UniPath: A uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles.

Smriti Chawla[1], Sudhagar Samydurai[2], Say Li Kong[2], Zhenxun Wang[2], Wai Leong TAM[2,3], Debarka SenGupta[1*] Vibhor Kumar[1,2*]
1. Department for Computational Biology, Indraprastha Institute of Information Technology, Delhi, 110020, India
2. Genome Institute of Singapore, Agency for Science Technology and Research, Singapore, Singapore
3. Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore

*** co-corresponding author**
Email: kumarv1@gis.a-star.edu.sg, vibhor@iiitd.ac.in, debarka@iiitd.ac.in

Abstract:
Here, we introduce UniPath, for representing single-cells using pathway and gene-set enrichment scores by transformation of their open-chromatin or expression profiles. Besides being robust to variability in drop-out, UniPath also provides consistency and scalability in estimating gene-set enrichment scores for every cell. UniPath also enables exploiting pathway continuum and dropping known covariate gene-sets for predicting temporal order of single-cells. Analyzing mouse cell atlas using pathway enrichment-scores revealed surprising but biologically-meaningful co-clustering of cell-types from distant organs and helped in annotating many unlabeled cells. By enabling unconventional analysis, UniPath also prooves to be useful in inferring context-specific regulation in cancer cells.

Introduction

Single-cell RNA sequencing (scRNA-seq) and single-cell open-chromatin profiling help us to decipher cellular heterogeneity of activity of coding and non-coding genomic elements[1, 2]. The heterogeneity in the activity of genomic sites among single-cells, is being regularly used to estimate cellular composition, finding rare cells and understanding the role of genes and transcription factors [2, 3]. However, new questions are being asked with an increase in throughput of scRNA-seq and single-cell open-chromatin profiling through ATAC-seq (single-cell assay for Transposase-Accessible Chromatin using sequencing). One such question is, how can we use single-cell transcriptome and epigenome profiles for new applications. Can single-cell epigenome and expression profile help in finding lineage potency of a cell? Can single-cell heterogeneity be used in choosing more specific target pathways for cancer therapeutics? The answers to such questions can be found by representing cell states with more abstract and biologically- meaningful terms to utilize heterogeneity among cells. Such as defining cell-state in terms of pathway activity scores could help us to have a meaningful perspective about its role and dynamic behaviour. However, most often enrichment of pathways is done using differential gene expression between two groups of cells and this procedure does not solve the purpose of studying heterogeneity of gene-set enrichment at single-cell resolution. Another category of methods like SVA[4], RUV[5], scLVM[3] and f-scLVM[6] provide relevance score for known and unknown dominating factors for a group of single-cells. Such methods are not meant to provide

enrichment and relevance of gene-sets in each single-cell like PAGODA[7]. However, PAGODA is not designed to handle scRNA-seq data from a non-heterogeneous collection of cells. The main hurdle in finding enriched pathways for each single-cell using scRNA-seq profile has been the default dependency on read-count data of genes. The statistical modelling of read-count of a genomic site across multiple cells is a non-trivial task, especially for single-cell open-chromatin and scRNA-seq profiles due to variability in drop-out rate and sequencing depth among cells[7, 8]. Moreover, there has been rarely any attempt to estimate pathway enrichment-scores for single-cells using their open-chromatin profiles for downstream analysis like classification and pseudo-temporal ordering. Hence, there is a need for a uniform method which can transform single-cell expression and open-chromatin profiles from both non-heterogeneous and heterogeneous samples to gene-set activity scores.

In this study, we have addressed the challenge of representing single-cells in terms of pathways and gene-set enrichment-scores estimated using scRNA-seq and open-chromatin profiles in spite of cell-to-cell variability in drop-out of genomic regions and sequencing depth. Unlike previously proposed methods for scRNA-seq profiles, we do not try to normalise or scale read-count of a gene across cells using parametric distributions like Poisson or negative binomial. Scaling read-count across cells with variable drop-out rate and sequencing depth increases chances of artefacts. Therefore, we use a common null model to estimate adjusted pathway enrichment scores while handling scRNA-seq profiles. Similarly, while using scATAC-seq profiles, we use the approach of highlighting enhancers by dividing read-counts of genomic sites with their global accessibility scores. We benchmarked our methods and null models for estimating single-cell gene-set enrichment using several published scRNA-seq and scATAC-seq datasets.

Using pathways and gene-set as features for single-cells creates new opportunities and challenges which we tried to explore further. Compared to raw read-counts, the pathway scores of single-cells are more likely to have less sparsity, noise and technical variation, which we exploited for classical procedures like classification and dimension reduction based visualisation. Next, we asked whether the temporal ordering of cells can be performed using gene-set enrichment scores as features since it can directly highlight the continuum of lineage potency with inflexion points defined by the activity of pathways. However, existing methods of temporal ordering for single-cell use read-count or gene-expression matrix where there is less flexibility to drop known covariate. Therefore, we develop and included a novel pseudo-temporal ordering method in UniPath which can use pathway scores and allow dropping gene-sets of known covariates. We applied UniPath on a large scRNA-seq data-set of mouse cell atlas (MCA) and performed clustering using pathway scores and annotated many unlabeled cells. We also explored the possibility of using enrichment and co-occurrence (co-enrichment) of pathways for understanding underlying regulation. While analysing scRNA-seq profile of differentiating hESC cells using pathway scores, we realized the strength of a new way of comparing different populations of cells. Therefore, we performed scRNA-seq for two cell lines of non-small cell lung cancer (NSCLC) and tried to understand the difference in their properties using the new way of comparison.

Results

For transforming scRNA-seq profiles to pathways score we treat each cell separately. Generally, in a single-cell, RPKM (read per Kilobase per million) or FPKM (fragment per Kilo per million) value of genes have a bimodal distribution, where one of the modes is around zero and other is for non-zero expression values (Supplementary Figure S1a). We used widely and theoretically accepted assumption that most of the time, non-zero RPKM and FPKM values within a sample

(or cell) follow log-normal distribution[9]. For a single-cell, we convert non-zero FPKM values of genes to P-values (right-tailed) using log-normal distribution. We apply Brown's method to combine P-values of genes in a gene-set to reduce the effect of covariation among genes. The combined P-value for every gene-set is adjusted using a null background model made using a systematic approach (see Method, Figure 1a). The objective of P-value adjustment using null model created by Monte-Carlo approach is to highlight cell-type-specific gene-set activity and reduce blurring due to background house-keeping function of cells. We call the adjusted P-value of a pathway (or gene-set) in a single-cell as its score.

Evaluation of UniPath's approach of transforming single-cell expression profiles to pathway enrichment scores

Due to the lack of gold standard, it is not trivial to assess gene-set enrichment methods for heterogeneous bulk samples. However, for single-cell from known cell-lines, the marker gene-set for cell-types can be used directly to test methods like UniPath. We used marker gene-sets for cell-types to compare our approach with an existing method PAGODA[7] on several data-sets. Systematic evaluation using scRNA-seq profiles from 10 studies (see Table S1) revealed that most of the time UniPath was better than PAGODA in terms of estimating enrichment of gene-sets for correct cell-types especially for the non-heterogenous collection of cells (Figure 2a, Supplementary Figure S1, Table S1)[10, 11]. The high accuracy of UniPath to highlight correct cell types also allows detection of doublets, which we confirmed using three simulated data-sets (supplementary Method). In our simulation-based tests, UniPath achieved accuracy of 64-72% for detecting doublets, which were substantially better than PAGODA (Supplementary Figure S2). Further, we made a collection of gene-sets of non-immune related pathway terms and as spike-in, we added two known B cell and T cell related pathway gene-sets (see Table S2). With this control experiment for both B cell[12] and T cell[13], UniPath revealed the correct respective pathway in top 5 enriched terms with substantially better accuracy than PAGODA (Figure 2b). We also assessed the consistency of enrichment of pathways by UniPath and PAGODA. We analyzed B cells (GM12878) scRNA-seq profile[10] while grouping them each time with different cell types. The scores for pathways and gene-sets from PAGODA were not consistent (see Figure 2c, Supplementary Figure S3) and for every cell the output was dependent on the composition of cell type in the data-set. However, UniPath based enrichment scores for a cell remains consistent and is not affected by other neighbouring cells (Figure 2c, Supplementary Figure S3). Thus, UniPath resolves the issues of highlighting correct gene-sets and relevant pathways with consistency for each single-cell irrespective of level of heterogeneity of cell-types in the provided scRNA-seq data.

Gene-set enrichment with UniPath as an alternative dimension-reduction method for single-cell ATAC-seq profile

For the transformation of open chromatin profile of single-cells to pathway enrichment scores, UniPath first highlights enhancers by normalizing read-count on peaks using their global accessibility scores (Figure 1b) (see Methods). The global accessibility scores have been independently calculated using available bulk open-chromatin profiles from many cell types (see Methods). The motivation behind normalising read-count of each peak using its global accessibility score is to have consistency and avoid adjustment of variability in sequencing depth and drop-out rate. For every cell, genomic sites with high normalized read-count are chosen as foreground set. Then, for every cell UniPath uses proximal genes of peaks in it's foreground set to estimate statistical significance (P-value) of enrichment gene-sets using Hypergeometric or Binomial test. We call the P-value of enrichment of a pathway or gene-set as its score. We

performed systematic evaluation using cell-type marker gene-set for both bulk ATAC-seq of immune cells[14] and multiple single-cell ATAC-seq profiles[15, 16]. Most of the time UniPath highlighted correct cell type among top 5 enriched gene-set for both bulk and single-cell ATAC-seq profiles (Figure 2d, Supplementary Figures S4 and S5). Making global list of peaks with accessibility score is possible due to the availability of bulk open-chromatin profiles for multiple species. In the absence of enough publicly available open-chromatin profiles for any species, one can also use UniPath by calculating local accessibility score (study-specific normalisation). However, local accessibility scores are dependent on composition of cells in the data-set and could lead to inconsistency in estimation of enrichment of gene-sets (shown in Figure 2e). Thus, UniPath calculates consistent and mostly correct enrichment scores for pathway and gene-sets for every cell using its scATAC-seq profile independently.

Handling drop-out and batch effect

Most often in single-cell scRNA-seq profile, there is heterogeneity in drop-out rate among cells. Such drop-out of genes could be random or systematic. The systematic drop-out rate often occurs due to differences in sequencing depth or RNA degradation level (frozen vs fresh) among different batches of samples. We tested whether UniPath is robust to systematic drop-out variability among cells. We simulated systematic drop-out rate using scRNA-seq data from 2 different studies[17, 18] (Figure 3a) (see Methods). We found that applying PCA on raw read-count lead to artefactual cluster formation due to systematic drop-out. Whereas, using UniPath based pathway scores, similar cells remained in same cluster irrespective of the non-random pattern in drop-out rate (Figure 3a, Supplementary Figure S6a). Besides being robust to systematic drop-out, UniPath allows correction for batch effect before calculating the adjusted p-value for enrichment of pathways (see methods and Supplementary Figure S6b). The framework of UniPath avoids normalisation artefact due to sequencing depth and drop-out rate, therefore, it could be used for efficient classification of single-cell. During hierarchical-clustering, UniPath based gene-set scores provided comparable or higher clustering-purity than raw FPKM based results[10] (Figure 3b, Supplementary Figure S6c-d). Clustering using pathway scores of imputed scATAC-seq profiles also resulted in high clustering purity (Figure 3c). The high accuracy in clustering with gene-set enrichment scores, prooves that defining cell-states in terms of pathway activity can be a reliable method for classifying single-cell epigenome and transcriptome data-sets.

Pseudo-temporal ordering using pathway enrichment scores and visualization of continuum of lineage potency and pathway co-occurrence

Pathway-scores based representation can provide new similarity measures among cells as well as help in avoiding few covariates like cell-cycle phase, tissue-microenvironment or culture conditions. However, current methods for temporal ordering [19] of cells are designed to handle FPKM and read-counts of genes and they are also not meant for visualisation of the continuum of pathway scores on temporally-ordered cells. Hence, we extended UniPath with a novel method for the pseudo-temporal ordering of single-cells which can utilize pathway scores based representations. For temporal ordering, we apply two levels of shrinking of distances between cells based on their pre-classification and continuum among their classes before finding a minimum spanning tree (MST). To find a continuum between different classes we use KNN based approach after initial classification so that correct temporal ordering among clusters of cells can be determined (see methods). Using published scRNA-seq and scATAC-seq profiles, we found that UniPath is indeed able to predict approximately correct order of cells using pathway scores derived from scATAC-seq and scRNA-seq profiles (see Figure 4 and Supplementary Figure S7).

We further used UniPath for temporal ordering of scRNA-seq profile of Human embryonic stem cells (hESC) and their differentiated states collected at time points of 0, 12, 36 , 72 and 96 hours during differentiation towards definitive endoderm (DE) (Figure 4)[17]. Using other tools (monocle, TSCAN, DiffusionMap, CellTree)[18, 20-22] for pseudo-temporal ordering with gene-expression (Transcript per million, TPM) matrix (Figure 4a), resulted in predicting wrong order of cells for same data-set. However with UniPath when we dropped gene-set associated with cell cycle, we achieved correct order of cells. We found that score of gene-set for cell cycle (S phase is shown here) is higher at 0 and 12 hours, possibly due to high level of proliferation (Figure 4c). The S phase gene-set score kept decreasing as the cells differentiated towards endoderm (Figure 4c). However at 36 hour we find two batches of cells such that one batch of cells had much lower level of cell cycle (S phase here) gene-set than other. Such batches of cells hint about possible impact of cell cycle as a covariate during prediction of temporal order. Besides handling known covariates, UniPath can also be used to visualise continuum of lineage potency and concurrence of two pathways on pseudo-temporally ordered tree. As shown in Figure 4d the endodermal lineage gene-set score increases as the cells differentiate towards endoderm (see methods). Such as Wnt/beta-catenin and BMP pathway scores seem to have mild co-enrichment at 0 and 12 hours. As cells differentiated towards mesendodermal stage at 24 and 36 hours, BMP signalling pathway seems to be getting more enriched compared to WNT/beta-catenin. Importance of BMP till mesendoderm stage had been shown before [23]. After 36 hours enrichment level of WNT/beta-catenin is slightly higher however its co-occurrence with BMP shows a slow increase towards of end of temporally ordered tree (Figure 4d). UniPath also enables analysis of co-occurrence pattern and detection of clusters of pathways which can be used to infer context-specific regulation (Supplementary Figures8, Supplementary Figure S9). Overall, UniPath tends to be beneficial for predicting correct temporal order of cells and making inference about stage-specific co-occurrence of pathways during differentiation of cells.

Enabling analysis of large atlas scale scRNA-seq data-set using pathway enrichment scores

The consistency due to use of global null models by UniPath provides horizontal scalability in calculating scores for pathways for single-cells. Even with single CPU the computation time needed by UniPath is much less than PAGODA on same number of cells (Figure 5a). The horizontal scalability, speed and consistency of UniPath allowed us to transform expression profiles of more than 61000 single-cells from mouse cell atlas (MCA) dataset[24], by dividing them into smaller groups of cells. Such step of the division of data-set and transformation to pathway scores with other similar tools (PAGODA) would provide inconsistent results, as explained and demonstrated above. We selected 49507 cells which have more than 800 genes with non-zero FPKM value. Further, t-SNE [25] based dimension reduction of pathway scores and subsequent application of dbscan[26] (Additional file 1), revealed a correct grouping of most of the cells according to their tissue (Figure 5b, Supplementary Figure S10a). As expected, some cells did not group with their source tissue cluster, but they formed a separate class. Such as immune cells from different organs grouped together in cluster 13,14,15 (see Figure 5b, Additional file1).

Surprisingly, co-clustering of few non-immune cells from different tissues revealed convergence which has been rarely reported before by single-cell analysis but is supported by earlier scientific studies. Such as in our analysis, cluster 40 has *Afp*+ fetal liver hepatocytes as well as *Afp*+ placental endodermal cells which were reported to belong to different classes in the original study of MCA (Additional file 2 and 3). Cluster 40 also has a few Fabp1+ hepatocytes. It has been shown previously that placenta-derived multipotent cells (PDMCs) with the expression of Afp (see Figure 6a) gene, has endodermal features and can differentiate easily towards hepatocytes like

cells[27]. We compared both types of cells (*Afp*+ placental endodermal cells and *Afp*+ fetal liver hepatocyte) in cluster 40 with other cells in MCA. Among top 50 pathways more enriched in *Afp*+ placental endodermal cells (from cluster 40) 22 were also present in 50 most differentially upregulated pathway in hepatocytes cells of cluster 40. These common 22 pathways (44% overlap) were mostly related to lipid metabolism (see Table S3). However, there was certainly a difference between hepatocytes and *Afp*+ placental endodermal cells of cluster 40, which is also visible in t-SNE based visualization (Figure 6a).

Another example of convergence is cluster 3 which has virgin mammary gland luminal-epithelial cells (including alveoli cells) and glandular epithelial cells from uterus. An interesting example of convergence is cluster 52 which has *Col10a1*+ and *Cmnd*+ bonemarrow mesenchyme stromal, pre-osteoblast and chondrocytes cells (Additional file 2). It is also well-known that bonemarrow mesenchyme stromal (also known as mesenchymal stem cells [28]) has high potency to transform to pre-osteoblast and chondrocytes cell state[29]. In contrast to such a result, *Cxcl1*+ MSC from in vitro culture grouped with trophoblast stem cells in cluster 21. It is to be noticed that the cell types from different organs, converging in a major cluster, did not overlap completely with each other but formed their own sub-cluster within their major class (Figure 6a). However, the convergence to a major class shows a reduction of covariates due to underlying tissue microenvironment in gene-set scores, which caused cells with similar state to group together. Overall UniPath, provided a new dimension to classify cells and revealed that even though an organ has a specific type of cells for its functioning, it also has some cells with regulatory state similar to cell-types from other parts of the body.

Revealing new minor classes using pathway scores and annotation of unlabeled cells

When analysis is done using FPKM or read-count of a large number of genes, feature selection could be required for proper clustering. However, currently, there is no optimal solution for selecting genes to highlight all relevant classes. Feature extraction in terms of pathway scores can help to reduce noise, sparsity and effect of few covariates. Thus, pathway scores can help to highlight clusters of cells which could not be detected by using raw read-counts. Such as, analysis using pathway scores of brain cells in MCA data-set, resulted in the detection of a new cluster among oligodendrocyte-precursor cells. Oligodendrocyte precursor cells belonging to the new small cluster had a higher expression for *Tuba1a, Sirt2, Cd9, Plp1* and *Bcas1* (Figure 6b). These genes are involved in the differentiation of oligodendrocyte-precursor towards mature oligodendrocytes[30-33]. On the same trend we found two new clusters of “unknown” cells from bladder in MCA data-set (Figure 6c). We could annotate cells in one of the newly detected clusters in bladder as Cd74_high dendritic cells.

In spite of tremendous effort by Han et al. [24] they could not annotate all cells in their MCA dataset, hence among 49507 cells, 5590 cells had annotation as “unknown”. We could find cell-type for 2188 cell with “unknown” label, using a two-pronged approach enabled by UniPath (see Figure 6d, supplementary methods, Additional file 4). Our approach is two-pronged as it utilizes UniPath scores for cell-type marker-set as well as the result of sub-clustering (supplementary method). When we used the same approach on 200 randomly picked cells with labels, we achieved a false detection rate (see Figure 6d) of less than 8% even for low-confidence annotation.

Application in inferring context specific regulation in cancer cells

We further explored how UniPath can be utilised for studying context-specific regulation in cancer cells which is often required for precision oncology. Recently Wang et al. [34] showed a

difference in metabolic profile among two types of NSCLC cell lines, non-adherent tumorspheres (TS) grown in serum-free culture conditions and adherent (Adh) cells cultured in serum-containing medium. They have demonstrated high tumorigenic potential of non-adherent TS cells in comparison to adherent ones using mouse xenograft models. We performed single-cell expression profiling of 80 TS and 82 Adh cells. We asked whether we can utilise the difference among TS and Adh cells in highlighting signalling pathways associated with tumorigenicity in lung cancer. After applying UniPath, the differential enrichment analysis using Wilkoxon Rank sum test revealed GPCR ligand binding gene-set (Figure 7a, Supplementary Figure S11a), IL23 pathway, cytochrome_P450_drug_metabolism and prolactin_receptor_signalling as having higher enrichment in TS cells (based on median fold change and Wilkoxon P-value < 0.01, Figure 7a) (Additional File 5). Distribution in TS and Adh cells and gradient of some pathways are shown in Supplementary Figure S11. GPCR and IL23 signalling are known to be associated with plasticity and proliferation of NSCLC[35-37]. Cytochrome P450 is also involved in promoting tumour development [38].

We further used an approach, rarely used for scRNA-seq. We performed co-occurrence and differential co-occurrence analysis for pathway and gene-set pairs. Wnt pathway had highest correlation with stemness gene-set in TS cells. However, in Adh cells Wnt was not among top correlated pathways with stemness gene-set. We found that Wnt/beta-catenin pathways had a significantly higher correlation with TGF-beta pathway in TS in comparison to Adh cells (P-value < 0.005, Jaccard index=0, see Table S4). Even though TGF-beta pathway itself did not have a significant difference in enrichment among TS and Adh cells (Supplementary Figure S11a). Both Wnt/beta-catenin and TGF-beta are known to promote state of epithelial to mesenchymal (EMT) state in cancer cells which is associated with high tumorigenicity [39]. Moreover, it has been previously shown that simultaneous over-activation of Wnt/beta-catenin and TGF-beta signalling promotes tumorigenicity and chemo-resistance in NSCLC cells [40]. Using hierarchical clustering of 31 chosen pathways, we found that TGF-beta, Wnt/beta-catenin and PDGFRB pathways co-clustered together in TS cells whereas in Adh cell WNT/beta-catenin pathway grouped with ERBB1 and PI3K1 signalling. The difference in co-occurrence pattern of Wnt/beta-catenin pathway in TS and Adh cells (Figure 7 b-c) and prior knowledge about the effect of their co-stimulation with TGF-beta in NSCLC hints about a possible cause of higher tumorogenicity in TS cells.

Wang et al. [34] also reported that glycolytic intermediates are more enriched in Adh cells. Our analysis revealed that among non-metabolic gene-sets, sonic hedgehog (SHH) pathway had the highest level of differential co-occurrence (P-value < 0.005, Jaccard index=0) with glycolysis gene-set (Figure 7d). SHH and glycolysis pathway had a correlation of 0.63 in Adh cell compared to -0.02 in TS cells (Figure 7e). SHH pathway has been shown to be promoting glycolysis in multiple types of cancer[41]. In our hierarchical clustering result (Figure 7b), SHH pathway also seems to group with cell-cycle related gene-set which hints about its involvement in regulation of proliferation in Adh cells. Previously SHH pathway has been associated with proliferation and drug-resistance in NSCLC[42]. However, our analysis reveals that it's role is context-specific and it could have a more dominating role in Adh like NSCLC cells compared to TS cells. Similarly, many more such differences could be revealed among Adh and TS cells. However, our analysis here, is meant to show that UniPath can help in building relevant hypothesis and help researchers in designing follow-up study of context-specific regulation in cancer cells.

Discussion

Exploiting single-cell heterogeneity using pathways and gene-set enrichment can give rise to multiple new applications. However, it needs an estimation of consistent enrichment scores for gene-set. UniPath fills the gap between the demand for consistent gene-set enrichment scores for a multitude of applications and availability of single-cell transcriptome and open-chromatin profiles. The novel approach of processing each cell separately using global null model provides unprecedented consistency and scalability to UniPath for calculating gene-set enrichment. UniPath is robust to systematic drop-out as well as it can handle batch effect in scRNA-seq profiles. For both scRNA-seq and scATAC-seq profiles, there is similarity in the downstream process after the transformation to gene-set enrichment score. Thus, UniPath provides a uniform platform for analyzing both single-cell transcriptome and open-chromatin profiles with the new dimension of pathway enrichment scores. Especially, for scATAC-seq profile, we have shown for the first time that transformation to pathway enrichment scores does not reduce the purity of classification of cells. UniPath also provides an alternative solution to transform more than one scATAC-seq read-count matrices to same feature space, despite differences in their peak list. In addition, we have shown how UniPath can help to use pathway scores for temporal ordering and displaying co-enrichment pattern among them.

Due to its horizontal scalability and consistency, UniPath helped in analysis of large MCA scRNA-seq data-set (> 49000 cells). Classification of MCA data-set using pathway scores revealed few clusters in which one of its member cell-type could be easily differentiated to other. Such as cluster 40 having *Afp*+ placental endodermal cells and fetal liver hepatocyte[27] and cluster 52 with bonemarrow mesenchyme stromal, pre-osteoblast and chondrocytes cells[29]. It could be the result of new regulatory distance defined by pathway scores and suppression of covariate due to tissue micro-environment. Such results hint that, biologist could use UniPath to find convergence and feasibility of convertibility between different cell-types.

There is vast literature on the non-trivial problem of analysing patterns of enrichment and co-occurrence of pathways using bulk expression profiles[43-45]. Exploiting heterogeneity among single-cells with UniPath can easily leverage such analysis. UniPath can be used to elucidate three kinds of differences between two populations of single cells. First one is differential enrichment of pathways which is quite regularly used. The other two possibilities with UniPath which have been rarely explored with single-cell scRNA-seq and scATAC-seq are analysis of modules (clusters) of pathways and enumeration of differential co-occurrence between two pathways. Such as, it revealed increase in enrichment of Nodal signaling with differentiation of hESC towards DE and patterns of it's co-occurrence with other pathways (SMAD2, Wnt/beta-catenin) which are corroborative with existing literature (see supplementary Methods, Supplementary Figure S8). Similarly, UniPath allowed us to study pathway co-occurrence in two types of NSCLC and differential co-occurrences of few pathway pairs (TGF-beta and Wnt/beta-catenin; SHH and glycolysis) which had literature support. Hence, besides improving classical procedure like classification and cell-type detection UniPath also open avenues for new analysis procedures for scRNA-seq and scATAC-seq profile.

Methods

Calculating Enrichment of gene-sets for scATAC-seq profiles

Multiple kinds of regulatory sites like promoters, enhancers and insulators have higher chromatin accessibility than background genomic regions. Most of these regulatory sites like insulators and active promoters tend to have high chromatin accessibility in the majority of cell types. However, to estimate differences among single-cells using open chromatin profiles, sites with cell-type-specific activity like enhancers could be more useful. Moreover, the profile of enhancers provides

a more clear perspective about active pathways in a cell. Therefore, UniPath first normalizes the tag count of scATAC-seq profiles of each cell to highlight enhancers. It has two methods for normalization to highlight enhancers. In the default first method, normalization is done using precalculated global accessibility score of genomic sites. For multiple organisms like human, mouse and Drosophila, bulk sample chromatin accessibility is available for many tissues and cell types. A union list of open-chromatin sites was made for human hg19 version genome and the accessibility score of union list of site was calculated. For example, for Human hg19 genome, we combined DNAse-seq and ATACseq peaks from ENCODE and IHEC consortiums[46], to achieve more than 1 million sites and calculated accessibility scores of combined peak list (see supplementary methods). The accessibility score is calculated for a site as the proportion of cell types or samples in which it was detected as open-chromatin peak.
For tagcount p_{ij} of a peak i in a single-cell j, the normalisation is done as

$$t_{ij} = p_{ij} / (a_i + \epsilon) \quad (1)$$

where ϵ stands for a pseudo-count and a_i is the global accessibility score for peak i. Thus, the first method of highlighting enhancers using global accessibility score does not need any inter-cell tag-count normalisation. Using the first method also makes it possible to have uniform transformation of scATAC-seq read-count matrix from different scientific groups without re-calculating tag-counts using the aligned DNA fragments (bam or sam files) on a common peak-list.
The second method can be used while analysing groups of cells with high heterogeneity among each other. In second method the quantile-normalisation of read-counts of cells is performed followed by the division of read-count of every genomic site by its mean read-count across all the cells. The second method needs heterogeneity among cells in order to be more effective.

