scTenifoldKnk: a machine learning workflow performing virtual knockout experiments on single-cell gene regulatory networks

Daniel Osorio[1], Yan Zhong[2], Guanxun Li[2], Qian Xu[1], Andrew Hillhouse[1], Jingshu Chen[3], Laurie A. Davidson[4], Yanan Tian[3], Robert S. Chapkin[4], Jianhua Z. Huang[2,*], James J. Cai[1,5,6,*]

[1]Department of Veterinary Integrative Biosciences, [2]Department of Statistics, [3]Department of Veterinary Physiology and Pharmacology, [4]Department of Nutrition, [5]Department of Electrical and Computer Engineering, [6]Interdisciplinary Program of Genetics, Texas A&M University, College Station, TX 77843, USA

[*]Corresponding authors: J.Z.H. (jianhua@stat.tamu.edu) and J.J.C. (jcai@tamu.edu)

Keywords: gene knockout, single-cell RNA sequencing, scRNA-seq, gene regulatory network, machine learning, principal component regression, tensor decomposition, quasi-manifold alignment

Abstract

Gene knockout (KO) experiments are a proven approach for studying gene function. A typical KO experiment usually involves the phenotypic characterization of KO organisms. The recent advent of single-cell technology has greatly boosted the resolution of cellular phenotyping, providing unprecedented insights into cell-type-specific gene function. However, the use of single-cell technology in large-scale, systematic KO experiments is prohibitive due to the vast resources required. Here we present scTenifoldKnk—a machine learning workflow that performs virtual KO experiments using single-cell RNA sequencing (scRNA-seq) data. scTenifoldKnk first uses data from wild-type (WT) samples to construct a single-cell gene regulatory network (scGRN). Then, a gene is knocked out from the constructed scGRN by setting weights of the gene's outward edges to zeros. ScTenifoldKnk then compares this "pseudo-KO" scGRN with the original scGRN to identify differentially regulated (DR) genes. These DR genes, also called virtual-KO perturbed genes, are used to assess the impact of the gene KO and reveal the gene's function in analyzed cells. Using existing data sets, we demonstrate that the scTenifoldKnk analysis recapitulates the main findings of three real-animal KO experiments and confirms the functions of genes underlying three Mendelian diseases. We show the power of scTenifoldKnk as a predictive method to successfully predict the outcomes of two KO experiments that involve intestinal enterocytes in Ahr$^{-/-}$ mice and pancreatic islet cells in Malat1$^{-/-}$ mice, respectively. Finally, we demonstrate the use of scTenifoldKnk to perform systematic KO analyses, in which a large number of genes are virtually deleted, allowing gene functions to be revealed in a cell type-specific manner.

Highlights

- scTenifoldKnk is a machine learning workflow to perform virtual KO experiments using data from single-cell RNA sequencing (scRNA-seq).

- scTenifoldKnk requires only one input matrix—a gene-by-cell expression matrix obtained by scRNA-seq from wild-type (WT) samples.
- scTenifoldKnk constructs a single-cell gene regulatory network (scGRN) from scRNA-seq data of the WT sample, and then produces a pseudo-KO scGRN by setting the KO target gene representation in the WT scGRN to zero.
- scTenifoldKnk compares the WT scGRN and the pseudo-KO scGRN using a quasi-manifold alignment method, to reveal the perturbation effect of gene KO and generate a perturbation profile for the KO gene.
- Using real-data examples, we show that scTenifoldKnk is a powerful and effective approach for performing virtual KO experiments with scRNA-seq data to elucidate gene function.

Summary

Gene knockout (KO) experiments are a proven approach for studying gene function. However, large-scale, systematic KO experiments are prohibitive due to the limitation of experimental and animal resources. Here we present scTenifoldKnk—a machine learning workflow for performing virtual KO experiments with data from single-cell RNA sequencing (scRNA-seq). We show that the scTenifoldKnk virtual KO analysis recapitulates findings from real-animal KO experiments and can be used to predict outcomes from real-animal KO experiments. scTenifoldKnk is a powerful and efficient virtual KO tool for gene function study, allowing a systematic deletion of a large number of genes individually in scRNA-seq data to reveal individual gene function in a cell type-specific manner.

Introduction

Gene knockout (KO) experiments are a proven approach for studying gene function. A typical KO experiment involves the phenotypic characterization of organisms that carry the gene KO. For example, in KO mice, a gene is made inoperative or knocked out using genetic techniques. Deletion of one or more alleles provides mechanistic insight as to how the target gene functions within a particular biological context. Gene function may be predicted from phenotypic differences between the KO and wild-type (WT) samples, and its expression is a molecular phenotype that can be quantitatively measured. Gene expression is regulated in a coordinated manner in all living organisms. Typically, regulation is observable as synchronized patterns of transcription in a gene regulatory network (GRN). Co-regulation and co-expression are often seen among genes associated with the same biological processes and pathways or regulated by the same master transcription factors [1]. If one gene is knocked out, functionally related genes can mediate a homeostatic response. Thus, in unraveling regulatory mechanisms and synchronized patterns of cellular transcriptional activities, network analysis of gene expression provides mechanistic insights.

The recent advent of single-cell technology has greatly improved cellular phenotyping resolution—e.g., droplet-based single-cell RNA sequencing (scRNA-seq) makes it possible to profile transcriptomes of thousands of individual cells per assay. Applications of single-cell technology in KO experiments allow investigators to probe gene function in a cell type-specific manner. Large-scale, systematic KO experiments aim at deleting many genes individually in an organism. The use of scRNA-seq in such a large-scale systematic KO experiment would allow the function of each gene in various cell types to be studied at a single-cell level of resolution. However, limited resources tend to prohibit this kind of experiments, especially when multiple genes are targeted. Also, essential genes cannot be knocked out. For that reason, predictive

computational methods stand as a possible solution for such prohibitive experiments [2-4]. Taking advantage of the synchronized gene expression variability allowing the detection of co-expression between genes belonging to the same biological processes or being under the effect of the same transcription factors, it is now possible to construct personalized transcriptome-wide GRNs that are sample, tissue, and cell type-specific [5]. The topology of those biological networks is known to serve as a basis to accurately predict perturbations, like those caused by a KO gene [2, 6].

Here we present a machine learning workflow, called scTenifoldKnk, that can be used to perform virtual KO experiments from the topology of the constructed GRNs. scTenifoldKnk takes expression data from scRNA-seq of wild-type (WT) samples as input and constructs a denoised single-cell GRN (scGRN) using principal component (PC) regression and tensor decomposition. The WT scGRN is copied and then converted to a pseudo-KO scGRN by artificially zeroing out the target gene in the adjacency matrix. A quasi-manifold alignment method [7, 8] is adapted to compare the two scGRNs (WT vs. pseudo-KO). Through comparing the two scGRNs, scTenifoldKnk predicts changes in transcriptional programs and assesses the impact of KO on the WT scGRN. This information is then used to elucidate functions of the KO gene in analyzed cells through enrichment analysis.

ScTenifoldKnk is computationally efficient enough to allow the method to be applied to systematic KO experiments. In such a systematic study, we assume that thousands of genes in analyzed cells will be knocked out one by one. As mentioned, due to the experimental and biological limitations, such systematic KO experiments would never be possible in a real-animal experimental setting. The other features of scTenifoldKnk include: (1) scTenifoldKnk requires no data from KO animals or cell lines, as it only utilizes the scRNA-seq data from WT samples; and (2) scTenifoldKnk can be used to perform double-KO experiments—that is, to knock out more than one gene at a time.

The remainder of this paper is organized as follows. We first present an overview of the workflow of scTenifoldKnk and use simulated data to demonstrate its basic functions. We then use existing data generated from authentic-animal KO experiments to highlight the use of scTenifoldKnk. These existing data sets contain scRNA-seq expression matrices from both WT and KO samples. Although the KO data sets were available, they were not used by scTenifoldKnk as input. Instead, the KO data sets were specifically used as positive controls to show that scTenifoldKnk can produce expected results. Next, we show applications of scTenifoldKnk to delete genes that cause three different Mendelian diseases with known phenotypic outcomes. Finally, we conduct our own KO animal experiments to validate the utility of scTenifoldKnk as a predictive tool and discuss the contribution of scTenifoldKnk to the future of predictive biology.

Results

Result 1: The scTenifoldKnk workflow

scTenifoldKnk takes a single gene-by-cell count matrix from the WT sample as input. Eventually, scTenifoldKnk employs a network comparison method to compare a pseudo-KO scGRN to the WT scGRN to identify differentially regulated genes. These genes are called *virtual-KO perturbed genes*. From the enriched function of these perturbed genes, the function of the KO gene (i.e., the gene that is virtually knocked out) can be inferred. scTenifoldKnk is implemented with a modular structure containing three core modules illustrated in **Figure 1** and briefly summarized in three steps as follows.

Step 1: Constructing scGRN with scRNA-seq data from a WT sample

With the scRNA-seq data from a WT sample, scTenifoldKnk first constructs an scGRN using a pipeline we proposed previously, namely scTenifoldNet [9]. This *network construction* step contains three sub-steps (**Figure 1A**):

Substep 1.1. Subsampling cells randomly

Denote $\boldsymbol{X}$ as the scRNA-seq expression data matrix, which contains the expression levels for p genes and n cells. Then, m (< n) cells in $\boldsymbol{X}$ are randomly sampled to form $\boldsymbol{X}'$ using an *m*-out-of-*n* bootstrap procedure. This subsampling process repeats t times to create t subsets of cells $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_t$.

Substep 1.2. Constructing a GRN for each subsampled set of cells

For each $\boldsymbol{X}'_i$, principal component (PC) regression is run p times to construct a GRN. Each time the expression level of one gene is used as the response variable and the expression levels for the remaining genes as dependent variables. The constructed GRN from $\boldsymbol{X}'_i$ is stored as a signed, weighted and directional graph, represented with a $p \times p$ adjacency matrix $\boldsymbol{W}_i$, each of whose columns stores the regression coefficients for the PC regression of a gene. $\boldsymbol{W}_i$ is then normalized via dividing by the maximal absolute value.

Substep 1.3. Denoising adjacency matrices to obtain the final GRN

Tensor decomposition is used to denoise the adjacency matrices $\{\boldsymbol{W}_i\}$ obtained from the PC regression step. First, the collection of $\{\boldsymbol{W}_i\}$ for t GRNs is processed as a third-order tensor Ξ, containing $p \times p \times t$ elements. Next, the CANDECOMP/PARAFAC (CP) decomposition is applied to decompose Ξ into components. Then, Ξ is reconstructed using top d components to obtain denoised tensors Ξ_d. Denoised $\{\boldsymbol{W}_i\}$ in Ξ_d are collapsed by taking the average of edge weights to obtain the final averaged matrix, $\boldsymbol{W}_d$.

Step 2: Generating pseudo-KO scGRN by virtually knocking out a gene

In the last step, the scRNA-seq expression data matrix from the WT sample, $\boldsymbol{X}$, is first used to construct the WT scGRN. In this step named *virtual KO*, the adjacency matrix of the WT scGRN, $\boldsymbol{W}_d$, is copied, and then the entire row of $\boldsymbol{W}_d$ corresponding to the target gene is set to 0 (**Figure 1B**). In this way, the virtual KO operation is performed on $\boldsymbol{W}_d$ directly. The modified $\boldsymbol{W}_d$ is denoted as $\widetilde{\boldsymbol{W}}_d$, that is, the adjacency matrix of pseudo-KO scGRN.

Step 3. Comparing scGRNs to identify virtual-KO perturbed genes

In this step, we assume that WT and pseudo-KO scGRNs, $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$, have been obtained. A quasi-manifold alignment method is then used to align $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$ (**Figure 1C**, see **Methods** for details). All genes included in the two scGRNs are projected in a k-dimensional space, where k << p. After the projection, each gene has two low-dimensional representations: one is in respect to $\boldsymbol{W}_d$ and the other $\widetilde{\boldsymbol{W}}_d$. For each gene j, d_j is the Euclidean distance between the gene's two projections. The greater d_j, the more significant the differential regulation. Finally, a χ^2 test is applied to detect significantly differentially regulated (DR) genes, i.e., *virtual-KO perturbed genes*.

A more detailed description of scTenifoldKnk modules is in the **Methods** section.

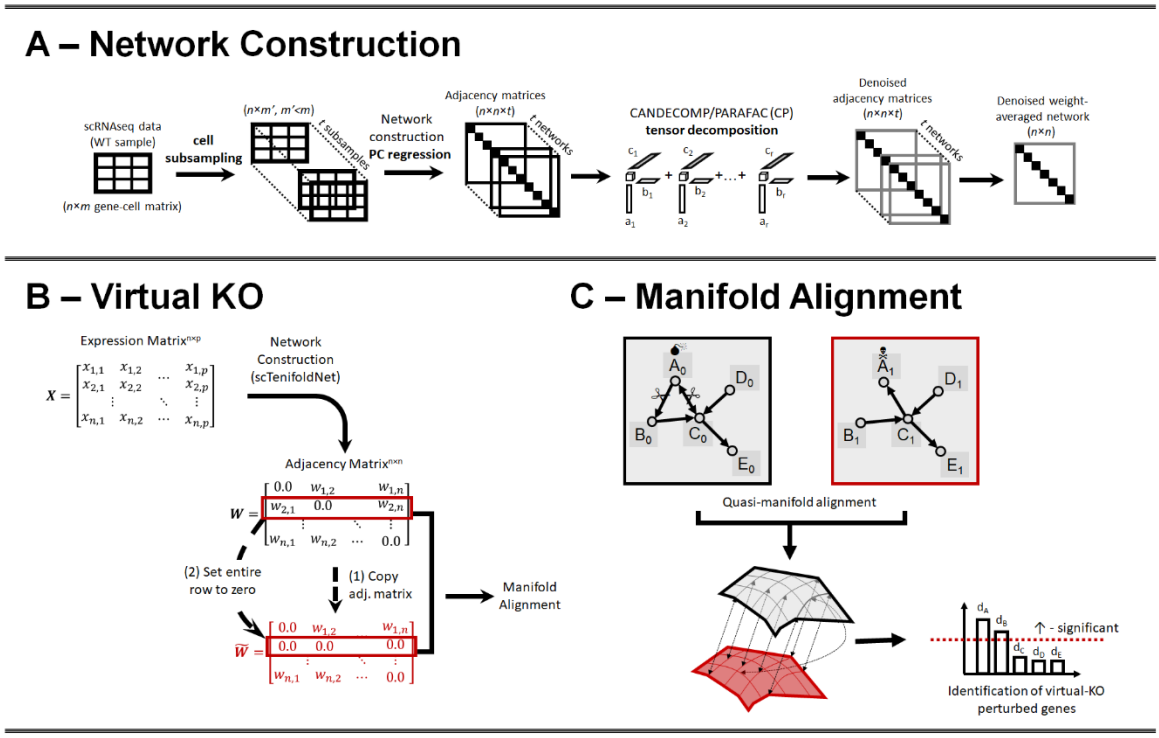

Figure 1. Overview of the scTenifoldKnk workflow.

scTenifoldKnk is a machine learning framework designed to perform virtual KO experiments with data from scRNA-seq. It consists of three main analytical modules: network construction, virtual KO, and manifold alignment. (**A**) Network construction. This module consists of three steps: cell subsampling, principal component regression, and tensor decomposition/denoising. (**B**) Virtual KO. This module starts by duplicating the WT adjacency matrix, $\boldsymbol{W}^{\langle 0 \rangle}$, to make $\boldsymbol{W}^{\langle 1 \rangle} = \boldsymbol{W}^{\langle 0 \rangle}$. Then, the entire row of $\boldsymbol{W}^{\langle 1 \rangle}$ corresponding to the KO gene is set to zero. The modified $\boldsymbol{W}^{\langle 1 \rangle}$ is called the pseudo-KO scGRN. (**C**) Quasi-manifold alignment. This method is used to learn latent representations of two networks, $\boldsymbol{W}^{\langle 0 \rangle}$ and $\boldsymbol{W}^{\langle 1 \rangle}$, and align them based on their underlying manifold structures. The distance between a gene's projections with respect to the two scGRNs on the low-dimensional latent representation is used to measure the level of differential regulation of the specific gene. The significantly differentially regulated genes are identified as *virtual-KO perturbed genes*.

