Identifying Metabolic Regulators of Human Stem Cell Differentiation using Machine Learning

Cameron Mirhossaini
Jason Cosgrove and Leïla Périe
L'Institut Curie
Paris, France

Introduction

Hematopoietic Stem Cell differentiation serves a key role in maintaining the body's homeostasis. Hematopoietic Stem Cells (HSC's) take many stages to differentiate before resulting in mature cells that give rise to all cells in the blood, many contributing to the body's immune system. The HSC differentiation tree contains four main lineages: lymphoid, myeloid, erythroid, and megakaryocyte. If this process becomes dysregulated at any point in their cell lifetimes, diseases such as lymphoblastic leukemia and myeloid leukemia may arise, which is why it is important that we study how these cells differentiate and which genes play key roles in differentiation.

Under the Périe team at the Curie Institute, we broadly study how the hematopoietic cell tree is shaped using bioinformatics approaches. We hypothesize that hematopoiesis is regulated by metabolism. Through this exploration, our goal was to examine the ways in which metabolic genes could influence HSC differentiation and if it were possible to see these influences in the progenitor stage of the cell.

Methodology

In order to examine the influence of metabolic genes on HSC differentiation, we used Affymatrix microarray data collected from human umbilical cord blood and adult peripheral blood used from the paper "Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis". These samples had already reduced batch effects using the ComBat method available from the sva Bioconductor package and had already been normalized via Robust Multi-Array Average (RMA) normalization. The ComBat method uses mean and variance from each microarray and each gene independently and estimates the batch effects by empirical Bayes in order to target and remove technical sources of variation introduced to the sample from human handling. The RMA method conducts a background correction, standardizes and normalizes the micro-array

data, and then performs quantile normalization, an iterative transformation in which the average from the most highly expressed genes in each sample is taken and replotted until the least expressed gene in each sample is reached. Both methods of data processing are integral to remainder of the analysis.

We then used the Gradient Boost Algorithm to create a sample classifier on the lymphoid, myeloid, and erythroid lineages. Gradient Boost is a machine learning technique that uses decision trees in a step-wise fashion in order to classify samples into categories. The algorithm first creates a single-leafed tree that predicts, without loss of generality, whether or not a sample will be in the lymphoid lineage. It does this by taking the log of odds and converting that value into a probability. With that probability, the algorithm calculates a residual—a metric that shows how accurate the prediction was—for each sample. The algorithm then creates a new tree based off those residual values and repeats until either the residual values are equal to zero or the maximum number of trees—inputted by the used—has been reached. We used Gradient Boost to train samples in the lymphoid, erythroid, and myeloid lineages with only metabolic genes, evaluated the training performance using k-folds cross validation, and then tested the classifier on progenitor cells.

Using Principle Component Analysis (PCA) in conjunction with the given samples that were consistently incorrectly classified using the Gradient Boost Algorithm, we targeted specific samples that expressed genes unlike other cells in that lineage and filtered them out in order to increase the efficiency of our classifier. We will discuss which samples were removed and why in the following section.

Finally, once our classifier achieved a desirable performance—measured by the AUC metric which will be discussed in the following section—we extracted the important metabolic genes that the algorithm used to classify samples by the collecting the top 141 genes which had the highest "gain" from Gradient Boost. Gain refers to the relative contribution each gene made in the classifier

by calculating each gene's contribution in every tree. We used the gain metric rather than the cover or frequency metrics because of its relevance in interpreting the relative performance of each gene rather than the overall performance. Finally, we cross-validated the top genes with a generated heat map and boxplots of gene expressions for each lineage. The pipeline used for this analysis was written using RStudio using the R coding language.

Results

Under the first run of the GB Algorithm, the classifier received an Area Under the Curve (AUC) of .870 with nfolds set equal to 10, nrounds set equal to 16, and the subsample set equal to .8. The classifier predicted 12/13 of the progenitor cells correctly.

