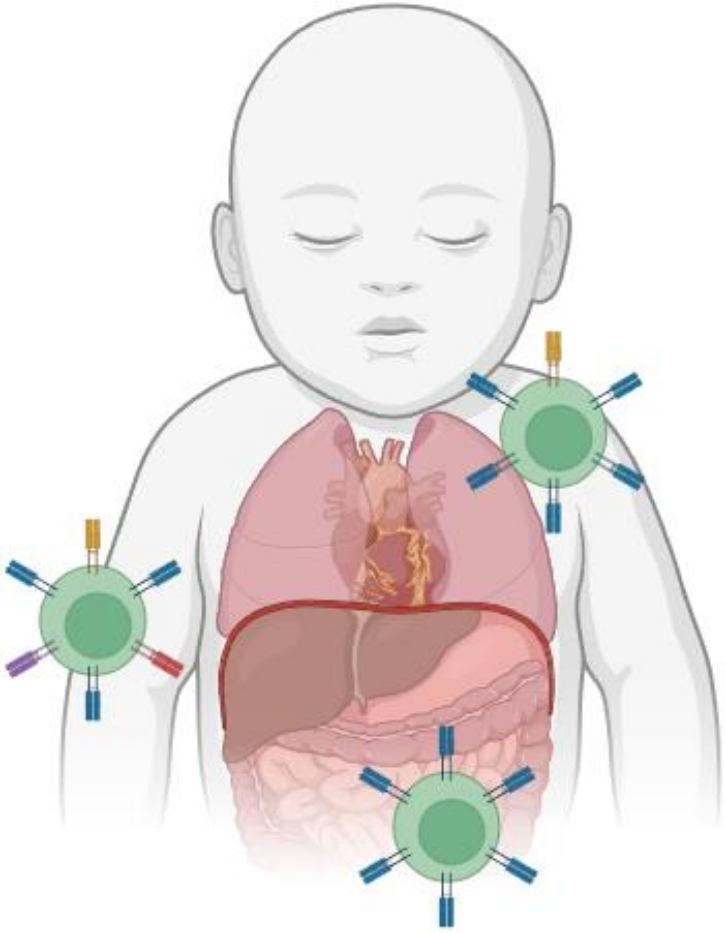
Identifying the immunological drivers of early life homing of human naïve T cells



Marle Lokerse
Utrecht University

7 September 2023

Student number: 6365280
E-mail address: m.e.lokerse@students.uu.nl
Daily supervisor: E.S. Saager
Examiner: F. van Wijk
Second examiner: E.M. Delemarre

Abstract

Naïve T cells are generally considered as a circulating, homogeneous and quiescent population of T cells. However, recent advancements show that the naïve T cell pool is much more heterogeneous and compartmentalized than thought before. This compartmentalization suggests selective homing of naïve T cells to specific tissues in the body, through the expression of homing markers. Homing of naïve T cells is thought to take place especially early in life. Until this day, the drivers of early-life homing and homing marker expression of naïve T cells are not yet fully understood. In this report, we investigate the immunological context of early life naïve T cell homing. For this purpose, we studied naïve T cell homing throughout early life immune development as well as in response to immunological triggers. First, we analyzed flow cytometry data of the immune system of preterm infants in the first 42 days postpartum. We show that preterm neonates have a high percentage of naïve T cells and that these T cells predominantly express gut homing markers integrin α4β7 and CCR9. Interestingly, the neonatal naïve T cell phenotype does not consistently change within the first 42 days postpartum. Second, we looked at the effect of Toll-Like Receptor (TLR) ligand exposure in different stimulatory settings. TCR and TLR co-stimulation seems to enhance naïve T cell proliferation, along with upregulation of homing markers. Altogether, we conclude that neonatal naive T cells indeed express homing markers and that this is influenced by immune development as well as TCR and TLR stimulation.

Plain language summary

Specialized immune cells called T cells play a key role in almost all immune responses. These cells can directly attack infected cells, but they can also help other immune cells become more effective at their jobs. Before a T cell can do any of this, it needs to be activated by other immune cells. They need to give the T cell three important pieces of information: What is going on? Where in the body is this going on? And how bad is the damage? Based on this information, the T cell will decide the best tactic to fight the infection.

In order to travel to the correct organ as quickly as possible, the activated T cells will start to express so-called homing markers. These are labels presented on the outside of the T cell and they show where in the body the cell has to go. For example, some markers help cells migrate to the gut, while other markers help cells travel to the skin. First, scientists thought that only activated T cells were able to travel to specific organs. They thought that unactivated T cells, called naïve T cells, could only circulate through our body, waiting to be activated. However, there is an increasing amount of evidence that in some cases, also naïve T cells can travel to specific organs. Especially young T cells seem to be able to label themselves with homing markers to travel to specific tissues all over the body. In this study, we wanted to find out in which situations young naïve T cells are able to migrate to specific organs.

First, we studied the T cells from newborn babies in their first weeks after birth. We saw that naïve T cells mostly express homing markers to go to the gut. Second, we aimed to test which immunological signals can change the homing marker expression of naïve T cells. We studied this by stimulating naïve T cells with immunological signals called TLR ligands, which are small parts derived from microbes. After stimulation with TLR ligands, we measured how many of the cells had homing markers on their surface. We show that TLR ligands can encourage cells to divide more and that the more a cell has divided, the more likely it is to present homing markers. As a result of this, we conclude that immunological signals can increase the number of naïve T cells with homing markers on their surface.

Altogether, we found more evidence that not only activated, but also naïve T cells can migrate to specific organs in the body. In addition, we saw that migration of naïve T cells is more likely to take place in newborns than in adults and that this can also be promoted by the presence of TLR ligands. This knowledge can help us to better understand the importance of naïve T cell homing in newborn babies and reminds us of the differences between the immune system of infants and adults.

Abbreviations

CBMC	Cord Blood Mononuclear Cell
CTV	Cell Trace Violet
FBS	Fetal Bovine Serum
Lisi	Local Inverse Simpson's Index
PBMCs	Peripheral Blood Mononuclear Cell
PTK7	Protein-Tyrosine Kinase 7
scRNAseq	Single-cell RNA sequencing
T_{CM}	Central Memory T cell
T_{EM}	Effector Memory T cell
T_{EMRA}	Effector Memory cell Re-expressing CD45RA
TLR	Toll-Like Receptor
T_N	Naïve T cell
TREC	T cell receptor excision circle
UMI	Unique Molecular Identifier

Introduction

Naïve T cells stand at the basis of all adaptive immune responses. Depending on the encountered pathogen and immunological context, they can steer towards a wide variety of immune responses. The general view is that naïve T cells are a circulating, homogeneous and quiescent population of T cells waiting for antigen encounter. However, accumulating evidence suggests that the naïve T cell pool is not nearly as homogeneous and quiescent as was believed (1). Instead, there seem to be distinct subpopulations of naïve T cells with unique phenotypical characteristics. Until this day, it is unknown what triggers are involved in the regulation of this heterogeneity and how these phenotypical characteristics relate to the functionality of naïve T cell subpopulations.

There are several phenotypical characteristics that are associated with the developmental stages of naïve T cells. For CD4+ naïve T cells, a commonly used distinction is made based on the surface expression of the platelet endothelial cell adhesion molecule CD31. As CD4+ naïve T cells seem to lose their CD31 surface expression over time, this suggests that we can classify the most recent thymic emigrants as CD31+ cells and the older naïve T cells as CD31- cells (2). The proportion of CD31+ T cells also decreases with advance in age, so the percentage of CD31+ T cells is high in neonates but declines with ageing. Within the youngest CD31+ population even further subcategorization has been proposed. Previous studies have shown that the most recent thymic emigrants express protein-tyrosine kinase 7 (PTK7) and produce IL-8 (3,4). Since CD31 is constantly expressed on almost all CD8+ naïve T cells, it cannot be used to subcategorize CD8+ naïve T cells. Instead, CD103 has been proposed as a marker to identify the youngest CD8+ naïve T cells (5). The search for more specific signatures within in the naïve T cell pool is still developing, so it is likely that there is even more naïve T cell diversity that has not yet been discovered.

In addition to heterogeneity during naïve T cell development there is evidence for tissue compartmentalization of naïve T cells. Even though the majority of naïve T cells can be found in peripheral blood, some studies have shown that naïve T cells can also be found in different nonlymphoid tissues, like the skin and the gut (6,7). The compartmentalization of the naïve T cell pool is especially seen early in life, after which the number of naïve T cells residing in tissues decreases with ageing. Especially migration of neonatal naïve T cells to the gut is thought to be important for the priming of naïve T cells and in situ tolerization (8). Indeed, development of the neonatal immune system was hampered in newborns with gut bacterial dysbiosis (9). There are also studies that propose that naïve T cells exhibit a specific phenotype based on the tissue they are located in. For example, a study on fetal naïve T cells has shown that most fetal naïve T cells in the skin are CD4+ and express markers characteristic for hematopoietic stem cells (CD34 and CD38) (10). In addition, a murine study suggests that CD8+ naïve T cells in the gut seem to adapt a similar surface phenotype as intraepithelial lymphocytes (7). Still, naïve T cells seem to keep some of their general characteristics, as CD31 expression does not differ between naïve T cells found in the gut or in the circulation (6). Naive T cell compartmentalization also suggests a role of selective homing of naïve T cell subsets to different tissues in the body. Indeed, multiple studies have reported the expression of homing markers on neonatal naïve T cells (6,7). Neonatal naïve T cell compartmentalization as described above promotes heterogeneity in the naïve T cell pool, especially early in life.

Apart from the phenotypic heterogeneity, there is thought to be functional heterogeneity in the neonatal naïve T cell pool. Particularly noteworthy is the bias of recent thymic emigrants towards innate immune signaling (11). For example, recent thymic emigrants seem to express innate receptors, like complement receptors and Toll-Like Receptors (TLRs) (12,13). Especially TLR expression on naïve T cells has been studied. TLRs are pattern-recognition receptors which are well-known for their role in innate immunity. Most TLRs are expressed on the cell surface, where they can

recognize distinct pathogenic patterns (14). For example, TLR1/2 binds to triacyl lipopeptides, TLR4 to LPS and TLR5 to flagellin. In contrast, intracellular TLRs like TLR3 and TLR7 can recognize double-stranded and single-stranded RNA respectively. Several studies indicate that T cells express TLRs especially after activation (15).

It is still largely unknown whether and how these receptors might play a role in establishing naïve T cell heterogeneity and functionality. Some studies suggest that TLR stimulation promotes the production of the innate-associated cytokine IL-8 by neonatal naïve T cells (4). TLR stimulation also seems promote naïve T cell maturation in general, as *in vitro* TLR-stimulated naïve cord blood cells showed enhanced proliferation and increased expression of maturation markers CD25 and CD45RO (16,17). Additionally, TLR stimulation resulted in an increased gut homing phenotype in term neonatal T cells, but not preterm neonatal T cells (17). Usually, naïve T cell differentiation and homing marker expression is induced by TCR stimulation and additional co-stimulation provided by dendritic cells. It is still to be elucidated how TLR stimulation interacts in this process, for example as a priming stimulus prior to stimulation or as co-stimulation.

Until this day, the drivers of naïve T cells heterogeneity and differential homing marker expression of neonatal naïve T cells are not yet fully understood. In this report, we investigate the immunological drivers of early-life homing of human naïve T cells. For this purpose, we studied naïve T cell homing throughout early-life immune development as well as in the response to TLR stimulation. We investigate the naïve T cell phenotype of preterm neonates in the first 42 days postpartum. In addition, we study the effect of *in vitro* TCR and TLR stimulation on the expression of homing markers on cord blood derived CD4+ naïve T cells.

Ultimately, we find that neonatal naïve T cells have a more extensive gut-homing phenotype than adult naïve T cells throughout immune development. In addition, exposure to TLR ligands seems to enhance cord blood derived CD4+ naïve T cell proliferation, along with upregulation of homing markers.

Methods part 1. Characterizing the T cell phenotype of preterm neonates

Sample preparation

Blood was collected from preterm neonates at 7, 21 and 42 days postpartum as part of the Nutribrain study of University Medical Centre Utrecht (n=25). The cells were incubated for 15 minutes at room temperature with BD Pharm Lyse™ diluted 1:10 in aquadest to lyse the red blood cells before staining.

