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

15 October 2009

Lecture 14: 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 ecolicdf).

```
> library(ecolicdf)
> library(ecoliLeucine) # loads affy, Biobase
> data(ecoliLeucine) # an AffyBatch
> eLeuRMA <- rma(ecoliLeucine)
```

So, What Do We Know?

```
> pData(eLeuRMA)
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_
/home/laurent/Affymetrix_data/ecoli_sample//nolrp_
/home/laurent/Affymetrix_data/ecoli_sample//wt_1.C
/home/laurent/Affymetrix_data/ecoli_sample//wt_2.C
/home/laurent/Affymetrix_data/ecoli_sample//wt_3.C
/home/laurent/Affymetrix_data/ecoli_sample//wt_4.C
rownames(pData(eLeuRMA)) <-
  substr(rownames(pData(eLeuRMA)),45,56)
colnames(exprs(eLeuRMA)) <- rownames(pData(eLeuRM
```

That the data was loaded by Laurent Gautier...

Setting up the Linear Model

In order to use `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.

```
> library(limma)
> designMatrix <-
    model.matrix(~pData(eLeuRMA$strain))
> # or
> # strain <- rep(c("lrp-","lrp+"),each=4)
> # design <- model.matrix(~factor(strain))
```

What Does This Produce?

```
> designMatrix
  (Intercept) pData(eLeuRMA)$strainlrp+
1           1                        0
2           1                        0
3           1                        0
4           1                        0
5           1                        1
6           1                        1
7           1                        1
8           1                        1
attr(,"assign") [1] 0 1
attr(,"contrasts")
attr(,"contrasts")$`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)
                  Length Class      Mode
coefficients     14624  -none-     numeric
rank                 1  -none-     numeric
assign               2  -none-     numeric
qr                   5  qr         list
df.residual       7312  -none-     numeric
sigma             7312  -none-     numeric
cov.coefficients     4  -none-     numeric
stdev.unscaled   14624  -none-     numeric
pivot                2  -none-     numeric
```

```
genes                1  data.frame list
Amean             7312  -none-     numeric
method               1  -none-     character
design              16  -none-     numeric
```

How is the second invocation different? What gets added?

How Do We Fit Things? (2)

```
summary(fit2)
                  Length Class      Mode
(as with fit1, plus)
df.prior               1 -none-     numeric
s2.prior               1 -none-     numeric
var.prior              2 -none-     numeric
proportion             1 -none-     numeric
s2.post             7312 -none-     numeric
t                  14624 -none-     numeric
p.value            14624 -none-     numeric
lods               14624 -none-     numeric
F                   7312 -none-     numeric
F.p.value           7312 -none-     numeric
```

How Do We Display Things?

```
options(digits=2)
topTable(fit2,coef=2,n=5,adjust="BH")
                    ID logFC AveExpr   t P.Value
      IG_821_1300838
4282 _1300922_fwd_st  -3.3    12.4 -23 7.2e-09
5365   serA_b2913_st   2.8    12.2  16 1.6e-07
1389   gltD_b3213_st   3.0    10.9  13 6.4e-07
4625    lrp_b0889_st   2.3     9.3  11 2.3e-06
1388   gltB_b3212_st   3.2    10.0  11 2.8e-06
   adj.P.Val   B
     5.3e-05 8.0
     6.0e-04 6.6
```

Double-Checking

```
> t(exprs(eLeuRMA)[c(4282,5365),])
                IG_821_1300838
                _1300922_fwd_st serA_b2913_st
nolrp_1.CEL              13.872         10.403
nolrp_2.CEL              14.253         10.745
nolrp_3.CEL              14.136         10.984
nolrp_4.CEL              13.811         11.195
wt_1.CEL                 10.504         13.561
wt_2.CEL                 10.960         13.739
wt_3.CEL                 10.637         13.415
wt_4.CEL                 10.699         13.722
```

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 `estrogen`.

```
dataDir <- file.path(.find.package("estrogen"),
  "extdata")
targets <-
  readTargets("phenoData.txt", path=dataDir,
              sep="", row.names="filename")
```

The Sample Info

```
targets
                    filename estrogen time.h
low10-1.cel   low10-1.cel   absent     10
low10-2.cel   low10-2.cel   absent     10
high10-1.cel high10-1.cel  present     10
high10-2.cel high10-2.cel  present     10
low48-1.cel   low48-1.cel   absent     48
low48-2.cel   low48-2.cel   absent     48
high48-1.cel high48-1.cel  present     48
high48-2.cel high48-2.cel  present     48
```

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
  Intercept Time E10 E48
1         1    0   0   0
2         1    0   0   0
3         1    0   1   0
4         1    0   1   0
5         1    1   0   0
6         1    1   0   0
7         1    1   0   1
8         1    1   0   1
```

Properly expanded to cover all samples...

Fit the Model(s)

