Comparative Analysis of Cancer Cell Lines with RUNX and NOTCH1 Inactivation/Inhibition/Knockdown

Erdi Küçük, Chiara Digravio, Nelson(Hao) Chen, Ogan Mancarci, Alastair Jamieson-Lane, and Sohrab Salehi

Background

NOTCH1 is a transcription factor that activates genes by forming chromatin-activated transcription complexes. When disregulated, it is an oncogenic driver for T-lymphoblastic leukaemia/lymphoma (TLL) and many other cancers [1]. Recently it has been shown that NOTCH1 binding sites significantly overlap with RUNX1 binding sites that is another transcription factor known to be oncogenic in leukemia [2].

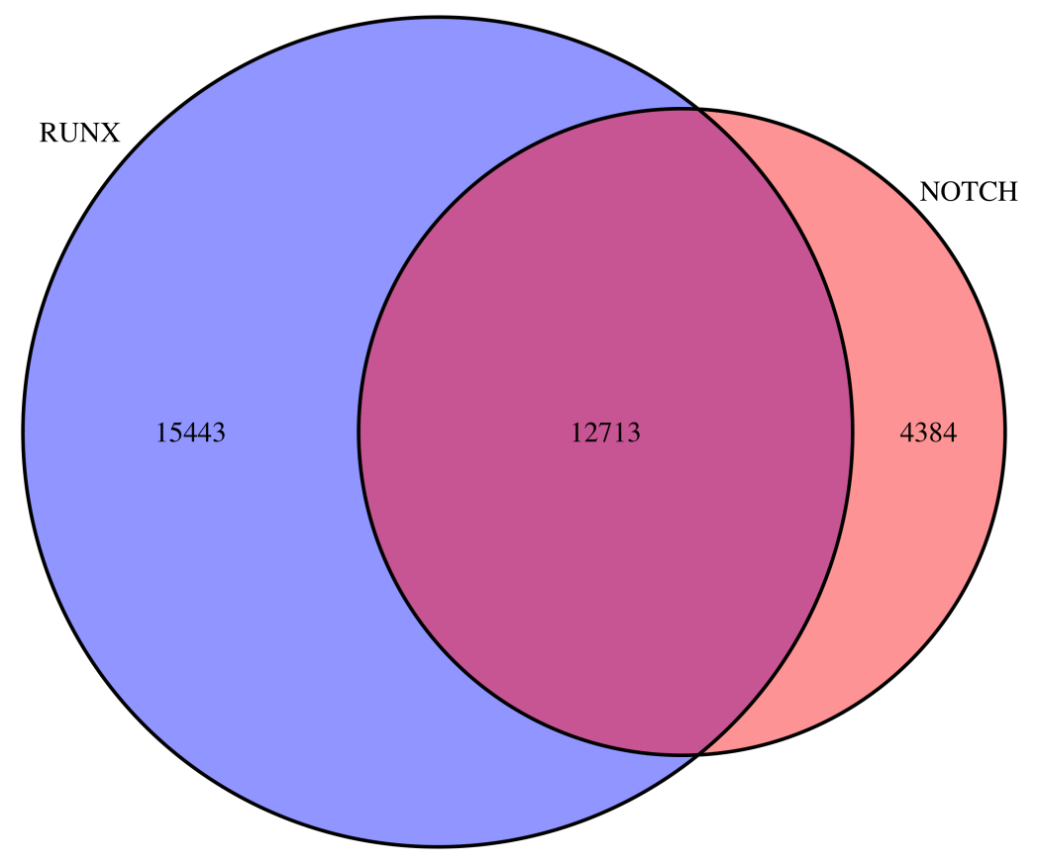

Fig 1. Significant overlap between NOTCH1& RUNX1 binding sites.

This suggests that RUNX1 and NOTCH1 may be binding cooperatively or in a hierarchical fashion to regulate the same set of genes [3]. In order to investigate the existence of a possible relationship between these two transcription factors we compared microarray results of RUNX1 knockdown and NOTCH1 pathway inhibition experiments to determine the overlap of differentially expressed genes. RUNX1 knockdown was performed by shRNA and on three different cell lines while NOTCH1 inhibition data is only composed of a single cell line and accomplished by enzymatic inhibition of activation of expressed NOTCH1.