Flow cytometry

The isolated immune cells were stained with two separate panels for the identification of different T cell subsets and the expression of homing markers. For the T cell subset panel, all cells were stained for 20 minutes at 4 °C in Brilliant Staining Buffer (BD) with CD25-FITC, CD4-PerCp-Cy5.5, CCR7-APC, CD3-AF700, CD127-BV421, CD27-BV510, CD31-BV605, CD45RO-BV711, TCRgd-PE, CD21-PE-CF594 and CD8-PE-Cy7. For the T cell homing panel, cells were stained for 20 minutes at 4 °C in Brilliant Staining Buffer (BD) with CD103-FITC, CD4-PerCp-Cy5.5, CCR9-APC, integrin β1-AF700, CD8a-APC-eF780, integrin β7-BV421, CD3-BV605, CD45RO-BV711, GPR15-PE and integrin α4-PE-Cy7. After washing in Dulbecco's Phospate Buffered Saline (PBS), the cells were fixed overnight at 4 °C using 2% paraformaldehyde. The next day, the cells were washed again and resuspended in FACS buffer (PBS containing 2% Fetal Bovine Serum (FBS) and 0.1% NaN_3) and the samples were measured on a BD LSRFortessa.

Data analysis

FlowJo software was used to analyze the data and compare it to adult reference data. For the T cell subset panel, T lymphocytes were characterized as CD3+ lymphocytes. We used the expression of CD45RO and CCR7 to distinguish between naïve (T_N ; CD45RO-, CCR7+), effector memory (T_{EM} ; CD45RO+, CCR7-), central memory (T_{CM} ; CD45RO+, CCR7+) and effector memory cells re-expressing CD45RA (T_{EMRA} ; CD45RO-, CCR7-). Additionally, true naïve CD4+ T cells were gated as CCR7+CD27+CD45RO- (figure S1.1). For the T cell homing panel, we gated for CD4+ and CD8+ CD3+ lymphocytes. For both CD4+ and CD8+ populations we used the expression of CD45RO to distinguish between naïve (CD45RO-) and memory (CD45RO+) T cells. We gated the double-positive population for a4b1 and a4b7 and the positive populations for CCR9, CD103, CLA and GPR15 (figure S1.2). CD103 and GPR15 were left out in all further analyses. Visualization of the data was performed in GraphPad Prism 9.3.0. We used a one-way ANOVA test for normally distributed samples and a Kruskal-Wallis test for non-normally distributed samples.

Methods part 2. Homing marker expression in naïve CBMCs and PBMCs in response to *in vitro* TCR and TLR stimulation

Sample preparation

Human cord blood was obtained from neonates in the Wilhelmina Kinderziekenhuis Utrecht (n=2). Peripheral blood was collected from healthy donors through the Minidonordienst of University Medical Centre Utrecht (n=2, age range = 27-32). Blood was obtained via venipuncture and collected in 9 mL natrium heparin tubes, after which CBMCs and PBMCs were isolated using Ficoll gradient centrifugation. Subsequently, naïve T cells were sorted from the isolated CBMCs and PBMCs using a naïve CD4 T cell isolation kit and magnetic sorting on an autoMACS (Miltenyi Biotec).

Cell stimulation

Freshly isolated naïve CD4 T cells were stimulated on the same day as isolation. Prior to stimulation, the cells were stained with Cell Trace Violet (CTV) to investigate proliferation in response to the cell stimulation. After this, PBMCs and CBMCs were stimulated according to one of the following two different stimulatory conditions:

Stimulatory condition 1: Sequential TLR and CD3/CD28 stimulation

Isolated CBMCs and PBMCs were stimulated overnight at 37 °C in culture medium (RPMI 1640 medium containing 10% FBS, 2% L-glutamine and 2% penicillin-streptomycin) with one of the following TLR ligands separately: Pam3Cys4 (TLR1/2 ligand, 1 µg/mL), poly I:C (TLR3 ligand, 50 µg/mL), flagellin (TLR5 ligand, 100 ng/mL) and imiquimod (TLR7 ligand, 5 µg/mL) . Subsequently, the TLR ligands were removed and the cells were resuspended in culture medium with IL-2 (20U/mL) in order to enhance survival. The cells were then stimulated for 24 hours with anti-CD3 (1 mg/mL, eBioscience™) and anti-CD28 (1 mg/mL, eBioscience™) coated on a 96-well MaxiSorp plate in a 1:1000 dilution with PBS. Ultimately, the cells were removed from the coated plate and transferred to a regular round bottom plate and left to incubate at 37 °C for another 24 hours in culture medium with IL-2 (20U/mL).

Stimulatory condition 2: TLR and CD3/CD28 co-stimulation

Isolated CMBCs and PBMCs were stimulated for 48 hours at 37 °C with one of the following TLR ligands: Pam3Cys4 (TLR1/2 ligand, 1 µg/mL), poly I:C (TLR3 ligand, 50 µg/mL), flagellin (TLR5 ligand, 100 ng/mL) and imiquimod (TLR7 ligand, 5 µg/mL) simultaneously with anti-CD3 and anti-CD28 in the concentrations described above. Ultimately, the cells were removed from the coated plate and transferred to a regular round bottom plate and left to incubate at 37 °C for another 48 hours in culture medium with IL-2 (20U/mL).

Flow cytometry

After the stimulation procedure, all cells were stained with immunofluorescent antibodies for flow cytometry as read-out of the cells' phenotypes. First, the cells were stained for 20 minutes at 4 °C with eBioscience™ Fixable Viability Dye eFluor™ 506 (1:1000 diluted in PBS) to be able to discriminate between living cells and dead cells. Then, the cells were stained for 20 minutes at 4 °C in FACS buffer with 8% Brilliant Staining Buffer (BD), 2% FcR-blocking reagent (Miltenyi Biotec) and the following antibodies: integrin β7-PerCp-Cy5.5, integrin β1-AF700, CD27-APC-eF780, CD31-BV605, CD45RA-BV711, CD8-PE, CD45RO-PE-CF594 and integrin α4-PE-Cy7. Cells were incubated with the eBioscience Foxp3/Transcription Factor Staining Buffer Set for 30 minutes at 4 °C for fixation and permeabilization. After this, the cells were stained intracellularly for 30 minutes at 4 °C in permeabilization buffer (Invitrogen) with 2% FcR-blocking reagent, CD3-FITC and CD4-BV785. CD3 and CD4 were stained intracellularly because they have shown to become internalized after cells get stimulated with aCD3/CD28. Samples were measured on a BD LSRFortessa.

Data analysis

FlowJo software was used to analyze the data. T lymphocytes were characterized as CD3+ lymphocytes. We defined naïve T cells as CD45RA+CD27+ and subcategorized these as CD31+ or CD31-. Additionally, we determined homing marker expression and CTV staining (figure S1.3). Visualization of the data was performed in GraphPad Prism 9.3.0.

Results part 1. Characterizing the T cell phenotype of preterm neonates

Naïve T cells predominate in preterm neonates throughout the first 42 days postpartum

To investigate immune development in preterm neonates we first determined the average percentage of naïve T cells in proportion to the percentages of T_{EM}, T_{CM} and T_{EMRA} in the first 42 days postpartum. By far the most CD4+ and CD8+ T cells are naïve in preterm neonates (80-85%), without significant correlation with gestational age (n.s. 24-26 vs 27-29 weeks) (figure 1.1 A). However, there is more variance in the 27-29 weeks cohort. Preterm neonates also have a more naïve T cell phenotype (80-85%) than healthy adults (45-55%). The average CD4+ and CD8+ T cell composition, with a strong predominance of naïve T cells, shows no consistent change in the first 42 days postpartum (n.s. 7 vs 21 vs 42 days postpartum), despite individual variation (figure 1.1B and C) There is also no consistent change in CD31 expression over time (n.s. 7 vs 21 vs 42 days postpartum). The percentage of CD31+ naïve T cells diverges slightly over time, but the mean prevalence does not consistently change (figure 1.1 D).

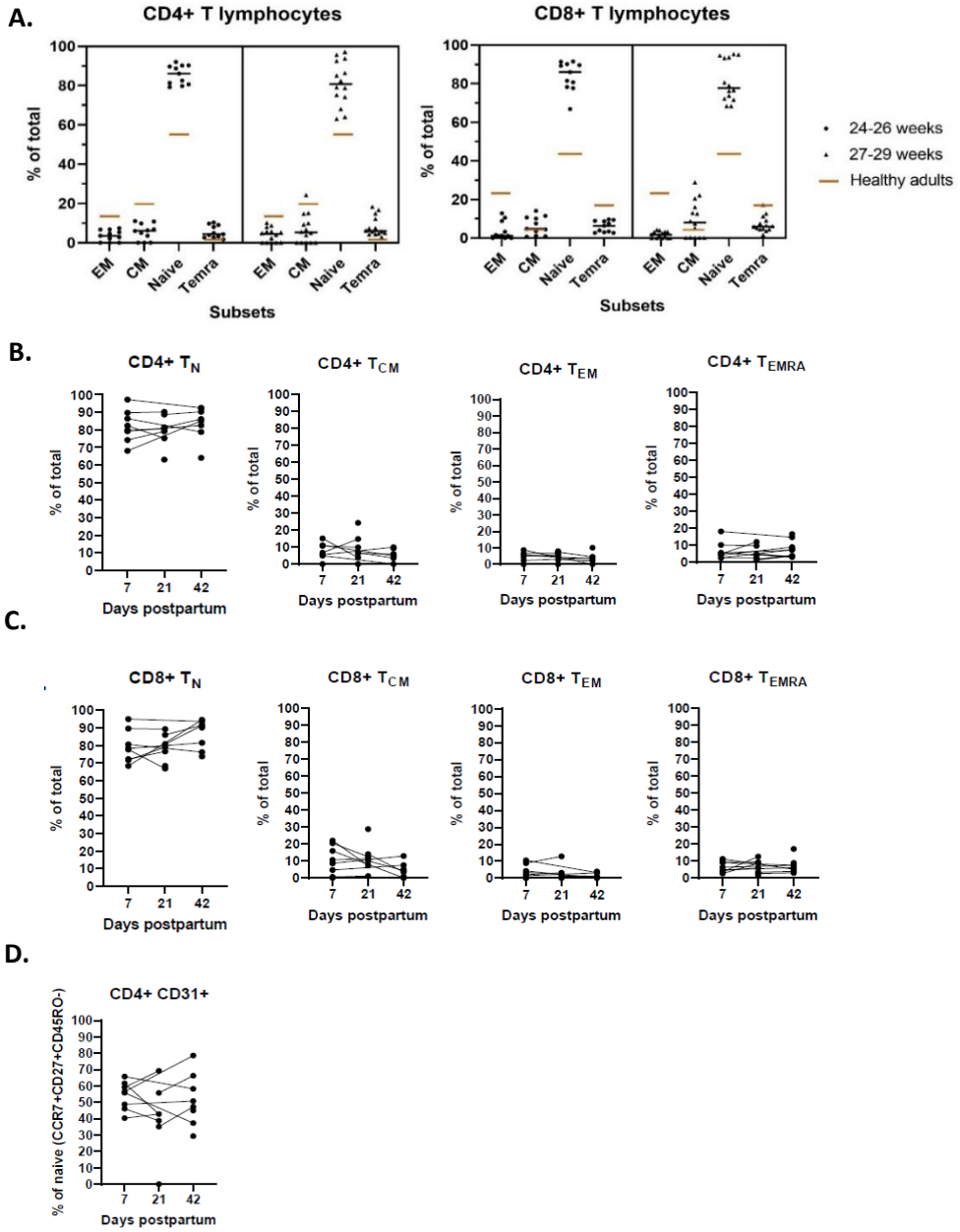


Figure 1.1: The relative abundance of CD4+ and CD8+ T cell subsets of preterm neonates. A) The percentage of naïve (T_N), effector memory (T_{EM}), central memory (T_{CM}) and effector memory cells re-expressing CD45RA (T_{EMRA}) in CD4+ T lymphocytes and CD8+ T lymphocytes compared to reference adult data (brown line). **B+C)** Individual CD4+ and CD8+ T cell ratio over time. **D)** The percentage of CD31+ T cells in the true naïve T cell population over time. Lines connect the samples of one individual at different timepoints.

