Garvan Institute of Medical Research

Multiplexing and Normalization of single cell data

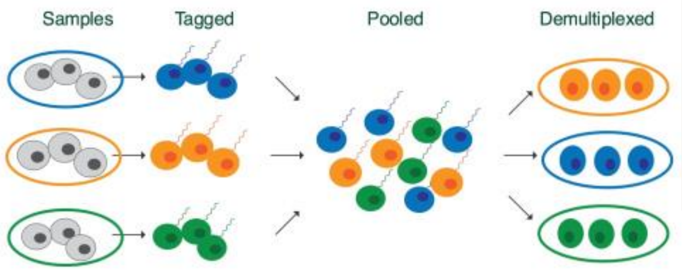
Associate Professor Joseph Powell
Director, Garvan-Weizmann Centre for Cellular Genomics
Deputy Director, UNSW Cellular Genomics Futures Institute

SISG - 2019

What is multiplexing?

Garvan Institute of Medical Research

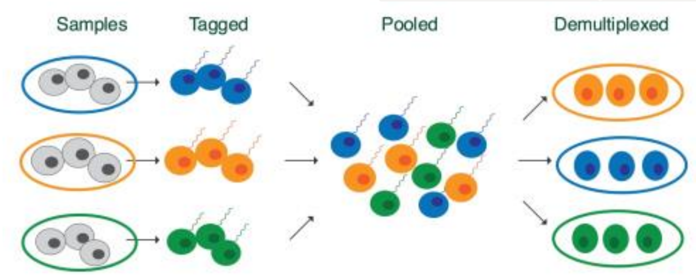
- Pooling cells from multiple samples before loading into a single cell capture



What is multiplexing?

Garvan Institute of Medical Research

- Pooling cells from multiple samples before loading into a single cell capture



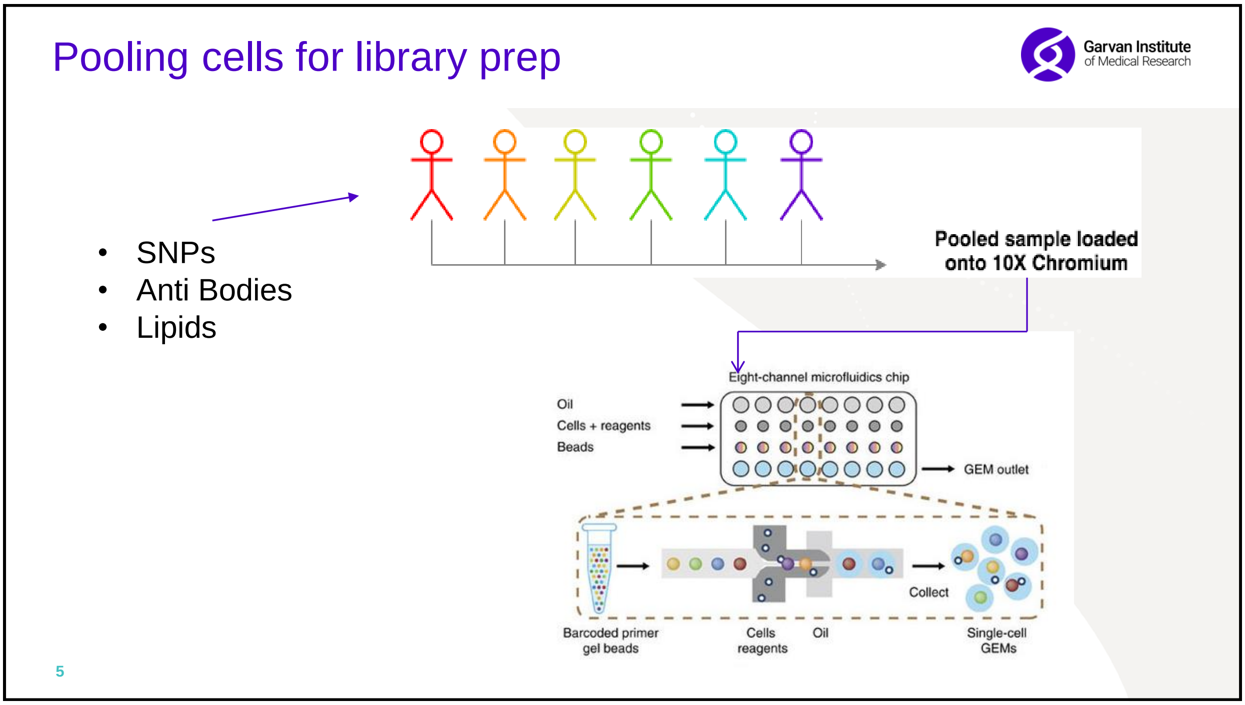
- Lower costs for library preparation
- Better doublet identification

Cell multiplexing methods

Garvan Institute of Medical Research

- SNP calling and matched genotyping (natural genetic barcoding)
- SNP calling without matched genotypes (natural genetic barcoding)
- Antibody hash tagging
- Lipid- and cholesterol-modified oligonucleotides

