ACM Conference Proceedings Format

|  | Adam Pater-Faranda  University of Delaware  Newark, United States  abf@udel.edu |  |
| --- | --- | --- |

# ABSTRACT

When applied to the analysis of gene expression, high throughput sequencing technology can yield systems-level insights into molecular mechanisms underlying complex biological phenomena. The temporal dynamics of transcriptomic responses are especially informative when it comes to unraveling the sequence of molecular events leading to observable changes in phenotype. In this project, hierarchical clustering methods were used to group genes based on changes in expression following

# INTRODUCTION & PROBLEM STATEMENT

In human cataract patients, posterior capsular opacification (PCO) is common post surgical complication [[1]](https://www.ncbi.nlm.nih.gov/pubmed/19013456).  The lens of the eye is a clear, ellipsoidal structure that focuses incoming light on the retina. It consists of a mass of lens fiber cells, surrounded by a capsular membrane.  The anterior inner surface of this membrane is lined with lens epithelial cells (LEC) [[1]](https://www.ncbi.nlm.nih.gov/pubmed/19013456).   When a cataract forms, these normally clear cells become cloudy and obscure the patient’s vision.  The procedure to remove a cataract involves the removal of the fibers through an incision in the capsular membrane, mechanical disruption and removal of opacified fibers, and placement of an artificial lens inside the capsular bag [[1]](https://www.ncbi.nlm.nih.gov/pubmed/19013456).

Lens epithelial cells left behind after surgery can undergo an epithelial to mesenchymal transition (EMT) and migrate behind the replacement lens, to the capsule’s posterior.  There they can form opaque, fibrotic plaques that interfere with patient’s vision [[1]](https://www.ncbi.nlm.nih.gov/pubmed/19013456). By studying the mechanisms through which the surgical procedure triggers EMT, we may find new ways to prevent, or reduce the likelihood of PCO in cataract patients after their surgery.

The process of EMT in residual LEC involves highly coordinated changes in gene expression.  These changes are believed to be dependent on a variety of extracellular signals that arise as a result of mechanical tissue damage [[2].](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6212802/)With high-throughput RNA sequencing assays (RNA-Seq) it is possible to measure changes in the expression level for thousands of genes simultaneously.  Based on observed changes in expression it may be possible to infer which signals are driving this profound change in phenotype. Using RNA-Sequencing data collected from a mouse model, the goal of this project is to evaluate the suitability of hierarchical clustering methods for identifying time dependent changes in gene expression following cataract surgery.

Hierarchical clustering methods assign data points to clusters based on their distances from one another. Data points that are more similar tend to be assigned to the same cluster. For each sample in an RNA Sequencing experiment, abundance measurements are collected for thousands of genes simultaneously. Clustering methods can be used to evaluate which samples, or which genes are most similar to one another. In this context, the data points being clustered can wither be genes or samples. In this work, various applications of hierarchical clustering were explored in order to mine

# Related work & Background

# METHOD

To simulate cataract surgery, the entire mass of lens fiber cells was surgically removed from the lens capsule. The lens capsule or ‘capsular-bag’ was left in place and the incision was closed with a suture. The capsular bag was later harvested, at intervals between 0 and 48 hours, for RNA extraction, library preparation and sequencing.

Three biological replicates were sequenced for each time point. Each replicate consisted of pooled capsular-bags from 5 individual mice. Simulated cataract surgeries were carried out over a two year period. Capsular bags were harvested at 0, 6, 24 and 48 hours after surgery. Sequencing was conducted at two different laboratories, the Delaware Biotechnology Institute (DBI) and DNA Link (DNA) (Table 1). For the 0 hour and 24 hour time points, each laboratory sequenced there independent replicates.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Interval | 0 Hours | 6 Hours | 24 Hours | 48 Hours |
| Sequencing Lab(s) | DBI, DNA | DNA only | DBI, DNA | DBI only |

**Table 1 Laboratories where sequencing was conducted for each of four time points: The Delaware Biotechnology Institute (DBI) or DNA Link (DNA).**

Raw sequencing reads were aligned to a reference genome using HISAT2. Prior to alignment, single end data generated at DBI was trimmed using TrimGalore. No trimming was applied to paired-end data generated by DNA Link. The genome index used for alignment was built with the most recent Ensembl mouse genome draft (GRCm38.96, retrieved from [ftp.ensembl.org](ftp://ftp.ensembl.org) on April 9, 2019). The GTF file used in this index includes feature for all definitive mouse chromosomes (1-19, X,Y and Mitochondrial) as well as un-assigned scaffolds (entries prefixed with GL\* and JH\*). Features or genes were quantified using htseq-count, to count the number of sequencing reads aligned to each gene. The same GTF file used to build the alignment index was also used with htseq-count for feature quantification.