Result 2: Virtual knockout experiments using simulated scRNA-seq data

We first used the simulated data to validate the relevance of our method. We generated a synthetic scRNA-seq data set using the simulator SERGIO—a single-cell expression simulator guided by GRNs [10]. To simulate the data, we supplied SERGIO with a predefined GRN with five co-expression modules of different sizes (containing 5, 10, 25, 40, and 20 genes, respectively). The simulated scRNA-seq data set was a sparse matrix (70% zeros) of 3,000 cells and 100 genes. We applied the PC regression method to the simulated data and constructed an scGRN. As expected, in the scGRN, genes were clearly clustered into five distinct modules (see **Figure 2A** for the adjacency matrix), mirroring the predefined modules given by the generative model of SERGIO. Genes in the same module were supposed to be functionally related or under the same regulation. Because SERGIO simulates gene expression in steady-state cells [10], we only used it to simulate co-expression modules rather than the regulatory processes of some upstream

regulators acting on these modules. We regarded the constructed network as the WT scGRN. Next, we set the weight of the 20th gene (gene #20) in the WT scGRN to zero to produce the pseudo-KO scGRN. Gene #20 belongs to the third module, which includes a total of 25 genes. We then used scTenifoldKnk to compare the pseudo-KO scGRN with the WT scGRN to identify genes significantly differentially regulated due to the KO. These genes were predicted likely to be perturbed along with gene #20. Because we knew the KO effect was due to the deletion of gene #20, we expected that the identified genes to be those closely correlated with gene #20. Indeed, as expected, scTenifoldKnk showed all significant genes were from the third module (**Figure 2B & 2C**, top), in which gene #20 is located. We repeated the analysis using genes #50 and #100 as another two examples. Again, the results were as expected (**Figure 2B & 2C**, middle and bottom). Thus, we concluded that when a member gene is knocked out from a tightly regulated module, other member genes in the same module should be detected by scTenifoldKnk. Algorithmically, genes in the same module with the KO gene were detected because their projected positions in low-dimensional latent representations of WT and pseudo-KO networks changed more than other genes not in the same module (see **Methods** for details).

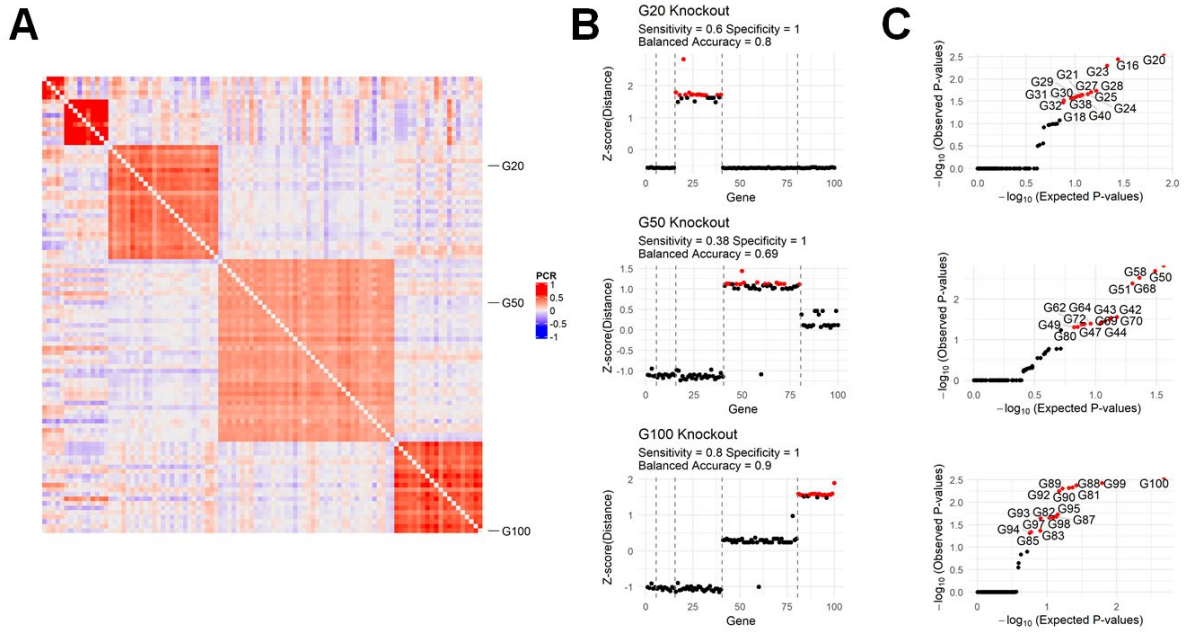

Figure 2. Simulations show that scTenifoldKnk specifically detects regulatory modules that include the KO gene.

(**A**) Heatmap of a 100 × 100 adjacency matrix of scGRN constructed from simulated scRNA-seq data of 100 genes and 3,000 cells. The color is scaled according to the normalized PC regression coefficient values between genes. The network contains 5 predefined co-regulated modules of different sizes, indicated by the blocks of gene pairs with a high correlation. The number of genes of each module is 5, 10, 25, 40, and 20, respectively. (**B**) The significance of the scTenifoldKnk DR test of genes in the simulated network is shown as -log(p-value). Red and black dots indicate whether genes are significant or not after false discovery correction. Assuming genes in the same module should be detected as significance genes, sensitivity is defined as = TP/(TP+FN), and specificity as = (TN)/(TN+FP), where T, P, F, and N stands for true, positive, false and negative, respectively. Balanced accuracy is defined as the average sensitivity and specificity values. (**C**) QQ-plots of expected vs. observed p-values of genes given by the DR tests.

Table 1. Summary of real-data applications of scTenifoldKnk analysis.

	KO gene	**Cell type**	**Main results from original and related studies**	**Original study**	**scTenifoldKnk predicted results** (i.e., enriched functions of virtual-KO perturbed genes)	**Figure and table in this article** (Number of virtual-KO perturbed genes)
Real-animal KO experiments with published data						
1	*Nkx2-1*	Pulmonary alveolar cells	Decreased expression of marker genes of alveolar type I (AT1) & II (AT2) cells; Increased expression of marker genes of gastrointestinal cells; *Nkx2-1* regulates expression of genes related to membrane composition, extracellular matrix, and cytoskeleton; mutations in Nkx2-1 interrupts AT2 cell function and identity [11, 12]	[11]	Marker genes of AT1 and AT2 cells; intestinal microvillus, cell cycle, and cytoskeleton; epithelial to mesenchymal transition led by the WNT signaling pathway members; surfactant homeostasis; lamellar body; cell adhesion molecules	**Figure 3A; Suppl. Table S1** (171 genes)
2	*Trem2*	Microglial cells	*Trem2* regulates expression of genes related to lipid transport and catabolism [13], lipid metabolism [14], microglial cell damage response [15], lysosome and phagosome function, Alzheimer's disease, and oxidative phosphorylation [16]; *Trem2* interacts with signaling transducer *Hcst* and adaptor *Tyrobp* [17]	[16]	Alzheimer's disease; oxidative phosphorylation; lysosome; TYROBP causal network; microglia pathogen phagocytosis pathway	**Figure 3B; Suppl. Table S2** (128 genes)
3	*Hnf4a & Hnf4g*	Intestinal villus	Increased expression of Goblet cell-	[18]	Enterocyte marker genes; electron transport chain; fat	**Figure 3C; Suppl.**

		epithelial cells	enriched genes (e.g., *Agr2, Spink4, Gcnt3* and *S100a6*) and genes in BMP/SMAD signaling pathway; decreased expression of enterocyte-enriched genes (e.g., *Npc1l1, Apoc3, Slc6a19* and *Lct*) genes involved in lipid metabolism, microvillus and absorption, and genes related to cytoplasm [18]		digestion and absorption; cholesterol metabolism; chylomicron assembly; cytoplasmic vesicle lumen	**Table S3** (65 genes)
Mendelian diseases						
4	*Cftr* (Cystic fibrosis)	Pulmonary alveolar cells	Expressed in epithelial cells of many organs [19]; mutations disrupt the function of the chloride channels, preventing them from regulating the flow of chloride ions and water across cell membranes [20, 21]	[22]	ABC transporter disorders and surfactant metabolism; ion transmembrane transporter activity, abnormal surfactant secretion, and alveolus morphology	**Figure 4A; Suppl. Table S4** (17 genes)
5	*Dmd* (Duchenne muscular dystrophy)	Skeletal myocytes	Mutation disrupts the linkage between the cytoskeleton and the glycoproteins of the extracellular matrix; impairment of muscle contraction; muscle cell necrosis [23, 24]	[25]	beta-1 integrin cell surface interaction, contractile actin filament bundle, actomyosin, extracellular matrix receptor interaction, extracellular matrix organization; abnormal collagen fibril morphology, abnormal skeletal muscle morphology, abnormal skeletal muscle fiber morphology	**Figure 4B; Suppl. Table S5** (190 genes)
6	*Mecp2* (Rett syndrome)	Neurons	Decreases the expression level of the genes involved in the BDNF signaling [26]; repressing the	[33]	Affected *REST* target genes; synaptic vesicle cycle, GABA synthesis, release, reuptake and degradation, syntaxin	**Suppl. Figure S2; Suppl. Table S6 and S7** (377 and 322

			transcription factor REST [26]; dysregulated maintenance of normal neuronal functions [27, 28]; altered synapses and synaptic vesicle proteins [29]; defects of GABAergic synapses [30]; autism-like stereotypies [31]; syntaxin-1 mutant phenotype [32]		binding, transmission across chemical synapses	genes, respectively)
Real-animal KO experiments (data generated in this study)						
7	*Ahr*	Enterocytes	*Ahr* deletion in intestinal epithelial cells results in an unreliable barrier and uncontrolled proliferation of intestinal cells, resulting in malignant transformation [34]. Enterocyte marker genes, which may lead to the alteration of enterocyte differentiation fate [35]. Intestinal-specific *Ahr* KO increase basal stem cell and crypt injury-induced cell proliferation and promotes colon tumorigenesis [36, 37].	This study	Myc active pathway, nuclear receptors meta-pathway, MHC class II protein complex binding, chaperone-mediated protein complex assembly, response to unfolded protein, RNA binding, translation elongation factor activity, cytoplasmic vesicle lumen, T cell receptor regulation of apoptosis and secretory granule lumen	**Figure 5; Suppl. Table S8** (53 genes)
8	*Malat1*	Pancreatic beta cells	Malat1 known to be associated adenocarcinoma; involved in the generation of reactive oxygen species, cell cycle regulation, cancer metastasis, cell migration, and hyperglycemia [38-42]. Real KO experiment showed: insulin	This study	Cori cycle; glycolysis and gluconeogenesis; regulation of insulin secretion, cellular carcinomas; cytoskeleton; circadian rhythm; hypoxia-inducible factor (HIF) signaling pathway	**Figure 6; Suppl. Table S10** (167 genes)

			secretion; cellular respiration; circadian rhythm			

Result 3: scTenifoldKnk virtual KO analysis recapitulates results of real KO experiments

As a virtual KO tool, scTenifoldKnk is expected to recapitulate results obtained using real KO experimental data. To prove this, we applied scTenifoldKnk to three scRNA-seq data sets from three real KO experiments. In all three cases, the scRNA-seq data sets of the original studies contained expression matrices from both WT and KO samples. When applying scTenifoldKnk to each of these data sets, we did not use the data from the KO samples, because the virtual KO does not need it. Instead, we used scTenifoldKnk to generate pseudo-KO scGRN, which was used as a surrogate to replace the scGRN generated from the real KO data.

Virtual vs. real KO experiment 1: *Nkx2-1* is required for the transcriptional control, development, and maintenance of alveolar type-1 and type-2 cells

NK homeobox 2-1 (*Nkx2-1*) is a transcription factor expressed in lung epithelial cells of alveolar type I (AT1) and type II (AT2). AT1 cells cover 95% of the gas exchange surface and are 0.1 μm thick to allow passive oxygen diffusion into the bloodstream. *Nkx2-1* is essential at all developmental stages of AT1 cells. Loss of *Nkx2-1* results in the impairment of three main defining features of AT1 cells, molecular markers, expansive morphology, and cellular quiescence [11]. AT2 cells are cuboidal and secrete surfactants to reduce surface tension. Mutations in *Nkx2-1* interrupt the expression of *Sftpb* and *Sftpc*, two genes related to AT2 cell function and molecular identity [11, 12].

To examine the molecular and cellular changes underlying the mutant phenotypes in mouse lungs caused by the *Nkx2-1*$^{-/-}$ KO, Little, Gerner-Mauro [11] generated a comprehensive set of data using the lung samples from WT and *Nkx2-1*$^{-/-}$ KO mice. Using bulk RNA-seq and immunostaining assays, they observed that the expression of marker genes of AT1 and AT2 cells was downregulated in the *Nkx2-1* mutant cells. They also found that the expression of marker genes for gastrointestinal cells was upregulated in *Nkx2-1*$^{-/-}$ mutant AT1 cells, which form dense microvilli-like structures apically. Using ChIP-seq, they found that *Nkx2-1* binds to a set of genes implicated in regulating the cytoskeleton, membrane composition, and extracellular matrix. Little, Gerner-Mauro [11] also generated scRNA-seq data for 2,312 and 2,558 epithelial cells from lung samples of the WT and *Nkx2-1*$^{-/-}$ KO mice, respectively.

We obtained the scRNA-seq data, generated by Little, Gerner-Mauro [11], and used the expression matrix of 8,647 genes × 2,312 cells from WT mice as the input for scTenifoldKnk. We constructed the WT scGRN and then knocked out *Nkx2-1*. The final report of scTenifoldKnk analysis contained 171 significant genes (FDR < 0.05, **Supplementary Table S1**). These virtual-KO perturbed genes included seven marker genes of AT1 cells (*Egfl6*, *Ager*, *Cldn18, Icam1, Crlf1, Gprc5a,* and *Aqp5*) and 25 marker genes of AT2 cells (highlighted in **Supplementary Table S1**). Enrichr functional enrichment test [43] indicated that these genes were enriched for the following functional categories: *epithelial to mesenchymal transition led by the WNT signaling pathway members*, *surfactant homeostasis*, *lamellar body,* and *cell adhesion molecules*. These enriched function annotations were known to be related to the function of AT2 cells. Next, we applied the interaction enrichment analysis to the 171 significant genes. The interaction enrichment analysis was based on the STRING database [44]. We found that these significant genes appeared in a

fully connected component in the STRING interaction network, which is unexpected by chance ($P < 0.01$, STRING interaction enrichment test), indicating a closely related functional relationship between those genes. We subsequently performed the gene set enrichment analysis (GSEA [45]) to evaluate the extent of perturbation caused by the *Nkx2-1* KO at the transcriptome-wide level. The GSEA analysis identified gene sets containing marker genes of AT1 and AT2 cells (FDR < 0.01 in both cases). That is to say, AT1 and AT2 marker genes were among the topmost perturbed genes caused by the deletion of *Nkx2-1*. GSEA analysis also showed that the *Nkx2-1*$^{-/-}$ KO impacted genes with functions related to intestinal microvillus (**Figure 3A**), cell cycle, and the cytoskeleton. These results are consistent with those reported in the original study [11].

Again, we emphasize that scTenifoldKnk did not use data generated from the KO mice. Instead, all results of scTenifoldKnk were derived in silico using data generated from the WT mice. Overall, the results of our scTenifoldKnk analysis are consistent with those of the original study [11], in which scRNA-seq data was experimentally derived from KO mice and combined with other experimental techniques to characterize the KO individual.

Virtual vs. real KO experiment 2: *Trem2* regulates microglial cholesterol metabolism

The Triggering Receptor Expressed on Myeloid cells 2 (*Trem2*) is a single-pass transmembrane immune receptor selectively expressed in microglia within the central nervous system. *Trem2* is known to be involved in late-onset Alzheimer's disease and plays a role in modulating proliferation, survival, immune response, calcium mobilization, cytoskeletal dynamics, mTOR signaling, autophagy, and energy metabolism [46]. The function of *Trem2* is known to be mediated via signaling transducer *Hcst* and adaptor *Tyrobp* [17]. *Trem2* is also known to play a role in regulating lipid metabolism, with most studies focusing on lipids in the form of either lipoprotein particles or cell surface-exposed signals, such as candidate *Trem2* ligands [13]. By comparing WT and *Trem2*$^{-/-}$ KO mice, Poliani, Wang [15] showed that *Trem2* regulates many genes, such as *Apoe* and *Lpl*, which control lipid transport and catabolism in microglia. *Trem2* was also found to modulate gene expression of macrophages in adipose and control blood cholesterol metabolism in obese mice [14], further linking the function of *Trem2* to lipid metabolism. To examine whether *Trem2* mediates myelin lipid processing in microglia, Nugent, Lin [16] isolated and characterized *Cd11b*$^{+}$*/Cd45*^{low} microglial cells from *Trem2*$^{+/+}$, *Trem2*$^{+/-}$, and *Trem2*$^{-/-}$ mice, fed with a 0.2% demyelinating cuprizone diet for 12 weeks. They analyzed a comprehensive set of analytical data using FACS, RNA-seq, scRNA-seq, and lipidomics. They reported that *Trem2* upregulates *Apoe* and other genes involved in cholesterol transport and metabolism, causing robust intracellular accumulation of a storage form of cholesterol upon chronic phagocytic challenge. *Trem2* was also shown to regulate the expression of genes associated with cell damage response, lysosome and phagosome function, Alzheimer's disease, and oxidative phosphorylation [16].

