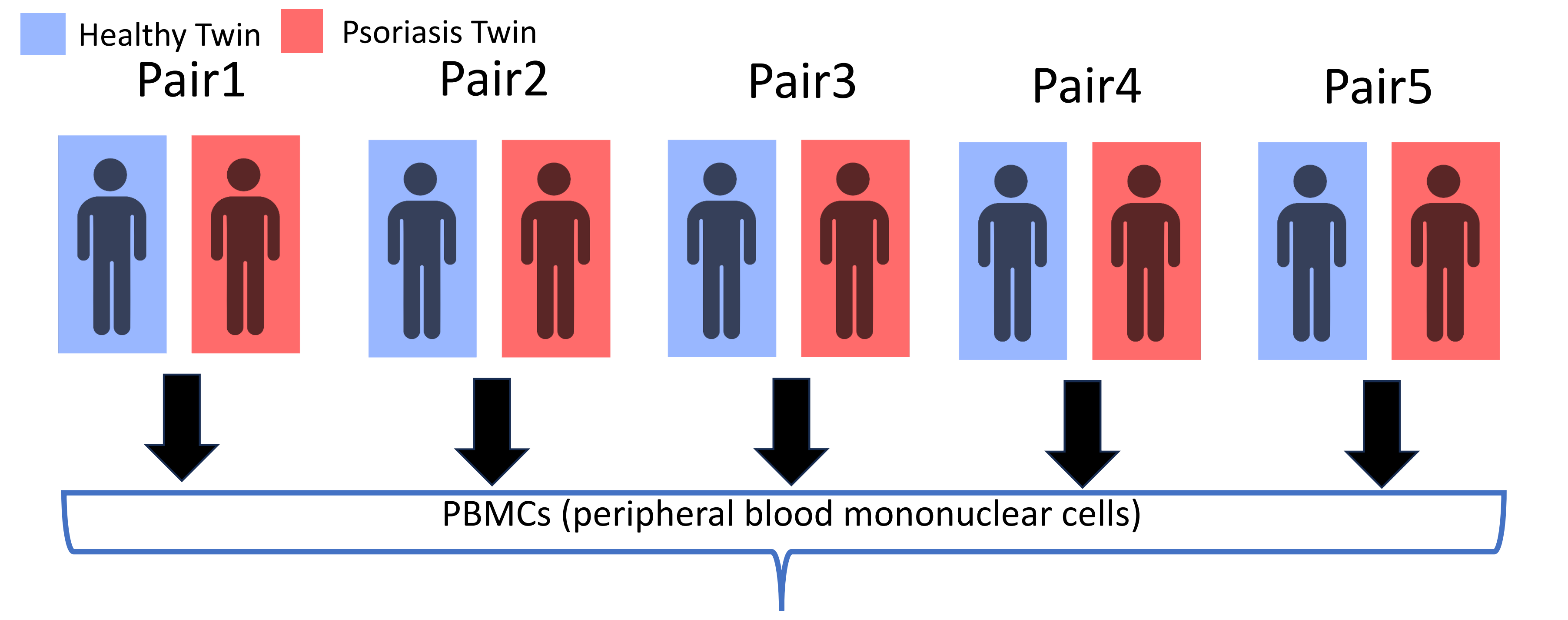
SMTB
hhmi Howard Hughes Medical Institute
ZIMIN FOUNDATION
UNIVERSITY OF TARTU 1632 UNIVERSITAS TARTUENSIS

Differential gene expression analysis of immune cells in monozygotic twins discordant for Psoriasis