Preterm neonates have a more gut-oriented homing phenotype compared to healthy adults

We observe that in naïve T cells from preterm neonates, expression of general homing markers is low, whereas around 50% of CD4+ and 90% of CD8+ naïve T cells expresses the gut homing marker integrin α4β7 (figure 1.2). CCR9 is only in expressed in CD8+ naïve T cells, of which around 85% expresses CCR9. Reference data of adult naïve T cells report a much lower expression of gut homing markers, as 10-30% of naïve T cells expresses integrin α4β7 and less than 10% of naïve T cells expresses CCR9. In contrast, the expression of integrin α4β1 and CLA in adult reference data is comparable to the expression in preterm neonates. In memory T cells, expression of the general homing markers integrin α4β1 and CLA is higher, with an expression of around 50% and 20% respectively. Around 35% of CD4+ and 65% of CD8+ memory T cells expresses integrin α4β7, which is remarkably lower than in naïve T cells. CCR9 expression is low in CD4+ memory cells, but above 90% in CD8+ memory cells. Compared to adult reference data, gut homing markers integrin α4β7 and CCR9 are expressed significantly more in neonatal memory T cells, whereas the expression pattern of integrin α4β1 and CLA is comparable in neonatal and adult memory T cells. Additionally, we observe that the homing marker expression in all neonatal subsets changes slightly, but not significantly within the first 42 days postpartum and that any variance observed is independent of gestational cohort.

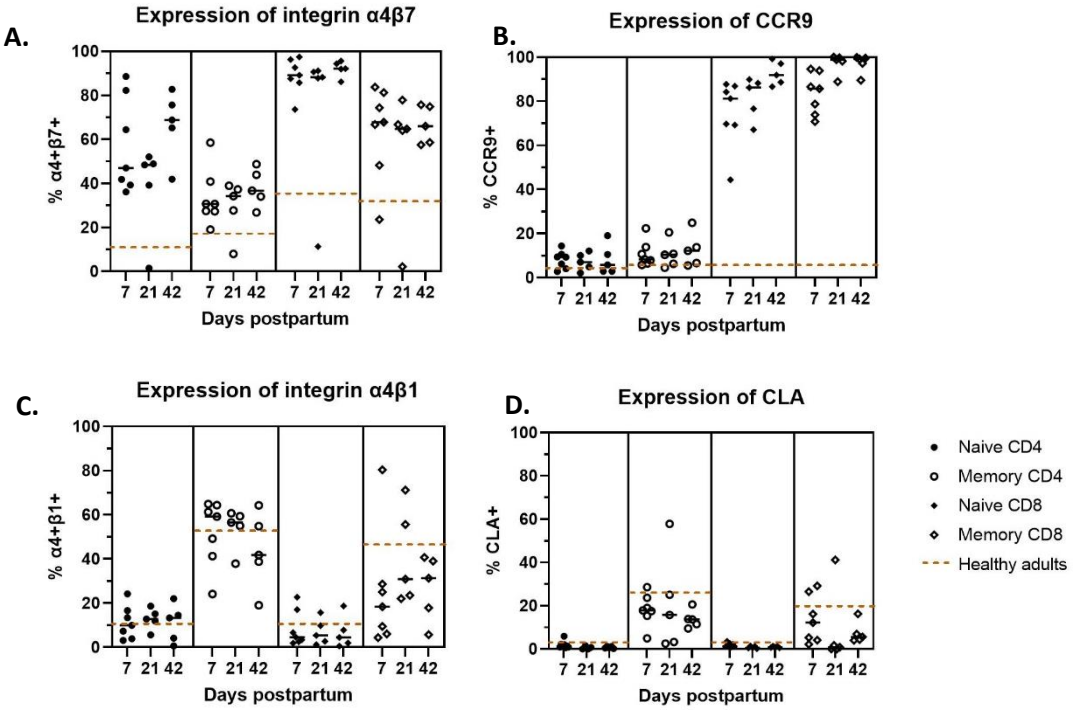


Figure 1.2: Expression of homing markers in preterm neonates in the first 42 days postpartum. The expression of integrin α4β7 (A), CCR9 (B), integrin α4β1 and CLA (D) for naïve CD4+, memory CD4+, naïve CD8+ and memory CD8+ T cells. Horizontal lines indicate the median expression per subset for neonates (black) and adults (brown - dotted).

Results part 2. Homing marker expression in naïve CBMCs and PBMCs in response to *in vitro* TCR and TLR stimulation

Cord blood naïve CD4+ T cells have increased integrin expression upon co-stimulation with aCD3/aCD28 and TLR ligands

In naïve CD4+ CBMCs, we observe that 48 hours of *in vitro* aCD3/aCD28 stimulation alone already increases the expression of both integrin α4β1 and integrin α4β7 (figure 2.1). We observe that simultaneous addition of TLR1/2 ligands (Pam3Cys4) and, to a lesser extent, TLR7 ligands (imiquimod) to aCD3/aCD28 further enhances the expression of integrin α4β7, but not integrin α4β1. The effect of stimulation is the same for CD31+ and CD31- cells, although the expression of integrin α4β1 is slightly higher in CD31- CBMCs in all conditions, including the unstimulated condition. The effect of TCR and TLR co-stimulation of naïve CD4+ CBMCs is remarkably different in naïve CD4+ PBMCs, which do not change their integrin expression in response to either aCD3/CD28 stimulation alone nor in combination with TLR ligands (figure 2.2 A+B). However, expression of CCR9 and CLA on PBMCs is altered in response to TCR and TLR co-stimulation. Combined aCD3/CD28 and TLR1/2 stimulation on PBCMs impairs CCR9 expression, while TLR4 (LPS) and TLR5 (flagellin) stimulation enhance CCR9 expression on PBMCs (figure 2.2 C). In addition, CLA expression on PBMCs seems to be upregulated in CD31+ cells after aCD3/CD28 and TLR1/2 co-stimulation (figure 2.2 D).

We also investigated whether pre-treatment of naïve CD4+ CBMCs with TLR ligands affects integrin expression after subsequent aCD3/aCD28 stimulation. We observe that 24 hours of *in vitro* aCD3/aCD28 stimulation of naïve CD4+ CBMCs alone increases the expression of integrin α4β1, but not integrin α4β7 (figure 2.3). In addition, the modulating effect of TLR1/2 and TLR7 on integrin α4β7 expression is lost when they are administered as pre-stimulation instead of co-stimulation, whereas pre-treatment with TLR ligands does still decrease integrin α4β1 expression. We also see that, just like with co-stimulation, integrin α4β1 expression is somewhat higher in CD31- CBMCs than CD31+ CBMCS. None of the stimulatory conditions results in integrin upregulation in PBMCs. This suggests that especially CBMCs are potent to upregulate the expression of both integrin α4β1 and α4b7 in response to TCR stimulation alone. In addition, TLR stimulation can enhance integrin α4β7 upregulation, but only as co-stimulation with TCR stimulation and not pre-stimulation.

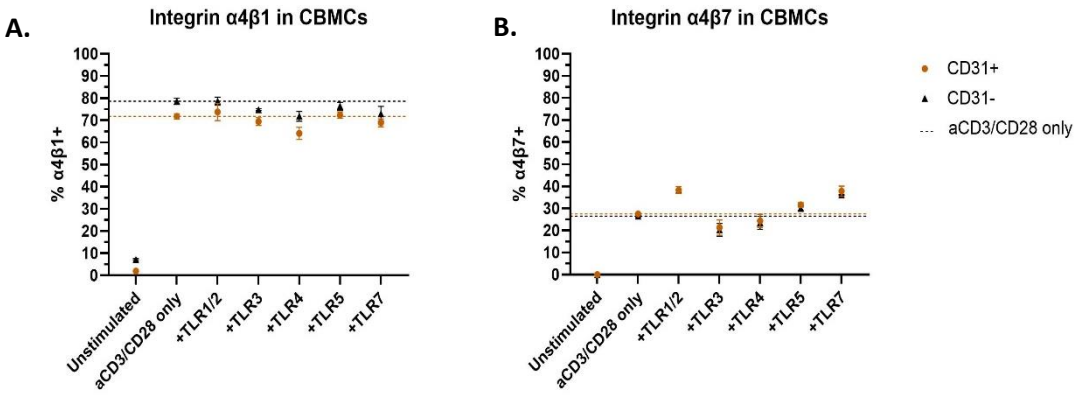


Figure 2.1: Expression of integrin α4β1 and α4β7 on naïve CD4+ T cells after aCD3/CD28 and TLR ligand co-stimulation. Mean percentage of cord blood naïve CD4+ T cells that express integrin α4β1 (A) or integrin α4β7 (B) after co-stimulation (n=1). Dotted lines indicate the mean percentage after aCD3/CD28 stimulation alone. Error bars indicate the standard-deviation of the duplicate measurements for each condition.

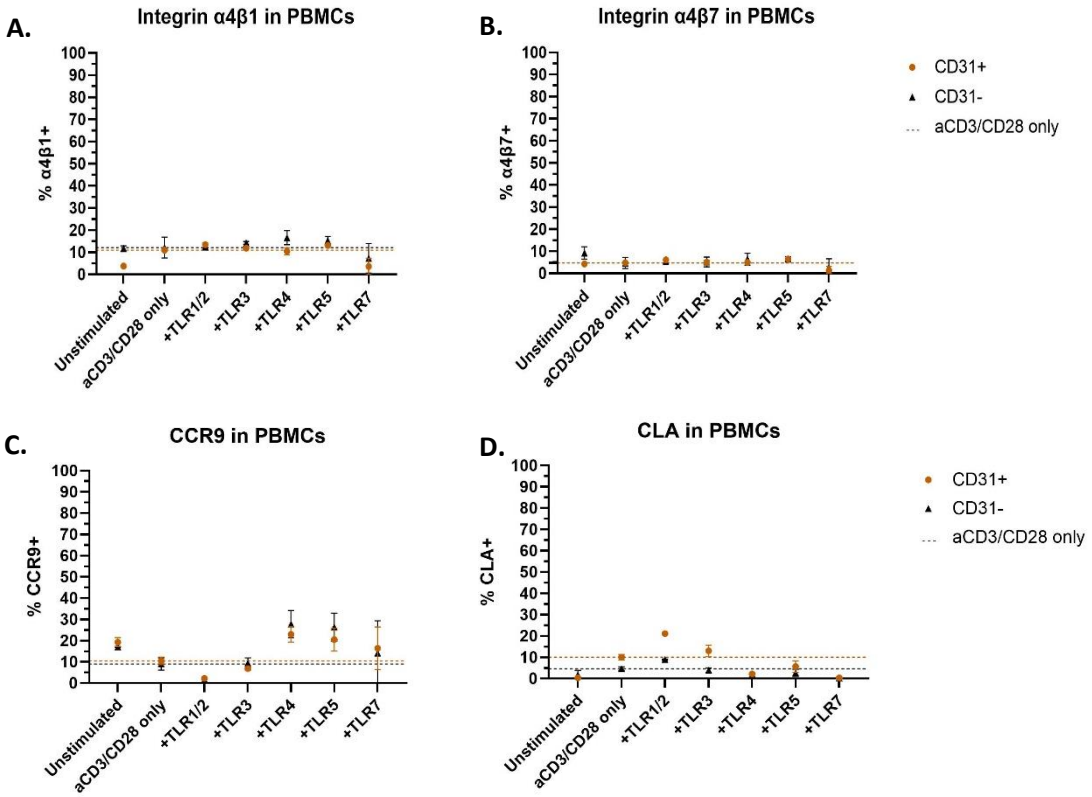


Figure 2.2: Expression of integrin α4β1, integrin α4β7, CCR9 and CLA on naïve CD4+ T cells after aCD3/CD28 and TLR ligand co-stimulation. Mean percentage of healthy adult peripheral blood naïve CD4+ T cells that express integrin α4β1 (A), integrin α4β7 (B), CCR9 (C) or CLA (D) after co-stimulation (n=1). Dotted lines indicate the mean percentage after aCD3/CD28 stimulation alone. Error bars indicate the standard-deviation of the duplicate measurements for each condition.

