**Analysis of Gene Expression across Time in Hematopoietic Cell Development**

Varun Gupta1\* and Kapil Kanagal1

1 Department of Statistics, Stanford University, 592 Mayfield Avenue, Stanford, CA 94309

\*To whom correspondence should be addressed. Email: vgupta16@stanford.edu

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**I. Background**

The use of single-cell gene expression analysis of embryonic progenitor cells can aid in determining the molecular pathways that control organ development. Single-cell gene expression analysis has already been used to study transcriptome dynamics during embryonic development; analysis has successfully identified clusters of co-expressed genes and their leading hub genes during preimplantation in mouse and human development[[1]](#footnote-1), cell fates from zygote to blastocyst according to particular transcription factors[[2]](#footnote-2), and differences between the gene expression signatures of human epiblast and in vitro hematopoietic stem cells.[[3]](#footnote-3) In particular, blood, or hematopoietic stem cells, can be used to study organ development due to recent technical advances in improving its accessibility and the availability for markers of specific embryonic cell populations[[4]](#footnote-4), as demonstrated by recent identifications of lineage commitment[[5]](#footnote-5) and transcriptional regulatory[[6]](#footnote-6) events in blood. Moreover, blood development occurs early enough during organogenesis, as primitive erythrocytes are required to support the growing embryo.[[7]](#footnote-7)

Given such success in single-cell expression analysis, Moignard et al. inspected hematopoietic development in the mouse embryo through unsupervised analyses of a dataset of 3,934 cells “with blood-forming potential captured at four time points between E7.0 and E8.5” and their levels of expression for 46 different genes.[[8]](#footnote-8) In particular, Moignard et al. used hierarchical (agglomerative) clustering, dimensionality reduction in the form of multidimensional scaling (MDS), diffusion mapping, and principal components analysis (PCA), and boolean network synthesis to reconstruct boolean update rules to analyze the molecular pathways behind development. They demonstrated that the Sox7 factor blocks primitive erythroid development and Sox and Hox factors control expression of the Erg factor. They also identified twenty highly connected transcription factors that stabilize in eight subclusters.

This project aims to develop a more basic understanding of the transcriptome dynamics in the mouse embryo over the given time period. Instead of using sophisticated unsupervised statistical learning methods, this project will focus on conducting tests of homogeneity and independence on the discretized setting of the data to identify factors that contribute the most to heterogeneity in gene expression levels over time and factors that are highly co-dependent during particular periods of time, respectively. Given these insights, our project will then quantify differences in factor expression and dependence from the continuous setting of the data using nonparametric tests such as the Wilcoxon signed-rank test and nonparametric methods such as bootstrapping. These more fundamental statistical tests will enable us to measure transcriptome changes that occur in cells transitioning from one period to the next and eventually differentiating into either endothelial or blood cells without making parametric assumptions. **(INSERT CONCLUSIONS)**

The cells had 5 potential types along 4 different time periods, which represented the different stages in their differentiation. The cells could either be primitive streak (PS), neural plate (NP), head fold (HF), four somite putative blood cells (4SG), or endothelial Flk1+GFP- cells (4SFG-). A diagram of the cell differentiation pathway can be seen below:

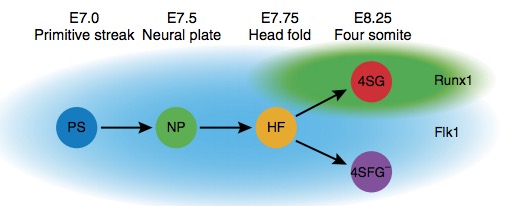


Figure 1: Cell differentiation pathway (Moignard, et. al 2013)

**II. Data and Statistical Methods**

The *continuous* data consists of gene expression levels of 46 genes for 3,934 cells captured at four time points between E7.0 and E8.5. Moignard et al. discretized the *continuous data* by setting the gene expression levels to either 0 or 1 based on whether the expression level exceeded a predetermined threshold specific to each gene. Then, Moignard et al. created the final *discrete* data by isolating those 1,448 cells that could be connected by single gene changes to form one state graph. Thus, the *discrete* data represent the largest connected component of the overall graph of the 3,934 binary states of all of the cells. In sum, the binary single-cell expression states correspond to the on and off patterns of transcription factor expression.

