scRNA-seq notes

Esthy Hung

July 15, 2019

Steps of single-celll RNA-seq analysis include pre-processing (quality control, normalisation, data correction, feature selection, and dimensionality reduction) and cell-and gene-level downstream analysis.

1 Workflow

1.1 Pre-processing

1. Raw data processing
2. Count matrices
3. Visualisation
    - Quality control
    - Normalisation
    - Data Correction
    - Feature selection

1.2 Downstream Analysis

1. clustering
2. marker identification
3. trajectory inference
4. Differential expression
5. gee dynamics

6. compositional analysis

7. metastable states

2 Pre-Processing

Raw data generated by sequencing machines are processed to obtain matrices of molecular counts (count matrices) or, alternatively, read counts (read matrices), depending on whether unique molecular identifiers (UMIs) were incorporated in single-cell libray construction protocol.

Raw data processing take care of read quality control, assigning reads to their cellular barcodes and mRNA molecules of origin, genome alignment, and quantification. Read and count matrices have dimensions (no. barcodes, no. transcripts).

Quality control

Three QC Covariates:

- Count depth: no. counts per barcode
- No. genes per barcode
- Fraction of counts from mitochondrial genes per barcode

Considering QC covariates in isoation can lead to misinterpretation of cellular signals so all three should be considered jointly when univariate thresholding decisions are made. In future, filtering models that account for multivariate QC dependencies may provide more sensitive QC options. Raw count matrices can include over 20,000 genes. This can be filtered out by filtering out genes not expressed in more than a few cells and thus not informative of cell hetrogeneity. To summarise: Perform QC by fidning outlier peaks in the number of genes, the count depth and the fraction of mitochondrial reads. Consider covariates jointly instead of separately.

2.1 Normalisation

Each count in a count matrix represent successful capture, reverse transcription and sequencing of a molecule of cellular mRNA. Normalisation scales count data to obtain correct relative gene expression abundances between cells.

Most commonly used normalisation protocol is count depth scaling - 'counts per million'

summary

- Scran is recommended for normalisation of non-full-length datasets An alternative is to evaluate normalisation approache via scone especially for plate-based datasets. Full-length scRNA-seq protocols can be corrected for gene length using bulk methods.
- There is no consensus on scaling genes to 0 mean and unit variance. We prefer not to scale gene expression.
- Normalised data should be log(x+1) -transformed for use with downstream analysis methods that assume data are normally distributed.

2.2 Data correction and integration

Normalised data may still contain unwanted variability. Data correction targets furtger technical and biological covariates like batch, dropout, or cell cycle effect. We consider correction of these effects separately.

Biological effects Most common correction- remove effect of cell cycle on transcriptome. Can be performed by a simple linear regression aggainst a cell cycle score. Removing cell cycle effects can improve inference of developmental trajectories but cell cycle signals can also be informative of the biology. EG proliferating cell populations can be identified based on cell cycle scores.

Technical effects Most prominent covariates are count depth and batch. Technical count effects may remain after normalisation as no scaling method can infer expression values of genes not detected due to poor sampling. Regressing out ount depth effects can improve erformance of trajectory inference algorithms, which rely on finding transitions between cells.

Batch effects and data integration Batch effects can occur when cells are handled in dinstinct groups; eg cells on different chips, different sequencing lanes or harvested at different times.

Batch correction Correcting for batch effects between samples or cells in the same experiment from bulk RNA-seq. Typically use linear methods.

Data integration Distinguish the batch correction from the integration of data from multiple experiements. Typically use non-linear approaches.

Expression recovery Measurements contain various sources of noise, a particular aspect of this being dropout. Expression recovery has been shown to improve esitmation of gene-gene correlations.

summary

- Regress out biological covariates for trajectory inference andif other biological processes of interest are not masked dby the regressed out biological covariate.
- Regress out technical and biological covariates jointly rather than serially
- Plate-based dataset pre-processing may require regressing out ccounts, normalisation via non-linear normalisation methods or downsampling.
- Data integration and batch correction should be performed by different methods. Data integration toold may over-correct simple batch effects
- be careful of signals found only after expression recovery

2.3 Feature selection, dimensionality reduction and visualisation

A single-cell RNA-seq dataset can contain expression values for up to 25,000 genes. Many genes will mostly contain zero-counts, many will not be informative. Even after filtering out zero-count genes in QC, feature space can have over 15,000 dimensions. So it's nice to reduce dimensionaity of dataset with other tools.

