Technische Universität Dresden
Biotechnologisches Zentrum (BIOTEC)
Molecular Bioengineering Master Program

PC-corr application for developmental gene expression pattern analysis

Name of the student: Ming-Ju Kuo
First supervisor: Dr. Carlo Vittorio Cannistraci
Second supervisor: Dr. Sara Ciucci
Name of lab head: Dr. Carlo Vittorio Cannistraci
Affiliation: BIOTEC, TU Dresden

Grade:
Remarks:

Date:
Signature of supervisor:

1. Introduction

With stem cell biology and regenerative medicine being gradually emphasized, the cell development in organisms is getting more attention and, with the help of new technologies, a considerable amount of data is now available. However, the developmental process is still unelucidated.

It is known that many genes are involved in organism development, therefore gene expression plays a major role in cell differentiation. A previous related study showed that patterns from gene expression data were suggested by unsupervised machine learning. Hence, tools for computational gene expression data may help to understand the process of development [1]. In this study, gene expression data was analyzed by two unsupervised machine learning methods: principal component analysis (PCA) and PC-corr. Patterns of three germ layers and novel gene regulatory networks in cell development are elucidated and discussed.

Gene expression

The central dogma of biology describes how genes information in DNA becomes its product, which is protein or RNA. In biological systems, the macromolecules that are needed for life are derived from gene expression, which involves important steps such as transcription, RNA splicing, and translation. The cells have different gene expression patterns at different time and space which regulate several functions of themselves, including cellular differentiation and morphogenesis.

Transcription

One gene is a series of code from a DNA fragment. Genome DNA is composed of two strands that complement each other and each strand contains 3’ end and 5’ end. For a gene, one strand is the template and the other the coding strand. The template strand can be used to produce an RNA molecule whose process is called transcription.

RNA splicing

In humans, RNA splicing is an important modification process after transcription. It refers to the formation of mature messenger RNA (mRNA) from pre-mRNA by removing introns and connecting exons together.

Translation

The process of translation decodes the mRNA through the triplet-base arrangement (codons) to produce the corresponding amino acids. The translation has three steps, initiation, elongation and termination. All these processes take place within the ribosome in the cytoplasm. The tRNAs are associated with the mRNAs and the mRNAs are brought to the ribosome to later create the polypeptide chains formed according to the base arrangement of the respective mRNA. Subsequently, the polypeptide chain will be folded into a stable protein structure that will be used in the organism to fulfill its biological purpose.

Organogenesis

Tissues or organs in an organism are derived from different cells of a fertilized egg, this developmental process is called embryogenesis. In the very beginning, two haploid cells merge together and form a diploid egg, which can be formed either by sexual or asexual reproduction. As next, the diploid single cell cleavages rapidly. In its developmental stage, organogenesis is of vital importance since it has to do with the differentiation of ectoderm, mesoderm and endoderm into different organs (Table 1) helped by the regulation of gene expression.

Table 1. Organs differentiated from three germ layers.

Endoderm	Mesoderm	Ectoderm
Thymus	Bone	Brain
Lung	Cartilage	Spinal cord
Pancreas	Skeletal muscle	Nervous system
Prostate	Tendon	
GI tract	Reproductive system	
Liver	Kidney	
Thyroid	Spleen	
	Limb	
	Heart	
	Smooth muscle	
	Blood	
	Adipose	

High-throughput Method Data

RNA sequencing

RNA sequencing (RNA-seq) is a transcriptomics research method based on next-generation sequencing and able to quantify RNA in a transcriptome sample. The first step in RNA-seq is to extract RNA from the biological sample and to transcript RNA to cDNA. To analyze the RNA-seq data, cDNA fragments are sequenced by proper technologies. Finally, cDNA sequencing data is compared with the genome data to identify the transcripts. RNA-seq can provide the gene expression data directly by quantifying mRNA in the cells. The method is widely used in transcriptomics research.

Microarray

Microarray is a silicon chip with a DNA array. Thousands of nucleic acid probes are on the few square centimeter area of the chip. The microarray is detected by fluorescence intensity when the sample RNA or DNA associates with the probe. Only with a single experiment, the microarray is able to generate plenty of gene information. It is a tool for genomics and genetics research, and it is often applied for gene expression test.

Dimension reduction: Principal component analysis (PCA)

Dimension reduction (DR) is a widely used strategy in the machine learning and statistics field. DR is defined as a method to obtain a set of principal variables from high-dimensional data by reducing its random variables in limited conditions. Principal component analysis (PCA) is one of the most used DR methods due to its simplicity. The dimensions of the new coordinate space are called principal components, which explain the variance in the data. Thus, the first principal component (PC1) is the linear combination of the data variables with the largest variance. The second principal component is the linear combination orthogonal to the first principal component with the second largest variance, and so on. This is able to reveal the inner structure of data provided in a lower dimensional space which can be visualized by humans. In this study, the PCA method was applied to the high throughput biological data which contains several tissue samples with their numerous gene expression values.

PC-corr network

PC-Corr is a method to construct discriminative correlation networks from multidimensional data based on PCA. The first step of the algorithm is to apply the PCA on a multidimensional dataset. PC-corr finds automatically the PCs that discriminate the data into the respective class groups, which in the example case is PC1 (Figure 1). The loading (weight) of each feature is obtained according to PC1 and scaled with a sigmoid function to be later combined with the Pearson correlation in the so called PC-corr formula. A cut-off is applied to filter out features with little contribution on PC1. Finally, the thresholded features and its PC-corr and loading values are used for the PCA discriminative correlation network construction (Figure 1) [2].