To perform the virtual KO analysis, we obtained scRNA-seq data from WT (*Trem2*$^{+/+}$) mice [16]. The expression matrix contained data of 7,715 genes and 765 *Cd11b*$^{+}$*/Cd45*^{low} microglial cells. We used this WT expression matrix as the input and used scTenifoldKnk to knock out *Trem2*. The final results of scTenifoldKnk analysis contained 128 virtual-KO perturbed genes (FDR < 0.05, **Supplementary Table S2**). The Enrichr analysis [43] showed that these genes were enriched with *Alzheimer's disease*, *oxidative phosphorylation*, *lysosome*, *TYROBP causal network*, *metabolic pathway of LDL, HDL and TG*, and *microglia pathogen phagocytosis pathway*. Such an enrichment indicates that the proteins were at least partially biologically connected as a group. These virtual-KO perturbed genes were highly interactive with each other, as shown by their positions on the STRING interaction network. The network of virtual-KO perturbed genes had

significantly more interactions than expected ($P < 0.01$, STRING interaction enrichment test), which means that gene products exhibited more interactions among themselves than what would be expected for a random set of proteins of similar size, drawn from the genome. This result suggests that the identified virtual-KO perturbed genes are closely related to shared functions. Collectively, these scTenifoldKnk findings provide insight into understanding *Trem2* function by revealing the list of genes perturbed following *Trem2* deletion (**Figure 3B**). We show that the scTenifoldKnk results recapitulate those reported in the original study [16], demonstrating the utility of scTenifoldKnk.

Next, to assess the robustness of scTenifoldKnk, we randomly subsampled 500 cells from each set of replicates and performed the virtual KO of *Trem2*. We repeated the process independently five times with each set. With each of the subsamples, scTenifoldKnk produced a ranked gene list. We found that these ranked gene lists were positively correlated with each other (average rank correlation coefficient = 0.55, Spearman's tests, **Supplementary Figure S3A**). GSEA analyses with these ranked gene lists produced similar results (**Supplementary Figure S3B**). The most enriched functions included *Oxidative phosphorylation*, *Alzheimer's disease*, *Cholesterol metabolism*, *Phagosome*, and *Lysosome*, which were shared among all subsamples. These results suggest scTenifoldKnk is robust against the heterogeneity of input cells.

Virtual vs. real KO experiment 3: *Hnf4a* and *Hnf4g* stabilize enterocyte identity

Hepatocyte nuclear factor 4 alpha and gamma, *Hnf4a* and *Hnf4g*, are transcription factors that regulate gene expression in the gut epithelium. *Hnf4a* and *Hnf4g* function redundantly, and thus, an independent deletion of one paralog causes no gross abnormalities [18, 47]. *Hnf4a* and *Hnf4g* double-KO $Hnf4ag^{DKO}$ mice exhibit fluid-filled intestines indicative of intestinal malfunction [18]. Epithelial cells in the $Hnf4ag^{DKO}$ mutants fail to differentiate. Using the bulk RNA-seq data, Chen, Toke [18] compared gene expression in duodenal epithelial cells isolated from WT mice and $Hnf4a^{KO}$, $Hnf4g^{KO}$ and $Hnf4ag^{DKO}$ mutants. They identified 2,892 DE genes in the $Hnf4ag^{DKO}$ mutant but only 560 and 77 in the $Hnf4a^{KO}$ and $Hnf4g^{KO}$ mutants, respectively (FDR < 0.05, absolute log2(fold change) > 1). The DE genes identified in the $Hnf4ag^{DKO}$ enterocytes were enriched for functions in digestive metabolisms such as lipid metabolism, microvillus, and absorption, as well as enterocyte morphology, cytoplasm, Golgi apparatus, and immune signaling. $Hnf4ag^{DKO}$ epithelium exhibited a robust shift in the transcriptome away from differentiated cells toward proliferating and Goblet cells, suggesting that $Hnf4ag^{DKO}$ impairs enterocyte differentiation and destabilize enterocyte identity. To validate their findings, Chen, Toke [18] used scRNA-seq to measure gene expression in intestinal villus epithelial cells. They obtained scRNA-seq data for 4,100 and 4,200 cells from $Hnf4ag^{WT}$ and $Hnf4ag^{DKO}$ respectively and confirmed that compared to the WT, mutant epithelial cells show increased Goblet cell-enriched genes, such as *Agr2, Spink4, Gcnt3* and *S100a6*, and decreased expressions of enterocyte-enriched genes, such as *Npc1l1, Apoc3, Slc6a19* and *Lct* (**Figure 3C** left panel). The $Hnf4ag^{DKO}$ mutant cells also showed increased expression for genes in the BMP/SMAD signaling pathway and decreased expression of genes involved in lipid metabolism, microvillus and absorption, and genes related to the cytoplasm [18].

We obtained the scRNA-seq expression matrix of 4,100 cells from the $Hnf4ag^{WT}$ samples used as input for scTenifoldKnk. We constructed the WT scGRN of 2,591 genes and then virtually knocked out both *Hnf4a* and *Hnf4g* genes at the same time. The final result of scTenifoldKnk contained 65 virtual-KO perturbed genes (FDR < 0.05, **Supplementary Table S3**). These genes were enriched with *electron transport chain*, *fat digestion and absorption*, *cholesterol metabolism*,

chylomicron assembly, and *cytoplasmic vesicle lumen*. A search of the STRING database [44] indicated that all virtual-KO perturbed genes form a fully connected network module. STRING is a database of known and predicted protein-protein interactions. The interaction enrichment test, a statistical test provided by the STRING database, indicated that such a full interconnection of 65 genes is less likely to be expected by chance ($P < 0.01$). Furthermore, GSEA analysis revealed that these virtual-KO perturbed genes were enriched with canonical marker genes of enterocytes (**Figure 3C**), which is consistent with the finding of the original study [18].

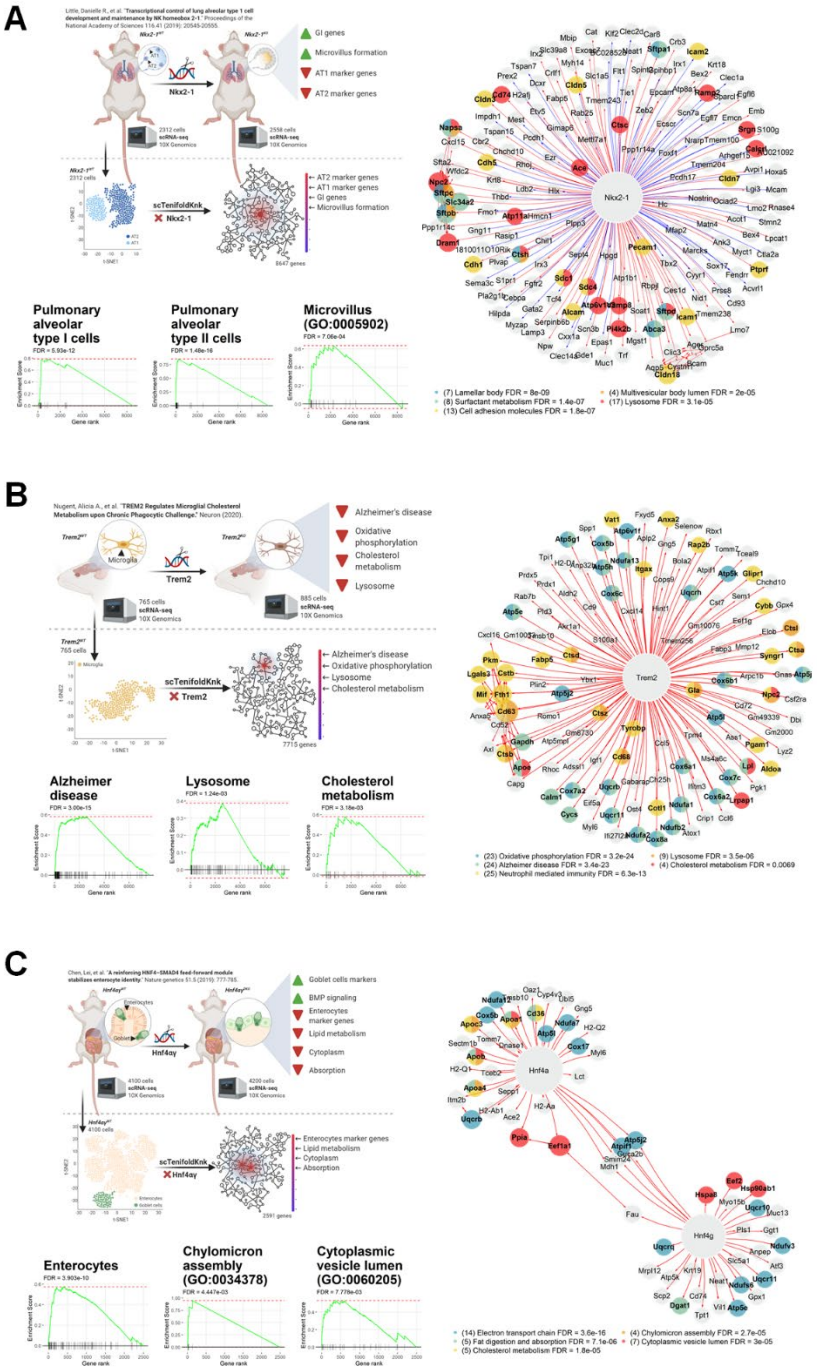

Figure 3. scTenifoldKnk virtual KO analysis recapitulates the findings of real KO experiments.
(**A**) Virtual KO of *Nkx2*-1 identifies gene expression program changes reported in the original study. GSEA analysis identifies significant gene sets, including *pulmonary alveolar type I cells*, *pulmonary alveolar type II cells*, and *microvillus* (GO:0005920 under cellular component *actin-based cell projection*). The egocentric plot shows virtual-KO perturbed genes connected to *Nkx2-1*. Nodes are color-coded by each gene's membership association with significantly enriched functional groups. (**B**) Virtual KO of *Trem2* identifies gene expression program changes reported in the original study. GSEA analysis identifies significant gene sets, including *Alzheimer's disease*, *lysosome*, and *cholesterol metabolism*. The egocentric plot shows virtual-KO perturbed genes connected to *Trem2*. Nodes are color-coded by each gene's membership association with significantly enriched functional groups. (**C**) Virtual double KO of *Hnf4a* and *Hnf4g* identifies gene expression program changes reported in the original study. GSEA analysis identifies significant gene sets, including *enterocytes, chylomicron assembly*, and *cytoplasmic vesicle lumen*. The egocentric plot shows virtual-KO perturbed genes connected to *Hnf4a* and *Hnf4g*. Nodes are color-coded by each gene's membership association with significantly enriched functional groups.

Result 4: scTenifoldKnk reveals functions of Mendelian disease genes

Mendelian diseases are a family of diseases caused by the loss or malfunctioning of a single gene. For many Mendelian diseases, we know a great deal about their genetic basis and pathophysiological phenotypes [48]. Thus, we decided to use three Mendelian diseases as "positive controls" to test the performance of scTenifoldKnk. We attempted to determine whether scTenifoldKnk can accurately predict the molecular phenotypic consequences of gene deletion. The three chosen Mendelian diseases were cystic fibrosis, Duchenne muscular dystrophy, and Rett syndrome. As described in more detail below, in each case, we performed scTenifoldKnk analysis using existing scRNA-seq data generated from cell types that are most relevant to the disease conditions (**Table 1**).

Cystic fibrosis

Cystic fibrosis (CF) is one of the most common autosomal recessive diseases [49]. It is caused by mutations in *CFTR*, a gene encoding for a transmembrane conductance regulator [20, 50], which functions as a channel across the membrane of cells that produce mucus, sweat, saliva, tears, and digestive enzymes. The CFTR protein also regulates the function of other channels. *CFTR* is expressed in epithelial cells of many organs, including the lung, liver, pancreas, and digestive tract [19]. The most common *CFTR* mutation that causes CF is the deletion of phenylalanine 508 (ΔF508), which disrupts the function of the chloride channels, preventing patients from regulating the flow of chloride ions and water across cell membranes [20]. The truncated CFTR protein leads to a reduction of surfactant and causes the build-up of sticky, thick mucus that clog the airways, increasing the risk of bacterial infections and permanent lung damage [21].

To test scTenifoldKnk, we obtained scRNA-seq data from 7,326 pulmonary alveolar type II (AT2) cells in the GEO database (access number: GSM3560282). The original data sets were generated by Frank, Penkala [22] to study the lineage-specific development of alveolar epithelial cells in mice. The original study was not directly focused on CF. Nevertheless, with the downloaded data, we constructed a WT scGRN that contained 7,107 genes. We then used scTenifoldKnk to knock out *Cftr*. The final results of scTenifoldKnk contained 17 virtual-KO perturbed genes: *Cftr*, *Birc5, Cldn10, Cxcl15, Dcxr, Hmgb2, Lamp3, Mgst1, Npc2, Pclaf, Pglyrp1, Sftpa1, Sftpb, Sftpc, Smc2, Tspan1* and *Tubb5* (FDR < 0.05, **Figure 4A, Supplementary Table S4**). Among them, *Sftpa1, Sftpb,* and *Sftpc* have functions associated with *ABC transporter disorders* and *surfactant*

metabolism. *Sftpb* and *Sftpd* encode for surfactant proteins implicated in CF and innate immunity [51, 52]. *Ctsh* encodes for cathepsin H, a cysteine peptidase, which is involved in the processing and secretion of the pulmonary surfactant protein B [53]. GSEA analysis of the ranked gene list was conducted using gene sets of the MGI mammalian phenotypes database as a reference. The result showed that the gene list scTenifoldKnk produced was significant in terms of *ion transmembrane transporter activity*, *abnormal surfactant secretion*, and *alveolus morphology* (FDR < 0.01 in all cases, **Figure 4A**). These results are consistent with the known pathophysiological changes resulting from the loss of *Cftr* function in the lungs.

To test the specificity of scTenifoldKnk, we decided to use genes that show highly similar expression profiles to that of *Cftr* to repeat the analysis. The selection of these genes was made by computing for each gene both the mean and variance of their expression levels across cells. Five closest genes to *Cftr—Akap7, Zranb1, Krcc1, Mta1* and *Rps8*—were selected. These selected genes may not necessarily have any functions related to that of *Cftr* in alveolar type II cells. When scTenifoldKnk was applied to predict perturbation profiles caused by their deletion, we found none of the sets of predicted functions similar to those specific functions (i.e*., ion transmembrane transporter function, abnormal surfactant secretion,* and *alveolus morphology*) that are associated with *Cftr*. A principal component analysis (PCA) was used to confirm that perturbation profiles of *Mta1*, *Akap7* and *Zranb1* differ most from that of *Cftr*, while the perturbation profiles of *Krcc1* and *Rps8* are relatively more similar to that of *Cftr* (**Supplementary Figure S4**). The two genes closer to *Cftr*, *Krcc1* and *Rps8* have been reported to play a role in regulating cellular plasticity and fibrotic process [54, 55]. These results suggest that scTenifoldKnk specifically reveals gene function using the KO gene's perturbation profile rather than the KO gene's expression profile.

Duchenne muscular dystrophy

Duchenne muscular dystrophy (DMD) arises as a result of mutations that affect the open reading frame of *DMD* [56, 57]. The *DMD* gene encodes dystrophin, a large cytoskeletal structural protein, which is mostly absent in DMD patients [23]. The absence of dystrophin results in a disturbance of the linkage between the cytoskeleton and the glycoproteins of the extracellular matrix, generating an impairment of muscle contraction, eventually leading to muscle cell necrosis [23, 24](**Figure 4B**, left).

We obtained the scRNA-seq data of 5,159 muscle cells from the mouse limb (quadriceps) in the GEO database (Access number: GSM4116571). The original data was generated to study gene expression patterns in skeletal muscle cells [25]. The original study was not focused on DMD. We used scRNA-seq data from normal tissue to construct the WT scGRN of 9,783 genes. We subsequently performed the scTenifoldKnk virtual KO analysis to predict the molecular phenotype due to the impact of the *Dmd* KO. The final results of scTenifoldKnk included 190 virtual-KO perturbed genes (FDR < 0.05, **Supplementary Table S5**). These genes were enriched with functions related to *beta-1 integrin cell surface interaction*, *contractile actin filament bundle*, *actomyosin*, *extracellular matrix receptor interaction*, and *extracellular matrix organization* (**Figure 4B**, middle). The GSEA analysis against the MGI mammalian phenotype database gave the following top hits (FDR < 0.01): *abnormal collagen fibril morphology*, *abnormal skeletal muscle morphology*, and *abnormal skeletal muscle fiber morphology* (**Figure 4B**, right). These phenotype terms represent consistently with known effects of the loss of DMD function in muscle cells, verifying that scTenifoldKnk can predict phenotypic effects caused by gene KO that are pertinent for the biological context.