Feature selection Processing of filtering out and keeping genes that are 'informative' of variability in the data. Highly variable genes (HGVs) are usually used. Genes with highest variance-to-mean ratio are selected as HGVs in each bin. This is preferably done after technical data correction,

Dimensionality Reduction The biological manifold on which cellular expressions lie can be sufficiently described by far fewer dimensions than the number of genes; dimensionality reduction aims to find these dimensions. There are two main objectives of dimensionality reduction:

Visualisation - attempt to optimally describe dataset in 2/3 dimensions. These used as coordinates on a scatter plot to obtain a visual representation of the data.

Summarisation - method of reducing data to its essential components by finding inherent dimensionality of the data - helpful for downstream analysis.

Reduced dimensions are generated with linear or non-linear gene expression vectors. Interpretability (especially in non-linear case) of reduced dimensions is sacrificed in the process. Two popular dimensionality reduction techniques that are principally summarisation methods are: Principal Component Analysis (PCA), and diffusion maps.

PCA - A linear approach, generates reduced dimensions by maximising the captured residual variance in each further dimension. It is the basis of many currently available analysis tools for clustering or trajectory inference.

- Summarises data set via its top N principal components, where N can be determined by eg 'elbow' heuristics.
- Advantage: Distances in reduced dimensional space have a consistent interpretation in all regions of this space - thus can correlate quantities of interest with principal components to assess their importance.

Diffusion maps - Non-linear data summarisation technique

- As diffusion components emphasise transitions in the data, they are principally used when continuous processes such as differentiation are of interest.
- Typically, each diffusion component i.e. diffusion map component highlights the heterogeneity of a different cell population

Visualisation

summary

- Select between 1000 and 5000 HGVs depending on complexity of dataset
- Feature selection methods using gene expression means and variances cannot be used when gene expression values have been normalised to zero mean and unit variance, or when residuals from model fitting are

used as normalised expression values. Thus, consideration of which pre-processing to perform is important before selecting HGVs.

- Dimensionality methods should be considered separately for summarisation and visualisation.

•

2.4 Stages of pre-processsed data

summary

3 Downstream analysis

Performed after pre-processing. We obtain descriptions of the underlying biological system described by the data by fitting interpretable models.

Examples:

- groups of cells with similar gene expression profiles representing cell-type clusters
- small changes in gene expression between similar cells denoting continuous (differentiation) trajectories
- genes with correlated expression profiles indicating co-regulation

Downstream analysis -can be divided into cell-level or gene-level approaches.

3.1 Cluster Analysis

The attempt to explain the heterogeneity in the data based on a categorisation of cells into groups.

Clustering Organising cells into clusters help us to infer the identity of member cells. Expression profile similarity is determined via distance metrics, which often take dimensionality-reduced representations as input.

Clustering - a classical unsupervised machine learning problem based directly on a distance matrix.

Community detection methods - grapth-partitioning algorithms and thus rely on a graph representation of single cell data. Often faster than clustering. A standard approach has become mulit-resolution modularity optimisation.

Cluster annotation On gene-level we can find gene signatures of each cluster by analysing clustered dat, labelling with a 'marker gene'.

summary

-

Compositional Analysis At cell level, we can analyse clustered data in terms of compositional structure - looking at the proportion of cells that fall into each cell-identity cluster. These proportions can change in response to disease. Statistical tests over changes in the proportion of a cell-identity cluster between samples are dependent on one another.

3.2 Trajectory analysis

We regard data as a snapshot of dynamic process and investigate the underlying process.

Trajectory inference Required in order to capture transitions between cell identities, branching differentiation processes, or gradual, unsynchronised changes in biological function, using dynamic models of gene expression.

We reconstruct the underlying processs by finding paths through cellular space that minimise transcriptional changes between neighbouring cells. Pseudotime variable describes the ordering of cells along paths

Examples of current models:

- Simple linear trajectories
- Sime bifurcating trajectories
- Complex graphs, trees or multifurcating trajectories

No individual method performs optimally for all trajectories.

Slingshot - Concluded to be the best for simple trajectories from linear to bi- and multi-furcating models.

PAGA - recommended for more complex trajectories.

TI is typically performed after clustering.

Gene expression dynamics - Genes that vary smoothly across pseudotime characterise the trajectory and can be used to identify the underlying biological process. This group of trajectory-associated genes is expected to contain the genes that regulate the modelled process.