

Stat 532 Statistical Genetics II Homework 3

Erik Erhardt

May 11, 2005

1 Part I

Determine a list of genes that show a significant difference in expression between the two strains of mice over the six time points.

First I describe the criteria I used for reducing the set of possible genes, then I provide a list of 40 significant genes. Because we are interested in identifying promising genes, I do not chose genes strictly by significance at a typical .05 or .10 level, but start by including every gene that shows promise, then refine the list from there.

From the original set of 8993 genes, I retained those with Sidak adjusted p-values that were less than 1, essentially those with any signal of significance. I combined the genes retained from the six time points into a single set of 293 records. Next, I retained those that either appeared on more than 1 time point or had a Bonferroni p-value less than 1, either likely be a prolonged response to the pathogen or had a slightly stronger signal (than based on Sidak). From the remaining 90 I removed those occuring at time 0h that had Bonferroni p-value 1, since a gene expressed at time 0 was not due to the experience of the pathogen. The result is a list of 79 records, which represent 40 genes, many representing multiple times.

Table 1 on page 3 presents the list of 40 most likely genes to be a result of differential expression between mouse strains. These are sorted by significance and role. The gene number column indicates the NCBI identifier; “none” means the gene number was not provided in the probeset list, and “absent” means that probe ID was not in the probeset list. The time column gives the time points at which the gene showed significant difference in expression, in increasing order of significance. The last time shown is the one associated with the other columns and is the most significant. The Sig column indicates those genes that are significant at the Bonferroni .10 (**), and .20 (*) levels, and those that show expression over several time points with Bonferroni p-values less than 1 (–). Columns Holm and Hoch (same as Bonf), SidSS (same as SidSD), BH and BY were not included in the inter-

est of space. The name of the gene and chromosome location are also given where known.

The Role column indicates Gene Ontology chosen based on *a priori* belief about what response the mice would undergo. From the over 100 known functions, processes, and components, represented by this list of 40 genes, these seven seemed reasonable to a nonbiologist to consider important.

I, immune response Any process involved in the immunological reaction of an organism to an immunogenic stimulus.

M, granulocyte macrophage colony-stimulating factor receptor binding (eats up bad stuff) Interacting selectively with the granulocyte macrophage colony-stimulating factor receptor.

R, DNA repair The process of restoring DNA after damage. Genomes are subject to damage by chemical and physical agents in the environment (e.g. UV and ionizing radiations, chemical mutagens, fungal and bacterial toxins, etc.) and by free radicals or alkylating agents endogenously generated in metabolism. DNA is also damaged because of errors during its replication. A variety of different DNA repair pathways have been reported that include direct reversal, base excision repair, nucleotide excision repair, photoreactivation, bypass, double-strand break repair pathway, and mismatch repair pathway.

D, response to DNA damage stimulus A change in the state or activity of a cell (in terms of enzyme production, gene expression, etc.) as a result of damage to its DNA from environmental insults or errors during metabolism.

F, inflammatory response The immediate defensive reaction (by vertebrate tissue) to infection or injury caused by chemical or physical agents. The process is characterized by local vasodilation, extravasation of plasma into intercellular spaces and accumulation of white blood cells and macrophages.

N, neutrophil chemotaxis The movement of a neutrophil cell, the most numerous polymorphonuclear leukocyte found in the blood, in response to an external stimulus, usually an infection or wounding.

C, negative regulation of cell cycle Any process that stops, prevents or reduces the rate of cell cycle activity.

The plot in Figure 1 on page 4 plots the $\log(\text{p-value})$ versus time for the top 40 genes, illustrating that time point 4 strangely does not exhibit the degree of differential expression as during times 2, 3, 5, and 6. Apparently, these genes are expressed similarly between strains at 7 days.

