UVIC
UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL
DE CATALUNYA

Master of Science in Omics Data Analysis

Master Thesis

Dissecting the genetic signatures of a multipotent hematopoietic progenitor's subpopulations

by

Llucia Inès Albertí Servera

Supervisor: Robert Ivánek, University of Basel

Guarantor: Jordi Villà, University of VIC

Department of Systems Biology

University of Vic – Central University of Catalonia

September 12, 2017

Acknowledgements

This project wouldn't have been possible without Dr. Robert Ivanek's guidance. You inspired me and generated my first interest in bioinformatics. Thank you for your constant help in data analysis, for forcing me to use bookdown and version control (GitHub) to generate this thesis and for your patience with all the editing work.

I also want to thank my former PhD mentor, prof Antonius Rolink, for all his scientific input and for encouraging me to undertake this master while providing me the time and necessary resources.

Thanks to this master in "Omics Data Analysis" I have acquired computational data analysis skills into a wide range of omics analysis. This will be extremely useful for my future professional work in bioinformatics. For that, I want to thank all the teachers of this master.

Last but not least, a big thanks goes to my boyfriend, Francisco Camacho, for his comprehension, constant support, and injeccion of energy to complete this master.

Abstract

In hematopoiesis and other developmental systems, there is an active debate regarding the heterogeneity of apparently phenotypically homogeneous progenitors having different lineage potentials. The host laboratory has previously characterized a B220^{+} c-Kit^{int} CD19^{-} and NK1.1^{-} uncommitted and multipotent hematopoietic progenitor with combined lymphoid and myeloid differentiation capacity that was called Early Progenitor with Lymphoid and Myeloid potential (EPLM). Under physiological conditions, EPLM was mainly described as a B-cell progenitor. More recently, with flow cytometry analysis, EPLM has been fractionated into at least four subpopulations based on the expression of Ly6D, SiglecH and CD11c cell surface markers, thus revealing phenotypic heterogeneity. The question remains whether these subpopulations are still multipotent or, instead, biased towards distinct hematopoietic lineages.

In this project, I have further studied the two EPLM subpopulations, namely Ly6D^{+} and triple negative (TN), that could possess B-cell developmental potential and/or be multipotent. The main goal was to elucidate if the phenotypic heterogeneity (differential expression of Ly6D) would reflect distinct and biologically meaningful molecular signatures that could indicate, for instance, different developmental potentials for the subpopulations. A second goal was to identify a potential EPLM fraction containing most of the B-cell differentiation capacity and being the precursor of the first B-cell committed stage, the pro-B cells. To address the previous goals, I performed population RNA sequencing and carried a detail analysis of the molecular signatures of the two EPLM subsets while comparing them with the transcriptome profile of the first B-cell committed progenitor, the pro-B. The results obtained in this project demonstrate that heterogeneous expression of Ly6D can be used to discriminate

among EPLM subpopulations that have distinct genetic signatures. Whereas the Ly6D$^+$ cells are lymphoid primed and have a strong B-cell genetic signature, the TN cells are myeloid primed. Therefore, EPLM is not only phenotypically but also genetically heterogeneous. I speculate that the lympho-myeloid developmental potential observed for the whole EPLM population could be constrained within the Ly6D$^+$ and TN fractions, respectively. Moreover, the Ly6D$^+$ cells, which have a closer transcriptome profile to pro-B than when the TN cells are compared with pro-B, could be the direct precursor of the first B-cell committed stage. Ultimately, this master project sets the basis for further functional experiments to resolve the developmental potentials of the EPLM subsets.

Contents

List of Tables

List of Figures

Abbreviations

BCR	B-cell receptor
Blk	B lymphocyte kinase
bp	base pair
Ccr2	chemokine receptor 2
CD	cluster of differentiation antigen
Cebpa	CCAAT/enhancer-binding protein alpha
Ciita	class II transactivator
CLP	common lymphoid progenitor
CMP	common myeloid progenitor
CPM	counts per million
Csf1r	colony-stimulating factor 1 receptor
Ctla	cytotoxic T-lymphocyte-associate antigen
Cts	cathepsin
DEG	differentially expressed gene
DL1	Delta-like 1
DNA	deoxyribonucleic acid
dsDNA	double strand DNA
Ebf1	early B-cell factor 1
EDTA	ethylenediaminetetraacetic acid
EPLM	early progenitor with lymphoid and myeloid potential
FACS	fluorescence-activated cell sorting
FC	fold change

FcR	fragment crystallizable receptor
FDR	false discovery rate
FGF4	fibroblast growth factor-4
Flt3	Fms-like tyrosine kinase 3
Flt3L	Flt3 ligand
Flt3Ltg	human Flt3l transgenic
FPKM	fragments per kilobase of transcript per million mapped reads
GO	gene ontology
Iga	immunoglobulin alfa
Igll1	immunoglobulin lambda-like polypeptide 1
IL	interleukin
IL7r	IL7 receptor
Lax	linker for activation of X cells
Lck	leukocyte C-terminal Src kinase
Ly	lymphocyte antigen
Mpo	myeloperoxidase
mRNA	messenger RNA
NGS	next generation sequencing
PAM	portioning around medoids
Pax5	Paired box protein 5
PCA	principal component analysis
PCR	polymerase chain reaction
Prtn3	proteinaise 3
QS	quality score
Rag	recombination-activating gene
RIN	RNA integrity number
RNA	ribonucleic acid
RNA-seq	RNA sequencing
sd	standard deviation
SEM	standard error of the mean

SiglecH	Sialic acid binding Ig-like lectin H
Sla2	Src-like-adapter protein
Stat5	Signal transducer and activator of transcription 5
TCR	T-cell receptor
Tlr	toll-like receptor
TN	triple negative
Trat1	T-cell receptor-associated transmembrane adapter 1
UMI	unique molecular identifiers
vs	versus
WT	wild type
Zap70	70 kDa zeta-chain associated protein

Chapter 1

Introduction

For decades, immunology has relied on the expression of limited cell surface markers to phenotypically characterize and classify immune cells with the use of flow cytometry analysis. For instance, the phenotype of the common lymphoid progenitor (CLP) is Lin$^-$ IL-7R$^+$ Thy-1$^-$ Sca-1^{lo} c-Kit(CD117)^{lo} (Motonari Kondo and Akashi 1997), whereas the first cell that under physiological conditions is committed to the B-cell lineage, the pro-B, is phenotypically characterized by the expression of CD19 and c-Kit and genotypically by the immunoglobulin heavy (IgH) chain loci being both D_H-J_H rearranged (Boekel E. and Rolink 1995).

While traditional flow cytometer-based technologies have been and still are very important in immunology, the explosion of high throughput technologies, such as RNA sequencing, now permit to quantify the expression of thousands of genes in parallel and in an unbiased manner. This better characterizes a cell population and enables the identification of molecular differences otherwise masked when only analysing few genes. Moreover, it has the potential to identify new and more robust markers to describe a population.

In 2005, a new progenitor cell was described being phenotypically closely related to the CLP but with the particularity that, apart of being able to differentiate into lymphoid cells, could also give rise to myeloid cells and was therefore called EPLM (Early Progenitor with Lymphoid and Myeloid developmental potential). EPLM cells were identified as B220$^+$ c-Kit^{int} CD19$^-$ and NK1.1$^-$ representing 0.2% of all nucleated bone marrow cells in wild

type (WT) mice (Balcicunaite G. and Rolink 2005). As mentioned before, in terms of phenotype, this progenitor is closely related to the CLP with the marked difference of B220 expression, EPLM being B220$^+$ whereas CLP are B220$^-$, and partially overlaps with the so-called Fraction A cells identified by Hardy and co-workers (Li YS. and Hardy 1996). Limiting dilution analysis of EPLMs cultured together with stromal cells and addition of appropriate cytokines, enabled the quantification of *in vitro* B, T and, myeloid precursor frequencies. EPLMs showed strong B-cell developmental potential and strong-to-moderate differentiation potential for T cells and myeloid cells (mostly macrophages). Therefore, this suggested that under physiological conditions the developmental fate of EPLM is mainly to become B cells. Reconstitution assays in order to assess the EPLM's *in vivo* developmental potentials revealed their ability to transiently reconstitute both B- and T-cell compartments in sublethally irradiated *Rag2*-deficient mice.

In line with the description of individual progenitor cells having multiple lineage potentials, there is an increasing debate regarding their heterogeneity. In fact, there is accumulating evidence that multipotent hematopoietic progenitors identified over the years are more heterogeneous than previously thought. For the common myeloid progenitor (CMP), differential cell-surface expression of Slamf1 (CD150), Endoglin (CD105), and Itga2b (CD41) was shown to be correlated with individual developmentally restricted lineage potentials for the granulocyte/macrophage, erythroid, and megakaryocytic lineages respectively (Cornelis J.H. Pronk and Bryder 2007). Regarding the CLPs, they were further sub-grouped after their initial description. For instance, re-analysis of the CLP compartment revealed absolute lymphoid multipotentiality only within the Flt3$^+$ proportion (Karsunky H 2008). Later on, Ly6D was identified and used to assign B-cell restricted progenitors within the Flt3$^+$ CLP population. Ly6D$^+$ CLPs were termed BLPs (B-cell biased lymphoid progenitor), whereas Ly6D$^-$ CLPs were named ALPs (all lymphoid progenitors), since they retain T- as well as NK-cell potential (Inlay MA 2009).

