

**Statistical Method Development and Design of
Computational Pipelines for Differential Analyses of
High-Throughput Data, Including DNA Microarrays,
RNA Sequencing and HDCyto Data**

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von

Malgorzata Franciszka Nowicka

aus

Polen

Promotionskommission

Prof. Dr. Mark D. Robinson (Vorsitz)

Prof. Dr. Torsten Hothorn

Prof. Dr. Magnus Rattray

Zürich, 2017

Preface

In first place, I want to thank my supervisor Mark D. Robinson for choosing me as his PhD student during the Swiss Institute of Bioinformatics (SIB) Fellowship interviews. The final interview part was combined with the SIB Days. I remember being amazed by all the great scientific talks that took place and the atmosphere of integrity between Swiss Bioinformaticians, and I very grateful that I could become a part of it.

Taking the chance, I would like acknowledge the funding from the SIB Fellowship.

However, these four years of my PhD would have not become fulfilled without the great supervision from Mark, for which I am sincerely grateful.

I also wish to thank, my other committee members, including Prof. Torsten Hothorn and Prof. Magnus Rattray for their advice during my committee meetings.

I give my most sincere thanks to my collaborators Edwin D. Hawkins and Cristina Lo Celso. Many thanks to Carsten Krieg for the very dynamic and exciting collaboration.

I cannot even express the extent to which I am thankful to my colleagues from the Robinson group for all the scientific and, most of all, for all the non-scientific discussions during lunches and Friday meetings, and for the great time and fun during our retreats. Special thank you to Lukas Weber, Helen Lindsay, Charlotte Soneson and Simone Tiberi for proof reading of my thesis. Vielen dank to Katharina Hembach and Lukas Weber for translating my abstract into German.

Moreover, I express my thanks to the former members of our group Charity Law, Katarina Matthes, Dania Machlab, Romy Schleiss and Xiaobei Zhou for all the inspiring talks.

I would like to also thank the members of von Mering and Baudis groups for all the constructive input during the Bioinfo meetings.

Accomplishing this PhD would not be possible without the great support from my climbing friends Caterina Specchia and Luca Lorenzelli. Climbing makes us cool at the University and PhD makes us cool in the climbing gym.

I extremely grateful to Marta Pittavino for being around and cheering me up in the more difficult times of my thesis writing. Many thanks to Kinga Czyzewska for being proud of me

even before completing the PhD.

Finally, I would like to thank my Polish and Swiss family and my wonderful boyfriend Patrick Hammer for being supportive during this intensive and challenging four years.

Zurich, May 2017

Malgorzata Nowicka

Zusammenfassung

Während meinen vier Jahren als Doktorandin habe ich mit diversen Arten von Hochdurchsatz-Daten gearbeitet, darunter DNA-Microarrays, RNA-Sequenzierungsdaten (RNA-seq) und Massenzytometrie (CyTOF). Diese Arten von Daten werden von Biologen oftmals für Vergleichsstudien gesammelt in denen Merkmale (z.B. Gene, Transkripte, Zellen, usw.) gesucht werden, deren Expression oder Häufigkeit mit dem gesuchten Phänotyp korrelieren. Solche Analysen werden als "Differentialanalysen" bezeichnet. Meine PhD Projekte haben sich somit vor allem mit der Entwicklung von massgeschneiderten statistischen Methoden und der Ausarbeitung von computergestützten Workflows für die Differentialanalyse der erwähnten Daten befasst.

Meine Dissertation beginnt mit der Beschreibung von wissenschaftlichen Hintergrundinformationen zu DNA-Microarrays, RNA-seq und Durchfluss- sowie Massenzytometrie. Für jede dieser Technologien wird jeweils die Charakteristik der zugrundeliegenden Daten erklärt und anschliessend aufgezeigt, wie ein typischer Workflow, dessen letzter Schritt immer eine Differentialanalyse darstellt, zusammengesetzt sein kann. Mein Fokus in dieser Dissertation wird dabei auf jener Art der Differentialanalyse liegen, welche in meinen Projekten zentraler Bestandteil war. Das heisst, bei den DNA-Microarrays ist dies die differentielle Genexpressionsanalyse (DGE-Analyse), für das Gebiet der RNA-seq ist es die differentielle Verwendung von Transkripten (DTU-Analyse, eine Erweiterung der DGE-Analyse), und im Kontext von CyTOF-Daten wird die differentielle Häufigkeit von Zellpopulationen und Marker-Signalen untersucht.

Nach der Einleitung präsentiere ich das Dirichlet-Multinomial (DM) Modell für die differentielle Verwendung von Transkripten (DTU) und Transkript-Verwendung Quantitative-Trait-Loci (tuQTL) Analysen von RNA-seq Daten. Beide Frameworks, die ich in einem R/Bioconductor-Paket namens DRIMSeq implementiert habe, nehmen als Eingabe (sie sind aber nicht darauf beschränkt) die Transkript-Quantifizierungen von den schnellen Pseudoalignment basierten Methoden wie zum Beispiel kallisto oder Salmon. Inspiriert von den existierenden Lösungsansätzen für RNA-seq, löst DRIMSeq die Herausforderung der Schätzung basierend auf wenigen Proben durch die Anwendung der "Cox-Reid adjusted profile likelihood" und Moderierungs-Schema in der Berechnung der Dispersion. Die aktualisierte Version von DRIMSeq erlaubt zusätzlich zur Gen-Level-Analyse mit DM nun auch die Analyse auf dem Transkript-Level mittels des Beta-Binomial (BB) Modells und somit die Identifikation von Transkripten, die die DTU antreiben. Die Erweiterung des Regressionsframeworks erleichtert die Analyse von beliebigen Experimental Designs. Zusatzinformationen zu diesem Ansatz, sowie eine Evaluation davon können im Paper I nachgelesen werden.

Im nächsten Teil beschreibe ich einen R/Bioconductor-Workflow zur Analyse von hochdimensionellen Massen- sowie Durchfluszytometriedaten (HDCyo). Dieser Workflow beinhaltet eine reproduzierbare und flexible Analyse, welche die Datenaufbereitung, graphische Diagnostik, Identifikation von Zellpopulationen, sowie die Differentialanalyse ermöglicht. Das Ziel dieser Arbeit war die Förderung von reproduzierbaren Forschungsergebnissen gestützt durch formelle statistische Inferenz. Ausserdem stellt dieses R/Bioconductor-Paket aufgrund der Kombination aus Code und detaillierter Beschreibung der Analysen eine interessante Referenzimplementation dar, die besonders für R-Neulinge hilfreich ist. Die einzelnen Schritte

des Workflows sind eine Verallgemeinerung des Workflows, der für die CyTOF-Analyse eines Kollaborationsprojektes verwendet wurde, das ich im nachfolgenden Teil der Dissertation vorstelle. Der Workflow wird im Paper II präsentiert.

Anschliessend beschreibe ich meine zwei Kollaborationsprojekte. Für das erste Projekt habe ich eine Differentialanalyse entworfen und durchgeführt, die auf Affymetrix GeneChip Mouse Gene Microarray-Daten basierte. Die Daten verglichen die Genexpression von Leukämieproben die vor und nach der Behandlung mit Dexamethasone entnommen wurden. Für das zweite Kollaborationsprojekt habe ich eine CyTOF-Analyse entwickelt und durchgeführt, deren Ziel es war, Biomarker zu identifizieren, die unterscheiden können ob die Immuntherapie bei Melanoma-Patienten Wirkung zeigt oder nicht. Eine der Hauptherausforderungen war dabei, dass die Analyse gepaarte Patienten-Daten und Gruppeneffekte berücksichtigen musste, da die Daten bei verschiedenen Färbungen und CyTOF-Experimenten gesammelt wurden. Gelöst wurde dies durch die Verwendung von Regressionsframeworks, wobei die HD-Cyto-Daten die erklärten Variablen sind. Im Detail heisst das, dass Linear Mixed Models und Generalized Linear Mixed Models für die Differentialanalysen der Zellpopulations-Häufigkeit und Signaling-Marker-Expression verwendet wurden. Das daraus resultierende Framework ermöglicht die Analyse von komplexen Experimental Designs. Zudem kann die Analyse dank ihres flexiblen und modularen Designs auch für andere Studien wiederverwendet werden. Eine detaillierte Beschreibung der zwei Kollaborationsprojekte befindet sich im Kapitel Collaboration Papers.

Das letzte Kapitel (Discussion and Perspectives) beschreibt die Details des in DRIMSeq implementierten Regressionsframeworks und das Beta-Binomial Modell für die Transkript-Level-Analysen. Zusätzlich werden mögliche Erweiterungen und Verbesserungen für DRIMSeq, wie zum Beispiel die Verwendung der statistischen Unsicherheit von Transkript-Schätzungen sowie die Unterstützung von weiteren Arten von multivariaten Daten besprochen.

Abstract

During the four years of my PhD, I had the opportunity to work with various types of high-throughput data, including DNA microarrays, RNA-sequencing (RNA-seq) and mass cytometry (CyTOF). Often those assays are used by biologists for comparative studies to detect features (genes, transcripts, cells, etc.) for which their expression or abundance correlate with phenotype of interest. Such comparisons are referred to as differential analyses. Overall, my PhD project involved development of tailored statistical methods and design of computational pipelines for differential analyses based on the mentioned types of data.

I start my dissertation by providing some necessary scientific background in the context of DNA microarrays, RNA-seq and flow and mass cytometry. For each of the technologies, the characteristics of the data are first described and followed by brief presentation of a typical analysis workflow where the final step is differential analysis. My focus is on highlighting the type of differential analysis that I was involved in during my PhD studies. That is, for the microarray data, I concentrate on differential gene expression (DGE). In the RNA-seq part, my focus is dedicated to differential transcript usage (DTU), which is an extension of DGE analysis. For CyTOF data, the differential abundance of cell populations and marker signals is interrogated.

Following the **Introduction**, I present the Dirichlet-multinomial (DM) model for DTU and transcript usage quantitative trait loci (tuQTL) analyses from RNA-seq data. Both frameworks, implemented as an R/Bioconductor package called **DRIMSeq**, take as input (but are not limited to) transcript quantifications from the fast pseudo-alignment based methods, such as *kallisto* or *Salmon*. Borrowing from the existing RNA-seq approaches, **DRIMSeq** handles the challenge of estimation in small sample size by exploiting the Cox-Reid adjusted profile likelihood and moderation scheme in the dispersion calculation. The recently updated version allows, additionally to gene-level analysis with DM, transcript-level analysis with the beta-binomial (BB) model, which enables identification of transcripts that drive the DTU. The extension to the regression framework facilitates the study of arbitrary experimental designs. Further information about the model and its evaluation can be found in **Paper I**.

Next, I introduce an R/Bioconductor workflow of analysis for high dimensional (mass and flow) cytometry (HDCyto) data. This pipeline covers a reproducible, flexible analysis that include data preprocessing, diagnostic plots, cell type identification and differential analysis. The goal of this work was to promote reproducible research supported with formal statistical inference. Moreover, by combining the code and detailed analysis description, it is a great resource especially for those who are newer to R. The individual steps of this workflow were generalized from the pipeline originally designed for the analysis of CyTOF data in the collaboration project introduced in the following paragraph. The workflow is presented in **Paper II**.

Subsequently, I move toward my two collaboration projects. In the first project, I designed and conducted a differential analysis based on Affymetrix GeneChip Mouse Gene microarray data that compared gene expression between leukemia samples collected before and after dexamethasone treatment. In the second collaborative project, I have designed and performed analysis of CyTOF data where the objective was to identify biomarkers discriminating between

melanoma patients who have responded to immunotherapy treatment and those who have not. One of the main challenges in this analysis was the necessity of accounting for patient pairing and batch effects, as the data were acquired in different stainings and CyTOF runs. These challenges were addressed by employing regression frameworks where the HDCyto data is the response. More specifically, linear mixed models and generalized linear mixed models were used for differential analysis of cell population abundance and signaling marker expression. The obtained framework enables analysis of complex experimental designs, and its flexibility and modularity makes it general enough to be applied to other studies. Detailed description of the two collaborative projects can be found in the **Collaboration Papers** chapter.

In the final chapter (**Discussion and Perspectives**), the details of the regression framework implemented in *DRIMSeq* and the beta-binomial model for transcript-level analysis are described. Additionally, some future improvements to *DRIMSeq* are discussed, such as incorporation of uncertainty of transcript estimates and applications to other types of multivariate data.

Thesis outline

Introduction

Paper I: **DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics**

Malgorzata Nowicka, Mark D. Robinson

Paper published in *F1000Research* (2016), 5(1356)

Paper II: **CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets**

Malgorzata Nowicka, Carsten Krieg, Lukas M. Weber, Felix J. Hartmann, Silvia Guglietta, Burkhard Becher, Mitch P. Levesque and Mark D. Robinson

Paper submitted to *F1000Research*

Collaboration Papers:

T-cell acute leukaemia exhibits dynamic interactions with bone marrow microenvironments

Edwin D. Hawkins, Delfim Duarte, Olufolake Akinduro, Reema A. Khorshed, Diana Passaro, Malgorzata Nowicka, Lenny Straszkowski, Mark K. Scott, Steve Rothery, Nicola Ruivo, Katie Foster, Michaela Waibel, Ricky W. Johnstone, Simon J. Harrison, David A. Westerman, Hang Quach, John Gribben, Mark D. Robinson, Louise E. Purton, Dominique Bonnet and Cristina Lo Celso

Paper published in *Nature* (2016), 538(7626), 518–522

High dimensional single cell analysis predicts response to anti-PD-1 immunotherapy

Carsten Krieg, Malgorzata Nowicka, Silvia Guglietta, Sabrina Schindler, Felix J. Hartmann, Lukas M. Weber, Reinhard Dummer, Mark D. Robinson, Mitchell P. Levesque and Burkhard Becher

Paper under review at *Nature Medicine*

Discussion and Perspectives

Contents

Introduction	13
1 Background	14
1.1 Biological background	14
1.2 DNA microarrays	17
1.3 DNA microarrays analysis	19
1.3.1 Preprocessing: background adjustment, normalization, summarization	19
1.3.2 Differential analysis	20
1.4 RNA-seq technology	21
1.5 RNA-seq data analysis	22
1.5.1 Quantification	24
1.5.2 Differential analysis	25
1.6 Strategies for modeling counts in small-sample size RNA-seq data	26
1.6.1 Two-stage estimation	26
1.6.2 Adjusted profile likelihood	27
1.6.3 Sharing information between genes	28
1.7 HDCyto data	28
1.8 HDCyto data analysis	30
1.8.1 Preprocessing: normalization, debarcoding, compensation, transformation	30
1.8.2 Dimension reduction	33
1.8.3 Cell population identification	34
1.8.4 Differential analysis	35
2 Research objectives	36
2.1 Multivariate model for DTU analysis	36
2.2 Workflow for differential analysis of HDCyto data	37
3 Research challenges	37
3.1 Inference in small sample size data	37
3.2 Modeling complex experimental designs	38
4 Thesis summary	39
References	41

Paper I	51
DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics	51
Paper II	77
CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets	77
Collaboration Papers	129
T-cell acute leukaemia exhibits dynamic interactions with bone marrow microenvironments	131
High dimensional single cell analysis predicts response to anti-PD-1 immunotherapy	151
Discussion and Perspectives	179
1 Dirichlet-multinomial regression framework	180
2 Transcript-level analysis with the beta-binomial model	185
3 Incorporating uncertainty of transcript abundance estimates	186
4 DRIMSeq application to other types of multivariate data	187
References	187

Introduction

1 Background

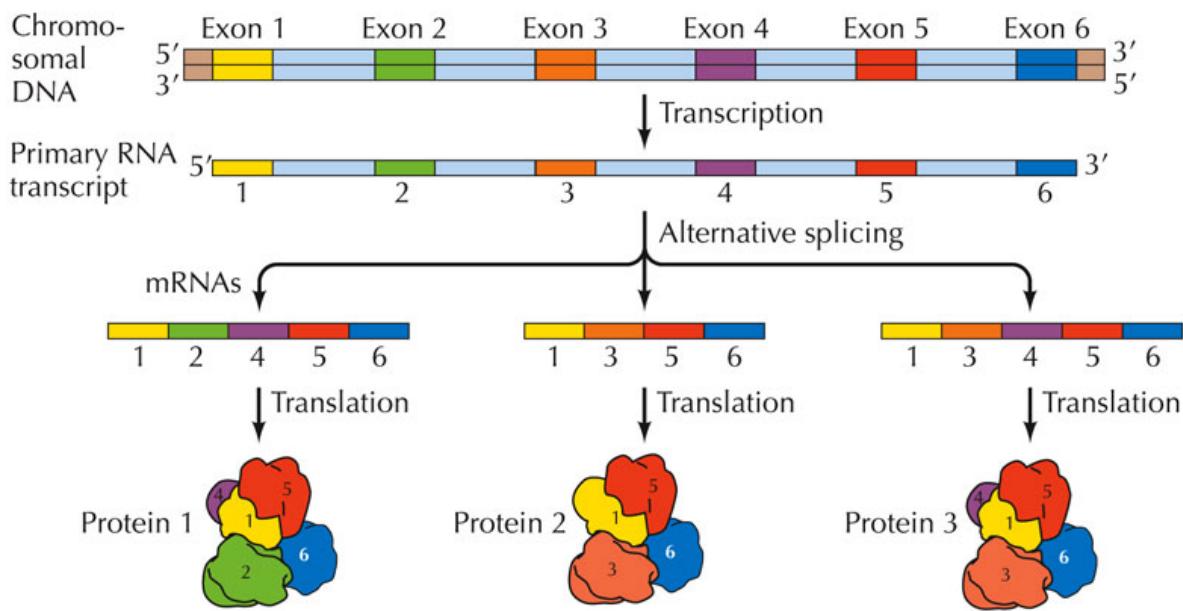
1.1 Biological background

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are molecules that are essential for all known forms of life. DNA carries the genetic information of organisms, which contains instructions for all of the life processes, and RNA plays an active role during the protein synthesis. DNA is composed of the paired chains (strands) of nucleotides (adenine, cytosine, guanine, and thymine), each typically of millions nucleotides long. Each strand has a polarity: they start with 5' and end with 3'. Since the two strands are complementary one will run from 5' to 3' and the other from 3' to 5' (see Figure 1). The nucleotides are paired in a complementary fashion: guanine always pairs with cytosine and adenine with thymine. The precise order of nucleotides in the chains encodes information. A short stretch of DNA that encodes for a protein is called a gene. In complex organisms, genes consist of sections called exons that code for proteins interspersed with non-coding section called introns that can be involved in gene regulation. RNA, on the other hand, is only 75-5000 nucleotides long, and usually is found in nature as a single-stranded chain of nucleotides in contrary to the double-stranded DNA and is less stable than DNA. RNA can also bind with a single strand of DNA, however RNA contains uracil instead of thymine. There are several types of RNA present in cells, for example, messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA).

The central dogma of molecular biology introduced by Francis Crick [1] defines the genetic information flow between DNA, RNA and protein within a biological system. DNA is copied to DNA during DNA replication, DNA information is copied into mRNA during transcription, and during translation, proteins are synthesized using the information from mRNA as a template. Collectively, the process of converting the DNA information into RNA and then from RNA into a protein is referred to as gene expression (see Figure 1).

In eukaryotes, a single gene can express diverse mRNA variants due to the differences in transcription start sites and polyadenylation sites or as a consequence of alternative splicing, which takes place in roughly 90% of human genes [2]. Alternative splicing is a regulated process taking place during the gene expression by which exons may be included or excluded from mRNA transcripts, resulting in different mature mRNA isoforms from a single gene locus (see Figure 1). The differences in mRNA isoforms composition, may affect their stability, localization or translation. Furthermore, the synthesized proteins may contain differences in their amino acid sequence, which can influence their biological function. Hence, alternative splicing contributes to the proteome complexity [3].

For a particular organism, the DNA content is the same in most of the cells. However, the amount of mRNA transcribed and proteins translated from this mRNA varies substantially between cells and also varies within a cell under different conditions. In the fields of biology and medicine, it is of interest to study the differences in gene expression between various cell types, tissues, experimental or environmental conditions, as they allow for understanding the function of genes. Ideally, the study of gene function would be done by exploring the function of proteins, as they are the final product of gene expression and are directly linked to the phenotype of an organism. However, as it is easier to handle and measure nucleic acids than proteins, most gene expression studies are based on the investigation of mRNA, assuming that the amounts of mRNA reflect the protein content. The mRNA expression can be measured



THE CELL, Fourth Edition, Figure 5.5 © 2006 ASM Press and Sinauer Associates, Inc.

Figure 1.: Scheme illustrating gene expression involving alternative splicing. Colored blocks represent exons, whereas light blue blocks represent introns.

genome-wide with DNA microarrays or with the more recent RNA-seq technology described in the following sections.

One of the main types of comparative analysis, based on the transcriptomic measurements, is differential gene expression (DGE), where the goal is to compare the overall expression of each gene between conditions. This can be further stratified into the differential transcript expression (DTE) analysis, where expression of each individual mRNA transcript is investigated. While the overall gene expression in two conditions may be the same, changes in the relative ratio of expressed transcript isoforms within that gene can have significant phenotypic consequences and its aberrations can be associated with various diseases [2, 4, 5]. Such relative-level analyses can aim at various subunits of genes to be compared and lead to differential transcript usage (DTU) or differential exon usage (DEU) analyses. Some of the methods focus on investigating specific splicing events, such as exon skipping or mutually exclusive exons. A summary of these methods is presented in Figure 2.

Studying differences in gene expression and alternative splicing patterns between experimental conditions is biologically meaningful because it can explain a variety of observed phenotypes. In addition, natural variation in gene expression and splicing patterns is present and occurs due to stochasticity and differences in DNA sequences. Genetic variation is an

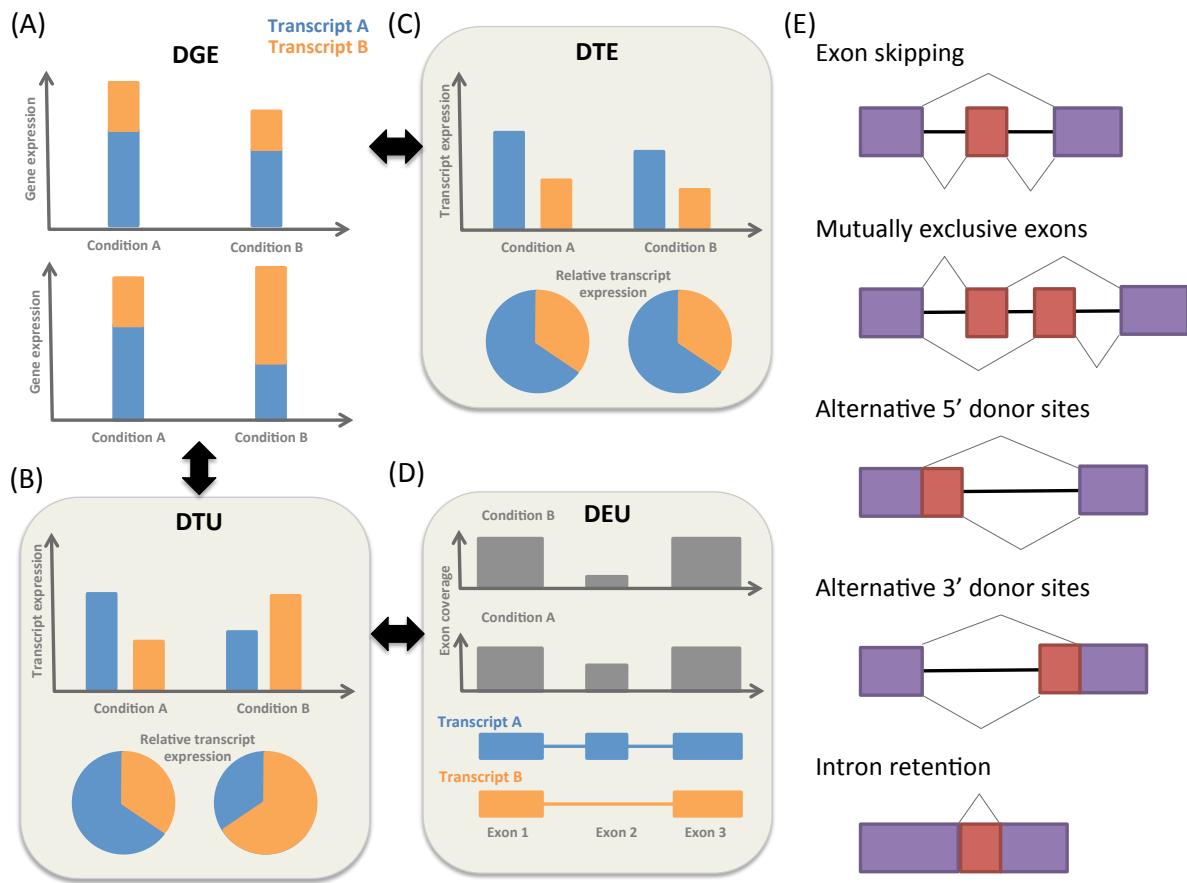


Figure 2.: Schematic illustration presenting types of comparative analyses based on transcriptomic measurements. (A) Differential gene expression. (B) Differential transcript usage. (C) Differential transcript expression. (D) Differential exon usage. (E) Examples of most commonly recognized modes of alternative splicing (splicing events).

important determinant of human phenotypic variation, however some DNA mutations may alter normal splicing and cause various diseases and disorders [5, 6] or affect the response to drugs [7]. Thus, genome-wide discovery of genetic variants that mediate alternative splicing is another step to deeply understand the regulatory processes taking place in cells and consequently, it can lead to greater understanding of complex traits and the causes or origin of disease. By combining transcriptomic measurements with the genotype information monitored in large cohorts, one can identify single nucleotide polymorphisms (SNPs) that are associated with a phenotype (quantitative trait) of interest, known as quantitative trait loci (QTL). When the phenotype is gene expression, the identified variants are called eQTLs, when associations with alternative splicing are interrogated, they are called sQTLs or asQTLs. As for differential analysis, the first QTL studies employed microarrays [8, 9]. With the introduction of RNA-seq, large scale RNA-seq datasets for individuals that were also genotyped have become available via the GEUVADIS [10] and GTEx [11] projects. The extensive characteristics of genomic variants were first obtained within the HapMap [12] project which currently is overtaken by the 1000 Genomes project [13].

There are many mature statistical tools and frameworks that perform any of the differential analyses mentioned above based on DNA microarrays or RNA-seq data and several of the

key methods are described in this thesis. Some of the principles and concepts developed for improving differential inference in microarrays translate to RNA-seq (e.g., moderation of variance parameters), despite the fact that abundances are measured as counts in RNA-seq and as continuous intensities in microarrays. However, in many instances, these differences posed additional challenges that had to be accounted for (e.g., the dispersion-mean trend in RNA-seq data). Currently, we can say that RNA-seq has overtaken microarrays, as it is more versatile, and the protocols and analysis pipelines are well established and allow known and non-model organisms to be studied at comparable costs to microarrays. In general, most of the differential methods are based on testing a null hypothesis for each gene or genomic feature of interest: that there is no difference in expression of that unit between the compared conditions. The hypothesis is rejected when there is a sufficient evidence for the differences between the two conditions, i.e., the differences are not due to the experimental variability which one can expect between different samples from the same treatment group. The result of the test is reported with a p-value, which is a probability of observing at least as strong difference between the two conditions assuming the null hypothesis is true.

This thesis is dedicated to differential analysis based on three types of high-throughput data, and the following sections provide some necessary information about them. First, we concentrate in on DNA microarrays and the differential gene expression analysis. Next, we move to RNA-seq, where our main focus is on the differential transcript usage analysis. At the end, we introduce the high dimensional (mass and flow) cytometry (HDCyto) experiments. In contrast to DNA microarrays and RNA-seq, which quantify transcriptomic content of cells, cytometry techniques directly measure protein abundance on the surface or inside cells. The differential analysis based on this type of data involve association of cell type abundance with a phenotype, or changes in signaling markers over all measured cells or within specific subpopulations.

1.2 DNA microarrays

One of the first technologies that enabled simultaneous measurement of the expression of tens of thousands of genes in a particular tissue or cell type were DNA microarrays, or gene chips, which were invented in the 1990s. A microarray device is basically a solid surface with specifically designed microscopic spots of DNA attached. There are many types of microarrays with varying manufacturing techniques and probe characteristics. For example, the cDNA arrays [14, 15], where probes corresponding to the entire transcripts are synthesized prior to being "spotted" onto glass, or high-density oligonucleotide arrays produced by Affymetrix [16, 17], where short oligonucleotide sequences are printed directly onto the array surface instead of being deposited as an intact sequence.

The main principle of microarray technology is hybridization, i.e., the phenomenon where single-stranded DNA or RNA molecules anneal to their complementary DNA or RNA on the chip. Thus, a piece of DNA can be used for finding and binding its matching sequence even in complex mixtures of millions of unrelated pieces of DNA. In microarrays, this observation is used for specific detection and quantitative measurement of targeted DNAs or RNAs. Furthermore, living systems do not discriminate between DNA synthesized chemically and natural DNA allowing researchers to use custom made oligonucleotides (short single-stranded molecules of DNA) in the microarray construction. However, the DNA sequence of the investigated organism must be known prior to the experiment, and DNA fragments complementary to the known targets designed, which makes the microarrays technology suitable mainly for

model organisms.

We focus on the description and analysis of the Affymetrix GeneChip array, as this was a microarray used by our collaborators. Moreover, this type of microarray dominated the market before microarray technology was surpassed by RNA-seq. On the Affymetrix GeneChip array, the oligonucleotides are complementary to the particular sections of target genes and are designed to be 25 bases long. Those 25-mers are called probes. Because of the short probe length, each gene is represented by 11 to 20 oligonucleotides, referred to as probesets, which aims at improving the specificity. One array can contain between 12,000 and 22,000 probesets. Probes are chosen from all transcript sequences such that they uniquely match only one transcript and have many mismatches for all other transcripts. Nevertheless, in reality, cross-hybridization (or non-specific hybridization), where single-stranded DNA sequence binds to a probe sequence which is not completely complementary, occurs and is unavoidable. Therefore additionally to the perfect match (PM) probes, which are exactly complementary to the sequence of interest, Affymetrix arrays contain another set of probes intended to measure the non-specific binding that can be used to estimate the probe-specific background. Initially, those were the mismatch (MM) probes. Each PM probe was paired with a MM probe that was created by changing the middle (13th) base. A PM and its corresponding MM probe were referred to as a probe pair. In the following generations of arrays, the MM probes were replaced with the antigenomic background probes: probes that were not present in any of the eight genomes (human, mouse, rat, Drosophila, C. elegans, S. cerevisiae, Arabidopsis and E. coli) and thus were not expected to cross-hybridize to transcribed sequences of interest. As hybridization strength differs with GC content, the background probes were designed so they varied in GC content from zero GCs out of a 25-mer sequence, to all 25 bases being GC bases. Each category was represented by approximately 1,000 probes.

The general workflow of microarray experiments consists of: i) extraction of the RNA of interest (total or ribosomal RNA or poly-A mRNA) from biological samples, ii) reverse transcription of mRNA into double-stranded complementary DNA (cDNA), iii) production of biotin-labeled cRNA from the cDNA, iv) fragmentation into typically 25-200 bases pieces, v) hybridization on a microarray for as long as 16-20 hours, vi) removal of non-hybridized cRNA, vii) reading of the fluorescent intensities with a laser that creates an image of emission levels by exciting the fluorescent dye and a detector that quantifies this image. It is expected that a higher amount of cRNA for a particular gene in the hybridization cocktail should result in more material attached to the probes corresponding to that gene, and subsequently the emitted signal should be brighter at locations where more cRNA has hybridized. However, the varying affinity of probes allows for relative measurements only, and no absolute quantification of transcript abundance is possible. Thus, one can not compare the measurements of different probes and numerical comparisons are only possible for the same probe on different arrays assuming that the probe affinity does not change from experiment to experiment.

Affymetrix GeneChip is a single-channel system. To compare gene expression between two conditions (e.g. diseased versus healthy tissue), two separate single-dye hybridizations are required. However, there are also two-color or two-channel microarrays, where cDNA prepared from two samples and colored with two different dyes (usually green and red) is mixed and hybridized on a single microarray. Intensities of the two fluorophores are measured and their relative expression can be then used in ratio-based analysis to identify up- and down-regulated genes. Usage of the two-colored arrays does not scale when the interest is

in comparing many samples. It is better to use the single-dye systems where each sample is hybridized separately, which allows also for studying more complex experimental designs.

In many applications, GeneChip microarrays were mainly used for gene expression profiling. However, a study by Robinson and Speed [18] showed that the Affymetrix Gene 1.0 ST platform could be also used to identify differential splicing events. There are also microarrays specifically designed to study different splicing isoforms and exon or transcript usage, such as the exon, exon junction or tilling microarrays.

1.3 DNA microarrays analysis

1.3.1 Preprocessing: background adjustment, normalization, summarization

Preprocessing of Affymetrix expression array results usually involves three steps: background adjustment, normalization and summarization; and can have a dramatic effect on differential expression results. Background correction, also referred to as background subtraction or signal adjustment, attempts to adjust for cross-hybridization and any other sources of background noise occurring during array processing. Normalization refers to the task of manipulating data so that the measurements from different arrays are comparable. Summarization of the multiple probe intensities for each probeset to produce an expression value is the final stage in preprocessing of Affymetrix GeneChip data. Various methods have been proposed in the microarray literature to accomplish those three steps for GeneChip arrays [19, 20, 21]. We describe those used by the Robust Multi-array Average (RMA) preprocessing methodology [19, 22, 23] (the *rma* function from the *oligo* package). RMA consists of the three particular steps: convolution background correction known also as normexp [24, 25], quantile normalization and summarization based on the robust multi-array average algorithm.

Background adjustment

In the Affymetrix arrays, the MM probes were originally introduced to adjust the PM probes by subtracting intensities of MM probes from the intensities of the corresponding PM probes. However, this procedure was problematic, as, in a typical array, almost 30% of the MM probes had intensities higher than their corresponding PM probes, leading to negative values. This in turn caused problems in data scaling with logarithms, which were proven to be useful for microarray data transformation. The RMA convolution model and its adaptation, the normexp procedure, ignore the MM intensities. The PM values are corrected, separately for each array, using a global model for the distribution of the observed PM probe intensities, which is motivated by the empirical distribution of the PM probe intensities. The observed intensities I are assumed to be a sum of a normally distributed background component B and an exponential signal component S : $I = B + S$. The normal distribution is truncated at zero to avoid the possibility of negative expression values. Under such a model, the background corrected PM probe intensities are given as the conditional expectations of the signals given the observations $E(S|I)$. As a consequence, the largest relative adjustments are applied to the smallest intensities, the order of probe intensities stays invariant and the corrected values are always positive. In the RMA algorithm, parameter estimation uses an ad hoc density kernel method, while the normexp implementation in the *limma* package maximizes the saddle-point approximation to the likelihood or directly the exact likelihood. An additional preprocessing step can be applied to the recent Affymetrix arrays that do not use the MM probes at all in favor of the anti-genomic probes. The latter can be used for the assessment of the post-normalization

background level over different GC content probes. Subsequently, probesets with intensities that are close to the background values can be removed from the differential analysis.

Normalization

There are various sources of obscure variation, due to, for example, sample preparation (total amount of RNA used) and processing of the arrays including labeling (dye-incorporation efficiency), hybridization, stringency of the washing and scanning, which can have many different effects on the observed intensities. Comparison of data from different arrays can lead to misleading results, unless arrays are appropriately normalized. Various methods have been proposed for normalization of GeneChip arrays, and the quantile normalization was found to perform best among them [19]. Quantile normalization imposes identical empirical distributions of intensities to each array, and is applied to the intensities at the probe level. It is model-free, i.e., it makes no assumptions on the cause of the biases that are removed. However, as for most of the normalization methods, to work properly, two general assumptions should be met: the majority of genes should be unchanged, and the amount of up- and down-regulated genes should be approximately the same.

Summarization

The key features of the RMA procedure are that it does not average the intensities of a probeset from a single sample, but fits a model using the measurements from all samples. Moreover, it takes into account that there is a strong probe effect within a probeset. To get a summary expression measure for a probeset, RMA models the background-adjusted, normalized, and log-transformed PM signal, denoted with Y , as a function of the probe and the sample. By fitting a linear additive model to the log-intensities, the model assumes a multiplicative probe effect with a multiplicative error:

$$Y_{ij} = \mu_i + \alpha_j + \epsilon_{ij}, \quad (1)$$

where α_j represents the affinity effect for probe $j = 1, \dots, J$, μ_i represents the log-scaled expression level for array $i = 1, \dots, I$, and ϵ_{ij} represents an independent identically distributed error term with mean 0. For identifiability of the parameters, it is assumed that $\sum_{j=1}^J \alpha_j = 0$, such that probes are chosen in a way that they represent on average the associated gene expression. The regression parameters are computed in a robust way, to protect against outlier probes and samples, using the median polishing algorithm.

1.3.2 Differential analysis

One of the most popular, flexible and powerful frameworks for the analysis of gene expression microarray data is *limma* [26]. It can handle both single- and two-channel microarrays, and it also provides frameworks for the analysis of data arising from the RNA-seq technology. The core feature of *limma* is usage of linear models to assess differential expression, which allows for analysis of experiments with arbitrarily complex designs with a variety of experimental conditions and predictors. *limma* enables simultaneous analysis of many genes and provides stable analyses, even when the sample size is small, by sharing information across genes with the Empirical Bayesian methods.

1.4 RNA-seq technology

RNA sequencing (RNA-seq) [27] refers to techniques used to determine the sequence of RNA. It employs high-throughput sequencing of cDNA molecules obtained by reverse transcription of RNA into DNA. Obtained sequences are then mainly used to assess the relative abundance of each RNA molecule in a biological sample.

A typical RNA-seq experiment includes: i) capture of mRNA subpopulation of interest (poly-A-enriched or ribo-depleted), ii) fragmentation of long RNA transcripts into 200-500 base long pieces, iii) reverse transcription of RNA into double-stranded cDNA primed by random hexamers, iv) sequencing using DNA sequencing protocols. Most commonly, the mature RNA-seq experiments produce millions of reads that can be sequenced from one (single-end reads) or both (paired-end reads) ends of DNA fragments, and are between 50 and 150 bases long. The subsequent analyses use the generated reads in various ways depending on the biological question of interest. Often, the goal is to quantify genes or other genomic features, such as transcripts or exons, so their expression or usage can be compared between conditions in differential analysis. Usually a step preceding quantification is mapping of the reads to the reference genome or transcriptome for organisms with existing annotation. The read coverage can then be used to quantify expression levels of features of interest. When the annotation is not available or is incomplete, RNA-seq reads can be also used to generate a catalog of transcripts via *de novo* assembly [28].

Various criteria should be carefully considered when designing an RNA-seq experiment, such as library type (single-end or paired-end), length of reads, sequencing depth and the number of replicates [29]. All of them can have a substantial impact on the following analysis and often a tradeoff between one or the other needs to be made for cost reasons.

RNA-seq has clear advantages over DNA microarrays and has revolutionized the manner in which eukaryotic transcriptomes are analyzed. Unlike microarrays, RNA-seq technology does not require knowledge of the genome annotation prior to the experiment, as it does not rely on predetermined probe sequences. Thus, it is compatible with any species and especially suitable for non-model organisms. It provides quantification at single-base resolution. RNA-seq is especially useful for discovery of novel transcripts, gene fusions, single nucleotide variants, indels (small insertions and deletions), splice junctions, allele-specific expression, and other previously unknown changes that arrays cannot detect [30]. In the microarray hybridization technology, the accuracy of gene expression measurement is limited by high background effects and non-specific binding at the low end of expression, and signal saturation of binding sites within the probe sets at the high end [31]. RNA-Seq technology offers a broader dynamic range, as it quantifies discrete, digital sequencing read counts and the sequencing coverage depth can easily be increased to improve the sensitivity of rare and low-abundant transcript detection. It avoids biases related to probe selection and their varying hybridization properties. However, RNA-seq is not free from biases either. Each step of the RNA-seq protocol may introduce biases into the resulting data [32]. Nevertheless, RNA-seq outperforms microarrays in gene expression profiling, mainly due to its improved accuracy for lowly abundant transcripts [33] and ability to discern splicing events.

1.5 RNA-seq data analysis

In this section, we concentrate on the workflow for the DTU analysis where it is assumed that the reference genome or transcriptome are available. Such a workflow typically consists of: quality control (QC) steps, read mapping/alignment, quantification and differential analysis.

Quality control

RNA-seq data can suffer from various sample-specific biases as a result of RNA extraction, library preparation and sequencing steps. Spatial biases which exist along the genome, caused by RNA degradation, mRNA selection techniques, size selection, differential binding efficiency of random hexamer primers or PCR artifacts, result in a non-uniform coverage of expressed transcripts, which poses a challenge for quantification methods [34]. Quality control checks at various steps of RNA-seq data analysis allow for identification, monitoring and potential removal of some of the biases. Quality control assessment of raw reads involves investigation of sequence quality, sequence content and presence of adaptors and duplicated reads in order to detect sequencing errors, GC content or PCR artifacts or contaminations. *FastQC* [35] and *NGSQC* [36] are the two most popular tools to perform such analyses. In general, the quality of reads drops towards the 3' end. In the case it becomes too low, to improve mappability, the poor-quality bases of reads can be trimmed with, for example *Trimmomatic* [37]. Once the reads are mapped to the reference genome or transcriptome, one can check the mapping quality which is reflected by the percentage of mapped reads and the degree of read multi-mapping. As mentioned already, it is common to observe variable coverage of RNA-seq fragments along transcripts due to the technical specifications of the RNA-seq experiment. Different quantification methods try to account for these biases [34]. Thus, it may be useful to verify the GC content and gene/transcript length biases of the quantified gene/transcript expression so it could be corrected if necessary.

Read alignment

Two alternatives are possible when a reference annotation is available: RNA-seq reads can be mapped to the genome or to the transcriptome. This task can be challenging due to the relatively short read lengths and constantly increasing throughput of the sequencing technologies. Additionally, two key requirements make read alignment computationally intensive. First, an alignment algorithm should allow mapping of reads containing a certain amount of mismatches, insertions and deletions, which may be caused by natural sequence variations, sequencing errors or biases created during library preparation, such as RNA priming bias. Second, it should be able to identify splice junctions and map RNA fragments that due to splicing consist of sequences that correspond to non-contiguous genomic regions. The latter challenge is specific for RNA-seq (when aligning to the genome) and crucial for studying alternative splicing and transcript structure. It can be avoided by mapping to the transcriptome at the risk of missing unannotated isoforms.

Mapping of RNA-seq reads to the transcriptome is equivalent to aligning DNA sequences to genome, except that more multi-mapping reads may be observed, as exonic reads will map to all the transcript isoforms containing those exons, and can be carried out with the well known short DNA (contiguous) sequence aligner *Bowtie* [38]. *TopHat* [39] spliced read mapper for RNA-seq was developed as an extension of *Bowtie*, and is able to align RNA-seq reads to the

genome without relying on known splice junctions. It can identify splice junctions without referring to an external database. In the first step, it maps RNA-seq reads to the genome and identifies potential exons, since many RNA-seq reads (non-junction reads) will be contiguously aligned to the genome. Based on this information, *TopHat* identifies the splice junction structures and then maps the initially unmapped reads against those junctions to confirm them. In contrast to the above approach, *STAR* (Spliced Transcripts Alignment to a Reference) [40] is able to align the non-contiguous sequences directly to the reference genome. It also showed a great improvement in computational speed in aligning millions of reads [41]. Additionally, it provides estimates of the relative probabilities of the alignments for reads that map to multiple loci.

The newest generation of transcript quantification tools (*Sailfish* [42], *kallisto* [43] and *Salmon* [44]) do not require read mapping to the transcriptome or the genome. Instead, these tools rely on a pseudo-mapping of the k-mers present within each read to the k-mer distributions from the transcript annotation. The expression of each transcript is then directly inferred with an expectation maximization algorithm. As consequence, alignment-free methods are much faster than traditional alignment-based approaches.

QTL analysis

The general outline of a QTL analysis is to define the phenotype (quantitative trait) and test for the correlation with genotype of the nearby variants. Usually, the genetic variants are narrowed to bi-allelic SNPs located in the surrounding of the gene that the association is tested with, and their genotype is translated into the number of minor alleles (0, 1 or 2). Typically, associations between SNPs and phenotype are interrogated by model fitting and testing, with the independent variable defined by the genotype and performed for each gene-SNP pair. In standard differential analysis, independent variables are defined by the design of the experiment and are the same for each gene. Thus, even though QTL analysis represents a different application, it is essentially the same as differential analysis between groups defined by genotypes. There are a few additional challenges to be handled in QTL analysis. Multiple testing corrections have to account for a large number of tests per gene with highly variable allele frequencies (models) and linkage disequilibrium (non-random association of alleles at different loci). Usually, this is done by employing a permutation approach to empirically construct the null distribution of associations and use it for the adjustment of nominal p-values. Because the sample size is typically much larger, it may be not necessary to share information between genes as is done in standard differential analysis (see Section 1.6.3). The sample size is usually much larger and there are multiple SNPs tested per considered feature resulting in much larger scale of analysis in terms of computing time. Perhaps because of that, rather independent sets of methods were developed for QTL analysis and differential analysis based on samples from different experimental conditions. However, we notice that some of the challenges are common for those two types of analyses, such as defining which phenotype best captures splicing variation. Further, some of the methods (e.g. *LeafCutter* [45]) use the same quantification strategy for sQTL and differential splicing analyses. In the following overview, we do not distinguish between applications but rather between the general concepts used to detect differences in splicing.

1.5.1 Quantification

Most commonly, RNA-seq data consists of sequences of short reads (75-150bp) that originate from the cDNA fragments, which are the result of RNA fragmentation. In RNA-seq, the measurement units are not defined in advance, and the set of reads can be used to estimate expression of many different types of features. Gene expression levels are usually represented as a single value. However, there are a variety of ways to represent the phenotype of an alternatively spliced gene. These can be encapsulated into three main categories: relative transcript usage, or in a more local context, exon or exon junction usage, or via the presence or absence of specific splicing events (e.g., exon skipping), and all have their advantages and disadvantages.

One of the most straightforward approaches to abundance quantification is based on counting the aligned reads overlapping the target genomic regions (e.g., *HTSeq* [46] and *featureCounts* [47]). This approach can be used to estimate the overall gene expression by counting the read-gene overlaps. To study alternative splicing, the target regions can be defined as exons or exonic bins: non-overlapping windows obtained by projecting all exons to the linear genome. However, this strategy does not utilize the full information from junction reads. Such reads are counted multiple times (in all exons that they overlap), artificially increasing the total number of counts per gene and ignoring that junction reads support the isoforms that explicitly contain the combinations of exons spanned by these reads.

Quantifying other kinds of features can preserve this information. For example, one could quantify exon-links (exon junctions) with *Altrans* [48], or calculate splicing event inclusion levels expressed as percentage spliced in (PSI) with *MISO* [49], *rMATS* [50], *SUPPA* [51] and *SGSeq* [52]. Such events capture not only cassette exons but also alternative 3' and 5' splice sites, mutually exclusive exons or intron retention. *GLiMMPS* [53] and Jia *et al.* [54], with quantification from PennSeq [55], use event inclusion levels for detecting SNPs that are associated with differential splicing, while *LeafCutter* [45] quantifies intron usage to solve the same problem. Nevertheless, there are (hypothetical) instances where changes in splicing pattern may not be captured by exon-level quantifications (for example, see Figure 1A in the paper by Monlog *et al.* [56]). Furthermore, detection of more complex transcript variations remains a challenge for exon junction or PSI methods (see Figure S5 in the paper by Ongen *et al.* [48]). Soneson *et al.* [57] considered counting that accommodates various types of local splicing events, such as exon paths traced out by paired reads, junction counts or events that correspond to combinations of isoforms; in general, the default exon-based counting resulted in strongest performance for DTU gene detection.

The above methods allow for detection of differential usage of local splicing features, which can serve as an indicator of differential transcript usage for a given gene but often without identifying specifically which isoforms are differentially regulated. This can be a disadvantage in cases where knowing the isoform ratio changes is important, since isoforms are the ultimate determinants of proteins. Moreover, exons are not independent transcriptional units but building blocks of transcripts. Thus, the main alternative is to make a calculation of differential splicing using isoform-level quantitations. A vast number of methods are available for gene isoform quantification, such as *MISO* [49], *BitSeq* [58], *casper* [59], *Cufflinks* [60], *RSEM* [61], *FlipFlop* [62] and more recent, extremely fast pseudoalignment-based methods, such as *Sailfish* [42], *kallisto* [43] and *Salmon* [44]. Additionally, *Cufflinks* allows for *de novo* transcriptome assembly, and *casper* and *FlipFlop* can identify the structure of expressed isoforms.

However, it remains a complex undertaking to quantify isoform expression from short cDNA fragments due to a high degree of overlap between transcripts in complex genes. In the case of incomplete transcript annotation, local approaches may be more robust and can detect differential changes due to transcripts that are not in the catalog. Transcript-level estimates can be represented in different units, such as RPKM (reads per kilobase per million reads), FPKM (fragments per kilobase per million reads) and a more recent TPM (transcripts per million), which is preferable, as it is more consistent across libraries. Introduction of such units aims to account for the impact of sequencing depth and feature length on the observed number of reads. However, the influence of those aspects on the quantification uncertainty is masked. To account for these characteristics, one could use the expected counts as input to the differential methods that are able to operate on counts. Though, the effect of using fractional counts resulting from partitioning reads aligning to multiple transcripts is still unknown. Notably, transcript-level abundance estimates can be summarized to gene-level abundances with *tximport* [63] leading to improved DGE detection as compared to using simple read-gene counting approach, especially for genes exhibiting DTU.

Despite the existing limitations in transcript abundance estimation, we believe that studying differential splicing at the resolution of isoforms is the ultimate goal. Thus, we developed the *DRIMSeq* [64] framework for DTU analysis presented in this thesis. With the emergence of longer reads, transcript quantifications will become more accurate and methods for multivariate transcript abundances, like *DRIMSeq*, will be needed even more.

1.5.2 Differential analysis

There are various methods designed to detect differential usage of gene features resulting from differential splicing, and they can be discriminated by the two dimensions of choice: what to compare (e.g., exons, transcripts) and how. Each method's design was mostly driven by the quantification input that it analyzes (transcripts, exons, or events). However, some of the methods are more flexible than others and their differential engines may effectively be used to analyze multiple types of input. For example, *DEXSeq* [65] and *voom-diffSplice* [66], which originally were designed to model exonic counts, can accept any table that contains the counts of reads falling into preferred features across all samples [57]. In general, differential usage analysis takes place when there are multiple subunits of a gene and one is interested in detecting whether their usage ratios vary between experimental conditions. The existing methods often treat each of the subunits separately or consider them all together as a multivariate output, and have different strategies to account for the overall expression effects.

Some of the first sQTL approaches model and test each transcript [10, 67] or exon [68] ratio separately. *Altrans* performs sQTL analysis separately for each of the exon junctions via a correlation-based approach with *FastQTL* [69] and *LeafCutter* for each intron usage quantification. *rMATS*, *GLiMMPS*, Jia *et al.* [54], Montgomery *et al.* [70] model and test each of the splicing events of a gene. Such approaches ignore the correlated structure of these quantities and lead to non-independent statistical tests, although the full effect of this on calibration (e.g., controlling the rate of false discoveries) is not known.

DEXSeq and *voom-diffSplice* undertake another approach, where the modeling is done per gene. *DEXSeq* fits a generalized linear model (GLM), assuming that (exonic) read counts follow the negative-binomial distribution. A bin is deemed differentially used when its cor-

responding group-bin interaction is significantly different from zero. The exact details of *voom-diffSplice* are not published. Nevertheless, exons are again treated as independent in the gene-level model.

MISO, *Cuffdiff* [71] and *sQTLseekerR* [56] model transcript usage as a multivariate response. *MISO* is designed for DS analyses only between two samples and therefore does not handle replicates. Variability among replicates is captured within *Cuffdiff* via the Jensen-Shannon divergence metric on probability distributions of isoform proportions as a measure of changes in isoform relative abundances between samples. *sQTLseekerR* tests for the association between genotype and transcript composition, using an approach similar to a multivariate analysis of variance (MANOVA) without assuming any probabilistic distribution and Hellinger distance as a dissimilarity measure between transcript ratios. *LeafCutter* performs the DS analyses on intron usage quantifications using the Dirichlet-multinomial model, similar to our *DRIMSeq* model.

sQTLseekerR, *Altrans*, *LeafCutter* and other earlier methods for the sQTL analysis employ feature ratios to account for the overall gene expression. A potential drawback of this approach is that feature ratios do not take into account whether they are based on high or low expression, while the latter have more uncertainty in them.

The statistical framework, called *DRIMSeq*, that we propose within this thesis, for discovering changes in isoform usage between conditions and SNPs that affect relative expression of transcripts using transcript-level quantification is based on the Dirichlet-multinomial distribution. The Dirichlet-multinomial model naturally accounts for the differential transcript usage without losing information about overall gene abundance and by joint modeling of isoform expression, it has the capability to account for their correlated nature.

1.6 Strategies for modeling counts in small-sample size RNA-seq data

There are some key concepts that have been well grounded in modeling of RNA-seq data. Some of them were developed for microarrays, such as sharing information between genes, but their principles can be translated into RNA-seq and improve differential inference.

1.6.1 Two-stage estimation

The negative binomial (NB) distribution is a well established model for gene expression counts from RNA-seq data and is used by mature software packages for differential gene expression, such as *edgeR* [72, 73] or *DESeq* [74] and its successor *DESeq2* [75]. In this setting, the gene expression counts Y_{gi} for gene g and sample i are assumed to follow a NB distribution with mean μ_{gi} and dispersion ϕ_g , denoted as $Y_{gi} \sim NB(\mu_{gi}, \phi_g)$. The negative binomial is a result of a Poisson-gamma mixture, i.e., it can be seen as a Poisson distribution with the rate parameter being itself a gamma distributed random variable. This takes into account the biological and technical variability that may cause differences in the relative abundance of genes between different RNA samples resulting in over-dispersion with respect to Poisson. The construction of the Dirichlet-multinomial (DM) distribution in *DRIMSeq* is based on a similar principle. The Dirichlet-multinomial is a hierarchical model where the transcript counts of a gene are assumed to follow a multinomial distribution with proportion parameters being random variables from a Dirichlet distribution, which allows the modeling of the over-dispersion observed

in RNA-seq data. Thus, the transcript counts $\mathbf{Y}_{gi} = (Y_{gi1}, \dots, Y_{g iq})$ for sample i and gene g with q transcripts and m_{gi} being the total counts of expressed transcripts are assumed to follow a DM distribution with proportions $\boldsymbol{\pi}_{gi} = (\pi_{gi1}, \dots, \pi_{g iq})$ and concentration γ_{g+} , denoted as $\mathbf{Y}_{gi} \sim DM(m_{gi}, \boldsymbol{\pi}_{gi}, \gamma_{g+})$.

A parametrization of the negative binomial model defined by its mean μ_{gi} and dispersion ϕ_g is extremely useful in genomics. By separate estimation of these parameters, one is able to perform various manipulations to the dispersion parameter, such as moderation. Moderating dispersion is one of the principal features used in genomics that allows methods to effectively analyze experiments with small number of replicates. The frameworks in *edgeR*, *DESeq* and *DESeq2* are constructed as a two-stage estimation. First, estimating the dispersion and then the generalized linear model (GLM) coefficients that are directly linked to the mean. The same strategy is also used in *DEXSeq* [65] designed for differential exon usage analysis. In the *DRIMSeq* implementation, we use a parameterization of the Dirichlet-multinomial that represents proportions $\boldsymbol{\pi}_{gi}$ and a concentration (also called precision) parameter γ_{g+} that is inversely proportional to the dispersion. This enables the two-stage estimation of concentration, with adjustments and moderation, in the first place, and then the estimation of proportions.

1.6.2 Adjusted profile likelihood

In the first step of the two-stage estimation for NB, the aim is to estimate the dispersion parameter, thus the regression coefficients become nuisance parameters. In DM, the concentration is a parameter of interest, and proportions are nuisance. Reliable inference in the presence of nuisance parameters is a widely encountered, but a difficult, problem. Estimation of nuisance parameters can strongly affect inference for the parameter of interest. Thus, one could try to produce a version of the likelihood that does not depend on the nuisance parameters, such as marginal (if there exists a distribution for the nuisance parameters, the marginal distribution for the parameter of interest could be calculated) or conditional likelihood (using a sufficient statistic for the nuisance parameters; conditioning on this statistic results in a likelihood that does not depend on nuisance parameters). However, often their construction is difficult or even impossible. Conditional likelihood with a sum of observed counts as a sufficient statistic was used to estimate the NB dispersion in the per group comparison, and showed good performance [76, 77], but its implementation within a GLM framework is infeasible.

An alternative approach is to construct so-called profile likelihood (PL) by maximizing out the nuisance parameters for fixed values of the parameters of interest. The profile likelihood is then treated as an ordinary likelihood function for inference about the parameters of interest. Unfortunately, with large numbers of nuisance parameters, this procedure can produce inefficient or even inconsistent estimates [78, 79], as standard maximum likelihood tends to underestimate variance parameters by not allowing for the fact that other unknown parameters are estimated from the same data. To reduce the bias introduced by ML in presence of nuisance parameters, one could maximize an adjusted profile likelihood (APL), as proposed by Cox and Reid [80]. The Cox-Reid adjusted profile likelihood is successfully used to estimate the dispersion parameter in the NB model in *edgeR* and *DESeq/DESeq2*.

The definition of Cox-Reid APL requires that the parameter of interest and the nuisance parameters are orthogonal with respect to expected Fisher information. This is not the case for the Dirichlet-multinomial distribution, as it does not belong to the exponential family. Despite

that, we still observe that the Cox-Reid adjustment, which is implemented within *DRIMSeq*, leads to an improved estimation of the concentration parameter, see Paper I.

1.6.3 Sharing information between genes

Due to the limited sample size of typical RNA-seq experiments, it is difficult to get accurate gene-wise (obtained from the data for a specific gene alone) estimators of dispersion. Thus, further efforts to improve the dispersion estimates were necessary. One of the first strategies was usage of common dispersion across all the genes. The estimates obtained by pooling the likelihood of all genes have been shown to be more stable, but they ignore the gene-specific variability. To compromise between the gene-wise and common estimators, a parametric empirical Bayes strategy was introduced in the analysis of microarray data in the context of variance of data from a normal distribution [26, 81]. With this approach, the stability and reliability of variance estimates was improved, especially for small sample size data, by sharing the structure from all the genes while allowing for some flexibility for each individual gene.

For the NB model used for RNA-seq data, the empirical Bayes approach can not be easily implemented, as NB does not belong to the exponential family, and no conjugate prior for the NB dispersion exists. An alternative solution for NB, which is an approximation to empirical Bayes, was proposed. The individual dispersion estimates are squeezed toward the common dispersion via a weighted combination of the common and individual likelihood [72, 77]. In the read counts of RNA-seq, a dispersion-mean trend can be commonly observed. The strategy to moderate dispersion via weighted likelihood can efficiently make use of this dispersion-mean relationship by allowing a certain level of information sharing only between genes with similar average expression, which is referred to as moderation, or shrinkage, toward the trend.

We have implemented the weighted likelihood method in *DRIMSeq* to estimate the concentration parameter. Similarly as for gene level counts, the dispersion-mean trend can be observed when estimating transcript ratios with the DM (see Figure 3). Application of this moderation toward the trend improves the performance of *DRIMSeq* by reducing the false discovery rate (FDR) and increasing the power.

1.7 HDCyto data

Flow cytometry and the more recently introduced CyTOF (cytometry by time-of-flight mass spectrometry or mass cytometry) are technologies that measure protein abundance on the surface or inside cells at high-throughput. We refer to them together as HDCyto (high-dimensional cytometry) experiments. Flow cytometry is a well established method based on labeling antibodies that are conjugated to proteins of interest with fluorescent tags [82]. The intensity of fluorescent dyes is excited with lasers and captured with photodetectors. CyTOF utilizes transition element isotopes, which do not occur in biological systems, to label antibodies, and abundances per cell are recorded with a time-of-flight mass spectrometer [83]. In both cases, the intensity fluorescent dyes or ion counts are assumed to be proportional to the expression level of antibody-targeted proteins of interest.

The differences in acquisition are a reason for substantial differences in the throughput of these methods in terms of the number of protein markers that can be measured and the num-

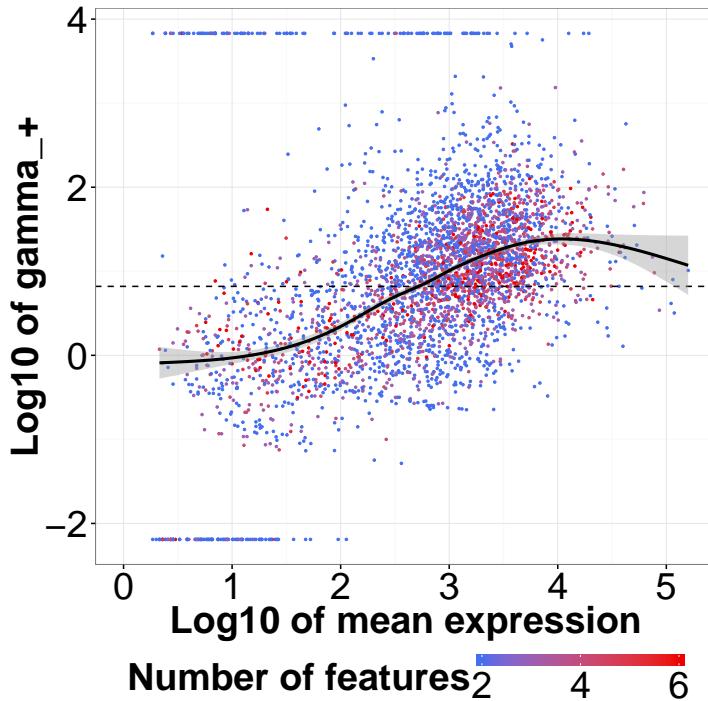


Figure 3.: Concentration parameter estimates versus mean gene expression for the DM model. Each point corresponds to a gene. The fitted line indicates the trend. The DM model was applied to the DTU analysis on the pasilla data based on *kallisto* counts. This Figure corresponds to Figure S31A in the Supplementary Materials of the *DRIMSeq* paper [64].

ber of cells. For flow cytometry, due to the issue of spectral overlap between fluorophores, a compensation step is needed, and the number of markers that can be studied at once is limited to 6-12 in the standard flow cytometry experiments with modern systems measuring up to 20 channels, while new developments promise to increase this capacity towards 50 [84]. By using rare metal isotopes in CyTOF, cell autofluorescence does not take place and the problem of spectral overlap is greatly reduced. However, spill-over still takes place, due to the varying sensitivity of mass spectrometry in measuring of metal impurities and oxide formations, and the compensation step cannot be entirely omitted. Nevertheless, CyTOF offers measurement of a greater number of parameters per cell. Currently, around 40 parameters can be detected and in principle, this number could be increased to more than 100 [85]. The throughput of flow cytometry is much higher than for CyTOF with tens to hundreds of thousands of single cells measured per second as compared to hundreds of cells per second. Also the operating costs per sample are lower in flow cytometry in comparison to CyTOF. In CyTOF, cells are destroyed during ionization while flow cytometry allows for capturing and sorting of the cells and usage for further experiments.

The ability of flow cytometry and mass cytometry to analyze individual cells at such a high-throughput has resulted in a wide range of biological and medical applications. For example, in immunology, the rich deep high-dimensional cell profiles are used to characterize unknown cell populations or to detect and quantify known cell populations of interest allowing researchers to catalog the diversity of cell types [86]. They are used for comparison of population abundance between different conditions, such as different patient groups for discovery of biomarkers [87] or various stimulus conditions to study the system response [88]. HDCyto

data allow the study of cell development, transitioning to different states and mechanisms of cell fate commitment by reconstructing trajectories or time-series experiments.

1.8 HDCyto data analysis

A typical analysis of HDCyto data starts with the preprocessing steps: normalization, debarcoding, compensation and data transformation. Usually, the debris, doublets and dead cells are removed and the subset of cells can be reduced to a population of interest. An important part of the analysis is dimension reduction which is mainly used for visualization but can also be a part of clustering algorithms. Cell population identification of known cell types or identification of new populations can be a final goal per se. Often it can be a basis for the further differential analysis where one wants to compare the abundance of different cell types and identify those that change between conditions. Or, one can study differential expression of specific markers. Some methods, like *CellCnn* [89], combine those two steps and identify only those populations that are differential.

Assembling an analysis workflow for differential analysis of HDCyto data, starting from the preprocessed marker expression values, was one of my main PhD projects. The exact preprocessing steps may vary for flow cytometry and CyTOF, however afterwards the data takes the same format: a table with expression values for markers in columns and cells in rows and can be treated in the same way for the downstream analysis. The key characteristics of the proposed workflow are that the differential analyses are performed using regression frameworks where HDCyto data is the response. Thanks to that, we are able to model arbitrary experimental designs, such as paired experiments or those with batch effects. Moreover, the workflow employs the methods that perform the best to our current knowledge in a modular way such that as new, better tools come along, they can be adapted into it if necessary.

1.8.1 Preprocessing: normalization, debarcoding, compensation, transformation

Preprocessing steps including normalization using bead standards, debarcoding and compensation can be completed with the *CATALYST* [90] R package, which provides an implementation of the debarcoding algorithm described by Zunder et al. [91] and the bead-based normalization from Finck et al. [92]. Alternatively, those steps can also be accomplished outside of R.

Normalization

There are two types of normalization: one that adjusts the data within an experiment, which is especially important for CyTOF where the efficiency of detection drops over time; and one that adjusts the data from different run batches to make it comparable prior to further analysis where the data is combined. Those kinds of normalization are similar to, for example, within-slide and between-slide normalization adjustments in microarrays.

In CyTOF, the ion detection sensitivity of a mass cytometer drifts during the instrument use and changes after each maintenance of the machine. These effects can be accounted for and corrected by applying a normalization technique based on bead standards [92], where polystyrene beads and cells are simultaneously measured by the mass cytometer. The signal variation can be observed in the plot of bead intensities over time. In case of no distortions, it

is expected that bead expression is constant over time. Thus, any differences from the uniform pattern reflect changes in the instrument performance, which presumably affect all the cells. The algorithm enables correction of both short- and long-term signal fluctuations by extracting the bead-based signature, which is interpolated to the cell events. Moreover, the variation in the intensity of the beads that remains after normalization may also be used to assess the data quality.

In flow cytometry, the fluorophore measurements are more stable over time. However, as a quality control step, it is a good practice to plot the scatter and marker values over time. One can filter out regions that show abnormal behavior as a consequence of clogging, speed change or air measurements when the tube is empty.

When the analyses are based on combination and comparison of data originating from different acquisitions, such as multi-center clinical trials, it is necessary to check for batch effects, which may occur due to technical variation in sample preparation and instrumentation differences. If batch effects are present, they should be adequately corrected to make the data comparable, which is essential in the downstream analysis, such as clustering or dimension reduction analysis of pooled data. There are some adjustment methods available that aim to remove the technical between-sample variation, such as equalization of the dynamic range between batches for each marker or usage of warping functions which eliminate non-linear distortions by aligning the landmarks of the raw data [93]. For CyTOF, one could also use the already mentioned bead-based normalization. However, a comprehensive evaluation of these approaches and their effect on subsequent analyses is still missing.

Debarcoding

Multiplexing, where samples are uniquely labeled with barcodes and pooled together for processing and measurement, is a general strategy used in many biological assays to reduce the intra-sample variability and improve the comparability of samples. As additional benefits, it may increase the throughput of an assay and reduce the reagent consumption. Subsequently, the results of a pooled measurement are debarcoded for further analysis, meaning the labels of individual samples are recovered for each of the measured units based on the uniquely identifiable barcodes.

In flow cytometry, the fluorescent cell barcoding (FCB) technique [94], which labels samples with unique combinations of fluorophores as barcodes, is employed. Then, samples are pooled together for antibody staining and flow cytometry analysis. In mass cytometry, the mass-tag cellular barcoding (MCB) technique [88, 91], which is an adaptation of FCB, is used. Cell samples are labeled with combinatorial barcodes of metal ion tags. As all the samples are exposed to the uniform antibody concentration in tubes and measured with one mass/flow cytometry run, technical biases that may affect marker expression should be the same in all samples, and they should cancel out in comparisons between samples from the multiplexed experiment.

Traditionally, debarcoding was conducted by Boolean gating, where manually drawn gates define the negative and positive cell populations for each barcode. The positive-negative combinations over all barcodes are then compared with the table containing the scheme of deconvolution to decipher sample labels. An obvious drawback of this approach is its subjectivity in defining the gates. Moreover, cells that are found outside the gates are discarded

which may result in a quite substantial dropout when variability in barcode staining intensity between pooled samples increases. Instead, an automated approach called single-cell debarcoding (SCD) [91] is available. Briefly, it uses the differences between intensities of barcodes to define positive and negative barcodes for each cell and applies a set of criteria to decide about cell rejection or assignment to a sample.

Compensation

In flow cytometry, spectral overlap is a reason for, so called, spillover or cross talk between channels, where the fluorescence emission of one fluorochrome is detected in a detector designed to measure signal from another fluorochrome. Such contribution of background signal to the neighboring channels can have substantial effects on resolving cell populations and any downstream analysis [95]. Compensation is a process for correcting the spillover from the primary signal measured in each secondary channel. Usually compensation analysis is based on the fact that the amount of spillover is a linear function, which is then used to align the population medians. However, compensation may not always be able to fix all the undesirable effects of spillovers. It is challenging especially for channels where dimmer signal is supposed to be detected. It is also of high importance to appropriately assign reagents to markers, for example, markers that are essential to discriminate a cell population of interest should be tagged with reagents that are further apart on the wavelength axis.

Usage of metal isotopes in mass cytometry drastically reduces the spectral overlap issue. However, the sensitivity of mass spectrometry results in the measurement of metal impurities and oxide formations, which need to be carefully considered in antibody panel design (e.g., through antibody concentrations and coupling of antibodies to neighboring metals). Leipold et al. [96] recently commented that minimal spillover does not equal no spillover. Taking from the principles of compensation in flow cytometry, compensation of CyTOF data is performed via a two-step algorithm [90]. First, it uses the single-cell debarcoding to identify single positive populations of single-stained beads. Second, it estimates a spillover matrix from the identified populations, and its inverse, called a compensation matrix, is used to compensate the observed cell intensities.

Improper compensation, for example, by too much signal subtraction, may result in artifacts as the interpretation of the data can become extremely difficult or impossible [97]. As a quality control step, one should investigate the compensated data for any strange behavior before proceeding with any further analysis.

Transformation

Usually, the raw marker intensities read by a cytometer have strongly skewed distributions with varying ranges of expression, thus making it difficult to distinguish between the negative and positive cell populations. Therefore, it is common practice to transform marker intensities using, for example, arcsinh (inverse hyperbolic sine) with cofactor 5 for mass cytometry and 150 for flow cytometry data [85, 98], to make the distributions more symmetric and to map them to a comparable range of expression, which is important for clustering.

