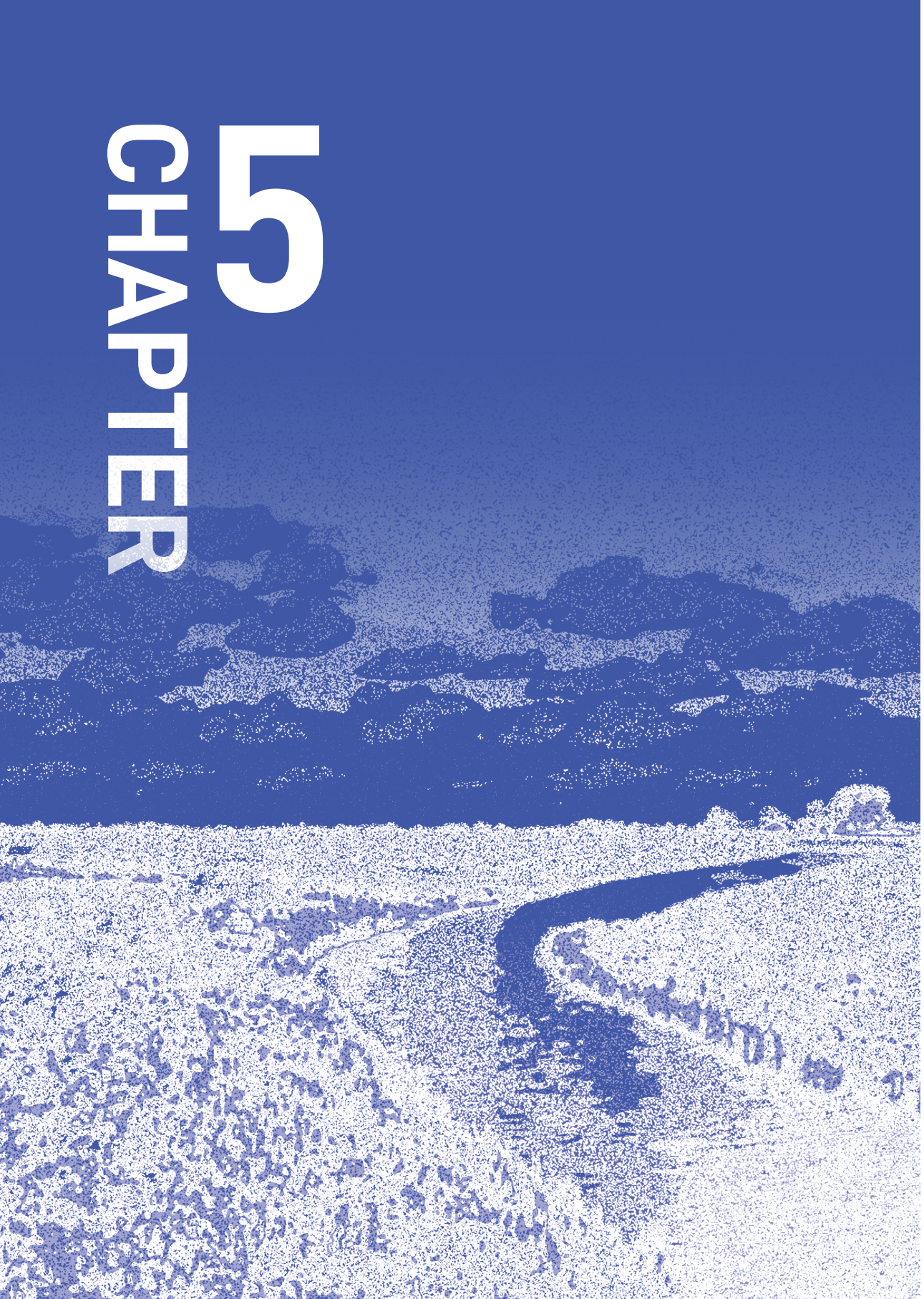
Systemic immune dynamics in cancer

Bakker, E.A.M.



