GS01 0163
Analysis of Microarray Data

Keith Baggerly and Brad Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

14 October 2010
Lecture 13: Limma and TCGA

- Linear Models – parallel fits, and borrowing strength
- Design Matrices and Contrast Matrices
- TCGA — What is it?
Looking at Contrasts in R

We talked earlier about incorporating multiple covariates into our modeling, and pointed out that the general statistical extension was the linear model.

Today, I want to introduce limma, which is, as you might guess, “linear models for microarrays”.

This takes many standard statistical tests and codes them rather efficiently for (a) massive parallelization and (b) borrowing across arrays.

Much of what follows today is taken straight from the User’s manual.
Example 1: Contrasting Two Groups

Our first case study involves an E. coli knockout experiment, as described in Hughes et al. *J Biol Chem*, **277**:40309-23, 2002. In it, 4 wild-type samples are contrasted with 4 samples from which Lrp has been knocked out. The dataset is available from BioConductor as ecoliLeucine (we also need ecolicddf).

```r
> library(ecolicddf)
> library(ecoliLeucine) # loads affy, Biobase
> data(ecoliLeucine) # an AffyBatch
> eLeuRMA <- rma(ecoliLeucine)
```
So, What Do We Know?

```r
> pData(eLeuRMA)
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_1.CEL
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_2.CEL
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_3.CEL
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_4.CEL
/home/laurent/Affymetrix_data/ecoli_sample//wt_1.CEL
/home/laurent/Affymetrix_data/ecoli_sample//wt_2.CEL
/home/laurent/Affymetrix_data/ecoli_sample//wt_3.CEL
/home/laurent/Affymetrix_data/ecoli_sample//wt_4.CEL
rownames(pData(eLeuRMA)) <-
    substr(rownames(pData(eLeuRMA)),45,56)
colnames(exprs(eLeuRMA)) <- rownames(pData(eLeuRMA))
```

That the data was loaded by Laurent Gautier...
Setting up the Linear Model

In order to use \texttt{limma}, we need three things: (1) an expression matrix, (2) a design matrix, and (3) a contrast matrix. The expression matrix we have. What about the other two? The design matrix basically states what treatments were applied to what samples.

\begin{verbatim}
> library(limma)
> designMatrix <-
    model.matrix(~pData(eLeuRMA$strain))
> # or
> # strain <- rep(c("lrp-","lrp+"),each=4)
> # design <- model.matrix(~factor(strain))
\end{verbatim}
What Does This Produce?

```r
> designMatrix
     (Intercept) pData(eLeuRMA)$strain lrp+
     1       1     0
     2       1     0
     3       1     0
     4       1     0
     5       1     1
     6       1     1
     7       1     1
     8       1     1

attr("assign") [1] 0 1
attr("contrasts")
attr("contrasts")$`pData(eLeuRMA)$strain` [1] "contr.treatment"
```
How Do We Fit Things?

colnames(designMatrix) <- c("lpr-","lpr+Diff")
fit1 <- lmFit(eLeuRMA, designMatrix)
fit2 <- eBayes(fit1)
summary(fit1)

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Class</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficients</td>
<td>14624</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>rank</td>
<td>1</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>assign</td>
<td>2</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>qr</td>
<td>5</td>
<td>qr</td>
<td>list</td>
</tr>
<tr>
<td>df.residual</td>
<td>7312</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>sigma</td>
<td>7312</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>cov.coefficients</td>
<td>4</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>stdev.unscaled</td>
<td>14624</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>pivot</td>
<td>2</td>
<td>-none-</td>
<td>numeric</td>
</tr>
</tbody>
</table>
How is the second invocation different? What gets added?
How Do We Fit Things? (2)

summary(fit2)

