Generative adversarial networks uncover epidermal regulators and predict single cell perturbations

Arsham Ghahramani[1,2,*], Fiona M. Watt[2,†] and Nicholas M. Luscombe[1,3,4,†]

[1]The Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom
[2]King's College London, Centre for Stem Cells and Regenerative Medicine, 28th Floor, Tower Wing, Guy's Hospital, Great Maze Pond, London SE1 9RT, United Kingdom
[3]UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom
[4]Okinawa Institute of Science & Technology Graduate University, Okinawa 904-0495, Japan

[*] Correspondence to: arsham.ghahramani@crick.ac.uk
[†] These authors contributed equally

Abstract

Recent advances have enabled gene expression profiling of single cells at lower cost. As more data is produced there is an increasing need to integrate diverse datasets and better analyse underutilised data to gain biological insights. However, analysis of single cell RNA-seq data is challenging due to biological and technical noise which not only varies between laboratories but also between batches. Here for the first time, we apply a new generative deep learning approach called Generative Adversarial Networks (GAN) to biological data. We show that it is possible to integrate diverse skin (epidermal) datasets and in doing so, our generative model is able to simulate realistic scRNA-seq data that covers the full diversity of cell types. In contrast to many machine-learning approaches, we are able to interpret internal parameters in a biologically meaningful manner. These include using internal GAN learned features for improved dimensionality reduction. Using our generative model we are able to obtain a universal representation of epidermal differentiation and use this to predict the effect of cell state perturbations on gene expression at high time-resolution. We show that our trained neural networks identify biological state-determining genes and through analysis of these networks we can obtain inferred gene regulatory relationships. In combination these attributes provide a powerful framework to progress the analysis of scRNA-seq data beyond exploratory analysis of cell clusters and towards integration of multiple datasets regardless of origin.

1. Introduction

The development of affordable single cell RNA-seq has enabled the measurement of transcript abundance in hundreds to thousands of individual cells[1–3]. These profiles provide an opportunity to define cell state and indirectly measure the signals and factors influencing cell fate. However, despite the richness of single cell measurements, these data are computationally challenging to analyse due to technical and biological noise meaning that traditional bulk expression analysis approaches are frequently not applicable. Current single cell RNA-seq dimensionality reduction methods can successfully reveal clustering and structure within data when technical noise is low; however, they cannot easily integrate diverse datasets produced using distinct protocols[4–6]. Furthermore, current approaches focus on differential expression and marker gene identification but do not yield functional gene regulatory relationships.

Deep learning algorithms enabled by advances in computational power have demonstrated the capability to analyse diverse datasets from images to genomics[7,8]. In particular, generative adversarial networks (GAN), first introduced in 2014, have shown promise in the field of computer vision[9]. Since their introduction, GANs have become an active area of research with multiple variants developed that have resulted in improved performance and training[10–14]. Common to all of these variants[15] is the concurrent training of two neural networks competing against one another, referred to as the generator and discriminator (Figure 1). The generator is tasked with generating simulated data, whereas the discriminator is tasked with evaluating whether data is authentic or not. Notably, only the discriminator directly observes real data while the generator improves its simulations through interaction with the discriminator. As training progresses both neural networks learn key features of the training data. Additionally, both discriminator and generator performance improves as they compete against each other.

In the field of computer vision, GANs have proved capable of generating visually convincing and novel images such as faces or furniture[16,17]. Furthermore, it has been shown that the generator network learns a meaningful latent space where visually similar data such as similar faces are clustered. A meaningful latent space for single cell RNA-seq is particularly appealing as this can be used in conjunction with existing dimensionality reduction methods for improved and more meaningful subpopulation extraction. Additionally, an advantageous feature of generative neural networks is that non-linear relationships between features of the data are learned through training and can be later extracted to provide further insights.

Here, we integrate disparate epidermal datasets produced by three separate laboratories. The mammalian epidermis comprises interfollicular epidermis (IFE), hair follicles, sebaceous glands and further associated adnexal structures: together they form a protective interface between the body and external environment. Hence, epidermal cell fate is determined through the integration of extracellular cues with transcriptional activity[18,19]. Under steady-state conditions, each epidermal compartment is maintained by distinct stem cell populations. However, under certain conditions such as wounding, each stem cell subpopulation is able to contribute to all differentiated lineages[20]. These distinct yet seemingly interchangeable subpopulations hint at a common, but as yet undefined epidermal gene regulatory network. For the first time, we have applied GANs to genomic data to uncover previously unknown gene regulatory relationships and regulators of epidermal cell state. We show that GANs produce biologically meaningful dimensionality reduction and using the trained generator neural network we demonstrate that GANs can be used to predict the effect of cell state perturbations on unseen single cells.

2. Results

2.1 Generative adversarial networks integrate diverse datasets

We used a generative adversarial network (GAN) to integrate multiple diverse mouse epidermal single cell RNA-seq datasets. These datasets originated from three labs, spanning mouse epidermal cells *in vitro*[21], whole epidermis *in vivo*[22] and isolated subpopulations *in vivo*[23] (Figure 1, upper right. See methods for GEO accession numbers). No adjustments were made to account for batch-to-batch variations within datasets and across labs. After removing non-epidermal cell types and outlier cells we retained 1763 cells for training and 500 unseen cells for testing and validation. Similarly to other analysis methods, we filtered genes by expression level, resulting in 6605 potentially informative genes (see methods).

GANs consist of two neural networks competing against each other. The generator is tasked with producing realistic output data from a random input vector z or latent variable. In our case the latent variable z is of lower dimension (100, arbitrarily chosen) than the output scRNA-seq data (6605 genes), hence the generator represents a mapping from a lower dimensional space to gene expression space. Each 100-dimensional latent vector maps to 6605 transcripts per million (TPM, log(TPM+1)) normalised gene expression values for one cell. The second network, named the discriminator, is tasked with discriminating between the real and generated data. The discriminator takes log(TPM+1) gene expression values as input and outputs a score related to its assessment of the gene expression input. These two networks are trained using a combined loss function leading to improved performance of both neural networks (Figure 1).

We tested several variations of GANs ultimately utilising a Wasserstein GAN with added gradient penalty term in the loss function as described previously[24] due to training stability. We trained our neural network for approximately 15,000 epochs per full training run. We evaluated the generator network output performance at multiple checkpoints using t-distributed stochastic neighbour embedding (t-SNE, Figure 2A) and correlation between real samples and generated samples as shown in Supplementary Figure 1. For each evaluation we generated 500 cells using 500 random latent variables. Early in training, the GAN struggles to produce a varied output representative of the breadth of cell types and states; generated cells are closely correlated and form a single cluster in the t-SNE plot (Figure 2A 1000 steps, Supplementary Figure S1A). After 5,000 steps the generator begins to produce a varied output with generated cells mapping to multiple clusters in the t-SNE plot covering the different cell types, cell states, cell origin and experimental batches present in our combined dataset. After 10,000 steps, the generator network is capable of producing a subset of cells with similar transcriptomic profiles to those of Yang and colleagues that form a distinct cluster. Further GAN training broadens the distribution of correlations between generated samples, indicating an increasingly diverse generator output (Supplementary Figure S1A). We observed that after 13,000 steps continued training does not continue to increase the output diversity of the generator network, as defined by the median distance between cell gene expression profiles. Figure 2B shows that GAN output diversity reaches a maximum between 12,500 and 15,000 training steps. Furthermore, additional training between 12,500 and 15,000 steps does not change the generated cell correlation distribution (Supplementary Figure S1A). At this point the discriminator loss function has also converged and therefore we cease further training (Supplementary Figure S2).

We computationally validate our GAN by examining the correlation between generated samples and unseen samples (Supplementary Figure S1B). The distributions obtained at different training steps show that generated samples are correlated with, but distinct from, real samples. At early training steps this distribution is centred around a Pearson correlation of 0.4. After 10,000 steps this distribution broadens, indicating the generator learns to simulate a diverse population of cells. This can also be seen from the expression of individual genes. Figure 2C shows single cell expression bar plots for three genes that distinguish subpopulations in our cell cohort, Ap1s3, Bcl2 and Nucks1. For each gene we show 500 unseen cells and 500 closely related generated cells with an average correlation of 0.71. Our generated cohort follows a similar pattern of expression to the real unseen cells, with Ap1s3 and Bcl2 highly expressed in differentiating cell types, and Nucks1 expression absent in vitro. In real cells where Ap1s3 or Bcl2 are not detected, the corresponding generated cells show low non-zero expression indicative of a form of imputation performed by the generator. Across all genes, expression variability in the simulated cells is similar to the real data. Together these results demonstrate that the generator network is not memorising and reproducing training samples but is instead inferring relationships between gene expression values in order to output convincing heterogeneous generated cells.

Focusing on the data from Joost and colleagues, the 1422 cells comprise eleven epidermal cell types originating from whole dorsal epidermis including previously uncharacterised cells. The generator neural network is able to simulate cells clustering with all *in vivo* epidermal cell types, ranging from sebaceous gland cells with distinct gene expression profiles to upper hair follicle and interfollicular epidermal cells (IFE) with similar transcriptomes but distinct spatial positions (Figure 2D). This range of cell type simulation is achieved despite training the GAN using three diverse datasets.

2.2 Discriminator network + t-SNE allows better dimensionality reduction

We hypothesised that the discriminator network learned biologically relevant features of scRNA-seq data, discarding uninformative genes in order to discriminate successfully between the generated and real samples. The discriminator neural network consists of a single hidden layer whose output is transformed to a discriminator or “critic” output value It transforms 6605 gene expression values into 200 learned internal features, effectively a reduced dimensional representation. Hence, we sought to understand whether the discriminator hidden layer output contained learned features that could be used to improve on existing dimensionality reduction. We performed t-SNE on the discriminator hidden layer output to visualise these learned features (Figure 3A) and compared this with two other approaches.

First we performed a principal component analysis (PCA) on the combined epidermal training data (Figure 3A, middle column); the main source of linearly separable variance in the first two components is the batch effect of dataset origin and laboratory and the approach generally fails to identify biologically meaningful clusters even at higher principal components. We removed this technical source of variation by subtracting the first two principal components and performing a t-SNE on the resulting dataset. The resulting t-SNE (Figure 3A, right column) no longer clusters cells by dataset origin, so removing variation caused by different experimental protocols, at the cost of concurrently removing informative biological variation in gene expression.