Single-cell RNA-sequencing of whole blood of patients with metastatic triple negative breast cancer and healthy donors: a chapter of inconclusive data

Noor A.M. Bakker[1,2,5], Ewald van Dyk[1,2,3], Lodewyk F.A. Wessels[2,3], Marleen Kok[1,4] and Karin E. de Visser[1,2,5]

Unpublished work

[1] Division of Tumor Biology & Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[2] Oncode Institute, Utrecht, The Netherlands
[3] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[4] Department of Medical Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[5] Department of Immunology, Leiden University Medical Centre, Leiden, The Netherlands

Introduction

In science, publishing negative or inconclusive data is essential for advancing knowledge. These results, which may not support a hypothesis or expected outcome, provide valuable insights into what does not work, helping to prevent the unnecessary repetition of experiments. Additionally, they guide future research towards a more promising direction. When researchers share negative results, they contribute to a fuller understanding of a field, help refine theoretical frameworks, and promote transparency, which are all fundamental to scientific progress. However, sharing negative or inconclusive data is not common. Many researchers face significant barriers, including a preference within academic journals for positive, groundbreaking findings, which are seen as more publishable and impactful. Inconclusive data are often perceived as less exciting or premature and, therefore, less likely to gain visibility or career advancement. Additionally, researchers may be hesitant to publish negative results for fear of being perceived as unsuccessful or facing scrutiny from their peers. This reluctance creates a publication bias that skews the scientific literature, hindering cumulative knowledge and leading to inefficiencies in research. Encouraging the sharing of negative data would promote a more balanced and honest scientific discourse, ultimately fostering a more reliable and effective research ecosystem.

In this chapter, a research project centered on our scientific endeavors involving single-cell RNA-sequencing experiments is discussed. Due to the lack of discriminating results, which was primarily a result of insufficient statistical power due to low sample numbers and substantial inter-individual heterogeneity, we chose not to publish the “inconclusive data” in one of the few journals that accept such studies, such as the Journal of Articles in Support of the Null Hypothesis or the Journal of Negative Results in BioMedicine. Instead, I opted to include this work as a chapter of inconclusive data in my thesis. Including this work in my thesis not only documents and shares the findings from this project but also allows me to reflect on what I would have done differently in hindsight. I hope that this reflection will help guide the experimental design of scientists with similar goals, ensuring more efficient use of time and resources while increasing the likelihood of success. Additionally, by sharing this less successful aspect of my PhD journey, I aim to contribute to a more transparent and realistic understanding of the challenges a PhD can entail, which I hope will benefit future PhD students.

The aim of our study was to investigate transcriptional differences in the systemic immune landscape, encompassing both adaptive and myeloid immune cells, using single-cell RNA-sequencing of fresh whole blood samples from metastatic triple-negative breast cancer (mTNBC) patients and healthy donors (HDs). This approach allowed us to capture the full complexity of the immune system, enabling the analysis of not only cell-specific differences but also potential correlations and interactions among various immune cell types. Additionally, collecting whole blood samples without pre-processing or enriching for

certain cell types increased the likelihood of successfully capturing neutrophils, which are highly sensitive and prone to rapid cell death. Neutrophils, making up approximately 70% of circulating white blood cells, play a critical role in systemic inflammation and cancer[1-4]. Despite their abundance, distinct neutrophil subsets have yet to be clearly defined. By comparing the single-cell RNA profiles of neutrophils from mTNBC patients and HDs, we aimed to identify unique transcriptional states or neutrophil subsets that differ between the two groups. In addition to single-cell RNA-sequencing, we performed matched TCR- and BCR-sequencing to obtain insight into clonality and diversity of the circulating T cell and B cell repertoire. This comprehensive analysis was designed to offer deeper insights into the systemic immune dysregulation associated with mTNBC

Methods and Materials

Patients and Healthy Donors

Blood samples from patients with mTNBC were collected at baseline of the Triple B clinical trial[5], (NCT01898117). All patients with mTNBC were chemotherapy naïve for metastatic disease, and four out of five patients were chemotherapy-naïve for their primary tumor. The study protocol was conducted in accordance with the ICH Harmonised Tripartite Guideline for Good Clinical Practice and the principles of the Declaration of Helsinki. Baseline blood samples were used after approval by the institutional review board of the Netherlands Cancer Institute. Fresh blood samples from the healthy women (healthy donors, HD) were obtained after approval by the local medical ethical committee (NCT03819829). All patients and HDs provided written informed consent before enrolment. HDs were age matched to mTNBC patients. Blood samples were drawn in the morning and blood draw times were comparable for HDs and mTNBC patients.

