Computational analysis of high throughput flow cytometry data of blood cells in cord blood

Antton Lamarca Arrizabalaga

Supervisor: Aitzkoa Lopez de Lapuente Portilla
Co-supervisor: Ludvig Ekdahl

Master's Thesis Project (60 ECTS)

17th June 2021

Contents

Abstract

The genetic regulation of hematopoietic stem and progenitor cells (HSPCs) plays a vital role in the development and function of the immune and circulatory systems. Genetic variants can lead to significant differences in HSPC-related phenotypes, and pathogenic variants can cause blood disorders and other maladies. Discovering these variants and understanding the mechanisms by which they affect an individual's blood could lead to advances in the treatment of such disorders. HSPCs are commonly studied in adult blood samples where the amount of progenitor cells is very small, making examining HSPCs difficult. However, umbilical cord blood samples extracted from the placenta or the umbilical cord after childbirth contain a much higher proportion of progenitor cells. Cord blood samples are significantly more challenging to obtain than adult peripheral blood samples, making large cord blood sample datasets uncommon. In this project, 3020 cord blood samples were analyzed using Flow Cytometry and pattern-recognition based gating. We carried out a replication analysis focusing on 14 SNPs known to have an effect in adult peripheral HSPCs. Additionally, a preliminary Genome-Wide Association Study was completed with 962 samples for which full genotype data was available. We were able to replicate 4 of the effects discovered in adults, and 4 potentially relevant loci were discovered containing SNPs in or close to genes that might have an effect in HSPC proliferation.

Introduction

Cord Blood and HSPCs

Umbilical cord blood is the blood that can be found in the placenta and in the umbilical cord after childbirth. Unlike adult blood, it contains high numbers of Hematopoietic Stem and Progenitor Cells (HSPCs). HSPCs are involved in the production of other blood cells, and are therefore crucial in the study of blood cell formation [1]. Because of this higher HSPC count, cord blood is more suitable than adult blood to study multipotent blood cells [2]. However, collecting cord blood samples is significantly more challenging than collecting regular adult blood samples. Because of this, large-scale cord blood sample collection is relatively uncommon, and most blood related genetic studies are performed in adult blood samples.

The Nilsson Laboratory at Lund University (https://www.hematogenomics.lu.se/) focuses on studying how genetic variation influences blood cell formation and blood cancer risk. To this end, they collect cord blood samples in order to investigate the development of HSPC subpopulations.

The main HSPC subpopulations are: hematopoietic stem cells (HSCs), multipotent progenitors (MPPs), common myeloid progenitors (CMPs), lymphoid-primed multipotent progenitors (LMPPs), common lymphoid progenitors (CLPs), granulocyte-macrophage progenitors (GMPs) and megakaryocyte-erythroid progenitors (MEPs) [1]. The differentiation hierarchy is usually represented by the hematopoietic tree, shown in Figure 1.

Flow Cytometry

Flow Cytometry (FCM) is a technique used to study and measure physical characteristics of a population of cells, such as the presence of particular cell-membrane proteins. In blood research, flow cytometry data can be used for immunophenotyping: identifying and separating the different cell-types present in a blood sample based on surface-markers. Cells are incubated with a mix of antibodies, each antibody being specific to a particular surface protein (marker) and labelled with a different fluorochrome. It is then possible to classify the cells according to their expression of combinations of surface markers by analysing them in the flow cytometer.

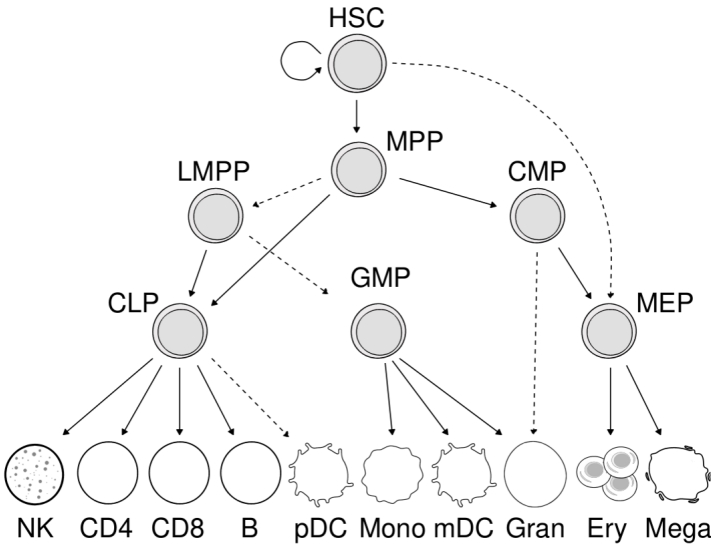

Figure 1: Hematopoietic Tree. *Figure by Ludvig Ekdahl*

Gating and gating-strategies

The process in which cells in a sample are identified and classified into different cell-types according to the surface expression of marker molecules is called flow cytometry gating, or just gating. Cells are plotted in 2-dimensions according to their marker-values, and "gates" are placed around specific populations, selecting regions and grouping cells with common characteristics. Gating can be a complex process and usually requires the application of several gates or filters in order to discern the regions of interest in the data. Performing gating manually is therefore a time-consuming process and can easily become a bottleneck in any FCM based study, particularly when dealing with large amounts of samples. Despite this, gating is still largely performed manually in flow cytometry research, using software tools such as FlowJo [3]. Automated flow cytometry gating tools have been developed [4], but gating strategies can not usually be generalized completely because of the diversity in sample characteristics and experimental setups. Because of the non-trivial nature of some classifications, flow cytometry gating is considered to be a subjective analysis.

Each type of sample requires specialised gating strategies, particularly when dealing with barely differentiated cells. Guidelines exist for the specific gating strategies of known sample types [5] and for applications such as immunophenotyping [6]. Value-ranges for specific markers are generally consistent, which allows for semi-automated gating in certain sample types if the surface-markers are known. This requires careful implementation of the automatic gates, as well as in depth knowledge about the characteristics of the cells under study.

AliGater

AliGater (https://github.com/LudvigEk/aligater) is a pattern-recognition based flow cytometry gating tool written in python. It is currently under development at the Hematogenomics Lab at Lund University. The package uses pattern recognition functions and machine-learning libraries so that the user can design the necessary flow cytometry gating strategies. It provides a framework with basic gating functionalities and can be used to build custom gating strategies based on mathematical and pattern recognition functions. Once the strategy is designed, it allows to gate several FCM files in a dynamic manner, eliminating the need for manual gating.

Genetic association studies and GWAS

Studies of genetic association intend to identify whether specific phenotypes and Single Nucleotide Polymorphisms (SNPs) co-occur more often than what could be expected by chance. To accomplish this, the genotypes and phenotypes of a cohort are analyzed to check for potential correlations between them. If a variant is more common in people with a certain phenotype, that variant is said to be *associated* with the trait.

Genome Wide Association Studies

Genome Wide Association Studies (GWAS) are, as the name indicates, association studies performed simultaneously on millions of SNPs across the entire genome. GWASs are used to identify variants that have an effect on phenotypic differences between individuals. They have been used to identify variants associated with hundreds traits and diseases, including blood-related traits [7, 8]. In the case of this project, we aim to find SNPs that have effect on the relative size of various HSPC populations observed in cord-blood samples. Phenotype data was obtained using AliGater, and genotype data for the samples was provided by DeCODE Genetics.

Methods

Designing a gating strategy

In Flow Cytometry, each of the cells that passes through the laser-sensors is referred to as an "event". The cytometer is able to detect the extent to which a surface protein is present in an event. Each event will have a vector of values associated, representing the measured intensity of each specific marker expressed on the cell surface. Different HSPC subpopulations can be defined by distinct combinations of surface markers. By hierarchically applying thresholds (gates) for these markers, specific HSPC subpopulations can be both defined and quantified. The hierarchical gating will thus result in data regarding relative counts of each kind of HSPC present in each sample.

Flow-cytometry analysis was performed in a BioRad ZE5™ machine. In flow cytometry, fluorescence intensity is not a fixed measure, and is instead variable depending on the cytometer machine and its settings. Because of this, gating is done relative to the fluorescence intensity values of all events in a single sample. This might introduce noise if the measurements are taken in different machines or over long periods of time.

The gating strategy was designed iteratively. It was initially based on the manual gating strategy developed by members of the Nilsson lab, and was later reproduced in AliGater. After the initial implementation, the cord blood samples were gated, the results evaluated, and the strategy was tweaked several times. In most cases, the tweaks were small, only altering the fixed values of certain thresholds or single steps within individual gates, but in a few cases entire gates were added or removed. In total, there were 8 major separate iterations of the gating strategy. The strategy design was assisted and approved by experts in flow cytometry every time.

Final Strategy

The final version of the gating strategy, employed for the analysis of HSPC subtypes in cord blood, consisted of 16 separate gates. A diagram of the complete strategy is presented in Figure 2. Real images of gates captured from AliGater are shown in Supplemental Figures 12 and 13.

