scRNA-seq: Challenges

Lähnemann *et al. Genome Biology* (2020) 21:31
https://doi.org/10.1186/s13059-020-1926-6

Genome Biology

REVIEW **Open Access**

Eleven grand challenges in single-cell data science

David Lähnemann[1,2,3], Johannes Köster[1,4], Ewa Szczurek[5], Davis J. McCarthy[6,7], Stephanie C. Hicks[8], Mark D. Robinson[9], Catalina A. Vallejos[10,11], Kieran R. Campbell[12,13,14], Niko Beerenwinkel[15,16], Ahmed Mahfouz[17,18], Luca Pinello[19,20,21], Pavel Skums[22], Alexandros Stamatakis[23,24], Camille Stephan-Otto Attolini[25], Samuel Aparicio[13,26], Jasmijn Baaijens[27], Marleen Balvert[27,28], Buys de Barbanson[29,30,31], Antonio Cappuccio[32], Giacomo Corleone[33], Bas E. Dutilh[28,34], Maria Florescu[29,30,31], Victor Guryev[35], Rens Holmer[36], Katharina Jahn[15,16], Thamar Jessurun Lobo[35], Emma M. Keizer[37], Indu Khatri[38], Szymon M. Kielbasa[39], Jan O. Korbel[40], Alexey M. Kozlov[23], Tzu-Hao Kuo[3], Boudewijn P.F. Lelieveldt[41,42], Ion I. Mandoiu[43], John C. Marioni[44,45,46], Tobias Marschall[47,48], Felix Mölder[1,49], Amir Niknejad[50,51], Lukasz Raczkowski[5], Marcel Reinders[17,18], Jeroen de Ridder[29,30], Antoine-Emmanuel Saliba[52], Antonios Somarakis[42], Oliver Stegle[40,46,53], Fabian J. Theis[54], Huan Yang[55], Alex Zelikovsky[56,57], Alice C. McHardy[3], Benjamin J. Raphael[58], Sohrab P. Shah[59] and Alexander Schönhuth[27,28]*

Un/solved Problems in scRNA-Seq

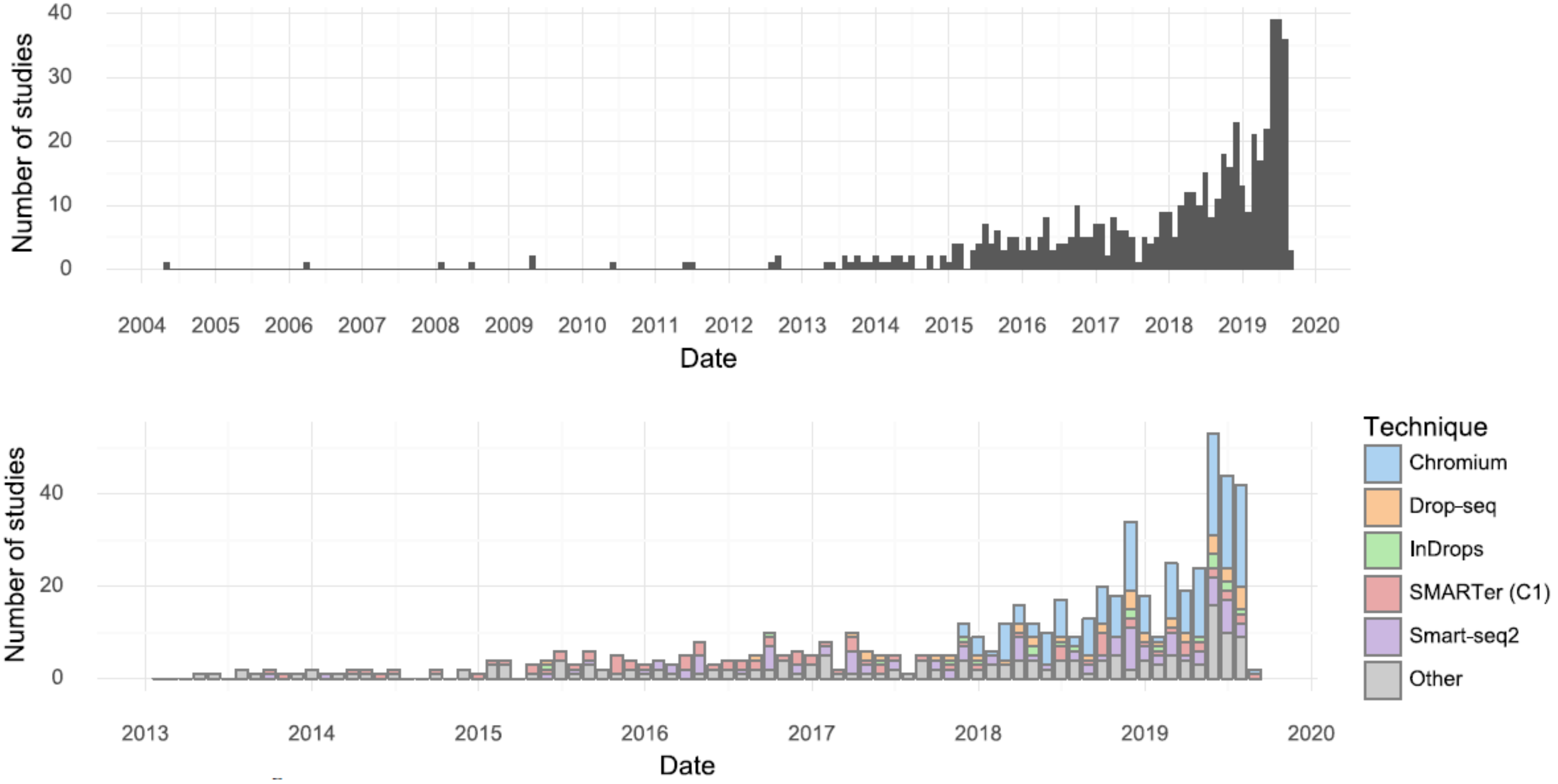
- How should the "atlas" be represented?
- How should clusters be determined?
- How should (marker) genes be identified?