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Class</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>df.prior</td>
<td>1</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>s2.prior</td>
<td>1</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>var.prior</td>
<td>2</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>proportion</td>
<td>1</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>s2.post</td>
<td>7312</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>t</td>
<td>14624</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>p.value</td>
<td>14624</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>lods</td>
<td>14624</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>F</td>
<td>7312</td>
<td>-none-</td>
<td>numeric</td>
</tr>
<tr>
<td>F.p.value</td>
<td>7312</td>
<td>-none-</td>
<td>numeric</td>
</tr>
</tbody>
</table>
How Do We Display Things?

```r
options(digits=2)
topTable(fit2,coef=2,n=5,adjust="BH")

<table>
<thead>
<tr>
<th>ID</th>
<th>logFC</th>
<th>AveExpr</th>
<th>t</th>
<th>P.Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IG_821_1300838</td>
<td>-3.3</td>
<td>12.4</td>
<td>-23</td>
<td>7.2e-09</td>
</tr>
<tr>
<td>_1300922_fwd_st</td>
<td>4282</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5365 serA_b2913_st</td>
<td>2.8</td>
<td>12.2</td>
<td>16</td>
<td>1.6e-07</td>
</tr>
<tr>
<td>1389 gltD_b3213_st</td>
<td>3.0</td>
<td>10.9</td>
<td>13</td>
<td>6.4e-07</td>
</tr>
<tr>
<td>4625 lrp_b0889_st</td>
<td>2.3</td>
<td>9.3</td>
<td>11</td>
<td>2.3e-06</td>
</tr>
<tr>
<td>1388 gltB_b3212_st</td>
<td>3.2</td>
<td>10.0</td>
<td>11</td>
<td>2.8e-06</td>
</tr>
</tbody>
</table>

adj.P.Val   B
5.3e-05     8.0
6.0e-04     6.6
```
Double-Checking

```r
> t(exprs(eLeuRMA)[c(4282,5365),])

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IG_821_1300838</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_1300922_fwd_st serA_b2913_st</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nolrp_1.CEL</td>
<td>13.872</td>
<td>10.403</td>
</tr>
<tr>
<td>nolrp_2.CEL</td>
<td>14.253</td>
<td>10.745</td>
</tr>
<tr>
<td>nolrp_3.CEL</td>
<td>14.136</td>
<td>10.984</td>
</tr>
<tr>
<td>nolrp_4.CEL</td>
<td>13.811</td>
<td>11.195</td>
</tr>
<tr>
<td>wt_1.CEL</td>
<td>10.504</td>
<td>13.561</td>
</tr>
<tr>
<td>wt_2.CEL</td>
<td>10.960</td>
<td>13.739</td>
</tr>
<tr>
<td>wt_3.CEL</td>
<td>10.637</td>
<td>13.415</td>
</tr>
<tr>
<td>wt_4.CEL</td>
<td>10.699</td>
<td>13.722</td>
</tr>
</tbody>
</table>
```

That’s most of what there is here.
Example 2: Two Factors

Here, we look at changes over time in MCF7 in response to exposure to estrogen. This involves 8 U95Av2 arrays in the BioConductor package \texttt{estrogen}.

\begin{verbatim}
dataDir <- file.path(.find.package("estrogen"), "extdata")
targets <- readTargets("phenoData.txt", path=dataDir, sep=" ", row.names="filename")
\end{verbatim}
The Sample Info

targets

<table>
<thead>
<tr>
<th>filename</th>
<th>estrogen</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>low10-1.cel</td>
<td>absent</td>
<td>10</td>
</tr>
<tr>
<td>low10-2.cel</td>
<td>absent</td>
<td>10</td>
</tr>
<tr>
<td>high10-1.cel</td>
<td>present</td>
<td>10</td>
</tr>
<tr>
<td>high10-2.cel</td>
<td>present</td>
<td>10</td>
</tr>
<tr>
<td>low48-1.cel</td>
<td>absent</td>
<td>48</td>
</tr>
<tr>
<td>low48-2.cel</td>
<td>absent</td>
<td>48</td>
</tr>
<tr>
<td>high48-1.cel</td>
<td>present</td>
<td>48</td>
</tr>
<tr>
<td>high48-2.cel</td>
<td>present</td>
<td>48</td>
</tr>
</tbody>
</table>
Getting Expression Values

library(hgu95av2cdf)
estRMA <- justRMA(celfile.path=dataDir)
dim(estRMA)
Features    Samples
  12625       9

colnames(exprs(estRMA))
[1]  "bad.cel"  "high10-1.cel"  "high10-2.cel"
[4]  "high48-1.cel"  "high48-2.cel"  "low10-1.cel"
[7]  "low10-2.cel"  "low48-1.cel"  "low48-2.cel"
estRMA2 <- estRMA[,c(2:9)]
estRMA <- justRMA(filenames=targets$filename,
celfile.path=dataDir)
Building a Design Matrix

treatmentCombos <- factor(rep(1:4,each=2),
    labels=c("e-10h","e+10h","e-48h","e+48h"))
contrasts(treatmentCombos)
    e+10h e-48h e+48h
e-10h  0    0    0
e+10h  1    0    0
e-48h   0    1    0
e+48h   0    0    1

contrasts(treatmentCombos) <- cbind(
    Time=c(0,0,1,1), E10=c(0,1,0,0),
    E48=c(0,0,0,1))
designMatrix <- model.matrix(˜treatmentCombos)
What Does the Design Matrix Look Like?

colnames(designMatrix) <-
  c("Intercept","Time","E10","E48")

designMatrix

<table>
<thead>
<tr>
<th>Intercept</th>
<th>Time</th>
<th>E10</th>
<th>E48</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Properly expanded to cover all samples...
Fit the Model(s)

```r
fit1 <- lmFit(estRMA, designMatrix)
fit1EB <- eBayes(fit1)

otherContrasts <- cbind(E10=c(0,0,1,0),
                        E48=c(0,0,0,1))
# note this is with respect to the design!

fit2 <- contrasts.fit(fit1, otherContrasts)
fit2EB <- eBayes(fit2)
```
What’s the Difference?