This context raised the need to revise the EPLM progenitor population with the aim of determining whether it is a homogeneous multipotent population or a mixture of individual lineage-restricted cells. Previous data of the host laboratory shows that EPLM express

heterogeneous levels of Ly6D, SiglecH and CD11c cell surface markers. Therefore, phenotypically, EPLM is an heterogeneous population that can be further divided into at least four subpopulations **(Figure 1.1A)**. Now the question arises whether these subpopulations are also genetically distinct and if they maintain the developmental potentials initially described by the whole EPLM population or, instead, correlate with more constrained lineage potentials.

In this present study, I have characterized in details two of the EPLM subpopulations at the molecular level: the Ly6D ($Ly6D^+$) single positive subset and the triple negative (TN) subset, which lacks expression of the three cell surface markers Ly6D, SiglecH and CD11c. The main goal of the project described in this thesis has been to elucidate if Ly6D discriminates two EPLM fractions that are truly molecularly distinct and, therefore, that could be functionally distinct. For that, I have performed bulk RNA sequencing (RNA-seq) and carried a detailed analysis of their expression profiles. Moreover, since EPLM was mainly described as a B-cell progenitor, a second goal was to identify a potential fraction mostly retaining the B-cell developmental potential and being the precursor of the first B-cell committed stage, the pro-B cells (Boekel E. and Rolink 1995). Therefore, I also compared the gene expression profile of the two EPLM subsets with the B-cell progenitor population pro-B.

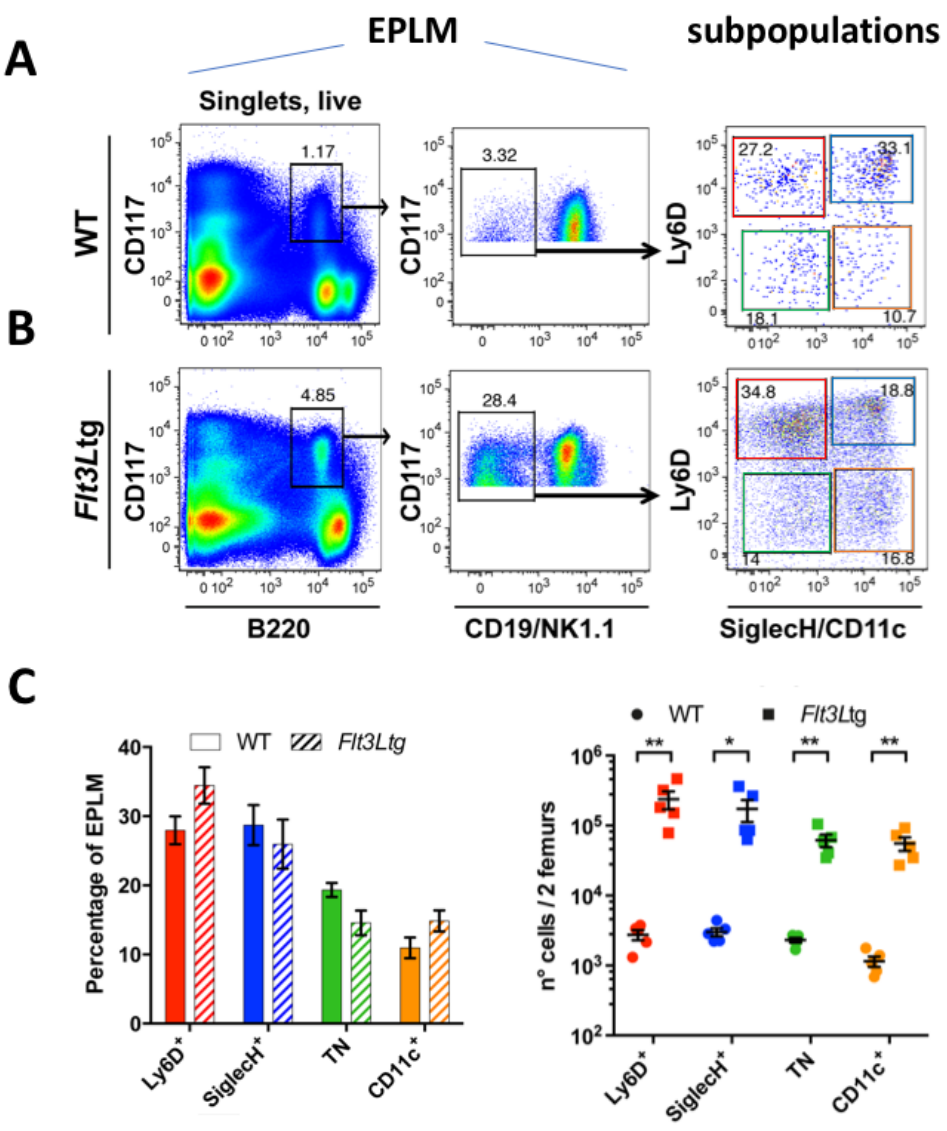

Figure 1.1: Gating strategy of EPLM progenitor population (B220$^+$ c-Kit^{int} CD19$^-$ and NK1.1$^-$) and heterogeneous expression of Ly6D, SiglecH and CD11c. (A,B) Representative FACS plot showing the gating strategy to identify EPLM and their subpopulations. Cells were sorted from the bone marrow of WT (A) and *Flt3L*tg (B) mice. (C) Summary of EPLM subpopulations as percentages (left) or numbers (right) from WT (n=5) and *Flt3L*tg (n=5) mice. Shown as mean ± SEM. Two-tailed unpaired Student t tests, $*P \leq 0.05$, $**P \leq 0.01$.

Chapter 2

Results

In order to characterize the EPLM subpopulations Ly6D$^+$ and TN at the molecular level, I performed bulk gene expression profiling (RNA-seq) of these two populations as well as CD117$^+$ CD19$^+$ pro-B cells from *Flt3L*tg mice. The latter population was included as an already B-cell lineage committed bone marrow population and thus downstream of EPLMs. For the project described in this thesis, I made use of a mouse model generated in the host laboratory, the *Flt3L*tg mouse line. These mice show the same EPLM subpopulations in comparable relative frequencies as WT mice but with a significant increase of about two orders of magnitude for each subset **(Figure 1.1)**. Therefore, the *Flt3L*tg mouse is a valid tool to isolate EPLM subpopulations in large numbers and perform functional and molecular experiments.

2.1 Adquisition of the samples, sequencing data and quality control

Ly6D$^+$ and TN EPLM subpopulations as well as pro-B cells were sorted from femurs of 2-pooled male *Flt3L*tg mice (6 to 8 weeks of age) as in **(Figure 1.1B)**. After each sort, cells were centrifuged, resuspended in 0.5ml of TRIzol reagent and stored at -80°C for later total RNA extraction.

Total RNA was extracted from *ex-vivo* sorted samples using TRIzol-based method (Chomczynski and Sacchi 1986, Chomczynski and Sacchi (2006)). Briefly, 1x10^5 to 3x10^5 cells were lysed in 0.5ml of TRIzol reagent and 0.1ml of chloroform was added per 0.5ml TRI reagent. After incubation and centrifugation for phase separation, the aqueous phase containing the RNA was recovered and mixed with isopropanol in a 1:1 ratio for RNA precipitation. Following 15min incubation and centrifugation, the supernatant was discarded while the RNA pellet was first washed with 75% ethanol and subsequently resuspended with 20μl of DEPC treated water. Concentration and 260/280 purity ratio was initially determined using NanoDrop 1000 Spectrophotometer (Witec AG). Selected RNA samples were stored at -80°C for later usage.

The isolated RNA, 500ng per sample, was sent to the Genomics Facility at the D-BSSE (Basel) for quality control, library preparation and sequencing. Quality and level of degradation of the extracted RNA was assessed with RNA integrity number (RIN) assigned by the Agilent 2100 Bioanalyzer instrument using either the Nano or the Pico Agilent RNA 6000 kit (Agilent Technologies). Samples with a RIN value over 8 and presenting clean peaks were considered for further analysis. The RNA quantity was measured by the Infinite M1000 PRO - Tecan instrument using the Quant-iT RiboGreen RNA Assay Kit.

For the generation of sequencing libraries, the TruSeq Stranded mRNA LT Sample Preparation kit was used following the manufacturer's guide (Tatiana Borodina and Sultan 2011). Briefly, the polyA containing mRNA molecules were purified using poly-T oligo attached magnetic beads and subsequently fragmented using divalent cations under elevated temperatures. Afterwards, the RNA fragments were copied into first strand cDNA using reverse transcriptase and random primers. Strand-specificity information was achieved by replacing dTTP with dUTP during the second strand cDNA synthesis. To prevent self-ligation of the double-stranded cDNA, the 3' ends of the blunt fragments were adenylated followed by ligation of barcoded adapters suitable for Illumina-based sequencing. The product was subjected to 15 cycles of PCR amplification. Size and purity of the library fragments was assessed by the Fragment Analyzer using the NGS Fragment 1-6000bp method (average fragment size 321bp, sd 20.36), while quantification was done with Quant-iT PicoGreen® dsDNA Assay Kit; Tecan instrument.

