Summer Research Fellowship Programme 2023
Final Report

Computational Analysis of Single-Cell Omics Breast Cancer Dataset

June-July, 2023

Principal Investigator: Dr. Kumardeep Chaudhary, Senior Scientist
CSIR – IGIB, New Delhi

Submitted by: Sunaina (LFS1781)
National Institute of Science Education and Research, Bhubaneswar

Abstract

Breast cancer is a complex and heterogeneous disease comprising of varied cellular subtypes that play a role in the manifestation of the disease. Every cell undergoes differential regulation mechanisms to express different cellular phenotypes such as cell markers or proteins. single-cell RNA sequencing has emerged as a powerful tool to understand this cellular heterogeneity. In this report, the computational analysis of single-cell omics between primary tumors and metastatic tumors of breast cancer patients has been performed. We visualise the apparent cellular heterogeneity of the tissue samples through state-of-the-art dimensional reduction techniques such as t-SNE and UMAP. We further map out the clusters using clustering algorithms on the reduced dimensional representation and annotate these clusters through reference datasets. We further study the cellular diversity through state-of-the-art statistical methods using over-representation analysis, gene set enrichment analysis and compositional studies. We found out that there were important cell types that varied in both conditions such as endothelial cells and epithelial cells. We also found epithelial-mesenchymal transition and N-F- Kappa B signalling pathway to be differentially expressed in both the conditions. Overall, the computational analysis of single-cell RNA sequencing has been extensively discussed and used to understand the molecular and cellular complexity of breast cancer. This technology may contribute to an increased understanding of the intricate mechanisms that take place for a cancer to be more invasive.

Acknowledgement

I would like to thank my guide, Dr. Kumardeep Chaudhary, for the guidance, mentorship, and support throughout this project. The completion of the project would not have been possible without their help and insights.

I would also like to express my sincere appreciation to my lab mates, specially to Ms. Akanksha who provided constructive criticism at every step of this project. She was extremely patient with all my queries. In addition, I'd like to thank: Ishita, Aisha, Puru, Vignesh, Balendu and Satyartha. I am grateful for the stimulating discussions, brainstorming sessions and the regular tea walks that we went on. Their diverse perspectives and expertise have broadened my horizons and contributed significantly to my growth as a researcher.

I would also like to thank IAS-INSA-NASI for providing me with this incredible opportunity to take part in the summer fellowship programme.

Lastly, I would like to express my deepest appreciation to CSIR-IGIB, New Delhi for providing me a comfortable stay for the last two months.

Regards,
Sunaina

Contents

Figure Index:

1: Introduction

Single-cell omics – in particular, whole-genome and transcriptome sequencing of single cells, was chosen as Nature Method of the Year 2013('Method of the Year 2013', 2013). After six years in 2019, single-cell multimodal omics – that is, integrating other modalities such as protein information with single cell data was chosen as Nature Method of the Year 2019 ('Method of the Year 2019: Single-Cell Multimodal Omics', 2020). The advent in technology has allowed us to progress from organismal level to cellular level at a very fast pace. Starting from genome of an organism, through bulk genome sequencing and bulk RNA sequencing, we can measure the gene expression profiles of different genes in a tissue. In addition, one can also focus on the structural and functional variants of a gene using tools such as AnnoVAR and ClinVAR. It is known that a tissue is not any homogenous mixture of cells; the molecular signatures of the cells and its compositional nature gets washed out in bulk sequencing. To preserve this heterogeneity, there have been substantial development in the techniques for resolving the genomic signature of cells. Mainly, there have been two types of methods for delineating the expression profile of cells – plate-based and droplet-based methods. Plate-based methods are inspired from the microarray-based probe methods where cells are separated into microwells. On the other hand, droplet-based methods are microfluidic-based methods that trap the cells inside hydrogels to uniquely append a cell with a molecular barcode. These methods have different cell coverage with droplet-based methods performing a lot better as shown in Fig 1.1. Plate-based methods like SMART-seq2 and MARS-seq perform full-length sequencing of the transcripts whereas droplet-based methods perform either 3' or 5' end sequencing where the tags are attached to 3' or 5' end. This marks an important difference between the two methods: although plate-based methods have a lower number of cell coverage, they have a higher recovered number of genes per cell. Thus, if one needs to study a single cell type on a deeper level, plate-based methods are preferred whereas if one needs to study cell heterogeneity, droplet-based methods are preferred.

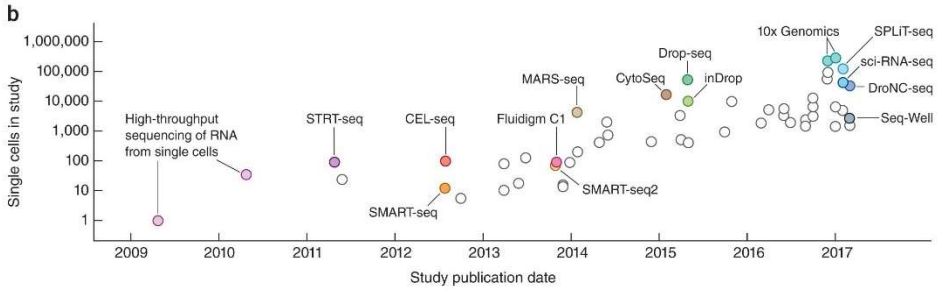


Fig. 1.1: Figure showing the cellular coverage using different methods. Here, 10x Genomics and Drop-seq is one of the droplet-based methods whereas MARS-seq and SMART-seq2 are plate-based methods and have a lower cell coverage (Taken from Svensson et al., 2018).

In this report, the computational analysis of single-cell RNA seq (scRNA-seq) data from 10x Genomics has been discussed to explore the cellular heterogeneity between primary tumors and metastatic tumors of breast cancer patients using scRNA-seq data. This is similar to bulk RNA seq in some respects but has additional steps to dissociate cells and deal with amplification bias. The experimental setup consists of microfluidic-based formation of droplets that ideally consist of a single cell and a molecular bead that is used to barcode the mRNA of the cell as shown in Fig 1.2 (A). After the tissue is dissociated into single cells, they are captured through microfluidics with barcoded beads. These barcoded beads consist of three molecular signatures: PCR handles that consist of sequencing adapters and primers, a cell barcode that is usually the size of a 14 bp sequence drawn from 7,50,000 unique barcode sequences, a 10 bp "unique molecular index" to index the mRNA molecules and a 30 bp oligo-dT to target poly-adenylated mRNA transcripts such that other types of RNA such as rRNA are excluded. When the cells are lysed, RNA from

the cytoplasm, nucleus and even mitochondria get ejected out into the droplet emulsion. The mRNAs consisting of poly-A tails prime with the oligo-dTs present on the bead. The droplets are further broken. The mRNAs are reverse transcribed to cDNA and the cDNA is amplified through PCR. This amplification process is not completely unbiased as cDNAs with a higher %GC content may get amplified poorly or the transcripts with a lower read count may not get amplified.(Zheng et al., 2017) The UMIs or the unique molecular index help in this regard as they lower this bias since we know the original count through the UMIs. If two transcripts bind with a UMI, then two reads are attributed to the gene that is aligned to the transcript.

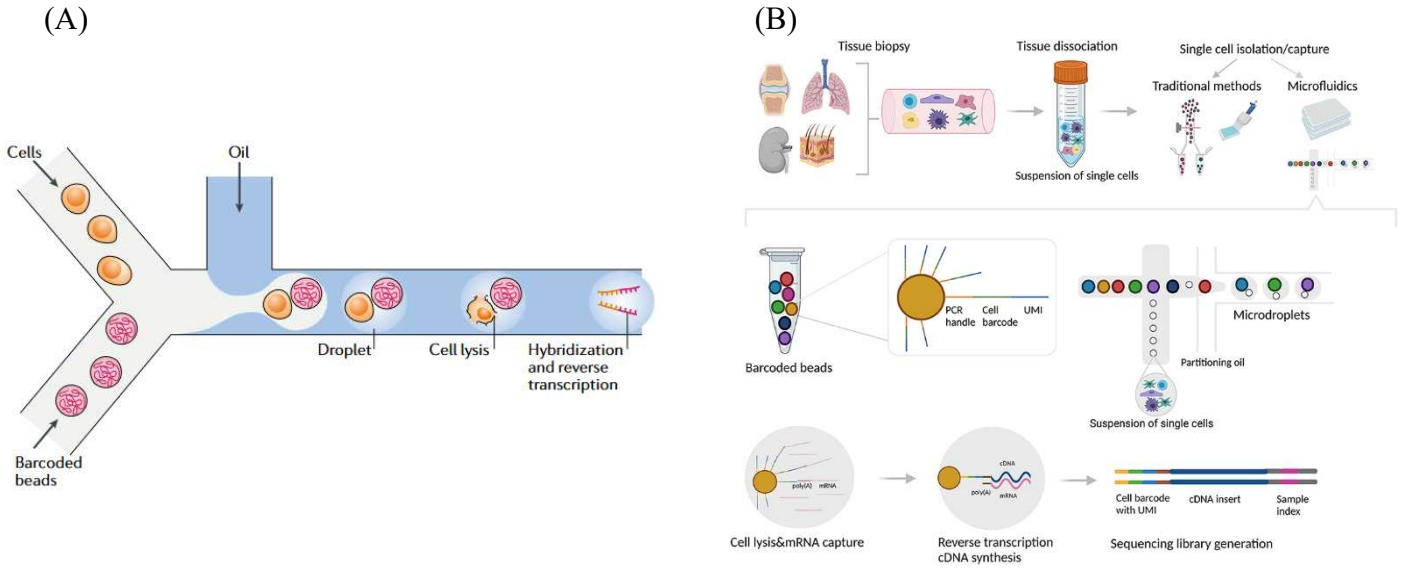


Fig. 1.2: Figure showing the experimental setup for single-cell RNA sequencing: A) Microfluidics system for production of oil droplets containing cells and barcoded beads. (Taken from Potter, 2018) B) Experimental workflow for single-cell RNA sequencing. (Taken from Kuret et al., 2022)

After amplification, the cDNA is fragmented into shorter reads and a sequencing library is generated as shown in Fig. 1.2 (B). For chromium-based 10X Genomics scRNA-seq workflow, these reads are present as R1 and R2: R1 corresponds to cellular barcodes used and R2 corresponds to mRNA transcripts ascribed to biological relevance. Using the cellRanger pipeline provided by 10X Genomics, these reads are aligned to a reference genome. Following these steps, counts matrix is generated and further processed through downstream analysis. Due to the high dimensional nature of the data, various dimensional-reduction visualisation methods such as Principal Component Analysis, t-SNE and UMAP are also used to get a glimpse of the data. Further, community detection algorithms are used for the generation of clusters from a low-dimensional embedding of the data. These clusters are then annotated using machine learning algorithms using reference data. Depending on the target of the analysis, one may delve into complex differential gene analysis and lineage studies such as calculating the RNA velocity and trajectory inference.(Nayak & Hasija, 2021) After identifying the cellular identities of the cells, we can proceed with the downstream analysis which may involve differential gene analysis, enrichment analysis, compositional analysis, cell to cell communication, gene regulatory pathways, perturbation analysis and trajectory inference. The downstream analysis is highly dependent on the experimental objective.

1. **Differential expression analysis**: Differential expression analysis identifies genes that are significantly upregulated or downregulated between different conditions. This is performed through statistical tests such as t-test, Wilcoxon Rank Sum test or a modified t-test that

overestimates the variance since in scRNA-seq studies, there is usually a lower sample size as the experimental setup is expensive.

2. **Cell trajectory analysis**: Cell trajectory analysis identifies the developmental state of various cells through gene expression. Thus, we can mark the trajectory of development using scRNA-seq data since at any given point of time, cells belonging to one cell type will be in different states. The gene expression counts of these "sub-types" can be plotted against a pseudotime axis which helps us to see the
3. **Cell-cell communication analysis**: scRNA-seq data can also provide information on cell-to-cell communication networks. One such example is of Cell chat that performs this by analysing the gene expression profiles of overexpressed genes and then using complex statistical models, it calculates the probability of two cell types communicating with each other using the law of mass action.
4. **Gene regulatory network analysis**: Gene regulatory network analysis identifies the gene regulatory relationships among various cell identities. Through this, we can focus on the transcription factors that regulate the regulation of various genes in different cells.
5. **Functional enrichment analysis**: Functional enrichment analysis assesses the enrichment of gene sets or gene ontology terms within a given set of genes. It helps to infer the biological functions and pathways associated with different cell types.
6. **Cell compositional analysis:** Compositions of different cell types can change between two conditions. Thus, there exist complex models to perform statistical analysis to find differential cell proportions between two or more conditions.

