X-Chromosome Inactivation in Human Adipose-Derived Mesenchymal Stem Cells using Single Cell RNA Sequencing

by Helen Ibeawuchi

B. Sc. in Biochemistry, April 1994, University of Calabar
MB BCh. in Medicine and Surgery, November 2001, University of Calabar

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

June 30, 2020

Thesis directed by

Anelia Horvath
Associate Research Professor of Biochemistry and Molecular Medicine

© Copyright 2020 by Helen Ibeawuchi
All rights reserved

Dedication

This thesis is dedicated to my late father Obong Udo Warrie Uso the pillar behind my academic pursuit, my late friend and brother Dr Chukwujindu Victor Mbakpuo, Dr and Mrs Michael Akpe, Dr and Mrs Nathaniel Adiakpan, and my children Izunna and Munachi for inspiring, encouraging, and supporting me throughout this academic journey.

Acknowledgments

I would like to thank my mentor and advisor Dr. Anelia Horvath for her guidance and support, and the members of Horvath lab; Prashant NM, Hongyu Liu, and Justin Sein. My appreciation also goes to Dr. Brett Shook for reviewing this work. Many thanks to Professor Jack Vanderhoek, Professor Raja Mazumder, and the George Washington University Department of Biochemistry and Molecular Medicine for giving me the opportunity to conduct this research. Finally, I would like to thank the George Washington University for providing me with the platform for another academic experience.

Abstract of Thesis

X-Chromosome Inactivation in Human Adipose-Derived Mesenchymal Stem Cells using Single Cell RNA Sequencing

Mesenchymal stem cells are multipotent stem cells that have the ability to differentiate into different cell types including adipocytes. Adipose derived mesenchymal stem cells (ADSCs) are considered to be the best source for mesenchymal cell isolation since adipose tissue is easily accessible and can be obtained using less invasive procedures. These cells have promising clinical applications in medicine. During embryonic development, X-inactivation occurs randomly in individual cells. ADSCs just like other cells undergo X-inactivation. X-chromosome inactivation (XCI) is a phenomenon in which one X chromosome in the female is randomly silenced. Some genes; however, escape this process and some of the escapee genes have been linked to genetic diseases.

The main goal of this study was to develop and assess an approach to estimate cell-type specific XCI from single cell RNA-sequencing (scRNA-seq) data generated using 3'protocol and utilizing Unique Molecular Identifiers (UMIs). To this end, the specific objectives of this study were to identify the different cell types present in adipose-derived mesenchymal stem cells, establish their XCI status, and compare the findings to existing knowledge in relation to disease and medicine, using single cell RNA sequencing (scRNA-seq) data. To perform this research, we used publicly available scRNA-seq data from 26,640 human adipose-derived mesenchymal stem cells from three healthy females. First, we estimated the different cells types. Across the three

individuals, we identified adipose cells, erythrocytes, neutrophils, and naïve B-cells. Within each cell type, we then assessed the expressed variant allele fraction (VAF_{RNA}) at all heterozygous Single Nucleotide Variant (SNV) positions on the X-chromosome. We identified cell-type specific genes that escape the chromosome X-inactivation. These included some genes not previously described as escapees in other tissues. The relationship between these escapee genes and their association with certain diseases was also established. Our future plan is to study the molecular mechanisms and biological pathways related to XCI escapees in health and disease. This will help in the understanding of the clinical implications and possibly suggests management strategies of these diseases to provide a better clinical outcome.

Table of Contents

Dedication……...…..iii

Acknowledgement…….…….....iv

Abstract of Thesis……..v

List of Figures…….....….ix

List of Tables……...……...…...xi

List of Equations……...…..xii

List of Symbols/Nomenclature……..……..……..……...….xiii

Glossary of Terms……...xiv

Chapter 1: Introduction……...……1

Chapter 2: Literature Review……...……...……4

Chapter 3: Methods…….....7

3.1 Data Acquisition……7

3.2 Data Processing……..8

3.3 Filtering Cells…....……11

3.4 Estimation of Variant Allele Fraction (VAF_{RNA})…...…….…….....13

3.5 Data Visualization……....13

Chapter 4: Results…….….14

4.1 Genome-wide Assessment of Allele Expression from scRNA-seq Data.…...17

4.2 Cell Type Identification……...27

4.3 Identification of Escapee Genes……...29

4.4 Escapee Genes and Clinical Abnormalities…….31

Chapter 5: Discussion and Conclusion……..36

5.1. Discussion……..….36

5.2. Conclusion…….38

References…….40

List of Figures

Figure 1. Analytical workflow for estimation of the variant allele fraction (VAF_{RNA}) from single-cell RNA-sequencing (scRNA-seq) data, cell-type identification from ADSCs, X-chromosome VAF_{RNA} assessment, and identification of XCI escapees.……...10

Figure 2. Results of filtering to remove low quality cells shows the number of genes, number of sequencing reads, and percent of mitochondrial genes for N8, N7, and N5…….....12

Figure 3. Distribution of statistical values for VAF_{RNA} at minR 10, 5, and 3 which were symmetrical at all thresholds with regards to VAF_{RNA} scale…....18

Figure 4. Genome-wide $ScVAF_{RNA}$ estimation at positions covered by a minimum of 10 sequencing reads, 5 sequencing reads, and 3 sequencing reads, across more than 1000 cells.……..21

Figure 5. $ScVAF_{RNA}$ distribution at positions covered by a minimum of 10 sequencing reads, and three sequencing reads across more than 1500 cells for genes reported by Borel et al[27]……24

Figure 6. Chromosome-wise $scVAF_{RNA}$ estimations at positions covered by a minimum of 10 sequencing reads, 5 sequencing reads, and 3 sequencing reads, across more than 1500 cells or in between 10 and 1500 cells with N8 sample……26

Figure 7. Clusters of different cell types identified from across the three ADSCs samples…….28

Figure 8. $ScVAF_{RNA}$ estimations on the X chromosome in the different cell types across all 3 samples at positions covered by a minimum of 10 sequencing reads shows RME…….....30

Figure 9. $ScVAF_{RNA}$ estimations on the X chromosome in the different cell types across all 3 samples at positions covered by a minimum of 5 sequencing reads shows both RME and biallelic patterns…...……...33

Figure 10. Gene set enrichment analysis (GSEA) of XCI escapee genes and associated clinical abnormalities……34

Figure 11. Gene disease association of some of XCI escapee genes and the link between the escapee genes and the associated diseases……...35

List of Tables

Table 1. Summary of the statistics on the scRNA-seq data of the 3 ADSC samples……16

Table 2. Percentage of monoallelic and bi-allelic expression of SNVs covered with different required minimum counts of sequencing reads……19

List of Equations

Equation 1: VAF_{RNA} calculation 13

List of Symbols/Nomenclature

1. XCI - X-Chromosome Inactivation
2. ASE - Allele Specific Expression
3. VAF - Variant Allele Frequency
4. SNV - Single Nucleotide Variant
5. ADSC - Adipose Derived Mesenchymal Stem Cell
6. scRNA-seq - Single cell RNA Sequencing
7. RME - Random monoallelic expression
8. minR - Minimum number of Reads

Glossary of Terms

Allele specific expression (ASE); A phenomenon where there is unbalanced expression of alleles at a genomic locus in a tissue that is diploid in origin. This is Allelic imbalance.

Minimum number of Reads (minR); The number of unique sequencing reads required for the SNV locus to qualify for VAF_{RNA} estimation.

Random monoallelic expression (RME); A phenomenon where there is sustained transcription from only one of the two parental alleles in diploid cells.

RNA-editing; A post-transcriptional, molecular process through which cells can change specific nucleotide sequences which may be insertion, deletion, and base substitution in RNA transcripts.

RNA variant allele frequency (VAF_{RNA}); A continuous metric equal to the proportion of RNA-sequencing read counts harboring a specific variant to the total read count (equation 1).

Single cell RNA Sequencing (scRNA- seq); A technique that provides the expression profiles of individual cells.

Single nucleotide variant (SNV); The replacement of one nucleotide with another at a single position in a DNA sequence. A base change.

.

Stochastic; A random determined process, having a random probability distribution that may be analyzed statistically without a precise prediction.

X-Chromosome Inactivation (XCI); A phenomenon in which one X-chromosome is transcribed in the female cell while the second is silenced.

XIST; A long non coding RNA (lncRNA) that coats the inactive X-chromosome.

Chapter 1: Introduction

X-chromosome inactivation (XCI) is a well-known phenomenon in which only one X-chromosome is transcribed in the female cell while the second is silenced [1-4]. This is a mechanism that allows for female dosage compensation since the female has two copies of X chromosome containing approximately 900 to 1500 genes while the male has one X and one Y chromosome with approximately 70 genes [5]. Some of the genes found on the inactive X-chromosome (Xi) escape this phenomenon and are therefore expressed. These genes are known as XCI escapees. The process of XCI starts during early embryogenesis at the X-Inactivation center (XIC) and gradually spreads along the silenced X-chromosome [1]. XIST, a long non coding RNA (lncRNA) coats the inactive X-chromosome and is required for XCI [7-10]. The initiation phase is also mediated by some cellular proteins that are recruited to participate in the process of XCI [6,11]. Chromatin repressive modifications are important hallmark of gene silencing in XCI [12]. Some of these modifications include; an increase in histone repressive marks such as H3K27me3 [13,14], DNA hypermethylation at CPG islands on Xi, and depletion of RNA polymerase II. However, genes that escapee XCI appear to be protected from these repressive modifications [69]. Studies have shown that escapees of XCI in females constitute about 15-25% of X-linked genes and 10% of these genes show a difference between cell types and individuals [15-18]. Some of these genes are documented as being associated with certain clinical abnormalities in patients [19,20,21-26] such as Kabuki syndrome and intellectual disabilities.

Recent studies have demonstrated the ability to study single cells and their ASE [27,77]. These studies have shown that scRNA-seq single nucleotide variant (SNV) assessment can be employed in the determination of random monoallelic expression (RME), dynamic transcriptional bursting of expression [27-31], X - chromosome inactivation [32,33], genetic heterogeneity in cancer [34-38], aneuploidy [39], haplotype inference [40], demultiplexing [41,42,43], and quantitative trait loci (QTL) assessments [44]. Therefore, this assessment is useful in identifying the genes that escape XCI by exploring these patterns at single cell level [45]. However, most of these studies carried out their analysis on full length transcripts and a systemic analysis of VAF_{RNA} estimations from 3′-based scRNA-seq data have not yet been explored. The unique advantage of using scRNA to study XCI as opposed to bulk is that since XCI happens randomly, bulk RNA cannot be used to distinguish which X-chromosome is silenced. In contrast, scRNA-seq allows the precise identification of the X-chromosome that is subject to XCI and the level to which it is silenced.

In this study we developed and tested a new approach to assess XCI from scRNA-seq data generated via 3'UTR protocol utilizing UMIs (10x Genomics). The rationale for using this approach is that 10x Genomics is becoming increasingly popular and has UMI allowing for precise estimation of the number of reads. Using this approach, we defined the different cell types in ADSCs, assessed the expressed variant allele fraction (VAF_{RNA}) at all heterozygous Single Nucleotide Variant (SNV) positions on the X-chromosome in each cell type, and identified cell-type specific genes that

escaped X-chromosome inactivation some of which have not been previously reported as XCI escapees.

Chapter 2: Literature Review

In the recent years, there has been many studies on Mesenchymal stem cells (MSCs) as these have promising applications in clinical medicine some of which include cell therapy and tissue regeneration [46]. Adipose tissue is considered ideal for mesenchymal stem cell isolation because it is ubiquitous in nature and contains more MSCs than other sources constituting up to 3% as opposed to bone marrow with 0.01%.[47,48] ADSCs have emerged as a class of multipotent stem cells that is useful in tissue repair some of which include; cutaneous tissue wound healing, cardiac regeneration, and soft tissue reconstruction [49,50]. The inability to fully understand cellular heterogeneity has impaired the clinical application of MSCs in medicine. However single cell RNA-seq (ScRNA-seq) has emerged as a powerful tool to analyze cellular heterogeneity without prior knowledge of the cell populations [51]. ScRNA-seq has also become available for genomic studies [52,53,54] by enabling transcriptome analyses at the cell level. The use of single cells provides the advantage of exploring and establishing intercellular relationships as opposed to using bulk RNA-seq. Since the emergence of single cell technologies, many studies have adopted this platform to assess and estimate genetic variation with more accurate results [55,56,57]. Using scRNA-seq data, genetic variation can be assessed by estimating the variant allele fraction (VAF_{RNA}) [43,58 ,59 ,60 ,61 ,62 ,63], which is a continuous measure that allows for a precise quantification of allele expression.