```
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))
   Intercept  Time   E10   E48
-1         0   136    55   181
0          0 11559 12065 11869
1      12625   930   505   575
> summary(classifyTestsF(fit2EB,
   p.value=0.0001))
     E10   E48
-1    40    76
0  12469 12410
1    116   139
```

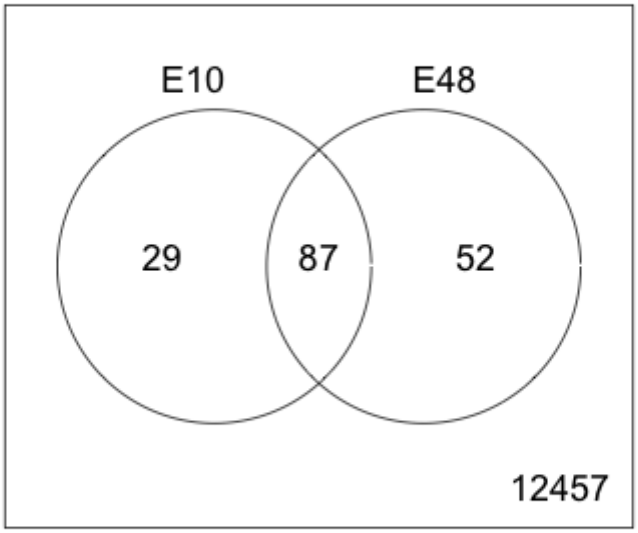
What's the overall model being tested?

Tabling the Results

