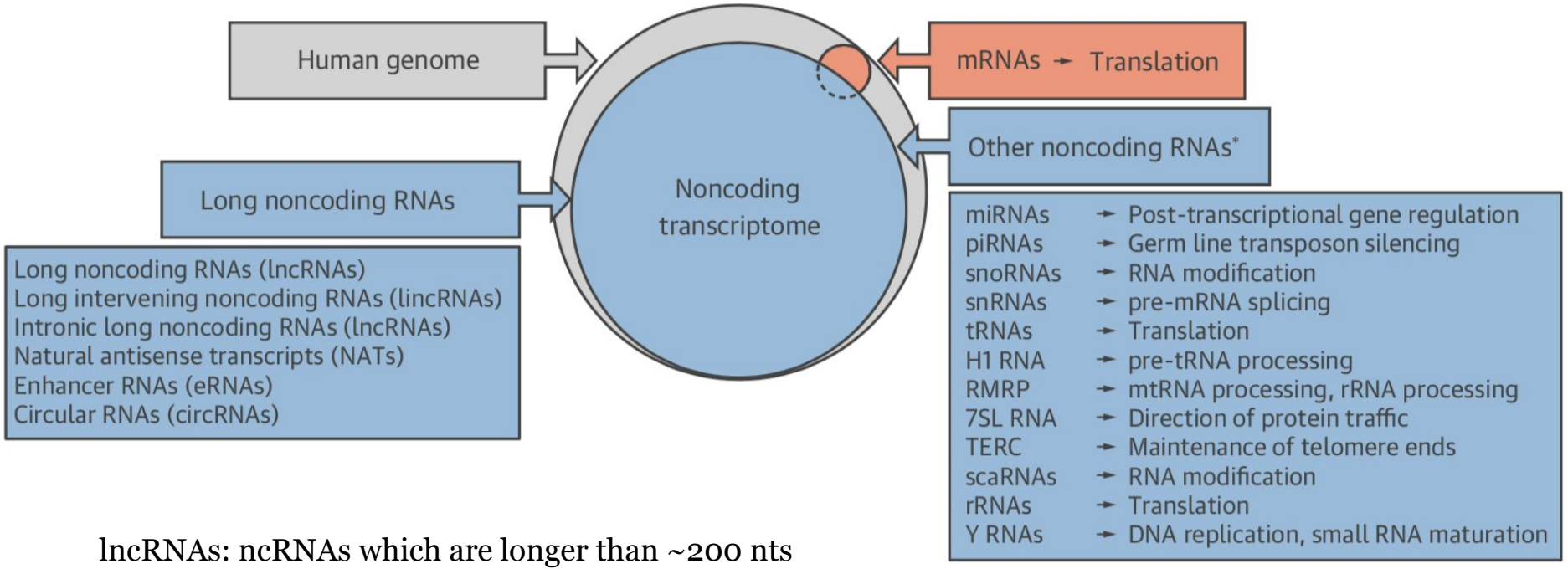
Identifying RNAs incorrectly labelled as non-coding

Progress Presentation
13 April 2020

A tiny bit of background



lncRNAs: ncRNAs which are longer than ~200 nts

Figure from: Boon, Reinier A., et al. "Long noncoding RNAs: from clinical genetics to therapeutic targets?." *Journal of the American College of Cardiology* 67.10 (2016): 1214-1226.

The Problem

Some RNAs that we accept as long non-coding actually contain small ORFs, which code for small proteins/micro-peptides.

The Problem

Some RNAs that we accept as non-coding actually contain small ORFs, which code for small proteins/micro-peptides.

- Traditional criteria of what makes an ORF: Has 100 amino acids in eukaryotes and 50 amino acids in bacteria.
- This arbitrary criteria excludes lot of ncRNAs that contain smaller ORFs.
- But there's growing evidence that ncRNAs produce biologically relevant micro-peptides (shown across species: humans, flies, bacteria).

The Problem

Some RNAs that we accept as non-coding actually contain small ORFs, which code for small proteins/micro-peptides.

- Traditional criteria of what makes an ORF: Has 100 amino acids in eukaryotes and 50 amino acids in bacteria.
- This arbitrary criteria excludes lot of ncRNAs that contain smaller ORFs.
- But there's growing evidence that ncRNAs produce biologically relevant micro-peptides (shown across species: humans, flies, bacteria).