Indexed DNA libraries were pooled in equal volumes and loaded on one NextSeq 500 High Output flow cell (Illumina). Single-end sequencing was performed on the Illumina NextSeq™ 500 Sequencing System (D-BSSE, Basel) for 81 cycles. Subsequently, the Genomics Facility performed de-multiplexing with the Illumina pipeline and reads were exported in the FastQ format. Quality control of the sequenced data was performed using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, version 0.11.5). All samples comprised high number of reads (> 21 millions) with median Quality Score (QS) of 35, a GC content distribution equivalent to the expected theoretical distribution (~52%) of mouse genome, a sequence duplication level typical for RNA-seq samples, and no adapter content present (no need for trimming of reads). **Figure 2.1** shows a representative example.

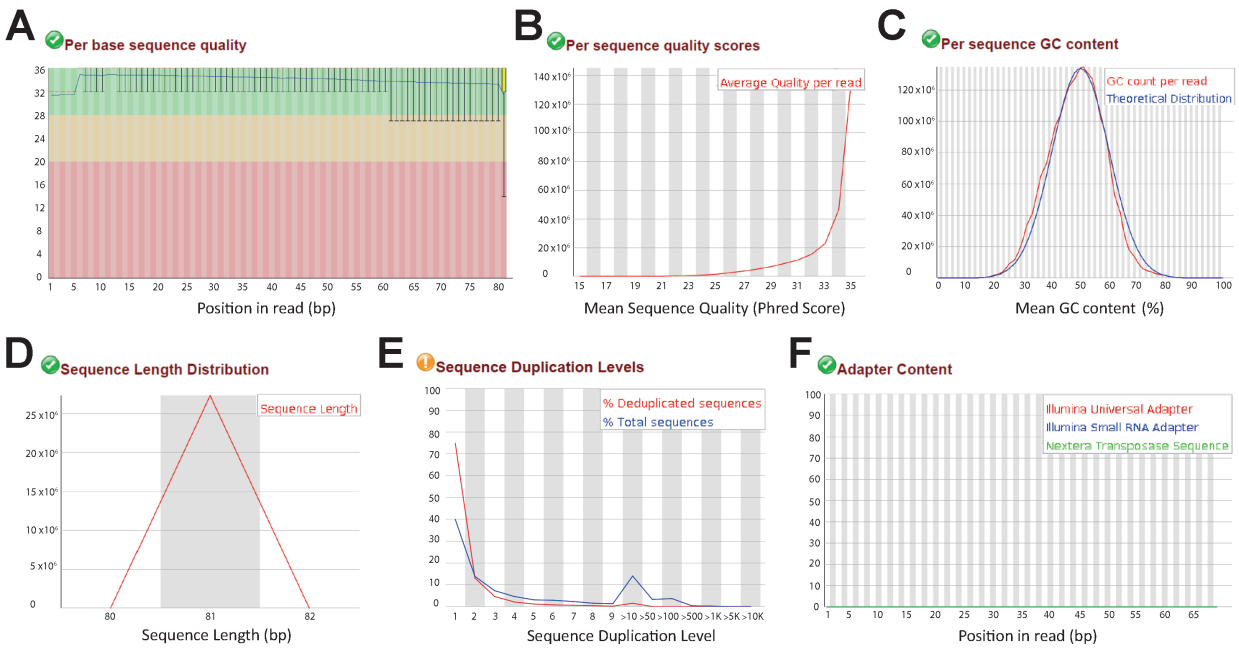

Figure 2.1: **Example of quality control of raw sequenced data (FASTQC). (A)** Quality scores for individual positions within read sequence (over all reads). **(B)** Quality score distribution over all sequences. **(C)** GC content distribution over all sequences. Blue: theoretical GC content (%); red: observed GC content (%). **(D)** Distribution of sequence length over all sequences. **(E)** Relative number of sequences with different degrees of duplication. **(F)** Frequency of contamination by sequencing adapters. Replicate 2 of Ly6D$^+$ group is taken as a representative example.

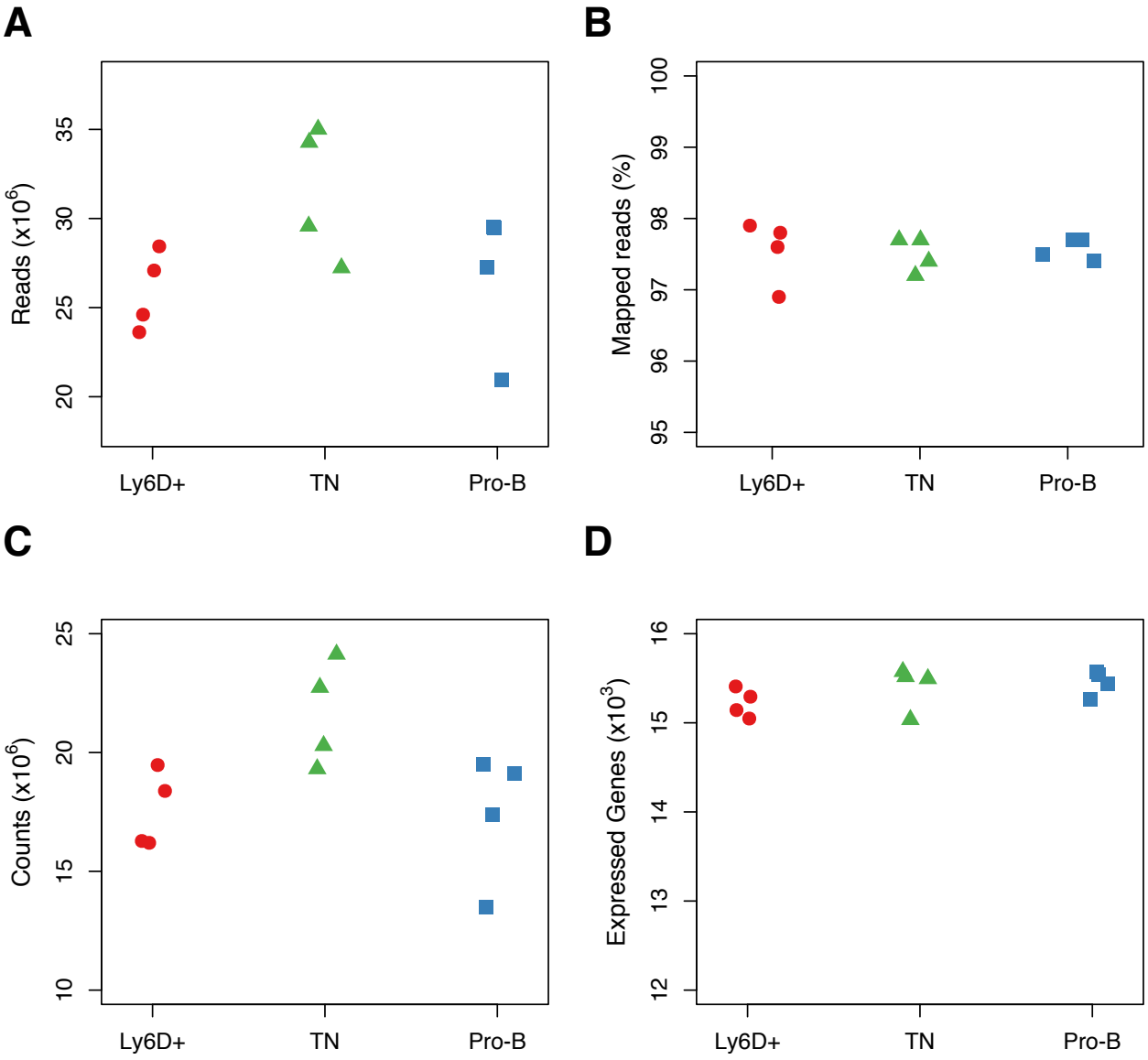

Figure 2.2: **Quantification of raw sequenced data for the Ly6D$^+$, TN and pro-B samples. (A)** Number of sequenced reads per sample. **(B)** Percentage of reads that mapped to the mouse genome (mm10). **(C)** Number of counts per sample (library size) considering reads mapped to genes (exons only). **(D)** The total number of detected genes per sample (with at least 1 count).

Obtained sequencing reads, 21 to 35 millions of reads (81-mers) per sample **(Figure 2.2A)**, were aligned to the mouse genome assembly, version mm10 (downloaded from UCSC http://genome.ucsc.edu), with STAR (Alexander Dobin and Gingeras 2013), run with a custom made R wrapper. The default parameters of STAR aligner were used, except for reporting for multi-mappers only one hit in the final alignment files (outSAMmultNmax=1) and filtering

reads without evidence in spliced junction table (outFilterType="BySJout"). All downstream analysis was performed using the open source R software (R Core Team 2017) accessed via RStudio server (https://www.rstudio.comR version 3.4.0).

More than 96% of total reads were successfully mapped for each sample **(Figure 2.2B)**. Subsequently, a count table with gene expression levels was generated using the qCount function from QuasR package v1.16.0 and coordinates of RefSeq mRNA genes (http://genome.ucsc.edu, downloaded in December 2013). The expression level was defined as the number of reads that started within any annotated exon of a gene (exon-union model). Total counts per sample ranged from 13 to 24 millions **(Figure 2.2C)**, the so-called library size. Genes with no counts across all samples were filtered out from the analysis. For 18,003 genes at least 1 read was detected across all samples, corresponding to ~15,400 genes per sample **(Figure 2.2D)**.

Raw counts were normalized between samples with the TMM method (weighted trimmed mean of M-values (Robinson and Oshlack 2010)). Counts per million of mapped reads (CPM), and especially log2-transformed CPM vaues ($\log_2$CPM) were used for data exploration.