For every cell, the peaks having high normalised tag-count are selected and used as set of positives (foreground) and the set of all peaks is used as background. Usually, we use a threshold of 1.25 above global accessibility score for choosing foreground peak, but it could vary depending on stringency needed. The chosen peaks in the positive set are highly likely to be enhancers and regulatory sites with cell-type-specific activity. Then for every peak most proximal gene within 1Mbp is found and peaks which do not have any gene within 1Mbp is dropped. To decide the most effective statistical test we used two different ways to calculate the significance of enrichment of pathways and benchmarked them by applying them using known set of markers (gene-sets) for different cell types. The statistical methods we use are binomial and hypergeometric tests. With binomial test, to calculate statistical significance (P-value) for a gene-set m whose genes appear proximal to k_m out of n peaks in foreground set, we use the formula below

$$\sum_{i=k_m}^{n} \binom{n}{i} p_m^i \, (1 - p_m)^{n-i} \quad (2)$$

Here p_m represents the probability of genes from the gene-set m to appear as proximal to peaks in the background list. With hypergeometric test the calculation of statistical significance (P-value) is
done using the formula

$$\sum_{i=k_m}^{\min(n,K_m)} \frac{\binom{K_m}{i}\binom{N-K_m}{n-i}}{\binom{N}{n}} \quad (3)$$

Where K_m is the number of times genes of gene-set m appear as proximal to peaks in the background, and N is total number of peaks in the background set. As above k_m represents the number of times out of n foreground peaks, the proximal genes are from gene-set m.

Normalisation free Gene-set enrichment for single-cell expression

For estimating the significance of enrichment of pathway (gene-set) using scRNA-seq, we use FPKM of genes and treat every cell independently from each other. Thus unlike other published methods, we avoid creating artefacts which can happen due to the unresolved issue of estimating the distribution of tag-count of a gene across multiple samples(or cells for normalisation. As scaling and normalisation across different cells can create artefacts due to a variable level of noise and gene drop-out rate among them. Rather we use the widely accepted fact that within a sample (cell) non-zero FPKM (or RPKM) values of genes follow approximately log-normal distribution (see Supplementary Figure S1a). We validated assumption (see Supplementary Figure S1a) for genes with non-zero FPKM values on data-sets from multiple studies and modelled log(FPKM) distribution as bimodal such that one mode corresponds to genes with zero FPKM and other mode correspond to normal distribution. Thus probability distribution function (pdf) for log(FPKM) value x in a cell can be written as

$$f(x) = p_0 I(x = 0) + (1 - p_0) N(x;\, \mu, \sigma) \qquad (4)$$

Where N(x; μ, σ) represent Gaussian pdf for genes with non-zero FPKM and I(x = 0) is the indicator function, whereas p_0 represents a fraction of genes with zero FPKM. The variables μ and σ represents the mean and standard deviation respectively in the logarithmic domain for non-zero FPKMs. Thus, for every cell we use its own value of μ and σ to convert the non-zero FPKM value of gene in to P-value (right-tailed) assuming Gaussian distribution. Then we combine p-values of genes belonging to a gene-set using Brown's method[47]. Brown's method is meant to combine p-values which have a dependence upon each other. Using Brown's method the combined p-value for a gene-set with k genes with non-zero FPKM can be given by

$$Pcombined = 1.0 - \Phi_{2f}(\psi/c) \qquad (5)$$

Where $\psi = -2\sum_{i=1}^{k} \log P_i$ such that P_i is p-value of log(FPKM) of gene i in a sample/cell and Φ_{2f} is the cumulative distribution function for the chi-square distribution χ^2_{2f}. Here f is the scaled degree of distribution and is calculated as $f = E[\psi]^2/var[\psi]$ [47]. The value of c in equation (5) is calculated as

$c = var[\psi]/2E[\psi]$, such that $E[\psi] = 2k$ and $var[\psi] = 4k + 2\sum_{(i<j)} cov(-2\log P_i, -2\log P_j)$.

This procedure leads to the calculation of combined P-value for each gene-set in every cell. In order to have robust estimate not affected by just 1 or 2 genes we use a threshold of minimum 5 genes with non-zero FPKM to calculate combined p-value for a gene-set. However, combined p-values could also have many unwanted effects from house-keeping genes, promiscuously enriched gene-set and multiple hypothesis testing. Hence, we corrected the p-values with a permutation-based test using a null model.

In order to make null model we first randomly chose cells from multiple studies so that we can have equal representation of multiple cell-type. Then we performed hierarchical clustering of chosen cells using genes selected using a criterion of coefficient-of-variation[48]. Using dynamic cutting of the hierarchical tree, we achieved clusters (or classes) of cells. We made 1000 pairs of cells such that in a pair the cells belonged to different classes. For each pair, we took average expression value for all gene. Thus, the null model consisted of 1000 expression vectors (false cells), each being average of gene-expression profiles of two cells. For every false-cell vector in null model, the combined p-values of gene-sets were calculated using the method mention above. Thus, for every pathway (gene-set) we achieved 1000 p-values corresponding to the number of false-cells in null model. For a pathway (gene-set) to calculate adjusted P-value in target cell we take the proportion of cells in null model which had lower combined P-value than the target cell.

UniPath's approach of temporal ordering of cells using pathway scores

Nearly all the methods developed so far for temporal ordering of single-cells use gene-expression or read-count data. Hence to utilize the continuity in pathway activity among cells and to get better insight using single-cell profiles, we developed novel temporal ordering method which can work efficiently using the pathway scores of single-cells. Our method first performs hierarchical clustering of cells before finding the order among the clusters of cells, followed by distance weighting and learning minimum spanning tree. As illustrated by Zhicheng and Hongkai, applying mimimum spanning tree detection directly on raw distances among cells like monocle-1 [18] can lead to false connection between cells due to noise or other bias[21]. However, following Zhicheng and Hongkai's method it is not feasible to get true ordering at single-cell resolution. Hence, we developed an approach, such that after initial classification of cells using pathway scores, we shrink (or weight) distances among every cell pair based on their belonging-ness to same class and using neighbourhood index among their classes. To calculate neighbourhood index among classes we first find top k nearest neighbour for every cell. Then for every class we count the number of times its cells have top k neighbours in other classes. For example, if cells in class A has total M neighbours in others classes out of which mb cells are from a class B then we calculate neighbourhood index of A with B (A ->B) as mb/M. We shrink the distances between the cells in class A and class B by mb/M. After two stages of shrinkage of distances among cells, we use shrinked-distance matrix to find minimum spanning tree. We plot the minimum spanning tree using the netbiov R library[49]. The minimum spanning tree drawn using our approach has fewer chances to be influenced by noise as the distances among cells are shrunk using consensus information.

Test for pathway score accuracy and consistency of UniPath

Even though we tested UniPath and PAGODA using cell-type markers, we also used spike-in method for measuring accuracy for pathway scores. For this, first we collected gene-sets for non-immune pathways. In this collection, we also added 2 gene-set specific to B cells and 2 pathways for T cells. Both PAGODA and UniPath were used with collected gene-set on FPKM of B cells and T cells. UniPath and PAGODA were evaluated based on the presence of relevant pathway among top 5 gene-sets in their output.

Simulating systematic drop-out

There could be many reasons of zero read-count for genes in scRNA-seq data-set. Those reasons include true biological cause, random and systematic drop-out. Variability due to uniform random drop-out can be handled by dimension reduction methods like PCA and t-SNE. However, systematic drop-out often leads to errors during dimension reduction and classification. Therefore, we simulated systematic drop-out to evaluate UniPath. To create systematic drop-out for evaluation, we first randomly chose some genes. We dropped the FPKM values of chosen genes from randomly selected cell; in other words, we made the FPKM values of chosen genes to zero. The case of systematic drop-out is actually different from most of the technical batch effect.

Differential co-occurrence analysis:

We use permutation to estimate the significance of difference of pathway co-occurrence among two groups of cells. For a pathway pair, we first calculate the difference between spearman correlation values of their enrichment scores (adjusted P-value) in two groups of cells. We call it as true difference. We perform random-shuffling of group-labels of cells and calculate difference in spearman-correlation of enrichment scores among two shuffled groups. Thus for a pair of pathways we make a collection of set of false difference in correlations using shuffled groups. The p-value is calculated as fraction of false differences which are greater than true difference in term of absolute value. Notice that, here, we use spearman correlation of adjusted P-value

(pathway score), not just the combined p-value of gene-sets. Using adjusted P-value increases robustness as it becomes rank based scores which helps in filtering out effect due to only 1 or 2 genes. Thus if two pathways are correlated using their adjusted P-value, the correlation has less chance to be affected by only 1 or 2 genes or outliers. Even though random permutation-based significance test reduces possible covariates, we have ignored differential co-occurrence among gene-sets which have high Jaccard index for overlap of genes. For hESC cells differentiating towards DE, we performed differential co-occurrence analysis for every time point of differentiation. Such as for differential co-occurrence analysis for time point of 12 hours, we compared them with cells not belonging to 12 hours.

Single-cell expression profiling for non-small lung cancer cells

The source and culture condition for Tumour sphere (TS) and Adherant cells(Adh) are mentioned in Wang et al. [34]. Tumour sphere (TS) line derived from lung cancer patient were maintained in medium with DMEM/F12 (US Biomedical), 4mg/ml Bovine Serum Albumin (Sigma), Non-essential amino acids, sodium pyruvate (Life Technologies) and 20ng/ml Epidermal Growth Factor, 4 ng/ml bovine Fibroblast Growth Factor and Insulin – Transferrin Selenium (Sigma).

Tumour sphere derived adherent (Adh) cells were grown in the same media as above, without EGF, bFGF, ITS and BSA. For Adh cells, media was supplemented with 10% fetal bovine serum.

RNA extraction, library construction, sequencing for NSCLC cells

NSCLC single-cells in suspension were dissociated using trypsin and loaded into C1 96 well-integrated microfluidic chip (IFC) as per manufactures guidelines. The single-cells were captured in C1 96 (large size) IFC using Fluidigm-C1 system. The captured single-cells were imaged using auto imaging fluorescent microscope to identify the viable single-cells and to omit the doublets. The reverse transcription and cDNA pre-amplification reagents were prepared using SMART-seq2 protocol and loaded into the IFC. Later reverse transcription and cDNA amplification were processed using SMART-seq2 script automatically in C1-Fluidigm machine. After harvesting cDNA from C1 chip, the samples were quantified using picogreen assay and normalized to the range 0.2-0.3ng/µL. The quality of the cDNA product was verified using high sensitivity DNA assay in Agilent bio-analyzer machine. The harvested single-cell cDNA was barcoded in 96 well plate using Nextera XT Library Prep kit (Illumina). Uniquely barcoded libraries from single-cells pooled together and sequenced using a HiSeq-Hi-output-2500 sequencer (Illumina).

Implementation and data availability

UniPath is implemented using R and is available at https://reggenlab.github.io/UniPathWeb/
As well as at https://github.com/reggenlab/UniPath
The FPKM data for single-cell RNA-seq for lung cancer cells is available with UniPath package. The raw sequences are being uploaded to BioSample database. The raw sequences contain genomic identity information of cancer patients, hence they can be accessed only with permission.

Authors contributions

VK and DS designed the study, WLT helped in designing study related to cancer cells. VK and SC wrote the code and the manuscript. SC and VK also executed the code on data-sets for generating results and figures. SS and SLK prepared the library for single-cell RNA-seq of lung cancer cell. ZW and WLT cultured non-small cell lung cancer cell-line and provided information about cultured cells and their behavior.

Acknowledgement:

This project was supported by YIG Grant (BMRC/YIG/1510851023) provided to Vibhor Kumar by BMRC, A-STAR, Singapore.

Ethics declarations
we used a previously characterized LC32 cell lines derived from resected primary non-small-cell lung cancer (NSCLC) adenocarcinoma samples

Competing Interest
The authors have no competing interests

Figure Caption

Figure 1: **Outline of UniPath** (a) Schematic workflow of UniPath for scRNA-seq data. UniPath works by transforming scRNA-seq gene expression profiles to P-values combined using Brown's method for each gene-set. The combined P-values are adjusted by using a null model to estimate final pathway score. The null models are made systematically to avoid redundancy. (b) Schematic workflow of UniPath for scATAC-seq data. UniPath transforms open chromatin profiles to pathway enrichment scores for gene-sets by highlighting enhancers and using their proximal genes for Hypergeometric or Binomial test. To highlight enhancers, UniPath normalizes read-count at a peak by its global accessibility score.

Figure 2: **Evaluation of UniPath using scRNA-seq and scATACseq profiles.** (a) Accuracy of highlighting relevant gene-set among top enriched terms. The terms here are cell-types and gene-sets are set of marker genes for cell-types. For evaluation of UniPath and PAGODA the results for scRNA-seq profiles of Epithelial cells and Astrocytes, as shown here. The evaluation was performed using both homogeneous and non-homogeneous data-sets (Table S1). More such examples are shown in Supplementary Figure S1. (b) Accuracy of results of pathway enrichment by UniPath and PAGODA for B cell and T cell scRNA-seq profiles. For systematic evaluation, two pathways relevant to T cell and other two gene-set for B cells were spiked-in into the list of gene-set related non-immune functions. (c) Consistency of UniPath pathway scores when B cells are grouped with epithelial or T cells in comparison to PAGODA. Gene-set enrichment scores provided by PAGODA change when the same cell is grouped with other cells. While UniPath's output remains consistent. (d) Evaluation of UniPath for highlighting correct gene-sets among top enriched term for single-cell ATAC-seq profile. For this purpose marker gene-sets for cell-type were used by UniPath on scATAC-seq profile of B cell (GM12878) and Monocyte. Here UniPath used global accessibility scores to highlight enhancers. (e) Consistency of UniPath based pathway enrichment score calculation using scATAC-seq. Here hematopoietic progenitor cells are grouped with B cells or monocytes. UniPath's approach of highlighting enhancers by using global accessibility score gives more consistent result than mean based normalisation (local accessibility score). While doing normalisation of scATAC-seq profile using local accessibility score, the enrichment-scores of pathways are highly dependent upon composition of neighbouring cells.

Figure 3: **Reduction of artifact by Unipath and clustering using pathway scores** (a) Principal component analysis (PCA) based visualization of Embryonic stem cells combined with Myoblast. Here PCA was done using scRNA-seq based gene expression. Simulation of systematic dropout of 10% genes in few embryonic stem cells lead to the formation of a separate group of ESCs when PCA based visualisation of gene-FPKM. However, PCA using pathway score from UniPath lead to grouping of all ESCs in same cluster irrespective of systematic dropout. (b) Clustering purity of scRNA-seq based gene expression without and with transformation to pathway scores for dataset published by Li et al.[10]. (c) Clustering purity of scATAC-seq profiles transformed into pathway space. Scatter plot of tSNE results for cells from two scATAC-seq datasets (Cusanovish et al. [50] and Buenrostro et al.[15]) are shown here.

Figure 4: **Pseudotemporal ordering using gene set enrichment scores and visualisation of potency and pathway co-occurrences.** The dataset used here, consisted of cells collected at different time points (0,12, 24, 36, 72 and 96 hours) of differentiation of human embryonic cells (hESC, 0 hours) towards definitive endoderm (DE) [17] (96 hours). (a) Imperfect prediction of temporal order using gene-expression by other tools. Monocle mixed 0 hours (hESCs) and 96 DE hours cells, Diffusion map also mixed 0 hour cells with 72 hours. TSCAN could not find a proper order in sequence of 0, 12, 24, 36, 72 and 96 hours , cellTree also could not find a proper temporal order among cells (b) Predicted temporal order of cells of differentiating human embryonic stem cells towards definitive endoderm. The order predicted is exactly according to true time-points of cells. (c) The enrichment score of gene-set for S phase in cells at different time point of differentiation. (d) The trend of endoderm lineage potency and co-occurrence of pathways at single-cell resolution on temporally ordered tree.

Figure 5: **Execution time of UnPath and analysis atlas scale single-cell RNA-seq data-set** (a) Comparison of the execution time of UniPath with PAGODA and PCA with varying number of cells. (b) Scatter plot of t-SNE results for 49507 cells of mouse cell atlas (MCA) data-set [24] represented using pathway. The transformation of MCA data-set to pathway enrichment scores was possible due to consistency and scalability provided by UniPath. The clusters detected using the shown tSNE result for MCA data-set are shown in Supplementary Figure S10a.

Figure 6: **Analysis of single-cell RNA-seq profile of Mouse cell Atlas (MCA)** (a) t-SNE scatter plot of cells co-clustering in cluster number 40. AFP high placenta endodermal cells and Afp+ hepatocytes do not overlap, but they lie closer to each other in t-SNE based plot. (b) t-SNE results of scRNA-seq profile from brain showing two clusters of oligodendrocyte precursor cells along with their enriched genes. These two clusters of cells were labeled as a single cell-type in the original study by Han et al. [24] (c) t-SNE based scatter plot for bladder cells in MCA data-set represented in terms of pathway enrichment scores. Two clusters of cells labeled as “unknown” were visible. Cells in one of the two clusters were identified as cd74_high_dendritic cells. (d) Pie chart showing confidence level for cell-type annotation on MCA data. Total 2188 cells with “unknown” label could be annotated with the help of UniPath. False detection rates for different confidence level are also shown for the same procedure when same procedure was applied to 200 randomly chosen but labeled cells.

Figure 7**: Differences in enrichment and co-occurrence of pathways in two types of cells of non-small cell lung cancer (NSCLC).** a) A global view of differential enrichment of

pathways using volcano plot. The x-axis shows log fold-change of median enrichment scores of gene-sets in tumorosphere (TS) and Adherent cells (Adh). P-values were calculated using the Wilcoxon rank-sum test. Few pathways which showed significant difference in enrichment between TS and Adh cells are shown here. (b) Heatmaps of correlation between pathways are shown with their hierarchical cluster in TS and Adh lung cancer cells. (c)Correlation values of WNT signaling pathway with gene-set of stemness in TS and Adh cells are shown as bar plot. The other bar-plot shows co-occurrence of WNT/beta-catenin with TGF-beta signaling pathways in Adh and TS cell lines. WNT/beta-catenin and TGF-beta had significant differential cooccurrence among TS and Adh. (P-value < 0.005) (d) Volcano plot showing differential co-occurrence of pathways with gene-set for Glycolysis_Gluconeogenesis among TS and Adh cells. Sonic Hedgehog (SHH) pathway and glycolysis gene-set had significant differential co-occurrence among TS and Adh cells. (e) Spearman correlation of scores of Gycolysis_Gluconegenesis pathway with sonic hedgehog pathway (SHH) pathway in TS and Adh cells. The P-value of differential co-occurrence is also shown.

References

1. Packer J, Trapnell C: **Single-cell multi-omics: an engine for new quantitative models of gene regulation.** *Trends in Genetics* 2018.
2. Jia G, Preussner J, Chen X, Guenther S, Yuan X, Yekelchyk M, Kuenne C, Looso M, Zhou Y, Teichmann S: **Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement.** *Nature communications* 2018, **9:**4877.
3. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O: **Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.** *Nature biotechnology* 2015, **33:**155.
4. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** *PLoS genetics* 2007, **3:**e161.
5. Gagnon-Bartsch JA, Speed TP: **Using control genes to correct for unwanted variation in microarray data.** *Biostatistics* 2012, **13:**539-552.
6. Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O: **f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq.** *Genome biology* 2017, **18:**212.
7. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan J-B, Zhang K, Chun J: **Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis.** *Nature methods* 2016, **13:**241.
8. Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, Pawitan Y: **Beta-Poisson model for single-cell RNA-seq data analyses.** *Bioinformatics* 2016, **32:**2128-2135.
9. Mukherjee S, Zhang Y, Fan J, Seelig G, Kannan S: **Scalable preprocessing for sparse scRNA-seq data exploiting prior knowledge.** *Bioinformatics* 2018, **34:**i124-i132.
10. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS: **Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors.** *Nature genetics* 2017, **49:**708.
11. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Gephart MGH, Barres BA, Quake SR: **A survey of human brain transcriptome diversity at the single cell level.** *Proceedings of the National Academy of Sciences* 2015, **112:**7285-7290.
12. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J: **Massively parallel digital transcriptional profiling of single cells.** *Nature communications* 2017, **8:**14049.

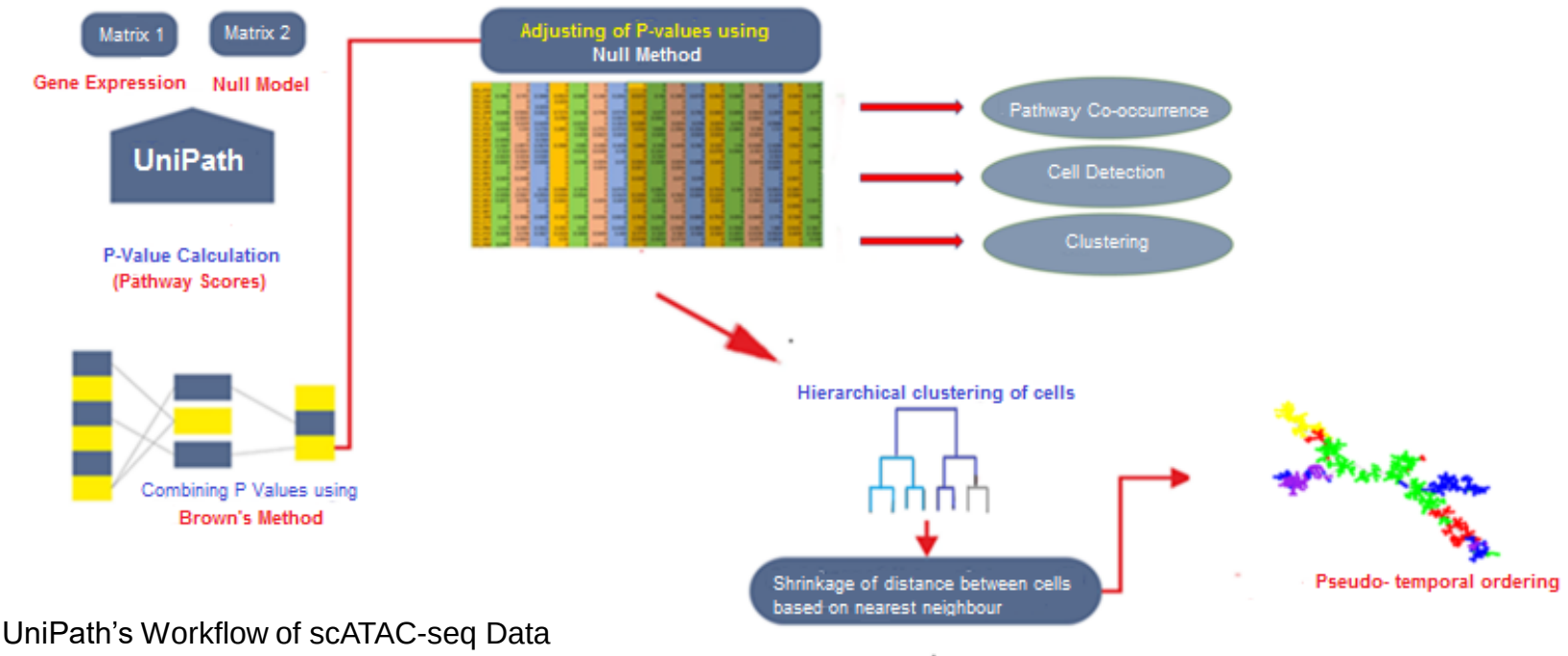
13. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su M-J, Melms JC, Leeson R, Kanodia A, Mei S, Lin J-R: **A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade.** *Cell* 2018, **175:**984-997. e924.
14. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ: **Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.** *Nature genetics* 2016, **48:**1193.
15. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ: **Single-cell chromatin accessibility reveals principles of regulatory variation.** *Nature* 2015, **523:**486.
16. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ: **Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation.** *Cell* 2018, **173:**1535-1548. e1516.
17. Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, Choi J, Kendziorski C, Stewart R, Thomson JA: **Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm.** *Genome biology* 2016, **17:**173.
18. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL: **The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.** *Nature biotechnology* 2014, **32:**381.
19. Saelens W, Cannoodt R, Todorov H, Saeys Y: **A comparison of single-cell trajectory inference methods.** *Nature biotechnology* 2019, **37:**547.
20. Yotsukura S, Nomura S, Aburatani H, Tsuda K: **CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data.** *BMC bioinformatics* 2016, **17:**363.
21. Ji Z, Ji H: **TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis.** *Nucleic acids research* 2016, **44:**e117-e117.
22. Haghverdi L, Buettner F, Theis FJ: **Diffusion maps for high-dimensional single-cell analysis of differentiation data.** *Bioinformatics* 2015, **31:**2989-2998.
23. Loh KM, Ang LT, Zhang J, Kumar V, Ang J, Auyeong JQ, Lee KL, Choo SH, Lim CY, Nichane M: **Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations.** *Cell stem cell* 2014, **14:**237-252.
24. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F: **Mapping the mouse cell atlas by microwell-seq.** *Cell* 2018, **172:**1091-1107. e1017.
25. Maaten Lvd, Hinton G: **Visualizing data using t-SNE.** *Journal of machine learning research* 2008, **9:**2579-2605.
26. Tran TN, Drab K, Daszykowski M: **Revised DBSCAN algorithm to cluster data with dense adjacent clusters.** *Chemometrics and Intelligent Laboratory Systems* 2013, **120:**92-96.
27. Chien CC, Yen BL, Lee FK, Lai TH, Chen YC, Chan SH, Huang HI: **In vitro differentiation of human placenta-derived multipotent cells into hepatocyte-like cells.** *Stem Cells* 2006, **24:**1759-1768.
28. Elsafadi M, Manikandan M, Atteya M, Hashmi JA, Iqbal Z, Aldahmash A, Alfayez M, Kassem M, Mahmood A: **Characterization of cellular and molecular heterogeneity of bone marrow stromal cells.** *Stem cells international* 2016, **2016**.
29. Ashton BA, Allen TD, Howlett C, Eaglesom C, Hattori A, Owen M: **Formation of bone and cartilage by marrow stromal cells in diffusion chambers in vivo.** *Clinical orthopaedics and related research* 1980**:**294-307.

