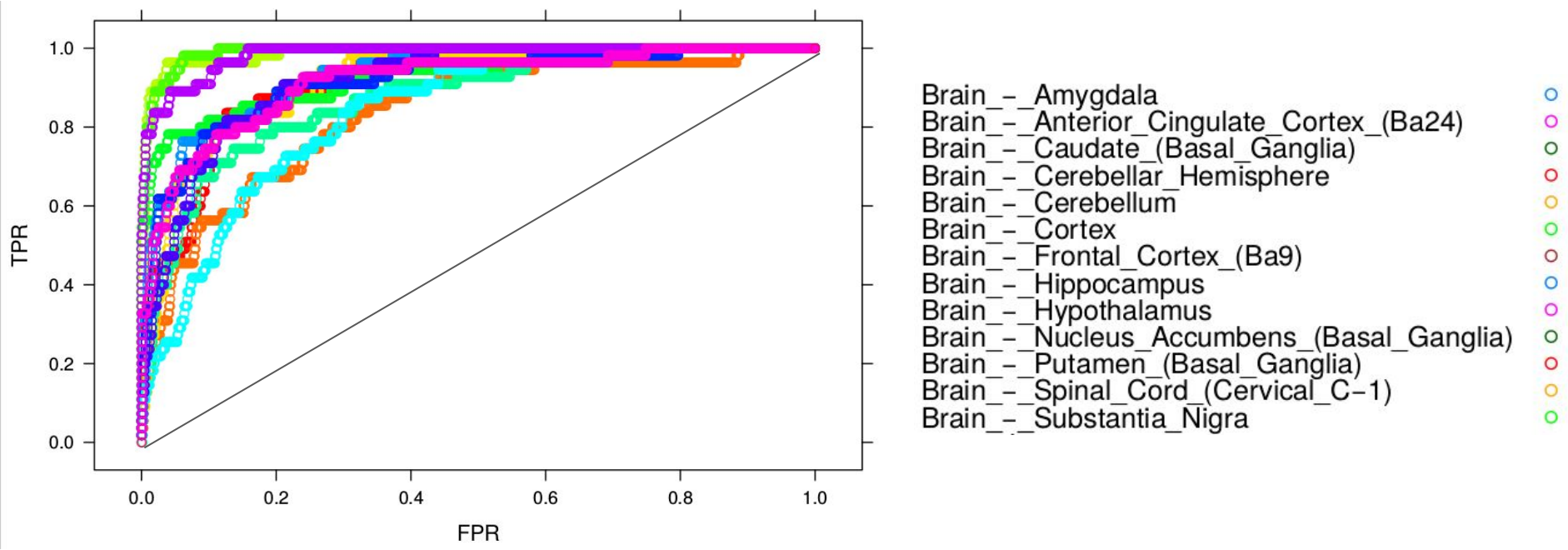
# Naive Bayes

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P(v_j) \prod_{r=1}^{n} P(a_r \mid v_j)$$

