Identification of co-regulated gene modules

Zhang Qiong Computational Biology Programme

1. Introduction

Genes do not act alone in a biological system. Instead, a group of genes and their protein products act together as a module to carry out their biological roles. The identification of ~~their potential~~ co-regulated modules ~~is~~ challenging problem. In this project, information of putative transcription binding site motifs, gene ontology annotations and gene regulation pathways, which all have implications of a group of genes being co-regulated, are integrated. A workflow is proposed with detailed steps described, to generate a list of potential co-regulated gene modules given any set of input genes of Homo Sapien. The size of the generated list is also adjustable according to need, by setting proper parameters.

2. Methodology

2.1 Putative TFBS Motifs and first round of bi-clique mining

The promoter annotations of each respiratory disease associated gene are obtained from online database of respiratory genomics (Chowdhary, 2011). In the motif-centred view, the information of putative transcription factor binding site (TFBS) motifs signatures is available for each target gene.

~~Such motif gene network is further structured into a bi-partite graph view, with one vertex set being all genes related to a particular disease, and the other set being all motifs that appear in at least one of the genes. An edge is added between a gene vertex in set V1 and a motif vertex in set V2, wherever the gene contains the motif. Such a bipartite G is called a bi-clique if, for every v1 ϵ V1 and v2 ϵ V2, there is an edge between v1and v2. And it is called a maximum bi-clique if it is not a proper subset of another bi-clique.~~ A complete set of maximum bi-cliques are mined in this bi-partite graph, with threshold of at least three genes in set V1 and at least one motif in V2. All data are formatted in a way that could readily utilize the algorithm and

codes developed by Liu Guimei (Jinyan Li, 2005). Each maximum bi-clique corresponds to a group of genes sharing some common motifs.

2.2 Gene Ontology annotations and second round of bi-clique mining

The Gene Ontology (GO) project is used to address the need for consistent descriptions of gene products properties in different databases. Its website provides associations between gene products and GO terms submitted by members and associates of the GO consortium. Used as preliminary sources, the annotation file downloaded is processed in the following steps. Firstly, GO annotations are filtered according to GO evidence code. Annotations of obsolete evidence without experimental study or computational analysis are erased from further analysis. These include annotations with author/curator statement, annotations without available biological data, untraceable or not-recorded annotations, and electronic automatically assigned annotations done by key word mapping (Guide to GO evidence codes, 2011). Secondly, GO annotations are filtered according to GO quantifiers. Those with "NOT", "contribute_to", and " colocalizes_with" are erased, which are know not to be annotated, annotated to complex unit whose function not yet known, or with inaccurate resolution (The Gene Ontology quanlifiers, 2011). Thirdly, as the GO terms are structured as a graph with parent-child relationship, only explicit annotations are recorded which corresponds to only the annotations of genes to the most specific GO terms it could annotate to. To restore the full annotations, the Gene Ontology Bioinformatics toolbox in Matlab is utilized to retrieve the ancestors of each GO term and same annotations are assigned to the ancestors as well. Last but not least, among all left GO annotations, only GO terms with high information content(IC) are ~~remained~~. The information content is based on frequency of terms p, which is mathematically represented as IC=-log(p). The frequency is calculated as the number of annotation instances for that GO term and its descendent terms divided by the total number of annotations. In our case, a threshold of 30(can be further increased to loosen the filter) is used, which is the total number of gene annotations a particular GO term has. If a GO term is too general and annotated to more than 30 genes, it is not considered as informative and thus removed from the annotation list.

Using the same idea of bi-clique mining, such data is further structured into a bi-partite graph view, with one vertex set being all genes related to a particular disease, and the other set being GO terms that appear in at least one of the genes. An edge is added between a gene vertex in set V1 and a motif vertex in set V2, wherever there is a GO annotation recorded. A complete set of maximum bi-cliques are mined in this bi-partite graph, with threshold of at least three genes in set V1 and at least one GO term in V2. Mining as such serves as a second filter of gene groups obtained in previous session of motifs bi-clique mining, since only overlapping groups are retained. Up to this stage, the bi-cliques are groups of genes with at least one common motif and one common GO annotation.

2.3 Gene pathways and test of significance value

For flexibility of further processing and integration of pathways data and their associated genes, the raw data from integrated pathways database (Hufeng, 2011) for Homo Sapien is used instead of KEGG pathway database. The data comes in a format of a list of gene pairs and the corresponding pathway they co-occur. And it is converted to a list of genes with the pathways each gene belongs to.

The bi-cliques obtained from the previous two sessions are considered as groups of candidates of co-regulated genes, with the significance level of each assessed in the following manner. To calculate the significance level of a size-n bi-clique, 100 million times of random drawings of size n is done from a same gene list considered in both motif and GO annotation bi-clique mining, or more accurately the overlapping of genes in case there are genes with no available information from either GO annotations or pathways (all genes have motifs information since they are originally imported in a gene list from the motifs data for a particular disease). The p-value is expressed as the number of random groups out of all drawings that have performance as good as or better than the gene group of interest. After mining bi-cliques, but this time with the second vertex set being gene pathways, the performance measure of a gene group being co-regulated is defined as the largest bi-clique it contains. In case of ties, performances are further distinguished by the number of common pathways within the largest bi-clique. Performances of random drawings for each size of gene groups are pre-calculated to a single check-up table in

order to save computation time, from where the new gene group to be assessed retrieves its ranking according to the calculated score, which leads to a p value. The list of co-regulated gene groups is ranked according to the significance level.

