Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis

Russkikh N.[1,2], Antonets D.[1,2,3]*, Shtokalo D.[1,2,4], Makarov A.[1], Zakharov A.[4] and Terentyev E.[4]

[1]AcademGene LLC, Novosibirsk, Russia,

[2]A.P.Ershov institute of informatics systems SB RAS, Novosibirsk, Russia,

[3]State Research Center of Virology and Biotechnology "Vector" Rospotrebnadzor, Koltsovo, Russia,

[4]Cancer Research Foundation, Moscow, Russia.

*Corresponding author: antonec@yandex.ru

Key words: biomarkers, NGS, single cell RNA sequencing, deep learning, autoencoders, disentanglement, harmonization, mouse.

ABSTRACT

The transcriptomic data is being frequently used in the research of biomarker genes of different diseases and biological states. Generally, researchers have data from hundreds, rarely thousands of specimens at hand. In most cases, the proposed candidate biomarker genes and corresponding decision rules fail in prospective research studies, especially for diseases with complex polygenic background. The naive addition of training data usually also does not improve performance due to batch effects, resulting from various discrepancies between different datasets. To get a better understanding of factors underlying the observed gene expression data variation, we applied a style transfer technique. The most of style transfer studies are focused on image data, and, to our knowledge, this is the first attempt to adapt this procedure to gene expression domain. As a style component, there might be used either technical factors of data variance, such as sequencing platform, RNA extraction protocols, or any biological details about the samples which we would like to control (gender, biological state, treatment etc.). The proposed solution is based on Variational Autoencoder artificial neural network. To disentangle the style components, we trained the encoder with discriminator in an adversarial manner. This approach can be useful for both data harmonization and data augmentation – for obtaining semisynthetic samples when the real data is scarce. We demonstrated the applicability of our framework using single cell RNA-Seq data from Mouse Cell Atlas, where we were able to transfer the mammary gland biological state (virgin, pregnancy and involution state)

between the samples with semantics (cell types) being preserved and with biologically relevant gene expression changes.

BACKGROUND

The new era of modern life sciences has begun with the development of high throughput nucleic acid sequencing methods – new generation sequencing (NGS) techniques. The amount of current genomic and transcriptomic data is tremendous and grows exponentially. The single cell sequencing methods enabled even more detailed description of a transcriptomic landscape that allowed to decipher the very complex nature of cellular subtypes, to analyze their developmental patterns and ancestry [1], [2].

However, current NGS data is highly fragmented due to different sources of technical variation associated with particular NGS platforms, sample acquisition and preparation procedures, subsequent analysis steps etc. The costs of transcriptomic experiments are still high and thus the really big datasets comprising thousands of samples are still rare. One of the most frequent tasks in transcriptomic data analysis is the identification of potential biomarker genes for various diseases and conditions. In most cases the researchers operate with data comprising from tens to hundreds, and, in rare cases, thousands of samples and the tens of thousands of genes or individual transcripts. The extremely high dimensionality and complex mutual interdependencies make it difficult to achieve the reproducibility in prospective studies. The problem is the insufficient volume of any single training dataset, an excessively large number of influencing factors and the lack of knowledge about the structure of molecular genetic systems. Thus, there is an urgent need in methodological approaches capable to analyze heterogenous and limited datasets of high dimensionality, suffering from technical noise and different kinds of batch effects. One of the available options is to harmonize the quality control procedures and the data analysis pipeline to make the resulting gene (transcripts) expression values more comparable. One of the best examples of this approach is DEE2 – Digital Expression Explorer 2 (DEE2) [3] – a repository of uniformly processed RNA-seq data obtained from NCBI Short Read Archive. There are also other examples: ARCHS4 the massive collection of uniformly processed murine and human public transcriptomic datasets [4], recount2 [5] etc. However, the most important task in transcriptomic data harmonization is the correction of batch effects and in general it remains unresolved.

Currently it is widely accepted that gene expression profiles of the living cells resulted from a complex mixture of different biological processes and technical parameters. At the moment, there were several attempts to model this kind of data as combinations of certain low-dimensional representations corresponding to various biological pathways and conditions [6]. In this work we test the hypothesis whether these attributes could be

reasonably and controllably changed *in silico* using the deep learning models. This type of transformation was mostly applied to image data to adopt the style of fine art paintings to generic images [7]. A large group of image style transfer frameworks uses the pretrained models to extract image descriptors in order to build the transfer objective. Due to absence of such pretrained models for gene expression data, we stick to adversarial approach: separating the features into style and semantic groups using the discriminator network. Learning the representations independent of domain with the help of discriminator using gradient reversal layer was proposed in [7]. The adversarial decomposition strategy was successfully applied to style transfer of texts, for example in this work [8]. In our work, we also used cycle-consistency loss for style transfer, which was proposed in [9]. The same technique used in domain adaptation for image segmentation tasks can be found in [10].

In this work we applied adversarial decomposition methodology to disentangle the biological and technical sources of variation in single cell RNA-Seq data. In our approach we used no prior dimensionality reduction as it makes strong assumptions about the data. For example, PCA tries to maximize the variance in projected dimensions and in case of heavy outliers and non-symmetric distribution the result becomes unstable at least if one doesn't apply the robust covariance estimates. Another common problem is that top PCs often extract the technical variation. Besides, we assume it's unlikely that biological states can be modelled by simple linear combinations of some low dimensional basis vectors since different sorts of non-linear relations are common for gene regulation circuits e.g., logical XOR patterns, various feedback loops and conditional dependencies etc. Given the highly hierarchical modular organization of cellular regulatory pathways and the clonal nature of the cells, deep neural network-based approaches seem to be the most feasible for the tasks involving gene expression. For example, an approach with deep generative modeling for scRNA-Seq data normalization and domain adaptation was recently proposed in [6]. Another approach to gene expression data modelling with autoencoders was presented in [11] – the authors induced the sparsity of network weights by connecting only the genes from the same functional group to the same hidden neuron. This is a step towards interpretability of autoencoder models. Authors of [12] and [13] successfully used variational autoencoders (VAE) as a non-linear dimensionality reduction method for gene expression data from different cancer subtypes and cell types, respectively. In [14] the autoencoder with ZINB-likelihood loss was effectively used as a denoising tool on gene expression data. Inspired by these results we decided to study if different components of gene expression data variance can be disentangled with adversarial decomposition methodology and if such disentangled representation might be of interest from a biological point of view.

MATERIALS AND METHODS

Dataset. The data was taken from Murine Cell Atlas (scMCA). This dataset comprising numerous murine single cell gene expression profiles was produced with cost-effective high throughput Microwell-seq platform [15], that allowed to analyze over 400,000 single cells from 51 mouse tissues and organs extracted from several animals at varying physiological conditions. The original scMCA data contains gene expression profiles for over 800 major murine cell types. The detailed annotation was provided by the authors for over 200,000 single cells. A detailed description of the data can be found in the original paper [15] and online [16]. This dataset was selected due to the following major reasons: (1) it contained the huge amount of data obtained with a consistent methodology by the same research group thus presumably making the technical dispersion less profound; (2) since the samples belong to different animals, distinct organs/tissues and physiological conditions one could build a model to decompose these sources of variation.