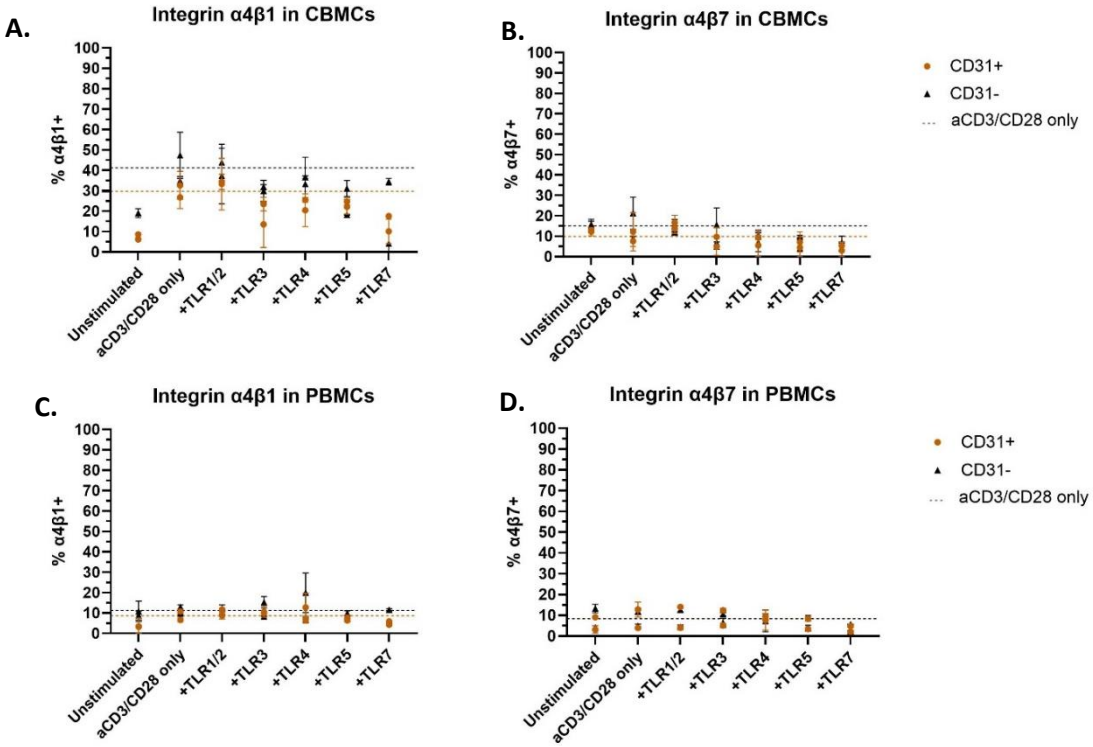


Figure 2.3: Expression of integrin α4β1 and α4β7 on naïve CD4+ T cells after TLR ligand pre-stimulation and subsequent aCD3/CD28 stimulation. Mean percentage of cord blood naïve CD4+ T cells that express integrin α4β1 (A) or integrin α4β7 (B) after stimulation and mean percentage of healthy adult peripheral blood naïve CD4+ T cells that express integrin α4β1 (C) or integrin α4β7 (D) after stimulation (n=2). Dotted lines indicate the mean percentage after aCD3/CD28 stimulation alone. Error bars are standard-deviation.

Increasing integrin expression in CD4+ naïve T cells correlates with proliferation

An important effect of aCD3/aCD28 stimulation is inducing proliferation. When comparing the expression of integrin α4β1 and α4b7 between CBMCs with a different proliferation history, we notice that expression of both integrin α4β1 and α4β7 increases the more cells proliferate, independent of the TLR ligand added (figure 2.4 A + B). Interestingly, the kinetics of integrin α4β1 upregulation differ from α4b7 upregulation. Integrin α4β1 is already highly expressed in cells that divided only once (CTV1), whereas the expression of integrin α4β7 only increases after multiple divisions. Also in PBMCs, we observe that rapid upregulation of integrin α4β1 and more gradual upregulation of integrin α4β7, independent of the TLR ligand exposure (figure 2.4 C+D). However, the maximum integrin expression reached is lower than in CBMCs. In contrast to integrin α4β1 and α4β7, the expression of CCR9 and CLA in PBCMs is TLR ligand-dependent, as TLR1/2 stimulation decreases the expression of CCR9 but increases the expression of CLA per division state, whereas TLR4 stimulation increases the expression of CCR9 and decreases the expression of CLA per division state (figure 2.4 E+F).

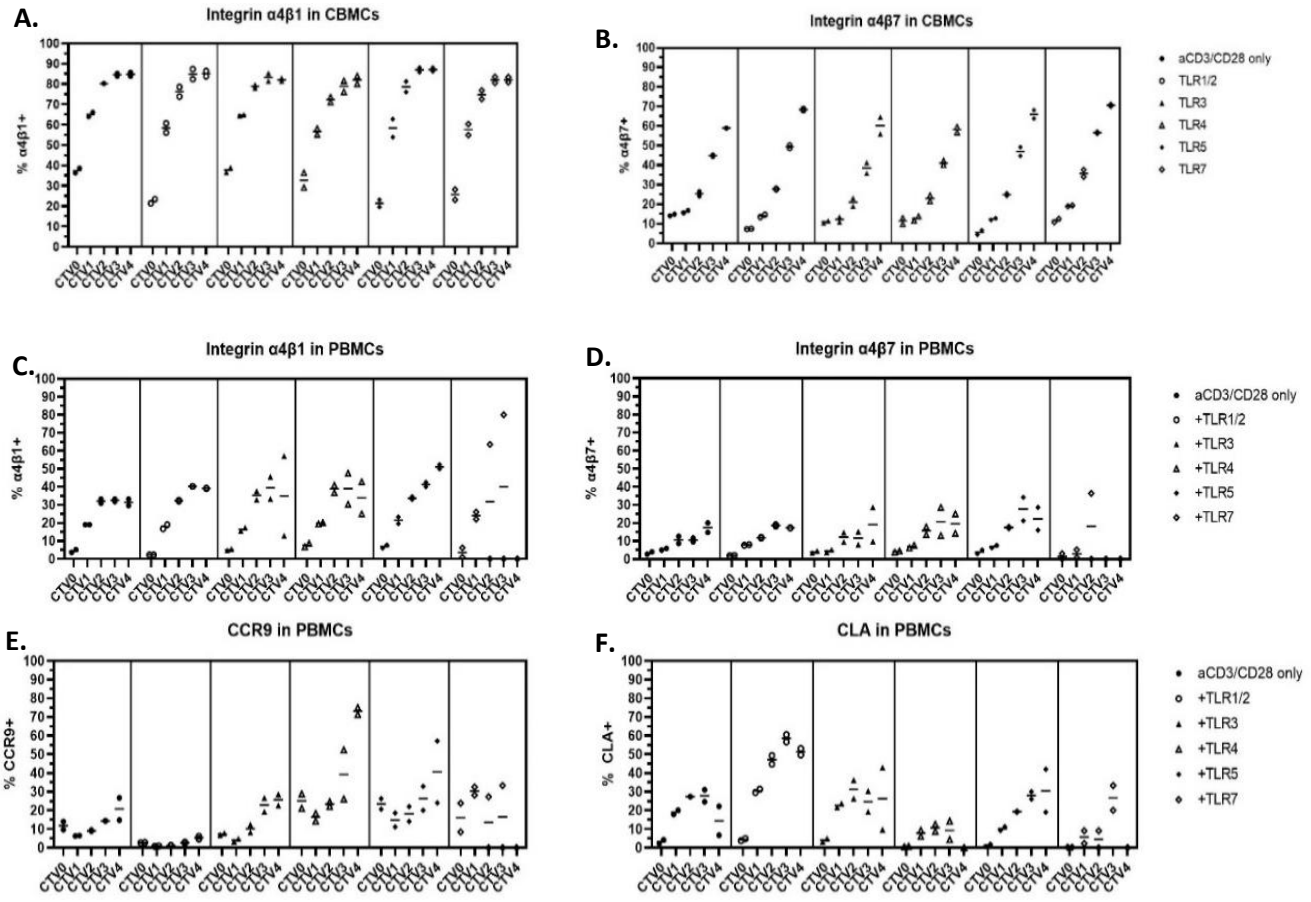


Figure 2.4: Homing receptor expression and proliferation after TCR and TLR co-stimulation
The percentage of cells that express integrin α4β1 in CBMCs (A), integrin α4β7 in CBMCs (B), integrin α4β1 in PBMCs (C), integrin α4β7 in PBMCs (D), CCR9 in PBMCs (E) or CLA in PBMCs (F) per number of proliferative rounds a cell has undergone (CTV0-4). The expression is determined after co-stimulation with aCD3/CD28 and additional stimulation of TLRs (TLR1/2, TLR3, TLR4, TLR5 and TLR7).

Exposure to TLR ligands stimulates TCR-driven proliferation in naïve CD4+ T cells

The data in figure 2.4 shows that integrin expression correlates almost exclusively with proliferation and not with the TLR ligand added, whereas the result from figures 2.1, 2.2 and 2.3 suggested that integrin expression was TLR ligand-dependent. We hypothesized that this TLR ligand-dependent

increase in integrin expression might be the result of increased proliferation in these samples. Indeed, proliferation of CBMCs significantly increases in response to TLR1/2 and, to a lesser extent, TLR5 and TLR7 co-stimulation with aCD3/CD28 for 48 hours (figure 2.5). This stimulatory effect is slightly higher in CD31- CBMCs, as they have an overall higher proliferation index than CD31+ CBMCs. In PBMCs, TLR1/2 co-stimulation also induces proliferation, whereas TLR5 and TLR7 co-stimulation does not seem to promote proliferation. In fact, TLR7 co-stimulation even seems to impair proliferation of PBMCs. In contrast to CBMCs, proliferation is higher in CD31+ PBMCs than CD31- PBMCs. However, CBMCs still proliferate more than PBMCs in all stimulatory conditions.

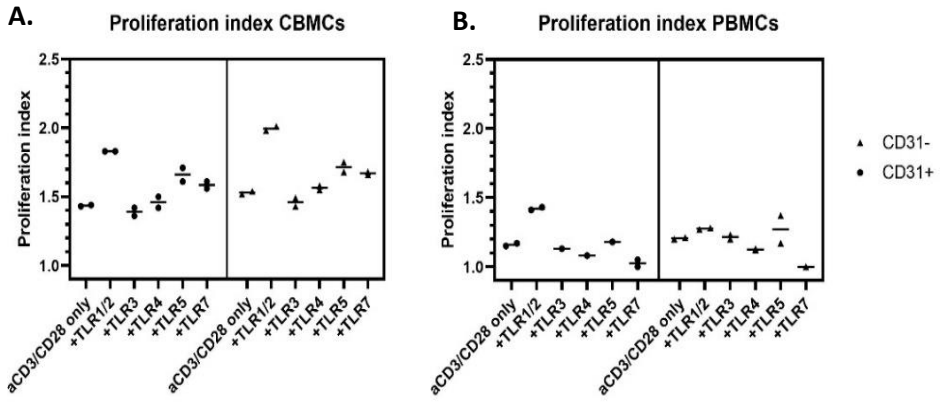


Figure 2.5: Proliferation indices of CBMCs and PBMCs after 48 hours of TCR and TLR co-stimulation. Proliferation is determined for CBMCs (A) and PBMCs (B) for all stimulatory conditions separately. The proliferation index is defined as the total number of divisions divided by the number of cells that underwent at least one division.

Discussion

In this study, we investigated the immunological drivers of early life homing of human naïve T cells. We show that the homing phenotype of naïve T cells is adaptive and depends on the immunological setting. Throughout early-life immune development, neonatal naïve T cells predominantly exhibit a gut homing phenotype, suggesting its importance in physiological immune development. We observe that skewing of the neonatal naïve T cell phenotype might also be driven by immunological triggers. We show that *in* vitro TLR stimulation promotes TCR-driven proliferation and general homing marker expression in cord blood derived CD4+ naïve T cells.

In the first 42 days postpartum of preterm neonates, we found a predominance of naïve T cells and elevated expression of the gut-homing markers integrin α4β7 and CCR9 in neonatal naïve T cells compared to adults. We found that integrin α4β7 is highly expressed in both CD4+ and CD8+ naïve T cells, which targets them to the intestinal mucosa and mesenteric lymph nodes (18). Besides its role as homing receptor, integrin α4β7 is thought to promote tolerogenic properties of the mucosal immune system, as β7 integrins were required to reconstitute tolerogenic gut mononuclear phagocytes in mice (19). CCR9 is expressed in almost all peripheral CD8+ but not CD4+ naïve T cells, as also described in previous studies on peripheral naïve T cells in mice (20). This implies specific tissue preferences of CD4+ and CD8+ naïve T cells, adding to the heterogeneity in the naïve T cell pool. On the other hand, there were no significant differences between neonates and adults in the expression of the general homing marker integrin α4β1 or the skin homing marker CLA. Even though we have no evidence that neonatal naïve T cells actually migrate to nonlymphoid tissues, there are studies that have shown that naïve T cells can also be found in different tissues, especially early in life (6,7). For example, gut-homing appears to be important in the neonatal immune system to ensure adequate immune development (8,9). Selective skin-homing of naïve T cells has also been suggested, although this has only been studied in fetal and not neonatal skin (10). These data suggest that the conventional hypothesis that naïve T cells are a homogeneous, circulating population might only be true for adults, but not so much for the neonatal immune system.