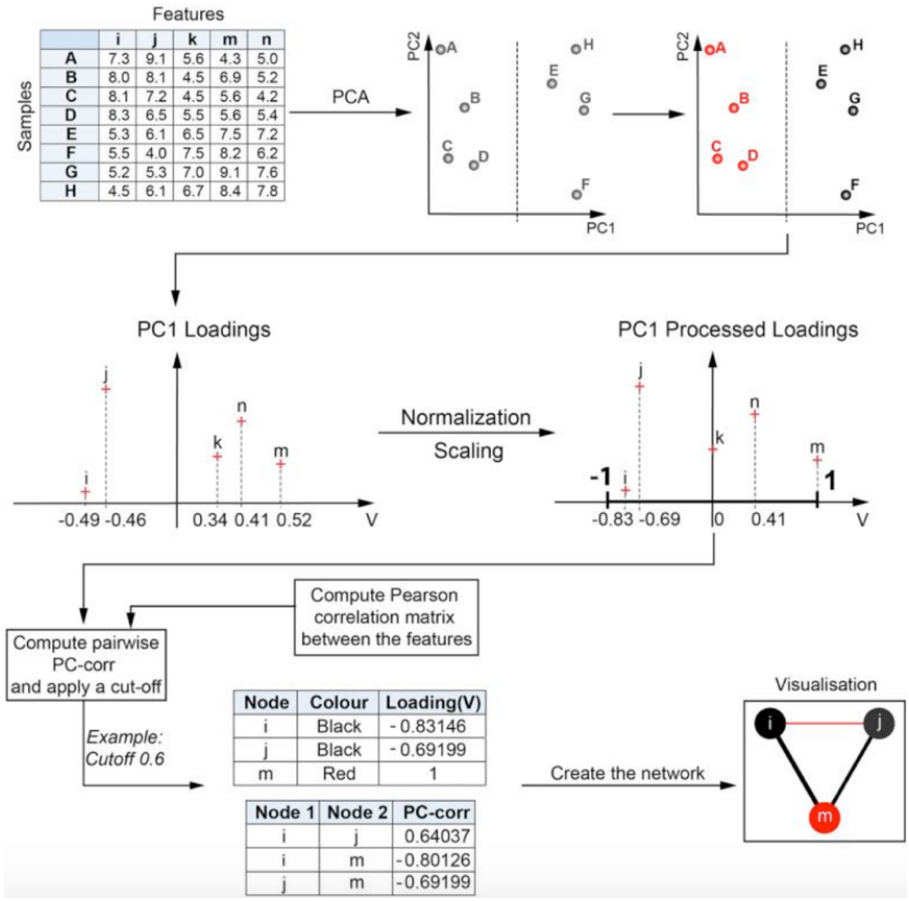

Figure 1. PC corr network. The dataset is analyzed by PCA which segregates the samples into two groups (red and black). The loading of each feature is calculated according to PC1. To construct the network, the Pearson correlation between features is calculated. Finally, the cut-off of loading and the correlation are used for the network construction. Red connection indicates a positive correlation while black connection indicates a negative correlation, while red nodes indicate higher expressed feature in the red group and black nodes in the black group [2].

2. Method

Dataset creation

Gene expression datasets were obtained from published articles of different species including pig, olive baboon and human [3-5]. Pig and human high-throughput data were obtained from microarray while olive baboon high-throughput data was obtained from RNA sequencing. The germ layer origin of tissues in each dataset was labeled and scored according to the physical contribution of the germ layers to the tissues. Score 0 is no direct contribution, score 0.5 is limited direct contribution and score 1 is direct contribution [1].

The interspecies dataset was merged from olive baboon and human dataset. In order to merge them, the common transcription factor genes of both species were chosen. Therefore, there were the same amount of features for each sample in the dataset.

Data analysis

The datasets were normalized and centered before the dimension reduction, analyzed in a two-dimensional space and divided into two classes, endoderm and mesoderm (positive class) against ectoderm (negative class). To choose the best combination of non-centered and centered PCA with different normalization, the nonparametric Mann-Whitney p-value was computed to indicate the class separability. The lowest p-value indicates the highest separability of samples. Subsequently, PC-corr algorithm was applied for the network construction after PCA analysis. During the network construction, different cut-offs were used for selecting features (genes) with significant PC-corr values.

3. Result

Pabio Anubis

The PCA analysis of the hidden pattern of olive baboon expression data is explored using different normalizations and its results are shown in Figure 2-4. The non-centered olive baboon dataset with quantile normalization is presented in Figure 2, and shows that endoderm and mesoderm are separated from ectoderm along with the second principal component (PC2). Interestingly, the tissues contributed for the immune system, bone marrow, spleen, lymph node and thymus, were separated from the other tissues along PC2. The non-centered olive baboon dataset with z-score normalization is shown in Figure 3. There, two separate groups are shown in the PCA space. The ectoderm was clearly separated from mesoderm and endoderm in PC2 and, moreover, the tissues of the immune system continued with the separative pattern. The other highly separative result is shown in Figure 4. The result shows the non-centered dataset with log normalization. If noticed, here the results presented similar patterns, two dimensions are enough to separate the patterns: immune tissues were separated from the others along PC2 as well as the ectoderm from mesoderm and endoderm.

