

# Day 3 Bioconductor 2008 Longwood course

March 7, 2008

## Contents

1	Data resources	1
2	Affy SNP chips	1
3	Illumina – raw MAQC data	4

## 1 Data resources

## 2 Affy SNP chips

Archive: gliPack.zip

Length	Date	Time	Name
-----	----	----	----
0	03-04-08	13:07	HIND/
114	03-04-08	13:06	HIND/doc.hist
25641293	02-15-08	14:29	HIND/GSM248457.CEL
25658564	02-15-08	14:29	HIND/GSM248458.CEL
25664717	02-15-08	14:29	HIND/GSM248468.CEL
25643596	02-15-08	14:29	HIND/GSM248469.CEL
25637357	02-15-08	14:29	HIND/GSM248522.CEL
25661493	02-15-08	14:29	HIND/GSM248524.CEL
13473013	03-02-08	21:53	HIND/hindCRLMM.rda
0	03-04-08	13:07	XBA/
116	03-04-08	13:07	XBA/dox.hist
25651694	02-15-08	14:28	XBA/GSM248244.CEL
25655151	02-15-08	14:28	XBA/GSM248245.CEL
25657892	02-15-08	14:29	XBA/GSM248255.CEL
25641899	02-15-08	14:29	XBA/GSM248256.CEL
25649886	02-15-08	14:29	XBA/GSM248309.CEL

```

25651094 02-15-08 14:29 XBA/GSM248311.CEL
13562608 03-02-08 21:53 XBA/xbaCRLMM.rda
27028881 03-02-08 22:12 gli100k.rda
      2194 03-04-08 13:05 workComb.hist
-----
361881562                20 files

```

It takes a long time on most laptops, but you could try to read some CEL files with code like:

```

> NSAMP = 2
> pd = new("AnnotatedDataFrame", data = data.frame(gender = rep("male",
+      NSAMP)))
> hindCRLMM = justCRLMM(dir(patt = "CEL")[1:NSAMP], phenoData = pd)

```

This generates an instance of SnpSetCallPlus. I have gone through the process of reading, preprocessing and summarizing the SNP data for both XBA and HIND chips yielding the gli100k.rda object.

```

> library(Biobase)
> library(oligo)
> library(pd.mapping50k.xba240)
> load("gli100k.rda")
> calls(gli100k)[1:5, 1:5]

```

	GSM248244.CEL	GSM248245.CEL	GSM248255.CEL	GSM248256.CEL
SNP_A-1507972	2	3	3	2
SNP_A-1510136	3	3	2	3
SNP_A-1511055	3	2	3	3
SNP_A-1518245	3	3	3	2
SNP_A-1641749	3	3	3	3
	GSM248309.CEL			
SNP_A-1507972	2			
SNP_A-1510136	2			
SNP_A-1511055	3			
SNP_A-1518245	3			
SNP_A-1641749	3			

To resolve the AFFY SNP identifiers, you need

```

> con = pd.mapping50k.xba240@getdb()
> con

```

```

<SQLiteConnection:(887,0)>

```

```
> dbListTables(con)
```

```
[1] "featureSet" "mmfeature" "pm_mm" "pmfeature" "qcommfeature"  
[6] "qcpm_qcmm" "qcpmfeature" "sequence" "sqlite_stat1" "table_info"
```

```
> dbGetQuery(con, "select * from featureSet limit 1000,5")
```

	fsetid	man_fsetid	affy_snp_id	dbsnp_rs_id	chrom	physical_pos	strand
1	1001	SNP_A-1710359	NA	rs898783	18	38154861	+
2	1002	SNP_A-1723176	NA	rs10495114	1	217155629	-
3	1003	SNP_A-1665918	NA	rs965437	14	33711388	-
4	1004	SNP_A-1746234	NA	rs315408	6	153591366	+
5	1005	SNP_A-1729788	NA	rs212970	1	48490395	+

	cytoband	allele_a	allele_b
1	q12.3	A	G
2	q41	A	C
3	q13.1	A	C
4	q25.2	A	G
5	p33	G	T

```
1  
2  
3 NM_022073 // upstream // 221351 // Hs.135507 // EGLN3 // 112399 // Homo sapiens egl n  
4  
5 NM_001011547 // downstream // 3492 // Hs.3789
```

	fragment_length	dbsnp	cnv
1	853	1	<NA>
2	1052	0	<NA>
3	1072	1	<NA>
4	761	0	<NA>
5	1302	0	<NA>

The sample phenoData (in text) are

```
primary glioblastoma 1: 244 X, 457 H  
primary glioblastoma 2: 245 X, 458 H  
anaplastic oligodendroglioma 1: 255 X, 468 H  
anaplastic oligodendroglioma 2: 256 X, 469 H  
anaplastic astrocytoma 5: 309 X, 522 H  
anaplastic astrocytoma 6: 311 X, 524 H
```

where nnn X denotes the suffix of the GSM number for Xba element of 100k set, nnn H is the suffix of the GSM number for Hind element.

