How to Get Started with Single Cell RNA Sequencing Data Analysis

Michael S. Balzer,[1,2,3] Ziyuan Ma,[1,2,3] Jianfu Zhou,[1,2,3] Amin Abedini,[1,2,3] and Katalin Susztak[1,2,3]

[1]Renal Electrolyte and Hypertension Division, Department of Medicine, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania
[2]Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania
[3]Institute for Diabetes, Obesity and Metabolism, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania

ABSTRACT

Over the last 5 years, single cell methods have enabled the monitoring of gene and protein expression, genetic, and epigenetic changes in thousands of individual cells in a single experiment. With the improved measurement and the decreasing cost of the reactions and sequencing, the size of these datasets is increasing rapidly. The critical bottleneck remains the analysis of the wealth of information generated by single cell experiments. In this review, we give a simplified overview of the analysis pipelines, as they are typically used in the field today. We aim to enable researchers starting out in single cell analysis to gain an overview of challenges and the most commonly used analytical tools. In addition, we hope to empower others to gain an understanding of how typical readouts from single cell datasets are presented in the published literature.

JASN 32: 1279–1292, 2021. doi: https://doi.org/10.1681/ASN.2020121742

BACKGROUND

The first description of single cell gene expression analysis on the basis of next-generation sequencing was in 1992.[1] In 2015, encapsulation and barcoding-based analysis was developed.[2] During the last 5 years, single cell analysis was democratized and most academic institutions have dedicated core facilities to perform single cell expression, epigenome, or other multimodal analysis. New statistical analytical methods have been developed rapidly. Several analytical platforms have also been developed, such as Seurat,[3] which is written in R (https://satijalab.org/seurat/get_started.html) and Scanpy,[4] written in Python (https://scanpy.readthedocs.io/en/stable/tutorials.html). Here, we review basic analytical tools and concepts. We focus on 10× Genomics data as they are more commonly used (Figure 1). The review is strongly on the basis of two case study tutorials (https://www.github.com/theislab/single-cell-tutorial and http://scrnaseqcourse.cog.sanger.ac.uk/website/index.html).[5,6]

DATA MATRIX GENERATION AND QUALITY CONTROL

A key technical advance in single cell analysis has been the development of barcoding, which allows massive parallelization while keeping costs at a minimum. The barcodes are added to the RNA molecules during reverse transcription, allowing the identification of both individual cells and unique molecules. The first analytical step is the generation of a data matrix, which represents a barcode (cell) by transcript database from the raw sequencing files. For 10× Genomics data, CellRanger (Table 1, summarizes tools, methods, and databases as mentioned in the text) is the most commonly used pipeline that includes demultiplexing and alignment of the sequencing reads to the genome, annotating the aligned reads to genes, and quantifying genes. Alternatives include, for example, unique molecular identifier (UMI) tools,[7] zUMIs,[8] kallisto,[9] STAR,[10] and STARsolo (https://github.com/alexdobin/STAR/blob/master/docs/STARsolo.md).

Each barcode could represent a single cell, a doublet, or an "empty" droplet containing no cells, but ambient RNA. An important issue to mention is that the standard pipeline aligns the sequencing data to the transcriptome, such as the processed mature mRNA. However, single nuclear RNA data or epigenome data (Assay for Transposase Accessible

Published online ahead of print. Publication date available at www.jasn.org.

Correspondence: Dr. Katalin Susztak, 12–123 Smilow Translational Research Center, 3400 Civic Center Boulevard, Philadelphia, PA 19104. Email: ksusztak@pennmedicine.upenn.edu

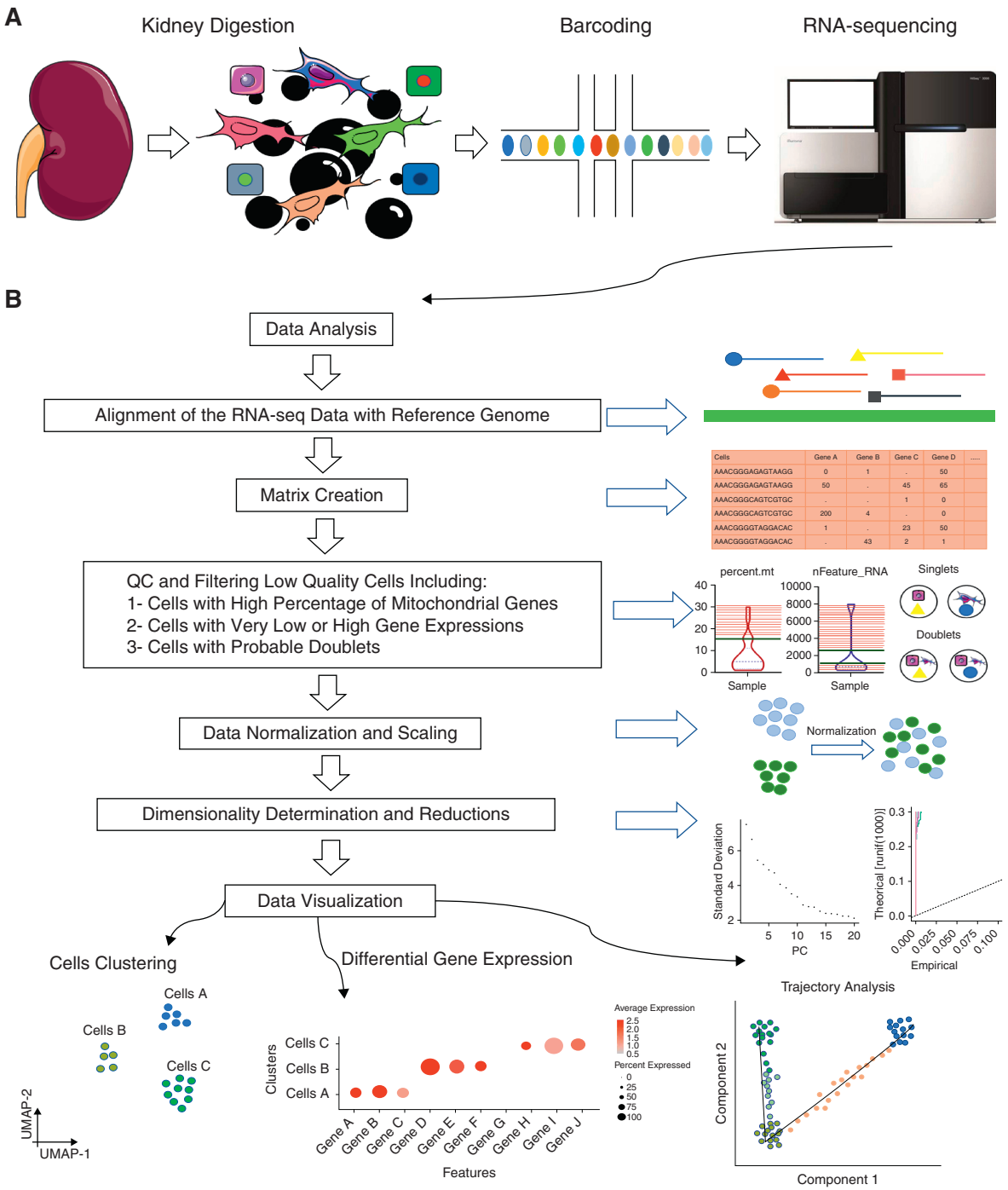

Figure 1. Workflow of renal scRNA-seq data creation and analysis. (A) Steps for preparation of kidney scRNA-seq data. (B) Typical steps for data analysis.

Chromatin by sequencing, ATAC-seq) should be aligned to the full genome, as the nucleus mostly contains pre-mRNA, which includes the intronic regions. Raw read counts usually also filter out genes detected in very few cells, effectively reducing data matrix size.