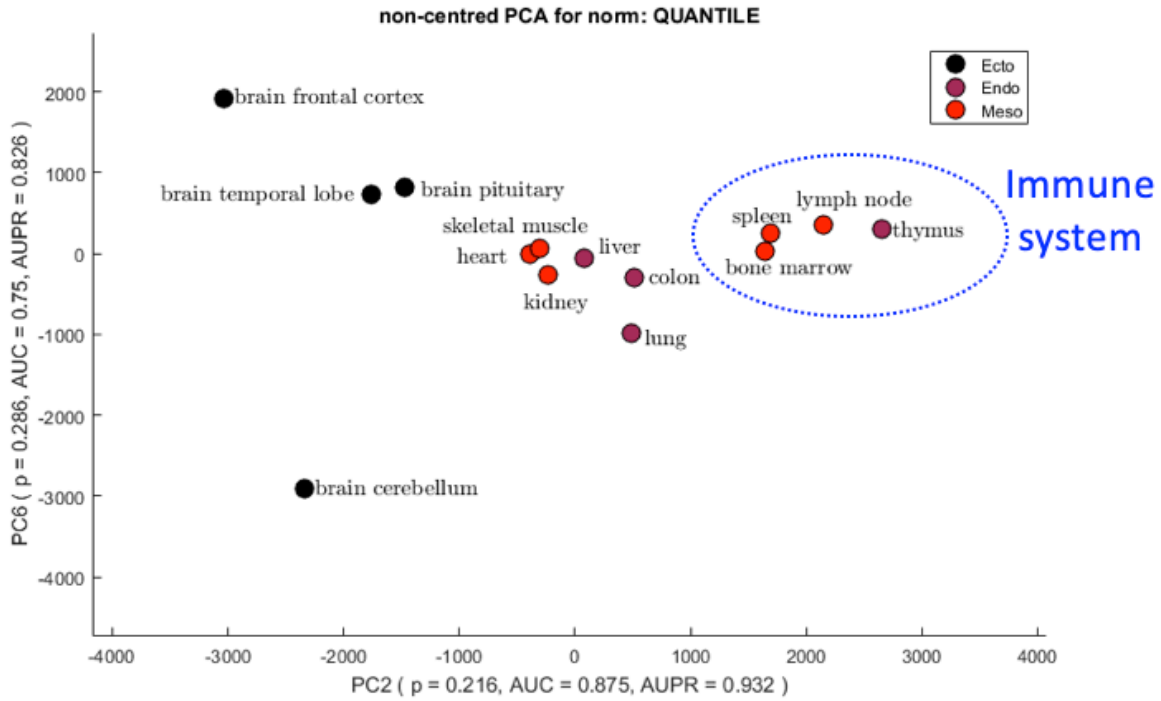

Figure 2. The result of PCA visualized in bi-dimension space (PC2 and PC6) from the olive baboon gene expression dataset with quantile normalization.

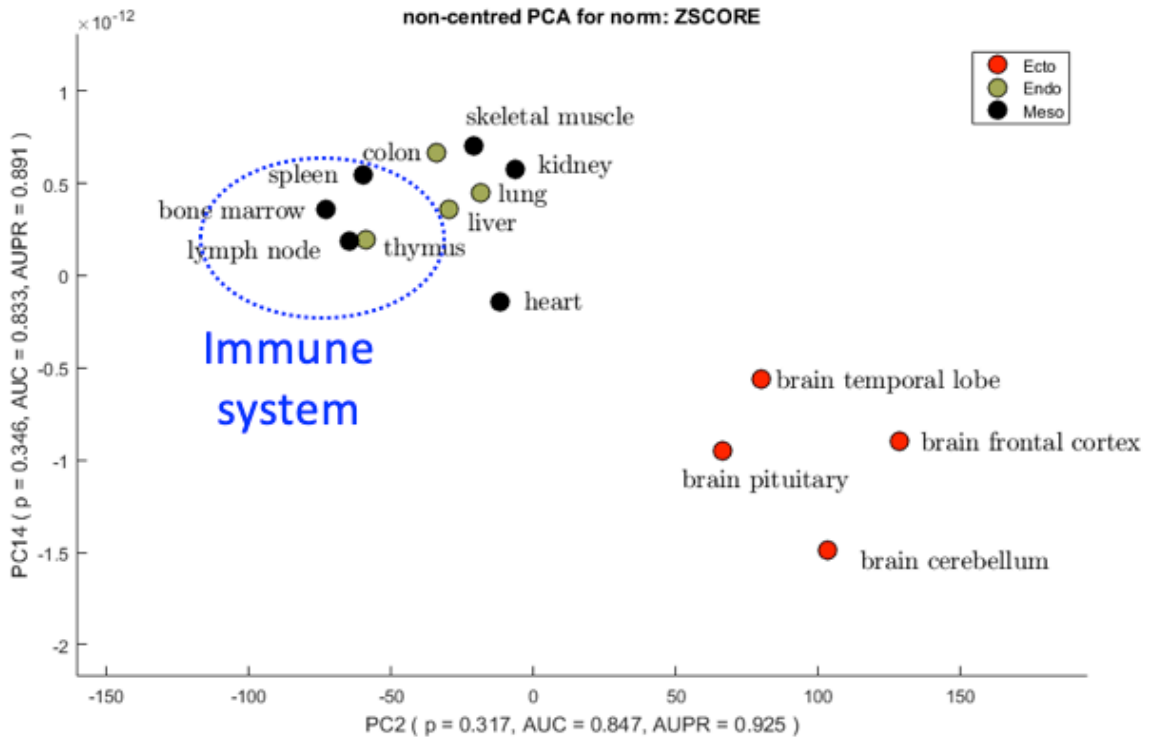

Figure 3. The result of PCA visualized in bi-dimension space (PC2 and PC14) from the olive baboon gene expression dataset with Z-score normalization.

