

MASARYK UNIVERSITY

FACULTY OF SCIENCE

Analysis of genomic data in patients with chronic lymphocytic leukemia

Ph.D. Dissertation

Mgr. Petr Taus

Supervisor: Mgr. Karla Plevova, Ph.D.

CEITEC Masaryk University

BRNO 2022

Bibliographic Entry

Author: Petr Taus, M.Sc.
Faculty of Science, Masaryk University
CEITEC Masaryk University

Title of Thesis: Analysis of genomic data in patients with chronic lymphocytic leukemia

Degree program: Life Science

Field of Study: Bio-omics

Supervisor: Mgr. Karla Plevova, Ph.D.

Academic Year: 2022/2023

Number of Pages: 124

Keywords: Chronic Lymphocytic Leukemia; Pathway mutation score, Single-cell RNA sequencing; Data analysis; B cells

I hereby declare on my honor that I have carried out the work on this dissertation thesis independently under the supervision of Mgr. Karla Plevova, Ph.D., using the published sources clearly indicated and quoted in the bibliographic section of this work.

.....

Mgr. Petr Taus

Acknowledgment

First of all, I would like to thank my supervisor Karla Plevova for her guidance, encouragement, and willingness to help throughout the years. Then, I wish to thank the head of the Medical Genomics research group Sarka Pospisilova for letting me work in this group. I wish to thank all collaborators that trusted me with their data. Namely, I want to thank Vitezslav Bryja and Pavlina Janovska for initiating the scRNA-seq project presented in this thesis and giving me the opportunity to participate in it. Furthermore, I would like to thank everyone from the GeneCore group at EMBL in Heidelberg and Julio-Rodriguez Saez's group at Heidelberg University for letting me spend some time in their laboratories and learn a lot from them. Finally, I would like to express my gratitude to all my colleagues and friends that have helped and supported me on this journey.

Abstract

Chronic lymphocytic leukemia (CLL) is an incurable lymphoproliferative malignancy of CD5+ B cells characterized by a highly variable clinical course and genetic complexity. Despite the fact that CLL genetic landscape has been well-characterized, translation of this knowledge into clinical practice remains elusive. The main culprit preventing patient stratification based on mutational profiles is the heterogeneity and sparsity of mutation data.

In this thesis, I set two main goals focusing on establishing the two distinct computational workflows primarily applied in the context of CLL. The first focused on stratifying patients based on their somatic mutation profile. We built the workflow using 506 CLL patients' whole exome sequencing data gathered by the International Cancer Genome Consortium. Firstly, we decreased data heterogeneity and sparsity by mapping mutated genes to biological processes and calculated pathway mutation scores. Then we applied ensemble clustering and extracted abnormal molecular pathways with a machine learning approach. We identified four patient groups with specific pathway mutational profiles differing in time to first treatment.

The second goal was to establish a computational workflow for the analysis of single-cell RNA sequencing (scRNA-seq) data. The main objective was to explore the cellular origin of CLL, on which multiple theories exist, but no consensus has been reached. Moreover, it is unclear whether the disease is derived from single or multiple precursors and at what stage the transformation occurs. We previously identified a rare B cell population in the peripheral blood of healthy adults expressing a CLL-specific gene *ROR1*. We decided to characterize this population that we hypothesized to be the cellular origin of CLL. In this project, we combined our datasets with publicly available data. We created a reference of healthy B cells and sorted ROR1+ B cells. To identify the most similar population to CLL cells, we leveraged the extreme gradient boosting algorithm to build a classification model for predicting the transcriptomic similarity of healthy B cell subsets to malignant cells. We found that cellular clusters enriched for ROR1+ B cells are more similar to CLL cells than any other in our reference. In addition, we observed that ROR1+ B cells resemble anergic-like cells, which is in line with the hypothesis suggesting the origin of CLL from autoreactive B cells.

In conclusion, this thesis presents two computational workflows and examples of their successful implementation in the study of CLL, and demonstrates their application beyond CLL to multiple different biological questions.

Abstrakt

Chronická lymfocytární leukémie (CLL) je nevléčitelná lymfoproliferativní malignita CD5+ B buněk charakterizovaná vysoce variabilním klinickým průběhem a genetickou komplexitou. Navzdory skutečnosti, že genetické pozadí CLL je dobře charakterizováno, přenos těchto poznatků do klinické praxe naráží na mnohá úskalí. Hlavní příčinou, která brání stratifikaci pacientů na základě mutačních profilů, je heterogenita a řídkost mutačních dat.

V této práci byly stanoveny dva hlavní cíle zaměřené na vytvoření odlišných výpočetních postupů použitých primárně v kontextu CLL. V rámci prvního cíle jsme stratifikovali pacienty na základě profilu somatických mutací. K vytvoření výpočetního postupu jsme využili sekvenční celoxomová data od 506 CLL pacientů shromážděná organizací International Cancer Genome Consortium. V prvním kroku jsme namapovali mutované geny na biologické procesy a vypočítali dráhové mutační skóre, čímž jsme snížili heterogenitu a řídkost dat. Poté jsme využili tzv. „ensemble“ klastrovací algoritmus a pomocí strojového učení extrahovali abnormální molekulární dráhy. Identifikovali jsme čtyři skupiny pacientů se specifickými profily dráhového mutačního skóre lišící se v čase do první léčby.

Druhý cíl byl zaměřen na vytvoření výpočetního postupu pro analýzu dat z RNA sekvenování jednotlivých buněk (scRNA-seq). Hlavním záměrem bylo identifikovat buněčný původ CLL, ke kterému se vztahuje mnoho teorií, stále však neexistuje žádný konsensus. Navíc není jasné, zda onemocnění vzniká z jednoho či více buněčných prekurzorů, a v jaké fázi buněčného vývoje k maligní transformaci dochází. V minulosti se nám podařilo identifikovat vzácnou populaci B buněk v periferní krvi zdravých dospělých, která exprimovala CLL-specifický gen *ROR1*. Tuto populaci, o níž jsme předpokládali, že by mohla být buněčným původem CLL, jsme se rozhodli detailně charakterizovat. V rámci tohoto projektu jsme využili vlastní data i veřejně dostupné datasety. Vytvořili jsme referenci zdravých B lymfocytů a sortovaných *ROR1*+ B lymfocytů. Pro detekci populace, která se nejvíce podobá CLL buňkám, jsme využili tzv. „extreme gradient boosting“ algoritmus, pomocí kterého jsme vytvořili klasifikační model pro predikci transkriptomické podobnosti různých podskupin zdravých B buněk s maligními buňkami. Zjistili jsme, že buněčné klastry obohacené o *ROR1*+ B buňky jsou podobnější buňkám CLL než kterékoli jiné buňky v naší referenci. Dále jsme pozorovali, že *ROR1*+ B buňky se podobají anergickým B buňkám, což je v souladu s hypotézou naznačující původ CLL z autoreaktivních B buněk.

Tato práce představuje dva výpočetní postupy a příklady jejich úspěšné implementace při studiu CLL, a současně demonstruje jejich aplikaci při hledání odpovědí na odlišné biologické otázky nad rámec CLL.

Table of Contents

Abstract	5
Abstrakt	6
1 Aims of the thesis.....	8
2 Normal B cell development	9
3 Chronic lymphocytic leukemia	12
3.1 CLL molecular heterogeneity.....	12
3.2 Monoclonal B cell lymphocytosis	14
3.3 CLL cell of origin	14
3.4 Expression of ROR1 as a stable marker of CLL.....	16
4 Genomic mutation data analysis	18
4.1 Genomic mutation subtype identification	18
4.2 Results and discussion	19
5 Single-cell RNA sequencing data analysis	20
5.1 Single-cell RNA-sequencing data analysis introduction	20
5.2 Data integration	24
5.3 Cell type annotation	25
5.4 Biological activity estimation	25
5.5 Trajectory inference	26
5.6 Results and discussion – the cellular origin of CLL	26
5.6.1 Establishing scRNA-seq method and computational pipeline	27
5.6.2 Characterization of ROR1+ B cells	28
5.6.3 Similarity prediction using a machine learning approach	29
5.6.4 Summary.....	35
6 Other articles related to the thesis	36
6.1 Commentary on published articles.....	36
6.2 Summary of a manuscript under review.....	38
6.3 Summary of unpublished work	40
References.....	41

Conference contributions

Research article 1

Research article 2

Research article 3

Research article 4

1 Aims of the thesis

The work presented in this thesis aimed to create two computational workflows to investigate the genetic and cellular complexity of CLL and its normal cellular counterpart.

While the genetic landscape of CLL has been well characterized, successful patients risk stratification based on mutation profile is missing. Using publicly available genetic data and clinical information, this thesis explored CLL patient stratification based on genetic data transformed into biological pathway mutation score.

Multiple hypotheses on the cell of origin of CLL have been proposed, but no consensus has been established. Using publicly available and in-house scRNA-seq data, this thesis interrogated a rare B cell population expressing a CLL-specific gene *ROR1* identified in healthy adults.

The scRNA-seq workflow presented in this thesis was primarily developed for the investigation of CLL and its normal cellular counterpart. Its secondary aim was to investigate diverse biological questions of the researchers from a consortium of collaborating laboratories in Brno.

The scope of this thesis can be summarized as follow:

1. Set up a computational workflow for CLL patient risk stratification based on publicly available mutation and clinical data.
2. Set up a computational workflow for scRNA-seq data analysis.
3. Investigate the cellular origin of CLL using publicly available and in-house scRNA-seq data.
4. Apply the workflow to distinct projects from a consortium of collaborating laboratories in Brno.

2 Normal B cell development

B cells are vital players of adaptive immunity providing protection against pathogens via the production of antibodies. Upon antigen recognition, they differentiate into antibody-producing plasma cells either through transient germinal center (GC) reaction or rapid extrafollicular pathway resulting in short-lived plasma cells¹.

In addition to the production of antibodies, B cells can regulate the immune response by secreting proinflammatory molecules such as TNF-alpha and IL-6 or immunosuppressive such as IL-10². Moreover, B cells can act as antigen-presenting cells, engage in T cell activation, participate in the innate part of the immune system¹, or even emulate the dendritic cells' role by transferring part of dendritic cells' membrane with molecular complexes to their surface³.

B cell development and commitment to the B cell lineages starts in the fetal liver and continuously transition to the bone marrow (BM), where most postnatal B cell development occurs. However, recent findings suggest that B lymphopoiesis might be more widespread than previously thought^{4,5,6}. In BM, immature B cells develop from hematopoietic stem cells (HSC). It involves several differentiation steps during which V, (D,) and J gene segments of immunoglobulin heavy and light chain are being rearranged (Figure 1). Immunoglobulins, together with CD79A/B molecules, form a B cell receptor that functions as an antigen receptor. This stochastic process is tightly regulated by mechanisms of central tolerance to eliminate naturally emerging autoreactive B cells. Most frequently, autoreactive B cells go through a process called receptor editing, which can rescue B cells from deletion⁷. Nevertheless, likely due to the limited presence of self-antigens in the BM microenvironment, not all autoreactive B cells are eliminated during the central tolerance checkpoint. In fact, it is estimated that around 40% of immature B cells that leave BM and migrate to secondary lymphoid organs are autoreactive⁸.

The immature B cells emerging from BM, called transitional B cells, continue their maturation process through the T1, T2, and T3 stages to naive B cells in the spleen, where they have to pass a second tolerance checkpoint⁹. During this phase, a percentage of autoreactive B cells within a mature B cell population is further reduced to ~20% either by clonal deletion or by induction of an anergic, functionally silenced state⁸. It is hypothesized that natural functions of weakly self-reactive B cells are, for instance, defense against pathogens that leverage molecular mimicry to evade the host immune system¹⁰, suppression of inflammatory processes by masking self-antigens, or removal of cellular debris from dying cells¹¹. Therefore, it is possible that the relatively high number of autoreactive B cells in the periphery is not a bug but a feature of the system.

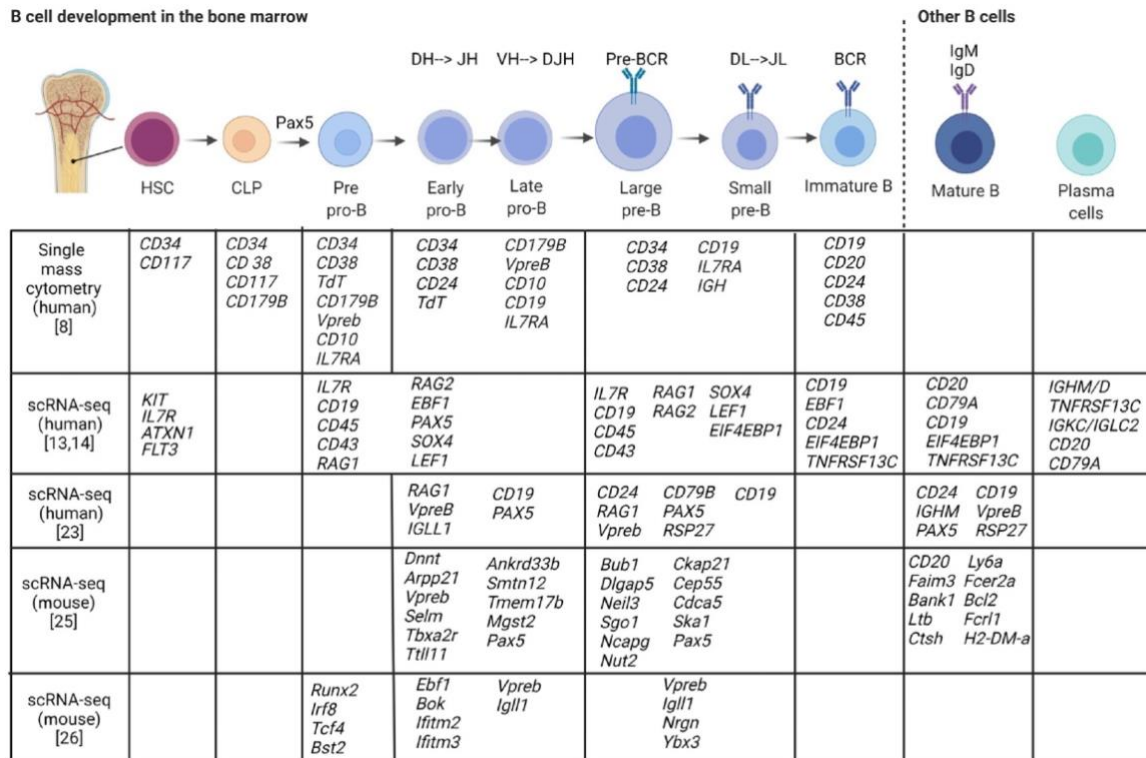


Figure 1: Gene signatures in B cell development (adopted from Morgan et al.¹²). HSC, hematopoietic stem cell; CLP, common lymphoid progenitor; BCR, B cell receptor.

One distinct B cell subset implicated in the autoreactive B cell response are the marginal zone B cells (MZB). MZBs are innate-like CD1c⁺ cells named based on the microanatomical niche in the spleen that they occupy. Additionally, MZ-like B cells were identified circulating in peripheral blood¹³. Due to the expression of memory marker CD27, they were thought to be IgM⁺ memory B cells (MBC). Nevertheless, Zho et al.¹⁴ presented data suggesting that MZBs migrate to gut-associated lymphoid tissue (GALT) to complete their maturation. Thus, it is possible that circulating MZ-like B cells represent MZBs transitioning between the spleen and GALT. Moreover, during the developmental process in the spleen, a subset of T2 cells characterized by a high level of IgM was suggested to be a precursor of MZBs¹⁵. Therefore, these recent findings indicate that MZBs do not share a differentiation route with MBCs. Still, the existence of a separate MZB lineage remains a subject of debate¹⁶.

Naive B cells become activated upon first antigen recognition, which results in the initiation of three distinct molecular programs generating either plasmablasts, early MBCs, or GC B cells¹. GC is a transient micro-anatomical structure consisting of two functionally different regions – the dark zone (DZ) and the light zone (LZ). In the DZ, B cells proliferate and undergo the somatic hypermutation process (SHM). Then the LZ B cells can undergo immunoglobulin class-switch recombination (CSR) and differentiate either into MBCs, plasma cells, or return to the DZ. The GC reaction is an iterative process

during which B cell survival depends on costimulatory signals from CD4⁺ follicular T helper cells. This is yet another tolerance checkpoint during which the autoreactive B cell pool is reduced. Interestingly, SHM can decrease B cell receptor (BCR) affinity for autoantigens of autoreactive B cells in the process termed clonal redemption¹⁷.

Notably, a growing body of literature challenges the dogma that SHM and CSR are limited to the GC^{1,18}. This, as will be shown later in the presented thesis, is especially relevant and important to note in the context of studying the cellular origin of CLL.

Recent advances in single-cell transcriptomics and proteomics led to the discovery of previously unappreciated heterogeneity within the B cell subsets defined by the immunophenotypic fluorescence-activated cell sorting (FACS) separation^{19,12}. Glass et al.²⁰ described twelve unique B cell subsets in peripheral blood (PB), tonsils, lymph nodes and BM using a time of flight cytometry. Of note, they identified a CD45RB⁺CD27⁻ early memory population that can be precisely detected only with a proteomic approach and not with RNA-seq. Furthermore, they did not find a population of cells phenotypically resembling a controversial subset of innate-like B1 B cells, which is prevalent and well-defined in mice²¹. However, a recent pan-organ scRNA-seq study focused on the prenatal development of the human immune system revealed and characterized the CD5⁺ B1 B cell population⁴. The two aforementioned observations are, in fact, aligned with the current knowledge derived from mice that B1 B cells are produced in the fetal and neonatal period and subsequently are predominantly located in body cavities²¹. Apart from the expression of previously reported markers, such as CLL diagnostic marker gene CD5, the human B1 B cells possessed a capacity for self-renewal, showed features of tonic BCR signaling, and spontaneously secreted antibodies⁴.

The following two scRNA-seq studies demonstrated the presence of atypical B cells in healthy adults^{22,23}. These cells are characterized by high expression of inhibitory receptors such as Fc-receptor-like molecules (FCRL4/5), the transcription factor T-bet (TBX21) and the integrin CD11c, and reduced expression of CD21, a co-receptor for BCR, and CD27. The atypical B cells are primarily generated via the extrafollicular pathway, however, data in the literature suggest that the GC pathway is also possible²⁴. The differentiation route is likely determined by the conditions of the immune response. The atypical B cells have been predominantly identified and studied in the context of aging, cancer, viral infections, or autoimmunity, and their nomenclature greatly differs between studies. For example, they have been called double negative (DN, i.e., lacking CD27 and IgD), age-associated, inflammatory, or exhausted. Recent studies showed that they should not be considered as an exhausted and dysfunctional B cell subset as they can respond to membrane-arrayed antigens and can function as potent antigen-presenting cells. However, their role in systemic immune responses remains largely unclear and needs to be further explored²⁴.

Sutton et al.²² presented data suggesting that atypical B cells, together with alternative MBCs, are part of a broader alternative developmental pathway. A high expression of IgM characterizes the alternative B cell lineage in PB of healthy donors,

which has also been described by Stewart et al.²³. They performed a scRNA-seq analysis of five immunophenotypically sorted B cell subsets (transitional, naive, IgM memory, classical memory, and DN). Interestingly, they found subpopulations of transcriptomically distinct DN cells associated with both classical and alternative lineages. However, only the population of DN cells within the alternative lineage matched the description of the atypical B cells. Moreover, the alternative lineage contained a subpopulation of cells strongly expressing a MZB marker CD1c. Overall, the studies above agreed on the existence of at least two distinct B cell lineages in PB. Although using distinct terminology, both Sutton et al.²² and Stewart et al.²³ described phenotypically similar B cell populations.

3 Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) is a lymphoproliferative malignancy characterized by $>5 \times 10^9/L$ peripheral CD5+ mature B cells with very heterogeneous biological and clinical behavior. CLL is the most prevalent form of adult leukemia in the Western world, with a higher incidence in men (6.8 cases per 100,000 in men vs. 3.5 cases per 100,000 in women) and the median age at diagnosis around 70 years²⁵. Despite the development of novel targeted inhibitors blocking pro-survival B cell receptor signaling (e.g., ibrutinib) or anti-apoptotic BCL2 signaling (venetoclax), CLL remains an incurable disease. The goals of therapy are to prolong survival and improve patients' well-being. The clinical manifestations range from a stable condition with no need for treatment to aggressive with frequent relapses and overall survival of less than three years. In rare cases, CLL even transforms into an aggressive lymphoma, such as diffuse large B cell lymphoma (so-called Richter transformation)²⁵.

The initial clinical prognostication in CLL relies on the Rai²⁶ or Binet²⁷ clinical staging systems. The following clinical examinations include molecular genetics tests for *TP53* mutation detection and analysis of the somatic hypermutation status of the immunoglobulin heavy chain variable region genes (IGHV)²⁸. Typically, patients with mutated IGHV genes have a better prognosis than patients with unmutated IGHV genes^{29,30}. Moreover, around 30-40% of all CLL cases can be grouped, based on IGHV usage and complementarity-determining regions, into so-called stereotyped subsets associated with a specific clinical course³¹.

3.1 CLL molecular heterogeneity

Over the past decade, molecular heterogeneity of CLL has been increasingly associated with the clinical outcome of patients²⁵. The most studied genetic alterations in CLL are single-nucleotide polymorphisms and chromosomal alterations. The complexity of the karyotype abnormalities, including recurrent deletions on chromosomes 11, 13, 17, and

trisomy 12, possesses not only prognostic value^{32,33} but also appears to be predictive in the context of treatment³⁴.

Advances in next-generation sequencing in the last decade brought insight into the mutational landscape^{35,36} and subclonal heterogeneity of CLL^{37,38}. However, the translation of these findings into clinical practice proved to be challenging. This is mainly due to a long tail of genes mutated at low frequencies where only a few are mutated in more than 5% of patients at diagnosis (*SF3B1*, *NOTCH1*, *TP53*, and *ATM*)^{35,36}. Only mutations in *TP53* and *ATM* genes were confidently associated with worse clinical outcomes. The reports of the prognostic relevance of other driver mutations (e.g., in *NOTCH1* and *SF3B1*) vary^{39,40}.

Despite an enormous genetic heterogeneity, when zooming out from the gene level to the molecular pathway level, we discover that mutations cluster in a handful of cellular processes (e.g., *NOTCH1* signaling, BCR receptor signaling, chromatin modifiers, and cell cycle)^{35,36} that have been associated with prognosis^{41,42} (Figure 2). In our work, we exploited this phenomenon to reduce data heterogeneity by calculating pathway mutation scores and identifying clinically relevant subtypes defined by their pathway mutation profiles⁴³.

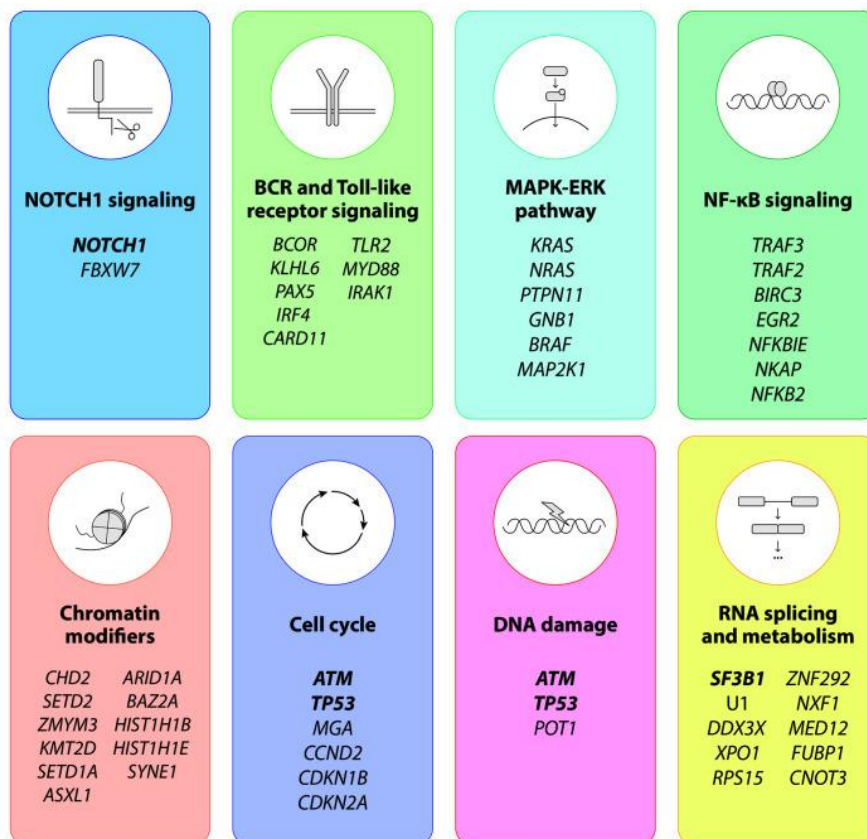


Figure 2: Recurrently affected molecular pathways in CLL (adopted from Delgado et al.²⁵). Genes highlighted in bold are mutated at higher frequencies in newly diagnosed patients (>5%).

In addition to genetic complexity, CLL was shown to be heterogeneous on an epigenomic level, too^{44,45,46}. Based on the similarity of CLL DNA methylation profiles to different B cell counterparts, CLL patients can be stratified into three epigenetic groups with prognostic relevance: naive B cell-like, memory B cell-like, and intermediate CLL^{44,47,48}. Nevertheless, genomic, epigenomic, and other biological layers, such as transcriptomics, are strongly interconnected. Therefore, analyzing omics data separately will likely never reveal a complete picture of the disease biology, as was beautifully demonstrated recently by Lu et al.²⁴. They discovered a new biological axis of heterogeneity in CLL using integrated analysis of genomic, transcriptomic, DNA methylation, and *ex vivo* drug response data.

3.2 Monoclonal B cell lymphocytosis

CLL is preceded by an asymptomatic condition called monoclonal B cell lymphocytosis (MBL)⁵⁰. MBL is, in most cases, indistinguishable at the genomic, transcriptomic, and epigenomic levels from CLL assigned to the same IGHV subgroup, yet MBL cells contain a lower burden of disease-driver alterations than CLL cells³⁶. Its diagnosis is based on the presence of a circulating monoclonal or oligoclonal B cell population in the PB, and it is further subdivided into low- or high-count based on whether the B cell count is above or below $0.5 \times 10^9/L$.

The progression rate of high-count MBL to CLL is around 1-2% per year, while the progression of low-count MBL to CLL is rare. MBL was detected in the healthy age-matched population with a frequency of 12%, which increases two- to threefold in relatives in families with familial CLL (i.e., a pedigree with at least two first-degree relatives with CLL)⁵¹. This suggests that MBL is not merely a physiological event occurring in most individuals with increasing age, but environmental or genetic predisposing factors exist for MBL. Indeed, several susceptibility loci for CLL were identified and found to be mostly mapping to active promoters or enhancers of transcription factors regulating immune response, apoptosis, or Wnt signaling^{52,53}.

3.3 CLL cell of origin

Despite several hypotheses being proposed, no consensus on the cell of origin (COO) of CLL has been reached. It has been long believed that CLL arises from a mature B cell. However, in 2011, Kikushige et al.⁵⁴ presented that HSC from CLL patients expanded to mono- or oligo-clonal CLL-like cells after xenogeneic transplantation into mice. Intriguingly, these B cell clones did not harbor genomic aberrations found in the original disease. Given the existence of the premalignant condition, i.e., MBL, and its molecular similarity to CLL, CLL leukemogenesis is likely a complex stepwise process starting from HSC (Figure 3). This theory is supported by findings generated in our group that confirmed the existence of independent oligoclonal B cell clones even in immunophenotypically

monoclonal CLL^{55,56}. The existence of mutated and unmutated IGHV CLL subgroups suggests that the final transformation event can occur at a different stage of the maturation process. Moreover, these subgroups are transcriptomically similar compared to a relatively large difference between normal B cells and CLL cells providing further evidence for a single population of origin. However, the question of CLL's normal B cell counterpart(s) remains unanswered.

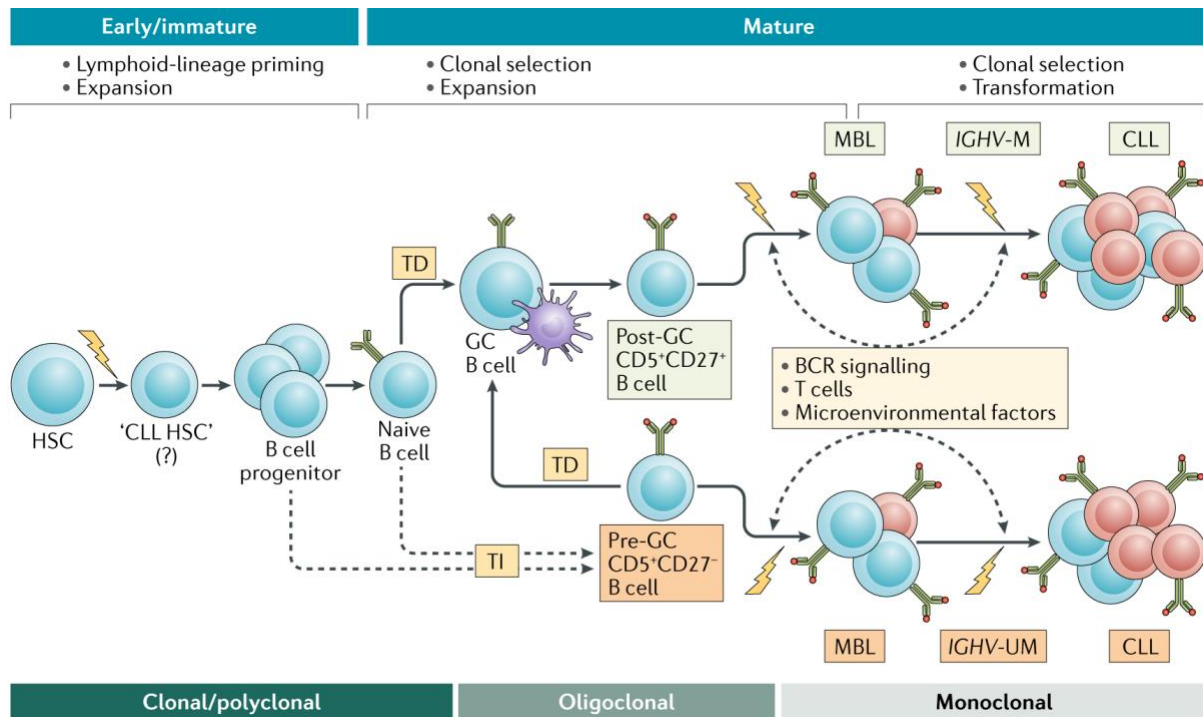


Figure 3: Development of CLL (adopted from Bosch et al.⁵⁷). Lightning symbols indicate genetic or epigenetic lesions leading to CLL. TD, T cell-dependent antigen; TI, T cell-independent antigen.

In the last decade, several publications described immunophenotypically distinct but partly overlapping B cell populations in some parts resembling CLL cells. The first studies demonstrated that CLL is most similar to memory-resembling B cells^{58,59}. Further studies detected features of CLL in B1 B cells found circulating in PB, transitional B cells, MZB, or in CD5⁺CD27⁺ B cells⁶⁰. However, none of the studies brought conclusive evidence. Oakes et al.⁶¹ suggested a hypothesis that CLL cells, based on their epigenetic profile, originate from a continuum of maturation stages between early memory and mature memory B cells. Intriguingly, CLL cells harbor autoreactive, polyreactive, and stereotyped BCRs. These data support the hypothesis suggesting the origin of CLL from autoreactive B cells that evaded tolerance checkpoints, which is in line with BCR anergy in CLL cells⁶².

It has been challenging to tackle these gaps in knowledge with classical cell immunobiology techniques profiling only a few molecular markers in individual cells. By

contrast, recent advances in high-throughput single-cell methods, enabling simultaneous measurement of the entire transcriptome across thousands of cells, have already revealed unappreciated heterogeneity within healthy B cell populations (see above). In this thesis, I will present the results of our ongoing efforts to leverage single-cell data to interrogate our hypothesis on the cellular origin of CLL.

3.4 Expression of ROR1 as a stable marker of CLL

Apart from gene expression of pan T cell antigen CD5 in CLL cells, upregulation of ROR1 (transmembrane receptor tyrosine kinase-like orphan receptor 1) was established as one of the most stable CLL markers^{63,64,65}. ROR1 is a membrane receptor, originally described as a pseudokinase involved in the non-canonical branch of the Wnt signaling pathway⁶⁶. Its activity in CLL cells may be regulated by post-translational modifications⁶⁷ and the presence of its ligands⁶⁸. A key role of ROR1, which is likely involved both in the development of normal B cells and malignant transformation to CLL, is the regulation of pre-BCR and BCR signaling^{69,70}. It was shown that ROR1 expression spikes sharply at the proliferative pre-BII large B cell stage^{69,71} and then decreases following maturation to naive B cells, which are generally considered as ROR1 negative⁶⁹ (Figure 4). The ROR1-specific expression on CLL cells makes it a promising candidate for targeted treatment with monoclonal antibodies⁷² and for monitoring CLL remission⁶⁵.

In our laboratory, we identified ROR1+ B cells in the peripheral blood of healthy adults. This finding sparked further efforts to characterize this rare population using scRNA-seq and advanced computational approaches. The results are discussed in section 5.6 of this thesis.

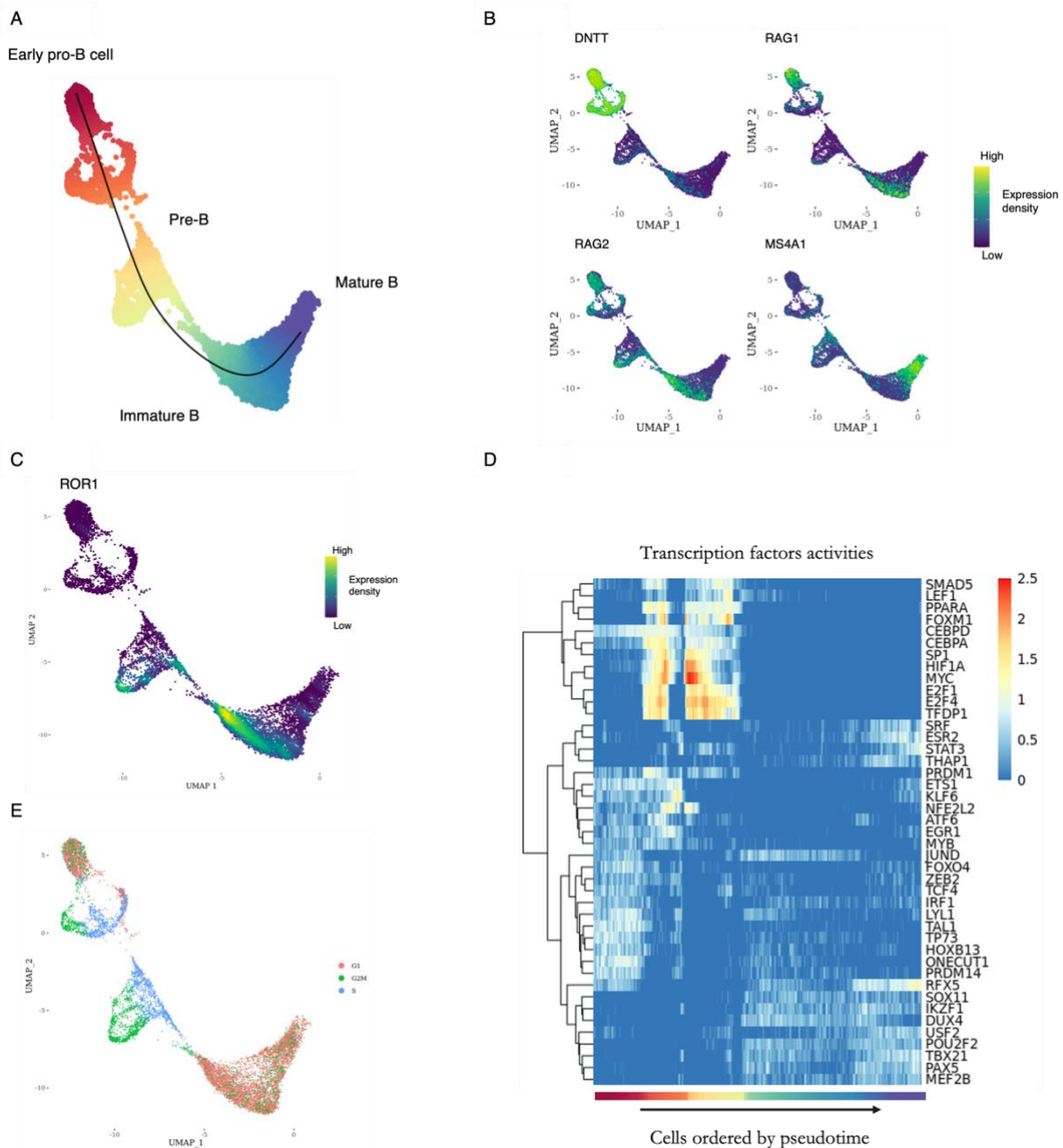


Figure 4: ROR1 expression during B cell development in the bone marrow. B cells were extracted from the bone marrow dataset from the Human Cell Atlas⁷³. A) Inferred developmental trajectory starting from early pro-B cells and continuing through pre-B and immature B cells to mature B cells. The rainbow color coding represents pseudotime values. B) Visualization of gene signatures in B cell development. C) The gene expression density distribution of ROR1 corresponds to the previous observations that ROR1 expression spikes sharply at the proliferative pre-BII large B cell stage^{69,71} and then decreases following maturation to mature B cells. D) Predicted activities of the most dynamic transcription factors. E) Cell cycle phase classification. The cell cycle phase was assigned using the *CellCycleScoring* function from the Seurat R package⁷⁴.

4 Genomic mutation data analysis

With the advent of next-generation sequencing, the mutation landscape of cancer was well characterized^{75,76,77}. However, our understanding of the genomic data, and therefore molecular biology of cancer, is far from complete. Most current clinical guidelines are based on profiling single-gene mutations (e.g., *TP53* gene mutation status in CLL²⁸). Nevertheless, mutations might have a distinct prognostic or predictive impact due to differential mutation background. Therefore, stratifying patients based on the entire mutation profile could enhance the accuracy of prognostication.

In contrast to subtyping based on mRNA^{78,79} and microRNA expression⁷⁹, or DNA methylation^{79,80}, not many studies have grouped patients based on somatic mutation data^{81,82}. The likely explanation lies in the fact that mutation data are extremely sparse and heterogeneous. The methods developed to alleviate these challenges will be discussed in the next section.

Among hematological malignancies, studies exploring mutational patterns have been carried out and led to the molecular classification of acute myeloid leukemia (AML)⁸³ and diffuse large B cell lymphoma⁸⁴. Papaemmanuil et al.⁸³ profiled mutations in 111 genes in 1540 patients and applied Bayesian Dirichlet processes to set classification rules segregating patients into subtypes displaying different clinical outcomes. However, in many other hematological malignancies, such as CLL, similar studies were until recently missing⁴³ (commented publication in section 4.2 of this thesis), although their genetic and clinical heterogeneity constitutes a prerequisite for further investigation.

4.1 Genomic mutation subtype identification

With the plummeting cost of sequencing, mutation data generated with mostly unbiased profiling of protein coding regions (i.e., whole exome sequencing, WES) or even a whole genome (i.e., whole genome sequencing, WGS) become available for large patient cohorts. In this thesis, we focused on extracting biological insights from WES data. A typical workflow of WES analysis includes raw data quality control, preprocessing, mapping, post-alignment processing, variant calling, annotation, and filtration⁸⁵. To predict the impact of the variants, computational tools such as the Ensembl Variant Effect Predictor⁸⁶ or Combined Annotation Dependent Depletion score might be used⁸⁷.

When searching for mutation subtypes, it was demonstrated that desparsification of the data using prior biological knowledge greatly enhanced the analysis. One way to reduce the dimensionality and sparsity of data is to map mutations to their respective biological processes. Kuijjer et al.⁸² developed a method for calculating pathway mutation score, which considers all genes in a pathway and quantifies the level of disruption of the corresponding pathway. The authors applied this method to analyze 23 cancer types from the TCGA and identified nine pan-cancer mutation subtypes. Additionally, they searched

for subtypes in each analyzed cancer type and discovered distinct prognostic subtypes in three entities, including AML, the only hematological malignancy analyzed within the study.

Other approaches for identifying mutational subtypes include the application of diffusion process or random walk and prior biological knowledge in the form of protein-protein interaction networks^{88,89,90}. For example, Hofre et al.⁸⁸ applied random walk with a restart to propagate mutations over its network neighborhood. Subsequently, smoothed networks are clustered using non-negative matrix factorization. A drawback of these approaches lies in the assumption that genes interact physically within biological processes, which is not always true.

4.2 Results and discussion

Identification of Clinically Relevant Subgroups of Chronic Lymphocytic Leukemia Through Discovery of Abnormal Molecular Pathways

Petr Taus¹, Sarka Pospisilova^{1,2,3}, Karla Plevova^{1,2,3}

¹Central European Institute of Technology, Masaryk University, Brno, Czechia

²Department of Internal Medicine – Hematology and Oncology, University Hospital Brno, Brno, Czechia

³Faculty of Medicine, Masaryk University, Brno, Czechia

Frontiers in Genetics, <https://doi.org/10.3389/fgene.2021.627964>

COMMENTARY

The commented article⁴³ aimed at identifying the prognostic subtypes of CLL based on a somatic mutation profile. The main objective of this project was to learn the analysis and build and test a computational workflow on a large patient cohort. The developed workflow is currently being applied in ongoing projects with in-house datasets (see section List of conference contributions - relevant abstracts denoted with an asterisk).

In the commented study, we leveraged the publicly available WES dataset of 506 CLL patients with shared clinical data³⁶. Similarly to Kuijjer et al.⁸², we mapped the somatic mutations to the molecular pathways and calculated the pathway mutation score. Pathways data from databases such as the Molecular Signature Database⁹¹, including data from various sources like Gene Ontology⁹², KEGG⁹³, and Reactome⁹⁴, are inherently redundant (e.g., genes often participate in multiple pathways and some pathway databases organize pathways hierarchically). In our work, we first applied a set theory algorithm⁹⁵ to decrease the redundancy of the curated pathway signatures from the MSigDB⁹¹. Then we calculated the pathway mutation score and performed ensemble clustering of the patient's samples based on the score. Ensemble clustering represents a combination of multiple clustering solutions through a consensus approach which helps to improve the robustness of the clustering result⁹⁶.

We identified four clusters that we found to differ in prognosis. Subsequently, we built a classification model using the extreme gradient boosting (XGBoost) algorithm⁹⁷. We extracted biological insights from the model to inform the interpretation of the identified clusters and described recurrently mutated pathways associated with subtypes such as DNA-damage response, RNA processing, inflammatory pathways, and calcium signaling. We found mutations in *ATM* and *TP53* associated with the respective subtypes but not mutations in other commonly mutated genes such as *SF3B1* and *NOTCH1*.

The drawback of this study is that the clustering was performed on a single patient cohort, as we could not get access to an independent dataset with available clinical data. Being aware of the sole dataset with clinical data, we randomly split the available dataset into the training set (80% of patients) and test set (20% of patients) to prevent overfitting of our classification model. We used only the training set for parameter tuning and feature selection using 5-fold cross-validation. After the model was trained, we evaluated its performance on the unseen test set (the remaining 20% of the dataset) and reported the performance metrics. Moreover, we applied the classification model to the available data gathered by the Dana-Farber Cancer Institute, and we observed similar distributions of our clusters within their dataset. However, this could hardly be considered a proper evaluation without available clinical data, so we decided not to include this result in the commented article.

I contributed to this work by designing the study, performing the analysis, and writing the manuscript.

5 Single-cell RNA sequencing data analysis

The second central part of my dissertation thesis was focused on establishing a computational workflow for the analysis of scRNA-seq data. This effort was a part of a newly created consortium of collaborating laboratories aiming to establish a scRNA-seq technology in Brno. As every single-cell dataset is different and therefore possesses unique challenges, it is necessary to be familiar with multiple methods for the analysis. In the following sections, I will summarize the most frequently used approaches in our projects.

5.1 Single-cell RNA-sequencing data analysis introduction

In 2009, Tang et al.⁹⁸ published a new protocol allowing the application of RNA-sequencing at the single-cell level (scRNA-seq). Since around 2014, when new methods making the technology more accessible emerged⁹⁹, the popularity of scRNA-seq has skyrocketed^{100,101}. Nowadays, there is a plethora of available protocols that can be categorized based on aspects of cell capture (e.g., microtiter-plate-based and microfluidic-droplet-based) and transcript quantification (full-length and tag-based).

Typically, the processing of raw scRNA-seq data generated by sequencing machines includes four main steps: read mapping to a reference, assigning reads to genes, assigning reads to the cellular barcodes, and unique RNA molecules quantification (i.e., unique molecular identifiers deduplication). As scRNA-seq deals with a lower amount of RNA than bulk RNA-seq, more PCR cycles are required to compensate for that. For this reason, the application of unique molecular identifiers that identify the specific RNA molecule¹⁰² has become standard practice.

Several tools are available for processing scRNA-seq data differing in speed, mapping, and quantification accuracy^{103,104,105}. Among others, CellRanger¹⁰³, a proprietary tool from 10x Genomics company, takes care of preprocessing of the most widely used 10x Genomics Chromium scRNA-seq data. CellRanger uses RNA-seq aligner STAR¹⁰⁶ for mapping reads to the reference genome, but the rest of the gene quantification pipeline is conducted with its own algorithms. An academic alternative to the CellRanger that is not locked into 10x Genomics products is the STARsolo tool, which seems to be more computationally efficient while generating similar results as a CellRanger pipeline¹⁰⁵. Other available tools, e.g., Alevin¹⁰⁷ and Kallisto¹⁰⁸, leverage a pseudoalignment-to-transcriptome approach that makes them computationally ultra-efficient, however, their mapping and quantification accuracy is usually lower compared to the tools mentioned above¹⁰⁵

As outlined in the scheme of a classical scRNA-seq analysis workflow (Figure 5), proper quality control and data normalization must be performed before downstream analyses. Cell quality control is usually based on the total number of molecules and number of unique genes detected within a cell and the percentage of transcripts mapping to mitochondrial genes. Cells with outlier peaks in these metrics' distribution are filtered out. In general, cells with a high fraction of mitochondrial genes correspond to dying cells as RNA enclosed in the mitochondria will retain in the cells while cytoplasmic RNA leaks through a broken membrane. Importantly, the threshold for this metric is cell type and state specific¹⁰⁹. Cells with a high number of unique genes might represent doublets. However, a more elegant and accurate approach for doublet detection is to apply one of the specifically developed doublet detection tools^{110,111,112}. For example, the DoubletDecon¹¹¹ uses a deconvolution method originally developed to infer cellular heterogeneity in bulk RNA-seq data. The DoubletFinder¹¹⁰ compares cell expression profiles to artificially created doublets to predict real doublets.