scRNA-seq Atlas Projects

- Human Cell Atlas (humancellatlas.org)
- JingleBells (jinglebells.bgu.ac.il)
- Conquer (imlspenticton.uzh.ch:3838/conquer)
- PanglaoDB (panglaodb.se)
- Single Cell Expression Atlas (ebi.ac.uk/gxa/sc)
- Single Cell Portal (singlecell.broadinstitute.org)

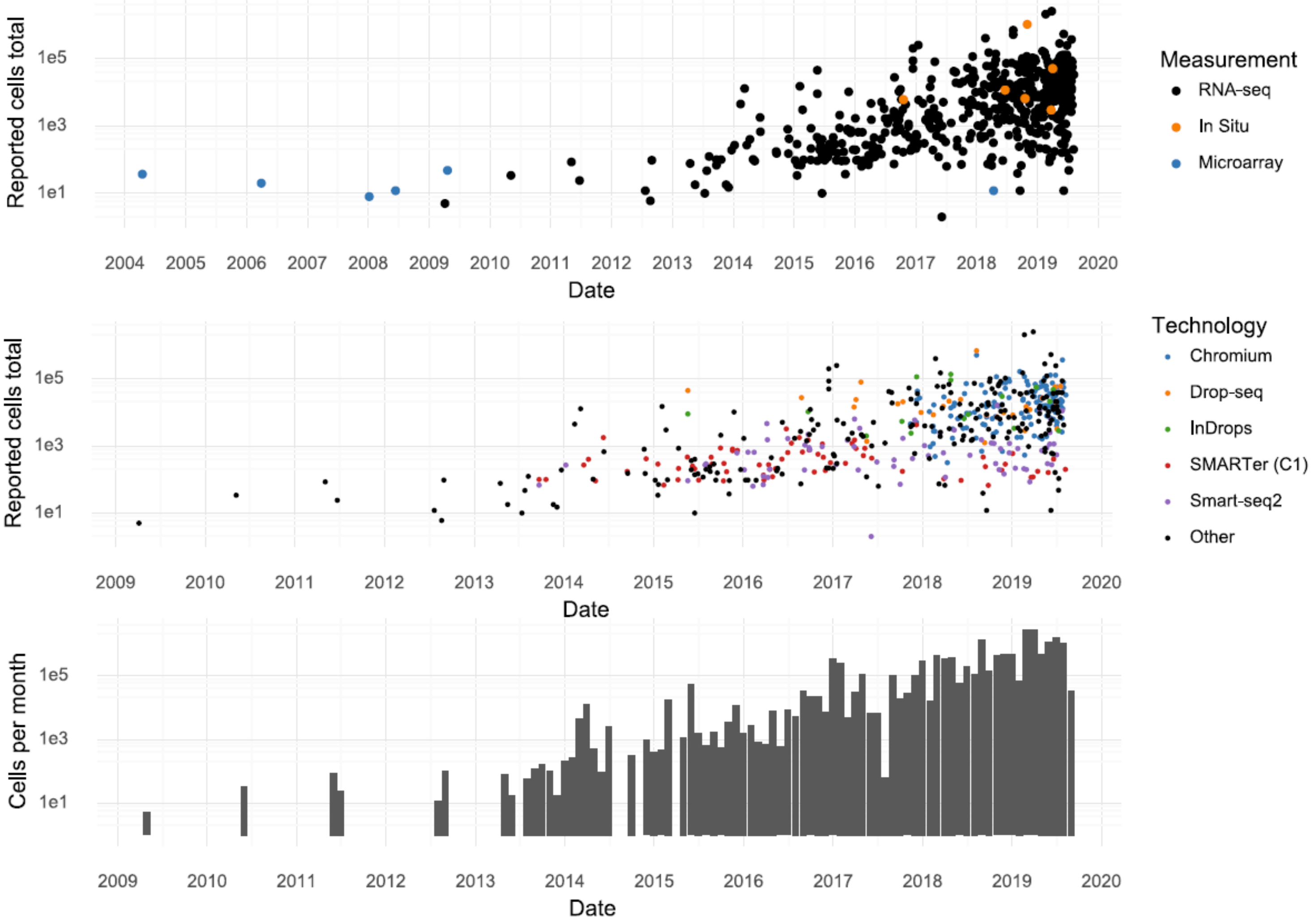
Svensson, V., et al. bioRxiv (2019)