```
mod2Results <- classifyTestsF(fit2EB,
   p.value=0.0001)
table(E10=mod2Results[,1], E48=mod2Results[,2])
    E48
E10     -1     0     1
  -1    29    11     0
  0     47 12370    52
  1      0    29    87
```

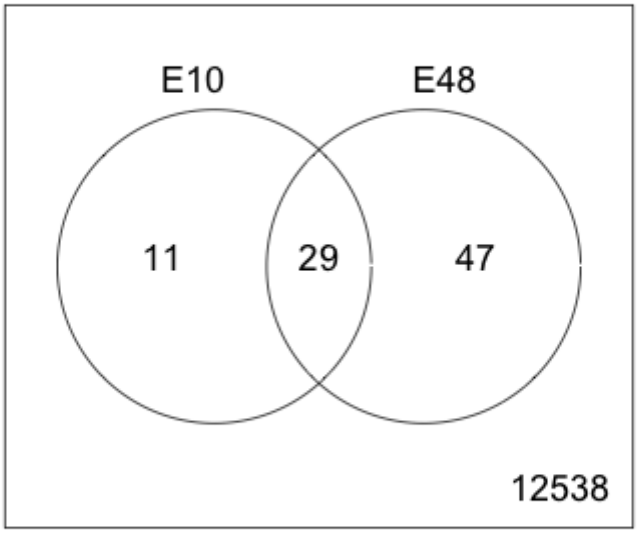
Or, if you prefer (1)

Venn Up

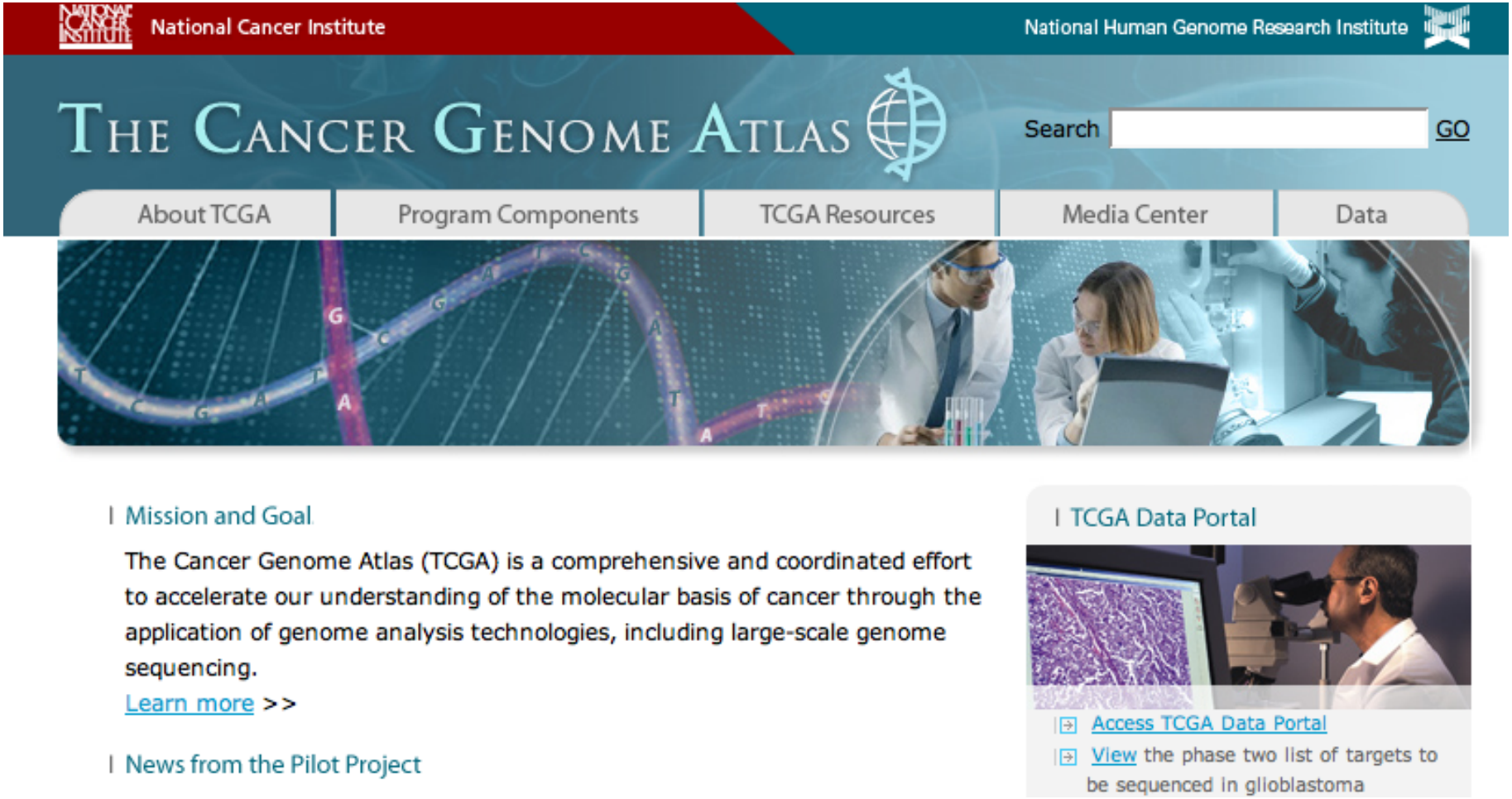


Or, if you prefer (2)

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

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: 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)

