Fall 2004 Genomics Exam #2
Genomic Medicine and Sequencing Tools

There is no time limit on this test, though I don't want you to spend too much time on this. You know I work hard to design challenging tests, but not ones that are excessive. You do not need to read any additional papers other than the ones I send to you. There are three pages for this test, including this cover sheet. You are not allowed discuss the test with anyone until all exams are turned in at 11:30 am on Friday November 5. **EXAMS ARE DUE AT CLASS TIME ON FRIDAY NOVEMBER 5**. You may use a calculator, a ruler, your notes, the book and the internet. You may take it in as many blocks of time as you need to. NOTE: I leave town on November 5 and I want to take the tests with me to grade. Submit your paper and electronic versions before 11:30 am so I can take them with me.

The **answers to the questions must be typed in a Word file and emailed to me as an attachment**. Be sure to backup your test answers just in case. You will need to capture screen images as a part of your answers which you may do without seeking permission since your test answers will not be in the public domain. Print this test but make sure the screen shots are big enough to be seen easily. Remember to explain your thoughts in your own words and use screen shots to support your answers. Screen shots without your words are worth very few points.

DO NOT DOWNLOAD ANY PAPERS FOR THIS EXAM. RELY ONLY ON THE FIGURES PROVIDED, YOUR EXPERIENCE AND YOUR SKILLS.

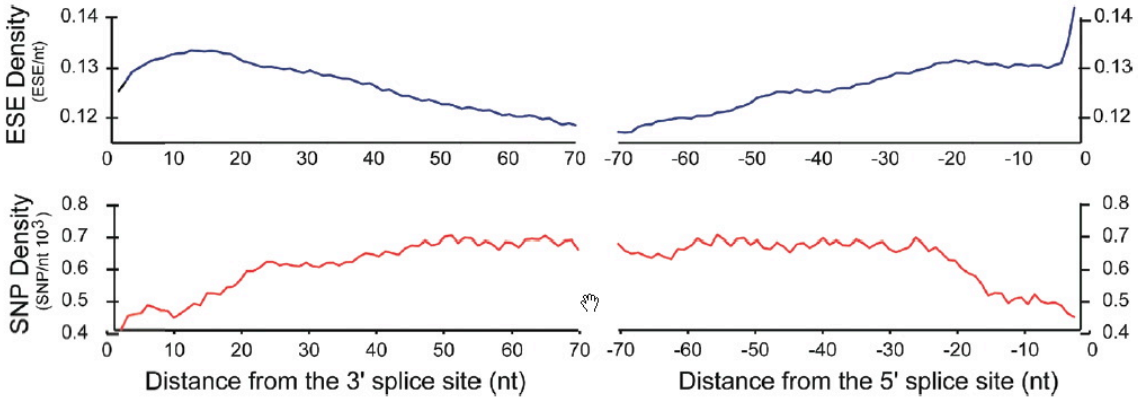
-3 pts if you do not follow this direction.
Please do not write or type your name on any page other than this cover page.
Staple all your pages (INCLUDING THE TEST PAGES) together when finished with the exam.

Name (please print):

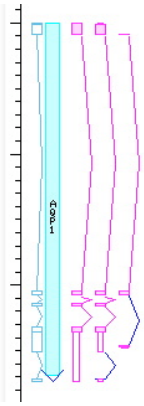
Write out the full pledge and sign:

How long did this exam take you to complete (excluding typing)?

20 pts.
1)



a) Interpret the figure above. Do not look for any online sources of information to answer part a.
Key Points: There are fewer SNPs where ESE tend to be.
This indicates a selection pressure to maintain ESEs without mutations (even 3rd base wobble) due to protein binding.
ESE tend to be closer to the splice sites when many, many genes are examined.

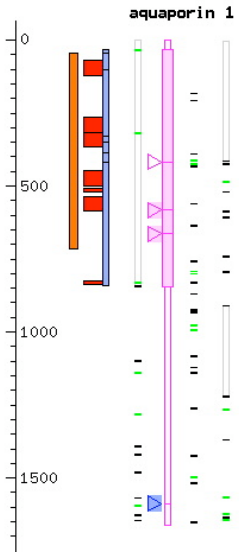


b) Go to this AceView database and query for the human gene aqp1.

http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html?human

c) Use the "gene on genome" and "annotated mRNA" tabs at the top to help you navigate.
d) How many mRNAs are produced from this gene? Support your answer with an image.
There are 4 mRNAs in the Sept. 04 version.

e) Use the bDec03 version to answer the following questions.
f) How many exons are in this mRNA?
5-7 depending on which Sept. version you chose.



g) Choose one exon only, copy and paste its sequence into the form you see on this page.

http://genes.mit.edu/burgelab/rescue-ese/

Show the ESE you found with a screen shot.

mRNA Sequence 1661 bp derived from the genome

cggcagcggtctcaggccaagccccctgccagcATGGCCAGCGAGTTCAA
GAAGAAGCTCTTCTGGAGGGCAGTGGTGGCCGAGTTCCTGGCCACGACCC
TCTTTGTCTTCATCAGCATCGGTTCTGCCCTGGGCTTCAAATACCCGGTG
GGGAACAACCAGACGGCGGTCCAGGACAACGTGAAGGTGTCGCTGGCCTT
CGGGCTGAGCATCGCCACGCTGGCGCAGAGTGTGGGCCACATCAGCGGCG
CCCACCTCAACCCGGCTGTCACACTGGGGCTGCTGCTCAGCTGCCAGATC
AGCATCTTCCGTGCCCTCATGTACATCATCGCCCAGTGCGTGGGGGCCAT
CGTCGCCACCGCCATCCTCTCAGGCATCACCTCCTCCCTGACTGGGAACT
CGCTTGGCCGCAATGACCTGGCTGATGGTGTGAACTCGGGCCAGGGCCTG
GGCATCGAGATCATCGGGACCCTCCAGCTGGTGCTATGCGTGCTGGCTAC
TACCGACCGGAGGCGCCGTGACCTTGGTGGCTCAGCCCCCCTTGCCATCG
GCCTCTCTGTAGCCCTTGGACACCTCCTGGCTATTGACTACACTGGCTGT
GGGATTAACCCTGCTCGGTCCTTTGGCTCCGCGGTGATCACACACAACTT
CAGCAACCACTGGATTTTCTGGGTGGGGCCATTCATCGGGGGAGCCCTGG
CTGTACTCATCTACGACTTCATCCTGGCCCCACGCAGCAGTGACCTCACA
GACCGCGTGAAGGTGTGGACCAGCGGCCAGGTGGAGGAGTATGACCTGGA
TGCCGACGACATCAACTCCAGGGTGGAGATGAAGCCCAAATAGaaggggt
ctggcccgggcatccacgtaggggcaggggcaggggcgggcggagggag
gggaggggtgaaatccatactgtagacactctgacaagctggccaaagtc
acttccccaagatctgccagacctgcatggtcaagcctcttatgggggtg
tttctatctctttctttctctttctgtttcctggcctcagagcttcctgg
ggaccaagatttaccaattcacccactcccttgaagttgtggaggaggtg
aaagaaagggacccacctgctagtcgcccctcagagcatgatgggaggtg
tgccagaaagtccccccctcgccccaaagttgctcaccgactcacctgcgc
aagtgcctgggattctaccgtaattgctttgtgcctttgggcacggccct
ccttcttttcctaacatgcaccttgctcccaatggtgcttggaggggaaa
gagatcccaggaggtgcagtggaggggcaagctttgctccttcagttct
gcttgctcccaagcccctgacccgctcggacttactgcctgaccttggaa
tcgtccctatatcagggcctgagtgacctccttctgcaaagtggcaggga
ccggcagagctctacaggcctgcagcccctaagtgcaaacacagcatggg
tccagaagacgtggtctagaccagggctgctctttccacttgccctgtgt
tctttccccaggggcatgactgtcgccacacgcctctgcatatatgtctc
tttggagttggaatttcattatatgttaagaaaataaaggaaaatgactt
gtaaggtcctt