Data

Table 1. Study design and number of samples.

	Treatment			Normal			
Experiment	HPBALL	RPMI	CUTLL1	HPBALL	RPMI	CUTLL1	Cell Lines
NOTCH1-KO	0	0	3	0	0	3	# of Samples
RUNX1-KO	2	2	2	2	2	2	

Exploratory Data Analysis

Looking for possible outliers

The dataset was normalized in advance (using RMA). Heat map of the sample correlation does not show any outlier.

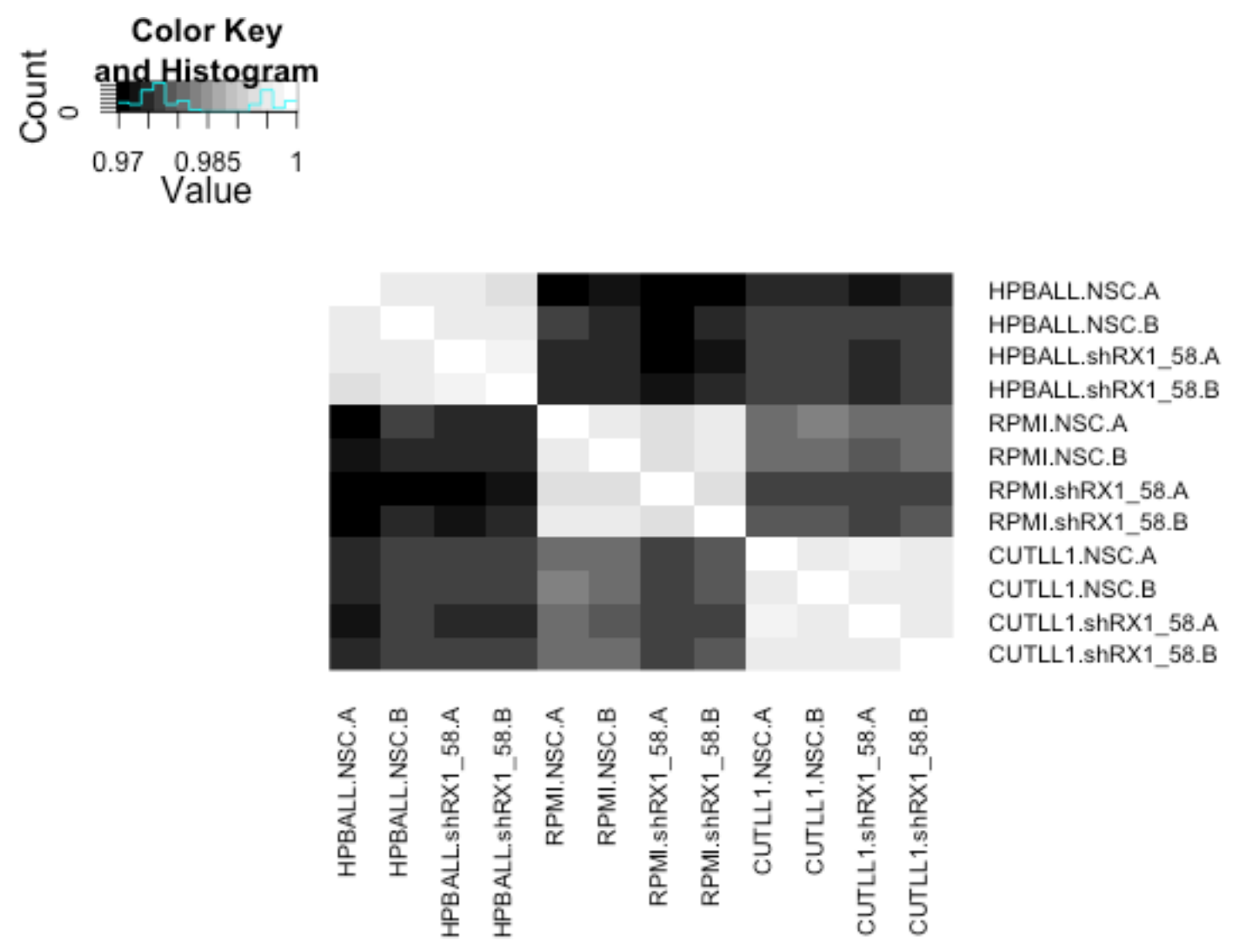

Fig 2. Heat map displaying the Pearson correlation of the samples.