Two major sources of variation in epidermal cells are their differentiation status and spatial position within the epidermis: a useful dimensionality reduction approach should capture at least one of these features, ideally both. To investigate this we overlaid expression of Krt14, a marker of undifferentiated keratinocytes, Krt10 a marker of commitment to differentiation and Rps29, a ubiquitously expressed ribosomal protein upregulated in the IFE in comparison to hair follicles.

Biologically, we expect cells with similar expression levels of these marker genes to cluster. PCA and the combined approach of PC removal + t-SNE perform poorly in this regard and expression of all three markers fails to distinguish cells across the dimensionally reduced space (Figure 3 middle and right column). In contrast, our GANt-SNE approach clusters Krt14^{high} cells deriving from the basal layer of the IFE (Figure 3B, Basal IFE) which are separated from the remaining IFE cells. A third cluster of hair follicle cells shows sporadic expression of Krt14 as this is not a functional marker for these cells. Our dimensionality reduction separates differentiated IFE cells highly expressing Krt10 from terminally differentiated and keratinized Krt10^{low} cells (Figure 3C). Furthermore, GANt-SNE correctly separates IFE-derived cells from hair follicle cells as seen by Rps29 expression (Figure 3D). These biological distinctions are not identified by the two alternative methods of dimensionality reduction.

These results lend credence to our hypothesis that the discriminator learns biologically relevant features of the data. Using the GANt-SNE approach we successfully separate cells by differentiation status and spatial position (IFE vs. hair follicle and differentiating vs. undifferentiated). This is achieved without a priori knowledge of technical variation and batch effects by training on the commonalities between training data. In contrast the PCA and t-SNE approaches fail to separate cells into these biologically meaningful clusters, as they crudely extract the strongest sources of variation - often a non-linear mixture of biological and technical variation.

2.3 Simulating cellular perturbations using latent space interpolation

To take advantage of the gene expression rules learned by the generator network we devised an algorithm for retrieving the latent space vector z from an arbitrary gene expression profile x such that $x = G(z)$. For each target cell we randomly sampled latent space vectors (z) and generated cells until a sufficiently similar cell is obtained (see methods). In other fields where GANs have been applied, such as image generation, the latent space representation of an image meaningfully represents visually similar images. Furthermore, vector arithmetic in the latent space leads to meaningful outputs. For example Radford and colleagues[10] have shown that subtracting the latent space vectors of a face wearing glasses from a face without glasses results in a differential vector representing glasses; adding this to a different face outputs a face with glasses (i.e.$G(z')$ produces a person wearing glasses where $z' = (z_{glasses} - z_{no\ glasses}) + z_{face}$).

We hypothesised that latent space arithmetic can be extended to cell types and cell states. To investigate this we simulated the process of differentiation at the single cell level using latent space arithmetic, as shown in Figure 4A. We sampled 30 terminally differentiated and undifferentiated pairs of cells from the unseen pool of cells in the Joost dataset using the original study labels, obtained their latent space vectors $z_{differentiated}$ and z_{basal} and calculated the difference between these vectors $\delta = z_{differentiated} - z_{basal}$. We sampled cells from both the interfollicular epidermis and hair follicles in order to obtain a universal differentiation latent vector. Unseen cells were used to demonstrate that the generator's learned latent space could generalise to previously unseen gene expression profiles. We then added δ, the latent space differentiation vector, to an unseen undifferentiated cell and interpolated 1000 timepoints between the undifferentiated latent space point and the simulated differentiation latent space endpoint. We used the generator network to produce gene expression for all 1000 timepoints and repeated this process for 30 cells with similar undifferentiated starting profiles. It is important to note that only the starting point (the undifferentiated cell gene expression profile) is real, whereas the differentiated endpoint and all timepoints in between are generated by the neural network.

Figure 4B shows simulated time-series gene expression profiles for a selection of genes. Two differentiation markers, Periplakin (Ppl) and Grainyhead-like protein 3 homolog (Grhl3), demonstrate the ability of our latent space arithmetic to simulate successfully simulate differentiation of single cells. Over differentiation time points the mean expression level of these two markers increases, spanning the expression levels observed in the undifferentiated and differentiated subpopulations (B - basal start point, D - differentiated end point median expression line). Similar to real scRNA-seq gene expression profiles, the generated time points for individual cells display substantial cell-to-cell variation in gene expression over time, with cells showing different gene expression curves dependent on initial expression level and cell state. Similarly, for basal IFE markers such as the examples in Figure 3B, Integrin-beta 1 (Itgb1) and Metallothionein 2A (Mt2) the reduction in expression is successfully captured by the GAN latent space interpolation.

Moreover, the expression curve is not necessarily monotonic or linear. We also observed a range of highly non-linear expression profiles such as genes only expressed in early or late timepoints (Uxs1), parabola-like expression (Actb, Ube2f) and genes which change in expression variation but not mean expression. These non-linear expression profiles demonstrate that latent space interpolation produces meaningful gene expression predictions based on gene expression interdependencies and does not simply average between gene expression profiles.

To investigate further the dynamics of gene expression over epidermal differentiation we clustered the simulated time-series profiles for all 6605 genes. Using k-means clustering we were able to distinguish eight types of dynamic expression profile (Figure 4C). Genes expressed early in differentiation were enriched for epidermal stem cell related gene ontology and concordantly late-expressed genes were enriched for epidermal differentiation (q-value<0.01 for both, Supplementary Figure S3). We focused on four clusters, two corresponding to increasing and decreasing expression (Figure 4C, Groups I and II) and two subsets of genes only transiently up- or down-regulated during commitment to differentiation (Figure 4C, Group III and IV genes). To validate these findings we used a recently published dataset investigating human epidermal differentiation[25]. Mishra and colleagues utilised methylcellulose suspension-induced differentiation to obtain temporal differentiation gene expression data from human keratinocytes at 0, 4, 8 and 12 hours. A majority of our predicted transiently expressed genes are dynamically expressed in this dataset with Groups I and II showing decreasing and increasing bulk expression over the differentiation time course (Figure 4D, Group I and II). Focusing on Group IV, these 223 genes are of particular biological interest as they are predicted to be transiently expressed during commitment to differentiation and are therefore likely to play a functional role in this process. From our predicted epidermal commitment genes, this group clusters into three subgroups based on bulk peak expression around 0 hours, between 4 and 8 hours, and 12 hours or later. We hypothesise that three groups are observed as our simulated differentiation assumes a synchronised population of differentiating cells. The asynchronous nature of *in vitro* differentiation and heterogeneous starting population reduces our ability to detect transiently expressed and subsequently downregulated genes. From the subgroup with peak expression between 4 and 8 hours we identified two protein phosphatases, PTPN1 and PTPN13 which are also identified and extensively validated by Mishra and colleagues. This is followed by transient MAF expression - a member of the AP1 subfamily of differentiation transcription factors - predicted by our latent space interpolation approach and observed in the Mishra dataset along with previous studies[26].

In summary, using our generative model we have simulated a perturbation to cell state in the form of epidermal differentiation and subsequently obtained high-resolution predicted gene expression profiles. Using the generator neural network we have obtained information on switch-like expression of genes, an aspect of gene expression which is difficult to infer from previous methods of single cell RNA-seq analysis such as pseudotime ordering. Pseudo-ordering of cells provides an understanding of the sequence of gene expression events, however, this can be confounded by highly transient gene expression events which may lead to erroneously missing or identifying cell state transitions. In comparison our generative method is guided by gene expression rules learned by the GAN, hence producing valid gene expression profiles at all timepoints. These results can be extended to any other perturbation captured by latent space.

2.4 Discriminator network identifies state-determining gene expression ranges

As GAN training progresses, the discriminator network learns to identify compatible ranges of gene expression levels and interrelationships for all genes. An advantage of our approach is that these relationships can be highly non-linear, for example the discriminator can learn gene expression interdependencies which only apply when genes are expressed in a certain range or combination. To extract this learned information we performed a sensitivity analysis on the discriminator network by taking real gene expression profiles from the unseen cell group and varying expression of genes individually from the lowest observed expression level in the cohort to the maximum observed expression level. Figure 5A shows the relationship between adjusted expression level and change in discriminator network output critic value. This analysis produced a sensitivity curve for each gene where absolute sensitivity value indicates a strong change in discriminator critic value and the gradient of the sensitivity curve denotes ranges of gene expression that the discriminator is sensitive to. For genes with no known role in epidermal cell state such as Actb there is little change in sensitivity across all expression levels. In contrast, for known markers or determinators of epidermal state, such as Ivl, Krt10, Krt14, Cav2, Brd4 and Sox9, there is a strong relationship between expression level and sensitivity. For some genes such as Ivl, an IFE differentiation marker, an increase in expression showed increased sensitivity, i.e. a greater effect on discriminator critic value. However, for other genes such as Krt10 medium levels of transcription resulted in a lower sensitivity than low and high levels indicating that for this gene there are two important ranges of transcription. The discriminator network identification of Krt10 import expression ranges is supported by strong expression of Krt10 in the interfollicular epidermis above the basal layer and subsequent downregulation in terminally differentiating cells[22]. Krt10 misexpression has been shown to cause epidermal barrier defects[27,28]. Furthermore, a cellular need for binary expression of Krt10 is clear as the protein constitutes approximately 40% of all cellular proteins in the suprabasal layer of the epidermis[29,30].

We sought to compare the relative importance of genes in determining cell transcriptional state across all epidermal cell types. Sensitivity varies non-linearly with expression level, hence, to compare all genes we calculated the maximum gradient of the sensitivity curve for each gene (Figure 5B). This metric is an indication there is at least one range of expression levels which has a large effect on discriminator network output. Strikingly, the top 100 genes identified are highly enriched for known epidermal regulators and comprise both low and highly expressed genes (Figure 5C). Amongst the top 10 are three keratin genes with spatially-restricted expression; Krt6a (inner bulge), Krt10 (suprabasal IFE) and Krt14 (basal IFE). Our analysis predicted Lars2 expression as one of the most important indicators of pan-epidermal cell state. Lars2 is a mitochondrial leucyl-tRNA synthetase and is upregulated during differentiation as seen by our differentiation predictions (Supplementary Figure S4) and from the Joost dataset[22]. Considering these analyses we hypothesise that Lars2 is one of the key regulators of cell metabolic considerations during differentiation. This sensitivity analysis approach enables identification of state-regulating genes without bias for transcript abundance.