Recent Publications and Technology



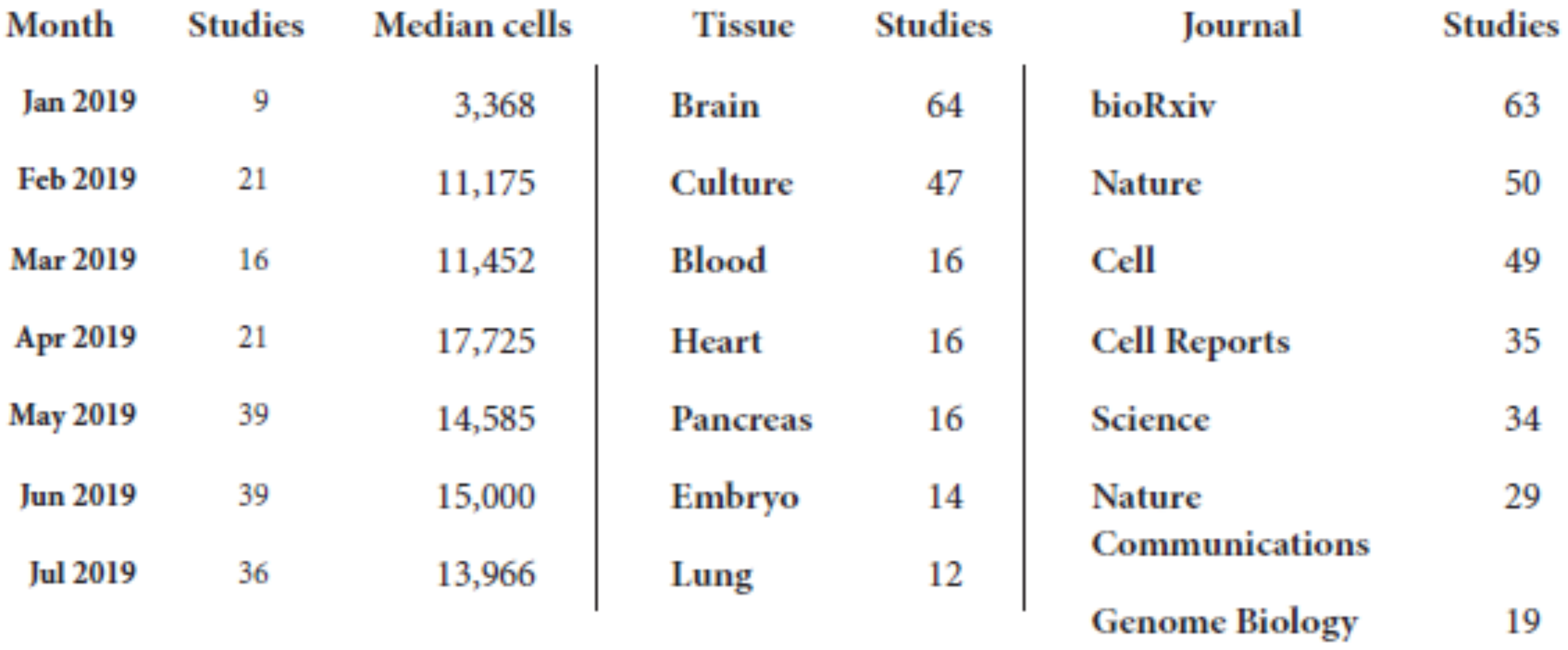
Svensson, V., et al. bioRxiv (2019)

Trends in the Studies



Svensson, V., et al. bioRxiv (2019)

