Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Biologie

Studijní obor: Imunologie



Tereza Podolská

IMUNOFENOTYP MALIGNÍCH BUNĚK DĚTSKÝCH AKUTNÍCH LEUKEMIÍ A JEHO VÝVOJ V PRŮBĚHU ONEMOCNĚNÍ

Leukaemia associated immunophenotype in childhood acute leukaemias and its development during the course of disease

Diplomová práce

Vedoucí diplomové práce: RNDr. Karel Fišer, Ph.D.

Praha, 2020

Ráda bych poděkovala svému školiteli RNDr. Karlu Fišerovi, Ph.D.
za vedení této práce a trpělivé uvedení do jazyka R a světa bioinformatické imunologie.
Velký dík patří rovněž mé rodině.

Prohlášení:

Prohlašuji, že jsem tuto bakalářskou práci zpracovala samostatně, pod vedením školitele RNDr. Karel Fišer, Ph.D., a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, datum 6. 6. 2020

……..

Tereza Podolská

Abstrakt

Akutní lymfoblastická leukémie (ALL) je nejčastější dětskou malignitou. K jednomu z nedávných zlepšení v léčbě ALL došlo zavedením monitoringu minimální reziduální nemoci (MRD), který umožňuje přizpůsobit léčbu na základě stratifikace rizika. Monitoring MRD rovněž umožňuje zvolit vhodnou léčbu při relapsu onemocnění, indikovat transplantaci kmenových buněk (SCT) a umožňuje personalizovaný přístup k pacientům podstupujícím SCT. Jedním ze základních způsobů detekce hladiny MRD je charakterizace leukemických blastů pomocí průtokové cytometrie.

Hlavním limitem průtokové cytometrie je však to, že analýza dat je manuální a závislá na vysoce expertních znalostech. Tento postup je subjektivní a ne zcela dostatečný pro analýzu komplexních datových souborů užívaných v současnosti. Přestože jsou dostupné nové výpočetní nástroje pro zpracování multidimenzionálních dat z průtokové cytometrie, je zároveň potřeba tyto metody otestovat a přizpůsobit je klinickému prostředí.

Cílem této práce je detekovat imunofenotypy asociované s leukémiemi a jejich vývojem pomocí počítačem asistované analýzy u sady diagnostických souborů vybraných na základě informací o více než třech stech tisíců datasetů z víceparametrové průtokové cytometrie. Pokročilé bioinformatické nástroje pomohou detekovat blastické a zdravé hematopoetické populace, odvodit jejich imunofenotypy a identifikovat příspěvky jednotlivých CD molekul. Výsledné leukemické imunofenotypy budou porovnány s klinicky reportovanými imunofenotypy.

Klíčová slova: Akutní lymfoblastická leukémie, imunofenotyp, průtoková cytometrie, bioinformatika

Abstract

Acute lymphoblastic leukaemia (ALL) is the most frequent childhood malignancy. One of the recent improvements in ALL treatment was the introduction of minimal residual disease (MRD) monitoring that enables risk stratification based treatment adaptation. The same MRD monitoring helps to choose relapse treatment, to guide indication for stem cell transplantation (SCT) and allows for a more personalized management of patients undergoing SCT. One of the main routes of MRD levels detection is characterisation of leukemic blasts using flow cytometry.

However, flow cytometry is limited by its mainly manual expertise-based analysis. Such analysis is subjective and clearly insufficient for current complex data. While new computational tools are available for multidimensional flow cytometry data, there is an urgent need to test and adapt them for the use in clinical environment.

The goal of this thesis is to detect immunophenotypes associated with leukaemia and their development by leveraging machine-assisted analysis of a set of diagnostic files selected based on information about more than three hundred thousand of multiparameter flow cytometry datasets. Advanced bioinformatic tools will help to detect blast and healthy haematopoietic populations, to derive their immunophenotypes and to identify individual CD marker contributions. Resulting leukemic immunophenotypes will be compared to the clinically reported immunophenotypes.

Keywords: Acute lymphoblastic leukaemia, immunophenotype, flow cytometry, bioinformatics

Contents

List of abbreviations vii
List of markers and their approved gene symbols viii
1 Introduction 1
2 Literature review 2
2.1 Acute lymphoblastic leukaemia 2
2.1.1 Diagnostics 4
2.1.2 Leukaemia associated immunophenotype 4
2.2 Flow cytometry 5
2.2.1 Flow cytometry in diagnostics 8
2.3 FCS file structure 8
2.4 Preprocessing of FC data 10
2.4.1 Compensation 10
2.4.2 Transformation 12
2.4.3 Normalisation 13
2.5 Algorithms for FC data analysis 14
2.5.1 HCA – Hierarchical Clustering Analysis 14
2.5.2 FlowSOM 15
3 Aims of the thesis 16
4 Materials and methods 17
4.1 Hardware 19
4.2 R libraries 19
4.2.1 flowCore 19
4.2.2 FlowSOM 20
4.2.3 base library 20
4.3 In house libraries 20
4.3.1 flowPeel 20
4.3.2 flowPheno 21
4.4 FCS files in CLIP database 23
5 Results 25
5.1 Processing of information about the FCS files 25
5.2 Testing and choice of an appropriate data transformation 32
5.3 Automatic detection of populations in files of a selected FC panel 33

5.3.1 HCA – hierarchical clustering 39
5.3.2 FlowSOM 54
5.4 Automatic reporting of immunophenotype of detected cell populations 59
5.5 Comparison of automatic and clinical immunophenotype report 81
5.6 Development of immunophenotype during the course of ALL 87
6 Discussion 90
6.1 Processing of information about the FCS files 90
6.2 Automatic detection of populations in files of a selected FC panel 90
6.3 Comparison of automatic and clinical immunophenotype report 91
6.4 Development of immunophenotype during the course of ALL 93
7 Conclusion 94
8 References 96
Supplement 1 – Matrices of scatterplots of the resulting HCA clustering of diagnostic files 2-19 100
Supplement 2 - Matrices of scatterplots of the resulting FlowSOM clustering of diagnostic files 2-19 119
Supplement 3 – Full list of all R libraries used 138

List of abbreviations

ALL	acute lymphoblastic leukaemia
AML	acute myeloid leukaemia
B-ALL	B-lymphoblastic leukaemia
CD	cluster of differentiation
CLIP	Childhood Leukaemia Investigation Prague, laboratory of the Second Faculty of Medicine of the Charles University and the Motol University Hospital
D0	day 0 of ALL treatment
D8	day 8 of ALL treatment
D15	day 15 of ALL treatment
FDE	Flow Diagnostics Essential code for reporting immunophenotype
FISH	fluorescence *in situ* hybridization
FC	flow cytometry
FCS	"flow cytometry standard", flow cytometry data file format
HCA	Hierarchical Clustering Analysis
HLA-DR	Human Leukocyte Antigen – DR isotype
Ig	immunoglobulin
PCR	polymerase chain reaction
MHCA	Mahalanobis Hierarchical Clustering Analysis
MHC-II	major histocompatibility complex class II
MPAL	mixed-phenotype acute leukaemia, acute lymphoblastic leukaemia of ambiguous lineage
MRD	minimal residual disease
T-ALL	T-lymphoblastic leukaemia
TCR	T cell receptor
WBC	white blood cell

List of markers and their approved gene symbols

Marker	Approved gene symbol according to HGNC
CD1a	CD1A
CD2	CD2
CD3	CD3D, CD3E, CD3G
CD4	CD4
CD5	CD5
CD7	CD7
CD8	CD8A, CD8B
CD10	MME
CD11b	ITGAM
CD13	ANPEP
CD15	FUT4
CD19	CD19
CD20	MS4A1
CD22	CD22
CD33	CD33
CD45	PTPRC
CD66c	CEACAM6
CD79a	CD79A
CD99	CD99
Tdt	DNTT

1 Introduction

Acute lymphoblastic leukaemia (ALL) is the most frequent childhood malignancy (Swerdlow *et al.*, 2016). ALL is diagnosed based on analysis of samples of patients mainly from peripheral blood and bone marrow aspirates. The main methods of the analysis include flow cytometry (FC) and various molecular genetic methods. During the treatment, the level of minimal residual disease (MRD) is monitored by FC and PCR. (Vrooman and Silverman, 2016) The MRD (the amount of remaining malignant cells) monitoring is useful especially for treatment stratification and has led to a large improvement of leukaemia treatment (Starza *et al.*, 2019).

The flow cytometry data are usually analysed manually by implementing a "gating" strategy, a process of defining boundaries around populations of certain characteristics. However, this approach requires experience, and is subjective and time consuming. Furthermore, the recent progress of flow cytometry technology has led to a growth in number of parameters measured by cytometers, which further complicates the FC data analysis by "gating". (Verschoor *et al.*, 2015; Dworzak *et al.*, 2018) There are many computational methods for analysing multiparameter cytometry data that provide a good alternative for "gating" methods. (Fišer *et al.*, 2012; Van Gassen *et al.*, 2015)

In the FC database of CLIP (Childhood Leukaemia Investigation Prague) laboratory of the Second Faculty of Medicine of the Charles University and the Motol University Hospital, there are approximately 350 thousand of FC files. From these files, a large portion includes measured data from ALL patient samples. What is more, a large patient database can be accessed to link the FC data files to clinically reported immunophenotypes.

Metainformation about all flow cytometric files in the CLIP database were extracted in this thesis, analysed and used for file selection.

In this thesis, an R analytical protocol for automatic preprocesing of FC data, automatic identification of relevant populations and automatic determination of immunophenotype of these populations is proposed. This protocol is applied on a selection of diagnostic files from the CLIP FC database and the results are compared to the clinically reported immunophenotype. Furthermore, files measured during day 0, day 8 and day 15 of the ALL treatment were analysed by this analytical protocol, proving that it is capable of detecting MRD populations and the changes in blast immunophenotype during treatment.

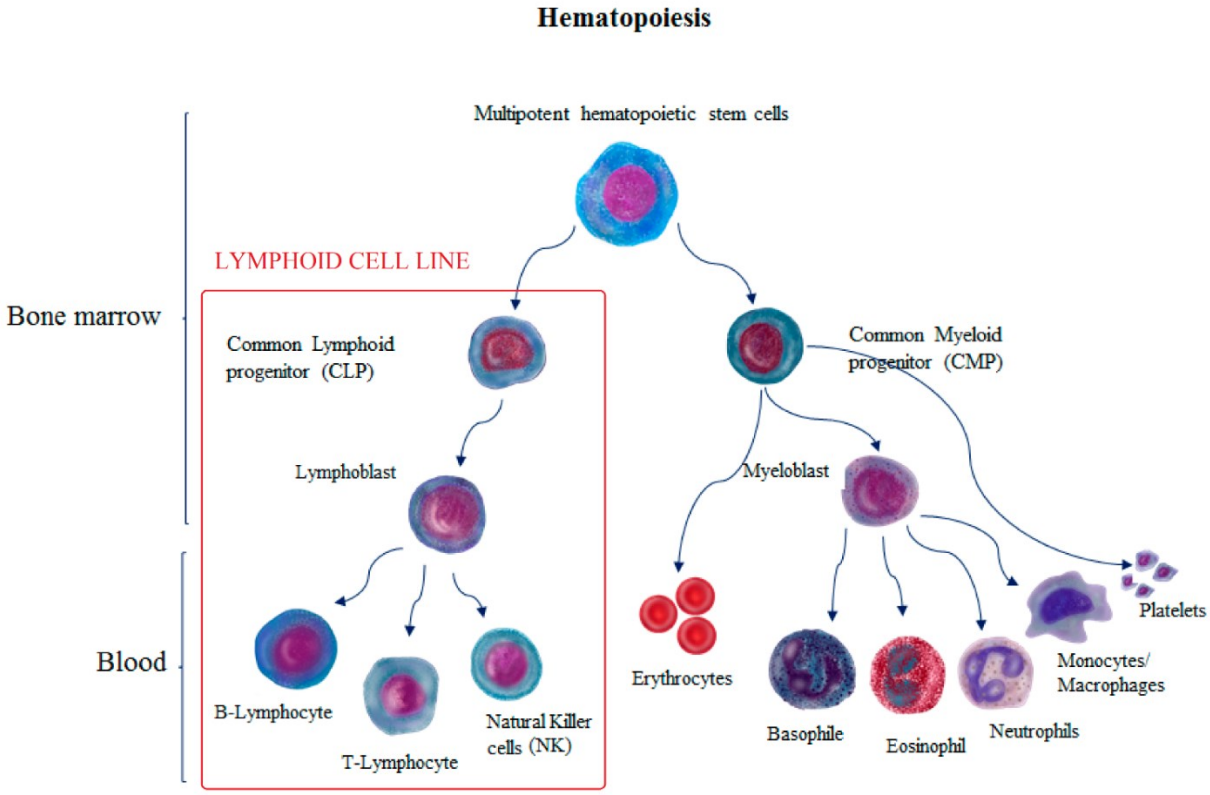
2 Literature review

2.1 Acute lymphoblastic leukaemia

Acute lymphoblastic leukaemia (ALL) is a type of blood cancer derived from the hematopoietic cell precursors of the lymphoid line. For the diagram of haematopoiesis with marked lymphoid cell line see Fig. 1. Acute leukaemias are caused by a series of mutations occurring in the process of haematopoiesis in the bone marrow and progress rapidly in comparison to chronic leukaemias. The genetic changes resulting in leukaemia may be various deletions, translocations and amplifications of chromosomal regions or mutations of single genes. These mutations result in changes in cellular morphology, an uncontrollable proliferation of the cells and lack of response to cell controlling mechanisms. (Rose-Inman and Kuehl, 2014) While these leukaemic cells proliferate abnormally, the volume of healthy blood cells decreases. The WHO differentiates the ALL into B-lymphoblastic leukaemia (B-ALL), derived from B cell precursors, and T-lymphoblastic leukaemia (T-ALL), derived from T cell precursors. Both types can be derived from several stages of precursor lymphoid cells, or in some cases even mature cells. The development stage further subspecifies the ALL typology (B-ALL into pro-B-ALL, common ALL, pre-B-ALL and mature B-ALL and T-ALL into pro-T-ALL, pre-T-ALL, cortical-T-ALL and mature-T-ALL). Acute lymphoblastic leukaemia of ambiguous lineage (or mixed-phenotype acute leukaemia - MPAL) can also be recognised as an ALL, because it originates either from both B-cell and T-cell precursors, or both lymphoid and myeloid precursor cells. (Swerdlow *et al.*, 2016)

The reduction of normal haematopoiesis causes many symptoms typical for ALL. The lack of erythrocytes results in anemia and causes weakness, paleness and fatigue (Teuffel *et al.*, 2008), lack of neutrophils and monocytes sensitivity to minor infections, especially during treatment (Ozdemir *et al.*, 2016) and the lack of thrombocytes frequent hemorrhages and easy bruising (Dixit *et al.*, 2007). Another common symptoms include splenomegaly (Manoharan *et al.*, 1980), hepatomegaly (Sayyed *et al.*, 2018), lymph node enlargement (Karanth *et al.*, 2008), excessive sweating (Rose-Inman and Kuehl, 2014), reduction of weight (Withycombe *et al.*, 2015) and bone pains caused by the expansion of bone marrow (Sakata *et al.*, 2014).

Fig. 1 - A diagram of differentiation of hematopoietic stem cells into individual blood cell types. The lymphoid cell line (B and T lymphocytes, Natural Killer cells, lymphoid progenitors) from which the ALL is derived is highlighted by a red frame. Adopted from (Alomari *et al.*, 1982)



Most of the cases of ALL develop in children. At the same time, it is the most frequent cancer among children. (Swerdlow *et al.*, 2016) In the Czech Republic, the incidence of ALL is about 7/100 000 for children under 4 years and 1-1.5/100 000 for adults. (Adam, Krejčí and Vorlíček, 2008)

The survival rate of acute lymphoblastic leukaemia patients has greatly increased in the past decades. While in the year 1975-1977 the 5-year relative survival rate for age group under 15 years was approximately 60 % in the USA, between the years 2006-2013 it was about 90 %. (Hunger and Mullighan, 2015; Howlader *et al.*, 2016) However, the survival rate for adults is much lower, about 40-50 % (Gökbuget and Hoelzer, 2009). The decline in ALL mortality can be largely attributed to the development of ALL diagnostics, disease monitoring and treatment methods. (Smith *et al.*, 2014) However, the relapse rate for the childhood ALL is about 20 % (Rose-Inman and Kuehl, 2014).

2.1.1 Diagnostics

The suspicion of leukaemia may arise after observations of clinical symptoms and abnormal WBC (white blood cell) count in the peripheral blood smear (Rose-Inman and Kuehl, 2014).

The definitive leukaemia diagnostics is based on the analysis of peripheral blood or bone marrow biopsy samples. The analysis consists of pathological examination, immunophenotypisation by flow cytometry (see chapter 2.2.1) and cytogenetic characteristics. The genetic characterisation involves the detection of karyotype and chromosomal aberrations by FISH and the detection of leukaemia-specific fusion transcripts and immunoglobuline chains by PCR and next generation sequencing. (Corradi *et al.*, 2008; Vrooman and Silverman, 2016)

After the introduction of therapy, it is fundamental to monitor the minimal residual disease (MRD), the number of malignant cells still present in body. The MRD is detected by the flow cytometry or PCR and helps to assess the effectiveness of treatment. The FC focuses on monitoring the leukaemia associated phenotype and the PCR on detecting the fusion transcripts or the rearrangements of TCR and leukemic Ig genes. (Starza *et al.*, 2019)

2.1.2 Leukaemia associated immunophenotype

The acute lymphoblastic leukaemia is divided into B cell ALL and T-cell ALL lineage.

Because of the nature of genesis of leukaemias by an accumulation of mutations, the expressed immunophenotype of B-ALL may be very diverse as it is directly dependent on particular genetics of the population of origin. However, a common sign is a co-expression of precursor and mature lymphocyte markers at the same time. Mature B-cell markers that are almost always expressed in B-ALL are CD19, CD22 and Cd79a. Very commonly precursor B cell markers CD10, CD34 and Tdt (terminal deoxynucleotidyl transferase) are present in B-ALL. (Onciu, 2009; Peters and Ansari, 2011) On the other hand, an expression pattern of common lymphocyte markers CD45 and CD38 and common B lymphocyte marker CD20 can be very variable from negative to strong (Peters and Ansari, 2011). Sometimes even myeloid markers are expressed, e.g. CD13, CD33, CD11b, CD15, and CD66c (Onciu, 2009; Peters and Ansari, 2011). Another commonly expressed marker in B-ALL is HLA-DR, an MHC-II component. (Peters and Ansari, 2011)

The B-ALL may also be classified into subtypes based on Immunoglobulin (Ig) expression – early pre-ALL with no Ig expression, pre-B-ALL with the expression of Ig heavy chain in

cytoplasm and transitional pre–B-ALL with Ig heavy chain expression in cytoplasm and surface. The light Ig chain is expressed very rarely. (Kansal *et al.*, 2004)

The T-ALL blastic cells may often express characteristic T-cell markers. CD4, usually expressed on T helper cells, and CD8, usually expressed on T cytotoxic cells, may be expressed by the T-ALL populations. In addition to these, other T-cell specific markers may also occur, such as CD1a, CD2 CD3, CD5 and CD7. (Onciu, 2009) However, based on the development stage of the precursor T-cell of origin and their diversity not all of these markers may be co-expressed. Based on this specific immunophenotype, T-ALL may by further classified. (Swerdlow *et al.*, 2016) The expression of Tdt and CD34 often correlates with immature T-cell origin of ALL. As in the case of B-ALL immunophenotype, the lymphocyte markers CD45 and CD38 may also be expressed with a vast variability. (Peters and Ansari, 2011) CD10, CD99 and HLA-DR markers may also be expressed (Onciu, 2009).

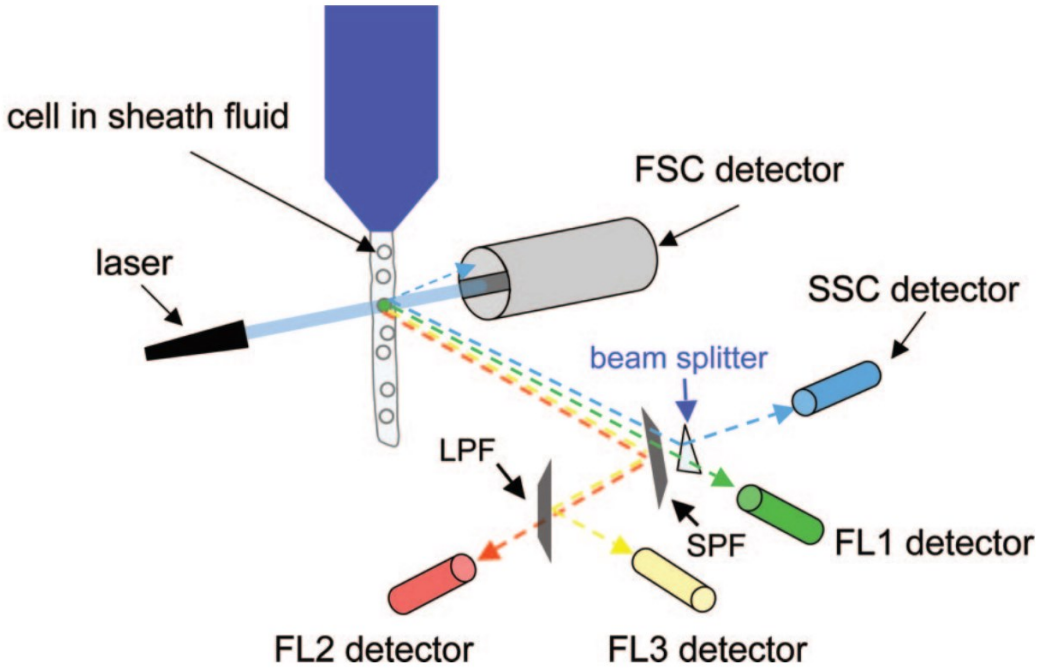
2.2 Flow cytometry

Flow cytometry (FC) is a technique for measuring multiple biological structures on the surface or inside of particles, usually cells. It is important that the measured sample by FC is a suspension. If it is needed to measure a solid sample, it needs to be dissociated into a single cell solution. The sample is stained by fluorescent antibodies that bind specifically on the desired structure to be measured. This suspension is drawn into the cytometer together with a sheath fluid which allows the flow of the sample through the cytometer. This fluid is flows in a focused stream of single particles, which are excited one-by-one by several lasers of a different wavelength. This allows the fluorescent labels to absorb the light of a specific wavelength and emit a fluorescent signal that passes through wavelength specific filters and mirrors before it is detected by a number of detectors. This signal is transformed into a digital signal and saved. For a diagram of a flow cytometer optical compartment see Fig. 2. (Givan, 2011; Jahan-Tigh *et al.*, 2012; Maciorowski, Chattopadhyay and Jain, 2017)

In addition to the detection of wavelength specific signal, the forward scatter (FSC) and side scatter (SSC) of individual particles are also measured. Both of these parameters measure light refraction while the particle is passing the laser, FSC in the direction of the laser and SSC in a 90 degree angle to the light source together with all other fluorescence intensity detectors (see Fig. 2). FSC is a characteristic that correlates with the size of the particle and SSC correlates with its complexity. (Leary *et al.*, 1979; Givan, 2011) In immunology, both of these characteristics are used to distinguish cell type of the population. E.g. granulocytes are cells of

a large size and with a high level of granularity compared to other cell types and therefore can be recognised as the subpopulation of high level of FSC and SSC. (Gorczyca *et al.*, 2011)

Fig. 2 - A simplified diagram of an optical compartment of a flow cytometer. Single cells are passing through the capillary and the laser excites the fluorochromes to emit light. This light is detected by a series of detectors: FSC detector (measuring in the direction of the laser), SSC detector and fluorescence intensity detectors of a specific wavelength - FL1-FL3 (measuring perpendicularly to the laser). Note, that there may be more than three fluorescence intensity detectors in the cytometer. The light emitted by the particles passes through a series of long pass and short pass filters (LPF and SPF) before it is measured by the detectors. Adopted from (Delude, 2005).



A fluorescence activated cell sorter (FACS) is capable of sorting the measured particles based on the measured data. The sample is separated into droplets containing a single particle after the measurement and charged based on the measured characteristics. The charged droplets enter an electromagnetic field, are deflected based on the charge and sorted into tubes. The sorting strategy can be defined in the software prior to measurement and includes characteristics such as size, shape and structure of particles, and the fluorescence intensities. (Givan, 2011) In fact, the first flow cytometers had the ability to sort cells and were developed for this reason. (Fulwyler, 1965; Kamentsky and Melamed, 1967)

There are many parameters that can be measured in FC. Most commonly, surface or cytoplasmic structures can be labelled by the fluorescent antibody and measured. This is used especially for immunophenotyping where a combination of specific markers present on or in a cell (especially CD molecules) can identify characteristics of various cell populations and their cell type. However, FC can be used for many other purposes, such as apoptosis indication by

staining apoptosis-specific biochemical and morphological structures (Wlodkowic, Skommer and Darzynkiewicz, 2009), measuring the intracellular pH or membrane potential by fluorescently labelling ions (Valli *et al.*, 2005; Shapiro, 2008) or determining the cell cycle by staining of DNA (Darzynkiewicz, Juan and Bedner, 1999).

The number of parameters that are possible to be measured at the same time grows rapidly, bringing new possibilities for the complex FC analysis and diagnostics. However, the rise of the massively multiparameter cytometry and complexity of its data are a challenge for the data analysis, data compensation (see chapter 2.4.1) and data visualisation. A cytometer capable of measuring dozens of parameters at the same time produces high dimensional data that could be difficult to be processed by conventional "gating" strategies. Therefore, the usage of various algorithms to process these data is encouraged. (O'Donnell, Ernst and Hingorani, 2013)

The most convenient samples to be measured in FC are liquid samples such as blood, bone marrow fluid or biopsies. However, even solid tissue samples can be measured after their homogenisation into a suspension. What is more, the FC is not only able to perform on cells, but also particles, e.g. labelled viruses (Lippé, 2017), cell nuclei (Guérin *et al.*, 2017), chromosomes (Doležel *et al.*, 2011) or DNA fragments (Hussels, Engel and Bock, 2019). Multiplex assays are also often used for measuring cytokine levels. The principle is, that fluorescent beads coated with a cytokine-specific antibody bind the cytokines and can be detected by the flow cytometer. (Castillo and MacCallum, 2012)

A novel method of mass cytometry combines the methods of flow cytometry and mass spectroscopy. It allows measuring of more than 40 parameters at the same time. The principle is to mark biological structures by conjugates of antibodies with stable isotopes not occurring in biological systems instead of fluorochromes. Each isotope has a different atomic weight and therefore can be detected with high sensitivity. The marked particles are nebulised into single cells, which are ionised into a cloud in an argon plasma. This cloud is then measured by time-of-flight (TOF) mass spectrometry. The advantages of the mass cytometry are high sensitivity, the ability to measure great amount of parameters and the reduction of spectral overlap in individual channels. (Bandura *et al.*, 2009) However, it is very challenging to visualise the mass cytometry data due to the high dimensionality (Kimball *et al.*, 2018). Also, the cell throughput in mass cytometry is limited to hundreds per second, compared to the flow cytometry throughput of tens of thousands of cells per second (Robinson *et al.*, 2017).

2.2.1 Flow cytometry in diagnostics

Together with molecular and genetic assays, the immunophenotypisation of cell populations by multiparameter flow cytometry is one of the main tools for the diagnostics and classification of leukaemia, lymphoma and other haematological malignancies. Furthermore, it is crucial for the monitoring the minimal residual disease (MRD), which is used for determining the disease progression and effectiveness of the treatment. (Peters and Ansari, 2011) MRD is defined as the quantity of malignant cells still remaining in body during treatment (Starza *et al.*, 2019). Generally, the samples used for the diagnostics are peripheral blood, bone marrow fluid or lymph node biopsies. The expression of individual markers is usually determined by a "gating" strategy, which is a series of selecting populations based on desired characteristics of markers and scatter parameters (FSC and SSC). (Peters and Ansari, 2011)

In order to standardise the immunophenotypisation diagnostics of haematological malignancies procedures several flow cytometry standards have been proposed. One of these proposed standard is EuroFlow. The EuroFlow Consortium suggested an 8-parameter flow cytometry with standardised procedures and fluorochrome panel designs. Following the EuroFlow recommendations results in high data reproducibility and the ability to compare results between all facilities implementing EuroFlow protocol. (Kalina *et al.*, 2012)

2.3 FCS file structure

The flow cytometry data are saved in the FCS file format. This binary format is defined by the FCS standard and is widely recognised by the flow cytometer software manufacturers. This norm has been first introduced in 1984 by the ISAC (International Society for the Advancement of Cytometry) and has been updated several times since then. The latest update of the standard was the FCS 3.1 in 2010. While the FCS version 3.1 is widely recognised by the software and is the most common FCS format today, many databases store older FCS files with the format of versions FCS 2.0 and 3.0 as well (Spidlen *et al.*, 2009)

The FCS file is divided into three binary-encoded segments - HEADER, TEXT and DATA. The HEADER segment defines the FCS version and length and position of other segments of the file. The TEXT segment contains string of pairs of keywords and their values. These keywords contain metainformation about the flow cytometry data. There are many defined keywords in the FCS standard, each starting with the '$' character. Any number of keyword-value pairs can be saved in the TEXT segment. However, there are sixteen keywords that are

required by the FCS standard (see Table 1). Moreover, there are 54 optional keywords that may be used in the FCS file. Notable optional keywords are shown in the Table 2. (Spidlen *et al.*, 2009)

Table 1 - The list of required keywords in the TEXT segment of the FCS file. The keywords $PnB, $PnE, $PnN and $PnR are present in the file for each parameter measured and the "n" is replaced by the parameter number. Adopted from the supplementary data of the FCS 3.1 standard definition paper (Spidlen *et al.*, 2009).

Keyword	Description
$BEGINANALYSIS	Byte-offset to the beginning of the ANALYSIS segment
$BEGINDATA	Byte-offset to the beginning of the DATA segment
$BEGINSTEXT	Byte-offset to the beginning of a supplemental TEXT segment
$BYTEORD	Byte order for data acquisition computer
$DATATYPE	Type of data in DATA segment (ASCII, integer, floating point)
$ENDANALYSIS	Byte-offset to the last byte of the ANALYSIS segment
$ENDDATA	Byte-offset to the last byte of the DATA segment
$ENDSTEXT	Byte-offset to the last byte of a supplemental TEXT segment
$MODE	Data mode (list mode - preferred, histogram - deprecated)
$NEXTDATA	Byte offset to next data set in the file
$PAR	Number of parameters in an event
$PnB	Number of bits reserved for parameter number n
$PnE	Amplification type for parameter n
$PnN	Short name for parameter n
$PnR	Range for parameter number n
$TOT	Total number of events in the data set

Table 2 - The list of selected keywords in the TEXT segment of the FCS file. The $PnS keyword is present in the file for each parameter measured and the "n" is replaced by the parameter number. Adopted from the supplementary data of the FCS 3.1 standard definition paper (Spidlen *et al.*, 2009).

Keyword	Description
$CYT	Type of flow cytometer
$DATE	Date of data set acquisition
$OP	Name of flow cytometry operator
$PnS	Name used for parameter n
$SPILLOVER	Fluorescence spillover matrix

The DATA segment stores the measured flow cytometry data. Preferably, the data are stored in the list mode. It is also possible to save the data in the histogram mode, but this is generally not encouraged. The used mode of storing the data is defined by the `$MODE` required keyword in the TEXT segment. (Spidlen *et al.*, 2009)

If needed, an optional ANALYSIS segment may be added to the file. This segment may contain the information about the data processing results, such as immunophenotype. The ANALYSIS segment has a keyword-value structure similar to the TEXT segment. However, there are no required keywords in this segment and only optional keywords are defined by the FCS standard. It is also possible to add OTHER segment to the FCS file, even though this is not used very often. (Spidlen *et al.*, 2009)

By definition, it is possible to store more FC datasets in one FCS file. In this case each dataset contains its own HEADER, TEXT and DATA (and optionally ANALYSIS and OTHER) segments. However, it is generally not recommended to save the data this way. (Spidlen *et al.*, 2009)

The analysis of the FC data may be done either in a proprietary FC software (FlowJo, FCSExpress, BC Kaluza or BD DiVa) or a freeware FC software (FCSalyzer, Flowing SW). Alternatively, the FCS file may be analysed in a statistical computing programs with a specialised FC library (R, Python, MATLAB). This approach offers more flexibility in the data processing. In this thesis, the FC data are analysed in R.

In R the FCS file can be loaded into a flowFrame object, a list type object with tree slots that respects the original FCS file organisation. The first slot of the list, `description`, includes all the keyword-value pairs included in the TEXT segment of the FCS file. The measured data from the DATA segment are saved in the second slot of the list, `exprs`. The `parameters` slot is an annotated data frame containing the metadata for the columns in `exprs` slot and is derived from the parameter specific keywords. (Meur, Hahne and Ellis, 2007)

2.4 Preprocessing of FC data

The FC raw data needs to undergo several preprocessing steps before any data analysis and visualisation. This includes compensation, transformation and also normalisation, if the data from several experiments are compared.

2.4.1 Compensation

The emission spectra of fluorochromes used in cytometry often overlap, which creates a problem in the analysis of multiparameter cytometry data. Ideally, each detector should measure only one signal from one fluorescence dye. However, the spectral overlaps of the fluorescence channels create background noise and make it hard to distinguish individual signals. This can

be corrected by measuring the spectral overlap of each fluorochrome into channels of each other fluorochrome. (Roederer, 2001) The measurement of spillover is achieved by measuring Fluorescence Minus One (FMO) controls, containing all but one fluorochrome from the panel. The acquired values are saved in the spillover table, which is stored in the FCS file as a value of the `$SPILLOVER` keyword. For the earlier FCS version files, the spillover table is saved in the `$SPILL` keyword. Alternatively, the `$COMP` keyword may be used instead for storing the compensation matrix, and inverted spillover table. (Bagwell and Barden, 1990; Spidlen *et al.*, 2010)

The compensation is then accomplished by finding and extracting the spillover matrix and calculating the inverted matrix of the spillover table, the compensation matrix:

$$compensation\ matrix = \ spillover\ matrix^{-1} = \begin{bmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{bmatrix}$$

Then, the raw measured data is multiplied by this compensation matrix. The compensated value of one channel is calculated as the sum of raw detected values multiplied by their compensation coefficient in the said channel:

$$compensated\ fulorescence = A_1 \times F_A + B_1 \times F_B + C_1 \times F_C$$

where A_1, B_1 and C_1 are the compensation coefficients of A, B and C fluorochrome in the A fluorescence channel and F_A, F_B and F_C are the measured values of A, B and C fluorochormes. (BAGWELL and ADAMS, 1993)

Even though the mass cytometry does not involve fluorescence measurement, the compensation step is still needed. The measurements of the isotopes in mass cytometry is very accurate, in comparison to fluorochromes in flow cytometry, which create a significant spectral overlap. Overlap between channels is still present in the mass cytometry, but compared to the fluorochrome spillover, it is very limited. The source of the mass cytometry spillover is oxidation, impurities of the isotopes and properties of the mass cytometer. According to Chevrier et al., 2018, the spillover in mass cytometry is linear and the data can be compensated in a similar manner as the fluorescence spillover. It can be done by measuring stained single-positive antibody-capture beads in the mass cytometer and calculating the spillover matrix post-measurement. The spillover calculation and compensation can be achieved by the `CATALYST` R package. (Chevrier *et al.*, 2018)

2.4.2 Transformation

It is characteristic for the raw FC data to have a log-normal distribution. This means, that the logarithm of the values has the Gaussian normal distribution. If this data was to be visualised on a linear scale, majority of values would be concentrated in the near-zero area and it would be almost impossible to distinguish the negative and positive populations. Therefore, one of the major steps of preprocessing is the data transformation. (Parks, Roederer and Moore, 2006)

For many years, the simple *logarithmic* transformation was used in the analysis (the formula of the transformation is shown in Table 3). While this method solves the problem of the log-normal distribution, it does not work ideally for the near-zero area values. These low values often form a peak close to zero known as the "log artifact" as a result of logarithmic transformation. (Herzenberg *et al.*, 2006) Another problem of the *logarithmic* transformation is, that it cannot be used for negative and zero values. These negative values may occur in the dataset directly when measured, or in the process of data compensation. All these events are lost and cannot be analysed further as a result of the *logarithmic* transformation, even though they may include a large amount of cells. This loss of data may be prevented by shifting the whole dataset to positive values before using the transformation. However, the shift might introduce a problem for the data comparison. (Parks, Roederer and Moore, 2006)

Instead of the *logarithmic* transformation several log-like transformations are usually used nowadays. Log-like transformations, such as *hyperlog*, *biexponential* and *logicle* (the formulas of the transformations are shown in Table 3), are specific by displaying low values in a linear scale and higher values in a logarithmic scale (Parks, Roederer and Moore, 2006; Finak *et al.*, 2010). This scaling solves the problem of the "log artifact" and of the discarded negative and zero values. The disadvantage of the *hyperlog* transformation is, that the linearization near zero is not optimal and tends to create a small "log artifact". (Parks, Roederer and Moore, 2006). The *biexponential* transformation is similar to the *hyperlog*. Its specification is an asymmetric linear scale around zero. (Finak *et al.*, 2010) Finally, the *logicle* transformation is a variation of the *biexponential.* The main difference is an extra parameter specifying the width of the linearization (Moore and Parks, 2012).

Another widely used transformation is the *arcsinh*, hyperbolic arcsine (the formula of the transformation is shown in Table 3). According to Finak *et al.*, the *arcsinh* performed very well for the scatter parameters, but not so much for the fluorescence parameters. (Finak *et al.*, 2010)

The transformations need to be parametrised before being used, but usually the default value is used and corrected only if needed. However, it is recommended for the automated bioinformatical FC analysis to use function for the parameter estimation, as the correctness of the transformations cannot be checked. (Finak *et al.*, 2010) E.g. for the parameter choice of the *logicle* transformation a `flowCore` function `estimateLogicle` can be used (Meur, Hahne and Ellis, 2007). Similarly, the `flowPheno` in house library that is used for the FC data preprocessing in this thesis uses its own estimation algorithm for the choice of *logicle* parameters. (Koblížek et al., manuscript in preparation)

Table 3 - The formulas of frequently used transformation functions. The x value is transformed by the function and the parameters a,b,c,d,e,f and r determine attributes of the transformations. Some of the transformation formulas are presented as inverse functions. Adopted from (Bagwell, 2005; Parks, Roederer and Moore, 2006; Meur, Hahne and Ellis, 2007)

Transformation	**Formula**
Logarithmic	$f(x) = log_b x \frac{r}{d}$
Arcinh	$f(x) = asinh(a + bx) + c$
Hyperlog	$f^{-1}(x) = ae^{bx} + cx - f$
Biexponential	$f^{-1}(x) = ae^{bx} - ce^{dx} + f$
Logicle	$f^{-1}(x) = ae^{bx} - ce^{-dx} + f$

2.4.3 Normalisation

There are often technical and biological differences in FC data acquisition which result in the variability between samples. It is therefore essential to normalise the fluorescence intensities before comparing cell subpopulations between datasets. The aim of normalisation is to transform the measured values to the same scale and standardise the fluorescence values. (Finak *et al.*, 2014)

One of the normalisations that are frequently used for cytometry data is z-scoring. Z-score normalisation scales the data with the centre at 0 and the standard deviation of 1. The equation for z-score normalisation is

$$Z = \frac{x_i - \bar{x}}{s_x}$$

with the x_i being the original value, $\bar{x}$ the mean and s_x the standard deviation. The advantage of using the z-score normalisation is, that it handles outliers very well. (Malo *et al.*, 2006)

2.5 Algorithms for FC data analysis

There is a number of algorithms for flow cytometry analysis, both unsupervised and supervised. The supervised algorithms need to be provided with the dataset to be analysed and the training data. (Chester and Maecker, 2015)

These computational methods in combination with strong visualisation methods can provide a good alternative to the manual "gating" strategies. The data from massively multiparameter cytometry are hard to interpret from series of 2D scatterplots produced by the conventional "gating" analysis. Therefore it is preferred to analyse the high dimensional data by computational tools (Chester and Maecker, 2015)

The algorithms can be based on several computational approaches including clustering, dimensionality reduction (Chester and Maecker, 2015). Clustering (e.g. FlowSOM, hierarchical clustering) analyses the data by linking cells with similar characteristics (Van Gassen, Callebaut and Saeys, 2019). The dimensionality reduction (e.g. t-SNE, ve-SNE) is based on reducing the number of parameters into two or three dimensions, so that they can be visualised at the same time (Toghi Eshghi *et al.*, 2019).

The clustering algorithms for unsupervised populations (clusters) used in the thesis are explained below.

2.5.1 HCA – Hierarchical Clustering Analysis

The process of Hierarchical Clustering Analysis compares single cells and the profile of all their parameters and clusters the most similar ones into a hierarchical dendrogram. Gradually more cells are conjoined until it forms a large population. (Fišer *et al.*, 2012)

In this thesis an MHCA (Mahalanobis Hierarchical Clustering Analysis) algorithm implemented in an in-house R library `flowPeel` (see chapter 4.3.1) was used. The MHCA is

based on Mahalanobis distance measure instead of Euclidean distance measure and is designed specifically for FC datasets. (Fišer *et al.*, 2012; Krishnan and Jagannathan, 2015)

2.5.2 FlowSOM

The FlowSOM algorithm is also designed to perform on multiparameter FC data. It is accessible in a public R library `FlowSOM`. The principle of the algorithm is using an artificial neural network to arrange the individual events into Self-Organising Maps (SOM), which can be visualised as a minimal spanning tree. The SOM is based on two-level clustering - it consists of single nodes that correspond with the cell populations that are arranged into metaclusters. (Van Gassen *et al.*, 2015) FlowSOM is recommended for its fast runtimes and good performance (Weber and Robinson, 2016a).

3 Aims of the thesis

Aims of the thesis were the following:

Major aims

- To create an analytical protocol that would automatically, without human input, process flow cytometry data files, detect biologically relevant populations present in the data and report their clinically relevant immunophenotypes
- To implement this protocol into a working software tool which would be able to identify major malignant cell populations in a set of diagnostic ALL samples and to generate correct immunophenotype representation as compared to expert generated immunophenotypes
- To use such tool to detect malignant cells in ALL samples at the time of diagnosis and in the follow up samples and to compare their immunophenotypes.

Minor aims

- To create a database of metainformation about all FCS files in the CLIP database
- To evaluate available transformation function and choose best candidate for log-normally distributed flow cytometry data
- To test at least two clustering algorithms for cell population detection

4 Materials and methods

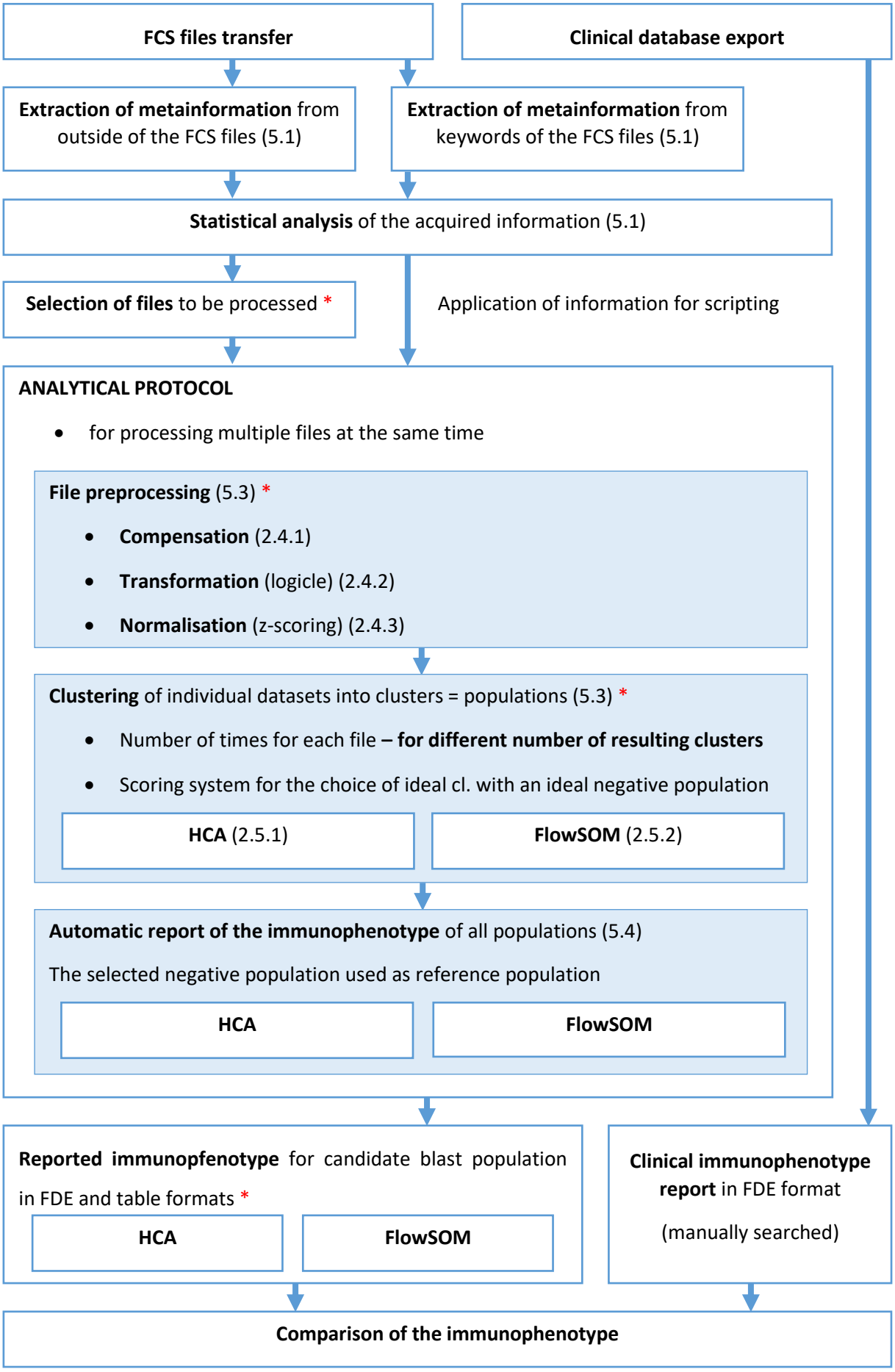
First of all, it is important to stress, that the results of this thesis were achieved by using *in silico* methods only. All code was written in the R open-source programming language in RStudio, an R integrated development environment. RStudio sessions used R version 3.6.2. An analytical protocol for processing FCS files and their metadata was created. Besides the original code, this protocol also implements many functions from public R libraries and *in house* functions, created in the CLIP laboratory.

The source of material for this thesis were are more than 350 thousand FCS files in the CLIP database (see chapter 4.4). A large portion of these files are diagnostic and follow up files of ALL patients. Clinically relevant data of each patient are present in the CLIP clinical database, including the FC sample information and clinically reported immunophenotypes.

The outline of the project is depicted in a diagram in Fig. 3. The first step of the project is FCS file transfer and extraction of metainformation from these files, such as size, measured panel of markers or the type of cytometer the file was measured on. These information were saved in a database and statistically analysed (see 5.1). Based on the database, the adequate files for further analysis were selected. The characteristics of the files were also relevant for scripting the functions for the data analysis, because they brought an insight into possible problems that might occur during the data processing.

Then, the analytical protocol for FCS files consisting of many functions was designed. The main aim was to process many files at the same time and produce a report of the immunophenotype of relevant size populations in these files. The first step of the protocol was data preprocessing consisting of compensation, logicle transformation and z-score normalisation of the data. Then, each dataset was clustered several times producing a variation of clustering sets each with a different number of clusters. There were two clustering algorithms used – HCA (see 2.5.1) and FlowSOM (see 2.5.2). For both clustering sets (HCA and FlowSOM) the best clustering of choice containing an ideal negative population (negative in all markers) in a spate cluster was chosen based on a novel scoring system (see chapter 5.3). This chosen iteration of clustering was used for determining the immunophenotype of relevant clusters with the selected negative population as a reference. The protocol produced an immunophenotype report in table and FDE format for HCA and FlowSOM clustering for all significant populations for all processed files. From these reports candidate blast populations

*Fig. 3 - Diagram of the workflow of the thesis project. * marks a single manual entry in the automatical protocol. All other steps are automated.*



were chosen manually. The reported immunophenotypes of these populations were then compared to the clinical report of immunophenotype from the clinical database.

Furthermore, an analysis of immunophenotype of follow up FCS files of several patients from day 0, day 8 and day 15 of the treatment by the analytical protocol was made to demonstrate the development of ALL immunophenotype during the course of disease (see 5.6).

4.1 Hardware

Computer intensive work for the thesis was done on PC:

- OS - Ubuntu version 18.10, CPU - Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz, RAM – 32 GB

4.2 R libraries

There were several publicly accessible R libraries used for coding in this thesis.

Data manipulation and visualisation was mainly accomplished by the `tidyverse` system of packages (Wickham *et al.* 2019), including `dplyr` (Wickham *et al.*, 2020) and `ggolot2` (Wickham, 2016). Bellow, the R libraries that were crucial for the R scripting in this thesis are explained into detail. For the full list of R libraries used and teir versions see Supplemet 3.

4.2.1 flowCore

The `flowCore` package is a widely used R library for FC data import, processing and manipulation. An FCS file can be imported by `flowCore` functions in the workspace in a flowFrame object, a list type object storing metainformation from FCS TEXT segment in the `description` slot and the measured event data in the `exprs` slot as an annotated data frame. The `flowCore` functions mainly perform compensation, various transformations and “gating” tasks. (Hahne *et al.*, 2009)

In this thesis, `flowCore` was used mainly for extracting metainformation from the TEXT segment by the `read.FCSheader` function and verifying the FCS format by the `isFCSfile` function. Several `flowCore` functions (`read.FCS`, `compensate`) are implemented in an in house library `flowPeel`, that was also used in the thesis. (Ellis *et al.*, 2019)

4.2.2 FlowSOM

The `FlowSOM` is a library for building self-organising maps with a two-level clustering that uses an unsupervised method designed by Van Gassen et al. The two-level clustering means that individual events are organised into clusters and these clusters into metaclusters associating the events into populations. The individual clusters can be visualised by `FlowSOM` plotting functions in the form of a minimal spanning tree with star charts representing the marker values of each cluster. (Van Gassen *et al.*, 2015; Van Gassen *et al.*, 2019)

4.2.3 base library

Various functions from the base R libraries were used in this thesis. Notably, the `cutree` function was used to cut the resulting dendrograms of the Hierarchical Clustering Analysis (see chapter 2.5.1) into clusters (populations). The attributes of `cutree` function can be determined by either the `k` variable, setting the resulting number of clusters, or the `h` variable, setting the height of the dendrogram, where it should be cut. (R Core Team, 2019)

4.3 In house libraries

In addition to public R libraries, libraries that were created in or for the CLIP laboratory were used.

4.3.1 flowPeel

`flowPeel` is an *in house* R library frequently used in CLIP for FC data analysis. It contains several functions for opening and processing an FCS file. `openFCS` function uses `flowCore read.FCS` function to open an FCS file and extract it in a `flowFrame` object. This function also saves the spillover table and parameter names from several keywords. These results can be used as an input for the `flowCLZ` function, which compensates, logicle transforms and z-scores the FC data. `flowPeel` uses the `compensate flowCore` function for compensation and the base `scale` function for z- scoring. However, the logicle transform step uses a novel algorithm that automatically estimates the values of parameters of the *logicle* `transform` function for each file. The output of the function are flowFrame objects of results of all three steps (compensated, compensated and transformed and compensated, transformed and z-scored data) and a data frame object of the final result. The options of the `flowCLZ` function include the choice to compensate or not, choice of parameters to be excluded from the data processing and limiting the event count for processing to 10 thousand.

Both functions, `openFCS` and `flowCLZ`, are implemented in a wrapper function `processFCS`. The number of events saved can be set to all, or 10 thousand. The `processFCS` function can be also called on all FCS files in a directory by applying the `processFCSdir` function. However, this choice was not used for the purpose of this thesis, as the files selected for the processing are located in various directories.

The preprocessed data can be directly used in another `flowPeel` function, `hcaFCS`, to cluster the data by the hierarchical clustering method. It uses a novel algorithm that implements the MHCA (Mahalanobis Hierarchical Clustering Analysis) based on the computation of the Mahalanobis inter-cluster distance. (Fišer *et al.*, 2012)

4.3.2 flowPheno

The immunophenotypisation of the FC data is commonly performed based on subjective rules for the population characteristics (Dworzak *et al.*, 2018). The aim of the `flowPheno` library is to create an objective immunophenotypisation tool based on mathematical parameters of histogram distribution of individual fluorescence channels. The `flowPheno` R library will be made accessible to the public and the paper will be published soon. (Koblížek et al., manuscript in preparation)

In this work two functions from the `flowPheno` library that produce an immunophenotype report for selected population in the FCS dataset, `makeReport` and `getFDE` were used. The `getFDE` function produces a string specifying the population percentage and the positive and negative markers identified for the population based on the WHO classification of immunophenotype (Dworzak *et al.*, 2018). This string is very convenient for a simple and short report. An example of the FDE format report is:

FDE:65.05%;STRONG:CD10,19,NEGATIVE:CD20,27,34,38,44,45; ……. .

The function `makeReport` produces more detailed characteristics of the immunophenotype. For each marker, not only the WHO immunophenotypisation is determined, but also the population details and various metrics determining the value of positivity of a populations (SI, KLD, iKLD, KST, MI, MIb, MEM, COMsum, COMdif, WROCI). Most of these metrics are based on determining the distance between two population densities. The results are saved in a data frame format. For the description of all metrics that are used see Table 4.