Because we could not make legitimate parametric assumptions about either the *discrete* or *continuous* data, we used nonparametric statistical methods. Typically, researchers use two particular parametric methods for measuring transcriptome changes over time. The first method is the “moderated t-test”, which is a variant of the simple t-test oriented to experiments with few biological replicates. While the simple t-test calculates the variance from the data that is available for each gene, the moderated t-test calculates variance by using information from all of the selected genes.[[9]](#footnote-9) Even though this test can be effective in settings where normality is not exact, the data must still trend towards a normal distribution for it to be accurate. The second method involves using Analysis of Variance (ANOVA) to test linear contrasts of the means of several groups. ANOVA also assumes that the data are distributed normally. Typically, large enough sample sizes can justify the assumption that qPCR data is distributed normally.[[10]](#footnote-10)

However, the size of our samples is not large enough relative to each gene in each different component of the cell differentiation to justify the normal assumption. The data contained ~4,000 cells and was tested for the independence and homogeneity of 46 genes across 5 different cell types. As a result, each gene only had = 17.104 ≈ 20 cells in each given cell type. Therefore, controlling for the genes and types of cells, the sample size is too small to assume a normal distribution for the qPCR gene expression data.[[11]](#footnote-11) Parametric options only yield a higher power if the parametric assumptions are satisfied, and since we cannot assume normality, we defer to nonparametric options.

We use four nonparametric methods to analyze the data. First, we use the Chi-squared test of homogeneity to test differences in gene expression proportions at different time periods. Second, we use the Chi-squared test of independence to test the expressive relationship between any two genes. Third, we compute bootstrapped correlations between particular pairs of genes for the purposes of robustness. Fourth, we use the Wilcoxon signed-rank test to check differences in expression levels of particular genes across time.

First, we use the Chi-squared test of homogeneity to analyze the similarity in gene expression levels at different time periods for a given set of genes. We first create a contingency table that contains genes as the rows and the cell types (time periods) as the columns. Since the Chi-squared test of homogeneity is a nonparametric test, we can determine which genes are most actively contributing to the statistic – and therefore heterogeneity in gene activation across time – without prior knowledge of the distribution. The null hypothesis for this test is that the proportions of gene activation are the same across time. Correspondence analysis can also be used on the *discrete data* to determine which genes contribute the greatest to inhomogeneity.

Second, we use the Chi-squared test of independence on the *discrete data* to determine if two genes in the same cell type are independent of one another. We supplement this independence test by estimating the distribution of correlations between the two genes in a given cell type using the bootstrap method. Association plots can also be used on the *discrete data* to visualize the deviation from independence of particular genes in a given cell type.

Third, we use the Wilcoxon signed-rank test on the *continuous* data to determine whether expression levels of a particular gene across time are drawn from different populations. In effect, this test will replace the t-test to quantify the difference in expression levels for a given gene between cell types. We adjust significance levels by controlling for the familywise error rate, as opposed to the false discovery rate, because we compute fewer than 100 tests, which is orders of magnitude lower than the thousands or millions of tests that necessitate controlling the false discovery rate to maintain a degree of power. We use the Bonferroni correction and the Holm correction to control the familywise error rate.