The next step in the analytical pipeline is quality control (QC), such as identifying the number of counts per barcode, the number of genes per barcode, and the fraction of counts from mitochondrial genes per barcode (Figure 2).[11,12] Low gene numbers and a high fraction of mitochondrial reads generally indicates poor-quality cells. Some cells, however, including the kidney proximal and distal convoluted tubule cells, are very rich in mitochondria. Unusually high read and gene counts could represent doublets. Several doublet

Table 1. Overview of software tools, methods, and databases

Tool/Method/Database	Source	Repository
Data matrix generation and quality control		
CellRanger	10× Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest
UMI-tools	Open	https://github.com/CGATOxford/UMI-tools
zUMIs	Open	https://github.com/sdparekh/zUMIs
kallisto	Open	https://github.com/pachterlab/kallisto
STAR	Open	https://github.com/alexdobin/STAR
STARsolo	Open	https://github.com/alexdobin/STAR/blob/master/docs/STARsolo.md
DoubletDecon	Open	https://github.com/EDePasquale/DoubletDecon
Scrublet	Open	https://github.com/AllonKleinLab/scrublet
DoubletFinder	Open	https://github.com/chris-mcginnis-ucsf/DoubletFinder
SoupX	Open	https://github.com/constantAmateur/SoupX
CellBender	Open	https://github.com/broadinstitute/CellBender
Normalization		
Scran	Open	https://github.com/MarioniLab/scran
Seurat	Open	https://github.com/satijalab/seurat
SCtransform	Open	https://github.com/ChristophH/sctransform
SCnorm	Open	https://github.com/rhondabacher/SCnorm
BayNorm	Open	https://github.com/WT215/bayNorm
Batch effect correction and data integration		
Seurat (CCA)	Open	https://github.com/satijalab/seurat
Seurat (RPCA)	Open	https://github.com/satijalab/seurat
Scanorama	Open	https://github.com/brianhie/scanorama
Harmony v. 1.0	Open	https://github.com/immunogenomics/harmony
LIGER	Open	https://github.com/welch-lab/liger
Visualization and clustering		
t-SNE	Open	https://github.com/oreillymedia/t-SNE-tutorial
UMAP	Open	https://github.com/lmcinnes/umap
Louvain	Open	https://github.com/vtraag/louvain-igraph
Leiden	Open	https://github.com/kharchenkolab/leidenAlg
DESC	Open	https://github.com/eleozzr/desc
Garnett	Open	https://github.com/cole-trapnell-lab/garnett
SingleR	Open	https://github.com/dviraran/SingleR
CHETAH	Open	https://github.com/jdekanter/CHETAH
MOANA	Open	https://github.com/yanailab/moana
Cell level analysis: cell fraction changes, decomposition, and trajectory analysis		
MuSiC	Open	https://github.com/xuranw/MuSiC
CIBERSORT	Open	https://github.com/jason-weirather/CIBERSORT
BSEQ-sc	Open	https://github.com/shenorrLab/bseqsc
BisqueRNA	Open	https://github.com/cran/BisqueRNA
Monocle	Open	https://github.com/cole-trapnell-lab/monocle-release
tradeSeq	Open	https://github.com/statOmics/tradeSeq
Slingshot	Open	https://github.com/kstreet13/slingshot
PHATE	Open	https://github.com/KrishnaswamyLab/PHATE
VelocytoR	Open	http://velocyto.org
Gene-level analysis: Differential expression, gene regulatory network, driver pathways, and cell-cell interaction		
MAST	Open	https://github.com/RGLab/MAST
GSEA	Open	https://github.com/GSEA-MSigDB/gsea-desktop
WGCNA	Open	https://github.com/cran/WGCNA
MSigDB	Open	http://www.gsea-msigdb.org/gsea/msigdb/index.jsp
GO	Open	http://geneontology.org
KEGG	Open	https://www.genome.jp/kegg/
Reactome	Open	https://reactome.org
CellPhoneDB	Open	https://github.com/Teichlab/cellphonedb
Connectome	Open	https://github.com/msraredon/Connectome

Table 1. Continued

Tool/Method/Database	Source	Repository
snATAC-seq analysis		
SnapATAC	Open	https://github.com/r3fang/SnapATAC
Signac	Open	https://github.com/timoast/signac
ArchR	Open	https://github.com/GreenleafLab/ArchR
MACS2	Open	https://github.com/taoliu/MACS
Cell Ranger ATAC	10× Genomics	https://support.10xgenomics.com/single-cell-atac/software/downloads/latest
HOMER	Open	http://homer.ucsd.edu/homer/motif/
chromVAR	Open	http://bioconductor.org/packages/release/bioc/html/chromVAR.html
Cicero	Open	https://github.com/cole-trapnell-lab/cicero-release
GREAT	Open	http://great.stanford.edu/public/html/
Webtools and datasets		
Human Cell Atlas	Open	https://www.humancellatlas.org
Human BioMolecular Atlas Program	Open	https://hubmapconsortium.org
Kidney Precision Medicine Project	Open	https://www.kpmp.org
Rebuilding a Kidney	Open	https://www.rebuildingakidney.org
KIT (Humphreys Lab)	Open	http://humphreyslab.com/SingleCell/
Susztak Lab	Open	http://susztaklab.com/sc http://susztaklab.com/VisCello/ http://susztaklab.com/developing_adult_kidney/snATAC http://susztaklab.com/developing_adult_kidney/scRNA/ /http://susztaklab.com/developing_adult_kidney/igv/
VisCello	Open	https://github.com/qinzhu/VisCello
Azimuth	Open	https://github.com/satijalab/azimuth

For a comprehensive overview of bioinformatic tools for scRNA-seq analysis see also https://www.scrna-tools.org/tools. Tools, methods, and databases are ordered as mentioned in the main text and given with a repository URL. RPCA, reciprocal PCA; t-SNE, *t*-distributed stochastic neighbor embedding; GSEA, gene set enrichment analysis; WGCNA, weighted correlation network analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology.

detection tools now are available including DoubletDecon,[13] Scrublet,[14] and DoubletFinder (Figure 2).[15] However, a critical issue in doublet detection is that transitional cells containing marker genes from, for example, both epithelial and mesenchymal origin might be tagged as doublet, sometimes resulting in false-positive detection. Furthermore, these tools identify only poorly homotypic doublets, namely, doublets formed from transcriptionally similar cells that cluster among their composite cell type singlets in the gene-expression space.

It also important to control for ambient RNA contamination. Ambient RNA is RNA that is present in the single cell solution and is incorporated into the oil droplet during encapsulation. We routinely use SoupX, which estimates ambient RNA contamination from empty droplets (Figure 2).[16] An alternative package is CellBender, which removes counts due to ambient RNA molecules and random barcode swapping from (raw) UMI-based single cell RNA sequencing (scRNA-seq) count matrices.[17] (preprint) In a typical analysis, we consider multiple QC parameters for filtering, which we use iteratively.

Normalization

Different types and levels of normalization are needed for single cell data (Figure 1). For example, the total sequencing read count number alters the raw count number, so gene counts should be scaled to the overall count depths. A commonly used method assumes each cell had the same initial number of transcripts, simply normalizing data into counts per million. Scran uses pooling-based size factor estimation and linear regression to normalize data, and it is one of the most popular methods[18] in addition to simple log normalization used by Seurat.[3] Other methods have been developed, such as SCtransform,[19] SCnorm,[20] and BayNorm.[21] After normalization, the data are log(x+1) transformed. It is common to regress out cell cycle–associated variation from the data, and it is included in the standard analysis platform in Seurat or Scanpy. The platform allows the regression of other technical or biologic variation as well.

Batch Effect Correction and Data Integration

Most often, several datasets are generated, necessitating additional batch correction and data integration methods (Figure 3). Larger datasets that contain multiple different experiments and different methods are typically integrated using nonlinear methods. In Seurat there is an option for reference-based integration, which uses the Canonical Correlation Analysis or reciprocal principal component analysis (PCA).[22] Scanorama is another popular and well-performing method used in Scanpy.[23] Recently, Harmony[24] has gained popularity and is rapidly becoming the most commonly used integration method for single cell datasets. In a first step, the