Once the count matrix is cleaned, data must be normalized to account for variable count depths. Among other strategies, most commonly, data are scaled by the sum of the read counts, multiplied by a scale factor (e.g., 10,000), and transformed with a natural logarithm. Alternatively, a regularized negative binomial regression can be leveraged for normalization where Pearson residuals are used in the subsequent downstream analyses¹¹³. After normalization, feature selection usually follows. For instance, an often-used approach is to select the most variable genes based on the variation in their expression across the cell populations. The aim is to keep informative genes while removing those only affected by technical noise or biological variation stemming from transcriptional bursting. Next, linear

transformation, such as PCA, is conducted to reduce the dimensionality of the data. Subsequently, principal components are fed into clustering algorithms¹¹⁴ to identify groups of transcriptomically similar cells and to nonlinear dimensionality reduction methods such as t-SNE¹¹⁵ or UMAP¹¹⁶ for visualization purposes.

The aforementioned steps can be performed using Bioconductor R packages¹¹⁷, Seurat R package⁷⁴, or Scanpy Python package¹¹⁸.

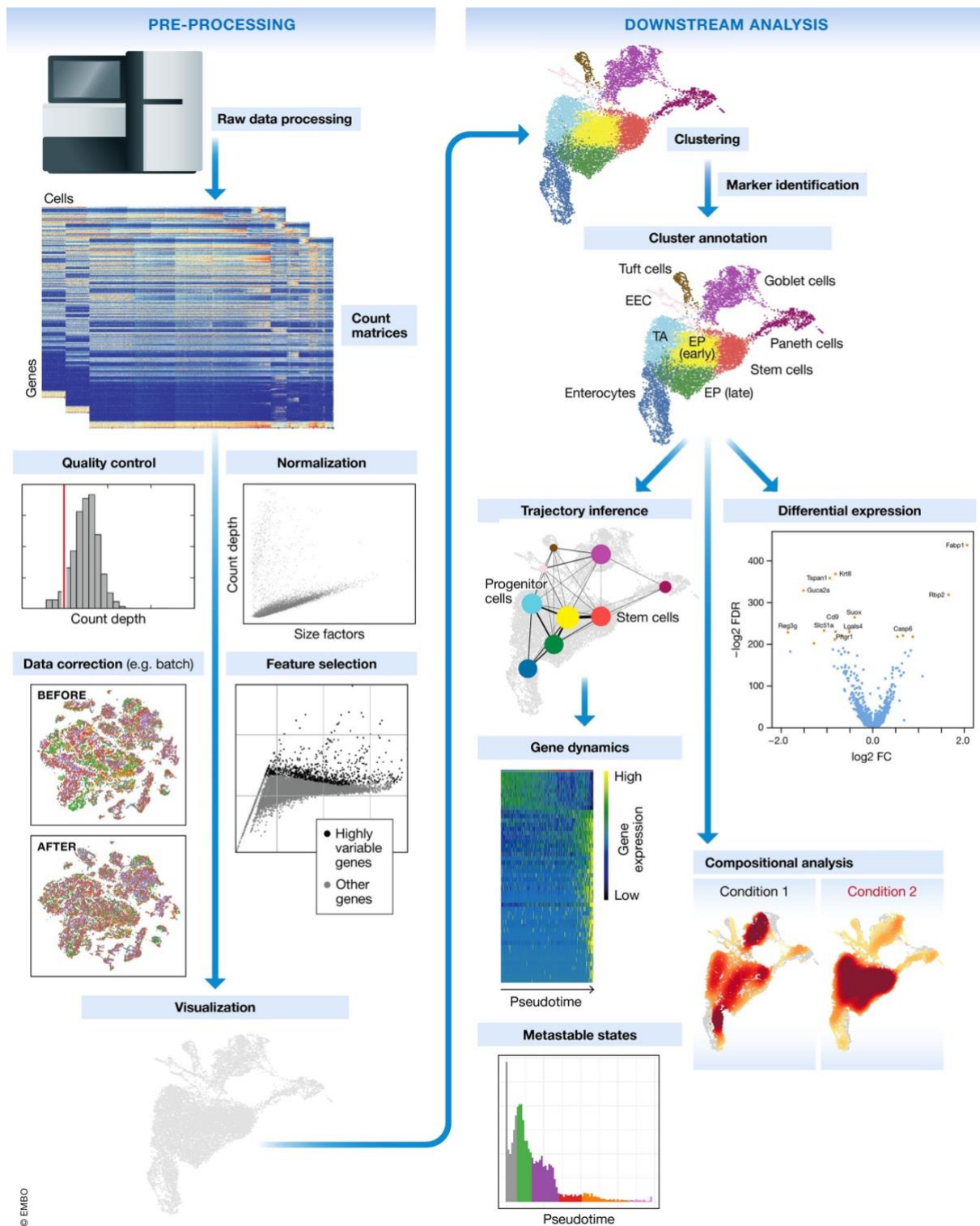


Figure 5: Scheme of a classical scRNA-seq analysis workflow (adopted from Luecken et al.¹¹⁹).

5.2 Data integration

Typically, one 10x Chromium single-cell gene expression library contains between 2000 and 10,000 cells. Datasets can be combined to increase detection power and the chance of detecting rare cell types. However, data generated with diverse technology, across multiple conditions or in different laboratories, will often contain technical or biological heterogeneity, the so-called batch effect, preventing straightforward integration and downstream analysis.

Compared to bulk RNA-seq, data integration scenarios in scRNA-seq data are more diverse and complex. Since the advent of scRNA-seq, a plethora of data integration tools have been developed, however, none of them perform well under all circumstances^{120,121}. The methods usually balance between incomplete batch effect removal and overcorrection, i.e., removing both technical and interesting biological variance. The tools can be categorized based on the representation of integrated cellular profiles as an integrated graph, joint embedding, or corrected feature matrix.

The best performers across multiple metrics are methods called Harmony¹²² (implemented in R) and Scanorama¹²³ (implemented in Python). Harmony constitutes the integrated profile in the form of joint embedding. It iteratively performs fuzzy clustering (meaning a cell can be assigned to multiple clusters) while optimizing for a diversity of batches within each cluster until convergence. Scanorama, inspired by algorithms for stitching panoramic photographs, searches for nearest neighbors between the datasets that are subsequently leveraged to integrate data. The Scanorama approach is a generalization of the mutual nearest neighbors integration method¹²⁴ to more complex scenarios containing multiple datasets.

The approach implemented in Seurat v3¹²⁵ projects datasets into a shared correlation structure across datasets using canonical correlation analysis and identifies pairs of mutual nearest neighbors across datasets. Subsequently, the weighted average of differences between the pairs is leveraged for batch correction. It represents integrated cellular profiles as a corrected feature matrix. The performance of this tool is variable as it tends to favor the removal of batch effects over the conservation of biological variance. However, such property can be desirable in some cases. For example, when integrating biological replicates or when our goal is to purposely remove inter donors' variability.

The last method presented here, the cluster similarity spectrum (CSS)¹²⁶, achieves integration by representing each cell by its transcriptome's similarity to every cell cluster in each sample. The CSS was included in neither of the recent benchmarking efforts^{120,121}, however, it was developed and tested in the context of scRNA-seq data from cerebral organoids. Therefore, we decided to implement it in our study of cerebral organoids derived from Alzheimer's disease patients (manuscript under review summarized in section 6.2 of this thesis).

In summary, there is no one method to rule them all, and it is beneficial to be familiar with diverse approaches that can be tried in the project. We can evaluate their

performance and choose the one with the best batch effect correction efficiency in the particular context. A straightforward way to compare integration results is to compute the Local inverse Simpson's index (LISI)¹²² that can be used to quantify the mixing of cells from different batches in the cell neighborhood.

5.3 Cell type annotation

Classically, cell type annotation is the next step following the identification of clusters of transcriptomically similar cells either in single or integrated datasets (Figure 5). Cell type annotation can be either approached manually using expert knowledge or automatically using reference datasets or gene signatures¹²⁷. Manual annotation is usually done by projecting the expression of canonical marker genes in 2D/3D representations such as UMAP¹¹⁶ or t-SNE¹¹⁵. This process can be time-consuming as detecting distinct subpopulations often requires subclustering of the originally identified clusters using a different set of variable genes, better separating specific subpopulations. Moreover, marker genes are often not sufficiently detected by scRNA-seq, preventing their direct visualization. This can be overcome by leveraging an approach based on gene-weighted kernel density estimation implemented in the Nebulosa R package¹²⁸. Automatic cell type annotation is well suited for major cell types but generally performs poorly in annotating rare cell types. A plethora of methods have been developed that can be grouped into three main categories – marker gene database-based (e.g., SCINA¹²⁹), correlation-based (e.g., SingleR¹³⁰), or supervised classification-based (e.g., SingleCellNet¹³¹).

5.4 Biological activity estimation

Functional interpretation of scRNA-seq data is inherently difficult, and multiple methods have been developed trying to recover functional insights from data using gene signatures. The signatures, sometimes called footprints, might be defined by the effect of molecules (i.e., transcription factors) or biological processes of interest (i.e., signaling pathways)¹³².

Footprint-based approaches rely on a combination of prior knowledge with a statistical method. In our work, we explored the influence of statistical methods on performance and discovered that simple linear approaches and consensus score outperform other methods¹³³. Nevertheless, a choice of prior knowledge seems even more important for activity estimation accuracy¹³⁴. It is encoded in the databases usually as gene signatures belonging to the same biological process, e.g., regulon of transcription factor (TF) in MSigDB⁹¹, ChEA3¹³⁵, or DoRothEA¹³⁶. The latter is a curated, comprehensive resource built upon different types of evidence (literature-curated resources, ChIP-seq peaks, TF binding site motifs, and interactions inferred directly from gene expression). These resources are tissue and sex agnostic, however, in reality, gene regulatory networks (GRN) are likely both tissue-¹³⁷ and sex-specific¹³⁸. The GRN can be accessed from the GRAND database¹³⁹. To the best of our knowledge, there is no benchmarking study

comparing the performance of tissue-specific and tissue-agnostic GRN for estimating TF activities. In theory, tissue-agnostic GRN could outperform tissue-specific GRN in the context of scRNA-seq due to the high dropout rate and other data quirks.

Additionally, GRN can even be cell-type-specific. Such GRN can be directly inferred from the data at hand¹⁴⁰. However, recent reviews revealed the overall poor performance of currently available methods for GRN inference^{141,142}.

Apart from estimating TF activities, it is possible to predict the activities of signaling pathways using their footprints. Pathway responsive genes for activity reference (PROGENy) is a resource for estimating the activity of 14 signaling pathways coupled with a linear model^{143,144}, and we leveraged it in one of our projects (publication summarized in section 6.1 of this thesis).

5.5 Trajectory inference

ScRNA-seq data represent a static snapshot of cellular states at a point in time. Therefore, studying cellular dynamic processes such as differentiation and cell activation poses a challenge. These processes can be modeled computationally using either pseudotime analysis^{145,146,147} or RNA velocity^{148,149}, or even a combination of the two¹⁵⁰. Pseudotime analysis, also called trajectory inference, pseudotemporally orders cells based on similarities of their expression profiles.

Developmental trajectories might be linear or more complex, with one or more bifurcation points or tree-like structures. There are now over 100 tools for ordering cells into lineage (as listed in the repository scRNA-tools.org¹⁵¹). Their optimal performance depends on the properties of the data¹⁵², and some prior knowledge about expected topology or the origin of trajectory is usually required.

The concept of RNA velocity¹⁴⁸, unraveling the speed and direction of the cell movement in transcriptomic space, enabled to study cellular dynamics without prior knowledge about the system. RNA velocity vectors are computed based on the ratio of spliced to unspliced counts. However, the accuracy of RNA velocity is prone to the inherent technical and biological noise of scRNA-seq data. For example, it was shown that the choice of preprocessing pipeline for quantifying spliced/unspliced counts significantly influences the results interpretation¹⁵³. The recently developed tool called CellRank¹⁵⁰ combines RNA velocity and pseudotime analysis to overcome each approach's drawbacks.

5.6 Results and discussion – the cellular origin of CLL

In the following sections, 5.6.1 – 5.6.4, I will describe my main scRNA-seq project focused on exploring the cellular origin of CLL. The initiation of this project was sparked by the identification of a rare ROR1+ B cell population in PB of healthy adults before I joined our research group. Mature ROR1+ B cells in PB may theoretically represent a pool of B cells that are either defective (and escaped regulation mechanisms), autoreactive, or

represent a specific functional B cell subset from which MBL, and subsequently CLL, originate. Using diverse analytical approaches and a combination of our datasets with published ones, we characterized the ROR1+ B cells and generated evidence supporting the hypothesis that this population is the cellular origin of CLL.

5.6.1 Establishing scRNA-seq method and computational pipeline

To interrogate our hypothesis, we screened a relatively large cohort of healthy individuals (N=68) and found a correlation between the prevalence of ROR1+ B cells and a donor's age. These findings prompted us to characterize the ROR1+ B cells in the context of B cell development and CLL transformation using 10x Genomics Chromium scRNA-seq. As a pilot experiment, we prepared two sequencing runs containing samples of FACS-sorted ROR1+ B cells, CLL cells, B cells from PB, and B cells from BM. We used antibody-based hashtag oligos for sample multiplexing¹⁵⁴, which allowed us to keep information about the sample origin for each cell.

As this was the first scRNA-seq experiment conducted at our institute, we had to optimize wet and dry lab protocols. To preprocess raw expression and hashtag oligo data, we used the CellRanger pipeline. For the basic data analysis, we mostly followed the Seurat workflow⁷⁴. We tested log normalization and SCTransform¹¹³ and observed minimal impact on the separation of B cell populations. To identify possible ambient RNA contamination we implemented SoupX¹⁵⁵ and for doublet detection we used DoubletFinder¹¹⁰.

We identified clusters of expected B cell populations including precursor pro- and pre-B cells expressing ROR1 as described in the literature^{69,71}. To integrate the datasets, we tested Seurat v3¹²⁵, Harmony¹²², and Scanorama¹²³. We evaluated integration results with the LISI score¹²² which was the best for results obtained with Harmony.

Subsequently, we inferred B cell developmental lineage structure and pseudotime by fitting a principal curve using the Slingshot package¹⁴⁷. Interestingly, we identified a trajectory leading from immature B cells through a subpopulation of FACS-sorted ROR1+ B cells to CLL cells. To detect genes differentially expressed along the lineage, we used random forest regression to predict pseudotime values from gene expression and leveraged the model's feature (=gene expression) importance to rank gene expression dynamics. When inspecting the most dynamic genes, we noticed that some of them correspond to stress response genes (such as *FOS*, *JUN*, and *JUNB*), suggesting the presence of possible artifacts in our data¹⁵⁶. Moreover, we could not separate memory from naive B cells from peripheral blood, and during the subclustering procedure, we identified clusters mainly differing in the expression of stress response genes. To further investigate this issue, we integrated our data with a published dataset of B cells from lymph node¹⁵⁷ and with two peripheral blood mononuclear cell (PBMC) datasets from the 10x Genomics website. We identified differentially expressed genes between the corresponding B cell populations and performed gene set enrichment analysis that revealed top enriched GO terms linked with

response to temperature stress (e.g., "response to temperature stimulus", "regulation of cellular response to heat", and "response to cold"). Unfortunately, these results strongly supported our hypothesis about the presence of sampling artifacts in our data. We suspected the culprit was the high complexity of our sequencing runs with multiple samples and antibody-based hashtag oligos. The prolonged sample preparation procedure exposed cells to stressful conditions for a long time, resulting in inevitable changes in their expression profiles. We concluded that there is no safe way to remove this artifact from data *in silico* as a stress-caused distortion of cellular expression profile can unpredictably influence the expression of other genes, primarily not associated with a stress response¹⁵⁶.

5.6.2 Characterization of ROR1+ B cells

Based on the abovementioned experience with an overly complex sample, we decided to prepare a simple sequencing run containing FACS-sorted ROR1+ B cells. The simplicity of the sample allowed us to work quickly with the cells and therefore eliminate the activation of a cellular stress response. We analyzed the heterogeneity of ROR1+ B cells and found that they represent a heterogeneous population containing clusters of cells with transitional-like, naive-like, and memory-like gene expression features. However, their proportion was skewed towards transitional B cells (TS) and IgM+ memory B cells.

Next, we were interested if we could find ROR1+ B cells in published B cell datasets and, if so, what their distribution among B cell populations is. We gathered data from Stewart et al.¹⁵⁸, Sutton et al.¹⁵⁹, and the 10x Genomics website, which were of the highest quality from the available datasets. We reanalyzed and unified their annotation and searched for ROR1+ B cells. In line with our previous observation, we found ROR1+ B cells primarily within the neighborhood of TS and IgM+ memory B cells denoted here as alternative memory B cells (AMB) (Figure 6). These findings suggest that the expression of ROR1 is not randomly distributed between B cell populations and is not solely limited to one specific B cell subtype. We think that the expression of ROR1 in B cells could denote a distinct cellular lineage or cell state, such as an anergic cell state.

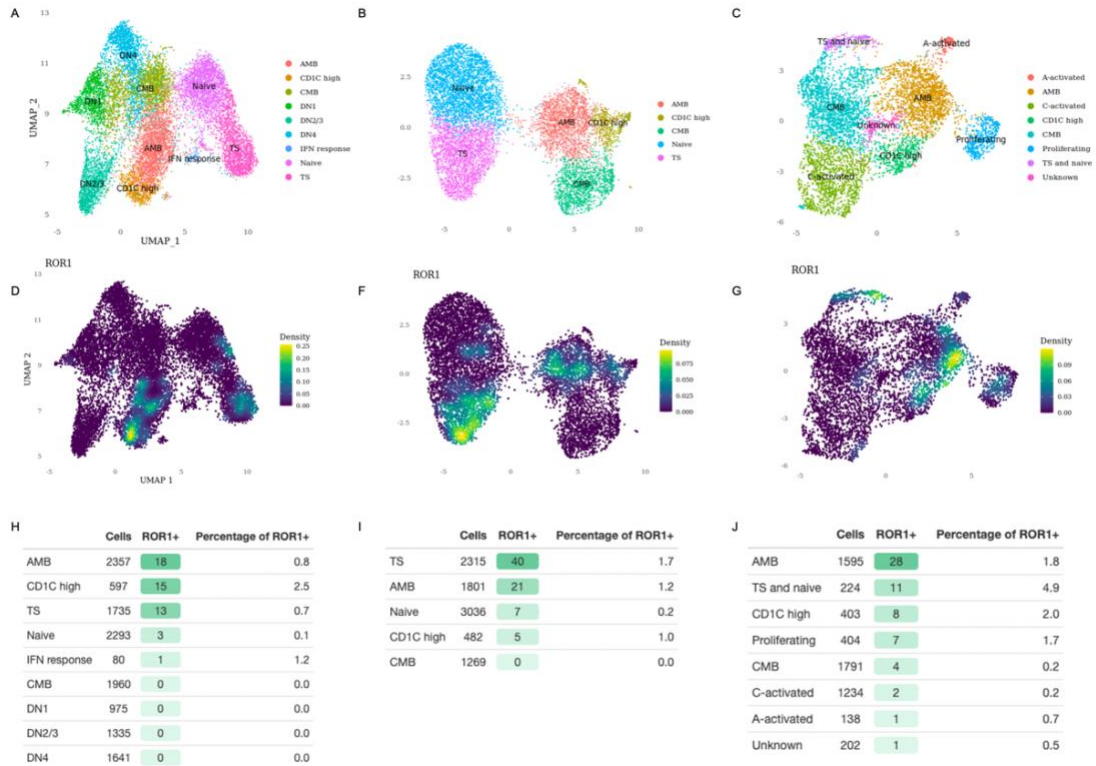


Figure 6: ROR1+ B cells in PB. A,D,H) Dataset of five immunophenotypically sorted B cell subsets (Transitional, Naive, IgM Memory, Classical Memory, and Double Negative) from Stewart et al.¹⁵⁸ B, F, I) Dataset of total B cells from 10x Genomics website. C, G, J) Dataset of atypical memory' B cells (CD20+ CD19+ IgD-) from Sutton et al.¹⁵⁹.

5.6.3 Similarity prediction using a machine learning approach

Further, we aimed to compare the transcriptomic resemblance of healthy B cell populations to CLL cells using the XGBoost-based similarity prediction model. The XGBoost algorithm is a machine learning approach that combines a large number of weak learners (i.e., slightly better than random guessing) based on decision trees into a single strong learner (i.e., a classifier)⁹⁷. The classifier can then be applied to a single sample to calculate a class probability that reflects its similarity to a given class. XGBoost algorithm is well suited for datasets containing correlated features, multiple classes, many samples, and many features which are characteristic properties of scRNA-seq data. In this study, samples represent expression profiles of single cells, features represent genes, and classes represent different B cell populations. For the analysis, we used the xgboost R package version 1.3.2.1.

To train the classifier, we split data randomly into a training set (80% of cells) and a hold-out test set (20% of cells). To limit noise from random fluctuations in the expression of uninformative genes, we used only 3000 of the most variable genes as input features for our classifier. Then we trained the XGBoost model with default hyperparameters to rank

the feature importance. Feature importance was ranked by its information gain, which corresponds to the relative contribution of the feature to a prediction. The training was stopped after 100 rounds without improvement of the multiclass logarithmic loss function (mlogloss), which was evaluated using 5-fold cross-validation. XGBoost, like any machine learning method, possesses several hyperparameters that need to be optimized for the best performance. A plethora of approaches exist to tune the hyperparameters. In our case, we performed a random grid search with 5-fold cross-validation to assess the quality of a specific parameter setting using the `mlr` R package version 2.19.0. Subsequently, the classifier's performance was evaluated on the hold-out test set using mlogloss, multiclass auROC, and multiclass aucPR.

We trained a model on the abovementioned dataset of B cells downloaded from the 10x Genomics website that was of the highest quality from the gathered datasets. Altogether, we built three distinct classifiers to predict the similarity of single CLL cells to different B cell populations based on their expression profiles. The first classifier was built to predict the similarity of CLL cells to the five B cell populations (Figure 7). To assess the model's generalizability to cells from independent datasets, we applied the classifier to B cells extracted from the PBMC dataset obtained from the 10x Genomics website (Figure 7D). Then we applied it to CLL cells and observed that majority of cells were predicted to be the most similar to AMB (55%), naive B cells (29%), and CMB (16%). However, the model was less confident about classifying CMB than AMB or naive B cells (Figure 7F). The CLL sample contained cells with and without specific chromosomal aberration significantly differing in their expression profile, creating two distinct clusters (Figure 7G). Interestingly, we did not observe changes in prediction distribution between these two subpopulations suggesting that prediction is robust to random changes in expression profiles.

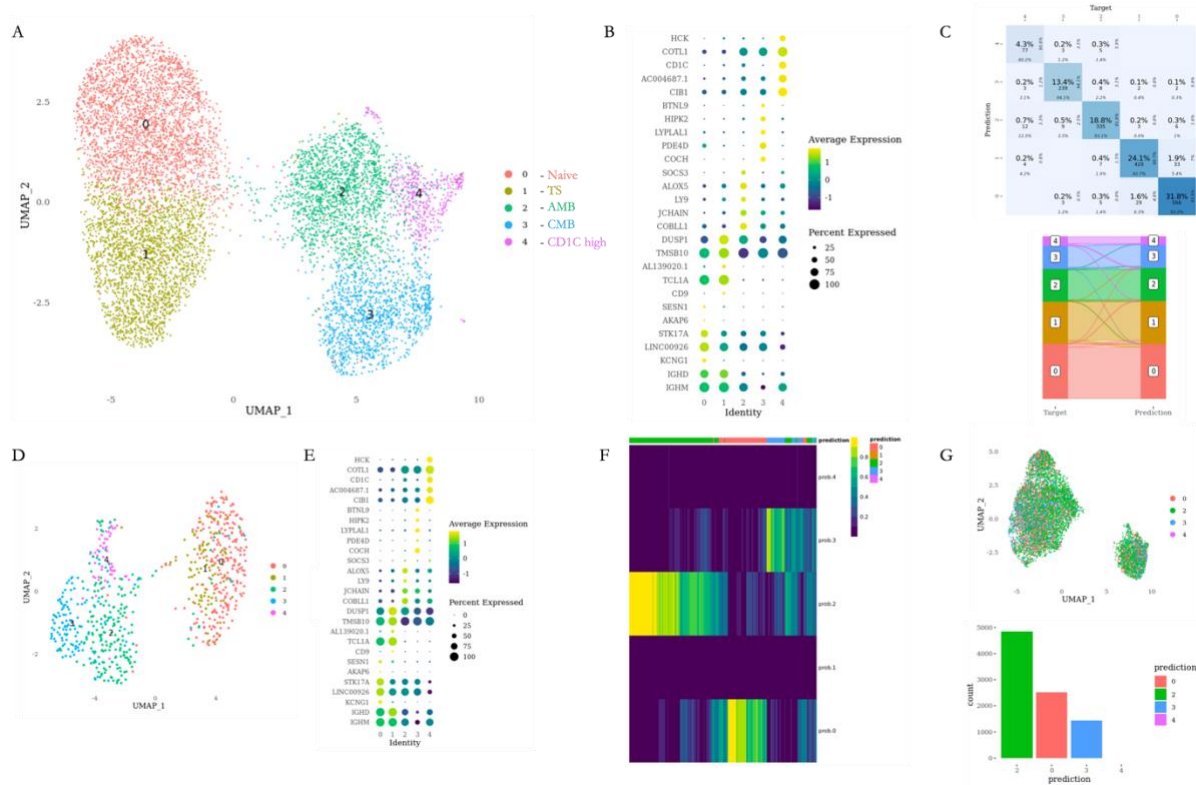


Figure 7: Similarity prediction of single CLL cells to healthy B cell populations. A) Dataset of all B cells from the PB used to train the classification model. The cells were annotated based on the comparison of top differentially expressed genes (B) with canonical gene markers. C) Model performance on the hold-out test set (20% of cells). D) Classification of the B cells extracted from the PBMC dataset obtained from the 10x Genomics website. Cells are colored by the results of the prediction. E) Marker genes expression between the predicted cells. F) Heatmap of classification probability estimates from the XGboost model for (G) similarity prediction of the CLL cells containing subpopulation with the specific chromosomal aberration.

Next, we decided to subcluster naive and memory B cells. We built two classifiers to predict the similarity of CLL cells predicted by the first classifier as naive or AMB/CMB to the identified subclusters of naive and memory B cells, respectively (Figure 8). We observed that in the case of the naive classifier, the majority of CLL cells were predicted to be the most similar to cells from the subcluster containing ROR1+ B cells (Figure 8A). Regarding the memory classifier, most cells were predicted to be the most like cells from the AMB cluster containing ROR1+ B cells (Figure 8B).

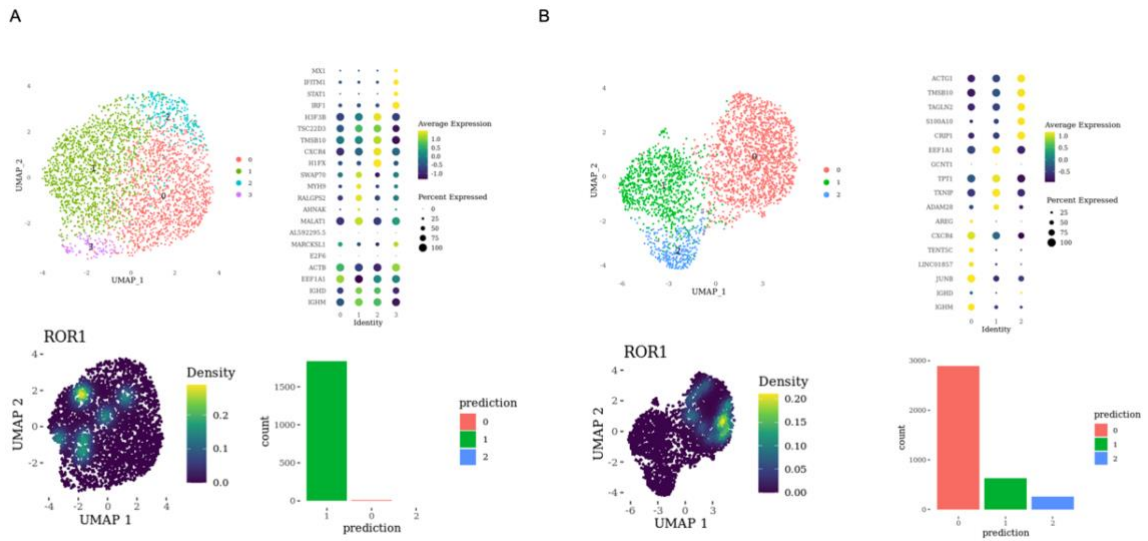


Figure 8: Similarity prediction of single CLL cells to healthy B cell subpopulations. A) The classification model for subclusters of naive B cells predicted virtually all naive-like CLL cells to be the most similar to subcluster of naive cells that share some gene expression markers with AMB (e.g., SWAP70, COBLL1, and RALGPS2) and contains ROR1+ B cells. B) The classification model for subclusters of AMB/CMB B cells predicted virtually all memory-like CLL cells to be the most like AMB that contains ROR1+ B cells.

In the subsequent phase of the project, we generated a scRNA-seq dataset of B cells from the PB of two donors of 48 and 63 years old. We integrated this dataset with the abovementioned FACS-sorted ROR1+ B cells and B cells from the 10x Genomics using the Seurat v3 approach and performed clustering analysis and annotation. We found clusters of TS, naive, AMB, CMB, and CD1C high cells. Additionally, we identified two clusters that contained cells almost exclusively from the dataset of FACS-sorted ROR1+ B cells (Figure 9).

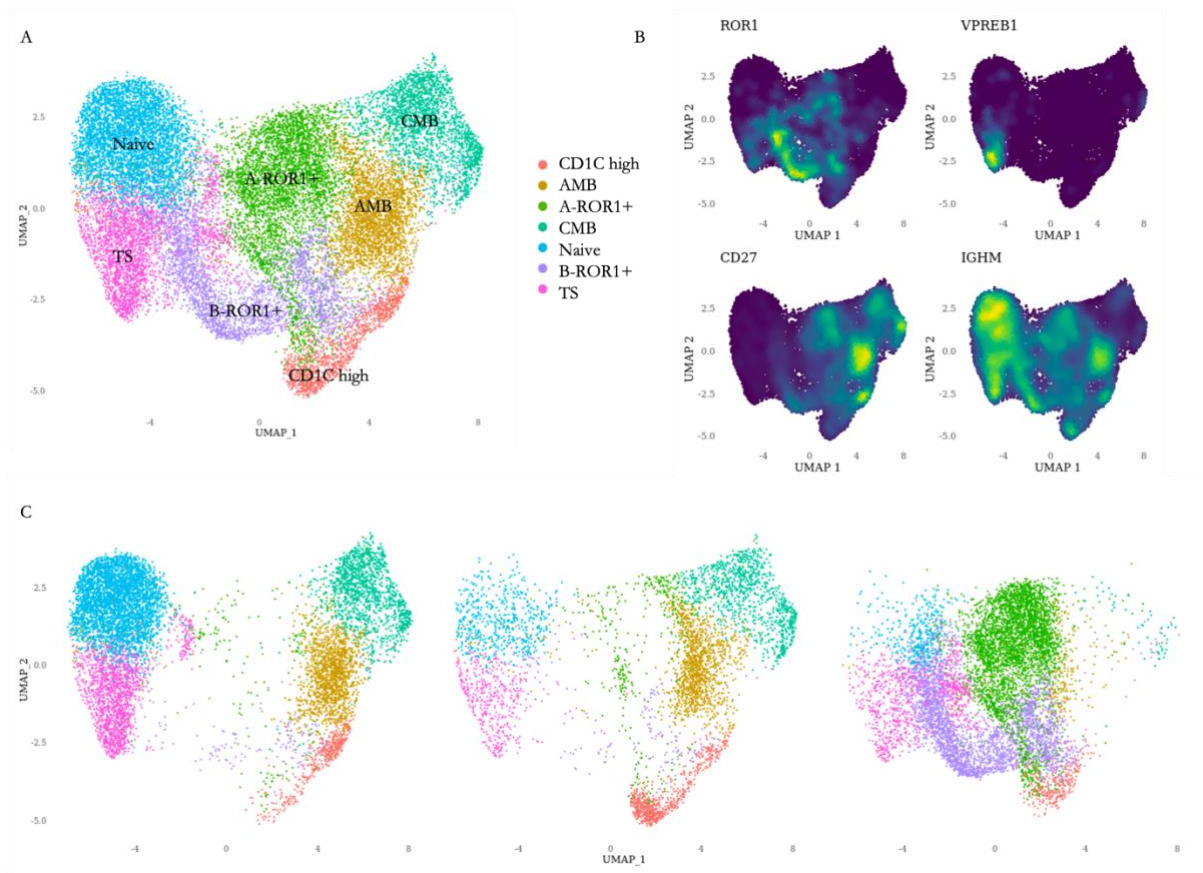


Figure 9: Integrated analysis. A) Integrated dataset of an in-house generated dataset of B cells from the PB of two donors, FACS-sorted ROR1+ B cells, and dataset of B cells from the 10x Genomics website. B) Visualization of selected marker genes. C) Integrated dataset split by dataset identity. Left, B cells from the 10x Genomics website; middle, an in-house generated dataset of B cells; right, a dataset of FACS-sorted ROR1+ B cells.

Then, again, we asked which of these clusters is the most similar to CLL cells. To answer this, we built an XGBoost classifier using gene expression profiles. Additionally, we trained a classifier with activities of TFs, estimated with the Dorothea R package, as input features. We assumed that we could extract further insights from the model about the transcriptional regulation of ROR1+ B cells. The expression-based model predicted the majority of CLL cells to be the most similar to cluster A of ROR1+ B cells. This subpopulation of ROR1+ B cells is the most related to IGHM+ memory B cells expressing CD27 and IGHM, while the subpopulation denoted as B is more naive-like but intriguingly expresses a low level of CD27, a canonical marker of the memory B cells. In the case of the TF activities-based model, most cells were also predicted to be the most like A-ROR1+ B cells, but the difference between the second and third most abundant predictions was small (Figure 10). One possible explanation of this result is that ROR1+ B cells share a core transcriptional regulatory network similar to that of TS that are naturally enriched for ROR1+ B cells.

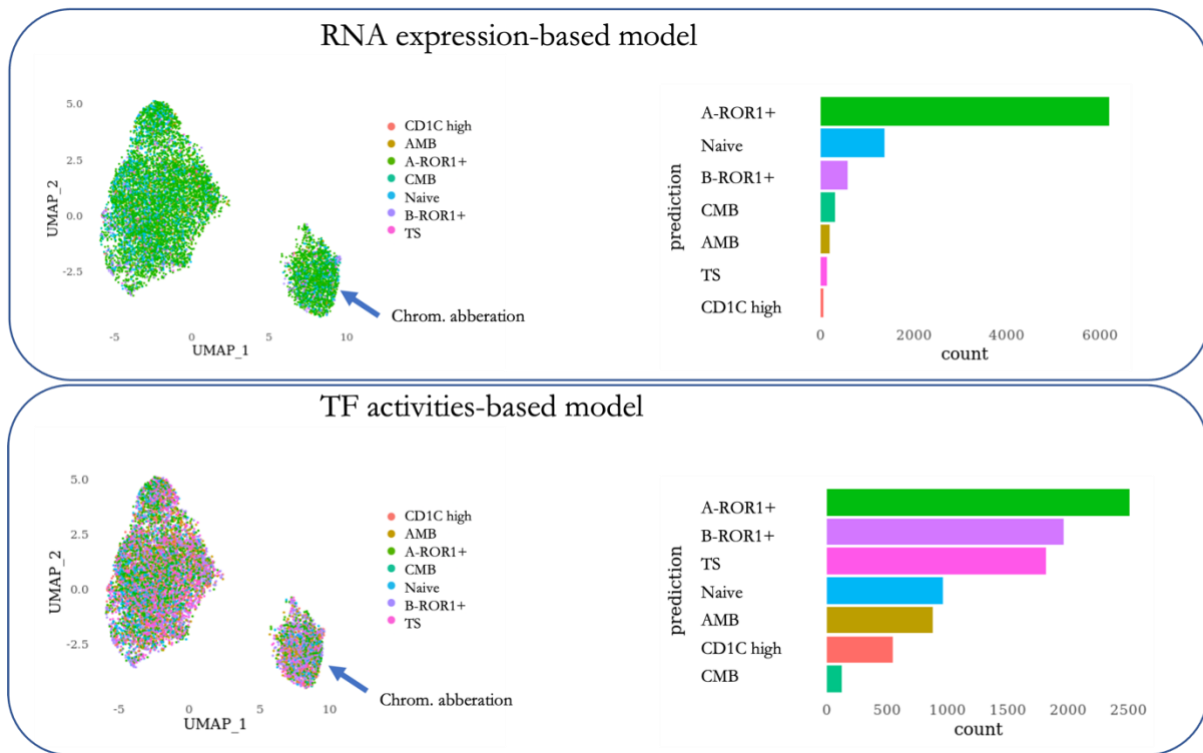


Figure 10: Similarity prediction of single CLL cells to normal and FACS-sorted ROR1+ B cell subpopulations.

We proceeded with the extraction of biological insights from the TF activities-based model. We explored the most informative features for the prediction and discovered several differentially activated TFs associated with ROR1+ clusters (Figure 11). For example, the most important feature for the model was FOXP1, whose activity was increased in ROR1+ clusters. FOXP1 was presented in the literature as a repressor of proapoptotic genes¹⁶⁰, suggesting that the ROR1+ B cells could be more resistant to apoptosis. Another was NFKB2, whose activity was decreased in ROR1+ clusters. Interestingly, NFKB2 was recently described to be downregulated in self-reactive anergic B cells in mice¹⁶¹. Furthermore, we observed decreased activities of PAX5 and RFX5 in ROR1+ clusters that we saw declined during the phase of B cell development in BM when ROR1 is transiently expressed (Figure 4D). PAX5 is a master regulator of B cell differentiation^{162,163} and RFX5 is a crucial regulator of MHCII gene expression¹⁶⁴. Overall, these findings suggest that ROR1+ B cells could be B cells in the anergic state which corresponds to one of the main characteristics of CLL cells⁶².

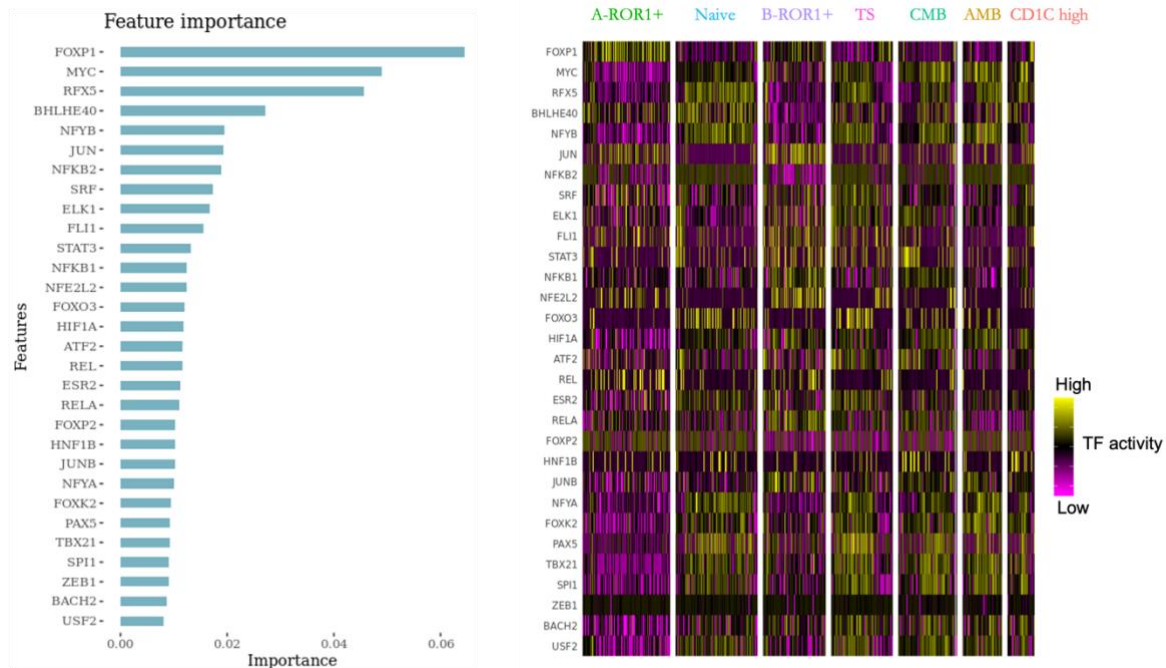


Figure 11: The most important features for the TF activities-based model. On the left is visualized TFs importance for the classification model and on the right is a heatmap of the TFs activities across the B cell subpopulations.

5.6.4 Summary

Here I presented the most exciting results of our ongoing effort to characterize the cellular origin of CLL. We established a computational workflow for the analysis of scRNA-seq data and demonstrated its multiple functionalities and ability to detect sampling artifacts. Using the workflow, we discovered that ROR1+ B cells are a heterogeneous population prevalently consisting of transitional-like and IGHM+ memory-like B cells. We further explored the published B cell datasets. We detected a small number of ROR1+ B cells, primarily among transitional B cells that naturally contain a high percentage of self-reactive B cells⁸ and among IGHM+ memory B cells. Next, we built a pipeline to identify the most similar population to malignant cells and found ROR1+ B cells to be the most similar to CLL cells. Moreover, we identified differentially activated TFs in ROR1+ B cells involved in the regulation of energy and apoptosis. In summary, our findings suggest that gene expression of ROR1 marks self-reactive anergic B cells that transcriptomically resemble CLL cells.

Of note, I must emphasize that all the analyses presented in this section are exploratory, performed on small sample size, and do not provide conclusive evidence. However, these analyses were essential for the project's progress and helped us to formulate better questions and design subsequent experiments. Currently, we are collecting samples of MBL where we are most interested in the characterization of healthy B cells

and comparison of MBL to ROR1+ B cells. Then we would like to perform bulk RNA-seq of sorted ROR1+ B cells from healthy adults of different ages. Moreover, in parallel to the presented effort, we generated cytometry by time of flight (CyTOF) data that supported our hypothesis.

I contributed to this project by establishing scRNA-seq workflow, designing and performing all the presented analyses, and putting the results into a biological context.

6 Other articles related to the thesis

In the following sections, 6.1 – 6.3, I will provide a short commentary on three published articles and one manuscript under review that I co-authored. Next, I will briefly summarize unpublished work related to the thesis.

6.1 Commentary on published articles

decoupleR: Ensemble of computational methods to infer biological activities from omics data

Pau Badia-i-Mompel¹, Jesús Vélez¹, Jana Braunger¹, Celina Geiss¹, Daniel Dimitrov¹, Sophia Müller-Dott¹, **Petr Taus**², Aurelien Dugourd¹, Christian H. Holland¹, Ricardo O. Ramirez Flores¹, Julio Saez-Rodriguez¹

¹Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Bioquant, Heidelberg, Germany

²Central European Institute of Technology, Masaryk University, Brno, Czechia

Bioinformatics Advances, <https://doi.org/10.1093/bioadv/vbac016>

COMMENTARY

The commented article describes the decoupleR, an R and Python package that collects distinct computational methods to estimate biological activities. In this publication, we applied the decoupleR to benchmark the performance of the methods by recovering perturbed regulators. We showed that simpler, linear approaches and the consensus score across top methods overcome other methods at predicting perturbed regulators.

Currently, the DecoupleR contains 11 methods, including AUCell¹⁴⁰, VIPER¹⁶⁵, SCIRA¹⁶⁶, and others. The code of the decoupleR is designed to enable the straightforward addition of other methods in the future. DecoupleR can be used with omics datasets such as phospho-proteomics or transcriptomics. It requires two inputs: One input is a matrix of any molecular readouts, either for single samples or from population comparisons, like normalized gene expression or log fold change. The second is a prior knowledge resource that encodes relations between target features and source biological entities (e.g., network).

The latter can be easily accessed from the meta-database OmniPath¹⁶⁷ using wrappers provided within the decoupleR.

For benchmarking, we developed the decoupleRBench R package that is built on the decoupleR. It allows to evaluate the performance of the statistical methods for the extraction of biological signatures using perturbation experiments.

I contributed to this project during my research stay in Julio Saez-Rodriguez's lab, mainly by implementing and testing one established (AUCell) and two novel methods within the framework. The novel methods were Univariate Decision Tree (UDT) and Multivariate Decision Tree (MDT). UDT is based on a decision tree algorithm that is fitted for each regulator, where associated weights of a given regulator are used to estimate the molecular readouts of the features in a sample. Consequently, the feature importance obtained from the model represents the prediction of the regulator activity. MDT is an ensemble of decision trees, also known as random forest, that, contrary to UDT, is fitted to all regulators of a given network to predict the molecular readouts of the features in a sample. Same as for UDT, the extracted features importance are the regulator activities.