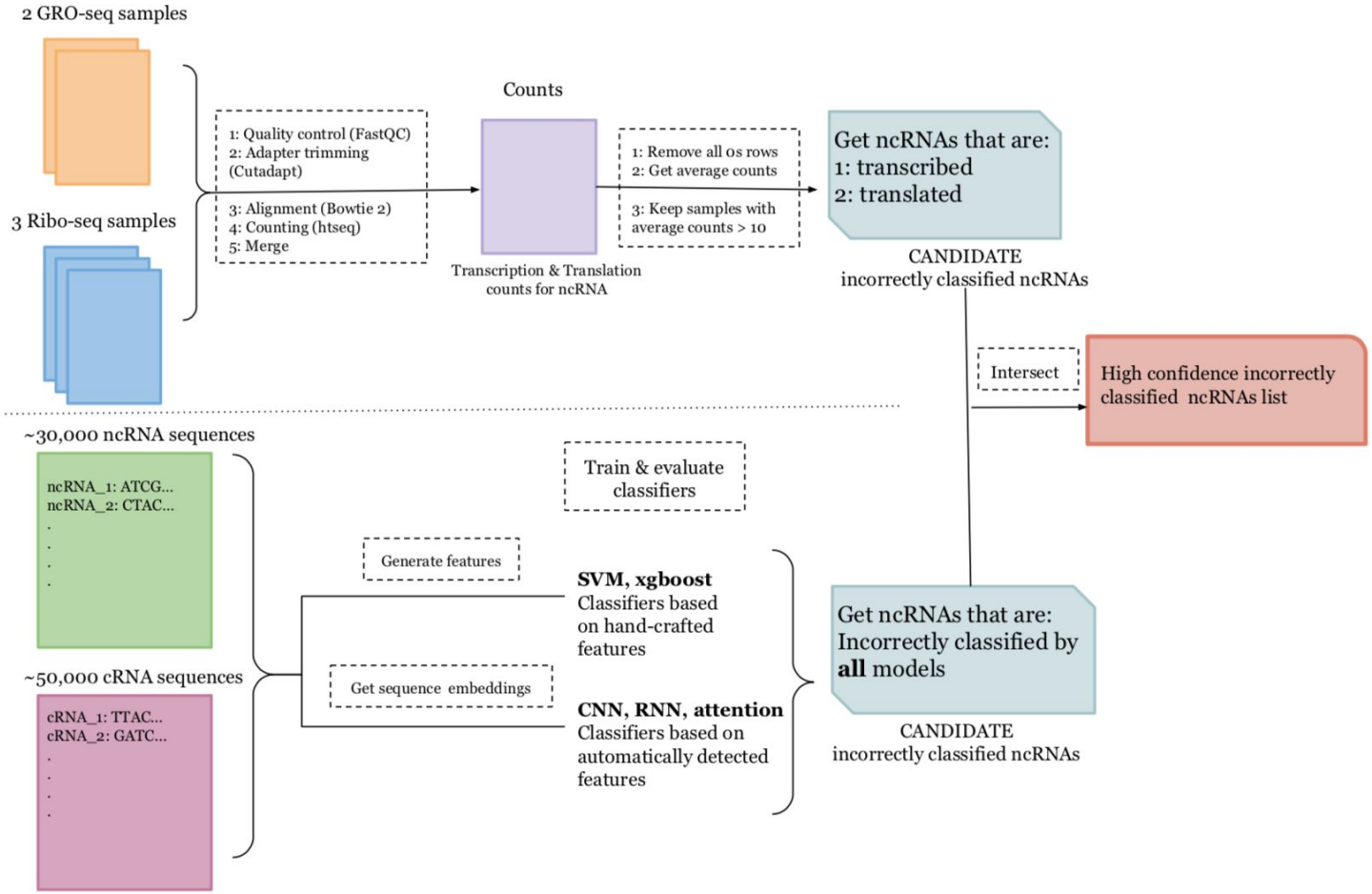
Aim

Identify RNAs incorrectly classified as non-coding (when they are actually translated).