> summary(classifyTestsF(fit1EB, p.value=0.0001))

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Time</th>
<th>E10</th>
<th>E48</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>136</td>
<td>55</td>
<td>181</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>11559</td>
<td>12065</td>
<td>11869</td>
</tr>
<tr>
<td>1</td>
<td>12625</td>
<td>930</td>
<td>505</td>
<td>575</td>
</tr>
</tbody>
</table>

> summary(classifyTestsF(fit2EB, p.value=0.0001))

<table>
<thead>
<tr>
<th></th>
<th>E10</th>
<th>E48</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>40</td>
<td>76</td>
</tr>
<tr>
<td>0</td>
<td>12469</td>
<td>12410</td>
</tr>
<tr>
<td>1</td>
<td>116</td>
<td>139</td>
</tr>
</tbody>
</table>

What’s the overall model being tested?
Tabling the Results

```r
mod2Results <- classifyTestsF(fit2EB,
    p.value=0.0001)

table(E10=mod2Results[,1], E48=mod2Results[,2])

<table>
<thead>
<tr>
<th>E48</th>
<th>E10</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>29</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>47</td>
<td>12370</td>
<td>52</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>29</td>
<td>87</td>
</tr>
</tbody>
</table>
```
Or, if you prefer (1)

Venn Up
Or, if you prefer (2)

Venn Down (Venn breaks for $> 3$ sets.)
TCGA: The Cancer Genome Atlas

http://cancergenome.nih.gov
What is it?

An attempt to do high-throughput studies right.

We’ve run a lot of high-throughput studies, but haven’t always learned as much as we’d hoped. Some common problems:

- small sample sizes
- variable sample quality
- poor clinical information
- batch effects
- looking just at one piece of the puzzle
- (poor experimental design)
A Big Science Pilot

time to think big (it worked for the genome project...)

$100M to start (actually up to a few now, but who’s counting?)

For a small number of tumor types, identify a large number of high-quality samples with good clinical information and some matched normal material. Some prospective collection may be required.

They picked 3 tumor types to start (now 20): brain (glioblastoma, GBM), lung (non-small cell), and ovary (serous adenocarcinoma).

For each, they’re seeking 500 samples, which will then be subjected to a barrage of assays.
The Assays (So Far)

- Sequencing of specific genes
- CGH Arrays (Agilent 244K)
- SNP Arrays (Affy 6/500K, Illumina 550K BeadArray)
- Expression Arrays (Affy U133+2, Agilent 44K)
- Exon Arrays (Affy)
- Methylation Arrays (Illumina)
- micro RNA (miRNA) Arrays (Agilent)
Where MDA Comes In

Biospecimen Sources

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom

GS01 0163: ANALYSIS OF MICROARRAY DATA
Where The Samples Go

TCGA Components

Biospecimen Core Resource
Where They Do Sequencing

Genome Sequencing Centers
Where They Run the Other Assays

Cancer Genome Characterization Centers
Where They Collect the Data

Data Coordinating Centers
Putting it All Together

TCGA Components

TCGA Map

© Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom
So, How’s It Going?

Progress 2009

Well, there’s good news and bad news...

Started with GBMs; samples from about 150 patients have been profiled.

They recently (late March) declared a data freeze to allow people to compare results at equal stages.

Concurrently, a “progress meeting” of sorts was held at the NCI. There’s an informative webcast available (http://cancergenome.nih.gov/media/workshops.asp).
News From the Front

They’d hoped to be further along.

Sample quality and access to corresponding normal material have been roadblocks.

The standards initially set (e.g., 80% tumor cells, less than 40% necrotic) may be unrealistic, and this may be worse with the other tumor types.

Shove more samples out to assays that may not require matched normal material (e.g., CGH).
More News From the Front

Sequencing and CGH are showing some successes due to sample size.

Gain of chr 7, loss of chr 10, several much more localized alterations.

The main known players (e.g., EGFR) are being found, and a few new ones are showing up as well (NF1, ERBB2). Clustering reveals 4 consistent subtypes.

Limited integration to date (one or two platforms); many studies involve results from other assays.
Progress 2010

GBMs now up to about 380.

Ovarian samples up to about 510.

Lung samples up to about 100, samples for about 10 other tissue types started.


Not too many biological shocks yet.

