Mackenzie Fleischer

CSC 7300 Project

Bioinformatics Applications in Osteoarthritis

Introduction:

Osteoarthritis (OA) is the most common chronic joint condition affecting 27 million Americans, causing a significant health care system burden. It is characterized, by the breakdown of cartilage that lines the bones of a joint causing symptoms of pain, stiffness, and swelling. To date, there is no known cure for OA but symptoms can typically be conservatively managed with medications and physical therapy. However, in advanced stages of OA many patients bone remodeling requiring joint replacement surgery to alleviate symptoms. Resultantly, research efforts are being made to identify the pathology leading to OA and potential therapeutic options. One useful area of research is the development of diagnostic methods to distinguish OA from other types of chronic arthritis, such as Rheumatoid Arthritis (RA). For example a study by Woetzel et al. (GSE55584) aimed to identify OA patients based of differentially expressed genes isolated from synovial fluid samples.

Data Description:

This study collected synovial fluid samples during joint replacement procedures from 16 patients clinically diagnosed with either RA (n= 10) or OA (n=6). Total RNA was isolated from these samples and processed for microarray analysis. The microarray data was processed using Affymetrix Microarray Suite and was made publically available on the GEO Database.

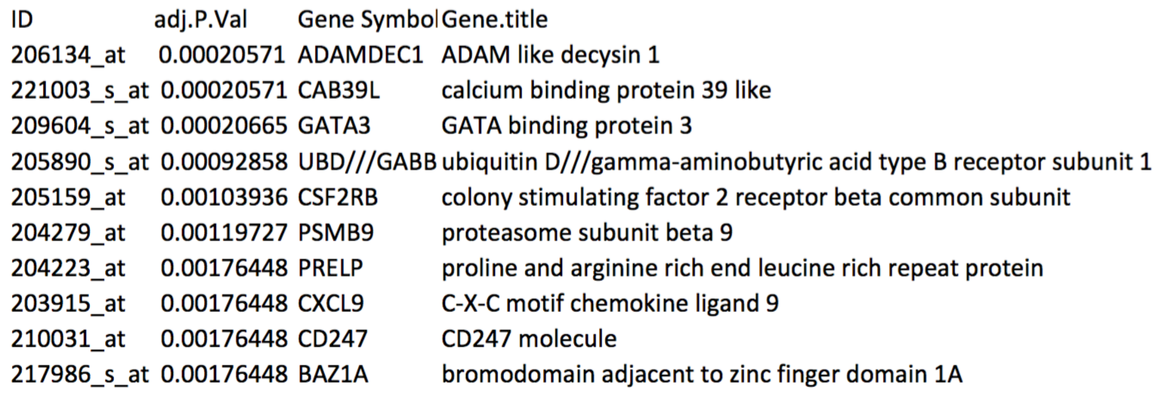
Analysis Methods:

Initial analysis of the microarray data was performed with GO2R. This analysis included a log2 transformation and the eBayes function (t-test) to identify differentially expressed genes. The top 250 genes with a p-value < 0.05 were used to employ clustering methods. A k-means clustering function was used as well as a c-means fuzzy clustering method. Next, a heat map was generated to further identify subjects with similar gene expression. Lastly the top differentially expressed genes will be analyzed using GeneMania to identify gene relationships.

These analysis methods differ slightly from the methods described in the paper because prior to clustering they used a training data set to establish a set of rules that assign a rank to a gene. For example the clustering methods may weight one gene more heavily than another gene when deciding which group to assign the subject. Since the pathway analysis was based on these rules and used a software for purchase GeneMania was used instead.