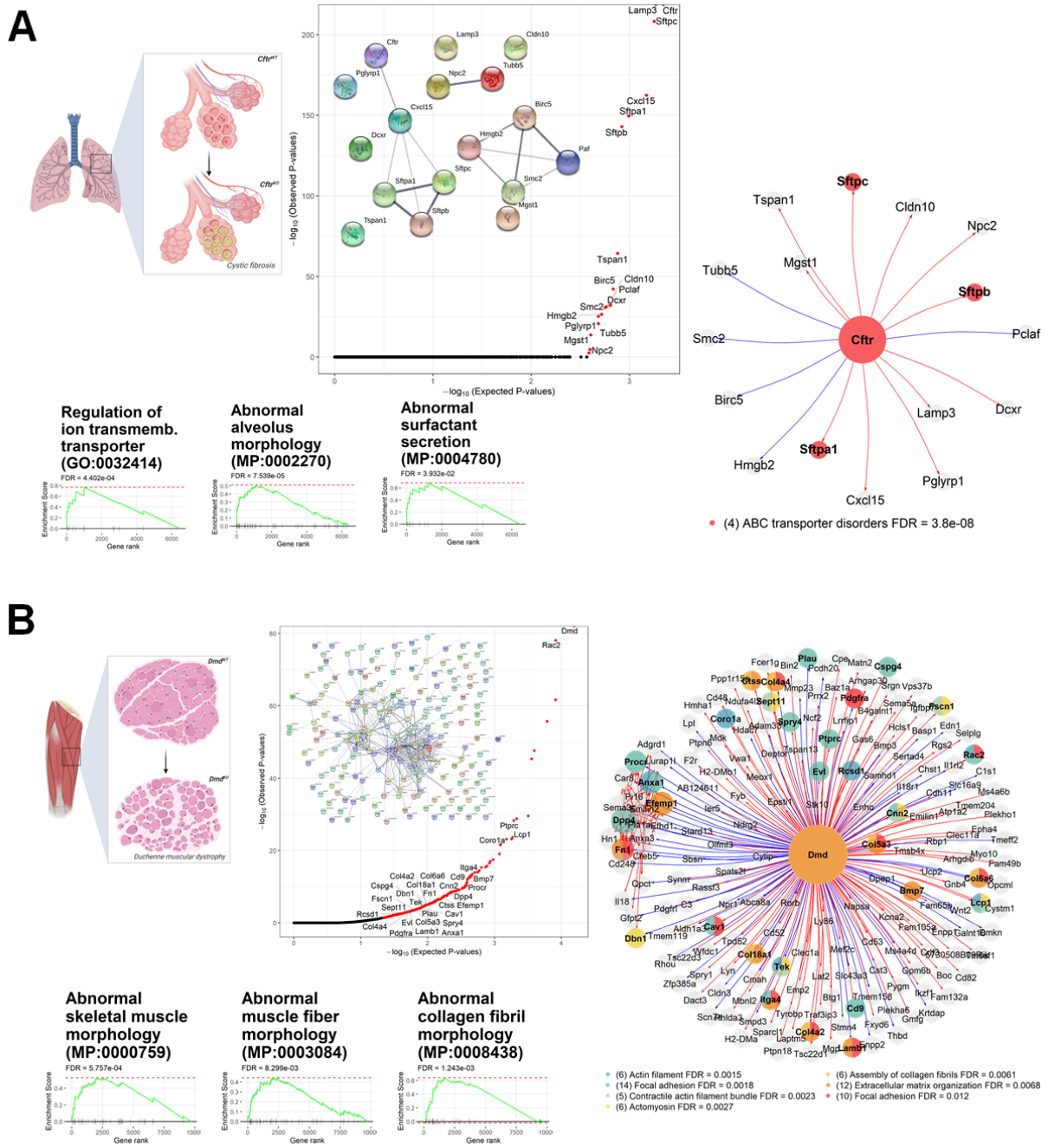

Figure 4. scTenifoldKnk virtual KO analysis reveals the function of Mendelian disease genes in relevant cell types.

(**A**) Virtual KO of *Cftr* in pulmonary alveolar type II cells identifies gene expression program changes associated with cystic fibrosis. GSEA analysis identifies significant gene sets, including *regulation of ion transmembrane transporter*, *abnormal alveolus morphology*, and *abnormal surfactant secretion*. The egocentric plot shows virtual-KO perturbed genes connected to *Cftr*. Nodes are color-coded by each gene's membership association with significantly enriched functional groups. QQ-plot of genes and interconnection of virtual-KO perturbed genes in STRING are given. (**B**) Virtual KO of *Dmd* in muscular cells identifies gene expression program changes associated with Duchenne muscular dystrophy. GSEA analysis identifies significant gene sets, including *abnormal skeletal muscle morphology*, and *abnormal collagen fibril morphology*. The egocentric plot shows virtual-KO perturbed genes connected to *Dmd*.

Rett syndrome

The third Mendelian disease we considered was the Rett syndrome (RTT, MIM 312750), which is a severe neurodevelopmental disease [58, 59]. RTT is known to be caused by mutations in *Mecp2,* a transcriptional repressor required to maintain normal neuronal functions [27, 28]. *Mecp2* deficiency in the brain decreases the expression level of genes involved in the *Bdnf* signaling pathway, mediated by repressing the transcription factor *Rest* [26].

We obtained scRNA-seq data generated from mouse neurons (SRA database access numbers: SRX3809326 and SRX3809327) for the mouse brain atlas project [33]. The two data sets contain 2,054 and 2,156 neurons, respectively, derived from two CD1 P19 female mice that served as biological replicates. We analyzed the two data sets independently to see whether scTenifoldKnk could, as expected, produce similar results with data gathered from biological replicates. Two scGRNs containing 8,652 and 8,555 genes were constructed first. The scTenifoldKnk analysis of virtual KO of *Mecp2* produced 377 and 322 virtual-KO perturbed genes, respectively (FDR < 0.05, **Supplementary Table S6 & S7**), including 211 shared genes. The number of shared genes was significantly greater than random expectation ($P < 1x10^{-5}$, hypergeometric test), indicating a high overlap rate between results of scTenifoldKnk analyses when applied to the two data sets from biological replicates. We also compared two ranked gene lists generated from the analysis of two replicate data sets. If scTenifoldKnk results are robust, we expected that the relative positions of the same genes in the two ranked lists should be more similar to each other than with a list of randomly ranked genes. Indeed, we found the correlation between rankings of two reported lists was positive (Spearman's correlation coefficient = 0.68) and highly significant ($P < 1x10^{-12}$).

Many of these 211 genes were found to be *targets of the transcription factor Rest* (FDR < 0.01). The enriched functions include *axon*, *synaptic vesicle cycle*, *GABA synthesis, release, reuptake and degradation*, *syntaxin binding*, and *transmission across chemical synapses*. GSEA analyses using ranked gene list as input showed *BDNF signaling pathway* was highly significant (FDR < 0.01, for both replicates, **Supplementary Figure S2**). These results are consistent with previous experimental results. For example, it is known that the most prominent alterations in gene products due to *Mecp2* KO are related to synapses and synaptic vesicle proteins [29]. At the phenotypic level, the *Mecp2* KO causes early defects in GABAergic synapses [30] and mediates autism-like stereotypies and RTT phenotypes [31]. Mutations in syntaxin-1 are known to elicit phenotypes similar to those found in RTT [32].

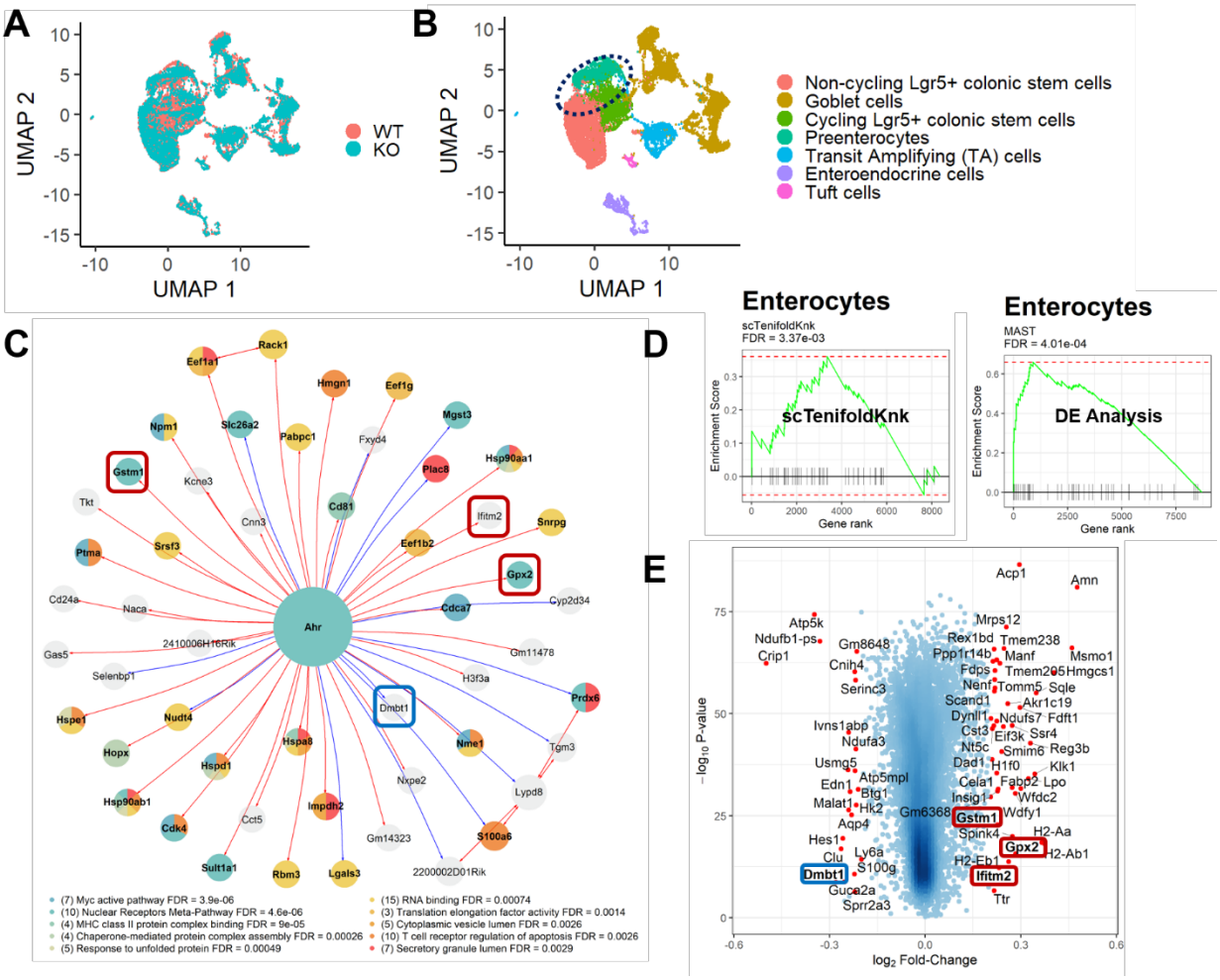

Figure 5. scTenifoldKnk predicts transcriptional changes in the enterocytes of Ahr$^{-/-}$ KO mice. (**A**) UMAP plot of cells distinguished by experimental groups: WT or KO. (**B**) Cell clustering and cell type assignment. Enterocytes included in the analysis are highlighted with the dashed oval. (**C**) Egocentric plot showing virtual-KO perturbed genes connected to *Ahr*. Nodes are color-coded by each gene's membership association with the top 10 significantly enriched functional groups. Proliferative genes (associated with the Myc pathway) were identified as the expected topmost perturbed genes after the deletion of *Ahr* in enterocytes. (**D**) GSEA analyses with the DE gene list and scTenifoldKnk gene list produce the same result—gene sets of *enterocytes*. (**E**) Volcano plot from the DE analysis comparing gene expression between WT and KO samples. In the plot, -log2(FC) and -log10(P-value) of genes are plotted against each other. Names of genes are given for those with a fold change (FC) > 10% and FDR < 0.05. Among them, four virtual-KO perturbed genes are highlighted with a red (*Gstm1*, *Gpx2* and *Ifitm2*) or a blue (*Dmbt1*) box, indicating up-regulated or down-regulated in the KO sample.

Result 5: scTenifoldKnk predicts real KO experimental results

In this section, we describe two *in vivo* KO experiments that we performed. In each of the experiments, we generated scRNA-seq data from both WT and KO mice. The WT data was used to perform scTenifoldKnk virtual KO analysis and predict the consequences of gene KO. The comparative studies between the data sets from WT and KO mice were conducted to validate the prediction made.

scTenifoldKnk predicts transcriptional changes in enterocytes of Ahr$^{-/-}$ KO mice

The Aryl hydrocarbon receptor (*AhR*) is a transcription factor that regulates the gene expression of metabolic enzymes, such as cytochrome P450s. It acts primarily as a sensor for both dietary and gut microbial-derived ligands. These include structurally diverse phytochemicals, pharmaceuticals, and endogenous compounds, including tryptophan metabolites and xenobiotic chemicals. *AhR* also plays a role in regulating immunity, stem cell maintenance, and cellular differentiation [34, 37, 60-62].

To study the role of Ahr in intestinal stem cell fate determination, we generated inducible, stem cell-specific *Ahr*$^{-/-}$ KO mice (see **Methods**). We then used scRNA-seq to generate data for five WT and five KO mice and obtained 13,626 and 8,425 sequenced cells from WT and KO mice, respectively. Data was pooled and projected to low-dimensional embedding, showing no global shift in overall gene expression profiles between cells from WT and KO mice (**Figure 5A**). We subsequently clustered cells into groups and identified seven distinct cell types: non-cycling Lgr5^{+} colonic stem cells, Goblet cells, cycling Lgr5^{+} colonic stem cells, pre-enterocytes, transit-amplifying cells, enteroendocrine cells, and Tuft cells (**Figure 5B**). Since enterocytes are the most abundant epithelial cell lineage in the intestine, we decided to conduct scTenifoldKnk using data from enterocytes (indicated by dashed oval in **Figure 5B**) to assess the function of *Ahr*.

We conducted scTenifoldKnk analysis to knock out *Ahr* from the scGRN of WT mice. The scGRN was constructed from an input expression matrix of 1,029 cells (i.e., WT enterocytes) and 8,478 genes. The output of scTenifoldKnk was a ranked gene list including 53 top genes, i.e., virtual-KO perturbed genes (FDR < 0.05, **Supplementary Table S8**). Enrichr functional enrichment tests indicated that these genes were associated with *Myc active pathway*, *nuclear receptors meta-pathway*, *MHC class II protein complex binding*, *Chaperone-mediated protein complex assembly*, *response to unfolded protein*, *RNA binding*, *translation elongation factor activity*, *cytoplasmic vesicle lumen*, *T cell receptor regulation of apoptosis*, and *secretory granule lumen* (**Figure 5C**). Next, a STRING database [44] search indicated that identified virtual-KO perturbed genes form two network modules (both $P < 0.01$, STRING interaction enrichment tests, **Supplementary Figure S5**). Gene members in each module are fully connected, further confirming closely related functional associations among identified genes. Finally, GSEA analysis against a comprehensive collection of gene sets (see **Methods**) identified 60 non-redundant functional gene sets (**Supplementary Table S9**), including many related to cell cycle and cell proliferation signaling.

GSEA analysis against the canonical marker gene sets (retrieved from PanglaoDB, see **Methods**) showed that *enterocytes* was the most significant hit (**Figure 5D**, left), suggesting that *Ahr* KO induces a change in the expression of enterocyte marker genes, which may be responsible for the alteration of enterocyte differentiation fate [35]. Taken together, our comprehensive functional annotation for the virtual-KO perturbed genes revealed the multifactorial functions of *Ahr* as a key player of enterocyte cell differentiation and proliferation. These results are consistent with the known functions of *Ahr*. For example, a previous study showed that *Ahr* deletion in intestinal epithelial cells resulted in an abnormal barrier function and uncontrolled proliferation of intestinal cells, promoting malignant transformation [34]. Organoids with dysregulated *Ahr* had substantially compromised differentiation to goblet and enterocyte cell fate [35]. Intestinal-specific Ahr KO increases basal stem cell and crypt injury-induced cell proliferation and promotes colon tumorigenesis [36]. *Myc* is a key transcriptional effector of the Wnt signaling pathway through its actions to promote cellular proliferation [63].

Without scTenifoldKnk, we would have to use differential expression (DE) analysis. Indeed, DE analysis has been a dominant method for assessing the impact of gene KO (e.g., as in [11, 16, 18]). Here, we put ourselves in such a hypothetical scenario—that is, scTenifoldKnk is not available. We resorted to comparing gene expression levels between the WT and *Ahr* KO mice. To identify DE genes, we used function FindMarkers in the R package Seurat [64] with its default parameter setting. This function calls an established method for scRNA-seq DE analysis—MAST [65]. A total of 68 DE genes were identified [FDR < 0.05 and log(FC) > 0.25], including 46 upregulated and 22 downregulated genes in the *Ahr*-KO enterocytes (**Figure 5E**). Three up-regulated DE genes, *Gstm1*, *Gpx2*, and *Ifitm2*, and one down-regulated DE gene, *Dmb1t*, are also among the list of 53 virtual-KO perturbed genes identified by scTenifoldKnk (highlighted with red and blue boxes, respectively, in **Figure 5C & 5E**). Note that, in the WT scGRN, the sign of edges from *Ahr* to the three up-regulated genes is positive (shown as red edges in **Figure 5C**), while the sign of edge from *Ahr* to the down-regulated *Dmb1t* is negative (shown using the blue edge between *Ahr* and *Dmb1t* in **Figure 5C**). The GSEA analysis of the DE gene list, with the comprehensive gene sets as the references (**Methods**), reported 35 non-redundant significant gene sets, with 16 overlaps with the 60 non-redundant significant gene sets in **Supplementary Table S9**, which were identified using scTenifoldKnk ranked gene list. The overlap is significantly larger than expected by chance ($P < 10^{-20}$, hypergeometric test, assuming $N = 10^9$ of non-redundant gene sets). In addition, GSEA analysis of the DE gene list with marker gene sets as references showed that *enterocytes* was one of the significant hits (**Figure 5D, right**), which is the same as the result of GSEA analysis of virtual-KO perturbed genes (**Figure 5D, left**). In summary, we obtained overall similar results using scTenifoldKnk and DE analysis. The difference is that scTenifoldKnk is a virtual KO analysis, whereas DE analysis requires data from the real KO experiment.