Sample preparation

Peripheral blood was collected in EDTA vacutainers (BD) and processed immediately after blood draw. After erythrocyte lysis (lysis buffer: dH2O, NH4Cl, NaHCCO3, EDTA), cells were resuspended in Cell Staining Buffer (BioLegend). We included five patients with mTNBC and five HDs. For each run, blood cells from a patient with mTNBC were combined with blood cells from an age-, sex- and BMI-matched HD. For this purpose, cells were labeled for 30 min. at 4°C with a barcode using hashing antibodies against LNH-94 and 2M2 from BioLegend. We used TotalSeq-C0257 (394673) and TotalSeq-C0258 (394675) both in a 1:200 dilution. The barcoded single-cell suspensions of a patient and a HD were mixed in equal ratio, and further processed using 10X genomics Chromium Next GEM Single Cell 5’ Library and Gel Bead Kit v1.1 and Chromium Next GEM Chip G Single Cell Kit, following manufactures’ instructions (CG000208 Rev E, 10X Genomics). All libraries were quantified and normalized

based on library QC data generated on the Bioanalyzer system according to manufacturer's protocols (G2938-90321 and G2938-90024, Agilent Technologies). For each library type, based on the expected target cell counts, a balanced library pool of all samples was composed. Then all 4 library pools (Single Cell 5' Gene expression, TotalSeq-C Cell hashing and both types of V(D)J Enriched libraries) were quantified by qPCR, according to the KAPA Library Quantification Kit Illumina® Platforms protocol (KR0405, KAPA Biosystems). The Single Cell 5' Gene Expression and TotalSeq-C Cell hashing libraries were sequenced together on a NextSeq 550 Instrument (Illumina) using a NextSeq 500/550 High Output Kit v2.5 (cat. no. 20024906, Illumina). Paired end sequencing was performed using 28 cycles for Read 1, 8 cycles for Read i7 and 56 cycles for Read 2. For the Single Cell 5' Gene Expression sequencing this resulted in an average sequencing depth of 30.000 reads pairs/cell. The V(D)J Enriched libraries were sequenced together on a NextSeq 550 Instrument (Illumina) using a NextSeq 500/550 Mid Output Kit v2.5 (cat. no. 20024904, Illumina). Paired end sequencing was performed using 28 cycles for Read 1, 8 cycles for Read i7 and 130 cycles for Read 2.

Computational analysis single-cell RNA-sequencing data

Demultiplexing with cell hashing

Demultiplexing was performed using the HTODemux function provided by Seurat with default parameters[6,7]. Cells were split into four categories: one for each sample, a negative category for cells with insufficient hash and a doublet category where appreciable hashtags were observed for both samples.

Quality control

Cells with negative or doublet labels from hash demultiplexing were removed from the analysis. Using hashtag doublets as a positive control, we were unable to see a correlation between read counts and doublet status. Furthermore, computational methods such as Scrublet[8] were unable to confidently predict doublets. As a consequence, we were unable to remove doublets sharing the same hashtags.

We removed cells with less than 10 total read counts or a mitochondrial percentage above 20%. The low read count threshold was used to retain cell types such as neutrophils that naturally contain very few read counts. Cells with low read counts were filtered out indirectly after applying a posterior threshold of 80% using a multinomial cell type classifier (see "Automatic cell type classification").