Exercise: make the phenoData structure and attach. Is there a SNP among the first 1000 on the chips that distinguishes primary glio from anaplastic astro in the sense that subjects with one of the tumor types are consistently homozygous rare and the subjects with the other one are consistently homozygous common? What is the SNP?

Note:

```
> sampleNames(gli100k)

[1] "GSM248244.CEL" "GSM248245.CEL" "GSM248255.CEL" "GSM248256.CEL"
[5] "GSM248309.CEL" "GSM248311.CEL"

> table(as.numeric(calls(gli100k)))

      1      2      3
259481 180787 256956
> distt = function(x) all((x[1:2] - x[5:6]) == 2)
> apply(calls(gli100k)[1:1000,], 1, distt) -> chk
> any(chk)
[1] TRUE
> which(chk)
SNP_A-1642215
      236
> calls(gli100k)[236,]
GSM248244.CEL GSM248245.CEL GSM248255.CEL GSM248256.CEL GSM248309.CEL
              3              3              3              1              1
GSM248311.CEL
              1
```

### 3 Illumina – raw MAQC data

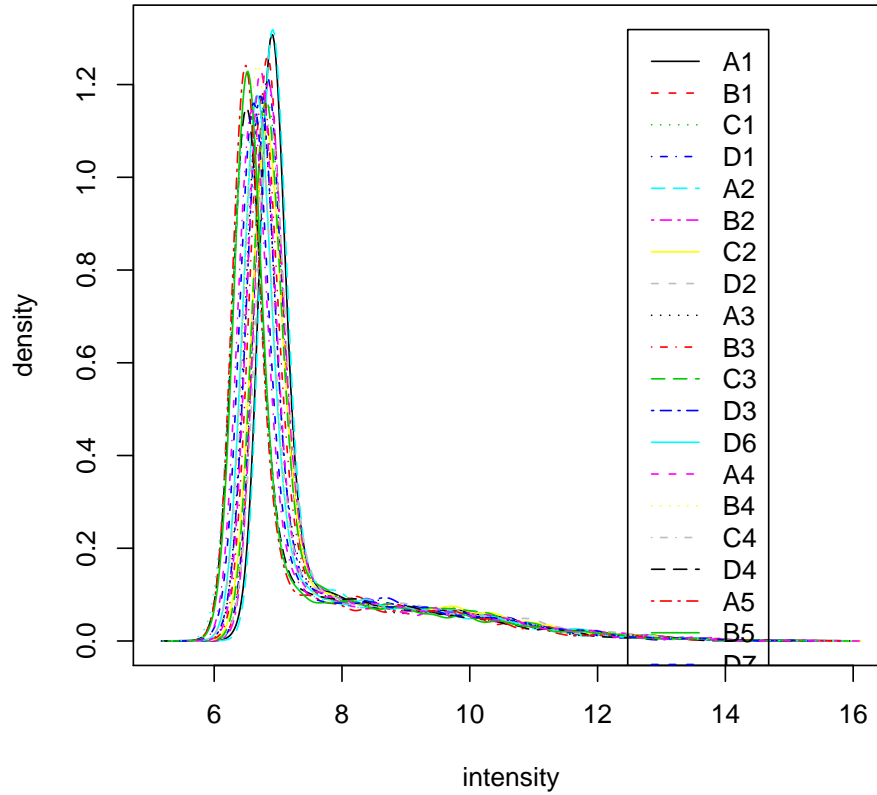
Use lumiR to capture data in ILM\_2\_ALL\_raw\_gene\_profile.csv. Perform background correction (use bgAdjust.affy) and normalization, and view the boxplots at all three stages: raw, background-corrected, and (default) normalized. Show the effect of normalization on expression measures of a selected gene.

```
> library(lumi)

This is mgcv 1.3-29

> library(genefilter)
> library(limma)
> ilm2 = lumiR(dir(patt = "ILM"))
> annotation(ilm2) = "illuminaHumanv1.db"
> plot(ilm2)
```