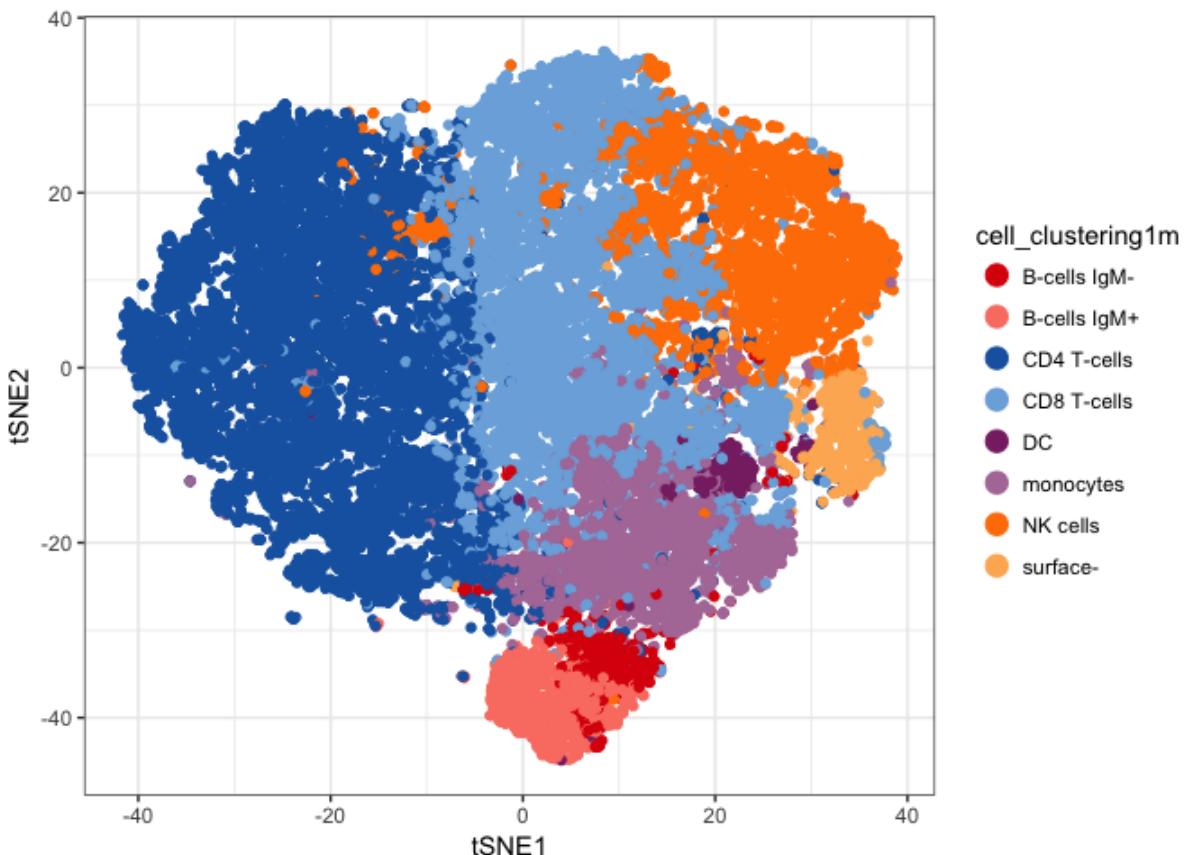


Figure 4.: An example 2D projection of peripheral blood mononuclear cells (PBMCs) obtained using t-SNE. Each point corresponds to a cell that originally was represented by 10 lineage markers. Cells are colored according to the identified cell type. This Figure corresponds to Figure 12 in **Paper II**.

1.8.2 Dimension reduction

It is a good practice to visually investigate the data after the pre-processing steps. One could examine the two-dimensional scatterplots of marker intensities. However, with increasing number of dimensions studying all the combinations becomes unfeasible. There are many alternative visualization techniques developed that give a better overview of the data. Many of them are based on dimensionality reduction in order to represent the original, high-dimensional data into usually a two-dimensional space, which is much simpler for humans to comprehend. However, there is a price to pay for the facilitated visualization: not all the details of the data structure can be preserved in the reduced dimensions.

One of the most popular dimension reduction methods for flow and mass cytometry data is t-SNE (t-distributed stochastic neighbor embedding) [99, 100] (see Figure 4), which was shown to work very well for representing cells in a lower, usually two-dimensional, space [101]. t-SNE aims at finding a lower-dimensional projection that best preserves the similarity from the original space. PCA (principal component analysis) is based on a different approach. It uses linear combinations of the original features to find orthogonal dimensions that show the highest levels of variability; the top 2 or 3 principal components can then be visualized. In general, one has to be careful with the interpretation of the lower-dimensional visualizations as it strongly depends on the algorithmic properties of the underlying methods. For example,

for the methods that are based on preserving similarities between cells, such as t-SNE, cells that are similar in the original space will be close in the 2D representation, but the opposite does not always hold [102].

1.8.3 Cell population identification

Cell population identification typically has been carried out by manual gating, a method based on visual inspection of two-dimensional scatterplots. At each step a subset of cells, either positive or negative for the given two markers, is selected and further stratified in the next iteration until a population with given marker characteristics is selected. Despite its popularity, manual gating has many drawbacks such as subjectivity, bias toward the favored cell types, high time workload and inefficiency when analyzing large datasets, which all contribute to the fact that manual gating is hard to reproduce [84]. Considerable effort has been made to improve and automate cell population identification, and many methods are currently available that address this question by means of unsupervised clustering of cells [103]. However, not all the methods scale well in terms of performance and speed from the lower dimensionality flow cytometry data to the higher dimensionality mass cytometry data [104], since clustering in higher dimensions can suffer from the “curse of dimensionality”. Beside the mathematical and algorithmic challenges of clustering, cell population identification may be difficult due to the chemical and biological aspects of the mass cytometry experiment itself. There may be cell subtypes that are difficult to discern because of the low sensitivity of antibodies (or lower than some dominating ones). The right choice of a marker panel used for clustering is also important. It should include markers that are relevant for cell type identification.

Clustering is one of the most challenging steps in the workflow and its accuracy is highly important, as it directly impacts the downstream differential analyses. In particular, getting the right resolution of clusters is critical since there can be situations where a biologically meaningful cell population may be differentially enriched between conditions, but was combined with another cell population that behaves differently, and the differential signal cannot be detected. Automatic approaches for selecting the number of clusters in HDCyto data do not always succeed [104]. Thus, it is generally recommended to allow for some level of over-clustering, and if necessary, manual merging of clusters. Such a hierarchical approach is especially suited when the goal is to detect smaller cell populations. This strategy is employed in *Citrus* [98], which tackles the problem by strong over-clustering of the data and building a hierarchy of clusters from very specific to general ones.

One of the best performing clustering approaches for high-dimensional cytometry data [104] is *FlowSOM* [105]. The *FlowSOM* workflow consists of three main steps. (I) Building of a self-organizing map (SOM), which is a work horse for clustering, where cells are assigned according to their similarities to 100 (by default) grid points (also referred to as codebook vectors or codes) of the SOM. (II) Building of a minimal spanning tree (MST), which is mainly used for graphical representation of clusters. And finally, (III) metaclustering of the SOM codes using hierarchical consensus clustering, which is performed with the *ConsensusClusterPlus* package [106]. Notably, *FlowSOM* scales easily to millions of cells and thus no subsetting of the data is required.

Once the clusters are identified, they need to be annotated. Cluster annotation is usually based on the manual investigation of the expression profiles of markers. Recently, a tool

was proposed for consistent and automated characterization of cell subsets using marker enrichment modeling (MEM) [107].

1.8.4 Differential analysis

While there is a vast number of clustering methods available, this part of the analysis is still under active development. Moreover, it is still an open question how serious some of the arising issues (e.g., normalization among batches) are and how to potentially handle them. There are three main types of differential analyses that we focus on:

1. differential abundance analysis that detects the association of cell type abundance with a phenotype,
2. differential marker expression analysis that seeks for changes in signaling markers overall,
3. or differential marker expression within specific subpopulations.

Those cell types or markers that link samples with the clinical outcome of interest are called biomarkers.

One of the approaches to differential analysis, to which we refer as a "classic" approach, first, requires the identification of cell populations of interest by the means of manual gating or automated clustering. Second, the cell subpopulations or protein markers associated with a phenotype (e.g., clinical outcome) of interest are determined using statistical tests. Typically, cell subpopulation abundance expressed as cluster cell counts, or proportions, or median marker expression would be used in the statistical model to relate to the sample-level phenotype. For two-group comparisons, the nonparametric Mann-Whitney-Wilcoxon test [87], which makes no assumptions about normality of the data, or the Student's t-test [108] and its variations, such as the paired t-test are used.

The differential analyses we propose within this thesis are based on regression frameworks where the HDCyto data is the response. Due to this, we are able to model arbitrary experimental designs, such as those with batch effects or paired designs. In particular, we apply generalized linear mixed models (GLMM) for analyses of cell population abundance or cell-population-specific analyses of signaling markers, allowing overdispersion in cell counts or aggregated marker signal across samples to be appropriately modeled.

Citrus [98] and *CellCnn* [89] model the patient response as a function of the measured HD-Cyto values. *Citrus* identifies clusters and markers that correlate with the outcome via model selection and regularization techniques. *CellCnn* learns the representation of clusters that are associated with the considered phenotype by means of convolutional neural networks, which makes it particularly applicable to detecting discriminating rare cell populations. The "filters" (populations) from *CellCnn* may identify one or more cell subsets that distinguish experimental groups. However, neither *Citrus* nor *CellCnn* are able to directly account for more complex experimental designs, such as paired experiments or presence of batches.

All of the mentioned approaches may perform poorly for extremely small sample sizes. Solutions similar to those widely accepted in transcriptomics that share information over variance parameters [26, 75, 77], would be needed. An example of such an approach is *cydar*, which performs the differential abundance analysis (on hypersphere counts) using *edgeR* [73].

In the differential marker expression analysis, mainly, the median expression of markers in each cluster and sample is used. One drawback of summarizing the protein marker intensity with median is that it does not give a full representation of the distribution. Characteristics such as bimodality, skewness and variance, are ignored. On the other hand, it results in a simple, easy to interpret approach, which in many cases is sufficient to detect interesting differences between conditions. Another issue that arises when using a summary statistic of a distribution is its uncertainty, which increases as the number of cells used to calculate it decreases. This could be partially handled in the GLMM approach by assigning weights corresponding to the size of clusters in each sample.

2 Research objectives

2.1 Multivariate model for DTU analysis

Observation of multiple isoforms being expressed from the same primary transcript leads to the idea of modeling this phenomenon as a multivariate response to have the capability to account for the correlated character of transcript expression. In addition, due to the count nature of RNA-seq used for measuring the abundance of transcript isoforms, it is natural to consider that these multivariate, count quantifications could be modeled with the multinomial distribution. However, as expected from the fact that the Poisson distribution is not able to model overdispersion present in the gene-level counts and therefore, the negative binomial is used, the same could apply to the multinomial distribution. To be able to account for overdispersion of transcript counts, we focused on the Dirichlet-multinomial (DM) distribution which is an extension of the multinomial. Notably, the beta-binomial is a univariate version of DM. To verify the applicability of DM to the DTU analysis based on RNA-seq, we have imposed the following goals:

- Design of an estimation scheme for the DM model that is robust in the small sample size analysis and allows for modeling of arbitrary complex experimental designs.
- Testing of the estimation and differential inference capabilities of the DM implementation mainly on small sample size data of increasing complexity:
 - Data simulated from the DM model itself. Simulations from the model, where one or few characteristics are changed at a time in a controlled manner, allow for better (and easier) understanding of method's sensitiveness to individual factors.
 - Simulations that mimic real data [57]. Even though, it is challenging to recapitulate all the characteristics of real data, such simulations are very useful, as the known truth of differential signal allows for usage of various metrics for methods assessment, such as achieved false discovery rate (FDR).
 - Real experimental data. In this case, evaluation of methods performance is not trivial, as it is hard to extract ground truth from the same dataset. Some external validation techniques may be employed, such as qPCR. However, they are rather laborious, and only few genes or transcripts are usually validated, which does not allow for representing a full spectrum of methods performance. Real data can be easily used to represent concordance among different methods and reproducibility capabilities (e.g. involving subsampling), but those metrics do not reflect abilities to detect truly differential genes.
- Adjustment to transcript usage QTL (tuQTL) analysis.
- Performance comparison of the DM approach to the current best performing methods

for DTU and tuQTL analyses.

- Evaluation of whether the model can be applied to other types of multivariate quantifications, such as exon counts.
- R/Bioconductor implementation of the method.

2.2 Workflow for differential analysis of HDCyto data

This project started as a collaboration work where the goal was to analyze CyTOF data obtained by our collaborators to identify biomarkers discriminating between melanoma patients who have responded to immunotherapy treatment and those who have not. In other words, the aim was to design a differential analysis strategy comparing signatures of responders and non-responders in CyTOF data. The data was acquired in two CyTOF runs and 4 independent stainings of samples. Each CyTOF run, which corresponded to a different set of patients, included samples obtained before and after the treatment initiation. Each batch includes samples for patients who had and had not responded to the treatment and healthy donors. The data placed in our hands was already pre-processed, meaning it was debarcoded and normalized. Our task was to decide which types of analysis should be performed further and to choose the best performing tools to conduct them and if needed, to tailor the existing approaches or develop new ones according to the demands of our analysis (e.g., to handle the batch effects). An important requirement for the assembled analysis pipeline was its flexibility and reproducibility, which most efficiently could be achieved by using methods that are available via a scripting language, such as R.

The established pipeline consisted of steps involving quality control spot-checks, cell population identification via clustering and differential analysis of cluster abundance and marker expression, which were adapted to account for the batch effects. We refer to such scheme of analysis as a "classic" approach, and many of its variants, using different combinations of methods, can be encountered in the literature, see Section 1.8.4. However, even with quite detailed explanation of the individual steps, it may be difficult for those who are entering the field of bioinformatics to perform such an analysis, as they may involve a variety of tools and programming skills. The Bioconductor project [109] has recently facilitated publications of workflows where the key idea, beside presenting the sequence of analysis and the tools (R/Bioconductor packages) used to carry them on, is to show the actual R code. With such a runnable workflow, one can learn how to conduct the analysis, and then use it as a template for the actual analysis on their own data. Since the pipeline that we have established fitted perfectly into this scheme, we have decided to present it as a Bioconductor workflow.

3 Research challenges

3.1 Inference in small sample size data

RNA-seq is a widely used method for transcript profiling. The costs are relatively low and the conducted experiments produce more replicates than those in the past, especially with the awareness that for standard gene profiling it is better to invest money in biological replicates than sequencing depth [110]. Despite that, the sample size in RNA-seq experiments is still lower than many classic statistical methods would require for reliable inference. One of the challenges in RNA-seq data analysis is to make the estimation methods robust in situations when only few replicates are available. Usually, estimates of the mean are not biased, but es-

timation of dispersion and variance may suffer substantially from the low sample size, which may affect the differential inference. Some excellent solutions to that problem were proposed for microarray data. They learn the general structure of variance from all the genes and use it for shrinking the individual estimates and they are adaptive, meaning that the level of shrinkage is learned from the data itself. The principles of sharing information between genes with similar characteristics, such as mean expression, to moderate dispersion are a widely used strategy also in RNA-seq data for differential gene expression analysis.

In DTU analysis based on the DM model, a similar challenge occurs when estimating concentration/dispersion. Borrowing from the existing concepts and solutions for the NB model, we have implemented a weighted likelihood method to moderate the DM concentration. However, calculating the level of moderation is not trivial.

The same challenge is encountered in the differential analysis of HDCyto data. In the "classic" approach, it is rather hard to apply information sharing as in RNA-seq. Usually, there are only few final cell populations that are tested, and it may be hard to identify a trend between, for example, the population abundance and the variation of a metric calculated on these populations. Moreover, it is a challenge to apply moderation in most of the mixed effects model settings.

3.2 Modeling complex experimental designs

Biologists are often interested in inferences beyond simple two-group comparisons. Time course changes, studying of interactions, comparisons between many types of conditions (e.g. tissues) are of increasing interest. Even when a simple two group comparison is the main goal, it is important to account for so-called batch effects in the model, if any are present. To take full advantage of such experiments, frameworks that allow analysis of complex designs are needed. For gene expression studies, linear models and generalized linear models are available and widely used. This task appears to be more challenging in differential transcript usage analyses. For example, *DEXSeq* uses the GLM engine, however since it employs interactions to model exon usage, eventually one can test only those contrasts that correspond to elimination of one or multiple coefficients from the model. This approach still covers many of the designs but not as many as a standard application of GLM allows. The multivariate approaches, such as *Cuffdiff2* or *sQTLseekerR* allow only multiple group comparisons.

The GLM framework cannot be used with DM as it does not belong to the exponential family, but a regression framework is feasible and arbitrary contrasts can be tested, see Discussion and Perspectives. *LeafCutter* [45] implements it using the Bayesian probabilistic programming language Stan for optimization. By using this Bayesian strategy, they are able to impose prior distributions for concentration, which they choose to be Gamma, to stabilize the estimates. In the new version of *DRIMSeq*, we have implemented a regression framework using an analogous optimization strategy as for the group comparison with the difference that now the regression coefficients are directly estimated instead of the proportions. However, parameter estimation is challenging, as the score and Hessian functions have relatively complicated forms, see Discussion and Perspectives. This may be a reason for computational instabilities, which influence the optimization results.

4 Thesis summary

This thesis consists of two main papers and two collaboration papers. At the end, we present the future perspectives and provide a discussion. The content of the papers and their contribution are briefly summarized below.

Paper I

DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics

by Malgorzata Nowicka and Mark D. Robinson

In this paper, we present an approach for modeling transcript expression of a gene as a multivariate response. Transcript quantifications (counts), which can be obtained with fast and accurate methods, such as *kallisto* or *Salmon*, are modeled using the Dirichlet-multinomial distribution. The DM model naturally accounts for the differential gene expression without losing information about overall gene abundance and by joint modeling of isoform expression, it has the capability to account for their correlated nature. In this paper, the DM model is employed for differential transcript usage (DTU) analysis between two experimental groups and for transcript usage QTL analysis. Both types of analyses are implemented within an R/Bioconductor package called *DRIMSeq*, which recently was extended to the regression framework allowing analysis of more complex experimental designs, and by employing the beta-binomial model to each transcript, one can identify which isoforms change within a gene with DTU. The challenge of obtaining robust estimates of model parameters when a limited numbers of replicates is available is approached by sharing information between genes. Performance of *DRIMSeq* is compared with other approaches and shows an improvement in terms of standard statistical performance metrics in the analysis based on transcript quantifications.

Paper II

CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets

by Malgorzata Nowicka, Carsten Krieg, Lukas M. Weber, Felix J. Hartmann, Silvia Guglietta, Burkhard Becher, Mitch P. Levesque and Mark D. Robinson

This paper is written in the form of a Bioconductor workflow, which along with the description of each step of analysis contains chunks of R code used to conduct them. The aim is to perform differential analyses of HDCyo data, such as differential abundance and marker expression that are associated with a phenotype. To represent it, we use a dataset used by the authors of the Citrus method, which is already pre-processed, and we do not concentrate on the pre-processing steps but show briefly how they could be performed. Notably, this workflow is equally applicable to flow or mass cytometry datasets for which the pre-processing steps are already performed. The scheme of analysis could be assigned to the "classic" approach described in Section 1.8.4. Cell population identification is conducted by means of unsupervised clustering using the *FlowSOM* and *ConsensusClusterPlus* packages, which together were among the best performing clustering approaches for high-dimensional cytometry data [104]. The differential analyses we show are based on regression frameworks

where the HDCyto data is the response; thus, we are able to model arbitrary designs, such as those with batch effects, paired designs, etc. In particular, we apply generalized linear mixed models for analyses of cell population abundance or cell-population-specific analyses of signaling markers, allowing overdispersion in cell counts or aggregated signal across samples to be appropriately modeled. To support the formal statistical analyses, we include various visualizations at every step of the analysis, including for quality control (e.g., multi-dimensional scaling plots), for reporting of clustering results (dimensionality reduction, heatmaps with dendograms) and for differential analyses (e.g., plots of aggregated signal). This workflow is not fully automatic and this is by design. It involves a step where the user can optionally manually merge and annotate clusters. In addition, the workflow is modular. In particular, alternative clustering algorithms and dimensionality reduction techniques can be used if preferred.

Collaboration papers

T-cell acute leukaemia exhibits dynamic interactions with bone marrow microenvironments

by *Edwin D. Hawkins, Delfim Duarte, Olufolake Akinduro, Reema A. Khorshed, Diana Passaro, Malgorzata Nowicka, Lenny Straszkowski, Mark K. Scott, Steve Rothery, Nicola Ruivo, Katie Foster, Michaela Waibel, Ricky W. Johnstone, Simon J. Harrison, David A. Westerman, Hang Quach, John Gribben, Mark D. Robinson, Louise E. Purton, Dominique Bonnet and Cristina Lo Celso*

In this project, human T-cell acute lymphoblastic leukaemia (T-ALL) interactions with the bone marrow (BM) microenvironment in mouse model were studied. Mainly the intravital microscopy was used to monitor the progression of disease within the bone marrow. Results of this study showed that T-ALL cells do not depend on specific bone marrow microenvironments for propagation of disease, nor for the selection of chemo-resistant clones. To investigate this proposed paradigm further, a microarray experiment was conducted to compare the gene expression profiles of T-ALL cell populations purified at the full infiltration stage to that of cells isolated from mice 7-10 days after initiation of dexamethasone treatment, a time point at which surviving cells have recolonised the BM space.

We designed the microarray experiment, and proposed and conducted a workflow of differential gene expression (DGE) analysis based on the obtained data. The Affymetrix GeneChip Mouse Gene 2.0 ST array platform was used for gene profiling. The final analysis pipeline consisted of the pre-processing steps performed with the RMA approach described in Section 1.3.1. Probe annotation was obtained from the Affymetrix NetAffx Query website. Differential analysis were conducted with the *limma* package. Various visualizations were proposed to support the statistical analysis, such as MDS and MA plots, Venn diagrams and heatmaps of differentially expressed genes.

The results of DGE analysis showed that gene expression profiles of all T-ALL samples were significantly more heterogeneous compared to those of control T-cell (CD4+ and CD8+), T-cell progenitor (CD4+CD8+ thymocytes) and whole BM populations as expected based on inter-clone variation in leukaemia cells. The transcriptome of resistant cells overlapped with that of T-ALL cells pre-treatment. Indeed, only 79 genes were differentially expressed in the post-treatment group, and consistent with the intravital imaging data, none of the differentially expressed genes were related to known cell-niche interaction candidates.

High dimensional single cell analysis predicts response to anti-PD-1 immunotherapy

by Carsten Krieg, Małgorzata Nowicka, Silvia Guglietta, Sabrina Schindler, Felix J. Hartmann, Lukas M. Weber, Reinhard Dummer, Mark D. Robinson, Mitchell P. Levesque and Burkhard Becher

In this study, the main goal was to identify biomarkers of clinical response to the targeted immunotherapy with anti-PD-1. Such immune signatures could then be used in a screening tool to select potential responders prior to treatment initiation. For that a CyTOF experiment on peripheral blood mononuclear cell (PBMC) from melanoma patients before and during anti-PD-1 immunotherapy was conducted.

With this data in hand, we have designed an interactive bioinformatics pipeline to generate profiles of the peripheral blood immune cells, which were then used to identify signatures of responsiveness to the therapy. Individual steps of this pipeline are presented in detail in Paper II.

As a result of the assembled differential analysis, we have found that the frequency of CD14+CD16-CD33+HLA-DRhi monocytes is a strong predictor of responsiveness to anti-PD-1 immunotherapy. We could observe strikingly higher abundance of CD14+CD16-CD33+HLA-DRhi classical monocytes in responders before therapy, which was also confirmed by regular flow cytometry.

References

- [1] Francis Crick and Others. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.
- [2] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, 2008.
- [3] Timothy W Nilsen and Brenton R Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.
- [4] Jamal Tazi, Nadia Bakkour, and Stefan Stamm. Alternative splicing and disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1792(1):14–26, 2009.
- [5] Guey-Shin Wang and Thomas a Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature reviews. Genetics*, 8(10):749–61, 2007.
- [6] Javier F Cáceres and Alberto R Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends in Genetics*, 18(4):186–193, 2002.
- [7] Erin L Heinzen, Woohyun Yoon, Sarah K Tate, Arjune Sen, Nicholas W Wood, Sanjay M Sisodiya, and David B Goldstein. Nova2 Interacts with a Cis-Acting Polymorphism to Influence the Proportions of Drug-Responsive Splice Variants of {SCN1A}. *The American Journal of Human Genetics*, 80(5):876–883, 2007.
- [8] Michael Morley, Cliona M Molony, Teresa M Weber, James L Devlin, Kathryn G Ewens, Richard S Spielman, and Vivian G Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747, 2004.
- [9] Tony Kwan, David Benovoy, Christel Dias, Scott Gurd, Cathy Provencher, Patrick Beaulieu, Thomas J Hudson, Rob Sladek, and Jacek Majewski. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet*, 40(2):225–231, 2008.

-
- [10] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar González-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G MacArthur, Monkol Lek, Esther Lizano, Henk P J Buermans, Ismael Padoleau, Thomas Schwarzmayr, Olof Karlberg, Halit Onyen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B Montgomery, Peter Donnelly, Mark I McCarthy, Paul Flicek, Tim M Strom, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E Antonarakis, Robert Häslar, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G Gut, Xavier Estivill, and Emmanouil T Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–11, 2013.
 - [11] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multi-tissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.
 - [12] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.
 - [13] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.
 - [14] J DeRisi, L Penland, P O Brown, M L Bittner, P S Meltzer, M Ray, Y Chen, Y A Su, and J M Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet*, 14(4):457–460, 1996.
 - [15] Patrick O Brown and David Botstein. Exploring the new world of the genome with DNA microarrays. *Nat Genet*, 21(1 Suppl):33–37, 1999.
 - [16] David J Lockhart, Helin Dong, Michael C Byrne, Maximillian T Follettie, Michael V Gallo, Mark S Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi, Heidi Norton, and Eugene L Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech*, 14(13):1675–1680, dec 1996.
 - [17] Robert J Lipshutz, Stephen P A Fodor, Thomas R Gingras, and David J Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20 – 24, 1999.
 - [18] Mark D Robinson and Terence P Speed. Differential splicing using whole-transcript microarrays. *BMC bioinformatics*, 10:156, 2009.
 - [19] B M Bolstad, R A Irizarry, M Åstrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185, 2003.
 - [20] Benjamin Milo Bolstad. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, UNIVERSITY OF CALIFORNIA, BERKELEY, 2004.
 - [21] B.M. Bolstad, R. A. Irizarry, L. Gautier, and Z. Wu. Preprocessing High-density Oligonucleotide Arrays. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. 2005.

-
- [22] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15–e15, feb 2003.
 - [23] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics Oxford England*, 4(2):249–264, 2003.
 - [24] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700, 2007.
 - [25] Jeremy D Silver, Matthew E Ritchie, and Gordon K Smyth. Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. *Biostatistics*, 10(2):352, 2009.
 - [26] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.
 - [27] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.
 - [28] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica Di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.
 - [29] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 2016.
 - [30] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17, 2008.
 - [31] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE*, 9(1):1–13, 2014.
 - [32] SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotech*, 32(9):903–914, 2014.
 - [33] Charles Wang, Binsheng Gong, Pierre R Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, Huixiao Hong, Jie Shen, Zhenqiang Su, Joe Meehan, Xiaojin Li, Lu Yang, Haiqing Li, Pawel P Labaj, David P Kreil, Dalila Megherbi, Stan Gaj, Florian Caiment, Joost van Delft, Jos Kleinjans, Andreas Scherer, Viswanath Devanarayan, Jian Wang, Yong Yang, Hui-Rong Qian, Lee J Lancashire, Marina Bessarabova,

Yuri Nikolsky, Cesare Furlanello, Marco Chierici, Davide Albanese, Giuseppe Jurman, Samantha Riccadonna, Michele Filosi, Roberto Visintainer, Ke K Zhang, Jianying Li, Jui-Hua Hsieh, Daniel L Svoboda, James C Fuscoe, Youping Deng, Leming Shi, Richard S Paules, Scott S Auerbach, and Weida Tong. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotech*, 32(9):926–932, 2014.

- [34] Michael I Love, John B Hogenesch, and Rafael A Irizarry. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat Biotech*, 34(12):1287–1291, 2016.
- [35] Simon Andrews. FastQC: A quality control tool for high throughput sequence data, 2010.
- [36] Manhong Dai, Robert C Thompson, Christopher Maher, Rafael Contreras-Galindo, Mark H Kaplan, David M Markovitz, Gil Omenn, and Fan Meng. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*, 11(4):S7, 2010.
- [37] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114, 2014.
- [38] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.
- [39] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.
- [40] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):bts635–, 2013.
- [41] Par G Engstrom, Tamara Steijger, Botond Sipos, Gregory R Grant, Andre Kahles, The RGASP Consortium, Gunnar Ratsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigo, and Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Meth*, 10(12):1185–1191, 2013.
- [42] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5):462–4, 2014.
- [43] Nicolas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, advance on, 2016.
- [44] Rob Patro, Geet Duggal, and Carl Kingsford. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*, page 021592, 2015.
- [45] Yang I Li, David A Knowles, and Jonathan K Pritchard. LeafCutter: Annotation-free quantification of RNA splicing. *bioRxiv*, 2016.

-
- [46] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.
 - [47] Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.
 - [48] Halit Ongen and Emmanouil T. Dermitzakis. Alternative Splicing QTLs in European and African Populations. *American Journal of Human Genetics*, 97(4):567–575, 2015.
 - [49] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12):1009–1015, 2010.
 - [50] Shihao Shen, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(51):E5593–601, 2014.
 - [51] Gael P Alamancos, Amadis Pages, Juan L Trincado, Nicolas Bellora, and Eduardo Eyras. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA (New York, N.Y.)*, 21(9):1521–1531, sep 2015.
 - [52] Leonard D Goldstein, Yi Cao, Gregoire Pau, Michael Lawrence, Thomas D Wu, Somasekar Seshagiri, and Robert Gentleman. Prediction and Quantification of Splice Events from RNA-Seq Data. *PLoS ONE*, 11(5):e0156132, 2016.
 - [53] Keyan Zhao, Zhi-Xiang Lu, Juw Won Park, Qing Zhou, and Yi Xing. GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome biology*, 14(7):R74, 2013.
 - [54] Cheng Jia, Yu Hu, Yichuan Liu, and Mingyao Li. Mapping Splicing Quantitative Trait Loci in RNA-Seq. *Cancer Informatics*, 13:35–43, 2014.
 - [55] Yu Hu, Yichuan Liu, Xianyun Mao, Cheng Jia, Jane F. Ferguson, Chenyi Xue, Muredach P. Reilly, Hongzhe Li, and Mingyao Li. PennSeq: Accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution. *Nucleic Acids Research*, 42(3), 2014.
 - [56] Jean Monlong, Miquel Calvo, Pedro G. Ferreira, and Roderic Guigó. Identification of genetic variants associated with alternative splicing using sQTLseeker. *Nature Communications*, 5(May):4698, 2014.
 - [57] Charlotte Soneson, Katarina L Matthes, Małgorzata Nowicka, Charity W Law, and Mark D Robinson. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, 17(1):1–15, 2016.
 - [58] Peter Glaus, Antti Honkela, and Magnus Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.
 - [59] David Rossell, Camille Stephan-Otto Attolini, Manuel Kroiss, and Almond Stöcker. Quantifying Alternative Splicing From Paired-End Rna-Sequencing Data. *The annals of applied statistics*, 8(1):309–330, 2014.
-

-
- [60] Cole Trapnell, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
 - [61] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12:323, jan 2011.
 - [62] Elsa Bernard, Laurent Jacob, Julien Mairal, and Jean-Philippe Vert. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics (Oxford, England)*, pages 1–9, 2014.
 - [63] C Soneson, M I Love, and M D Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research*, 4(1521), 2016.
 - [64] Małgorzata Nowicka and Mark D Robinson. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 2; referees: 2 approved]. *F1000Research*, 5(1356), 2016.
 - [65] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, 2012.
 - [66] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, 2014.
 - [67] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, Kenneth B. Beckman, Jianxin Shi, Rui Mei, Alexander E. Urban, Stephen B. Montgomery, Douglas F. Levinson, and Daphne Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24, 2014.
 - [68] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010.
 - [69] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics (Oxford, England)*, pages 1–7, 2015.
 - [70] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777, 2010.
 - [71] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, 2013.
 - [72] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, 2010.
-

-
- [73] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.
 - [74] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.
 - [75] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.
 - [76] Mark D. Robinson and Gordon K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008.
 - [77] Mark D. Robinson and Gordon K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.
 - [78] N. Reid and D. A. S. Fraser. Likelihood inference in the presence of nuisance parameters. page 7, dec 2003.
 - [79] Peter McCullagh and Robert Tibshirani. A Simple Method for the Adjustment of Profile Likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(2):325–344, 1990.
 - [80] D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1):1–39, 1987.
 - [81] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:Article3, 2004.
 - [82] J Paul Robinson and Mario Roederer. Flow cytometry strikes gold. *Science*, 350(6262):739–740, 2015.
 - [83] Dmitry R Bandura, Vladimir I Baranov, Olga I Ornatsky, Alexei Antonov, Robert Kinach, Xudong Lou, Serguei Pavlov, Sergey Vorobiev, John E Dick, and Scott D Tanner. Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. *Analytical Chemistry*, 81(16):6813–6822, aug 2009.
 - [84] Yvan Saeys, Sofie Van Gassen, and Bart N Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol*, 16(7):449–462, jul 2016.
 - [85] Sean C Bendall, Erin F Simonds, Peng Qiu, El-ad D Amir, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, Robert S Balderas, Sylvia K Plevritis, Karen Sachs, Dana Pe\textquoterightighter, Scott D Tanner, and Garry P Nolan. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030):687–696, 2011.
 - [86] Cole Trapnell. Defining cell types and states with single-cell genomics, 2015.
 - [87] Felix J Hartmann, Raphaël Bernard-Valnet, Clémence Quériault, Dunja Mrdjen, Lukas M Weber, Edoardo Galli, Carsten Krieg, Mark D Robinson, Xuan-Hung Nguyen, Yves

Dauvilliers, Roland S Liblau, and Burkhard Becher. High-dimensional single-cell analysis reveals the immune signature of narcolepsy. *Journal of Experimental Medicine*, 213(12):2621–2633, 2016.

- [88] Bernd Bodenmiller, Eli R Zunder, Rachel Finck, Tiffany J Chen, Erica S Savig, Robert V Bruggner, Erin F Simonds, Sean C Bendall, Karen Sachs, Peter O Krutzik, and Garry P Nolan. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature Biotechnology*, 30(9):858–867, 2012.
- [89] Eirini Arvaniti and Manfred Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *bioRxiv*, 2016.
- [90] Helena Lucia Crowell, Mark Robinson, and Vito Zanotelli. *CATALYST: Cytometry dATA anALYSis Tools*, 2017.
- [91] Eli R Zunder, Rachel Finck, Gregory K Behbehani, El-ad D Amir, Smita Krishnaswamy, Veronica D Gonzalez, Cynthia G Lorang, Zach Bjornson, Matthew H Spitzer, Bernd Bodenmiller, Wendy J Fantl, Dana Pe'er, and Garry P Nolan. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nature Protocols*, 10(2):316–333, 2015.
- [92] Rachel Finck, Erin F Simonds, Astraea Jager, Smita Krishnaswamy, Karen Sachs, Wendy Fantl, Dana Pe'er, Garry P Nolan, and Sean C Bendall. Normalization of mass cytometry data with bead standards. *Cytometry Part A*, 83A:483–494, 2013.
- [93] Florian Hahne, Alireza Hadj Khodabakhshi, Ali Bashashati, Chao-Jen Wong, Randy D Gascoyne, Andrew P Weng, Vicky Seyfert-Margolis, Katarzyna Bourcier, Adam Asare, Thomas Lumley, Robert Gentleman, and Ryan R Brinkman. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, 77A(2):121–131, 2010.
- [94] Peter O Krutzik and Garry P Nolan. Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. *Nat Meth*, 3(5):361–368, 2006.
- [95] Mario Roederer. Spectral compensation for flow cytometry: Visualization artifacts, limitations, and caveats. *Cytometry*, 45(3):194–205, 2001.
- [96] Michael D Leipold. Another step on the path to mass cytometry standardization. *Cytometry Part A*, 87(5):380–382, 2015.
- [97] Stephen P Perfetto, David Ambrozak, Richard Nguyen, Pratip Chattopadhyay, and Mario Roederer. Quality assurance for polychromatic flow cytometry. *Nat. Protocols*, 1(3):1522–1530, nov 2006.
- [98] Robert V Bruggner, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences of the United States of America*, 111(26):E2770–7, 2014.
- [99] L J P Van Der Maaten and G E Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 2008.
- [100] Laurens van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

-
- [101] El-ad David Amir, Kara L Davis, Michelle D Tadmor, Erin F Simonds, Jacob H Levine, Sean C Bendall, Daniel K Shenfeld, Smita Krishnaswamy, Garry P Nolan, and Dana Pe'er. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, 31(6):545–52, 2013.
 - [102] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to Use t-SNE Effectively. *Distill*, 2016.
 - [103] Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nat Meth*, 10(3):228–238, mar 2013.
 - [104] Lukas M Weber and Mark D Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096, 2016.
 - [105] Sofie Van Gassen, Britt Callebaut, Mary J Van Helden, Bart N Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 87(7):636–45, 2015.
 - [106] Matthew D Wilkerson and D Neil Hayes. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572, 2010.
 - [107] Kirsten E Diggins, Allison R Greenplate, Nalin Leelatian, Cara E Wogsland, and Jonathan M Irish. Characterizing cell subsets using marker enrichment modeling. *Nat Meth*, 14(3):275–278, mar 2017.
 - [108] David Pejoski, Nicolas Tchitchev, André Rodriguez Pozo, Jamila Elhmouzi-Younes, Rahima Yousfi-Bogniaho, Christine Rogez-Kreuz, Pascal Clayette, Nathalie Dereuddre-Bosquet, Yves Lévy, Antonio Cosma, Roger Le Grand, and Anne-Sophie Beignon. Identification of Vaccine-Altered Circulating B Cell Phenotypes Using Mass Cytometry and a Two-Step Clustering Analysis. *The Journal of Immunology*, 196(11):4814–4831, 2016.
 - [109] Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.
 - [110] Nicholas J. Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, and Geoffrey J. Barton. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851, jun 2016.

PAPER I

**DRIMSeq: a Dirichlet-multinomial framework for
multivariate count outcomes in genomics**

Malgorzata Nowicka and Mark D. Robinson

Paper published in *F1000Research* (2016), 5(1356)



METHOD ARTICLE

REVISED DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 2; referees: 2 approved]Malgorzata Nowicka^{1,2}, Mark D. Robinson^{1,2}¹Institute for Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland²SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, 8057, Switzerland**V2** First published: 13 Jun 2016, 5:1356 (doi: [10.12688/f1000research.8900.1](https://doi.org/10.12688/f1000research.8900.1))Latest published: 06 Dec 2016, 5:1356 (doi: [10.12688/f1000research.8900.2](https://doi.org/10.12688/f1000research.8900.2))**Abstract**

There are many instances in genomics data analyses where measurements are made on a multivariate response. For example, alternative splicing can lead to multiple expressed isoforms from the same primary transcript. There are situations where differences (e.g. between normal and disease state) in the relative ratio of expressed isoforms may have significant phenotypic consequences or lead to prognostic capabilities. Similarly, knowledge of single nucleotide polymorphisms (SNPs) that affect splicing, so-called splicing quantitative trait loci (sQTL) will help to characterize the effects of genetic variation on gene expression. RNA sequencing (RNA-seq) has provided an attractive toolbox to carefully unravel alternative splicing outcomes and recently, fast and accurate methods for transcript quantification have become available. We propose a statistical framework based on the Dirichlet-multinomial distribution that can discover changes in isoform usage between conditions and SNPs that affect relative expression of transcripts using these quantifications. The Dirichlet-multinomial model naturally accounts for the differential gene expression without losing information about overall gene abundance and by joint modeling of isoform expression, it has the capability to account for their correlated nature. The main challenge in this approach is to get robust estimates of model parameters with limited numbers of replicates. We approach this by sharing information and show that our method improves on existing approaches in terms of standard statistical performance metrics. The framework is applicable to other multivariate scenarios, such as Poly-A-seq or where beta-binomial models have been applied (e.g., differential DNA methylation). Our method is available as a Bioconductor R package called DRIMSeq.



This article is included in the **Bioconductor** channel.

Open Peer Review

Referee Status: ✓ ✓

Invited Referees	
1	2
REVISED	✓
version 2 published 06 Dec 2016	report ↑

REVISED	✓
version 2 published 06 Dec 2016	report ↑

version 1	✓
published 13 Jun 2016	report ?

1 **Alejandro Reyes**, European Molecular Biology Laboratory Germany

2 **Robert Castelo**, Pompeu Fabra University Spain

Discuss this article

Comments (0)



This article is included in the **RPackage** channel.

Corresponding author: Mark D. Robinson (mark.robinson@imls.uzh.ch)

How to cite this article: Nowicka M and Robinson MD. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 2; referees: 2 approved] *F1000Research* 2016, 5:1356 (doi: [10.12688/f1000research.8900.2](https://doi.org/10.12688/f1000research.8900.2))

Copyright: © 2016 Nowicka M and Robinson MD. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: MN acknowledges the funding from a Swiss Institute of Bioinformatics (SIB) Fellowship. MDR would like to acknowledge funding from a Swiss National Science Foundation (SNSF) Project Grant (143883).

Competing interests: No competing interests were disclosed.

First published: 13 Jun 2016, 5:1356 (doi: [10.12688/f1000research.8900.1](https://doi.org/10.12688/f1000research.8900.1))

REVISED Amendments from Version 1

In version 2 of the manuscript, we have reworded sections in the Introduction to clarify the scope of existing methods, with respect to the term 'differential splicing'. We have added additional analyses for differential splicing analyses, to better understand how the null P-value distributions compare across different simulation scenarios and dispersion estimators. For the detected tuQTLs, we added an analysis with respect to enrichment of splicing-related features.

See referee reports

Introduction

With the development of digital high-throughput sequencing technologies, the analysis of count data in genomics has become an important theme motivating the investigation of new, more powerful and robust approaches that handle complex overdispersion patterns while accommodating the typical small numbers of experimental units.

The basic distribution for modeling univariate count responses is the Poisson distribution, which also approximates the binomial distribution. One important limitation of the Poisson distribution is that the mean is equal to the variance, which is not sufficient for modeling, for example, gene expression from RNA sequencing (RNA-seq) data where the variance is higher than the mean due to technical sources and biological variability^{1–5}. A natural extension of the Poisson distribution that accounts for overdispersion is the negative-binomial distribution, which has been extensively studied in the small-sample situation and has become an essential tool in genomics applications^{1–3}.

Analogously, the fundamental distribution for modeling multivariate count data is the multinomial distribution, which models proportions across multiple features. To account for overdispersion, the multinomial can be extended to the Dirichlet-multinomial (DM) distribution⁶. Because of its flexibility, the DM distribution has found applications in forensic genetics⁷, microbiome data analysis⁸, the analysis of single-cell data⁹ and for identifying nucleosome positions¹⁰. Another extension of the multinomial is the Dirichlet negative multinomial distribution¹¹, which allows modeling of correlated count data and was applied in the analysis of clinical trial recruitment¹². Notably, the beta-binomial distribution, such as those used in differential methylation from bisulphite sequencing data^{13–15}, represent a special case of the DM.

Genes may express diverse transcript isoforms (mRNA variants) as a consequence of alternative splicing or due to the differences in transcription start sites and polyadenylation sites¹⁶. Hence, gene expression can be viewed as a multivariate expression of transcripts or exons and such a representation allows the study of not only the overall gene expression, but also the expressed variant composition. Differences in the relative expression of isoforms can have significant phenotypic consequences and aberrations are associated with disease^{17,18}. Thus, biologists are interested in using RNA-seq data to discover differences in transcript usage between conditions or to study the specific molecular mechanisms that mediate these

changes, for example, alternative splice site usage. In general terms, we collect all these together under the term "differential splicing" (DS)¹⁹.

Alternative splicing is a process regulated by complex protein-RNA interactions that can be altered by genetic variation. Knowledge of single nucleotide polymorphisms (SNPs) that affect splicing, known as splicing quantitative trait loci (sQTL), can help to characterize this layer of regulation.

In this article, we propose the DM distribution to model relative usage of isoforms. The DM model treats transcript expression as a multivariate response and allows for flexible small-sample estimation of overdispersion. We address the challenge of obtaining robust estimates of the model parameters, especially dispersion, when only a small number of replicates is available by applying an empirical Bayes approach to share information, similar to those proven successful in negative-binomial frameworks^{1,20}. In particular, weighted likelihood is used to moderate the gene-wise dispersion toward a common or trended value.

The Dirichlet-multinomial framework, implemented as a *Bioconductor* R package called *DRIMSeq*, is applicable to both differential transcript usage (DTU) analysis between conditions and transcript usage quantitative trait loci (tuQTL) analysis. It has been evaluated and compared to the current best methods in extensive simulations and in real RNA-seq data analysis using transcript and exon counts, highlighting that *DRIMSeq* performs best with transcript counts. Furthermore, the framework can be applied to other types of emerging multivariate genomic data, such as PolyA-seq where the collection of polyadenylated sites for a given gene are measured²¹ and to settings where the beta-binomial is already applied (e.g., differential methylation, allele-specific differential gene expression).

Approaches to DS and sQTL analyses

RNA-seq has provided an attractive toolbox to unravel alternative splicing outcomes. There are various methods designed explicitly to detect DS based on samples from different experimental conditions^{19,22,23}. Independently, a set of methods was developed for detecting genetic variation associated with changes in splicing (sQTLs). While sQTL detection represents a different application, it is essentially DS between groups defined by genotypes. In the following overview, we do not distinguish between applications but rather between the general concepts used to detect differences in splicing.

DS can be studied in three main ways: as differential transcript usage (DTU) or, in a more local context, as differential exon or exon junction usage (DEU) or as specific splicing events (e.g., exon skipping), and all have their advantages and disadvantages. A survey of the main methods can be found in Table S1 (Supplementary File). From the quantification perspective, exon-level abundance estimation is straightforward since it is based on counting read-region overlaps (e.g., *featureCounts*²⁴). Exons from different isoforms may have different boundaries, thus the authors of *DEXSeq*²⁵ quantify with *HTSeq*²⁶ non-overlapping windows defined by projecting all exons to the linear genome.

However, this strategy does not utilize the full information from junction reads. Such reads are counted multiple times (in all exons that they overlap with), artificially increasing the total number of counts per gene and ignoring that junction reads support the isoforms that explicitly contain the combinations of exons spanned by these reads. This issue is captured in *Altrans*²⁷, which quantifies exon-links (exon junctions) or in *MISO*²⁸, *rMATS*²⁹, *SUPPA*³⁰ and *SGSeq*³¹, all of which calculate splicing event inclusion levels expressed as percentage spliced in (PSI). Such events capture not only cassette exons but also alternative 3' and 5' splice sites, mutually exclusive exons or intron retention. *GLiMMPS*³² and Jia *et al.*³³, with quantification from *PennSeq*³⁴, use event inclusion levels for detecting SNPs that are associated with differential splicing. However, there are (hypothetical) instances where changes in splicing pattern may not be captured by exon-level quantifications (Figure 1A in the paper by Monlog *et al.*³⁵). Furthermore, detection of more complex transcript variations remains a challenge for exon junction or PSI methods (see Figure S5 in the paper by Ongen *et al.*²⁷). Soneson *et al.*²³ considered counting which accommodates various types of local splicing events, such as exon paths traced out by paired reads, junction counts or events that correspond to combinations of isoforms; in general, the default exon-based counting resulted in strongest performance for DS gene detection.

The above methods allow for detection of differential usage of local splicing features, which can serve as an indicator of differential transcript usage but often without knowing specifically which isoforms are differentially regulated. This can be a disadvantage in cases where knowing the isoform ratio changes is important, since isoforms are the ultimate determinants of proteins. Moreover, exons are not independent transcriptional units but building blocks of transcripts. Thus, the main alternative is to make a calculation of DS using isoform-level quantitations. A vast number of methods is available for gene isoform quantification, such as *MISO*²⁸, *BitSeq*³⁶, *casper*³⁷, *Cufflinks*³⁸, *RSEM*³⁹, *FlipFlop*⁴⁰ and more recent, extremely fast pseudoalignment-based methods, such as *Sailfish*⁴¹, *kallisto*⁴² and *Salmon*⁴³. Additionally, *Cufflinks*, *casper* and *FlipFlop* allow for *de novo* transcriptome assembly. Recently, performance of various methods was extensively studied^{44,45}, including a webtool⁴⁵ to allow further comparisons. Regardless of this progress, it remains a complex undertaking to quantify isoform expression from short cDNA fragments since there is a high degree of overlap between transcripts in complex genes; this is a limitation of the technology, not the algorithms. In the case of incomplete transcript annotation, local approaches may be more robust and can detect differential changes due to transcripts that are not in the catalog^{23,27}. Nevertheless, DS at the resolution of isoforms is the ultimate goal within the *DRIMSeq* framework, and with the emergence of longer reads (fragments), transcript quantifications will become more accurate and methods for multivariate transcript abundances will be needed.

Whether the differential analysis is done at the transcript or local level, modeling and testing independently each transcript^{46,47} or exon ratio⁴⁸ ignores the correlated structure of

these quantities (e.g., proportions must sum to 1). Similarly, separate modeling and testing of exon junctions (*Altrans*²⁷) or splicing events (*rMATS*²⁹, *GLiMMPS*³², Jia *et al.*³³, Montgomery *et al.*⁴⁹) of a gene leads to non-independent statistical tests, although the full effect of this on calibration (e.g., controlling the rate of false discoveries) is not known. Nevertheless, with the larger number of tests, the multiple testing correction becomes more extreme. In sQTL analyses, this burden is even larger since there are many SNPs tested for each gene. There, the issue of multiple comparisons is usually accounted for by applying a permutation scheme in combination with the false discovery rate (FDR) estimation^{27,32,35,46,48–50}.

DEXSeq and *voom-diffSplice*^{4,5} undertake another approach, where the modeling is done per gene. *DEXSeq* fits a generalized linear model (GLM), assuming that (exonic) read counts follow the negative-binomial distribution. A bin is deemed differentially used when its corresponding group-bin interaction is significantly different. The exact details of *voom-diffSplice* are not published. Nevertheless, exons are again treated as independent in the gene-level model.

In contrast, *MISO*²⁸, *Cuffdiff*^{38,51} and *sQTLseekerR*³⁵ model alternative splicing as a multivariate response. *MISO* is designed for DS analyses only between two samples and does not handle replicates. Variability among replicates is captured within *Cuffdiff* via the Jensen-Shannon divergence metric on probability distributions of isoform proportions as a measure of changes in isoform relative abundances between samples. *sQTLseekerR* tests for the association between genotype and transcript composition, using an approach similar to a multivariate analysis of variance (MANOVA) without assuming any probabilistic distribution and Hellinger distance as a dissimilarity measure between transcript ratios. Very recently, *LeafCutter*⁵² gives intron usage quantifications that can be used for both DS analyses (also using the DM model) and sQTL analyses via a correlation-based approach with *FastQTL*⁵⁰.

sQTLseeker, *Altrans*, *LeafCutter* and other earlier methods for the sQTL analysis^{35,46–48} employ feature ratios to account for the overall gene expression. A potential drawback of this approach is that feature ratios do not take into account whether they are based on high or low expression, while the latter have more uncertainty in them. *DRIMSeq* naturally builds this in via the multinomial model.

Dirichlet-multinomial model for relative transcript usage

In the application of the DM model to DS, we refer to *features* of a gene. These features can be transcripts, exons, exonic bins or other multivariate measurable units, which for DS, contain information about isoform usage and can be quantified with (estimated) counts.

Assume that a gene has q features with relative expression defined by a vector of proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_q)$, and the feature counts $\mathbf{Y} = (Y_1, \dots, Y_q)$ are random variables. Let $\mathbf{y} = (y_1, \dots, y_q)$ be the observed counts and $m = \sum_{j=1}^q y_j$. Here, m is treated as an ancillary statistic since it depends on the sequencing depth and gene

expression, but not on the model parameters. The simplest way to model feature counts is with the multinomial distribution with probability function defined as:

$$f_M(y; \pi) = \binom{m}{y} \prod_{j=1}^q \pi_j^{y_j}, \quad (1)$$

where the mean and the covariance matrix of \mathbf{Y} are $\mathbb{E}(\mathbf{Y}) = m\pi$ and $\mathbb{V}(\mathbf{Y}) = \text{diag}(\pi) - \pi\pi^T$, respectively.

To account for overdispersion due to true biological variation between experimental units as well as technical variation, such as library preparation and errors in transcript quantification, we assume the feature proportions, Π , follow the (conjugate) Dirichlet distribution, with density function:

$$f_D(\pi; \gamma) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^q \Gamma(\gamma_j)} \prod_{j=1}^q \pi_j^{\gamma_j-1}, \quad (2)$$

where $\gamma_j, j = 1, \dots, q$ are the Dirichlet parameters and $\gamma_+ = \sum_{j=1}^q \gamma_j$. The mean and covariance matrix of random proportions Π are $\mathbb{E}(\Pi) = \gamma/\gamma_+ = \pi$ and $\mathbb{V}(\Pi) = \{\gamma_+ \text{diag}(\gamma) - \gamma\gamma^T\}/(\gamma_+^2(\gamma_+ + 1))$, respectively. We can see that proportions Π are proportional to γ and their variance is inversely proportional to γ_+ , which is called the concentration or precision parameter. As γ_+ gets larger, the proportions are more concentrated around their means.

We can derive the marginal distribution of \mathbf{Y} by multiplying densities (1) and (2) and integrating over π . Then, feature counts \mathbf{Y} follow the DM distribution⁶ with probability function defined as:

$$\begin{aligned} f_{DM}(y; \gamma) &= \int_{\pi} f_M(y; \pi) f_D(\pi; \gamma) d\pi \\ &= \binom{m}{y} \frac{\Gamma(\gamma_+)}{\Gamma(m + \gamma_+)} \prod_{j=1}^q \frac{\Gamma(\gamma_j + \gamma_+)}{\Gamma(\gamma_j)}. \end{aligned} \quad (3)$$

The mean of \mathbf{Y} is unchanged at $\mathbb{E}(\mathbf{Y}) = \mathbb{E}\{\mathbb{E}(\mathbf{Y}|\Pi)\} = \mathbb{E}(m\Pi) = m\gamma/\gamma_+ = m\pi$, while the covariance matrix of \mathbf{Y} is given by $\mathbb{V}(\mathbf{Y}) = cm\{\text{diag}(\pi) - \pi\pi^T\}$, where $c = (m + \gamma_+)/(1 + \gamma_+)$ is an additional factor when representing the Dirichlet-multinomial covariance to the ordinary multinomial covariance. c depends on concentration parameter γ_+ which controls the degree of overdispersion and is inversely proportional to variance of \mathbf{Y} .

We can represent the DM distribution using an alternative parameterization: $\pi = \gamma/\gamma_+$ and $\theta = 1/(1 + \gamma_+)$; then, the covariance of \mathbf{Y} can be represented as $\mathbb{V}(\mathbf{Y}) = n\{\text{diag}(\pi) - \pi\pi^T\} \{1 + \theta(n - 1)\}$, where θ can be interpreted as a dispersion parameter. When θ grows (γ_+ gets smaller), the variance becomes larger. From the knowledge of the gamma function, $x\Gamma(x) = \Gamma(x + 1)$, we can write $\frac{\Gamma(x+s)}{\Gamma(x)} = \prod_{r=1}^x \{x + (r - 1)\}$. Hence, the DM density function becomes:

$$f_{DM}(y; \pi, \theta) = \binom{m}{y} \frac{\prod_{j=1}^q \prod_{r=1}^{\gamma_j} \{\pi_j(1-\theta) + (r-1)\theta\}}{\prod_{r=1}^m \{1 - \theta + (r-1)\theta\}}, \quad (4)$$

such that for $\theta = 0$, DM reduces to multinomial.

Detecting DTU and tuQTLs with the Dirichlet-multinomial model

Within DRIMSeq, the DM method can be used to detect the differential usage of gene features between two or more conditions. For simplicity, suppose that features of a gene are transcripts and the comparison is done between two groups. The aim is to determine whether transcript ratios of a gene are different in these two conditions. Formally, we want to test the hypothesis $H_0: \pi_1 = \pi_2$ against the alternative $H_1: \pi_1 \neq \pi_2$. For the convenience of parameter estimation, we decide to use the DM parameterization with precision parameter γ_+ , which can take any non-negative value, instead of dispersion parameter θ , which is bounded to values between 0 and 1. Because our goal is to compare the proportions from two groups, γ_+ is a nuisance parameter that gets estimated in the first step (see the following Section). Let $l(\pi_1, \pi_2, \gamma_+)$ be the joint log-likelihood function. Assuming $\gamma_+ = \hat{\gamma}_+$, the maximum likelihood (ML) estimates of π_1, π_2 are the solution of $\frac{dl}{d(\pi_1, \pi_2)}(\pi_1, \pi_2, \gamma_+ = \hat{\gamma}_+) = 0$. Under the hypothesis $H_1: \pi_1 = \pi_2 = \pi$, the ML estimate of π is the solution of $\frac{dl}{d\pi}(\pi_1 = \pi, \pi_2 = \pi, \gamma_+ = \hat{\gamma}_+) = 0$. We test the null hypothesis using a likelihood ratio statistic of the form

$$\begin{aligned} D &= 2l(\pi_1 = \hat{\pi}_1, \pi_2 = \hat{\pi}_2, \gamma_+ = \hat{\gamma}_+) \\ &\quad - 2l(\pi_1 = \hat{\pi}_1, \pi_2 = \hat{\pi}, \gamma_+ = \hat{\gamma}_+), \end{aligned} \quad (5)$$

which asymptotically follows the chi-squared distribution χ_{q-1}^2 with $q - 1$ degrees of freedom. In comparisons across c groups, the number of degrees of freedom is $(c - 1) \times (q - 1)$. After all genes are tested, p-values can be adjusted for multiple comparisons with the Benjamini-Hochberg method.

In a DTU analysis, groups are defined by the design of an experiment and are the same for each gene. In tuQTL analyses, the aim is to find nearby (bi-allelic) SNPs associated with transcript usage of a gene. Model fitting and testing is performed for each gene-SNP pair, and grouping of samples is defined by the genotype, typically translated into the number of minor alleles (0, 1 or 2). Thus, tuQTL analyses are similar to DTU analyses with the difference that multiple models are fitted and tested for each gene. Additional challenges to be handled in tuQTL analyses include a large number of tests per gene with highly variable allele frequencies (models) and linkage disequilibrium, which can be accounted for in the multiple testing corrections. As in other sQTL studies^{35,49,50}, we apply a permutation approach to empirically assess the null distribution of associations and use it for the adjustment of nominal p-values (see Supplementary Note 2 in [Supplementary File](#)). For computational efficiency, SNPs within a given gene that exhibit the same genotypes are grouped into blocks. In this way, blocks define unique models to be fit, reducing computation and the degree of multiple testing correction.

Dispersion estimation with adjusted profile likelihood and moderation

Accurate parameter estimation is a challenge when only a small number of replicates is available. Following the *edgeR* strategy^{1,2,53}, we propose multiple approaches for dispersion estimation, all based on the maximization and adjustment of the profile likelihood, since standard maximum likelihood (ML) is known to produce biased estimates as it tends to underestimate variance parameters by not

allowing for the fact that other unknown parameters are estimated from the same data^{54,55}.

In the DM model parameterization of our choice, we are interested in estimating the precision (concentration) parameter, γ_+ (inverse proportional to dispersion θ). Hence, at this stage, proportions π_1 and π_2 can be considered nuisance parameters and the profile log-likelihood (PL) for γ_+ can be constructed by maximizing the log-likelihood function with respect to proportions π_1 and π_2 for fixed γ_+ :

$$PL(\gamma_+; \hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \mathbf{y}) = \max_{\pi_1, \pi_2} l(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \gamma_+, \mathbf{y}). \quad (6)$$

The profile likelihood is then treated as an ordinary likelihood function for estimation and inference about parameters of interest. Unfortunately, with large numbers of nuisance parameters, this approach can produce inefficient or even inconsistent estimates^{54,55}. To correct for that, one can apply an adjustment proposed by Cox and Reid⁵⁶ and obtain an adjusted profile likelihood (APL):

$$APL(\gamma_+; \hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \mathbf{y}) = PL(\gamma_+; \hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \mathbf{y}) - \frac{1}{2} \log(\det mI), \quad (7)$$

where \det denotes determinant and I is the observed information matrix for $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$. The interpretation of the correction term in APL is that it penalizes values of γ_+ for which the information about $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ is relatively large. When data consists of many samples, one can use gene-wise dispersion estimates, i.e., the dispersion is estimated for each gene $g = 1, \dots, G$ separately:

$$\arg \max \{APL_g(\gamma_+^g)\} = \arg \max \{APL(\gamma_+^g; \hat{\boldsymbol{\pi}}_1^g, \hat{\boldsymbol{\pi}}_2^g, \mathbf{y}^g)\}. \quad (8)$$

These estimates become more unstable as the sample size decreases. At the other extreme, one can assume a common dispersion for all genes and use all genes to estimate it:

$$\arg \max \left\{ \frac{1}{G} \sum_{g=1}^G APL_g(\gamma_+^g) \right\}. \quad (9)$$

Common dispersion estimates are more stable but the assumption of a single dispersion for all genes is rather strong, given that some genes are under tighter regulation than others^{57,58}. Thus, moderated dispersion is a trade-off between gene-wise and common dispersion and estimates are calculated with an empirical Bayes approach, which uses a weighted combination of the common and individual likelihood:

$$\arg \max \left\{ APL_g(\gamma_+^g) + W \cdot \frac{1}{G} \sum_{g=1}^G APL_g(\gamma_+^g) \right\}. \quad (10)$$

If a dispersion-mean trend is present (see Figure S16, Figure S17, Figure S28 and Figure S29 in Supplementary File), as commonly

observed in gene-level differential expression analyses^{1,3}, one can apply shrinkage towards this trend instead of to the common dispersion:

$$\arg \max \left\{ APL_g(\gamma_+^g) + W \cdot \frac{1}{|C|} \sum_{g \in C} APL_g(\gamma_+^g) \right\}, \quad (11)$$

where C is a set of genes that have similar gene expression as gene g and W is a weight defining the strength of moderation (see Supplementary Note 1 for further details).

Estimation and inference: simulations from the Dirichlet-multinomial model

We first investigated the performance of the DM model and the approach for parameter estimation and inference in the case where only few replicates are available. We performed simulations that correspond to a two-group comparison with no DTU (i.e. null model) where feature counts were generated from the DM distribution with identical parameters in both groups. Simulations were repeated 50 times for 1000 genes. In these simulations, we can vary the overall expression (m), number of features (q), proportions (prop) and sample size in one condition (n). Proportions follow a uniform or decaying distribution or are estimated based on *kallisto* transcripts or *HTSeq* exon counts from Kim *et al.* and Brooks *et al.* data (more details on these datasets below). In the first case, all genes have the same (common) dispersion, and in the second one, each gene has different (genewise) dispersion. Simulations for evaluating the dispersion moderation are intended to better resemble a real dataset. For these instances (repeated 25 times for 5000 genes), genes have expression, dispersion and proportions that were estimated from the real data. See Supplementary Note 3 for the additional details.

Figure 1A and **Figure S1** confirm that using the Cox-Reid adjustment (CR) improves the estimation (in terms of median absolute error and extreme error values) of the concentration parameter γ_+ in comparison to raw profile likelihood (PL) estimates. Additionally, the median error of concentration estimates for Cox-Reid APL is always lower than for PL or maximum likelihood (ML) used in the *dirmult* package⁷ (**Figure 1C**, **Figure S2**). This translates directly into the inference performance where the CR approach leads to lower false positive (FP) rate than other approaches (**Figure 1B**, **Figure S3**).

Accurate estimates of dispersion do not always lead to expected control of FP rate. Notably, using the true concentration parameters in genes with many features (with decaying proportions) results in higher than expected nominal FP rates (**Figure 1B**, **Figure S3** and **Figure S6A**). Meanwhile, for genes with uniform proportions, even with many features, the FP rate for true dispersion is controlled (**Figure S3** and **Figure S6B**). Also, the Cox-Reid adjustment tends to underestimate the concentration (overestimate dispersion) for genes with many features and decaying proportions, especially

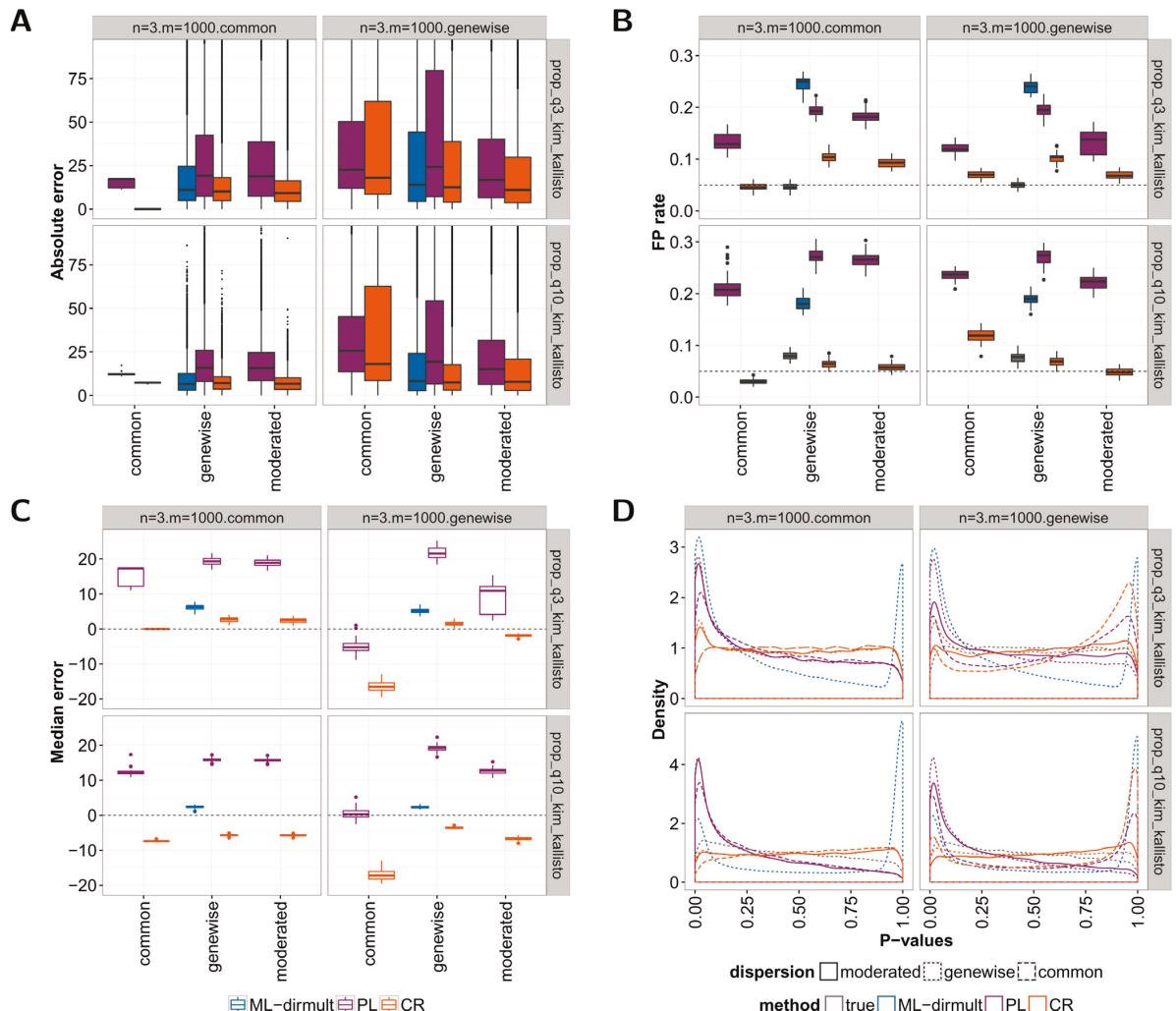


Figure 1. Results of two-group (3 versus 3 samples) DS analyses on data simulated from the DM null model. In the first scenario, all genes have the same (common) dispersion, and in the second one, each gene has a different (genewise) dispersion. All genes have expression equal to 1000 and 3 or 10 features with the same proportions estimated from *kallisto* counts from Kim *et al.* data set. For each of the scenarios, common, genewise, with and without moderation to common dispersion is estimated with maximum likelihood using the *dirmult* R package, the raw profile likelihood and the Cox-Reid APL. **A:** Absolute error of concentration γ_+ estimates. **B:** False positive (FP) rate for the p-value threshold of 0.05 of the null two-group comparisons based on the likelihood ratio statistics. Dashed line indicates the 0.05 level. **C:** Median raw error of γ_+ estimates. **D:** Distributions of p-values of the null two-group comparisons based on the likelihood ratio statistics. Additionally, results when true concentration estimates are used are indicated with the gray color.

for very small sample size (Figure 1C, Figure S2, Figure S5A, Figure S5E), which leads to accurate FP rate control not achieved even with the true dispersion (Figure S6A).

As expected, common dispersion estimation is effective when all genes indeed have the same dispersion, though this cannot be generally assumed in most real RNA-seq datasets (see results of

simulations in the following section). In contrast, pure gene-wise estimates of dispersion lead to relatively high estimation error in small sample sizes (Figure 1A, Figure S1 and Figure S8). Thus, sharing information about concentration (dispersion) between genes by moderating the gene-wise APL is applied. This improves concentration estimation in terms of median error (Figure 1C and Figure S8) and by shrinking extremely large values (on the

boundary of the parameter space, see Figure S7) toward common or trended concentration. Therefore, moderated gene-wise estimates lead to better control of the nominal FP rate (Figure 1B and Figure S10).

In these simulations, the overall best performance of the DM model is achieved when dispersion parameters are estimated with the Cox-Reid APL and the dispersion moderation is applied. This strategy leads to p-value distributions that in most of the cases are closer to the uniform distribution (Figure 1D, Figure S4 and Figure S11).

Comparison on simulations that mimic real RNA-seq data

Next, we use the simulated data from Soneson *et al.*²³, where RNA-seq reads were generated such that 1000 genes had isoform switches between two conditions of the two most abundant transcripts. For each condition three replicates were simulated resulting in 3 versus 3 comparisons. Altogether, we summarize results for three scenarios: i) Drosophila melanogaster with no differential gene expression; ii) Homo sapiens without differential gene expression; iii) Homo sapiens with differential gene expression.

The aim of these analyses is to compare the performance of DRIMSeq against DEXSeq, which emerged among the top

performing methods for detection of DTU from RNA-seq data²³. For DRIMSeq, we consider different dispersion estimates: common, gene-wise with no moderation and with moderation-to-common and to-trended dispersion. We use the exonic bin counts provided by HTSeq (same input to the DEXSeq pipeline), and transcript counts obtained with kallisto. Additionally, we use HTSeq and kallisto counts that are re-estimated after the removal of lowly expressed transcripts (less than 5% in all samples) from the gene annotation (pre-filtering) as proposed by Soneson *et al.*²³ and kallisto filtered counts that exclude the lowly expressed transcripts (also less than 5% in all samples). DRIMSeq returns a p-value per gene. To make results comparable, we used the module within DEXSeq that summarizes exon-level p-values to a gene-level adjusted p-value.