4

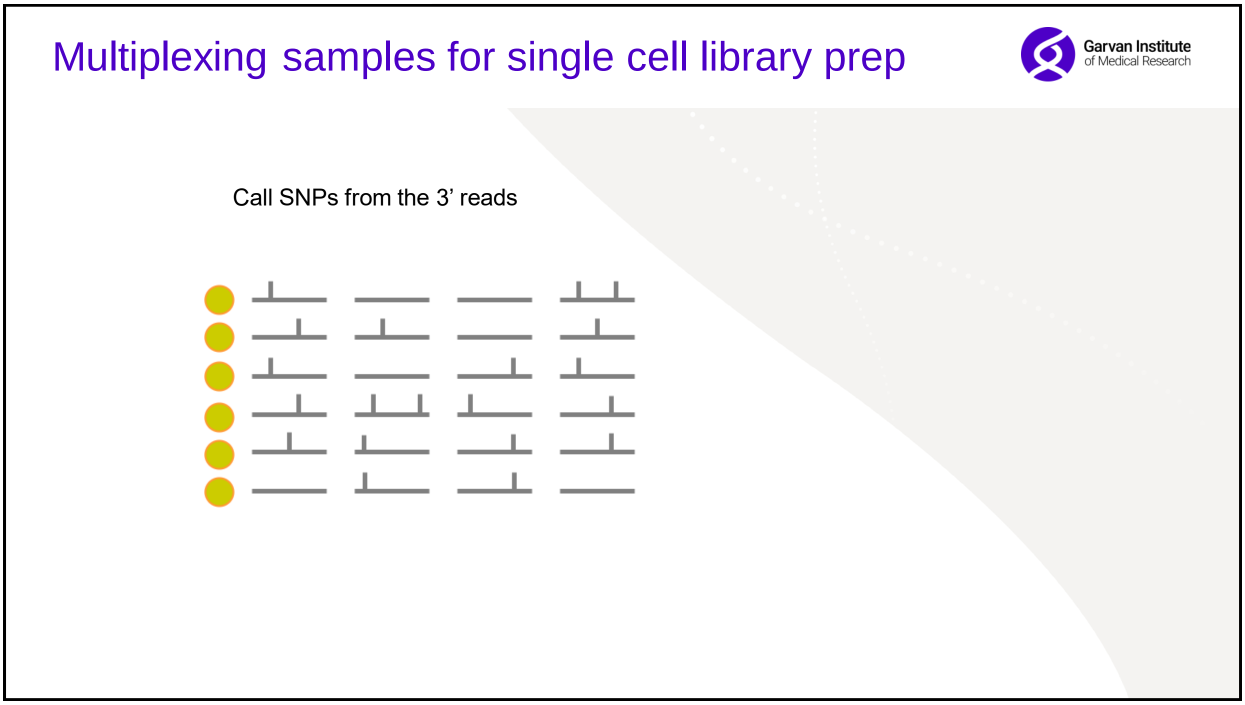


Pooling cells for library prep

Garvan Institute
of Medical Research

- SNPs
- Anti Bodies
- Lipids

Pooled sample loaded
onto 10X Chromium



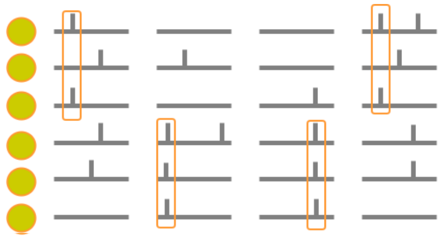
Multiplexing samples for single cell library prep

Garvan Institute
of Medical Research

Call SNPs from the 3’ reads

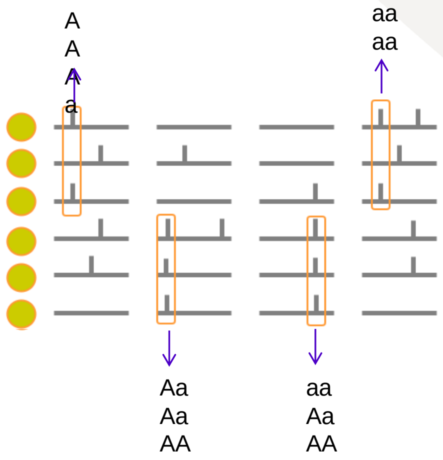
Multiplexing samples for single cell library prep

Garvan Institute of Medical Research



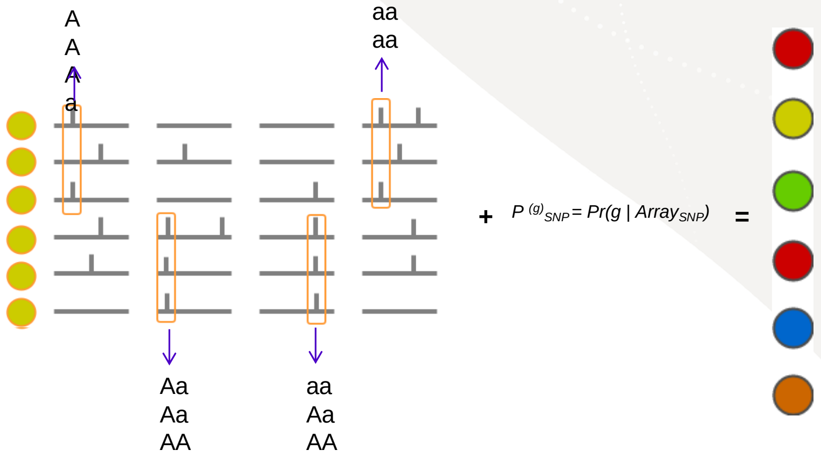
Multiplexing samples for single cell library prep

Garvan Institute of Medical Research



Multiplexing samples for single cell library prep

Garvan Institute of Medical Research

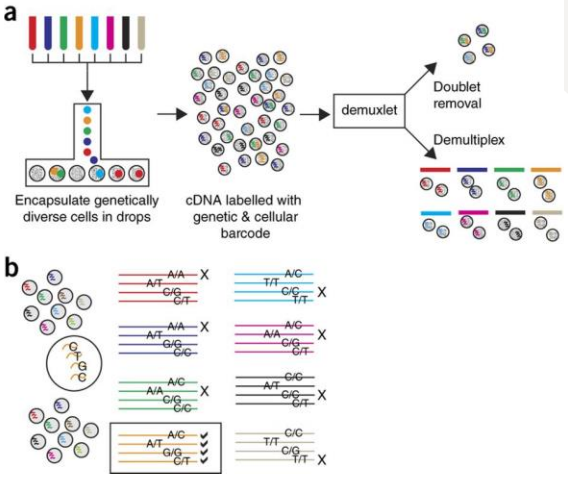


+ $P^{(g)}_{SNP} = Pr(g \mid Array_{SNP})$ =

Powell et al. Nature Reviews Genetics 2010

Demuxlet – Kang et al. Nat Biotech 2018

Garvan Institute of Medical Research



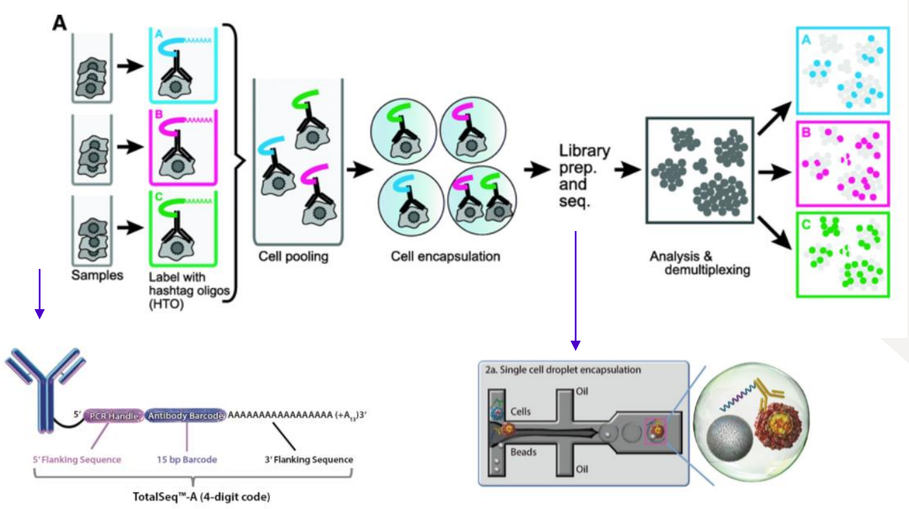
https://www.protocols.io/view/instructional-tutorial-for-using-demuxlet-233gqqn

https://github.com/statgen/demuxlet.