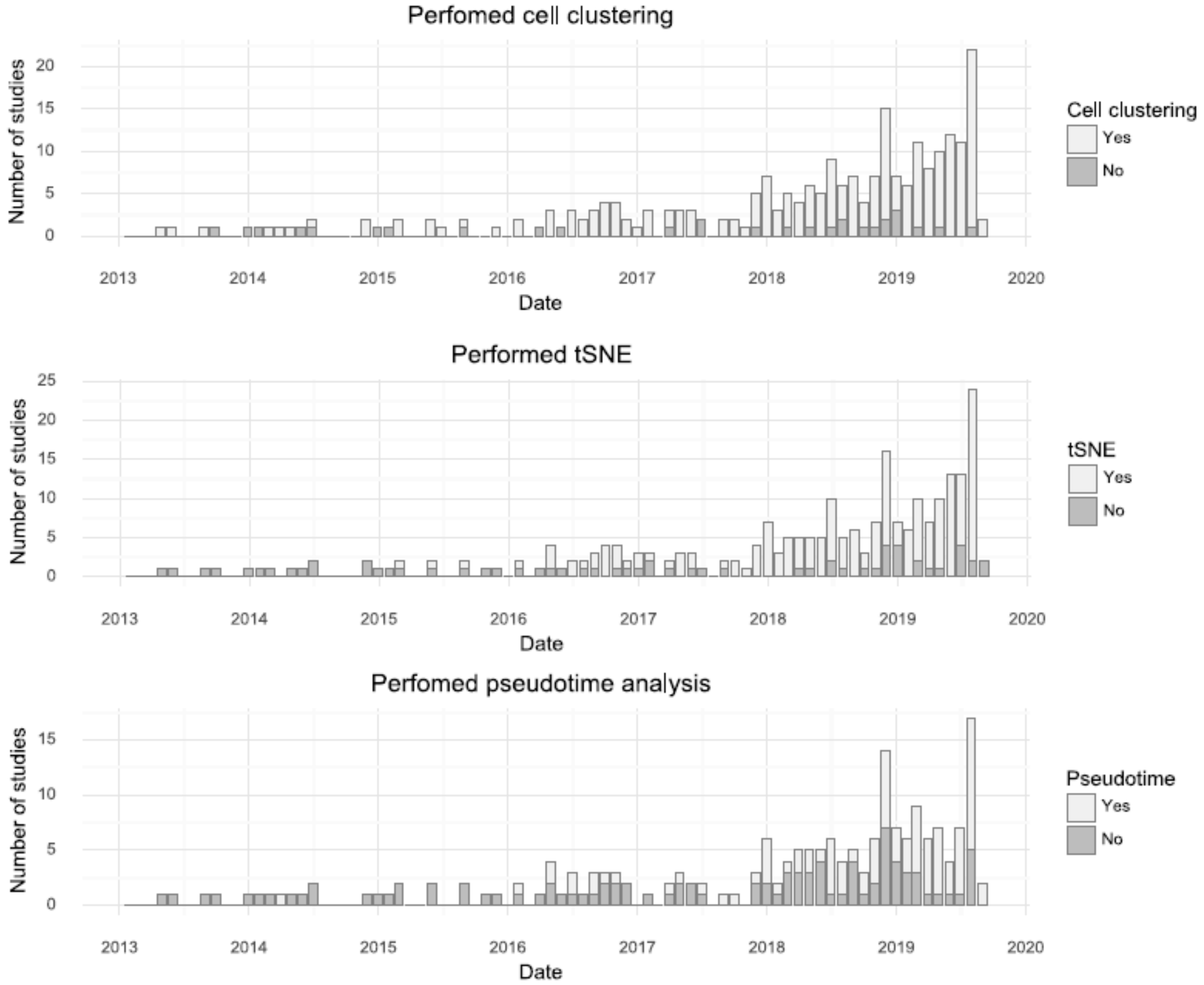
Trends in the Studies



Month	Studies	Median cells	Tissue	Studies	Journal	Studies
Jan 2019	9	3,368	Brain	64	bioRxiv	63
Feb 2019	21	11,175	Culture	47	Nature	50
Mar 2019	16	11,452	Blood	16	Cell	49
Apr 2019	21	17,725	Heart	16	Cell Reports	35
May 2019	39	14,585	Pancreas	16	Science	34
Jun 2019	39	15,000	Embryo	14	Nature Communications	29
Jul 2019	36	13,966	Lung	12		
					Genome Biology	19

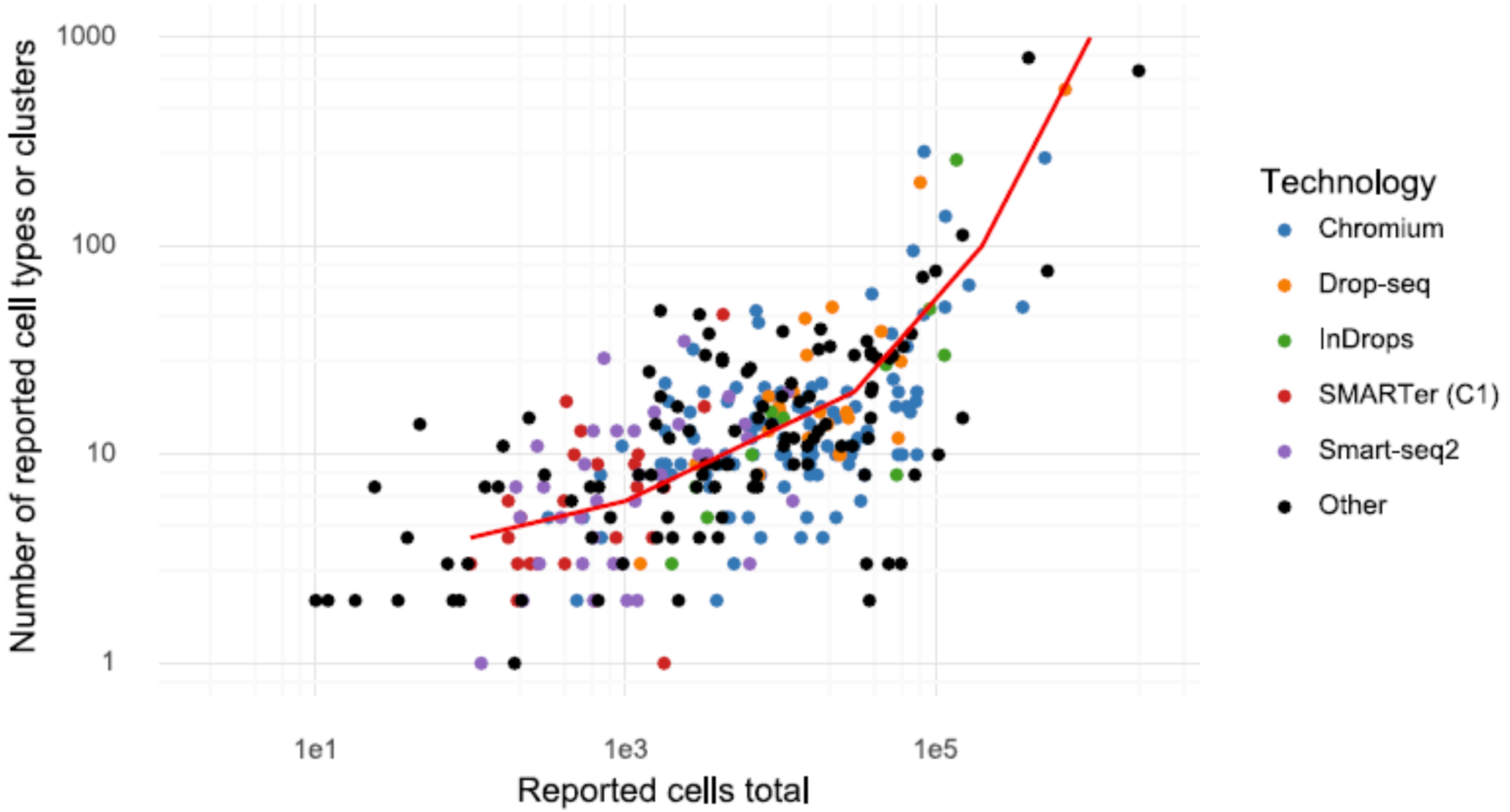
Svensson, V., et al. bioRxiv (2019)