Automatic cell type classification

We performed automatic cell type identification based on the blood derived cell type categories in the CIBERSORT LM22 dataset[9] using a multinomial model similar to that

proposed in[10]. We extended the CIBERSORT signature gene list with platelet specific genes PF4, ITGA2B, F13A1 and NCOA4 since they are frequently expressed in platelets and not in other cell types as specified in PanglaoDB. In total, we used a set S containing 548 signature genes to differentiate between cell types. For each cell, we compute the likelihood of observed read counts for each cell type: $P(x_b, \bar{x}|c) \propto p_b^{x_b} \prod_{i \in S} p_i^{x_i}$ where c represents the cell type under consideration, $\bar{x}$ represents the vector of read counts of genes in the signature list S with x_i representing the read counts of the i-th gene, x_b represents a single background read count (i.e. total number of reads not from genes in the signature list) and the parameters p (and p_b) that represent the expected proportion of reads in signature genes (and background). These model parameters are estimated from bulk sequencing datasets provided by CIBERSORT and the Human Primary Cell Atlas[11]. Posterior probabilities were derived from the likelihoods using Bayes' theorem. Average proportions of different cell types across our flow cytometry dataset were used as priors except for neutrophils and platelets, where the priors were both set to 30%. Neutrophil frequencies are typically high (>50%) in human blood, but due to their low read counts, many do not survive our original quality control, which justifies a lower prior.

Batch alignment

Batches were aligned using the fastMNN algorithm[12]. After alignment we modeled the remaining batch effect and sample type (mTNBC vs HD) simultaneously using the generalized linear modeling framework in MiloR (see differential abundance analysis section). Since all batches were processed in the same core facility, fastMNN resulted in superior performance compared to the diagonalized CCA method provided by Seurat.

Count normalization and dimensionality reduction

For the analysis that included all cell types, we retained the top 2000 highest varying genes based on the vst method in Seurat. Read counts were normalized by per-cell library size and log (+1) normalized. The top 50 principal components were extracted for further downstream analysis. For visualization in two dimensions, we performed UMAP manifold learning[13]. For the differential abundance analysis, cell types were separated based on the multinomial classifier described earlier. For each cell type, the top 2000 highest varying genes were recomputed as before, except for the neutrophils where we used the top 500 varying genes due to low overall read counts. For each cell type we extracted the top 10 principal components, except for the T cells. T cells exhibit higher heterogeneity than other cell types and we extracted the top 20 principal components.

Clustering

Unsupervised clustering was performed using a smart local moving algorithm based on

shared nearest neighbor (SNN) networks[14]. Optimization of parameters was collectively controlled with a resolution parameter in the Seurat package.

Differential Gene Expression Analysis

Analyzing single-cell RNA-sequencing data using mixed-effects models can be computationally prohibitive due to the large number of observations. To simplify the process, we aggregated unique molecular identifier (UMI) counts for each immune cell population within each sample, generating pseudobulk data. This approach allows for differential expression analysis using well-established methods originally developed for bulk RNA-sequencing. Prior to this conversion, we created subsets of the data corresponding to distinct immune cell populations, ensuring that DE analysis could be performed for each population separately. For immune cell populations that exhibit significant differences in gene expression between patients with mTNBC and HDs, we could further conduct gene set enrichment analysis and other pathway analyses to interpret the biological significance of the differentially expressed genes.

Differential abundance analysis

We detected groups of similar cell states and tested for differentially abundance when comparing triple-negative breast cancer samples and healthy donors using MiloR[15]. Traditionally, clusters of cells sharing similar states are identified prior to performing a differential abundance analysis. However, MiloR is independent of clustering and can potentially pick up differential regions in transition phases. Batch effects were modeled simultaneously as a nuisance parameter (see “Batch alignment”). The TCR and BCR diversity was calculated using Pielou’s evenness procedure.

Results

No differential gene expression and cell state abundances between mTNBC and HDs

To gain deeper insights into the cellular states of circulating cells in TNBC patients, we performed single-cell RNA-sequencing along with matched single-cell BCR- and single-cell TCR-sequencing on fresh blood samples from five mTNBC patients and five age-matched HDs (Figure 1a). The samples were processed in five batches, with each batch comprising white blood cells from a fresh blood sample of both a mTNBC patient and a HD. After hash tagging, the two samples in each batch were pooled, and all batches were sequenced in a single run.