10

Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics

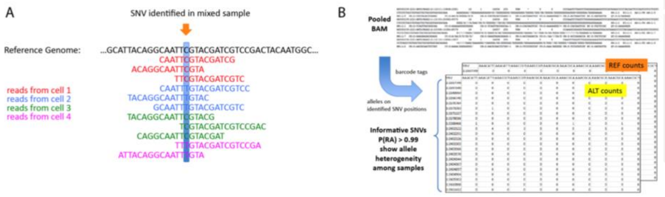
Garvan Institute of Medical Research

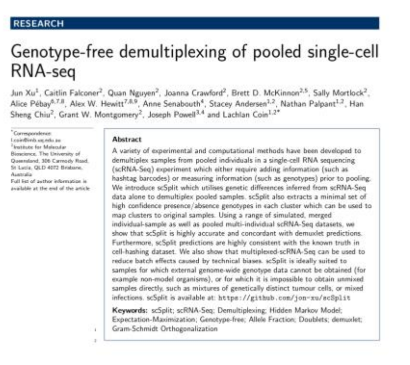

A
A
B
C
Samples
Label with hashtag oligos (HTO)
Cell pooling
Cell encapsulation
Library prep. and seq.
Analysis & demultiplexing
A
B
C
5′ PCR Handle Antibody Barcode AAAAAAAAAAAAAAAAAA (+A$_{11}$)3′
5′ Flanking Sequence
15 bp Barcode
3′ Flanking Sequence
TotalSeq™-A (4-digit code)
2a. Single cell droplet encapsulation
Oil
Cells
Beads
Oil

Stoeckius et al. Genome Biology

scSplit - Genotyping Free Demultiplexing

Garvan Institute of Medical Research




RESEARCH

Genotype-free demultiplexing of pooled single-cell RNA-seq

Jun Xu[1], Caitlin Falconer[2], Quan Nguyen[2], Joanna Crawford[2], Brett D. McKinnon[2,5], Sally Mortlock[2], Alice Pébay[6,7,8], Alex W. Hewitt[7,8,9], Anne Senabouth[4], Stacey Andersen[1,2], Nathan Palpant[1,2], Han Sheng Chiu[2], Grant W. Montgomery[2], Joseph Powell[3,4] and Lachlan Coin[1,2*]

*Correspondence: l.coin@imb.uq.edu.au
[1]Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, QLD 4072 Brisbane, Australia
Full list of author information is available at the end of the article

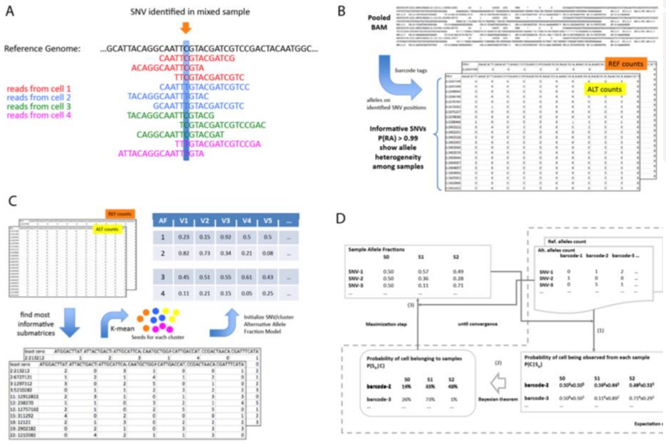
Abstract

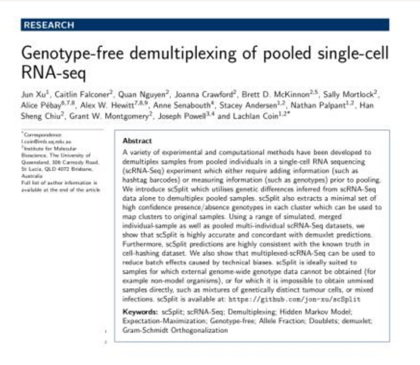
A variety of experimental and computational methods have been developed to demultiplex samples from pooled individuals in a single-cell RNA sequencing (scRNA-Seq) experiment which either require adding information (such as hashtag barcodes) or measuring information (such as genotypes) prior to pooling. We introduce scSplit which utilises genetic differences inferred from scRNA-Seq data alone to demultiplex pooled samples. scSplit also extracts a minimal set of high confidence presence/absence genotypes in each cluster which can be used to map clusters to original samples. Using a range of simulated, merged individual-sample as well as pooled multi-individual scRNA-Seq datasets, we show that scSplit is highly accurate and concordant with demuxlet predictions. Furthermore, scSplit predictions are highly consistent with the known truth in cell-hashing dataset. We also show that multiplexed-scRNA-Seq can be used to reduce batch effects caused by technical biases. scSplit is ideally suited to samples for which external genome-wide genotype data cannot be obtained (for example non-model organisms), or for which it is impossible to obtain unmixed samples directly, such as mixtures of genetically distinct tumour cells, or mixed infections. scSplit is available at: `https://github.com/jon-xu/scSplit`