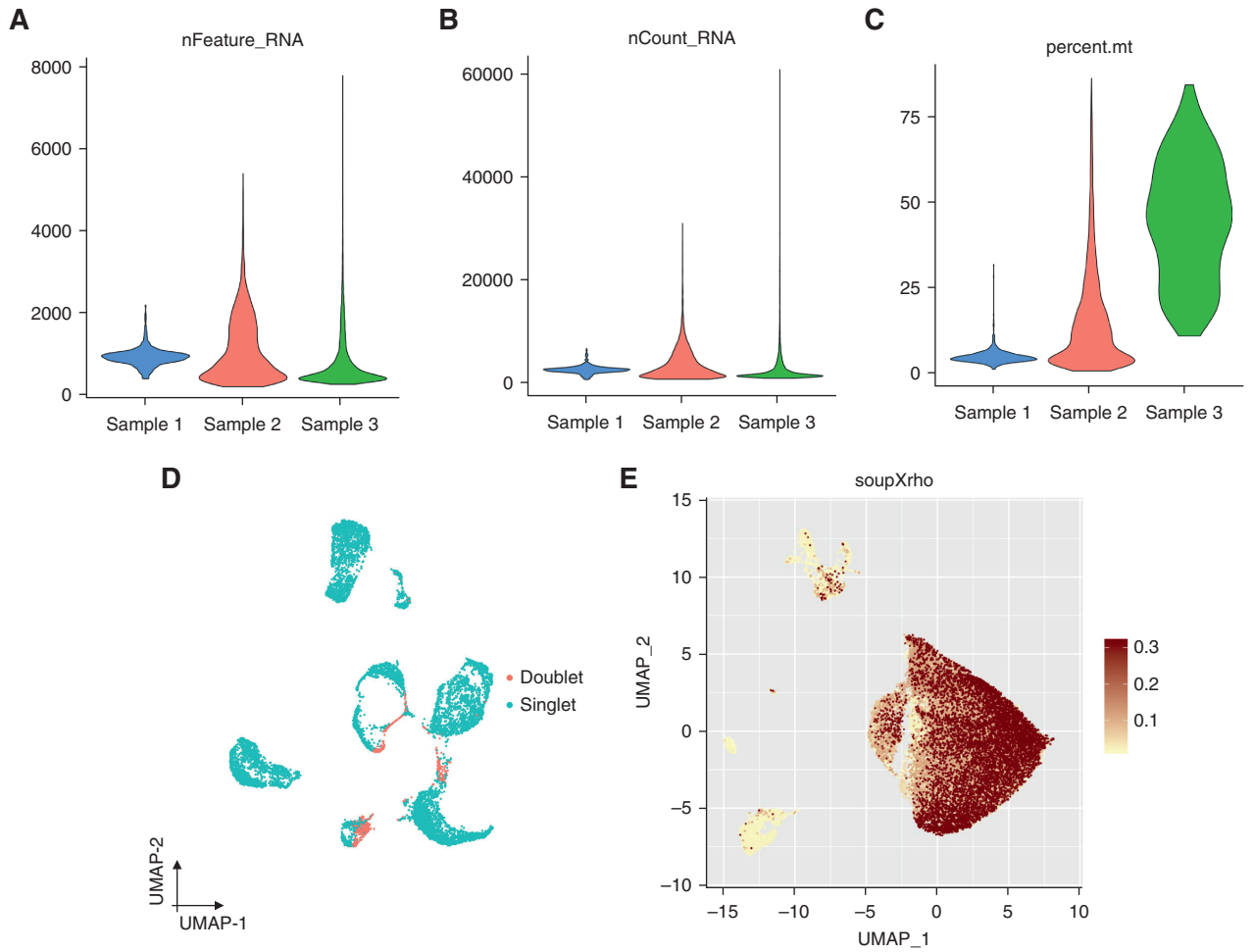

Figure 2. Quality control. Common QC metrics include (A) the number of unique genes (features) detected in each cell, (B) the total number of molecules detected within a cell, and (C) the percentage of reads that map to the mitochondrial genome. (D) DoubletFinder identifies the doublets in a single cell dataset. Doublets are shown in UMAP plot by pink. (E) SoupX highlights ambient RNA contamination. In this mock dataset, one cell cluster has relatively high ambient RNA contamination (rho=fraction of contamination), whereas other clusters demonstrate very low contamination.

PCA-derived embeddings matrix and batch metadata are used for scaling so that each cell unit is given a length parameter. Then, cluster centroids are initialized with regular k-means clustering on the scaled data. Finally, batch effects are removed by iteratively pulling batch-specific centroid to cluster centroid until convergence. Linked Inference of Genomic Experimental Relationships[25] identifies shared and dataset-specific factors through integrative non-negative matrix factorization. After normalization by the number of UMIs, gene expression is scaled but not centered. Different integration methods could possibly show different results. In general, we expect the same cell types from different experiments integrate, specifically the control cells should align from multiple experiments. The interested reader is referred to two excellent recent papers by Tran *et al.*[26] and Chen *et al.*[27] that provide some of the best evidence in favor of Harmony, Linked Inference of Genomic Experimental Relationships, and Seurat regarding batch effect correction.

Visualization and Clustering

The first step of visualization is feature selection when informative genes (1000–5000) are retained and others are filtered out, which is implemented in both Seurat and Scanpy (Figures 4 and 5). Visualization is an attempt to summarize the dataset in a low dimensional space to observe patterns. In general, dimension reduction is achieved by linear and nonlinear methods. PCA is the basis of clustering and trajectory inference and is a linear transformation that preserves the Euclidian distances between the cells in the full PCA. In the commonly used Seurat pipeline, PCA is used in the preprocessing stage. PCs can be projected into technical and biologic covariates to understand their performance. Using a permutation-test–based jackstraw method, the PCA is summarized for the top PCs and the number of PCs selected by the "elbow" heuristic method (Figure 4).

Single cell data visualization mostly uses other nonlinear dimension-reduction methods, such as *t*-distributed stochastic neighbor embedding.[28] This method is

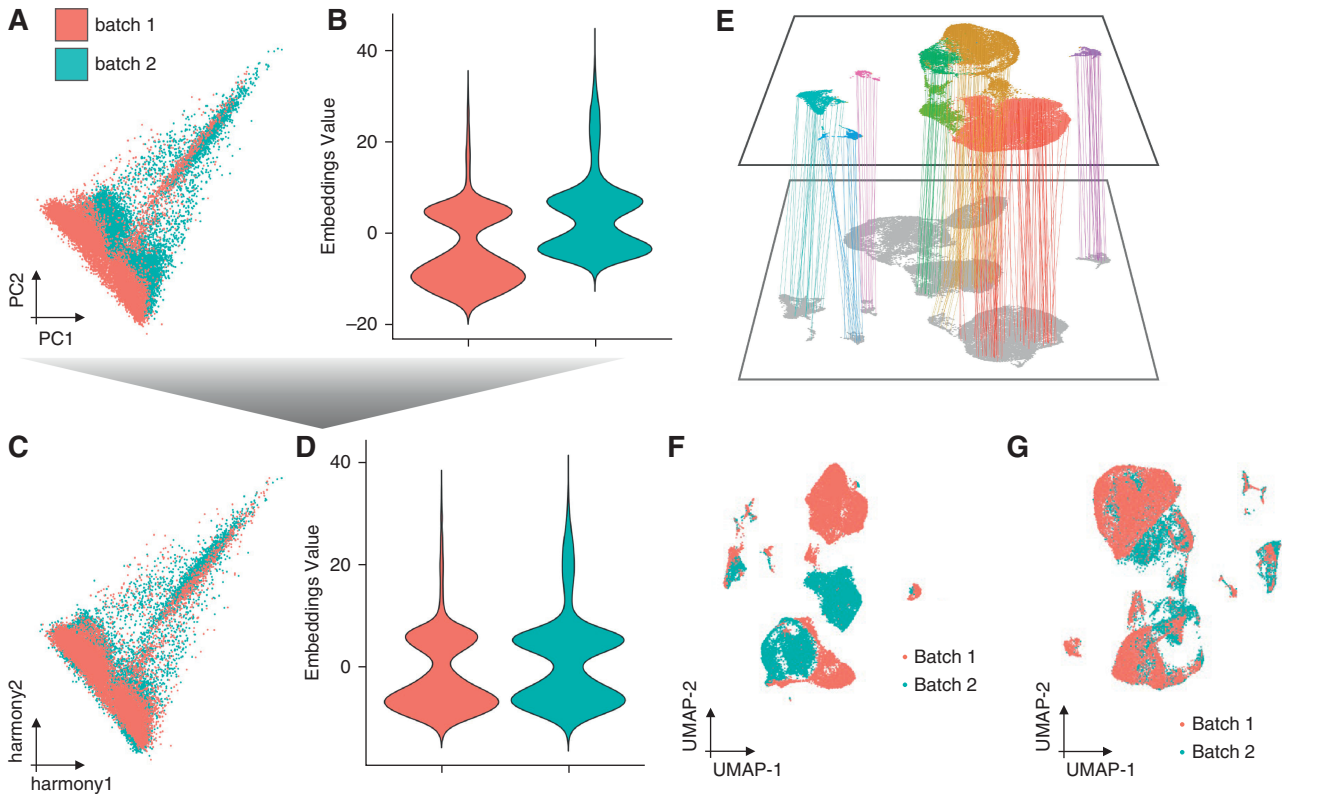

Figure 3. Batch effect correction and data integration. Batch effects are common in single cell datasets, whether they pertain to technical replicates or biologic samples. (A) Two-dimensional visualization of PCs corresponding to two separate batches analyzed in the same dataset. Note that cells clearly separate by individual batches. (B) The embedding value is a surrogate measure of similarity of the PCs. (C) After batch correction, cells overlap in the PC space and (D) embedding values are similar between the two batches, making the batches more comparable within the dataset. (E) Data integration algorithms such as Canonical Correlation Analysis (CCA) use anchors for batch integration. (F and G) The same dataset as in (A–D) is visualized in UMAP-embedded space (F) before and (G) after data integration with CCA.

focused on capturing local similarities at the expense of global structure. The Uniform Approximation and Projection (UMAP) method has gained popularity also due to its speed.[29] UMAP appears to capture underlying data structure better and can summarize data in more than two dimensions; therefore, it is now most commonly used for single cell data visualization. A key limitation of UMAP and *t*-distributed stochastic neighbor embedding is that they strongly depend on user-defined parameters, and the results are highly sensitive for these parameters. Most important to note is that neither visualization preserves cell-cell distances, so the resulting embedding should not be used directly by downstream analysis (Figure 4).