Distinct p53 phosphorylation patterns in chronic lymphocytic leukemia patients are reflected in circumjacent pathways' activation upon DNA damage

Veronika Mancikova^{1,2}, Michaela Pesova^{1,2}, Sarka Pavlova^{1,2}, Robert Helma^{1,2}, Kristyna Zavacka^{1,2}, Vaclav Hejret¹, **Petr Taus**¹, Jakub Hynst¹, Karla Plevova^{1,2,3}, Jitka Malcikova^{1,2}, Sarka Pospisilova^{1,2,3}

¹Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic

²Department of Internal Medicine - Hematology and Oncology, Faculty of Medicine, Masaryk University and University Hospital Brno, Czech Republic

³Institute of Medical Genetics and Genomics, Faculty of Medicine, Masaryk University and University Hospital Brno, Czech Republic

Molecular Oncology, <https://doi.org/10.1002/1878-0261.13337>

COMMENTARY

The commented article focused on the effect of p53 protein phosphorylation on its function as a possible inactivation mechanism. We demonstrated that doxorubicin treatment of CLL tumor-derived cells leads to two distinct phosphorylation patterns. Samples from the group with less phosphorylated p53 had a lower capability to activate p53 target genes and were transcriptomically similar to *TP53*-mutated samples.

We performed bulk RNA-seq for samples before and after treatment. PROGENY analysis revealed that the activity of hypoxia signaling is associated with the type of

disruption of p53. The highest level was observed in *TP53*-mutated samples, followed by samples with less phosphorylated p53, and the lowest activity was in the samples with heavily phosphorylated p53. These findings were further supported by the same pattern in the activity of TF HIF1A, which is the main regulator of hypoxia. I contributed to the project by suggesting and implementing PROGENy and TF activity analyses.

Single-cell RNA sequencing analysis of T helper cell differentiation and heterogeneity

Radim Jaroušek^{1,2}, Antónia Mikulová^{1,2}, Petra Daďová^{1,2}, **Petr Tauš**³, Terézia Kurucová^{2,3}, Karla Plevová^{3,4}, Boris Tichý³, Lukáš Kubala^{1,2}

¹Institute of Biophysics, Czech Academy of Sciences, Brno, Czech Republic

²Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic

³Central European Institute of Technology, Masaryk University, Brno, Czech Republic

⁴Institute of Medical Genetics and Genomics, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic

BBA Molecular Cell Research, <https://doi.org/10.1016/j.bbamcr.2022.119321>

COMMENTARY

In the commented article, we explored the heterogeneity of Th1, Th2, Th17, and Treg cells utilizing standard *in vitro* cytokine-mediated differentiation of human T cells isolated from human PB by scRNA-seq and identified specific gene signatures for their identification. I contributed to this study by processing raw data, providing scripts for the downstream analysis, and training of scRNA-seq analysis to the first author.

6.2 Summary of a manuscript under review

Cerebral organoids derived from Alzheimer's disease patients with PSEN1/2 mutations have defective tissue patterning and altered development

Tereza Vanova^{1,2}, Jiri Sedmik¹, Jan Raska¹, Katerina Amruz Cerna¹, **Petr Tauš**³, Veronika Pospisilova¹, Marketa Nezvedova⁵, Veronika Fedorova¹, Hana Klimova¹, Michaela Capandova¹, Petra Orviska¹, Petr Fojtik¹, Simona Vochyanova¹, Karla Plevova^{3,4}, Zdenek Spacil⁵, Hana Hribkova¹, Dasa Bohaciakova^{1,2}

¹Department of Histology and Embryology, Faculty of Medicine, Masaryk University, Brno, Czech Republic.

²International Clinical Research Center (ICRC), St. Anne's University Hospital, Brno, Czech Republic.

³Central European Institute of Technology, Masaryk University, Brno, Czech Republic.

⁴Department of Internal Medicine - Hematology and Oncology, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic.

⁵RECETOX, Faculty of Science, Kotlarska 2, Brno, Czech Republic.

Under review in the Cell Reports

COMMENTARY

The commented manuscript describes the development and characterization of the Alzheimer's disease-cerebral organoid (AD-CO) model using induced pluripotent stem cells derived from patients with the familial form of Alzheimer's disease. Apart from other experimental approaches presented in the manuscript, we used scRNA-seq to characterize AD-CO and my responsibility was to extract biological insights from scRNA-seq data.

We generated two scRNA-seq datasets of 60-days old AD-CO and matched healthy control (ND-CO) and processed data as described above in section 5. Each of the samples contained 18 COs to compensate for their heterogeneity. To identify cluster-specific genes, we leveraged the natural language processing concept of term frequency-inverse document frequency using the *quickMarkers* function from the SoupX R package¹⁵⁵. We compared our data with published COs data from Kanton et al.¹⁶⁸ using the SingleR R package¹³⁰. To characterize the cell fate decision process, firstly, we estimated count matrices of unspliced and spliced abundances with the loompy/kallisto counting pipeline^{169,170}. Then we calculated RNA velocity and combined it with pseudotime analysis using the CellRank Python package¹⁵⁰.

We proceeded with data integration by applying the CSS method¹²⁶. Projecting the LISI score to the UMAP embedding, we noticed that some clusters are enriched for cells belonging to AD- or ND-CO, and we tested the significance of the difference by permutation testing using the scProportionTest R package¹⁷¹. Then we inferred the developmental trajectory for the cells of the neuronal lineage using the Slingshot R package and searched for dynamical genes along the trajectory differing between AD- and ND-CO using the tradeSeq R package¹⁷².

Finally, to make all the analyses open and reproducible, we used Rmarkdown and WorkflowR R package¹⁷³ for R scripts and Jupyter Notebook for Python scripts that will be available upon publication at https://petrsh.github.io/AD_CO_scRNAseq. I contributed to this study by processing raw scRNA-seq data and suggesting and performing all the analyses of scRNA-seq data. Moreover, to continuously share the results of the analyses and to make them accessible to collaborators with no coding skills I leveraged the PAGODA2 interface¹⁷⁴.

6.3 Summary of unpublished work

Here I will briefly summarize my contribution to two collaborative projects with manuscripts in preparation.

The first project is being conducted in Jan Křivánek's laboratory of Dental Development and Regeneration at the Department of Histology and Embryology at Masaryk University. It is a continuation of the dental cell type atlas project in which Krivanek et al.¹⁷⁵ revealed and characterized stem cell type in a mouse continuously growing teeth. I contributed to this ongoing effort by analyzing two scRNA-seq datasets of FACS-sorted, lineage-traced stem cells from normal and injured continuously growing teeth. The analyses included clustering, cell type annotation, differential cell type abundance, pseudotime, RNA velocity, and CellRank analyses. For example, using the CellRank toolkit, we identified previously unknown cellular subpopulation dedifferentiating during the injury. The findings obtained from scRNA-seq data are currently being experimentally validated.

The next project is a collaboration with Marcela Buchtova's laboratory of Molecular Morphogenesis at the Department of Experimental Biology at Masaryk University. The project is focused on the characterization of Lgr5+ cells during molar development. I contributed to this project by analyzing four distinct scRNA-seq datasets. The analyses included data integration, clustering, cell type annotation, pseudotime, RNA velocity, and CellRank analyses.

References

1. Elsner, R. A. & Shlomchik, M. J. Germinal Center and Extrafollicular B Cell Responses in Vaccination, Immunity, and Autoimmunity. *Immunity* **53**, 1136–1150 (2020).
2. Shen, P. & Fillatreau, S. Antibody-independent functions of B cells: a focus on cytokines. *Nat Rev Immunol* **15**, 441–451 (2015).
3. Schriek, P. *et al.* Marginal zone B cells acquire dendritic cell functions by trogocytosis. *Science* **375**, eabf7470 (2022).
4. Suo, C. *et al.* Mapping the developing human immune system across organs. *Science* **376**, eabo0510 (2022).
5. Massoni-Badosa, R. *et al.* An Atlas of Cells in the Human Tonsil. 2022.06.24.497299 Preprint at <https://doi.org/10.1101/2022.06.24.497299> (2022).
6. Brioschi, S. *et al.* Heterogeneity of meningeal B cells reveals a lymphopoietic niche at the CNS borders. *Science* **373**, eabf9277 (2021).
7. Nemazee, D. Mechanisms of central tolerance for B cells. *Nat Rev Immunol* **17**, 281–294 (2017).
8. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
9. Cashman, K. S. *et al.* Understanding and measuring human B cell tolerance and its breakdown in autoimmune disease. *Immunol Rev* **292**, 76–89 (2019).
10. Watanabe, A. *et al.* Self-tolerance curtails the B cell repertoire to microbial epitopes. *JCI Insight* **4**, e122551.
11. Maddur, M. S. *et al.* Natural Antibodies: from First-Line Defense Against Pathogens to Perpetual Immune Homeostasis. *Clin Rev Allergy Immunol* **58**, 213–228 (2020).

12. Morgan, D. & Tergaonkar, V. Unraveling B cell trajectories at single cell resolution. *Trends Immunol* S1471-4906(22)00003-5 (2022) doi:10.1016/j.it.2022.01.003.
13. Bagnara, D. *et al.* A Reassessment of IgM Memory Subsets in Humans. *J Immunol* **195**, 3716–3724 (2015).
14. Zhao, Y. *et al.* Spatiotemporal segregation of human marginal zone and memory B cell populations in lymphoid tissue. *Nat Commun* **9**, 3857 (2018).
15. Tull, T. J. *et al.* Human marginal zone B cell development from early T2 progenitors. *J Exp Med* **218**, e20202001 (2021).
16. Nemazee, D. Natural history of MZ B cells. *J Exp Med* **218**, e20202700 (2021).
17. Burnett, D. L., Reed, J. H., Christ, D. & Goodnow, C. C. Clonal redemption and clonal anergy as mechanisms to balance B cell tolerance and immunity. *Immunol Rev* **292**, 61–75 (2019).
18. King, H. W. *et al.* Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. *Sci Immunol* **6**, eabe6291 (2021).
19. Sanz, I. *et al.* Challenges and Opportunities for Consistent Classification of Human B Cell and Plasma Cell Populations. *Front Immunol* **10**, 2458 (2019).
20. Glass, D. R. *et al.* An Integrated Multi-omic Single-Cell Atlas of Human B Cell Identity. *Immunity* **53**, 217-232.e5 (2020).
21. Baumgarth, N. A Hard(y) Look at B-1 Cell Development and Function. *J Immunol* **199**, 3387–3394 (2017).
22. Sutton, H. J. *et al.* Atypical B cells are part of an alternative lineage of B cells that participates in responses to vaccination and infection in humans. *Cell Rep* **34**, 108684 (2021).
23. Stewart, A. *et al.* Single-Cell Transcriptomic Analyses Define Distinct Peripheral B Cell Subsets and Discrete Development Pathways. *Front Immunol* **12**, 602539 (2021).

24. Courey-Ghaouzi, A.-D., Kleberg, L. & Sundling, C. Alternative B Cell Differentiation During Infection and Inflammation. *Front Immunol* **13**, 908034 (2022).
25. Delgado, J., Nadeu, F., Colomer, D. & Campo, E. Chronic lymphocytic leukemia: from molecular pathogenesis to novel therapeutic strategies. *Haematologica* **105**, 2205–2217 (2020).
26. Rai, K. R. *et al.* Clinical staging of chronic lymphocytic leukemia. *Blood* **46**, 219–234 (1975).
27. Binet, J. L. *et al.* A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer* **48**, 198–206 (1981).
28. Eichhorst, B. *et al.* Chronic lymphocytic leukaemia: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* **32**, 23–33 (2021).
29. Damle, R. N. *et al.* Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* **94**, 1840–1847 (1999).
30. Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. & Stevenson, F. K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848–1854 (1999).
31. Agathangelidis, A. *et al.* Higher-order connections between stereotyped subsets: implications for improved patient classification in CLL. *Blood* **137**, 1365–1376 (2021).
32. Döhner, H. *et al.* Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J Med* **343**, 1910–1916 (2000).
33. Leeksma, A. C. *et al.* Genomic arrays identify high-risk chronic lymphocytic leukemia with genomic complexity: a multi-center study. *Haematologica* **106**, 87–97 (2021).
34. Baliakas, P. *et al.* Cytogenetic complexity in chronic lymphocytic leukemia: definitions, associations, and clinical impact. *Blood* **133**, 1205–1216 (2019).

35. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
36. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
37. Nadeu, F. *et al.* Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood* **127**, 2122–2130 (2016).
38. Nadeu, F. *et al.* Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia* **32**, 645–653 (2018).
39. Lazarian, G., Guièze, R. & Wu, C. J. Clinical Implications of Novel Genomic Discoveries in Chronic Lymphocytic Leukemia. *J Clin Oncol* **35**, 984–993 (2017).
40. Hallek, M. Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment. *Am J Hematol* **94**, 1266–1287 (2019).
41. Martínez-Trillos, A. *et al.* Mutations in TLR/MYD88 pathway identify a subset of young chronic lymphocytic leukemia patients with favorable outcome. *Blood* **123**, 3790–3796 (2014).
42. Giménez, N. *et al.* Mutations in the RAS-BRAF-MAPK-ERK pathway define a specific subgroup of patients with adverse clinical features and provide new therapeutic options in chronic lymphocytic leukemia. *Haematologica* **104**, 576–586 (2019).
43. Taus, P., Pospisilova, S. & Plevova, K. Identification of Clinically Relevant Subgroups of Chronic Lymphocytic Leukemia Through Discovery of Abnormal Molecular Pathways. *Front Genet* **12**, 627964 (2021).
44. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* **44**, 1236–1242 (2012).

45. Rendeiro, A. F. *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun* **7**, 11938 (2016).
46. Beekman, R. *et al.* The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat Med* **24**, 868–880 (2018).
47. Queirós, A. C. *et al.* A B cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29**, 598–605 (2015).
48. Wojdacz, T. K. *et al.* Clinical significance of DNA methylation in chronic lymphocytic leukemia patients: results from 3 UK clinical trials. *Blood Adv* **3**, 2474–2481 (2019).
49. Lu, J. *et al.* Multi-omics reveals clinically relevant proliferative drive associated with mTOR-MYC-OXPPOS activity in chronic lymphocytic leukemia. *Nat Cancer* **2**, 853–864 (2021).
50. Rawstron, A. C. *et al.* Monoclonal B cell lymphocytosis and chronic lymphocytic leukemia. *N Engl J Med* **359**, 575–583 (2008).
51. Strati, P. & Shanafelt, T. D. Monoclonal B cell lymphocytosis and early-stage chronic lymphocytic leukemia: diagnosis, natural history, and risk stratification. *Blood* **126**, 454–462 (2015).
52. Berndt, S. I. *et al.* Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nat Commun* **7**, 10933 (2016).
53. Speedy, H. E. *et al.* Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. *Nat Commun* **10**, 3615 (2019).
54. Kikushige, Y. *et al.* Self-renewing hematopoietic stem cell is the primary target in pathogenesis of human chronic lymphocytic leukemia. *Cancer Cell* **20**, 246–259 (2011).

55. Plevova, K. *et al.* Multiple productive immunoglobulin heavy chain gene rearrangements in chronic lymphocytic leukemia are mostly derived from independent clones. *Haematologica* **99**, 329–338 (2014).
56. Brazdilova, K. *et al.* Multiple productive IGH rearrangements denote oligoclonality even in immunophenotypically monoclonal CLL. *Leukemia* **32**, 234–236 (2018).
57. Bosch, F. & Dalla-Favera, R. Chronic lymphocytic leukaemia: from genetics to treatment. *Nat Rev Clin Oncol* **16**, 684–701 (2019).
58. Damle, R. N. *et al.* B cell chronic lymphocytic leukemia cells express a surface membrane phenotype of activated, antigen-experienced B lymphocytes. *Blood* **99**, 4087–4093 (2002).
59. Oppezio, P. *et al.* Do CLL B cells correspond to naive or memory B-lymphocytes? Evidence for an active Ig switch unrelated to phenotype expression and Ig mutational pattern in B-CLL cells. *Leukemia* **16**, 2438–2446 (2002).
60. Darwiche, W., Gubler, B., Marolleau, J.-P. & Ghamlouch, H. Chronic Lymphocytic Leukemia B cell Normal Cellular Counterpart: Clues From a Functional Perspective. *Front Immunol* **9**, 683 (2018).
61. Oakes, C. C. *et al.* DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat Genet* **48**, 253–264 (2016).
62. García-Muñoz, R., Roldan Galiacho, V. & Llorente, L. Immunological aspects in chronic lymphocytic leukemia (CLL) development. *Ann Hematol* **91**, 981–996 (2012).
63. Baskar, S. *et al.* Unique cell surface expression of receptor tyrosine kinase ROR1 in human B cell chronic lymphocytic leukemia. *Clin Cancer Res* **14**, 396–404 (2008).

64. Klein, U. *et al.* Gene Expression Profiling of B Cell Chronic Lymphocytic Leukemia Reveals a Homogeneous Phenotype Related to Memory B Cells. *J Exp Med* **194**, 1625–1638 (2001).
65. Kotašková, J. *et al.* ROR1-based immunomagnetic protocol allows efficient separation of CLL and healthy B cells. *Br J Haematol* **175**, 339–342 (2016).
66. Green, J., Nusse, R. & van Amerongen, R. The role of Ryk and Ror receptor tyrosine kinases in Wnt signal transduction. *Cold Spring Harb Perspect Biol* **6**, a009175 (2014).
67. Kaucká, M. *et al.* Post-translational modifications regulate signalling by Ror1. *Acta Physiol (Oxf)* **203**, 351–362 (2011).
68. Janovska, P. *et al.* Autocrine Signaling by Wnt-5a Deregulates Chemotaxis of Leukemic Cells and Predicts Clinical Outcome in Chronic Lymphocytic Leukemia. *Clin Cancer Res* **22**, 459–469 (2016).
69. Bicocca, V. T. *et al.* Crosstalk between ROR1 and the Pre-B cell receptor promotes survival of t(1;19) acute lymphoblastic leukemia. *Cancer Cell* **22**, 656–667 (2012).
70. Zhang, Q. *et al.* Cutting Edge: ROR1/CD19 Receptor Complex Promotes Growth of Mantle Cell Lymphoma Cells Independently of the B Cell Receptor-BTK Signaling Pathway. *J Immunol* **203**, 2043–2048 (2019).
71. Hudecek, M. *et al.* The B cell tumor-associated antigen ROR1 can be targeted with T cells modified to express a ROR1-specific chimeric antigen receptor. *Blood* **116**, 4532–4541 (2010).
72. Choi, M. Y. *et al.* Phase I Trial: Cirmtuzumab Inhibits ROR1 Signaling and Stemness Signatures in Patients with Chronic Lymphocytic Leukemia. *Cell Stem Cell* **22**, 951–959.e3 (2018).
73. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, e27041 (2017).

74. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
75. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).
76. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* **37**, 367–369 (2019).
77. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
78. Hedegaard, J. *et al.* Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. *Cancer Cell* **30**, 27–42 (2016).
79. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
80. Noshmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010).
81. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696 (2015).
82. Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W. & Quackenbush, J. Cancer subtype identification using somatic mutation data. *Br J Cancer* **118**, 1492–1501 (2018).
83. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209–2221 (2016).
84. Schmitz, R. *et al.* Genetics and Pathogenesis of Diffuse Large B cell Lymphoma. *N Engl J Med* **378**, 1396–1407 (2018).
85. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271–281.e7 (2018).

86. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
87. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886–D894 (2019).
88. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108–1115 (2013).
89. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106–114 (2015).
90. Le Morvan, M., Zinovyev, A. & Vert, J.-P. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Comput Biol* **13**, e1005573 (2017).
91. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
92. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330–D338 (2019).
93. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361 (2017).
94. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res* **50**, D687–D692 (2022).
95. Stoney, R. A., Schwartz, J.-M., Robertson, D. L. & Nenadic, G. Using set theory to reduce redundancy in pathway sets. *BMC Bioinformatics* **19**, 386 (2018).
96. Ronan, T., Qi, Z. & Naegle, K. M. Avoiding common pitfalls when clustering biological data. *Sci. Signal.* **9**, re6–re6 (2016).

97. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016) doi:10.1145/2939672.2939785.
98. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382 (2009).
99. Method of the Year 2013. *Nat Methods* **11**, 1–1 (2014).
100. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13**, 599–604 (2018).
101. Zappia, L. & Theis, F. J. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biol* **22**, 301 (2021).
102. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72–74 (2011).
103. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
104. Petukhov, V. *et al.* dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol* **19**, 78 (2018).
105. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. 2021.05.05.442755 Preprint at <https://doi.org/10.1101/2021.05.05.442755> (2021).
106. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
107. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol* **20**, 65 (2019).
108. Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39**, 813–818 (2021).

109. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).
110. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e4 (2019).
111. DePasquale, E. A. K. *et al.* DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Rep* **29**, 1718-1727.e8 (2019).
112. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291.e9 (2019).
113. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296 (2019).
114. Yu, L., Cao, Y., Yang, J. Y. H. & Yang, P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol* **23**, 49 (2022).
115. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun* **10**, 5416 (2019).
116. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* (2018) doi:10.1038/nbt.4314.
117. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* **17**, 137–145 (2020).
118. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).

119. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746 (2019).
120. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21**, 12 (2020).
121. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022).
122. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).
123. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* **37**, 685–691 (2019).
124. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421–427 (2018).
125. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
126. He, Z., Brazovskaja, A., Ebert, S., Camp, J. G. & Treutlein, B. CSS: cluster similarity spectrum integration of single-cell genomics data. *Genome Biology* **21**, 224 (2020).
127. Clarke, Z. A. *et al.* Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat Protoc* **16**, 2749–2764 (2021).
128. Alquicira-Hernandez, J. & Powell, J. E. Nebulosa recovers single-cell gene expression signals by kernel density estimation. *Bioinformatics* **37**, 2485–2487 (2021).
129. Zhang, Z. *et al.* SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes (Basel)* **10**, 531 (2019).
130. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).

131. Tan, Y. & Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Syst* **9**, 207-213.e2 (2019).
132. Dugourd, A. & Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. *Curr Opin Syst Biol* **15**, 82–90 (2019).
133. Badia-i-Mompel, P. *et al.* decoupleR: Ensemble of computational methods to infer biological activities from omics data. 2021.11.04.467271 Preprint at <https://doi.org/10.1101/2021.11.04.467271> (2021).
134. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol* **21**, 36 (2020).
135. Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res* **47**, W212–W224 (2019).
136. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**, 1363–1375 (2019).
137. Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21**, 1077–1088 (2017).
138. Lopes-Ramos, C. M. *et al.* Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. *Cell Rep* **31**, 107795 (2020).
139. Ben Guebila, M. *et al.* GRAND: a database of gene regulatory network models across human conditions. *Nucleic Acids Res* **50**, D610–D621 (2022).
140. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).
141. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* **17**, 147–154 (2020).

142. Stone, M. *et al.* Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. 2021.06.01.446671 Preprint at <https://doi.org/10.1101/2021.06.01.446671> (2022).
143. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat Commun* **9**, 20 (2018).
144. Holland, C. H., Szalai, B. & Saez-Rodriguez, J. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**, 194431 (2020).
145. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
146. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845–848 (2016).
147. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
148. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
149. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408–1414 (2020).
150. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* **19**, 159–170 (2022).
151. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput Biol* **14**, e1006245 (2018).
152. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019).

153. Sonesson, C., Srivastava, A., Patro, R. & Stadler, M. B. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLoS Comput Biol* **17**, e1008585 (2021).
154. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* **19**, 224 (2018).
155. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* **9**, giaa151 (2020).
156. Massoni-Badosa, R. *et al.* Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol* **21**, 112 (2020).
157. Roider, T. *et al.* Dissecting intratumour heterogeneity of nodal B cell lymphomas at the transcriptional, genetic and drug-response levels. *Nat Cell Biol* **22**, 896–906 (2020).
158. Stewart, A. *et al.* Single-Cell Transcriptomic Analyses Define Distinct Peripheral B Cell Subsets and Discrete Development Pathways. *Front. Immunol.* **12**, (2021).
159. Sutton, H. J. *et al.* Atypical B cells are part of an alternative lineage of B cells that participates in responses to vaccination and infection in humans. *Cell Rep* **34**, (2021).
160. van Keimpema, M. *et al.* FOXP1 directly represses transcription of proapoptotic genes and cooperates with NF- κ B to promote survival of human B cells. *Blood* **124**, 3431–3440 (2014).
161. Masle-Farquhar, E. *et al.* Uncontrolled CD21^{low} age-associated and B1 B cell accumulation caused by failure of an EGR2/3 tolerance checkpoint. *Cell Rep* **38**, 110259 (2022).
162. Cobaleda, C., Schebesta, A., Delogu, A. & Busslinger, M. Pax5: the guardian of B cell identity and function. *Nat Immunol* **8**, 463–470 (2007).
163. Hill, L. *et al.* Wapl repression by Pax5 promotes V gene recombination by Igh loop extrusion. *Nature* **584**, 142–147 (2020).

164. Garvie, C. W. & Boss, J. M. Assembly of the RFX complex on the MHCII promoter: role of RFXAP and RFXB in relieving autoinhibition of RFX5. *Biochim Biophys Acta* **1779**, 797–804 (2008).
165. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838–847 (2016).
166. Teschendorff, A. E. & Wang, N. Improved detection of tumor suppressor events in single-cell RNA-Seq data. *NPJ Genom Med* **5**, 43 (2020).
167. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol Syst Biol* **17**, e9923 (2021).
168. Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
169. Loom. <http://loompy.org/>.
170. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).
171. Miller, S. A. *et al.* Lysine-Specific Demethylase 1 Mediates AKT Activity and Promotes Epithelial-to-Mesenchymal Transition in PIK3CA-Mutant Colorectal Cancer. *Mol Cancer Res* **18**, 264–277 (2020).
172. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun* **11**, 1201 (2020).
173. Blischak, J. D., Carbonetto, P. & Stephens, M. Creating and sharing reproducible research code the workflowr way. Preprint at <https://doi.org/10.12688/f1000research.20843.1> (2019).
174. Barkas, N., Petukhov, V. & Kharchenko, P. pagoda2: Single Cell Analysis and Differential Expression. R package version 1.0.10. (2021).

175. Krivanek, J. *et al.* Dental cell type atlas reveals stem and differentiated cell types in mouse and human teeth. *Nat Commun* **11**, 4816 (2020).

Conference contributions

Taus P, Kuručová T, Zavačková K, Pal K, Valisová M, Kotásková J, Pospíšilová S, Bryja V, Janovská P, Plevová K. Interrogating the molecular heterogeneity of chronic lymphocytic leukemia through computational approaches. In CEITEC PhD Conference, Brno. 2022

Taus P, Kuručová T, Valisová M, Brychtová Y, Plevová K, Kotásková J, Bryja V, Janovská P. Data mining of publicly available scRNA-seq datasets: application of machine learning to interrogate normal cellular counterparts of chronic lymphocytic leukemia. In 11th International Conference Analytical Cytometry, Ostrava. 2021

Taus P, Henzl J, Plevová K. Identification of novel chronic lymphocytic leukemia subtypes using germline pathway mutation scores and ensemble clustering. In EMBL Conference: Cancer Genomics, Heidelberg, Germany. 2019.

Taus P, Plevová K, Darzentas N, Pal K, Pospíšilová S. Identification of novel chronic lymphocytic leukemia subtypes using pathway mutation scores and consensus clustering. In ESHG 2019, Gothenburg, Sweden. 2019.

Taus P, Darzentas N, Plevová K. Identification of novel chronic lymphocytic leukemia subtypes using pathway mutation scores and machine learning. In Joint PhD retreat 2019, Kouty u Ledče nad Sázavou. 2019.

Taus P, Plevová K, Hynst J, Pospíšilová S. Identification of new prognostic subtypes of chronic lymphocytic leukemia using pathway mutation score and machine learning. In XLIII Brno Oncology Days. 2019.

Selected co-authored conference contributions:

Krivánek J, Lavický J, **Taus P**, Bogdanović A, Gonzalez Lopez M, Rakultsev V, Fedr R, Souček K, Buchtová M. Revealing the complexity of mesenchymal stem cell niche of continuously growing teeth. In 14th conference of the Tooth Morphogenesis and Differentiation 2022, Prague. 2022

Krivánek J, Wentworth Winchester E, **Taus P** Single cell transcriptomics - Single cell analysis (invited workshop). In 14th conference of the Tooth Morphogenesis and Differentiation 2022, Prague. 2022

*Zavačková K, **Taus P**, Pal K, Stránská K, Pavlova S, Malčíková J, Zenatová M, Panovská A, Doubek M, Kotásková J, Pospíšilová S, Plevová K. Exploring Clonal Evolution in CLL: Analysis By Whole Exome Sequencing. In 2nd Translational Research Conference: Chronic Lymphocytic Leukaemia. 2022.

*Zavacka K, **Taus P**, Pal K, Stranska K, Pavlova S, Panovska A, Pospisilova S, Plevova K. Exploring clonal evolution and genetic causes of therapy failure in chronic lymphocytic leukemia. In ESHG Conference 2021. 2021.

*Zavacka K, **Taus P**, Pal K, Stranska K, Pavlova S, Malcikova J, Panovska A, Doubek M, Pospisilova S, Plevova K. Detailed analysis of treatment-related clonal evolution in CLL through the identification of abnormal molecular pathways. In XIX International Conference on Chronic Lymphocytic Leukemia (iwCLL). 2021.

Krivanek J, Lavicky J, **Taus P**, Gonzalez Lopez M, Rakultsev V, Kurucova T, Dunajova A, Buchtova M, Sulcova M. Unique stem cell subpopulation which ensures mesenchymal regeneration of continuously growing teeth contributes to various tissue organogenesis. In 2nd Conference of the Visegrád Group Society for Developmental Biology. Szeged, Hungary, 2021.

*Plevova K, Zavacka K, Rausch T, Hynst J, Jarosova M, **Taus P**, Pavlova S, Doubek M, Benes V, Pospisilova S. Complex structural variants and de novo fusion gene expression in chronic lymphocytic leukemia. In EMBL Conference: Cancer Genomics, Heidelberg, Germany. 2019.

Romzova M, Smitalova D, **Taus P**, Mayer J, Culen M. High Throughput Immunophenotyping and Expression Profiling at Single Cell Level Reveal BCR-ABL1 Dependent Surface Markers of Chronic Myeloid Leukemia Stem Cells. In 61st ASH Annual Meeting and Exposition, Orlando, Florida, USA. 2019



Research article 1

Identification of Clinically Relevant Subgroups of Chronic Lymphocytic Leukemia Through Discovery of Abnormal Molecular Pathways

Petr Taus¹, Sarka Pospisilova^{1,2,3*} and Karla Plevova^{1,2,3*}

¹ Central European Institute of Technology, Masaryk University, Brno, Czechia, ² Department of Internal Medicine – Hematology and Oncology, University Hospital Brno, Brno, Czechia, ³ Faculty of Medicine, Masaryk University, Brno, Czechia

OPEN ACCESS

Edited by:

Kimberly Glass,
Brigham and Women's Hospital
and Harvard Medical School,
United States

Reviewed by:

Joseph Paulson,
Genentech, Inc., United States
Maud Fagny,
UMR 7206 Eco Anthropologie et
Ethnobiologie (EAE), France

*Correspondence:

Karla Plevova
karla.plevova@mail.muni.cz
Sarka Pospisilova
sarka.pospisilova@ceitec.muni.cz

Specialty section:

This article was submitted to
Genomic Medicine,
a section of the journal
Frontiers in Genetics

Received: 10 November 2020

Accepted: 04 May 2021

Published: 28 June 2021

Citation:

Taus P, Pospisilova S and
Plevova K (2021) Identification
of Clinically Relevant Subgroups
of Chronic Lymphocytic Leukemia
Through Discovery of Abnormal
Molecular Pathways.
Front. Genet. 12:627964.
doi: 10.3389/fgene.2021.627964

Chronic lymphocytic leukemia (CLL) is the most common form of adult leukemia in the Western world with a highly variable clinical course. Its striking genetic heterogeneity is not yet fully understood. Although the CLL genetic landscape has been well-described, patient stratification based on mutation profiles remains elusive mainly due to the heterogeneity of data. Here we attempted to decrease the heterogeneity of somatic mutation data by mapping mutated genes in the respective biological processes. From the sequencing data gathered by the International Cancer Genome Consortium for 506 CLL patients, we generated pathway mutation scores, applied ensemble clustering on them, and extracted abnormal molecular pathways with a machine learning approach. We identified four clusters differing in pathway mutational profiles and time to first treatment. Interestingly, common CLL drivers such as ATM or TP53 were associated with particular subtypes, while others like NOTCH1 or SF3B1 were not. This study provides an important step in understanding mutational patterns in CLL.

Keywords: chronic lymphocytic leukemia, pathway mutation score, ensemble clustering, extreme gradient boosting, mutation subtypes

INTRODUCTION

Chronic lymphocytic leukemia (CLL) is a genetically and clinically heterogeneous disease. The disease manifestations range from asymptomatic with no need for therapy to an aggressive disease associated with therapeutic resistance and overall survival of less than 3 years (Kipps et al., 2017). CLL is divided into two main diagnostic subgroups based on the somatic hypermutation status of the immunoglobulin heavy chain variable region genes (IGHV; Damle et al., 1999; Hamblin et al., 1999). Clinical heterogeneity within both groups is substantial, nevertheless, patients with unmutated IGHV typically experience a more aggressive disease (Sutton et al., 2017). Over the past decade, genomic studies in CLL have discovered several putative drivers (Landau et al., 2013, 2015; Puente et al., 2015). Mutations in some of the drivers (e.g., mutations in TP53 and ATM genes)

are associated with worse clinical outcomes whereas, in other instances, reports of prognostic relevance vary (e.g., NOTCH1 and SF3B1) (Lazarian et al., 2017; Hallek, 2019). Many of the driver genes cluster in specific signaling pathways (Landau et al., 2013, 2015; Puente et al., 2015), however, in a significant proportion of patients, no recurrent mutation has been found (Puente et al., 2015). Still, only a limited set of molecular pathways may be abnormal due to the contribution of non-recurrent mutations that are commonly present, but their impact remains elusive and deserves further elaboration.

Stratification of CLL patients based on the entire mutation profile could improve the accuracy of prognostication as it has been shown in the context of other diagnoses (Papaemmanuil et al., 2016; Schmitz et al., 2018). In acute myeloid leukemia, patients assigned into subgroups based on patterns of co-mutations in 111 driver genes displayed different clinical outcomes (Papaemmanuil et al., 2016). However, this approach is challenging for a disease as genetically heterogeneous as CLL. An alternative approach is to use prior knowledge of a protein-protein interaction network to reduce the heterogeneity and classify patients into subtypes (Hofree et al., 2013; Leiserson et al., 2015; Le Morvan et al., 2017). For example, mutations can be aggregated in network neighborhoods using network propagation that spreads the signal from mutated drivers to other functionally related genes in network space (Hofree et al., 2013). A limitation of such approaches, using the protein-protein interaction network, is that the genes involved in a biological process do not always interact physically.

Kuijjer et al. (2018) developed a method for reducing heterogeneity of mutation data using biological pathways. This approach takes into account all genes in a pathway and quantifies the level of disruption of the pathway function. Based on this approach, the authors identified nine pan-cancer mutation subtypes across the 23 cancer types from The Cancer Genome Atlas (Kuijjer et al., 2018). To the best of our knowledge, either network- or pathway-based stratification of CLL patients using mutation data has not been performed until now.

Unsupervised learning, also known as clustering, has been extensively used to gain insight into the underlying structure of complex biological data and has led to discoveries of various cancer molecular subtypes (Noushmehr et al., 2010; Cancer Genome Atlas Research Network, 2011; Hedegaard et al., 2016). However, there are several pitfalls, stemming from the nature of biological data, which must be considered during the clustering analysis to obtain robust and meaningful results (Ronan et al., 2016). These pitfalls may be overcome by the application of a combination of multiple clustering solutions through a consensus approach (i.e., ensemble clustering). In this study, we used sequencing data gathered by the International Cancer Genome Consortium (ICGC) for 506 CLL patients to generate pathway mutation scores and applied ensemble clustering. We extracted abnormal molecular pathways with a machine learning approach and identified groups of CLL patients that differ in pathway mutational profiles, as reflected by the clinical behavior of the disease.

RESULTS

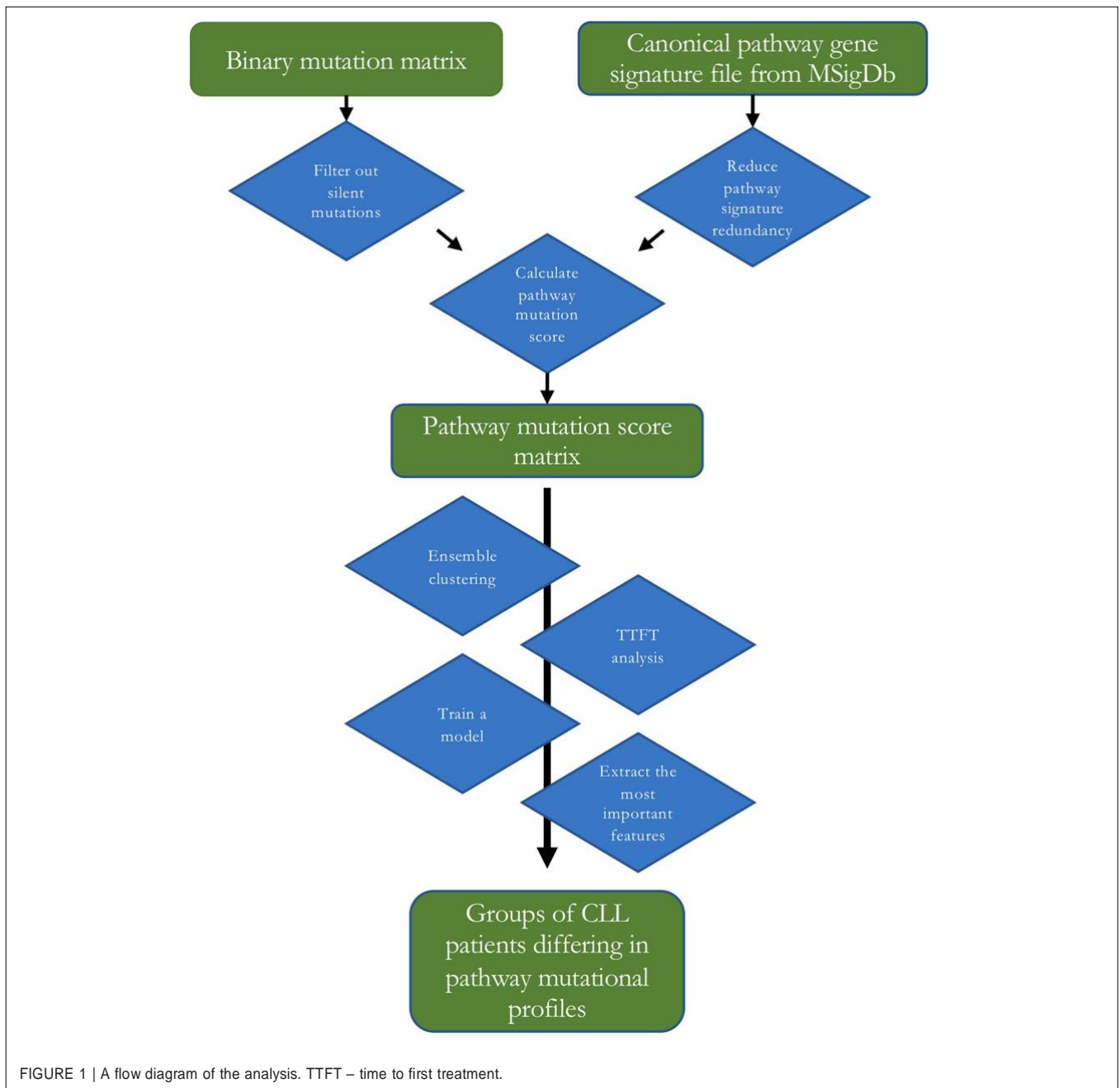
Reducing Pathway Signature Redundancy to Enhance Prognostic Subtype Identification

In the present work, we used 1,329 canonical pathway signatures (covering 8,904 genes) from the collection of curated gene sets (i.e., pathways) from the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) gathered from various sources including e.g., BioCarta, KEGG, and Reactome. Combining multiple sources of pathway information often leads to redundancy in the combined dataset that can hinder the downstream analysis. We explored the canonical signature dataset and found out that each gene belonged to 7.6 pathways on average and that the pathway sizes ranged from 6 to 1,028 genes with the median pathway size of 29 genes. This means that most of the pathways contain tens of genes encompassing specific biological processes (see **Figure 1** for a flow diagram of the presented analysis).

A set theory algorithm (Stoney et al., 2018) aimed to identify a minimum subset of gene sets required to cover genes in the combined pathway database. We expected that the application of the algorithm would reduce redundancy, decrease dimensionality and lead to the exclusion of large uninformative gene sets. We tested two algorithms, i.e., the hitting set cover and the proportional set cover, that approach pathway reduction in a slightly different way with their unique biases (Stoney et al., 2018). We applied these algorithms with 100 and 99% gene coverage on the canonical signature dataset. Using 99% gene coverage means that we allowed the algorithms not to cover the remaining one percent of genes as the covering of the remaining genes, which tend to have the most overlap with other gene sets, is often at the expense of redundancy reduction. However, this resulted only in marginal improvement of the reduction of redundancy (**Table 1**), and the excluded genes were mutated in the tested CLL patient samples. In order not to lose this information, for further analyses, we decided to use a reduced pathway dataset with all genes from canonical pathway signatures generated by the hitting set cover algorithm. The hitting set cover algorithm resulted in a 67% reduction of redundancy (from 7.6 to 3.2) and a 58% reduction of dimensionality (from 1,329 to 564) and thus outperformed the proportional set cover algorithm in both the reduction of overall redundancy and decreasing dimensionality (**Table 1**).

Identification of Prognostic Mutation Subtypes Using SAMBAR

In the next step, we tested a method called Subtyping Agglomerated Mutations By Annotation Relations (SAMBAR; Kuijjer et al., 2018), utilizing hierarchical clustering with binomial distance. We applied SAMBAR in default settings, i.e., with subsetting to cancer-associated genes, which resulted in the loss of 22% ($n = 113$) samples without mutation



in any of these genes from our patient dataset ($n = 506$). Therefore, we decided not to subset genes in the next analyses. We cut the dendrogram at $k = 2-7$ which means that we grouped the patients into 2–7 groups containing cases with the most similar pathway mutation profiles. We removed clusters of size <20 and tested time to first treatment (TTFT) differences between the subtypes. We identified those solutions with significant differences bearing potential clinical relevance. These concerned $k = 3$ and 5 that, after filtering out clusters of size <20 , contained only two clusters (**Supplementary Figure 1**).

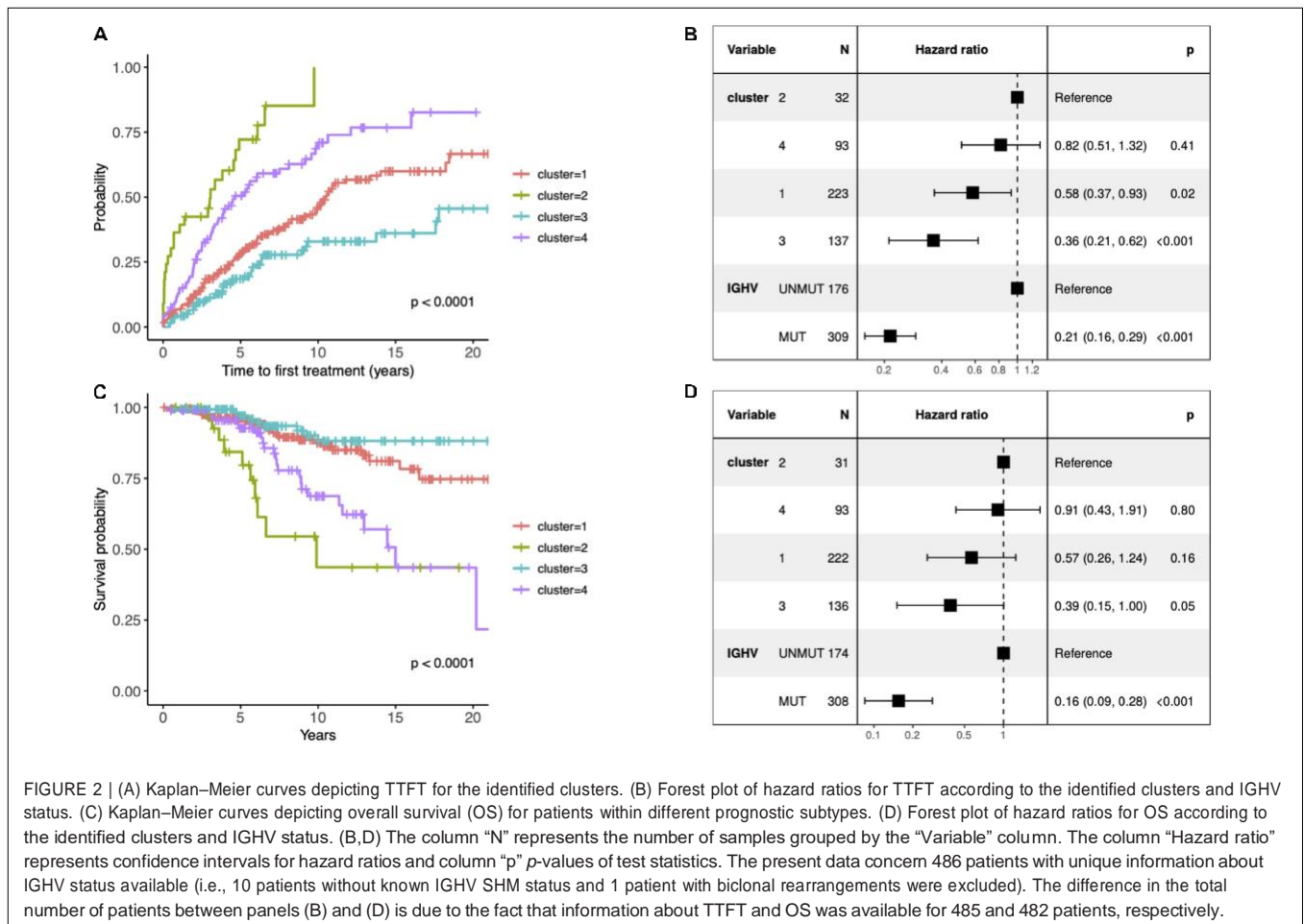
Identification of Prognostically Relevant Patient Subtypes Using Ensemble Clustering

We further explored whether we could identify subtypes with a greater prognostic value in our cohort that would be defined by distinct pathway mutation profiles. We used a combination of multiple clustering solutions through a consensus approach to cluster pathway mutation scores. We chose distinct clustering algorithms in order to maximize the diversity of the ensemble and therefore to reduce biases due to the selected algorithms

TABLE 1 | Reducing redundancy using two different set theory algorithms (hitting set cover and proportional set cover) with 100 and 99% gene coverage. The original canonical pathway signatures dataset is described in the first row.