Table 4 - Description of metrics used by flowPheno package to assess immunophenotype

Metrics	Description
WHO	Immunophenotypisation based on WHO immunophenotype classification • *Strong*, *Weak*, *Negative*
Detail	Description of the histogram distribution of the population: • Bright, Medium, Heterogeneous, Partially positive 2 - correspond with WHO Strong • Partially positive 1, Dim - correspond with WHO Weak • *Negative* - corresponds with WHO *Negative*
SI	Separation index
KLD	Kullback-Leibler divergence of positive population from negative population
iKLD	Inverted KLD
KST	Kolmogrov-Smirnov test of positive population and negative population
MI	Moment index of the population
MIb	Bilateral moment index
MEM	Marker Enrichment Modeling
COMsum	Centre of mass of positive and negative population sum
COMdif	Centre of mass of positive and negative population difference
WROCI	Weighted ROC (receiver operator characteristics) index of population

The input for these functions is a data frame of preprocessed data, a vector identifying each cell to a certain population, the population identifier of the population to be immunophenotyped and the population identifier of the population declared as negative. The negative population is used to set the negative value standard for each parameter. The `flowPheno` functions need correct negative population reference to be able to provide accurate results. For this reason, a

substantial part of the analytical protocol was dedicated to selection of a negative population with the lowest values from a series of clustering sets (see chapter 5.3).

4.4 FCS files in CLIP database

CLIP (Childhood Leukaemia Investigation Prague) is a diagnostic and research laboratory of the Second Faculty of Medicine of the Charles University and the Motol University Hospital. CLIP has two major departments for diagnostics of particularly leukaemic patients – the flow cytometry department and the molecular genetics department. The source of the samples diagnosed is e.g. blood, bone marrow or in some cases even tumour interstitial fluid. The patient data samples are either from patients that are suspected to have a certain disease (e.g. leukaemia) or they are follow up samples of the patients that were positively diagnosed. The majority of samples in the CLIP laboratory are of leukaemic patients. However, the diagnostics is also made for patients with primary immunodeficiencies, lymphomas and myelodysplastic syndromes. The leukemic patient files can be filtered based on acquired information about the files. This is the reason, why the first step of my thesis was creating a large database of metainformation obtained about all the FCS files. Some processed samples are from patients post bone marrow transplantation or from fluids from solid tumour biopsies.

The 355 220 flow cytometry standard files used in this thesis were acquired from the CLIP cytometry database. Most of the files were acquired during the years 2006 and 2019. For the analysis of metainformation of the FCS files see chapter 5.1. Most of the FCS files are saved in a current version of the FCS standard 3.1 FCS. This is very important for the data processing, because older FCS versions contain many structural differences that complicate the data analysis. The database contains both files from measured patient samples and some technical files, including testing and calibration files. The technical files are specific by a small file size and a small number of events and parameters measured.

The files in the CLIP database have a defined system of naming. The name contains the patient identification code, sample type (e.g. bone marrow or blood) and often the code of the cytometer that was used for measuring the files.

In the flow cytometry department, an optimal diagnostic panel of markers is chosen for each sample, depending on the suspected disease or the progression of the disease. For the diagnostic leukaemic samples, one of base leukaemic panels is measured firstly. Based on the results a panel designed for specific diagnosis was used, e.g. B cell lineage, T cell lineage or AML

oriented panel. Often, the samples are measured for more than one panel. The follow up leukaemic samples are taken at day 8, day 15, day 33 and week 12 of the treatment and the minimal residual disease (MRD) present in these samples is monitored. MRD represents the remaining blast cells remaining in blood or bone marrow during treatment. MRD level is a major indicator of the disease progression and based on it, the treatment is adjusted. The MRD panel is chosen based on the immunophenotype of diagnostic blast population of the patient.

The clinical database of CLIP includes all clinical information of each patient, including the information about processing and source of the individual samples, reported immunophenotypes in the FDE format and clinical comments. This database is relational and is managed by the PostgreSQL system.

4.5 Deviations from the FCS standard

Even though the FCS standard is defined and generally recognised, some differences in the usage of this standard may occur. For example the software of Cyan cytometer, Summit, tends to save all parameters two or three times, without any further information as to why. Furthermore, the software generates the `$PAR` keyword (the number of parameters) based on all the versions of the markers, which results in this number being double or triple of the correct value. This is a complication in identifying the individual markers, especially in the big FC data analysis.

Another challenge for the FC data processing is a human error. Even a small misspell in the name of the marker saved as a `$PnS` or `$PnN` keyword results in an inability to link this FCS file to other files that measure this marker. This error may seem insignificant when analysing one file only, but when trying to computationally analyse FCS files on a massive scale, it represents a vast problem and basically discards the said file from further analysis. Similarly, a change in marker names notation is inconvenient. However, these are easier to overcome as there are at least some rules to these changes.

5 Results

5.1 Processing of information about the FCS files

The aim of the thesis is creating an analytical protocol that is able to automatically process a large number of FCS files in terms of correct preprocessing of the data, identification of immunologically relevant cell populations and reporting of their immunophenotype in a clinically meainingful format.

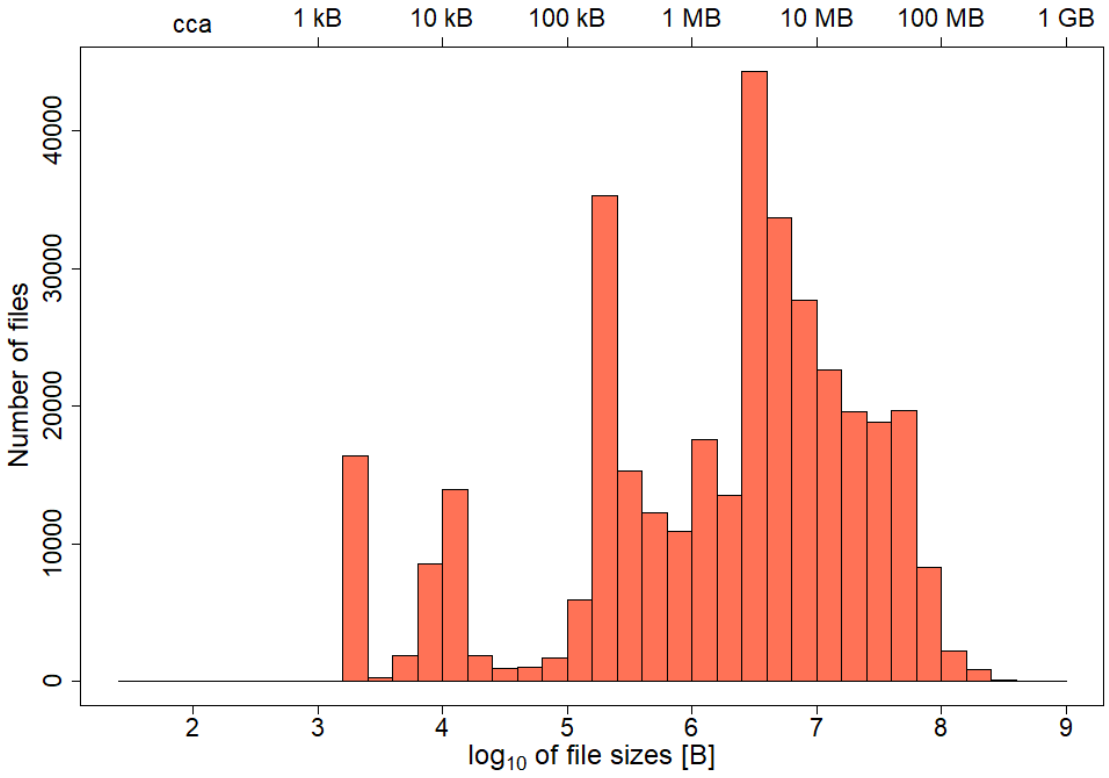
The first step of the thesis was acquiring and analysing available metadata relevant to the files. This was important for the selection FCS file for further processing. The main criteria for the file selection were the sample types, antibody panels and the FCS version.

The figures shown in this chapter represent data analysis of all FCS files stored in CLIP FC database on 4th November 2019. Such analysis allowed the selection of the files suitable for further analysis and also to obtain the information useful for further processing of the data. The information were part of the whole analytical protocol and in some cases served as an input for algorithmic decision making. For example, the cytometer and the acquisition software used for FCS file acquisition determines slight differences in the FCS standard and the structure of the file. This and other similar irregularities need to be considered when processing the data and to overcome these problems computationally. Therefore, it is essential to extract and analyse these information and store them in structural form.

The information about the files were acquired from two sources. The first source was the information present outside of the file, such as the information included in the name of the files (patient identification code, sample type), file creation date corresponding with the data acquisition and file sizes. The R base function `file.info` was used to extract the file acquisition dates and file sizes of the files and the `flowCore` function `isFCSfile` was used to confirm that the file was indeed an FCS file. Also, the md5sum of the file was saved using the function `md5sum` from the `tools` package for the future verification that the file was unchanged. The second source was the extraction of all keywords from TEXT segment of all FCS files. The keyword information include e.g. cytometer type, acquisition software, number of events (cells measured), number of parameters measured or names of markers. (Spidlen *et al.*, 2010) To load the TEXT segment of the files into the R environment, the `flowCore`

function `read.FCSheader`[1] was used. There are usually higher decades of keywords in the TEXT segment per file. However, only some of them are relevant for further analysis.

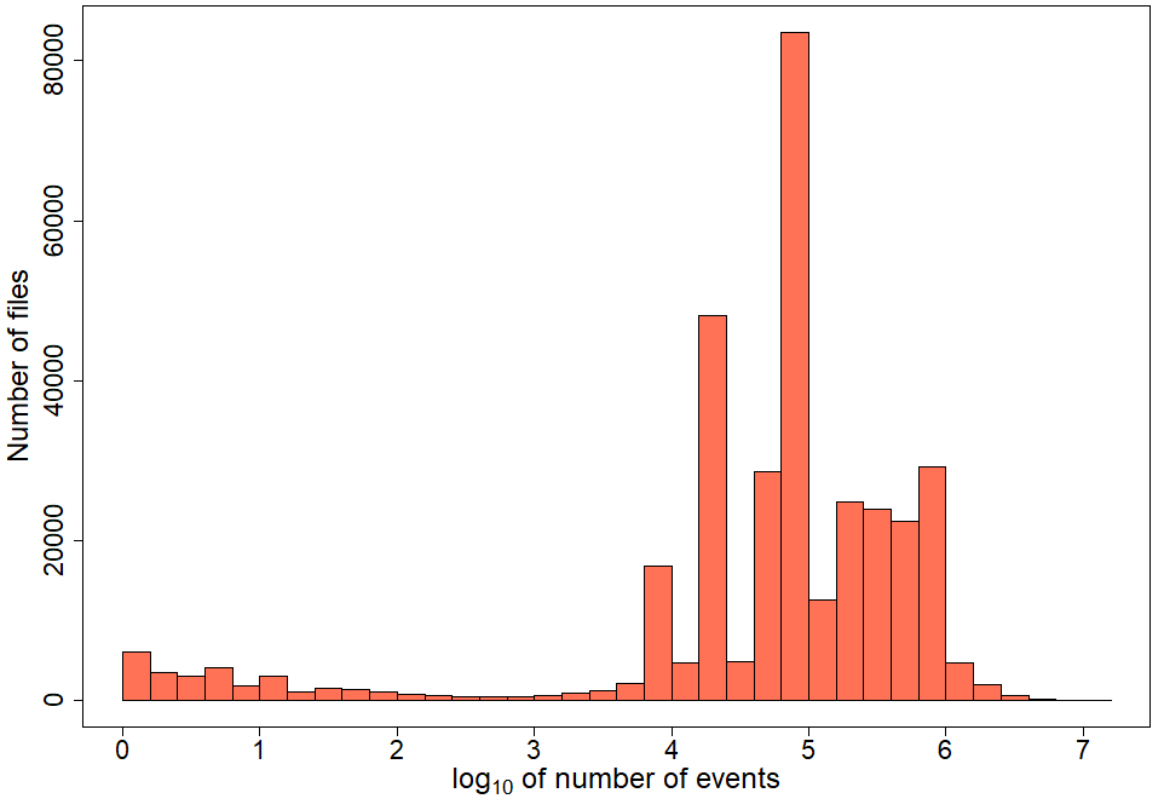
Fig. 4 - Distribution of file sizes of all FCS files in the CLIP database in Bytes shown on a log_{10} scale



Extraction of the sizes of FCS files in the database is a quick method to use for determining empty files. Also, the size of the file determines the time it takes to open them. Knowing the sizes is therefore important for further analysis. In Fig. 4, the distribution of file sizes of FCS files in the CLIP database is depicted In Bytes. Note, that the size of files is shown in a log_{10} scale. It is clear, that the most frequent size of files is around 5 MB. There is also a number of files with a size under 10kB. These files, however, are mostly testing files or files with low number of events and markers measured. Therefore they have no significance for further analysis.

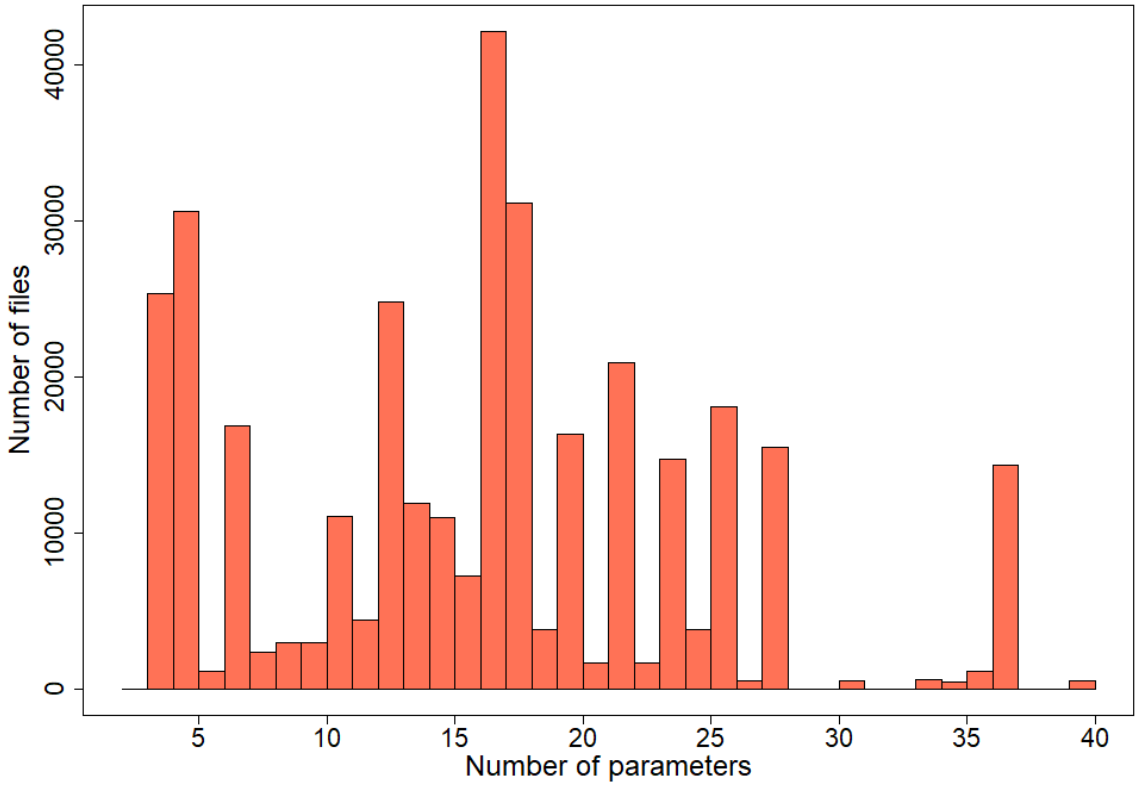
[1] Note, that the `read.FCSheader` function has not been named correctly, as the function does not open the HEADER segment of an FCS file, but the TEXT segment. While the HEADER segment defines the length and position of the TEXT and DATA segments, the TEXT segment stores the metainformation about the file.

Fig. 5 - Distribution of number of events of FCS files in the CLIP database shown on a log_{10} scale



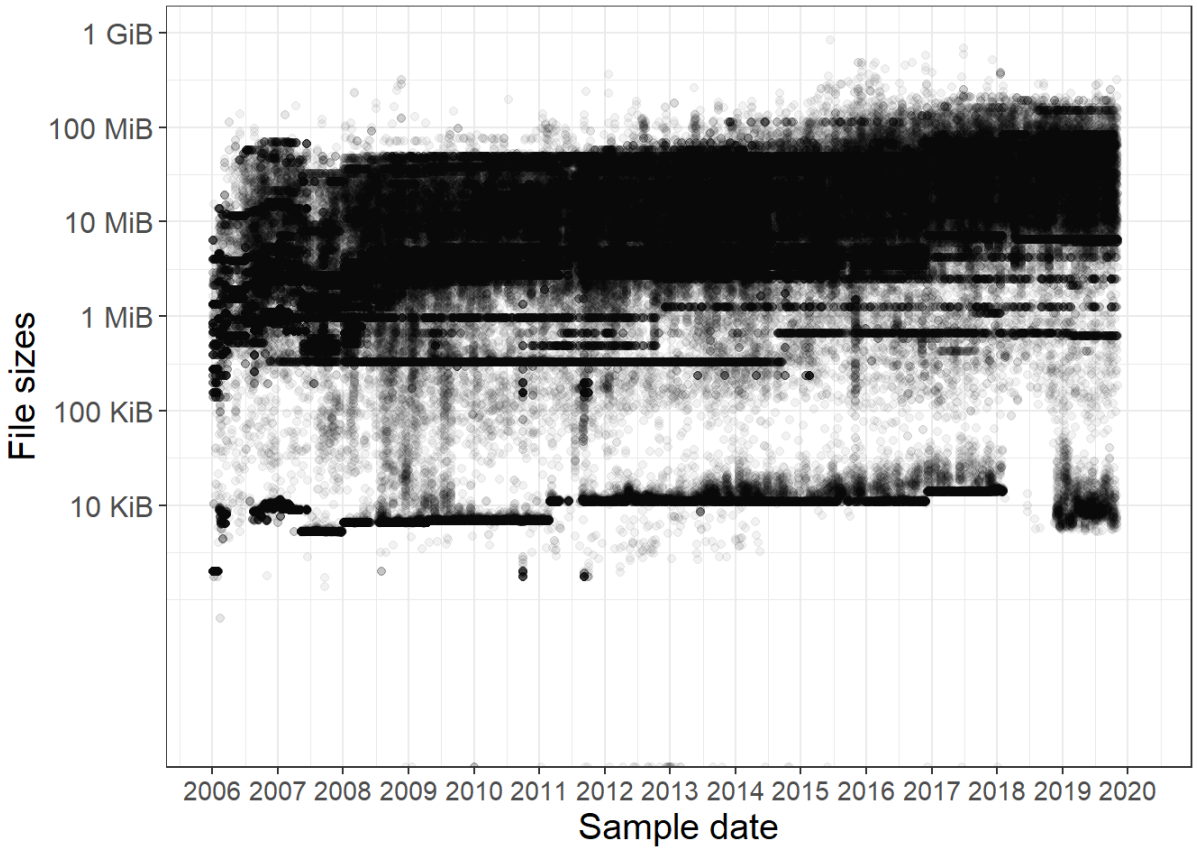
Another information giving us more insight into file complexity is the number of events (see Fig. 5). An event is anything the cytometer detects and stores as single measurement, but mostly one event corresponds to single cell measured. The data was extracted from the keyword `$TOT` saved in the TEXT segment of the FCS files. Usually, this number is set by the cytometrist before the experiment. The most common number of events to set is 100 000. Furthermore, there are several files with number of events smaller than hundreds. These are mostly testing and calibration files and are excluded from further analysis.

Fig. 6 - Distribution of number of parameters of FCS files in the CLIP database



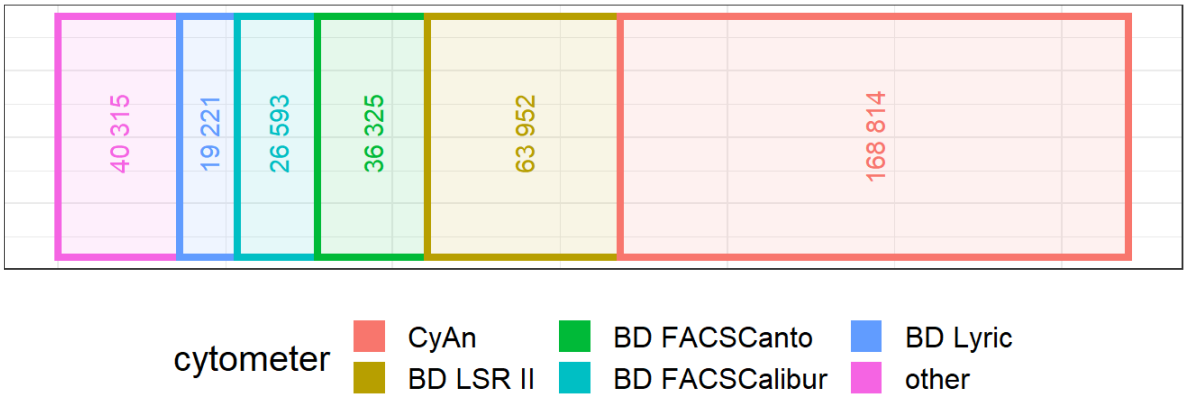
Similarly to the number of events, the number of parameters measured in FCS files was extracted form a keyword $PAR. The most common number of parameters measured for the FCS files in the CLIP database is 17 (see Fig. 6). Note, that the Cyan cytometer acquisition software stores measured data for each marker in two or three versions. For this reason the software multiplies the real number of parameters and saves the number into a $PAR keyword. Therefore the higher values of number of parameters can be considered as artificial.

Fig. 7- Development of sizes of FCS files in the CLIP database over time



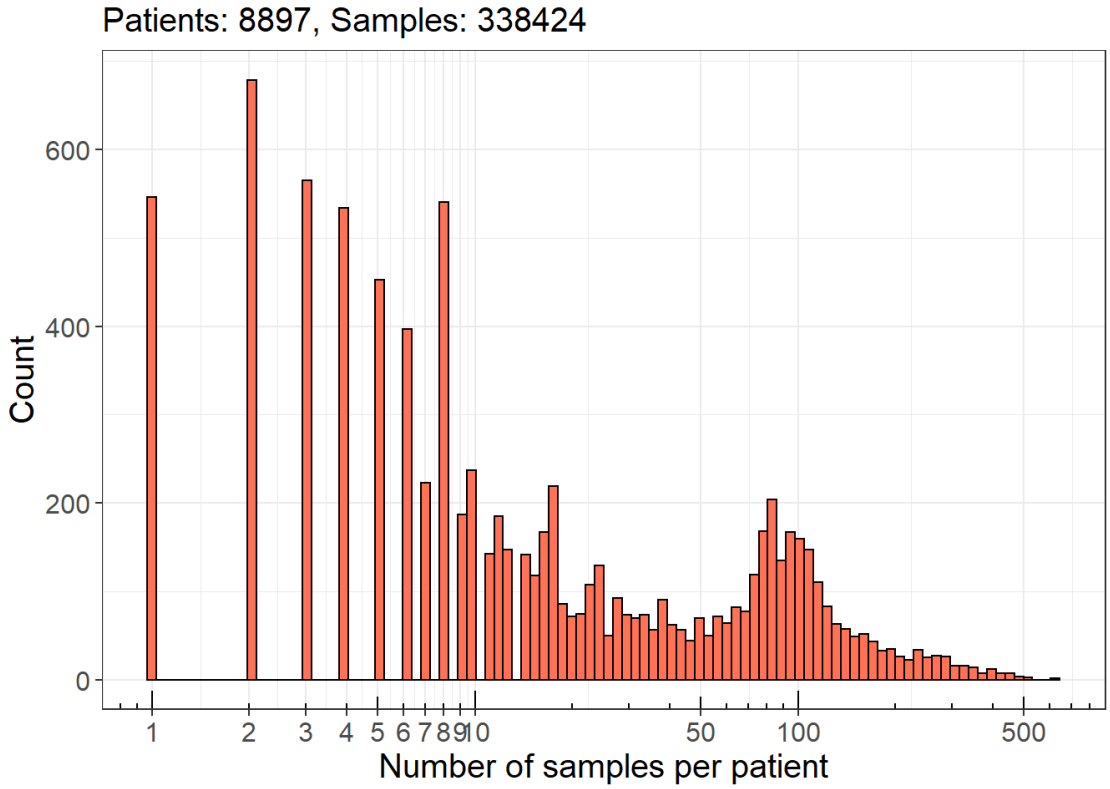
The dependency of sizes of FCS files in the CLIP database on the acquisition date is depicted in Fig. 7. It is clear, that the size of files slightly grows in time. This is probably due to growing number of parameters measured. The tendency to prefer multiparameter cytometry has been a trend in cytometry for years. The files of size around 10 kB are test and calibration files and have been used steadily, except for the year 2018. Fig. 7 also shows that acquisition dates of the available data in the CLIP database ranges from 2006. The FC leukaemia diagnostics has been performed for much longer than that in CLIP, but the older files are unavailable.

Fig. 8 - Distribution of FCS files per cytometer type in CLIP FC database. The group "other" includes cytometers BD FACSCelesta, BD FACSAria, Cell Lab QUANTA, CyTOF and SONY SP6800. On these cytometers, less than 16 000 files were measured in total.



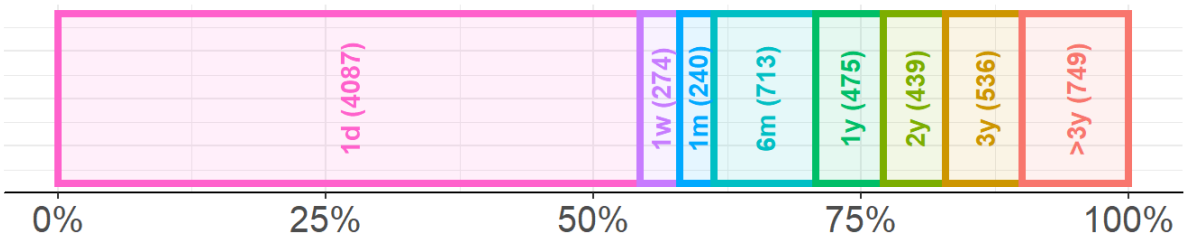
In Fig. 8, the distribution of FCS files in the CLIP database per cytometer is shown. The most used cytometer in the CLIP database is *Cyan ADP* manufactured by Dako, Glostrup, Denmark with 168 814 of the total 355 220 of FCS files. Files measured on modern digital instruments from BD Biosciences, San Jose, CA, USA form second big portion of the CLIP FC database – cca 64 thousand of files measured on *BD FACS LSR II*, cca 36 thousand on *BD FACS Canto II* and cca 19 thousand on *BD FACS Lyric*. Approximately 26 thousand files measured by *BD FACSCalibur* manufactured by BD Bioscience, San Jose, CA, USA cannot be used easily, because the software already compensates and transforms the data. There are also several cytometers that measured less than 16 thousand files in the CLIP laboratory in total. They include cytometers *BD FACSCelesta* and *BD FACSAria III* manufactured by BD Bioscience, San Jose, CA, USA, *Cell Lab QUANTA SC MPL* by Beckman Coulter, Brea, CA, USA, *CyTOF 2* by DVS Sciences/Fluidigm and *SONY SP6800* by SONY Biotechnology, Inc. In addition, approximately 37 thousands of files lacked the information about the acquisition cytometer.

Fig. 9- Distribution of number of FCS samples per patient in the CLIP FC database



To be able to analyse the data on a per-patient basis the information about which data were acquired form samples of same patients needed to be extracted. For a quick overview, the distribution of number of samples per patient is shown in Fig. 9. From approximately nine thousand of patients almost half of them had only up to 10 samples measured. This is strongly indicative of patients who were not diagnosed with ALL or other disease which would need the measurement of follow up samples for minimal residual disease monitoring. However, many patients have had hundreds of files measured over a long period of time. Approximately 10 thousand of files lacked the information to link the file to the patient.

Fig. 10 - Patient follow up duration time passed on the metainformation about FCS files in the CLIP database.. The percentage represents the percentage of patients whose samples were followed up to a certain period of time. The total number of patients is 8897.



In Fig. 10, the follow up duration time is depicted, for a more complete view of the patient data. The follow up was computed by associating all files of individual patients and determining the dates of acquisition of their first and last FCS file. From these dates, the length of follow up in days was counted. It is clear, that for more than 50 % of patients, the follow up was 1 day, suggesting they were diagnosed negative or removed from follow up for other reasons. For approximately 25 % of patients, the follow up was one or even several years. These were the patients whose minimal residual disease (MRD) was monitored.

5.2 Testing and choice of an appropriate data transformation

One of the very important steps of FC data preprocessing is the choice of correct data transformation. There are various transformations that can be used for the FC data (see chapter 2.4.2).

For this thesis several options for transformation were considered (book chapter in preparation, ISBN 978-80-88214-18-2, edited by Tomáš Kalina and Jan Zuna). Based on the literature several transformations were tested on FC data. First, the *logarithmic* transformation was ruled out because in our data the spread of values to negativity was too high for simple data shifting. The *arcsinh*, even though reported to be fast and without the need to estimate parameters (Finak *et al.*, 2010) proved to be too sensitive to parameter selection. This left only the *hyperlog, biexponential* a *logicle* transformations to consider. They all preformed relatively well. Among them the *biexponential* a *logicle* transformations can be easily parametrised automatically for each FCS file by available R functions, which is an advantage for big data processing. As the remaining two transformations were both good candidate, the *logicle* was chosen for this work to maintain compatibility with previous work.

5.3 Automatic detection of populations in files of a selected FC panel

The main result of this thesis is an analytical protocol which preprocesses any number of input FCS files, identifies their populations using clustering algorithms and determines which of the populations are potential populations of interest and which of the populations will serve as the referential "negative population". Note, that the negative population is term established in this thesis for describing a cluster with the lowest median values of markers in comparison to other populations and is not necessarily negative in all markers. In further text, it will be referred to as negative cluster or negative population. The negative cluster is used by the `flowPheno` R library functions to set the negative values of markers to determine the immunophenotype of all other identified populations (see chapter 5.4) (Koblížek et al., manuscript in preparation). The whole protocol of clustering and choice of clusters was saved in wrapper functions `choose_cluster_hca` for HCA clustering and `choose_cluster_fSOM` for FlowSOM clustering.

For the automatic processing, an altered `processFCS` function from the `flowPeel` *in house* R library is used. The `processFCS` includes loading of the FCS file and compensation, *logicle* transformation and z-scoring of the data. `flowPeel` uses the `flowCore` function `read.FCS` for opening and extracting the FCS files to a flowFrame object. From the flowFrame, the parameter names and the spillover table are saved. The compensation of the data is performed by a `flowCore` function, `compensate`, while using the extracted spilllover table. Logicle transformation is performed by an original in-built algorithm of `flowPeel` that includes an automatic evaluation of the logicle transformation parameters for each file. The z-scoring is performed by the `scale` function of the R base library. The output of the `processFCS` function is a flowFrame of the result of each step (compensation, transformation and z-scoring) and a data frame of the final result. For this thesis, the `processFCS` function was slightly rescripted in terms of outputs. Instead of saving the output in an RData file in the source folder, the function now directly loads the results as a list object in the R workspace. The reasons for this rescripting were the complications of the data organisation, as the processed data in RData files are saved in the original folder of the FCS file. Furthermore, in this approach the data would need to be saved and loaded and this process would take too much computational capacity. The output list includes a slot for each FCS file set as input and inside the compensated, transformed and z-scored data are stored in the dataframe and flowFrame format. These preprocessed data were directly used for population clustering. The only manual input of the preprocessing step is the setting of columns to exclude from the analysis.

For the automatic detection of populations, two types of clustering algorithms were used: hierarchical clustering analysis with Mahalanobis distance measure (Fišer et al., 2012) and FlowSOM (Van Gassen et al., 2015). The reason for choosing to use FlowSOM in this thesis was, that multiple comparison studies recommend FlowSOM for very high precision and reproducibility on multiparameter FC data. At the same time, it maintains significantly faster runtimes than other clustering and dimensionality reduction methods (e.g. flowMeans, X-shift, immunoClust, PhenoGraph or ClusterX). (Weber and Robinson, 2016b; Liu et al., 2019) Hierarchical clustering analysis was chosen for good performance for ALL blast population detection, the high precision and the ability to distinguish overlapping populations very well. (Fišer *et al.*, 2012). The details about the clustering methods are explained in chapters 2.5.1 and 2.5.2.

It was assumed, that the antibody panel used for the samples was designed to clearly separate the populations in the samples and also, that the most relevant population (i.e. blastic) would be one of the major, relatively homogenous populations in the diagnostic FCS files.

However, the background populations can be much smaller, but also significantly different from both the major populations and each other. This can severely complicate the clustering process, as the number of cluster is set and there is a preference to cluster primarily the significantly different populations. This results in one large cluster including all the populations of interests and many small clusters containing the outliers. Such population should be a relatively easy target for a clustering method. However other cell populations, namely negative populations, needed for immunophenotype construction had to be correctly detected as well. Therefore the analytical protocol had to be robust in regards of differences in population sizes and homogeneity distributions. For example, in some samples there can be a number of outliers (artificial or biologically relevant) which would form separate clusters and need to be excluded. On the other hand, the candidate negative population can be relatively small and still needs to be identified and kept for further use. Therefore the use of programming was leveraged to produce a number of clustering sets each with a different number of clusters from a set range. Also, the heuristics to choose from the variation of clusters for each dataset was implemented. The range of the output clusters can be set in a variable of the `choose_cluster_hca` or `choose_cluster_fSOM` wrapper functions and iteratively alter either the grid setting variables in `flowSOM`, `xdim` and `ydim`, or the `k` variable for the `cutree` function used to create the desired number of clusters from HCA output. The output of hierarchical clustering is a dendrogram, which is cut into desired number of individual clusters by the `cutree` function.

The only manual entry to `choose_cluster_hca` or `choose_cluster_fSOM` function is setting of the numbers of scatter columns.

From these clusters, only the populations of a relevant size are chosen for analysis. The minimal population size was set as 3 % in this thesis, but the threshold can be set differently as a parameter of the function. From the chosen clusters, the most probable negative cluster is chosen based on the following heuristics:

- For each cluster in each iteration of clustering, the median values are counted for each marker (see Table 5 A and B showing intermediate results for 1 file)
- Because it is possible, that there will not be a cluster with the lowest median value for all markers, an evaluation process was set. Each marker median values is scored from cluster with lowest (score = 1) to highest value. Then, for each cluster, the sum of all scores is saved and the cluster with the lowest sum is determined as the negative population for this clustering (see Table 5 C showing intermediate results for 1 file).
- If there are multiple populations with the same sum of scores, the population with the lower number of events is chosen, because this population has a higher probability of being homogenous. Moreover, an additional condition was added to ensure, that large populations will not be chosen as negative. These populations are usually blastic in case of leukaemic diagnostic files, or at least form relevant immune cells population that should not be excluded from the analysis. A threshold for the population size was set to be 25 %. If the cluster with the lowest score is larger than 25 % of the dataset, the population with the second lowest score is chosen instead.
- Then, from these negative populations of each iteration of clustering the most suitable negative population is selected. Firstly, if the same population is selected by more iterations of clustering, the cluster from the iteration of clustering with the lowest number of clusters is chosen (see Table 5 D and E showing intermediate results for 1 file). Again, the same evaluating system is used – scoring of median values for each marker, counting sum of the scores for each cluster and selecting the population with the lowest sum (see Table 5 F-H showing intermediate results for 1 file).
- If there are two or more clusters resulting in the same sum of score, the iteration of clustering of the lower number of clusters is preferred.

Table 5 - Demonstration of the negative cluster choice heuristics of HCA iteration of clustering. k value in tables D-F is parameter of cutree function and is equivalent to the number of clusters of the iteration of clustering.

A – Table of median immunophenotype values of clusters of 16-cluster iteration of clustering and their size.

B – Clusters of 16-cluster iteration of clustering reduced to clusters with the size of at least 3 % of events, their median immuophenotype values and size.

C - Clusters of 16-cluster iteration of clustering reduced to clusters with the size of at least 3 % of events, their scored median immunophenotype values from lowest to highest value, their size and the sum of scores for each cluster. Cluster number 2 is determined as the negative cluster.

D – Selected negative clusters of all HCA iterations of clustering (with at least two clusters bigger than 3 % of events), their median immunophenotype values and sizes.

E – Reduced number of selected negative clusters in D.

F – Selected negative clusters in E, their scored median immunophenotype values from lowest to highest, their size and the sum of scores for each cluster. Cluster number 16 is determined as the negative cluster.

G – Selected negative cluster and the number of clusters of the iteration of clustering.

H – Table of median immunophenotype clusters of the selected iteration of clustering (32 clusters) and their size.

Steps A-C is performed for each iteration of clustering.

A

cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	# events
1	0.779	1.014	-0.664	-0.844	-0.740	0.254	-0.375	0.808	1154
2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
3	0.821	-1.226	-0.625	1.550	-0.707	-0.101	2.927	0.723	169
4	-0.690	-1.033	-0.515	-0.853	-0.173	-1.712	1.047	-0.387	2
5	0.625	-0.161	0.710	-0.553	-0.135	-1.711	3.014	1.205	1
6	-0.765	0.518	1.425	1.185	1.594	1.057	-0.312	-0.869	3
7	0.595	-0.465	0.430	-0.714	0.132	-1.794	2.303	0.870	3
8	0.209	-0.449	0.394	-0.827	-0.121	-0.762	1.137	0.977	2
9	0.622	1.614	0.196	1.707	-1.526	2.979	-0.012	0.662	1
10	2.176	-0.857	1.376	1.416	1.019	0.284	3.404	0.299	1
11	0.504	-0.009	0.863	-0.540	0.304	-0.966	2.716	1.261	1
12	3.387	3.093	2.313	2.598	1.969	3.704	4.592	2.862	1
13	-1.006	-1.159	-0.514	-0.854	0.071	-1.584	1.044	-1.042	1
14	0.113	-0.184	1.962	-0.976	0.800	-0.228	1.828	0.536	2
15	1.458	1.424	-0.607	-0.587	-0.860	2.315	-0.259	1.113	1
16	-0.791	-1.055	-0.489	-1.037	-0.344	0.573	-0.309	0.039	1

B

cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	# events
1	0.779	1.014	-0.664	-0.844	-0.740	0.254	-0.375	0.808	1154
2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
3	0.821	-1.226	-0.625	1.550	-0.707	-0.101	2.927	0.723	169

C

cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	# events	score
1	2	3	1	1	1	3	2	3	1154	16
2	1	2	3	2	3	2	1	1	657	15
3	3	1	2	3	2	1	3	2	169	17

D

k	cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	# events
13	2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
14	2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
15	2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
16	2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
17	2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
18	2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
19	1	0.758	1.043	-0.679	-0.847	-0.751	0.332	-0.387	0.818	1078
20	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
21	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
22	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
23	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
24	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
25	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
26	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
27	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
28	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
29	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
30	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
31	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
32	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
33	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
34	2	-1.093	-1.159	-0.506	-0.817	-0.675	-1.604	-0.148	-1.338	145
35	2	-1.093	-1.159	-0.506	-0.817	-0.675	-1.604	-0.148	-1.338	145

E

k	cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	# events
13	2	-1.192	-0.302	1.373	1.003	1.381	0.004	-0.377	-1.227	657
19	1	0.758	1.043	-0.679	-0.847	-0.751	0.332	-0.387	0.818	1078
20	2	-1.090	-1.158	-0.506	-0.817	-0.671	-1.604	-0.148	-1.337	147
34	2	-1.093	-1.159	-0.506	-0.817	-0.675	-1.604	-0.148	-1.338	145

F

k	cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	# events
13	2	1	3	4.0	4.0	4	3.0	2.0	3	657
19	1	4	4	1.0	1.0	1	4.0	1.0	4	1078
20	2	3	2	2.5	2.5	3	1.5	3.5	2	147
34	2	2	1	2.5	2.5	2	1.5	3.5	1	145

G

chosen k	34
negative cluster	2

H

cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	# events
1	0.815	1.089	-0.730	-0.853	-0.759	0.481	-0.402	0.844	917
2	-1.093	-1.159	-0.506	-0.817	-0.675	-1.604	-0.148	-1.338	145
3	-1.192	-0.308	1.371	0.998	1.380	-0.018	-0.379	-1.231	648
4	0.826	-1.229	-0.634	1.550	-0.717	-0.100	2.927	0.722	161

The output of this function is a large list, with a slot for each file that was set to be processed. Inside, the iteration of clustering that includes the chosen negative cluster is saved - for FlowSOM the whole FlowSOM object, for HCA an object identifying each cell to a cluster. Along with the clustering, the number of the clusters of this clustering, the identification of the negative cluster and the identification of clusters of a relevant size are saved. For the purposes of control, intermediate results are also saved. Also, the compensated, logicle transformed and z-scored data are saved as data frames and flowFrames. These data are used for the immunophenotyping in chapeter 5.4.

To demonstrate the application of this protocol, several diagnostic files were chosen based on the following criteria:

- All cca 350 000 FC files were grouped by the patient identification code and the length of follow up was determined based on the first and the last acquisition date.
- Only the files of patients with the follow up longer than 1 day were selected. These patients were probably diagnosed positive and there is a chance of including blastic populations.
- From this selection, the files with the oldest acquisition date were chosen for each patient. These files are the diagnostic samples and have the highest probability of including larger populations of blastic cells. The FCS files measured later include data of samples during or after treatment and the blastic cells are reduced.
- Finally, from these files, those which measured an example leukaemic diagnostic panel CD44 CD27 CD34 CD19 CD10 CD38 CD20 CD45 were chosen. Practically, this was done by filtering the `$`TUBE NAME`` keyword of the files to "44 27 34 19 10 38 20 45". Alternatively, the keyword `$`PANEL`` or the separate keywords for marker names could be used.
- This resulted in the selection of 19 diagnostic files.

5.3.1 HCA – hierarchical clustering

After the preprocessing step of compensation, logicle transformation and z-scoring the raw data, the output was saved in the data frame format. This output is used for both HCA and FlowSOM clustering.

For the hierarchical clustering, the `hcaFCS` function from R `flowPeel` library was used. For the reason of the lengthy time of processing, only 10 000 events were clustered for each file. This reduced the time of processing to approximately 80 minutes per file.

The wrapper function for automatic population detection was named `choose_cluster_hca`. It includes cutting of the general hierarchical dendrograms of selected files and the identification of an optimal cut producing clustering with an optimal negative cluster. The HCA dendrogram is cut by the `cutree` base R function to produce clustering of a number of clusters set in `k` variable. This dendrogram cutting is performed several times to produce clustering iterations for each `k` in range that was set in the `cluster_range` variable of the `choose_cluster_hca` function.

Next, the negative cluster is chosen based on the heuristics described above.

The median values of clusters chosen as optimal by the `choose_cluster_hca` function for `cluster_range` 5-100 for all 19 selected diagnostic files are shown in Table 6. The sizes of clusters and the negative cluster used as reference for the immunophenotypisation in chapter 5.4 are also shown.

Table 6 - The selected HCA iteration of clustering of diagnostic files 1-19 selected by choose_cluster_hca function with range of clusters set to 5-100, median values of their markers and their size. Only clusters of size bigger than 3 % are shown and are used for immunophenotypisation in chapter 5.4. For each file, the total size and total number of clusters is shown. The number of clusters is equivalent to the k variable of the cutree function chosen as optimal. The negative clusters used as a reference for immunophenotypisation is highlighted in blue.

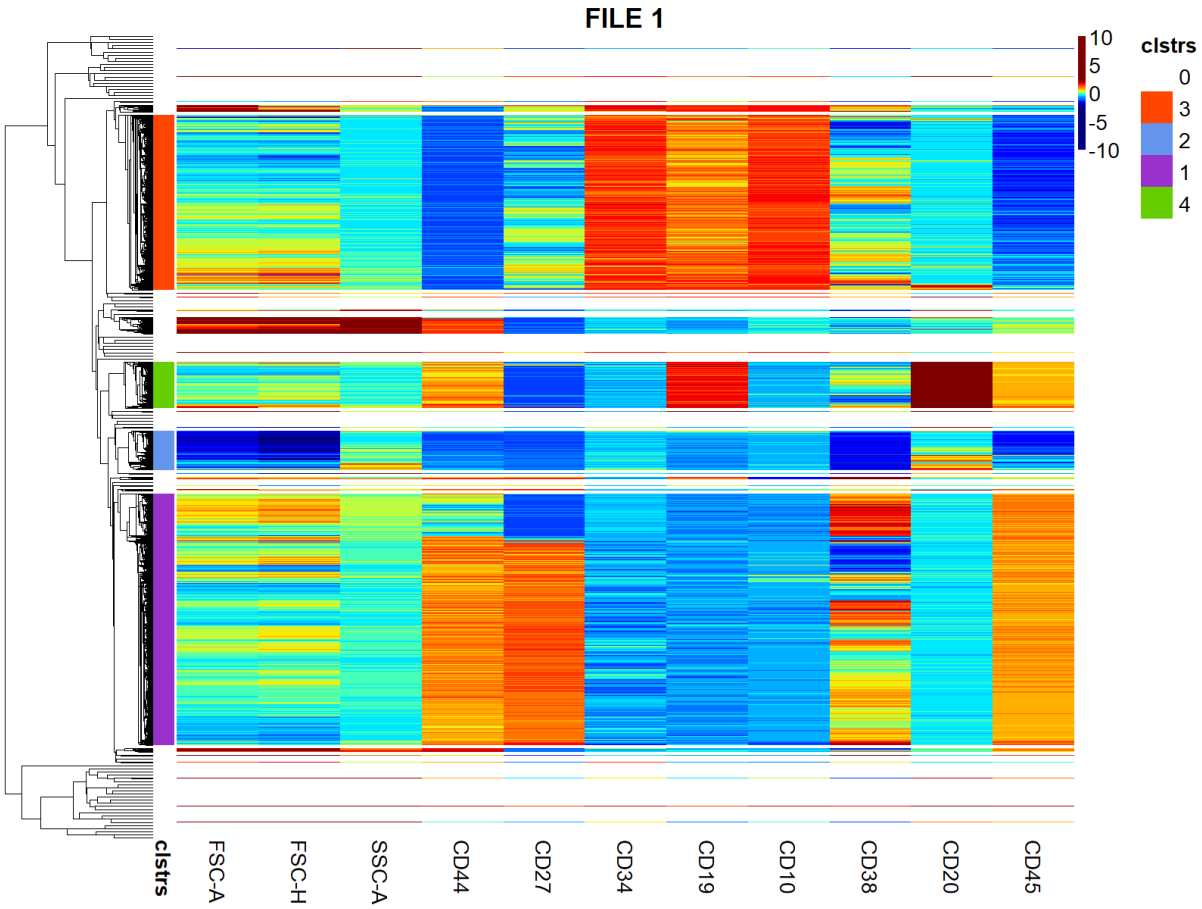
FILE 1 – 10 000 cells									
Number of clusters - 82									
cluster	**median CD44**	**median CD27**	**median CD34**	**median CD19**	**median CD10**	**median CD38**	**median CD20**	**median CD45**	**number of cells**
1	0.825	1.097	-0.732	-0.863	-0.771	0.464	-0.400	0.842	4518
2	-1.090	-1.156	-0.519	-0.863	-0.672	-1.612	-0.260	-1.321	701
3	-1.184	-0.302	1.361	0.987	1.351	-0.048	-0.395	-1.239	3148
4	0.798	-1.232	-0.663	1.522	-0.734	-0.169	2.872	0.710	839
FILE 2 – 10 000 cells									
Number of clusters - 63									
1	-0.465	-0.274	0.360	0.215	0.311	-0.191	-0.214	-0.438	8737
4	1.845	4.141	-2.825	-2.307	-2.023	-0.106	-0.172	2.614	478
FILE 3 – 10 000 cells									
Number of clusters - 21									
1	0.101	-0.307	0.462	0.319	0.378	0.289	-0.252	-0.316	8957
2	0.026	3.361	-1.531	-2.752	-2.670	-2.187	0.553	2.851	878
FILE 4 – 10 000 cells									
Number of clusters - 17									
1	-0.244	-0.291	0.394	0.313	0.353	0.305	-0.056	-0.271	8618
3	-0.711	-0.393	-1.983	-2.934	-2.789	-2.295	-1.147	-1.691	367
5	1.077	4.107	-2.368	-2.866	-2.773	-1.871	-1.170	2.324	430
FILE 5									
Number of clusters - 95									
3	-0.489	-0.215	0.547	0.537	0.642	0.118	-0.370	-0.296	7084
4	2.218	-0.168	-1.766	-1.869	-1.242	-1.055	0.504	0.880	888
5	-0.695	-0.213	-1.690	-1.890	-1.377	-1.191	0.702	-1.204	953
FILE 6 – 10 000 cells									
Number of clusters - 85									
1	-1.187	-0.200	-0.379	-1.990	-0.683	-2.485	0.460	-2.573	685
3	0.209	-0.198	-0.236	0.487	-0.242	0.446	-0.263	-0.008	7990
6	-0.052	-0.046	-0.097	-1.714	0.867	-1.388	0.872	0.106	507

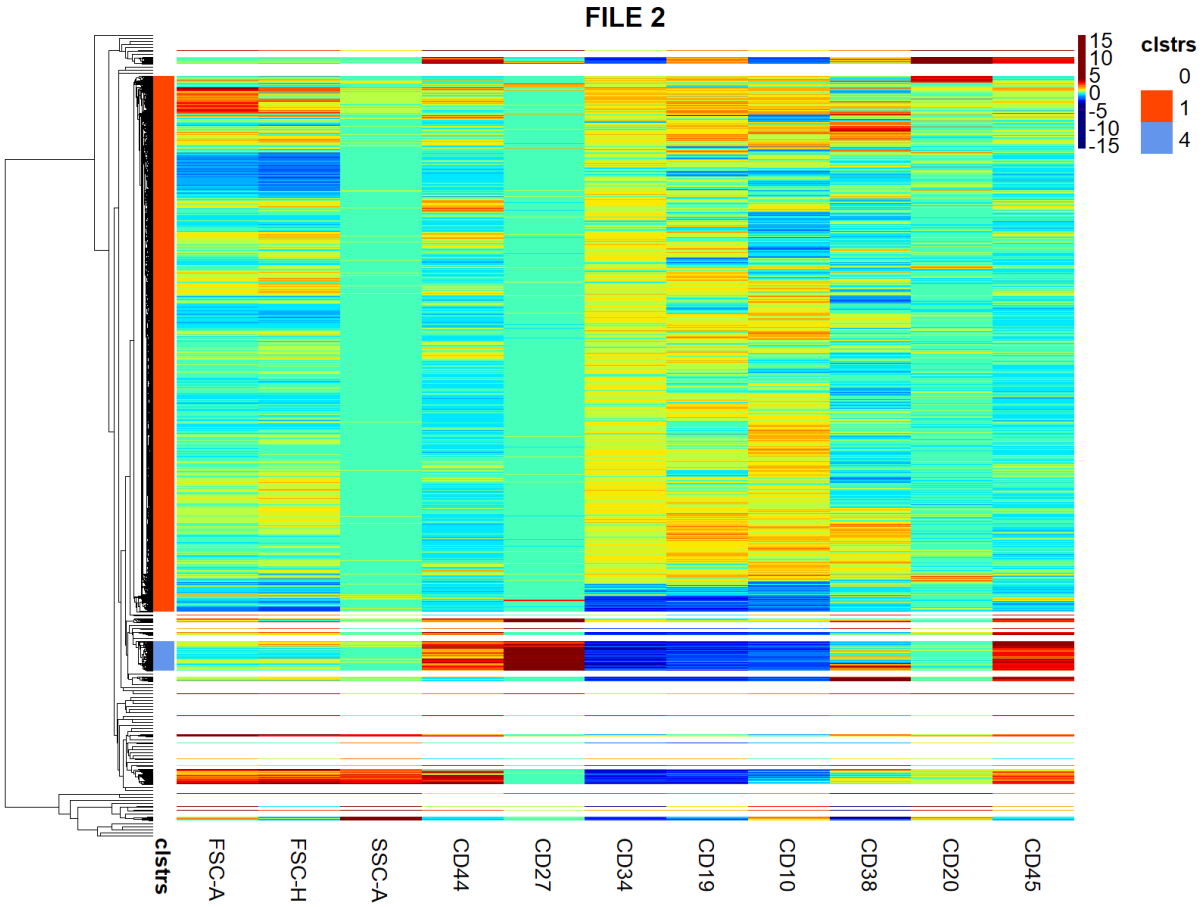
FILE 7 – 10 000 cells									
Number of clusters - 55									
cluster	median CD44	median CD27	median CD34	median CD19	median CD10	median CD38	median CD20	median CD45	number of cells
1	-0.364	-0.255	0.391	0.308	0.406	0.235	-0.150	-0.328	9059
2	2.642	4.282	-1.601	-1.136	-1.572	-0.724	-0.133	3.399	434
FILE 8 – 10 000 cells									
Number of clusters - 10									
1	0.081	-0.224	0.264	0.221	0.275	0.203	-0.234	-0.218	9419
2	-1.547	-0.230	-3.056	-4.117	-3.754	-2.698	-0.940	0.329	339
FILE 9 – 10 000 cells									
Number of clusters - 90									
1	-0.581	-0.494	-0.108	0.593	0.727	0.583	-0.401	-0.784	5949
2	0.841	2.159	-0.314	-1.475	-1.383	-0.348	-0.398	1.482	1684
3	0.861	-0.564	0.407	-1.586	-0.380	-1.368	-0.211	0.649	687
4	0.753	-0.677	-0.245	0.744	-1.250	-0.911	2.956	1.339	761
11	-1.320	-0.642	0.175	-1.381	-1.282	-1.803	-0.306	-0.818	318
FILE 10 – 10 000 cells									
Number of clusters - 75									
1	-0.194	-0.415	-1.109	-1.395	1.234	-1.321	0.271	-0.400	1332
2	-0.152	-0.483	0.736	0.725	-0.506	0.569	-0.390	-0.273	6264
3	0.949	2.287	-1.479	-1.340	-0.513	-0.505	-0.529	1.511	1638
5	-1.813	-0.464	-1.232	-1.475	-0.440	-1.971	-0.135	-2.172	335
FILE 11 – 10 000 cells									
Number of clusters - 76									
1	-0.391	-0.300	-0.259	0.377	0.416	0.358	-0.268	-0.380	8339
3	-0.262	-0.227	-0.540	-2.489	-2.537	-2.468	-0.617	0.048	548
5	2.866	3.650	-0.916	-2.246	-2.360	-0.548	-0.855	2.958	523
FILE 12 – 10 000 cells									
Number of clusters - 54									
5	-0.573	-1.126	-0.853	-2.485	-2.425	-2.522	-0.174	-0.618	633
6	-0.358	-0.185	-0.005	0.388	0.446	0.418	-0.320	-0.291	8368
8	2.873	2.770	-1.185	-2.319	-2.457	-1.684	-0.496	2.741	571
FILE 13 – 10 000 cells									
Number of clusters - 81									
1	-0.378	-0.528	0.602	0.490	0.652	0.165	-0.419	-0.591	6583
2	0.957	2.028	-1.599	-1.734	-1.518	-0.142	-0.518	1.500	1876
4	1.300	-0.470	-1.365	0.775	-1.557	-0.441	3.072	1.382	442
6	0.550	-0.449	0.955	0.864	1.003	1.006	0.312	0.210	374
FILE 14 – 10 000 cells									
Number of clusters - 90									
1	-0.393	-0.307	0.438	0.431	0.43	0.431	-0.372	-0.291	8388
2	1.412	-0.909	-1.468	-2.329	-2.71	-2.141	-0.196	-0.806	452
3	3.019	3.118	-1.818	-2.199	-2.35	-1.495	-0.319	2.785	315