Cell clusters, formed on the basis of their similarities of gene expression, are the first immediate results of the analysis. Cell clustering allows inference of cell types by grouping cells on the basis of similarities of gene expression. Clustering is an unsupervised machine learning process that is on the basis of a distance matrix. The default clustering method in the community is the Louvain community detection on a single cell K-nearest neighbor approach. Cells are represented as nodes in the graph. Each cell is connected to its K most similar cells, which are typically obtained using Euclidean distances on the PC-reduced expression space. One critical issue is that the user determines the resolution in Louvain clustering, and the resolution determines the number of clusters or cell types identified in the dataset. We recommend performing subclustering, such as subsetting certain clusters from the initial dataset and then reclustering without the other cell types. This allows the emergence of finer, more granular data structure within the cell types. The Leiden community detection algorithm,[30] as incorporated in the Leidenbase package, is an alternative to the Louvain algorithm and is used as default in Monocle trajectory analysis (see below). New clustering methods use neural networks and artificial intelligence, for example Deep Embedding for Single-cell Clustering uses a deep neural network, with network weights and initial clustering obtained from an autoencoder.[31]

Clusters do not necessarily mean cell types. This is critically important to highlight, because user-defined cluster resolution parameters determine the number of observed clusters. Annotation and clustering are strongly linked. Clustering and annotation is conducted in an iterative fashion, which is time consuming. At present, there is no

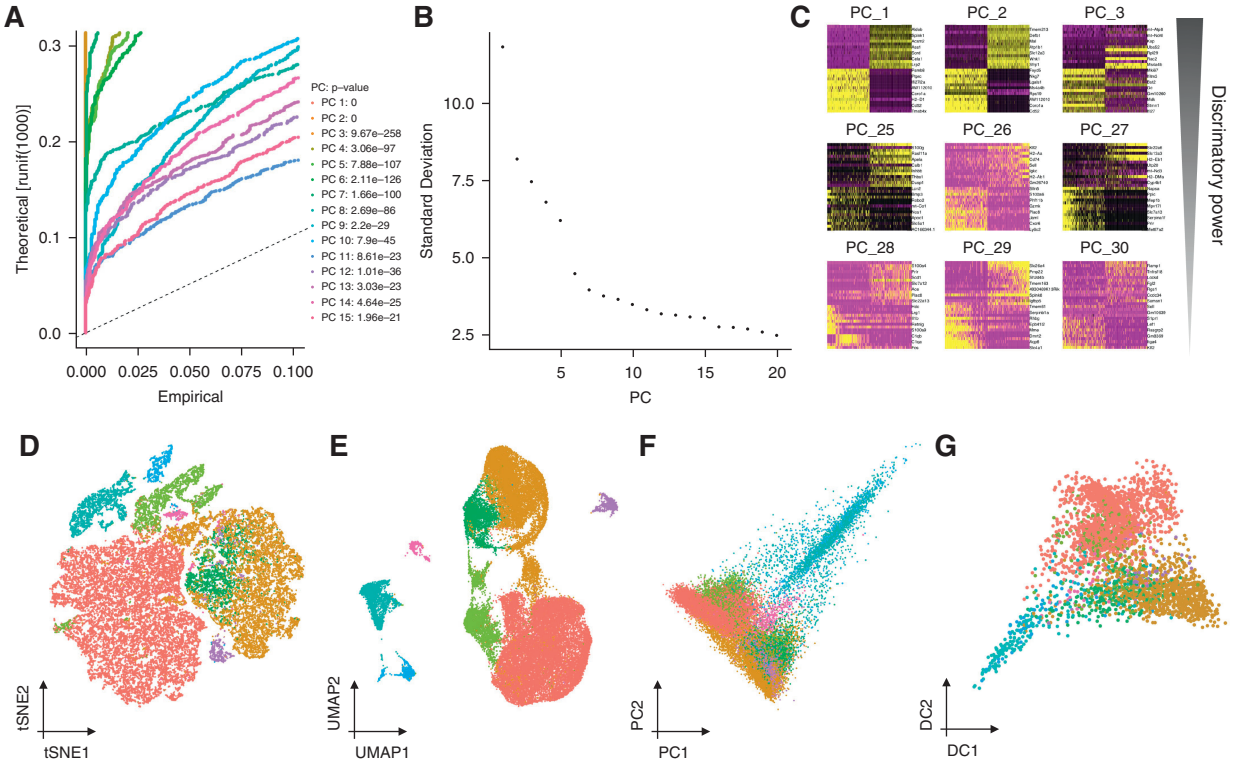

Figure 4. Visualization. To visualize high-dimensional single cell datasets, dimension reduction is used. (A) The jackstraw method performs association tests between known (empirical) values and estimated (theoretical) latent variables. The dashed line denotes a uniform distribution for each PC, against which the distribution of *P* values for each PC is compared. *P* values aid in choosing the number of informative PCs. (B) An elbow plot demonstrates the degree of variance explained by each individual component. Looking for the "elbow" in the plot is usually a good indication of where usefulness of additional PCs is minimal. (C) Heatmaps showing the enrichment of top genes loading on the first three PCs and PCs 25 through 30. Sharpness of separation as a surrogate of discriminatory power is decreasing with increasing PCs. (D–G) A mock kidney dataset that is projected on to tSNE, UMAP, PCA, and Diffusion map spaces to demonstrate the different visualization properties of the respective dimension reduction techniques.

consensus around optimal clustering parameters. Therefore, multiple versions of clustering and interpretation of the same data are acceptable. Wilcoxon rank-sum test is used to rank genes by difference in expression among groups.

Classic cell type annotations use an external dataset, which is considered ground truth. The growing number of external datasets for kidney cell type annotation include Susztak Lab,[32–35] Humphreys Lab,[36,37] Tabula Muris,[38] Human Cell Atlas,[39] Renal Epithelial Cell Ontology webpage,[40–42] and ImmGen Consortium.[43] Recently, automated cell annotations have been developed, such as Garnett,[44 (preprint)] SingleR,[45] CHETAH,[46] and MOANA,[47] which offer a more holistic and probabilistic method of cell identity annotation. Marker genes for the same cell types may differ between datasets.

Cell-level Analysis: Cell Fraction Changes, Decomposition, and Trajectory Analysis

Changes of cell fractions (proportions of each cell type in the dataset) show strong association with disease state, which is one of the most simplistic outputs of the single cell analysis. These numbers can provide relative estimates between conditions, but cell fractions inferred from single cell data might be inaccurate, due to bias in cell capture of the single cell library preparation. Also, the proportion of, for example, proximal tubule cells will be higher in samples obtained from the kidney cortex compared with samples taken from the medulla. To infer cell type composition of bulk RNA-seq data, MuSiC[48] is a recently developed method for bulk tissue cell type deconvolution with single cell expression data as reference. MuSiC uses weighted non-negative least squares regression to estimate cell type proportions.[49] Alternative methods include CIBERSORT,[50] BSEQ-sc,[51] and BisqueRNA.[52] Statistical tests over changes in the proportion of a cell identity cluster between samples are dependent on one another, and, because as the proportion of one cell identity cluster changes, the proportions of all others will have changed as well. Alternatively, a permutation-based statistical testing approach could be used for differential proportion analysis, in which cluster proportions are compared with a random proportion of total cells.[53]