First, the debris from dead cells was removed using 7-aminoactinomycin D (7-AAD) and forward scatter area (FSC-A). Singlet cells were gated based on FSC-A and forward scatter height (FSC-H). From singlets, cord blood mononuclear cells (CBMCs) were gated based on FSC-A and side scatter area (SSC-A). From

CBMCs, $CD34^{+}45^{low}$ cells and $CD45^{+}$ cells were separated. $CD34^{+}$ cells form a discrete cluster if the events are visualized using CD34 and CD45 markers as axes (see Supplemental Figure 12, *d* and *g*). The levels of $CD34^{+}$ cells were recorded as the number of $CD34^{+}45^{low}$ cells divided by the number of $CD45^{+}$ CBMCs.

Singlets were separated using an ellipsoid gate based on principal component analysis (PCA), and the CBMC gate used the Dijkstraa's shortest path algorithm to separate granulocytes from CBMCs. Debris with low scatter values were also removed at this gate. The $CD34^{+}$ cluster was identified by finding the highest intensity value among the $CD45^{+}34^{+}$ cells, which often corresponds to the center of the cluster. This point was used to identify the area where CD34 negative and positive cells are divided, and the cluster was then isolated by diagonally gating at 45 degrees above and below this area.

The CBMCs, minus the $CD45^{low}CD34^{-}$ cells, were then gated four consecutive times to ensure they had low values of the lineage markers CD3, CD19, CD14, CD16 and CD56. The resulting lineage negative cells were gated again with the exact same gate used to separate $CD34^{+}$ events, this time resulting in a $Lin^{-}CD34^{+}$ population. These events were later divided into CD38 positive and negative cells.

The four gates involving lineage markers used the lowest density point in the corresponding marker-axis to separate the cells into two groups each time. Marker axes represent fluorescence intensity, and these "valleys" therefore correspond to a clear lack of cells with intermediate fluorescence values. The CD38 gate was fixed, considering the 30% of cells with the lowest CD38 values to be $CD38^{-}$ cells, while the rest were determined to be $CD38^{+}$. B/NK progenitors, CMPs, GMPs, MEPs, HSCs, MLPs and MPPs were determined in each case by different thresholds on CD10, CD135 and CD45RA values. These thresholds were determined by first identifying the highest density point in the distribution of the relevant marker values, and then locating a point in a predefined interval where the density dropped below a specific fraction of the highest value.

Experimenting with the CD38 gate

CD38 is a glycoprotein found on the membrane of many immune cells and absent in *bona fide* hematopoietic stem cells (HSC) with self-renewal capacity. As such, it is one of the most important markers to define HSCs. While designing the gating strategy, one of the steps that was discussed most extensively was the CD38 gate. Unlike most of the other gates, which are designed in such a way so that the populations can be easily discerned visually, the standard CD38 gate presents a single cluster of events (see Supplemental Figure 13 *l*). As stated above, this is a fixed gate, where the events with the bottom 30% of the values are considered to be CD38 negative, while the rest of the values are taken as CD38 positive.

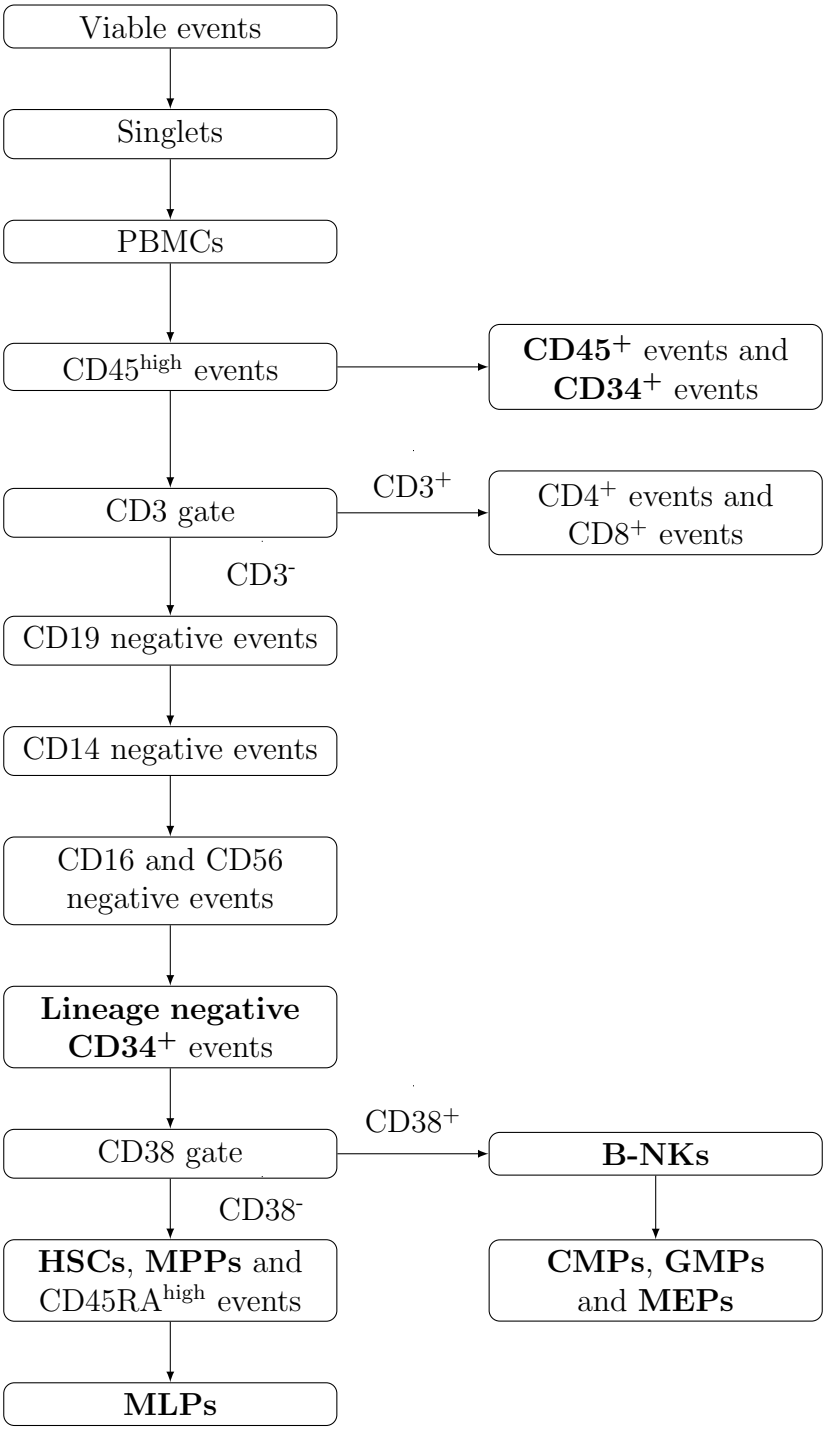

Figure 2: Diagram of the complete final gating strategy.

Most authors use the 30% threshold to determine CD38 positivity [9], but other thresholds and approaches have also been proposed, such as using the mean value (50% threshold) [10] or using pixel intensity [11]. In order to check if there was any way of separating the two populations (CD38 positive and negative) in a cleaner way, we looked at the problem multi-dimensionally. We calculated correlation between CD38 and known associated markers like CD90, and looked at 3D plots of promising marker pairs. We also examined if there were any clusters when doing Principal Component Analysis (PCA) or drawing t-Distributed Stochastic Neighbor Embedding (t-SNE) maps [12].

Unfortunately, at least for our FCM marker setup, no marker or marker combination appears to be more effective than the current standard for the binary classification of primitive and differentiated cells. While there is no clear divide, a low CD38 threshold will consistently result in a subset significantly enriched for primitive cells.

Quality Control and Repeat Analysis Correlation

Technical errors can occur during sample collection, flow cytometry and gating steps. Natural data variability alone can lead to inaccurate gating. All of this can introduce noise in the results when calculating the final phenotype values of a sample.

While technical errors that occurred before gating could not be corrected, the gating strategy itself was designed with the intention of keeping gating mistakes to a minimum. Even so, inter-sample variability caused a few of the samples to be incorrectly processed by AliGater. These, on top of samples with faulty FCM data, resulted in a small subset of samples whose phenotype data was incorrect. Knowing this, extensive quality control was performed by checking the images resulting from the gating process to try to identify and discard such samples.

AliGater contains native Quality Control functionalities that allow for the storage and comparison of down-sampled images of the desired gating steps. In addition to this, full-resolution images of the gating were stored for each sample at each major gate. The down-sampled, 128x128 resolution images were stored for the CBMC, $CD34^+$ and Lin^-CD34^+ gates. These images were later used in a PCA to detect samples that differed considerably from the rest. Those samples with maximal and minimal values for each of the principal components were selected as potentially faulty. The full resolution images were then used to manually check the samples detected as outliers in the PCA. A total of 239 samples were deemed faulty and were discarded.