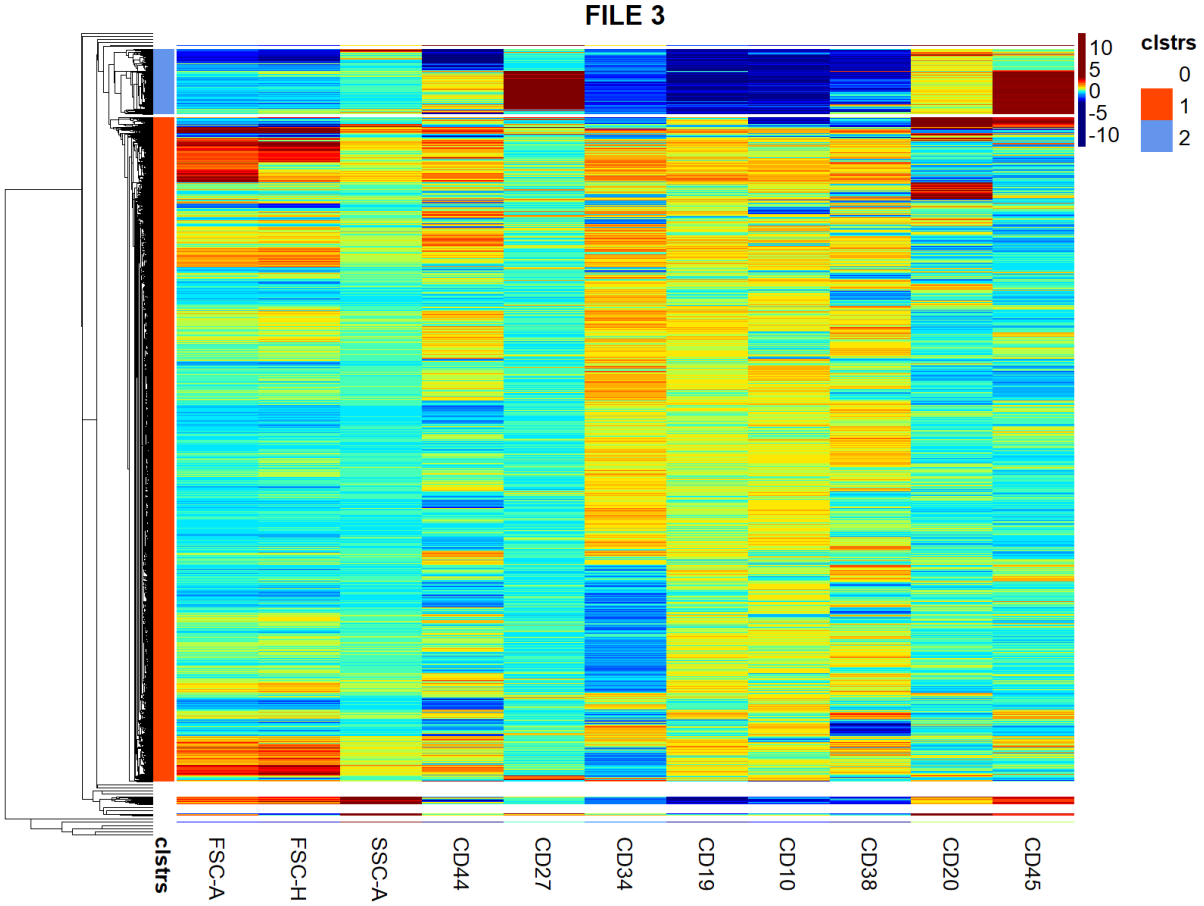
FILE 15 – 10 000 cells									
Number of clusters - 80									
cluster	**median CD44**	**median CD27**	**median CD34**	**median CD19**	**median CD10**	**median CD38**	**median CD20**	**median CD45**	**number of cells**
1	-0.539	-0.163	0.627	0.535	0.613	0.127	-0.277	-0.585	6882
2	1.871	1.863	-1.367	-1.636	-1.577	0.520	-0.325	1.868	1463
3	-0.438	-1.143	-1.178	-1.699	-1.507	-1.703	-0.088	-0.701	681
FILE 16 – 10 000 cells									
Number of clusters - 94									
2	-0.476	-0.377	-0.086	0.236	0.398	0.396	-0.432	-0.452	7965
4	2.264	3.529	-0.390	-2.548	-2.399	-2.144	-0.744	2.326	376
9	2.074	-0.540	-0.277	0.406	-2.206	-1.797	2.508	2.037	513
FILE 17 – 10 000 cells									
Number of clusters - 83									
1	-0.545	0.277	0.415	0.379	0.449	0.119	-0.216	-0.441	7274
3	-0.057	0.921	0.772	0.869	0.798	1.010	0.009	-0.030	753
5	0.701	-1.839	-1.981	-2.090	-1.578	-0.469	0.399	1.403	320
6	-0.211	-1.912	-2.207	-2.212	-2.206	-1.659	-0.227	-0.662	588
7	2.318	1.082	-2.157	-2.074	-2.149	0.156	-0.158	2.879	483
FILE 18 – 10 000 cells									
Number of clusters - 63									
1	-0.429	-0.319	-0.207	-0.296	-0.437	-0.442	-0.291	-0.498	7879
2	1.338	3.998	-1.288	-0.403	-0.788	0.635	-0.843	1.680	374
6	1.712	-0.392	-0.210	-0.327	-0.366	2.028	-0.564	1.281	431
7	0.669	1.396	2.311	-0.037	1.039	0.077	1.838	1.684	309
FILE 19 – 10 000 cells									
Number of clusters - 85									
1	-0.461	-0.304	-0.068	0.396	0.414	0.278	-0.297	-0.514	8036
2	2.089	4.028	-1.051	-2.128	-2.299	-0.168	-0.303	2.515	458
3	1.472	-0.271	0.550	-2.273	-1.707	-0.830	0.146	1.449	432
4	0.180	-0.421	-0.049	-2.408	-2.290	-1.697	-0.349	-0.533	326

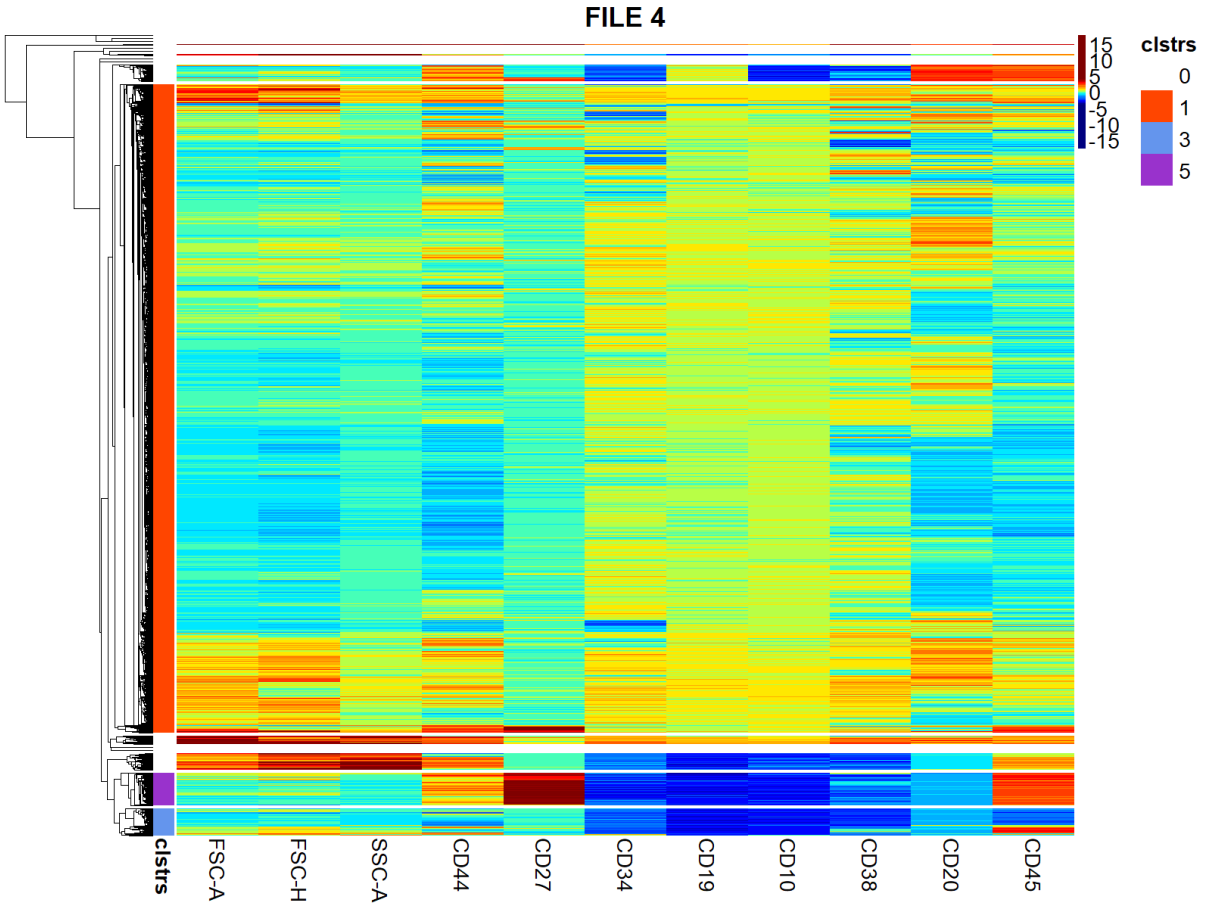
The clustering result for all files were depicted in the matrices of scatterplots and in the heatmaps with joined dendrograms for parameters FSC-A (forward scatter – area), FSC-H (forward scatter – height), SSC-A (side scatter – area), CD44, CD27, CD34, CD19, CD10, CD38, CD20 and CD45. The resulting HCA iteration of clusters with size over 3 % were highlighted in colour – red for blastic population, blue for the negative population used as reference for immunophenotyping and purple, green and yellow for remaining populations. The blastic B-ALL populations were selected manually based on the results of automated immunophenotypisation in chapter 5.4. The heatmap visualisation for diagnostic files 1-19 are shown in Fig. 11. The matrix of scatterplots of file 1 is shown in the Fig. 12. The matrices of scatterplots of files 2-19 are shown in the Supplement 1.

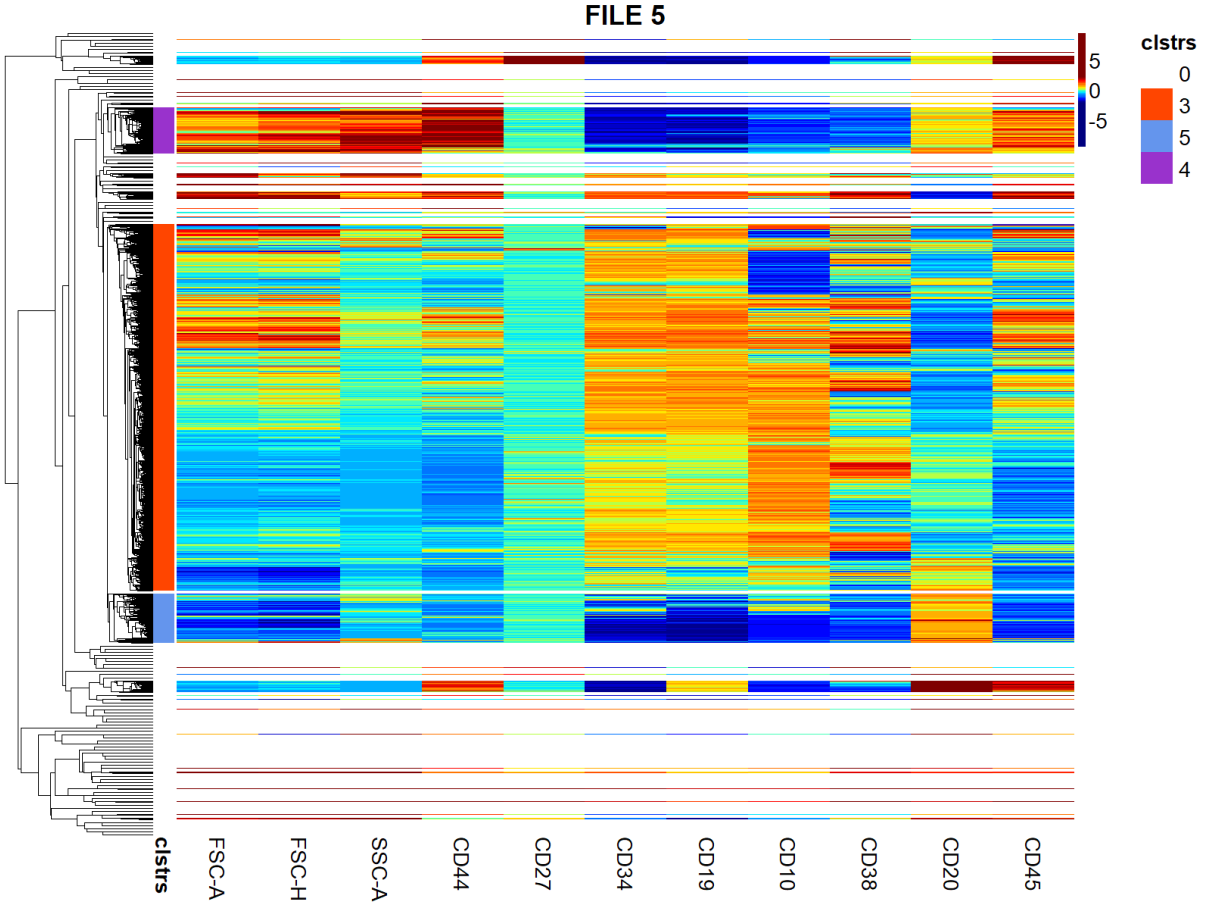
Fig. 11 - Heatmap visualisation with joined dendrogram of resulting HCA clustering of diagnostic files 1-19 selected by the choose_cluster_hca function with range of clusters set to 5-100. The populations of this clustering of size bigger than 3 % are highlighted by colours on the side of the heatmap - red for blastic population, blue for referential negative population and purple, green and yellow for other populations. The blastic B-ALL populations were selected manually based on the results of automated immunophenotypisation in chapter 5.4. In file 18, no blastic population was identified. The colours of the populations correspond to the ones in matrices of scatterplots (see Fig. 12 and Supplement 1).

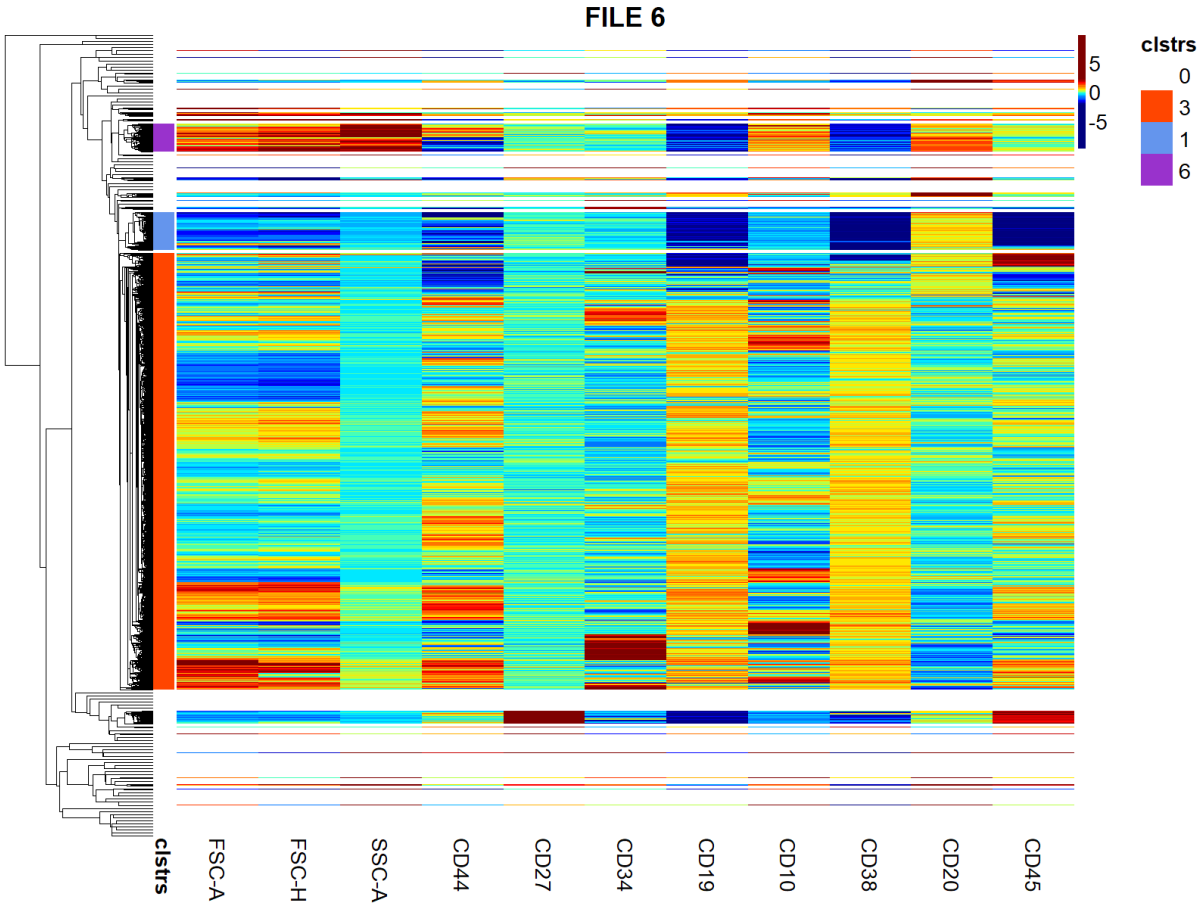


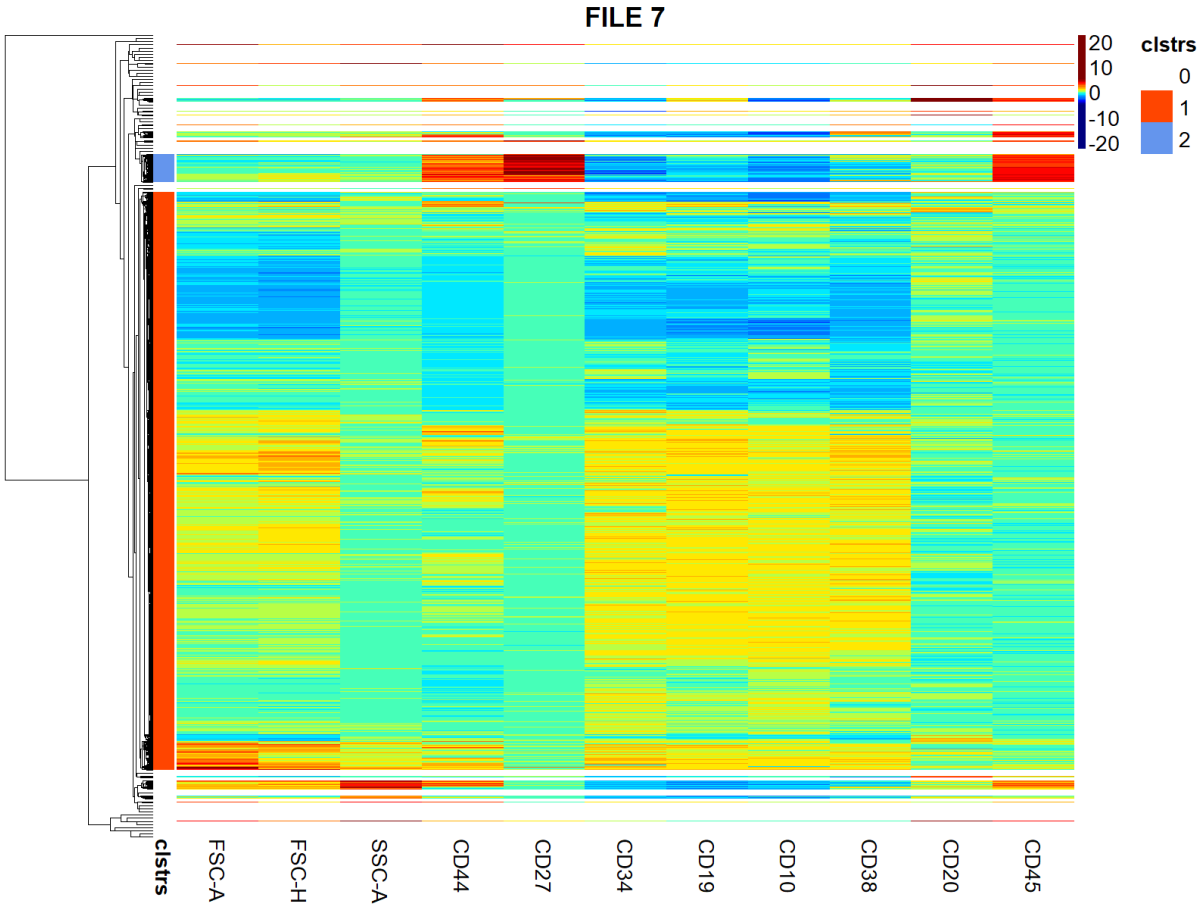

FILE 2
15
10
5
0
-5
-10
-15
clstrs
0
1
4
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

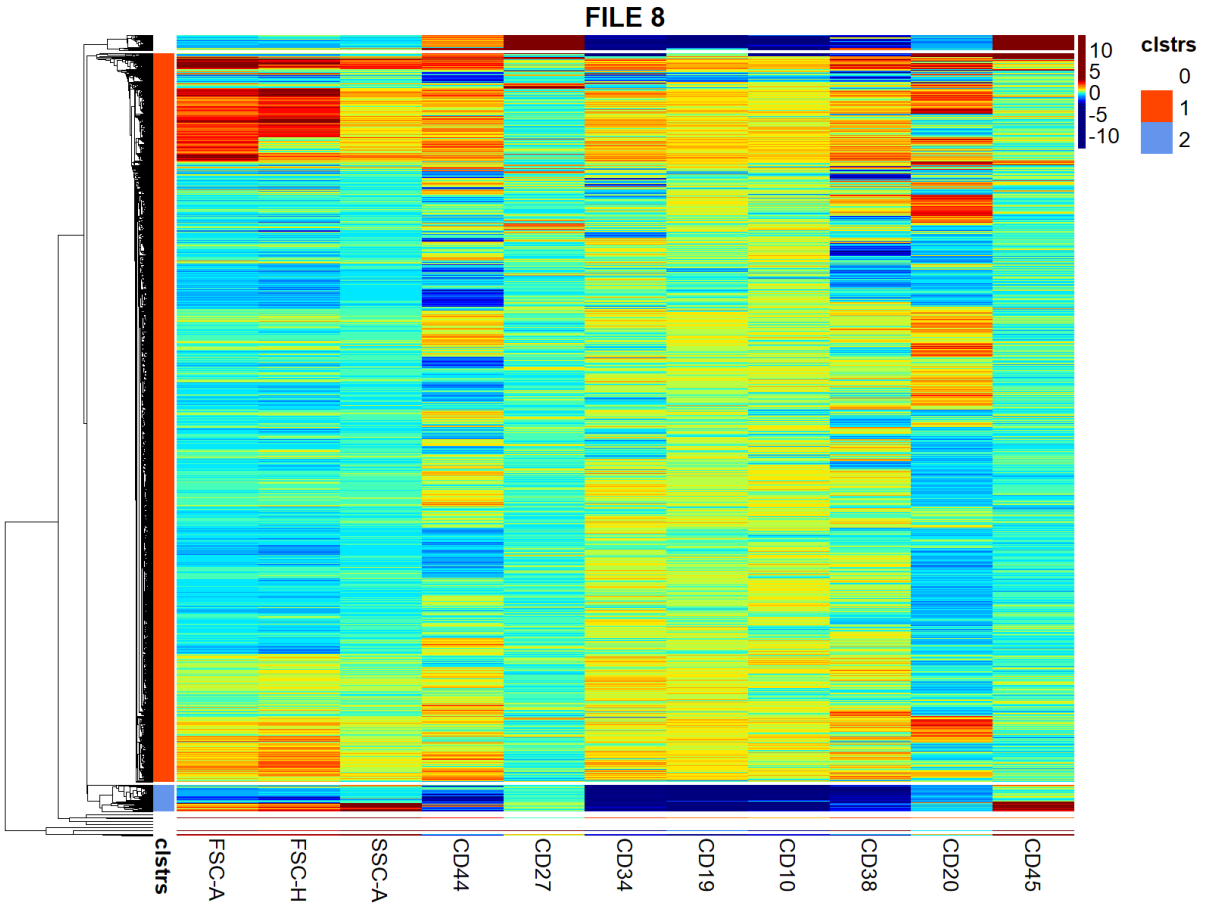

FILE 3
10
5
0
-5
-10
clstrs
0
1
2
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

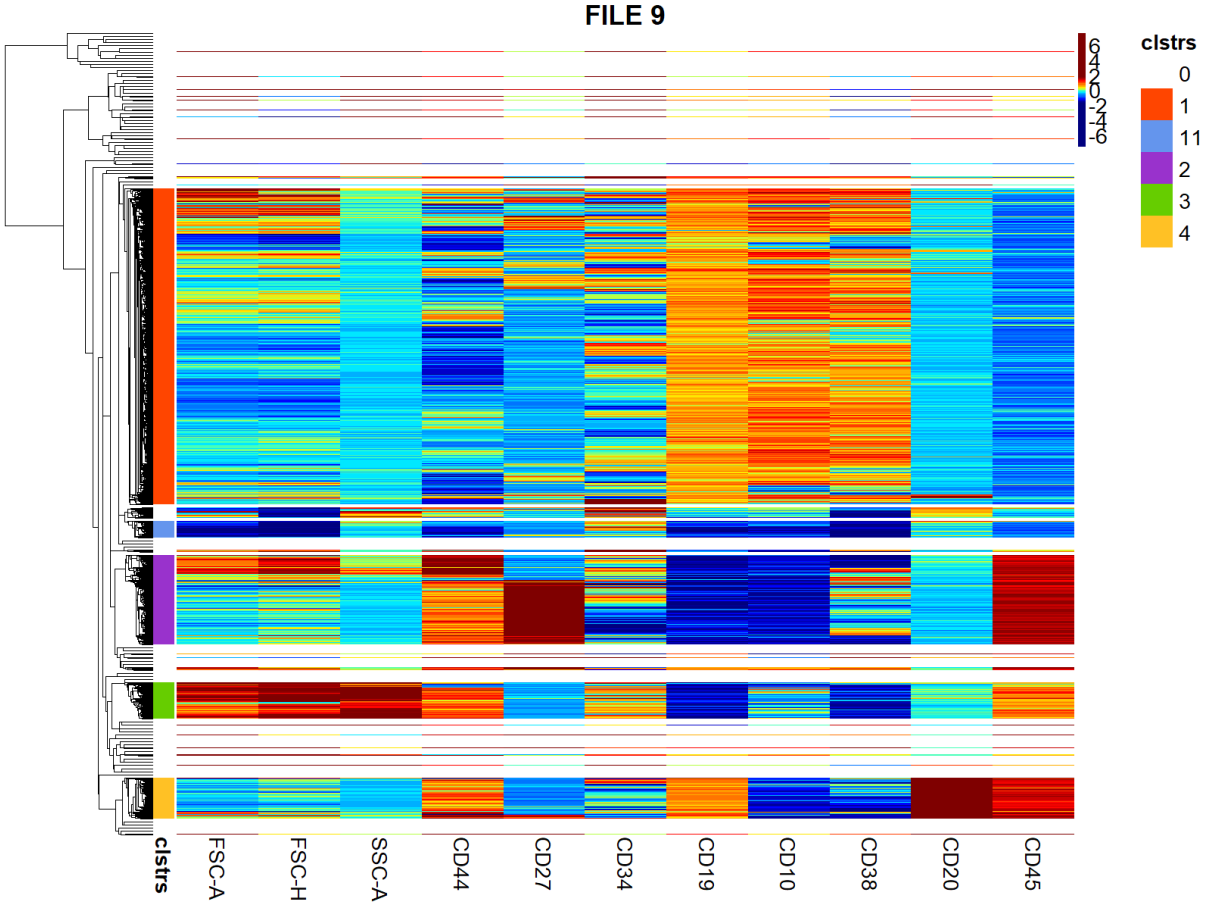

FILE 4
15
10
5
0
-5
-10
-15
clstrs
0
1
3
5
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

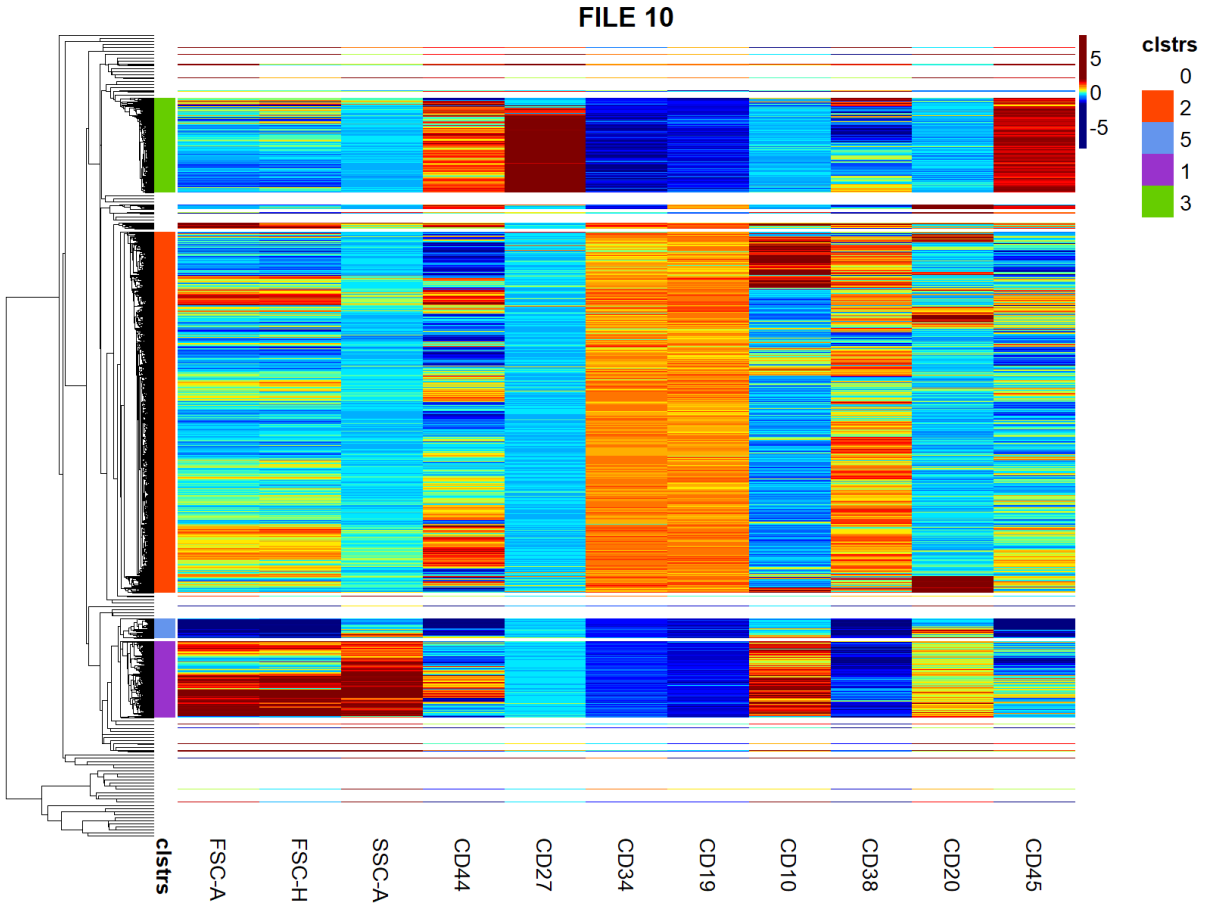

FILE 5
5
0
-5
clstrs
0
3
5
4
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

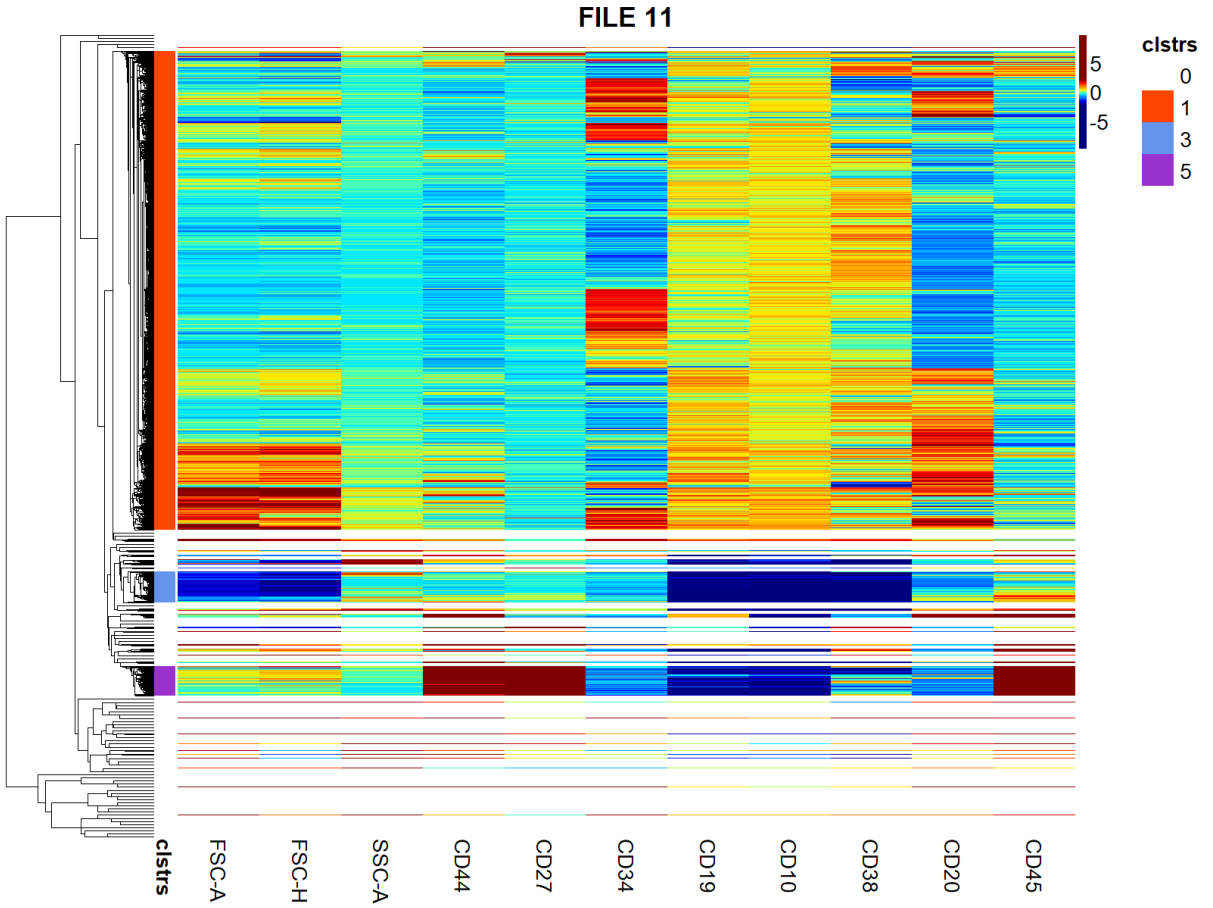

FILE 6
clstrs
0
3
1
6
5
0
-5
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

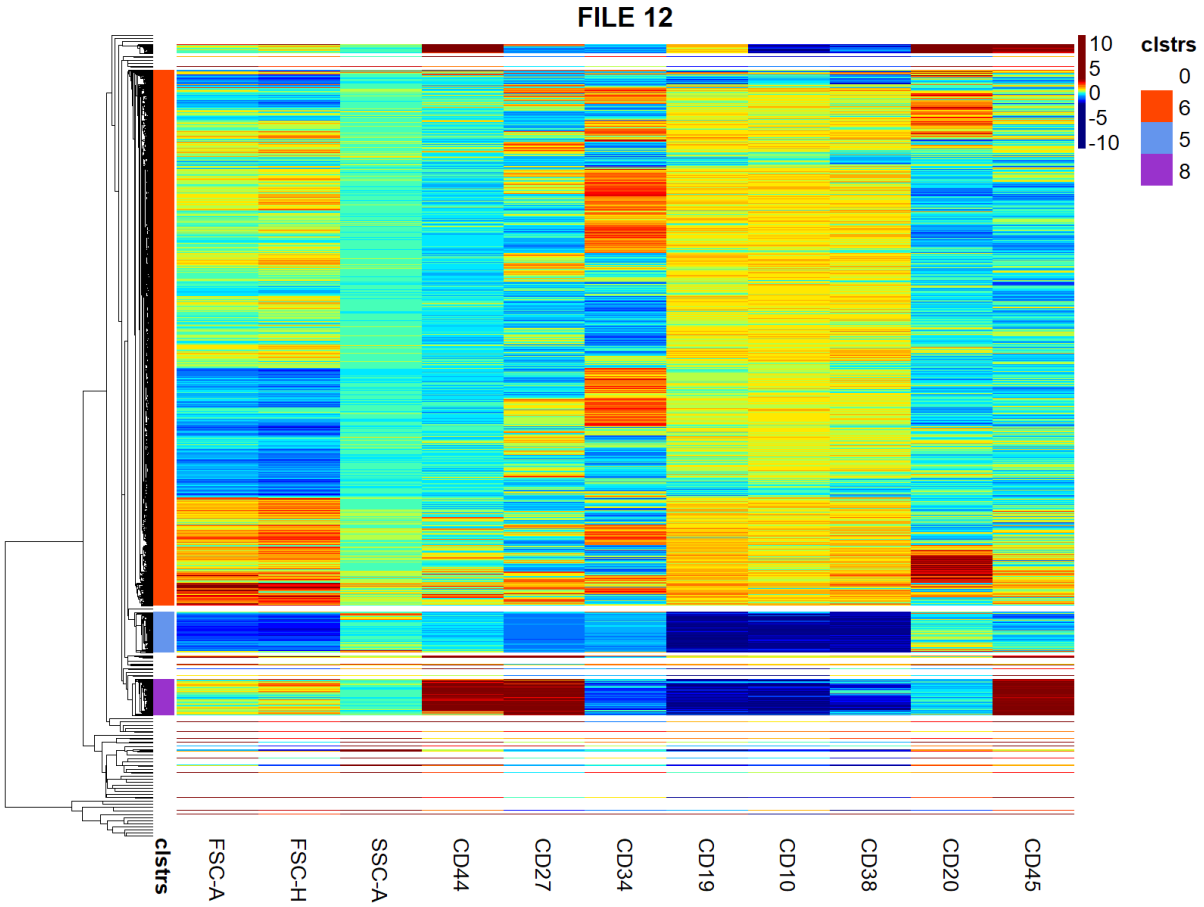

FILE 7
clstrs
0
1
2
20
10
0
-10
-20
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

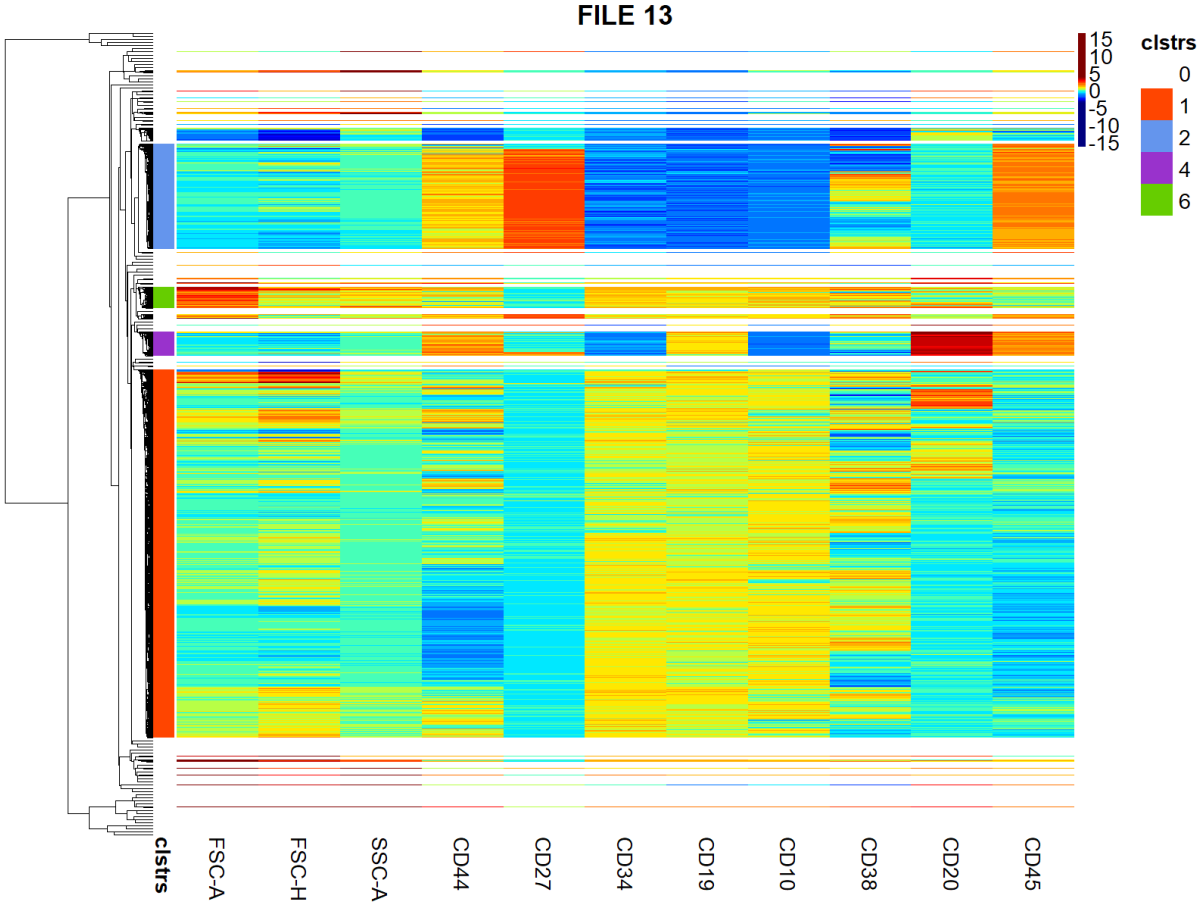

FILE 8
clstrs
0
1
2
10
5
0
-5
-10
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

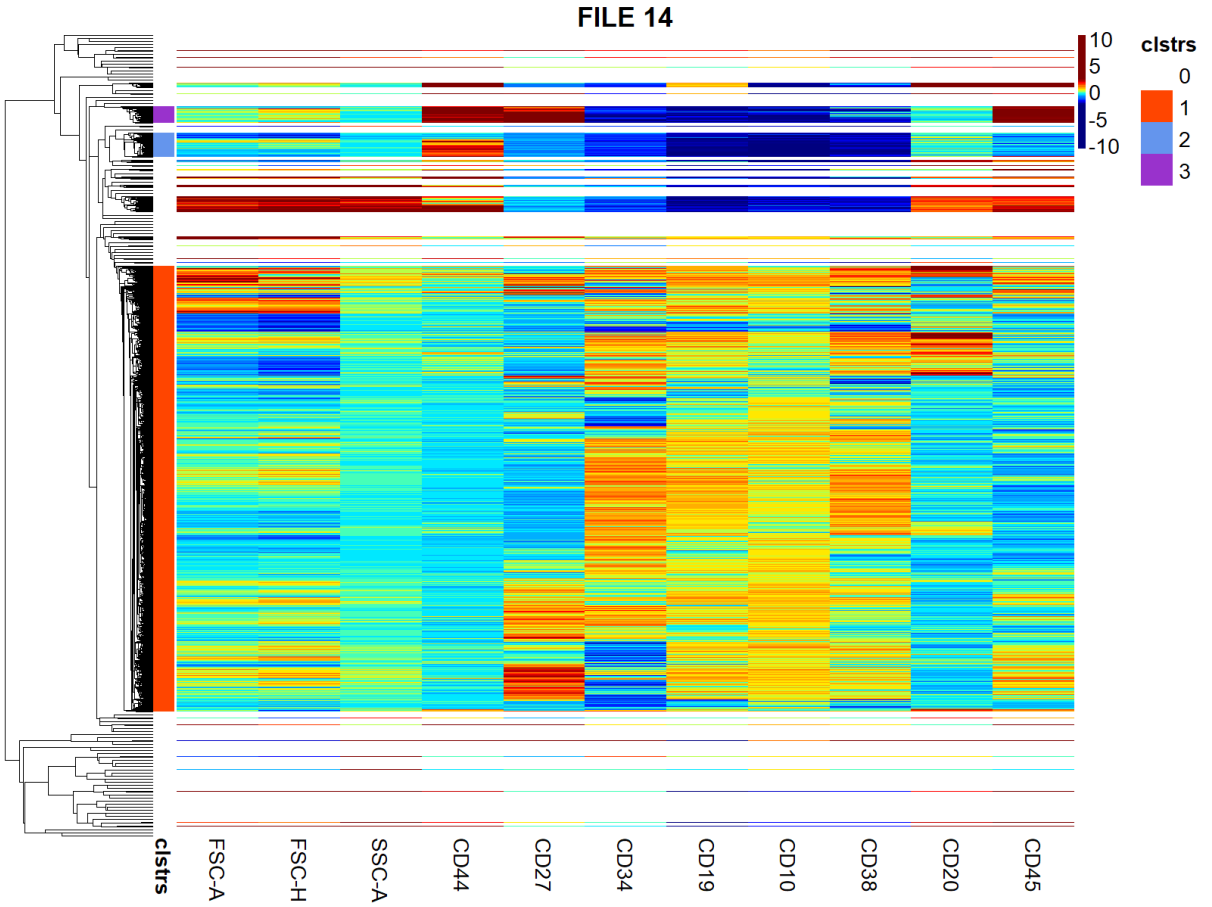

FILE 9
clstrs
0
1
11
2
3
4
6
4
2
0
-2
-4
-6
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

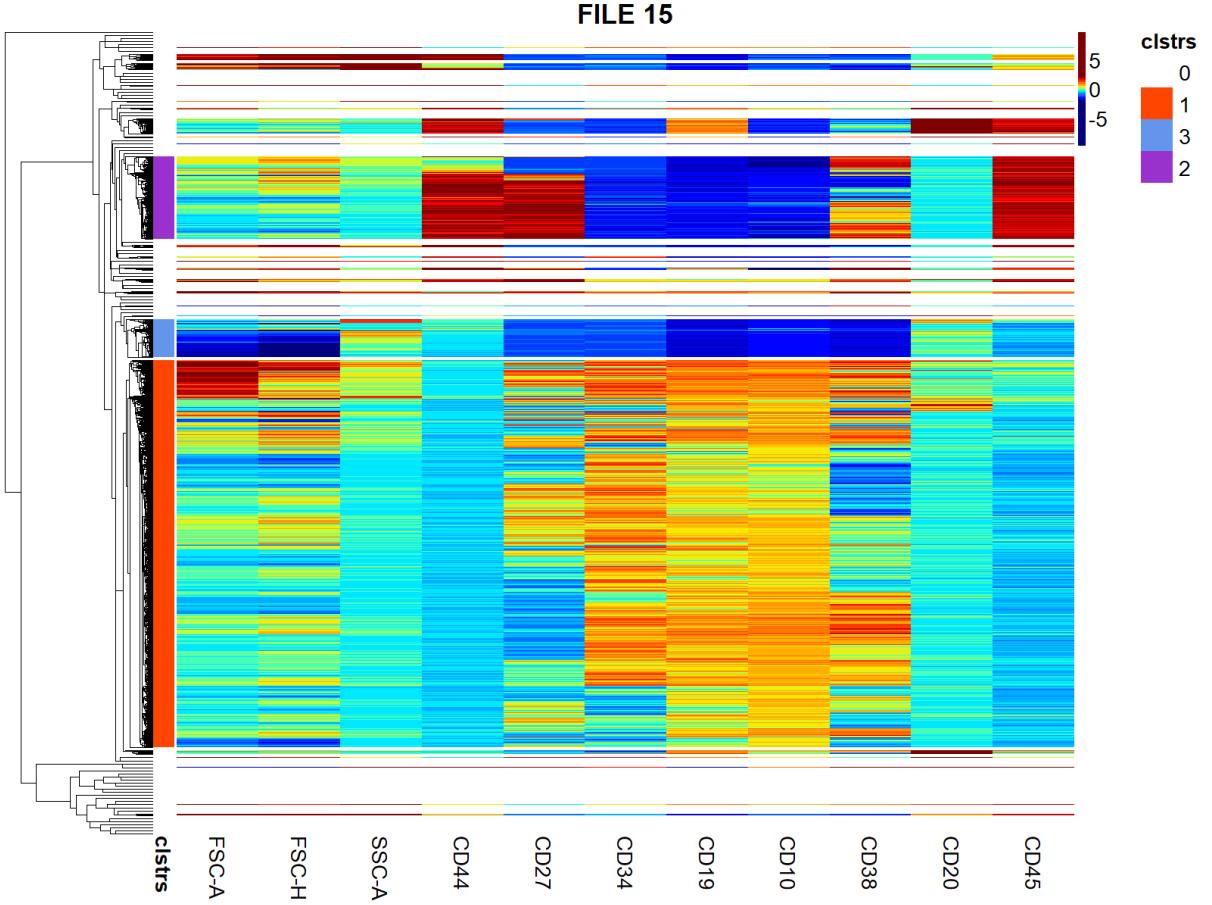

FILE 10
clstrs
0
2
5
1
3
5
0
-5
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

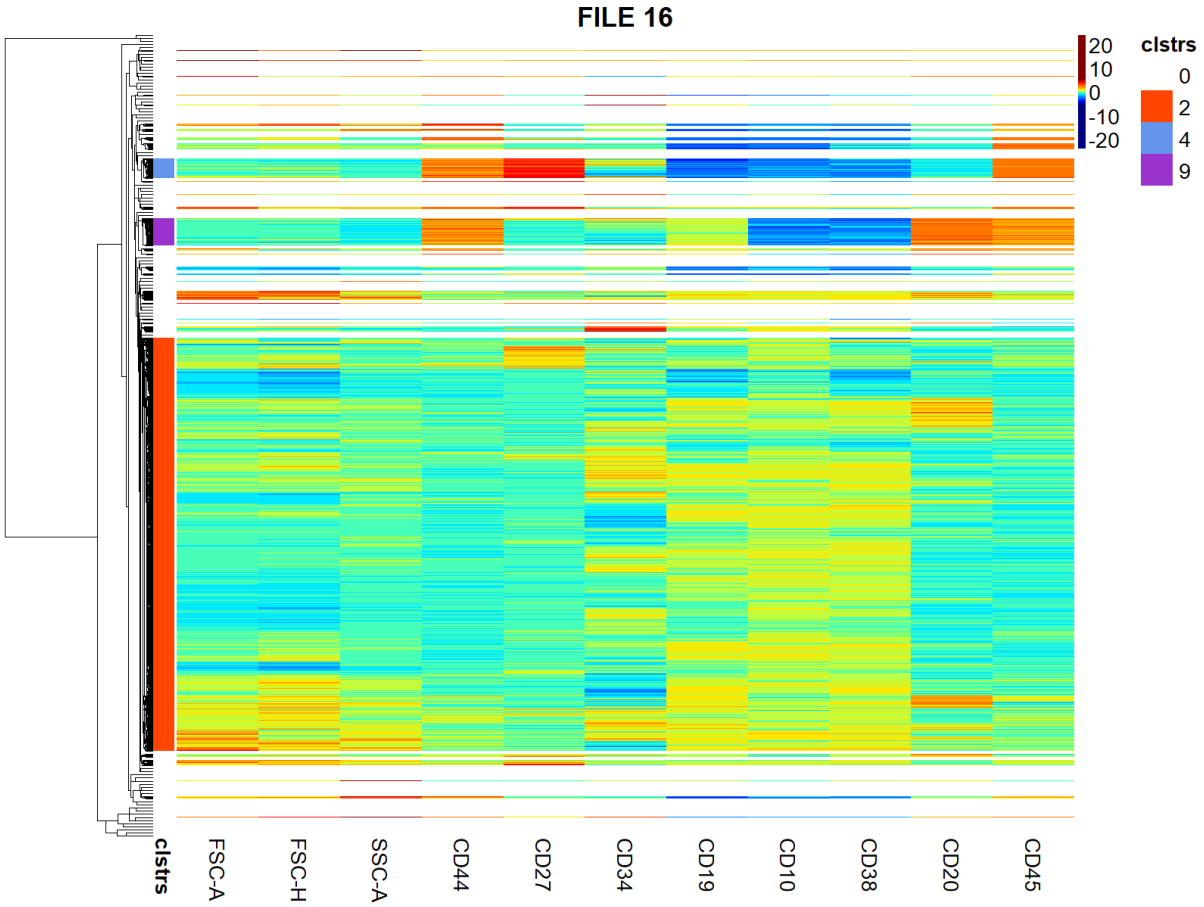

FILE 11
clstrs
0
1
3
5
5
0
-5
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