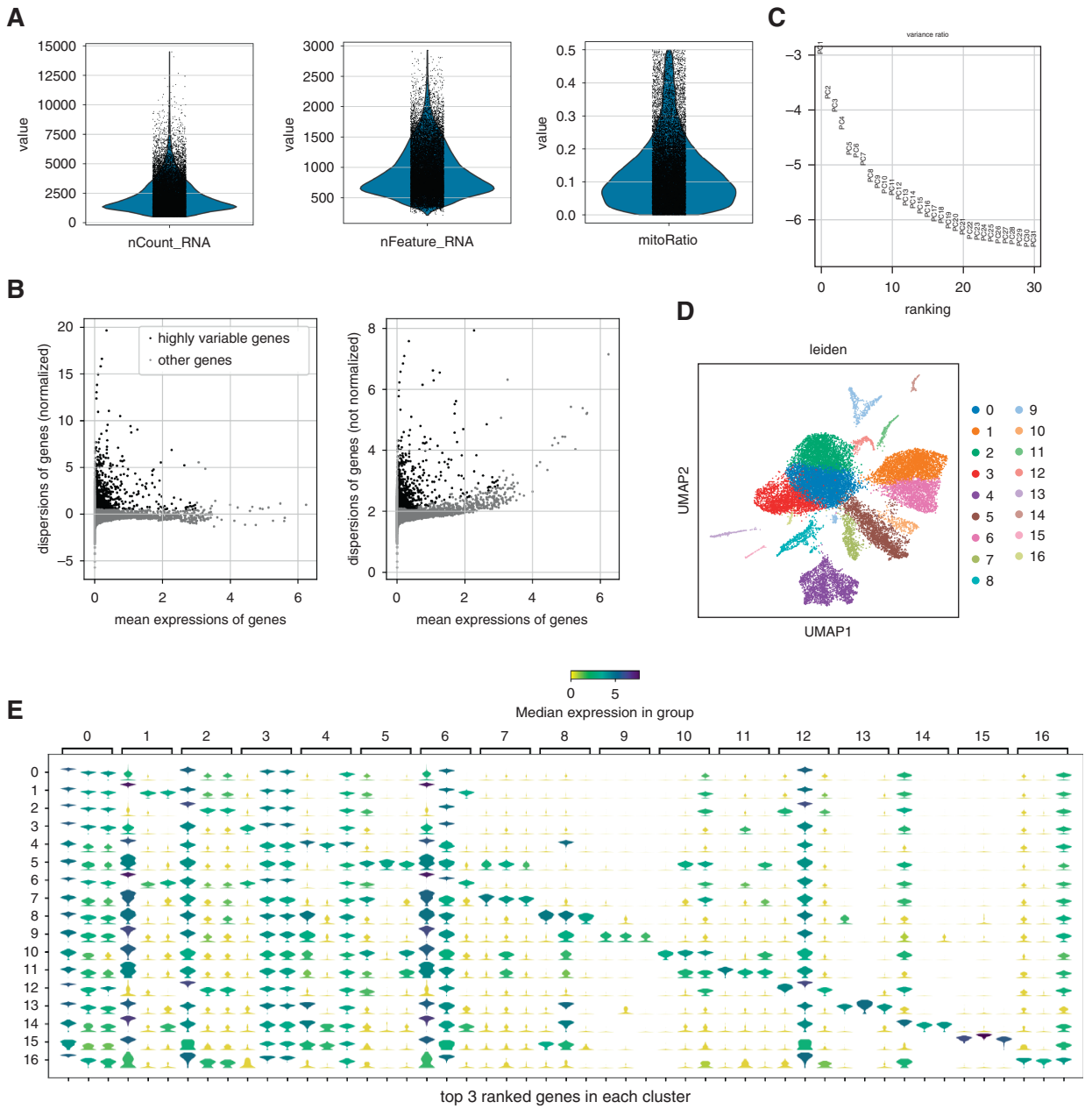

Figure 5. Basic workflow of single cell analysis in Scanpy. (A) Scanpy uses common metrics, such as the total number of molecules, the number of unique genes, and the percentage of reads mapped to the mitochondrial genome detected in each cell for quality control. (B) Scanpy finds highly variable genes within the normalized data. (C) Scanpy reduces the dimensionality of the data by running PCA, followed by the calculation of cell neighborhood graphs. (D) Leiden graph-clustering method is run on UMAP to separate cells. (E) Scanpy can run a Wilcoxon rank-sum test to calculate a ranking for the highly differential genes within each cluster, which helps identify its cell type.

Cellular diversity cannot sufficiently be described by a discrete classification system such as clustering. Trajectory analysis captures the salient characteristics of cells during transitions, such as during organ development along several time points, or between disease states, cellular history, or topological information. The biologic processes that drive the observed heterogeneity are continuous.[54] Thus, capturing transitions between cell identities, branching differentiation processes or gradual, unsynchronized changes in biologic function requires dynamic models of gene expression. Monocle is a machine learning method to reconstruct the sequence of gene expression changes each cell must execute as it transitions from one state to another.[55–57] It is on the basis of reverse-graph embedding, a highly scalable nonlinear manifold learning technique. After the method learns

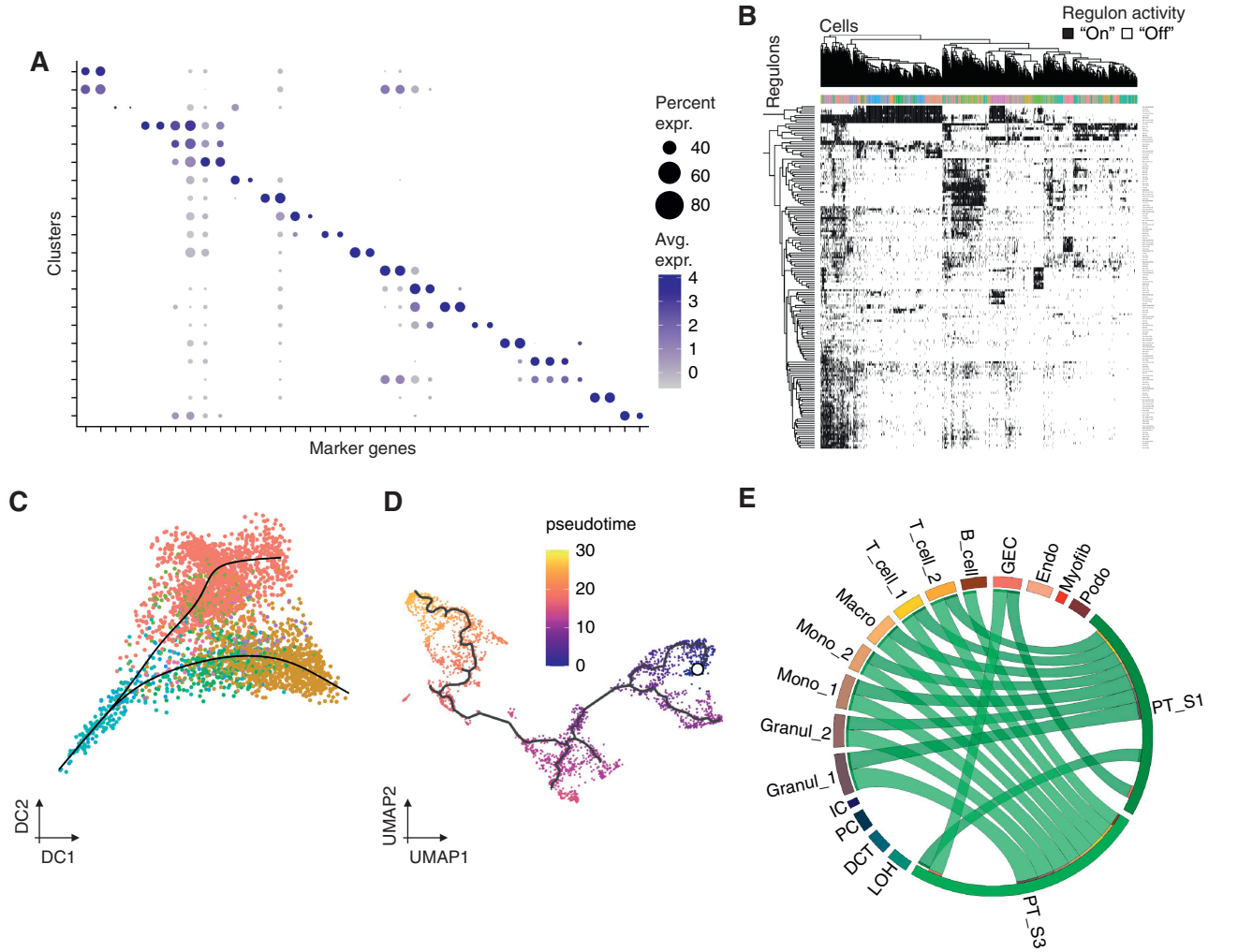

Figure 6. Downstream analyses. (A) Gene expression values are compared across cell cluster identities to identify similarities or differences between cell clusters. In this example, cells are grouped into respective clusters along the *y*-axis and gene expression of selected marker genes on the *x*-axis is shown. Dot size denotes the percentage of cells expressing the marker gene, whereas the color corresponds to the average expression level in each cluster. (B) Heatmap demonstrating SCENIC-derived regulon activity, a measure that is derived by *cis*-regulatory analysis predicting target genes of low-expression transcription factors. (C) Slingshot-derived trajectory in diffusion maps embedding space. (D) Corresponding Monocle-derived trajectory in UMAP embedding space. The color scale corresponds to increasing pseudotime. (E) Circos plot quantifying the number of cell-cell interactions of kidney cell clusters, as derived by CellphoneDB analysis.