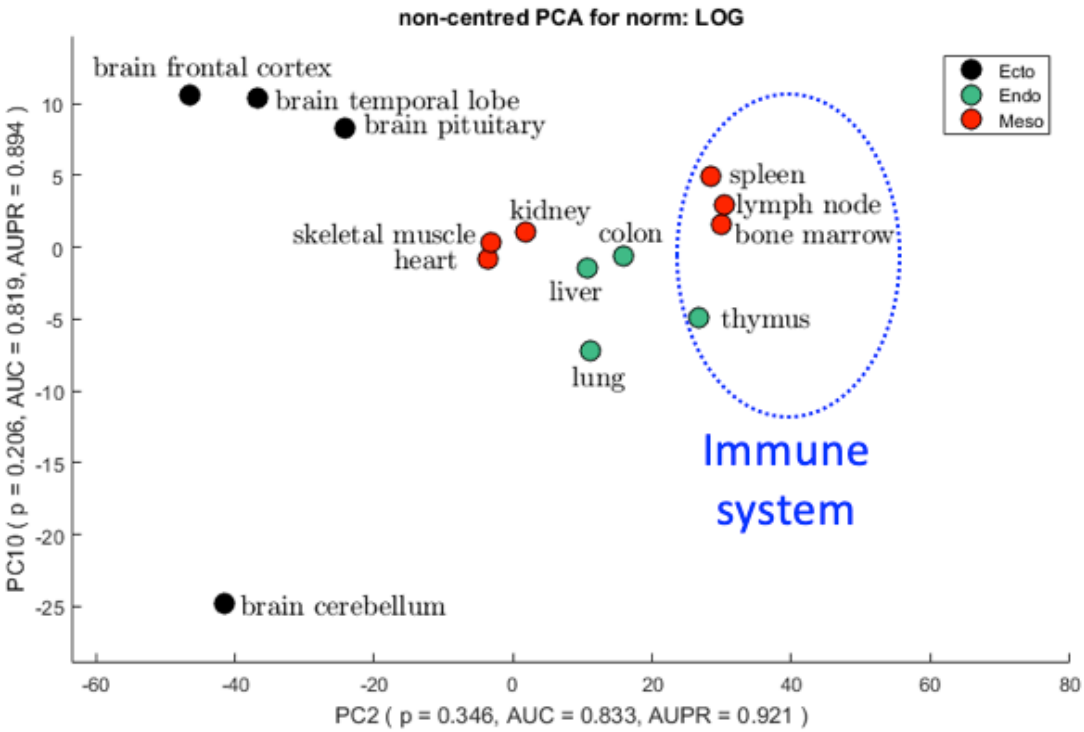

Figure 4. The result of PCA visualized in bi-dimension space (PC2 and PC10) from the olive baboon gene expression dataset with log normalization.

To construct the PC-corr network (Figure 6), we took the PCA result related to the non-centered dataset with log normalization. The samples in the olive baboon dataset were re-grouped into two classes, ectoderm for negative class, and endoderm and mesoderm for positive class (Figure 5). Since with this new grouping, the PC2 discriminate perfectly both groups, it was taken for the PC-corr analysis and the respective network construction. Red genes in the network highly expressed in the endoderm and mesoderm while black genes expressed higher in the ectoderm's group.

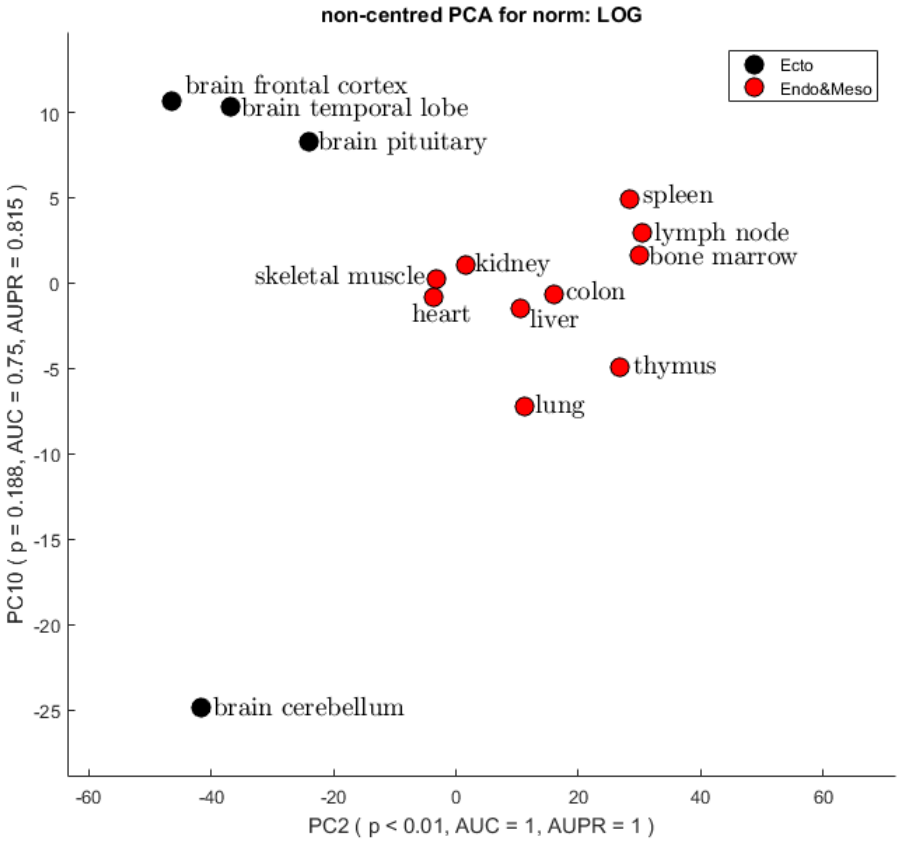

Figure 5. The result of PCA visualized in bi-dimension space (PC2 and PC10) from the olive baboon gene expression dataset with log normalization. Positive class (red): endoderm and mesoderm. Negative class (black): ectoderm.