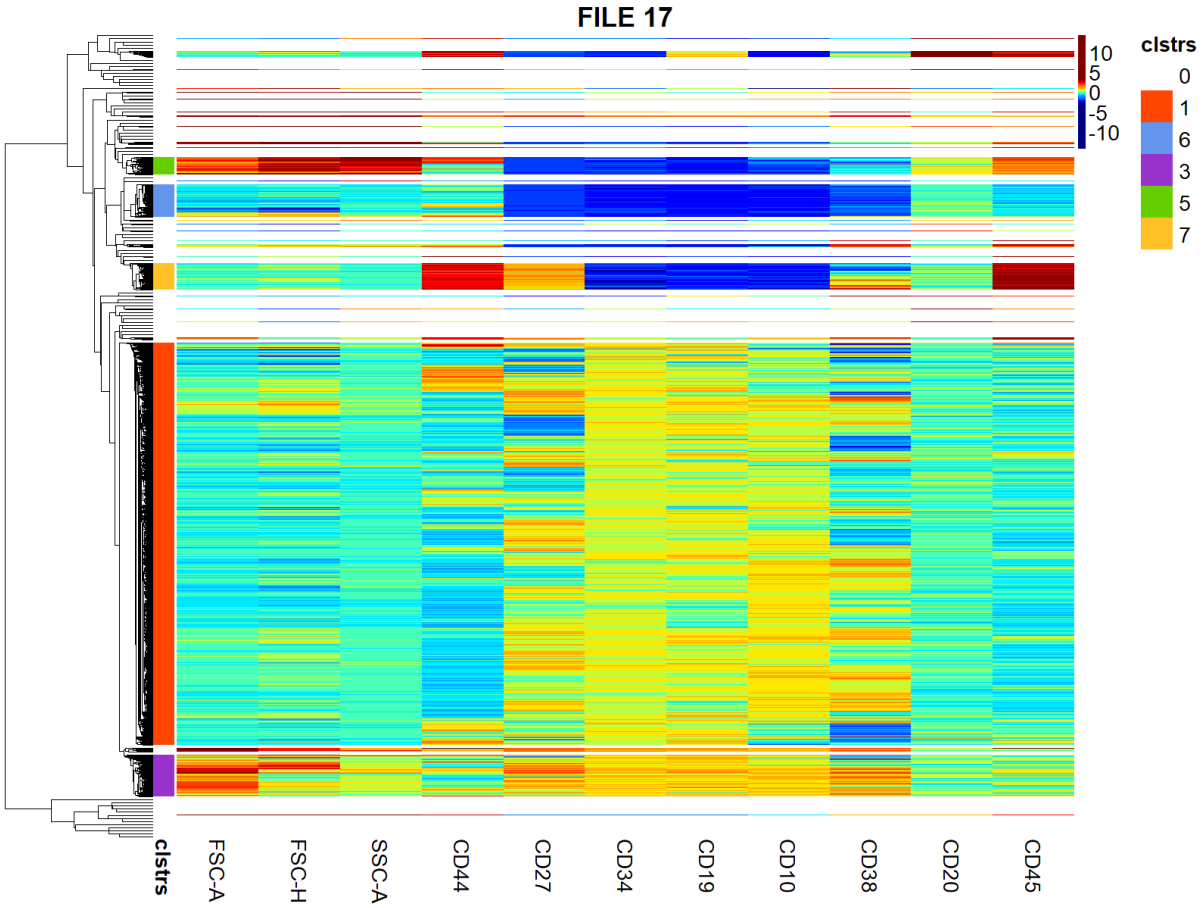

FILE 12
clstrs
0
6
5
8
10
5
0
-5
-10
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

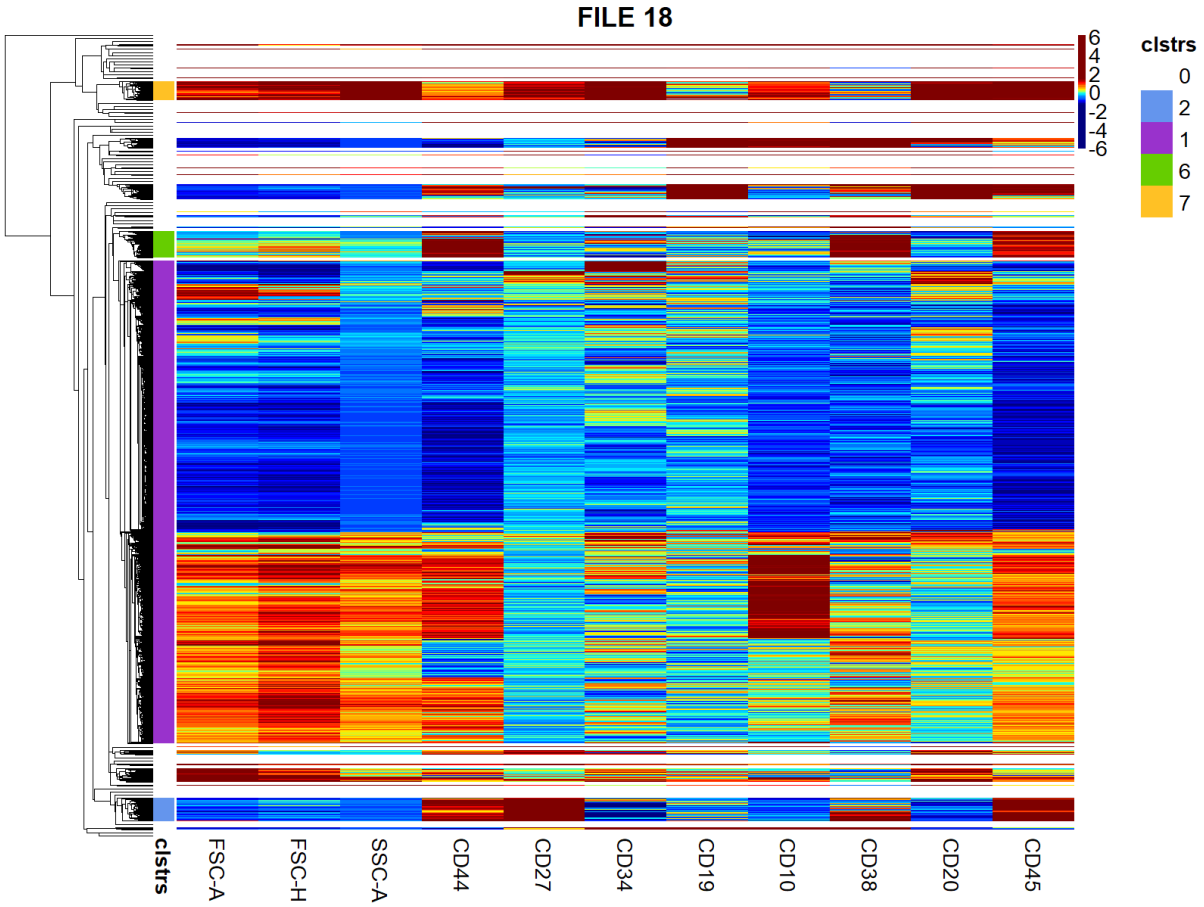

FILE 13
clstrs
0
1
2
4
6
15
10
5
0
-5
-10
-15
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 14
clstrs
0
1
2
3
10
5
0
-5
-10
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 15
clstrs
0
1
3
2
5
0
-5
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 16
20
10
0
-10
-20
clstrs
0
2
4
9
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 17
10
5
0
-5
-10
clstrs
0
1
6
3
5
7
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 18
6
4
2
0
-2
-4
-6
clstrs
0
2
1
6
7
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

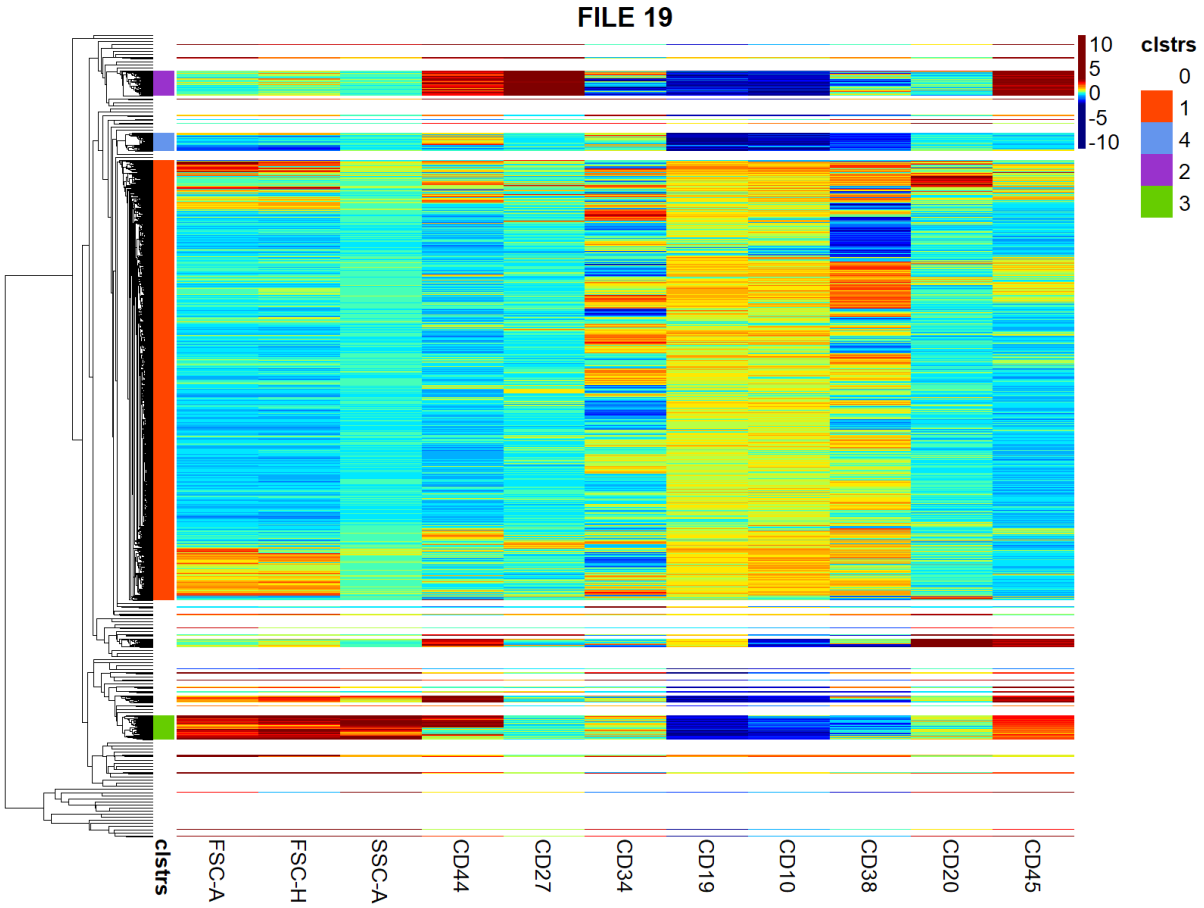
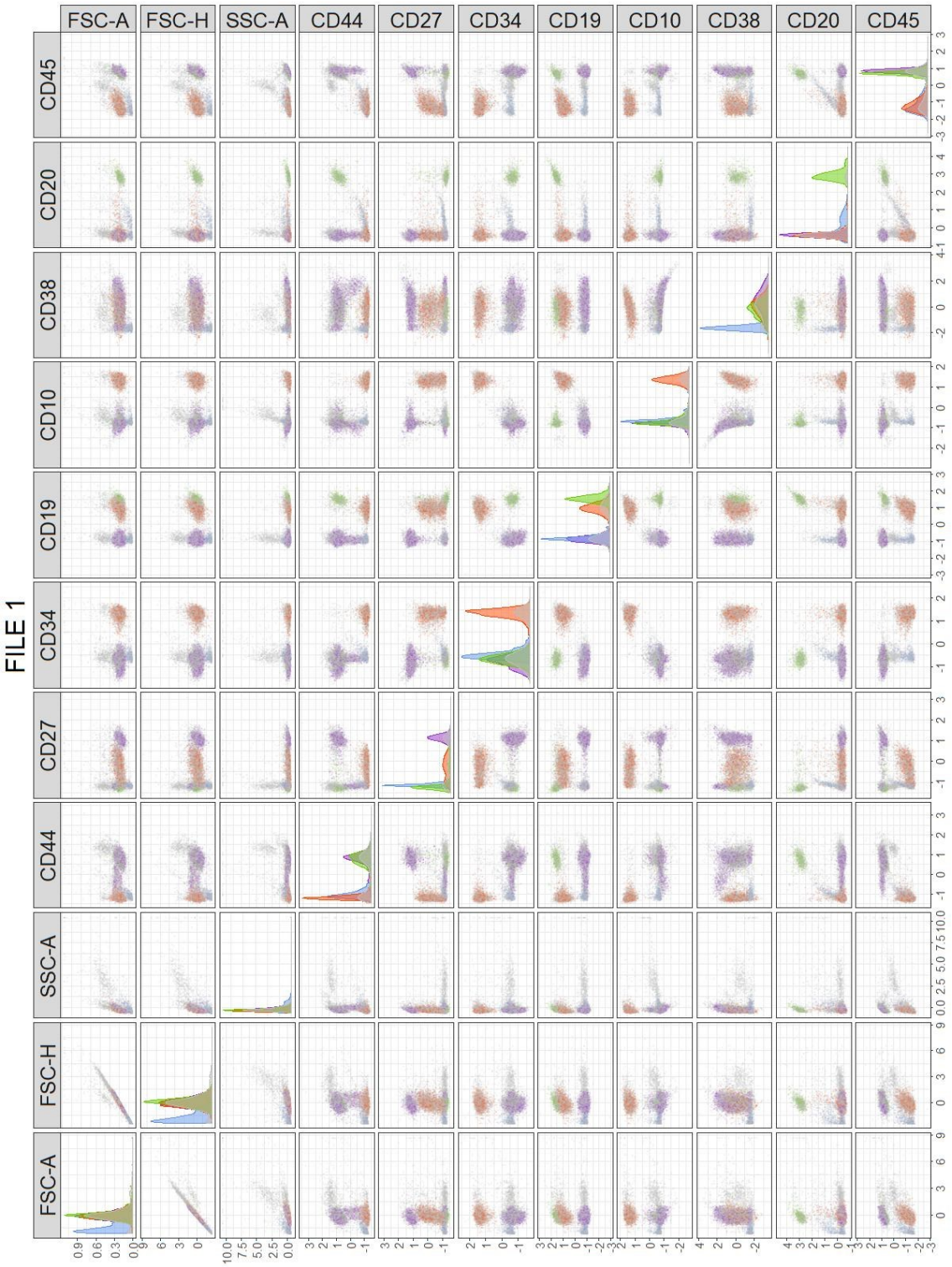

FILE 19
10
5
0
-5
-10
clstrs
0
1
4
2
3
clstrs
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

Fig. 12 – Matrix of scatterplots of the resulting HCA clustering of diagnostic file 1 in parameters FSC-A, FSC-H, SSC-A, CD44, CD27, CD34, CD19, CD10, CD38, CD20 and CD45. For the clustering, the choose_cluster_hca function was used with range of clusters set to 5-100 and minimal size of the populations set to 3 %. The histograms of each parameter are on the diagonal. The clusters of size over 3 % from the resulting HCA iteration of clustering are highlighted in colour – red for blastic population, blue for referential negative population and purple and green for other populations. The blastic B-ALL population was selected manually based on the results of automated immunophenotypisation in chapter 5.4. The colours of the populations correspond to the ones in heatmap with associated dendrogram for file 1 (see Fig. 11). For matrices of scatterplots of HCA clustering of diagnostic files 2-19 see Supplement 1.



5.3.2 FlowSOM

The wrapper function for automatic population detection by FlowSOM was named `choose_cluster_fSOM`. It includes, iteration of clustering of selected files by FlowSOM and the identification of an optimal clustering with an optimal negative cluster. The input of this function is a data frame of preprocessed (compensated, logicle transformed and z-scored) data used also for HCA clustering.

The `FlowSOM` function from R `FlowSOM` library is used for clustering (Van Gassen *et al.*, 2015). The input is the compensated, logicle transformed and z-scored data in the flowFrame format. Similarly to the hierarchical clustering, the FlowSOM was set to process 10 thousand of events in the selected files (the processing time of `choose_cluster_fSOM` function for a file containing 10 thousand events was 2 minutes). For `FlowSOM` function, the number of clusters to be produced is determined by the density of the clustering grid. The dimensions of the grid is set by the `xdim` and `ydim` variables. These dimensions were set by the `cluster_range` variable of the `choose_cluster_fSOM` function to produce iteration of clustering of a ranged number of clusters. The `nClus` variable, determining the number of metaclusters containing the individual clusters, was set to NULL. This option allows the optimisation of the number of metaclusters. The max number of clusters to try for the metaclustering (parameter `maxMeta`) was set to 25.

The median values of clusters chosen as optimal by the `choose_cluster_fSOM` function for the range of dimensions of clustering grid form 6 to 16 for all 19 selected files are shown in Table 7. Only 10 000 events were clustered for each file. The sizes of clusters and the negative cluster used as reference for the immunophenotypisation in chapter 5.4 are also shown.

The clustering result for processed files were depicted in the matrix of scatterplots for parameters FSC-A (forward scatter – area), FSC-H (forward scatter – height), SSC-A (side scatter – area), CD44, CD27, CD34, CD19, CD10, CD38, CD20 and CD45. The resulting FlowSOM clusters with size over 3 % were highlighted in colour – red for blastic population, blue for the negative population used as reference for immuonophenotyping and purple and green for remaining populations. The blastic B-ALL populations were selected manually based on the results of automated immunophenotypisation in chapter 5.4. On the diagonal there is a histogram of fluorescence for each fluorescence parameter. The matrix of scatterplots of file 1 is shown in the Fig. 13. The matrices of scatterplots of file 2-19 are shown in the Supplement 2.

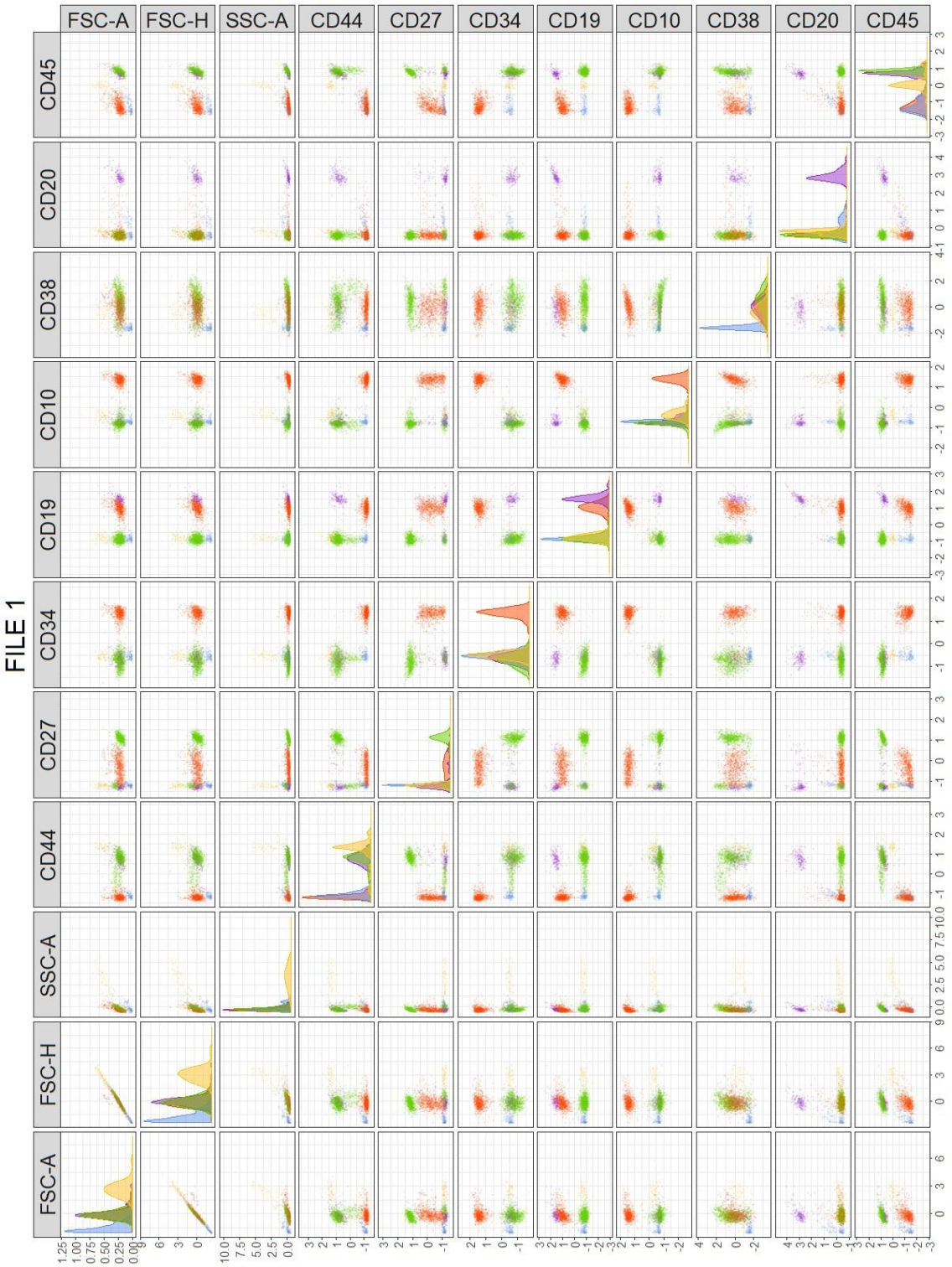
Table 7 - The selected FlowSOM iteration of clustering of diagnostic files 1-19 chosen by choose_cluster_fSOM function with the range of dimensions of clustering grid set to 6-16, median values of their markers and their size. Only clusters of size bigger than 3 % are shown and are used for immunophenotypisation in chapter 5.4. For each file, the total size and number of clusters is shown. The dimensions of clustering grid is equivalent to the xdim and ydim variables of the FlowSOM function chosen as optimal. The negative clusters used as a reference for immunophenotypisation is highlighted in blue.

FILE 1 – 10 000 cells									
Dimensions of clustering grid - 6									
clus-ter	**median CD44**	**median CD27**	**median CD34**	**median CD19**	**median CD10**	**median CD38**	**median CD20**	**median CD45**	**number of cells**
1	-1.191	-0.302	1.389	1.015	1.378	-0.028	-0.389	-1.231	3249
2	-1.106	-1.166	-0.516	-0.846	-0.665	-1.611	-0.266	-1.342	670
4	0.793	-1.239	-0.654	1.550	-0.719	-0.164	2.915	0.713	850
5	0.820	1.080	-0.719	-0.845	-0.755	0.459	-0.393	0.842	4646
8	1.339	-1.166	-0.534	-0.760	-0.319	-0.564	-0.220	0.038	312
FILE 2 – 10 000 cells									
Dimensions of clustering grid - 7									
1	1.824	3.990	-2.319	-2.052	-1.771	-0.140	-0.182	2.615	546
3	2.365	-0.231	-2.126	-2.111	-1.573	0.617	0.043	1.979	430
6	-0.425	-0.280	0.411	0.296	0.352	-0.189	-0.216	-0.387	8708
FILE 3 – 10 000 cells									
Dimensions of clustering grid - 12									
1	0.083	-0.312	0.471	0.325	0.389	0.317	-0.278	-0.328	8733
2	-3.274	-0.423	-1.456	-2.898	-2.712	-2.486	0.505	-0.050	308
6	0.424	3.831	-1.738	-2.722	-2.648	-1.719	0.514	2.937	530
FILE 4 – 10 000 cells									
Dimensions of clustering grid - 6									
1	1.884	0.404	1.024	0.898	0.809	1.261	1.115	1.500	318
2	-0.270	-0.300	0.387	0.311	0.347	0.286	-0.085	-0.289	8430
4	-0.457	-0.374	-1.932	-2.886	-2.686	-2.244	-1.096	-1.008	463
5	1.099	4.054	-2.352	-2.832	-2.737	-1.865	-1.166	2.296	438
FILE 5 – 10 000 cells									
Dimensions of clustering grid - 11									
1	-0.510	-0.200	0.535	0.513	0.661	0.117	-0.339	-0.336	6820
2	0.730	-0.191	0.965	0.977	0.801	0.944	-1.136	1.238	517
3	-0.748	-0.207	-1.780	-1.954	-1.448	-1.200	0.677	-1.237	838
4	1.939	-0.140	-1.697	-1.825	-1.188	-0.962	0.517	0.708	1175
FILE 6 – 10 000 cells									
Dimensions of clustering grid - 9									
1	0.165	-0.196	-0.243	0.470	-0.266	0.449	-0.234	0.047	7882
2	1.248	-0.101	-0.049	0.780	0.752	0.801	-0.698	0.943	345
4	-0.022	-0.019	-0.056	-1.704	0.880	-1.271	0.855	0.231	578
5	-1.327	-0.189	-0.376	-1.994	-0.671	-2.439	0.474	-2.512	712

FILE 7 - 10 000 cells									
Dimensions of clustering grid - 6									
clus-ter	**median CD44**	**median CD27**	**median CD34**	**median CD19**	**median CD10**	**median CD38**	**median CD20**	**median CD45**	**number of cells**
1	-0.030	-0.242	0.714	0.790	0.746	0.727	-0.200	-0.314	5674
2	-0.796	-0.274	-0.855	-0.917	-0.612	-1.102	-0.079	-0.340	3401
9	2.642	4.382	-1.687	-1.173	-1.624	-0.765	-0.163	3.385	410
FILE 8 – 10 000 cells									
Dimensions of clustering grid - 9									
2	0.061	-0.23	0.263	0.227	0.269	0.188	-0.272	-0.232	9287
4	-1.347	-0.25	-3.054	-4.007	-3.715	-2.662	-0.911	0.397	336
FILE 9 – 10 000 cells									
Dimensions of clustering grid - 12									
1	-0.482	-0.494	-0.096	0.574	0.707	0.584	-0.378	-0.733	6466
2	-1.278	-0.646	0.150	-1.354	-1.244	-1.790	-0.290	-0.802	329
3	1.251	-0.573	0.288	-1.556	-0.856	-1.433	-0.203	0.980	939
4	0.765	2.252	-0.691	-1.439	-1.415	-0.259	-0.424	1.551	1214
6	0.808	-0.678	-0.272	0.741	-1.243	-0.881	3.031	1.396	771
FILE 10 – 10 000 cells									
Dimensions of clustering grid - 16									
1	0.943	2.358	-1.514	-1.339	-0.518	-0.549	-0.529	1.567	1547
2	-0.145	-0.468	0.725	0.714	-0.515	0.588	-0.391	-0.224	6220
3	-1.855	-0.453	-1.240	-1.485	-0.461	-1.963	-0.129	-2.166	309
4	-0.538	-0.410	-1.159	-1.414	0.821	-1.545	0.237	-0.815	610
5	-0.032	-0.381	0.587	0.786	0.213	0.027	3.493	0.456	326
6	0.046	-0.393	-1.090	-1.385	1.433	-1.152	0.307	-0.179	757
FILE 11 – 10 000 cells									
Dimensions of clustering grid - 16									
1	-0.384	-0.294	-0.257	0.381	0.415	0.349	-0.238	-0.363	8529
3	2.814	3.479	-0.811	-2.236	-2.323	-0.404	-0.838	2.954	683
4	-0.226	-0.207	-0.536	-2.459	-2.487	-2.469	-0.560	0.082	619
FILE 12 – 10 000 cells									
Dimensions of clustering grid - 6									
1	-0.390	-0.190	0.121	0.362	0.444	0.388	-0.395	-0.354	7158
4	0.036	-0.156	-0.441	0.611	0.421	0.572	1.590	0.321	1148
7	-0.583	-1.129	-0.862	-2.514	-2.446	-2.468	-0.170	-0.621	584
8	2.830	2.804	-1.153	-2.316	-2.450	-1.596	-0.469	2.750	618
FILE 13 – 10 000 cells									
Dimensions of clustering grid - 11									
1	0.952	2.010	-1.607	-1.749	-1.536	-0.125	-0.514	1.502	1903
2	1.283	-0.472	-1.381	0.760	-1.571	-0.439	3.042	1.386	455
3	0.574	-0.452	0.849	0.835	0.863	0.966	0.059	0.226	570
6	-0.391	-0.531	0.588	0.466	0.637	0.143	-0.416	-0.591	6492
8	-1.879	-0.500	-1.347	-1.866	-1.493	-2.085	0.178	-0.598	303

FILE 14 – 10 000 cells									
Dimensions of clustering grid - 6									
clus-ter	**median CD44**	**median CD27**	**median CD34**	**median CD19**	**median CD10**	**median CD38**	**median CD20**	**median CD45**	**number of cells**
1	-0.411	-0.315	0.400	0.376	0.432	0.314	-0.384	-0.318	7844
2	-0.091	0.469	0.656	0.832	0.711	0.736	-0.315	0.509	312
4	-0.271	-0.467	0.743	0.715	0.291	1.040	2.279	0.020	376
6	1.676	-0.706	-1.221	-2.248	-1.877	-1.810	1.260	1.731	369
8	2.910	3.006	-1.812	-2.202	-2.336	-1.548	-0.309	2.769	331
9	1.199	-0.877	-1.455	-2.331	-2.677	-2.222	-0.157	-0.765	525
FILE 15 – 10 000 cells									
Dimensions of clustering grid - 8									
1	1.897	1.864	-1.359	-1.608	-1.585	0.574	-0.318	1.918	1580
6	-0.533	-0.154	0.626	0.546	0.617	0.142	-0.274	-0.559	6961
7	1.840	-1.122	-1.243	0.813	-1.494	-0.467	4.733	1.806	318
8	-0.444	-1.134	-1.179	-1.681	-1.512	-1.687	-0.085	-0.694	672
FILE 16 – 10 000 cells									
Dimensions of clustering grid - 14									
1	2.237	3.458	-0.326	-2.548	-2.365	-2.158	-0.721	2.298	425
3	-0.491	-0.377	-0.068	0.197	0.385	0.364	-0.447	-0.470	7867
4	2.086	-0.515	-0.217	0.449	-2.103	-1.666	2.491	2.016	577
6	0.121	-0.166	0.281	0.900	0.734	0.998	0.583	0.219	663
FILE 17 – 10 000 cells									
Dimensions of clustering grid - 6									
1	2.351	1.053	-2.151	-2.066	-2.155	0.178	-0.159	2.884	523
2	-0.212	-1.915	-2.198	-2.198	-2.187	-1.664	-0.200	-0.646	625
3	-0.501	0.319	0.438	0.420	0.474	0.161	-0.185	-0.404	8048
5	0.780	-1.838	-1.953	-2.072	-1.609	-0.362	0.425	1.478	445
FILE 18 – 10 000 cells									
Dimensions of clustering grid - 8									
1	0.779	-0.287	0.042	-0.237	0.283	0.350	-0.107	0.581	4142
2	-0.967	-0.369	-0.418	-0.360	-0.755	-0.756	-0.506	-1.055	4027
5	0.719	1.305	2.183	0.157	1.020	0.093	1.756	1.627	395
7	1.332	3.968	-1.273	-0.404	-0.799	0.675	-0.830	1.671	391
FILE 19 – 10 000 cells									
Dimensions of clustering grid - 11									
2	2.106	-0.313	0.426	-2.259	-1.783	-0.845	0.156	1.538	564
3	2.098	3.909	-0.925	-2.053	-2.252	-0.158	-0.252	2.520	533
4	0.276	-0.423	-0.063	-2.426	-2.281	-1.712	-0.347	-0.536	304
5	-0.464	-0.307	-0.052	0.398	0.417	0.281	-0.265	-0.503	8257

Fig. 13 - Matrix of scatterplots of the resulting FlowSOM clustering of diagnostic file 1 in parameters FSC-A, FSC-H, SSC-A, CD44, CD27, CD34, CD19, CD10, CD38, CD20 and CD45. For the clustering, the choose_cluster_fSOM function was used with the range of dimensions of clustering grid set to 6-16 and minimal size of the populations set to 3 %. The histograms of each parameter are on the diagonal. The clusters of size over 3 % from the resulting FlowSOM iteration of clustering are highlighted in colour – red for blastic population, blue for referential negative population and purple, green and yellow for other populations. The blastic B-ALL population was selected manually based on the results of automated immunophenotypisation in chapter 5.4. For matrices of scatterplots of FlowSOM clustering of diagnostic files 2-19 see Supplement 2.



5.4 Automatic reporting of immunophenotype of detected cell populations

For determining the immunophenotype of the populations of iteration of clustering selected by the `choose_cluster_hca` or `choose_cluster_fSOM` function another function was scripted. The HCA version of the function is named `report_all_clusters_hca` and the FlowSOM `report_all_clusters_fSOM`.

For the direct reporting of the immunophenotype, the `report_all_clusters_hca` and `report_all_clusters_fSOM` functions use the `makeReport` function from the *in house* `flowPheno` R library.

The `report_all_clusters_hca` and `report_all_clusters_fSOM` functions determine the immunophenotype of all clusters based on the clustered data in the data frame format. The phenotype is determined for clusters of relevant size only and the assignment of each cell to a cluster was based on the selected iteration of clustering object (FlowSOM or HCA clustering object). The `makeReport` function requires the negative population to be set for the reasons of scaling other values of populations against the values of the negative population. As this negative population the negative cluster selected by the `choose_cluster` functions is set. All the necessary input for the `report_all_clusters` functions is saved in the `choose_cluster` functions output (clusters of relevant size, selected iteration of FlowSOM or HCA clustering, selected negative cluster, preprocessed data in data frame).

The results of the `report_all_clusters_hca` and the `report_all_clusters_fSOM` functions is a report in two formats for all clusters with the population bigger than 3 % of all events (as chosen by the functions `choose_cluster_hca` or `choose_cluster_fSOM`). The table format reports the immunophenotype of each marker in the WHO standard, Detail and several metrics. The second reporting format is the FDE format which is a clinically codified string form of reporting the immunophenotype (Dworzak *et al.*, 2018). Even though the FDE format is very convenient, it lacks the valuable information about the population detail, which describes the immunophenotype more clearly. Also, for each cluster an identifying string is added by the `report_all_clusters_hca` or `report_all_clusters_fSOM` functions. It represents the number of the cluster, the number of the negative cluster it is compared to in terms of phenotype by the `flowPheno` function, its population percentage and total number of events. The example of the string is “pos.1 neg.2 - 45.85 %, 917events”, meaning that the report shows the immunophenotype of cluster number 1 of a size of 917 cells and 45.85 % of the total events in the file and was compared to negative cluster number 2.

The report of the `report_all_clusters_hca` or `report_all_clusters_fSOM` functions can be easily used for identifying any population. For example, the `report_all_clusters_hca` function results were used to identify the blast population in the diagnostic file 1 based on sizes and immunophenotypes of the clusters. The data of the file 1 were visualised in the matrix of scatterplots using `ggpairs` function of the `GGally` library and highlighted the cells of this B-ALL blast population in red (see Fig. 12 and Fig. 13).

The immunophenotype results for all 19 files for HCA clustering are shown in Table 8 and for the FlowSOM clustering in Table 9 in FDE format and in table format including the WHO immunophenotype classification of the population and Detail, to gather fulfilling AIEOP-BFM (Associazione Italiana Ematologia Oncologia Pediatric Group, Berlin-Frankfurt-Munster) clinical reporting guidelines (Dworzak *et al.*, 2018). The metrics of the table format report were left out.

Table 8 - Immunophenotype determined by the report_all_clusters_hca funtion for all HCA clusters clusters bigger than 3 % of all files. The clustering was chosen automatically from iterations of clustering with the range of clusters from 5 to 100 in chapter 5.3.1. For each cluster, its number, the number of the negative reference cluster, its population percentage and total number of events is shown (in the format "pos.1 neg.2 - 45.85 %, 917events"). The phenotype is reported for each marker in the WHO immunophenotype classification and population detail (describing the histogram distribution of the population). The FDE format of report of immunophenotype is shown on the bottom of each file section.

FILE 1				
	pos.1 neg.2 - 45.18 %, 4518 events (1)		**pos.2 neg.2 - 7.01 %, 701 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Medium	NEGATIVE	Negative
CD27	STRONG	Partially positive 2	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Partially positive 2	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Bright	NEGATIVE	Negative
	(1) FDE:45.18%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .			
	(2) FDE:7.01%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	pos.3 neg.2 - 31.48 %, 3148 events (3)		**pos.4 neg.2 - 8.39 %, 839 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	STRONG	Bright
CD27	STRONG	Partially positive 2	WEAK	Dim
CD34	STRONG	Bright	NEGATIVE	Negative
CD19	STRONG	Bright	STRONG	Bright
CD10	STRONG	Bright	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	STRONG	Bright
CD45	NEGATIVE	Negative	STRONG	Bright
	(3) FDE:31.48%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .			
	(4) FDE:8.39%;STRONG:CD19,20,38,44,45,WEAK:CD27,NEGATIVE:CD10,34; ……. .			

FILE 2				
	pos.1 neg.4 - 87.37 %, 8737 events (1)		**pos.4 neg.4 - 4.78 %, 478 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Medium	NEGATIVE	Negative
CD19	STRONG	Heterogeneous	NEGATIVE	Negative
CD10	STRONG	Heterogeneous	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:87.37%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .			
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 3				
	pos.1 neg.2 - 89.57 %, 8957 events (1)		pos.2 neg.2 - 8.78 %, 878 events (2)	
parameter	WHO	Detail	WHO	Detail
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Partially positive 2	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative
CD10	STRONG	Medium	NEGATIVE	Negative
CD38	WEAK	Dim	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:89.57%;STRONG:CD10,19,34,WEAK:CD38,NEGATIVE:CD20,27,44,45; ……. .			
	(2) FDE:8.78%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 4						
	pos.1 neg.3 - 86.18 %, 8618 events (1)		pos.3 neg.3 - 3.67 %, 367 events (2)		pos.5 neg.3 - 4.3 %, 430 events (3)	
parame-ter	WHO	Detail	WHO	Detail	WHO	Detail
CD44	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Bright
CD34	WEAK	Dim	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD10	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD38	WEAK	Dim	NEGATIVE	Negative	NEGATIVE	Negative
CD20	STRONG	Heteroge-neous	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:86.18%;STRONG:CD10,19,20,WEAK:CD34,38,NEGATIVE:CD27,44,45; ……. .					
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(3) FDE:<5%;STRONG:CD27,NEGATIVE:CD10,19,20,34,38,44,45; ……. .					

FILE 5						
	pos.3 neg.5 - 70.84 %, 7084 events (1)		pos.4 neg.5 - 8.88 %, 888 events (2)		pos.5 neg.5 - 9.53 %, 953 events (3)	
parame-ter	WHO	Detail	WHO	Detail	WHO	Detail
CD44	NEGATIVE	Negative	STRONG	Partially posi-tive 2	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Medium	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Medium	NEGATIVE	Negative	NEGATIVE	Negative
CD10	WEAK	Partially posi-tive 1	NEGATIVE	Negative	NEGATIVE	Negative
CD38	WEAK	Dim	NEGATIVE	Negative	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	WEAK	Dim	STRONG	Heteroge-neous	NEGATIVE	Negative
	(1) FDE:70.84%;STRONG:CD19,34,WEAK:CD10,38,45,NEGATIVE:CD20,27,44; ……. .					
	(2) FDE:8.88%;STRONG:CD44,45,NEGATIVE:CD10,19,20,27,34,38; ……. .					
	(3) FDE:9.53%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					

FILE 6						
	pos.1 neg.1 - 6.85 %, 685 events (1)		**pos.3 neg.1 - 79.9 %, 7990 events (2)**		**pos.6 neg.1 - 5.07 %, 507 events (3)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	WEAK	Dim	WEAK	Partially posi-tive 1
CD27	NEGATIVE	Negative	NEGATIVE	Negative	WEAK	Dim
CD34	NEGATIVE	Negative	WEAK	Dim	NEGATIVE	Negative
CD19	NEGATIVE	Negative	STRONG	Medium	NEGATIVE	Negative
CD10	NEGATIVE	Negative	WEAK	Dim	STRONG	Heteroge-neous
CD38	NEGATIVE	Negative	STRONG	Medium	WEAK	Dim
CD20	NEGATIVE	Negative	NEGATIVE	Negative	WEAK	Dim
CD45	NEGATIVE	Negative	WEAK	Dim	WEAK	Dim
	(1) FDE:6.85%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(2) FDE:79.9%;STRONG:CD19,38,WEAK:CD10,34,44,45,NEGATIVE:CD20,27; ……. .					
	(3) FDE:5.07%;STRONG:CD10,WEAK:CD20,27,38,44,45,NEGATIVE:CD19,34; ……. .					

FILE 7				
	pos.1 neg.2 - 90.59 %, 9059 events (1)		**pos.2 neg.2 - 4.34 %, 434 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Partially positive 2	NEGATIVE	Negative
CD19	STRONG	Heterogeneous	NEGATIVE	Negative
CD10	STRONG	Heterogeneous	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:90.59%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .			
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 8				
	pos.1 neg.2 - 94.19 %, 9419 events (1)		**pos.2 neg.2 - 3.39 %, 339 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Medium	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative
CD10	STRONG	Medium	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:94.19%;STRONG:CD10,19,34,38,NEGATIVE:CD20,27,44,45; ……. .			
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 9						
	pos.1 neg.11 - 59.49 %, 5949 events (1)		**pos.2 neg.11 - 16.84 %, 1684 events (2)**		**pos.3 neg.11 - 6.87 %, 687 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	STRONG	Heterogeneous	STRONG	Partially positive 2
CD27	WEAK	Dim	STRONG	Partially positive 2	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD10	STRONG	Medium	NEGATIVE	Negative	WEAK	Dim
CD38	STRONG	Medium	STRONG	Partially positive 2	WEAK	Dim
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	STRONG	Bright	STRONG	Bright
	(1) FDE:59.49%;STRONG:CD10,19,38,WEAK:CD27,44,NEGATIVE:CD20,34,45; ……. .					
	(2) FDE:16.84%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .					
	(3) FDE:6.87%;STRONG:CD44,45,WEAK:CD10,38,NEGATIVE:CD19,20,27,34; ……. .					
	pos.4 neg.11 - 7.61 %, 761 events (4)		**pos.11 neg.11 - 3.18 %, 318 events (5)**			
parameter	**WHO**	**Detail**	**WHO**	**Detail**		
CD44	STRONG	Medium	NEGATIVE	Negative		
CD27	NEGATIVE	Negative	NEGATIVE	Negative		
CD34	NEGATIVE	Negative	NEGATIVE	Negative		
CD19	STRONG	Bright	NEGATIVE	Negative		
CD10	NEGATIVE	Negative	NEGATIVE	Negative		
CD38	STRONG	Heterogeneous	NEGATIVE	Negative		
CD20	STRONG	Bright	NEGATIVE	Negative		
CD45	STRONG	Bright	NEGATIVE	Negative		
	(4) FDE:7.61%;STRONG:CD19,20,38,44,45,NEGATIVE:CD10,27,34; ……. .					
	(5) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					

FILE 10				
	pos.1 neg.5 - 13.32 %, 1332 events (1)		**pos.2 neg.5 - 62.64 %, 6264 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Partially positive 1	WEAK	Dim
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	STRONG	Bright
CD19	NEGATIVE	Negative	STRONG	Bright
CD10	STRONG	Heterogeneous	WEAK	Dim
CD38	WEAK	Dim	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	WEAK	Dim	STRONG	Heterogeneous
	(1) FDE:13.32%;STRONG:CD10,WEAK:CD38,44,45,NEGATIVE:CD19,20,27,34; ……. .			
	(2) FDE:62.64%;STRONG:CD19,34,38,45,WEAK:CD10,44,NEGATIVE:CD20,27; ……. .			
	pos.3 neg.5 - 16.38 %, 1638 events (3)		**pos.5 neg.5 - 3.35 %, 335 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Medium	NEGATIVE	Negative
CD27	STRONG	Partially positive 2	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Bright	NEGATIVE	Negative
	(3) FDE:16.38%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .			
	(4) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 11						
	pos.1 neg.3 - 83.39 %, 8339 events (1)		**pos.3 neg.3 - 5.48 %, 548 events (2)**		**pos.5 neg.3 - 5.23 %, 523 events (3)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Heteroge-neous
CD27	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Bright
CD34	WEAK	Partially posi-tive 1	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative	WEAK	Dim
CD10	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Heteroge-neous	NEGATIVE	Negative	WEAK	Dim
CD20	WEAK	Partially posi-tive 1	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Bright
	(1) FDE:83.39%;STRONG:CD10,19,38,WEAK:CD20,34,NEGATIVE:CD27,44,45; ……. .					
	(2) FDE:5.48%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(3) FDE:5.23%;STRONG:CD27,44,45,WEAK:CD19,38,NEGATIVE:CD10,20,34; ……. .					

FILE 12						
	pos.5 neg.5 - 6.33 %, 633 events (1)		**pos.6 neg.5 - 83.68 %, 8368 events (2)**		**pos.8 neg.5 - 5.71 %, 571 events (3)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	WEAK	Dim	STRONG	Heteroge-neous
CD27	NEGATIVE	Negative	STRONG	Heteroge-neous	STRONG	Bright
CD34	NEGATIVE	Negative	WEAK	Partially posi-tive 1	NEGATIVE	Negative
CD19	NEGATIVE	Negative	STRONG	Bright	NEGATIVE	Negative
CD10	NEGATIVE	Negative	STRONG	Medium	NEGATIVE	Negative
CD38	NEGATIVE	Negative	STRONG	Bright	STRONG	Heteroge-neous
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Bright
	(1) FDE:6.33%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(2) FDE:83.68%;STRONG:CD10,19,27,38,WEAK:CD34,44,NEGATIVE:CD20,45; ……. .					
	(3) FDE:5.71%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .					

FILE 13				
	pos.1 neg.2 - 65.83 %, 6583 events (1)		**pos.2 neg.2 - 18.76 %, 1876 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Bright	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative
CD10	STRONG	Bright	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	WEAK	Dim	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:65.83%;STRONG:CD10,19,34,WEAK:CD20,NEGATIVE:CD27,38,44,45; ……. .			
	(2) FDE:18.76%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	pos.4 neg.2 - 4.42 %, 442 events (3)		**pos.6 neg.2 - 3.74 %, 374 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	STRONG	Bright
CD19	STRONG	Bright	STRONG	Bright
CD10	NEGATIVE	Negative	STRONG	Bright
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	STRONG	Bright	STRONG	Partially positive 2
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(3) FDE:<5%;STRONG:CD19,20,WEAK:CD44,NEGATIVE:CD10,27,34,38,45; ……. .			
	(4) FDE:<5%;STRONG:CD10,19,20,34,NEGATIVE:CD27,38,44,45; ……. .			

FILE 14						
	pos.1 neg.2 - 83.88 %, 8388 events (1)		**pos.2 neg.2 - 4.52 %, 452 events (2)**		**pos.3 neg.2 - 3.15 %, 315 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Heterogeneous
CD27	STRONG	Heterogeneous	NEGATIVE	Negative	STRONG	Bright
CD34	STRONG	Heterogeneous	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD10	STRONG	Bright	NEGATIVE	Negative	WEAK	Dim
CD38	STRONG	Heterogeneous	NEGATIVE	Negative	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	WEAK	Dim	NEGATIVE	Negative	STRONG	Bright
	(1) FDE:83.88%;STRONG:CD10,19,27,34,38,WEAK:CD45,NEGATIVE:CD20,44; ……. .					
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(3) FDE:<5%;STRONG:CD27,38,44,45,WEAK:CD10,NEGATIVE:CD19,20,34; ……. .					

FILE 15						
	pos.1 neg.3 - 68.82 %, 6882 events (1)		**pos.2 neg.3 - 14.63 %, 1463 events (2)**		**pos.3 neg.3 - 6.81 %, 681 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	STRONG	Partially positive 2	NEGATIVE	Negative
CD27	STRONG	Heterogeneous	STRONG	Partially positive 2	NEGATIVE	Negative
CD34	STRONG	Heterogeneous	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD10	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	STRONG	Partially positive 2	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	STRONG	Bright	NEGATIVE	Negative
	(1) FDE:68.82%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .					
	(2) FDE:14.63%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .					
	(3) FDE:6.81%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					

FILE 16						
	pos.2 neg.4 - 79.65 %, 7965 events (1)		**pos.4 neg.4 - 3.76 %, 376 events (2)**		**pos.9 neg.4 - 5.13 %, 513 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Heterogeneous	NEGATIVE	Negative	STRONG	Bright
CD10	STRONG	Medium	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Medium	NEGATIVE	Negative	NEGATIVE	Negative
CD20	WEAK	Dim	NEGATIVE	Negative	STRONG	Bright
CD45	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:79.65%;STRONG:CD10,19,38,WEAK:CD20,NEGATIVE:CD27,34,44,45; ……. .					
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(3) FDE:5.13%;STRONG:CD19,20,NEGATIVE:CD10,27,34,38,44,45; ……. .					

FILE 17						
	pos.1 neg.6 - 72.74 %, 7274 events (1)		**pos.3 neg.6 - 7.53 %, 753 events (2)**		**pos.5 neg.6 - 3.2 %, 320 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	WEAK	Dim	WEAK	Partially positive 1
CD27	STRONG	Heterogeneous	STRONG	Bright	WEAK	Dim
CD34	STRONG	Bright	STRONG	Bright	WEAK	Dim
CD19	STRONG	Bright	STRONG	Bright	NEGATIVE	Negative
CD10	STRONG	Bright	STRONG	Bright	WEAK	Dim
CD38	STRONG	Heterogeneous	STRONG	Heterogeneous	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	WEAK	Dim	STRONG	Medium
CD45	WEAK	Dim	STRONG	Medium	STRONG	Bright
	(1) FDE:72.74%;STRONG:CD10,19,27,34,38,WEAK:CD45,NEGATIVE:CD20,44; ……. .					
	(2) FDE:7.53%;STRONG:CD10,19,27,34,38,45,WEAK:CD20,44; ……. .					
	(3) FDE:<5%;STRONG:CD20,38,45,WEAK:CD10,27,34,44,NEGATIVE:CD19; ……. .					
	pos.6 neg.6 - 5.88 %, 588 events (4)		**pos.7 neg.6 - 4.83 %, 483 events (5)**			
parameter	**WHO**	**Detail**	**WHO**	**Detail**		
CD44	NEGATIVE	Negative	STRONG	Bright		
CD27	NEGATIVE	Negative	STRONG	Bright		
CD34	NEGATIVE	Negative	WEAK	Dim		
CD19	NEGATIVE	Negative	WEAK	Dim		
CD10	NEGATIVE	Negative	WEAK	Dim		
CD38	NEGATIVE	Negative	STRONG	Heterogeneous		
CD20	NEGATIVE	Negative	NEGATIVE	Negative		
CD45	NEGATIVE	Negative	STRONG	Bright		
	(4) FDE:5.88%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(5) FDE:<5%;STRONG:CD27,38,44,45,WEAK:CD10,19,34,NEGATIVE:CD20; ……. .					

FILE 18				
	pos.1 neg.2 - 78.79 %, 7879 events (1)		**pos.2 neg.2 - 3.74 %, 374 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	WEAK	Partially positive 1	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	WEAK	Dim	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:78.79%;,WEAK:CD10,20,NEGATIVE:CD19,27,34,38,44,45; ……. .			
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	pos.6 neg.2 - 4.31 %, 431 events (3)		**pos.7 neg.2 - 3.09 %, 309 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	WEAK	Dim	STRONG	Bright
CD19	NEGATIVE	Negative	WEAK	Dim
CD10	WEAK	Dim	STRONG	Bright
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	NEGATIVE	Negative	STRONG	Bright
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(3) FDE:<5%;,WEAK:CD10,34,44,NEGATIVE:CD19,20,27,38,45; ……. .			
	(4) FDE:<5%;STRONG:CD10,20,34,WEAK:CD19,NEGATIVE:CD27,38,44,45; ……. .			

FILE 19				
	pos.1 neg.4 - 80.6 %, 8060 events (1)		**pos.2 neg.4 - 4.58 %, 458 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	STRONG	Medium
CD27	WEAK	Dim	STRONG	Bright
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Medium	NEGATIVE	Negative
CD10	STRONG	Medium	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	STRONG	Partially positive 2
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	STRONG	Bright
	(1) FDE:80.36%;STRONG:CD10,19,38,WEAK:CD27,NEGATIVE:CD20,34,44,45; ……. .			
	(2) FDE:<5%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .			
	pos.3 neg.4 - 4.32 %, 432 events (3)		**pos.4 neg.4 - 3.26 %, 326 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Partially positive 1	NEGATIVE	Negative
CD27	WEAK	Dim	NEGATIVE	Negative
CD34	WEAK	Dim	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	WEAK	Dim	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Bright	NEGATIVE	Negative
	(3) FDE:<5%;STRONG:CD45,WEAK:CD27,34,38,44,NEGATIVE:CD10,19,20; ……. .			
	(4) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

Table 9 - Immunophenotype determined by the report_all_clusters_fSOM funtion for all FlowSOM clusters bigger than 3 % of all files. The clustering was chosen automatically from iterations of clustering with the range of dimensions from 6 to 16 in chapter 2.5.2. For each cluster, its number, the number of the negative reference cluster, its population percentage and total number of events is shown (in the format "pos.1 neg.2 - 45.85 %, 917events"). The phenotype is reported for each marker in the WHO standard and detail (describing the histogram distribution of the population). The FDE format of report of immunophenotype is shown on the bottom of each file section.