PCA analysis

PCA based on the first three principal components, reveals that samples cluster based on the different cell lines. This result has been taken into account in the process of finding DE genes.

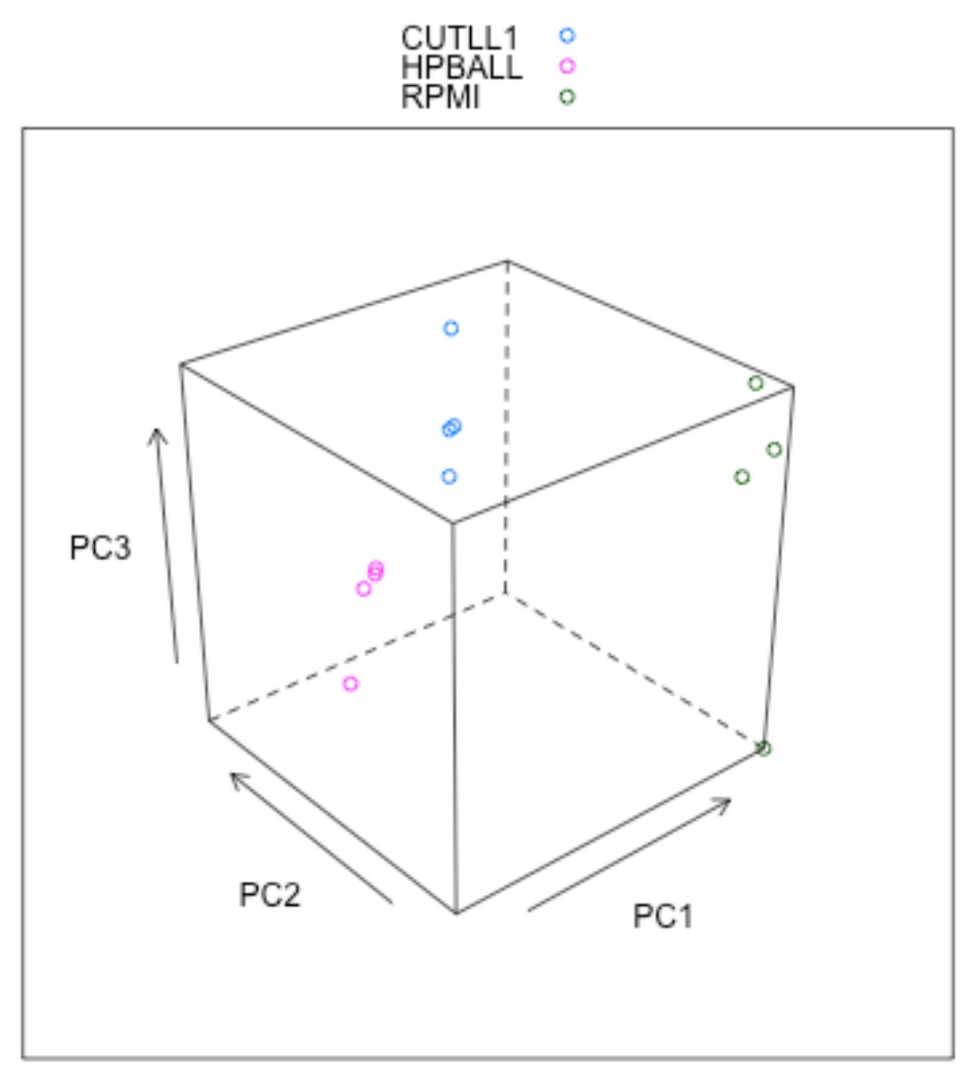

Fig 3. PCA based on the first three principal components. Points are colored based on the different cell lines.

