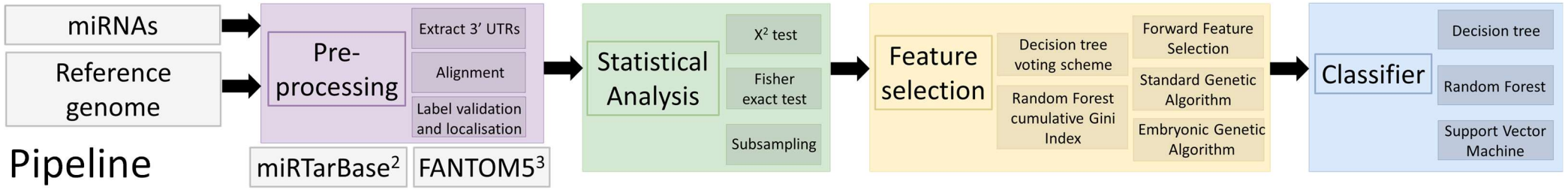
# Validation-driven features predictive of miRNA/mRNA interactions are captured with genetic algorithms-based selection using *feamiR*

**E.C. Williams** & I. Mohorianu@

Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, UK

## Pipeline

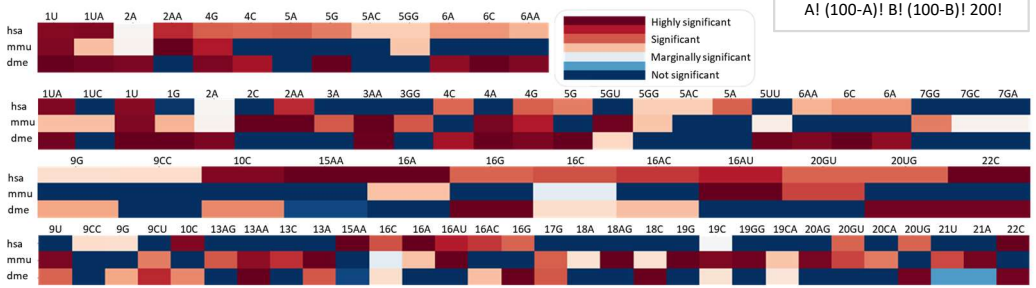



## Aims

- Existing tools for predicting miRNA-mRNA interactions are based on fixed rules and achieve limited sensitivity, e.g. TargetScan at 0.643
- Additional information is available through miRTarBase[2] [validations] and FANTOM5[3] [localisation] which allow increased accuracy and sensitivity
- Current approaches focus on seed matches, compensatory and flanking regions coupled with localisation improve predictions (feamiR[1] )

## Statistical background

Heatmaps illustrating the conservation of significant seed- (top) and compensatory- (bottom) features between *H. sapiens*, *M. musculus* and *D. melanogaster* assessed using Fisher exact BH-adjusted p-values.




| | Validated | All | |
|---|---|---|---|
| G | A | B | =A+B |
| Not G | 100-A | 100-B | =200-A-B |
| | =100 | =100 | =200 |

Fisher exact p-value:

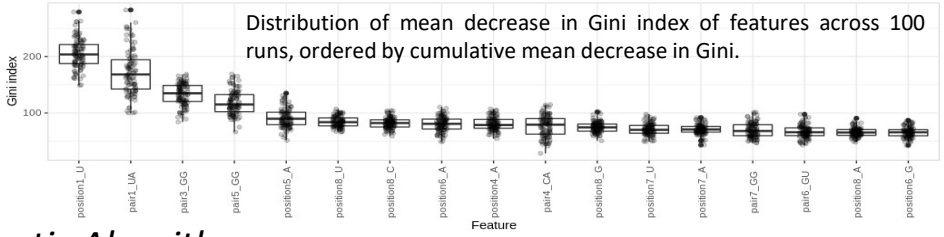
$$\frac{(A+B)!\,(200-A-B)!\,100!\,100!}{A!\,(100-A)!\,B!\,(100-B)!\,200!}$$

The majority of features with significant differences between validated and non-validated interactions are conserved across 1 or more species.

## Feature selection

- feamiR[1] includes 6 feature selection approaches: statistical, tree/entropy-based and Genetic Algorithms-based approaches.
- These methods identify both common and specific features.