Algorithm	Gene coverage [%]	No. of pathways	Mean pathways per gene	Min pathway size [genes]	Max pathway size [genes]	Median pathway size [genes]
	100	1,329	7.6	6	1,028	29
* Hitting set cover	100	564	3.2	8	389	34
Proportional set cover	100	669	3.5	6	389	30
Hitting set cover	99	513	2.8	8	389	32
Proportional set cover	99	603	2.9	6	389	27

Star denotes the final solution.



(see section “Materials and Methods”). We split data into 2–7 groups and evaluated differences in TTFT for the three best solutions selected based on the proportion of ambiguous clustering (PAC; Şenbabaoğlu et al., 2014). We identified subtypes with significantly different TTFT (log-rank test $p < 0.05$) for clustering solutions splitting data into 5 and 7 groups (Supplementary Figure 2). Clustering samples in 5 and 7 groups produced subtypes of 228, 33, 142, 5, 94 and 141, 57, 93, 47, 41, 66, 57 patients, respectively. As in the previous step, we removed clusters of size < 20 , therefore, after this filtering step, the clustering solution originally splitting data into 5 groups, contained only 4 groups (Figure 2A).

Since the multiclass classification that we subsequently performed was challenging, we further elaborated the solution with the fewer (i.e., 4) groups in all downstream analyses. First, we evaluated the effect of each subtype characterized by distinct pathway mutation profiles on the TTFT. The subtype with the most favorable prognosis differed from the one with the worst outcome by 20 years in the median TTFT (3 vs 23.4 years) independently of the IGHV status (Figure 2B). We also checked differences in OS, however, they were not independent of the IGHV status in the multivariate analysis (Figures 2C,D).

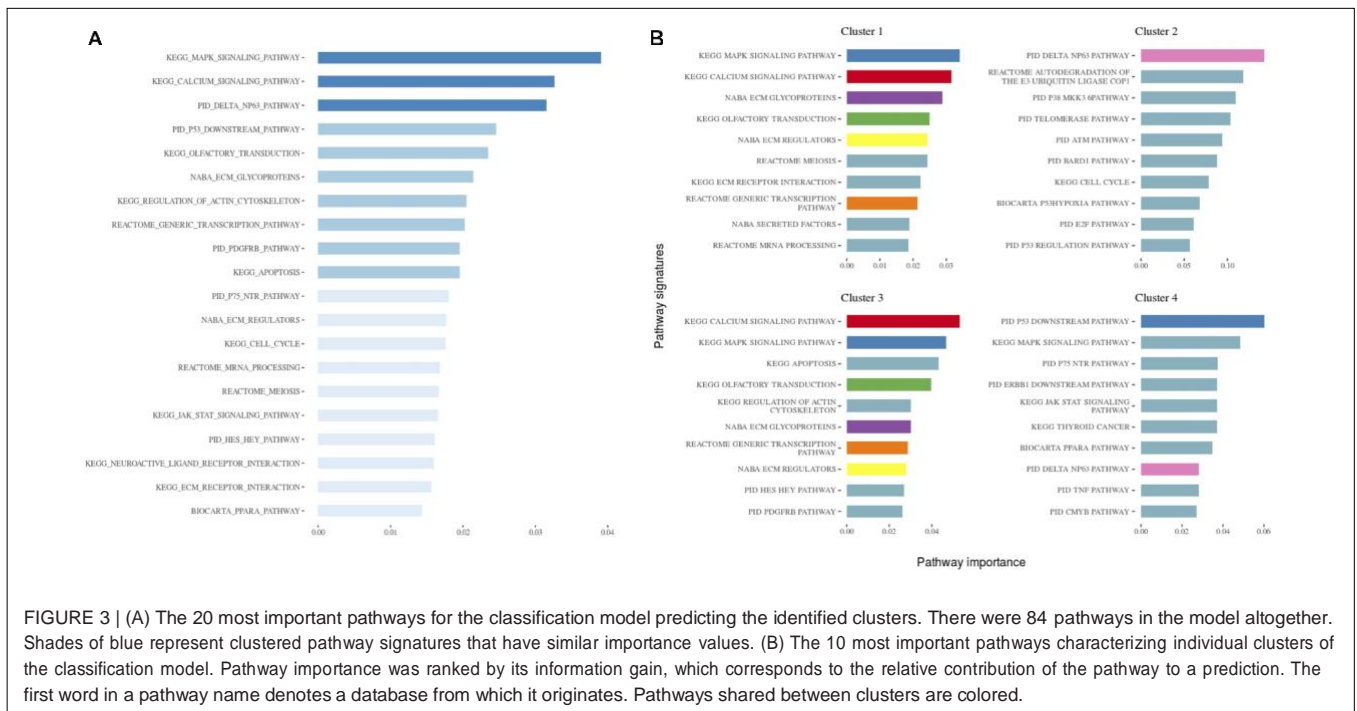


FIGURE 3 | (A) The 20 most important pathways for the classification model predicting the identified clusters. There were 84 pathways in the model altogether. Shades of blue represent clustered pathway signatures that have similar importance values. (B) The 10 most important pathways characterizing individual clusters of the classification model. Pathway importance was ranked by its information gain, which corresponds to the relative contribution of the pathway to a prediction. The first word in a pathway name denotes a database from which it originates. Pathways shared between clusters are colored.

Abnormal Molecular Pathways Extraction

We next wanted to build a classification model for the identified subtypes, which would be able to assign new cases into existing subtypes. We selected the best model based on a well suited evaluation metric for imbalanced multiclass classification $mlogLoss$ from the five-fold cross-validation, which was 0.54. Next, we evaluated the performance of the final model on a hold-out dataset ($n = 100$), i.e., samples that were not used in any step of the model development, thus representing new, unseen data. The final model used 84 pathway signatures and achieved high prediction performance (0.51 $mlogLoss$, 0.96 multiclass $auROC$, and 0.87 multiclass $aucPR$). The 84 pathway signatures contained 1,504 mutated genes in the dataset. We analyzed protein–protein interactions of mutated genes from each cluster and described gene communities using the fast greedy community detection algorithm. To interpret gene communities, we performed text mining of the column with the description of gene function for each gene and visualized networks (Supplementary Figures 3–6). Then, we extracted the top ten most important features for the model and each subtype separately (Figures 3, 4).

When investigating the most important pathway signatures for each cluster we noticed that the top ten most important pathways in Cluster 2, the cluster with the worst prognosis, all contained the *ATM* gene. *ATM* is one of the most commonly mutated genes in CLL (Puente et al., 2015) and the tested cohort, 31 out of 33 patients in Cluster 2 had *ATM* mutations. This finding prompted us to check the distributions of other common CLL driver genes (Landau et al., 2015; Puente et al., 2015) (i.e., *TP53*, *NOTCH1*, *SF3B1*, *MYD88*, *BIRC3*, *RPS15*, *FBXW7*, *BRAF*, *EGR2*, *NFKBIE*, *XPO1*, *POT1*, *ZMYM3*, and *MGA*) in all subtypes (Table 2). We found mutations in *TP53* to be solely associated with Cluster 4,

containing 94 patients, but no other mutations were specific for a particular subtype.

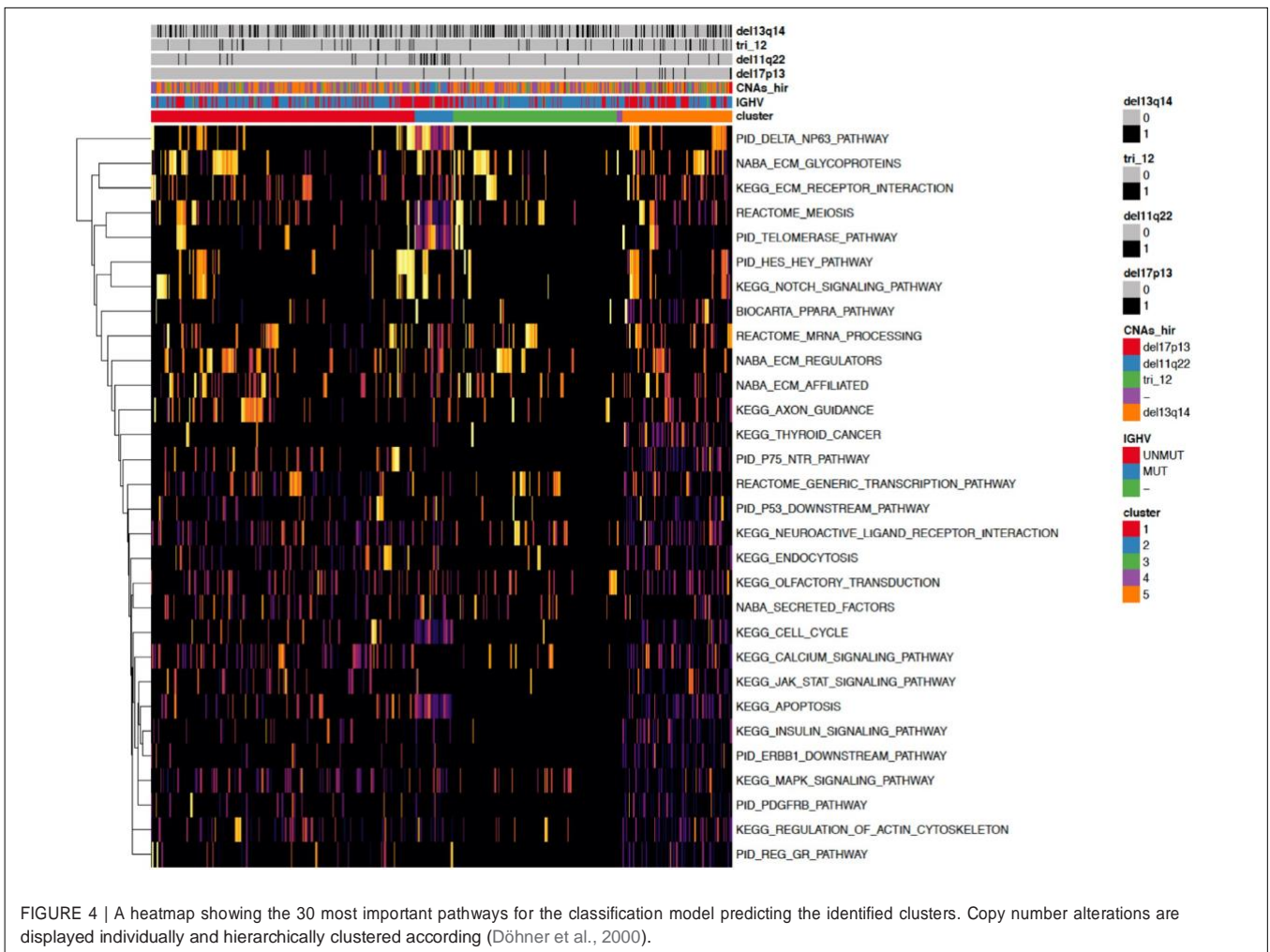
Identification of Prognostically Relevant Patient Subtypes Within IGHV Subgroups

Considering the substantial impact of IGHV somatic hypermutation status, we then explored whether we could identify subtypes separately within IGHV-mutated vs -unmutated subgroups using the ensemble clustering (Table 3). We found two subtypes among patients with unmutated IGHV differing significantly in median TTFT (3 vs 5.3 years; $p = 0.0052$; Figure 5A), but no separate subtypes among patients with mutated IGHV. The subtype with a more favorable prognosis among IGHV-unmutated cases (median TTFT 5.3 years) consisted of 61 patients, whereas the other one with a worse prognosis (median TTFT 3 years) consisted of 117 patients. Again, we checked the distribution of common CLL driver genes and found mutations in *ATM* and *TP53* only in the cluster with a worse prognosis (Table 4).

Finally, we built a classification model for the identified subtypes and extracted the most important pathway signatures for the model (Figure 5B). The final model used 35 pathway signatures (containing 1,004 mutated genes in the dataset) and achieved good prediction performance (0.92 $auROC$ and 0.85 $aucPR$).

DISCUSSION

In the present study, we built a combination of multiple clustering solutions through a consensus approach and applied it to the pathway mutation scores of CLL patients. We identified four



clusters differing in pathway mutational profiles and TTFT. Although the identification of prognostic mutation subtypes in the pan-cancer analysis by clustering pathway mutation scores has already been carried out (Kuijjer et al., 2018), to our best knowledge, this is the first attempt to apply a similar approach to a CLL dataset.

We developed machine learning models which classified CLL cases into the identified mutation subtypes with high performance. We leveraged feature importance assigned to pathway signatures by the models to extract subtype-specific pathway mutation profiles. Among the most important pathway signatures, biological processes previously described as recurrently mutated in CLL appeared frequently: namely DNA-damage response, RNA processing, and inflammatory pathways (Hallek, 2019). More importantly, we also identified processes, which have not been described as recurrently mutated in CLL but are known to play a vital role in CLL biology, such as calcium signaling (Lawrence et al., 2013; Martincorena and Campbell, 2015) and pathways involved in cellular motility and interaction (Lazarian et al., 2017). Interestingly, common CLL drivers such as ATM or TP53 were associated with

particular subtypes, while others like NOTCH1 or SF3B1 were not (Lazarian et al., 2017). These results suggest that the clinical effect of well-known CLL driver genes depends on mutation background.

We anticipate that the findings of our study will have implications for the improved identification of patients with high-risk CLL, even without well-known CLL drivers. In addition, using pathway mutation scores rather than single-gene approaches could help to identify groups of CLL patients who might respond to specific targeted therapies. This is of importance especially in the light of current treatment options (Hallek, 2019). For example, we hypothesize that patients with affected pathways involved in calcium signaling could respond differently to the treatment with Bruton's tyrosine kinase inhibitors since calcium signaling can be triggered by BCR pathway stimulation (Chiu and Talhouk, 2018). We believe that our findings will pave the way for the design of new personalized treatment strategies focusing not only on well-known driver genes but also taking into account mutational patterns in particular biological pathways.

TABLE 2 | Distribution of common CLL driver genes among the identified clusters.

Cluster	No. of patients	TP53	ATM	NOTCH 1	SF3B1	MYD88	BIRC3	RPS15	FBXW7	BRAF	EGR2	NFKBIE	XPOI	POT1	ZMYM3	MGA
1	228	0 (0%)	0 (0%)	22 (10%)	19 (8%)	14 (6%)	3 (1%)	3 (1%)	1 (0%)	0 (0%)	5 (2%)	3 (1%)	7 (3%)	6 (3%)	2 (1%)	5 (2%)
2	33	0 (0%)	31 (94%)	6 (18%)	8 (24%)	0 (0%)	1 (3%)	0 (0%)	1 (3%)	1 (3%)	3 (9%)	1 (3%)	0 (0%)	2 (6%)	1 (3%)	2 (6%)
3	142	0 (0%)	0 (0%)	4 (3%)	6 (4%)	0 (0%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	5 (4%)	2 (1%)	1 (1%)
4	94	15 (16%)	0 (0%)	16 (17%)	8 (9%)	4 (4%)	5 (5%)	0 (0%)	3 (3%)	9 (10%)	1 (1%)	1 (1%)	2 (2%)	4 (4%)	2 (2%)	4 (4%)

TABLE 3 | Distribution of IGHV somatic hypermutation status among the identified clusters.

Cluster	No. of patients	MUT	UNMUT
1	228	155 (68%)	69 (30.3%)
2	33	4 (12.1%)	28 (84.8%)
3	142	107 (75.4%)	30 (21.1%)
4	94	43 (45.7%)	50 (53.2%)

MATERIALS AND METHODS

Processing of Somatic Mutation Data

Somatic mutation data were downloaded from a published study (Puente et al., 2015) containing 506 pre-treatment patient samples. Among these, 452 patients were diagnosed with CLL and 54 with MBL. By IGHV somatic hypermutation status, there were 316 IGHV-mutated cases and 179 IGHV-unmutated cases, 1 biclonal, and 10 undetermined cases. Silent mutations were filtered out and only mutations in protein-coding regions and splice sites were kept. Then mutational matrix was binarized. The average number of affected genes per patient was 14.1. If not stated otherwise all analyses were performed using R software v3.4.4 (R Core Team, 2020). The **supplementary Figures 3–6** were prepared using R software v3.4.4 (R Core Team, 2020) and Cytoscape software v3.7.1 (Shannon et al., 2003).

Reducing Pathway Signature Redundancy

Proportional and hitting set cover algorithms (Stoney et al., 2018) were applied on the canonical pathway gene signature file “c2.cp.v6.2.symbols.gmt” downloaded from MSigDb (Liberzon et al., 2011). The gene coverage threshold was set to 100 and 99%, meaning that one percent of the genes from the original dataset would be missing in the resulting reduced datasets. Then, the excluded genes were checked, whether they were mutated in the patient cohort, and properties of the pathway sets (such as median pathway size, mean paths per gene, min/max pathway size, and the number of pathways) were calculated and compared before and after reduction. Based on this evaluation, a pathway signature dataset was created by the application of a hitting set cover algorithm with a 100% gene coverage threshold was chosen for further analysis.

Mutation Subtype Identification Using SAMBAR R Package

The *sambar* function from the SAMBAR package v0.2 was used to identify CLL mutation subtypes. The function subsets somatic mutation data to 2,352 cancer-associated genes, divides the number of mutations by the gene length, and calculates gene mutation score. Then, it corrects for sample-specific mutation rate and for the number of pathways each gene belongs to, and de-sparsifies gene mutation score into pathway mutation score when it corrects for pathway length. In the final step, it performs hierarchical clustering with binomial distance on the pathway mutation score.

However, gene length normalization is only a partial correction for the background mutation rate, which depends on other features including 3D structure, gene expression level, and GC content (Martincorena and Campbell, 2015). Additionally, we hypothesized that gene length normalization is relevant in tumor types with a high mutation rate but in tumors with low mutation rates, including CLL (Lawrence et al., 2013), this correction could introduce noise in the data. Therefore, we decided to omit this correction and binarized the mutation score. The function was further modified to exclude subsetting to cancer-associated genes. Then, it was applied on the whole patient cohort following the instruction on <https://github.com/mararie/SAMBAR> and in Kuijjer et al. (2018) with the reduced pathway signature file as a signature input for the *sambar* function. Two to seven subtypes were assessed.

Identification of CLL Subtypes Using Ensemble Clustering

The pathway mutation score was calculated using the *sambar* function but without gene length correction and subsetting to cancer-associated genes. De-sparsification of somatic mutation data resulted in a data matrix containing 503 patients and 553 pathway signatures. The pathway signatures that were affected in less than 10 patients were removed, leaving us with 502 patients and 344 pathways. Ensemble clustering was applied on pathway mutation score for the whole cohort and the cohorts with mutated and unmutated IGHV using the R package diceR v0.5.2 (Chiu and Talhouk, 2018). Four distance-based and two non-distance-based methods were included. The distance-based methods were the following: Ward linkage hierarchical clustering, divisive analysis clustering, partition around medoids, and k-means. As the distance metrics for these algorithms, binomial and Mahalanobis distance and random forests proximity converted to distance were used. The non-distance-based methods were the following: spectral clustering

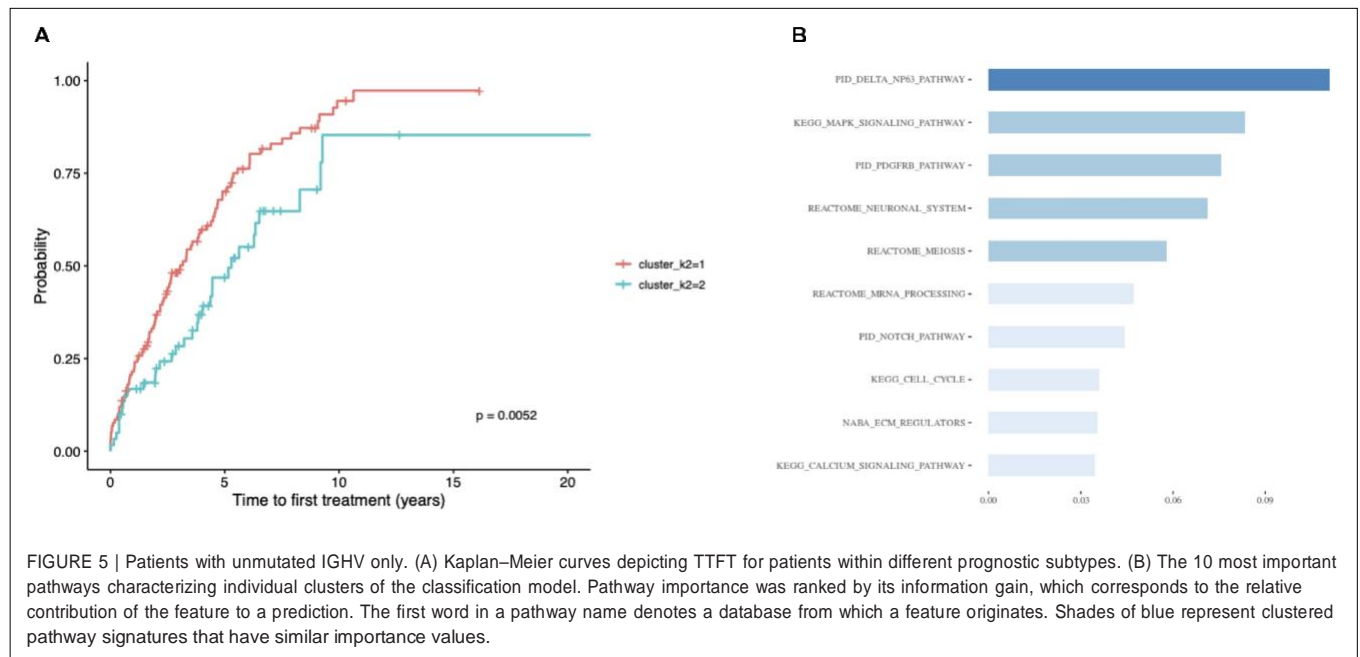


FIGURE 5 | Patients with unmutated IGHV only. (A) Kaplan–Meier curves depicting TTFT for patients within different prognostic subtypes. (B) The 10 most important pathways characterizing individual clusters of the classification model. Pathway importance was ranked by its information gain, which corresponds to the relative contribution of the feature to a prediction. The first word in a pathway name denotes a database from which a feature originates. Shades of blue represent clustered pathway signatures that have similar importance values.

TABLE 4 | Distribution of common CLL driver genes between the clusters identified within the unmutated IGHV subgroup.

Cluster	N of patients	TP53	ATM	NOTCH 1	SF3B1	MYD88	BIRC3	RPS15	FBXW7	BRAF	EGR2	NFKBIE	XPO1	POT1	2MYM3	MGA
1	117	8 (7%)	26 (22%)	34 (29%)	14 (12%)	0 (0%)	6 (5%)	3 (3%)	4 (3%)	8 (7%)	7 (6%)	3 (3%)	6 (5%)	11 (9%)	4 (3%)	7 (6%)
2	61	0 (0%)	0 (0%)	6 (10%)	9 (15%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)	0 (0%)	0 (0%)	1 (2%)	3 (5%)	5 (8%)	2 (3%)	4 (7%)

using radial-basis kernel function and self-organizing map with hierarchical clustering. Ninety percent (90%) resampling on five replicates was performed and the 2–7 subtypes were evaluated. The average PAC across the clustering results was assessed and half of the solutions with the lowest PAC were selected for further evaluation. Subsequently, the K-modes algorithm was applied to combine the results of the clustering.

Associations With Clinical Parameters

Publicly available clinical data were downloaded from the ICGC Data Portal and information about TTFT as an important clinical parameter was extracted. A log-rank test was used to identify whether the found subtypes differed in TTFT (p -value < 0.05). All the P values were adjusted for multiple comparisons using the Benjamini–Hochberg correction. If more solutions differed in TTFT statistically significantly, the one with the least subtypes was chosen for further analysis. A Multivariate Cox regression model was fitted to assess the independent prognostic impact of IGHV somatic hypermutation status of each subtype in the outcome of the patients.

A Classification Model for the Identified Subtypes

The Extreme gradient boosting algorithm (Chen and Guestrin, 2016) is a machine learning approach that combines a large number of weak learners (i.e., slightly better than random

guessing) based on decision trees into a single strong learner (i.e., a prediction model). The prediction model can then be applied to a single sample to calculate a group probability. Here we aimed to build a classification model for the identified subtypes and to extract the most important features for each cluster in the prediction model. The extreme gradient boosting algorithm from R package xgboost v0.82.1 was implemented using pathway mutation scores as the input features. Before a model tuning, highly correlated features ($r > 0.7/r < -0.7$) and clusters smaller than 10 patients were removed leaving us with 497 patients and 317 pathway signatures. Then, data were split randomly into a training set (80% of patients) and a test set (20% of patients). To find the best number of rounds for the algorithm, it was run with subsample parameter set to 0.25 and the following parameter settings of learning rate and depth of trees were tested: 0.01, 0.05, 0.1, 0.3, and 4, 6, 9, respectively. The algorithm was stopped after 100 rounds without improvement of multiclass Logarithmic Loss function (mlogloss), which was evaluated using a five-fold CV. The algorithm was run again with an optimized number of rounds and selected parameter setting, which minimized mlogloss. Feature importance was ranked by its information gain, which corresponded to the relative contribution of the feature to a prediction. The process of the parameter tuning was repeated with half of the most important features and then in the following repetitions with 3/4 of the most important features until mlogloss started increasing. The performance of the model with optimized parameters and extracted features was tested using mlogloss,

multiclass auROC, and multiclass aucPR. An information gain of the features was extracted for each subtype separately.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

Ethical review and approval was not required for the study due to the secondary use of published data. The original written informed consents with the research use of the data were collected by the ICGC consortium.

AUTHOR CONTRIBUTIONS

PT designed the research and performed the analysis. PT and KP analyzed the results and wrote the manuscript. KP and SP supervised the study and critically evaluated the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

- Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166
- Chen, T., and Guestrin, C. (2016). “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY: ACM).
- Chiu, D. S., and Talhouk, A. (2018). diceR: an R package for class discovery using an ensemble driven approach. *BMC Bioinform* 19:11. doi: 10.1186/s12859-017-1996-y
- Damle, R. N., Wasil, T., Fais, F., Ghiotto, F., Valetto, A., Allen, S. L., et al. (1999). Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 94, 1840–1847. doi: 10.1182/blood.v94.6.1840
- Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., et al. (2000). Genomic aberrations and survival in chronic lymphocytic leukemia. *N. Engl. J. Med.* 343, 1910–1916.
- Hallek, M. (2019). Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment. *Am. J. Hematol.* 94, 1266–1287. doi: 10.1002/ajh.25595
- Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G., and Stevenson, F. K. (1999). Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* 94, 1848–1854. doi: 10.1182/blood.v94.6.1848
- Hedegaard, J., Lamy, P., Nordentoft, I., Algaba, F., Høyer, S., Ulhøi, B. P., et al. (2016). Comprehensive transcriptional analysis of early-stage urothelial carcinoma. *Cancer Cell* 30, 27–42.
- Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651
- Kipps, T. J., Stevenson, F. K., Wu, C. J., Croce, C. M., Packham, G., Wierda, W. G., et al. (2017). Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Primer* 3:16096.
- Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., and Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. *Br. J. Cancer* 118, 1492–1501. doi: 10.1038/s41416-018-0109-7

FUNDING

This research was supported by the projects MH-CZ AZV NU21-08-00237 and MH-CZ DRO FNBr 65269705, and MEYS-CZ MUNI/A/1595/2020.

ACKNOWLEDGMENTS

The authors greatly appreciate the computational resources supplied by the project “e-Infrastruktura CZ” (e-INFRA LM2018140) provided within the program Projects of Large Research, Development, and Innovations Infrastructures by MEYS-CZ. The sequencing data from the ICGC consortium were provided by the Data Access Compliance Office (DACO) under application no. DACO-1062301. PT is a holder of Brno Ph.D. Talent Scholarship funded by the Brno City Municipality. The content of this manuscript is part of the doctoral thesis of PT.

SUPPLEMENTARY MATERIAL


The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.627964/full#supplementary-material>

- Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152, 714–726.
- Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., et al. (2015). Mutations driving CLL and their evolution in progression and relapse. *Nature* 526, 525–530. doi: 10.1038/nature15395
- Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218.
- Lazarian, G., Guièze, R., and Wu, C. J. (2017). Clinical implications of novel genomic discoveries in chronic lymphocytic leukemia. *J. Clin. Oncol.* 35, 984–993. doi: 10.1200/jco.2016.71.0822
- Le Morvan, M., Zinovyev, A., and Vert, J.-P. (2017). NetNorM: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Comput. Biol.* 13:e1005573. doi: 10.1371/journal.pcbi.1005573
- Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168
- Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260
- Martincorena, I., and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science* 349, 1483–1489. doi: 10.1126/science.aab4082
- Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17, 510–522.
- Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* 374, 2209–2221.
- Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519–524.

- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: <https://www.R-project.org/>.
- Ronan, T., Qi, Z., and Naegle, K. M. (2016). Avoiding common pitfalls when clustering biological data. *Sci. Signal.* 9:re6. doi: 10.1126/scisignal.aad1932
- Schmitz, R., Wright, G. W., Huang, D. W., Johnson, C. A., Phelan, J. D., Wang, J. Q., et al. (2018). Genetics and pathogenesis of diffuse large B-Cell lymphoma. *N. Engl. J. Med.* 378, 1396–1407.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Stoney, R. A., Schwartz, J.-M., Robertson, D. L., and Nenadic, G. (2018). Using set theory to reduce redundancy in pathway sets. *BMC Bioinformatics* 19:386. doi: 10.1186/s12859-018-2355-3
- Sutton, L.-A., Hadzidimitriou, A., Baliakas, P., Agathangelidis, A., Langerak, A. W., Stilgenbauer, S., et al. (2017). Immunoglobulin genes in chronic lymphocytic leukemia: key to understanding the disease and improving risk stratification. *Haematologica* 102, 968–971. doi: 10.3324/haematol.2017.165605
- Şenbabaoğlu, Y., Michailidis, G., and Li, J. Z. (2014). Critical limitations of consensus clustering in class discovery. *Sci. Rep.* 4:6207.
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2021 Taus, Pospisilova and Plevova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Gene regulation

decoupleR: ensemble of computational methods to infer biological activities from omics data

Pau Badia-i-Mompel ^{1,2}, Jesús Vélez Santiago^{1,2}, Jana Braunger^{1,2}, Celina Geiss^{1,2}, Daniel Dimitrov^{1,2}, Sophia Müller-Dott^{1,2}, Petr Taus³, Aurelien Dugourd^{1,2}, Christian H. Holland^{1,2}, Ricardo O. Ramirez Flores^{1,2} and Julio Saez-Rodriguez^{1,2,*}

¹Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, BioQuant, Heidelberg 69120, Germany, ²Institute for Computational Biomedicine, Heidelberg University Hospital, BioQuant, Heidelberg 69120, Germany and ³Central European Institute of Technology, Masaryk University, Brno 601, Czechia

*To whom correspondence should be addressed.

Associate Editor: Marieke Lydia Kuijjer

Received on January 25, 2022; revised on February 28, 2022; editorial decision on March 1, 2022; accepted on March 4, 2022

Abstract

Summary: Many methods allow us to extract biological activities from omics data using information from prior knowledge resources, reducing the dimensionality for increased statistical power and better interpretability. Here, we present decoupleR, a Bioconductor and Python package containing computational methods to extract these activities within a unified framework. decoupleR allows us to flexibly run any method with a given resource, including methods that leverage mode of regulation and weights of interactions, which are not present in other frameworks. Moreover, it leverages OmniPath, a meta-resource comprising over 100 databases of prior knowledge. Using decoupleR, we evaluated the performance of methods on transcriptomic and phospho-proteomic perturbation experiments. Our findings suggest that simple linear models and the consensus score across top methods perform better than other methods at predicting perturbed regulators.

Availability and implementation: decoupleR's open-source code is available in Bioconductor (<https://www.bioconductor.org/packages/release/bioc/html/decoupleR.html>) for R and in GitHub (<https://github.com/saezlab/decoupler-py>) for Python. The code to reproduce the results is in GitHub (https://github.com/saezlab/decoupleR_manuscript) and the data in Zenodo (<https://zenodo.org/record/5645208>).

Contact: pub.saez@uni-heidelberg.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics Advances* online.

1 Introduction

Omics datasets, such as transcriptomics or phospho-proteomics, provide unbiased high-dimensional molecular profiles. However, their big dimensionality, combined with the highly connected nature of the molecules that are measured, makes it difficult to interpret them in a mechanistically relevant manner. Leveraging prior knowledge, we can use computational methods to infer which biological activities are relevant. For example, the activity of transcription factors (TFs) and kinases can be inferred robustly from downstream transcripts and phosphosite targets, respectively (Dugourd and Saez-Rodriguez, 2019). Over the past decade, a plethora of methods that infer biological activity has emerged, each with its own assumptions and biases.

Although comparisons and collections of these methods exist (Alhamdoosh *et al.*, 2017; Geistlinger *et al.*, 2016; Va'remo *et al.*,

2013; [Supplementary Table S1](#)), they do not incorporate recent methodological developments, such as modeling activities based on weighted mode of regulation ([Supplementary Table S2](#)). Here, we present decoupleR, an R and Python package containing a collection of methods adapted for biological activity estimation in bulk, single-cell and spatial omics data.

2 Implementation

Currently, decoupleR contains 11 different methods ([Fig. 1A](#)), these include popular methods such as AUCell (Aibar *et al.*, 2017), fast GSEA (Korotkevich *et al.*, 2021), GSVA (Ha'nzelmann *et al.*, 2013), over-representation analysis, univariate linear model (ULM) adapted from Teschendorff and Wang (2020), VIPER (Alvarez *et al.*, 2016) and others ([Supplementary Table S1](#)). The inputs of decoupleR are: (i) a matrix containing molecular feature values,

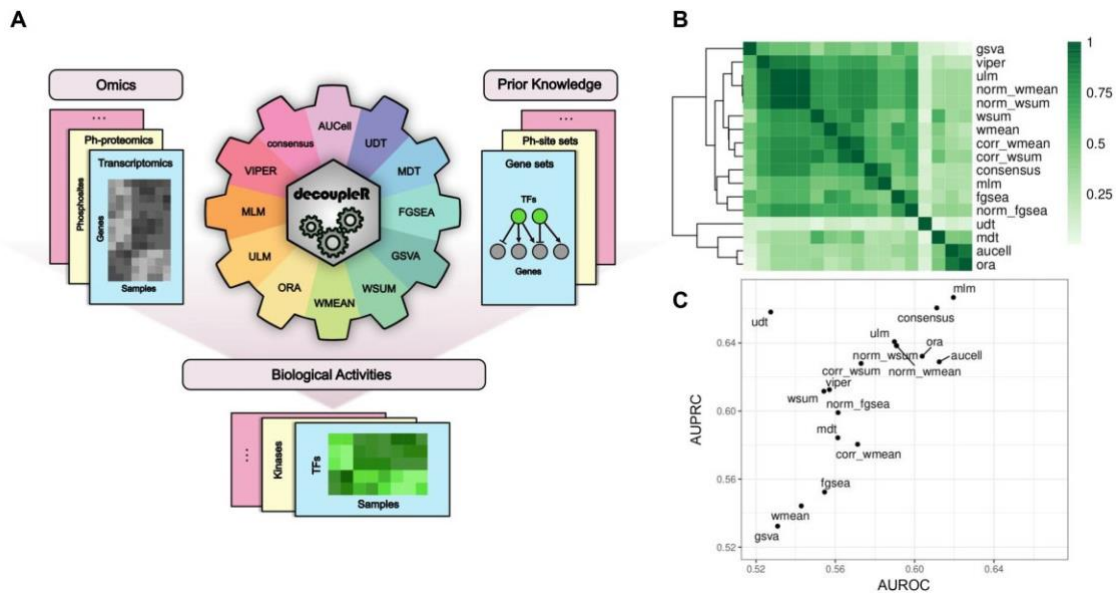


Fig. 1. Inference of biological activities with decoupleR's workflow. (A) decoupleR's workflow, it contains a collection of computational methods that coupled with prior knowledge resources estimates biological activities from omics data molecular readouts such as normalized counts or log fold changes. (B) Spearman correlation across methods and (C) predictive performance across methods in the RNA-seq data-set

either for single samples or from population comparisons, like normalized gene expression counts per sample or log fold changes and (ii) a prior knowledge resource such as a collection of gene sets. The user can then choose any method alone or many simultaneously. decoupleR also provides a consensus score obtained by computing a mean z-score across methods (Supplementary Note). Additionally, decoupleR offers easy to use wrappers to query the meta-database OmniPath (Türei et al., 2021), making it easy to flexibly access processed resources such as cell-type marker databases, gene regulatory networks or pathway footprints, and estimate biological activities from them.

3 Benchmark design

We used decoupleR to evaluate the performance of individual methods by recovering perturbed regulators—TFs and kinases—from two independent collections of transcriptomics (Holland et al., 2020) and phospho-proteomics (Hernandez-Armenta et al., 2017) datasets (Supplementary Note), respectively, upon single-gene perturbation experiments. As resources, we used the gene regulatory network DoRothEA (Garcia-Alonso et al., 2019) and a kinase substrate network (Hernandez-Armenta et al., 2017), respectively.

We built a benchmarking pipeline with decoupleR (Supplementary Note), which evaluates the performance of regulator activity scores from different methods, mainly focused on the sensitivity of methods. Furthermore, to evaluate the robustness of the methods to noise, we added or deleted a percentage of edges from the prior knowledge resources.

4 Results

Methods return different distributions of activities (Supplementary Fig. S1) but display general similarities (Supplementary Fig. S2), with a median Spearman correlation of activities between methods of 0.52, and 0.65 for transcriptomics and phospho-proteomics, respectively (Fig. 1B). There was also a moderate agreement between methods in the top 5% ranked regulators (median Jaccard indexes of 0.23 and 0.21, respectively; Supplementary Fig. S2).

Despite these similarities, methods showed different performances at predicting perturbed regulators (Supplementary Fig. S3). Some of them performed consistently better than the others

(Supplementary Table S3; Fig. 1C), the top three being: consensus, multivariate linear model and ULM. Moreover, methods that leverage weights perform better when those are taken into account (P -value $< 2.2e-16$; one-sided Wilcoxon signed-rank test; Supplementary Fig. S4).

Deleting edges in the resource had a greater effect than adding them across methods (Supplementary Fig. S5); with a median Spearman correlation of activities to the original ones of 0.84 and 0.77 for the addition and deletion, respectively (P -value $< 2.2e-16$; one-sided Wilcoxon signed-rank test). Additionally, adding or deleting edges decreased predictability, and deleting edges had a worse effect than adding (adjusted P -values $< 2.2e-16$ for normal-addition, $< 2.2e-16$ for normal-deletion and $< 2.2e-16$ for deletion-addition; $F \frac{1}{4} 131$; Tukey's HSD *post hoc* test) (Supplementary Fig. S6).

Finally, we evaluated decoupleR's speed and found that methods run relatively fast in the R version, and orders of magnitude faster in the Python one [median across methods of 1.44 and 0.44 ms per sample and regulator in R and Python, respectively, with an Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz; Supplementary Fig. S7], enabling their use with larger datasets such as single-cell or spatial omics.

5 Conclusion

In summary, decoupleR combines a variety of methods to infer biological activities into one efficient, robust, and user-friendly tool in the two most used programming languages for omics data analysis. With a common syntax for different methods, types of omics datasets, and knowledge sources available via OmniPath, it facilitates the exploration of different approaches and can be integrated in many workflows.

We observed that the majority of methods return adequate estimates of regulator activities, but that their aggregation into a consensus score and linear models perform better than other methods. We welcome the addition of further methods by the community.

Acknowledgements

We thank Celia Lerma-Martin for the design of the main figure, Atila Gabor for the technical support and Fan Zheng for the insightful discussion about regularization of multivariate models.

Funding

D.D. was supported by the European Union's Horizon 2020 research and innovation program (860329 Marie-Curie ITN 'STRATEGY-CKD').

Conflict of Interest: J.S.-R. reports funding from GSK and Sanofi and consultant fees from Travers Therapeutics and Astex Pharmaceutical.

References

- Aibar,S. *et al.* (2017) Scenic: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
- Alhamdoosh,M. *et al.* (2017) Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, **33**, 414–424.
- Alvarez,M.J. *et al.* (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.*, **48**, 838–847.
- Dugourd,A. and Saez-Rodriguez,J. (2019) Footprint-based functional analysis of multiomic data. *Curr. Opin. Syst. Biol.*, **15**, 82–90.
- Garcia-Alonso,L. *et al.* (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, **29**, 1363–1375.
- Geistlinger,L. *et al.* (2016) Bioconductor's enrichment browser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics*, **17**, 45.
- Hahnzelmann,S. *et al.* (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.
- Hernandez-Armenta,C. *et al.* (2017) Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, **33**, 1845–1851.
- Holland,C.H. *et al.* (2020) Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.*, **21**, 36.
- Korotkevich,G. *et al.* (2021) Fast gene set enrichment analysis. *bioRxiv*. DOI: <https://doi.org/10.1101/060012>.
- Teschendorff,A.E. and Wang,N. (2020) Improved detection of tumor suppressor events in single-cell RNA-seq data. *NPJ Genomic Med.*, **5**, 43.
- Türei,D. *et al.* (2021) Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.*, **17**, e9923.
- Va'remo,L. *et al.* (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.*, **41**, 4378–4391.

Research article 3

Distinct p53 phosphorylation patterns in chronic lymphocytic leukemia patients are reflected in the activation of circumjacent pathways upon DNA damage

Veronika Mancikova^{1,2*}, Michaela Pesova^{1,2*}, Sarka Pavlova^{1,2}, Robert Helma^{1,2}, Kristyna Zavacka^{1,2},
Vaclav Hejret¹, Petr Taus¹, Jakub Hynst¹, Karla Plevova^{1,2,3}, Jitka Malcikova^{1,2}, Sarka Pospisilova^{1,2,3}

¹Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic

²Department of Internal Medicine - Hematology and Oncology, Faculty of Medicine, Masaryk University and University Hospital Brno, Czech Republic

³Institute of Medical Genetics and Genomics, Faculty of Medicine, Masaryk University and University Hospital Brno, Czech Republic

*V.M and M.P. contributed equally to this study.

#Correspondence: prof. RNDr. Sarka Pospisilova, PhD, ORCID ID 0000-0001-7136-2680, Central European Institute of Technology (CEITEC), Center of Molecular Medicine, Masaryk University, Brno, Czech Republic. E-mail: sarka.pospisilova@ceitec.muni.cz

Running title: p53 phosphorylation in CLL

Keywords: p53, phosphorylation, CLL

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the [Version of Record](#). Please cite this article as doi: [10.1002/1878-0261.13337](https://doi.org/10.1002/1878-0261.13337)

This article is protected by copyright. All rights reserved

ABSTRACT

TP53 gene abnormalities represent the most important biomarker in chronic lymphocytic leukemia (CLL). Altered protein modifications could also influence p53 function, even in the wild-type protein. We assessed the impact of p53 protein phosphorylations on p53 functions as an alternative inactivation mechanism. We studied p53 phospho-profiles induced by DNA-damaging agents (fludarabine, doxorubicin) in 71 *TP53*-intact primary CLL samples. Doxorubicin induced two distinct phospho-profiles: profile I (heavily phosphorylated) and profile II (hypophosphorylated). Profile II samples were less capable of activating p53 target genes upon doxorubicin exposure, resembling *TP53*-mutant samples at the transcriptomic level, whereas standard p53 signaling was triggered in profile I. *ATM* locus defects were more common in profile II. The samples also differed in the basal activity of the hypoxia pathway: the highest level was detected in *TP53*-mutant samples, followed by profile II and profile I. Our study suggests that wild-type *TP53* CLL cells with less phosphorylated p53 show *TP53*-mutant-like behavior after DNA damage. p53 hypophosphorylation and the related lower ability to respond to DNA damage are linked to *ATM* locus defects and the higher basal activity of the hypoxia pathway.

1. INTRODUCTION

The p53 transcription factor exerts its central genome-protecting role by coordinating a regulatory circuit that senses and reacts to a wide range of stimuli, including DNA damage, abnormal oncogenic signals, or hypoxia[1]. p53 protein's stability and activity are tightly regulated through a multitude of posttranslational modifications. To date, over 50 individual p53 posttranslational modifications produced by a wide range of stress-sensing enzymes have been described. Significant differences exist in the modifications' spectra triggered by distinct stress-inducing agents, creating a highly complex and flexible signaling network[2]. Diverse combinations of these modifications allow for fine-tuning the cell response and eventually determine the final cell fate[3].

Phosphorylation belongs to the most essential p53 posttranslational modifications as it is crucial for protein stabilization and its consequent activity. Human p53 harbors an array of serine and threonine residues that can be phosphorylated by an extensive collection of kinases. Phosphorylation on the p53 N terminus shows a remarkable redundancy (multiple kinases can modify a single site, and a single kinase can phosphorylate multiple residues), highlighting the "fail-proof" layered regulation of the p53

pathway due to its central role in tumor suppression[3,4]. Once activated, p53 triggers specific transcriptional programs that control cell cycle arrest, DNA damage response, cell metabolism and apoptosis to prevent a potentially compromised cell from proliferation and, thus, propagation of mutations. Nevertheless, half of all human tumors escape this guardian mechanism by either direct mutations in the *TP53* gene or aberrations of other p53 pathway's components (e.g. MDM4 amplification[5]). However, the complete landscape of p53 pathway alterations operating in tumorigenesis is likely far from being fully portrayed.

Defects in the *TP53* gene represent the most important biomarker of chronic lymphocytic leukemia (CLL) – clinically and genetically highly heterogeneous and incurable disease. The *TP53* gene status (deletion of *TP53* locus 17p and/or *TP53* gene mutations) affects the prognosis of CLL patients and their response to therapy. Therefore, the *TP53* gene status testing has been introduced into routine clinical practice[6], and positive results provide grounds for applying targeted inhibitors of B-cell receptor or Bcl2 pathways that have shown the ability to induce a response in these difficult-to-treat patients[7].