Differential Expression Analysis

The biggest issue in dealing with this data set was the large variance between cell lines. Since cell lines could represent a source of variation, it is important to include them in the model in order to improve power.

$$Y_g = X_g\boldsymbol{\alpha}_g + \boldsymbol{\epsilon}_g, \quad g = 1, 2, \ldots, N.$$
$$\boldsymbol{\beta}_g = C^T\boldsymbol{\alpha}_g,$$

We first used a linear model with no interaction terms (A+B style linear model), treating cell lines as fixed effect. In this setting, the response variable is the gene expression, while the two covariates are treatment (non silent and clone 58), and cell line (HPBALL, CUTLL1, RPMI). X is a matrix with two columns accounting for two covariates.

Differentially Expressed Genes Identified

Along with a 0.05 threshold for our false discovery rate, we identify 236 potential differentially expressed genes. 204 of these were under expressed in the RUNX1 knock out treatment, and 32 were over expressed:

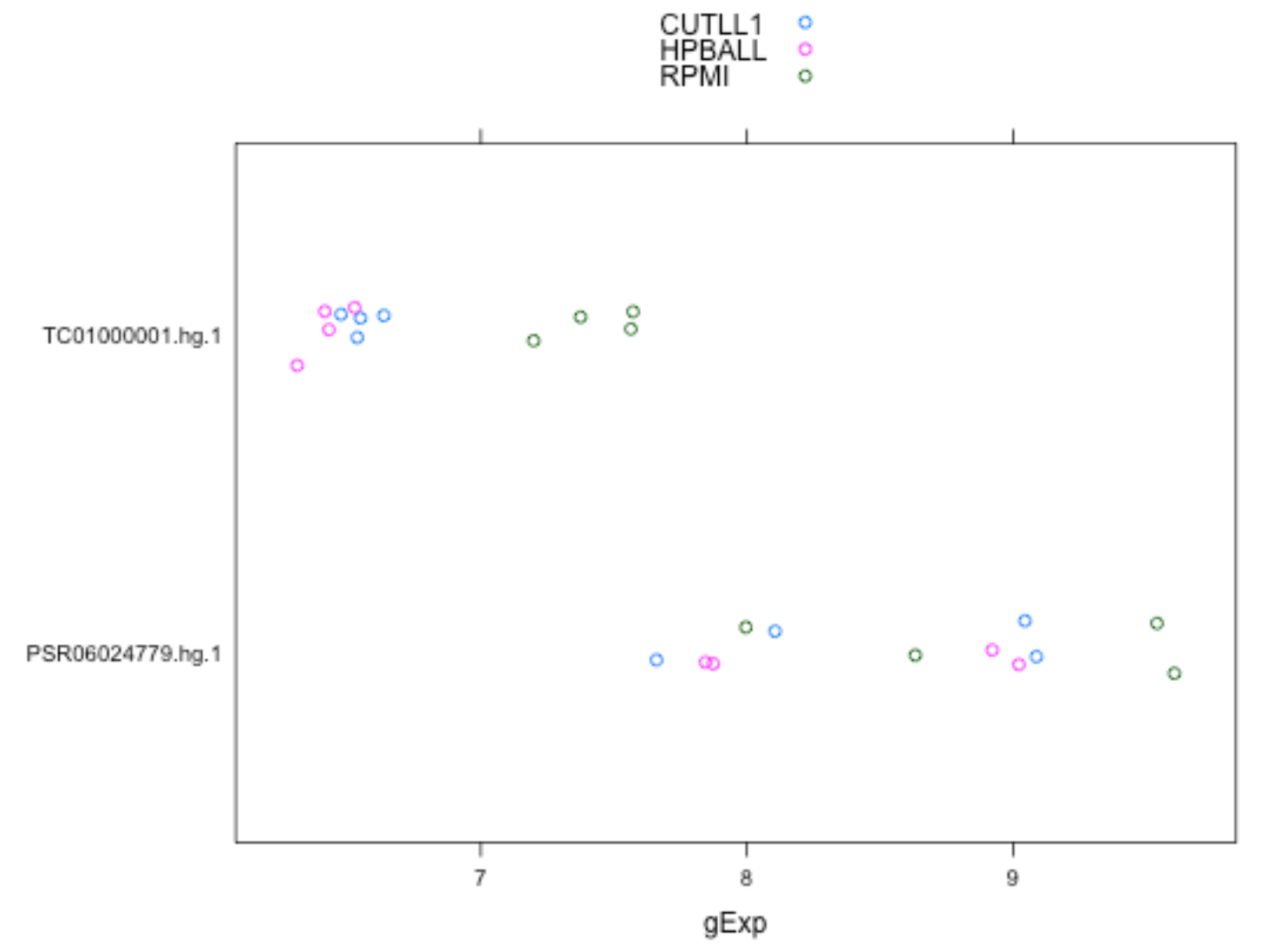

Fig 4. Stripplot of gene expression of a "hit" (PSR06024779.hg.1), and of a non hit (TC01000001.hg.1). Points are colored based on different cell lines.

