UPPSALA
UNIVERSITET

Single Cell Methods and Cell Hashing for High Throughput Drug Screens

Alva Annett

Degree project in bioinformatics, 2021
Examensarbete i bioinformatik 30 hp till masterexamen, 2021
Biology Education Centre and Department of Medical Sciences, Uppsala University
Supervisor: Jessica Nordlund

Abstract

Acute lymphoblastic leukemia (ALL) is one of the most prevalent childhood cancers and even if survival rates have improved over the last decades problems relating to relapse and drug resistance persist. One way to gain better understanding of the mechanisms underlying these challenges is with high throughput drug screens where individual cells from patient samples are sequenced before and after drug exposure. To be able to process many biological samples, cells and treatments in parallel a cell-hashing approach will be employed to mark individual cells exposed to different conditions. Cell-hashing is the addition of treatment-specific barcodes that allow pooling of treatments into one sequencing library, which brings down the cost and increases throughput. Before moving on to patient samples a few pilot studies were performed to ensure that the cell protocols and the cell-hashing barcodes work as intended. A so called "barnyard" experiment was used to validate the integrity of the barcodes. Human and horse cells were barcoded, mixed at a 1:1 ratio and sequenced by single-cell RNA-seq using the protocol from 10x Genomics. The cells were then evaluated to see if they could be correctly identified by their barcode and by the species. The single-cell assay for transposase active chromatin from 10x genomics was tested on two cell lines and a novel library preparation method for single-cell whole genome bisulfite sequencing was tested on a cell line and liver cells from a patient sample. In addition to testing the methods, analysis pipelines were built for the different data types that will be used in the data analysis of future drug screen experiments. While there were several technical problems with the cell-hashing barcodes, the different single-cell methods tested were successful. The development of the analysis pipelines and the knowledge gained from the barcoding experiment will be useful as the project moves forward.

Taggade celler för storskaliga läkemedelsscreens

Populärvetenksaplig sammanfattning
Alva Annett

Akut lymfatisk leukemi (ALL) är en form av blodcancer som ofta drabbar barn. Under de senaste decennierna har prognosen för dessa patienter blivit allt bättre men många råkar fortfarande ut för återfall och resistens mot den behandling de ges. För att undersöka de mekanismer som leder till återfall och resistens kommer blodceller från patienter med ALL testas mot ett stort antal läkemedel och andra substanser i en så kallad storskalig läkemedels screen. För att ta reda på vad som händer i cellerna när de utsätts för olika substanser sekvenserar man deras genetiska innehåll innan och efter behandling. Detta ger information om vilka genetiska varianter och mönster som leder till att celler överlever eller inte. Dessa varianter och mönster kan sedan användas för att dels öka förståelsen för hur ALL fungerar och dels för att individualisera behandling för framtida patienter. För att på ett tids- och kostnadseffektivt sätt kunna utföra läkemedels screens för ett stort antal patienter, krävs det att de celler som behandlats med olika läkemedel kan samlas ihop och processeras tillsammans. Detta kan åstadkommas genom att märka in cellerna med unika ”taggar” som visar vilken behandling de fått. Dessa taggar är DNA sekvenser som läses av tillsammans med cellens övriga genetiska innehåll. Detta gör att både cellens genetiska profil och den behandling cellen genomgått kan identifieras. På detta sätt kan man alltså processeras celler som genomgått olika behandlingar tillsammans.

För att kontrollera att taggarna fungerar på korrekt sätt har vi genomfört en pilot studie. Taggarn utvärderades med hjälp av ett så kallat ladugårds experiment där celler från två olika arter blandas. Vi använde celler från häst och celler från människa som märkts in med varsin tagg. Dessa celler sekvenserades sedan och baserat på det gener som identifierades klassificerades de som antingen häst- eller människoceller. Cellerna klassificerades också utifrån vilken tagg de innehöll och dessa två klassificeringar jämfördes sedan för att avgöra om taggarna har fungerat. På grund av flera tekniska problem med experimentet överlappande de två klassificeringarna för väldigt få celler. För det första verkar det som att häst cellerna dött och gått sönder i något steg. Detta betyder att genetiskt material från häst spridits och blandats med människo cellerna. På grund av detta var det svårt att avgöra vilka celler som kom från vilken art. För det andra verkar det som att den tagg som använts för människo cellerna inte absorberats på korrekt sätt. På grund av detta var det inte möjligt att avgöra vilken tagg många av cellerna hade märkts med. Resultaten från den första pilotstudien var inte tillräckliga för att avgöra om taggarna fungerar på korrekt sätt. Trots detta har resultaten varit användbara för att planera de nästkommande stegen i projektet. Den data som genererats har även använts för att testa och utveckla de metoder för dataanalys som kommer att användas i projektet framöver.

Examensarbete i bioinformatik 30 hp till masterexamen, 2021
Institutionen för biologisk grundutbildning och Institutionen för medicinska vetenskaper
Handledare: Jessica Nordlund

Tableof Contents

Abbreviations

ALL	acute lymphoblastic leukemia
ATAC	assay for transposase-accessible chromatin
CB	cell barcode
DNA	deoxyribonucleic acid
GEMs	gel beads in emulsion
GSEA	gene set enrichment analysis
HPC	high performance computing
IQR	interquartile range
LMO	lipid-modified oligonucleotide
NGS	next generation sequencing
QC	quality control
RNA	ribonucleic acid
SC	single-cell
SNN	shared nearest neighbor
SPLAT	splinted ligation adapter tagging
SVD	singular value decomposition
TF-IDF	term frequency inverse document frequency
TSS	transcription start site
t-SNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier
WGBS	whole genome bisulfite sequencing

Introduction

Applications And Opportunities for Analysis of Single Leukemia Cells

Acute lymphoblastic leukemia (ALL) originates in the bone marrow from proliferative leukocytes stuck at an early stage of maturation (Malard & Mohty 2020). The malignant cells are derived from either the B-cell lineage or more rarely the T-cell lineage. ALL is currently one of the most prevalent childhood cancers. Even though overall survival rates have greatly improved in the last decades the problem of resistance, adverse drug reactions and relapse leading to reduced survival rate persists (Iacobucci & Mullighan 2017). There are several ALL subtypes characterized by different chromosomal rearrangements and secondary mutations such as copy number variations and single nucleotide polymorphisms. In addition to this, ALL is characterized by heterogeneity in gene expression and epigenetic markers. The subtypes are closely linked to prognosis and a better understanding of these subtypes is vital for future diagnosis and treatment of ALL (Liu *et al.* 2016, Iacobucci & Mullighan 2017).

Understanding the underlying molecular mechanisms of drug sensitivity, resistance and relapse is an important key in improving survival rates and quality of life for ALL patients. One way this can be achieved is through high throughput drug screens where cells obtained from patients are exposed to a wide range of concentrations and combinations of drugs (Pauli *et al.* 2017). High throughput drug screens have several applications. Firstly, they can be used to discover new treatments by testing many substances, combinations, and concentrations at the same time. The development in sequencing methods makes this feasible in terms of time and money. Secondly, they can be used in a clinical setting to identify which drugs different subgroups of cells are sensitive to. It has been shown that the response to *ex vivo* treatment correlates well with patient outcome (Pauli *et al.* 2017). This information could be used to guide treatment at an early stage of the disease by screening patient samples. Thirdly, investigating how cancer cells from a range of patients react to different treatments will give us an overall better understanding of the disease.

Single-cell sequencing is a method where the ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) form individual cells are tagged with cell specific barcodes. The RNA or DNA is then processed in the same way as bulk sequencing but each read can be traced back to the original cell (Shapiro *et al.* 2013). Single-cell sequencing has greatly increased the resolution at which a cell population can be understood. Rare cell types that would be masked in bulk sequencing can be identified and the heterogeneity between cells investigated. In cancer research this has the potential to improve our understanding of why certain cell populations survive treatment and subsequently lead to relapse and may allow for a more detailed investigation of clonal evolution and how specific subgroups of cells respond to different drugs (Hwang *et al.* 2018). Single-cell sequencing can be used in combination with high throughput drug screens to give us an even better understanding of the surviving cells within a sample. In the future this will hopefully help us individualize treatment for patients and reduce the risk of relapse and drug resistance.

To avoid having to prepare one sequencing library per treatment and biological sample in a high throughput drug screen, a method called cell-hashing can be employed. Cell-hashing is the addition of condition-specific barcodes that are introduced prior to single-cell isolation in a microfluidics device where the cell-specific barcodes are added. This gives an extra layer of information which enables pooling of many cells and conditions at an early stage of library preparation. This greatly reduces the cost of library preparation and saves time and reagents. Batch effects due to differences in handling individual treatments and samples can also be reduced by early pooling of treatments. There have been several cell-hashing methods proposed such as CITE-seq, which uses antibody tagged oligonucleotides or demuxlet, which uses the natural genetic variation between samples to identify which cells the reads come from (Stoeckius *et al.* 2017, Kang *et al.* 2018). However, most of these methods still face substantial challenges with throughput and cost. The cell-hashing method used in this project is built on the MULTI-seq protocol (McGinnis *et al.* 2019). The MULTI-seq method uses lipid-modified oligonucleotides (LMOs) hybridized with DNA barcodes, which have been introduced as an affordable alternative for cell-hashing. The LMOs are introduced into the cell membrane of living cells and the cell-hashing barcodes are subsequently linked to the cell specific barcodes during reverse transcription. Before sequencing, size exclusion is used to separate the cDNA and the cell-hashing barcodes during the library preparation step. At this stage both the cDNA and the cell-hashing barcodes are linked to the cell barcodes. The two libraries are sequenced using next generation sequencing (NGS) and can then be used to identify both individual cells and their treatment (McGinnis *et al.* 2019). A schematic overview of the reads from the two libraries can be seen in Figure 1.

