Fall 2005 Genomics Exam #2 – Answer Key
Genomic Variation and Microarrays

There is no time limit on this test, though I don't want you to spend too much time on this. 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 4. **EXAMS ARE DUE AT CLASS TIME ON FRIDAY NOVEMBER 4**. You may use a calculator, a computer, but only the web pages that appear in this exam. You are NOT allowed to explore the internet to take this exam. This is a new policy and is required if I am to shorten the length of the exams. You may take it in as many blocks of time as you need to. NOTE: I leave town on November 4 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 along with paper versions.

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 read 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.

THIS IS A CLOSED BOOK EXAM TO HELP SHORTEN THE TEST.

-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:

On my honor I have neither given nor received unauthorized information regarding this work, I have followed and will continue to observe all regulations regarding it, and I am unaware of any violation of the Honor Code by others.

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

20 pts.

1) Genome Variations question:

"The demonstration of association between common genetic variants and chronic human diseases such as obesity could have profound implications for the prediction, prevention, and treatment of these conditions. Unequivocal proof of such an association, however, requires independent replication of initial positive findings. Recently, three (rs2236418, rs928197, and rs992990) single nucleotide polymorphisms (SNPs) within glutamate decarboxylase 2 (GAD2) were found to be associated with class III obesity (body mass index >40 kg/m^2). The association was observed among188 families (612 individuals) segregating the condition, and a case-control study of 575 cases and 646 lean controls….We found no evidence for a relationship between the three GAD2 SNPs and obesity."

You may use this web site only: http://www.ncbi.nlm.nih.gov/SNP/ and the pages your searching of this site directly leads you to.

a) Give me the DNA sequences for these 3 mutations. Provide the sequences in a readable screen shot. Copy and pasting the sequence is not acceptable.

rs2236418

>gnl|dbSNP|rs2236418|allelePos=256|totalLen=511|taxid=9606|snpclass=1|alleles='A/G'|mol=Genomic|build=123

CGCCCGCACT TCCCGCCTCT GGCTCGCCCG AGGACGCGCT GGCACGCCTC CCACCCCCTC
ACTCTGACTC CAGCTGGCGT GCATGGTCTG CCTCGCATCC TCACGACTCA GCTCCCTCCC
TCTCTCGTGT TTTTTTCCTC CGCCGCCCCC TCATTCATCC CCACTGGGCT CCCTTTCCCT
CAAATGCTCT GGGGCTCTCC GCGCTTTCCT GAGTCCGGGC TCCGAGGACC CTTAGGTAGT
CCCGGTCTCT TTTAA
R
GCTCCCCGGC TTCCAAAGGG TTGCCACGTC CCTAAACCCT GTCTCCAGCT CGCATACACA
CACGCACAGA CACGCACGTT TTCTGTTCCT GCGTGACACC CGCCCTCGCC GCTCGGCCCC
GCCGGTCCCC GCGCGGTGCC CTCCTCCCGC CACACGGGCA CGCACGCGCG CGCAGGGCCA
AGCCCGAGGC AGCTCGCCCG CAGCTCGCAC TCGCAGGCGA CCTGCTCCAG TCTCCAAAGC
CGATGGCATC TCCGG

rs928197

>gnl|dbSNP|rs928197|allelePos=256|totalLen=511|taxid=9606|snpclass=1|alleles='A/T'|mol=Genomic|build=123

ACTGTGGGGG GAAAATGCCC CCAAACGTCT TGCTAACCCA TTTAGCTTGG GGCCAATACT
AGATTCATCC CATCCTCCCA AAACACTAAC TGGAAAGTCA AGGACAAGGT GGCAGGCAGC
TGATAGTCTA TCACTTATTA TTCCTCTTAT CACTTGCAGG ATCTTGAATG TGTTAGACTG
TTCTAATTCT CTGATCCCAA GAAAACTTGG AGGTAACTCT TTCAAGAGAA AACAATAAGG
TTCTGACTGT TGAGC
W
AAAAACTAAA GACGCTGCTT GCTGTTGGGT TCTTTGACTC AGGGGAGAGT CCCAGGAGAA
AGTCACCATG CTGATATGGT CTGTCCCACA GGTGGCTCCA GTGATTAAAG CCAGAATGAT
GGAGTATGGA ACCACAATGG TCAGCTACCA ACCCTTGGGA GACAAGGTCA ATTTCTTCCG
CATGGTCATC TCAAACCCAG CGGCAACTCA CCAAGACATT GACTTCCTGA TTGAAGAAAT
AGAACGCCTT GGACA