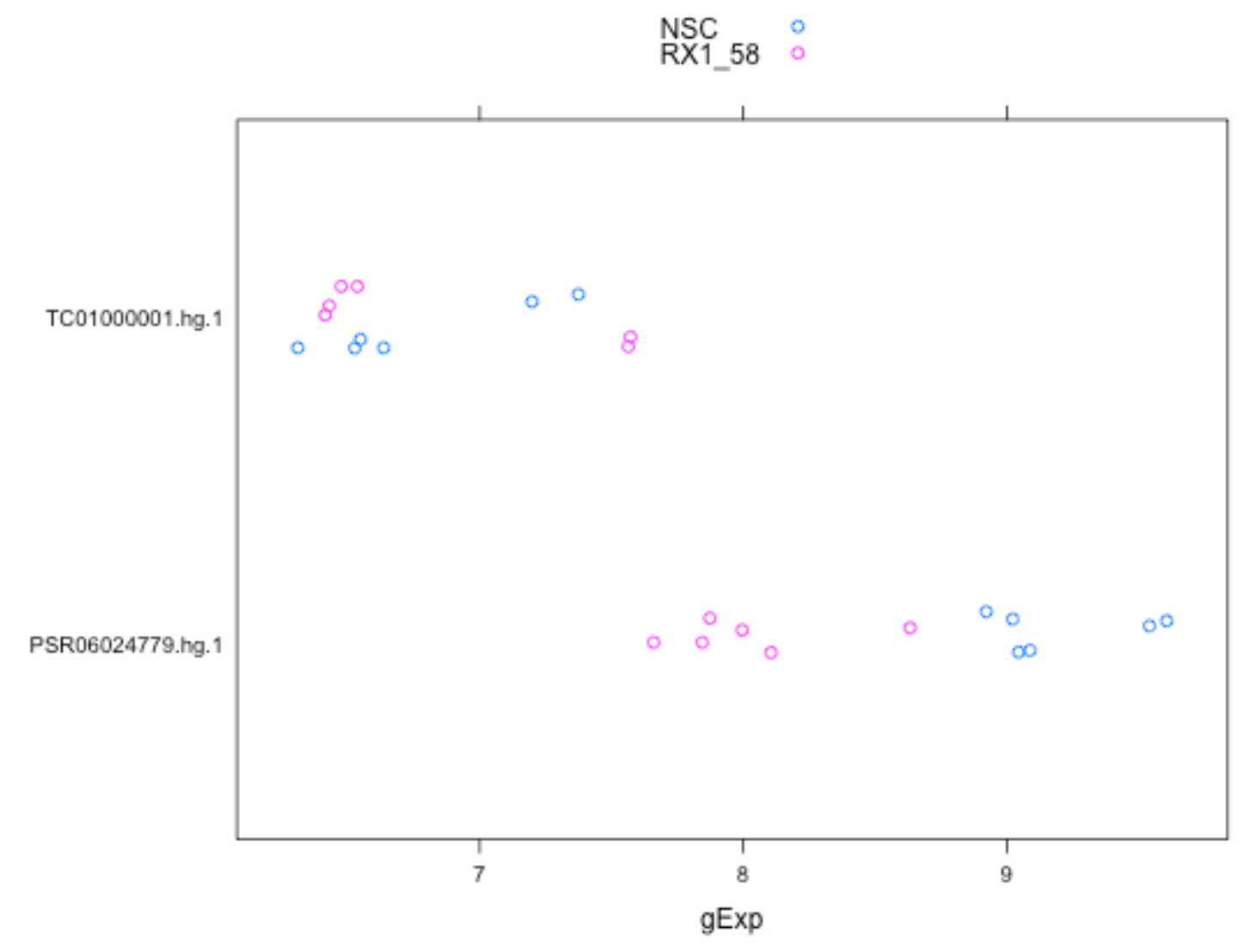

Fig 5. Stripplot of gene expression of a "hit" (PSR06024779.hg.1) of a non hit (TC01000001.hg.1). Points are colored based on different treatment.

Linear Mixed Effect Model Selection

Each cell line is simply a random effect: representing samples drawn from a population. Since we are not interested in a particular cell line, but we would like to draw inferences about the population from which the cell lines are drawn we perform a linear mixed effect model including both fixed (treatment) and random effects (cell line). *X* is a matrix with two columns accounting for two covariates.

$$Y_g = X_g\boldsymbol{\alpha}_g + Z_g\boldsymbol{b}_g + \boldsymbol{\epsilon}_g, \quad g = 1, 2, \ldots, N.$$
$$\boldsymbol{b}_g \sim N(0, D), \quad \boldsymbol{\epsilon}_g \sim N(0, R_i),$$

Along with a 0.05 threshold for our FDR we identify 117 potential differentially expressed genes: the simpler linear model also identifies 110 of those.