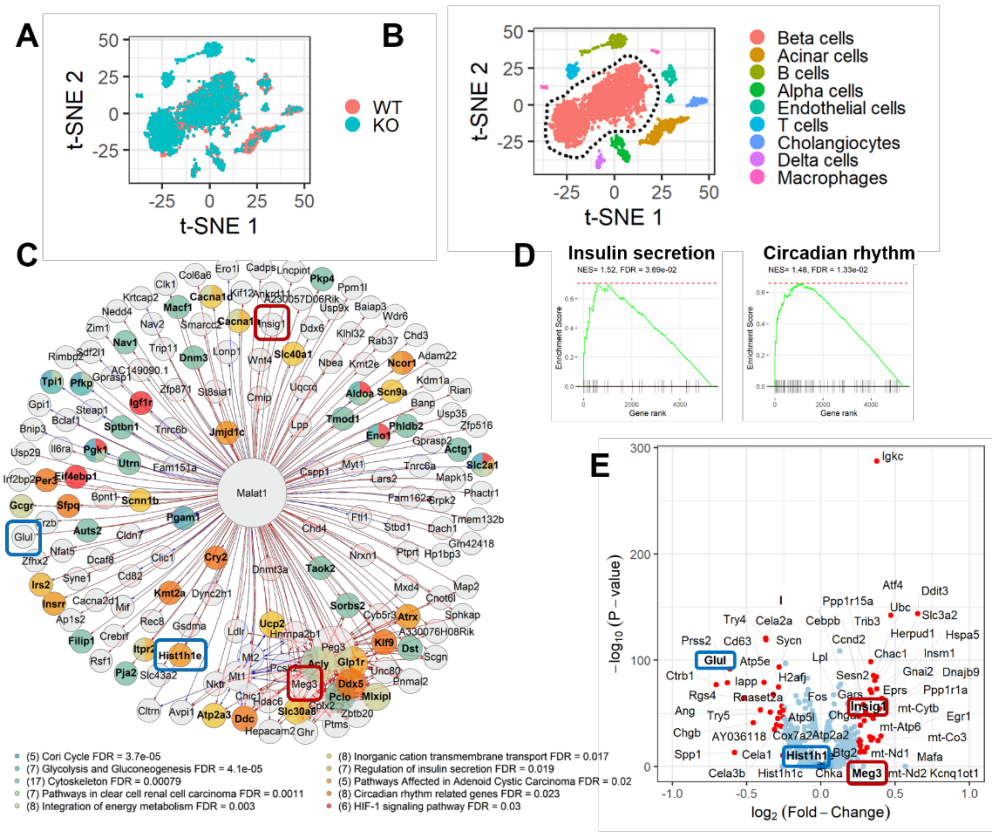

Figure 6. scTenifoldKnk predicts transcriptional changes in beta cells of Malat1$^{-/-}$ KO mice
(**A**) UMAP plot of cells distinguished by experimental groups: WT or KO. (**B**) Cell clustering and cell type assignment. Beta cells included in the analysis are highlighted with a dashed outline border. (**C**) Egocentric plot showing virtual-KO perturbed genes connected to *Malat1*. Nodes are color-coded by each gene's membership association with the top 10 significantly enriched functional groups. (**D**) GSEA analysis with scTenifoldKnk gene list identifies significant gene sets, including *insulin secretion* and *circadian rhythm*. (**E**) Volcano plot from the DE analysis comparing gene expression between WT and KO samples. In the plot, -log2(FC) and -log10(P-value) of genes are plotted against each other. Names of genes are given for those with a fold change (FC) > 10% and FDR < 0.05. Among them, four virtual-KO perturbed genes are highlighted with a red (*Insig1* and *Meg3*) or a blue (*Glu1* and *Hist1h1*) box, indicating up-regulated or down-regulated in the KO sample.

scTenifoldKnk predicts transcriptional changes in pancreatic beta cells of Malat1$^{-/-}$ KO mice

Malat1 is a long, non-coding RNA gene that acts as a transcriptional regulator for numerous genes, which are involved in the generation of reactive oxygen species, cell cycle regulation, cancer metastasis, and cell migration [39-42]. Malat1 upregulation is also associated with hyperglycemia [38]. We have previously shown that the ablation of *Malat1* changes insulin responses in vivo by suppressing *Jnk* activity and inducing the activation of the insulin receptor substance 1 (*IRS-1*) [42]. The ablation of Malat1 also induces the phosphorylation of *Akt* in a process mediated by increased reactive oxygen species (ROS) [42].

To explore the regulatory role of *Malat1* in pancreatic cell functions, we generated *Malat1*$^{-/-}$ KO mice (see **Methods**). We performed scRNA-seq experiments with isolated pancreatic cells from WT (C57BL/6) and *Malat1*$^{-/-}$ KO mice (**Figure 6A**). The beta cells were the most abundant cells in both samples, 1,142 WT and 1,695 KO (**Figure 6B**).

We used scTenifoldKnk to knock out *Malat1* over the WT scGRN and identified 167 virtual-KO perturbed genes (FDR < 0.05, **Supplementary Table S10**). Enrichr functional enrichment analysis indicated that these genes were significantly associated with glucose metabolisms functions such as *Cori cycle*, *glycolysis and gluconeogenesis*, and the *regulation of insulin secretion*, as well as with pathways related to *cellular carcinomas* (**Figure 6C**). This result is consistent with the known function of *Malat1*. For example, *Malat1*, which stands for Metastasis Associated Lung Adenocarcinoma Transcript 1, is known to be overexpressed in lung carcinoma. In addition, *Malat1* is known to have pleiotropic functions in various physiological and pathophysiological processes [42, 66-72], including *cytoskeleton*, *circadian rhythm*, and *hypoxia-inducible factor (HIF) signaling pathway*, as predicted (**Figure 6C**). A STRING database search showed that virtual-KO perturbed genes form a highly connected interaction network (P < 0.01, STRING interaction enrichment test, **Supplementary Figure S6**). GSEA analysis with scTenifoldKnk-reported gene list indicated that 79 non-redundant gene sets were enriched significantly (FDR < 0.05, **Supplementary Table S11**), including *insulin secretion* and *circadian rhythm* (**Figure 6D**).

To validate these predictions made using scTenifoleKnk, we performed DE analysis using MAST to compare gene expression in beta cells between WT and KO mice. We identified 1,695 DE genes (FDR < 0.05, **Figure 6E**). Two up-regulated genes (*Insig1* and *Meg3*) and two down-regulated genes (*Glu1* and *Hist1h1*) were among the 167 virtual-KO perturbed genes (**Figure 6C**). GSEA analysis with the ranked list of DE genes sorted by the fold-change produced only 17 non-redundant enriched gene sets, including *insulin secretion*.

Result 6: Systematic KO of all genes in a given sample to generate the landscape of genes' perturbation profiles

As mentioned, scTenifoldKnk is designed and implemented to be computationally efficient. Indeed, we benchmarked the performance of scTenifoldKnk. When the WT scGRN is given, scTenifoldKnk could complete virtual KO analysis at a rate of one minute per gene on a standard workstation. That is to say, for a given data set of 6,000 genes, scTendifoldKnk could knock out all of these genes individually in 4 to 5 days. The process can also be split and run in parallel to increase the speed of computing. The outcome of such a systematic KO experiment is a collection of perturbation profiles of all genes. For each gene (e.g., gene δ), the perturbation profile of the gene (i.e., gene δ) is a vector of distances of all other genes ($\{\delta^{-}\}$) produced by scTenifoldKnk (see **Figure 1C**). The distance value quantifies the level of a gene (i.e., a gene in $\{\delta^{-}\}$) being perturbed by the deletion of the KO gene (i.e., gene δ). **Figure 7A** illustrates an analytic flowchart when scTenifoldKnk is used in a systematic KO experiment. For a given WT scGRN with *n* genes, scTenifoldKnk can be used to delete individual genes from 1 to n. For the *i*-th gene deletion, G_i^{KO}, scTenifoldKnk produces a KO perturbation profile for the *i*-th gene. The KO perturbation profile is a vector of distances: $\left[d_{i,1}, d_{i,2}, \ldots d_{i,n}\right]^T$, where *i* = 1, 2, …, n. Combining all genes' KO perturbation profiles into an n × n matrix, called *KO perturbation profile matrix*, is followed by t-SNE embedding and clustering of genes.

To demonstrate the use of the systematic KO functionality, we downloaded scRNA-seq data from the brain immune atlas and obtained the expression matrix of 6,853 genes and 5,271 microglial cells (see **Methods**). These microglial cells were derived from WT homeostatic mice [73]. After knocking out all genes, we obtained the KO perturbation landscape of all 6,853 genes, as shown in the t-SNE embedding (**Figure 7B**). Each point represents a perturbation profile caused by KO of a gene. Genes with similar perturbation profiles will be closely located in the low dimensional embedding. Therefore, examining genes closely located will allow us to discover potential functional associations between them [74]. We selected three clusters of different sizes to examine the member of genes in each cluster and the functional relationships between these genes. We found that all three clusters contain genes that show a significantly higher level of functional associations than expected by chance ($P < 0.01$, STRING interaction enrichment tests). We then focused on the smallest cluster, which contains 16 genes (**Figure 7C**). According to the STING database, 12 of these genes (*Apoc1, Apoe, Clec12a, Clec4n, Cp, Fth1, Lilrb4, Mrc1, Ms4a6c, Ms4a7, Pilra* and *Pla2g7*) are functionally associated with each other. These associations are supported by evidence from published literature [75-77]. All these supporting references are related to the microglia study. The remaining 4 genes (*Htra3*, *Tgfbi*, *Pf4* and *Ifitm2*) are not connected to this 12-gene sub-network. Nevertheless, a new study [78] showed that *Htra3* is overexpressed in repopulating microglia. Therefore, these "isolated" genes may be worthy of further scrutiny for their functions in microglia with the potential links to the other genes in the 12-gene sub-network.

Instead of using genes' KO perturbation profiles, using genes' expression information can also produce a landscape of genes. Indeed, we applied t-SNE to the UMI count matrix, which was transformed and standardized using the same procedure (see **Methods**), and obtained the embedding plot of genes (**Supplementary Figure S7B**). The difference between this embedding plot derived from gene expression and the one derived from the gene's KO perturbation profile is obvious. The former has no structure among cells but is just a cloud of data points. Performing clustering using the embedding derived from gene expression ought to not produce clusters

containing more true interactions than using the embedding derived from gene expression profile across cells. To test whether similar results were obtained using expression and perturbation profiles, we calculated the average distance between genes that belong to the same KEGG gene set. We found that, across all KEGG gene sets, the average distance in the embedding derived from a KO perturbation profile was significantly smaller than that in the embedding derived from the expression profile (**Supplementary Figure S7C**). Genes in the three example clusters, as shown in **Figure 7B**, were found to be scattered in different clusters in the embedding derived from gene expression (**Supplementary Figure S7B**). We also conducted the same test using the embedding derived from the genes' profile of the weight of edges in the WT scGRN. The same pattern was uncovered—that is, the average distance is smaller in the embedding derived from the KO perturbation profile than in that derived from the scGRN edge weight. These results suggested that gene sets identified using the gene's KO perturbation profile were more likely to be functionally connected than gene sets identified using other types of gene profiles.

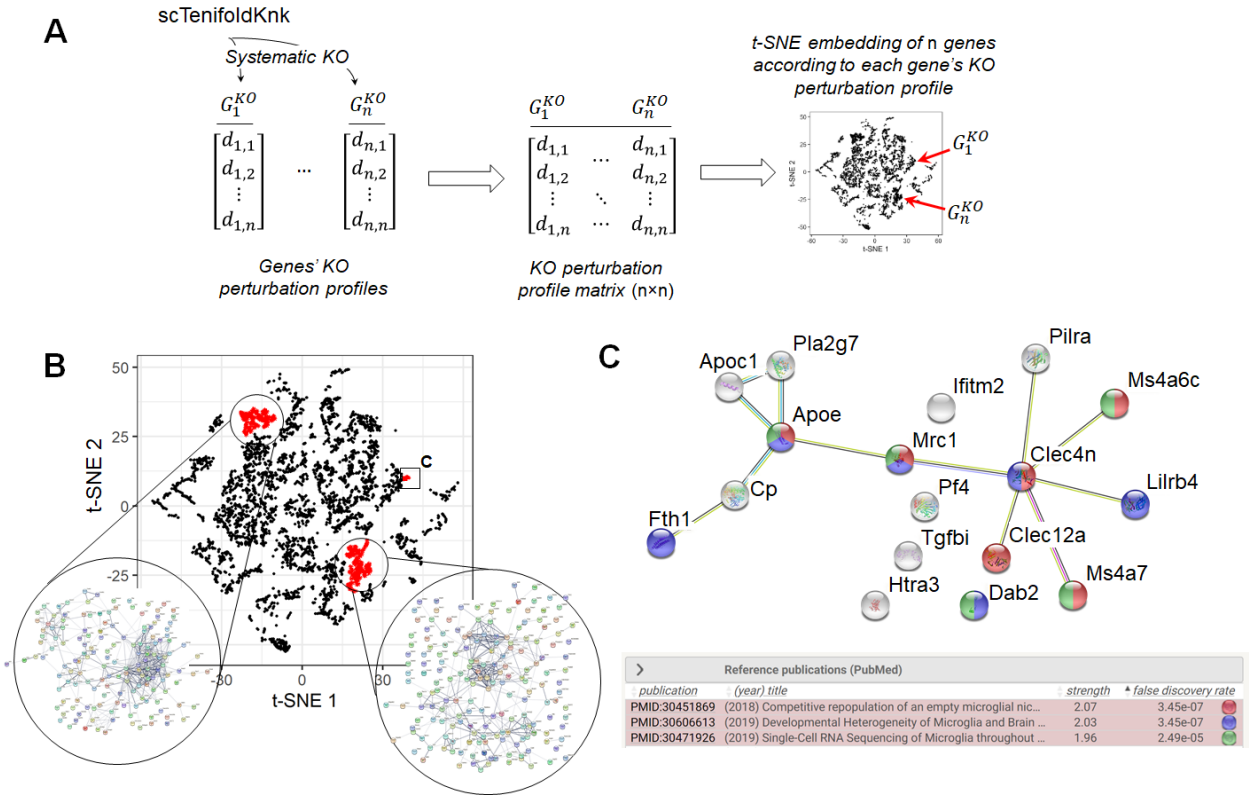

Figure 7. scTenifoldKnk enables systematic KO experiment in microglia and establishment of KO perturbation profile landscape

(**A**) An illustration of systematic virtual-KO analysis using scTenifoldKnk. Two KO genes are shown as an example. Due to the difference in their profiles, two genes are embedded in two different locations, indicated by the red arrows, in the dimensionality reduction visualization. (**B**) t-SNE embedding of 6,853 genes expressed in microglia based on these genes' perturbation profiles. Genes in three clusters in the embedding are highlighted in red. The connections between genes are retrieved from the STING database. Clusters are zoomed to show the STING sub-networks. The rectangle indicates a sub-network shown in **C**. (**C**) STING sub-network of 12 genes (*Apoc1, Apoe, Clec12a, Clec4n, Cp, Fth1, Lilrb4, Mrc1, Ms4a6c, Ms4a7, Pilra* and *Pla2g7*) in the cluster highlighted with the rectangle in **B**. The references [75-77], from which the associations between genes are established, are given in the table.

Result 7: Systematic KO of a gene from different cell types

ScTenifoldKnk can be used to systematically knock out a gene in different cell types. In this way, it is possible to identify cell-type-specific functions of the gene, as well as functions shared across multiple cell types. Here we use *MYDGF* (Myeloid Derived Growth Factor) as an example KO gene to illustrate this application of scTenifoldKnk. *MYDGF*, also known as *C19orf10*, is a 142-residue protein broadly expressed in multiple tissues and cell types [79, 80]. Mydgf has been shown in a mouse model to enhance cardiac myocyte survival, tissue repair, and angiogenesis caused by myocardial infarction [81]. To elucidate *MYDGF*'s functional roles, we downloaded multiple human scRNA-seq data sets from the PanglaoDB database [82]. From the downloaded data sets, we extracted cells from 45 different cell types. Subsequently, a virtual KO of *MYDGF* for each of these cell types and recovered 45 cell-type-specific perturbation profiles was performed. We first examined the perturbation profile of endothelial cells, in which the function of *MYDFG* has been studied [83]. GSEA analysis with the ranked list of genes showed that the enriched functions of *MYDGF* include *Cell cycle*, *VEGFA-VEGFR2 pathway*, and *Intra-Golgi traffic and activation* (**Figure 8A**). These results are consistent with previous findings [83], suggesting that scTenifoldKnk can recapitulate results from a cell-type-specific gene function study. We also concatenated perturbation profiles of 45 different cell types. There were 1,288 genes expressed across all tested cell types. As expected, several cell types exhibited perturbation profiles similar to that of endothelial cells (**Figure 8B**). These cell types include T-cell, hepatocyte, cardiomyocyte, and kidney cell. GSEA analysis identified enriched functions, including *AKT signaling*, *endothelial NOS activation*, *apoptosis*, *muscle contraction*, and *ALK2* pathway (**Figure 8C**) [84]. These findings are consistent with the fact that overexpression of *Mydgf* increases *AKT* phosphorylation and cell proliferation via *AKT/MAPK* signaling pathways [85]. Identified functions were also related to *promoting survival and growth,* as previously reported [79-81, 83, 85]. In summary, we used *Mydgf* as an example gene to demonstrate that scTenifoldKnk can be used to knock out a gene across multiple cell types in order to identify shared as well as cell type-specific functions of the KO gene.