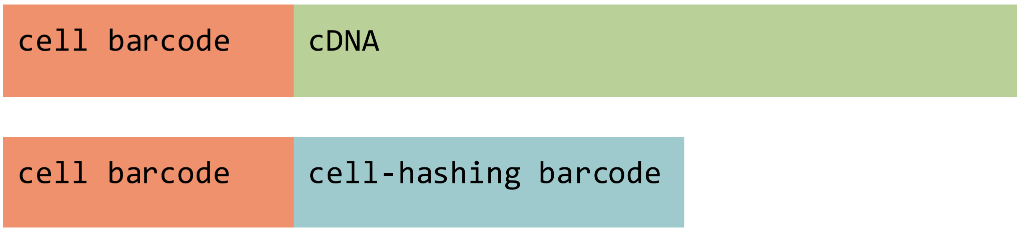

Figure 1. *Overview of two different types of libraries constructed when using MULTI-seq barcodes. Cell barcodes will be linked both to cDNA and to the MULTI-seq cell-hashing barcodes. The information can be used to assign cell barcodes to the sample or treatment of origin.*

Droplet Based Technologies for Single-cell Library Preparation

10x Genomics provides a microfluidics platform for single-cell RNA sequencing and single-cell assay for transposase-accessible chromatin (scATAC-seq). RNA sequencing is used to measure at what level each gene is expressed and ATAC sequencing is used to measure which genes are accessible to the transposase and therefore could be transcribed. For scRNA-seq a diluted solution of cells is loaded on the platform together with enzymes, gel beads and partitioning oil to create droplets or GEMs (gel beads in emulsion) containing single-cells and the chemicals needed for transcription of cDNA. When the GEMs have been created the GEMs are dissolved releasing primers. Any cells present in the droplet will be lysed and the production of cDNA from RNA starts. The primers contain a cell barcode and unique molecular identifier (UMI) which makes it possible to track each read both to cell and transcript of origin. The GEMs are then broken, and the cDNA purified and amplified to create a library that can be sequenced (Zheng *et al.* 2017). The scATAC-seq protocol is similar, except that the nuclei are isolated and exposed to transposase in bulk. Exposed chromatin is cut and tagged for further processing. The purified nuclei are then loaded on the microfluidics device in the same way as described above to yield a DNA library for sequencing (Satpathy *et al.* 2019).

Single-cell Whole Genome Bisulfite Sequencing

Methylated cytosines are an important epigenetic marker involved in cellular processes such as gene regulation. The methylation patterns of cells can be investigated by whole genome bisulfite sequencing (WGBS) where unmethylated cytosines converted into uracils. When the genome is subsequently sequenced the methylation state of cytosines can be identified at a single base resolution. In recent years, the interest in single-cell epigenomics has grown. Moving from bulk sequencing to single-cell allows for identification of rare cell types and a better understanding of the heterogeneity of methylation patterns within a cell population such as a tumor (Wen & Tang 2018). Many of the methods available for scWGBS are both low throughput, low coverage, and high cost, which limits the widespread use. To overcome some of these problems a new method for scWGBS library preparation based on the previously published method Splinted Ligation Adapter Tagging (SPLAT) (Raine *et al.* 2017) has been developed. Single-cell SPLAT is based on using splinted ligation to barcode individual cells in a 384-well plate format, which are subsequently pooled, and a single library is prepared in bulk (Raine et al, in preparation). This approach is a promising alternative to achieve high coverage scWGBS data at a low cost.

As single-cell technologies have increased in accessibility, so has the need for analysis methods specifically designed for single-cell WGBS. Single-cell WGBS often results in a very sparse data which require computational methods that can deal with both missing data and sequencing errors (P. E. de Souza *et al.* 2020). A number of non-probabilistic methods for clustering of scWGBS data have been proposed such as using Euclidean distance, hierarchical clustering and genome wide pairwise dissimilarity (PDclust) (Smallwood *et al.* 2014, Angermueller *et al.* 2016, Hui *et al.* 2018). These methods are however not able to both deal with missing data and cluster the cells at the same time. In their article from 2020 de Souza *et al.* propose a probabilistic framework for clustering of single-cell whole genome bisulfite sequencing data called Epiclomal. Epiclomal uses mixed hierarchical Bernoulli distributions trained by a variational Bayesian algorithm to infer both the true hidden state for each CpG site and cluster assignments for each cell (P. E. de Souza *et al.* 2020).

Workflow managements systems

Workflow management systems are used to automate analysis workflows and ensure reproducibility. They function by running a specified set of tasks in a certain order without human interference. There are several systems available for scientific computing such as Nextflow and Snakemake, with different pros and cons (Köster & Rahmann 2012, Di

Tommaso *et al.* 2017). The main differences between the systems are which language they are written in, how the user interacts with the system (command line or graphical interface), how readable the code is and how flexible they are (Mölder *et al.* 2021).

Snakemake has some advantages compared to other systems. It is run through the command line and is easily integrated with job scheduling systems such as SLURM on high performance computing (HPC) clusters. It is written in Python which is one of the most used programing languages for bioinformatics and the code is highly readable for someone not familiar with either Python or Snakemake. It is flexible and any code that can be run in the command line can be run by Snakemake. The Snakemake script also functions as a regular Python script which means that any Python code can be easily executed. Snakemake allows the user to specify the required software versions or environments. The pipeline is defined as a set of rules with input files and output files and the rules are linked together by Snakemake based on the names of the output files. The pipeline can be generalized with variables and wildcards and a 'yaml' type configuration file where input parameters can be easily changed. This means that a pipeline can be reused for the same type of analysis by just changing parameters and sample names in the configuration file (Köster & Rahmann 2012).

Purpose of Thesis

The main goals of the thesis project were to test software and set up analysis pipelines for various single cell data types including scRNA-seq, scATAC-seq and scWGBS. The pipelines were developed using Snakemake and will be used in future single cell drug screen studies of ALL patients. In addition to this the novel protocol for cell hashing called MULTI-seq was tested and evaluated. This protocol will also be used in future drug screens.

Materials and Methods

scRNA-seq and MULTI-seq Barcodes

To validate that the MULTI-seq barcodes work as intended a pilot study was performed using the REH cell line, which is a commercially available cell line derived from a B-cell ALL patient (Rosenfeld *et al.* 1977) and primary horse cells. A "barnyard" experiment was conducted with a 1:1 mix of horse cells and human cells from the REH cell line. The REH cells and the horse cells were tagged with MULTI-seq barcodes. If the REH cells and the horse cells can be identified both by the barcode and by which genome the reads map to the barcodes work. The barnyard experiment can also be used to evaluate the multiplet rate. If both human and horse genes are expressed in a cell that would indicate that the droplet contained more than one cell.

Ampules of 1-5 million cells were thawed, washed, and subsequently marked with different barcodes (Table1) according to the protocol outlined by McGinnis *et al.* The resulting barcoded cells were mixed in a 1:1 ratio and a scRNA-seq library was made using the 10x Genomics platform for single-cell RNA sequencing (v3) according to the manufacturer's specifications. Two libraries were created, first the gene expression library containing the 3' sequences from the expressed genes, and second a MULTI-seq library containing the sample barcodes. The two sequencing libraries were quality controlled on an Agilent TapeStation, pooled in a 95/5 ratio for GEX/Barcode libraries, loaded on an Illumina NovaSeq6000 sequencer and sequenced on a SP flowcell according to the following sequencing recipe: Read 1: 28bp, I7 index: 8bp, I5 index: 0 bp, Read 2: 98bp at the SNP&SEQ Technology Platform, National Genomics Infrastructure at SciLifeLab.

The sequencing resulted in two FASTQ files, one contacting the GEX reads and one containing the MULTI-seq barcode reads. The two files were pre-processed with the Cellranger software from 10x Genomics (v 5.0.1). The barcode sequences can be seen in Table1. A custom reference package was built using genomes and annotation files from Ensembl for GRCh38 and EquCab3. The function Cellranger mkgtf was used to filter the gtf files for protein coding transcripts of interest. The function Cellranger mkref was then used to create a joint indexed reference genome for human and horse. Cellranger count was run for the two libraries to perform trimming, alignment to the reference genome, filtering based on quality metrics, cell calling and identification of the MULTI-seq barcodes. Cellranger outputs count matrices for the gene expression library and the barcode library and classifies cells as doublets or singlets based on the proportion of reads mapped to either genome. The 10th percentile of barcodes where either UMI count for horse or human transcripts exceeds the other is used as the threshold for classifying the cell as that cell type.

Table1. *The MULTI-seq barcode names and sequences which were used for the REH sample and horse cells.*

Cell Type	Barcode Name	Barcode Sequence
REH	Barcode 1	GGAGAAGA
Horse	Barcode 2	CCAACCGG

Two different methods were tested for demultiplexing the cell type from the barcode library. Firstly, the R package Seurat (v4.0) and secondly the R package deMULTIplex (Satija *et al.* 2015, McGinnis *et al.* 2019). The results were similar for the two methods and the barcode assignment from Seurat was used in the downstream analysis. Demultiplexing with Seurat was done in two steps. The barcode library was first normalised by centred log-ratio transformation and then the Seurat function *HTOdemux* was used to assign cells to the MULTI-seq barcodes by inferring the negative binomial distribution for each barcode. If there is not enough of either barcode the cell is classified as negative. The cells were filtered based on the quality control (QC) parameters defined in Table2. The gene expression count matrix was then log normalised and used to identify variable features. The data was scaled, and the 2000 most variable features were used to run PCA. The 10 first dimensions from the PCA were visualised using uniform manifold approximation and projection (UMAP). The cells were clustered by constructing a shared nearest neighbour graph and inferring cluster identity from the graph by optimizing the modularity function. Differential gene expression analysis was then performed using Wilcoxon Rank Sum test to identify the top marker genes for each cluster.