TGGCTGATGGTGTGAACTCGGGCCAGGGCCTG
GGCATCGAGATCATCGGGACCCTCCAGCTGGTGCTATGCGTGCTGGCTAC
TACCGACCGGAGGCGCCGTGACCTTGGTGGCTCAGCCCCCCTTGCCATCG
GCCTCTCTGTAGCCCTTGGACACCTCCTGGCT

h) Was the ESE you found above consistent with the data presented in the figure for part a? Explain your answer.

For the example above, the exon was consistent with the data. For other exons, the consistency varied.

20 pts.

2) Several studies using microarrays have been performed on DLBCL. The figures below summarize much of this work.

a) Interpret figure 1 below.

6 genes appear to be predictors of DLBCL survivability. The three at the top and the three at the bottom. The z scores are explained in clinical terms, but we do not know which way you want these genes to be expressed in the cancer (induced v. repressed). Simply having the genes is not a correct answer.

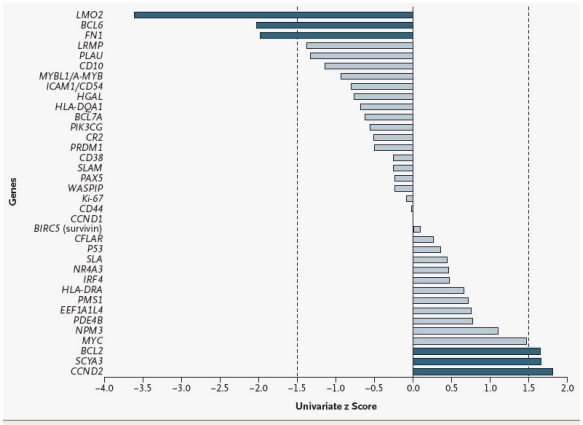
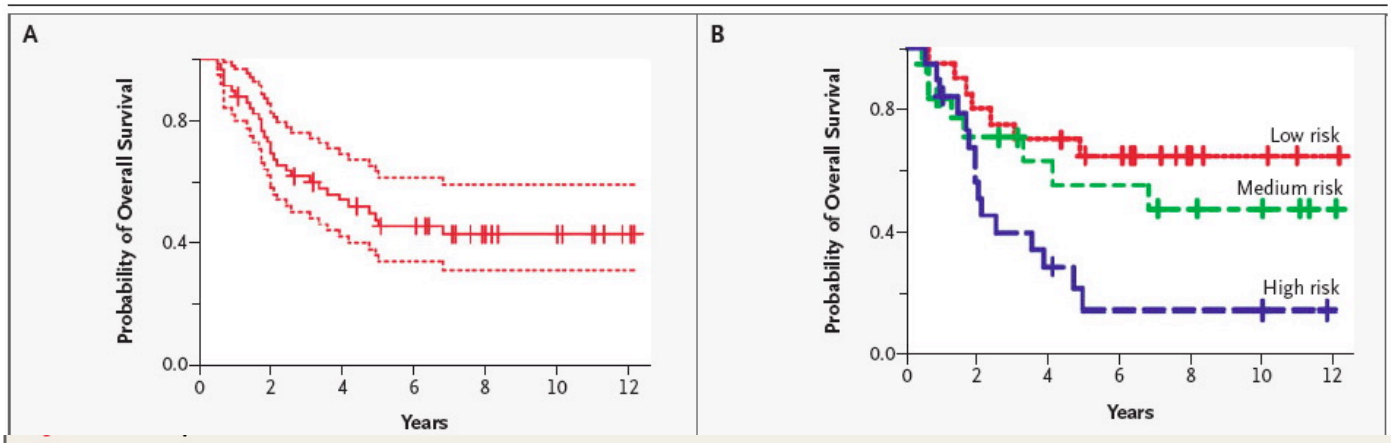


Figure 1. **Univariate Analysis of Expression of 36 Genes with Overall Survival as a Dependent Variable.**
The genes are ranked on the basis of their predictive power (univariate z score), with a negative score associated with longer overall survival and a positive score associated with shorter overall survival. The dashed lines represent an absolute univariate z score of ±1.5. The prediction model is based on the weighted expression of six genes and is expressed

b) Interpret this figure below



Panel A shows Kaplan–Meier estimates of overall survival in the 66 patients with diffuse large-B-cell lymphoma, analyzed by quantitative reverse-transcriptase polymerase chain reaction with TaqMan probe-based assays. The dotted lines represent 95 percent confidence intervals. Panel B shows Kaplan–Meier curves for overall survival in the three groups (at low, medium, and high risk of death) as defined by a prediction

Similar to the case in the book, probability of surviving can be predicted. All 66 patients were considered as one group in panel A. In panel B, they were divided into 3 groups, based on the 6 genes in the figure above. The medium group is new, the low group is similar to the book. The high group is much more accurate than the book's version of same. Compare the numerical probabilities.

This web site explains this type of graph.
http://www.cancerguide.org/scurve_km.html

c) Given the different method for classifying risk of survival with DLBCL compared to the version in your genomics textbook, do you think this meta-analysis method is better, worse, or the same? Explain your answer to get full credit.

See above.

d) Go to the lymphoma search page
http://genome-www.stanford.edu/lymphoma/search.shtml

and see if you can validate any of the genes indicated as important from this meta-analysis.