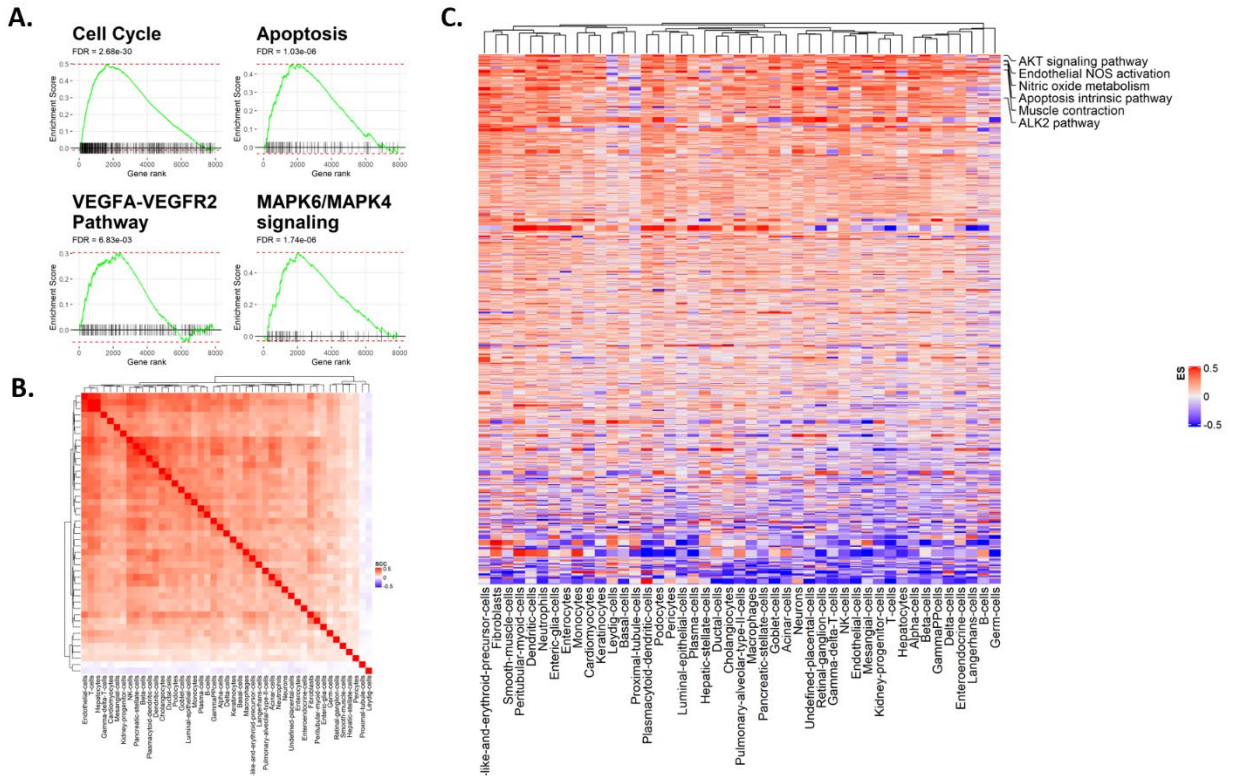

Figure 8. scTenifoldKnk enables the identification of universal and cell type-specific functions of MYDGF across different human cell types.

(**A**) Results of GSEA analysis with ranked gene lists generated by virtual KO of *MYDGF* in endothelial cells. Significant gene sets include *cell cycle*, *apoptosis*, *VEGFA-VEGFR2 pathway*, and *MAPK6/4 signaling pathway*. (**B**) Correlation matrix showing similarity between *MYDGF* perturbation profiles of 45 cell types. (**C**) Heatmap showing enrichment scores for gene sets. Each row is a gene set in the BioPlanet database. Rows are sorted in reverse order of average enrichment score across cell types. Names of several gene sets with functions known to be associated with *MYGDF* are shown.

Discussion

Gene expression is almost always under coordinated regulation in cells of living organisms. Inferring GRNs is the key to a better understanding of such coordinated regulation. However, inferring GRNs is a very difficult process—there are always a large number of unknown variables in the system, and the power of inference is limited by the sample size. The development of single-cell technology has brought new “oil” to network science. We have previously shown that scRNA-seq information can be leveraged to fuel the machine learning algorithms for reliable scGRN construction [9].

In a GRN, the regulatory effect manifests as observable synchronized patterns of expression between genes. These genes are associated with the same biological process, pathway, or under the control of the same set of transcription factors [1]. When a gene involved in a process is perturbed (e.g., knocked out), the expected first responders for such perturbation are those functionally closely related to the KO gene. Thus, modeling influence patterns in a GRN, such as using topological models to approximate perturbation patterns [6], can be used to bypass expensive experimental measurement. Thus, in principle, GRN-based perturbation analysis can contribute to the planning and design of real-animal experimental work for testing or validating

hypotheses. There is evidence showing that gene expression data need not necessarily be collected from perturbation experiments for GRN-based analysis to be successful [86, 87].

Our contribution is to provide an scGRN-based perturbation analytical system—scTenifoldKnk. By performing virtual KO analyses with a series of existing scRNA-seq data, we showed that scTenifoldKnk can be used to identify KO responsive genes. The reported genes were found to be enriched with molecular functions that are similar to those enriched in genes identified by the authentic KO experiments. We also tested scTenifoldKnk using data from three different cell types known to be affected in three Mendelian diseases, respectively. While these diseases represent conditions involving distinct causal genes and different dysregulated molecular processes, scTenifoldKnk demonstrated its value in all three cases. The top hits reported by scTenifoldKnk are highly interpretable, suggesting scTenifoldKnk can be used to reveal molecular mechanisms underlying less characterized pathophysiological processes. The predictive power of scTenifoldKnk was further demonstrated with scRNA-seq data from our own KO experiments, involving enterocytes in Ahr$^{-/-}$ mice and islet cells in Malat1$^{-/-}$ mice. Finally, we showed two case studies of systematic KO experiments.

Despite some obvious limitations associated with the virtual KO method, we start by discussing its advantages. First, the virtual KO method, as we implemented in scTenifoldKnk, is species agnostic—it works with scRNA-seq data from humans and animals alike. This feature gives the method a huge advantage for KO experiments focusing on human samples. In the lack of human KO samples, the KO animals are used as surrogates. The evolutionary divergence between humans and model animals is assumed to play a neglectable role in shaping orthologue gene function—but we know this is not always the case. While applying scTenifoldKnk to human scRNA-seq data, researchers can avoid many pitfalls caused by extending the conclusions from animal KO experiments to humans. Second, scTenifoldKnk allows for any gene to be knocked out for functional analysis as long as the gene expression is detectable in the WT sample. One may want to knock out all genes or a set of genes one by one to obtain an effect profile for each of the KO genes. The effect profile can be as simple as a virtual-KO disrupted gene list produced by scTenifoldKnk. Genes have similar effect profiles that are most likely to share molecular functions or are involved in the same signaling pathways. Genes of known functions can be used as positive controls to gauge the performance of scTenifoldKnk in the tested system and cell type. Third, using scTenifoldKnk, it is feasible to virtually delete more than one gene from a system at the same time to study synergistic KO effects. An example is the double-KO settings for *Hnf4ag*^{DKO}. This is significant because the number of possible combinations of multiple KO targets is often too large to allow researchers to explore the impact of even a small fraction of the multi-gene KO possibilities. In this circumstance, scTenifoldKnk will be an invaluable tool to predict the consequences of these KO combinations and provide experimental design guidance by suggesting the gene combinations that should be prioritized. Fourth, scTenifoldKnk can be used to study the effects of gene KO across multiple cell types. A typical scRNA-seq experiment generates expression data for multiple cell types. In this case, the scTenifoldKnk analysis will allow researchers to study diverse phenotypes associated with the KO gene in all cell types as long as the gene is expressed. Finally, scTenifoldKnk can be used to study the function of essential genes, for which the gene KO causes lethal outcomes, making it impossible to establish the KO animals. When scTenifoldKnk is applied to these essential genes, especially with embryonic expression data of the genes from the WT samples, developmental functions of these genes can be studied.

The resultant gene list of scTenifoldKnk analysis may differ substantially from that of the differential gene expression analysis, with few overlapping genes, as seen in the analyses of *Ahr*$^{-/-}$ and *Malat1*$^{-/-}$ KO mice. We attribute this difference to that scTenifoldKnk detects direct and early responders to the perturbation caused by the gene KO, whereas differentially expressed genes tend to be late responders and reflect a final steady state of the response. Given the hierarchical structure in GRNs [87, 88], it is not unexpected that early and late responders are not the same set of genes. Furthermore, the simple and effective design of scTenifoldKnk does not mean that the analysis can be achieved by simply "picking up" genes that have a large weight in the adjacency matrix of the original scGRN. We examined the correlation between the scTenifoldKnk-reported distance and the weight of genes. We found that, in all real data analyses, scTenifoldKnk picked genes with a wide range of weight (see **Supplementary Figure S8** for two examples: *Trem2* KO and *Dmd* KO). These results showed that scTenifoldKnk's manifold alignment algorithm does not depend upon the weight of the edge between a gene and the KO gene to assert responder genes.

scTenifoldKnk can be used in extended research areas other than KO experiments. For example, biologists often need to know whether a manipulation has an effect or not. scTenifoldKnk can be applied not only to detect novel targets but also to prioritize known targets prior to in vivo or in vitro studies with different experimental techniques. One may use drugs to block the transcription of a gene in a candidate pathway. If the drug has an effect, one would conclude that the drug works on that pathway involved; otherwise, the pathway is not affected. The scTenifoldKnk-based analysis represents a novel analytical method that can be applied to genome-wide association studies (GWASs). Throughout the last decade, GWASs have been successful in detecting associations between variants and phenotypes; however, a phenotypic trait is usually associated with many variants, presumably influencing gene expression regulation. scTenifoldKnk may be used to help geneticists to assess functional consequences in order to prioritize actionable gene targets. We point out that the prediction of scTenifoldKnk does not have to be perfect. As we can see, with many example data sets, scTenifoldKnk could recapitulate major findings reported in real-KO experiments. We showed the landscape of KO perturbation profiles of all genes in microglia. Experimentally, such systematic perturbation analysis has only been done in yeast [89]. With scTenifoldKnk, systematic KO analysis *in silico* can be performed in a cell-type-specific manner in organismal systems.

Limitations of scTenifoldKnk are inherited from being a virtual KO method. scTenifoldKnk cannot be used to predict the consequence of gene overexpression, which is also a commonly used method for gene function study. Also, as the power of scTenifoldKnk is rooted from the WT scGRN, the regulatory network from the WT sample, the prediction that scTenifoldKnk may "favor" regulatory rather than structural genes. Nevertheless, it is still possible to make adjustments to some details in the implementation of scTenifoldKnk to make it better fit user analytical needs. For example, instead of knocking out a target gene by setting its values to zeros in the adjacency matrix, a random shuffle of a gene's expression values in the expression matrix may mimic gene dysregulation.

Prediction of gene expression responses to perturbation using scRNA-seq data is an active research area. To the best of our knowledge, there are two software tools that have been developed for this purpose: scGen [3] and CellOracle [4]. scGen is a package implemented in Python, using TensorFlow variational autoencoders combined with vector arithmetic to predict gene expression changes in cells. Overall, scGen works like a neural network-empowered

regression tool that predicts the changes of gene expression in cells in response to specific perturbations such as disease and drug treatment. scGen requires training data sets from samples before and after being exposed to the same perturbation. CellOracle is a workflow, developed in Python with several R dependencies, that integrates scRNA-seq and single-cell chromatin accessibility data (scATAC-seq) data to infer GRN and predict the changes of gene expression in response to specific perturbations. CellOracle constructs a GRN that accounts for the relationship between transcription factors and their target genes based on sequence motif analysis using the information provided by the scATAC-seq data. After that, the constructed GRN is further refined using regularized Bayesian regression models to remove weak connections and is adjusted to infer the context-dependent GRN using the scRNA-seq data. scTenifoldKnk has a different design compared to either scGen or CellOracle. scTenifoldKnk is specifically designed for performing virtual KO experiments. It requires only the scRNA-seq data from the WT sample: unlike scGen, scTenifoldKnk does not need training data; unlike CellOracle, scTenifoldKnk does not require extra scATAC-seq data to obtain information of transcription factors and their targets. The minimalist design of scTenifoldKnk should allow for adoption widely in most scenarios.

Overall, we have shown evidence that our machine learning workflow represents a powerful and efficient method for conducting virtual KO experiments. The highly efficient implementation of scTenifoldKnk allows systematic deletion of a large number of genes in scRNA-seq data. The prediction power offered by scTenifoldKnk enables accurate prediction of perturbations in regulatory networks caused by the deletion of a gene, revealing its function in a cell type-specific manner. We anticipate that scTenifoldKnk will be adopted and widely applied in the predictive science of single-cell biomedical research.

Methods

scTenifoldKnk is a machine learning workflow for virtual KO experiments with scRNA-seq data. It utilizes a scRNA-seq expression matrix from a WT sample as input, without using any data from a KO sample, to predict regulatory network changes and perturbed genes caused by the KO of a gene.

Construction of the WT scGRN

scTenifoldKnk uses a method we proposed previously in scTenifoldNet [90] to construct scGRNs. The procedure of scGRN construction consists of three steps: cell subsampling, principal component (PC) regression, and tensor decomposition. In cell subsampling, cells are randomly sampled to form subsets. This strategy assures that the outliers in the sample are not included by many sub-sample sets and reduce the influences of outliers. When constructing the scGRN for each sub-sample set, both the accuracy and computational efficiency of the algorithms should be considered. Among existing methods, PCNet based on PC regression is an efficient approach to construct scGRN. It can also surpass the multicollinearity problem when explanatory variables are linearly correlated and provides stable constructing results. Finally, to aggregate multiple GRNs, scTenifoldNet uses the CP tensor decomposition algorithm. CP decomposition factorizes the multilayer scGRN and regenerates all adjacency matrices using top components, which achieves denoising and enhancing, i.e., making main signals stronger and other signals weaker. Overall, scTenifoldNet provides relatively stable and accurate results in constructing GRN. We briefly describe each step below.

(1) *Cell subsampling*: Initially, scTenifoldNet builds several subsets of cells via random sampling. Denote $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ as the scRNA-seq data matrix that reflects gene expression levels for p genes

in n cells. A sub-sample set of cells is constructed via randomly sampling *m* (< *n*) cells in $\boldsymbol{X}$. By repeating this subsampling process for *t* times, *t* sub-sample sets of cells are derived, denoted as $\boldsymbol{X}'_1, \dots, \boldsymbol{X}'_t \in \mathbb{R}^{m \times p}$.

(2) *Network construction*: For each $\boldsymbol{X}'_i$, scTenifoldNet builds a GRN with an adjacency matrix $\boldsymbol{W}^i$ via PC regression, where a PC analysis (PCA) is applied to the original explanatory variables, and then the response variable is regressed on a few leading PCs. Since PC regression only utilizes *d* PCs (*d* << *n*) as the covariates in regression, it mitigates over-fitting and reduces the computation time. To build an scGRN, each time scTenifoldNet focuses on one gene (referred to as the response gene) and applies PC regression. The expression level of the response gene is used as the response variable, and the expression levels of other genes are used as the explanatory variables in PC regression. scTenifoldNet repeats this process for another *N* - 1 times, with one different gene as the response gene each time. In the end, scTenifoldNet collects the coefficients of N regression models together and forms a *p* × *p* adjacency matrix $\boldsymbol{W}^i$, whose (*i*, *j*) entry saves the coefficient of the *i*-th gene on the *j*-th gene. $\boldsymbol{W}^i$ could reflect the interaction strengths between each pair of genes.