For building the models we selected a subset of 45497 samples corresponding to single cells derived from murine mammary glands of virgin and pregnant mice and also from involution state (24395, 9737 и 11365 samples respectively). This subset was selected both due to its volume and the ease of interpretation of distinctions between conditions. The samples from lactating animals were excluded as they were clearly isolated (data not shown) and different cellular types were barely distinguishable. We kept 20% of the data (with keeping the same proportion of biological states) as a test set, and 15% of the remaining data was used for validation. The original raw gene expression counts were used as inputs. 15987 genes were taken into consideration. The gene expression tables used in our study can be found in supplementary tables (ST1). To reduce the cell type labelling complexity, we decided to switch to more general cell types categories, presuming that the expression patterns between the cells of common origin should be more consistent than those of different cellular types. The complete data annotation used in our experiments is listed in supplementary table ST2. We considered the cellular types as the element of data semantics. Among the major goals was to control the preservation of related cell types under the style transfer procedure.

Autoencoder architecture. We use beta-VAE [17] (with beta=0.0003) as a backbone for our encoder-decoder architecture. Beta-VAE is a simple modification of vanilla VAE with additional hyperparameter aimed to weight a contribution of Kullback-Leibler divergence with prior distribution to the total loss. We train our encoder with discriminator in

an adversarial fashion in order to eliminate the information about one categorical variable (namely, biological state). To make the decoder be able to reconstruct the initial expression with absence of this information in the latent variables, we add learned representation of one-hot encoding of the category to the latents before feeding it to the decoder. This kind of architecture makes us able to perform style transfer: after encoding of the initial expression, we can choose a target category before decoding. We use LeakyReLU nonlinearities and batch normalization in the encoder layers. The architecture scheme is presented on Fig. 1. Discriminator scheme is FC(1024)-BatchNorm-LeakyReLU-FC(1024)-BatchNorm-LeakyReLU-FC(3).

Autoencoder training. For the training of our autoencoder, we use mean squared error (MSE) as a reconstruction loss function. Moreover, a cyclic consistency loss (with weight 0.2) is used: we obtain the encodings for a batch, make a random style transfer, and then transfer the style back at the second forward-pass through the autoencoder. Reconstruction loss between the values obtained this way and the initial expression is a cycle consistency loss. In order to enforce the hidden representation to not contain any information about biological state, we maximize Shannon entropy of discriminator predictions as generator loss. This adversarial loss contributes to the overall loss with weight 0.07. Discriminator is trained with log-loss objective.

For the regularization we use L2 weight penalty with weight of 0.001 for autoencoder along with VAE-regularization. For adversarial training stabilization, we have used gaussian instance noise [18] with variance 0.01 for discriminator.

Autoencoder and discriminator were trained for 1000 epochs with batch size of 128 with RAdam optimizer [19] with learning rate 0.0001, and per-epoch learning rate decay factor of 0.992. Also, clipping the gradient down to unite norm was used.

For the downstream analysis of autoencoder outputs, we substitute the predicted negative values with zero. Several experiments with ReLU activation used as the last layer to prevent the appearance of negative outputs were conducted, but these led to poor model convergence.

MA-plots construction. Each point on the MA-plot is a gene. Sum of expression of each gene was calculated across all samples belonging to the particular cell type in the same state and 1.0 was added to avoid division by zero problem. The abscissa is calculated as an average of log2-transformed expression of a gene in two compared states. The ordinate is the log2 transformation of the fold change of expression between two compared states.

Differential gene expression and gene set enrichment analysis. Differential gene expression analysis was performed using RPM-normalized expression counts. The statistical significance was assessed with Mann-Whitney test with multiple testing p-value correction using FDR procedure. Several cellular types were processed separately: (1) Stromal/Luminal/Alveolar cells – those functionally involved in mammary gland development and lactation and (2) Dendritic cells – antigen presenting cells that were expected to display less profound differences between virgin, pregnant and involution states. GO- and KEGG-enrichment analysis were performed with the online resource ShinyGO (v0.60) [20]. The lists of murine genes, associated with certain GO-categories were taken from Gene Ontology Browser at Mouse Genome Informatic portal [21].

RESULTS

Our research was aimed to disentangle the information about the cell type and biological state in the low-dimensional representation of gene expression data. While information about biological state is determined by its one-hot encoding which is fed to decoder, to keep the representation disentangled, we must eliminate this information from the latent variables obtained with the encoder. To achieve this, we used the dedicated discriminator, and the worse was its performance on the testing set, the better was the obtained disentanglement (minimum 33% since we have three classes).

It turns out that after training, the discriminator accuracy on the testing set was 63.2% with the following confusion matrix:

	Involution	**Pregnancy**	**Virgin**
Involution	660	248	325
Pregnancy	16	964	225
Virgin	73	482	719

Table 1. The confusion matrix of biological state prediction on the test dataset.

The disentanglement can be also illustrated with the following examples. Fig. 2 and Fig. 3 depict the 2D projections of the testing samples obtained with tSNE using either the original gene expression values or the recovered expression obtained with our model, respectively. The samples are colored according to cell types (A) and to physiological states (B). One can readily see the clusters corresponding to cell types and to biological states on both these plots. However, when similar visualization was built using the extracted latent

representations of the samples as input (Fig. 4), there were no clusters corresponding to different physiological states, but the clusterization of cell types was still observed.

Style transfer validation. In order to validate the style transfer, we train neural network classifier to predict the biological state on the raw expression from the train set, transfer the style of all test examples in all possible ways (including keeping the original style, therefore obtaining 3 times larger test set, because we are considering three styles) and evaluate the accuracy of classifier prediction on this set. As the ground truth in this experiment we take a style in which the sample was transferred. The architecture was the following: Input(200)-FC(512)-FC(256)-Output(3) with LeakyReLU nonlinearities; we used Adam optimizer with learning rate of 0.0003, batch size of 2 and 10 epochs. High accuracy of such classification points that synthetic samples of some category share category-defining features with its raw counterpart. The prediction accuracy was 89.2% with the following confusion matrix:

	Involution	**Pregnancy**	**Virgin**
Involution	3365	114	233
Pregnancy	266	3433	13
Virgin	459	111	3142

Table 2. The confusion matrix of biological state prediction on the testing dataset, transformed with autoencoder into all possible states with style transfer procedure.

The detailed description of the training procedure can be found at project home on Github.

Calibration procedure. Yet another, simpler approach to validate our model is what we call a *calibration procedure*. It is designed to control that keeping the original sample style while passing the sample through the model provides less deviation of expression than an arbitrary style transfer. Namely, we take a sample, transfer its style in all possible ways and check if L2-distance between the original and decoded expression achieves the smallest value when the initial sample style is used. Turns out that it's the case for 78.4% of the test samples and 93.6% of training samples. Worth mentioning that in the testing set the average difference between the best and the second-best category is 0.008 for correctly calibrated

samples and eight times lower (0.001) for incorrectly calibrated ones. It means that switching to another style does not affect such samples that much.