Previous studies used different approaches to identify XCI Escapee genes from whole tissue. These include detecting allelic imbalance in clonal lymphoblast and

fibroblast cell lines (15), use of methylation profiles to identify active and inactive transcription start sites [64], comparison of X-linked gene expression between males and females [65], and among females with X-chromosome aneuploidies [66]. Recently, some studies have used scRNA-seq to established XCI in various tissues. Garieri et al [67] studied X-linked ASE in fibroblast and lymphoblast and established XCI. Tukiainen et al [68] studied X-inactivation in lymphoblasts and dendritic cells and partially validated their observation. Some escapee genes identified by these studies are known to be associated with clinical abnormalities in patients diagnosed with X-chromosome aneuploidy, including Turner and Klinefelter syndromes [69]. Some variants that cause disease have been identified to be associated with many phenotypes in female carriers, notably Kabuki syndrome [KABUK1 (OMIM 147920)] [19, 20], and intellectual disabilities [21–26].

ScRNA-seq is an increasingly popular platform for various analyses. The use of scRNA-seq to define genes that escape XCI and their association with clinical abnormalities is currently being explored. No study has used the approach of estimating VAF_{RNA} from 3′-based scRNA-seq libraries to answer biological questions as yet. The prevalence of random monoallelic expression across the human genome is a process that can be determined using VAF_{RNA} estimations. Recently, many studies have used scRNA-seq to describe widespread RME in both human and mouse models [27,28,29,30,31]. The Majority of these studies used full length transcripts from hundreds of cells to perform their analysis. Estimation of VAF_{RNA} from scRNA-seq data generated on 10x Genomics platform allows assessments of thousands of cells per

sample. In this study, we explored the feasibility of XCI assessment using scRNA-seq data generated via10x Genomics 3′-based protocol. To do that, we defined the different cell types in ADSCs, estimated their allele specific expression by assessing the expressed VAF (VAF_{RNA}) at all heterozygous Single Nucleotide Variant (SNV) positions on the X-chromosome, and identified cell-type specific genes that escaped X-chromosome inactivation.

Chapter 3: Methods

Bioinformatics analysis was carried out on cells that have been previously isolated and cultured. The data was acquired and processed to removed confounders like doublets, batch effect, and cell cycle effect giving a clean matrix that was used for further analysis. The process for data acquisition and processing is outlined in this section. It involved use of a pipeline which included quality assessment, filtering, read processing, UMI and cell barcode processing, alignment, variation call, allele-specific bias removal, VAF_{RNA} estimation, gene expression estimation and normalization, identification of cell types and across-cell-types comparative XCI assessment. This pipeline was used to identify the different cell types and further analysis was done to establish the XCI status of the identified cell types.

3.1. Data Acquisition

The data was obtained from publicly available scRNA-seq data from 26,640 human adipose-derived mesenchymal stem cells from three healthy females designated as N8, N7, and N5. The scRNA-seq data was generated on 10x Genomics Chromium v2 platform.[46]. Detailed description of library construction and sequencing is documented in a previous study [46]. Briefly, it involved loading of the cells onto 10x Genomics Single Cell 3′ Chips, addition of barcodes to index cells (14 bp), and incorporation of UMIs (10 bp UMI). The resulting libraries were sequenced on an Illumina NovaSeq 6000 System in a 2 × 150 nucleotides paired-end mode.

3.2. Data Processing

The data was preprocessed and a read count matrix per cell was generated and used for the analysis. The following steps were employed according to the pipeline shown in Figure 1. The first step was the extraction of cell barcodes and unique molecular identifiers (UMIs) from the combined raw sequencing reads.[71]. The next step was the alignment of the reads using STAR v 2.7.3.c.[72]. The reads were aligned to the latest version of the human genome reference (GRCh38, Dec 2013) in 2-pass mode with transcript annotations from the assembly GRCh38.79. Deduplication was performed using UMI-tools. The reads with the highest mapping scores were retained. After deduplication, SNVs were called in the pooled alignments using GATK v.4.1.4.1 [59]. The criteria used for the selection of heterozygous SNVs was based on the presence of a minimum of 50 high quality reads in the pooled alignments in support of each, the reference and alternative nucleotides. From this pool, SNVs that met criteria for high quality heterozygous calls were retained for further analysis. The criteria included; QUAL (Phred-scaled probability) > 100, MQ (mapping quality) > 60, QD (quality by depth) > 2, and FS (Fisher's exact test estimated strand bias) = 0.000.

Other annotation tools used to remove single nucleotide variant loci found in intergenic or repetitive regions included SeattleSeq v.13.00 and Single Nucleotide Polymorphism database (dbSNP) build 153. A second SNV-aware alignment was done with the SNV list for each donor using STAR with the WASP-option [70,72] to remove reads mapped ambiguously due to the variant nucleotides. Estimation of raw gene counts

was done using featureCount [73], followed by normalization and scaling with Seurat v.3.0 [74].

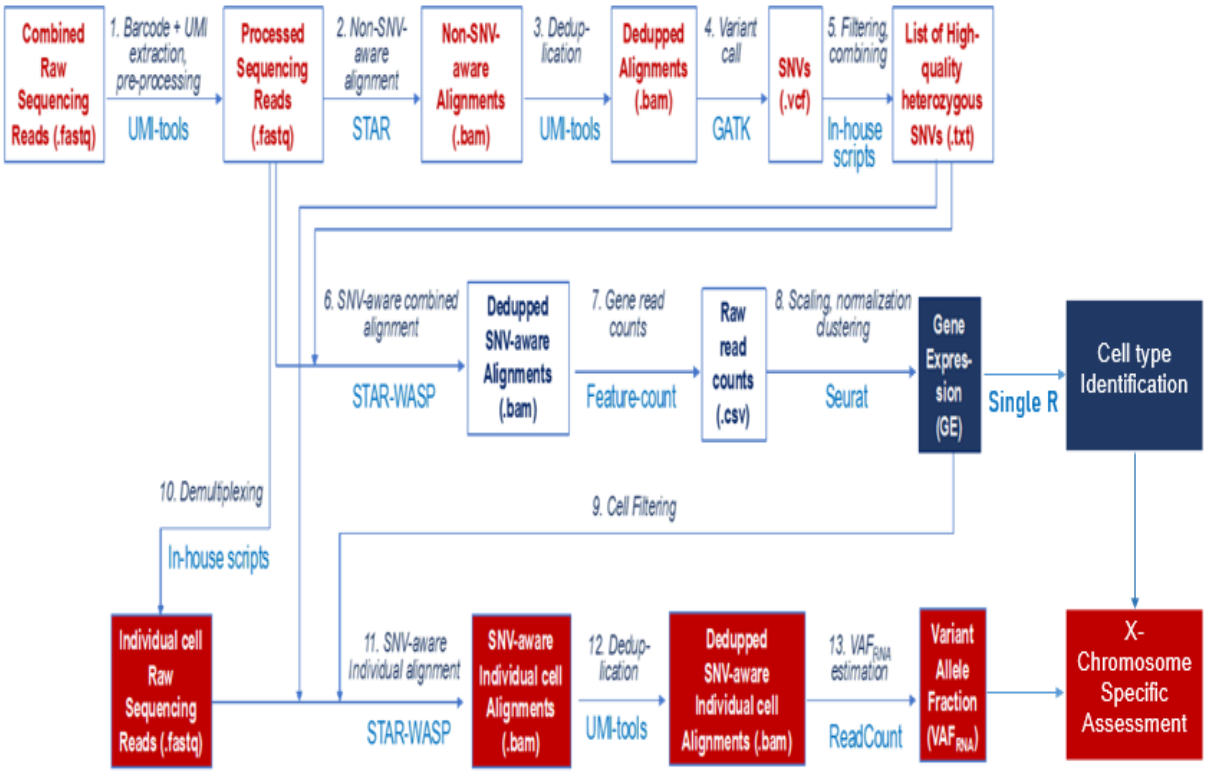


Figure 1. Pipeline for scVAF$_{RNA}$ estimation. ScVAF$_{RNA}$ is estimated from deduplicated (UMI-tools[2]), SNV-aware alignments in each cell separately. To generate SNV-aware alignment, high-quality heterozygote calls produced by GATK[3] from the non-SNV-aware alignments generated by STAR[4], are used. These SNVs were combined into a list of unique SNV loci and used for a second-pass STAR-alignment, this time using the option for WASP[5] to remove reads mapped ambiguously due to the variant nucleotide. This second, SNV-aware alignment was performed in parallel on the combined (per individual/sample) and demultiplexed (per cell) raw sequencing data. The combined alignment was then used to estimate row gene counts using FeatureCounts[6]. Counts scaling and normalization was performed using Seurat[7], and cells with low quality data were filtered out. On the high quality individual alignments, we estimated scVAF$_{RNA}$ using ReadCounts[8] at three thresholds for required minimum number of unique RNA-seq reads;10, 5 and 3.[85]. Cell type identification was done followed by X-chromosome specific assessment for XCI escapee genes.

3.3. Filtering Cells

This is an important step in analyzing single cell RNA-Seq data. Filtering removes low quality cells like doublets, cells that are damaged during isolation, or cells that have few reads. Cells with very high read count or high number of detected genes are likely to be doublets while those with a high percentage of mitochondrial genes are probably damaged or dying cells. Read counts and number of detected genes per cell are used to filter out doublets while the percentage of mitochondrial genes are used to filter non-viable cells. Filtering was done with the filters set to remove cells that had low quality data; that is cells with less than 3000 detected genes, or greater than 6% Mitochondrial genes. Figure 2 shows the RNA-seq reads and distribution of genes before and after filtering.

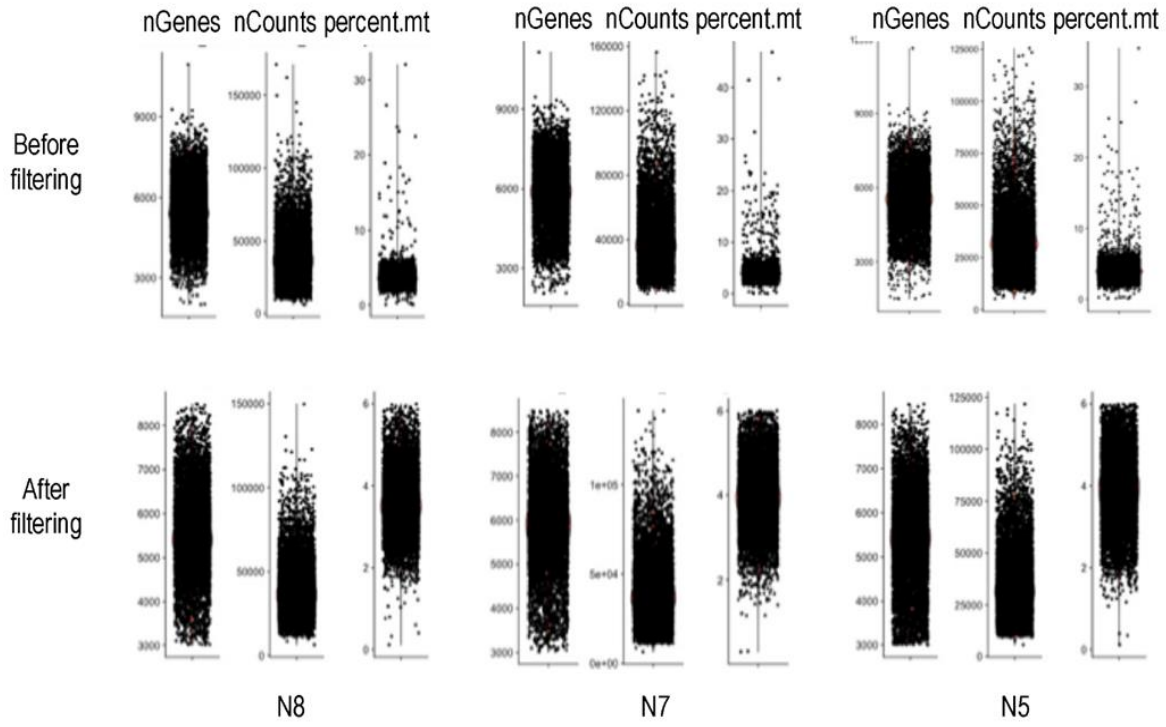


Figure 2. Filtering to remove low quality cells. The number of genes, number of sequencing reads, and percent of mitochondrial genes for N8, N7, and N5 are shown, before (on the top) and after (on the bottom) filtering [85]. Within each individual, the cells were filtered to remove low quality data, the following filters were applied: minimum cells = 300, minimum features =2000, Filter set for nFeature_RNA > 3000 & nFeature_RNA < 8500 & percent mitochondrial genes < 6.