Much more sequencing data coming.
Where Can We Get the Data?

This has changed quite a bit over time.

http://tcga-data.nci.nih.gov/datarreports

is good to explore, in particular the latest archive:


You can also browse the publicly available data, which contains earlier releases of some of the data.
Controlled Access Data

So, what is “controlled”?

More flattery than is warranted...

Sequence data, SNP data, exon data, clinical data.

Faculty need to sign up to get the data.

We want it, but we need to restrict access to it if we use it here.
Things About the Data

There’s a *lot* of it.

Samples were sent out to the characterization centers in batches; roughly 30-45 patients per. The same batch went to each center (mostly).

Data (“raw” and processed) is grouped by Batch into gzipped tarballs, which can be 10s of gigabytes in size. This is why we see only a few files from the archive pulldown.

Descriptions of some of the processing applied can be found at [http://cancergenome.nih.gov/data/types/genomic:description](http://cancergenome.nih.gov/data/types/genomic:description)

Sample mappings are available at [http://tcga-data.nci.nih.gov/tcga/findArchives.htm](http://tcga-data.nci.nih.gov/tcga/findArchives.htm)
What Does the Data Look Like?

Pretty good, but not perfect.

We did a quick survey of some of the public TCGA data (mostly ovarian) over the past few weeks, identifying some areas where clarification could help.

In brief, these involve code, clinical data, annotation, basics, and bulletins.
What We Did

Divvied up data types:

miR (Keith*)
Clinical (Mary Edgerton)
Methylation (Wenyi Wang, Anna Unruh)
Illumina Transcriptome (Peng Qiu)
Affy Expression (Brad Broom)
Agilent Expression (Nianxiang Zhang)

Other (Rehan Akbani, John Weinstein, Chad Creighton, Li Zhang)

Data acquired through “Browse Public Data” (not Search).
Mimicking Cross-Level Processing is Hard

L2 to L3: log2-transform, average, center. *Forgot one.*
Level 2 Fits: Are We OK?

Misfit Between Predicted and Observed Level 2 Values

Log2(Sum of Absolute Deviations)
Checking Some Misfits...

Correlations > 0.999 For 6 Odd Samples
Level 2 Illumina Transcriptome

Predicting Level 2 Illumina Data From Level 1

Overlap as Fraction of Predicted

0.0 0.2 0.4 0.6 0.8 ... 1.0

Note: this uses our reconstruction, not sdrf.
Clinical Definition Varies by Site
Clinical Entries can be Idiosyncratic

One patient with “null” VitalStatus but DaysToDeath

PrimaryTherapyOutcomeSuccess vs TumorResidualDisease

<table>
<thead>
<tr>
<th></th>
<th>&gt;20mm</th>
<th>11–20mm</th>
<th>1–10mm</th>
<th>NoMacDis</th>
<th>null</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMP</td>
<td>35</td>
<td>10</td>
<td>136</td>
<td>75</td>
<td>34</td>
</tr>
<tr>
<td>PART</td>
<td>17</td>
<td>5</td>
<td>32</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>STAB</td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>PROG</td>
<td>10</td>
<td>5</td>
<td>16</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>null</td>
<td>21</td>
<td>6</td>
<td>41</td>
<td>15</td>
<td>31</td>
</tr>
</tbody>
</table>

PFS vs RFS?

Do we have CA125?
## Drug Names and Classes Vary

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pacilatxel</td>
<td>1</td>
</tr>
<tr>
<td>Paciltaxal</td>
<td>1</td>
</tr>
<tr>
<td>Paciltaxel</td>
<td>50</td>
</tr>
<tr>
<td>Paciltaxle</td>
<td>2</td>
</tr>
<tr>
<td>Pacitaxel</td>
<td>1</td>
</tr>
<tr>
<td>Pacliatxel</td>
<td>1</td>
</tr>
<tr>
<td>paclitaxel</td>
<td>4</td>
</tr>
<tr>
<td>Paclitaxel</td>
<td>193</td>
</tr>
<tr>
<td>Paclitaxel; Albumin-Bount</td>
<td>1</td>
</tr>
<tr>
<td>Paclitaxil</td>
<td>1</td>
</tr>
<tr>
<td>taxol</td>
<td>5</td>
</tr>
<tr>
<td>Taxol</td>
<td>346</td>
</tr>
</tbody>
</table>

### Targeted/Hormonal
What Are Genes Called?

Two annotation files, and Excel.
Interesting Things Within Batch

Can we get run date? Plate? Batch number?
Summary Recommendations

1. Post code for cross-level transitions.
2. Audit the clinical data.
3. Post explicit definitions for column headers.
4. Use common annotation within & across platforms.
5. Include run date/batch number/plate in L1 or SDRF.

Ask questions...