(3) *Network denoising*: The adjacency matrices of the t networks $\boldsymbol{W}^1, \dots, \boldsymbol{W}^t$ can be stacked to form a third-order tensor $\Xi \in \mathbb{R}^{(p \times p \times t)}$. To remove noise and construct an overall adjacency matrix, scTenifoldNet applies CANDECOMP/PARAFAC (CP) tensor decomposition to Ξ to extract important latent factors. More specifically, scTenifoldNet approximates Ξ by Ξ_R:

$$\Xi \approx \Xi_R = \sum_{r=1}^{R} \lambda_r a_r \circ b_r \circ c_r,$$

where $\circ$ denotes the outer product, $a_r \in \mathbb{R}^p$, $b_r \in \mathbb{R}^p$, and $c_r \in \mathbb{R}^t$ are unit-norm vectors, and λ_r is a scalar. The reconstructed tensor $\Xi_R \in \mathbb{R}^{(p \times p \times t)}$ includes t denoised adjacency matrices, and by taking the average of them, scTenifoldNet obtains the overall stable adjacency matrix. After further normalizing its entries by dividing them by their maximum absolute value, scTenifoldNet generates the final adjacency matrix of scGRN for the given sample. For later use, denote it as $\boldsymbol{W}_d$.

(4) *Adjusting edge weights of the directed network:* We provided an option to adjust edge weights for the constructed scGRN. PC regression constructs directed networks. To obtain the strictly directed network, with a given, denoised network $\boldsymbol{W}_d$, for each gene pair (i, j) and (j, i), only the entry with a larger absolute weight is kept. More specifically, we defined the (i, j) entry for the strictly directed network $\boldsymbol{W}_s$ by $W_s(\mathrm{i},\mathrm{j}) = W_d(\mathrm{i},\mathrm{j})$, if $|W_d(\mathrm{i},\mathrm{j})| > |W_d(\mathrm{j},\mathrm{i})|$ and $W_s(\mathrm{i},\mathrm{j}) = 0$ otherwise. Note that if $|W_d(\mathrm{i},\mathrm{j})| < |W_d(\mathrm{j},\mathrm{i})|$, then we set $W_s(\mathrm{i},\mathrm{j}) = 0$, then the information in $W_d(\mathrm{i},\mathrm{j})$ is removed. In order to keep the information of $W_d(\mathrm{i},\mathrm{j})$, instead of removing the information completely, we defined a new parameter λ. Using this new parameter, we can integrate $\boldsymbol{W}_d$ and $\boldsymbol{W}_s$ to generate an "interpolated network" $\boldsymbol{W}_\mathrm{i}$, which contains part of the information of $W_\mathrm{d}(\mathrm{i},\mathrm{j})$. Given parameter λ, $\boldsymbol{W}_\mathrm{i}$ is define as:

$$\boldsymbol{W}_\mathrm{i} = \lambda \boldsymbol{W}_\mathrm{s} + (1-\lambda)\boldsymbol{W}_\mathrm{d},$$

where $\lambda \in [0,1]$. It is easy to check that when $\lambda = 0$, we get the original denoised network $\boldsymbol{W}_d$ and when $\lambda = 1$, it is to go back to the strictly directed network $\boldsymbol{W}_s$.

Deletion of the KO gene from the WT scGRN

We propose the virtual KO method that directly works on the WT scGRN (**Figure 1B**). The adjacency matrix $\boldsymbol{W}_d$ represents the scGRN constructed using the WT data. In the virtual KO method, the entire row of the adjacency matrix $\boldsymbol{W}_d$ for the gene is set to zero. We denote the adjacency matrix of the scGRN generated as $\widetilde{\boldsymbol{W}}_d$.

Comparison between the WT and pseudo-KO scGRNs

After obtaining $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$, two comparable low-dimensional feature vectors of each gene in the two networks are built and then compared to detect affected genes. Our approach of creating low-dimensional feature vectors is inspired by manifold alignment [7, 8] and its application [91]; our approach is referred to as quasi-manifold alignment because the adjacency matrices used here are not symmetric matrices while they are required to be symmetric in the original procedure. Here $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$ serve as the inputs for manifold alignment and the outputs are the low-dimensional features $F \in \mathbb{R}^{p\times k}$ and $\tilde{F} \in \mathbb{R}^{p\times k}$ of genes before and after knocking out the target gene, where $k << n$. Before giving the details of the alignment procedure, we point out that $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$ may include negative values, which reflect the negative correlation between genes. Before doing alignment, we add 1 to all entries in $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$, and the range of $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$ is transformed from [-1,1] to [0,2].

To perform quasi-manifold alignment, we first construct a joint adjacency matrix $\boldsymbol{W}$ by combining $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$ together, where $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_d & \frac{\lambda}{2}I \\ \frac{\lambda}{2}I & \widetilde{\boldsymbol{W}}_d \end{bmatrix}$. We can treat $\boldsymbol{W}$ as the adjacency matrix of a joint network formed by linking the corresponding genes in two networks. The off-diagonal block of this matrix reflects the corresponding genes between two networks. λ is a tuning parameter. In practice, we select λ as the mean of the row summations of $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$. We further build $\mathbb{F} = [\begin{smallmatrix} F \\ \tilde{F} \end{smallmatrix}] \in \mathbb{R}^{2p\times k}$, and the manifold alignment problem of two networks characterized by the adjacency matrices $\boldsymbol{W}_d$ and $\widetilde{\boldsymbol{W}}_d$ is equivalent to the manifold learning problem that finds the low dimensional features $\mathbb{F}$ for the joint network characterized by the adjacency matrix $\boldsymbol{W}$. For the sake of convenience, we denote $\mathbb{F}_i \in \mathbb{R}^k$ as the *i*-th row of $\mathbb{F}$ that reflects the projection corresponding to the *i*-th gene in the large network. The next step is to build a "Laplacian" matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n}(\boldsymbol{W})_{j,i}$. Denote $f_1, f_2, \dots, f_d$ as the eigenvectors corresponding to the d smallest nonzero eigenvalues of $\boldsymbol{L}$. Note that $\boldsymbol{L}$ is not a symmetric matrix. We found that the usual solution of symmetrizing $\boldsymbol{L}$ does not work well with either simulated or real data. We, therefore, use asymmetric matrix $\boldsymbol{L}$ in our quasi-manifold alignment procedure. Since $\boldsymbol{L}$ is not symmetric, there may be imaginary parts in the eigendecomposition. Based on our experiment, taking only the real part of eigenvectors with respect to the eigenvalue that has the smallest real part will give better overall results. The final low dimensional representation is $\mathbb{F} = [Re(f_1), Re(f_2), \dots, Re(f_d)]$, where $Re(f)$ means the real part of f.

Test for significance of virtual-KO perturbed genes

The virtual-KO perturbed genes are identified as genes with significant differences in their regulatory patterns in two scGRNs constructed from the WT and KO data. The method for testing the significance of the difference for each gene is described here. With $\mathbb{F} = \begin{bmatrix} F \\ \tilde{F} \end{bmatrix} = [f_1, f_2, \dots, f_d]$ obtained in manifold alignment, for each gene, we calculate the distance d_j between its two

projected feature vectors from two networks. The rankings of d_j are used to help identify significant genes. To avoid arbitrariness in deciding the number of selected genes, we proposed a χ^2 test. Specifically, since d_j^2 is calculated by taking the summation of squares of the differences of projected representations of two samples, its distribution could be approximately χ^2. Instead of d_j^2, we use the scaled fold-change $df \cdot d_j^2/\overline{d^2}$ as the test statistic for each gene j to adjust the scale of the distribution, where df is the degree of freedom. $df \cdot d_j^2/\overline{d^2}$ approximately follows a χ^2 distribution with the degree of freedom (*df*) if the gene does not perform differently before and after knocking out the target gene. By using the upper tail ($P[X>x]$) of the χ^2 distribution and the Benjamini–Hochberg (B–H) FDR correction [92] for multiple testing correction, we assign a *P*-value for each gene. To determine *df*, since the number of the selected significant genes will increase as *df* increases, we choose *df* = 1 to make a conservative selection of genes with high confidence.

Gene functional annotation and enrichment tests

Identified virtual-KO perturbed genes were evaluated and assessed for their molecular functions and biological processes using Enrichr functional enrichment analysis. The enrichment test results of Enrichr were derived from a large number of reference gene sets. In this study, we reported results derived from the following gene set collections: *KEGG_2019_Human, KEGG_2019_Mouse, GO_Biological_Process_2018, GO_Cellular_Component_2018, GO_Molecular_Function_2018, BioPlanet_2019, WikiPathways_2019_Human, WikiPathways_2019_Mouse*, and *Reactome_2016*. GSEA analysis was conducted using a ranked list of all genes as input and comparing against either KEGG_2019_Mouse or KEGG_2019_Human gene sets. The genes were ranked according to their distance between corresponding projection on the aligned manifold between WT and pseudo-KO scGRNs. To remove redundant hits, identified gene sets were compared pairwise. For each pair of identified gene sets, e.g., A and B, if Jaccard index, $J = |A \cap B|/|A \cup B|$, is greater than 0.8, then this pair of gene sets were merged into one. In this way, non-redundant sets of identified gene sets were created and used for functional inference. GSEA analysis was also conducted against gene sets of marker genes. The marker genes were obtained from PanglaoDB [82]. The protein interaction enrichment tests were performed using the web tool of the STRING database [44].

Systematic KO analysis using scRNA-seq data in microglia

The scRNA-seq data used in the systematic KO analysis was obtained from the brain immune atlas (https://www.brainimmuneatlas.org), which is a scRNA-seq resource for assessing and capturing the diversity of the brain immune compartment, as published in [73]. The data was generated using the 10x genomics chromium platform, including more than 61,000 CD45+ immune cells from whole brains or isolated dura mater, subdural meninges, and choroid plexus of mice. The downloaded data is called the *full aggregate data set* (combining cells of whole brain and choroid plexus cells from WT + Irf8 KO mice), stored in the file named filtered_gene_bc_matrices_mex_irf8_fullAggr.zip. The downloaded matrix was processed, and the sub-matrix contained 5,271 microglia from the WT mice. For all genes, the KO perturbation profile of each gene, i.e., a vector of DR distances, was transformed using Box-Cox transformation and then was standardized using z-score transformation. The processed KO perturbation profiles of all genes were combined into one matrix for t-SNE embedding.

Ahr$^{-/-}$ KO mouse strain and crypt cell isolation

Animals were housed under conventional conditions, adhering to the guidelines approved by the Institutional Animal Care and Use Committee at Texas A&M University. Stem cell-targeted Lgr5-EGFP-IRES-Cre^{ERT2}, AhR$^{f/f}$ and tdTomato$^{f/f}$ mouse strains have all been previously described [37]. The mouse genotypes used in this study include: tamoxifen inducible - Lgr5-EGFP-Cre^{ERT2} X Tomato$^{f/f}$ (WT, control), AhR$^{f/f}$ X Lgr5-EGFP-Cre^{ERT2} X Tomato$^{f/f}$ (KO). Male mice (n=5 per genotype, 8-10 weeks of age) were intraperitoneally injected with 2.5 mg of tamoxifen (Sigma, T5648) dissolved in corn oil (25 mg/ml) once a day, for four consecutive days. Mice were maintained on an AIN-76A, semi-purified diet (Research Diets, D12450B), fed ad libitum and housed on a 12 h light-dark cycle. For all experiments, littermate controls were cohoused with the KO mice. Two weeks post tamoxifen injection, colons were removed, washed with cold PBS without calcium and magnesium (PBS-/-), everted on a disposable mouse gavage needle (Instech Laboratories) and incubated in 15 mM EDTA in PBS-/- at 37°C for 35 min as previously described [37]. Subsequently, following transfer to chilled PBS-/-, crypts were mechanically separated from the connective tissue by vigorous vortexing. Cell suspensions were then filtered through a 40-µm mesh and Tomato-expressing cells (includes GFP^{+}/Tom^{+} as well as GFP negative/Tom^{+}) were collected using a MoFlo Astrios Cell Sorter (Beckman Coulter) or a Bio-Rad S3e Cell Sorter. Tomato positive cells represent colonic stem cells and their progeny. Dead cells were excluded by staining with propidium iodide or 7-AAD. Samples were subsequently processed using the 10x Genomics scRNA-seq pipeline described below. A total of 62,741 cells from 10 mice were sequenced. These included 34,889 sorted colonocytes from the WT and 27,852 from the KO mice. The average number of genes detected per sample was 20,141. From all sequenced cells, 40,690 (21,263 from WT and 19,427 from KO samples) were removed with the scRNA-seq quality control procedure. These cells had either a high proportion of mitochondrial reads (greater than 10%) or exhibited an extremely large or small library size.

Malat1$^{-/-}$ KO mouse strain and pancreatic islet cell isolation

Malat1 null mice were obtained from Texas A&M Institute for Genomic Medicine (TIGM). Islets were isolated based on published procedures [93-95]. Briefly, mice were killed by cervical dislocation. The pancreas was inflated with 3 mL of cold collagenase solution (0.3 mg/mL) (Collagenase type 4, Worthington Biochemical Corporation) through the common bile duct with a 20 G needle, starting at the gallbladder. The pancreas was then removed from the body and placed in a siliconized vial containing 2 mL of 1 mg/mL collagenase solution and digested at 37°C in a water bath for 12~15 min. After three washes of the digested pancreas with centrifugation (97 × g at 4°C for 1 min) to collect the tissues, islets were purified by density gradient centrifugation (cold polysucrose/sodium diatrizoate solution 1.1119 g/ml, 560 × g at 4 °C for 15 min) and then dispersed in 96 well plates and cultured in RPMI (Gibco) with 10% FBS for further experiments. To determine islet function and confirm success of isolation of the functional islets, the glucose-stimulated insulin secretion level was measured using a highly sensitive sandwich ELISA assay kit following the manufacturer's instructions (EMD Millipore, EZRMI-13K). After confirming the function of isolated islets, the islets were washed in 10 ml of PBS and dissociated using 2ml of Accutase (Sigma-Aldrich, Cat#A6964) and incubated at room temperature for 20 mins with gentle intermittent mixing by pipetting. The dissociated islets were confirmed using microscopy and were stopped with islet culture media (RPMI with 10% FBS). The dissociated islets were then filtered through a 40-mm filter to remove disassociated tissue and obtain a single-cell suspension of pancreatic islet cells. A total of approximately 20,000 cells were resuspended in 30 ml PBS with

0.05% BSA. Samples were subsequently processed using the 10x Genomics scRNA-seq pipeline described below.

10x Genomics scRNA-seq

Single-cell sample preparation was conducted according to Sample Preparation Demonstrated Protocol provided by 10x Genomics as follows: 1 ml of cell suspension from each mouse genotype was pelleted in Eppendorf tubes by centrifugation (400 g, 5 min). The supernatant was discarded, and the cell pellets were then resuspended in 1x PBS with 0.04% BSA, followed by two washing procedures by centrifugation (150 g, 3 min). After the second wash, cells were resuspended in ~500 μL 1x PBS with 0.04% BSA followed by gently pipetting 10–15 times. Cells were counted using an Invitrogen Countess automated cell counter (Thermo Fisher Scientific, Carlsbad, CA), and the viability of cells was assessed by Trypan Blue staining (0.4%). Subsequently, single-cell GEMs (Gel bead in EMulsion) and sequencing libraries were prepared using the 10x Genomics Chromium Controller in conjunction with the single-cell 3′ v3 kit. Cell suspensions were diluted in nuclease-free water to achieve a targeted cell count of 5,000 for each cell line. The cDNA synthesis, barcoding, and library preparation were then carried out according to the manufacturer's instructions. Libraries were sequenced in the North Texas Genome Center facilities on a NovaSeq6000 sequencer (Illumina, San Diego). For the mapping of reads to transcripts and cells, sample demultiplexing, barcode processing, and unique molecular identifier (UMI) counting were performed using the 10x Genomics pipeline CellRanger v.2.1.0 with default parameters. Specifically, for each library, raw reads were demultiplexed using the pipeline command 'cellranger mkfastq' in conjunction with 'bcl2fastq' (v2.17.1.14, Illumina) to produce two fastq files: the read-1 file containing 26-bp reads, consisting of a cell barcode and a unique molecule identifier (UMI), and the read-2 file containing 96-bp reads including cDNA sequences. Reads then were aligned to the mouse reference genome (mm10), filtered, and counted using "cellranger count" to generate the gene-barcode matrix.

Data and code availability

The scRNA-seq data of KO mice reported in this paper has been deposited in the NCBI's Gene Expression Omnibus database under the accession numbers: GSE166391 and GSE167595. scTenifoldKnk has been implemented in R, as well as in Matlab. The source code is available at https://github.com/cailab-tamu/scTenifoldKnk. The R package is available at the CRAN repository at https://cran.r-project.org/web/packages/scTenifoldKnk/.

Author Contributions

J.J.C. conceived and designed the workflow and data analysis and implemented the software. D.O. designed the workflow and implemented the software. Y.Z., G.L. and Q.X. contributed to the software design and data analysis. J.J.C. and J.Z.H. supervised the data analysis. L.A.D. and R.S.C. performed Ahr KO experiment; J.C. and Y.T. performed Malat1 KO experiment; A.H. performed the scRNA-seq experiments. All authors contributed to the writing of the manuscript.

Declaration of Interests

The authors declare no competing interests.