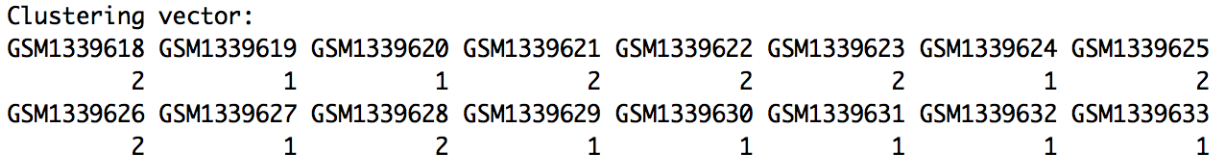
Results & Discussion:

The top genes resulting from GO2R analysis were different from those described in the paper as a result of implementing the "rules". The top 10 genes obtained are listed below. Research of these genes indicates many of them play a role in inflammation and T-cell regulation. This makes sense as arthritis is known to be an inflammatory setting.

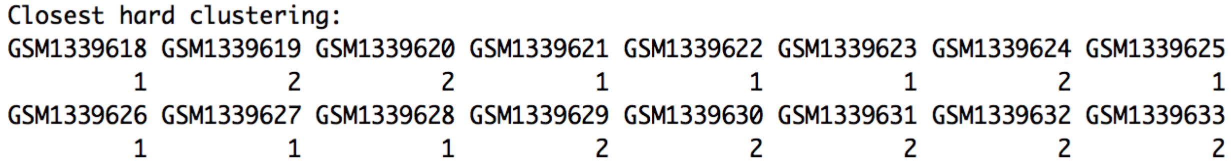


ID	adj.P.Val	Gene Symbol	Gene.title
206134_at	0.00020571	ADAMDEC1	ADAM like decysin 1
221003_s_at	0.00020571	CAB39L	calcium binding protein 39 like
209604_s_at	0.00020665	GATA3	GATA binding protein 3
205890_s_at	0.00092858	UBD///GABB	ubiquitin D///gamma-aminobutyric acid type B receptor subunit 1
205159_at	0.00103936	CSF2RB	colony stimulating factor 2 receptor beta common subunit
204279_at	0.00119727	PSMB9	proteasome subunit beta 9
204223_at	0.00176448	PRELP	proline and arginine rich end leucine rich repeat protein
203915_at	0.00176448	CXCL9	C-X-C motif chemokine ligand 9
210031_at	0.00176448	CD247	CD247 molecule
217986_s_at	0.00176448	BAZ1A	bromodomain adjacent to zinc finger domain 1A

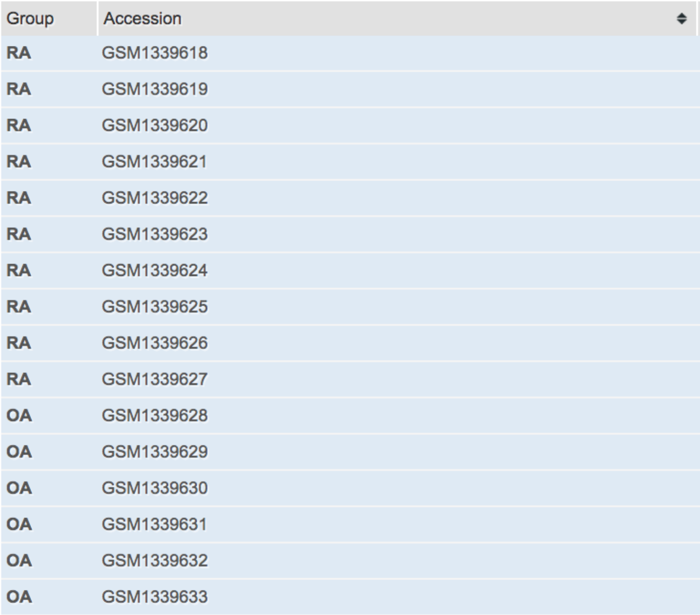
Next, I performed k-means clustering. The following results were obtained in which 9 subjects were sorted into the first cluster and 7 into the 2nd cluster. Compared to the known clinical diagnosis, 4 subjects were placed into the wrong group yielding a 70% accuracy. A c-means clustering was also performed as described in the paper. This is a fuzzy method in which, a data point can be assigned to one or more cluster until the final iteration in which a hard cluster is chosen. The c-means method sorted the subjects into two groups of 8. Compared to the known clinical diagnosis, 5 subjects were placed into the wrong group yielding a 68% accuracy. Both of these methods produced an accuracy less than what was obtained in the referenced paper. By applying the "rules" prior to clustering, the authors were able to obtain an overall accuracy of 91% using fuzzy c-means.



```
Clustering vector:
GSM1339618 GSM1339619 GSM1339620 GSM1339621 GSM1339622 GSM1339623 GSM1339624 GSM1339625 
         2          1          1          2          2          2          1          2 
GSM1339626 GSM1339627 GSM1339628 GSM1339629 GSM1339630 GSM1339631 GSM1339632 GSM1339633 
         2          1          2          1          1          1          1          1 
```