3. Results and Discussions

3.1. Bi-clique mining using the information of putative TFBS motifs gives potential co-regulated gene modules but in large number

Table 1. Bi-clique mining results of a list of 61 allergic-rhinitis associated genes, being vertically increasing minimum bi-clique gene size and horizontally increasing minimum bi-clique motif size

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
2	2491	2398	1711	1017	602	404	287	201	151	108	68	54	34	26	18	13	13	13	13	13	13	10
3	1958	1865	1214	620	319	200	129	91	65	43	22	16	7	5	5	5	5	5	5	5	5	4
4	1317	1225	696	293	144	80	48	31	17	7	4	1	1	1	1	1	1	1	1	1	1	0
5	815	726	341	105	40	17	9	4	1	1	0	0	0	0	0	0	0	0	0	0	0	
6	470	389	123	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	263	190	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	147	80	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	79	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	58	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	33	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	28	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Bi-clique mining is a useful technique for its capability to search for group of genes with common promoter motifs region. It has been suggested that the attempt to analyze a group of genes' promoter sequences to identify transcription factors is very crucial in identifying possible coordinated regulation of the group of genes (Veerla, 2006). And it has been shown, in a similar approach, that promoter clustering based

on motif similarity greatly enhances the identification of common TFBS patterns in co-expressed genes (Veerla, 2006). Given the assumption that co-regulated genes show enrichment for common transcription binding sites in their promoter regions, it is indicated that the common motifs detected using bi-clique mining give sufficient insights to the possible co-regulation of the genes.

Though providing evidence on the possibility of genes' co-regulated with each other in a same bi-clique, with the motifs information alone, a large number of gene groups are obtained. It expands a wide range of gene group size to up 21, even when just using a short list of 61 allergic-rhinitis associated genes, as shown in Table 1. In particular, 1958 bi-cliques are obtained with at least three genes containing one common motif, among which many may not be really co-regulated due to possible false positives of motifs information. Besides, sizes of gene groups need to be reduced first before further proceed to assess and rank their significance levels. It is due to the time constraint in generating random drawings and table computation for each size of gene group ranging from 3 all the way up to 21, even for a very small group of disease gene list. To further narrow the groups of co-regulated genes candidates, a second bi-clique mining procedure is carried out with information of Gene Ontology annotations.

3.2 Integration of Gene Ontology annotations and gene pathway information helps narrow down and give rankings to the co-regulated gene module candidates

Table 2. List of filtered co-regulated gene groups in the order of decreasing significance level, 3th and 4th columns combined as the performance measure. Only the highest 20 ranking groups and lowest 5 ranking groups are listed. Input genes are a list of 379 asthma associated genes.

index	p value	gene group size	largest bi-clique size	number of common pathways in largest bi-clique	gene1	gene2	gene3	gene4
1	4.00E-05	4	4	1	ADIPOQ	TGFB1	PPARG	GATA3
2	6.00E-05	3	3	1	TNF	CASP8	FASLG	

3	5.24E-03	4	3	1	TGFBR3	TGFBR1	SMAD3	SMAD2
4	7.69E-03	3	3	1	GATA2	PPARG	GATA3	
5	7.69E-03	3	3	1	ADIPOQ	TGFB1	GATA3	
6	7.69E-03	3	3	1	TLR4	TLR6	TLR1	
7	7.69E-03	3	3	1	IL8	CCL2	TNFSF4	
8	7.69E-03	3	3	1	IL6	TLR9	TLR7	
9	7.69E-03	3	3	1	NOD1	TLR6	TLR1	
10	7.69E-03	3	3	1	IL1B	TLR7	TLR2	
11	7.69E-03	3	3	1	IL12A	IL12B	CCL5	
12	7.69E-03	3	3	1	IL12A	IL12B	MYD88	
13	7.69E-03	3	3	1	IL6	IL12A	CCL5	
14	1.52E-02	3	2	5	MIF	TNF	IFNG	
15	1.52E-02	3	2	5	IL17A	TNF	IFNG	
16	1.73E-02	3	2	4	MYB	TLR4	TNF	
17	1.73E-02	3	2	4	MYD88	TLR4	TNFSF4	
18	1.73E-02	3	2	4	IL4	TLR4	TNF	
19	1.73E-02	3	2	4	TNF	TLR4	ACVR1	
20	2.46E-02	3	2	3	CCL5	EDN1	TNF	
			…					
117	9.11E-01	3	1	1	TNFSF4	IRF4	IL9	
118	9.11E-01	3	1	1	IL33	GHRL	GATA3	
119	9.11E-01	3	1	1	TGM2	SLC11A1	CCL2	
120	9.11E-01	3	1	1	SLC11A1	CCL2	HLA-DPA1	
121	9.11E-01	3	1	1	HLA-DRB4	HLA-DRB5	HLA-DRB3	

Effective GO terms are extracted from the gene ontology annotation database, with ambiguous annotations and uninformative terms removed. Explicit annotations are also recovered by retrieving the ancestors of each GO term in an annotation pair. A second round of bi-clique mining is applied with the information of gene ontology annotations for each gene, after which bi-cliques obtained from the previous session with no common gene ontology annotations are removed from the final candidate list. In this manner, 11433 bi-cliques mined using the TFBS motifs information are largely reduced to 121 groups for further significance level ranking and exhaustively-possible experimental validation. One important note is that according to need, size of this list could be further increased to be more than 121, by flexibly setting the threshold of defining a informative GO term to be larger than 30.

Though coordinated regulated genes could generally contain similar promoter regions and share common gene ontology annotations, however, the opposite way, the conclusion of a given gene group with common motifs and gene ontology annotations

being co-regulated is not automatically reached. Given candidate 121 gene groups filtered by second step of bi-clique mining using gene ontology information, which are suspected to being co-regulated because they share some common motifs and GO terms, our null hypothesis is that they are not co-regulated at all and their performance are as good as randomly selected gene groups of the same size. Such test is carried out using an additional property of gene regulation pathways. It is ranked according to p value calculated, which is the proportion of random gene groups that have as good as or performance than the candidate gene groups. Randomization is done to ensure that each gene is independently drawn. As is shown in the table 2, the highest ranking 20 groups have fairly significant levels of their being co-regulated, with p values all smaller than 0.03. Whereas for the lowest 5 ranking groups, the significance levels are extremely low that we could well rule out their possibility of being co-regulated. However, since this manner of carrying p value calculation gives very strict initial null hypothesis of assuming they are not co-regulated at all despite of the observations that they indeed share some common motifs and gene ontology annotations(for example, in a bayesian setup the prior probability of such hypothesis can be distinguished to be rather low), thus the p values calculated are only used as a reference to give the rankings. To reject a null hypothesis, a common p value of 0.05 might be used as a threshold. In the case of a p value larger than 0.05, one might still be interested in examining through the list, for the reasons just given.