rs992990

>gnl|dbSNP|rs992990|allelePos=201|totalLen=613|taxid=9606|snpclass=1|alleles='A/C'|mol=Genomic|build=123

TGCAGGGCTT TTTGCCATCT TTATGCCTCT GAGAGGGAGG TGGGACAGAG AATTCAGTGA
CAGGTAGTTG GGGGCCTTGG CAGCTTCTCT TCTAAAAAGA CAAATAGGCC CCCACGTAGA
GATAAACACC ACAGCCAGAC ATGGAAGACA GCTGTTTTCC TCTCCCATCA GGCATTCTTA
CTGACAAAGC TGAGTTTATC
M
GAATTAGACA TCTAGCCATA GAACATGATG GAATGTATAG AATGGCCATG TGTACGTGCA
TGCACGGTGT CACCAAGCTC ACAAATGACA GAGATGAAAT CCATAGCaaa agccaacctt
atttaatgct tactgcatgt taagcacagt tctaagcact tcacCtaaat acatattatt
attctccttt tctagttgag gaaatctagg cacagagagg ttaagtaact tgctcaaagt
cacacagcca ggaagtgatg aaacctgaac gcaaacccat actatctggc tcctgagggc
ttcccccttа actattatgT CTGCTTTATA GCCCAGGGTC CTGACTCCAG GGTCATTCTA
CAAAGCAAGG AGGTATTTTT TTGGAGGGAA CTTTCCAATA CCTCAATGCT GT

b) What is the frequency for each SNP? Use a screen shot to show me your data.

rs2236418

Average estimated heterozygosity: 0.499
Average Allele Frequency:
G 0.474
A 0.526

rs928197

Average estimated heterozygosity: 0.415
Average Allele Frequency:
T 0.706
A 0.294

rs992990

Average estimated heterozygosity: 0.471
Average Allele Frequency:
A 0.380
C 0.620

c) Describe any differences of frequency between populations for each of these SNPs? Support your answer with data from this web site.

The main point is to show that an average frequency is a meaningless number once you look at different populations. You can see in this screen shot, different populations have very different frequencies. Therefore human-wide variations mask the distinctions of populations.

ss3190812	Submitter's Id	IMS-JST016723			Orientation to rs	fwd	
	Handle-Population Id	2n	Allele Freq	Genotype Freq	Hardy-Weinberg		
	YUSUKE-JBIC-allele	1496	G 0.407 A 0.593	N/A	N/A		
ss4020966	Submitter's Id	GAD2-4			Orientation to rs	fwd	
	Handle-Population Id	2n	Allele Freq	Genotype Freq	Hardy-Weinberg		
	JDRF_WT_DIL-UK	400	G 0.138 A 0.863	N/A	N/A		
ss12584303	Submitter's Id	GAD2-001882			Orientation to rs	fwd	
	Handle-Population Id	2n	Allele Freq	Genotype Freq	Hardy-Weinberg		
	EGP_SNPS-PDR90	162	G 0.364 A 0.636	G/G 0.21 A/G 0.309 A/A 0.481	Chi Square 9.012		
	CSHL-HAPMAP-HapMap-CEU	120	G 0.183 A 0.817	G/G 0.05 A/G 0.267 A/A 0.683	Chi Square 0.721		
	CSHL-HAPMAP-HapMap-YRI	120	A 0.067 G 0.933	A/G 0.133 G/G 0.867	Chi Square 0.335		
ss24200482	Submitter's Id	afd2225624			Orientation to rs	fwd	
	Handle-Population Id	2n	Allele Freq	Genotype Freq	Hardy-Weinberg		
	PERLEGEN-AFD_EUR_PANEL	48	G 0.167 A 0.833	A/G 0.333 A/A 0.667	Chi Square 0.96		
	PERLEGEN-AFD_AFR_PANEL	46	A 0.109 G 0.891	A/G 0.217 G/G 0.783	Chi Square 0.342		
	PERLEGEN-AFD_CHN_PANEL	48	G 0.333 A 0.667	G/G 0.125 A/G 0.417 A/A 0.458	Chi Square 0.094		

