About the Homework Data

Keith Baggerly

We want you to work with some real data, so we tried to pull some together.

In particular, we've tried to collect all of the Level 3 (public) data we could find for some assay types (RNA-Seq, SNP6 Copy Number, and Methylation) from TCGA pertaining to the 25 most frequently mutated genes (as indicated by the TumorPortal).

Since there's a lot of data here, we wanted to describe it in a bit more detail.

These are measurements from about 10K patient samples, across many tumor types.

Not all samples were measured with all assays, so the numbers of measurements differ by assay.

The "primary key" for linking information across tables is the TCGA Barcode.

Let's take a look at how the barcode is used in the data matrices.

```r
load("homework.RData")

dim(rnaSeq2Data)
```

```
## [1]   25 9971
```

```r
dim(snp6Data)
```

```
## [1]    25 22453
```

```r
dim(methyData)
```

```
## [1]   995 10011
```

```r
dim(clinicData)
```

```
## [1] 11054     6
```

```r
colnames(rnaSeq2Data)[1:3]
```

```
## [1] "TCGA-OR-A5J1-01A-11R-A29S-07" "TCGA-OR-A5J2-01A-11R-A29S-07"
## [3] "TCGA-OR-A5J3-01A-11R-A29S-07"
```

```r
colnames(snp6Data)[1:3]
```

```
## [1] "TCGA-OR-A5J1-01A-11D-A29H-01" "TCGA-OR-A5J1-10A-01D-A29K-01"
## [3] "TCGA-OR-A5J2-01A-11D-A29H-01"
```

```r
colnames(methyData)[1:3]
```

```
## [1] "TCGA-AV-A03D-20A-02D-A29J-05" "TCGA-OR-A5J1-01A-11D-A29J-05"
## [3] "TCGA-OR-A5J2-01A-11D-A29J-05"
```

```r
rownames(clinicData)[1:3]
```

```
## [1] "TCGA-02-0001" "TCGA-02-0003" "TCGA-02-0004"
```

The TCGA barcodes all start out quite long, but in some cases (e.g., the clinical data), only an abbreviated form is used. This is because the different dashed fields represent different types of information about the sample being profiled. Characters 1-4 are always “TCGA”. Characters 6-7 indicate which tissue source site (TSS) the sample came from - “02” is the code for MD Anderson. Characters 9-12 supply an anonymized patient identifier. Characters 14-15 specify the “type” of sample being profiled: “01” is “Primary Solid Tumor” (the vast bulk of the data), whereas “10” is “Blood Derived Normal”. Further fields and characters specify which aliquot of material was used, which PCR plate was used for archiving, and the like. Thus, in the names above, the names for the clinical data only go out to character 12, because that’s all that’s needed to uniquely identify the patient - information such as length of patient survival doesn’t change if it’s being linked to a tumor sample or a normal sample. The assay names are longer because the differences here do matter (at least out to character 15). The barcodes don’t specify which type of tumor (e.g., breast, ovarian) the patient had, so we’ve supplied SampleAnnotation matrices for each assay. Let’s check the first few entries of these for RNA-Seq and SNP6.

```r
dim(rnaSeq2SampleAnnotation)
```

```
## [1] 9971    6
```

```r
rnaSeq2SampleAnnotation[1:3,]
```

```
##                                                    sample      patient
## TCGA-OR-A5J1-01A-11R-A29S-07 TCGA-OR-A5J1-01A-11R-A29S-07 TCGA-OR-A5J1
## TCGA-OR-A5J2-01A-11R-A29S-07 TCGA-OR-A5J2-01A-11R-A29S-07 TCGA-OR-A5J2
## TCGA-OR-A5J3-01A-11R-A29S-07 TCGA-OR-A5J3-01A-11R-A29S-07 TCGA-OR-A5J3
##                              diseaseCode            diseaseString typeCode
## TCGA-OR-A5J1-01A-11R-A29S-07         ACC Adrenocortical carcinoma       01
## TCGA-OR-A5J2-01A-11R-A29S-07         ACC Adrenocortical carcinoma       01
## TCGA-OR-A5J3-01A-11R-A29S-07         ACC Adrenocortical carcinoma       01
##                                       typeString
## TCGA-OR-A5J1-01A-11R-A29S-07 Primary solid Tumor
## TCGA-OR-A5J2-01A-11R-A29S-07 Primary solid Tumor
## TCGA-OR-A5J3-01A-11R-A29S-07 Primary solid Tumor
```

```r
dim(snp6SampleAnnotation)
```

```
## [1] 22453     6
```

```r
snp6SampleAnnotation[1:3,]
```

```
##                                                    sample      patient
## TCGA-OR-A5J1-01A-11D-A29H-01 TCGA-OR-A5J1-01A-11D-A29H-01 TCGA-OR-A5J1
## TCGA-OR-A5J1-10A-01D-A29K-01 TCGA-OR-A5J1-10A-01D-A29K-01 TCGA-OR-A5J1
## TCGA-OR-A5J2-01A-11D-A29H-01 TCGA-OR-A5J2-01A-11D-A29H-01 TCGA-OR-A5J2
##                              diseaseCode            diseaseString typeCode
## TCGA-OR-A5J1-01A-11D-A29H-01         ACC Adrenocortical carcinoma       01
## TCGA-OR-A5J1-10A-01D-A29K-01         ACC Adrenocortical carcinoma       10
## TCGA-OR-A5J2-01A-11D-A29H-01         ACC Adrenocortical carcinoma       01
##                                        typeString
## TCGA-OR-A5J1-01A-11D-A29H-01  Primary solid Tumor
## TCGA-OR-A5J1-10A-01D-A29K-01 Blood Derived Normal
## TCGA-OR-A5J2-01A-11D-A29H-01  Primary solid Tumor
```

In many discussions of TCGA data, the "disease codes" (short abbreviations) are used, so you may hear me refer to "BRCA" when I mean "breast cancer". The annotation tables give the full verbal expansions for each, and similarly contain descriptions of the various types of samples that are seen.