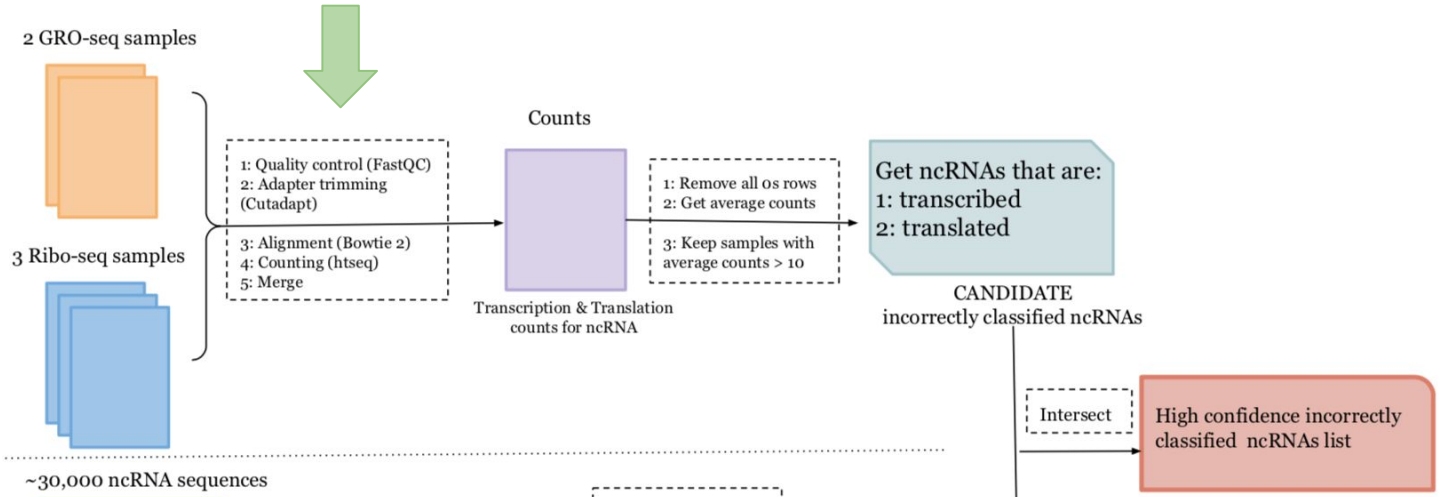
Proposed Solution

Part 1: NGS

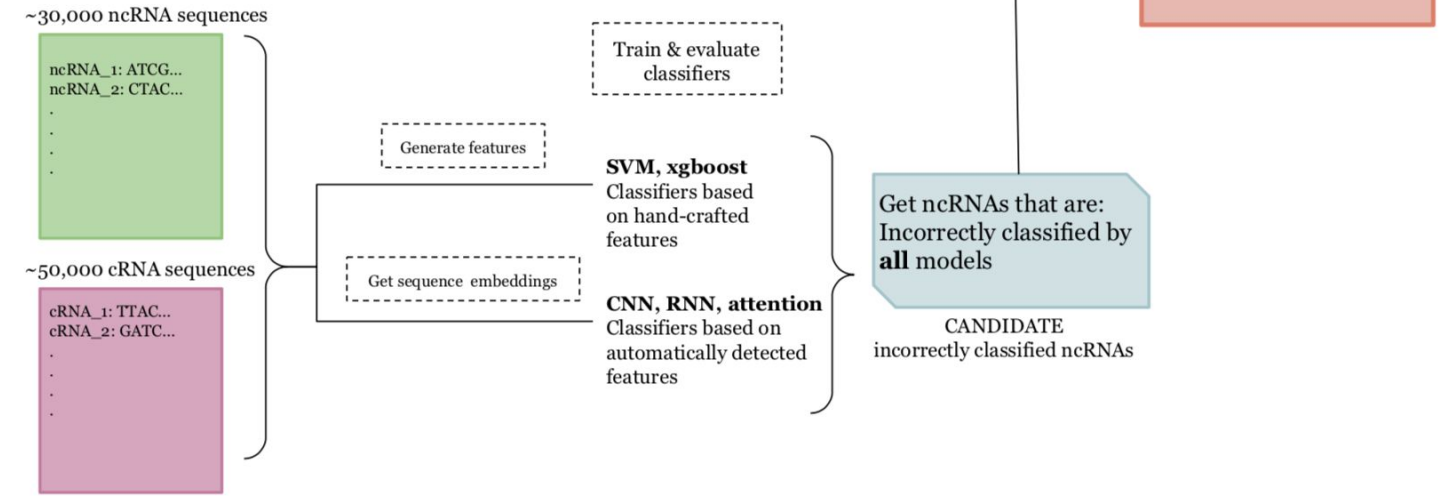
Part 2: ML



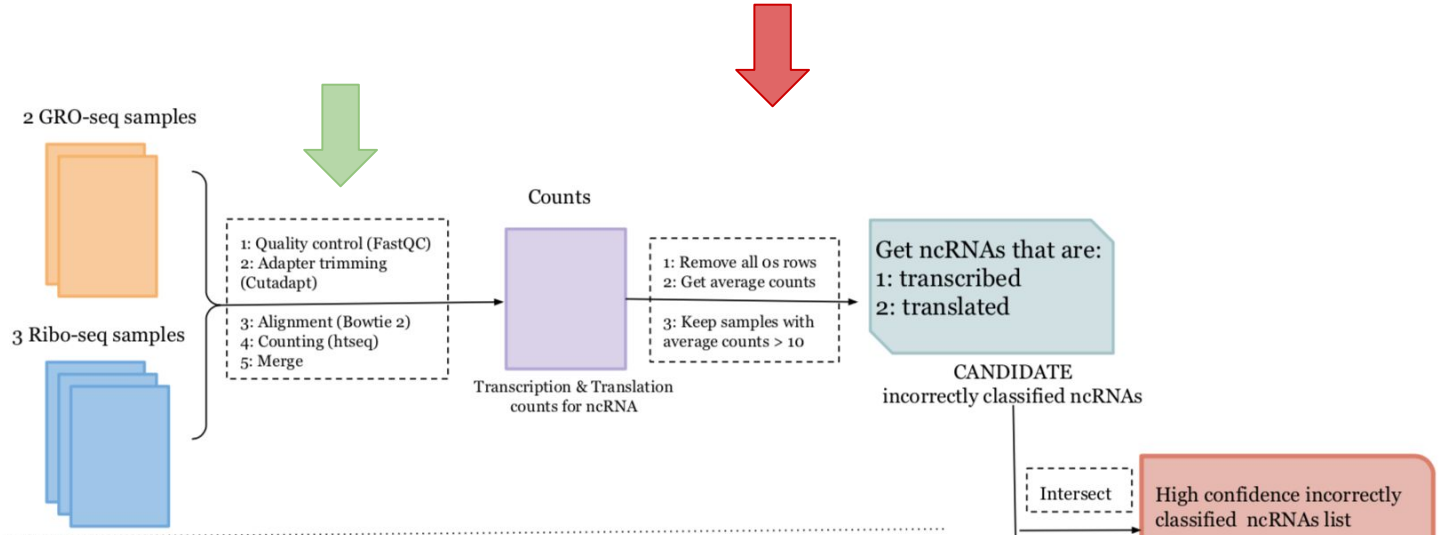
Part 1: NGS



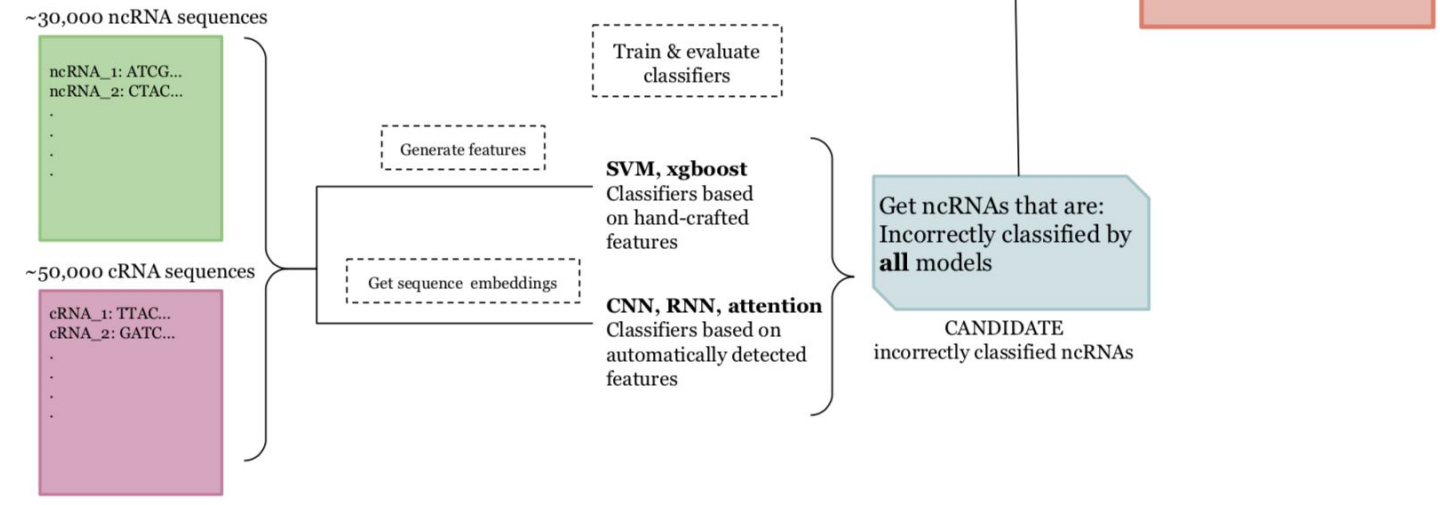
Part 2: ML



Part 1: NGS



Part 2: ML



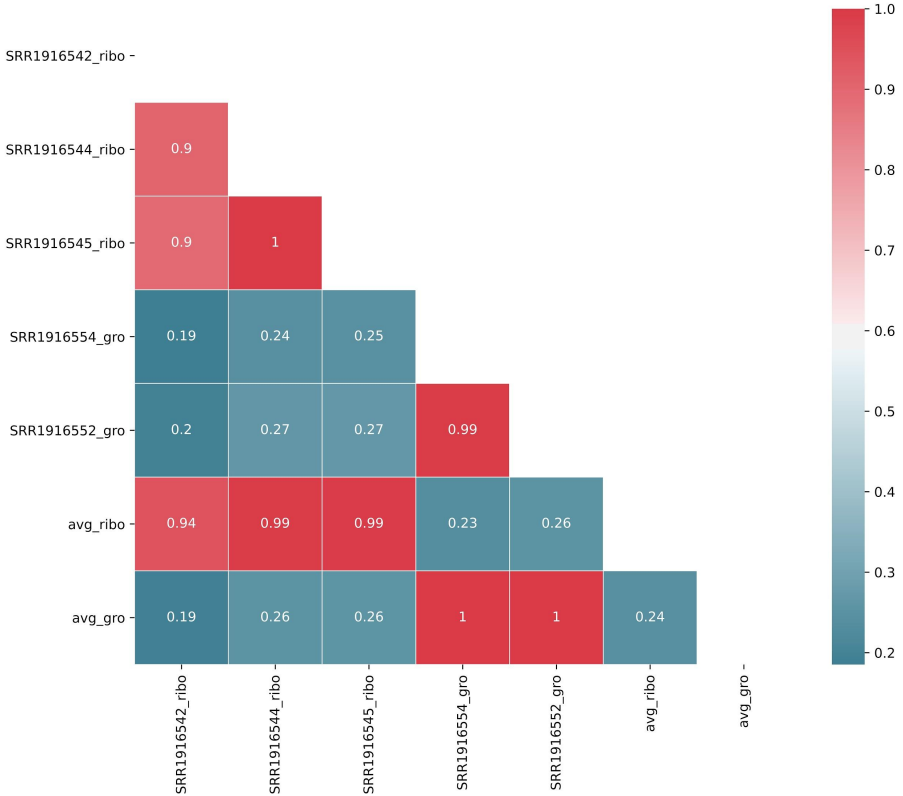
Part 1: NGS Data Counts Matrix

	lncRNA	SRR1916542_ribo	SRR1916544_ribo	SRR1916545_ribo	SRR1916554_gro	SRR1916552_gro	avg_ribo	avg_gro
15652	ENSG00000285646.2	8216	11120	11684	4374	3965	10340.000000	4169.5
6845	ENSG00000251562.8	1548	7158	7874	18224	19654	5526.666667	18939.0
7976	ENSG00000255717.7	2917	3599	3860	8730	7320	3458.666667	8025.0
11911	ENSG00000269900.3	3482	1852	1894	2342	2739	2409.333333	2540.5
8917	ENSG00000259001.3	2459	2249	2372	280	372	2360.000000	326.0

