Clustering of Cells Using Complete Gene Expression Profiles

Lior Ziv, Nitzan Bodenheimer, Amir Bar, Dana Perez

Abstract

The developments in genomic sequencing tools and especially single cell RNA-seq has brought new opportunities to study developmental biology and cell differentiation. In this study we revisit a paper by Satija et al who used landmark gene expression profiles to determine single cell tissue, and try to separate cells using the complete gene expression profiles. We believe that complete gene expression profiles may hold valuable information and will generalize the approach used by Satija et al. Lacking the data to verify our results our method success remains unknown.

Introduction

Ontogeny is the study of an organism's origin and development. Its emphasis is on the development of various cell types within an organism [1]. It is a useful field in many disciplines such as developmental biology and developmental psychology.   
A main challenge in the study of ontogenesis, is in analysis of the characteristics of different cell types in a tissue. Up until recently, gene expression analysis was only possible large-scale and contained the average expression levels of many different cells and cell types which compose a tissue. However the development and improvement of new technologies, such as single-cell RNA-seq have enabled us, for the first time, to determine an extensive gene expression profile of specific cell types within complex tissues.  
This new data type opens up many research opportunities. It also requires the development of new tools and protocols for pre-processing and analysis. One such main issue is using the data collected in order to locate the spatial location within the organism from which the specific cell has originated.  
Satija et al. (2015) [2] have developed a computational strategy tool named Seurat to do just that. They took information gathered from RNA-seq over 1152 single cells, analyzed and normalized it to extract a comparable gene expression levels which they used in order to cluster the cells. Then, using in-situ hybridizations they determined the expression or lack of expression of 47 'landmark' genes in certain spatial locations; and used the expression pattern in each bin to match the cluster of cells drawn from it.  
For our project, we used the single cell RNA-seq data to try and cluster the different cells by their gene expression levels, hoping to locate groups with the same cell type origin as accurately as possible. While we did not have access to labeled data and were therefore unable to validate our findings or speculate the tissue from which they came, we developed the Tux pipeline which we believe enables the clustering of same type cells out of single-cell RNA-seq data.

Results

**Generation of comparable gene expression table**

We started by analysing 1152 single-cell RNA-seq SRA files. Using TopHat we created a BAM file for each cell. Before we continued processing the data we considered several normalization methods for gene expression including normalization by “housekeeping genes”, differential expression (such as TMM and DESeq), library size methods such as total counts and fragments per kilobase of exon per million reads mapped (FPKM). We finally chose to normalize by FPKM since it takes into consideration both the gene's length and library size which provided us with comparable expression levels among the different genes in each cell and between different experiments. The computation of the FPKM was done using Cufflinks over all BAM files and the resulted values were arranged in a 1152x13,902 table representing the relative expression levels of all genes in each of the single cells. This table has allowed us to compare the genes expression levels of all cells and characterize expression profiles.

-Switched to article table + filtering of cells with too low level of expression/seq

**Applying PCA to find the highest variability components**

We now had an order of magnitude more genes than cells, and so to be able to detect expression profiles we needed to reduce the profile variables to those which best explain our data. To achieve this we applied PCA over our data. To choose the number of components we plotted the sum of percentage of variance explained by all principal components for 100 - 900 (in 100 components intervals) first components (Figure 1).