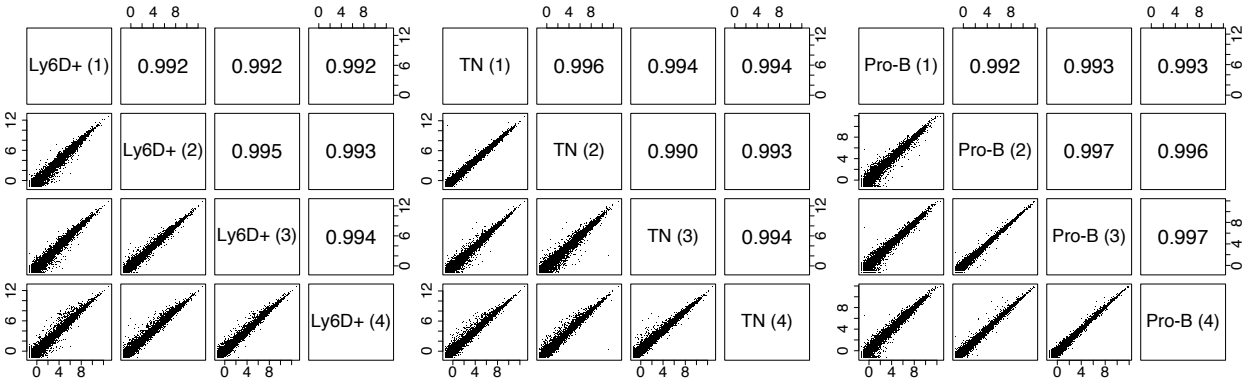

Figure 2.3: **Quality of the Ly6D$^+$ (left), TN (middle) and pro-B (right) replicates.** Lower panels show the pair-wise scatter plots with the expression profiles ($\log_2$CPM) of the replicates per population. Upper panels show the corresponding pair-wise Pearson's transcriptome correlation per population.

To assess the quality of the replicates, I calculated the pair-wise Pearson's correlations among the Ly6D$^+$, TN and Pro-B replicates and illustrated them in pair-wise scatter plots per population (lower panels) with the corresponding correlation value in the upper panels

(function pairs()) **(Figure 2.3)**. All biological replicates showed very high transcriptome correlation (r > 0.990) **(Figure 2.3)**. Therefore, all samples passed the basic quality control (high frequency of mapped reads, elevated number of detected genes and high level of transcriptome correlation among the replicates) and I proceeded with the subsequent analysis using all samples.

2.2 Exploratory analysis reveals that Ly6D$^+$ is the EPLM subset more transcriptionally related to the pro-B cells

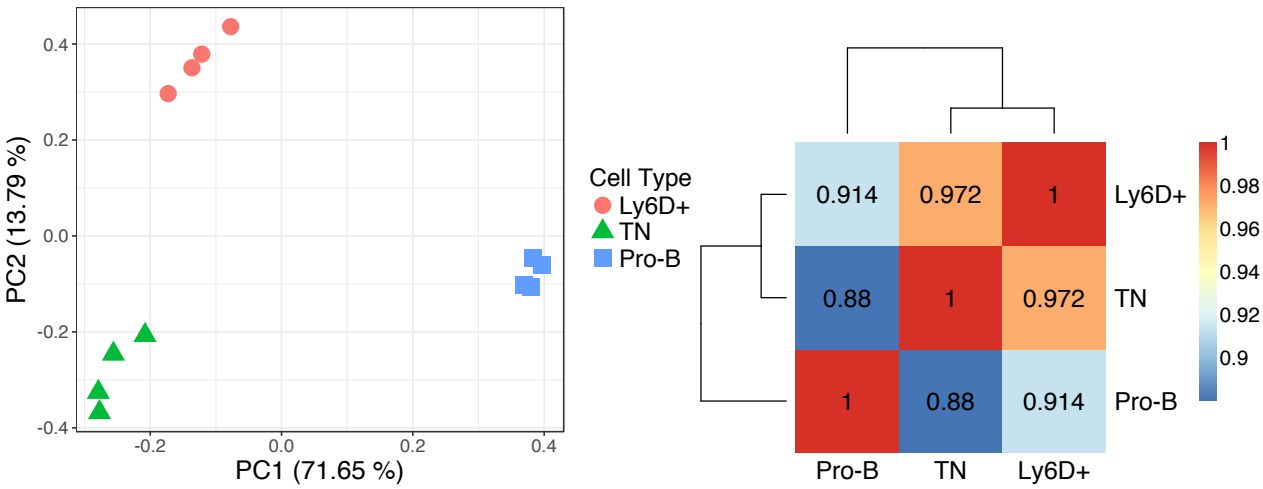

Figure 2.4: **Of the two EPLM subpopulations, Ly6D$^+$ cells have a closer transcriptome profile to the pro-B cells.** Principal component analysis (left) and heatmap with pair-wise Pearson's transcriptome correlation (right) of Ly6D$^+$, TN and Pro-B replicates and averaged populations respectively. The top 50% of genes with highest variance across analysed dataset (calculated as inter-quartile range) were used.

In order to visualize the data I applied dimensionality reduction technique: Principal Component Analysis (PCA, prcomp() function. For that, only the top 50% of genes with highest variance across analysed dataset (calculated as inter-quartile range) were used. The expression of every gene was centered to zero and the final plot PCA plot was generated with the ggplot2 v2.2.1 R package. As reflected in the PCA plot, the highest variation among samples (PC1

axis captures 71.65% of the total variation) corresponds to their developmental stage; with the uncommitted EPLM subpopulation (Ly6D$^+$ and TN) on the left and the committed pro-B cells on the right **(Figure 2.4 left)**, thus suggesting that EPLM subsets are significantly different to a B-cell committed transcriptional state. Hierarchical clustering of the subsets based on Pearson's correlation coefficients and illustrated in a correlation heatmap (function aheatmap() of NMF R package v0.20.6) revealed that, in line with PCA, Ly6D$^+$ and TN were the two populations with the highest transcriptome association (r=0.972). Moreover, from the two EPLM subpopulations, the Ly6D$^+$ cells had higher transcriptome correlation to pro-B cells (r = 0.914), than that of the TN subset (r = 0.88) **(Figure 2.4 right)**. As a result, this exploratory analysis indicates that Ly6D$^+$ cells are closer to the B-cell lineage compared with the TN cells.

2.3 Differentially expressed genes

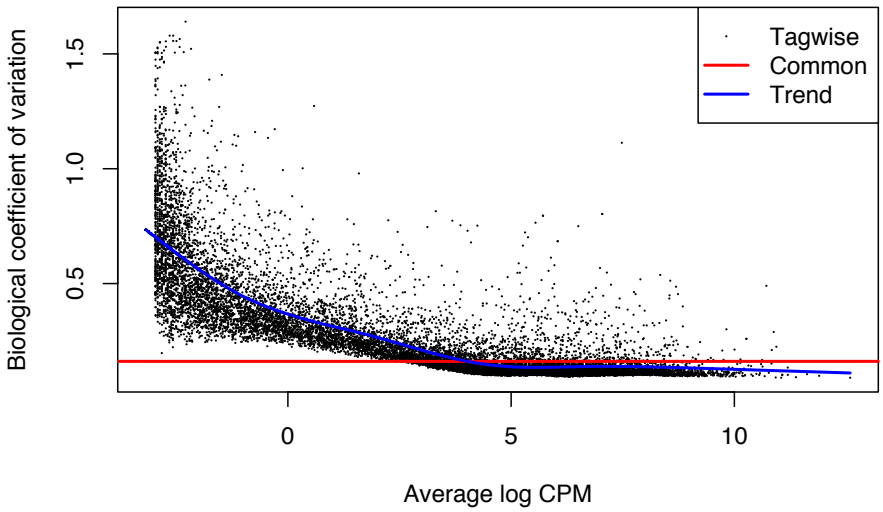

Figure 2.5: Estimated common, trended and tag-wise (per gene) biological coefficient of variation against the average gene expression level ($\log_2$CPM).

I further studied expression differences by performing differential expression analysis using edgeR v3.18.1 R package. **Figure 2.5** shows estimated dispersion as function of gene expression levels. It is clear from the figure that the genes with low expression levels have the highest dispersion. The dispersion estimation is done in three steps: first a single value

Table 2.1: Summary table of differential expression analysis containing the differentially expressed, up-regulated and down-regulated genes in number and percentage of each pair-wise transcriptome comparison. The percentage of up- and down-regulated genes is relative to the total number of DEG.

		Total	DEG	Up-regulated	Down-regulated
Ly6D+ vs TN	n	18002	972	476	496
	%	100	5.4	48.97	51.03
Pro-B vs Ly6D+	n	18002	2944	1488	1456
	%	100	16.35	50.54	49.46
Pro-B vs TN	n	18002	3639	1753	1886
	%	100	20.21	48.17	51.83

Table 2.2: A table of the top 10 differentially expressed genes (DEG) in the comparison Ly6D+ vs TN. logFC (log2 fold change): group mean expression ratio; logCPM (log2 counts per million): average expresion of tge gene in the dataset; LR (likelihood ratio): comparison of full versus null model, where one coefficiant is droppped out; PValue: significance level; FDR (false discovery rate): adjusted p-value. The sign of the logFC column indicates the direction: positive, up-regulated or higher expressed in the first population; negative: down-regulated or higher expressed in the latter population.