3.4. **Estimation of Variant Allele Fraction (VAF_{RNA})**

After filtering, the number of reads harboring the variant (nvar) or the reference (nref) nucleotides were estimated from the individual cell alignments using ReadCounts [75]. ReadCounts filtered aligned reads based on alignment quality metrics to retain only reads with the requested quality values [75]. VAF_{RNA} was estimated using the following formula;VAF_{RNA} = nvar/(nvar + nref) at three different thresholds on requirement for minimum number of reads (minR) covering the locus of interest; minimum of ten reads (minR = 10), minimum of five reads (minR = 5), and minimum of three reads (minR = 3). Positions covered by reads less than the minR were excluded from the analysis. For each analysis, minR was kept constant across the genome. The remaining reads were categorized as having the reference or the variant allele.

3.5. Data Visualization

To visualize the clusters for the different cell types identified, an R package was used to generate UMAP (Uniform Manifold Approximation and Projection) and 3D t-SNE (t-Distributed Stochastic Neighbor Embedding) for the data using Plotly and cell ranger Rkit. Stacked bar plots were used to visualize VAF_{RNA} estimations for the individual chromosomes.

Chapter 4: Results

The number of detected genes per cell, number of read counts per cell, and the percentage of mitochondrial reads per cell are shown on Figure 2. Each point on the plots represents a cell and the clusters show how the values on the y-axis metric are distributed before and after quality filtering. For N8, N7, and N5, the number of cells retained after filtering were 9115, 8125, and 8533 respectively. From the high quality cells, VAF_{RNA} estimation was done. 50,532 heterozygous SNV genomic positions were assessed in N8, 61,407 in N7, and 38,822 in N5. These were the number of genomic positions retained after filtering for heterozygosity, the quality of the cell, and the position in intragenic non-repetitive regions. Analyses were performed at all threshold for minimum required sequencing reads: minR = 10, minR=5 and minR = 3. Naturally, the confidence of the VAF_{RNA} estimation increases with the number of sequencing reads required to cover each locus of interest. As in our analysis minR=10 supports the highest confidence in the VAF_{RNA} estimation, we selected it to demonstrate our approach. Accordingly, unless otherwise indicated, all the results hereafter are obtained using minR = 10. To further facilitate the confidence of the estimations, we retained analyses performed from a minimum of 10 cells (per donor).

The absolute number of these positions with minR = 10 was 366 for N8, 431 for N7, and 277 for N5. This number was approximately 4-times higher for positions that were assessed at minR = 5 and up to 20-times higher for positions assessed at minR = 3; as shown on Table 1 (the relaxed threshold is inclusive of the more stringent ones). About 6% to 14% of single nucleotide variants selected using this stringent level are known to

be associated previously with a clinical phenotype or highlighted in genome-wide associated studies [Table 1].

Table 1. Summary of the statistics for scRNA-seq data of the three ADSC samples. We estimated the number of single nucleotide variant (SNV) loci covered in at least 10 individual cells (per donor) with different thresholds for the minimum number of unique reads (minR), and from those, the number of SNVs associated with a phenotype (phenotype and clinical associations were extracted via SeattleSeq database annotation (v13, dbSNP build 153))[85].

Sample	N Cells	N Reads	Mean Reads/ Cell	Median Genes/ Cell	N cells Post filtering	No. of Heterozygous SNVs (min 10 cells)			No. of SNVs with Phenotype/Clinical associations		
						minR 10	minR 5	minR 3	minR 10	minR 5	minR 3
N8	9256	1,285,218,728	138,852	5559	9,115	366	1,567	7253	47	181	552
N7	8478	1,579342505	186,287	6049	8,125	431	1994	9032	57	184	568
N5	8906	1,071,156,174	120,273	5439	8,533	277	1134	5375	23	134	422

4.1 Genome-wide Assessment of Allele Expression from scRNA-seq Data

As mentioned above, VAF_{RNA} statistics was computed using at least 10 individual cells for all SNVs loci that was analyzed using the required minimum number of reads which were minR = 10, minR = 5, and minR =3. Statistical assessments included mean and median VAF_{RNA} estimations, and percentage of cells that exhibited monoallelic and biallelic expression. The distribution of these statistical values for VAF_{RNA} were symmetrical at all thresholds with regards to VAF_{RNA} scale as shown in Figure 3. In addition, more than half of the VAF_{RNA} estimations ranged between $0.2 < VAF_{RNA} < 0.8$, corresponding to bi-allelic expression (Table 2). More than 50% of VAF_{RNA} estimation at minR = 3 corresponded to bi-allelic expression with an increase in this proportion to approximately 90% with VAF_{RNA} estimation limited to minR = 10.

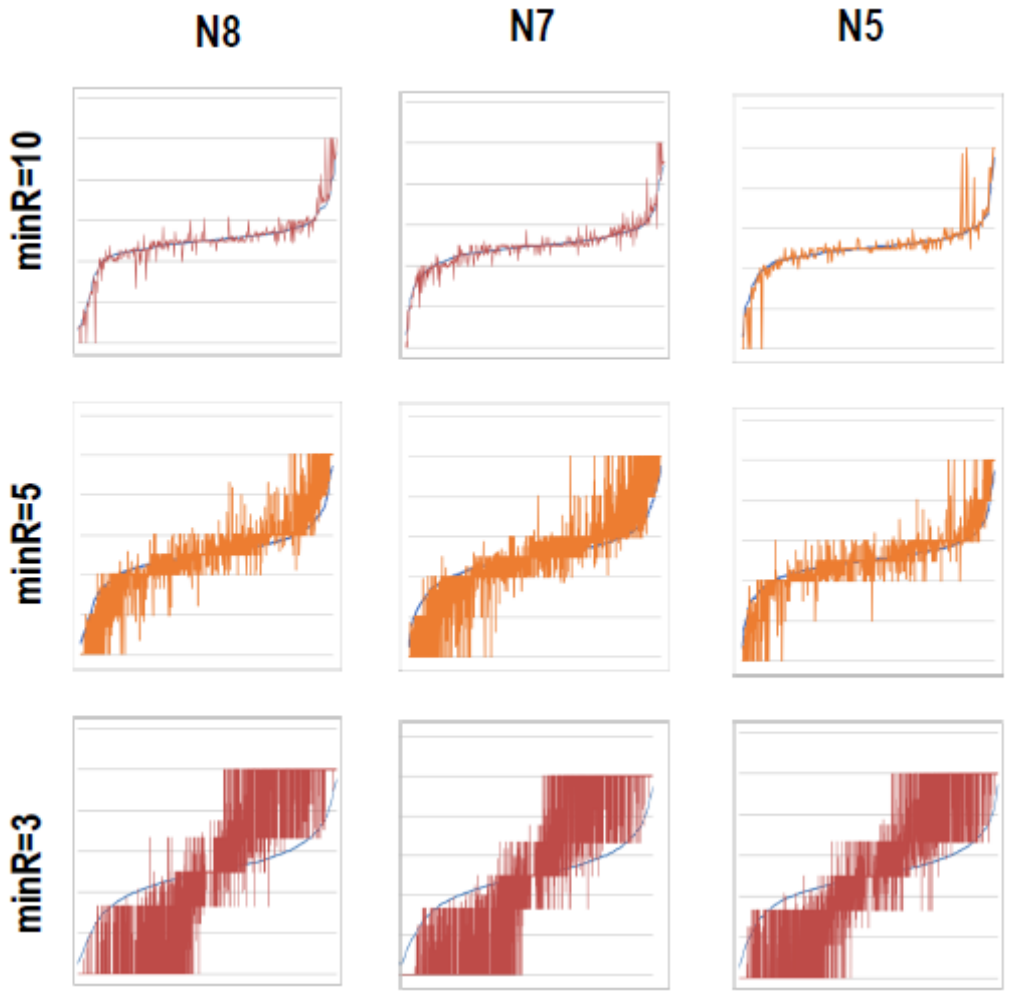


Figure 3. Distribution of statistical values for VAF_{RNA} at minR = 10 (top), minR = 5 (middle), and minR = 3 (bottom). Mean (blue) and median (orange) VAF_{RNA} were symmetrical at all thresholds with regards to VAF_{RNA} scale [85].

Table 2. Percentage of monoallelic and biallelic expression of SNVs that are covered at different required minimum number of sequencing reads.[85]

Sample	% strictly monoallelic VAF_{RNA} = 0 or 1			% predominantly monoallelic VAF_{RNA} = 0 - 0.2 or 0.8 -1			% biallelic VAF_{RNA} = 0.2 - 0.8		
	minR 10	minR 5	minR 3	minR 10	minR 5	minR 3	minR 10	minR 5	minR 3
N8	4.4	15.1	38.2	16.1	30.6	45.5	83.9	69.4	54.5
N7	2.9	15.8	41.3	13.1	30.3	47.9	86.9	69.7	52.1
N5	2.4	12.5	35.9	10	26.1	42.0	90	76.9	58.0

The distribution of VAF_{RNA} for SNVs in the entire genome is shown on Figure 4. At minR = 10, the majority of the SNV position had a significant number of cells with VAF_{RNA} estimations between 0.2 and 0.8. This observation corresponds to biallelic expression for genes with relatively high expression (i.e. covered by a minimum of 10 unique sequencing reads. At lower minR, the percentage of cells with biallelic expression decreased. The later observation can result from two factors. The first one is technical and reflects the well-known phenomenon "randomness of sampling" which refers to the fact that at low number of reads (i.e. minR = 3) there is a considerable likelihood to randomly obtain reads harboring only one of the two alleles (i.e. bearing only the reference or only the alternative nucleotide). The second is the biological phenomenon referring to the bursting mode of gene expression. The burst-like expression has been recently analyzed via scRNA-seq studies, where it shows that it is more characteristic for genes with low expression. In our analysis, genes with low expression correspond to the genes captured with lower minR (minR = 3 and minR = 5). Our results showed that the distribution of VAF_{RNA} depends on the minimum number of reads that is required for an SNV to be included in the analysis.