As expected, common dispersion estimates lead to worse performance (lower power and higher FDR) compared to gene-wise dispersions. DRIMSeq achieves the best performance with moderated gene-wise dispersion estimates, while the difference in performance between moderating to common or to trended dispersion is quite small, with moderated-to-trend dispersion estimates being slightly more conservative (Figure 2 and Figure S15).

As noted by Soneson *et al.*²³, detecting DTU in human is harder than in fruit fly due to the more complex transcriptome of the first

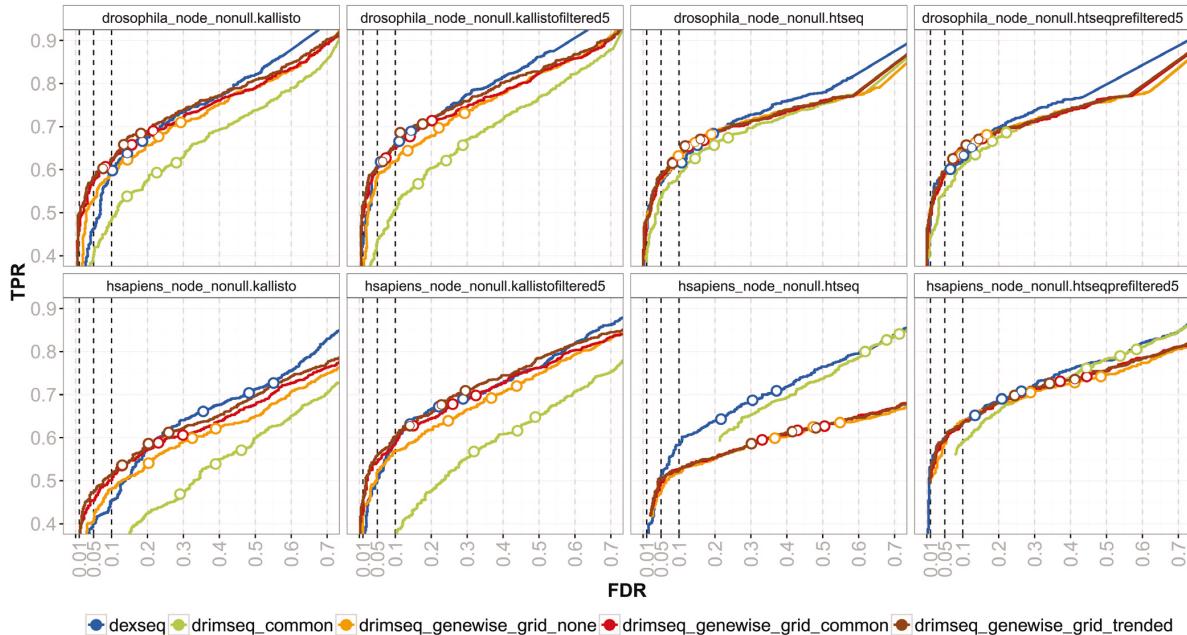


Figure 2. True positive rate (TPR) versus achieved false discovery rate (FDR) for three FDR thresholds (0.01, 0.05 and 0.1) obtained by DEXSeq and DRIMSeq. DRIMSeq was run with different dispersion estimation strategies: common dispersion and gene-wise dispersion with no moderation (genewise_grid_none), moderation to common dispersion (genewise_grid_common) and moderation to trended dispersion (genewise_grid_trended). Results presented for Drosophila melanogaster and Homo sapiens simulations with DTU (nonull) and no differential gene expression (node) using transcript counts from kallisto and exonic counts from HTSeq. Additionally, filtered counts (kallistofiltered5, htseqprefiltered5) are used. When the achieved FDR is smaller than the threshold, circles are filled with the corresponding color, otherwise, they are white.

one; all methods have much smaller false discovery rate (FDR). Nevertheless, none of the methods manages to control the FDR at a given threshold in either of the simulations.

Annotation pre-filtering, suggested as a solution to mitigate high FDRs²³, affects *DEXSeq* and *DRIMSeq* in a different way. For *DEXSeq*, it strongly reduces the FDR. For *DRIMSeq*, it increases power without a strong reduction of FDR. Moreover, the results for *kallisto* filtered and pre-filtered are almost identical (Figure S15 and Figure S24), which means that the re-estimation step based on the reduced annotation is not necessary for *kallisto* when used with *DRIMSeq* or *DEXSeq*. Additionally, we have considered how other filtering approaches affect DTU detection.

From Figure S24, we can see that DS analysis based on transcript counts are more robust to different variations of filtering and indeed some filtering improves the inference. For exonic counts, filtering should be less stringent and the pre-filtering approach is the best performing strategy.

DRIMSeq performs well when coupled with transcript counts from *kallisto*. In the case when no filtering is applied to the data, it outperforms *DEXSeq*. When transcript counts are pre-filtered, both methods have very similar performance (Figure S15). For both differential engines, the performance decreases substantially with increasing number of transcripts per gene, with *DRIMSeq* having slightly more power when genes have only a few transcripts (Figure S17). *DRIMSeq* has poor performance for the exonic counts in the human simulation, where achieved FDRs of more than 50% are observed for an expected 5%; consequently, we recommend the use of *DRIMSeq* on transcript counts only. On the other hand, the concordance of *DRIMSeq* and *DEXSeq* top-ranked genes is quite high and similar even for exonic counts (Figure S16).

The p-value distributions highlight a better fit of the DM model to transcript counts compared to exonic counts (it is more uniform with a sharp peak close to zero). Similarly, dispersion estimation gives better results for transcript counts (Figure S19 and Figure S20). In particular, for exonic counts, a large number of genes have concentration parameter estimates at the boundary of the parameter space, unlike the situation for transcript counts (Figure S19 and Figure S20).

DS analyses on real datasets

To compare the methods further, we consider two public RNA-seq data sets. The first is the *pasilla* dataset⁵⁹ (Brooks *et al.*). The aim was to identify genes regulated by *pasilla*, the Drosophila ortholog of mammalian splicing factors *NOVA1* and *NOVA2*. In this experiment, libraries were prepared from seven biologically independent samples: four control samples and three samples in which *pasilla* was knocked down. Libraries were sequenced using a mixture of single-end and paired-end reads as well as different read lengths. The second data set is from matched human lung normal and adenocarcinoma samples from six Korean female nonsmoking patients⁶⁰, using paired-end reads (Kim *et al.*).

Both datasets have a more complex design than those used in the simulations; in addition to the grouping variable of interest, there are additional covariates to adjust for (e.g., library layout for the

fruit fly data, patient identifier for the paired human study). In order to account for such effects, one should rather use a regression approach, which currently is not supported by *DRIMSeq*, but can be applied within *DEXSeq*'s GLM framework. To make the comparison fair, we fit multiple models. For the *pasilla* dataset, we compare four control samples versus three *pasilla* knock-down samples without taking into account the library layout (model full) as well as compare only the paired-end samples, which removes the covariate. To not diminish *DEXSeq* for its ability to fit more complex models, we run it using a model that does the four control versus three knock-down comparison with library layout as an additional covariate (model full 2). For the adenocarcinoma data, we do a two-group comparison of six normal versus six cancer samples (model full) and for *DEXSeq*, we fit an extra model that takes into account patient effects (model full 2). Additionally, we do so-called “mock” analyses where samples from the same condition are compared (model null), and the expectation is to detect no DS since it is a within-condition comparison (see Supplementary Note 5 for the exact definition of these null models).

In the full comparisons with transcript counts, *DRIMSeq* calls similar or fewer DS genes than *DEXSeq*, and a majority of them are contained within the *DEXSeq* calls (Figure S26, Figure S27) showing high concordance between *DRIMSeq* and *DEXSeq* and slightly more conservative nature of *DRIMSeq*. Accounting for covariates in *DEXSeq* (model full 2) or performing the analysis on a subgroup without covariates (model full paired) results in more DS genes detected (Figure S28, Figure S29 and Figure S30).

In the “mock” analyses, as expected, both methods detect considerably fewer DS genes, except in two cases. First, for the *pasilla* data (model null 3), where the two *versus* two control samples were from single-end library in one group and from paired-end library in the second group, leading to a comparison between batches in which both of the methods found more DS genes than in the comparison of control versus knock-down showing that the “batch” effect is very strong. Second, in the adenocarcinoma data (model null normal 1), where the two groups of individuals (each consisting of three women) happened to be very distinct (Figure S25). Therefore, we do not include these two cases when referring to the null models.

Overall, in the full comparisons, there are more DS genes detected based on differential transcript usage than differential exon usage (Figure S26). For *DEXSeq*, this is also the case in the null comparisons, which shows that *DEXSeq* works better with exonic counts than with transcript counts. *DRIMSeq*, on the other hand, has better performance on transcript counts, for which it calls less DS genes in the null analysis than with exon counts. In particular, the p-value distributions under the null indicate that DM fits better to transcript counts than exon counts (Figure S14, Figure S31 and Figure S32).

Method comparisons based on real data are very challenging as the truth is simply not known. In this sense the *pasilla* data is very precious, as the authors of this study have validated alternative usage of exons in 16 genes using RT-PCR. Of course, these validations represent an incomplete truth, and ideally, large-scale independent validation would be needed to comprehensively compare the DTU detection methods. In Figure 3, Figure S33, Figure S34 and

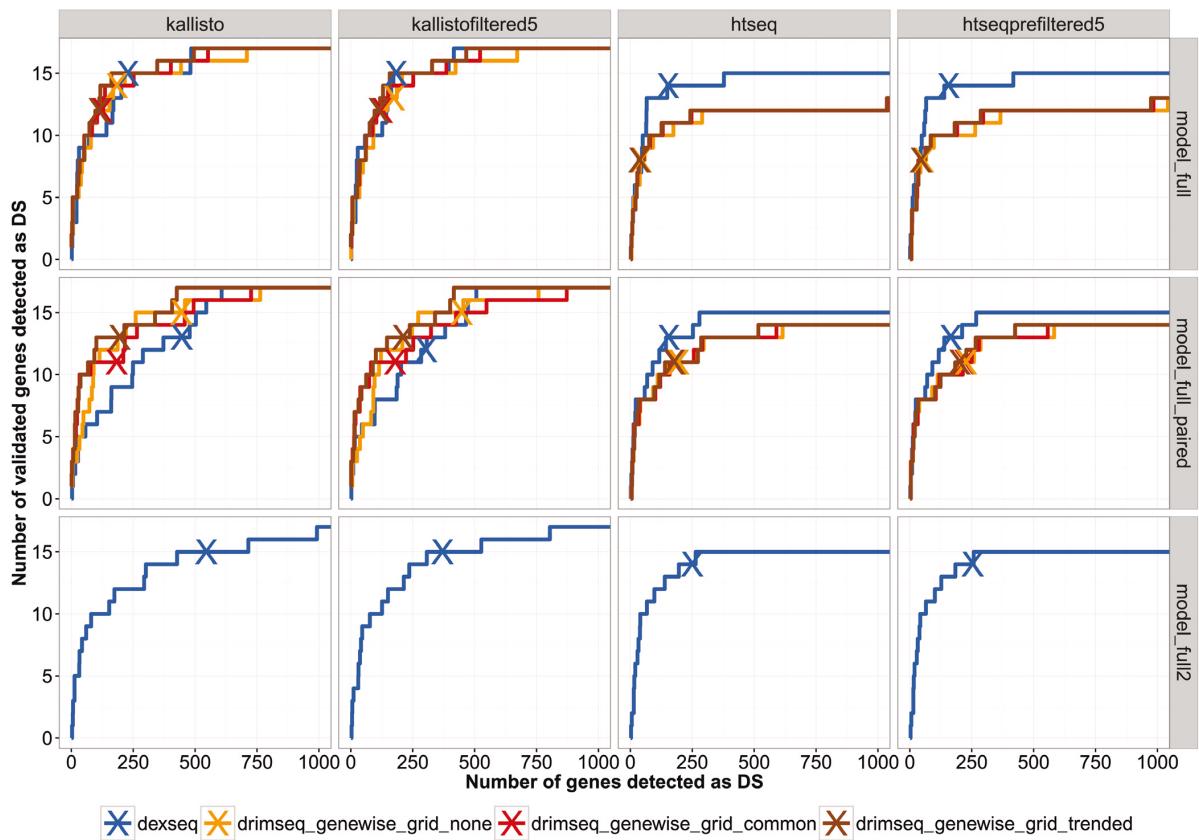


Figure 3. Results of DS analysis on the pasilla dataset showing how many of the 16 validated genes are called by DRIMSeq and DEXSeq using different counting strategies and different models. On each curve, “X” indicates the number of DS genes detected for the FDR of 0.05. Model full - comparison of 4 control samples versus 3 knock-down. Model full paired - comparison of 2 versus 2 paired-end samples. Model full 2 - as model full but including the information about library layout (no results for DRIMSeq because currently, it is not able to fit models with multiple covariates).

Figure S35 again it is shown that DRIMSeq is slightly more conservative than DEXSeq. DRIMSeq performs poorly on exon-level but returns strong performance on transcript-level quantification (e.g., kallisto) and even outperforms DEXSeq when the sample size is very small (model full paired).

tuQTL analyses

To demonstrate the application of DRIMSeq to tuQTL analysis, we use the data from the GEUVADIS project⁴⁶ where 465 RNA-seq samples from lymphoblastoid cell lines were sequenced, 422 of which were sequenced in the 1000 Genomes Project Phase 1. Here, we present the analysis of 91 samples corresponding to the CEU population and 89 samples from the YRI population. Expected transcript counts (obtained with Flux Capacitor) and genotype data were downloaded from the GEUVADIS project website. We choose to compare the performance of DRIMSeq with sQTLseekerR, because it is a very recent tool that performs well³⁵, can be directly applied to transcript count data and models transcript usage as a multivariate outcome.

For both of the methods, we investigate only the bi-allelic SNPs with a minor allele present in at least five samples (minor allele frequency approximately equal to 5%) and at least two alleles present in a population. Given a gene, we keep the SNPs that are located within 5 Kb upstream or downstream of the gene. We use the same pre-filtered counts in DRIMSeq and sQTLseekerR to have the same baseline for the comparison of the statistical engines offered by these packages. We keep the protein coding genes that have at least 10 counts in 70 or more samples and at least two transcripts left after the transcript filtering, which keeps the one that has at least 10 counts and proportion of at least 5% in 5 or more samples. The numbers of tested and associated genes and tuQTLs are indicated in Figure 4, Figure S38 and Figure S39.

In Figure 4A and Figure S40 we can see that the concordance between DRIMSeq and sQTLseekerR is quite high and reaches 75%. Nevertheless, there is considerable difference between the number and type of genes that are uniquely identified by each method. sQTLseekerR finds more genes with alternative splicing associated

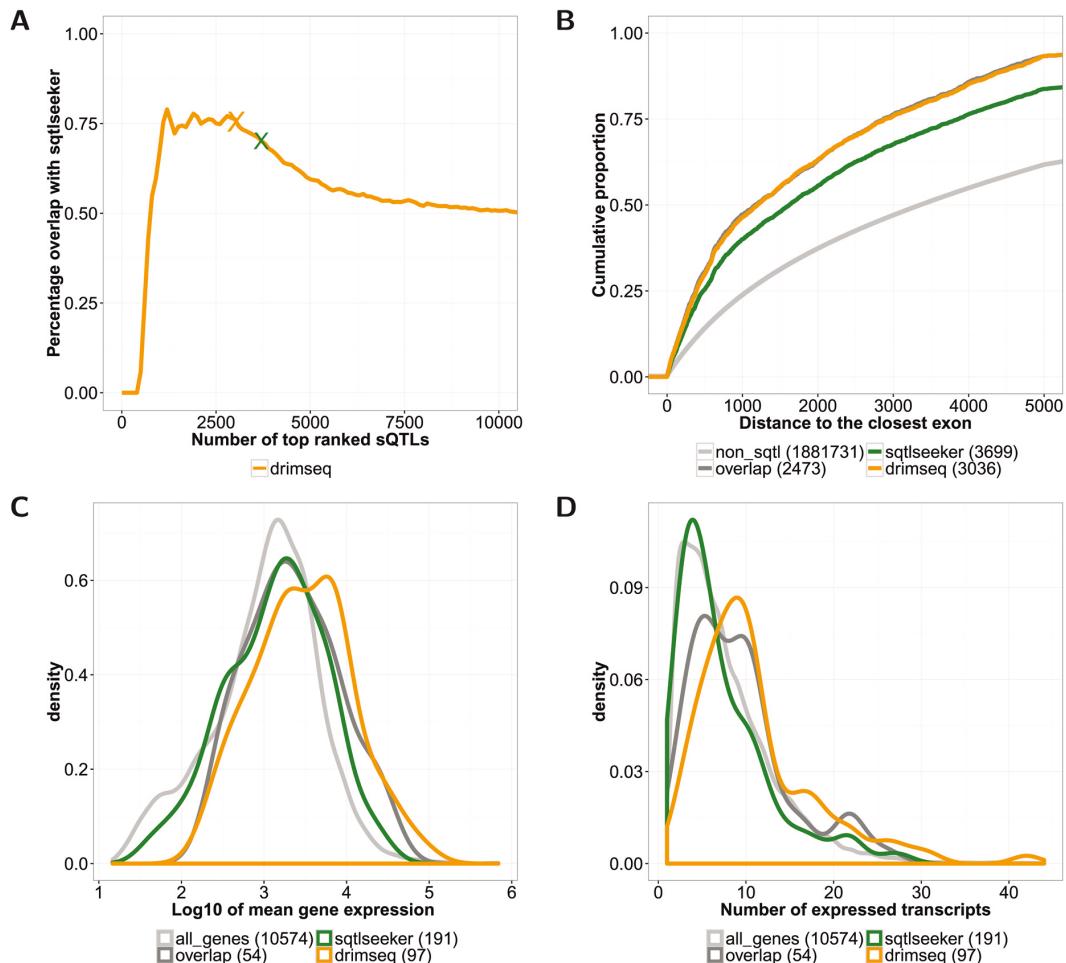


Figure 4. Results of the tuQTL analysis on the CEU population from the GEUVADIS data. **A:** Concordance between *sQTLseeker* and *DRIMSeq*. "X" indicates number of tuQTLs for FDR = 0.05. Panel **B**, **C** and **D** show characteristics of tuQTLs and genes detected by *sQTLseeker* or *DRIMSeq* for FDR = 0.05. Values in brackets indicate numbers of tuQTLs or genes in a given set. Dark gray line corresponds to tuQTLs or genes that were identified by both of the methods (overlap). The light gray line (non_sqtl) corresponds to intronic tuQTLs that were not called by any of the methods. **C:** Distribution of mean gene expression for genes that are associated with tuQTLs. **D:** Distribution of the number of expressed transcripts for genes that are associated with tuQTLs. The light gray lines (all_genes) represent corresponding features for all the analyzed genes.

to genetic variation (Figure S38 and Figure S39), but these genes have fewer transcripts expressed and lower overall expression in comparison to genes detected by *DRIMSeq* (Figure 4C, Figure 4D, Figure S40C and Figure S40D). To further investigate characteristics of detected tuQTLs, we measured enrichment of splicing-related features as used in a previous comparison³⁵. This includes

the location of tuQTLs within exons, within splice sites, in the surrounding of GWAS SNPs and distance to the closest exon. tuQTLs detected by *DRIMSeq* show higher enrichment for all splicing related features (Table 1 and Figure 4B), than *sQTLseeker* tuQTLs, suggesting that by accounting for the overall gene expression, one can detect more meaningful associations.

Table 1. Enrichment in splicing related features for tuQTLs detected by *DRIMSeq* and *sQTLseeker* in CEU and YRI populations for FDR = 0.05.

	% within exons		% within splice sites		% within 1Kb of a GWAS	
	CEU	YRI	CEU	YRI	CEU	YRI
DRIMSeq	26.09	35.89	19.76	21.42	12.75	15.43
sQTLseeker	20.95	25.43	13.52	17.4	10.22	10.09
Overlap	26.85	40.58	16.17	25.36	13.42	18.14
Non tuQTLs	5.25	5.24	1.75	1.53	1.15	0.97

Discussion

We have created a statistical framework called *DRIMSeq* based on the Dirichlet-multinomial distribution to model alternative usage of transcript isoforms from RNA-seq data. We have shown that this framework can be used for detecting differential isoform usage between experimental conditions as well as for identifying tuQTLs. In principle, the framework is suitable for differential analysis of any type of multinomial-like responses. From our simulations and real data analyses towards DS and sQTL analyses, *DRIMSeq* seems better suited to model transcript counts rather than exonic counts.

Overall, there are many tradeoffs to be made in DS analyses. For example, deriving transcript abundances from RNA-seq data is more difficult (e.g., complicated overlapping genes at medium to low expression levels) than directly counting exon inclusion levels of specific events. On the other hand, local splicing events may be not able to capture biologically interesting splice changes (e.g., switching between two different transcripts) but have ultimately more ability to detect DS in case when the transcript catalog is incomplete. Despite these tradeoffs and given the results observed here, *DRIMSeq* finds its place as a method to make downstream calculations on transcript quantifications. With emerging technologies that sequence longer DNA fragments (either truly or synthetically), we may see in the near future more direct counting of full-length transcripts, making transcript-level quantification more robust and accurate. Even with current standard RNA-seq data, ultrafast and lightweight methods make transcript counting more accessible and users will want to make comparative analyses directly from these estimates.

In principle, existing DS methods that allow multiple group comparisons could be adapted to the sQTL framework and *vice versa*; *DRIMSeq* is one of few tools that bridge these two applications. In particular, parameter estimation with *DRIMSeq*

is suited for a situation where only a few replicates are available per group (common in DS analysis) as well as analyses over larger sample sizes (typical in sQTL analysis). For small sample sizes, accurate dispersion estimation is especially challenging. Thus, we incorporate estimation techniques analogous to those used in negative binomial frameworks, such as Cox-Reid APL; perhaps not surprisingly, raw profile likelihood or standard maximum likelihood approaches do not perform as well in our tests of estimation performance. In addition, as with many successful genomics modeling frameworks, sharing information across genes leads to more stable and accurate estimation and therefore better inference (e.g., better control of nominal FP rates).

In comparison to other available methods, *DRIMSeq* seems to be more conservative than both *DEXSeq* (using transcript counts) and *sQTLseeker*, identifying fewer DTU genes and tuQTLs, respectively. On the other hand, *DEXSeq* is known to be somewhat liberal²³. Moreover, the sQTL associations detected by *DRIMSeq* have more enrichment in splicing-related features than *sQTLseeker* tuQTLs, which could be due to the fact that *DRIMSeq* accounts for the higher uncertainty of lowly expressed genes by using transcript counts instead of transcript ratios.

Our developed DM framework is general enough that it can be applied to other genomic data with multivariate count outcomes. For example, PolyA-seq data quantifies the usage of multiple RNA polyadenylation sites. During polyadenylation, poly(A) tails can be added to different sites and thus more than one transcript can be produced from a single gene (alternative polyadenylation); comparisons between groups of replicates can be conducted with *DRIMSeq*. As mentioned, the DM distribution is a multivariate generalization of the beta-binomial distribution, as the binomial and beta distributions are univariate versions of the multinomial and Dirichlet distributions, respectively. Although untested here, the *DRIMSeq* framework could be applied to analyses where the beta-binomial distribution are used with the advantage of naturally accommodating small-sample datasets. Interesting beta-binomial-based analyses include differential methylation using bisulfite sequencing data, where counts of methylated and unmethylated cytosines (a bivariate outcome) at specific genomic loci are compared, or allele-specific gene expression, where the expression of two alleles (again, a bivariate outcome) are compared across experimental groups.

One particularly important future enhancement is a regression framework, which would allow direct analysis of more complex experimental designs. For example, covariates such as batch, sample pairing or other factors could be adjusted for in the model. In the tuQTL analysis, it would allow studying samples from the pooled populations, with the subpopulation as a covariate, allowing larger

sample sizes and increased power to detect interesting changes. Another potential limitation is that *DRIMSeq* treats transcript estimates as fixed, even though they have different uncertainty, depending on the read coverage and complexity of the set of transcripts within a gene. Although untested here, propagation of this uncertainty could be achieved by incorporating observational weights that are inversely proportional to estimated uncertainties or, in case of fast quantification methods like *kallisto*, by making effective use of bootstrap samples. At this stage, there is no consensus on how these approaches will perform and ultimately may require considerable additional computation.

Software availability

The Dirichlet-multinomial framework described in this paper is implemented within an R package called DRIMSeq. In addition to the user friendly workflow for the DTU and tuQTL analyses, it provides plotting functions that generate diagnostic figures such as the dispersion versus mean gene expression figures and histograms of p-values. User can also generate figures of the observed proportions and the DM estimated ratios for the genes of interest to visually investigate their individual splicing patterns.

The release version of *DRIMSeq* is available on Bioconductor <http://bioconductor.org/packages/DRIMSeq>, and the latest development version can be found on GitHub <https://github.com/markrobinsonuzh/DRIMSeq>.

Data availability

Data for simulations that mimic real RNA-seq was obtained from Soneson *et al.*²³, where all the details on data generation and accessibility are available.

Differential splicing analyses were performed on the publicly available pasilla dataset, which was downloaded from the NCBI's Gene Expression Omnibus (GEO) under the accession number GSE18508, and adenocarcinoma dataset under the accession number GSE37764.

Data for the tuQTL analyses was downloaded from the GEUVADIS project website.

All the details about data availability and preprocessing are described in the [Supplementary Materials](#).

Archived source code as at the time of publication

DRIMSeq analyses for this paper were done with version 0.3.3 available on Zenodo [https://zenodo.org/record/53084⁶¹](https://zenodo.org/record/53084) and Bioconductor release 3.2. Source code used for the analyses in this paper is available on Zenodo [https://zenodo.org/record/167305⁶²](https://zenodo.org/record/167305).

Author contributions

MN drafted the manuscript, designed the analyses, analyzed the data and implemented the *DRIMSeq* R package. MDR drafted the manuscript and designed the overall study. All authors read and approved the final manuscript and have agreed to the content.

Competing interests

No competing interests were disclosed.

Grant information

MN acknowledges the funding from a Swiss Institute of Bioinformatics (SIB) Fellowship. MDR would like to acknowledge funding from an Swiss National Science Foundation (SNSF) Project Grant (143883).

Acknowledgments

The authors wish to thank Magnus Rattray, Torsten Hothorn and members of the Robinson lab for helpful discussions with special acknowledgment for Charlotte Soneson and Lukas Weber for careful reading of the manuscript.

Supplementary material

Supplementary File 1. Contains supplementary figures and tables referred to in the text. It also contains descriptions of dispersion moderation and p-value adjustment in tuQTL analysis and details about the simulations and real data analyses.

[Click here to access the data](#)

References

1. McCarthy DJ, Chen Y, Smyth GK: Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 2012; 40(10): 4288–4297.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Robinson MD, Smyth GK: Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*. 2008; 9(2): 321–332.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Anders S, Huber W: Differential expression analysis for sequence count data. *Genome Biol.* 2010; 11(10): R106.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Ritchie ME, Phipson B, Wu D, *et al.*: Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015; 43(7): e47.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

5. Law CW, Chen Y, Shi W, et al.: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol.* 2014; 15(2): R29. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Mosimann JE: **On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions.** *Biometrika*. 1962; 49(1–2): 65–82. [Publisher Full Text](#)
7. Tvedebrink T: **Overdispersion in allelic counts and θ -correction in forensic genetics.** *Theor Popul Biol.* 2010; 78(3): 200–210. [PubMed Abstract](#) | [Publisher Full Text](#)
8. Chen J, Li H: **Variable Selection for Sparse Dirichlet-Multinomial Regression With an Application To Microbiome Data Analysis.** *Ann Appl Stat.* 2013; 7(1): 418–442. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Finak G, McDavid A, Chattopadhyay P, et al.: **Mixture models for single-cell assays with applications to vaccine studies.** *Biostatistics*. 2014; 15(1): 87–101. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Samb R, Khadraoui K, Belleau P, et al.: **Using informative Multinomial-Dirichlet prior in a t-mixture with reversible jump estimation of nucleosome positions for genome-wide profiling.** *Stat Appl Genet Mol Biol.* 2015; 14(6): 517–532. [PubMed Abstract](#) | [Publisher Full Text](#)
11. Mosimann JE: **On the Compound Negative Multinomial Distribution and Correlations Among Inversely Sampled Pollen Counts.** *Biometrika*. 1963; 50(1–2): 47–54. [Publisher Full Text](#)
12. Farewell DM, Farewell VT: **Dirichlet negative multinomial regression for overdispersed correlated count data.** *Biostatistics*. 2013; 14(2): 395–404. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Sun D, Xi Y, Rodriguez B, et al.: **MOABS: model based analysis of bisulfite sequencing data.** *Genome Biol.* 2014; 15(2): R38. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Park Y, Figueroa ME, Rozek LS, et al.: **MethylSig: a whole genome DNA methylation analysis pipeline.** *Bioinformatics*. 2014; 30(17): 2414–22. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Feng H, Connelly KN, Wu H: **A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data.** *Nucleic Acids Res.* 2014; 42(8): e69. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Wang ET, Sandberg R, Luo S, et al.: **Alternative isoform regulation in human tissue transcriptomes.** *Nature*. 2008; 456(7221): 470–6. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Wang GS, Cooper TA: **Splicing in disease: disruption of the splicing code and the decoding machinery.** *Nat Rev Genet.* 2007; 8(10): 749–61. [PubMed Abstract](#) | [Publisher Full Text](#)
18. Tazi J, Bakkour N, Stamm S: **Alternative splicing and disease.** *Biochim Biophys Acta*. 2009; 1792(1): 14–26. [PubMed Abstract](#) | [Publisher Full Text](#)
19. Hooper JE: **A survey of software for genome-wide discovery of differential splicing in RNA-Seq data.** *Hum Genomics*. 2014; 8(1): 3. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics*. 2010; 26(1): 139–140. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Derti A, Garrett-Engele P, Macisaac KD, et al.: **A quantitative atlas of polyadenylation in five mammals.** *Genome Res.* 2012; 22(6): 1173–1183. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Alamancos GP, Agire E, Eyras E: **Methods to study splicing from high-throughput RNA sequencing data.** *Methods Mol Biol.* 2014; 1126: 357–397. [PubMed Abstract](#) | [Publisher Full Text](#)
23. Soneson C, Matthes KL, Nowicka M, et al.: **Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage.** *Genome Biol.* 2016; 17(1): 12. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Liao Y, Smyth GK, Shi W: **FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics*. 2014; 30(7): 923–930. [PubMed Abstract](#) | [Publisher Full Text](#)
25. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from RNA-seq data.** *Genome Res.* 2012; 22(10): 2008–2017. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Anders S, Pyl PT, Huber W: **HTSeq—a Python framework to work with high-throughput sequencing data.** *Bioinformatics*. 2015; 31(2): 166–169. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Ongen H, Dermitzakis ET: **Alternative Splicing QTLs in European and African Populations.** *Am J Hum Genet.* 2015; 97(4): 567–575. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. Katz Y, Wang ET, Airoldi EM, et al.: **Analysis and design of RNA sequencing experiments for identifying isoform regulation.** *Nat Methods.* 2010; 7(12): 1009–1015. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Shen S, Park JW, Lu ZX, et al.: **IMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.** *Proc Natl Acad Sci U S A.* 2014; 111(51): E5593–601. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Alamancos GP, Pagès A, Trincado JL, et al.: **Leveraging transcript quantification for fast computation of alternative splicing profiles.** *RNA*. 2015; 21(9): 1521–1531. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Goldstein LD, Cao Y, Pau G, et al.: **Prediction and Quantification of Splice Events from RNA-Seq Data.** *PLoS One.* 2016; 11(5): e0156132. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Zhao K, Lu ZX, Park JW, et al.: **GLIMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data.** *Genome Biol.* 2013; 14(7): R74. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Jia C, Hu Y, Liu Y, et al.: **Mapping Splicing Quantitative Trait Loci in RNA-Seq.** *Cancer Inform.* 2014; 13(Suppl 4): 35–43. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Hu Y, Liu Y, Mao X, et al.: **PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution.** *Nucleic Acids Res.* 2014; 42(3): e20. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Monlong J, Calvo M, Ferreira PG, et al.: **Identification of genetic variants associated with alternative splicing using sQTLseekerR.** *Nat Commun.* 2014; 5: 4698. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
36. Glauš P, Honkela A, Patrav M: **Identifying differentially expressed transcripts from RNA-seq data with biological variation.** *Bioinformatics*. 2012; 28(13): 1721–1728. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
37. Rossell D, Stephan-Otto Attolini C, Kroiss M, et al.: **Quantifying Alternative Splicing From Paired-End RNA-Sequencing Data.** *Ann Appl Stat.* 2014; 8(1): 309–330. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
38. Trapnell C, Williams BA, Pertea G, et al.: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol.* 2010; 28(5): 511–515. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics*. 2011; 12: 323. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Bernard E, Jacob L, Mairal J, et al.: **Efficient RNA Isoform Identification and quantification from RNA-Seq data with network flows.** *Bioinformatics*. 2014; 30(17): 2447–2455. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Patro R, Mount SM, Kingsford C: **Salishfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.** *Nat Biotechnol.* 2014; 32(5): 462–4. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Bray NL, Pimentel H, Melsted P, et al.: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; 34(5): 525–7. [PubMed Abstract](#) | [Publisher Full Text](#)
43. Patro R, Dugall G, Kingsford C: **Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment.** *bioRxiv*. 2015; 021592. [PubMed Abstract](#)
44. Kanitz A, Gypas F, Gruber AJ, et al.: **Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data.** *Genome Biol.* 2015; 16(1): 150. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
45. Teng M, Love MI, Davis CA, et al.: **A benchmark for RNA-seq quantification pipelines.** *Genome Biol.* 2016; 17(1): 74. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
46. Lappalainen T, Sammeth M, Friedländer MR, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature*. 2013; 501(7468): 506–11. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
47. Battle A, Mostafavi S, Zhu X, et al.: **Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.** *Genome Res.* 2014; 24(1): 14–24. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
48. Pickrell JK, Marioni JC, Pai AA, et al.: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature*. 2010; 464(7289): 768–772. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, et al.: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature*. 2010; 464(7289): 773–777. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
50. Ongen H, Buil A, Brown AA, et al.: **Fast and efficient QTL mapper for thousands of molecular phenotypes.** *Bioinformatics*. 2016; 32(10): 1479–85. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Trapnell C, Hendrickson DG, Sauvageau M, et al.: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nat Biotechnol.* 2013; 31(1): 46–53. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
52. Li YI, Knowles DA, Pritchard JK: **LeafCutter: Annotation-free quantification of RNA splicing.** *bioRxiv*. 2016. [PubMed Abstract](#) | [Publisher Full Text](#)

53. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics*. 2007; 23(21): 2881–2887.
[PubMed Abstract](#) | [Publisher Full Text](#)
54. Reid N, Fraser DAS: **Likelihood inference in the presence of nuisance parameters.** 2003; 7.
[Reference Source](#)
55. McCullagh P, Tibshirani R: **A Simple Method for the Adjustment of Profile Likelihoods.** *J R Stat Soc Series B Stat Methodol.* 1990; 52(2): 325–344.
[Reference Source](#)
56. Cox DR, Reid N: **Parameter orthogonality and approximate conditional inference.** *J R Stat Soc Series B Stat Methodol.* 1987; 49(1): 1–39.
[Reference Source](#)
57. Choi JK, Kim YJ: **Intrinsic variability of gene expression encoded in nucleosome positioning sequences.** *Nat Genet.* 2009; 41(4): 498–503.
[PubMed Abstract](#) | [Publisher Full Text](#)
58. Singh A, Soltani M: **Quantifying intrinsic and extrinsic variability in stochastic gene expression models.** *PLoS One.* 2013; 8(12): e84301.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
59. Brooks AN, Yang L, Duff MO, et al.: **Conservation of an RNA regulatory map between *Drosophila* and mammals.** *Genome Res.* 2011; 21(2): 193–202.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
60. Kim SC, Jung Y, Park J, et al.: **A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers.** *PLoS One.* 2013; 8(2): e55596.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
61. Nowicka M, Robinson MD: **Source code of the R package used for analyses in “DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics” paper.** *Zenodo.* 2016.
[Data Source](#)
62. Nowicka M, Robinson MD: **Source code of the analyses in the “DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics” paper.** *Zenodo.* 2016.
[Data Source](#)

Open Peer Review

Current Referee Status:  

Version 2

Referee Report 20 December 2016

doi:[10.5256/f1000research.11139.r18253](https://doi.org/10.5256/f1000research.11139.r18253)



Robert Castelo

Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, Spain

I appreciate that the authors have made an effort to address my comments and I'm particularly happy to see that my suggestion to check the overlap of tuQTLs with splice site binding sites reveals an improved enrichment by DRIMSeq. I also understand now the decision you took about not using the 'SummarizedExperiment' class. For the future development of DRIMSeq you may want to consider using the MultiAssayExperiment class (<http://bioconductor.org/packages/MultiAssayExperiment>) that allows multiple assay types over multiple sample sets.

The authors say that it is not worth made a comparison with Cuffdiff because in the study by Soneson¹ *et al.* (2016), where both authors of DRIMSeq were involved, Cuffdiff was very conservative in detecting differential isoform/transcript usage (DTU). In that paper the authors assess DTU by switching the two most abundant isoforms and show that Cuffdiff has a low true positive rate (TPR) at small magnitudes of the difference in relative abundance between the two most abundant isoforms per gene. However, in Supplementary Figure 10 of that paper, the authors show that at larger magnitudes of that difference, the TPR of Cuffdiff improves substantially while correctly controlling the false discovery rate (FDR).

In this paper the authors assess DTU following the same strategy of switching the two most abundant isoforms and I think it would be again very interesting to see how Cuffdiff and DRIMSeq compare at different magnitudes of the change in isoform usage. The authors also argue that Frazee² *et al.* (2014) find that Cuffdiff is very conservative. However, as far as I understand that paper, Frazee and co-workers are not evaluating DTU but differential transcript expression (DTE), and therefore, in my view, the experiments conducted on that paper do not warrant the conclusion that Cuffdiff is overly conservative for DTU.

The authors decided not to perform an enrichment analysis of tuQTLs on ESEs and ESSs because Lalonde³ *et al.* (2011) concluded that ESE predictions themselves are a poor indicator of the effect of SNPs on splicing patterns. However, Lalonde and co-workers scored ESE motifs with ESEfinder 3.0 (Cartegni⁴ *et al.* 2003), a method based on SELEX experiments conducted about 10 years ago and I would expect some advance in this field in the last decade. A recent study that seems to successfully use more recent ESE and ESS data to assess their enrichment with respect to polymorphisms is Supek⁴ *et al.* (2014).

While these two aspects remain, in my opinion, open, I think the statistical model of DRIMSeq proposed for DTU makes a lot sense and people interested in addressing biological questions that involve DTU

should give it try, and I'm happy to approve the paper.

References

1. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol.* 2016; **17**: 12 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT: Flexible isoform-level differential expression analysis with Ballgown. *bioRxiv*. 2014. 003665 [Publisher Full Text](#)
3. Lalonde E, Ha KC, Wang Z, Bemmo A, Kleinman CL, Kwan T, Pastinen T, Majewski J: RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.* 2011; **21** (4): 545-54 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B: Synonymous mutations frequently act as driver mutations in human cancers. *Cell.* 2014; **156** (6): 1324-35 [PubMed Abstract](#) | [Publisher Full Text](#)

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 06 July 2016

doi:[10.5256/f1000research.9577.r14580](https://doi.org/10.5256/f1000research.9577.r14580)



Robert Castelo

Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, Spain

This article introduces a new statistical method, called DRIMSeq and implemented in a R/Bioconductor package of the same name, to detect isoform expression changes between two conditions from RNA-seq data. The same method can be used to search for significant associations between SNPs and isoform quantifications obtained also from RNA-seq data (sQTLs). The main novelty of this method with respect to the existing literature on this problem, is the joint modelling of transcript quantification values derived from isoforms of the same gene, by using a Dirichlet-multinomial model. This allows the method to account of the intrinsic dependency between quantification values of these isoforms.

The assessment of DRIMSeq on differential isoform usage provides a comparison of its performance with DEXSeq¹, a statistical method for differential exon inclusion from RNA-seq data, as function of two different "isoform" quantification strategies: exonic-bin (not really "isoform") count values calculated with HTSeq and transcript-quantification values calculated with kallisto².

The experimental results make perfect sense, DRIMSeq works better than DEXSeq with transcript-quantification values and DEXSeq works better than DRIMSeq with exonic-bin count values. However, while both methods, and both types of "isoform" quantification input data, allow one to study the post-transcriptional processing of RNA transcripts, the kind of questions that can be addressed with each of them are different. Exonic-bin count values and DEXSeq can be used to investigate differential exon

inclusion across conditions, which is a consequence of differential isoform usage, while transcript-quantification values and DRIMSeq can be used to directly investigate differential isoform usage.

A potentially interesting outcome of this comparison in the paper could be some sort of guidelines about when it is more sensible to investigate differential exon inclusion or differential isoform usage, depending on factors such as the biological question at hand, sequencing depth or number of biological replicates. However, this is apparently beyond the scope of this paper and the experimental results are in principle geared towards convincing the reader that DRIMSeq improves on existing approaches to discover changes in isoform usage, as suggested in the abstract. In my view, the experimental results do not address this question and I would suggest the authors to compare DRIMSeq with methods that also work with transcript-quantification values and assess differential isoform usage such as, for instance, Cuffdiff³ or sleuth⁴.

The experimental results on searching for sQTLs compare favourably DRIMSeq with an existing tool for that purpose, sQLseekR⁵. Evaluating performance in this context is challenging and the idea of assessing enrichment with respect to splicing-related features is a good one. However, the (two) presented features in Table 1 could be made more precise. It is unclear that a SNP close to a GWAS hit should be necessarily related to splicing and it is also unclear why one should expect splicing-related enrichment more than a few hundred nucleotides away from the intervening exon. While it is technically interesting to see a method being used to address two completely different research questions, in my view, mixing both types of analyses makes the article less focused. I would argue that both questions deserve separate papers, and that would allow the authors to investigate in depth critical aspects of both types of analysis that are currently not addressed in the current article.

In summary, this article provides an interesting new methodology for the analysis of differential isoform usage from RNA-seq data, it is well-written and the implemented software runs smoothly and is well documented. However, in my view, the current experimental results of the article are not that informative for the reader to learn what advantages DRIMSeq provides over other tools for differential isoform usage analysis, and to decide whether he/she should be doing a differential isoform usage, or a differential exon inclusion analysis, if this were a goal of the comparison with DEXSeq.

Minor comments:

1. I would replace the term "edgeR ideology" in page 5 by "edgeR strategy".
2. In page 9 it is described that the distributions of raw p-values shown in Supplementary Figures S28 and S29 fit "better" when derived from transcript quantification values than from exonic-bin count values, but in fact in both cases the distributions are non-uniform for p-values distributed under the null hypothesis. This can be easily shown with the data from vignette of the DRIMSeq package when skipping the step that reduces the transcript set to analyze to speed up the building time of the vignette. This is not openly discussed in the paper but I would argue that it is quite critical to know under what technical assumptions the proposed hypothesis test leads to uniform raw p-values under the null, as this has a direct consequence on the control of the probability of the type-I error.
3. The sQTL analysis described in pages 9, 10 and 11 uses transcript-quantification values from FluxCapacitor. If the entire first part of the paper shows the performance metrics of DRIMSeq using kallisto, in my view, it would make more sense to use kallisto for this analysis as well.
4. With regard to the implementation in the R/Bioconductor software package DRIMSeq, the authors have implemented a specialized S4 object class called 'dmDSdata' to act as a container for counts and information about samples. Since the package forms part of the Bioconductor project, I think it

would better for both, the end-user and the developer authors, that the package re-uses the 'SummarizedExperiment' class as container for counts and sample information. This would facilitate the integration of DRIMSeq into existing or new workflows for the analysis of RNA-seq data. As an example of the limitations derived from providing a completely new specialized object class, the dimensions of a 'dmDSdata' object in terms of number of features and number of samples cannot be figured out using the expected call to the 'dim()' accessor method. Of course the authors may add that method to the 'dmDSdata' object class but, in general, there are obvious advantages derived from enabling data interoperability through the use of common data structures across Bioconductor software packages⁶.

References

1. Anders S, Reyes A, Huber W: Detecting differential usage of exons from RNA-seq data. *Genome Res.* 2012; **22** (10): 2008-17 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Bray NL, Pimentel H, Melsted P, Pachter L: Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* 2016; **34** (5): 525-7 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* 2013; **31** (1): 46-53 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Pimentel H: Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv*. 2016; **058164**. [Publisher Full Text](#) | [Reference Source](#)
5. Monlong J, Calvo M, Ferreira PG, Guigó R: Identification of genetic variants associated with alternative splicing using sQTLseeker. *Nat Commun.* 2014; **5**: 4698 [PubMed Abstract](#) | [Publisher Full Text](#)
6. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M: Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015; **12** (2): 115-21 [PubMed Abstract](#) | [Publisher Full Text](#)

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 25 Nov 2016

Mark Robinson, University of Zurich, Switzerland

Thank you for taking the time to read and review our paper.

DEXSeq is a package designed for the differential exon usage (DEU) and returns exon-level p-values, which can be also summarized to the gene level. In principle, *DEXSeq*'s implementation could be used to address the question of differential isoform/transcript usage (DTU) as well, which was done, for example, in the simulation study by Soneson *et al.* [1]. They use different counting strategies, among them transcript quantifications from *kallisto* [2], coupled with *DEXSeq*'s differential engine to detect differential transcript usage. *DRIMSeq*, based on the Dirichlet-multinomial model, was developed to detect differential usage of any kind of multivariate genomic features at the gene-level. Thus potentially, both *DEXSeq* and *DRIMSeq* can be applied to exon counts and to transcript quantifications. However, from our comparisons, which were

performed at the gene-level, the performance of *DEXSeq* and *DRIMSeq* is different on these different types of counts. *DEXSeq* performs better on exon counts and *DRIMSeq* on transcript counts.

We have not used *Cuffdiff* [3] in our comparisons here because in the study by Soneson *et al.* [1], it performed poorly compared to *DEXSeq*. In particular, *Cuffdiff* was very conservative having low false discovery rate (FDR) at the cost of very low power for detecting DTU. The conservative nature of *Cuffdiff* for differential transcript expression, was also pointed out by Frazee *et al.* [4]. We decided to compare *DRIMSeq* only to the top performing method, *DEXSeq*. The other tool proposed by the Reviewer, sleuth [5], is meant for differential transcript expression analyses, not DTU.

The scope of this paper was not to justify exon or transcript level analysis, for that one could refer to the comparison paper by Hooper [6], but to propose a methodologically-sound tool for differential isoform usage analysis or detect transcript usage QTLs based on transcript quantifications. We propose to use *DRIMSeq* since it outperformed *DEXSeq* in this type of analysis and there are no other tools for differential transcript usage that were intended for transcript level quantifications from the latest generation of fast quantification tools, such as *kallisto* [2] or *Salmon* [7].

Importantly, *DEXSeq* returns p-values per feature (exon or transcript), which can be also summarized to the gene level. *DRIMSeq* performs gene-level tests and returns p-values per gene only. When the interest is in detecting specific exons or isoforms that change, one should use *DEXSeq* because currently *DRIMSeq* does not provide any post hoc analysis (although in many cases, the relevant information can be deduced from looking at the relative transcript expression from *DRIMSeq*'s plots). We have not investigated the differences in performance due to sequencing depth or number of biological replicates, but we believe that the requirements would be basically the same in these terms for both of the methods. What matters is the completeness of annotation. Detecting DTU based on exon counts is generally more robust than that based on transcript quantifications when the annotation is incomplete, which was investigated in detail by Soneson *et al.* [1].

To compare the performance of *DRIMSeq* and *sQTLseeker*, we use the splicing-related features that were also used in the *sQTLseeker* paper [8] to compare *sQTLseeker* against other methods. The Reviewer suggested to consider other splicing-related features, such as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs) and splice sites. We have added the frequency of tuQTL overlapping with the splice sites to Table 1. However, we have not performed analyses on ESEs and ESSs since Lalonde *et al.* [9] concluded from their study that "ESE predictions themselves are a poor indicator of the effect of SNPs on splicing patterns".

By addressing differential splicing and sQTLs in one paper, our aim was to show that methods used for these analyses are based on statistical approaches that in the end tackle ultimately the same question: differential splicing between conditions. Both analyses employ the same methods for gene feature quantification and potentially one main differential engine could be used with slight analysis-specific adjustments, such as information sharing between genes for small sample size data or using genotypes as grouping factor, which is done in *DRIMSeq*. We believe we have addressed in sufficient depth aspects of both of these analyses providing comparisons on simulated and real data.

Addressing the minor comments:

- We have replaced the term "edgeR ideology" in page 5 by "edgeR strategy".
- As suggested, we have investigated in more depth, based on simulations from the DM model, the DRIMSeq p-value distributions under the null hypothesis of no differential transcript usage (Figures 1, S4, S6, S11, S14). Overall, using the Cox-Reid adjusted profile likelihood and the dispersion moderation leads to p-value distributions that in most cases are closer to the uniform distribution (Figures 1D, S4 and S11). The better fit of the DM model to transcript counts in comparison to exon counts can be seen in Figure S14, where the p-value distributions are more uniform for simulations that mimic *kallisto* counts than for simulations that mimic *HTseq* counts.
- Yes, using *kallisto* counts would be more consistent with the rest of our manuscript. Nevertheless, we decided to use the Flux Capacitor counts because they were already available on the GEUVADIS project website and have been used extensively in other projects, for example, in the sQTLseekeR paper. Moreover, we think that using other counts should not affect the comparison between DRIMSeq and sQTLseekeR.
- We had already considered the SummarizedExperiment class while developing the DRIMSeq package. However, it does not provide features and functionality that we need for storing the count data and DRIMSeq results. In particular, the dimensions of Assays in SummarizedExperiment must be the same. That is not the case for us for two reasons. Firstly, each gene has multiple transcripts and, for example, the table with proportion estimates per transcript is larger than a table with dispersion estimates which are available per gene. Second, in the QTL analysis, table with transcript counts has different dimensions than table with genotypes. Additionally, we use matrices instead of data frames to store our data because the former occupies less space. Specifically, we have created a class called MatrixList, which is adjusted to store data where each gene has multiple features quantified and allows a quick access to these counts in per gene basis. We have not implemented the dim() method on dmDSdata or dmSQLdata because we want to keep consistency between them and, for example, dmSQLdata contains transcript counts and genotypes which have different dimensions. Thus we decided to make the dim() methods available for the counts and genotypes slots in these classes but not for the classes themselves.

References

- [1] Charlotte Soneson, Katarina L Matthes, Małgorzata Nowicka, Charity W Law, and Mark D Robinson. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, 17(1):1–15, 2016.
- [2] Nicolas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, advance on, Apr 2016.
- [3] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, 2013.
- [4] A. C. Frazee, G. Pertea, a. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek. Flexible isoform-level differential expression analysis with Ballgown. *bioRxiv*, pages 0–13, 2014.
- [5] Harold J Pimentel, Nicolas Bray, Suzette Puente, Pall Melsted, and Lior Pachter. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv*, Jun 2016.
- [6] Joan E Hooper. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Human genomics*, 8:3, 2014.

- [7] Rob Patro, Geet Duggal, and Carl Kingsford. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*, page 021592, 2015.
- [8] Jean Monlong, Miquel Calvo, Pedro G. Ferreira, and Roderic Guigo'. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications*, 5(May):4698, Aug 2014.
- [9] Emilie Lalonde, Kevin C H Ha, Zibo Wang, Amandine Bemmo, Claudia L Kleinman, Tony Kwan, Tomi Pastinen, and Jacek Majewski. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Research*, 21(4):545–554, Apr 2011.

Competing Interests: No competing interests were disclosed.

Referee Report 24 June 2016

doi:[10.5256/f1000research.9577.r14338](https://doi.org/10.5256/f1000research.9577.r14338)



Alejandro Reyes

Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

Nowicka and Robinson propose a novel method, called DRIMSeq, to test for differential transcript usage between groups of samples using RNA-seq. The method is based on the Dirichlet-multinomial distribution.

The authors evaluate different existing approaches to estimate the parameters of their model using simulated experiments with a small number of replicates, which is a common scenario of high-throughput sequencing experiments. Furthermore, Nowicka *et al.* provide a proof of principle of their method by applying it to both simulated and real RNA-seq data. They also compare the performance of DRIMSeq with DEXSeq and sQTLseekeR in detecting differential transcript usage and splicing quantitative trait loci (sQTLs), respectively. DRIMSeq shows high concordance with DEXSeq. Furthermore, the authors demonstrate that DRIMSeq performs better than DEXSeq when using transcript-level counts. DRIMSeq and sQTLseekeR were also highly concordant. Nevertheless, sQTL genes detected by DRIMSeq were expressed higher than those detected by sQTLseekeR, and sQTLs detected by DRIMSeq were in closer proximity to exons compared to sQTLs detected by sQTLseeker. DRIMSeq is implemented as an R/Bioconductor package.

Overall, the manuscript is well presented and is scientifically sound. The description of the method is clear, the comparisons are fair, and the conclusions are supported by data and analyses.

Below some minor comments:

1. Transcription of multiple isoforms from a single gene can be the consequence of differences in the following molecular mechanisms: transcription start sites, splicing, and termination of transcription. The terms “differential splicing” and “splicing QTLs”, which are used throughout the manuscript and the package vignette, focus only on splicing. Consider a hypothetical example of an isoform switch between conditions in which the two isoforms only diverge by the transcription start site of the first exon. DRIMSeq should also detect this difference, and this would not be due to differential splicing. Thus, the authors could use more generic terminology that describes all possible interpretations of the outcome of their test. Perhaps “differential transcript usage” or “transcript usage QTLs”?

2. In equations 6-11, *PL* and *APL* are understandable from the context but are not defined in the text.
3. It would be useful for the reader to include more information of the simulated data from Soneson *et al.* (2016) in the main text of this manuscript (for example, number of replicates per condition).
4. The authors describe how DEXSeq can account for additional covariates in complex experimental designs. This paragraph, as well as the figures and supplementary material associated to it, could be understood as if DEXSeq fits GLMs only for complex experimental designs. In reality, DEXSeq always fits GLMs, even for simple two-group comparisons.
5. There are some panels from the supplementary figures where data are missing. Specifically, Fig. S13 has 3 empty panels and Fig. S21 the left panels are missing the data for “dexseq.prefilter5” and “drimseq_genewise_grid_trended.filter5”.
6. The list of software for splicing event quantification is already very extensive, however a citation to the Bioconductor package SGSeq (Goldstein *et al.*, 2016) could also be added.
7. As for the readability of the supplementary information, some abbreviations are not defined in each supplementary figure caption. For example, in Fig. S5, n, m, DM, FP and nr_features are not defined in its caption (some of them, however, are defined in previous captions). Since many abbreviations repeat several times through the supplementary information, it would be useful to include a glossary of all abbreviations at the beginning of all supplementary figures.

References

1. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol.* 2016; **17**: 12 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, Gentleman R: Prediction and Quantification of Splice Events from RNA-Seq Data. *PLoS One.* 2016; **11** (5): e0156132 [PubMed Abstract](#) | [Publisher Full Text](#)

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 25 Nov 2016

Mark Robinson, University of Zurich, Switzerland

Thank you for taking the time to read and review our paper.

As per your suggestion, we have now stressed that DRIMSeq can be applied to differential transcript usage (DTU), which accounts for not only differential splicing but also the differences in transcription start sites and differential transcript termination. In the QTL analysis, as we test for associations between genotypes and transcript usage and not only splicing, following your suggestion, we have also changed the term from splicing QTLs (sQTLs) to transcript usage QTLs

(tuQTLs).

We have addressed all the other minor comments which include:

- defining the abbreviations of profile likelihood (PL) and adjusted profile likelihood (APL),
- adding the sample size information about the simulations from Soneson *et al.* [1],
- in order to remove the misleading suggestion that DEXSeq fits GLMs only in the complex designs, we have changed the names of the models used in real data analysis from "model full glm" to "model full 2" and paraphrased the corresponding manuscript sections,
- we have included results for the panels with missing data in the Supplementary Figures S15, S16 and S24,
- we have included the citation to SGSeq [2] - the Bioconductor package for analyzing splice events from RNA-seq data,
- in the Supplementary Materials, we have prepared a section explaining abbreviations used in the subsequent Supplementary Figures.

References

- [1] Charlotte Soneson, Katarina L Matthes, Małgorzata Nowicka, Charity W Law, and Mark D Robinson. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, 17(1):1–15, 2016.
- [2] Leonard D Goldstein, Yi Cao, Gregoire Pau, Michael Lawrence, Thomas D Wu, Somasekar Seshagiri, and Robert Gentleman. Prediction and Quantification of Splice Events from RNA-Seq Data. *PLoS ONE*, 11(5):e0156132, may 2016.

Competing Interests: No competing interests were disclosed.

PAPER II

CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets

Malgorzata Nowicka, Carsten Krieg, Lukas M. Weber, Felix J. Hartmann, Silvia Guglietta, Burkhard Becher, Mitch P. Levesque and Mark D. Robinson

Paper submitted to *F1000Research*

CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets

Malgorzata Nowicka^{*1,2}, Carsten Krieg³, Lukas M. Weber^{1,2}, Felix J. Hartmann³, Silvia Guglietta⁴, Burkhard Becher³, Mitch P. Levesque⁵, and Mark D. Robinson^{†1,2}

¹Institute for Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland

²SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland

³Institute of Experimental Immunology, University of Zurich, 8057 Zurich, Switzerland

⁴Department of Experimental Oncology, European Institute of Oncology, Via Adamello 16, I-20139 Milan, Italy

⁵Department of Dermatology, University Hospital Zurich, CH-8091 Zurich, Switzerland

Abstract High dimensional (mass and flow) cytometry (HDCyto) experiments have become a method of choice for high throughput interrogation and characterization of cell populations at high throughput. Here, we present an R-based pipeline for differential analyses of HDCyto data, largely based on Bioconductor packages. We computationally define cell populations using FlowSOM clustering and facilitate an optional but reproducible strategy for manual merging of algorithm-generated clusters. Our workflow offers different analysis paths, including association of cell type abundance with a phenotype, or changes in signaling markers within specific subpopulations or differential analyses of aggregated signals. Importantly, the differential analyses we show are based on regression frameworks where the HDCyto data is the response; thus, we are able to model arbitrary experimental designs, such as those with batch effects, paired designs and so on. In particular, we apply generalized linear mixed models to analyses of cell population abundance or cell-population-specific analyses of signaling markers, allowing overdispersion in cells count or aggregated signal across samples to be appropriately modeled. To support the formal statistical analyses, we encourage exploratory data analysis at every step, including for quality control (e.g., multi-dimensional scaling plots), for reporting of clustering results (dimensionality reduction, heatmaps with dendograms) and for differential analyses (e.g., plots of aggregated signal).

Keywords

CyTOF, flow cytometry, differential analysis

^{*}gosia.nowicka@uzh.ch
[†]mark.robinson@imls.uzh.ch

Introduction

Flow cytometry and the more recently introduced CyTOF (cytometry by time-of-flight mass spectrometry or mass cytometry) are high-throughput technologies that measure protein abundance on the surface or within cells. In flow cytometry, antibodies are labeled with fluorescent dyes and fluorescence intensity is measured using lasers and photodetectors. CyTOF utilizes antibodies tagged with metal isotopes from the lanthanide series, which have favorable chemistry and do not occur in biological systems; abundances per cell are recorded with a time-of-flight mass spectrometer. In either case, fluorescence intensities (flow cytometry) or ion counts (mass cytometry) are assumed to be proportional to the expression level of the antibody-targeted antigens of interest.

Due to the differences in acquisition, further distinct characteristics should be noted. Conventional fluorophore-based flow cytometry is non-destructive and thus can be used to sort cells for further analysis. However, because of the spectral overlap between fluorophores, compensation of the data needs to be performed (Roederer 2001), which also limits the number of parameters that can be measured simultaneously. Thus, standard flow cytometry experiments measure 6-12 parameters with modern systems measuring up to 20 channels (Mahnke and Roederer 2007), while new developments (e.g., BD FACSymphony) promise to increase this capacity towards 50. Moreover, flow cytometry offers the highest throughput with 10s of thousands of cells measured per second at relatively low operating costs per sample.

By using rare metal isotopes in CyTOF, cell autofluorescence can be avoided and spectral overlap is drastically reduced. However, the sensitivity of mass spectrometry results in the measurement of metal impurities and oxide formations, which need to be carefully considered in antibody panel design (e.g., through antibody concentrations and coupling of antibodies to neighboring metals). Leipold et al. recently commented that *minimal spillover does not equal no spillover* (Leipold 2015). Nonetheless, CyTOF offers a high dimension of parameters measured per cell, with current panels using ~40 parameters and the promise of up to 100. Throughput of CyTOF is lower, at the rate of hundreds of cells per second and cells are destroyed during ionization.

The ability of flow cytometry and mass cytometry to analyze individual cells at high-throughput scales has resulted in a wide range of biological and medical applications. For example, these “immunophenotyping” assays are used to detect and quantify cell populations of interest, to uncover new cell populations and to compare abundance of cell populations between different conditions, such as between patient groups (Unen et al. 2016). That is, it can be used as a biomarker discovery tool.

Various methodological approaches aim for biomarker discovery (Saeys, Gassen, and Lambrecht 2016). A common strategy, which we will refer through this workflow as the “classic” approach, is to first identify cell populations of interest by manual gating or automated clustering (Hartmann et al. 2016; Pejoski et al. 2016). Second, using statistical tests, one can determine which of the cell subpopulations or protein markers are associated with a phenotype (e.g., clinical outcome) of interest. Typically, cell subpopulation abundance expressed as cluster cell counts or median marker expression would be used in the statistical model to relate to the sample-level phenotype.

Importantly, there are many alternatives to what we propose below and several new methods are emerging. For example, *Citrus* (Bruggner et al. 2014) tackles the differential discovery problem by strong over-clustering of the cells and building a hierarchy of clusters from very specific to general ones. Using model selection and regularization techniques, clusters and markers that associate with the outcome are identified. A new machine learning approach, *CellCnn* (Arvaniti and Claassen 2016), learns the representation of clusters that are associated with the considered phenotype by the means of convolutional neural networks, which makes it particularly applicable to detecting discriminating rare cell populations. However, there are tradeoffs to consider. *Citrus* performs feature selection but does not provide significance levels, such as p-values, for the strength of associations. Due to its computational requirements, *Citrus* can not be run on entire mass cytometry datasets and one typically must analyze a subset of the data. The “filters” from *CellCnn* may identify one or more cell subsets that distinguish experimental groups, while these groups may not necessarily correspond to any of the canonical cell types, since they are learned with a data-driven approach.

A noticeable distinction between the machine learning approaches and our classical regression approach is how the model is designed. *Citrus* and *CellCnn* model the patient response as a function of the measured HDCyto values, whereas the classical approach models the HDCyto data itself as the response, thus putting the distributional assumptions on the experimental HDCyto data. This carries the distinct advantage that covariates (e.g., age, gender, batch) can be included, together with finding associations of the phenotype to the predictors of interest (e.g., cell type abundance). Specifically, neither *Citrus* nor *CellCnn* are able to directly account for complex designs, such as paired experiments or presence of batches.

Within the classical approach, hybrid methods are certainly possible, where discovery of interesting cell populations is done with one algorithm and quantifications or signal aggregations are modeled in standard regression frameworks. For instance, *CellCnn* provides p-values from a t-test or Mann-Whitney U-test conducted on the frequencies of previously detected cell populations. The models we propose below are flexible extensions of this strategy.

Step by step, this workflow presents differential discovery analyses assembled from a suite of tools and methods that, in our view, leads to a high level of flexibility, robust statistically-supported and interpretable results.

Cell population identification is conducted by means of unsupervised clustering using the *FlowSOM* and *ConsensusClusterPlus* packages, which together were among the best performing clustering approaches for high-dimensional cytometry data (Weber and Robinson 2016). Notably, *FlowSOM* scales easily to millions of cells and thus no subsetting of the data is required.

To be able to analyze arbitrary experimental designs (e.g., batch effects, paired experiments, etc.), we show how to conduct the differential analysis of cell population abundances using the generalized linear mixed models (GLMM) and marker intensities using linear models (LM) and linear mixed models (LMM). Model fitting is performed with *lme4* and *stats* packages and hypothesis testing with the *multcomp* package.

We use the *ggplot2* package as our graphical engine. Notably, we propose a suite of useful visual representations of HDCyto data characteristics, such as an MDS (multidimensional scaling) plot of aggregated signal for exploring sample similarities. The obtained cell populations are visualized using dimension reduction techniques (e.g., t-SNE via the *Rtsne* package) and heatmaps (via the *pheatmap* package) to represent characteristics of the annotated cell populations and identified biomarkers.

The workflow is intentionally not fully automatic. First, we strongly advocate for exploratory data analysis to get an understanding of data characteristics before formal statistical modeling. Second, the workflow involves an optional step where the user can manually merge and annotate clusters, see Section Cluster merging and annotation, but in a way that is easily reproducible. The CyTOF data used here, see Section Data description, is already preprocessed, i.e., the normalization and debarcoding as well as removal of doublets, debris and dead cells were already performed. To see how such an analysis could be performed, go to Section Data preprocessing.

Notably, this workflow is equally applicable to flow or mass cytometry datasets for which the preprocessing steps have already been performed. In addition, the workflow is modular and can be adapted as new algorithms or new knowledge about how to best use existing tools comes to light. In particular, alternative clustering algorithms, such as the popular PhenoGraph algorithm (Levine et al. 2015) (e.g., via the *Rpheno-graph* package) and dimensionality reduction techniques, such as diffusion maps (L. Haghverdi, Buettner, and Theis 2015) via the *destiny* package (Angerer et al. 2016)) and SIMLR (B. Wang et al. 2017) via the *SIMLR* package, could be inserted to the workflow. We point to various alternatives and considerations in the text below.

Data description

We use a subset of CyTOF data originating from Bodenmiller *et al.* (Bodenmiller et al. 2012) that was also used in the *Citrus* paper (Bruggner et al. 2014). Specifically, we perform our analysis on 16 samples of peripheral blood mononuclear cells (PBMCs) from 8 healthy donors where 8 were unstimulated and another 8 were stimulated for 30 min with B cell receptor/Fc receptor crosslinking (BCR/FcR-XL). For each sample, 14 signaling markers and 10 cell surface markers were measured.

The original data can be downloaded from the Cytobank report. The subset used here is also available from the Citrus Cytobank repository.

In both the Bodenmiller *et al.* and *Citrus* manuscripts, the 10 lineage markers were used to identify cell subpopulations, which were then investigated for differences between the reference and stimulated separately for each of the 14 functional markers. The same strategy is used in this workflow; 10 lineage markers are used for cell clustering and 14 functional markers are tested for the differential expression between the reference and BCR/FcR-XL stimulation. Even though differential analysis of cell abundance was not in the scope of the Bodenmiller *et al.* experiment, we present them here to highlight the generality of the discovery.

Data preprocessing

Conventional flow cytometers and mass cytometers produce .fcs files that can be manually analyzed using programs such as FlowJo [TriStar] or Cytobank (Kotecha, Krutzik, and Irish 2001). During this initial analysis step, dead cells are removed, compensation is checked and with simple two dimensional scatter plots (e.g., marker intensity versus time), marker expression patterns are checked. Often, modern experiments are barcoded in order to remove analytical biases due to individual sample variation or acquisition time. Preprocessing steps including normalization using bead standards, debarcoding and compensation can be completed with the *CATALYST* package, which provides an implementation of the debarcoding algorithm described by Zunder *et al.* (Zunder et al. 2015) and the bead-based normalization from Finck *et al.* (Finck et al. 2013). Of course, preprocessing steps can occur using custom scripts within R or outside of R (e.g., Normalizer (Finck et al. 2013)).

Data import

We recommend as standard practice to keep an independent record of all the samples collected, with additional information about the experimental condition, including sample or patient identifiers, processing batch and so on. That is, we recommend having a trail of *metadata* for each experiment. In our example, the metadata file, PBMC8_metadata.xlsx, can be downloaded from the Robinson Lab server with the `download.file` function.

For the workflow, the user should place in the current working directory (`getwd()`). Here, we load it into R with the `read_excel` function from the `readxl` package and save it into a variable called `md`, but other file types and interfaces to read them in are possible.

The data frame `md` contains the following columns:

- `file_name` with names of the .fcs files corresponding to the reference (suffix “Reference”) and BCR/FcR-XL stimulation (suffix “BCR-XL”) samples,
- `sample_id` with shorter unique names for each sample containing information about conditions and patient IDs,
- `condition` describes whether samples originate from the reference (Ref) or stimulated (BCRXL) condition,
- `patient_id` defines the IDs of patients.

The `sample_id` variable is used as row names in metadata and will be used all over the workflow to label the samples. It is important to carefully check whether variables are of the desired type (factor, numeric, character), since input methods may convert columns into different data types. For the statistical modeling, we want to make the condition variable a factor with the reference (Ref) samples being the reference level, where the order of factor levels can be defined with the `levels` parameter of the `factor` function. We also specify colors for the different conditions in a variable `color_conditions`.

```
library(readxl)
url <- "http://imlspenticton.uzh.ch/robinson_lab/cytoWorkflow"
metadata_filename <- "PBMC8_metadata.xlsx"
download.file(file.path(url, metadata_filename), destfile = metadata_filename)
md <- read_excel(metadata_filename)

## Make sure condition variables are factors with the right levels
md$condition <- factor(md$condition, levels = c("Ref", "BCRXL"))
head(data.frame(md))

##          file_name sample_id condition patient_id
## 1 PBMC8_30min_patient1_BCR-XL.fcs    BCRXL1    BCRXL   Patient1
## 2 PBMC8_30min_patient1_Reference.fcs     Ref1      Ref   Patient1
## 3 PBMC8_30min_patient2_BCR-XL.fcs    BCRXL2    BCRXL   Patient2
## 4 PBMC8_30min_patient2_Reference.fcs     Ref2      Ref   Patient2
## 5 PBMC8_30min_patient3_BCR-XL.fcs    BCRXL3    BCRXL   Patient3
## 6 PBMC8_30min_patient3_Reference.fcs     Ref3      Ref   Patient3

## Define colors for conditions
color_conditions <- c("#6A3D9A", "#FF7FOO")
names(color_conditions) <- levels(md$condition)
```

The .fcs files listed in the metadata can be downloaded manually from the Citrus Cytobank repository or automatically from the Robinson Lab server where they are saved in a compressed archive file, `PBMC8_fcs_files.zip`.

```
fcs_filename <- "PBMC8_fcs_files.zip"
download.file(file.path(url, fcs_filename), destfile = fcs_filename)
unzip(fcs_filename)
```

To load the content of the .fcs files into R, we use the `flowCore` package and read in all files into a `flowSet` object, which is a general container for HDCyo data. Importantly, `read.flowSet` and the `read.FCS` functions, by default, may transform the marker intensities and remove cells with extreme positive values. We keep these options off to be sure that we control the exact preprocessing steps.

```
library(flowCore)
fcs_raw <- read.flowSet(md$file_name, transformation = FALSE,
                       truncate_max_range = FALSE)
fcs_raw
```

In our example, information about the panel is also available in a file called `PBMC8_panel.xlsx` and can be downloaded from the Robinson Lab server and loaded into a variable called `panel1`. It contains columns for

`Isotope` and `Metal` that define the atomic mass number and the symbol of the chemical element conjugated to the antibody, respectively and `Antigen`, which specifies the protein marker that was targeted; two additional columns specify whether a channel belongs to the lineage or surface type of marker.