Correlation between repeated measurements

In order to ensure the robustness of AliGater, we compared AliGater gating and manual gating. Of the automatically gated samples, 697 cord blood samples had also been manually gated beforehand, and of these, 56 samples had been analyzed by FCM in at least two separate occasions. This produced two separate Flow Cytometry Standard (FCS) files for each of the samples, which could be used to compare the consistency of both gating methods.

The $Lin^-CD34^+/CD45^+$, $HSC/CD45^+$ and HSC/Lin^-CD34^+ phenotypes were collected for both FCS files of the 56 samples, using both manual and AliGater gating. For each method, Pearson correlation was computed between the two instances of each unique sample. Finally, the data was represented in a cumulative distribution plot. This allowed us to observe how consistent the results of a repeat analysis were for each of the gating methods.

Genetic Association

The Nilson lab has performed GWAS studies using adult peripheral blood samples [13]. These studies pointed at 14 SNPs in different regions of the genome that could have an effect on the overall $CD34^+$ levels in peripheral blood (see Table 1). There was interest in examining whether or not these effects were replicated in the cord blood samples, so before carrying out a complete GWAS on the cord blood data, an association study for the 14 candidate SNPs was performed.

A collection of several HSPC subpopulation frequencies, 8 out of the 33 observed phenotypes, were selected for the association study (see Table 2). A gating strategy to measure the values for these phenotypes based on FCM data was designed in AliGater. A total of 2631 samples were gated, and the AliGater output values were later normalized using logarithmic normalization.

Genotype data was provided by DeCODE Genetics for 760 cord blood samples, which allowed us to determine the specific alleles at the locations of the candidate SNPs. After joining the data and performing quality control and normalization, we had a total of 694 samples for which we had phenotype and genotype data. This allowed us to look for genetic association.

A particular phenotype will usually be affected by several genes, which is referred to as an additive genetic effect. A singular SNPs will therefore be behind just a fraction of the total variability on the phenotype. In order to test if any one of the candidate SNPs had an effect on any of our HSPC subpopulations, ordinary least-squares (OLS) regression was used with allelic content as the independent variable and the phenotype values as the dependent variable for each of the SNP-phenotype combinations.

It is important to keep in mind that OLS is quite susceptible to outliers, particularly when the sample size is relatively small. This makes the results very dependent on quality control and normalisation.

Table 1: List of Candidate SNPs.

Variant rsID	**Proxy rsID**	**Associated gene**
rs2047094	-	ENO1
rs309137	-	CXCR4
rs11688530	-	CXCR4
rs555647251	-	CXCR4
rs10193623	rs6726457	CXCR4
rs201494641	s17227404	ITGA9
rs7705526	-	TERT
rs1029094211	-	TTC1
rs1991866	-	CCDC26
rs699585	-	PPM1H
rs117701013	-	STXBP6
rs35532684	-	ARHGAP45
rs12975577	-	CEBPA

Table 2: List of HSPC subpopulations used as phenotypes.

Phenotype name	Marker combination
CD34$^+$/CD45$^+$	$45^{high}34^+$ out of 45^+
Lin$^-$CD34$^+$/CD45$^+$	Lin^-34^+ out of 45^+
HSC/CD34$^+$	$Lin^-34^+38^-45RA^-90^+$ out of $45^{high}34^+$
MPP/CD34$^+$	$Lin^-34^+38^-45RA^-90^-$ out of $45^{high}34^+$
B-NK/CD34$^+$	$Lin^-34^+38^+10^+$ out of $45^{high}34^+$
B-NK/CD38$^+$	$Lin^-34^+38^+10^+$ out of $Lin^-34^+38^+$
CMP/CD34$^+$	$Lin^-34^+38^+10^-45RA^-135^+$ out of $45^{high}34^+$
MEP/CD34$^+$	$Lin^-34^+38^+10^-45RA^-135^-$ out of $45^{high}34^+$

SNPs and Proxies

Five of the candidate SNPs provided by the Nilsson lab in Lopez de Lapuente et al. [13] were not present in the genotype data due to them not being included in the genotyping array nor being included in the imputation. Two of those, however, had SNPs in high Linkage Disequilibrium (LD) that did appear in the genotyping. For the two variants rs10193623 and rs201494641, the analysis was performed with 'proxy' SNPs instead (rs6726457 and rs17227404 respectively).

Conditional analysis for CXCR4

Of the remaining 11 candidate SNPs, three were located in chromosome 2, roughly in a 20kb range, and were suspected to be related to the gene CXCR4. These three SNPs are not in LD, and are therefore inherited independently. Their effects are therefore independent, and even occur in different directions for the same phenotype. Therefore, we performed conditional analysis when checking for effects in cord blood [14]. The association analysis was repeated adjusting for every SNP believed to be associated with CXCR4, correcting for the effect of the other SNPs each time.

Additional analysis of PPM1H variants

After an overall analysis of the candidate SNPs, it was determined that the variant rs699585, most probably related to the gene PPM1H, required deeper study. The gene was suspected to be involved in the regulation of overall CD34+ levels, and the association analysis of the SNP with 694 samples in cord blood marginally replicated effects in two distinct HSPC subpopulations: B-NK progenitors and CMPs. Deeper understanding of the effects of this variant could help unveil the mechanisms by which PPM1H operates, and was therefore of great interest.

Because of the limited sample size, however, it was difficult to draw definitive conclusions. In order to solve this, additional 939 samples were manually genotyped at the specific locus of the SNP. The increased sample size allowed for a more robust association study, with a total of 1680 samples to build the linear regression model on.

Issues with multiple testing

While looking for effects in all of the 33 collected phenotypes might seem appealing, we had to be mindful of the multiple testing problem. The number of observed phenotypes had to be kept to a minimum, and Bonferroni correction was applied to determine the correct threshold of p-value for a significant effect, adjusting for the 8 observed phenotypes.

Pilot GWAS

Cord blood samples phenotyped using Flow Cytometry gating were used to complete a small scale pilot GWAS. The GWAS was carried out on Hail [15] version 0.2.60., using data from 692 cord blood samples.

Data preprocessing

Population stratification can confound a GWAS study because of systematic, ancestry-based differences on the data. While no additional sample-donor ancestry data existed, principal components obtained through PCA can be used as covariates fulfilling a similar role [16].

Before performing a PCA, the variants were filtered based on allele frequency (AF), taking only variants with AF > 0.05 and thus removing rare variants. SNPs with Hardy-Weinberg equilibrium (HWE) exact test p-value below 1e-6, far from HWE, were also removed to avoid potentially incorrectly genotyped variants. LD pruning was also performed to lighten the computational load by taking as few variants as possible from each LD block, with r^2 imputation > 0.3 and a window size of 1000kb. The filtering, as well as a PCA on the filtered cord blood genotype data, was carried out in PLINK 2.0 [17].

The phenotype data were obtained by FCM using AliGater. It was normalized using rank-based inverse normal transformation (INT). INT is the most common normalization approach for non-normally distributed phenotype data in GWAS [18].

Hail

After the outlying samples were removed, the genotype data of the rest was loaded into Hail. In Hail, a Genome Wide Association Study was performed using the cord blood phenotype data obtained by FCM. The analysis was repeated for each one of the 8 phenotypes described in Table 2.

The previously obtained principal components were used as covariates to minimize the impact of population stratification. The results were displayed in Hail and later a Manhattan plot was drawn for each of the phenotypes using the python tool "manhattan_generator" (https://github.com/pgxcentre/manhattan_generator).

Results

Robustness and Reproducibility

Figure shows a comparison between the distributions of each phenotype for manual gatig versus AliGater gating. The cumulative distribution plots of the repeat-analysis correlation (Figure 4) show that, overall, gating with AliGater is more consistent than manual gating. Repeat correlation data for each individual phenotype can be found on Supplementary Figure 14.

