Predicting Cell Age from Synthetic scRNAseq Data

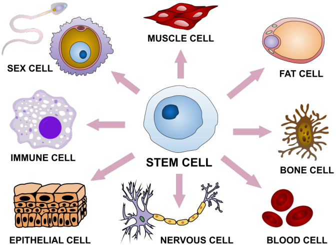
Thomas Kerby

Department of Mathematics and Statistics
Utah State University

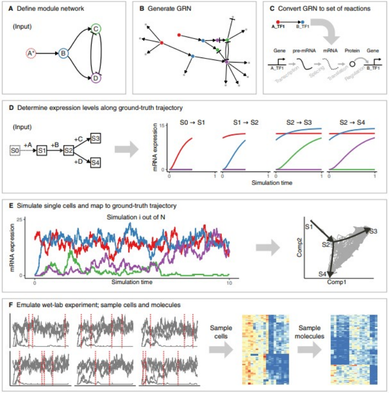
April 29, 2022

Motivating Question

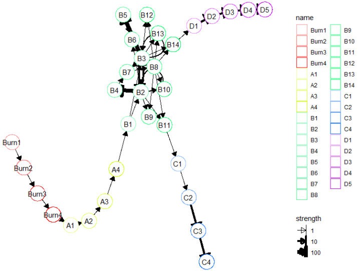
- Understanding Cell trajectories is a large research space.
- What genes drive cell differentiation?
- Moment of differentiation is difficult to distinguish



Data Overview

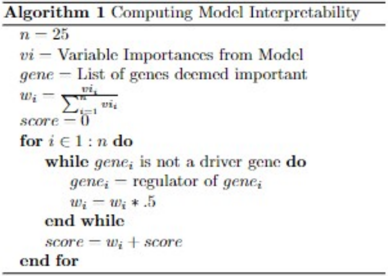


2000 Cells and 5031 Genes



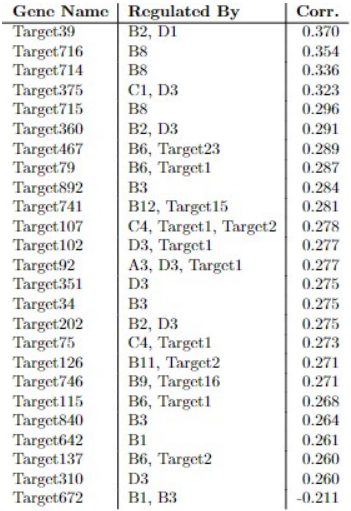
Determining Metrics

- Predictive Capacity
  - MSE
- Model Informativity
  - Based on driving genes
  - Scale between 0 and 1



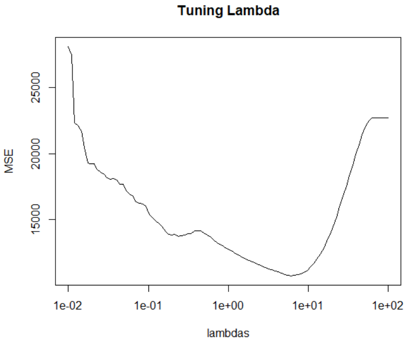
Algorithm 1 Computing Model Interpretability

```
$n = 25$
$vi$ = Variable Importances from Model
$gene$ = List of genes deemed important
$w_i = \frac{vi_i}{\sum_{i=1}^{n} vi_i}$
$score = 0$
for $i \in 1 : n$ do
    while $gene_i$ is not a driver gene do
        $gene_i$ = regulator of $gene_i$
        $w_i = w_i * .5$
    end while
    $score = w_i + score$
end for
```