We observed a large variation in RNA-content between different immune cell types; both in the number of genes detected in each cell (nFeature_RNA) as well as in the total number of RNA molecules detected within a cell (nCount_RNA) (Figure 1b). Cells in clusters

0, 1, 4 and 7 have low RNA content and are identified as neutrophil clusters, as anticipated (Figure 1b). Cluster 10 is also comprised of cells with a low RNA content and was identified as a platelet cluster.

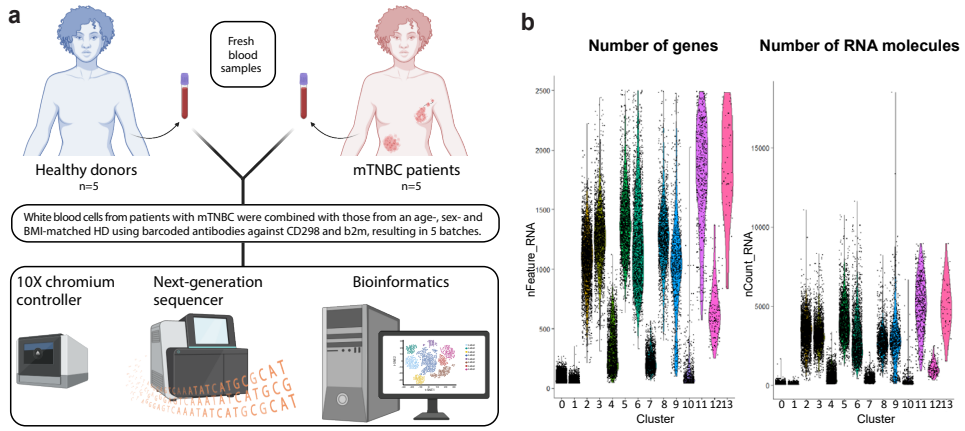


Figure 1: *Experimental setup* ***(a)*** *and quality control* ***(b)*** *of single-cell RNA-sequencing experiments. nFeature_RNA is the number of genes detected in each cell. nCount_RNA is the total number of RNA molecules detected within a cell. Neutrophils are divided over clusters 0, 1, 4 and 7.*

In all samples, we captured a great diversity of immune cell types, including the technically challenging neutrophil population (Figure 2a). For most immune cell populations, we found multiple subpopulations. Intriguingly, eight continuous neutrophil states were identified in our dataset, suggestive of neutrophil subset diversity (Figure 2a). This is in agreement with what has previously been described[10,16,17]. However, none of these states were unique to either HDs or mTNBC patients and no statistically significant enrichments or depletions in those cell states were found (Figure 2b).

After subsetting the major immune cell populations (Figure 2a) and generating pseudobulk data for each cell type, we conducted differential gene expression analysis comparing mTNBC patients and HDs. However, this analysis did not reveal a list of significantly differentially expressed genes between the two groups. As a result, subsequent analyses such as pathway enrichment and gene set enrichment analysis on differentially expressed genes could not be performed. We found that the variance between the two groups in gene expression at the level of the selected immune cell populations was similar to the variance within each group. This suggests that gene expression differences between mTNBC patients and HDs potentially requires an alternative analytical approach.

Such an alternative analytical approach is the application of the MiloR algorithm, which achieves greater power by reducing sparsity and noise, improving the detection of subtle abundance changes, and increasing sensitivity to shifts in cell population composition across groups or conditions. After extensive analysis of this single-cell transcriptomic dataset using the MiloR package, we did not find specific cell states or subsets that were enriched or depleted in mTNBC when compared to HDs (Figure 2b). Moreover, analyzing TCR clonality and clone sizes did not reveal statistically significant differences between patients with mTNBC and HDs (Figure 2c). Calculating TCR diversity measured by Pielou's evenness, revealed equal TCR diversity in HDs and mTNBC patients (Figure 2d). BCR clonality and clone sizes did not differ in a statistically significant manner between patients with mTNBC and HDs (Figure 2e). Calculating BCR diversity measured by Pielou's evenness, showed comparable BCR diversity in HDs and mTNBC patients (Figure 2f). Important to mention is the substantial degree of inter-individual and inter-batch heterogeneity that was observed (Figures 2c, e), making it increasingly challenging to identify breast cancer driven differences between the two groups. Combining transcriptional profiling with single-cell BCR or single-cell TCR sequencing did not yield any discriminatory information between HDs and mTNBC patients.