2.5 GAN-derived gene association networks predict Gata6 targets

Finally we examined the internal features of the generator network to extract gene expression interdependencies to complement our sensitivity analysis ranking of state-determining genes by providing context on how these genes are coregulated. The final layer of the generator network non-linearly transforms an arbitrary number of internal features to the final 6605 gene expression values using a leaky rectified linear unit activation function. We found empirically that 600 internal features produced stable and diverse generator output, representing a maximum of 600 features that, when non-linearly combined, produce each gene expression output value. Since these internal features are an order of magnitude fewer than the number of gene expression values, the generator network is forced to learn the most salient relationships between genes in order to produce a convincing output. Hence, genes with correlated generator final layer values are predicted to correspond to co-regulated genes. This is distinct from correlation analysis of gene expression, which infers linear and directly correlated regulation.

We visualised the correlation of the final layer weights for the 600 features between all genes using a force directed network (Figure 5D, middle network) and also examined the structure and clustering of the genes using a hierarchically clustered heatmap (Supplementary Figure S5). On a macroscopic scale genes are segregated by their overall positive or negative effect on gene expression. On a local scale the majority of genes cluster into small groups of between 10 and 50 closely associated genes. We used these local gene association networks to examine Lars2, a gene predicted from our analysis to be highly important for epidermal cell state. Lars2 is a mitochondrial leucyl-tRNA synthetase and our analysis shows it to be a member of an extremely closely regulated group of ribosomal and translation related genes. We also examined the local gene association networks for three known epidermal regulators examined in our previous sensitivity analysis Krt14, Itga6 and Ivl. Itga6 and Krt14 are both known basal IFE markers; the local network for these genes highlights the power of our machine learning approach to identify coregulated genes. Several known basal IFE regulators are present both local networks, such as Krt5, Krt14, Itga3, Cav2 and Col17a1. We observed a similar pattern of enriched associated genes for Ivl, an IFE differentiation marker which is co-associated with several other differentiation genes and Rps29 a marker of IFE-derived cells which is exclusively associated with ribosomal genes in our local networks. Furthermore, the local network for Notch ligand Jag2 contains several hair follicle bulge and basal IFE regulators, corresponding to its role in these epidermal locations[31]. Strikingly, Wnt3 is present in this local network indicative of Wnt-Notch signaling interplay[32,33], which cannot be seen at the RNA level using conventional gene expression analysis. These co-regulated epidermal genes are robustly identified despite little correlation of expression in the scRNA-seq data.

Finally, in order to explore the predictive power of our neural network approach we derived the local gene association network for the GATA-binding factor 6 transcription factor, GATA6 (Figure 5E). We predicted that as a transcription factor, the GATA6 network would contain many genes which are directly regulated by GATA6. Hence, we hypothesised that genes in the network with a positive final layer correlation should be upregulated in GATA6$^+$ cells. Using a previous dataset from the lab, we contrasted bulk gene expression of GATA6$^+$ and GATA6$^-$ cells derived from the junctional zone and sebaceous duct of the epidermis (Figure 5F)[34]. Figure 4G shows the log fold-change distribution of genes between the GATA6$^+$/GATA6$^-$ subpopulations. Genes derived from our GATA6 gene regulatory network show significantly higher expression in the GATA6$^+$ cells when compared to all other genes ($p < 0.05$, Kolmogorov-Smirnov test). Taken together these results suggest GANs are capable of elucidating complex gene interrelationships beyond the limits of linear correlation analyses.

3. Discussion

We have addressed a major goal of single cell gene expression analysis; the desire to obtain functional gene relationships and a predictive model of transcriptional state in single cells. This serves as a framework for understanding cell fate transitions, disease-causing perturbations and state-determining genes across tissue types.

Here we demonstrate the effectiveness of a new deep learning approach in integrating multiple gene expression datasets, despite no prior adjustment for technical and batch-to-batch variation. To the best of our knowledge, this is the first time GANs have been applied to genomic data. Our approach of combining data from cells in different environments with diverse technical and biological noise allows robust inference of gene associations intrinsic to epidermal cells rather than experimental conditions. We focused on skin due to the wealth of knowledge available regarding subpopulations and cell state. However, our generative model can be extended to integrate cells from multiple tissues and we anticipate this will advance our understanding of gene expression relationships across tissues.

Currently, our analysis is constrained by available data. As further scRNA-seq data becomes available, particularly through large scale projects such as the Human Cell Atlas[35] we envisage that GANs will be a viable strategy to analyse all human cell types en masse. In our study we have shown that a relatively small dataset of under 2000 cells is sufficient to uncover state-determining genes and gene expression networks in the epidermis. GANs applied to a greater dataset covering multiple tissues could be used to resolve organism-wide gene expression relationships and to predict previously unseen cell state transitions.

4. Methods

Datasets and data preparation

All datasets analysed in this study are publicly available under the GEO accessions GSE90848, GSE67602 and GSE99989. We combined these three datasets using transcripts per million (TPM) normalisation and removed labelled non-epithelial cells from the Joost dataset. Data was filtered for: cells with more than 1000 genes detected $\log_2(\text{TPM}+1) > 1$ and genes expressed in more than 500 cells at $\log_2(\text{TPM}+1) > 1$. All neural network input data was $\log_2(\text{TPM}+1)$ and the combined dataset is available as a CSV matrix at https://github.com/luslab/scRNAseq-WGAN-GP/tree/master/data.

Tools and code availability

Data preparation was performed using R. Gene ontology enrichment was evaluated using enrichR[36]. All other analysis and implementation of the generative adversarial networks was performed using Python (numpy, sklearn and networkx). We used Google's Tensorflow[37] deep learning framework to implement, train and monitor our neural networks. We have provided a Jupyter Notebook containing our implementation of the generative adversarial network available at https://github.com/luslab/scRNAseq-WGAN-GP.

Generative adversarial network algorithm

We adapted our generative adversarial network algorithm from four previous works[9,10,13,24]. Both generator and discriminators are fully connected neural networks with one hidden layer. We used a Leaky Rectified Linear Unit (LReLU) as the activation function for both networks using a coefficient of 0.2.

The generator input consists of a 100-dimensional latent variable (z) and a hidden layer size of 600 fully connected units. In order to reduce training time and data required for the generator network to simulate the Poisson distributed nature of scRNA-seq count data we trained the generator network using an additive Poisson and Gaussian distributed latent variable where $z = N(\mu = 0, \sigma = 0.1 * max) + P(\lambda = 1)$. Generator output is a 6605 dimensional vector representing the 6605 genes in the training cohort.

The discriminator network input is a 6605 dimensional vector (gene expression profile) and has a hidden layer size of 200 fully connected units. The final layer of the discriminator does not apply an activation function, in line with other Wasserstein GANs using a gradient penalty loss function.

We used backpropagation and RMSProp to train the two neural networks (learning rate of 5e-5) and a batch size of 32 cells. Additionally, data were augmented during training by randomly permuting the expression of 10 genes expressed at $\log_2$(TPM+1) < 3. We used several loss functions during this study before finalising on using the Wasserstein-GAN (WGAN) with gradient penalty loss function. WGAN replaces the original GAN where data distributions of generated and real data are compared using the Jensen-Shannon divergence for the loss function[9]. In WGANs the loss function is calculated using the Wasserstein distance which gives improved training stability[13]. Gulrajani and colleagues further improved on this variant of GANs by eliminating weight-clipping in favor of a gradient penalty in the form of the gradient between pairs of real and generated samples[24].

The two loss functions to minimise are:

$$L_{Discriminator} = E[D(G(z))] - E[D(x)] + \lambda E[(||\nabla D(x')||_2 - 1)^2]$$

$$L_{Generator} = -E[D(G(z))] \,,$$

where G is the generator network and $D$is the discriminator network such that $G: z \rightarrow \hat{x}$, $z$is the generator latent variable, $\hat{x}$ is a generated gene expression profile, x is a real gene expression profile and x' is uniformly sampled between pairs of generated and real gene expression profiles.

Convergence of the loss function is not sufficient to evaluate completed training of GANs. For each timepoint we evaluated the diversity of cells produced by the generator by simulating 500 cells and calculating the median distance (Pearson), i.e. median cell-cell distance.

For a list of GAN parameters used see Supplementary Table 2.

Dimensionality reduction and clustering

We evaluated clustering and structure of real and generated single cell RNA-seq profiles in three ways 1) PCA (principal component analysis), 2) t-distributed stochastic neighbour embedding (t-SNE) on data with the top two principal components removed. 3) t-SNE on the hidden layer output values of the discriminator network. For all t-SNEs we used a default perplexity parameter of 30.

Simulated time-series gene expression profiles were clustered by k-means clustering with 20 expected clusters. Clustered dynamic gene expression profiles were subsequently visualised and two pairs of clusters were manually merged where they were deemed to be visually similar.

Latent space mapping and interpolation

Terminally differentiated cells and basal IFE cells were sampled from the combined dataset on using labels from the original studies. In order to generate corresponding latent space variable z_{basal} and $z_{differentiated}$ for cells x_{basal} and $x_{differentiated}$ we randomly generated cell expression profiles until the correlation between x and the generated cell $G(z)$reached a threshold of 0.7. Figure S1B shows that a correlation of 0.7 between two cells is rare, hence, we assume that this threshold is sufficient to generate a gene expression profile representative of the original cell state. We next calculated the difference between the basal and differentiated cell in latent space i.e. $\delta = z_{basal} - z_{differentiated}$. This latent space representation of differentiation was used to simulate differentiation on a further unseen cell where we obtained $z_{unseen}$and calculated a predicted terminally differentiated latent space point $z_{unseen\ differentiated} = z_{unseen} + \delta$. To obtain simulated gene expression time series data we interpolated 1000 points between $z_{unseen}$and $z_{unseen\ differentiated}$and generated 1000 intermediate gene expression profiles from these.