Even though our data shows that homing phenotype of neonatal naïve T cells is different from adult naïve T cells, we found that the gut-oriented T cell phenotype in preterm neonates does not consistently change in the first 42 days postpartum. This is in line with previous studies, which showed that, in contrast to B cells, dendritic cells and natural killer cells, the T cell compartment in infants maintains a distinct phenotype from adults for at least the initial 2 years of life (6,9). Although we found that the average T cell population did not change over the given time, preterm neonates showed high inter-individual variability in T cell subset composition over time, independent of gestational age at birth. Such discrepancies could be due to individual events like disease. Even though we excluded neonates with acute immunological complications from our data, it has been shown that the majority of preterm neonates suffers from immunological complications and pro-inflammatory signatures (9,21). The inter-individual variability in T cell phenotype suggests that skewing of the neonatal naïve T cell phenotype might also be driven by immunological triggers.

To study the effect of immunological triggers on the neonatal naïve T cell homing phenotype, we stimulated cord blood derived naïve CD4+ T cells with TCR and TLR stimulation. We observed that TCR stimulation alone can induce upregulation of integrins in cord blood derived but not adult naïve CD4+ T cells. This suggests that CBMCs have a lower threshold for aCD3/CD28 stimulation alone. This is in line with several studies which have shown that both CD4+ and CD8+ neonatal T cells proliferate more rapidly upon stimulation than adult-derived T cells (22–24). Interestingly, there was only a negligible difference in response between CD31+ and CD31- naïve T cells upon TCR stimulation. This suggests that the characteristics of naïve T cells correlate most with ageing of individuals and not so much the developmental stages of naïve T cells.

Previous research has stated that dendritic cells are required for T cells to respond to TLR ligand exposure (25). However, our data suggests that also MACS-sorted naïve CD4+ T cells have the potency to respond to TLR stimulation in the absence of dendritic cells. We show that especially TLR1/2 stimulation can significantly boost TCR-driven proliferation and homing marker expression of neonatal naïve T cells. This is in line with previous findings that human naïve T cells upregulate TLR2 expression after activation (13). The modulating effect of TLR ligand exposure on neonatal but not adult naïve CD4+ T cells could be attributed to the fact that neonatal naïve T cells have a higher baseline TLR expression, whereas adult naïve T cells only upregulate their TLR expression after activation (15).

Although neonatal integrin α4β1 upregulation is already promoted after 24 hours of TCR stimulation, integrin α4β7 is only upregulated after 48 hours of TCR stimulation. Notably, a previous mice study showed that integrin α4β7 upregulation upon TCR stimulation is dependent on simultaneous interaction with CD103 dendritic cells (26). This possibly explains why prolonged TCR stimulation is required in absence of CD103 dendritic cells to force sufficient activation for integrin α4β7 upregulation. Interestingly, the kinetics of integrin α4β1 upregulation differ from α4β7 upregulation. We observed that integrin α4β1 is already highly expressed upon one proliferation round, whereas the expression of integrin α4β7 only increases after multiple divisions. This difference in kinetics can be explained by the chemical properties of their common subunit α4. Since α4 has a higher affinity for β1 than β7, it will predominantly form heterodimers with β1 (27). Possibly, the faster kinetics of α4β1 compared to α4β7 upregulation upon proliferation are related to their biological function, as integrin α4β1 is a more general homing marker, whereas integrin α4β7 is a more specific homing marker.

The kinetics of integrin upregulation seemed to be independent of TLR ligand exposure. Interestingly, upregulation of CCR9 and CLA did seem to be influenced by TLR stimulation. However, direct comparison between the TLR ligands in our study is difficult, since we only took along one concentration per TLR ligand. Therefore, future studies should investigate a broader range of TLR ligand concentrations, as well as the effect of combinations of different TLR ligands. Additionally, our data could be compared to the effect of other immunological triggers on the neonatal naïve T cell phenotype, for example in response to cytokines or complement.

Altogether, this study emphasizes that neonatal naive T cells are not just a homogeneous, circulating T cell population as suggested about the adult naïve T cell population. In contrast, neonatal naïve T cells exhibit distinct homing phenotypes which depend on the immunological context. Our data indicates that neonatal T cells are especially prone to gut homing during early-life immune development. In addition, exposure to TLR ligands seems to enhance TCR-driven naïve T cell proliferation, along with upregulation of homing markers. Even though it is still largely unknown how these phenotypical qualities relate to the functionality of naïve T cells, it does underline that the neonatal immune system cannot be directly compared to the adult immune system.

References

1. Van Den Broek T, Borghans JAM, Van Wijk F. The full spectrum of human naive T cells. Vol. 18, Nature Reviews Immunology. Nature Publishing Group; 2018. p. 363–73.

2. Kimmig S, Przybylski GK, Schmidt CA, Laurisch K, Möwes B, Radbruch A, et al. Brief Definitive Report Two Subsets of Naive T Helper Cells with Distinct T Cell Receptor Excision Circle Content in Human Adult Peripheral Blood [Internet]. Vol. 195. 2002. Available from: http://www.jem.org/cgi/content/full/195/6/789

3. Haines CJ, Giffon TD, Lu LS, Lu X, Tessier-Lavigne M, Ross DT, et al. Human CD4 + T cell recent thymic emigrants are identified by protein tyrosine kinase 7 and have reduced immune function. Journal of Experimental Medicine. 2009 Feb 16;206(2):275–85.

4. Gibbons D, Fleming P, Virasami A, Michel ML, Sebire NJ, Costeloe K, et al. Interleukin-8 (CXCL8) production is a signatory T cell effector function of human newborn infants. Nat Med. 2014 Oct 1;20(10):1206–10.

5. Mcfarland RD, Douek DC, Koup RA, Picker LJ. Identification of a human recent thymic emigrant phenotype [Internet]. Vol. 97. PNAS; 2000. Available from: www.pnas.orgcgidoi10.1073pnas.070061597

6. Thome JJC, Bickham KL, Ohmura Y, Kubota M, Matsuoka N, Gordon C, et al. Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues. Nat Med. 2016 Jan 1;22(1):72–7.

7. Staton TL, Habtezion A, Winslow MM, Sato T, Love PE, Butcher EC. CD8+ recent thymic emigrants home to and efficiently repopulate the small intestine epithelium. Nat Immunol. 2006 May;7(5):482–8.

8. Lewis M, Tarlton JF, Cose S. Memory versus naive T-cell migration. Vol. 86, Immunology and Cell Biology. 2008. p. 226–31.

9. Olin A, Henckel E, Chen Y, Lakshmikanth T, Pou C, Mikes J, et al. Stereotypic Immune System Development in Newborn Children. Cell. 2018 Aug 23;174(5):1277-1292.e14.

10. Reitermaier R, Ayub T, Staller J, Kienzl P, Fortelny N, Vieyra-Garcia PA, et al. The molecular and phenotypic makeup of fetal human skin T lymphocytes. Development. 2022 Apr 15;149(8).

11. Das A, Rouault-Pierre K, Kamdar S, Gomez-Tourino I, Wood K, Donaldson I, et al. Adaptive from Innate: Human IFN-γ+CD4+ T Cells Can Arise Directly from CXCL8-Producing Recent Thymic Emigrants in Babies and Adults. The Journal of Immunology. 2017 Sep 1;199(5):1696–705.

12. Pekalski ML, García AR, Ferreira RC, Rainbow DB, Smyth DJ, Mashar M, et al. Neonatal and adult recent thymic emigrants produce IL-8 and express complement receptors CR1 and CR2. JCI Insight. 2017 Aug 17;2(16).

13. Komai-Koma M, Jones L, Ogg GS, Xu D, Liew FY. TLR2 is expressed on activated T cells as a costimulatory receptor [Internet]. 2004. Available from: www.pnas.orgcgidoi10.1073pnas.0400171101

14. El-Zayat SR, Sibaii H, Mannaa FA. Toll-like receptors activation, signaling, and targeting: an overview. Bull Natl Res Cent. 2019 Dec;43(1).

15. MacLeod H, Wetzler LM. T cell activation by TLRs: a role for TLRs in the adaptive immune response. Sci STKE. 2007;2007(402).

16. McCarron M, Reen DJ. Activated Human Neonatal CD8+ T Cells Are Subject to Immunomodulation by Direct TLR2 or TLR5 Stimulation. The Journal of Immunology. 2009 Jan 1;182(1):55–62.

17. Crespo M, Martinez DG, Cerissi A, Rivera-Reyes B, Bernstein HB, Lederman MM, et al. Neonatal T-cell maturation and homing receptor responses to Toll-like receptor ligands differ from those of adult naive T cells: Relationship to prematurity. Pediatr Res. 2012 Feb;71(2):136–43.

18. Mora JR, Von Andrian UH. Specificity and plasticity of memory lymphocyte migration. Curr Top Microbiol Immunol. 2006;308:83–116.

19. Villablanca EJ, De Calisto J, Torregrosa Paredes P, Cassani B, Nguyen DD, Gabrielsson S, et al. β7 integrins are required to give rise to intestinal mononuclear phagocytes with tolerogenic potential. Gut. 2014 Sep;63(9):1431–40.

20. Wurbel MA, Malissen B, Campbell JJ. Complex regulation of CCR9 at multiple discrete stages of T cell development. Eur J Immunol. 2006 Jan;36(1):73–81.

21. Goldenberg RL, Culhane JF, Iams JD, Romero R. Preterm Birth 1 Epidemiology and causes of preterm birth [Internet]. Vol. 371, www.thelancet.com. 2008. Available from: www.thelancet.com

22. Reynaldi A, Smith NL, Schlub TE, Venturi V, Rudd BD, Davenport MP. Modeling the dynamics of neonatal CD8+ T-cell responses. Immunol Cell Biol. 2016 Oct;94(9):838–48.

23. Mold JE, Venkatasubrahmanyam S, Burt TD, Michaëlsson J, Rivera JM, Galkina SA, et al. Fetal and Adult Hematopoietic Stem Cells Give Rise to Distinct T Cell Lineages in Humans. Science (1979). 2010 Dec 17;330(6011):1695–9.

24. Wang J, Wissink EM, Watson NB, Smith NL, Grimson A, Rudd BD. Fetal and adult progenitors give rise to unique populations of CD8+ T cells. Blood. 2016 Dec 29;128(26):3073–82.

25. Heidegger S, Kirchner SK, Stephan N, Bohn B, Suhartha N, Hotz C, et al. TLR Activation Excludes Circulating Naive CD8+ T Cells from Gut-Associated Lymphoid Organs in Mice. The Journal of Immunology. 2013 May 15;190(10):5313–20.

26. Mora JR, Bono MR, Manjunath N, Weninger W, Cavanagh LL, Rosemblatt M, et al. Selective imprinting of gut-homing T cells by Peyer's patch dendritic cells. Nature. 2003 Jul;424(6944):88–93.

27. DeNucci CC, Pagán AJ, Mitchell JS, Shimizu Y. Control of α4β7 Integrin Expression and CD4 T Cell Homing by the β1 Integrin Subunit. The Journal of Immunology. 2010 Mar 1;184(5):2458–67.

28. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun;184(13):3573-3587.e29.