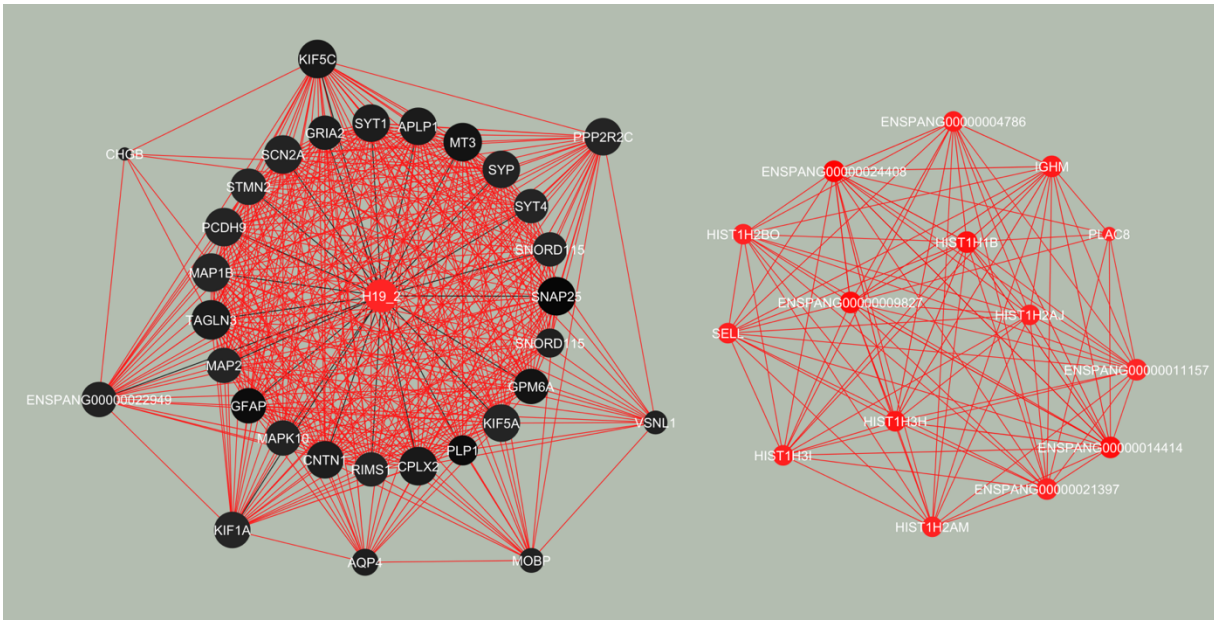

Figure 6. The 0.85 cut-off correlation network was constructed according to PC2 in Figure 5. The PC-corr network has 47 nodes and 516 edges. Red nodes indicate higher gene expression for the positive class while black nodes indicate higher gene expression from negative class.

Sus Scrofa

In this case, we present the results of the pig data. The PCA results are shown in Figure 7 and Figure 8. In Figure 7, the samples were labeled in three groups, endoderm, mesoderm, and ectoderm. Mesoderm and endoderm were separated from the ectoderm in PC2 while PC3 separated mesoderm from endoderm. The samples were divided into two class, positive class for mesoderm and endoderm, and negative class for ectoderm (Figure 8). Based on this PCA result, the discriminative correlation network was constructed from PC2. The network results with 0.85 cut-off appears in Figure 9. The black nodes were the genes contributed higher to ectoderm development. On the other hand, there is an absence of red nodes, which are genes that would contribute higher to mesoderm and endoderm development. In this case, all genes present a positive correlation (red edges).

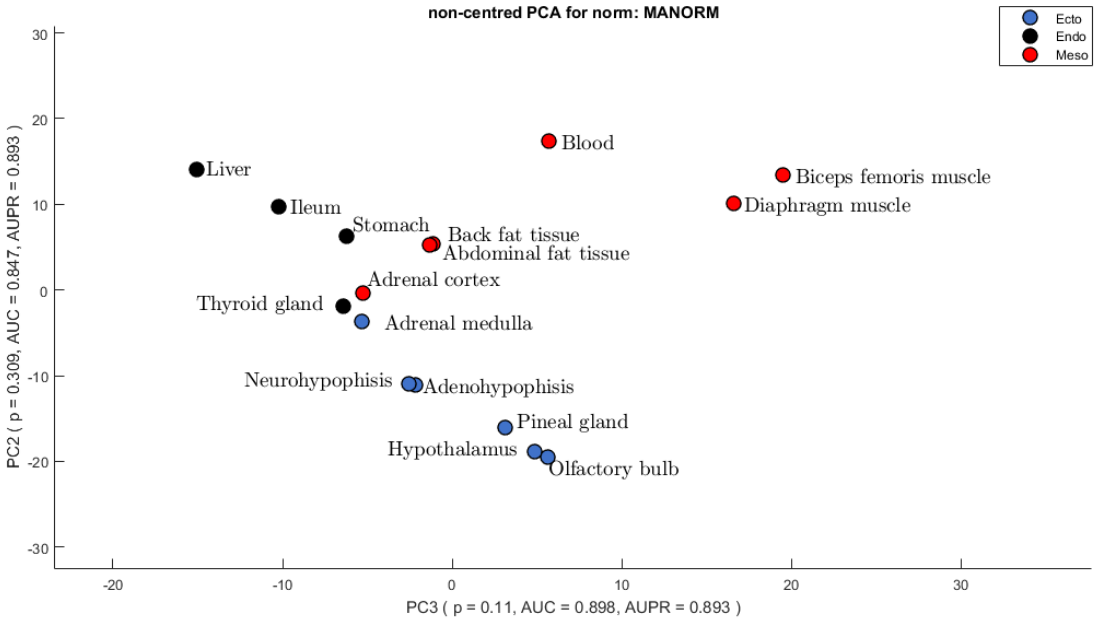

Figure 7. The result of PCA visualized in bi-dimension space (PC2 and PC3) from the pig gene expression dataset with Manorm normalization.

