2.3 Datasets

The datasets Kumar, Trapnell and Koh were downloaded from the *conquer* repository `http://imlspenticton.uzh.ch:3838/conquer`. The Zheng data was downloaded from the 10xChromium repository `https://support.10xgenomics.com/single-cell-gene-expression/datasets`. For the Kumar, Trapnell and Koh data count-scale, length-scaled Transcripts Per kilobase Million (TPM) are used. These are on a count-scale and independent from the average length of the transcript (Soneson et al., 2015). The Zheng data consists of UMI counts.

Kumar et al. (2014) The Kumar dataset consists of *Dgcr*8-knockout and V6.5 variotypes from mouse embryonic stem cells (mESCs). Cells were cultured on a serum with a Leukaemia Inhibitory Factor (LIF) or under Erk and GSK3 signalling inhibition (2Li). The authors investigated the expression of pluripotency factors and their involvement in the heterogeneity of pluripotent stem cells. Sample preparation and whole transcriptome amplification was done using a Fluidigm C1 system and following a SMARTer protocol. Sequencing was done using Illumina system with paired-end reads. Sequencing depth was 1 million reads per cell.

Trapnell et al. (2014) Trapnell et al. (2014) used human skeletal muscle myoblast cells to investigate temporal differentiation. Cells were expanded under high-mitogen conditions. Differentiation is induced by switching to low-serum medium. Cells were captured before switching to low-serum medium (T0), after 24 h (T24) and 48h (T48). Between 49 and 77 cells were isolated at each time point and used for single mRNA-Seq library preparation by the use of a Fluidigm C1 system and following a SMARTer protocol. Libraries were sequenced with paired-end sequencing on a HiSeq 2500 (Illumina) platform. Sequencing depth was 4 million reads per library. The Authors excluded libraries that contained fewer than 1 million reads.

Koh et al. (2016) H7 human embryonic stem cells (hESCs) were used to study human mesoderm development. Starting from undifferentiated stem cells, several differentiation stages, sorted by time point and further refined by fluorescence-activated cell sorting (FACS) were isolated. Finally nine different cell lines were obtained: undifferentiated H7 hESCs (H7hESC), anterior primitive streak populations (APS) , mid primitive streak populations (MPS), lateral mesoderm (D2LtM), FACS-purified DLL1+ paraxial mesoderm populations (DLL1pPXM), early somite progenitor populations (ESMT) , PDGFR+ sclerotome populations (Sclrtm) and two different dermomyotome populations (D5CntrlDRmmtm). In total ten different cell types were then sequenced on a Fluidigm C1 system and following a SMARTer protocol. Libraries were sequenced with paired-end sequencing on a HiSeq 2500 (Illumina) platform. Sequencing depth was $1 - 2$ million reads per cell.

Zheng et al. (2017) FACS-purified fresh peripheral blood mononuclear cells (PBMCs) were used to assess the performance of the 10xChromium system. Sample preparation and library construction were done with a 10xChromium system. The libraries were sequenced using an Illumina system. For this study, the datasets for CD19+B, CD8+CD45RA+ naive cytotoxic, CD14+ monocytes and CD4+/CD25+ regulatory T cells were used to construct an artificial population. From each library, 200 cells were sampled and merged to obtain a single expression matrix.

Simulated dataset Using the Splatter package expression data were simulated (Oshlack et al., 2017). Parameters for the simulation were estimated from a subpopulation of the Kumar dataset. Embryonic stem cell variotypes V6.5 with signalling inhibition and LIF were used for the estimation. SimDataKumar consists of 500 cells with four subpopulations. The fractions per subpopulations were 0.1, 0.15, 0.5, 0.25 of the total cell population. The probability that a gene is differentially expressed is

0.05,0.1,0.2 and 0.4 in the four different groups. Similarly, SimDataKumar2 consist of four subgroups with fractions of 0.2, 0.15, 0.4 and 0.25 of a total of 500 cells. The fractions of differentially-expressed genes were lower with a probability of 0.01, 0.05, 0.05 and 0.08. The spike-in RNA is excluded before the parameter estimation.