3.3 Examining down from the top of the list gives implications of novel gene modules and examining up from the bottom of the list gives implications of possible documentation inconsistencies

Some of the top listed gene modules already have experimental evidence among some or all of them. Some examples of examining the first several groups give the following observations. Examination of 1st group reveals interesting fact of ADIPOQ and PPARG actually being genetic polymorphism, where there is experimental evidence for interaction between them in determining type 2 diabetes intermediate phenotypes (Lovisa, 2009). And another pair of genes, PPARG and GATA3, is also coordinated in the process of suppressing adipocyte differentiation (Dalgin, 2005). Besides pair wise association, there are also established evidence for

the whole 2nd group being co-regulated with each other, belonging to the same tumor necrosis superfamily (TNF) of ligands and receptors. Besides, in cells of the embryonic palate, functional activation of the Smad pathway by TGFβ1, TGFβ2, and TGFβ3 was demonstrated and detected, which gives evidence to the whole 3rd group (Robert M. Greene, 2003). The rest groups could be examined in the similar way, either to verify existing co-regulated gene modules, or to propose potential novel association among genes for further experimental study. On top of examining down the list, an opposite direction could give some implications as well. Examining last several gene groups does not give any significant evidence specifying their associations, despite of their sharing some common motifs and GO terms. It may due to an unrevealed role of one gene in a pathway that some other genes exist, which is the incomplete record of the pathway information. It may also be, subject to further experimental verification, that there are inaccuracies in either the Gene Ontology annotations documentation or motifs discovered that may further require corrections.

3.4 Verified co-regulated gene modules give implications on possible associated transcription factor of and associated function of novel motifs discovered via computational tools

Table 3. Group of genes with common motifs and gene ontology annotations

Gene names	Motifs	Gene ontology terms
ADIPOQ, TGFB1,PPARG, GATA3	127 (unkown)	GO:0045598(regulation of fat cell differentiation)
TNF,CASP8, FASLG	97 (M00080, M00082, M00079)	GO:0008624(induction of apoptosis by extracellular signals)
TGFBR3,TGFBR1, SMAD3, SMAD2	246 (M01107)	GO:0030501(positive regulation of bone mineralization); GO:0007179(transforming growth factor beta receptor signaling pathway)

The TFBS motifs for a particular set of associated genes are consist of both discovered motifs with correspondence to known database such as TRANSFAC and JASPAR and novel motifs that annotated by computational pipeline that comprises an ensemble of the most state-of-the-art programs developed for promoter analysis. In our case, the gene group candidates with experimental evidence in the Table 2 list could be regarded as successfully identified co-regulated gene modules. After that, further inference could be made concerning the common motifs they contain, among which the novel TFBS motifs may be associated with the function of or have the binding site of the known transcription factors associated with the discovered TFBS motifs. In the case if all common motifs are unknown or there is just one common motif, implications could be made by looking at the common gene ontology annotation. In table 3, top listed three groups are examined in this manner as an example. The common gene ontology term of ADIPOQ, TGFB1, PPARG, GATA3 is the regulation of fat cell differentiation, which is in consist with the literature verified in section3.3. The common motif 97, which is unknown in database, is thus very likely to be associated with this function as well. Common motifs of other two gene groups are already known, and the corresponding transcription factors are indicated in the table.

4. Conclusion and further remarks

The described workflow in this project provides methods to identify a list of potential co-regulated gene modules with rankings. The size of the list is adjustable subject to practical needs, by increasing the gene ontology informative content threshold value. During the process, information of putative transcription binding site motifs, gene ontology annotations and gene regulation pathways, which all have implications of a group of genes being co-regulated, are integrated. The generated list of gene modules could be used to propose new co-regulated gene associations when examined top down, or to refine the inconsistency among different documentations when examined from bottom up, of either TFBS motifs, gene ontology annotations or gene regulatory pathways.

The workflow comes in distinct steps as templates that are easy for future modification. Other possible ways of re-arranging the workflow include use the motifs information alone for bi-clique mining and ranking the list with two sets of p values calculated from gene ontology annotations and gene regulatory pathways respectively. This is especially suitable in situation where the need to have a set complete potential co-regulated gene modules is high. A second possible way of modification would be to use gene ontology annotations and gene regulatory pathways to carry first two rounds of bi-clique mining, and use motifs information to calculate p values.

Bibliography

Chowdhary, R. (2011). *Transcriptional Regulation Analysis of Respiratory Associated Disease Genes* . Retrieved from http://www.respiratorygenomics.com/ASDScan/index.html

Dalgin, T. (2005). Interaction between GATA and the C/EBP family of transcription factors is critical in GATA-mediated suppression of adipocyte differentiation. *Molecular cell biology , 25* (2), 706-715.