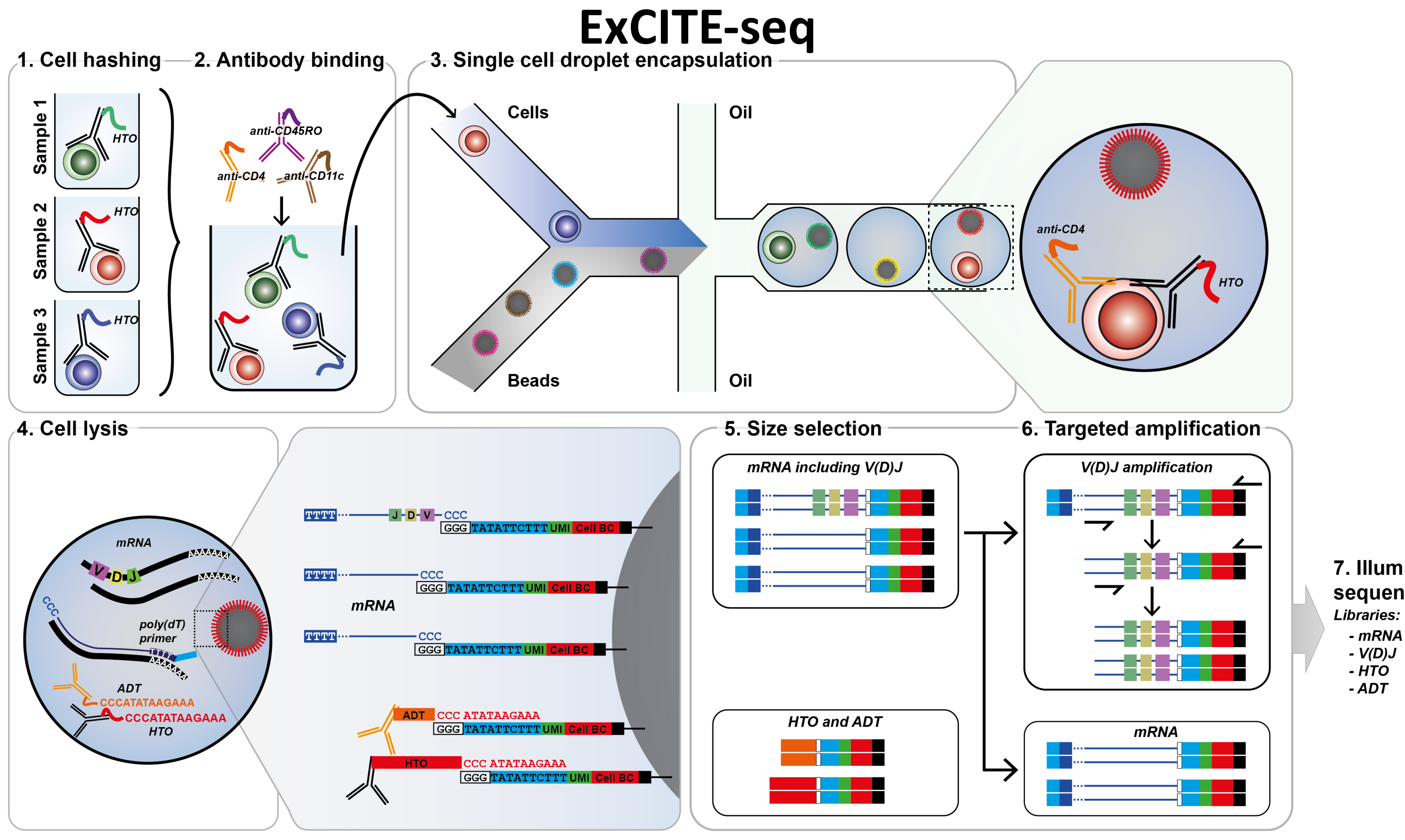
Introduction

From the very beginning of the development of genetic research technologies, monozygotic twins have played a key role in them, and ours was no exception. The object of our study is human autoimmune diseases - our goal is to identify genes whose expression changes have a significant impact on the development of diseases such as psoriasis. We aim to expand human understanding of the associated and causal regulation of genomic DNA for these diseases. As a basis for data collection, as mentioned above, we chose pairs of monozygotic twins discordant for psoriasis and performed single cell **ExCITE-seq** (**Ex**pand **C**ellular **I**ndexing of **T**ranscriptomes and **E**pitopes by sequencing). Single-cell sequencing allows for the comparison of gene expression of similar cell types. Our use of identical twins gives us a unique possibility to control for genomic DNA and therefore isolate transcriptional differences that may contribute to disease. Despite this, our data is not perfect: the batch effect, the twin effect, and the difference in sequencing depth require a careful and sensitive statistical approach to the data in order to reduce minor errors caused by the inaccuracies of the data collection methodology used. In this study we performed a primary and secondary analysis on single-cell RNA data for immune cells from monozygotic twins discordant for Psoriasis.