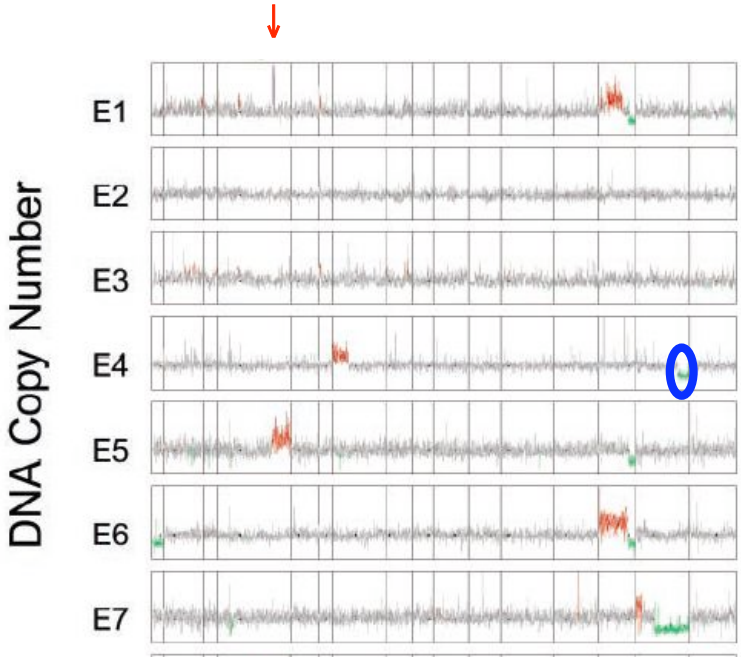
Two hints at no extra charge:
Use BCL-6 instead of BCL6.
Use ttg-2 instead of LMO2
Many of the genes were in this database, but not all. You only needed to focus on the top and bottom 6. Screen shots were required.

e) There is some ambiguity from the figure 1 for part a of this question. However, after doing these searches (part d), you should be able to answer this question now. For each z score, is it better for the patient's survival to have the best indicator genes induced or repressed. Explain how you determined your answer. Refer to figure 1 as necessary.

This was the hardest part. You should have noticed from part d above that BCL-2 was repressed in GC-like DLBCL, while BCL-6 and LMO2 (ttg-2) were induced.
Since we know that GC-like cancers lead to a good clinical outcome, you can deduce that the top three (negative z scores) should be induced and the bottom three should be repressed for a good clinical outcome. Conversely, the opposite expression ratios probably happens in 60% of the patients which leads to unsuccessful treatment.

20 pts.
3) In the figure below, you see the expression ratio for every gene for 8 evolved (E) yeast strains. All the strains were grown in constant glucose-limited conditions for hundreds of generations. The 8 strains began their evolution as genetically identical isolates from the same parental strain.

a) Find the gene HXT6 in these graphs. Do the best you can to narrow the exact location and draw an arrow on this test where you think it is located. Explain the rationale you used to pinpoint HXT6's location. You must tell me where you found the information required for your answer in order to receive full credit. SGD and other sources.
b) Find a region of the genome that experienced a lot of aneuploidy. On this test, circle your region of choice.



c) Using the table to the right, find the exact location for one end of the aneuploidy for one of the strains in the table. Look at chromosome and describe any features that like good candidates for the site of the chromosomal rearrangement. You may want to compare what you found to a few other sites.
The goal was to have you zoom in on a break point near a gene and find areas that look like causes for deletions or insertions. Many of you noted correctly the existence of transposons, Long Terminal Repeats (LTRs), or tRNA genes. These all are repetitive and could lead to recombination events.

d) Find one location for the aneuploidy and search the Tup1 and Yap+++ databases from the DeRisi paper to see if you can verify similar sites of aneuploidy in the either mutant strain used by DeRisi. You do not have to find it exists, but you do have to document how you showed it does or does not exist in both mutant strains. I strongly encourage you to combine screen shots with your explanations.
Many possible answers. You had to show screen shots from the deletion databases of more than one gene to survey for aneuploidy. Then you had to explain quantitatively how you determined whether the ratio indicated aneuploidy or not. This means you should have compared the ratios to the known duplication or deletions documented in these cases.

20 pts.
4) First, I want to tell you this question is the most experimental one I have written in the test. If you try to just grunt it out and systematically do every permutation, you will waste WAY too much time. I strongly suggest you do a quick pass with the list on both sites, navigate the sites some to see what you can find, and grab a few screen shots to remind you of your options. Then, answer the questions below.

Go to these two web sites
http://transcriptome.ens.fr/ymgv/access_2.php
http://db.yeastgenome.org/cgi-bin/expression/expressionConnection.pl
to analyze the genes in this list.

ELM1
SRC1
FKH1
YOL128C
GSP2
WSC4
GPA1
YBL009W
YML125C
YGL101W
STU2
YOR073W
CIN8
YEL017W
YLR455W
STR3
Cdc28

CRC1
YRO2
MET17
STB1

Below are a series of screen shots that help answer these questions. There were several clusters of related genes. You only had to find one cluster to get full credit.

a) Do these genes have anything in common? In other words, can you use their expression profiles to find any patterns? You do not have to consider all of these genes as a group. Feel free to pull out subsets that appear to be associated. To answer this question for full credit you must:

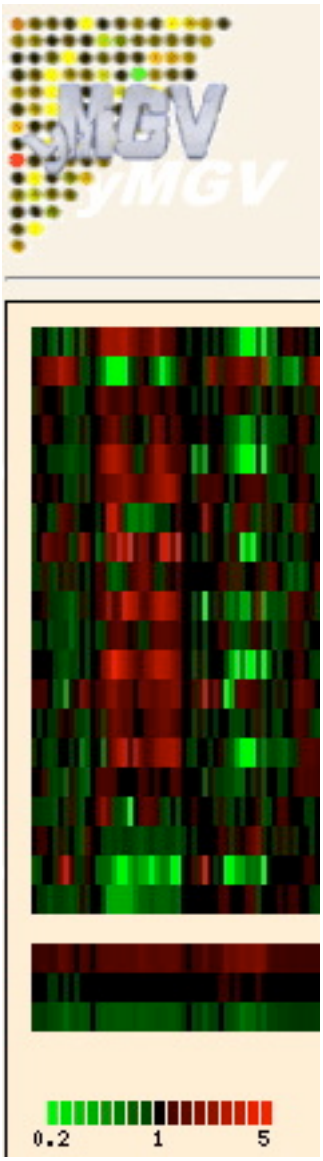
b) Show me the data you used to support your answer.