Trends in the Studies: Analysis



Svensson, V., et al. bioRxiv (2019)

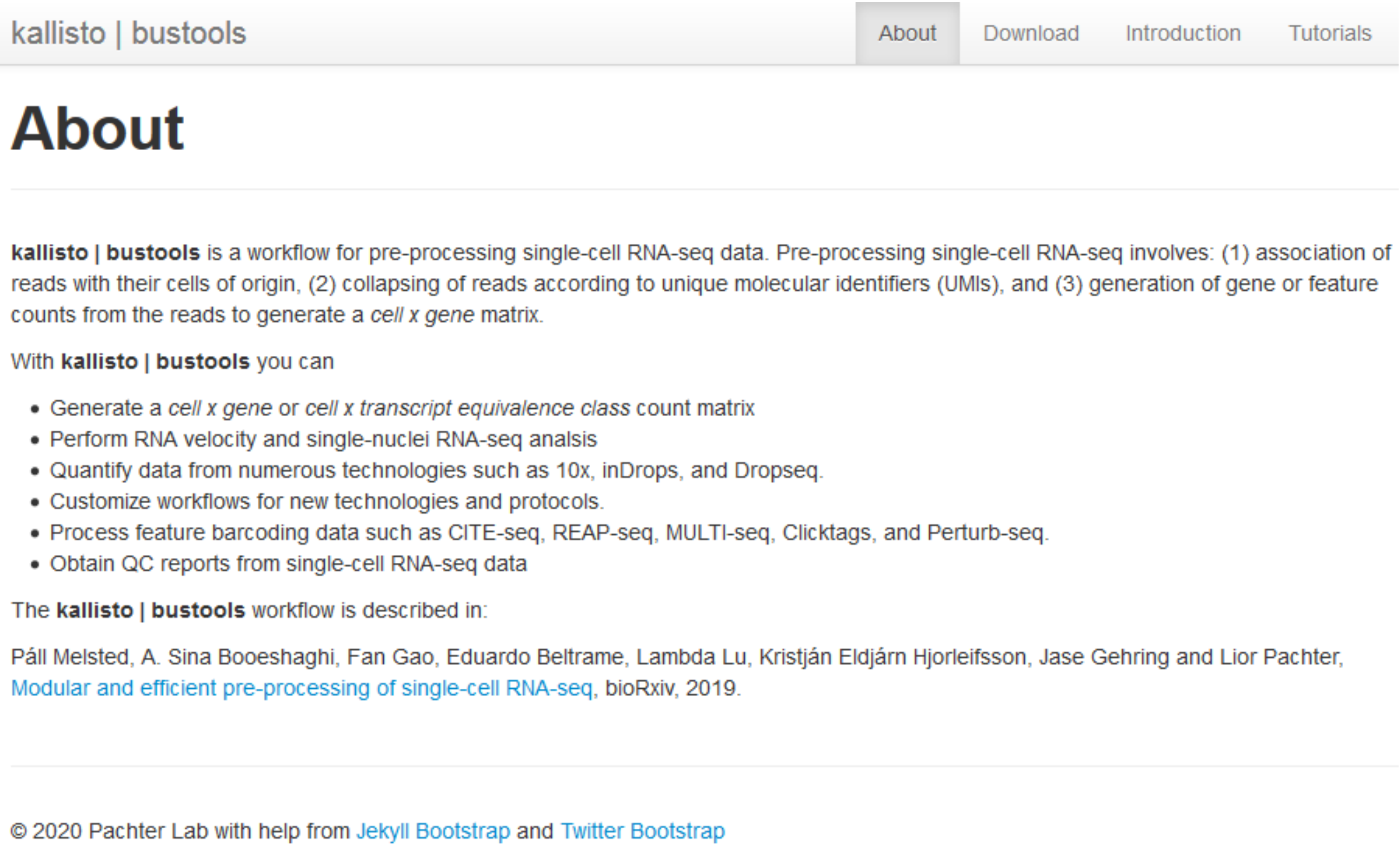
Trends in the Studies: Number of Cells vs Clusters



Svensson, V., et al. bioRxiv (2019)

colab

bustools



About

kallisto | bustools is a workflow for pre-processing single-cell RNA-seq data. Pre-processing single-cell RNA-seq involves: (1) association of reads with their cells of origin, (2) collapsing of reads according to unique molecular identifiers (UMIs), and (3) generation of gene or feature counts from the reads to generate a *cell x gene* matrix.