Apart from a direct genetic impairment, the p53 pathway can be dysregulated by other mechanisms. In this regard, it has been described that decreased p53 phosphorylation can lead to changes in protein conformation affecting interaction partners of p53 protein in breast tumors, resembling a cancer-associated p53 mutated state[8]. However, whether alternative p53 phosphorylation plays a role in CLL pathogenesis remains to be explored.

Herein, we screened for the first time the p53 phosphorylation patterns of 71 *TP53*-intact primary CLL samples treated by two DNA damaging agents (fludarabine, doxorubicin) and studied the impact of DNA damage on the CLL transcriptome. We describe that while fludarabine induces a relatively uniform phospho-pattern, samples treated with doxorubicin show two different profiles. The transcriptomic analysis revealed that samples having one of these profiles fail to activate p53 signaling after DNA damage, resembling those with genetically impaired *TP53*.

2. MATERIAL AND METHODS

2.1 Human primary samples, cell lines, and culture conditions

The study has been approved by the Ethics committee of Masaryk University (number of ethics committee case EKV-2018-017). Eighty clinically-characterized primary CLL samples were provided from the biobank of the Department of Internal Medicine – Hematology and Oncology, University Hospital Brno (CZ). In this biobank, all samples were collected after written informed patient's consent, approved by the hospital ethics committee, in accordance with the Declaration of Helsinki. All patients fulfilled the iwCLL/NCI diagnostic criteria for CLL[9]. Peripheral blood samples were processed by gradient centrifugation using Ficoll-Paque PLUS (GE Healthcare) combined with RosetteSep Kit (StemCell). Obtained high-purity B lymphocytes (>98%) were vitally frozen (viability after thawing >80%). Seventy-one samples with intact *TP53* were used to study phosphorylation patterns (Table 1). Nine samples with fully expanded biallelic defect of the *TP53* locus were used as controls in mRNA expression analyses (mutation variant allele frequency (VAF) range 88-100%; *TP53* mutation accompanied with either del(17p) or cn-LOH 17p; Supplementary Table S1).

Once thawed, primary cells were kept in RPMI-1640 medium (Biosera). Additionally, the HG3 cell line was used herein (a generous gift from Prof. R. Rosenquist, Sweden). HG3 is a cell line derived from a human CLL through EBV-transformation with the wt-*TP53* gene[10] and unmutated IGHV. HG3 was also maintained in the RPMI-1640 medium. All media were supplemented with 10% (v/v) heat-inactivated fetal bovine serum (FBS; Biosera) and 1% (v/v) penicillin/streptomycin (MP Biomedicals). To induce DNA damage, cells were incubated with 1.5 μ M doxorubicin or 15 μ M fludarabine. HG3 cell line was treated for 1h, 3h, 6h, 12h and 24h; primary CLL cells were treated for 24h.

2.2 Phos-tag analysis and western blots

After treatment, cells were lysed in RIPA buffer (50 mM Tris pH 7.4, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 1% NP-40, 1 mM sodium vanadate and 50 mM NaF). Half of the sample was directly heated with 2 \times LDS loading buffer (ThermoFisher), while the other half was treated with a

1:1 mixture of Alkaline phosphatase (ThermoFisher): λ protein phosphatase (New England Biolabs) for 30 min at 30°C, and then heated with loading buffer. Lysates were then resolved by both SDS-PAGE and Phos-tag PAGE. Phos-tag (Wako Pure Chemical Industries) analysis was performed according to the manufacturer's protocol using a neutral-pH gel system and a Zinc(II) complex. The antibodies used in this study are listed in Supplementary Table S2. Imaging and quantification of western blots were performed with a UVITEC imaging system and the ImageJ program.

2.3 Genetic characterization of the samples

Somatic hypermutations in the IGHV locus were routinely screened as described previously[11,12]. Variants in the *TP53* gene were studied using in-house amplicon-based next-generation sequencing (NGS)[13,14]. Recurrent chromosomal aberrations (i.e., deletion of 17p13, 11q22.3 and 13q14.2, trisomy 12) were analyzed by FISH. To detect variants in 70 genes associated with lymphoid malignancies (Supplementary Table S3) and additional chromosomal defects, targeted NGS was performed using a custom LYNX panel with the limit of detection of 5% VAF[15]. The somatic origin of all found variants in the *ATM* gene was verified by Sanger sequencing of germline DNA isolated from buccal swabs.

SNVs and indels in exons and adjacent splice sites were identified. Additionally, the 3'UTR region of *NOTCH1* and introns of *MYC* were covered and analyzed. Variants with a minimum coverage of 100 \times , ≥ 5 variant reads, and $\geq 5\%$ VAF were called. Next, the functional impact of variants classified as missense, frameshift, in-frame, splice donor/acceptor, start loss, and stop gain were analyzed further. Only variants with population frequency $< 1\%$ or unknown in the population databases gnomAD and 1000 genomes were considered. The information about detected variants in dbSNP, COSMIC, ClinVar, VarSome, and available literature was used during variant interpretation. Finally, frameshift variants were visually inspected in the IGV program to exclude potential artifacts.

CNVs were evaluated with the limit of detection of 20% and the resolution of 300kB-1Mb for recurrent deletions on 17p, 11q, and 13q loci and 6 Mb in the rest of the genome. For this study, we focused on relevant CLL-related aberrations in chromosomes 11, 12, 13, and 17.

2.4 RNA isolation, library preparation, and NGS sequencing

RNA was isolated from CLL cells left intact in the culture medium for 24h or maintained in 1.5 μ M doxorubicin for 24h. Total RNA was isolated using TRIzol (ThermoFisher) according to the manufacturer's instructions. The RNA integrity was assessed by the Fragment Analyzer system (Agilent). Only RNAs with RIN>7.0 were processed further. RNA-Seq libraries were prepared using Lexogen QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina with polyA selection and sequenced on Illumina NextSeq 500 sequencer (read length 1 \times 75 nt). The adapters and quality trimming of raw fastq reads were performed using Trimmomatic v0.36[16]. Trimmed RNA-Seq reads were mapped against the human genome reference (hg38) annotations using STAR v2.7.3a[17]. UMIs were used for the deduplication of aligned reads[18]. Quality control after alignment concerning the number and percentage of uniquely- and multi-mapped reads, rRNA contamination, mapped regions, read coverage distribution, strand specificity, gene biotypes, and PCR duplication was performed using several tools, namely RSeQC v2.6.2[19], Picard toolkit v2.18.27 and Qualimap v.2.2.2[20], and BioBloom tools v 2.3.4-6-g433f[21].

2.4.1 Differential expression analysis

The differential gene expression was calculated based on the gene counts produced using featureCounts tool v1.6.3[22] and using Bioconductor package DESeq2 v1.20.0[23]. Volcano plots were produced using the ggplot v3.3.3 package, and MA plots were generated using the ggpubr v0.4.0 package. Heatmap was generated from selected top differentially regulated genes using R package pheatmap v1.0.10. DESeq2 normalized gene counts for all individual samples were visualized. Genes with baseMean coverage ≥ 25 and $\log_2(\text{fold-change}) \geq 1$ or ≤ -1 from comparisons of treated profile I vs. control profile I, and treated profile II vs. control profile II were considered. Ordering in such heatmap was determined by the biggest $\log_2(\text{fold-change})$ differences between profile I and profile II in descending direction. Row scaling was applied to emphasize differences between conditions.

2.4.2 PROGENy & DoRothEA

We used a footprint-based method called PROGENy (Pathway RespOnsive GENes)[24,25] to estimate signaling pathway activities based on consensus gene signatures obtained from perturbation experiments. PROGENy contains signatures for 14 signaling pathways (Androgen, EGFR, Estrogen, Hypoxia, JAK-STAT, MAPK, NFkB, p53, PI3K, TGFb, TNFa, Trail, VEGF, and WNT). The gene counts produced using featureCounts tool v1.6.3 were log₂ transformed. Then, we inspected the log₂(counts) distribution and removed transcripts with log₂(counts)<3, usually containing genes expressed under the RNAseq detection threshold. The cleaned data were normalized using vsn R package v3.60.0. Pathway activity score was calculated with the function *progeny* from the Progeny R package v1.14.0 using the 100 most responsive genes per pathway. The unpaired two-sided Student's t-test was used to compare differences in the pathway activity between the conditions. Heatmaps were generated using the R package pheatmap v1.0.10.

Additionally, we used the DoRothEA R package v1.4.1[26] to infer the HIF1A activity from the expression of its target genes. The DoRothEA is a curated, comprehensive resource built upon different types of evidence (literature-curated resources,ChIP-seq peaks, transcription factors' binding site motifs and interactions inferred directly from gene expression).

2.5 Real-time PCR analysis

The expression levels of p53 target genes *BAX*, *BBC3*, *CDKN1A*, and *GADD45A* were studied. First, 500 ng of total RNA isolated from treated and untreated cultivated cells was reverse-transcribed using Superscript II (ThermoFisher) and oligo(dT)₁₄ primer following the manufacturer's instructions. The level of target mRNA was quantified by real-time PCR using TaqMan assays (ThermoFisher), TaqMan Gene Expression Master Mix (ThermoFisher), and the QuantStudio 12 Flex Real-Time PCR system (ThermoFisher). Assays were carried out in triplicates, and negative controls were included in all PCR series. The $\Delta\Delta C_t$ method was used for the determination of mRNA content. The geometrical mean of house-keeping genes *HPRT1A* and *TBP* cycle threshold (Ct) was used as an internal standard.

For miRNA-34a expression analysis, 4 ng of total RNA isolated from non-cultivated untreated cells was reverse-transcribed using TaqMan MicroRNA Assays (ThermoFisher) and specific primers

for miRNA-34a and RNU38B following the manufacturer's instructions. Quantification of miRNA was performed by real-time PCR using TaqMan assays (ThermoFisher), ABsolute QPCR Mix, ROX (ThermoFisher), and 7500 Fast Real-Time PCR System (ThermoFisher). All reactions were carried out in triplicates with respective negative controls. The obtained miRNA-34a expression levels were normalized to RNU38B and interpreted as $2^{-\Delta C_t} * 100\%$.

2.6 Whole exome sequencing

Sequencing libraries were prepared from 100 ng of DNA using TruSeq Exome Kit (Illumina) according to the manufacturer's instructions and sequenced on NextSeq 500 machine (Illumina).

Raw sequencing data in fastq format were processed using the bcbio pipeline manager version 1.2.3.[27]. The pipeline consists of read trimming, performed by the Atropos tool[28], read alignment to the human reference genome GRCh38, performed with bwa mem[29], samtools[30] and sambamba[31], and somatic variant calling performed by mutect2[32], strelka2[33], and vardict[34] variant callers. The resulting variants were annotated using the VEP annotation software version 100.2[35]. The resulting annotated VCF files were converted to a table format using an in-house conversion script.

All detected somatic variants were manually filtered and inspected in the respective bam files using IGV software[36].

2.7 Flow-cytometric analysis

γ -H2AX phosphorylation at Ser139 was assessed using flow cytometry. Representative samples from profile I (N = 4) and profile II (N = 8) were cultured for 30 min or 24 h *in vitro* with or without 1.5 μ M doxorubicin. Afterwards, cells were collected, fixed with 4% paraformaldehyde (PFA), permeabilized with 1 \times PBS, 5% FBS and 0.5% Tween20, stained using anti-phospho-Histone H2AX (Ser139) primary antibody, clone JBW301 (Sigma) and visualized with an AlexaFluor647-conjugated secondary antibody (Invitrogen). Samples were measured using FACS Verse flow cytometer (BD Biosciences). Data were analyzed in FlowJo v.10 software.

2.8 Statistical analysis

All statistical analyses were performed using GraphPad Prism v5 and SPSS version 25. Specific statistical tests used for different study variables are described in the figure legends. All tests were two-sided. The Gaussian distribution of data was assessed. The Kaplan-Meier survival analysis was used to assess the probability of time to second treatment (TTST) from the start of first-line treatment to the initiation of second-line therapy or death of any cause. Overall survival (OS) was estimated from the initiation of treatment to death of any cause. *P*-values <0.05 were considered statistically significant.

3. RESULTS

3.1 Induction of p53 phosphorylation by DNA damaging agents in HG3 cells

Under normal conditions, the level of p53 protein is kept low; however, it is readily stabilized and activated by phosphorylation upon stress[3]. Herein, we have applied two DNA damaging agents, doxorubicin and fludarabine, to induce stabilization of p53 *in vitro*. HG3 cells' exposure to these drugs led to a gradual increase in the p53 level accompanied by phosphorylation of different serine residues over 24 hours (Figure 1A). While we observed phosphorylation of all studied sites after doxorubicin, we only detected increased phosphorylation of serine 15, 315, and 392 after fludarabine treatment, which we attributed to the lower level of total p53 protein after the induction. Additionally, we have applied Zinc(II)-Phos-tagTM PAGE analysis to readily screen the complete phospho-profile (Figure 1B). This method provides characteristic separation patterns for phosphoforms according to the number and/or site of modifications[37]. A typical control in this electrophoretic method is treating the protein lysates with a mixture of phosphatases, which helps identify the dephosphorylated form of the studied protein. In our case, Zinc(II)-Phos-tagTM method revealed that both drugs caused abundant phosphorylation of the entire fraction of p53 protein, which was only partially eliminated by the phosphatases' treatment. For further *in vitro* experiments with primary CLL cells, we selected the most extended time point (24h) when a significant p53 activation was observed for both drugs.

3.2 Primary CLL cells display two distinct p53 phospho-profiles after doxorubicin treatment

In order to assess if alternative p53 phosphorylation plays a role in CLL pathogenesis, we used the Zinc(II)-Phos-tagTM PAGE to screen the p53 phospho-profile of 71 clinically and biologically characterized CLL cases with the intact *TP53* gene (Table 1). All samples were treated separately with

doxorubicin and fludarabine *in vitro*. Phos-tag analysis revealed three major DNA damage-induced phosphoforms of p53 in primary CLL cells (marked as phosphoform p+, p++, and p+++; Figure 2A). Each of these is supposed to represent a p53 protein with different phosphorylation levels. We noticed marked differences in the phospho-profiles caused by doxorubicin and fludarabine in primary CLL cells. In detail, the fludarabine-induced pattern was relatively homogeneous among the screened samples, with phosphoform p++ being the most pronounced one in the majority of samples. Conversely, we identified two phospho-profiles, termed I and II, after doxorubicin treatment. While phosphoforms p++ and p+++ were more abundant in profile I, the hypophosphorylated p+ was the most prominent in profile II (Figure 2A, more examples of phospho-profiles are shown in Supplementary Figure S1). We also noticed that profile II samples had significantly higher basal p53 protein levels than profile I samples, where p53 was generally only detectable upon DNA damage (Figure 2B, Supplementary Figure S2).

In representative samples, the profiles showed a trend to differ in the level of p53 phosphorylation at least at two sites: profile II was less phosphorylated at serine 15 and serine 392 (Figure 2C). Out of the 71 screened CLL cases, we unequivocally assigned the doxorubicin-induced profile in 55 samples (77%). No or minimal p53 stabilization was achieved for the remaining cases, which impeded profile assignment. Association analyses between the identified profiles and important clinico-biological features are listed in Table 1. Profile II samples were enriched in those harboring deletions in 11q ($P = 0.002$). Additionally, profile II samples showed a trend towards lower basal miR-34a expression when compared to profile I (Supplementary Figure S3).

3.3 CLL samples showing phospho-profile II fail to activate the p53 signaling pathway under doxorubicin treatment

Next, we studied if the two distinct doxorubicin-induced phospho-profiles translate into transcriptomic differences in CLL cells exposed to doxorubicin. For the analysis by RNAseq, we have selected 11 representative samples with profile I, 10 samples with profile II, and 9 samples with biallelic defect of the *TP53* locus, the latter representing dysfunctional p53 (Supplementary Table S1).

Firstly, we compared untreated and doxorubicin-treated conditions in paired samples within each experimental group (profile I, profile II, and *TP53*-mutated samples). This analysis revealed 113 significantly upregulated and 35 significantly downregulated genes in samples from profile I after doxorubicin treatment ($FDR < 0.05$; $\log_2fc \geq |1|$). The differentially expressed genes identified in profile

I were enriched in the p53 signaling pathway ($P_{\text{Adjusted}} = 1.6 \times 10^{-8}$), as shown by the DAVID functional annotation analysis[38]. Surprisingly, no such genes and only three significantly downregulated genes were identified in profile II and *TP53*-mutated samples, respectively (Figure 3A, Supplementary Table S4).

Next, we searched for the most differentially expressed genes after doxorubicin treatment between profiles I and II (Figure 3B, top 55 genes). The heatmap showed an apparent change in mRNA levels of these genes upon treatment in profile I samples, while no such change was observed in *TP53*-mutated samples. Profile II samples could be characterized by an intermediate pattern (Figure 3B). Upon closer inspection, many of the top 55 genes belonged to the p53 pathway (*BBC3*, *CDKN1A*, *FDXR*, *GADD45A*, etc.).

These findings were validated by TaqMan qRT-PCR assays in an extended cohort of 38 CLL RNA samples, composed of 27 samples initially included in the RNAseq and 11 additional samples (4 profile I and 7 profile II samples). We assessed the expression of *BAX*, *BBC3*, *CDKN1A*, and *GADD45A*, the four known downstream effectors of the p53 signaling pathway[39]. We observed a significantly lower induction of expression of all selected genes in *TP53*-mutated and profile II samples upon doxorubicin treatment, as opposed to a higher induction in profile I samples (Figure 3C; $P = 0.04$, <0.0001 ; <0.0001 and 0.0001 for *BAX*, *BBC3*, *CDKN1A*, and *GADD45A*, respectively; Kruskal-Wallis test). After fludarabine treatment, the differences were not so prominent; however, profile II samples again tend to show intermediate induction of expression (Supplementary Figure S4).

Besides the standard differential gene expression analysis, we additionally applied the PROGENy[40] package to assess the overall activity of selected cancer-related pathways. Compared to conventional pathway analysis methods, this footprint-based approach is well generalizable across experimental conditions and reflects the effects of posttranslational modifications such as phosphorylation. This approach confirmed the previous findings concerning different activation of the p53 pathway among the patient subgroups (Figure 3D). Additionally, this method revealed that basal activity of the hypoxia pathway in untreated cells significantly differed among our experimental groups; the highest activity of the hypoxia pathway was found in *TP53*-mutated cells, followed by intermediate levels in profile II samples, and the lowest activity of the hypoxia pathway was in profile I. This pattern was also maintained upon treatment; DNA damage did not have any additional effect on the activity of

the hypoxia pathway (Figure 3D, Supplementary Table S5). Differentially affected genes of the hypoxia pathway, as calculated by PROGENy, are depicted in Supplementary Figure S5A, and listed in Supplementary Table S6. Additionally, we have used another resource, DoRothEA[26], to confirm our conclusions regarding the hypoxia pathway. The latter approach allowed us to calculate the activity of individual transcription factors (e.g. HIF1A) by looking at the expression patterns of their targets. This analysis confirmed the dysregulation of HIF1A among our experimental groups - HIF1A was the most active in *TP53* mutated cells, followed by profile II samples and HIF1A activity was the lowest in profile I (Supplementary Figure S5B).

Finally, we explored whether the biological differences between the two phospho-profiles affected the clinical outcome of the patients. To analyze the potential differences in treatment response, we assessed the remission duration as a function of time to second treatment. As different treatment regimens have different effectiveness and response rates, we analyzed a sub-cohort of patients uniformly treated by the chemoimmunotherapy regimen fludarabine+chlorambucil+rituximab (FCR; N=20) that represented a standard-of-care in this retrospective cohort. We were not able to show differences between patients assigned in the two phospho-profiles (Supplementary Figure S6A). For overall survival (OS), patients were stratified as to whether they had received targeted inhibitor treatment at any time during the course of the disease. We confirmed that this treatment strategy improved the outcome of the patients regardless of the profile, but no difference between profile I and profile II was observed (Supplementary Figure S6B).

3.4 Profile II samples are enriched with *ATM* locus and *MED12* aberrations

In order to gain more insight into the possible genetic drivers of the observed phospho-patterns, targeted NGS of tumor DNA was applied. In total, 70 genes (Supplementary Table S3) associated with lymphoid malignancies were studied in all but one sample with a clearly determined profile. Generated data allowed not only to identify SNVs and indels but also recurrent CLL-related CNVs (Figure 4A). The two studied profiles differed only in aberrations affecting *ATM* locus [*ATM* gene mutation and/or del(11q22.3)] – these were significantly more frequent in profile II (Figure 4B, Supplementary Table S7). Since *ATM* is a central kinase in sensing double-strand breaks, we assessed the level of

phosphorylation of γ -H2AX (Ser139) as a read-out of ATM activity and DNA damage. Profile II samples showed a mild, albeit non-significant, dampening of overall DNA damage signaling (Supplementary Figure S7).

Furthermore, in 5 patients from profile II without identified *ATM* defects, we performed exome sequencing to analyze other aberrations that could potentially contribute to the profile II phenotype. Interestingly, in 2/5 patients, we detected somatic mutations in the *MED12* gene. These pathogenic variants in *MED12* were previously reported to be recurrent in CLL patients[41]. In addition, a somatic mutation in *MED12L*, a *MED12* paralog, was found in another patient (Supplementary Table S8).

4. DISCUSSION

The tumor suppressor protein p53, encoded by the *TP53* gene localized on chromosome 17, plays a key role in the pathology of chronic lymphocytic leukemia (CLL). As its genetic inactivation by either a locus deletion and/or gene mutations is directly associated with chemo-refractoriness[42,43], it is one of the few CLL biomarkers routinely analyzed in the clinical practice. Defective protein phosphorylation has also been shown to induce a mutant-like p53 behavior[8]. Herein, we studied if alternative posttranslational phosphorylation can impair the function of wild-type p53 protein in CLL. To confirm the hypothesis, we induced p53 phosphorylation by two DNA-damaging drugs in a large set of *TP53* wild-type primary CLL samples and screened the p53 phospho-patterns. We were able to associate hypophosphorylated profile II with disrupted activation of p53 signaling as assessed by RNA sequencing and real-time PCR analysis. Moreover, we linked this p53-mutant-like state with a higher activity of the hypoxia pathway and defects in the *ATM* gene locus.

Under stress-free conditions, p53 protein's stability and activity are tightly regulated and kept low[44] through MDM2-mediated timely degradation[45]. For this master regulatory loop, the N-terminal end of p53 is essential. If it is unmodified, MDM2 readily binds it[46], triggering p53 ubiquitination and proteasomal degradation. N-terminus is targeted by a plethora of stress-sensing kinases, which phosphorylate it at multiple sites upon various forms of DNA damage, thus increasing the protein half-life[47]. In this regard, chemotherapeutic drugs, including purine analogs (such as fludarabine) or topoisomerase inhibitors (such as doxorubicin), have been shown to increase p53 level in CLL cells effectively[48,49]. We observed differences in the p53 phospho-profiles induced by fludarabine or doxorubicin in CLL cells. Moreover, doxorubicin treatment led to two distinct phospho-

profiles. This confirms the different mechanisms of action of both used drugs, which is in line with previously published findings[50]. The altered signaling resulting in p53 hypophosphorylation in profile II after doxorubicin is likely not crucial for fludarabine response, as we observed a rather homogeneous pattern of p53 phosphorylation after fludarabine. Besides, reduced phosphorylation of p53, as we observe in profile II (after doxorubicin treatment), has been related to a state that resembles mutated p53[8]. Having this in mind, CLL phospho-profile II could encompass those samples that bear genetically wild-type p53, but whose activity is impaired at the protein level by inadequate posttranslational modifications.

Indeed, we observed that, like the samples carrying an inactivated *TP53* gene, profile II samples failed to trigger p53 signaling upon DNA damage on the transcriptomic level. In this regard, it has already been described that wild-type p53 can undergo conformational changes into a mutant form with an unavailable DNA-binding domain and is thus incapable of induction of its target genes[51,52]. In detail, phospho-profile II could represent an intermediate state between wild-type and mutant p53 since the induction of the studied downstream effector genes after doxorubicin treatment was much lower in profile II than in profile I samples, but still higher than in *TP53*-mutated samples. In line with this, the expression of miR-34a, whose downregulation is a well-known indicator of deleted and/or mutated *TP53* gene[53], also showed this intermediate expression pattern in profile II samples.

In order to uncover the underlying mechanisms plausibly giving rise to the different p53 phospho-profiles, we next used PROGENy to assess the basal activity of selected cancer-related pathways in untreated cells. This analysis pointed to the importance of the hypoxia pathway, an established inducer of p53 accumulation[54]. We detected its highest activity in *TP53* mutated samples, which is in line with the recent findings[55]. Correspondingly with the above-mentioned, the hypoxia pathway's activity in profile II samples was between the high levels found in *TP53* mutants and low activity in profile I. Under hypoxic conditions, p53 is not appropriately degraded through the MDM2-mediated process[56], leading to its elevated protein levels that accumulate in the cell. In this scenario, p53 is known to be hypophosphorylated and transcriptionally inactive[52]. Our data suggest that elevated basal activity of the hypoxia pathway in profile II samples could contribute to the presence of clearly detectable levels of hypophosphorylated p53 protein in these cells and also to the p53 inability to respond to DNA damage caused by doxorubicin. Thus, increased hypoxia could contribute to the

emergence of phospho-profile II and its p53-mutant-like character. Given the increasing evidence of the importance of the hypoxia pathway in CLL pathogenesis and its potential druggability[57,58], our results point to the possibility of hypoxia pathway targeting even in wt-*TP53* patients.

Besides, we noticed that profile II samples were enriched in those harboring *ATM* defects. *ATM* is a kinase that senses and reacts to DNA double-strand breaks and stabilizes p53 through phosphorylation, especially at Ser15[59]. Although the *ATM*-p53 axis is disrupted in most profile II samples, double-strand breaks' sensing was only mildly affected, and we were still able to detect p53 protein in primary CLL cells after doxorubicin treatment. It suggests that p53 must be stabilized via an alternative pathway. Apart from *ATM*, DNA-PK is involved in response to DNA double-strand breaks[59–61]. DNA-PK was reported to be overexpressed in CLL cells with del11q (encompassing *ATM* gene)[60], and DNA-PK activity was described to be crucial for the survival of primary CLL cells with *ATM* defects[61]. In detail, after exposing cells to DNA damage, DNA-PK might act via DNA-PK/AKT/GSK3/MDM2 axis resulting in MDM2 hypophosphorylation and, consequently, p53 accumulation[59]. p53 stabilized this way (in the absence of a fully functional *ATM*) was reported to be hypophosphorylated on Ser15[59], which is in line with our results. Thus, the activity of an alternative DNA damage response pathway after using doxorubicin could contribute to p53 accumulation in *ATM* defective samples. Moreover, *ATM* loss results in chronic oxidative stress, which might be responsible for the increased biogenesis of the HIF1 protein, a key component of the hypoxia pathway[62]. This finding can thus partially explain the observed increased activity of the hypoxia pathway in profile II *ATM*-defective samples.

Mutations in *MED12/MED12L* genes could also be of importance since they were found in 3 of 5 profile II patients not having *ATM* defects. This high proportion is noteworthy, considering that mutations in the *MED12* gene were previously described in 5-8% of CLL cases[41,63,64]. *MED12* is a part of the mediator kinase module complex involved in p53 signal transduction, more specifically, it is a stimulus-specific positive coregulator of p53 target genes[65]. Moreover, it has been shown that mutation in *MED12* lead to downregulation of p53 signaling[66].

Although we clearly demonstrated biological differences between the two identified phospho-profiles, the impaired function of the p53 pathway in profile II was likely overcome by other mechanisms *in vivo*. The differences between profiles did not translate into the clinical outcome; neither

the time to second treatment nor overall survival differed between the respective patient subgroups. The lack of difference in remission duration can be explained by the small sample size as only sub-cohort of patients treated with the same regimen (FCR) could be compared. Moreover, the FCR regimen has a different mechanism of action compared to doxorubicin alone and other mechanisms might compensate for the insufficient p53 pathway activity. This is even more valid for overall survival because patients receive multiple treatment lines and different treatment regimens during the course of the disease.

5. CONCLUSIONS

Our study highlights the importance of correct p53 phosphorylation to perform its tumor suppressor roles in primary CLL cells properly. We describe a complex regulatory circuit in which higher hypoxic activity and impaired DNA double-strand breaks' sensing lead to hypophosphorylation of p53 and accumulation of this dysfunctional form in CLL cells, rendering them less responsive to acute DNA damage.

6. DECLARATIONS

Acknowledgement

We would like to thank all clinicians and nurses involved in collecting patients' samples (Prof. M. Doubek, Dr. Y. Brychtova) and all the patients. We acknowledge the CF Genomics of CEITEC supported by the NCMG research infrastructure (LM2015091) and CF Bioinformatics of CEITEC for their support with obtaining the scientific data presented in this paper. We would like to thank Tomas Loja for help with the FACS analysis. The study was funded by grant nr. 19-15737S from the Czech Science Foundation, Grants RVO 65269705, NV19-03-00091 of the Ministry of Health of the Czech Republic, and grant nr. MUNI/A/1330/2021 from Masaryk University.

Authorship contributions

S.Pospisilova, S.Pavlova, J.M. and K.P. designed the study. V.M., M.P and R.H. performed western blot analyses, real-time analyses, prepared libraries for RNAseq and LYNX panel sequencing and performed data analyses. K.Z. performed whole-exome sequencing. V.H. analyzed the RNAseq data. P.T. provided PROGENy analysis. J.H. and M.P. analyzed LYNX panel NGS data. All authors contributed to the drafting of the manuscript.

Conflict of interest

All authors declare no conflicts of interest.

Data availability statement

Further data are available upon request.

7. REFERENCES

- [1] H.F. Horn, K.H. Vousden (2007) Coping with stress: Multiple ways to activate p53, *Oncogene*. <https://doi.org/10.1038/sj.onc.1210263>.
- [2] S. Saito, H. Yamaguchi, Y. Higashimoto, C. Chao, Y. Xu, A.J. Fornace, E. Appella, C.W. Anderson (2003) Phosphorylation site interdependence of human p53 post-translational modifications in response to stress, *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M305135200>.
- [3] D.W. Meek, C.W. Anderson (2009) Posttranslational modification of p53: cooperative integrators of function., *Cold Spring Harb Perspect Biol*. <https://doi.org/10.1101/cshperspect.a000950>.
- [4] Š. Pospíšilová, V. Brázda, K. Kuchaříková, M.G. Luciani, T.R. Hupp, P. Skládal, E. Paleček, B. Vojtěšek (2004) Activation of the DNA-binding ability of latent p53 protein by protein kinase C is abolished by protein kinase CK2, *Biochem J.* 378, 939–947. <https://doi.org/10.1042/BJ20030662>.
- [5] D. Danovi, E. Meulmeester, D. Pasini, D. Migliorini, M. Capra, R. Frenk, P. de Graaf, S. Francoz, P. Gasparini, A. Gobbi, K. Helin, P.G. Pelicci, A.G. Jochemsen, J.-C. Marine (2004) Amplification of Mdmx (or Mdm4) Directly Contributes to Tumor Formation by Inhibiting p53 Tumor Suppressor Activity, *Mol Cell Biol*. <https://doi.org/10.1128/mcb.24.13.5835-5843.2004>.
- [6] S. Pospisilova, D. Gonzalez, J. Malcikova, M. Trbusek, D. Rossi, A.P. Kater, F. Cymbalista, B. Eichhorst, M. Hallek, H. Döhner, P. Hillmen, M. Van Oers, J. Gribben, P. Ghia, E. Montserrat, S. Stilgenbauer, T. Zenz (2012) ERIC recommendations on TP53 mutation analysis in chronic lymphocytic leukemia, *Leukemia*. <https://doi.org/10.1038/leu.2012.25>.
- [7] A.R. Mato, B.T. Hill, N. Lamanna, P.M. Barr, C.S. Ujjani, D.M. Brander, C. Howlett, A.P. Skarbnik, B.D. Cheson, C.S. Zent, J.J. Pu, P. Kiselev, K. Foon, J. Lenhart, S. Henick Bachow, A.M. Winter, A.L. Cruz, D.F. Claxton, A. Goy, C. Daniel, K. Isaac, K.H. Kennard, C. Timlin,

- M. Fanning, L. Gashonia, M. Yacur, J. Svoboda, S.J. Schuster, C. Nabhan (2017) Optima 1 sequencing of ibrutinib, idelalisib, and venetoclax in chronic lymphocytic leukemia: results from a multicenter study of 683 patients, *Ann Oncol*. <https://doi.org/10.1093/annonc/mdx031>.
- [8] N. Furth, N.B. Ben-Moshe, Y. Pozniak, Z. Porat, T. Geiger, E. Domany, Y. Aylon, M. Oren (2015) Down-regulation of LATS kinases alters p53 to promote cell migration, *Genes Dev*. <https://doi.org/10.1101/gad.268185.115>.
- [9] M. Hallek, B.D. Cheson, D. Catovsky, F. Caligaris-Cappio, G. Dighiero, H. Döhner, P. Hillmen, M.J. Keating, E. Montserrat, K.R. Rai, T.J. Kipps (2008) Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: A report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines, *Blood*. <https://doi.org/10.1182/blood-2007-06-093906>.
- [10] A. Rosén, A.C. Bergh, P. Gogolák, C. Evaldsson, A.L. Myhrinder, E. Hellqvist, A. Rasul, M. Björkholm, M. Jansson, L. Mansouri, A. Liu, B.T. Teh, R. Rosenquist, E. Klein (2012) Lymphoblastoid cell line with B1 cell characteristics established from a chronic lymphocytic leukemia clone by in vitro EBV infection, *Oncoimmunology*. <https://doi.org/10.4161/onci.1.1.18400>.
- [11] K. Brazdilova, K. Plevova, H. Skuhrova Francova, H. Kockova, M. Borsky, V. Bikos, J. Malcikova, A. Oltova, J. Kotaskova, B. Tichy, Y. Brychtova, J. Mayer, M. Doubek, S. Pospisilova (2018) Multiple productive IGH rearrangements denote oligoclonality even in immunophenotypically monoclonal CLL, *Leukemia*. <https://doi.org/10.1038/leu.2017.274>.
- [12] K. Plevova, H.S. Francova, K. Burckova, Y. Brychtova, M. Doubek, S. Pavlova, J. Malcikova, J. Mayer, B. Tichy, S. Pospisilova (2014) Multiple productive immunoglobulin heavy chain gene rearrangements in chronic lymphocytic leukemia are mostly derived from independent clones, *Haematologica*. <https://doi.org/10.3324/haematol.2013.087593>.
- [13] J. Malcikova, K. Stano-Kozubik, B. Tichy, B. Kantorova, S. Pavlova, N. Tom, L. Radova, J. Smardova, F. Pardy, M. Doubek, Y. Brychtova, M. Mraz, K. Plevova, E. Diviskova, A. Oltova,

- J. Mayer, S. Pospisilova, M. Trbusek (2015) Detailed analysis of therapy-driven clonal evolution of TP53 mutations in chronic lymphocytic leukemia, *Leukemia*. <https://doi.org/10.1038/leu.2014.297>.
- [14] S. Pavlova, J. Smardova, N. Tom, M. Trbusek (2019) Detection and Functional Analysis of TP53 Mutations in CLL, in: *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-8876-1_6.
- [15] V. Navrkalova, K. Plevova, J. Hynst, K. Pal, A. Mareckova, T. Reigl, H. Jelinkova, Z. Vrzalova, K. Stranska, S. Pavlova, A. Panovska, A. Janikova, M. Doubek, J. Kotaskova, S. Pospisilova (2021) LYmphoid NeXt-Generation Sequencing (LYNX) Panel: A Comprehensive Capture-Based Sequencing Tool for the Analysis of Prognostic and Predictive Markers in Lymphoid Malignancies, *Journal of Molecular Diagnostics*. <https://doi.org/10.1016/j.jmoldx.2021.05.007>.
- [16] A.M. Bolger, M. Lohse, B. Usadel (2014) Trimmomatic: A flexible trimmer for Illumina sequence data, *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btu170>.
- [17] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras (2013) STAR: Ultrafast universal RNA-seq aligner, *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bts635>.
- [18] T. Smith, A. Heger, I. Sudbery (2017) UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy, *Genome Res*. <https://doi.org/10.1101/gr.209601.116>.
- [19] L. Wang, S. Wang, W. Li (2012) RSeQC: Quality control of RNA-seq experiments, *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bts356>.
- [20] K. Okonechnikov, A. Conesa, F. García-Alcalde (2016) Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data, *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btv566>.

- [21] J. Chu, S. Sadeghi, A. Raymond, S.D. Jackman, K.M. Nip, R. Mar, H. Mohamadi, Y.S. Butterfield, A.G. Robertson, I. Birol (2014) BioBloom tools: Fast, accurate and memory-efficient host species sequence screening using bloom filters, *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btu558>.
- [22] Y. Liao, G.K. Smyth, W. Shi (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btt656>.
- [23] M.I. Love, W. Huber, S. Anders (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol.* <https://doi.org/10.1186/s13059-014-0550-8>.
- [24] M. Schubert, B. Klinger, M. Klünemann, A. Sieber, F. Uhlitz, S. Sauer, M.J. Garnett, N. Blüthgen, J. Saez-Rodriguez (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression, *Nat Commun.* <https://doi.org/10.1038/s41467-017-02391-6>.
- [25] C.H. Holland, B. Szalai, J. Saez-Rodriguez (2020) Transfer of regulatory knowledge from human to mouse for functional genomics analysis, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. <https://doi.org/10.1016/J.BBAGRM.2019.194431>.
- [26] L. Garcia-Alonso, C.H. Holland, M.M. Ibrahim, D. Turei, J. Saez-Rodriguez (2019) Benchmark and integration of resources for the estimation of human transcription factor activities, *Genome Res.* 29. <https://doi.org/10.1101/gr.240663.118>.
- [27] <https://zenodo.org/record/3743344#.YkMHJnpByUl>.
- [28] J.P. Didion, M. Martin, F.S. Collins (2017) Atropos: specific, sensitive, and speedy trimming of sequencing reads, *PeerJ.* 5. <https://doi.org/10.7717/PEERJ.3720>.
- [29] H. Li (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, <https://doi.org/10.48550/arxiv.1303.3997>.

- [30] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*. 25, 2078–2079. <https://doi.org/10.1093/BIOINFORMATICS/BTP352>.
- [31] A. Tarasov, A.J. Vilella, E. Cuppen, I.J. Nijman, P. Prins (2015) Sambamba: fast processing of NGS alignment formats, *Bioinformatics*. 31, 2032–2034. <https://doi.org/10.1093/BIOINFORMATICS/BTV098>.
- [32] <https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2>.
- [33] <https://github.com/Illumina/strelka>.
- [34] <https://github.com/AstraZeneca-NGS/VarDict>.
- [35] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, F. Cunningham (2016) The Ensembl Variant Effect Predictor, *Genome Biol.* 17. <https://doi.org/10.1186/S13059-016-0974-4>.
- [36] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov (2011) Integrative Genomics Viewer, *Nat Biotechnol.* 29. <https://doi.org/10.1038/NBT.1754>.
- [37] E. Kinoshita, E. Kinoshita-Kikuta (2011) Improved Phos-tag SDS-PAGE under neutral pH conditions for advanced protein phosphorylation profiling, *Proteomics*. <https://doi.org/10.1002/pmic.201000472>.
- [38] D.W. Huang, B.T. Sherman, R.A. Lempicki (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc.* <https://doi.org/10.1038/nprot.2008.211>.
- [39] V. Navrkalova, L. Sebejova, J. Zemanova, Z. Jaskova, M. Trbusek (2013) The p53 pathway induction is not primarily dependent on Ataxia Telangiectasia Mutated (ATM) gene activity after fludarabine treatment in chronic lymphocytic leukemia cells, *Leuk Lymphoma*. <https://doi.org/10.3109/10428194.2013.796056>.

- [40] M. Schubert, B. Klinger, M. Klünemann, A. Sieber, F. Uhlitz, S. Sauer, M.J. Garnett, N. Blüthgen, J. Saez-Rodriguez (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression, *Nat Commun.* <https://doi.org/10.1038/s41467-017-02391-6>.
- [41] B. Wu, M. Słabicki, L. Sellner, S. Dietrich, X. Liu, A. Jethwa, J. Hüllelein, T. Walther, L. Wagner, Z. Huang, M. Zapatka, T. Zenz (2017) MED12 mutations and NOTCH signalling in chronic lymphocytic leukaemia, *Br J Haematol.* 179, 421–429. <https://doi.org/10.1111/BJH.14869>.
- [42] D. Gonzalez, P. Martinez, R. Wade, S. Hockley, D. Oscier, E. Matutes, C.E. Dearden, S.M. Richards, D. Catovsky, G.J. Morgan (2011) Mutational status of the TP53 gene as a predictor of response and survival in patients with chronic lymphocytic leukemia: Results from the LRF CLL4 trial, *Journal of Clinical Oncology.* <https://doi.org/10.1200/JCO.2010.32.0838>.
- [43] S. Stilgenbauer, A. Schnaiter, P. Paschka, T. Zenz, M. Rossi, K. Döhner, A. Bühler, S. Böttcher, M. Ritgen, M. Kneba, D. Winkler, E. Tausch, P. Hoth, J. Edelmann, D. Mertens, L. Bullinger, M. Bergmann, S. Kless, S. Mack, U. Jäger, N. Patten, L. Wu, M.K. Wenger, G. Fingerle-Rowson, P. Lichter, M. Cazzola, C.M. Wendtner, A.M. Fink, K. Fischer, R. Busch, M. Hallek, H. Döhner (2014) Gene mutations and treatment outcome in chronic lymphocytic leukemia: Results from the CLL8 trial, *Blood.* <https://doi.org/10.1182/blood-2014-01-546150>.
- [44] I. Cordone, S. Masi, F.R. Mauro, S. Soddu, O. Morsilli, T. Valentini, M.L. Vegna, C. Guglielmi, F. Mancini, S. Giuliacci, A. Sacchi, F. Mandelli, R. Foa (1998) p53 expression in B-cell chronic lymphocytic leukemia: A marker of disease progression and poor prognosis, *Blood.* https://doi.org/10.1182/blood.v91.11.4342.411k39_4342_4349.
- [45] Y. Haupt, R. Maya, A. Kazaz, M. Oren (1997) Mdm2 promotes the rapid degradation of p53, *Nature.* <https://doi.org/10.1038/387296a0>.
- [46] J. Lin, J. Chen, B. Elenbaas, A.J. Levine (1994) Several hydrophobic amino acids in the p53 amino-terminal domain are required for transcriptional activation, binding to mdm-2 and the adenovirus 5 E1B 55-kD protein, *Genes Dev.* <https://doi.org/10.1101/gad.8.10.1235>.