Keywords: scSplit; scRNA-Seq; Demultiplexing; Hidden Markov Model; Expectation-Maximization; Genotype-free; Allele Fraction; Doublets; demuxlet; Gram-Schmidt Orthogonalization

scSplit - Genotyping Free Demultiplexing

Garvan Institute of Medical Research





RESEARCH

Genotype-free demultiplexing of pooled single-cell RNA-seq

Jun Xu[1], Caitlin Falconer[2], Quan Nguyen[2], Joanna Crawford[2], Brett D. McKinnon[2,5], Sally Mortlock[2], Alice Pébay[6,7,8], Alex W. Hewitt[7,8,9], Anne Senabouth[4], Stacey Andersen[1,2], Nathan Palpant[1,2], Han Sheng Chiu[2], Grant W. Montgomery[2], Joseph Powell[3,4] and Lachlan Coin[1,2*]

*Correspondence: l.coin@imb.uq.edu.au
[1]Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, QLD 4072 Brisbane, Australia
Full list of author information is available at the end of the article

Abstract

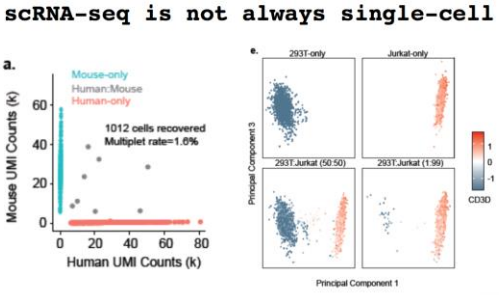
A variety of experimental and computational methods have been developed to demultiplex samples from pooled individuals in a single-cell RNA sequencing (scRNA-Seq) experiment which either require adding information (such as hashtag barcodes) or measuring information (such as genotypes) prior to pooling. We introduce scSplit which utilises genetic differences inferred from scRNA-Seq data alone to demultiplex pooled samples. scSplit also extracts a minimal set of high confidence presence/absence genotypes in each cluster which can be used to map clusters to original samples. Using a range of simulated, merged individual-sample as well as pooled multi-individual scRNA-Seq datasets, we show that scSplit is highly accurate and concordant with demuxlet predictions. Furthermore, scSplit predictions are highly consistent with the known truth in cell-hashing dataset. We also show that multiplexed-scRNA-Seq can be used to reduce batch effects caused by technical biases. scSplit is ideally suited to samples for which external genome-wide genotype data cannot be obtained (for example non-model organisms), or for which it is impossible to obtain unmixed samples directly, such as mixtures of genetically distinct tumour cells, or mixed infections. scSplit is available at: `https://github.com/jon-xu/scSplit`

Keywords: scSplit; scRNA-Seq; Demultiplexing; Hidden Markov Model; Expectation-Maximization; Genotype-free; Allele Fraction; Doublets; demuxlet; Gram-Schmidt Orthogonalization