Study Design:



ExCITE-seq



Primary Analysis

Methods: Twin pairs were integrated and clustered using Harmony reduction. Following integration and cell phenotyping, RNA counts were normalized pairwise using Seurat's SCT Transform Function. We then performed differential gene expression for each twin pair comparing diseased twin to healthy for each immune cell population. Differential genes associated with psoriasis within each twin pair, were then compared across twin pair.

Results:

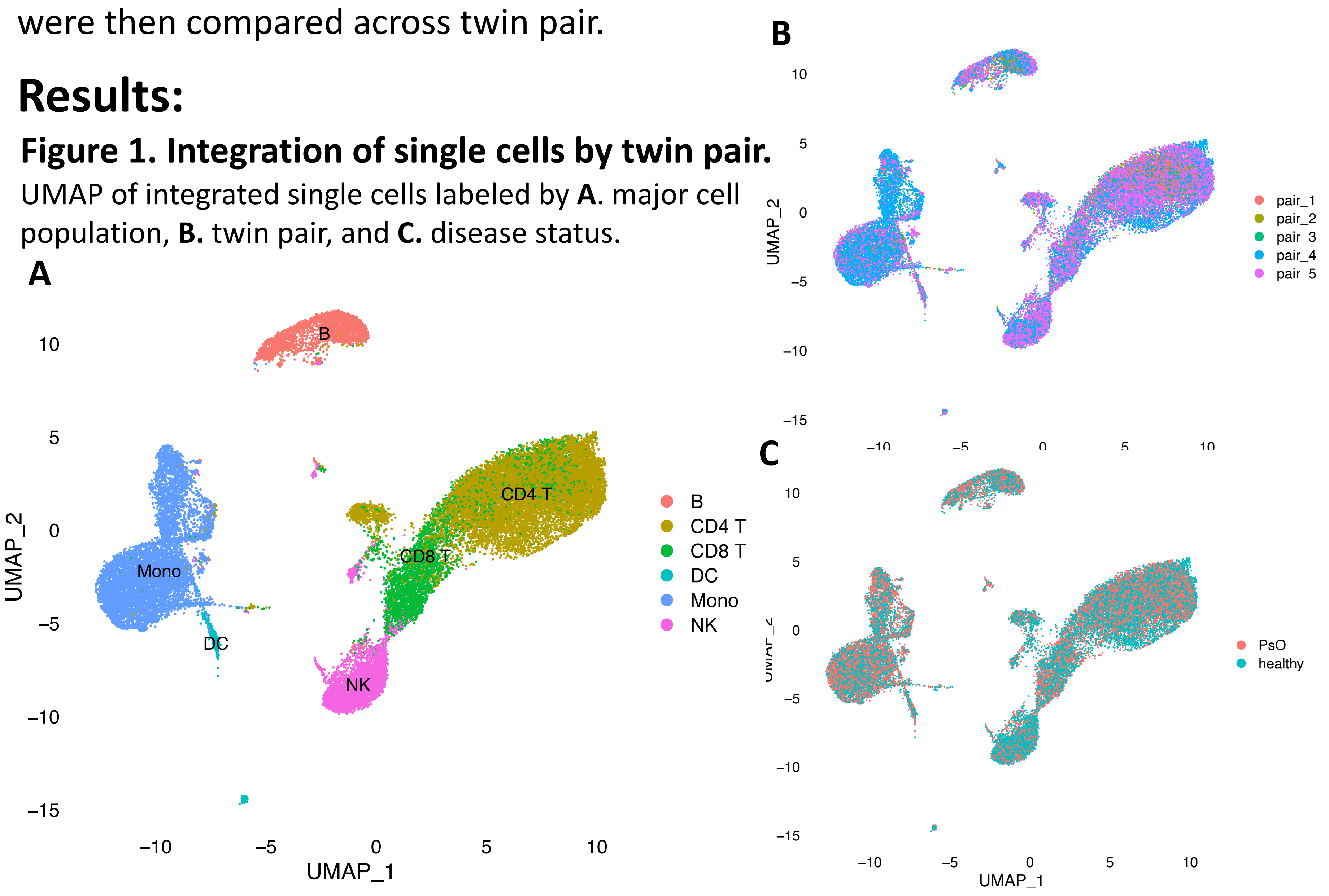
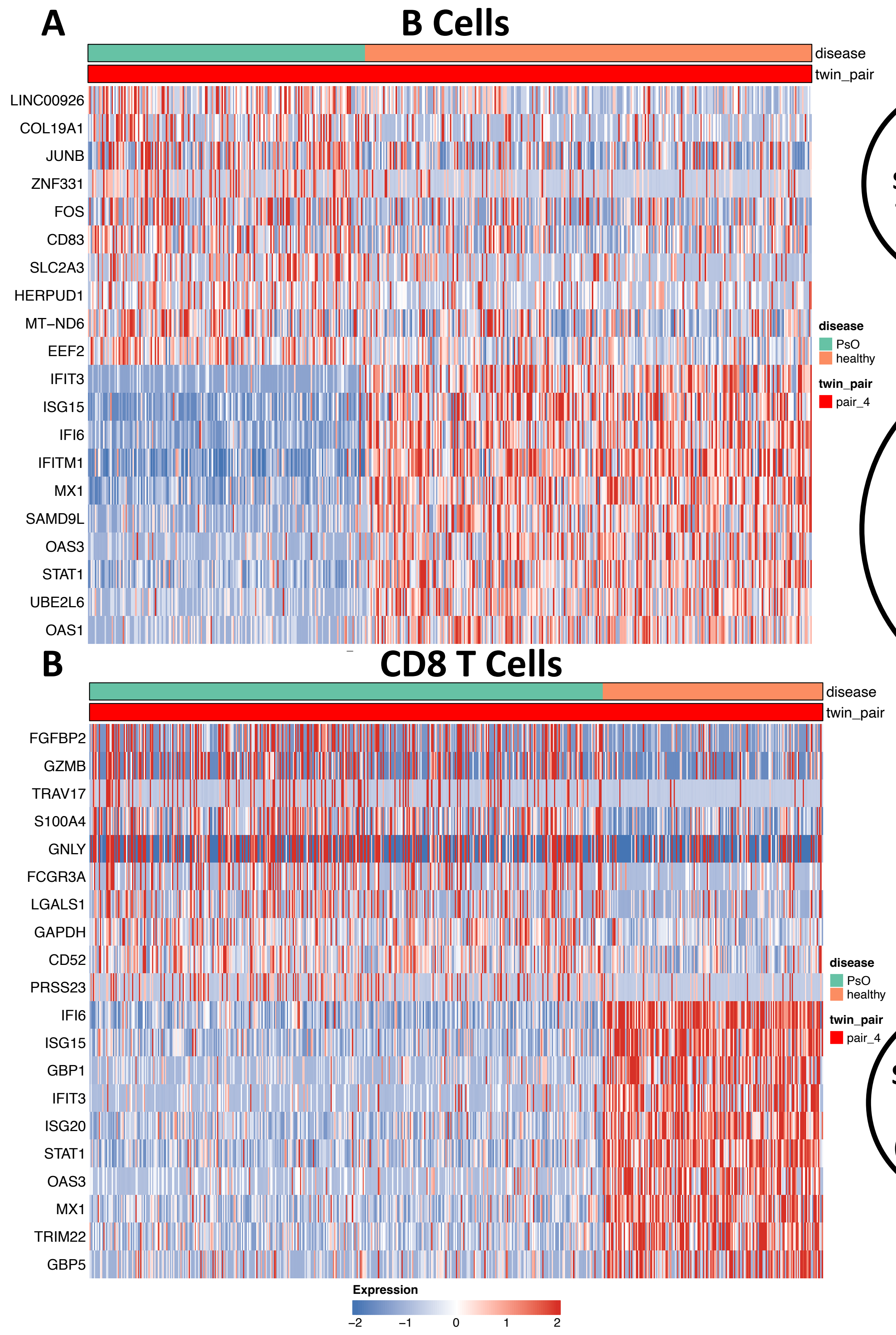


Figure 1. Integration of single cells by twin pair. UMAP of integrated single cells labeled by **A.** major cell population, **B.** twin pair, and **C.** disease status.