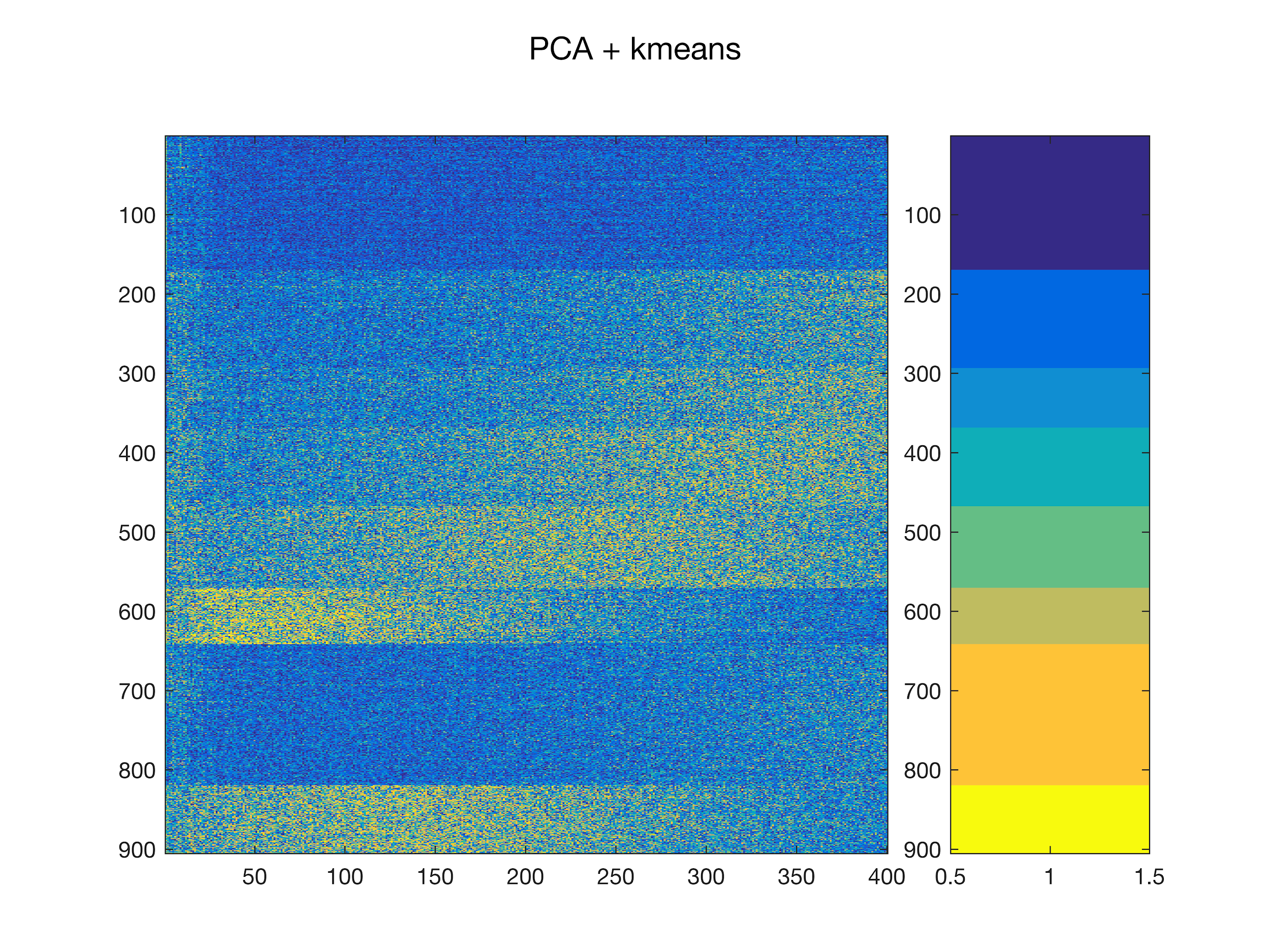
We decided to take the top 400 components, which explained ~91% of the variance in our data.

****

**Figure 1: Components explained variance.** The graph shows the additive effect component number on the explained variance in the data. Above 400 components the slope of the function decreases towards plateau.

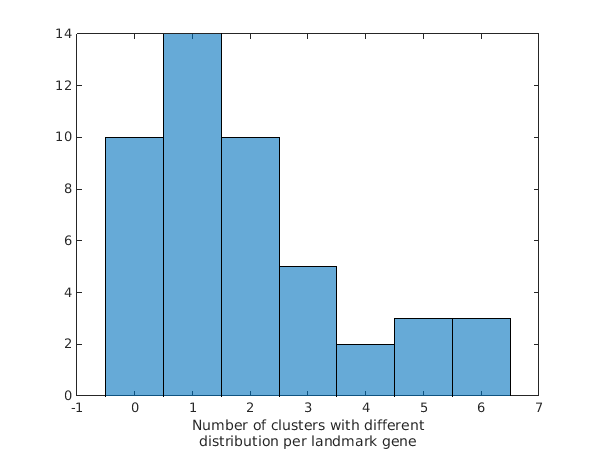
**Cells are clustered into different groups**

We continued to cluster the data, using the components we chose in the previous step. By applying the necessary normalization (see methods) and k-means clustering, we have successfully separated the cells into 8 distinct groups (Figure 2). To measure whether our



**Figure 2: k-means clustering.** The rows are components representing gene expression, the columns are zebrafish cells. The PCA vectors are normalized to values between 1 to -1, and the cells are sorted to match the clusters (rightmost bar).

results are meaningful, we applied a statistical test for each landmark gene out of the 47 used in Satija et al.. For each landmark gene we asked whether the expression levels within a cluster are of the same distribution of the background (expression in the rest of the clusters). By applying Wilcoxon rank sum test with a p-value of 0.05 we got that ~51% of the landmark genes are mostly unique to one or two clusters (Figure 3). This shows that our results are somewhat related to the literature and are not completely random.



**Figure 3: Distribution of landmark gene in clusters.** The x-axis is the number of clusters where a landmark gene was found to originate from a distribution different from the background. The y-axis is the amount of landmark genes who are significantly over- or under- expressed in x clusters.