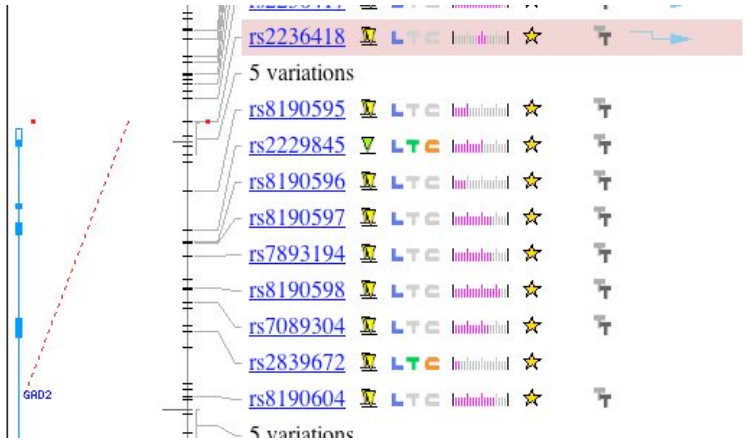
d) What evidence is there to validate these 3 SNPs? Use text to support your answer.
Multiple populations and many individuals.
rs2236418
Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.

rs928197
Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.

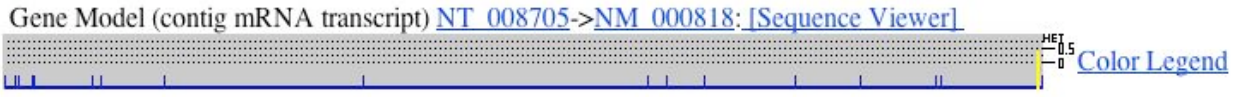
rs992990
Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.

It is worth noting that not all the submissions for a given SNP were validated. This leaves them open to a small amount of doubt.

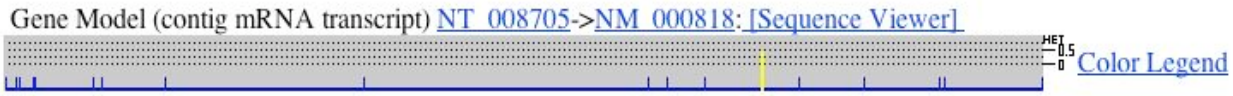
e) Do any of these 3 SNPs alter the protein primary structure? Support your answer with data from this web site.
rs2236418 = non-coding portion, but within intron. Different data required to show this.



rs928197 = non-coding portion, but within intron.



rs992990 = non-coding portion, but within intron.



Altering alternative splicing within an intron is possible, but less likely.

Now go to http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap/ and answer two more

questions.
f) Would you expect these 3 SNPs to be in linkage disequilibrium in any population or populations? Support your answer with data from this site.

For 2236418:		A	G
	CEU	.817	.183 (Utah Europeans)
	YRI	.067	.933
for 928197:	CEU	.833	.167
	YRI	.608	.392 (Yoruba in Ibadan, Nigeria)

If these alleles/SNPs were in LD, then you would expect them to be in similar ratios within a single popuation. You can see that the Yoruba population does not retain linkage, and thus it does not appear to be in LD. However, the European population does retain similar ratios, so this might be another example of population-specific differences.

g) Find a SNP for which this is no variation. Support your answer with data from this site (even though this question sounds like an oxymoron).

refSNP rs8190752 with alleles A/G in dbSNP (dbSNP report | Ensembl SNPview)