Guide to GO evidence codes. (2011). Retrieved from The Gene Ontology : http://www.geneontology.org/GO.evidence.shtml

Hufeng, Z. (2011, June 20th). *Integrated Pathways H.sapiens*. Retrieved 2011, from Integrated Pathways: https://sites.google.com/site/integratedpathways/

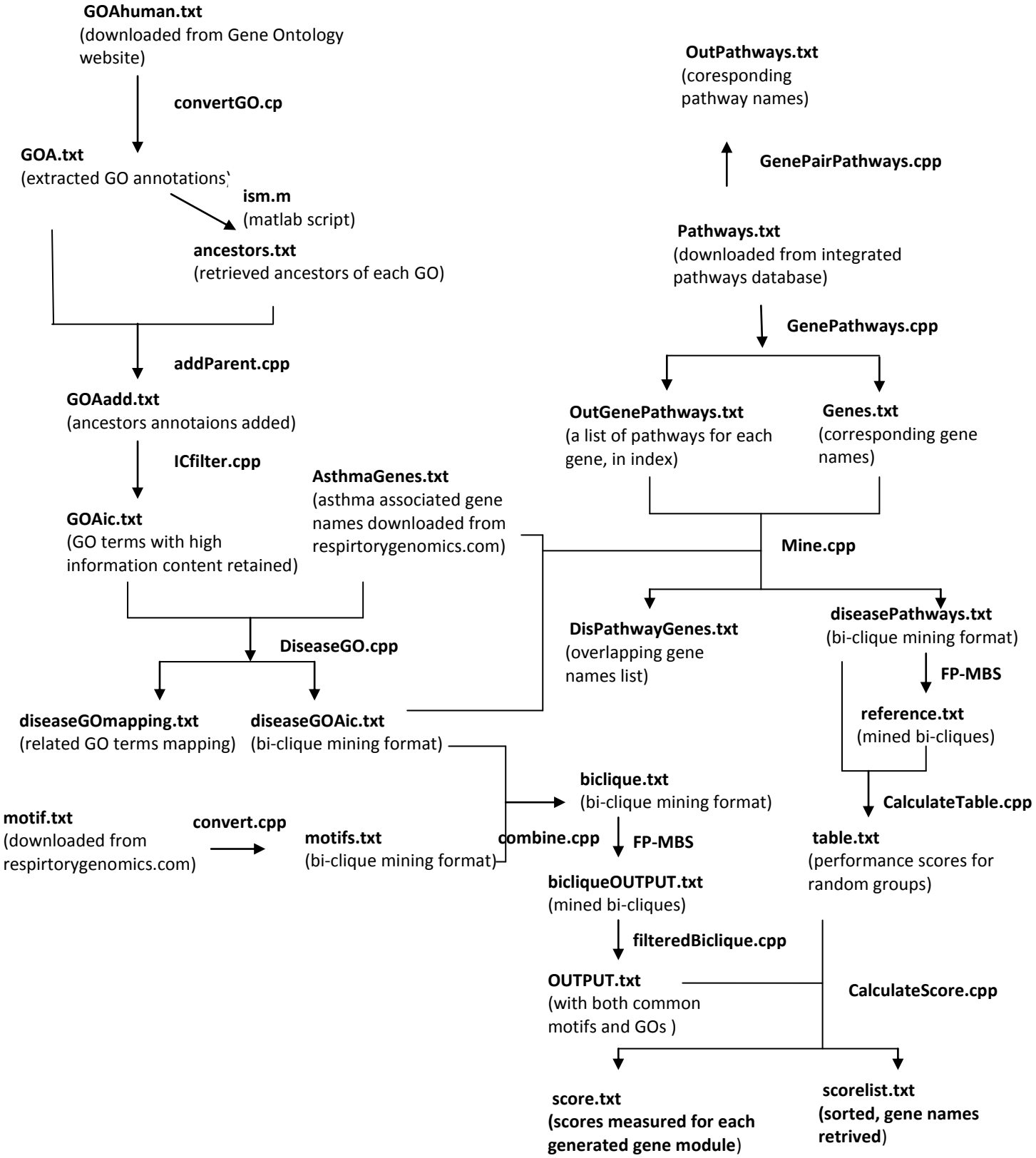
Jinyan Li, H. L. (2005, Octorber). A Correspondence Between Maximal Complete Bipartite Subgraphs and closed patterns. *Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases* , 146--156.

Lovisa, J. (2009). Interaction between PPARG Pro12Ala and ADIPOQ G276T concerning cholestrol in childhood obesity. *International Journal of Pediatric Obesity , 4* (2), 119-121.

Robert M. Greene, P. N. (2003). Intracellular dynamics of Smad-mediated TGFβ signaling. *Journal of cell physiology* , 10.

The Gene Ontology quanlifiers. (2011). Retrieved from The Gene Ontology: http://www.geneontology.org/GO.qualifiers.shtml

Veerla, S. (2006). Analysis of promoter regions of co-expressed genes identified by microarray analysis. *BMC Bioinformatics* , 7:384.


GOAhuman.txt
(downloaded from Gene Ontology website)
convertGO.cp
GOA.txt
(extracted GO annotations)
ism.m
(matlab script)
ancestors.txt
(retrieved ancestors of each GO)
addParent.cpp
GOAadd.txt
(ancestors annotaions added)
ICfilter.cpp
GOAic.txt
(GO terms with high information content retained)
AsthmaGenes.txt
(asthma associated gene names downloaded from respirtorygenomics.com)
DiseaseGO.cpp
diseaseGOmapping.txt
(related GO terms mapping)
diseaseGOAic.txt
(bi-clique mining format)
motif.txt
(downloaded from respirtorygenomics.com)
convert.cpp
motifs.txt
(bi-clique mining format)
biclique.txt
(bi-clique mining format)
combine.cpp
FP-MBS
bicliqueOUTPUT.txt
(mined bi-cliques)
filteredBiclique.cpp
OUTPUT.txt
(with both common motifs and GOs)
CalculateScore.cpp
score.txt
(scores measured for each generated gene module)
scorelist.txt
(sorted, gene names retrived)
OutPathways.txt
(coresponding pathway names)
GenePairPathways.cpp
Pathways.txt
(downloaded from integrated pathways database)
GenePathways.cpp
OutGenePathways.txt
(a list of pathways for each gene, in index)
Genes.txt
(corresponding gene names)
Mine.cpp
DisPathwayGenes.txt
(overlapping gene names list)
diseasePathways.txt
(bi-clique mining format)
FP-MBS
reference.txt
(mined bi-cliques)
CalculateTable.cpp
table.txt
(performance scores for random groups)