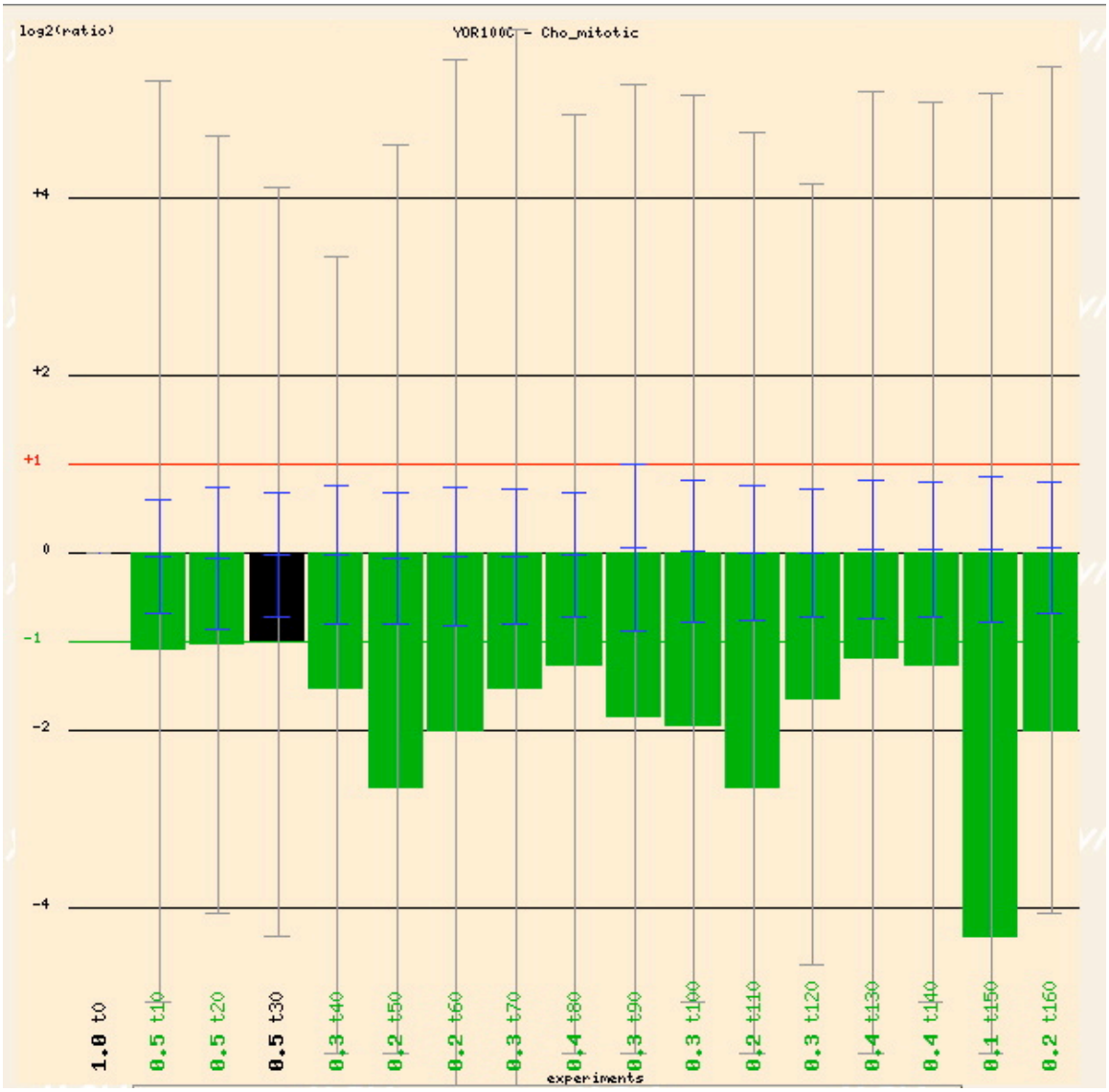
c) Explain the logic you used to produce your answer.

d) Compare the data produced by the two databases and decide which was more helpful.

e) Explain why one database was more helpful to you than the other.

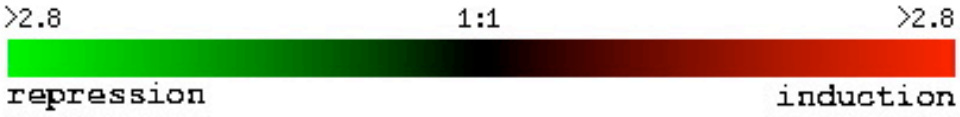
f) Validate your conclusion with data from another site and show me the data.




log2(ratio)
YOR100C - Cho_mitotic
+4
+2
+1
0
-1
-2
-4
1.0 t0
0.5 t10
0.5 t20
0.5 t30
0.3 t40
0.2 t50
0.2 t60
0.3 t70
0.4 t80
0.3 t90
0.3 t100
0.2 t110
0.3 t120
0.4 t130
0.4 t140
0.1 t150
0.2 t160
experiments

Similarly Expressed Genes: Expression duri

Scale



Click on a color strip to see 20 genes expressed similarly to that gene

Up to 20 similar genes are shown, with a Pearson correlation of > 0.

