scRNA-seq Workflow

For the initial processing of the data we use the Cellranger software (version 5.0.1). The raw data in bcl format is first converted to fastq format and demultiplexed using cellranger mkfastq. Gene counts are then obtained by aligning reads to the human genome (GRCh38) using cellranger count with the option –include-introns to ensure that reads aligning to introns are mapped. The transcriptome annotation used is ENSEMBL v98 (reference available from the CellRanger website). Libraries are then aggregated from all samples using cellranger aggr with the option –normalize=none, which means we avoid downsampling to give equal numbers of reads per cell mapped to the transcriptome across the samples.

For subsequent quality control and downstream analyses we utilize the R package Seurat. In order to remove low quality cells from the dataset, we exclude from downstream analysis any cells that have 1, a total number of unique genes detected greater than three median average deviations above the dataset median or less than 200; 2, greater than 5% of their reads mapping to the mitochondrial chromosome, because in a broken cell cytoplasmic RNA will be lost, while RNAs that are enclosed in the mitochondria will be retained. Finally, genes detected in less than three cells are excluded. After applying these filtering steps, the resulting dataset is subject to a library size normalization on a per cell basis where the scale factor is 10,000 and then log-transformed via ln(x+1) where x is the normalized count value for a given gene. The top 2000 most variable genes are found and the normalized counts are then scaled using the formula (TotalCounts(Cell) - mean(TotalCounts(AllCells)))/S.D(TotalCounts(AllCells)). Principal component analysis is performed using GLM-PCA, and the Harmony algorithm used to correct for batch effects (and integrate data from different scRNA-seq runs where necessary). For single batch analyses (a single scRNA-seq sequencing run) the sample ID and the disease state are passed as covariates to Harmony, whereas for the analyses where 2 or more batches are integrated together, batch, sex and the age bracket are also included. In each case, the theta value passed to Harmony for each covariate is equal to 2/Number of Covariates. An Elbow plot is generated in order to identify the significant principle components and define a cut off for the ‘dimensionality’ of the dataset i.e. the number of principle components to carry forward. Clustering (using the Seurat functions ‘FindNeighbours(), which uses the K-nearest neighbors algorithm and FindClusters(), using the Leiden algorithm) is then performed over the number of PCs we define to be significant, and UMAP run over the significant principal components for cluster visualization. We then assess the expression of a cohort of known marker genes across all the clusters, and based on this we assign each cluster a cell-type label. Finally, within each cell type cluster we perform differential expression analyses between cells for each pairwise combination of healthy control, incidental Lewy body diease and Parkinson’s disease, in order to assess how gene expression changes across different cell types as disease progresses. For this we use the package MAST, with batch, sex, age, post-mortem interval, RNA integrity number and cellular detection rate ((TotalUniqueGenes(Cell) - mean(TotalUniqueGenes(AllCells)))/S.D(TotalUniqueGenes (AllCells))) as covariates.