30. Aiken J, Buscaglia G, Bates EA, Moore JK: **The α-tubulin gene TUBA1A in brain development: a key ingredient in the neuronal isotype blend.** *Journal of developmental biology* 2017, **5:**8.
31. Lourenço T, De Faria JP, Bippes CA, Maia J, Lopes-da-Silva JA, Relvas JB, Grãos M: **Modulation of oligodendrocyte differentiation and maturation by combined biochemical and mechanical cues.** *Scientific reports* 2016, **6:**21563.
32. Fard MK, van der Meer F, Sánchez P, Cantuti-Castelvetri L, Mandad S, Jäkel S, Fornasiero EF, Schmitt S, Ehrlich M, Starost L: **BCAS1 expression defines a population of early myelinating oligodendrocytes in multiple sclerosis lesions.** *Science translational medicine* 2017, **9:**eaam7816.
33. Ji S, Doucette JR, Nazarali AJ: **Sirt2 is a novel in vivo downstream target of Nkx2. 2 and enhances oligodendroglial cell differentiation.** *Journal of molecular cell biology* 2011, **3:**351-359.
34. Wang Z, Yip LY, Lee JHJ, Wu Z, Chew HY, Chong PKW, Teo CC, Ang HY-K, Peh KLE, Yuan J: **Methionine is a metabolic dependency of tumor-initiating cells.** *Nature medicine* 2019, **25:**825.
35. Kuzumaki N, Suzuki A, Narita M, Hosoya T, Nagasawa A, Imai S, Yamamizu K, Morita H, Suzuki T, Okada Y: **Multiple analyses of G-protein coupled receptor (GPCR) expression in the development of gefitinib-resistance in transforming non-small-cell lung cancer.** *PLoS One* 2012, **7:**e44368.
36. Kastner S, Voss T, Keuerleber S, Glöckel C, Freissmuth M, Sommergruber W: **Expression of g protein-coupled receptor 19 in human lung cancer cells is triggered by entry into s-phase and supports g2–m cell-cycle progression.** *Molecular Cancer Research* 2012, **10:**1343-1358.
37. Baird A-M, Leonard J, Naicker KM, Kilmartin L, O'Byrne KJ, Gray SG: **IL-23 is pro-proliferative, epigenetically regulated and modulated by chemotherapy in non-small cell lung cancer.** *Lung Cancer* 2013, **79:**83-90.
38. Oyama T, Sugio K, Uramoto H, Onizuka T, Iwata T, Nozoe T, Takenoyama M, Hanagiri T, Isse T, Kawamoto T: **P2-049: Cytochrome P450 expression in non-small cell lung cancer.** *Journal of Thoracic Oncology* 2007, **2:**S509-S510.
39. Heldin C-H, Vanlandewijck M, Moustakas A: **Regulation of EMT by TGFβ in cancer.** *FEBS letters* 2012, **586:**1959-1970.
40. Cai J, Fang L, Huang Y, Li R, Xu X, Hu Z, Zhang L, Yang Y, Zhu X, Zhang H: **Simultaneous overactivation of Wnt/β-catenin and TGFβ signalling by miR-128-3p confers chemoresistance-associated metastasis in NSCLC.** *Nature communications* 2017, **8:**15870.
41. Ge X, Lyu P, Gu Y, Li L, Li J, Wang Y, Zhang L, Fu C, Cao Z: **Sonic hedgehog stimulates glycolysis and proliferation of breast cancer cells: Modulation of PFKFB3 activation.** *Biochemical and biophysical research communications* 2015, **464:**862-868.
42. Yuan Z, Goetz J, Singh S, Ogden S, Petty W, Black C, Memoli V, Dmitrovsky E, Robbins DJ: **Frequent requirement of hedgehog signaling in non-small cell lung carcinoma.** *Oncogene* 2007, **26:**1046.
43. Ramanan VK, Shen L, Moore JH, Saykin AJ: **Pathway analysis of genomic data: concepts, methods, and prospects for future development.** *TRENDS in Genetics* 2012, **28:**323-332.
44. Pita-Juarez Y, Altschuler G, Kariotis S, Wei W, Koler K, Green C, Tanzi R, Hide W: **The pathway Coexpression network: revealing pathway relationships.** *PLoS computational biology* 2018, **14:**e1006042.
45. Li Y, Agarwal P, Rajagopalan D: **A global pathway crosstalk network.** *Bioinformatics* 2008, **24:**1442-1447.

46. Bujold D, de Lima Morais DA, Gauthier C, Côté C, Caron M, Kwan T, Chen KC, Laperle J, Markovits AN, Pastinen T: **The international human epigenome consortium data portal.** *Cell systems* 2016, **3:**496-499. e492.
47. Poole W, Gibbs DL, Shmulevich I, Bernard B, Knijnenburg TA: **Combining dependent P-values with an empirical adaptation of Brown's method.** *Bioinformatics* 2016, **32:**i430-i436.
48. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC: **Accounting for technical noise in single-cell RNA-seq experiments.** *Nature methods* 2013, **10:**1093.
49. Tripathi S, Dehmer M, Emmert-Streib F: **NetBioV: an R package for visualizing large network data in biology and medicine.** *Bioinformatics* 2014, **30:**2834-2836.
50. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J: **Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing.** *Science* 2015, **348:**910-914.

Fig. 1

a UniPath's Workflow of scRNA-seq Data



b UniPath's Workflow of scATAC-seq Data

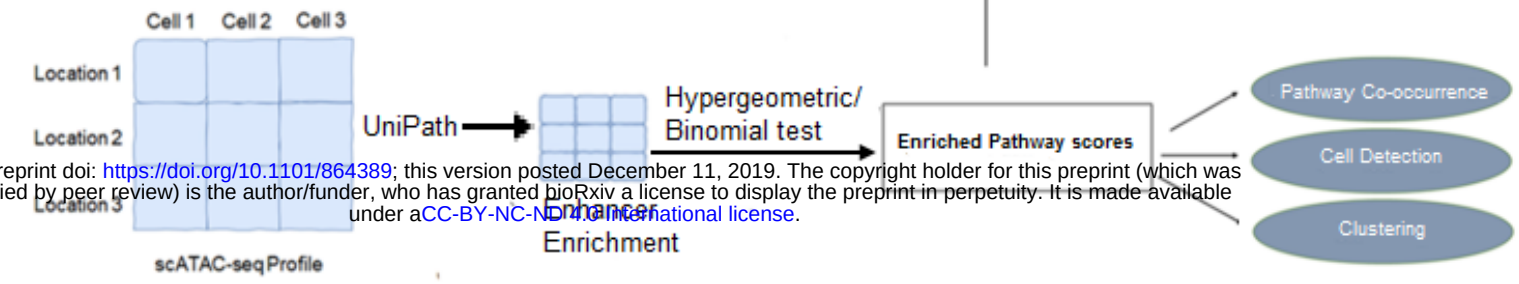
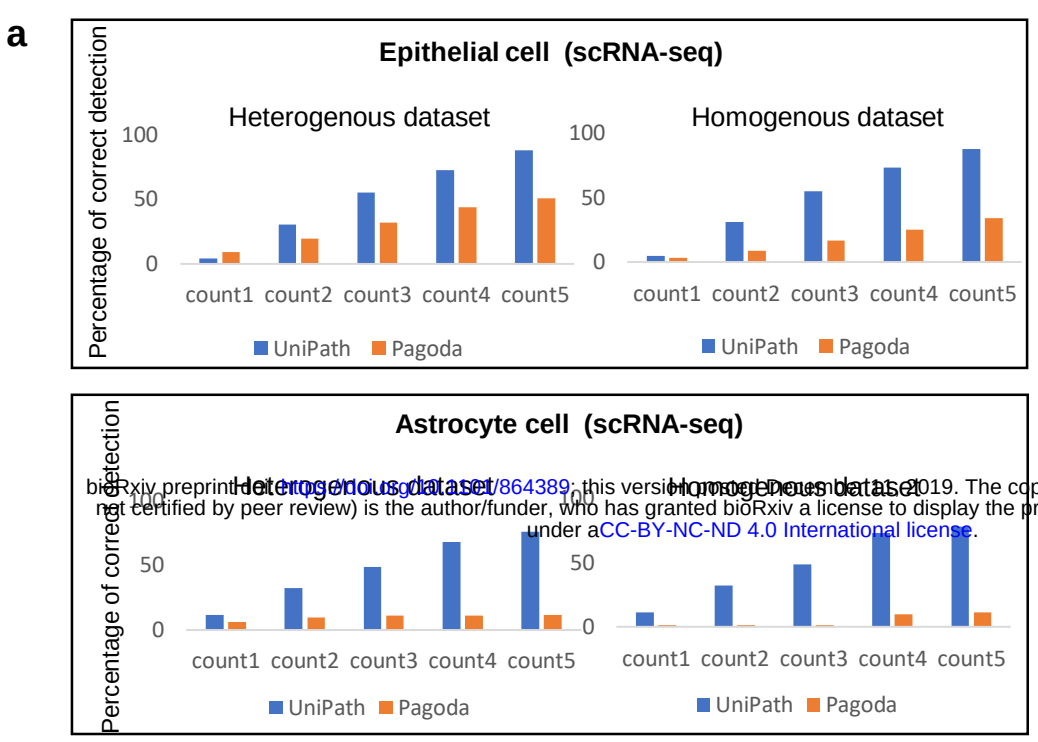
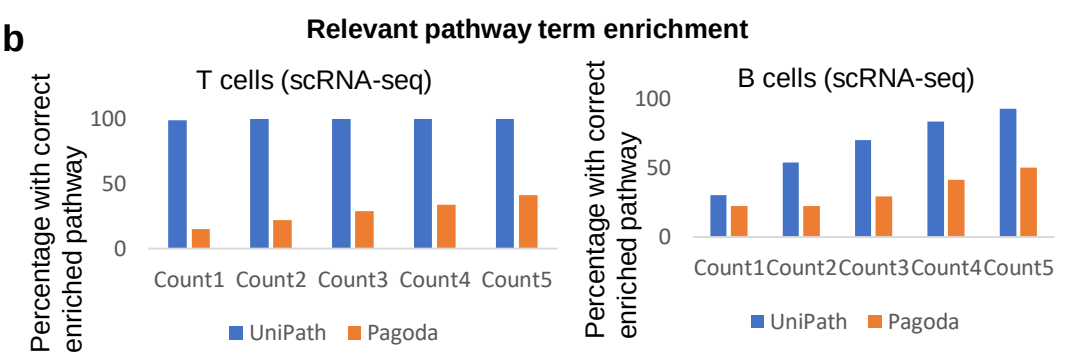
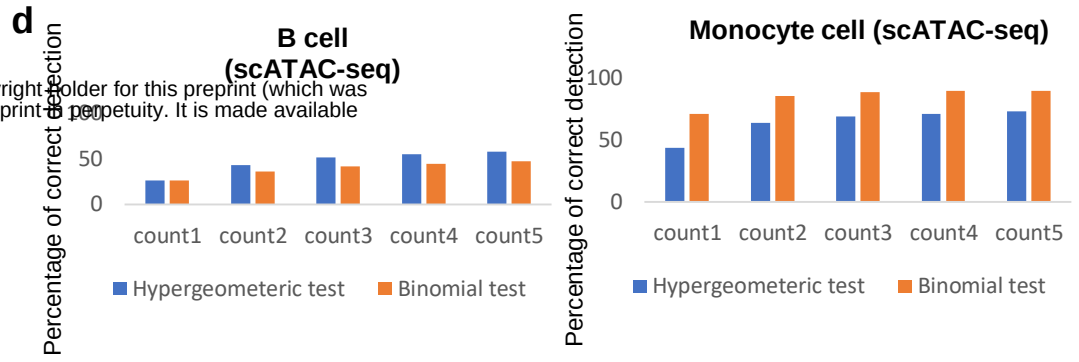
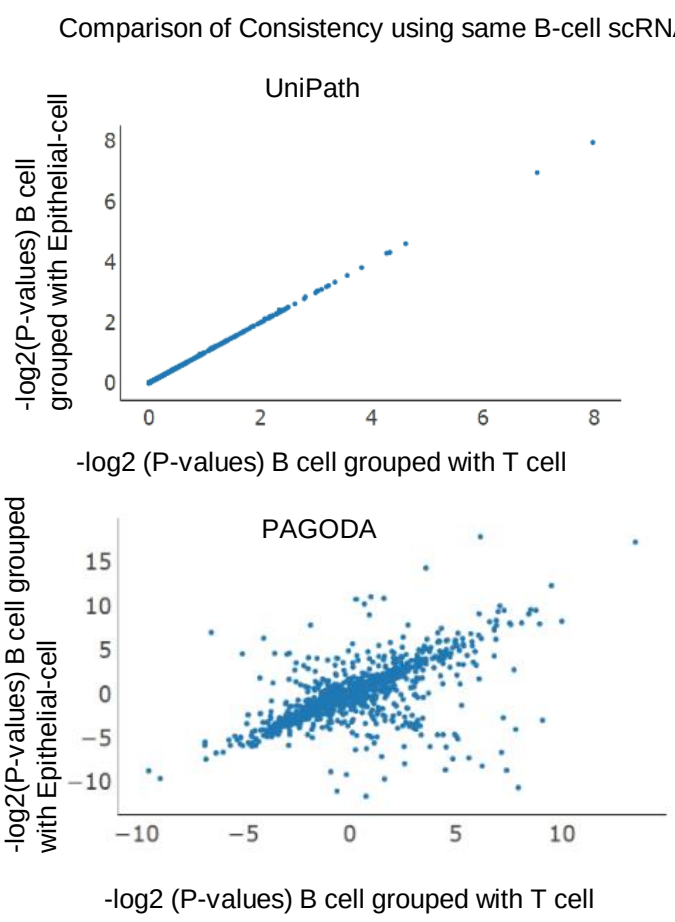


Fig. 2









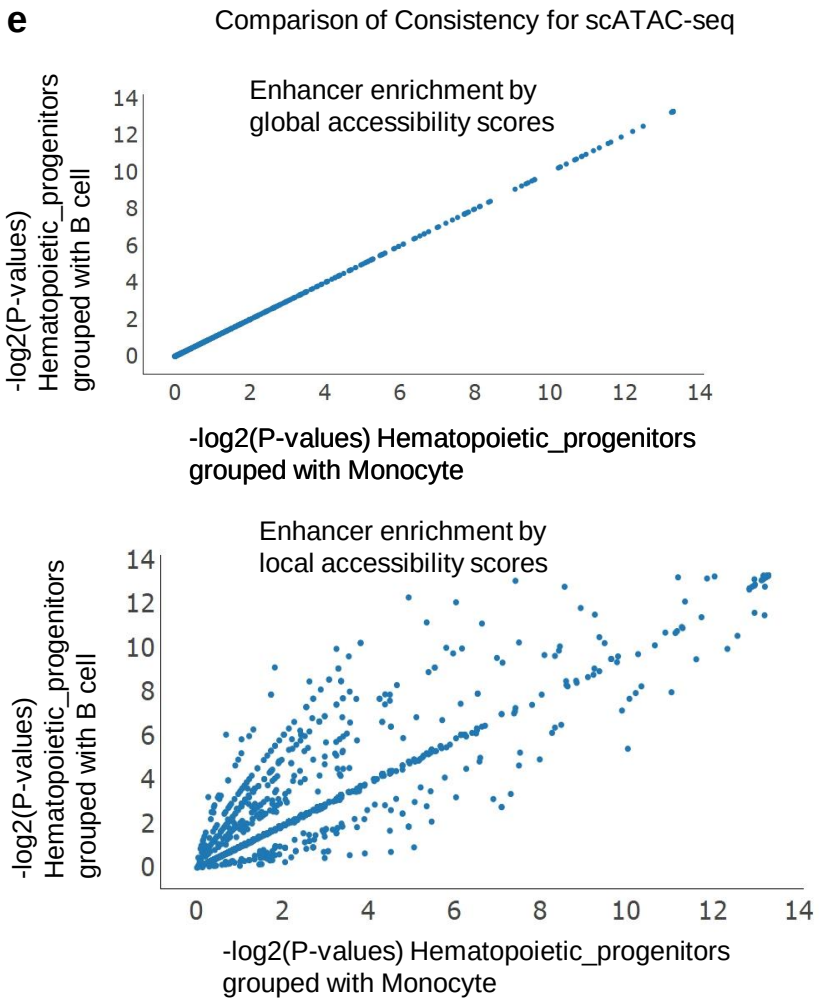
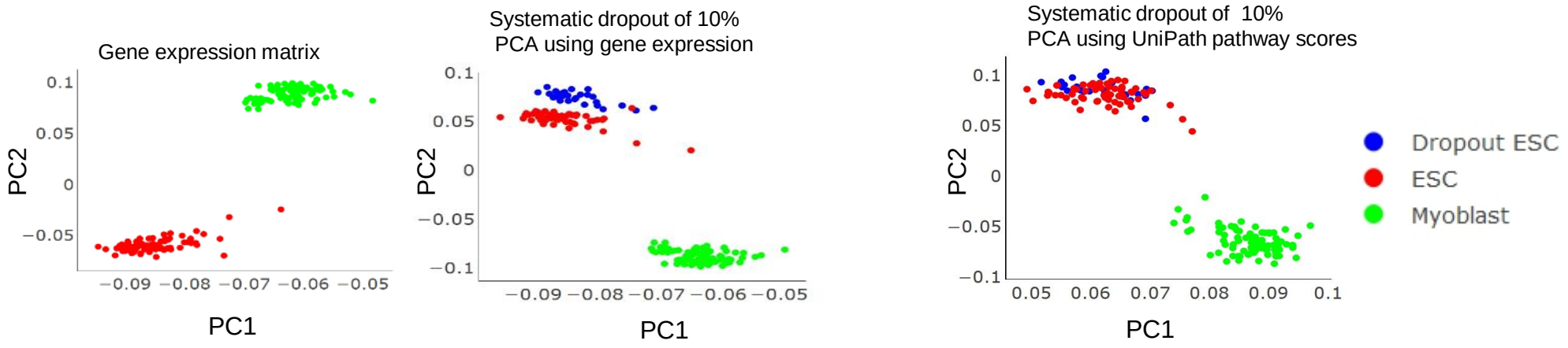
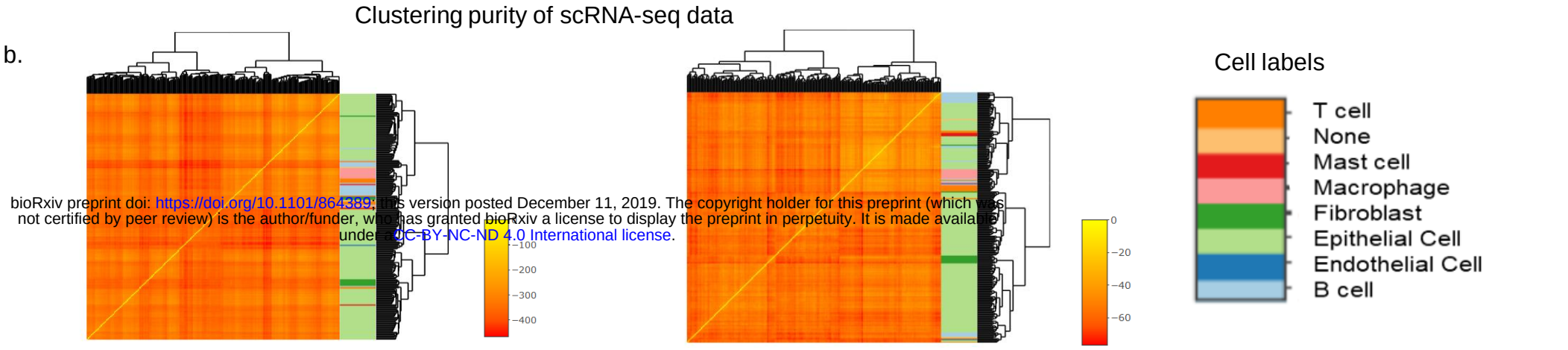


Fig. 3

a.



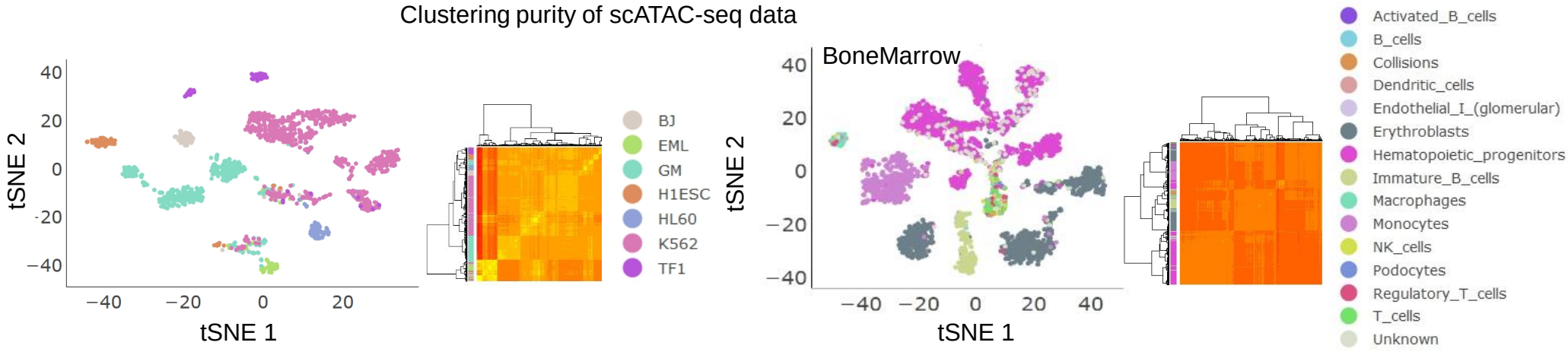
b.



Gene Expression : Cluster purity 84.65116 %

Pathway space : Cluster purity 86.51163 %

c.

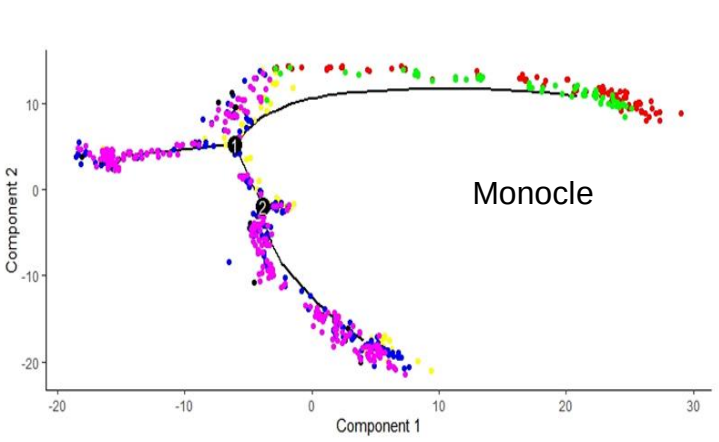


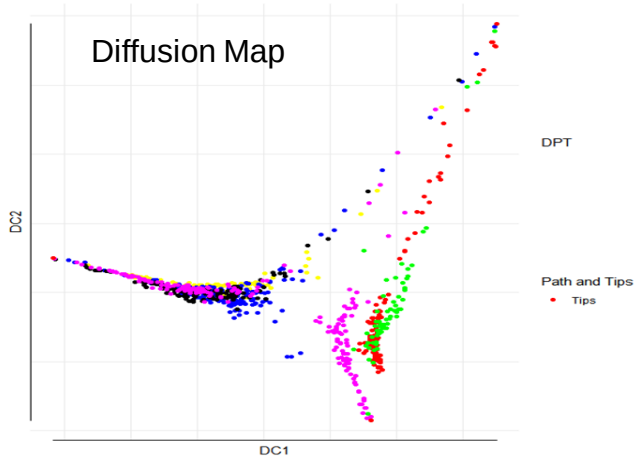
Pathway space : Cluster purity 79.10 %

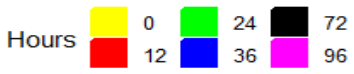
Pathway space : Cluster purity 87.25 %

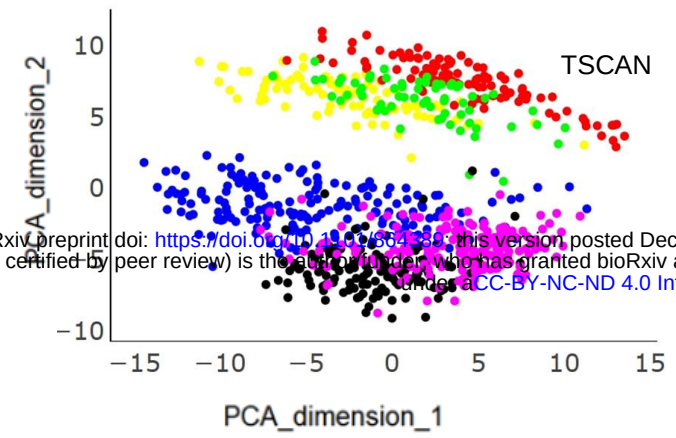
Fig. 4

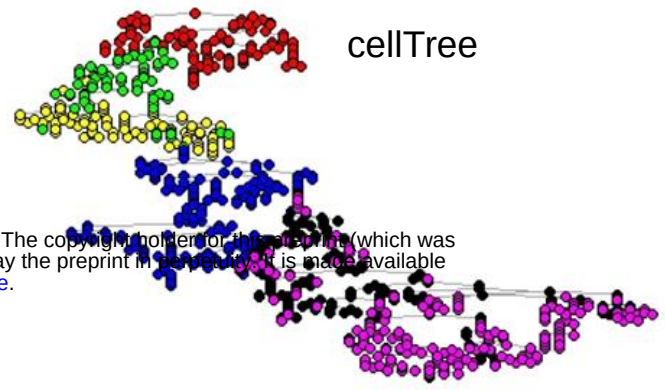
a Pseudo temporal ordering using gene-expression