...

	lncRNA	SRR1916542_ribo	SRR1916544_ribo	SRR1916545_ribo	SRR1916554_gro	SRR1916552_gro	avg_ribo	avg_gro
6973	ENSG00000253288.2	7	4	5	16	15	5.333333	15.5
10564	ENSG00000263327.6	6	7	3	10	19	5.333333	14.5
10207	ENSG00000261537.1	6	4	6	7	13	5.333333	10.0
11880	ENSG00000269793.7	5	8	3	11	4	5.333333	7.5
7004	ENSG00000253357.1	8	6	2	10	2	5.333333	6.0

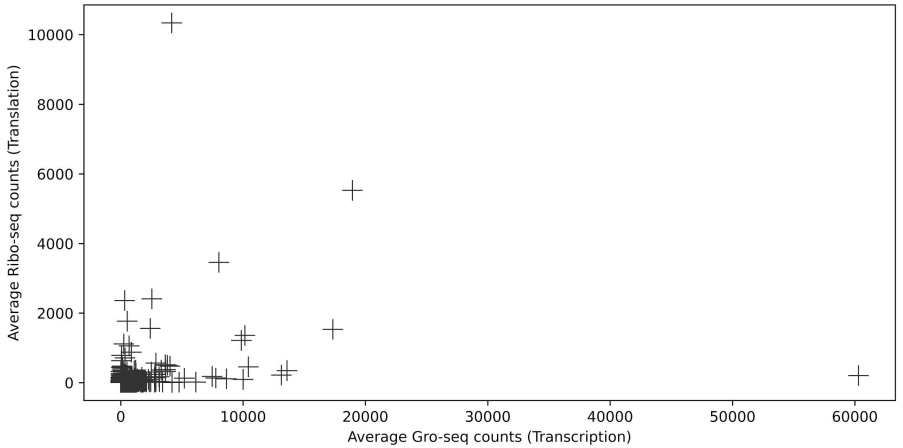
Part 1: NGS Data Correlation between NGS sample counts