In the report, differential expression analysis, functional enrichment analysis and compositional analysis have been implemented using Python (version 3.9.6) and R (version 4.3).

Objectives

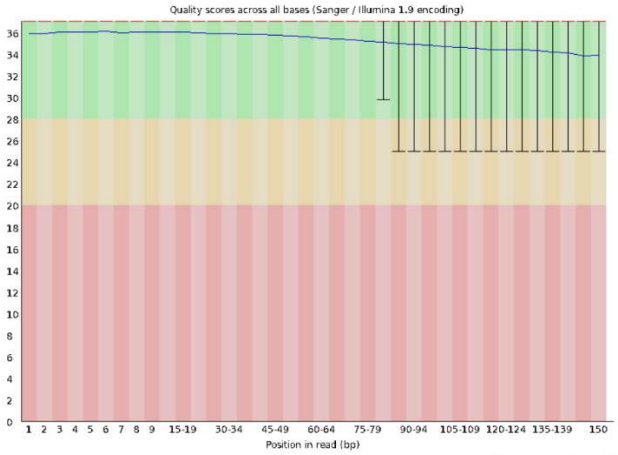
1. To construct and implement a computational pipeline for analyzing single-cell RNA sequencing dataset of breast cancer patients.
2. To evaluate the cellular heterogeneity using state-of-the art clustering methods and differential gene expression techniques.

2: Methodology

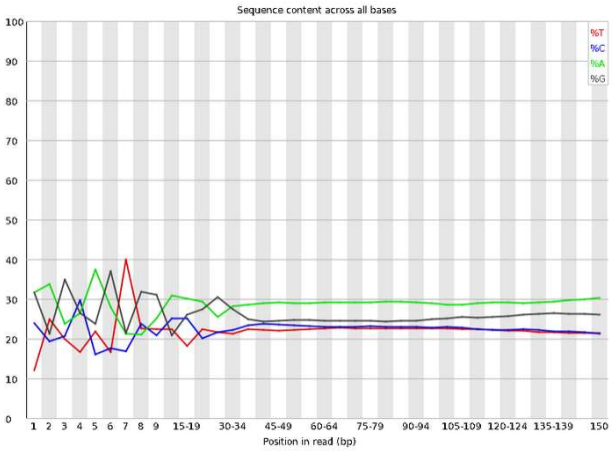
2.1 Data acquisition:

The scRNA-seq data of 37,554 primary tumor cells and 29,898 metastatic tumor cells from three breast cancer patients were downloaded from the *Sequence Read Archive* (SRA) database under the accession code GSE225600 (Liu et al., 2023) (https://www.ncbi.nlm.nih.gov/sra). This data consisted of FASTQ files containing sequenced reads and their corresponding quality scores for each read. To check the quality of the fragmented reads, FASTQC was used on Read 2 for each sample since Read 1 mostly contained artificial sequences corresponding to the gel chemistry of the beads used in chromium 10X Genomics experiment.

A)



B)



C)

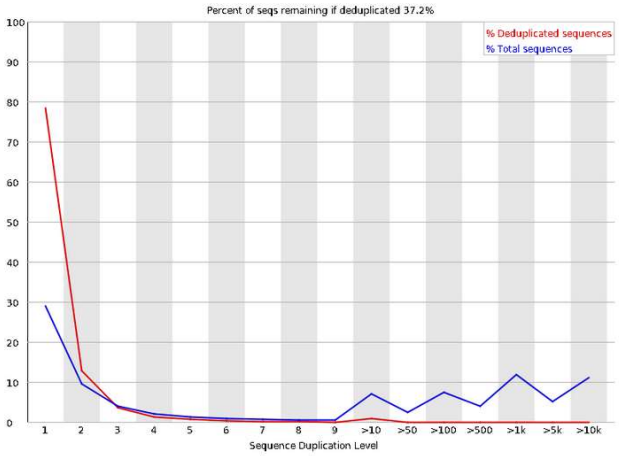


Fig. 2.1: Figure showing FASTQC metrics for one of the primary tumor samples: A) Per-base sequence quality scores; B) Per-base sequence content; C) Sequence duplication levels

FASTQC (version 0.12.0) was used to check the overall features of the raw data. In Fig 2.1, various quality tests are shown that were performed for all the samples: A) for per-base sequence quality scores, most of the scores should lie in the green zone, B) for scRNA-seq data, we may see bias in the per-base sequence content for the first 15-30 bases because even in bulk RNA-seq, the priming is not completely random

and certain random primers may be biased for priming over others. Thus, after the first few bases, we observe parallel lines for each base as expected, C) For expression level studies using scRNA-seq, there is an amplification step that amplifies the genes of each cell – thus, genes that have a higher expression may have a higher number of transcripts due to which one may observe higher sequence duplication level. But this percentage decreases when we use the pipeline provided by Cell Ranger and use FASTQC on the aligned files.

2.2 Generating the count matrices:

The count matrix is generated using 10x Genomics Cell Ranger 7.1.0 tool. The cellranger count pipeline built on STAR aligner version 2.7.2a takes in FASTQ file for a particular sample and aligns it to a reference genome. Here, we have used the GRCh38 human reference genome assembly (https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build#GRCh38_2020A). The pipeline performs both transcriptome and genome alignment. Firstly, all the reads that don't have proper barcodes are removed. Secondly, using the valid barcodes, the reads that have not mapped to the reference genome are removed. Reads with low alignment scores are also removed. Thirdly, all the intergenic reads – that is, those reads that are present between the genes are also removed. After this step, we have a subset of exonic and intronic reads out of which those reads that may be mapping to more than one gene and reads that are aligned to the opposite strand, that is, the anti-sense strand, are also removed. This is shown in Fig. 2.2. After performing the transcriptomic alignment, each UMI is subjected to quality check where a subset of UMIs with low base quality score are removed. After this, a cutoff for the UMI counts is used to distinguish between background (empty barcodes) and actual cells. Initially, there is one raw matrix which is further filtered to remove these empty barcodes to generate a filtered matrix file. This whole process is summarised in Fig. 2.3 and all these quantitative measures are reported in the Cell Ranger Web Summary report.

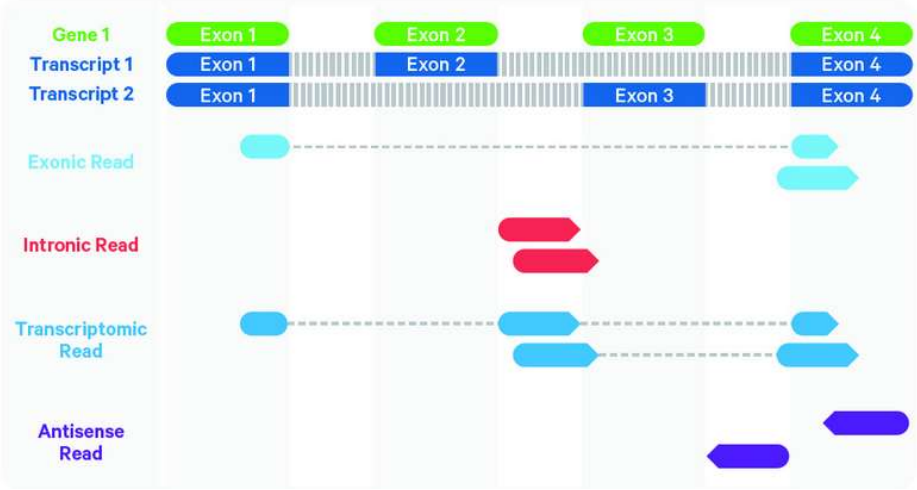


Fig. 2.2: Figure showing the different kinds of reads present in the sample: exonic reads, intronic reads, transcriptomics reads containing both exon and intron and antisense reads (Figure taken from Gene Expression Algorithms Overview -Software -Single Cell Gene Expression -Official 10x Genomics Support)

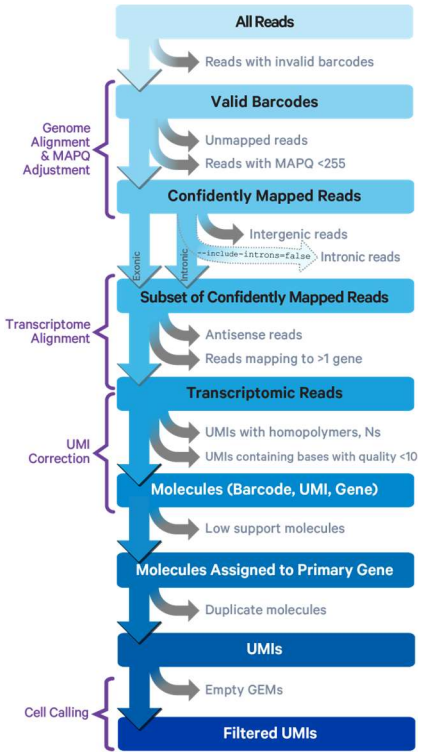


Fig. 2.3: General steps followed by the Cell Ranger count pipeline. (Figure taken from Gene Expression Algorithms Overview -Software -Single Cell Gene Expression -Official 10x Genomics Support)

The cellranger count tool outputs four important files: genome-aligned reads in BAM format, a .tsv file containing all the UMIs, a .tsv file containing all the cellular barcodes, a .mtx file containing the sparse matrix representing the total number of UMIs, total number of cells and the counts for each of them, a cloupe file and a websummary file. The cloupe file can be used with cell ranger's downstream analysis pipeline. The websummary file has many important metrics which can be used to understand the quality of the data. It contains the barcode rank plot which preferably must have a defined cliff and knee shape that separates the cells from the empty barcodes as shown in Fig. 2.4. This can be further refined in the background processing step where we remove ambient mRNA.

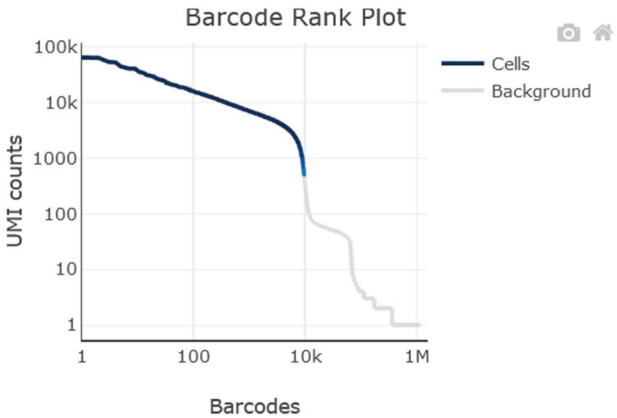


Fig. 2.4: Barcode rank plot for one of the metastatic samples

Other metrics that should be checked are(Interpreting Cell Ranger Web Summary Files for Single Cell Expression Assay):

1. Valid barcodes that are mapped to the barcode list provided by cell ranger for the specific experiment should be more than 75%. There are various barcode lists which are relevant to different chemistries and different experiments.
2. Q30 bases for both barcode and RNA reads should be more than 30%
3. Reads that have been mapped confidently to transcriptome should ideally be more than 30%
4. Median genes per cell should be more than 1000
5. Fraction reads in cells should be more than 70%

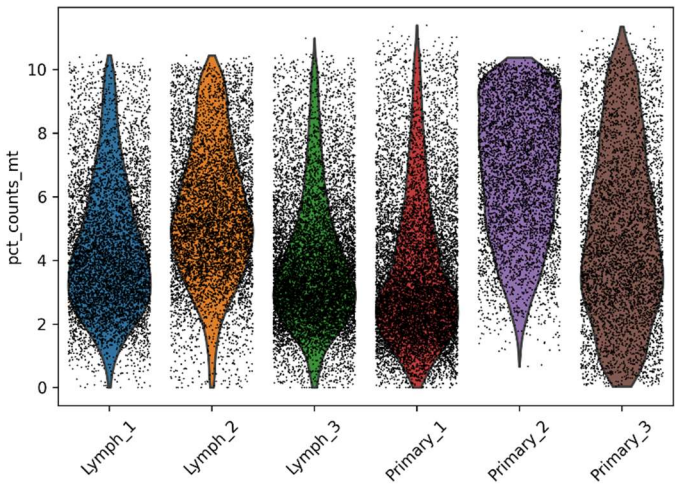
2.3 Quality control:

After the matrices has been generated, the filtered matrix is imported in matrix as an AnnData(Virshup et al., 2021, 2023) object using ScanPy(Wolf et al., 2018) in Python (version 3.9.6). At this point, there are three important ways in which quality check can be performed at the cellular level –

1. n_genes_by_counts: This represents the number of genes with positive counts in a cell
2. total_counts: This is the total number of counts for a cell,
3. pct_counts_mt: This represents the percentage of mitochondrial counts for a cell

Usually, cells that have more than 10% of mitochondrial genes and cells with lower number of total counts are discarded as these might represent dead and decaying cells. These cut-offs should be used with caution as cells that have a higher percentage of mitochondrial genes may also represent oxidative function whereas low number of counts might represent quiescent cells. These quality control metrics are applied together rather than separately. In addition to these quality checks, the percentage of ribosomal genes and haemoglobin genes are also calculated. (Luecken & Theis, 2019)

(A)



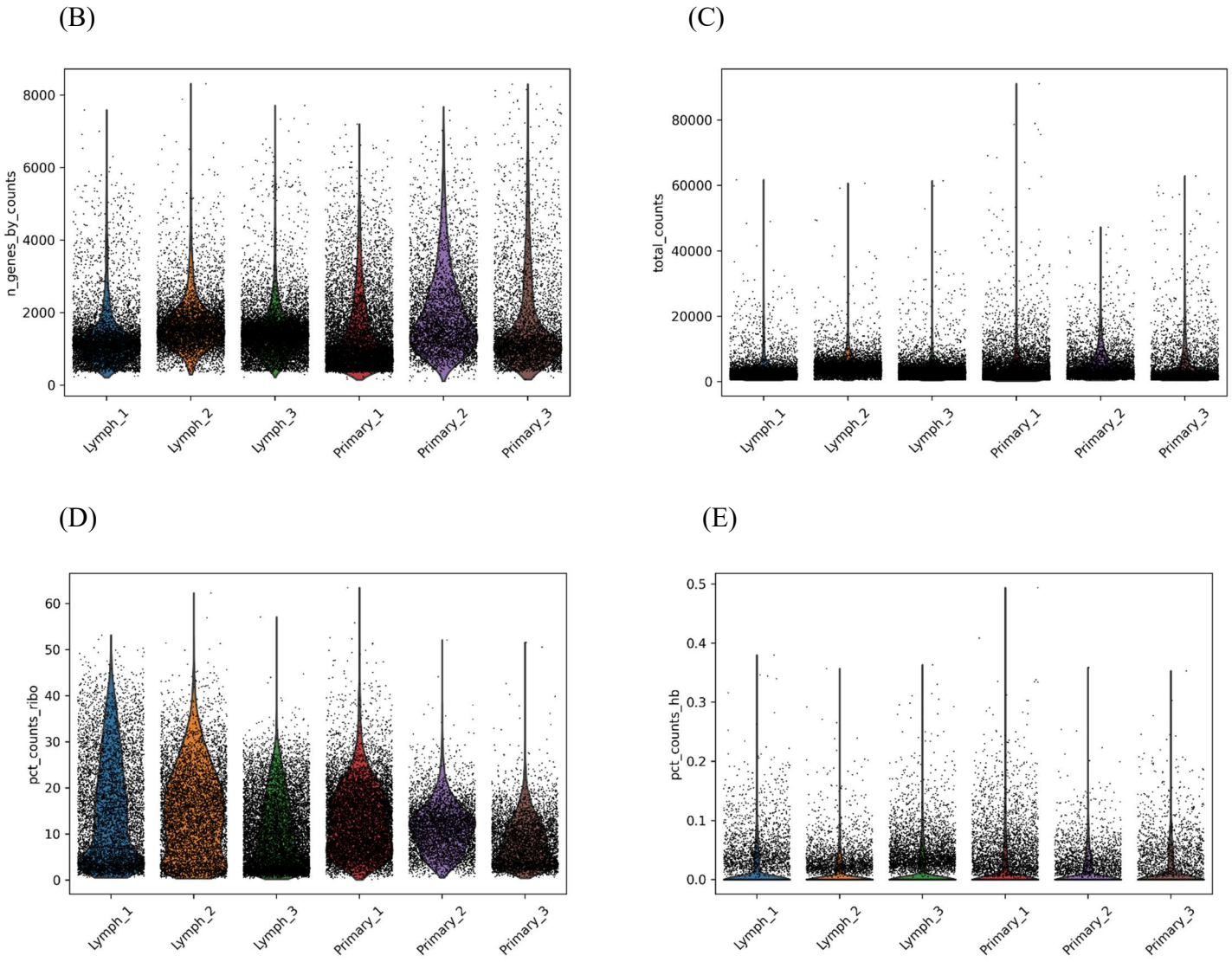


Fig. 2.5: Quality checks at the cellular level of all the samples: A) Percentage of mitochondrial genes, B) Total number of genes which have positive counts for each cell, C) Total cell counts, D) Percentage of haemoglobin genes E) Percentage of ribosomal genes in each cell for all the samples.

For each sample, 10% of mitochondrial genes was used as a cutoff. At the gene level, those genes that are not present in at least 20 cells were removed. Those cells that did not have at least 200 genes were also removed.

2.3.1: Removing ambient mRNA:

Within each droplet, ideally there is one cell which is lysed, and the mRNAs released are annealed to the oligonucleotides. But since the cells are suspended in a suspension medium, due to stress or other issues, cells may release ambient RNA as shown in Fig 2.6. Contamination rate, even in highly controlled experiments is around 1 %. Thus, there is a need to remove this ambient mRNA. Following steps are used to clean this using the SoupX package version 1.6.2 (Young & Behjati, 2020).

1. Empty droplets are used to measure the concentration of ambient mRNA using the raw matrix file generated by cellranger. In this step, the expression profile of contamination is determined using those barcodes which have a lower number of UMIs. This algorithm requires clustering.

2. A contamination fraction is estimated based on the marker genes available in each cluster. An estimate of more than 1% is considered as a bad dataset.
3. The expression of each cell is corrected using the ambient mRNA expression profile.

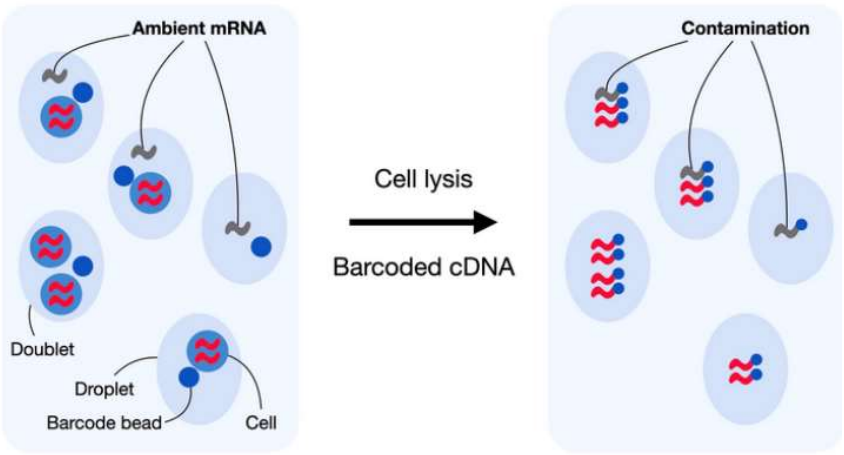


Fig. 2.6: Contamination by ambient mRNA (Taken from Young & Behjati, 2020)

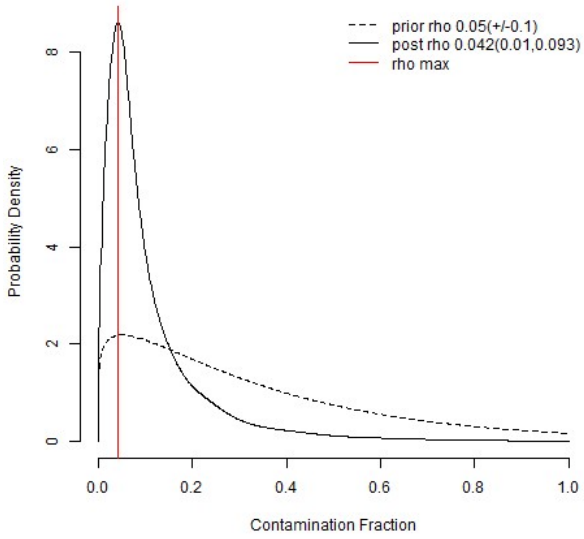


Fig. 2.7: Figure showing the estimate of contamination rate for one of the samples is 0.042 as shown by the highest peak.

2.3.2 Finding doublets:

Using scDblFinder version 1.14.0 (McGinnis et al., 2019), single droplets containing two cells with a single barcode were removed. In these cells, the mRNA counts in the matrix will be the average of the two cells. The package creates artificial doublet droplets and compares their expression profiles with all the cells. Based on the scoring scheme, we found 4-5% doublets in each sample.

2.4 Different types of normalization:

Highly expressed genes have more variance in their distribution as compared to lowly expressed genes. Due to this difference in variance between different genes among different cells or *heteroskedasticity*, there are different methods to correct it:

1. Log-shift transformation:

$$f(y) = \log(\frac{y}{s} + y_0)$$

This function is used to produce log-shift transformation of the raw count matrix. Here y represents the raw counts, y_0 is pseudocount that is added so that the function is not undefined when raw counts are zero and s is a size factor that is dependent on the library size L which is usually calculated as median gene counts.

$$s = \frac{\sum y}{L}$$

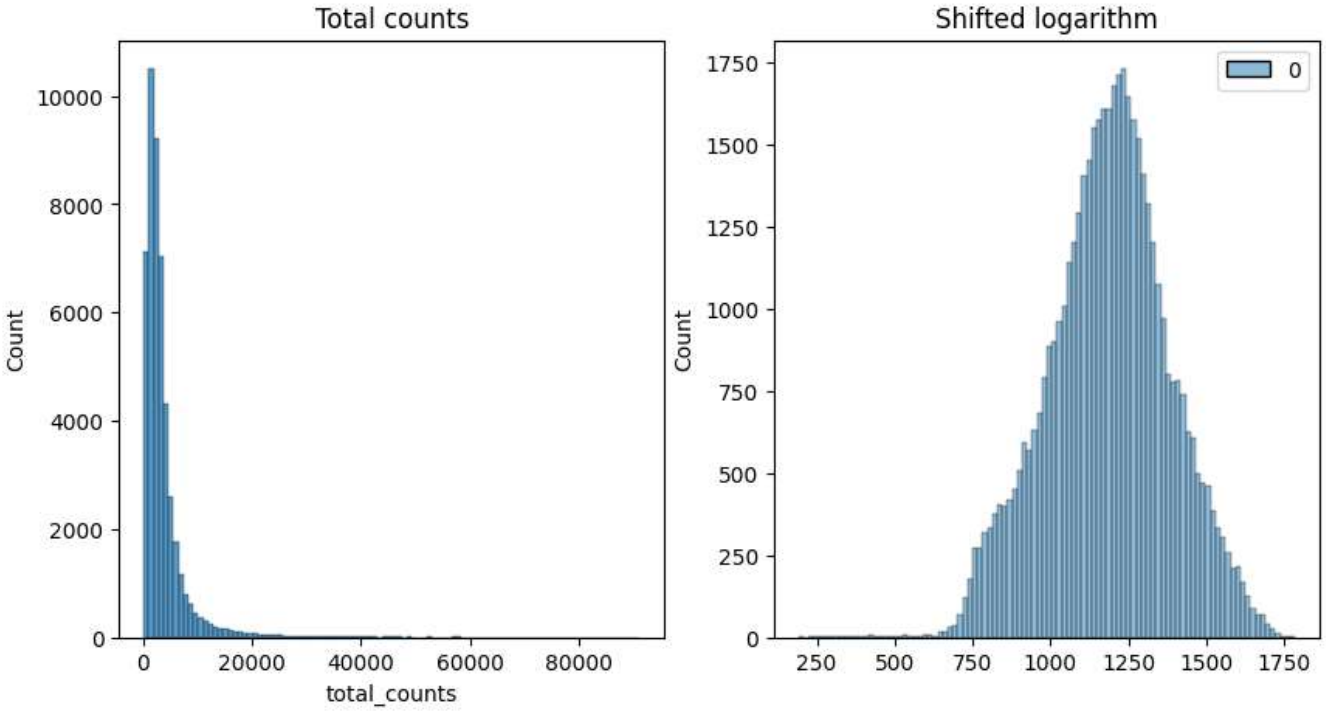


Fig. 2.8: Log-shift transformation of all the samples

2. Analytic Pearson's residuals: The raw counts are fit to a negative binomial regression.

Pearson's residuals are calculated from that to generate estimates for technical noise. Residuals refer to the difference between the actual and predicted values. Here, standard deviation of a binomial distribution is estimated. Pearson's residuals for linear regression are calculated as:

$$r_i = \frac{y_i - \hat{\mu}_i}{SE(y_i)}$$

The numerator represents the difference between actual and predicted values whereas SE represents the square root of variance for the actual values. Inspired from this, the Pearson's residuals for binomial distribution are similar but they are scaled by the standard deviation as calculated from a binomial distribution(Townes et al., 2019) where r_{ij} represents the Pearson's residuals, y_{ij} is the observed UMIs for each gene in a cell, and $\hat{\mu}$ represents the mean for a gene j:

$$r_{ij}^{(p)} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij} - \frac{1}{n_i}\hat{\mu}_{ij}^2}}$$