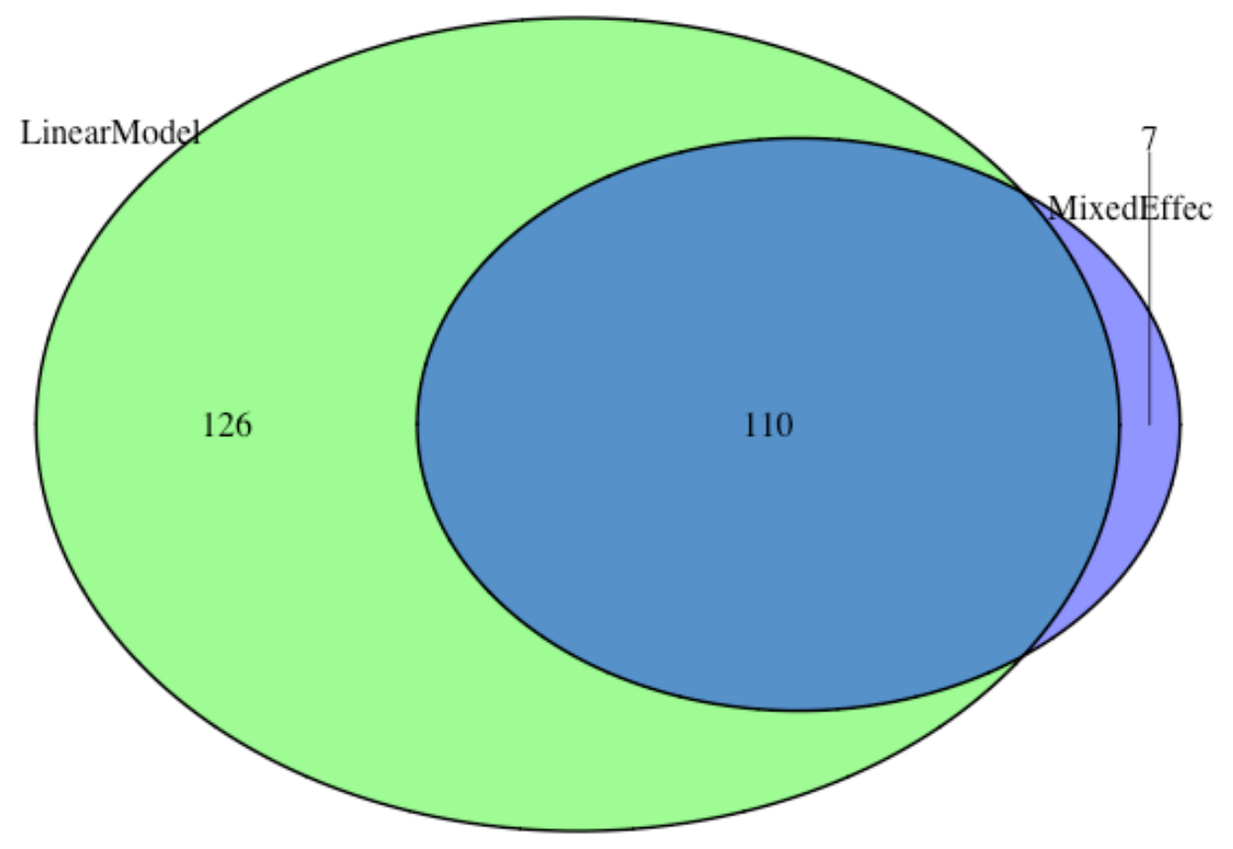

Fig 6. Venn diagram representing the number of differentially expressed genes found using limma (green part) and using a linear mixed effect model (blue part).

Conclusions

Table 2. Summary of results from individual data analysis methods.

Method	Model	Total Probes	D.E. Probes	Total Genes	D.E. Genes
NOTCH: *limma*		54675	789	19852	553
RUNX1:*limma*		70523	236	26757	135
RUNX1: L.M.M.		70523	117	26757	70

Having run differential expression analysis on both data sets, and identified which probes were differentially expressed in each, we return to our original question of interest, namely, what is the relationship between the RUNX1 and NOTCH transcription factors?

To do this, we limit our investigation to only genes that were tested in both experimental conditions. We construct a Null model, by assuming independent differential expression in each analysis. We test out Null model by applying the exact binomial test to the "DE-DE" genes of each pairing.

Table 3. Probabilities under the null model.

Null Prediction: pq genes		Method One	
		D.E.	Boring
Method Two	D.E.	pq	p(1-q)
	Boring	(1-p)q	(1-p)(1-q)

Table 4. Comparing models on RUNX1 data set.

Null Prediction: 0.45 genes		RUNX1 Limma	
		D.E.	Boring
RUNX1 LMM	D.E.	62	3
	Boring	65	18023

Table 5. Contingency table for Notch (limma) gene results vs RUNX1(L.M.M.) gene results.

Null Prediction: 1.89 genes		Notch Limma	
		D.E.	Boring
RUNX1 LMM	D.E.	5	60
	Boring	522	17566

Table 6. Contingency table for Notch(limma) gene results vs RUNX1(limma) gene results.

Null Prediction: 3.69 genes		Notch Limma	
		D.E.	Boring
RUNX1 limma	D.E.	6	3
	Boring	65	18023

