Doublet Detection in Single-Cell RNA Sequencing Data and Dexamethasone Induced Bone Toxicity

by

Tristan Vermaat

January 2024
Prinses Maxima Center for Pediatric Oncology
Supervisor: Philip Lijnzaad
Examiner: Claudia Janda

Table of Contents

Abstract

Dexamethasone is an anti-inflammatory drug commonly used during treatment of paediatric cancers. However, prolonged exposure to dexamethasone can lead to impaired bone development. To study the detrimental effect of dexamethasone on the developing bone single-cell RNA sequencing datasets were created. The libraries used to create these datasets contained too many cells, which lead to the formation of technical artifacts called “doublets”. Doublets cannot be recognized easily and compromise the data analysis. This study aimed to evaluate and improve upon existing doublet detection methods. So that the single-cell RNA sequencing datasets could be cleaned and the effect of dexamethasone on the developing bone could be investigated. Existing doublet detection methods such as *DoubletFinder* and *scDblFinder* create artificial doublets and look at the distances to real cells in principle component space to identify real doublets. The performance of these methods was tested on datasets for which a ground truth reference is known. The single-cell RNA sequencing datasets were then cleaned of doublets and projected onto a bone reference dataset to refine cell typing. Hereafter, differential gene expression and overrepresentation analysis were applied to identify enriched processes that might be related to dexamethasone induced bone toxicity. This study reports that *scDblFinder* can be used to predict doublets with great speed and higher accuracy than *DoubletFinder*. Furthermore, *scDblFinder* annotations and subsequent removal successfully cleaned the single-cell RNA sequencing datasets and allowed for improvements to be made in cell typing. Finally, dexamethasone negatively affected the chondrocyte, osteoblast, lymphatic endothelial and B-cell populations and caused the pre-osteoblast population to have decreased differentiation and bone development. In conclusion, this study suggests that the effect of dexamethasone on the developing bone is primarily a decrease in bone formation.

Introduction

Bone Toxicity

Dexamethasone is a glucocorticoid used as anti-inflammatory drug during treatment of childhood cancers (Ferrara *et al*. 2019). However, this treatment has serious side-effects, including impaired bone development (Oray *et al.* 2016). Existing studies on dexamethasone treatment have used cell lines or skeletally mature mice. However, the effect of dexamethasone on the developing bone might be substantially different from mature mice (Ward, 2020). To study the molecular mechanisms of dexamethasone induced bone toxicity the Janda group (Warmink *et al*.) employed single-cell RNA sequencing (single-cell RNA-seq) on tissue samples of skeletally immature mice. These mice were treated with high dosages of dexamethasone. The single-cell RNA-seq datasets from this study were then used to investigate the detrimental effect of dexamethasone on a single-cell level, since the effects of dexamethasone treatment might be cell type specific.

Single-cell RNA-sequencing

Single-cell RNA-seq is a technique used to study the expression profiles of single cells. single-cell RNA-seq is able to achieve a high-resolution view of cell-to-cell variation. This level of detail cannot be obtained by bulk RNA sequencing as this method can only look at the expression profile of an entire sample. Consequently, the expression profiles of rare cell populations will be lost in the data. Whereas single-cell RNA-seq can discover these cellular differences. 10x genomics is a major provider of single-cell RNA-seq tools, which will be used throughout this study (10xgenomics.com). One crucial step in single-cell RNA-seq is the isolation of single cells. Capture techniques make use aqueous droplets that form on contact with an oily suspension to capture single cells together with barcoded gel beads (Figure 1). These barcoded gel beads are unique and used to identify the cells. However, sometimes not one but two cells are captured, forming a doublet (Germain *et al*. 2022). The doublet formation process depends on the number of cells loaded. To reduce doublet formation, cells are deliberately underloaded compared to barcoded gel beads so that on occasion a cell is captured but most droplets will be empty, effectively reducing the doublet formation rate (Bloom, 2018). However, the bone toxicity libraries used in this study were overloaded and contained a relatively large number of doublets. Re-sequencing is expensive, wasteful and not always possible due to the limited availability of biological material. Hence, there is a need for doublet removal in these bone toxicity datasets.

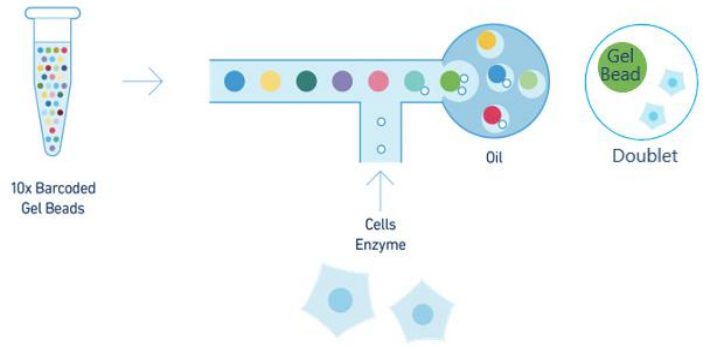

Figure 1: Single-cell RNA-seq capture technique. *10x Genomics approach to capture single cells in droplets containing barcoded gel beads. Doublets are formed when not one, but two cells, are captured in a single droplet (Adapted from 10x Genomics).*

Doublets

When two cells are captured together, they are sequenced together as well. The resulting doublet consists of a combined expression profile from both cells. Depending on the cell types of the captured cells a doublet can be called either homotypic or heterotypic. Homotypic doublets form when two

cells of the same cell type mix. Homotypic doublets generally do not hamper data analysis, since both expression profiles are near identical. However, this characteristic makes it difficult to detect homotypic doublets based on solely transcriptomics. Heterotypic doublets form when two different cell types mix. The expression profiles of the captured cells are likely to be different and the resulting doublet will have a mixture of both. Heterotypic doublets reduce the differences between clusters, which compromises the biological analysis. Furthermore, they can easily be mistaken for intermediate populations or transitory states (Germain *et al*. 2022; Amezquita *et al*. 2022). Thus, it is crucial to remove doublets to prevent them from affecting the biological analysis.

Doublets contain on average more mRNA transcripts than single cells, hence it might seem reasonable to set an upper threshold based on the number of mRNA transcripts (Stoeckius *et al*. 2018). And while quantitative measure of gene expression is possible with unique molecular identifiers (UMI) which are uniquely associated with individual mRNA transcripts (Islam *et al*. 2014). This approach is insufficient, since there is still a large overlap between doublets and single cells, referred to as singlets, which is also observed in the intestinal organoid data that will be used throughout this study (Figure 2). Furthermore, this approach does not account for technical variability in capture efficiency nor biological variability between cell types and individual cells (Kang *et al*. 2018). Hence, there is a need for proper doublet detection and subsequent removal.

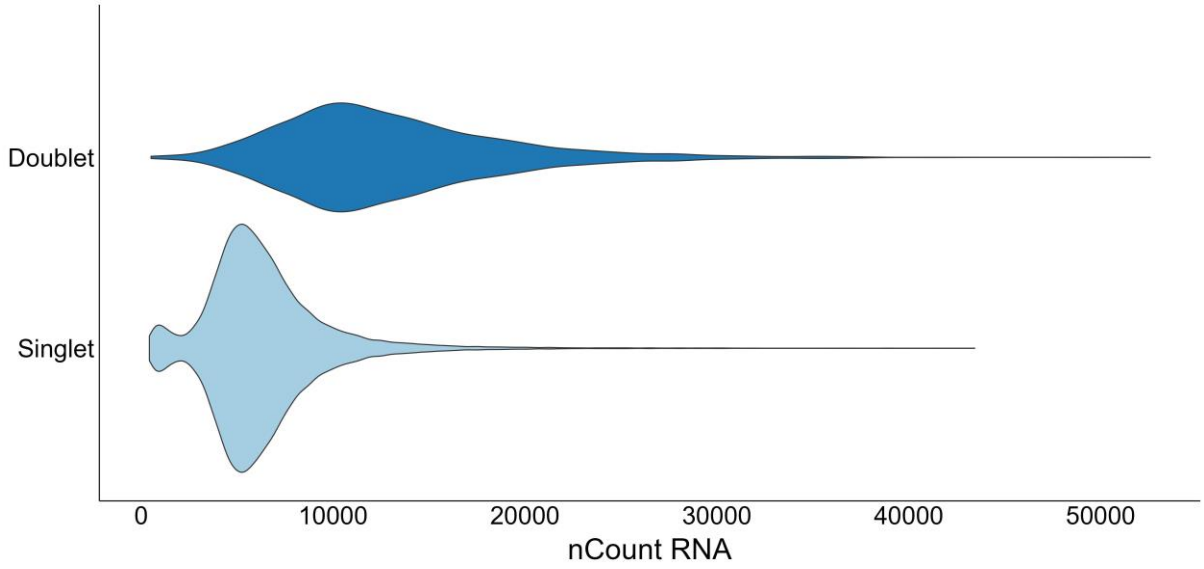

Figure 2: Transcript count distribution across doublets and singlets in intestinal organoid data. *Doublets contain on average a higher transcript count, however there is a large overlap in the distribution, which confirms that a doublet predictor based on solely the transcript count is insufficient.*

Doublet detection methods

First of all, an attempt was made to apply simple logistic regression as a doublet detection approach. Simple logistic regression is a machine learning approach in which a logistic function is fitted to the data with the purpose of predicting a dichotomous variable. However, this approach proved inferior to existing doublet detection methods, such as *DoubletFinder* (McGinnis et al. 2019) and *scDblFinder* (Germain et al. 2022). At the time of writing these methods were found to be most promising. Furthermore, the *scDblFinder* paper has performed their own validation and comparison study. This study found that the *DoubletFinder* method was performing better than other doublet detection methods. However, they also report that to date no method was found to systematically outperform the others, which was their reasoning for creating *scDblFinder* (Germain et al. 2022). And they report that *scDblFinder* performs better than other methods, including *DoubletFinder*, on most datasets.

Hence, *DoubletFinder* and *scDblFinder* were identified as potential doublet removal techniques for the bone toxicity data.

These approaches make use of artificial doublets generated by mixing real cells together (Figure 3, 4). The logic behind this approach is that doublets are a combination of two real cells and the resulting doublet gene expression is likely similar to the gene expression of an artificial cell consisting of two similar real cells mixed together. Hence, these artificial doublets will likely cluster together with real doublets, which can be used to identify real doublets. This is done by constructing a k Nearest Neighbour network, which effectively looks at these distances in principle component space. *scDblFinder* then uses the ratio of doublets in a cells neighbourhood and gathers additional statistics to use in gradient-boosted trees to train a classifier (Figure 4) (Germain et al. 2022). Whereas *DoubletFinder* trains a classifier on the proportion of artificial nearest neighbours directly (McGinnis et al. 2019).

DoubletFinder and s*cDblFinder* can be run on solely gene expression data and do not require external information nor experimental techniques. Hence, *DoubletFinder* and *scDblFinder* can be used to predict doublets in our datasets. However, these methods first need to be evaluated on their doublet detection capabilities.

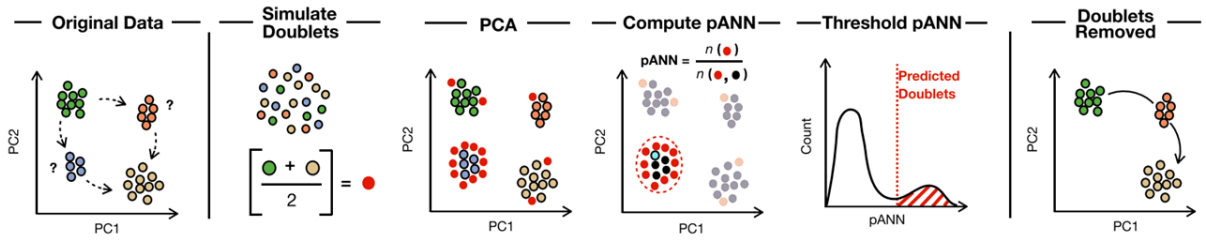

Figure 3: DoubletFinder workflow. *Original data is used to simulate artificial doublets, which are incorporated in the existing data and processed together. Consequently, a cells' neighbourhood is defined and the proportion of artificial nearest neighbours (pANN) is computed. Cells with high pANN values are likely doublets and consequently removed (blue cells). The doublets in this dataset represented an artefactual intermediate cell state (McGinnis et al. 2019).*

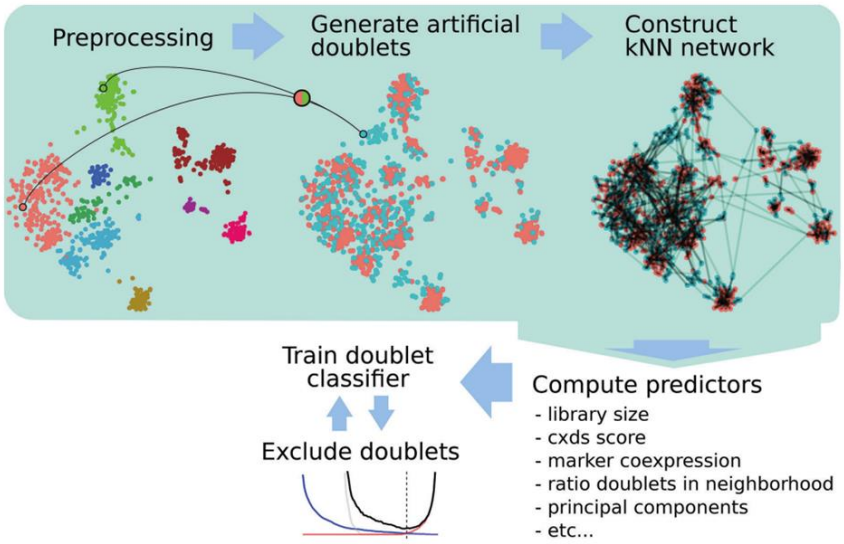

Figure 4: scDblFinder workflow*. scDblFinder randomly selects cells and adds them together to generate artificial doublets. Thereafter a k nearest neighbours (kNN) is constructed. scDblFinder then gathers statistics at various neighbourhood sizes to build predictors to use in gradient-boosted trees to train a classifier than can be used to annotate doublets. (Germain et al. 2022)*