Discriminator network sensitivity analysis

Contribution of genes to output and sensitivity of the discriminator network was performed by taking unseen cells and varying the output of each gene between the minimum and maximum observed across all cells. For important genes this results in a change of the discriminator output $D(x_p)$where $x_p$is a perturbed gene expression profile. We performed 100 linearly interpolated perturbations. Sensitivity curves were calculated by taking the mean normalised $D(x_p)$at each expression level as unseen cells have can differ in their baseline or unperturbed discriminator score $D(x)$. Max sensitivity gradient (see Figure 5B) was calculated by taking the absolute value of maximum gradient of these curves.

Local gene association networks

Gene co-regulatory relationships were inferred by analysing the weights of the generator neural network final layer. The Pearson correlation matrix of the final layer was used as a network adjacency matrix. To construct association networks for a gene we selected all local genes with a final layer correlation > 0.5 and retained any connections within the local network also with final layer correlation > 0.5.

Author Contributions

AG performed analyses and wrote the manuscript. NML and FMW supervised the project and co-wrote the paper.

Acknowledgements

We are grateful to Victor Negri, Aylin Cakiroglu and Dean Plumbley for helpful advice and discussions. This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC010110), the UK Medical Research Council (FC010110), and the Wellcome Trust (FC010110), Wellcome Trust funding to NML (103760/Z/14/Z), and MRC Medical Bioinformatics Award eMedLab to NML (MR/L016311/1).

References

1. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2,** 666–673 (2012).
2. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).
3. Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12,** 44–73 (2017).
4. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16,** 241 (2015).
5. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14,** 414–416 (2017).
6. Lin, C., Jain, S., Kim, H. & Bar-Joseph, Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* **45,** e156 (2017).
7. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118 (2017).
8. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18,** 67 (2017).
9. Goodfellow, I. J. *et al.* Generative Adversarial Networks. *arXiv [stat.ML]* (2014).
10. Radford, A., Metz, L. & Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv [cs.LG]* (2015).
11. Chen, X. *et al.* InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *arXiv [cs.LG]* (2016).
12. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral Normalization for Generative Adversarial Networks. in *ICML Workshop on Implicit Models* (2017).
13. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN. *arXiv [stat.ML]* (2017).

14. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. Improved Training of Wasserstein GANs. *arXiv [cs.LG]* (2017).

15. Creswell, A. *et al.* Generative Adversarial Networks: An Overview. *arXiv [cs.CV]* (2017).

16. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv [cs.NE]* (2017).

17. Zamyatin, E. & Filchenkov, A. Learning to Generate Chairs with Generative Adversarial Nets. *arXiv [cs.CV]* (2017).

18. Watt, F. M. & Hogan, B. L. Out of Eden: stem cells and their niches. *Science* **287,** 1427–1430 (2000).

19. Hsu, Y.-C., Li, L. & Fuchs, E. Emerging interactions between skin stem cells and their niches. *Nat. Med.* **20,** 847–856 (2014).

20. Page, M. E., Lombard, P., Ng, F., Göttgens, B. & Jensen, K. B. The epidermis comprises autonomous compartments maintained by distinct stem cell populations. *Cell Stem Cell* **13,** 471–482 (2013).

21. Ghahramani, A., Donati, G., Luscombe, N. M. & Watt, F. M. Epidermal Wnt signalling regulates transcriptome heterogeneity and proliferative fate in neighbouring cells. *Genome Biol.* **19,** 3 (2018).

22. Joost, S. *et al.* Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity. *cels* **3,** 221–237.e9 (2016).

23. Yang, H., Adam, R. C., Ge, Y., Hua, Z. L. & Fuchs, E. Epithelial-Mesenchymal Micro-niches Govern Stem Cell Lineage Choices. *Cell* **169,** 483–496.e13 (2017).

24. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C. Improved Training of Wasserstein GANs. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 5763–5773 (Curran Associates, Inc., 2017).

25. Mishra, A. *et al.* A protein phosphatase network controls the temporal and spatial dynamics of differentiation commitment in human epidermis. *Elife* **6,** (2017).

26. Lopez-Pajares, V. *et al.* A LncRNA-MAF:MAFB transcription factor network regulates epidermal differentiation. *Dev. Cell* **32,** 693–706 (2015).

27. Müller, F. B. *et al.* A human keratin 10 knockout causes recessive epidermolytic hyperkeratosis. *Hum. Mol. Genet.* **15,** 1133–1141 (2006).

28. Cheng, J. *et al.* The genetic basis of epidermolytic hyperkeratosis: a disorder of differentiation-specific epidermal keratin genes. *Cell* **70,** 811–819 (1992).

29. Fuchs, E. & Green, H. Changes in keratin gene expression during terminal differentiation of the keratinocyte. *Cell* **19,** 1033–1042 (1980).

30. Fuchs, E., Esteves, R. A. & Coulombe, P. A. Transgenic mice expressing a mutant keratin 10 gene reveal the likely genetic basis for epidermolytic hyperkeratosis. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 6906–6910 (1992).

31. Powell, B. C., Passmore, E. A., Nesci, A. & Dunn, S. M. The Notch signalling pathway in hair growth. *Mech. Dev.* **78,** 189–192 (1998).

32. Estrach, S., Ambler, C. A., Lo Celso, C., Hozumi, K. & Watt, F. M. Jagged 1 is a beta-catenin target gene required for ectopic hair follicle formation in adult epidermis. *Development* **133,** 4427–4438 (2006).

33. Shi, Y. *et al.* Wnt and Notch signaling pathway involved in wound healing by targeting c-Myc and Hes1 separately. *Stem Cell Res. Ther.* **6,** 120 (2015).

34. Donati, G. *et al.* Wounding induces dedifferentiation of epidermal Gata6(+) cells and acquisition of stem cell properties. *Nat. Cell Biol.* **19,** 603–613 (2017).

35. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550,** 451–453 (2017).

36. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–7 (2016).

37. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv [cs.DC]* (2016).

Figures, Figure Legends and Tables

Figure 1

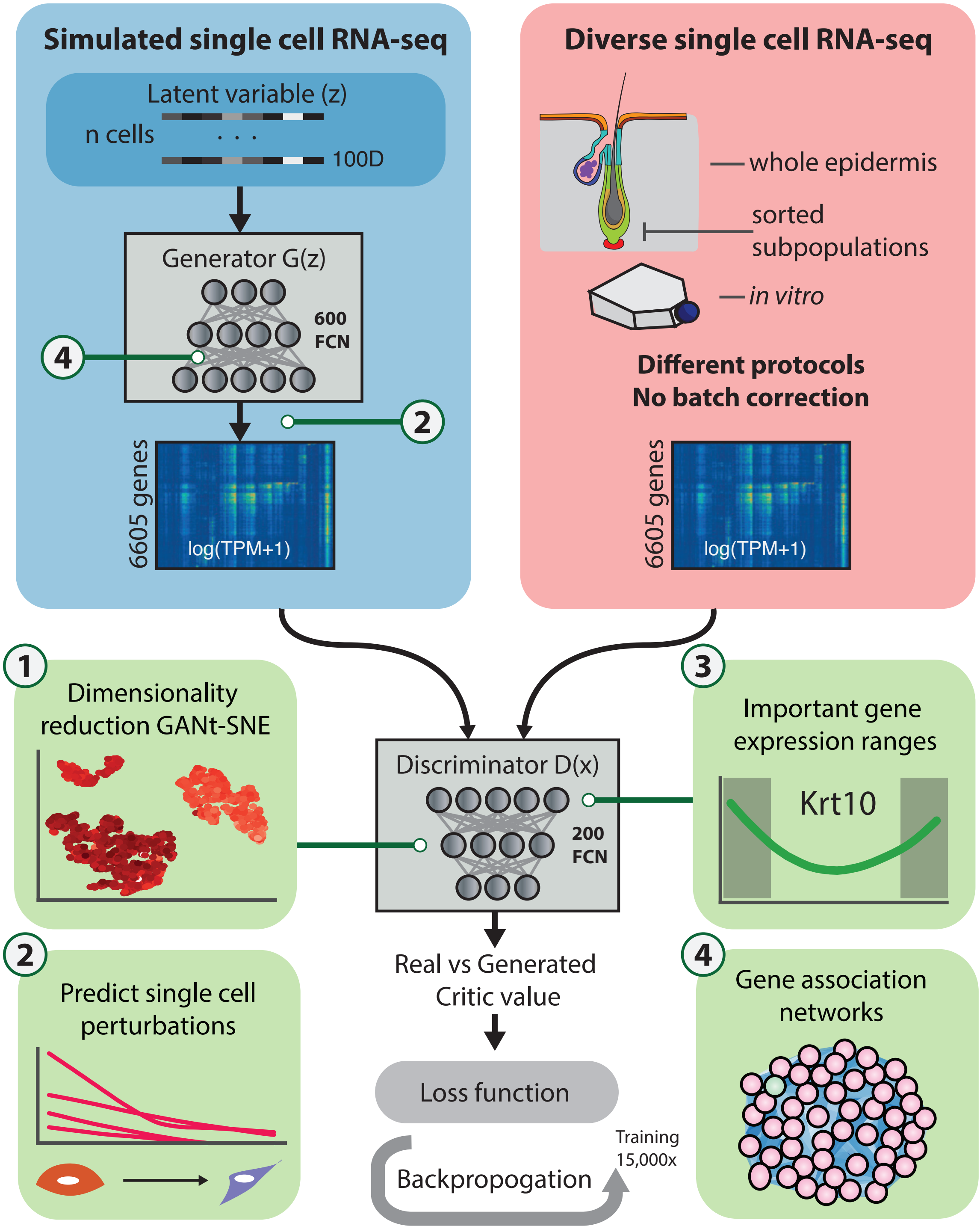


Figure 1. Overview of single cell RNA-seq analysis with generative adversarial networks Generative adversarial networks consist of two neural networks concurrently training whilst competing against one another. These networks referred to as the generator and discriminator networks each have a distinct task. The generator is tasked with generating data by transforming a 100-dimensional latent variable into a single cell gene expression profile. In turn the discriminator is tasked with evaluating whether data is authentic or generated. Only the discriminator network sees the real single cell RNA-seq data, which have not been corrected for batch effects or technical variation. Once trained, we are able to extract biologically meaningful information from the generator and discriminator networks such as gene association networks, predict single cell time-courses, important gene expression ranges and biologically meaningful dimensionality reduction. Fully connected neurons, FCN.