With **kallisto | bustools** you can

- Generate a *cell x gene* or *cell x transcript equivalence class* count matrix
- Perform RNA velocity and single-nuclei RNA-seq analsis
- Quantify data from numerous technologies such as 10x, inDrops, and Dropseq.
- Customize workflows for new technologies and protocols.
- Process feature barcoding data such as CITE-seq, REAP-seq, MULTI-seq, Clicktags, and Perturb-seq.
- Obtain QC reports from single-cell RNA-seq data

The **kallisto | bustools** workflow is described in:

Páll Melsted, A. Sina Booeshaghi, Fan Gao, Eduardo Beltrame, Lambda Lu, Kristján Eldjárn Hjorleifsson, Jase Gehring and Lior Pachter, Modular and efficient pre-processing of single-cell RNA-seq, bioRxiv, 2019.

kallistobus.tools

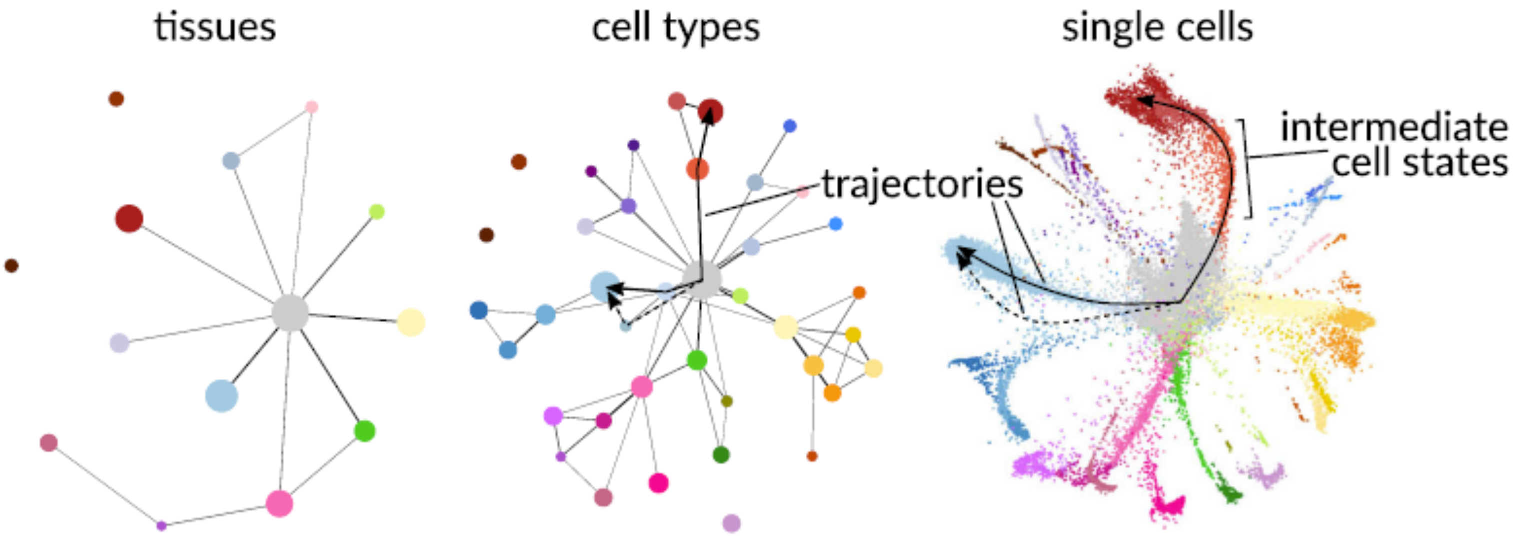
Grand Challenges

- Single-cell transcriptomics
  - Sparsity
  - Flexible statistical frameworks
  - Reference atlas
  - Trajectory
  - Spatial data
- Single-cell genomics
  - Errors and missing data
- Single-cell phylogenomics
  - Scaling phylogenetic models
  - Integrating
  - Inferring pop genetic parameters
- Overarching
  - Integration of sc data
  - Validating/benchmarking analysis tools

Common Themes

- Not specific to sc-seq
  - Quantify uncertainty
  - Benchmark methods systematically
- Specific to sc-seq
  - Scale to higher dimension data
  - Level of (sc) resolution

Levels of Resolutions



I: Handling Sparsity in scRNA-seq

- "dropout": use it only for technical-effect
- Status:
  - Statistical models (recommended)
  - Imputation
    - Model-based imputation methods
    - Data smoothing methods
    - Data reconstruction methods
- Open problems:
  - circularity