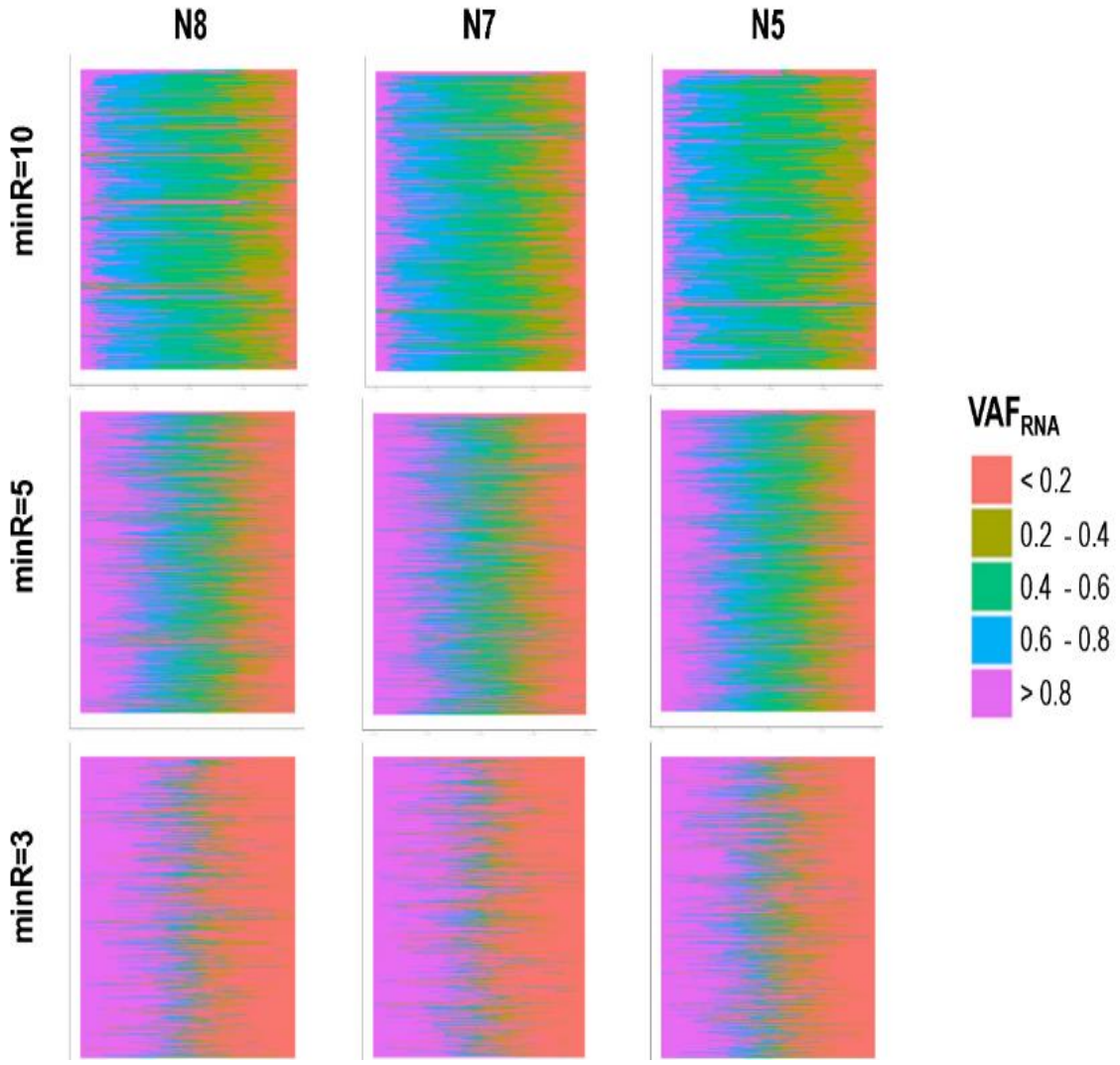


Figure 4. Genome-wide scVAF$_{RNA}$ estimated at positions covered by a minimum of 10 sequencing reads (top), 5 sequencing reads (middle), and 3 sequencing reads (bottom), across more than 1000 cells. For the majority of positions, VAF$_{RNA}$ showed bi-allelic expression, with a substantial proportion of the scVAF$_{RNA}$ estimations in the interval 0.4–0.6.[85]

As mentioned earlier, when positions covered by a high number of reads are included, the majority of VAF_{RNA} estimations show biallelic expression. This analysis however was limited to the low number of SNVs covered by a minimum of 10 unique sequencing reads. We note that current scRNA-seq datasets generated using 3'UTR protocol and UMI (such as 10x Genomics which we used for our analysis) have relatively low number of SNVs positions covered by a minimum of 10 unique sequencing reads. This number, however is expected to quickly grow in the near future with the constantly improved scRNA-seq technologies and protocols. When the requirement for minimum number of reads is lowered, the number of SNVs included in the analysis increases. This increases the positions where mono allelic patterns are observed. However, since reducing the minimum number of reads can have a stochastic effect, thereby affecting VAF_{RNA} estimations, the increased number of monoallelic expression by the SNVs could be due in addition to biological factors, to technical noise. To evaluate the effect of technical factors, VAF_{RNA} estimations carried out at different SNV loci on the same gene were compared for consistency and this was done across the three samples. The analysis was also compared with previous studies by Borel et al [27].

To assess the allele specific expression at the gene level, estimation of VAF_{RNA} was done at all three thresholds beginning at the minimum required sequencing reads of 10. From the analysis, the 21 genes present showed biallelic expression and this correlated with studies by Borel et al where they had 60 genes. The 21 genes from this study were among the 60 genes shown by Borel et al [27] (Figure 5 below). At minR = 5, the autosomal genes showed random monoallelic expression. This was comparable to

findings from the study by Borel et al [27]. Some of these genes are shown in Figure 5**,** example; RAD52 which is strictly monoallelic. Genes on the X chromosome show strong random monoallelic pattern, for example TSPAN6.

Single cell VAF_{RNA} estimation at minR = 10 for autosomal SNVs showed no random monoallelic expression (N8 is shown on Figure 6). With a minimum of 3 sequencing reads as expected, frequent monoallelic signals were observed for all autosomes. However, on X chromosome, strong monoallelic patterns were suggested at all three thresholds [Figure 6], indicating that only one allele is transcribed.

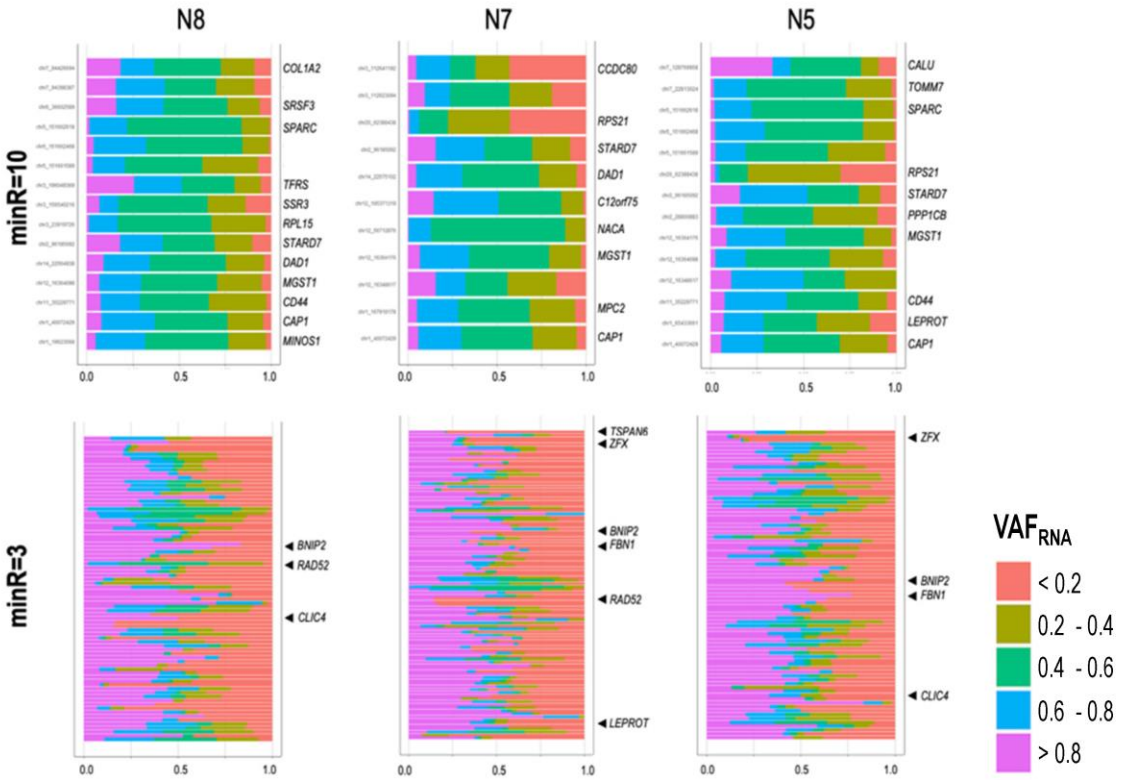


Figure 5. scVAF$_{RNA}$ distribution at positions covered by a minimum of 10 sequencing reads (top), and three sequencing reads (bottom), across more than 1500 cells for genes reported by Borel et al. For the positions with minR = 10, no RME was suggested by the scVAF$_{RNA}$ distribution for autosomal genes (i.e., for the majority of the estimations scVAF$_{RNA}$ values were between 0.2 and 0.8), while positions covered with minR = 3 showed frequent monoallelic signals (scVAF$_{RNA}$ > 0.8 or scVAF$_{RNA}$ < 0.2). As expected, chromosome X shows strong RME patterns (see gene *TSPAN6*) [85].

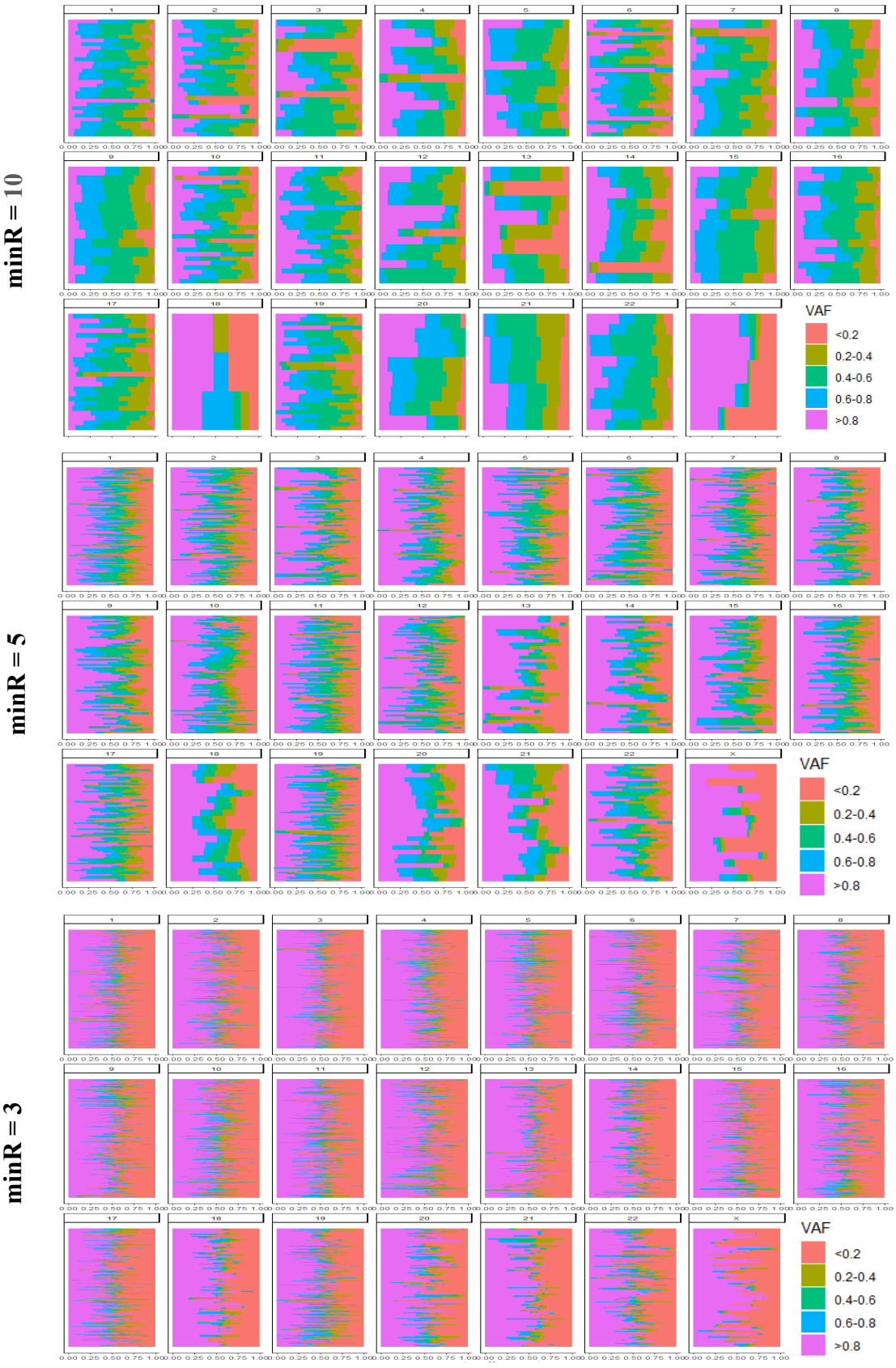

minR = 10
minR = 5
minR = 3
VAF
<0.2
0.2-0.4
0.4-0.6
0.6-0.8
>0.8

Figure 6. Chromosome-wise $scVAF_{RNA}$ estimations (stacked) at positions covered by a minimum of 10 sequencing reads (top), 5 sequencing reads (middle) and 3 sequencing reads (bottom). across more than 1500 cells (top) or in between 10 and 1500 cells (bottom); the sample is N8. For the majority of the positions with minR = 10, no RME was suggested by the $scVAF_{RNA}$ distribution (i.e. multiple $scVAF_{RNA}$ values between 0.4 – 0.6, while positions covered with minR = 3 showed frequent monoallelic signals ($scVAF_{RNA} > 0.8$ or $scVAF_{RNA} < 0.2$. At all thresholds, chromosome X shows strong RME patterns.[85]