	Erythroid	Lymphoid	Meg	Myeloid
Erythroid	31	0	3 MEGA1.3 MEGA1.4 MEGA2.10	1 GRAN1.23
Lymphoid	1 ERY2.7	91	0	6 BASO1.2 BASO1.5 DEND1.5 EOS.17 GRAN1.21 GRAN1.22
Meg	0	0	7	0
Myeloid	1 ERY2.10	1 Pre.BCELL3.46	2 MEGA1.2 MEGA2.5	36

Figure 1; Confusion Matrix, metabolic genes; AUC = .870

Rows represent predicted lineage; columns represent true lineage

True Cell Type	Algorithm Prediction
GMP	Myeloid ✅
GMP	Myeloid ✅
GMP	Myeloid ✅
GMP	Erythroid ⛔
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅
MEP	Erythroid ✅

Figure 2; Gradient Boost Predictions

As shown by Figure 1, we see that the cells that are mostly being misclassified are megakaryocytes and myeloid cells. Basophils, Eosinophils, Granulocytes, and Plasmacytoid Dendritic Cells are the myeloid cells being misclassified as lymphoid cells. Thus, in the next iteration of the algorithm, we decided to remove Plasmacytoid Dendritic Cells because the literature has stated that they should not be classified as myeloid cells. When those samples were filtered out, we received still a similar AUC and decided to further filter out the megakaryocytes because they had been consistently misclassified in every iteration. Additionally, megakaryocytes had the least characteristic cluster in the PC plots compared to the three other lineages. As seen in Figure 3, the green dots representing megakaryocytes are not as close together as the other lineages.

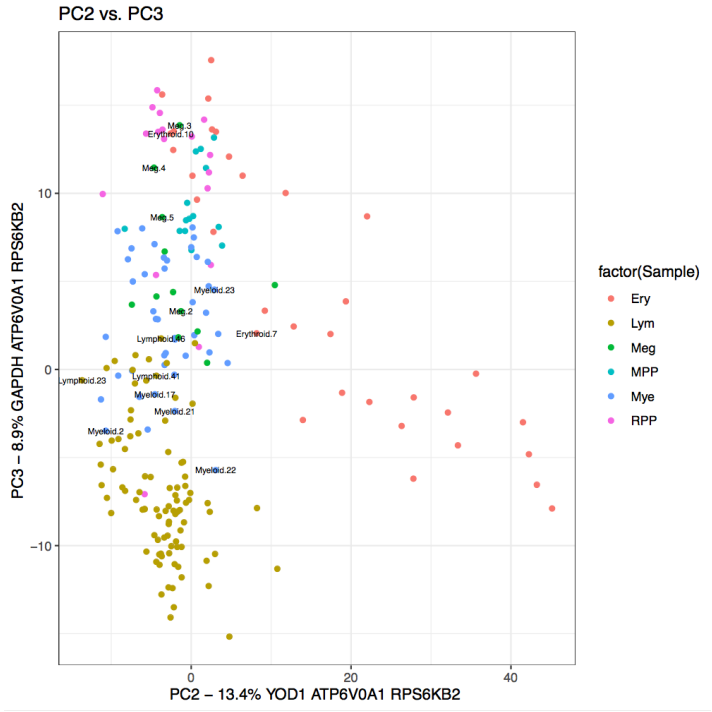

Figure 3; PCA with only metabolic genes, no cells filtered out
**the labeled samples are ones misclassified by GB*

After running the algorithm on all genes, we achieved an AUC of .954 and an AUC of .941 with only metabolic genes, as shown by Figures 4 and 5. The cells that were still being misclassified were Basophils and Granulocytes, all myeloid cells which were classified as lymphoid cells. After this filtering, the classifier still predicted 12/13 progenitors correctly, and the one it predicted incorrectly was the same as the first iteration (a GMP predicted as an erythroid cell), so the prediction chart was identical to Figure 2.