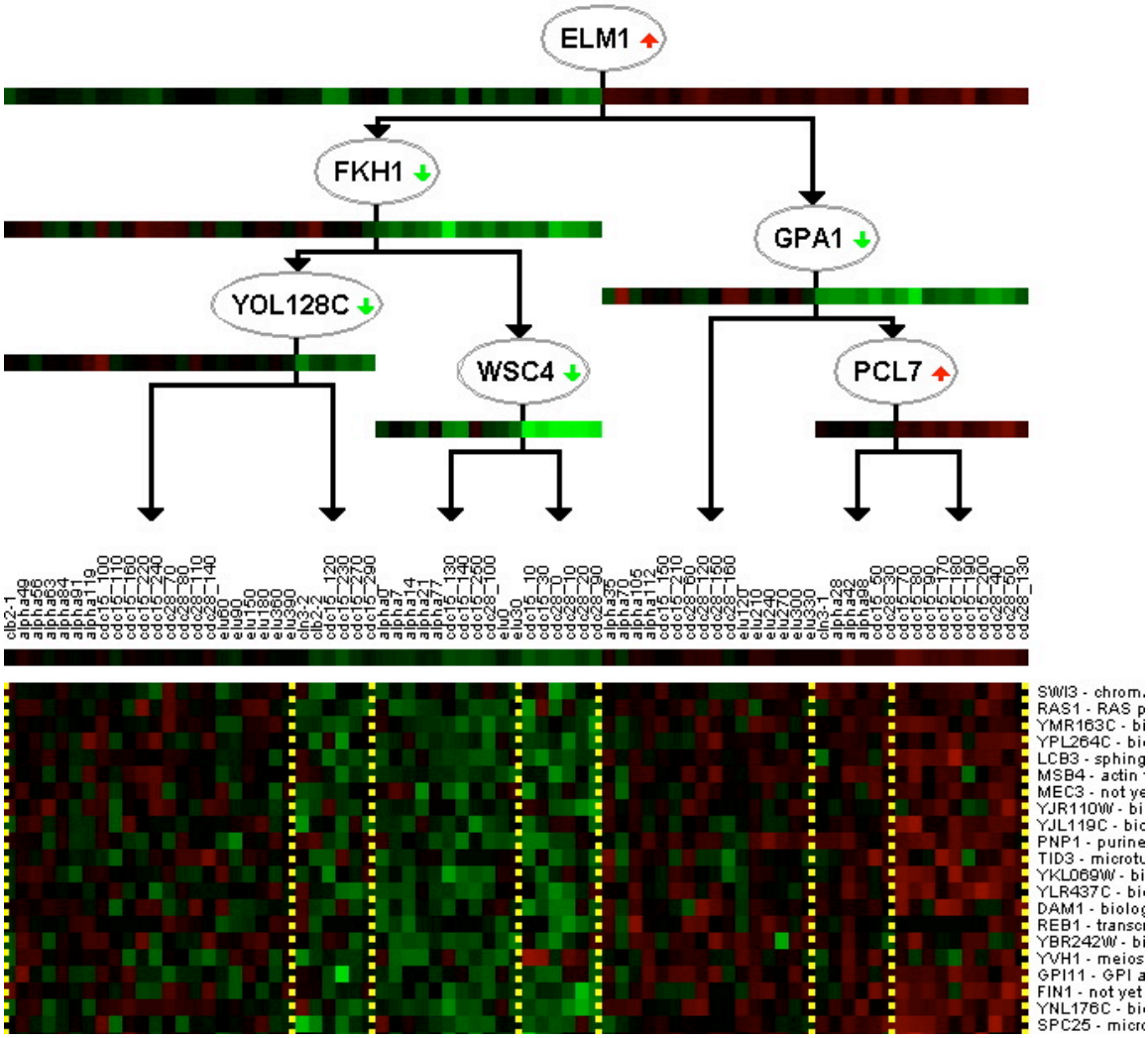
ORF	Gene	0 hrs 0.5 hrs 2 hrs 5 hrs 7 hrs 9 hrs 11.5 hrs	Go Process term
YML034W	SRC1		mitotic sister chromatid segregation
YJR099W	YUH1		protein deubiquitination
YLR220W	CCC1		iron ion homeostasis*
YOL068C	HST1		chromatin silencing*
YGR266W	-		biological_process unknown
YKR002W	PAP1		mRNA polyadenylylation*
YMR184W	-		biological_process unknown
YLL005C	SPO75		sporulation (sensu Fungi)
YOR073W	SGO1		mitotic sister chromatid segregation*
YHR152W	SPO12		regulation of exit from mitosis*
YDL239C	ADY3		protein complex assembly*
YLL004W	ORC3		DNA replication initiation*
YPL027W	SMA1		spore wall assembly (sensu Fungi)
YNL225C	CNM67		microtubule nucleation
YKL121W	-		biological_process unknown
YEL061C	CIN8		mitotic sister chromatid segregation*
YJL068C	-		formaldehyde catabolism
YDR263C	DIN7		DNA repair*
YDR117C	-		biological_process unknown
YNL187W	-		transport*
YOL069W	NUF2		chromosome segregation*

Module 17 (39 Genes) - Cell Cycle Analysis

Jump to: Significant GO annotations | Significant Motif Binding Sites.

Predicted Regulatory Module

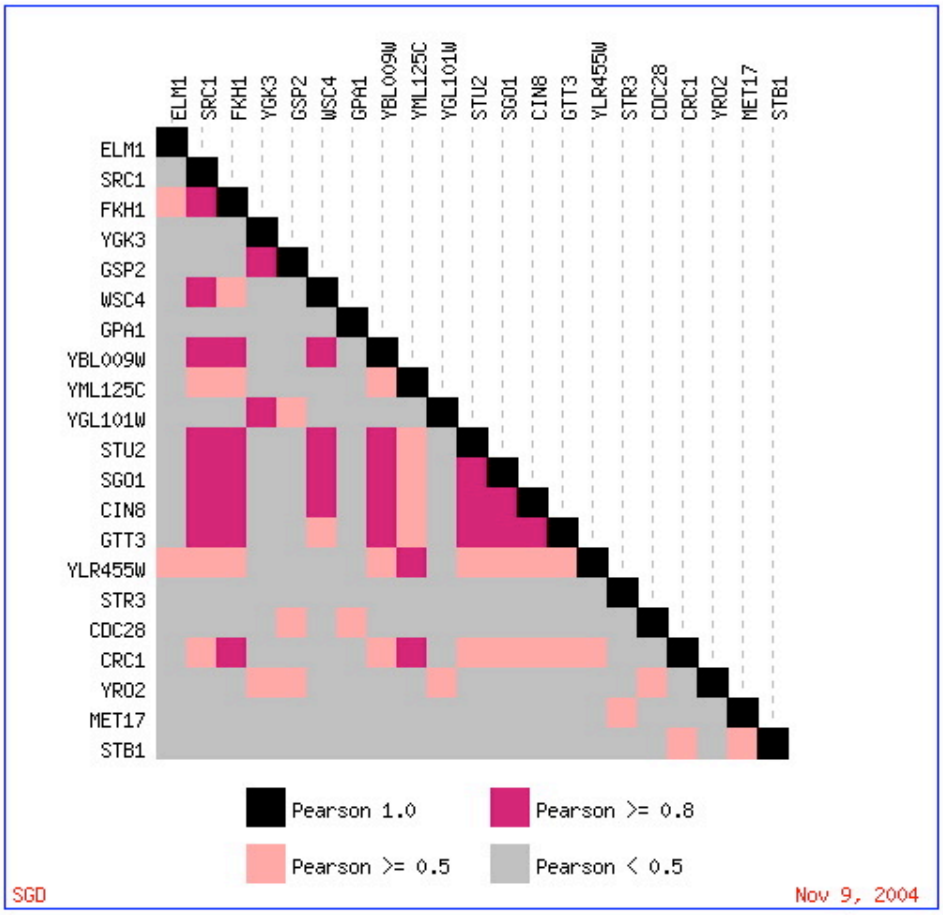
The control program is a set of gene products predicted to regulate the cluster of genes below. The prediction is base genes in the cluster as described in Segal et al. In several instances, there is direct experimental evidence to corrobo products) are in a hierarchy according to their effects on gene expression in the experiments directly below them.