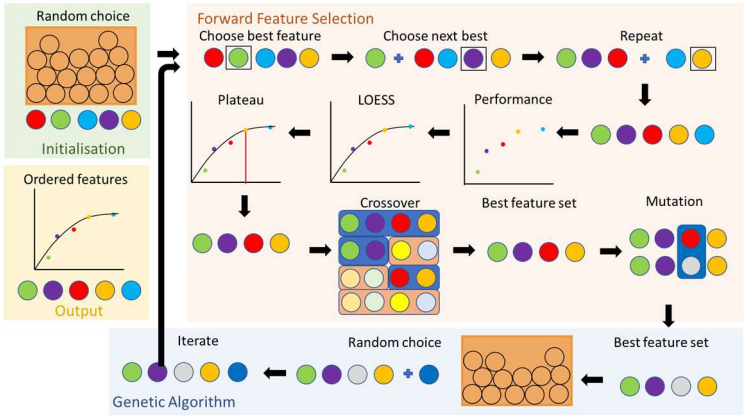
### *Entropy-based feature selection*




Distribution of mean decrease in Gini index of features across 100 runs, ordered by cumulative mean decrease in Gini.

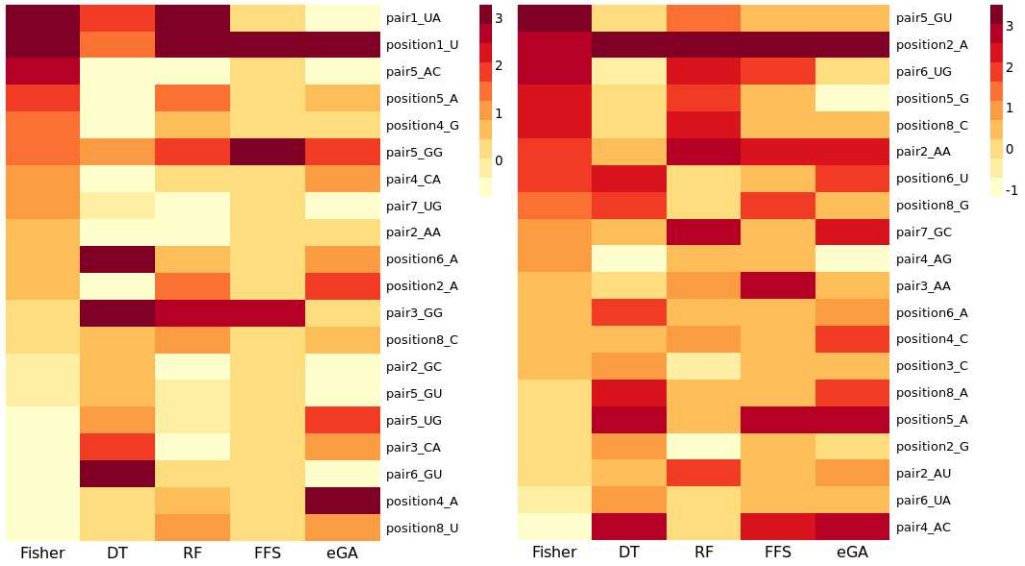
### *Genetic Algorithms*

As well as a standard Genetic Algorithm with a custom fitness function and Forward Feature Selection, feamiR[1] implements an embryonic Genetic Algorithm which combines these approaches