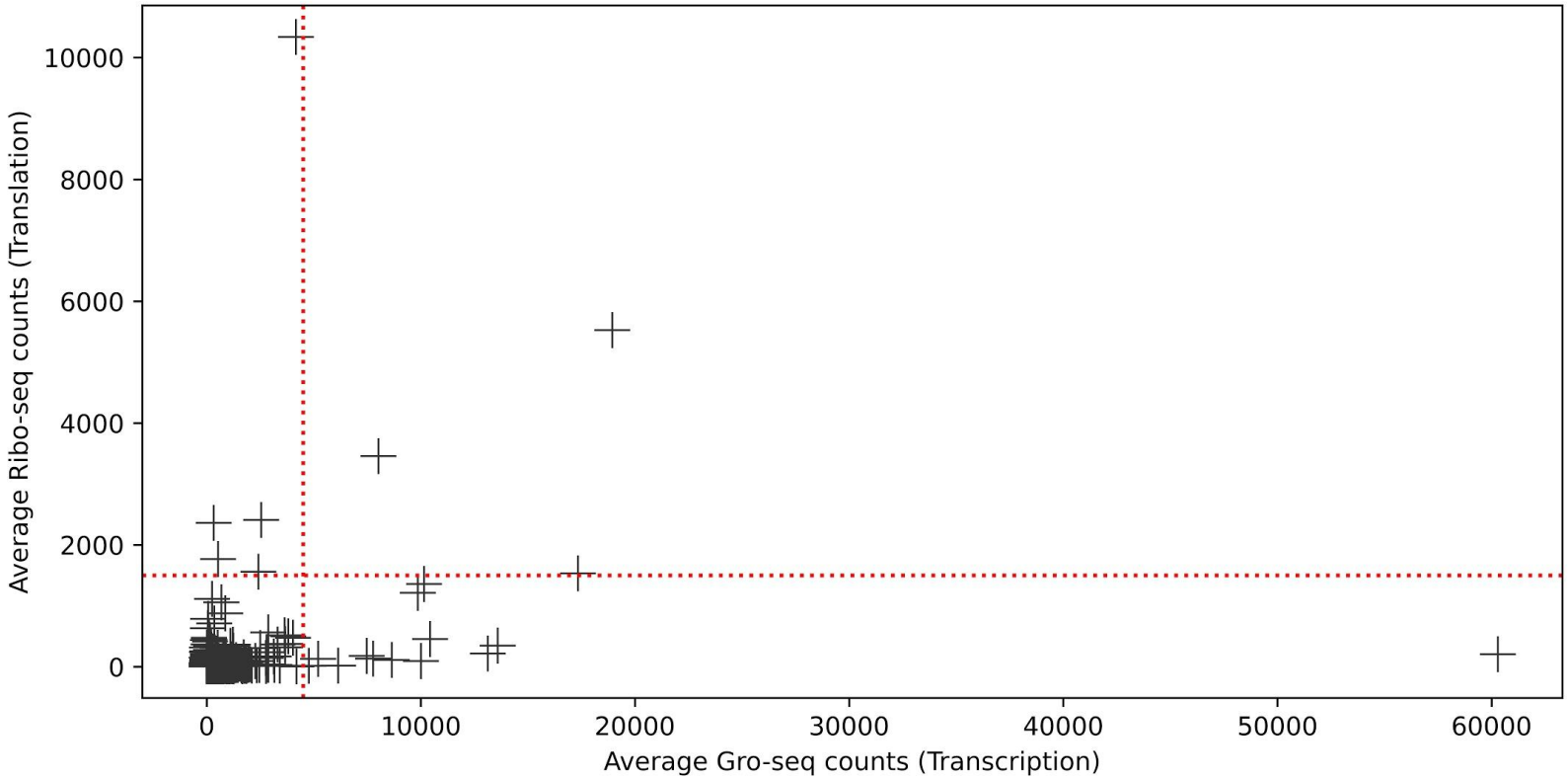


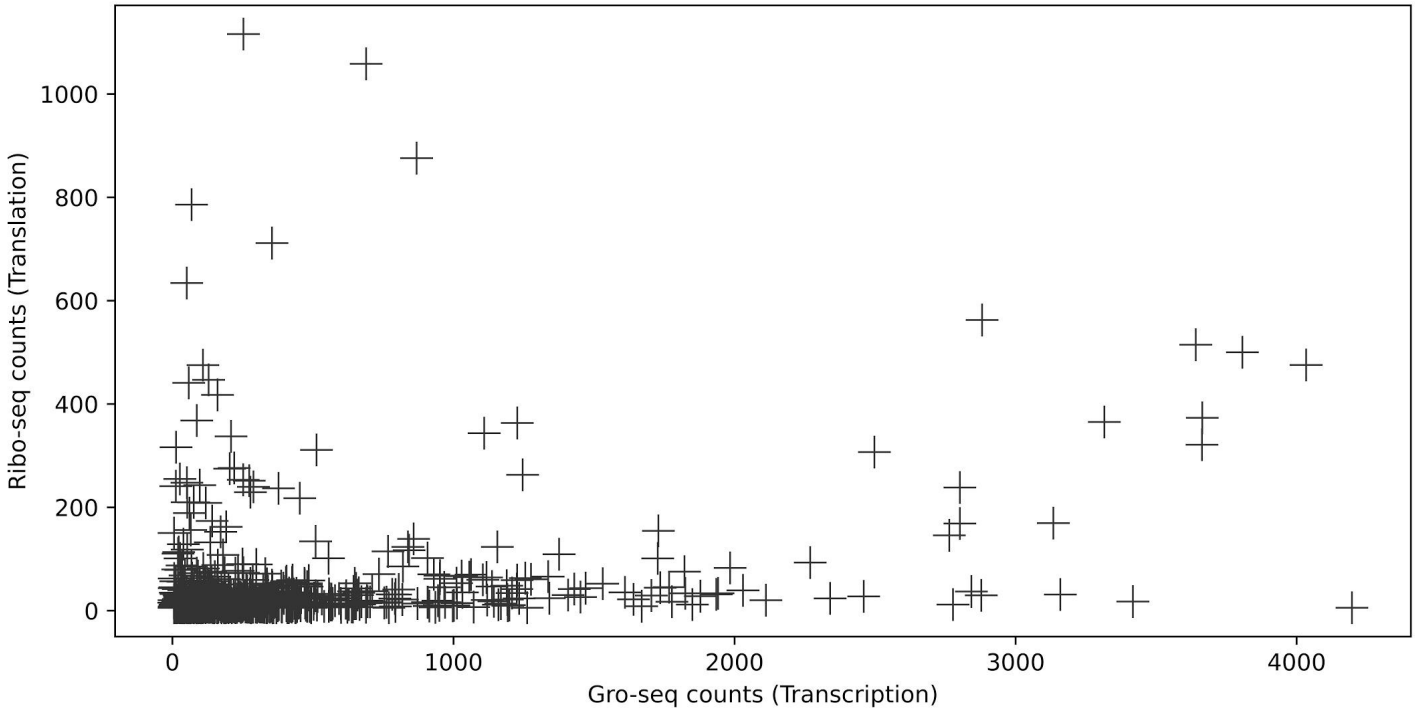
High correlation between ribo-ribo & gro-gro samples

Part 1: NGS Data Counts

	lncRNA	SRR1916542_ribo	SRR1916544_ribo	SRR1916545_ribo	SRR1916554_gro	SRR1916552_gro	avg_ribo	avg_gro
15652	ENSG00000285646.2	8216	11120	11684	4374	3965	10340.000000	4169.5
6845	ENSG00000251562.8	1548	7158	7874	18224	19654	5526.666667	18939.0
7976	ENSG00000255717.7	2917	3599	3860	8730	7320	3458.666667	8025.0
11911	ENSG00000269900.3	3482	1852	1894	2342	2739	2409.333333	2540.5
8917	ENSG00000259001.3	2459	2249	2372	280	372	2360.000000	326.0



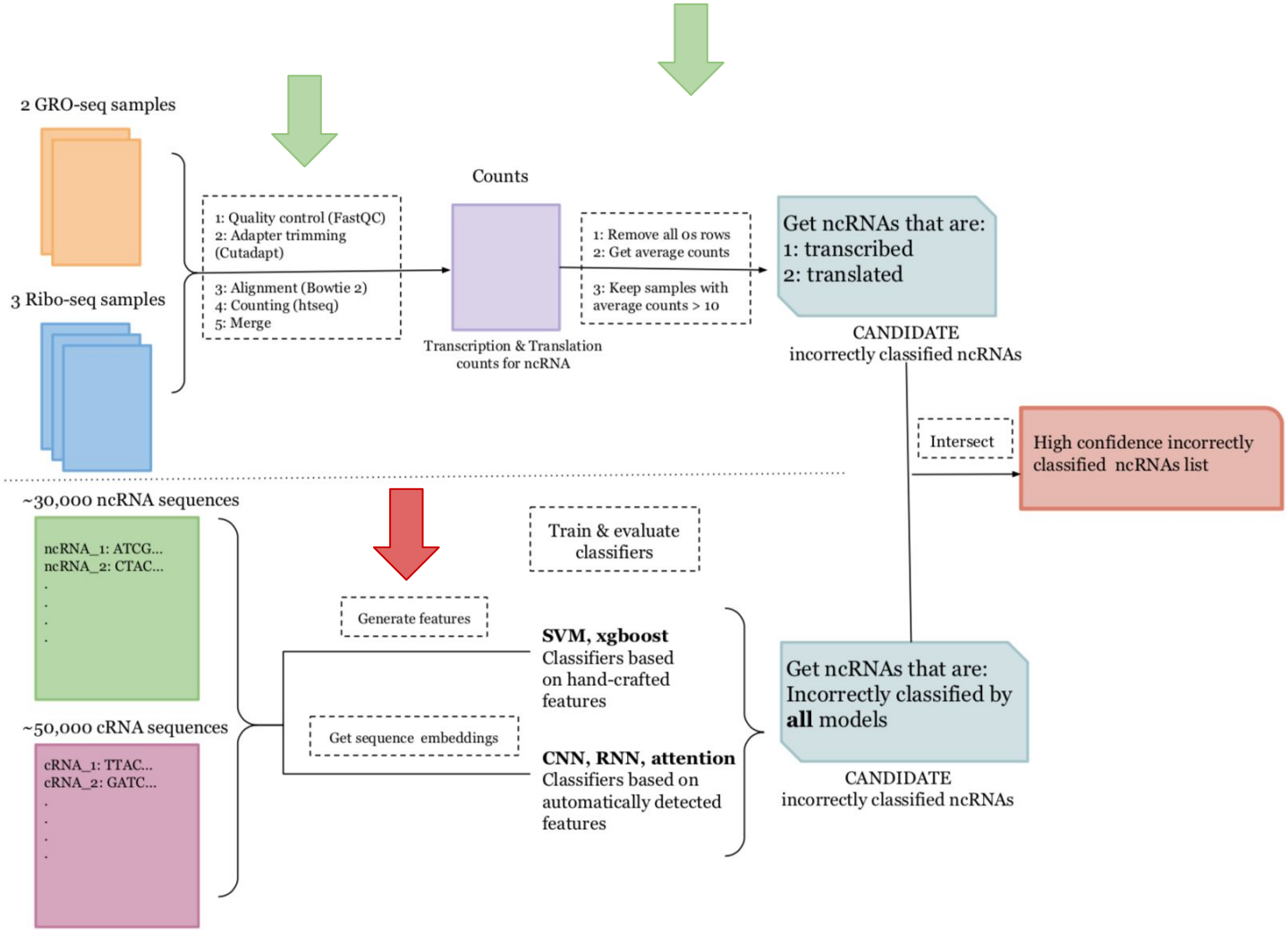

Average Ribo-seq counts (Translation)
10000
8000
6000
4000
2000
0
0
10000
20000
30000
40000
50000
60000
Average Gro-seq counts (Transcription)



RNA that are supposedly both transcribed and translated
836 candidates!