Ground truth

To evaluate the performance of *DoubletFinder* and *scDblFinder* a ground truth is required. However, this is not available for the bone toxicity datasets. Hence, both methods will first be tested and compared to one another on datasets where a ground truth is available. Cell multiplexing techniques such as 10x CellPlex (*Cell Ranger*) and genotypically multiplex data (Heaton et al. 2020) can be used for this purpose. CellPlex is a technique included in *Cell Ranger,* which is a set of analysis pipelines from 10x Genomics. CellPlex works with cell multiplexing oligonucleotides (CMO) attached to lipids (Figure 5). These lipids embed into the membrane and tag the samples with different CMOs. The samples are then sequenced together. During data analysis these samples can be deconvolved based on their CMOs. However, if a cell contains CMOs from two different samples it is annotated as doublet. Alternatively, genotypically multiplex data is created by combining two different patient samples. These patient samples can be distinguished using *Souporcell* (Heaton et al. 2020). *Souporcell* remaps the reads of the dataset and performs its own single nucleotide polymorphism (SNP) calling. *Souporcell* then identifies patient specific SNPs and uses this to cluster cells by genotypes. However, if a cell contains patient specific SNPs from both patients, it is likely a doublet. Therefore, 10x CellPlex and genotypically multiplexed datasets can be used to assess the performance of *DoubletFinder* and *scDblFinder*. However, these methods do have some limitations, they are not able to predict doublets within samples of the same CMO or conversely between cells from the same patient. Furthermore, it is unsure how well these methods predict doublets.

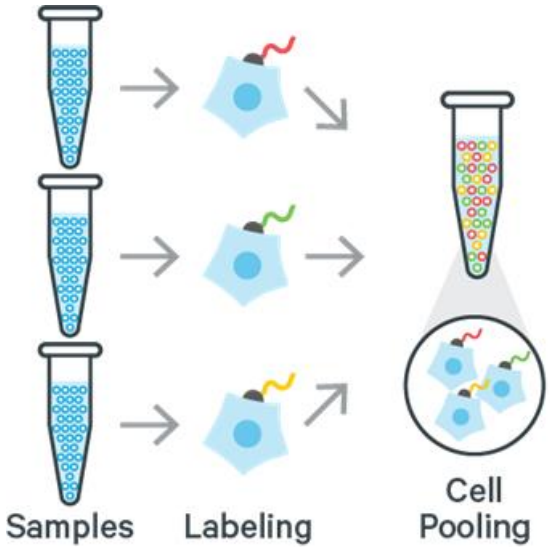

Figure 5: Ground truth labeling with cell multiplexing oligonucleotides. *Cells from different samples are labelled with specific oligonucleotides before pooling and sequencing. This allows for deconvolution of sample origin during data analysis (Adapted from 10x Genomics).*

Bone toxicity analysis

Once the optimal doublet detection approach is identified, doublets in the bone toxicity datasets can be removed and the data analysed. However, to analyse these datasets, they first need to be annotated. Preliminary annotations were available, but after doublet removal cell typing will have to be confirmed. Besides, preliminary annotations performed by Warmink *et al*. were obtained using singleR (Aran *et al.* 2019), which is an automatic reference-based annotation tool for single-cell RNA-seq data. This approach is limited in its bone marrow references. Hence, cell projection (Stuart *et al*. 2019) with a bone marrow reference dataset will be applied to the bone toxicity data to refine cell typing. Projection works by first integrating the reference and query dataset into a shared subspace defined by a shared correlation structure across the datasets (Figure 6). Within this shared subspace

cell pairwise correspondences can be identified between single cells across datasets, these cell pairwise correspondences are referred to as anchors. Using these anchors, the labels of a reference dataset can be transferred onto the query dataset. After cell typing is refined, the effects of dexamethasone will be investigated on a population level. Furthermore, differential gene expression will be applied on the bone toxicity datasets to perform overrepresentation analysis with the purpose of identifying enriched processes that might be related to dexamethasone induced bone toxicity.

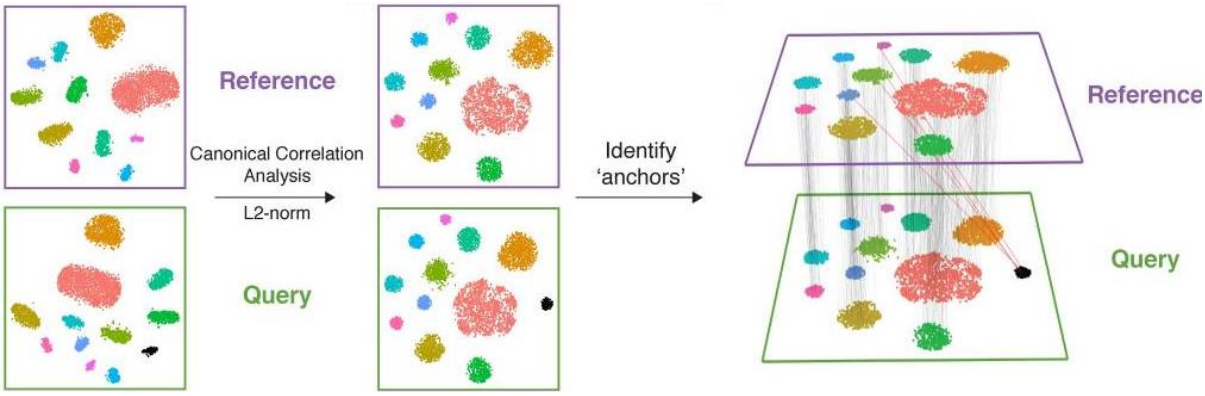

Figure 6: Projection overview. *Dimensionality of the reference and query datasets is reduced. Hereafter, correlation analysis is performed and the correlation vectors are normalized. This creates a subspace defined by shared correlation structure across the datasets. Within this space cell pairwise correspondences can be identified between single cells. (Stuart et al. 2019).*

Aim

The aim of this study is to evaluate and improve upon existing methods to identify doublets in single-cell RNA-seq data. This approach will then be applied to remove doublets from the bone toxicity datasets. Subsequently, single-cell analysis will be performed on the bone toxicity datasets with the aim of revealing underlying molecular mechanisms of the detrimental effect of dexamethasone on bone development in skeletally immature mice.

Results

scDblFinder outperforms DoubletFinder on multiplexed intestinal organoid data

First *scDblFinder* and *DoubletFinder* were evaluated to determine the best doublet removal approach. To explore the performance of both methods, an existing dataset of intestinal organoids generated with the 10x CellPlex technique was used (*Cell Ranger*). This dataset functioned as ground truth to calculate the sensitivity and specificity of these methods. The dataset consisted of 6 pooled intestinal organoid samples each tagged with a different oligonucleotide. The ground truth reference consisted of 3,331 doublets in 21,347 annotated cells (Figure 7A). This dataset was selected because it contained a high number of cells and consequently a large number of doublets.

Both methods were first optimized prior to performance comparison, which will be explained in the Methods section. After filtering the intestinal organoid dataset consisted of 21,924 cells, 577 more than in the ground truth reference. Cells not present in the ground truth annotations had a transcript count below 2,500 mRNA transcripts. These cells were included during doublet annotation by *scDblFinder* and *DoubletFinder* but were excluded during performance evaluation, since ground truth annotations were unavailable. Furthermore, *scDblFinder* was run multiple times in order to create a consensus annotation with cells called at least half of the time. This was done to create a more consistent annotation, since there was a 10% variation in doublet calls between independent runs, which will be explained in the next chapter.

Hereafter, each prediction method was compared to the ground truth annotations and the Jaccard index calculated. The Jaccard index measures the similarity between two different sets by dividing the size of the intersection of both sets by that of their union. This works best when both sets are of equal size. *DoubletFinder* identified 2,663 doublets correctly and achieved a Jaccard index of 0.586 (Figure 7B,F). Whereas *scDblFinder* identified 2,787 doublets correctly and achieved a Jaccard index of 0.631 (Figure 7C,F). Furthermore, 2,597 doublets were annotated correctly by both prediction methods. However, a set of 772 cells were incorrectly annotated by both *scDblFinder* and *DoubletFinder*. These cells are likely doublets tagged with the same oligonucleotide and hence not annotated by the ground truth. All in all, *scDblFinder* achieved a higher overlap than *DoubletFinder*, annotating more doublets correctly and fewer singlets incorrectly, based on the ground truth.

Furthermore, the models' performance was evaluated using the Receiver Operating Characteristic (ROC) and the area under the curve (AUC), which is frequently used as a quality measure for classifiers. Comparing the ROCs of *scDblFinder* and *DoubletFinder* it becomes evident that both models achieve a high AUC (Figure 7D). However, the ratio between singlets and doublets is imbalanced. Therefore, a Precision-recall curve (PRC) is likely more informative, since it takes class imbalance into account (Saito and Rehmsmeier, 2015). *scDblFinder* consistently outperformed *DoubletFinder* on both the ROC and PRC curves (Figure 7E). Moreover, *scDblFinder* achieved much greater speed than *DoubletFinder,* requiring only 3 minutes per run as opposed to 10. Furthermore, *scDblFinder* was much easier to use as it did not need dataset specific parameter optimisation whereas *DoubletFinder* did, this took 45 minutes. All together, this made *scDblFinder* the ideal candidate for doublet removal.

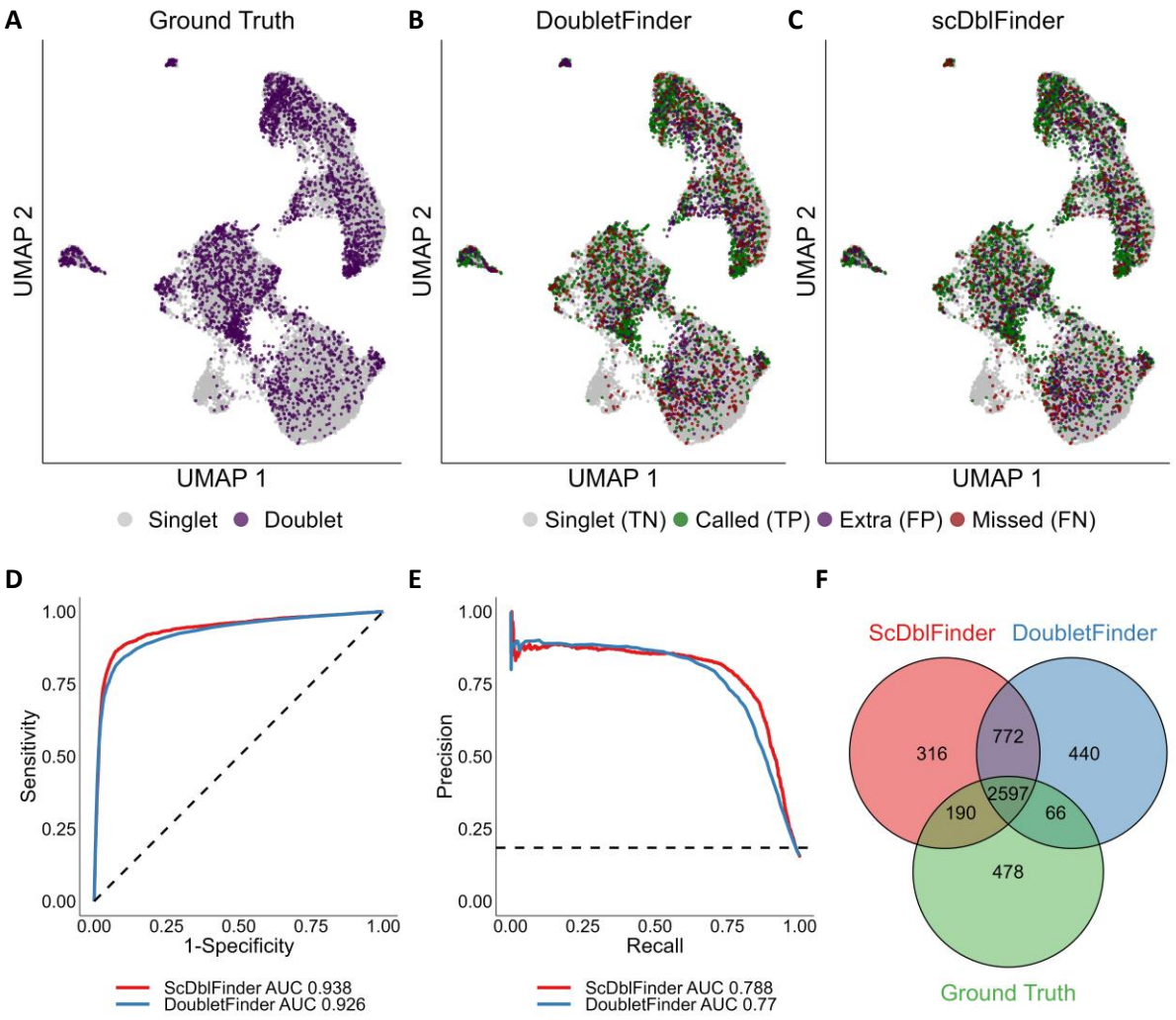

Figure 7: Overview of doublet predictions in intestinal organoid data. (A-C) *UMAP projection of doublet predictions and CellPlex ground truth reference.* ***(D)*** *Receiver Operating Characteristic and the area under the curve (AUC) displaying the performance of both prediction methods.* ***(E)*** *Precision-Recall curve and the AUC displaying the performance of both prediction methods.* ***(F)*** *Venn diagram displaying the doublet calls between both prediction methods and the ground truth reference.*

scDblFinder performance on multiplexed liver organoid data with genotypes

Next, *scDblFinder* was evaluated using an existing multiplexed liver organoid dataset. *DoubletFinder* has not been performed on this dataset, because *scDblFinder* performed consistently better and faster than *DoubletFinder* on the intestinal organoid dataset. And there was no reason to expect a better performance of *DoubletFinder* on the liver organoids. The liver organoid dataset consisted of 12 pooled liver organoid samples labelled with CMOs. In addition, *Souporcell* was run on this dataset, since it contained different genotypes. *Souporcell* uses this information to cluster cells by genotype and additionally calls doublets (Heaton et al. 2020). Hence, this dataset contained 2 ground truths. *scDblFinder* was compared to the union of these references (Figure 8A). This approach should reduce the limitation of the cell multiplexing approach, which cannot annotate doublets within the same sample. Furthermore, the use of 12 pooled samples also reduces the chance of doublets forming within the same sample as opposed to the 4 samples used in the intestinal organoid dataset. The CellPlex ground truth consisted of 394 doublets and the *Souporcell* ground truth consisted of 683 doublets in 7,759 annotated cells. The Jaccard index between the *Souporcell* and CellPlex reference was 0.23 (Figure 8E). This resulted in a total of 876 unique ground truth doublets. The ground truths have only a small overlap and hence supplement each other.