Preservation of cell types. In order to demonstrate the preservation of semantics, we used cell types annotation. We expect that the cell type must be almost invariant to style transfer. To demonstrate this, again, we trained a neural network classifier to recognize the cell type on the raw expression from the training set and evaluated it on a transferred test set expression. The architecture was the following: Input(200)-FC(512)-FC(256)-Output(13) with LeakyReLU nonlinearities; we used Adam optimizer with learning rate of 0.00001, batch size of 2 and 1 epoch. The accuracy equals to 77% for samples which passed through the autoencoder with no style transfer (and equals to 86,5% for the raw test set data taken as-is, with no autoencoder involved). So that, accuracy drop is 9% while we have reduced dimensionality in more than 70 times, from 15987 to 210. For transferred samples, the accuracy decreased to 59.5%. This value is fairly low due to severe class imbalance of different cell types among different biological states in the training set. Namely, for Pregnancy state there are 1652 Epithelial cells, which is a major class, but in Involution state they are totally absent. After discarding such cases form evaluation, the accuracy increased to 74.0%. The accuracy of a trivial baseline (always predicting the major class) is 24.2%.

Biological examination of gene expression changes after encoding and decoding transformation. The verification was performed using differential gene expression analysis and gene set enrichment analysis with GO and KEGG categories. Differential gene expression analysis was performed using RPM-normalized expression counts. The statistical significance was assessed with Mann-Whitney test with multiple testing p-value correction using FDR procedure. Several cellular types were processed separately: (1) Stromal/Luminal/Alveolar cells – those functionally involved in mammary gland development and lactation and (2) Dendritic cells – antigen presenting cells that were expected to display less profound differences between virgin, pregnant and involution states.

GO-enrichment analysis of Stromal/Luminal/Alveolar cells of top 100 genes found to be differentially expressed in samples of virgin and pregnant mice				
Enrichment FDR	**Genes in list**	**Total genes**	**Functional Category**	**GO ID**
4.1E-06	22	1117	Epithelium development	GO:0060429
1.0E-05	16	621	Epithelial cell differentiation	GO:0030855
4.0E-05	37	3437	Animal organ development	GO:0048513
4.9E-04	24	1855	Tissue development	GO:0009888

5.3E-04	12	503	Morphogenesis of an epithelium	GO:0002009
5.3E-04	12	500	Gland development	GO:0048732
1.8E-03	7	167	Mammary gland development	GO:0030879
2.1E-03	10	404	Epithelial cell proliferation	GO:0050673
2.6E-03	39	4627	System development	GO:0048731
2.6E-03	2	2	Proximal/distal pattern formation involved in metanephric nephron development	GO:0072272
GO-enrichment analysis of Stromal/Luminal/Alveolar cells of top 100 genes found to be differentially expressed in samples of pregnant mice and animals with mammary gland involution				
Enrichment FDR	**Genes in list**	**Total genes**	**Functional Category**	**GO ID**
2.2E-04	36	3522	Response to stress	GO:0006950
2.2E-04	12	448	Response to oxidative stress	GO:0006979
2.2E-04	27	2100	Cell death	GO:0008219
2.2E-04	28	2378	Response to external stimulus	GO:0009605
2.2E-04	25	1949	Programmed cell death	GO:0012501
4.5E-04	24	1909	Apoptotic process	GO:0006915
4.6E-04	22	1654	Positive regulation of molecular function	GO:0044093
5.4E-04	26	2247	Catabolic process	GO:0009056
5.9E-04	24	1985	Cellular catabolic process	GO:0044248
6.5E-04	22	1728	Regulation of cell death	GO:0010941

Table 3. GO-enrichment analysis of top 100 differentially expressed genes observed in Stromal/Luminal/Alveolar cells in virgin vs. pregnant and involution vs. pregnant comparisons. The top 10 enriched categories are shown.

GO-enrichment analysis demonstrated that used data contained relevant biological signals (Table 3). When Stromal/Luminal/Alveolar cells taken from mammary glands of pregnant mice were compared against those of virgin mice, the top 100 upregulated differentially expressed genes were found to be significantly enriched with epithelium development, epithelial cell differentiation, mammary gland development GO categories. The top 200 upregulated genes were also found to be significantly associated with progesterone-mediated oocyte maturation and prolactin signaling KEGG pathways. The top 100 upregulated differentially expressed genes found with comparison of Stromal/Luminal/Alveolar cells from mice with mammary gland involution against those of pregnant animals were found to be significantly enriched with GO categories related to apoptosis, stress response and catabolism (Table 3). When similar analysis was performed

using dendritic cell samples, the top 100 differentially expressed upregulated genes were found to be significantly enriched with GO categories related to defense and immune responses, cytokine production, dendritic cell differentiation etc. (data not shown).

Besides the examination of original expression profiles, we also made a comparison between the samples after the "style transfer" procedure: when samples of pregnant mice were transformed into a virgin or involution state, virgin – to pregnant or involution, involution – to pregnant or virgin. As an example, Fig. 5 shows MA-plots with comparison of Virgin versus Pregnant states of stromal cells (shown with blue dots), and Virgin versus artificial Virgin state created from Pregnant by style transfer (shown with orange). The overlay of these MA-plots provides a clear illustration that gene expression of original Virgin state is closer to that of artificially obtained Virgin than to original Pregnant samples.

However, the similar GO- and KEGG-enrichment analysis of recovered gene expression and semisynthetic samples obtained with style transfer was less straightforward since there were numerous changes associated with basic GO categories. Thus, we decided to compare the variation of gene expression associated with relevant GO categories: mammary gland development (GO:0030879) and positive regulation of apoptotic process (GO:0043065). The highest variance in expression of genes involved in mammary gland development was observed in samples from pregnant mice (Fig. 6). The similar results were observed both in stromal and luminal cells (A) and also with using all the cells (B). The recovered expression was similar to original values, but the most interesting is that the style transfer from Virgin state to Involution and Pregnancy and from Involution to Virgin and Pregnancy resulted in biologically relevant changes in gene expression (the two lower panels of Fig. 6A and Fig. 6B). The similar analysis of genes involved in apoptosis regulation revealed two different pictures (Fig. 7). When only stromal, luminal and alveolar cells were considered the maximal variance was observed in samples from pregnant mice, and the second-high values were observed in samples from virgin mice (Fig. 7A, the top panels), however when all the cell types were considered the maximal variance was observed in samples from involuting mammary gland – as it was expected (Fig. 7B, the top panels). However, the results of style transfer (Fig. 7, the bottom panels) also demonstrate that the variance in apoptosis-related genes is higher in Involution state. The contradiction observed when only stromal, luminal and alveolar cells were considered might be due to the striking differences in proportions of various cell types. Thus, from here we can propose the additional advantage of style transfer procedure as it might be of help in studying gene expression changes resulted from certain biological or technical traits using the same initial data and treating the resulting samples as paired data.