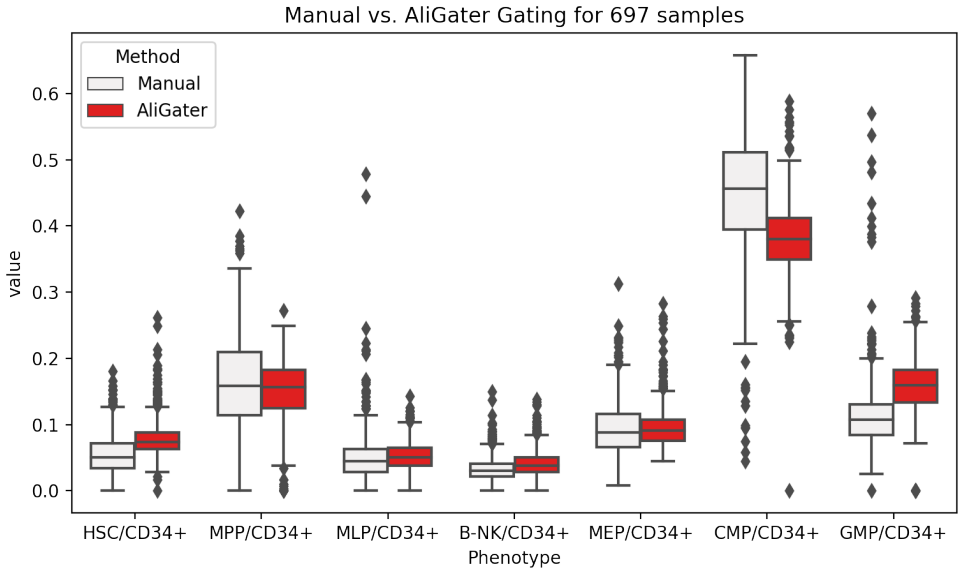

Figure 3: Manual and AliGater gating results distribution comparison.

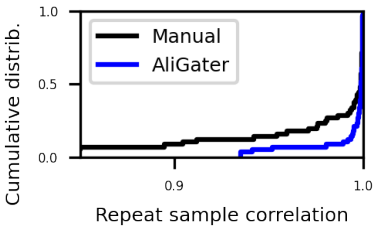

Figure 4: Repeat Measurement Correlation. Cummulative distribution plot.

Data exploration: comparison to Mantri et al.

Collection of cord blood samples is relatively rare, and as such no big databases exist on cord blood data. One pilot study exploring the relative quantities of each separate HSPC subpopulation in cord blood samples was Mantri et al. [19]. This study followed an experimental design similar to our own, using flow cytometry data of 50 cord blood samples to classify cells based on the expression of cell surface antigens. While the phenotypes they used were not perfectly comparable in some cases, most of them were similar enough to warrant checking if the values reported in their paper were close to the ones observed in our samples.

Once gating and quality control were completed, the percentages of the main HSPC subpopulations out of total $CD34^+$ cells were collected for 2631 cord blood samples. These values were compared to those presented by Mantri et al. Calculations out of total $CD34^+$ cells are shown in Table 3. The median value is reported in each case, as well as the complete range of values. The results are shown compared to those reported by Mantri et al.

Table 3: Percentage of HSPC subpopulations. Comparison to data presented in Mantri et al. Formatted as *median (range)*.

Cell Type	**Immunophenotype**	**Percentages Mantri et al.***	**Percentages**
HSC	$Lin^-34^+38^-45RA^-90^+$	6.7(0.1-17.3)	6.46(0.61-21.22)
MPP	$Lin^-34^+38^-45RA^-90^-$	10.6(1.1-37.5)	12.33(0.45-22.43)
CMP	$Lin^-34^+38^+10^-45RA^-135^+$	33.6(3.2-45.8)	31.37(7.46-47.88)
GMP	$Lin^-34^+38^+10^-45RA^+135^+$	20.8(7.1-51.9)	12.09(1.64-25.88)
MEP	$Lin^-34^+38^+10^-45RA^-135^-$	11.4(3.8-22.9)	8.14(1.2-31.85)
B-NK	$Lin^-34^+38^+10^+$	1.2(0.2-4.1)	2.85(0.14-12.42)
MLP	$Lin^-34^+38^-45RA^+90^-10^+$	1.1(0.3-4.6)	2.51(0.06-12.59)

**data from Mantri et al.* [19]

The study by Mantri and colleagues claimed there was no correlation between the total $CD34^+$ cells and HSC levels in a sample. In order to test this claim, a linear regression plot between the two phenotypes was drawn for our 2631 cord blood samples. Contrary to what the paper states, our analysis points to a clear linear relationship between the two, with a r^2 of 0.7149.

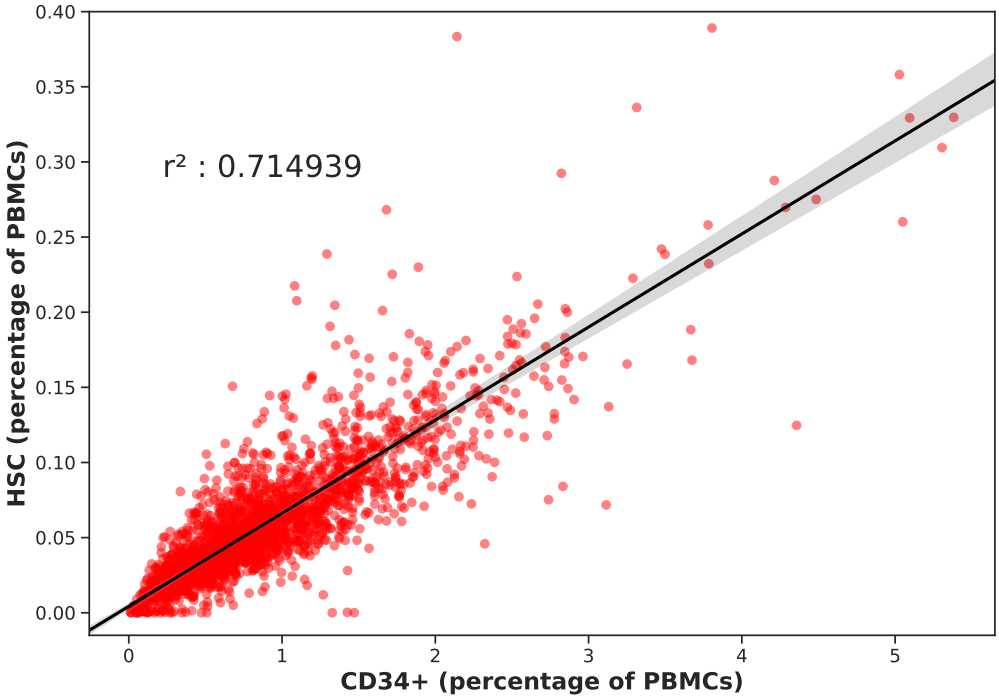

Figure 5: Linear regession between HSCs and total $CD34^+$ cells.

Effects of candidate SNPs

Figure 6 shows the most significant effects for the candidate SNPs. The SNP rs2047094, believed to be related to the gene ENO1, seems to have a significant effect on the $CD34^+/CD45^+$ phenotype.

For the SNP rs699585, located inside of the gene PPM1H, it is possible that there is an effect on B-NK progenitors in cord blood. Both the B-NK/$CD38^+$ and B-NK/$CD34^+$ phenotypes show an upwards tendency for the G allele.

The variant rs11688530 is believed to be related to the gene CXCR4. It could have an effect on the MPP/$CD34^+$ phenotype. Additionally, it also shows an upwards tendency for the A allele in the $Lin^-CD34^+/CD45^+$ phenotype. Interestingly enough, variant rs309137, also close to CXCR4 shows an effect on the same phenotype but with opposite effect.