All counts in the data set were incremented by a constant,   
C = 2, to avoid issues with log transformation of 0. Adjusted counts were normalized using edgeR’s TMM method, which normalizes counts based on library size (total number of reads counted for a given sample and a scaling factor estimated from a given data set (matrix of counts for all samples) << robinson, olshack 2010 Genome Biology>>. Gene expression values used in subsequent clustering and principal components analyses were TMM normalized counts per million (CPM). The Surrogate variable analysis methods provided by the “ComBat” function in the R package “sva” were used to correct for batch effects related to sequencing being done at 2 different labs <<W.E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray data using empirical bayes methods. Biostatistics>>. The effects of batch correction were evaluated by examining PCA plots, and clusters before and after correction. Heirarchical clustering in R was accomplished using the ‘hclust’ function which implements an agglomerative algorithm that accommodates a variety of distance and linkage methods (R documentation). Clustering performance was evaluated based on silhouette width. For Sample clusters (as opposed to gene clusters) performance was evaluated

|  |  |  |
| --- | --- | --- |
| Software (version) | Purpose | Environment |
| HISAT2  () | Align short reads to genome | Biomix HPC |
| TrimGalore () | Trim adapters from raw reads | Biomix HPC |
| htseq-count () | Count reads for each gene | Biomix HPC |
| R (3.3.3) | Data Analysis | OSX 10.9 Desktop |
| edgeR (3.16.5) | Normalize Counts | OSX 10.9 Desktop (R environment) |
| sva  (3.22.0) | Batch Correction. | OSX 10.9 Desktop (R environment) |

**Table X Software used to pre-process RNA Sequencing data, and perform statistical calculations**

# Preliminary experiments & Results.

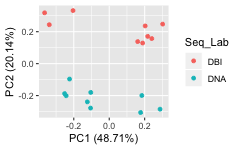
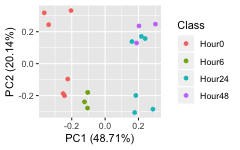
# Related work & Background

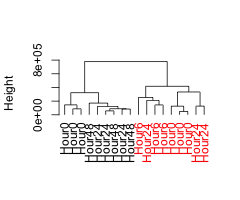
The instructions for protein synthesis are encoded in the genome. The synthesis of a protein depends on an intermediate step, transcription of an mRNA transcript from its corresponding gene in the genome. The mRNA transcript serves as a template used to synthesize the protein << FRANCIS CRICK 1970 Central Dogma>>. Proteins mediate all cellular processes required to sustain life, from the organization of tissues to the metabolism of nutrients. Living cells regulate the relative expression of different genes in order to meet specific metabolic needs and respond to environmental changes.

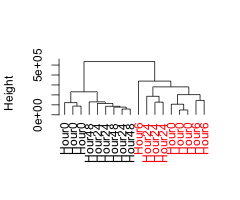
In one of the earliest examples (if not the first) of clustering applied to gene expression, investigators demonstrated how hierarchical clustering based on Euclidean distance with average linkage could be used to highlight coordinated changes in gene expression <<pq014863.pdf>>. Their study demonstrated two important concepts in systems biology that are still relevant today; that changes in gene expression are highly coordinated, and that groups of co-regulated genes tend to have a common biological function <<pq014863.pdf>>. In this specific case Eisen et al’s cluster analysis illustrates how in budding yeast, cell cycle related genes upregulate during sporulation while genes involved in the cell cycle simultaneously down regulate.

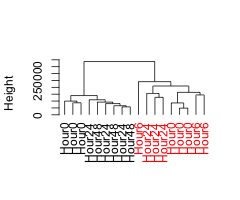
Since then, cluster analysis has become ubiquitous in transcriptomic experiments, with a wide variety of algorithms available based on a number of different computational paradigms. As of 2016, one group of investigators identified nearly 30 different algorithms that have been applied to transcriptomic analysis <<Applications Paper>>. Many of the more traditional algorithms identify clusters based only similarity between gene-wise measurements, however there are also bi-clustering and even tri-clustering based methods that incorporate additional factors such as sample class or time interval. <<Applications Paper>>.

Different types of algorithms each have their own relative strengths and weakness. In the case of hierarchical clustering it is simple to implement, easily interpretable and and for a given set of data yields the same solution every time. Disa <<Applications Paper>>. .