Table2. *The QC parameters used to filter the cells with Seurat.*

QC Parameters	REH	Horse
% MT features	< 6 %	< 6 %
Number of RNA features	> 700 & < 6000	> 700 & < 6000

A Snakemake pipeline was built to achieve automation of each step for future analysis of samples (https://github.com/Molmed/multiseq-ds). The pipeline first runs the Cellranger steps mkgtf and mkref to create a reference genome used as input for Cellranger count. This step can be removed if a reference genome already exists. The next step is to create a library csv file containing the path to the gene expression library and the barcode library. The file also contains sample names and the type of library. For a MULTI-seq barcode library the library type is set to “Custom”. The reference genome, the library csv and a csv file describing the custom barcodes are then used as input for Cellranger count. The output from Cellranger count is then used to demultiplex the cell type or treatment from the MULTI-seq barcodes. This step is done by the R package Seurat but an additional script with code for using the R package deMULTIplex is also provided. The summary rule will concatenate summary statistics for all the samples processed in one file. The pipeline comes with a configuration file which makes it easy to reuse the pipeline since it is only necessary to edit the configuration file. An overview of the pipeline can be seen in Figure 2.

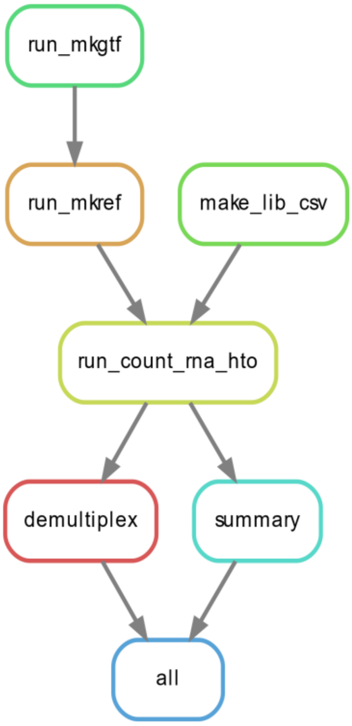

Figure 2. *Overview of the Snakemake pipeline developed to process and demultiplex scRNA data. The pipeline uses Cellranger to process the raw FASTQ files and Seuart for demultiplexing.*

scATAC-seq

To test out the scATAC protocol two different cell lines were used, REH and GM12878 which is a genome in the bottle cell line derived from B-lymphocytes (Zook *et al.* 2016). Nuclei were extracted from the REH and GM12878 cells and the nuclei were used to generate scATAC-seq libraries according to 10x Genomics protocol. The two resulting libraries were quality controlled on an Agilent TapeStation and subsequently sequenced on an Illumina NovaSeq flowcell at the SNP&SEQ Technology Platform, National Genomics Infrastructure at SciLifeLab.

Cellranger atac count (v1.2.0) from 10x genomics was used to process the raw FASTQ files. Cellranger atac count performs read filtering and alignment to the reference genome (GRCh38), identify transposase cut sites, peak calling, cell calling, and calculation of sequencing QC metrics. The samples were integrated in two different ways. Firstly, using Cellranger atac aggr and secondly using Seurat. The two methods were evaluated, and since Cellranger subsamples the reads to correct differences in sequencing depth, it was decided to move forward using Seurat for integration of samples. Seurat was used to create a peak set with common coordinates for the samples which was then used to quantify the peaks. Seurat was then used to further process the samples by calculating the QC parameters, including TSS enrichment score, nucleosome signal, blacklist ratio and the number of reads in identified peaks. The predicted gene activity was also calculated using the combined counts assigned to gene bodies and promoter regions. The gene activity matrix was log normalized and used in downstream analysis.

The data was filtered using the parameters specified in Table3. The remaining cells were then normalized across both cells and peaks using term frequency inverse document frequency (TF-IDF). Variable features were identified and singular value decomposition (SVD) was performed using these variable features to reduce the dimensionality of the data. A shared nearest neighbor (SNN) graph was constructed using the 15 first dimensions from the SVD and clusters identified based on the graph. UMAP was used for visualization of the clusters. Seurat was also used to identify markers for each cluster. Based on the predicted gene activity, gene set enrichment analysis (GSEA) and pathway enrichment analysis was then performed for the inhouse GM12878 sample using the R package ClusterProfiler (v3.18.1).

An additional publicly available GM12878 scATAC-seq dataset was downloaded from 10x Genomics. This second data set was preprocessed with Cellranger atac count in the same way as the two samples above. The two GM12878 samples were then integrated using a common peak set extracted by Seurat. The downstream processing and filtering were done in the same way as previously mentioned. The filtering parameters can be seen in Table3.

Table3. *Parameter values used to filter the scATAC-seq data.*

QC Parameters	**REH**	**GM12878**	**GM12878.10x**
TSS Enrichment	> 2	> 2	> 2
Blacklist Ratio	< 0.05	< 0.05	< 0.05
Nucleosome Signal	< 4	< 4	< 4
% Reads in Peaks	> 15	> 15	> 15
Total Reads in Peaks	> 3000 & < 20000	> 3000 & < 20000	> 3000 & < 20000

A Snakemake pipeline was developed that can run multiple samples in parallel. The pipeline starts by running Cellranger atac count which produces a count matrix with peaks as one dimension and cells as the other. The count matrix is then used to calculate QC parameters such as TSS enrichment score, blacklist ratio and nucleosome signal. The preprocessing step will return a Seurat object. If multiple samples are processed, they can be merged using Seurat. The samples can either be merged on a common peak set or on their individual peak sets. These two steps can be easily turned on or off in the configuration file. The summary rule will concatenate sequencing statistics from Cellranger atac count if multiple samples are processed. As with the gene expression pipeline this pipeline comes with a configuration file which makes it easy to edit and reuse. As there was no scATAC-seq data with MULTI-seq barcodes available yet the demultiplexing of this data type has not been tested and included in the pipeline. The steps will however be similar the scRNA-seq pipeline and can easily be added when this data becomes available. An overview of the pipeline without the demultiplexing step included can be seen in Figure 3 (https://github.com/Molmed/multiseq-ds/).

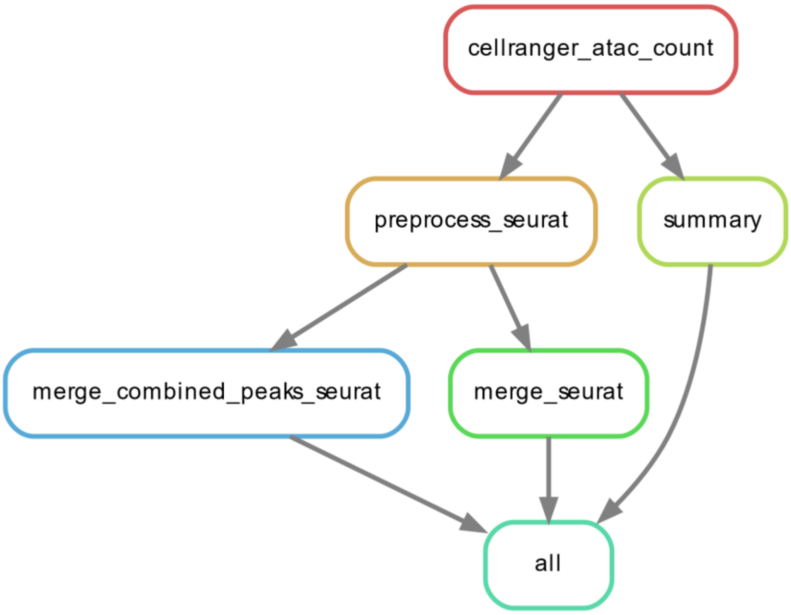

Figure 3. *Overview of the Snakemake pipeline developed to process and demultiplex scATAC data. The pipeline uses Cellranger to process the raw FASTQ files and Seurat for downstream analysis.*

scWGBS

Cells from a human leukemia K562 cell line (Lozzio & Lozzio 1975) and liver cells from a patient sample were processed and sequenced using the scSPLAT protocol (Raine et al, in preparation). In total 240 K562 cells and 309 liver cells were sequenced on an Illumina NovaSeq 6000 S4 flowcell to a depth of 2-60 M raw read pairs per cell. A custom pipeline using the software Bismark (v0.22.1) was used to process the raw FASTQ files. Bismark performs alignment to the reference genome (GRCh38) with the Bowtie aligner (Langmead & Salzberg 2012) and performs methylation calling (Krueger & Andrews 2011). Bismark can be run for single-cell data and will in that case output individual methylation call files for each cell.