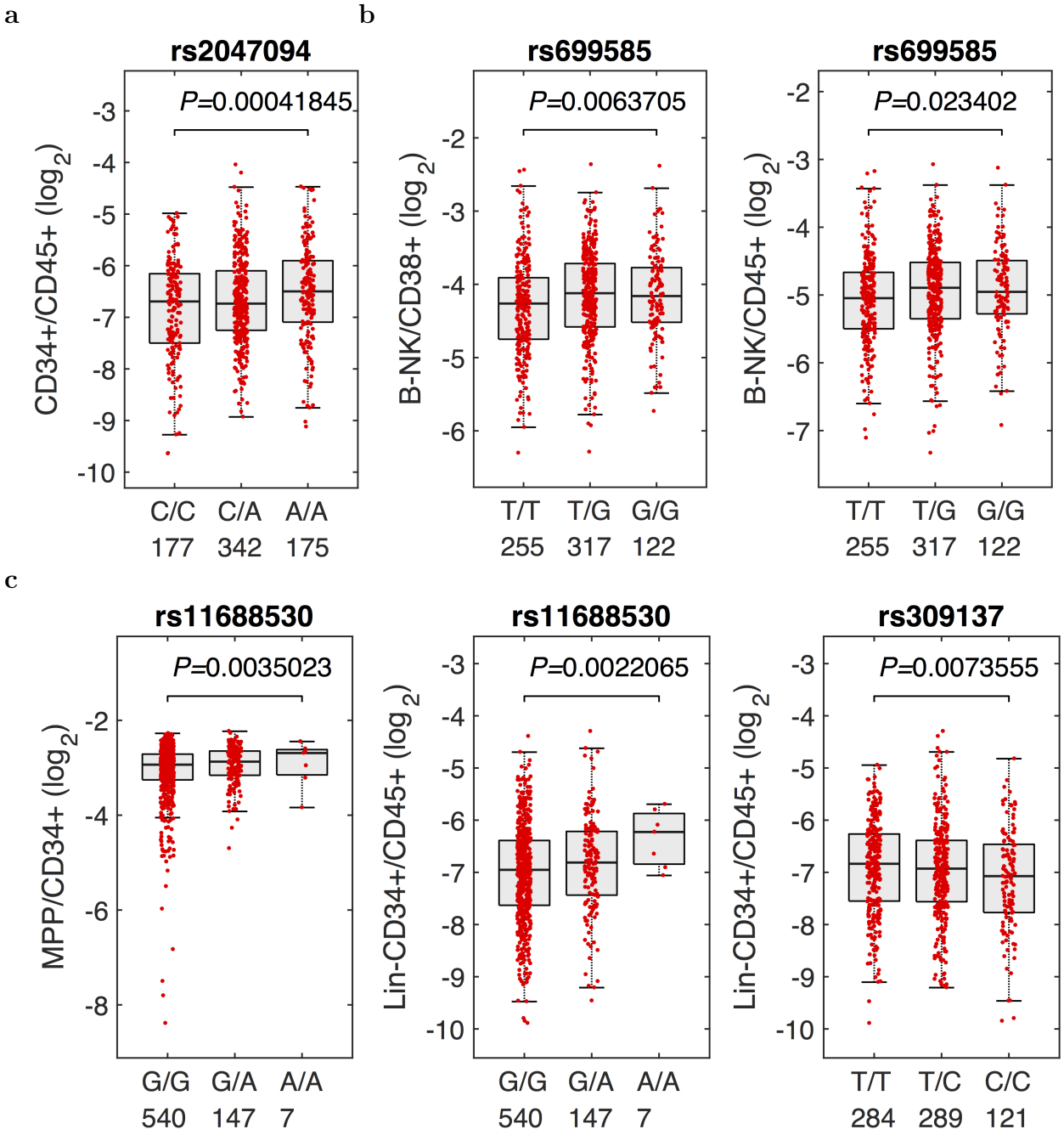

Figure 6: **a)** Variant rs2047094, related to gene ENO1. **b)** SNP rs699585 in gene PPM1H. **c)** Variants rs11688530 and rs309137, both related to gene CXCR4.

The effects for the variant in PPM1H with the increased sample size can be seen in Figure 7. When increasing the sample size SNP rs699585, located in the gene PPM1H, shows an effect in the CMP subpopulation. The previously observed effect in B-NK progenitors (Figure 6 *b)*) however seems to significantly decrease in strength.

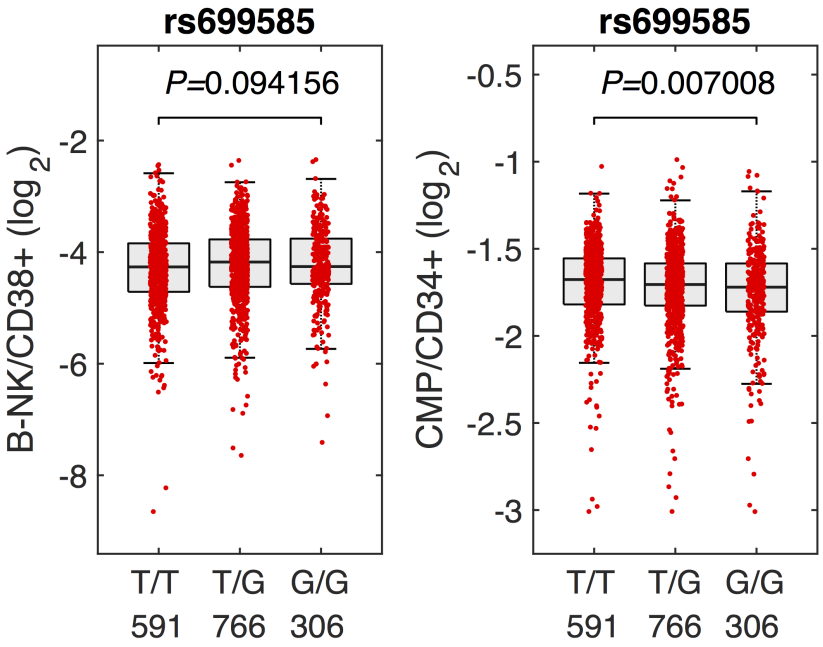

Figure 7: Effect for SNP rs699585, in gene PPM1H, for 1680 samples. The effect on the B-NK phenotype disappears after the sample size increase. The effect in CMPs becomes much more significant.

GWAS hits

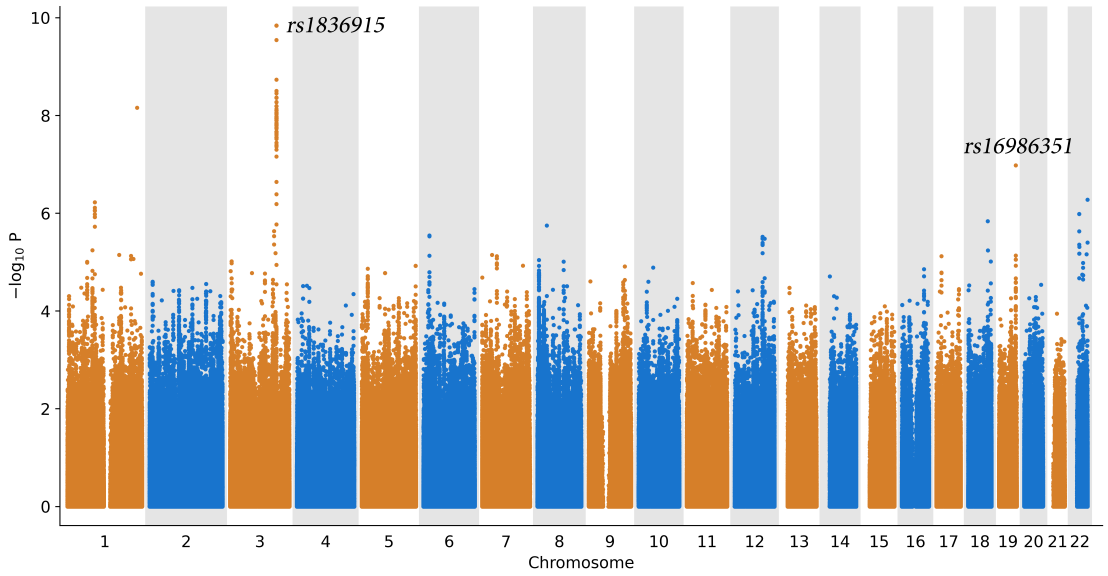

Figure 8: Manhattan plot of pilot GWAS results for the B-NK/CD34$^+$ pehenotype. There were two main hits for the B-NK phenotype: rs1836915 in chromosome 3, and rs16986351, in chromosome 19.

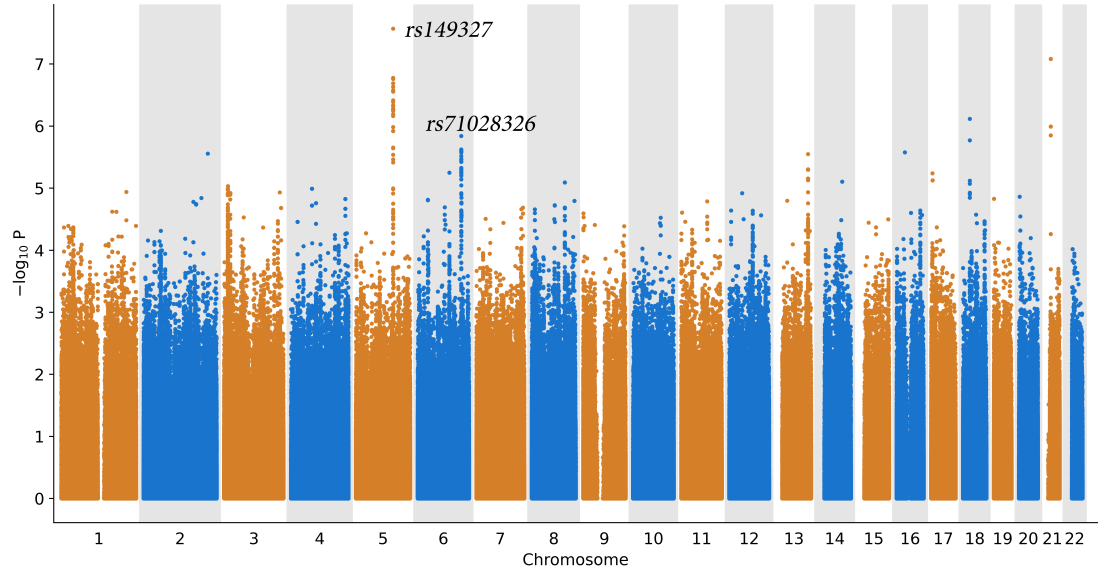

Figure 9: Manhattan plot of pilot GWAS results for the CD34$^+$/CD45$^+$ pehenotype. The main two hits were rs149327, located in chromosome 5, and rs71028326, in chromosome 6.