Primary Analysis

Figure 2. Differential gene expression in discordant psoriatic twins



C

CD4 T Cells
Shared DGEs:
TXNIP (Pair1, Pair5)

B Cells
Shared DGEs:
TXNIP (Pair1, Pair3, Pair5)
MARCH1 (Pair1, Pair5)

CD8 T Cells
Shared DGEs:
GZMH (Pair2, Pair4)
GZMB (Pair2, Pair4)
LITAF (Pair2, Pair4)
PIK3IP1 (Pair4, Pair5)
SERINC5 (Pair4, Pair5)
FOXP1 (Pair4, Pair5)

NK Cells
Shared DGEs:
HLA-DPB1 (Pair1, Pair3, Pair4)
S100A4 (Pair1, Pair4)
LDHB (Pair2, Pair4)
RPS16 (Pair2, Pair4)
KLRB1 (Pair4, Pair5)
SELL (Pair4, Pair5)

DCs
Shared DGEs:
TSC22D3
(Pair3, Pair5)

Monocytes
Shared DGEs:
TXNIP (Pair1, Pair5)
HLA-DPA1 (Pair2, Pair3)
CD52 (Pair3, Pair4)
FOS (Pair4, Pair5)

A Heatmap of the top differentially expressed genes in B cells across twin pair 4. **B.** Heatmap of the top differentially expressed genes in CD8 T ells across twin pair 4. **C.** Summarization by cell population of the top upregulated genes associated with psoriasis that are shared in more than one twin pair.

Secondary Analysis

The problem I had to solve was the need to create a data structure in which the data will not mix and have even a minimal impact on their statistical distribution and, as a result, further analysis, but without using Hilbert space tenders.
In order to achieve at least a somewhat similar picture, I proposed to use the multimetric space as the mathematical basis for further methodology. The multimetric space, as you might guess, is a structure built on a multiset, the use of which is the main idea.

A multiset is a structure consisting, of course, of a set and a function that assigns to each element of this set the number of repetitions of this element. Of course, with the necessary add-ons, such a structure is an excellent basis for working with single cell RNA-seq data, as it offers a wider range of mathematical operations available. Basically, the counts that we obtain as the CountMatrix, which usually serves the main object for data manipulations in single-cell studies, have their mostly identical analog in terms of multisets. Therefore, we can rewrite the CountMatrix as a set of different sequences of our interest with their multiplicity being represented as the $Count(x)$ function:

$$MmCs = \{s_1^{n_1}, s_2^{n_2}, \dots, s_m^{n_N} \mid n \coloneqq \{n\}_1^N = C_{Counts}(s_n)\},$$

Where MmCs – being the Multi-metrical Counts Set,
s – being the sample IDs,
n – being the number of same sequences in the set
And

$$\{MmCs, C_{Counts}(s_n)\} = [MmCs]^n$$

In which special functionals as multimetric and multitopology can be defined to allow special interclusteral data manipulations. Further studies on this theoretical approach are yet to be done, but so far it offers promising methods in terms of statistical and analytical purity.

Conclusions & Future Directions

- Sequencing depth greatly effects gene recovery for differential gene expression analysis. This was evident across certain twin pairs and batches.

In our future work, we plan to delve into the study of the prospects of both the theoretical secondary analysis presented above, and other methods of data normalization, refactoring, and restructuring of data that can potentially deepen our understanding of the processes under study. In addition, our studies, especially theoretical ones, can shed light on as yet unknown properties and features of single cell RNA-seq datasets.