FILE 1						
	pos.1 neg.2 - 32.49 %, 3249 events (1)		**pos.2 neg.2 - 6.7 %, 670 events (2)**		**pos.4 neg.2 - 8.5 %, 850 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Bright
CD27	STRONG	Partially positive 2	NEGATIVE	Negative	WEAK	Dim
CD34	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative	STRONG	Bright
CD10	STRONG	Bright	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	NEGATIVE	Negative	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Bright
CD45	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Bright
	(1) FDE:32.49%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .					
	(2) FDE:6.7%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(3) FDE:8.5%;STRONG:CD19,20,38,44,45,WEAK:CD27,NEGATIVE:CD10,34; ……. .					
	pos.5 neg.2 - 46.46 %, 4646 events (4)		**pos.8 neg.2 - 3.12 %, 312 events (5)**			
parameter	**WHO**	**Detail**	**WHO**	**Detail**		
CD44	STRONG	Medium	STRONG	Medium		
CD27	STRONG	Partially positive 2	NEGATIVE	Negative		
CD34	NEGATIVE	Negative	NEGATIVE	Negative		
CD19	NEGATIVE	Negative	NEGATIVE	Negative		
CD10	NEGATIVE	Negative	WEAK	Dim		
CD38	STRONG	Partially positive 2	STRONG	Medium		
CD20	NEGATIVE	Negative	NEGATIVE	Negative		
CD45	STRONG	Bright	STRONG	Partially positive 2		
	(4) FDE:46.46%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .					
	(5) FDE:<5%;STRONG:CD38,44,45,WEAK:CD10,NEGATIVE:CD19,20,27,34; ……. .					

FILE 2						
	pos.1 neg.1 - 5.46 %, 546 events (1)		**pos.3 neg.1 - 4.3 %, 430 events (2)**		**pos.6 neg.1 - 87.08 %, 8708 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	WEAK	Partially positive 1	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Medium
CD19	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Heterogeneous
CD10	NEGATIVE	Negative	WEAK	Dim	STRONG	Heterogeneous
CD38	NEGATIVE	Negative	WEAK	Partially positive 1	NEGATIVE	Negative
CD20	NEGATIVE	Negative	WEAK	Dim	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:5.46%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(2) FDE:<5%;,WEAK:CD10,20,38,44,NEGATIVE:CD19,27,34,45; ……. .					
	(3) FDE:87.08%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .					

FILE 3						
	pos.1 neg.2 - 87.33 %, 8733 events (1)		**pos.2 neg.2 - 3.08 %, 308 events (2)**		**pos.6 neg.2 - 5.3 %, 530 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	NEGATIVE	Negative	STRONG	Medium
CD27	WEAK	Dim	NEGATIVE	Negative	STRONG	Partially positive 2
CD34	STRONG	Partially positive 2	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative	WEAK	Dim
CD10	STRONG	Bright	NEGATIVE	Negative	WEAK	Dim
CD38	WEAK	Dim	NEGATIVE	Negative	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:87.33%;STRONG:CD10,19,34,WEAK:CD27,38,44,NEGATIVE:CD20,45; ……. .					
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(3) FDE:5.3%;STRONG:CD27,44,WEAK:CD10,19,NEGATIVE:CD20,34,38,45; ……. .					

FILE 4				
	pos.1 neg.4 - 3.18 %, 318 events (1)		**pos.2 neg.4 - 84.3 %, 8430 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	NEGATIVE	Negative
CD27	STRONG	Partially positive 2	NEGATIVE	Negative
CD34	STRONG	Heterogeneous	STRONG	Medium
CD19	STRONG	Bright	STRONG	Bright
CD10	STRONG	Bright	STRONG	Bright
CD38	STRONG	Heterogeneous	WEAK	Dim
CD20	STRONG	Heterogeneous	STRONG	Heterogeneous
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:<5%;STRONG:CD10,19,20,27,34,38,WEAK:CD44,NEGATIVE:CD45; ……. .			
	(2) FDE:84.3%;STRONG:CD10,19,20,34,WEAK:CD38,NEGATIVE:CD27,44,45; ……. .			
	pos.4 neg.4 - 4.63 %, 463 events (3)		**pos.5 neg.4 - 4.38 %, 438 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	STRONG	Bright
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(3) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	(4) FDE:<5%;STRONG:CD27,NEGATIVE:CD10,19,20,34,38,44,45; ……. .			

FILE 5				
	pos.1 neg.3 - 68.2 %, 6820 events (1)		**pos.2 neg.3 - 5.17 %, 517 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	STRONG	Heterogeneous
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Medium	STRONG	Medium
CD19	STRONG	Medium	STRONG	Medium
CD10	WEAK	Partially positive 1	STRONG	Heterogeneous
CD38	WEAK	Dim	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	WEAK	Dim	STRONG	Heterogeneous
	(1) FDE:68.2%;STRONG:CD19,34,WEAK:CD10,38,45,NEGATIVE:CD20,27,44; ……. .			
	(2) FDE:5.17%;STRONG:CD10,19,34,38,44,45,NEGATIVE:CD20,27; ……. .			
	pos.3 neg.3 - 8.38 %, 838 events (3)		**pos.4 neg.3 - 11.75 %, 1175 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	STRONG	Partially positive 2
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	NEGATIVE	Negative	WEAK	Dim
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	WEAK	Dim
	(3) FDE:8.38%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	(4) FDE:11.75%;STRONG:CD44,WEAK:CD38,45,NEGATIVE:CD10,19,20,27,34; ……. .			

FILE 6				
	pos.1 neg.5 - 78.82 %, 7882 events (1)		**pos.2 neg.5 - 3.45 %, 345 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	STRONG	Medium
CD27	NEGATIVE	Negative	WEAK	Dim
CD34	WEAK	Dim	WEAK	Dim
CD19	STRONG	Medium	STRONG	Bright
CD10	WEAK	Dim	STRONG	Heterogeneous
CD38	STRONG	Medium	STRONG	Bright
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Medium	STRONG	Bright
	(1) FDE:78.82%;STRONG:CD19,38,45,WEAK:CD10,34,44,NEGATIVE:CD20,27; ……. .			
	(2) FDE:<5%;STRONG:CD10,19,38,44,45,WEAK:CD27,34,NEGATIVE:CD20; ……. .			
	pos.4 neg.5 - 5.78 %, 578 events (3)		**pos.5 neg.5 - 7.12 %, 712 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Partially positive 1	NEGATIVE	Negative
CD27	WEAK	Dim	NEGATIVE	Negative
CD34	WEAK	Dim	NEGATIVE	Negative
CD19	WEAK	Dim	NEGATIVE	Negative
CD10	STRONG	Heterogeneous	NEGATIVE	Negative
CD38	WEAK	Dim	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Medium	NEGATIVE	Negative
	(3) FDE:5.78%;STRONG:CD10,45,WEAK:CD19,27,34,38,44,NEGATIVE:CD20; ……. .			
	(4) FDE:7.12%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 7						
	pos.1 neg.9 - 56.74 %, 5674 events (1)		**pos.2 neg.9 - 34.01 %, 3401 events (2)**		**pos.9 neg.9 - 4.1 %, 410 events (3)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Medium	WEAK	Dim	NEGATIVE	Negative
CD19	STRONG	Medium	NEGATIVE	Negative	NEGATIVE	Negative
CD10	STRONG	Medium	STRONG	Heteroge-neous	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:56.74%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .					
	(2) FDE:34.01%;STRONG:CD10,WEAK:CD34,NEGATIVE:CD19,20,27,38,44,45; ……. .					
	(3) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					

FILE 8				
	pos.2 neg.4 - 92.87 %, 9287 events (1)		**pos.4 neg.4 - 3.36 %, 336 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Medium	NEGATIVE	Negative
CD19	STRONG	Medium	NEGATIVE	Negative
CD10	STRONG	Bright	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:92.87%;STRONG:CD10,19,34,38,NEGATIVE:CD20,27,44,45; ……. .			
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 9						
	pos.1 neg.2 - 64.66 %, 6466 events (1)		**pos.2 neg.2 - 3.29 %, 329 events (2)**		**pos.3 neg.2 - 9.39 %, 939 events (3)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	NEGATIVE	Negative	STRONG	Heteroge-neous
CD27	WEAK	Dim	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Medium	NEGATIVE	Negative	NEGATIVE	Negative
CD10	STRONG	Medium	NEGATIVE	Negative	WEAK	Dim
CD38	STRONG	Medium	NEGATIVE	Negative	WEAK	Partially posi-tive 1
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Partially posi-tive 2
	(1) FDE:64.66%;STRONG:CD10,19,38,WEAK:CD27,44,NEGATIVE:CD20,34,45; ……. .					
	(2) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	(3) FDE:9.39%;STRONG:CD44,45,WEAK:CD10,38,NEGATIVE:CD19,20,27,34; ……. .					
	pos.4 neg.2 - 12.14 %, 1214 events (4)		**pos.6 neg.2 - 7.71 %, 771 events (5)**			
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**		
CD44	STRONG	Medium	STRONG	Medium		
CD27	STRONG	Bright	NEGATIVE	Negative		
CD34	NEGATIVE	Negative	NEGATIVE	Negative		
CD19	NEGATIVE	Negative	STRONG	Bright		
CD10	NEGATIVE	Negative	NEGATIVE	Negative		
CD38	STRONG	Heteroge-neous	STRONG	Heteroge-neous		
CD20	NEGATIVE	Negative	STRONG	Bright		
CD45	STRONG	Bright	STRONG	Bright		
	(4) FDE:12.14%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .					
	(5) FDE:7.71%;STRONG:CD19,20,38,44,45,NEGATIVE:CD10,27,34; ……. .					

FILE 10						
	pos.1 neg.3 - 15.47 %, 1547 events (1)		**pos.2 neg.3 - 62.2 %, 6220 events (2)**		**pos.3 neg.3 - 3.09 %, 309 events (3)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Medium	WEAK	Dim	NEGATIVE	Negative
CD27	STRONG	Medium	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	STRONG	Bright	NEGATIVE	Negative
CD19	NEGATIVE	Negative	STRONG	Bright	NEGATIVE	Negative
CD10	NEGATIVE	Negative	WEAK	Dim	NEGATIVE	Negative
CD38	WEAK	Dim	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Bright	STRONG	Heterogeneous	NEGATIVE	Negative
	(1) FDE:15.47%;STRONG:CD27,44,45,WEAK:CD38,NEGATIVE:CD10,19,20,34; ……. .					
	(2) FDE:62.2%;STRONG:CD19,34,38,45,WEAK:CD10,44,NEGATIVE:CD20,27; ……. .					
	(3) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					
	pos.4 neg.3 - 6.1 %, 610 events (4)		**pos.5 neg.3 - 3.26 %, 326 events (5)**		**pos.6 neg.3 - 7.57 %, 757 events (6)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Dim	WEAK	Partially positive 1	WEAK	Partially posi-tive 1
CD27	NEGATIVE	Negative	WEAK	Dim	NEGATIVE	Negative
CD34	NEGATIVE	Negative	STRONG	Partially positive 2	NEGATIVE	Negative
CD19	NEGATIVE	Negative	STRONG	Partially positive 2	NEGATIVE	Negative
CD10	STRONG	Heterogeneous	WEAK	Partially positive 1	STRONG	Heteroge-neous
CD38	NEGATIVE	Negative	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	NEGATIVE	Negative	STRONG	Heterogeneous	NEGATIVE	Negative
CD45	WEAK	Dim	STRONG	Heterogeneous	STRONG	Medium
	(4) FDE:6.1%;STRONG:CD10,WEAK:CD44,45,NEGATIVE:CD19,20,27,34,38; ……. .					
	(5) FDE:<5%;STRONG:CD19,20,34,38,45,WEAK:CD10,27,44; ……. .					
	(6) FDE:7.57%;STRONG:CD10,45,WEAK:CD44,NEGATIVE:CD19,20,27,34,38; ……. .					

FILE 11						
	pos.1 neg.4 - 85.29 %, 8529 events (1)		**pos.3 neg.4 - 6.83 %, 683 events (2)**		**pos.4 neg.4 - 6.19 %, 619 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	STRONG	Heterogeneous	NEGATIVE	Negative
CD27	NEGATIVE	Negative	STRONG	Partially positive 2	NEGATIVE	Negative
CD34	WEAK	Partially positive 1	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	WEAK	Dim	NEGATIVE	Negative
CD10	STRONG	Bright	WEAK	Dim	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	WEAK	Partially positive 1	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	STRONG	Medium	NEGATIVE	Negative
	(1) FDE:85.29%;STRONG:CD10,19,38,WEAK:CD20,34,NEGATIVE:CD27,44,45; ……. .					
	(2) FDE:6.83%;STRONG:CD27,38,44,45,WEAK:CD10,19,NEGATIVE:CD20,34; ……. .					
	(3) FDE:6.19%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					

FILE 12				
	pos.1 neg.7 - 71.58 %, 7158 events (1)		**pos.4 neg.7 - 11.48 %, 1148 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	WEAK	Dim
CD27	STRONG	Heterogeneous	STRONG	Heterogeneous
CD34	STRONG	Partially positive 2	WEAK	Partially positive 1
CD19	STRONG	Bright	STRONG	Bright
CD10	STRONG	Bright	STRONG	Bright
CD38	STRONG	Bright	STRONG	Bright
CD20	NEGATIVE	Negative	STRONG	Heterogeneous
CD45	NEGATIVE	Negative	WEAK	Dim
	(1) FDE:71.58%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .			
	(2) FDE:11.48%;STRONG:CD10,19,20,27,38,WEAK:CD34,44,45; ……. .			
	pos.7 neg.7 - 5.84 %, 584 events (3)		**pos.8 neg.7 - 6.18 %, 618 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	STRONG	Heterogeneous
CD27	NEGATIVE	Negative	STRONG	Heterogeneous
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	WEAK	Dim
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	NEGATIVE	Negative	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	STRONG	Bright
	(3) FDE:5.84%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	(4) FDE:6.18%;STRONG:CD27,38,44,45,WEAK:CD19,NEGATIVE:CD10,20,34; ……. .			

FILE 13						
	pos.1 neg.8 - 19.03 %, 1903 events (1)		**pos.2 neg.8 - 4.55 %, 455 events (2)**		**pos.3 neg.8 - 5.7 %, 570 events (3)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Medium	STRONG	Medium	STRONG	Heterogeneous
CD27	STRONG	Medium	WEAK	Dim	WEAK	Partially positive 1
CD34	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Medium
CD19	NEGATIVE	Negative	STRONG	Bright	STRONG	Bright
CD10	NEGATIVE	Negative	NEGATIVE	Negative	STRONG	Medium
CD38	STRONG	Partially positive 2	STRONG	Heterogeneous	STRONG	Bright
CD20	NEGATIVE	Negative	STRONG	Bright	WEAK	Dim
CD45	STRONG	Medium	STRONG	Medium	WEAK	Partially positive 1
	(1) FDE:19.03%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .					
	(2) FDE:<5%;STRONG:CD19,20,38,44,45,WEAK:CD27,NEGATIVE:CD10,34; ……. .					
	(3) FDE:5.7%;STRONG:CD10,19,34,38,44,WEAK:CD20,27,45; ……. .					
	pos.6 neg.8 - 64.92 %, 6492 events (4)		**pos.8 neg.8 - 3.03 %, 303 events (5)**			
parameter	**WHO**	**Detail**	**WHO**	**Detail**		
CD44	WEAK	Dim	NEGATIVE	Negative		
CD27	NEGATIVE	Negative	NEGATIVE	Negative		
CD34	STRONG	Medium	NEGATIVE	Negative		
CD19	STRONG	Medium	NEGATIVE	Negative		
CD10	STRONG	Medium	NEGATIVE	Negative		
CD38	STRONG	Heterogeneous	NEGATIVE	Negative		
CD20	NEGATIVE	Negative	NEGATIVE	Negative		
CD45	NEGATIVE	Negative	NEGATIVE	Negative		
	(4) FDE:64.92%;STRONG:CD10,19,34,38,WEAK:CD44,NEGATIVE:CD20,27,45; ……. .					
	(5) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					

FILE 14						
	pos.1 neg.9 - 78.44 %, 7844 events (1)		**pos.2 neg.9 - 3.12 %, 312 events (2)**		**pos.4 neg.9 - 3.76 %, 376 events (3)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD27	STRONG	Heterogeneous	STRONG	Partially positive 2	WEAK	Dim
CD34	STRONG	Heterogeneous	STRONG	Heterogeneous	STRONG	Heteroge-neous
CD19	STRONG	Medium	STRONG	Bright	STRONG	Bright
CD10	STRONG	Bright	STRONG	Bright	STRONG	Bright
CD38	STRONG	Heterogeneous	STRONG	Bright	STRONG	Bright
CD20	NEGATIVE	Negative	NEGATIVE	Negative	WEAK	Dim
CD45	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:78.44%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .					
	(2) FDE:<5%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .					
	(3) FDE:<5%;STRONG:CD10,19,34,38,WEAK:CD20,27,NEGATIVE:CD44,45; ……. .					
	pos.6 neg.9 - 3.69 %, 369 events (4)		**pos.8 neg.9 - 3.31 %, 331 events (5)**		**pos.9 neg.9 - 5.25 %, 525 events (6)**	
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD44	WEAK	Partially positive 1	STRONG	Heterogeneous	NEGATIVE	Negative
CD27	NEGATIVE	Negative	STRONG	Heterogeneous	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD10	WEAK	Dim	WEAK	Dim	NEGATIVE	Negative
CD38	WEAK	Dim	WEAK	Dim	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Medium	STRONG	Medium	NEGATIVE	Negative
	(4) FDE:<5%;STRONG:CD45,WEAK:CD10,38,44,NEGATIVE:CD19,20,27,34; ……. .					
	(5) FDE:<5%;STRONG:CD27,44,45,WEAK:CD10,38,NEGATIVE:CD19,20,34; ……. .					
	(6) FDE:5.25%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .					

FILE 15				
	pos.1 neg.8 - 15.8 %, 1580 events (1)		**pos.6 neg.8 - 69.61 %, 6961 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Partially positive 2	NEGATIVE	Negative
CD27	STRONG	Partially positive 2	STRONG	Heterogeneous
CD34	NEGATIVE	Negative	STRONG	Heterogeneous
CD19	NEGATIVE	Negative	STRONG	Bright
CD10	NEGATIVE	Negative	STRONG	Bright
CD38	STRONG	Partially positive 2	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Bright	NEGATIVE	Negative
	(1) FDE:15.8%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .			
	(2) FDE:69.61%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .			
	pos.7 neg.8 - 3.18 %, 318 events (3)		**pos.8 neg.8 - 6.72 %, 672 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Medium	NEGATIVE	Negative
CD27	WEAK	Dim	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	STRONG	Bright	NEGATIVE	Negative
CD45	STRONG	Bright	NEGATIVE	Negative
	(3) FDE:<5%;STRONG:CD19,20,38,44,45,WEAK:CD27,NEGATIVE:CD10,34; ……. .			
	(4) FDE:6.72%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 16				
	pos.1 neg.1 - 4.25 %, 425 events (1)		**pos.3 neg.1 - 78.67 %, 7867 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	STRONG	Heterogeneous
CD10	NEGATIVE	Negative	STRONG	Medium
CD38	NEGATIVE	Negative	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	WEAK	Dim
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	(2) FDE:78.67%;STRONG:CD10,19,38,WEAK:CD20,NEGATIVE:CD27,34,44,45; ……. .			
	pos.4 neg.1 - 5.77 %, 577 events (3)		**pos.6 neg.1 - 6.63 %, 663 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	STRONG	Bright	STRONG	Bright
CD10	WEAK	Dim	STRONG	Bright
CD38	WEAK	Partially positive 1	STRONG	Bright
CD20	STRONG	Bright	STRONG	Heterogeneous
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(3) FDE:5.77%;STRONG:CD19,20,WEAK:CD10,38,NEGATIVE:CD27,34,44,45; ……. .			
	(4) FDE:6.63%;STRONG:CD10,19,20,38,NEGATIVE:CD27,34,44,45; ……. .			

FILE 17				
	pos.1 neg.2 - 5.23 %, 523 events (1)		**pos.2 neg.2 - 6.25 %, 625 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Medium	NEGATIVE	Negative
CD27	STRONG	Medium	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	NEGATIVE	Negative	NEGATIVE	Negative
CD38	STRONG	Heterogeneous	NEGATIVE	Negative
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	STRONG	Bright	NEGATIVE	Negative
	(1) FDE:5.23%;STRONG:CD27,38,44,45,NEGATIVE:CD10,19,20,34; ……. .			
	(2) FDE:6.25%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	pos.3 neg.2 - 80.48 %, 8048 events (3)		**pos.5 neg.2 - 4.45 %, 445 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	WEAK	Partially positive 1
CD27	STRONG	Heterogeneous	NEGATIVE	Negative
CD34	STRONG	Bright	NEGATIVE	Negative
CD19	STRONG	Bright	NEGATIVE	Negative
CD10	STRONG	Bright	WEAK	Dim
CD38	STRONG	Heterogeneous	STRONG	Partially positive 2
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	STRONG	Medium
	(3) FDE:80.48%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .			
	(4) FDE:<5%;STRONG:CD38,45,WEAK:CD10,44,NEGATIVE:CD19,20,27,34; ……. .			

FILE 18				
	pos.1 neg.7 - 41.42 %, 4142 events (1)		**pos.2 neg.7 - 40.27 %, 4027 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	NEGATIVE	Negative
CD10	STRONG	Partially positive 2	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	WEAK	Dim	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(1) FDE:41.42%;STRONG:CD10,WEAK:CD20,NEGATIVE:CD19,27,34,38,44,45; ……. .			
	(2) FDE:40.27%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	pos.5 neg.7 - 3.95 %, 395 events (3)		**pos.7 neg.7 - 3.91 %, 391 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	NEGATIVE	Negative
CD34	STRONG	Heterogeneous	NEGATIVE	Negative
CD19	WEAK	Dim	NEGATIVE	Negative
CD10	STRONG	Bright	NEGATIVE	Negative
CD38	NEGATIVE	Negative	NEGATIVE	Negative
CD20	STRONG	Medium	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(3) FDE:<5%;STRONG:CD10,20,34,WEAK:CD19,NEGATIVE:CD27,38,44,45; ……. .			
	(4) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			

FILE 19				
	pos.2 neg.4 - 5.64 %, 564 events (1)		**pos.3 neg.4 - 5.33 %, 533 events (2)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	STRONG	Partially positive 2	STRONG	Medium
CD27	WEAK	Dim	STRONG	Medium
CD34	WEAK	Dim	NEGATIVE	Negative
CD19	NEGATIVE	Negative	WEAK	Partially positive 1
CD10	WEAK	Dim	NEGATIVE	Negative
CD38	WEAK	Dim	STRONG	Partially positive 2
CD20	WEAK	Dim	NEGATIVE	Negative
CD45	STRONG	Partially positive 2	STRONG	Bright
	(1) FDE:5.64%;STRONG:CD44,45,WEAK:CD10,20,27,34,38,NEGATIVE:CD19; ……. .			
	(2) FDE:5.33%;STRONG:CD27,38,44,45,WEAK:CD19,NEGATIVE:CD10,20,34; ……. .			
	pos.4 neg.4 - 3.04 %, 304 events (3)		**pos.5 neg.4 - 82.57 %, 8257 events (4)**	
parameter	**WHO**	**Detail**	**WHO**	**Detail**
CD44	NEGATIVE	Negative	NEGATIVE	Negative
CD27	NEGATIVE	Negative	WEAK	Dim
CD34	NEGATIVE	Negative	NEGATIVE	Negative
CD19	NEGATIVE	Negative	STRONG	Bright
CD10	NEGATIVE	Negative	STRONG	Bright
CD38	NEGATIVE	Negative	STRONG	Heterogeneous
CD20	NEGATIVE	Negative	NEGATIVE	Negative
CD45	NEGATIVE	Negative	NEGATIVE	Negative
	(3) FDE:<5%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .			
	(4) FDE:82.57%;STRONG:CD10,19,38,WEAK:CD27,NEGATIVE:CD20,34,44,45; ……. .			

5.5 Comparison of automatic and clinical immunophenotype report

The clinical database contains various information about the processing and source of the individual samples. For some samples, the official immunophenotype report in FDE format is present. This PostgreSQL relational database was accessed and the data for each analysed patient was exported to xlsx. For each of the 19 files processed in chapters 5.1, 5.3 and 5.4 the xlsx file was searched manually for the reported immunophenotype. Because the selected files were diagnostic, the FDE report was present for all of them, except for file 18. However, the cell population structure in file 18 is quite atypical and based on the notes in the database, the diagnosis was concluded to be of a solid tumour, not ALL. The presence of the immunophenotype report in the clinical database for 18 out of 19 files confirms the correct choice of diagnostic files for the analysis.

For each file, the automatic reports from HCA and FlowSOM clustering of all clusters bigger than 3 % (see chapter 5.4) were viewed over and the candidate B-ALL blastic population was chosen manually based on the reported immunophenotype and size. Its HCA and FlowSOM immmunophenotype report in FDE format was then compared with the official reported

immunophenotype from the clinical database (see Table 10). The complete comparison is shown in Table 11. Some markers were missing in the clinical FDE report and therefore could not be compared.

Table 10 - FDE immunophenotype reports for files 1-19 from CLIP clinical database and automatic FDE reports of manually selected candidate B-ALL blastic populations based on HCA and FlowSOM clustering in chapter 5.4. Range of clusters set for clustering from which the ideal cluster was chosen was 5-100 for HCA. The range of dimensions of clustering grid for FlowSOM was 6-16.

FILE 1 - FDE	
Clinical immunophenotype	FDE:32%;STRONG:CD10,19,27,34,38, WEAK:44, NEGATIVE: 20; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:31.48%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:32.49%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .
FILE 2 - FDE	
Clinical immunophenotype	FDE:85%;STRONG:CD10,19,34,WEAK:CD38,NEGATIVE:20; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:87.37%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:87.08%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .
FILE 3 - FDE	
Clinical immunophenotype	FDE:77%; STRONG:CD10,19,34,38, NEGATIVE:20; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:89.57%;STRONG:CD10,19,34,WEAK:CD38,NEGATIVE:CD20,27,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:87.33%;STRONG:CD10,19,34,WEAK:CD27,38,44,NEGATIVE:CD20,45; ……. .
FILE 4 - FDE	
Clinical immunophenotype	FDE:85%;STRONG:CD10,19,20,34,38,44, WEAK:CD27; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:86.18%;STRONG:CD10,19,20,WEAK:CD34,38,NEGATIVE:CD27,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:84.3%;STRONG:CD10,19,20,34,WEAK:CD38,NEGATIVE:CD27,44,45; ……. .
FILE 5 - FDE	
Clinical immunophenotype	FDE:83%;STRONG:CD10,19,34,38,44, NEGATIVE: 20,27; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:70.84%;STRONG:CD19,34,WEAK:CD10,38,45,NEGATIVE:CD20,27,44; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:68.2%;STRONG:CD19,34,WEAK:CD10,38,45,NEGATIVE:CD20,27,44; ……. .
FILE 6 - FDE	
Clinical immunophenotype	FDE:78%;STRONG:CD19,38, WEAK: 10,NEGATIVE: 20,34; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:79.9%;STRONG:CD19,38,WEAK:CD10,34,44,45,NEGATIVE:CD20,27; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:78.82%;STRONG:CD19,38,45,WEAK:CD10,34,44,NEGATIVE:CD20,27; ……. .

FILE 7 - FDE	
Clinical immunophenotype	FDE:91%;STRONG:CD10,19, 34,38, NEGATIVE: 20,27, 44; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:90.59%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:56.74%;STRONG:CD10,19,34,NEGATIVE:CD20,27,38,44,45; ……. .
FILE 8 - FDE	
Clinical immunophenotype	FDE:90%;STRONG:CD10,19,34,38, WEAK: 20; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:94.19%;STRONG:CD10,19,34,38,NEGATIVE:CD20,27,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:92.87%;STRONG:CD10,19,34,38,NEGATIVE:CD20,27,44,45; ……. .
FILE 9 - FDE	
Clinical immunophenotype	FDE:70%;STRONG:CD10,19, 38,44, WEAK: 27, NEGATIVE: 20, 34; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:59.49%;STRONG:CD10,19,38,WEAK:CD27,44,NEGATIVE:CD20,34,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:64.66%;STRONG:CD10,19,38,WEAK:CD27,44,NEGATIVE:CD20,34,45; ……. .
FILE 10 - FDE	
Clinical immunophenotype	FDE:59,2%;STRONG:CD19, 34,38,44, WEAK: 20, NEGATIVE: 10, 27; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:62.64%;STRONG:CD19,34,38,45,WEAK:CD10,44,NEGATIVE:CD20,27; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:62.2%;STRONG:CD19,34,38,45,WEAK:CD10,44,NEGATIVE:CD20,27; ……. .
FILE 11 - FDE	
Clinical immunophenotype	FDE:87%;STRONG:CD10,19, 38 ,WEAK: 20, 34,44, NEGATIVE: 27; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:83.39%;STRONG:CD10,19,38,WEAK:CD20,34,NEGATIVE:CD27,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:85.29%;STRONG:CD10,19,38,WEAK:CD20,34,NEGATIVE:CD27,44,45; ……. .
FILE 12 - FDE	
Clinical immunophenotype	FDE:81%;STRONG:CD10,19, 34,38, WEAK:CD20; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:83.68%;STRONG:CD10,19,27,38,WEAK:CD34,44,NEGATIVE:CD20,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:71.58%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .
FILE 13 - FDE	
Clinical immunophenotype	FDE:76%;STRONG:CD10,19, 34,38, WEAK: 20,45; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:65.83%;STRONG:CD10,19,34,WEAK:CD20,NEGATIVE:CD27,38,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:64.92%;STRONG:CD10,19,34,38,WEAK:CD44,NEGATIVE:CD20,27,45; ……. .

FILE 14 - FDE	
Clinical immunophenotype	FDE:75% STRONG:,CD10,19, 27,34,38, WEAK:CD20, NEGATIVE: 44; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:83.88%;STRONG:CD10,19,27,34,38,WEAK:CD45,NEGATIVE:CD20,44; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:78.44%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .
FILE 15 - FDE	
Clinical immunophenotype	FDE:67%;STRONG:CD10,19, 27,34,38, WEAK: 44,NEGATIVE: 20; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:68.82%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:69.61%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .
FILE 16 - FDE	
Clinical immunophenotype	FDE:73%;STRONG:CD10,19, 38, WEAK:CD20, NEGATIVE: 34; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:79.65%;STRONG:CD10,19,38,WEAK:CD20,NEGATIVE:CD27,34,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:78.67%;STRONG:CD10,19,38,WEAK:CD20,NEGATIVE:CD27,34,44,45; ……. .
FILE 17 - FDE	
Clinical immunophenotype	FDE:79%;STRONG:CD10, 19, 27, 34,38, WEAK: 44,45, NEGATIVE: 20; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:72.74%;STRONG:CD10,19,27,34,38,WEAK:CD45,NEGATIVE:CD20,44; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:80.48%;STRONG:CD10,19,27,34,38,NEGATIVE:CD20,44,45; ……. .
FILE 18 - FDE	
Clinical immunophenotype	
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:78.79%;,WEAK:CD10,20,NEGATIVE:CD19,27,34,38,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:40.27%;,NEGATIVE:CD10,19,20,27,34,38,44,45; ……. .
FILE 19 - FDE	
Clinical immunophenotype	FDE:88%;STRONG:CD10,19, 38, WEAK: 44,45,NEGATIVE: 20,27, 34; ……. .
Automatic immunophenotype – HCA (range of clusters 5-100)	FDE:80.36%;STRONG:CD10,19,38,WEAK:CD27,NEGATIVE:CD20,34,44,45; ……. .
Automatic immunophenotype – FlowSOM (range of dim. 6-16)	FDE:82.57%;STRONG:CD10,19,38,WEAK:CD27,NEGATIVE:CD20,34,44,45; ……. .

Table 11 – Comparison of the automatic FDE reports of manually selected candidate B-ALL blastic populations for files 1-19 based on HCA and FlowSOM clustering in chapter 5.4 and the FDE immunophenotype report from CLIP clinical database. Positive markers are labeled with a "+", negative with a "–" and weak with a "W". Expression level of markers corresponding with the clinical database are highlighted in green colour, marker expressions that do not match in red and those that are close to the clinical phenotype in yellow (weak instead of positive or negative, or vice versa). Range of clusters set for clustering from which the cluster was selected was 5-100 for HCA. The range of dimensions of clustering grid for FlowSOM was 6-16. Note, that some markers were not reported in the CLIP database and cannot be compared with our results.

FILE 1								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	W	+	+	+	+	+	-	
HCA (range of clusters 5-100)	-	+	+	+	+	+	-	-
FlowSOM (range of dim. 6-16)	-	+	+	+	+	+	-	-
FILE 2								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical			+	+	+	W	-	
HCA (range of clusters 5-100)	-	-	+	+	+	-	-	-
FlowSOM (range of dim. 6-16)	-	-	+	+	+	-	-	-
FILE 3								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical			+	+	+	+	-	
HCA (range of clusters 5-100)	-	-	+	+	+	W	-	-
FlowSOM (range of dim. 6-16)	W	W	+	+	+	W	-	-
FILE 4								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	+	W	+	+	+	+	+	
HCA (range of clusters 5-100)	-	-	W	+	+	W	+	-
FlowSOM (range of dim. 6-16)	-	-	+	+	+	W	+	-
FILE 5								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	+	-	+	+	+	+	-	
HCA (range of clusters 5-100)	-	-	+	+	W	W	-	W
FlowSOM (range of dim. 6-16)	-	-	+	+	W	W	-	W
FILE 6								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical			-	+	W	+	-	
HCA (range of clusters 5-100)	W	-	W	+	W	+	-	W
FlowSOM (range of dim. 6-16)	W	-	W	+	W	+	-	+

FILE 7								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	-	-	+	+	+	+	-	
HCA (range of clusters 5-100)	-	-	+	+	+	-	-	-
FlowSOM (range of dim. 6-16)	-	-	+	+	+	-	-	-
FILE 8								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical			+	+	+	+	W	
HCA (range of clusters 5-100)	-	-	+	+	+	+	-	-
FlowSOM (range of dim. 6-16)	-	-	+	+	+	+	-	-
FILE 9								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	+	W	-	+	+	+	-	
HCA (range of clusters 5-100)	W	W	-	+	+	+	-	-
FlowSOM (range of dim. 6-16)	W	W	-	+	+	+	-	-
FILE 10								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	+	-	+	+	-	+	W	
HCA (range of clusters 5-100)	W	-	+	+	W	+	-	+
FlowSOM (range of dim. 6-16)	W	-	+	+	W	+	-	+
FILE 11								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	W	-	W	+	+	+	W	
HCA (range of clusters 5-100)	-	-	W	+	+	+	W	-
FlowSOM (range of dim. 6-16)	-	-	W	+	+	+	W	-
FILE 12								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical			+	+	+	+	W	
HCA (range of clusters 5-100)	W	+	W	+	+	+	-	-
FlowSOM (range of dim. 6-16)	-	+	+	+	+	+	-	-
FILE 13								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical			+	+	+	+	W	W
HCA (range of clusters 5-100)	-	-	+	+	+	-	W	-
FlowSOM (range of dim. 6-16)	W	-	+	+	+	+	-	-
FILE 14								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	-	+	+	+	+	+	W	
HCA (range of clusters 5-100)	-	+	+	+	+	+	-	W
FlowSOM (range of dim. 6-16)	-	+	+	+	+	+	-	-

FILE 15								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	W	+	+	+	+	+	-	
HCA (range of clusters 5-100)	-	+	+	+	+	+	-	-
FlowSOM (range of dim. 6-16)	-	+	+	+	+	+	-	-
FILE 16								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical			-	+	+	+	W	
HCA (range of clusters 5-100)	-	-	-	+	+	+	W	-
FlowSOM (range of dim. 6-16)	-	-	-	+	+	+	W	-
FILE 17								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	W	+	+	+	+	+	-	W
HCA (range of clusters 5-100)	-	+	+	+	+	+	-	W
FlowSOM (range of dim. 6-16)	-	+	+	+	+	+	-	-
FILE 18								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical								
HCA (range of clusters 5-100)	+	+	+	+	W	+	W	+
FlowSOM (range of dim. 6-16)	+	+	+	+	+	+	+	+
FILE 19								
Immonophenotype	**CD44**	**CD27**	**CD34**	**CD19**	**CD10**	**CD38**	**CD20**	**CD45**
Clinical	W	-	-	+	+	+	-	W
HCA (range of clusters 5-100)	-	W	-	+	+	+	-	-
FlowSOM (range of dim. 6-16)	-	W	-	+	+	+	-	-

5.6 Development of immunophenotype during the course of ALL

To characterise the immunophenotype development of the blastic population in ALL patients during the course of the disease, we searched for suitable files of the patients whose diagnostic files were used in the previous analysis. The choice of the files was based on the database of metainformation acquired from all files in the CLIP database (see chapters 5.1, 4.4, 2.3). The files were chosen based on the identical patient ID to the diagnostic files analysed in chapters 5.1, 5.3, 5.4 and 5.5, the acquisition date of the files corresponding with the day 0 (D0), day 8 (D8) and day 15 (D15) of the treatment and an identical panel for minimal residual disease (MRD) monitoring on D0, D8 and D15. An MRD panel CD58 CD66c CD10 CD45 CD19 CD20 CD34 CD38 was present in D0, D8 and D15 files of the patients 7, 8, 12, 16 and 19. These files were used for the analysis.

The selected files were preprocessed (compensation, logicle transformation and z-score normalisation) and HCA clustered into a dendrogram identically to the analysis of the diagnostic files. The `choose_cluster_hca` function was used the same way as in the diagnostic files analysis in chapter 5.3.1. The range of clusters for the iteration of cluster selection was set to 5-100 clusters and the threshold for the size of populations was set to 2 % instead of 3 %. The chosen iteration of clustering was used for the immunophenotype report by the function `report_all_clusters_hca` which was also used in immunophenotype report of the diagnostic files (see chapter 5.4).

The immunophenotype of the candidate blast population for the D0, D8 and D15 of treatment for patients was selected manually and is shown in Table 12. However, for three out of the five analysed patient datasets, the MRD level in D8 and D15 follow up files was under the detection threshold of the method. This is the reason, why the immunophenotype of D0, D8 and D15 of treatment is only shown for two patients. Furthermore, the comparison of immunophenotype of MRD on day 0 and day 8 of the treatment of patients 8 and 19 and the percentage of the cell population is presented in Table 13.

Table 12 - Immunophenotypes of MRD on day 0, 8 and 15 of treatment of patients 8 and 19 reported by report_all_clusters_hca function based on the resulting iteration of clustering by choose_cluster_hca function (range of clusters set to 5-100, threshold of the size of the clusters to 2 %). The FDE report, number of clusters of selected iteration of clustering, number of reported MRD cluster, number of reference negative cluster, population percentage and number of events are also shown. For day 15, the blast population was not detectable for both patients.

PATIENT 8						
	D0 - pos.1 neg.6 - 81.7 %, 8170 events (1)		**D8 - pos.2 neg.3 - 13.36 %, 1336 events (2)**		**D15 (3)**	
	Number of clusters - 97		Number of clusters - 32			
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD58	STRONG	Heterogeneous	NEGATIVE	Negative		
CD66c	STRONG	Heterogeneous	STRONG	Heterogeneous		
CD10	STRONG	Bright	STRONG	Bright		
CD45	WEAK	Dim	NEGATIVE	Negative		
CD19	STRONG	Bright	STRONG	Bright		
CD20	STRONG	Heterogeneous	STRONG	Bright		
CD34	STRONG	Bright	WEAK	Dim		
CD38	STRONG	Heterogeneous	NEGATIVE	Negative		
	(1) FDE:81.7%;STRONG:CD10,19,20,34,38,58,66c,WEAK:CD45; ……. .					
	(2) FDE:13.36%;STRONG:CD10,19,20,66c,WEAK:CD34,NEGATIVE:CD38,45,58; ……. .					
	(3) UNDETECTABLE					

PATIENT 19						
	D0 - pos.1 neg.2 - 79.45 %, 7945 events (1)		**D8 - pos.13 neg.8 - 2.51 %, 251 events (2)**		**D15 (3)**	
	Number of clusters - 71		Number of clusters - 80			
parame-ter	**WHO**	**Detail**	**WHO**	**Detail**	**WHO**	**Detail**
CD58	STRONG	Medium	WEAK	Dim		
CD66c	NEGATIVE	Negative	NEGATIVE	Negative		
CD10	STRONG	Medium	STRONG	Partially positive 2		
CD45	WEAK	Dim	NEGATIVE	Negative		
CD19	STRONG	Medium	STRONG	Bright		
CD20	NEGATIVE	Negative	NEGATIVE	Negative		
CD34	NEGATIVE	Negative	NEGATIVE	Negative		
CD38	STRONG	Heteroge-neous	STRONG	Heterogeneous		
	(1) FDE:79.45%;STRONG:CD10,19,38,58,WEAK:CD45,NEGATIVE:CD20,34,66c; ……. .					
	(2) FDE:<5%;STRONG:CD10,19,38,WEAK:CD58,NEGATIVE:CD20,34,45,66c; ……. .					
	(3) UNDETECTABLE					

Table 13 - Comparison of MRD immunophenotype on day 0 and day 8 of the treatment of patients 8 and 19 reported by report_all_clusters_hca function based on the resulting iteration of clustering by choose_cluster_hca function (range of clusters set to 5-100, threshold of the size of the clusters to 2 %) and the MRD percentage in the population. The green colour highlights the marker expressions corresponding on D0 and D8 of the treatment, red the expression differing on D0 and D8 and yellow the expression differing sligthly on D0 and D8 (weak to negative or positive or vice versa).

PATIENT 8								
Immonophenotype	**CD58**	**CD66c**	**CD10**	**CD45**	**CD19**	**CD20**	**CD34**	**CD38**
D0 - 81.7 %	+	+	+	W	+	+	+	+
D8 - 13.36 %	-	+	+	-	+	+	W	-
PATIENT 19								
Immonophenotype	**CD58**	**CD66c**	**CD10**	**CD45**	**CD19**	**CD20**	**CD34**	**CD38**
D0 - 79.45 %	+	-	+	W	+	-	-	+
D8 - 2.51 %	W	-	+	-	+	-	-	+

6 Discussion

6.1 Processing of information about the FCS files

More than 350 thousand of FCS files present in the CLIP database were analysed based on the acquired metainformation about the files. They served as source of important information for scripting to avoid possible obstacles in analysis.

Furthermore, based on these information 19 diagnostic files of patients with ALL with the panel CD44 CD27 CD34 CD19 CD10 CD38 CD20 CD45 were selected out of the FCS files database in CLIP. When searching for the reported immunophenotypes in the clinical database in chapter 5.5, it was confirmed that these files are indeed the diagnostic files, as the immunophenotype of other samples is not usually reported.

6.2 Automatic detection of populations in files of a selected FC panel

The automatically preprocessed files were clustered by both, HCA and FlowSOM clustering algorithms. The clustering was done several times, each time producing an iteration of clustering with a different number of clusters. From these clustering sets, the iteration of clustering containing the most negative population was chosen. This negative population was later used as a reference for immunophenotypisation.

The resulting clusters of HCA for range of clusters from 5 to 100 are presented in the form of a heatmaps with joined dendrograms in Fig. 11 and in matrices of scatterplots in Fig. 12 for file 1 and in Fig. 14 in Supplement 2 for files 2-19. During the analysis, it was determined, that a bigger range of clusters produces better clustering results. However, sometimes the negative cluster was not chosen ideally and was positive in a few markers. This is caused by the completely negative population not being present at all, or being so small it cannot be detectable as a separate population. E.g. in file 3 the truly negative population is present as a subpopulation of the current negative reference cluster. In this case, the correct determination could be achieved by setting a bigger range of clusters for clustering in `choose_cluster_hca` function with a maximum value higher than 100. In files 2, 5, 6, 7, 13, 14 and 16 there is at least one marker positive in the negative reference cluster. However, a completely negative cluster is missing in these files (as can be seen in Fig. 11 and Fig. 14) and therefore the proposed approach will probably not lead to better results even if adjusted. In future, this could be solved by creating an artificial negative population with parameter values of a defined quantile of the

values of the analysed dataset and using this population as negative reference for the `makeReport` function of the `flowPheno` library. This approach was already tested during the work on this thesis and did not produce plausible results yet. However, upon further fine-turning this method could still solve the problem of negative cluster choice in some cases.

The resulting clusters of FlowSOM for range of dimensions of the grid from 6 to 16 are shown in matrices of scatterplots in Fig. 13 for file 1 and in Fig. 15 in Supplement 2 for files 2-19. Similarly to HCA clustering, the negative reference cluster was also not adequate in a few cases. The clusters were positive for file 2 in marker CD27, for file 16 in CD34 and for file 7 in markers CD27 and CD34. The FlowSOM clustering was more reliable in negative reference cluster determination that the HCA clustering.

Notably, the file 18 did not produce representative clustering results for both HCA and FlowSOM algorithms. However, there was no reported immunophenotype in the clinical database. Only a note was present, stating that the cells of the patient express an atypical phenotype possibly indicating a solid tumour. This sample did not contain ALL blastic populations and therefore the determination of negative reference cluster was not discussed.

6.3 Comparison of automatic and clinical immunophenotype report

The automatic immunophenotype determination was done for HCA and FlowSOM clustering of all 19 diagnostic files. The reports in FDE format were compared to an FDE report present in the clinical database. It is important to state, that the FDE report in the clinical database was not complete and some markers were missing from the report, most often CD44, CD45 or CD27. The automatically reported immunophenotype could not be compared with the clinical immunophenotype in these cases. The comparison of FDE string report for both automatic reporting and the clinical reporting is shown in Table 10. The graphic comparison of these reports is presented in Table 11.

There were some variations present between the automatic HCA and FlowSOM FDE reports and the clinical report, especially when immunophenotyping markers as weak instead of positive or negative, or vice versa. This sort of misidentification occurred for 20.9 % of all markers with phenotype known from the clinical database for both, HCA and FlowSOM clustering. However, the agreement with reported immunophenotype was 75.7 % for HCA and 76.5 % for FlowSOM. Determination of the immunophenotype as negative instead of positive was scarce. This happened only for 3.5 % of all markers that were reported in the clinical

database for HCA and in 2.6 % for FlowSOM (CD44 for file 4 and 5, CD38 for file 7 for both HCA and FlowSOM clustering and CD38 for HCA clustering of file 13). This occurred mostly for the files where the reference negative cluster for immunophenotypisation was chosen poorly, being negative in most markers, but also being significantly positive in one or two markers. The reasons for this was already discussed in 6.2. No negative marker was determined as positive. However, this is probably due to the methodology of immunophenotype determination by `flowPheno`, when the negative population is set and based on this, all other values are mathematically classified. A negative population that was not chosen ideally and had a higher value in one parameter would not affect the immunophenotypisation of true negative markers.

Often the automatically reported immunophenotype for the markers CD44 and CD45 did not correspond with the clinical report and was weak or negative instead of positive. However, in some cases these markers were missing from clinical FDE report. Also, sometimes the clinical report consisted of more markers than could fit single tube. This could hint either the use of CD44 and CD45 values from other tubes or some correction based on values from multiple tubes. Therefore there can be no certainty, that the CD44 and CD45 immunophenotype was not indeed different in the diagnostic files analysed in this thesis.

Overall the level of agreement between automatic approach and expert reporting of B-ALL blast immunophenotyping was surprising. While the more important aim of this work was to prepare the framework and show if it is possible to automate the work traditionally purely in the expert domain, the results in respect to the actual malignant immunophenotypes are encouraging.

There are three manual inputs needed for the whole analytical protocol from the point of processing of metainformation about the FC database:

1) setting the requirements of the file selection
2) setting the columns to be excluded from the analysis prior to preprocessing and
3) setting the columns of scatter parameters prior to the clustering step.

Apart from this, the protocol is fully automated to the point of immunophenotype report. However, both manual inputs of column selection could be automated in the future.

6.4 Development of immunophenotype during the course of ALL

The analysis of CD58, CD66c, CD10, CD45, CD19, CD20, CD34 and CD38 immunophenotype of MRD on day 0 (D0), day 8 (D8) and day 15 (D15) was performed for 5 patients and presented in Table 12 and

Table 13. For three patients, the MRD level was under the detection threshold set in the analytical protocol on D8 and D15 and therefore these files are not shown. For two patients, the blastic populations are identified on D0 and D8 and the development of their immunophenotype can be seen.

As expected, the MRD population percentage decreased drastically in the course of the treatment of ALL, which was reflected in a much smaller, but clearly blastic population clusters on D8. For the files presented in Table 12 and

Table 13, the percentage dropped from approximately 90 % on D0 of treatment to 14 % on D8 and from 80 % do 3 %. On D15, the blastic population was not detectable at all. These percentages of blastic population on D0 and D15 correspond with the clinically reported percentage – 90 % on D0 and 0.022 % on D15 for patient 8 and 75 % on D0 and 0.016 % on D15. The percentage of blast population was not reported for D8 for both patients. For patient 8, markers CD34, CD38, CD45 and CD58 were downmodulated on D8 in comparison to D0. For patient 19, the expression of markers CD45 and CD58 decreased on D8 in comparison to D0. According to literature, markers CD58 (Ramalingam *et al.*, 2020), CD45 (Gaipa *et al.*, 2005; Dworzak *et al.*, 2010; Ramalingam *et al.*, 2020), CD38 (Ramalingam *et al.*, 2020) and CD34 (Gaipa *et al.*, 2005; Dworzak *et al.*, 2010; Burnusuzov *et al.*, 2016) can indeed be downmodulated during and after treatment. The expression of marker CD20 (Ramalingam *et al.*, 2020) usually remains stable.

The results demonstrate that it is possible to detect the imunophenotype development of MRD during the course of the disease and treatment by the proposed analytical protocol. Furthermore, the analytical protocol was successfully tested for another panel of markers.

However, what is most important is, that the analytical protocol that was proposed in this thesis clearly analyses the immunophenotypes of MRD populations. In theory, these results could be presented visually to the clinician in the form of heatmaps with conjoined dendrograms or in matrices of scatterplots (similarly to Fig. 11 and Fig. 12) to provide the overview of the population characteristics.

7 Conclusion

To conclude, a novel automated analytical protocol for reporting the immunophenotype of immunologically relevant populations for multiple FC files at the same time was constructed. It includes preprocessing of the files (compensation, logicle transformation and z-score normalisation), clustering the datasets into populations by two clustering algorithms, HCA and FlowSOM, choice of the best clustering set and an automated immunophenotype report for all major clusters.

The protocol is not fully automated yet – it requires to set the columns to be excluded from the preprocessing step and further analysis and set the column names in the clustering step. Another unautomated step is the choice of candidate blast population from the final immunopheotype report. Needless to say, the requirements for the selection of files to be analysed (e.g. choice of panel) also needs a human input.

The CLIP cytometric database containing approximately 350 thousand of FCS files was analysed in terms of metainformation about the files. Based on these information a set of 19 diagnostic files of ALL patients with the same panel of markers was chosen. The correct choice of the diagnostic files was later confirmed when searching in the CLIP clinical database.

These 19 files were processed by the proposed analytical protocol and immunophenotypes of all relevant populations were reported for both HCA and FlowSOM clustering methods. The report was compared to the officially reported immunophenotype saved in the CLIP clinical database. Both reports were markedly similar. The blastic populations were identified correctly for each file and the population percentage and immunophenotype were similar to the report in the clinical database. However, this approach is still somewhat problematic, especially in the cases where no population negative for all parameters is present in the dataset.

Moreover, files of five ALL patients from day 0, day 8 and day 15 of the treatment were immunophenotyped by the analytical protocol and the resulting immunophenotypes were compared between the files from different stages of the disease. The blast populations were detectable on day 8 for two patients only and on the day 15 for none of them. While no statistically relevant conclusions can be drawn for this number of files, it demonstrates the ability to determine malignant immunophenotypes during disease progression.

While still preliminary this or similar analytical protocol can, in the future, enhance and ease flow cytometry analysis in clinical settings.

8 References

Adam, Z., Krejčí, M. and Vorlíček, J. (2008) *Hematologie: přehled maligních hematologických nemocí*. 2nd edn. Prague: Grada Publishing.

Alomari, M. *et al.* (1982) 'Role of Lipid Rafts in Hematopoietic Stem Cells'.

Bagwell, B. and Barden, L. (1990) 'Data file standard for flow cytometry', *Cytometry*, 11(3), pp. 323–332. doi: 10.1002/cyto.990110303.

Bagwell, C. B. (2005) 'HyperLog - A flexible log-like transform for negative, zero, and positive valued data', *Cytometry Part A*, 64(1), pp. 34–42. doi: 10.1002/cyto.a.20114.

BAGWELL, C. B. and ADAMS, E. G. (1993) 'Fluorescence Spectral Overlap Compensation for Any Number of Flow Cytometry Parameters', *Annals of the New York Academy of Sciences*, 677(1 Clinical Flow), pp. 167–184. doi: 10.1111/j.1749-6632.1993.tb38775.x.

Bandura, D. R. *et al.* (2009) 'Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry', *Analytical Chemistry*, 81(16), pp. 6813–6822. doi: 10.1021/ac901049w.

Bendall, S. C. *et al.* (2014) 'Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development', *Cell*, 157(3), pp. 714–725. doi: 10.1016/j.cell.2014.04.005.

Burnusuzov, H. A. *et al.* (2016) 'Immunophenotypic Modulation of the Blast Cells in Childhood Acute Lymphoblastic Leukemia Minimal Residual Disease Detection', *Folia medica*, 58(1), pp. 28–35. doi: 10.1515/folmed-2016-0004.

Castillo, L. and MacCallum, D. M. (2012) 'Cytokine Measurement Using Cytometric Bead Arrays', in, pp. 425–434. doi: 10.1007/978-1-61779-539-8_29.

Chester, C. and Maecker, H. T. (2015) 'Algorithmic Tools for Mining High-Dimensional Cytometry Data', *The Journal of Immunology*, 195(3), pp. 773–779. doi: 10.4049/jimmunol.1500633.