Time	Sig	Name	Gene	Role	Name	Ch	Loc	t-stat	unadj-p	Bonf	SidSD
352	**	160163	94184	—	AA415817	16	A1	-32.54	5.31E-06	0.047	0.046
263	**	95974	14468	I	Gbp1	3	67.4 cM	30.00	7.35E-06	0.066	0.063
2	**	92948	12981	IM	Csf2	11	29.5 cM	27.81	9.94E-06	0.089	0.085
2	*	102171	12355	C	Nr1i3	1	92.6 cM	-25.00	1.52E-05	0.136	0.127
365	—	103918	57738		Slc15a2	16	B3	-17.05	6.94E-05	0.623	0.464
653	—	104444	77090		9430098E02Rik	8	B3.3	18.89	4.62E-05	0.415	0.339
26		103486	16176	IFN	Ii1b	2	73.0 cM	10.84	0.000409	1	0.974
56		94774	26388	I	Ifi202b	1	95.2 cM	10.42	0.000477	1	0.986
35		101027	30939	RD	Pttg1	11	A5	-11.45	0.000331	1	0.949
2		94396	26356	C	Ing1	8	A1.1	16.89	7.20E-05	0.647	0.476
32156	**	160799	none					-34.23	4.35E-06	0.039	0.038
6231	*	97713	94184					26.15	1.27E-05	0.114	0.107
13	*	103409	66898		Baiap2l1	5	G2	23.51	1.94E-05	0.174	0.160
136	—	96156	73824		1110008H02Rik	1		-17.38	6.43E-05	0.578	0.438
162	—	98084	107566		Arl2bp	8	C5	-15.57	9.92E-05	0.892	0.590
5163	—	101020	13032		Ctsc	7	D3-E1.1	-16.69	7.54E-05	0.677	0.491
1563	—	101044	17025		Alad	4	30.6 cM	17.07	6.90E-05	0.620	0.462
36		93145	none					-9.96	0.000570	1	0.993
63		93861	13034		Ctse	1	69.1 cM	-10.73	0.000426	1	0.978
21		93866	17313		Mglap	6	G1	-20.32	3.46E-05	0.311	0.267
56		94312	94184					-15.85	9.25E-05	0.831	0.564
5		94390	56399		Akap8	17	B2	18.48	5.04E-05	0.453	0.364
6		94426	233802		Thumpd1	7	F1	-17.18	6.72E-05	0.604	0.453
3		95142	none					-17.88	5.74E-05	0.516	0.402
2		97983	20910		Stxbp1	2	B	20.39	3.41E-05	0.307	0.264
5		99009	18115		Nnt	13	64.0 cM	20.27	3.50E-05	0.314	0.269
2		99652	193742		Bat5	17	B1	-17.14	6.79E-05	0.610	0.456
5		99833	73647		Capn9	8	E2	-16.34	8.20E-05	0.737	0.521
4		100885	18005		Nek2	1	103.0 cM	-15.45	0.000102	0.920	0.601
3		100916	absent					19.94	3.73E-05	0.335	0.284
2		101741	77132		2810433D01Rik	11	E1	-19.71	3.90E-05	0.350	0.295
5		102927	15194		Hdh	5	20.0 cM	-20.70	3.21E-05	0.288	0.250
65		103471	66687		Tbc1d15	10	D2	-8.14	0.001233	1	0.999
56		103828	224705		Vps52	17	B1	-11.04	0.000381	1	0.967
263		104263	100177		9330177P20Rik	4	D2.2	-10.60	0.000448	1	0.982
23		160179	69981		D9Wsu20e	9	E1	9.73	0.000623	1	0.996
52		160605	74841		Usp38	8	C3	-10.56	0.000453	1	0.982
6		161075	272636		D9Ertd280e	9	52.0 cM	19.43	4.14E-05	0.371	0.310
5		161265	16818		Lck	4	59.0 cM	-16.74	7.45E-05	0.670	0.488
3		161603	13824		Epb4.1l4a	18	B1	-21.18	2.94E-05	0.263	0.231

Table 1: Top 40 genes by significance.