Comparison between different transformations showed that this method removed sampling noise. Both positive and negative values are present in the output. For positive values, the more counts are observed than expected compared to the gene's expression and vice versa. The transformation for one of the samples is shown in Fig 2.9.

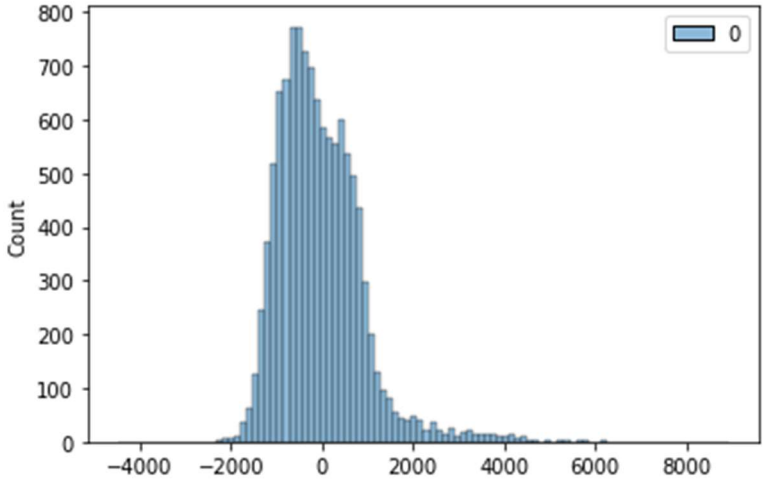


Fig. 2.9: Analytic Pearson's residuals transformation of all the samples

2.5 Feature selection:

Dispersion: Since scRNA-seq data is concentrated with dropouts (that is -- has too many zeroes since gene expression profile from each cell is calculated and a lot of genes may have very low number of transcripts), informative features (genes) need to be selected. This can be performed by calculating the coefficient of variance (CV):

$$CV = \frac{\sigma}{\mu}$$

Where σ represents standard deviation and μ represents mean for each gene across all the cells

Another metric called dispersion can also be calculated:

$$\text{Dispersion} = \frac{\sigma^2}{\mu}$$

With the default value of flavor = "Seurat", ScanPy is used to generate a list of highly variable genes by calculating dispersion for each gene and then ranking the genes. But these methods expect logarithmized data which may be biased since we are adding arbitrary pseudo-counts. This was performed while providing the batch key so that the variable genes are found not just among the various cells of one sample but for all the batches.

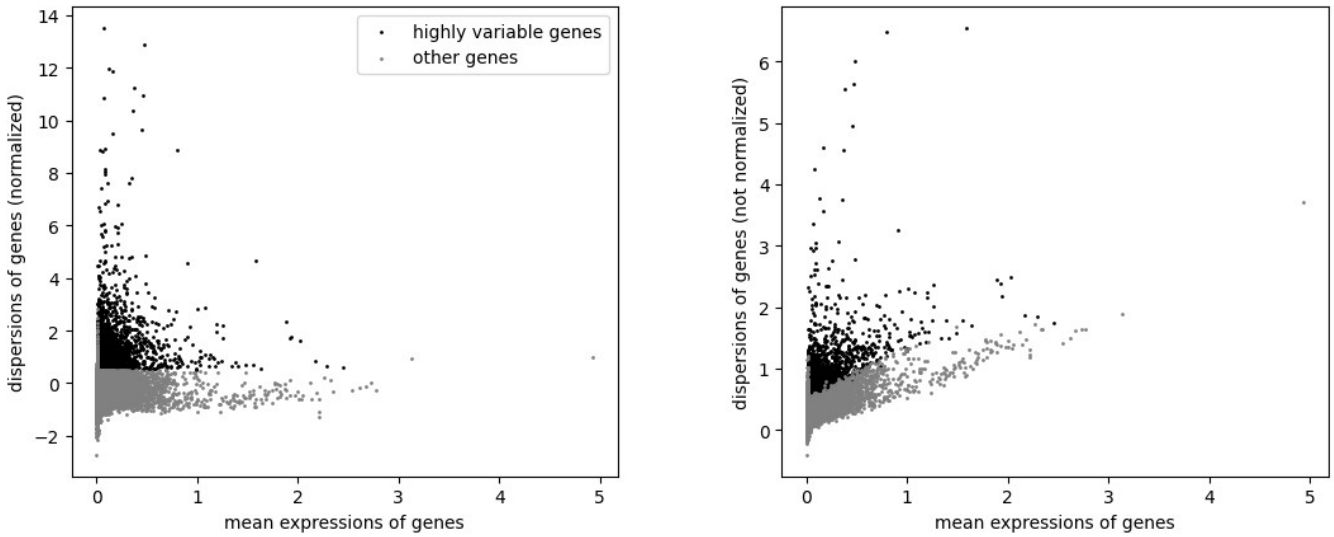


Fig. 2.10: Using dispersion, highly variable genes are ranked and shown.

Deviance: Using scry version 1.12.0 (Street K et al., 2023), a "null-model" is used to approximate the expression profiles of genes which are constant for different cells. Then, the genes having a high variability will have a higher "deviance". This is based upon negative binomial distribution modelling since the raw number of counts in scRNA-seq data is distributed around zero. Using deviance as a measure, top 4000 variable genes were selected. The residuals for deviance(Lause et al., 2021; Townes et al., 2019) are similar to Pearson's residuals:

$$r_{ij}^{(d)} = \text{sign}(y_{ij} - \hat{\mu}_{ij})\sqrt{2y_{ij}\log\frac{y_{ij}}{\hat{\mu}_{ij}} + 2(n_i - y_{ij})\log\frac{n_i - y_{ij}}{n_i - \hat{\mu}_{ij}}}$$

Where r_{ij} represents the deviance residuals for gene j in cell i, y_{ij} represents the observed UMI counts for cell i, $\hat{\mu}$ represents the mean for each gene j and n_i is the total UMIs in the sample.

2.6 Dimensionality reduction:

Various dimensionaly reduction methods are incorporated since they are highly useful to visualise the dataset.

1. PCA: For scRNA-seq analysis, PCA is not used for dimensionality reduction but for the selection of principal components that are used for downstream analysis. It is performed done after log-shift normalization. Since the raw counts have dropouts, the raw count PCA is not meaningful and does not show clusters based on cell types as shown in Fig 2.11. This is also attributed to the non-linear nature of scRNA-seq data.

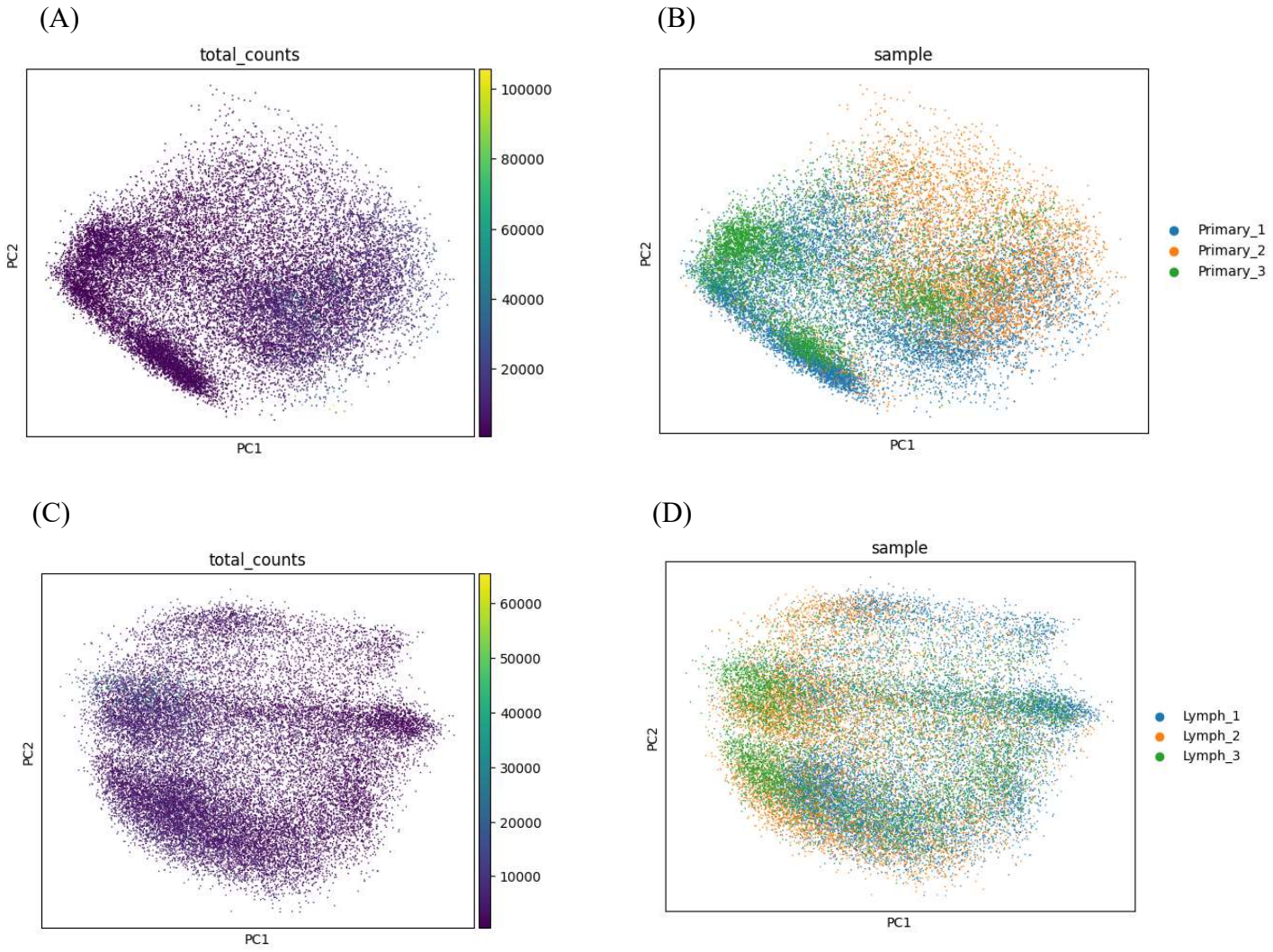


Fig. 2.11: Figures showing the PCA plots for all samples: (A) PCA plot coloured by sample (B) PCA plot coloured by total counts per cell of primary tumor cells (C) PCA plot coloured by total counts (D) and sample of lymph tumor cells

3. t-SNE: t-SNE refers to t-distributed stochastic neighbor embedding. t-SNE (Van Der Maaten & Hinton, 2008) is a graph based, non-linear dimensionality reduction technique which projects the high dimensional data onto 2D or 3D components. The method defines a Gaussian probability distribution based on the high-dimensional Euclidean distances between data points. Subsequently, a Student's t-distribution is used to recreate the probability distribution in a low dimensional space. But t-SNE does not take the inter-cluster information into account. The t-SNE graphs coloured by sample and total counts are shown in Fig. 2.12.