Part 1: NGS

Part 2: ML



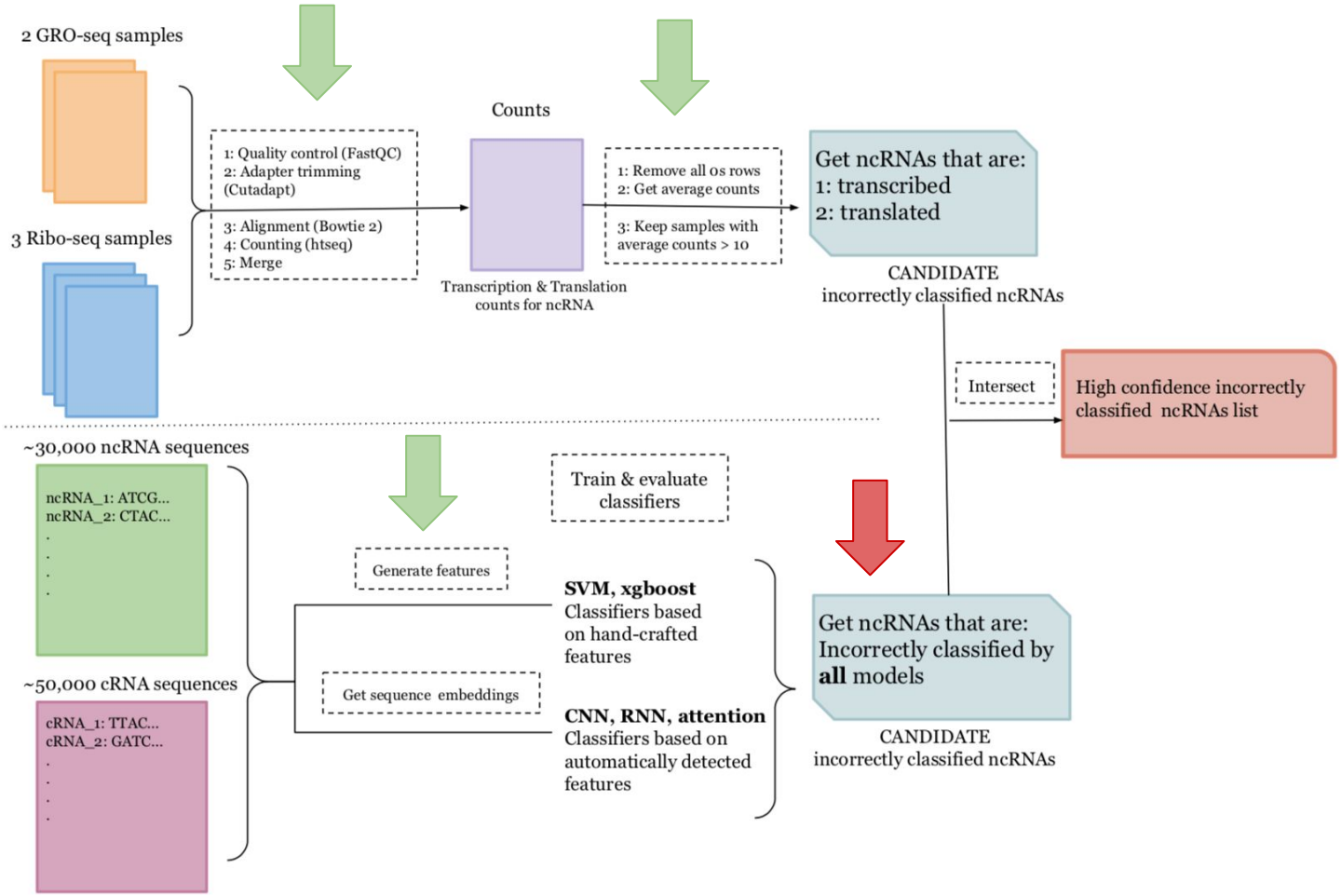
Part 2: Machine Learning Feature Generation

Property	Description	Number of features added
ORF length	length of the longest possible ORF	1
ORF coverage	quality of ORF	1
Fickett score	codon bias for 4 nucleotides	1
Hexamer score	hexamer usage bias	1
ORF integrity	Binary, whether ORF contains start and stop codon	1
Isoelectric point	pH at which molecule carries no net charge	1
Gravy	average hydropathicity of predicted peptide	1
Instability	estimated stability of predicted peptide	1
Composition	percentage of each of the 4 nucleotides	4
Transition	percent frequency of transition from each of nt to other nts	6
Distributio	distribution for each nt 25% intervals along sequence	20

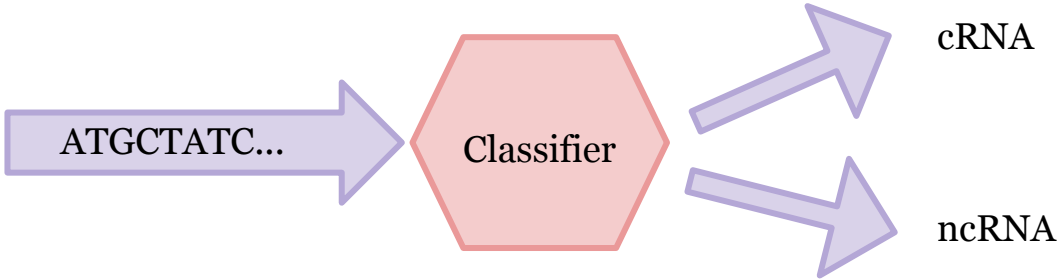
Tong, Xiaoxue, and Shiyong Liu. "CPPred: coding potential prediction based on the global description of RNA sequence." *Nucleic acids research* 47.8 (2019): e43-e43.