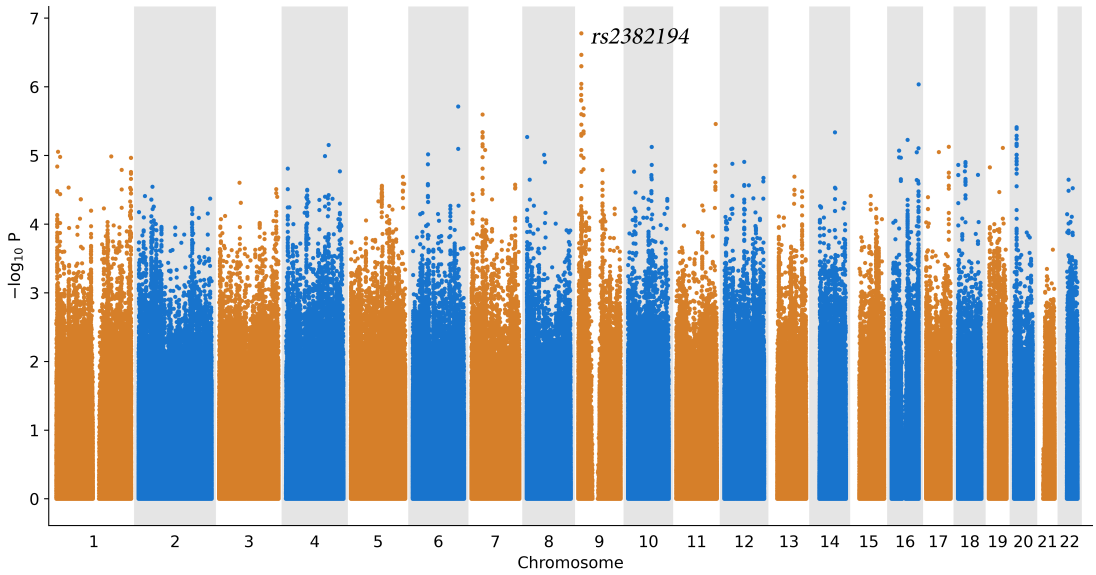

Figure 10: Manhattan plot of pilot GWAS results for the HSC/CD34$^+$ pehenotype. The main hit was rs2382194, in chromosome 9.

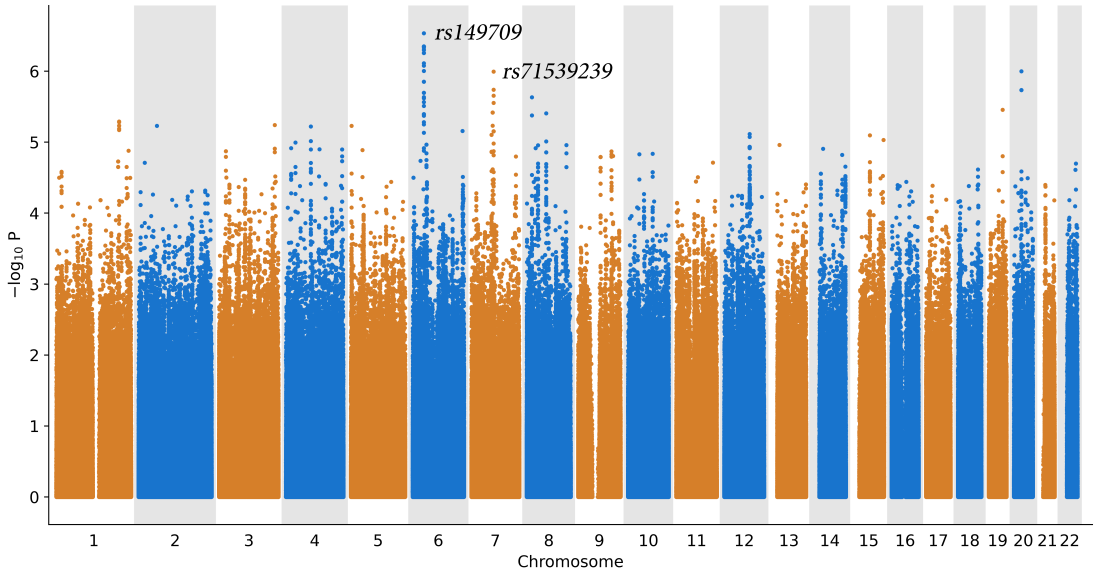

Figure 11: Manhattan plot of pilot GWAS results for the MEP/CD34$^+$ pehenotype. The main hits are rs149709 in chromosome 6 and rs71539239 in chromosome 7.

Discussion

Comparison to the results of Mantri et al.

The main conclusion obtained from comparing our results to those obtained by Mantri et al. [19] is that there seems to be, contrary to what is stated in the paper, a linear relationship between the overall $CD34^+$ cell population and HSC levels in cord blood. It appears that the restricted sample size was the reason why this was not observed before.

Beyond this, the results obtained in this study seem consistent with what was reported in their paper (see Table 3). We see high inter-sample variability, with the medians and ranges being similar in both cases. The HSPC subpopulation with the biggest discrepancy in median values are GMPs. This might be due to the differences in the markers used to define the population.

Association of candidate SNPs

While only some of the 14 SNPs with an effect in adult blood seem to replicate in cord blood, four of them do show promising effects. In particular, variant rs2047094, close to gene ENO1, seems to have a clear effect in the overall $CD34^+$ cell population. This effect was also observed in adults [13].

The effect of SNPs rs309137 and rs11688530, both related to gene CXCR4, seem to both have an effect in the $Lin^-CD34^+/CD45^+$ phenotype. Interestingly, however, their effects seem to be in opposite directions. Lopez de Lapuente et al. reports opposing directions for overall CXCR4 expression for these two SNPs as well. Variant rs11688530 might also have an effect in the MPP/$CD34^+$ and B-NK/$CD34^+$ phenotypes.

A few other of the observed variants seem to have slight trends for one of the alleles in several different phenotypes, but do not cross the threshold for significance. It might be that increasing the sample size would make these effects more pronounced.

While not Bonferroni significant, the SNP rs699585 showed a considerable upwards tendency for the G allele in the B-NK/$CD34^+$ phenotype. It was determined that more cord blood samples would be genotyped at that locus to increase the amount of available data for the association, hoping that the signal would become stronger. However, the increased sample size made the signal in the B-NK phenotypes disappear, which suggests the association was coincidental. Instead, it made the signal in CMPs go from a slight tendency to a reasonably convincing effect.

This is also consistent with what was reported in the adult blood GWAS study [13].

GWAS

The results of the pilot GWAS showed a number of interesting peaks despite the small sample size. The hit in chromosome 3 for the B-NK phenotype (Figure 8) is the most significant, and the most convincing after some followup analysis. The SNP rs1836915 is located within the gene MME, a protein coding gene whose protein is present on leukemic cells of pre-B phenotype. It is expressed in whole blood, mostly in polymorphonuclear cells from bone marrow and peripheral blood (PMN-BM and PMN-PB). The variant itslef has also been linked to Alzheimers' disease in a previous association study [20].

A few of the other hits also show features that could potentially increase their beliveabilty. The second hit on the same phenotype, rs16986351, is located in the TMEM150B gene. This gene codes for a membrane protein in the "damage-regulated autophagy modulator" (DRAM) family, and it is predominantly expressed in the small intestine, but also in whole blood. Variant rs149327, which appeared in the CD34 phenotype, is one of the hits with best significance value. It is located close to the ZNF608 gene. The hit in HSCs, rs2382194 is in gene PTPRD, which is somewhat expressed in HSCs. Finally, the SNP rs149709, which appeared in the MEP phenotype, was identified as being linked to blood presure by a previous GWAS [21].

Sample Size

This pilot GWAS was carried out using only 698 cord blood samples, which is generally considered utterly insufficient for this kind of analysis. Phenotype data for over 3000 samples is already available as presented in the data exploration segment of this report. However, these samples are yet to be genotyped. Once the genotype data exists, the GWAS analysis can be easily upscaled to meet the requirements for a standard genome wide association study.

Limitations and future work

The most obvious limitation of this study was the limited sample size. This work will be expanded upon in the near future when genotype data for additional samples becomes available.

Some of the data filtering steps common in GWAS were skipped in this study to avoid decreaseing the sample size further. In a full scale GWAS, Identity by Descent (IBD) could be used to filter out related individuals such as siblings or cousins. Technical covariates, such as time of collection of a sample, or plate number, could be incorporated into the study to avoid confounding or control for batch effects. And biological covariates, such as the gender and ethnic background of the sample donors should also be used. Tools such as ADMIXTURE [22] would allow for more in depth ancestry analisys.

On the phenotying side, the gating strategy could still be perfected further to properly deal with samples that the current strategy simply discards. Cord blood samples are rare, and we should discard as few of them as is possible. It would also be interesting to further analyze the possibility of a more precise CD38 gate.