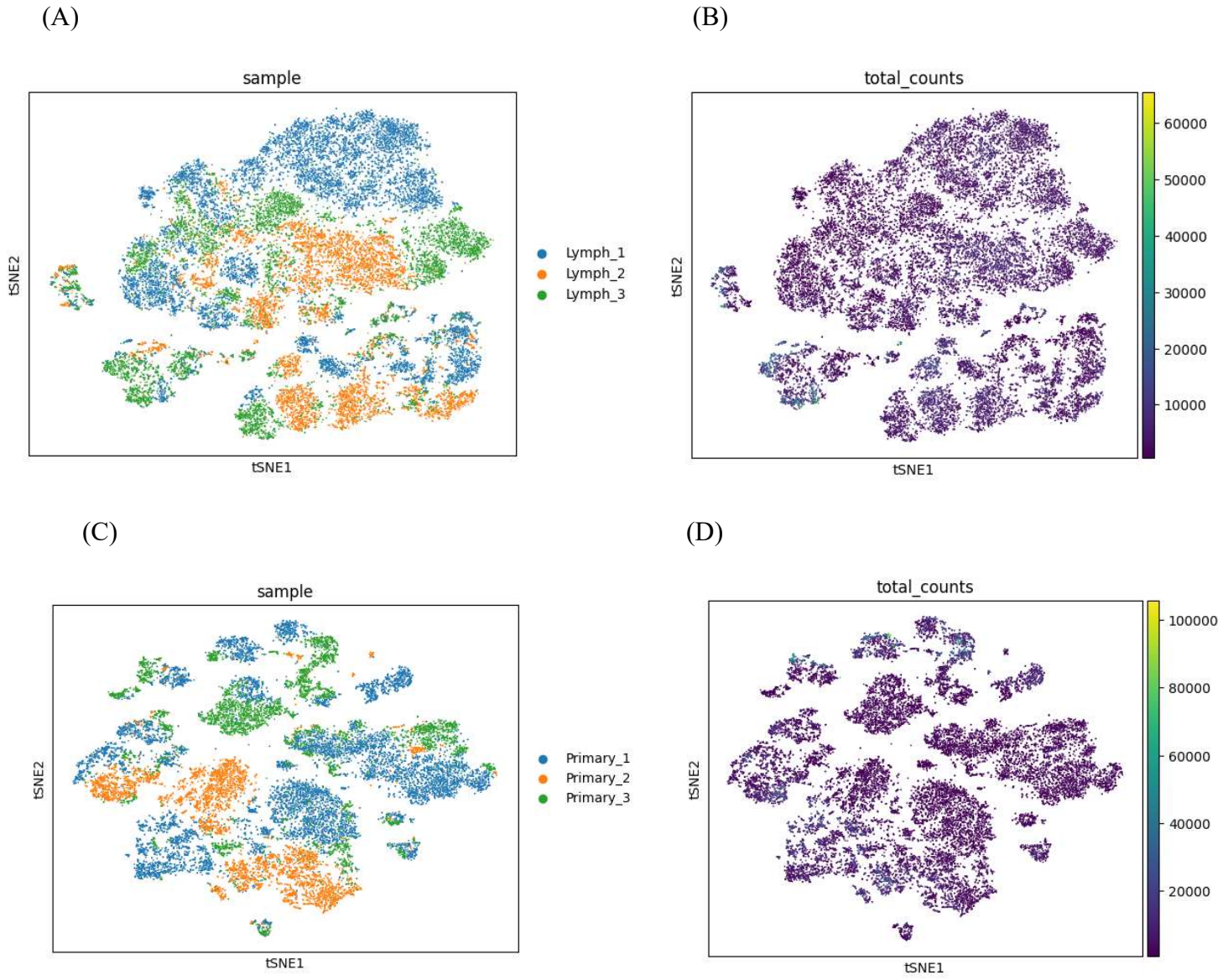


Fig. 2.12: Figure showing t-SNE plots of all samples: A) t-SNE plot coloured by sample and B) total counts for lymph tumor cells (C)t-SNE plot coloured by sample and (D) and total counts for primary tumor cells

3. UMAP: UMAP refers to Uniform Manifold Approximation and Projection. This is also a graph-based non-linear dimensionality reduction technique similar to t-SNE but it preserves the global structure of the clusters.(Becht et al., 2018) The t-SNE graphs coloured by sample and total counts are shown in Fig. 2.13.

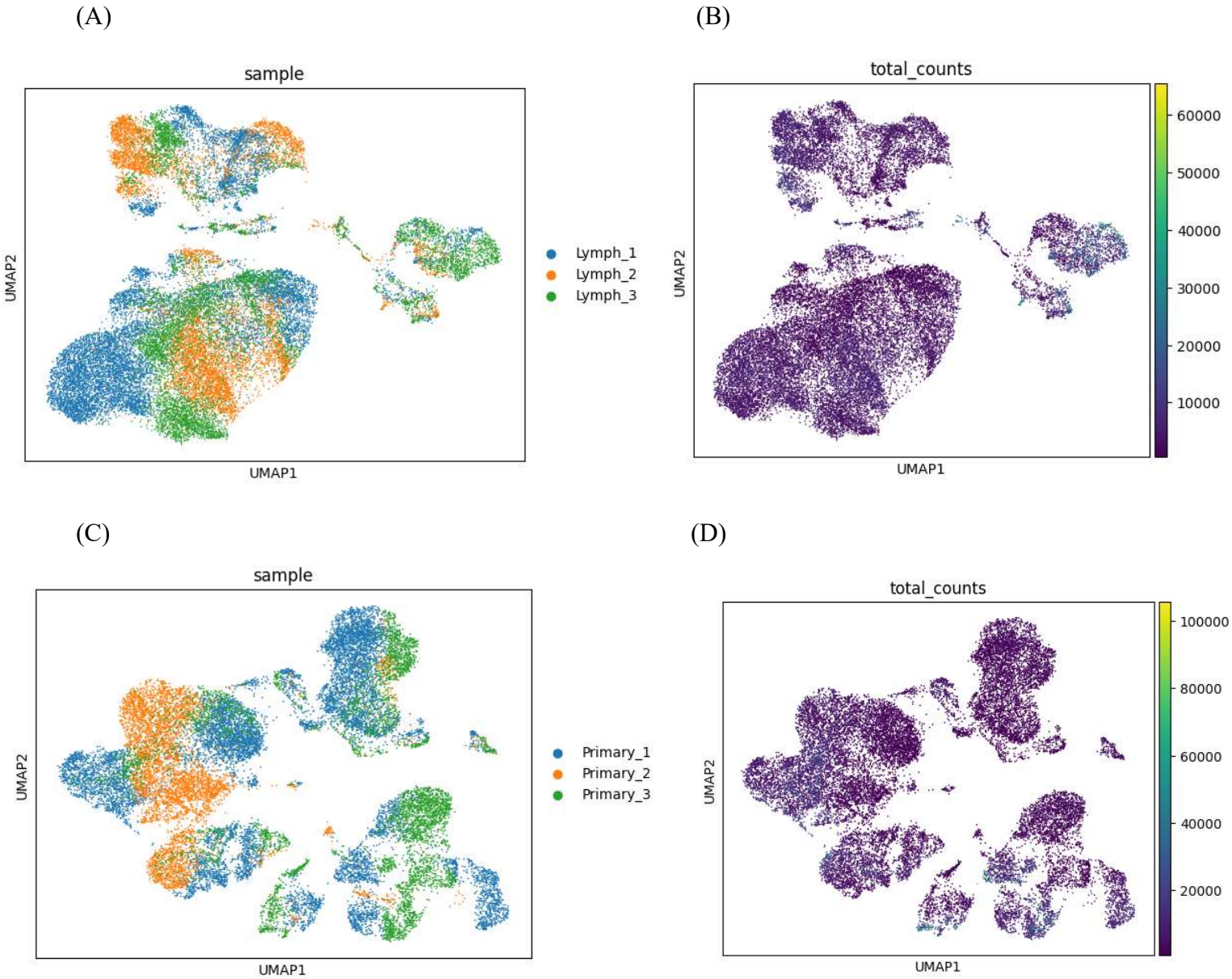


Fig. 2.13: Figure showing UMAP plots of lymph tumor cells coloured by A) sample and B) total counts of each cell and UMAP plot of primary tumor cells coloured by C) sample and D) total counts of each cell

2.7 Batch-correction:

scRNA-seq dataset is highly prone to batch correction since the heterogeneity of the samples is very high due to various factors such as: circadian rhythms, transcriptional bursting and low capture efficiency(Lähnemann et al., 2020). In addition, the samples can be processed at different times. Due to these constraints, there are different methods for batch correction. Before performing it, the raw data is visualised through a dimensional reduction process. Each sample is assumed to be a single batch. In this report, three tools for batch integration have been explored and shown in Fig 2.14 for primary tumor samples. (Tran et al., 2020):

1. COMBAT: This is a popular batch-integration method used for bulk RNA-seq data. It assumes linear combination of the genes for each sample. Due to non-linear nature, we don't get a good result for the sample integration. (Johnson et al., 2007)
2. Batch-balanced k-Nearest Neighbors (BBKNN): This is a graph-based method which forces connections among different batches. (Polański et al., 2020)

3. Mutual Nearest Neighbors (MNN): This is a linear embedding model that only considers a subset of the interactions to be closely related with other batches. (Haghverdi et al., 2018)

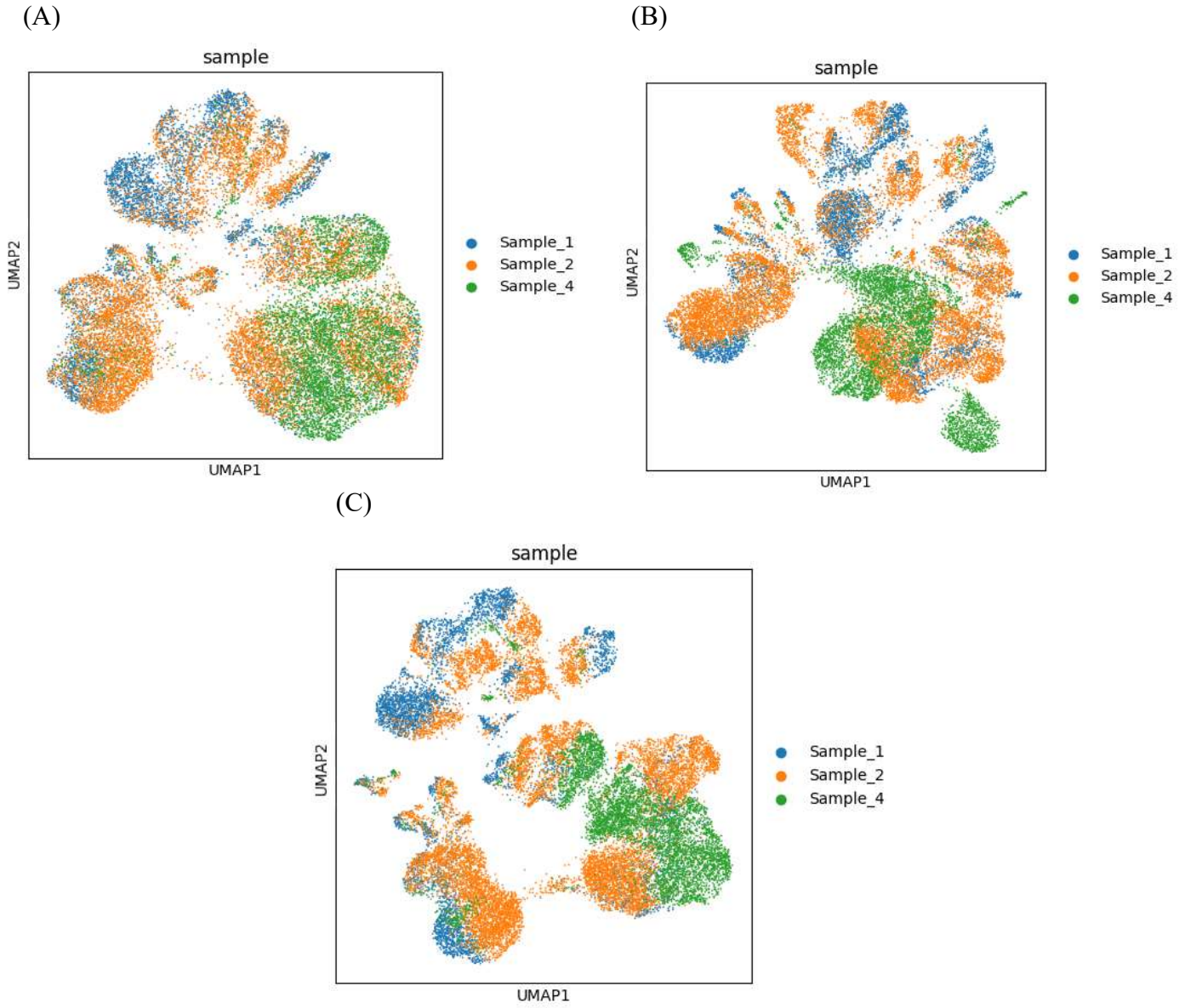


Fig. 2.14: Figure showing batch correction of the primary tumor samples using A) BBKNN, B) COMBAT and C) MNN from left to right.