Supplement 1. Supplementary figures

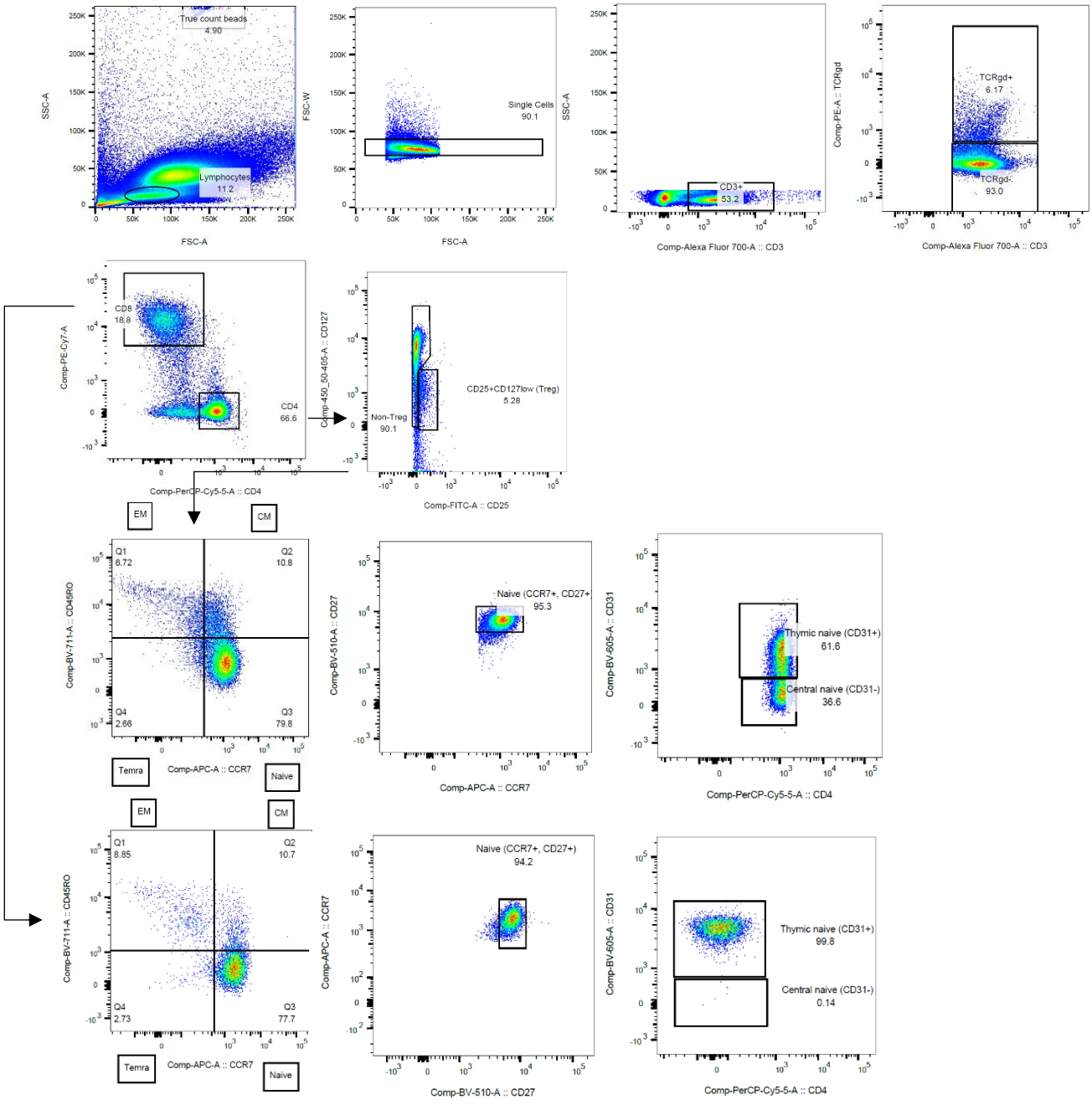


Figure S1.1: Gating strategy part 1 T cell subset panel. We gated for CD4+ and CD8+ CD3+TCRgd- lymphocytes. For the CD4+ population, we used CD25 and CD127 to select non-Tregs. For both CD4+ and CD8+ populations we used the expression of CD45RO and CCR7 to distinguish between naïve (TN ; CD45RO-, CCR7+), effector memory (T_{EM} ; CD45RO+, CCR7-), central memory (T_{CM} ; CD45RO+, CCR7+) and effector memory cells re-expressing CD45RA (T_{EMRA} ; CD45RO-, CCR7-). Additionally, true naïve CD4+ T cells were gated as CCR7+CD27+CD45RO-. Thymic naïve cells were gated as CD31+ and central naïve as CD31-.

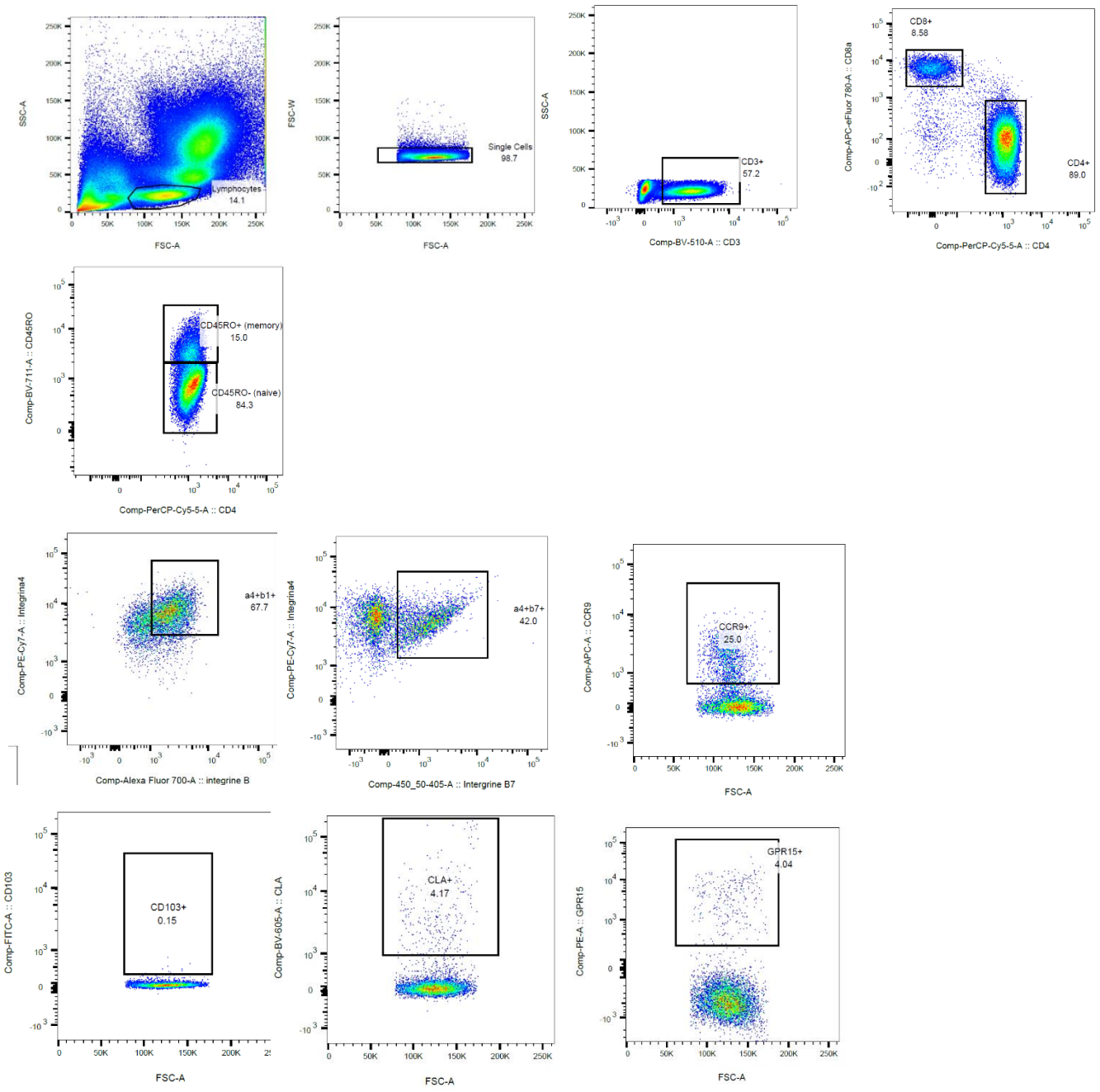


Figure S1.2: Gating strategy part 1 T cell homing panel. We gated for CD4+ and CD8+ CD3+ lymphocytes. For both CD4+ and CD8+ populations we used the expression of CD45RO to distinguish between naïve (CD45RO-) and memory (CD45RO+) T cells. We gated the double-positive population for a4b1 and a4b7 and the positive populations for CCR9, CD103, CLA and GPR15. CD103 and GPR15 were left out in all further analyses.

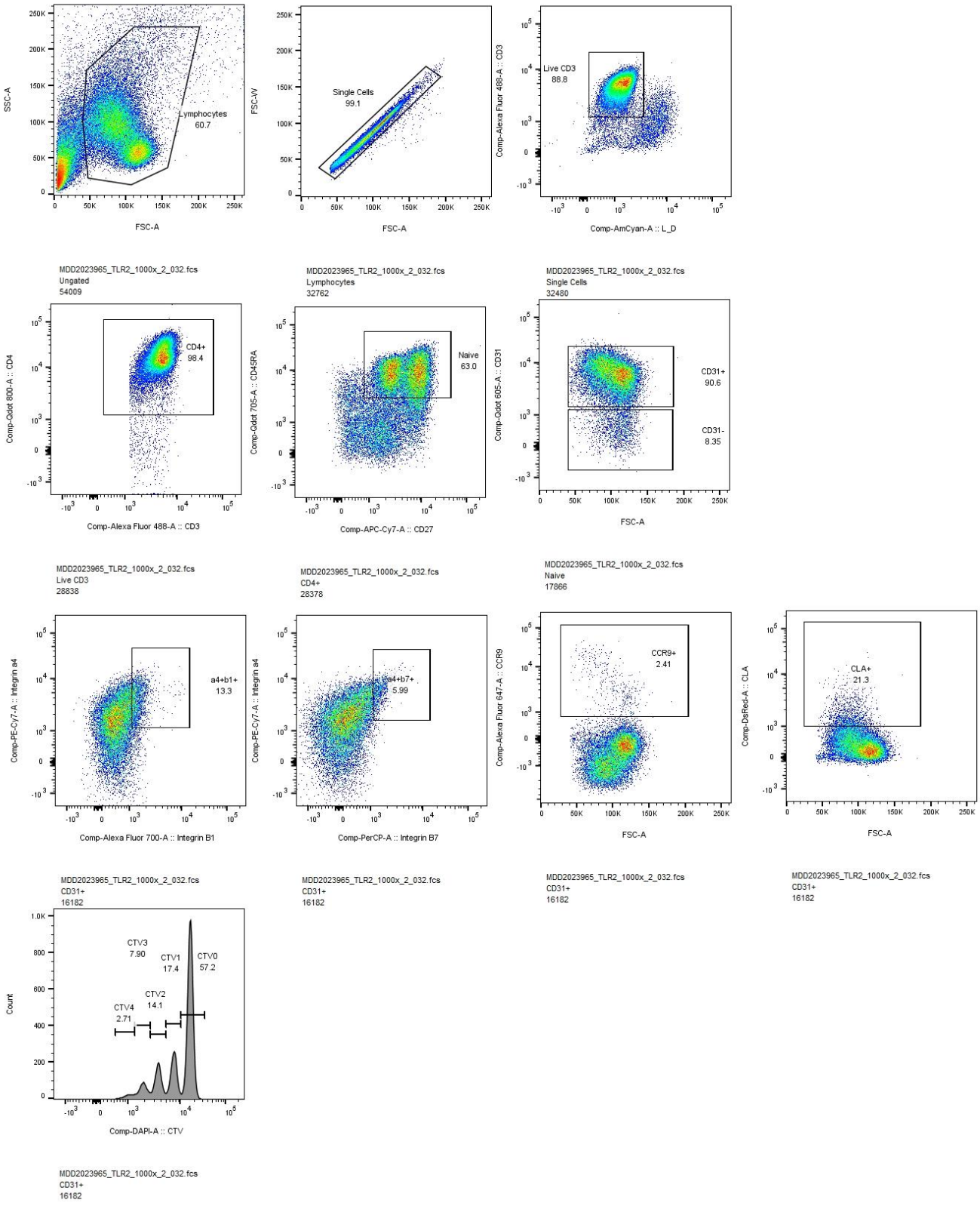


Figure S1.3: Gating strategy part 2 TLR stimulation. We gated for live CD3+CD4+ lymphocytes. We defined naïve cells as CD45RA+CD27+ and subcategorized CD31+ and CD31- cells. We gated the double-positive population for a4b1 and a4b7 and the positive populations for CCR9 and CLA. We used the CTV staining to categorize cells based on the number divisions undergone.

Supplement 2. Integration and clustering parameter adjustment for single-cell RNA sequencing

Introduction

Single-cell RNA sequencing (scRNAseq) is a popular technique to study immune cell heterogeneity at a single-cell transcriptomic level. Because this approach generates large amounts of data, subsequent bioinformatic analysis is required to process the data. In order to gain an unbiased insight in the full diversity of gene expression within and between cells, it is important to distinguish between technical variability between samples and actual biological heterogeneity. For this purpose, it is essential to first minimalize the background signal caused by technical variability, which is reflected as batch effect. A major pitfall in this analysis workflow is overcorrection of the data, which inherently decreases the detectability of biological heterogeneity. Consequently, a major aim of scRNAseq analysis is to find the optimal balance between batch effect correction while retaining biologically relevant output.