A modified version of the Epiclomal preprocessing pipeline was then run to identify non-redundant regions. The pipeline takes individual methylation call files for each cell and a set of specified regions of interest. Non-redundant regions are identified by first filtering the regions with less then 5% coverage in 10% of cells and then inferring the interquartile range (IQR) and keeping the most variable regions. The output files from the preprocessing pipeline were used as input for Epiclomals clustering pipeline. In the first step of the pipeline four different non-probabilistic methods of clustering are used to initialise clusters assignments for the probabilistic part of the pipeline. These methods include EuclidianClust and DensityCut which both use the mean methylation levels across the retained regions and HammingClust and PerasonClust which both use the methylation levels at individual CpG sites across the retained regions. The four methods are evaluated, and the optimal number of clusters are then used in combination with random values to initialise the probabilistic clustering. Epiclomal uses mixed hierarchical Bernoulli distributions trained by a variational Bayesian algorithm to infer the true hidden state for each CpG site and cluster assignments for each cell. Epiclomal was run in two different ways, EpiclomalBasic and EpiclomalRegion. EpiclomalBasic assumes that the true hidden methylation states of CpG sites share the same distribution across all regions while EpiclomalRegion allows for variations in distribution (P. E. de Souza *et al.* 2020). An overview of the pipelines for preprocessing and clustering is outlined Figure 5.

Three different types of regions were used for clustering. Firstly, the coordinates for CpG islands for GRCh38 were downloaded from the UCSC Genome Browser (https://genome.ucsc.edu/) and used to cluster both cell types. To further investigate the heterogeneity observed in the patient liver cells, the coordinates for gene bodies and the coordinates for 1000 base pairs upstream and downstream from the transcription start sites (TSS) for GRCh38 was downloaded from UCSC Genome Browser. The liver cells were then clustered based on both these regions. The mean methylation levels in each region were used to visualise the cells with UMAP. The plots were evaluated to check for batch effects and the cluster assignments were compared for the different input regions.

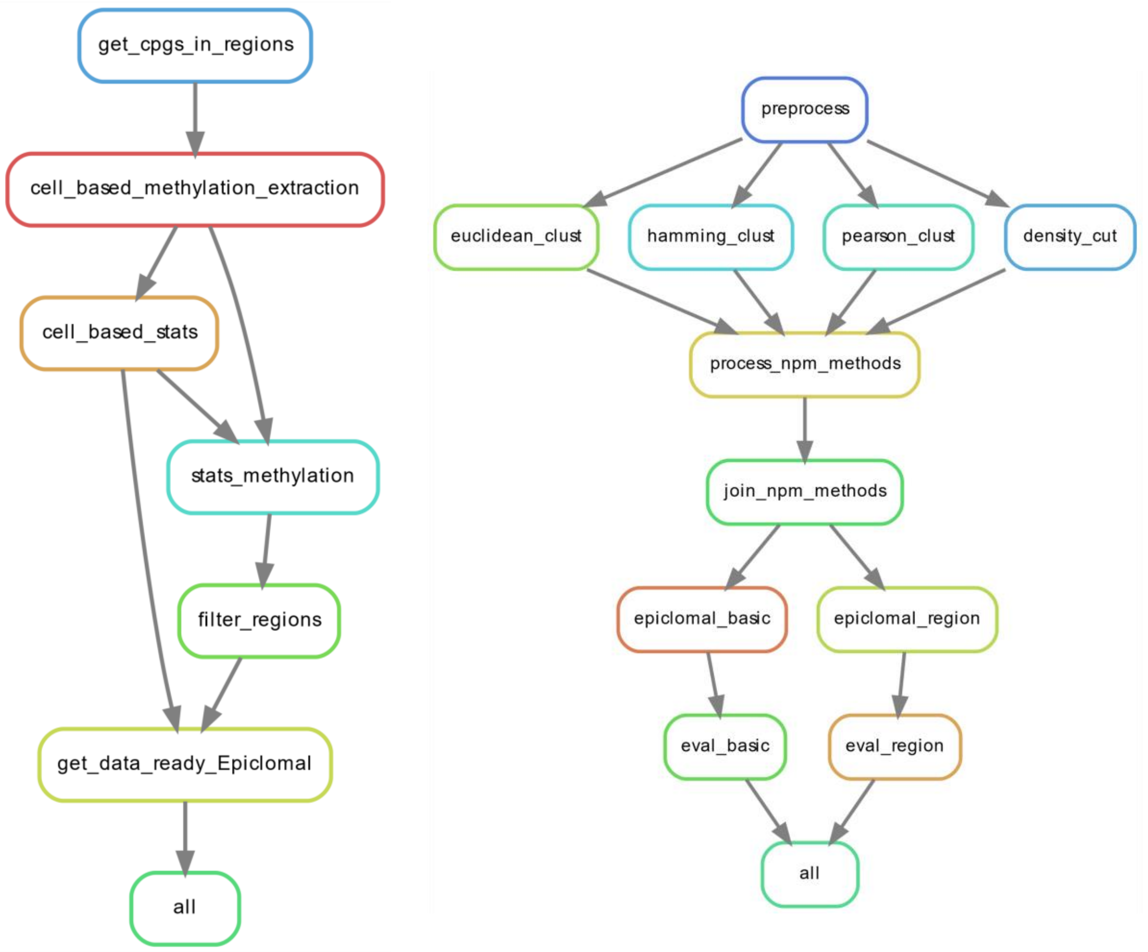

Figure 4. *To the left the modified Epiclomal pipeline for preprocessing single-cell methylation call files can be seen. To the right the pipeline for performing clustering with EpiclomalBasic and EpiclomalRegion can be seen.*

Results

scRNA-seq and MULTI-seq Barcodes

A barnyard experiment was conducted with a 1:1 mix of REH cells and primary horse cells to validate that the MULTI-seq barcodes worked. The two cell types were tagged with MULTI-seq barcodes prior to mixing. Two sequencing libraries were obtained, one GEX library and one MULTI-seq barcode library. The fraction of bases with a Phred score of 30 or more (Q30) in UMIs, the cell barcodes and the RNA reads were all over 90%. For the MULTI-seq barcodes the fraction was only 57% (Table4). To verify that MULTI-seq barcodes were present in the barcode library a manual search for the barcode sequences was done. Approximately 20% of the reads contained an exact match to barcode 2 and approximately 2% of reads contained an exact match to barcode 1. Almost 100% of the cell barcodes identified in the gene expression library were also present in the MULTI-seq barcode library.

Table4. *Sequencing and mapping statistics for the GEX and MUTI-seq libraries. Fraction of Q30 bases is split by UMI, cell barcode (CB) and read (RNA or MULTI-seq barcode).*

	Gene Expression	MULTI-seq Barcodes
% Bases > Q30 (UMI / CB / read)	97.9 / 96.6 / 93.8	97.2 / 97.1 / 57.2
N Reads	564,124,824	31,996,386
% Reads with Valid CB	97.8	99.1
% Valid UMI	100.0	100.0
% Mapped to EquCab3	13.2	-
% Mapped to GRCh38	79.2	-
Median UMI Counts / Cell EquCab3	1,517	-
Median UMI Counts / Cell GRCh38	9,416	-

The 10x Genomic's software Cellranger was used to map and annotate the reads and cells. In total 12,186 cell-barcodes associated with at least one cell could be identified. Out of these, 109 cells were classified as horse, 4770 were classified as human, and the remaining 7307 were classified as multiplets. A multiplet is a cell barcode with reads from more than one cell. Cellranger classifies multiplets as cell barcodes with significant amounts of reads mapped to both genomes. Out of the total number of reads 79.2% mapped confidently to the human genome and 13.2% mapped confidently to the horse genome, which is different than the expected 50:50 ratio. The MULTI-seq library was then used to demultiplex the cell type and classify the cells based on the barcode. The demultiplexing was performed with the Seurat function *HTOdemux* which uses the negative binomial distribution of each barcode to assign cells to the correct barcode. 7958 cells were classified as multiplets, 232 as having MULTI-seq barcode 1, 198 as having MULTI-seq barcode 2 and 3798 as negative. A negative cell does not have enough of either barcode to make a classification. The classifications from Cellranger and Seurat were compared to validate that the cells could be identified both by the gene expression and the barcodes. The number of cells in each class can be seen in Table5.

Table5. *Number of cells classified as human or horse by the reads in the gene expression library and by the MULTI-seq barcodes.*

	GRCh38/Barcode 1	EquCab3/Barcode 2	Multiplet	Negative
GEX	4770	198	7307	-
MULTI-seq	232	109	7958	3798

In Figure 6 the number of features detected for each cell, the total number of RNA counts and the percentage of mitochondrial genes out of the total number of features can be observed. The cells were filtered based on these QC parameters as stated in Table2. After filtering, 3026 cells out of 12186 remained and were used in the downstream analysis. As can be seen in Figure 6 many of the cells had quite a high fraction of mitochondrial features which is an indication of dying and/or lysed cells and were thus removed. The multiplets have a much higher number of identified features and RNA counts which indicates that they are indeed multiples.