	Erythroid	Lymphoid	Myeloid
Erythroid	32	0	1 GRAN1.23
Lymphoid	0	92	4 BASO1.2 BASO1.3 BASO1.5 GRAN1.21
Myeloid	1 ERY2.10	0	33

Figure 4; Confusion Matrix, all genes; AUC = .954

	Erythroid	Lymphoid	Myeloid
Erythroid	32	0	1 GRAN1.23
Lymphoid	0	90	6 BASO1.2 BASO1.3 BASO1.5 EOS.17 GRAN1.21 GRAN1.22
Myeloid	1 ERY2.10	2 NKa4.35 Pre-BCELL46	31

Figure 5; Confusion Matrix, metabolic genes; AUC = .941

We then extracted 141 genes that were used in the algorithm and targeted the top three genes per lineage. Those top genes were, Glutathione Peroxidase (GPX1), Transaldolase 1 (TALDO1), catalase (CAT) for the erythroid lineage; Aldehyde Dehydrogenase 2 (ALDH2), Major Histocompatibility Complex, Class II, DR Beta 4 (HLA-DRB4), and Hexokinase 2 (HK2) for the myeloid lineage; and Amylase Alpha 1A (AMY1A), Polypeptide N-Acetylgalactosaminyltransferase 11 (GALNT11), and 3'-Phosphoadenosine 5'-Phosphosulfate Synthase 2 (PAPSS2) in the lymphoid lineage. While looking at the clusters created by the kclusters algorithm, we see that the groups it created do not align with how the boxplots show the highest expressed lineage for each gene, thus we cannot use kclustering as a valid way to group genes into lineages. However, from the heat map we do see distinct groupings between the three lineages, thus we know which genes are highly expressed in each lineage respectively, shown by Figure 6. We see that the genes are grouped similarly using both the heat map and the box plots.