Chr10:26574684..26574684, (+) strand relative to the human reference sequence

Population	Genotype frequencies: Ref-homozygote genotype	freq	count	Heterozygote genotype	freq	count	Other-homozygote genotype	freq	count	Total count	Allele frequencies: Ref-allele	freq	count	Other-allele	freq	count	Total count	
CEU	G/G	1.000	60	A/G	0.000	0	A/A	0.000	0	60	G	1.000	120	A	0.000	0	120	retrieve genotypes
CHB	G/G	1.000	44	A/G	0.000	0	A/A	0.000	0	44	G	1.000	88	A	0.000	0	88	retrieve genotypes
JPT	G/G	1.000	44	A/G	0.000	0	A/A	0.000	0	44	G	1.000	88	A	0.000	0	88	retrieve genotypes
YRI	G/G	0.967	58	A/G	0.033	2	A/A	0.000	0	60	G	0.983	118	A	0.017	2	120	retrieve genotypes

Note: the 'reference' allele is the base observed in the reference genome sequence at this location

From the first view, it looks like there is no variation. But when you drill down, you see there is about a 2% frequency in the Yoruba population for the minor allele. The simplifying graphic was not sensitive enough to show small percentage.

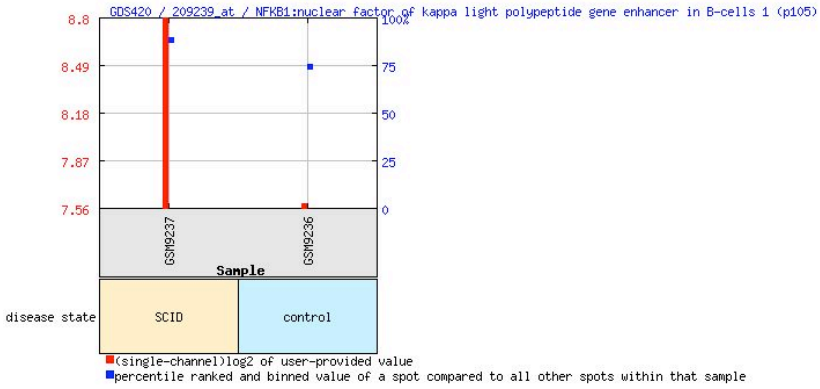
20 pts.
2) Use the attached Figure 1 PDF file to answer this question. Interpret figure 1 as completely as you can. Interpret the data and tell me what you can deduce about the biology being revealed. Principle components analysis is a way to objectively identify the portions of the data that are responsible for the most amount of inter-sample variation.

Some key points:
Panel a: It is hard to separate bins 0, 1, and 2. Bin 3 almost looks like a mistake was made with 2 of the 4 replicates being exchanged, but we will assume this did not happen.
Bin 4 shows the 2 hour effect with repressed genes moving towards ratio of 1. Bins 5 – 9 show the time cascade of different genes being induced during this immune challenge. Bins 0 – 4 plus 9 look indistinguishable for the placebo samples. Not sure why bins 5 – 8 are so different from the other genes for the placebo alone.
Panel b: Principle components show us which variables are most different from each other. Time points 0 and 24 are much like the placebos. Time point 2hrs is the most different from all others, indicating the initial gene response is very different from all subsequent gene responses. Times 4, 6 and 9 are roughly the same, though we see from panel a that different sets of genes (rows) are induced as time progresses. That makes panel a seem contradictory to the principle components analysis. If we believe the PC analysis, then the number and values of induced genes during early time points must be substantially different.

20 pts.

3) Use only this web site to answer the following questions: http://www.ncbi.nlm.nih.gov/geo/ . Search for this gene: NFKB1. (Read question #4 too so you will not have to redo any of this question.) Use screen shots to show one microarray example when this human gene was:

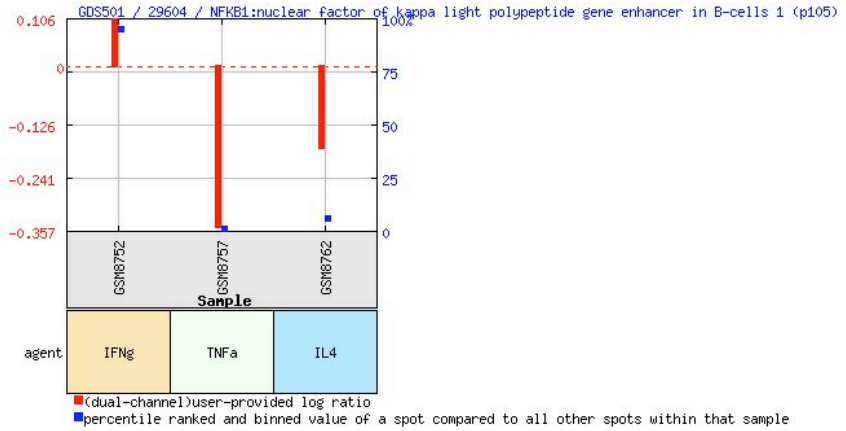
a) Strongly induced in one condition but not another. What were the conditions?



Human Severe Combine Immune Difficient (SCID) vs. wt human T cells. Single channel (Affy) chips, log transformed.

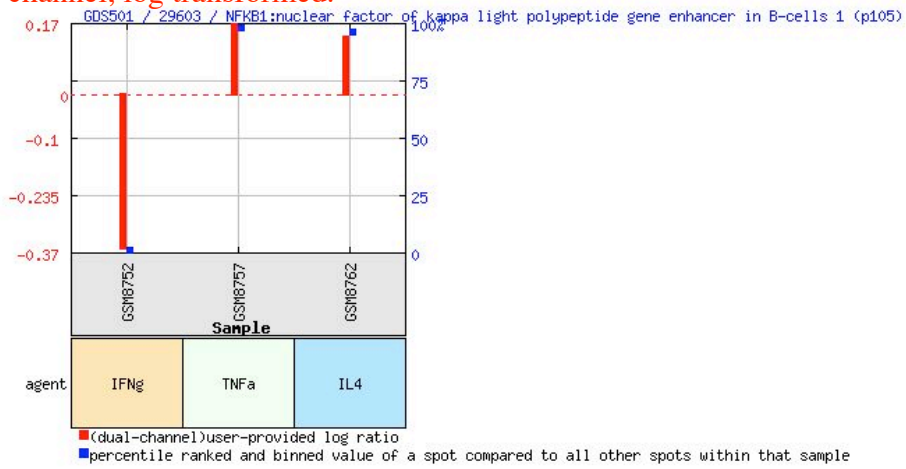
b) Repressed in both conditions. What were the conditions?

Examination of gene expression induced by interferon gamma (IFNg), tumor necrosis factor alpha (TNFa) and interleukin 4 (IL4) inflammatory cytokines on primary dermal endothelial cells. Dual channel arrays, log transformed.



c) Strongly repressed in only one condition. What were the conditions?

Examination of gene expression induced by interferon gamma (IFNg), tumor necrosis factor alpha (TNFa) and interleukin 4 (IL4) inflammatory cytokines on primary dermal endothelial cells. Dual channel, log transformed.



d) What are the meanings of the red and the blue symbols? Explain your answer in terms a Bio111 student could understand.

Red = log2 transformed signal after normalization (single channel) or ratios (dual channels). Normalization allows you to compare genes across different arrays.

Blue = percentage of signal for this gene compared to the microarray as a whole.

e) What is the value to knowing the answer to part d above?

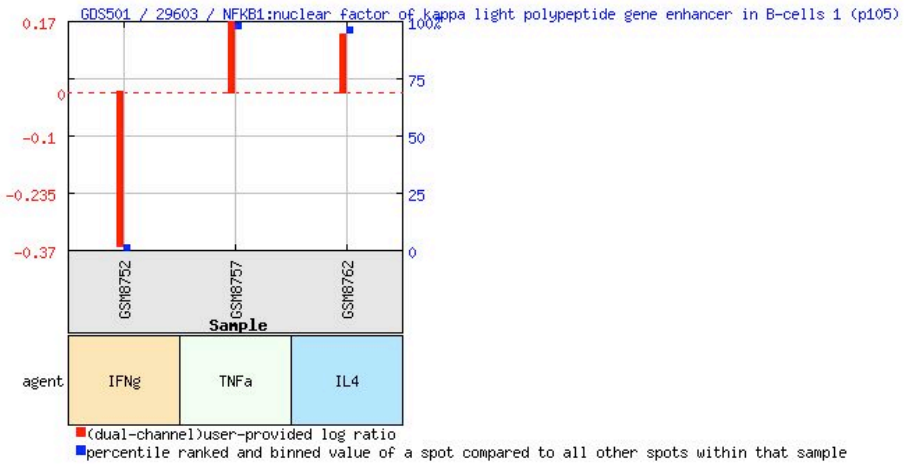
You want to make sure your spot is not in the bottom percentage for the blue dot. If it is, then the ratio for dual channel chips is not reliable.