the transition path, or trajectory, it places each cell at the correct position along it, which is called pseudotime, a measure of how far a cell has moved through biologic progress. A newer method to analyze cell history is implemented in the recently developed RNA velocity analysis, as in the package velocyto.[57] RNA velocity is the time derivative of the gene expression state and can be directly estimated by distinguishing between unspliced and spliced mRNAs in common scRNA-seq protocols.[57,58] RNA velocity is a high-dimensional vector that predicts the future state of individual cells on a timescale of hours. TradeSeq,[59] which is on the basis of a prior method called Slingshot,[60] outperforms other methods for simple trajectory analysis. Another useful package is PHATE, a visualization method that captures both local and global nonlinear structure using an information-geometric distance between data points.[61,62] Inferred trajectories do not necessarily have to represent a biologic process and further sources of evidence should be collected to interpret a trajectory derived from these methods (Figure 6).

Gene-level Analysis: Differential Expression, Gene Regulatory Network, Driver Pathways, and Cell-Cell Interaction

Differential expression (DE) analysis is performed on uncorrected data by including technical and biologic covariates. Seurat uses different models for DE analysis (Figure 6). MAST uses a hurdle model to account for drop-out.[63] To correlate scRNA-seq dataset information

with other phenotypic variables, regression-based models can combine several samples and their associated phenotypic characteristics to correlate gene expression changes in certain cell types (such as proximal tubular cells), with a respective quantitative measured phenotype (for example GFR, albuminuria). Although DE testing tools typically allow the user the flexibility to incorporate confounders, users must be vigilant as to which variables are added to the model. For example, in most single cell experimental set-ups, the sample and condition covariates are confounded, because it is rarely possible to obtain a single sample under multiple conditions. Gene-level analysis can also be combined with gene set enrichment analysis methods, such as gene set enrichment analysis or weighted correlation network analysis.[64]

To interpret DE results, we typically group genes on the basis of involvement in common biologic processes. Biologic process labels are stored in databases such as MSigDB,[65] the Gene Ontology,[66,67] or the Kyoto Encyclopedia of Genes and Genomes[68] and Reactome[69] databases. Although one needs to keep in mind that enrichment for gene expression of some pathway members might not necessarily be associated with pathway activity, enrichment of annotations on the gene list can be tested using a vast array of tools, which have been reviewed and compared elsewhere.[70,71]

A recent development in the single cell analysis field is the use of paired gene labels to perform ligand–receptor analysis.[72] Here, interaction between cell clusters is inferred from the expression of receptors and their cognate ligands. Ligand–receptor pair labels can be obtained from recent databases, such as CellPhoneDB[73] or Connectome,[74] (preprint) and used to interpret highly expressed genes across clusters using statistical models.[75–77]

Gene Regulation at Single Cell Resolution

Single nuclei Assay for Transposase Accessible Chromatin by sequencing (snATAC-seq) allows for the analysis of the epigenomic landscape in single cells by profiling chromatin accessibility (Figure 7). Multiple tools have been developed for snATAC-seq analysis. The best known are SnapATAC developed by the Ren Lab,[78] (preprint) Signac developed by the Satija Lab,[79] (preprint) and ArchR developed by the Greenleaf Lab.[80] (preprint) We prefer SnapATAC, which is a nonlinear dimensionality reduction method. After generating the barcode-by-cell matrix in CellRanger, we preprocess the matrix by binarizing the fragments into uniformly sized cell-by-bin matrix using SnapATAC. The QC steps include filtering poor-quality cells or doublets with a read depth that is too low or too high, and removing reads in genomic blacklist regions. Important QC criteria are the enrichment of transcription start sites, fraction of reads in peaks, and the ratio of reads in promoter regions. To identify cell types in heterogeneous tissue, SnapATAC utilizes diffusion maps. The low dimensional embeddings obtained from the diffusion maps are used as inputs into Harmony to remove the batch effect. Clustering is then performed with the Louvain algorithm using selected k values from k Nearest Neighbor algorithm as input. For cluster annotation, cell-gene activity score matrices from selected kidney cell type–specific marker genes were generated. Predefined promoter regions (*e.g.*, from the Ensembl regulatory build) or gene body + 2 kb region were used to integrate all fragments overlapped with gene transcripts. To call peaks from each cell type, all fragments obtained from the same cell types were aggregated to build a pseudo-bulk ATAC dataset and MACS2,[81] conducted separately for each cell type. ArchR implements an improved method by calling peaks on independent samples and then retaining reproducible peaks.[80] Fisher's exact test in edgeR tested between cell clusters to reveal differentially accessible regions for each cell type. To identify enriched motifs in different cell types, HOMER[82] or chromVAR[83] can be used for transcription factor analysis for the snATAC-seq data, although the genetic background will heavily influence which transcription factor motifs are enriched. To study how open chromatin changes are associated with cell differentiation and cell fate decision, Monocle3[56] for trajectory analysis was used by reducing dimensions using Latent Semantic Indexing and visualizing by UMAP. To understand open chromatin and target gene expression changes, a peak-peak correlation study is conducted by analyzing the coaccessibility of two peaks implemented in Cicero.[84] This strategy aggregates similar cells to obtain a set of "metacells" and addresses the issue of sparsity in the snATAC-seq data. Or, peaks can be imputed into GREAT[85] to identify nearest genes.

Webtools and Datasets

A large number of human and mouse kidney datasets have been generated over the last couple of years. The raw datasets are usually available for download from GEO. Large comprehensive reference human kidney annotation will be available as part of the Human Cell Atlas project[39] and the Human BioMolecular Atlas Program.[86] The Kidney Precision Medicine Project[87] (preprint) aims to generate datasets for a variety of human kidney disease conditions. The Rebuilding a Kidney consortium[88] will analyze developing human kidney samples and *in vitro* differentiated kidney organoids. In addition, several investigators have generated visualization tools for small single experimental datasets. The Humphreys Lab's KIT site allows quick visualization of their extensive data (http://humphreyslab.com/SingleCell/). The McMahon and Kim laboratories used VisCello to visualize data from developing and adult mice by also comparing differences between male and female animals. Using the VisCello[89] platform, our laboratory visualized developing, adult, healthy, and disease mouse model data (http://susztaklab.com/VisCello/)[35] and open chromatin epigenome data, which is also available for the same timepoints (http://susztaklab.com/developing_adult_kidney/igv/). These sites do not allow comprehensive analyses and the clustering parameters (which are somewhat subjective) are fixed, but they are