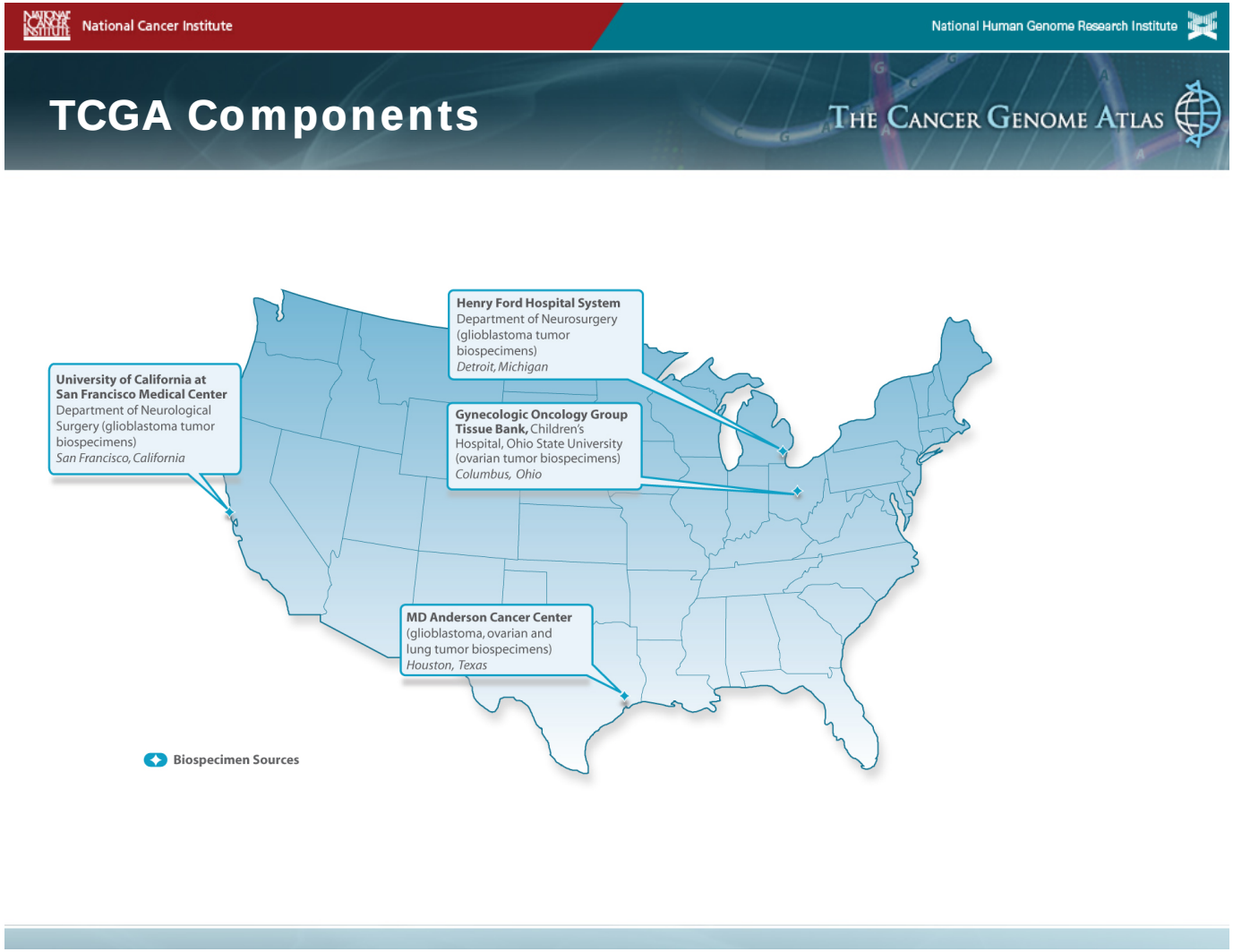
Cool Stuff

The data will all be made public*.

We may get to look at some of it early.

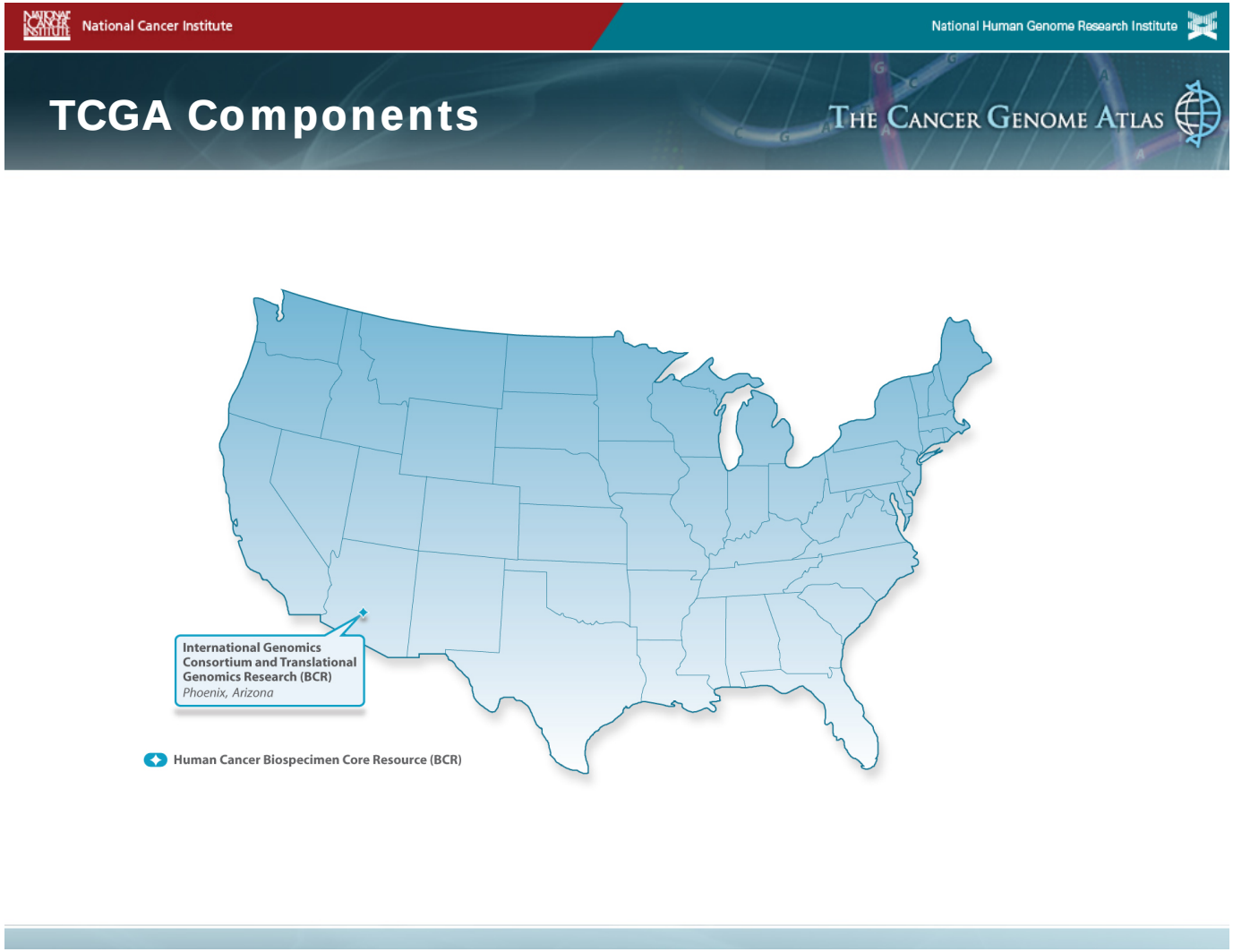
* we'll come back to this.

Where MDA Comes In



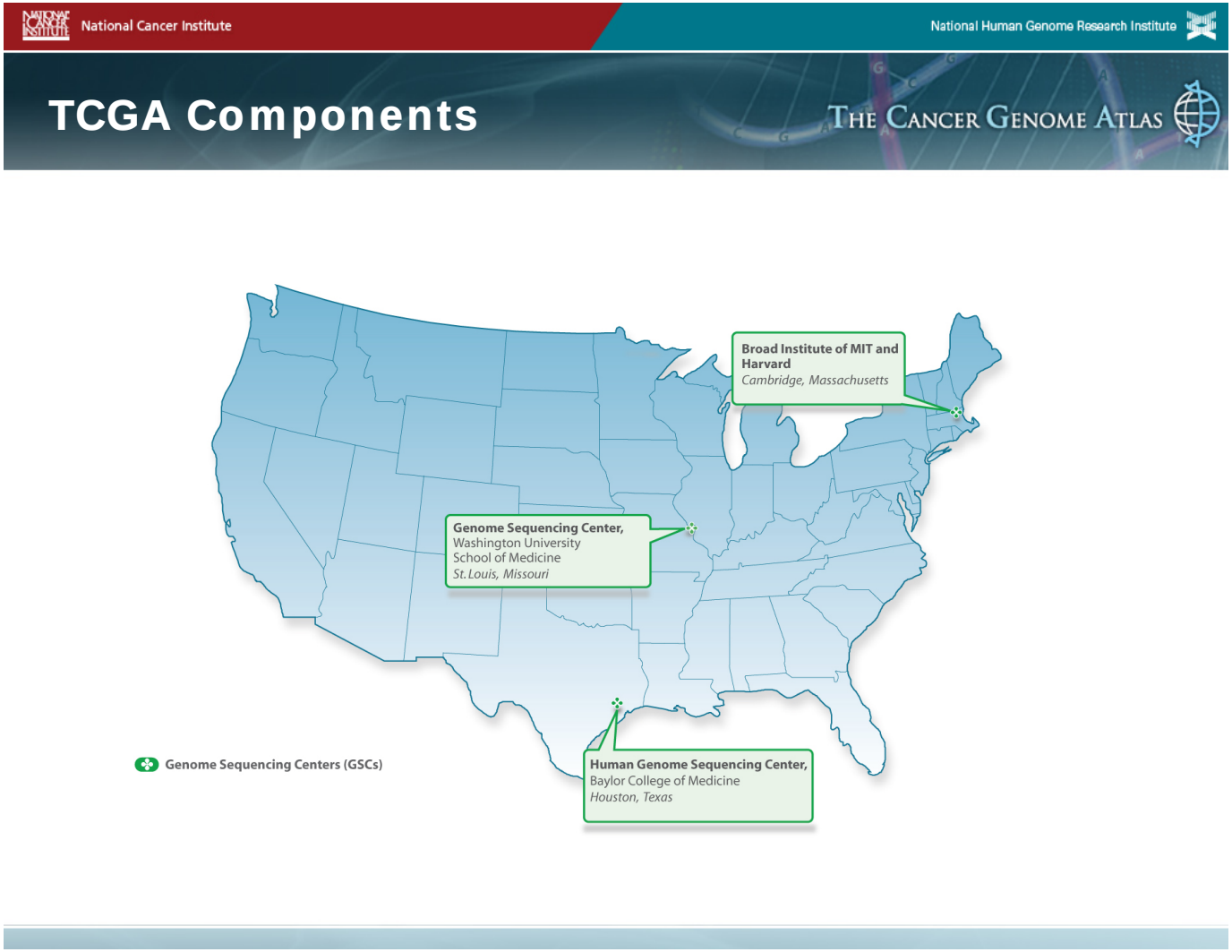
Biospecimen Sources

Where The Samples Go



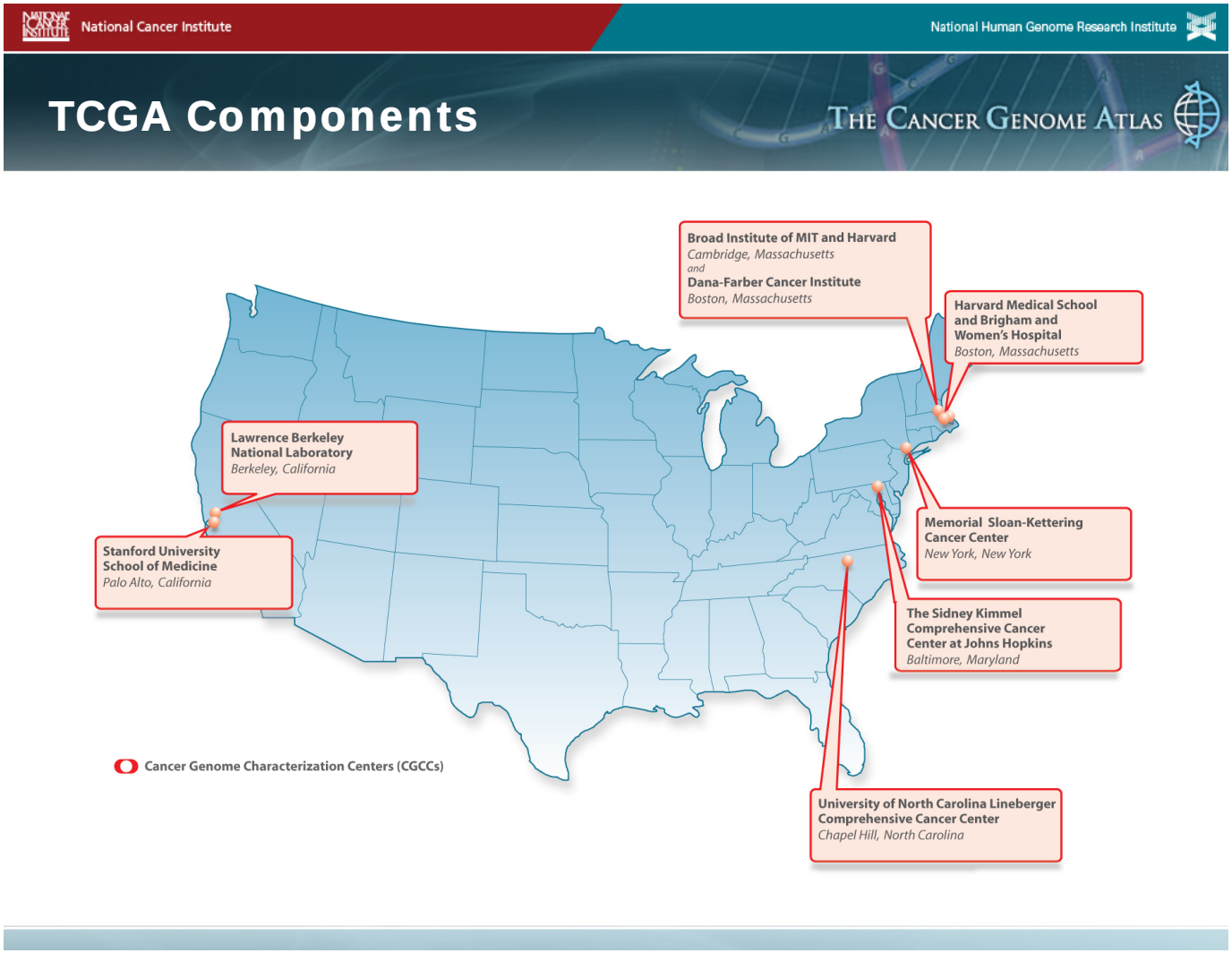
Biospecimen Core Resource

Where They Do Sequencing



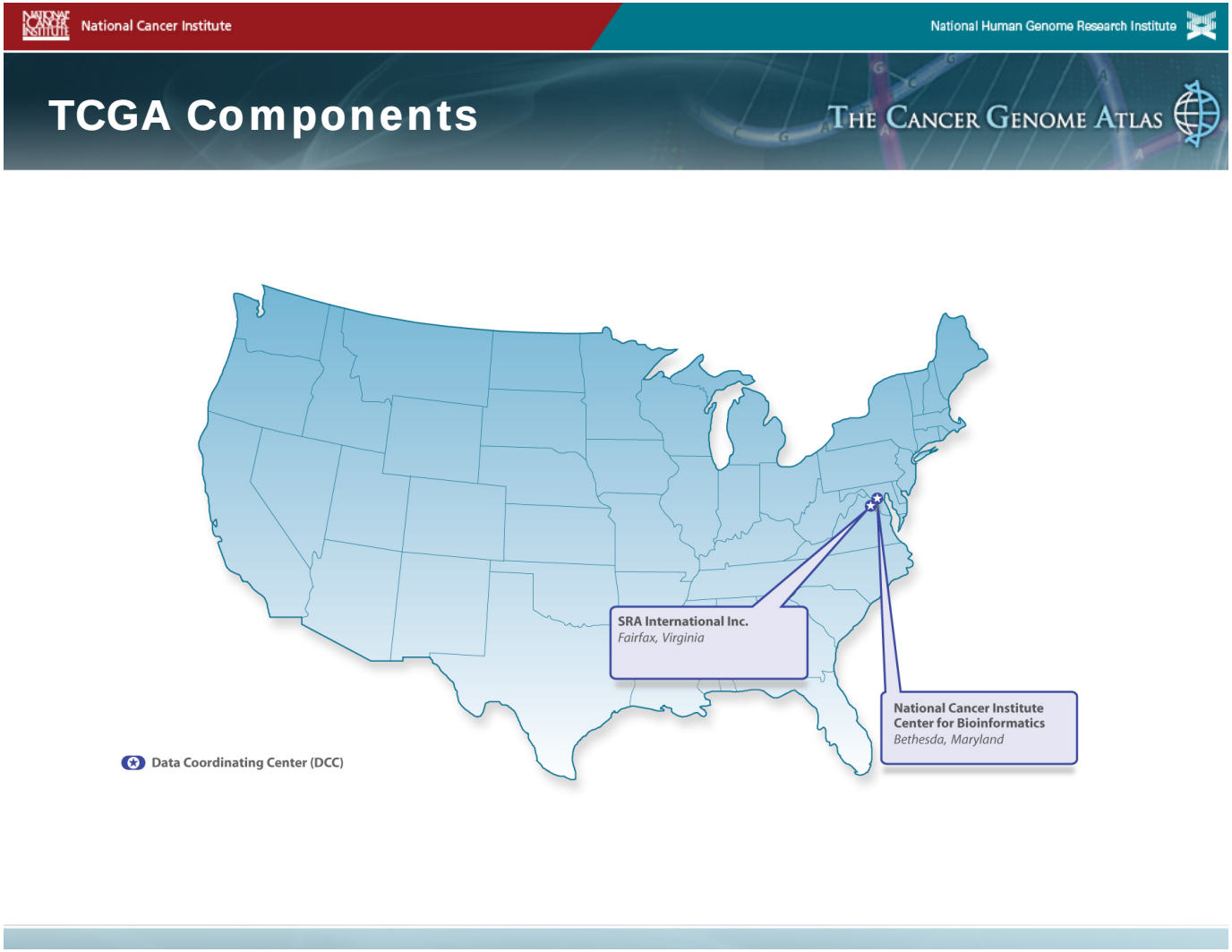
Genome Sequencing Centers

Where They Run the Other Assays



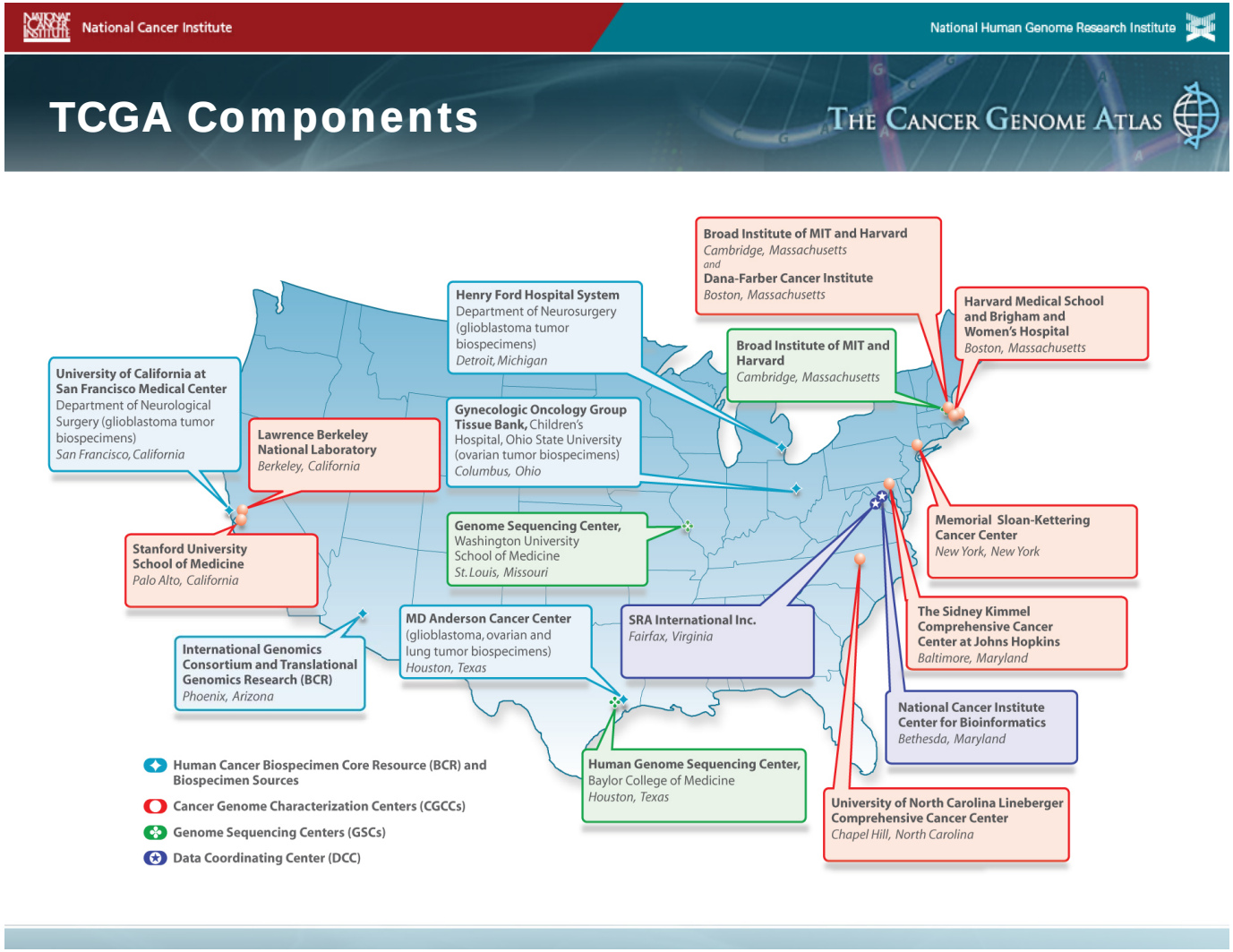