Similarly, the same process was followed for the metastatic samples. Since primary tumor samples were processed using BBKNN, same procedure was followed for metastatic tumors as shown in Fig. 2.15.

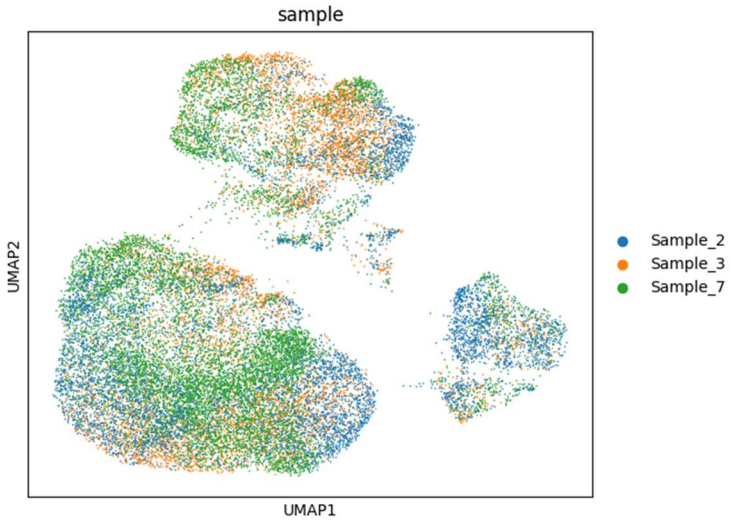


Fig. 2.15: Figure showing batch correction of the metastatic samples using BBKNN

3: Results and discussion:

3.1 Clustering cellular identities:

From a low-dimensional embedding, we further delineate the number of clusters using community detection algorithms. There are mainly two algorithms used for it: Louvain and Leiden. Since Louvain has been known to generate disconnected clusters, we have used Leiden clustering.(Traag et al., 2019) Even though we can see certain clusters in the pre-processing steps of our dataset, they have not been distinctly separated into actual clusters. Rather than relying on intuition or visual maps, since we are dealing with around 4000 genes for each cell, there are community-detection algorithms for the evaluation of clusters. One of the popular approaches is using the Leiden algorithm. A high-level depiction of the algorithm is provided:

1. Each node (cell) starts as its own cluster.
2. These nodes go through a local moving process where nodes move to optimise a "quality function" such as modularity of the network to measure the connectedness of the various nodes.
3. Then the partitions formed through these nodes produce a cluster-map which further undergoes a refining process where nodes which increase the modularity of any community (any community that is randomly chosen that increases the modularity function) are joined with those communities
4. Further the different nodes are aggregated under each of the clusters while still maintaining an optimum level of modularity.

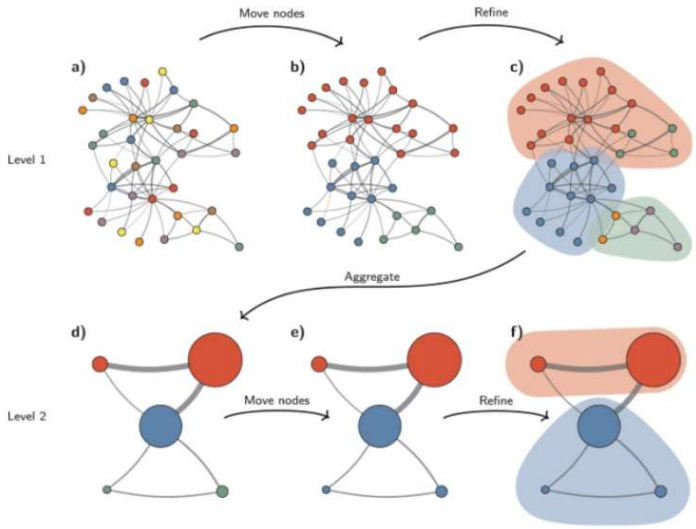


Fig. 3.1: Leiden clustering algorithm(Traag et al., 2019)

It is possible to choose different resolutions to increase or decrease the number of clusters as shown in Fig. 3.2 for lymph tumor cells and Fig 3.3 for primary tumor cells.

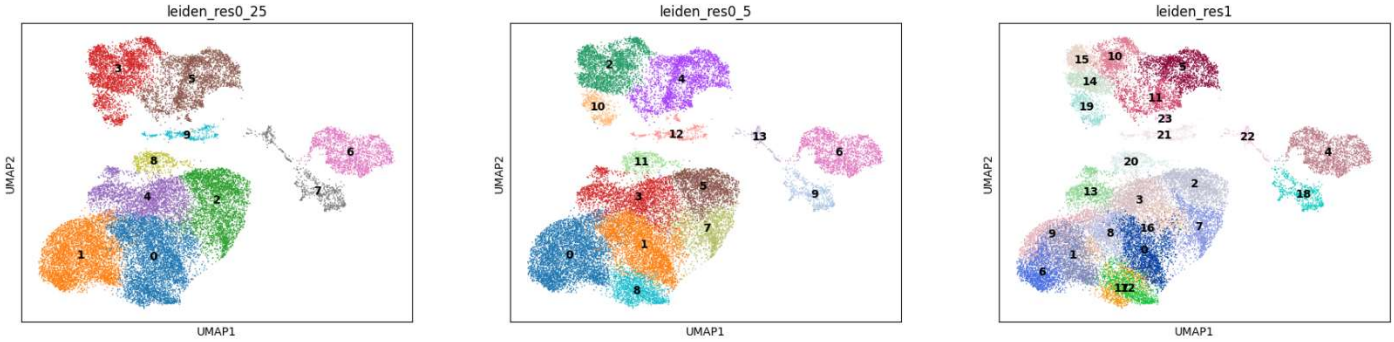


Fig. 3.2: Figure showing Leiden clustering of lymph tumor cells at different resolutions of 0.25, 0.5 and 1 from left to right.

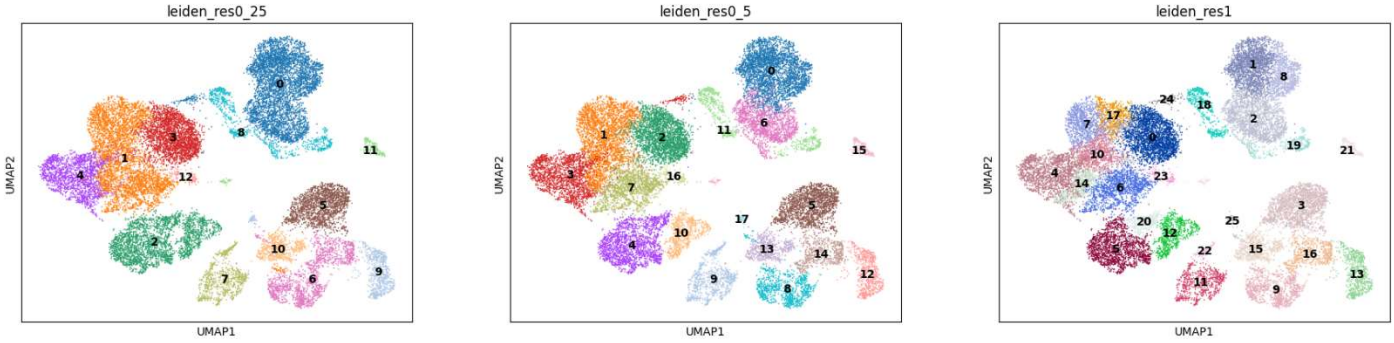


Fig. 3.3: Figure showing Leiden clustering of primary tumor cells at different resolutions of 0.25, 0.5 and 1 from left to right.

3.2 Annotating the clusters:

After the clusters have been formed, there is a need to annotate the clusters. One can either manually annotate the clusters using specific gene markers for each cell or automatically annotate them through machine learning. Here, CellTypist(Domínguez Conde et al., 2022) has been used to annotate clusters which is trained on previously annotated data using logistic regression. It has different models that are already incorporated in it. We use "Immune_All_Low.pkl" and "Immune_All_High.pkl" which are immune populations combined from 20 tissues of 18 studies. The former has a greater number of cellular subtypes. Fig 3.4 shows the results of the annotations. A dendrogram can be used as a quality check on the kind of cellular annotations that we observe. For instance, in Fig 3.5, we see that most of the T-cells subtypes have clustered together as expected since they are similar to each other.

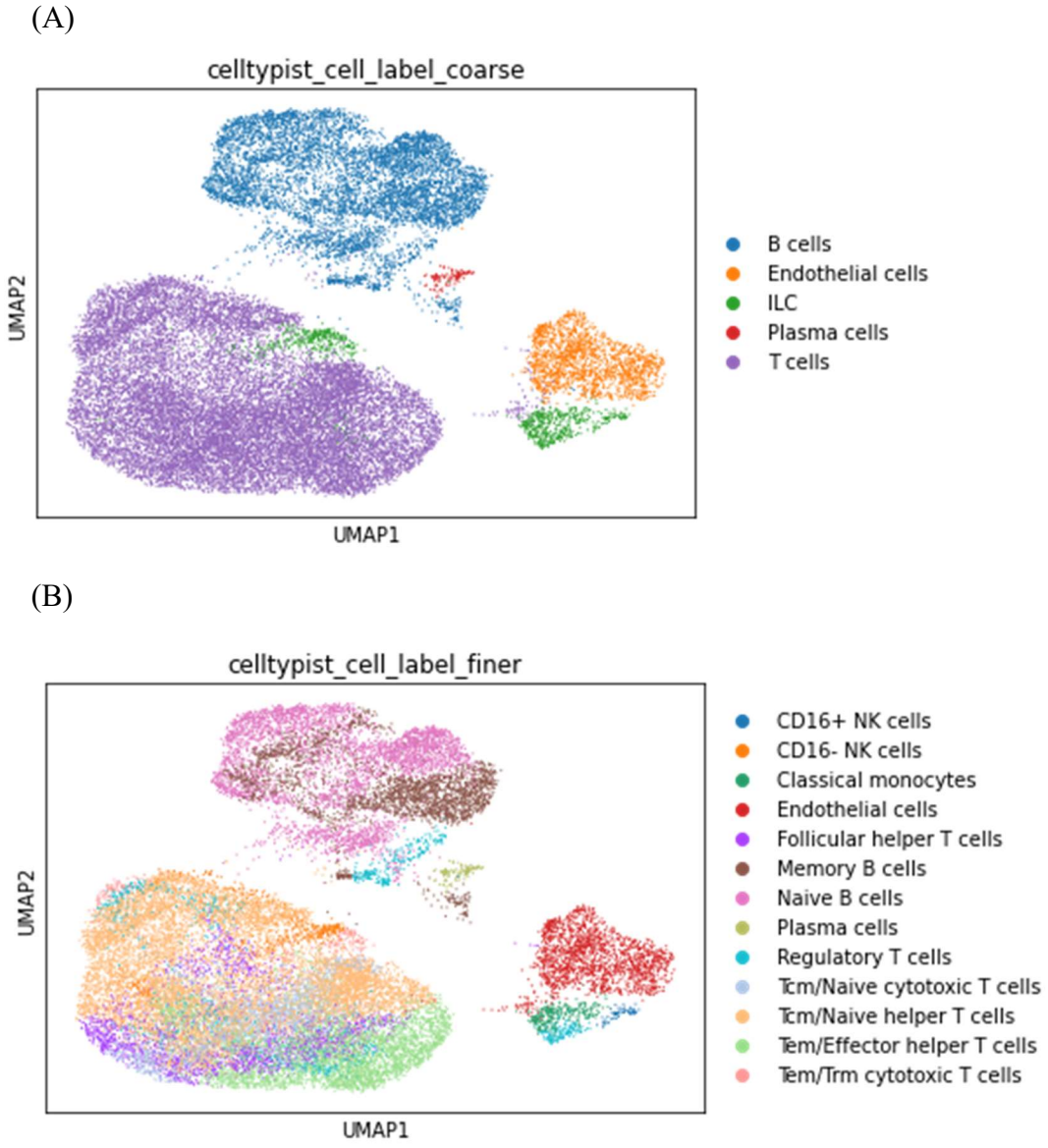


Fig. 3.4: Cell type annotations using (A) Immune_Low (coarse) and (B) Immune_High (finer) model.