4.2. Cell type Identification

The different cell types across all samples were identified using an R package single R (version 1.0.5 [76]), and the different clusters were obtained. The expression profile of each single cell was correlated to expression data derived from: (1) Human Primary Cell Atlas, containing 713 samples representing 38 main cell types and 157 subtypes, and, (2) BluePrint + ENCODE dataset, derived from 259 bulk RNA-seq samples representing 24 main cell types and 43 subtypes. Across the three study samples, four major cell types were identified: adipose cells, erythrocytes, neutrophils, and naïve-B-cells. Adipose cells and erythrocytes were found in all three samples, whereas naïve-B-cells were seen in N5 and N7 and neutrophils in N8 (Figure 7). Adipose cells and erythrocytes are neuroepithelial cells, epithelial cells, and stromal cells consisting of MSC, fibroblasts, chondrocytes, and smooth muscle cells.

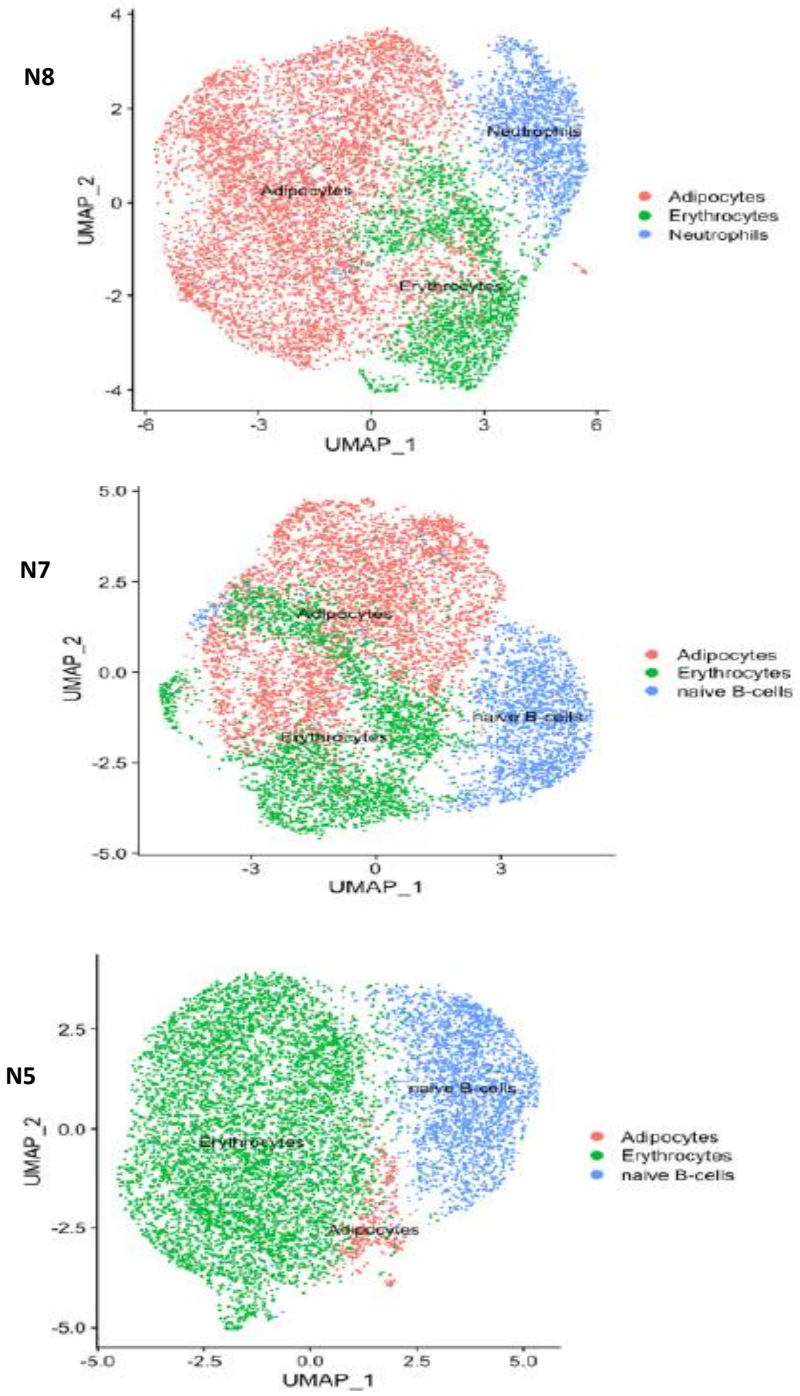


Figure 7. Four major cell types were identified from across the three ADSCs samples; adipose cells, erythrocytes, and neutrophils were identified from N8 while adipose cells, erythrocytes, and naïve B cells were identified from both N5 and N7 samples.

SeattleSeq v.13.00 annotation tool was used to identify cell type specific genes that escaped XCI at SNV loci. These were assessed from across N8 sample and the genes from the different cell types were identified.

4.3. Identification of Escapee genes

Genes subject to XCI were defined as X-chromosome positioned genes showing monoallelic expression (VAF_{RNA} values approximating 0 or 1). Accordingly, XCI Escapee genes were defined as those X-chromosome positioned genes whose VAF_{RNA} deviated from 0 and 1. More precisely, XCI-escapees were defined as genes where the mean VAF_{RNA} estimated across all the cells satisfying the requirements for the analysis was between 0.2 and 0.8. VAF_{RNA} allows for precise allele quantitation and bias in allele mapping may compromise accurate VAF_{RNA} estimation especially if the loci have low number of sequencing reads. To increase the confidence of our observations, allele mapping bias was corrected in the alignments by using the WASP package [70] which is implemented in the latest version of STAR alignment tool [72]. Fully inactivated genes expressed allele ratio equal to 1. Estimation of VAF_{RNA} for all the samples at a minimum of 10 sequencing reads was done across all SNV loci within all the cell types and as expected, random monoallelic expression was observed for X chromosome [Figure 8].

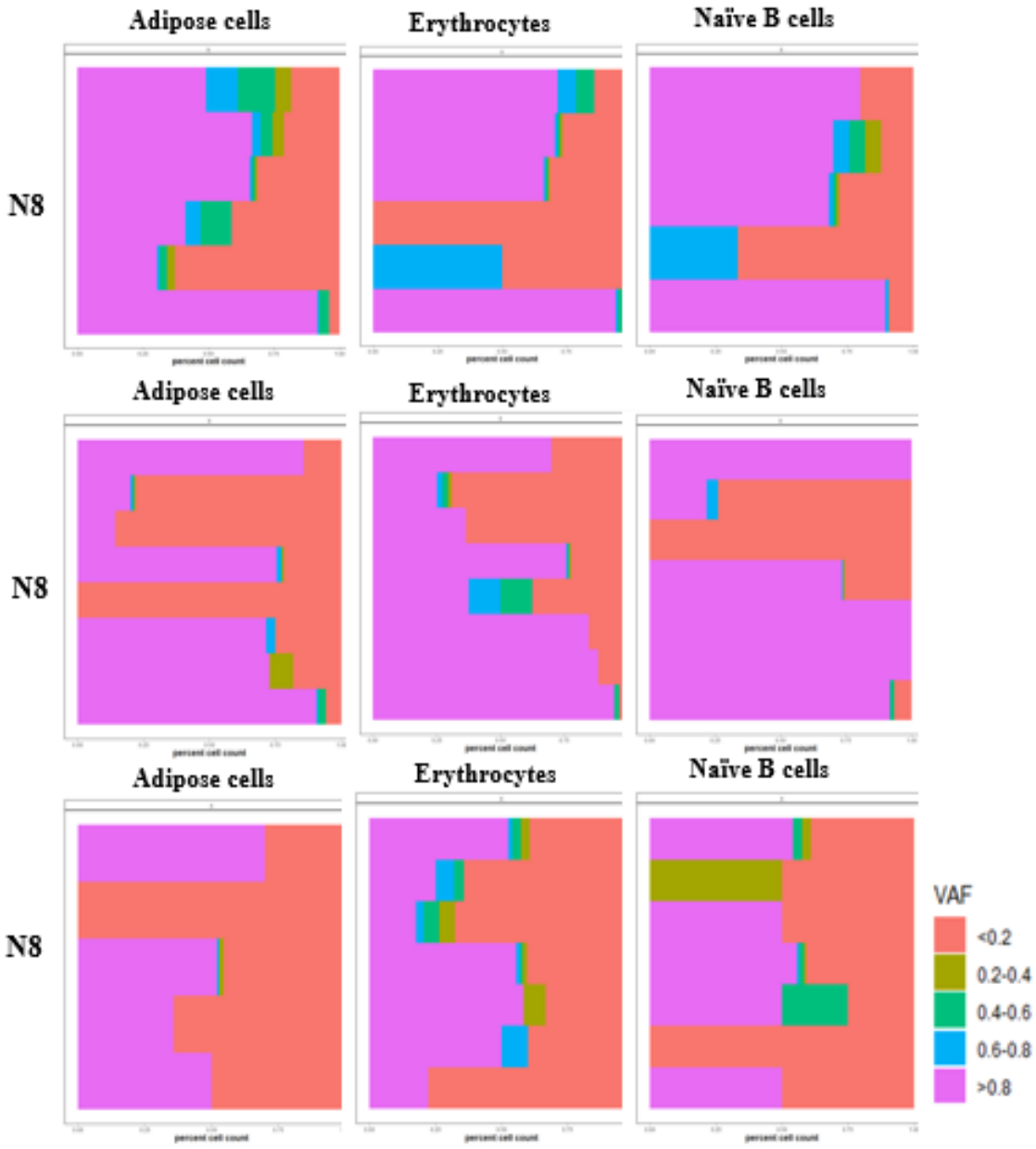


Figure 8; $ScVAF_{RNA}$ estimations on the X chromosome in the different cell types across all 3 samples at positions covered by a minimum of 10 sequencing reads shows RME. Escapees are not identified, rather some genes that do not express either the variant or the reference nucleotide are noted. XCI escapees are identified as those that show biallelic patterns with VAF_{RNA} estimation values between 0.2 and 0.8.

To include a higher number of SNVs in the analysis, further assessment of XCI escapees was carried out on the samples at a minimum of 5 sequencing reads. At minR = 5, the genes that escaped XCI were identified as those who were observed to show biallelic expression with VAF_{RNA} estimation values between 0.2 and 0.8. Across each sample, some genes were observed to escape XCI and were expressed as biallelic in one cell type while the same gene showed RME with complete inactivation in a different cell type [Figure 9]. This shows that the XCI is heterogenous in different cell types as some genes escape XCI in one cell type while the same gene is silenced in another cell type [see PUDP Figure 9].

4.4. Escapee Genes and Clinical Abnormalities

Gene set enrichment analysis (GSEA) suggests that some of XCI escapee genes are associated with certain clinical abnormalities [Figure 10]. This association is seen between six of the escapee genes that are related to Bipolar disorder. ARMCX2 is one of those genes shown to have a significant expression positively correlated with oligodendrocyte density in the brain of patients with bipolar disorder. Gene disease association analysis showed other clinical conditions associated with some of the novel XCI escapees. The escapees and their associated clinical abnormalities are shown in Figure 11.