- [47] S.Y. Shieh, M. Ikeda, Y. Taya, C. Prives (1997) DNA damage-induced phosphorylation of p53 alleviates inhibition by MDM2, *Cell*. [https://doi.org/10.1016/S0092-8674\(00\)80416-X](https://doi.org/10.1016/S0092-8674(00)80416-X).
- [48] B. Bellosillo, N. Villamor, D. Colomer, G. Pons, E. Montserrat, J. Gil (1999) In vitro evaluation of fludarabine in combination with cyclophosphamide and/or mitoxantrone in B-cell chronic lymphocytic leukemia, *Blood*. https://doi.org/10.1182/blood.v94.8.2836.420k35_2836_2843.
- [49] I. Sturm, A.G. Bosanquet, S. Hermann, D. Güner, B. Dörken, P.T. Daniel (2003) Mutation of p53 and consecutive selective drug resistance in B-CLL occurs as a consequence of prior DNA-damaging chemotherapy, *Cell Death Differ*. <https://doi.org/10.1038/sj.cdd.4401194>.
- [50] V. Navrkalova, L. Sebejova, J. Zemanova, J. Kminkova, B. Kubsova, J. Malcikova, M. Mraz, J. Smardova, S. Pavlova, M. Doubek, Y. Brychtova, D. Potesil, V. Nemethova, J. Mayer, S. Pospisilova, M. Trbusek (2013) Atm mutations uniformly lead to ATM dysfunction in chronic lymphocytic leukemia: Application of functional test using doxorubicin, *Haematologica*. <https://doi.org/10.3324/haematol.2012.081620>.
- [51] M. Weinmann, V. Jendrossek, D. Güner, B. Goecke, C. Belka (2004) Cyclic exposure to hypoxia and reoxygenation selects for tumor cells with defects in mitochondrial apoptotic pathways, *The FASEB Journal*. <https://doi.org/10.1096/fj.04-1918fje>.
- [52] R. Gogna, E. Madan, P. Kuppusamy, U. Pati (2012) Re-oxygenation causes hypoxic tumor regression through restoration of p53 wild-type conformation and post-translational modifications, *Cell Death Dis*. <https://doi.org/10.1038/cddis.2012.15>.
- [53] M. Mraz, K. Malinova, J. Kotaskova, S. Pavlova, B. Tichy, J. Malcikova, K. Stano Kozubik, J. Smardova, Y. Brychtova, M. Doubek, M. Trbusek, J. Mayer, S. Pospisilova (2009) miR-34a, miR-29c and miR-17-5p are downregulated in CLL patients with TP53 abnormalities, *Leukemia*. <https://doi.org/10.1038/leu.2008.377>.
- [54] C. Koumenis, R. Alarcon, E. Hammond, P. Sutphin, W. Hoffman, M. Murphy, J. Derr, Y. Taya, S.W. Lowe, M. Kastan, A. Giaccia (2001) Regulation of p53 by Hypoxia: Dissociation of

- Transcriptional Repression and Apoptosis from p53-Dependent Transactivation, *Mol Cell Biol*.
<https://doi.org/10.1128/mcb.21.4.1297-1310.2001>.
- [55] V. Griggio, C. Vitale, M. Todaro, C. Riganti, J. Kopecka, C. Salvetti, R. Bomben, M.D. Bo, D. Magliulo, D. Rossi, G. Pozzato, L. Bonello, M. Marchetti, P. Omedè, A.A. Kodipad, L. Laurenti, G. Del Poeta, F.R. Mauro, R. Bernardi, T. Zenz, V. Gattei, G. Gaidano, R. Foà, M. Massaia, M. Boccadoro, M. Coscia (2020) HIF-1 α is over-expressed in leukemic cells from TP53-disrupted patients and is a promising therapeutic target in chronic lymphocytic leukemia, *Haematologica*.
<https://doi.org/10.3324/haematol.2019.217430>.
- [56] C. Koumenis, R. Alarcon, E. Hammond, P. Sutphin, W. Hoffman, M. Murphy, J. Derr, Y. Taya, S.W. Lowe, M. Kastan, A. Giaccia (2001) Regulation of p53 by Hypoxia: Dissociation of Transcriptional Repression and Apoptosis from p53-Dependent Transactivation, *Mol Cell Biol*.
<https://doi.org/10.1128/mcb.21.4.1297-1310.2001>.
- [57] V. Griggio, C. Vitale, M. Todaro, C. Riganti, J. Kopecka, C. Salvetti, R. Bomben, M.D. Bo, D. Magliulo, D. Rossi, G. Pozzato, L. Bonello, M. Marchetti, P. Omedè, A.A. Kodipad, L. Laurenti, G. Del Poeta, F.R. Mauro, R. Bernardi, T. Zenz, V. Gattei, G. Gaidano, R. Foà, M. Massaia, M. Boccadoro, M. Coscia (2020) HIF-1 α is over-expressed in leukemic cells from TP53-disrupted patients and is a promising therapeutic target in chronic lymphocytic leukemia, *Haematologica*.
<https://doi.org/10.3324/haematol.2019.217430>.
- [58] M. Seiffert (2020) HIF-1 α : a potential treatment target in chronic lymphocytic leukemia, *Haematologica*. 105, 856–858. <https://doi.org/10.3324/HAEMATOL.2019.246330>.
- [59] K.A. Boehme, R. Kulikov, C. Blattner (2008) p53 stabilization in response to DNA damage requires Akt/PKB and DNA-PK, *Proc Natl Acad Sci U S A*. 105, 7785–7790.
<https://doi.org/10.1073/pnas.0703423105>.
- [60] E. Willmore, S.L. Elliott, T. Mainou-Fowler, G.P. Summerfield, G.H. Jackson, F. O’Neill, C. Lowe, A. Carter, R. Harris, A.R. Pettitt, C. Cano-Soumillac, R.J. Griffin, I.G. Cowell, C.A. Austin, B.W. Durkacz (2008) DNA-dependent protein kinase is a therapeutic target and an

- indicator of poor prognosis in B-cell chronic lymphocytic leukemia, *Clinical Cancer Research*.
<https://doi.org/10.1158/1078-0432.CCR-07-5158>.
- [61] A. Riabinska, M. Daheim, G.S. Herter-Sprie, J. Winkler, C. Fritz, M. Hallek, R.K. Thomas, K.A. Kreuzer, L.P. Frenzel, P. Monfared, J. Martins-Boucas, S. Chen, H.C. Reinhardt (2013) Therapeutic targeting of a robust non-oncogene addiction to PRKDC in ATM-defective tumors, *Sci Transl Med*. <https://doi.org/10.1126/scitranslmed.3005814>.
- [62] M. Ousset, F. Bouquet, F. Fallone, D. Biard, C. Dray, P. Valet, B. Salles, C. Muller (2010) Loss of ATM positively regulates the expression of hypoxia inducible factor 1 (HIF-1) through oxidative stress: Role in the physiopathology of the disease, *Cell Cycle*. <https://doi.org/10.4161/cc.9.14.12253>.
- [63] K. Kämpjärvi, T.M. Järvinen, T. Heikkinen, A.S. Ruppert, L. Senter, K.W. Hoag, O. Dufva, M. Kontro, L. Rassenti, E. Hertlein, T.J. Kipps, K. Porkka, J.C. Byrd, A. de la Chapelle, P. Vahteristo (2015) Somatic MED12 mutations are associated with poor prognosis markers in chronic lymphocytic leukemia, *Oncotarget*. 6, 1884–1888.
<https://doi.org/10.18632/ONCOTARGET.2753>.
- [64] R. Guièze, P. Robbe, R. Clifford, S. De Guibert, B. Pereira, A. Timbs, M.S. Dilhuydy, M. Cabel, L. Ysebaert, A. Burns, F. Nguyen-Khac, F. Davi, L. Véronèse, P. Combes, M. Le Garff-Tavernier, V. Leblond, H. Merle-Béral, R. Alsolami, A. Hamblin, J. Mason, A. Pettitt, P. Hillmen, J. Taylor, S.J.L. Knight, O. Tournilhac, A. Schuh (2015) Presence of multiple recurrent mutations confers poor trial outcome of relapsed/refractory CLL, *Blood*. 126, 2110–2117.
<https://doi.org/10.1182/BLOOD-2015-05-647578>.
- [65] A.J. Donner, S. Szostek, J.M. Hoover, J.M. Espinosa (2007) CDK8 is a stimulus-specific positive coregulator of p53 target genes, *Mol Cell*. 27, 121–133.
<https://doi.org/10.1016/J.MOLCEL.2007.05.026>.
- [66] F. Klatt, A. Leitner, I. V. Kim, H. Ho-Xuan, E. V. Schneider, F. Langhammer, R. Weinmann, M.R. Müller, R. Huber, G. Meister, C.D. Kuhn (2020) A precisely positioned MED12 activation

FIGURE LEGENDS

Figure 1. Effect of DNA damage-inducing agents (doxorubicin, fludarabine) on p53 phosphorylation in the HG3 cell line. **A.** HG3 cells were incubated with either 1.5 μ M doxorubicin or 15 μ M fludarabine for 1h, 3h, 6h, 12h, and 24h, and subsequently lysed to extract proteins. Phosphorylation of serine 6, 9, 15, 20, 46, 315, and 392 was studied by the western blot analysis, which was also used to assess the protein level of total p53. Blots shown are representative of 2 technical replicates. The exposure times were as follows: Ser6–10min, Ser9–5min, Ser15–9s, Ser20–5min, Ser46–10min, Ser315–25s, Ser392–25s, p53(total)–6s. β -actin was used as a loading control (exposure time 8s). **B.** Phos-Tag analysis of protein lysates from A. Each protein lysate was loaded untreated and treated with a mixture of phosphatases (marked with +). The phosphatase treatment serves as a control and reveals the dephosphorylated form of the studied protein. It is possible to appreciate that p53 in HG3 cells treated with the selected drugs is phosphorylated to such a high degree that only partial dephosphorylation was achieved. Residually phosphorylated isoforms are present above the unphosphorylated form, marked with a grey arrow. Images shown are representative of 3 technical replicates.

Figure 2. p53 phospho-profiling of primary CLL cells. **A.** Doxorubicin and fludarabine induced distinct phospho-profiles in the studied samples (N=71), as assessed by Phos-Tag analysis and quantified using ImageJ. While there was a consistent pattern after fludarabine, doxorubicin induced two different phospho-profiles, termed I and II. The most phosphorylated p53 phosphoform is marked as p+++ , the least phosphorylated as p+ . Phosphatase treatment (marked with +) was used to identify the unphosphorylated form of p53. Each sample was considered a biological replicate. **B.** Western blot analysis was used to compare basal levels of p53 expression in protein lysates from unstimulated CLL cells cultured for 24h (N=52). Actin was used as a loading control to normalize the signal intensity. Profile II samples had significantly higher basal p53 expression ($P = 0.039$ [*], Mann-Whitney test). Each sample was considered a biological replicate. **C.** Western blot analysis of 6 representative CLL samples was used to identify serine residues whose phosphorylation differs among the two phospho-profiles. Only Ser15 and 392 out of the panel of serine residues (Supplementary Table S2) were

evaluable. Total p53 was used as a loading control to normalize the signal intensity of phospho-antibodies. All measurements were normalized to the signal detected in sample 1393. Images of serine 15 and 392 are quantified in the right panel of the figure. Mann-Whitney test was used to evaluate the statistical significance of the results ($P = 0.100$ for both sites, representative of 3 technical replicates).

Figure 3. Transcriptomic analysis of doxorubicin-induced phospho-profiles. **A.** Volcano plots representing results of differential expression analysis comparing untreated and treated paired samples within each experimental group by RNA sequencing (individual samples were considered biological replicates: profile I N=11 samples, profile II N=10 samples, TP53 mutated N=9 samples). Fold-change is depicted on the x-axis, while significance is on the y-axis. Significantly downregulated genes after doxorubicin treatment ($FDR \leq 0.05$ and $\text{fold change} \leq 0.5$) are depicted in red, while upregulated ($FDR \leq 0.05$ and $\text{fold-change} \geq 2$) are in green. Each dot in volcano plots represents the mean value for all samples in the designated group. Even though DESeq2 analyzed over 32 000 data points in each of the three experimental groups, most of the data points had very similar or identical non-significant FDR values and/or log change values in TP53 mutant samples and profile II samples. Thus, the plots seem to depict fewer points. **B.** Heat map of 55 genes (rows) showing the highest difference between differentially expressed genes in profile I and profile II samples by RNA sequencing analysis. Columns represent level of gene expression in individual patient samples in control or doxorubicin conditions. TP53-mutated samples depict the state when p53 protein is not functional. **C.** TaqMan qRT-PCR validation of RNAseq findings. The validation included twenty-seven samples included in the RNAseq (filled symbols) and 11 additional samples (4 profile I and 7 profile II samples, open symbols), run in triplicates for each gene assayed. The expression of *BAX*, *BBC3*, *CDKN1A*, and *GADD45A* was calculated relative to the mean of two housekeeping genes using the $\Delta\Delta C_t$ method. Mann-Whitney test was used for comparing two out of three groups, while all three groups were compared by the Kruskal-Wallis test. P -values < 0.05 are coded as *, those < 0.01 as ** and < 0.001 as ***. n. s. = non-significant. Horizontal red dashed lines at $y=1$ depict no response to treatment. **D.** PROGENy analysis of transcriptomic data. Activity of selected pathways (rows) in each patient sample (columns) is shown in basal and doxorubicin-treated conditions. In the basal state, all three groups significantly differ in the

activity of the hypoxia pathway ($P = 0.009$ for profile I vs. profile II comparison, $P = 0.0009$ for profile I vs. *TP53* mut, and $P = 0.003$ for profile II vs. *TP53* mut). After doxorubicin treatment, the activity of the hypoxia pathway does not change further in either of the studied groups. Moreover, the p53 pathway is most significantly activated by doxorubicin in profile I samples ($P = 1.7 \times 10^{-8}$), followed by profile II ($P = 0.028$) and *TP53* mut ($P = 0.51$), where no significant activation of the pathway was observed.

Figure 4: A. Overview of results from NGS targeted gene panel focused on lymphoid malignancies. This panel covered 70 genes and was applied in all but one sample with a clearly determined profile (54 samples in total were sequenced). **B.** Comparison of the presence of various *ATM* locus defects in profile I and profile II samples ($P = 0.0001$ [***], Fisher's exact test).

Supporting information section

Supplementary Table S1: Overview of samples carrying *TP53* aberrations.

Supplementary Table S2: Antibodies used in the study.

Supplementary Table S3: List of NGS panel target genes focused on lymphoid malignancies.

Supplementary Table S4: List of differentially expressed genes identified when untreated and doxorubicin-treated conditions in paired samples within each experimental group were compared.

Supplementary Table S5: List of Progeny *P* values.

Supplementary table S6: List of hypoxia-related genes used in PROGENy analysis.

Supplementary Table S7: List of variants detected by targeted NGS panel (LYNX).

Supplementary table S8: List of validated somatic variants detected by whole exome sequencing.

Supplementary Figure S1: Phosphorylation patterns detected by Zn(II) Phos-Tag technique.

Supplementary Figure S2: Western blot analysis of basal p53 protein levels.

Supplementary Figure S3: Relative miR34-a expression levels in uncultured primary CLL cells.

Supplementary Figure S4: qRT-PCR of p53 targets after fludarabine treatment.

Supplementary Figure S5: Activity of hypoxia pathway and HIF1A transcription factor.

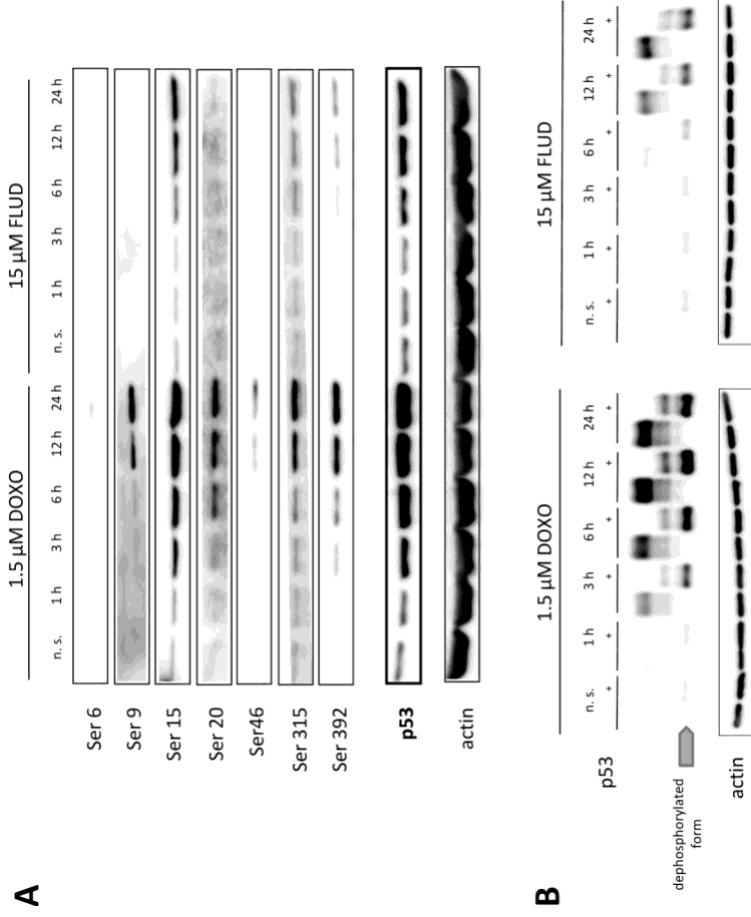
Supplementary Figure S6: Patients' clinical outcome in relation to phospho-profiles.

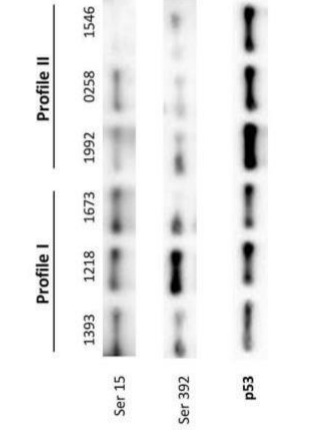
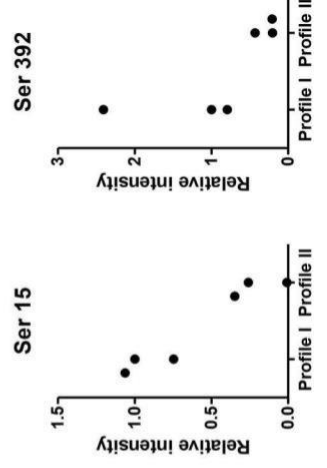
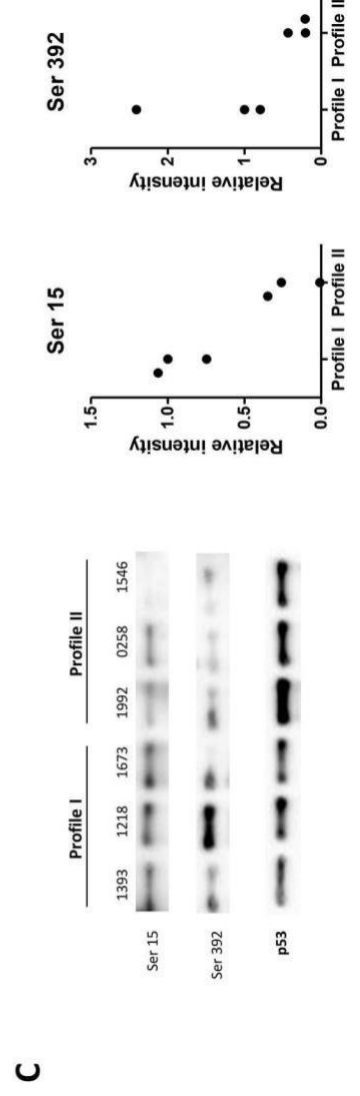
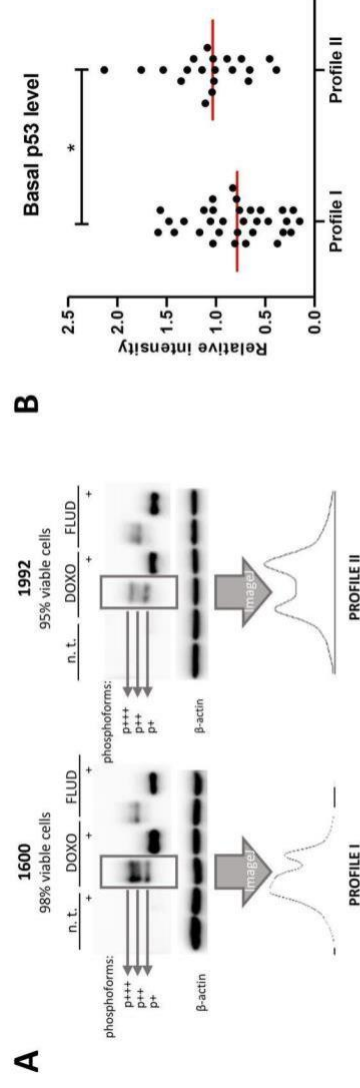
Supplementary Figure S7: H2AX phosphorylation.

Table 1. Clinico-biological characteristics of the studied patients (N=71) and association analyses between these characteristics and identified phospho-profiles. Significant *P* values are in bold. * Only

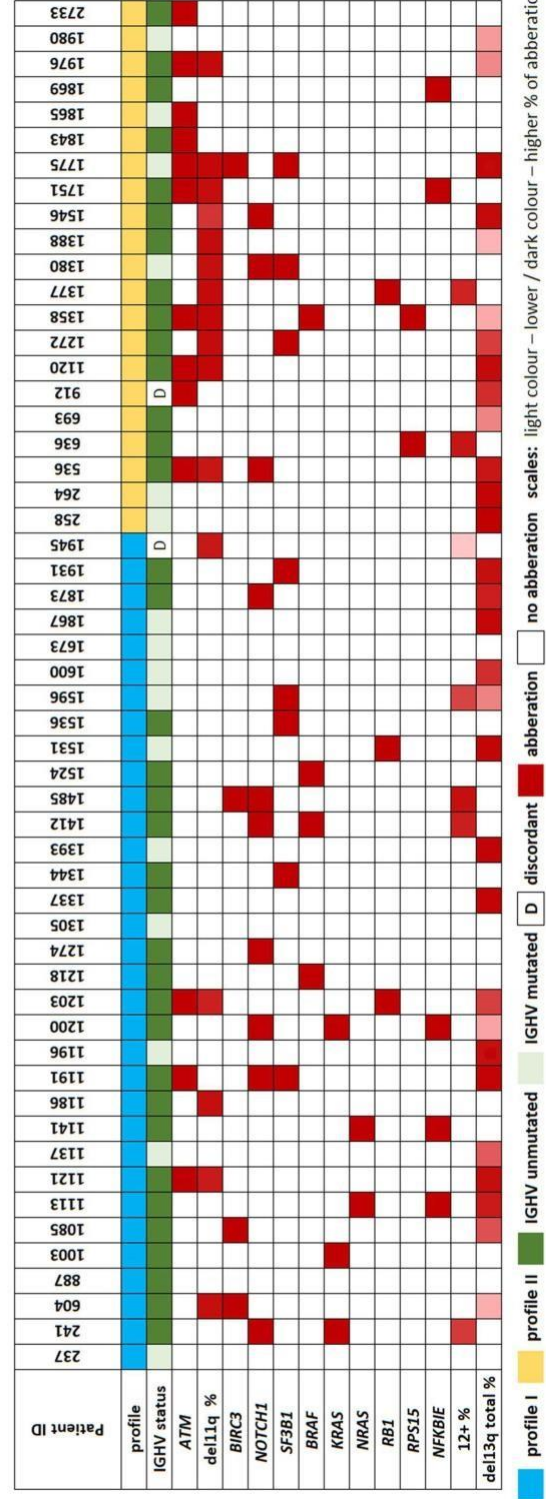
	Study cohort (N=71)	Profile I (N=33)	Profile II (N=22)	<i>P</i> -value
<i>Gender</i>				
Male (%)	48 (68)	26 (79)	9 (41)	0.009
Female (%)	23 (32)	7 (21)	13 (59)	
<i>Age at diagnosis</i>				
Median (range)	62.8 (43.2 – 85.4)	64.7 (43.5 – 82.8)	60.5 (43.2 – 85.4)	0.399
<i>Status at sampling</i>				
Never treated (%)	5 (7)	3 (9)	2 (9)	
Before treatment (%)	55 (77)	26 (79)	17 (77)	
After a therapy (%)	11 (16)	4 (12)	3 (14)	
<i>RAI staging at diagnosis</i>				
Low: 0 (%)	23 (32)	10 (35)	10 (53)	0.267*
Intermediate: I + II (%)	27 (38)	12 (41)	4 (21)	
High: III + IV (%)	12 (17)	5 (17)	2 (10)	
Unknown (%)	9 (13)	2 (7)	3 (16)	
<i>Time to first treatment from diagnosis (days)</i>				
Median (range)	1148 (34–8296)	728 (34 – 8170)	1589 (63 – 8296)	0.132*
<i>IGHV status</i>				
Unmutated (%)	48 (68)	22 (67)	15 (68)	1.0**
Mutated (%)	20 (28)	10 (30)	6 (27)	
Unknown (%)	3 (4)	1 (3)	1 (5)	
<i>11q-#</i>				
Yes (%)	24 (34)	5 (15)	12 (55)	0.002**
No (%)	45 (63)	28 (85)	8 (36)	
Unknown (%)	2 (3)	0 (0)	2 (9)	
<i>12+#</i>				
Yes (%)	9 (13)	5 (15)	2 (9)	0.694**
No (%)	58 (82)	27 (82)	18 (82)	
Unknown (%)	4 (5)	1 (3)	2 (9)	
<i>13q-#</i>				
Yes (%)	39 (55)	17 (52)	14 (64)	0.253**
No (%)	30 (42)	16 (48)	6 (27)	
Unknown (%)	2 (3)	0 (0)	2 (9)	

calculated for the samples taken prior to starting any CLL-related therapy. **Calculated only for those samples where data were available. # Assessed by FISH.

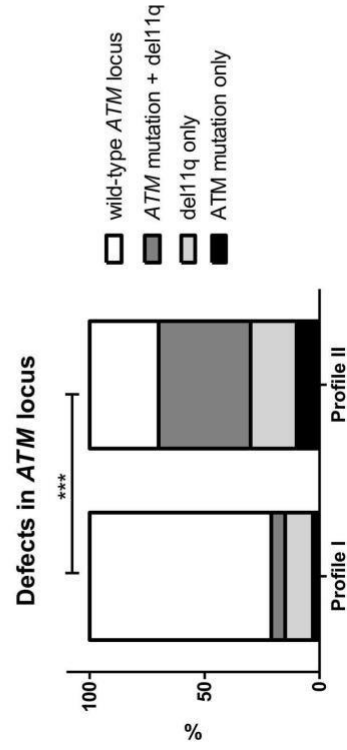




A



B



MOL2_13337_Figure 4.jpg

Research paper

Single-cell RNA sequencing analysis of T helper cell differentiation and heterogeneity



Radim Jaroušek^{a,b,1}, Antónia Mikulová^{a,b,1}, Petra Dařová^{a,b}, Petr Tauš^c, Terézia Kurucová^{b,c}, Karla Plevová^{c,d}, Boris Tichý^c, Lukáš Kubala^{a,b,*}

^a Institute of Biophysics, Czech Academy of Sciences, Brno, Czech Republic

^b Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic

^c Central European Institute of Technology, Masaryk University, Brno, Czech Republic

^d Institute of Medical Genetics and Genomics, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic

ARTICLE INFO

Keywords:

T helper cells
Activation
Differentiation
Plasticity
Single-cell RNA sequencing
Gene expression profiling
Signature genes
Differential expression
Cell cycle regression
Correction for batch effect
Data analysis

ABSTRACT

Single-cell transcriptomics has emerged as a powerful tool to investigate cells' biological landscape and focus on the expression profile of individual cells. Major advantage of this approach is an analysis of highly complex and heterogeneous cell populations, such as a specific subpopulation of T helper cells that are known to differentiate into distinct subpopulations. The need for distinguishing the specific expression profile is even more important considering the T cell plasticity. However, importantly, the universal pipelines for single-cell analysis are usually not sufficient for every cell type. Here, the aims are to analyze the diversity of T cell phenotypes employing classical in vitro cytokine-mediated differentiation of human T cells isolated from human peripheral blood by single-cell transcriptomic approach with support of labelled antibodies and a comprehensive bioinformatics analysis using combination of *Seurat*, *Nebulosa*, *GGplot* and others. The results showed high expression similarities between Th1 and Th17 phenotype and very distinct Th2 expression profile. In a case of Th2 highly specific marker genes SPINT2, TRIB3 and CST7 were expressed. Overall, our results demonstrate how donor difference, Th plasticity and cell cycle influence the expression profiles of distinct T cell populations. The results could help to better understand the importance of each step of the analysis when working with T cell single-cell data and observe the results in a more practical way by using our analyzed datasets.

1. Introduction

The concept of immune cell differentiation is based on the progressive formation of a distinct effector immunophenotype responsible for a particular function within the complex system of immune response mechanisms. Another important feature of most immune cells is the ability to transdifferentiate into a different cell subset, which is often connected to the regulation of stimulated immune response. The most widely studied are T helper cells (Th) that have distinct immunophenotypes related to their effector functions and their roles in the immune response. Therefore, distinguishing T cell subsets is crucial to understanding the status of specific immunity.

Single-cell RNA sequencing (scRNASeq) analysis is considered to be a revolutionary method in multiple biological fields as it provides a significant insight into how individual cells work at the transcriptomic

level. With the knowledge obtained from scRNASeq analysis, changes in the transcription profiles of T cells and identification of subsets based on their gene expression patterns can be determined. Regarding Th cells specifically, a more detailed gene expression profiling can contribute to the classification system that was established upon the description of the first two Th cell subsets – Th1 and Th2 [1]. Recent studies widened the T cell classification system, which now includes the following subsets: Th1, Th2, Th9, Th17, Th22, follicular T cells (Tfh), and regulatory T cells (Treg) [2].

The Th1, Th2, Th17, and Treg subsets belong to the most abundant and most studied subsets of T cells. It is known that each of these subsets orchestrates highly specific effector functions that enable the immune system to adapt its response to various pathogens. The Th1 cells belong to populations of pro-inflammatory immune cells that primarily coordinate an antiviral immune response (type 1). Their main effector

* Corresponding author at: Institute of Biophysics, Academy of Sciences of the Czech Republic, v.v.i., Kralovopolska 135, CZ-612 65 Brno, Czech Republic.

E-mail address: kubalal@ibp.cz (L. Kubala).

¹ These authors have contributed equally to this work and share the first authorship

cytokine interferon γ (IFN- γ) is produced upon successful differentiation that is maintained by the signal transducer and activator of transcription (STAT) 1 [3] and STAT4 signaling [4]. Furthermore, the signaling pathway of the mammalian target of rapamycin (mTOR)/Akt leads to the phosphorylation of the master regulator of Th1 differentiation, T-box transcription factor (T-bet) [5].

The Th2 cells play an important role in regulation of the immune response against extracellular parasites and inducing the humoral immune response that produces interleukin (IL-) 4, IL-5, and IL-13. The differentiation process is associated with their main differentiation cytokine, IL-4, and its master regulator, GATA-binding protein 3 (GATA3). Regarding signaling pathways that facilitate Th2 cell differentiation, STAT6 [6] and STAT5 ([7], p.) were found to be crucial protagonists in this process. During their differentiation, the Th2 cells benefit from a positive feedback loop that is orchestrated through IL-4. Since the IL-4 functions not only as the main Th2-specific effector cytokine but also as the Th2-specific differentiation cytokine, it was found that the Th2 cells enhance their differentiation by strong machinery of IL-4 production [8].

The Th17 subset is another proinflammatory cell type that orchestrates immune response against extracellular parasites and fungi. It is characteristic of its proinflammatory potential induced via the production of effector cytokines IL-17, IL-21, and IL-22. Th17 differentiation depends on cytokine stimuli, such as IL-6, IL-1 β , IL-23, and transforming growth factor β (TGF- β) [9]. Another key player in the Th17 differentiation is STAT3 [10] that facilitates the expression of their master regulator, RAR-related orphan receptor γ (ROR γ t) [11]. ROR γ t plays an essential role in the production of Th17 effector cytokines and reaches functionality upon phosphorylation by S6 kinase 2, as a part of the mTOR/Akt signaling pathway [12].

The Treg subset has an essential role in the suppression of immune responses and maintenance of immune tolerance in the organism. It comprises natural Treg cells (nTreg) and induced Treg cells (iTreg). Their lineage commitment is determined by a response to transforming growth factor β (TGF- β) and retinoic acid-abundant microenvironment. Therefore, the essential cytokines produced by Treg are the anti-inflammatory IL-10, IL-35, and TGF- β [13]. Regarding the Th17 and Treg cell subsets, differentiation plasticity is a frequently discussed phenomenon. Very recent studies confirmed the hypothesis considering TGF- β cooperation with other cytokines as a crucial factor in the cell differentiation axis. Whereas in the Treg differentiation program increased levels of TGF- β in combination with IL-2 expression are required, low levels of TGF- β in association with IL-6, IL-1 β , and IL-23 are crucial for triggering Th17 lineage commitment [9].

Importantly, the balance between the differentiation of the Th cells into distinct subsets is crucial for the maintenance of the physiological state of a host organism. Recent studies highlighted roles of different Th cell subsets in the progression of various pathologies, such as Th1 cells in lupus erythematosus [14], Th2 cells in asthma and allergies [15], Th17 cells in autoimmune diseases [16] and the role of Treg subset within the progression of malignant tumors [17]. Thus, the Th cell differentiation has become a frequented therapeutic target in the treatment of such pathologies, which brings the Th cells into the spotlight in highly sophisticated analyses, e. g. scRNASeq. Therefore, conventional immunological identification of the main Th cell subsets is being pushed more into the background as many publications point out the importance of signature genes as more meaningful information [18].

In our study, we analyzed Th cells from the peripheral blood of two healthy donors. Based on our scRNASeq analysis, we first distinguished Th cell subsets in relation to their gene expression profiles. We also elucidated their differentiation mechanisms and plasticity between the chosen Th cell subsets – Th1, Th2, and Th17. Finally, we focused more closely on the methodological part of Th cell differentiation analysis with an emphasis on discussing the most crucial steps along the scRNASeq data analysis pipeline.

2. Material and methods

2.1. Cell isolation

Cells were isolated from a buffy coat by additional purification of leukocytes using the solution of 1 % dextran from *Leuconostoc* sp. (Sigma-Aldrich) for 45 min. Next, gradient centrifugation on Histopaque-1077 (Sigma-Aldrich) was performed to isolate peripheral mononuclear cells (PBMCs). Fraction of PBMCs underwent hemolysis step by sterile H₂O to prevent residual erythrocyte contamination of the sample. PBMCs were then collected in Roswell Park Memorial Institute 1640 Medium with GlutaMAX™ Supplement (RPMI, Gibco). The separation of monocytes from the fraction of lymphocytes was performed by PBMC incubation for 30 min at 37 °C in Petri dishes, to achieve adhesion of monocytes onto plastic and their depletion from the PBMC suspension.

Lymphocytes were resuspended in RPMI with 10 % fetal bovine serum (FBS, BioTech), optimized amount of fluorescently labelled antibodies was added (0.5 μ l of antibody per 20 million cells), incubated on ice for 30 min, washed in PBS, and filtered through 70 μ m filter (Süd-Laborbedarf GmbH). Markers used for sorting Th cells were Pacific Blue™ anti-human CD4, phycoerythrin/cyanine7 (PE-Cy7) anti-human CD25, and phycoerythrin (PE) anti-human CD45RA (all from SONY Biotechnology). Th cells were sorted using BD FACSAria II (BD Bioscience) based on their surface marker constitution CD4⁺CD25⁻CD45RA⁺.

2.2. In vitro stimulation

Sorted cells were used for a set of treatment procedures that were carried out with the aim to activate the Th cells as well as to polarize them into distinct lineages (Th1, Th2 and Th17). To achieve that, 6-well plates coated with 1 ml of PBS with the addition of 1 μ g/ml anti-CD3 antibody (ExBio) were used. The density of cells on cell culture dishes was aimed at 4–5 \times 10⁵ cells/ml. RPMI medium used for cultivation of cytokine-stimulated Th cells contained 10 % FBS and 1 % penicillin-streptomycin (Gibco). The activation of the Th cells was performed upon contact of the cells with coated anti-CD3 and 10 ng/ml of soluble anti-CD28 (ExBio) that lasted 30 min. Subsequently, the activated cells were incubated for 5 days in presence of following differentiation cytokines: 1) IL-12 (25 ng/ml, PeproTech) and neutralizing antibody anti-IL-4 (1 μ g/ml, eBioscience) for Th1; 2) IL-4 (10 ng/ml, PeproTech) and neutralizing antibody anti-IFN- γ (1 μ g/ml, eBioscience) for Th2; 3) IL-6 (25 ng/ml), IL-23 (25 ng/ml), IL-1 β (25 ng/ml), TGF- β (0.25 ng/ml, all PeproTech) and neutralizing antibodies anti-IFN- γ (1 μ g/ml, eBioscience) and anti-IL-4 (5 μ g/ml, eBioscience) for Th17 differentiation.

2.3. Sample preparation for 10 \times chromium controller encapsulation

This scRNASeq data analysis is based on Th cells isolated from 2 healthy donors. Both populations of the sorted Th cells underwent the same workflow in terms of isolation, subset-specific cytokine polarization, cultivation, sample preparation (including labelling with barcoded TotalSeq-B antibodies, BioLegend) that allowed multiplexing of different samples into one scRNASeq run, sequencing, and data analysis. From both donors, a population of 10,000 cells was aimed to be encapsulated via 10 \times Chromium Controller. This number of cells interspersed between samples of activated Th cells (ACT_1), Th2 cells (TH2_1), and Th17 cells (TH17_1) in the case of donor_1, whereas in the case of donor_2, the number of total encapsulated cells split between samples of non-treated Th cells (NT-control), activated Th cells (ACT_2), Th1 cells (TH1_2), Th2 cells (TH2_2), and Th17 cells (TH17_2). The portfolio of samples during our study varied due to limited financial resources and the accessibility of the method.

Upon 5-day cultivation, all samples were collected into separate test tubes and resuspended in cold PBS with 3 % FBS and 18.2 % of Blocking Buffer (BioLegend). The cells underwent 10 min incubation on ice and

Table 1

Barcode sequences of TotalSeq-B antibodies used for labelling of distinct samples during scRNASeq sample preparation workflow (includes data from two scRNASeq runs that were held separately).

Sample	#	Product ID	Barcode sequence
NT-control	1	TotalSeq-B0251 anti-human Hashtag	GTCAACTCTTTAGCG
ACT_1	6	TotalSeq-B0256 anti-human Hashtag	GGTTGCCAGATGTCA
ACT_2	2	TotalSeq-B0252 anti-human Hashtag	TGATGGCCTATTGGG
TH1_2	4	TotalSeq-B0254 anti-human Hashtag	AGTAAGTTCAGCGTA
TH2_1	2	TotalSeq-B0252 anti-human Hashtag	TGATGGCCTATTGGG
TH2_2	6	TotalSeq-B0256 anti-human Hashtag	GGTTGCCAGATGTCA
TH17_1	4	TotalSeq-B0254 anti-human Hashtag	AGTAAGTTCAGCGTA
TH17_2	8	TotalSeq-B0258 anti-human Hashtag	CTCCTCTGCAATTAC

were labelled with TotalSeq-B anti-human antibodies conjugated with unique barcode (Hashtag oligos - HTO) that had been previously diluted in PBS with 3 % FBS in ratio 1:24 (Table 1). Next, the cells were incubated in presence of TotalSeq-B anti-human antibodies on ice for 45 min, washed 3 times in the solution of PBS with 3 % FBS, and the concentration of separate samples was measured. Finally, the cells from distinct samples were pooled together in equal ratios, creating a final mix ready for loading onto 10X Chromium Controller.

The collected samples were processed in two separate scRNASeq runs. Prior to loading on 10X Chromium Controller, the concentration of each final mix was estimated once again (700 cells/ μ l) and the sample of cell suspension was mixed with nuclease-free water (ThermoFisher Scientific) and master mix (10X Genomics). Gel Bead-In Emulsions (GEM) suspension was collected and reverse transcription was performed. The output of the reverse transcription was processed, homogenized and washed. Such suspension underwent the cDNA amplification step (in C1000 Touch Thermal Cycler by BioRad) and was divided into the 3' gene expression library and cell surface protein library that were handled separately during library preparation.

2.4. Library preparation and sequencing

The entire process of library preparation was performed according to the Chromium Next GEM Single Cell 3' Reagent Kit v3.1 User Guide with Feature Barcoding technology for Cell Surface Protein (10X Genomics). Library construction quality control revealed optimal parameters for both constructed libraries. Sequencing was carried out via the NextSeq 500/550 platform (Illumina) with the usage of NextSeq 500/550 High Output Kit v2.5 for 75 cycles (Illumina). During sequencing, 28 cycles belonged to read 1, 55 cycles to read 2 and 8 cycles to library index i7.

2.5. Data processing and analysis

Both runs of scRNASeq data were pre-processed using a standardized protocol from the Cell Ranger Single-Cell Software Suite (v3.1.0; 10X Genomics) [18] (G. X. Y. Zheng et al., 2017a). Upon the first quality control (QC), each read was assigned to cells and aligned to the reference, i.e., hg38 assembly of the human genome (GRCh38-3.0.0). For each run, the gene expression was quantified using the UMI counts in specific cell barcode sequences. The final overall fraction of cells was approximately 90 % of the targeted cell count.

The pre-processed data from the Cell Ranger [18] were imported into RStudio (v1.4.17) and analyzed using the Seurat [46] package (v4.0.3). The first step included control of the presence of Hashtag oligos. To each cell feature barcode, a sample label corresponding to our experimental

group was assigned. The same approach was used for both of our datasets. The Seurat object was created using two filtration parameters. First, all genes expressed in less than three cells were excluded from the analysis. Second, cells bearing a low number of expressed genes were excluded to avoid the dead cells in our analysis. The threshold was based on the distribution of the number of unique genes for each dataset. Both runs were joined using the function embedded in the Seurat package. Data revealed no obvious batch effect because of the similarity of pre-processing and sample preparation. Final figures were created using GGplot2 package that is compatible with Seurat.

The HTO assay was normalized using a default CLR normalization method. After that, data were demultiplexed by HTODemux function with default parameters of 0.99 positive quantiles. Identified negative cells and doublets were excluded from the analysis. The fraction of cells classified as singlets was approximately 80 % of the overall cell count. The PercentageFeatureSet function was used to calculate the percentage of mitochondrial gene transcripts for each cell in the dataset. After QC, we filtered out all cells with >20 % of mitochondrial gene transcripts in the sequenced transcriptome, which was due to positive correlation of high percentage of mitochondrial gene transcripts and potential apoptotic processes ongoing within the cell. The minimum threshold for mitochondrial gene transcripts was not used because of possible biological bias. For both datasets, after filtering out the cells with high levels of mitochondrial transcripts, viable singlet cells formed around 65 % of the raw cell count. Then, we identified outliers on a mean variability plot. The number of gene threshold was set to 2000 per cell, as close as possible to the observed median values for each dataset. In the following analytic step, we reduced our datasets to only protein-coding genes, using the Biostrings-based genome data package [19] (v1.58.0) and genome annotated sequences for Homo sapiens (BSgenome.Hsapiens.NCBI.GRCh38) provided by NCBI [20] (GenBank accession number: GCA_000001405.28; 2019-02-28).

SCTransform [21] algorithm was used to normalize and scale the data. The default parameters were changed according to observed values in QC (Fig. 1) and the number of cells parameter was set to the highest quantile observed in the dataset. Principal component analysis (PCA) [22] with default parameters was run to reduce dimensionality on transformed data. Based on PCA, uniform manifold approximation and projection for dimension reduction (UMAP) [23] was performed. The proximity of cells was analyzed using expression profiles as established by the FindNeighbors function, where 50 dimensions were used to compute neighborhood overlaps via sharing nearest neighbor (SNN) algorithm [91]. Clusters were identified using the resolution threshold of 0.7 to identify distinct cell phenotypes. Visualized UMAP graphs were adjusted by the GGplot2 package [24].

Using the Seurat FindMarkers function, the most differentially expressed genes were identified in each cluster (in comparison to the background, i.e., the mean expression of the dataset), and thus, we were able to distinguish artificially formed clusters from biologically relevant clusters in the dataset. Chosen subsets and clusters underwent a differential expression (DE) analysis using Model-based Analysis of Single Cell Transcriptomics package (MAST) ([25], v1.16.0). Setting up the positive/negative change in expression for a chosen subset was performed using the "only.pos" parameter. The X-fold difference between compared subsets was set to logFC = 0.2 to see false positives.

2.6. Differential expression analysis

The 30 most differentially expressed genes for each subset were visualized by the adjusted Heatmap function embedded in Seurat with the GGplot2 package and final figures were organized by the pubR package [26]. Figures of density plots visualized the most specifically expressed genes (markers) for each identified cluster and each subset labelled by a cell feature barcode. Final figures were created using the Nebulosa package by the wkde method [27] and the GGplot2 package [24]. The last step was the enrichment of genes using the enrichR

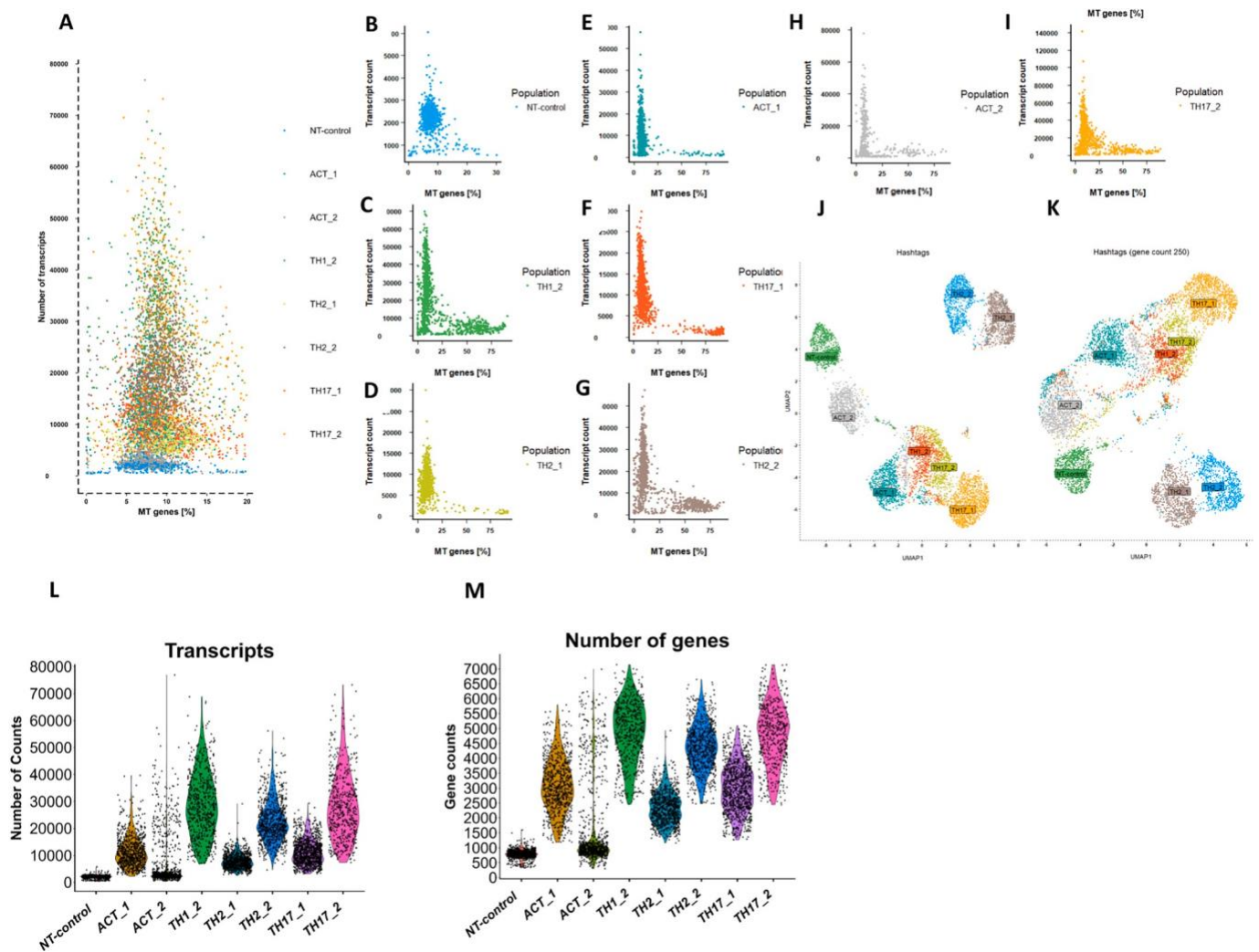


Fig. 1. Primary processing and filtration of single-cell RNA sequencing (scRNASeq) data from T cells.

A–I) Dim plots displaying the number of transcripts originating from mitochondrial genes, in bulk visualization (as shown in A) and separately for each cluster in the dataset (as shown in B–I). Colors and labels correspond to clusters separated according to the cell feature barcodes.

J) UMAP of scRNASeq data from differentiating T cells. The dataset was filtered according to the separately estimated gene count thresholds (>250 gene counts per cell for NT-control and ACT_2, >1200 gene counts per cell for ACT_1, TH2_1 and TH17_1, and >2500 gene counts per cell for TH1_2, TH2_2 and TH17_2). Colors and labels correspond to separate clusters distinguished according to the cell feature barcodes.