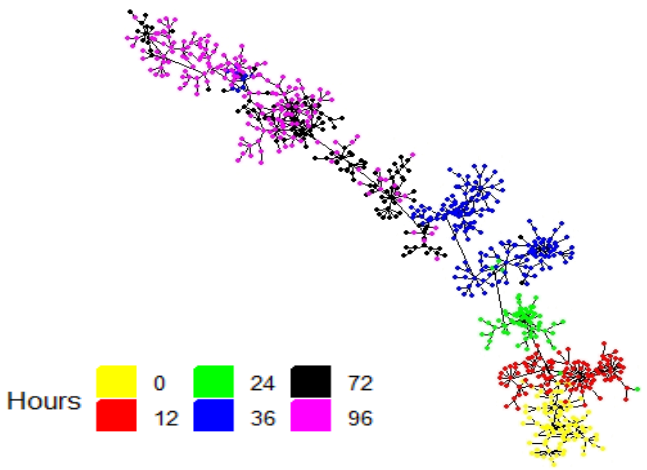




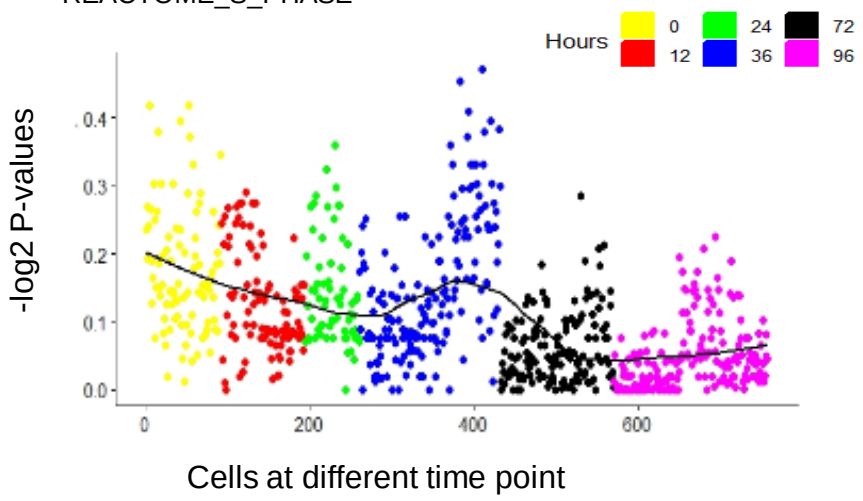




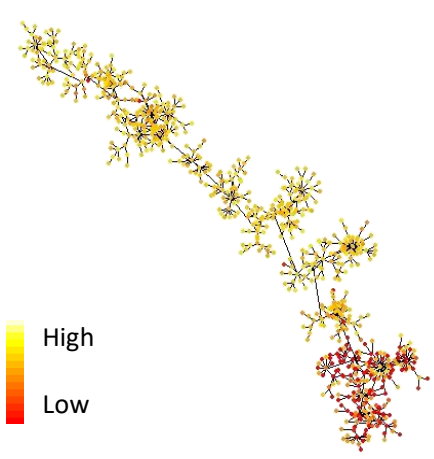
b Temporal by UniPath using pathway scores



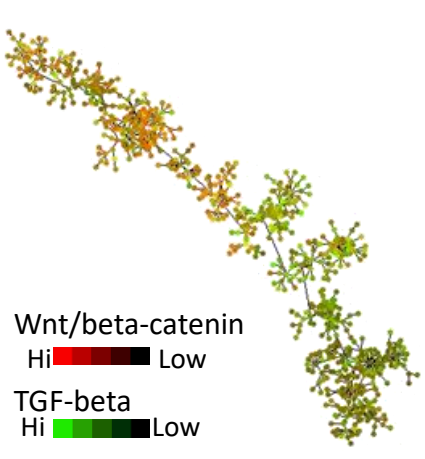
c REACTOME_S_PHASE



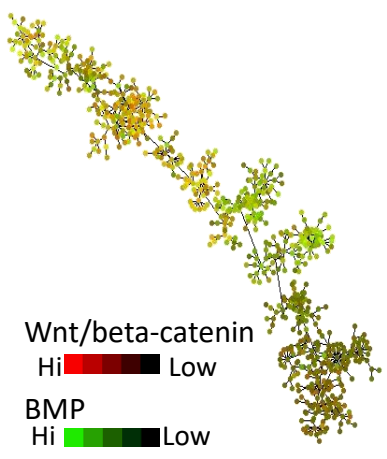
d



Endodermal scores



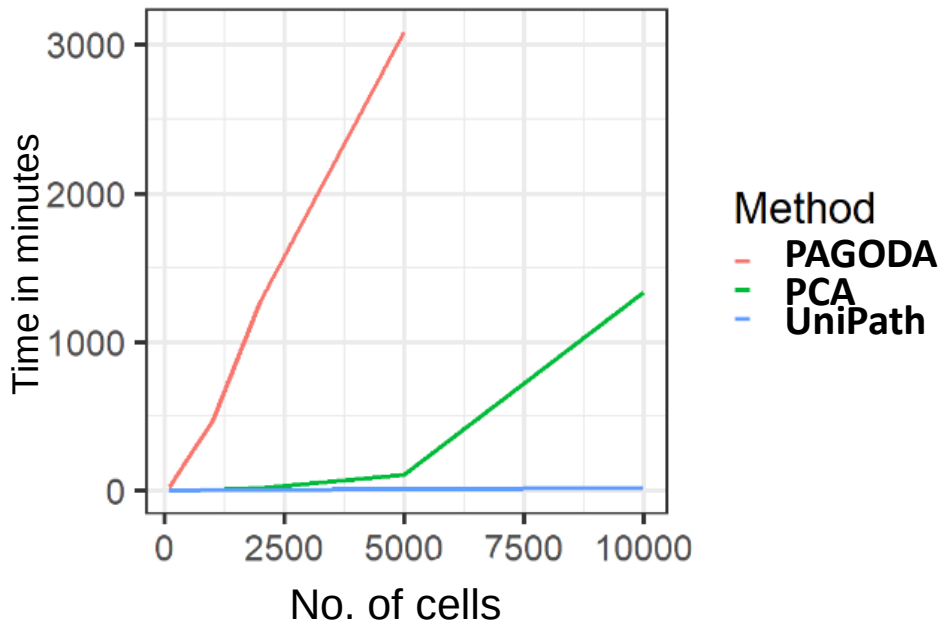
Cooccurrence of Wnt/beta-catenin and TGF-beta pathway scores



Cooccurrence of Wnt/beta-catenin and BMP pathway scores

Fig. 5

a



b

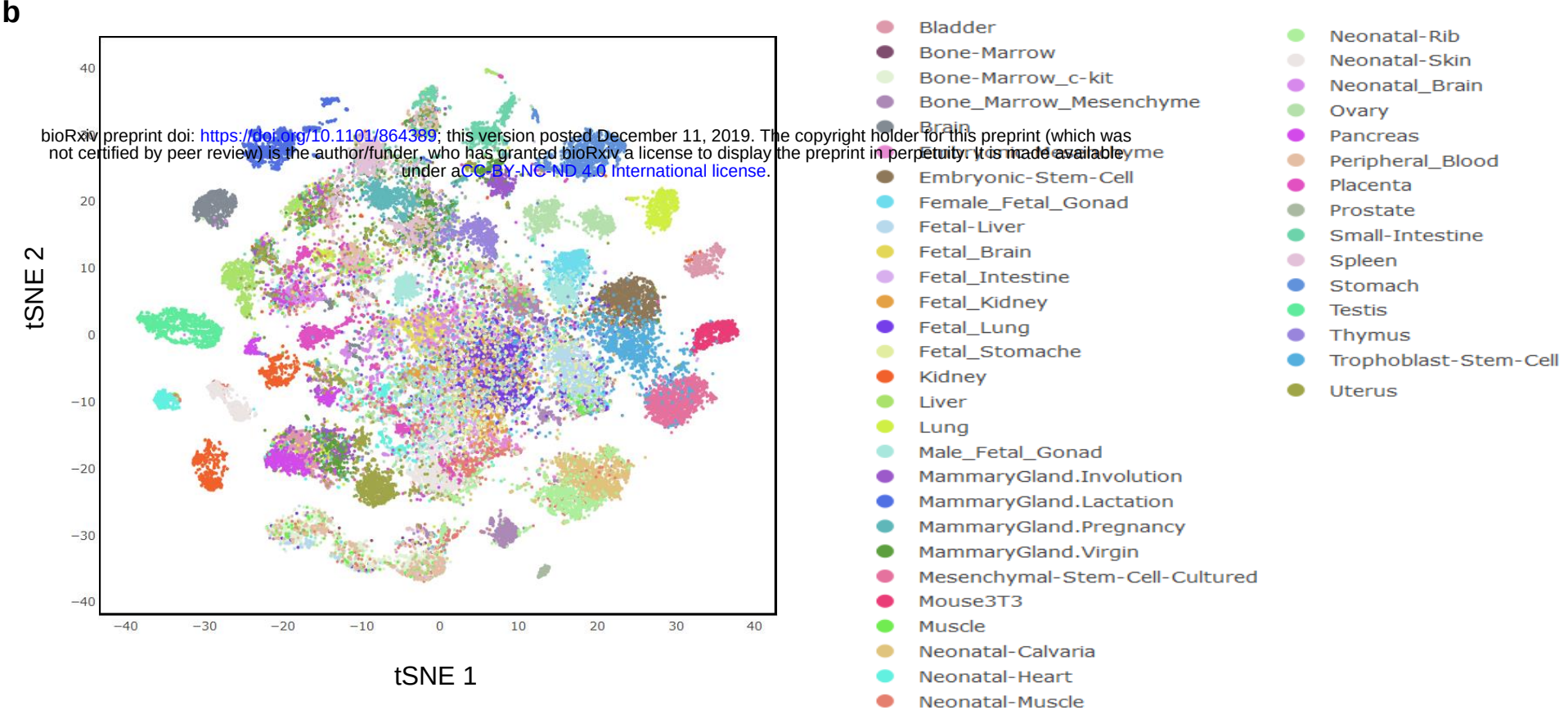
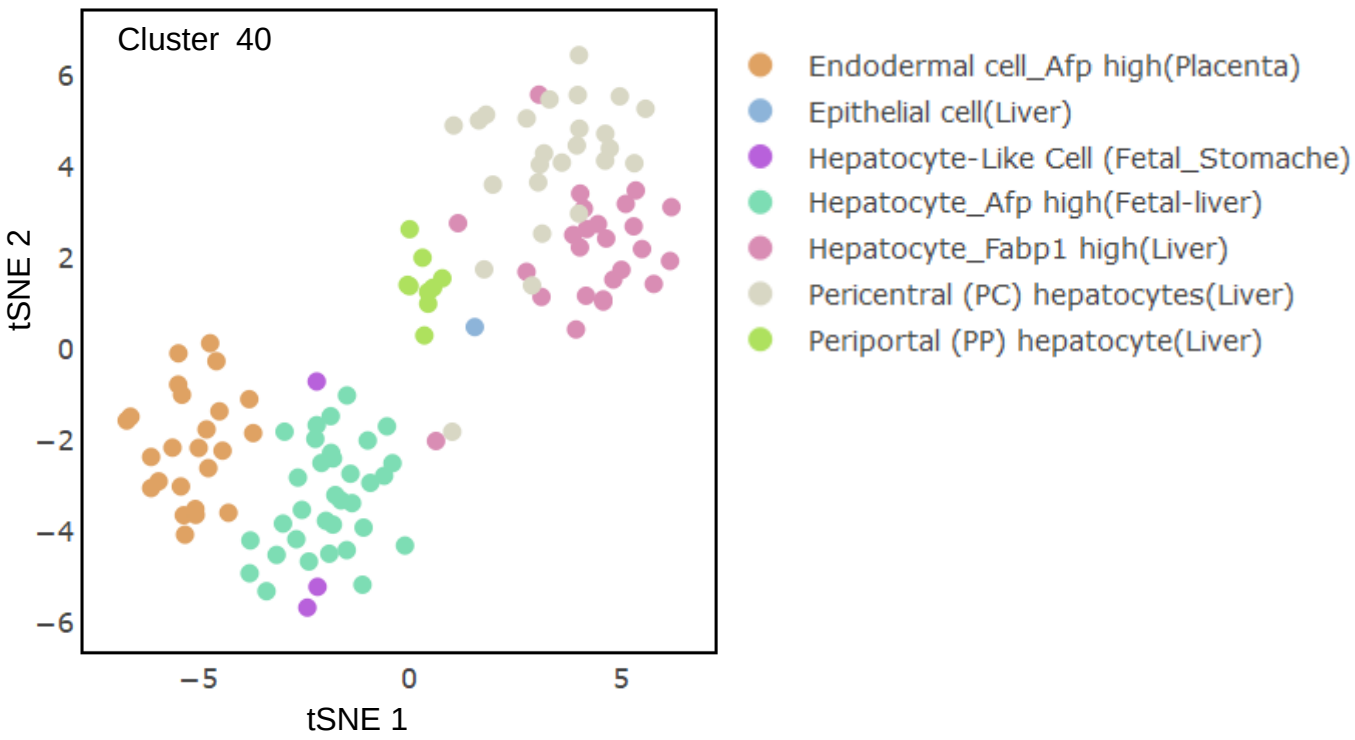
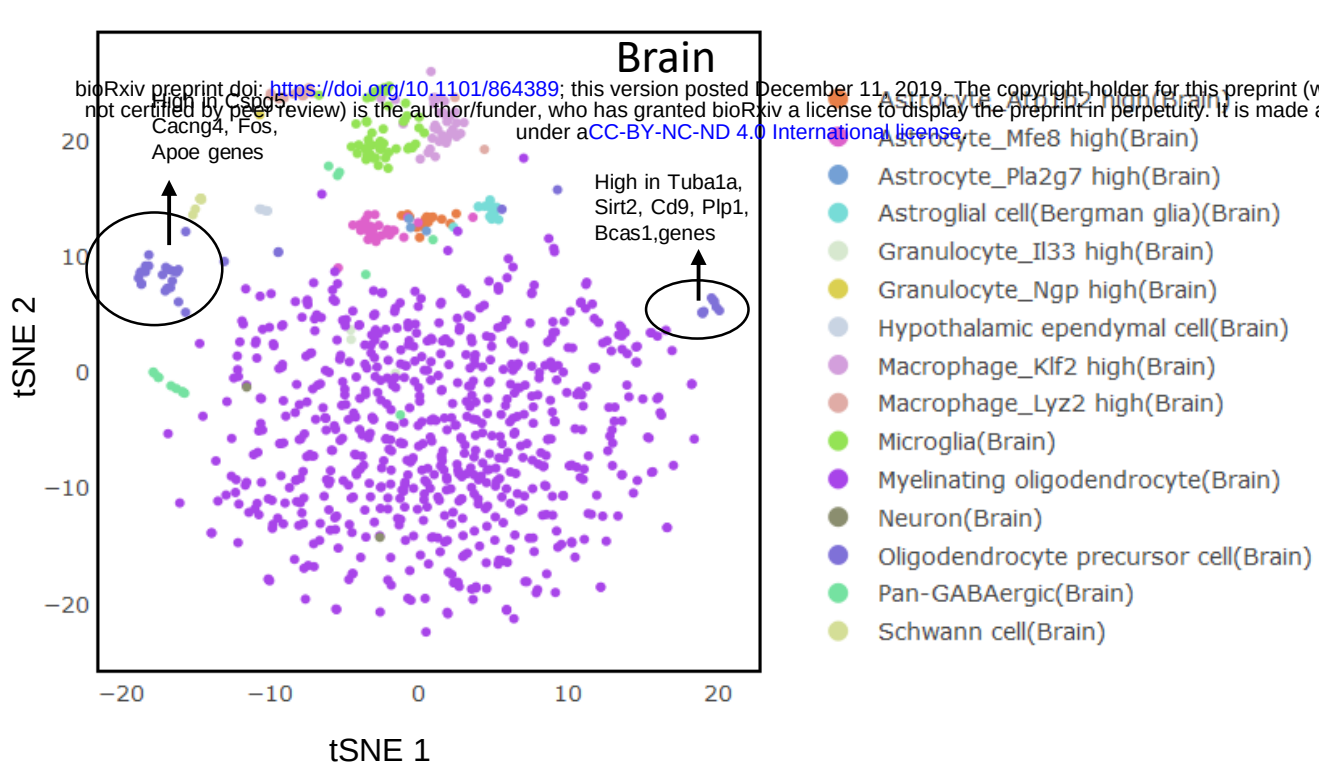


Fig. 6

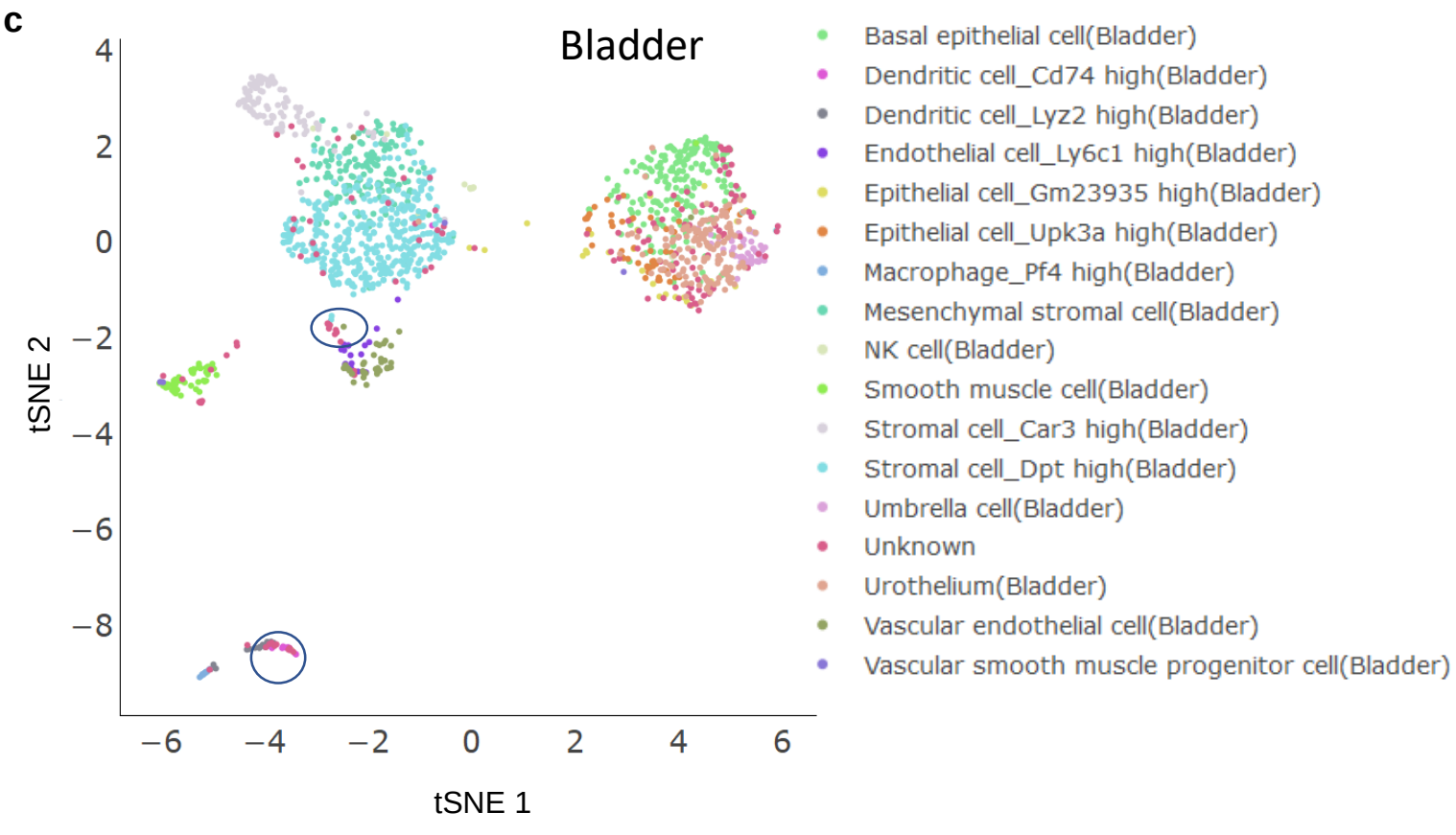
a



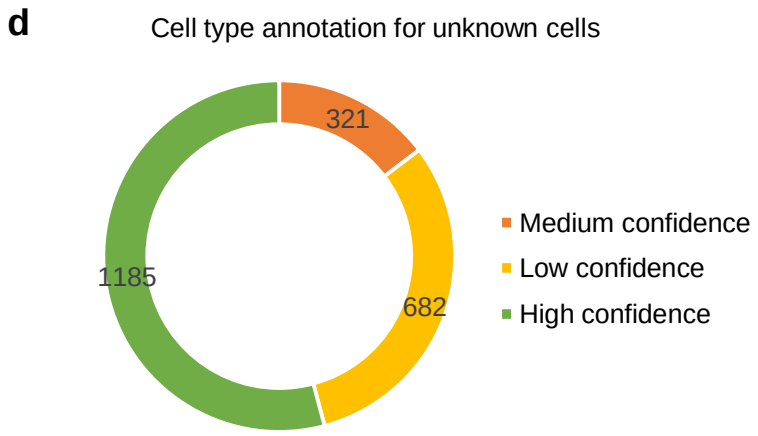
b

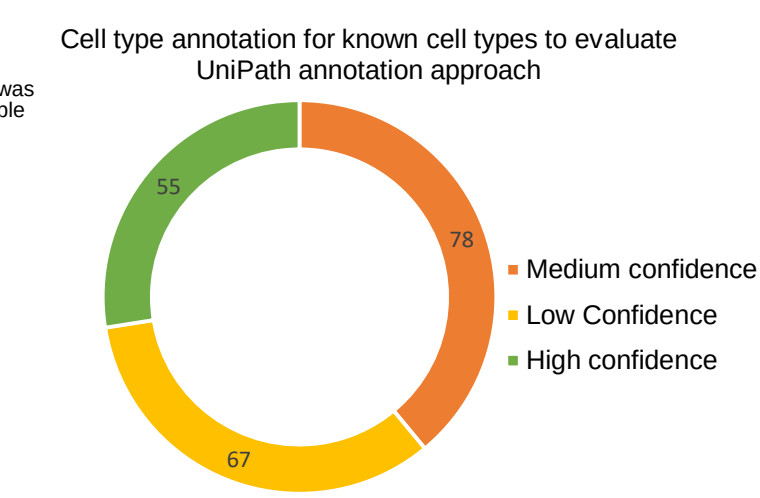


c



d

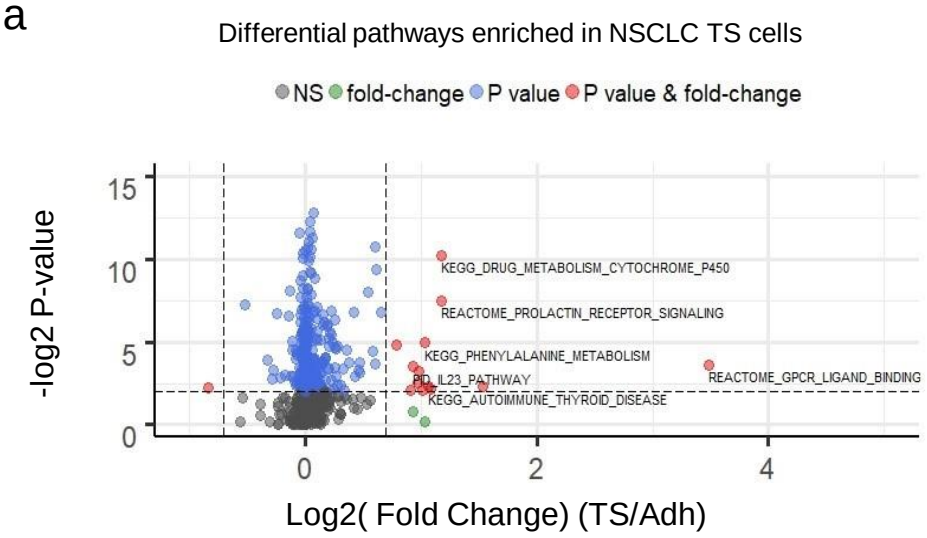




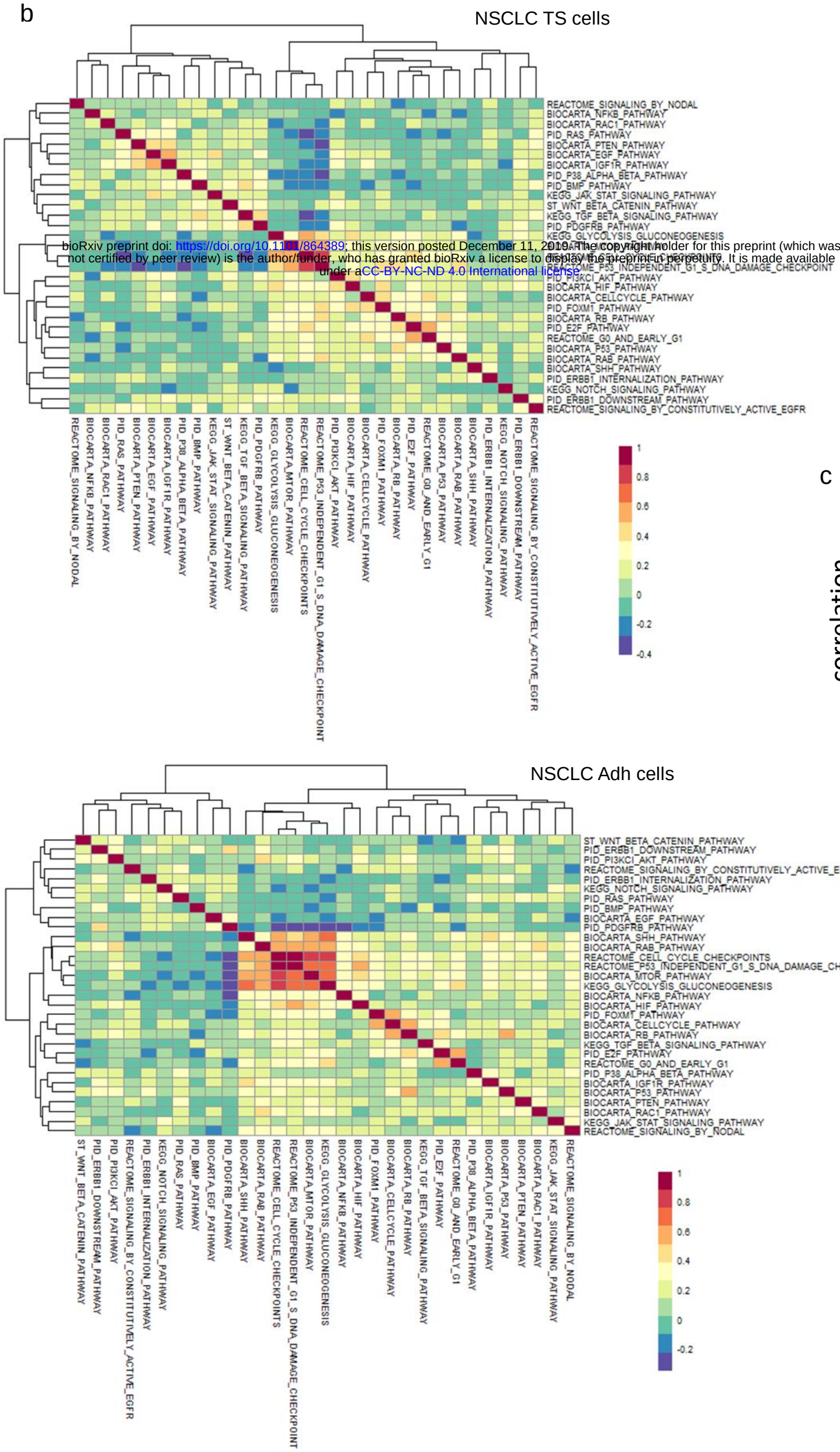
		Wrong Annotation	FDR
Medium confidence	78	1	0.01
Low Confidence	67	5	0.07
High confidence	55	0	0

