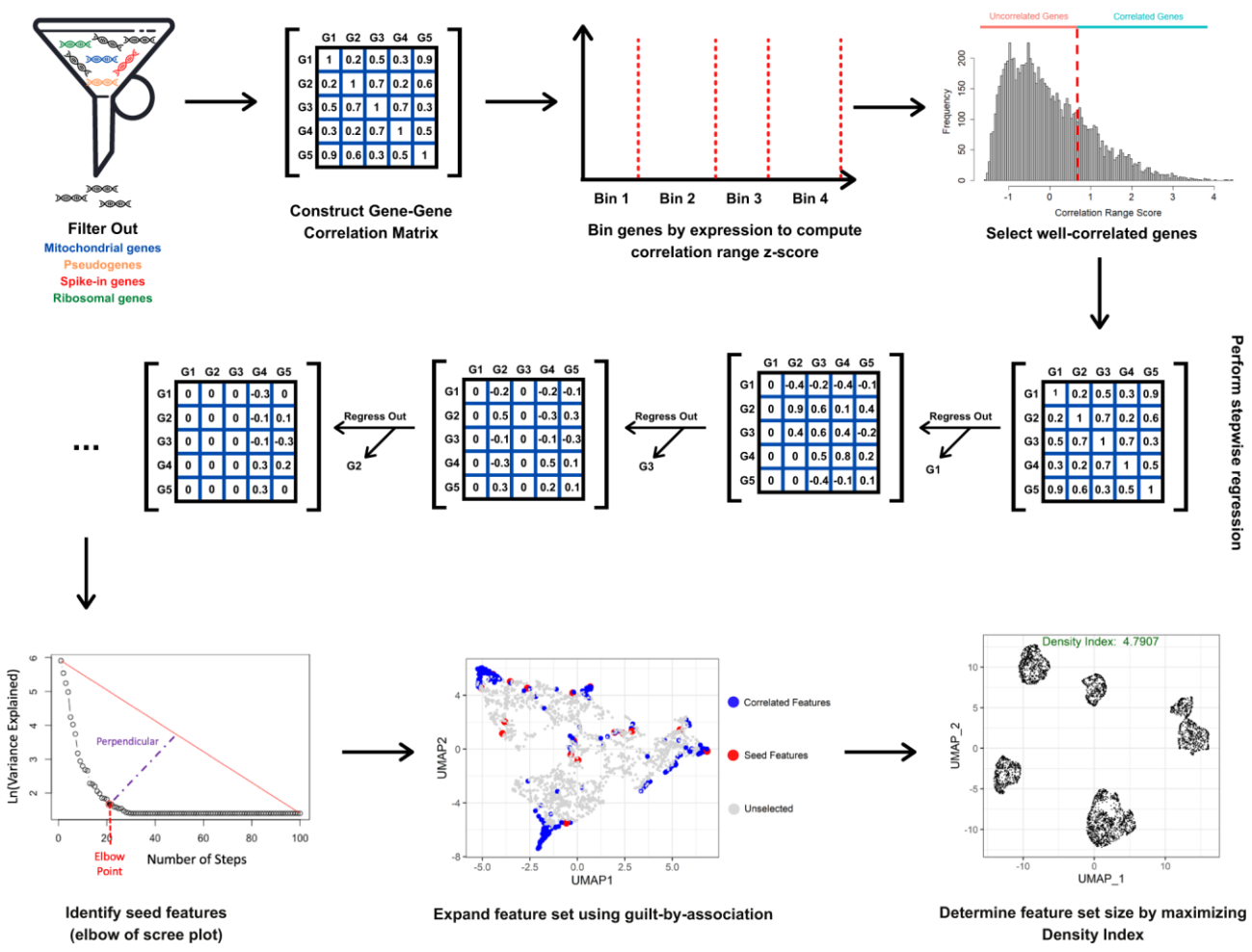
DUBStepR

A NOVEL FEATURE SELECTION ALGORITHM SYSTEM



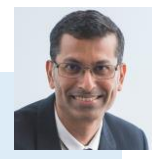
After filtering out mitochondrial, ribosomal, spike-in, and pseudogenes, DUBStepR constructs a GGC matrix and bins genes by expression to compute their correlation range z-scores, which are used to select well-correlated genes. DUBStepR then performs stepwise regression on the GGC matrix to identify a minimally redundant subset of seed features, which are then expanded by adding correlated features (guilt-by-association). The optimal feature set size is determined using the Density Index metric.

Since the first single-cell experiment was published in 2009, single-cell RNA sequencing (scRNA-seq) has become the quasi-standard for transcriptomic profiling of heterogeneous data sets. In contrast to bulk RNA-sequencing, scRNA-seq is able to elucidate transcriptomic heterogeneity at an unmatched resolution, and thus, allows downstream analyses to be performed in a cell-type-specific manner easily. This has been proven to be especially important for instance in case-control studies or in studying tumor heterogeneity.

In this study, the team from the Genome Institute of Singapore (GIS) developed a novel feature selection algorithm for scRNA-seq data, DUBStepR (**D**etermining the **U**nderlying **B**asis using **Step**wise **R**egression), that leverages gene-gene correlations with a novel measure of inhomogeneity in feature space, termed the Density Index (DI).

Heterogeneity in single-cell RNA sequencing (scRNA-seq) datasets is frequently characterised by identifying cell clusters in gene expression space, wherein each cluster represents a distinct cell type or cell state. Feature selection (selecting the set of genes on which cells are separated into clusters) is a critical step in the canonical clustering workflow. A good feature selection algorithm is one that selects cell-type-specific (DE) genes as features, and rejects the remaining genes, with the objective of optimising the separation between biologically distinct clusters.

DE genes specific to the same cell types tend to be highly correlated with each other (highly and lowly expressed in the same cells), whereas those specific to distinct cell types are likely to be anti-correlated.



"The team leveraged this property of gene expression correlations to develop DUBStepR. It provides a novel metric (termed DI) for determining the optimal number of features needed to separate cells into distinct clusters. They performed an objective benchmarking analysis to demonstrate the improvement DUBStepR provides over existing methods in the field based on cluster separation and selection of marker genes."

Dr Shyam Prabhakar, Associate Director,
Laboratory of Systems Biology & Data Analytics, GIS

The most striking improvement provided by DUBStepR was observed in analysis of scRNA-seq data from blood samples of rheumatoid arthritis (RA) patients. DUBStepR detected novel rare cell types and cryptic cell states that were either completely or partially undetected by all other methods. Based on previous literature, some of these cell types/states may even be relevant in the pathophysiology of the disease.

DUBStepR also possesses the unique property of being directly extensible to single-cell ATAC sequencing (scATAC-seq) data. Existing feature selection methods developed for scRNA-seq data are unable to overcome the technical constraints of scATAC-seq data, which DUBStepR is able to circumvent.

"DUBStepR provides a substantial improvement in identifying hematopoietic differentiation trajectories, thus opening up the possibility of incorporating a feature selection step in single-cell epigenomics analysis pipelines. "

Prof Patrick Tan, Executive Director, GIS

The paper "DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data" was published in *Nature Communications* on 6 October 2021.