20 pts.

4) Read all the parts to this question before you begin.

a) Use your answer to question 3 above that had a fold change the furthest from 1. Tell me which condition you chose, and supply me with a screen shot of the one you have chosen.

I chose



because my answer to part a was not a ratio, but a raw number. Note the Y axis in part a.

b) Tell me the fold change for your chosen gene and the experimental conditions.
1.29 fold repressed, which is not very much. I converted log2 to fold repression.

Examination of gene expression induced by interferon gamma (IFNg), tumor necrosis factor alpha (TNFa) and interleukin 4 (IL4) inflammatory cytokines on primary dermal endothelial cells. Dual channel, log transformed.

c) Convert the fold change to a ratio of two numbers that is consistent with your data.
1000/1290

d) If control is green and experimental is red, what color spot would you see on the microarray, assuming this is not an Affy chip? To answer this question, you must draw the circle and color in the spot here →

More green than red, but a mixture.

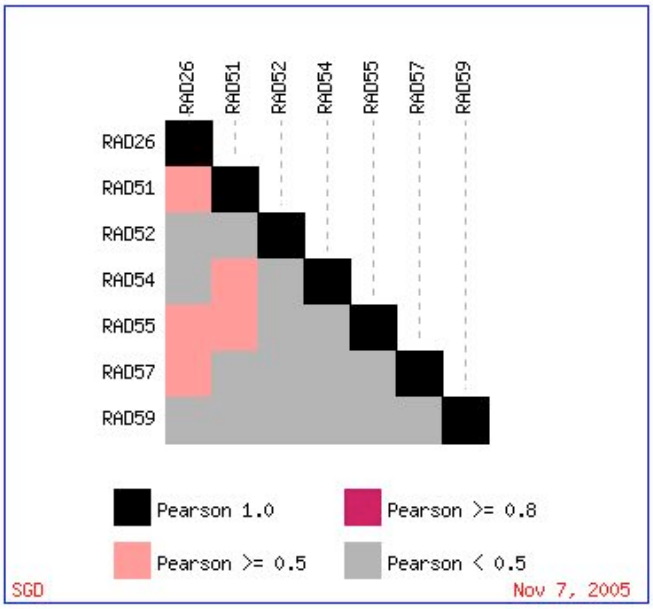
e) Draw an arrow on this color scale to indicate the color you'd choose for your example's ratio:

20 pts.