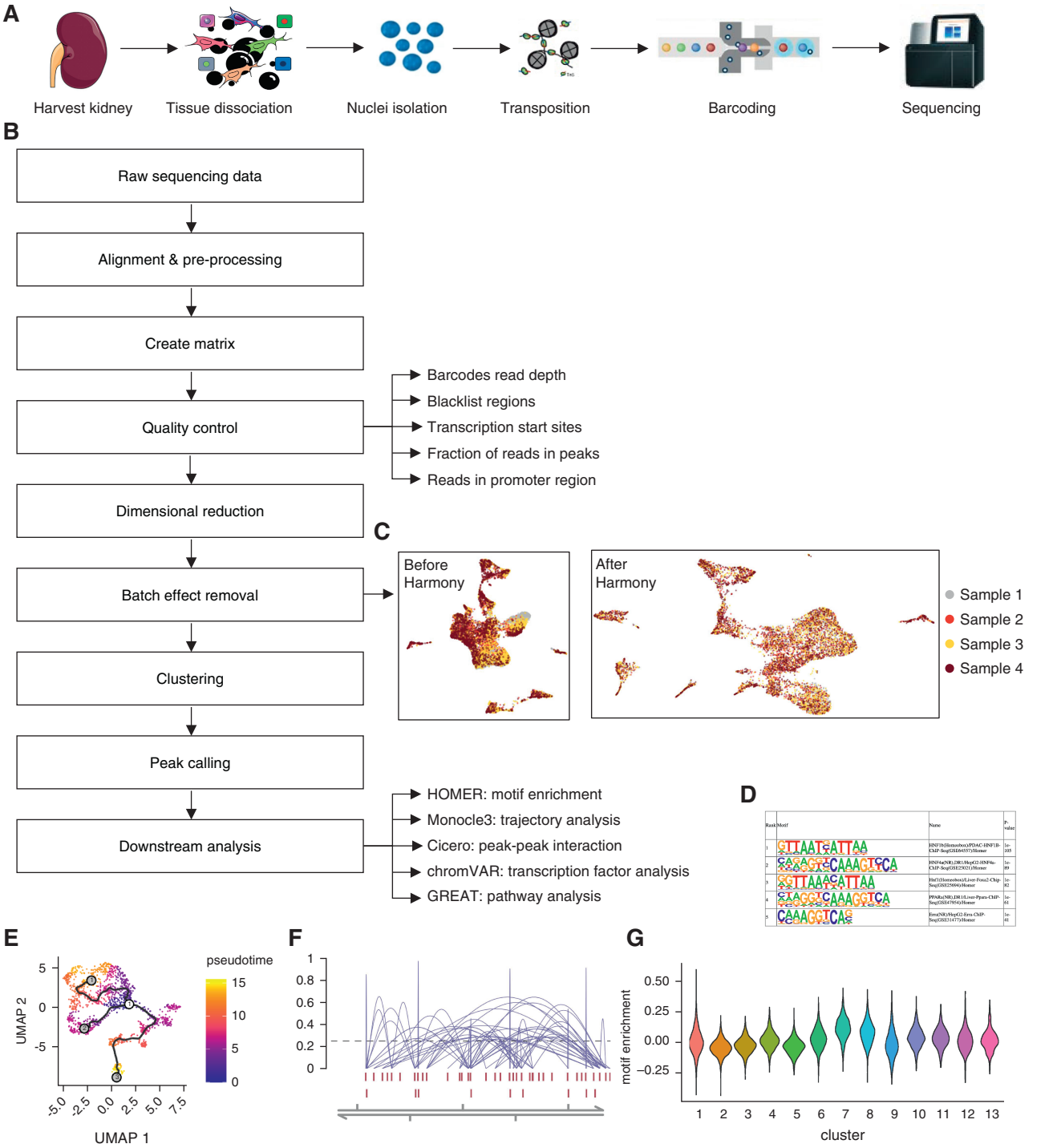

Figure 7. snATAC analysis pipeline. (A) Steps for preparation of kidney snATAC-sequencing data. (B) Typical steps for snATAC data analysis. (C) Batch effect removal by Harmony. (D) Motif enrichment analysis by Homer. (E) Trajectory analysis by Monocle3. (F) Peak-peak correlation analysis by Cicero. (G) Transcription factor analysis by chromVAR.

extremely useful for look-ups and comparisons. Another important development in data analysis automation is the stand-alone analysis application by the Satija Lab (http://azimuth.satijalab.org/app/azimuth),[90] (preprint) which allows projection of individual datasets on to a reference dataset, in which RNA and surface protein expression have been simultaneously measured in single cells. The investigator can upload locally generated datasets and the package automatically performs all steps outlined above and clusters with the human blood reference data.

Spatial and Multiomics Datasets

For the emerging fields of integration of spatial and multiomics datasets, we would like to refer to excellent current reviews[91–94] and to Supplemental Materials 1–3.

CONCLUSIONS

At present, kidney diseases are grouped on the basis of their temporal course, such as acute or chronic, or histologic descriptions, defined by color and shape homologies developed several centuries ago. These descriptions are unable to capture molecular mechanisms that underlie disease-driving molecular pathways. Therefore, they are not suited for target identification and drug development.[95,96] Single cell methods can resolve changes in disease states, allowing novel molecular disease classification and potential target identification.

DISCLOSURES

K. Susztak reports consultancy agreements with AstraZeneca, Bayer, Jnana, and Maze; reports receiving research funding from Bayer, Boehringer Ingelheim, Gilead, GSK, Lilly, Merck, Novo Nordisk, and Regeneron; reports receiving honoraria from Bayer, Jnana, and Maze; and reports being a scientific advisor or membership with the editorial board of *Cell Metabolism*, the *Journal of Clinical Investigation*, *JASN*, *Jnana*, and *Kidney International*. All remaining authors have nothing to disclose.

FUNDING

Work in the Susztak Lab is supported by the National Institutes of Health grants DK076077, DK087635, and DK105821. M.S. Balzer is supported by German Research Foundation (Deutsche Forschungsgemeinschaft) grant BA6205/2-1. We thank the University of Pennsylvania Diabetes Research Center for the use of the Core (P30-DK19525).

SUPPLEMENTAL MATERIAL

This article contains the following supplemental material online at http://jasn.asnjournals.org/lookup/suppl/doi:10.1681/ASN.2020121742/-/DCSupplemental.

Supplemental Material 1. Spatially resolved single cell datasets.

Supplemental Material 2. Integration of multiomics datasets: epigenome, protein expression, and beyond.

Supplemental Material 3. Supplemental references.

REFERENCES

1. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, et al.: Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* 89: 3010–3014, 1992
2. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al.: Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161: 1202–1214, 2015
3. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36: 411–420, 2018
4. Wolf FA, Angerer P, Theis FJ: SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 19: 15, 2018
5. Luecken MD, Theis FJ: Current best practices in single-cell RNA-seq analysis: A tutorial. *Mol Syst Biol* 15: e8746, 2019
6. Kiselev VY, Andrews TS, Hemberg M: Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 20: 273–282, 2019
7. Smith T, Heger A, Sudbery I: UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 27: 491–499, 2017
8. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I: zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* 7: giy059, 2018
9. Bray NL, Pimentel H, Melsted P, Pachter L: Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34: 525–527, 2016
10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.: STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21, 2013
11. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al.: Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* 17: 29, 2016
12. Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC: Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat Commun* 9: 2667, 2018
13. DePasquale EAK, Schnell DJ, Van Camp P-J, Valiente-Alandi I, Blaxall BC, Grimes HL, et al.: DoubletDecon: Deconvoluting doublets from single-cell RNA-sequencing data. *Cell Rep* 29: 1718–1727.e8, 2019
14. Wolock SL, Lopez R, Klein AM: Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst* 8: 281–291.e9, 2019
15. McGinnis CS, Murrow LM, Gartner ZJ: DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst* 8: 329–337.e4, 2019
16. Young MD, Behjati S: SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience 9: giaa151, 2020
17. Fleming SJ, Marioni JC, Babadi M: CellBender remove-background: A deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *bioRxiv* 10.1101/791699 (Preprint posted October 3, 2019)
18. Lun AT, Bach K, Marioni JC: Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 17: 75, 2016
19. Hafemeister C, Satija R: Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 20: 296, 2019
20. Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, et al.: SCnorm: Robust normalization of single-cell RNA-seq data. *Nat Methods* 14: 584–586, 2017
21. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, Marguerat S, et al.: bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics* 36: 1174–1181, 2020
22. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al.: Comprehensive integration of single-cell data. *Cell* 177: 1888–1902.e21, 2019
23. Hie B, Bryson B, Berger B: Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* 37: 685–691, 2019
24. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al.: Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16: 1289–1296, 2019
25. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ: Single-cell multiomic integration compares and contrasts features of brain cell identity. *Cell* 177: 1873–1887.e17, 2019
26. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al.: A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 21: 12, 2020
27. Chen Wanqiu, Zhao Yongmei, Chen Xin, Yang Zhaowei, Xu Xiaojiang, Bi Yingtao, et al.: A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. *Nat Biotechnol*, 2020 10.1038/s41587-020-00748-9
28. van der Maaten L, Hinton G: Visualizing Data using t-SNE. *J Mach Learn Res* 9: 2579–2605, 2008