scDblFinder predicted 910 doublets and identified 628 doublets correctly, resulting in a Jaccard-index of 0.54 (Figure 8B,E). Furthermore, *scDblFinder* achieved an AUC of 0.89 on the ROC curve and an AUC of 0.75 on the PRC curve (Figure 8C,D). Performance of *scDblFinder* on the liver organoid dataset was slightly lower compared to the performance on the intestinal organoids. However, the liver organoid dataset contained a lot less doublets and the ground truth reference was created by combining two approaches. Whilst these approaches complement each other, the limitations also add up, making the reference set bigger, at the cost of a lower Jaccard Index, a lower true positive rate (TPR) and a higher false positive rate (FPR). All in all, *scDblFinder* proved a robust performance and could now be confidently applied to the bone toxicity datasets, for which no ground truth reference is available.

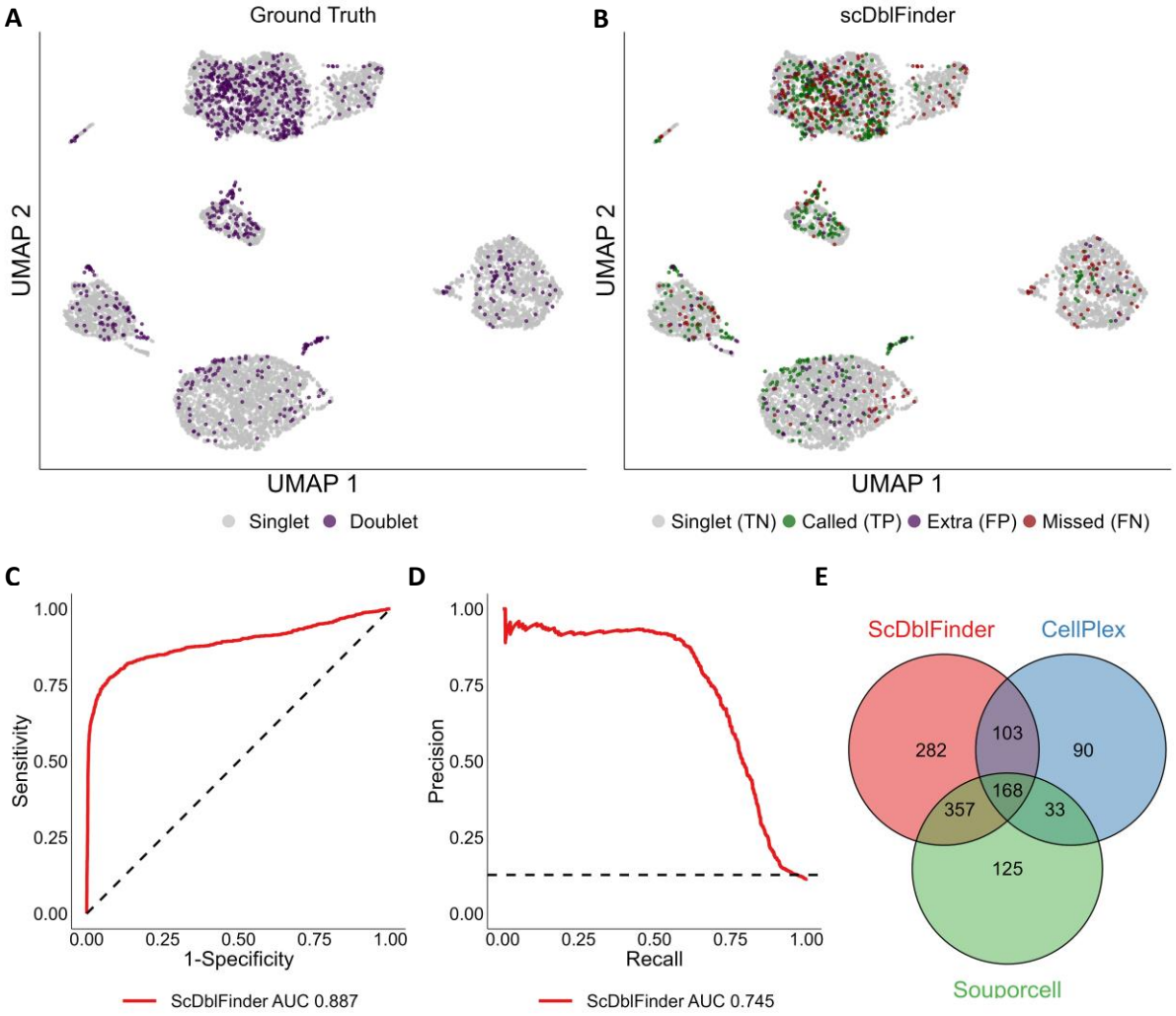

Figure 8: Overview of doublet predictions in liver organoid data. (A-B) *UMAP projection of scDblFinder predictions and combined ground truth reference.* ***(C)*** *Receiver Operating Characteristic and the area under the curve (AUC) displaying the performance of scDblFinder.* ***(D)*** *Precision-Recall curve and the AUC displaying the performance of scDblFinder.* ***(E)*** *Venn diagram displaying the doublet calls between scDblFinder and both ground truth references.*

Addressing scDblFinder variability between runs

So far scDblFinder has shown great potential, identifying about 80% of the ground truth doublets correctly. However, these results shown in the previous chapters are only after *scDblFinder* had been optimised for consistency. The *scDblFinder* doublet calls across multiple runs on the intestinal organoid data were not consistent and had on average a 0.10 difference in Jaccard overlap. These differences can be explained by the randomness in the generation of artificial doublets as well as in the gradient boosted trees approach (Germain et al. 2022).

scDblFinder was run 9 times to investigate the doublet call distribution across multiple runs. The number 9 was chosen for visualization purposes only. This revealed that on average about 78% of the doublets calls were unanimous (Figure 9A). However, the remaining 22% was not and about 12% of the total doublets were called in less than half of the runs. Next, the ground truth distribution across these runs was investigated to identify whether there was a correlation with the number of times a cell was called as doublet and this distribution (Figure 9B). This revealed that the majority of ground truth doublets were called unanimously. And that there was indeed a correlation between the ground truth distribution and the number of times a cell was called as doublet across the independent runs.

Next, the expected doublet frequency was added to the plots as baseline (Figure 9B). This baseline can be used to evaluate whether the observed ground truth distribution is significantly different from a random sample, since a random sample would contain on average the same number of ground truth doublets as the doublet frequency. This revealed that the ground truth frequency is higher than the expected doublet frequency for cells annotated at least half of the time and lower for cells annotated less than half of the time. This suggests that the observed ground truth frequency is not random. The *scDblFinder* variability between runs can effectively be reduced by creating a consensus annotation, since this averages the differences between runs. Putting the threshold for such an annotation at doublets called at least half of the time would seem logical as the ground truth contribution is higher than the expected doublet frequency up until this point. Furthermore, the doublet probability scores of each cell were averaged and plotted against the number of times a cell was called as doublet (Figure 9C). This revealed a perfect correlation, this meant that either of the two could be used to effectively threshold the doublet calls and create a consensus doublet annotation. This consensus doublet annotation has been used in the previous chapters and resulted in consistent doublet calls.

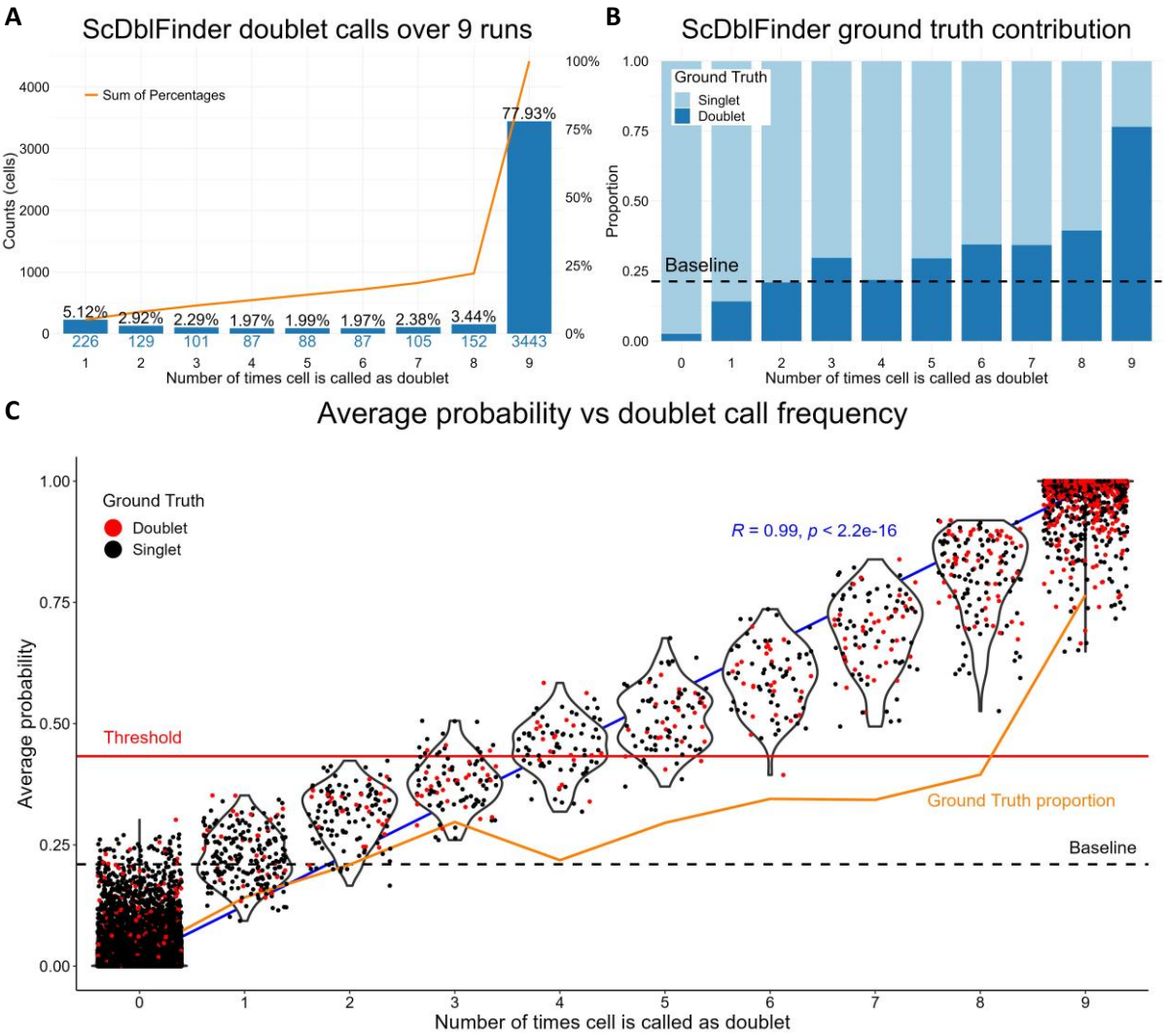

Figure 9: Overview of scDblFinder doublet calls across 9 independent runs. (A) *Barplot of doublet call frequency over 9 runs.* ***(B)*** *Ground truth contribution within each doublet call frequency.* ***(C)*** *Violin plots displaying the distribution of cells across the doublet call frequency. Furthermore, the correlation between the average doublet probability of cells and the number of times the cell was called as doublet. The baseline displays the expected doublet rate based on library size. The threshold displays the average cutoff value used by scDblFinder in these 9 runs.*

Doublet detection on bone toxicity data

The best approach to remove doublets from the bone toxicity datasets has been identified. Hence, the bone toxicity datasets, which contained many doublets, could be cleaned. Prior to doublet removal the bone toxicity datasets were merged into 1 bone toxicity dataset. This dataset contained a vehicle control group and 3 dexamethasone conditions (5, 20 and 50 mg/kg), all sequenced separately. *scDblFinder* was run on the combined bone toxicity dataset with the 4 samples supplied, since doublets cannot form between them. This meant that artificial doublet generation and processing was done separately for each sample. Nonetheless, the classifier was trained globally but the thresholds were optimized per sample (Germain *et al*. 2022). *scDblFinder* has also been run numerous times on the separate datasets but this resulted in a greater inconsistency in doublet predictions, hence the combined approach was used.

scDblFinder annotated 4,212 doublets across the 4 samples, which contained 29,994 cells in total. However, the library of dexamethasone condition 50 mg/kg contained only 2,074 cells among which 120 doublets. Furthermore, the vehicle control library contained 5,934 cells, whereas the other conditions contained around 10,000 cells. *scDblFinder* predictions revealed large doublet densities in the UMAP gene expression space. The majority of doublets were associated with the neutrophil

cluster, which was also the largest cluster in this dataset. Furthermore, the doublets are situated at the periphery of clusters, which is to be expected. There is one density of neutrophil cells entirely annotated as doublet, which is separated from the large neutrophil cluster. Furthermore, there is a density of neutrophil cells labelled as doublets positioned on top of the monocyte cluster. Furthermore, there are several protrusions or loosely positioned cells next to other clusters, annotated as doublets. The doublet probabilities are most distinct in the large libraries, whereas the smallest library contained a lot of uncertainties. However, this library contained only 2,074 cells and only few doublets are to be expected. All in all, these doublet predictions are in line with the expectation that doublets sometimes form clusters themselves but are often found at the periphery of clusters they most resemble (Germain *et al*. 2022).