**III. Results and Discussion**

Chi-Squared Analysis of Homogeneity

First, we converted the data into a readable format and created a contingency table with the rows being the different genes and the columns being the different cell types (an indicator of the time, since the different cell types represent the differentiation from one cell type to another). We parsed and grouped the data into 5 different categories – PS, NP, HF, 4SFGA, and 4SGA – each representing a different cell type. Each of the categories had a recorded score equal to the sum of the number of cells that had expressed a particular gene above its given threshold. As a result, the contingency table looked as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 4SFGA | 4SGA | HF | NP | PS |
| Cbfa2t3h | 182 | 359 | 402 | 190 | 112 |
| Erg | 181 | 29 | 371 | 151 | 81 |
| Ets1 | 241 | 154 | 433 | 210 | 156 |
| Ets2 | 259 | 377 | 436 | 212 | 158 |
| Etv2 | 223 | 69 | 431 | 205 | 152 |
| Etv6 | 259 | 322 | 436 | 212 | 156 |
| Fli1 | 255 | 278 | 430 | 203 | 136 |
| FoxH1 | 250 | 382 | 436 | 212 | 158 |
| FoxO4 | 254 | 383 | 435 | 212 | 158 |
| Gata1 | 1 | 368 | 48 | 76 | 35 |
| Gfi1 | 5 | 18 | 59 | 70 | 26 |
| Gfi1b | 15 | 379 | 128 | 104 | 52 |
| Hhex | 244 | 177 | 416 | 192 | 122 |
| HoxB2 | 1 | 0 | 0 | 0 | 0 |
| HoxB4 | 160 | 10 | 297 | 127 | 62 |
| HoxD8 | 0 | 0 | 0 | 0 | 0 |
| Ikaros | 15 | 379 | 208 | 151 | 74 |
| Ldb1 | 253 | 379 | 432 | 211 | 158 |
| Lmo2 | 118 | 141 | 295 | 149 | 93 |
| Lyl1 | 214 | 353 | 393 | 178 | 101 |
| Mecom | 133 | 6 | 194 | 38 | 27 |
| Meis1 | 211 | 84 | 422 | 175 | 125 |
| Mitf | 28 | 152 | 71 | 32 | 28 |
| Myb | 34 | 320 | 218 | 138 | 85 |
| Nfe2 | 9 | 381 | 119 | 103 | 51 |
| Notch1 | 251 | 29 | 427 | 199 | 148 |
| Runx1 | 153 | 383 | 363 | 191 | 146 |
| Sfpi1 | 78 | 182 | 243 | 137 | 73 |
| Sox17 | 205 | 1 | 284 | 72 | 35 |
| Sox7 | 247 | 19 | 420 | 195 | 127 |
| Tal1 | 249 | 383 | 431 | 204 | 133 |
| Tbx20 | 170 | 16 | 377 | 184 | 150 |
| Tbx3 | 135 | 32 | 223 | 135 | 131 |

Because the contingency matrix contained zeros in the binary expression data for HoxB2 and HoxD8, we removed them from the table to perform an accurate Chi-squared analysis. Their lack of expression across the different cell types might be due to biological regulation by other genes in the analysis. More importantly, the Chi-squared statistic cannot incorporate these genes because the statistic would be undefined otherwise.

The modified contingency matrix was then used for the Chi-squared test of homogeneity. An overall test of homogeneity yields the following result:

X-squared = 4873.5, df = 120, p-value < 2.2e-16

As expected, we reject the original null hypothesis that the distribution of gene expression levels across time are the same for a given gene in favor of the alternative hypothesis that the gene expression levels for a given gene are not constant across time. Intuitively, different cell types would naturally require different genes to be activated in order to correctly function. In fact, the activation and deactivation of particular genes drives the differentiation process.

To better visualize and analyze which specific genes were contributing the most to inhomogeneity, we performed a correspondence analysis. We also recomputed the Chi-squared statistics after removing each gene individually for the same purpose. To control for the final differentiation (into either 4SG or 4SFG), we performed a Chi-Squared Analysis on the data after removing these two columns from the table. This would tell us which genes were most important in differentiation prior to the final differentiation, and which genes were most important in determining if an HF cell became an endothelial progenitor or an erythroid progenitor. By plotting the Chi-Squared statistics in the manner described above, we yielded the following two plots:

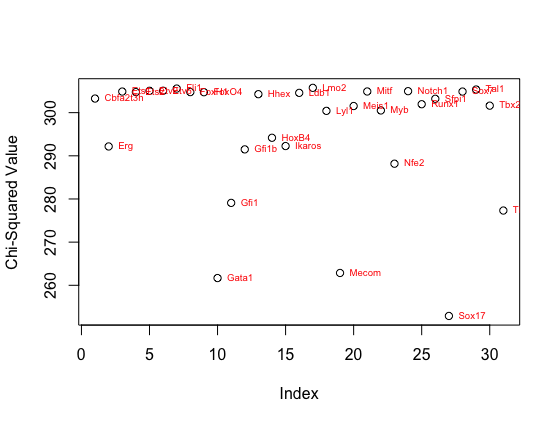


Figure 2: This graph represents the Chi-squared statistics of the table after removing the given genes *prior* to the final differentiation into an endothelial cell or an erythroid cell.



Figure 3: This graph represents the Chi-Squared Statistics of the table after removing the given genes *after* the final differentiation into an endothelial cell or an erythroid cell.

Note the overall difference in the range of Chi-squared statistic values between the two graphs, which indicates that the vast majority of difference in genetic expression is due to the final step of development into either erythroid or endothelial progenitors. With these two graphs, we determined which genes have relatively the same level of impact on the Chi-squared statistics both before and after the final differentiation. We labelled these genes as background genes that did not contribute heavily to inhomogeneity. We then performed the rest of the analysis focusing only on the genes that contributed significantly to the inhomogeneity. We found that Cbfa2t3h, Ets2, Etv6, Fli1, FoxH1, Hhex, Ikaros, Ldb1, Lmo2, Runx1, Sfpi1, and Tal1 contributed the least to the inhomogeneity in the Chi-Squared statistic and as such, removed them from our data table before commencing with our correspondence analysis.

To determine specifically which genes contributed the most to inhomogeneity, we used correspondence analysis on the remaining genes, which yielded the following graph:

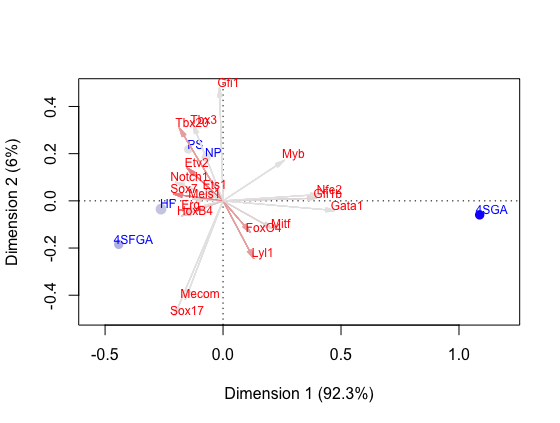


Figure 4: Correspondence Analysis Graph of Genes using Binary Data to determine which genes contribute the most to inhomogeneity.

We can make three observations from the correspondence analysis plot, which indicates that the first dimension captures over 92 percent of the variance in the data. First, the first three time periods – and thus the first three cell stages in hematopoietic development – do not have very different gene activation patterns, even if the expression levels of those activation patterns are different. This is because the three stages – PS, NP, and HF – are clustered close together in the correspondence analysis.

Second, Mecom, Sox17, Nfe2, Gfi1b, and Gata1 contribute a significant amount to the heterogeneity in activation patterns between 4SFGA – the endothelial progenitor – and the rest of the types, particularly 4SGA, the erythroid progenitor. These results make sense. Sox17 is one of the Sox and Hox factors that Moignard et al. note controls the expression of Erg, which is critical to the development of the endothelium. Mecom has been shown to prevent the terminal differentiation of hematopoietic stem cells into erythroid cells. Thus, 4SFGA is defined by higher activation of Sox17 and Mecom. On the other hand, Nfe2, Gfi1b, and Gata1 complex together to form some of the “zinc finger” proteins necessary for erythroid function, which explains why 4SGA is defined by higher activation of those genes.