Figure 2

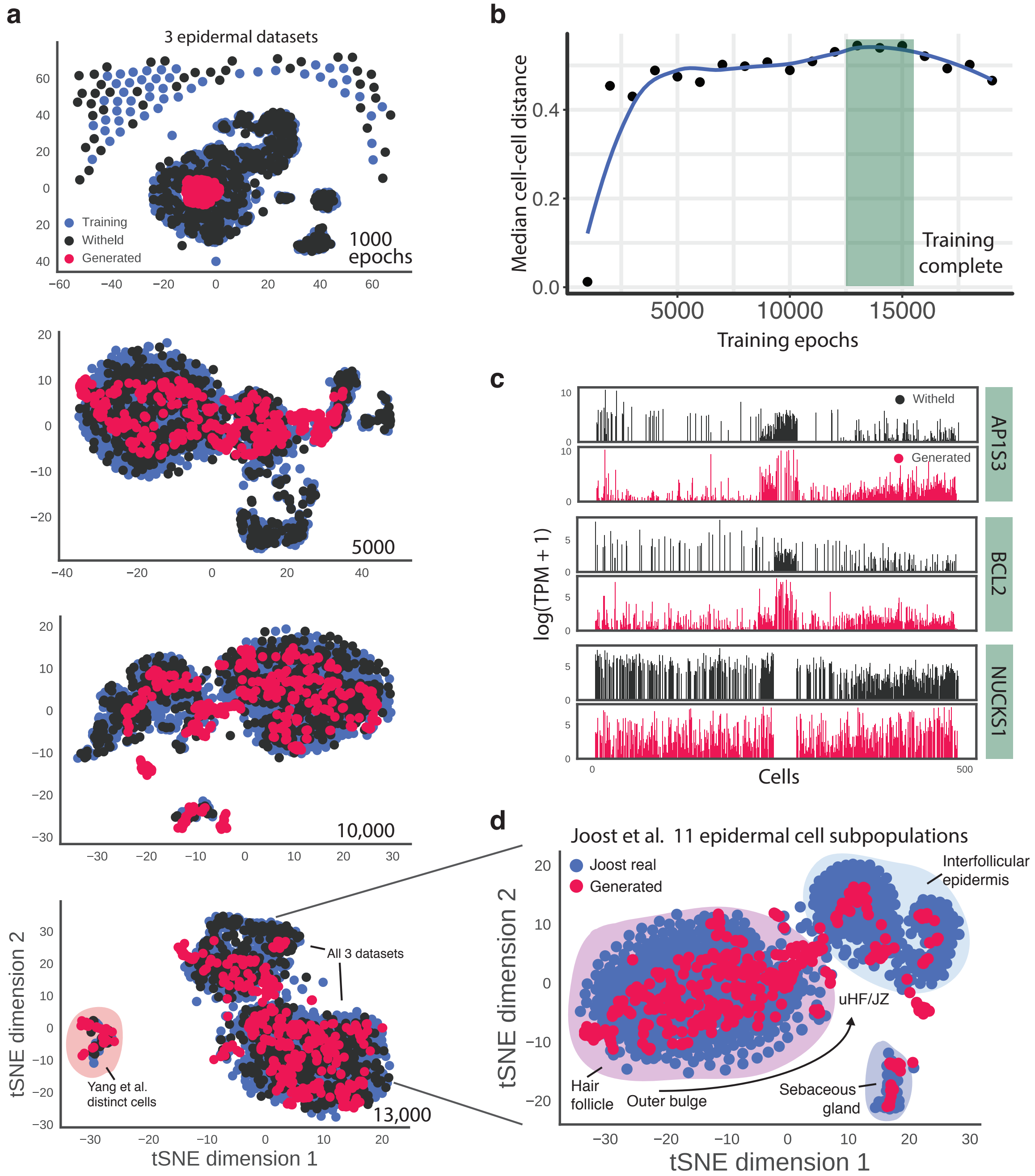

a
3 epidermal datasets
Training
Witheld
Generated
1000 epochs
60
40
20
0
20
40
−60
−40
−20
0
20
40
60
5000
−40
−20
0
20
40
20
10
0
−10
−20
10,000
−30
−20
−10
0
10
20
30
13,000
All 3 datasets
Yang et al. distinct cells
tSNE dimension 2
tSNE dimension 1
b
Median cell-cell distance
0.0
0.2
0.4
5000
10000
15000
Training epochs
Training complete
c
Witheld
Generated
log(TPM + 1)
AP1S3
BCL2
NUCKS1
10
5
0
Cells
0
500
d
Joost et al. 11 epidermal cell subpopulations
Joost real
Generated
Interfollicular epidermis
uHF/JZ
Hair follicle
Outer bulge
Sebaceous gland
tSNE dimension 2
tSNE dimension 1
20
10
0
−10
−20
−30
−20
−10
0
10
20
30

Figure 2. GANs model epidermal single cell gene expression

a) t-SNE projection of 1763 real training cells, 500 real withheld cells and 500 generated single cell expression data: each data point represents a cell and cells with similar expression profiles are positioned close together. Generated cells are clustered at the beginning, but with training gradually occupy the entire range of expression outputs for different cell types across three studies. Cells are visualised at 1000, 5000, 10,000 and 13,000 training steps.

b) Relationship between GAN training epochs and output diversity as measured by median cell-cell distance. Blue line is a LOESS regression. The greatest diversity of gene expression is achieved between 13,000 and 15,000 steps.

c) Bar plots showing expression of Ap1s3, Bcl2 and Nucks1 in 500 unseen real cells and 500 closely related generated cells. Cells are ordered by hierarchical clustering of the unseen real cells using all 6605 genes.

d) t-SNE projection of subset of real cells from Joost et al. and corresponding generated cells. Despite training with a larger and diverse dataset, the GAN captures cell diversity at a detailed level described by Joost. Labels show the corresponding cell type for each cluster. Upper hair follicle, uHF. Junctional zone, JZ

Figure 3

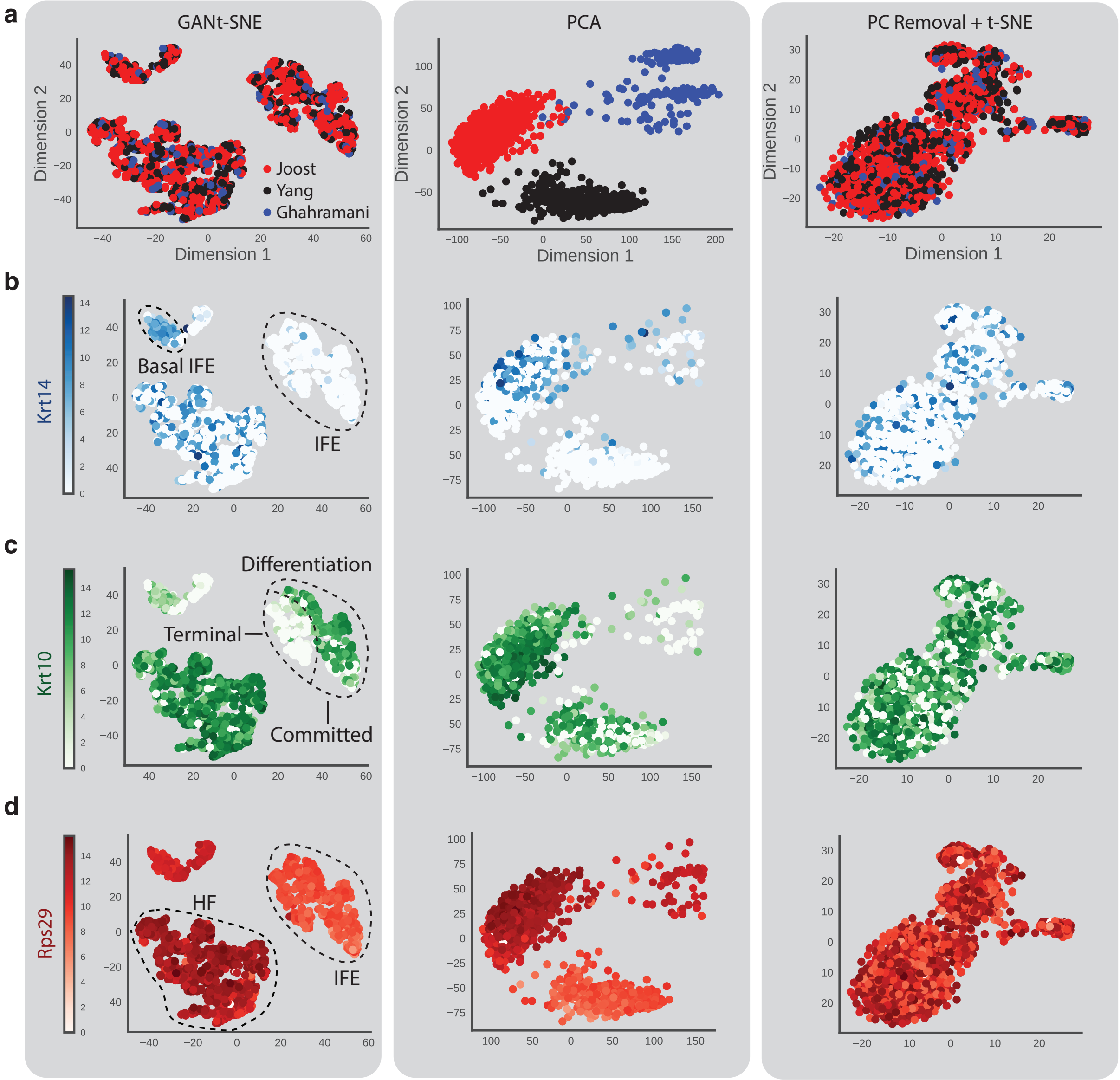

a
GANt-SNE
PCA
PC Removal + t-SNE
Dimension 2
Dimension 1
Joost
Yang
Ghahramani
40
20
0
−20
−40
60
100
50
−50
−100
150
200
30
10
−10
b
Krt14
Basal IFE
IFE
14
12
10
8
6
4
2
75
25
−75
c
Krt10
Differentiation
Terminal
Committed
d
Rps29
HF
IFE

Figure 3. GANt-SNE clusters biologically similar cells

a) GANt-SNE visualises 2263 real cells by performing t-SNE on 200 features captured by the discriminator neural network's hidden layer. This is compared with principal component analysis (PCA) and t-SNE after removal of linear batch effects using PCA (PC removal + t-SNE). PCA is heavily affected by batch effect and clusters the three datasets separately.

b) Keratin-14 expression overlaid onto projections. Krt-14 is a marker for basal interfollicular epidermis (IFE). GANt-SNE distinguishes between basal IFE and non-basal IFE cells, whereas PCA and PC removal + t-SNE fail to distinguish between these cell types.

c) Keratin-10 is a marker for differentiation interfollicular epidermis (IFE). . GANt-SNE successfully clusters separates supra-basal and terminally differentiated IFEs.

d) Ribosomal Protein S29 is known to be expressed highly in interfollicular epidermis (IFE) and at medium levels in hair follicles (HF). GAN-tSNE successfully distinguishes between these two major cell types.