ENTREZID	SYMBOL	GENENAME	logFC	logCPM	LR	PValue	FDR
19152	Prtn3	proteinase 3	-4.38	6.75	630.38	4.12e-139	7.42e-135
74145	F13a1	coagulation factor XIII, A1 subunit	-4.95	6.91	615.57	6.87e-136	6.18e-132
244234	5830411N06Rik	RIKEN cDNA 5830411N06 gene	5.31	4.32	474.63	3.15e-105	1.89e-101
17068	Ly6d	lymphocyte antigen 6 complex, locus D	4.75	8.40	458.29	1.13e-101	5.09e-98
11745	Anxa3	annexin A3	-3.32	4.15	454.26	8.53e-101	3.07e-97
212032	Hk3	hexokinase 3	-4.18	5.38	333.91	1.35e-74	4.06e-71
246707	Emilin2	elastin microfibril interfacer 2	-3.99	4.29	296.94	1.53e-66	3.93e-63
54483	Mefv	Mediterranean fever	-4.23	3.31	287.45	1.79e-64	4.03e-61
11820	App	amyloid beta (A4) precursor protein	-2.64	3.96	282.91	1.74e-63	3.48e-60
23936	Lynx1	Ly6/neurotoxin 1	2.45	5.18	249.05	4.19e-56	7.54e-53

Table 2.3: A table of the top 10 differentially expressed genes (DEG) in the comparison pro-B vs Ly6D+ as in Table 2.2

ENTREZID	SYMBOL	GENENAME	logFC	logCPM	LR	PValue	FDR
276829	Smtnl2	smoothelin-like 2	7.36	5.19	1002.82	4.39e-220	7.90e-216
56198	Heyl	hairy/enhancer-of-split related with YRPW motif-like	7.30	4.56	976.73	2.06e-214	1.85e-210
68149	Otub2	OTU domain, ubiquitin aldehyde binding 2	5.00	5.21	847.89	2.09e-186	1.25e-182
14255	Flt3	FMS-like tyrosine kinase 3	-3.36	9.21	793.21	1.61e-174	7.27e-171
14732	Gpam	glycerol-3-phosphate acyltransferase, mitochondrial	4.23	7.34	743.23	1.19e-163	4.28e-160
12490	Cd34	CD34 antigen	-5.20	8.17	705.33	2.08e-155	6.23e-152
16000	Igf1	insulin-like growth factor 1	6.04	3.58	675.31	7.01e-149	1.80e-145
12511	Cd6	CD6 antigen	-5.57	4.68	668.99	1.66e-147	3.73e-144
12043	Bcl2	B cell leukemia/lymphoma 2	-4.22	8.10	652.16	7.59e-144	1.52e-140
72324	Plxdc1	plexin domain containing 1	4.91	6.08	611.68	4.82e-135	8.68e-132

Table 2.4: A table of the top 10 differentially expressed genes (DEG) in the comparison Pro-B vs TN as in Table 2.2

ENTREZID	SYMBOL	GENENAME	logFC	logCPM	LR	PValue	FDR
276829	Smtnl2	smoothelin-like 2	7.13	5.19	1007.15	5.00e-221	9.00e-217
56198	Heyl	hairy/enhancer-of-split related with YRPW motif-like	6.99	4.56	972.75	1.50e-213	1.35e-209
68149	Otub2	OTU domain, ubiquitin aldehyde binding 2	5.19	5.21	914.80	5.96e-201	3.58e-197
12490	Cd34	CD34 antigen	-5.68	8.17	805.86	2.87e-177	1.29e-173
12043	Bcl2	B cell leukemia/lymphoma 2	-4.75	8.10	791.00	4.90e-174	1.76e-170
53623	Gria3	glutamate receptor, ionotropic, AMPA3 (alpha 3)	-4.22	7.35	780.52	9.30e-172	2.79e-168
14732	Gpam	glycerol-3-phosphate acyltransferase, mitochondrial	4.22	7.34	745.66	3.53e-164	9.07e-161
104175	Sbk1	SH3-binding kinase 1	3.82	6.93	739.13	9.26e-163	2.08e-159
20562	Slit1	slit homolog 1 (Drosophila)	5.28	4.14	729.16	1.36e-160	2.73e-157
16000	Igf1	insulin-like growth factor 1	6.10	3.58	727.54	3.07e-160	5.53e-157

representing common dispersion (estimateGLMCommonDisp() function, horizontal line), trended dispersion, which takes into account the dependecy on the expression level (esti-

mateGLMTrendedDisp() function) and finally gene specific dispersion (estimateGLMTagwiseDisp() function). When fitting the design using the negative binomial model (glmFit() function) with the tag-wise dispersion, I set a prior count (pseudocount) to 8 in order to minimize the large log-fold changes for genes with very small number of counts. Finally, after the dispersion estimate and fit of the negative binomial model, I proceeded with determining differential expression for each comparison (contrast of interest) using the likelihood ratio test (glmLRT() function).

Table 2.1 summarizes the analysis by reporting the total number of Differentially Expressed Genes (DEG) for each pair-wise comparison as well as the number and the fraction corresponding to up-regulated and down-regulated genes. To be considered as DEG, the gene expression had to be at least two times up/down-regulated (abs($\log_2$FoldChange) >1) and this change in expression had to be significant (false discovery rate (FDR) corrected p-value, FDR <0.05). **Tables 2.2, 2.3, and 2.4** shows the top 10 differentially expressed genes per comparison ranked according to the significance. Positive fold change values, log2FC >0, correspond to genes higher expressed by the Ly6D$^+$ **(Table 2.2)** whereas negative fold change values, log2FC <0, to genes higher expressed by the TN cells **(Table 2.2)**. The complete table is provided in a supplementary excel file (1_DEGlists).

A considerable fraction (20% and 16%) of genes was differentially expressed when comparing pro-B with either Ly6D$^+$ or TN subpopulations respectively whereas only about 5% (972 genes) had a significant change in expression between the two EPLM subpopulations **(Table 2.1)**, thus suggesting again that Ly6D$^+$ and TN cells are more related to each other than to pro-B cells. In addition, the graphical representation of the differential expression analysis (volcano plots), **(Figure 2.6)**) also revealed that the overall fold changes and significance levels of the DEGs were lower when comparing the Ly6D$^+$ vs TN subpopulations **(Figure 2.6** left plot) than when they were individually compared with the pro-B population **(Figure 2.6** middle and right plots). Interestingly, there was a similar fraction of up-regulated and down-regulated genes in each comparison, indicating no predominant activation or repression of genetic programs from one hematopoietic stage to the other **(Table 2.1 and Figure 2.6**.

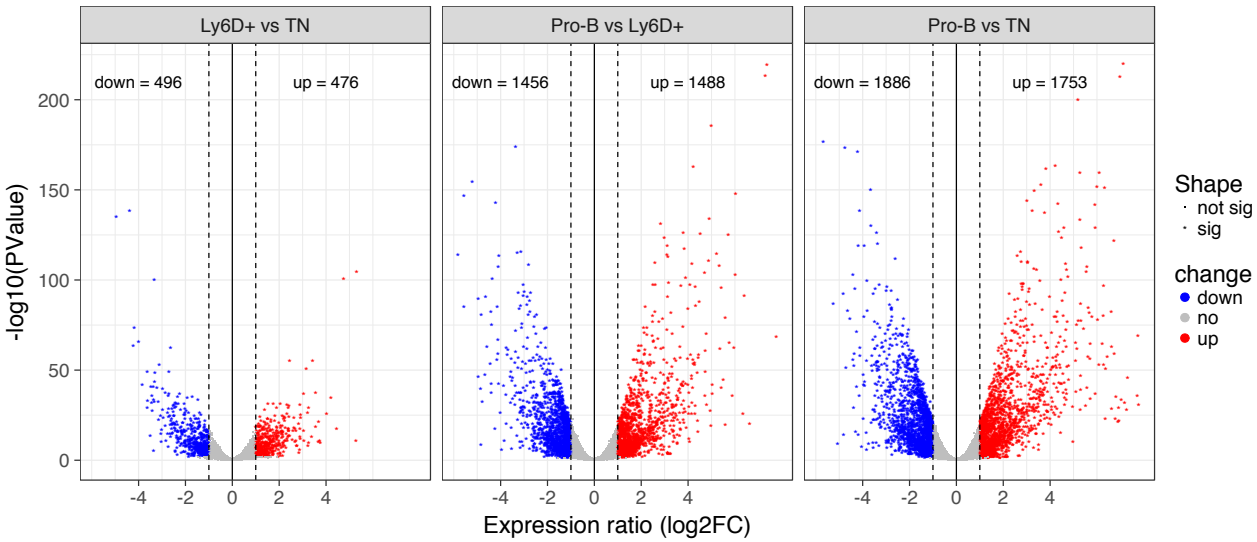

Figure 2.6: Volcano plot (plotted significance against expression ratio) of each pair-wise transcriptome comparison. Each dot/star represents a gene. Grey dots: not DEGs; red star: up-regulated genes; blue stars: down-regulated genes.