DISCUSSION

Construction of information-rich, low-dimensional representations of gene expression profiles, remains a challenging task. Availability of such representations is a gateway to successful data harmonization, domain adaptation and deeper understanding of interconnections between expression of various genes. The proposed framework allows to investigate gene expression profile shifts when some specific, pre-defined categorical factor of variation changes. The framework performs dimensionality reduction of gene expression data in such a way that hidden variables are disentangled into two separate domains where one subgroup is fully interpretable and accounted for chosen, pre-defined factors of variation and another, larger group of hidden variables is designed to contain no information from controlled factors. So that, we can controllably change the factor(s) and see the impact on the gene expression level. This leads us to possibilities to harmonize the data by using batch codes, or sequencer model as factors of style and to perform the downstream analysis on the latent variables (which also dramatically reduces the dimensionality, and therefore helps to control overfitting) instead of raw expression, or transferring all of the samples to the same style. Ability to observe gene expression of synthetic samples makes possible their analysis with classical bioinformatic approaches, for example to check which genes show differential expression when you switch the technical factors of variation with style transfer.

The proposed approach can help to solve different problems associated with real transcriptomic data, e.g., to reduce the variance associated with batch effects, to check the data for outliers, to reduce the data dimensionality retaining the relevant biological information. Generative adversarial neural networks could also be used for data augmentation, and, which is of particular interest, with a style transfer approach one can generate realistic examples to upsample the rare cases or even to produce the cases lacked in the current data.

Our future efforts on the framework will be mostly conducted towards increasing the fidelity of the generated samples and evaluating our approach on different datasets and comparing its performance with the existing frameworks. Also, future research will include the induction of sparsity in both encoder and decoder weights to figure out its effect on performance of the framework in terms of disentanglement and applicability of generated samples to downstream pipelines. Moreover, the sparse weights promise some insights on what genes affect each other expression and affected by choice of the style.

SUPPLEMENTARY DATA

Supplementary tables with data and annotation can be found online:

DOI: 10.6084/m9.figshare.9925115

This data contains the following tables:

ST1 – original_expression.csv – the selected subset of original raw gene expression counts from scRNA-Seq experiments from DOI: 10.1016/j.cell.2018.02.001 (https://figshare.com/articles/MCA_DGE_Data/5435866) – only mammary gland samples were used;

ST2 – transfer_annotation.csv – samples annotation table (both original and those obtained with style transfer)

ST3 – reconstructed_expression.csv – gene expression values of original samples obtained with the developed VAE model;

ST4 – transferred_expression.csv – semisynthetic samples obtained with the developed style transfer procedure.

The source code can be found at: https://github.com/NRshka/stvae-source

REFERENCES

[1] A.-E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, “Single-cell RNA-seq: advances and future challenges,” *Nucleic Acids Res.*, vol. 42, no. 14, pp. 8845–8860, Aug. 2014.

[2] R. Stark, M. Grzelak, and J. Hadfield, “RNA sequencing: the teenage years,” *Nat. Rev. Genet.*, Jul. 2019.

[3] M. Ziemann, A. Kaspi, and A. El-Osta, “Digital expression explorer 2: a repository of uniformly processed RNA sequencing data,” *Gigascience*, vol. 8, no. 4, Apr. 2019.

[4] A. Lachmann *et al.*, “Massive mining of publicly available RNA-seq data from human and mouse.,” *Nat. Commun.*, vol. 9, no. 1, p. 1366, 2018.

[5] L. Collado-Torres, A. Nellore, and A. E. Jaffe, “recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor,” *F1000Research*, vol. 6, p. 1558, Aug. 2017.

[6] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, “Harmonization and Annotation of Single-cell Transcriptomics data with Deep Generative Models,” *bioRxiv*, p. 532895, Jan. 2019.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge, “A Neural Algorithm of Artistic Style,” Aug. 2015.

[8] A. Romanov, A. Rumshisky, A. Rogers, and D. Donahue, “Adversarial Decomposition of Text Representation,” Aug. 2018.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” Mar. 2017.

[10] J. Hoffman *et al.*, “CyCADA: Cycle-Consistent Adversarial Domain Adaptation,” Nov. 2017.

[11] M. P. Gold, A. LeNail, and E. Fraenkel, “Shallow Sparsely-Connected Autoencoders for Gene Set Projection,” in *Biocomputing 2019*, 2018, pp. 374–385.

[12] G. P. Way and C. S. Greene, “Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders,” *Biocomput. 2018*, pp. 80–91, 2018.

[13] C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, and O. Winther, “scVAE: Variational auto-encoders for single-cell gene expression data,” *bioRxiv*, p. 318295, 2018.

[14] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, “Single-cell RNA-seq denoising using a deep count autoencoder,” *Nat. Commun.*, vol. 10, no. 1, p. 390, Dec. 2019.

[15] X. Han *et al.*, “Mapping the Mouse Cell Atlas by Microwell-Seq,” *Cell*, vol. 172, no. 5, pp. 1091-1107.e17, Feb. 2018.

[16] G. Guo, “MCA DGE Data.” 22-Oct-2018.

[17] I. Higgins *et al.*, “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework,” in *ICLR*, 2017.

[18] L. M. Mescheder, “On the convergence properties of {GAN} training,” *CoRR*, vol. abs/1801.04406, 2018.

[19] L. Liu *et al.*, “On the Variance of the Adaptive Learning Rate and Beyond,” Aug. 2019.

[20] S. X. Ge and D. Jung, “ShinyGO: a graphical enrichment tool for ani-mals and plants,” *bioRxiv*, p. 315150, Jan. 2018.

[21] C. J. Bult *et al.*, “Mouse Genome Database (MGD) 2019,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D801–D806, Jan. 2019.

FIGURE LEGENDS

Fig. 1. The autoencoder architecture scheme.

Fig. 2. Visualization of original samples with tSNE. Raw expression values were used, samples were colored according to cell types (A) and physiological state (B). tSNE perplexity was set to 30.