Acknowledgements

I would like to thank my supervisor, Dr. Aitzkoa Lopez de Lapuente Portilla for her guidance and advice during the development of this project. My warmest thanks to my co-supervisor, Ludvig Ekdahl, for his invaluable improvised lectures, thoughtful explanations and infinite patience. My thanks to Professor Björn Nillson, for welcoming me into his research team. Special thanks to Natsumi Miharada and Aurelie Baudet for their hard work on manual flow cytometry gating and sample data collection. Thanks as well to all other members of the Nilsson lab. Finally, thanks to everyone involved in the cord blood sample collection process, as well as to everyone whose code I used during the project.

Bibliography

[1] Faiyaz Notta et al. 'Distinct routes of lineage development reshape the human blood hierarchy across ontogeny'. In: *Science* 351.6269 (2016).

[2] Jennifer D Newcomb et al. 'Umbilical cord blood research: current and future perspectives'. In: *Cell transplantation* 16.2 (2007), pp. 151–158.

[3] Dickinson Becton and Company. *FlowJo™ Software*. [software application]. Ashland, OR. 2019.

[4] Chris P Verschoor et al. 'An introduction to automated flow cytometry gating tools and their implementation'. In: *Frontiers in immunology* 6 (2015), p. 380.

[5] D Barnet et al. 'Guideline for the flow cytometric enumeration of CD34+ haematopoietic stem cells'. In: *Clin Lab Haem* 21 (1999), pp. 301–308.

[6] Janet Staats et al. 'Guidelines for gating flow cytometry data for immunological assays'. In: *Immunophenotyping*. Springer, 2019, pp. 81–104.

[7] Jacob C Ulirsch et al. 'Systematic functional dissection of common genetic variation affecting red blood cell traits'. In: *Cell* 165.6 (2016), pp. 1530–1545.

[8] Jacob C Ulirsch et al. 'Interrogation of human hematopoiesis at single-cell and single-variant resolution'. In: *Nature genetics* 51.4 (2019), pp. 683–693.

[9] Joke G Boonstra et al. 'CD38 as a prognostic factor in B cell chronic lymphocytic leukaemia (B-CLL): Comparison of three approaches to analyze its expression'. In: *Cytometry Part B: Clinical Cytometry: The Journal of the International Society for Analytical Cytology* 70.3 (2006), pp. 136–141.

[10] Prashant Ramesh Tembhare et al. 'Flow cytometric evaluation of CD38 expression levels in the newly diagnosed T-cell acute lymphoblastic leukemia and the effect of chemotherapy on its expression in measurable residual disease, refractory disease and relapsed disease: an implication for anti-CD38 immunotherapy'. In: *Journal for immunotherapy of cancer* 8.1 (2020).

[11] Piers EM Patten et al. 'CD38 expression in chronic lymphocytic leukemia is regulated by the tumor microenvironment'. In: *Blood, The Journal of the American Society of Hematology* 111.10 (2008), pp. 5173–5181.

[12] Laurens Van der Maaten and Geoffrey Hinton. 'Visualizing data using t-SNE.' In: *Journal of machine learning research* 9.11 (2008).

[13] Aitzkoa Lopez de Lapuente Portilla et al. 'Genome-wide association study on 13,167 individuals identifies regulators of hematopoietic stem and progenitor cell levels in human blood'. In: *bioRxiv* (2021). DOI: 10.1101/2021.03.31.437808. eprint: https://www.biorxiv.org/content/early/2021/04/03/2021.03.31.437808.full.pdf. URL: https://www.biorxiv.org/content/early/2021/04/03/2021.03.31.437808.

[14] Jian Yang et al. 'Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits'. In: *Nature genetics* 44.4 (2012), pp. 369–375.

[15] Hail Team. *Hail 0.2.60-de1845e1c2f6*. https://github.com/hail-is/hail/commit/de1845e1c2f6.

[16] Alkes L Price et al. 'New approaches to population stratification in genome-wide association studies'. In: *Nature Reviews Genetics* 11.7 (2010), pp. 459–463.

[17] Christopher Chang et al. 'Second-generation PLINK: Rising to the challenge of larger and richer datasets'. In: *GigaScience* 4 (Oct. 2014). DOI: 10.1186/s13742-015-0047-8.

[18] Zachary R McCaw et al. 'Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies'. In: *Biometrics* 76.4 (2020), pp. 1262–1272.

[19] Sruthi Mantri et al. 'CD34 expression does not correlate with immunophenotypic stem cell or progenitor content in human cord blood products'. In: *Blood advances* 4.21 (2020), pp. 5357–5361.

[20] Linda S Wood et al. 'Association between neprilysin polymorphisms and sporadic Alzheimer's disease'. In: *Neuroscience letters* 427.2 (2007), pp. 103–106.

[21] Daniel Levy et al. 'Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness'. In: *BMC medical genetics* 8.1 (2007), pp. 1–11.

[22] David H Alexander, John Novembre and Kenneth Lange. 'Fast model-based estimation of ancestry in unrelated individuals'. In: *Genome research* 19.9 (2009), pp. 1655–1664.

Supplement

The code used in the development of this project can be found at:
https://github.com/AnttonLA/Masters-Thesis-Code.

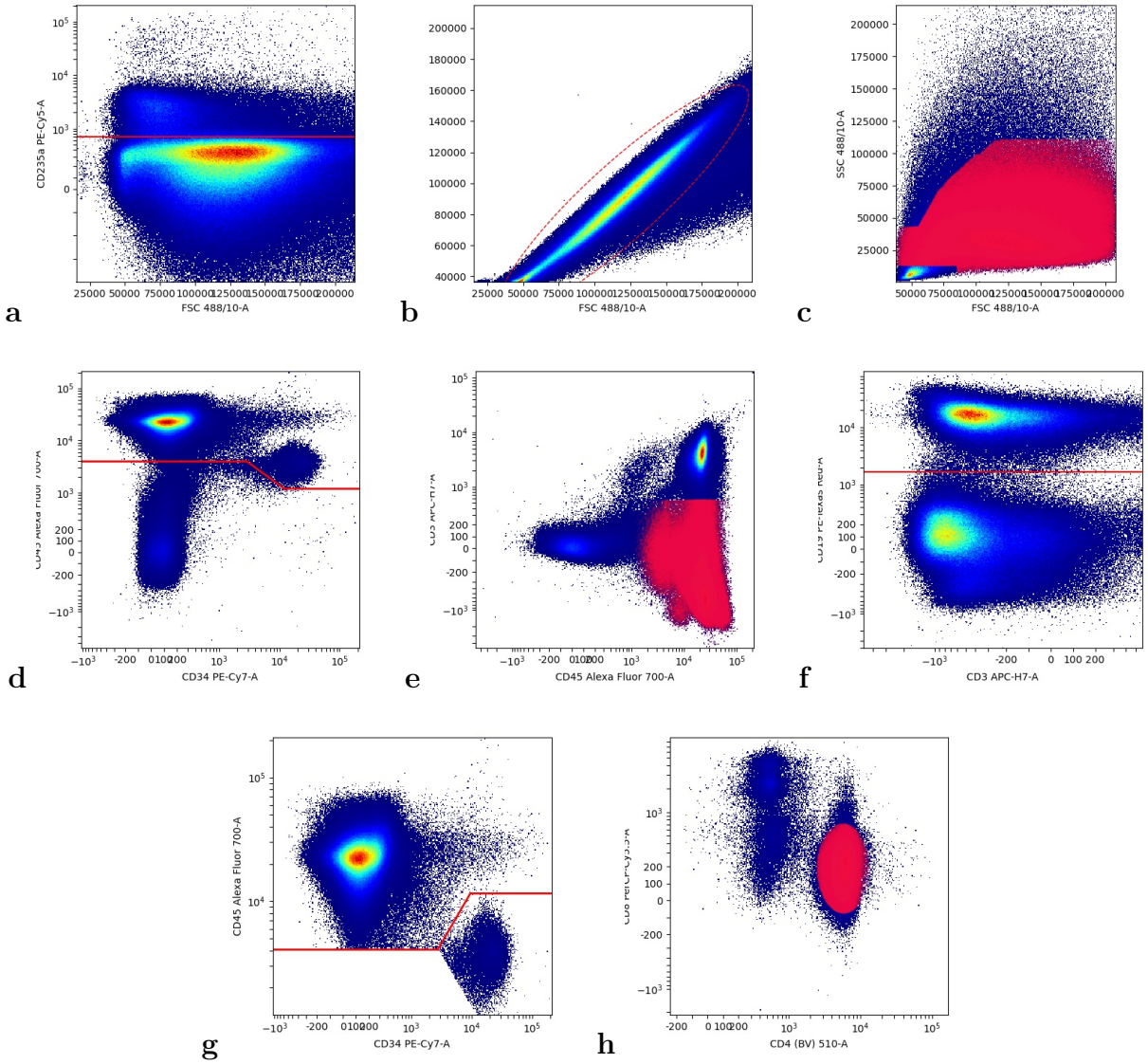

Figure 12: **a)** Debree gate. Lower part taken. **b)** Singlet gate. **c)** PBMC gate. **d)** CD45 and CD34 gate. Upper part taken. **e)** CD3 gate. **f)** CD19 gate. Lower part taken. **g)** CD34 cluster gate. Comes from d). **h)** CD4 CD8 gate. Comes from e), $CD3^+$ events.