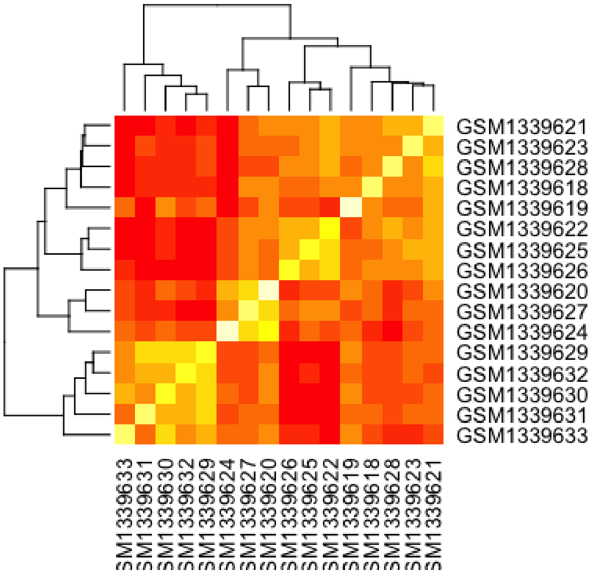


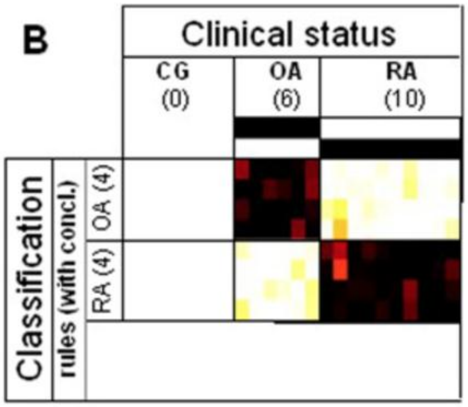
```
Closest hard clustering:
GSM1339618 GSM1339619 GSM1339620 GSM1339621 GSM1339622 GSM1339623 GSM1339624 GSM1339625 
         1          2          2          1          1          1          2          1 
GSM1339626 GSM1339627 GSM1339628 GSM1339629 GSM1339630 GSM1339631 GSM1339632 GSM1339633 
         1          1          1          2          2          2          2          2 
```



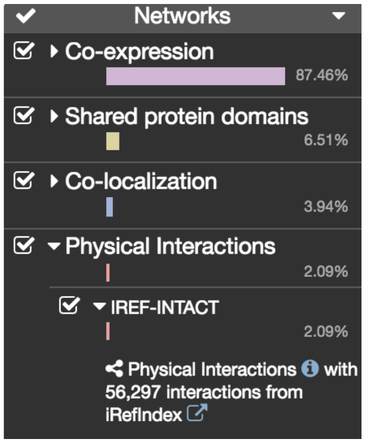
Group	Accession
RA	GSM1339618
RA	GSM1339619
RA	GSM1339620
RA	GSM1339621
RA	GSM1339622
RA	GSM1339623
RA	GSM1339624
RA	GSM1339625
RA	GSM1339626
RA	GSM1339627
OA	GSM1339628
OA	GSM1339629
OA	GSM1339630
OA	GSM1339631
OA	GSM1339632
OA	GSM1339633

A heat map was generated in R using the top 250 differentially expressed genes. As seen below, the resulting clustering is similar to the heat map in the manuscript. The areas of light (white, yellow, and orange) indicate samples that had the most similar gene expression while the dark areas (red) indicated subjects with the most different gene expression. As a result, the groups have 5 and 11 subjects each. This clustering and shading is similar to the heat map seen in the paper suggesting similar results were produced.