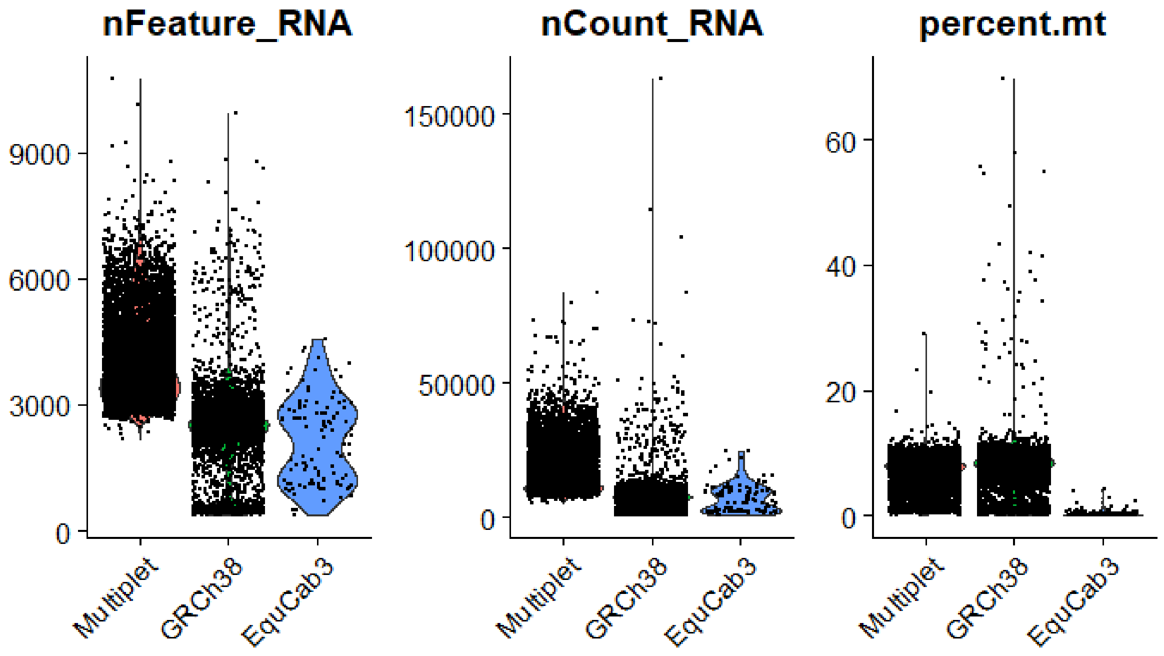

Figure 5. *The total number of genes detected, the total amount of RNA and the percent of reads mapping to mitochondrial genes for each cell split by Cell ranger classification.*

A visualization of the cells and the classifications from Cellranger and Seurat can be seen in Figure 7. Most cells are classified as either multiplets or negative based on the barcodes and as multiplets based on the gene expression. To further investigate the identity of the cells, Seurat was used to cluster the cells and perform differential gene expression analysis. The average gene expression of the top 100 marker genes for each cluster was calculated. The results of the clustering and the average gene expression per cluster can be seen in Figure 7. The markers for cluster 2 are exclusively horse genes. Cluster 3 show low expression levels for the horse genes and high expression levels for the human genes. For cluster 1 and 0 the expression levels are high for human genes, but they also have higher expression levels for the horse genes than cluster 3. NoTableis also that cluster 0 show very high expression levels for human mitochondrial genes which might indicate that these are dying cells.

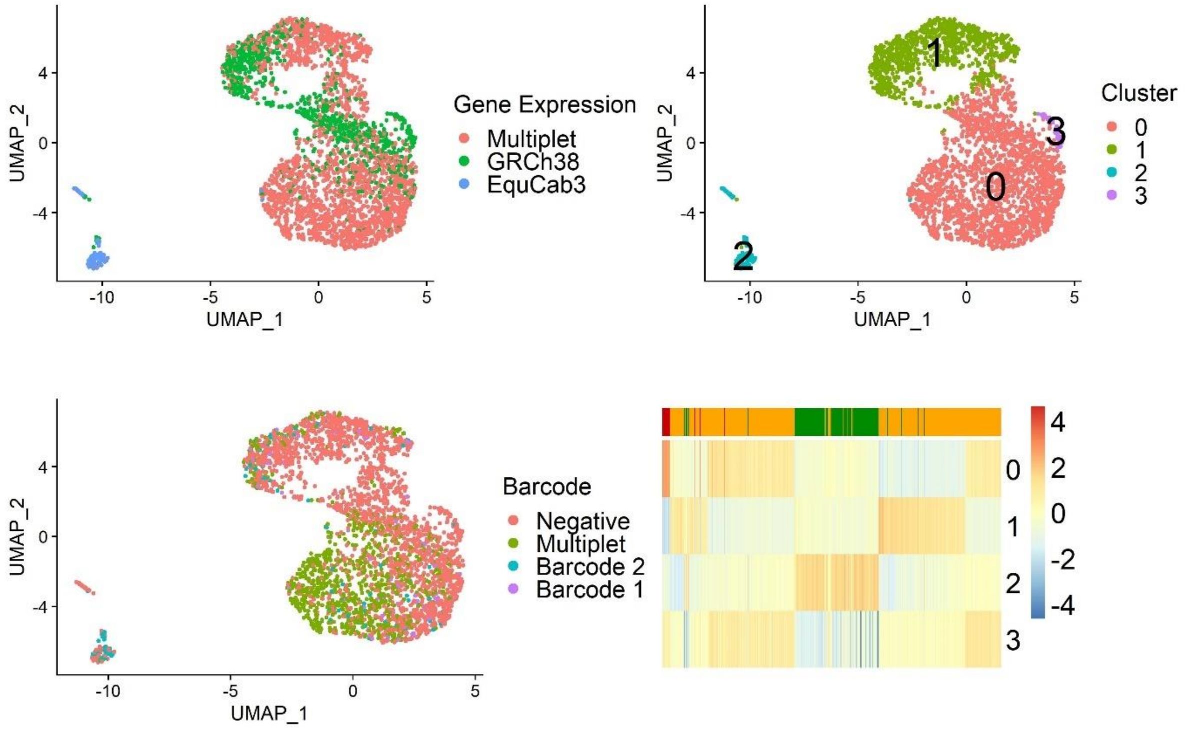

Figure 6. *(TL) The cells classified as human or horse based on which genome the reads mapped to. Multiplets have a significant amount of reads from both genomes. (TR) The cluster assignments by Seurat. (BL) The classification of cells based on which barcode they contained. Barcode 1 was used for the REH cells and barcode 2 was used for the horse cells. Multiplets have significant amounts of both barcodes and negative cells do not have either barcode. (BL) A heatmap of the top 100 marker genes for each cluster. Genes shown in yellow are human genes, genes shown in green are horse genes and genes shown in red are mitochondrial genes.*

scATAC-seq

Two different cell lines (REH and GM12878) were sequenced using the 10x scATAC-seq protocol to test if the method worked as intended. An additional data set with scATAC-seq data from a GM12878 sample was downloaded from 10x Genomics and processed in the same way as the inhouse samples. For all three samples the sequencing quality and depth was good. The fraction of Q30 bases was over 90% for all three samples. The number of cells detected in each sample and the number of cells that remained after filtering can be observed in Table6. The GM12878 samples were merged on a common peak set and clustered and visualised with Seurat (Figure 8).

Table6. *Sequencing statistics for the three ATAC-seq samples.*

	REH	GM12878 (inhouse)	GM12878 (10x)
N Cells	3561	4584	4776
N Reads	186,331,842	221,056,783	339,414,585
% Bases > Q30 (R1/R2/BC)	94.4 / 94.1 / 89.0	94.7 / 94.5 / 89.3	95.0 / 94.3 / 90.5
% Valid CB	98.3	98.3	98.2
Passed Filters	2176 (61%)	3078 (67%)	3862 (81%)

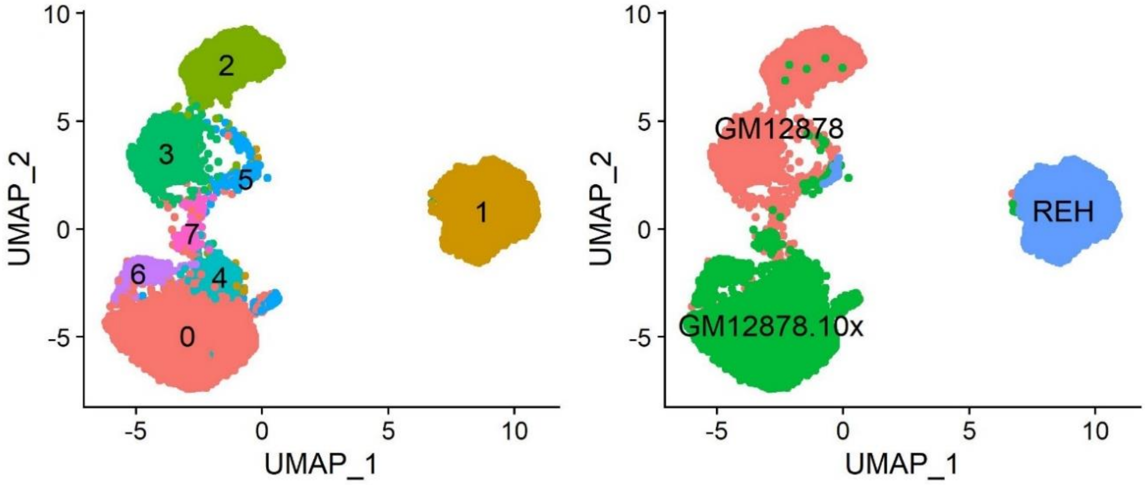

Figure 7. *To the left the REH and GM12878 samples can be seen visualised with UMAP and colored by cluster. To the right the cells are colored by sample. The REH and GM12878 samples were prepared inhouse and the GM12878.10x data was downloaded from 10x Genomics.*

To investigate if the heterogeneity in the inhouse GM12878 sample could be linked to the cell cycle state of cells Seurat's built-in annotation of cell cycle genes was used. Seurat scores each cell based on the expression levels of the genes linked to each cell cycle state and assigns the cells a state. A chi2 test was performed and it determined that the distribution of cell cycle states was different across clusters with a p-value of less than 2.6e-16. There is some enrichment for cells in phase G1 for cluster 5 and some enrichment for phase G2M and S in cluster 2 and 3. The result of the cell phase classification can be seen in Figure 9. Differential gene activity analysis between the clusters was also performed on the predicted gene activity. The top differentially active genes were used to perform GSEA and pathway enrichment analysis. An example of the GSEA results can be seen in Figure 9. The top 10 enriched terms and the fold change for the genes contributing to each term for cluster 5 is shown. Cluster 5 has lower expression levels for many genes linked to cellular ion homeostasis than the three four clusters.