The `isotope`, `metal` and `antigen` information that the instrument receives is also stored in the `flowFrame` (container for one sample) or `flowSet` (container for multiple samples) objects. You can type `fcs_raw[[1]]` to see the first `flowFrame`, which contains a table with columns `name` and `desc`. Their content can be accessed with functions `pData(parameters())`, which is identical for all the `flowFrame` objects in the `flowSet`. The variable `name` corresponds to the column names in the `flowSet` object, you can type in R `colnames(fcs_raw)`.

One should make sure that the elements from `panel` can be matched to their corresponding entries in the `flowSet` object to make the analysis less prone to subsetting mistakes. Here, for example, the entries in `panel$Antigen` have their exact equivalents in the `desc` columns of the `flowFrame` objects. In the following analysis, we will often use marker IDs as column names in the tables containing expression values. As a cautionary note, object conversion from one type to another (e.g., creation of `data.frame` from a matrix), some characters (e.g., dashes) in the dimension names are replaced with dots and this may cause problems in matching. To avoid this problem, we replace all the dashes with underscores. Also, we define two variables that indicate the lineage and functional markers.

```
panel_filename <- "PBMC8_panel.xlsx"
download.file(file.path(url, panel_filename), destfile = panel_filename)
panel <- read_excel(panel_filename)
head(data.frame(panel))
```

	Metal	Isotope	Antigen	Lineage	Functional
## 1	Cd	110:114	CD3	1	0
## 2	In	115	CD45	1	0
## 3	La	139	BC1	0	0
## 4	Pr	141	BC2	0	0
## 5	Nd	142	pNFkB	0	1
## 6	Nd	144	pp38	0	1

```
# Replace problematic characters
panel$Antigen <- gsub("-", "_", panel$Antigen)

panel_fcs <- pData(parameters(fcs_raw[[1]]))
head(panel_fcs)
```

	name	desc	range	minRange	maxRange
## \$P1	Time	Time	2377271	0.00000	2377270
## \$P2	Cell_length	Cell_length	66	0.00000	65
## \$P3	CD3(110:114)Dd	CD3	1212	-13.66756	1211
## \$P4	CD45(In115)Dd	CD45	2654	0.00000	2653
## \$P5	BC1(La139)Dd	BC1	13357	0.00000	13356
## \$P6	BC2(Pr141)Dd	BC2	39	-66.97583	38

```
# Replace problematic characters
panel_fcs$desc <- gsub("-", "_", panel_fcs$desc)

# Lineage markers
(lineage_markers <- panel$Antigen[panel$Lineage == 1])
```

```
## [1] "CD3"      "CD45"     "CD4"      "CD20"     "CD33"     "CD123"    "CD14"
## [8] "IgM"      "HLA_DR"   "CD7"
```

```
# Functional markers
(functional_markers <- panel$Antigen[panel$Functional == 1])
```

```
## [1] "pNFkB"    "pp38"     "pStat5"   "pAkt"     "pStat1"   "pSHP2"    "pZap70"
## [8] "pStat3"   "pSlp76"   "pBtk"     "pPlcg2"   "pErk"     "pLat"     "pS6"
```

```
# Spot checks
all(lineage_markers %in% panel_fcs$desc)
```

```
## [1] TRUE

all(functional_markers %in% panel_fcs$desc)

## [1] TRUE
```

Data transformation

Usually, the raw marker intensities read by a cytometer, have strongly skewed distributions with varying ranges of expression, thus making it difficult to distinguish between the negative and positive cell populations. Therefore, it is common practice to transform CyTOF marker intensities using, for example, arcsineh (hyperbolic inverse sine) with cofactor 5 (Bendall et al. 2011 Figure S2; Bruggner et al. 2014) to make the distributions more symmetric and to map them to a comparable range of expression, which is important for clustering. A cofactor of 150 has been promoted for flow cytometry, but users are free to implement an alternative transformation, some of which are available from the `transform` function of the `flowCore` package. In the following step, we include only those channels that correspond to the lineage and functional markers. We also rename the columns in the `flowSet` to the antigen names from `panel$desc`.

```
## arcsineh transformation and column subsetting
fcs <- fsApply(fcs_raw, function(x, cofactor=5){
  colnames(x) <- panel_fcs$desc
  expr <- exprs(x)
  expr <- asinh(expr[, c(lineage_markers, functional_markers)] / cofactor)
  exprs(x) <- expr
  x
})
fcs

## A flowSet with 16 experiments.
##
##   column names:
##   CD3 CD45 CD4 CD20 CD33 CD123 CD14 IgM HLA_DR CD7 pNFKB pp38 pStat5 pAkt pStat1 pSHP2 pZap70 pSt
```

For some of the further analysis, it is more convenient for us to work using a matrix (called `expr`) that contains marker expression for cells from all samples. We create such a matrix with the `fsApply` function that extracts the expression matrices (function `exprs`) from each element of the `flowSet` object.

```
## Extract expression
expr <- fsApply(fcs, exprs)
dim(expr)

## [1] 172791      24
```

As the ranges of marker intensities can vary substantially, we apply another transformation that scales expression of all markers to values between 0 and 1 using low (e.g., 1%) and high (e.g., 99%) percentiles as the boundary. This additional transformation of the `asinh`-transformed data can sometimes give better representation of relative differences in marker expression between annotated cell populations, however, it is only used here for visualization.

```
library(matrixStats)
rng <- colQuantiles(expr, probs = c(0.01, 0.99))
expr01 <- t((t(expr) - rng[, 1]) / (rng[, 2] - rng[, 1]))
expr01[expr01 < 0] <- 0
expr01[expr01 > 1] <- 1
```

Diagnostic plots

We propose some quick checks to verify whether the data we analyze globally represents what we expect, such as whether samples that are replicates of one condition are more similar and are distinct from samples from another condition. In addition, another important check is to verify that marker expression distributions do not have any abnormalities, such as, having different ranges or distinct distributions for a subset of the samples. This could highlight problems with the sample collection or HDCyto acquisition or batch effects that were unexpected. Depending on the situation, one can then consider removing problematic markers or

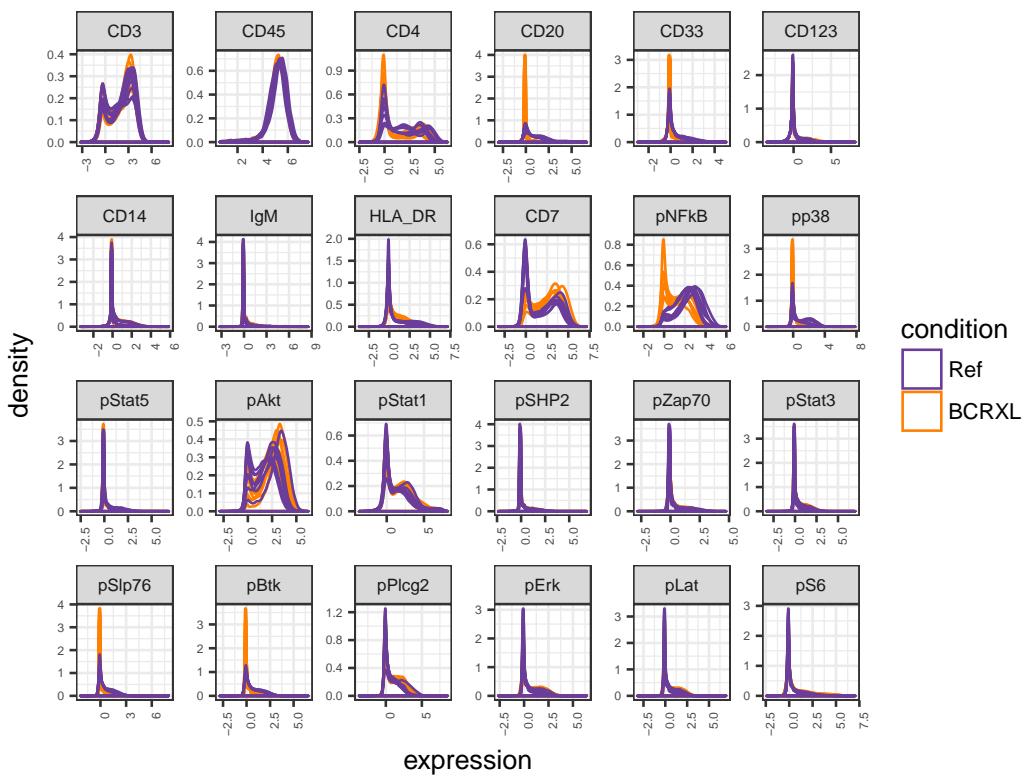


Figure 1. As a diagnostic, per-sample marker expression distributions are shown. Samples are colored according to experimental condition.

samples from further analysis; in the case of batch effects, a covariate column could be added to the metadata table and used below in the statistical analyses.

The step below generates a plot with per-sample marker expression distributions, colored by condition. Here, we can already see distinguishing markers, such as pNFkB and CD20, between stimulated and unstimulated conditions.

```
## Generate sample IDs corresponding to each cell in the 'expr' matrix
sample_ids <- rep(md$sample_id, fsApply(fcs_raw, nrow))

library(ggplot2)
library(reshape2)

ggdf <- data.frame(sample_id = sample_ids, expr)
ggdf <- melt(ggdf, id.var = "sample_id",
            value.name = "expression", variable.name = "antigen")
mm <- match(ggdf$sample_id, md$sample_id)
ggdf$condition <- md$condition[mm]

ggplot(ggdf, aes(x = expression, color = condition,
                 group = sample_id)) +
  geom_density() +
  facet_wrap(~ antigen, nrow = 4, scales = "free") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        strip.text = element_text(size = 7), axis.text = element_text(size = 5)) +
  guides(color = guide_legend(ncol = 1)) +
  scale_color_manual(values = color_conditions)
```

Another spot check is the number of cells per sample. This plot can be used as a guide together with other readouts (see below) to identify samples for which not enough cells were assayed.

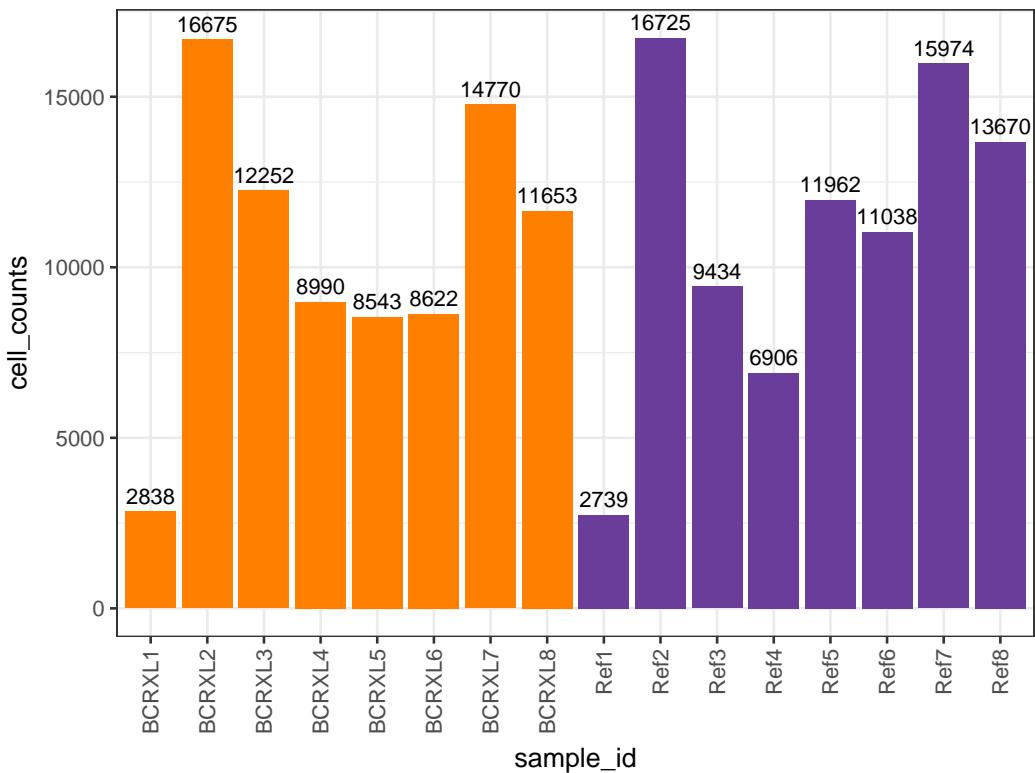


Figure 2. Barplot with the number of cells per sample.

```
cell_table <- table(sample_ids)

ggdf <- data.frame(sample_id = names(cell_table),
  cell_counts = as.numeric(cell_table))
mm <- match(ggdf$sample_id, md$sample_id)
ggdf$condition <- md$condition[mm]

ggplot(ggdf, aes(x = sample_id, y = cell_counts, fill = condition)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = cell_counts), hjust=0.5, vjust=-0.5, size = 3) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
    legend.position = "none") +
  scale_fill_manual(values = color_conditions, drop = FALSE) +
  scale_x_discrete(drop = FALSE)
```

MDS plot

In transcriptomics applications, one of the most utilized exploratory plots is the multi-dimensional scaling (MDS) plot or a principal component analysis (PCA) plot. Such plots show similarities between samples measured in an unsupervised way and give a sense of how much differential expression can be detected before conducting any formal tests. An MDS plot can be generated with the `plotMDS` function from the `limma` package. In transcriptomics, distances between samples are calculated based on the expression of the top varying genes. We propose a similar plot for HDCyto data using median marker expression over all cells to calculate dissimilarities between samples (other aggregations are also possible and one could reduce the number of top varying markers to include in the calculation). Ideally, samples should cluster well within the same condition, although this depends on the magnitude of the difference between experimental conditions. With this diagnostic, one can identify the outlier samples and eliminate them if the circumstances warrant it.

In our MDS plot on median marker expression values, we can see that the first dimension (MDS1) separates the unstimulated and stimulated samples reasonably well. The second dimension (MDS2) represents, to some degree, differences between patients. Most of the samples that originate from the same patient are placed at a similar point along the y-axis, for example, samples from patients 7, 5, and 8 are at the top of the plot,

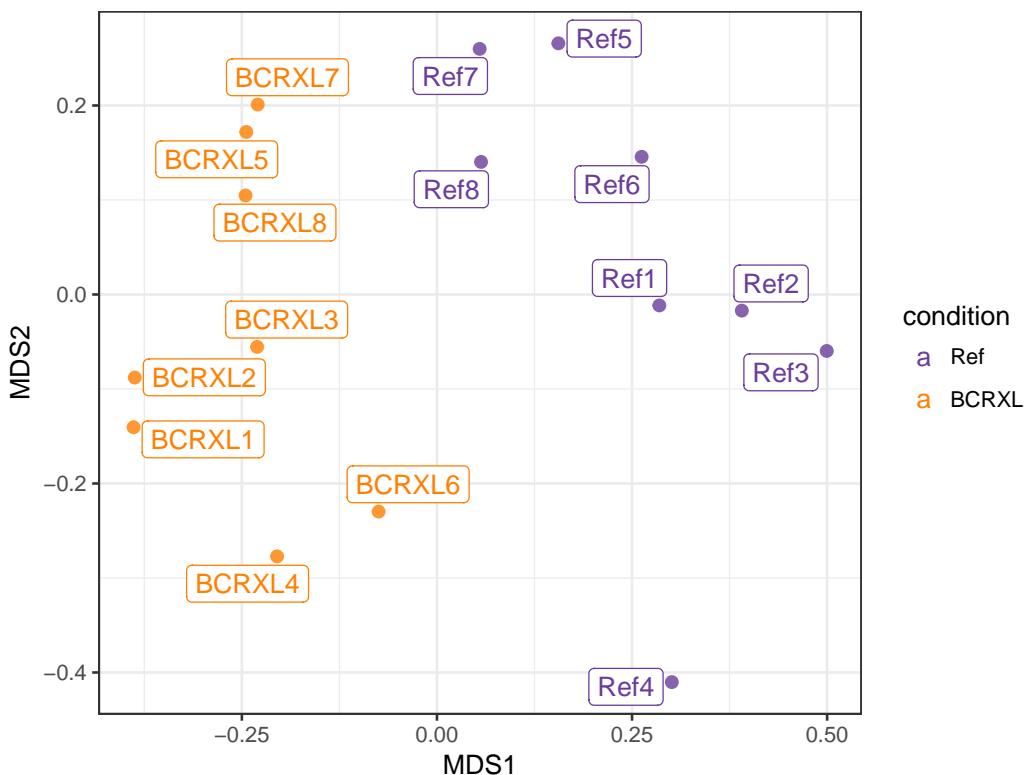


Figure 3. MDS plot based on the median marker expression.

samples from patient 4 are located at the bottom of the plot. This also indicates that the marker expression of individual patients is driving similarity and perhaps should be formally accounted for in the downstream statistical modeling.

```
# Get the median marker expression per sample
library(dplyr)

expr_median_sample_tbl <- data.frame(sample_id = sample_ids, expr) %>%
  group_by(sample_id) %>%
  summarize_each(funs(median))

expr_median_sample <- t(expr_median_sample_tbl[, -1])
colnames(expr_median_sample) <- expr_median_sample_tbl$sample_id

library(limma)
mds <- plotMDS(expr_median_sample, plot = FALSE)

library(ggrepel)
ggdf <- data.frame(MDS1 = mds$x, MDS2 = mds$y,
  sample_id = colnames(expr_median_sample))
mm <- match(ggdf$sample_id, md$sample_id)
ggdf$condition <- md$condition[mm]

ggplot(ggdf, aes(x = MDS1, y = MDS2, color = condition)) +
  geom_point(size = 2, alpha = 0.8) +
  geom_label_repel(aes(label = sample_id)) +
  theme_bw() +
  scale_color_manual(values = color_conditions)
```

In contrast to genomic applications, the number of variables measured for each sample is much lower in HDCyo data. In the former, thousands of genes are surveyed, where in the latter, ~20-50 antigens are typically targeted. Similar to the MDS plot above, a heatmap of the same data also gives insight into the structure of the data. The heatmap shows median marker intensities with clustered columns (samples) and rows (markers).

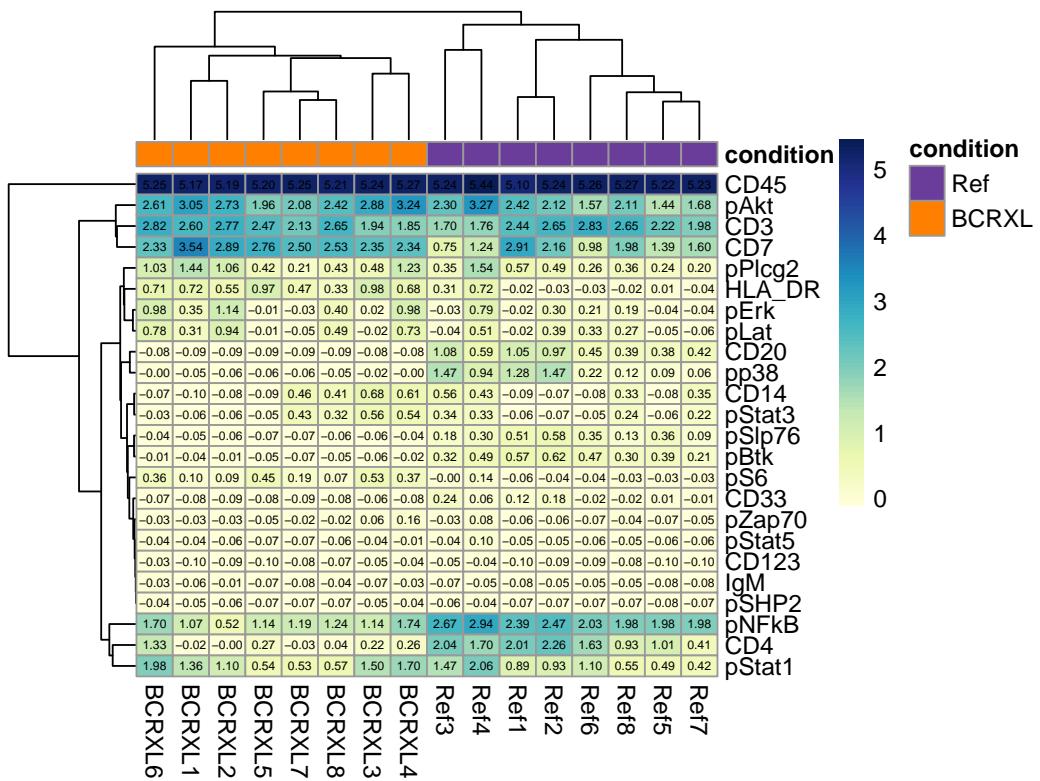


Figure 4. Clustering of samples and markers based on the median marker expression.

We have used hierarchical clustering with average linkage and euclidean distance, but also Ward's linkage could be used (Bruggner et al. 2014), and in CyTOF applications, a cosine distance shows good performance (Bendall et al. 2014). In this plot, we can see which markers drive the observed clustering of samples.

As with the MDS plot, the dendrogram separates the reference and stimulated samples very well. Also, similar groupings of patients within experimental conditions are observed (patients 1-2, 5-7-8 and 3-4 are together in both conditions).

```
library(RColorBrewer)
library(pheatmap)

# Column annotation for the heatmap
mm <- match(colnames(expr_median_sample), md$sample_id)
annotation_col <- data.frame(condition = md$condition[mm],
  row.names = colnames(expr_median_sample))
annotation_colors <- list(condition = color_conditions)

# Colors for the heatmap
color <- colorRampPalette(brewer.pal(n = 9, name = "YlGnBu"))(100)

pheatmap(expr_median_sample, color = color, display_numbers = TRUE,
  number_color = "black", fontsize_number = 5, annotation_col = annotation_col,
  annotation_colors = annotation_colors, clustering_method = "average")
```

Marker ranking based on the non-redundancy score

In this step, we identify the ability of markers to explain the variance observed in each sample. In particular, we calculate the PCA-based non-redundancy score (NRS) from Levine *et al.* (Levine et al. 2015). Markers with higher score explain a larger portion of variability present in a given sample.

The average NRS can be used to select a subset of markers that are non-redundant in each sample but at the same time capture the overall diversity between samples. Such a subset of markers can be then used for cell population identification analysis (i.e., clustering). We note that there is no precise rule on how to choose the right cutoff for marker inclusion, but one of the approaches is to select a suitable number of the

top-scoring markers. The number can be chosen by analyzing the plot with the NR scores, shown below, where the markers are sorted by the decreasing average NRS. One can drop out markers that are not likely to distinguish cell populations of interest, even if they have high scores and add in markers with low scores but known to be important in discerning cell subgroups (Levine et al. 2015).

In the dataset considered here, we want to use all the 10 lineage markers, so there is no explicit need to restrict the set of cell surface markers. However, there may be other situations where this feature selection step would be of interest.

```
## Define a function that calculates the NRS per sample
NRS <- function(x, ncomp = 3){
  pr <- prcomp(x, center = TRUE, scale. = FALSE)
  score <- rowSums(outer(rep(1, ncol(x)),
    pr$sdev[1:ncomp]^2) * abs(pr$rotation[,1:ncomp]))
  return(score)
}

## Calculate the score
nrs_sample <- fsApply(fcs[, lineage_markers], NRS, use.exprs = TRUE)
rownames(nrs_sample) <- md$sample_id
nrs <- colMeans(nrs_sample, na.rm = TRUE)

## Plot the NRS for ordered markers
lineage_markers_ord <- names(sort(nrs, decreasing = TRUE))
nrs_sample <- data.frame(nrs_sample)
nrs_sample$sample_id <- rownames(nrs_sample)

ggdf <- melt(nrs_sample, id.var = "sample_id",
  value.name = "nrs", variable.name = "antigen")

ggdf$antigen <- factor(ggdf$antigen, levels = lineage_markers_ord)
mm <- match(ggdf$sample_id, md$sample_id)
ggdf$condition <- md$condition[mm]

ggplot(ggdf, aes(x = antigen, y = nrs)) +
  geom_point(aes(color = condition), alpha = 0.9,
    position = position_jitter(width = 0.3, height = 0)) +
  geom_boxplot(outlier.color = NA, fill = NA) +
  stat_summary(fun.y = "mean", geom = "point", shape = 21, fill = "white") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  scale_color_manual(values = color_conditions)
```

Cell population identification with FlowSOM and ConsensusClusterPlus

Cell population identification typically has been carried out by manual gating, a method based on visual inspection of a series of two-dimensional scatterplots. At each step, a subset of cells, either positive or negative for the two visualized markers, is selected and further stratified in the subsequent iterations until populations of interest across a range of marker combinations are captured. However, manual gating has drawbacks, such as subjectivity, bias toward well-known cell types and inefficiency when analyzing large datasets, which also contribute to a lack of reproducibility (Saeys, Gassen, and Lambrecht 2016).

Considerable effort has been made to improve and automate cell population identification, such as unsupervised clustering (Aghaeepour et al. 2013). However, not all methods scale well in terms of performance and speed from the lower dimensionality flow cytometry data to the higher dimensionality mass cytometry data (Weber and Robinson 2016), since clustering in higher dimensions can suffer the “curse of dimensionality”.

Beside the mathematical and algorithmic challenges of clustering, cell population identification may be difficult due to the chemical and biological aspects of the cytometry experiment itself. Therefore, caution should be taken when designing panels aimed at detecting rare cell populations by assigning higher sensitivity metals to rare markers. The right choice of a marker panel used for clustering can also be important. It should include all markers that are relevant for cell type identification.

In this workflow, we conduct cell clustering with *FlowSOM* (Van Gassen et al. 2015) and *ConsensusClusterPlus* (Wilkerson and Hayes 2010), which appeared amongst the fastest and best performing clustering approaches in a recent study of HDCyto datasets (Weber and Robinson 2016). This ensemble showed strong performance in detecting both high and low frequency cell populations and is one of the fastest methods to run, which enables its interactive usage. We use a slight modification of the original workflow presented in the *FlowSOM*

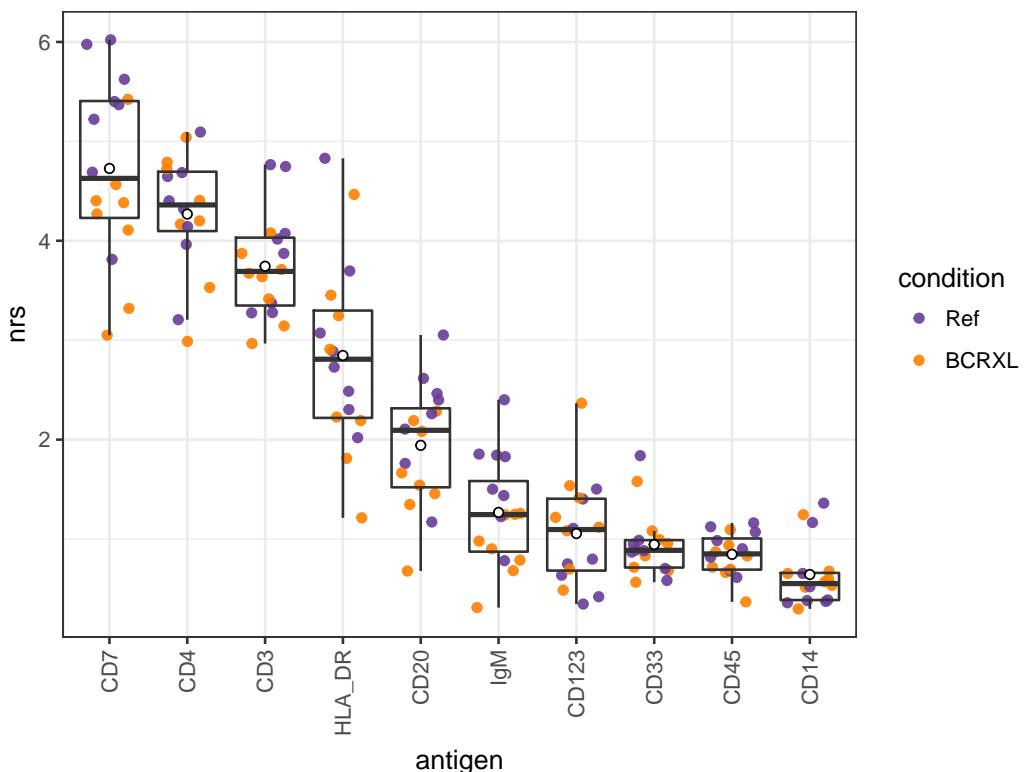


Figure 5. Non-redundancy scores. The colored points represent the per-sample NR scores, while open white circles indicate the mean NR scores from all the samples. Markers on the x-axis are sorted according to the decreasing average NRS.

vignette, which we find more flexible. In particular, we directly call the `ConsensusClusterPlus` function that is embedded in `metaClustering_consensus`. Thus, we are able to access all the functionality of the `ConsensusClusterPlus` package to identify the number of clusters.

The `FlowSOM` workflow consists of three main steps. First, a self-organizing map (SOM) is built using the `BuildSOM` function, where cells are assigned according to their similarities to 100 (by default) grid points (or, so-called codebook vectors or codes) of the SOM. The building of a minimal spanning tree, which is mainly used for graphical representation of the clusters, is skipped in this pipeline. And finally, *metaclustering* of the SOM codes, is performed directly with the `ConsensusClusterPlus` function. Additionally, we add an optional round of manual expert-based merging of the metaclusters and allow this to be done in a reproducible fashion.

`FlowSOM` output can be sensitive to random starts (Weber and Robinson 2016). To make results reproducible, one must specify the seed for the random number generation in R using function `set.seed`. It is also advisable to rerun analyses with multiple random seeds for two reasons. First, one can see how robust the detected clusters are, and second, when the goal is to find smaller cell populations, it may happen that, in some runs, random starting points do not represent rare cell populations as a chance of selecting starting cells from them is low and they are merged into a larger cluster.

It is important to point out that we cluster all cells from all samples together. This strategy is beneficial, since we label cell populations only once and the mapping of cell types between samples is automatically consistent. In our analysis, cell populations are identified using only the 10 lineage markers as defined in the `BuildSOM` function with the `colsToUse` argument.

```
library(FlowSOM)

fsom <- ReadInput(fcs, transform = FALSE, scale = FALSE)
set.seed(1234)
som <- BuildSOM(fsom, colsToUse = lineage_markers)
```

Automatic approaches for selecting the number of clusters in HDcyt data do not always succeed (Weber and Robinson 2016). In general, we therefore recommend some level of over-clustering, and if desired, manual merging of clusters. Such a hierarchical approach is especially suited when the goal is to detect smaller cell populations.

The SPADE analysis performed by Bodenmiller et al. (Bodenmiller et al. 2012) identified 6 main cell types (T-cells, monocytes, dendritic cells, B-cells, NK cells and surface- cells) that were further stratified into 14 more specific subpopulations (CD4+ T-cells, CD8+ T-cells, CD14+ HLA-DR high monocytes, CD14+ HLA-DR med monocytes, CD14+ HLA-DR low monocytes, CD14- HLA-DR high monocytes, CD14- HLA-DR med monocytes, CD14- HLA-DR low monocytes, dendritic cells, IgM+ B-cells, IgM- B-cells, NK cells, surface- CD14+ cells and surface- CD14- cells). In our analysis, we are interested in identifying the 6 main PBMC populations including: CD4+ T-cells, CD8+ T-cells, monocytes, dendritic cells, NK cells and B-cells. Following the concept of over-clustering we perform the metaclustering of the (by default) 100 SOM codes into more than expected number of groups. For example, stratification into 20 groups should give enough resolution. We can explore the clustering with a wide variety of visualizations: t-SNE plots, heatmaps and a plot generated by **ConsensusClusterPlus** called “delta area”.

We call **ConsensusClusterPlus** with maximum number of clusters `maxK = 20` and other clustering parameters set to the values as in the `metaClustering_consensus` function. Again, to ensure that the analyses are reproducible, we define the random seed.

```
## Metaclustering into 20 clusters with ConsensusClusterPlus
library(ConsensusClusterPlus)

codes <- som$map$codes
plot_outdir <- "consensus_plots"
nmc <- 20

mc <- ConsensusClusterPlus(t(codes), maxK = nmc, reps = 100,
  pItem = 0.9, pFeature = 1, title = plot_outdir, plot = "png",
  clusterAlg = "hc", innerLinkage = "average", finalLinkage = "average",
  distance = "euclidean", seed = 1234)

## Get cluster ids for each cell
code_clustering1 <- mc[[nmc]]$consensusClass
cell_clustering1 <- code_clustering1[som$map$mapping[, 1]]
```

We can then investigate characteristics of identified clusters with heatmaps that illustrate median marker expression in each cluster. As the range of marker expression can vary substantially from marker to marker, we use the 0-1 transformed data for some visualizations. However, to stay consistent with *FlowSOM* and *ConsensusClusterPlus*, we use the (asinh-transformed) unscaled data to generate the dendrogram of the hierarchical structure of metaclusters.

Instead of using only medians, which do not give a full representation of cluster specifics, one can plot the entire marker expression distribution in each cluster. Such a plot gives more detailed profile of each cluster, but represents an increase in the amount of information to interpret. Heatmaps give the overall overview of clusters, are quicker and easier to interpret and together with the dendrogram can be a good basis for further cluster merging, see Section Cluster merging and annotation.

Since we will use the heatmap and density plots again later in this workflow, in code chunks below, we create wrapper functions that generate these two types of plots.

```
color_clusters <- c("#DC050C", "#FB8072", "#1965B0", "#7BAFDE", "#882E72",
  "#B17BA6", "#FF7F00", "#FDB462", "#E7298A", "#E78AC3",
  "#33A02C", "#B2DF8A", "#55A1B1", "#8DD3C7", "#A6761D",
  "#E6AB02", "#7570B3", "#BEAED4", "#666666", "#999999",
  "#aa8282", "#d4b7b7", "#8600bf", "#ba5ce3", "#808000",
  "#aeae5c", "#1e90ff", "#00bfff", "#56ff0d", "#ffff00")

plot_clustering_heatmap_wrapper <- function(expr, expr01,
  cell_clustering, color_clusters, cluster_merging = NULL){

  # Calculate the median expression
  expr_median <- data.frame(expr, cell_clustering = cell_clustering) %>%
    group_by(cell_clustering) %>%
    summarize_each(funs(median))
  expr01_median <- data.frame(expr01, cell_clustering = cell_clustering) %>%
    group_by(cell_clustering) %>%
    summarize_each(funs(median))

  # Calculate cluster frequencies
  clustering_table <- as.numeric(table(cell_clustering))
```

```

# This clustering is based on the markers that were used for the main clustering
d <- dist(expr_median[, colnames(expr)], method = "euclidean")
cluster_rows <- hclust(d, method = "average")

expr_heat <- as.matrix(expr01_median[, colnames(expr01)])
rownames(expr_heat) <- expr01_median$cell_clustering

labels_row <- paste0(rownames(expr_heat), " (",
  round(clustering_table / sum(clustering_table) * 100, 1), "%)")

labels_col <- colnames(expr_heat)

# Row annotation for the heatmap
annotation_row <- data.frame(cluster = factor(expr01_median$cell_clustering))
rownames(annotation_row) <- rownames(expr_heat)

color_clusters <- color_clusters[1:nlevels(annotation_row$cluster)]
names(color_clusters) <- levels(annotation_row$cluster)
annotation_colors <- list(cluster = color_clusters)
annotation_legend <- FALSE

if(!is.null(cluster_merging)){
  cluster_merging$new_cluster <- factor(cluster_merging$new_cluster)
  annotation_row$cluster_merging <- cluster_merging$new_cluster
  color_clusters <- color_clusters[1:nlevels(cluster_merging$new_cluster)]
  names(color_clusters) <- levels(cluster_merging$new_cluster)
  annotation_colors$cluster_merging <- color_clusters
  annotation_legend <- TRUE
}

# Colors for the heatmap
color <- colorRampPalette(rev(brewer.pal(n = 9, name = "RdYlBu")))(100)

pheatmap(expr_heat, color = color,
  cluster_cols = FALSE, cluster_rows = cluster_rows,
  labels_col = labels_col, labels_row = labels_row,
  display_numbers = TRUE, number_color = "black",
  fontsize = 8, fontsize_number = 4,
  annotation_row = annotation_row, annotation_colors = annotation_colors,
  annotation_legend = annotation_legend)

}

plot_clustering_heatmap_wrapper(expr = expr[, lineage_markers_ord],
  expr01 = expr01[, lineage_markers_ord],
  cell_clustering = cell_clustering1, color_clusters = color_clusters)

```

```

plot_clustering_distr_wrapper <- function(expr, cell_clustering){

  cell_clustering <- factor(cell_clustering)
  expr_median <- data.frame(expr, cell_clustering = cell_clustering) %>%
    group_by(cell_clustering) %>%
    summarize_each(funs(median))

  d <- dist(expr_median[, colnames(expr)], method = "euclidean")
  cluster_rows <- hclust(d, method = "average")
  cell_clustering <- factor(cell_clustering,
    levels = levels(cell_clustering)[cluster_rows$order])

  freq_clust <- table(cell_clustering)
  freq_clust <- round(as.numeric(freq_clust)/sum(freq_clust)*100, 1)
  cell_clustering <- factor(cell_clustering,
    labels = paste0(levels(cell_clustering), " \n(", freq_clust, "%)"))
}

```

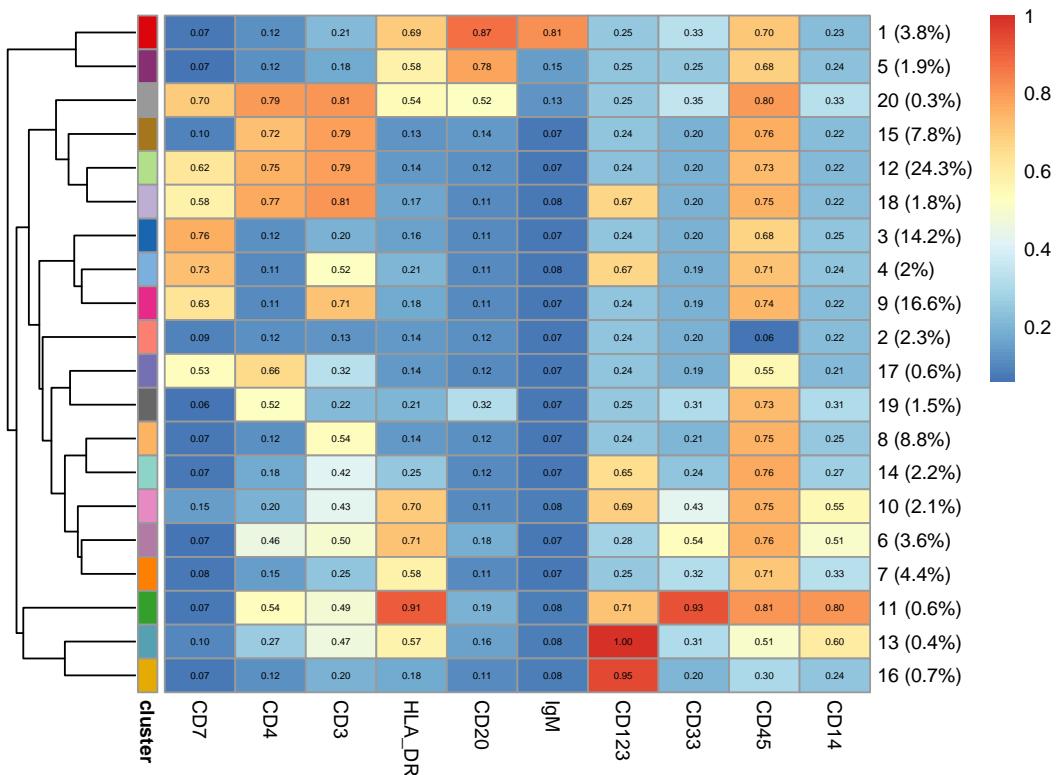


Figure 6. Heatmap of the median marker intensities in 20 cell populations obtained from the metaclustering step. The dendrogram on the left represents the hierarchical similarity between the metaclusters.

```

ggd <- melt(data.frame(cluster = cell_clustering, expr),
  id.vars = "cluster", value.name = "expression",
  variable.name = "antigen")
ggd$antigen <- factor(ggd$antigen, levels = colnames(expr))

ggplot(data = ggd, aes(x = expression, y = ..scaled..)) +
  geom_density(data = transform(ggd, cluster = NULL),
    color = "darkgrey", fill = "black", adjust = 1, alpha = 0.3) +
  geom_density(color = "blue", fill = "blue", adjust = 1, alpha = 0.3) +
  facet_grid(cluster ~ antigen, scales = "free") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
    axis.text = element_text(size = 5),
    strip.text = element_text(size = 5))

}

plot_clustering_distr_wrapper(expr = expr[, lineage_markers_ord],
  cell_clustering = cell_clustering1)

```

Visual representation with t-SNE

One of the most popular plots to represent single cell data are t-SNE plots, where each cell is represented in a lower, usually two-dimensional, space computing using t-stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton 2008). More generally, dimensionality reduction techniques represent the similarity of points in 2 or 3 dimensions, such that similar objects in high dimensional space are also similar in lower dimensional space. Mathematically, there are a myriad of ways to define this similarity. For example, principal components analysis (PCA) uses linear combinations of the original features to find orthogonal dimensions that show the highest levels of variability; the top 2 or 3 principal components can then be visualized.

Nevertheless, there are few notes of caution when using tSNE or any other dimensionality reduction technique. Since they are based on preserving similarities between cells, those that are similar in the original space will

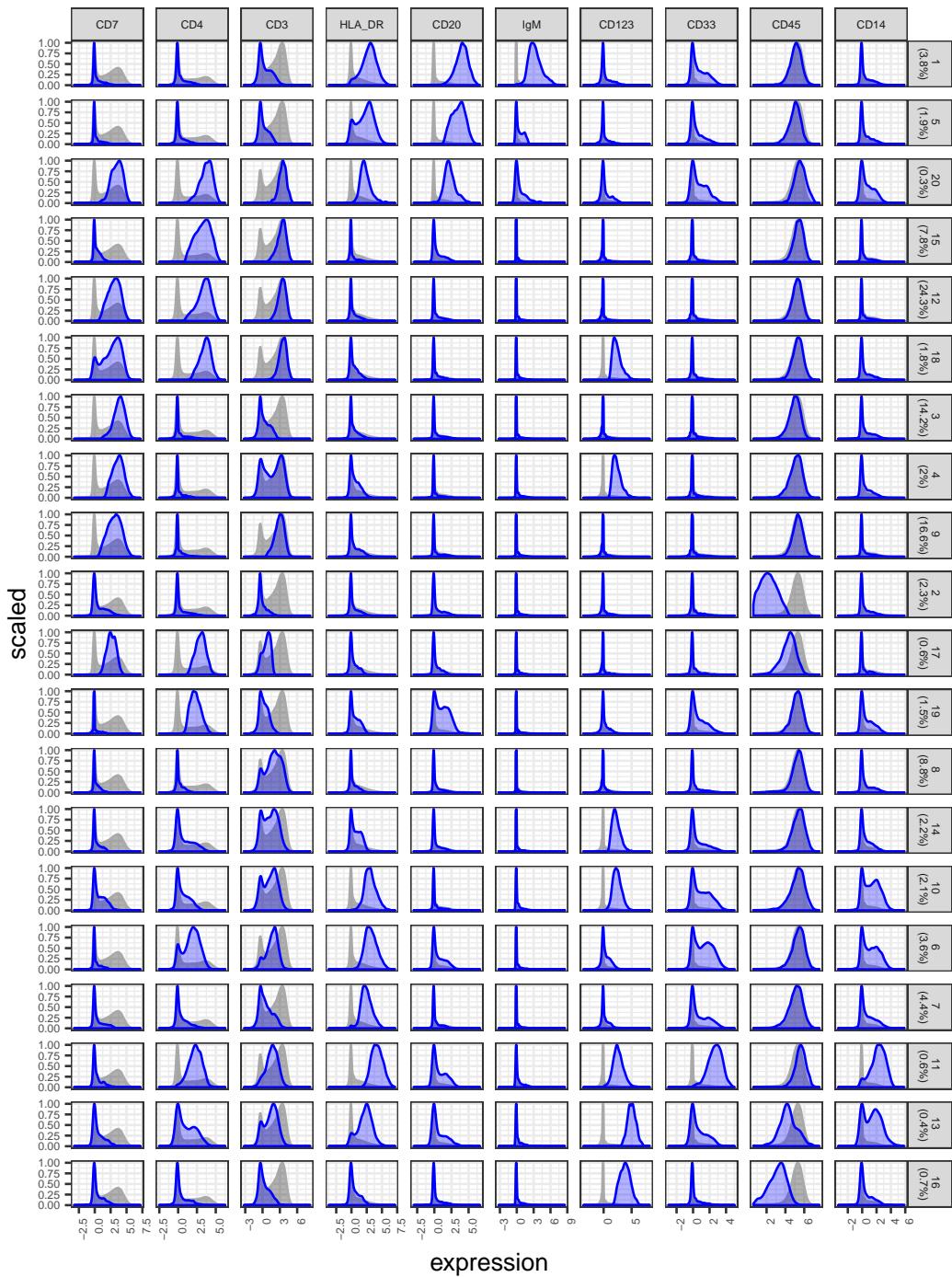


Figure 7. Distributions of marker intensities in 20 cell populations obtained from the metaclustering step. Blue densities represent marker expression for cells in given clusters. Grey densities are calculated from all the cells and serve as a reference.

be close in the 2D/3D representation, but the opposite does not always hold. In our experience, t-SNE with default parameters for HCYTO data is often suitable; for more guidance on the specifics of t-SNE, see How to Use t-SNE Effectively (Wattenberg, Viégas, and Johnson 2016). Due to the stochastic nature of t-SNE optimization, rerunning the method will result in different lower dimensional projections, thus it is advisable to run it a few times to identify the common trends and get a feeling about the variability of the results. As with other methods, to be sure that the analysis is reproducible, the user can define the random seed.

t-SNE is a method that requires significant computational time to process the data even for tens of thousands of cells. CyTOF datasets are usually much larger and thus to keep running times reasonable, one can use a subset of cells; for example, we use here 2000 cells from each sample. The t-SNE map below is colored according to the expression level of the CD4 marker, highlighting that the CD4+ T-cells are placed to the left side of the plot. In this way, one can use a collection of markers to highlight where cell types of interest are located on the *map*.

Instead of t-SNE, one could also use other dimension reduction techniques, such as PCA, diffusion maps, SIMLR (B. Wang et al. 2017) or isomaps, some of which are conveniently available via the `cytobio_dimReduction` function from the `cytobio` package (H. Chen et al. 2016). To speed up the t-SNE analysis, one could use a multicore version that is available via the `Rtsne.multicore` package. Alternative algorithms, such as `largeVis` [REF] (available via the `largeVis` package), can be used for dimensionality reduction of very large datasets without downsampling. Alternatively, the dimensionality reduction can be performed on the *codes* of the SOM, at a resolution specified by the user (see below).

```
## Find and skip duplicates
dups <- which(!duplicated(expr[, lineage_markers]))

## Data subsampling: create indices by sample
inds <- split(1:length(sample_ids), sample_ids)

## How many cells to downsample per-sample
tsne_ncells <- pmin(table(sample_ids), 2000)

## Get subsampled indices
set.seed(1234)
tsne_inds <- lapply(names(inds), function(i){
  s <- sample(inds[[i]], tsne_ncells[i], replace = FALSE)
  intersect(s, dups)
})

tsne_inds <- unlist(tsne_inds)

tsne_expr <- expr[tsne_inds, lineage_markers]

## Run t-SNE
library(Rtsne)

set.seed(1234)
tsne_out <- Rtsne(tsne_expr, check_duplicates = FALSE, pca = FALSE)

## Plot t-SNE colored by CD4 expression
dr <- data.frame(tSNE1 = tsne_out$Y[, 1], tSNE2 = tsne_out$Y[, 2],
  expr[tsne_inds, lineage_markers])

ggplot(dr, aes(x = tSNE1, y = tSNE2, color = CD4)) +
  geom_point() +
  theme_bw() +
  scale_color_gradientn("CD4",
    colours = colorRampPalette(rev(brewer.pal(n = 11, name = "Spectral")))(50))
```

We can color the cells by cluster and ideally, cells of the same color should be close to each other. When the figure is stratified by sample, we can verify whether similar cell populations are present in replicates and if differences in cell abundance are strong, identify distinguishing clusters.

```
dr$sample_id <- sample_ids[tsne_inds]
dr$cell_clustering1 <- factor(cell_clustering1[tsne_inds], levels = 1:nmc)

## Plot t-SNE colored by clusters
```

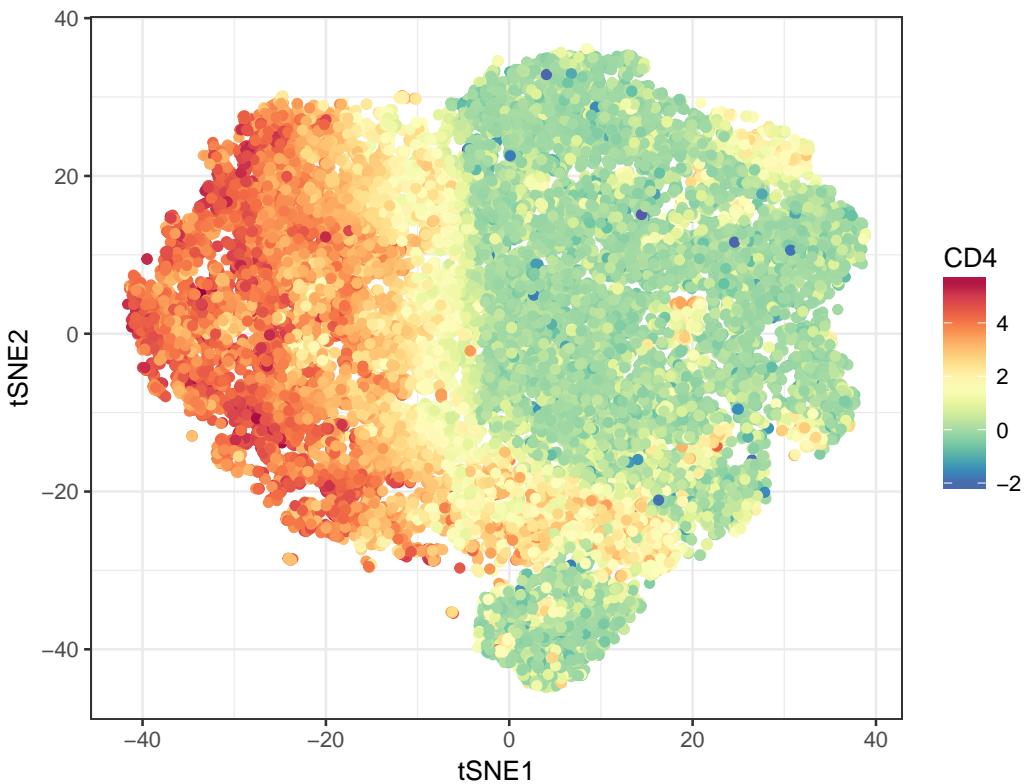


Figure 8. t-SNE plot with cells colored according to the expression level of the CD4 marker.

```

ggp <- ggplot(dr, aes(x = tSNE1, y = tSNE2, color = cell_clustering1)) +
  geom_point() +
  theme_bw() +
  scale_color_manual(values = color_clusters) +
  guides(color = guide_legend(override.aes = list(size = 4), ncol = 2))
ggp

```

```

## Facet per sample
ggp + facet_wrap(~ sample_id)

```

The SOM codes represent characteristics of the 100 (by default) clusters generated in the first step of the *FlowSOM* pipeline. Their visualization can also be helpful in understanding the cell population structure and determining the number of clusters. Ultimately, the metaclustering step uses the codes and not the original cells. We treat the codes as new representative cells and apply the t-SNE dimension reduction to visualize them in 2D. The size of the points corresponds to the number of cells that were assigned to a given code. The points are colored according to the results of metaclustering. Since we have only 100 data points, the t-SNE analysis is fast.

As there are multiple ways to mathematically define similarity in high dimension space, it is always good practice to visualize projections from other methods to see how consistent are the observed patterns. For example, we represent the *FlowSOM* codes also via the first two principal components.

```

## Get code sizes; sometimes not all the codes have mapped cells so they will have size 0
code_sizes <- table(factor(som$map$mapping[, 1], levels = 1:nrow(codes)))
code_sizes <- as.numeric(code_sizes)

```

```

## Run t-SNE on codes
set.seed(1234)
tsne_out <- Rtsne(codes, pca = FALSE)
## Run PCA on codes
pca_out <- prcomp(codes, center = TRUE, scale. = FALSE)

```

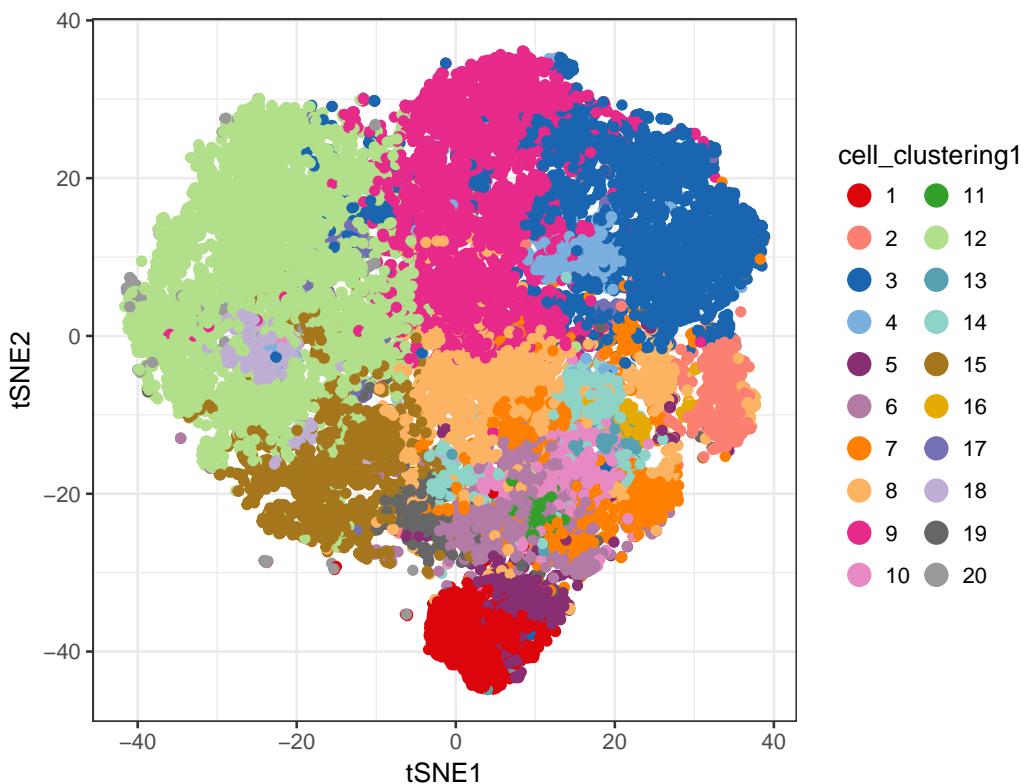


Figure 9. t-SNE plot with cells colored according to the 20 metaclusters.

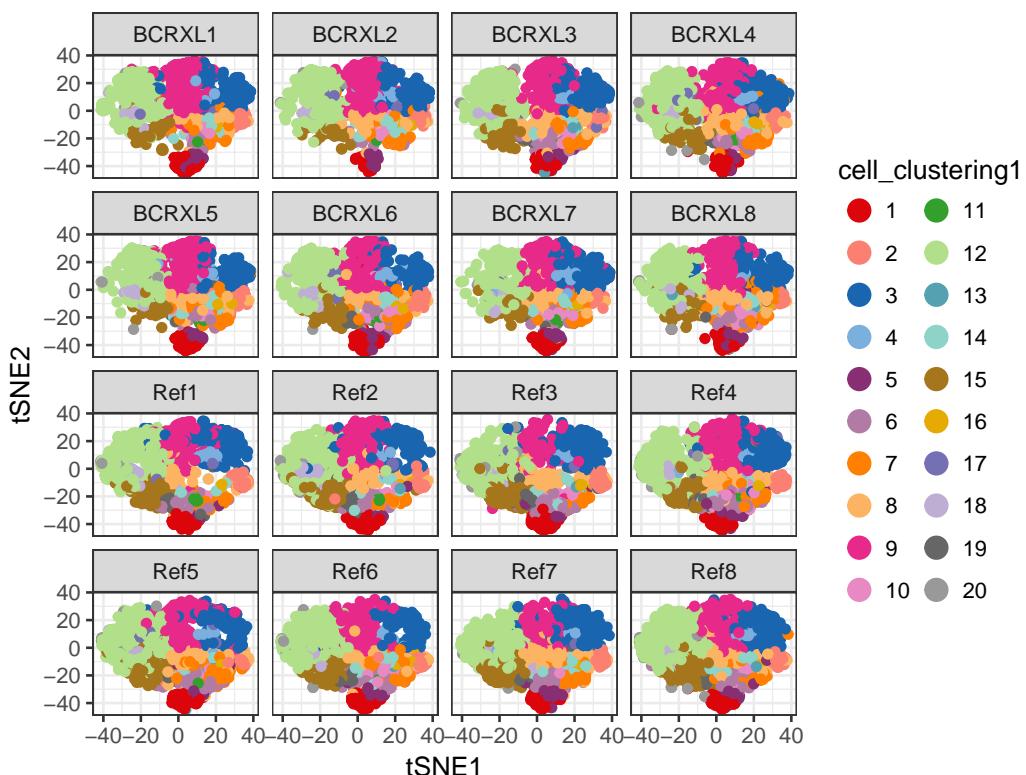


Figure 10. t-SNE plot stratified by sample.

```

codes_dr <- data.frame(tSNE1 = tsne_out$Y[, 1], tSNE2 = tsne_out$Y[, 2],
  PCA1 = pca_out$x[, 1], PCA2 = pca_out$x[, 2])
codes_dr$code_clustering1 <- factor(code_clustering1)
codes_dr$size <- code_sizes

## Plot t-SNE on codes
gg_tsne_codes <- ggplot(codes_dr, aes(x = tSNE1, y = tSNE2,
  color = code_clustering1, size = size)) +
  geom_point(alpha = 0.9) +
  theme_bw() +
  scale_color_manual(values = color_clusters, guide=FALSE) +
  theme(legend.position="bottom")

## Plot PCA on codes
gg_pca_codes <- ggplot(codes_dr, aes(x = PCA1, y = PCA2,
  color = code_clustering1, size = size)) +
  geom_point(alpha = 0.9) +
  theme_bw() +
  scale_color_manual(values = color_clusters) +
  guides(color = guide_legend(override.aes = list(size = 3),
    ncol = 10)) +
  scale_size(guide=FALSE) +
  theme(legend.position="top")

library(cowplot)
plot_grid(gg_tsne_codes, gg_pca_codes, ncol = 1, labels = c('A', 'B'))

```

Cluster merging and annotation

In our experience, manual merging of clusters leads to slightly different results compared to an algorithm with a specified number of clusters. In order to detect somewhat rare populations, some level of over-clustering is necessary so that the more subtle populations become separated from the main populations. In addition, merging can always follow an over-clustering step, but splitting of existing clusters is not generally feasible.

In our setup, over-clustering is also useful when the interest is in identifying the “natural” number of clusters present in the data. Additionally to the t-SNE plots, one could investigate the delta area plot from the *ConsensusClusterPlus* package and the hierarchical clustering dendrogram of the over-clustered subpopulations, as shown below.

In our example, we expect around 6 specific cell types, and we have performed *FlowSOM* clustering into 20 groups as a reasonable over-estimate. After analyzing the heatmaps and t-SNE plots, we can clearly see that stratification of the data into 20 clusters may be too strong. Many clusters are placed very close to each other indicating that they could be merged together. The same can be deduced from the heatmaps, highlighting that marker expression patterns for some neighboring clusters are very similar. Cluster merging and annotating is somewhat manual, based partially on visual inspection of t-SNE plots and heatmaps and thus, benefits from expert knowledge of the cell types.

Manual cluster merging and annotating based on heatmaps

In our experience, the main reference for manual merging of clusters is the heatmap of marker characteristics across metaclusters with dendrograms showing the hierarchy of similarities. Importantly, such plots aggregate information over many cells, thus showing an average picture of each cluster, but combined with dimensionality reduction, gives a good insight into the data structure. Given expert knowledge of the cell types and markers, it is then left to the researcher to decide how exactly to merge clusters (e.g., with higher weight given to some markers).

The dendrogram highlights the similarity between the metaclusters and can be used explicitly for the merging. However, there are reasons why we would not always follow exactly the dendrogram. In general, when it comes to clustering, blindly following the hierarchy of codes will lead to identification of populations of similar cells, but it does not necessarily mean that they are of biological interest. The distances between metaclusters are calculated over all the markers and it may be that some markers carry higher weight for certain cell types. In addition, different linkage methods may lead to different hierarchy, especially when clusters are not fully distinct. Another aspect to consider in cluster merging is the cluster size, represented in the parentheses next to the cluster label in our plots. If the cluster size is very small, but the cluster seems relevant and distinct, one can keep it as separate. However, if it is small and different from the neighboring clustering only in a somewhat unimportant marker, it could be merged. And, if some of the metaclusters do not represent any

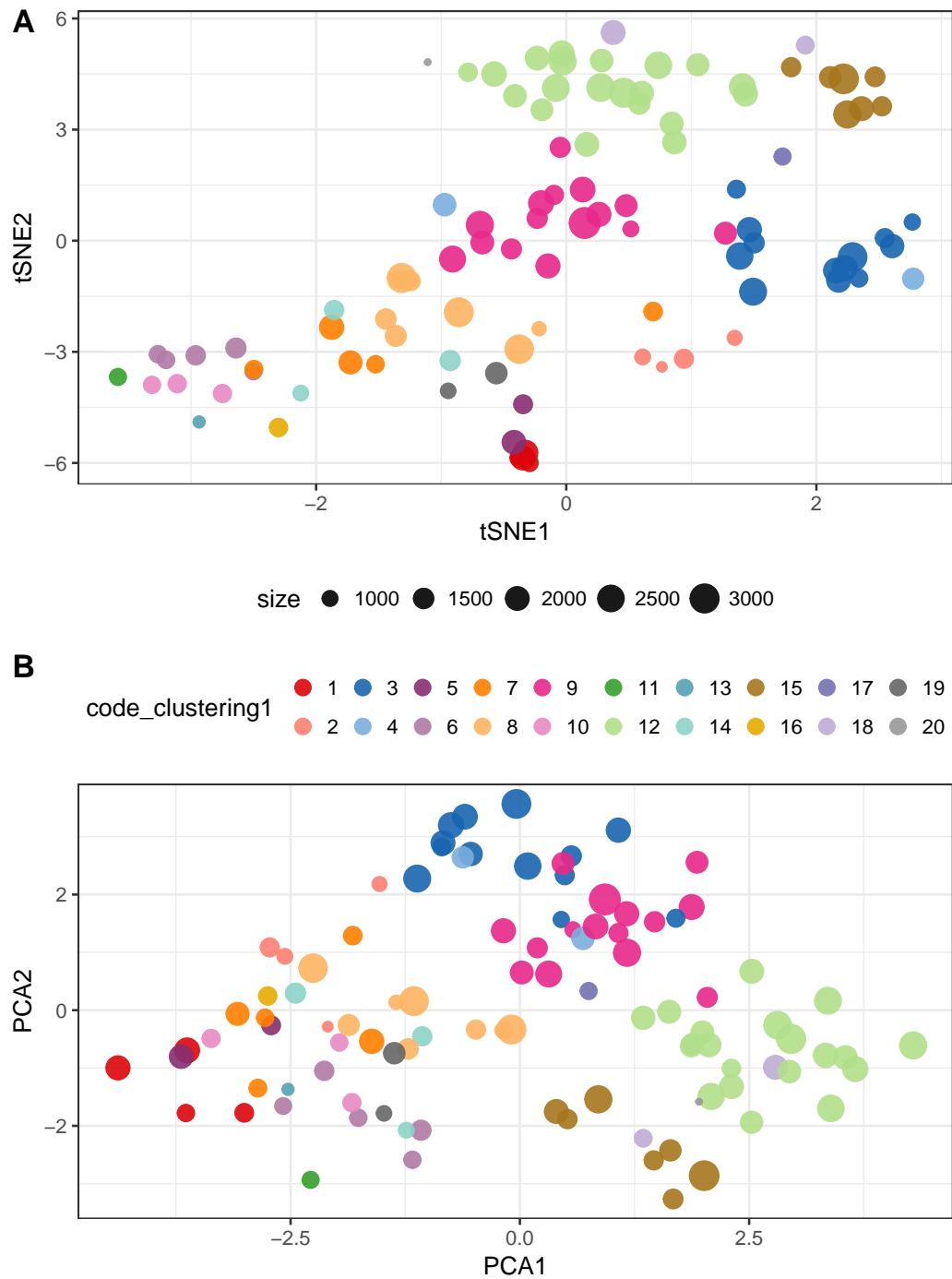


Figure 11. t-SNE plot (A) and PCA plot (B) representing the 100 SOM codes colored according to the metaclustering into 20 cell populations.

specific cell types, they could be dropped out of the downstream analysis instead of being merged. However, in case an automated solution for cluster merging is truly needed, one could use the `cutree()` function applied to the dendrogram.

Based on the seed that was set, cluster merging of the 20 metaclusters is defined in the `PBMC8_cluster_merging1.xlsx` file on the Robinson Lab server with the IDs of the original clusters and new cluster names, and we save it as a `cluster_merging1` data frame. The expert has annotated 8 cell populations: CD8 T-cells, CD4 T-cells, B-cells IgM-, B-cells IgM+, NK cells, dendritic cells (DC), monocytes and surface negative cells; monocytes could be further subdivided based on HLA-DR into high, medium and low subtypes.

```
cluster_merging1_filename <- "PBMC8_cluster_merging1.xlsx"
download.file(file.path(url, cluster_merging1_filename),
  destfile = cluster_merging1_filename)
cluster_merging1 <- read_excel(cluster_merging1_filename)
data.frame(cluster_merging1)

##   original_cluster  new_cluster
## 1                 1 B-cells IgM+
## 2                 2 surface-
## 3                 3 NK cells
## 4                 4 CD8 T-cells
## 5                 5 B-cells IgM-
## 6                 6 monocytes
## 7                 7 monocytes
## 8                 8 CD8 T-cells
## 9                 9 CD8 T-cells
## 10                10 monocytes
## 11                11 monocytes
## 12                12 CD4 T-cells
## 13                13 DC
## 14                14 CD8 T-cells
## 15                15 CD4 T-cells
## 16                16 DC
## 17                17 CD4 T-cells
## 18                18 CD4 T-cells
## 19                19 CD4 T-cells
## 20                20 CD4 T-cells

## New clustering1m
mm <- match(cell_clustering1, cluster_merging1$original_cluster)
cell_clustering1m <- cluster_merging1$new_cluster[mm]

mm <- match(code_clustering1, cluster_merging1$original_cluster)
code_clustering1m <- cluster_merging1$new_cluster[mm]
```

We update the t-SNE plot with the new annotated cell populations.

```
dr$cell_clustering1m <- factor(cell_clustering1m[tsne_inds])
ggplot(dr, aes(x = tSNE1, y = tSNE2, color = cell_clustering1m)) +
  geom_point() +
  theme_bw() +
  scale_color_manual(values = color_clusters) +
  guides(color = guide_legend(override.aes = list(size = 4)))
```

One of the usefull representations of merging is a heatmap of median marker expression in each of the original clusters, which are labeled according to the proposed merging.

```
plot_clustering_heatmap_wrapper(expr = expr[, lineage_markers_ord],
  expr01 = expr01[, lineage_markers_ord], cell_clustering = cell_clustering1,
  color_clusters = color_clusters, cluster_merging = cluster_merging1)
```

To get a final summary of the annotated cell types, one can plot a heatmap of median marker expression, calculated based on the cells in each of the annotated populations.

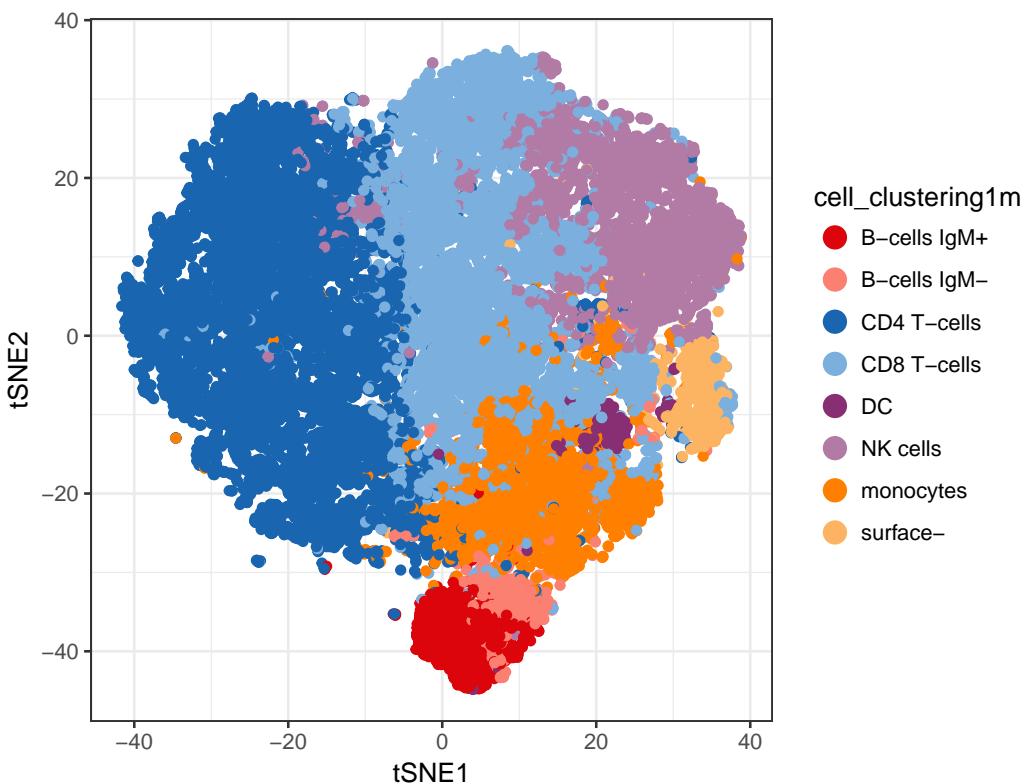


Figure 12. t-SNE plot with cells colored according to the merging of 20 metaclusters.

```
plot_clustering_heatmap_wrapper(expr = expr[, lineage_markers_ord],
                                expr01 = expr01[, lineage_markers_ord], cell_clustering = cell_clustering1m,
                                color_clusters = color_clusters)
```

Reducing the number of clusters in ConsensusClusterPlus

The *ConsensusClusterPlus* package provides visualizations that can help to understand the metaclustering process and the characteristics of the analyzed data. For example, the delta area plot (see below) highlights the amount of extra cluster stability gained when clustering into k groups as compared to k-1 groups (from k=2 to k=20). It can be expected that high stability of clusters can be reached when clustering into the number of groups that best fits the data. Thus, this score could be used as a method for finding the “natural” number of clusters present in the data and this corresponds to the value of k where there is no appreciable increase in stability. This strategy can be referred as the “elbow criterion”. For more details about the meaning of this plot, the user can refer to the original description of the consensus clustering method (Monti et al. 2003).