Expression during sporulation

ısure of how similar the expression of two gene products is. Pearson = 1.0 represents a perfect cor

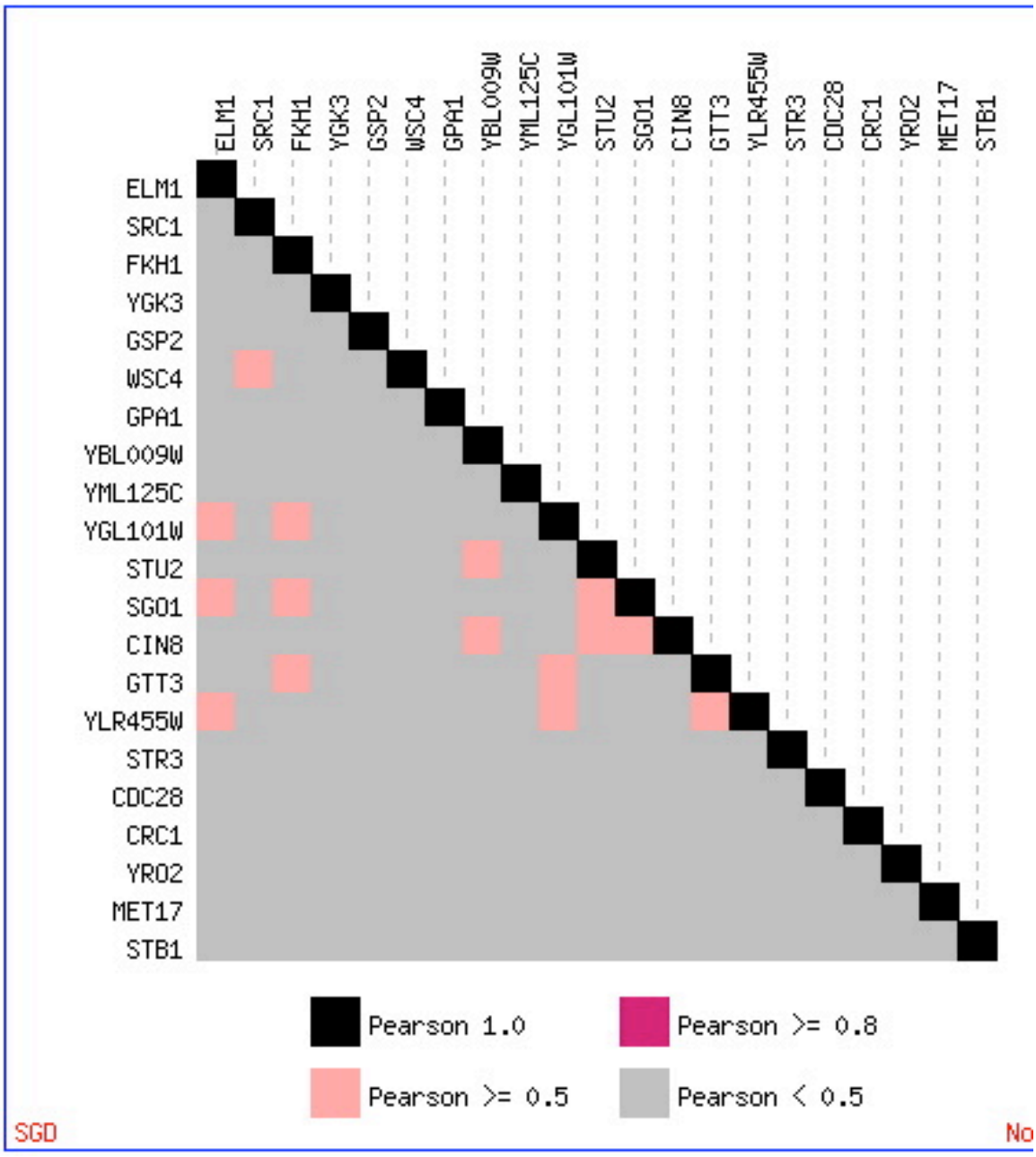
to find common processes, functions, or localizations for the two genes.