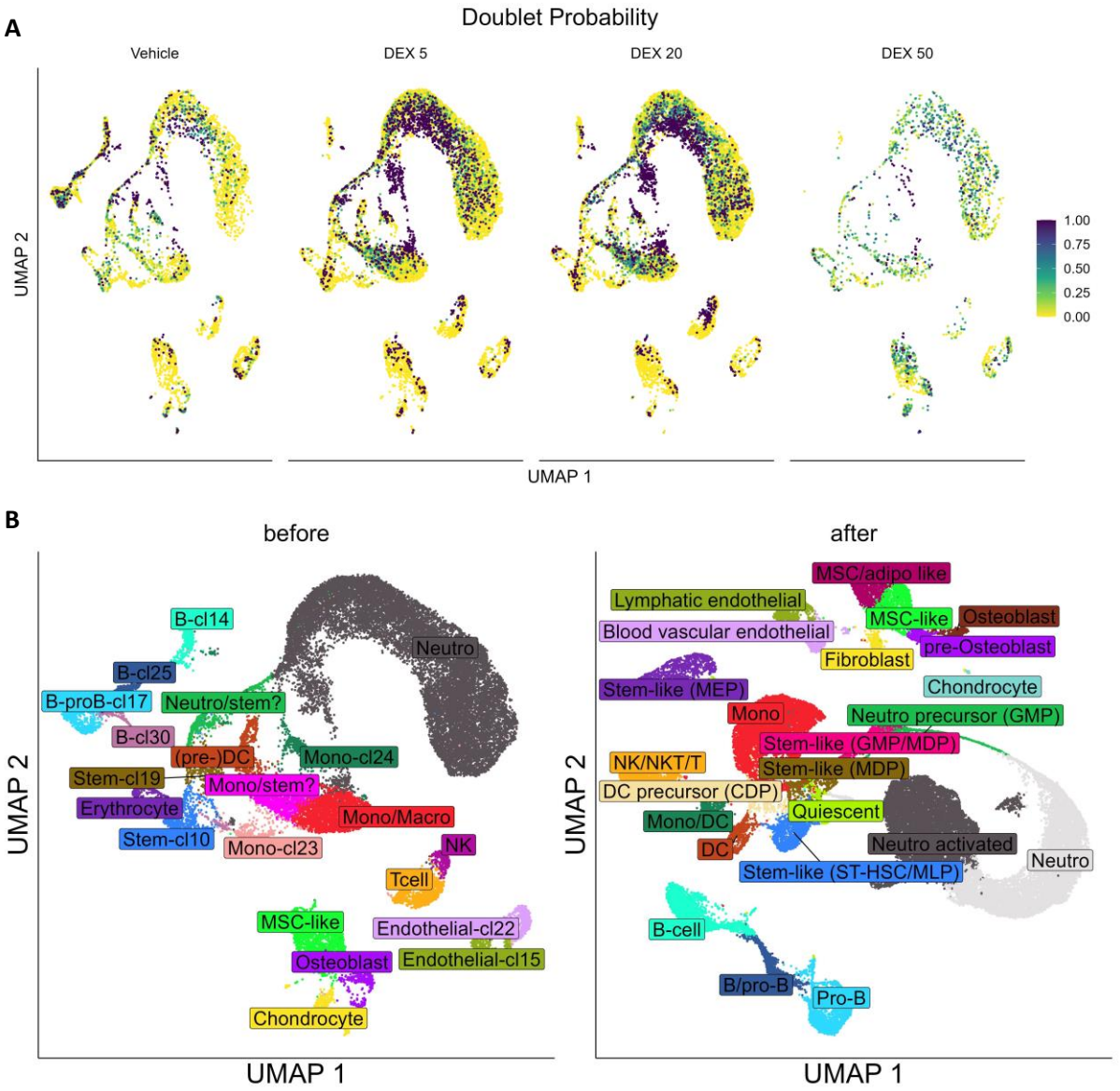

Figure 10: Doublet detection and effect of removal on bone toxicity data. (A) *UMAP projection of bone toxicity data, coloured by doublet probability score across vehicle control and dexamethasone (DEX) treatment conditions, according to ScDblFinder. Mice (n = 10/group) treated with daily dexamethasone injections of 5, 20 or 50 mg/kg for 28 days.* ***(B)*** *UMAP projection of bone toxicity data from before and after doublet removal. Clusters colours from before doublet removal are inherited by the largest overlapping cluster after doublet removal. New clusters have been given a separate colour. Cell type annotations from after doublet removal are final.*

scDblFinders doublet predictions were used to remove doublets from the combined bone toxicity dataset. The cleaned dataset was then reprocessed and a new UMAP calculated. This new UMAP

consisted of better-defined clusters compared to the old UMAP. Furthermore, the presence of loosely positioned cells had been greatly reduced. Lastly, several improvements in cell typing could be made after doublet removal which will be explained in the next chapter (Figure 10B).

Projection and cell typing of bone toxicity data

Cell labels are required to interpret the effect of dexamethasone treatment on a single-cell level. Preliminary annotations were available prior to doublet removal. However, the data has been reprocessed and cluster composition has changed, rendering the preliminary labels insufficient. In addition, 2 extra datasets have been included in the combined bone toxicity dataset. The additional datasets consist of one additional vehicle control and one untreated aged control. These datasets originated from an earlier bone toxicity project and doublet removal has been performed for both. Hence, cell typing had to be revised and these new labels were used in the UMAP from figure 10C. Hereafter, projection was used to confirm the new cell typing. The Skeletal Cell Atlas was used as a reference, since it is composed of publicly available single-cell RNA-seq datasets (Herpelinck *et al*. 2022). However, the Skeletal Cell Atlas consists of primarily data from embryonic and neonatal mice. Consequently, the predicted cell types were embryonic, which is not informative for 10 weeks old mice. Hence, an existing dataset of bone marrow stroma was selected form the Skeletal Cell Atlas and used as reference (Baryawno *et al*. 2019). This dataset was selected since it consisted of 8-10 weeks old mice and was created using the same 10x sequencing platform as the bone toxicity datasets. However, the cell type labels in this reference dataset were shallow (Figure 11A). Nonetheless, the labels were successfully transferred to the combined bone toxicity dataset, resulting in a predicted cell type based on the reference (Figure 11B). The predicted cell types confirmed the new cell typing (Figure 10A) and allowed the uncertainties from prior to doublet removal to be refined (Figure 10B). For example, the earlier identified chondrocyte population consisted of primarily fibroblasts in the new labeling, which was confirmed by the predicted cell types. Furthermore, the projected labels contained a pericyte population, which was not identified by the new cell typing. However, this population was small and not represented as a separate cluster in the bone toxicity data. Hence, this pericyte label was not transferred to the final annotations (Figure 10B). All in all, cell typing improved after doublet removal and projection of the bone marrow stroma dataset confirmed the new cell typing.

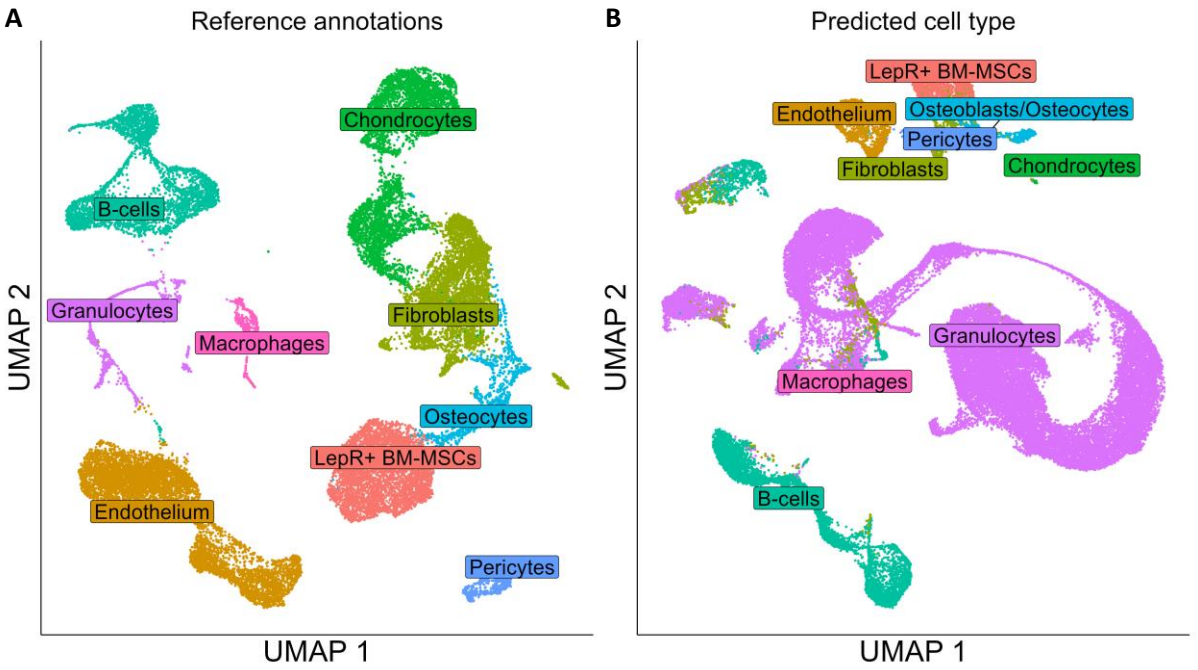

Figure 11: Projection of baryawno dataset on the bone toxicity data. (A) *UMAP projection labelled with the cell types of the baryawno dataset.* ***(B)*** *UMAP projection of the bone toxicity dataset labelled with the predicted cell types from the baryawno dataset.*

Effect dexamethasone treatment on a population level

Next, the effects of various dosages of dexamethasone treatment were investigated on a population level. The combined bone toxicity dataset consisted of 8 stromal cell populations and 16 immune cell populations (Figure 12A-C). The dataset contained 42,957 cells in total, which were distributed unequally across the 3 conditions and 2 controls (Figure S1). The immune cells were most prevalent, even though the samples were enriched for stromal cells in a 1:1 ratio pre-sequencing.

Dexamethasone treatment caused a proportional decrease in chondrocyte and lymphatic endothelial cells and nearly depleted the osteoblast population compared to the vehicle control (Figure 12A,B,D). On the contrary, the proportion of pre-osteoblast cells was increased in all dexamethasone conditions compared to the control (Figure 12A,D). The untreated aged group contained no chondrocytes but revealed a similar, yet weaker, trend for both the osteoblast and pre-osteoblast populations. However, the lymphatic endothelial cells were decreased even more in the aged control group compared to the dexamethasone conditions (Figure 12A,D).

Furthermore, dexamethasone treatment caused a proportional decrease of B-cells and depleted the pro-B and mixed B/pro-B cell populations compared to the vehicle control (Figure 12A,C,D). The aged group had less pro-B and mixed B/pro-B cells compared to the vehicle control but consisted of more B-cells compared to the vehicle and dexamethasone conditions. Next, the population of activated neutrophils increased proportionally across the dexamethasone conditions (Figure 12A,C,D). However, the proportion of activated neutrophils did not increase significantly in the aged group compared to the vehicle control. All in all, the osteoblast and pro-B cell populations appear to be most vulnerable to dexamethasone treatment, as these populations are nearly or completely depleted. Whereas the proportion of activated neutrophils and pre-osteoblast cells increased with dexamethasone treatment.