A considerable fraction (20% and 16%) of genes was differentially expressed when comparing pro-B with either Ly6D$^+$ or TN subpopulations respectively whereas only about 5% (972 genes) had a significant change in expression between the two EPLM subpopulations **(Table 2.1)**, thus suggesting again that Ly6D$^+$ and TN cells are more related to each other than to pro-B cells. In addition, the graphical representation of the differential expression analysis (volcano plots generated with custom R scripts, **(Figure 2.6)**) also revealed that the overall fold changes and significance levels of the DEGs were lower when comparing the Ly6D$^+$ vs TN suppopulations **(Figure 2.6** left plot) than when they were individually compared with the pro-B population **(Figure 2.6** middle and right plots). Interestingly, there was a similar fraction of up-regulated and down-regulated genes in each comparison, indicating no predominant activation or repression of genetic programmes from one hematopoietic stage to the other **(Table 2.1 and Figure 2.6**).

2.4 Functional analysis reveals that Ly6D$^+$ and TN cells have distinct genetic signatures

Next, I investigated into detail the expression differences between the apparently transcriptionally related Ly6D$^+$ and TN EPLM subpopulations. For this, I studied the nature of the DEG (Ly6D$^+$ vs TN) by gene ontology (GO) enrichment analysis performed with the DAVID v6.8 bioinformatics database. This analysis is based on Fisher's Exact method (Huang da W. and Lempicki 2009a, Huang da W. and Lempicki (2009b)) and I considered gene ontology terms of DEG to be significantly enriched when p-value <0.05.

Results revealed that the ~500 up-regulated genes in Ly6D$^+$ cells were enriched for lymphoid biological processes such as activation, proliferation and differentiation of B and T cells, VDJ recombination or immunoglobulin production **(Table 2.5 up)**. Moreover, the genes that accounted for B-cell related biological processes were highly expressed and most of them within the top up-regulated genes, as reflected in the labelled MA plot of **Figure 2.7**, generated with ggplot2 v2.2.1 and ggrepel v0.6.5 packages. Therefore, Ly6D$^+$ cells already express important genes for B-cell development. Among these, there are genes encoding the B-cell related transcription factors *Pax5*, *Ebf1* and *Pou2af1* (*Obf1*), the recombinase machinery *Rag1* and *Rag2*, the surrogate light chains *VpreB1*, *VpreB2*, *VpreB3*, and *Igll1* (lambda 5) of the pre B-cell receptor (pre-BCR), the signalling immunoglobulin α (*Cd79a*) and β (*Cd79b*) chains of the pre-BCR complex, the non-receptor tyrosine kinase *Blk* involved in BCR signaling, the receptor for interleukin-7 *Il7r*, and other lymphoid related genes such as *Dntt*, *Lax* and, as expected, *Ly6d* itself. Interestingly, although Ly6D$^+$ cells were sorted as CD19$^-$ cells, mRNA expression of the B cell co-receptor *CD19* was already detected. Taken together, these results suggest that qualitatively, Ly6D$^+$ cells express a B-cell genetic signature characteristic of the CD19$^+$ pro-B cell stage. However, quantitatively, the overall expression of these genes is markedly lower than the pro-B cells, as exemplified in the heatmap (generated with the function aheatmap() of NMF R package v0.20.6) of **Figure 2.8**.

Table 2.5: Selection of enriched Biological Processes (eBP) in up-regulated genes of the Ly6D+ versus TN comparison. Complete list in excel file 2 (2_eBP_Ly6DvsTN.xls).

Category	Term	Count	PValue	Fold.Enrichment
GOTERM_BP_FAT	GO:0046649 lymphocyte activation	19	4.00e-07	4.35
GOTERM_BP_FAT	GO:0042113 B cell activation	10	6.83e-05	5.60
GOTERM_BP_FAT	GO:0042100 B cell proliferation	5	1.14e-04	18.20
GOTERM_BP_FAT	GO:0033151 V(D)J recombination	4	8.92e-04	19.42
GOTERM_BP_FAT	GO:0042110 T cell activation	10	1.30e-03	3.77
GOTERM_BP_FAT	GO:0030217 T cell differentiation	8	1.70e-03	4.60
GOTERM_BP_FAT	GO:0050853 B cell receptor signaling pathway	4	3.60e-03	12.48
GOTERM_BP_FAT	GO:0016444 somatic cell DNA recombination	4	1.31e-02	7.94
GOTERM_BP_FAT	GO:0045165 cell fate commitment	9	1.95e-02	2.67
GOTERM_BP_FAT	GO:0030183 B cell differentiation	5	2.04e-02	4.75
GOTERM_BP_FAT	GO:0002377 immunoglobulin production	4	3.03e-02	5.83

Table 2.6: Selection of enriched Biological Processes (eBP) in down-regulated genes of the Ly6D+ versus TN comparison. Complete list in excel file 2 (2_eBP_Ly6DvsTN.xls).

Category	Term	Count	PValue	Fold.Enrichment
GOTERM_BP_FAT	GO:0009611 response to wounding	41	0.00e+00	4.22
GOTERM_BP_FAT	GO:0006954 inflammatory response	30	0.00e+00	4.77
GOTERM_BP_FAT	GO:0006909 phagocytosis	12	1.00e-07	8.76
GOTERM_BP_FAT	GO:0002274 myeloid leukocyte activation	9	4.10e-06	9.19
GOTERM_BP_FAT	GO:0032680 regulation of TNF production	8	6.80e-06	10.59
GOTERM_BP_FAT	GO:0006897 endocytosis	18	2.06e-05	3.42
GOTERM_BP_FAT	GO:0045087 innate immune response	13	4.25e-05	4.34
GOTERM_BP_FAT	GO:0016064 immunoglobulin mediated immune response	9	3.45e-04	5.11
GOTERM_BP_FAT	GO:0045576 mast cell activation	5	6.37e-04	11.92
GOTERM_BP_FAT	GO:0042742 defense response to bacterium	11	8.64e-04	3.64
GOTERM_BP_FAT	GO:0030099 myeloid cell differentiation	10	1.40e-03	3.72
GOTERM_BP_FAT	GO:0006957 complement activation, alternative pathway	4	2.20e-03	14.30
GOTERM_BP_FAT	GO:0019882 antigen processing and presentation	9	2.90e-03	3.70
GOTERM_BP_FAT	GO:0001878 response to yeast	3	1.08e-02	17.88
GOTERM_BP_FAT	GO:0009620 response to fungus	4	1.09e-02	8.41
GOTERM_BP_FAT	GO:0030593 neutrophil chemotaxis	4	1.09e-02	8.41
GOTERM_BP_FAT	GO:0042116 macrophage activation	3	3.01e-02	10.73

In contrast to B-cell related genes, the genes accounting for T-cell related biological processes in the ~500 up-regulated genes in Ly6D$^+$ cells presented overall lower expression ratios and variable expression intensities **(Figure 2.7)**. Among these genes were the T-cell transcription factor *Bcl11b*, the *Notch1* receptor, a master regulator of T-cell commitment whose signaling represses the expression of genes related with other lineages, the signalling CD3 zeta chain (*Cd247*) of the T-cell receptor (TCR) complex, genes involved in pre-TCR signalling such as *Lck* (non-receptor tyrosine kinase), *Rhoh* (related GTP-binding protein), *Zap70* (tyrosine kinase), and *Sla2* (Src-like-adapter protein), *Trat1* (an adaptor protein that stabilizes the TCR/CD3 complex at the surface of T-cells), the *tnfsf11* cytokine and *Nlrc3* (positive regulators of T cell activation), and the inhibitory T-cell related receptors *Ctla4* and *Ctla2b*. This T-cell genetic signature is exclusive to the Ly6D$^+$ subpopulation **(Figure 2.8)** and is the "feature" that separates them from both the pro-B and TN cells along the PC2 of the principal component analysis **(Figure 2.4 left)**.