Acknowledgments

This research was funded by Texas A&M University X-Grant (2019), T3-Grant (2019), and DoD GW200026 for J.J.C., the Allen Endowed Chair in Nutrition & Chronic Disease Prevention, NIH

R01-ES025713, R01-CA202697, R35-CA197707 grants for R.S.C., and Texas A&M Institute of Data Science (TAMIDS) Data Resource Development Program Award (2020) for D.O. and Y.Z. A sincere thank you to the editing team members of the English 320 course project, Kamryn Watson, Kanza Akhtar, Marisa Morris, and Mary Frances Nance, for diligent proofreading of this paper.

References

1. Michalak, P., *Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes.* Genomics, 2008. **91**(3): p. 243-8.
2. Li, D. and J. Gao, *Towards perturbation prediction of biological networks using deep learning.* Sci Rep, 2019. **9**(1): p. 11941.
3. Lotfollahi, M., F.A. Wolf, and F.J. Theis, *scGen predicts single-cell perturbation responses.* Nat Methods, 2019. **16**(8): p. 715-721.
4. Kamimoto, K., C.M. Hoffmann, and S.A. Morris, *CellOracle: Dissecting cell identity via network inference and in silico gene perturbation.* bioRxiv, 2020: p. 2020.02.17.947416.
5. Chalancon, G., et al., *Interplay between gene expression noise and regulatory network architecture.* Trends Genet, 2012. **28**(5): p. 221-32.
6. Santolini, M. and A.L. Barabasi, *Predicting perturbation patterns from the topology of biological networks.* Proc Natl Acad Sci U S A, 2018. **115**(27): p. E6375-E6383.
7. Wang, C. and S. Mahadevan, *A General Framework for Manifold Alignment*, in *AAAI Fall Symposium: Manifold Learning and Its Applications*. 2009, Association for the Advancement of Artificial Intelligence. p. FS-09-04.
8. Vu, H.T., C.J. Carey, and S. Mahadevan, *Manifold warping: manifold alignment over time*, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012, AAAI Press: Toronto, Ontario, Canada. p. 1155-1161.
9. Osorio, D., et al., *scTenifoldNet: a machine learning workflow for constructing and comparing transcriptome-wide gene regulatory networks from single-cell data.* Patterns, 2020. **1**(100139).
10. Dibaeinia, P. and S. Sinha, *SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks.* Cell Syst, 2020. **11**(3): p. 252-271 e11.
11. Little, D.R., et al., *Transcriptional control of lung alveolar type 1 cell development and maintenance by NK homeobox 2-1.* Proc Natl Acad Sci U S A, 2019. **116**(41): p. 20545-20555.
12. Attarian, S.J., et al., *Mutations in the thyroid transcription factor gene NKX2-1 result in decreased expression of SFTPB and SFTPC.* Pediatr Res, 2018. **84**(3): p. 419-425.
13. Atagi, Y., et al., *Apolipoprotein E Is a Ligand for Triggering Receptor Expressed on Myeloid Cells 2 (TREM2).* J Biol Chem, 2015. **290**(43): p. 26043-50.
14. Jaitin, D.A., et al., *Lipid-Associated Macrophages Control Metabolic Homeostasis in a Trem2-Dependent Manner.* Cell, 2019. **178**(3): p. 686-698 e14.
15. Poliani, P.L., et al., *TREM2 sustains microglial expansion during aging and response to demyelination.* J Clin Invest, 2015. **125**(5): p. 2161-70.
16. Nugent, A.A., et al., *TREM2 Regulates Microglial Cholesterol Metabolism upon Chronic Phagocytic Challenge.* Neuron, 2020. **105**(5): p. 837-854 e9.
17. Guerreiro, R., et al., *TREM2 variants in Alzheimer's disease.* N Engl J Med, 2013. **368**(2): p. 117-27.
18. Chen, L., et al., *A reinforcing HNF4-SMAD4 feed-forward module stabilizes enterocyte identity.* Nat Genet, 2019. **51**(5): p. 777-785.
19. Yoshimura, K., et al., *Expression of the cystic fibrosis transmembrane conductance regulator gene in cells of non-epithelial origin.* Nucleic Acids Res, 1991. **19**(19): p. 5417-23.
20. Collins, F.S., *Cystic fibrosis: molecular biology and therapeutic implications.* Science, 1992. **256**(5058): p. 774-9.
21. Stoltz, D.A., D.K. Meyerholz, and M.J. Welsh, *Origins of cystic fibrosis lung disease.* N Engl J Med, 2015. **372**(4): p. 351-62.
22. Frank, D.B., et al., *Early lineage specification defines alveolar epithelial ontogeny in the murine lung.* Proc Natl Acad Sci U S A, 2019. **116**(10): p. 4362-4371.

23. Campbell, K.P., *Three muscular dystrophies: loss of cytoskeleton-extracellular matrix linkage.* Cell, 1995. **80**(5): p. 675-9.
24. Hoffman, E.P., R.H. Brown, Jr., and L.M. Kunkel, *Dystrophin: the protein product of the Duchenne muscular dystrophy locus.* Cell, 1987. **51**(6): p. 919-28.
25. Rubenstein, A.B., et al., *Single-cell transcriptional profiles in human skeletal muscle.* Sci Rep, 2020. **10**(1): p. 229.
26. Abuhatzira, L., et al., *MeCP2 deficiency in the brain decreases BDNF levels by REST/CoREST-mediated repression and increases TRKB production.* Epigenetics, 2007. **2**(4): p. 214-22.
27. Nan, X., F.J. Campoy, and A. Bird, *MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin.* Cell, 1997. **88**(4): p. 471-81.
28. Lyst, M.J. and A. Bird, *Rett syndrome: a complex disorder with simple roots.* Nat Rev Genet, 2015. **16**(5): p. 261-75.
29. Johnston, M.V., et al., *Neurobiology of Rett syndrome: a genetic disorder of synapse development.* Brain Dev, 2001. **23 Suppl 1**: p. S206-13.
30. Medrihan, L., et al., *Early defects of GABAergic synapses in the brain stem of a MeCP2 mouse model of Rett syndrome.* J Neurophysiol, 2008. **99**(1): p. 112-21.
31. Chao, H.T., et al., *Dysfunction in GABA signalling mediates autism-like stereotypies and Rett syndrome phenotypes.* Nature, 2010. **468**(7321): p. 263-9.
32. Romaniello, R., et al., *A de-novo STXBP1 gene mutation in a patient showing the Rett syndrome phenotype.* Neuroreport, 2015. **26**(5): p. 254-7.
33. Zeisel, A., et al., *Molecular Architecture of the Mouse Nervous System.* Cell, 2018. **174**(4): p. 999-1014 e22.
34. Metidji, A., et al., *The Environmental Sensor AHR Protects from Inflammatory Damage by Maintaining Intestinal Stem Cell Homeostasis and Barrier Integrity.* Immunity, 2018. **49**(2): p. 353-362 e5.
35. Ko, C.I., et al., *Pluripotency factors and Polycomb Group proteins repress aryl hydrocarbon receptor expression in murine embryonic stem cells.* Stem Cell Res, 2014. **12**(1): p. 296-308.
36. Han, H., et al., *Loss of aryl hydrocarbon receptor potentiates FoxM1 signaling to enhance self-renewal of colonic stem and progenitor cells.* The EMBO Journal. **n/a**(n/a): p. e104319.
37. Han, H., et al., *Loss of aryl hydrocarbon receptor potentiates FoxM1 signaling to enhance self-renewal of colonic stem and progenitor cells.* EMBO J, 2020: p. e104319.
38. Puthanveetil, P., et al., *Long non-coding RNA MALAT1 regulates hyperglycaemia induced inflammatory process in the endothelial cells.* J Cell Mol Med, 2015. **19**(6): p. 1418-25.
39. Samimi, H., et al., *Transcript-level regulation of MALAT1-mediated cell cycle and apoptosis genes using dual MEK/Aurora kinase inhibitor "BI-847325" on anaplastic thyroid carcinoma.* Daru, 2019. **27**(1): p. 1-7.
40. Sun, Y. and L. Ma, *New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis.* Cancers (Basel), 2019. **11**(2).
41. Wang, Q., G. Lu, and Z. Chen, *MALAT1 promoted cell proliferation and migration via MALAT1/miR-155/MEF2A pathway in hypoxia of cardiac stem cells.* J Cell Biochem, 2019. **120**(4): p. 6384-6394.
42. Chen, J., et al., *Long noncoding RNA MALAT1 regulates generation of reactive oxygen species and the insulin responses in male mice.* Biochem Pharmacol, 2018. **152**: p. 94-103.
43. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
44. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-D613.

45. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
46. Shi, Y. and D.M. Holtzman, *Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight.* Nat Rev Immunol, 2018. **18**(12): p. 759-772.
47. Babeu, J.P., et al., *Hepatocyte nuclear factor 4alpha contributes to an intestinal epithelial phenotype in vitro and plays a partial role in mouse intestinal epithelium differentiation.* Am J Physiol Gastrointest Liver Physiol, 2009. **297**(1): p. G124-34.
48. Carter, C.O., *Monogenic disorders.* J Med Genet, 1977. **14**(5): p. 316-20.
49. Andersen, D.H. and R.G. Hodges, *Celiac syndrome; genetics of cystic fibrosis of the pancreas, with a consideration of etiology.* Am J Dis Child, 1946. **72**: p. 62-80.
50. Liu, F., et al., *Molecular Structure of the Human CFTR Ion Channel.* Cell, 2017. **169**(1): p. 85-95 e8.
51. Lin, Z., et al., *Genetic Association of Pulmonary Surfactant Protein Genes, SFTPA1, SFTPA2, SFTPB, SFTPC, and SFTPD With Cystic Fibrosis.* Front Immunol, 2018. **9**: p. 2256.
52. von Bredow, C., A. Wiesener, and M. Griese, *Proteolysis of surfactant protein D by cystic fibrosis relevant proteases.* Lung, 2003. **181**(2): p. 79-88.
53. Buhling, F., et al., *Gene targeting of the cysteine peptidase cathepsin H impairs lung surfactant in mice.* PLoS One, 2011. **6**(10): p. e26247.
54. Dwivedi, S.K.D., et al., *KRCC1: A potential therapeutic target in ovarian cancer.* FASEB J, 2020. **34**(2): p. 2287-2300.
55. Li, Y., et al., *Anti-fibrosis activity of combination therapy with epigallocatechin gallate, taurine and genistein by regulating glycolysis, gluconeogenesis, and ribosomal and lysosomal signaling pathways in HSC-T6 cells.* Exp Ther Med, 2018. **16**(6): p. 4329-4338.
56. Cohn, R.D. and K.P. Campbell, *Molecular basis of muscular dystrophies.* Muscle Nerve, 2000. **23**(10): p. 1456-71.
57. Brinkmeyer-Langford, C., et al., *Expression profiling of disease progression in canine model of Duchenne muscular dystrophy.* PLoS One, 2018. **13**(3): p. e0194485.
58. Neul, J.L., et al., *Rett syndrome: revised diagnostic criteria and nomenclature.* Ann Neurol, 2010. **68**(6): p. 944-50.
59. Amir, R.E., et al., *Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2.* Nat Genet, 1999. **23**(2): p. 185-8.
60. Kawajiri, K. and Y. Fujii-Kuriyama, *The aryl hydrocarbon receptor: a multifunctional chemical sensor for host defense and homeostatic maintenance.* Exp Anim, 2017. **66**(2): p. 75-89.
61. Stockinger, B., et al., *The aryl hydrocarbon receptor: multitasking in the immune system.* Annu Rev Immunol, 2014. **32**: p. 403-32.
62. Ramadoss, P., C. Marcus, and G.H. Perdew, *Role of the aryl hydrocarbon receptor in drug metabolism.* Expert Opin Drug Metab Toxicol, 2005. **1**(1): p. 9-21.
63. Hipfner, D.R. and S.M. Cohen, *Connecting proliferation and apoptosis in development and disease.* Nat Rev Mol Cell Biol, 2004. **5**(10): p. 805-15.
64. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.
65. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome Biol, 2015. **16**: p. 278.
66. Zhang, H., et al., *MALAT1 accelerates the development and progression of renal cell carcinoma by decreasing the expression of miR-203 and promoting the expression of BIRC5.* Cell Prolif, 2019. **52**(5): p. e12640.

67. Malakar, P., et al., *Long Noncoding RNA MALAT1 Regulates Cancer Glucose Metabolism by Enhancing mTOR-Mediated Translation of TCF7L2.* Cancer Res, 2019. **79**(10): p. 2480-2493.
68. Malakar, P., et al., *Long Noncoding RNA MALAT1 Promotes Hepatocellular Carcinoma Development by SRSF1 Upregulation and mTOR Activation.* Cancer Res, 2017. **77**(5): p. 1155-1167.
69. Yan, C., J. Chen, and N. Chen, *Long noncoding RNA MALAT1 promotes hepatic steatosis and insulin resistance by increasing nuclear SREBP-1c protein stability.* Sci Rep, 2016. **6**: p. 22640.
70. Liu, H., et al., *Long non-coding RNA MALAT1 mediates hypoxia-induced pro-survival autophagy of endometrial stromal cells in endometriosis.* J Cell Mol Med, 2019. **23**(1): p. 439-452.
71. Sun, R., et al., *Down-regulation of MALAT1 inhibits cervical cancer cell invasion and metastasis by inhibition of epithelial-mesenchymal transition.* Mol Biosyst, 2016. **12**(3): p. 952-62.
72. Cai, X., et al., *Long noncoding RNA MALAT1 as a potential therapeutic target in osteosarcoma.* J Orthop Res, 2016. **34**(6): p. 932-41.
73. Van Hove, H., et al., *A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment.* Nat Neurosci, 2019. **22**(6): p. 1021-1035.
74. Dorrity, M.W., et al., *Dimensionality reduction by UMAP to visualize physical and genetic interactions.* Nat Commun, 2020. **11**(1): p. 1537.
75. Lund, H., et al., *Competitive repopulation of an empty microglial niche yields functionally distinct subsets of microglia-like cells.* Nat Commun, 2018. **9**(1): p. 4845.
76. Li, Q., et al., *Developmental Heterogeneity of Microglia and Brain Myeloid Cells Revealed by Deep Single-Cell RNA Sequencing.* Neuron, 2019. **101**(2): p. 207-223 e10.
77. Hammond, T.R., et al., *Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes.* Immunity, 2019. **50**(1): p. 253-271 e6.
78. Willis, E.F., et al., *Repopulating Microglia Promote Brain Repair in an IL-6-Dependent Manner.* Cell, 2020. **180**(5): p. 833-846 e16.
79. Bortnov, V., et al., *Myeloid-derived growth factor is a resident endoplasmic reticulum protein.* Journal of Biological Chemistry, 2018. **293**(34): p. 13166-13175.
80. Bortnov, V., et al., *Solution structure of human myeloid-derived growth factor suggests a conserved function in the endoplasmic reticulum.* Nature communications, 2019. **10**(1): p. 1-14.
81. Korf-Klingebiel, M., et al., *Myeloid-derived growth factor (C19orf10) mediates cardiac repair following myocardial infarction.* Nature medicine, 2015. **21**(2): p. 140-149.
82. Franzen, O., L.M. Gan, and J.L.M. Bjorkegren, *PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data.* Database (Oxford), 2019. **2019**.
83. Wang, Y., et al., *Mydgf promotes Cardiomyocyte proliferation and Neonatal Heart regeneration.* Theranostics, 2020. **10**(20): p. 9100.
84. Huang, R., et al., *The NCATS BioPlanet - An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics.* Front Pharmacol, 2019. **10**: p. 445.
85. Sunagozaka, H., et al., *Identification of a secretory protein c19orf10 activated in hepatocellular carcinoma.* International journal of cancer, 2011. **129**(7): p. 1576-1585.
86. Hartemink, A.J., *Reverse engineering gene regulatory networks.* Nat Biotechnol, 2005. **23**(5): p. 554-5.
87. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, 2005. **37**(4): p. 382-90.

88. Erwin, D.H. and E.H. Davidson, *The evolution of hierarchical gene regulatory networks.* Nat Rev Genet, 2009. **10**(2): p. 141-8.
89. Busby, B.P., et al., *Genetic interaction networks mediate individual statin drug response in Saccharomyces cerevisiae.* NPJ Syst Biol Appl, 2019. **5**: p. 35.
90. Osorio, D., et al., *scTenifoldNet: A Machine Learning Workflow for Constructing and Comparing Transcriptome-wide Gene Regulatory Networks from Single-Cell Data.* Patterns (N Y), 2020. **1**(9): p. 100139.
91. Nguyen, N.D., I.K. Blaby, and D. Wang, *ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks.* BMC Genomics, 2019. **20**(Suppl 12): p. 1003.
92. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J Roy Statist Soc Ser B (Methodological), 1995. **57**: p. 289-300.
93. Szot, G.L., P. Koudria, and J.A. Bluestone, *Murine pancreatic islet isolation.* JoVE (Journal of Visualized Experiments), 2007(7): p. e255-e255.
94. Chen, J., et al., *A Murine Pancreatic Islet Cell-based Screening for Diabetogenic Environmental Chemicals.* Journal of visualized experiments: JoVE, 2018(136).
95. Marquina-Sanchez, B., et al., *Single-cell RNA-seq with spike-in cells enables accurate quantification of cell-specific drug effects in pancreatic islets.* Genome Biology, 2020. **21**(1): p. 1-22.