The elbow criterion is quite subjective since the “appreciable” increase is not defined exactly. For example, in the delta plot below, we could say that the last point before plateau is for k=6, or for k=5, or for k=3, depending on our perception of sufficient decrease of the delta area score. Moreover, the exact point where a plateau is reached may vary for runs with different random seeds, the drop may not always be so sharp and and the function is not guaranteed to be decreasing. It is advised to investigate more of those plots and the resulting t-SNE and heatmaps before drawing any conclusions about the final number of “natural” clusters.

Manual merging of up to 20 clusters can be laborious. To simplify this task, one could reduce the strength of over-clustering and allow the metaclustering method to do a part of the merging, which then can be completed manually. Analyzing the delta plot from the right side, we can see how much we can reduce the strength of over-clustering while still obtaining stable clusters. In parallel, one should check the heatmaps to see whether the less stringent stratification is able to capture cell populations of interest.

As an example, we choose to reduce the strength of metaclustering to 12 groups.

```
## Get cluster ids for each cell
nmc2 <- 12
code_clustering2 <- mc[[nmc2]]$consensusClass
cell_clustering2 <- code_clustering2[som$map$mapping[, 1]]
```

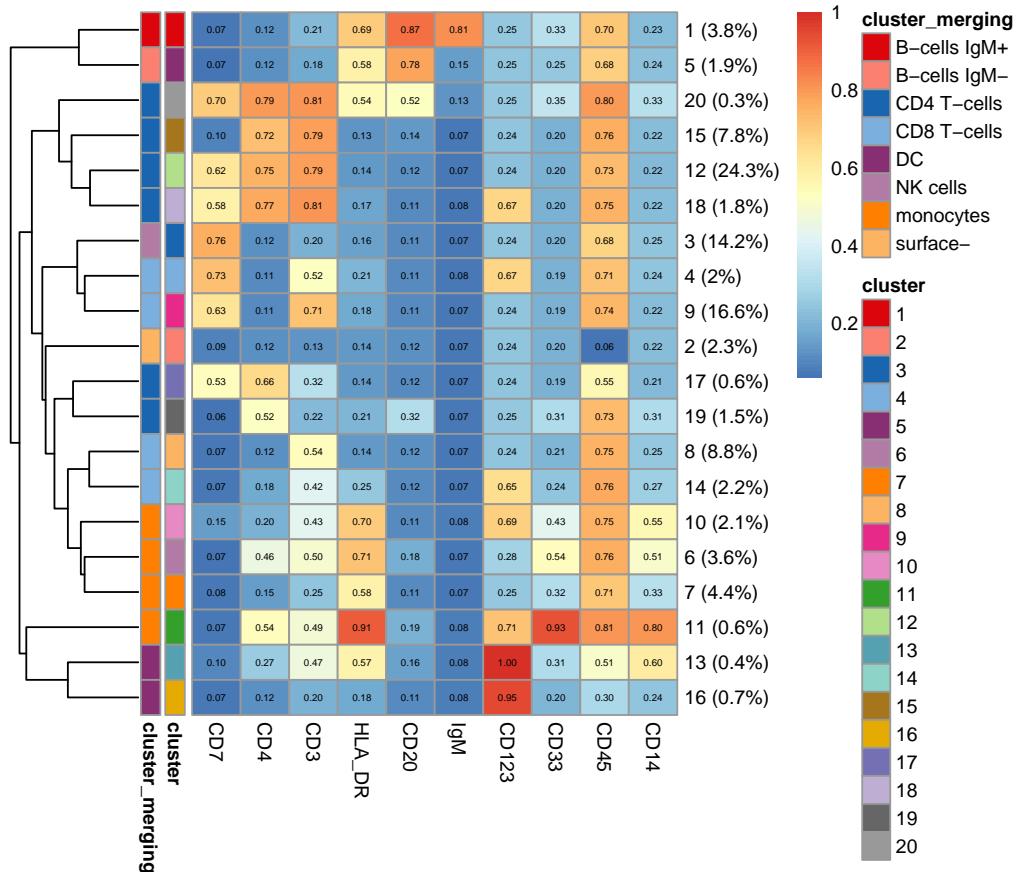


Figure 13. Heatmap showing how the 20 metaclusters are merged.

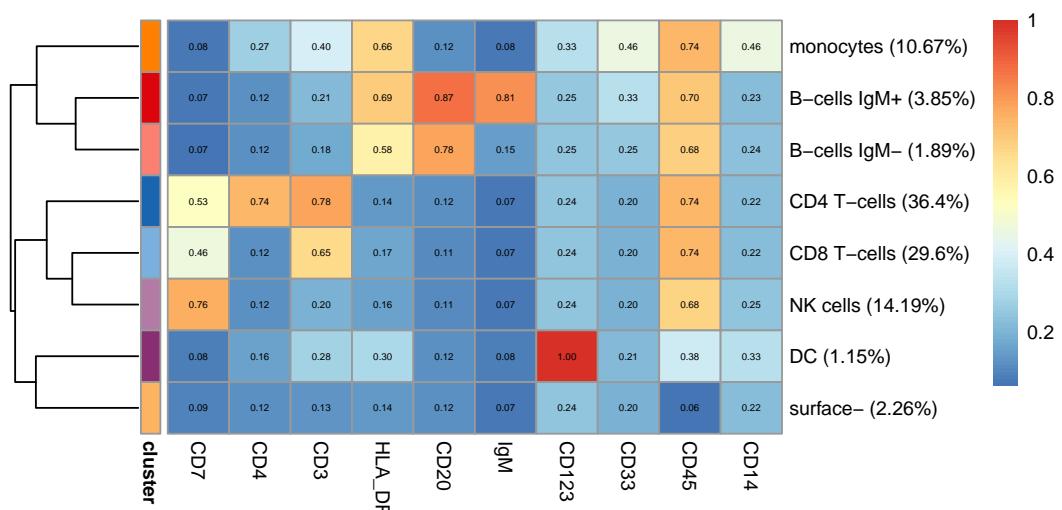


Figure 14. Heatmap of cell populations obtained from merging the 20 metaclusters.

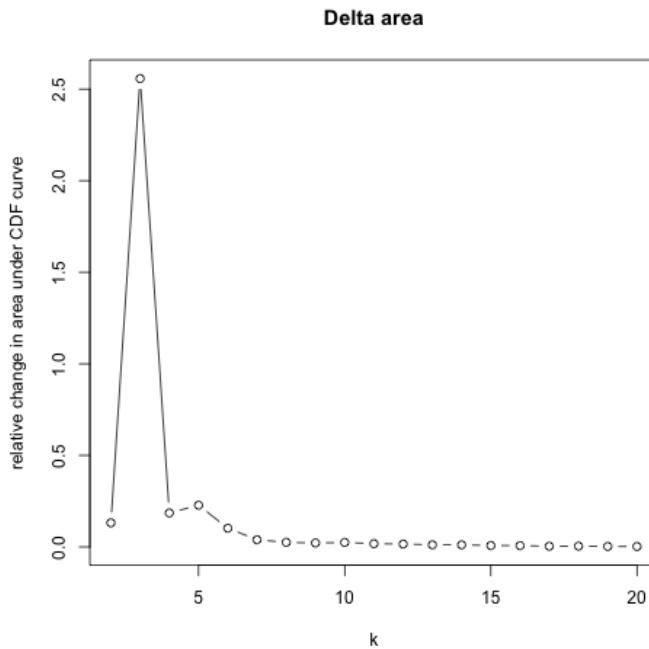


Figure 15. The delta area plot from ConsensusClusterPlus indicating the relative increase in cluster stability obtained when clustering into k groups.

In the t-SNE plot, we can see that many small clusters obtained when stratifying data into 20 groups are now merged together, which should simplify the new cluster annotation.

```
dr$cell_clustering2 <- factor(cell_clustering2[tsne_inds], levels = 1:nmc2)
ggplot(dr, aes(x = tSNE1, y = tSNE2, color = cell_clustering2)) +
  geom_point() +
  theme_bw() +
  scale_color_manual(values = color_clusters) +
  guides(color = guide_legend(override.aes = list(size = 4), ncol = 2))
```

```
plot_clustering_heatmap_wrapper(expr = expr[, lineage_markers_ord],
  expr01 = expr01[, lineage_markers_ord], cell_clustering = cell_clustering2,
  color_clusters = color_clusters)
```

Over-clustering into as few as 12 groups still allows us to identify the same 8 cell populations as when merging 20 clusters, but is simpler to define since fewer profiles need to be manually scanned. The expert-based merging is saved in the PBMC8_cluster_merging2.xlsx file.

```
cluster_merging2_filename <- "PBMC8_cluster_merging2.xlsx"
download.file(file.path(url, cluster_merging2_filename),
  destfile = cluster_merging2_filename)
cluster_merging2 <- read_excel(cluster_merging2_filename)
data.frame(cluster_merging2)
```

```
##   original_cluster new_cluster
## 1                 1 B-cells IgM+
## 2                 2 surface-
## 3                 3 NK cells
## 4                 4 CD8 T-cells
## 5                 5 B-cells IgM-
## 6                 6 monocytes
## 7                 7 CD8 T-cells
## 8                 8 CD8 T-cells
```

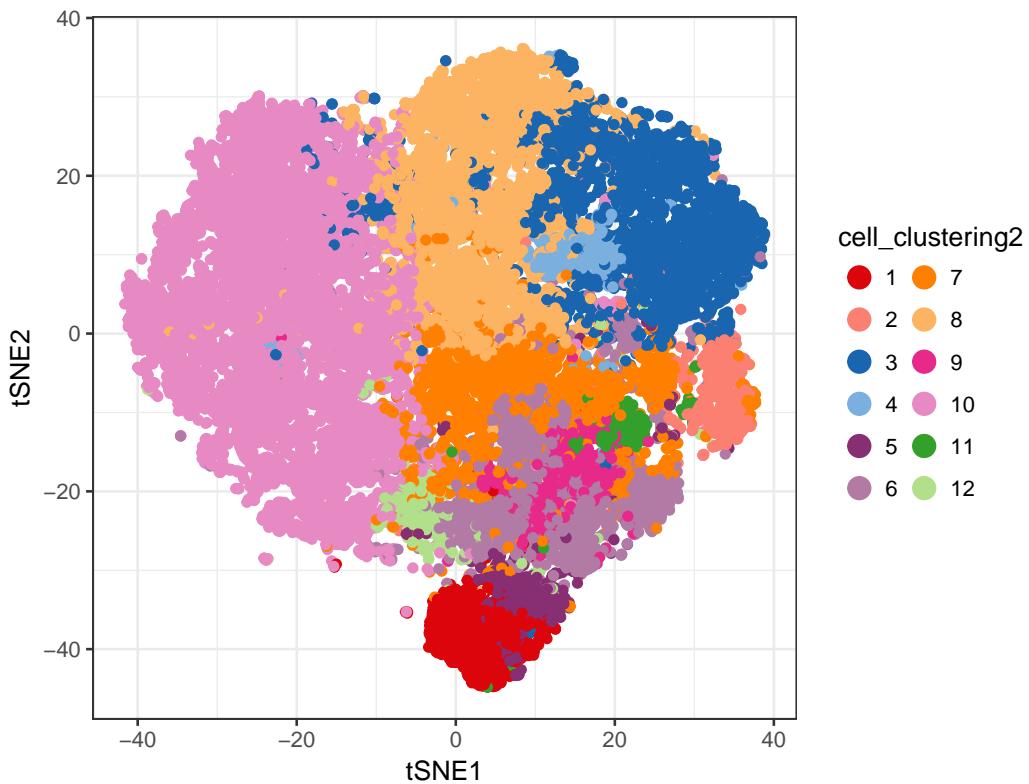


Figure 16. t-SNE plot with cells colored according to the 12 metaclusters.

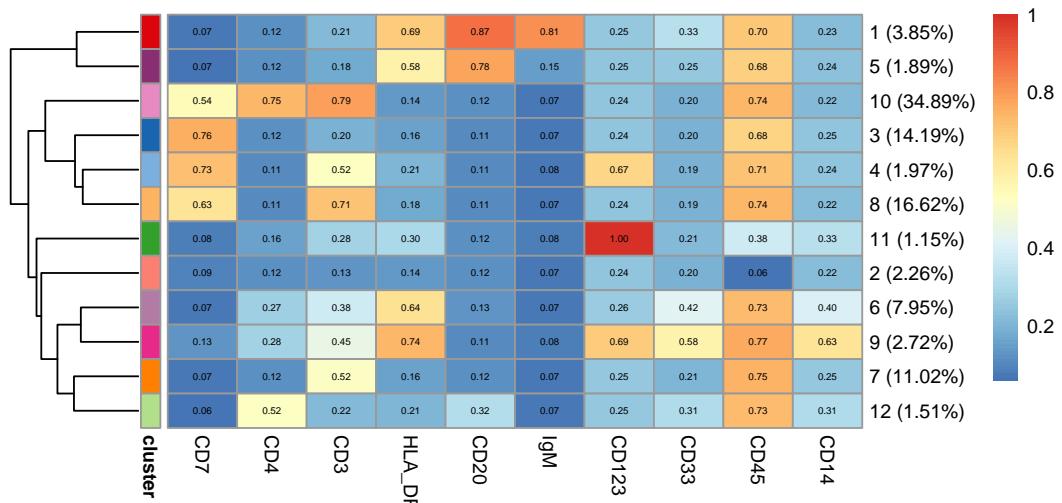


Figure 17. Heatmap of the median marker intensities in 12 cell populations obtained from the meta-clustering step.

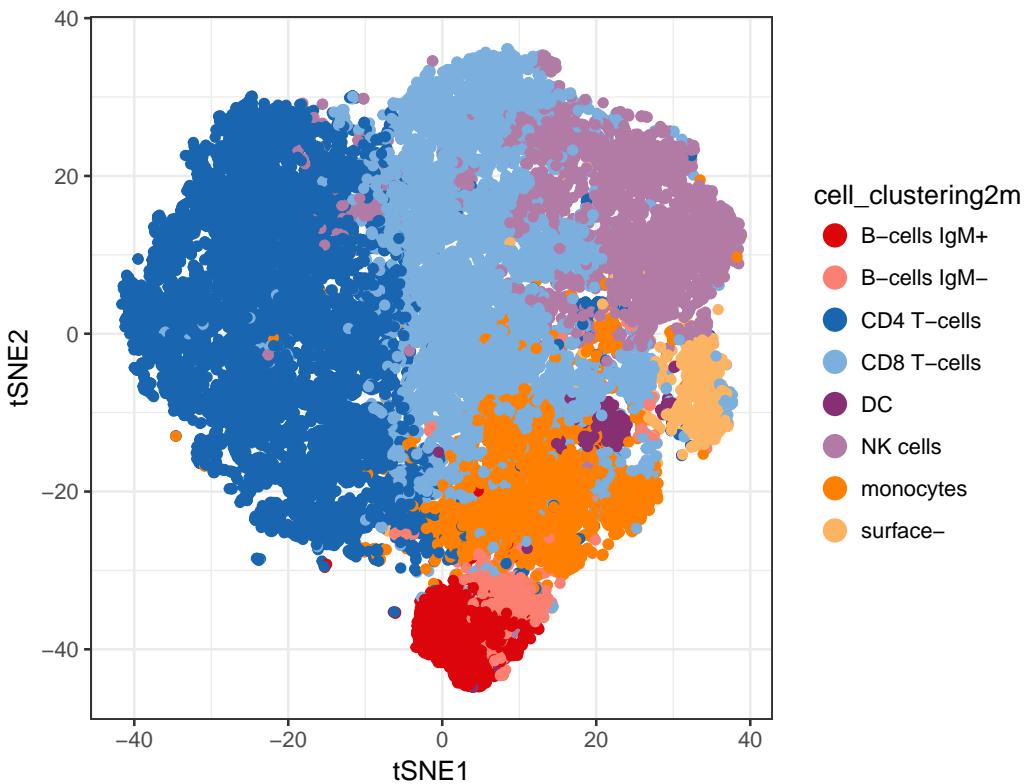


Figure 18. t-SNE plot with cells colored according to the merging of 12 metaclusters.

```

## 9      9    monocytes
## 10     10   CD4 T-cells
## 11     11     DC
## 12     12   CD4 T-cells

## New clustering2m
mm <- match(cell_clustering2, cluster_merging2$original_cluster)
cell_clustering2m <- cluster_merging2$new_cluster[mm]

dr$cell_clustering2m <- factor(cell_clustering2m[tsne_inds])
gg_tsne_cl2m <- ggplot(dr, aes(x = tSNE1, y = tSNE2, color = cell_clustering2m)) +
  geom_point() +
  theme_bw() +
  scale_color_manual(values = color_clusters) +
  guides(color = guide_legend(override.aes = list(size = 4)))
gg_tsne_cl2m

plot_clustering_heatmap_wrapper(expr = expr[, lineage_markers_ord],
  expr01 = expr01[, lineage_markers_ord], cell_clustering = cell_clustering2,
  color_clusters = color_clusters, cluster_merging = cluster_merging2)

plot_clustering_heatmap_wrapper(expr = expr[, lineage_markers_ord],
  expr01 = expr01[, lineage_markers_ord], cell_clustering = cell_clustering2m,
  color_clusters = color_clusters)

```

Comparison of automated and manual merging

The manual merging of 20 (or 12) clusters by an expert resulted in identification of 8 cell populations. To highlight the impact of manual merging versus algorithm-defined subpopulations, we compare to the results of an automated cluster merging that is set to stratify the data also into 8 clusters. We extract the result from

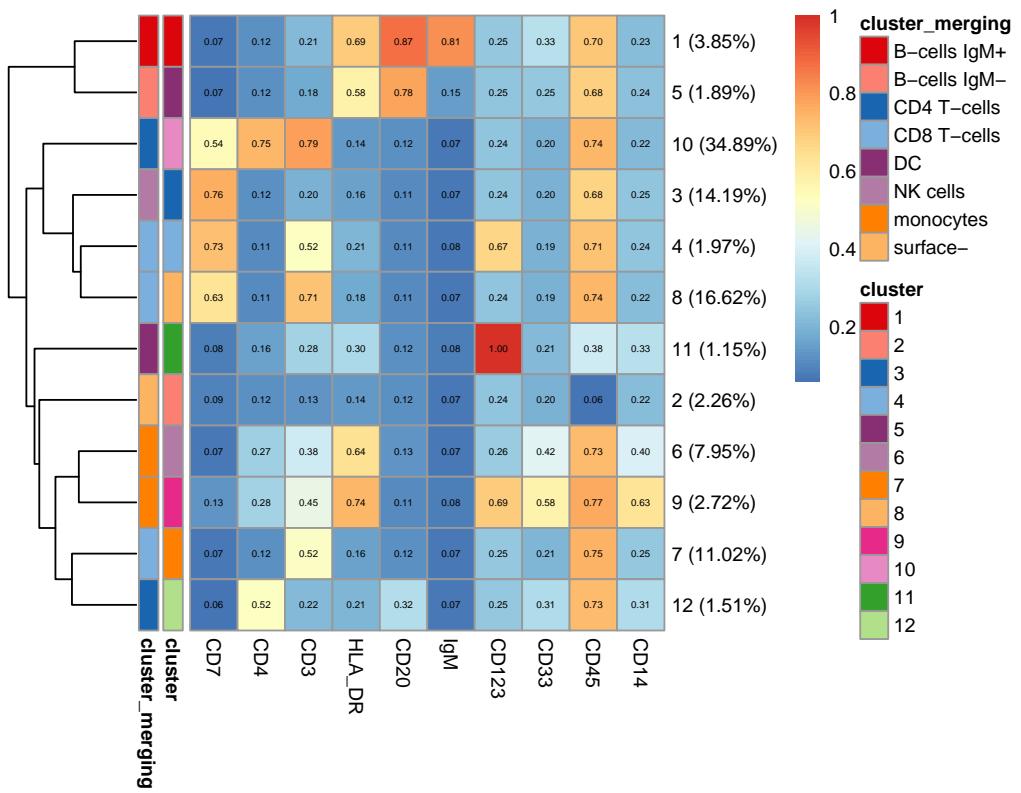


Figure 19. Heatmap of the median marker intensities in 12 cell populations obtained from the meta-clustering step along with the annotation of cell populations.

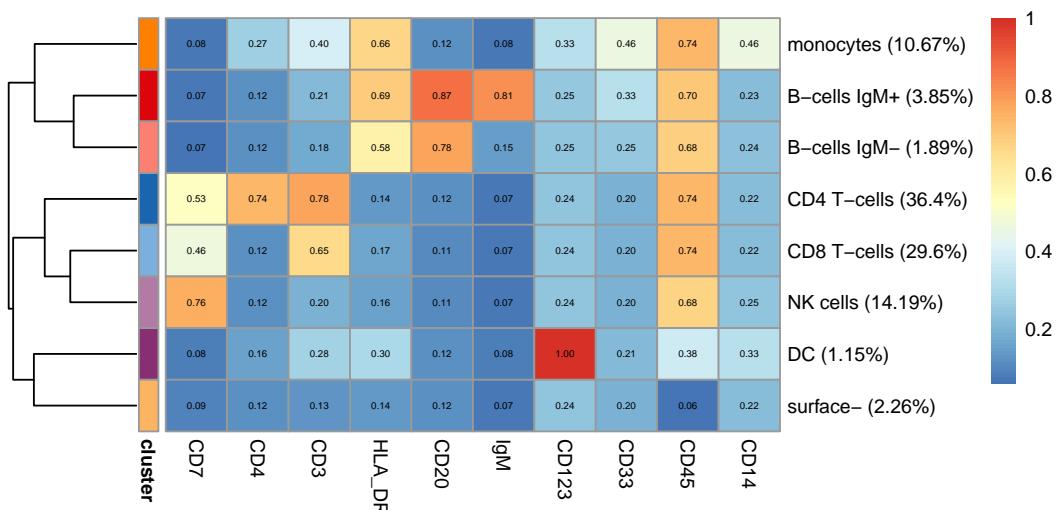


Figure 20. Heatmap of cell populations obtained from merging the 12 metaclusters.

the ConsensusClusterPlus output. Out of interest, we can see which clusters are split by tabulating the cell labels.

```
## Get cluster ids for each cell
nmc3 <- 8
code_clustering3 <- mc[[nmc3]]$consensusClass
cell_clustering3 <- code_clustering3[som$map$mapping[, 1]]

# tabular comparison of cell_clustering3 and cell_clustering2m
table(algorithm=cell_clustering3, manual=cell_clustering2m)

##          manual
## algorithm B-cells IgM+ B-cells IgM- CD4 T-cells CD8 T-cells DC NK cells
##      1           6651            0       0       0       0       0       0       0       0
##      2             0            0       0       0       0       0       0       0       0
##      3             0            0       0       0       32112    0     24518
##      4             0            0       3265      0       0       0       0       0       0
##      5             0            0       0       0       0       0       0       0       0
##      6             0            0       0       2603    19038    0       0       0       0
##      7             0            0       0       60287      0       0       0       0       0
##      8             0            0       0       0       0       0     1980      0       0

##          manual
## algorithm monocytes surface-
##      1           0            0
##      2           0            3901
##      3           0            0
##      4           0            0
##      5           18436        0
##      6           0            0
##      7           0            0
##      8           0            0
```

In the t-SNE map, we can see that part of the new cell populations (cluster 7, 1 and 4, 2, 5 and 8) overlaps substantially with populations obtained by the means of manual merging (CD4 T-cells, B-cells, surface-, monocytes and DC). However, cells that belong to cluster 3 and 6 are subdivided in a different manner according to the manual merging. Cluster 3 consists of CD8 T-cells and NK cells, and the latter can not be identified anymore based on the heatmap corresponding to clustering into 8 groups.

```
dr$cell_clustering3 <- factor(cell_clustering3[tsne_inds], levels = 1:nmc3)
gg_tsne_cl3 <- ggplot(dr, aes(x = tSNE1, y = tSNE2, color = cell_clustering3)) +
  geom_point() +
  theme_bw() +
  scale_color_manual(values = color_clusters) +
  guides(color = guide_legend(override.aes = list(size = 4)))
plot_grid(gg_tsne_cl2m, gg_tsne_cl3, ncol = 1, labels = c('A', 'B'))
```

```
plot_clustering_heatmap_wrapper(expr = expr[, lineage_markers_ord],
  expr01 = expr01[, lineage_markers_ord], cell_clustering = cell_clustering3,
  color_clusters = color_clusters)
```

The example above highlights the difference between automatic clustering and manual merging of algorithm-generated clusters in the search for biologically meaningful cell populations. Automated and manual merging may give different weight to marker importance and thus result in different populations being detected. However, in our view, the manual merging here done in a reproducible fashion results in a more biologically meaningful cell stratification.

Differential analysis

For the dataset used in this workflow, we perform three types of analyses that aim at identifying subsets of PBMCs and signaling markers that respond to BCR/FcR-XL stimulation compared to unstimulated. First, we describe the differential abundance of the defined cell populations, followed by differential analysis of marker expression within each cluster. Finally, differential expression of the overall aggregated marker expression could also be of interest.

The PBMC subset analyzed in this workflow originates from a paired experiment, where samples from 8 patients were treated with 12 different stimulation conditions for 30 min, together with unstimulated reference

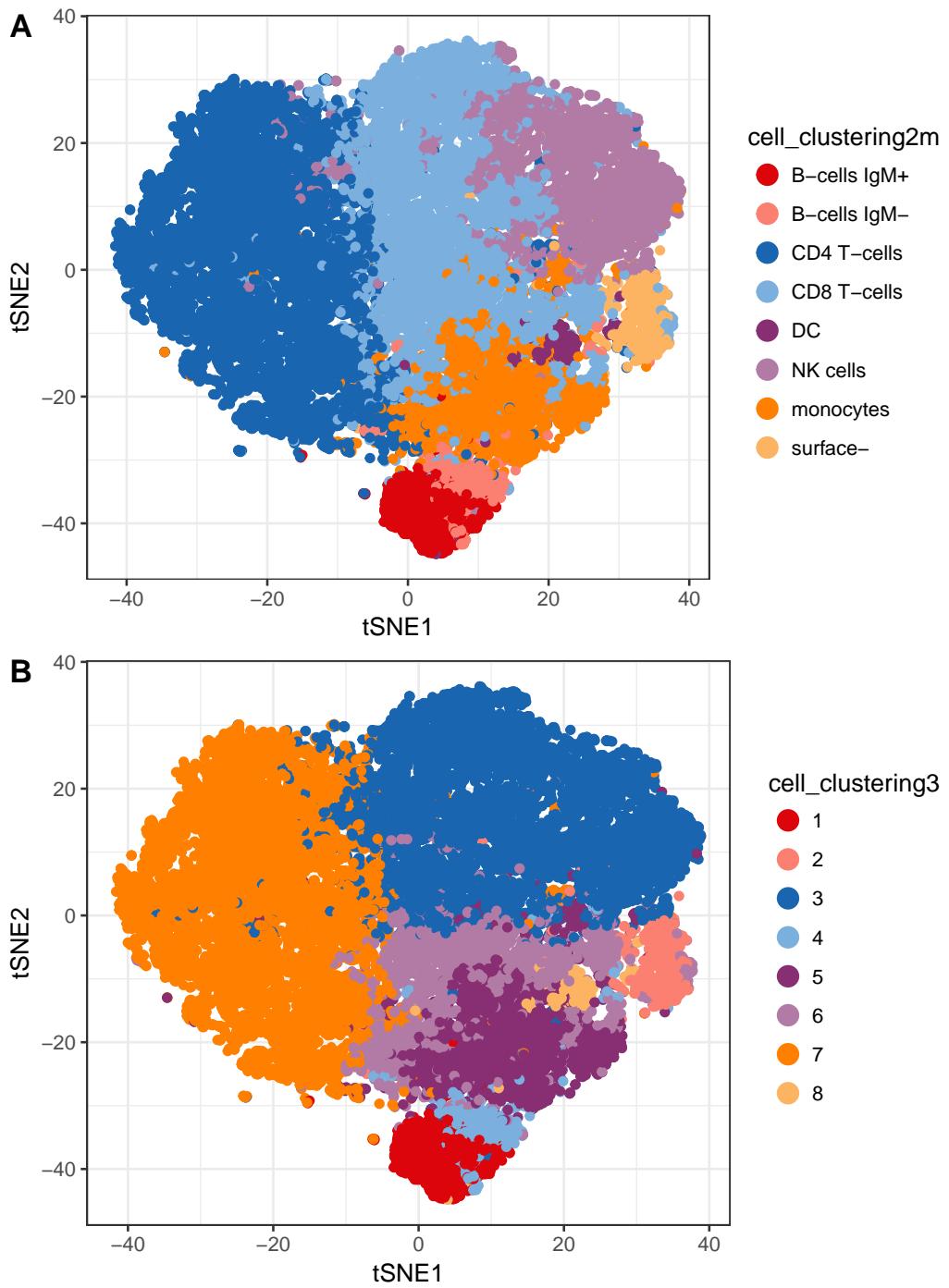


Figure 21. t-SNE plot with cells colored according to (A) the expert merging of 12 metaclusters into 8 cell populations and (B) the 8 automatically detected metaclusters.

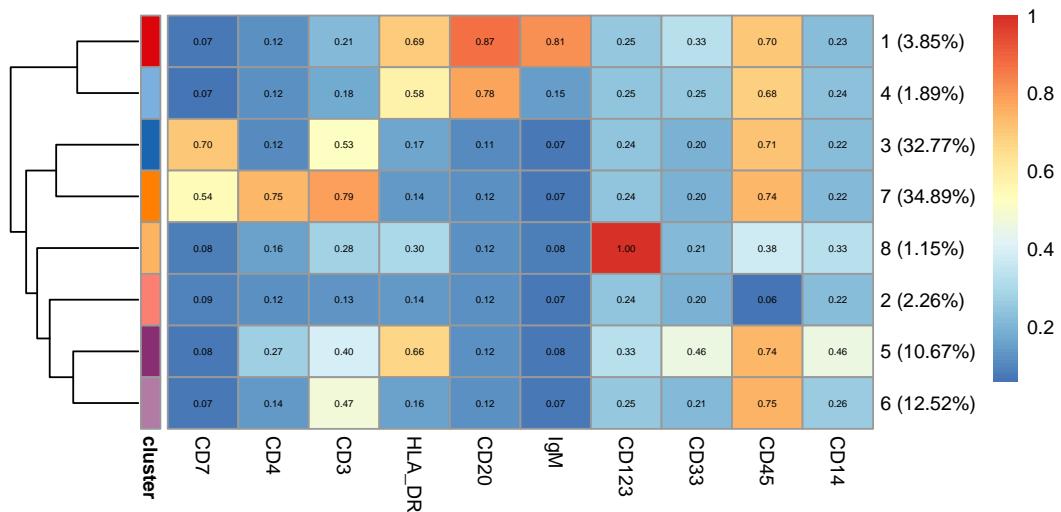


Figure 22. Heatmap of the median marker intensities in 8 cell populations obtained from the metaclustering step.

samples. This is a natural example where one would choose a mixed model to model the response (abundance or marker signal), where patients would be treated as a random effect. In this way, one can formally account for within-patient variability, as noted above in the MDS plot and should give a gain in power to detect differences between conditions.

We use the *stats* and *lme4* packages to fit the fixed and mixed models, respectively, and the *multcomp* package for hypothesis testing. In all differential analyses here, we want to test for differences between the reference (Ref) and BCR/FcR-XL treatment (BCRXL). The fixed model formula is straightforward: `~ condition`, where `condition` indicates the treatment group. The corresponding full model design matrix consists of the intercept and dummy variable indicating the treated samples. In the presence of batches, one can include them in the model by using a formula `~ condition + batch`, or if they affect the treatment, a formula with interactions `~ condition * batch`.

For testing, we use the general linear hypotheses function `glht`, which allows testing arbitrary hypotheses. The `linfct` parameter specifies the linear hypotheses to be tested. It should be a matrix where each row corresponds to one comparison (contrast), and the number of columns must be the same as in the design matrix. In our analysis, the contrast matrix indicates that the regression coefficient corresponding to `conditionBCRXL` is tested to be equal to zero, i.e., we test the null hypothesis that there is no effect of the BCR/FcR-XL treatment. The result of the test is a p-value, which is a probability of observing an as strong (or stronger) difference between the two conditions assuming the null hypothesis is true (i.e., that there is no difference between treated and untreated).

In our analysis, testing is performed on each cluster and marker separately, resulting in 8 tests for differential abundance (one for each merged population), 14 tests for overall differential marker expression analysis and 8×14 tests for differential marker expression within-populations. Thus, to account for the multiple testing correction, we apply the Benjamini & Hochberg adjustment to each of them using an FDR cutoff of 5%.

```

library(lme4)
library(multcomp)
## Model formula without random effects
model.matrix(~ condition, data = md)

##      (Intercept) conditionBCRXL
## 1              1              1
## 2              1              0
## 3              1              1
## 4              1              0
## 5              1              1
## 6              1              0
## 7              1              1
## 8              1              0
## 9              1              1
## 10             1              0
## 11             1              1

```

```

## 12      1      0
## 13      1      1
## 14      1      0
## 15      1      1
## 16      1      0
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$condition
## [1] "contr.treatment"

## Create contrasts
contrast_names <- c("BCRXLvsRef")
k1 <- c(0, 1)
K <- matrix(k1, nrow = 1, byrow = TRUE, dimnames = list(contrast_names))
K

##          [,1] [,2]
## BCRXLvsRef    0    1

FDR_cutoff <- 0.05

```

Differential cell population abundance

Differential analysis of cell population abundance compares the proportions of cell types across experimental conditions and aims at highlighting populations that are present at different ratios. First, we calculate two tables: one that contains cell counts for each sample and population and one with the corresponding proportions of cell types by sample. The proportions are used only for plotting, since the statistical modeling takes the cell counts by cluster and sample as input.

```

counts_table <- table(cell_clustering1m, sample_ids)
props_table <- t(counts_table) / colSums(counts_table) * 100

counts <- as.data.frame.matrix(counts_table)
props <- as.data.frame.matrix(props_table)

```

For each sample, we plot its PBMC cell type composition represented with colored bars, where the size of a given stripe reflects proportion of the corresponding cell type in a given sample.

```

ggdf <- melt(data.frame(cluster = rownames(props), props),
  id.vars = "cluster", value.name = "proportion", variable.name = "sample_id")

## Add condition info
mm <- match(ggdf$sample_id, md$sample_id)
ggdf$condition <- factor(md$condition[mm])

ggplot(ggdf, aes(x = sample_id, y = proportion, fill = cluster)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ condition, scales = "free_x") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_fill_manual(values = color_clusters)

```

It may be quite hard to see the differences in the cluster abundances in the plot above, especially for clusters with very low frequency. And, since boxplots cannot represent multimodal distributions, we show boxplots with jittered points of the sample-level cluster proportions overlaid. The y-axes are scaled to the range of data plotted for each cluster to better visualize the differences in frequency of lower abundance clusters. For this experiment, it may be interesting to additionally depict the patient information. We do this by plotting a different point shape for each patient. Indeed, we can see that often the direction of abundance changes between the two conditions is concordant among the patients.

```

ggdf$patient_id <- factor(md$patient_id[mm])

ggplot(ggdf) +

```

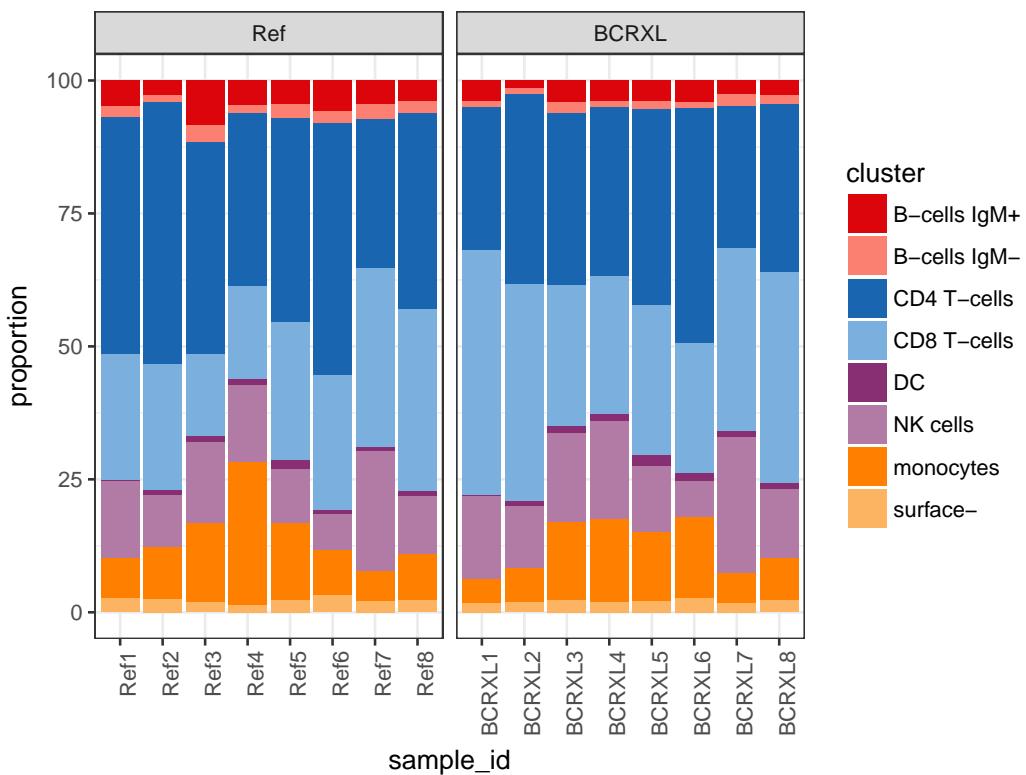


Figure 23. Relative abundance of PBMC populations in each sample represented with a barplot.

```
geom_boxplot(aes(x = condition, y = proportion, color = condition,
                 fill = condition), position = position_dodge(), alpha = 0.5,
                 outlier.color = NA) +
  geom_point(aes(x = condition, y = proportion, color = condition,
                 shape = patient_id), alpha = 0.8, position = position_jitterdodge()) +
  facet_wrap(~ cluster, scales = "free", nrow = 2) +
  theme_bw() +
  theme(axis.text.x = element_text(size = 6)) +
  scale_color_manual(values = color_conditions) +
  scale_fill_manual(values = color_conditions) +
  scale_shape_manual(values = c(16, 17, 8, 3, 12, 0, 1, 2))
```

As our goal is to compare proportions, one could take these values, transform them (e.g., logit) and use them as a dependent variable in a linear model. However, this approach does not take into account the uncertainty of proportion estimates, which is higher when ratios are calculated for samples with lower total cell counts. A distribution that naturally accounts for such uncertainty is the binomial distribution (i.e., logistic regression), which takes the cell counts as input (relative to the total for each sample). Nevertheless, as in the genomic data analysis, the pure logistic regression is not able to capture the overdispersion that is present in HDCyo data. A natural extension to model the extra variation is the generalized linear mixed model (GLMM) where the random effect is defined by the sample ID (Zhao et al. 2013; Jia et al. 2014). In our example, additionally the patient pairing could be accounted in the model by incorporating a random intercept for each patient. Thus, we present two GLMMs. Both of them comprise a random effect defined by the sample ID to model the overdispersion in proportions. The second model, includes additionally a random effect defined by the patient ID to account for the experiment pairing.

In our model, the blocking variable is patient ID $i = 1, \dots, n$, where $n = 8$. For each patient, there are n_i samples measured, and $j = 1, \dots, n_i$ indicates the sample ID. Here, $n_i = 2$ for all i (one from reference and one from BCR/FcR-XL stimulated).

We assume that for a given cell population, the cell counts Y_{ij} follow a binomial distribution $Y_{ij} \sim Bin(m_{ij}, \pi_{ij})$, where m_{ij} is a total number of cells in a sample corresponding to patient i and condition j . The generalized linear mixed model with observation-level random effects ξ_{ij} is defined as follows:

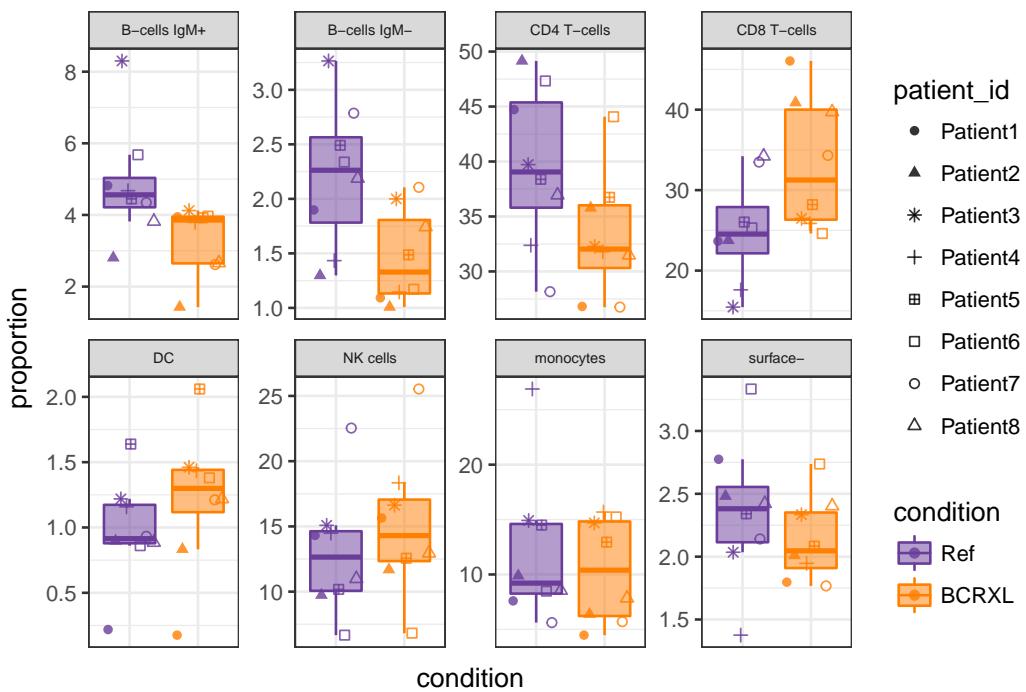


Figure 24. Relative abundance of PBMC populations in each sample plotted using boxplots.

$$E(Y_{ij}|\beta_0, \beta_1, \xi_{ij}) = \text{logit}^{-1}(\beta_0 + \beta_1 x_{ij} + \xi_{ij}),$$

where $\xi_{ij} \sim N(0, \sigma_\xi^2)$ and x_{ij} corresponds to the condition BCRXL column in the design matrix and indicates whether a sample i_j belongs to the reference ($x_{ij} = 0$) or treated condition ($x_{ij} = 1$). Since $E(Y_{ij}|\beta_0, \beta_1, \xi_{ij}) = \pi_{ij}$, the above formula can be written as follows:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + \xi_{ij}.$$

The generalized linear mixed model that furthermore accounts for the patient pairing incorporates additionally a random intercept for each patient i :

$$E(Y_{ij}|\beta_0, \beta_1, \gamma_i, \xi_{ij}) = \text{logit}^{-1}(\beta_0 + \beta_1 x_{ij} + \gamma_i + \xi_{ij}),$$

where $\gamma_i \sim N(0, \sigma_\gamma^2)$.

```
formula_glmer_binomial1 <- y/total ~ condition + (1|sample_id)
formula_glmer_binomial2 <- y/total ~ condition + (1|patient_id) + (1|sample_id)
```

The wrapper function below takes as input a data frame with cell counts (each row is a population, each column is a sample), the metadata table, the formula and performs the differential analysis specified with contrast K for each population separately and returns a table with non-adjusted and adjusted p-values.

```
differential_abundance_wrapper <- function(counts, md, formula, K){
  ## Fit the GLMM for each cluster separately
  ntot <- colSums(counts)
  fit_binomial <- lapply(1:nrow(counts), function(i){

    data_tmp <- data.frame(y = as.numeric(counts[i, md$sample_id]),
                           total = ntot[md$sample_id], md)

    fit_tmp <- glmer(formula, weights = total, family = binomial,
                     data = data_tmp)

    ## Fit contrasts one by one
    out <- apply(K, 1, function(k){
      contr_tmp <- glht(fit_tmp, linfct = matrix(k, 1))
      }
```

```

        summ_tmp <- summary(contr_tmp)
        pval <- summ_tmp$test$pvalues
        return(pval)
    })
    return(out)
})
pvals <- do.call(rbind, fit_binomial)
colnames(pvals) <- paste0("pval_", contrast_names)
rownames(pvals) <- rownames(counts)
## Adjust the p-values
adjp <- apply(pvals, 2, p.adjust, method = "BH")
colnames(adjp) <- paste0("adjp_", contrast_names)
return(list(pvals = pvals, adjp = adjp))
}

```

We fit both of the GLMMs specified above. We can see that accounting for the patient pairing increases the sensitivity to detect differentially abundant cell populations.

```

da_out1 <- differential_abundance_wrapper(counts, md = md,
    formula = formula_glmmer_binomial1, K = K)
apply(da_out1$adjp < FDR_cutoff, 2, table)

##      adjp_BCRXLvsRef
## FALSE      5
## TRUE       3

da_out2 <- differential_abundance_wrapper(counts, md = md,
    formula = formula_glmmer_binomial2, K = K)
apply(da_out2$adjp < FDR_cutoff, 2, table)

##      adjp_BCRXLvsRef
## FALSE      2
## TRUE       6

```

An output table containing the observed cell population proportions in each sample and p-values can be assembled (and optionally written to a file).

```

da_output2 <- data.frame(cluster = rownames(props), props,
    da_out2$pvals, da_out2$adjp, row.names = NULL)
print(head(da_output2), digits = 2)

```

	cluster	BCRXL1	BCRXL2	BCRXL3	BCRXL4	BCRXL5	BCRXL6	BCRXL7	BCRXL8
## 1	B-cells IgM+	3.95	1.43	4.1	3.8	3.9	4.0	2.6	2.7
## 2	B-cells IgM-	1.09	1.01	2.0	1.1	1.5	1.2	2.1	1.7
## 3	CD4 T-cells	26.81	35.78	32.3	31.8	36.7	44.1	26.8	31.5
## 4	CD8 T-cells	46.05	40.87	26.5	25.9	28.2	24.6	34.3	39.7
## 5	DC	0.18	0.83	1.5	1.4	2.1	1.4	1.2	1.2
## 6	NK cells	15.64	11.69	16.6	18.3	12.6	6.8	25.5	12.9
##	Ref1 Ref2 Ref3 Ref4 Ref5 Ref6 Ref7 Ref8						pval_BCRXLvsRef		
## 1	4.82 2.8 8.3 4.7 4.4 5.68 4.34 3.82							3.5e-08	
## 2	1.90 1.3 3.3 1.4 2.5 2.34 2.79 2.19							2.2e-11	
## 3	44.72 49.1 39.7 32.4 38.4 47.33 28.16 36.94							1.9e-03	
## 4	23.66 23.8 15.5 17.6 26.0 25.31 33.49 34.21							1.2e-03	
## 5	0.22 0.9 1.2 1.2 1.6 0.86 0.93 0.89							7.1e-05	
## 6	14.31 9.7 15.1 14.5 10.2 6.67 22.54 10.99							4.5e-13	
##	adjp_BCRXLvsRef								
## 1	9.2e-08								
## 2	8.8e-11								
## 3	2.5e-03								
## 4	1.9e-03								
## 5	1.4e-04								
## 6	3.6e-12								

We use a heatmap to report the differential cell populations. Proportions are first scaled with the arcsine-square-root transformation (as an alternative to logit that does not return NAs when ratios are equal to zero or one). Then, Z-score normalization is applied to each population to better highlight the relative differences between compared conditions. We created two wrapper functions: `normalization_wrapper` performs the normalization of submitted expression `expr` to mean 0 and standard deviation 1, and `plot_differential_heatmap_wrapper` generates a heatmap of submitted expression `expr_norm`, where samples are grouped by condition, indicated with a color bar on top of the plot. Additionally, labels of clusters contain the adjusted p-values in parenthesis.

```

normalization_wrapper <- function(expr, th = 2.5){
  expr_norm <- apply(expr, 1, function(x){
    sdx <- sd(x, na.rm = TRUE)
    if(sdx == 0){
      x <- (x - mean(x, na.rm = TRUE))
    }else{
      x <- (x - mean(x, na.rm = TRUE)) / sdx
    }
    x[x > th] <- th
    x[x < -th] <- -th
    return(x)
  })
  expr_norm <- t(expr_norm)
}

plot_differential_heatmap_wrapper <- function(expr_norm, sign_adjp,
                                               condition, color_conditions, th = 2.5){
  ## Order samples by condition
  oo <- order(condition)
  condition <- condition[oo]
  expr_norm <- expr_norm[, oo, drop = FALSE]

  ## Create the row labels with adj p-values and other objects for pheatmap
  labels_row <- paste0(rownames(expr_norm), " (", sprintf(" %.02e", sign_adjp), ")")
  labels_col <- colnames(expr_norm)
  annotation_col <- data.frame(condition = factor(condition))
  rownames(annotation_col) <- colnames(expr_norm)
  annotation_colors <- list(condition = color_conditions)
  color <- colorRampPalette(c("#87CEFA", "#56B4E9", "#56B4E9", "#0072B2",
                             "#000000", "#D55E00", "#E69F00", "#E69F00", "#FFD700"))(100)
  breaks = seq(from = -th, to = th, length.out = 101)
  legend_breaks = seq(from = -round(th), to = round(th), by = 1)
  gaps_col <- as.numeric(table(annotation_col$condition))

  pheatmap(expr_norm, color = color, breaks = breaks,
           legend_breaks = legend_breaks, cluster_cols = FALSE, cluster_rows = FALSE,
           labels_col = labels_col, labels_row = labels_row, gaps_col = gaps_col,
           annotation_col = annotation_col, annotation_colors = annotation_colors,
           fontsize = 8)
}

## Apply the arcsine-square-root transformation
asin_table <- asin(sqrt((t(t(counts_table) / colSums(counts_table)))))
asin <- as.data.frame.matrix(asin_table)
## Keep significant clusters and sort them by significance
sign_clusters <- names(which(sort(da_out2$adjp[, "adjp_BCRXLvsRef"]) < FDR_cutoff))
## Get the adjusted p-values
sign_adjp <- da_out2$adjp[sign_clusters, "adjp_BCRXLvsRef", drop=FALSE]
## Keep samples for condition A and normalize to mean = 0 and sd = 1
asin_norm <- normalization_wrapper(asin[sign_clusters, ])

mm <- match(colnames(asin_norm), md$sample_id)
plot_differential_heatmap_wrapper(expr_norm = asin_norm, sign_adjp = sign_adjp,
                                   condition = md$condition[mm], color_conditions = color_conditions)

```

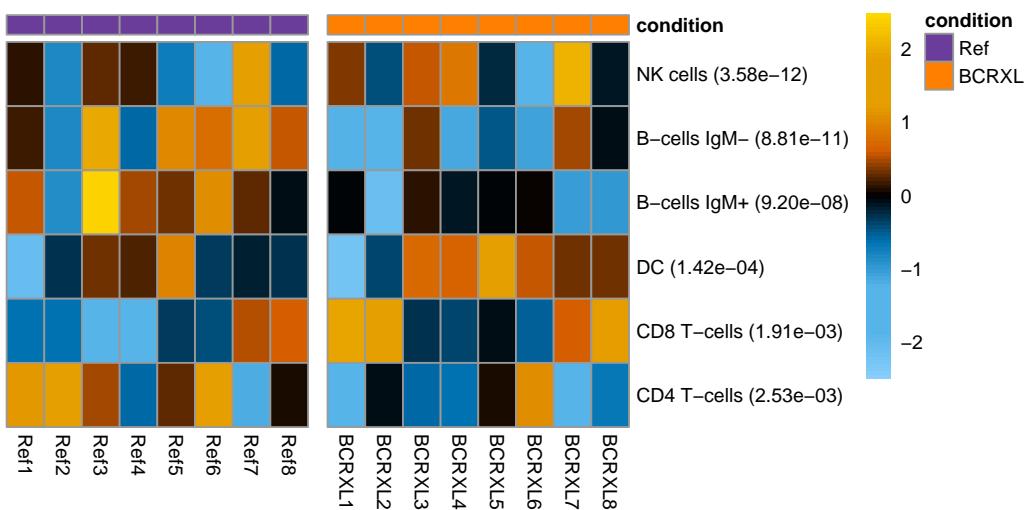


Figure 25. Normalized proportions of PBMC cell populations that are significantly differentially abundant between BCR/FcR-XL stimulated and unstimulated condition.

Differential analysis of marker expression stratified by cell population

For this part of the analysis, we calculate the median expression of the 14 signaling markers in each cell population (merged cluster) and sample. These will be used as the response variable Y_{ij} in the linear model (LM) or linear mixed model (LMM), for which we assume that the median marker expression follows a Gaussian distribution (on the already asinh-transformed scale). The linear model is formulated as follows:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and the mixed model includes a random intercept for each patient:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_i + \epsilon_{ij},$$

where $\gamma_i \sim N(0, \sigma_\gamma^2)$. In the current experiment, we have an intercept (basal level) and a single covariate, x_{ij} , which is represented as a binary (stimulated/unstimulated) variable. For more complicated designs or batch effects, additional columns of a design matrix can be used.

One drawback of summarizing the protein marker intensity with a median over cells is that all the other characteristics of the distribution, such as bimodality, skewness and variance, are ignored. On the other hand, it results in a simple, easy to interpret approach, which in many cases will be able to detect interesting changes. Another issue that arises from using a summary statistic is the level of uncertainty, which increases as the number of cells used to calculate it decreases. In the statistical modeling, this problem could be partially handled by assigning observation weights (number of cells) to each cluster and sample. However, since each cluster is tested separately, these weights do not account for the differences in size between clusters.

There might be instances of small cell populations for which no cells are observed in some samples or where the number of cells is very low. For clusters absent from a sample (e.g., due to biological variance or insufficient sampling), NAs are introduced because no median expression can be calculated; in the case of few cells, the median may be quite variable. Thus, we apply a filter to remove samples that have fewer than 5 cells. We also remove cases where marker expression is equal to zero in all the samples, as this leads to an error during model fitting.

```
## Get median marker expression per sample and cluster
expr_median_sample_cluster_tbl <- data.frame(expr[, functional_markers],
  sample_id = sample_ids, cluster = cell_clustering1m) %>%
  group_by(sample_id, cluster) %>%
  summarize_each(funs(median))

## Melt
expr_median_sample_cluster_melt <- melt(expr_median_sample_cluster_tbl,
  id.vars = c("sample_id", "cluster"), value.name = "median_expression",
  variable.name = "antigen")

## Rearrange so the rows represent clusters and markers
expr_median_sample_cluster <- dcast(expr_median_sample_cluster_melt,
  cluster ~ antigen ~ sample_id, value.var = "median_expression")
rownames(expr_median_sample_cluster) <- paste0(expr_median_sample_cluster$cluster,
```

```

  " ", expr_median_sample_cluster$antigen)
## Eliminate clusters with low frequency
clusters_keep <- names(which((rowSums(counts < 5) == 0)))
keepLF <- expr_median_sample_cluster$cluster %in% clusters_keep
expr_median_sample_cluster <- expr_median_sample_cluster[keepLF, ]
## Eliminate cases with zero expression in all samples
keep0 <- rowSums(expr_median_sample_cluster[, md$sample_id]) > 0
expr_median_sample_cluster <- expr_median_sample_cluster[keep0, ]

```

It is helpful to plot the median expression of all the markers in each cluster for each sample colored by condition to get a rough image of how strong the differences might be. We do this by combining boxplots and jitter.

```

ggdf <- expr_median_sample_cluster_melt[expr_median_sample_cluster_melt$cluster
  %in% clusters_keep, ]
## Add info about samples
mm <- match(ggdf$sample_id, md$sample_id)
ggdf$condition <- factor(md$condition[mm])
ggdf$patient_id <- factor(md$patient_id[mm])
ggplot(ggdf) +
  geom_boxplot(aes(x = antigen, y = median_expression,
    color = condition, fill = condition),
    position = position_dodge(), alpha = 0.5, outlier.color = NA) +
  geom_point(aes(x = antigen, y = median_expression, color = condition,
    shape = patient_id), alpha = 0.8, position = position_jitterdodge(),
    size = 0.7) +
  facet_wrap(~ cluster, scales = "free_y", ncol=2) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  scale_color_manual(values = color_conditions) +
  scale_fill_manual(values = color_conditions) +
  scale_shape_manual(values = c(16, 17, 8, 3, 12, 0, 1, 2))

```

We created a wrapper function `differential_expression_wrapper` that performs the differential analysis of marker expression. The user needs to specify a data frame `expr_median` with marker expression, where each column corresponds to a sample and each row to a cluster/marker combination. One can choose between fitting a regular linear model `model = "lm"` or a linear mixed model `model = "lmer"`. The `formula` parameter must be adjusted adequately to the model choice. The wrapper function returns the non-adjusted and adjusted p-values for each of the specified contrasts K for each cluster/marker combination.

```

differential_expression_wrapper <- function(expr_median, md, model = "lmer", formula, K){
  ## Fit LMM or LM for each marker separately
  fit_gaussian <- lapply(1:nrow(expr_median), function(i){
    data_tmp <- data.frame(y = as.numeric(expr_median[i, md$sample_id]), md)
    switch(model,
      lmer = {
        fit_tmp <- lmer(formula, data = data_tmp)
      },
      lm = {
        fit_tmp <- lm(formula, data = data_tmp)
      })
    ## Fit contrasts one by one
    out <- apply(K, 1, function(k){
      contr_tmp <- glht(fit_tmp, linfct = matrix(k, 1))
      summ_tmp <- summary(contr_tmp)
      pval <- summ_tmp$test$pvalues
      return(pval)
    })
    return(out)
  })
  pvals <- do.call(rbind, fit_gaussian)
  colnames(pvals) <- paste0("pval_", contrast_names)
  rownames(pvals) <- rownames(expr_median)
  ## Adjust the p-values
  adjp <- apply(pvals, 2, p.adjust, method = "BH")
}

```

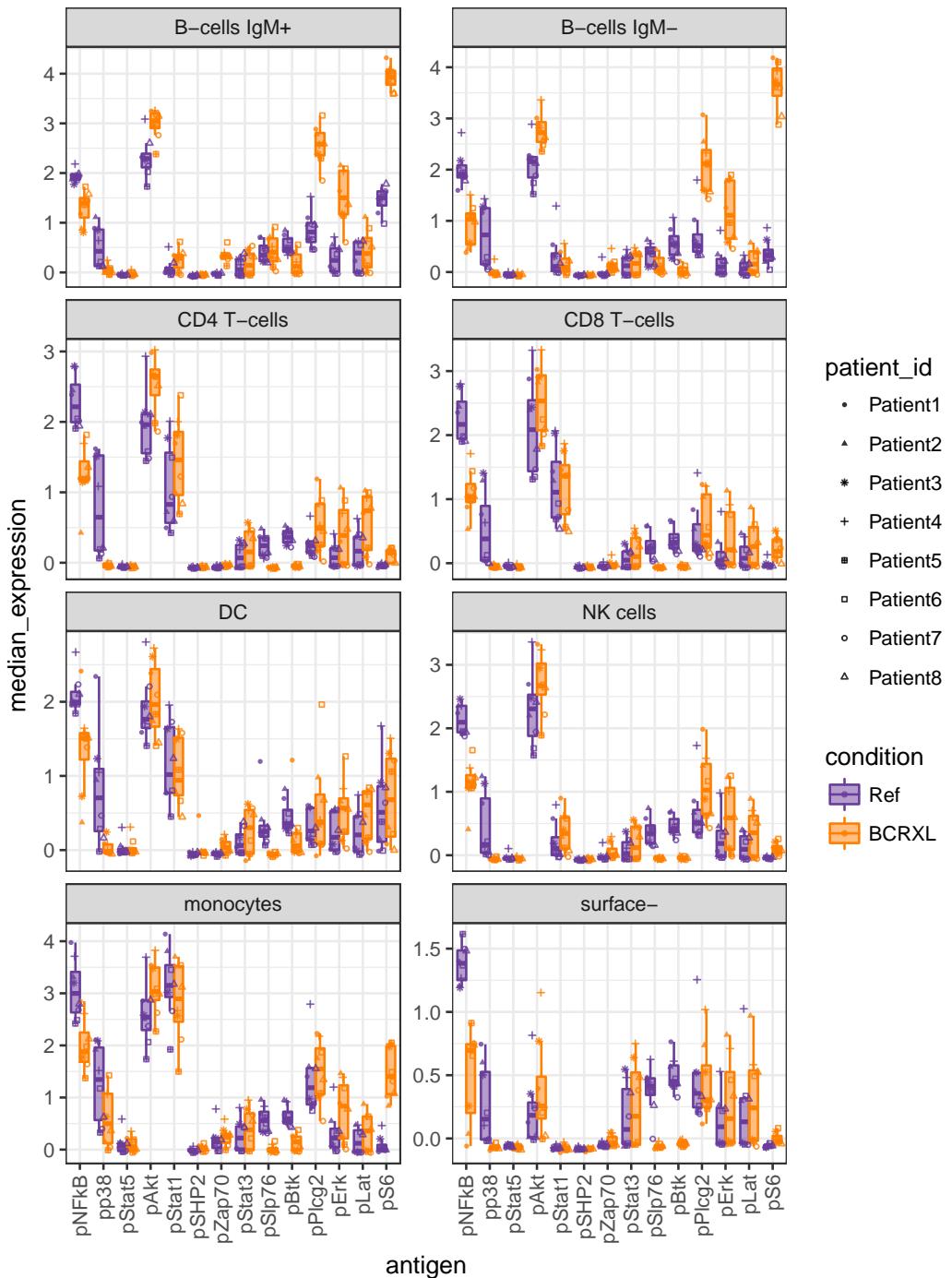


Figure 26. Median expression of 14 signaling markers in the identified PBMC cell populations.

```

  colnames(adjp) <- paste0("adjp_", contrast_names)
  return(list(pvals = pvals, adjp = adjp))
}

```

To present how accounting for the within patient variability with the mixed model increases the sensitivity, we also fit a regular linear model. The linear mixed model has a random intercept for each patient.

```

formula_lm <- y ~ condition
formula_lmer <- y ~ condition + (1|patient_id)

```

By accounting for the patient effect, we detect almost twice as many cases of differential signaling compared to the regular linear model.

```

de_out1 <- differential_expression_wrapper(expr_median = expr_median_sample_cluster,
  md = md, model = "lm", formula = formula_lm, K = K)
apply(de_out1$adjp <- FDR_cutoff, 2, table)

##           adjp_BCRXLvsRef
## FALSE                  51
## TRUE                   42

de_out2 <- differential_expression_wrapper(expr_median = expr_median_sample_cluster,
  md = md, model = "lmer", formula = formula_lmer, K = K)
apply(de_out2$adjp <- FDR_cutoff, 2, table)

##           adjp_BCRXLvsRef
## FALSE                  23
## TRUE                   70

```

One can assemble together an output table with the information about median marker expression in each cluster and sample and the obtained p-values.

```

de_output2 <- data.frame(expr_median_sample_cluster,
  de_out2$pvals, de_out2$adjp, row.names = NULL)
print(head(de_output2), digits = 2)

##           cluster antigen BCRXL1 BCRXL2 BCRXL3 BCRXL4 BCRXL5 BCRXL6 BCRXL7
## 1 B-cells IgM+   pNFkB  1.179  0.880  0.808  1.47  1.361  1.725  1.436
## 2 B-cells IgM+   pp38  0.109 -0.012  0.044  0.24 -0.046  0.083 -0.039
## 3 B-cells IgM+   pAkt  3.247  2.960  2.951  3.26  2.382  3.184  2.762
## 4 B-cells IgM+   pStat1 0.343  0.126  0.242  0.33 -0.010  0.616 -0.050
## 5 B-cells IgM+   pZap70 0.317  0.287  0.351  0.40  0.132  0.604  0.267
## 6 B-cells IgM+   pStat3 -0.047 -0.059  0.451  0.35 -0.058 -0.026  0.534
##           BCRXL8 Ref1 Ref2 Ref3 Ref4 Ref5 Ref6 Ref7 Ref8
## 1      1.5747 1.9639 1.869 1.7726 2.1833 1.861 1.953 1.915 1.979
## 2     -0.0055 0.8891 1.113 0.8534 0.6424 0.126 0.210 0.128 0.126
## 3      3.1439 2.3195 2.310 2.2688 3.0858 1.729 2.024 2.145 2.603
## 4      0.3795 -0.0058 0.064 0.0079 0.5151 -0.047 0.030 -0.034 0.191
## 5      0.3202 -0.0198 -0.033 -0.0336 -0.0056 -0.061 -0.060 -0.032 -0.017
## 6      0.3092 -0.0479 -0.082 0.2652 0.1567 -0.060 -0.066 0.275 0.381
##           pval_BCRXLvsRef adjp_BCRXLvsRef
## 1          6.1e-11        2.7e-10
## 2          7.5e-04        1.6e-03
## 3          2.6e-11        1.3e-10
## 4          6.2e-02        7.5e-02
## 5          1.6e-14        1.0e-13
## 6          5.6e-02        7.1e-02

```

To report the significant results, we use a heatmap. Instead of plotting the absolute expression, we display the normalized expression, which better highlights the direction of marker changes. Additionally, we order the cluster-marker instances by their significance and group them by cell type (cluster).

```

## Keep the significant markers, sort them by significance and group by cluster
sign_clusters_markers <- names(which(de_out2$adjp[, "adjp_BCRXLvsRef"] < FDR_cutoff))
oo <- order(expr_median_sample_cluster[sign_clusters_markers, "cluster"],
  de_out2$adjp[sign_clusters_markers, "adjp_BCRXLvsRef"])
sign_clusters_markers <- sign_clusters_markers[oo]

## Get the significant adjusted p-values
sign_adjp <- de_out2$adjp[sign_clusters_markers , "adjp_BCRXLvsRef"]

## Normalize expression to mean = 0 and sd = 1
expr_s <- expr_median_sample_cluster[sign_clusters_markers,md$sample_id]
expr_median_sample_cluster_norm <- normalization_wrapper(expr_s)

mm <- match(colnames(expr_median_sample_cluster_norm), md$sample_id)
plot_differential_heatmap_wrapper(expr_norm = expr_median_sample_cluster_norm,
  sign_adjp = sign_adjp, condition = md$condition[mm],
  color_conditions = color_conditions)

```

DA of the overall marker expression

The analysis of *overall* expression is analogous to the previous section, except that median marker expression is aggregated from all the cells in a given sample.

```

ggdf <- melt(data.frame(expr_median_sample[functional_markers, ],
  antigen = functional_markers), id.vars = "antigen",
  value.name = "median_expression", variable.name = "sample_id")
## Add condition info
mm <- match(ggdf$sample_id, md$sample_id)
ggdf$condition <- factor(md$condition[mm])
ggdf$patient_id <- factor(md$patient_id[mm])
ggplot(ggdf) +
  geom_boxplot(aes(x = condition, y = median_expression, color = condition,
    fill = condition), position = position_dodge(), alpha = 0.5,
    outlier.color = NA) +
  geom_point(aes(x = condition, y = median_expression, color = condition,
    shape = patient_id), alpha = 0.8, position = position_jitterdodge()) +
  facet_wrap(~ antigen, scales = "free", nrow = 4) +
  theme_bw() +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
  scale_color_manual(values = color_conditions) +
  scale_fill_manual(values = color_conditions) +
  scale_shape_manual(values = c(16, 17, 8, 3, 12, 0, 1, 2))

```

Similar to the analysis above, we identify more markers being differentially expressed with the LMM, which accounts for the within patient variability.

```

## Fit a linear model
de_out3 <- differential_expression_wrapper(expr_median =
  expr_median_sample[functional_markers, ],
  md = md, model = "lm", formula = formula_lm, K = K)
apply(de_out3$adjp < FDR_cutoff, 2, table)

##      adjp_BCRXLvsRef
## FALSE          9
## TRUE          5

## Fit a linear mixed model with patient ID as a random effect
de_out4 <- differential_expression_wrapper(expr_median =
  expr_median_sample[functional_markers, ],
  md = md, model = "lmer", formula = formula_lmer, K = K)
apply(de_out4$adjp < FDR_cutoff, 2, table)

##      adjp_BCRXLvsRef
## FALSE          3
## TRUE          11

```

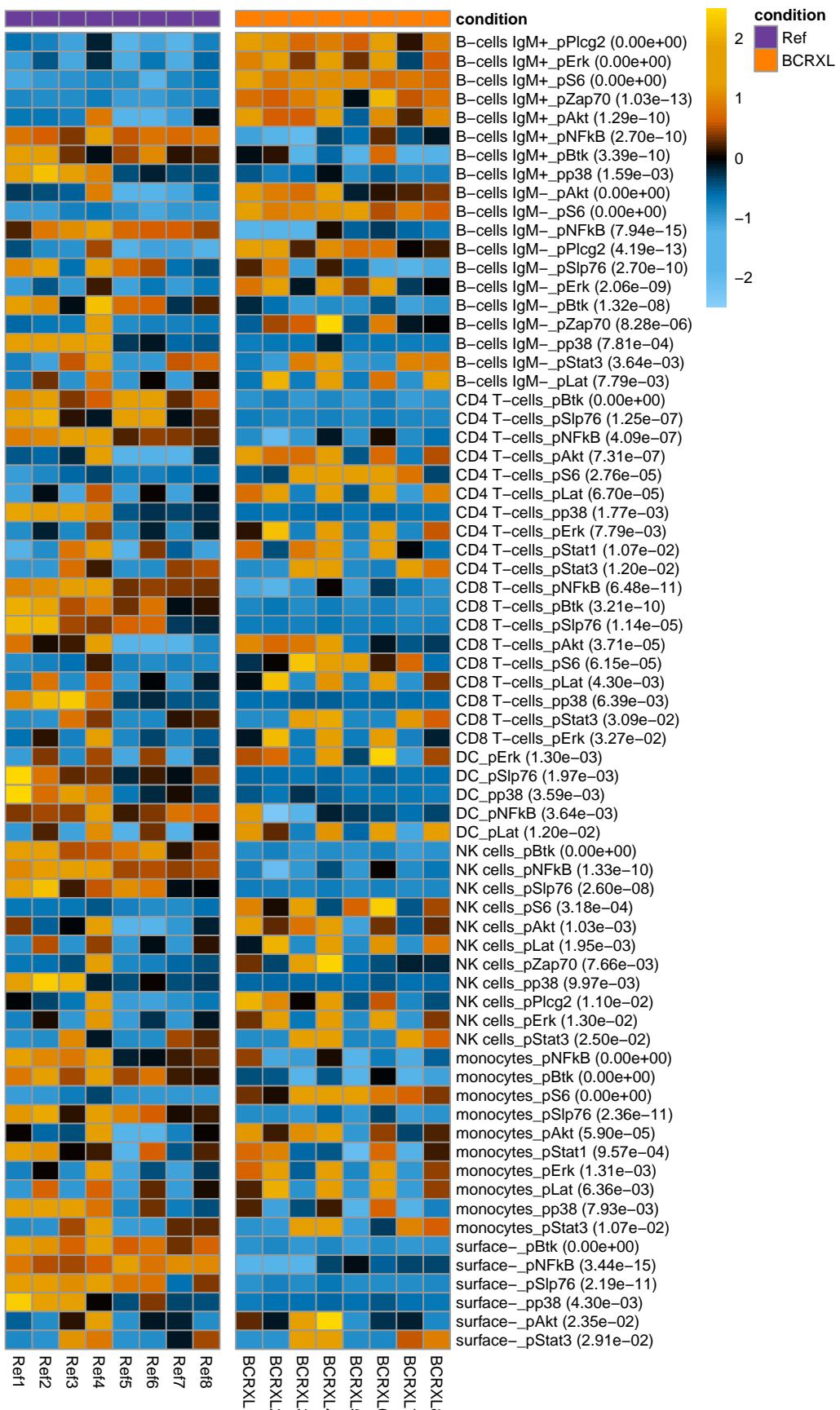


Figure 27. Normalized expression of signaling markers that are significantly differentially expressed between BCR/FcR-XL stimulated and unstimulated condition.