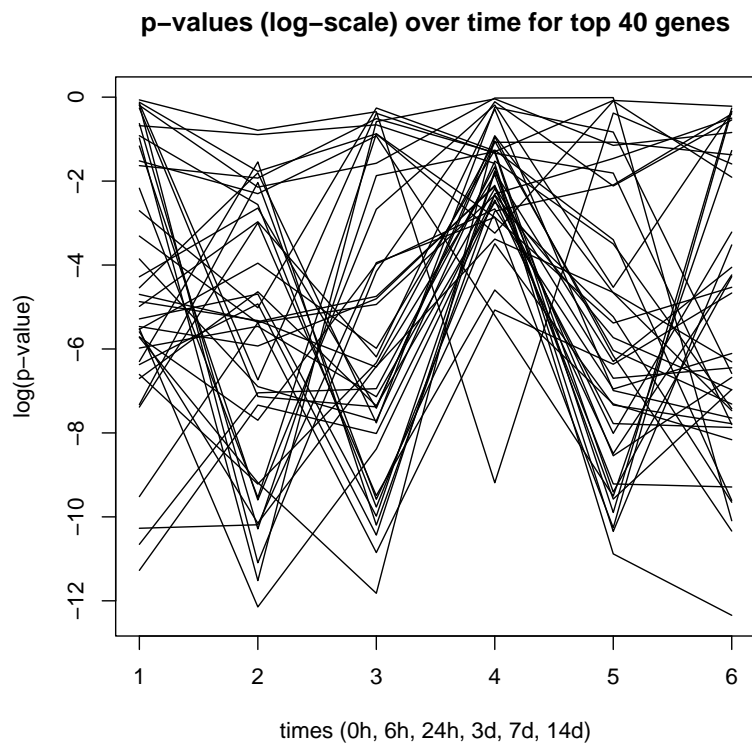


Figure 1: $\log(\text{p-value})$ versus time for the top 40 genes

2 Part II

What 5 (or so) genes should be pursued?

Referencing Table 1 on page 3, the top 40 list has been divided into four groups. The top group are Significant without time point 1, the second group has biological Roles, the third group is Sig but includes the first time point so may simply be a inherently differentially expressed protein between mouse strains, and the last group is the complement of the first three groups.

Based on the above microarray analysis, I would present the top two groups of (6 + 4 = 10) genes for further investigation, allowing Julie to continue based on statistical significance or known biological role.

Based on the previous QTL analysis, I would suggest there is at least one QTL on chromosome 6 near D6mit188, and a second near D6mit86 and D6Mit138. There is possibly another on chromosome 3 near D3Mit215. Gene 14468 (Gbp1) located on Chromosome 3 at 67.4 cM is outside the band of 50–60 cM identified by QTL analysis. However, the interval mapping was based on markers located far from 67.4 cM (closest are 58.8 and 71.8), so QTL has little information about this region of the chromosome. The one gene in my top 40 on chromosome 6 has a location identifier (G1) that I don't know how to convert to cM, so can not relate that to the QTL analysis.

My final sorted list of genes are given in Table 2 on page 5; these are my top 10 in groups of 5. The top gene does not have a listed function in the Gene Ontology (—), so this may play an important role and Julie could be the first to identify this particular function.

Time	Sig	Name	Gene	Role	Name	Ch	Loc	t-stat	unadj-p	Bonf	SidSD
352	**	160163	94184	—	AA415817	16	A1	-32.54	5.31E-06	0.047	0.046
2	**	92948	12981	IM	Csf2	11	29.5 cM	27.81	9.94E-06	0.089	0.085
263	**	95974	14468	I	Gbp1	3	67.4 cM	30.00	7.35E-06	0.066	0.063
26		103486	16176	IFN	Il1b	2	73.0 cM	10.84	0.000409	1	0.974
56		94774	26388	I	Ifi202b	1	95.2 cM	10.42	0.000477	1	0.986
365	—	103918	57738		Slc15a2	16	B3	-17.05	6.94E-05	0.623	0.464
653	—	104444	77090		9430098E02Rik	8	B3.3	18.89	4.62E-05	0.415	0.339
35		101027	30939	RD	Pttg1	11	A5	-11.45	0.000331	1	0.949
2	*	102171	12355	C	Nr1i3	1	92.6 cM	-25.00	1.52E-05	0.136	0.127
2		94396	26356	C	Ing1	8	A1.1	16.89	7.20E-05	0.647	0.476

Table 2: Top 10 genes for further investigation based on significance and known function.