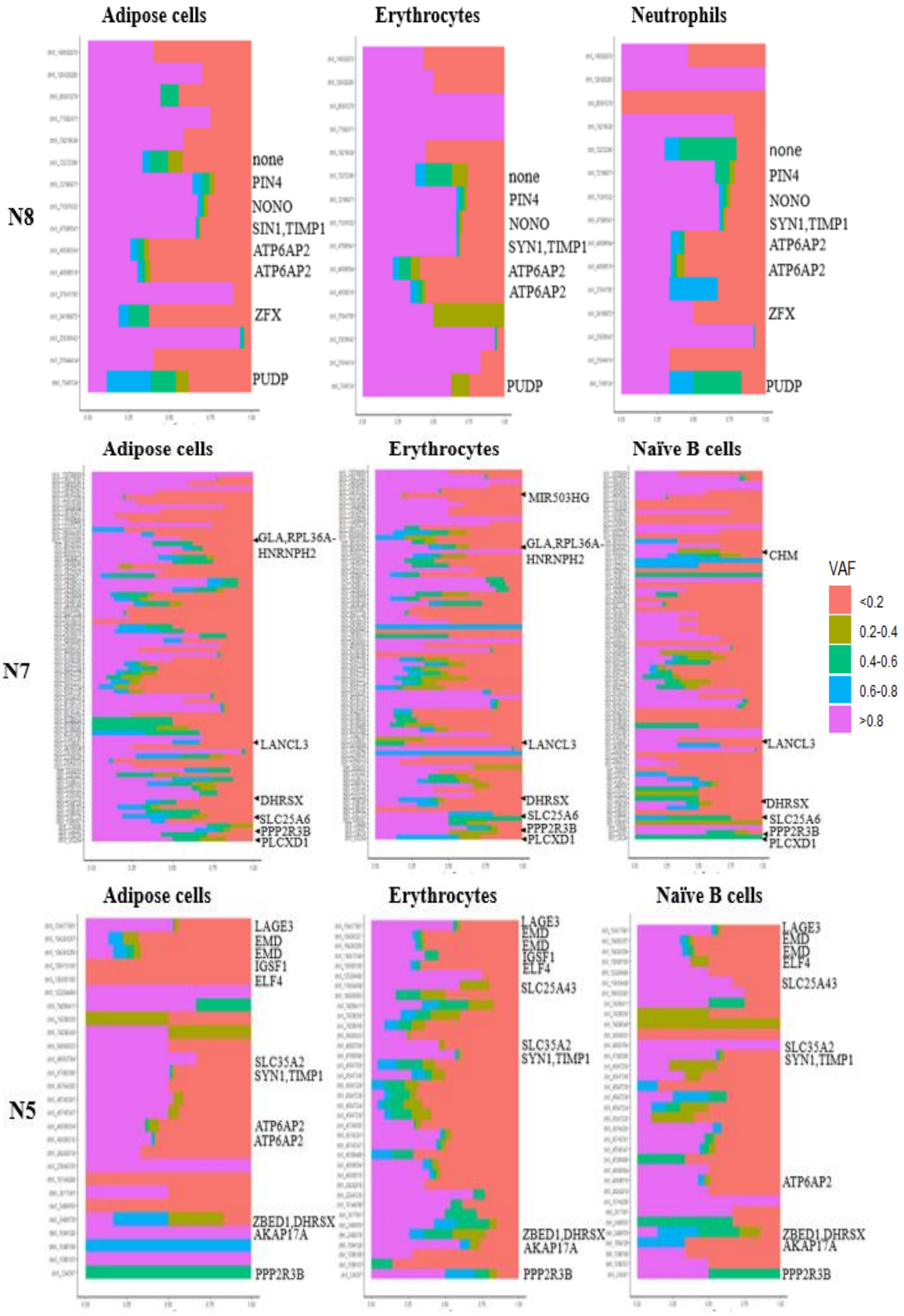
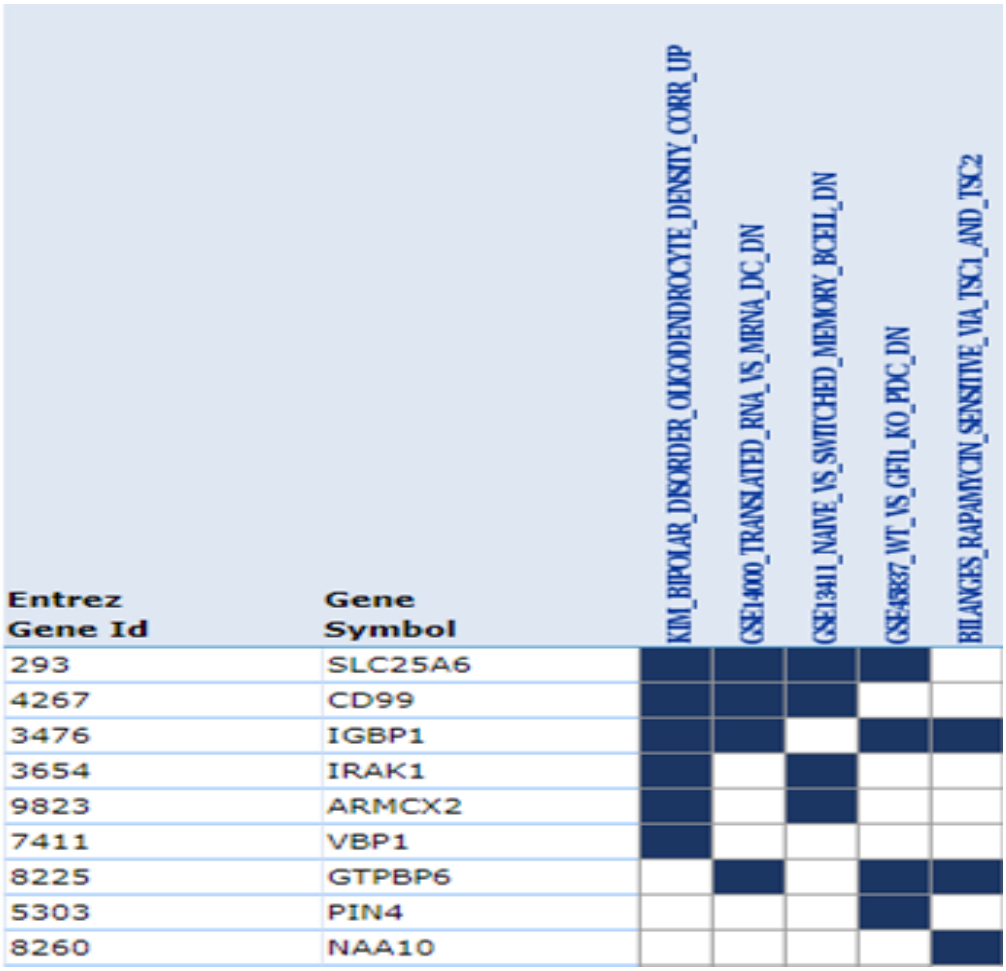

Adipose cells
Erythrocytes
Neutrophils
N8
none
PIN4
NONO
SIN1,TIMP1
ATP6AP2
ATP6AP2
ZFX
PUDP
none
PIN4
NONO
SYN1,TIMP1
ATP6AP2
ATP6AP2
PUDP
none
PIN4
NONO
SYN1,TIMP1
ATP6AP2
ATP6AP2
ZFX
PUDP
Adipose cells
Erythrocytes
Naïve B cells
N7
GLA,RPL36A-
HNRNPH2
LANCL3
DHRSX
SLC25A6
PPP2R3B
PLCXD1
MIR503HG
GLA,RPL36A-
HNRNPH2
LANCL3
DHRSX
SLC25A6
PPP2R3B
PLCXD1
CHM
LANCL3
DHRSX
SLC25A6
PPP2R3B
PLCXD1
VAF
<0.2
0.2-0.4
0.4-0.6
0.6-0.8
>0.8
Adipose cells
Erythrocytes
Naïve B cells
N5
LAGE3
EMD
EMD
IGSF1
ELF4
SLC35A2
SYN1,TIMP1
ATP6AP2
ATP6AP2
ZBED1,DHRSX
AKAP17A
PPP2R3B
LAGE3
EMD
EMD
IGSF1
ELF4
SLC25A43
SLC35A2
SYN1,TIMP1
ZBED1,DHRSX
AKAP17A
PPP2R3B
LAGE3
EMD
EMD
ELF4
SLC25A43
SLC35A2
SYN1,TIMP1
ATP6AP2
ZBED1,DHRSX
AKAP17A
PPP2R3B

Figure 9. $ScVAF_{RNA}$ estimations on the X chromosome in the different cell types across all 3 samples at positions covered by a minimum of 5 sequencing reads shows both RME and biallelic patterns. XCI escapees are identified as those that show biallelic patterns with VAF_{RNA} estimation values between 0.2 and 0.8. Across each sample, some genes were observed to escape XCI and expressed as biallelic in one cell type while the same gene shows RME with complete inactivation in another cell type. With N8 sample, PUDP is shown to escape XCI in adipose cells and neutrophils but not in erythrocytes. ZFX escapes XCI in adipose cells but not in erythrocytes and neutrophils. N7 sample shows PLCXD1 gene as an escapee in the three different cell types while LANCL3 is an escapee in both adipose cells and erythrocytes but not in naïve B cells. Also, in N5 sample, IGSF1 and ELF4 are both inactivated in adipose cells but noted as XCI escapees in erythrocytes.



Entrez Gene Id	Gene Symbol	KIM_BIPOLAR_DISORDER_OLIGODENDROCYTE_DENSITY_CORR_UP	GSE14000_TRANSLATED_RNA_VS_MRNA_DC_DN	GSE13411_NAIVE_VS_SWITCHED_MEMORY_BCELL_DN	GSE45837_WT_VS_GFI1_KO_PDC_DN	BILANGES_RAPAMYCIN_SENSITIVE_VIA_TSC1_AND_TSC2
293	SLC25A6	■	■	■	■	
4267	CD99	■	■	■		
3476	IGBP1	■	■		■	■
3654	IRAK1	■		■		
9823	ARMCX2	■		■		
7411	VBP1	■				
8225	GTPBP6		■		■	■
5303	PIN4				■	
8260	NAA10					■

Figure 10. Gene set enrichment analysis (GSEA) suggests some of XCI escapee genes are associated with certain clinical abnormalities. The relationship between the escapee genes and some of these diseases is identified. The first column shows 6 genes that are related to Bipolar disorder. ARMCX2 is one of those genes shown to have a significant expression which is positively correlated with oligodendrocyte density in layer 4 of the BA9 brain region of patients with Bipolar disorder.