Figure 4

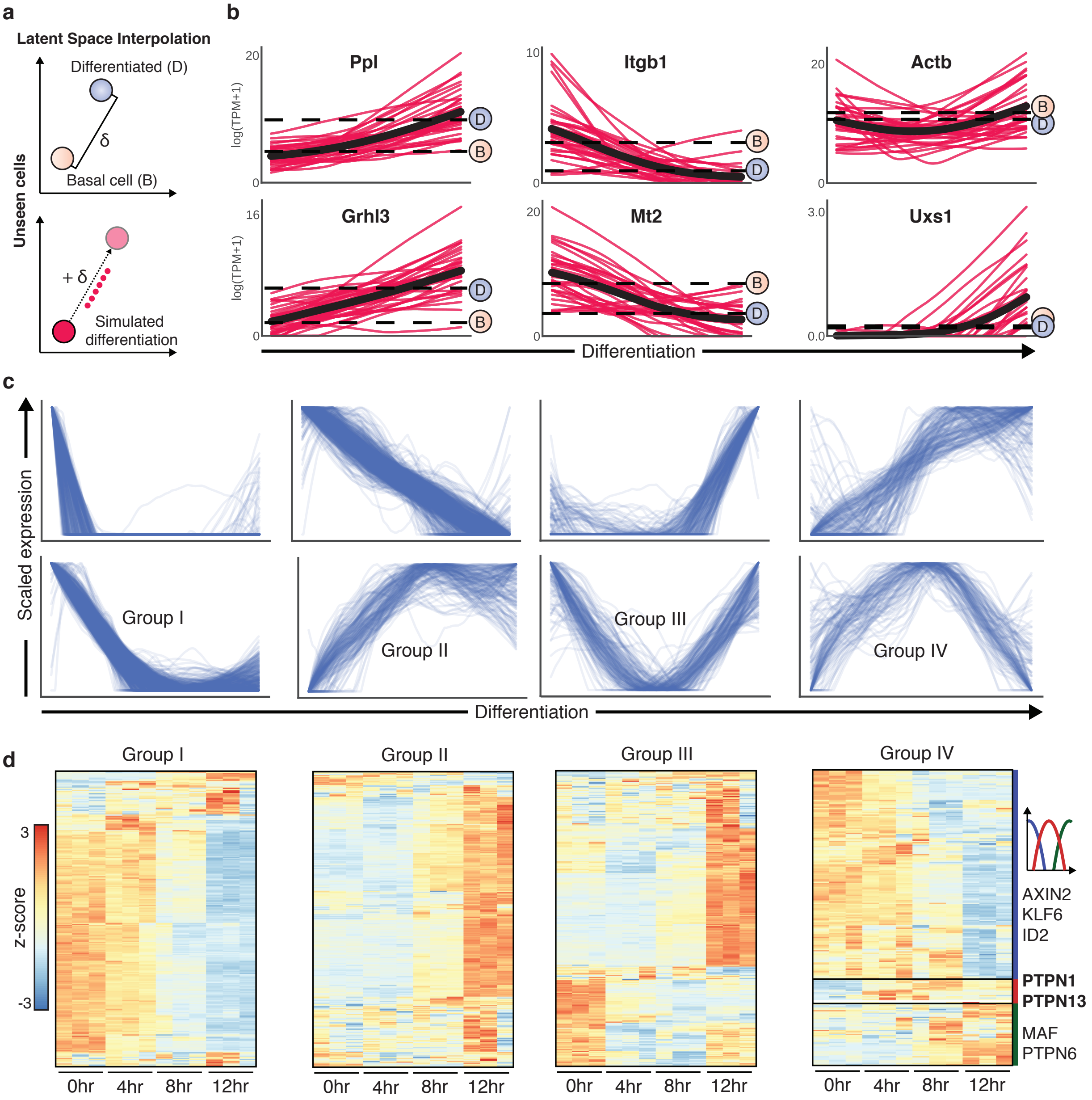

a
Latent Space Interpolation
Unseen cells
Differentiated (D)
δ
Basal cell (B)
+ δ
Simulated differentiation
b
log(TPM+1)
Ppl
Itgb1
Actb
Grhl3
Mt2
Uxs1
20
10
16
3.0
0.0
0
B
D
Differentiation
c
Scaled expression
Group I
Group II
Group III
Group IV
Differentiation
d
Group I
Group II
Group III
Group IV
z-score
3
-3
0hr
4hr
8hr
12hr
AXIN2
KLF6
ID2
PTPN1
PTPN13
MAF
PTPN6

Figure 4. Latent space interpolation allows simulation of single cell differentiation
a) Schematics showing how to simulate differentiation for single cells using latent space interpolation. Expression data for pairs of undifferentiated and differentiated cells are transformed to their respective latent space vectors (z_{basal}, $z_{differentiated}$)) The difference between the vectors (δ) represents differentiation in latent space (upper panel). This difference vector is then applied to the latent space vector of another unseen cell in order to predict the terminally differentiated state (lower panel). We also simulate 1000 differentiation timepoints.
b) Simulated expression profiles for 6 representative genes in 30 cells over 1000 differentiation timepoints. Red lines indicate expression profiles in individual cells and lack lines provide the mean expression over all cells. Dotted lines correspond to the median expression of basal undifferentiated (B) and differentiated (D) cells from the Joost dataset. Periplakin (Ppl) and Grainyhead like transcription factor 3 (Grhl3) are known differentiation whose simulations show increasing expression. Integrin beta-1 (Itgb1) and Metallothionein 2 (Mt2) are basal IFE marker which display gradually decreasing expression Actin beta (Actb) and UDP-Glucuronate Decarboxylase 1 (Uxs1) are not traditionally considered to be markers but display non-linear expression profiles with substantial cell to cell variability.
c) K-means clustering of simulated differentiation expression profiles for 6605 genes. We identify eight broad types of temporal gene expression, of which four of the dynamic expression patterns are labelled Groups I-IV.
d) Heatmaps showing bulk expression levels of genes in Groups I-IV from Mishra et al. at 0, 4, 8 and 12 hours of suspension-induced differentiation. Most genes display similar expression profiles to the simulations, despite the differences in experimental set ups.