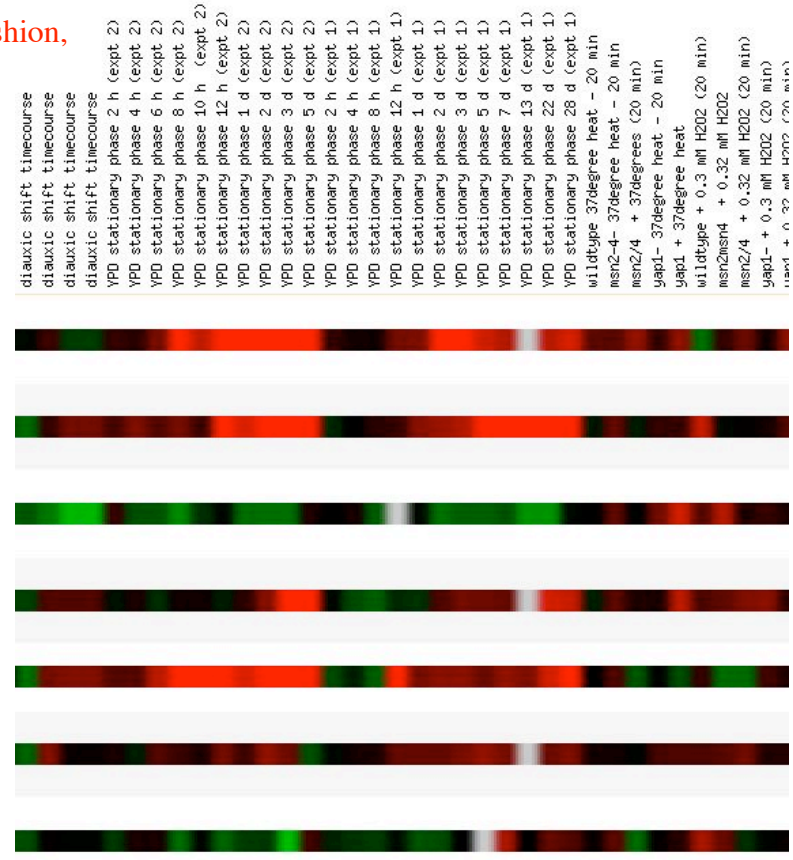
5) Use http://db.yeastgenome.org/cgi-bin/expression/expressionConnection.pl to answer the following questions concerning this list of yeast genes: *Rad26, Rad51, Rad52, Rad54, Rad55, Rad57,* and *Rad59*.

a) Are these genes transcribed in a coordinated fashion when exposed to environmental stresses? Support your answer with data from this web site only.

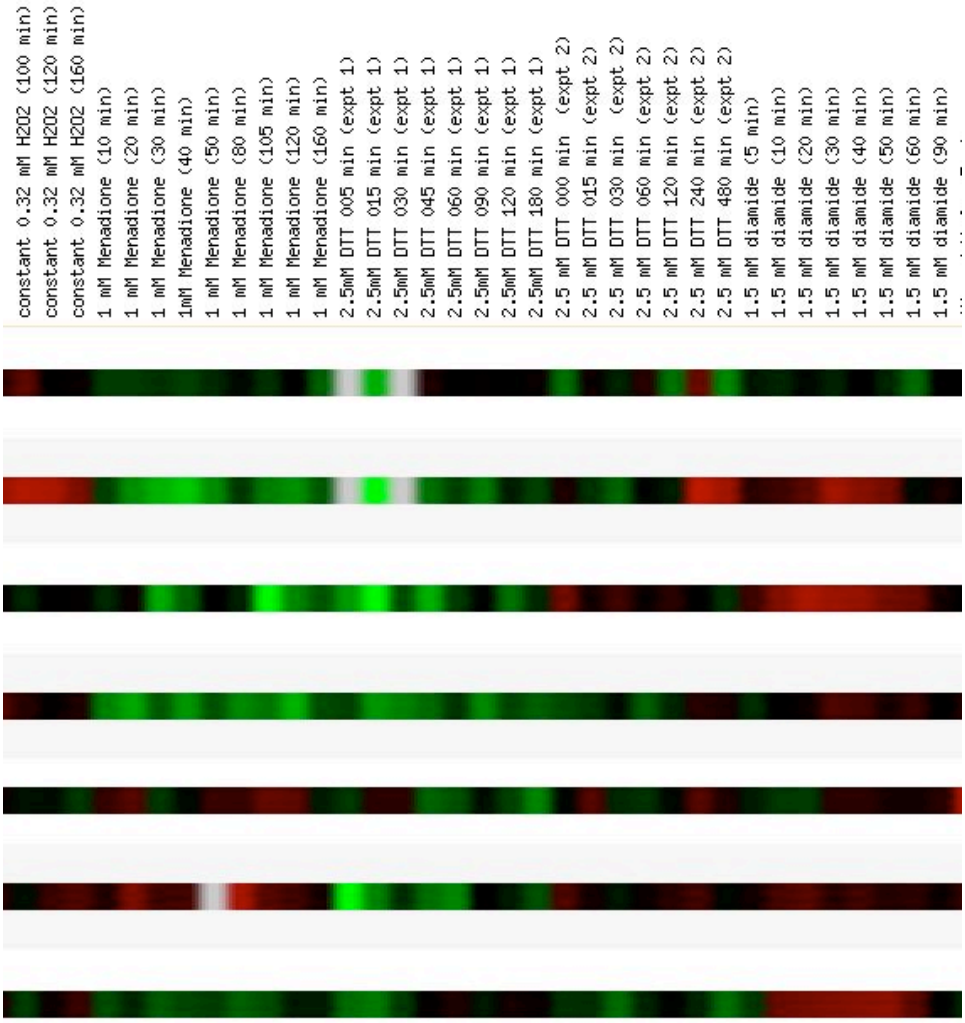
No, they are not well coordinated. You can tell this easily by looking at the correlation coefficient. They are below 50% or below 80%.