Third, activation of several other genes likely push the first three cell stages towards an endothelial fate, as opposed to an erythroid fate. Note that because the first dimension captures the vast majority of the variance in the data, we give more weight to the position of the genes and the cell types with respect to the first dimension than with the second. As a result, the activation of genes to the left of the line at 0, which includes Sox7, Erg, Notch1, and the Etv/Ets family, pushes cells towards an endothelial fate. This matches model predictions, as Sox7 overexpression is known to block progression towards an erythroid fate, and Erg, Etv, and Ets – members of the Ets transcription actor family – reflect endothelial development. Moreover, the endothelium has recently been identified as the target for Notch1 signaling (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633594/>). Overall, the Chi-squared analysis of homogeneity confirms model predictions of which activations are necessary for development towards an erythroid or endothelial fate.

Chi-Squared Analysis of Independence and Bootstrapped Correlations

We then performed a Chi-squared test of independence to determine particular pairs of genes that operate independently or jointly. Using a correlation matrix, we identified pairs of genes that either exhibited high degrees of correlation or almost no correlation at all. We selected 3 highly correlated genes and 1 low/uncorrelated gene to examine how high correlation and expression dependence were linked to one another. First, we selected Sox7 and HoxB4 because Sox7 blocks preliminary erythroid development, and HoxB4 is one of the Hox factors that directly regulate the expression of the hematopoietic stem-cell regulator, Erg. As a result, both Sox7 and HoxB4 are known to be endothelial genes. Second, we selected Gata1 and Gfi1b, which are both hematopoietic genes that are known to complex together in “zinc fingers.” Third, we selected Notch1 and Ets1 as random controls, since Moignard et al. do not specifically examine the pair. Fourth, we selected Myb and Erg to check for independence, as Myb is a hematopoietic gene, while Erg is an endothelial gene. For tests of independence, we obtained the corresponding Chi-squared statistics and association plots:

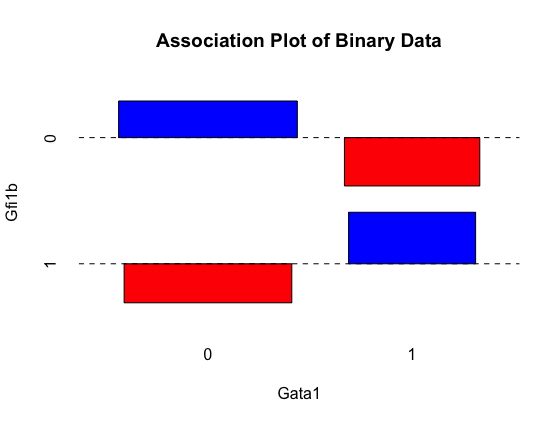
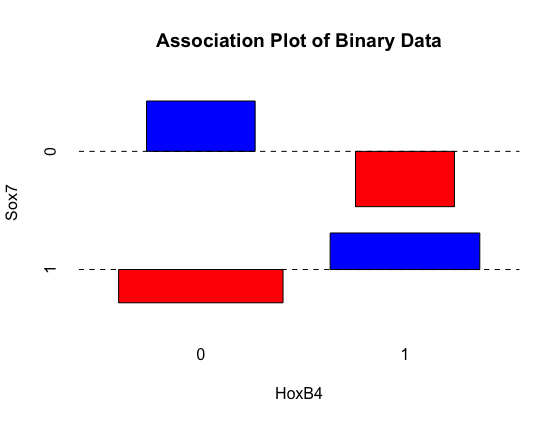
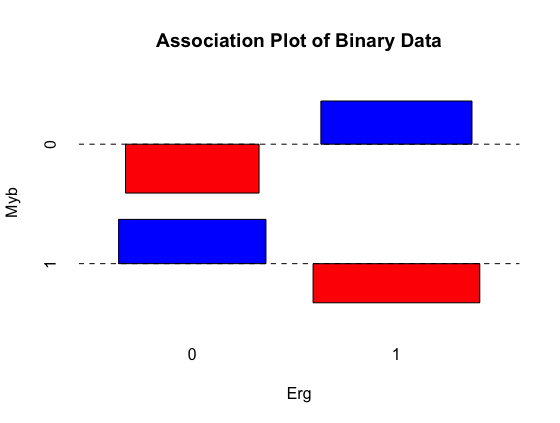