Appendix

R code used for the above analysis

```

# options for printing
ps.options(horizontal=FALSE,height=6,width=6)
print.switch= 0; # 0=display only, 1=home directory, 2=school directory
#### if(print.switch==1){postscript("f:\\USERS\\Erik\\UNM\\statgen\\sghw1_2.ps");}
#### if(print.switch==2){postscript("x:\\statgen\\sghw1_2.ps");}
####plot(x,prior1,type="l")
####title("Prior with mean=1, std=.5");
#### if(print.switch>0){dev.off()};

#####
#####
#
# Make the packages available in R
library(multtest);

# help on functions in multtest
# help("mt.teststat")
# help("mt.rawp2adjp")

#####
# 1. MVN with 2 means

# read data
# Screen out all absent genes and start normalization and transformation in R
newdat.org<-read.table("F:\\USERS\\Erik\\UNM\\statgen\\unnorm_screen_use.txt",header=T,sep="\t")
newdat = newdat.org[,2:37]
# Check data
dim(newdat)
newdat[1:10,]

# Normalize by dividing by the median
newdat2<-matrix(0,ncol=36,nrow=8993)
for (i in 1:36) newdat2[,i]<-newdat[,i]/median(newdat[,i],na.rm=T)

# Now log transform
newdat3<-matrix(0,ncol=36,nrow=8993)
for (i in 1:36) newdat3[,i]<-log(newdat2[,i])

# Create a data frame so that gene names will be associated
datause<-data.frame(newdat3,row.names=newdat.org[,1])

# boxplots for each group
boxplot(datause[,4],datause[,5],datause[,6],datause[,10],datause[,11],datause[,12],
        datause[,16],datause[,17],datause[,18],datause[,19],datause[,23],datause[,24],
        datause[,28],datause[,29],datause[,30],datause[,34],datause[,35],datause[,36])
boxplot(datause[,1],datause[,2],datause[,3],datause[,7],datause[,8],datause[,9],
        datause[,13],datause[,14],datause[,15],datause[,19],datause[,20],datause[,21],
        datause[,25],datause[,26],datause[,27],datause[,31],datause[,32],datause[,33])

# plot means within each group against one another
plot(rowMeans(datause[,1:3]),rowMeans(datause[,4:6]),xlab="Mean Expression Values for B6 Mice at Time 0",ylab="Mean

# get fold-changes and find those that exceed a threshold
time0fold=rowMeans(datause[,1:3])/rowMeans(datause[,4:6])
bigtime0fold_b6=subset(time0fold,time0fold>10)
bigtime0fold_b6
summary(bigtime0fold_b6)

# make the sigtest function available in R
source("F:\\USERS\\Erik\\UNM\\statgen\\sigtest.txt")

# perform statistical tests for differential expression, adjusting for multiple comparisons
times=1*6-5;sigtest(times,times+3,"F:\\USERS\\Erik\\UNM\\statgen\\sghw3_sigtime1.txt","t","n",1000)
times=2*6-5;sigtest(times,times+3,"F:\\USERS\\Erik\\UNM\\statgen\\sghw3_sigtime2.txt","t","n",1000)
times=3*6-5;sigtest(times,times+3,"F:\\USERS\\Erik\\UNM\\statgen\\sghw3_sigtime3.txt","t","n",1000)
times=4*6-5;sigtest(times,times+3,"F:\\USERS\\Erik\\UNM\\statgen\\sghw3_sigtime4.txt","t","n",1000)
times=5*6-5;sigtest(times,times+3,"F:\\USERS\\Erik\\UNM\\statgen\\sghw3_sigtime5.txt","t","n",1000)
times=6*6-5;sigtest(times,times+3,"F:\\USERS\\Erik\\UNM\\statgen\\sghw3_sigtime6.txt","t","n",1000)

# the 40 genes, unadjusted p-values at all 6 time points
unad.p.40 = c(
  1 , 0.376633255, 9.94E-06, 0.558626798, 0.958843762, 0.316899361, 0.430866367, # "92948_at"
  2 , 0.013835371, 0.071122057, 0.000603853, 0.39877106, , 0.009437568, 0.000570399, # "93145_at"
  3 , 0.010650535, 0.164615645, 0.000426288, 0.783204982, 0.435172179, 0.001383645, # "93861_f_"
  4 , 3.46E-05, 3.76E-05, 0.01937147, , 0.057066949, 0.004598789, 0.01077959, # "93866_s_"
  5 , 0.004230959, 0.002660182, 0.008198525, 0.122171535, 9.95E-05, 9.25E-05, # "94312_at"
  6 , 0.195569297, 0.145362559, 0.599898771, 0.266999856, 5.04E-05, 0.770615079, # "94390_at"
  7 , 0.114123531, 7.20E-05, 0.416822845, 0.039128341, 0.912005535, 0.148366423, # "94396_at"
  8 , 0.003955668, 0.01914366, , 0.002519657, 0.283455072, 0.03347787, , 6.72E-05, # "94426_at"
  9 , 0.006835847, 0.051362049, 0.002053115, 0.74924837, , 0.000920768, 0.000477978, # "94774_at"
  10 , 0.399299046, 0.077291909, 5.74E-05, 0.066041785, 0.123368321, 0.671024178, # "95142_s_"