One approach for scRNAseq analysis is using the R package Seurat, developed by the Satija lab (28). In the Seurat version 4 workflow, first all individual samples are filtered for low-quality cells. These cells are filtered out based on low number of gene counts, low number of cell counts and high percentage of mitochondrial genes. Subsequently, the individual samples are normalized for the percentage of mitochondrial genes, the number of S phase and G2M phase genes and optionally the percentage of ribosomal genes (***pct.ribo***). Then, all individual samples can be merged into a single Seurat object. However, depending on your sample origin, this can result in significant batch effects which can complicate correct data interpretation. Therefore, an alternative approach offered by the Seurat package is to perform integration with batch effect correction when combining different samples.

First, the samples are screened for genes with high cell-to-cell variation in expression, called features. Depending on the size and heterogeneity of the samples that were used, the number of features to return can be adjusted **(*n.features*)**. Then, the average gene expression in all samples is compared using universally expressed anchor genes. These are genes which have a similar expression in all samples, so discrepancies in their expression can be used to detect technical background signal. Low quality anchors can be filtered out by the argument ***k.filter***. Finally, dimension reduction of the integrated dataset facilitates further analysis. In this function, the number of dimensions to be used for the reduction can be adjusted ***(dim)*.** Together, the integration parameters described above determine the degree of batch effect correction during integration. The amount of remaining batch effect can be quantified using the Local Inverse Simpson's Index (Lisi) test. This test provides information about how well-mixed the samples are. In case of a well-mixed distribution across samples, this would ideally yield in a Lisi score close to the number of samples that were integrated. After batch effect correction, the remaining differential gene expression can be determined, which is then considered as the actual biological variation in the samples. For this purpose, the integrated object can be clustered with a specific ***resolution***, which reflects the clustering depth. This is followed by differential gene expression analysis to identify top differentially expressed genes per cluster.

Although default values are available for all the parameters described above, these do not always result in the optimal balance between batch effect correction and biologically relevant output. Therefore, these parameters can be adjusted to improve the outcome of scRNAseq analysis. It is, however, still debatable how this parameter adjustment should be approached and how specific parameter settings affect the outcome of the analysis. In this study, we develop and test a bioinformatic tool to easily explore the effects of integration and clustering parameters in Seurat's scRNAseq analysis.

Methods

We performed SORT-seq of CD4+ True Naive (CD45RA+CD27+CD95-) T cells from patients with Down's Syndrome as part of the PRIDE study of University Medical Centre Utrecht. As input data for the scRNAseq analysis in Seurat, we used the raw, Unique Molecular Identifier (UMI)-corrected counts provided by Single Cell Discoveries of four separate samples (s005, s008, s009, s010). For each of the individual samples we performed quality control and normalization with and without regressing for ***pct.ribo***. The final objects were then integrated with a selection of different parameter settings for ***k.filter*** (100, 200, 300 and 400) and ***n.features*** (1000, 2000, 3000) (supplement 3). These values were selected based on the default settings for these arguments (***k.filter=200, n.features=2000***) and a range of values in the same order of magnitude. Preliminary analysis revealed that the ***dim*** argument had a negligible effect on the amount of batch effect. Therefore, we focus on the results of adjusting the arguments ***k.filter*** and ***n.features*** only. The Lisi test was performed to quantify the amount of batch effect and this was used to compare the different combinations of parameter settings. For this purpose, we used an automated for-loop to easily explore the different parameter combinations, including a standardized output format (supplement 3). Subsequently, we performed clustering on the integrated objects using values between 0.1 and 0.9 for the ***resolution*** argument. After this, the defining markers per cluster were determined using Seurat's ***FindAllMarkers*** function and MAST test. Significant genes ($p<0.05$) were selected and sorted based on their log2 fold change.

Results

Integration of samples results in less batch effect than merging samples

We compared the effect of integration and merging on batch effect with and without regression for ***pct.ribo*** separately. We found less batch effect in nearly all integrated objects compared to the merged objects (figure S2.1 and table S2.1). In addition, the samples that were not normalized for ***pct.ribo*** showed less batch effect after either merging or integrating than the samples that were normalized for ***pct.ribo***.

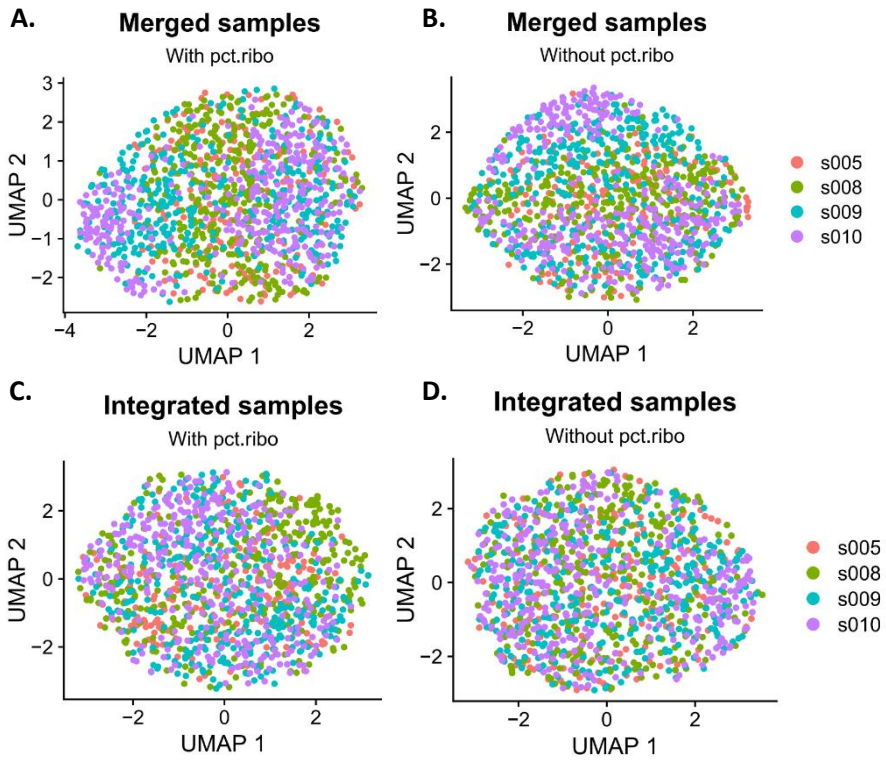


Figure S2.1 UMAPs of batch effect in merged and integrated samples. The samples were clustered based on original identity after merging (A+B) or integrating (C+D) the samples. For the integrated samples the default integration settings are shown ***(k.filter=200, n.features=2000)***

Table S2.1: Batch effect in merged and integrated samples. The amount of batch effect is quantified as the Lisi test score. This was compared between samples with and without normalization for ***pct.ribo*** after either merging or integrating the samples. The range for the integrated samples covers the range in Lisi scores obtained with the different integration parameter settings.

	Merged	Integrated
With pct.ribo	2,649	2,644-3,365
Without pct.ribo	2,750	2,782-3,399

Batch effect correction is mostly determined by *n.features*

Although sample integration results in less batch effect than merging regardless of the integration parameter settings used, there is still a broad range of remaining batch effect. We show that the amount of remaining batch effect is mostly determined by the value of ***n.features***, both for the samples with and without regression for ***pct.ribo*** (table S2.2 and S2.3). Interestingly, there seems to be no linear correlation for ***k.filter***. Lisi score decreases with ***n.features***, but not linearly.

Table S2.2 and S2.3: Batch effect after integration with different settings for *k.filter* and *n.features*. The amount of batch effect is quantified as the Lisi test score for integration of the samples with and without regression for ***pct.ribo***.

Without pct.ribo

		N.features		
		1000	2000	3000
K.filter	100	3.224450652	2.963149527	2.781883933
	200	3.211560943	3.131807327	3.000971743
	300	3.359055219	3.169682359	3.052746692
	400	3.399024295	3.197298010	3.086605477

With pct.ribo

		N.features		
		1000	2000	3000
K.filter	100	3.347848350	2.993724544	2.769168720
	200	3.185957478	2.929128462	2.644296833
	300	3.317477891	3.199982175	2.720265794
	400	3.365077459	3.181236234	2.665079598

Batch effect correction affects the detectability of biological heterogeneity

We further looked the detectability of biological heterogeneity after different approaches for batch effect correction. For this purpose, we compared the clustering and number of differentially expressed genes of objects with different Lisi scores. We observe that integrated objects with little remaining batch effect and a high Lisi score (table S2.4) yield in less differentially expressed genes and thus less biological information than integrated objects with more remaining batch effect (table S2.5). In addition, the object with high Lisi score identifies two different clusters at a resolution of 0.5, whereas the object with the Lisi score of 2.64 already detects two clusters at a clustering resolution of 0.3. Moreover, clusters with only a limited number of differentially expressed genes, as in table S2.4, do not provide much more insight in the heterogeneity of the samples, suggesting overcorrection and subsequent overclustering of these samples. We also compared integrated objects with the same amount of batch effect (***Lisi score = 3.18***), but different integration settings (***n.features = 1000, k.filter=200*** and ***n.features = 2000, k.filter = 400).*** Interestingly, clustering of these objects resulted in a notably different biological output, as they largely differ in number of clusters and differentially expressed genes (table S2.6 and S2.7). So the choice of integration parameter settings can influence the detectability of biological heterogeneity, even if the remaining batch effect of two object is the same.

Table S2.4: Clustering and differential gene expression analysis for different resolutions (with pct ribo, nfeatures = 1000, k.filter = 200, Lisi score = 3.37). NA = not available

Resolution	Nr of clusters	Nr of genes cluster 1	Nr of genes cluster 2	Nr of genes cluster 3	Nr of genes cluster 4
0.1	1	NA	NA	NA	NA
0.2	1	NA	NA	NA	NA
0.3	1	NA	NA	NA	NA
0.4	1	NA	NA	NA	NA

0.5	2	2	2	NA	NA
0.6	2	2	2	NA	NA
0.7	2	2	2	NA	NA
0.8	3	8	2	2	NA
0.9	3	2	3	6	NA

Table S2.5: Clustering and differential gene expression analysis for different resolutions (with pct.ribo, k.filter=200, n.features=3000, Lisi score = 2.64). NA = not available

Resolution	Nr of clusters	Nr of genes cluster 1	Nr of genes cluster 2	Nr of genes cluster 3	Nr of genes cluster 4
0.1	1	NA	NA	NA	NA
0.2	1	NA	NA	NA	NA
0.3	2	105	105	NA	NA
0.4	2	119	119	NA	NA
0.5	2	109	109	NA	NA
0.6	3	12	113	11	NA
0.7	3	11	121	15	NA
0.8	4	101	5	4	17
0.9	4	106	5	5	20

Table S2.6: Clustering and differential gene expression analysis for different resolutions (with pct.ribo, k.filter=200, n.features=1000, Lisi score = 3.18). NA = not available

Resolution	Nr of clusters	Nr of genes cluster 1	Nr of genes cluster 2	Nr of genes cluster 3	Nr of genes cluster 4
0.1	1	NA	NA	NA	NA
0.2	1	NA	NA	NA	NA
0.3	2	55	55	NA	NA
0.4	2	39	39	NA	NA
0.5	2	33	33	NA	NA
0.6	2	32	32	NA	NA
0.7	3	31	13	7	NA
0.8	3	31	12	7	NA
0.9	4	55	6	12	8

Table S2.7: Clustering and differential gene expression analysis for different resolutions (with pct.ribo, k.filter=400, n.features=2000, Lisi score = 3.18). NA = not available

Resolution	Nr of clusters	Nr of genes cluster 1	Nr of genes cluster 2	Nr of genes cluster 3	Nr of genes cluster 4
0.1	1	NA	NA	NA	NA
0.2	1	NA	NA	NA	NA
0.3	2	64	64	NA	NA
0.4	2	71	71	NA	NA
0.5	2	22	22	NA	NA
0.6	3	3	21	64	NA
0.7	3	2	21	66	NA
0.8	4	0	9	8	72
0.9	4	2	9	10	75

Conclusion

In this study, we investigated the effects of adjusting integration and clustering parameters in scRNAseq analysis using the R package Seurat. For this purpose, we developed and tested a bioinformatic tool to easily explore the effects of integration and clustering parameters in Seurat's scRNAseq analysis. Besides the development of this tool for future analyses, we show that it is advisable to perform integration instead of merging of our samples to reduce the amount of batch effect. For these specific samples normalization without regressing for ***pct.ribo*** also results in less batch effect. Additionally, the amount of batch effect correction and thus the amount of remaining batch effect is mostly determined by the argument ***n.features*** and then ***k.filter***. We also show that overcorrection of the data can decrease the detectability of biological heterogeneity. Interestingly, the values chosen for ***n.features*** and ***k.filter*** during integration influence the biological output even when the batch effect is the same. This suggests that Seurat's integration functions manipulate the data in more ways than only batch effect correction.