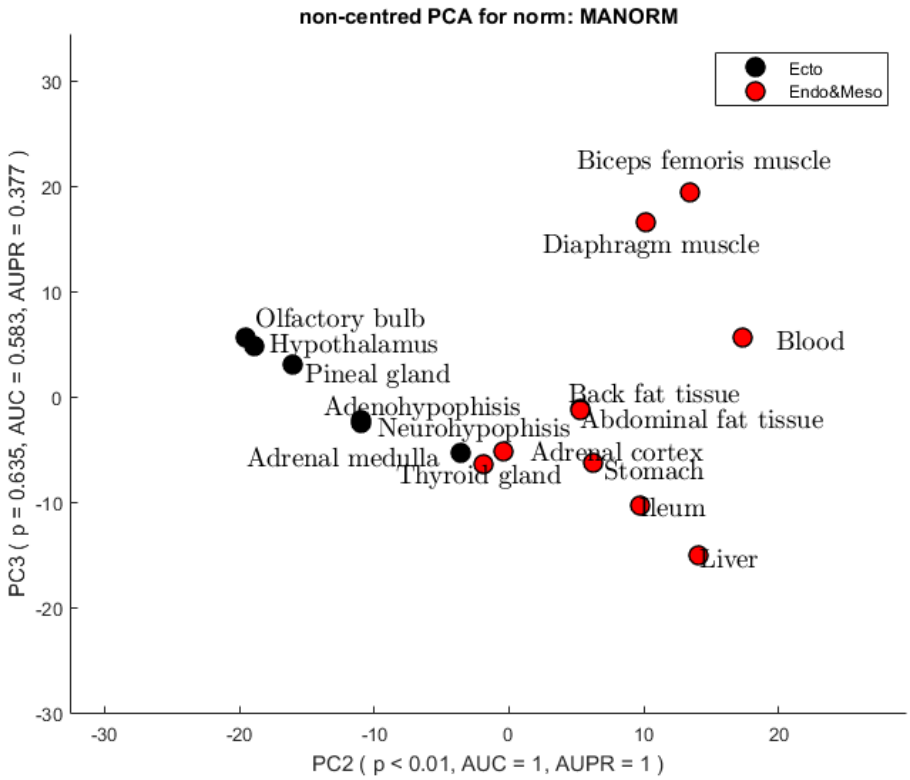

Figure 8. The result of PCA visualized in bi-dimension space (PC2 and PC3) from the pig gene expression dataset with log normalization. Positive class (red): endoderm and mesoderm. Negative class (black): ectoderm.

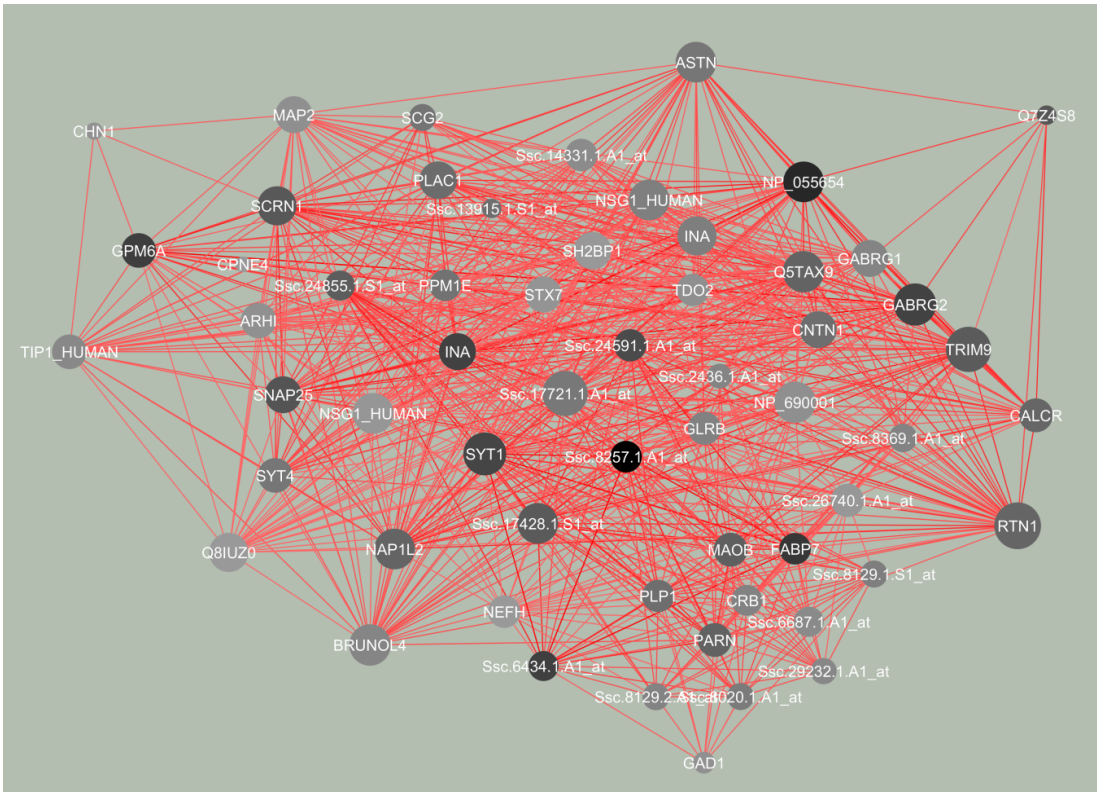

Figure 9. 0.85 cut-off correlation network was constructed according to PC2 of Figure 8. The PC-corr network has 58 nodes and 758 edges. Black nodes indicate higher gene expression from negative class, the intensity of the color represent the intensity of the expression for the negative class.