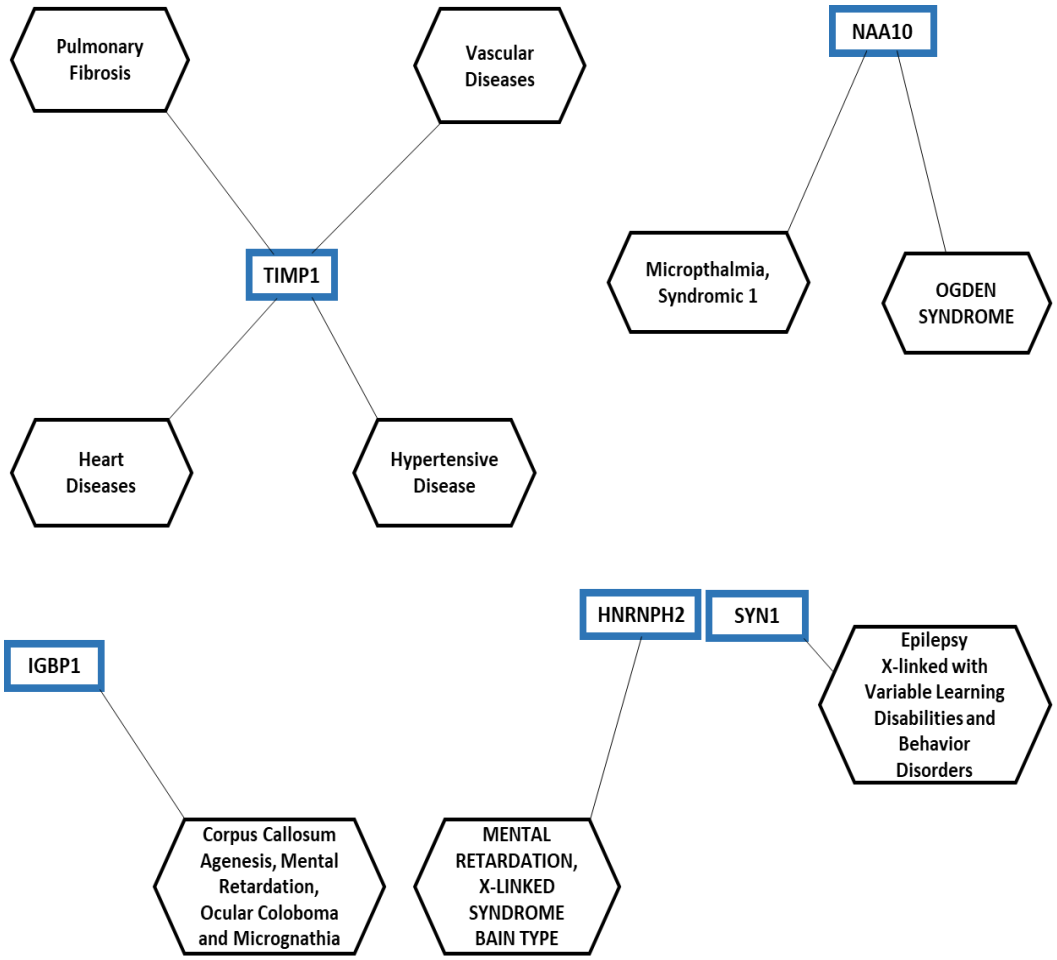


Figure 11. Gene disease association of some of XCI escapee genes with certain clinical abnormalities. There is a link between the escapee genes and the associated diseases as shown; SYN1 is associated with X-linked variable learning disabilities and behavior disorders, HNRNPH2 with mental retardation and x-linked syndrome, IGBP1 with corpus callosum agenesis, mental retardation, ocular coloboma and micrognathia, NAA10 with OGDEN syndrome, and TIMP1 with pulmonary fibrosis, vascular diseases, and heart diseases.

Chapter 5: Discussion and Conclusion

5.1 Discussion

Single cell VAF_{RNA} is a powerful tool useful in identifying novel escapee genes. This study used single cell RNA-seq data sets to identify cell clusters in ADSCs as well as genes that escape X-Chromosome inactivation. Genes that show random monoallelic expression code for proteins with a variety of functions. In neural progenitor cells, these genes have been known to be higher in people at risk for schizophrenia. Copy number variations associated with schizophrenia and autism have been shown to contain a high number of genes that exhibit monoallelic expression when compared to control regions [77]. Therefore, monoallelic gene expression whether genetic or random may play a role or partially explain the variable penetrance of certain diseases that are associated with mutations [78]. In our study, a total of 25 genes were identified as XCI Escapees, with 18 of them previously known to escape XCI. The 7 undescribed escapees (MIR503HG, GTPBP6, ARMCX2, HNRNPH2, RPL36A, IRAKI, and SYN1) have not been previously reported as Escapees in other tissues. Some of the genes noted to escape XCI in our study were documented as inactive in previous studies; ARMCX2 previously known to be an inactivated gene in other tissues by Tukiainen et al 2017 [68] was found to escapee XCI. This could be explained by the fact that XCI has not only been observed to vary from gene to gene, but has been noted to be heterogenous in different tissues and amongst individuals [67].

Some escapee genes have been found to be associated with many clinical abnormalities; ARMCX2 (ALEX1) is one of those genes shown to have a significant expression which is positively correlated with oligodendrocyte density in layer 4 of the BA9 brain region of patients with bipolar disorder [Figure 10]. This gene is required for proper development of the embryo and maintenance of tissue integrity. It is also known to be associated with tumorigenesis in cervical cancer [79]. According to Kurochkin et al, it is known to be expressed differentially in normal tissues and in carcinoma. Its mRNA is lost or markedly reduced in certain tissues including ovarian carcinomas, likely promoting suppression of tumors of epithelial origin. MIC2 (CD99) is associated with many tumors like Ewing's Sarcoma and malignant gliomas where overexpression is known to contribute to oncogenesis [80]. It is involved in many cellular processes including leukocyte transendothelial migration, T cell adhesion, and caspase independent apoptosis. SYN1 a novel escapee was noted to be associated with X-linked variable learning disabilities and behavior disorders [Figure 11]. However, its role and implication in neuro developmental disorder and MECP2 duplication syndrome has not been well documented. According to Signorini et al, overexpression of methyl-CpG binding protein 2 (MECP2) likely results in a state of enhanced oxidative stress [81]**.** Other novel escapees associated with certain clinical abnormalities are shown in Figure 11; HNRNPH2 is linked to mental retardation and x-linked syndrome, IGBP1 is associated with corpus callosum agenesis, mental retardation, ocular coloboma and micrognathia, NAA10 with OGDEN syndrome, and TIMP1 with pulmonary fibrosis, vascular disease, and hypertensive heart diseases.

The findings from this study compares to observations from other studies on XCI. Garieri et al investigated the cellular heterogeneity of X-inactivation in other tissues like fibroblast and lymphoblasts [67] and classified some gene as known XCI escapees while others were undescribed escapees. STK26 was described as an escapee gene in both tissues. Low levels of STK26 expression is implicated in Grave's disease [82] which is an autoimmune disease with abnormal thyroid function resulting in hyperthyroidism. Mutations in XCI escapees are also known to be associated with different pathological conditions. Specifically, mutations in HUWE1 have been found to be associated with individuals diagnosed with Turner-type intellectual disability [MRX17 9MIM 300706)] and another study suggested that HUWE1 plays a role as a tumor suppressor gene in colonic cancer [83]. Patients with turner syndrome are known to be predisposed to developing hypothyroidism [84] a condition where patients produce low levels of the thyroid hormone. A recent study by Dunford et al showed that some escapee genes contribute to cancer sex bias [26] and another study confirmed the status of these XCI escapees suggesting that inactivation profile of some of these genes varies within an individual [67] thereby accounting for the potential difference in cancer predisposition among female individuals.

5.2. Conclusion

This study has provided an approach to assess cell type specific genes that escape XCI from single cell RNA-seq data and identified the different cell types present in adipose-derived mesenchymal stem cells (ADSCs). In addition, it has provided more insight on the use of VAF_{RNA} to determine the allele specific expression of genes and

successfully identified XCI escapee genes in the different cell types. When comparing our findings to existing knowledge, we concluded that since XCI is known to be heterogenous in different tissues, this may explain the disparity in our findings of different escapees in the different cell types. Overall, these results have proffered a better understanding of XCI in the different cell types in ADSCs and is suggestive of a possible influence on phenotypic variance of X-linked single cell disorders and female sex bias in many diseases including cancer.

References

1. Lyon M. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* 1961; 190:372–373.

2. Penny, G., Kay, G., Sheardown, S. *et al.* Requirement for *Xist* in X chromosome inactivation. *Nature* 1996; 379, 131–137.

3. Chow, J., Heard, E. X inactivation and the complexities of silencing a sex chromosome. *Curr Opin Cell Biol* 2009; 21:359–366.

4. Bartolomei, M.S., Ferguson-Smith, A.C. (2011) Mammalian genomic imprinting. Cold Spring Harb Perspect *Biol* 2011; 3a002592.

5. Disteche, C. M. Dosage compensation of the sex chromosomes and autosomes. *Semin. Cell Dev. Biol.* 2016; 56, 9–18.

6. Pinheiro, I., Heard E., X chromosome inactivation: New players in the initiation of gene silencing. F1000 Res 2017; **6**:344.

7. Jonkers, I., Monkhorst, k., Rentmeester, E., Grootegoed, J.A., Grosveld, F., Grosveld, J. Xist RNA is confined to the nuclear territory of the silenced X chromosome throughout the cell cycle. *Mol Cell Biol* 2008; **28**:5583–5594.

8. Brown, C. J., Ballabio, A., Rupert, J.L., Lafrenière, R.G.,Grompe., Tonlorenzi, R., Willard, H.F. A gene from the region of the human X inactivation center is expressed exclusively from the inactive X chromosome. *Nature* 1991; **349**:38–44.

9. Brown, C. J., Hendrich, B. D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence. J., Willard, H.F. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus.*Cell* 1992; **71**:527–542.

10. Ballabio, A., Willard, H.F. Mammalian X-chromosome inactivation and the XIST gene. *Curr Opin Genet Dev* 1992; **2**:439–447.

11. Wang, C.Y., Jegu, T., Chu, H.P., Oh, H. J., Lee, J.T. SMCHD1 merges chromosome compartments and assists formation of super-structures on the inactive X. *Cell* 2018; **174**:406–421.e25.

12. Gendrel, A.V., Heard, E. Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. Annu Rev *Cell Dev Biol* 2014; **30**:561–580.

13. Jeffries, A.R., Perfect, L.W., Ledderose, J., Schalkwyk, L.C., Bray, N.J., Mill, J., Price, J. Stochastic Choice of Allelic Expression in Human Neural Stem Cells. Stem Cells. 2012; 30:1938–1947.

14. Gendrel, A.V., Attia, M., Chen, C.J., Diabangouaya, P., Servant, N., Barillot, E., Heard, E. Developmental dynamics and disease potential of random monoallelic gene expression. Developmental Cell. 2014; 28:366–380

15. Cotton, A.M., Ge, B., Light, N., Adoue, V., Pastinen, T. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. Genome Biol . 2013; **14**:R122.

16. Carrel, L., Willard, H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 2005; **434**:400–404.

17. Yang, F., Babak, T., Shendure. J., Disteche, C.M. Global survey of escape from X inactivation by RNA-sequencing in mouse. Genome Res 2010; **20**:614–622.

18. Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat Genet. 2015; **47**:353–360.

19. Lederer, D., Grisart, B., Digilio, M. C., Benoit, V., Crespin, M., Ghariani, S. C., Maystadt, I., Dallapiccola, B., & Verellen-Dumoulin, C Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. Am J Hum Genet. 2012; **90**:119–124.

20. Miyake, N., Mizuno, S., Okamoto, N., Ohashi, H., Shiina, M., Ogata, K., Tsurusaki, Y., Nakashima, M., Saitsu, H., Niikawa, N., Matsumoto, N. KDM6A point mutations cause Kabuki syndrome. Hum Mutat 2013; **34**:108–110.

21. Gropman, A., Samango-Sprouse, C.A., Neurocognitive variance and neurological underpinnings of the X and Y chromosomal variations. Am J Med Genet C Semin Med Genet 2013; **163C**:35–43.

22. Zhang, Y, *et al.* Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving. Mol Biol Evol (2013) **30**:2588–2601.

23. van Haaften G, *et al.* Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet (2009) **41**:521–523.

24. Grasso, C.S., *et al.* The mutational landscape of lethal castration-resistant prostate cancer. Nature (2012) **487**:239–243.

25. Jones DT, *et al.* Dissecting the genomic complexity underlying medulloblastoma. Nature (2012) **488**:100–105.

26. Dunford, A., *et al.* Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. Nat Genet (2016) **49**:10–16.

27. Borel, C.; Ferreira, P.G.; Santoni, F.; Delaneau, O.; Fort, A.; Popadin, K.Y.; Garieri, M.; Falconnet, E.; Ribaux, P.; Guipponi, M.; et al. Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* **2015**, *96*, 70–80.

28. Larsson, A.J.M.; Johnsson, P.; Hagemann-Jensen, M.; Hartmanis, L.; Faridani, O.R.; Reinius, B.; Segerstolpe, Å.; Rivera, C.M.; Ren, B.; Sandberg, R. Genomic encoding of transcriptional burst kinetics. *Nature* **2019**, *565*, 251–254.