Chevrier, S. *et al.* (2018) 'Compensation of Signal Spillover in Suspension and Imaging Mass Cytometry', *Cell Systems*, 6(5), pp. 612-620.e5. doi: 10.1016/j.cels.2018.02.010.

Corradi, B. *et al.* (2008) 'Efficient detection of leukemia-related fusion transcripts by multiplex PCR applied on a microelectronic platform', *Leukemia*, 22(2), pp. 294–302. doi: 10.1038/sj.leu.2404987.

Darzynkiewicz, Z., Juan, G. and Bedner, E. (1999) 'Determining Cell Cycle Stages by Flow Cytometry', *Current Protocols in Cell Biology*, 1(1). doi: 10.1002/0471143030.cb0804s01.

Delude, R. L. (2005) 'Flow cytometry', pp. 426–428. doi: 10.1097/01.CCM.0000186781.07221.F8.

Dixit, A. *et al.* (2007) 'Disseminated Intravascular Coagulation in Acute Leukemia at Presentation and During Induction Therapy', *Clinical and Applied Thrombosis/Hemostasis*, 13(3), pp. 292–298. doi: 10.1177/1076029607302435.

Doležel, J. *et al.* (2011) 'Chromosome Analysis and Sorting Using Flow Cytometry', in, pp. 221–238. doi: 10.1007/978-1-61737-957-4_12.

Dworzak, M. N. *et al.* (2010) 'Modulation of antigen expression in B-cell precursor acute lymphoblastic leukemia during induction therapy is partly transient: Evidence for a drug-induced regulatory phenomenon. Results of the AIEOP-BFM-ALL-FLOW-MRD-study group', *Cytometry Part B - Clinical Cytometry*, 78(3), pp. 147–153. doi: 10.1002/cyto.b.20516.

Dworzak, M. N. *et al.* (2018) 'AIEOP-BFM consensus guidelines 2016 for flow cytometric immunophenotyping of Pediatric acute lymphoblastic leukemia', *Cytometry Part B - Clinical Cytometry*, 94(1), pp. 82–93. doi: 10.1002/cyto.b.21518.

Finak, G. *et al.* (2010) 'Optimizing transformations for automated, high throughput analysis of flow cytometry data', *BMC Bioinformatics*, 11(1), p. 546. doi: 10.1186/1471-2105-11-546.

Finak, G. *et al.* (2014) 'High-throughput flow cytometry data normalization for clinical trials', *Cytometry Part A*, 85(3), pp. 277–286. doi: 10.1002/cyto.a.22433.

Fišer, K. *et al.* (2012) 'Detection and monitoring of normal and leukemic cell populations with hierarchical clustering of flow cytometry data', *Cytometry Part A*, 81A(1), pp. 25–34. doi: 10.1002/cyto.a.21148.

Fulwyler, M. J. (1965) 'Electronic Separation of Biological Cells by Volume', *Science*, 150(3698), pp. 910–911. doi: 10.1126/science.150.3698.910.

Gaipa, G. *et al.* (2005) 'Drug-induced immunophenotypic modulation in childhood ALL: Implications for minimal residual disease detection', *Leukemia*, 19(1), pp. 49–56. doi: 10.1038/sj.leu.2403559.

Van Gassen, S. *et al.* (2015) 'FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data', *Cytometry Part A*, 87(7), pp. 636–645. doi: 10.1002/cyto.a.22625.

Van Gassen, S., Callebaut, B. and Saeys, Y. (2019) 'FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data'.

Givan, A. L. (2011) 'Flow Cytometry: An Introduction', in, pp. 1–29. doi: 10.1007/978-1-61737-950-5_1.

Gökbuget, N. and Hoelzer, D. (2009) 'Treatment of Adult Acute Lymphoblastic Leukemia', *Seminars in Hematology*, 46(1), pp. 64–75. doi: 10.1053/j.seminhematol.2008.09.003.

Gorczyca, W. *et al.* (2011) *Immunophenotypic Pattern of Myeloid Populations by Flow Cytometry Analysis*, *Methods in Cell Biology*. Elsevier Inc. doi: 10.1016/B978-0-12-385493-3.00010-3.

Guérin, F. *et al.* (2017) 'Flow cytometry sorting of nuclei enables the first global characterization of Paramecium germline DNA and transposable elements', *BMC Genomics*, 18(1), p. 327. doi: 10.1186/s12864-017-3713-7.

Hahne, F. *et al.* (2009) 'flowCore: a Bioconductor package for high throughput flow cytometry', *BMC Bioinformatics*, 10(1), p. 106. doi: 10.1186/1471-2105-10-106.

Herzenberg, Leonore A *et al.* (2006) *Interpreting flow cytometry data: a guide for the perplexed*. doi: 10.1038/ni0706-681.

Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, C. K. (eds). (2016) 'SEER Cancer Statistics Review 1975-2013 National Cancer Institute SEER Cancer Statistics Review 1975-2013 National Cancer Institute', *SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016.*, pp. 1992–2013. Available at: http://seer.cancer.gov/csr/1975%7B_%7D2013/results%7B_%7Dmerged/sect%7B_%7D24%7B_%7Dstomach.pdf.

Hunger, S. P. and Mullighan, C. G. (2015) 'Acute lymphoblastic leukemia in children', *New England Journal of Medicine*, 373(16), pp. 1541–1552. doi: 10.1056/NEJMra1400972.

Hussels, M., Engel, S. and Bock, N. (2019) 'Investigation of direct counting and sizing of DNA fragments in flow applying an improved data analysis and correction method', *Biomolecular Detection and Quantification*. Elsevier, 17(August 2018), p. 100083. doi: 10.1016/j.bdq.2019.100083.

Jahan-Tigh, R. R. *et al.* (2012) 'Flow cytometry', *Journal of Investigative Dermatology*. Elsevier Masson SAS, 132(10), pp. 1–6. doi: 10.1038/jid.2012.282.

Kalina, T. *et al.* (2012) 'EuroFlow standardization of flow cytometer instrument settings and immunophenotyping protocols', *Leukemia*, 26(9), pp. 1986–2010. doi: 10.1038/leu.2012.122.

Kalina, T. and Zuna, J. (eds) (no date) 'Current investigations in blood and cancer cells', in.

Kamentsky, L. A. and Melamed, M. R. (1967) 'Spectrophotometric Cell Sorter', *Science*, 156(3780), pp. 1364–1365. doi: 10.1126/science.156.3780.1364.

Kansal, R. *et al.* (2004) 'Precursor B Lymphoblastic Leukemia With Surface Light Chain Immunoglobulin Restriction', *American Journal of Clinical Pathology*, 121(4), pp. 512–525. doi: 10.1309/WTXCQ5NRACVXTYBY.

Karanth, N. *et al.* (2008) 'Mediastinal lymphadenopathy in a patient with previously treated T-cell acute lymphoblastic leukaemia', *Medical Journal of Australia*, 188(2), pp. 117–118. doi: 10.5694/j.1326-5377.2008.tb01537.x.

Kimball, A. K. *et al.* (2018) 'A Beginner's Guide to Analyzing and Visualizing Mass Cytometry Data', *The Journal of Immunology*, 200(1), pp. 3–22. doi: 10.4049/jimmunol.1701494.

Krishnan, R. and Jagannathan, S. (2015) 'Hierarchical mahalanobis distance clustering based technique for prognostics in applications generating big data', *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, pp. 516–521. doi: 10.1109/SSCI.2015.82.

Leary, J. F. *et al.* (1979) 'Laser flow cytometric light scatter and fluorescence pulse width and pulse rise-time sizing of mammalian cells.', *Journal of Histochemistry & Cytochemistry*, 27(1), pp. 315–320. doi: 10.1177/27.1.374594.

Lippé, R. (2017) 'Flow Virometry: a Powerful Tool To Functionally Characterize Viruses', *Journal of Virology*. Edited by B. A. Glaunsinger, 92(3). doi: 10.1128/JVI.01765-17.

Liu, X. *et al.* (2019) 'A comparison framework and guideline of clustering methods for mass cytometry data', *Genome Biology*. Genome Biology, 20(1), pp. 1–18. doi: 10.1186/s13059-019-1917-7.

Maciorowski, Z., Chattopadhyay, P. K. and Jain, P. (2017) 'Basic multicolor flow cytometry', *Current Protocols in Immunology*, 2017(April), pp. 5.4.1-5.4.38. doi: 10.1002/cpim.26.

Malo, N. *et al.* (2006) 'Statistical practice in high-throughput screening data analysis', *Nature Biotechnology*, 24(2), pp. 167–175. doi: 10.1038/nbt1186.

Manoharan, A. *et al.* (1980) 'SIGNIFICANCE OF SPLENOMEGALY IN CHILDHOOD ACUTE LYMPHOBLASTIC LEUKÆMIA IN REMISSION', *The Lancet*, 315(8166), pp. 449–452. doi: 10.1016/S0140-6736(80)90998-8.

Meur, N. Le, Hahne, F. and Ellis, B. (2007) *FlowCore: data structures package for flow cytometry data*, *Bioconductor Project*. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:flowCore+:+data+structures+package+for+flow+cytometry+data#2.

Moore, W. A. and Parks, D. R. (2012) 'Update for the logicle data scale including operational code implementations', *Cytometry Part A*, 81 A(4), pp. 273–277. doi: 10.1002/cyto.a.22030.

O'Donnell, E. A., Ernst, D. N. and Hingorani, R. (2013) 'Multiparameter Flow Cytometry: Advances in High Resolution Analysis', *Immune Network*, 13(2), p. 43. doi: 10.4110/in.2013.13.2.43.

Onciu, M. (2009) 'Acute Lymphoblastic Leukemia', *Hematology/Oncology Clinics of North America*. Elsevier Ltd, 23(4), pp. 655–674. doi: 10.1016/j.hoc.2009.04.009.

Ozdemir, N. *et al.* (2016) 'Febrile neutropenia in children with acute lymphoblastic leukemia: single center experience', *Türk Pediatri Arşivi*, 51(2), pp. 79–86. doi: 10.5152/TurkPediatriArs.2016.2757.

Parks, D. R., Roederer, M. and Moore, W. A. (2006) 'A new "logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data', *Cytometry Part A*, 69(6), pp. 541–551. doi: 10.1002/cyto.a.20258.

Peters, J. M. and Ansari, M. Q. (2011) 'Multiparameter flow cytometry in the diagnosis and management of acute leukemia', *Archives of Pathology and Laboratory Medicine*, 135(1), pp. 44–54. doi: 10.1016/j.ypat.2011.07.004.

Ramalingam, T. R. *et al.* (2020) 'Immunophenotypic modulation in pediatric B lymphoblastic leukemia and its implications in MRD detection', *Leukemia and Lymphoma*. Taylor & Francis, 0(0), pp. 1–7. doi: 10.1080/10428194.2020.1742902.

Robinson, M. D. *et al.* (2017) 'CyTOF workflow: Differential discovery in high-throughput high-dimensional cytometry datasets', *F1000Research*, 6(May), pp. 1–53. doi: 10.12688/f1000research.11622.2.

Roederer, M. (2001) 'Spectral compensation for flow cytometry: Visualization artifacts, limitations, and caveats', *Cytometry*, 45(3), pp. 194–205. doi: 10.1002/1097-0320(20011101)45:3<194::AID-CYTO1163>3.0.CO;2-C.

Rose-Inman, H. and Kuehl, D. (2014) 'Acute leukemia', *Emergency Medicine Clinics of North America*. Elsevier Inc, 32(3), pp. 579–596. doi: 10.1016/j.emc.2014.04.004.

Sakata, H. *et al.* (2014) 'Acute leukemia presenting as bone pain with normal white blood cell count', *Acute Medicine & Surgery*, 1(4), pp. 249–249. doi: 10.1002/ams2.46.

Sayyed, A. H. *et al.* (2018) 'Acute Lymphoblastic Leukemia Presenting with Liver Infiltration and Severe Lactic Acidosis', *American Journal of Case Reports*, 19, pp. 453–457. doi: 10.12659/AJCR.907383.

Shapiro, H. M. (2008) 'Flow cytometry of bacterial membrane potential and permeability.', *Methods in molecular medicine*, 142, pp. 175–186. doi: 10.1007/978-1-59745-246-5_14.

Smith, M. A. *et al.* (2014) 'Declining childhood and adolescent cancer mortality', *Cancer*, 120(16), pp. 2497–2506. doi: 10.1002/cncr.28748.

Spidlen, J. *et al.* (2009) 'Data File Standard for Flow Cytometry, version FCS 3.1', *Cytometry Part A*, 9999A, p. NA-NA. doi: 10.1002/cyto.a.20825.

Spidlen, J. *et al.* (2010) 'Data File Standard for Flow Cytometry Normative Reference', pp. 1–34.

Starza, I. Della *et al.* (2019) 'Minimal residual disease in acute lymphoblastic leukemia: Technical and clinical advances', *Frontiers in Oncology*, 9(August), pp. 1–17. doi: 10.3389/fonc.2019.00726.

Swerdlow, S. H. *et al.* (2016) 'The 2016 revision of the World Health Organization classification of lymphoid neoplasms', *Blood*, 127(20), pp. 2375–2390. doi: 10.1182/blood-2016-01-643569.

Teuffel, O. *et al.* (2008) 'Anemia and survival in childhood acute lymphoblastic leukemia', *Haematologica*, 93(11), pp. 1652–1657. doi: 10.3324/haematol.13156.

Toghi Eshghi, S. *et al.* (2019) 'Quantitative Comparison of Conventional and t-SNE-guided Gating Analyses', *Frontiers in*

Immunology, 10. doi: 10.3389/fimmu.2019.01194.

Valli, M. *et al.* (2005) 'Intracellular pH Distribution in', *Microbiology*, 71(3), pp. 1515–1521. doi: 10.1128/AEM.71.3.1515.

Verschoor, C. P. *et al.* (2015) 'An introduction to automated flow cytometry gating tools and their implementation', *Frontiers in Immunology*, 6(JUL), pp. 1–9. doi: 10.3389/fimmu.2015.00380.

Vrooman, L. M. and Silverman, L. B. (2016) 'Treatment of Childhood Acute Lymphoblastic Leukemia: Prognostic Factors and Clinical Advances', *Current Hematologic Malignancy Reports*. Current Hematologic Malignancy Reports, 11(5), pp. 385–394. doi: 10.1007/s11899-016-0337-y.

Weber, L. M. and Robinson, M. D. (2016a) 'Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data', *Cytometry Part A*, 89(12), pp. 1084–1096. doi: 10.1002/cyto.a.23030.

Weber, L. M. and Robinson, M. D. (2016b) 'Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data', *Cytometry Part A*, 89(12), pp. 1084–1096. doi: 10.1002/cyto.a.23030.

Withycombe, J. S. *et al.* (2015) 'Weight change during childhood acute lymphoblastic leukemia induction therapy predicts obesity: A report from the Children's Oncology Group', *Pediatric Blood & Cancer*, 62(3), pp. 434–439. doi: 10.1002/pbc.25316.

Wlodkowic, D., Skommer, J. and Darzynkiewicz, Z. (2009) 'Flow Cytometry-Based Apoptosis Detection', in *Business*, pp. 19–32. doi: 10.1007/978-1-60327-017-5_2.

Citations of R and R libraries:

Ellis, B. *et al.* (2019) *FlowCore: FlowCore: Basic structures for flow cytometry data*. R package version 1.52.1.

Finak, G. & Jiang, M. (2020) *FlowWorkspace: Infrastructure for representing and interacting with gated and ungated cytometry data sets.* R package version 3.34.1.

Grolemund, G. & Wickham, H. Dates and times made easy with lubridate. (2011) *Journal of Statistical Software* **40,** 1–25. http://www.jstatsoft.org/v40/i03/.

Kolde, R. (2019) *Pheatmap: Pretty heatmaps*. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap.

Neuwirth, E. (2014) *RColorBrewer: ColorBrewer palettes*. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schloerke, B. *et al.* (2020) *GGally: Extension to 'ggplot2'*. R package version 1.5.0. https://CRAN.R-project.org/package=GGally

Vaidyanathan, R., Xie, Y., Allaire, J., Cheng, J. & Russell, K. (2019) *Htmlwidgets: HTML widgets for r*. R package version 1.5.1. https://CRAN.R-project.org/package=htmlwidgets.

Van Gassen, S., Callebaut, B. & Saeys, Y. (2019) *FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data*. http://www.r-project.org, http://dambi.ugent.be.

Wickham, H. (2016) *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H. *et al.* (2019) Welcome to the tidyverse. *Journal of Open Source Software* **4,** 1686, https://doi.org/10.21105/joss.01686.

Wickham, H., François, R., Henry, L. & Müller, K. (2020) *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr.

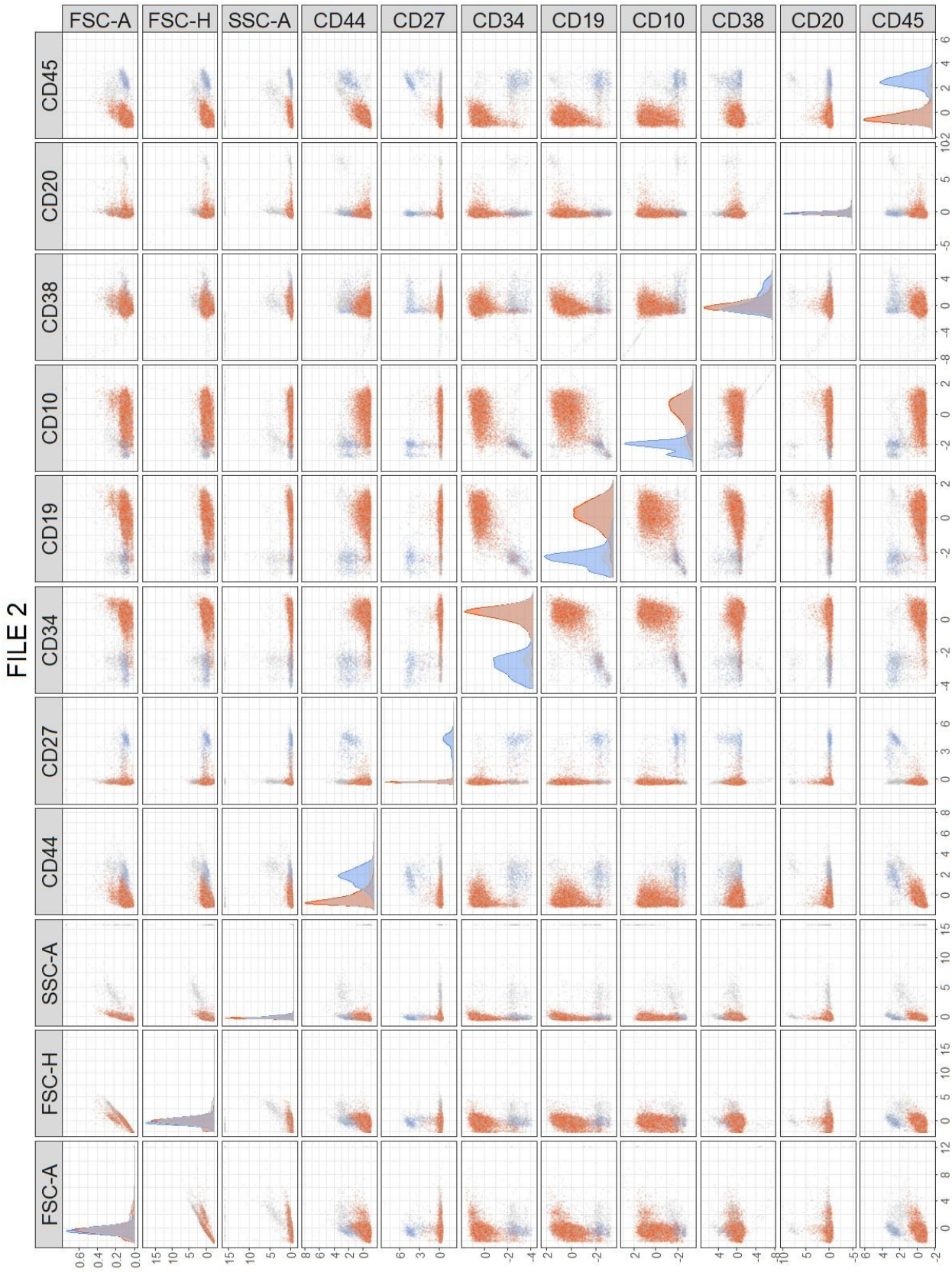
Wickham, H. & Henry, L. (2020) *Tidyr: Tidy messy data*. R package version 1.0.2. https://CRAN.R-project.org/package=tidyr.

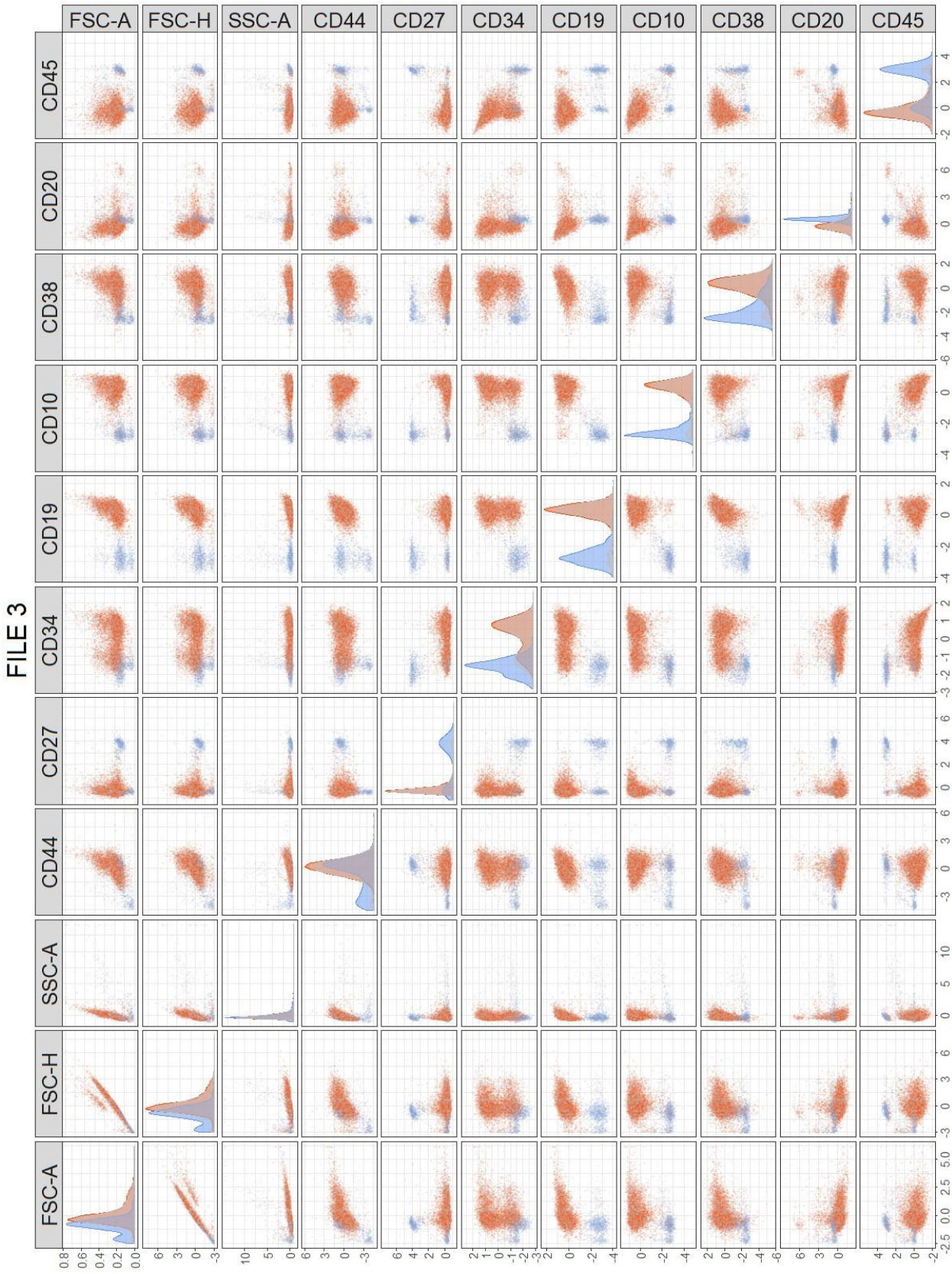
Xie, Y. (2020) *Knitr: A general-purpose package for dynamic report generation in r*. R package version 1.28.

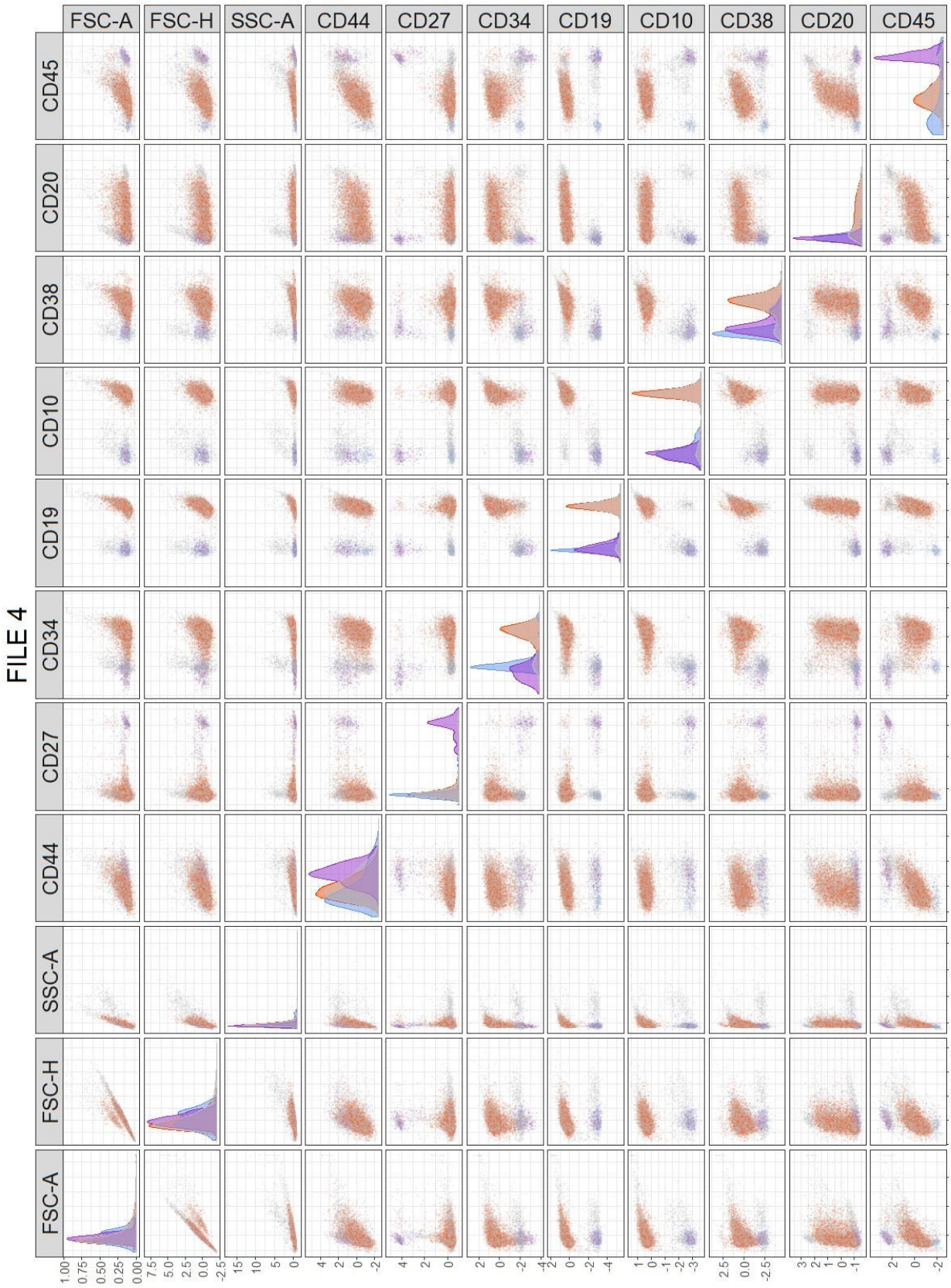
Zhu, H. (2019) *KableExtra: Construct complex table with 'kable' and pipe syntax*. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra.

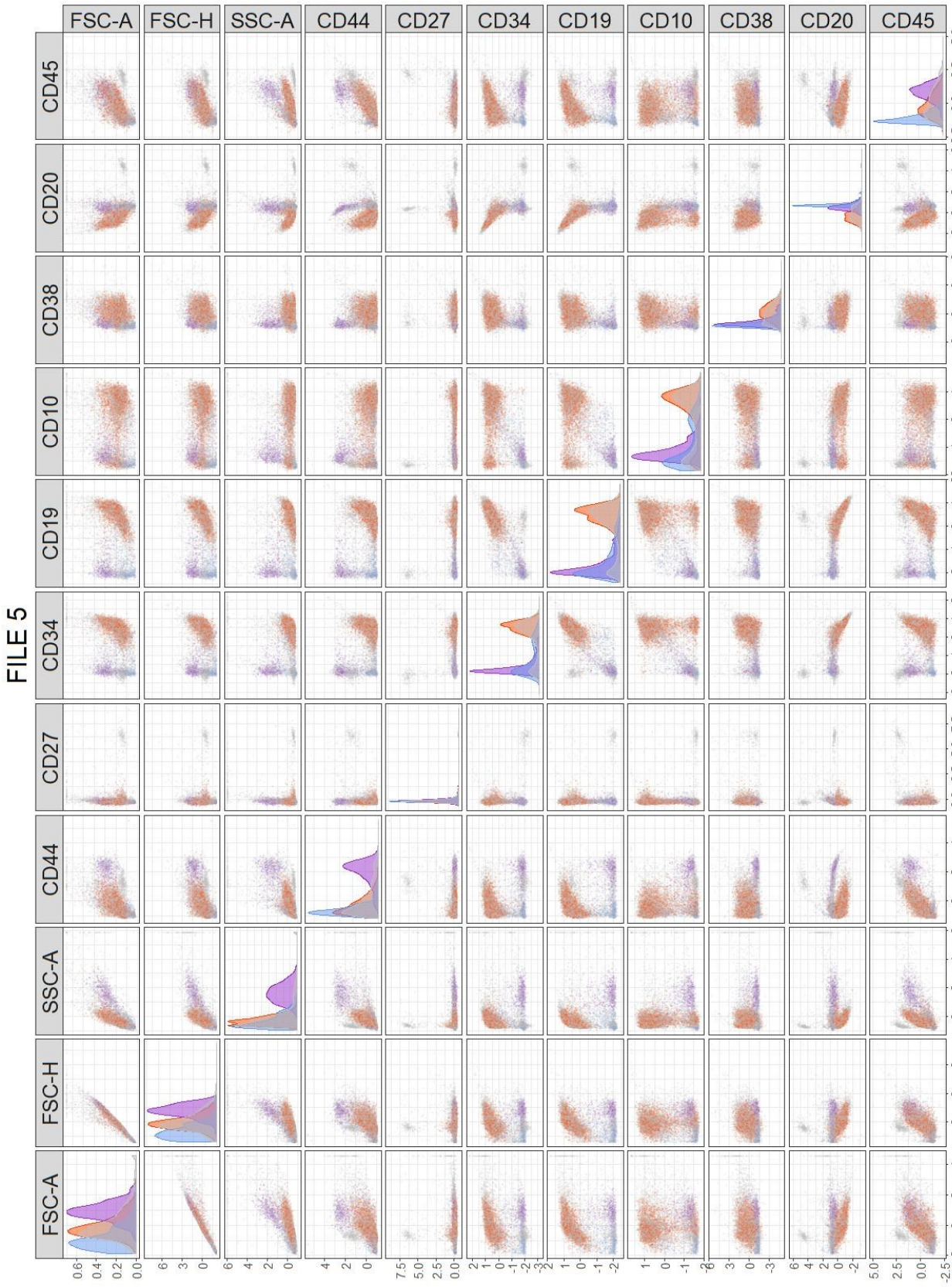
Supplement 1 – Matrices of scatterplots of the resulting HCA clustering of diagnostic files 2-19

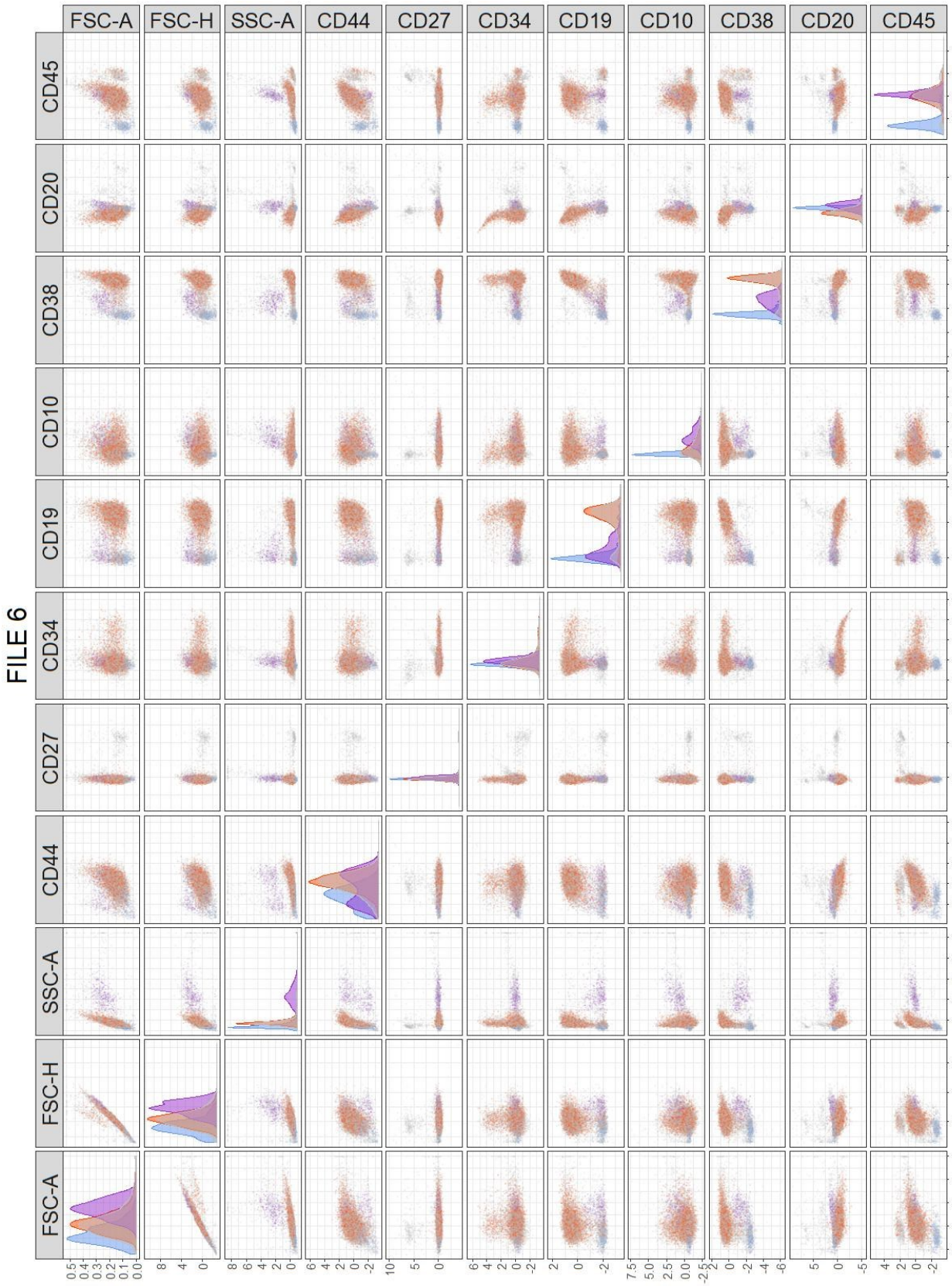
Fig. 14 - Matrices of scatterplots of the resulting HCA clustering of diagnostic files 2-19 in parameters FSC-A, FSC-H, SSC-A, CD44, CD27, CD34, CD19, CD10, CD38, CD20 and CD45. For the clustering, the choose_cluster_hca function was used with range of clusters set to 5-100 and minimal size of the populations set to 3 %. The histograms of each parameter are on the diagonal. The clusters of size over 3 % from the resulting HCA iteration of clustering are highlighted in colour – red for blastic population, blue for referential negative population and purple, green and yellow for other populations. The blastic B-ALL population was selected manually based on the results of automated immunophenotypisation in chapter 5.4. In file 18, no blastic population was identified. The colours of the populations correspond to the ones in heatmap with associated dendrogram (see Fig. 11).