Interspecies: Papio Anubis and Homo Sapiens

In this section, the dataset is composed of transcription factors of the olive baboon and the human dataset. In the first part, the dataset was divided into four groups, ectoderm, mesoderm, endoderm and immune system. The result of PCA showed that PC1 separated three germ layers and PC2 separated the immune system form three germ layers (Figure 10). To construct the discriminative correlation network, the samples were divided into two classes (Figure 11), where the positive class was three germ layers and the negative class the immune system. The correlation network of the three germ layers against the immune system is shown in Figure 12.

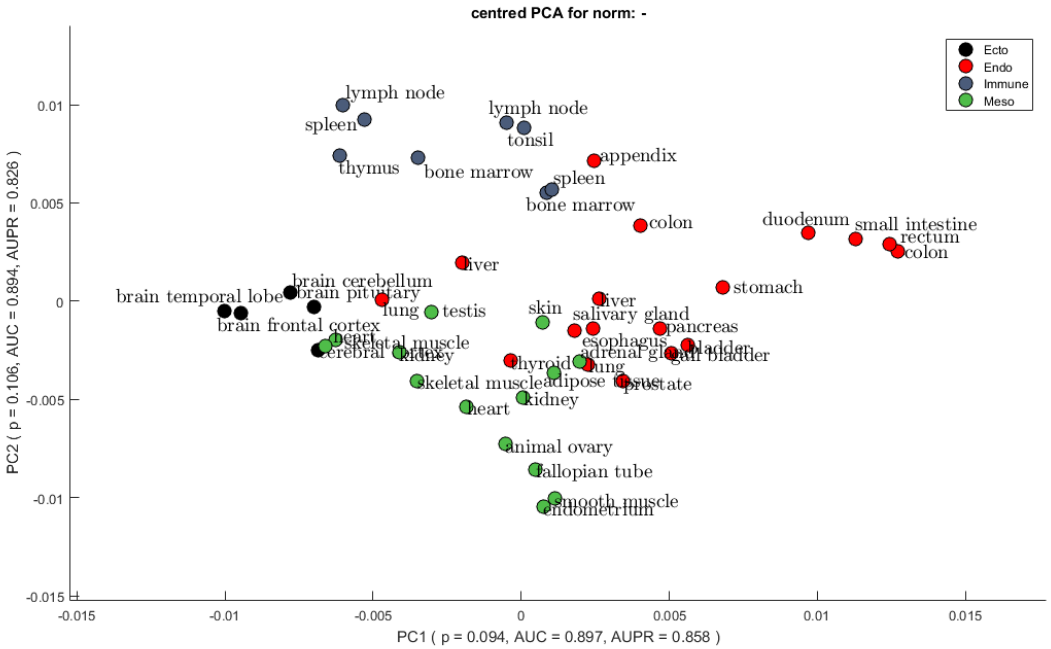

Figure 10. The result of PCA visualized in bi-dimension space (PC1 and PC2) from olive baboon and human composed transcription factor gene expression centered dataset.

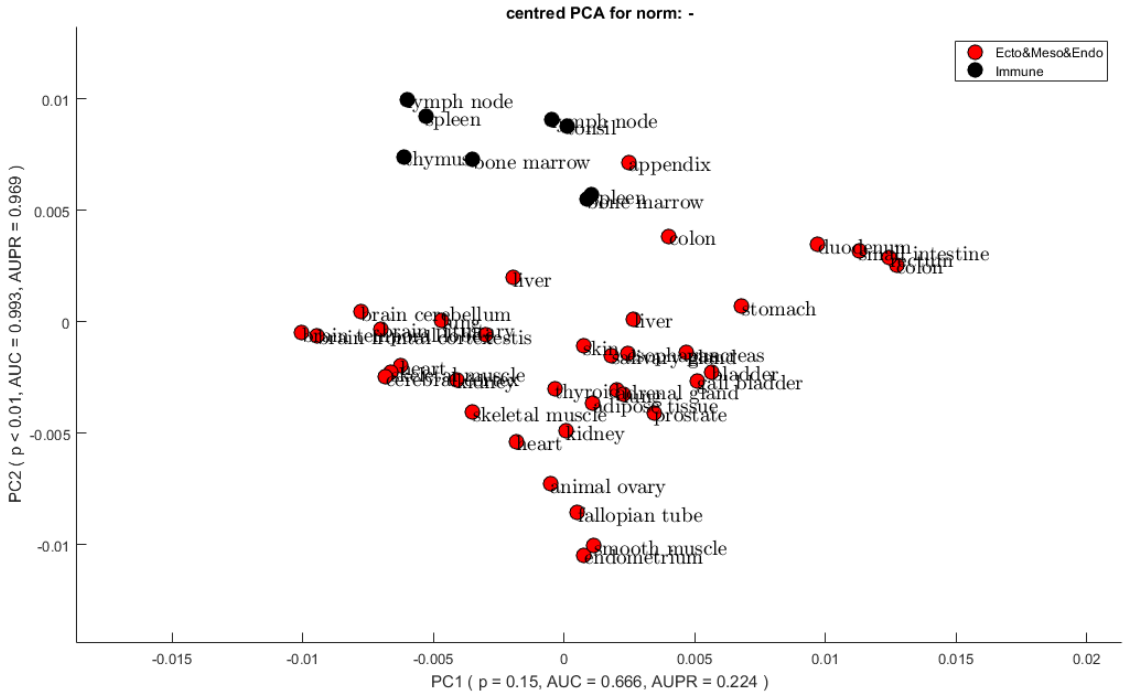

Figure 11. The result of PCA visualized in bi-dimension space (PC1 and PC2) from olive baboon and human transcription factor gene expression dataset. Positive class (red): three germ layers. Negative class (black): the immune system.