29. Kim, J.K.; Marioni, J.C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* **2013**, *14*, R7.

30. Deng, Q.; Ramsköld, D.; Reinius, B.; Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science.* **2014**, *343*, 193–196.

31. Kim, J.K.; Kolodziejczyk, A.A.; Ilicic, T.; Teichmann, S.A.; Marioni, J.C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **2015**, *6*, 8687.

32. Moreira de Mello, J.C.; Fernandes, G.R.; Vibranovski, M.D.; Pereira, L.V. Early X chromosome inactivation during human preimplantation development revealed by single-cell RNA-sequencing. *Sci Rep.* **2017**, *7*, 10794.

33. D'Antonio-Chronowska, A.; Donovan, M.K.R.; Greenwald, W.W.; Nguyen, J.P.; Fujita, K.; Hashem, S.; Matsui, H.; Soncin, F.; Parast, M.; Ward, M.C.; et al. Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. *Stem Cell Reports.* **2019**, *13*, 924–938.

34. Poirion, O.; Zhu, X.; Ching, T.; Garmire, L.X. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nat. Commun.* **2018**, *9*, 4892.

35. Vu, T.N.; Nguyen, H.N.; Calza, S.; Kalari, K.R.; Wang, L.; Pawitan, Y. Cell-level somatic mutation detection from single-cell RNA-sequencing. *Bioinformatics.* **2019**.

36. Rodriguez-Meira, A.; Buck, G.; Clark, S.A.; Povinelli, B.J.; Alcolea, V.; Louka, E.; McGowan, S.; Hamblin, A.; Sousos, N.; Barkas, N.; et al. Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Mol. Cell.* **2019**, *73*, 1292–1305.

37. Reinius, B.; Mold, J.E.; Ramsköld, D.; Deng, Q.; Johnsson, P.; Michaëlsson, J.; Frisén, J.; Sandberg, R. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* **2016**, *48*, 1430–1435.

38. Lee, M.P. Understanding Cancer Through the Lens of Epigenetic Inheritance, Allele-Specific Gene Expression, and High-Throughput Technology. *Front Oncol.* **2019**, *9*, 794.

39. Griffiths, J.A.; Scialdone, A.; Marioni, J.C. Mosaic autosomal aneuploidies are detectable from single-cell RNAseq data. *BMC Genomics.* **2017**, *18*, 904.

40. Edsgärd, D.; Reinius, B.; Sandberg, R. Scphaser: Haplotype inference using single-cell RNA-seq data. *Bioinformatics.* **2016**, *32*, 3038–3040

41. Huang, Y.; McCarthy, D.J.; Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **2019**, *20*, 273.

42. Xu, J.; Falconer, C.; Nguyen, Q.; Crawford, J.; McKinnon, B.D.; Mortlock, S.; Senabouth, A.; Andersen, S.; Chiu, H.S.; Jiang, L.; et al. Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol.* **2019**, *20*, 290.

43. Kang, H.M.; Subramaniam, M.; Targ, S.; Nguyen, M.; Maliskova, L.; McCarthy, E.; Wan, E.; Wong, S.; Byrnes, L.; Lanata, C.M.; et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **2018**, *36*, 89–94.

44. van der Wijst, M.G.; Brugge, H.; de Vries, D.H.; Deelen, P.; Swertz, M.A.; Franke, L. Single-cell RNA sequencing identifies cell type-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **2018**, *50*, 493–497.

45. Chen, G., Schell, J. P., Benitez, J. A., Petropoulos, S., Yilmaz, M., Reinius, B., Alekseenko, Z., Shi, L., Hedlund, E., Lanner, F., Sandberg, R., & Deng, Q. Single-cell

analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res*. 2016;26(10):1342–1354.

46. Liu, X.; Xiang, Q.; Xu, F.; Huang, J.; Yu, N.; Zhang, Q.; Long, X.; Zhou, Z. Single-cell RNA-seq of cultured human adipose-derived mesenchymal stem cells. *Sci. Data.* **2019**, *6*, 190031.

47. Fraser, J. K., Zhu, M., Wulur, I. & Alfonso, Z. Adipose-derived stem cells. *Methods Mol Biol* 2008; 59–67 (2008).

48. Fraser, J. K., Wulur, I., Alfonso, Z. & Hedrick, M. H. Fat tissue: an underappreciated source of stem cells for biotechnology. *Trends Biotechnol* 2006; **24**, 150–154.

49. Ma, T. et al. A brief review: adipose-derived stem cells and their therapeutic potential in cardiovascular diseases. *Stem Cell Res Ther* 2017; **8**, 124.

50. Naderi, N. et al. The regenerative role of adipose-derived stem cells (ADSC) in plastic and reconstructive surgery. *Int Wound J* 2017; **14**, 112–124.

51. Wen, L., Tang, F. Single-cell sequencing in stem cell biology. *Genome Biol* 2016; **17**, 71.

52. Kulkarni, A.; Anderson, A.G.; Merullo, D.P.; Konopka, G. Beyond bulk: A review of single cell transcriptomics methodologies and applications. *Curr Opin Biotechnol.* **2019**,*58*, 129–136.

53. Stuart, T.; Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **2019**, *20*, 257–272.

54. Zafar, H.; Wang, Y.; Nakhleh, L.; Navin, N.; Chen, K. Monovar: Single-nucleotide variant detection in single cells. *Nat. Methods.* **2016**, *13*, 505–507.

55. Schnepp, P.M.; Chen, M.; Keller, E.T.; Zhou, X. SNV identification from single-cell RNA sequencing data. *Hum. Mol. Genet.* **2019**, *28*, 3569–3583.

56. Dong, M.; Jiang, Y. Single-Cell Allele-Specific Gene Expression Analysis. *Methods Mol. Biol.* **2019**, *1935*, 155–174.

57. Liu, F.; Zhang, Y.; Zhang, L.; Li, Z.; Fang, Q.; Gao, R.; Zhang, Z. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol* **2019**, *19*, 242.

58. Horvath, A.; Pakala, S.B.; Mudvari, P.; Reddy, S.D.; Ohshiro, K.S.; Pires, R.; Fuqua, S.A.; Toi, M.; Costa, L.; Nair, S.S.; et al. Novel insights into breast cancer genetic variance through RNA sequencing. *Sci Rep.* **2013**, *3*, 2256.

59. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to

high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* **2013**, *43*.

60. Deelen, P.; Zhernakova, D.V.; de Haan, M.; van der Sijde, M.; Bonder, M.J.; Karjalainen, J.; van der Velde, K.J.; Abbott, K.M.; Fu, J.; Wijmenga, C.; et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **2015**, *7*, 30.

61. Kravitz, S.N.; Gregg, C. New subtypes of allele-specific epigenetic effects: Implications for brain development, function and disease. *Curr Opin Neurobiol.* **2019**, *59*, 69–78.

62. Spurr, L.; Alomran, N.; Bousounis, P.; Reece-Stremtan, D.; Prashant, N.M.; Liu, H.; Słowiński, P.; Li, M.; Zhang, Q.; Sein, J.; et al. ReQTL: Identifying correlations between expressed SNVs and gene expression using RNA-sequencing data. *Bioinformatics.* **2019**.

63. Sein, J.; Spurr, L.; Bousounis, P.; Prashant, N.M.; Liu, H.; Alomran, N.; Bernot, J.; Ibeawuchi, H.; Reece-Stremtan, D.; Horvath, A. RsQTL: Correlation of expressed SNVs with splicing using RNA-sequencing data. *Bioinformatics* **2019**. Under Review.

64. Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S., & Brown, C. J. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Human molecular genetics*, 2015; *24*(6), 1528–1539.

65. Yasukochi, Y, *et al.* X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils. Proc Natl Acad Sci USA 2010; **107**:3704–3709.

66. Sudbrak R, et al. X chromosome-specific cDNA arrays: Identification of genes that escape from X-inactivation and other applications. Hum Mol Genet. 2001; 10:77–83.

67. Garieri, M., Stamoulis, G., Blanc, X., Falconnet, E., Ribaux, P., Borel, C., Santoni, F., & Antonarakis, S. E. Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proceedings of the National Academy of Sciences of the United States of America*, 2018; *115*(51), 13015–13020.

68. Tukiainen, T., Villani, A. C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B. B., Castel, S. E., Karczewski, K. J., Aguet, F., Byrnes, A., GTEx Consortium, Laboratory, Data Analysis &Coordinating Cotton AM, et al. (2015) Landscape of DNA methylation on the X chromosome re-flects CpG density, functional chromatin state and X-chromosome inactivation. Hum Mol Genet 24:1528–1539.

69. Berletch, J.B., Yang, F., Xu, J., Carrel, L., Disteche, C.M. Genes that escape from X inactivation. Hum Genet 2011; **130**:237–245.

70. van de Geijn, B.; McVicker, G.; Gilad, Y.; Pritchard, J.K. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **2015**, *12*, 1061–1063.

71. Smith, T.; Heger, A.; Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **2017**, *27*, 491–499.

72. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21.

73. Liao, Y.; Smyth, G.K.; Shi, W. Feature counts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930.

74. Butler, A.; Hoffman, P.; Smibert, P.; Papalexi, E.; Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **2018**, *36*, 411–420.

75. Movassagh, M.; Alomran, N.; Mudvari, P.; Dede, M.; Dede, C.; Kowsari, K.; Restrepo, P.; Cauley, E.; Bahl, S.; Li, M.; et al. RNA2DNAlign: Nucleotide resolution allele asymmetries through quantitative assessment of RNA and DNA paired sequencing data. *Nucleic Acids Res.* **2016**, *44*, e161.

76. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M. “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.” *Nat. Immunol.*, 2019; **20**, 163-172.

77. Jeffries, A. R., Collier, D. A., Vassos, E., Curran, S., Ogilvie, C. M., & Price, J. Random or stochastic monoallelic expressed genes are enriched for neurodevelopmental disorder candidate genes. *PloS one*, 2013; *8*(12), e85093.

78. Gendrel, A.V., Attia M., Chen C.J., Diabangouaya P., Servant, N., Barillot E., Heard, E Developmental dynamics and disease potential of random monoallelic gene expression. Developmental Cell. 2014; 28:366.

79. Kurochkin, I. V., Yonemitsu, N., Funahashi, S., Nomura, H. ALEX1, a novel human armadillo repeat protein that is expressed differentially in normal tissues and carcinomas. Biochem. Biophys. Res. Commun. 2001; 280: 340-347.

80. Pasello, M., Manara, M. C., & Scotlandi, K. CD99 at the crossroads of physiology and pathology. *Journal of cell communication and signaling*, 2018; *12*(1), 55–68.

81. Signorini, C., De Felice, C., Leoncini, S., Moller, R.S., Zollo, G., Buoni, S., Cortelazzo A., Hayek, J. MECP2 duplication syndrome: Evidence of enhanced oxidative stress. A comparison with Rett syndrome *PLoS ONE*, 2016; 11 (3), art. no. e0150101

82. Guo, A., Tan, Y., Liu, C., & Zheng, X. MST-4 and TRAF-6 expression in the peripheral blood mononuclear cells of patients with Graves' disease and its significance. *BMC endocrine disorders*, 2017; *17*(1), 11.

83. Myant, K. B., Cammareri, P., Hodder, M. C., Wills, J., Von Kriegsheim, A., Győrffy, B., Rashid, M., Polo, S., Maspero, E., Vaughan, L., Gurung, B., Barry, E., Malliri, A., Camargo, F., Adams, D. J., Iavarone, A., Lasorella, A., & Sansom, O. J. HUWE1 is a critical colonic tumour suppressor gene that prevents MYC signalling, DNA damage accumulation and tumour initiation. *EMBO molecular medicine*, 2017; *9*(2), 181–197.

84. El-Mansoury M, et al. Hypothyroidism is common in turner syndrome: Results of a five-year follow-up. J Clin Endocrinol Metab. 2005; 90:2131–2135.

85. M, P. N., Liu, H., Bousounis, P., Spurr, L., Alomran, N., Ibeawuchi, H., Sein, J., Reece-Stremtan, D., Horvath, A. Estimating the Allele-Specific Expression of SNVs from 10× Genomics Single-Cell RNA-Sequencing Data. *Genes*, 2020;*11*(3), 240.