```

```

11 , 0.003269654, 0.000102795, 7.35E-06, 0.390206902, 0.002107495, 3.24E-05, # "95974_at"
12 , 0.001228162, 0.009685554, , 0.000639756, 0.081719379, 0.005340435, 6.43E-05, # "96156_at"
13 , 1.27E-05, 0.000650763, 0.000330714, 0.097453381, 0.003249834, 0.000936962, # "97713_at"
14 , 0.532555002, 3.41E-05, 0.154096181, 0.272675041, 0.924470476, 0.805288242, # "97983_s"
15 , 0.001351364, 9.92E-05, 0.00178401, , 0.350926057, 0.002401598, 0.000667567, # "98084_at"
16 , 0.885405983, 0.170505737, 0.417466425, 0.057082534, 3.50E-05, 0.713440097, # "99009_at"
17 , 0.311802302, 6.79E-05, 0.069184598, 0.828028868, 0.010776282, 0.704949251, # "99652_at"
18 , 0.066777952, 0.006616924, , 0.771893338, 0.277649813, 8.20E-05, 0.013904663, # "99833_at"
19 , 0.936554725, 0.453868971, 0.706469426, 0.00010236, , 0.683793519, 0.203575119, # "100885_a"
20 , 0.001840305, 0.213655993, 3.73E-05, 0.099877767, 0.231895722, 0.605879873, # "100916_a"
21 , 0.000621306, 0.050050744, 7.54E-05, 0.010099096, 0.000669998, 0.000285288, # "101020_a"
22 , 0.002533289, 0.004250413, 0.000794665, 0.11318939, , 0.000331906, 0.040113131, # "101027_s"
23 , 0.000659068, 0.130758118, 6.90E-05, 0.030432699, 0.000652334, 0.000418288, # "101044_a"
24 , 0.003727392, 3.90E-05, 0.00164046, , 0.033997866, 0.009525432, 0.001824809, # "101741_a"
25 , 0.834695296, 1.52E-05, 0.01569933, , 0.895780364, 0.120419764, 0.580893943, # "102171_r"
26 , 0.506799014, 0.409572324, 0.517160072, 0.205804952, 3.21E-05, 0.029561274, # "102927_s"
27 , 7.36E-05, 0.004634361, 1.94E-05, 0.006270679, 0.001723806, 0.009402885, # "103409_a"
28 , 0.009108807, 0.004753475, 0.008686423, 0.118832338, 0.001233563, 0.001581485, # "103471_a"
29 , 0.862532545, 0.001187515, 0.681183113, 0.262732107, 0.163777121, 0.000409782, # "103486_a"
30 , 0.036510882, 0.004625912, 0.007161012, 0.096576352, 0.000419247, 0.000381783, # "103828_a"
31 , 0.007663385, 0.004878039, 0.001609172, 0.157697793, 6.94E-05, 0.00094395, # "103918_a"
32 , 0.021182627, 0.001011821, 0.000448255, 0.217416473, 0.030273257, 0.000597552, # "104263_a"
33 , 0.005074474, 0.009172739, 4.62E-05, 0.172704954, 0.000194773, 0.001259921, # "104444_a"
34 , 0.004015421, 5.31E-06, 0.00022843, , 0.086185388, 0.000205351, 0.014462097, # "160163_a"
35 , 0.769241458, 0.000798605, 0.000623564, 0.340940926, 0.341773178, 0.253396329, # "160179_a"
36 , 0.003312162, 0.000453523, 0.018770829, 0.069699764, 0.000955262, 0.002227236, # "160605_s"
37 , 2.36E-05, 0.00087163, , 0.000962514, 0.19035429, , 1.88E-05, 4.35E-06, # "160799_a"
38 , 0.821978799, 0.117091519, 0.207012778, 0.978306028, 0.988691091, 4.14E-05, # "161075_a"
39 , 0.219059969, 0.100291373, 0.397263698, 0.005516332, 7.45E-05, 0.280014854, # "161265_f"
40 , 0.001708978, 0.007244161, 2.94E-05, 0.167546518, 0.001838511, 0.017358713) # "161603_r"

unad.p.40.m = matrix(data=unad.p.40,ncol=7,byrow=TRUE)
unad.p.40.times = matrix(data=rep(0:6,40),ncol=7,byrow=TRUE)
if (print.switch==1){postscript("f:\\USERS\\Erik\\UNM\\statgen\\sghw3_1.ps");}
if (print.switch==2){postscript("x:\\statgen\\sghw3_1.ps");}
plot(unad.p.40.times[,2:7],log(unad.p.40.m[,2:7]),"n",
     main="p-values (log-scale) over time for top 40 genes",
     xlab="times (0h, 6h, 24h, 3d, 7d, 14d)",
     ylab="log(p-value)")
for (i in 1:40)
{
  lines(unad.p.40.times[i,2:7],log(unad.p.40.m[i,2:7]))
}
if (print.switch>0){dev.off();}

# plot for gene 102904
group1=matrix(c(datause[1547,2:4],datause[1547,8:10],datause[1547,14:16],datause[1547,20:22],datause[1547,26:28],datause[1547,32:34]),nrow=1)
group2=matrix(c(datause[1547,5:7],datause[1547,11:13],datause[1547,17:19],datause[1547,23:25],datause[1547,29:31],datause[1547,35:37]),nrow=1)
timevec=matrix(c(0,0,0,6,6,6,24,24,24,72,72,72,168,168,168,336,336,336),ncol=18,nrow=1)
plot(timevec,group1,col="blue",xlab="Time (hours)",ylab="Normalized Expression Value",main="Gene 102904")
points(timevec,group2,col="red")

#####
#When starting R, type library(multtest)
#Data must be stored in data frame called datause
#start1 is the column number to begin at for sample 1
#start2 is the column number to begin at for sample 2
#fout is the name of the output file to which results will be written
#whichtest determines which of the tests will be used
#np = "y" for nonparametric tests, "n" for parametric tests
#numboot is the number of permutation replicates

sigtest<-function(start1,start2,fout,whichtest,np,numboot)
{
#Determine what columns of data matrix to use
end = start1+2
data1u<-datause[,start1:end]
end = start2+2
data2u<-datause[,start2:end]
alldata<-cbind(data1u,data2u);
data.cl<-c(rep(0,3),rep(1,3));
#Compute test statistics for each gene.
#The parameter whichtest determines what test is used.
#The parameter np determines whether the parametric or nonparametric version is used.
#See help("mt.teststat") in R for details of these parameters.
teststat<-mt.teststat(alldata,data.cl,test=whichtest,nonpara=np)
#Get p-values and adjusted p-values.
rawp0<-2*(1-pt(abs(teststat),4));