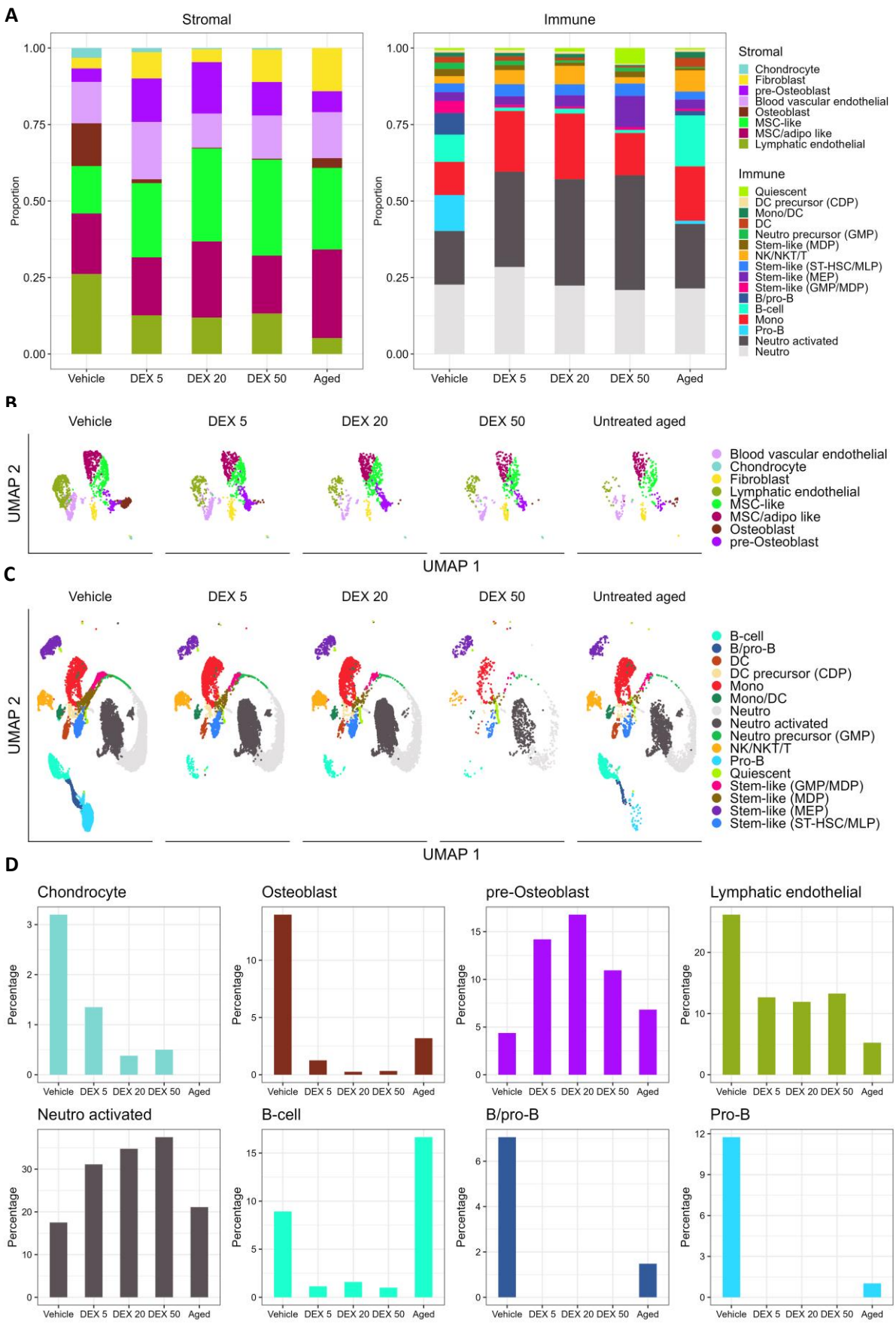
A
Stromal
Immune
Proportion
1.00
0.75
0.50
0.25
0.00
Vehicle
DEX 5
DEX 20
DEX 50
Aged
Stromal
Chondrocyte
Fibroblast
pre-Osteoblast
Blood vascular endothelial
Osteoblast
MSC-like
MSC/adipo like
Lymphatic endothelial
Immune
Quiescent
DC precursor (CDP)
Mono/DC
DC
Neutro precursor (GMP)
Stem-like (MDP)
NK/NKT/T
Stem-like (ST-HSC/MLP)
Stem-like (MEP)
Stem-like (GMP/MDP)
B/pro-B
B-cell
Mono
Pro-B
Neutro activated
Neutro
B
Vehicle
DEX 5
DEX 20
DEX 50
Untreated aged
UMAP 2
UMAP 1
Blood vascular endothelial
Chondrocyte
Fibroblast
Lymphatic endothelial
MSC-like
MSC/adipo like
Osteoblast
pre-Osteoblast
C
Vehicle
DEX 5
DEX 20
DEX 50
Untreated aged
UMAP 2
UMAP 1
B-cell
B/pro-B
DC
DC precursor (CDP)
Mono
Mono/DC
Neutro
Neutro activated
Neutro precursor (GMP)
NK/NKT/T
Pro-B
Quiescent
Stem-like (GMP/MDP)
Stem-like (MDP)
Stem-like (MEP)
Stem-like (ST-HSC/MLP)
D
Chondrocyte
Osteoblast
pre-Osteoblast
Lymphatic endothelial
Neutro activated
B-cell
B/pro-B
Pro-B
Percentage
Vehicle
DEX 5
DEX 20
DEX 50
Aged

Figure 12: Overview of stromal and immune cell populations across treatment conditions in the bone toxicity dataset. (A) *Bar plots displaying the proportional cell type composition of the stromal and immune cell subsets across dexamethasone treatment conditions and control groups.* ***(B,C)*** *UMAP projection of Immune and stromal cell populations.* ***(D)*** *Bar plots displaying the proportional contribution of independent cell types across dexamethasone treatment conditions and control groups.*

Differential expression compared to vehicle control

Several populations have been identified to be affected by dexamethasone treatment based on the cell type composition. This includes the (activated) neutrophils, (pro-)B cells, (pre-) osteoblast, chondrocytes and lymphatic endothelial cells. These populations were then investigated using the single-cell gene expression matrix to obtain molecular insights into dexamethasone induced bone toxicity. The differentially expressed genes were identified per treatment condition versus the vehicle control. Next, over-representation analysis was performed with Gene Ontology terms on each set of significantly up or downregulated differentially expressed genes (Guangchuang *et al.* 2012).

For the pre-osteoblast cells the top enriched processes were filtered on bone and osteoblast related terms to reduce ambiguous processes. This revealed an enrichment of down-regulated genes in bone development, osteoblast differentiation and bone mineralization processes (Figure 14A). The down-regulated genes within this enrichment are *Bglap*, *Bglap2* and *Rpl13* (Figure 14B). *Bglap* and *Bglap2* encode for osteocalcin production, one of the most abundant non-collagenous proteins in the bone. Furthermore, osteocalcin is used as a biochemical marker for bone formation. The increase of pre-osteoblasts cells could be explained by a decrease of pre-osteoblast differentiation, which contributes to osteoblast depletion. All in all, the decreased pre-osteoblast differentiation, bone development and bone mineralization of the pre-osteoblast cluster together with a depletion of osteoblasts indicate a reduced osteoblast activity.

The neutrophil cells had shown increased activation upon dexamethasone treatment. The top enriched processes for these clusters revealed a similar pattern (Figure 14B,C). Pattern recognition and immune signalling pathway processes are significantly enriched for up-regulated genes in both the neutrophil and activated neutrophil cluster. In addition, the neutrophil cluster was significantly enriched for leukocyte chemotaxis. Both the neutrophil and activated neutrophil cluster were significantly upregulated for the inflammatory related genes *Nfkbia and Nfkbiz* in all treatment conditions (Figure 14F, G). In addition, the neutrophils showed a significant upregulation of *Cxcl2*, encoding for the *Cxcl2* chemokine and the activated neutrophils were significantly upregulated for *Clec4e*, which encodes for a pattern recognition receptor. Furthermore, the activated neutrophil population was enriched for keratinocyte proliferation and regulation in the 50 mg/kg treatment group (Figure 14C). However, in the context of neutrophils, the up-regulated gene within this enrichment, *Lrg1*, was found to relate to granule formation (Figure 14G). All in all, the top enriched processes for the neutrophil populations confirm the observed increase in activated neutrophils.

The percentage of lymphatic endothelial cells decreased upon dexamethasone treatment. The down-regulated genes for this cluster were significantly enriched in protein production processes, such as cytoplasmic translation and ribosome biogenesis (Figure 14D). The differentially expressed genes involved in these processes were abundant and all ribosomal. To visualize the differential expression, a module score of cytoplasmic translation was calculated (Figure 14H). This revealed a decrease of cytoplasmic translation across dexamethasone treatment compared to the vehicle control. Furthermore, the module score of the aged control group was also found to be decreased (Figure 14H). All in all, lymphatic endothelial cells were found to have a decreased function in dexamethasone treated conditions compared to the vehicle control.

Pro-B cell and B/pro-B cell clusters were completely depleted in dexamethasone treated conditions. Hence, differential gene expression is not possible for these clusters. Furthermore, the B-cell cluster consisted of too few cells per condition to obtain insightful results. Instead, the B-cell development of

the vehicle condition was investigated to confirm the cell typing and consequently the absence of pre- and pro-B cells in the dexamethasone treated conditions. For this purpose, trajectory inference was used to determine the pseudotime ordering of B-cells (Figure 13A) (Kelly *et al.* 2018). Subsequently, the genes were correlated against this pseudotime. The top pseudotime correlated genes consisted of several MHC class II related genes, which are involved in B-cell affinity, maturation and activation (Figure 13B). Furthermore, several genes involved in B-cell receptor development. As well as several genes encoding for cluster of differentiation molecules, with some expressed exclusively in pro- and pre-B cells (*Vpreb1* and *Igll1*). All in all, the pseudotime ordering correlates with gene expression found in adult B-cells and anticorrelates with gene expression found in pro- and pre-B cells. This confirms the presence of a B-cell trajectory and consequently confirms the cell typing. The absence of pre- and pro-B cells in the dexamethasone treated conditions suggest a detrimental effect of dexamethasone treatment on B-cell differentiation in the bone marrow.

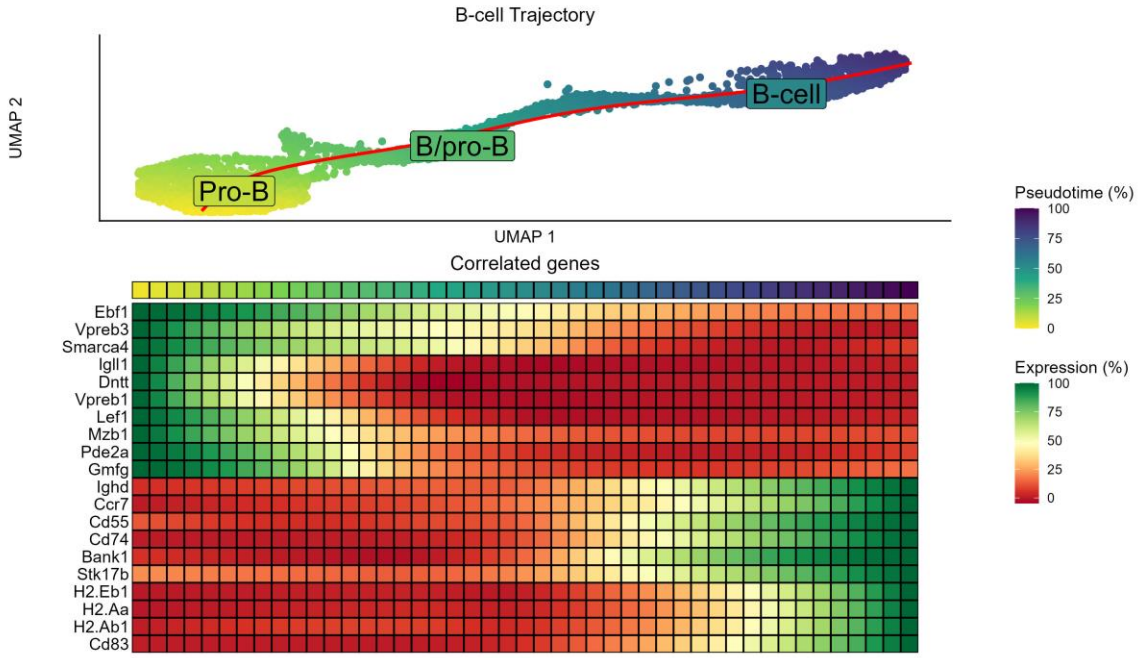

Figure 13: UMAP of B-cell trajectory and heatmap of pseudotime correlated genes. (A) *Trajectory of B-cell development from pro-B to Mature B-cells. Cells have been assigned a pseudotime score to represents their stage of B-cell development. UMAP 1 coordinates has been flipped to represent the development from left to right.* ***(B)*** *Heatmap visualization of the top 10 pseudotime correlated and anticorrelated genes.*