Discussion

In this study, we explored the possibility of obtaining an unbiased clustering of different cells types out of heterogeneous tissues by using complete gene expression profiles. We developed a computational method termed Tux which process single cell RNA-seq data, estimates gene expression levels profiles and clusters cells. By applying Tux to the data of Satija et al. we managed to separate the data to visible 8 different clusters. While we cannot validate our results as we do not hold the labeling for the different cells, we believe our results at least support that such an approach is worth investigation.

**Tux pros and cons**

Tux advantage relies on the use of complete gene expression profiles for cell clustering. We believe that a complete profile holds a valuable information that might be lost by selection of specific genes as representatives for clustering. As a consequence of using the complete profile, we are able to apply this method on organisms of which we have limited knowledge. While those are impressive advantages, the Achilles heel of the method is our lack of data to assess the method’s accuracy. Another issue is the amount of samples required for the success of the analysis. Here we were able to reduce the dimension of the genes using PCA from 13,902 to 300 while still having ~96% of the variance in the data explained. This shows that we need a significant amount of samples just to apply the method and avoid dimensionality issues. Another pitfall regarding the application of PCA is the possibility that most of the data would be explained by few components. In such a radical case, the problem would be equivalent to segmentation of dots on a single line which sometime cannot be separated.

**Validation and applications**

An access to a labeled dataset would have allowed us to test the method’s sensitivity and specificity and to compare those to Satija et al. method. Moreover, such data can be utilized to apply machine learning approaches. With large enough dataset of labeled data, if the data is indeed separable as the clustering suggest, we should be able to generate a multi-label classifier for recognition of cell type.

**Summary**

Tux is an ambitious method but the lack of data to validate it is a huge hurdle it has to cross before being applied on any data. However it shows that such a may be possible and if applicable it would be a powerful tool.

Methods

**Data gathering, mapping and expression estimation:**

We downloaded the raw data from NCBI given by the GEO ID GSM1628102. The data is composed of 1152 single cell paired-end RNA-seq experiments, each one is given in a separated sra file. We converted the sra files to fastq format, extracting the reads, using sratoolkit version 2.8.2-1 with the command **fastq-dump <sra-file>**. We did not use the split flag of the program since the data was missing the second pair mate of each read.

Next we mapped the reads to the Zebrafish genome, GRCz10 (GCA\_000002035.3), with the corresponding GFF file using tophat[3] version 2.0.13 by the command **tophat -o <out-directory> --GTF <gff-file> -T <fastq-file>**. We used the program’s default parameters since we did not want to restrict the mapping further as we refer to the paired-end as single-end reads. This process resulted with 25 bam files, one for each chromosome in the Zebrafish genome.

Quantifying the gene expression in a manner that would balance the different experiments was a major challenge. First, the submitted data didn’t contain the random molecular tag (RMT), a sequence attached to the second pair mate of each read which allow us to measure the effect of the PCR amplification. Second, with our approach we want to be as agnostic as possible regarding to prior knowledge of landmark genes. With those restrictions we decided to compute the FPKM for each gene. With that, we achieve 1152 (one per cell) vectors of relative gene expression. The FPKM was computed using cufflinks[4] version 2.2.1 with the command **cufflinks -o <out-directory> <chromosome-bam-file> -G <chromosome-gff-file>**. The complete scripts used to process the data can be found in the supplementary files.

**Gene expression table filtering:**

Our following analysis were performed over the Satija et al. gene expression supplementary table. Following their suggestion, we excluded cells that showed low expression levels, that is cells with less than a threshold of 2,000 genes expressed. This process have narrowed the dataset to 905 expression vectors.

**Principal Component analysis:**

We used principal component analysis (PCA) to reduce the dimension of each expression vector from 13,902 to 400 components using the Matlab function **pca**. Those reduced vectors will be used later to cluster the cells to their specific type.

**PCA normalization and Clustering:**

As we cannot directly cluster the PCA transformation as it represents the variance across the first 400 components we applied two step normalization. The first was the handle the variance along the component axis and was done by applying Z-score for each component. At this stage, as we still couldn’t see any clustering since some of the values were off-scale and we could not see the similarities. To overcome this problem, we applied a logistic function and limiting the values to the range [-1, 1]. Over the normalized data, we applied clustering using the Matlab function **k-means** with 1,000 iterations.

References

[1] Thiery, Jean Paul (1 December 2003). "Epithelial–mesenchymal transitions in development and pathologies". *Current Opinion in Cell Biology*. 15 (6): 740–746. PMID 14644200. doi:10.1016/j.ceb.2003.10.006.

[2] Satija, Rahul, et al. "Spatial reconstruction of single-cell gene expression data." *Nature biotechnology* 33.5 (2015): 495-502.‏

[3] Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq." *Bioinformatics* 25.9 (2009): 1105-1111.‏

[4] Trapnell, Cole, et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." *Nature biotechnology* 28.5 (2010): 511-515.‏