# Classifiers

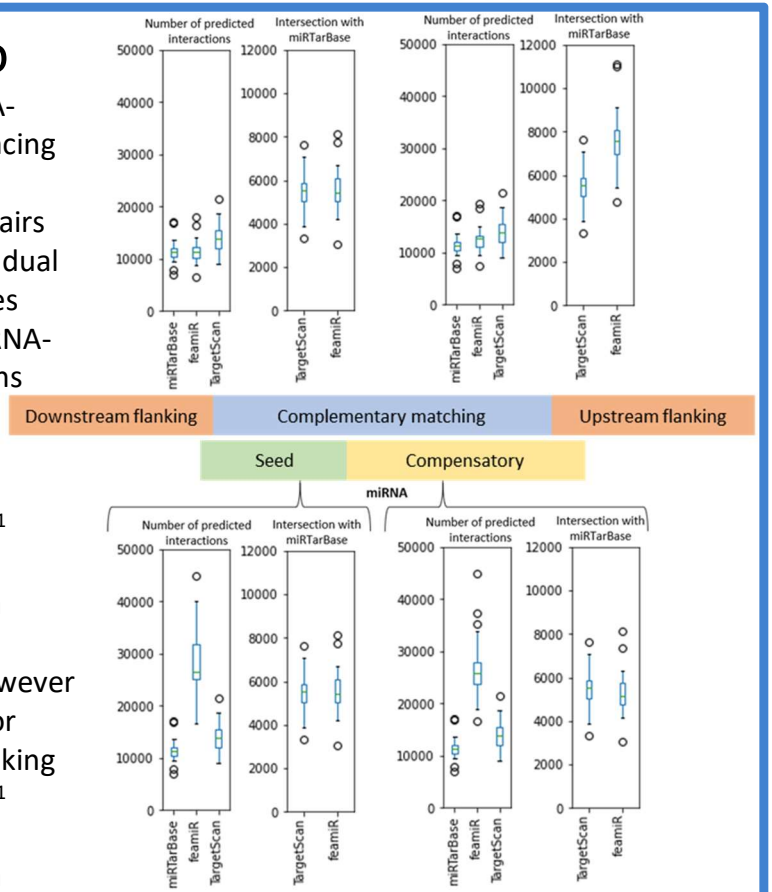
Heatmap of discriminative power of features across 5 feature-selection methods on the full positive set and positive set where miRNAs are localised to pluripotent stem cells: Fisher exact p-values, decision tree (DT) voting scheme, random forest (RF) cumulative mean decrease in Gini, forward feature selection (FFS) and embryonic Genetic Algorithm (eGA). The top 10 features across methods were included; all scores were quantile normalised on all 144 features per method, for comparability.




- Each feature selected by eGA is also significant for another method
- Oher approaches choose specific features e.g. **pair5_AC** identified using Fisher exact tests or **pair6_GU** in DTs
- The novel eGA captures a consistent set of highly ranked features, e.g. **pair1_UA**, **position1_U** and **position5_A** and **pair1_UG**.
- SVMs trained on the features selected by the eGA achieve stable accuracy around 0.8, higher than on the features chosen by FFS.

# Spatial info

- Single-cell miRNA-mRNA co-sequencing is used to find miRNAs/mRNA pairs localised to individual cells and improves prediction of miRNA-mRNA interactions
- Using seed or compensatory features, feamiR[1] predicts more interactions than TargetScan and miRTarBase[2], however using upstream or downstream flanking features, feamiR[1] predicts fewer interactions than TargetScan yet the intersection with miRTarBase[2] is larger, supporting the observed increased sensitivity.




The distribution of number of interactions predicted by TargetScan, miRTarBase[2] and feamiR[1] (using seed, compensatory, upstream and downstream flanking features) across 19 cells, in a single cell co-expression experiment. We show the number of interactions per prediction tool and the size of the intersection of TargetScan and feamiR[1] with miRTarBase[2] (considered the true positive set).

# References

[1] E.C. Williams, A. Calinescu and I. Mohorianu. *feamiR: Feature selection based on Genetic Algorithms for predicting miRNA-mRNA interactions*, bioRxiv; https://doi.org/10.1101/2020.12.23.424130 CRAN: https://CRAN.R-project.org/package=feamiR GitHub: https://github.com/Core-Bioinformatics/feamiR

[2] H-Y. Huang, Y.D. Lin, J. Li, K.Y. Huang, S. Shrestha, H-C Hong, Y. Tang, Y-G. Chen, C-N. Jin, Y. Yu et al. *miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database.* Nucleic Acids Research, 48(D1):D148–D154, 2019. https://doi.org/10.1093/nar/gkz896

[3] M. Lizio, I. Abugessaisa, S. Noguchi, A. Kondo, A. Hasegawa, C.C. Hon, M. de Hoon, J. Severin, S. Oki, Y. Hayashizaki et al. *Update of the FANTOM web resource: expansion to provide additional transcriptome atlases.* Nucleic Acids Research, 47(D1):D752–D758, 11 2018. https://doi.org/10.1093/nar/gky1099