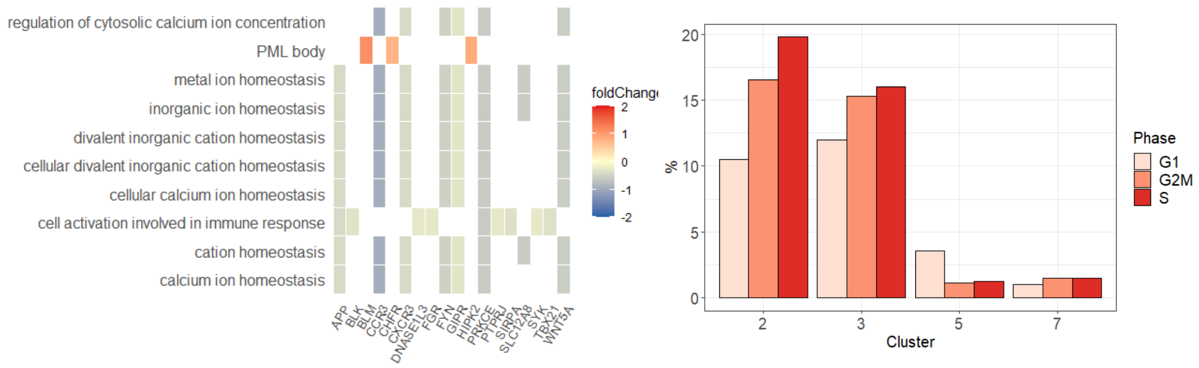

Figure 8. *To the left the result of the GSEA for the top differentially active genes for cluster 5 can be seen. The top 10 terms and the fold change is shown. To the left the distribution of cell cycle phase for each cluster within the inhouse GM12878 sample is plotted. The cluster correspond to the assignment in Figure 7.*

scWGBS

340 cells from the K562 cell line and 310 liver cells from a patient sample were prepared using the scSPLAT library preparation method (Raine et al, in preparation). The libraries were sequenced and the methylation status of individual CpG sites was called using Bismark (Krueger & Andrews 2011). In total, 503 cells (273 patient liver cells and 230 K562 cells) passed sequencing quality control and were used for downstream analysis with the Epiclomal clustering pipeline (P. E. de Souza *et al.* 2020). Epiclomal removes cells with no information available for the regions of interest and all cells passed this filtering step. Clustering was first performed using the methylation levels of individual CpG sites in the CpG Islands for both cell types. In addition to this the liver cells were clustered using gene bodies and TSS as the input regions of interest.

One cluster was identified by Epiclomal within the K262 cells. The K562 cells were prepared in 8 pools with scSPLAT protocol and sequenced in 2 batches. In addition to this, a third batch which was prepared with a different library preparation method called snmC-seq2 (Luo *et al.* 2018) and sequenced. As shown in Figure 9, batches 1 and 2 separate along the second dimension, however the cells prepared with the snmC-seq2 approach do not seem to cluster together and are distributed between the two scSPLAT batches. The preparation pools do not seem to cluster within each batch (Figure 9).

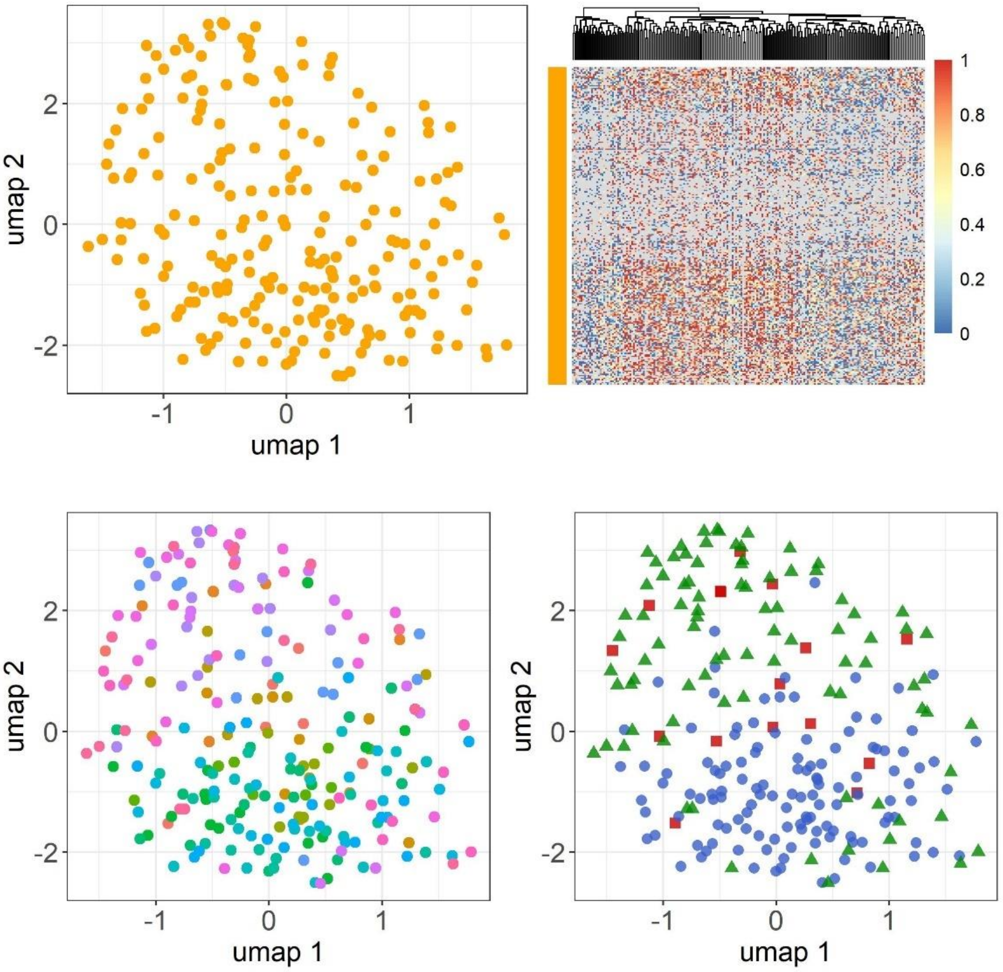

Figure 9. *In the top left the UMAP visualization of the K562 cells colored by cluster can be seen. The same data can be seen as a heatmap at the top right. In the bottom left the cells are colored by pool and to the right by batch. The snmC-se2 cells can be seen in red in the bottom left Figure.*

Two clusters with distinct methylation profiles could be identified among the liver cells. The difference in methylation levels can be observed in Figure 10. The liver cells were prepared using the scSPLAT protocol in 9 different pools and sequenced in three different batches. The distribution of the pools and batches within the two clusters can be seen in Figure 10. No batch effects were observed for either the pool or the sequencing batch. To further investigate the heterogeneity observed in the liver cells clustering was performed using gene bodies and 1000 bp upstream and downstream from the TSS as the input regions of interest. Both gene bodies and the transcription start sites resulted in close to identical cluster assignments to the CpG Islands. No batch effects could be observed when clustering based on either TSS or gene bodies.

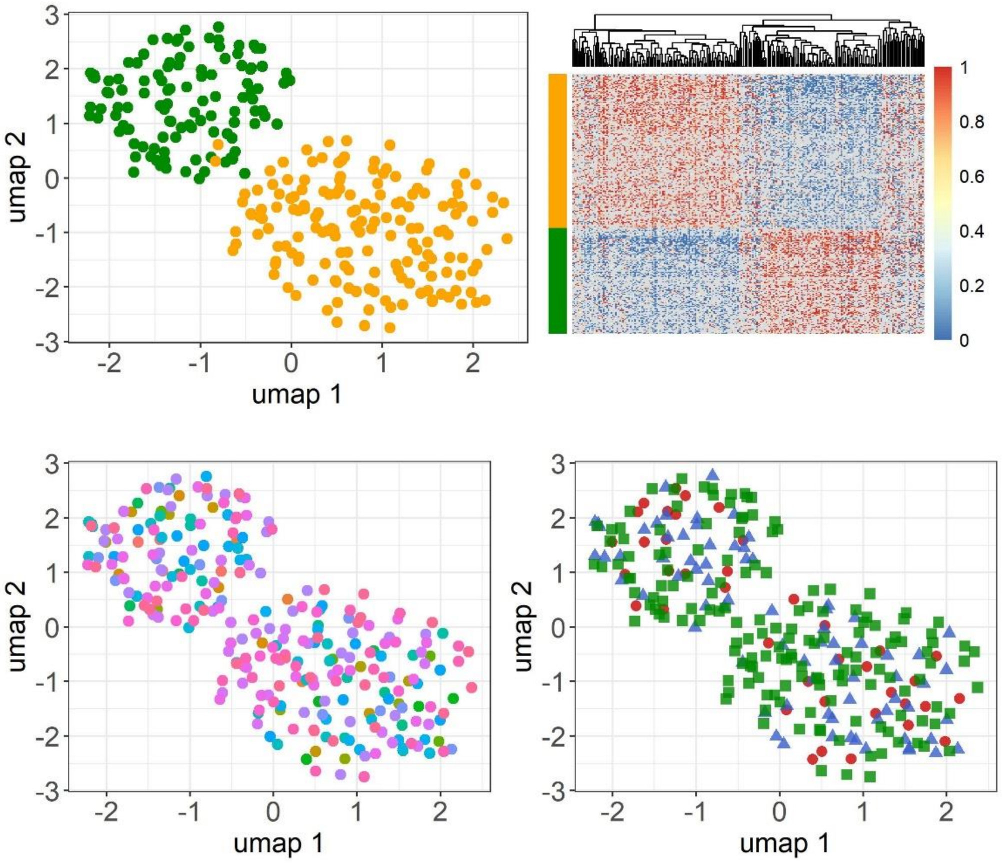

Figure 10. *The UMAP visualization and heatmap for the liver cells colored by Epiclomal cluster can be seen in the top row. In the bottom left Figure the cells are colored by pool and in the bottom right they are colored by batch.*