```

```

procs<-c("Bonferroni", "Holm", "Hochberg", "SidakSS", "SidakSD", "BH", "BY");
res<-mt.rawp2adjp(rawp0,procs);
adjp<-res$adjp[order(res$index),];
#Get p-values and adjusted p-values using step-down procedures.
#maxT
resT<-mt.maxT(alldata,data.cl,test=whichtest,B=numboot,nonpara=np);
ord1<-order(resT$index);
rawpT<-resT$rawp[ord1];
maxT<-resT$adjp[ord1];
teststat2<-resT$teststat[ord1];
#minP
resP<-mt.minP(alldata,data.cl,test=whichtest,B=numboot,nonpara=np);
ord2<-order(resP$index);
rawpP<-resP$rawp[ord2];
minP<-resP$adjp[ord2];
teststat3<-resP$teststat[ord2];cat("\n");
#Create output dataset
dataout<-matrix(0,nrow=8993,ncol=14);
dataout[, 1]<-teststat;
dataout[, 2]<-rawp0;
dataout[, 3]<-adjp[,2];
dataout[, 4]<-adjp[,3];
dataout[, 5]<-adjp[,4];
dataout[, 6]<-adjp[,5];
dataout[, 7]<-adjp[,6];
dataout[, 8]<-adjp[,7];
dataout[, 9]<-adjp[,8];
dataout[,10]<-rawpT;
dataout[,11]<-maxT;
dataout[,12]<-minP;
dataout[,13]<-(start1+5)/6; # time group
dataout<-data.frame(dataout,row.names=newdat.org[,1])
dataoutnames<-c("name","teststat","unadjp","Bonferroni","Holm","Hochberg","SidakSS","SidakSD","BH","BY","unadjpermp")
write.table(dataout,fout,row.names=TRUE,col.names=dataoutnames, sep="\t", quote=FALSE, na="");
#return(dataout);s
}
# end

```