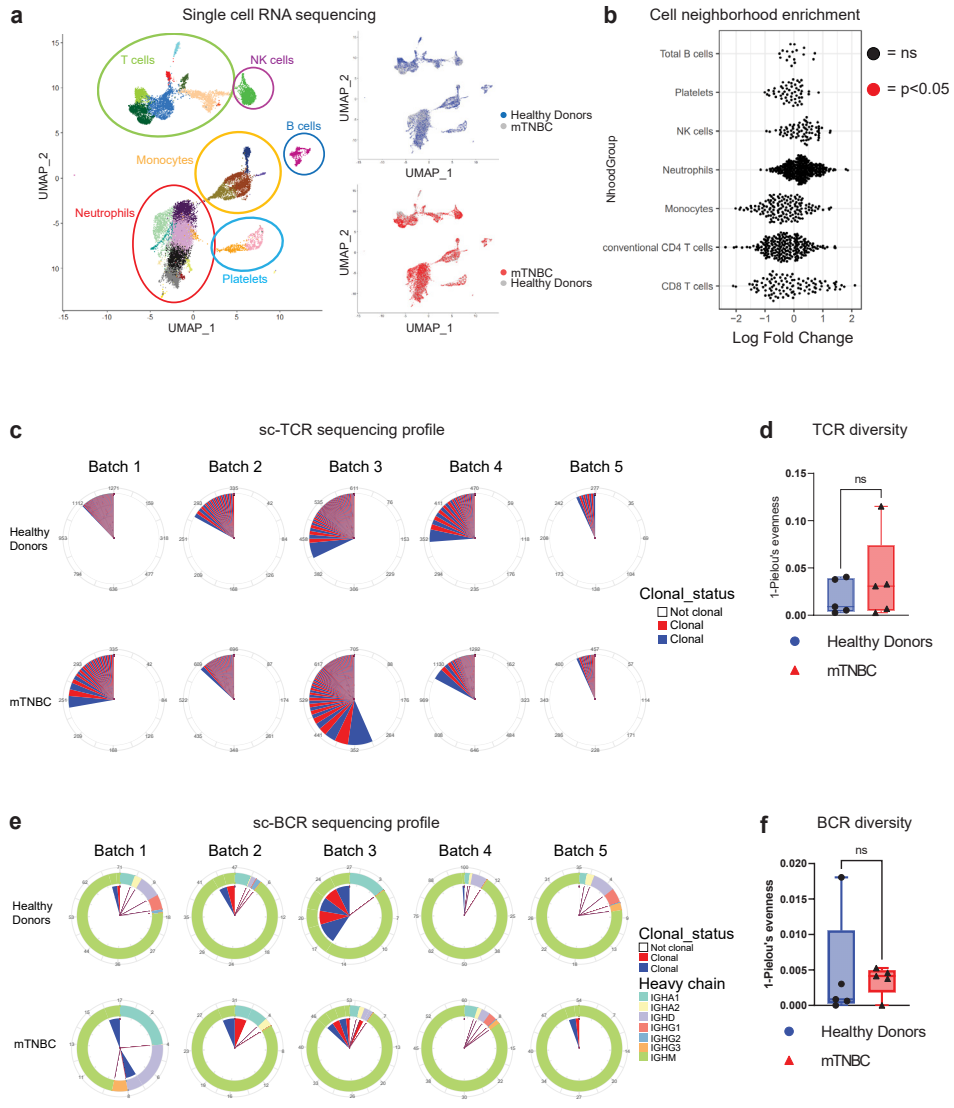


Figure 2. Single-cell RNA-sequencing and matched TCR- and BCR-sequencing on fresh whole blood samples. (a) *UMAP showing aggregated cells from HDs (n=5) and patients with mTNBC (n=5). On the left, unsupervised clustering was performed using shared nearest neighbor networks, resulting in multiple cellular states. On the right, UMAPs are colored according to sample type; blue are cells from HDs and red are cells from patients with mTNBC.* ***(b)*** *Differentially abundant cell counts grouped into Milo neighborhoods after correcting for batch effect. No significant neighborhoods were found.* ***(c)*** *TCR clonality derived from single-cell TCR sequencing for five HDs (top row) and five mTNBC patients (bottom row). Each vertical pair of pie charts corresponds to a batch of two samples that were processed together. Alternating red and blue pieces of the pie chart represent T cells that are clonal (>1 cell with identical TCR sequences).* ***(d)*** *T cell receptor repertoire diversity measure for HDs and patients with mTNBC. Diversity is defined as 1 – Pielou's evenness (i.e. 1 – normalized entropy). High values represent less diversity and therefore more clonality.* ***(e)*** *BCR clonality derived from single-cell BCR sequencing for five HDs (top row) and five mTNBC patients (bottom row). Each vertical pair of pie charts corresponds to a batch of two samples that were processed together. Alternating red and blue pieces of the pie chart represent B cells that are clonal (>1 cell with identical BCR sequences). The outer ring of the pie chart is colored according to the heavy chain.* ***(f)*** *B cell receptor repertoire diversity measure for HDs and patients with mTNBC. Diversity is defined as 1 – Pielou's evenness (i.e. 1 – normalized entropy). High values represent less diversity and therefore more clonality.* ▶▶▶

Discussion

To gain more insight in potential differences in the cellular states from various circulating immune cell populations between mTNBC patients and HDs, single-cell RNA-sequencing and matched single-cell TCR and single-cell BCR sequencing was performed on fresh leukocytes of five mTNBC patients and five HDs. Extensive analysis of this dataset did not