Cancer Genome Characterization Centers

Where They Collect the Data



Data Coordinating Centers

Putting it All Together



TCGA Map

So, How's It Going?

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

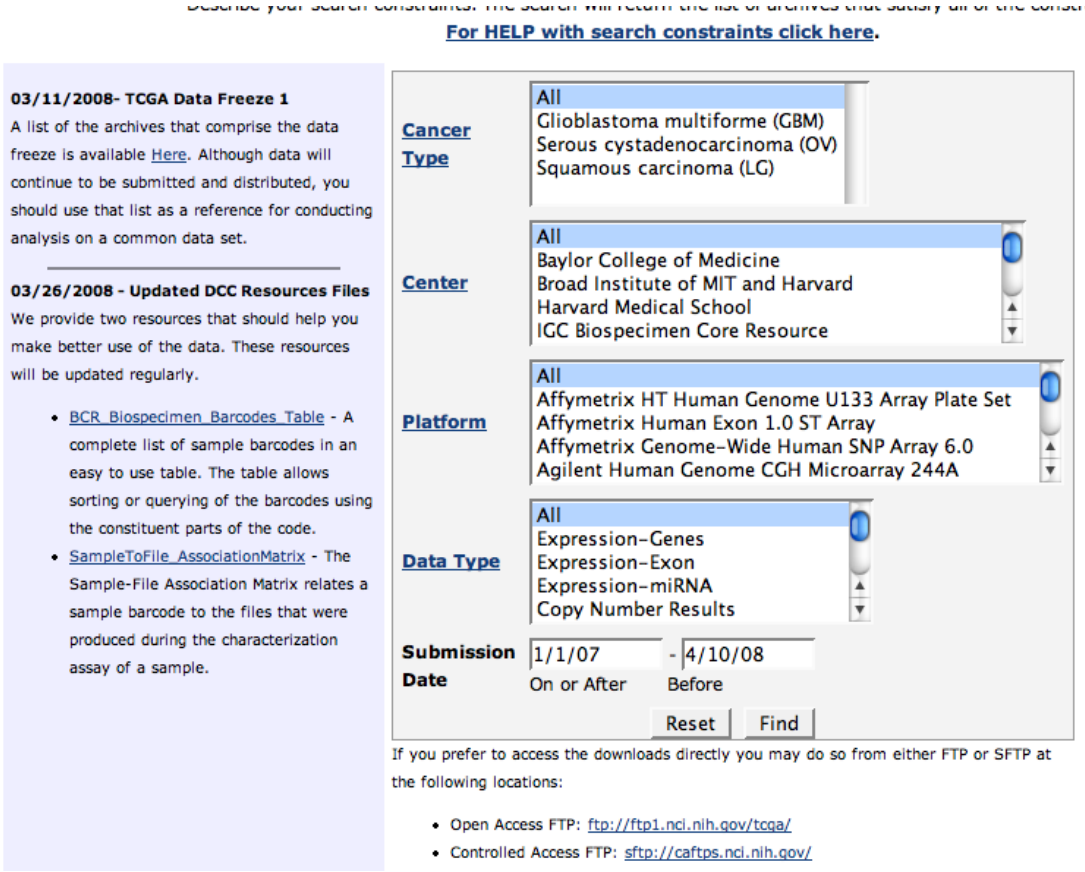
Sequencing and CGH are showing some successes due to sample size.

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.

Where Can We Get the Data?



`http://tcga-data.nci.nih.gov/tcga/findArchives.htm`

(don't try "Data Access" or "Data Portal" directly)

Other Ways to Get the Data

Open Access FTP:
`ftp://ftp1.nci.nih.gov/tcga/`

Controlled Access FTP:
`sftp:/caftps.nci.nih.gov/`

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 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 the processing applied can be found at `http://cancergenome.nih.gov/data/types/genomic/description`

Sample mappings are available at `http://tcga-data.nci.nih.gov/tcga/findArchives.htm`

What Questions Do We Want to Ask?

Where can we make a positive contribution? Most likely by combining results from multiple assays.

A Starting Example:

Can we exploit data of similar types?

Start with the DNA-level assays: Sequencing, CGH, SNPs

Focus on the question of copy number.

What's been done?

The CGH Data

Open access – Agilent 244K arrays.

Both Harvard and MSKCC ran the same samples.

Do the results agree? Are there reasons that they might not?

Harvard: Reports 9 batches. 35 arrays in batch 1.

MSKCC reports 7 batches. 26 arrays in batch 1.

What Was Run? SDRF files

Harvard's first 3:

```
Labeled Extract Name Label
TCGA-02-0001-01C-01D-0185-02 Cy3
Promega ref DNA Cy5
TCGA-02-0002-01A-01D-0185-02 Cy3
Promega ref DNA Cy5
TCGA-02-0002-10A-01D-0185-02 Cy3
Promega ref DNA Cy5
```

What Was Run? SDRF files (2)

MSKCC's first 3:

```
Labeled Extract Name Label
Ex_TCGA-02-0001-01C-01D-0183-04 Cy5
Ex_Promega ref DNA Cy3
Ex_TCGA-02-0002-01A-01D-0183-04 Cy5
Ex_Promega ref DNA Cy3
Ex_TCGA-02-0003-01A-01D-0183-04 Cy5
Ex_TCGA-02-0003-10A-01D-0183-04 Cy3
```

Some Notes

Raw quantification data is 180M/array.

MSKCC didn't get everything Harvard got.

When paired tumor/normal samples were available, Harvard ran them on two arrays, MSKCC on one.

At Harvard, the experimental sample was always Cy3 for the first batch.

At MSKCC, the tumor samples were always in Cy5 for the first batch.

Some Broader Notes

Both analyses proceed by segmentation first, followed by cross-sample comparison.

For both, raw data uses one set of annotation, but processed data uses another.

The format of some annotation files is not consistent across batch.

Early batches use straight tumor (01) and normal (10) material. Later batches use whole-genome amplified normal material as well (11).

First few batches are all from MD Anderson, later ones from Henry Ford.

What Can We Get With SNPs As Well?

Better localization of changepoints (requires using a common set of annotation)

Assessment of consistency/dye bias

LOH, genotype, haplotype info

What Questions Do We Want to Ask?

Can we find eQTL data? Correlations between genotype (from SNPs) with expression (from expression arrays)?

Can we infer regulation? Correlation between microRNA levels and expression levels?

Can we track translocations or chromosomal rearrangements? Correlation of copy number changes across the genome, odd sequencing behavior, SKY? Gene fusions?

For genes of interest, can we track proportions of inactivation by mechanism (copy number loss, expression loss, methylation)?

Coupling changes to outcome?

Things to Keep in Mind

Folks at MDA have already been working on pieces of this – see the tools available from Jonas' Integrative Biology Laboratory.

Some new assays will likely be added and/or updated.

There's other data out there for checking, or using in a training/testing context.

We can try out other integration techniques using data available in house (e.g., the NCI60 cell lines).

Logistics

Who wants in?

What level of effort?

How much do we need to do something right?

What Questions Do We Want to Ask?

Your turn now...