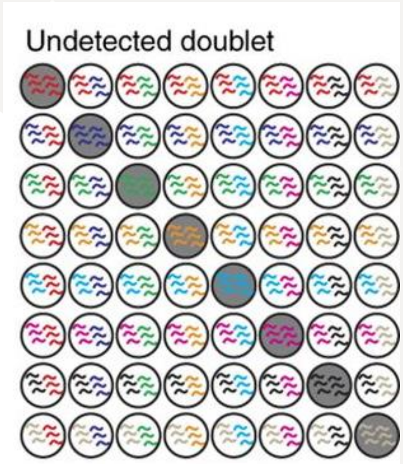
13

Detecting doublets

Garvan Institute of Medical Research

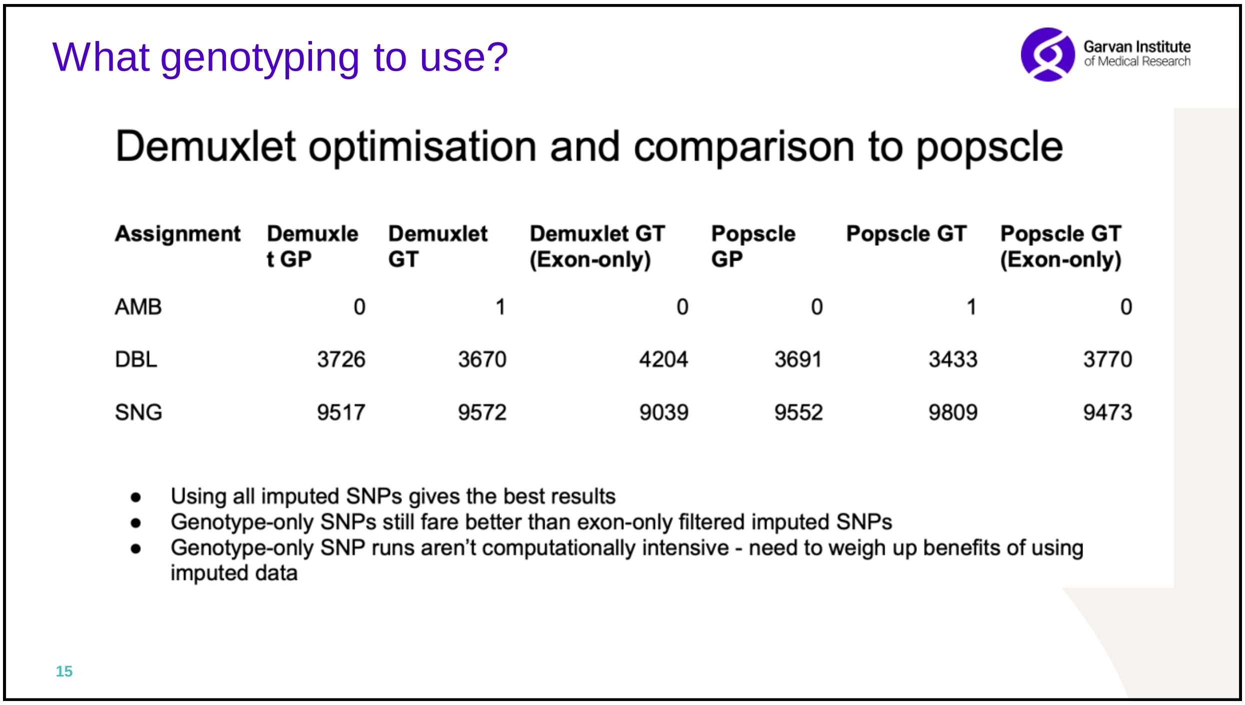


Each color comes from a different individual/sample



Demuxlet seems to over estimate doublets

14



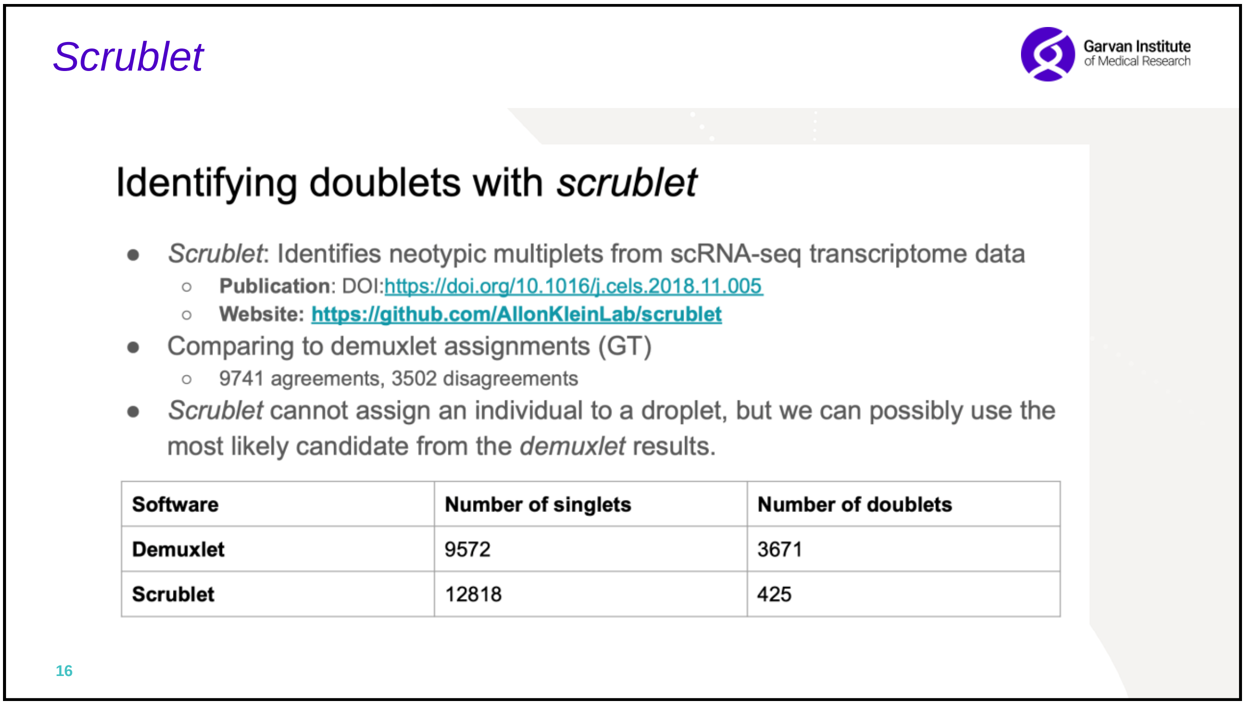
What genotyping to use?

Demuxlet optimisation and comparison to popscle

Assignment	Demuxlet GP	Demuxlet GT	Demuxlet GT (Exon-only)	Popscle GP	Popscle GT	Popscle GT (Exon-only)
AMB	0	1	0	0	1	0
DBL	3726	3670	4204	3691	3433	3770
SNG	9517	9572	9039	9552	9809	9473

- Using all imputed SNPs gives the best results
- Genotype-only SNPs still fare better than exon-only filtered imputed SNPs
- Genotype-only SNP runs aren't computationally intensive - need to weigh up benefits of using imputed data

15



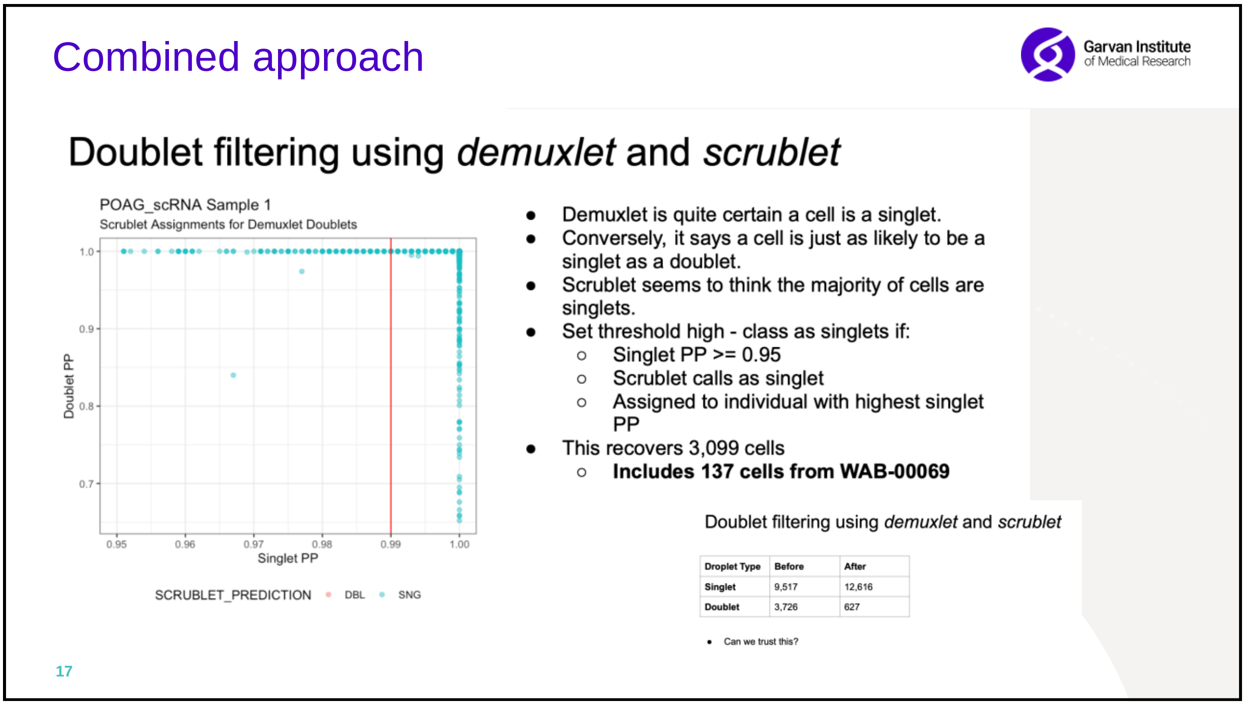
Scrublet

Identifying doublets with *scrublet*

- *Scrublet*: Identifies neotypic multiplets from scRNA-seq transcriptome data
  - **Publication**: DOI:https://doi.org/10.1016/j.cels.2018.11.005
  - **Website: https://github.com/AllonKleinLab/scrublet**
- Comparing to demuxlet assignments (GT)
  - 9741 agreements, 3502 disagreements
- *Scrublet* cannot assign an individual to a droplet, but we can possibly use the most likely candidate from the *demuxlet* results.