Fig. 7

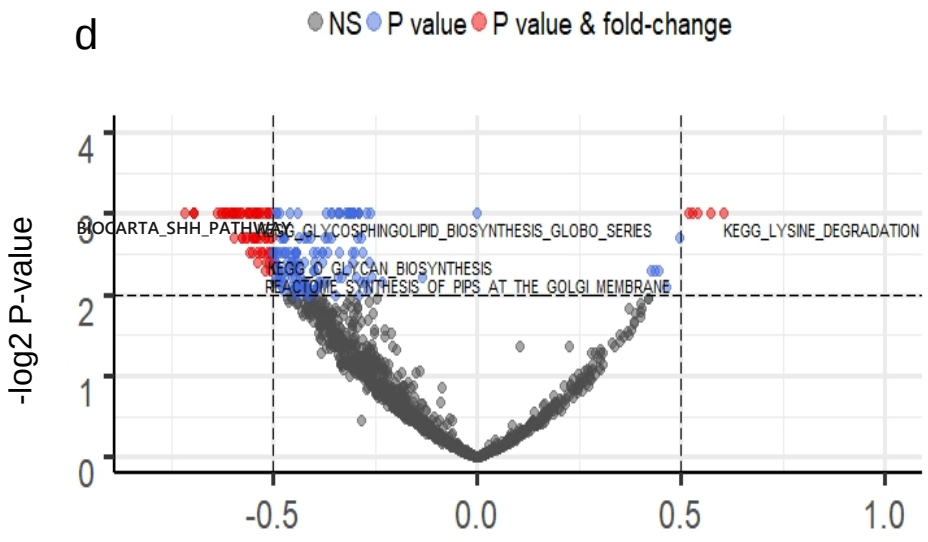
a



b

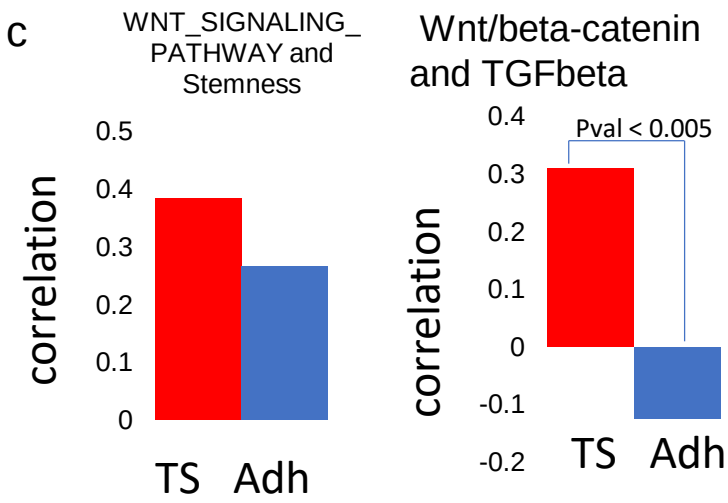


d

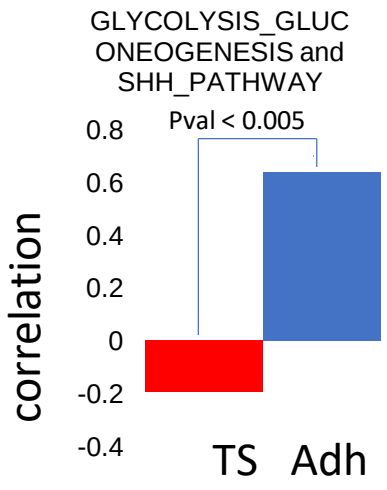


Differential cooccurrence with GLYCOLYSIS_GLUCONEOGENESIS

c



e



Supplementary Information

Supplementary Methods

Analysis of enrichment and cooccurrence pattern of of pathways in differentiating human embryonic stem cells

There is a big body of literature related to problem of studying level of enrichment and co-activity of signalling pathways using bulk gene-expression profile for differentiating stem cells. We found that, due to it's consistency, UniPath could provide a reliable solution to such problem by utilising heterogeneity among cells. Therefore, we analysed the pattern of gene-set scores of cells at different time of differentiation toward DE starting from hESC stage (Chu et al., 2016)[1]. We investigated correlation pattern for all possible pairs of pathways at different time points of differentiation. We manually searched and found literature support for few cooccurrence patterns observed. Such as, Nodal signalling is known to act via smad2/smad3 for differentiation towards mesendodermal lineage. It can be seen in supplementary Figure S8b that correlation of smad2/smad3 and Nodal signaling gene-set scores, was more at 12 and 24 hours then it decreased at the start of 36 hours which is consistent with existing literature[2-4]. Another example is about known synergistic effect of Nodal and Wnt signalling required for induction of mesendoderm and differentiation towards endoderm[5]. In our analysis also, correlation between Wnt and Nodal pathway score increases as differentiation proceeds towards definitive endoderm till 96 hours (supplementary Figure S8b). We also performed differential cooccurrence analysis for pathways at 6 time points of differentiation (supplementary Figure S9b). One of the example to be mentioned here is of mTOR and FGF signalling. FGF induces activation of mTOR and mTOR is involved in suppressing endodermal related activities. It could be seen in supplementary Figure S9b, that differential co-occurrence of FGF and mTOR pathway is significant at 0 and 12 hours, which is consistent with the literature that mTOR maintains ESCs pluripotency[6]. Another example is of BMP and Nodal signalling which have highest correlation at 36 hours before cells gain pure endodermal features. Wu et al.[7] have shown that combination of BMP and Nodal is involved in determining early embryonic stem cell lineages.The increase in enrichment score of Nodal signalling pathway[7] with differentiation towards endoderm is just as reported previously by several groups. The decrease in co-enrichment between BMP and Nodal signalling at 72 and 90 hours is according the finding reported by Kyel et al.[8] that BMP inhibits differentiation of primitive streak cells towards definitive endoderm. The co-occurrence analysis results supported by previous literature hint that systematic analysis with UniPath can reveal useful insight about dependencies among pathways as cells progress from one state to other.

Batch effect correction in UniPath

The framework of UniPath can reduce some amount of batch effect due to technical variations. However, beyond certain limit, the batch effect needs to be corrected. UniPath performs batch effect correction on combined P-values for pathways before adjusting them using null model. For batch correction of combined p-values, it uses standard Limma package[9]. It adjusts batch-corrected combined p-values further to achieve adjusted P-values. To demonstrate batch effect correction task (supplementary Figure S6b), we used scRNAseq datasets of mouse embryonic stem cell lines processed using various protocols by Ziegenhain et al. (GEO ID: GSE75790)[10].

We combined mESC dataset with scRNAseq profiles of hematopoietic stem cells from another study (GEO ID: GSE71794)[11] and used UniPath with batch correction mode.

Gene-sets used for pathway transformation

For transformation of gene expression data into pathway scores we have used two gene-set marker files from msigdb. First one is Cannonical gene set containing 1329 pathways and other one is GO biological process consisting of 4436 pathways. (http://software.broadinstitute.org/gsea/msigdb/collections.jsp#H)

Evaluation of UniPath using cell-type markers

We evaluated UniPath for accuracy of finding enrichment of known genuine gene-sets. For this purpose we used marker gene-sets for cell-types. We tested accuracy of UniPath using scRNAseq, scATACseq and bulk ATAC-seq data using appropriate cell marker gene-set. We curated cell makers from CellMarker[12] database and BioGPS[13]. Thus our list had more than 460 cell-type gene-sets. In case of scRNA-seq, both homogeneous and heterogeneous datasets[14-23] were used and percentage of cells with correct cell types among top 5 enriched terms was estimated. For scRNA-seq profiles UniPath was compared with PAGODA. In most of the cases UniPath was able highlight correct gene-set among top 5 enriched terms (Fig. 2a, supplementary Figure S1b-k). Details of study IDs along with cell types used for this evaluation are also provided in supplementary table S1.

For evaluating performance on scATACseq profile of K562 cells (from study GSE65360[24]) the set of positive gene-set included marker of Granulocyte, Leukemia-chronic Myelogenous K562, CD71-EarlyErythroid. We used 3 cell-type marker-set as K562 cell possess[21] properties of granulocytes and erythrocytes[25]. Other cell types used from same study were B cells. Other data-sets were also used for evaluating performance for bulk ATAC-seq profile as well as scATAC-seq (GEO ID: GSE68103[26] , GSE96769[27], GSE74912[28] bulk ATAC-seq).

Clustering purity estimation

Given true classes of samples as W = [w_1, w_2, …, w_k] and predicted clusters as C = [c_1, c_2, .., c_m] the clustering purity is calculated as

$$purity(W, C) = \frac{1}{N}\sum_{j=1}^{k} \max_j(w_k \cap c_j)$$

Parameters used for running PAGODA

In order to evaluate UniPath performance in detecting correct cell type and correctly enriched gene sets, we compared it with PAGODA using same marker-gene sets and pre-defined pathway gene sets (Fig. 2a). Data was pre-processed using clean.counts and then error models for each cell were fitted using one core with min.count.threshold of 2 and non-failed measurements per gene of at least 5 and value of k to be ¼ of number of cells in the data. It was followed by normalization of variance with the maximum adjusted variance of 5 and trim value of 3 divided by number of columns of the data. The function PAGODA.subtract.aspect was used as guided in the manual of PAGODA. Then we computed weighted first principle components scores for each

gene set. Using UnIPath and PAGODA enrichment scores, we computed the accuracy of cell type and pathway detection in top 5 enriched terms.

Doublet simulation and evaluation of UnIPath

UnIPath can help in detecting doublet cells in scRNA-seq data based on cell markers, given that we know possible cell-type coming together as doublet. We took 100 cells from each study id: Macrophages and Natural killer cells from study with GEO ID:GSE115978, microglial and endothelial cells from GEO ID: GSE67835 and B and T cells from GEO ID:GSE81861. Simulations on all the three studies were performed by taking average of each of the cell pairs. Accuracy for doublet detection was computed by counting number of times appropriate gene sets enriched for each of the cell pair together in top 5 terms.

Evaluation of Consistency of UnIPath

To ensure consistency of UnIPath pathway scores, we used profile published by Li et al. (GEO ID: GSE81861). First, we grouped B cells with T cells and estimated pathway enrichment using UnIPath and PAGODA and then we grouped B cells with epithelial cells and repeated the pathway score calculation (supplementary Figure S3) . For Fig. 2c, pathway scores of same B cell when grouped with T and epithelial cell were plotted. Pathway score calculated using UniPath were consistent. On the other hand, when this task was performed on same B-cell using PAGODA and plotted against each other, scores were quite varying. Similar task was performed using epithelial cells and changing its grouping partner, constant results were obtained with UniPath.

Parameters used for imputation and t-SNE of scATAC-seq

For evaluating clustering purity using pathway score, scATAC-seq (GEO ID: GSE65360) dataset were imputed using DrImpute. Imputed datasets were transformed into pathway space using UniPath. Singular value decomposition of pathway score matrix was performed. First 10 singular vectors were passed to Rtsne function from Rtnse R package with perplexity of 1000.

Data used for pseudo-temporal ordering

Chu et al (GEO ID: GSE75748)

The second dataset is time course differentiation data of human embryonic stem cells (HESCs) towards definitive endoderm (DE) [1](Fig. 4, Supplementary Figure S8-S9). The original study pin-pointed molecular mechanisms involved in formation of definitive endoderm. They have created a single cell trajectory tracing pluripotent state from mesendoderm to DE along with monitoring gene expression patterns in every stage. In our study we have analyzed scRNA-seq time course data having 758 cells. After transformation of data into pathway space and removal of pathways having 'cycle' word in term-names, we used temporal ordering function of UniPath. The parameter of number of clusters as set as 6, as there are 6 time points for the data-set. The parameter for number nearest neighbors was set as 4 (k=4). We also curated markers for ectoderm, mesendoderm, endoderm and mesoderm (https://www.bdbiosciences.com/documents/BD_Stem_Cell_Resource_poster.pdf), https://www.rndsystems.com/research-area/early-endodermal-lineage-markers. Using these markers -sets fo early developmental stages, UniPath facilitates viewing of lineage potency on the pseudo-temporally ordered tree (Fig. 4d).

Olfactory epithelium data (GEO ID: GSE95601)

Other dataset (supplementary S7a) we have used, is olfactory epithelium data for tracing trajectory of horizontal basal stem cells (HBCs) into different lineages[29]. This dataset consisted of 616 cells which were transformed into pathway scores and single cell trajectory was created using 13 clusters and k nearest neighbor 4. For pseudo-temporal ordering original data labels were used. In pathway space also we obtained neuronal and sustentacular lineages from HBCS via intermediate stages.

Other datasets used for pseudo-temporal ordering

We further determined temporal order of cells for data-sets by Trapnell et al. (GEO ID :GSE52529)[30]. Single cell trajectory from human skeletal muscle myoblasts to mature myoblasts were created using pathway scores (supplementary Figure S7). In other dataset (GEO ID: GSE52583)[31], mouse lung development trajectory was created using pathway scores (supplementary Figure S7). Another differentiation dataset from study by deng et al. (GEO ID:GSE45719)[32] was used for constructing preimplantation trajectory. For our analysis we have used 286 cells from this study and transformed them into pathway space and reconstructed developmental lineage from zygote to late-blast. We have used 10 clusters and K=5 nearest neighbors to shrink distance matrix to obtain a lineage tree labelled with original cell stage (supplementary Figure S7).

Pseudo-temporal ordering for scATAC-seq data

For tracing lineage of cardiac progenitor cells, scATAC-seq data published by Jia et al. (ENA ID: PRJEB23303) [33] was transformed into pathway space using global accessibility score. Single-cell lineage was constructing using 5 clusters and K=4 nearest neighbors (supplementary Figure S7e).

Parameters for temporal ordering

For pseudo temporal ordering, UniPath requires user to specify number of K nearest neighbor and number of classes.

Parameters for temporal ordering for different methods

Cell tree was ran with maptpx method and K.topics of 4 and log scale true. Monocle was ran using expression family of tobit and DDRtree with genes above threshold of 0.1 and getting expressed in at least 10 cells. TSCAN and Diffusion maps were run with their default settings.

Mouse single cell atlas analysis (GSE108097)

Mouse cell atlas data was transformed into pathway space using UniPath. Final data for down-stream analysis consisted of 49507 cells which had more than 800 genes with non-zero FPKM

value. First 50 principle components of pathway score matrix for 49507 cells, were passed to Rtsne function with perplexity of 1000 and maximum iterations of 1000. Using the output from Rtsne, we performed dbscan based clustering using the option of minimum points=50 for annotation of unlabeled cells. Clustering resulted in classification of 2518 as noise. Among remaining cells, 5590 were labelled as unknown in MCA dataset. Based sub-clusters information and UniPath's cell marker based cell detection we tried to annotate cells with "Unknown" label. We used three different kinds of marker files to annotate 2188 cells. Our annotation had 3 different confidence level. High confidence category included cells annotated using sub-clustering as well as by one of the marker files coming in top 5. On the other hand, medium confidence category involved cells which were detected in top 5 in any one of the two of the three marker files. Lastly low medium category cells were either coming as noise or were annotated by either one of the approach i.e. subclustering or marker based. We couldn't annotate remaining 3402 cells and tagged them in zero confidence category.

Source of lung cancer tumour cells and approvals

We have already mentioned that tumours sphere cells were provided by Wang et al. As mentioned in the manuscript published by Wang et al.,[34] we used a previosly characterized LC32 cells derived from resected primary non-small-cell lung cancer (NSCLC) adenocarcinoma samples. Their lab-grown version as non-adherent tumorspheres in serum-free medium is called as TS32. As mention in Wang et al., relevant ethical regulations pertaining to the IRB have been followed. The participating patient diagnosed with non-small-cell adenocarcinoma has signed written informed consent before surgical resection or biopsy.

Processing single-cell RNA-seq data of lung cancer cells

The raw read in FASTQ files achieved from sequencer were aligned to the human genome (hg19 assembly) using Tophat. Cuffdiff was applied to calculate FPKM using the aligned reads.

Supplementary Tables:

Table S1. Study ids and cell types used for cell detection along with the protocol used for processing of the data.

Study id	**Cell type**	**Heterogenous or Homogenous dataset**	**Total No. of target cells**	**Total No. of cells**	**Protocol**
GSE64016	ESC	Homogeneous	460	460	Fluidigm C1 platform.
GSE71858	ESC	Homogeneous	45	45	FRISCR and TritonX-100 Lysis
GSE73727	Beta cell	Heterogeneous	12	72	Single-cell RNA-seq
GSE44618	B cell	Homogeneous	62	62	SMART-seq
GSE98638	T cell	Homogeneous	198	5063	Smart-seq2 and Tang2010 protocol
GSE36552	ESC	Heterogeneous	34	123	Single cell RNA-seq technique (Tang protocol)
Zheng et. al	B cell	Homogeneous	128	128	10x genomics
GSE63818	Primordial germ cell	Heterogeneous	242	327	Single cell RNA-seq technique (Tang protocol)

GSE81861	B cell, T cell ,Macrophage, Epithelial cell	Heterogeneous	B cell: 18 T cell: 11 Macrophage:10 Epithelial cell:160	213	Fluidigm based single cell RNA-seq protocol
GSE67835	Endothelial, Microglial cells, Astrocytes	Heterogeneous	Endothelia:20 Microglial:15 Astrocytes:62	461	Single cell RNA-seq for Fluidigm C1 protocol

Table S2. The pathways which were added as spike-in with other gene-set of non-immune function for evaluatating performance of UniPath (in Fig. 2b)

Cell Type	Pathway Terms
T cell	KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY, ST_T_CELL_SIGNAL_TRANSDUCTION
B cell	ST_B_CELL_ANTIGEN_RECEPTOR, KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY

Table S3. Common terms among top 50 differential pathways (compared to all MCA cells) for AFP+ hepatocytes and AFP-high placental endodermal cells from cluster 40 of scMCA data (GSE108097).

	Pathways
1	GO_DIGESTION
2	GO_PHOSPHATIDYLCHOLINE_METABOLIC_PROCESS
3	GO_LIPOPROTEIN_METABOLIC_PROCESS
4	GO_PHOSPHATIDYLCHOLINE_BIOSYNTHETIC_PROCESS
5	GO_NEGATIVE_REGULATION_OF_PRODUCTION_OF_MOLECULAR_MEDIATOR_OF_IMMUNE_RESPONSE
6	GO_RETINOL_METABOLIC_PROCESS
7	GO_PROTEIN_LIPID_COMPLEX_ASSEMBLY
8	GO_CHOLESTEROL_EFFLUX
9	GO_MACROMOLECULAR_COMPLEX_REMODELING
10	GO_PLASMA_LIPOPROTEIN_PARTICLE_CLEARANCE
11	GO_CELLULAR_HORMONE_METABOLIC_PROCESS
12	GO_REGULATION_OF_INTERLEUKIN_8_PRODUCTION
13	GO_NEGATIVE_REGULATION_OF_LIPASE_ACTIVITY
14	GO_REGULATION_OF_LIPASE_ACTIVITY
15	GO_NEUTRAL_LIPID_CATABOLIC_PROCESS
16	GO_PROTEIN_LIPID_COMPLEX_SUBUNIT_ORGANIZATION
17	GO_REGULATION_OF_LIPID_TRANSPORT
18	GO_DIGESTIVE_SYSTEM_PROCESS
19	GO_NEGATIVE_REGULATION_OF_CYTOKINE_SECRETION
20	GO_POSITIVE_REGULATION_OF_LIPID_CATABOLIC_PROCESS
21	GO_TRIGLYCERIDE_CATABOLIC_PROCESS
22	GO_REGULATION_OF_STEROL_TRANSPORT

Table S4. Differential cooccurrence of few pathways in NSCLC data along with their correlation value, statistical significance of coenrichement (P-value), difference of the pathways in TS and Adh cells. Jaccard index represneting overlap of genes among the pathways are also shown in the table.

Coenriched pathways	Correlation	P-value	Difference (TS - Adh)	Jaccard Index
ST_WNT_BETA_CATENIN_PATHWAY-KEGG_TGF_BETA_SIGNALING_PATHWAY	0.3097043	0.009	0.4347173	0
ST_WNT_BETA_CATENIN_PATHWAY-KEGG_NOTCH_SIGNALING_PATHWAY	0.1527704	0.513	-0.1061958	0.01265823
BIOCARTA_SHH_PATHWAY-KEGG_GLYCOLYSIS_GLUCONEOGENESIS	-0.1935556	0	-0.8306809	0

Additional file 1. This .xlsx file contains 64 sheets representing 64 clusters along with the cells belonging to each cluster of single-cells from mouse cell atlas (GSE108097).

Additional file 2. Wilcoxon Rank sum test based differential pathways between AFP+ placental endodermal cells of subcluster 40 of mouse single cell atlas dataset (GSE108097) and rest of the cells.

Additional file 3. Wilcoxon Rank sum test based differential pathways between AFP+ fetal liver hepatocyte of subcluster 40 of mouse single cell atlas dataset (GSE108097) and rest of the cells.

Additional file 4. This file contains ell type annotations of Unknown and Known cells (for FDR) of mouse single cell atlas dataset (GSE108097).

Adiitional file 5. Wilcoxon Rank sum test based differential pathways between NSCLC and NSCLC Adherent(Adh) cells.

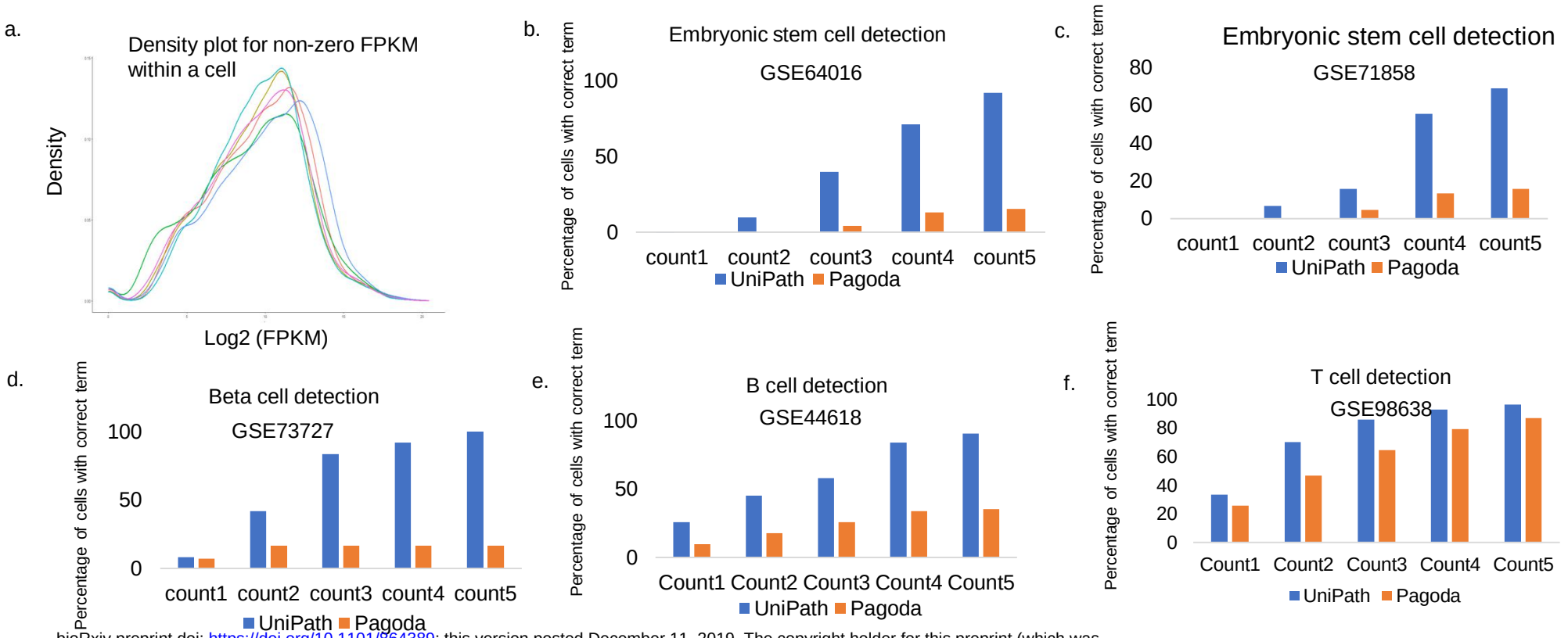
References

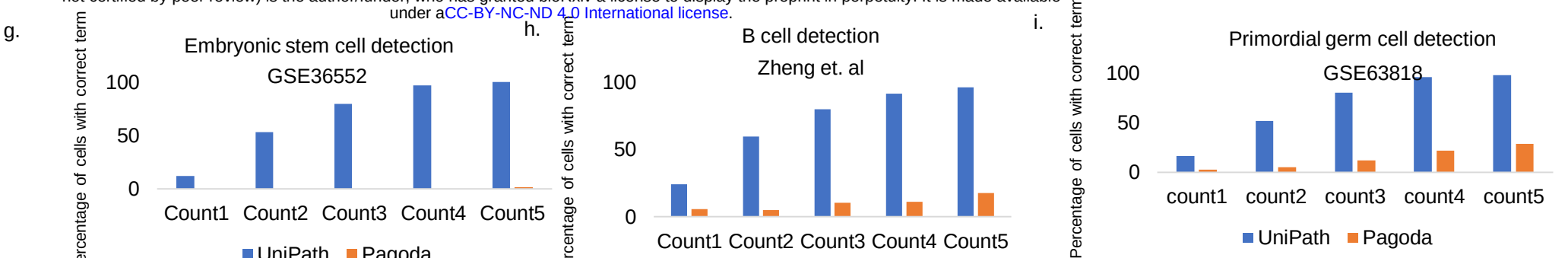
1. Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, Choi J, Kendziorski C, Stewart R, Thomson JA: **Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm.** *Genome biology* 2016, **17:**173.
2. Senft AD, Costello I, King HW, Mould AW, Bikoff EK, Robertson EJ: **Combinatorial Smad2/3 activities downstream of Nodal signaling maintain embryonic/extra-embryonic cell identities during lineage priming.** *Cell reports* 2018, **24:**1977-1985. e1977.
3. Fei T, Zhu S, Xia K, Zhang J, Li Z, Han J-DJ, Chen Y-G: **Smad2 mediates Activin/Nodal signaling in mesendoderm differentiation of mouse embryonic stem cells.** *Cell research* 2010, **20:**1306.
4. Brown S, Teo A, Pauklin S, Hannan N, Cho CHH, Lim B, Vardy L, Dunn NR, Trotter M, Pedersen R: **Activin/Nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors.** *Stem cells* 2011, **29:**1176-1185.
5. Chhabra S, Liu L, Goh R, Warmflash A: **The timing of signaling events in the BMP, WNT, and Nodal cascade determines self-organized fate patterning in human gastruloids.** *bioRxiv* 2018**:**440164.
6. Zhou J, Su P, Wang L, Chen J, Zimmermann M, Genbacev O, Afonja O, Horne MC, Tanaka T, Duan E: **mTOR supports long-term self-renewal and suppresses mesoderm and endoderm activities of human embryonic stem cells.** *Proceedings of the National Academy of Sciences* 2009, **106:**7840-7845.
7. Wu Z, Zhang W, Chen G, Cheng L, Liao J, Jia N, Gao Y, Dai H, Yuan J, Cheng L: **Combinatorial signals of activin/Nodal and bone morphogenic protein regulate the early lineage segregation of human embryonic stem cells.** *Journal of Biological Chemistry* 2008, **283:**24991-25002.
8. Loh KM, Ang LT, Zhang J, Kumar V, Ang J, Auyeong JQ, Lee KL, Choo SH, Lim CY, Nichane M: **Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations.** *Cell stem cell* 2014, **14:**237-252.