K) UMAP of scRNASeq data from differentiating T cells. The dataset was filtered according to the chosen gene count threshold (>250 gene counts per cell). Colors and labels correspond to separate clusters distinguished according to the cell feature barcodes.

L) Distribution of numbers of transcripts per cell in distinct clusters upon filtration of the dataset by the thresholds fitted to the groups of samples due to their activity in gene expression (>250 gene counts per cell for NT-control and ACT_2, >1200 gene counts per cell for ACT_1, TH2_1 and TH17_1, and >2500 gene counts per cell for TH1_2, TH2_2 and TH17_2). Colors and labels correspond to cell feature barcodes.

M) Distribution of numbers of gene counts per cell in distinct clusters upon filtration of the dataset by the thresholds fitted to the groups of samples due to their activity in gene expression (>250 gene counts per cell for NT-control and ACT_2, >1200 gene counts per cell for ACT_1, TH2_1 and TH17_1, and >2500 gene counts per cell for TH1_2, TH2_2 and TH17_2). Colors and labels correspond to cell feature barcodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

package [28] (data not shown) and heatmap of shared transcription factors formed by the *Dorothea* package [29].

2.7. Trajectory analysis

Upon successfully verifying the identities of distinct clusters in the dataset, we used results from the DE analysis to find and propose novel signature genes characteristics for each of the present Th cell subsets – Th1, Th2, and Th17. To perform a trajectory analysis, we used the *monocle3* package [30] and *garnett* [31]. Based on the set of genes identified during the analysis, we decided to choose most promising genes not in terms of upregulation but specificity. The expression of these genes in particular Th cell phenotypes was verified by the RT-qPCR

analysis.

2.8. ELISA

Supernatants from each sample was collected. Effector cytokines IFN- γ , IL-13, and IL-17A were detected by the IFN gamma Human Uncoated ELISA Kit, IL-13 Human Uncoated ELISA, and IL-17A Human Uncoated ELISA Kit (all *ThermoFisher Scientific, Waltham, MA, USA*), respectively, according to the manufacturer's instructions, employing the Sunrise microplate reader (*Tecan, Zürich, Switzerland*).

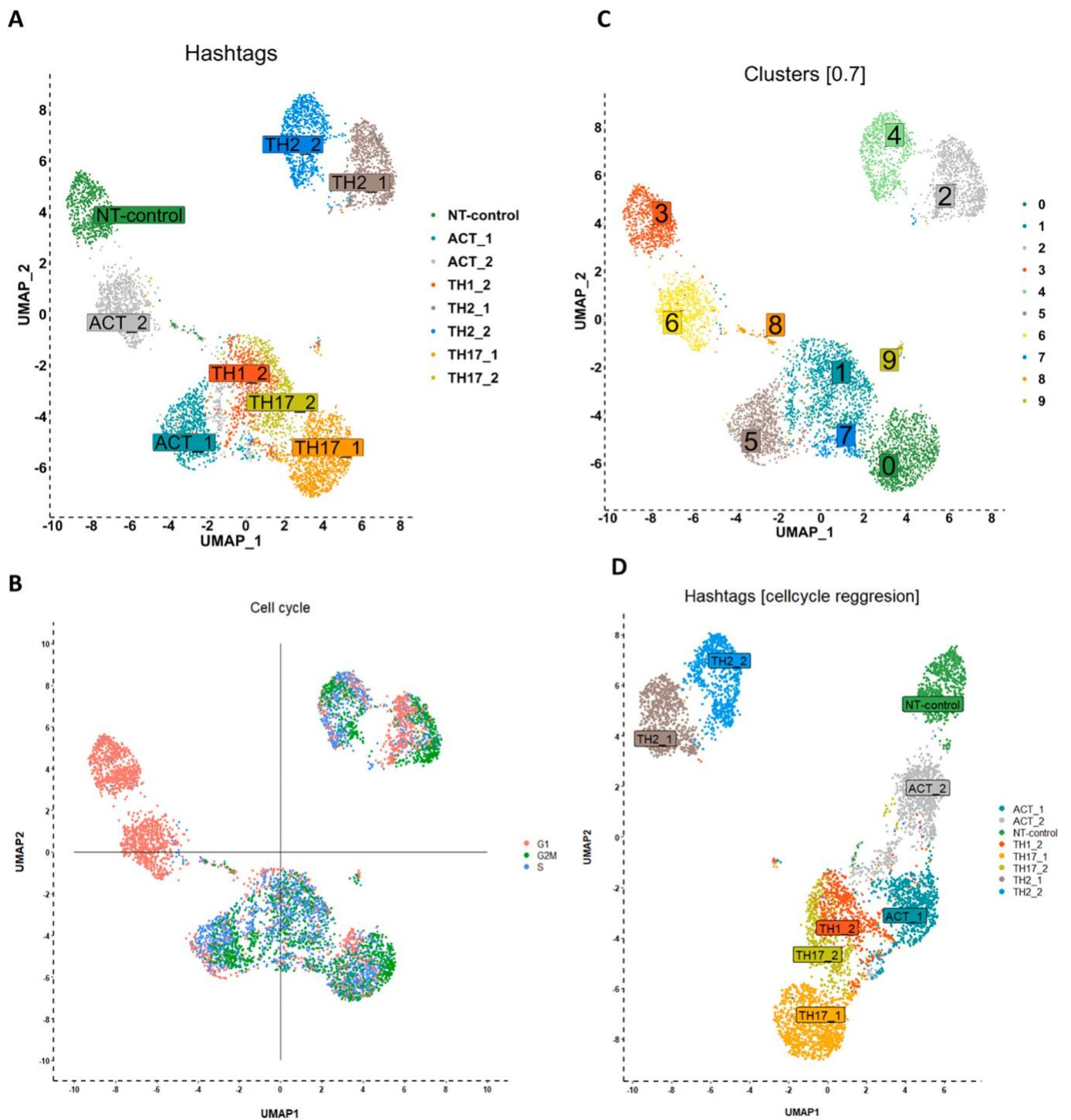


Fig. 2. Analysis of the impact of cell cycle phase on clustering of the dataset.

A) UMAP of scRNASeq data from Th cells. Colors and labels represent cells in clusters corresponding to the particular cell feature barcodes that were used for multiplexing during the sample preparation.

B) UMAP of scRNASeq data of differentiating Th cells. Colors represent cells assigned to the G1 phase (red), S phase (blue), and G2/M phase (green) of the cell cycle.

C) UMAP of scRNASeq data from Th cells. Colors represent cells in the ten clusters defined using top variable genes and unsupervised clustering. Labelling represents the numbering assigned to the clusters distinguished by the FindNeighbours function (Seurat) under resolution = 0.7.

D) UMAP of scRNASeq data from differentiating Th cells. Comparison between the dataset without performing the cell cycle regression step (A) and the dataset that underwent cell cycle regression (D). Colors and labels correspond to the separate clusters distinguished according to the cell feature barcodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.9. Gene expression analysis

Total RNA was isolated from cells using the Cell/Tissue RNA Kit (CatchGene, New Taipei City, Taiwan). Isolated RNA was reversely transcribed into cDNA using the Reverse Transcription Kit according to the manufacturer's protocol (GeneriBiotech, Hradec Kralove, Czech Republic). Real-Time quantitative PCR reaction was performed in a LightCycler480 instrument (Roche, Basel, Switzerland) using Universal Probe Library and a LightCycler480 Probe Master (both Roche, Basel, Switzerland) with the following program: initial step at 95 °C for 15 min (denaturation), followed by 45 cycles at 95 °C for 15 s and 60 °C for 1 min (amplification) and a final step at 40 °C for 1 min. Specific primers for signature genes were used and data were normalized to *RPL13A* mRNA presented as $2^{-\Delta\Delta C_t}$. The primers were designed to detect all splicing variants.

3. Results and discussion

3.1. Study design

This study is based on Th cells isolated from peripheral blood of healthy donors activated with anti-CD3 and anti-CD28 for 30 min and differentiated using subset-specific cytokines (IL-12 for Th1, IL-4 for Th2, and combination of IL-6, IL-1 β , IL-23, and TGF- β for Th17). Treated cells were cultivated for 5 days and then collected and processed for scRNASeq (Supplementary Fig. 1).

The data analysis aimed to cover both Th subset heterogeneity and biological repeatability. Therefore, the final dataset was created to link two separate scRNASeq runs based on the same treatment of Th cells originating from two different donors. In the first run, the samples of activated Th cells (ACT_1), Th2 cells (TH2_1), and Th17 cells (TH17_1) were included, and in the second run, the samples of non-treated Th cells (NT-control), activated Th cells (ACT_2), Th1 cells (TH1_2), Th2 cells (TH2_2) and Th17 cells (TH17_2) were included. The portfolio of samples during our study varied due to limited financial resources and the accessibility of the method. All samples underwent all stages of sample preparation, Illumina sequencing, and data analysis in the same way. The observed heterogeneity in terms of genes per cell and counts per cell may appear due to different transcriptional activity of non-treated and activated cells in comparison to the cells undergoing the process of differentiation and induction of their effector functions.

3.2. Selection of viable cells for further analysis

The first step in scRNASeq analysis consisted of dataset QC. Generally, a pointer for distinguishing viable and dead cells during the QC is the percentage of mitochondrial genes expressed in the distinct cells. Several publications highlighted the importance of the specification of the thresholds for distinct cell types [32,33]. To estimate the right balance between excluding non-viable cells from the dataset and preserving significant gene expression profiles, generating a plot visualization of mitochondrial gene transcripts distribution in the dataset is recommended. The threshold for Th cells depends on treatment design and protocol [34,35]. Based on current literature [36,37], the expected percentage of mitochondrial gene transcripts in viable cells lies within the range of 5–15 %. The upper limit is considered the main parameter since higher levels of mRNA transcripts originating from mitochondria are often considered a marker of ongoing apoptosis [38,39]. However, an out-of-range percentage of mitochondrial mRNA transcripts could reflect a technological bias. We observed substantial heterogeneity of mitochondrial gene expression in different Th cell subsets (Fig. 1A–I). This observation led to an assumption that the threshold mentioned above for filtration of mitochondrial mRNA transcripts (5–15 %) might not be suitable for filtration of viable cells in different Th cell subsets. As shown, the suggested range would be applicable to ACT_1 and ACT_2, TH17_1, and TH2_1, and NT-control (Fig. 1B, D, E, F, and H) samples. However, due to apparent alterations in the percentages of

mitochondrial genes in TH1_2, TH17_2, and TH2_2 (Fig. 1C, G, and I) during the QC, we applied the top threshold 20 % to avoid losing more cells than necessary.

Regarding the bottom threshold, cells not expressing mitochondrial genes could be considered a pointer to a specific cell cycle phase [40]. Moreover, there is always a chance that cells could be metabolically resting [40,41]. On the contrary, the more apoptotic cells remain in the dataset, the more redundant and ambiguous output may be received [32,42]. Interestingly, the number of mitochondrial gene transcripts present in the PCA in downstream data analysis steps can also be used to double-check for the adequacy of the chosen threshold.

The threshold for gene count filtration can be estimated as a default parameter or for each sample separately. Our dataset comprised a very heterogeneous population of cells (especially when pooling fully differentiated Th cells into one dataset) (Fig. 1L and M). When comparing the final outputs of clustering based on a dataset with default gene count threshold (i.e., minimum of 250 genes for all samples) (Fig. 1K) and the dataset with the gene count threshold fitted specifically to distinct samples (250 for NT-control and ACT_2, 1200 for ACT_1, and 2500 for TH1_2, TH2_2, and TH17_2) (Fig. 1J), we observed substantial differences. Applying the first filtration variant created interconnections between clusters of interest. In some cases, such clustering even produced clusters unrelated to Th cell differentiation only based on the expression of translation-associated genes. On the contrary, the second filtration variant gave rise to a more straightforward clustering and minimized the number of clusters created independently from differentiation-associated gene expression profile similarities. Considering the main aim of this study, the dataset chosen for further analysis was the one that underwent filtration at multiple gene count thresholds (Fig. 1J).

The lower gene count similarity of cells within clusters might illustrate the biological process of differentiation more realistically. On the other hand, the dataset based on filtration with separately fitted gene count thresholds may obliterate essential signals of biological phenomena. However, this variant may lead to a more straightforward identification of various cell populations and a more simplistic characterization of their traits (transcriptional profiling) (Fig. 1J and K), and it is necessary in the case of different quality of distinct runs of sequencing. In conclusion, both ways of dataset processing can be highly beneficial for the correct fitting of data analysis to a specific biological question [43,44].

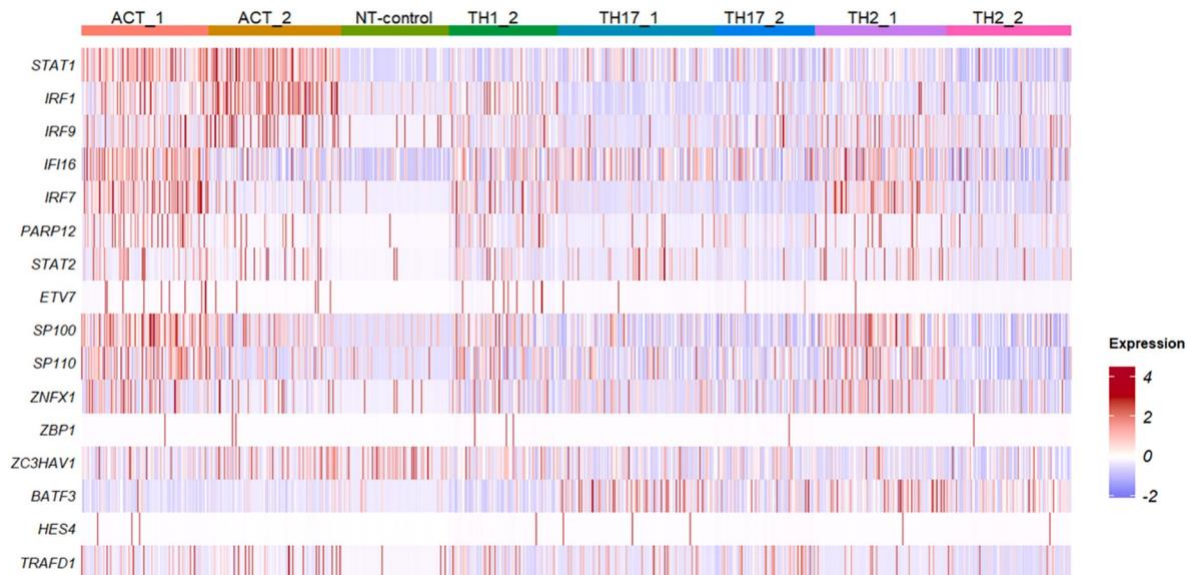
3.3. Cell cycle scoring and filtration of translation-associated genes

Cell cycle regression of scRNASeq data is an important step to consider during every data analysis [41,43]. It provides detailed information on what phase the cells were captured at the point of encapsulation. Hence, it can also distort the clustering of the dataset. During scRNASeq analysis, identification of the cell cycle phase can be achieved via a scoring step assigning a specific cell cycle phase to each cell according to the expression of cell cycle-associated genes [44,45]. This scoring step can be followed by normalization via SCTransform as it is performed as the crucial part of the scRNASeq analysis in the *Seurat* package [46,47]. However, the impact of cell cycle genes on final clustering should always be considered with respect to the dataset and type of treatment [48].

In our study, we performed cell cycle regression and compared the filtered dataset to the initial dataset that was not influenced by this processing step (Fig. 2A and D). The filtered dataset was found to produce a highly similar output to the unfiltered dataset in terms of clustering and composition of principal components (Fig. 2C and D). Thus, we decided to continue with the analysis and characterization of clustered data based on the unfiltered dataset, i.e. including the expression of genes related to the cell cycle.

Knowing the phenotype of analyzed cells is essential for a decision whether to perform the cell cycle filtration step or not. In this case, Th

A



B

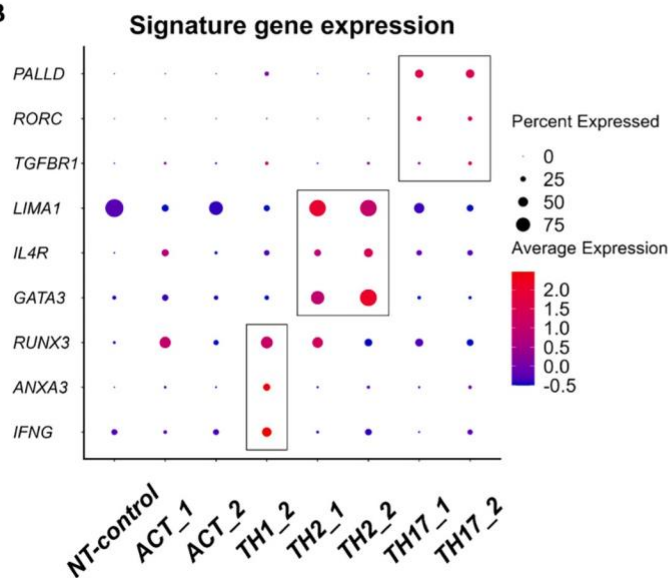


Fig. 3. Cluster analysis and Th subset characterization using differential expression.

A) Heatmap based on the top differentially expressed transcription factors in activated Th cells (ACT_1, ACT_2). The X axis represents ordered cells, assigned to distinct clusters created according to cell feature barcode classification (HTO classification). The colors represent the scaled (Z-scored) gene expression for each selected gene in each cell. The scale is displayed in the right color bar.

B) Dotplot of artificial signature genes from differentiating Th cells. Colors correspond to the median expression level in expressing cells and a circle size corresponds to the proportion of expressing cells (as shown in the legend on the right). Columns represent clusters separated based on their cell feature barcodes, rows represent selected signature genes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cells are known for their rapid proliferation and gene expression in response to activation [34,35]. Thus, the cell cycle may be an essential marker for distinguishing the G0-phase (resting) Th cells from their proliferating counterparts in various cell cycle phases. Cell cycle status reproduce a biological effect in the case of activation and differentiation of Th cells in vitro (Fig. 2B and D). Here, the identification of expression specific for cell cycle progression provided the initial proof of Th cell proliferation in our dataset. Therefore, its regression might delete important information and even mislead subsequent analyses of distinct Th subsets. In contrast to our Th cell differentiation in vitro, in the case

of the experiment with cultivated cell lines (e.g. stabilized cell lines) that are meant to be synchronized in the cell cycle, a substantial deviation in the expression of cell cycle-related genes should be addressed as a batch effect [42].

Another filtration parameter considered before cell clustering was the analysis of translation-associated genes, i.e. genes participating in translation mechanisms and ribosomal genes [49,50], starting with a pattern *RPL-/RPS-/EEF-* [51]. During the analysis, we noticed that the genes explaining the variability of the first three principal components included just a few ribosomal genes starting with *RPL-* or *RPS-* pattern,

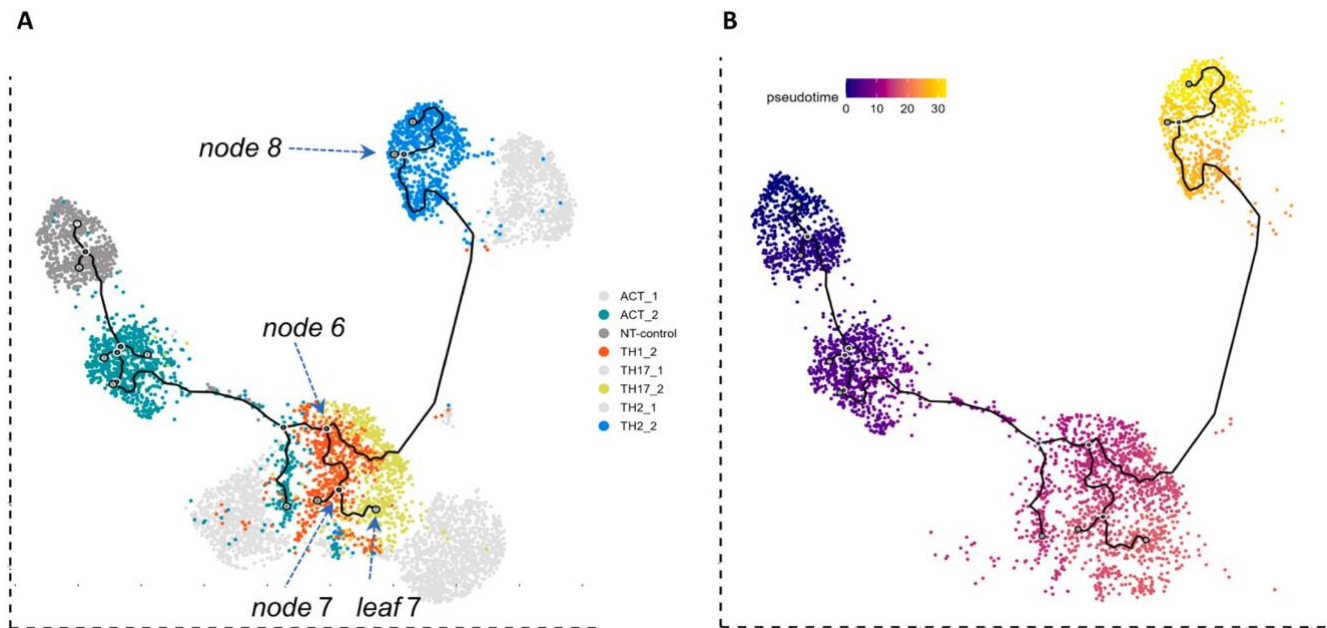


Fig. 4. Trajectory analysis.

A) Trajectory analysis of samples from donor_2. Colors and labels represent cells in clusters corresponding to the particular cell feature barcode. Black corresponds to trajectory pathway computed by *monocle3*. Grey clusters correspond to samples from donor_1 and were not used in trajectory analysis. Arrows highlights most important nodes along the pathway.

B) Pseudotime values measured along the trajectory pathway. Position in UMAP corresponds to the separate clusters distinguished according to the cell feature barcodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which indicated a less significant influence of cell translational activity on clustering (Supplementary Fig. 2). We assumed such clusters could be defined by non-specific gene expression patterns unrelated to Th cell differentiation. The scRNASeq community has recently emphasized the necessity to perform a DE analysis among clusters to validate previously performed QC steps [50,52]. The DE analysis serves as a control that translation-associated gene expression may prevail over other expressed genes in specific clusters, thus creating the only reason for joining specific cells into a cluster in the dataset. Also, the expression pattern of genes encoding ribosomal subunits can cause a bias in the dataset [49,50]. In both cases, the expression of translation-associated genes and cell cycle gene expression are suitable parameters for decision-making during each data analysis step [43,44].

3.4. Cluster analysis

The SNN clustering performed with resolution 0.7 produced 10 clusters, most of them corresponded to sample-specific barcodes (Fig. 2A and C). Cluster 3 corresponded to non-treated Th cells (NT-control), clusters 3 and 6 overlapped with activated Th cells (ACT_1, ACT_2). Clusters 2 and 4 overlapped with Th2 cells from both donors (TH2_1, TH2_2). Cluster 0 corresponded to Th17 cells of the first donor (TH17_1). In contrast, some clusters were created solely due to gene expression profile unity, e.g. cluster 7 represents a mixture of samples from activated Th cells, Th1 cells, and Th17 cells. Finally, cluster 1 was based on the conjunction of Th1 and Th17 cells from the second donor (TH1_2, TH17_2) with a small part of the sample belonging to activated Th cells (ACT_2). Moreover, a specific population of cells (cluster 7, Fig. 2C) displayed a selective pattern in terms of Treg cell subset-specific gene expression, which will be discussed in following section.

The activation of Th cells from both donors (ACT_1, ACT_2) was documented by the increased gene expression of classical activation markers *CD25*, *IL-2* receptor α chain (*IL-2RA*), (Supplementary Fig. 3A), and the downregulation of *CD62L*, also known as L-selectin (*SELL*) characteristic for naïve Th cells (Supplementary Fig. 3B). The DE analysis between activated Th cells from both runs (ACT_1, ACT_2) and

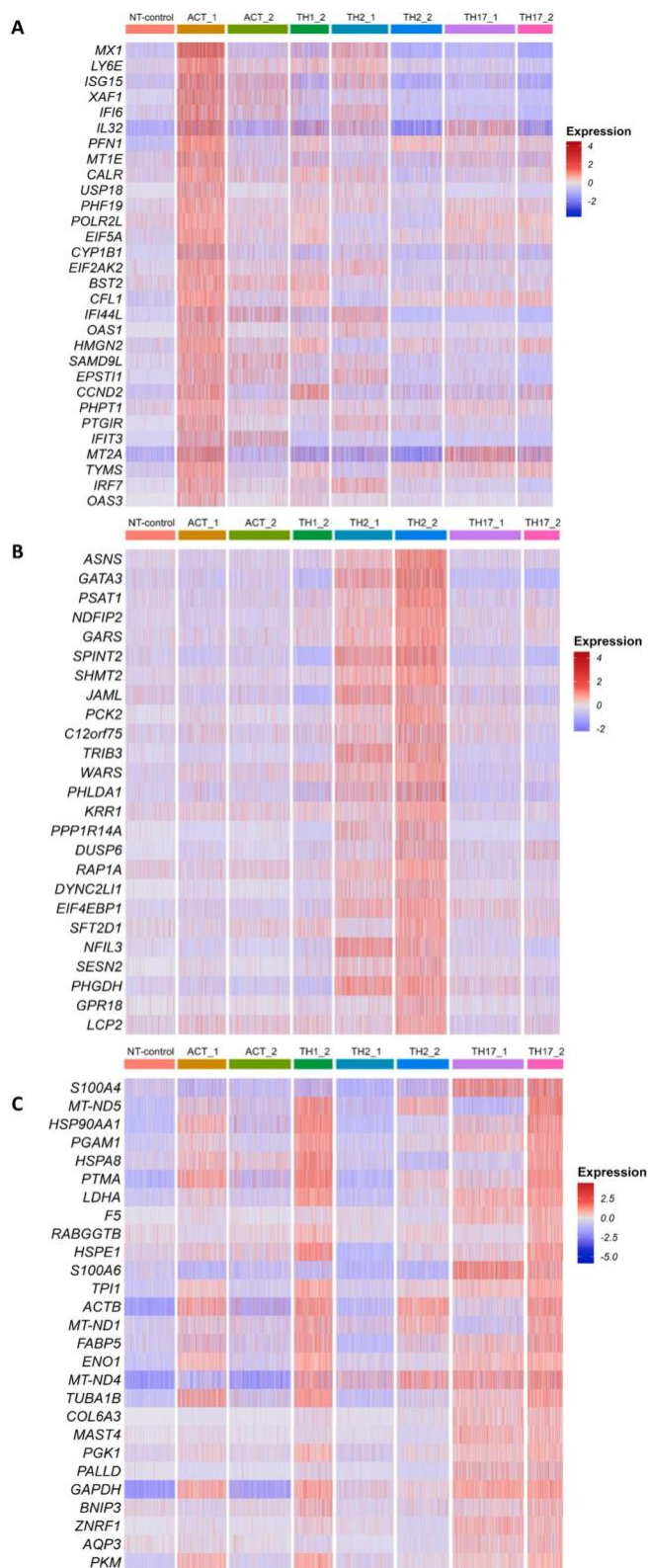
comparison to non-treated Th cells (NT-control) revealed an upregulation of transcription factors such as *STAT1*, *IRF1*, *IRF7*, *IRF9*, and *IFI16* in the case of the activated Th cells from both donors (ACT_1, ACT_2) (Fig. 3A). Multiple studies linked the expression of genes mentioned above to T cell activation and expansion, particularly with *STAT1* encoding signal transducer and transcription activator [53], *IRF*-encoding the interferon-regulatory factors [54], and *IFI16* encoding IFN- γ inducible factor 16 [55]. According to Twhig et al., *STAT1* regulates the production of inflammatory cytokines, transcription factors, and immune checkpoint regulators [53]. Moreover, *IFI16* was found to contribute to the formation of the effector T cell phenotype [55].

3.5. Heterogeneity among donors

One of the phenomena we observed was the difference of gene count of labelled cells between donors (Fig. 1L and M). As can be observed from Fig. 1K and L, the alteration between the number of counts in ACT_2 and the number of counts in NT-control from donor_2 was minor. Interestingly, the number of gene counts in ACT_1 was higher than in both controls (NT-control, ACT_2) from donor_2 (Fig. 1M). We speculated that this is due to different transcriptional activity of individual donors and thus, it could be based on the biological effect of priming. Priming is described as the first contact of an antigen-specific Th cell with an antigen that induces enhanced gene expression [56]. At the same time, gene expression profile similarity between ACT_1 and TH1_2 (Fig. 2A) was observed. One could suggest that the actual difference between the runs can be caused by the batch effect [42,44,57]. However, as we followed the same protocols in sample preparation and data analysis, we inclined to the priming theory.

3.6. T helper cell subset verification

To verify the polarization of Th cells into subsets on the level of transcriptome, a set of conventional signature genes was used [58]. The identification of Th1 cells was based on the expression of genes such as *IFNG* encoding IFN- γ (Th1 subset-specific effector cytokine), *ANXA3*



(caption on next column)

Fig. 5. Cluster analysis and Th subset characterization using differential expression.

A) Heatmap based on the top differentially expressed genes of ACT_1 subset, according to the Model-based Analysis of the Single-cell Transcriptomics (MAST) algorithm. The X axis represents the ordered cells, grouped according to their cell feature barcodes (columns). Different colors correspond to the scaled (Z-scored) expression of each selected gene in each cell, as shown in the color bar on the right.

B) Heatmap based on the top differentially expressed genes of Th2 subset, according to the MAST algorithm. The X axis represents the ordered cells, grouped according to their cell feature barcodes (columns). Different colors correspond to the scaled (Z-scored) expression of each selected gene in each cell, as shown in the color bar on the right.

C) Heatmap based on the top differentially expressed genes of Th1 and Th17 subsets, according to the MAST algorithm. The X axis represents the ordered cells, grouped according to their cell feature barcodes (columns). Different colors correspond to the scaled (Z-scored) expression of each selected gene in each cell, as shown in the color bar on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

encoding annexin 3, and *RUNX3* encoding RUNX family transcription factor 3 (Fig. 3B). As for Th2 cells, a set of conventional signature genes comprised *GATA3* (the main transcription factor and master regulator of Th2 cell differentiation), *IL4R* encoding the IL-4 receptor (receptor for the main Th2 subset-specific differentiation and effector cytokine), and *LIMA1* encoding LIM domain and actin-binding protein 1 (Fig. 3B). Finally, Th17 cells were identified using *RORC* encoding ROR γ t (the main transcription factor and master regulator of Th17 differentiation), *TGFBR1* encoding TGF- β receptor (involved in Th17 differentiation induction), and *PALLD* encoding palladin (Fig. 3B).

To validate the functionality of the differentiated cells, cytokine production was measured by ELISA. It was observed that cells polarized into the Th1 subset produced a significantly higher amount of IFN γ in comparison to non-treated control and activated Th cells. Similarly, the cells stimulated with IL-4 and thus polarized into the Th2 subset produced a significantly higher amount of IL-13 (one of the main effector cytokines of Th2 cells). Finally, cells polarized into the Th17 subset revealed the production of significantly higher levels of IL-17A, the main effector cytokine of Th17 cells (Supplementary Fig. 4).

3.7. Gene expression profiling, trajectory analysis, and analysis of novel signature genes

Due to the differences between donors, we decided to perform the trajectory analysis only on the donor_2 to preserve consistency in our dataset. Moreover, the number of analyzed subsets was higher so the use of the donor_2 provided us more variability in the terms of cell phenotype. For the expression of particular genes, we used visualization of log₁₀ gene expression on a trajectory path.

It is visible that part of the cells from ACT_2 and NT-control undergo a shift of expression closer to the cluster where the TH17_2 and TH1_2 subset is present (Fig. 4A, cluster 8 in Fig. 2C). The trajectory analysis revealed the possible routes between the Th subsets. The TH1_2 and TH17_2 subsets have already shown similar expression profiles using the differential expression analysis (Fig. 5A). Compared to visualization based on hashtags (Fig. 4A) part of the TH1_2 cells from node 7 became more similar to TH17_2, especially at the leaf 7 (Fig. 4A). This route suggests trans-differentiation potential between these two phenotypes as suggested previously in the literature [59]. The route from node 6 to node 8 lies within the border of TH17_2 and TH1_2 and also supports our result from the differential expression heatmap (Fig. 5A and C).

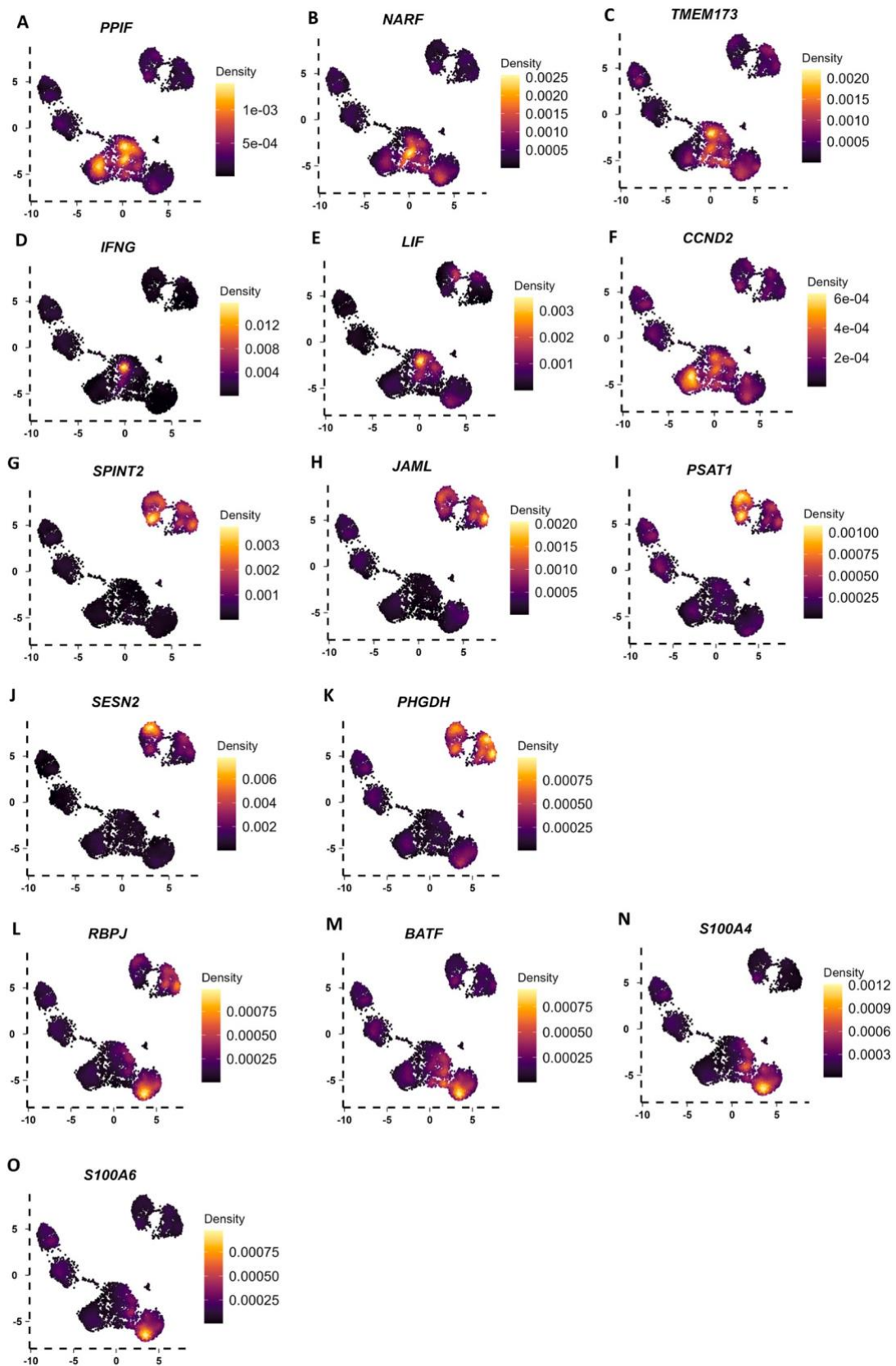


Fig. 6. Density plots highlighting the most differentially expressed genes from both donor's datasets for each Th subset. A–F) Th1-specific genes (TH1_2), G–K) Th2-specific genes (TH2_1; TH2_2), L–O) Th17-specific genes (TH17_1; TH17_2). Right color bar: a relative density scaling. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

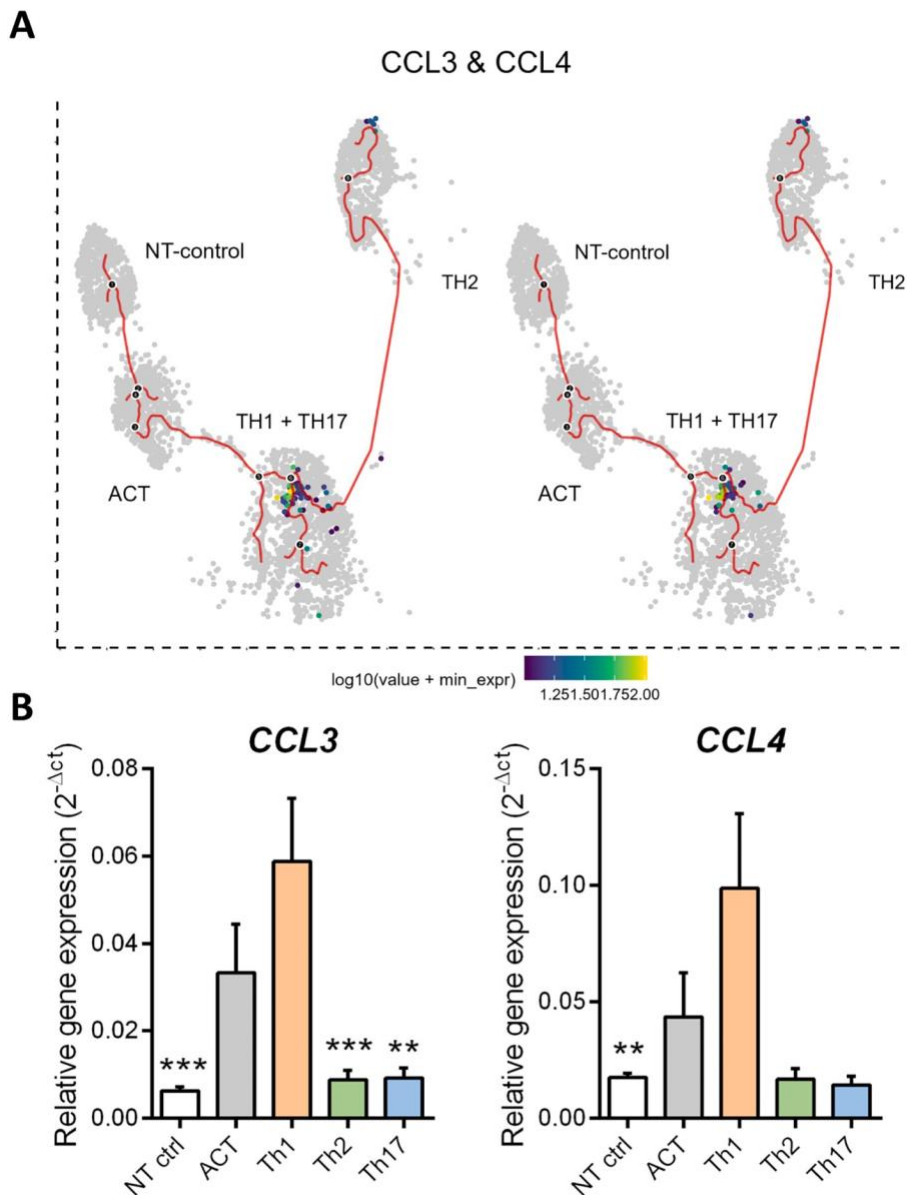


Fig. 7. Analysis of chosen signature genes for Th1 subset.

A) Visualization of chosen signature genes for Th1 cells (TH1_2). Colored dots represent cell localization on UMAP and scale based on \log_{10} expression along the pathway. Red line corresponds to trajectory pathway computed by *monocle3*. Artificial clusters are colored grey.

B) Verification of the selection of Th1 signature genes (*CCL3* and *CCL4*) based on the relative gene expression analysis covering various donors ($n = 5-6$). Statistical significance was examined by one-way ANOVA. Mean \pm SEM, ** $p < 0.01$, *** $p < 0.001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.7.1. Th1 cell subset

Besides conventional Th1-specific signature genes, the cells from ACT_1 displayed upregulated expression of several genes connected to Th1 phenotype, including *MX1* encoding MX dynamin-like GTPase 1, *LY6E* encoding the lymphocyte antigen 6 family member E, *IFI44L* encoding interferon-induced protein 44 like, and *ISG15* encoding the ISG15 ubiquitin-like modifier. The *MX1* was found necessary in differentiating the Th1 cell subset as it was assigned to one of the interferon production proteins [54]. Moreover, involvement of *IFI44L* and *LY6E* was assigned to an interferon-inducible gene interaction network ([60]; H. [61]). Finally, the role of *ISG15* in the production of IFN γ by T cells has been recently elucidated [85]. The cells from the TH1_2 expressed very specifically *PPIF*, *NARF*, *LIF*, *IFNG*, *TMEM173* and the *CCND2* gene which expression overlap with ACT_1 (Fig. 6A–F). The *PPIF* gene (Peptidyl-prolyl cis-trans isomerase) was found to be expressed in activated T cells and constitutes the central channel to balance the needs for and potential harm from reactive oxygen species ROS [62]. *NARF* (Nuclear Prelamin A Recognition Factor) was previously connected to down-regulation of WNT signaling and regulates the ubiquitylation and degradation of T cell factor/lymphoid enhancer factor family [63]. The

TMEM173 (*STING*) expression likely influences the T cell recruitment in response to a TNF- α [64]. Finally, the *LIF* gene (leukemia inhibitory factor) is a member of the interleukin-6 (IL-6) cytokine family. LIF is known for counter-regulation of Treg and Th17 development and could work as a core regulatory circuitry of T cells [65,66].

Using the trajectory analysis and *gamett*, we were able to identify additional two highly specific genes in the TH1_2 subset, *CCL3* and *CCL4* (C-C Motif Chemokine Ligand 3 and 4, Fig. 7A). The expression of both genes is limited to node 6 of the trajectory (Fig. 4A). According to the literature, both genes interact with the CCR5 receptor, which is expressed on the surface of Th1 cells [67]. Hence both *CCL* genes were not previously connected to the Th1 subset, the overall specificity value from the DE and PCR suggests a good potential to be denoted as novel signature genes for the Th1 phenotype (Fig. 7B).

3.7.2. Th2 cell subset

In the case of Th2 cells, the uniqueness of the gene expression profile separated this subset from all other analyzed samples since the DE analysis highlighted multiple genes related to the Th2 cell differentiation. These genes comprised *SPINT2* encoding serine peptidase inhibitor

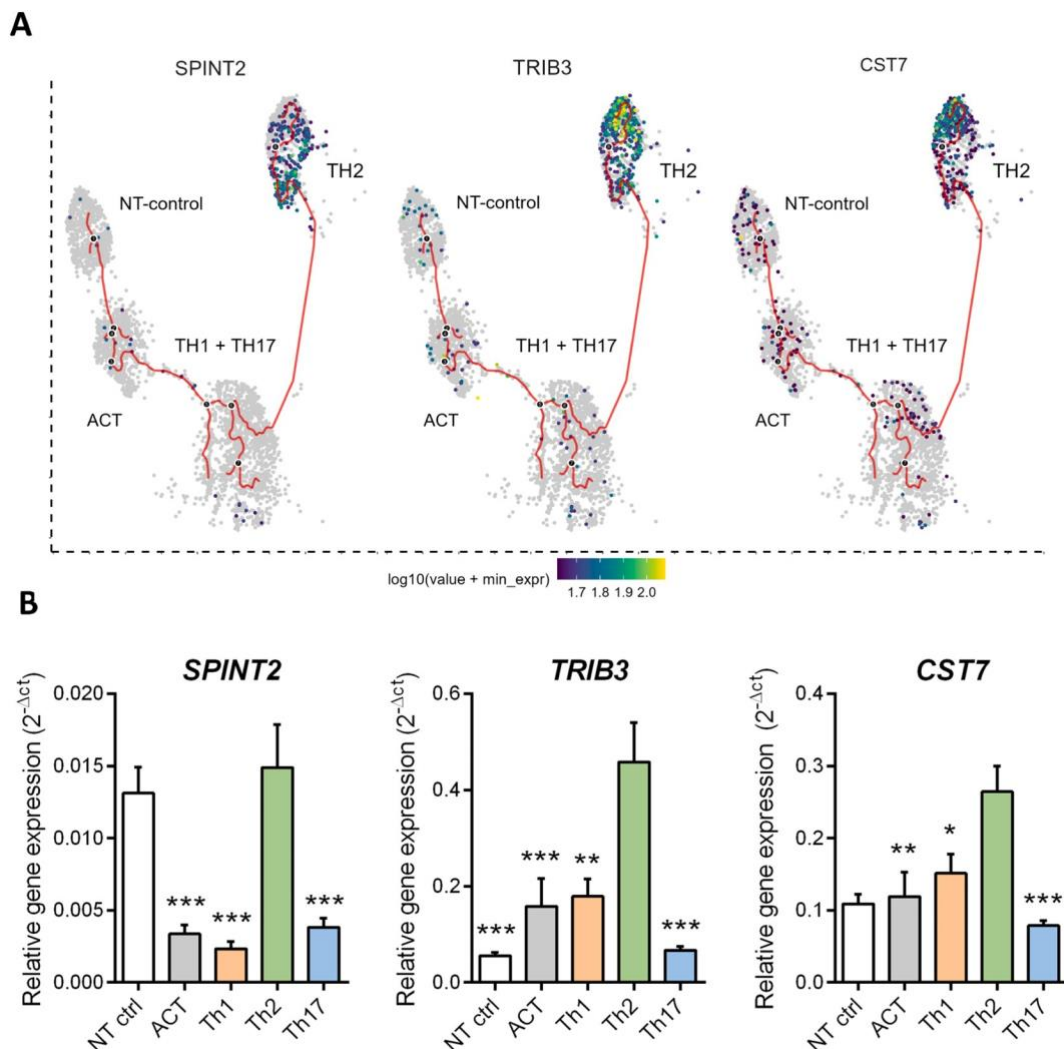


Fig. 8. Analysis of chosen signature genes for Th2 subset.