yield any substantial differences between mTNBC and HD transcriptomes or the identification of differentially abundant cell states between the two groups. We believe the lack of discriminating results is mostly an issue of limited group sizes. Given the high degree of heterogeneity within the five HDs and five patients with mTNBC, identifying differences between the two groups became near impossible.

Interestingly, eight distinct neutrophil states were identified in our dataset, suggesting possible subset diversity; however, none of these states were exclusive to either healthy donors (HDs) or metastatic triple-negative breast cancer (mTNBC) patients, and no differences in their abundances were detected. We speculate that the lack of discriminating features of the neutrophils is – in addition to the low n-number of individuals – a result of the low read-depth within this cell type. Because we chose not to enrich for any specific cell type and neutrophils have very low RNA content, most reads in the dataset came from other immune cells. Since the number of genes coming from e.g. lymphocytes and monocytes (Figure 1b clusters 2, 3, 5, 6, 8, 9, 11-13) is so much larger compared to those coming from neutrophils (Figure 1b clusters 0, 1, 4 and 7), increasing the number of sequencing runs would mainly lead to more reads of genes that are already covered. It is expected that with increased read depth in the neutrophils and an increased sample size, there are transcriptional differences in neutrophils from HDs and mTNBC patients. Therefore, we hypothesize that subjecting purified neutrophils from well defined, untreated mTNBC patients to single-cell RNA-sequencing in comparison to purified neutrophils from HDs, may provide valuable information about which neutrophil states are associated with mTNBC. By combining this with barcoded antibodies to identify surface markers associated with a particular cell state, followed by live cell sorting, one could possibly even link certain over- or underrepresented neutrophil cell states to their functional properties. This would further advance our understanding about the role of neutrophils in cancer, and nearer the step towards modulation of myeloid cells in cancer patients.