Figure 5

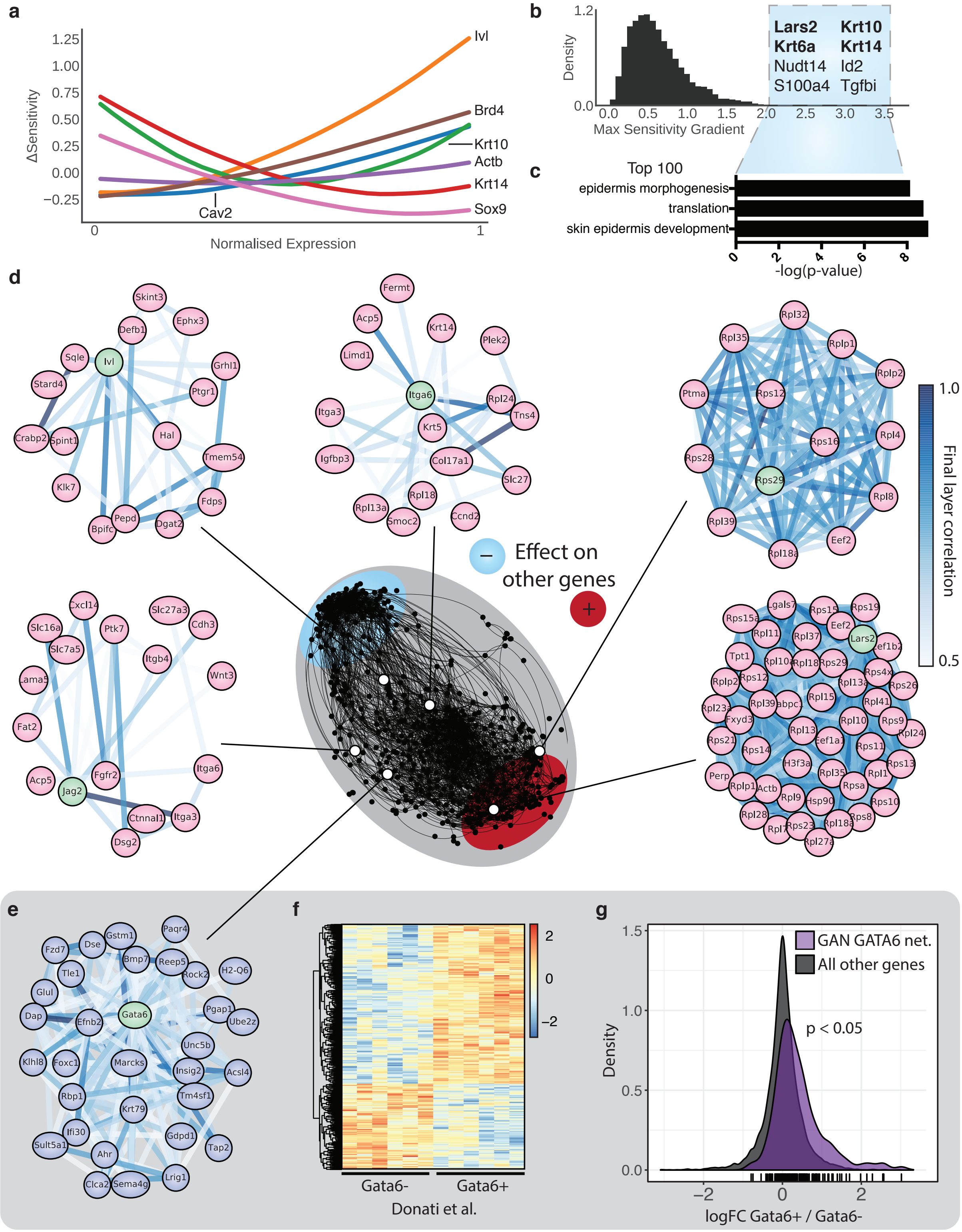

a
ΔSensitivity
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0
1
Normalised Expression
Ivl
Brd4
Krt10
Actb
Krt14
Sox9
Cav2
b
1.2
0.0
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Max Sensitivity Gradient
Lars2
Krt10
Krt6a
Krt14
Nudt14
Id2
S100a4
Tgfbi
c
Top 100
epidermis morphogenesis
translation
skin epidermis development
0
2
4
6
8
-log(p-value)
d
Skint3
Ephx3
Defb1
Sqle
Ivl
Grhl1
Stard4
Ptgr1
Crabp2
Spint1
Hal
Tmem54
Klk7
Fdps
Pepd
Bpifc
Dgat2
Fermt
Acp5
Krt14
Plek2
Limd1
Itga6
Rpl24
Itga3
Tns4
Krt5
Igfbp3
Col17a1
Slc27
Rpl18
Rpl13a
Smoc2
Ccnd2
Rpl32
Rpl35
Rplp1
Rplp2
Ptma
Rps12
Rpl4
Rps16
Rps28
Rps29
Rpl8
Rpl39
Eef2
Rpl18a
1.0
0.5
Final layer correlation
Effect on other genes
−
+
Cxcl14
Slc27a3
Slc16a
Ptk7
Cdh3
Slc7a5
Itgb4
Lama5
Wnt3
Fat2
Itga6
Acp5
Fgfr2
Jag2
Ctnnal1
Itga3
Dsg2
Lgals7
Rps15
Rps19
Rps15a
Eef2
Rpl11
Rpl37
Lars2
Eef1b2
Tpt1
Rpl10a
Rpl18
Rps29
Rps4x
Rplp2
Rps12
Rpl13a
Rps26
Rpl23
Rpl39
Fabpc
Rpl15
Rpl41
Fxyd3
Rpl10
Rps9
Rpl13
Rpl24
Rps21
Rps14
Eef1a1
Rps11
H3f3a
Rps13
Perp
Rpl35
Rpl1
Rplp1
Actb
Rpsa
Rpl9
Hsp90
Rps10
Rpl28
Rpl18a
Rps8
Rpl7
Rps23
Rpl27a
e
Paqr4
Gstm1
Fzd7
Dse
Bmp7
Reep5
Tle1
Rock2
H2-Q6
Glul
Gata6
Pgap1
Dap
Efnb2
Ube2z
Unc5b
Klhl8
Foxc1
Marcks
Insig2
Acsl4
Rbp1
Tm4sf1
Krt79
Ifi30
Gdpd1
Sult5a1
Tap2
Ahr
Lrig1
Clca2
Sema4g
f
2
0
−2
Gata6-
Gata6+
Donati et al.
g
1.5
1.0
0.5
0.0
Density
GAN GATA6 net.
All other genes
p < 0.05
−2
0
2
logFC Gata6+ / Gata6-

Figure 5. Sensitivity analysis identifies state-determining genes and local gene association networks

a) Sensitivity curves for seven known markers for epidermal or determinants of epidermal state: Actin beta (Actb), Involucrin(Ivl), Keratin 10 (Krt10), Keratin 14 (Krt14), Caveolin 2 (Cav2), Bromodomain-containing protein 4 (Brd4), SRY-Box 9 (Sox9). Some genes such as Ivl an IFE differentiation marker display higher sensitivity at increased expression levels whereas others display the opposite trend.

b) Histogram of maximum sensitivity curve gradients of 6605 genes, with the top eight genes labelled; known markers are highlighted in bold.

c) Gene ontology enrichment for the top 100 genes by maximum sensitivity gradient as scored by -log10(p-value).

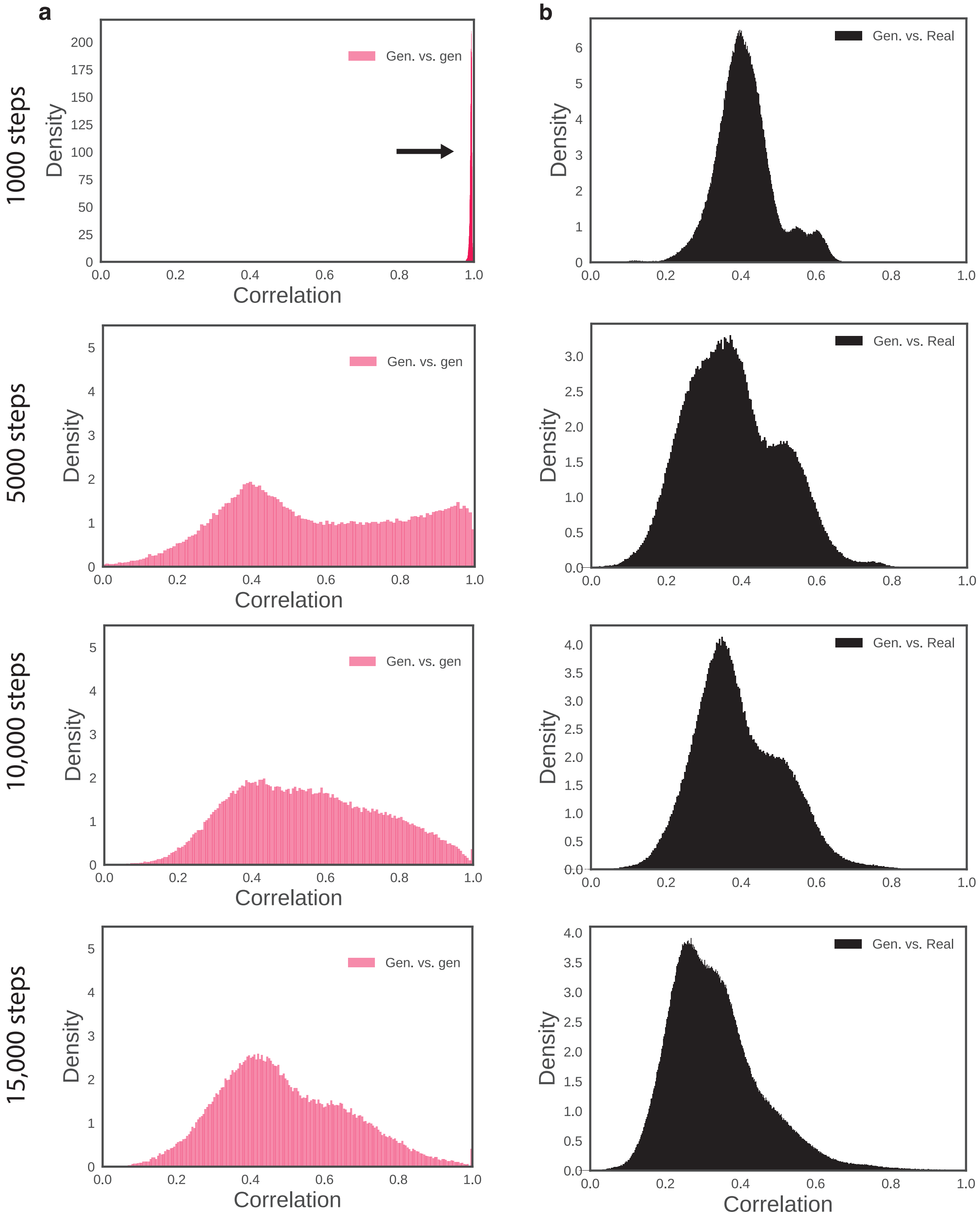
d) Network representation of the correlation between discriminator neural network final layer values for 6605 genes (centre panel); genes are depicted as nodes and correlations between them are shown as edges (threshold final layer correlation: 0.5). Network regions containing genes with overall positive and negative effect on other genes are highlighted in red and blue respectively. Surrounding local network are local gene association networks for Lars2, Itga6, Ivl, Rps29 and Jag2; edges are colored by generator final layer correlation.

e) Local gene association network for Gata6 displayed as described above.

f) Heatmap showing differentially expressed gene expression z-scores for Gata6$^+$ and Gata6$^-$ junctional zone and sebaceous duct cells from Donati et al.