GeneMania was used to identify genes relationships including co-expression, shared protein domains, co-localization, and physical interations. Using the top 25 differentially expressed genes, 87.5 were found to be co-expressed with another gene on the list. 6.51% of the genes shared protein domains meaning the proteins encoded for by the gene are of a similar protein family (example: CD247 and FCER1G). 3.94% of the genes are known to be co-localized meaning they are expressed in the same tissues (example:). Lastly, 2.09% of the genes were positive for a physical interaction meaning their protein products were found to interact with each other (example: CCR7 and CCL21). GeneMania also provides useful information about what studies these interactions have been identified and what role they are known to play.



Networks

☑ Co-expression 87.46%

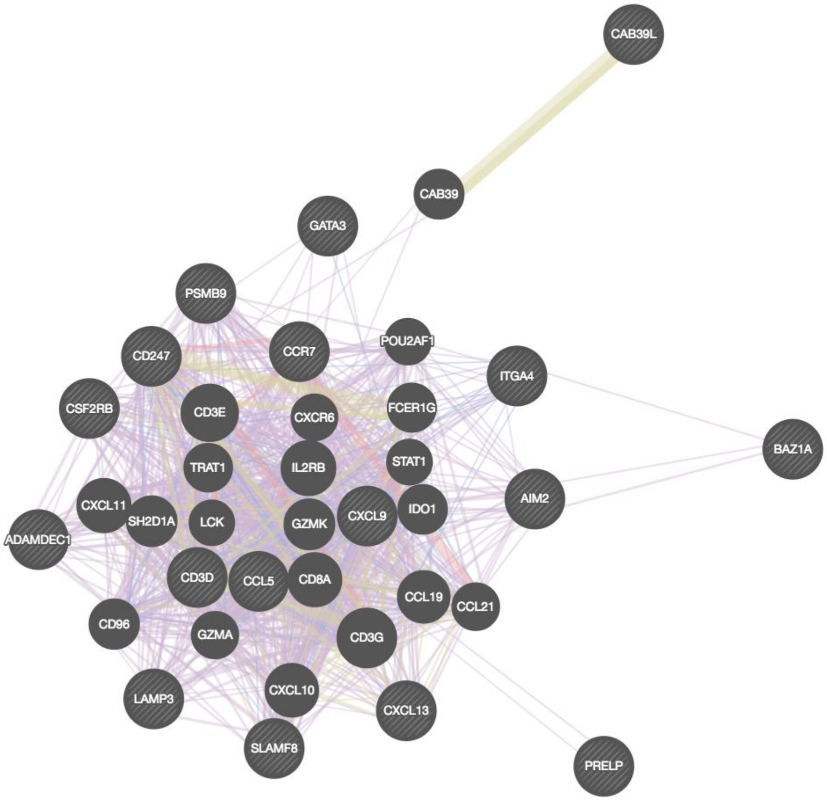
☑ Shared protein domains 6.51%

☑ Co-localization 3.94%

☑ Physical Interactions 2.09%

☑ IREF-INTACT 2.09%

Physical Interactions with 56,297 interactions from iRefIndex



Conclusion:

In conclusion, identifying patients with OA purely on the basis of differential gene expression and clustering methods is not very accurate. However, employing rule-based clustering may help increase the accuracy, which may allow for identification of pathogenetically or therapeutically relevant genetic targets.

Bibliography:

1. http://www.arthritis.org/about-arthritis/types/osteoarthritis/
2. https://www.niams.nih.gov/health_info/osteoarthritis/osteoarthritis_ff.asp
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060460/#B46
4. http://www.webmd.com/osteoarthritis/osteoarthritis-causes
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694558/
6. http://genemania.org/