Identification of co-regulated gene modules

Zhang Qiong
Computational Biology Year 3

OUTLINE

- Background
- Methods and Results
- Conclusions

Background

Background

- Co-regulated genes:

  A group of genes that function together as a module

- -> Motif profile

  -> gene ontology terms

  -> gene regulatory pathways

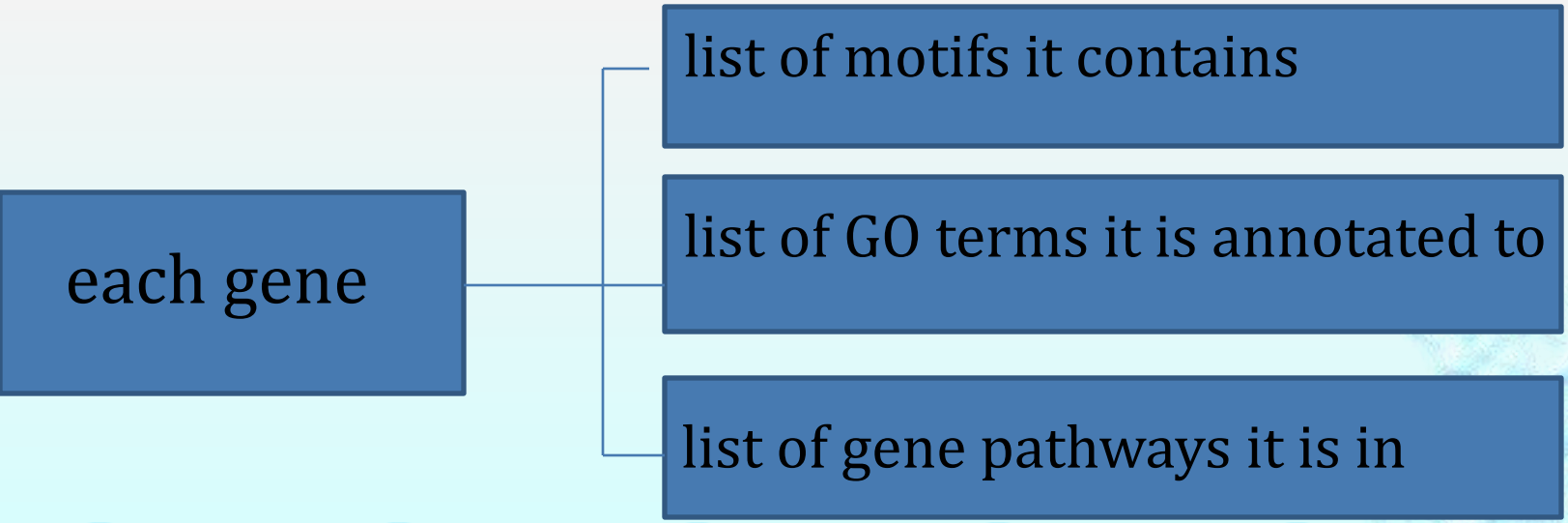
  ->gene expression data,

  -> protein interaction network

Methods and results

1. Data collection and processing

- Putative TFBS Motifs(Respiratory Genomics)
- Gene ontology terms(Gene Ontology)
- Gene regulatory pathways(IntPath)



Gene ontology terms

Original gene ontology annotations

↓ Filtered by GO evidence code

↓ Filtered by GO quantifiers

↓ Full annotations recovered using Matlab

↓ Only annotations with high information content IC=-log(p) is retained. A threshold of 30.

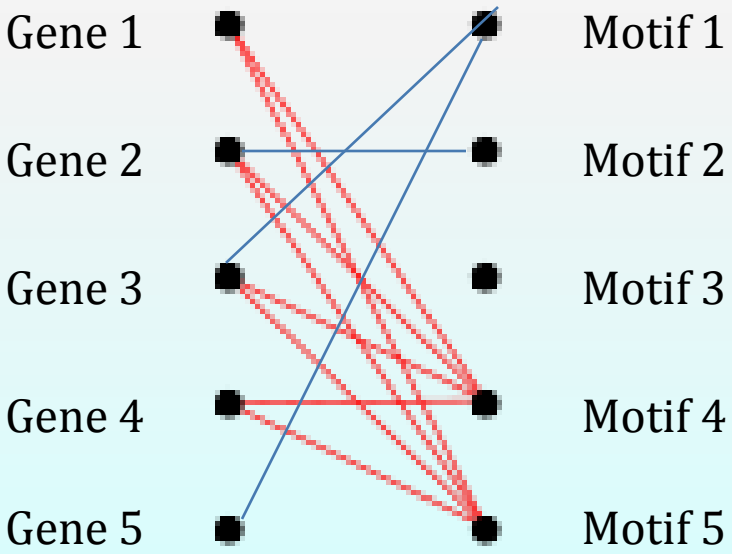
Processed gene ontology annotations

Information content: IC=-log(p)

- P: the frequency, calculated as the number of annotation instances for this GO divided by the total number of annotations
- Setting the threshold to be 30 (adjustable)
- All other GO terms with annotation instances more than 30 are deleted

2. Maximum bi-clique mining

- Use a bi-graph to describe the motif gene network



Number of Mined Maximum Bi-cliques

(61 input genes in disease allergic-rhinitis)