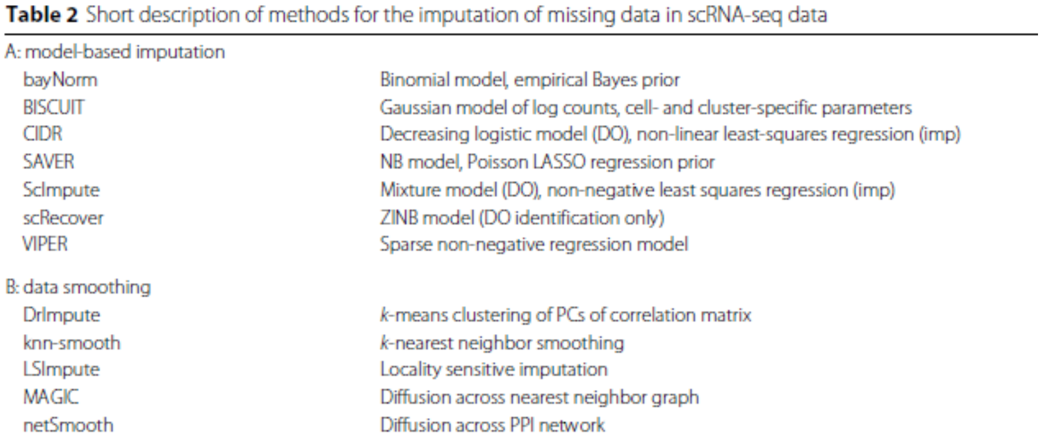


Table 2 Short description of methods for the imputation of missing data in scRNA-seq data

A: model-based imputation	
bayNorm	Binomial model, empirical Bayes prior
BISCUIT	Gaussian model of log counts, cell- and cluster-specific parameters
CIDR	Decreasing logistic model (DO), non-linear least-squares regression (imp)
SAVER	NB model, Poisson LASSO regression prior
ScImpute	Mixture model (DO), non-negative least squares regression (imp)
scRecover	ZINB model (DO identification only)
VIPER	Sparse non-negative regression model
B: data smoothing	
DrImpute	*k*-means clustering of PCs of correlation matrix
knn-smooth	*k*-nearest neighbor smoothing
LSImpute	Locality sensitive imputation
MAGIC	Diffusion across nearest neighbor graph
netSmooth	Diffusion across PPI network

II: Defining flexible statistical frameworks for discovering complex differential patters in gene expression

- Status:
  - Most methods assume groups of cells to be compared are known *a priori.*
  - Pseudo-bulk analysis
- Open problems
  - Account for uncertainty
  - Integrative approach: simultaneously perform clustering and differential testing

III: Mapping single cells to a reference atlas

- Need for classifying cells into cell types/states
  - intermediate states
- Status:
  - Reference-free approaches: unsupervised clustering
  - Manual annotation
- Open problems:
  - Mapping cells/profiles on to reference (atlas)

III: Reference Atlas

Organism	Scale of Cell Atlas	Ref/Links
Nematode (C.*elegans*)	Whole organism	atlas.gs.washington.edu SOMA data portal (incl. other organisms)
Planaria (S.*mediterranea*)	Whole organism	Fincher, C.T., et al. (2019) radiant.wi.mit.edu Plass, M., et al. (2018) shiny.mdc-berlin.de (incl. other organisms)
Fruit fly	Whole organism (emb.)	Karaiskos, N. et al. (2017) flycellatlas.org
Zebrafish	Whole organism (emb.)	Farrell, J.A., et al. (2018) cells.ucsc.edu (UCSC Cell Browser, incl. others)
Frog	Whole organism (emb.)	tinyurl.com/scXen2018 kleintools.hms.harvard.edu/tools/spring.html
Mouse	Whole (adult/brain)	dropviz.org (brain) mousebrain.org tabula-muris.ds.czbiohub.org bis.zju.edu.cn/MCA portal.brain-map.org (incl. human)

Adapted from Table 3

IV: Generalizing trajectory inference

- Continuous/dynamic changes in cell types/states
- Status
  - infer pseudotime (incl. branching trajectories)
    - MST
    - Curve/graph fitting
    - Random walks
    - Diffusion
- Open problems:
  - Using/integrating other non-transcriptomic data, e.g. methylation or chromosome acc. (scATAC-seq)
  - Assess the different methods robustly, including suitable metrics

V: Finding patterns in spatially resolved measurements

- Retaining spatial coordinates of cell, or transcripts, within a tissue
- Status:
  - Slide-seq
  - starMAP
  - SeqFISH/MERFISH
- Open problems:
  - integrating spatial information

VI: Dealing with errors and missing data in the identification of variation from sc DNA sequencing

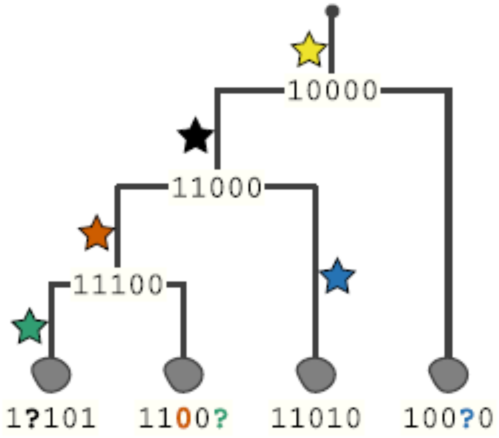
- Track somatic evolution at single cell resolution
- Errors introduced in the WGA process
- Status:
  - SNV: Monovar, SCcaller, SCAN-SNV
  - CNV: Aneufinder, Ginkgo
- Open problems:
  - incorporate WGA errors/bias
  - indel callers
  - benchmarks

VII: Scaling phyogenetic models to many cells and sites

- Inference of phylogenetic trees
- Leaves/taxa will represent cells or subclones
- Open problems:
  - Most population genetics methods will work on a maximum of ~20 cells.