9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic acids research* 2015, **43:**e47-e47.
10. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W: **Comparative analysis of single-cell RNA sequencing methods.** *Molecular cell* 2017, **65:**631-643. e634.
11. Lu Y-F, Cahan P, Ross S, Sahalie J, Sousa PM, Hadland BK, Cai W, Serrao E, Engelman AN, Bernstein ID: **Engineered murine HSCs reconstitute multi-lineage hematopoiesis and adaptive immunity.** *Cell reports* 2016, **17:**3178-3192.
12. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M: **CellMarker: a manually curated resource of cell markers in human and mouse.** *Nucleic acids research* 2018, **47:**D721-D728.
13. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A: **The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins.** *Database* 2016, **2016**.
14. Leng N, Chu L-F, Barry C, Li Y, Choi J, Li X, Jiang P, Stewart RM, Thomson JA, Kendziorski C: **Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments.** *Nature methods* 2015, **12:**947.
15. Thomsen ER, Mich JK, Yao Z, Hodge RD, Doyle AM, Jang S, Shehata SI, Nelson AM, Shapovalova NV, Levi BP: **Fixed single-cell transcriptomic characterization of human radial glial diversity.** *Nature methods* 2016, **13:**87.
16. Li J, Klughammer J, Farlik M, Penz T, Spittler A, Barbieux C, Berishvili E, Bock C, Kubicek S: **Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types.** *EMBO reports* 2016, **17:**178-187.
17. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ: **From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing.** *Genome research* 2014, **24:**496-510.
18. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, Kang B, Hu R, Huang JY, Zhang Q: **Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing.** *Cell* 2017, **169:**1342-1356. e1316.
19. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J: **Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells.** *Nature structural & molecular biology* 2013, **20:**1131.
20. Guo F, Yan L, Guo H, Li L, Hu B, Zhao Y, Yong J, Hu Y, Wang X, Wei Y: **The transcriptome and DNA methylome landscapes of human primordial germ cells.** *Cell* 2015, **161:**1437-1452.
21. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su M-J, Melms JC, Leeson R, Kanodia A, Mei S, Lin J-R: **A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade.** *Cell* 2018, **175:**984-997. e924.
22. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS: **Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors.** *Nature genetics* 2017, **49:**708.
23. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Gephart MGH, Barres BA, Quake SR: **A survey of human brain transcriptome diversity at the single cell level.** *Proceedings of the National Academy of Sciences* 2015, **112:**7285-7290.
24. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ: **Single-cell chromatin accessibility reveals principles of regulatory variation.** *Nature* 2015, **523:**486.

25. Guo Q, Zhang L, Li F, Jiang G: **The plasticity and potential of leukemia cell lines to differentiate into dendritic cells.** *Oncology letters* 2012, **4:**595-600.
26. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J: **Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing.** *Science* 2015, **348:**910-914.
27. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ: **Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation.** *Cell* 2018, **173:**1535-1548. e1516.
28. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ: **Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.** *Nature genetics* 2016, **48:**1193.
29. Fletcher RB, Das D, Gadye L, Street KN, Baudhuin A, Wagner A, Cole MB, Flores Q, Choi YG, Yosef N: **Deconstructing olfactory stem cell trajectories at single-cell resolution.** *Cell stem cell* 2017, **20:**817-830. e818.
30. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL: **The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.** *Nature biotechnology* 2014, **32:**381.
31. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR: **Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.** *Nature* 2014, **509:**371.
32. Deng Q, Ramsköld D, Reinius B, Sandberg R: **Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells.** *Science* 2014, **343:**193-196.
33. Jia G, Preussner J, Chen X, Guenther S, Yuan X, Yekelchyk M, Kuenne C, Looso M, Zhou Y, Teichmann S: **Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement.** *Nature communications* 2018, **9:**4877.
34. Wang Z, Yip LY, Lee JHJ, Wu Z, Chew HY, Chong PKW, Teo CC, Ang HY-K, Peh KLE, Yuan J: **Methionine is a metabolic dependency of tumor-initiating cells.** *Nature medicine* 2019, **25:**825.

Figure S1





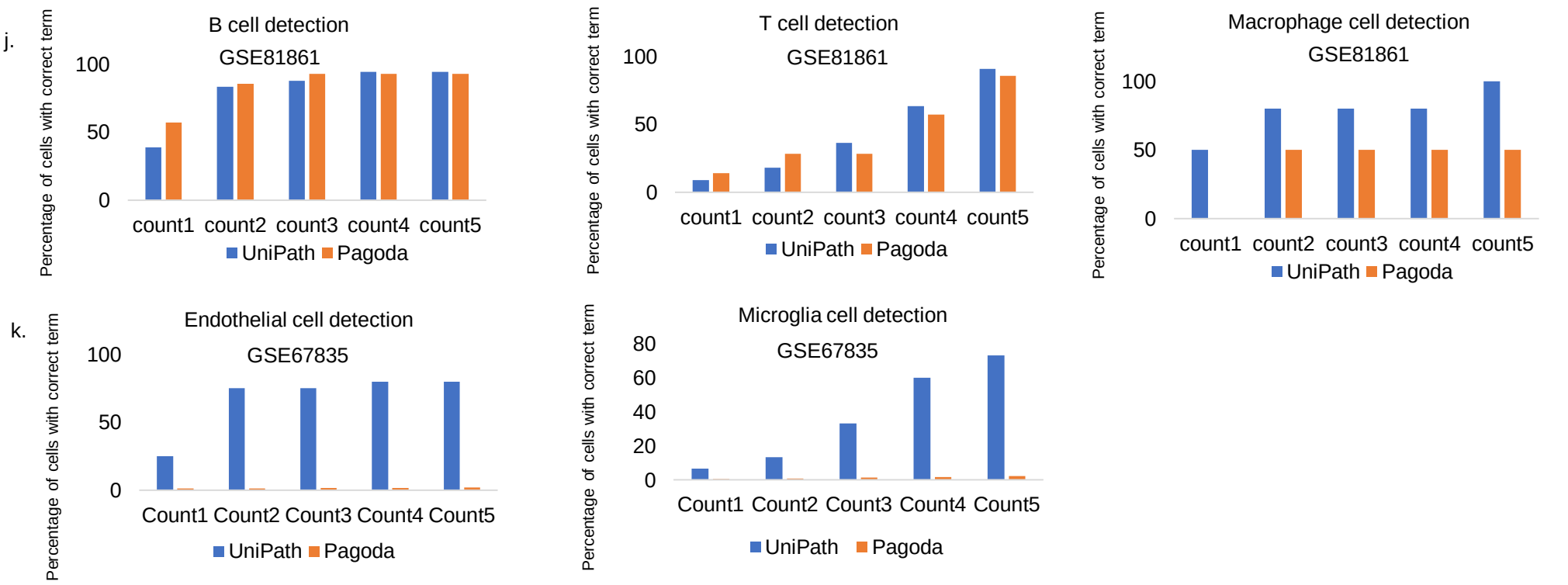
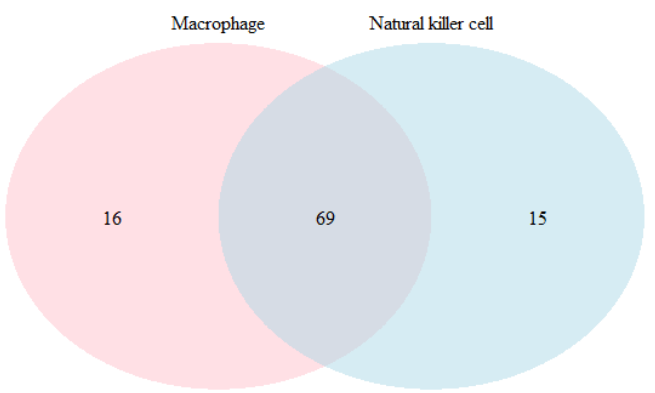
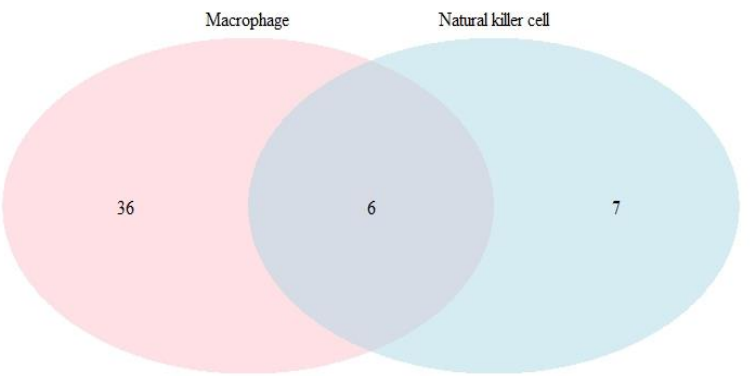


Figure S1. Comprehensive evaluation of UniPath for single-cell RNAseq for highlighting correct terms in top enriched results. The terms here are cell-types and gene-sets of terms are set of marker genes for corresponding cell-type. **(a)** Density plot for non-zero FPKM in a cell using single cell RNA seq data (GSE52529). **(b)** Embryonic stem cell (ESC) percentage detected in homogeneous dataset hESC scRNAseq (Fluidigm C1 platform). The bars show percentage of cells with correct cell-type among top enriched terms. 'count1' shows percent of cells with correct cell-type as the first enriched term. Similarly count5 shows percentage of cells with correct cell-type among top-5 enriched terms. **(c)** Accuracy for cell-type detection in homogeneous dataset of ESC processed using FRISCR and TritonX-100 Lysis. **(d)** correct cell-type detection percentage for heterogeneous dataset of Beta Cell **(e)** correct cell-type detection percentage for homogeneous dataset of B-cell (Smart-seq protocol.) **(f)** percentage of correct detection in homogeneous scRNAseq dataset of T-cell (Smartseq2 and Tang et al., 2010 protocol). (**g**). For heterogeneous dataset of Embryonic stem cell (Tang protocol). **(h)** correct detection in homogeneous dataset of B-cell **(i)** Primordial germ cell ; heterogeneous dataset (Tang protocol). **(j)** Cell type detection for heterogeneous dataset of B-cell, T-cell and macrophages (fluidigm based scRNA-seq protocol). **(k)** Detection of cell-types from scRNAseq profiles of Endothelial and Microglial cell processed using fluidigm C1 based protocol.

Figure S2

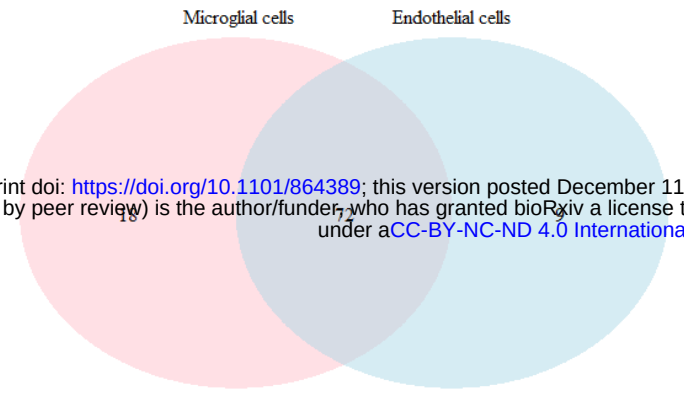


UniPath



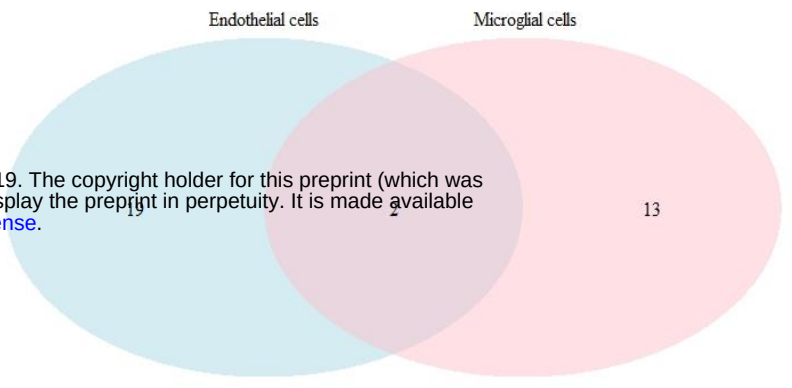
PAGODA

GSE115978

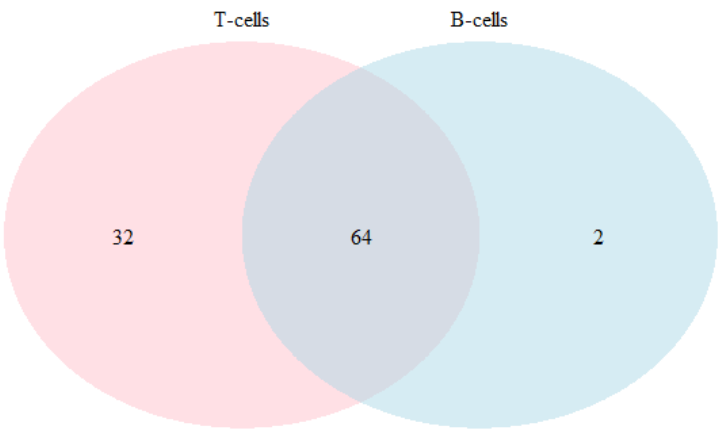


UniPath

GSE67835

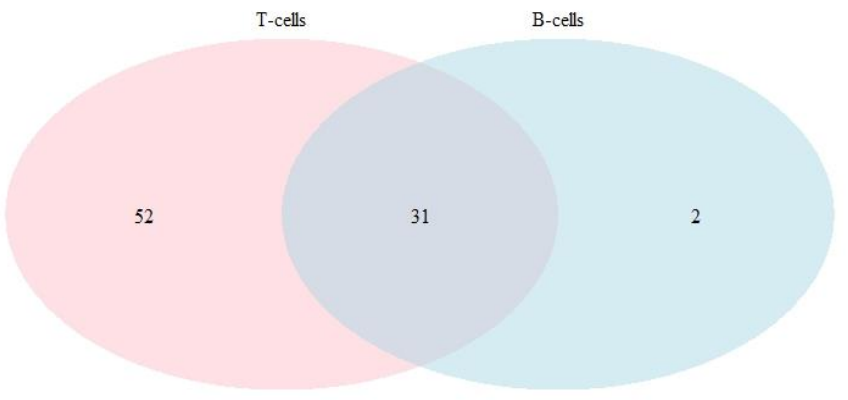


PAGODA



UniPath

GSE81861



PAGODA

Figure S2: Evaluation of UniPath for detection of doublet. 100 cells from each of the study with GEO ID: GSE115978, GSE67835 and GSE81861, were used for simulating doublets by averaging two cells together. Macrophage and Natural killer cells doublets were detected with 69% accuracy, Microglial and endothelial cells doublets were detected with 72% and T and B cells doublets were detected with 64% accuracy. Whereas for PAGODA the accuracy of detection of doublet is lower than UniPath.

Figure S3

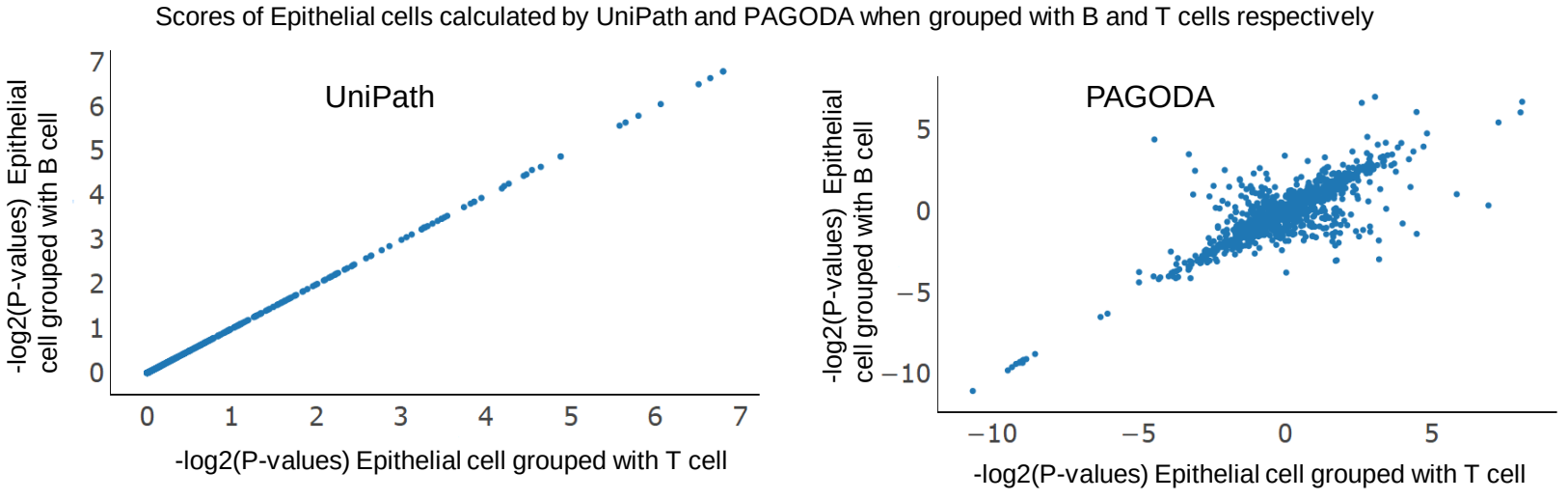


Figure S3: Test of Consistency of UniPath and PAGODA for enrichment of pathways in epithelial cell when it is grouped with B cells or T cells. The data-set used here was adapted from study with GEO ID: GSE81861. Output of UniPath for a cell is consistent and is not affected by the type of neighboring cells. Whereas for PAGODA the estimate of dispersion (equivalent to enrichment) for pathway is dependent upon composition of cell-type in the data-set.

Figure S4

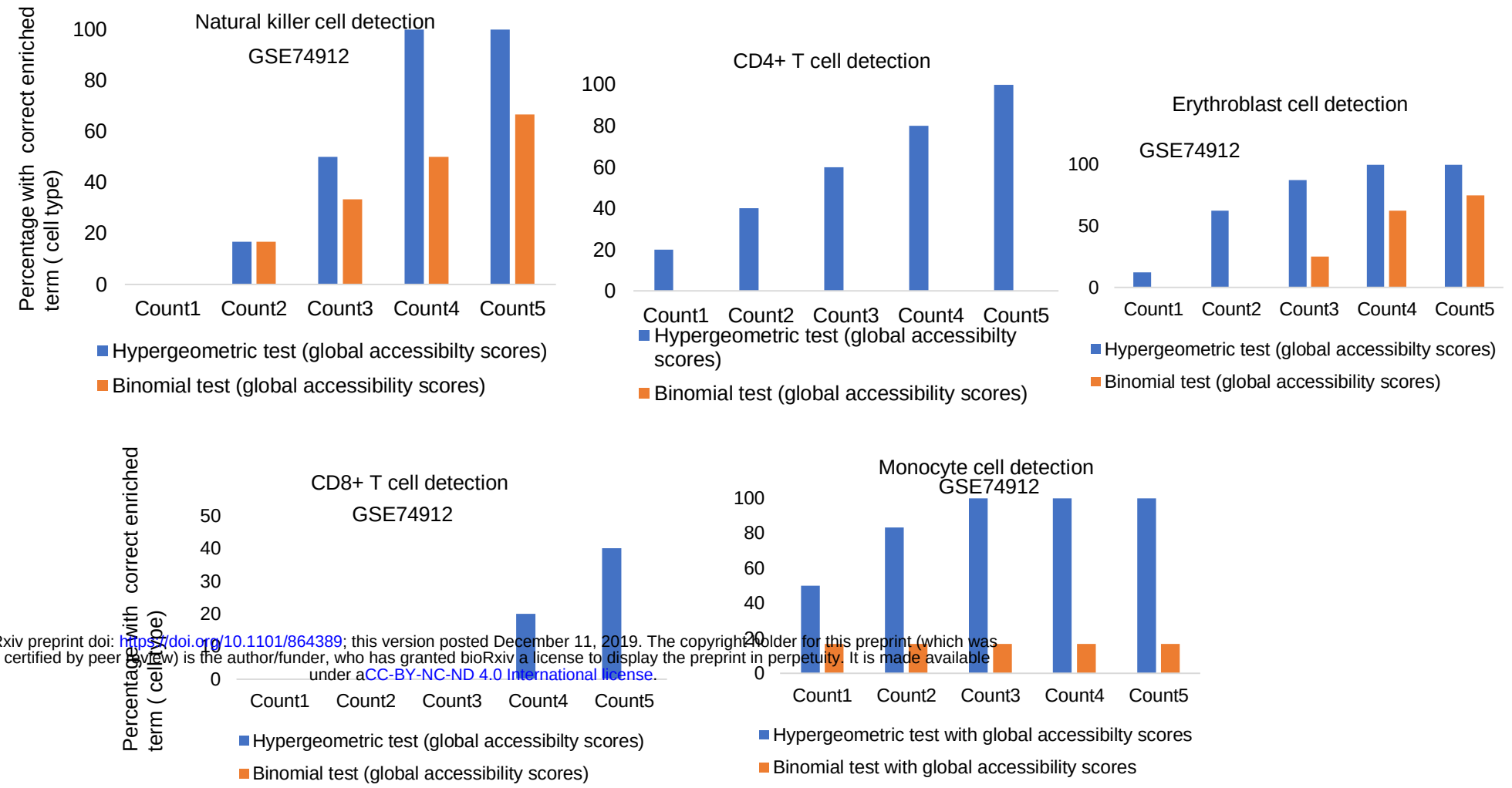


Figure S4: Evaluation of accuracy of UniPath for estimating gene-set enrichment using open-chromatin profile of bulk sample of cell lines. The gene-sets of marker for cell-types are used here for benchmarking. Shown here is percentage of cells with correct cell-type gene-set among top enriched terms. ‘count1’ shows percent of cells with correct cell-type as the first enriched term. Similarly, count3 shows percentage of cells with correct cell-type among top-3 enriched terms. The bulk ATAC-seq were adapted from study with (GEO ID: GSE74912). Enrichment score for Natural killer cell, CD4+ T cell, CD8+ T cell, Monocyte and Erythroblast cells were calculated by UniPath using hypergeometric and binomial test. For every cell-type results are shown for both situation, when enhancers were enriched using division global accessibility scores.

Figure S5

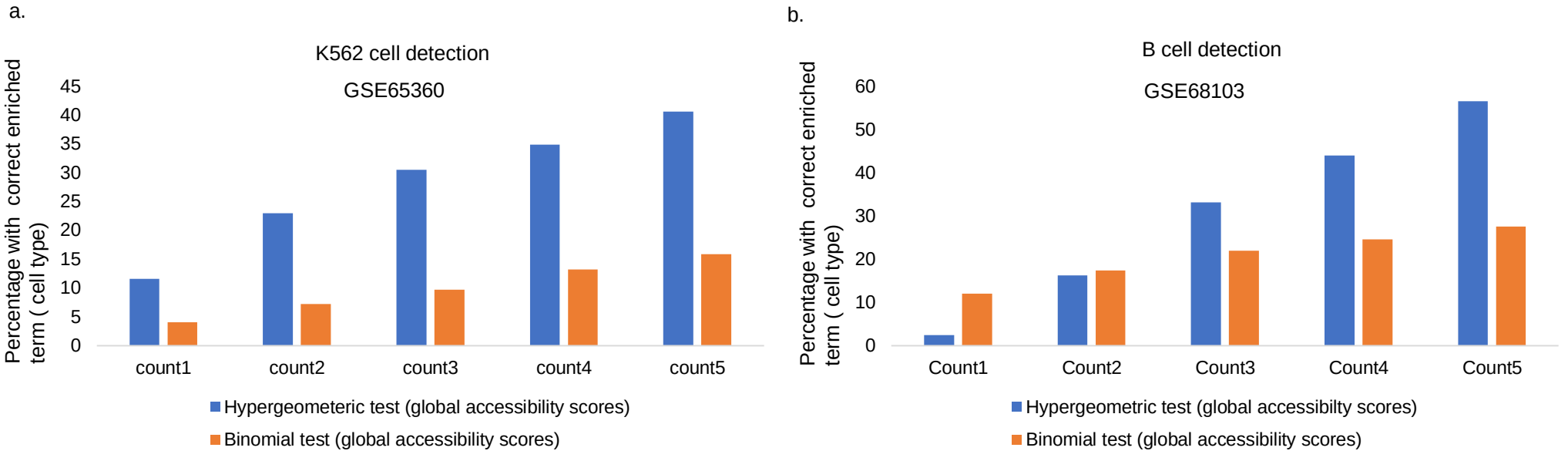
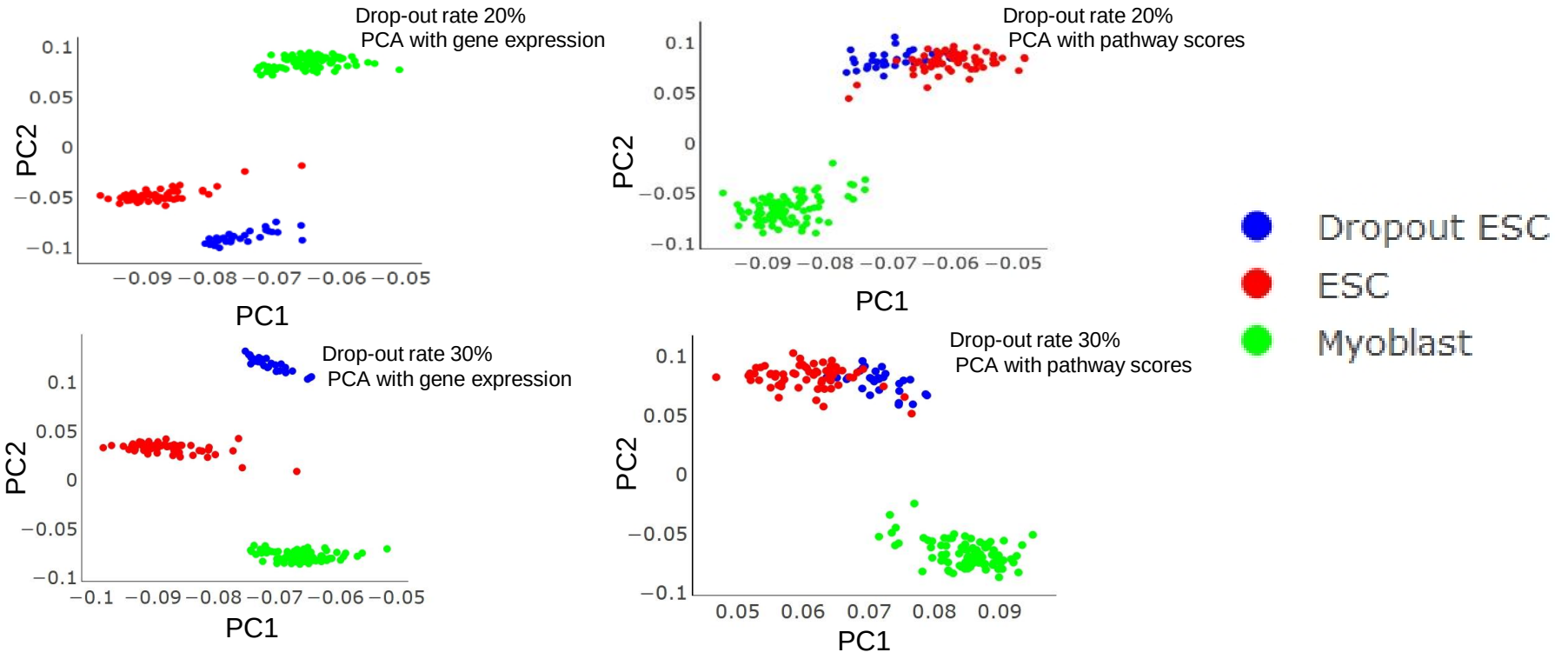


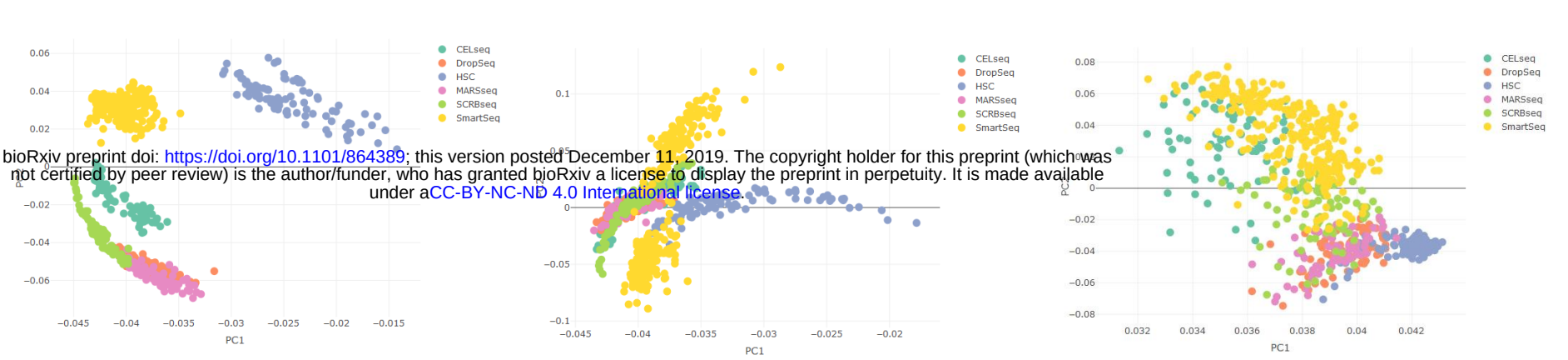
Figure S5: Accuracy of UniPath for enrichment of correct gene-set in top results for single cell open chromatin profile. Evaluation of accuracy is done using marker gene-set for cell-types while querying ChIP-atlas or scATAC-seq profiles. **(a)** K562 cell detection using hypergeometric test and binomial test using enhancer highlighted using global accessibility scores (data set GEO ID: GSE65360). **(b)** percentage of Correct cell-type detection using scATAC-seq of B-cell (GEO ID: GSE68103).