Density plot of intensity



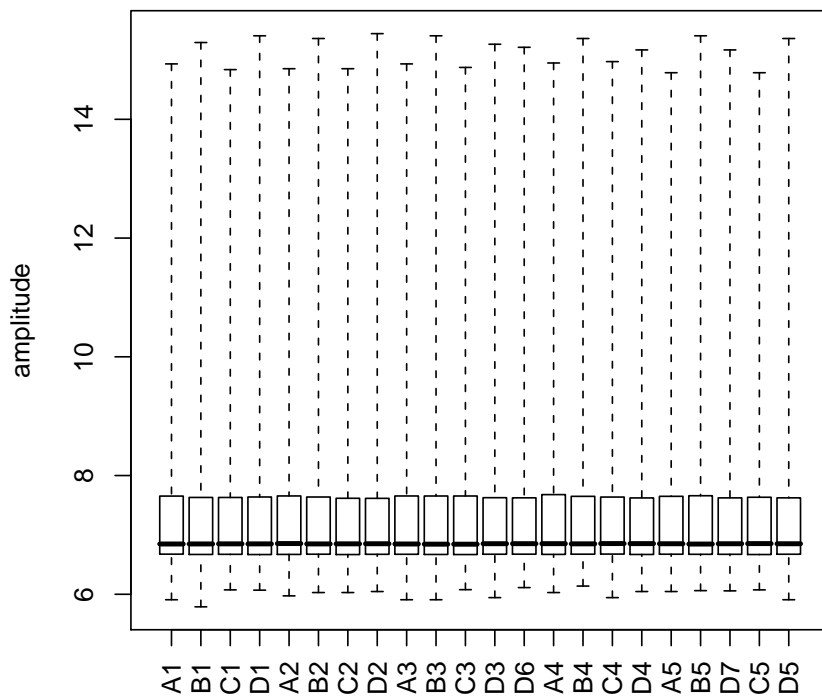
```
> boxplot(ilm2)
```

```
> args(lumiN)
```

```
function (x.lumi, method = c("quantile", "rsn", "ssn", "loess",  
"vsn"), ...)
```

```
NULL
```

```
> boxplot(lumiN(ilm2))
```



1. install new AnnotationDbi 1.1.24 and illuminaHumanv1.db (in illuminaHumanv1.zip, sic)
2. Create phenoData with pctBrain correctly assigned to the samples.
3. Find genes that are differentially expressed in the A vs B comparison.
4. Find the GO category and KEGG pathway assignments of the top 10 genes. Build an HTML table for these genes.

```

> library(illuminaHumanv1.db)

> bilm2 = lumiB(ilm2, method = "bgAdjust.affy")
> nilm2 = lumiN(bilm2)
> mads = apply(exprs(nilm2), 1, mad)
> bm = which(mads > quantile(mads, 0.8))
> isAB = which(substr(sampleNames(nilm2), 1, 1) %in% c("A", "B"))
> f2 = nilm2[bm, isAB]

```

```

> isB = 1 * (substr(sampleNames(f2), 1, 1) == "B")
> f2$isBrain = factor(isB)
> mm = model.matrix(~isBrain, data = pData(f2))
> l1 = lmFit(f2, mm)
> e11 = eBayes(l1)

> tt = topTable(e11, 2, 10)
> tt

```

	TargetID	ID	logFC	AveExpr	t	P.Value
4404	GI_32401419-S	GI_32401419-S	6977.395	3507.104	665.5817	6.113056e-23
5129	GI_34335281-I	GI_34335281-I	6689.040	3747.648	185.0331	8.351982e-18
7272	GI_4504138-S	GI_4504138-S	3950.762	1983.158	183.5596	8.992055e-18
7525	GI_4507456-S	GI_4507456-S	-4971.832	3713.065	-168.3521	1.998659e-17
3511	GI_28302130-S	GI_28302130-S	-37262.322	18662.817	-156.7556	3.863565e-17
1009	GI_16554578-S	GI_16554578-S	-1949.824	1047.892	-154.7828	4.342935e-17
2586	GI_22748688-S	GI_22748688-S	2941.165	1596.995	151.1170	5.418943e-17
7384	GI_4505466-S	GI_4505466-S	-2017.994	1131.891	-141.8770	9.704072e-17
7160	GI_4502858-S	GI_4502858-S	-2686.078	1397.780	-137.3527	1.308944e-16
3221	GI_26665889-I	GI_26665889-I	-2585.702	2375.802	-130.8629	2.046552e-16
	adj.P.Val	B				
4404	5.782340e-19	-4.464246				
5129	2.835195e-14	-4.464283				
7272	2.835195e-14	-4.464283				
7525	4.726329e-14	-4.464291				
3511	6.846637e-14	-4.464298				
1009	6.846637e-14	-4.464300				
2586	7.322540e-14	-4.464303				
7384	1.147385e-13	-4.464311				
7160	1.375700e-13	-4.464315				
3221	1.858200e-13	-4.464322				

```

> library(annaffy)
> nn = tt[, 1]
> myt = aafTableAnn(nn, "illuminaHumanv1.db")
> saveHTML(myt, file = "lkbr.html")

```