Software	Number of singlets	Number of doublets
Demuxlet	9572	3671
Scrublet	12818	425

16



Combined approach

Doublet filtering using *demuxlet* and *scrublet*

POAG_scRNA Sample 1
Scrublet Assignments for Demuxlet Doublets

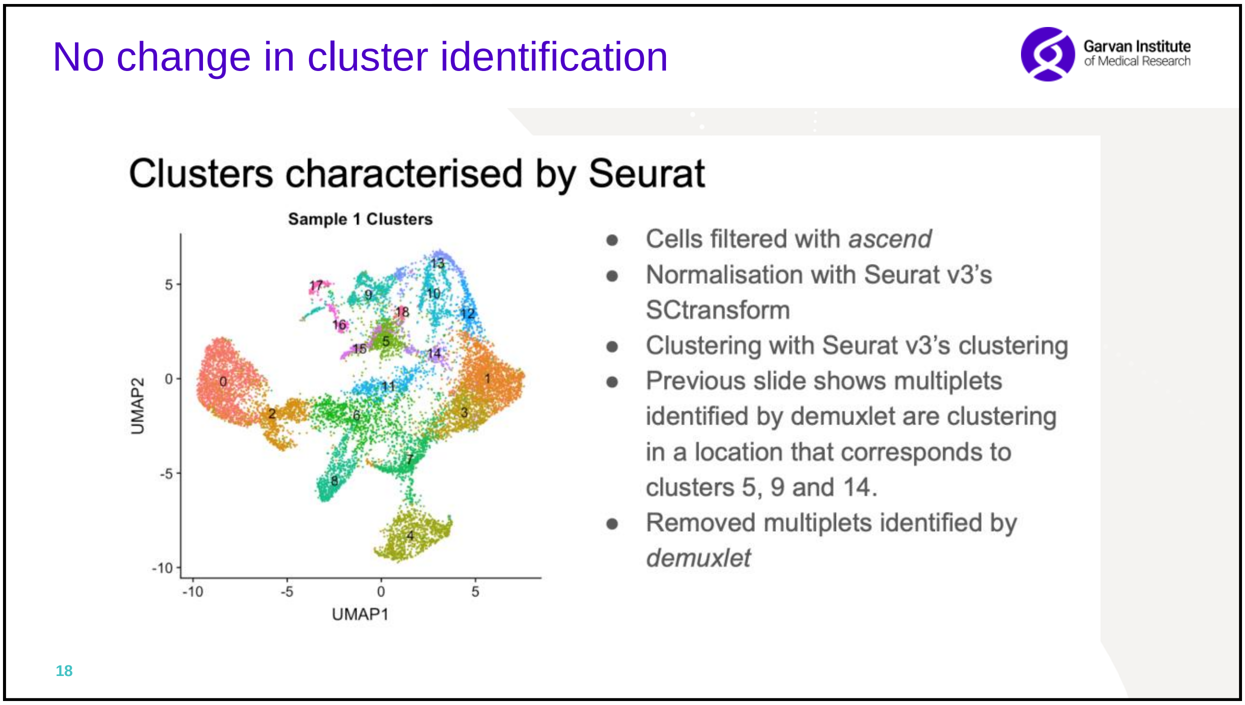
- Demuxlet is quite certain a cell is a singlet.
- Conversely, it says a cell is just as likely to be a singlet as a doublet.
- Scrublet seems to think the majority of cells are singlets.
- Set threshold high - class as singlets if:
  - Singlet PP >= 0.95
  - Scrublet calls as singlet
  - Assigned to individual with highest singlet PP
- This recovers 3,099 cells
  - **Includes 137 cells from WAB-00069**

Doublet filtering using *demuxlet* and *scrublet*

Droplet Type	Before	After
Singlet	9,517	12,616
Doublet	3,726	627

- Can we trust this?

17



No change in cluster identification

Clusters characterised by Seurat

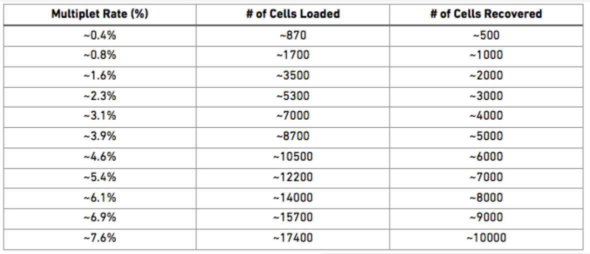
- Cells filtered with *ascend*
- Normalisation with Seurat v3's SCtransform
- Clustering with Seurat v3's clustering
- Previous slide shows multiplets identified by demuxlet are clustering in a location that corresponds to clusters 5, 9 and 14.
- Removed multiplets identified by *demuxlet*

18

Effects on cost

Garvan Institute of Medical Research

- Costs for generating 1 library 10x (1x chip, 1x reagent)
  - List price: $2,830-3,215
  - Doublet rate



Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~870	~500
~0.8%	~1700	~1000
~1.6%	~3500	~2000
~2.3%	~5300	~3000
~3.1%	~7000	~4000
~3.9%	~8700	~5000
~4.6%	~10500	~6000
~5.4%	~12200	~7000
~6.1%	~14000	~8000
~6.9%	~15700	~9000
~7.6%	~17400	~10000

  - Sequencing to 50,000 reads per cell
    - NovaSeq S4 = $0.25
    - NextSeq = $0.50

- Cost per sample (3,000 cells) = $2,830+750=$3,580

19

Effects on cost

Garvan Institute of Medical Research

- Pool cells from 20 samples
- Aim for 20,000 cells = ~**1,000** cells per sample
- Cost per sample
  - Library prep =$280
  - SNP Chip =$47
  - Sequencing =$150
  - **Total** =$477