Part 1: NGS

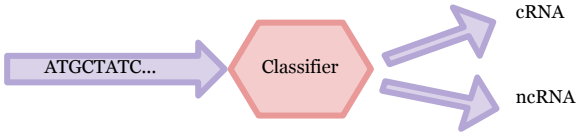
Part 2: ML



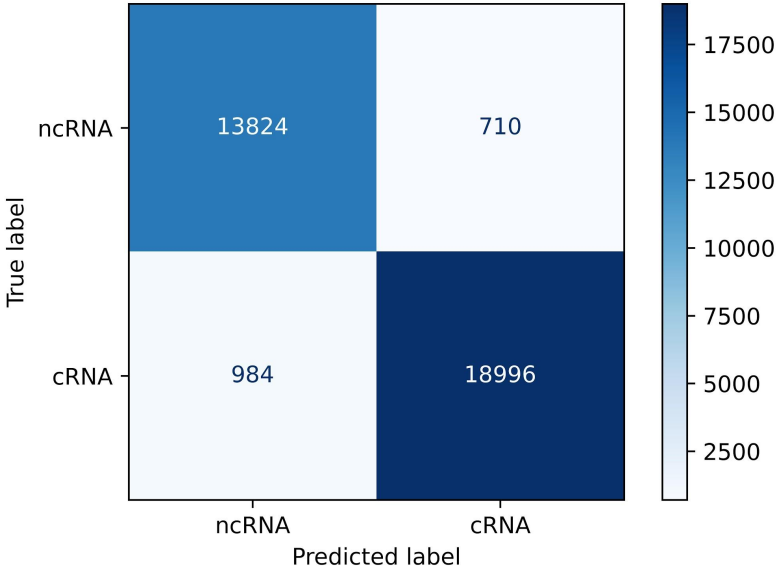
A tiny bit of background



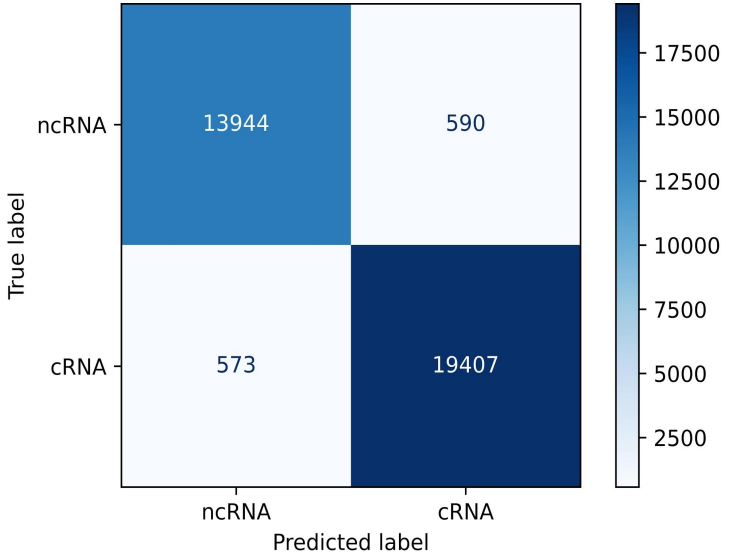
Part 2: Machine Learning Models



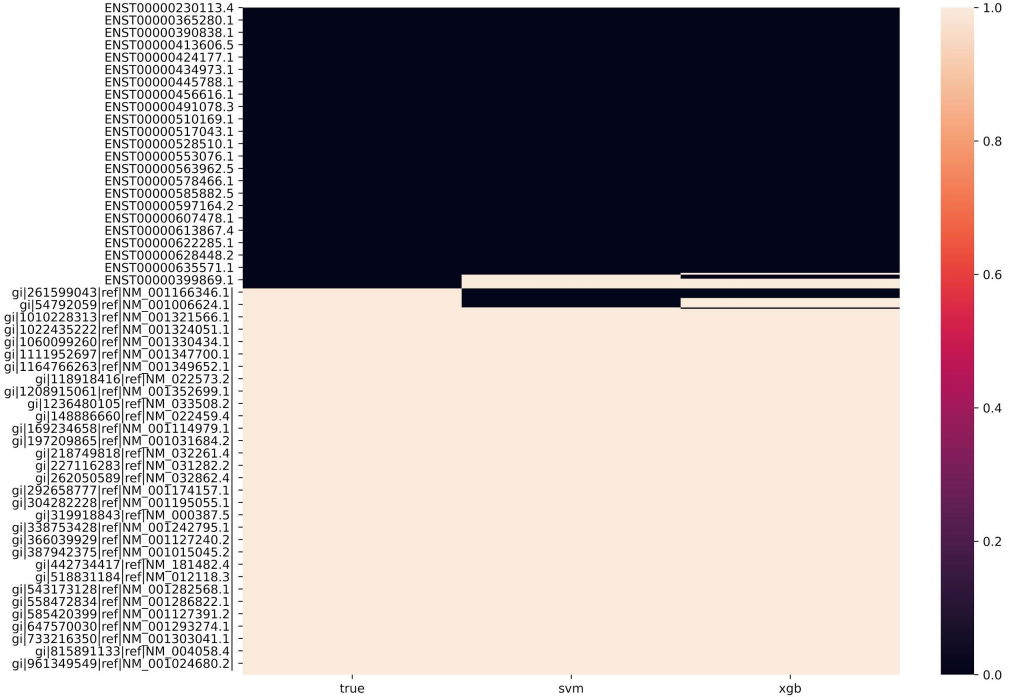
Classifier: SVM



Classifier: Xgboost



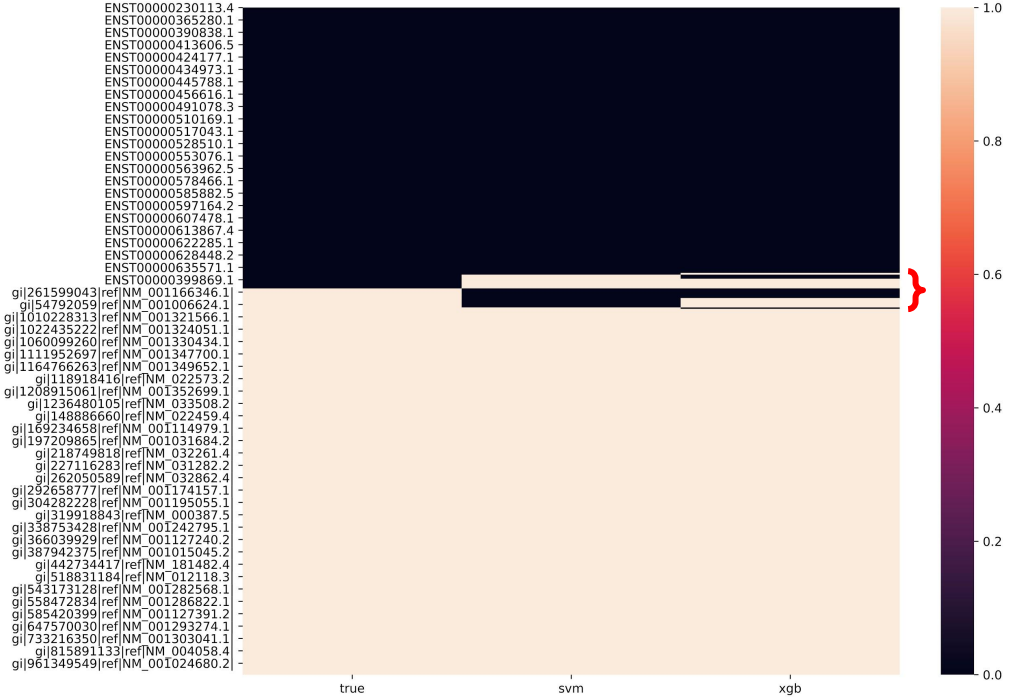
Part 2: Machine Learning Predictions



← 1 refers to cRNA