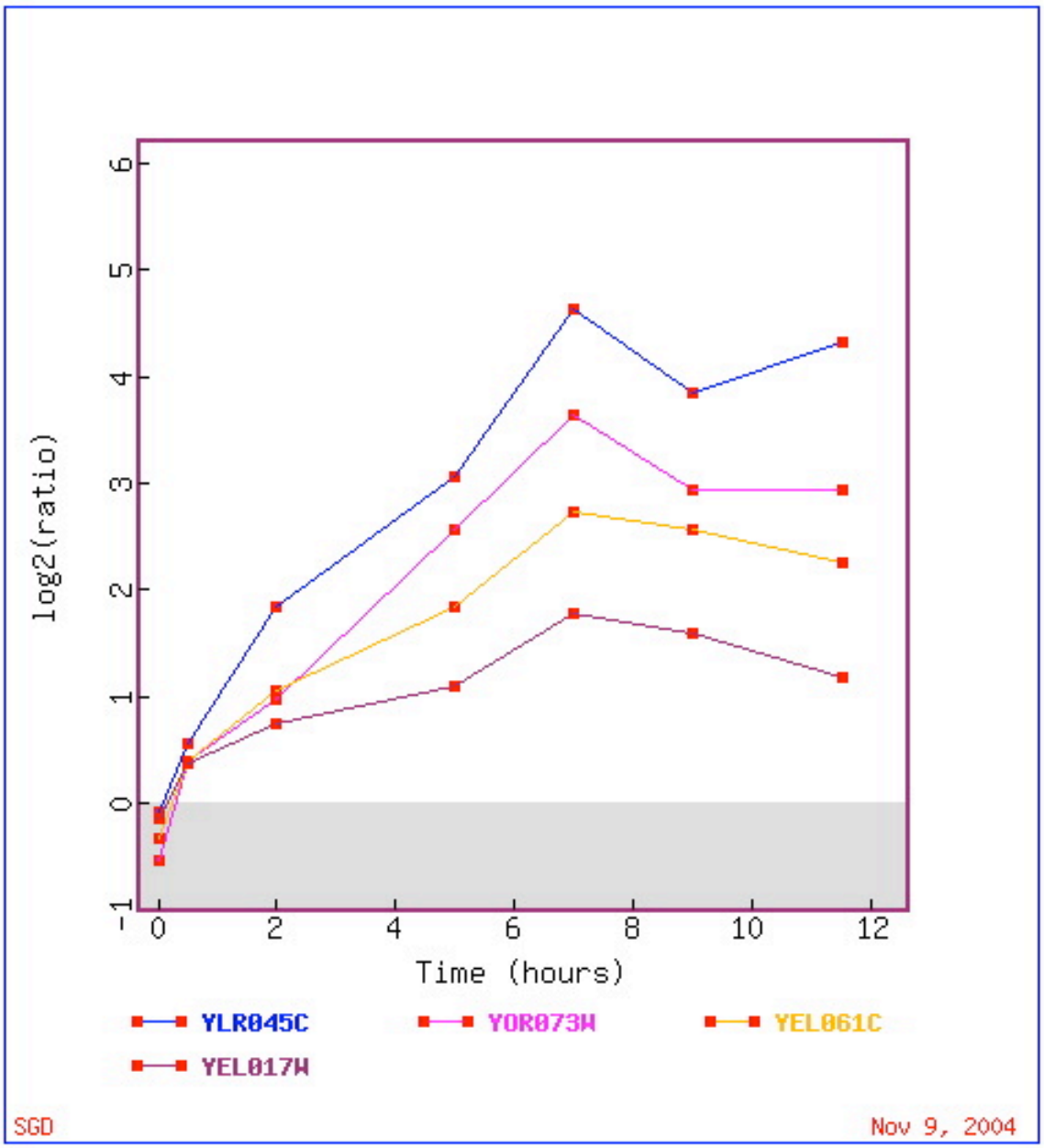


on during the cell cycle

imilar the expression of two gene products is. Pearson = 1.0 represent

ɔn processes, functions, or localizations for the two genes.





Search results in table form for top 25 hits

Terms from the Process Ontology

Gene Ontology term	Cluster frequency	Genome frequency of use	P-value	Genes annotated to the term
cell cycle | AmiGO	9 out of 21 genes, 42.8%	516 out of 7286 annotated genes, 7.0%	5.99e-06	ELM1, SRC1, FKH1, YBL009W, STU2, SGO1, CIN8, CDC28, STB1
cell proliferation | AmiGO	9 out of 21 genes, 42.8%	594 out of 7286 annotated genes, 8.1%	1.88e-05	ELM1, SRC1, FKH1, YBL009W, STU2, SGO1, CIN8, CDC28, STB1
mitotic cell cycle | AmiGO	6 out of 21 genes, 28.5%	210 out of 7286 annotated genes, 2.8%	2.14e-05	SRC1, STU2, SGO1, CIN8, CDC28, STB1
nuclear division | AmiGO	6 out of 21 genes, 28.5%	243 out of 7286 annotated genes, 3.3%	4.84e-05	SRC1, YBL009W, STU2, SGO1, CIN8, CDC28
M phase | AmiGO	6 out of 21 genes, 28.5%	259 out of 7286 annotated genes, 3.5%	6.89e-05	SRC1, YBL009W, STU2, SGO1, CIN8, CDC28
sister chromatid segregation | AmiGO	3 out of 21 genes, 14.2%	33 out of 7286 annotated genes, 0.4%	0.00011	SRC1, SGO1, CIN8
mitotic sister chromatid segregation | AmiGO	3 out of 21 genes, 14.2%	33 out of 7286 annotated genes, 0.4%	0.00011	SRC1, SGO1, CIN8
mitotic anaphase | AmiGO	3 out of 21 genes, 14.2%	41 out of 7286 annotated genes, 0.5%	0.00021	SRC1, SGO1, CIN8

There were several correct answers. Above I have collected screen shots of ways to use the two databases, but mostly Expression Connection which was easier for me to search in a meaningful way (quantifiable correlations). GO allowed me to validate my clusters of genes. Some genes were in known pathways that Expression Connection highlighted.

20 pts.
5) This question has you work with an excel sheet to approximate what real microarray data analysis is like. You must download this excel file before you can answer any of these questions.

URL: http://www.bio.davidson.edu/courses/genomics/exams/2004/Exam2_04.xls

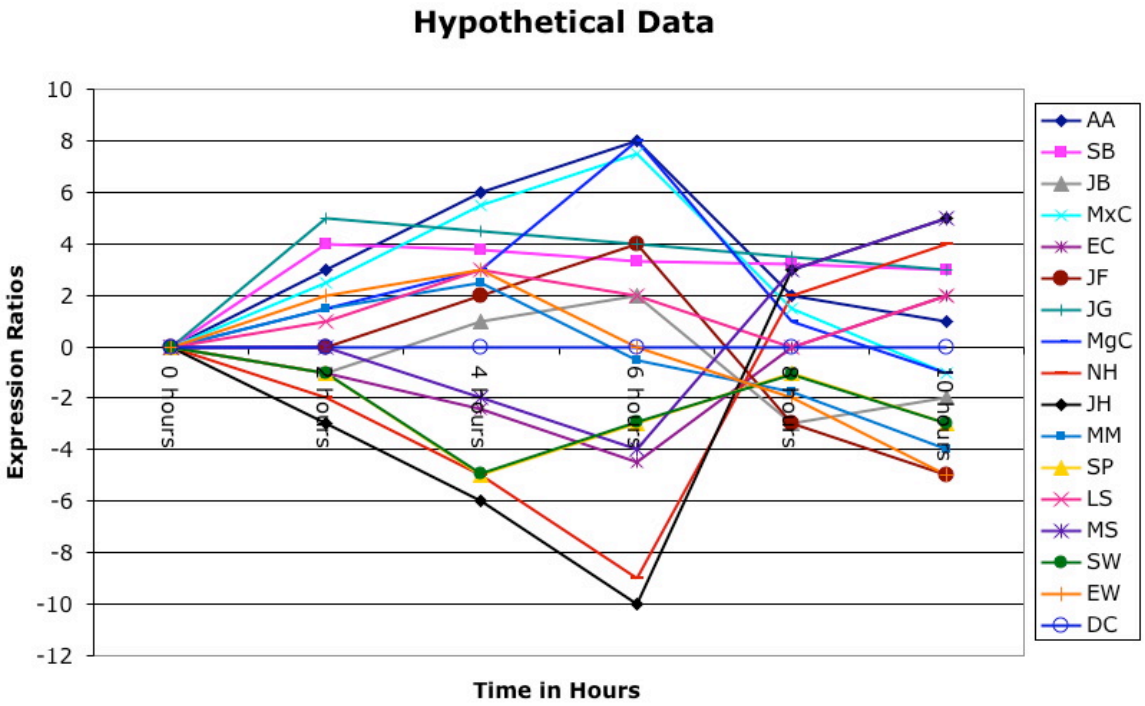
Do all your work on this excel file and submit it separately by email along with your final exam Word file. Print the two worksheets out, but make sure you limit the printouts to single pages (i.e. don't waste paper or time printing unnecessary pages.