Taken together the above provide little to no evidence that interfering with RUNX1 and NOTCH will affect the same genes. This is surprising given the very large overlap between genes which contained binding sites for both transcription factors. Due to asymmetric nature of our data we do not have any direct proof, but inspection of the overlapping genes suggests that RUNX1 and NOTCH1 may co-regulate some aspect of cell proliferation.

Problems and Limitations

Limited sample size and multiple cell lines (in the RUNX1 case), forced us to accept a FDR of 0.05, much higher than we would have liked.

Our data was further complicated by having different gene lists sampled by each experiment which we dealt with by limiting our final intersection analysis to genes that were tested in both experiments. It's possible this should have been done earlier in the analysis.

Finally, during the last days of the project we noticed that certain numbers did not appear to add up to what they were supposed to. Given that the discrepancy was only 9 out of a list of more than 500 genes we suspect that this difference will have had minimal effects on the major results, but given more time, this problem would have been dealt with more thoroughly.

Follow-up studies should focus on acquiring a symmetrical data allowing proper comparisons of matching datasets and DEA in a double knock-down.

References

1. Wang et al. (2011), Genome-wide analysis reveals conserved and divergent features of Notch1/RBPJ binding in human and murine T-lymphoblastic leukemia cells. PNAS 08(36): 1490814913.

2. Okuda T, Nishimura M, Nakao M, Fujita Y (2001). "RUNX1/AML1: a central player in hematopoiesis". Int. J. Hematol. 74 (3): 2527.PMID 11721959.

3. Wang et al.(2014),NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. PNAS 111(2):705-10