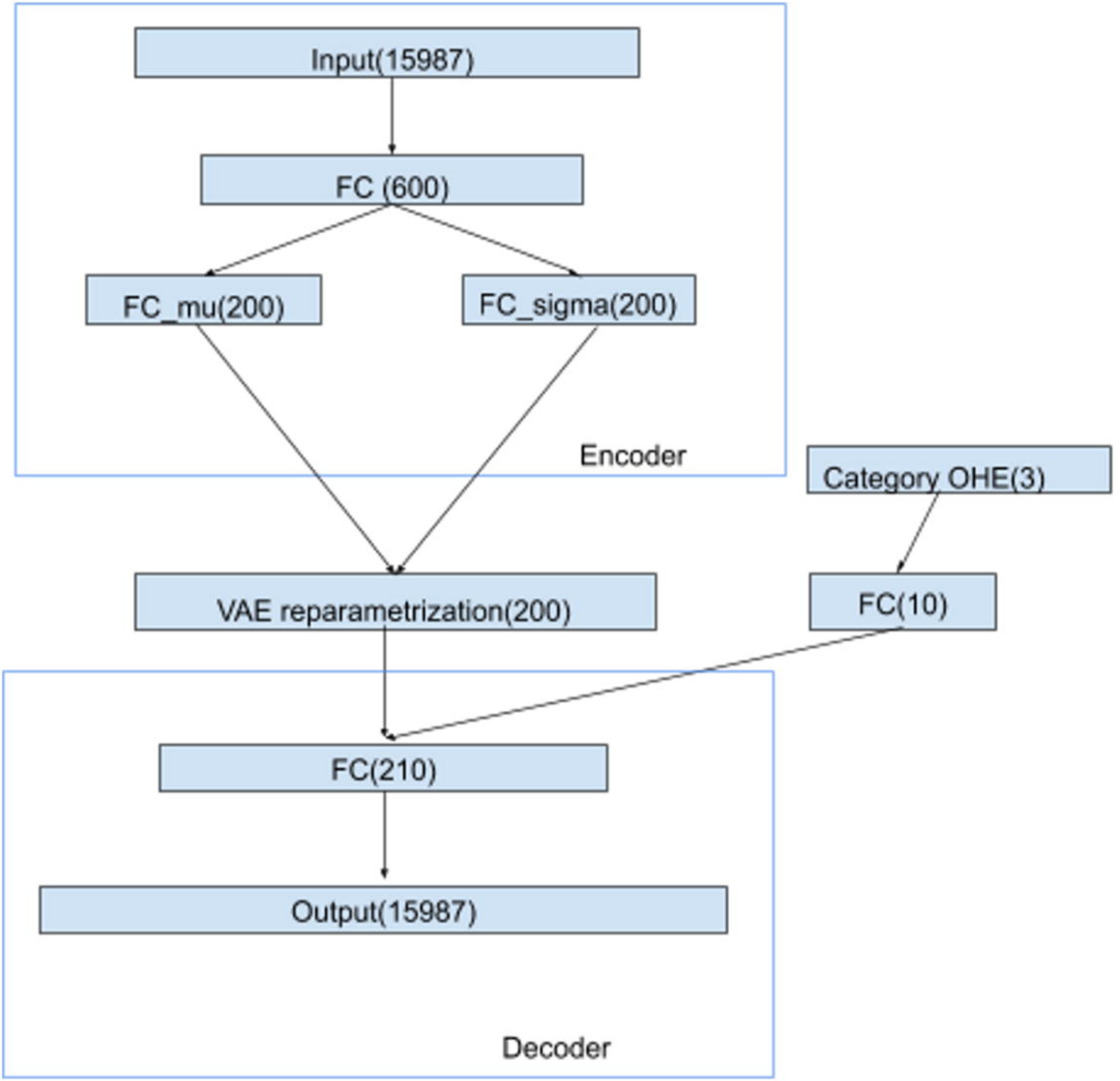
Fig. 3. Visualization of reconstructed samples with tSNE. Gene expression values reconstructed with VAE model were used, samples were colored according to cell types (A) and physiological state (B). tSNE perplexity was set to 30.

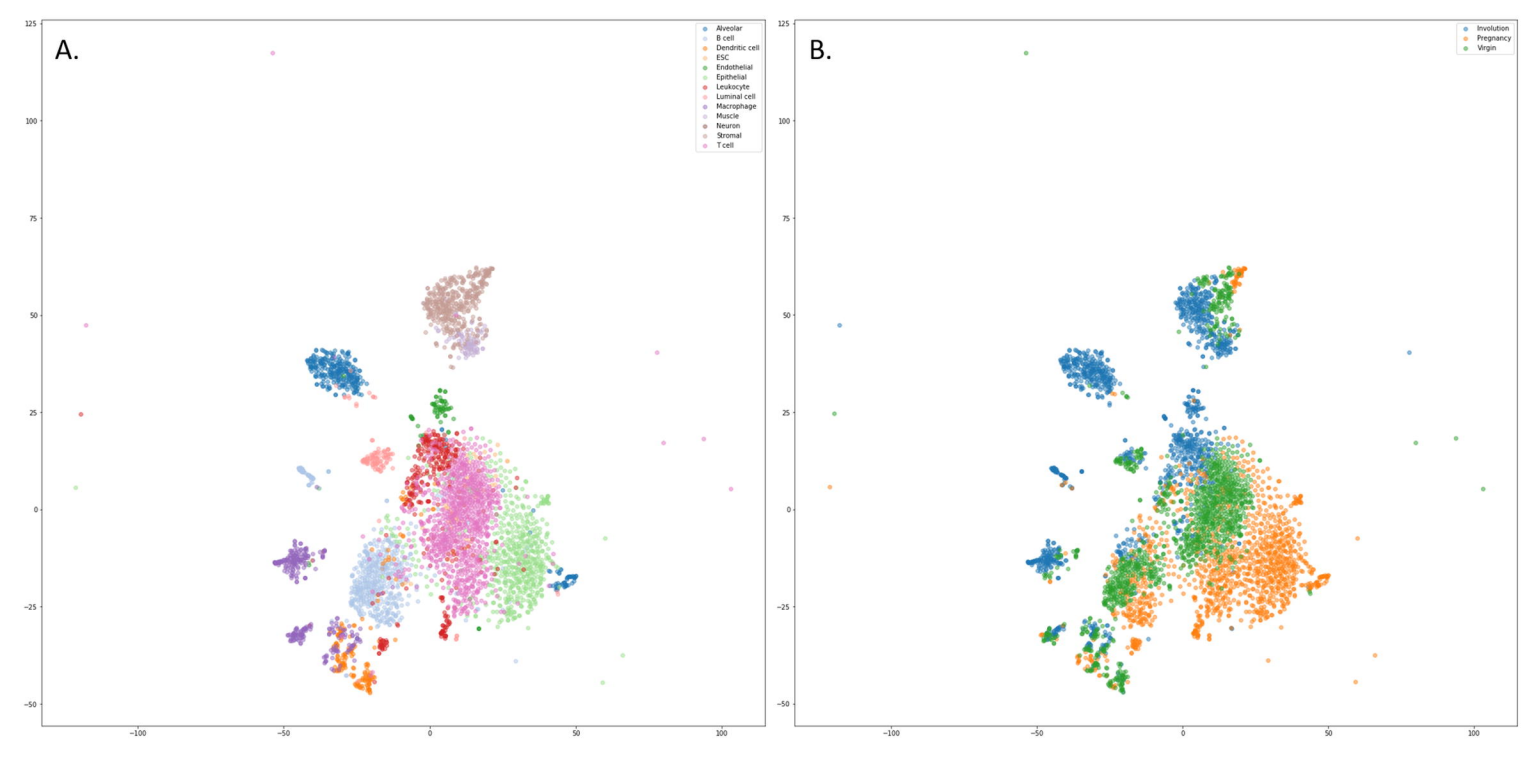
Fig. 4. Visualization of the samples with tSNE using the learned latent representation. The latent variables of the testing samples were obtained with pre-trained encoder. The samples were colored according to cell types (A) and physiological state (B). tSNE perplexity was set to 30.