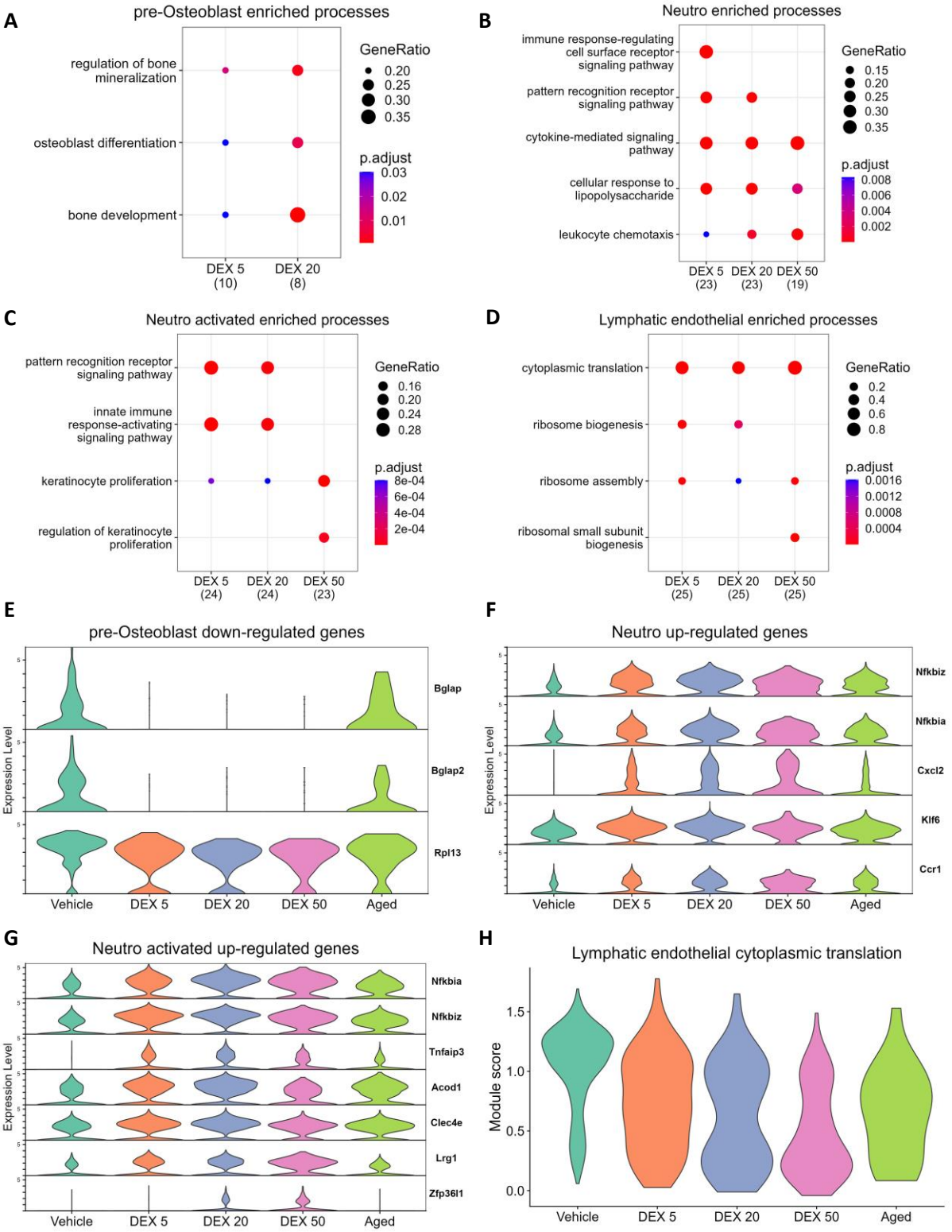

Figure 14: Dot plots of enriched processes and associated differentially expressed genes compared to the vehicle control. ***(A-D)*** *Combined overview of the top processes enriched with differentially expressed genes for each treatment condition versus the vehicle control. pre-Osteoblast results were filtered on bone and osteoblast related terms. The other results were filtered on the top 2 most significantly enriched processes for each treatment condition. The dot size relates to the number of differentially expressed genes. The total number of differentially expressed genes is indicated below each treatment condition.* ***(E-F)*** *Violin plots displaying the expression of the differentially expressed genes involved in the enriched processes per sample. The displayed genes are significantly expressed in at least 2 of the 3 treatment conditions, with the exception of neutrophils. The neutrophil genes are significantly expressed in all 3 treatment conditions.* ***(H)*** *Module score of cytoplasmic translation in the lymphatic endothelial cell cluster per sample.*

Differential expression across treatment groups

The effects of dexamethasone treatment have now been investigated on a single-cell level per sample condition compared to the vehicle control. The next step was to identify dosage dependent effects by looking at the differential gene expression across treatment conditions. This was done by calculating the Spearman's rank correlation of the differential gene expression per treatment condition versus the dexamethasone dosage. Hereafter, the significantly up- or down-regulated genes across treatment conditions were investigated using over-representation analysis. Furthermore, a module score of these genes was calculated to visualize a potential gradient of dexamethasone treatment in the recalculated UMAP gene expression space of each cluster. For many cell types the enriched processes were related to common cell processes, such as cell cycle, and did not show a clear pattern of dexamethasone induced bone toxicity.

However, the up-regulated genes of the neutrophil cluster were again enriched for immune related processes, suggesting heightened neutrophil activation across dexamethasone treatment (Figure 15A). The module score of up-regulated genes in the neutrophil cluster displayed a gradient from left to right In the UMAP space (Figure 15B). However, this gradient was also present in the vehicle control and correlated strongly with the cell cycle phase, suggesting that this gradient is likely related to cell cycle progression and not to dexamethasone treatment (Figure 15C).

Finally, the down-regulated genes of the B-cell cluster were enriched for DNA damage repair as well as G0 to G1 transition (Figure 15D). This could mean that the B-cells are becoming more dormant with dexamethasone treatment. However, previous results displayed a rapid decline of B-cells upon treatment with 5 mg/kg dexamethasone (Figure 12D). Thus, this result is based on the remaining B-cells only, since the majority of B-cells have already died. Hence, the remaining B-cells might become more dormant and eventually go into apoptosis or were less susceptible because they were more dormant. All in all, the cell cycle progression is impaired. Furthermore, there appeared to be no gradient related to the module score of down-regulated genes in the UMAP space of the B-cell cluster (Figure 15E,F).

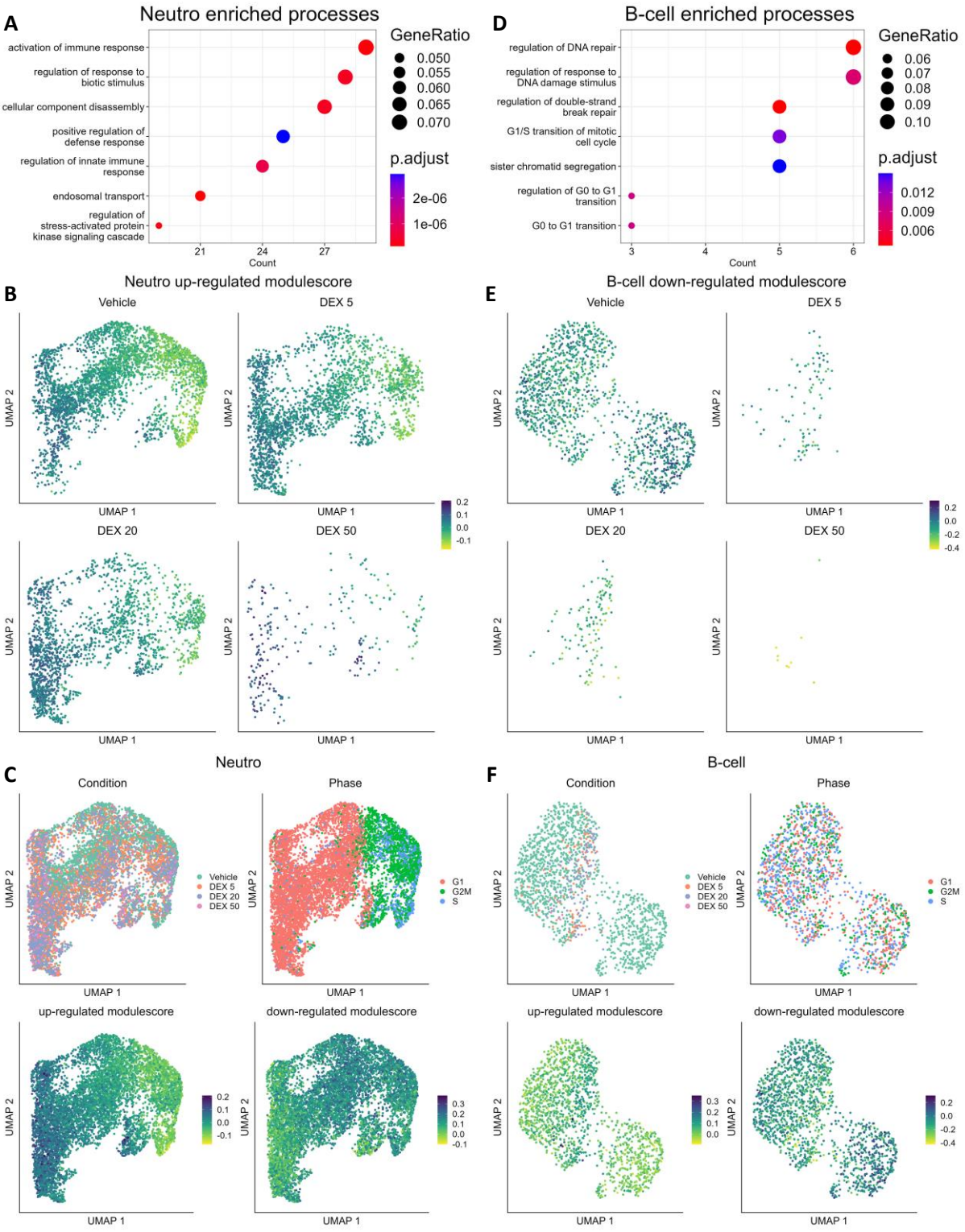

Figure 15: Differential expression across treatment conditions versus vehicle control. (A,D) *Dot plots of the top processes enriched for differentially expressed genes per cluster across treatment groups versus the vehicle control. Count displays the number of genes associated with each process.* ***(B,E)*** *UMAP visualization of module score created with all the up-regulated or down-regulated differentially expressed genes.* ***(C,F)*** *UMAP visualization of treatment condition, cell cycle phase and the module scores for up and down-regulated differentially expressed genes.*

Discussion

The aim of this study was to first evaluate and improve upon existing methods to identify doublets in single-cell RNA-seq data. Secondly, to apply the best approach to remove doublets to the bone toxicity datasets. And finally, to perform subsequent single-cell analysis of the bone toxicity datasets with the aim of revealing the underlying molecular mechanisms of dexamethasone induced bone toxicity in skeletally immature mice.

DoubletFinder and *scDblFinder* were selected as best candidates for doublet removal at the time of writing (McGinnis *et al*. 2019, Germain *et al*. 2022). This study revealed that *scDblFinder* consistently outperformed *DoubletFinder* with higher speed and greater accuracy on the multiplex intestinal organoid data. This finding was expected, since the *scDblFinder* paper reports to outperform *DoubletFinder* on most datasets (Germain *et al*. 2022). However, since there is no method that systemically outperforms the others and since data generation in the Maxima might favour one method over the other, this statement had to be investigated. The reason that *scDblFinder* performs better than *DoubletFinder* is likely because *scDblFinder* does not use a fixed neighbourhood size but allows the downstream classifier to select the most informative size. Furthermore, because *scDblFinder* does not average the expression of artificial doublets, whereas *DoubletFinder* does. Besides, *scDblFinder* performs consecutive rounds of doublet removal in which confidently called doublets are removed each round to prevent overfitting (Germain *et al*. 2022). All in all, *scDblFinder* is more optimized than *DoubletFinder* and achieves a higher accuracy with much greater speed. For this reason, *scDblFinder* was selected as optimal doublet detection approach and was further investigated using the liver organoid dataset.

However, the *scDblFinder* doublet calls were not consistent across multiple runs. There was on average a 0.10 difference in Jaccard overlap between independent runs. This difference could effectively be reduced by summing *scDblFinder* calls over multiple independent runs and creating a new annotation in which a cell has to be annotated as doublet in at least half of the runs. Using this approach, *scDblFinder* consistently identified well over 80% of the ground truth doublets, which is comparable to the performance reported in the scDblFinder paper (Germain *et al*. 2022). The *DoubletFinder* paper reports that *DoubletFinder* can accurately predict heterotypic doublets with more than 90% sensitivity (McGinnis *et al*. 2019). This study observed a lower sensitivity. However, the statement of the *DoubletFinder* paper only considers the heterotypic doublets, hence the performance including homotypic doublets is likely comparable.

Both the *scDblFinder* and *DoubletFinder* paper imply the clustered doublet generation approach to perform best (McGinnis *et al*. 2019, Germain *et al*. 2022). However, in this study the random doublet generation approach yielded slightly better results. Supposedly the clustered approach ignores homotypic doublets, which results in a higher performance as both methods cannot accurately predict homotypic doublets due to their insignificant divergence from real cells in gene expression space. Yet, the random approach worked best for both the Intestinal and Liver organoid datasets. It is likely that both methods can still identify at least part of the homotypic doublets. And since this resulted in a slightly higher performance, the random approach was used throughout this study.

However, the ground truth on which these results are based is not perfect, so it is likely that the performance of both *scDblFinder* and *DoubletFinder* is slightly higher than observed. The ground truth used for the intestinal organoid data was created with the 10x CellPlex technique that makes use of oligonucleotides (*Cell Ranger*). This technique is limited by its inability to annotate doublets within the same sample. Whereas the prediction methods can predict doublets within the same sample, if there is a significant difference in gene expression. Yet, these doublets will be wrongly annotated as false positives. This likely happened in the intestinal organoid data as there was an observed 772 cell overlap

between *DoubletFinder* and *scDblFinder* of cells that were not annotated by the ground truth. However, it is likely that these cells are doublets since they were confidently called by both approaches.

To overcome this limitation *scDblFinder* was also tested on a liver organoid dataset for which a CMO based ground truth as well as a SNP based ground truth was available. While these ground truths did complement each other, the performance of *scDblFinder* was slightly lower. This decrease in performance might be related to the combination of the 2 ground truths, as the limitations of both methods add up. The genotypically multiplexed ground truth is limited by its inability to annotate doublets with the same genotype (Kang *et al.* 2018). However, the difference in performance between the intestinal and liver organoid dataset could also be biological or technical, since the intestinal organoid dataset is twice as big as the liver organoid dataset. All in all, taking into account the limitations of both ground truths and the limitation of *scDblFinder* in its ability to detect homotypic doublets, *scDblFinder* showed a robust performance annotating 73% of the ground truth doublets correctly.