Figure 5: Association plot of Gata1-Gfi1b. X-squared = 829.81, df = 1, p-value < 2.2e-16

Figure 6: Association plot of HoxB4-Sox7. X-squared = 366.73, df = 1, p-value < 2.2e-16

For the Gata1-Gfi1b pair, we see that Gata1 and Gfi1b have more 0,0 and 1,1 associations than would be expected under the null distribution, with fewer 1,0 and 0,1 associations, for any given cell type. The p-value of the Chi-squared test is essentially 0, which indicates that they tend to be jointly activated and expressed, as model predictions confirm.

We see a similar pattern in the HoxB4 and Sox7 pair, in which the two genes have a high 0,0 and 1,1 association and a low 0,1 and 1,0 association. Here, the p-value also tends to 0, implying that the activation of the two genes is likely dependent or joint.

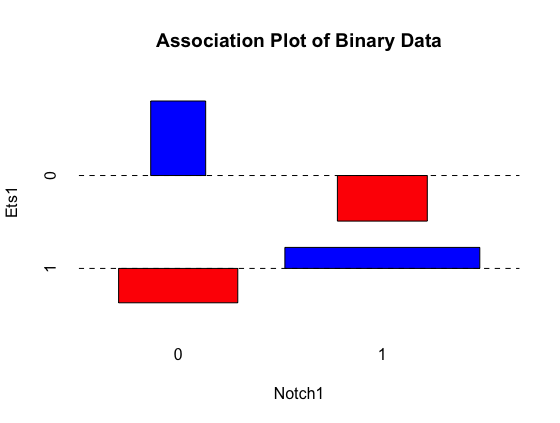


Figure 8: Association plot of Erg-Myb. X-squared = 10.103, df = 1, p-value = 0.00148

Figure 7: Association plot of Notch1-Ets1. X-squared = 552.47, df = 1, p-value < 2.2e-16

We see a similar trend with the association plot between Notch1 and Ets1, showing a dependency between the expression levels of these two genes with both the plot and the small p-value. This plot has different widths for the 0,0 and 1,1 associations because (EXPLAIN WHY THEY HAVE DIFFERENT WIDTHS)

This association plot and the corresponding p-value reveal that the expression levels or Erg and Myb are likely independent of one another, as the plot shows higher 1,0 and 0,1 expression levels for the two genes than would we expected under the null hypothesis.

We used a Bonferroni correction as a conservative correction for the p-value with which to determine significance. There were ) possible combinations which yielded a significance threshold of = 0.0001. As a result, we found that Gata1-Gfi1b, HoxB4-Sox7, and Notch1-Ets1 all had p-values well below this threshold implying that we could reject the null hypothesis that these pairs of genes were expressed independently from one another. However, Erg-Myb had a p-value above the threshold meaning we fail to reject the null hypothesis. As a result, we can see from the correlation graph that these two genes have a near 0 correlation implying that they might indeed be expressed independently of one another, since we fail say that these genes are necessarily dependent. These results are also visible from the association plots.

Furthermore, we used the bootstrap method to estimate the correlations between the various genes in each of the cell types. (Bootstrap)

Wilcoxon Signed-Rank Test

(Wilcoxon Results and Discussion)

**IV. Future Work**

(More future work)

**V. Works Cited**

Baron, M. H., J. Isern, and S. T. Fraser. “The Embryonic Origins of Erythropoiesis in Mammals.” *Blood* 119, no. 21 (May 24, 2012): 4828–37. doi:10.1182/blood-2012-01-153486.

Goni, Ramon, Patricia Garcia, and Sylvain Foissac. “The qPCR Data Statistical Analysis.” *Integromics*, September 2009.

Guo, Guoji, Mikael Huss, Guo Qing Tong, Chaoyang Wang, Li Li Sun, Neil D. Clarke, and Paul Robson. “Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst.” *Developmental Cell* 18, no. 4 (April 2010): 675–85. doi:10.1016/j.devcel.2010.02.012.