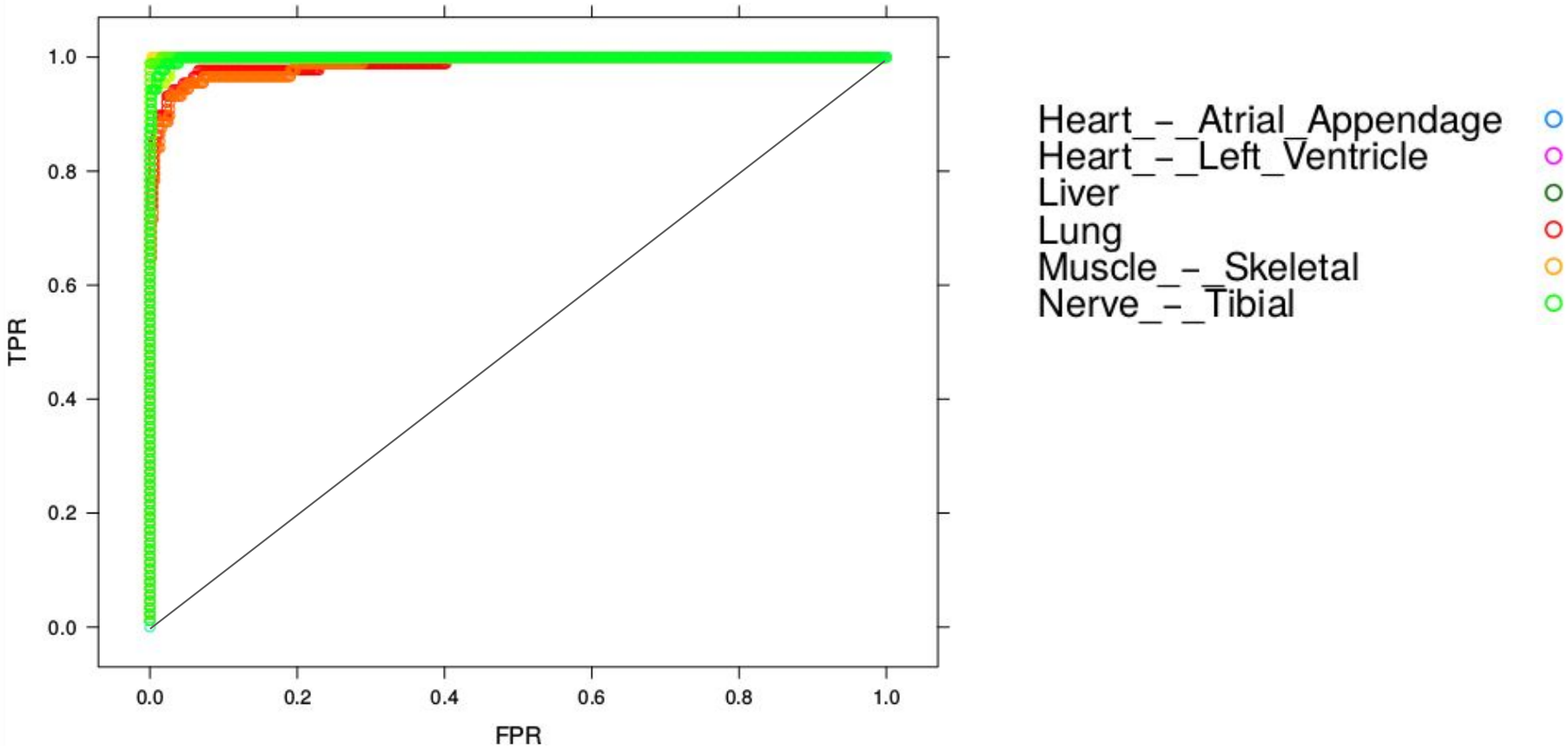
## Pseudocounts

$$P(a) = \frac{n_a}{n} \longrightarrow P(a) = \frac{n_a + 1}{n + m}$$

# ROC curves (brain)

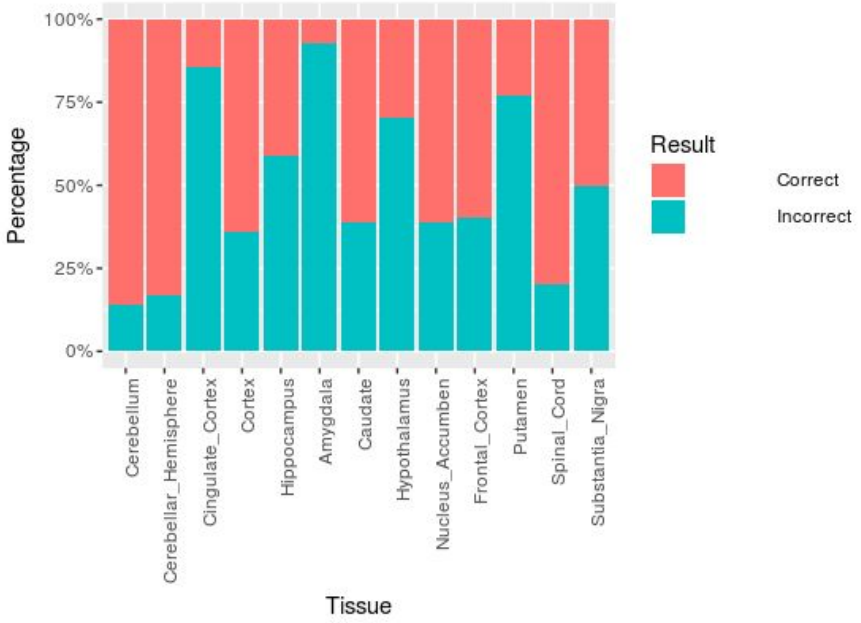


# ROC curves(other tissues)

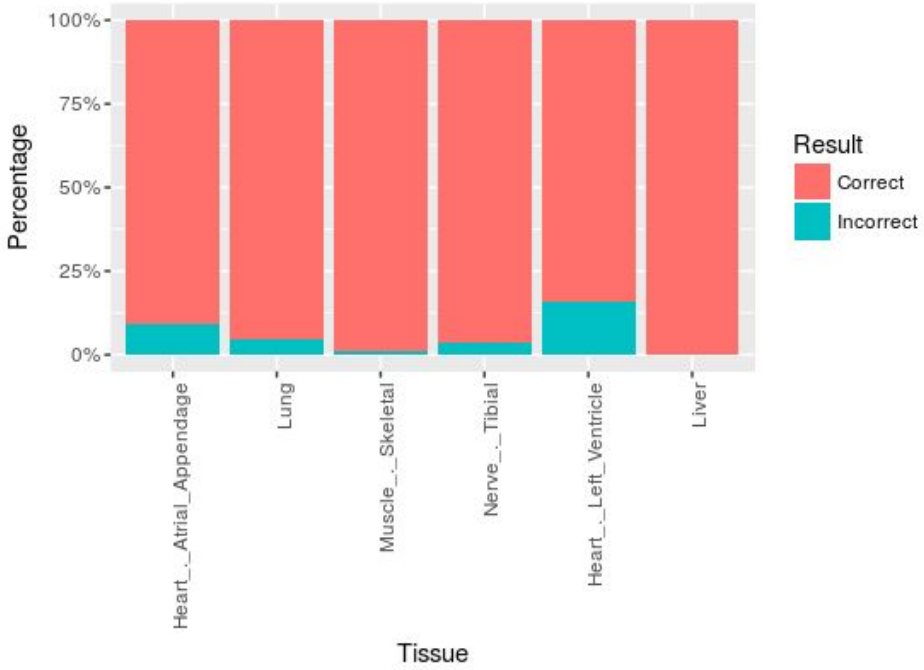


# Correct versus incorrect predictions

Brain




Other tissues



# Final Results

## BRAIN

- The Number of correct guesses are: 249
- The Number of incorrect guesses are: 188
- The correct ratio is 0.57

## OTHER TISSUES

- The Number of correct guesses are: 498
- The Number of incorrect guesses are: 30
- The correct ratio is 0.94

# Possible improvements

- Symmetrical uncertainty $$SU(S,A) = \frac{2MI(S,A)}{H(S)+H(A)}$$
- K-fold cross-validation
- Different approaches for pseudocounts
  - Prior estimate $$P(a) = \frac{n_a + mp}{n+m}$$
- Remove tissues with lower prediction accuracy: amygdala, cingulate cortex and putamen

# Thank you