- Run all 8 lanes on a 10x chip =$427
- Use BGI sequencer (~$75) =$352

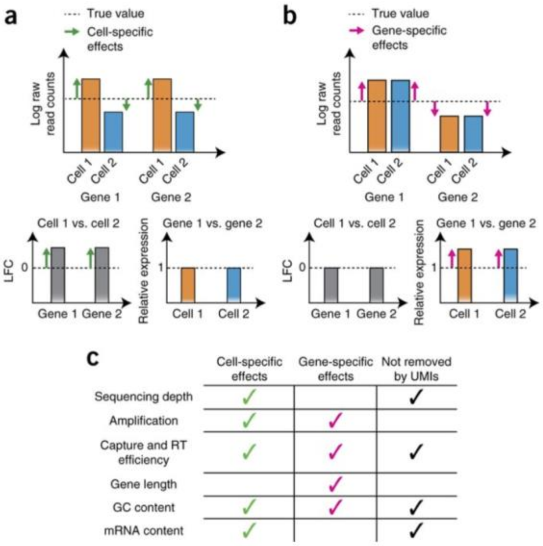
20

Normalization

21

Normalization – cell and gene level

This is the next stage of normalization after the ‘batch’ effects have been accounted for



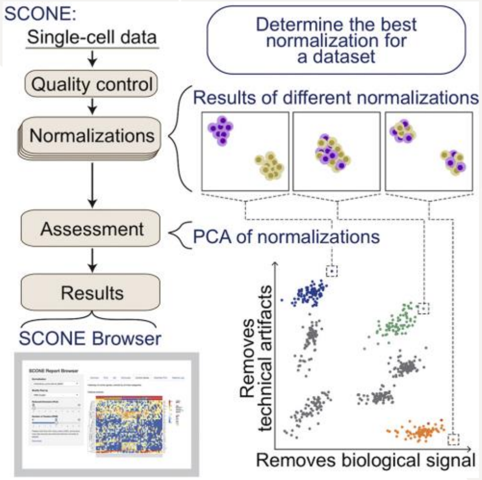
	Cell-specific effects	Gene-specific effects	Not removed by UMIs
Sequencing depth	✓		✓
Amplification	✓	✓	
Capture and RT efficiency	✓	✓	✓
Gene length		✓	
GC content	✓	✓	✓
mRNA content	✓		✓

22 Vellejos et al. Nature Methods, 2017

Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq

Flexible generalized regression method

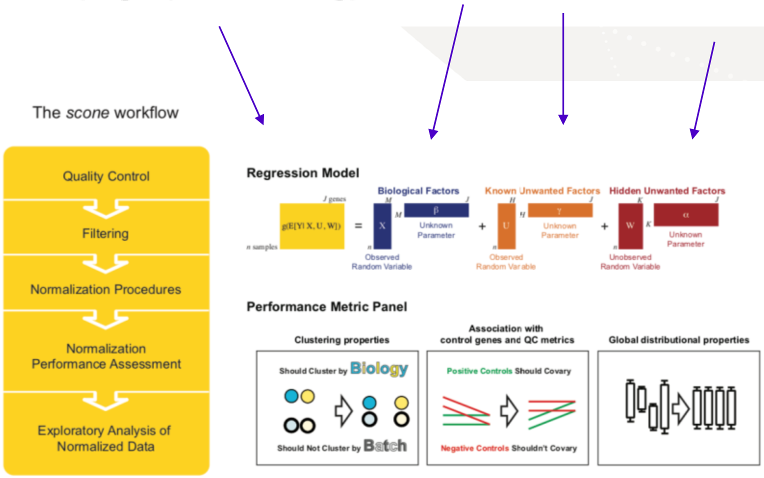
Implemented in the *scone* package


SCONE:
Single-cell data
Quality control
Normalizations
Assessment
Results
SCONE Browser
Determine the best normalization for a dataset
Results of different normalizations
PCA of normalizations
Removes technical artifacts
Removes biological signal

23 Cole et al. Cell Systems 2018

scone normalization

$$g(\mathrm{E}[Y|X,U,W]) = X\beta + U\gamma + W\alpha,$$


The *scone* workflow
Quality Control
Filtering
Normalization Procedures
Normalization Performance Assessment
Exploratory Analysis of Normalized Data
Regression Model
Biological Factors
Known Unwanted Factors
Hidden Unwanted Factors
J genes
n samples
g(E[Y| X, U, W])
X
β
U
γ
W
α
Unknown Parameter
Observed Random Variable
Unobserved Random Variable
Performance Metric Panel
Clustering properties
Association with control genes and QC metrics
Global distributional properties
Should Cluster by Biology
Should Not Cluster by Batch
Positive Controls Should Covary
Negative Controls Shouldn't Covary

24

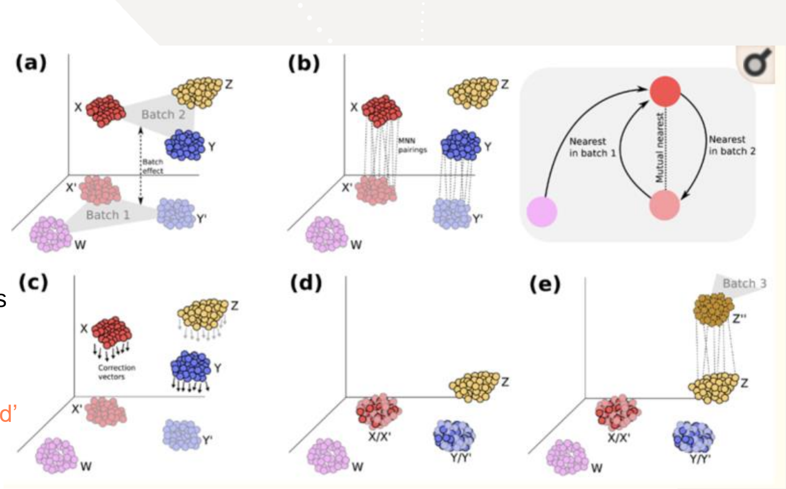
Normalization by nearest neighbor

Garvan Institute of Medical Research

Batch correction based on the detection of mutual nearest neighbours (MNNs) in the high-dimensional expression space

Makes the assumption that cells are similar between batches

If there are unique cell types in a sample then they will be 'adjusted' Closer to other cell types.