The majority of true positive doublets are positioned at the periphery or edges of clusters in the UMAP gene expression space, as well as in small doublet clusters. This description fits with the expected location of heterotypic doublets (Germain *et al*. 2022). The majority of false negatives are located within clusters, these doublets are likely homotypic. The performance of *scDblFinder* in detecting homotypic doublets is limited, which is to be expected. Perhaps a method like *BIRD*, which relies on heterologous SNPs could be investigated on its ability to detect homotypic doublets (Wainer-Katsir *et al*. 2020). Furthermore, the recently published deep learning *SoCube* approach could be investigated as it reports to outperform *scDblFinder,* especially on larger libraries (Zhang *et al*. 2023). The *SoCube* method includes a homotypic-doublets-first simulation, which might result in a better performance on homotypic doublets. Nevertheless, *scDblFinder* achieved great accuracy in detecting heterotypic doublets and these doublets are most important, since homotypic doublets are less harmful.

scDblFinder was then applied to the bone toxicity data, this revealed large doublet densities at the periphery of clusters, lots of loosely positioned doublets as well as several doublet clusters or protrusions on the UMAP projection. All these locations are in line with the expected positions of doublets. Furthermore, the number of doublets per library were also in line with the *scDblFinder* expected doublet frequency of 1% per 1,000 cells (Germain *et al*. 2022). However, on the smaller libraries the uncertainty in doublet calls was slightly elevated. Nonetheless, the *scDblFinder* predictions were plausible and annotated doublets were removed from the libraries. Hereafter, clustering of bone toxicity data visually improved, with better defined cluster edges and less loosely positioned cells. Furthermore, uncertainties in cell typing could be refined after doublet removal. This was confirmed by projection of the Baryawno dataset (Baryawno *et al.* 2019). However, the labels of this dataset were only superficial. The Skeletal Cell Atlas is more extensive, but consists of primarily embryonic data, which is not the best representation of the bone toxicity dataset. Moreover, the full Skeletal Cell Atlas has yet to be published (Herpelinck *et al*. 2019). Perhaps in the future, the full reference dataset can be used to further refine cell typing of the bone toxicity dataset.

Next, the effects of dexamethasone treatment on the developing bone were investigated. This study identified a consistent decrease of osteoblasts, chondrocytes, lymphatic endothelial cells and B-cells across all dosages of dexamethasone treatment. Furthermore, the proportion of pre-osteoblasts and activated neutrophil cells was found to be increased across all dosages of dexamethasone treatment.

The increase of pre-osteoblasts cells however was accompanied by a decrease in bone mineralization, osteoblast differentiation and bone development. The corresponding downregulated genes (*Bglap* and *Bglap2*) encode for osteocalcin, which is used as a marker for bone formation. This suggests that in skeletally immature mice dexamethasone treatment causes a diminished bone formation. Diminished

bone formation has also been identified in skeletally mature mice (Kim *et al.* 2007), however skeletally mature mice also display heightened bone resorption, which has not been identified in this study. This could be a key difference of dexamethasone induced bone loss in skeletally immature mice. Furthermore, the decrease in osteoblast differentiation likely contributes to the low osteoblast numbers in the dexamethasone treated conditions, since no new osteoblast cells are formed. Moreover, the reduction in osteoblasts has been experimentally validated by the Janda group (Figure S2). The number of osteoblasts per square millimetre is significantly less in dexamethasone treated conditions compared to the vehicle control.

The neutrophil cells demonstrated increased activation upon dexamethasone treatment as well as a proportional increase in cell number. This was accompanied by upregulation of immune processes and immune related genes. Glucocorticoids such as dexamethasone are known to prevent apoptosis in neutrophil cells. Furthermore, glucocorticoids are known to have both anti and pro-inflammatory effects on the neutrophil population (Ronchetti *et al*. 2018). However, a direct correlation between dexamethasone induced bone toxicity and neutrophil cells has not been identified in this study. Furthermore, the B-cell population decreased significantly and B-cell precursors were completely absent in dexamethasone treated conditions. This suggest that dexamethasone has a detrimental effect on B-cell development. While differential gene expression was not possible for B-cells due to the low cell numbers, differential expression across treatment groups identified a decrease of DNA damage repair as well as cell cycle progression in the remaining B-cells. This suggests that these B-cells are under stress and are either becoming dormant or going into apoptosis as well. The detrimental effects of dexamethasone treatment on developing B-cells have been investigated in a study by Gruver-Yates. This studied revealed that immature B-cells abundantly express glucocorticoid receptors (Gruver-Yates *et al.* 2014). This explains the immediate decrease of B-cell precursors upon dexamethasone treatment. All in all, dexamethasone treatment disrupted B-cell development and consequently impaired the immune system. While this study did not identify a clear relationship between B-cells and dexamethasone induced bone development, the bone marrow microenvironment is clearly affected by dexamethasone treatment.

Conclusion

In conclusion, this study reports that *scDblFinder* can be used to predict doublets with greater speed and higher accuracy than *DoubletFinder*. However, the randomness in *scDblFinders* doublet generation and machine learning approach causes it to have a high variability between runs. This variability can effectively be reduced by summing multiple *scDblFinder* runs and annotating cells that are called half of the time. Using this approach *scDblFinder* consistently identified about 80% of the ground truth doublets. *scDblFinder* was successfully applied to the bone toxicity data, this cleaned the data and allowed for several improvements in cell typing to be made. Furthermore, this study suggests that dexamethasone treatment in developing bones resulted in a decrease of several populations, including chondrocytes, osteoblasts, lymphatic endothelial cells and B-cells. Next, neutrophil cells increased proportionally and displayed increased immune activation. The pre-osteoblast cells also increased proportionally but showed decreased differentiation and bone development with reduced osteocalcin production. This data suggests that dexamethasone induced bone toxicity in the developing bone is caused primarily by a decrease of bone formation. Whereas in adult bone increased bone resorption is observed as well. All together, dexamethasone treatment alters the bone marrow microenvironment and disrupts healthy bone development.

Materials & Methods

Data acquisition

Bone toxicity datasets used throughout this study were kindly provided by the Janda group (Warmink *et al*.). These datasets were obtained with droplet-based single-cell RNA-seq using the 10x Genomics platform. Samples used for single-cell RNA-seq come from the isolated bone of young mice (6-10 weeks) treated with oral dosages of dexamethasone (5, 20 and 50 mg/kg). Furthermore, a vehicle control group as well as an untreated aged group were included. The bone toxicity datasets were pre-processed and preliminary cluster annotations were available. Clustering was performed with the FindNeighbors and FindClusters functions from the Seurat package. And cluster annotations were obtained using SingleR (1.10.0) (Aran *et al.* 2019) with MouseRNAseqData and ImmGenData reference datasets.

The intestinal organoid dataset stems from an in-house study on the comparison of human intestinal organoids and was kindly provided by the Single-Cell Genomic facility (Prinses Máxima Centrum). This dataset was generated with the 10x Genomics CellPlex technique and included 6 CMO's. Demultiplexing was performed during data analysis using *Cell Ranger*. Cells that were positively assigned to different CMO's were annotated as doublet. This annotation was used as a ground-truth reference in this study.

The liver organoid dataset comes from an in-house project on molecular characterization of hepatoblastomas and normal postnatal livers and was kindly provided by the Single-Cell Genomics facility (Prinses Máxima centrum). This dataset was genotypically multiplexed (Heaton et al. 2020) and also included 12 CMO's. Demultiplexing was performed using *Cell Ranger* and *Souporcell* was used to call genotypes. Cells that match two genotypes or are positively assigned to different CMO's were annotated as doublet. This combined doublet annotation was used as ground-truth reference throughout this study.

Data processing

Data processing for the Intestinal and Liver organoid dataset were performed in R (4.3.1) using the Seurat package (4.3.0) (Hao *et al*. 2021). The Read10X_h5 function from Seurat was used to load count data. Hereafter counts were filtered on mitochondrial reads and a percentage of mitochondrial expression was calculated. Cells with >20% mitochondrial expression were discarded. Next, count data was filtered on having more than 200 UMI. This filtering threshold was deliberately kept low to increase *scDblFinder* performance as described in the *scDblFinder* vignette (Germain *et al*. 2021). Setting the UMI threshold higher will affect the expected doublet rate resulting in an underestimation of doublets by *scDblFinder*.

Next data was normalized and scaled using the NormalizeData and ScaleData function from Seurat with the 'LogNormalize' method. The top 2000 variable features were identified using the FindVariableFeatures function. For visualisation purposes data was also transformed using the SCTransform function. Deconfounding was performed by excluding known confounders and correlated genes from the variable genelist. Known confounders were identified with the *Cell Ranger* reference data of GRCh38_3.0.0 included in the SCutils (1.120) package. Hereafter, cell phase was identified with the CellCycleSoring function from Seurat. Cell cycle correlated genes were identified with the metadataCorrelations function from the SCutils package. Hereafter, the top 30 principal components were identified using the RunPCA function and a UMAP was added using the RunUMAP function with dims = 20. Next doublets were identified and removed according to the methodology described in the Doublet detection section. After doublet removal, data was refiltered on UMI count using a violin plot

to determine the optimal threshold and reprocessed in the same manner. Clusters were identified using the FindNeighbors and FindClusters functions from Seurat.

Doublet detection and analysis

DoubletFinder (2.0.3) (McGinnis *et al*. 2019) was applied to the filtered intestinal organoid data. This was done according to the workflow described in the *DoubletFinder* github page. pK identification was performed without ground-truth to test *DoubletFinder* when no ground-truth is available. The parameter sweep identified an optimal pK value of 0.005. The expected doublet rate of 1% per 1000 cells was used instead of the homotypic doublet proportion estimate since this yielded better results. *DoubletFinder* has not been applied to the filtered liver organoid data, since *scDblFinder* performed better and much faster than *DoubletFinder* and there was no reason to suspect a better performance of *DoubletFinder* on the liver organoids.

scDblFinder (1.15.1) (Germain *et al*. 2021) was applied to both the filtered intestinal organoid data as well as the filtered liver organoid data. This was done using the optimised default settings of version 1.15.1, these settings are different from the base settings in earlier versions of *scDblFinder* (<1.10). The optimised settings yielded higher results than the base settings from earlier versions. *scDblFinder* was run on the RNA assay and data slot of the Seurat object using the GetAssayData function from Seurat. *scDblFinder* was run 9 times independently on both datasets. Hereafter, cells annotated in 5 or more independent runs were annotated as doublet. This was done to reduce the variability between independent *scDblFinder* calls as explained in the Results section. There was no significant difference between running *scDblFinder* 9 or 10 times, hence 9 was chosen for visualisation purposes.

Hereafter, the doublet calls were compared to the ground truth references. This was done by calculating the Jaccard index. Model performance was evaluated using the ROC, PRC and the AUC. Furthermore, both models were evaluated on their speed and usability. The metrics of both methods were then compared to each other. *scDblFinder* outperformed *DoubletFinder* on the intestinal organoid data and achieved much greater speed, even when running it 9 times. Furthermore, *scDblFinder* did not require a parameter sweep. Hence, *scDblFinder* was applied to detect doublets in the bone toxicity datasets.

Projection

Cell typing of the bone toxicity datasets was refined with cell projecting (Stuart *et al.* 2019) of a reference dataset onto the bone toxicity data. This was done with the FindTransferAnchors and TransferData functions from Seurat. The mouse bone marrow stroma dataset from Baryawno *et al*. was used for this purpose. This dataset is included in the Skeletal Cell Atlas, which was not used since the projected labels were non informative. The Baryawno reference was constructed according to the scripts available at the Skeletal Cell Atlas github page (Herpelinck *et al*. 2022).

Differential gene expression

Differentially expressed genes were identified using FindMarkers from Seurat with logfc.threshold = log2(1.5) and min.cells.group = 10. Differentially expressed genes were identified per cluster and condition versus the vehicle. The genes were then filtered on a p.val.adj < 0.1. The top 25 differentially expressed genes were submitted to gene ontology overrepresentation analysis using enrichGO from clusterProfiler (4.8.2) (Wu *et al.* 2021). Differential gene expression across treatment groups was identified by first calculating the average fold change of genes per cluster in each treatment condition versus the vehicle. The features were filtered on a minimum expression of 10 percent in each cluster. Hereafter, the Spearman's rank correlation was calculated over the average fold change of each gene per cluster versus the treatment condition group (1:4). Lastly, genes were filtered on a perfect Spearman's rank correlation of +1 or -1. The expression of genes with a perfect rank correlation either

increases or decreases across treatment groups versus the vehicle control. These genes were then submitted to gene ontology overrepresentation analysis using enrichGO.

B-cell trajectory inference

The B-cell trajectory was identified by first creating a separate Seurat object for only the B-cells. Hereafter, the B-cells were re-clustered with a Gaussian mixture model using Mclust (6.0.0) (Scrucca *et al*. 2018). Then, Slingshot (2.8.0) (Street et al. 2018) was used to infer the pseudo time of B-cells. Lastly, the Pearson correlation was calculated between gene expression and the pseudo time to identify the top 10 pseudo time correlating and anticorrelating genes. These genes were then investigated on their involvement in B-cell development.