2.4 Data transformation and normalisation

Data transformation RNA-seq data may suffer from heteroscedasticity and skewness (Zwiener et al., 2014). Genes with higher mean have on average a higher variance across cells leading to unequal variances between different genes. To handle this property different transformation were considered. A binary logarithmic transformation with a pseudo-count of one, arcus sinus transformations and a variance-stabilising transformation (VST) from the DESeq package (Huber et al., 2002). Log transformations will have an impact on extreme values. However, it will not address the problem of heteroscedasticity. Arcus sinus transformation should deal with extreme values and equalise the variances. After transformation, the mean and the variances should be independent. VST address the problem of extreme values and unequal variances across genes. After such transformation, the mean and the variances of the genes should be independent. For the study, a binary logarithmic transformation plus a pseudo-count of one is used. The data still shows heteroscedasticity. There are only subtle differences between the arcus sinus and log transformations in all the datasets. The Compared with the Arcus sinus transformation. The mean-SD dependence for different transformations is shown in Figure 1.

Filtering and normalisation The quality control of the data sets follows Lun et al. (2016b). In a first step genes that are not expressed in any cell (systematic-zeros) are removed to reduce the size of the expression matrix. To find potential outliers, PCA on the phenotype characteristic of each cell can be used (Figure 7, 8, 9, 10; a). Cells were filtered based on the library size and the total number of genes. Cells with log10-library sizes that are more than 3 median absolute deviations (MADs) below the median log-library size were filtered out (Figure 2, 3, 4, 5). The same filtered was used with respect to the total number of genes per cell. For the Kumar and the Zheng dataset ERCCs and MT counts were available. Cells with large proportions of ERCC or mitochondrial RNA are seen as low-quality cells. In the Kumar dataset cells with an ERCC proportion above 3 MADs are as well removed. The same filter was used for mitochondrial RNA in the Zheng data.

Metadata for the Trapnell dataset contained information about the cell quality. In this dataset cells that were marked as debris or if a single library consist of more than one cell were as well filtered out. Leaving 531 cells in the Koh dataset, 246 in the Kumar dataset and 222 in the Trapnell dataset. The filtering was less strict in the Koh dataset compared to the original analysis were they retained 498 cells.

Low-abundance genes influence the mean-variance trend. Here low-abundance genes are filtered by their average counts (see Figure 7, 8, 9, 10; d). For the Kumar, Trapnell, simDataKumar and Zheng data genes with average counts less than one are removed. The Zheng data set had a shallower sequencing depth. A different filter is used, and features which are not expressed in at least two cells are excluded. To find batch effects a linear model regressing the PC values against the total features was used Lun et al. (2016b).

Another examination of the technical variation was done using the marginal variancesLun et al. (2016b). For that, a linear model with the expression values per gene as response variables and a chosen explanatory variable is fitted. The correlation coefficient can then be seen as the marginal explained variance for the explanatory variables.

A wide variety of normalisation methods exist based on bulk RNA methods. These methods are usually not designed for dealing with the zero-inflated nature of scRNA-seq dataLun et al. (2016a). Methods for normalisation of scRNA-seq data are based on spike-in or the RNA counts. Spike-in RNA is added before the library preparation. Any changes in the spike-inn coverage are assumed to be due to technical factors. The normalisation is done by scaling the counts to level the spike-in. However, this approach is not feasible as no or only a limited number of spike-inn counts were present in the datasets.

Here normalisation through pooled cells was used, where the problem of excess zero counts is reduced by the pooling of multiple cells Lun et al. (2016a). The normalisation procedure can briefly be described as follows; (i) Different pools of cells are defined. (ii) The expression values are summed across the cell pools. (iii) The cell pool is normalised against an average of the summed expression values. (iv) This step is repeated several times to construct a linear system. The summed count size is then used to estimate corrected size factor. The size factors for the pooled cells were then deconvoluted" into cell-based factors.