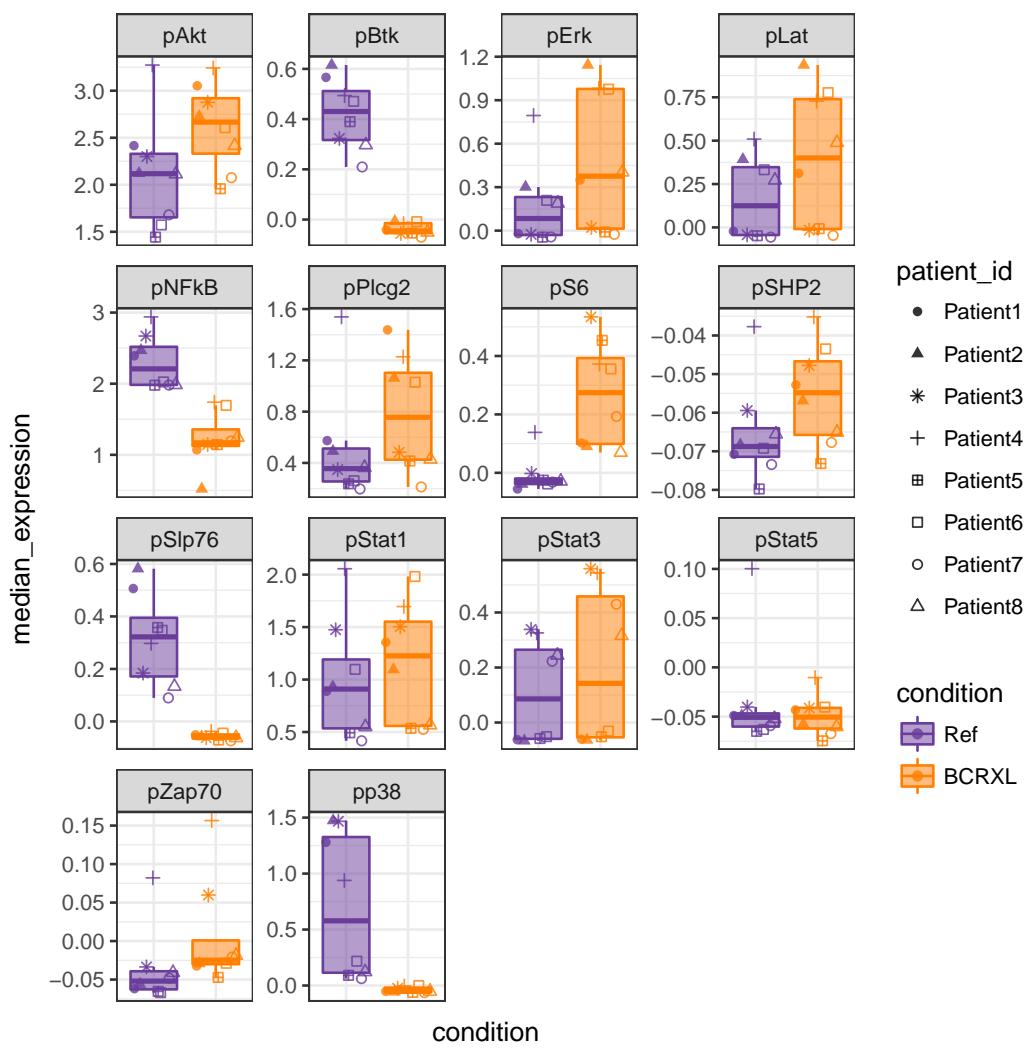


Figure 28. Median expression of 14 signaling markers calculated from all the cells in a given sample.

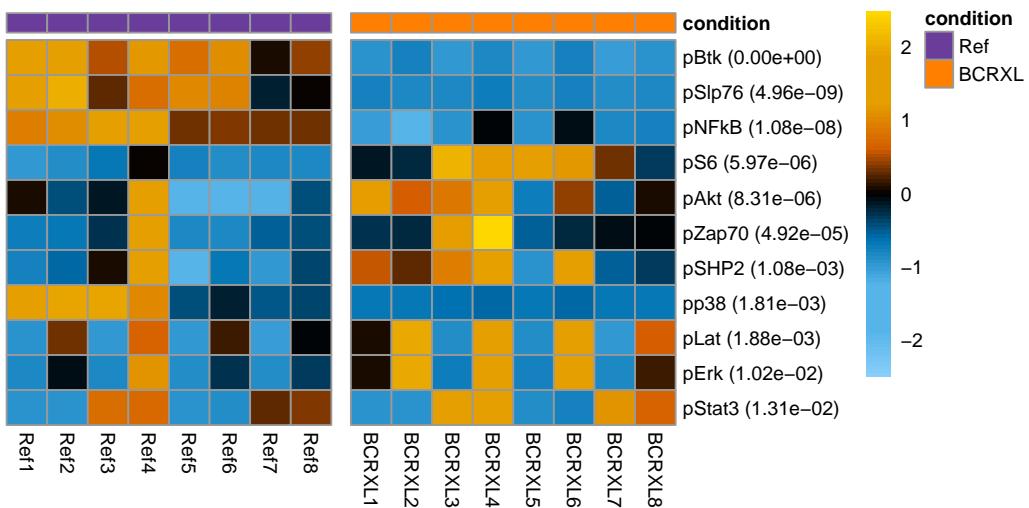


Figure 29. Normalized expression for signaling markers that are significantly differentially expressed between BCR/FcR-XL stimulated and unstimulated condition.

As before, we create an output table with the median marker expression calculated in each sample and the p-values, and we plot a heatmap with the significant markers sorted by their statistical significance.

```
de_output4 <- data.frame(antigen = functional_markers,
  expr_median_sample[functional_markers, ], de_out4$pvals, de_out4$adjp)
print(head(de_output4), digits=2)

##          antigen BCRXL1 BCRXL2 BCRXL3 BCRXL4 BCRXL5 BCRXL6 BCRXL7 BCRXL8
## pNFkB      pNFkB  1.070  0.520  1.144  1.7397  1.143  1.6951  1.195  1.245
## pp38       pp38  -0.052  -0.057  -0.024  -0.0034  -0.061  -0.0006  -0.064  -0.053
## pStat5     pStat5  -0.043  -0.058  -0.041  -0.0103  -0.075  -0.0404  -0.067  -0.060
## pAkt        pAkt   3.053  2.727  2.876  3.2424  1.958  2.6068  2.075  2.416
## pStat1     pStat1  1.356  1.096  1.504  1.6960  0.535  1.9823  0.526  0.566
## pSHP2      pSHP2  -0.053  -0.057  -0.048  -0.0352  -0.073  -0.0435  -0.068  -0.065
##           Ref1   Ref2   Ref3   Ref4   Ref5   Ref6   Ref7   Ref8
## pNFkB      2.392  2.469  2.670  2.940  1.979  2.025  1.980  1.985
## pp38       1.280  1.474  1.467  0.939  0.091  0.218  0.062  0.122
## pStat5    -0.049  -0.049  -0.040  0.100  -0.065  -0.063  -0.059  -0.052
## pAkt        2.416  2.122  2.300  3.273  1.443  1.573  1.680  2.114
## pStat1     0.889  0.930  1.474  2.056  0.493  1.097  0.416  0.549
## pSHP2     -0.071  -0.068  -0.059  -0.038  -0.080  -0.069  -0.073  -0.066
##           pval_BCRXLvsRef adjp_BCRXLvsRef
## pNFkB      2.3e-09      1.1e-08
## pp38       1.0e-03      1.8e-03
## pStat5     3.0e-01      3.0e-01
## pAkt       3.0e-06      8.3e-06
## pStat1     1.9e-01      2.1e-01
## pSHP2      5.4e-04      1.1e-03

## Keep the significant markers and sort them by significance
sign_markers <- names(which(sort(de_out4$adjp[, "adjp_BCRXLvsRef"]) < FDR_cutoff))
## Get the adjusted p-values
sign_adjp <- de_out4$adjp[sign_markers , "adjp_BCRXLvsRef"]
## Normalize expression to mean = 0 and sd = 1
expr_median_sample_norm <- normalization_wrapper(expr_median_sample[sign_markers, ])

mm <- match(colnames(expr_median_sample_norm), md$sample_id)
plot_differential_heatmap_wrapper(expr_norm = expr_median_sample_norm,
  sign_adjp = sign_adjp, condition = md$condition[mm],
  color_conditions = color_conditions)
```

Discussion

In this workflow, we have presented a pipeline for diverse differential analyses of HDCyto datasets. First, we highlight quality control steps where aggregate characteristics of the samples are visualized (e.g., an MDS plot), allowing verification of the experimental design, detection of batch effects and outlying samples. Next, cell population identification was carried out via clustering, which forms the basis for subsequent differential analyses of cell population abundance, differential marker expression within a population or overall marker expression differences. The approaches to differential analyses proposed here are very general and thus able to model complex experimental designs via design matrices, such as factorial experiments, paired experiments or adjustment for batch effects. We have presented a range of visualizations that help in understanding the data and reporting the results of clustering and differential analyses. The wrapper functions presented in this workflow may need to be tailored to the needs of a different experiment.

Clustering is one of the most challenging steps in the workflow and its accuracy is critical to the downstream differential analyses. In particular, getting the right resolution of clusters is crucial since there can be situations where a biologically meaningful cell population may be differentially enriched between conditions, but in an automatic clustering, was combined with another cell population that behaves differently. While we have a good understanding of how computational algorithms recapitulate manual gating in high dimensions (Weber and Robinson 2016), one of the open areas of research remains how to best cluster across samples. For example, a recent approach uses a combination of high dimensional density estimation, hierarchical clustering and network inference and comparison to extract clusters across samples, with a possibility to handle batch effects (Y. H. Li et al. 2017). In our approach, we aggregated all cells together before clustering. An alternative would be to cluster within each sample and then aggregate a collection of metaclusters across samples. Further research is required to better understand these effects, especially when batch effects are present.

The data analyzed here was generated using sample barcoding; this strategy reduces intersample variability, since all samples are exposed to the same antibody cocktail and measured in a single acquisition (Zunder et al. 2015). Thus, the range of marker expression for each channel should, in principle, be within a similar range across samples. Certainly, additional challenges may arise when combining data from different instrument acquisitions and additional preprocessing treatments may need to be applied. Despite adjustments through bead-based normalization (Finck et al. 2013), the observed marker expression may be affected by the varying efficiency of antibody binding in each batch and by the ion detection sensitivity after machine calibration. Beyond normalization, other strategies have been proposed, such as equalizing the dynamic range between batches for each marker or the use of warping functions to eliminate non-linear distortions (see *cydar* vignette). However, a comprehensive evaluation of these approaches and their effect on downstream analyses is still missing. Overall, we expect that as a general rule, including batch parameters (or other covariates) in the linear modeling largely mitigates the problem.

We presented a classical statistical approach where preprocessing of the HDCyto data leads to tables of summaries (e.g., cell counts) or aggregated measurements (e.g., cluster-specific signal) for each samples, which become the input to statistical model. Of course, there are a variety of alternative computational approaches available to the user. We have mentioned *Citrus* and *CellCnn*, which are both machine-learning approaches that fit a reverse model to ours (i.e., phenotype of interest as the response variable). Neither of these approaches are directly able to account for batch effects or complicated designs. However, they may have advantages in the search for rare distinguishing populations, which could be used together with our framework for formal statistical testing.

We have shown that some level of over-clustering is convenient for detecting meaningful cell populations, with the background that automatic detection of the number of natural clusters is difficult (Weber and Robinson 2016). However, there are tradeoffs between the resolution of clustering and the labor involved in aggregating them to biologically meaningful clusters. Overall, we take an interactive but flexible algorithm-guided approach together with subject-area experts to arrive at sensible cell populations. In particular, we rely on various visualizations, such as dendograms, t-SNE maps or other dimension reduction techniques to guide the process. However, alternative strategies could be combined with the statistical inference we present, such as over-clustering combined with data-driven aggregation to the optimal resolution.

One of the main goals of this workflow was to highlight how a model-based approach that is able to handle complex experimental designs. This becomes important in many experimental situations where covariates (e.g., age, gender, batch) may affect the observed HDCyto data. Thus, the classical regression framework allows also to flexibly test situations well beyond two-group differences. Of course, alternatives exist for two group comparisons, such as the nonparametric Mann-Whitney-Wilcoxon test (Hartmann et al. 2016), which makes no assumptions about normality of the data, or the Student's t-test (Pejosi et al. 2016) and its variations, such as the paired t-test.

We note that the the LM, LMM and GLMM, may perform poorly for extremely small samples. Solutions similar to those widely accepted in transcriptomics that share information over variance parameters (M. D. Robinson and Smyth 2007; Love, Huber, and Anders 2014; Ritchie et al. 2015) could be leveraged. An example of such an approach is *cydar*, which performs the differential abundance analysis (on hypersphere counts) using the generalized linear modeling capabilities of *edgeR* (McCarthy, Chen, and Smyth 2012).

As noted, the approach presented in this workflow is not fully automated due to the cluster merging and annotating as well as extensive exploratory data analysis steps. In general, our philosophy is that fully automated

analyses are to be avoided, but rather a battery of diagnostic checks can be designed, as we have promoted here. Cluster annotation remains a manual step in many other approaches as well. Recently, a tool was proposed for consistent characterization of cell subsets using marker enrichment modeling (MEM) (Diggins et al. 2017).

To keep the analysis of this workflow reproducible, one needs to define a random seed before running *FlowSOM* and t-SNE. It is especially important in the clustering step, where the order of clusters may change with different seeds; the cluster merging needs to be match to the see used.

Workflow availability

Code to perform this analysis is available in the Bioconductor workflow package cytofWorkflow from [LINK]. The corresponding manuscript is available on the F1000 website [LINK].

Author contributions

MN and MDR designed and ran analyses. MN drafted the manuscript with input from MDR. CK and SG performed experiments, analyzed data and gave feedback on clinical applications. CK performed the cluster merging. FJH and LMW contributed code and ideas for the analyses. MPL interpreted data and provided clinical perspective. BB gave feedback on the manuscript and the bioinformatics. All authors read and approved the final manuscript and have agreed to the content.

Competing interests

No competing interests were disclosed.

Grant information

MN acknowledges the funding from a Swiss Institute of Bioinformatics (SIB) Fellowship.

Acknowledgments

The authors wish to thank members of the Robinson, Bodenmiller and von Mering groups for helpful discussions.

Session Information

```
sessionInfo()

## R version 3.4.0 (2017-04-21)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.4
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## locale:
## [1] C
##
## attached base packages:
## [1] stats      graphics   grDevices  utils      datasets   methods    base
##
## other attached packages:
## [1] multcomp_1.4-6          TH.data_1.0-8
## [3] MASS_7.3-47             survival_2.41-3
## [5] mvtnorm_1.0-6           lme4_1.1-13
## [7] Matrix_1.2-9            cowplot_0.7.0
## [9] Rtsne_0.13               ConsensusClusterPlus_1.40.0
## [11] FlowSOM_1.8.0           igraph_1.0.1
## [13] pheatmap_1.0.8          RColorBrewer_1.1-2
## [15] ggrepel_0.6.5           limma_3.32.2
## [17] dplyr_0.5.0              reshape2_1.4.2
## [19] ggplot2_2.2.1            matrixStats_0.52.2
## [21] flowCore_1.42.0          readxl_1.0.0
```

```

## [23] BiocStyle_2.4.0           knitr_1.15.1
##
## loaded via a namespace (and not attached):
## [1] Rcpp_0.12.10            lattice_0.20-35
## [3] corpcor_1.6.9           zoo_1.8-0
## [5] assertthat_0.2.0        rprojroot_1.2
## [7] digest_0.6.12          R6_2.2.0
## [9] cellranger_1.1.0       plyr_1.8.4
## [11] backports_1.0.5        stats4_3.4.0
## [13] pcaPP_1.9-61          evaluate_0.10
## [15] highr_0.6              httr_1.2.1
## [17] lazyeval_0.2.0         minqa_1.2.4
## [19] nloptr_1.0.4           rmarkdown_1.5
## [21] labeling_0.3            splines_3.4.0
## [23] stringr_1.2.0          munsell_0.4.3
## [25] compiler_3.4.0         BiocGenerics_0.22.0
## [27] htmltools_0.3.6        tibble_1.3.0
## [29] bookdown_0.3            codetools_0.2-15
## [31] XML_3.98-1.7           rrcov_1.4-3
## [33] grid_3.4.0              nlme_3.1-131
## [35] tsne_0.1-3              gtable_0.2.0
## [37] DBI_0.6-1               magrittr_1.5
## [39] scales_0.4.1           BiocWorkflowTools_1.2.0
## [41] graph_1.54.0            stringi_1.1.5
## [43] robustbase_0.92-7       sandwich_2.3-4
## [45] tools_3.4.0             Biobase_2.36.2
## [47] DEoptimR_1.0-8          parallel_3.4.0
## [49] yaml_2.1.14             colorspace_1.3-2
## [51] cluster_2.0.6

```

References

- Aghaeepour, Nima, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. 2013. “Critical assessment of automated flow cytometry data analysis techniques.” *Nat Meth* 10 (3). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 228–38. <http://dx.doi.org/10.1038/nmeth.2365> <http://www.nature.com/nmeth/journal/v10/n3/abs/nmeth.2365.html{\#}supplementary-information>.
- Angerer, Philipp, Laleh Haghverdi, Maren Büttner, Fabian J Theis, Carsten Marr, and Florian Buettner. 2016. “destiny: diffusion maps for large-scale single-cell data in R.” *Bioinformatics* 32 (8): 1241–3. doi:10.1093/bioinformatics/btv715.
- Arvaniti, Eirini, and Manfred Claassen. 2016. “Sensitive detection of rare disease-associated cell subsets via representation learning.” *bioRxiv*, March. <http://biorkxiv.org/content/early/2016/03/31/046508.abstract>.
- Bendall, Sean C, Kara L Davis, El-Ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe’er. 2014. “Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.” *Cell* 157 (3). Elsevier: 714–25. doi:10.1016/j.cell.2014.04.005.
- Bendall, Sean C, Erin F Simonds, Peng Qiu, El-ad D Amir, Peter O Krutzik, Rachel Finck, Robert V Bruggner, et al. 2011. “Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum.” *Science* 332 (6030). American Association for the Advancement of Science: 687–96. doi:10.1126/science.1198704.
- Bodenmiller, Bernd, Eli R Zunder, Rachel Finck, Tiffany J Chen, Erica S Savig, Robert V Bruggner, Erin F Simonds, et al. 2012. “Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators.” *Nature Biotechnology* 30 (9). Nature Publishing Group: 858–67. doi:10.1038/nbt.2317.
- Bruggner, Robert V, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. 2014. “Automated identification of stratifying signatures in cellular subpopulations.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (26): E2770–7. doi:10.1073/pnas.1408792111.
- Chen, Hao, Mai Chan Lau, Michael Thomas Wong, Evan W Newell, Michael Poidinger, and Jinmiao Chen. 2016. “Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline.” *PLOS Computational Biology* 12 (9). Public Library of Science: 1–17. doi:10.1371/journal.pcbi.1005112.
- Diggins, Kirsten E, Allison R Greenplate, Nalin Leelatian, Cara E Wogsland, and Jonathan M Irish. 2017. “Characterizing cell subsets using marker enrichment modeling.” *Nat Meth* 14 (3). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 275–78. <http://dx.doi.org/10.1038/nmeth.4149> <http://www.nature.com/nmeth/journal/v14/n3/abs/nmeth.4149.html{\#}supplementary-information>

information.

- Finck, Rachel, Erin F Simonds, Astraea Jager, Smita Krishnaswamy, Karen Sachs, Wendy Fantl, Dana Pe'er, Garry P Nolan, and Sean C Bendall. 2013. "Normalization of mass cytometry data with bead standards." *Cytometry Part A* 83A: 483–94. doi:10.1002/cyto.a.22271.
- Haghverdi, L., F. Buettner, and F. J. Theis. 2015. "Diffusion maps for high-dimensional single-cell analysis of differentiation data." *Bioinformatics* 31 (May): 2989–98. doi:10.1093/bioinformatics/btv325.
- Hartmann, Felix J, Raphaël Bernard-Valnet, Clémence Quériault, Dunja Mrdjen, Lukas M Weber, Edoardo Galli, Carsten Krieg, et al. 2016. "High-dimensional single-cell analysis reveals the immune signature of narcolepsy." *Journal of Experimental Medicine* 213 (12). Rockefeller University Press: 2621–33. doi:10.1084/jem.20160897.
- Jia, Cheng, Yu Hu, Yichuan Liu, and Mingyao Li. 2014. "Mapping Splicing Quantitative Trait Loci in RNA-Seq." *Cancer Informatics* 13: 35–43. doi:10.4137/CIN.S13971.Received.
- Kotecha, Nikesh, Peter O Krutzik, and Jonathan M Irish. 2001. "Web-Based Analysis and Publication of Flow Cytometry Experiments." In *Current Protocols in Cytometry*. John Wiley & Sons, Inc. doi:10.1002/0471142956.cy1017s53.
- Leipold, Michael D. 2015. "Another step on the path to mass cytometry standardization." *Cytometry Part A* 87 (5): 380–82. doi:10.1002/cyto.a.22661.
- Levine, Jacob H., Erin F Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, et al. 2015. "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis." *Cell* 162 (1). Elsevier: 184–97. doi:10.1016/j.cell.2015.05.047.
- Li, Ye Henry, Dangna Li, Nikolay Samusik, Xiaowei Wang, Leying Guan, Garry P Nolan, and Wing Hung Wong. 2017. "Scalable Multi-Sample Single-Cell Data Analysis by Partition-Assisted Clustering and Multiple Alignments of Networks." *bioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/116566.
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology* 15 (12). BioMed Central: 550. <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>.
- Mahnke, Yolanda D, and Mario Roederer. 2007. "Optimizing a Multicolor Immunophenotyping Assay." *Clinics in Laboratory Medicine* 27 (3): 469–85. doi:<http://doi.org/10.1016/j.cll.2007.05.002>.
- McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research* 40 (10): 4288–97.
- Monti, Stefano, Pablo Tamayo, Jill Mesirov, and Todd Golub. 2003. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data." *Machine Learning* 52 (1): 91–118. doi:10.1023/A:1023949509487.
- Pejoski, David, Nicolas Tchitchek, André Rodriguez Pozo, Jamila Elhmouzi-Younes, Rahima Yousfi-Bogniaho, Christine Rogez-Kreuz, Pascal Clayette, et al. 2016. "Identification of Vaccine-Altered Circulating B Cell Phenotypes Using Mass Cytometry and a Two-Step Clustering Analysis." *The Journal of Immunology* 196 (11). American Association of Immunologists: 4814–31. doi:10.4049/jimmunol.1502005.
- Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research* 43 (7). Oxford University Press: e47.
- Robinson, Mark D., and Gordon K. Smyth. 2007. "Moderated statistical tests for assessing differences in tag abundance." *Bioinformatics* 23 (21): 2881–7.
- Roederer, Mario. 2001. "Spectral compensation for flow cytometry: Visualization artifacts, limitations, and caveats." *Cytometry* 45 (3). John Wiley & Sons, Inc.: 194–205. doi:10.1002/1097-0320(20011101)45:3<194::AID-CYTO1163>3.0.CO;2-C.
- Saeys, Yvan, Sofie Van Gassen, and Bart N Lambrecht. 2016. "Computational flow cytometry: helping to make sense of high-dimensional immunology data." *Nat Rev Immunol* 16 (7). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 449–62. <http://dx.doi.org/10.1038/nri.2016.56> <http://10.0.4.14/nri.2016.56>.
- Unen, Vincent van, Na Li, Ilse Molendijk, Mine Temurhan, Thomas Höllt, Andrea E van der Meulen-de Jong, Hein W Verspaget, et al. 2016. "Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets." *Immunity* 44 (5): 1227–39. doi:<http://dx.doi.org/10.1016/j.immuni.2016.04.014>.
- Van Der Maaten, L J P, and G E Hinton. 2008. "Visualizing high-dimensional data using t-sne." *Journal of Machine Learning Research*. doi:10.1007/s10479-011-0841-3.
- Van Gassen, Sofie, Britt Callebaut, Mary J Van Helden, Bart N Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. 2015. "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data." *Cytometry. Part A : The Journal of the International Society for Analytical Cytology* 87 (7): 636–45. doi:10.1002/cyto.a.22625.
- Wang, Bo, Daniele Ramazzotti, Luca De Sano, Junjie Zhu, Emma Pierson, and Serafim Batzoglou. 2017. "SIMLR: A Tool for Large-Scale Single-Cell Analysis by Multi-Kernel Learning." *bioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/118901.
- Wattenberg, Martin, Fernanda Viégas, and Ian Johnson. 2016. "How to Use t-SNE Effectively." *Distill*.

doi:10.23915/distill.00002.

Weber, Lukas M, and Mark D Robinson. 2016. "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data." *Cytometry Part A* 89 (12): 1084–96. doi:10.1002/cyto.a.23030.

Wilkerson, Matthew D, and D Neil Hayes. 2010. "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking." *Bioinformatics* 26 (12): 1572. doi:10.1093/bioinformatics/btq170.

Zhao, Keyan, Zhi-Xiang Lu, Juw Won Park, Qing Zhou, and Yi Xing. 2013. "GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data." *Genome Biology* 14 (7). BioMed Central Ltd: R74. <http://www.ncbi.nlm.nih.gov/pubmed/23876401>.

Zunder, Eli R, Rachel Finck, Gregory K Behbehani, El-ad D Amir, Smita Krishnaswamy, Veronica D Gonzalez, Cynthia G Lorang, et al. 2015. "Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm." *Nature Protocols* 10 (2). Nature Publishing Group: 316–33. doi:10.1038/nprot.2015.020.

Collaboration Papers

T-cell acute leukaemia exhibits dynamic interactions with bone marrow microenvironments

Edwin D. Hawkins, Delfim Duarte, Olufolake Akinduro, Reema A. Khorshed, Diana Passaro, Malgorzata Nowicka, Lenny Straszkowski, Mark K. Scott, Steve Rothery, Nicola Ruivo, Katie Foster, Michaela Waibel, Ricky W. Johnstone, Simon J. Harrison, David A. Westerman, Hang Quach, John Gribben, Mark D. Robinson, Louise E. Purton, Dominique Bonnet and Cristina Lo Celso

Paper published in *Nature* (2016), 538(7626), 518–522

T-cell acute leukaemia exhibits dynamic interactions with bone marrow microenvironments

Edwin D. Hawkins^{1,2,3*}, Delfim Duarte^{1,4*}, Olufolake Akinduro¹, Reema A. Khorshed¹, Diana Passaro⁵, Małgorzata Nowicka⁶, Lenny Straszkowski⁷, Mark K. Scott^{1,2}, Steve Rothery⁸, Nicola Ruivo¹, Katie Foster⁵, Michaela Waibel^{9,10}, Ricky W. Johnstone^{9,10}, Simon J. Harrison^{9,10}, David A. Westerman^{9,10}, Hang Quach^{11,12}, John Gribben¹³, Mark D. Robinson⁶, Louise E. Purton^{7,12}, Dominique Bonnet⁵ & Cristina Lo Celso^{1,4}

It is widely accepted that complex interactions between cancer cells and their surrounding microenvironment contribute to disease development, chemo-resistance and disease relapse. In light of this observed interdependency, novel therapeutic interventions that target specific cancer stroma cell lineages and their interactions are being sought. Here we studied a mouse model of human T-cell acute lymphoblastic leukaemia (T-ALL) and used intravital microscopy to monitor the progression of disease within the bone marrow at both the tissue-wide and single-cell level over time, from bone marrow seeding to development/selection of chemo-resistance. We observed highly dynamic cellular interactions and promiscuous distribution of leukaemia cells that migrated across the bone marrow, without showing any preferential association with bone marrow sub-compartments. Unexpectedly, this behaviour was maintained throughout disease development, from the earliest bone marrow seeding to response and resistance to chemotherapy. Our results reveal that T-ALL cells do not depend on specific bone marrow microenvironments for propagation of disease, nor for the selection of chemo-resistant clones, suggesting that a stochastic mechanism underlies these processes. Yet, although T-ALL infiltration and progression are independent of the stroma, accumulated disease burden leads to rapid, selective remodelling of the endosteal space, resulting in a complete loss of mature osteoblastic cells while perivascular cells are maintained. This outcome leads to a shift in the balance of endogenous bone marrow stroma, towards a composition associated with less efficient haematopoietic stem cell function¹. This novel, dynamic analysis of T-ALL interactions with the bone marrow microenvironment *in vivo*, supported by evidence from human T-ALL samples, highlights that future therapeutic interventions should target the migration and promiscuous interactions of cancer cells with the surrounding microenvironment, rather than specific bone marrow stroma, to combat the invasion by and survival of chemo-resistant T-ALL cells.

The importance of cancer cell interactions with their surrounding microenvironment has gained increased attention due to the hypothesis that specific supportive cells may regulate quiescence, survival and self-renewal of cancer cells themselves. This relationship may underlie a critical mechanism that facilitates both the initiation of disease and chemo-resistance. Leukaemia develops within the bone marrow (BM), where it has been suggested to take part in complex crosstalk that in some cases results in microenvironment remodelling^{2–8}. Therapeutic targeting of leukaemia-supportive niches has been proposed^{9,10},

therefore it is critical that we understand both the spatial and kinetic nature of leukaemia–BM interactions. However, our current knowledge of leukaemia biology is predominantly derived from *ex vivo* flow cytometric analysis, and static images that cannot capture information on the location and dynamics of leukaemia interactions with BM structures and cells over time.

We studied a Notch-driven mouse model of T-ALL, which recapitulates human disease both phenotypically (Extended Data Fig. 1) and genetically^{11,12}. Twenty-five per cent of paediatric and 40% of adult T-ALL patients develop aggressive relapsed disease originating from chemo-resistant clones¹³. Thus, there is a pressing need to understand if T-ALL cells migrate to, and interact with particular BM stroma during the propagation of disease and/or selection of chemo-resistance, or if T-ALL can remodel the BM microenvironment in its favour.

To address these questions, we monitored leukaemia growth in mouse calvarium bone marrow by intravital microscopy^{14–16}. We used a tile-based imaging approach that allows tissue-wide visualization of heterogeneous BM microenvironments (Fig. 1a, b) while maintaining a resolution that permits the measurement of single leukaemia cell interactions with BM cells and structures by time-lapse microscopy¹⁵ (Fig. 1c and Supplementary Video 1). To characterize T-ALL interactions systematically *in vivo*, we injected transformed leukaemic cells isolated from primary hosts into secondary recipients, and observed highly synchronous disease progression (Fig. 1d and Extended Data Fig. 1). In secondary recipients, T-ALL preferentially infiltrated the BM before expansion to peripheral lymphoid organs (Extended Data Fig. 1). This is consistent with expression of CXCR4, albeit at variable levels, on leukaemic cells (Extended Data Fig. 1). We visualized T-ALL cells relative to osteoblasts (green fluorescent protein (GFP) or cyan fluorescent protein (CFP) positive in Col2.3–GFP/CFP reporter mice^{17,18}), perivascular mesenchymal stem/progenitor cells (GFP positive in nestin–GFP reporter mice¹⁹) and vasculature (by injecting Cy5 dextran) in cohorts of recipient mice during disease progression and treatment (Fig. 1d).

By day 10 after transplantation we could reproducibly observe single, sparse T-ALL cells in the BM at a frequency of 1–30 cells per calvarium (Fig. 1e, f). We measured the proximity of leukaemia cells to osteoblastic and nestin–GFP⁺ cells and vasculature. We used randomly positioned dots as a control for the specificity of observed associations, as these do not possess any inherent ability to localize to a particular BM stroma component (Fig. 1g). The distribution of T-ALL cells was equivalent to that of the random dots and the actual distances

¹Department of Life Sciences, Sir Alexander Fleming Building, Imperial College London, London SW7 2AZ, UK. ²The Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria 3052, Australia. ³Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia. ⁴The Francis Crick Institute, 1 Midland Road, London NW1A 1AT, UK. ⁵The Francis Crick Institute, Haematopoietic Stem Cell Laboratory, 1 Midland Road, London NW1A 1AT, UK. ⁶SIB Swiss Institute of Bioinformatics and Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. ⁷Stem Cell Regulation Unit, St Vincent's Institute of Medical Research, 41 Victoria Parade Fitzroy, Victoria 3065 Australia. ⁸Imperial College Facility for Imaging by Light Microscopy, Sir Alexander Fleming Building, Imperial College London, London SW7 2AZ, UK. ⁹Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Victoria 3052, Australia. ¹⁰Peter MacCallum Cancer Centre, Department of Haematology, Melbourne, Victoria 3000, Australia. ¹¹Department of Haematology, St Vincent's Hospital, Fitzroy, Victoria 3065, Australia. ¹²Department of Medicine, The University of Melbourne, Fitzroy, Victoria 3065, Australia. ¹³Centre of Haemat-Oncology, Cancer Research UK Clinical Centre, Barts Cancer Institute, St Bartholomew's Hospital, Queen Mary University of London, London EC1M 6BQ, UK.

*These authors contributed equally to this work.

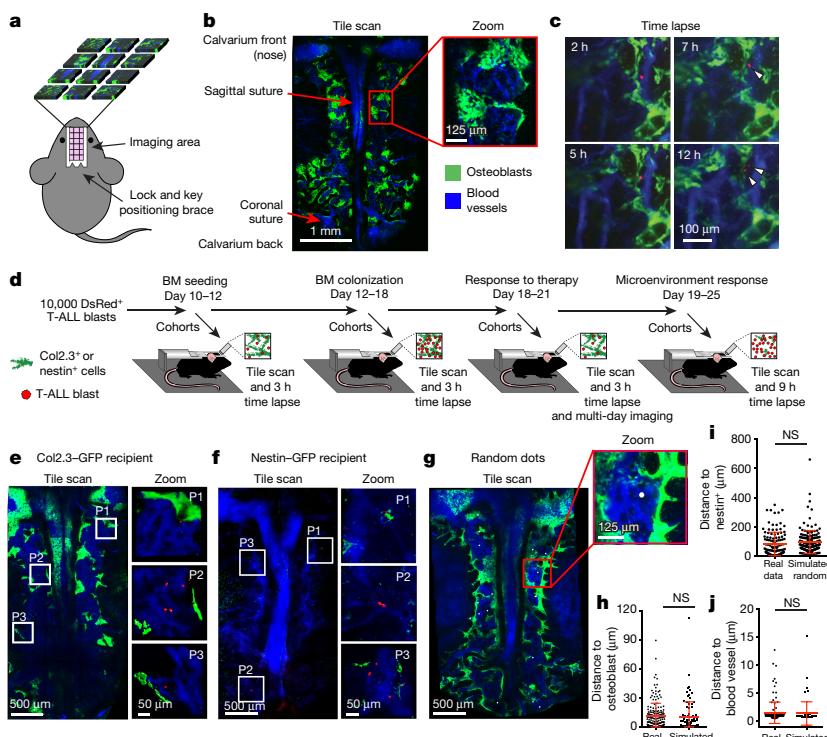


Figure 1 | Experimental set up and T-ALL BM seeding. **a**, Image data sets were formed of multiple, overlapping z-stacks covering the entire calvarium BM space. **b**, Tile scans preserve single-cell resolution. **c**, Long-term single-cell time-lapse microscopy (14 h). Arrows indicate division and daughter cells. **d**, Intra-vital imaging schedule **e**, **f**, Representative maximum projection tile scans showing T-ALL distribution in Col2.3-GFP (**e**) and nestin-GFP (**f**) recipient mice calvarium bone marrow, and corresponding high-magnification three-dimensional renders. P, position. **g**, Simulated cells (white) were randomly distributed within BM space, for control positional measurements. **h–j**, T-ALL cell location relative to osteoblasts (**h**), nestin cells (**i**) and blood vessels (**j**) compared with randomly positioned dots overlaid on tile scans. Red: T-ALL cells; green: osteoblasts/nestin cells; blue: vasculature. $n = 190, 117, 135$ cells and $91, 168, 70$ random dots, respectively in **h**, **i**, **j**; data are representative/pooled from seven (**e**, **f**, **h**, **i**) and four (**j**) independent mice (biological replicates) injected with cells from two independent primary donors. Error bars: mean \pm standard deviation (s.d.). NS, not significant.

recorded inversely correlated with the abundance of each component (Fig. 1h–j). These results demonstrate that seeding T-ALL in the BM is stochastically distributed relative to osteoblasts, nestin-GFP⁺ cells and vasculature.

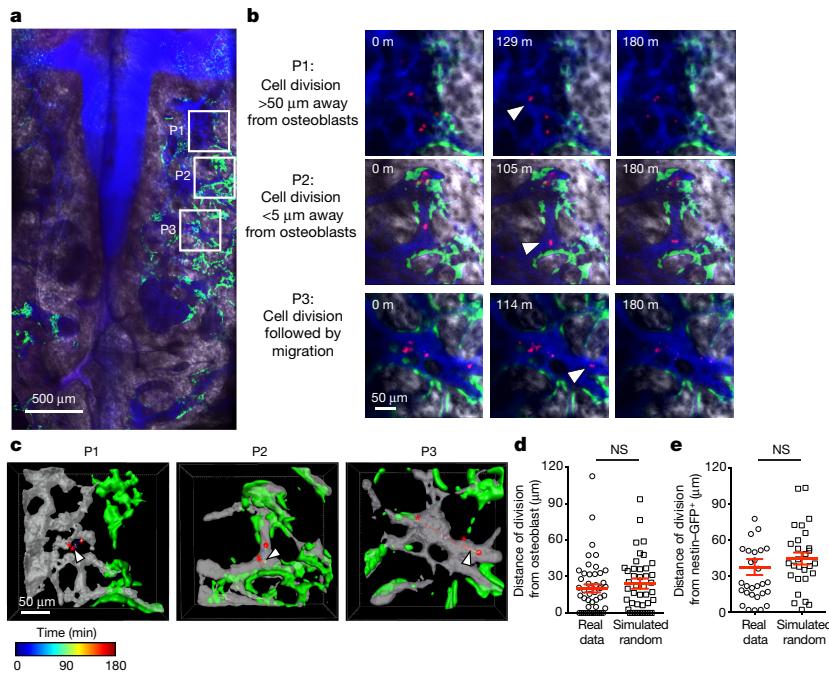
To determine whether T-ALL expansion was supported by specific constituents of the BM, we monitored the dynamics of single T-ALL cells (Fig. 2a) for 3 h (Fig. 2b, c, Extended Data Fig. 2 and Supplementary Videos 2, 3). This revealed that the vast majority of T-ALL cells were motile, in stark contrast with previous observations of transplanted haematopoietic stem cells in BM¹⁵, and that movement was rarely restricted to the proximity of any specific cell types or structures (Fig. 2b, c, position 2 and 3, and Supplementary Videos 2, 3). Notably, the speed of any given cell over time was also heterogeneous, and thus no single migratory behaviour was associated with osteoblastic, nestin-GFP⁺ cells or vasculature (Fig. 2c, Extended Data Fig. 2 and Supplementary Videos 2, 3). Tracking single T-ALL cells enabled us to measure the location of mitotic events, revealing the same stochastic distribution of dividing cells and suggesting that proximity to these stroma components is not key for T-ALL expansion (Fig. 2d, e, Extended Data Fig. 2 and Supplementary Videos 2, 3). Additionally, daughter cells migrated large distances after division, illustrating that clones and their progeny are not restricted to foci within the local microenvironment (Fig. 2e and Supplementary Video 2, position 3). These observations were consistent with tile scans performed at later stages of BM colonization (day 12–18 after transplantation), where we detected pockets of high and low infiltration juxtaposed in all types of microenvironments (Extended Data Figs 3 and 4). Interestingly, the motility patterns displayed by single, isolated cells were also consistently observed when tracking individual cells located in densely infiltrated areas (Supplementary Video 4). Combined, these analyses demonstrate that T-ALL seeding and colonization of BM do not select for, or depend on specific BM stroma.

The question remained whether certain BM regions could create ‘hotspots’ for chemo-resistance through provision of a protective

environment. To address this issue, we adapted our imaging protocol to follow the same BM areas over multiple days to track leukaemia dynamics from complete BM infiltration and throughout therapy. We used the lock-and-key mechanism of the imaging window to re-position mice precisely on the microscope stage over multiple imaging sessions. Mice were tile scanned 18 days after T-ALL transplantation to confirm full BM infiltration. Dexamethasone was then administered daily (Fig. 3a), and immediately after the third therapy dose we observed a staggering reduction in disease burden²⁰ (Fig. 3b–d and Extended Data Fig. 5a, b), while non-leukaemic, dexamethasone-treated control mice maintained robust BM and stroma cellularity (Extended Data Fig. 5d, e). Strikingly, surviving cells were scattered throughout the BM space and not preferentially associated with osteoblastic, nestin-GFP⁺ cells or vasculature compared with simulated data (Fig. 3f and Extended Data Fig. 5a–c).

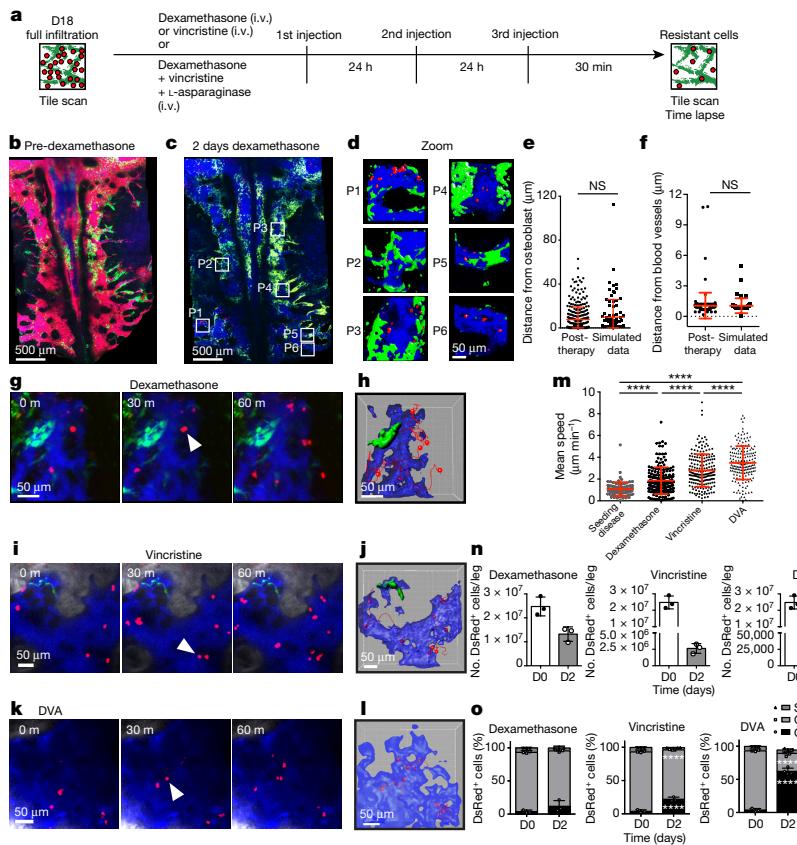
To test whether initial T-ALL loss was independent from specific stroma components, we increased the imaging temporal resolution to include the first day of treatment. The distribution of T-ALL cells was maintained as disease gradually succumbed to therapy (Extended Data Fig. 6). In contrast to predictions based on previous publications^{3,9,10}, time-lapse imaging immediately after administration of the third dose of dexamethasone demonstrated that single surviving cells were highly migratory and did not maintain long-lasting associations with osteoblastic (Fig. 3g, h and Supplementary Video 5) or nestin-GFP⁺ cells (Supplementary Video 6). Furthermore, surviving cells travelled at significantly faster speeds than early infiltrating cells (days 10–15) (Fig. 3m). Finally, residual T-ALL cells were still capable of undergoing division (Fig. 3g and Supplementary Videos 5, 6) at times when other cells within the same mouse were still undergoing death (Supplementary Video 6). This behaviour, coupled with the observation that mice maintained on dexamethasone for a 7-day period harboured an almost completely repopulated BM space (Extended Data Fig. 7a), suggests that these cells were genuinely resistant to dexamethasone.

Collectively, these data contradict the prevailing hypothesis that therapy-resistant leukaemic cells depend on a particular microenvironment



for survival. To investigate this proposed paradigm, we compared the gene expression profile of T-ALL cells purified at full infiltration to those isolated from mice 7–10 days after initiation of dexamethasone

treatment, when surviving cells have re-colonized the BM (Extended Data Fig. 7a). Gene expression profiles of all T-ALL samples were more heterogeneous compared with control T cells (CD4⁺ and CD8⁺),



T-cell progenitor ($CD4^+CD8^+$ thymocytes) and whole BM populations, as expected on the basis of inter-clone variation in leukaemia cells²¹. The transcriptome of resistant cells overlapped with that of T-ALL cells pre-treatment (Extended Data Fig. 7c). Indeed, only 79 genes were differentially expressed in the post-treatment group and, consistent with our intravital imaging data, none of the differentially expressed genes were related to known cell–niche interaction candidates (Supplementary Table 1 and Extended Data Fig. 7b). Together, our imaging and gene expression data suggest that dexamethasone treatment does not select for a subpopulation of T-ALL cells that have been directed to a specific niche.

To investigate the importance of these data in a human context, we assessed the migratory behaviour of xenotransplanted primary human T-ALL (Supplementary Table 2) in NOD/SCID/ γ mice. After dexamethasone treatment, even human wild-type NOTCH receptor T-ALL cells exhibited migratory behaviour equivalent to murine cells (Extended Data Fig. 8a–d and Supplementary Video 7). We noted an even distribution of quiescent cells throughout the BM as measured by Ki-67 staining (Extended Data Fig. 8e, f). To test whether the migratory behaviour was a specific trait induced by dexamethasone, we treated murine T-ALL burdened animals with vincristine. The T-ALL response followed similar kinetics to that of dexamethasone-treated animals. Enhanced migration of resistant cells and cell division events were evident immediately after the third dose of vincristine (Fig. 3i, j, m and Supplementary Video 8). This behaviour was even more pronounced when we combined dexamethasone, vincristine and L-asparaginase (Fig. 3k–m, Extended Data Fig. 9 and Supplementary Video 9). In this context, we observed the highest migration speed of all therapies tested, despite over 50% of surviving cells being in G0 (Fig. 3n, o), demonstrating that quiescence is not a feature of niche-restricted immotile cells after chemotherapy. These findings were universal in all therapies tested in our studies, suggesting that migration and lack of long-lasting interactions with the surrounding microenvironment are conserved features of chemo-resistance in T-ALL.

To assay whether T-ALL may affect BM structures, we performed time-lapse imaging in heavily disease-burdened, untreated mice. After day 18 we observed striking remodelling of osteoblasts (Fig. 4a, b). A 9 h period of time-lapse imaging after full infiltration (day 19 onwards) revealed that osteoblastic cells underwent dramatic shrinking and blebbing (Fig. 4a and Supplementary Video 10), in stark contrast to the early stages of disease (Supplementary Video 1) and to control Col2.3-GFP mice reconstituted with red fluorescent healthy BM, where osteoblastic cells were unaffected (Fig. 4a, b, Extended Data Fig. 10a, b and Supplementary Video 10). TdT-mediated dUTP nick end labelling (TUNEL) and cleaved caspase-3 histological stainings indicated that osteoblasts were undergoing apoptosis (Fig. 4c). The loss of osteoblasts was so dramatic that virtually no GFP⁺ osteoblasts remained apparent by days 22–25 after transplantation (Fig. 4b, d). Similar results were also found for osterix⁺ osteo-progenitors (Fig. 4f); however, nestin-GFP⁺ cells were maintained (Fig. 4e and Extended Data Fig. 10c, d), and blood vessels could still be visualized (Extended Data Fig. 10e). Importantly, these findings were consistent in analyses performed using human T-ALL samples (Supplementary Table 2). We observed a significant reduction in osterix⁺ osteo-progenitors in NOD/SCID/ γ recipient mice xenotransplanted with human T-ALL samples (Fig. 4f). Additionally, histological analysis of BM trephine biopsies from T-ALL patients with greater than 75% infiltration demonstrated an almost complete loss of osteoblasts (Fig. 4g).

We demonstrate that T-ALL infiltrates the BM and survives chemotherapy independently of stable interactions with specific microenvironments. However, T-ALL has a profound effect on osteoblastic cells. Our results suggest that to avoid the development of chemo-resistance, novel therapeutic interventions may not need to target specific BM stroma components, but rather the ability of

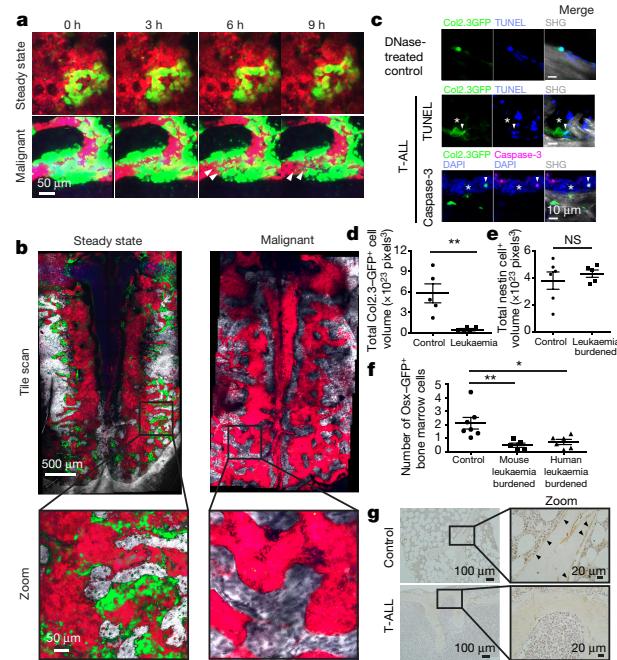


Figure 4 | T-ALL rapidly remodels the endosteal niche. **a**, Nine-hour time-lapse of Col2.3-GFP mice calvaria (green: GFP⁺ osteoblasts) transplanted with tomato⁺ bone marrow (red) >8 weeks earlier (top) or DsRed⁺ T-ALL blasts 19 days earlier (bottom). Arrows: osteoblastic membrane blebbing. **b**, Representative tile scans of Col2.3-GFP recipients during steady-state haematopoiesis (left) or in a malignant state (>22 days T-ALL, right). Bottom, high-magnification of boxed areas. Grey: bone; green: osteoblasts; blue: vasculature. Data are representative of five healthy and malignant mice (biological replicates). **c**, Bone sections from Col2.3-GFP mice stained for TUNEL or cleaved caspase-3. DNase pre-treated sections (top) were TUNEL-positive controls. Grey: bone; green: GFP; blue: TUNEL/4',6-diamidino-2-phenylindole (DAPI); purple: cleaved caspase-3; arrows: apoptotic osteoblasts; asterisks: surviving osteoblasts. Representative from three heavily infiltrated mice (biological replicates) injected with T-ALL from two primary donors. **d**, Quantified osteoblast volume from tile scans shown in **b**. **e**, Quantified nestin volume in nestin-GFP⁺ mice 22–25 days after T-ALL transplant. **a**, **b**, **d**, **e**, $n = 6/5$ mice (biological replicates) from three independent experiments and T-ALL from two primary donors. **f**, Osterix (Osx)-GFP⁺ cells quantified by flow cytometry in mice transplanted with murine and human T-ALL primary cells, at high tumour burden. $n = 7$ control mice, 5 mice with murine T-ALL from 2 primary donors and 6 mice with primary human T-ALL from 2 independent donors. **g**, BM trephine biopsies from healthy or T-ALL patients immunostained for osteocalcin (brown). Data shown are representative of three healthy controls and four T-ALL patients (biological replicates) with >75% BM blast infiltration. Error bars: mean \pm s.e.m.

T-ALL cells to interact with and migrate through the BM. Our studies also reveal that T-ALL has the intrinsic potential to remodel the BM microenvironment via apoptosis throughout the osteolineage. As these cells are associated with haematopoietic fitness^{14,22,23}, including in the context of leukaemia^{1,2,24}, the remodelling we uncovered may contribute to the loss of healthy haematopoiesis observed in leukaemia patients^{1,2}. Collectively, our observations suggest that a shift in therapeutic design may be advisable, such that any intervention targeting osteoblastic cells should focus on promoting survival of this lineage, rather than seeking to modulate a direct influence of the BM on the cancer itself. Therefore, a better focus for novel anti-cancer therapeutics may be on disrupting the ability of T-ALL cells to interact transiently with multiple components of the BM microenvironment.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 June 2015; accepted 24 August 2016.

Published online 17 October 2016.

1. Bowers, M. *et al.* Osteoblast ablation reduces normal long-term hematopoietic stem cell self-renewal but accelerates leukemia development. *Blood* **125**, 2678–2688 (2015).
2. Colmone, A. *et al.* Leukemic cells create bone marrow niches that disrupt the behavior of normal hematopoietic progenitor cells. *Science* **322**, 1861–1865 (2008).
3. Duan, C. W. *et al.* Leukemia propagating cells rebuild an evolving niche in response to therapy. *Cancer Cell* **25**, 778–793 (2014).
4. Hanoun, M. *et al.* Acute myelogenous leukemia-induced sympathetic neuropathy promotes malignancy in an altered hematopoietic stem cell niche. *Cell Stem Cell* **15**, 365–375 (2014).
5. Ishikawa, F. *et al.* Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region. *Nature Biotechnol.* **25**, 1315–1321 (2007).
6. Saito, Y. *et al.* Induction of cell cycle entry eliminates human leukemia stem cells in a mouse model of AML. *Nature Biotechnol.* **28**, 275–280 (2010).
7. Schepers, K. *et al.* Myeloproliferative neoplasia remodels the endosteal bone marrow niche into a self-reinforcing leukemic niche. *Cell Stem Cell* **13**, 285–299 (2013).
8. Zhang, B. *et al.* Altered microenvironmental regulation of leukemic and normal stem cells in chronic myelogenous leukemia. *Cancer Cell* **21**, 577–592 (2012).
9. Jin, L. *et al.* CXCR4 up-regulation by imatinib induces chronic myelogenous leukemia (CML) cell migration to bone marrow stroma and promotes survival of quiescent CML cells. *Mol. Cancer Ther.* **7**, 48–58 (2008).
10. Weisberg, E. *et al.* Inhibition of CXCR4 in CML cells disrupts their interaction with the bone marrow microenvironment and sensitizes them to nilotinib. *Leukemia* **26**, 985–990 (2012).
11. Sanda, T. *et al.* Interconnecting molecular pathways in the pathogenesis and drug sensitivity of T-cell acute lymphoblastic leukemia. *Blood* **115**, 1735–1745 (2010).
12. Weng, A. P. *et al.* Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269–271 (2004).
13. Pui, C. H., Robison, L. L. & Look, A. T. Acute lymphoblastic leukaemia. *Lancet* **371**, 1030–1043 (2008).
14. Lo Celso, C. *et al.* Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. *Nature* **457**, 92–96 (2009).
15. Rashidi, N. M. *et al.* *In vivo* time-lapse imaging shows diverse niche engagement by quiescent and naturally activated hematopoietic stem cells. *Blood* **124**, 79–83 (2014).
16. Scott, M. K., Akinduro, O. & Lo Celso, C. *In vivo* 4-dimensional tracking of hematopoietic stem and progenitor cells in adult mouse calvarial bone marrow. *J. Vis. Exp.* **91**, e51683 (2014).
17. Kalajzic, I. *et al.* Use of type I collagen green fluorescent protein transgenes to identify subpopulations of cells at different stages of the osteoblast lineage. *J. Bone Miner. Res.* **17**, 15–25 (2002).
18. Paic, F. *et al.* Identification of differentially expressed genes between osteoblasts and osteocytes. *Bone* **45**, 682–692 (2009).
19. Méndez-Ferrer, S. *et al.* Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* **466**, 829–834 (2010).
20. Inaba, H. & Pui, C. H. Glucocorticoid use in acute lymphoblastic leukaemia. *Lancet Oncol.* **11**, 1096–1106 (2010).
21. Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377–1380 (2008).
22. Lassailly, F., Foster, K., Lopez-Onieva, L., Currie, E. & Bonnet, D. Multimodal imaging reveals structural and functional heterogeneity in different bone marrow compartments: functional implications on hematopoietic stem cells. *Blood* **122**, 1730–1740 (2013).
23. Visnjic, D. *et al.* Hematopoiesis is severely altered in mice with an induced osteoblast deficiency. *Blood* **103**, 3258–3264 (2004).
24. Krevvata, M. *et al.* Inhibition of leukemia cell engraftment and disease progression in mice by osteoblasts. *Blood* **124**, 2834–2846 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements E.D.H. was supported by a fellowship from the European Haematology Association and an RD Wright Career fellowship from the National Health and Medical Research Council of Australia (NHMRC) as well as a project grant from Bloodwise (12033). D.D. was supported by the GABBA PhD program (FCT fellowship SFRH/BD/52195/2013). L.E.P. was supported by a National Health and Medical Research Council of Australia Senior Research Fellowship. H.Q. is supported by Victorian Cancer Agency Early Career Seed Grant. D.B., D.P. and K.F. were funded by Cancer Research UK (CRUK) and the Francis Crick Institute. C.L.C. was funded by Bloodwise (12033), Human Frontier Science Program (RGP0051/2011), CRUK (C36195/A1183), Biotechnology and Biological Sciences Research Council (BB/I004033/1) Kay Kendall Leukaemia Fund (KKL460) and the European Research Council (337066). We are grateful to M. Spitaler, D. Keller and L. Carlin for support with imaging. We also thank S. Piperelis, E. Ibarguen, W. Steel and H. Goyal for logistical help; A. Ivan for support from the genomics facility; J. Srivastava and C. Simpson for support from the flow cytometry facility; M. Wall for her advice on the human cytogenetics; D. Ibanez for access to software; H. Fleming, V. Greco, R. E. Sinden, S. Mostowy and P. O'Donovan for critical feedback on the manuscript.

Author Contributions E.D.H. and C.L.C. conceived the project. E.D.H., C.L.C., O.A. and M.K.S. designed and refined the intravital imaging systems. R.A.K. designed the image analysis platform; R.A.K., E.D.H. and D.D. performed all image analysis. S.R. designed the random data simulator. E.D.H., D.D. and N.R. performed imaging experiments. N.R. maintained all animal lines. M.N. and M.D.R. designed the microarray experiment and analysed the data. E.D.H., M.W. and R.W.J. performed experiments in Extended Data Fig. 1. D.P., K.F. and D.B. designed, performed and analysed osterix-GFP experiments. S.J.H., D.A.W. and J.G. provided human T-ALL samples. L.S. and L.E.P. performed histological analysis of human BM trephines. H.Q. funded human biopsy studies. D.P. and D.B. designed and D.P. performed the human T-ALL xenograft experiments and analysed data. E.D.H., D.D. and C.L.C. analysed data and wrote the manuscript. All authors contributed revisions to the manuscript. Note that O.A., R.A.K. and D.P. contributed equally to this work, and the same is for M.D.R., L.E.P. and D.B.

Author Information Gene expression data have been deposited in ArrayExpress under accession number E-MTAB-4889. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.D.H. (Hawkins.e@wehi.edu.au) or C.L.C. (c.lo-celso@imperial.ac.uk).

Reviewer Information *Nature* thanks T. Look, C. Mullighan, J. van Rheenen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Mice. All animal experimentation in this study was approved and performed according to the standards of the animal ethics committee at Imperial College London and to UK Home Office regulations (ASPA 1986). C57Bl/6 mice were purchased from Harlan UK Ltd; Col2.3-GFP, Col2.3-CFP¹⁸, nestin-GFP and mTmG²⁵ mice were bred and housed at Imperial College London. For imaging experiments, female Col2.3-GFP and nestin-GFP mice >8 weeks old were used. osterix-CreGFP mice were provided by A. McMahon and backcrossed over eight generations into NSG mice and maintained at the Francis Crick Institute, Cancer Research UK²⁶. Equal proportions of male and female osterix-CreGFP mice aged 11–14 weeks were used.

Generation of murine T-ALL disease. T-ALL was generated as previously described²⁷. Briefly, timed matings were established between C57Bl/6 mice and embryos harvested at E14.5. Single-cell suspensions were prepared from whole fetal livers isolated from the embryos. Suspensions were cultured in IL-3, IL-6, and stem cell factor conditioned media with 20% FCS for 3 days. Lin-xE cells were transfected by calcium phosphate with MigR1 plasmids containing either DsRed only or DsRed with NotchICNΔRamΔP as described previously²⁸. We also used GFP-tagged plasmids when required. Supernatants containing recombinant retrovirus were removed and spun by centrifugation onto non-tissue-culture-treated plates coated with 1.5 µg/ml retronectin (Takara Clontech, CA). Fetal liver cells were cultured in the presence of virus for 3 days and transduction was assessed by flow cytometry. Primary lethally irradiated recipient mice (two doses of 5.5 Gy administered greater than three hours apart) were transplanted with 1 × 10⁶ DsRed⁺ fetal liver cells by intravenous injection into the tail vein. Recipient mice were maintained on baytril-treated water to prevent infection for >6 weeks post-transplantation. Cohorts of reconstituted mice were the result of three independent fetal liver isolations and three independent transfections. Transformation of Notch-transduced non-malignant cells is highly heterogeneous *in vivo*, with onset of primary disease ranging from 6–25 weeks²⁸ (Extended Data Fig. 1). More than 4 weeks post-reconstitution, peripheral blood was isolated from mice, red blood cells were lysed, and successful reconstitution determined by presence of DsRed⁺ cells. Mice reconstituted with NotchICNΔRamΔP-transduced fetal livers were monitored daily for signs of leukaemia onset or other signs of ill health. Mice were euthanized when any one or a combination of the following signs were observed: hunched posture, laboured breathing, weight loss, enlarged lymph nodes and/or spleen, peripheral white blood cell cellularity of 13 × 10⁹ per litre or greater. No experiment exceeded the tumour burden approved by the Home Office and Imperial College ethics committee. Peripheral lymphoid organs were analysed by flow cytometry for DsRed or GFP, CXCR4, CD3, CD4 and CD8 expression. All FACS data was collected on a Fortessa flow cytometer (BD Biosciences, CA). Secondary recipients were sub-lethally irradiated (two doses of 3 Gy administered greater than three hours apart) and injected with 10,000 thawed, Ficoll purified T-ALL blasts and monitored as described earlier. In selected cases 10,000 secondary T-ALL cells were transplanted into tertiary recipients. For therapy experiments, mice were injected i.v. daily with 15 mg/kg dexamethasone sodium phosphate^{29,30} (Selleckchem, MA) alone, 0.15 mg/kg vincristine sulfate salt (Sigma) alone or with a combination of 15 mg/kg dexamethasone, 0.15 mg/kg vincristine and 1,000 IU/kg L-asparaginase (medac; obtained from the Imperial College Healthcare NHS Trust Pharmacy).

Bone marrow chimaeras. Reconstitution with MigR1 DsRed-transduced cells yields <50% chimaerism. For this reason, analysis of healthy BM cells by microscopy is inaccurate. Therefore, to obtain >95% chimaerism of healthy, red fluorescent BM to be used for imaging control experiments, whole BM mononuclear cells were isolated from femurs, hips and tibia of mTmG donor mice, suspended in phosphate balanced salt solution and administered intravenously to recipient mice at a dose of 2 × 10⁶ cells/mouse. Recipient Col2.3-GFP or C57Bl/6 mice had been lethally irradiated (two doses of 5.5 Gy irradiation greater than 3 h apart) immediately before the transplant, and were maintained on baytril-treated water to prevent infection >6 weeks post-transplantation.

Intravital microscopy. Intravital microscopy was performed using a Leica SP5 and a Zeiss LSM 780 upright confocal microscope with a motorized stage. The SP5 was fitted with the following lasers: Argon, 546, 633 and a tunable infrared multiphoton laser (Spectraphysics Mai Tai 690–1020). The Zeiss LSM 780 was fitted with the following lasers: Argon, 561, 633 and a tunable infrared multiphoton laser (Spectraphysics Mai Tai DeepSee 690–1040). Signal was visualized with a Leica HCX IRAPO L × 25 water immersion lens (0.95 N.A.) and a W Plan-Apochromat × 20 DIC water immersion lens (1.0 N.A.). Collagen bone second harmonic generation signal and GFP and CFP signals were generated through excitation at 840 and 870 nm and detected with external detectors. Internal detectors were used to collect DsRed and Cy5 signal (and on some occasions, GFP). Prior to surgery, mice were administered analgesia with buprenorphine (0.1 mg/kg intraperitoneally (i.p.)).

Anaesthesia was induced in mice with 4% isoflurane mixed with pure oxygen. This was gradually reduced to approximately 1% as anaesthesia stabilized. Surgery to attach the headpiece was then performed as described previously¹⁶. Large three-dimensional ‘tile scans’ of the entire BM cavity space were acquired by stitching adjacent, high-resolution z-stack images using a surgically implanted imaging window that ensures steady positioning of mice on the microscope. The calvarium has been demonstrated to be equivalent to the long bones such as the femur with regards to haematopoietic stem cell frequency, function and localization^{14,22}, and is the only BM compartment that allows longitudinal imaging through minimally invasive surgery^{16,31}. Blood vessels were highlighted by i.v. injection of 50 µl of 8 mg/ml 500 kDa Cy5-Dextran (Nanocs, MA). Cy5-Dextran was re-injected every 1–2 h to maintain blood vessel signal and cross reference for registration of blood vessel data in time-lapse analysis. For repeated imaging, protective intrasite gel (Smith & Nephew) was applied to the imaging window to preserve the bone integrity and prevent scar formation. The window was bandaged, and mice were allowed to recover from anaesthesia. Owing to the lock-and-key mechanism of the imaging window¹⁶, mice could then be re-anæsthetized and accurately repositioned on the microscope stage and the same BM areas re-imaged. After each imaging, analgesia was administered via oral buprenorphine in raspberry jelly at a dose of approximately 0.8 mg/kg.

Image quantification. Microscopy data was processed using multiple platforms. Tile scans were stitched using Leica Application Systems (LAS; Leica Microsystems, Germany) and ZEN black (Zeiss, Germany) softwares. Raw data were visualized and processed using Fiji/Image J. Simulated data was prepared using FIJI macros to create, and overlay z-stack images on original tile-scan data. Using the internal random number algorithm, spheres matching the size of T-ALL cells (11–15 µm) were placed at random x,y,z coordinates. Simulated data FIJI macro is available on request. Automated cell segmentation, distance and volume measurements were performed in Definiens (Definiens Developer 64, Germany) using local heterogeneity segmentation³² to isolate osteoblast and nestin cells as well as vasculature, and a combination of seed detection algorithm and morphological growing and shrinking operations to detect leukaemia cells. Definiens rulesets for these functions are available upon request. Distance measurements from this segmentation were performed as described previously³². Cell tracking was performed using Imaris (Bitplane, Switzerland) and the FIJI plugin MTrackJ. For accuracy in cell tracking data, videos were registered when required before using four-dimensional data protocols implemented in Fiji³³. Three-dimensional data rendering and measurement of cell division distances were performed in Velocity (Perkin Elmer, MA) and Definiens (Definiens Developer 64, Germany).

Microarrays. T-ALL samples were harvested from bone marrow and FACS sorted based on fluorescent protein expression (DsRed or GFP) unless infiltration of bone marrow was complete. Control samples for microarray were prepared by FACS sorting for splenic CD4⁺ T cells, CD8⁺ T cells and CD4⁺CD8⁺ thymocytes from 8–14 week female C57Bl/6 mice. RNA was purified from samples using the Qiagen RNeasy mini kit (Netherlands) as per the manufacturer's instructions. Purified RNA was prepared for hybridization using the Genechip WT Plus reagent kit (Affymetrix, CA) as per the manufacturer's instructions and hybridized with genechip Mouse gene 2.0 ST array (Affymetrix, CA) by the MRC Genomics Facility (Imperial College London). Analysis was performed using R version 3.1.1. Data were normalized and summarized to ‘core’ level using the RMA method from the oligo package (version 1.30.0)³⁴. Annotation was downloaded from the Affymetrix NetAffx Query website. Differential expression was determined using limma version 3.22.7 (ref. 35). Genes with Benjamini–Hochberg-adjusted P value < 0.05 and absolute log-fold-change >1 were deemed significant. Heatmaps of gene expression were generated with pheatmap package version 1.0.2.

Human T-ALL xenografts. Primary human T-ALL samples were obtained from Barts Hospital (London) after informed consent via a protocol approved by the East London Research Ethics Committee and carried out in accordance with the principles of the Helsinki declaration (see Supplementary Table 2 for details), before treatment being administered to the patients. Primary cells from two distinct patients were immunophenotyped, and CD45⁺/CD7⁺/CD4^{-low}/CD8^{-low} cells sorted and infused i.v. in non-conditioned osterix-CreGFP/NOD/SCID/γ recipient mice. Primary xenograft transplantation was assessed via peripheral blood sampling and/or BM aspiration. BM and spleen-derived primary xenografts were infused i.v. in non-conditioned NOD/SCID/γ secondary recipient mice for therapy experiments. Intravital imaging was performed as described earlier. Human T-ALL cells were labelled by injecting 10 µg of PE-conjugated human CD45 antibody (clone HI30, Biologend) 15–30 min before the imaging session. For dexamethasone therapy experiments, mice were treated with daily injections of 15 mg/kg i.v.³⁰. Number of human T-ALL cells in therapy experiments was quantified using reference beads as described previously³⁶.

Immunofluorescence. Hips and tibias were harvested and post-fixed overnight in periodate-lysine-parafomaldehyde fixative, at 4°C. Bones were then washed with 0.1 M phosphate buffer, cryoprotected in sucrose (10–30% gradient), for 48 h, frozen in optimal cutting temperature compound (TissueTek) and stored at –80°C. Sections were cut in a Leica Cryostat, using the CryoJane tape transfer system (Leica Microsystems) and stored at –80°C. For staining, slides were rehydrated in PBS, permeabilized in 0.1% Triton X-100, blocked in 5% goat serum and incubated with primary antibodies overnight, at 4°C. After washing in PBS, slides were incubated with secondary antibodies, counter-stained with DAPI (Invitrogen), washed in 0.1% Triton X-100 and mounted using Prolong Diamond antifade (Invitrogen). The following antibodies were used: Alexa Fluor 647 mouse anti-Ki-67 (B56, BD Biosciences, 1:50), PE-conjugated human CD45 antibody (HI30, BD Biosciences, 1:100), rabbit anti-cleaved caspase-3 (Asp175, Cell Signaling, 1:100), goat anti-rabbit IgG Alexa Fluor 633 (Life Technologies, 1:400). TUNEL labelling was performed to detect apoptotic cells, according to the manufacturer's instructions (DeadEnd Colorimetric TUNEL System, Promega). Images were obtained using a Zeiss LSM 780 upright confocal/two-photon combined microscope and analysed using Fiji/ImageJ. Cell counting was performed manually using the FIJI plugin Cell Counter.