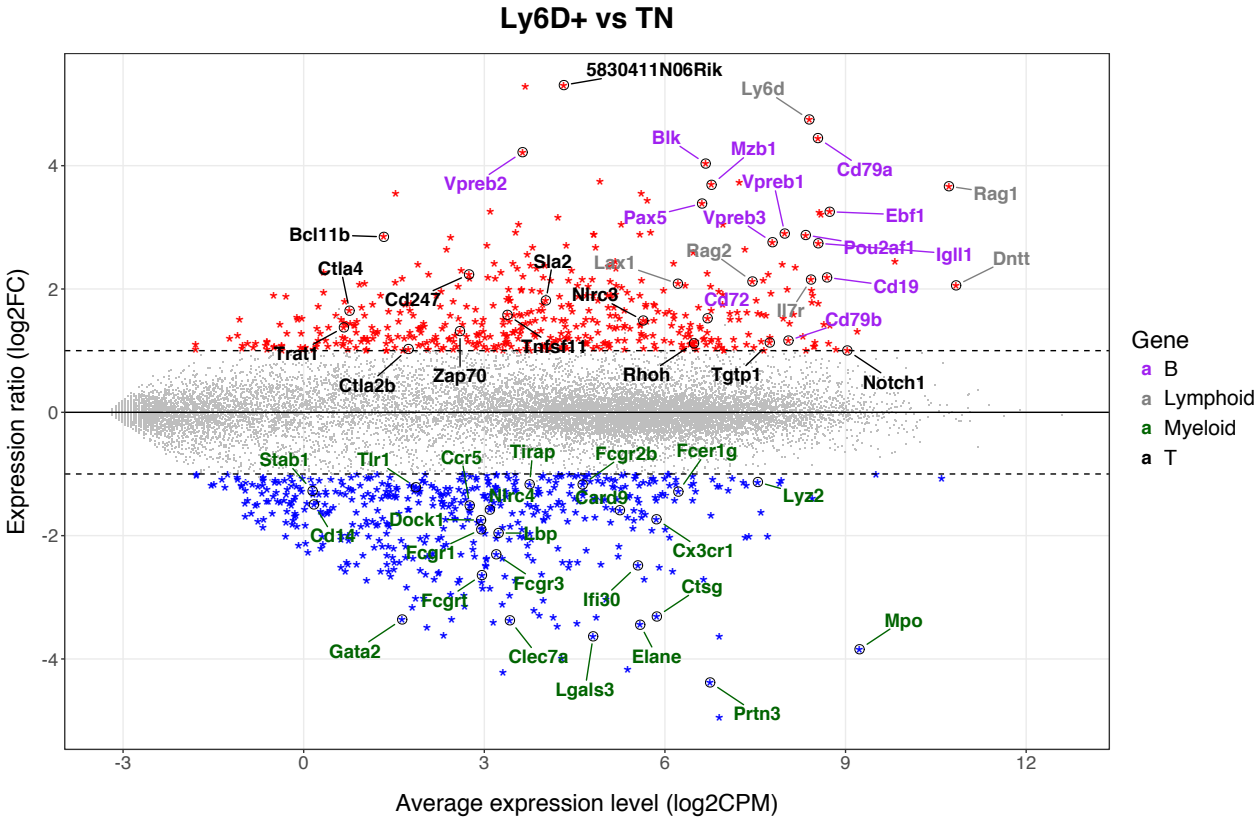

Figure 2.7: **Ly6D$^+$ and TN EPLM subpopulations have distinct genetic signatures.** MA plot (plotted expression ratio against average expression) of Ly6D$^+$ vs TN transcriptome comparison. B-cell, T-cell, lymphoid and myeloid related genes are indicated.

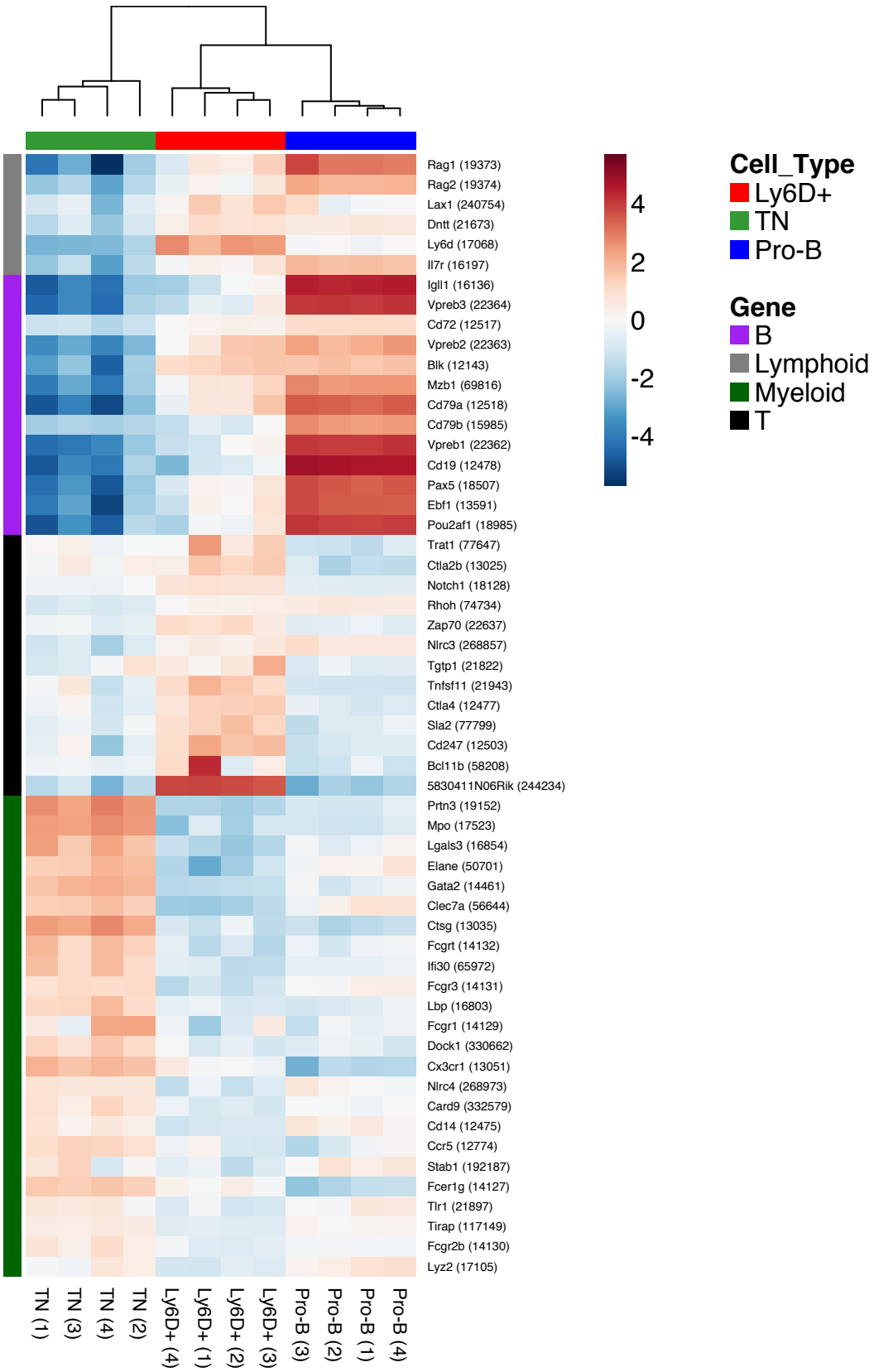

Figure 2.8: **Heatmap of expression of the lineage-specific genes indicated in the MAplot**. The colour gradient illustrates the relative gene expression level (gene-wise centered $\log_2$CPM values).

Analysis of the ~500 down-regulated genes revealed that they were largely related with myeloid and innate biological processes such as inflammation, phagocytosis, responses to bacteria, yeast and fungi, and macrophage activation **(Table 2.6, down)**. Some key genes accounting for these processes were *Mpo*, *Elane* and *Ctsg* (enzymes with microbicidal activity), *Prtn3* (a serine protease that degrades elastin, fibronectin, laminin, vitronectin, and collagen), the phagocytic Fc receptors *Fcgr2b*, *Fcgrt*, *Fcer1g* and *Fcgr1*, the pathogen recognition receptor *Tlr1* (Toll-like receptor 1), *Gata2* (transcriptional regulator of phagocytosis), *Clec7a* (involved in TLR2-mediated inflammatory responses), the polysaccharide binding protein *Lbp* and the chemokine receptor *Cx3cr1* involved in myeloid leukocyte activation **(Figure 2.7)** and **(Figure 2.8)**. Therefore, TN cells seem to present a myeloid genetic signature.

From this transcriptional analysis, I conclude that i) EPLM subpopulations are distinct from one another and are both distinct from pro-B cells ii) of the two EPLM subsets, Ly6D$^+$ cells are closer to pro-B cells, and iii) whereas the Ly6D$^+$ subset has a largely lymphoid genetic signature, that of the TN subset is more myeloid.

Chapter 3

Conclusions

In the present study, I have investigated whether the hematopoietic progenitor EPLM, previously characterized as a multipotent and phenotypically homogeneous B220^{+} c-Kit^{int} CD19^{-} and NK1.1^{-} population (Balcicunaite G. and Rolink 2005), possesses truly combined lymphoid and myeloid developmental potentials or, instead, if it is composed by a mixture of cells with more constrained differentiation capacities. By using three cell surface markers (Ly6D, SiglecH and CD11c) already known to be associated with distinct hematopoietic lineages (Inlay MA 2009, Blasius AL (2006), Zhang J (2006), Singh-Jasuja H (2013)), the host laboratory recently fractionated the EPLM into at least four subpopulations **(Figure 1.1)**, thus envisaging that EPLM is a heterogeneous population, at least phenotypically. Two of the EPLM subpopulations (blue and orange in **Figure 1.1**) might be, as expression of SiglecH and cD11c suggests, already committed (some cells even differentiated) to the plasmacytoid and conventional dendritic cell lineages, respectively (Blasius AL 2006, Zhang J (2006), Singh-Jasuja H (2013)). Therefore, and since EPLM was mainly described as a B-cell progenitor population but still retaining myeloid potential (Balcicunaite G. and Rolink 2005), I directed my analysis towards the two fractions, namely Ly6D^{+} and TN, that seemed to retain multipotentiality (including B-cell developmental potential). Ly6D has been already described as a marker that enriches B-cell progenitors (Inlay MA 2009) whereas the triple negative fraction, which lacks expression of the three lineage-related cell surface markers Ly6D, SiglecH and CD11c is a good multipotent candidate. In this project,

I have characterized these two EPLM subpopulations at the molecular level by performing population RNA-seq and comparing their transcriptomes with the first B-cell committed population, the pro-B.