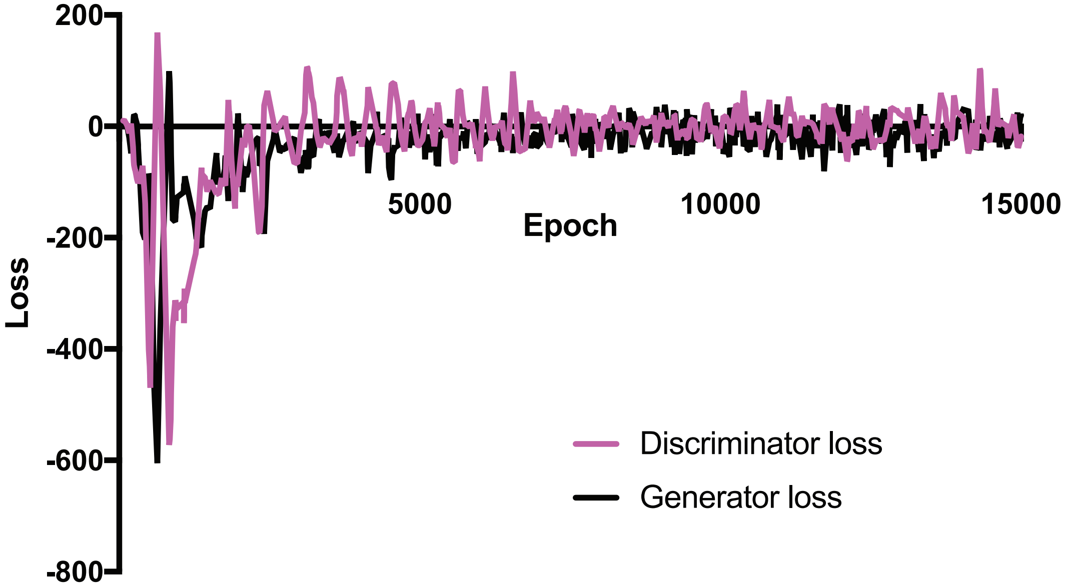
g) Probability density plot of the log-fold change in gene expression between Gata6$^+$ and Gata6$^-$ cells for genes in the predicted GATA6 local association network (violet) and all other genes (grey). Network genes display greater expression changes between Gata6$^+$ and Gata6$^-$ cells (p-value = 1.3e-4).

Figure S1


a
b
1000 steps
5000 steps
10,000 steps
15,000 steps
Gen. vs. gen
Gen. vs. Real
Density
Correlation
0
25
50
75
100
125
150
175
200
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0

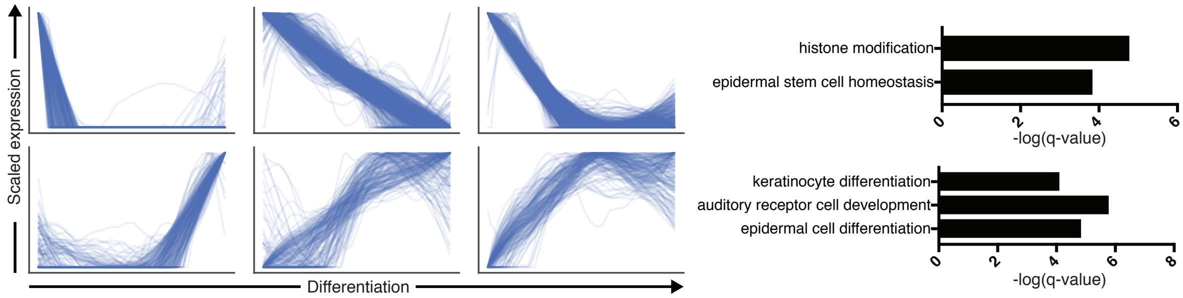
Supplementary Figure S1. GAN training improves expression output diversity
a) Distribution of pairwise correlation between pairs of real or generated cells after 1000, 5000, 10,000 and 15,000 training steps.
b) Distribution of pairwise correlation between pairs of real and generated cells after 1000, 5000, 10,000 and 15,000 training steps.

Figure S2



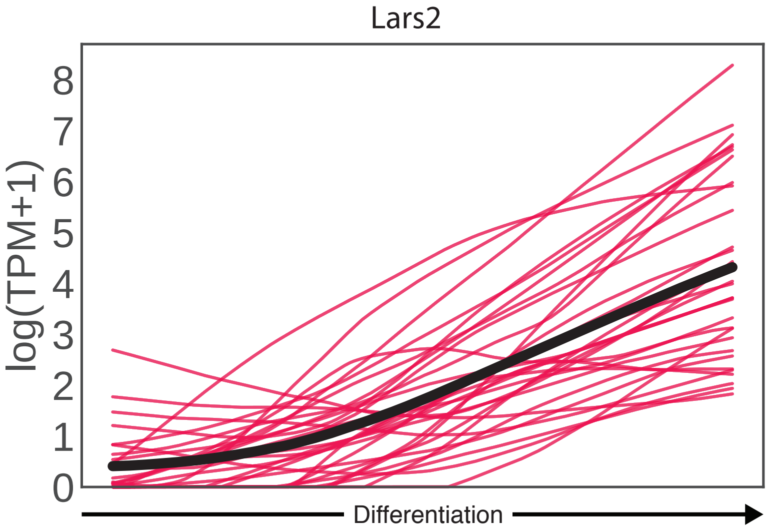
Supplementary Figure S2. Generator and discriminator network loss curves
Pink line shows discriminator network loss curve, black line shows generator network loss curve. Unlike most neural networks, convergence of loss for GANs is not sufficient to indicate that training is complete.

Figure S3



Supplementary Figure S3. Gene ontology enrichment for predicted differentiation genes Gene ontology enrichment for genes that generally decrease (upper) or increase (lower) over differentiation as predicted by latent space interpolation.

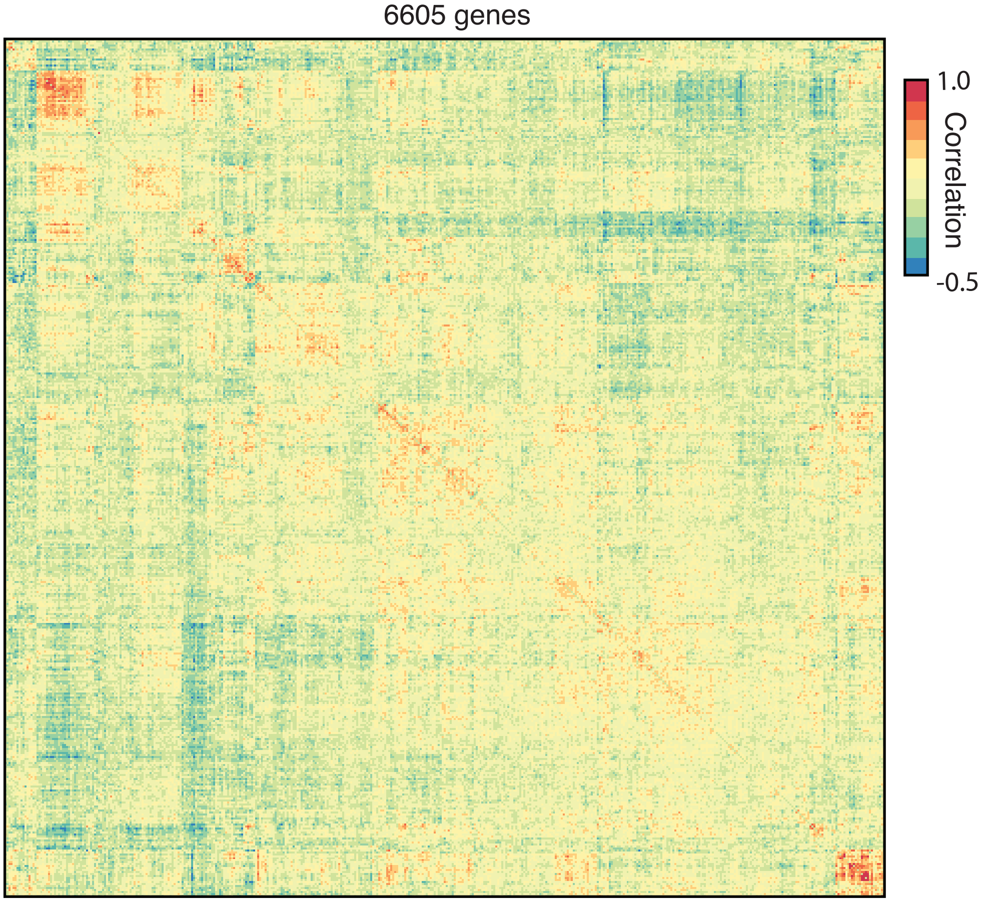
Figure S4



Supplementary Figure S4. Lars2 expression simulated over differentiation

Lars2 expression over differentiation for 30 generated single cells. Red lines are 1000 simulated differentiation timepoints. Black line represents mean expression over all cells.

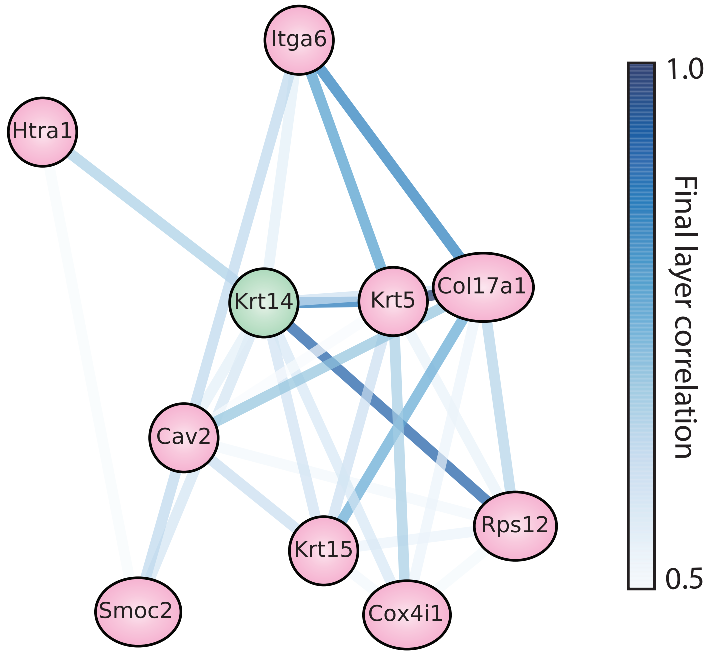
Figure S5



Supplementary Figure S5.

Correlation matrix of the generator neural network final layer weights visualised using a heatmap (Pearson correlation).

Figure S6



Supplementary Figure S6.

Local gene association networks for Krt14 visualised using a Fruchterman-Reingold force-directed network. Edges are colored by generator final layer correlation.

Supplementary Table 1. Genes identified by sensitivity analysis.

List of genes ranked by maximum discriminator network sensitivity gradient.
Related to Figure 5B

Supplementary Table 2. List of GAN neural network parameters.