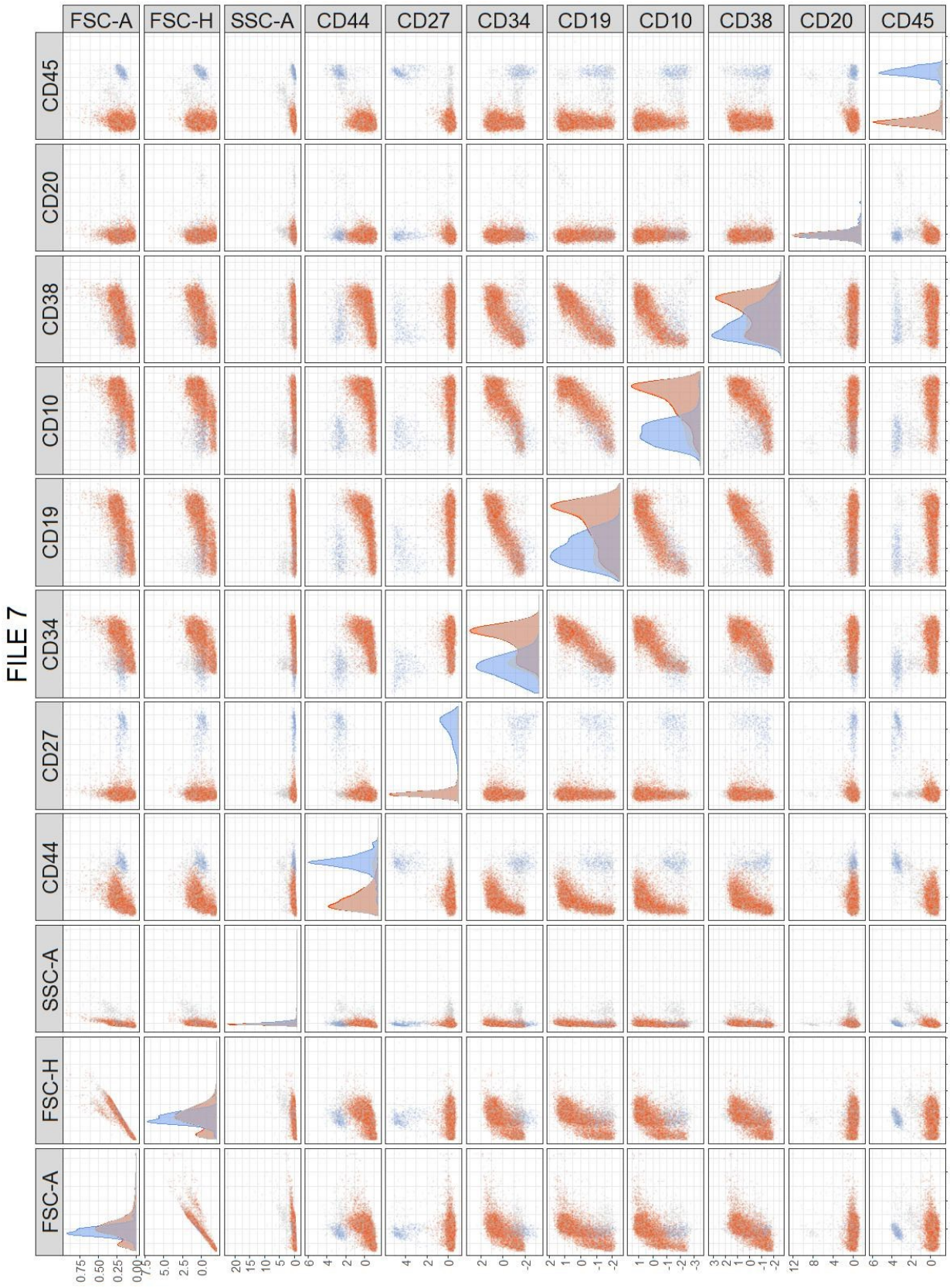

FILE 2
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

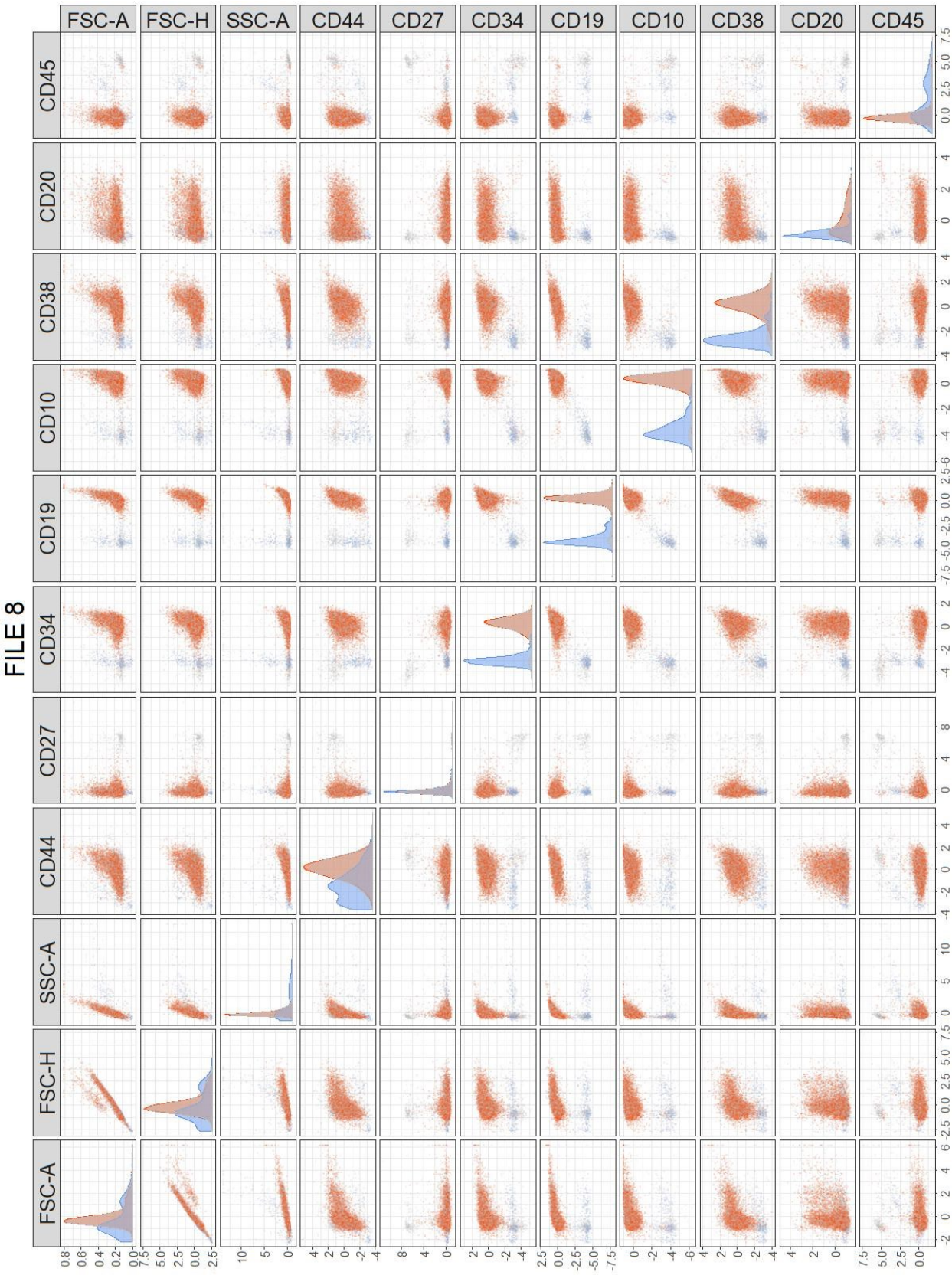

FILE 3
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

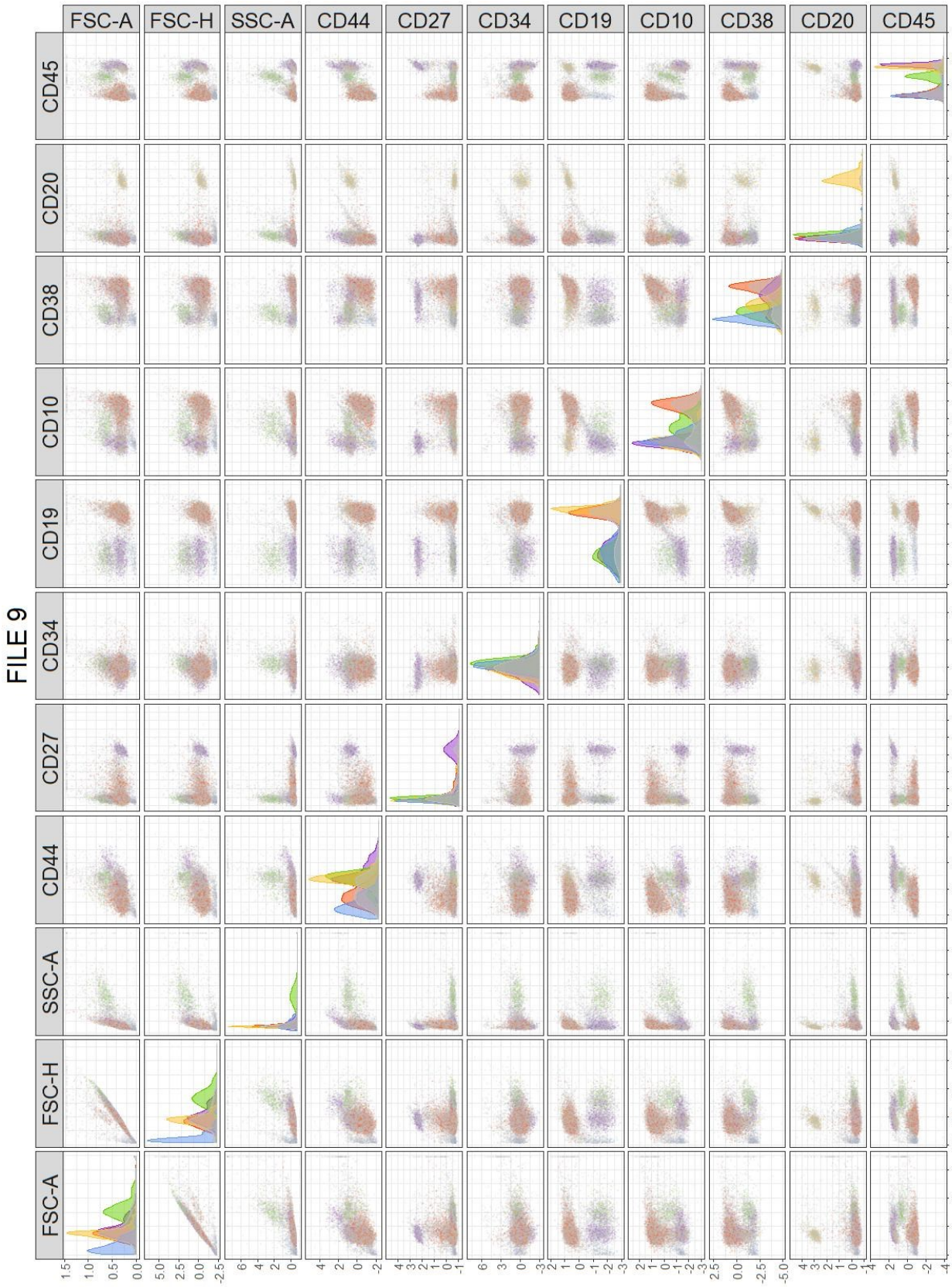

FILE 4
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

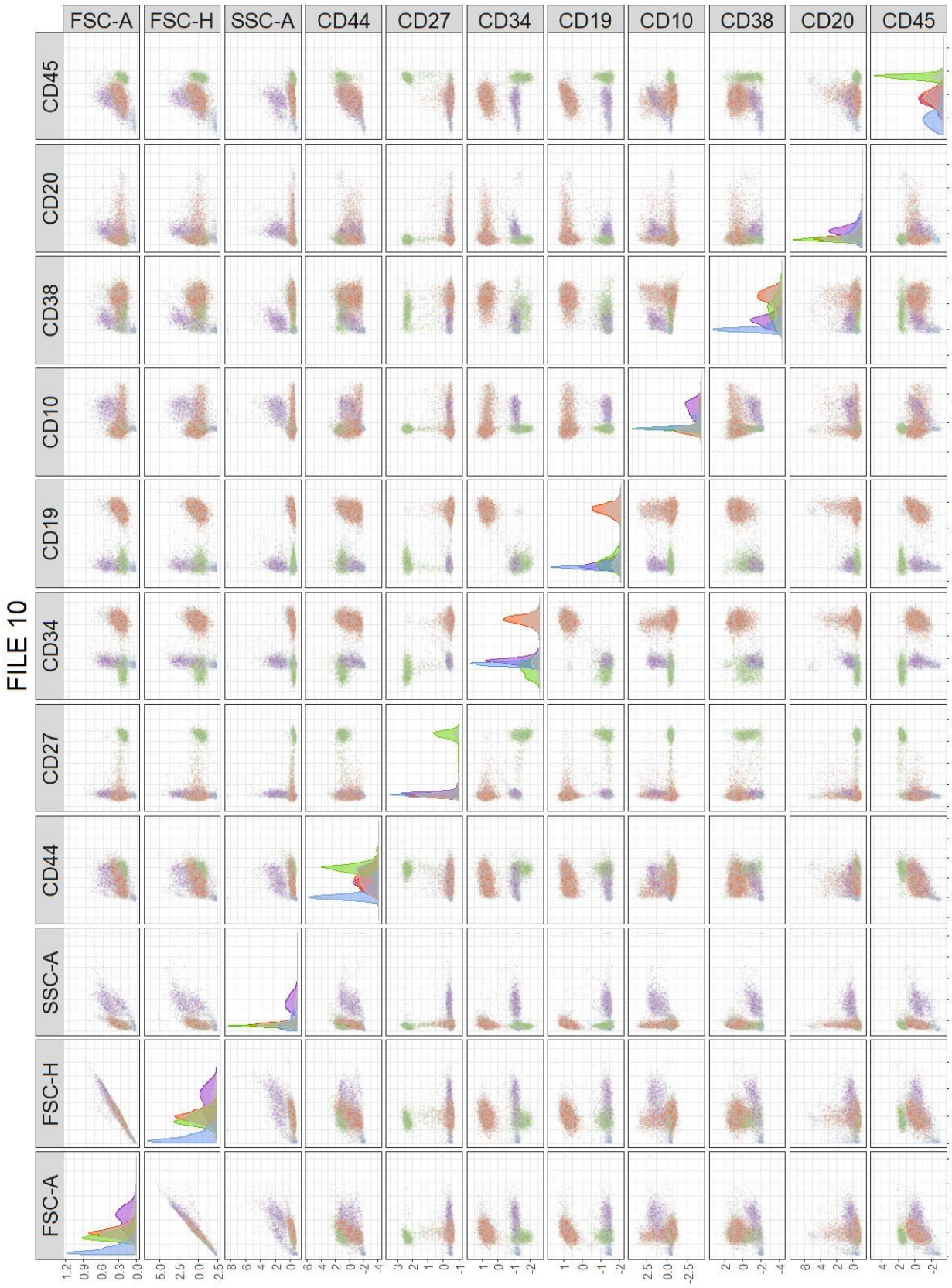

FILE 5
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

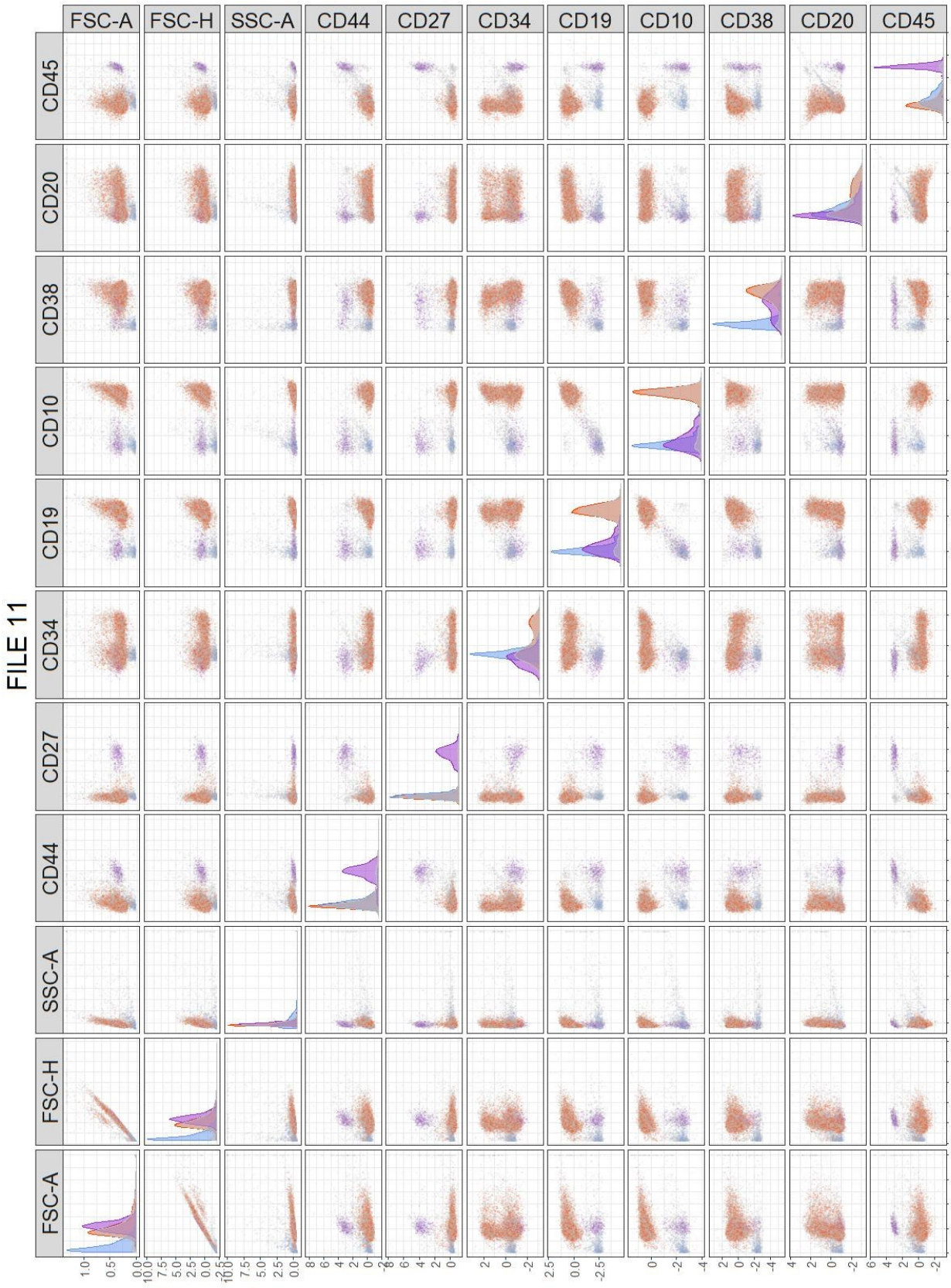

FILE 6
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

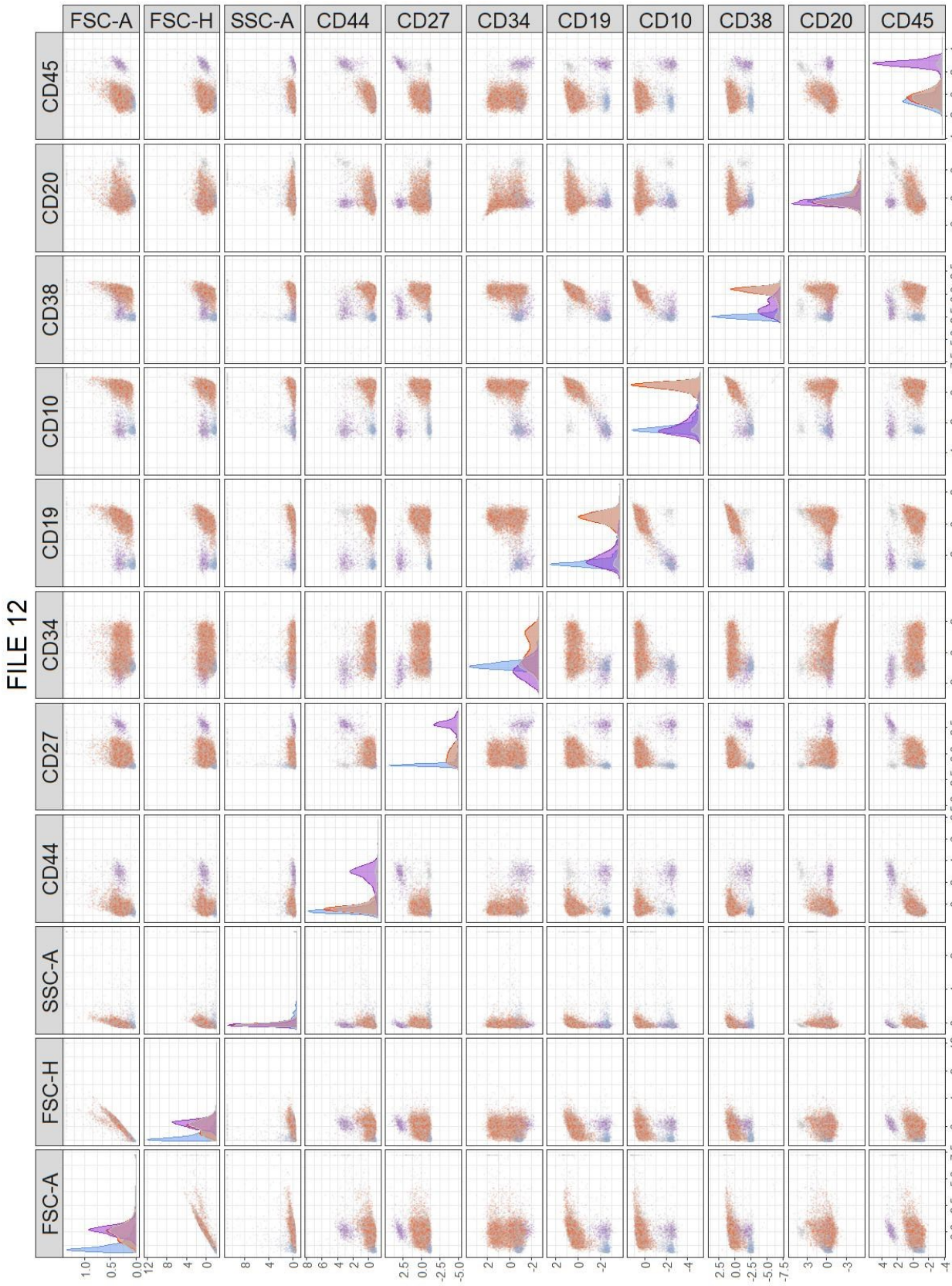

FILE 7
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

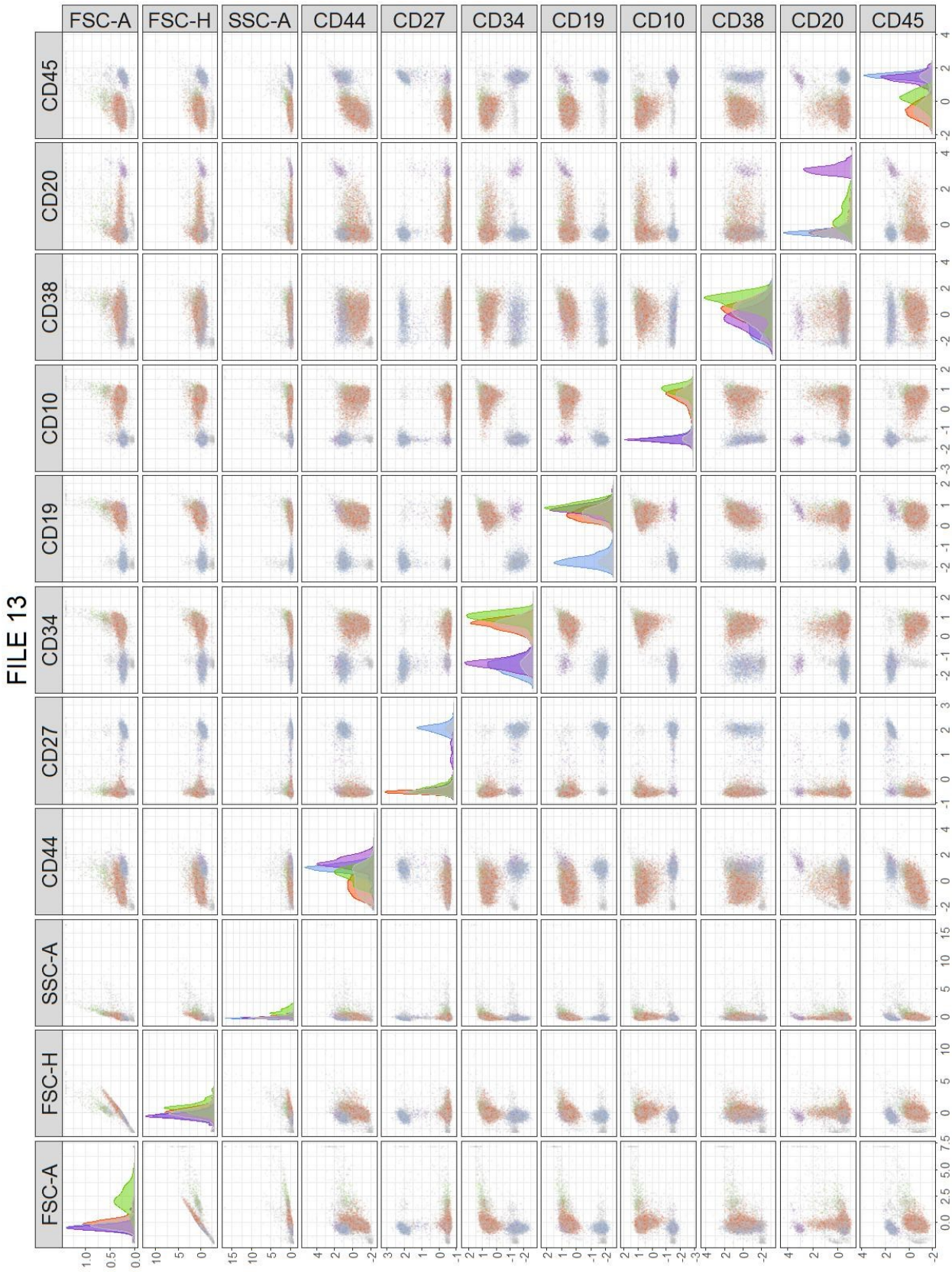

FILE 8
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

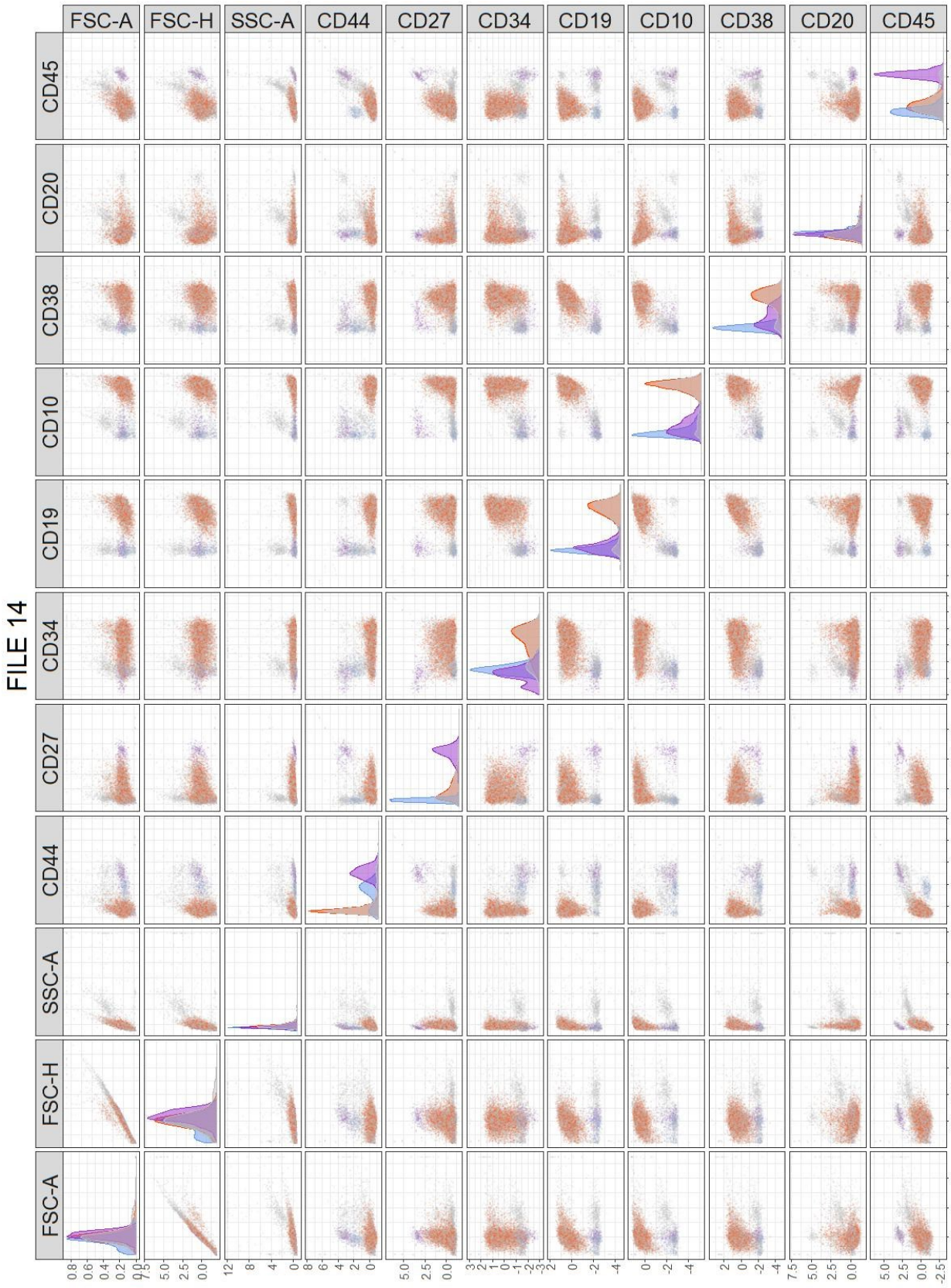

FILE 9
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

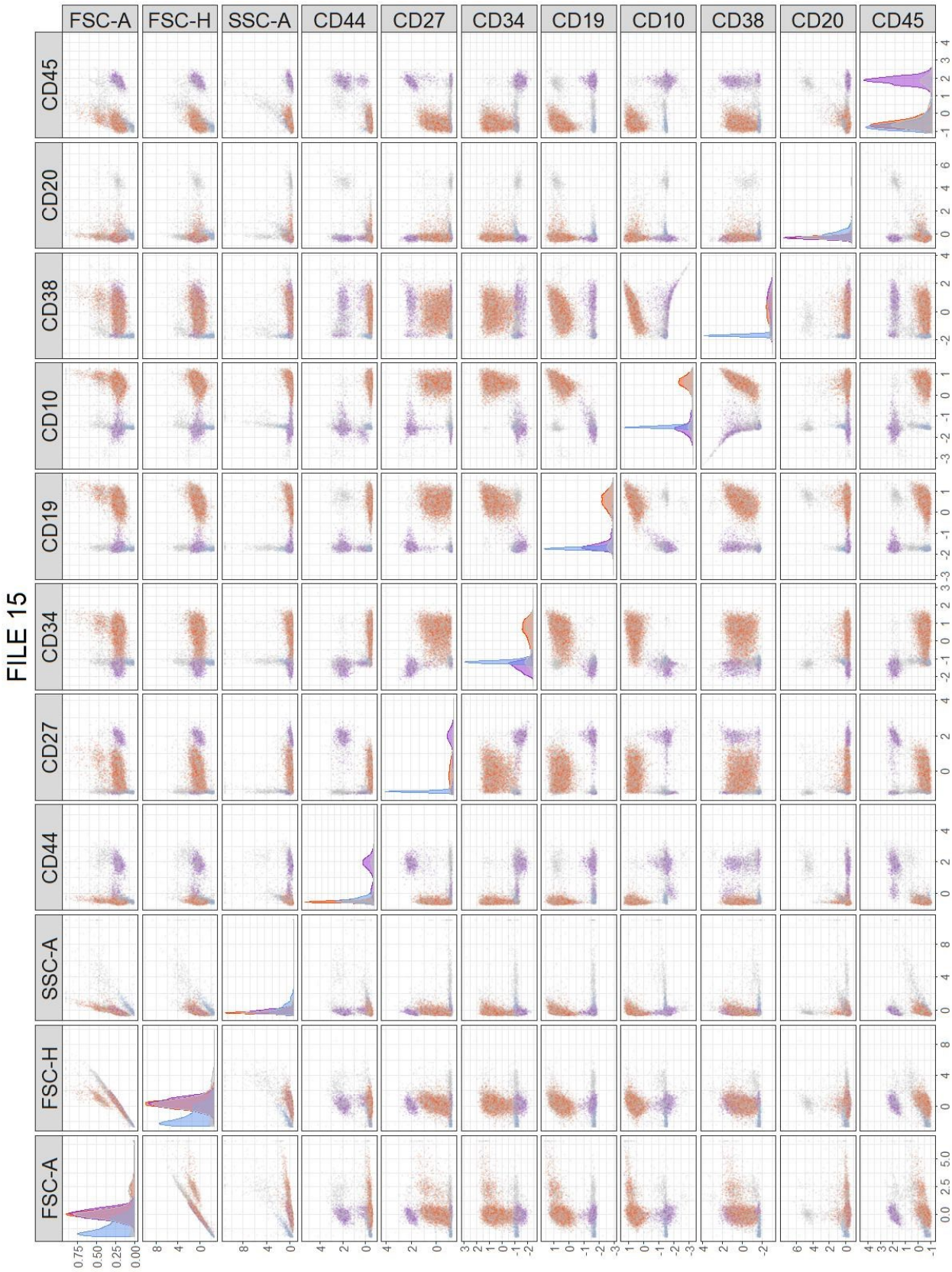

FILE 10
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

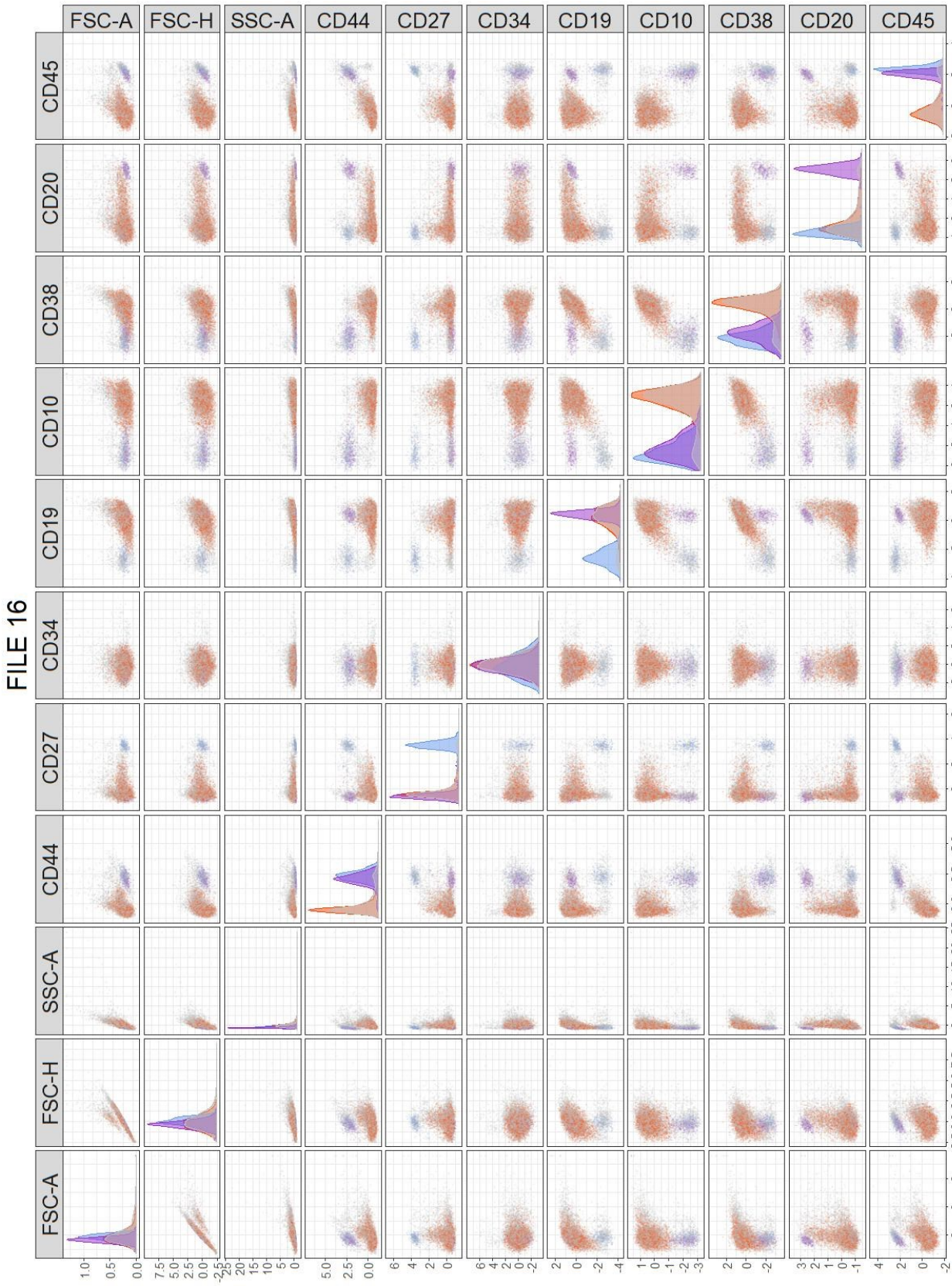

FILE 11
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

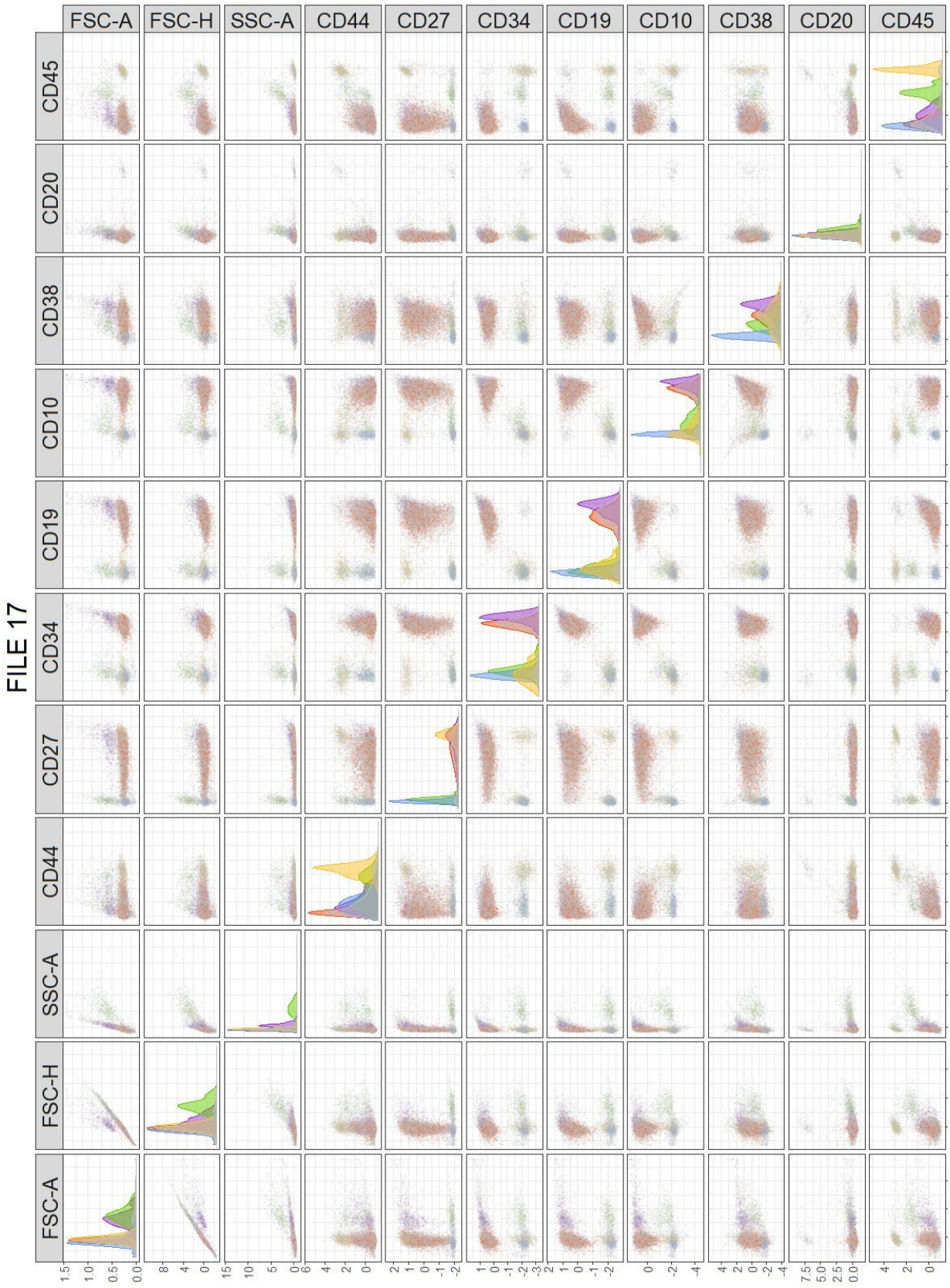

FILE 12
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

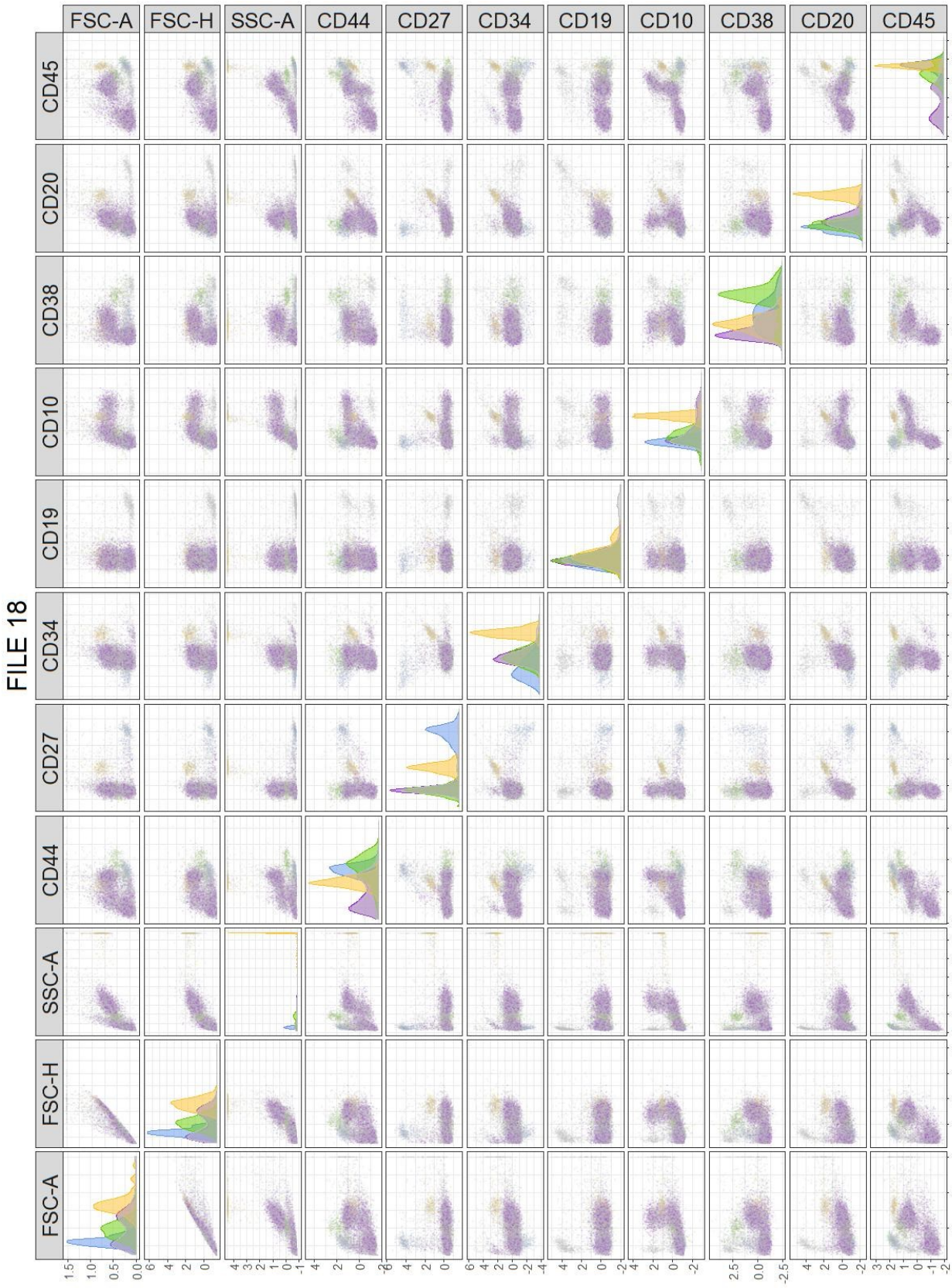

FILE 13
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

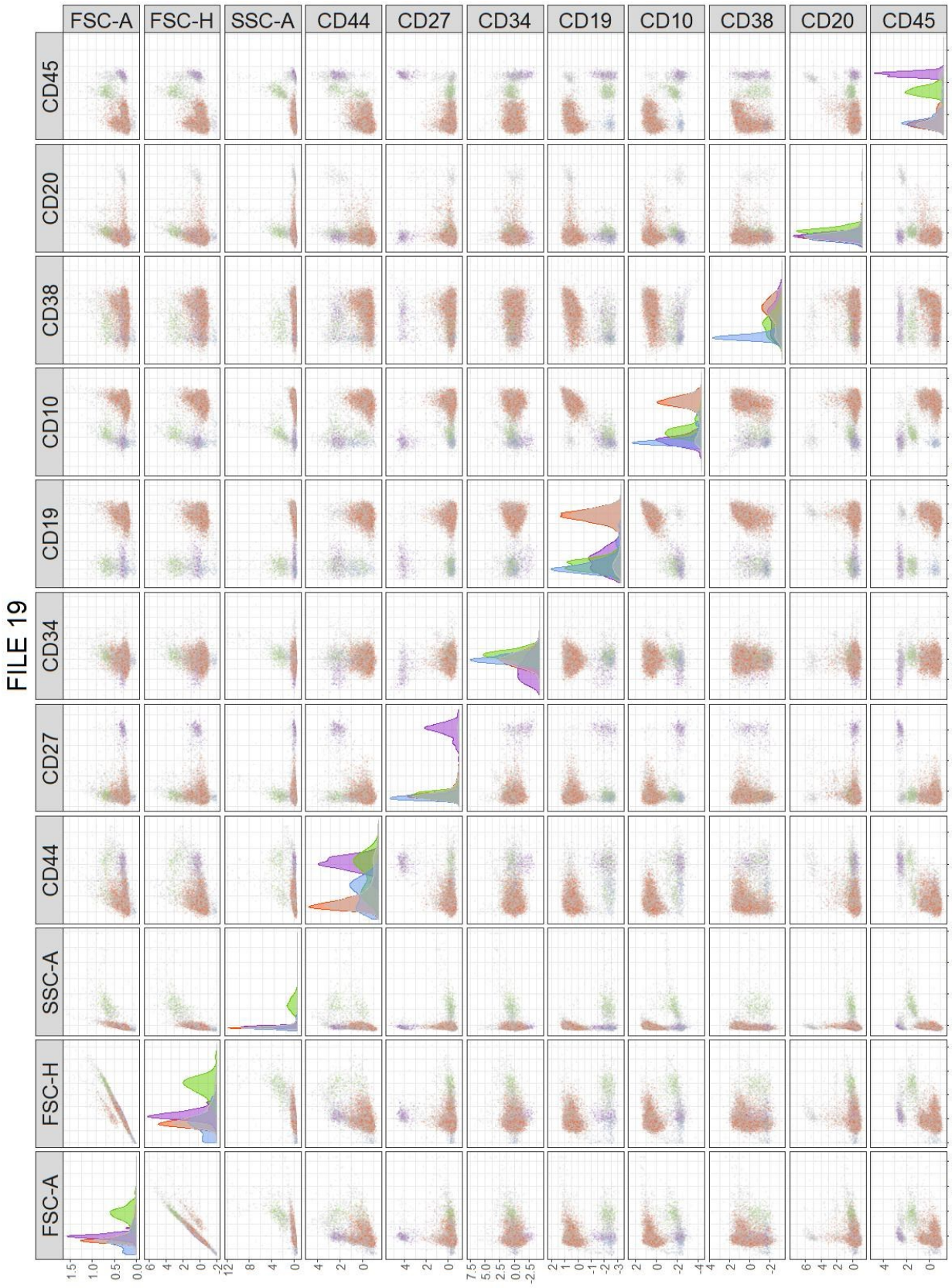

FILE 14
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 15
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 16
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

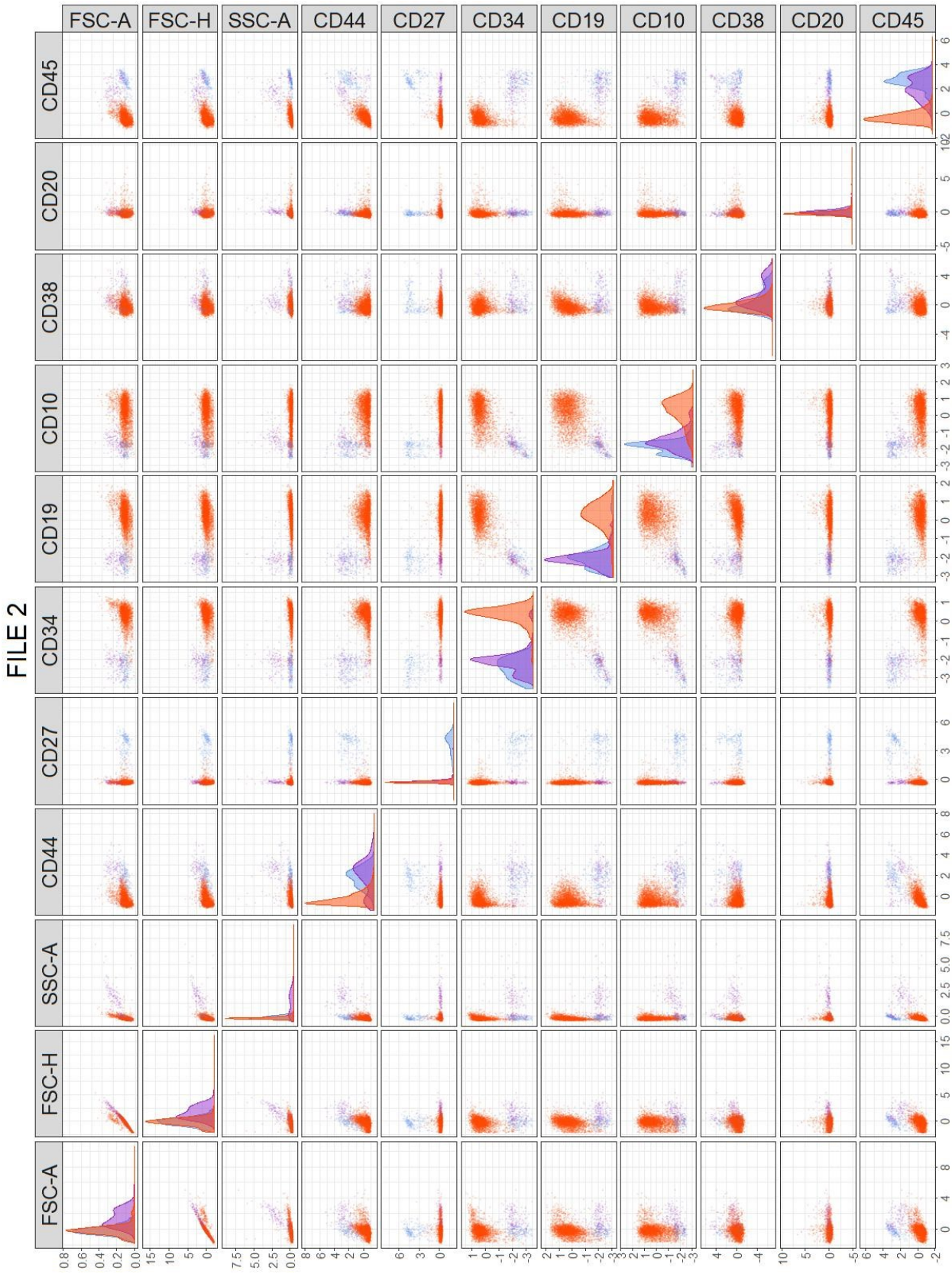

FILE 17
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

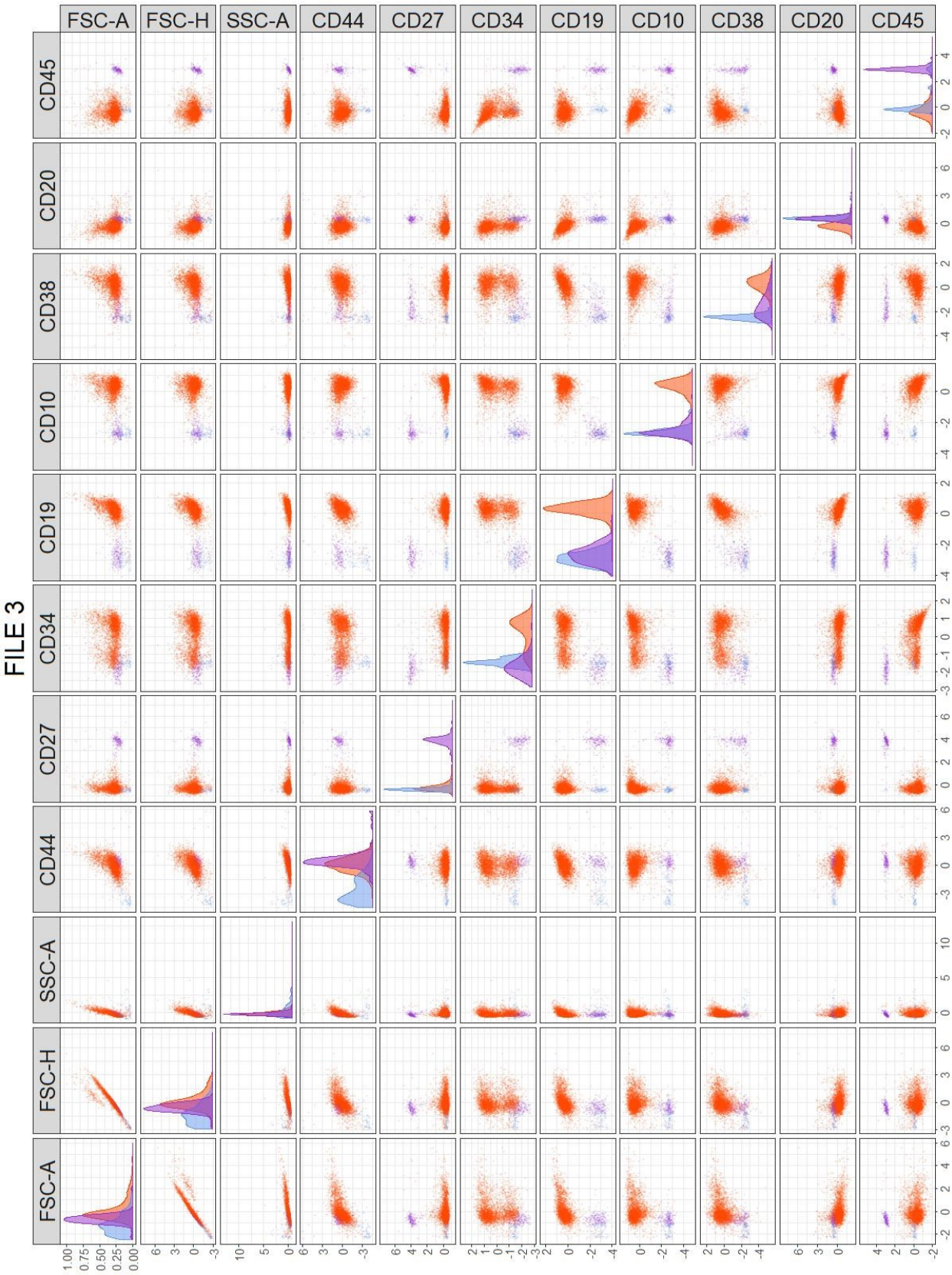

FILE 18
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

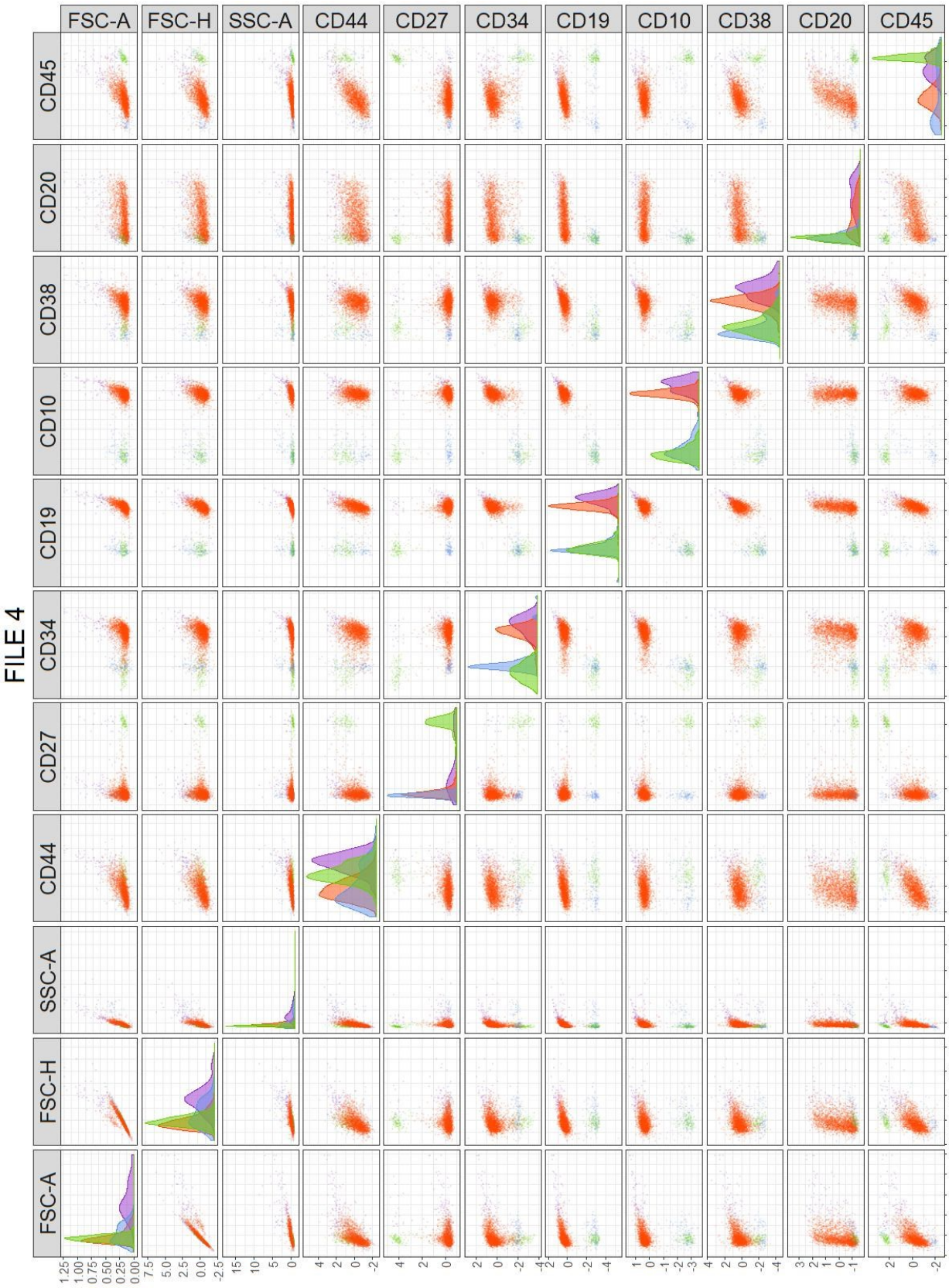

FILE 19
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

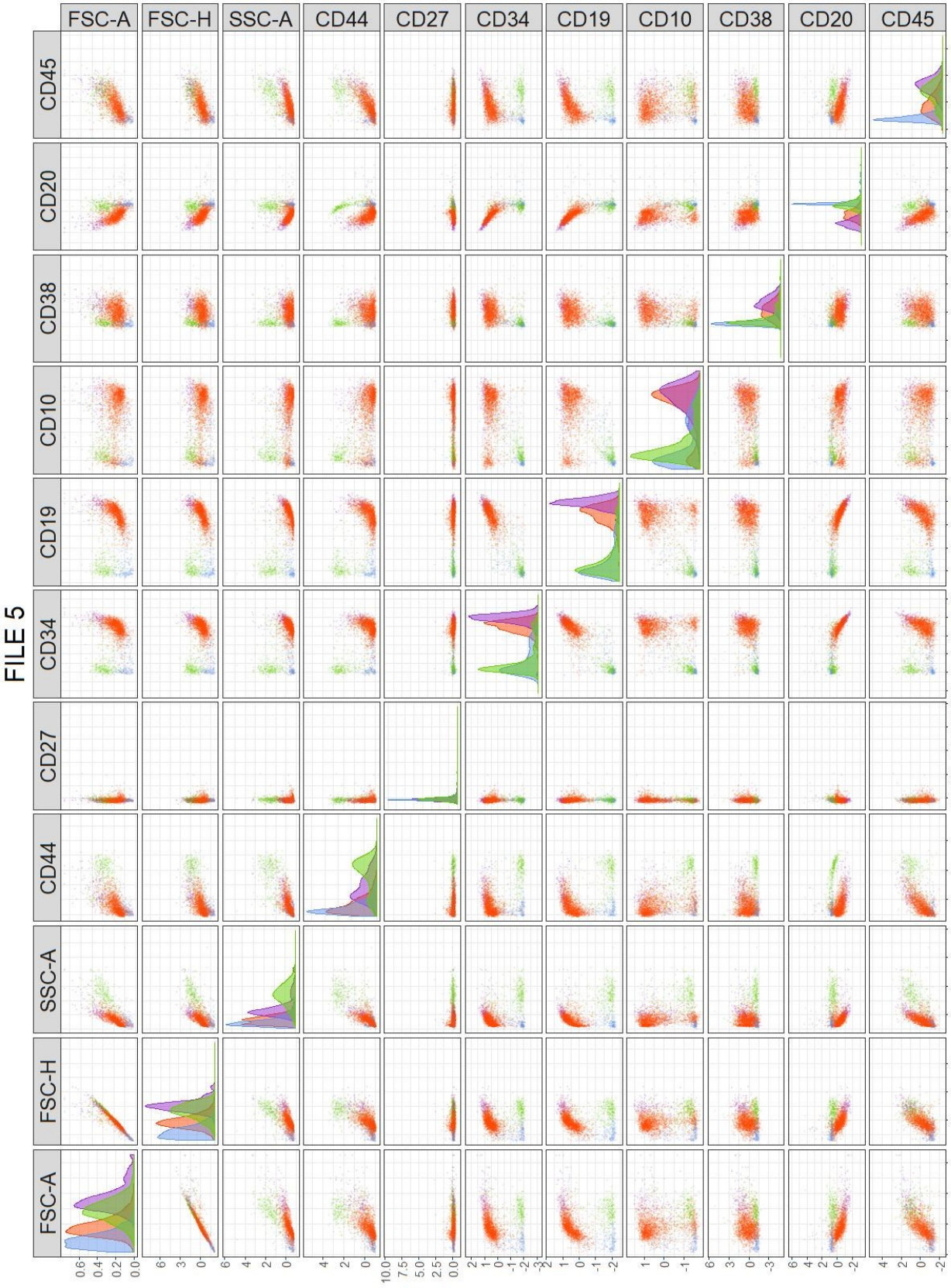
Supplement 2 - Matrices of scatterplots of the resulting FlowSOM clustering of diagnostic files 2-19

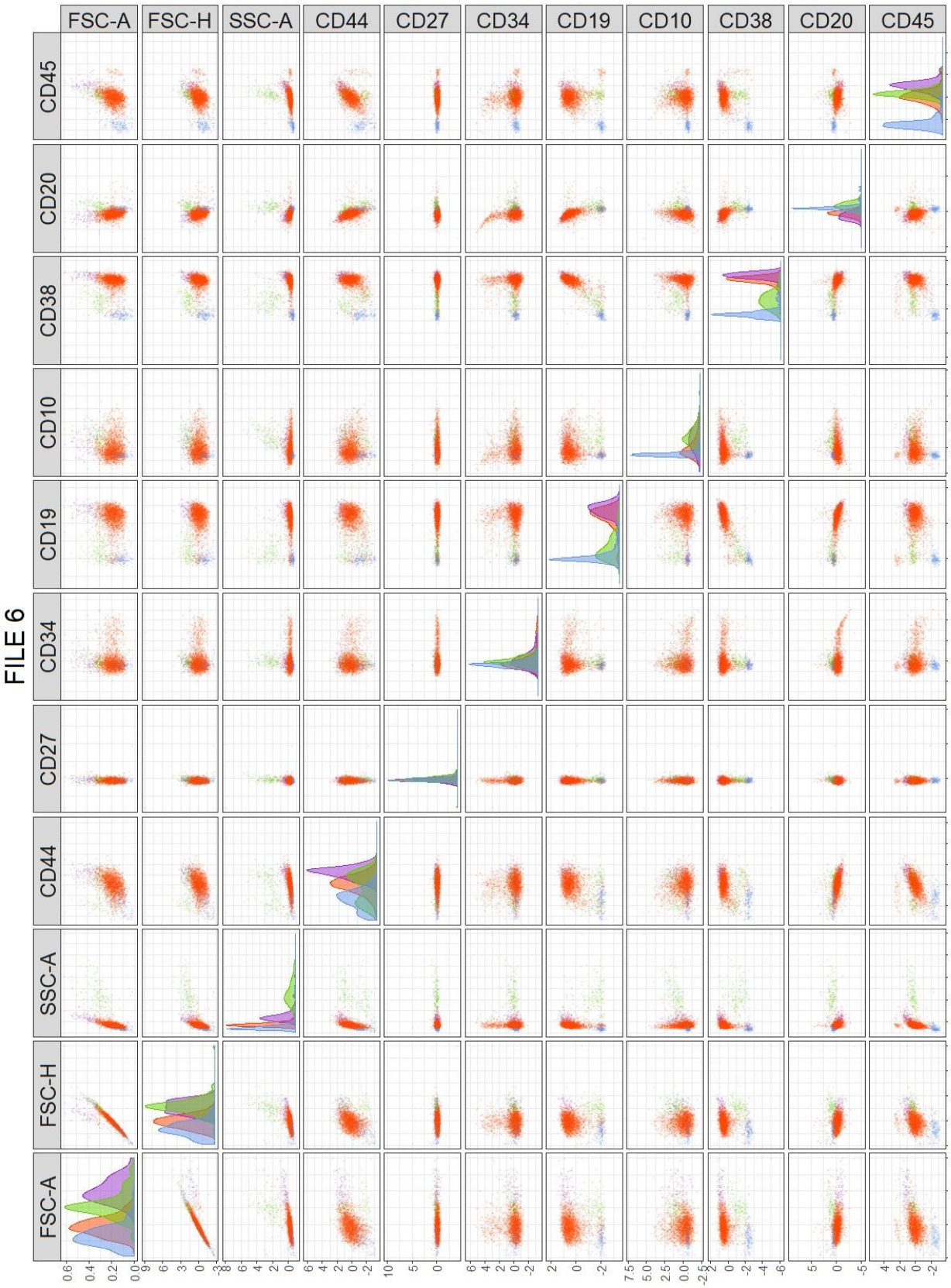
Fig. 15 - Matrices of scatterplots of the resulting FlowSOM clustering of diagnostic files 2-19 in parameters FSC-A, FSC-H, SSC-A, CD44, CD27, CD34, CD19, CD10, CD38, CD20 and CD45. For the clustering, the choose_cluster_fSOM function was used with the range of dimensions of clustering grid set to 6-16 and minimal size of the populations set to 3 %. The histograms of each parameter are on the diagonal. The clusters of size over 3 % from the resulting FlowSOM iteration of clustering are highlighted in colour – red for blastic population, blue for referential negative population and purple, green, yellow and cyan for other populations. The blastic B-ALL population was selected manually based on the results of automated immunophenotypisation in chapter 5.4. In file 18, no blastic population was identified.