Mahajan, Maya M., Betty Cheng, Ashley I. Beyer, Usha S. Mulvaney, Matt B. Wilkinson, Marina E. Fomin, and Marcus O. Muench. “A Quantitative Assessment of the Content of Hematopoietic Stem Cells in Mouse and Human Endosteal-Bone Marrow: A Simple and Rapid Method for the Isolation of Mouse Central Bone Marrow.” *BMC Hematology* 15, no. 1 (December 2015). doi:10.1186/s12878-015-0031-7.

Moignard, Victoria, Iain C. Macaulay, Gemma Swiers, Florian Buettner, Judith Schütte, Fernando J. Calero-Nieto, Sarah Kinston, et al. “Characterization of Transcriptional Networks in Blood Stem and Progenitor Cells Using High-Throughput Single-Cell Gene Expression Analysis.” *Nature Cell Biology* 15, no. 4 (March 24, 2013): 363–72. doi:10.1038/ncb2709.

Moignard, Victoria, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, et al. “Decoding the Regulatory Network of Early Blood Development from Single-Cell Gene Expression Measurements.” *Nature Biotechnology* 33, no. 3 (February 9, 2015): 269–76. doi:10.1038/nbt.3154.

Pina, Cristina, Cristina Fugazza, Alex J. Tipping, John Brown, Shamit Soneji, Jose Teles, Carsten Peterson, and Tariq Enver. “Inferring Rules of Lineage Commitment in Haematopoiesis.” *Nature Cell Biology* 14, no. 3 (February 19, 2012): 287–94. doi:10.1038/ncb2442.

Smyth, Gordon K. “Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.” *Statistical Applications in Genetics and Molecular Biology* 3, no. 1 (January 12, 2004): 1–25. doi:10.2202/1544-6115.1027.

Xue, Zhigang, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-yan Jiang, Yun Feng, Zhenshan Liu, et al. “Genetic Programs in Human and Mouse Early Embryos Revealed by Single-Cell RNA Sequencing.” *Nature* 500, no. 7464 (July 28, 2013): 593–97. doi:10.1038/nature12364.

Yan, Liying, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, et al. “Single-Cell RNA-Seq Profiling of Human Preimplantation Embryos and Embryonic Stem Cells.” *Nature Structural & Molecular Biology* 20, no. 9 (August 11, 2013): 1131–39. doi:10.1038/nsmb.2660.

1. Xue et al., “Genetic Programs in Human and Mouse Early Embryos Revealed by Single-Cell RNA Sequencing.” [↑](#footnote-ref-1)
2. Guo et al., “Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst.” [↑](#footnote-ref-2)
3. Yan et al., “Single-Cell RNA-Seq Profiling of Human Preimplantation Embryos and Embryonic Stem Cells.” [↑](#footnote-ref-3)
4. Mahajan et al., “A Quantitative Assessment of the Content of Hematopoietic Stem Cells in Mouse and Human Endosteal-Bone Marrow.” [↑](#footnote-ref-4)
5. Pina et al., “Inferring Rules of Lineage Commitment in Haematopoiesis.” [↑](#footnote-ref-5)
6. Moignard et al., “Characterization of Transcriptional Networks in Blood Stem and Progenitor Cells Using High-Throughput Single-Cell Gene Expression Analysis.” [↑](#footnote-ref-6)
7. Baron, Isern, and Fraser, “The Embryonic Origins of Erythropoiesis in Mammals.” [↑](#footnote-ref-7)
8. Moignard et al., “Decoding the Regulatory Network of Early Blood Development from Single-Cell Gene Expression Measurements.” [↑](#footnote-ref-8)
9. Smyth, “Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.” [↑](#footnote-ref-9)
10. Goni, Garcia, and Foissac, “The qPCR Data Statistical Analysis.” [↑](#footnote-ref-10)
11. Ibid. [↑](#footnote-ref-11)