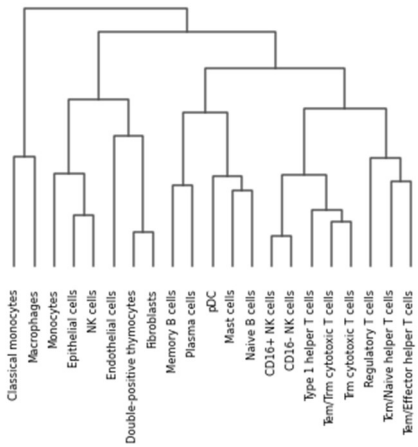


Fig. 3.5: Dendrogram of the cell type annotations

3.3: Differential gene analysis and gene set enrichment analysis:

In Fig. 3.6, we can observe the differences in gene expression profiles of metastatic and primary tumors in breast cancer patients. Wilcoxon's rank sum test, a non-parametric t-test, can be used to statistically get a list of genes that are significantly different in both the conditions. Various cell types have a differential expression in primary tumor as compared to metastatic tumor as seen in Fig. 3.7. Clear segregation between primary and metastatic tumors can also be visualised as a dotplot as seen in Fig 3.8. After finding the differential genes between the two groups, overrepresentation analysis using Enrichr(Chen et al., 2013) is performed on the list of genes using "MSigDB_Hallmark_2020" as the gene set using hypergeometric distribution.(Liberzon et al., 2011, 2015; Subramanian et al., 2005)

$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

Here, the probability of k genes being associated with a gene set for all genes in the gene list (n) with an a-priori set of all human genes available in the human genome (N) and another a-priori set of all the genes that have been associated with the gene set in literature (K). Pathways such as TNF-alpha (Tumor Necrosis Factor) signalling, apoptosis and epithelial-mesenchymal(EM) transition were overrepresented. Of particular importance is EM transition which is the main driver behind metastasis through which epithelial cells transform into mesenchymal cells.(Ribatti et al., 2020) For these pathways, functional scoring was performed and enrichment scores were calculated using GSEApy(version 1.0.5):

If the gene is not present in the gene set given, the running sum statistic is given as:

$$X_i = -\sqrt{\frac{N_S}{N - N_S}}$$

If the gene is present in the gene set, then the running sum statistic is:

$$X_i = \sqrt{\frac{N - N_s}{N_S}}$$

Here N represents the total number of genes, N_S is the number of genes in the given gene set. This running sum statistic is summed up over for all the genes to give the enrichment score.

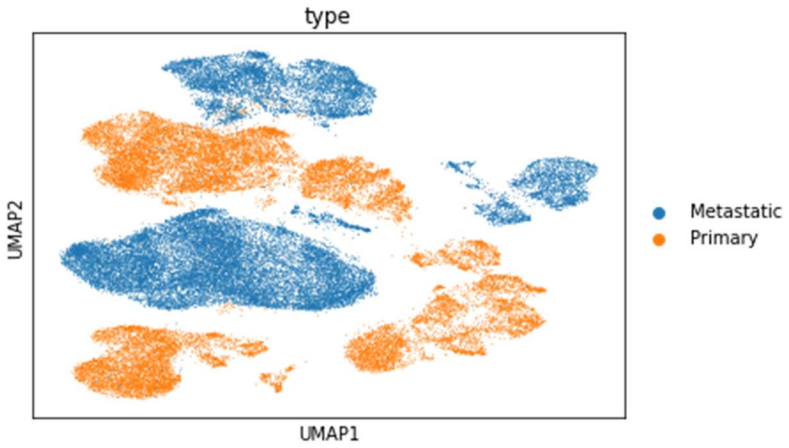


Fig. 3.6: UMAP plot showing the genetic heterogeneity in metastatic and primary tumors

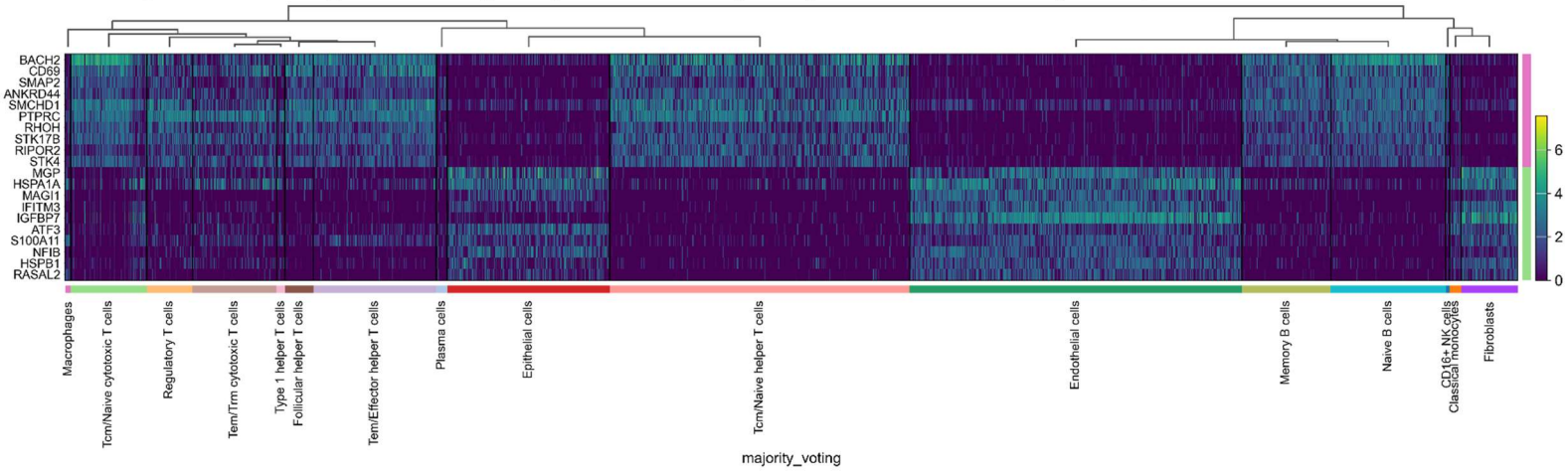


Fig. 3.7: Heatmap showing the differential expression analysis between the primary (upper) and metastatic (lower) tumors using Wilcoxon Rank Sum test

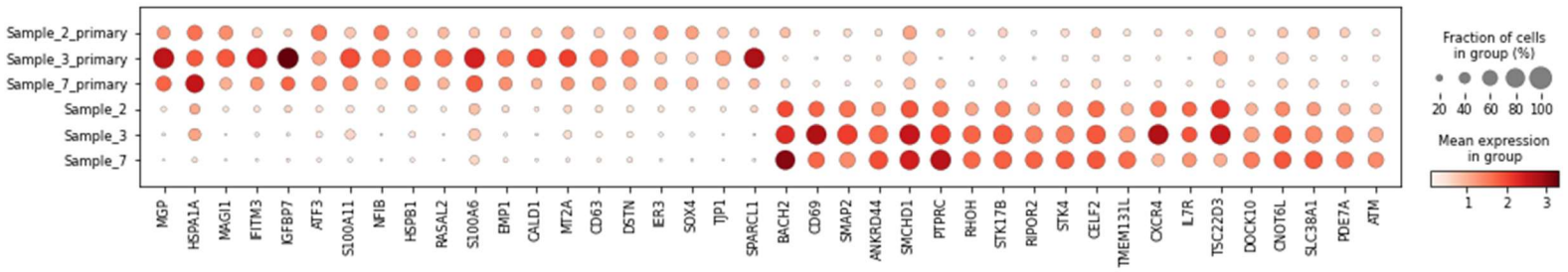
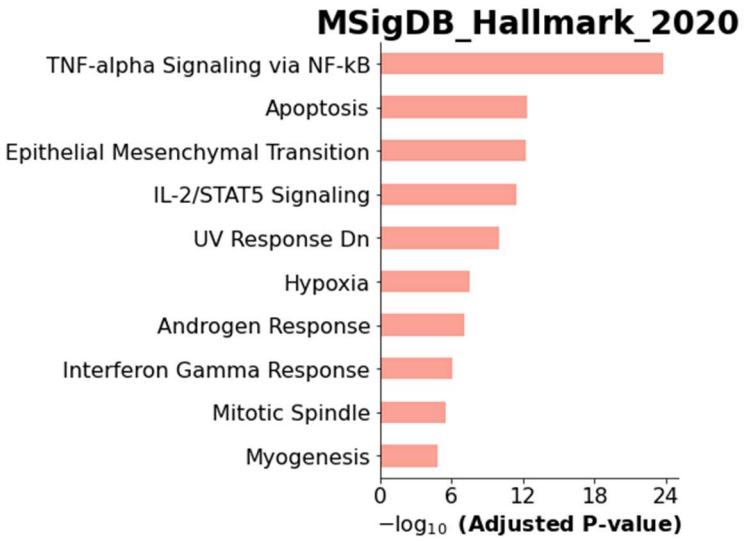
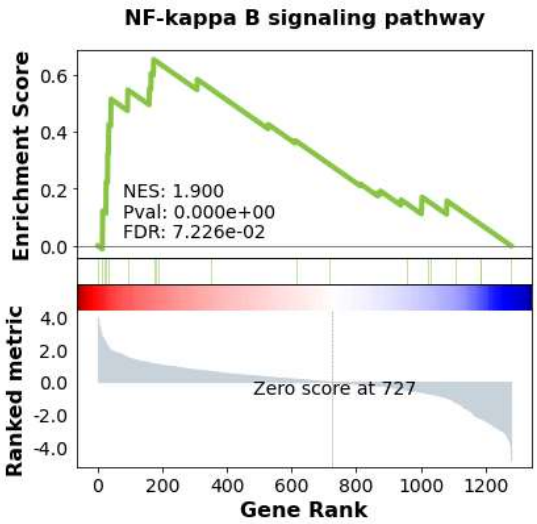


Fig. 3.8: Dotplot showing differential gene expression in primary and metastatic samples for top 20 genes that are differentially expressed in both groups. The first three rows refer to the primary tumor cells and the last three rows refer to the lymph tumor cells

(A)



(B)



(C)

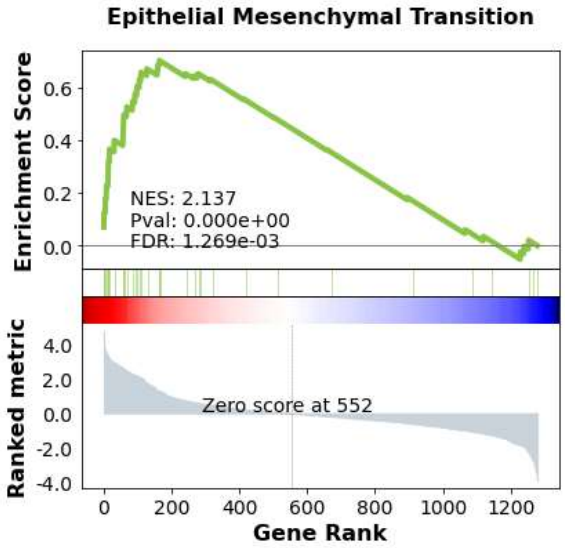


Fig. 3.9: (A) Figure showing pathways that were overrepresented since they had a higher number of genes associated with them, (B) Figure showing enrichment score calculated by gene set enrichment analysis of NF-kappa B signalling pathway (C) Figure showing enrichment score for Epithelial Mesenchymal Transition

Through further analysis using functional scoring methods we found that there was a higher occurrence of upregulated genes in EM transition as shown by the number of bars to the left in Fig 3.9 (C).