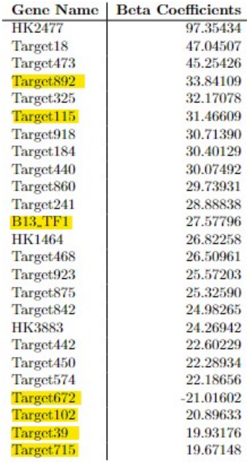
Gene Name	Regulated By	Corr.
Target39	B2, D1	0.370
Target716	B8	0.354
Target714	B8	0.336
Target375	C1, D3	0.323
Target715	B8	0.296
Target360	B2, D3	0.291
Target467	B6, Target23	0.289
Target79	B6, Target1	0.287
Target892	B3	0.284
Target741	B12, Target15	0.281
Target107	C4, Target1, Target2	0.278
Target102	D3, Target1	0.277
Target92	A3, D3, Target1	0.277
Target351	D3	0.275
Target34	B3	0.275
Target202	B2, D3	0.275
Target75	C4, Target1	0.273
Target126	B11, Target2	0.271
Target746	B9, Target16	0.271
Target115	B6, Target1	0.268
Target840	B3	0.264
Target642	B1	0.261
Target137	B6, Target2	0.260
Target310	D3	0.260
Target672	B1, B3	-0.211

Analysis - Lasso



Metrics

- MSE: 10723.23
- Model Informativity: 0.4956

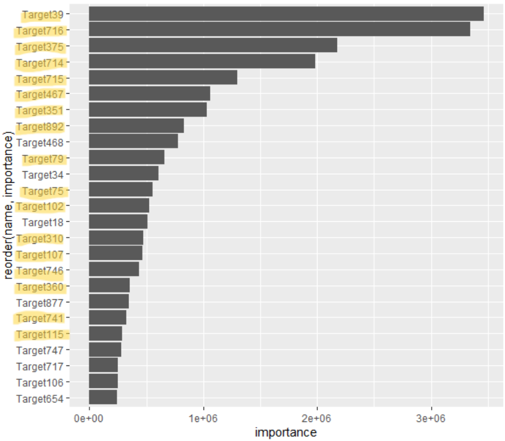


Gene Name	Beta Coefficients
HK2477	97.35434
Target18	47.04507
Target473	45.25426
Target892	33.84109
Target325	32.17078
Target115	31.46609
Target918	30.71390
Target184	30.40129
Target440	30.07492
Target860	29.73931
Target241	28.88838
B13_TF1	27.57796
HK1464	26.82258
Target468	26.50961
Target923	25.57203
Target875	25.32590
Target842	24.98265
HK3883	24.26942
Target442	22.60229
Target450	22.28934
Target574	22.18656
Target672	-21.01602
Target102	20.89633
Target39	19.93176
Target715	19.67148

Analysis - Random Forest

Metrics

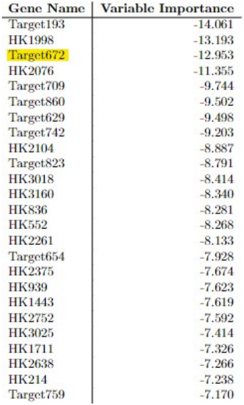
- MSE: 984.50
- Model Informativity: 0.7032



Analysis - Neural Network

Metrics

- MSE: 25066.65
- Model Informativity: 0.2702
  - Variable Importance calculated similarly to a random forest model

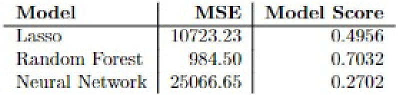


Gene Name	Variable Importance
Target193	-14.061
HK1998	-13.193
Target672	-12.953
HK2076	-11.355
Target709	-9.744
Target860	-9.502
Target629	-9.498
Target742	-9.203
HK2104	-8.887
Target823	-8.791
HK3018	-8.414
HK3160	-8.340
HK836	-8.281
HK552	-8.268
HK2261	-8.133
Target654	-7.928
HK2375	-7.674
HK939	-7.623
HK1443	-7.619
HK2752	-7.592
HK3025	-7.414
HK1711	-7.326
HK2638	-7.266
HK214	-7.238
Target759	-7.170

Conclusion

Takeaways

- This is a difficult problem
- Random Forest performed the best
- Variable Interactions are important



Model	MSE	Model Score
Lasso	10723.23	0.4956
Random Forest	984.50	0.7032
Neural Network	25066.65	0.2702