Number of common motifs contained

Number of Genes in a set

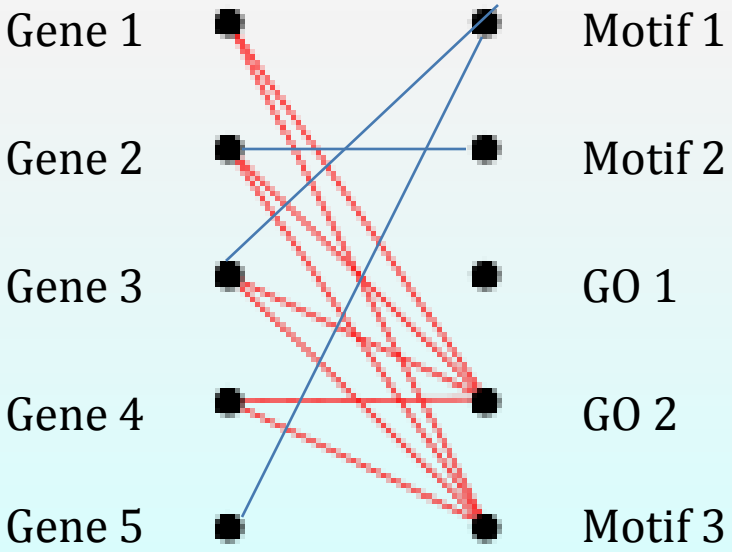
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
2	2491	2398	1711	1017	602	404	287	201	151	108	68	54	34	26	18	13	13	13	13	13	13	10
3	1958	1865	1214	620	319	200	129	91	65	43	22	16	7	5	5	5	5	5	5	5	5	4
4	1317	1225	696	293	144	80	48	31	17	7	4	1	1	1	1	1	1	1	1	1	1	0
5	815	726	341	105	40	17	9	4	1	1	0	0	0	0	0	0	0	0	0	0	0	
6	470	389	123	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	263	190	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	147	80	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	79	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	58	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	33	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	28	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

POTENTIAL CO-REGULATED GENE MODULES BUT IN A LARGE NUMBER

(FP-MBC program, developed by Guimei LIU)

Add gene ontology annotations to the bi-graph

- Only bi-cliques with at least one common motif and one common gene ontology annotation are maintained



A total of 121 gene modules of size 3\4 are obtained from an input of 379 asthma associated genes

3. Rankings of potential co-regulated gene modules using gene pathways information

- Null hypothesis
- p-value= proportion of randomly drawn gene groups of same size n, that have same or higher scores
- score=[size of largest bi-clique, number of common pathways in the largest bi-clique]

index	p value	gene group size	largest bi-clique size	number of common pathways in largest bi-clique	gene1	gene2	gene3	gene4
1	a4.00E-05	4	4	1	ADIPOQ	TGFB1	PPARG	GATA3
2	6.00E-05	3	3	1	TNF	CASP8	FASLG	
3	5.24E-03	4	3	1	TGFBR3	TGFBR1	SMAD3	SMAD2
4	7.69E-03	3	3	1	GATA2	PPARG	GATA3	
5	7.69E-03	3	3	1	ADIPOQ	TGFB1	GATA3	
6	7.69E-03	3	3	1	TLR4	TLR6	TLR1	
7	7.69E-03	3	3	1	IL8	CCL2	TNFSF4	
8	7.69E-03	3	3	1	IL6	TLR9	TLR7	
9	7.69E-03	3	3	1	NOD1	TLR6	TLR1	
10	7.69E-03	3	3	1	IL1B	TLR7	TLR2	
11	7.69E-03	3	3	1	IL12A	IL12B	CCL5	
12	7.69E-03	3	3	1	IL12A	IL12B	MYD88	
13	7.69E-03	3	3	1	IL6	IL12A	CCL5	
14	1.52E-02	3	2	5	MIF	TNF	IFNG	
15	1.52E-02	3	2	5	IL17A	TNF	IFNG	
16	1.73E-02	3	2	4	MYB	TLR4	TNF	
17	1.73E-02	3	2	4	MYD88	TLR4	TNFSF4	
18	1.73E-02	3	2	4	IL4	TLR4	TNF	
19	1.73E-02	3	2	4	TNF	TLR4	ACVR1	
20	2.46E-02	3	2	3	CCL5	EDN1	TNF	
			…					
117	9.11E-01	3	1	1	TNFSF4	IRF4	IL9	
118	9.11E-01	3	1	1	IL33	GHRL	GATA3	
119	9.11E-01	3	1	1	TGM2	SLC11A1	CCL2	
120	9.11E-01	3	1	1	SLC11A1	CCL2	HLA-DPA1	
121	9.11E-01	3	1	1	HLA-DRB4	HLA-DRB5	HLA-DRB3	

Examining the list

- From top down:

-> verification of existing co-regulation

-> propose potential novel co-regulated gene modules for experimental verification

- From bottom up:

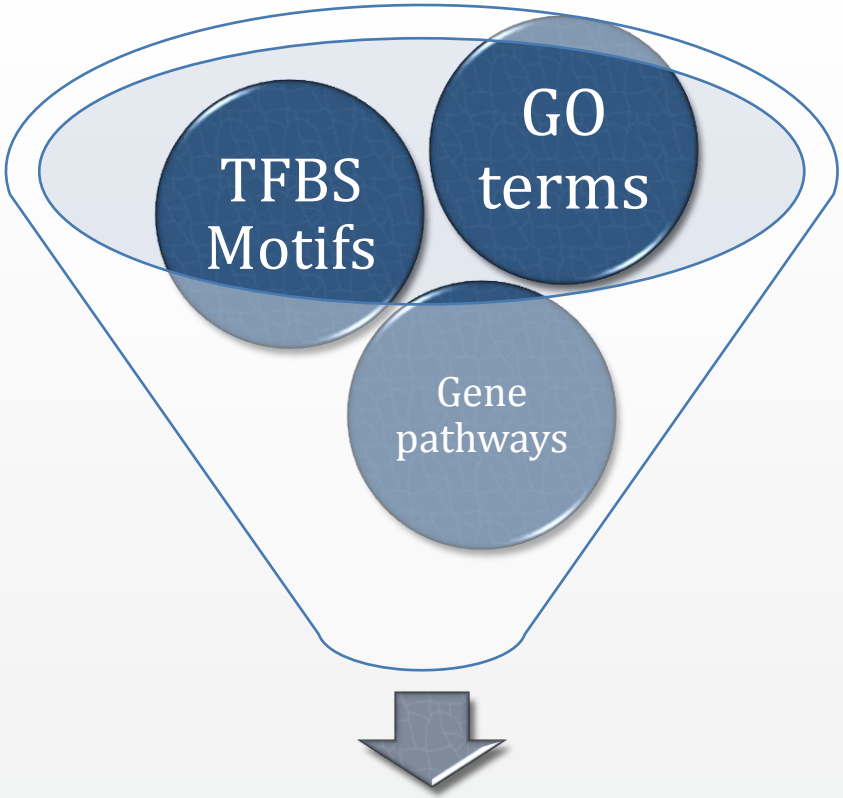
Implication for possible missing pathways