3.4: Compositional analysis:

Sccoda was used for performing compositional analysis.(Büttner et al., 2021) It's a Bayesian model based on Dirichlet-multinomial distribution. This model is based upon microbial compositional studies since they also deal with a high count of zeroes in the matrix. Based on the distribution, it calculates an expected parameter for each cell type. Further, it performs fold change based on the expected parameter for intercept and for the covariate. Here, for instance, the covariate is whether the tumor is primary or metastatic in nature. The results are shown in Fig. 3.10. The compositional changes in the cellular populations and the cellular heterogeneity of both the conditions can be seen in Fig 3.11.

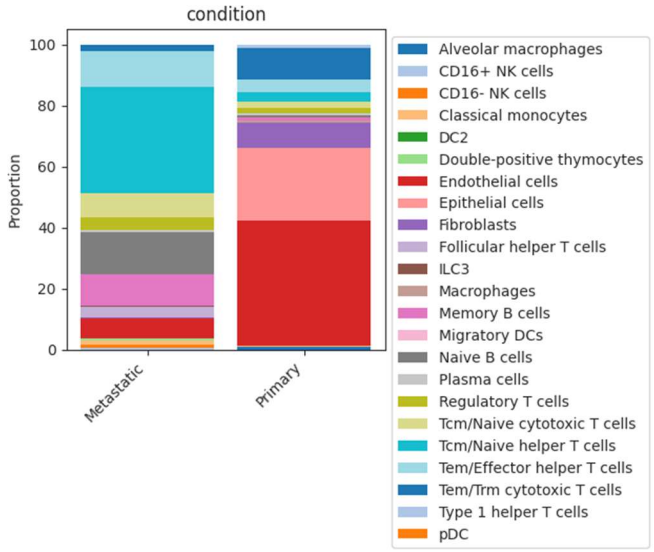


Fig. 3.10: Stacked bar plot showing cell type proportions between metastatic and primary tumor

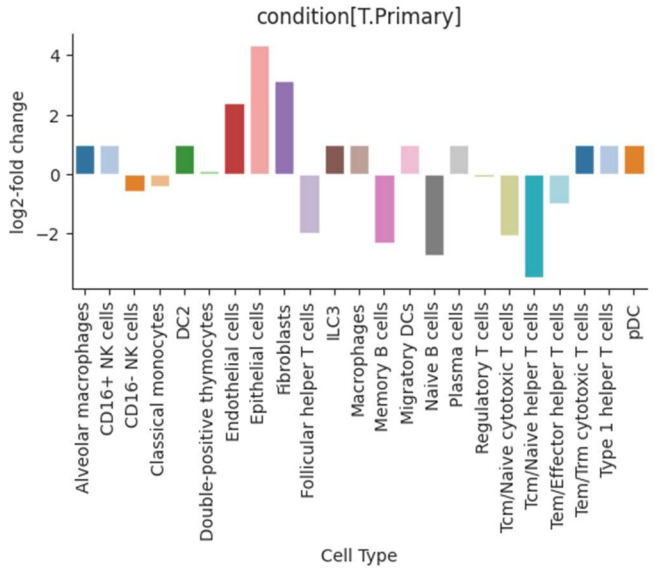


Fig. 3.11: Figure showing log2 fold changes between the two conditions as calculated from the model

4: Conclusion

This study aimed to get an insight into the cellular heterogeneity that is present in breast cancer tissues between primary tumors and metastatic tumors. There are intricate dynamics that are involved in a primary tumor changing to a more malignant and invasive form of metastatic tumor. These are important processes present both at a cellular and a genetic level. In fact, the cellular heterogeneity is the result of the gene expression hetereogeneity. Through scRNA-seq data, we have been able to capture both dimensions. At the cellular level, we found that there were different cell types present such as endothelial cells, epithelial cells, monocytes, helper T cells, killer T cells and fibroblasts in both the conditions using a reference dataset. Certain cell types like endothelial cells, epithelial cells and helper T cells showed a highly significant differences in their log fold-change value between the two conditions. Epithelial cells were found to be the most upregulated cell types followed by fibroblasts and endothelial cells. This could point out to the cellular mechanisms that promote metastatic tumor formation and may even give an insight into the cell-to-cell communication process. Further exploration can be performed using intercellular communication tools such as CellChat which infer such crosstalks between a “source” and a “receiver” cell type. At the genetic level, we found that there were genes such as *CD69*, *BACH2*, *SOX4* and *NFIB* that were significantly different in both the groups. We also found pathways like NF-Kappa B signalling pathway and epithelial mesenchymal transition that were found to be the major mechanisms through enrichment analysis. These pathways have also previously been implicated in breast cancer studies.(Greten & Karin, 2004; Sarkar et al., 2013; Shostak & Chariot, 2011; W. Wang et al., 2015; Y. Wang & Zhou, 2011; Wu et al., 2016) Thus, using scRNA-seq data, we were able to understand the cellular landscape of breast cancer at a highly resolved level.

Bibliography:

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology 2018 37:1*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Büttner, M., Ostner, J., Müller, C. L., Theis, F. J., & Schubert, B. (2021). scCODA is a Bayesian model for compositional single-cell data analysis. *Nature Communications 2021 12:1*, *12*(1), 1–10. https://doi.org/10.1038/s41467-021-27150-6

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*(1), 1–14. https://doi.org/10.1186/1471-2105-14-128/FIGURES/3

Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., King, H. W., Mamanova, L., Huang, N., Szabo, P. A., Richardson, L., Bolt, L., Fasouli, E. S., Mahbubani, K. T., Prete, M., Tuck, L., … Teichmann, S. A. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, *376*(6594). https://doi.org/10.1126/SCIENCE.ABL5197/SUPPL_FILE/SCIENCE.ABL5197_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Gene Expression Algorithms Overview -Software -Single Cell Gene Expression -Official 10x Genomics Support. (n.d.). Retrieved 12 July 2023, from https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview

Greten, F. R., & Karin, M. (2004). The IKK/NF-κB activation pathway - A target for prevention and treatment of cancer. *Cancer Letters*, *206*(2), 193–199. https://doi.org/10.1016/J.CANLET.2003.08.029

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology 2018 36:5*, *36*(5), 421–427. https://doi.org/10.1038/nbt.4091

Interpreting Cell Ranger Web Summary Files for Single Cell Expression Assay. (n.d.).

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/BIOSTATISTICS/KXJ037

Kuret, T., Sodin-Šemrl, S., Leskošek, B., & Ferk, P. (2022). Single Cell RNA Sequencing in Autoimmune Inflammatory Rheumatic Diseases: Current Applications, Challenges and a Step Toward Precision Medicine. *Frontiers in Medicine*, *8*, 822804. https://doi.org/10.3389/FMED.2021.822804/BIBTEX

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S. O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology 2020 21:1*, *21*(1), 1–35. https://doi.org/10.1186/S13059-020-1926-6

Lause, J., Berens, P., & Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, *22*(1), 1–20. https://doi.org/10.1186/S13059-021-02451-7/TABLES/2

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems*, *1*(6), 417. https://doi.org/10.1016/J.CELS.2015.12.004

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. https://doi.org/10.1093/BIOINFORMATICS/BTR260

Liu, Y. M., Ge, J. Y., Chen, Y. F., Liu, T., Chen, L., Liu, C. C., Ma, D., Chen, Y. Y., Cai, Y. W., Xu, Y. Y., Shao, Z. M., & Yu, K. Da. (2023). Combined Single-Cell and Spatial Transcriptomics Reveal the Metabolic Evolvement of Breast Cancer during Early Dissemination. *Advanced Science*, *10*(6). https://doi.org/10.1002/ADVS.202205395

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6), e8746. https://doi.org/10.15252/MSB.20188746

McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, *8*(4), 329-337.e4. https://doi.org/10.1016/J.CELS.2019.03.003

Method of the Year 2013. (2013). *Nature Methods 2014 11:1*, *11*(1), 1–1. https://doi.org/10.1038/nmeth.2801

Nayak, R., & Hasija, Y. (2021). A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics*, *113*(2), 606–619. https://doi.org/10.1016/J.YGENO.2021.01.007

Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A., & Park, J. E. (2020). BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, *36*(3), 964–965. https://doi.org/10.1093/BIOINFORMATICS/BTZ625

Potter, S. S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. *Nature Reviews Nephrology 2018 14:8*, *14*(8), 479–492. https://doi.org/10.1038/s41581-018-0021-7

Ribatti, D., Tamma, R., & Annese, T. (2020). Epithelial-Mesenchymal Transition in Cancer: A Historical Overview. *Translational Oncology*, *13*(6), 100773. https://doi.org/10.1016/J.TRANON.2020.100773

Sarkar, D. K., Jana, D., Patil, P. S., Chaudhari, K. S., Chattopadhyay, B. K., Chikkala, B. R., Mandal, S., & Chowdhary, P. (2013). Role of NF-κB as a Prognostic Marker in Breast Cancer : A Pilot Study in Indian Patients. *Indian Journal of Surgical Oncology*, *4*(3), 242. https://doi.org/10.1007/S13193-013-0234-Y

Shostak, K., & Chariot, A. (2011). NF-κB, stem cells and breast cancer: The links get stronger. *Breast Cancer Research*, *13*(4). https://doi.org/10.1186/BCR2886

Street K, Townes F, Risso D, & Hicks S. (2023). scry: Small-Count Analysis Methods for High-Dimensional Data. In *scry: Small-Count Analysis Methods for High-Dimensional Data. R package version 1.12.0*.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols 2018 13:4*, *13*(4), 599–604. https://doi.org/10.1038/nprot.2017.149

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, *20*(1), 1–16. https://doi.org/10.1186/S13059-019-1861-6/FIGURES/5

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports 2019 9:1*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-41695-z

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1), 1–32. https://doi.org/10.1186/S13059-019-1850-9/TABLES/1

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Angerer, P., Bergen, V., Boyeau, P., Büttner, M., Eraslan, G., Fischer, D., Frank, M., Hong, J., Klein, M., Lange, M., Lopez, R., … Theis, F. J. (2023). The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology 2023 41:5*, *41*(5), 604–606. https://doi.org/10.1038/s41587-023-01733-8

Virshup, I., Rybakov, S., Theis, F. J., Angerer, P., & Wolf, F. A. (2021). anndata: Annotated data. *BioRxiv*, 2021.12.16.473007. https://doi.org/10.1101/2021.12.16.473007

Wang, W., Nag, S. A., & Zhang, R. (2015). Targeting the NFκB Signaling Pathways for Breast Cancer Prevention and Therapy. *Current Medicinal Chemistry*, *22*(2), 264. https://doi.org/10.2174/0929867321666141106124315

Wang, Y., & Zhou, B. P. (2011). Epithelial-mesenchymal transition in breast cancer progression and metastasis. *Chinese Journal of Cancer*, *30*(9), 603. https://doi.org/10.5732/CJC.011.10226

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 1–5. https://doi.org/10.1186/S13059-017-1382-0/FIGURES/1

Wu, Y., Sarkissyan, M., & Vadgama, J. V. (2016). Epithelial-Mesenchymal Transition and Breast Cancer. *Journal of Clinical Medicine*, *5*(2). https://doi.org/10.3390/JCM5020013

Young, M. D., & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, *9*(12), 1–10. https://doi.org/10.1093/GIGASCIENCE/GIAA151

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications 2017 8:1*, *8*(1), 1–12. https://doi.org/10.1038/ncomms14049