Discussion

Herein, we tested three methods for sequencing of single-cells and established the bioinformatics pipelines for high throughput analysis of single-cells in a translational research environment. The bioinformatics pipelines have been established and well-documented, enabling robust and high throughput analysis of single-cells for future studies. However, further optimization in the wet-lab is required, which is discussed in detail below.

Analysis of scRNA-seq Data and MULTI-seq Barcodes

The purpose of the scRNA experiment was to test and verify that the MULTI-seq approach (McGinnis *et al.* 2019) for barcoding cells in the 10x Genomics Protocol worked as intended. Our approach was based on a “barnyard” experiment where cells from two different species, in our case human and horse, are barcoded, mixed at a 1:1 ratio and prepared in one sequencing library. The expected result is two clearly defined clusters based on the genome that the reads map to and based on the sample barcode the cell was tagged with prior to library preparation. Unfortunately, the results revealed at least four technical issues with the experiment, which are summarized below.

i. As can illustrated in Figure 6, more than half of the cells were classified as “multiples”, meaning that RNA from both cell types was observed in the same gel bead. Furthermore, very few horse cells were observed based on gene expression reads mapping to the horse genome EquCab3. It is unclear why such a low fraction of reads mapped to EquCab3. Both the horse cells and human cells were checked for viability before they were entered into the 10x Genomics microfluidics device, and during this QC step, both cell types appeared to have high viability >80%. Previous single-cell experiments using the same type of horse cells have been successful when mapping the reads to EquCab3, which speaks against reference bias. Over 95% of the total number of reads were mapped to either genome but only 13% were mapped to horse. This suggests that a large amount of horse RNA was lost before the amplification step. Over 90 percent of the bases in the RNA reads have a sequencing quality score of 30 or more, which suggests that the problem is not with poor quality of the sequences, rather that something has happened to the cells during the 10x Genomics protocol.

ii. Based on the mean gene expression levels for each cluster (Figure 6) we determined that the 118 cells in cluster 2 are correctly classified as horse cells by Cellranger. The 51 cells in Cluster 3 can be assumed to be human cells based on the annotation from Cellranger and the mean gene expression values. Clusters 0 and 1 are probably a combination of true multiplets and cells mixed with RNA from, presumptively, dying or lysed cells. The high expression levels for mitochondrial genes in cluster 0 corroborates this theory. High levels of mitochondrial RNA are a known indicator of cell death and lysis in single-cell experiments (Luecken & Theis 2019). It is therefore possible that a large fraction of both the human and horse cells lysed and mixed leading to the large number of multiplets and mix of human and horse RNA.

iii. The cells should have been clearly defined based on their unique MULTI-seq barcodes. Approximately 50% of the 109 horse cells could be identified correctly by the MULTIseq-barcode and the rest were negative, meaning they did not carry a barcode sequence. Of the cells that were most likely human, most did not carry a MULTI-seq barcode. In cluster 0 and 1 (Figure 6) most of the cells were either classified as barcode-negative or contained both barcodes. This also indicates that there are many lysed cells since both barcodes are found together in many of the cells.

iv. As previously mentioned, there is also an imbalance between the fraction of reads containing barcode 1 (horse) and barcode 2 (human), which deviates from the expected 1:1 distribution. The sequencing quality of the MULTI-seq barcode reads is lower (57% > Q30) than the quality of the RNA reads (94% > Q30). An exact match to barcode 2 was found in approximately 20% of the reads while barcode 1 was present in only 2% of the reads. If many horse cells lysed and released barcode 2, it may have masked the presence of barcode 1 in the human cells or caused the software to call the cell as a multiplet since there was very little of barcode 1 to begin with in the human cells. The fact that there is significantly less of barcode 1 could suggest that this barcode was not absorbed correctly by the human cells or that it for some reason had degraded at a higher rate than barcode 2.

These results indicate that we have two major issues, and the most probable scenario seems to be that 1) the horse cells have lysed and much of the horse RNA was washed away and 2) the barcode used for the human cells was not absorbed correctly and therefore also lost. In addition to this there are high expression levels of human mitochondrial genes which suggest that the human cells were also dying and possibly lysing. This would explain the high multiplet rate both based on gene expression and barcodes. Since there were problems both with the cells and the barcodes it is hard to evaluate how well the barcodes have worked.

A second barnyard experiment will be performed and based on the problems observed with this experiment some changes could be suggested. Firstly, it could be beneficial to use a species with a more extensively mapped transcriptome such as mouse. Since the human cells come from REH cell line it could also be useful to use a cell line for the second species as well. These two changes could potentially result in less difference in fraction of reads mapped to each genome and thus ensure that the cells get assigned the correct cell type by Cellranger. To ensure that the correct cell type can be assigned, even though there is some bleeding of barcodes between cells or degradation of the barcode sequence, two barcodes could be used per cell type. This would mean that even if one of the barcodes does not work as intended the second barcode can be used to demultiplex the cell type. Using two barcodes for the second barnyard experiment would allow us to compare the difference in fraction of correctly assigned cells using one or two barcodes and could thus aid in assessing if more than one barcode per treatment could be beneficial in the future.

Despite the technical problems with the MULTI-seq barcoding, we did develop an analysis pipeline with Snakemake that takes raw FASTQ files as input and outputs a demultiplexed Seurat object that can be used for downstream analysis (https://github.com/Molmed/multiseq-ds/). The pipeline can run multiple samples in parallel on a HPC cluster and will aggregate sequencing statistics for all samples. Since the pipeline comes with a configuration file it is easy to use and modify for different projects without changing the actual pipeline. The Snakemake pipeline will facilitate fast processing of future samples and ensure reproducibility.

Analysis of scATAC-seq Data

The scATAC experiments were performed to evaluate the protocol and to develop an analysis pipeline for this type of data. Two different cell lines were sequenced using 10x genomics protocol for scATAC-sequencing. As can be seen in Figure 7 the REH cell line is very homogenous while the GM12878 sample show some heterogeneity. To investigate if this could be a technical artifact or real biological variation, the predicted gene activity for each gene was used to perform differential gene activity analysis, gene set and pathway enrichment analysis and to investigate the cell cycle phase for each gene. There were some statistically significant differences in number of cells in the different cell phases between clusters which suggests that this could be a contributing factor to the observed heterogeneity observed in the GM12878 sample. The differential gene activity analysis and subsequent GSEA and pathway enrichment show that each cluster did have a set of significant marker genes enriched for certain terms. In Figure 8 the GSEA results for cluster 5 are shown as an example. Many of the terms are related to ion homeostasis which has been shown to play a part in regulation of the cell cycle which further strengthens the conclusion that this is biological variation possibly related to the cell cycle (Marakhova *et al.* 2019).

A second publicly available GM12878 scATAC-seq dataset was also downloaded from the 10x Genomics website and processed using our in-house pipeline. The two GM12878 samples were integrated using a common peak set and the cells visualised using UMAP. Heterogeneity could be observed in both samples. The samples did however mostly cluster separately and integrating them on a common peak set did not correct for this batch effect. There are several reasons that could contribute to the strong batch effects, such as the version of the 10x genomics kits used, differences in sample preparation, and in sequencing method. ScATAC data are generally not as consistent as gene expression data and is often analysed as a complement to gene expression data rather than on its own. Since heterogeneity could be observed in both GM12878 samples and the results from the GSEA and the cell cycle phase analysis it may be assumed that it is not a technical artifact but true biological variation. However, additional scATAC-seq samples should be run to evaluate the degree of technical variation and how much it will affect future studies where the samples will have to be processed in batches.

The scRNA-seq data the REH and GM12878 scATAC data was used to develop an analysis pipeline that can be used for future samples. The pipeline is built in Snakemake and comes with a configuration file which makes the pipeline general and easy to edit (https://github.com/Molmed/multiseq-ds/). Demultiplexing has not been tested yet but will work similarly to the scRNA data and because of this it will be easy to include this step in the pipeline as soon as the data becomes available. The pipeline will facilitate easy analysis of future samples within the project and ensure reproducibility.

Clustering of scWGBS Data with Epiclomal

A novel library preparation method for scWGBS called scSPLAT was tested on K562 and liver cells from a patient sample. The methylation call files were then used as input to the Epiclomal clustering pipeline. Epiclomal is a probabilistic framework for clustering of scWGBS data that takes methylation call files for individual cells and a specified set of regions of interest as input. The K562 and patient liver cells were first clustered based on the methylation levels in the CpG islands. The liver cells were also clustered based on gene bodies and transcription start sites.