Figure S6

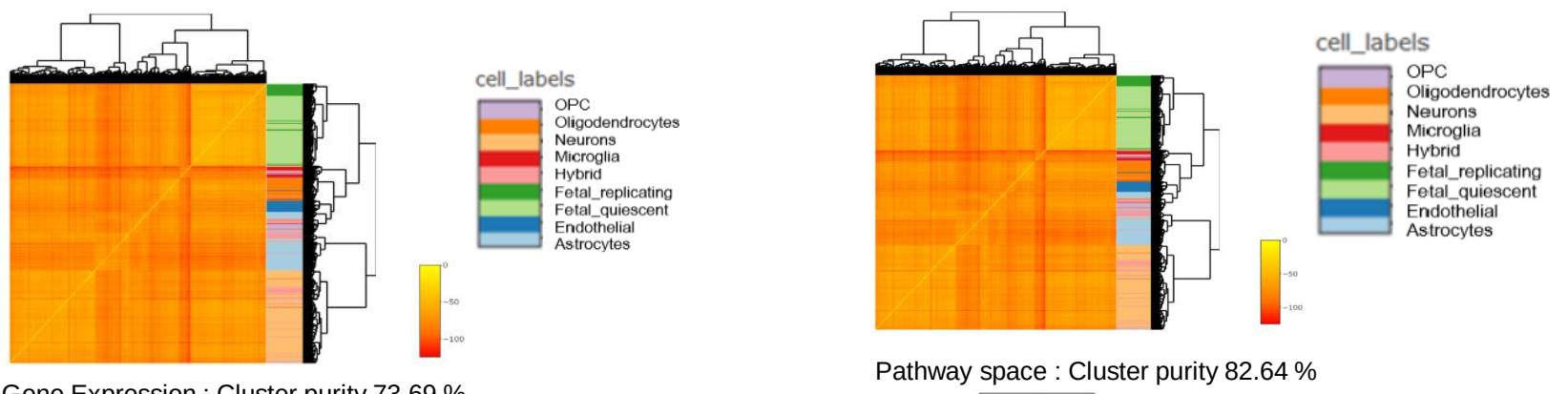
a.



b.



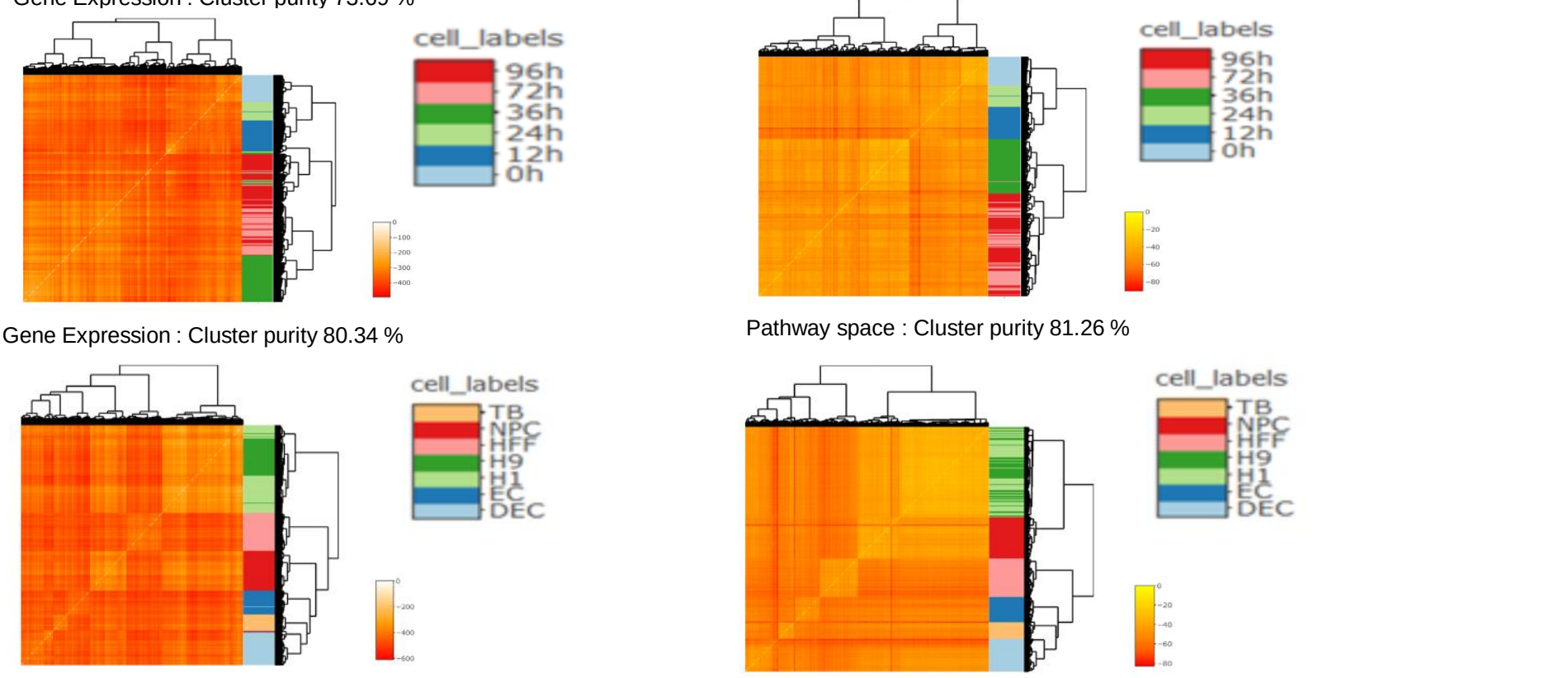
c.



Gene Expression : Cluster purity 73.69 %

Pathway space : Cluster purity 82.64 %

d.



Gene Expression : Cluster purity 80.34 %

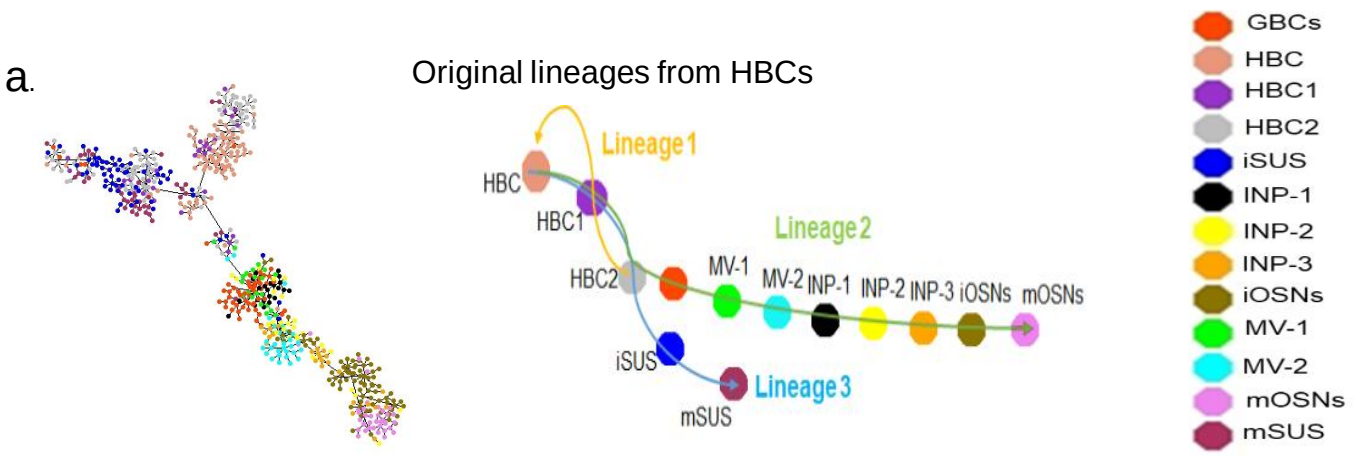
Pathway space : Cluster purity 81.26 %

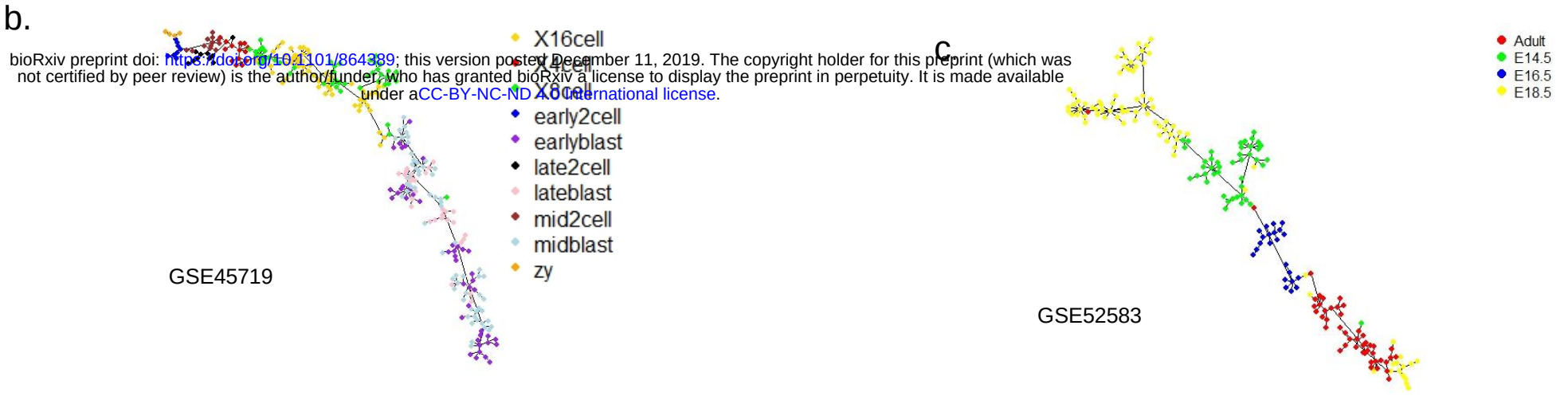
Gene Expression : Cluster purity 82.90 %

Pathway space : Cluster purity 83.79 %

Figure S6. Batch correction by UniPath and clustering purity using pathway scores **(a)** Combined human ESCs from a study (GSE75748) and Myoblast from another study (GSE52529). Few ESCs were subjected to systematic drop-out of 10% and 30%. Principal component analysis (PCA) based dimension reduction and visualization using gene expression shows two separate clusters of human ESC. However UniPath is robust to systematic variability in drop-out rate of 20% and 30%. In PCA based dimension-reduction of pathway scores by UniPath, all the hESC cells come together in one group. **(b)** PCA based visualization of batch effect in mouse embryonic stem cell lines processed using various protocols (GSE75790) combined with Hematopoietic stem cells from study GSE71794 in gene expression space. Batch effect removal in gene expression space using Limma. Batch effect removal done in pathway space using Limma shows mixing of ESCs to some extent and separation of HSCs. **(c)** Heatmaps showing comparable clustering purity in gene expression and pathway space for various cells types from study GSE67835. **(d)** Heatmaps showing comparable clustering purity in gene expression and pathway space for various cells types from study GSE75748.

Figure S7





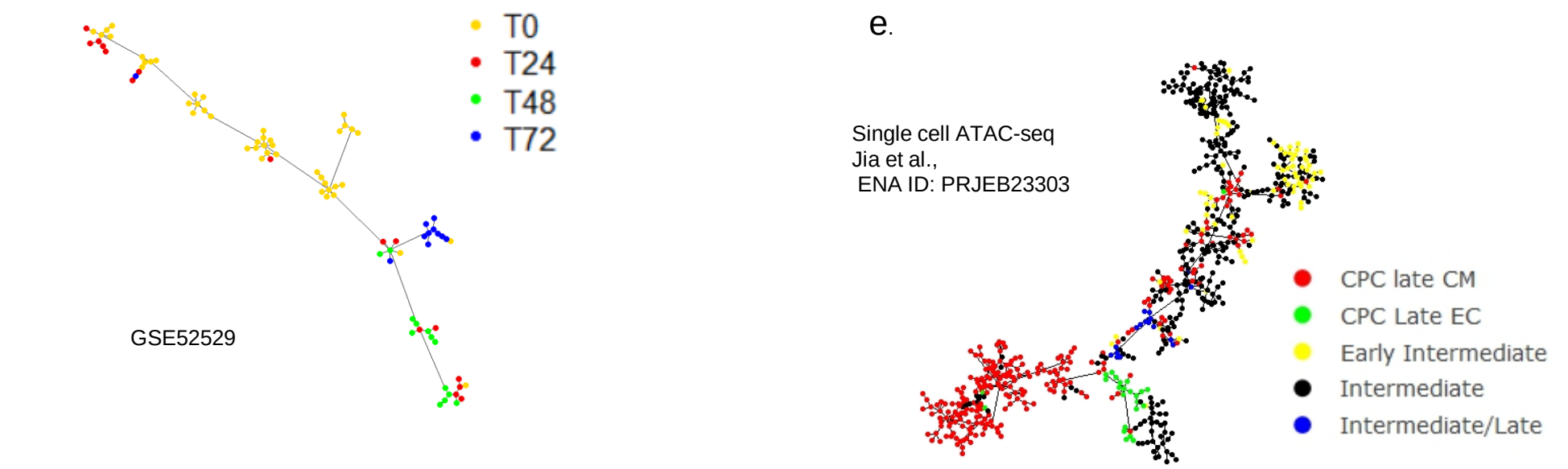


Figure S7: UniPath's results for pseudo-temporal ordering of cell using their pathway score. **(a)** Pseudo-temporal ordering for tracing differentiation trajectory of Horizontal basal cell (HBC) to neuronal and sustentacular cell lineages (GEO ID: GSE95601) **(b)** Pseudo-temporal ordering using pathway scores derived from scRNA-seq profiles of early developmental cells starting from zygote to late blastocyst (data from Deng et al. data , GEO ID: GSE45719). **(c)** Pseudo-temporal ordering showing mouse lung developmental stages (GEO ID: GSE52583). **(d)** Pseudo-temporal ordering showing myoblast cell differentiation at different time points (Trapnell et al, 2014, GEO ID: GSE52529). **(e)** Pseudo temporal ordering in pathway score derived from scATAC-seq profile of cardiac progenitor data (Jia et al., ENA ID: PRJEB23303).

Figure S8

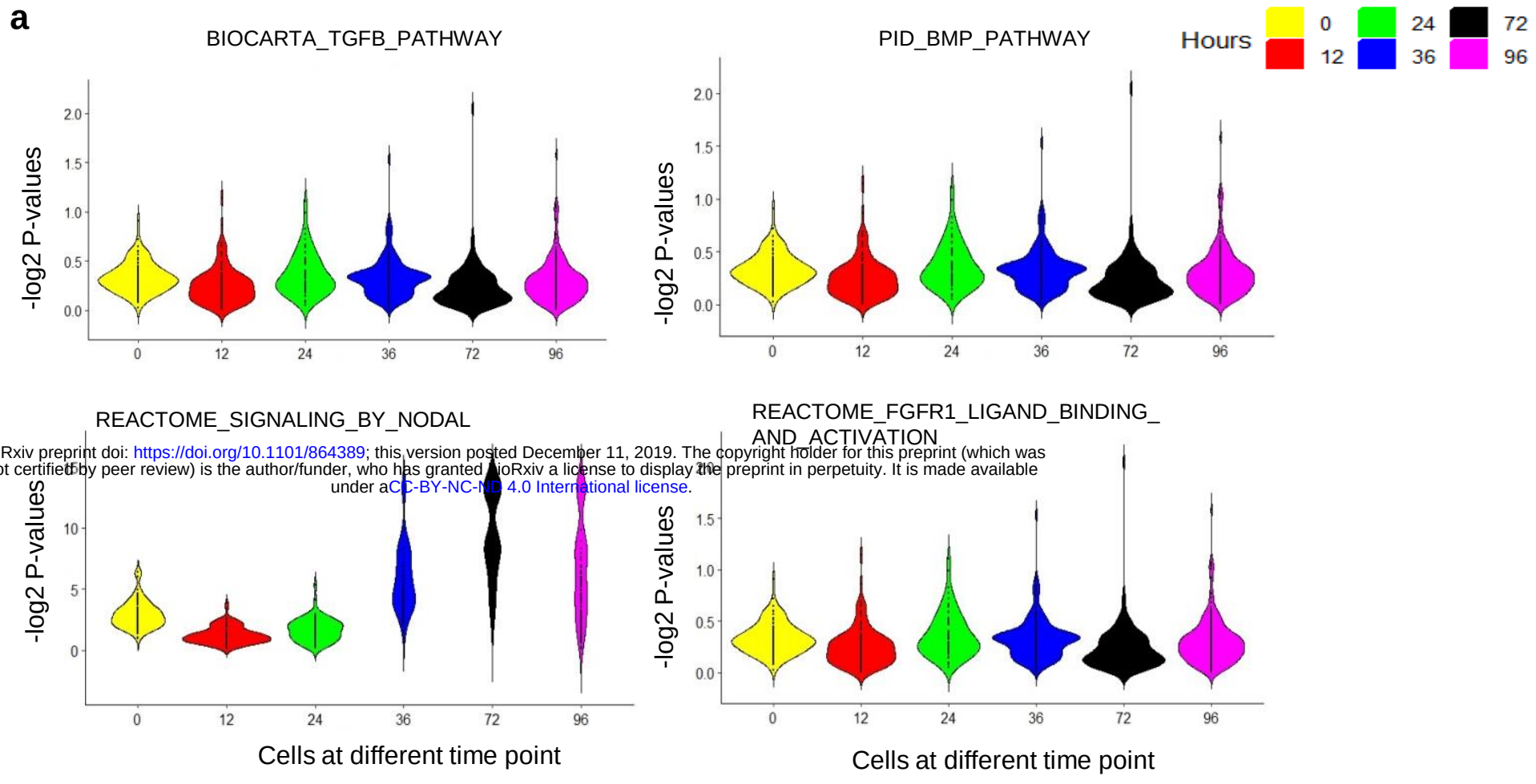
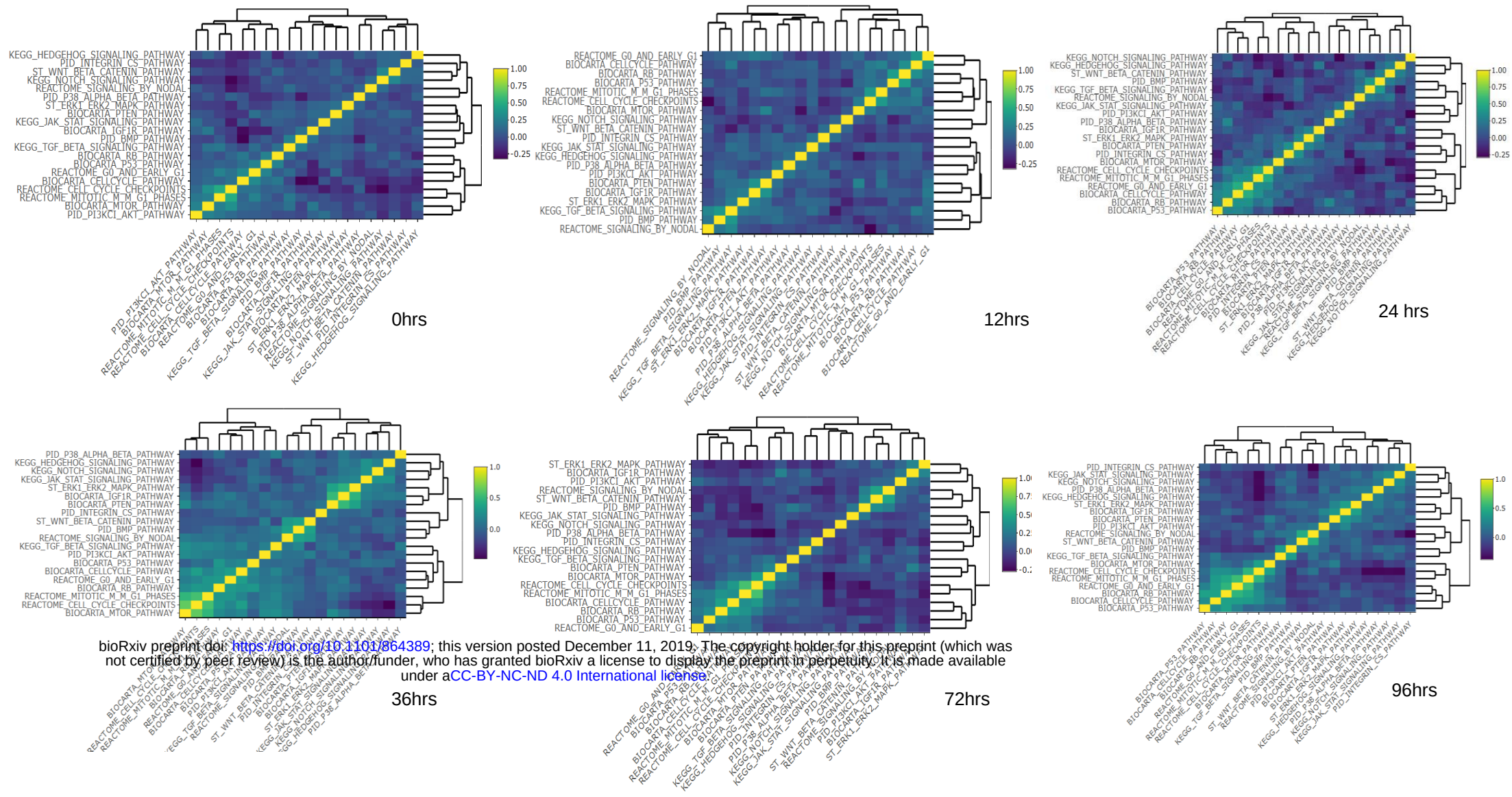


Figure S8. Enrichment and co-occurence of pathways during differentiation towards endoderm (a) Violin plot of enrichment score of pathways at different time point of differentiation. As expected, NODAL signaling pathways had higher score at 72 and 96 hours compared to 0 and 12 hours. Whereas BMP, TGF-beta and FGFR1 signaling pathway scores had bimodal distribution at 36 and 72 hours indicating the heterogeneity in regulation at same time point. (b) Spearman correlation between score of Nodal signalling and other pathways at different time points. Nodal signaling is known to act via smad2/smad3 for differentiation towards mesendodermal lineage (Fei et al. 2010). Here Nodal and smad2 signalling have highest correlation at 24 hours, then it decreased at the start of 36 hours which is consistent with existing literature (Fei, et al, 2010). Similarly Nodal and BMP has highest correlation at 36 hours after which it decreases. BMP is known to support differentiation towards mesendodermal lineage. Decrease in cooccurrence between BMP and nodal signaling at 72 and 90 hours is according the finding reported by Kyle et al. (2013) that high level of BMP inhibits differentiation of primitive streak cells towards definitive endoderm. Correlation between Wnt and Nodal signalling pathway scores increased as differentiation proceeded towards definitive endoderm till 96 hours and such trend support previous reports (Chabra, et al, 2018).