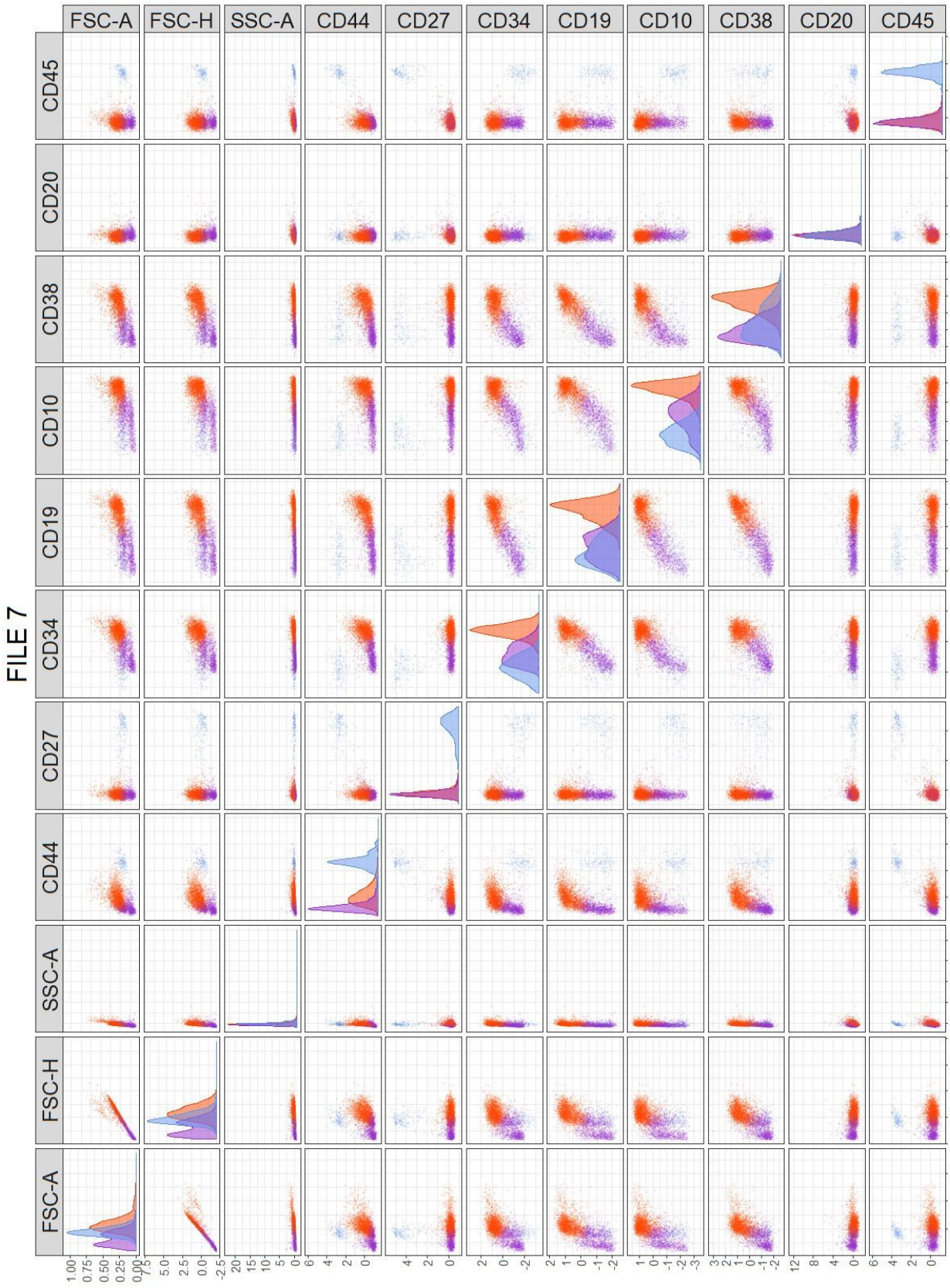

FILE 2
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

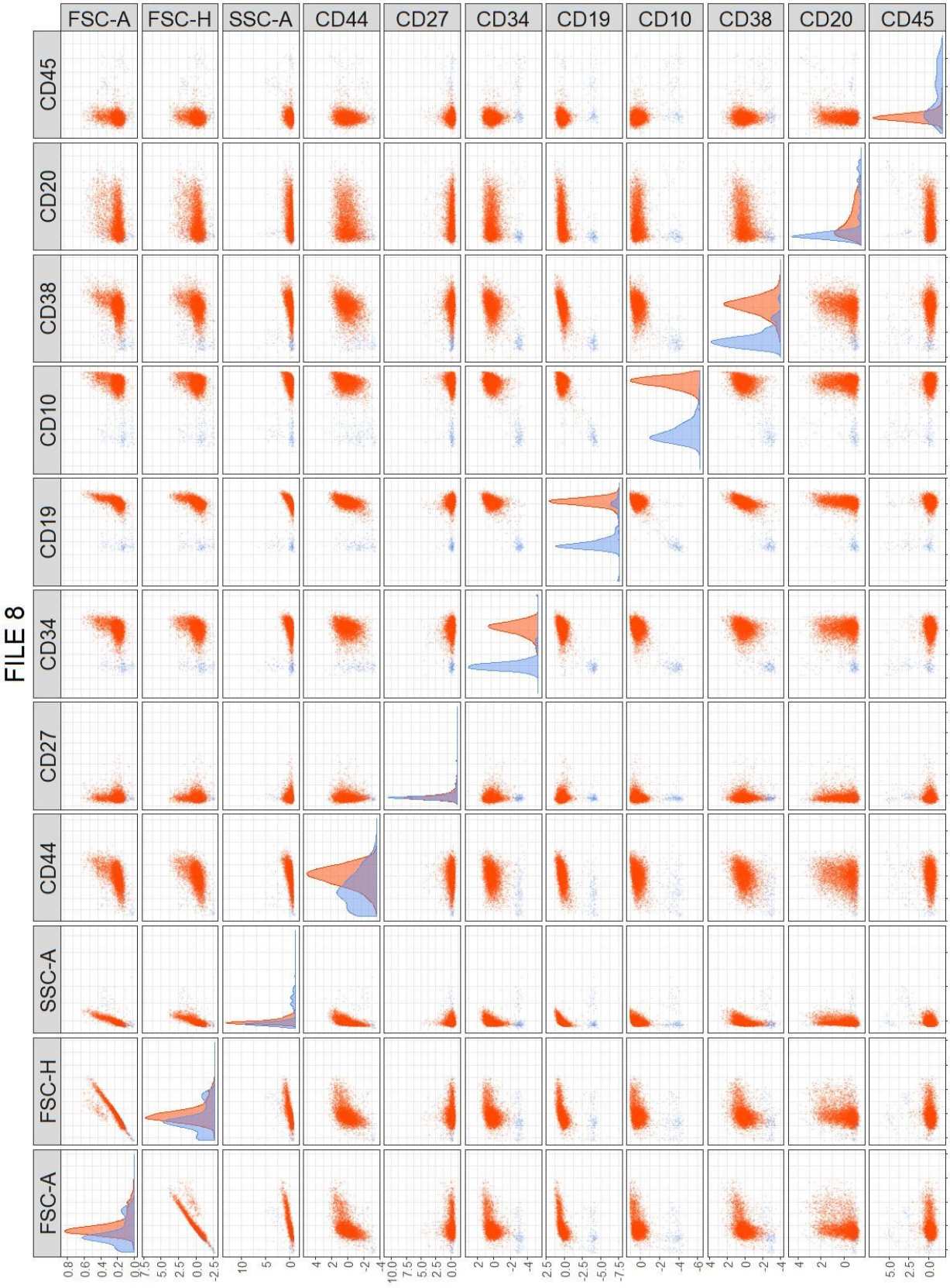

FILE 3
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

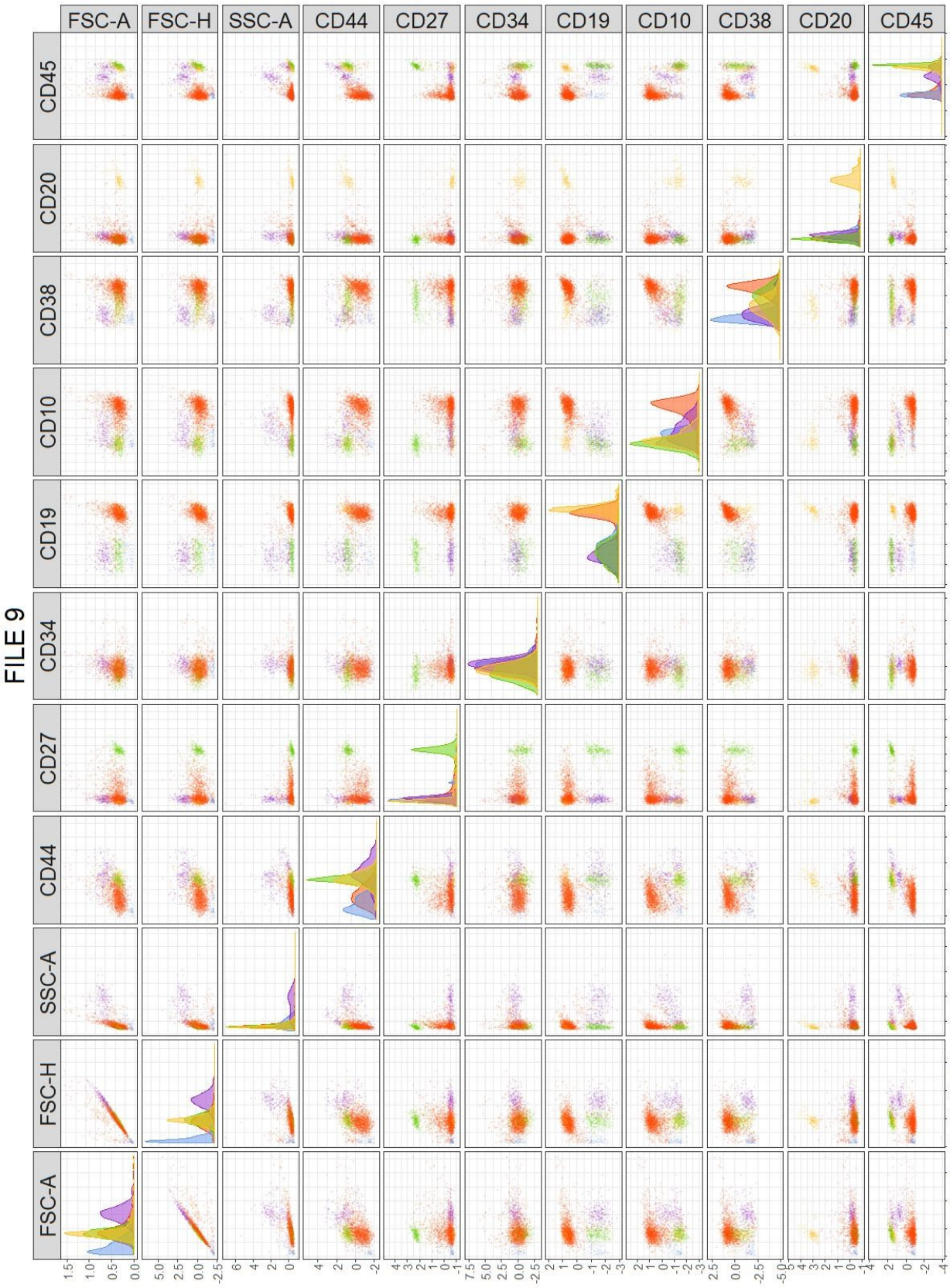

FILE 4
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

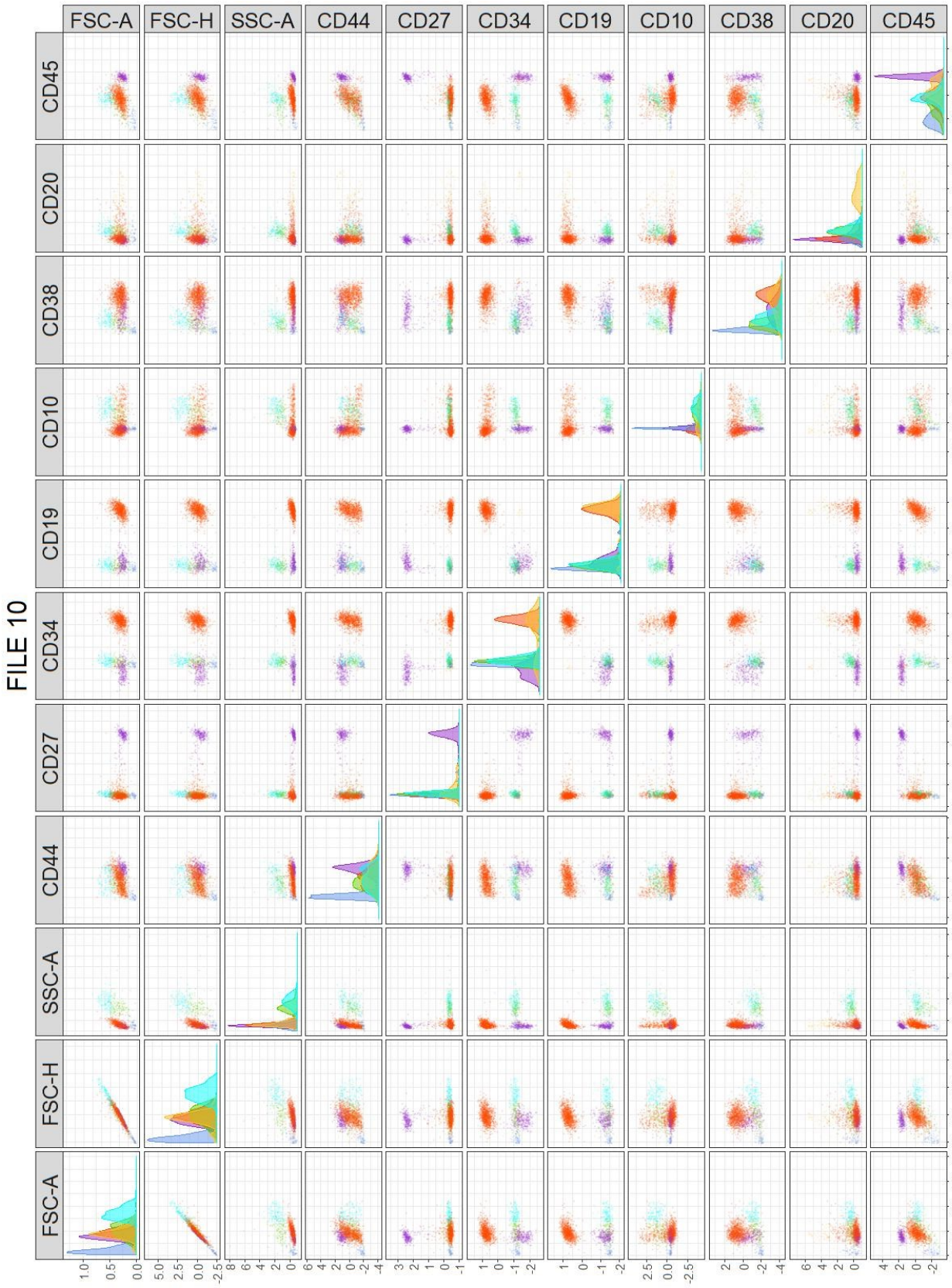

FILE 5
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

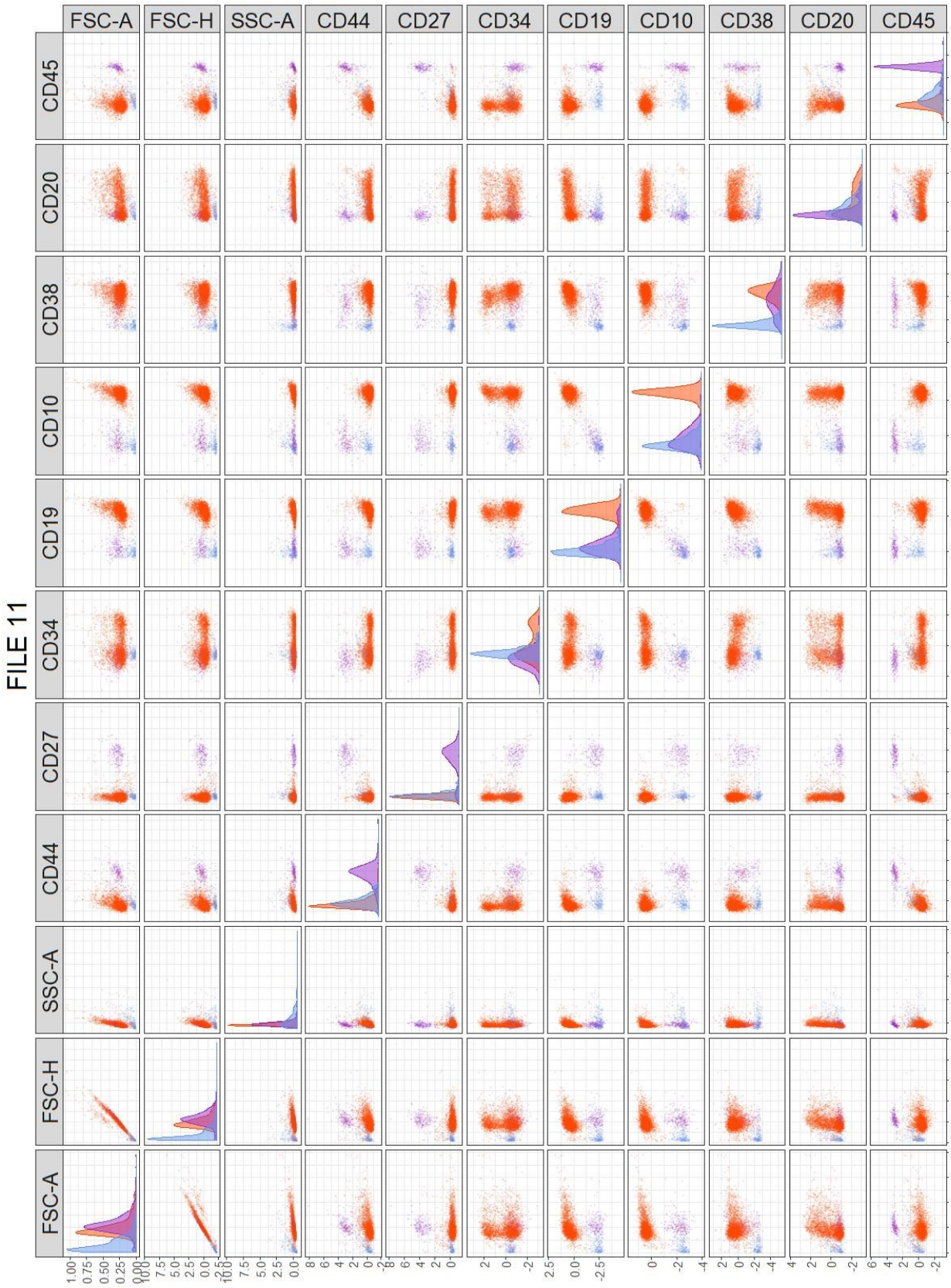

FILE 6
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

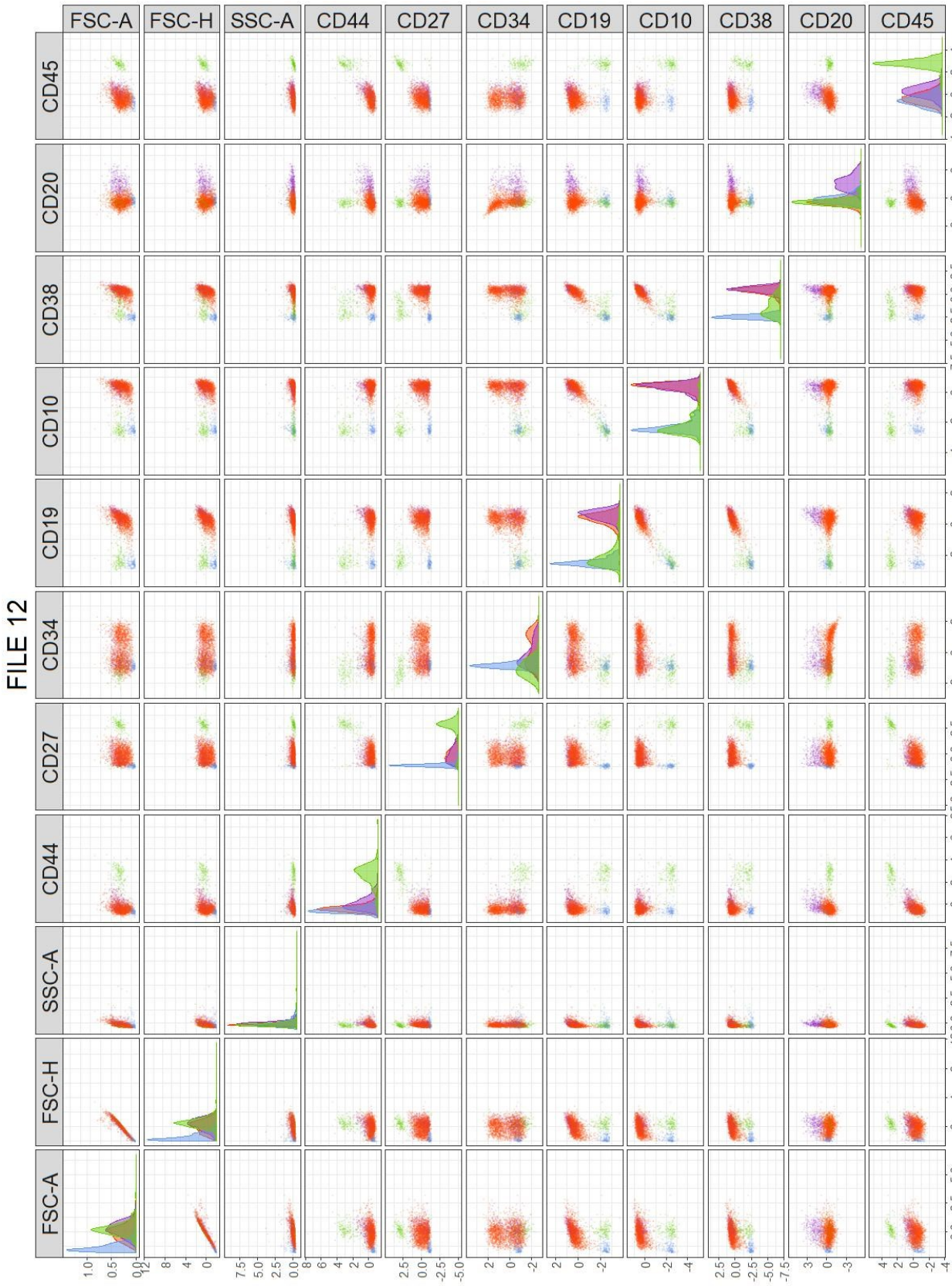

FILE 7
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

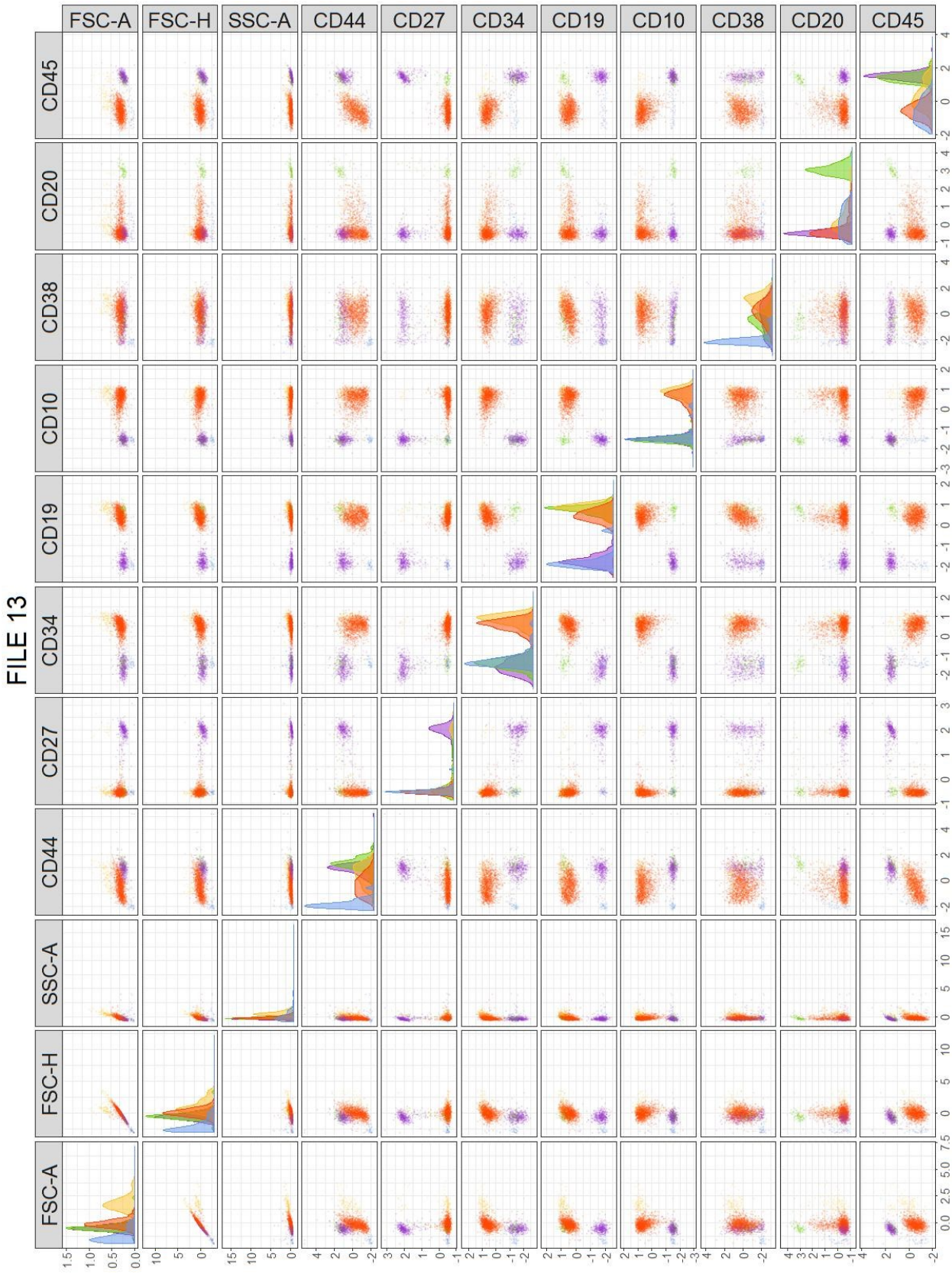

FILE 8
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

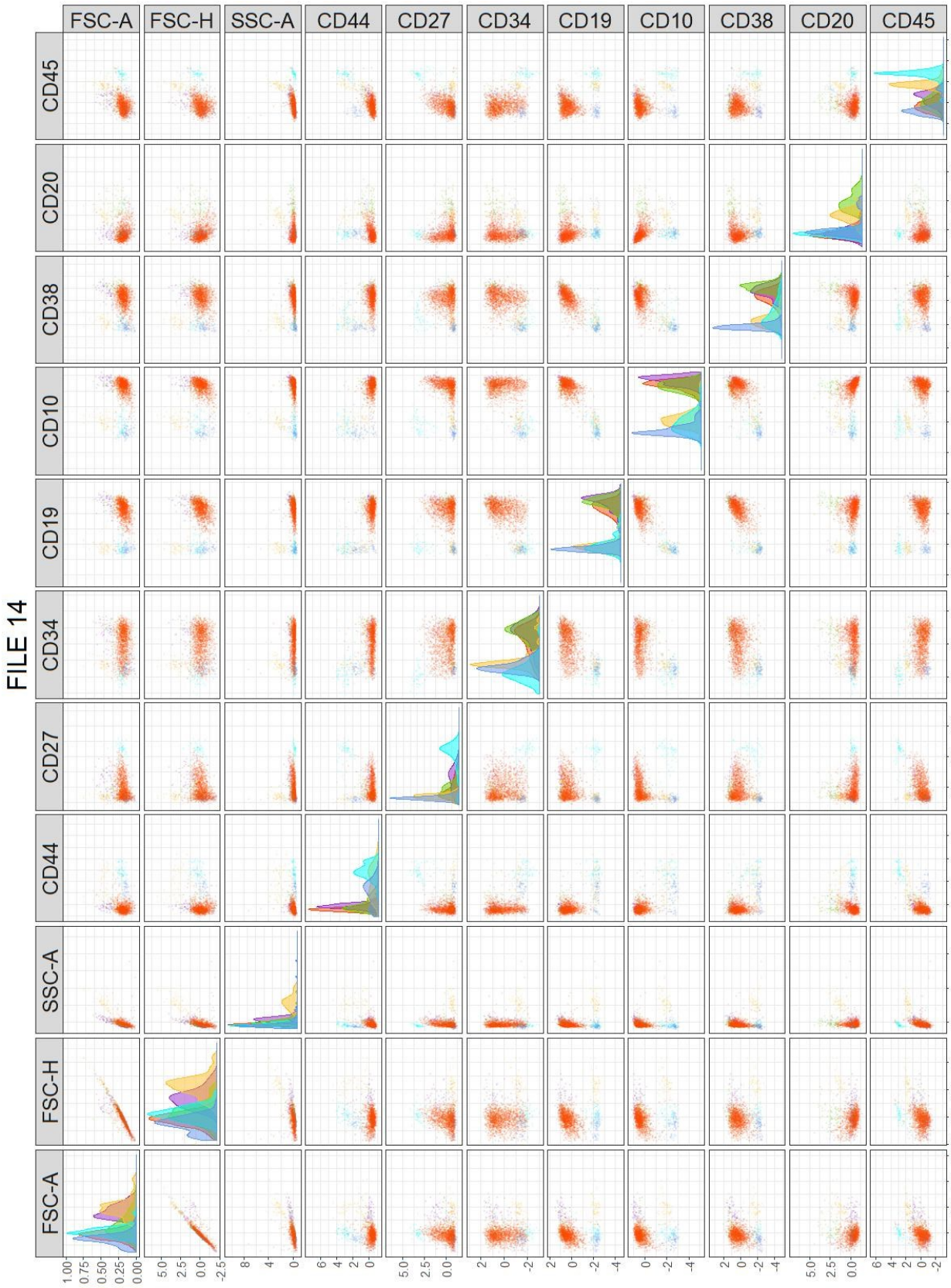

FILE 9
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

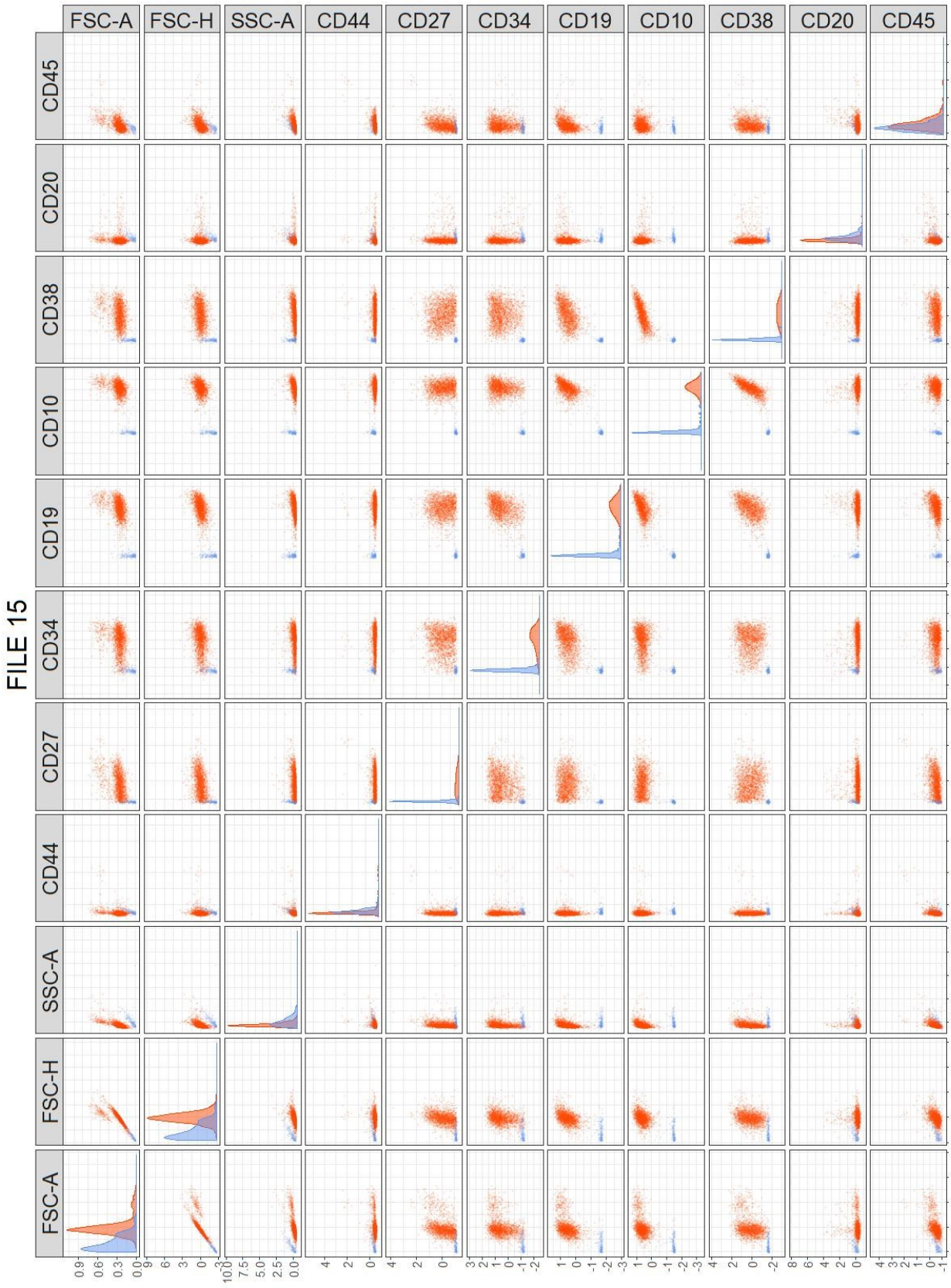

FILE 10
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

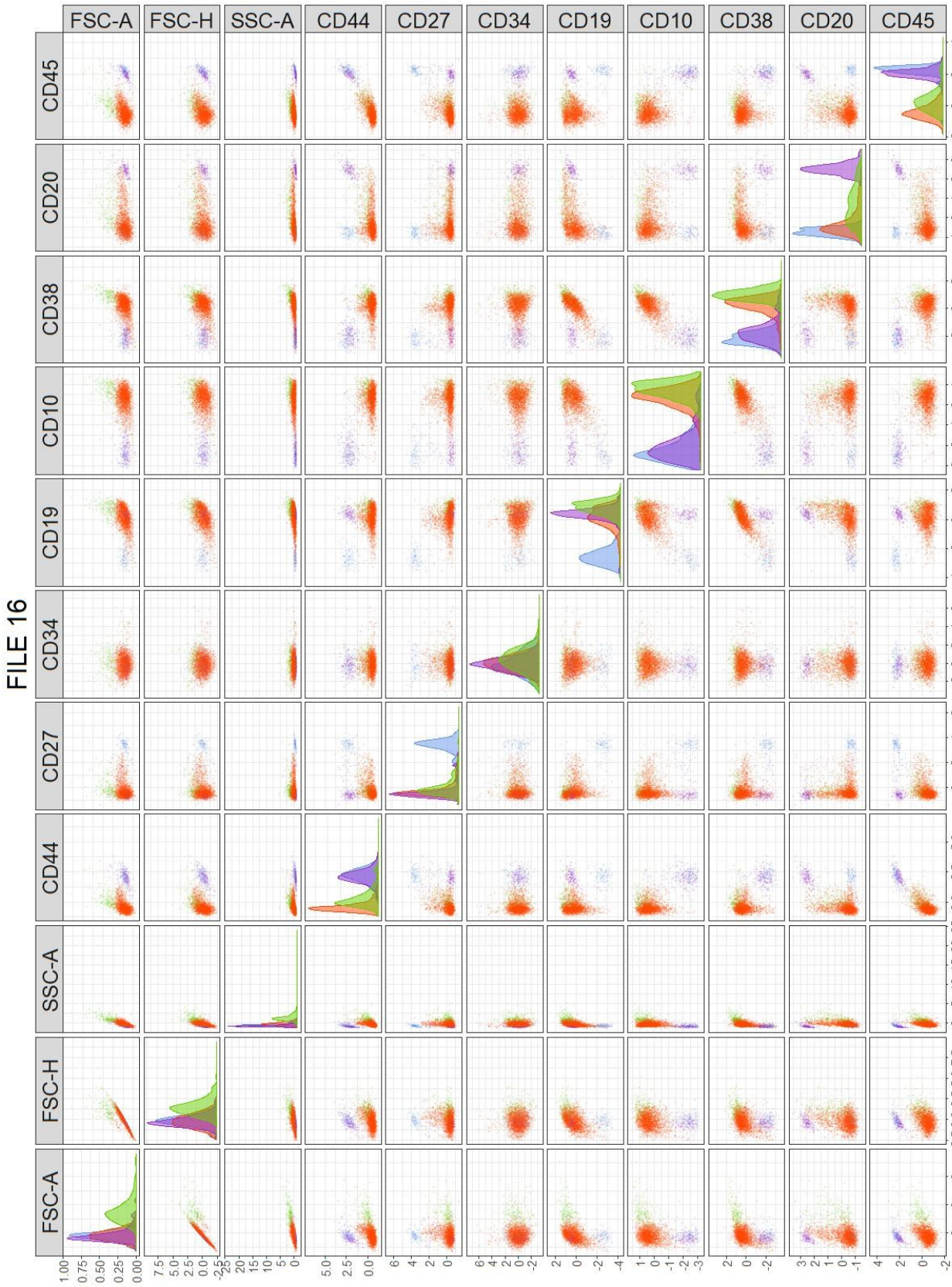

FILE 11
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

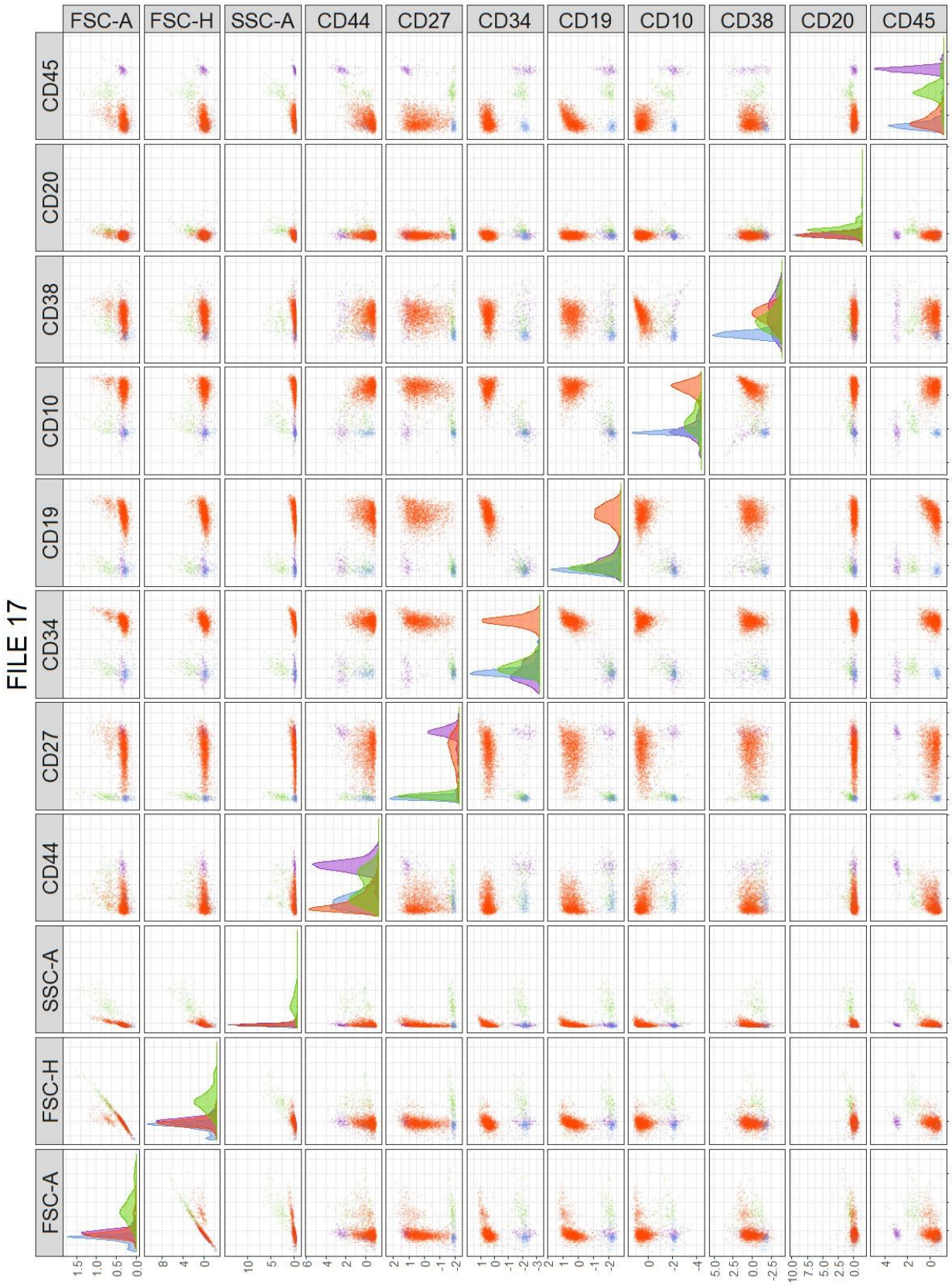

FILE 12
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

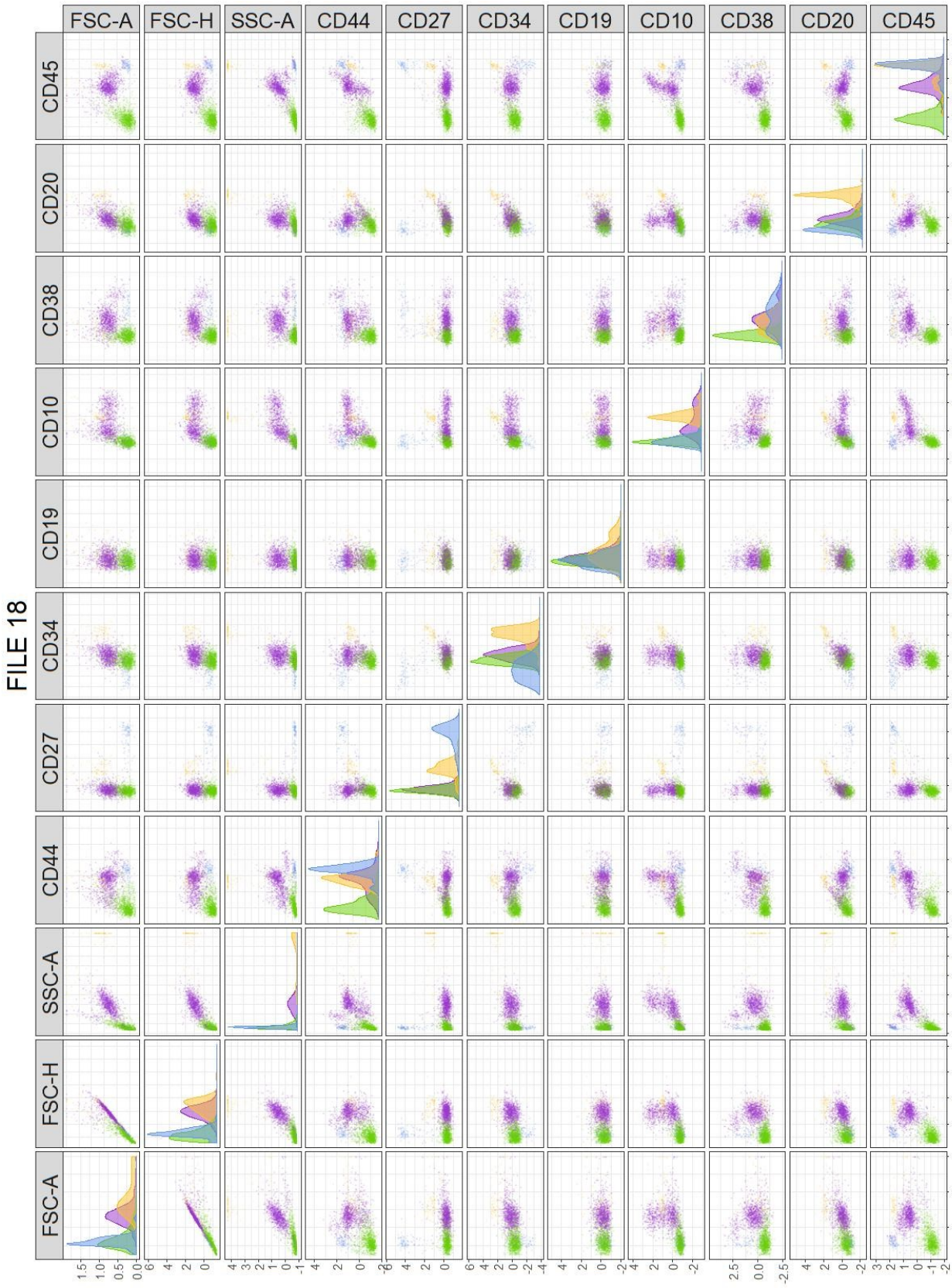

FILE 13
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

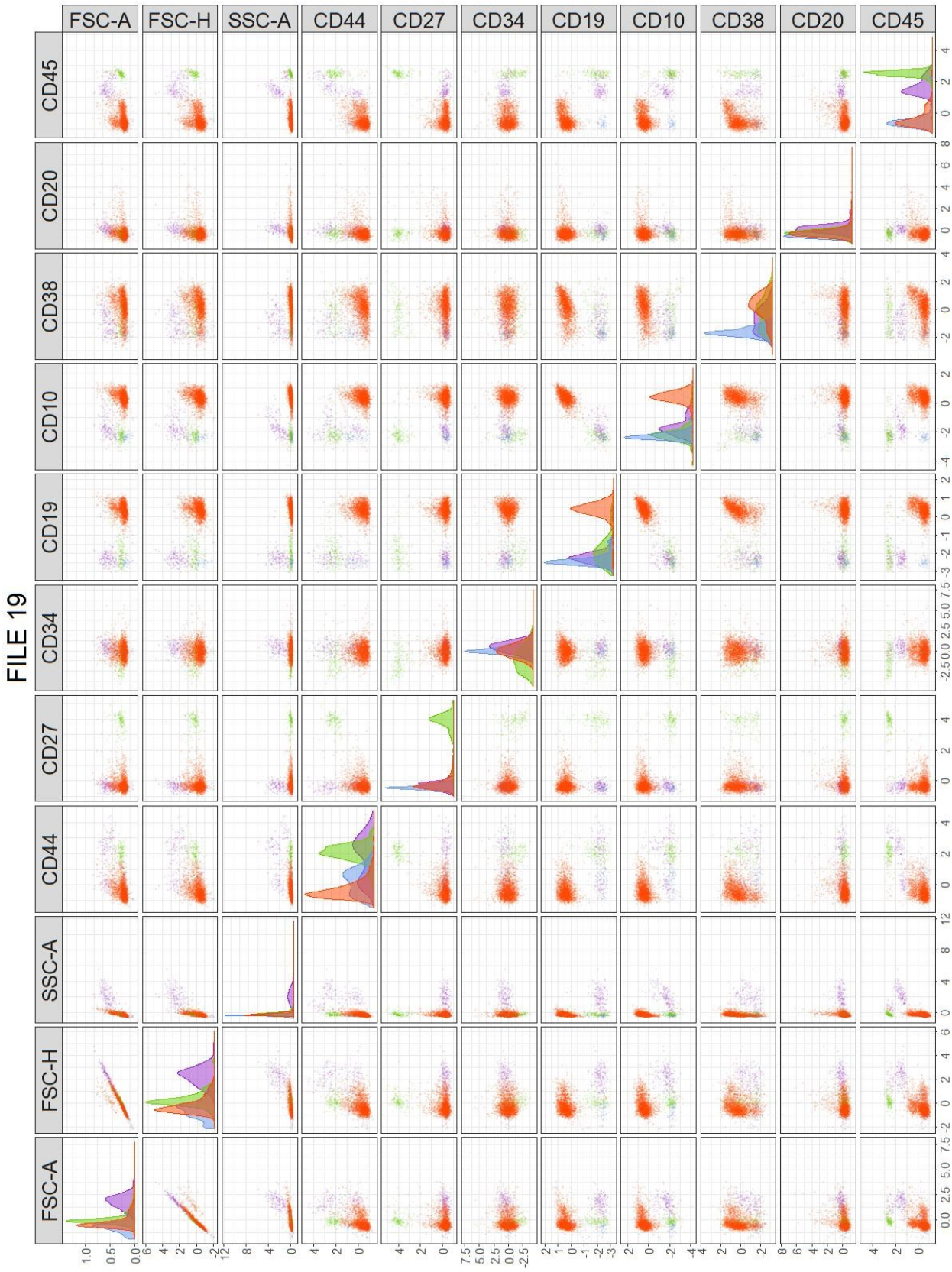

FILE 14
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 15
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 16
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 17
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 18
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45


FILE 19
FSC-A
FSC-H
SSC-A
CD44
CD27
CD34
CD19
CD10
CD38
CD20
CD45

Supplement 3 – Full list of all R libraries used

Table 14 - Full list of R libraries and their dependencies used in this thesis, their versions, source and date installed

R library	version	installation date	library source
BH	1.72.0-3	08.01.2020	CRAN (R 3.6.2)
Biobase	2.46.0	29.10.2019	Bioconductor
BiocGenerics	0.32.0	29.10.2019	Bioconductor
broom	0.5.4	27.01.2020	CRAN (R 3.6.2)
cellranger	1.1.0	27.07.2016	CRAN (R 3.6.1)
cluster	2.1.0	19.06.2019	CRAN (R 3.6.2)
colorspace	1.4-1	18.03.2019	CRAN (R 3.6.1)
ConsensusClusterPlus	1.48.0	02.05.2019	Bioconductor
data.table	1.12.8	09.12.2019	CRAN (R 3.6.2)
DBI	1.1.0	15.12.2019	CRAN (R 3.6.2)
dbplyr	1.4.2	17.06.2019	CRAN (R 3.6.1)
DEoptimR	1.0-8	19.11.2016	CRAN (R 3.6.0)
dplyr	0.8.4	31.01.2020	CRAN (R 3.6.2)
evaluate	0.14	28.05.2019	CRAN (R 3.6.1)
flowCore	1.52.1	04.12.2019	Bioconductor
flowPeel	0.1	12.02.2020	bitbucket (karel_fiser/flowPeel@8961cf7)
FlowSOM	1.16.0	02.05.2019	Bioconductor
flowViz	1.48.0	02.05.2019	Bioconductor
flowWorkspace	3.34.1	02.01.2020	Bioconductor
forcats	0.4.0	17.02.2019	CRAN (R 3.6.1)
gdata	2.18.0	06.06.2017	CRAN (R 3.6.0)
generics	0.0.2	29.11.2018	CRAN (R 3.6.1)
GGally	1.5.0	25.03.2020	CRAN (R 3.6.3)
ggcyto	1.12.0	02.05.2019	Bioconductor
ggplot2	3.2.1	10.08.2019	CRAN (R 3.6.1)
graph	1.62.0	02.05.2019	Bioconductor
gridExtra	2.3	09.09.2017	CRAN (R 3.6.1)
gtable	0.3.0	25.03.2019	CRAN (R 3.6.1)
gtools	3.8.1	26.06.2018	CRAN (R 3.6.0)
haven	2.2.0	08.11.2019	CRAN (R 3.6.1)
hexbin	1.28.1	03.02.2020	CRAN (R 3.6.2)
hms	0.5.3	08.01.2020	CRAN (R 3.6.2)
htmltools	0.4.0	04.10.2019	CRAN (R 3.6.1)
httr	1.4.1	05.08.2019	CRAN (R 3.6.1)
IDPmisc	1.1.20	21.01.2020	CRAN (R 3.6.2)
igraph	1.2.4.2	27.11.2019	CRAN (R 3.6.2)
jpeg	0.1-8.1	24.10.2019	CRAN (R 3.6.1)
jsonlite	1.6.1	02.02.2020	CRAN (R 3.6.2)
kableExtra	1.1.0	16.03.2019	CRAN (R 3.6.1)
KernSmooth	2.23-16	15.10.2019	CRAN (R 3.6.2)
knitr	1.28	06.02.2020	CRAN (R 3.6.2)

R library	version	installation date	library source
lattice	0.20-38	04.11.2018	CRAN (R 3.6.2)
latticeExtra	0.6-29	19.12.2019	CRAN (R 3.6.2)
lazyeval	0.2.2	15.03.2019	CRAN (R 3.6.1)
lifecycle	0.1.0	01.08.2019	CRAN (R 3.6.1)
lubridate	1.7.4	11.04.2018	CRAN (R 3.6.1)
MASS	7.3-51.4	31.03.2019	CRAN (R 3.6.2)
matrixStats	0.55.0	07.09.2019	CRAN (R 3.6.1)
modelr	0.1.5	08.08.2019	CRAN (R 3.6.1)
munsell	0.5.0	12.06.2018	CRAN (R 3.6.1)
mvtnorm	1.0-12	09.01.2020	CRAN (R 3.6.2)
ncdfFlow	2.33.0	20.02.2020	Github (RGLab/ncdfFlow@cc8ac65)
nlme	3.1-142	07.11.2019	CRAN (R 3.6.2)
pcaPP	1.9-73	14.01.2018	CRAN (R 3.6.0)
pheatmap	1.0.12	04.01.2019	CRAN (R 3.6.2)
pillar	1.4.3	20.12.2019	CRAN (R 3.6.2)
pkgconfig	2.0.3	22.09.2019	CRAN (R 3.6.1)
plyr	1.8.5	10.12.2019	CRAN (R 3.6.2)
png	0.1-7	03.12.2013	CRAN (R 3.6.0)
purrr	0.3.3	18.10.2019	CRAN (R 3.6.2)
RColorBrewer	1.1-2	07.12.2014	CRAN (R 3.6.0)
RcppArmadillo	0.9.850.1.0	09.02.2020	CRAN (R 3.6.2)
RcppParallel	4.4.4	27.09.2019	CRAN (R 3.6.1)
readr	1.3.1	21.12.2018	CRAN (R 3.6.1)
readxl	1.3.1	13.03.2019	CRAN (R 3.6.1)
reprex	0.3.0	16.05.2019	CRAN (R 3.6.1)
reshape	0.8.8	23.10.2018	CRAN (R 3.6.3)
Rgraphviz	2.28.0	02.05.2019	Bioconductor
rmarkdown	2.1	20.01.2020	CRAN (R 3.6.2)
robustbase	0.93-5	12.05.2019	CRAN (R 3.6.1)
rrcov	1.5-2	16.01.2020	CRAN (R 3.6.2)
rvest	0.3.5	08.11.2019	CRAN (R 3.6.1)
scales	1.1.0	18.11.2019	CRAN (R 3.6.3)
stringi	1.4.6	17.02.2020	CRAN (R 3.6.2)
stringr	1.4.0	10.02.2019	CRAN (R 3.6.1)
tibble	2.1.3	06.06.2019	CRAN (R 3.6.1)
tidyr	1.0.2	24.01.2020	CRAN (R 3.6.2)
tidyselect	1.0.0	27.01.2020	CRAN (R 3.6.2)
tidyverse	1.3.0	21.11.2019	CRAN (R 3.6.2)
tkrplot	0.0-24	19.07.2018	CRAN (R 3.6.0)
tsne	0.1-3	15.07.2016	CRAN (R 3.6.0)
vctrs	0.2.2	24.01.2020	CRAN (R 3.6.2)
viridisLite	0.3.0	01.02.2018	CRAN (R 3.6.1)
webshot	0.5.2	22.11.2019	CRAN (R 3.6.1)
xfun	0.12	13.01.2020	CRAN (R 3.6.2)
XML	3.99-0.3	20.01.2020	CRAN (R 3.6.2)

R library	version	installation date	library source
xml2	1.2.2	09.08.2019	CRAN (R 3.6.1)
zlibbioc	1.32.0	29.10.2019	Bioconductor