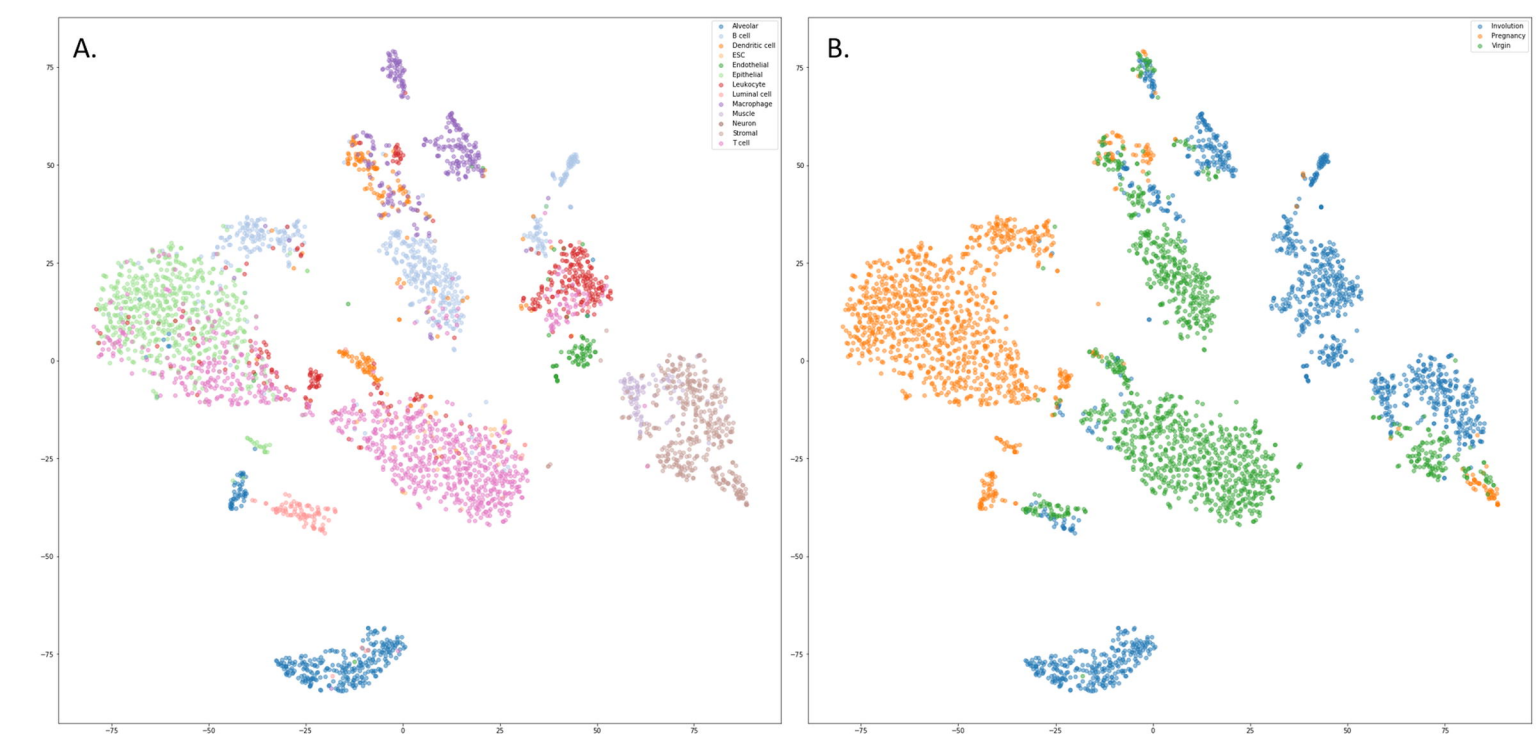
Fig. 5. MA-plots comparing the gene expression in stromal cells from murine mammary glands in original and transformed samples. The comparison of original samples is shown with blue; the comparison between the original virgin state and the virgin state produced from pregnancy with style transfer is shown with orange color.

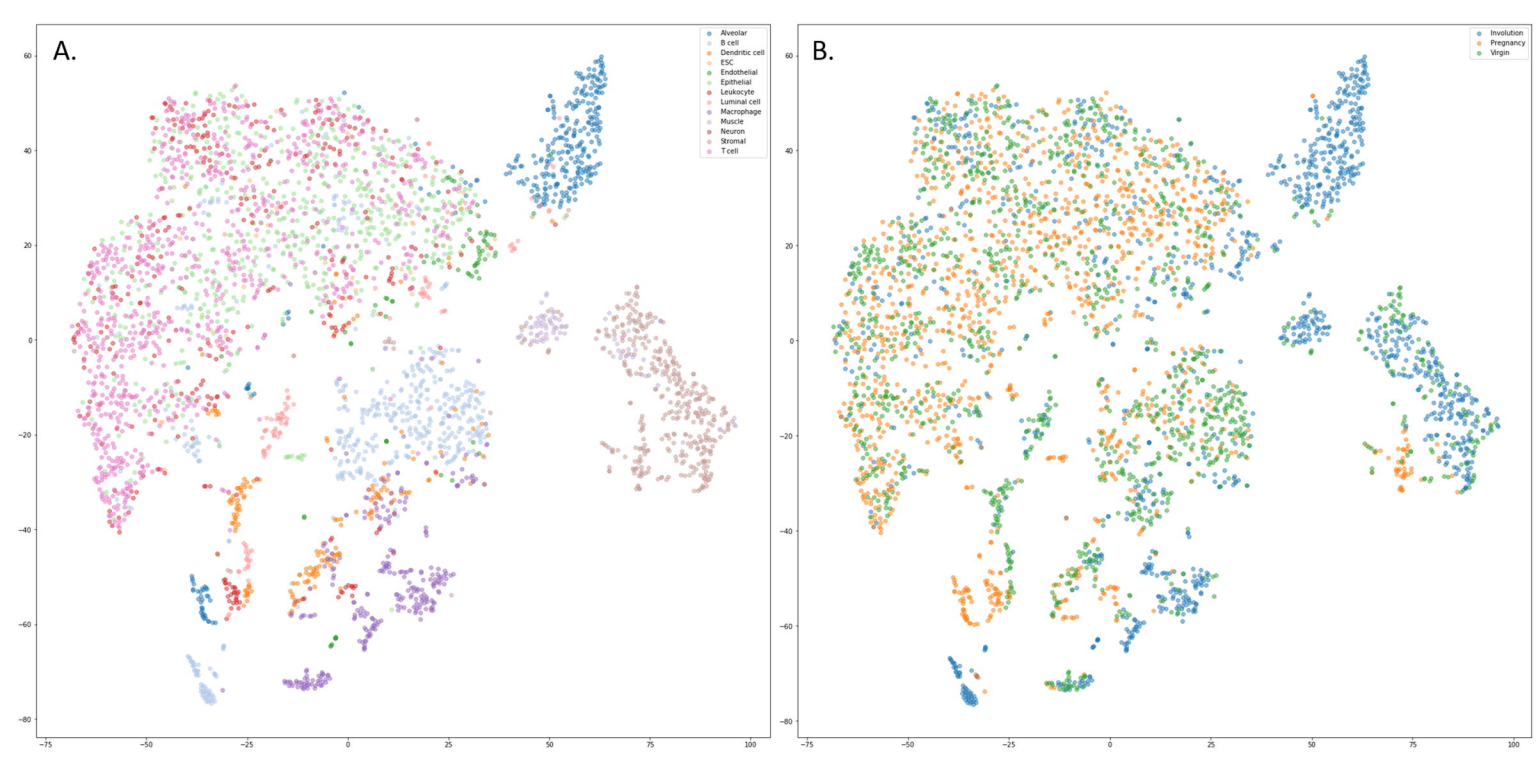
Fig. 6. Variation in gene expression related to mammary gland development (GO:0030879) in Stromal and Luminal cells (A) and in all cells (B).