First, I was able to isolate the populations from the bone marrow (two femurs) of *Flt3L*tg mice in great numbers (>1x10^5 cells) and to extract at least 500ng of total RNA per sample in order to be able to use a cost-effective and leading stranded library preparation protocol ((Tatiana Borodina and Sultan 2011, Joshua Z Levin (2010)) and 1st section of the results). The ability to differentiate sense and anti-sense transcripts is very important to avoid false positives with reads mapping to the wrong strand. All samples had an optimal number of sequenced reads, presented good sequencing quality **(Figure 1.1)**, and showed a very high mapping frequency to the mm10 reference genome **(Figure 2.2B)**, thus demonstrating the appropriateness of the aligner used, STAR (Alexander Dobin and Gingeras 2013). The preliminary analysis revealed that a great number of RefSeq genes was detected per sample (> 15,000, **Figure 2.2D**, indicating sufficient sequencing depth (~ 28 millions of read per sample) for the purpose of the project (analyze the change in expression of as many genes as possible among samples) and anticipating that the library preparation and sequencing protocols used not only captured the highly expressed genes but also the lowly expressed genes such as transcription factors, which play crucial roles in biological processes. Importantly, all biological replicates showed high transcriptome correlation (r >0.990, **Figure 2.3**) and clustered together **(Figure 2.4 left)**, revealing, as expected, higher inter-population transcriptional variation.

Next, I analyzed into detail the genetic signatures of the two EPLM subsets by performing differential expression analysis and, subsequently subjecting the DEG to functional analysis to identify which biological processes were enriched. With this approach, I was able to unravel marked genetic biases between the two EPLM subsets indicative of molecular priming towards distinct fates. Whereas the Ly6D$^+$ population showed a lymphoid genetic signature that was more prominent to the B-cell lineage, with robust expression of B-cell related genes (*CD79a*, *Vpreb* genes, *Igll1*, *CD19*, *Ebf1*, *Pax5* or *Blnk* **Figure 2.7**), the TN population exhibited a myeloid genetic signature reflected by expression of myeloid genes (*Mpo*, *Prtn3*, *Elane*, *Ctsg*, *Cx3cr1*, *Gata2* or the Fc receptors **Figure 2.7**) related with innate biological processes **(Table**

2.6, down), thus suggesting functional heterogeneity among EPLM subpopulations. There is extensive scientific evidence that, during a differentiation process such as hematopoiesis, genetic specification or molecular priming precedes commitment (Hu M 1997, Nimmo RA (2015), Zandi S (2012)) meaning that a cell first up-regulates lineage-related genes that reveal its spectrum of developmental potentials and, later on, as a consequence of its genetic program as well as the contribution of environmental cues such as cytokines, the cell makes a choice to gradually restrict, mature and differentiate towards a specific lineage. Therefore, although with caution, I speculate that Ly6D$^+$ EPLM subpopulation is a lymphoid progenitor and would predominantly differentiate towards the B-cell lineage, whereas the TN cells would mainly differentiate into myeloid cells. To confirm this hypothesis, functional experiments such as *in vitro* limiting dilution assays or *in vivo* reconstitution of sublethaly-irradiated mice would be required. Moreover, it is important to keep in mind that the gene expression levels obtained in an RNA-seq experiment are relative to the populations included in the analysis. For instance, although Ly6D$^+$ cells up-regulate B-cell genes compared with the TN cells, these genes are in turn expressed in lower levels when they are compared to the committed B-cell progenitor pro-B. Therefore, Ly6D$^+$ EPLM is a different population that could be placed in an earlier B-cell developmental stage than the pro-B cells. In addition, its T-cell signature, absent in the committed pro-B cells, suggests that Ly6D$^+$ might be still an uncommitted population.

Although there is no known function for the Ly6D cell surface protein, the data resulted from this project confirms that its expression can be used to discriminate among EPLM subpopulations that have distinct genetic programs and that is a good marker to enrich for B-cell biased populations (as previously demonstrated by Inlay et al. with the Ly6D$^+$ CLP fraction (Inlay MA 2009)). Therefore, EPLM is not only phenotypically but also genetically heterogeneous, thus suggesting that the lympho-myeloid developmental potential observed for the whole EPLM population could be constrained within the Ly6D$^+$ and TN fractions, respectively. To confirm that EPLM is not lympho-myeloid bipotent at the clonal level single-cell RNA sequencing, which has emerged as the master tool to dissect heterogeneity and now is widely accessible, would be required (Treutlein B. 2014, Zeisel (2015), Zheng GX (2015)). However, although highly interesting, this goes beyond the time frame of this

project.

As a conclusion, in this master project, by performing bulk RNA sequencing, I have found that the heterogeneous expression of the cell surface marker Ly6D by EPLM correlates with differential gene expression programs. Whereas the Ly6D$^+$ cells are lymphoid primed, have a closer transcriptome profile to the pro-B cells and could even be their direct precursor, the TN cells are myeloid primed. To resolve whether the distinct genetic signatures result into distinct differentiation capacities, further investigation is needed. More generally, this study provides a good example to support the concept that previously described homogeneous multipotent populations based on the expression of few cell surface markers, can result heterogeneous when they are analyzed with additional cell surface markers and even at the whole transcriptome scale.

Chapter 4

References

Alexander Dobin, Felix Schlesinger, Carrie A. Davis, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal Rna-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Balcicunaite G., Massa S., Ceredig R., and A. Rolink. 2005. "A B220+ Cd117+ Cd19- Hematopoietic Progenitor with Potent Lymphoid and Myeloid Developmental Potential." *Eur J Immunol.* 35 (7): 2019–30.

Blasius AL, Maldonado J, Cella M. 2006. "Siglec-H Is an Ipc-Specific Receptor That Modulates Type I Ifn Secretion Through Dap12." *Blood* 107 (6): 2472–6.

Boekel E., Melcher F. ten, and A. Rolink. 1995. "The Status of Ig Loci Rearrangements in Single Cells from Different Stages of B Cell Development." *Int Immunol.* 7 (6): 1013–9.

Chomczynski, Piotr, and Nicoletta Sacchi. 1986. "Single-Step Method of Rna Isolation by Acid Guanidinium Thiocyanate-Phenol-Chloroform Extraction." *Analytical Biochemistry* 162 (1): 156–59.

———. 2006. "The Single-Step Method of Rna Isolation by Acid Guanidinium Thiocyanate-Phenol-Chloroform Extraction: Twenty-Something Years on." *Nature Protocols* 1 (2): 581–85.

Cornelis J.H. Pronk, Robert Mansson, Derrick J. Rossi, and David Bryder. 2007. "Elucidation of the Phenotypic, Functional, and Molecular Topography of a Myeloerythroid Progenitor

Cell Hierarchy." *Cell Stem Cell* 11 (4): 428–42.

Hu M, Greaves M, Krause D. 1997. "Multilineage Gene Expression Precedes Commitment in the Hemopoietic System." *Genes Dev.* 11 (6): 774–85.

Huang da W., Sherman B. T., and R. A. Lempicki. 2009a. "Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists." *Nucleic Acids Res* 37 (1): 1–13.

———. 2009b. "Systematic and Integrative Analysis of Large Gene Lists Using David Bioinformatics Resources." *Nat Protoc* 4 (1): 44–57.

Inlay MA, Sahoo D, Bhattacharya D. 2009. "Ly6d Marks the Earliest Stage of B-Cell Specification and Identifies the Branchpoint Between B-Cell and T-Cell Development." *Genes Dev.* 23 (20): 2376–81.

Joshua Z Levin, Xian Adiconis, Moran Yassour. 2010. "Comprehensive Comparative Analysis of Strand-Specific Rna Sequencing Methods." *Nature Methods* 7 (9): 709–15.

Karsunky H, Serwold T, Inlay MA. 2008. "Flk2+ Common Lymphoid Progenitors Possess Equivalent Differentiation Potential for the B and T Lineages." *Blood* 15 (12): 5562–70.

Li YS., Hayakawa K., Wasserman R., and RR. Hardy. 1996. "Identification of the Earliest B Lineage Stage in Mouse Bone Marrow." *Immunity* 5 (6): 527–35.

Motonari Kondo, Irving L. Weissman, and Koichi Akashi. 1997. "Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow." *Cell* 91 (5): 661–72.

Nimmo RA, Enver T, May GE. 2015. "Primed and Ready: Understanding Lineage Commitment Through Single Cell Analysis." *Trends Cell Biol.* 11 (8): 459–67.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, Mark D., and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of Rna-Seq Data." *Genome Biology* 11 (3): R25.

Singh-Jasuja H, Ribon M, Thiolat A. 2013. "The Mouse Dendritic Cell Marker Cd11c Is down-

Regulated Upon Cell Activation Through Toll-Like Receptor Triggering." *Immunobiology* 218 (1): 28–39.

Tatiana Borodina, James Adjaye, and Marc Sultan. 2011. "A Strand-Specific Library Preparation Protocol for Rna Sequencing." *Methods Enzymology* 500: 79–98.

Treutlein B., Wu A. R., Brownfield D. G. 2014. "Reconstructing Lineage Hierarchies of the Distal Lung Epithelium Using Single-Cell Rna-Seq." *Nature* 8 (7500): 371–6.

Zandi S, Tsapogas P, Ahsberg J. 2012. "Single-Cell Analysis of Early B-Lymphocyte Development Suggests Independent Regulation of Lineage Specification and Commitment in Vivo." *Proc Natl Acad Sci* 109 (39): 15871–6.

Zeisel, A. B. Codeluppi, A. Munoz-Manchado. 2015. "Brain Structure. Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell Rna-Seq." *Science* 347 (6226): 1138–42.

Zhang J, Sugita N, Raper A. 2006. "Characterization of Siglec-H as a Novel Endocytic Receptor Expressed on Murine Plasmacytoid Dendritic Cell Precursors." *Blood* 107 (9): 3600–3608.

Zheng GX, Belgrader P, Terry JM. 2015. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications* 347 (14049).