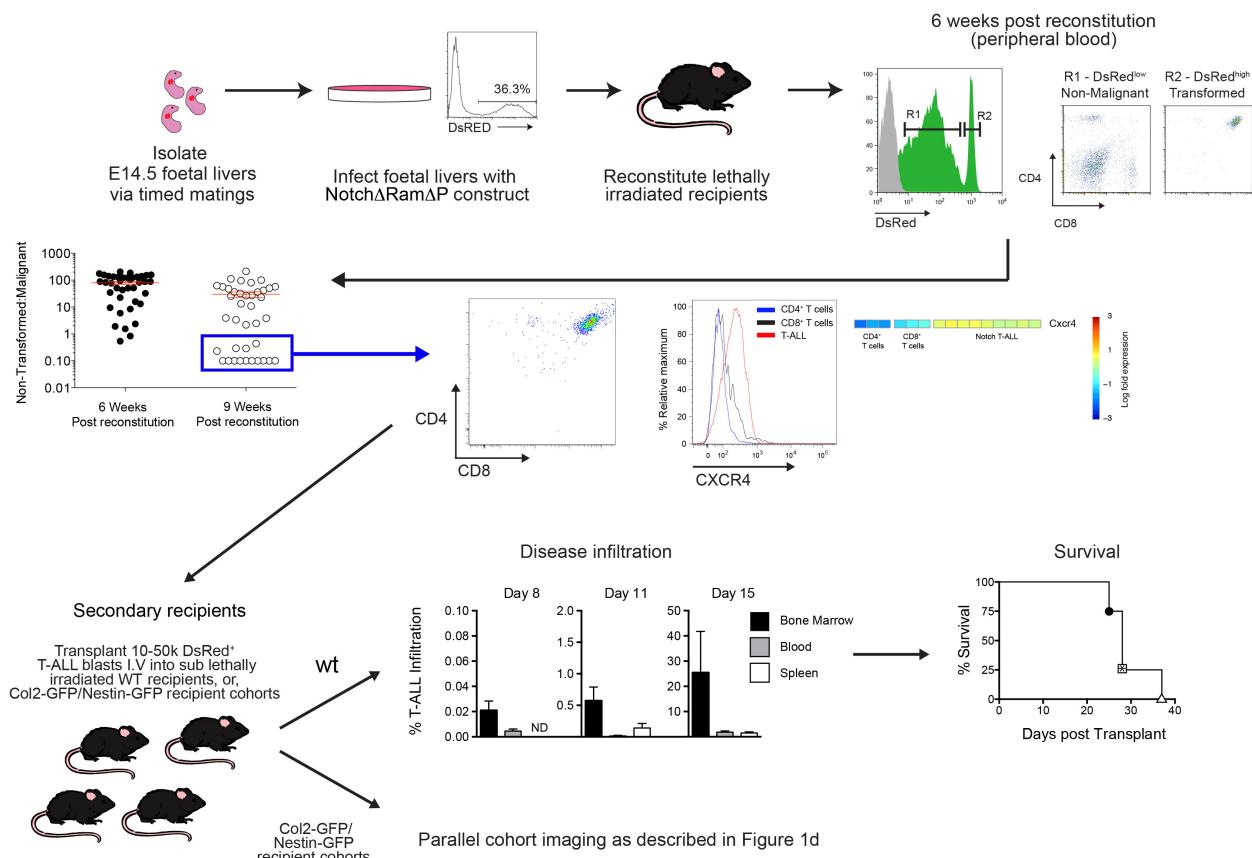
T-ALL cell number and cell cycle analysis. BM from human T-ALL xenotransplanted, untreated and treated mice was harvested and stained with DAPI (Invitrogen) and FITC mouse anti-Ki-67 set (BD Biosciences), according to the manufacturer's instructions. Cells were analysed by flow cytometry and absolute numbers were obtained using reference beads as described previously³⁶.

Osterix quantification. T-ALL engraftment and infiltration was confirmed via peripheral blood sampling and/or tibia puncture. Once mice presented with signs of ill health (as described earlier), mice were euthanized and bones were digested with a DNase I/Collagenase (Sigma) solution. The total number of Osx-GFP⁺ cells was assessed by flow cytometry analysis using counting beads (CountBright, Life Technologies).

Human trephine histology. Samples were obtained from patients after informed consent had been obtained, under full ethical approval by the Peter MacCallum Cancer Centre Human Research Ethics Committee. De-waxed human trephine biopsy sections (3 µM) were stained with osteocalcin antibody (Abcam ab93876, Cambridge), counterstained and mounted for viewing. All areas of each section were monitored for visible osteoblasts.

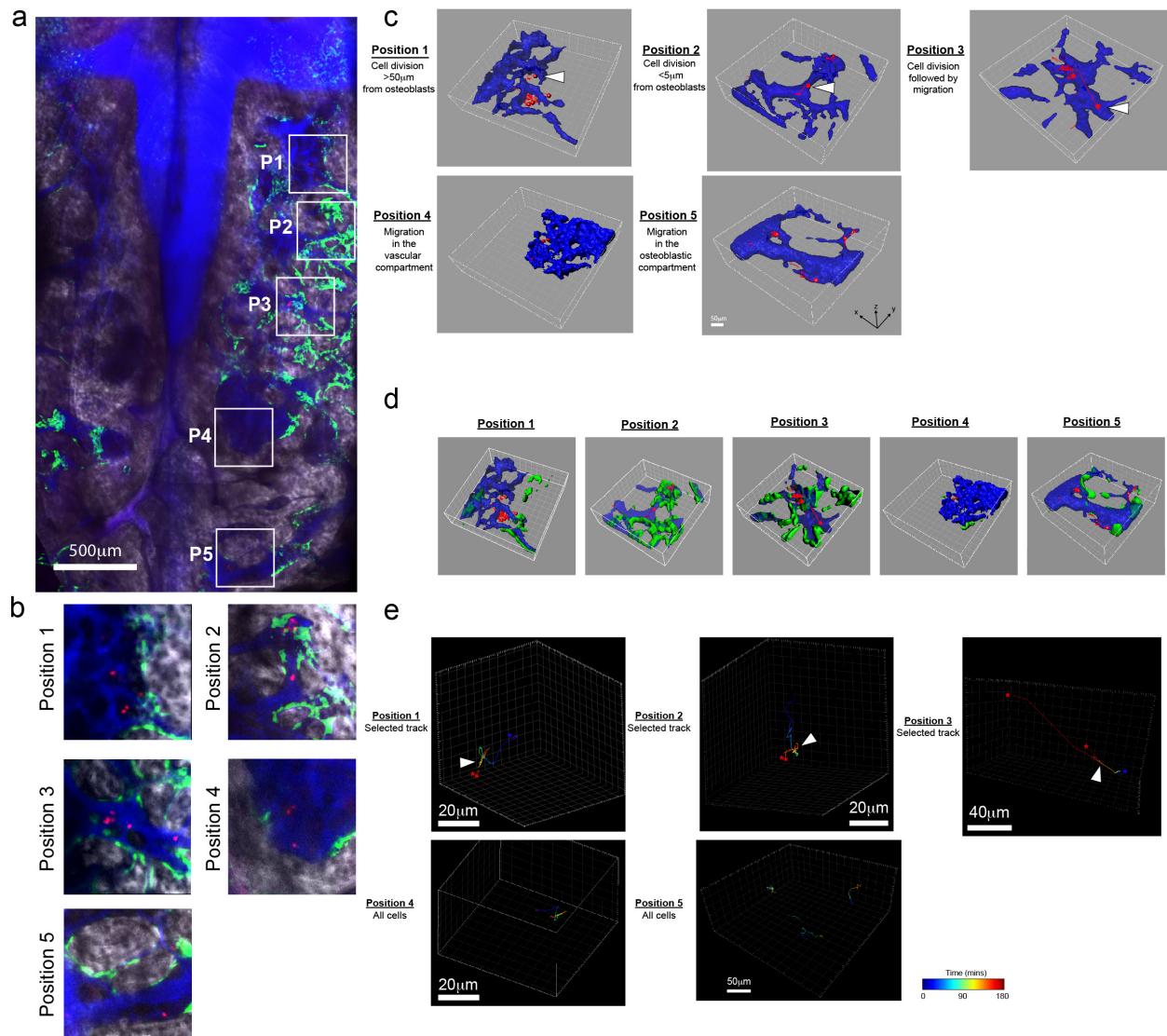
Statistics. The sample size required for the experiments was estimated based on the results of preliminary data. Blinding or randomization for animal experiments were not necessary due to the nature of the experiments. Statistical differences between the means of two data groups was determined by using two-tailed unpaired Student's *t*-test, and *P* values < 0.05 were considered significant. Multiple group comparisons were performed using ANOVA with a Bonferroni correction, *P* values < 0.05 were considered significant.

25. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
26. Rodda, S. J. & McMahon, A. P. Distinct roles for Hedgehog and canonical Wnt signaling in specification, differentiation and maintenance of osteoblast progenitors. *Development* **133**, 3231–3244 (2006).
27. Hawkins, E. D. *et al.* Lethal giant larvae 1 tumour suppressor activity is not conserved in models of mammalian T and B cell leukaemia. *PLoS ONE* **9**, e87376 (2014).
28. Aster, J. C. *et al.* Essential roles for ankyrin repeat and transactivation domains in induction of T-cell leukemia by notch1. *Mol. Cell. Biol.* **20**, 7505–7515 (2000).
29. Real, P. J. *et al.* Gamma-secretase inhibitors reverse glucocorticoid resistance in T cell acute lymphoblastic leukemia. *Nature Med.* **15**, 50–58 (2009).
30. Chiu, P. P., Jiang, H. & Dick, J. E. Leukemia-initiating cells in human T-lymphoblastic leukemia exhibit glucocorticoid resistance. *Blood* **116**, 5268–5279 (2010).
31. Lo Celso, C., Lin, C. P. & Scadden, D. T. In vivo imaging of transplanted hematopoietic stem and progenitor cells in mouse calvarium bone marrow. *Nature Protocols* **6**, 1–14 (2011).
32. Khorshed, R. A. *et al.* Automated identification and localization of hematopoietic stem cells in 3D intravital microscopy data. *Stem Cell Reports* **5**, 139–153 (2015).
33. Preibisch, S., Saalfeld, S., Schindelin, J. & Tomancak, P. Software for bead-based registration of selective plane illumination microscopy data. *Nature Methods* **7**, 418–419 (2010).
34. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).
35. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
36. Hawkins, E. D. *et al.* Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. *Nature Protocols* **2**, 2057–2067 (2007).



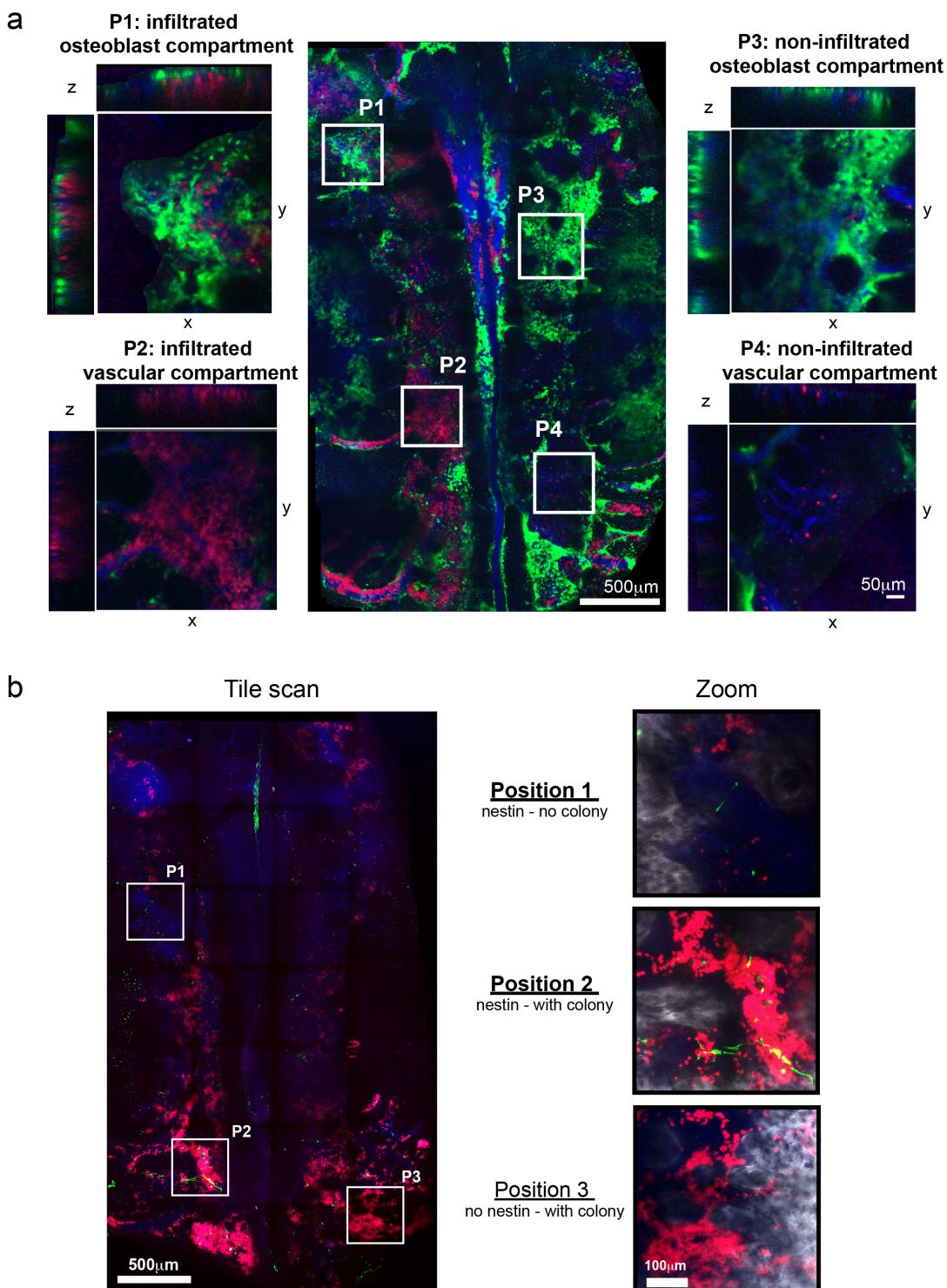
Extended Data Figure 1 | T-ALL disease experimental model. Fetal liver single-cell suspensions were isolated from embryonic day (E)14.5 wild-type (WT) embryos and transduced with DsRed alone or DsRed with NotchICN Δ Ram Δ P then transplanted into primary lethally irradiated recipient mice. Recipient mice typically accumulated CD4 $^+$ CD8 $^+$ cells in the peripheral blood from 4 weeks after transplant. Transformed leukaemic cells could be distinguished from non-malignant cells based on DsRed expression levels, where DsRED lo cells were transduced with the Notch construct yet were non-malignant as they had not yet acquired secondary mutations to drive leukaemogenesis, whereas DsRed hi cells were fully malignant. DsRed lo cells contained single-positive CD4 $^+$ and CD8 $^+$ T-cell populations, whereas DsRed hi cells had predominantly the leukaemic CD4 $^+$ CD8 $^+$ phenotype. The accumulation of transformed leukaemic populations displayed large variation over time as shown in the ratio of transformed to non-malignant cells in peripheral blood of primary recipient mice at 6 and 9 weeks. Data shown are mean \pm s.e.m. When DsRed hi cells dominated peripheral cell populations, mice were burdened with typical CD4 $^+$ CD8 $^+$ T-ALL (now simply referred to as DsRed $^+$). When primary recipient mice displayed enlarged lymph nodes and/or spleen, they were euthanized and DsRed $^+$ cells were harvested, stored frozen and transplanted into secondary recipients. CXCR4 expression was measured in CD4 $^+$ T cells, CD8 $^+$ T cells and T-ALL by flow cytometry

(T-ALL from four primary donors) and microarray gene expression analysis (triplicate biological replicates are shown for control T cells, and samples from nine individual secondary recipients injected with five independent primary T-ALL samples). To track disease progression, 10,000 primary T-ALL cells were transplanted into cohorts of sub-lethally irradiated secondary recipient mice. Secondary transplanted cells colonized the BM primarily, before spreading to peripheral organs and blood ($n = 4$ mice per time point; data shown are mean \pm s.e.m.) and developed disease more rapidly and synchronously than primary recipients as injected cells were already transformed. Secondary recipients survived for up to 38 days (shown are survival data of mice injected with four independent primary T-ALL samples, $n = 2$ mice per primary sample represented by the following symbols: filled circle, open square, open triangle and asterisk). In selected cases, secondary T-ALL blasts were transplanted into tertiary recipients, which developed disease more rapidly but showed similar responses to chemotherapy. The primary samples used in this study were from primary recipients 151, 907, B2M2, B2M3, B2M10, B3M3 and B3M30. This nomenclature refers to the mouse numbering from three independent fetal liver transductions. First transduction: mouse 151 and 907; second transduction (B2): mouse 2, 3 and 10; third transduction (B3): mouse 3 and mouse 30.



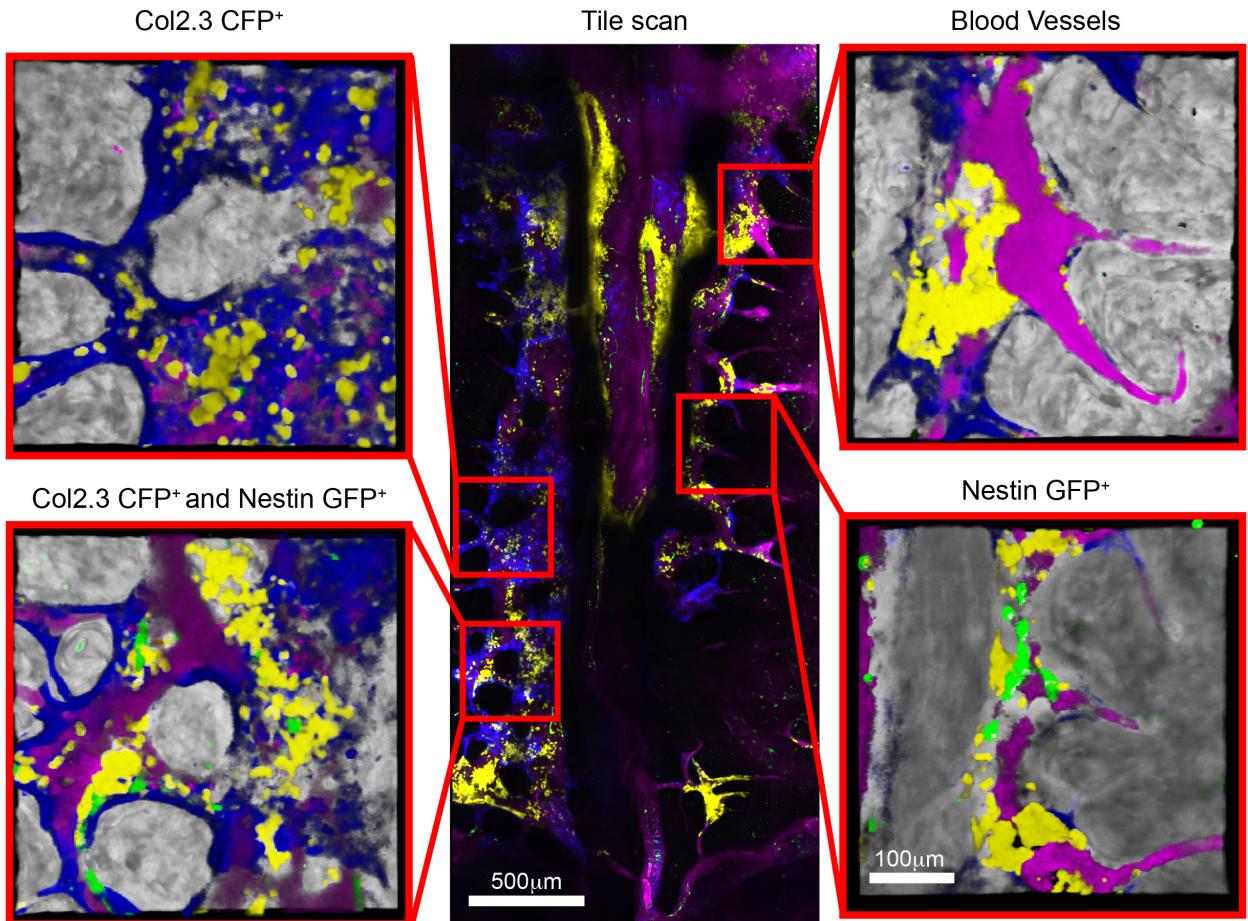
Extended Data Figure 2 | Four-dimensional multi-position imaging of leukaemia cells in the BM space. **a**, Representative maximum projection tile scan of a Col2.3-GFP recipient mouse (from Fig. 2) 12 days after transplantation of DsRed⁺ T-ALL. Red: T-ALL cells; green: GFP⁺ osteoblastic cells; blue: blood vessels; grey: bone collagen. **b**, Individual positions (framed in **a**) were selected and imaged at 3 min intervals for 3 h to measure cell migration and division. Shown here are a single time frame/position. For full time-lapse data see Supplementary Video 2. **c**, Three-dimensional rendering of the three-dimensional tracks of

individual leukaemia cells, measured using semi-automated tracking (red), overlaid to vasculature (blue). Spheres represent the beginning of each track. **d**, Data from **c** are shown with osteoblasts included in green. **e**, Tracks from **c** colour-coded based on time. Long stretches of the same colour correspond to faster movement, while rapid colour shifts represent slower movement. Data are representative of >30 time-lapse videos collected from eight secondary recipients (biological replicates) injected with three independent primary T-ALL samples.



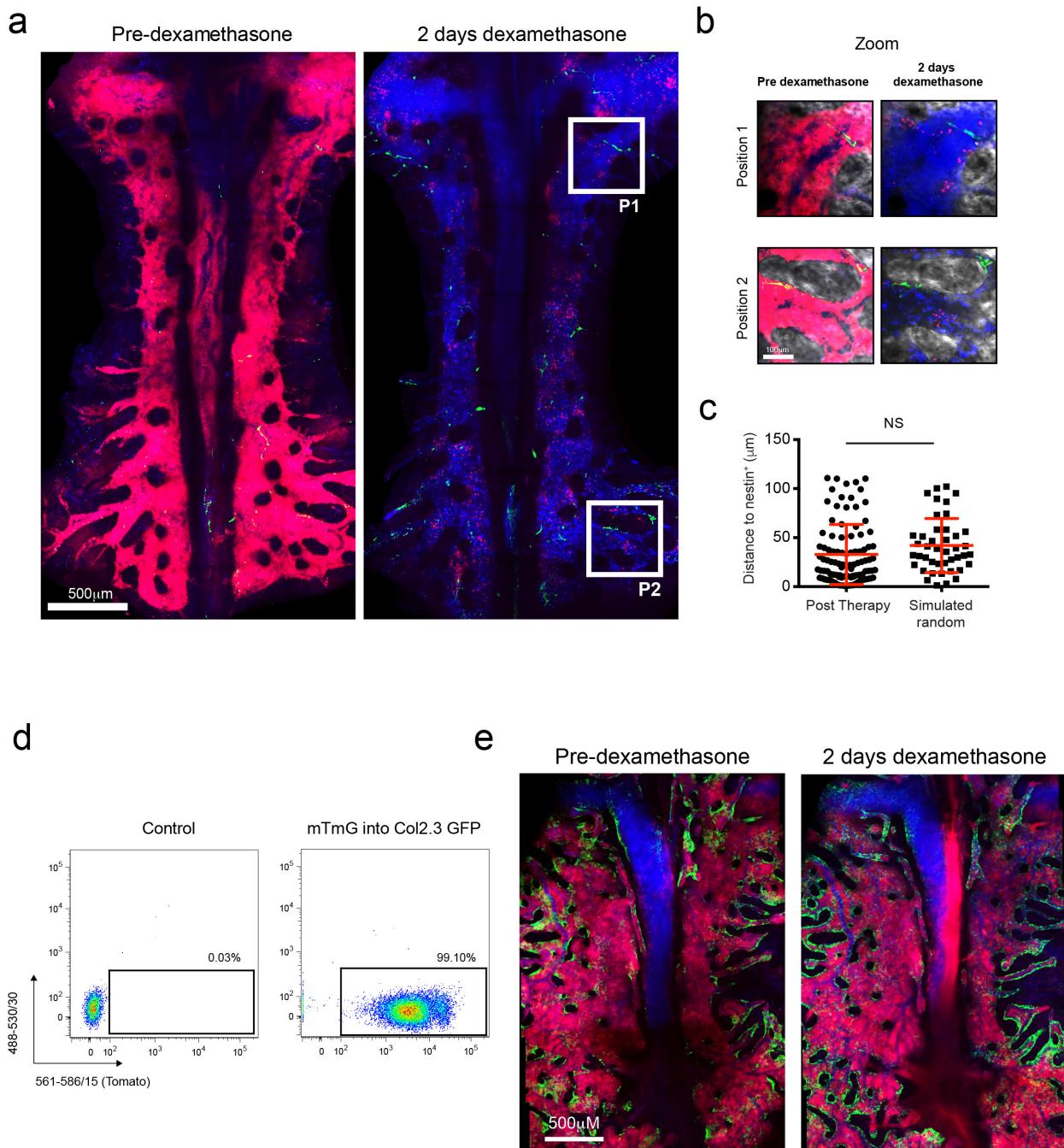
Extended Data Figure 3 | T-ALL expansion is not associated with Col2.3-GFP⁺ osteoblastic cells and nestin-GFP⁺ cells. **a**, Representative tile scan of a Col2.3-GFP mouse 15 days after transplantation of T-ALL. Zooms P1–P4 illustrate that the expansion of disease is not associated with the presence or absence of GFP⁺ osteoblastic cells. Red: T-ALL; green: osteoblastic cells; blue: blood vessels. Image is representative of five mice injected with two individual T-ALL primary samples. **b**, Representative

tile scan of a nestin-GFP mouse 15 days after transplantation of T-ALL. Zooms P1–P3 illustrate that the expansion of disease is not associated with the presence or absence of nestin-GFP⁺ cells. Red: T-ALL; green: nestin⁺ cells; blue: blood vessels; grey: bone collagen second harmonic generation (SHG) signal. Image is representative of three individual mice (biological replicates).



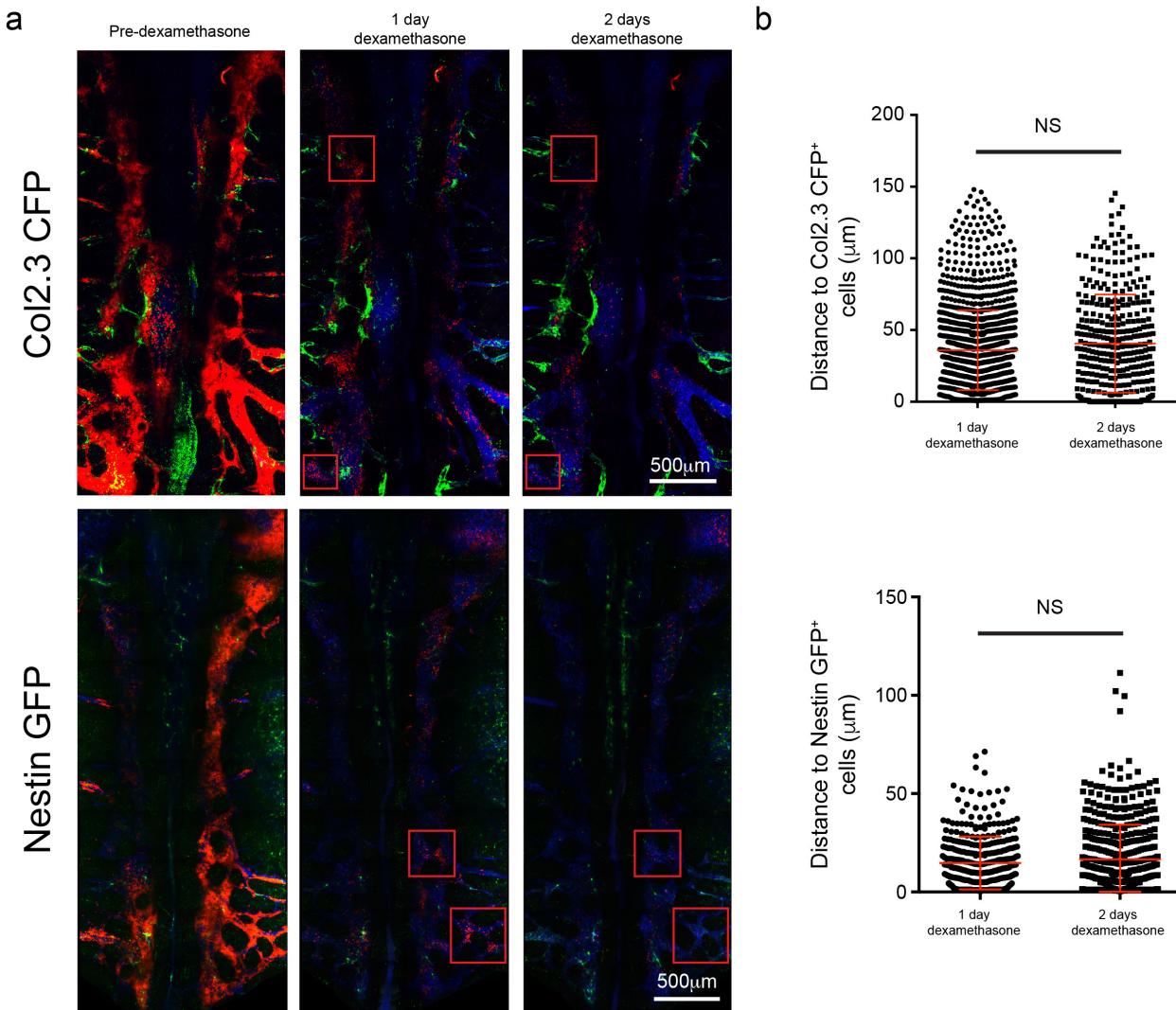
Extended Data Figure 4 | T-ALL expansion is not associated with BM areas containing nestin-GFP⁺, Col2.3-GFP⁺ cells or any combination of them. Representative tile scan of a Col2.3-CFP/nestin-GFP double-transgenic mouse 15 days after transplantation of T-ALL. Zooms P1–P3 illustrate that the expansion of disease is not associated with the presence

or absence of any combination of Col2.3-CFP⁺ or nestin-GFP⁺ cells. Yellow: T-ALL; green: nestin-GFP⁺ cells; blue: Col2.3-CFP⁺ cells; magenta: Cy5-labelled blood vessels; grey: bone collagen SHG signal. Image is representative of four mice (biological replicates).



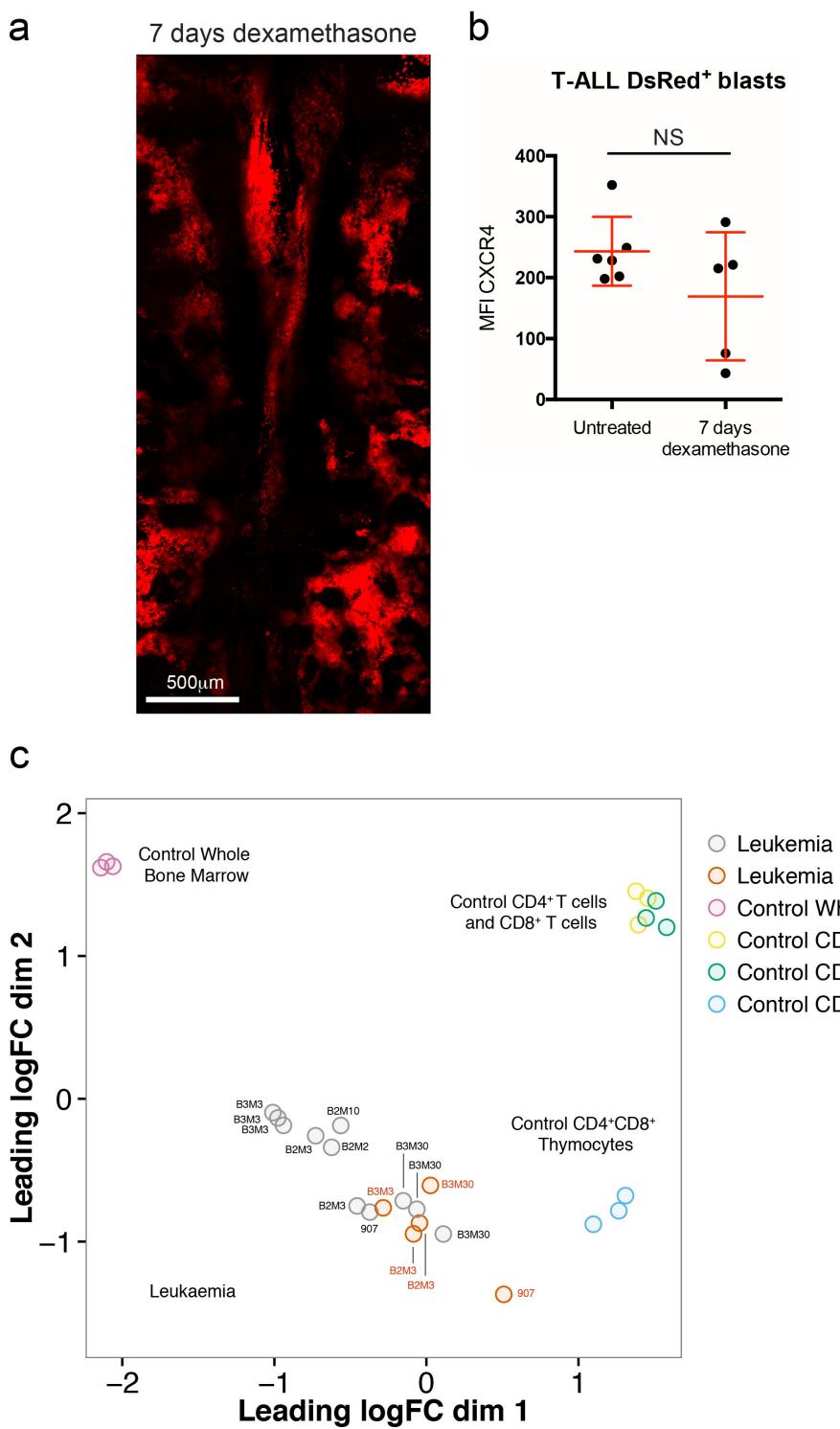
Extended Data Figure 5 | Dexamethasone-resistant T-ALL cells do not associate with nestin-GFP⁺ cells. **a**, Representative nestin-GFP mouse transplanted with T-ALL cells and imaged 18 days post-transplant to confirm complete BM infiltration (left). Tile-scan imaging was repeated after 2 days of treatment with 15 mg kg⁻¹ dexamethasone i.v. (right). **b**, Magnified view of representative positions, framed in **a**. **c**, We observed no preferential positioning of T-ALL surviving cells relative to nestin⁺ cells compared to simulated data. Red: T-ALL; green: nestin-GFP⁺ cells; blue: blood vessels. Data are representative of four individual mice injected with two T-ALL primary samples. Error bars: mean \pm s.d.

d, Chimaeric mice were generated by transplanting mTmG-tomato⁺ BM into Col2.3-GFP recipients. The high reconstitution efficiency of mTmG cells provided a more robust traceable marker of steady-state haematopoiesis for intravital imaging than MigR1-DsRed-transduced fetal liver cells (<40% reconstitution, data not shown). **e**, Representative tile scans of a mTmG/Col2.3-GFP chimaeric mouse performed before and after three doses of dexamethasone treatment showing that healthy BM is not affected by dexamethasone treatment or sub-lethal irradiation. Red: tomato⁺, healthy mTmG BM; green: GFP⁺ osteoblastic cells; blue: blood vessels. $n = 2$ mice. NS, not significant.



Extended Data Figure 6 | Multi-day time course of response to chemotherapy. **a**, Representative maximum projections of tile scans of calvarium BM of one Col2.3–CFP and one nestin–GFP mouse at 18 days after T-ALL transplant (pre-treatment) and 1 and 2 days of dexamethasone treatment (15 mg kg^{-1}). Red squares indicate some areas of cell loss from day 1 to day 2. **b**, Three-dimensional measurement of the position of

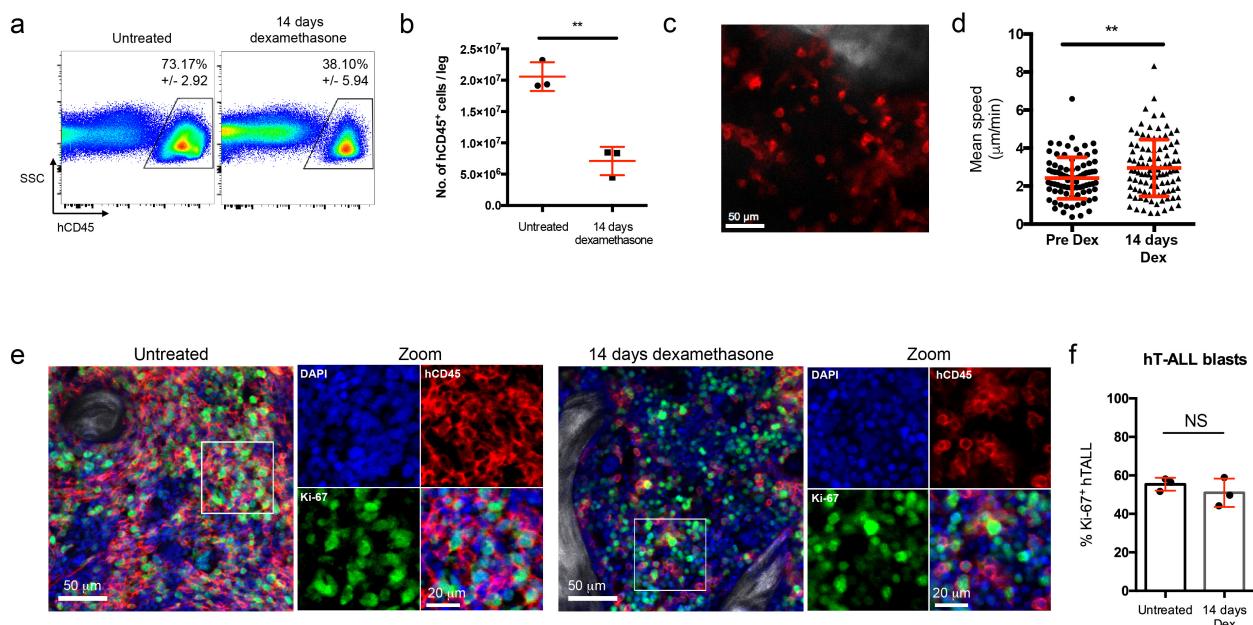
surviving cells in mice imaged at both 1 and 2 days of dexamethasone treatment. $n = 1,499$ and 352 T-ALL cells measured to osteoblastic CFP⁺ cells at day 1 and day 2, respectively, and 363 and 496 T-ALL cells measured to nestin–GFP⁺ cells at day 1 and day 2, respectively. Data are representative from three individual mice (biological replicates) of each genotype. Error bars: mean \pm s.d. NS, not significant.



Extended Data Figure 7 | See next page for caption.

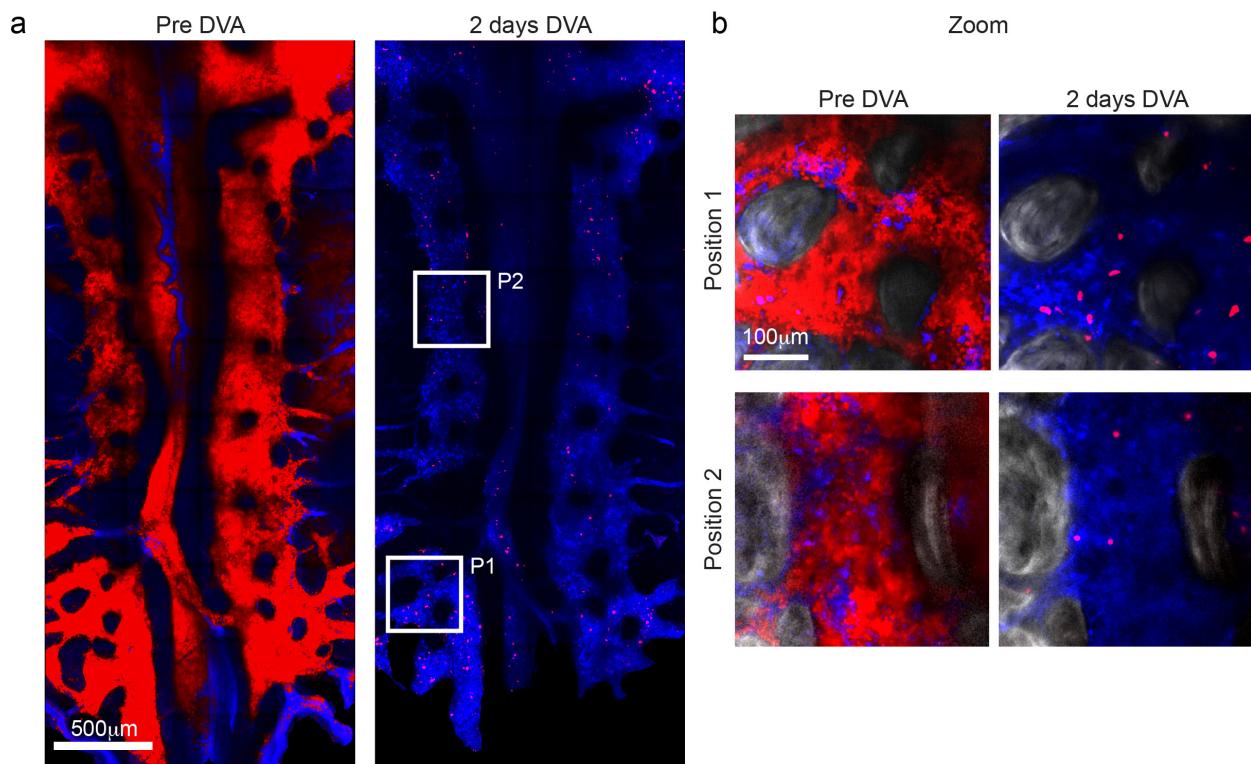
Extended Data Figure 7 | Development of resistance to dexamethasone and gene-expression-based clustering of leukaemia samples collected before and after dexamethasone treatment. **a**, Representative maximum projection of tile scan of calvarium bone marrow of a mouse after 7 days of daily dexamethasone treatment (15 mg kg^{-1} , i.v.). Red, DsRed⁺ T-ALL cells. Data are representative of four independent mice injected with four independent T-ALL primary samples. **b**, Mice with T-ALL were either kept untreated or treated with dexamethasone for 7 days, at which point they were culled and the expression of CXCR4 on T-ALL cells analysed by flow cytometry. There was no statistically significant difference in the mean fluorescence intensity (MFI) of CXCR4 between the two groups. Data are representative of six untreated and five treated mice, injected with three independent T-ALL primary samples. Error bars: mean \pm s.d.

c, Multi-dimensional scaling (MDS) plot of control CD4⁺ T cells, CD8⁺ T cells, CD4⁺8⁺ thymocytes, whole BM and T-ALL samples with no treatment (grey) or after treatment with dexamethasone (red) based on microarray transcriptomics data for the 1,000 most variable genes. The name of the primary T-ALL sample used to inject each mouse is indicated next to the dot marking its position relative to all other samples. This nomenclature refers to the mouse numbering from three independent fetal liver transductions. First transduction: mouse 907; second transduction (B2): mouse 2, B2M2, mouse 3, B2M3 and 10, B2M10; third transduction (B3): mouse 3, B3M3 and mouse 30, B3M30. Control samples are purified by flow cytometry from three biological replicate mice and each circle represents an individual sample. NS, not significant.



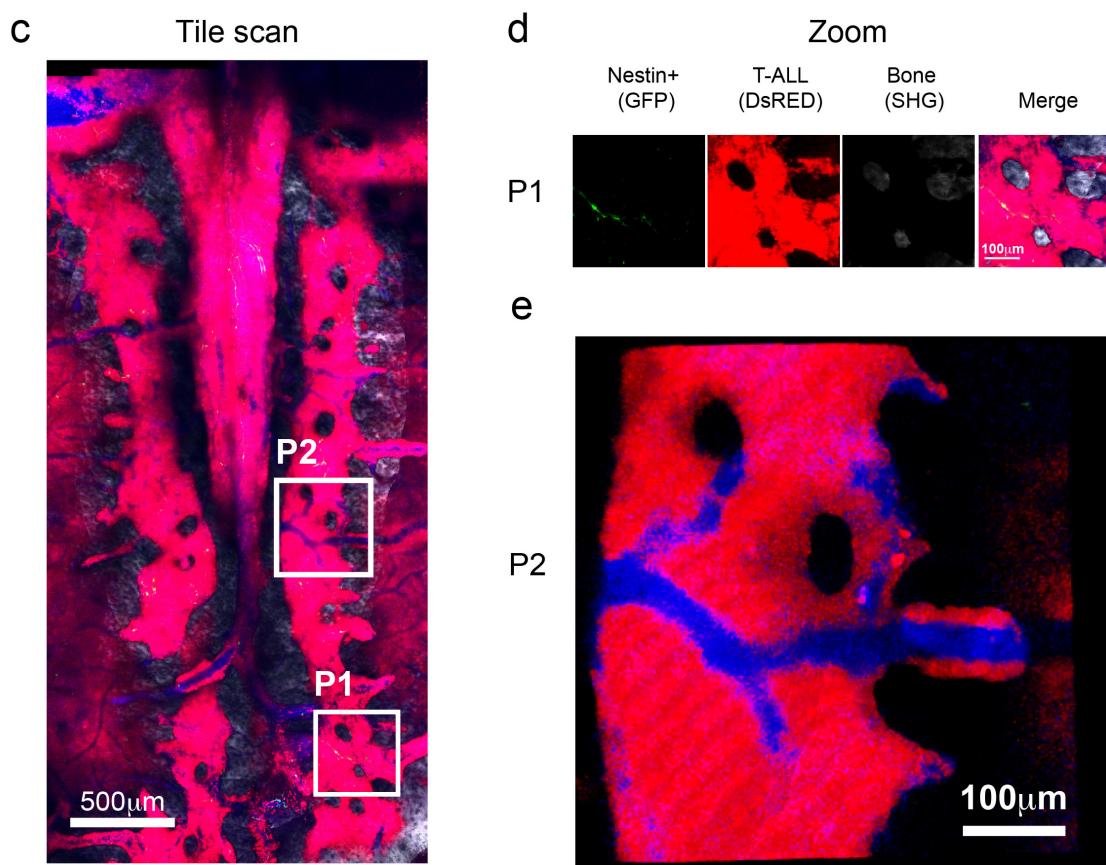
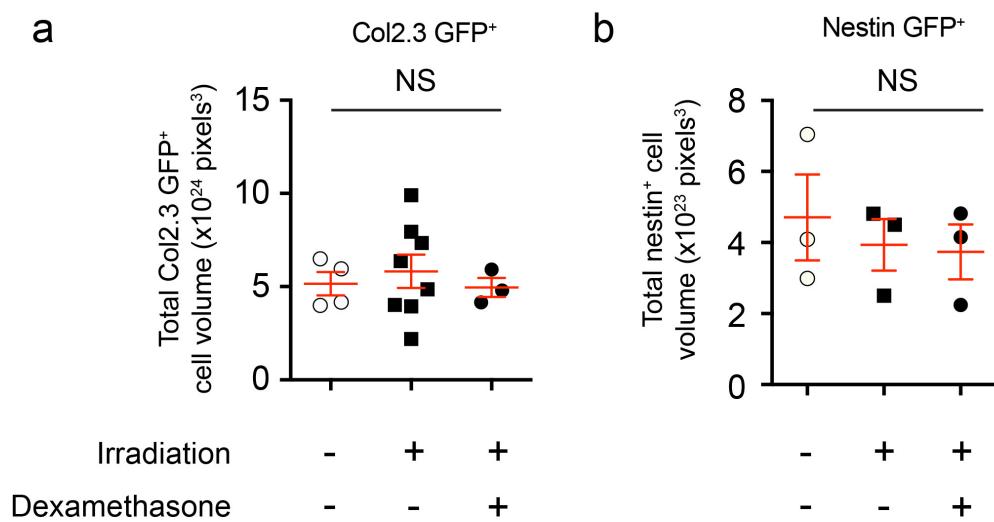
Extended Data Figure 8 | Analysis of human T-ALL cells during response to chemotherapy in NSG xenotransplant recipients.
a, b, Human T-ALL samples were transplanted into NOD/SCID/ γ (NSG) mice and 12 days post-transplant, daily dexamethasone treatment at 15 mg kg^{-1} was initiated. Fourteen days later, the response was measured by flow cytometry. **c**, For intravital imaging, human T-ALL cells were labelled by injection of $10 \mu\text{g}$ anti-human CD45-PE 15–30 min before imaging. **d**, Cells were imaged at 3 min intervals for >60 min and migration was measured by manual tracking either before or after dexamethasone treatment. Pre-dexamethasone: $n = 82$ cells from

2 independent mice, 14 days; dexamethasone: $n = 100$ from 3 independent mice. Shown are cells from patient JH, wild-type NOTCH. **e**, BM sections were prepared from untreated and treated NSG mice and stained for human CD45 (red) and Ki-67 (green). In addition, nuclei were visualized using DAPI (blue) and bone by SHG signal (grey). Zooms are of the areas framed by the white boxes on their left. **f**, Analysis of 2,338 (untreated) and 1,576 (14 days dexamethasone) human CD45⁺ cells in sections from three mice per condition reveals no change in the fraction of proliferating Ki-67⁺ cells after dexamethasone treatment. NS, not significant.
 $**P < 0.01$. Error bars: mean \pm s.d. (**b, d, f**).



Extended Data Figure 9 | Combined dexamethasone, vincristine and L-asparaginase treatment effectively reduces T-ALL burden.
a, Representative tile scan of a mouse calvarium fully infiltrated with T-ALL (pre-DVA) and after 2 days of combination therapy (dexamethasone, vincristine and L-asparaginase (DVA)). **b**, Zooms

P1 and P2 illustrate effectiveness of DVA treatment and the small number of surviving T-ALL cells. Red: T-ALL; blue: blood vessels; grey: SHG bone collagen. Image is representative of four mice (biological replicates) injected with one individual T-ALL secondary sample.



Extended Data Figure 10 | Analysis of the response of bone marrow structures to irradiation and dexamethasone treatment and of nestin-GFP⁺ cells to T-ALL. a, b, Col2.3-GFP (a) or nestin-GFP (b) mice were treated with combinations of sublethal irradiation (administered >18 days before measurement) or dexamethasone treatment (administered for 2 days before measurement) as indicated. Then, using three-dimensional image analysis of tile scans, the total volume of GFP⁺ cells was quantified. Groups were analysed using analysis of variance (ANOVA) with Bonferroni correction for multiple groups. Error bars: mean \pm s.d.

c, Representative tile scan of nestin-GFP mouse transplanted with T-ALL 21 days earlier. At infiltration levels that eradicated osteoblasts, we still observed healthy nestin-GFP⁺ cells. d, Higher magnification of area P1 framed in A, with the signal from each channel split for clarity. e, Three-dimensional render at higher magnification of area P2 framed in A, showing healthy blood flow within the highly infiltrated BM space. Red: T-ALL; green: nestin-GFP⁺ cells; blue: blood vessels; grey: bone collagen SHG signal. $n = 5$ independent mice injected with two independent T-ALL primary samples. NS, not significant.

High dimensional single cell analysis predicts response to anti-PD-1 immunotherapy

Carsten Krieg, Małgorzata Nowicka, Silvia Guglietta, Sabrina Schindler, Felix J. Hartmann,

Lukas M. Weber, Reinhard Dummer, Mark D. Robinson, Mitchell P. Levesque and

Burkhard Becher

Paper under review at *Nature Medicine*

Title: High dimensional single cell analysis predicts response to anti-PD-1 immunotherapy

Authors: Carsten Krieg^{*1}, Małgorzata Nowicka^{2,3}, Silvia Guglietta⁴, Sabrina Schindler⁵, Felix J. Hartmann¹, Lukas M. Weber^{2,3}, Reinhard Dummer⁵, Mark D. Robinson^{2,3}, Mitchell P. Levesque^{#*5}, Burkhard Becher^{#*1}.

Affiliations:

¹Institute of Experimental Immunology, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland

²Institute of Molecular Life Sciences, University of Zurich, Winterthurerstr. 190, CH-8057 Zurich, Switzerland

³SIB Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstr. 190, CH-8057 Zurich, Switzerland

⁴Department of Experimental Oncology, European Institute of Oncology, Via Adamello 16, I-20139 Milan, Italy

⁵Department of Dermatology, University Hospital Zurich, CH-8091 Zurich, Switzerland.

***Correspondence to:** krieg@immunology.uzh.ch, mitchell.levesque@usz.ch or becher@immunology.uzh.ch

#these authors contributed equally

One sentence summary: The frequency of CD14⁺CD16⁻HLA-DR^{hi} classical monocytes predicts response of melanoma patients to anti-PD-1 immunotherapy.

ABSTRACT

Immune checkpoint blockade has revolutionized cancer therapy. In particular, inhibition of programmed cell death protein 1 (PD-1) has proven to be effective for the treatment of metastatic melanoma and other cancers, but despite a dramatic increase in progression-free survival, only a minority of patients shows durable clinical benefit. Therefore, predictive biomarkers of clinical response are desperately needed. Here, we employed high-dimensional single cell cytometry by mass cytometry and an unbiased, custom algorithm-assisted bioinformatics pipeline for the in depth characterization of the immune compartment in liquid biopsies from the same metastatic melanoma patient before and after anti-PD-1 immunotherapy. We could observe a clear treatment response to immunotherapy in the T cell compartment. However, a strong predictor of responsiveness to anti-PD-1 immunotherapy, before treatment initiation, was the frequency $CD14^+CD16^-HLA-DR^{hi}$ monocytes. We could confirm this by regular flow cytometry and propose this as a novel predictive biomarker for therapy decisions in the clinic.

INTRODUCTION

Immunotherapy with anti-PD-1 aims to block the interaction of tumor-reactive T cells with PD-1 ligands (PD-L1 and PD-L2) expressed on various cell types including leukocytes and the tumor cells themselves¹. Clinical trials on PD-1 and PD-L1 blockade for patients with advanced melanoma have demonstrated consistent therapeutic responses, thus prompting their application to several other cancers²⁻⁸.

Despite these encouraging results, clinical outcomes remain highly variable, with only a fraction of patients showing durable responses, some with early progression and others with late response, while the majority of treated patients show no beneficial clinical response^{2,9}. Reliable criteria to discriminate responders from non-responders prior to treatment initiation are urgently needed. Predictive biomarkers would allow for the selection of patients who are more likely to respond and to provide predicted non-responders with alternative therapeutic options early on. Some recent reports used single-cell analysis to evaluate the expression of PD-1 and downstream signaling molecules on tumor infiltrating and circulating CD8⁺ T cells, with the aim of identifying such predictive biomarkers^{10,11}. However, these approaches are hampered by the limited accessibility of patient material, low dimensionality, overfitting due to the absence of independent validation cohorts, and lack of systematic, unbiased bioinformatics analysis pipelines resulting in a paucity of predictive biomarkers to date¹². Single-cell analysis of human immune compartments has so far been limited by the parameters that can be visualized by conventional flow cytometry¹³.

In this study, we used peripheral blood mononuclear cells (PBMC) from metastatic melanoma patients before and during therapy as a readily accessible and minimally invasive biopsy that has been shown to be more representative than tumor biopsies to probe immune signatures associated with responsiveness to anti-PD-1 immunotherapy¹⁴. High dimensional, single cell mass cytometry was used along with optimized immune marker panels and a customized, interactive bioinformatics pipeline to generate a thorough analysis of the peripheral blood immune cells in an effort to identify a responsiveness-associated predictive signature.

RESULTS

Stratification of responders versus non-responders using single cell mass cytometry

We performed the initial analysis with 40 cryopreserved PBMC samples isolated from the blood of a cohort of 20 melanoma patients as well as 10 age- and sex-matched healthy donors. Baseline samples and samples obtained after 12 weeks of anti-PD1 therapy originated from the same patients (table 1 and Figure 1A).

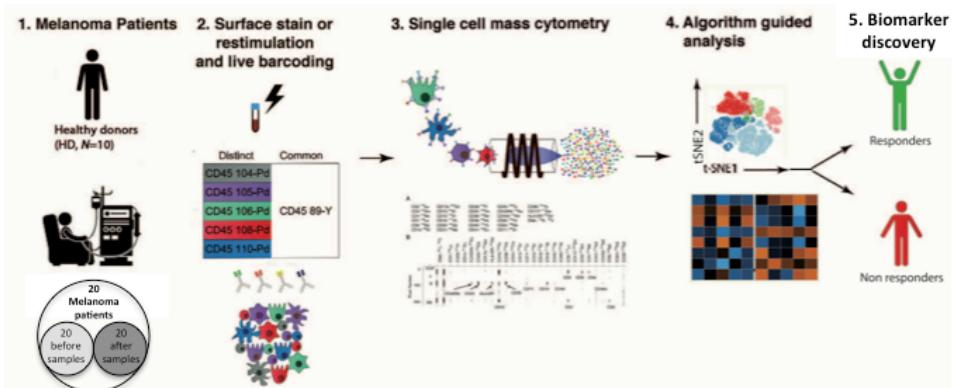
Table 1. Characteristics of blood samples from melanoma patients and healthy donors used for the biomarker discovery study. Numbers in parentheses display the age range of subjects.

Healthy Donors									
	Donors	samples		samples		Sample TOTAL			
N	10	2x5		2x5		20			
Age (years)	60.3 (46-71)								
Sex (male/female)	6/4								
Melanoma patients									
	<i>Responder</i>			<i>Non-Responder</i>					
	Patients	samples before therapy	samples after therapy	Patients	samples before therapy	samples after therapy	Sample TOTAL		
N	11	11	11	9	9	9	40		
Age (years)	62.0 (42-81)		57.8 (45-75)						
Sex (male/female)	9/2		5/4						
Pre-treatments									
Radiotherapy	6/11			5/9					
Chemotherapy	3/11			3/9					
Ipilimumab	9/11			7/9					
Other	0/11			2/9					

For the CyTOF analysis, frozen PBMCs were thawed and stained (Figure 1A and Supplementary Table 1) using three separate and partially overlapping mass cytometry panels, one for the phenotypic characterization of lymphocytes, one for T cell function and one specifically for the in-depth characterization of myeloid cells. The first staining panel contained 30 leukocyte markers to identify all major immune cell populations and cover all stages of T cell differentiation and activation (Supplementary Table 1). After acquisition, each sample was de-barcoded using Boolean gating. Staining quality was evaluated by defining a biological positive and negative control (Supplementary Figure 1).

After data pre-processing, we performed hierarchical clustering on normalized median marker expression values on CD45⁺ live cells in every patient before and after therapy. As demonstrated in Figure 1B, the dendrogram displayed two major clades. The left branch contained 15 samples, of which all were responders, whereas the right branch consisted of 18/25 samples from non-responders (72%). Thus, normalized median marker expression was sufficient to robustly separate most responders from non-responders. The unbiased clustering approach stratified the patients into responders and non-responders prior to therapy, which encouraged us to analyze the dataset more deeply.

A Workflow



B Patient stratification by dendrogram

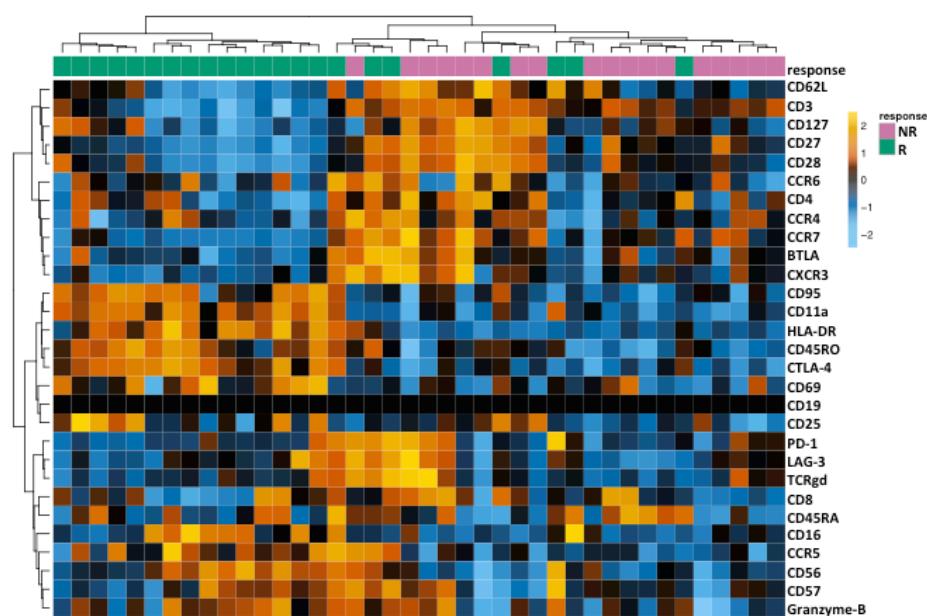
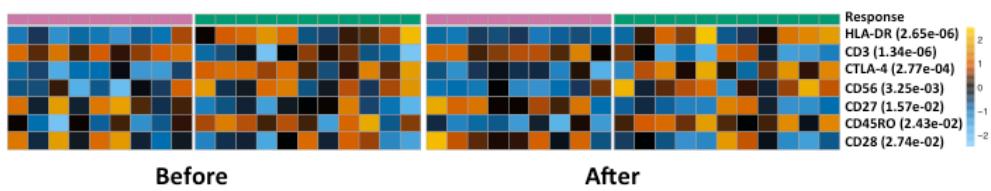


Figure 1. Stratification of responders and non-responders and identification of differences in immune cell populations using mass cytometry. (A) Experimental setup for the processing of frozen PBMC from matched samples before and after PD-1 immunotherapy from 20 melanoma patients using metal-labeled antibodies and acquisition by mass cytometry. (B) Dendrogram tree built on hierarchical clustering using Ward linkage of the normalized median marker expression from CD45⁺ single live cells of thawed patient PBMCs. Bars on top of the heatmap represent individual samples from responders (green) versus non-responders (red). Each column represents one patient sample from one time point (n patients=20, n samples=40).

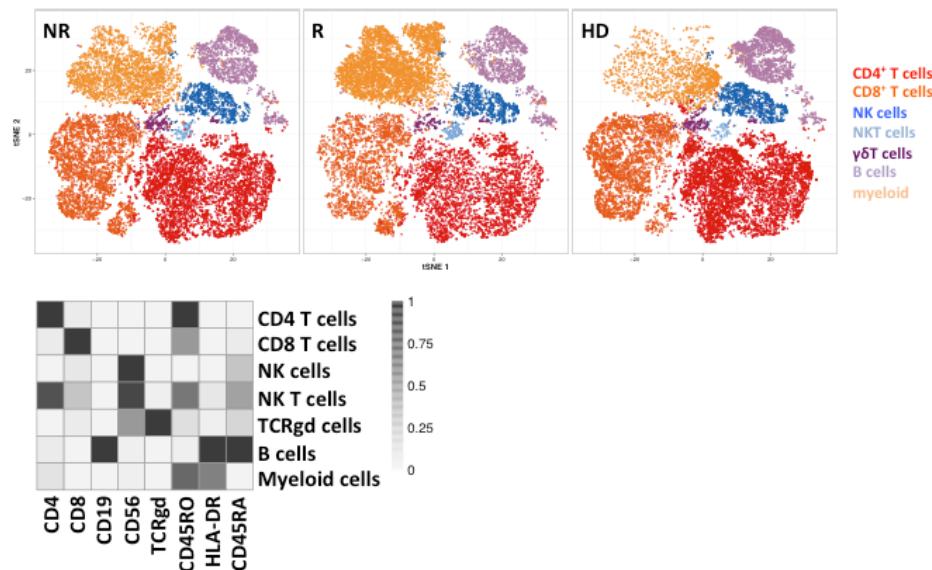
Altered T cell memory compartment before therapy in responders

We next tested the hypothesis that the differences in normalized median marker expression were driven by alterations in the relative abundance of the various cell populations between responders and non-responders. Therefore, we analyzed the differential median expression of the 29 markers, comparing responders and non-responders, before and after therapy initiation (Figure 1C). Significant increases in the expression of HLA-DR, CTLA-4, CD56, and CD45RO and decreased amounts of CD3, CD27, and CD28 were observed in responders versus non-responders.

C Global differential marker expression



D tSNE groups



E Immune cell frequencies

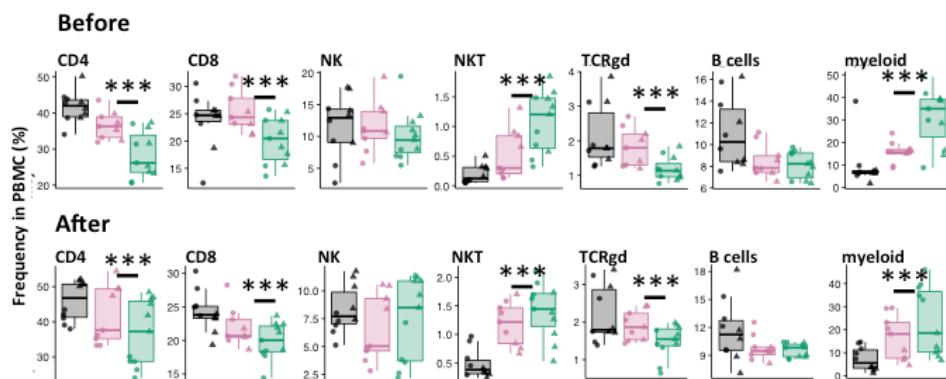


Figure 1 (continued). Stratification of responders and non-responders and identification of differences in immune cell populations using mass cytometry.

(C) Heatmap of significantly differentially expressed markers between responders and non-responders before and 12 weeks after therapy initiation, in pre-processed live single cells. Heat scale shows normalized median marker expression ranging from under-expressed (blue) to over-expressed (orange) where changes in marker expression between responders and non-responders was significant ($p < 0.05$). Colored bars on top of the heatmap represent individual samples from responders (green) and non-

responders (red). (D) Cells from healthy donors and patients were used as an input for the FlowSOM algorithm. Thirteen algorithm-chosen markers were used to form 7 machine-assisted clusters. Visualization of 15'000 events in non-responders (NR), responders (R), and healthy donors (HD) using the tSNE algorithm. The heatmap represents the expression of respective markers within a cellular cluster and was used to annotate clusters, which were overlaid in color code (panel on the right). (E) Direct comparison of cluster frequencies in healthy donors (HD, black), non-responders (NR, pink) and responders (R, green) from batch 1 (circles) and batch 2 (triangles) extracted from tSNE immune clusters in C. Numbers in brackets indicate p-values (**p=<0.01).

Next, we sought to identify which cell populations best described the cellular frequency differences between responders and non-responders. Markers were selected using the PCA informativeness score established by Levine et al.¹⁸, cells were clustered using the FlowSOM algorithm^{19,20} with consensus clustering and a two-dimensional t-stochastic neighbour embedding (tSNE) projection was used for visualization, as shown in Figure 1D²¹. Based on the marker intensities detected in the clusters, we manually annotated the seven major cell populations (CD4 T cells, CD8 T cells, NK cells, NKT cells, B cells, gdT cells, and myeloid cells) and then separated them into the three groups (HD, NR, R). We subsequently examined differences in frequencies in between groups of the identified clusters (Figure 1E) using a generalized mixed model (see Methods). Of note, the barcoding allowed the model to track the patients and match baseline and on-treatment samples for the analysis. In responders, the frequency of CD4⁺ T cells and CD8⁺ T cells was lower, while the frequency of CD19- HLA-DR+ myeloid cells was significantly elevated (p-values = 1.55e-05, 1.74e-03 and 1.74e-03, respectively) compared to the non-responders (before and during treatment). We also observed a higher frequency of NKT cells and a lower frequency of gdT cells (p=3.07e-03 and 2.52e-03) in responders versus non-responders at both time points.

Since T cells are described to be the major target of anti-PD-1 immunotherapy, and given the altered T cell composition in responders before immunotherapy, we next compared the normalized median marker expression on T cells between non-responders and responders before and after therapy. CD4⁺ T cells in responders showed an up-regulation of CTLA-4, HLA-DR, CD69, and BTLA (Fig. 2A) already at baseline before therapy. CD8⁺ T cells in responders showed a higher expression of CD45RO, CTLA-4, CD62L, CD69, CD11a, and CCR4 expression (Fig. 2B).

To determine whether there were differences in T cell subpopulations, we next extracted CD4⁺ T cells and CD8⁺ T cells from the FlowSOM-generated clusters in Fig. 1C and subdivided them into CD45RO⁻CD62L⁺ naïve, CD45RO⁻CD62L⁻ effector cells (TE), CD45RO⁺CD62L⁻ effector memory (EM) cells, CD45RO⁺CD62L⁺ central memory (CM) cells or CD127⁻CD25⁺ regulatory T cells (T_{regs}) using FlowSOM (Figures 2C and D).