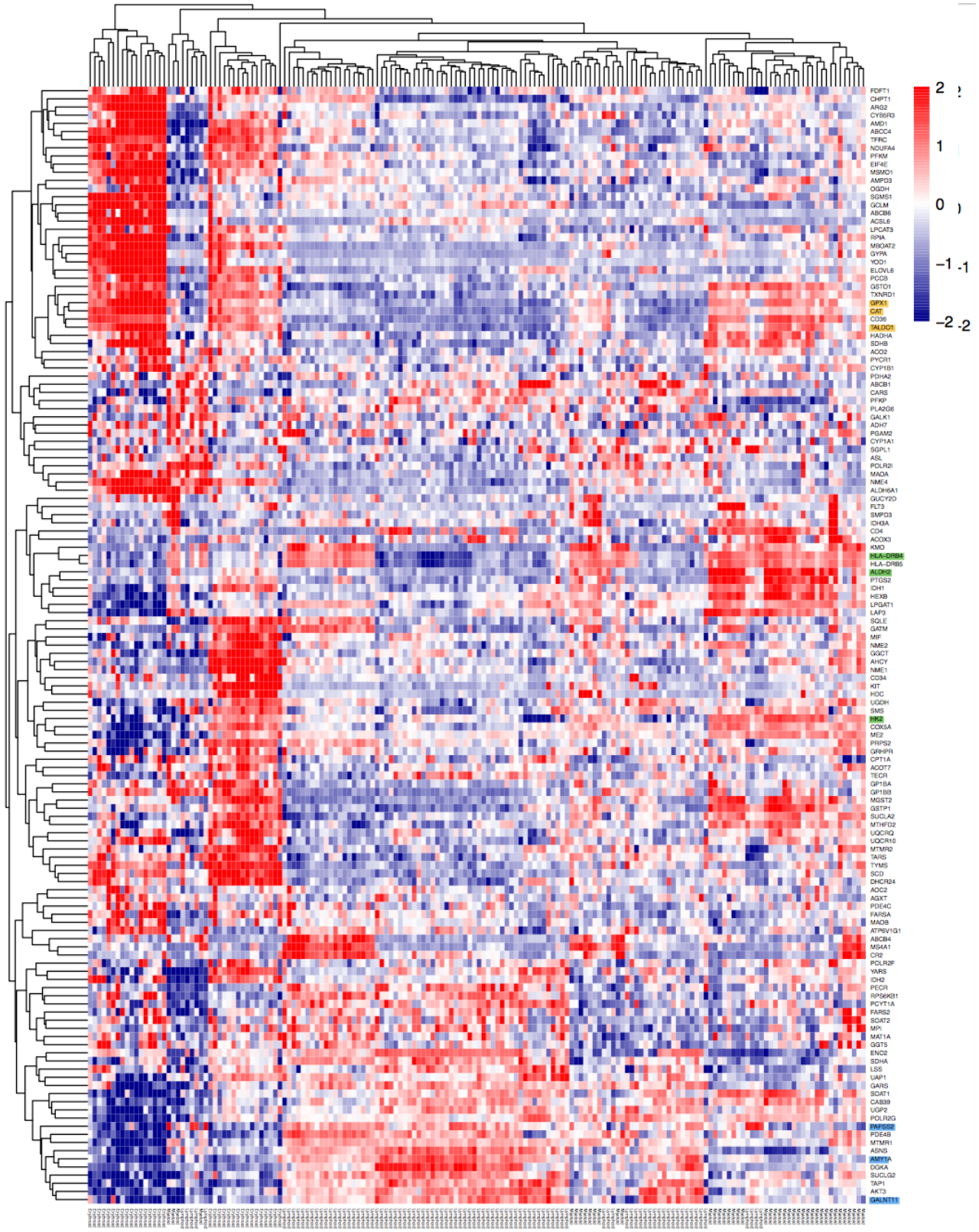

Figure 6; heat map of metabolic genes, PDC's and Megakaryocytes filtered
Yellow genes are highly expressed in erythroid lineage
Green genes are highly expressed in myeloid lineage
Blue genes are highly expressed in Lymphoid lineage

Conclusion

Using the heat map in conjunction with the box plots, we can safely group genes important to each of the three lineages. From the top 141 out of 724 metabolic genes used to classify by the GB algorithm, we arbitrarily selected to examine the top 3 metabolic genes from each lineage, listed above.

The first detail that stood out about the top genes was the fact that almost all of the first 10 genes were ones that were mostly expressed in erythroid cells, and that there were very few top genes highly expressed in myeloid cells, and even fewer in lymphoid cells. From this we can conclude that the algorithm depended on erythroid characteristics.

The top erythroid genes, GPX1, CAT, and TALDO1 are clustered right next to each other in the heat map. GPX1 is responsible for producing the enzyme that catalyze the reduction of organic hydro-peroxides and hydrogen peroxide (H2O2) by glutathione, and thereby protect cells against oxidative damage. According to the gene card database, GPX1 protects hemoglobin in erythrocytes from oxidative breakdown, explaining why it would be highly expressed in the erythroid lineage. CAT, which encodes the heme enzyme catalase, serves to protect cells from the toxic effects of hydrogen peroxide. From these two results, we may hypothesize that erythroid cells have high levels of H_2O_2 production, a toxic reagent that must be handled immediately. In fact, one of the other top hits for erythroid cells not mentioned was TXNRD1, responsible for the production of thioredoxin 1, constitutes the part of a powerful oxidoreductase system, which removes peroxides or H2O2 by using NADPH as an electron donor (source). Finally, TALDO1 encodes the key enzyme of the non-oxidative pentose phosphate pathway providing ribose-5-phosphate for nucleic acid synthesis and NADPH for lipid biosynthesis.

While the top myeloid hits—HLA-DRB4, ALDH2, and HK2—were not as close to each other on the heat map as the erythroid cells, they still are clustered in the same general area. HLA-

DRB4 belongs to the HLA class II beta chain paralogues and plays a central role to the immune system by expressing class II molecules on antigen-presenting cells, such as macrophages and dendritic cells, which are classified as myeloid cells. Aldehyde dehydrogenase is the second enzyme of the major oxidative pathway of alcohol metabolism. HK2 is responsible for hexokinases which phosphorylate glucose to produce glucose-6-phosphate, the first step in most glucose metabolism pathways. We have yet to find specific links between ALDH2, HK2, and myeloid cells.

Finally, the top lymphoid hits, AMY1A, GALNT11, and PAPSS2, were grouped together in the heat map. AMY1A produces an amylase which catalyzes the first step in digestion of dietary starch and glycogen. Salivation and digestive enzymes are often the innate responses to bacteria, responses that lymphoid cells contribute to. GALNT11 is related to O-linked glycosylation, which activates NOTCH1, and Mucin type O-glycan biosynthesis. PAPSS2 mediates 2 steps in the sulfate activation pathway: both ATP sulfurylase and APS kinase activity.