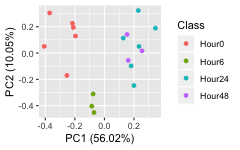
****

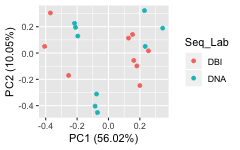
****

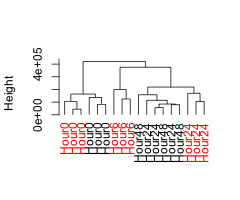
****

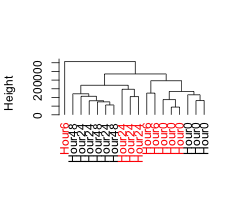
****

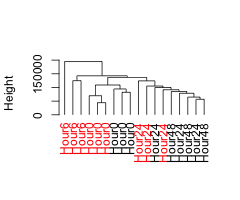
**Figure 1 Evaluation of TMM Normalized CPM for all samples using the top 200 genes (by overall variance). Principal components (A, B). Samples were clustered using agglomerative hierarchical clustering with either complete (C), average (D) or single(E) linkage and Manhattan distance. Leaves with red labels correspond to samples sequenced at DNA Link.**

****

****







**Figure 2 Evaluation of Batch Corrected CPM for all samples using the top 200 genes (by overall variance). Principal Components (A, B). Samples were clustered using agglomerative hierarchical clustering with either complete (C), average (D) or single (E) linkage and Manhattan distance. Leaves with red labels correspond to samples sequenced at DNA Link.**

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 2 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 1.000 | 0.739 | 9(0.82), 9(0.658) |
| Euclidean, Average | 1.000 | 0.739 | 9(0.82), 9(0.658) |
| Euclidean, Single | 1.000 | 0.739 | 9(0.82), 9(0.658) |
| Manhattan, Complete | 1.000 | 0.607 | 9(0.687), 9(0.526) |
| Manhattan, Average | 1.000 | 0.607 | 9(0.687), 9(0.526) |
| Manhattan, Single | 1.000 | 0.607 | 9(0.687), 9(0.526) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 3 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.882 | 0.610 | 9(0.792), 6(0.518), 3(0.251) |
| Euclidean, Average | 0.948 | 0.633 | 9(0.808), 8(0.515), 1(0) |
| Euclidean, Single | 0.948 | 0.633 | 9(0.808), 8(0.515), 1(0) |
| Manhattan, Complete | 0.869 | 0.450 | 9(0.647), 5(0.291), 4(0.206) |
| Manhattan, Average | 0.948 | 0.467 | 9(0.672), 8(0.293), 1(0) |
| Manhattan, Single | 0.948 | 0.467 | 9(0.672), 8(0.293), 1(0) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 4 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.765 | 0.427 | 3(0.423), 6(0.426), 6(0.518), 3(0.251) |
| Euclidean, Average | 0.869 | 0.561 | 9(0.792), 6(0.342), 2(0.462), 1(0) |
| Euclidean, Single | 0.850 | 0.499 | 9(0.785), 3(0.526), 5(0.067), 1(0) |
| Manhattan, Complete | 0.830 | 0.489 | 9(0.621), 3(0.68), 2(0.485), 4(0.051) |
| Manhattan, Average | 0.850 | 0.495 | 9(0.653), 5(0.387), 3(0.366), 1(0) |
| Manhattan, Single | 0.850 | 0.495 | 9(0.653), 5(0.387), 3(0.366), 1(0) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 5 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.752 | 0.378 | 3(0.423), 6(0.426), 6(0.342), 2(0.462), 1(0) |
| Euclidean, Average | 0.817 | 0.515 | 9(0.782), 4(0.292), 2(0.262), 2(0.27), 1(0) |
| Euclidean, Single | 0.810 | 0.516 | 9(0.785), 3(0.41), 3(0.067), 2(0.399), 1(0) |
| Manhattan, Complete | 0.712 | 0.444 | 3(0.501), 6(0.546), 3(0.68), 2(0.485), 4(0.051) |
| Manhattan, Average | 0.732 | 0.434 | 3(0.501), 6(0.546), 5(0.387), 3(0.366), 1(0) |
| Manhattan, Single | 0.810 | 0.488 | 9(0.631), 3(0.574), 3(0.277), 1(0), 2(0.275) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 6 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.699 | 0.337 | 3(0.423), 6(0.426), 4(0.292), 2(0.262), 2(0.27), 1(0) |
| Euclidean, Average | 0.699 | 0.337 | 3(0.423), 6(0.426), 4(0.292), 2(0.262), 2(0.27), 1(0) |
| Euclidean, Single | 0.758 | 0.279 | 8(0.349), 1(0), 3(0.41), 3(0.067), 2(0.399), 1(0) |
| Manhattan, Complete | 0.693 | 0.430 | 3(0.501), 6(0.546), 3(0.659), 2(0.443), 3(0.035), 1(0) |
| Manhattan, Average | 0.693 | 0.438 | 3(0.501), 6(0.546), 3(0.574), 3(0.277), 1(0), 2(0.275) |
| Manhattan, Single | 0.797 | 0.467 | 9(0.621), 3(0.574), 2(0.298), 1(0), 1(0), 2(0.248) |
|  |  |  |  |
| **Table 1 Statistics for hierarchical clustering of TMM normalized counts. Samples were clustered using the top 200 genes (ranked by overall variance). No batch correction was applied prior to clustering. Rand indices were calculated using the categorical time point as the true class.** | | | |