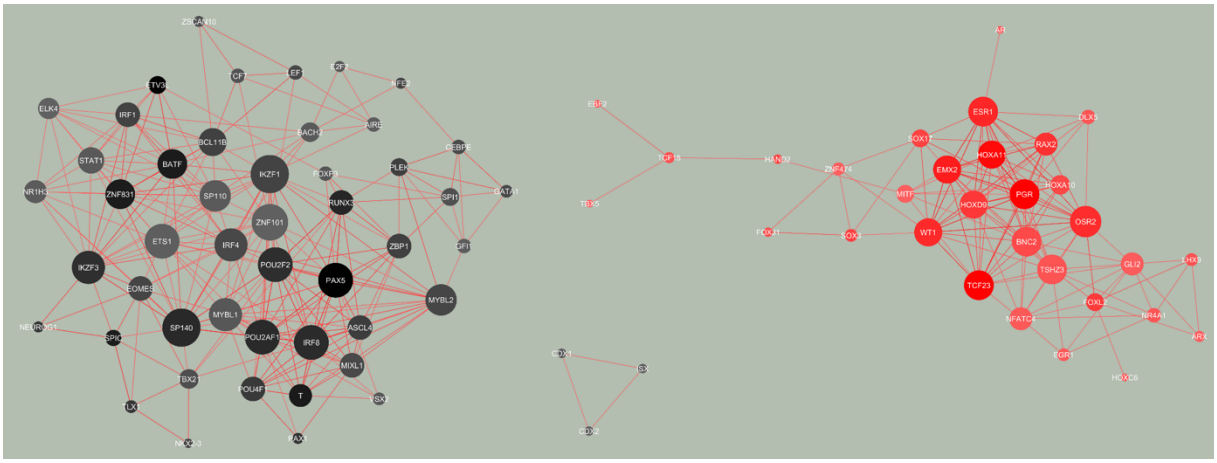

Figure 12. 0.85 cut-off correlation network was constructed according to PC2 of Figure 11. The PC-corr network contains 82 nodes and 361 edges. Red nodes indicate higher gene expression for the positive class while black nodes indicate higher gene expression from negative class.

4. Discussion

Gene expression plays an important factor in cell differentiation during organism development. Therefore, gene expression data can be of great help to unfold information from this process. In our result, PCA is able to segregate the immune system from other systems, suggesting that there are significantly different expression patterns between them during the organism development. The mesoderm and endoderm were clearly separated from the ectoderm. However, the boundary between mesoderm and endoderm was unclear which indicated that mesoderm and endoderm might have interaction during development stage, since mesoderm and endoderm were derived from the primitive streak region of the epiblast in the early stage [6].

In the 0.85 cut-off PC-corr result of the olive baboon, 47 genes were obtained. To understand the function and biological process of the gene set obtained from PC-corr analysis, a gene ontology analysis was carried out on the ectoderm class gene (black nodes), because we already know that the ectoderm is contributed to neuro system development only. The enriched biological terms indicated that the function of the genes in the network are relative to the neuro system and neuron development (Table 2). Notably, the neuro system and brain are derived from the ectoderm. Therefore, we can highlight that the results of PC-corr were consistent with the discovery in previous biological research [6].

Table 2. Gene ontology analysis from DAVID. Significant gene term was chosen (Benjamini < 0.05).

Category	Term	Benjamini
GOTERM_CC_DIRECT	GO:0043209~myelin sheath	0.003
GOTERM_CC_DIRECT	GO:0042734~presynaptic membrane	0.004
GOTERM_CC_DIRECT	GO:0043005~neuron projection	0.004
GOTERM_CC_DIRECT	GO:0048471~perinuclear region of cytoplasm	0.022
GOTERM_BP_DIRECT	GO:1903861~positive regulation of dendrite extension	0.024
GOTERM_BP_DIRECT	GO:0014002~astrocyte development	0.031

5. References

1. Sanchez L M M, et al. Detection of common developmental patterns in vertebrate by machine learning analysis of tissue gene expression. Mater thesis at BIOTEC, TU Dresden. Published 2015 August 31.
2. Ciucci S, Ge Y, Durán C, et al. Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies. *Sci Rep*. 2017;7:43946. Published 2017 Mar 13. doi:10.1038/srep43946
3. Ferraz AL, Ojeda A, López-Béjar M, et al. Transcriptome architecture across tissues in the pig. *BMC Genomics*. 2008;9:173. Published 2008 Apr 16. doi:10.1186/1471-2164-9-173
4. Pipes L, Li S, Bozinoski M, et al. The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res*. 2013;41(Database issue):D906–D914. doi:10.1093/nar/gks1268
5. Uhlén M, et al. Tissue-based map of human proteome. *Science*. 2015 Jan 23;347(6220):1260419. doi: 10.1126/science.1260419.
6. Gilbert, Scott F. Developmental Biology. 9th ed. Sunderland, MA: Sinauer Associates, 2010: 333-370. Print.