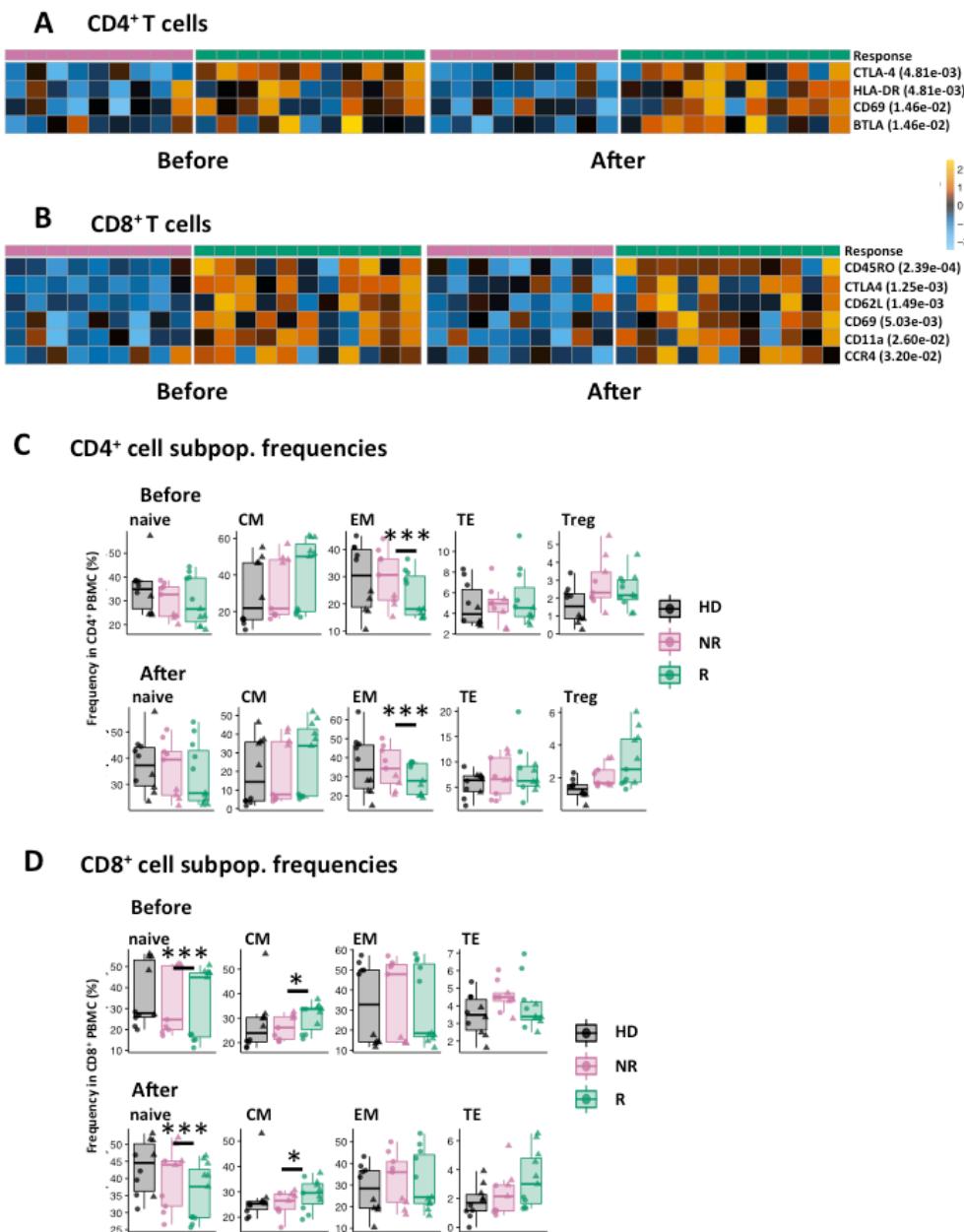


Figure 2. Differences in T cell activation status and in the frequency of the T cell subpopulations before and after 12 weeks of therapy in responders and non-responders. Heatmaps showing

significantly different ($p=<0.05$) normalized median marker expression in responders (green bar on top) and non-responders (red bar on top) in CD4 $^{+}$ T cells (**A**) and CD8 $^{+}$ T cells (**B**) before and after therapy. (**C** and **D**) FlowSOM was used to generate indicated T cell subpopulations and resultant cluster frequencies from batch 1 (circles) and batch 2 (triangles) are plotted as in Fig. 1E. Heat scale shows over-expression (orange) or under-expression (blue) of the respective marker. Numbers in brackets show p-values (* $p=<0.1$, *** $p=<0.01$).

We then compared the frequencies of resultant T cell sub-clusters between responders and non-responders before and 12 weeks after therapy. The patients who eventually responded to the therapy showed a significantly lower frequency of CD4 $^{+}$ EM T cells, as well as a lower frequency of CD8 $^{+}$ naïve T cell at baseline and after treatment (p-values: 8.21e-03, 6.95e-03). Additionally, the CD8 $^{+}$ T cell subpopulation of responders had a higher frequency of CM T cells before and after treatment as compared to non-responder patients (Figure 2C and D).

Anti-PD-1 immunotherapy alters the properties within the T cell compartment

In order to compare the functional properties of T cells between non-responders and responders, we designed a second mass cytometry panel to investigate cytokine production (Supplementary Figure 2) in polyclonally activated cells. PBMCs were processed as described above. Briefly, single cell suspensions were cultured for 4h in the presence of PMA/Ionomycin, barcoded, stained, fixed and analyzed by mass cytometry. In order to get a functional profile from antigen-experienced T cells, activated CD69 $^{+}$ memory and effector T cells ($T^{\text{mem/eff}}$ cells) were extracted and cytokine (IL-2, IL-4, IL-10, IL-13, IL-17A, GM-CSF, TNF- α , IFN- γ , Grz-B), PD-1 and CTLA-4 positive T cell subpopulations were identified. Importantly, we found no significant difference in cytokine production between responders and non-responders prior to therapy. However, after therapy T cells from responders presented with up-regulation of PD-1, IL-4, and granzyme-B in CD4 $^{+}$ CD69 $^{+\text{mem/eff}}$ T cells, while IL-17A-positive cells were less abundant (Fig. 3A). For CD8 $^{+}$ CD69 $^{+\text{mem/eff}}$ T cells, an up-regulation of CTLA4 and granzyme-B was detected in responders (Fig. 3B). In order to link these signatures to a specific cell population, we then created a matrix containing all possible marker combinations in CD4 $^{+}$ CD69 $^{+\text{mem/eff}}$ T cells (Fig. 3C) or CD8 $^{+}$ CD69 $^{+\text{mem/eff}}$ T cells (not shown). Using this approach, no differences in the CD8 $^{+}$ T cell subpopulations were observed. Figure 3D shows the different cell populations from this matrix when comparing CD4 $^{+}$ T cell subsets in responders to

non-responders. For CD4⁺CD69^{+mem/eff} T cells, we found that 6 clusters were more highly represented and 2 clusters were underrepresented in responders. Among the enlarged clusters, the most common signature was CTLA-4⁺, granzyme-B⁺, TNF- α ⁺, PD-1⁺ and IL-2⁺, which was present in clusters 10, 16 and 19 (p-values= 2.37e-2, 4.22e-02, and 3.95e-02). The lower frequency of IL-2+ in CD4⁺CD69^{+mem/eff} T cells from cluster 5 and the expansion of cluster 48 in responders reflect the higher activation status in the CD4⁺ T cell compartment that we observed in panel 1.

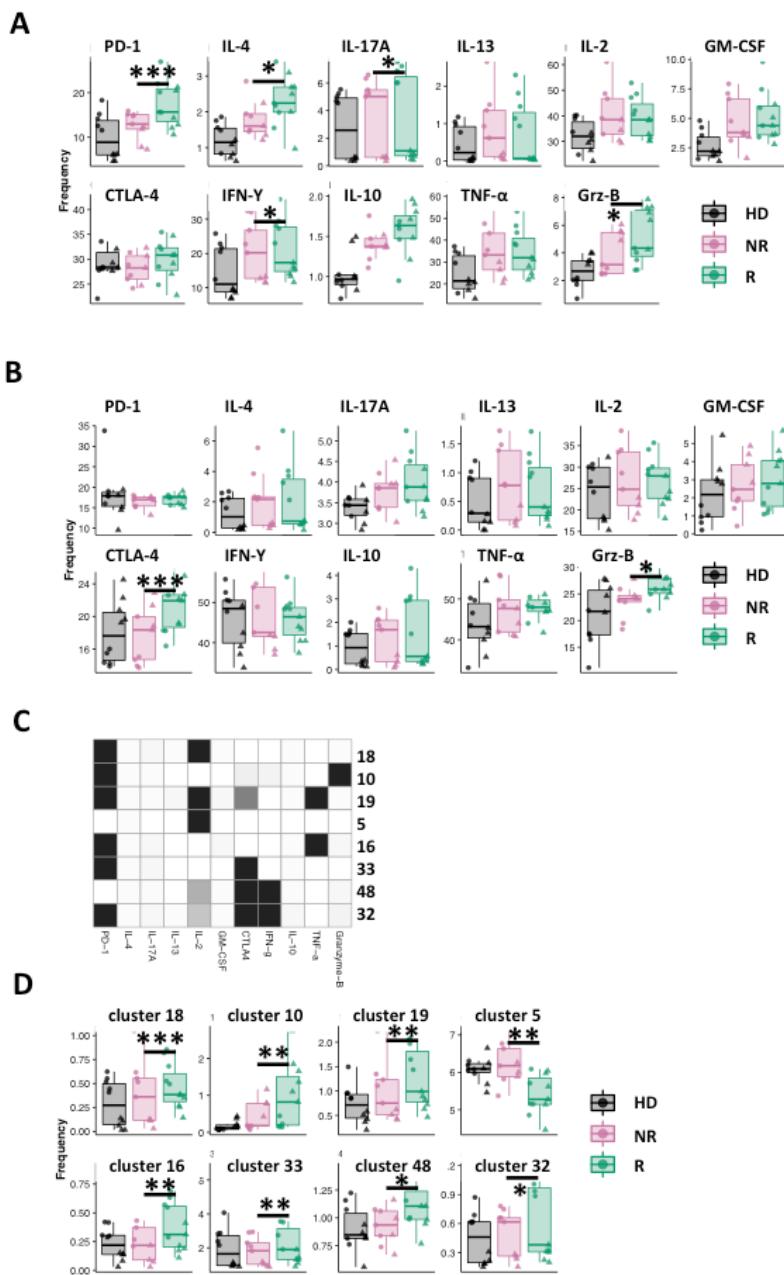


Figure 3. Increased activation in CD4⁺ or CD8⁺ CD69⁺T^{mem/eff} cells after immunotherapy start in responders. CD4⁺ and CD8⁺ memory/effector T cells after therapy were extracted and activated polyclonally. Frequencies of PD-1, CTLA-4, and cytokines in (A) CD4⁺ CD69⁺ T^{mem/eff} cells and (B) CD8⁺ CD69⁺ T^{mem/eff} cells in responders (green) and non-responders (pink) were compared. Healthy subjects (grey) served as controls. (C) A matrix using the afore-mentioned markers was created and cells were sorted into this matrix using FlowSOM. Shown are the significantly different combinations after comparing responders to non-responders in CD4⁺ T cells. (D) Bar graphs displaying differences in cluster frequencies derived from C. Samples from batch 1 are indicated by circles and from batch 2 by triangles (*p=<0.1, **p=<0.05, ***p=<0.01).

Myeloid cell frequencies predict responsiveness to anti-PD-1 immunotherapy

Since we found higher frequencies of myeloid cells in anti-PD-1 therapy responders before therapy (Fig. 1D), we generated a third myeloid centric panel (Supplementary Figure 3). FlowSOM was used to separate 7 subpopulations, which were annotated as T cells, B cells, NK cells, CD14+(CD11b+HLA-DRhi) myeloid cells, CD14-(CD11b+HLA-DRlo) myeloid cells, classical CD1c⁺CD11c⁺HLA-DR⁺ dendritic cells cDC, and plasmacytoid dendritic cells (CD123⁺CD303⁺HLA-DR⁺CD11c⁻ pDC). From the annotated clusters, cell frequencies were extracted and the composition of the individual samples was plotted (Supplementary Figure 4 and Fig. 4A). As already shown in Figure 1D a significant lower frequency of T cells (p=1.59e-03) and a higher frequency of CD14+ myeloid cells was observed (p=5.82e-03, Supplementary Figure 4). While we found that CD14- myeloid cells were indeed higher in cancer patients compared to healthy donors, we did not observe differences in this cell population between responders and non-responders. Next, to better characterize these myeloid cells, we extracted live myeloid cells by manually gating out CD3⁺ and CD19⁺ cells and excluding CD7⁺ and CD56⁺ cells and marker expression from further analysis. Unsupervised clustering of normalized median marker expression values in myeloid cells again separated patients into two distinct clusters, with one clade being mostly composed of (86%) non-responders (n = 12/14) and the other clade consisting of 76% (n = 19/25) of responders (Figure 4B).

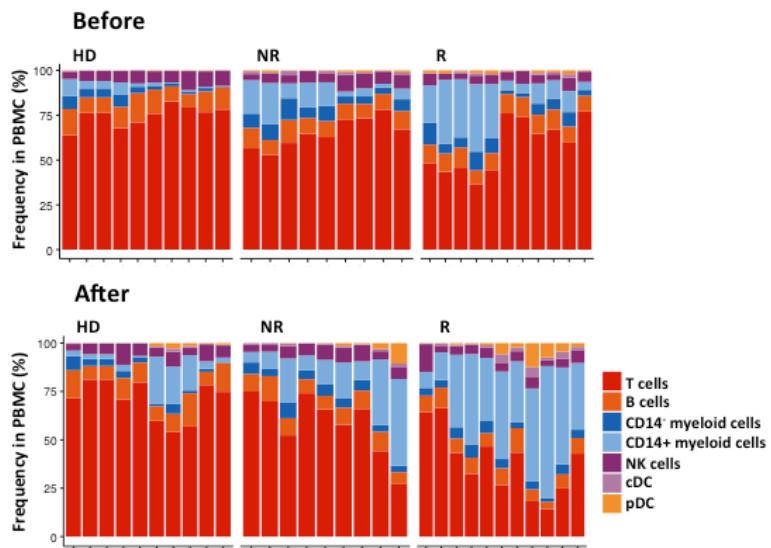
We next searched for changes in normalized median marker expression between non-responders and responders, before and 12 weeks after therapy, and we found that 16 markers (i.e., CD86, HLA-DR, CD141, ICAM-1, CD11c, PD-L1, CD38, CD16, CD33, CD11b, CD303, CD62L, CD1c, CD64, CD14, and CD34) were significantly up-regulated in the myeloid compartment of responders (Figure 4C). Next, FlowSOM

was used to subdivide the myeloid compartment into 4 major clusters, which were annotated as CD14⁺CD16⁻HLA-DR^{hi} classical monocytes, CD14⁻CD33^{low}CD11b⁺HLA-DR^{lo} myeloid cells, plasmacytoid dendritic cells (CD123⁺CD303⁺HLA-DR⁺CD11c⁻ pDC) and classical CD1c⁺CD11c⁺HLA-DR⁺ dendritic cells (cDC, Figure 4D).

Identification of a monocyte signature using CellCnn

The unsupervised FlowSOM algorithm allowed us to identify and characterize CD14⁺CD16b⁻HLA-DR^{hi} monocytes as being elevated in responders compared to non-responders prior to anti-PD-1 immunotherapy, but it also suggested that this is a heterogeneous cell population. To identify a core myeloid signature within CD14⁺CD16b⁻HLA-DR^{hi} cells that would allow us to predict responsiveness to anti-PD-1 immunotherapy without prior assumptions, we used the machine learning algorithm CellCnn, which is based on a representation learning approach using convolutional neural networks and is designed to detect rare cell populations associated with disease status²² (Figure 4E). In a data driven way, CellCnn automatically detects or “learns” several combinations of markers (“filters”, which need not correspond to known populations), whose presence or frequency discriminates between two groups. Results are visualized as plots showing differences in population frequencies and marker expression profiles for the most discriminating filters. We ran CellCnn on all baseline samples and were able to identify a robust small, discriminating cell population with a relative abundance of 1.5% +/- 1.1% (mean +/- standard deviation) in responders compared to 0.6% +/- 0.7% in non-responders. Although the variability within each group was relatively large, we found the difference in abundance to be statistically significant ($p < 0.01$, using an observation-level random effects model or OLRE to correctly model overdispersed binomial proportion data). In terms of marker expression, we found that this automatically detected population contained a core signature of CD14⁺, CD33⁺, HLA-DR^{hi}, ICAM-1⁺, CD64⁺, CD141⁺, CD86⁺, CD11c⁺, CD38⁺, PD-L1 and CD11b monocytes (Figure 4D right).

A



B

Patient stratification by dendrogram

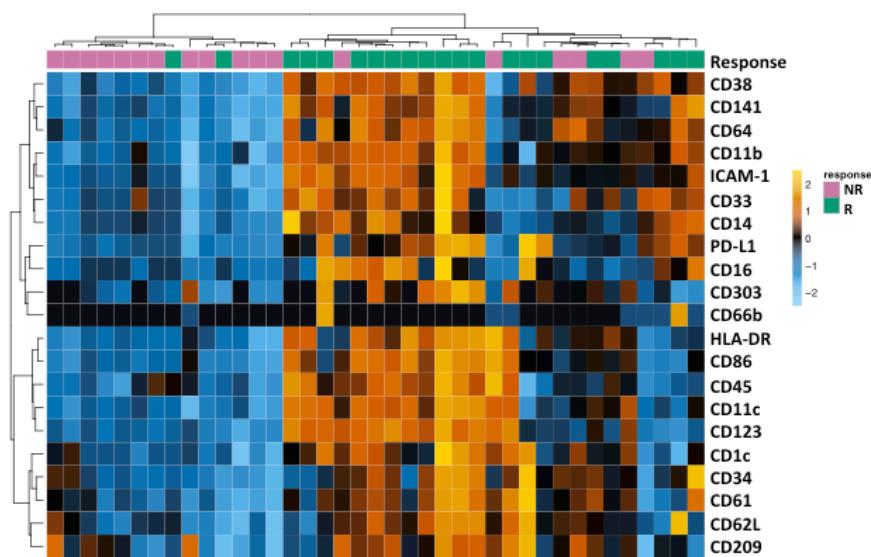


Figure 4. Patient stratification based on myeloid cell markers and expansion and enhanced activation of classical monocytes in responders. (A) Comparison of sample composition in healthy donors (HD) non-responders (NR) and responders (R) before and after anti-PD-1 therapy. (B) Dendrogram tree built using hierarchical clustering and Ward linkage on all cells using myeloid markers as in Fig 1B. Each column represents one patient sample from one time point (n patients=20, n samples=39). One baseline sample with a cell numbers <50 was excluded form further analysis.

Cox-proportional regression identifies clinical parameters associated with progression free survival

Finally, using a multivariate Cox-proportional hazards model (including all factors with $p < 0.05$ from the univariate analysis), we assessed the independent prognostic value of 53 standard clinical parameters plus the frequency of measured classical monocytes with progression-free survival (PFS, Supplementary Figure 5 and 6) in our discovery cohort. Previous targeted therapy (hazard rate: 13.9; 95%CI: 1.51-128; $P=0.02$) was identified as independent factor associated with progression under anti-PD1 therapy.

Validation of cellular immune signature by Citrus and conventional flow cytometry

As an independent validation of the computational results, we employed Citrus, which is a clustering-based supervised algorithm that identifies stratifying signatures, to compare the identified cell types and marker expression differences that could distinguish between non-responders and responders before therapy (Supplementary Figures 7 and 8). Citrus independently confirmed the lower frequency observed in the T cell compartment and the elevation of the myeloid compartment before therapy, as shown in panels 1 and 3.

To facilitate the translation of our observations into clinical practice, we designed a flow cytometry-based validation panel using a reduced number of markers. We selected a combination of markers that were significantly differentially expressed in Figures 1C and 4C and markers that define the cellular composition in the blood (Supplementary Figure 9). A blinded validation was performed on PBMCs from a second independent cohort of 31 melanoma patients containing 15 responders and 16 non-responders before anti-PD-1 therapy (Table 2).

As for the discovery cohort, we assessed in the validation cohort the correlation between commonly measured clinical factors and patients' PFS under treatment, including the monocyte frequencies from the validation panel (Supplementary Figure 10). The multivariate analysis of all the clinical variables from the independent FACS validation cohort of 31 patients revealed that classical monocytes were the most significant independent factor ($p=0.004$) associated with survival. Importantly, perhaps because this validation cohort was bigger, this significance held up in a multivariate analysis of all variables that were significant from the univariate results,

with classical monocytes having the most significant value of all 53 parameters (hazard rate: 0.461; 95%CI: 0.27-0.783; P=0.004). This was an exhaustive analysis of all features that are routinely recorded in a standard clinical setting and confirms the relevance of our discovery that the measurement of classical monocyte frequencies could provide a simple yet powerful tool to clinical practice.

Table 2 – Characteristics of melanoma patients and healthy donors used for the validation study.
Numbers in parenthesis display the age range of subjects.

Healthy donors							
N	14						
Age (years)	63.4 (46-91)						
Sex (male/female)	7/7						
Melanoma patients							
	Responder			Non-Responder			
	Patients	samples before therapy	samples after therapy	Patients	samples before therapy	samples after therapy	Sample TOTAL
N	31	15	0	16	16	0	31
Age	58.9 (31-93)			61.9 (27-89)			
Sex (male/female)	9/6			8/8			
Pre-treatments							
Radiotherapy	10/15			5/16			
Chemotherapy	0/15			4/16			
Ipilimumab	10/15			13/16			
Melanoma Inhibitor	2/15			2/16			
Other	0/15			5/16			

The data confirmed the lower frequency of T cells (CD3⁺ CD56⁻, p=1.67e-02) and the higher frequency of CD14⁺ monocytes (CD3⁻CD19⁻CD14⁺CD16⁺HLA-DR⁺, p=1.99e-06) before therapy in responders, as already shown by mass cytometry (Figure 4F). In order to visualize and assess the survival benefit conferred by a higher frequency of classical monocytes prior to treatment, we calculated the optimal cutoff point in monocytes frequency which best stratifies responders and non-responders. The calculated cutoff of 19.38% was then used to compute a cumulative hazard function for the high and low monocytes frequency groups. The resulting plot shows a clear difference in hazard between patients who have a high frequency or a low frequency of classical monocytes at baseline. Our model thus indicates that a classical monocyte

frequency higher than 19.38%, before anti-PD1 therapy initiation, is predictive of a better treatment response and survival (Figure 4G).

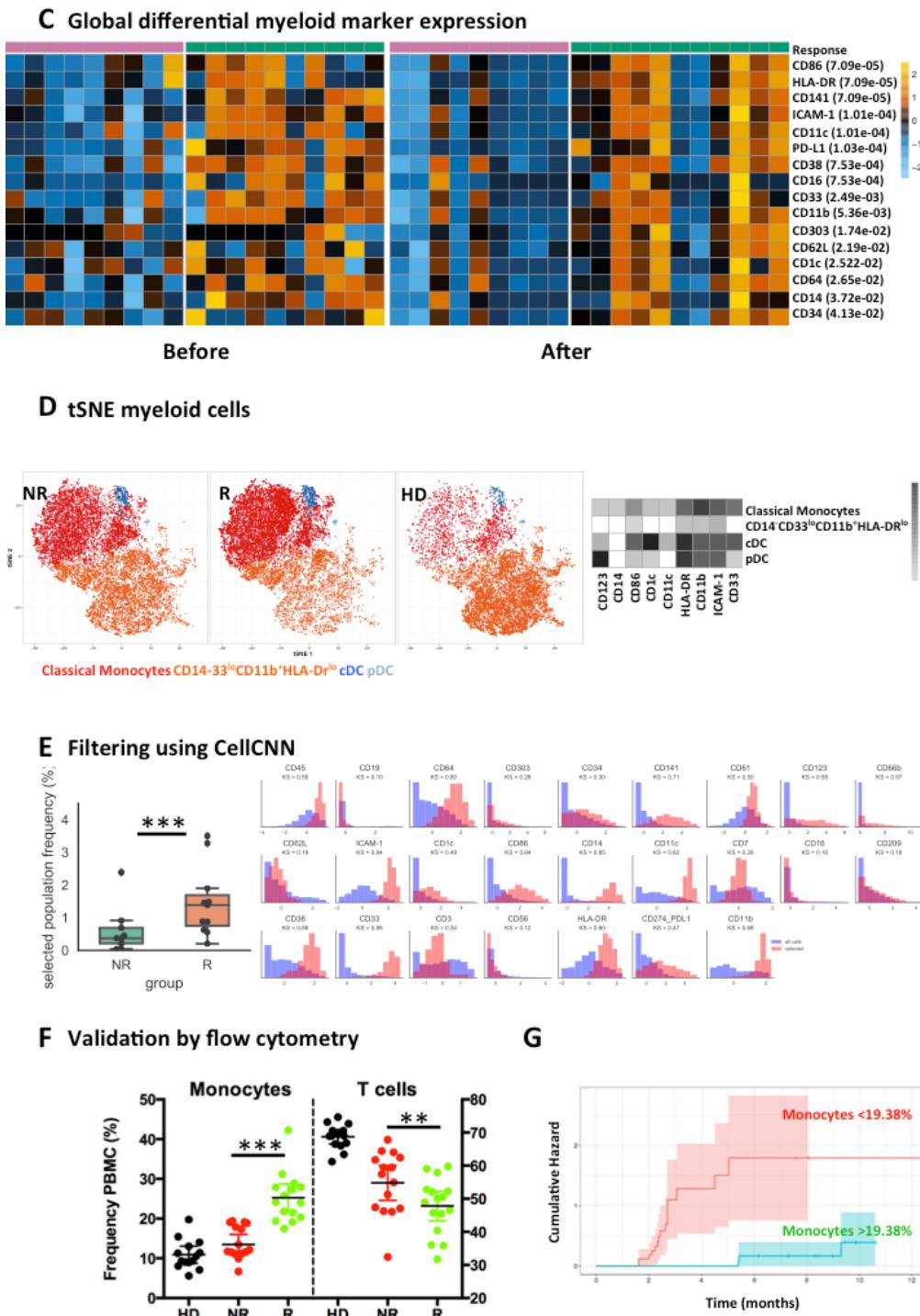


Figure 4 (continued). Patient stratification based on myeloid cell markers and expansion and enhanced activation of classical monocytes in responders. (C) Heatmap of significantly differentially expressed markers in the myeloid compartment ($p=0.05$, $CD3^-CD19^-$). Heat represents

median marker expression normalized to the mean of 0 and a standard deviation of 1. (D) Visualization of FlowSOM-generated myeloid clusters (CD3 negative CD19 negative) in non-responders (NR), responders (R) and health donors (HD) using tSNE. Per plot, 10'000 cells are displayed. CD7 and CD56 positive cells were excluded from analysis. The heatmap represents the expression of respective markers within a cellular cluster. (E) Frequency of cells discovered using CellCnn in non-responders (NR) and responders (R) with relative distribution of filtered marker expressions (all sample – blue, significant different population – red). (F) Validation of results on a second independent cohort of 31 patients using flow cytometry and CD3, CD4, CD11b, CD14, CD19, CD16, CD33, CD56, HLA-DR markers. (G) Cumulative hazard in patients with monocyte frequencies above (green) and below (red) 19.38% over time (months). Bars on top of the heatmaps represent individual samples from responders (green) versus non-responders (pink). Numbers in brackets show adjusted p-values (**p=<0.05, ***p=<0.01).

DISCUSSION

The successful therapeutic responses in patients with advanced melanoma encouraged the application of anti-PD-1 immunotherapy to several other cancers, such as non-small cell lung carcinoma (NSCLC), metastatic renal cell carcinoma, metastatic squamous NSCLC, Hodgkin's lymphoma, advanced gastric cancer, advanced bladder cancer, head and neck cancer, and triple negative breast cancer^{2,3,23,5-7,9,24,25}. Despite increasing overall survival in 33-40% of melanoma patients, anti-PD1 treatment is not effective in the majority of treated patients and results in disease progression at a median follow-up of 21 months in only 25% of patients^{26,27}. Moreover, given the broadening of its application, we can anticipate that the rate of non-responding and relapsing patients to anti-PD-1 therapy will further increase. In this context, the identification of biomarkers able to discriminate between responders and non responders before therapy initiation may tailor the application of this treatment only to those patients that are likely to benefit from it, while providing alternative treatments to the patients that are unlikely to show a response. In our study, by using single cell mass cytometry combined with a custom bioinformatics analysis, we searched for differential immune signatures in responders versus non-responders before therapy. Besides a modest alteration of the lymphocyte compartment before therapy, i.e. lymphopenia of CD4 and CD8 T cells, $\gamma\delta$ T cells and a slight elevation of NKT cells, we could clearly show that classical CD14⁺CD16⁻CD33⁺HLA-DR^{hi}

monocytes were the strongest predictor of responsiveness to anti-PD-1 immunotherapy.

The higher frequency of $CD14^+CD16^-CD33^+HLA-DR^{hi}$ classical monocytes in responders before therapy is striking. In recent years, the role of myeloid cells in cancer has been extensively debated and numerous studies have addressed the role of the so called myeloid derived suppressor cells, which have been shown to arise during chronic inflammation and cancer²⁸. However, the phenotypic, morphological and functional heterogeneity of these cells generates confusion when investigating their roles in anti-cancer immune responses.. It has been proposed that high frequencies of myeloid cells with immunosuppressive features, defined as $CD33^+CD11b^+HLA-DR^{lo/-}$, may lead to T cell dysfunction and failure to respond to immunotherapy. Accordingly, a reduction of suppressive myeloid cells correlated with an increase in the objective clinical responses and long-term survival²⁹⁻³¹. In our study, we used machine-assisted bioinformatics to define cell populations and found that the frequency of a cellular cluster most likely to resemble $CD33^{low}CD11b^+HLA-DR^{lo}$ myeloid cells shows no differences between responders and non-responders and remains constant before and after therapy. In addition, in responding patients classical monocytes ($CD14^+CD16^-$) were highly activated as shown by increased ICAM-1 and HLA-DR levels. This suggests that during anti PD-1 immunotherapy monocytes may sustain the development of an effective anti-tumor immune response similarly to what has been described for $CD14^-CD16^+$ monocytes during anti-CTLA-4 treatment³². Further support for a critical role of monocytes in anti-tumor immune responses comes from a study in which untreated melanoma patients with the highest tumor burden harbored dysregulated intermediate ($CD14^+CD16^+$) and non-classical monocytes ($CD14^-CD16^+$) characterized by a dramatic decrease of HLA-DR and inflammatory markers³³. Further, the up-regulation of PD-L1 on the monocytes from responders before therapy could be a result of the higher activation status of these cells. It is well described that IFN- γ can induce the up-regulation of PD-1 and PD-L1 on T and myeloid cells, respectively^{34,35}.

Further, in line with the hypothesis that the presence of highly activated classical monocytes may be a prerequisite for a successful response during anti-PD-1 immunotherapy we reported higher frequencies of central memory T cells and NKT

cells in circulation and a more activated (CTLA-4⁺, TNF- α ⁺, PD-1⁺, granzyme-B⁺ and IL-2⁺) T cell compartment after therapy in responding patients.

Given the shift of frequency from naïve to central memory T cells in responders before therapy and the increase in CTLA-4, IFN- γ , IL-17A, granzyme-B and PD-1 after therapy, our findings indicate that anti-PD-1 immunotherapy supports functionally activated T cells. This is in line with recent publications showing that higher levels of CTLA-4 on intra-tumoral T cells correlated with better response to anti-PD-1 treatment and that resistance to anti-PD-1 immunotherapy was associated with defects in the pathways of antigen presentation and interferon-receptor signalling^{10,34}. Indeed, besides being a regulator during T cell expansion, CTLA-4 is also a marker of activated T cells³⁶. Further, enhanced NK T cell frequencies after immunotherapy correlated with positive clinical responses in melanoma patients, while elevated frequencies of some $\gamma\delta$ T cell subsets following anti-CTLA4 treatment correlated with decreased clinical benefit^{37,38}. Lastly, we found a consistent and significant reduction of T cells in the peripheral blood of responders compared to non-responders (Fig. 2C and D). This phenomenon may be due to their enhanced ability to migrate to the tumor site³⁹. Indeed, in the CD8⁺ T cell compartment of responder patients, we also found an up-regulation of CD11a, which has been shown to be essential for migration to lymph nodes and distal sites^{36,40}. Lastly, Th17 cells were recently demonstrated to be potent apoptosis-resistant anti-tumor effector cells^{41,42}.

Altogether, we provided evidence for a responsiveness-associated immune signature in metastatic melanoma patients during anti-PD-1 immunotherapy. Future studies should confirm these signatures in larger, multi-center cohorts of melanoma patients as well as in patients with other cancer types for which anti-PD-1 treatment has been approved. Our finding might further help to elucidate the mechanism underlying anti-PD-1 activity.

ONLINE METHODS

Patient Samples

Fifty-one cryopreserved peripheral blood mononuclear cells (PBMC) samples of melanoma patients before and about 12 weeks after (median: 84 days, range: 23-162 days, average: 87.3 days) anti-PD-1 immunotherapy initiation were provided by the Department of Dermatology, University Hospital Zurich, Switzerland (see Table1). Patients were treated with 3mg/kg Nivolumab every 2 weeks or 2mg/kg Pembrolizumab every 3 weeks for 12 weeks when their clinical status was assessed again. Response was defined as the patient's disease control rate (DCR) in the course of treatment. That is, the responder group comprises every patient who showed signs of clinical benefit within the first 15 weeks of treatment, which includes partial response (PR), complete response (CR), and stable disease (SD) thus better capturing "real-world-patients". The non-responder group included every patient who discontinued treatment due to disease progression or showed signs of progression within the first 15 weeks of treatment. Progression was defined as either a measurable increase in tumor size, new metastatic sites, or the need to treat the patient with a secondary treatment such as radiotherapy.

Stimulations, stainings, and mass cytometry acquisition

PBMC stimulations, staining and acquisition by mass cytometry were performed as described previously¹⁵. Frozen PBMCs were used in this retrospective study to balance cohorts in terms of response and to reduce batch-effects through a unique bar-coding strategy. Data were stored using the Flow Repository¹⁶ which can be accessed under: <https://flowrepository.org/experiments/1124>. Full methods are included in the Supplementary Appendix.

Statistical Analysis

Data acquired by mass cytometry was normalized using the standalone MATLAB normalizer (Version 2013b)¹⁷, marker expression was controlled in FlowJo (Version10.1r5) and patient samples were de-barcoded using Boolean gating. For further analysis we developed a customized R workflow in order to discover different biomarkers when comparing marker expression between responders and non-responders. The workflow is described in the appendix and the R code can be

accessed under: https://github.com/gosianow/carsten_cytof_code. Additional biostatistical analysis using CellCnn can be found under: https://github.com/lmweber/PD1_analysis_CellCnn.

Validation by flow cytometry

Validation of the CyTOF data was done with a combination of markers with significantly different expression from the initial discovery mass cytometry approach, and markers that defined the cellular composition in blood using flow cytometry (Supplementary Figure 9). A set of PBMCs was analyzed in a blinded fashion from a second, independent cohort of 31 melanoma patients containing 15 responders and 16 non-responders before anti-PD-1 therapy. The panel is described in the Appendix. At least 100'000 live cells were acquired using Diva software on a Fortessa flow cytometer (BD) and analyzed using FlowJo software (TriStar). From FlowJo data, the frequencies of CD3⁺ T cells and CD14⁺CD16⁺HLA-DR^{hi} monocytes were extracted from the three groups (R, NR, HD). For statistical testing, we applied a generalized linear model (GLM) and cutpoint calculations as described in the appendix.

References

1. Topalian, S. L., Drake, C. G. & Pardoll, D. M. Targeting the PD-1/B7-H1(PD-L1) pathway to activate anti-tumor immunity. *Curr. Opin. Immunol.* **24**, 207–212 (2012).
2. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366**, 2443–2454 (2012).
3. Powles, T. *et al.* MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. *Nature* **515**, 558–562 (2014).
4. Brahmer, J. *et al.* Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **373**, 123–135 (2015).
5. Motzer, R. J. *et al.* Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* **373**, 1803–1813 (2015).
6. Rizvi, N. A. *et al.* Activity and safety of nivolumab, an anti-PD-1 immune checkpoint inhibitor, for patients with advanced, refractory squamous non-small-cell lung cancer (CheckMate 063): a phase 2, single-arm trial. *Lancet Oncol.* **16**, 257–265 (2015).
7. Ansell, S. M. *et al.* PD-1 blockade with nivolumab in relapsed or refractory Hodgkin's lymphoma. *N. Engl. J. Med.* **372**, 311–319 (2015).
8. Center for Drug EvaluationResearch. Approved Drugs - Hematology/Oncology (Cancer) Approvals & Safety Notifications. (2016).
9. Hamid, O. *et al.* Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N. Engl. J. Med.* **369**, 134–144 (2013).
10. Daud, A. I. *et al.* Tumor immune profiling predicts response to anti-PD-1 therapy in human melanoma. *J. Clin. Invest.* **126**, 3447–3452 (2016).
11. Dronca, R. S. *et al.* T cell Bim levels reflect responses to anti-PD-1 cancer therapy. *JCI Insight* **1**, (2016).
12. Wistuba-Hamprecht, K. *et al.* Establishing High Dimensional Immune Signatures from Peripheral Blood via Mass Cytometry in a Discovery Cohort of Stage IV Melanoma Patients. *J. Immunol.* **198**, 927–936 (2017).
13. Mair, F. *et al.* The end of gating? An introduction to automated analysis of high dimensional cytometry data. *European Journal of Immunology* **46**, 34–43 (2016).
14. Pérez-Callejo, D., Romero, A., Provencio, M. & Torrente, M. Liquid biopsy based biomarkers in non-small cell lung cancer for diagnosis and treatment monitoring. *Transl Lung Cancer Res* **5**, 455–465 (2016).
15. Hartmann, F. J. *et al.* High-dimensional single-cell analysis reveals the immune signature of narcolepsy. *J. Exp. Med.* **213**, 2621–2633 (2016).
16. Spidlen, J. & Brinkman, R. R. Use FlowRepository to share your clinical data upon study publication. *Cytometry B Clin Cytom* (2016). doi:10.1002/cyto.b.21393
17. Finck, R. *et al.* Normalization of mass cytometry data with bead standards. *Cytometry Part A* **83**, 483–494 (2013).
18. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
19. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* **87**, 636–645 (2015).
20. Weber, L. M. & Robinson, M. D. *Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data*. (2016). doi:10.1101/047613
21. Maaten, L. V. D. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
22. Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. (2016). doi:10.1101/046508
23. Brahmer, J. R. *et al.* Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumors: safety, clinical activity, pharmacodynamics, and immunologic correlates. *J. Clin. Oncol.* **28**, 3167–3175 (2010).
24. Brahmer, J. R. *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* **366**, 2455–2465 (2012).
25. Weber, J. S. *et al.* Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial. *Lancet Oncol.* **16**, 375–384 (2015).
26. Robert, C. *et al.* Pembrolizumab versus Ipilimumab in Advanced Melanoma. *N. Engl. J. Med.* **372**, 2521–2532 (2015).
27. Ribas, A. *et al.* Association of Pembrolizumab With Tumor Response and Survival Among

- Patients With Advanced Melanoma. *JAMA* **315**, 1600–1609 (2016).
- 28. Ostrand-Rosenberg, S. & Sinha, P. Myeloid-derived suppressor cells: linking inflammation and cancer. *J. Immunol.* **182**, 4499–4506 (2009).
 - 29. Gebhardt, C. *et al.* Myeloid Cells and Related Chronic Inflammatory Factors as Novel Predictive Markers in Melanoma Treatment with Ipilimumab. *Clin. Cancer Res.* **21**, 5453–5459 (2015).
 - 30. Meyer, C. *et al.* Frequencies of circulating MDSC correlate with clinical outcome of melanoma patients treated with ipilimumab. *Cancer Immunol. Immunother.* **63**, 247–257 (2014).
 - 31. Sade-Feldman, M. *et al.* Clinical Significance of Circulating CD33+CD11b+HLA-DR-Myeloid Cells in Patients with Stage IV Melanoma Treated with Ipilimumab. *Clin. Cancer Res.* (2016). doi:10.1158/1078-0432.CCR-15-3104
 - 32. Romano, E. *et al.* Ipilimumab-dependent cell-mediated cytotoxicity of regulatory T cells ex vivo by nonclassical monocytes in melanoma patients. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6140–6145 (2015).
 - 33. Chavan, R. *et al.* Untreated stage IV melanoma patients exhibit abnormal monocyte phenotypes and decreased functional capacity. *Cancer Immunology Research* **2**, 241–248 (2014).
 - 34. Zaretsky, J. M. *et al.* Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).
 - 35. Bellucci, R. *et al.* Interferon- γ -induced activation of JAK1 and JAK2 suppresses tumor cell susceptibility to NK cells through upregulation of PD-L1 expression. *Oncoimmunology* **4**, e1008824 (2015).
 - 36. Herbst, R. S. *et al.* Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).
 - 37. Ibarrondo, F. J. *et al.* Natural killer T cells in advanced melanoma patients treated with tremelimumab. *PLoS ONE* **8**, e76829 (2013).
 - 38. Wistuba-Hamprecht, K. *et al.* Proportions of blood-borne V δ 1+ and V δ 2+ T-cells are associated with overall survival of melanoma patients treated with ipilimumab. *Eur. J. Cancer* **64**, 116–126 (2016).
 - 39. Kluger, H. M. *et al.* Characterization of PD-L1 Expression and Associated T-cell Infiltrates in Metastatic Melanoma Samples from Variable Anatomic Sites. *Clin. Cancer Res.* **21**, 3052–3060 (2015).
 - 40. Andrian, von, U. H. & Mempel, T. R. Homing and cellular traffic in lymph nodes. *Nat. Rev. Immunol.* **3**, 867–878 (2003).
 - 41. Bowers, J. S. *et al.* Th17 cells are refractory to senescence and retain robust antitumor activity after long-term ex vivo expansion. *JCI Insight* **2**, e90772 (2017).
 - 42. Neitzke, D. J. *et al.* Murine Th17 cells utilize IL-2 receptor gamma chain cytokines but are resistant to cytokine withdrawal-induced apoptosis. *Cancer Immunol. Immunother.* **4**, 128–15 (2017).

Acknowledgements

We thank Drs. V. Tosevski and T.M. Brodie from the mass cytometry core facility, University Zurich, Alice Langer from Dermatology Department, University Zurich for excellent technical assistance and Drs. Bithi Chatterjee and Cornelia Gujer from the Department of Experimental Immunology, University Zurich and Alix Zollinger from the Swiss Institute of Bioinformatics, Lausanne for discussion. This work received funding from the University Research Priority Program (URPP) in Translational Cancer Research, the Swiss National Science Foundation (310030_146130 and 316030_150768 to B.B.) and the European Union FP7 project ATECT (BB).

Competing Financial Interest

The authors declare no competing financial interest.

Author's contributions

C.K., M.P.L. and B.B. conceived the study and analyzed data.

C.K., S.G. and B.B. designed and performed the experiments.

F.H. and S.G. assisted with the experiments.

S.S., R.D. and M.P.L. provided clinical samples and performed statistical analysis of clinical parameters.

M.N., L.M.W. and M.D.R. provided analysis algorithms and analyzed data.

C.K., S.G. wrote and M.P.L., M.D.R. and B.B. edited the manuscript.

All authors read and gave final approval to submit the manuscript.

Discussion and Perspectives

1 Dirichlet-multinomial regression framework

The Dirichlet-multinomial (DM) model is introduced in **Paper I**. In this section, we describe the Dirichlet-multinomial regression framework for incorporating the covariate effects in differential transcript usage (DTU) analysis. The model is presented for a given gene with q transcripts ($j = 1, \dots, q$). We assume that transcript counts $\mathbf{Y} = (Y_1, \dots, Y_q)$ follow a DM distribution with proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_q)$, concentration γ_+ and $m = \sum_{j=1}^q y_j$ treated as an ancillary statistic (as it depends on the sequencing depth and gene expression but not on the model parameters), which we denote as $\mathbf{Y}|m \sim DM(m, \gamma_+, \boldsymbol{\pi})$. Thus, the probability of observing $\mathbf{y} = (y_1, \dots, y_q)$ is defined as:

$$P_{DM}(\mathbf{y}|m, \gamma_+, \boldsymbol{\pi}) = \binom{m}{\mathbf{y}} \frac{\Gamma(\gamma_+)}{\Gamma(m + \gamma_+)} \prod_{j=1}^q \frac{\Gamma(y_j + \pi_j \gamma_+)}{\Gamma(\pi_j \gamma_+)}. \quad (1)$$

Let $\mathbf{y} = (y_{ij})_{n \times q}$ be the observed transcript count matrix for n samples ($i = 1, \dots, n$), and $\mathbf{X} = (x_{ik})_{n \times p}$ be the design matrix where each row records characteristics of sample i with respect to the p covariates. For a defined concentration parameter γ_+ , we assume that the proportion parameters π_{ij} for all $j = 1, \dots, q - 1$ depend on the covariates via the following logit-linear model:

$$\text{logit}(\pi_{ij}) = \log \left(\frac{\pi_{ij}}{\pi_{iq}} \right) = \sum_{k=1}^p x_{ik} \beta_{kj}, \quad (2)$$

where β_{kj} is a regression coefficient for transcript j with respect to the covariate k . As in the multinomial regression, there is no need for estimating all the q proportions. Since $\sum_{j=1}^q \pi_{ij} = 1$, by knowing π_{ij} for $j = 1, \dots, q - 1$, the q th value can be calculated as $\pi_{iq} = 1 - \sum_{j=1}^{q-1} \pi_{ij}$. The transcript proportions can be represented via the regression components as:

$$\pi_{ij} = \frac{\exp(\sum_{k=1}^p x_{ik} \beta_{kj})}{1 + \sum_{j=1}^{q-1} \exp(\sum_{k=1}^p x_{ik} \beta_{kj})}, \text{ for } j = 1, \dots, q - 1 \text{ and} \quad (3)$$

$$\pi_{iq} = \frac{1}{1 + \sum_{j=1}^{q-1} \exp(\sum_{k=1}^p x_{ik} \beta_{kj})}, \text{ for } j = q. \quad (4)$$

We can also represent the logit-linear model using a compact matrix notation that involves only the $q - 1$ transcripts:

$$\text{logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}, \quad (5)$$

where $\boldsymbol{\pi} = (\pi_{ij})_{n \times q-1}$ and $\boldsymbol{\beta} = (\beta_{kj})_{p \times q-1}$. The log-likelihood function, ignoring the constant part that does not involve the parameters, is given by:

$$\begin{aligned} l(\boldsymbol{\beta}; \gamma_+, \mathbf{X}, \mathbf{y}) \propto & \sum_{i=1}^n \left[\tilde{\Gamma}(\gamma_+) - \tilde{\Gamma}(m_i + \gamma_+) + \sum_{j=1}^{q-1} \{ \tilde{\Gamma}(y_{ij} + \gamma_+ \pi_{ij}) - \tilde{\Gamma}(\gamma_+ \pi_{ij}) \} + \right. \\ & \left. + \tilde{\Gamma}(y_{iq} + \gamma_+ \pi_{iq}) - \tilde{\Gamma}(\gamma_+ \pi_{iq}) \right], \end{aligned} \quad (6)$$

where $\tilde{\Gamma}(.)$ is the log gamma function. The q th components are separated from the sum because such a representation will be more convenient for the calculation of likelihood derivatives, since the partial derivatives for the proportions have different forms:

$$\frac{\partial \pi_{ij}}{\partial \beta_{kj}} = x_{ik} \pi_{ij} (1 - \pi_{ij}) = x_{ik} \pi_{ij} - x_{ik} \pi_{ij} \pi_{ij}, \quad (7)$$

$$\frac{\partial \pi_{ij}}{\partial \beta_{kj'}} = -x_{ik} \pi_{ij} \pi_{ij'}, \quad (8)$$

$$\frac{\partial \pi_{ij}}{\partial \beta_{kq}} = -x_{ik} \pi_{ij} \pi_{iq}. \quad (9)$$

The score function can be calculated as:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{kj'}} &= \sum_{i=1}^n \left[\sum_{j=1}^{q-1} \tilde{\Gamma}(y_{ij} + \gamma_+ \pi_{ij}) \gamma_+ (-x_{ik'} \pi_{ij} \pi_{ij'}) + \tilde{\Gamma}(y_{ij'} + \gamma_+ \pi_{ij'}) \gamma_+ x_{ik'} \pi_{ij'} + \right. \\ &\quad - \sum_{j=1}^{q-1} \tilde{\Gamma}(\gamma_+ \pi_{ij}) \gamma_+ (-x_{ik'} \pi_{ij} \pi_{ij'}) - \tilde{\Gamma}(\gamma_+ \pi_{ij'}) \gamma_+ x_{ik'} \pi_{ij'} + \\ &\quad \left. + \tilde{\Gamma}(y_{iq} + \gamma_+ \pi_{iq}) \gamma_+ (-x_{ik'} \pi_{ij'} \pi_{iq}) - \tilde{\Gamma}(\gamma_+ \pi_{iq}) \gamma_+ (-x_{ik'} \pi_{ij'} \pi_{iq}) \right] \quad (10) \\ &= \sum_{i=1}^n \left[\gamma_+ x_{ik'} \pi_{ij'} \left\{ - \sum_{j=1}^q \tilde{\Gamma}(y_{ij} + \gamma_+ \pi_{ij}) \pi_{ij} + \tilde{\Gamma}(y_{ij'} + \gamma_+ \pi_{ij'}) + \right. \right. \\ &\quad \left. \left. + \sum_{j=1}^q \tilde{\Gamma}(\gamma_+ \pi_{ij}) \pi_{ij} - \tilde{\Gamma}(\gamma_+ \pi_{ij'}) \right\} \right], \end{aligned}$$

where $\tilde{\Gamma}(.)$ is the first derivative of the log gamma function. We denote $g(\boldsymbol{\beta}) = -\sum_{j=1}^q \tilde{\Gamma}(y_{ij} + \gamma_+ \pi_{ij}) \pi_{ij} + \tilde{\Gamma}(y_{ij'} + \gamma_+ \pi_{ij'}) + \sum_{j=1}^q \tilde{\Gamma}(\gamma_+ \pi_{ij}) \pi_{ij} - \tilde{\Gamma}(\gamma_+ \pi_{ij'})$. Then the Hessian can be calculated as:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{k'j'} \partial \beta_{k''j''}} = & \sum_{i=1}^n \left[\gamma_+ x_{ik'} \left\{ -x_{ik''} \pi_{ij'} \pi_{ij''} g(\boldsymbol{\beta}) + \right. \right. \\
& + \pi_{ij'} \left(- \sum_{j=1}^q \tilde{\tilde{\Gamma}}(y_{ij} + \gamma_+ \pi_{ij})(-x_{ik''} \pi_{ij} \pi_{ij''}) \gamma_+ \pi_{ij} - \tilde{\tilde{\Gamma}}(y_{ij''} + \gamma_+ \pi_{ij''}) x_{ik''} \pi_{ij''} \gamma_+ \pi_{ij''} + \right. \\
& - \sum_{j=1}^q \tilde{\tilde{\Gamma}}(y_{ij} + \gamma_+ \pi_{ij})(-x_{ik''} \pi_{ij} \pi_{ij''}) - \tilde{\tilde{\Gamma}}(y_{ij''} + \gamma_+ \pi_{ij''}) x_{ik''} \pi_{ij''} + \\
& + \tilde{\tilde{\Gamma}}(y_{ij'} + \gamma_+ \pi_{ij'})(-x_{ik''} \pi_{ij'} \pi_{ij''}) \gamma_+ + \\
& + \sum_{j=1}^q \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij})(-x_{ik''} \pi_{ij} \pi_{ij''}) \gamma_+ \pi_{ij} + \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij''}) x_{ik''} \pi_{ij''} \gamma_+ \pi_{ij''} + \\
& + \sum_{j=1}^q \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij})(-x_{ik''} \pi_{ij} \pi_{ij''}) + \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij''}) x_{ik''} \pi_{ij''} + \\
& \left. \left. - \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij'})(-x_{ik''} \pi_{ij'} \pi_{ij''}) \gamma_+ \right) \right\} \right], \tag{11}
\end{aligned}$$

where $\tilde{\tilde{\Gamma}}(\cdot)$ is the second order derivative of the log gamma function. The diagonal elements of the Hessian contain extra terms in the first, fourth and last line of the following equation:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{k'j'} \partial \beta_{k'j'}} = & \sum_{i=1}^n \left[\gamma_+ x_{ik'} \left\{ (-x_{ik'} \pi_{ij'} \pi_{ij''} + x_{ik'} \pi_{ij'}) g(\boldsymbol{\beta}) + \right. \right. \\
& + \pi_{ij'} \left(- \sum_{j=1}^q \tilde{\tilde{\Gamma}}(y_{ij} + \gamma_+ \pi_{ij})(-x_{ik'} \pi_{ij} \pi_{ij'}) \gamma_+ \pi_{ij} - \tilde{\tilde{\Gamma}}(y_{ij'} + \gamma_+ \pi_{ij'}) x_{ik'} \pi_{ij'} \gamma_+ \pi_{ij'} + \right. \\
& - \sum_{j=1}^q \tilde{\tilde{\Gamma}}(y_{ij} + \gamma_+ \pi_{ij})(-x_{ik'} \pi_{ij} \pi_{ij'}) - \tilde{\tilde{\Gamma}}(y_{ij'} + \gamma_+ \pi_{ij'}) x_{ik'} \pi_{ij'} + \\
& + \tilde{\tilde{\Gamma}}(y_{ij'} + \gamma_+ \pi_{ij'})(-x_{ik'} \pi_{ij'} \pi_{ij'} + x_{ik'} \pi_{ij'}) \gamma_+ + \\
& + \sum_{j=1}^q \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij})(-x_{ik'} \pi_{ij} \pi_{ij'}) \gamma_+ \pi_{ij} + \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij'}) x_{ik'} \pi_{ij'} \gamma_+ \pi_{ij'} + \\
& + \sum_{j=1}^q \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij})(-x_{ik'} \pi_{ij} \pi_{ij'}) + \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij''}) x_{ik'} \pi_{ij''} + \\
& \left. \left. - \tilde{\tilde{\Gamma}}(\gamma_+ \pi_{ij'})(-x_{ik'} \pi_{ij'} \pi_{ij'} + x_{ik'} \pi_{ij'}) \gamma_+ \right) \right\} \right]. \tag{12}
\end{aligned}$$

The regression coefficients $\boldsymbol{\beta}$ are a matrix of dimension $p \times q - 1$. However, for the optimization purposes, it is more convenient to represent them in a vector form. Thanks to that, partial derivatives in the score function can also be organized as a vector, and the Hessian can be represented as a matrix. Let $\tilde{\boldsymbol{\beta}}$ be a concatenated matrix of regression coefficients of dimension $p(q - 1) \times 1$:

$$\tilde{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{q-1} \end{pmatrix} \quad \beta_j = \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \\ \vdots \\ \beta_{pj} \end{pmatrix}. \quad (13)$$

In a similar fashion, we concatenate the matrix of observed counts \mathbf{y} and its expected values π :

$$\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{q-1} \end{pmatrix} \quad \mathbf{y}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{nj} \end{pmatrix} \quad \tilde{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{q-1} \end{pmatrix} \quad \pi_j = \begin{pmatrix} \pi_{1j} \\ \pi_{2j} \\ \vdots \\ \pi_{nj} \end{pmatrix}. \quad (14)$$

To be able to reconstruct the matrix operations that represent the model, a new design matrix of dimension $n(q - 1) \times p(q - 1)$ has to be created:

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X} \end{pmatrix}. \quad (15)$$

We can again write the model, this time using the new concatenated objects:

$$\text{logit}(\tilde{\pi}) = \tilde{\mathbf{X}}\tilde{\beta}. \quad (16)$$

The first order derivatives can be organized in a $p \times q - 1$ matrix:

$$\mathbf{S} = \mathbf{X}^\top \varphi, \quad (17)$$

where $\varphi = (\varphi_{ij'})_{n \times q - 1}$ and from equation (10):

$$\varphi_{ij'} = \gamma_+ \pi_{ij'} \left\{ - \sum_{j=1}^q \tilde{\Gamma}(y_{ij} + \gamma_+ \pi_{ij}) \pi_{ij} + \tilde{\Gamma}(y_{ij'} + \gamma_+ \pi_{ij'}) + \sum_{j=1}^q \tilde{\Gamma}(\gamma_+ \pi_{ij}) \pi_{ij} - \tilde{\Gamma}(\gamma_+ \pi_{ij'}) \right\}. \quad (18)$$

Again a concatenation strategy can be applied where:

$$\tilde{\mathbf{S}} = \begin{pmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_{q-1} \end{pmatrix} \quad \mathbf{S}_j = \begin{pmatrix} S_{1j} \\ S_{2j} \\ \vdots \\ S_{pj} \end{pmatrix} \quad \tilde{\varphi} = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{q-1} \end{pmatrix} \quad \varphi_j = \begin{pmatrix} \varphi_{1j} \\ \varphi_{2j} \\ \vdots \\ \varphi_{nj} \end{pmatrix}. \quad (19)$$

Then the score can also be calculated as:

$$\tilde{\mathbf{S}} = \tilde{\mathbf{X}}^\top \tilde{\varphi}. \quad (20)$$

The Hessian can be represented as a matrix $\tilde{\mathbf{H}}$ of dimension $p(q - 1) \times p(q - 1)$:

$$\tilde{\mathbf{H}} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{W}} \tilde{\mathbf{X}}, \quad (21)$$

where

$$\tilde{\mathbf{W}} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \cdots & \mathbf{W}_{1(q-1)} \\ \mathbf{W}_{21} & \mathbf{W}_{22} & \cdots & \mathbf{W}_{2(q-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{(q-1)1} & \mathbf{W}_{(q-1)2} & \cdots & \mathbf{W}_{(q-1)(q-1)} \end{pmatrix}. \quad (22)$$

The sub-matrices $\mathbf{W}_{j'j''}$ are diagonal matrices of order $n \times n$, and based on equations (11) and (12) their diagonal elements are equal to:

$$\begin{aligned} \text{diag}(\mathbf{W}_{j'j''})_i = \gamma_+ \Bigg\{ & (-\pi_{ij'}\pi_{ij''} + \delta_{j'j''}\pi_{ij'})g(\beta) + \\ & + \pi_{ij'} \left(- \sum_{j=1}^q \tilde{\tilde{\Gamma}}(y_{ij} + \gamma_+\pi_{ij})(-\pi_{ij}\pi_{ij''})\gamma_+\pi_{ij} - \tilde{\tilde{\Gamma}}(y_{ij''} + \gamma_+\pi_{ij''})\pi_{ij''}\gamma_+\pi_{ij''} + \right. \\ & - \sum_{j=1}^q \tilde{\tilde{\Gamma}}(y_{ij} + \gamma_+\pi_{ij})(-\pi_{ij}\pi_{ij''}) - \tilde{\tilde{\Gamma}}(y_{ij''} + \gamma_+\pi_{ij''})\pi_{ij''} + \\ & + \tilde{\tilde{\Gamma}}(y_{ij'} + \gamma_+\pi_{ij'})(-\pi_{ij'}\pi_{ij''} + \delta_{j'j''}\pi_{ij'})\gamma_+ + \\ & + \sum_{j=1}^q \tilde{\tilde{\Gamma}}(\gamma_+\pi_{ij})(-\pi_{ij}\pi_{ij''})\gamma_+\pi_{ij} + \tilde{\tilde{\Gamma}}(\gamma_+\pi_{ij''})\pi_{ij''}\gamma_+\pi_{ij''} + \\ & + \sum_{j=1}^q \tilde{\tilde{\Gamma}}(\gamma_+\pi_{ij})(-\pi_{ij}\pi_{ij''}) + \tilde{\tilde{\Gamma}}(\gamma_+\pi_{ij''})\pi_{ij''} + \\ & \left. - \tilde{\tilde{\Gamma}}(\gamma_+\pi_{ij'})(-\pi_{ij'}\pi_{ij''} + \delta_{j'j''}\pi_{ij'})\gamma_+ \right\}, \end{aligned} \quad (23)$$

where $\delta_{j'j''}$ is the Kronecker delta which equals 1 if $j' = j''$ and 0 otherwise.

Matrices $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{X}}$ are rather large, of orders, $n(q-1) \times n(q-1)$ and $n(q-1) \times p(q-1)$, respectively, but are otherwise quite sparse and possess a neat structure. It is computationally inefficient to calculate the Hessian as proposed in equation (21). Instead, another approach can be chosen which exploits this structured sparsity similarly as in the multinomial logit models [1, 2]. The Hessian matrix can be represented of blocks of derivatives with respect to chunks of coefficients:

$$\tilde{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \cdots & \mathbf{H}_{1(q-1)} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \cdots & \mathbf{H}_{2(q-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{(q-1)1} & \mathbf{H}_{(q-1)2} & \cdots & \mathbf{H}_{(q-1)(q-1)} \end{pmatrix}, \quad (24)$$

where each of the blocks can be computed independently of the other blocks:

$$\mathbf{H}_{j'j''} = \mathbf{X}^\top \mathbf{W}_{j'j''} \mathbf{X}. \quad (25)$$

Additionally, for the matrix operations in R, it is not necessary to use the entire matrices $W_{j'j''}$, as the same results can be obtained by organizing the diagonal elements of those matrices in vectors, which further reduces the memory usage. By observing the symmetry property of the Hessian blocks $\mathbf{H}_{j'j''} = \mathbf{H}_{j''j'}^\top$, the computations can be further reduced, as we only need to compute roughly half of the blocks.

In the current implementation of *DRIMSeq*, the regression coefficients are estimated by maximizing the log-likelihood function with the general optimization methods available in R. Specifically, we use the `optim` function with the BFGS method, which builds up a picture of the surface to be optimized from the log-likelihood and score function values. The Hessian, on the other hand, is necessary only for calculating the Cox-Reid adjustment term in the profile likelihood function:

$$APL(\gamma_+) = l(\gamma_+; \boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) - \frac{1}{2} \log |nI|, \quad (26)$$

where $|.|$ denotes the determinant, and I is the Fisher information matrix for $\boldsymbol{\beta}$, which is equal to $-\tilde{\mathbf{H}}$.

The optimization problem could also be solved with the Newton-Raphson (NR) method. The Newton-Raphson algorithm is an iterative procedure where each individual update has the following form:

$$\tilde{\boldsymbol{\beta}}^{new} = \tilde{\boldsymbol{\beta}}^{old} - \left(\frac{\partial^2 l(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}^\top} \right)^{-1} \frac{\partial l(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}, \quad (27)$$

where $\tilde{\boldsymbol{\beta}}^{old}$ is a vector of initial approximations, and the derivatives are evaluated at $\tilde{\boldsymbol{\beta}}^{old}$. Using matrix notation it can also be represented as:

$$\tilde{\boldsymbol{\beta}}^{new} = \tilde{\boldsymbol{\beta}}^{old} - \tilde{\mathbf{H}}^{-1} \tilde{\mathbf{S}}. \quad (28)$$

The vector $-\tilde{\mathbf{H}}^{-1} \tilde{\mathbf{S}}$ is called a full Newton step. Usually, the log-likelihood increases after each iteration. However, if the value becomes smaller, then a half step can be used as an update. This line search procedure is repeated with a half of the previous step until the new log-likelihood is not lower than the value for $\tilde{\boldsymbol{\beta}}^{old}$. Applying such a line search procedure guarantees convergence of the NR algorithm.

2 Transcript-level analysis with the beta-binomial model

With the Dirichlet-multinomial model we are able to identify genes with DTU (gene-level analysis), meaning genes for which all or part of the transcripts are expressed at different ratios between conditions. However, this model does not indicate what transcripts actually change. To do so, we provide also a transcript-level analysis. In this case, each transcript is modeled separately assuming the marginal distribution of counts for a given transcript is a beta-binomial distribution, which is a one-dimensional version of Dirichlet-multinomial, as the binomial and beta distributions are univariate versions of the multinomial and Dirichlet distributions, respectively. Based on the fact that when $(Y_1, \dots, Y_q) \sim DM(m, \gamma_+, \pi_1, \dots, \pi_q)$, where $m = \sum_{j=1}^q Y_j$, then $Y_j \sim BB(m, \gamma_+, \pi_j)$ for $j = 1, \dots, q$ [3], we do not need to re-estimate the beta-binomial parameters, only the likelihoods for each transcript need to be recalculated.

DRIMSeq returns gene-level and transcript-level p-values that can be used as input to a stage-wise testing procedure [4] (implemented in the *stageR* package) as screening and confirmation p-values, respectively. As pointed by the authors of *stageR*, interpreting both gene-level and transcript-level adjusted p-values does not provide appropriate false discovery rate (FDR) control and should be avoided. However, applying a stage-wise testing provides a useful biological interpretation of these results and improved statistical performance. Such an approach provides increased power to identify transcripts that are actually differentially used in a gene detected as gene with DTU.

In short, the procedure consists of a screening stage and a confirmation stage. In the screening stage, gene-level BH-adjusted p-values are screened to detect genes for which the hypothesis of interest is rejected. Only those genes are further considered in the confirmation stage, where for each gene separately, transcript-level p-values are adjusted to control for the family-wise error rate (FWER) and Benjamini–Hochberg adjusted significance level of the screening stage.

3 Incorporating uncertainty of transcript abundance estimates

Transcript abundance estimation is a challenging task due to the reads aligned to sequences shared by multiple transcripts, which is a prevalent phenomenon in alternatively spliced transcriptomes. It is not possible to unambiguously tell which transcript such reads originated from. Hence, they are a reason for uncertainty in transcript quantifications. When such quantifications are then used in the downstream differential analysis, it is somewhat intuitive that the variability of the fold change estimates for the high uncertainty transcripts should be larger than for the low uncertainty ones.

Currently, the problem of assigning multi-matching reads to transcripts is approached in a probabilistic manner, for example, using the EM algorithm (*Cufflinks* [5], *RSEM* [6]) or a Bayesian approach via MCMC algorithm (*BitSeq* [7]). Some of the methods for differential analysis, such as *Cuffdiff2* [8], *MetaDiff* [9], *EBSeq* [10], *BitSeq* are able to incorporate the uncertainty of transcript abundance estimates in their significance calculations. For example, in *Cuffdiff2*, the differential gene and transcript analyses are based on the beta negative-binomial model, which combines the estimates of the uncertainty and the cross-replicate variability of the transcript quantifications obtained by *Cufflinks*.

Recently, a new generation of methods was introduced to directly estimate transcript abundance from raw reads without relying on the computationally costly read alignment step (e.g. *Sailfish* [11], *Salmon* [12], *kallisto* [13]). The substantial gain in speed allows them to effectively generate so called bootstrap quantifications obtained from transcript abundance re-estimation based on resampled (with replacement) reads. These bootstrap samples may serve to estimate the transcript abundance uncertainty for each original sample. The naturally arising challenge is: how to propagate such information into the differential analysis. A few approaches already exist. *Sleuth* [14] makes use of the bootstrap to directly estimate the inferential variance, which is then passed to a response error measurement model. This model is defined for the log-transformed transcript counts, which are assumed to follow a normal distribution. The very recent, tool *RATs* [14] uses the G-test of independence for DTU between two conditions. The same test is applied to the bootstrap samples to asses the reproducibility of identified DTU calls.

Estimates of transcript counts could potentially be used as input in count-based methods, such as *edgeR* [15, 16] or *DESeq2* [17], designed for differential gene expression analysis. Those methods offer many adjustments, which are necessary for accurate gene-count modeling (e.g. dispersion moderation). However, as the gene-level counts are a result of direct counting (not estimation), they are not adapted to incorporate the uncertainty that is present in transcript count estimates. The effect of the uncertainty for DTE analysis is not yet well understood. Methods that can handle it are theoretically superior, however their performance on real data is unknown. Applying *DEXSeq* [18], originally designed to model exon counts, to transcript quantifications showed good performance in DTU analysis [19]. So far, *DRIMSeq* does not incorporate the estimation uncertainty, but accounting for it is one of the future goals. It seems reasonable to employ the bootstrap samples for that. One of the ideas is to use a Wald test and incorporate the bootstrap uncertainty into it, but this strategy requires deeper evaluation.

4 DRIMSeq application to other types of multivariate data

By allowing regression models, gene-level and feature-level analysis, *DRIMSeq* has become a powerful tool for differential transcript usage analysis based on RNA-seq data. Moreover, the general structure of the developed DM framework makes it applicable to other genomic data with multivariate count outcomes, such as PolyA-seq data which quantifies the usage of multiple RNA polyadenylation sites [20]. As the DM distribution is a multivariate generalization of the beta-binomial distribution, *DRIMSeq* could be applied to settings where the beta-binomial is already used with the advantage of being adjusted for small-sample size datasets. Those analyses involve differential methylation using bisulphite sequencing data, where counts of methylated and unmethylated cytosines at specific genomic loci are compared [21, 22, 23], or allele-specific gene expression, where the expression of two alleles are compared across experimental groups [24, 25, 26].

Another potential application could be in the differential abundance analysis of cell populations from HDCyto data. We could observe that a simple binomial distribution (logistic regression) is not able to account for the overdispersion observed in this type of data. In the proposed workflow, we approach this problem by applying mixed models with observation-level random effects. However, overdispersion in binomial data could be also accounted for with the beta-binomial model [27]. In *DRIMSeq*, cell clusters could be seen as transcripts of a single gene. Using the DM model, one could identify whether any of the clusters is differentially abundant, or by employing the BB model, investigate each of the cell populations separately. Of course, any of the proposed applications requires deeper evaluation.

References

- [1] Asad Hasan, Zhiyu Wang, and Alireza Mahani. Fast Estimation of Multinomial Logit Models: R Package *mnlogit*. *Journal of Statistical Software*, 75(1):1–24, 2016.
- [2] Scott A Czepiel. Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. *Class Notes*, pages 1–23, 2012.
- [3] Peter J Danaher. Parameter estimation for the dirichlet-multinomial distribution using

-
- supplementary beta-binomial data. *Communications in Statistics - Theory and Methods*, 17(6):1777–1788, jan 1988.
- [4] Koen Van den Berge, Charlotte Soneson, Mark D Robinson, and Lieven Clement. A general and powerful stage-wise testing procedure for differential expression and differential transcript usage. *bioRxiv*, feb 2017.
 - [5] Cole Trapnell, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
 - [6] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12:323, jan 2011.
 - [7] Peter Glaus, Antti Honkela, and Magnus Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.
 - [8] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, 2013.
 - [9] Cheng Jia, Weihua Guan, Amy Yang, Rui Xiao, W H Wilson Tang, Christine S Moravec, Kenneth B Margulies, Thomas P Cappola, Chun Li, and Mingyao Li. MetaDiff: differential isoform expression analysis using random-effects meta-regression. *BMC Bioinformatics*, 16(1):208, 2015.
 - [10] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics Oxford England*, 29(8):1035–1043, 2013.
 - [11] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5):462–4, 2014.
 - [12] Rob Patro, Geet Duggal, and Carl Kingsford. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*, page 021592, 2015.
 - [13] Nicolas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, advance on, 2016.
 - [14] Harold J Pimentel, Nicolas Bray, Suzette Puente, Pál Melsted, and Lior Pachter. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv*, jun 2016.
 - [15] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, 2010.
 - [16] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.
-

-
- [17] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.
 - [18] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, 2012.
 - [19] Charlotte Soneson, Katarina L Matthes, Małgorzata Nowicka, Charity W Law, and Mark D Robinson. Isoform prefILTERING improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, 17(1):1–15, 2016.
 - [20] Adnan Derti, Philip Garrett-Engel, Kenzie D. MacIsaac, Richard C. Stevens, Shreedharan Sriram, Ronghua Chen, Carol A. Rohl, Jason M. Johnson, and Tomas Babak. A quantitative atlas of polyadenylation in five mammals. *Genome Research*, 22(6):1173–1183, 2012.
 - [21] Hao Feng, Karen N Conneely, and Hao Wu. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69, 2014.
 - [22] Yongseok Park, Maria E Figueroa, Laura S Rozek, and Maureen A Sartor. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics (Oxford, England)*, pages 1–8, 2014.
 - [23] Deqiang Sun, Yuanxin Xi, Benjamin Rodriguez, Hyun Jung Park, Pan Tong, Mira Meong, Margaret A Goodell, and Wei Li. MOABS: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):R38, 2014.
 - [24] Daniel A Skelly, Marnie Johansson, Jennifer Madeoy, Jon Wakefield, and Joshua M Akey. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, 21(10):1728–1737, oct 2011.
 - [25] Oleg Mayba, Houston N Gilbert, Jinfeng Liu, Peter M Haverty, Suchit Jhunjhunwala, Zhaoshi Jiang, Colin Watanabe, and Zemin Zhang. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biology*, 15(8):405, 2014.
 - [26] Rong Lu, Ryan M Smith, Michal Seweryn, Danxin Wang, Katherine Hartmann, Amy Webb, Wolfgang Sadee, and Grzegorz A Rempala. Analyzing allele specific RNA expression using mixture models. *BMC Genomics*, 16(1):566, 2015.
 - [27] Xavier A Harrison. A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ*, 3:e1114, jul 2015.