Figure S9

a



b

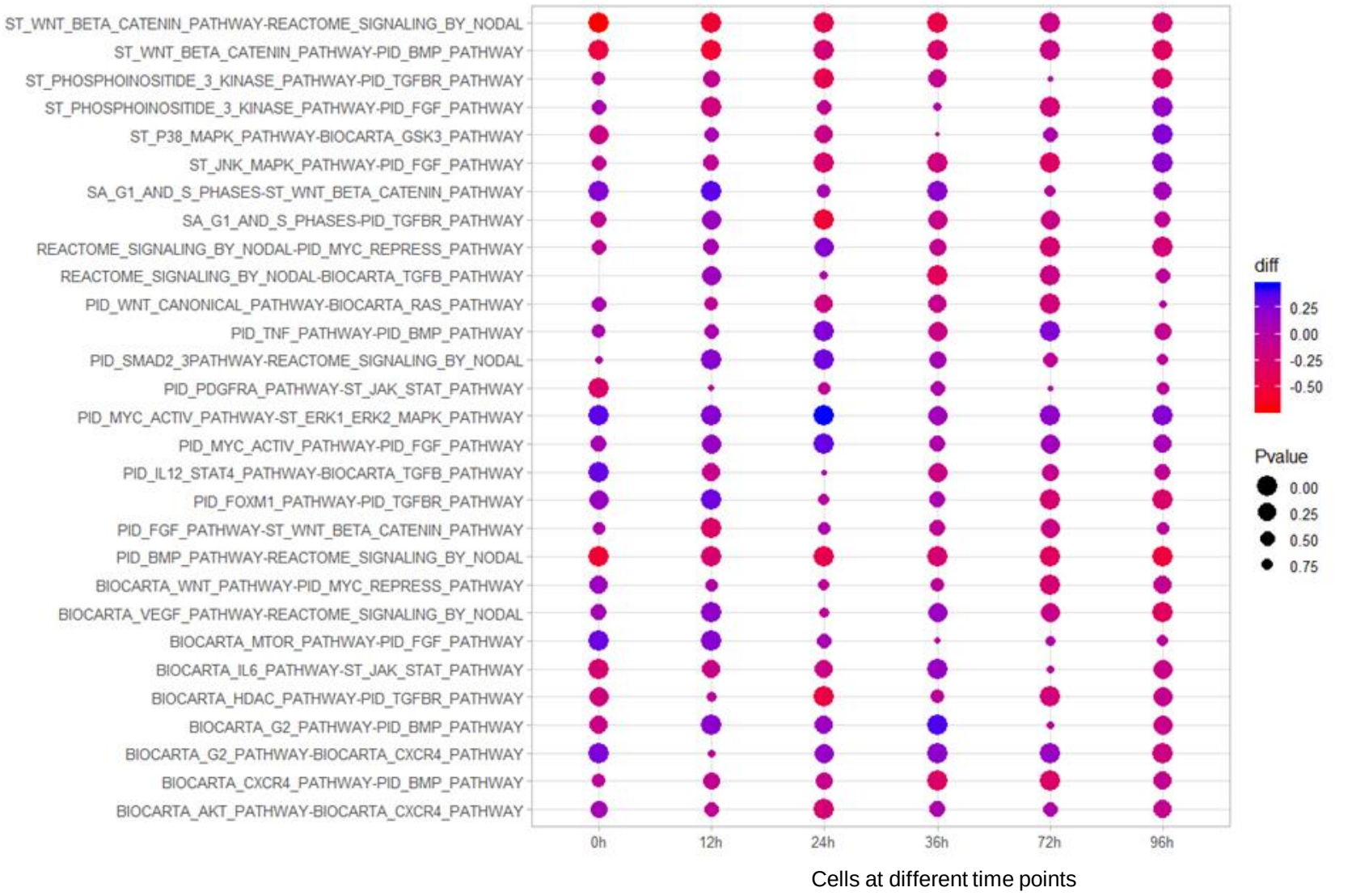
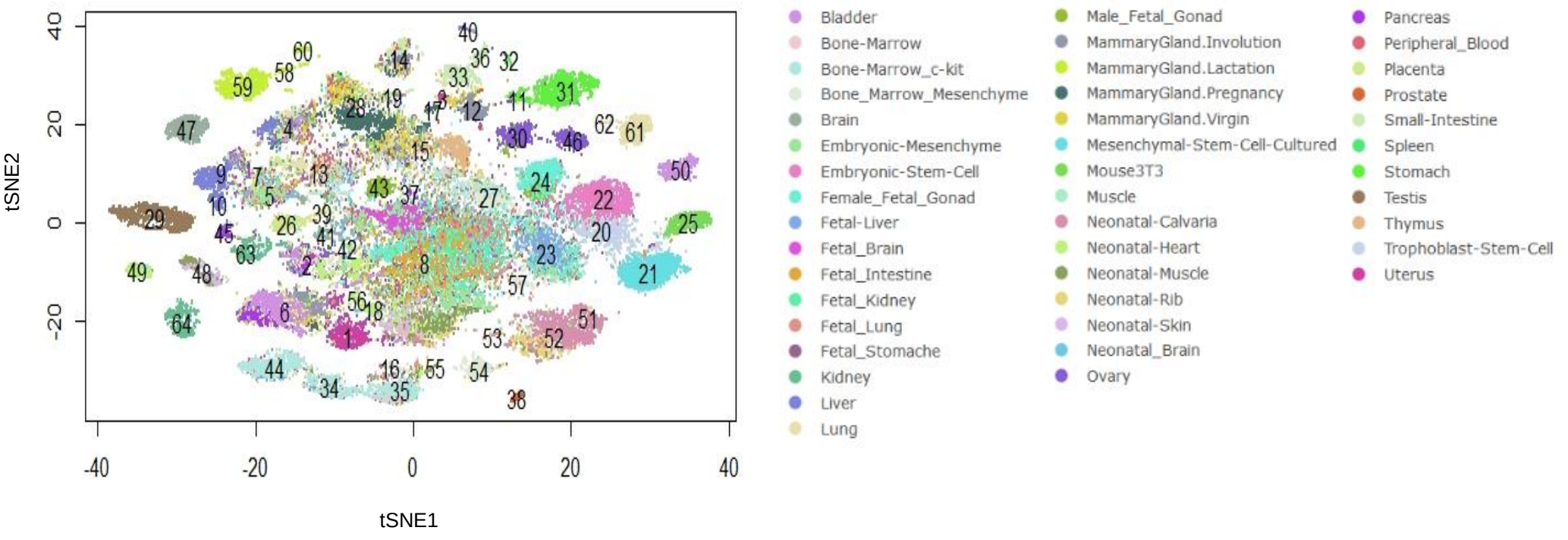


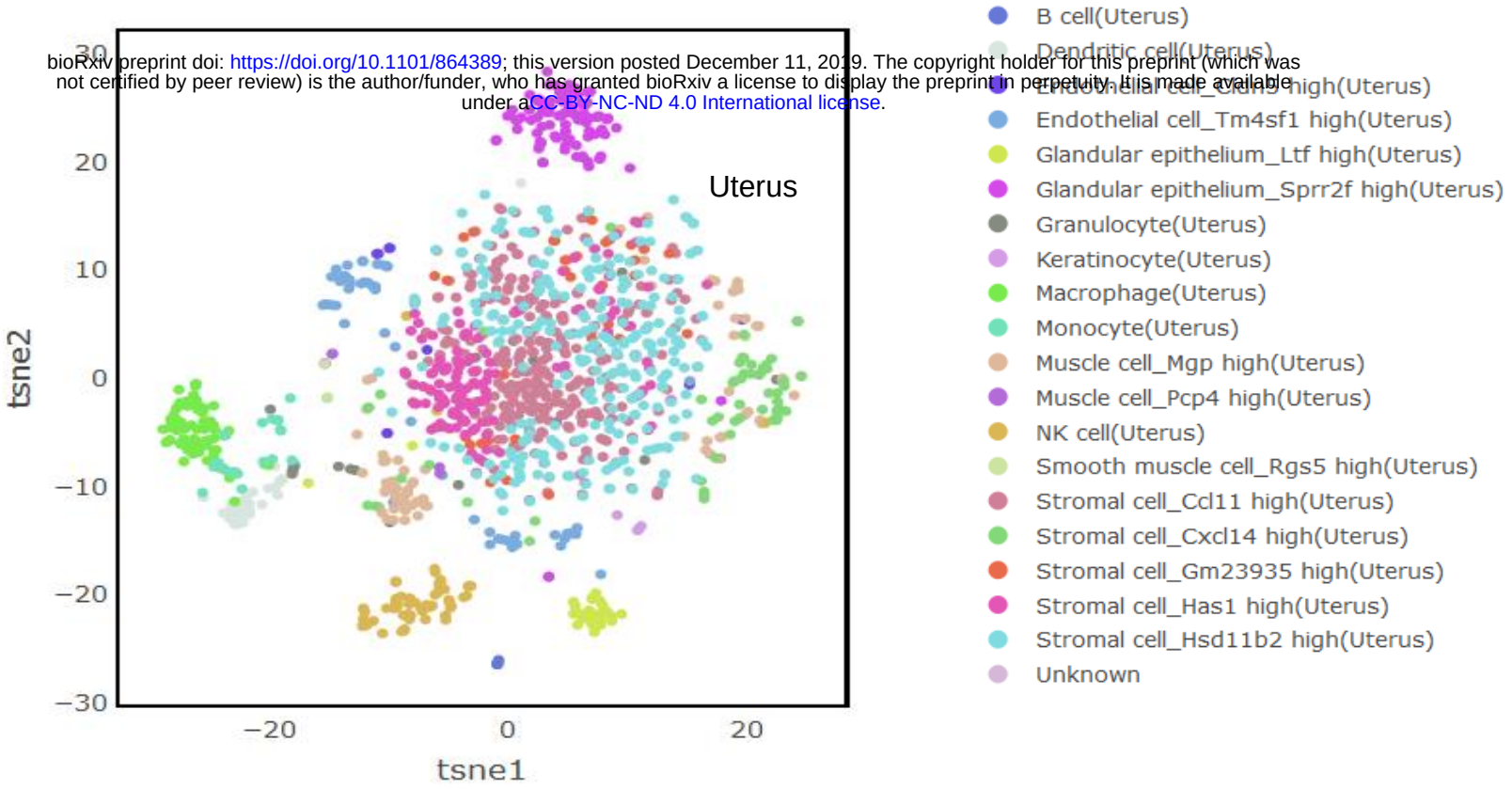
Figure S9: Analysis of change in pathway modules and co-occurences in differentiating human embryonic stem cells data (Chu et al, GEO ID: GSE75748) at different time points (0 hours,12 hours, 24 hours, 36 hours and 72 hours. **(a)** Heatmap of correlation between pathways at 6 time of differentiation for differentiating hESC. It shows how modules of pathways change at different time point of differentiation. It can be noticed that Nodal signaling pathways did not co-occur with TGF-beta and BMP pathways at 0 hour. However at 12 hr, 24 hr and 36 hr Nodal, TGF-beta and BMP pathways co-occurred together. **(b)** Differential correlation (co-occurrence/co-enrichment) of pathways, top pathways are selected from each time point and represented by dot-plot. Color of dots represents difference among pathways and size of dot represent P-values. Some of the co-occurrence patterns have been reported previously. Such as FGF induces activation of mTOR and mTOR is involved in involved ESCs pluripotency and suppression of mesoderm and endodermal related activities (Zhou et al., 2009). It can be seen that differential co-occurence of FGF and mTOR pathway is significant at 0 and 12 hours.

Figure S10

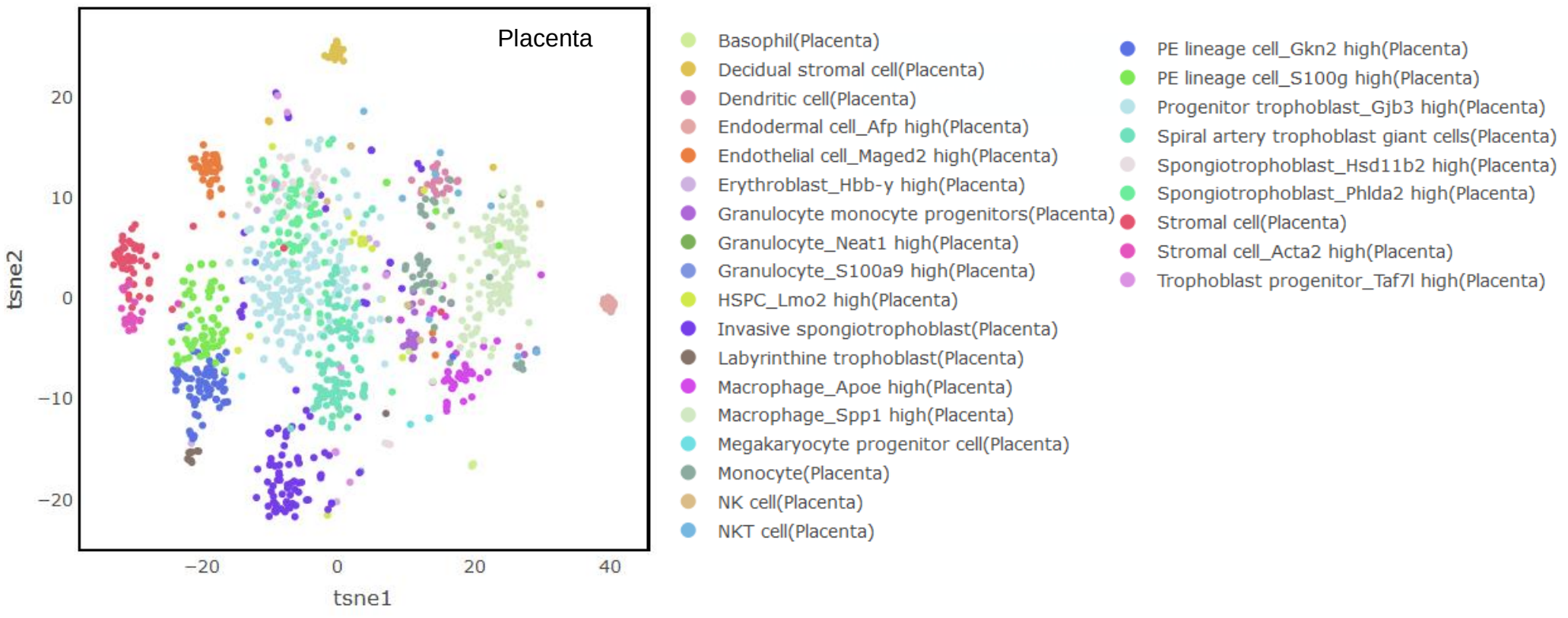
a



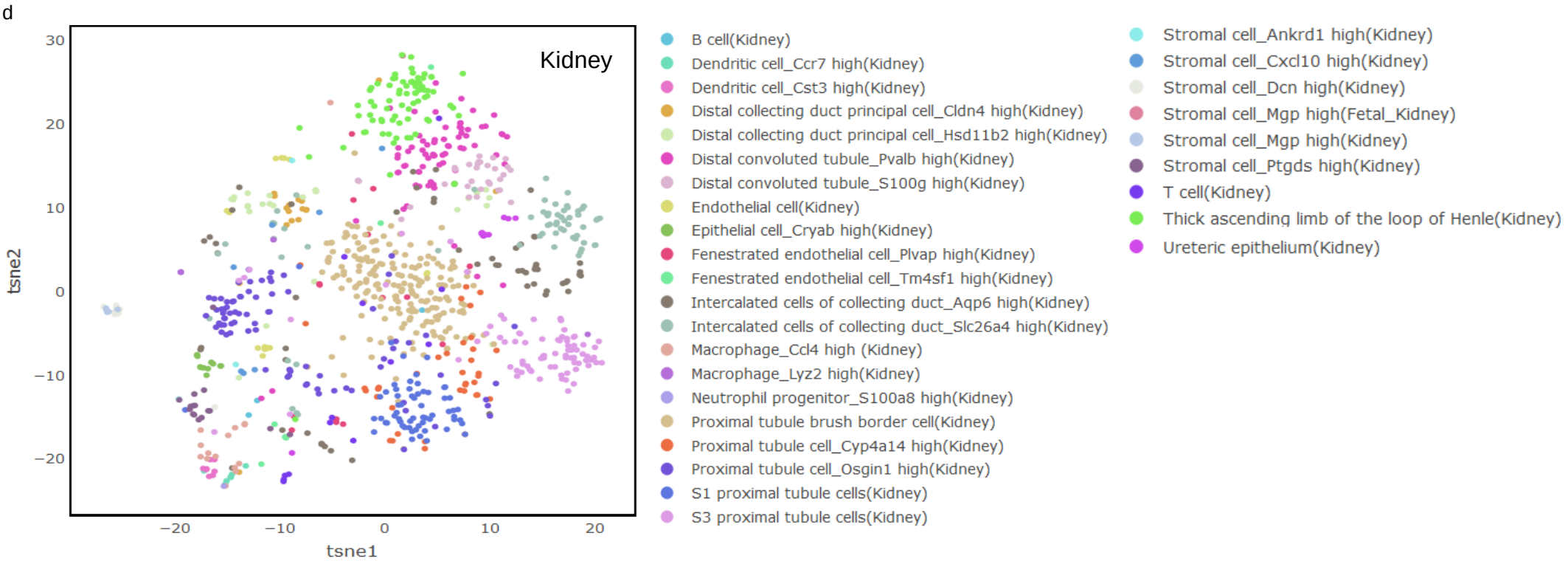
b



c



d



e

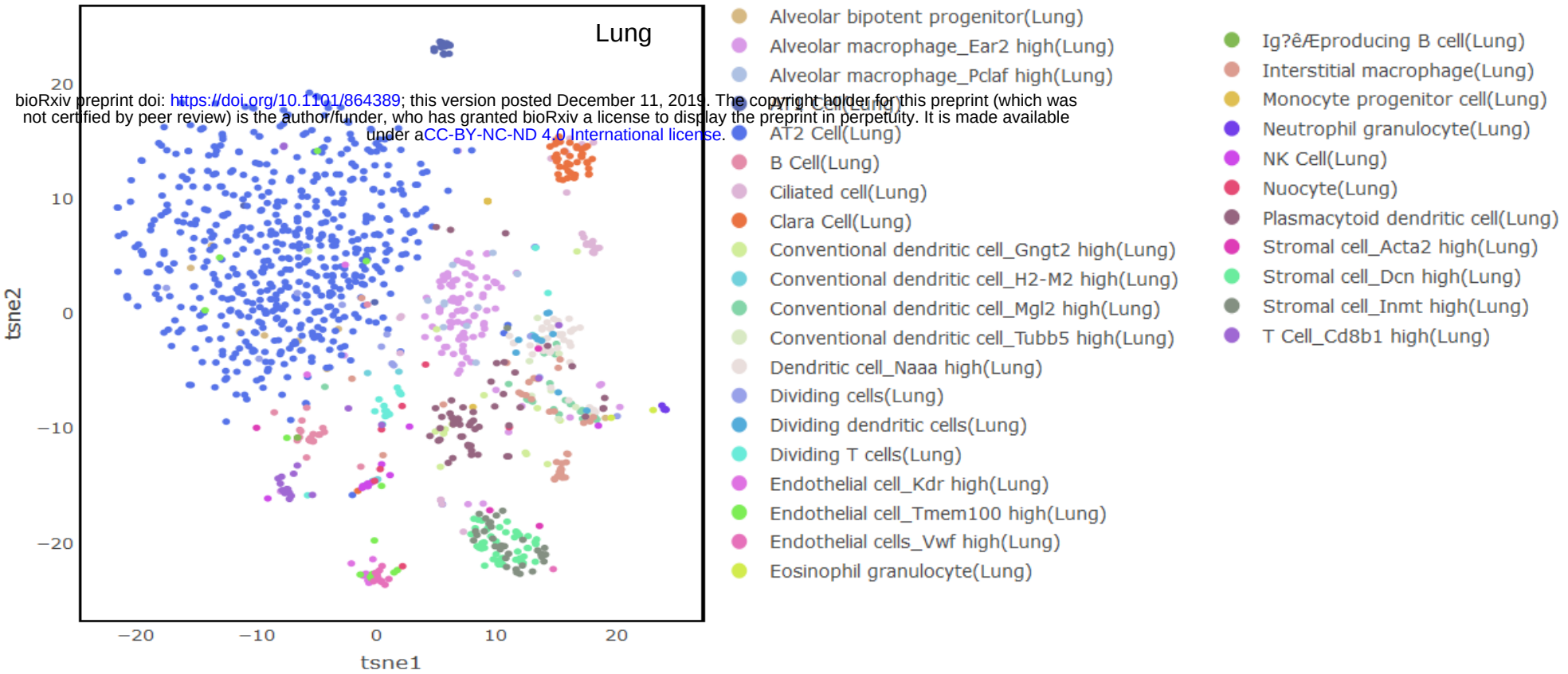
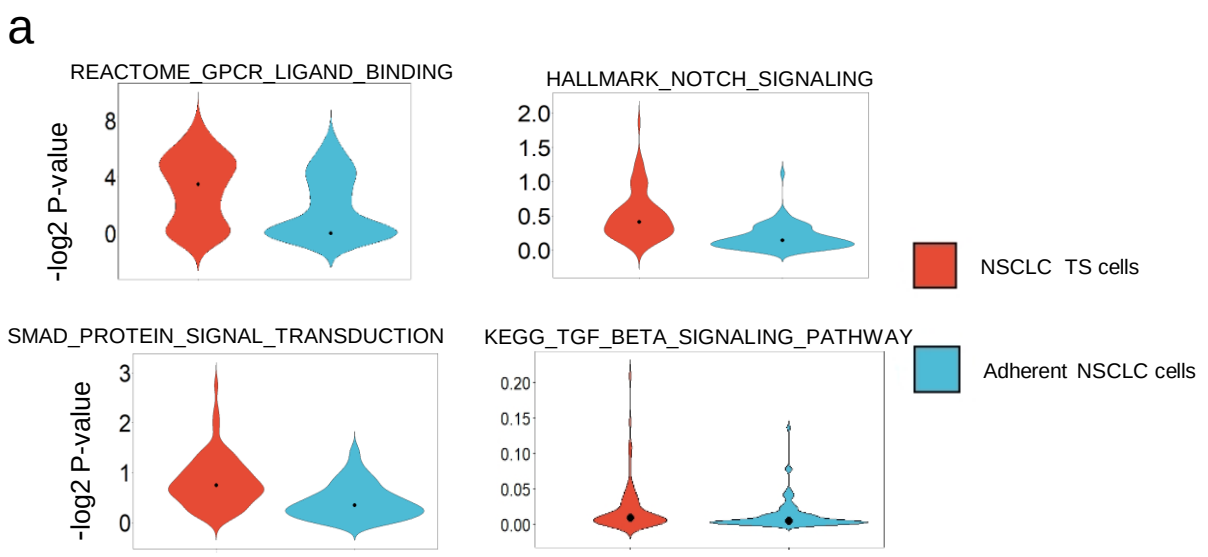


Figure S10: Pathway scores can be used successfully to distinguish each cell-type as distinct group. Scatter plot of t-SNE based dimension reduction using pathway scores of single-cells from different organs from mouse cell atlas (Han et al., GEO ID: GSE108097). **(a)** Scatter plot of t-SNE based dimension reduction of pathway score profile of mouse cell atlas (MCA). The major cluster numbers are also shown **(b)** t-SNE for cells from Uterus. **(c)** t-SNE for placenta cells. **(d)** t-SNE of kidney cell. **(e)** t-SNE of lung cell.

Figure S11



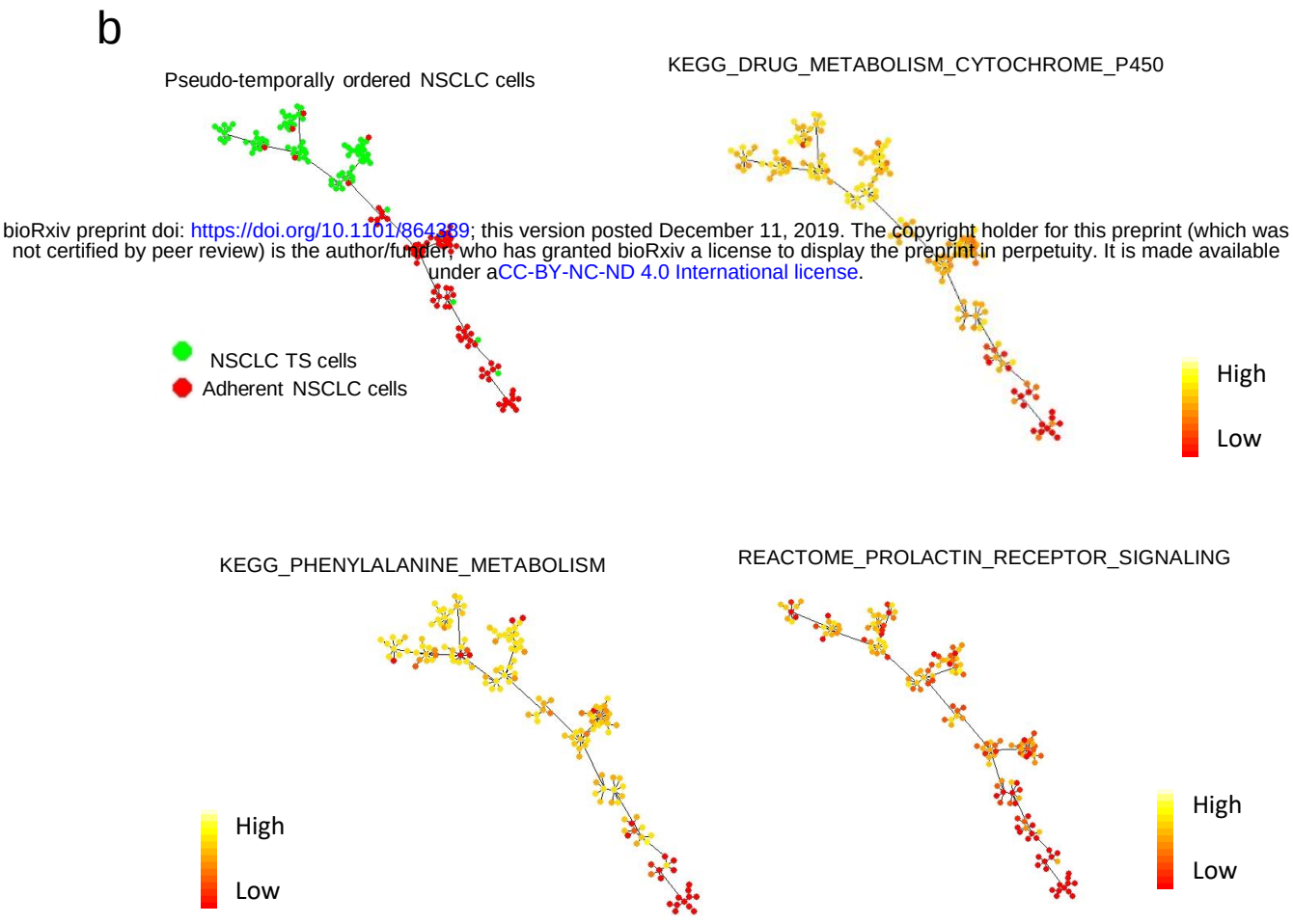


Figure S11: Visualization of distribution and gradient of enrichment score of pathways in non-small cell lung cancer (NSCLC) cells (a) Violin plot of enrichment scores for 4 pathways in Adh and TS cells of NSCLC. NOTCH, GPCR and SMAD signalling showed a significant difference (P-value < 0.01) in enrichment scores in TS and Adh cells. Whereas TGF-beta signalling pathway scores remained similar in both cell lines. (b) Pseudo-temporal ordering of lung cancer cells. The gradient of pathway scores for three different pathways/gene-sets are also shown.