Possible function implication of novel motifs

Gene names	Motifs	Gene ontology terms
ADIPOQ, TGFB1,PPARG, GATA3	127 (unkown)	GO:0045598(regulation of fat cell differentiation)
TNF,CASP8, FASLG	97 (M00080, M00082, M00079)	GO:0008624(induction of apoptosis by extracellular signals)
TGFBR3,TGFBR1, SMAD3, SMAD2	246 (M01107)	GO:0030501(positive regulation of bone mineralization); GO:0007179(transforming growth factor beta receptor signaling pathway)

(Putative transcription binding factor match in TRANFAC/JARSPAR)

Conclusions



Input:
A list of Homo Sapien genes

Identification of co-regulated gene modules

Output:
A list of ranked co-regulated gene modules

References

- Chowdhary, R. (2011). *Transcriptional Regulation Analysis of Respiratory Associated Disease Genes* . Retrieved from http://www.respiratorygenomics.com/ASDScan/index.html
- Dalgin, T. (2005). Interaction between GATA and the C/EBP family of transcription factors is critical in GATA-mediated suppression of adipocyte differentiation. *Molecular cell biology , 25* (2), 706-715.
- *Guide to GO evidence codes*. (2011). Retrieved from The Gene Ontology : http://www.geneontology.org/GO.evidence.shtml
- Hufeng, Z. (2011, June 20th). *Integrated Pathways H.sapiens*. Retrieved 2011, from Integrated Pathways: https://sites.google.com/site/integratedpathways/
- Jinyan Li, H. L. (2005, Octorber). A Correspondence Between Maximal Complete Bipartite Subgraphs and closed patterns. *Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases* , 146--156.
- Lovisa, J. (2009). Interaction between PPARG Pro12Ala and ADIPOQ G276T concerning cholestrol in childhood obesity. *International Journal of Pediatric Obesity , 4* (2), 119-121.
- Robert M. Greene, P. N. (2003). Intracellular dynamics of Smad-mediated TGFβ signaling. *Journal of cell physiology* , 10.
- *The Gene Ontology quanlifiers*. (2011). Retrieved from The Gene Ontology: http://www.geneontology.org/GO.qualifiers.shtml
- Veerla, S. (2006). Analysis of promoter regions of co-expressed genes identified by microarray analysis. *BMC Bioinformatics* , 7:384.

Thank you!

4e-005 4 4 1 ADIPOQ TGFB1 PPARG GATA3
6e-005 3 3 3 TNF CASP8 FASLG
0.00524 4 3 1 TGFBR3 TGFBR1 SMAD3 SMAD2
0.00769 3 3 1 GATA2 PPARG GATA3
0.00769 3 3 1 ADIPOQ TGFB1 GATA3
0.00769 3 3 1 TLR4 TLR6 TLR1
0.00769 3 3 1 IL8 CCL2 TNFSF4
0.00769 3 3 1 IL6 TLR9 TLR7
0.00769 3 3 1 NOD1 TLR6 TLR1
0.00769 3 3 1 IL1B TLR7 TLR2
0.00769 3 3 1 IL12A IL12B CCL5
0.00769 3 3 1 IL12A IL12B MYD88
0.00769 3 3 1 IL6 IL12A CCL5
0.015245 3 2 5 MIF TNF IFNG
0.015245 3 2 5 IL17A TNF IFNG
0.01729 3 2 4 MYB TLR4 TNF
0.01729 3 2 4 MYD88 TLR4 TNFSF4
0.01729 3 2 4 IL4 TLR4 TNF
0.01729 3 2 4 TNF TLR4 ACVR1
0.02455 3 2 3 CCL5 EDN1 TNF
0.053965 3 2 2 IL1B PTGS2 TLR4
0.053965 3 2 2 IL1B IL6 TNFSF4
0.053965 3 2 2 FOXP3 TGFB1 IL1B
0.053965 3 2 2 IL10 IL6 IL1RN
0.053965 3 2 2 IL12B ADIPOQ IL6
0.053965 3 2 2 HMOX1 IL12A TNF
0.053965 3 2 2 IL6 IFNG TGFBR1
0.053965 3 2 2 FGF2 VEGFA OSM
0.053965 3 2 2 IFNG IL6 CCL5
0.053965 3 2 2 IL1B PPARG TGFB1
0.053965 3 2 2 IL1B IL6 ADIPOQ
0.053965 3 2 2 ADIPOQ TLR7 IL1B
0.053965 3 2 2 IFNG CD40LG IL12B
0.053965 3 2 2 IFNG CD40LG TNFSF4
0.053965 3 2 2 IL12A TNF TNFSF4
0.053965 3 2 2 IFNG IL4 IL17A
0.053965 3 2 2 EFNA1 SMAD3 SMAD2
0.053965 3 2 2 IL1B TLR4 TNFSF4
0.0742 4 2 2 PTGER3 PTGS2 IL6 TLR4
0.0742 4 2 2 EDN1 IL12B ADIPOQ IL6
0.15153 4 2 1 F2RL1 IL6 SLC6A4 TLR7
0.15153 4 2 1 ICAM1 TNFRSF4 F2RL1 TNFSF4
0.20099 3 2 1 TNF TLR7 ELANE
0.20099 3 2 1 CD14 IL8 TNFSF4
0.20099 3 2 1 ELANE IL1B TLR2
0.20099 3 2 1 CD14 IL8 CCL2
0.20099 3 2 1 ADIPOQ TLR7 IL6
0.20099 3 2 1 SERPINE1 IL8 TNFSF4
0.20099 3 2 1 F2RL1 TLR7 TLR8
0.20099 3 2 1 F2RL1 IL6 TLR7
0.20099 3 2 1 F3 KDR VEGFA
0.20099 3 2 1 MAPK3 GATA3 EGR1
0.20099 3 2 1 PTGS1 MIF ALOX5