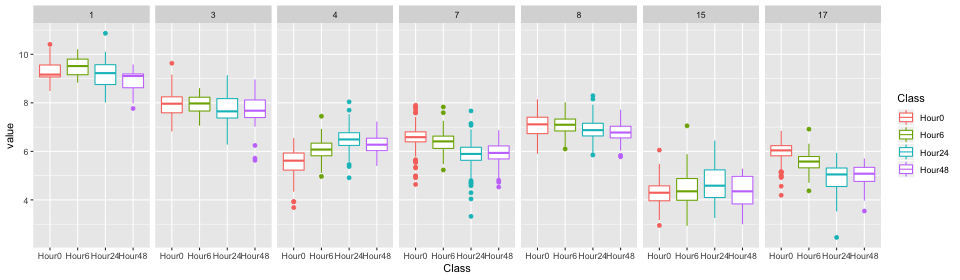
|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 2 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.634 | 0.413 | 13(0.489), 5(0.217) |
| Euclidean, Average | 0.477 | 0.509 | 17(0.539), 1(0) |
| Euclidean, Single | 0.477 | 0.509 | 17(0.539), 1(0) |
| Manhattan, Complete | 0.471 | 0.312 | 6(0.514), 12(0.212) |
| Manhattan, Average | 0.477 | 0.325 | 17(0.344), 1(0) |
| Manhattan, Single | 0.477 | 0.325 | 17(0.344), 1(0) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 3 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.608 | 0.341 | 13(0.385), 4(0.282), 1(0) |
| Euclidean, Average | 0.608 | 0.341 | 13(0.385), 4(0.282), 1(0) |
| Euclidean, Single | 0.490 | 0.289 | 16(0.325), 1(0), 1(0) |
| Manhattan, Complete | 0.529 | 0.388 | 6(0.496), 9(0.366), 3(0.238) |
| Manhattan, Average | 0.516 | 0.336 | 8(0.323), 9(0.386), 1(0) |
| Manhattan, Single | 0.516 | 0.132 | 15(0.112), 1(0), 2(0.349) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 4 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.608 | 0.301 | 6(0.28), 7(0.398), 4(0.238), 1(0) |
| Euclidean, Average | 0.582 | 0.244 | 1(0), 12(0.278), 4(0.264), 1(0) |
| Euclidean, Single | 0.484 | 0.068 | 15(0.082), 1(0), 1(0), 1(0) |
| Manhattan, Complete | 0.647 | 0.355 | 6(0.444), 6(0.42), 3(0.186), 3(0.215) |
| Manhattan, Average | 0.516 | 0.346 | 6(0.397), 9(0.353), 1(0), 2(0.34) |
| Manhattan, Single | 0.516 | 0.346 | 6(0.397), 9(0.353), 1(0), 2(0.34) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 5 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.582 | 0.257 | 6(0.262), 7(0.36), 2(0.186), 2(0.078), 1(0) |
| Euclidean, Average | 0.562 | 0.174 | 1(0), 11(0.191), 1(0), 4(0.258), 1(0) |
| Euclidean, Single | 0.497 | 0.108 | 14(0.139), 1(0), 1(0), 1(0), 1(0) |
| Manhattan, Complete | 0.634 | 0.336 | 6(0.386), 6(0.42), 3(0.179), 1(0), 2(0.34) |
| Manhattan, Average | 0.634 | 0.336 | 6(0.386), 6(0.42), 3(0.179), 1(0), 2(0.34) |
| Manhattan, Single | 0.510 | 0.246 | 6(0.336), 9(0.268), 1(0), 1(0), 1(0) |