Among the K562 cells only one cluster could be identified while liver cells split in two clusters with distinct methylation profiles. This is expected since the liver cells come from a tissue sample consisting of multiple cell types while the K562 cells are expected to be homogenous. The two clusters identified in the liver cells most likely represent two different cell types. To investigate if other cell types could be identified among the liver cells, additional input regions where used for Epiclomal. The CpG sites found in gene bodies and transcription start sites were used to cluster the cells. Both these Epiclomal runs identified two clusters and the cluster assignments were close to identical regardless of the input region. This further strengthens the conclusion that these two clusters are different cell types found in liver tissue. There are more than two cell types found in the liver. In scRNA-seq studies up to 20 distinct cell types have been identified and different types of hepatocytes are most dominant. Based on this one of the larger clusters are most likely hepatocytes and the smaller cluster could be endothelial cells (MacParland *et al.* 2018, Aizarani *et al.* 2019). Since the total number of cells is relatively small it is hard to tell if these rarer cell types are not present or if their methylation profile is not distinct enough from the more dominant cell types to be distinguished by Epiclomal. To further investigate the identity of the cell types the genes around the differentially methylated CpG islands would need to be explored.

As can be seen in Figure 9 no batch effect can be seen in the liver cells. For the K562 cells there is some clustering of the cells based on batch (Figure 10). The cells prepared with a separate protocol does however seem to be spread out amongst batch 2. This suggests that the effect of library preparation method is small. Since there is no true heterogeneity amongst the K562 cells the sequencing batch might be the most significant difference between the cells and Epiclomal is able to pick up on this variation. When using Epiclomal on cells with real biological variation the batch becomes less important and does not show up in the clustering.

Conclusions and Future Perspectives

The main goals of the project were to test out the MULTI-seq barcodes and single-cell protocols that will be used in future high throughput drug screens on pediatric ALL patient samples. The aim was also to develop analysis pipelines with Snakemake for the different data types. The MULTI-seq barcodes were tested with a barnyard experiment with a mix of human and horse cells. There were several technical issues with this experiment and based on the results it was not possible to properly evaluate the barcoding. A second barnyard experiment will be performed before the project moves on to the next phase and the lessons learnt from this initial test will be taken into consideration. Even though the barcodes did not work as intended, a robust analysis pipeline has been set up for analysis and demultiplexing of scRNA-seq data. In addition to this the scATAC-seq protocol was tested on two cell lines (REH and GM12878) and an analysis pipeline was developed. A demultiplexing step can easily be added to this pipeline in the same way as for the scRNA-seq data. Finally, clustering of scWGBS data was explored with the Epiclomal framework to evaluate a novel library preparation method called scSPLAT. The hope is to be able to use this method in the context of high throughput drug screens in the future. All the methods tested, and the analysis pipelines developed within this project will hopefully aid in future drug screen experiments.

References

Aizarani N, Saviano A, Sagar, Mailly L, Durand S, Pessaux P, Baumert TF, Grün D. 2019. A Human Liver Cell Atlas: Revealing Cell Type Heterogeneity and Adult Liver Progenitors by Single-Cell RNA-sequencing. bioRxiv 649194.

Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, Kelsey G, Stegle O, Reik W. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature Methods 13: 229–232.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. Nature Biotechnology 35: 316–319.

Hui T, Cao Q, Wegrzyn-Woltosz J, O'Neill K, Hammond CA, Knapp DJHF, Laks E, Moksa M, Aparicio S, Eaves CJ, Karsan A, Hirst M. 2018. High-Resolution Single-Cell DNA Methylation Measurements Reveal Epigenetically Distinct Hematopoietic Stem Cell Subpopulations. Stem Cell Reports 11: 578–592.

Hwang B, Lee JH, Bang D. 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine 50: 1–14.

Iacobucci I, Mullighan CG. 2017. Genetic Basis of Acute Lymphoblastic Leukemia. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 35: 975–983.

Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, Gate RE, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology 36: 89–94.

Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28: 2520–2522.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571–1572.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

Liu Y-F, Wang B-Y, Zhang W-N, Huang J-Y, Li B-S, Zhang M, Jiang L, Li J-F, Wang M-J, Dai Y-J, Zhang Z-G, Wang Q, Kong J, Chen B, Zhu Y-M, Weng X-Q, Shen Z-X, Li J-M, Wang J, Yan X-J, Li Y, Liang Y-M, Liu L, Chen X-Q, Zhang W-G, Yan J-S, Hu J-D, Shen S-H, Chen J, Gu L-J, Pei D, Li Y, Wu G, Zhou X, Ren R-B, Cheng C, Yang JJ, Wang K-K, Wang S-Y, Zhang J, Mi J-Q, Pui C-H, Tang J-Y, Chen Z, Chen S-J. 2016. Genomic Profiling of Adult and Pediatric B-cell Acute Lymphoblastic Leukemia. EBioMedicine 8: 173–183.

Lozzio CB, Lozzio BB. 1975. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood 45: 321–334.

Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology 15: e8746.

Luo C, Rivkin A, Zhou J, Sandoval JP, Kurihara L, Lucero J, Castanon R, Nery JR, Pinto-Duarte A, Bui B, Fitzpatrick C, O'Connor C, Ruga S, Van Eden ME, Davis DA, Mash DC, Behrens MM, Ecker JR. 2018. Robust single-cell DNA methylome profiling with snmC-seq2. Nature Communications 9: 3824.

MacParland SA, Liu JC, Ma X-Z, Innes BT, Bartczak AM, Gage BK, Manuel J, Khuu N, Echeverri J, Linares I, Gupta R, Cheng ML, Liu LY, Camat D, Chung SW, Seliga RK, Shao Z, Lee E, Ogawa S, Ogawa M, Wilson MD, Fish JE, Selzner M, Ghanekar A, Grant D, Greig P, Sapisochin G, Selzner N, Winegarden N, Adeyi O, Keller G, Bader GD, McGilvray ID. 2018. Single-cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nature Communications 9: 4383.

Malard F, Mohty M. 2020. Acute lymphoblastic leukaemia. The Lancet 395: 1146–1162.

Marakhova I, Domnina A, Shatrova A, Borodkina A, Burova E, Pugovkina N, Zemelko V, Nikolsky N. 2019. Proliferation-related changes in K + content in human mesenchymal stem cells. Scientific Reports 9: 346.

McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, Hu JL, Murrow LM, Weissman JS, Werb Z, Chow ED, Gartner ZJ. 2019. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods 16: 619–626.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. 2021. Sustainable data analysis with Snakemake. F1000Research, doi 10.12688/f1000research.29032.2.

P. E. de Souza C, Andronescu M, Masud T, Kabeer F, Biele J, Laks E, Lai D, Ye P, Brimhall J, Wang B, Su E, Hui T, Cao Q, Wong M, Moksa M, Moore RA, Hirst M, Aparicio S, Shah SP. 2020. Epiclomal: Probabilistic clustering of sparse single-cell DNA methylation data. PLOS Computational Biology 16: e1008270.

Pauli C, Hopkins BD, Prandi D, Shaw R, Fedrizzi T, Sboner A, Sailer V, Augello M, Puca L, Rosati R, McNary TJ, Churakova Y, Cheung C, Triscott J, Pisapia D, Rao R, Mosquera JM, Robinson B, Faltas BM, Emerling BE, Gadi VK, Bernard B, Elemento O, Beltran H, Demichelis F, Kemp CJ, Grandori C, Cantley LC, Rubin MA. 2017. Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. Cancer Discovery 7: 462–477.

Raine A, Manlig E, Wahlberg P, Syvänen A-C, Nordlund J. 2017. SPlinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. Nucleic Acids Research 45: e36.

Rosenfeld C, Goutner A, Venuat AM, Choquet C, Pico JL, Dore JF, Liabeuf A, Durandy A, Desgrange C, De The G. 1977. An effective human leukaemic cell line: Reh. European Journal of Cancer (1965) 13: 377–379.

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression. Nature biotechnology 33: 495–502.

Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, Shah P, Bell JC, Jhutty D, Nemec CM, Wang J, Wang L, Yin Y, Giresi PG, Chang ALS, Zheng GXY, Greenleaf WJ, Chang HY. 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nature Biotechnology 37: 925–936.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics 14: 618–630.

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature Methods 11: 817–820.

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single-cells. Nature Methods 14: 865–868.

Wen L, Tang F. 2018. Single-cell epigenome sequencing technologies. Molecular Aspects of Medicine 59: 62–69.

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. 2017. Massively parallel digital transcriptional profiling of single-cells. Nature Communications 8: 14049.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre ABR, Chandramohan D, Chen F, Jaeger E, Moshrefi A, Pham K, Stedman W, Liang T, Saghbini M, Dzakula Z, Hastie A, Cao H, Deikus G, Schadt E, Sebra R, Bashir A, Truty RM, Chang CC, Gulbahce N, Zhao K, Ghosh S, Hyland F, Fu Y, Chaisson M, Xiao C, Trow J, Sherry ST, Zaranek AW, Ball M, Bobe J, Estep P, Church GM, Marks P, Kyriazopoulou-Panagiotopoulou S, Zheng GXY, Schnall-Levin M, Ordonez HS, Mudivarti PA, Giorda K, Sheng Y, Rypdal KB, Salit M. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data 3: 160025.