0.20099 3 2 1 CD40LG FOXP3 TNFSF4
0.20099 3 2 1 ADIPOQ GATA3 EGR1
0.20099 3 2 1 CCR2 TNXB CCL2
0.20099 3 2 1 IL12B IL15 IL21
0.20099 3 2 1 IL21 IL15 IL2
0.20099 3 2 1 IL15 IL2 IL18
0.20099 3 2 1 VCAM1 SELE ITGA4
0.20099 3 2 1 ITK IL6 IL4
0.20099 3 2 1 IL1B EGF EDN1
0.20099 3 2 1 TNF TNFSF4 FOXP3
0.20099 3 2 1 IL1RN ADIPOQ IL6
0.20099 3 2 1 TNXB KDR ICAM1
0.20099 3 2 1 SLC11A1 ICAM1 F2RL1
0.20099 3 2 1 TNFRSF4 ICAM1 F2RL1
0.20099 3 2 1 IL7 FGF2 IL6
0.20099 3 2 1 IL2 IL7 TNFSF4
0.20099 3 2 1 IL6 SLC6A4 TLR7
0.20099 3 2 1 IL1B TNFSF4 TNFRSF4
0.20099 3 2 1 TGFBR3 TGFBR1 SMAD2
0.20099 3 2 1 GSTP1 IL8 TNFSF4
0.20099 3 2 1 CCL2 CD14 TNFSF4
0.38222 3 1 4 F2RL1 IL4 TLR4
0.38222 3 1 4 F2RL1 TLR4 TNFSF4
0.38222 3 1 4 F2RL1 MYB TLR4
0.38222 3 1 4 F2RL1 IL10 TNFSF4
0.38222 3 1 4 TLR9 CCL2 ADH5
0.474365 3 1 3 STAT3 RBP4 COL1A1
0.474365 3 1 3 ELANE CCL5 TGFBR1
0.474365 3 1 3 IL12A SLC11A1 TNFSF4
0.474365 3 1 3 EFNA1 TGFBR3 TGFBR1
0.474365 3 1 3 ADIPOQ IL12B TGFBR1
0.474365 3 1 3 IL12A HMOX1 TGFBR1
0.58051 4 1 3 IL15 IL12A IL21 TNFSF4
0.68207 3 1 2 ELANE C5 ADIPOQ
0.68207 3 1 2 F2RL1 IL1B CHIA
0.68207 3 1 2 C5 F2RL1 TLR7
0.68207 3 1 2 F2RL1 CHIA TLR7
0.68207 3 1 2 IL2 TNFSF4 IL21
0.68207 3 1 2 MIF TNFSF4 IL1B
0.68207 3 1 2 IL2 TNFSF4 FOXP3
0.68207 3 1 2 IL2 PPARG TNFSF4
0.68207 3 1 2 TGFB2 TNFSF4 IRAK3
0.68207 3 1 2 TGFB2 IRAK3 FOXP3
0.68207 3 1 2 F2RL1 TNFSF4 IRAK3
0.68207 3 1 2 F2RL1 SLC6A4 TLR7
0.68207 3 1 2 IL1B SERPINE1 PTGS2
0.68207 3 1 2 IL2 IL18 TNFSF4
0.68207 3 1 2 TNFSF4 SLC11A1 CTLA4
0.68207 3 1 2 FGF2 BMPR2 KDR
0.68207 3 1 2 IL18 IL2 GHRL
0.68207 3 1 2 PTGS2 IL1B C5
0.68207 3 1 2 IL12B IL18 TNFSF4
0.68207 3 1 2 IL12B TNFSF4 IRAK3
0.68207 3 1 2 IL4 TNFRSF4 MIF

0.68207 3 1 2 CHRNA7 ADIPOQ FOXP3
0.68207 3 1 2 ADAM8 IL1B TNFSF4
0.68207 3 1 2 PTGS2 TNFSF4 FOXP3
0.90873 4 1 1 HLA-DRB5 HLA-DRB4 HLA-DRB3 HLA-DRB1
0.91051 3 1 1 PTGER3 SERPINE1 TNFSF4
0.91051 3 1 1 CCL11 ITGB3 F3
0.91051 3 1 1 CHIA GCLC ANGPT1
0.91051 3 1 1 TGFB2 PON1 GATA3
0.91051 3 1 1 HLA-DRB1 HLA-DRB5 HLA-DRB3
0.91051 3 1 1 TNFSF4 IRF4 IL9
0.91051 3 1 1 IL33 GHRL GATA3
0.91051 3 1 1 TGM2 SLC11A1 CCL2
0.91051 3 1 1 SLC11A1 CCL2 HLA-DPA1
0.999885 3 0 7 HLA-DRB4 HLA-DRB5 HLA-DRB3