29. McInnes L, Healy J, Saul N, Großberger L: UMAP: Uniform manifold approximation and projection. *J Open Source Softw* 3: 861, 2018
30. Traag VA, Waltman L, van Eck NJ: From Louvain to leiden: Guaranteeing well-connected communities. *Sci Rep* 9: 5233, 2019
31. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al.: Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat Commun* 11: 2338, 2020
32. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, et al.: Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* 360: 758–763, 2018
33. Park J, Liu CL, Kim J, Susztak K: Understanding the kidney one cell at a time. *Kidney Int* 96: 862–870, 2019
34. Miao Z, Balzer MS, Ma Z, Liu H, Wu J, Shrestha R, et al.: Single cell resolution regulatory landscape of the mouse kidney highlights cellular differentiation programs and renal disease targets. *Nat Commun* 20212041-1723
35. Dhillon P, Park J, Hurtado del Pozo C, Li L, Doke T, Huang S, et al.: The nuclear receptor ESRRA protects from kidney disease by coupling metabolism and differentiation. *Cell Metab* 33: 379–394.e8, 2021
36. Wu H, Malone AF, Donnelly EL, Kirita Y, Uchimura K, Ramakrishnan SM, et al.: Single-cell transcriptomics of a human kidney allograft biopsy specimen defines a diverse inflammatory response. *J Am Soc Nephrol* 29: 2069–2080, 2018
37. Wilson PC, Wu H, Kirita Y, Uchimura K, Ledru N, Rennke HG, et al.: The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proc Natl Acad Sci U S A* 116: 19619–19625, 2019
38. Tabula Muris Consortium; Overall Coordination; Logistical Coordination; Organ Collection and Processing; Library Preparation and Sequencing; Computational Data Analysis; Cell Type Annotation; Writing Group; Supplemental Text Writing Group; Principal Investigators: Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562: 367–372, 2018
39. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA: The human cell atlas: From vision to reality. *Nature* 550: 451–453, 2017
40. Chen L, Clark JZ, Nelson JW, Kaissling B, Ellison DH, Knepper MA: Renal-tubule epithelial cell nomenclature for single-cell RNA-sequencing studies. *J Am Soc Nephrol* 30: 1358–1364, 2019
41. Clark JZ, Chen L, Chou CL, Jung HJ, Lee JW, Knepper MA: Representation and relative abundance of cell-type selective markers in whole-kidney RNA-Seq data. *Kidney Int* 95: 787–796, 2019
42. Lee JW, Chou CL, Knepper MA: Deep sequencing in microdissected renal tubules identifies nephron segment-specific transcriptomes. *J Am Soc Nephrol* 26: 2669–2677, 2015
43. Heng TS, Painter MW; Immunological Genome Project Consortium: The immunological genome project: Networks of gene expression in immune cells. *Nat Immunol* 9: 1091–1094, 2008
44. Michielsen L, Reinders MJT, Mahfouz A: Hierarchical progressive learning of cell identities in single-cell data. *bioRxiv* 10.1101/2020.03.27.010124 (Preprint posted July 29, 2020)
45. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al.: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 20: 163–172, 2019
46. de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FCP: CHETAH: A selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 47: e95, 2019
47. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, et al.: A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 20: 194, 2019
48. Wang X, Park J, Susztak K, Zhang NR, Li M: Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* 10: 380, 2019
49. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, et al.: Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci U S A* 111: 13924–13929, 2014
50. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al.: Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 37: 773–782, 2019
51. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al.: A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* 3: 346–360.e4, 2016
52. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al.: Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun* 11: 1971, 2020
53. Farbehi N, Patrick R, Dorison A, Xaymardan M, Janbandhu V, Wystub-Lis K, et al.: Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *eLife* 8: e43882, 2019
54. Tanay A, Regev A: Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541: 331–338, 2017
55. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, et al.: Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 34: 637–645, 2016
56. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al.: The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32: 381–386, 2014
57. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al.: RNA velocity of single cells. *Nature* 560: 494–498, 2018
58. Svensson V, Pachter L: RNA velocity: Molecular kinetics from single-cell RNA-seq. *Mol Cell* 72: 7–9, 2018
59. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al.: Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun* 11: 1201, 2020
60. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al.: Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19: 477, 2018
61. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al.: Visualizing structure and transitions in high-dimensional biological data [published correction appears in *Nat Biotechnol* 38: 108, 2020 10.1038/s41587-019-0395-5]. *Nat Biotechnol* 37: 1482–1492, 2019
62. Saelens W, Cannoodt R, Todorov H, Saeys Y: A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 37: 547–554, 2019
63. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al.: MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16: 278, 2015
64. Beckerman P, Qiu C, Park J, Ledo N, Ko YA, Park AD, et al.: Human kidney tubule-specific gene expression based dissection of chronic kidney disease traits. *EBioMedicine* 24: 267–276, 2017
65. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP: Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27: 1739–1740, 2011
66. The Gene Ontology Consortium: Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 45[D1]: D331–D338, 2017
67. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al.; The Gene Ontology Consortium: Gene ontology: Tool for the unification of biology. *Nat Genet* 25: 25–29, 2000
68. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45[D1]: D353–D361, 2017
69. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al.: The reactome

pathway knowledgebase. *Nucleic Acids Res* 46[D1]: D649–D655, 2018
70. Tarca AL, Bhatti G, Romero R: A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One* 8: e79217, 2013
71. Huang W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1–13, 2009
72. Shao X, Lu X, Liao J, Chen H, Fan X: New avenues for systematically inferring cell-cell communication: Through single-cell transcriptomics data. *Protein Cell* 11: 866–880, 2020
73. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al.: Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563: 347–353, 2018
74. Raredon MSB, Yang J, Garritano J, Wang M, Kushnir D, Schupp JC, et al.: Connectome: Computation and visualization of cell-cell signaling topologies in single-cell systems data. *bioRxiv* 10.1101/2021.01.21.427529 (Preprint posted January 21, 2021)
75. Zhou L, Todorovic V, Kakavas S, Sielaff B, Medina L, Wang L, et al.: Quantitative ligand and receptor binding studies reveal the mechanism of interleukin-36 (IL-36) pathway activation. *J Biol Chem* 293: 403–411, 2018
76. Cohen M, Giladi A, Gorki AD, Solodkin DG, Zada M, Hladik A, et al.: Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. *Cell* 175: 1031–1044.e18, 2018
77. Zepp JA, Zacharias WJ, Frank DB, Cavanaugh CA, Zhou S, Morley MP, et al.: Distinct mesenchymal lineages and niches promote epithelial self-renewal and myofibrogenesis in the lung. *Cell* 170: 1134–1148.e10, 2017
78. Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, et al.: SnapATAC: A comprehensive analysis package for single cell ATAC-seq. *bioRxiv* 10.1101/615179 (Preprint posted August 17, 2020)
79. Stuart T, Srivastava A, Lareau C, Satija R: Multimodal single-cell chromatin analysis with Signac. *bioRxiv* 10.1101/2020.11.09.373613 (Preprint posted November 10, 2020)
80. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al.: ArchR: An integrative and scalable software package for single-cell chromatin accessibility analysis. *bioRxiv* 10.1101/2020.04.28.066498 (Preprint posted April 29, 2020)
81. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al.: Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9: R137, 2008
82. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al.: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38: 576–589, 2010
83. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ: chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 14: 975–978, 2017
84. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al.: Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell* 71: 858–871.e8, 2018
85. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al.: GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 28: 495–501, 2010
86. HuBmap Consortium: The human body at cellular resolution: The NIH human biomolecular atlas program. *Nature* 574: 187–192, 2019
87. Hansen J, Sealfon R, Menon R, Eadon MT, Lake BB, Steck B, et al.: Towards building a smart kidney atlas: Network-based integration of multimodal transcriptomic, proteomic, metabolomic and imaging data in the Kidney Precision Medicine Project. *bioRxiv* 10.1101/2020.07.23.216507 (Preprint posted July 24, 2020)
88. Oxburgh L, Carroll TJ, Cleaver O, Gossett DR, Hoshizaki DK, Hubbell JA, et al.: (Re) Building a kidney. *J Am Soc Nephrol* 28: 1370–1378, 2017
89. Packer JS, Zhu Q, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, et al.: A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* 365: eaax1971, 2019
90. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al.: Integrated analysis of multimodal single-cell data. *bioRxiv* 10.1101/2020.10.12.335331 (Preprint posted October 12, 2020)
91. Lee J, Hyeon DY, Hwang D: Single-cell multiomics: Technologies and data analysis methods. *Exp Mol Med* 52: 1428–1442, 2020
92. Stuart T, Satija R: Integrative single-cell analysis. *Nat Rev Genet* 20: 257–272, 2019
93. Asp M, Bergenstråhle J, Lundeberg J: Spatially resolved transcriptomes-next generation tools for tissue exploration. *BioEssays* 42: e1900221, 2020
94. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al.: Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 361: 1380–1385, 2018
95. Breyer MD, Coffman TM, Flessner MF, Fried LF, Harris RC, Ketchum CJ, et al.; Kidney Research National Dialogue (KRND): Diabetic nephropathy: A national dialogue. *Clin J Am Soc Nephrol* 8: 1603–1605, 2013
96. Breyer MD, Susztak K: The next generation of therapeutics for chronic kidney disease. *Nat Rev Drug Discov* 15: 568–588, 2016