I had created data that could be calculated easily once you did the division. Log2 transformation was multiples of 2 and could be done on paper on in your head. Some of you used the excel sheet to do these calculations – that's great.

a) On the 10 hour sheet, determine the color for each of the genes. To answer, simply copy and paste the color from the table into an appropriately labeled column. You may create as many columns as you see fit. Be sure to label each column so I know what you are thinking.

	Gene Name	Cy3 Dye	Cy5 Dye	Ratio	log base 2 trans.	color
2	AA	2228	4457	2.0000	1	1
3	SB	2668	21346	8.0000	3	3
4	JB	3612	903	0.2500	-2	-2
5	MxC	10246	5123	0.5000	-1	-1
6	EC	4641	18564	4.0000	2	2
7	JF	8896	278	0.0312	-5	-5
8	JG	2764	22109	8.0000	3	3
9	MgC	4009	2004	0.5000	-1	-1
10	NH	1992	31876	16.0000	4	4
11	JH	1434	45876	32.0000	5	5
12	MM	7008	438	0.0625	-4	-4
13	SP	7848	981	0.1250	-3	-3
14	LS	1421	5684	4.0000	2	2
15	MS	2056	65798	32.0000	5	5
16	SW	3599	450	0.1250	-3	-3
17	EW	23552	736	0.0312	-5	-5
18	DC	2376	2375	1.0000	0	0

b) Copy your 10 hour data and paste it into the appropriate column on the All Times sheet.
c) Finish the graph by including the data you added to the 10 hour column.



By far, the most complete answer was from Allison who calculated all the correlation coefficients using excel. It was so complete that I used her table (below) to help me grade all the possible pairs suggested by students. Thanks Allison!

Name	0 hr	2 hrs	4 hrs	6 hrs	8 hrs	10 hrs	ave	sd
AA	0	3	6	8	2	1	3.333	3.077
SB	0	4	3.75	3.3	3.2	3	2.875	1.456
JB	0	-1	1	2	-3	-2	-0.5	1.871
MxC	0	2.5	5.5	7.5	1.5	-1	2.667	3.266
EC	0	-1	-2.4	-4.5	0	2	-0.98	2.245
JF	0	0	2	4	-3	-5	-0.33	3.266
JG	0	5	4.5	4	3.5	3	3.333	1.78
MgC	0	1.5	3	8	1	-1	2.083	3.2
NH	0	-2	-5	-9	2	4	-1.67	4.761
JH	0	-3	-6	-10	3	5	-1.83	5.636
MM	0	1.5	2.5	-0.5	-1.75	-4	-0.38	2.322
SP	0	-1	-5	-3	-1	-3	-2.17	1.835
LS	0	1	3	2	0	2	1.333	1.211
MS	0	0	-2	-4	3	5	0.333	3.266
SW	0	-1	-4.9	-2.9	-1.1	-3	-2.15	1.783
EW	0	2	3	0	-2	-5	-0.33	2.875
DC	0	0	0	0	0	0	0	0

Correlation Data: I used the function =correl to deterimine the correlation between each pair of genes

	AA	SB	JB	MxC	EC	JF	JG	MgC	NH	JH	MM	SP	LS	MS	SW	EW	DC
AA	1	0.57	0.73	0.98	-0.92	0.79	0.63	0.93	-0.91	-0.89	0.43	-0.66	0.66	-0.79	-0.65	0.49	
SB	0.57	1	-0.05	0.47	-0.30	0.08	**0.99**	0.34	-0.27	-0.27	0.15	-0.56	0.55	-0.08	-0.57	0.17	
JB	0.73	-0.05	1	0.77	-0.83	0.92	**0.06**	0.76	-0.90	-0.90	0.56	-0.44	0.53	-0.92	-0.41	0.59	
MxC	0.98	0.47	0.77	1	**-0.97**	0.88	0.56	0.95	-0.96	-0.94	0.56	-0.55	0.54	-0.88	-0.53	0.61	
EC	-0.92	-0.30	-0.83	-0.97	1	-0.95	-0.41	-0.96	**0.99**	0.97	-0.59	0.37	-0.39	0.95	0.36	-0.66	
JF	0.79	0.08	0.92	0.88	-0.95	1	0.21	0.85	**-0.97**	-0.97	0.73	-0.28	0.34	**-1**	-0.25	0.77	
JG	0.63	**0.99**	**0.06**	0.56	-0.41	0.21	1	0.42	-0.38	-0.40	0.30	-0.53	0.54	-0.21	-0.54	0.32	
MgC	0.93	0.34	0.76	0.95	-0.96	0.85	0.42	1	-0.93	-0.90	0.34	-0.39	0.40	-0.85	-0.37	0.42	
NH	-0.91	-0.27	-0.90	-0.96	**0.99**	**-0.97**	-0.38	-0.93	1	**1**	-0.62	0.42	-0.47	0.97	0.40	-0.68	
JH	-0.89	-0.27	-0.90	-0.94	0.97	**-0.97**	-0.40	-0.90	**1.00**	1	-0.67	0.41	-0.48	0.97	0.39	-0.72	
MM	0.43	0.15	0.56	0.56	-0.59	0.73	0.30	0.34	-0.62	-0.67	1	-0.17	0.21	-0.73	-0.16	**1**	
SP	-0.66	-0.56	-0.44	-0.55	0.37	-0.28	-0.53	-0.39	0.42	0.41	-0.17	1	-0.96	0.28	**1**	-0.16	
LS	0.66	0.55	0.53	0.54	-0.39	0.34	0.54	0.40	-0.47	-0.48	0.21	-0.96	1	-0.34	-0.95	0.21	
MS	-0.79	-0.08	-0.92	-0.88	0.95	**-1**	-0.21	-0.85	0.97	0.97	-0.73	0.28	-0.34	1	0.25	-0.77	
SW	-0.65	-0.57	-0.41	-0.53	0.36	-0.25	-0.54	-0.37	0.40	0.39	-0.16	**1**	-0.95	0.25	1	-0.15	
EW	0.49	0.17	0.59	0.61	-0.66	0.77	0.32	0.42	-0.68	-0.72	**1**	-0.16	0.21	-0.77	-0.15	1	
DC																	1

d) List two pairs of genes that have a correlation near 1.0. Explain your answer.
e) List one pair of genes that have a correlation exactly 1.0. Explain your answer.
f) List one pair of genes that have a correlation very near –1.0. Explain your answer.
g) List one pair of genes that have a correlation very near zero. Explain your answer.
h) What would have helped you answer these questions more efficiently? Explain your answer.

You should have wanted a way to generate a table like the one above. Doing it automatically would have been wonderful. If you take the lab course next semester, you will get to use MAGIC Tool (www.bio.davidson.edu/MAGIC which does this very quickly. Then you can visually have the software find the most similar genes, etc.