|  |  |  |  |
| --- | --- | --- | --- |
| Number of Clusters, k = 6 | | | |
| Method | Rand Index | Mean Silhouette | Size (Mean Silhoute by cluster) |
| Euclidean, Complete | 0.556 | 0.254 | 6(0.235), 1(0), 6(0.445), 2(0.169), 2(0.078), 1(0) |
| Euclidean, Average | 0.536 | 0.133 | 1(0), 11(0.163), 1(0), 2(0.225), 2(0.078), 1(0) |
| Euclidean, Single | 0.549 | 0.156 | 12(0.205), 1(0), 2(0.172), 1(0), 1(0), 1(0) |
| Manhattan, Complete | 0.621 | 0.305 | 6(0.386), 6(0.354), 2(0.188), 1(0), 1(0), 2(0.334) |
| Manhattan, Average | 0.621 | 0.305 | 6(0.386), 6(0.354), 2(0.188), 1(0), 1(0), 2(0.334) |
| Manhattan, Single | 0.536 | 0.234 | 6(0.336), 8(0.275), 1(0), 1(0), 1(0), 1(0) |
|  |  |  |  |
| **Table 2 Statistics for hierarchical clustering of Batch Corrected, TMM normalized counts. Samples were clustered using the top 200 genes (ranked by overall variance). Surrogate variable analysis was used to correct for variation introduced by sequencing of samples at different labs. Rand indices were calculated using the categorical time point as the true class.** | | | |



r citation) and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced (e.g., “[Borriello, personal communication]”).

References should be in ACM citation format: <http://acm.org/publications/submissions/latex_style>. This includes citations to internet resources [**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**] according to ACM format, although it is often appropriate to include URLs directly in the text, as above.

Occasionally MS Word generates larger-than-necessary PDF files when images inserted into the document are manipulated in MS Word. To minimize this problem, use an image editing tool to resize the image at the appropriate Occasionally MS Word generates larger-than-necessary PDF files when images inserted into the document are manipulated in MS Word. To minimize this problem, use

**Figure: Intersection between gene lists for pairwise contrasts in samples analyzed at DBI Only**

**Figure: Intersection between gene lists for pairwise contrasts in samples analyzed at DNA Link Only**

**Figure: Intersection between gene lists for pairwise contrasts in samples analyzed at both sequencing labs.**

an image editing tool to resize the image at the appropriate Occasionally MS Word generates larger-than-necessary PDF files when images inserted into the document are manipulated in MS Word. To minimize this problem, use an image editing tool to resize the image at the appropriate printing resolution (usually 300 dpi), and then insert the image into Word using Insert | Picture | From printing resolution (usually 300 dpi), and then insert the image into Word using Insert | Picture | From printing resolution (usually 300 dpi), and then insert the image into Word using Insert | Picture | From File.

# METHOD

The heading of a section should be in Arial 9-point bold, all in capitals (Heading 1 style). Sections should not be numbered. Subsections

Headings of subsections should be in Arial 9-point bold with initial letters capitalized (Heading 2 style). For sub-sections and sub-subsections, a word like *the* or *of* is not capitalized unless it is the first word of the heading.

### Sub-subsections

Headings for sub-subsections should be in Arial 9-point italic with initial letters capitalized (Heading 3 style).

# Preliminary experiments & Results

Place figures and tables at the top or bottom of the appropriate column or columns, on the same page as the relevant text (see Figure 1). A figure or table may extend across both columns to a maximum width of two columns, or 17.78 cm (7 in.).

Captions should be Times New Roman 9-point bold (Caption style). They should be numbered (e.g., “Table 1” or “Figure 2”), centered, and placed beneath the figure or table. The words “Figure” and “Table” should be spelled out (e.g., “Figure” rather than “Fig.”) wherever they occur.

All figures should also include alt text for improved accessibility. In Word, right click the figure, and select Format Picture | Layout | Alt Text). Papers and notes may use color figures, which are included in the page limit; the figures must be usable when printed in black-and-white in the proceedings.

The paper may be accompanied by a short video figure up to five minutes in length. However, the paper should stand on its own without the video figure, as the video may not be available to everyone who reads the paper.

## Inserting Images

Occasionally MS Word generates larger-than-necessary PDF files when images inserted into the document are manipulated in MS Word. To minimize this problem, use an image editing tool to resize the image at the appropriate