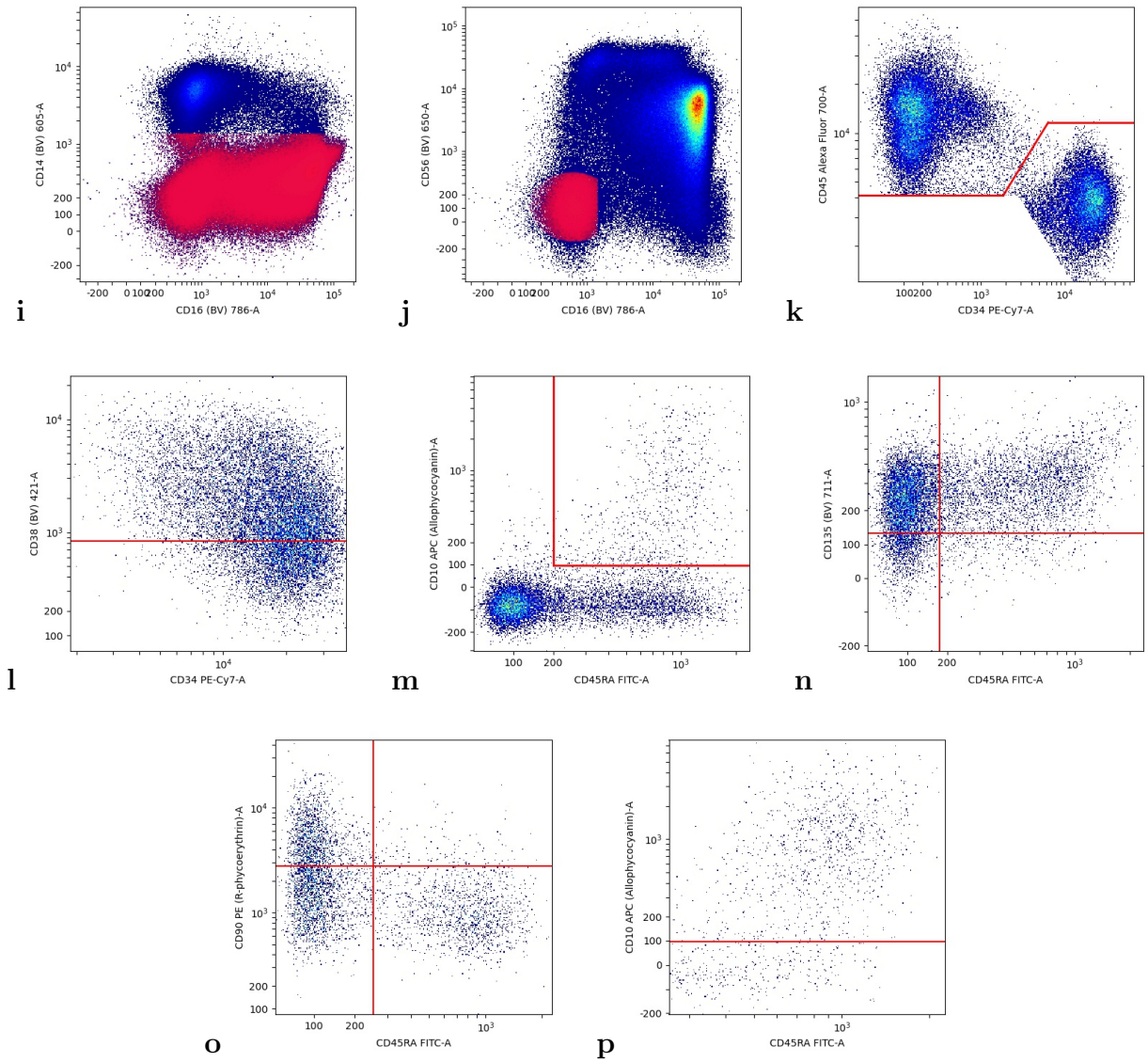

Figure 13: **i)** CD14 and CD16 gate. Comes from f). **j)** CD16 and CD56 gate. **k)** Linage negative gate. Identical to g). **l)** CD38 gate. **m)** B-NK gate. Comes from CD38 positives. Upper part are B-NK events. **n)** CMP, GMP and MEP gate. Comes from lower part of m). **o)** HSC, MPP and pre-MLP gate. Comes from l), CD38 negative events. **p)** MLP gate.

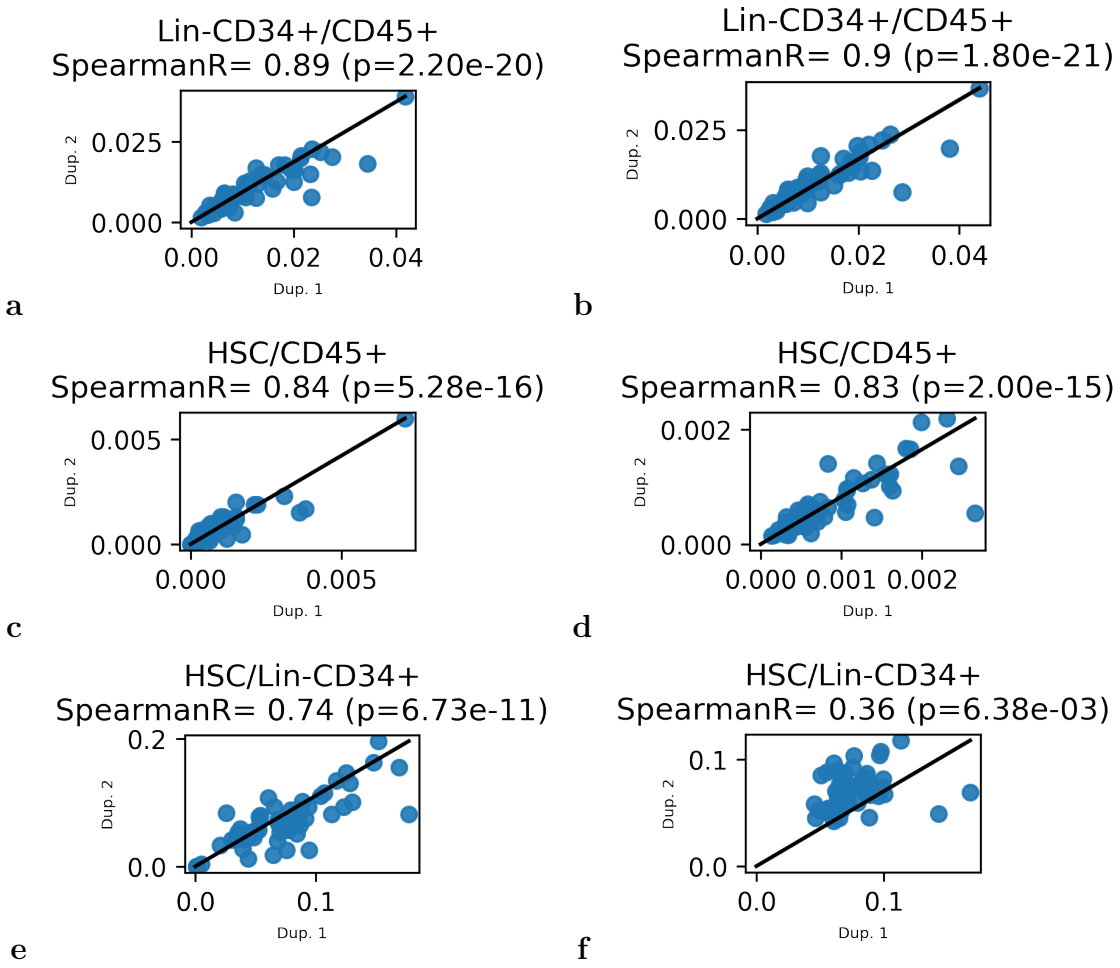

Figure 14: Correlation between the values of the same phenotypes in samples that were analyzed twice. **a)** Manual gating, Lineage negative CD34$^+$. **b)** AliGater gating, Lineage negative CD34$^+$. **c)** Manual gating, HSCs out of CD45$^+$. **d)** AliGater gating, HSCs out of CD45$^+$. **e)** Manual gating. HSCs out of Lineage negative CD34$^+$. **f)** AliGater gating. HSCs out of Lineage negative CD34$^+$.