This analysis does not offer a universal solution for optimal parameter settings, as the choice of parameter settings depends on the specific samples used in each individual analysis. In every case, the balance between batch effect correction and biologically relevant top genes should be carefully assessed when adjusting integration and clustering parameter settings. This study does give an overview of the impact of specific arguments in the Seurat workflow and thus helps to predict the effect of specific parameter alterations. In addition, it emphasizes the importance of understanding the different arguments in a function, since integration approaches with similar batch effects can still result in different clustering if different parameter settings are used.

Supplement 3. R script for Seurat v4 analysis

```
#### 0. Loading required packages to run this R script for Seurat v4 ####
  library("Seurat")
  library('dplyr')
  library('sctransform')
  library('ggplot2')
  library('MAST')
  library('metap')
  library('IDEATools')
  library('DESeq2')
  library('lisi')
  #library('kBET') # An alternative approach to quantify batch effect
  library('openxlsx')

#### 1. Setting working directory and loading data ####
  ## Set working directory to correct folder with the individual objects for
integrated script
  setwd("C:/Users/Marle Lokerse/Google Drive/I&I/Major internship/scRNAseq/Analyse
PRIDE/Objects for integrated script")

  ## Load the samples that have been normalized (SCTransform), regressed for %
mitochondrial genes+
  ## partial cell cycle regression ( + optional: % ribosomal genes)

  load(file = "C:/Users/Marle Lokerse/Google Drive/I&I/Major
internship/scRNAseq/Analyse PRIDE/Objects for integrated
script/s005SCT_ML_pctribo.Robj")
  load(file = "C:/Users/Marle Lokerse/Google Drive/I&I/Major
internship/scRNAseq/Analyse PRIDE/Objects for integrated
script/s008SCT_ML_pctribo.Robj")
  load(file = "C:/Users/Marle Lokerse/Google Drive/I&I/Major
internship/scRNAseq/Analyse PRIDE/Objects for integrated
script/s009SCT_ML_pctribo.Robj")
  load(file = "C:/Users/Marle Lokerse/Google Drive/I&I/Major
internship/scRNAseq/Analyse PRIDE/Objects for integrated
script/s010SCT_ML_pctribo.Robj")

#### 2. Integration and batch correction using Seurat integration function ####
## A: Merge without batch correction
  ML.merged <- merge(s005, c(s008, s009, s010), merge.data = T)

  setwd("C:/Users/Marle Lokerse/Google Drive/I&I/Major internship/scRNAseq/Analyse
PRIDE/Parameter analysis/Integrated objects")
  save(ML.merged, file = "DATE_Uncorrected_withribo_merged.Robj")

  #If the samples have been merged and saved before, reload the merged object
  load(file = "C:/Users/Marle Lokerse/Google Drive/I&I/Major
internship/scRNAseq/Analyse PRIDE/Parameter analysis/Integrated
objects/230605_Uncorrected_withribo_merged.Robj")

  #Runa PCA and UMAP on merged data
  ML.merged <- RunPCA(ML.merged, verbose = TRUE, features =
rownames(ML.merged@assays[["SCT"]]@scale.data))
  ML.merged <- RunUMAP(ML.merged, dims = 1:30)

  #Perform Lisi test to quantify the batch effect
  df_m <- as.data.frame(ML.merged@meta.data$orig.ident)
  colnames(df_m)[1] <- 'origID'
```

```
  res_m <- compute_lisi(as.matrix(ML.merged@reductions$umap@cell.embeddings), df_m
, label_colnames = 'origID')
  mean(res_m$origID)

## B: Integrate with batch correction
  #Set ident to patient/project/plate name
  s005 <- SetIdent(s005, value = "orig.ident")
  s008 <- SetIdent(s008, value = "orig.ident")
  s009 <- SetIdent(s009, value = "orig.ident")
  s010 <- SetIdent(s010, value = "orig.ident")

  #Create integrated Seurat object
  integrate.list <- objects()
  integrate.list$s005 <- s005
  integrate.list$s008 <- s008
  integrate.list$s009 <- s009
  integrate.list$s010 <- s010

  #Create vectors for the integration parameter values to be tested in the for
loop
  my_nfeatures <- c(1000, 2000, 3000) #These values can be adjusted to any value
to be tested
  my_kfilter <- c(100, 200, 300, 400) #These values can be adjusted to any value
to be tested

  #Create an empty dataframe where the batch effect will be pasted for all
different parameter settings
  batch_effect <- matrix(data = NA,ncol = length(my_nfeatures),nrow =
length(my_kfilter))

  df_batch_effect <- data.frame(batch_effect)
  colnames(df_batch_effect) <- my_nfeatures
  rownames(df_batch_effect) <- my_kfilter

  #Set working directory to save integrated objects
  setwd("C:/Users/Marle Lokerse/Google Drive/I&I/Major internship/scRNAseq/Analyse
PRIDE/Parameter analysis/Integrated objects")

  #Identify anchors
  reference.list <- integrate.list[c("s005", "s008", "s009", "s010")]

  for (i in 1:length(my_nfeatures)) {
    integrate.features <- SelectIntegrationFeatures(object.list = reference.list,
nfeatures = my_nfeatures[i])
    reference.list <- PrepSCTIntegration(object.list = reference.list,
anchor.features = integrate.features, verbose = TRUE)

    for (j in 1:length(my_kfilter)) {
      #Integrate data
      integrate.anchors <- FindIntegrationAnchors(object.list = reference.list,
normalization.method = "SCT", anchor.features = integrate.features, verbose =
TRUE, k.filter = my_kfilter[j])
      ML.integrated <- IntegrateData(anchorset = integrate.anchors,
normalization.method = "SCT", verbose = TRUE)

      #Runa PCA and UMAP on integrated samples
      ML.integrated <- RunPCA(ML.integrated, verbose = TRUE, features =
rownames(ML.integrated@assays[["SCT"]]@scale.data))
      ML.integrated <- RunUMAP(ML.integrated, dims = 1:30)
```

```
      #Perform Lisi test to quantify the batch effect
      df <- as.data.frame(ML.integrated@meta.data$orig.ident)
      colnames(df)[1] <- 'origID'
      res <-
compute_lisi(as.matrix(ML.integrated@reductions$umap@cell.embeddings), df,
label_colnames = 'origID')
      df_batch_effect[j,i] <- mean(res$origID)

      #Save integrated object
      save(ML.integrated,
           file = paste("DATE_Uncorrected_withpctribo_nf",
                        my_nfeatures[i],
                        "_kf",
                        my_kfilter[j],
                        ".Robj",
                        sep=""
           )
      )
    }
  }

  #Save the datframe with the batch effect results
  setwd("C:/Users/Marle Lokerse/Google Drive/I&I/Major internship/scRNAseq/Analyse
PRIDE/Parameter analysis")
  write.xlsx(df_batch_effect, file =
"DATE_Uncorrected_withpctribo_dim30_lisitest.xlsx", rowNames = TRUE)

#### 3. Clustering ####
  ## Load an integrated dataset (specific nfeatures and kfilter) to be clustered
  setwd("C:/Users/Marle Lokerse/Google Drive/I&I/Major internship/scRNAseq/Analyse
PRIDE/Parameter analysis/Integrated objects")
  load(file="230604_Uncorrected_withpctribo_nf2000_kf400.Robj")

  ## Create vectors for the parameter values to be tested in the for loop
  my_resolutions <- c(seq(from = 0.1, to = 0.9, by = 0.1)) #These values can be
adjusted to any value to be tested

  ## Create an empty dataframe where the clusters per conditions will later be
placed
  clusters <- matrix(data = NA,ncol = 2, nrow = length(my_resolutions))
  df_clusters <- data.frame(clusters)
  colnames(df_clusters) <- c("Resolution", "Number of clusters")
  df_clusters[, 1] <- my_resolutions
  df_clusters_pos <- df_clusters

  for (i in 1:length(my_resolutions)) {
    ## Prepare data
    DefaultAssay(object = ML.integrated) <- "integrated"

    ML.integrated <- RunPCA(ML.integrated, verbose = TRUE)
    ML.integrated <- RunUMAP(ML.integrated, dims = 1:30)

    ## Find clusters
    ML.integrated <- FindNeighbors(
      object = ML.integrated,
      reduction = "pca",
      dims = 1:30)
```

```
    ML.integrated <- FindClusters(
      object = ML.integrated,
      resolution = my_resolutions[i])

    DimPlot(
      object = ML.integrated,
      reduction = "umap",
      pt.size = 1.2)

    DimPlot(
      object = ML.integrated,
      reduction = "umap",
      group.by = 'orig.ident',
      pt.size = 1.2
    )

    ## Save the number of clusters to the empty dataframe
    df_clusters[i, 2] <- nlevels(ML.integrated)
    df_clusters_pos[i,2] <- nlevels(ML.integrated)

    ## Set the amount of clusters (if you have tried multiple resolutions for
clustering, they are all saved as ML.integrated$...._res0.8/...res0.4 etc,)
    key <- paste("integrated_snn_res.", my_resolutions[i], sep = "")
    ML.integrated <- SetIdent(ML.integrated, value = key)

    ## Order defined clusters based on phylogenetic analysis of identity classes
    if (nlevels(ML.integrated) > 1) {
      ML.integrated <-
        BuildClusterTree(
          ML.integrated,
          reorder = TRUE,
          reorder.numeric = TRUE,
          dims = 1:30,
          verbose = TRUE
        )

      #### 4. Finding cluster defining markers ####
      DefaultAssay(object = ML.integrated) <- "RNA"
      ML.integrated <- SetIdent(ML.integrated, value = "tree.ident")
      ML.integrated <- PrepSCTFindMarkers(object = ML.integrated)

      ##Find Markers that are specific for each cluster
      Batchedmarkers.mast <-
        FindAllMarkers(
          ML.integrated,
          test.use = "MAST",
          slot = 'data',
          logfc.threshold = 0.25,
          min.cells.feature = 5,
          only.pos = FALSE,
          min.diff.pct = 0.10
        )

       ## Create list
      listDEgenes_mast <-
        split(Batchedmarkers.mast, f = Batchedmarkers.mast$cluster)

      ## Filter on adj.P-value
      listDEgenes_mast <-
```

```
        lapply(listDEgenes_mast, function(x) {
          dplyr::filter(x, p_val_adj < 0.05)
        })

      # Sort on logFC
      listDEgenes_mast <-
        lapply(listDEgenes_mast, function(x) {
          x <- x[order(x$avg_log2FC, decreasing = T), ]
        })

      ## Save number of genes per cluster to dataframe
      for (j in 1:nlevels(ML.integrated)){
        if (nlevels(ML.integrated) >= j) {
          df_clusters[i, paste("Nr of genes cluster ", j, sep= "")] <-
nrow(listDEgenes_mast[[j]])
        } else {
          df_clusters[i, paste("Nr of genes cluster ", j, sep= "")] <- "NA"
        }
      }

      ## Save number of positive log fold changed genes per cluster to dataframe
      for (j in 1:nlevels(ML.integrated)){
        if (nlevels(ML.integrated) >= j){
          positive_foldchange <- listDEgenes_mast[[j]][,"avg_log2FC"] > 0
          df_clusters_pos[i, paste("Nr of positive genes cluster ", j, sep= "")]
<- sum(positive_foldchange)
        } else {
          df_clusters_pos[i, paste("Nr of positive genes cluster ", j, sep= "")]
<- "NA"
        }
      }

      ## Write to Excel
       setwd("C:/Users/Marle Lokerse/Google Drive/I&I/Major
internship/scRNAseq/Analyse PRIDE/Parameter analysis/Cluster defining markers")
       write.xlsx(listDEgenes_mast,file =
paste("DATE_Uncorrected_withpctribo_nf2000_kf400_dim30_GE_",
my_resolutions[i],".xlsx"))
       write.xlsx(df_clusters, file =
"DATE_Uncorrected_withpctribo_nf2000_kf400_dim30_clusters.xlsx")
       write.xlsx(df_clusters_pos, file =
"DATE_Uncorrected_withpctribo_nf2000_kf400_dim30_posclusters.xlsx")
    }
  }
```