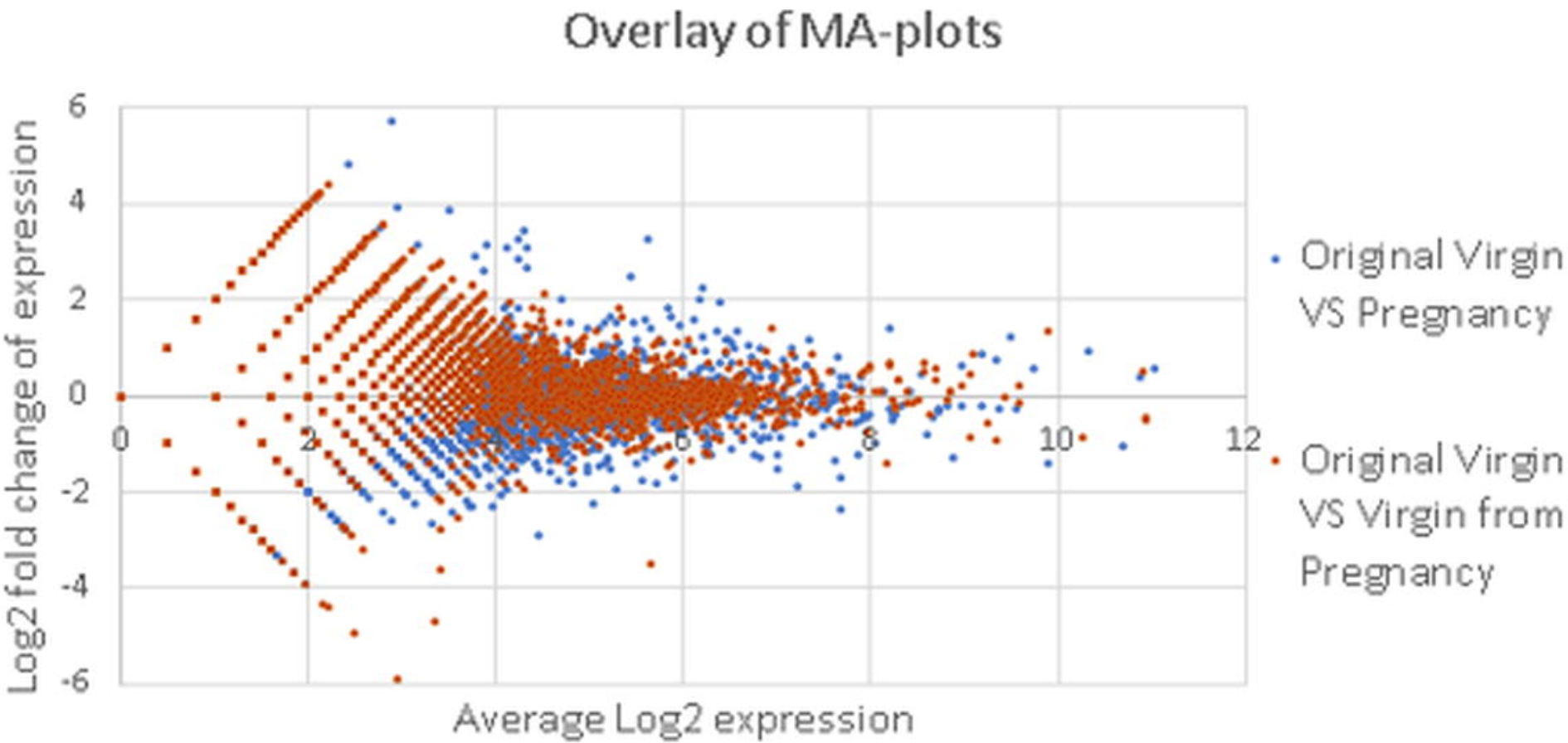
Fig. 7. Variation in gene expression related to positive regulation of apoptotic process (GO:0043065) in Stromal, Luminal and Alveolar cells (A) and in all cells (B).