Code availability

The source code used to perform the analysis and to generate the figures as well as an html file to reproduce the doublet detection of *scDblFinder* on the intestinal organoids are available from: https://github.com/Tristan891/Master-Internship

Acknowledgements

First of all, I would like to thank Philip for his supervision during my major research project. Thank you for making this interesting project possible. Furthermore, I want to thank Kelly and Claudia for letting me be involved in their bone toxicity project. Thank you for this opportunity to create many great figures and broaden my R skills. Next, I want thank Patrick for being my second examiner. Furthermore, I want to thank the people of the Single-Cell Genomics facility for their input and feedback on my project. Lastly, I would like to thank the corner office for letting me be involved in their social activities.

References

Amezquita RA, Lun ATL, Becht E, et al. Orchestrating single-cell analysis with Bioconductor [published correction appears in Nat Methods. 2019 Dec 11;:]. Nat Methods. 2020;17(2):137-145. doi:10.1038/s41592-019-0654-x

Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163-172. doi:10.1038/s41590-018-0276-y

Baryawno N, Przybylski D, Kowalczyk MS, et al. A Cellular Taxonomy of the Bone Marrow Stroma in Homeostasis and Leukemia. Cell. 2019;177(7):1915-1932.e16. doi:10.1016/j.cell.2019.04.040

Bloom JD. Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. PeerJ. 2018;6:e5578. Published 2018 Sep 3. doi:10.7717/peerj.5578

Cell Ranger. What is Cell Multiplexing? -Software -Single Cell Gene Expression -Official 10x Genomics Support. Retrieved February 15, 2023, from https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-CellPlex

Ferrara G, Petrillo MG, Giani T, et al. Clinical Use and Molecular Action of Corticosteroids in the Pediatric Age. Int J Mol Sci. 2019;20(2):444. Published 2019 Jan 21. doi:10.3390/ijms20020444

Germain PL, Lun A, Garcia Meixide C, Macnair W, Robinson MD. Doublet identification in single-cell sequencing data using scDblFinder. F1000Res. 2021;10:979. Published 2021 Sep 28. doi:10.12688/f1000research.73600.2

Gruver-Yates AL, Quinn MA, Cidlowski JA. Analysis of glucocorticoid receptors and their apoptotic response to dexamethasone in male murine B cells during development. Endocrinology. 2014;155(2):463-474. doi:10.1210/en.2013-1473

Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29. doi:10.1016/j.cell.2021.04.048

Heaton H, Talman AM, Knights A, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat Methods. 2020;17(6):615-620. doi:10.1038/s41592-020-0820-1

Herpelinck, T. et al. (2022) An integrated single-cell atlas of the skeleton from development through adulthood, bioRxiv. Available at: https://www.biorxiv.org/content/10.1101/2022.03.14.484345v1

Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11(2):163-166. doi:10.1038/nmeth.2772

Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation [published correction appears in Nat Biotechnol. 2020 Nov;38(11):1356]. Nat Biotechnol. 2018;36(1):89-94. doi:10.1038/nbt.4042

Kim HJ, Zhao H, Kitaura H, et al. Glucocorticoids and the osteoclast. Ann N Y Acad Sci. 2007;1116:335-339. doi:10.1196/annals.1402.057

McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019;8(4):329-337.e4. doi:10.1016/j.cels.2019.03.003

Oray M, Abu Samra K, Ebrahimiadib N, Meese H, Foster CS. Long-term side effects of glucocorticoids. Expert Opin Drug Saf. 2016;15(4):457-465. doi:10.1517/14740338.2016.1140743

Ronchetti S, Ricci E, Migliorati G, Gentili M, Riccardi C. How Glucocorticoids Affect the Neutrophil Life. Int J Mol Sci. 2018;19(12):4090. Published 2018 Dec 17. doi:10.3390/ijms19124090

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. Published 2015 Mar 4. doi:10.1371/journal.pone.0118432

Scrucca L., Fraley C., Murphy T. B. and Raftery A. E. (2023) Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman & Hall/CRC, ISBN: 978-1032234953, https://mclust-org.github.io/book/

Stoeckius M, Zheng S, Houck-Loomis B, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018;19(1):224. Published 2018 Dec 19. doi:10.1186/s13059-018-1603-1

Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477. Published 2018 Jun 19. doi:10.1186/s12864-018-4772-0

Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

Wainer-Katsir K, Linial M. BIRD: identifying cell doublets via biallelic expression from single cells. Bioinformatics. 2020;36(Suppl_1):i251-i257. doi:10.1093/bioinformatics/btaa474

Ward LM. Glucocorticoid-Induced Osteoporosis: Why Kids Are Different. Front Endocrinol (Lausanne). 2020;11:576. Published 2020 Dec 16. doi:10.3389/fendo.2020.00576

Warmink K, Janda C, et al. Dexamethasone induced bone toxicity. Manuscript in preparation

Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141. Published 2021 Jul 1. doi:10.1016/j.xinn.2021.100141

Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-287. doi:10.1089/omi.2011.0118

Zhang H, Lu M, Lin G, et al. SoCube: an innovative end-to-end doublet detection algorithm for analyzing scRNA-seq data. Brief Bioinform. 2023;24(3):bbad104. doi:10.1093/bib/bbad104

Supplementary Material

Layman's Summary

Treatment of paediatric cancer can result in late effects. Late effects are health problems that occur after cancer treatment has ended. One drug commonly used during treatment of paediatric cancers is dexamethasone. Dexamethasone has an anti-inflammatory effect. However, prolonged exposure to dexamethasone can lead to various health problems, including impaired bone development. The effects of dexamethasone treatment have been well studied in adult bone but not yet in young bone. There might be substantial differences in the effects of dexamethasone treatment between adult and young bone. To study this effect in young bone, young mice were treated with various dosages of dexamethasone. Hereafter, single-cell RNA sequencing was used on the tissue samples of these mice. Single-cell RNA sequencing is a technique used to capture and sequence individual cells. This results in cell-specific expression profiles. The expression profiles of dexamethasone treated and untreated cells can be compared to identify cell type specific effects of dexamethasone treatment. However, single-cell RNA sequencing data contains doublets. Doublets are technical artifacts which form when two cells are captured together. The doublet formation depends on the number of cells used during sequencing. Too many cells were loaded during the sequencing of the mice tissue samples, hence this data contains many doublets. Doublets cannot be recognized easily and compromise the data analysis. Therefore, they have to be detected and removed. *DoubletFinder* and *scDblFinder* are two existing doublet detection methods. These methods create artificial doublets by randomly picking and combining two real cells into one. Then these methods look at the gene expression of cells and identify the differences between the artificial doublets and the real cells. Cells that are closely related to the artificial doublets are likely to be doublets and hence annotated as doublet. *DoubletFinder* and *scDblFinder* differ in their machine learning approach to annotate doublets as well as in the generation of artificial doublets. *scDblFinder* sums and reweights cells, whereas *DoubletFinder* averages cells. The aim of this study was to evaluate the performance of *DoubletFinder* and *scDblFinder*. The best approach could then be applied to clean the single-cell RNA sequencing data. This study identified *scDblFinder* to perform better and faster than *DoubletFinder* on existing datasets for which an experimental ground truth reference was available. Hereafter, the single-cell RNA sequencing data was cleaned of doublets and the effect of dexamethasone was investigated. This was done by first annotating the cell types and then looking at the differences in expression profiles between the same cell types in treated and untreated cells. Annotating the cell types was done by comparing the expression profiles to a reference dataset for which the cell types are known and then importing the known labels of similar expression profiles to the dataset. This study reports that dexamethasone negatively affects several cell types, including chondrocytes and osteoblasts, these cells are found within the bone and contribute to bone formation. In adult mice dexamethasone has been reported to cause both a decrease in bone formation and an increase in bone resorption. The latter has not been identified in this study and could be a key difference between adult and young bone.

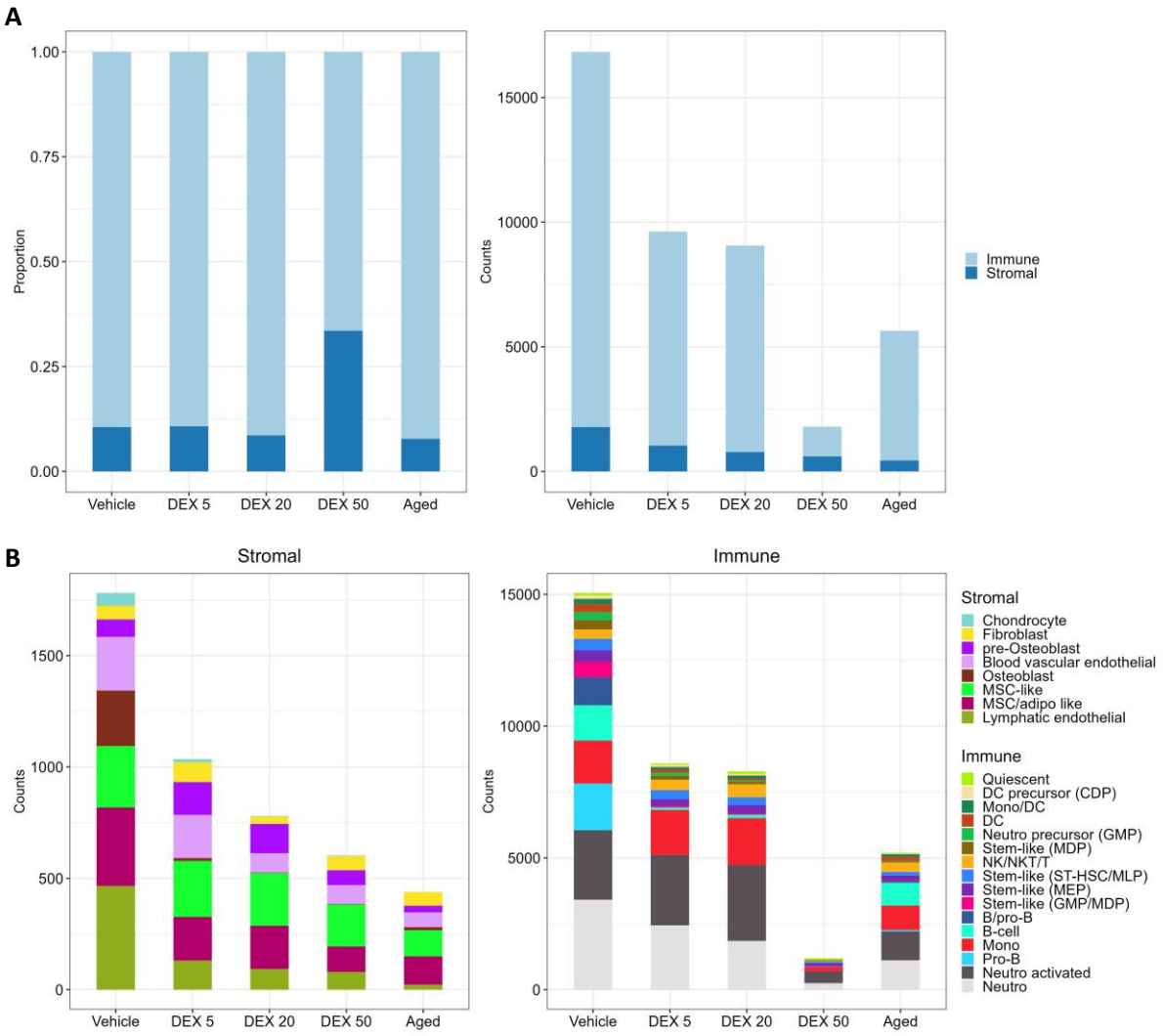

Figure S1: Absolute overview of stromal and immune cell populations in the bone toxicity dataset. (A) *Bar plots displaying the proportional and absolute cell counts of stromal versus immune cells across dexamethasone treatment conditions and control groups.* ***(B)*** *Bar plots displaying the absolute cell counts of stromal and immune cells across dexamethasone treatment conditions.*

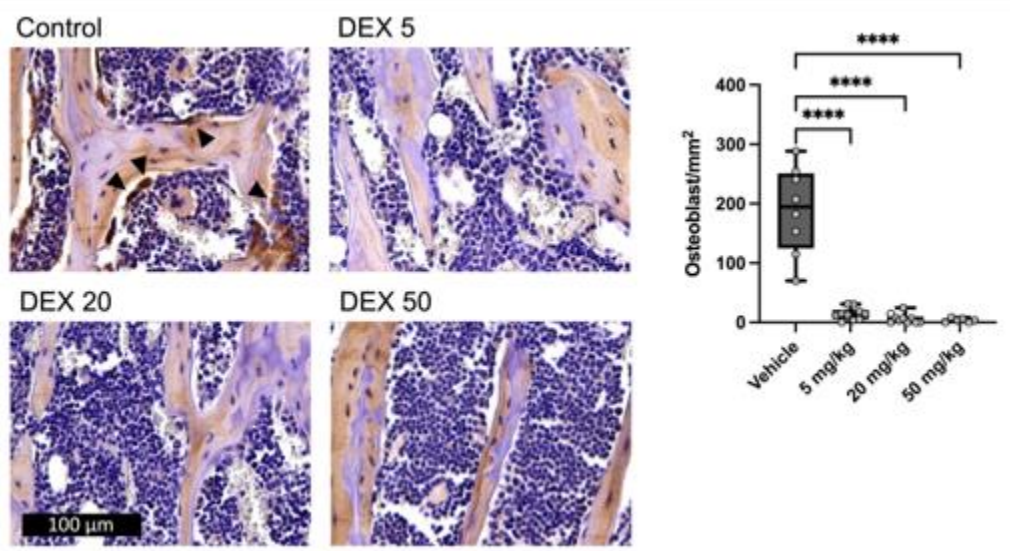

Figure S2: Staining of osteoblast cells across treatment conditions. *Osteoblast cells were stained and the number of osteoblast cells per square millimetre was found to decrease significantly upon dexamethasone treatment.*