VIII: integrating multiple types of variation into phylogenetic models

- Incorporate SNVs, small/large indels, and CNVs.
- Status:
  - SNVs: OncoNEM, SCITE, SiFit, SciCloneFit
- Open problems
  - integrating CNV or indel callers



IX: inferring population genetic parameters of tumor het. by model integration

- Mathematical models of tumor evolution
- Status:
  - no specific software
  - analyzing tumor subclones as populations
- Open problems:
  - Integrate spatial information, esp from other studies, are subclones co-located?
  - Incorporate other parameters such as,

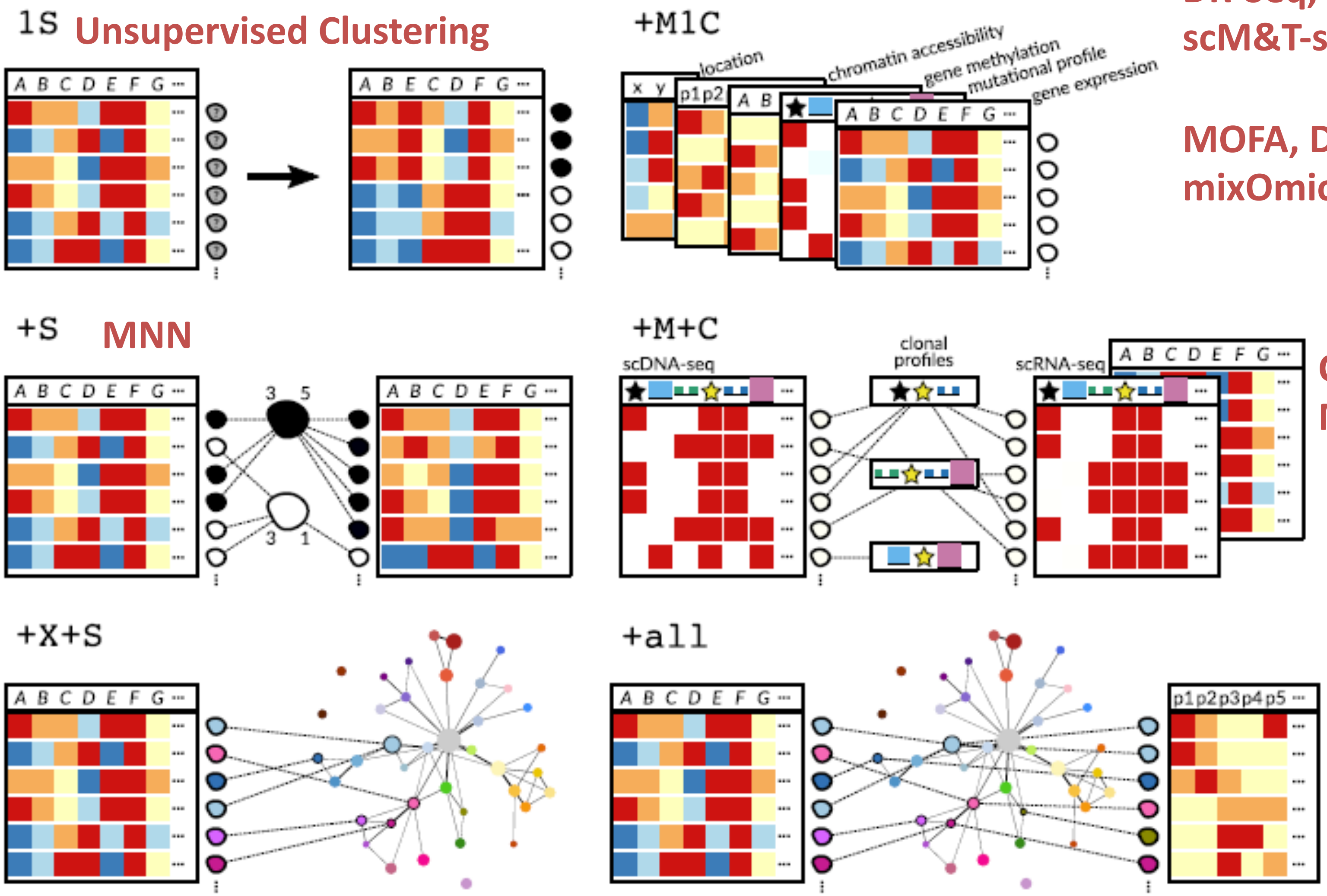
  i) rates of proliferation and mutation

  ii) microenvironment

X: Integration of sc data across samples, experiments, and types of measurements

- Issues with integration:
  - varying level of resolution
  - uncertainty in measurements
  - scaling to more cells

X: Status & Summary of Integration Options



Unsupervised Clustering

DR-Seq, G&T-seq
scM&T-seq

MOFA, DIABLO,
mixOmics, MINT

MNN

Cardelino
MATCHER

scmap, Conos,ClusterMap, BBKNN,
Moana, scID, scAlign, LAmbDA

X: Integrating

- Open problems
  - Missing data from different measurements due to limited sample

XI: Validating and benchmarking analysis tools for sc measurements

- Systematic benchmarking and evaluation
- Benchmarking datasets with known ground truth
- Status:
  - Single-cell data simulation
    - Splatter, powsimR, SymSim
- Open problems:
  - Non-transcriptomic data e.g. accuracy of phylogenetic inference
  - Evaluation metrics

scRNA-seq Challenges?

- QC'ing
  - Batch-effect
  - Integrating data from different labs/experiments
- Clustering
- Trajectory/pseudotime
- DE genes