However, if you look at subsets of conditions, you can see some coordination:
For example, the stationary phase seems to induce several of them in a coordinated fashion, and this is reproducible.

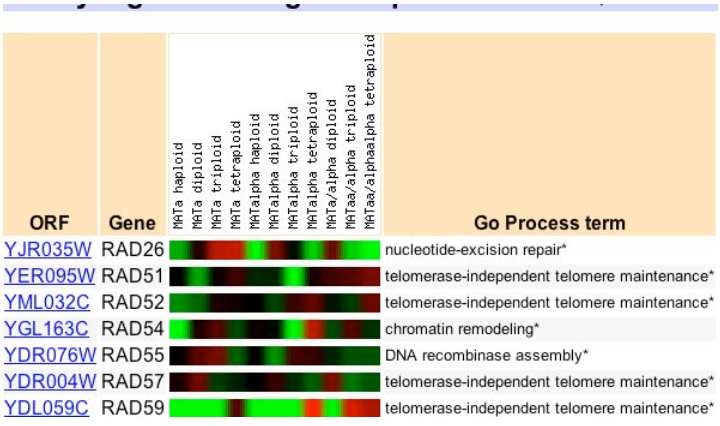


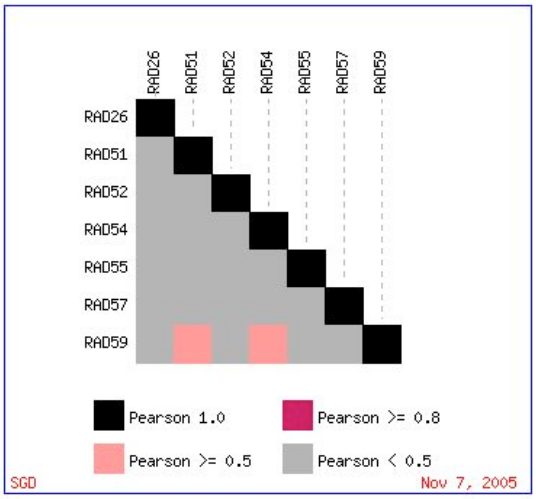
But other experiments are not as reproducible and thus the overall correlation coefficient is not a meaningful evaluation since it hides a coordination at stationary phase and uses the lack of reproducibility to evaluate overall co-regulation.



b) Are these genes transcribed in a coordinated fashion when the genome ploidy is altered? Support your answer with data from this web site only.

There is even less co-regulation with different ploidy.





c) Use data on this web site only to support the claim that the expression profiles for these 7 genes under the two conditions above (parts a and b) accurately represents independent gene regulation and not either of two common microarray artifacts. Name each artifact then show and describe data that demonstrate each artifact is not in play for these 7 genes.
One artifact is isozyme binding and since they are not co-regulated they must not be cross-reacting to inappropriate spots.
Another artifact is aneuploidy. From their ORF names, we can see that only the last 3 are on the same chromosome, and fairly near each other. However, they are not co-regulated either, so this does not seem to be a major factor in this analysis.

d) One artifact cannot be argued away with these genes. What artifact is this and what information do you need in order to evaluate its presence or absence?
We do not know how much signal there was for each spot and since spots with low signal can have widely different ratios, this artifact requires pixel values to determine whether low signal played a roll in appearing to be not co-regulated.