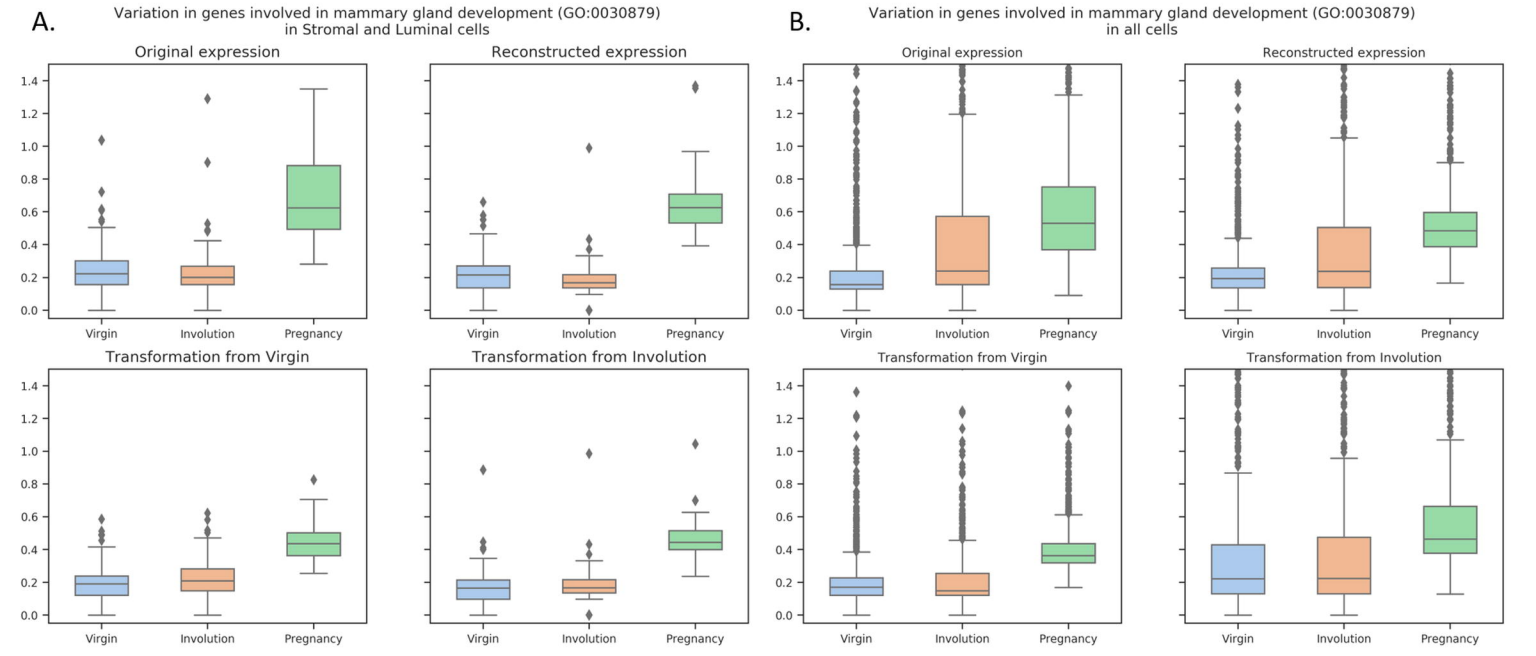

Input(15987)
FC (600)
FC_mu(200)
FC_sigma(200)
Encoder
Category OHE(3)
VAE reparametrization(200)
FC(10)
FC(210)
Output(15987)
Decoder

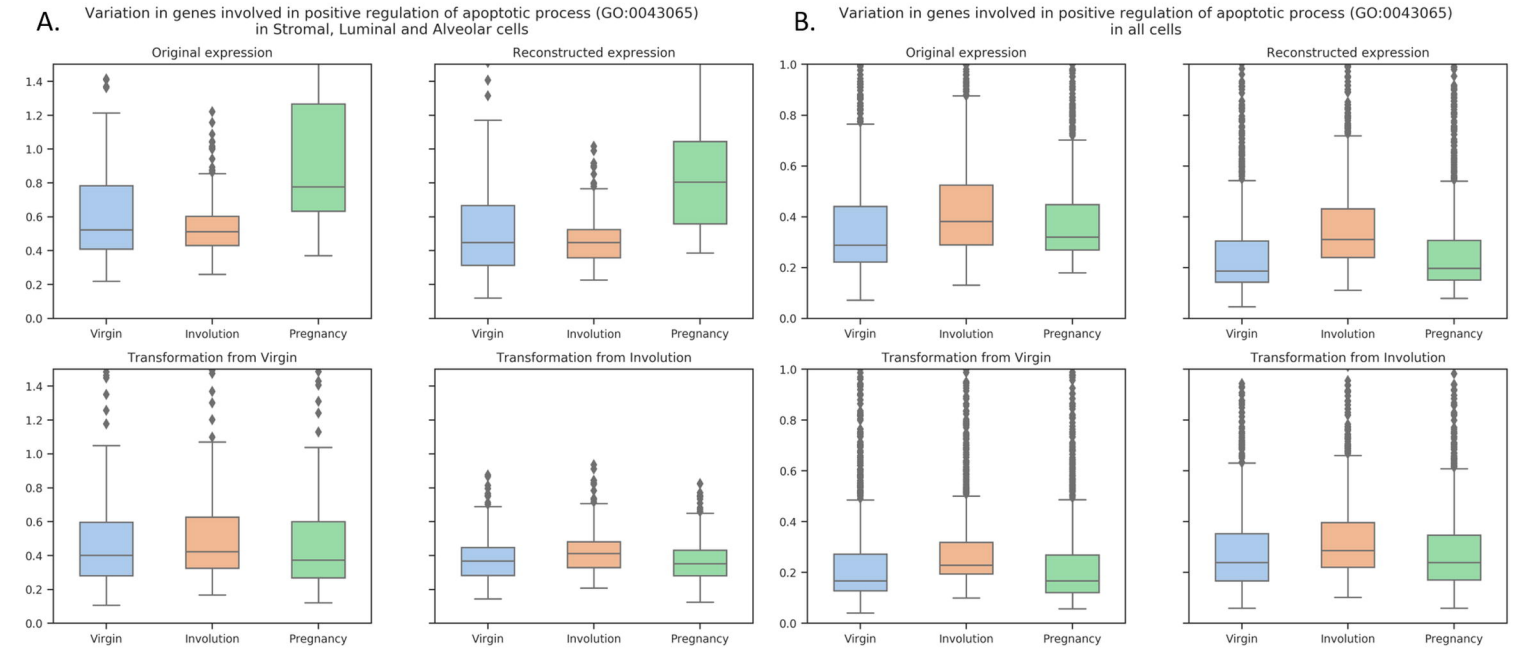

A.
Alveolar
B cell
Dendritic cell
ESC
Endothelial
Epithelial
Leukocyte
Luminal cell
Macrophage
Muscle
Neuron
Stromal
T cell
125
100
75
50
25
0
−25
−50
−100
−50
0
50
100
B.
Involution
Pregnancy
Virgin
125
100
75
50
25
0
−25
−50
−100
−50
0
50
100


A.
Alveolar
B cell
Dendritic cell
ESC
Endothelial
Epithelial
Leukocyte
Luminal cell
Macrophage
Muscle
Neuron
Stromal
T cell
B.
Involution
Pregnancy
Virgin


A.
Alveolar
B cell
Dendritic cell
ESC
Endothelial
Epithelial
Leukocyte
Luminal cell
Macrophage
Muscle
Neuron
Stromal
T cell
B.
Involution
Pregnancy
Virgin


Overlay of MA-plots
Log2 fold change of expression
6
4
2
0
-2
-4
-6
0
2
4
6
8
10
12
Average Log2 expression
Original Virgin VS Pregnancy
Original Virgin VS Virgin from Pregnancy


A.
Variation in genes involved in mammary gland development (GO:0030879)
in Stromal and Luminal cells
Original expression
Reconstructed expression
Transformation from Virgin
Transformation from Involution
B.
Variation in genes involved in mammary gland development (GO:0030879)
in all cells
Original expression
Reconstructed expression
Transformation from Virgin
Transformation from Involution
Virgin
Involution
Pregnancy


A.
Variation in genes involved in positive regulation of apoptotic process (GO:0043065)
in Stromal, Luminal and Alveolar cells
Original expression
Reconstructed expression
Transformation from Virgin
Transformation from Involution
Virgin
Involution
Pregnancy
B.
Variation in genes involved in positive regulation of apoptotic process (GO:0043065)
in all cells
Original expression
Reconstructed expression
Transformation from Virgin
Transformation from Involution
Virgin
Involution
Pregnancy