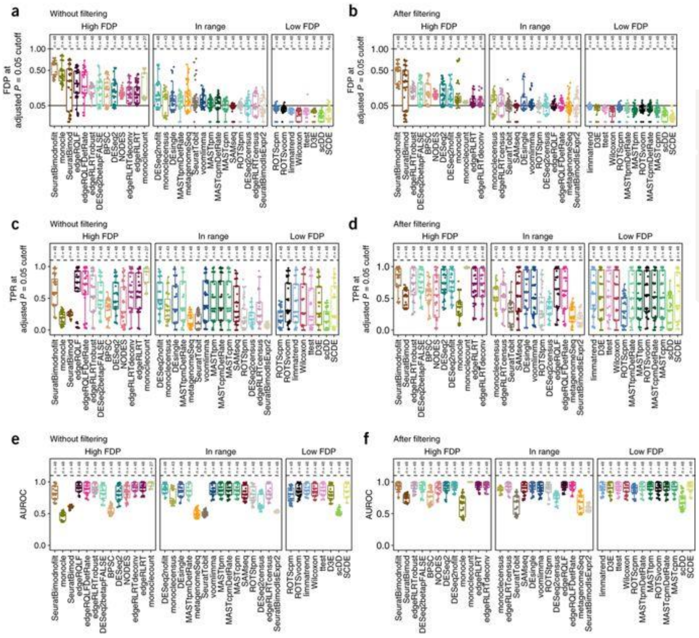
Haghverdi et al. Nature Biotechnology 2018

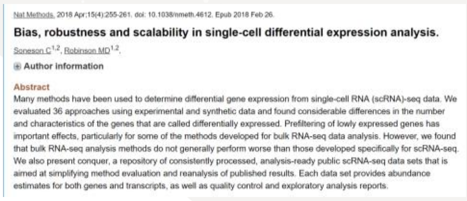
Resources

Garvan Institute of Medical Research

- *ascend*: Senabouth et al. GigaScience 2019 (https://github.com/powellgenomicslab/ascend)
- seurat: Butler *et al*. Nature Biotechnology 2018 (https://github.com/satijalab/seurat)
- scater: McCarthy *et al*. Bioinformatics 2017 (https://bioconductor.org/packages/release/bioc/html/scater.html)

Garvan Institute of Medical Research





Nat Methods. 2018 Apr;15(4):255-261. doi: 10.1038/nmeth.4612. Epub 2018 Feb 26.

Bias, robustness and scalability in single-cell differential expression analysis.

Soneson C[1,2], Robinson MD[1,2].

⊕ Author information

Abstract

Many methods have been used to determine differential gene expression from single-cell RNA (scRNA)-seq data. We evaluated 36 approaches using experimental and synthetic data and found considerable differences in the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes has important effects, particularly for some of the methods developed for bulk RNA-seq data analysis. However, we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq. We also present conquer, a repository of consistently processed, analysis-ready public scRNA-seq data sets that is aimed at simplifying method evaluation and reanalysis of published results. Each data set provides abundance estimates for both genes and transcripts, as well as quality control and exploratory analysis reports.

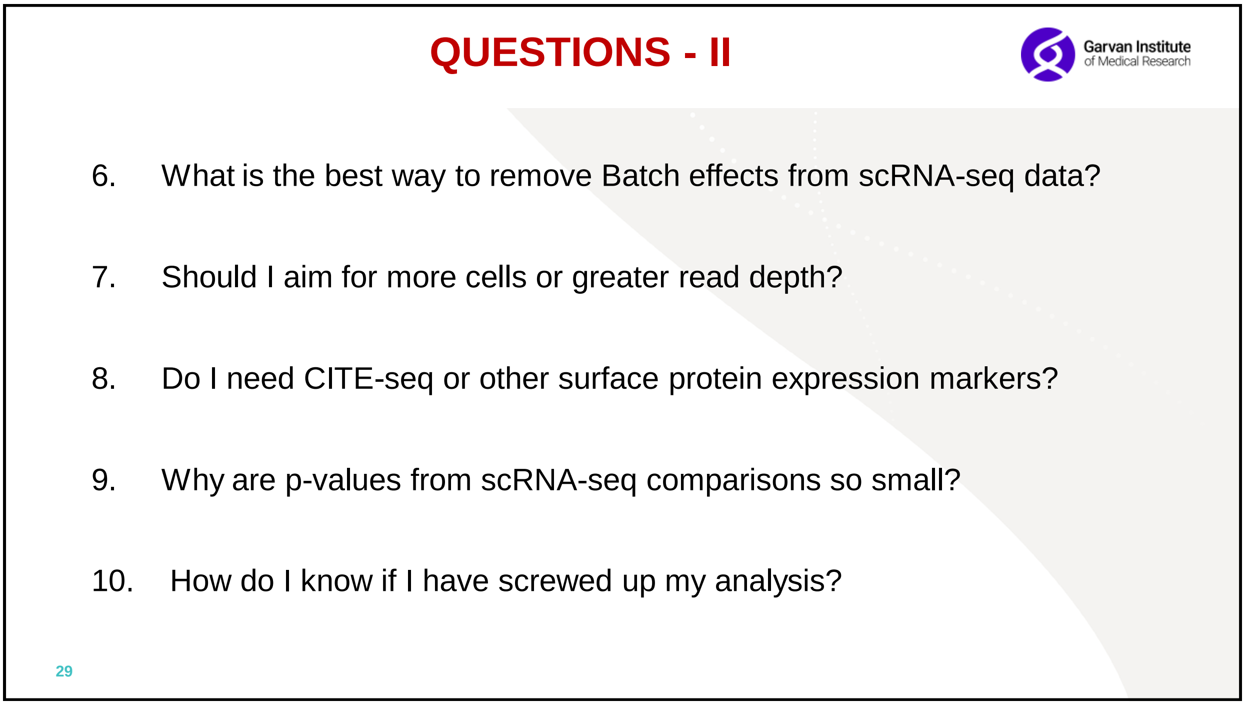
27

QUESTIONS - I

Garvan Institute of Medical Research

1. EdgeR or DESeq2 or … ?
2. Seurat or Monocle or … ?
3. tSNE or UMAP or … ?
4. What is the biggest impediment to robust bulk RNA-seq analysis?
5. What is the biggest impediment to robust scRNA-seq analysis?

28



QUESTIONS - II

Garvan Institute of Medical Research

6. What is the best way to remove Batch effects from scRNA-seq data?

7. Should I aim for more cells or greater read depth?

8. Do I need CITE-seq or other surface protein expression markers?

9. Why are p-values from scRNA-seq comparisons so small?

10. How do I know if I have screwed up my analysis?