A) Visualization of chosen signature genes for Th2 cells (TH2₁; TH2₂). Colored dots represent cell localization on UMAP and scale based on log₁₀ expression along the pathway. Red line corresponds to trajectory pathway computed by *monocle3*. Artificial clusters are colored grey.

B) Verification of the selection of Th2 signature genes (*SPINT2*, *TRIB3* and *CST7*) based on the relative gene expression analysis covering various donors ($n = 6-7$). Statistical significance was examined by one-way ANOVA. Mean \pm SEM, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Kunitz type 2, *PHGDH* encoding phosphoglycerate dehydrogenase, *PSAT1* encoding phosphoserine aminotransferase 1, *JAML* encoding junction adhesion molecule-like protein, and *SESN2* encoding sestrin 2 (Figs. 5B, 6G–K).

The observation is in line with recent studies, pointing out the potential specificity of analyzed markers for Th2 cells. Th2 differentiation was previously linked to serine metabolism, indicating the importance of *SPINT2* upregulation in the formation of the Th2 phenotype [68]. Also, *SESN2* was found to support mTORC2-Akt activation [69], while the mTORC2 signaling pathway is frequently discussed in terms of explicit function in Th2 differentiation [86]. Along with these observations, another study highlighted the importance of *PHGDH* and *PSAT1* upregulation in serine biosynthesis [70]. Finally, *PSAT1* was identified as a potential suppressor of Th1 cell differentiation, which is achieved via reduction of IFN γ production [71].

Beside the high specificity of *SPINT2* gene, we were able to identify 2 more genes that were not present in other clusters of the analyzed dataset (Fig. 8A). The *TRIB3* gene (Tribbles pseudokinase 3) seems to be highly specific for the Th2 phenotype. In the literature the abundance of *TRIB1* and *TRIB3* in T cells was experimentally proven. It was previously

published that *TRIB3* reduces CD8⁺ T cell infiltration [72]. Hence the information, the *TRIB3* have not been connected to Th2 phenotype yet. The *CST7* gene (cystatin-like metastasis-associated protein) was highly upregulated in TH2₂ with notably increased specificity. The high expression of CST proteins was previously observed in T cells [73].

Based on this set of genes, we decided to choose *SPINT2*, *TRIB3* and *CST7* for the PCR analysis (Fig. 8B). Overall, the separation of Th2 cells is in line with previous studies highlighting the rigid phenotype observed in single cell sequencing analysis [74].

3.7.3. Th17 cell subset

In the context of the complete dataset, gene expression profile similarities localized Th17 cells in proximity of the Th1 cells (Fig. 5C). Despite the fact that Th17 cells shared high expression levels of multiple genes with Th1 cells, the DE analysis between Th1 and Th17 cells revealed several potential markers of Th17 cells. Th17 cell subset-specific genes included *RBPJ* encoding recombination signal binding protein for immunoglobulin kappa J region important in Notch signaling, *BATF* gene encoding basic leucine zipper ATF-like transcription factor that is activated downstream of TGF- β signaling. Another

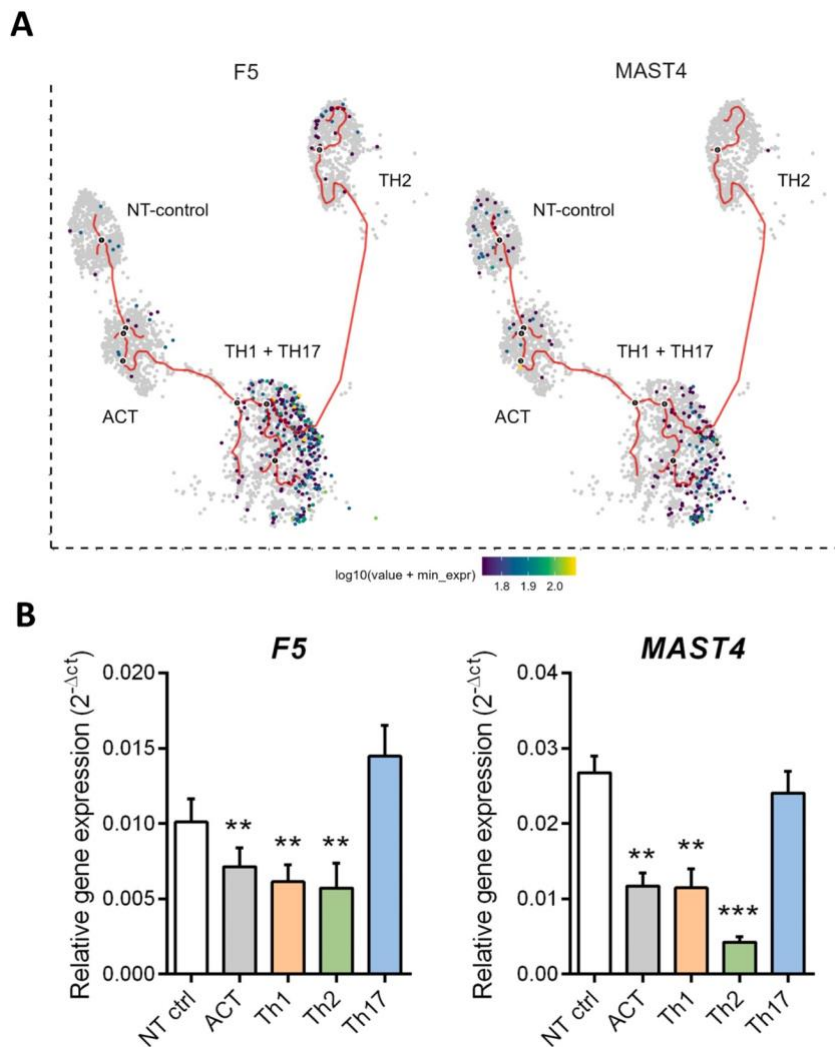


Fig. 9. Analysis of chosen signature genes for Th17 subset.

A) Visualization of chosen signature genes for Th17 cells (TH17_1; TH17_2). Colored dots represent cell localization on UMAP and scale based on \log_{10} expression along the pathway. Red line corresponds to trajectory pathway computed by *monocle3*. Artificial clusters are colored grey.

B) Verification of the selection of Th17 signature genes (F5, MAST4) based on the relative gene expression analysis covering various donors (n = 6–7). Statistical significance was examined by one-way ANOVA. Mean \pm SEM, **p < 0.01, ***p < 0.001. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

example is *S100A4* and *S100A6* encoding S100 calcium-binding proteins A4 and A6 (Figs. 5C, 6L–O). Recent studies revealed the role of *RBPJ* in the enhancement of Th17 differentiation [75]. Also, *BATF* as a transcription factor of TGF- β pathway and a direct downstream effector of Smad4 was linked to the formation of Th17 phenotype [9].

A more detailed analysis of Th1 and Th17 transcriptional similarities revealed that both Th cell subsets shared high expression levels of multiple genes, such as *PKM*, *ENO1*, *LDHA*, *PGAM1*, *PTMA*, *HSPA8* and *FABP5* (Fig. 5) that indicated similar molecular mechanisms regulating the differentiation of the Th1 and Th17 cell subsets. A recent study pointed out the formation of Th1/Th17 immunophenotype in response to *S. aureus*. This also indicated the potential plasticity between Th1 and Th17 cells, based on their similarities in the production of their effector cytokines – IFN γ and IL-17 [76]. Further, Th17 cells were shown to easily shift into the Th1 phenotype in the presence of IL-12 and/or TNF- α [77]. The TH1_2 subset forms a larger cluster together with a part of TH17_2 cells (Fig. 2C).

The difference in UMAP visualization is crucial for addressing the similarities of gene expression profiles between distinct clusters. In this case, similar gene expression patterns in earlier stages of differentiation can be associated with a trans-differentiation potential between cells in the observed clusters. The DE analysis supported this assumption, revealing substantial expression similarities between the TH1_2 and

TH17_2 cells from the same donor (Fig. 5C). As shown in Fig. 2A, a part of cluster 6 (ACT_2) went through activation and proliferation. The two different routes through TH17_2 and TH1_2 also correspond to a theory of easy trans-differentiation between these two phenotypes. Surprisingly, cluster 7 is also formed by a few cells from the TH2_2 subset.

The finding of Th17-specific genes was challenging due to their overlap with TH1 cells in our analysis. Still, we were able to identify two genes with a high specificity for the subset (Fig. 9A). The F5 gene (Coagulation Factor V) was associated with activation of T cells but the mRNA was abundant only in activated T cells not resting [78]. The second gene was MAST4 of the Serine/threonine kinases family. The MAST4 possesses a very high specificity toward the TH17_2 subset. The gene expression values were also measured on PCR to verify the specificity of these two genes (Fig. 9B).

3.7.4. Identification of cells with characteristics of regulatory T cells

The Treg peculiar expression was previously connected to genes such as *FOXP3* encoding forkhead box P3 (the main transcription factor and master regulator of Treg subset), *CTLA4* encoding cytotoxic T-lymphocyte associated protein 4, which led to a hypothesis about these cells being spontaneously differentiated Treg cells. According to the literature, all signature genes mentioned in this paragraph were considered conventional in terms of Th subset identification [58].

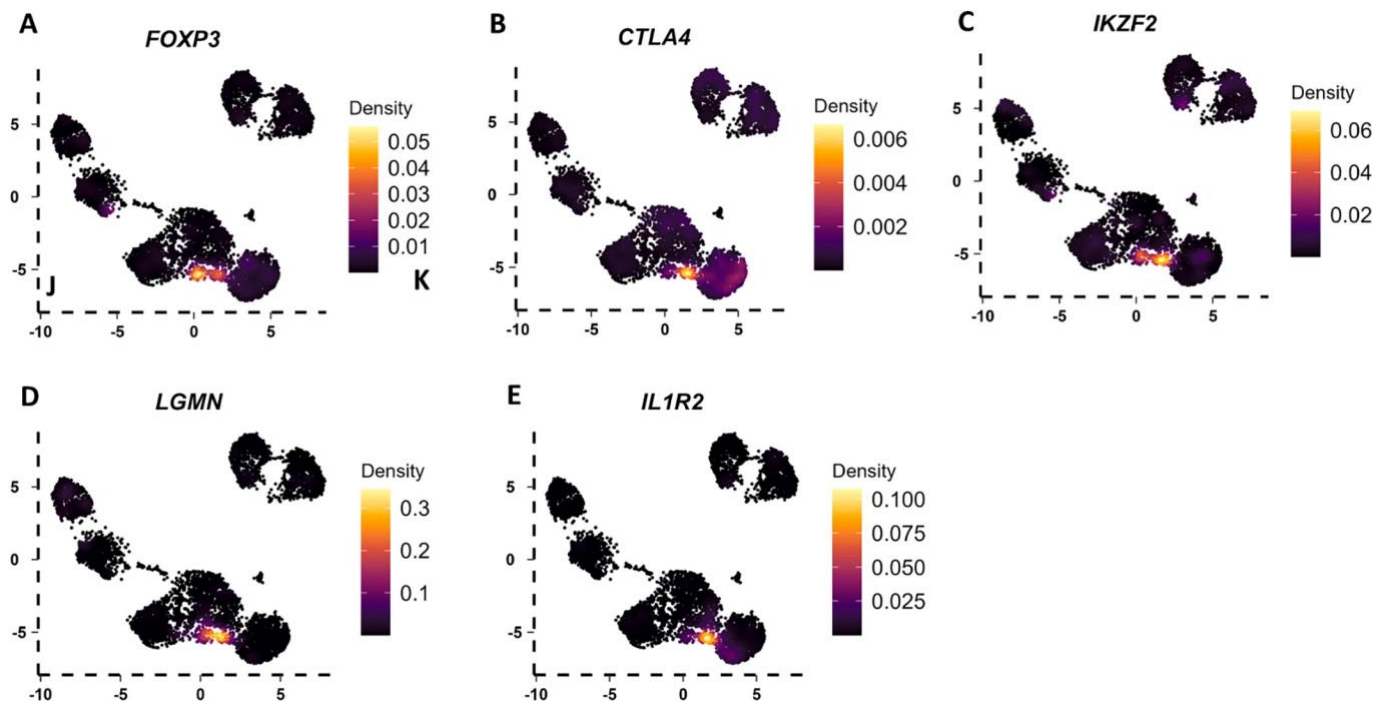


Fig. 10. Visualization of specific genes for Treg subpopulation.

A–E) Density plots highlighting gene expression intensity for chosen genes in the dataset. Right color bar: a relative density scaling.

As for the spontaneously differentiated Treg cells, a highly specific expression of genes such as *IKZF2* encoding the IKAROS family zinc finger 2 (also known as Helios), *LGMN* encoding legumain and *IL1R2* encoding interleukin 1 receptor 2 (Figs. 2C and 10A–E). Several of these genes were already discussed in terms of Treg differentiation. *IKZF2* encoding Helios was shown to contribute to the formation of Treg cells in cooperation with their main transcription factor, *FOXP3* [87]. Next, *IL1R2* and *LGMN* were found specific for Treg cells ([79,80,88]). According to the literature, Th2 differentiation is the default trans-differentiation program in both human and mouse Treg cells after the downregulation of *Foxp3* [81]. Likely, sorted TH2_2 cells may be Th2-like Treg cells [82,83].

4. Summary

This study focused on the analysis of heterogeneity among Th1, Th2, Th17 and Treg cell subsets and summarization of the best practices regarding the analysis of scRNASeq data fitted to Th cell datasets (e.g. gene count threshold estimation, cell cycle scoring, and regression or filtering of expression of mitochondria-associated genes). This study was based on the analysis of Th cells isolated from peripheral blood of healthy donors. The scRNASeq analysis elucidated the existence of a rigid immunophenotype of Th2 cells that was broadly separated from the rest of the dataset. Moreover, gene expression profile similarity of Th1, Th17, and spontaneously differentiated Treg cells indicated a potential plasticity between these subsets, which is consistent with recent studies regarding Th1/Th17/Treg trans-differentiation [76,84]. Moreover, this study was conducted with an emphasis on the identification of possible novel signature genes for the analyzed Th cell subsets, highlighting upregulation in *ACT_1* in genes such as *MX1*, *IFI44L*, *ISG15*, and *LY6E* that are likely to be Th1 cells [2,54,60,85], *SPINT2*, *SES2*, *PHGDH*, and *PSAT1* in the case of Th2 cells [86,69–71], *BATF*, *RBPJ*, *S100A4*, and *S100A6* in the case of Th17 ([75,9]) and *IKZF2*, *LGMN*, and *IL1R2* in the case of Treg cells [79,80,87–90], which is in line with recent studies aimed at Th cell subset phenotyping. Moreover, we performed trajectory analysis, which revealed possible routes in terms of trans-differentiation according to expression profile changes. In

combination of different tools, we identified highly specific Th1 genes, *CCL3*, and *CCL4* and supported this result by qPCR. For Th2 we were able to identify three genes *SPINT2*, *TRIB3* and *CST7*. All of them seem to be expressed in Th2 with very high specificity. For Th17, two genes were suggested as a signature for this subset – *F5* and *MAST4*.

Overall, this study may contribute to the simplification of scRNASeq data analysis optimization by discussing current best practices adjusted to Th cell scRNASeq data and providing insight into the characterization of Th cell subsets based on a combination of conventional and potential novel signature genes.

CRedit authorship contribution statement

Radim Jaroušek: Conceptualization, Methodology, Investigation, Software, Formal analysis, Visualization, Data curation, Writing – original draft. **Antónia Mikulová:** Conceptualization, Software, Methodology, Investigation, Formal analysis, Data curation, Visualization, Writing – original draft. **Petra Daďová:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Petr Tauš:** Methodology, Software, Formal analysis, Data curation, Visualization. **Terézia Kurucová:** Methodology, Investigation. **Karla Plevová:** Supervision, Writing – review & editing. **Boris Tichý:** Methodology, Supervision, Resources. **Lukáš Kubala:** Conceptualization, Supervision, Resources, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the institutional support from the Institute of Biophysics of the Czech Academy of Sciences, from European Structural and Investment Funds, Operational Program Research, Development and Education – “Preclinical Progression of New Organic Compounds with Targeted Biological Activity” (Preclinprogress) - CZ.02.1.01/0.0/0.0/16_025/0007381, and from the Ministry of Health of the Czech Republic (No. NU20-08-00314). Further, we acknowledge the CF Genomics CEITEC MU supported by the NCMG research infrastructure (LM2018132 funded by MEYS CR) for their support with obtaining scientific data presented in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bbamcr.2022.119321>.

References

- T.R. Mosmann, H. Cherwinski, M.W. Bond, M.A. Giedlin, R.L. Coffman, Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins, *J. Immunol.* (Baltimore, Md.: 1950) 136 (7) (1986) 2348–2357.
- J. Zhu, T helper cell differentiation, heterogeneity, and plasticity, *Cold Spring Harb. Perspect. Biol.* 10 (10) (2018), a030338, <https://doi.org/10.1101/cshperspect.a030338>.
- L.A. Lieberman, M. Banica, S.L. Reiner, C.A. Hunter, STAT1 plays a critical role in the regulation of antimicrobial effector mechanisms, but not in the development of Th1-type responses during toxoplasmosis, *J. Immunol.* 172 (1) (2004) 457–463, <https://doi.org/10.4049/jimmunol.172.1.457>.
- N. Hagberg, M. Joellsson, D. Leonard, S. Reid, M.-L. Eloranta, J. Mo, M.K. Nilsson, A.-C. Syvänen, Y.T. Bryceson, L. Rönnblom, The STAT4 SLE risk allele rs7574865 [T] is associated with increased IL-12-induced IFN- γ production in T cells from patients with SLE, *Ann. Rheum. Dis.* 77 (7) (2018) 1070–1077, <https://doi.org/10.1136/annrheumdis-2017-212794>.
- O. Chornoguz, R.S. Hagan, A. Haile, M.L. Arwood, C.J. Gamper, A. Banerjee, J. D. Powell, MTOC1 promotes T-bet phosphorylation to regulate Th1 differentiation, *J. Immunol.* 198 (10) (2017) 3939–3948, <https://doi.org/10.4049/jimmunol.1601078>.
- E.J. Scheinman, O. Avni, Transcriptional regulation of Gata3 in T helper cells by the integrated activities of transcription factors downstream of the Interleukin-4 receptor and T cell receptor, *J. Biol. Chem.* 284 (5) (2009) 3037–3048, <https://doi.org/10.1074/jbc.M807302200>.
- J. Cote-Sierra, G. Foucras, L. Guo, L. Chiodetti, H.A. Young, J. Hu-Li, J. Zhu, W. E. Paul, Interleukin 2 plays a central role in Th2 differentiation, *Proc. Natl. Acad. Sci.* 101 (11) (2004) 3880–3885, <https://doi.org/10.1073/pnas.0400339101>.
- R.V. Luckheeram, R. Zhou, A.D. Verma, B. Xia, CD4⁺ T cells: differentiation and functions, *Clin. Dev. Immunol.* 2012 (2012) 1–12, <https://doi.org/10.1155/2012/925135>.
- A.S. Rapaport, W. Ouyang, TRIMMING TGF- β signals in Th17 cells, *J. Exp. Med.* 215 (7) (2018) 1775–1776, <https://doi.org/10.1084/jem.20180986>.
- P. Fasching, M. Stradner, W. Graninger, C. DeJaco, J. Fessler, Therapeutic potential of targeting the Th17/Treg Axis in autoimmune disorders, *Molecules* 22 (1) (2017) 134, <https://doi.org/10.3390/molecules22010134>.
- H. Chang, F. Zhao, X. Xie, Y. Liao, Y. Song, C. Liu, Y. Wu, Y. Wang, D. Liu, Y. Wang, J. Zou, Z. Qi, PPAR α suppresses Th17 cell differentiation through IL-6/STAT3/ROR γ t pathway in experimental autoimmune myocarditis, *Exp. Cell Res.* 375 (1) (2019) 22–30, <https://doi.org/10.1016/j.yexcr.2018.12.005>.
- Y. Kurebayashi, S. Nagai, A. Ikejiri, M. Ohtani, K. Ichiyama, Y. Baba, T. Yamada, S. Egami, T. Hoshii, A. Hirao, S. Matsuda, S. Koyasu, PI3K-Akt-mTORC1-S6K1/2 Axis controls Th17 differentiation by regulating Gfi1 expression and nuclear translocation of ROR γ , *Cell Rep.* 1 (4) (2012) 360–373, <https://doi.org/10.1016/j.celrep.2012.02.007>.
- C. Asseman, S. Mauze, M.W. Leach, R.L. Coffman, F. Powrie, An essential role for interleukin 10 in the function of regulatory T cells that inhibit intestinal inflammation, *J. Exp. Med.* 190 (7) (1999) 995–1004, <https://doi.org/10.1084/jem.190.7.995>.
- U. Ellinghaus, A. Cortini, C.L. Pinder, G. Le Friec, C. Kemper, T.J. Vyse, Dysregulated CD46 shedding interferes with Th1-contraction in systemic lupus erythematosus, *Eur. J. Immunol.* 47 (7) (2017) 1200–1210, <https://doi.org/10.1002/eji.201646822>.
- P. Licona-Limón, L.K. Kim, N.W. Palm, R.A. Flavell, TH2, allergy and group 2 innate lymphoid cells, *Nat. Immunol.* 14 (6) (2013) 536–542, <https://doi.org/10.1038/ni.2617>.
- Y. Zheng, L. Sun, T. Jiang, D. Zhang, D. He, H. Nie, TNF α promotes Th17 cell differentiation through IL-6 and IL-1 β produced by monocytes in rheumatoid arthritis, *J. Immunol Res* 2014 (2014) 1–12, <https://doi.org/10.1155/2014/385352>.
- N.G. Núñez, J. Tosello Boari, R.N. Ramos, W. Richer, N. Cagnard, C. D. Anderfuhren, L.L. Niborski, J. Bigot, D. Meseure, P. De La Rochere, M. Milder, S. Viel, D. Loirat, L. Pérol, A. Vincent-Salomon, X. Sastre-Garau, B. Burkhardt, C. Sedlik, O. Lantz, E. Piaggio, Tumor invasion in draining lymph nodes is associated with Treg accumulation in breast cancer patients, *Nat. Commun.* 11 (1) (2020) 3272, <https://doi.org/10.1038/s41467-020-17046-2>.
- G.X.Y. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S. B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, M.T. Gregory, J. Shuga, L. Montesclaros, J.G. Underwood, D.A. Masquelier, S.Y. Nishimura, M. Schnall-Levin, P.W. Wyatt, C.M. Hindson, J.H. Bielas, Massively parallel digital transcriptional profiling of single cells, *Nat. Commun.* 8 (1) (2017) 14049, <https://doi.org/10.1038/ncomms14049>.
- H. Pagès, P. Aboyoun, R. Gentleman, S. DebRoy, Biostrings: Efficient Manipulation of Biological Strings (2.58.0) [Computer software], *Bioconductor version: Release (3.12)*, 2021, <https://doi.org/10.18129/B9.bioc.Biostrings>.
- NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.* 46 (D1) (2018) D8–D13, <https://doi.org/10.1093/nar/gkx1095>.
- C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression, *Genome Biol.* 20 (1) (2019) 296, <https://doi.org/10.1186/s13059-019-1874-1>.
- I.T. Jolliffe, Principal component analysis and factor analysis, in: I.T. Jolliffe (Ed.), *Principal Component Analysis*, Springer, 1986, pp. 115–128, https://doi.org/10.1007/978-1-4757-1904-8_7.
- L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, In *ArXiv e-prints*, <https://ui.adsabs.harvard.edu/abs/2018arXiv180203426M>, 2018.
- H. Wickham, *Ggplot2: Elegant Graphics for Data Analysis*, 2nd ed., Springer Publishing Company, Incorporated, 2009.
- G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, C.K. Slichter, H. W. Miller, M.J. McElrath, M. Prlic, P.S. Linsley, R. Gottardo, MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data, *Genome Biol.* 16 (1) (2015) 278, <https://doi.org/10.1186/s13059-015-0844-5>.
- Ggplot2 Based Publication Ready Plots (n.d.). Retrieved October 18, 2021, from, <https://rpkgs.datanovia.com/ggpubr/index.html>.
- J. Alquicira-Hernandez, J.E. Powell, Nebulosa recovers single-cell gene expression signals by kernel density estimation, *Bioinformatics* 37 (16) (2021) 2485–2487, <https://doi.org/10.1093/bioinformatics/btab003>.
- E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, Ma'ayan, A., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, *BMC Bioinformatics* 14 (2013) 128, <https://doi.org/10.1186/1471-2105-14-128>.
- L. Garcia-Alonso, C.H. Holland, M.M. Ibrahim, D. Turei, J. Saez-Rodriguez, Benchmark and integration of resources for the estimation of human transcription factor activities, *Genome Res.* 29 (8) (2019) 1363–1375, <https://doi.org/10.1101/gr.240663.118>.
- C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N.J. Lennon, K.J. Livak, T.S. Mikkelsen, J.L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, *Nat. Biotechnol.* 32 (4) (2014) 381–386.
- H.A. Pliner, J. Shendure, C. Trapnell, Supervised classification enables rapid annotation of cell atlases, *Nat. Methods* 16 (10) (2019) 983–986.
- T. Ilcic, J.K. Kim, A.A. Kolodziejczyk, F.O. Bagger, D.J. McCarthy, J.C. Marioni, S. A. Teichmann, Classification of low quality cells from single-cell RNA-seq data, *Genome Biol.* 17 (2016) 29, <https://doi.org/10.1186/s13059-016-0888-1>.
- H. Medini, T. Cohen, D. Mishmar, Mitochondrial gene expression in single cells shape pancreatic beta cells' sub-populations and explain variation in insulin pathway, *Sci. Rep.* 11 (1) (2021) 466, <https://doi.org/10.1038/s41598-020-80334-w>.
- P.A. Szabo, H.M. Levitin, M. Miron, M.E. Snyder, T. Senda, J. Yuan, Y.L. Cheng, E. C. Bush, P. Dogra, P. Thapa, D.L. Farber, P.A. Sims, Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease, *Nat. Commun.* 10 (1) (2019) 4706, <https://doi.org/10.1038/s41467-019-12464-3>.
- A.-C. Villani, K. Shekhar, Single-cell RNA sequencing of human T cells, *Methods Mol. Biol.* (Clifton, N.J.) 1514 (2017) 203–239, https://doi.org/10.1007/978-1-4939-6548-9_16.
- T.R. Mercer, S. Neph, M.E. Dinger, J. Crawford, M.A. Smith, A.-M.J. Shearwood, E. Haugen, C.P. Bracken, O. Rackham, J.A. Stamatoyannopoulos, A. Filipovska, J. S. Mattick, The human mitochondrial transcriptome, *Cell* 146 (4) (2011) 645–658, <https://doi.org/10.1016/j.cell.2011.06.051>.
- D. Osorio, J.J. Cai, Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control, *Bioinformatics (Oxford, England)* 37 (7) (2021) 963–967, <https://doi.org/10.1093/bioinformatics/btaa751>.
- S. Márquez-Jurado, J. Díaz-Colunga, R.P. das Neves, A. Martínez-Lorente, F. Almazán, R. Guantes, F.J. Iborra, Mitochondrial levels determine variability in cell death by modulating apoptotic gene expression, *Nat. Commun.* 9 (1) (2018) 389, <https://doi.org/10.1038/s41467-017-02787-4>.
- Q. Zhao, J. Wang, I.V. Levichkin, S. Stasinopoulos, M.T. Ryan, N.J. Hoogenraad, A mitochondrial specific stress response in mammalian cells, *EMBO J.* 21 (17) (2002) 4411–4419, Scopus, <https://doi.org/10.1093/emboj/cdf445>.
- W. Xiong, Y. Jiao, W. Huang, M. Ma, M. Yu, Q. Cui, D. Tan, Regulation of the cell cycle via mitochondrial gene expression and energy metabolism in HeLa cells, *Acta Biochim. Biophys. Sin.* 44 (4) (2012) 347–358, <https://doi.org/10.1093/abbs/gms006>.

- [41] A. McDavid, G. Finak, R. Gottardo, The contribution of cell cycle to heterogeneity in single-cell RNA-seq data, *Nat. Biotechnol.* 34 (6) (2016) 591–593, <https://doi.org/10.1038/nbt.3498>.
- [42] H.T.N. Tran, K.S. Ang, M. Chevrier, X. Zhang, N.Y.S. Lee, M. Goh, J. Chen, A benchmark of batch-effect correction methods for single-cell RNA sequencing data, *Genome Biol.* 21 (1) (2020) 12, <https://doi.org/10.1186/s13059-019-1850-9>.
- [43] A.A. AlJanahi, M. Danielsen, C.E. Dunbar, An introduction to the analysis of single-cell RNA-sequencing data, *Mol. Ther. Methods Clin. Develop.* 10 (2018) 189–196, <https://doi.org/10.1016/j.omtm.2018.07.003>.
- [44] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial, *Mol. Syst. Biol.* 15 (6) (2019), e8746, <https://doi.org/10.15252/msb.20188746>.
- [45] S. Nestorowa, F.K. Hamey, B. Pijuan Sala, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D.G. Kent, B. Göttgens, A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation, *Blood* 128 (8) (2016) e20–e31, <https://doi.org/10.1182/blood-2016-05-716480>.
- [46] Y. Hao, S. Hao, E. Andersen-Nissen, W.M. 3rd Mauck, S. Zheng, A. Butler, et al., Integrated analysis of multimodal single-cell data, *Cell* 184 (13) (2021) 3573–3587.e29.
- [47] I. Tirosh, B. Izar, S.M. Prakadan, I.I. Marc H. Wadsworth, D. Treacy, J.J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A.S. Genshaft, T.K. Hughes, C.G.K. Ziegler, L. A. Garraway, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq, *Science*. (2016) <https://www.science.org/doi/abs/10.1126/science.aad0501>.
- [48] M. Moussa, I.I. Mändoiu, Computational Cell Cycle Analysis of Single Cell RNA-Seq data, p. 2020.11.21.392613, 2020, <https://doi.org/10.1101/2020.11.21.392613>.
- [49] E.I. Athanasiadis, J.G. Botthof, H. Andres, L. Ferreira, P. Lio, A. Cvejic, Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis, *Nat. Commun.* 8 (1) (2017) 2045, <https://doi.org/10.1038/s41467-017-02305-6>.
- [50] Björklund, Å. (n.d.). Quality Control of scRNAseq Data. 47.
- [51] J.C. Guimaraes, M. Zavolan, Patterns of ribosomal protein expression specify normal and malignant human cells, *Genome Biol.* 17 (2016) 236, <https://doi.org/10.1186/s13059-016-1104-z>.
- [52] B. Hwang, J.H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, *Exp. Mol. Med.* 50 (8) (2018) 1–14, <https://doi.org/10.1038/s12276-018-0071-8>.
- [53] J.P. Twohig, A. Cardus Figueras, R. Andrews, F. Wiede, B.C. Cossins, A. Derrac Soria, M.J. Lewis, M.J. Townsend, D. Millrine, J. Li, D.G. Hill, J. Uceda Fernandez, X. Liu, B. Szomolay, C.J. Pepper, P.R. Taylor, C. Pitzalis, T. Tiganis, N.M. Williams, S.A. Jones, Activation of naive CD4₊ T cells re-tunes STAT1 signaling to deliver unique cytokine responses in memory CD4₊ T cells, *Nat. Immunol.* 20 (4) (2019) 458–470, <https://doi.org/10.1038/s41590-019-0350-0>.
- [54] M.C. Gerner, L. Niederstaetter, L. Ziegler, A. Bileck, A. Slany, L. Janker, R.L. J. Schmidt, C. Gerner, G. Del Favero, K.G. Schmetterer, Proteome analysis reveals distinct mitochondrial functions linked to interferon response patterns in activated CD4₊ and CD8₊ T cells, *Front. Pharmacol.* 10 (2019) 727, <https://doi.org/10.3389/fphar.2019.00727>.
- [55] D. Hotter, M. Bosso, K.L. Jonsson, C. Krapp, C.M. Stürzel, A. Das, E. Littwitz-Salomon, B. Berkhout, A. Russ, S. Wittmann, T. Gramberg, Y. Zheng, L.J. Martins, V. Planelles, M.R. Jakobsen, B.H. Hahn, U. Dittmer, D. Sauter, F. Kirchhoff, IFI16 targets the transcription factor Sp1 to suppress HIV-1 transcription and latency reactivation, *Cell Host Microbe* 25 (6) (2019) 858–872.e13, <https://doi.org/10.1016/j.chom.2019.05.002>.
- [56] S. Hugues, A. Boissonnas, S. Amigorena, L. Fetler, The dynamics of dendritic cell–T cell interactions in priming and tolerance, *Curr. Opin. Immunol.* 18 (4) (2006) 491–495, <https://doi.org/10.1016/j.coi.2006.03.021>.
- [57] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nat. Biotechnol.* 36 (5) (2018) 411–420, <https://doi.org/10.1038/nbt.4096>.
- [58] E. Cano-Gamez, B. Soskic, T.I. Roumeliotis, E. So, D.J. Smyth, M. Baldrighi, D. Willé, N. Nakic, J. Esparza-Gordillo, C.G.C. Larminie, P.G. Bronson, D.F. Tough, W.C. Rowan, J.S. Choudhary, G. Trynka, Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4₊ T cells to cytokines, *Nat. Commun.* 11 (1) (2020) 1801, <https://doi.org/10.1038/s41467-020-15543-y>.
- [59] C.T. Weaver, One road to the TH17 pathway: how TH1 led to TH17 (and vice versa), and first came last, *Nat. Immunol.* 21 (8) (2020) 819–821.
- [60] J. Yu, C. Liang, S.-L. Liu, Interferon-inducible LY6E protein promotes HIV-1 infection, *J. Biol. Chem.* 292 (11) (2017) 4674–4685, <https://doi.org/10.1074/jbc.M116.755819>.
- [61] H. Zhu, L.-F. Wu, X.-B. Mo, X. Lu, H. Tang, X.-W. Zhu, W. Xia, Y.-F. Guo, M.-J. Wang, K.-Q. Zeng, J. Wu, Y.-H. Qiu, X. Lin, Y.-H. Zhang, Y.-Z. Liu, N.-J. Yi, F.-Y. Deng, S.-F. Lei, Rheumatoid arthritis-associated DNA methylation sites in peripheral blood mononuclear cells, *Ann. Rheum. Dis.* 78 (1) (2019) 36–42, <https://doi.org/10.1136/annrheumdis-2018-213970>.
- [62] B. Zhang, S.Q. Liu, C. Li, E. Lykken, S. Jiang, E. Wong, Z. Gong, Z. Tao, B. Zhu, Y. Wan, Q.J. Li, MicroRNA-23a curbs necrosis during early T cell activation by enforcing intracellular reactive oxygen species equilibrium, *Immunity* 44 (3) (2016) 568–581.
- [63] M. Yamada, J. Ohnishi, B. Ohkawara, S. Iemura, K. Satoh, J. Hyodo-Miura, K. Kawachi, T. Natsume, H. Shibuya, NARF, a nemo-like kinase (NLK)-associated ring finger protein regulates the ubiquitylation and degradation of T cell factor/lymphoid enhancer factor (TCF/LEF), *J. Biol. Chem.* 281 (30) (2006) 20749–20760.
- [64] B. Larkin, V. Ilyukha, M. Sorokin, A. Buzdin, E. Vannier, A. Poltorak, Cutting edge: activation of STING in T cells induces type I IFN responses and cell death, *J. Immunol.* 199 (2) (2017) 397–402.
- [65] K. Janssens, C. Van den Haute, V. Baekelandt, S. Lucas, J. Van Horsen, V. Somers, B. Van Wijmeersch, P. Stinissen, J.J.A. Hendriks, H. Slaets, N. Hellings, Leukemia inhibitory factor tips the immune balance towards regulatory T cells in multiple sclerosis, *Brain Behav. Immun.* 45 (2015) 180–188.
- [66] S.M. Metcalfe, LIF in the regulation of T-cell fate and as a potential therapeutic, *Genes Immun.* 12 (3) (2011) 157–168.
- [67] J. Eberlein, B. Davenport, T.T. Nguyen, F. Victorino, K. Jhun, V. van der Heide, M. Kuleshov, A. Ma'ayan, R. Kedl, D. Homann, Chemokine signatures of pathogen-specific T cells I: effector T cells, *J. Immunol.* 205 (8) (2020) 2169–2187.
- [68] E.H. Ma, G. Bantug, T. Griss, S. Condotta, R.M. Johnson, B. Samborska, N. Mainolfi, V. Suri, H. Guak, M.L. Balmer, M.J. Verway, T.C. Raissi, H. Tsui, G. Boukhaled, S. Henriques da Costa, C. Frezza, C.M. Krawczyk, A. Friedman, M. Manfredi, R. G. Jones, Serine is an essential metabolite for effector T cell expansion, *Cell Metab.* 25 (2) (2017) 345–357, <https://doi.org/10.1016/j.cmet.2016.12.011>.
- [69] A.H. Kowalsky, S. Namkoong, E. Mettetal, H.-W. Park, D. Kazyken, D.C. Fingar, J. H. Lee, The GATOR2–mTORC2 axis mediates Sestrin2-induced AKT Ser/Thr kinase activation, *J. Biol. Chem.* 295 (7) (2020) 1769–1780, <https://doi.org/10.1074/jbc.RA119.010857>.
- [70] D.Y. Jun, D. Taub, F.J. Chrest, Y.H. Kim, Requirement of the expression of 3-phosphoglycerate dehydrogenase for traversing S phase in murine T lymphocytes following immobilized anti-CD3 activation, *Cell. Immunol.* 287 (2) (2014) 78–85, <https://doi.org/10.1016/j.cellimm.2013.12.003>.
- [71] Y.-C. Chan, Y.-C. Chang, H.-H. Chuang, Y.-C. Yang, Y.-F. Lin, M.-S. Huang, M. Hsiao, C.-J. Yang, K.-T. Hua, Overexpression of PSAT1 promotes metastasis of lung adenocarcinoma by suppressing the IRF1-IFN γ axis, *Oncogene* 39 (12) (2020) 2509–2522, <https://doi.org/10.1038/s41388-020-1160-4>.
- [72] S. Shang, Y.W. Yang, F. Chen, L. Yu, S.H. Shen, K. Li, B. Cui, X.X. Lv, C. Zhang, C. Yang, J. Liu, J.J. Yu, X.W. Zhang, P.P. Li, S.T. Zhu, H.Z. Zhang, F. Hua, TRIB3 reduces CD8 α T cell infiltration and induces immune evasion by repressing the STAT1-CXCL10 axis in colorectal cancer, *Sci. Transl. Med.* 14 (626) (2022).
- [73] M.P. Nanut, J. Sabotic, A. Jewett, J. Kos, Cysteine cathepsins as regulators of the cytotoxicity of NK and T cells, *Front. Immunol.* 5 (2014).
- [74] C.A. Tibbitt, J.M. Stark, L. Martens, J. Ma, J.E. Mold, K. Deswarte, G. Oliynyk, X. Feng, B.N. Lambrecht, P. De Bleser, S. Nylen, H. Hammad, M. Arsenian Henriksson, Y. Saey, J.M. Coquet, Single-cell RNA sequencing of the T helper cell response to house dust mites defines a distinct gene expression signature in airway Th2 cells, *Immunity* 51 (1) (2019) 169–184.e165.
- [75] G. Meyer zu Horste, C. Wu, C. Wang, L. Cong, M. Pawlak, Y. Lee, W. Elyaman, S. Xiao, A. Regev, V.K. Kuchroo, RBP1 controls development of pathogenic Th17 cells by regulating IL-23 receptor expression, *Cell Rep.* 16 (2) (2016) 392–404, <https://doi.org/10.1016/j.celrep.2016.05.088>.
- [76] A. Ferraro, S.M. Buonocore, P. Auquier, I. Nicolas, H. Wallemaq, D. Boutriau, R. G. van der Most, Role and plasticity of Th1 and Th17 responses in immunity to *Staphylococcus aureus*, *Human Vaccines Immunother.* 15 (12) (2019) 2980–2992, <https://doi.org/10.1080/21645515.2019.1613126>.
- [77] L. Cosmi, F. Liotta, E. Maggi, S. Romagnani, F. Annunziato, Th17 and non-classic Th1 cells in chronic inflammatory disorders: two sides of the same coin, *Int. Arch. Allergy Immunol.* 164 (3) (2014) 171–177, <https://doi.org/10.1159/000363502>.
- [78] M. Arai, M.B. Prystowsky, J.A. Cohen, Expression of the T-lymphocyte activation gene, F5, by mature neurons, *J. Neurosci. Res.* 33 (4) (1992) 527–537.
- [79] M. De Simone, A. Arrigoni, G. Rossetti, P. Guarini, V. Ranzani, C. Politano, R.J. P. Bonnal, E. Provasi, M.L. Sarnicola, I. Panzeri, M. Moro, M. Crosti, S. Mazzara, V. Vaira, S. Bosari, A. Palleschi, L. Santambrogio, G. Bovo, N. Zucchini, M. Pagani, Transcriptional landscape of human tissue lymphocytes unveils uniqueness of tumor-infiltrating T regulatory cells, *Immunity* 45 (5) (2016) 1135–1147, <https://doi.org/10.1016/j.immuni.2016.10.021>.
- [80] E. Nikolouli, Y. Elfaki, S. Herppich, C. Schelmbauer, M. Delacher, C. Falk, I. A. Mufazalov, A. Waisman, M. Feuerer, J. Huehn, Recirculating IL-1R2⁺ Tregs fine-tune intrathymic Treg development under inflammatory conditions, *Cell. Mol. Immunol.* 18 (1) (2021) 182–193, <https://doi.org/10.1038/s41423-019-0352-8>.
- [81] L. Hansmann, C. Schmid, J. Kett, L. Steger, R. Andreesen, P. Hoffmann, M. Rehli, M. Edinger, Dominant Th2 differentiation of human regulatory T cells upon loss of FOXP3 expression, *J. Immunol.* 188 (3) (2012) 1275–1282, <https://doi.org/10.4049/jimmunol.1102288>.
- [82] N. Gao, W. Cui, L.M. Zhao, T.T. Li, J.H. Zhang, L.L. Pan, Contribution of Th2-like Treg cells to the pathogenesis of Takayagi's arteritis, *Clin. Exp. Rheumatol.* 38 (Suppl 124(2)) (2020) 48–54.
- [83] L. Halim, M. Romano, R. McGreggor, I. Correa, P. Pavlidis, N. Grageda, S.-J. Hoong, M. Yuskel, W. Jassem, R.F. Hanner, M. Ong, O. Mckinney, B. Hayee, S. N. Karagiannis, N. Powell, R.I. Lechler, E. Nova-Lamperti, G. Lombardi, An atlas of human regulatory T helper-like cells reveals features of Th2-like Tregs that support a tumorigenic environment, *Cell Rep.* 20 (3) (2017) 757–770, <https://doi.org/10.1016/j.celrep.2017.06.079>.
- [84] M. Kleinewietfeld, D.A. Hafler, The plasticity of human Treg and Th17 cells and its role in autoimmunity, *Semin. Immunol.* 25 (4) (2013) 305–312, <https://doi.org/10.1016/j.jsmim.2013.10.009>.
- [85] C.D. Swaim, L.A. Canadeo, K.J. Monte, S. Khanna, D.J. Lenschow, J.M. Huijbregtse, Modulation of extracellular ISG15 signaling by pathogens and viral effector proteins, *Cell Rep.* 31 (11) (2020), 107772, <https://doi.org/10.1016/j.celrep.2020.107772>.
- [86] J.M. Stark, C.A. Tibbitt, J.M. Coquet, The metabolic requirements of Th2 cell differentiation, *Front. Immunol.* 10 (2019) 2318, <https://doi.org/10.3389/fimmu.2019.02318>.

- [87] F. Ocklenburg, D. Moharreggh-Khiabani, R. Geffers, V. Janke, S. Pfoertner, H. Garritsen, L. Groebe, J. Klemptner, K.E.J. Dittmar, S. Weiss, J. Buer, M. Probst-Kepper, UBD, a downstream element of FOXP3, allows the identification of LGALS3, a new marker of human regulatory T cells, *Lab. Investig.* 86 (7) (2006) 724–737, <https://doi.org/10.1038/labinvest.3700432>.
- [88] M. Probst-Kepper, R. Geffers, A. Kröger, N. Viegas, C. Erck, H.-J. Hecht, H. Lünsdorf, R. Roubin, D. Moharreggh-Khiabani, K. Wagner, F. Ocklenburg, A. Jeron, H. Garritsen, T.p. Arstila, E. Kekäläinen, R. Balling, H. Hauser, J. Buer, S. Weiss, GARP: a key receptor controlling FOXP3 in human regulatory T cells, *J. Cell. Mol. Med.* 13 (9b) (2009) 3343–3357, <https://doi.org/10.1111/j.1582-4934.2009.00782.x>.
- [89] H. Takatori, H. Kawashima, A. Matsuki, K. Meguro, S. Tanaka, T. Iwamoto, Y. Sanayama, N. Nishikawa, T. Tamachi, K. Ikeda, A. Suto, K. Suzuki, S. Kagami, K. Hirose, M. Kubo, S. Hori, H. Nakajima, Helios enhances Treg cell function in cooperation with FoxP3, *Arthritis Rheumatol.* 67 (6) (2015) 1491–1502, <https://doi.org/10.1002/art.39091>.
- [90] H.-F. Tsai, C.-S. Wu, Y.-L. Chen, H.-J. Liao, I.-T. Chyuan, P.-N. Hsu, Galectin-3 suppresses mucosal inflammation and reduces disease severity in experimental colitis, *J. Mol. Med.* 94 (5) (2016) 545–556, <https://doi.org/10.1007/s00109-015-1368-x>.
- [91] RA Jarvis, EA Patrick, Clustering Using a Similarity Measure Based on Shared Near Neighbors, *IEEE Transactions on Computers C-22* (11) (1973) 1025–1034, <https://doi.org/10.1109/T-C.1973.223640>.