Lastly, I would like to point out an unexpected advantage of the approach we took. To minimalize batch effects and reduce costs, we combined a patients sample with an age- and sex matched HD sample for each run by making use of barcoded antibodies against two antigens that are present on virtually all cells. During quality control, cells with low RNA yield are typically filtered out to exclude apoptotic or dying cells, which often show poor RNA quality. However, this process unintentionally removes neutrophils, as they naturally have low RNA content despite being viable. To avoid losing neutrophils during this filtering step,

we leveraged the use of cell barcode handles to identify intact cells. Within this pool of cells that were initially filtered out, neutrophils can be accurately identified and retained, preserving them in the analysis despite their low RNA levels. Additionally, the barcodes provided a straightforward handle to filter out doublets. Most published doublet-removal algorithms are based of RNA content, but when doublets are formed with low RNA content cells (like neutrophils or eosinophils), this will not be picked up. However, when a doublet is formed of cells coming from two donors, they can be filtered out because they carry two different barcodes. This still does not solve the doublet issue for situations in which a doublet is formed involving low RNA content cells from the same donor, but it certainly cleans up the dataset. To take this idea a step further, one could even consider dividing each sample over multiple wells, staining each well with a different barcode, increasing the effectiveness of this doublet removal approach even further, and potentially even allow for increasing the number of cells that can be loaded per experiment.

Overall, we conclude that in this limited cohort of five mTNBC patients and five HDs, no statistically significant discriminatory results were observed within the single-cell RNA-sequencing dataset. However, we acknowledge the potential for significant differences to emerge with an expanded sample size. By increasing the number of individuals and refining our pre-processing methods to separate immune cells based on RNA content—specifically targeting high- and low-RNA populations to ensure consistent read-depth across all immune cell types—we may uncover statistically significant differences in the single-cell RNA-sequencing profiles between mTNBC patients and HDs.

References

1 Carnevale, S. *et al.* Neutrophil diversity in inflammation and cancer. *Front Immunol* **14**, 1180810, doi:10.3389/fimmu.2023.1180810 (2023).

2 Coffelt, S. B., Wellenstein, M. D. & de Visser, K. E. Neutrophils in cancer: neutral no more. *Nat Rev Cancer* **16**, 431-446, doi:10.1038/nrc.2016.52 (2016).

3 Siwicki, M. & Pittet, M. J. Versatile neutrophil functions in cancer. *Semin Immunol* **57**, 101538, doi:10.1016/j.smim.2021.101538 (2021).

4 Wellenstein, M. D. *et al.* Loss of p53 triggers WNT-dependent systemic inflammation to drive breast cancer metastasis. *Nature* **572**, 538-542, doi:10.1038/s41586-019-1450-6 (2019).

5 van Rossum, A. G. J. *et al.* Carboplatin-Cyclophosphamide or Paclitaxel without or with Bevacizumab as First-Line Treatment for Metastatic Triple-Negative Breast Cancer (BOOG 2013-01). *Breast Care (Basel)* **16**, 598-606, doi:10.1159/000512200 (2021).

6 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).

7 Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol* **19**, 224, doi:10.1186/s13059-018-1603-1 (2018).

8 Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291 e289, doi:10.1016/j.cels.2018.11.005 (2019).

9 Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol Biol* **1711**, 243-259, doi:10.1007/978-1-4939-7493-1_12 (2018).

10 Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317-1334 e1310, doi:10.1016/j.immuni.2019.03.009 (2019).

11 Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C. & Hume, D. A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics* **14**, 632, doi:10.1186/1471-2164-14-632 (2013).

12 Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-427, doi:10.1038/nbt.4091 (2018).

13 Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*, doi:10.1038/nbt.4314 (2018).

14 Waltman, L. a. v. E., N. J. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal* **471**, doi:doi:10.1140/epjb/e2013-40829-0 (2013).

15 Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat Biotechnol* **40**, 245-253, doi:10.1038/s41587-021-01033-z (2022).

16 Xie, X. *et al.* Single-cell transcriptome profiling reveals neutrophil heterogeneity in homeostasis and infection. *Nat Immunol* **21**, 1119-1133, doi:10.1038/s41590-020-0736-z (2020).

17 Grieshaber-Bouyer, R. *et al.* The neutrotime transcriptional signature defines a single continuum of neutrophils across biological compartments. *Nat Commun* **12**, 2856, doi:10.1038/s41467-021-22973-9 (2021).