← 0 refers to ncRNA

Part 2: Machine Learning Predictions

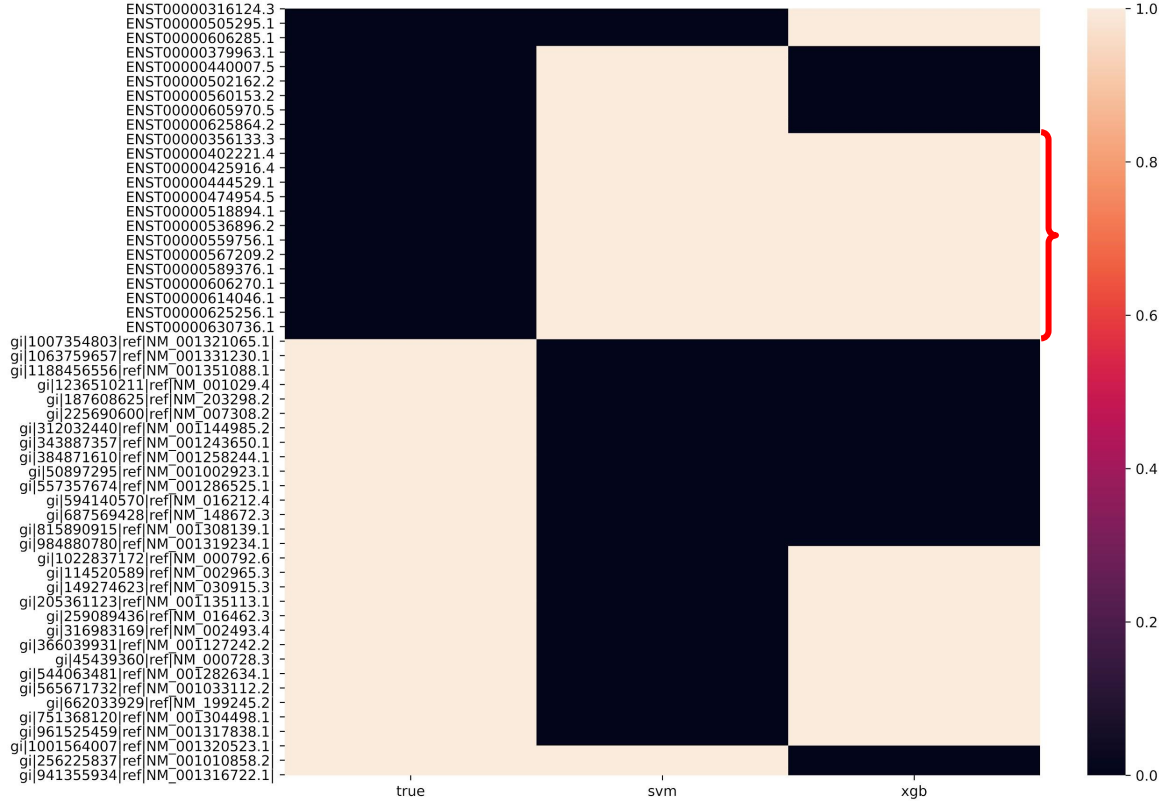


← 1 refers to cRNA

← Zoom into this region

← 0 refers to ncRNA

Part 2: Machine Learning Predictions



← 1 refers to cRNA

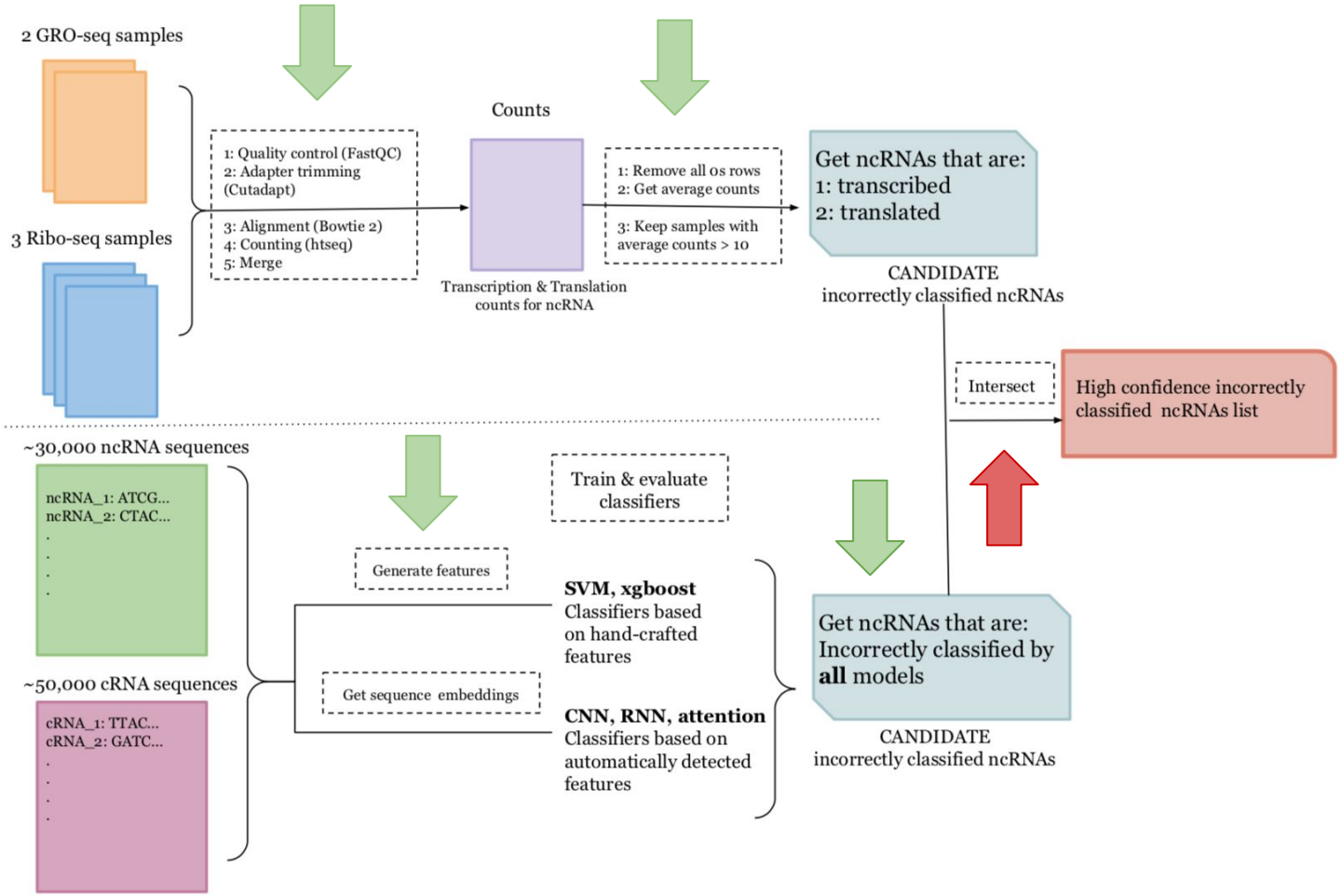
Labelled as ncRNA but both model classify as cRNA

499 candidates!

← 0 refers to ncRNA

Part 1: NGS

Part 2: ML



Part 1: NGS

Part 2: ML