Looking at the first few SNP6 entries, we see that the first two samples are from the same patient (the first 12 characters are identical), but are from different types of tissue (primary tumor as opposed to blood derived normal). This is because the "single nucleotide polymorphism" (SNP) assays are being used to measure the numbers of DNA copies of a gene we see in a typical cell, and this number normally doesn't vary with tissue type - we should have two copies of every gene on a non-sex chromosome. While assuming "two copies" is pretty safe most of the time, there are variations between healthy individuals in some cases, so for this assay a paired normal sample was run as a control that might better "factor out" these normal variants. SNP6 arrays were used to measure tumor/normal pairs for almost every patient, which is why we have data on about 20K samples as opposed to 10K. This type of pairing wasn't generally done with RNA or methylation, because levels of these analytes are *tissue specific*, and taking samples of some "normal" tissues (e.g., brain, lung, kidney) is problematic at best. That said, we do have paired "normal tissue" (code 11) material to accompany some primary tumor samples profiled with RNA-Seq and methylation assays. In most cases this is "adjacent normal" (it was physically next to the tumor, and had to be excised during surgery), so it may not be completely "normal", but it's what we have.

The numbers of measurements differ by assay.

For both the RNA-Seq and SNP6 data, we've supplied one measurement "per gene". We haven't done this with methylation because we really can't. Methylation is a process by which a methyl group (CH_3) becomes attached to a "CG" dinucleotide sequence somewhere in the vicinity of a gene, which can affect the ability of the transcriptional machinery to access and thus produce copies of the gene in question. However, the effect produced can depend on where in the gene the CG pair resides! Thus, we've tried to supply, for each gene, measurements of methylation at all of the CG sites which are "close enough" to the gene that they might have an effect on transcription. Just how close is "close enough" is a bit of a guess right now, so we shot for a few kilobases (kb) on either side of the full gene sequence. Because of the "many to one" mapping of methylation sites to genes, the sites are indexed by "probe Id", and we've supplied some annotation to clarify what we're working with.

```r
dim(methyGenomeAnnotation)
```

```
## [1] 995   6
```

```r
methyGenomeAnnotation[1:4,]
```

```
##      probeId fromGene chromosome strand bpLocation    geneStructure
## 1 cg00011350       NA         12      F   49444296      MLL2 | Body
## 2 cg00039463       NA         16      R    3931489 CREBBP | TSS1500
## 3 cg00059930       NA         13      R   48894382       RB1 | Body
## 4 cg00067720       NA          5      R   67521141 PIK3R1 | TSS1500
```

```r
temp <-
  methyGenomeAnnotation[
    grep("MLL2", methyGenomeAnnotation[, "geneStructure"]),]
dim(temp)
```

```
## [1] 19  6
```

```
temp <- temp[order(temp[, "bpLocation"]), ]

temp[, c("probeId", "chromosome", "bpLocation", "geneStructure")]
```

```
##        probeId chromosome bpLocation                  geneStructure
## 803 cg22153481         12   49412872 MLL2 | 3'UTR : PRKAG1 | TSS1500
## 424 cg11229610         12   49412926 MLL2 | 3'UTR : PRKAG1 | TSS1500
## 227 cg05581469         12   49413435 MLL2 | 3'UTR : PRKAG1 | TSS1500
## 107 cg02831219         12   49416516                    MLL2 | Body
## 280 cg07218487         12   49418559                    MLL2 | Body
## 921 cg25729807         12   49420108                    MLL2 | Body
## 590 cg15790839         12   49420118                    MLL2 | Body
## 319 cg08163578         12   49420291                    MLL2 | Body
## 375 cg09643371         12   49420459                    MLL2 | Body
## 704 cg19253410         12   49420669                    MLL2 | Body
## 405 cg10441828         12   49422292                    MLL2 | Body
## 313 cg08089780         12   49426543                    MLL2 | Body
## 787 cg21580220         12   49435106                    MLL2 | Body
## 848 cg23521919         12   49444157                    MLL2 | Body
## 1   cg00011350         12   49444296                    MLL2 | Body
## 19  cg00522588         12   49449063                MLL2 | 1stExon
## 471 cg13007988         12   49449136                 MLL2 | TSS200
## 948 cg26686975         12   49449791                MLL2 | TSS1500
## 793 cg21787386         12   49450126                MLL2 | TSS1500
```

The scales of measurement differ by assay.

RNA-Seq tries to count the number of mRNA copies of a gene present in a given sample by sequencing a bunch of short nucleotide segments; these counts are typically then scaled by the numbers of “million reads” checked for the entire sample and in some cases for the length of the gene itself in kilobases, producing the “reads per kilobase per megaread” or “RPKM” value. This can then often be further scaled or transformed for comparisons across samples. Here, we have log2 transformed the Level 3 data, so the values you see should go up to 20 or so at most. Strictly, we applied log2(x + k) using a very small value of k (typically a small quantile of the observed set of RPKM values). If the value’s negative, it’s too small to be trusted.

SNP6 measurements are generally scaled to show the log2 ratio between the number of copies seen and the number of copies a paired normal sample would be expected to produce; if paired normal material is not available, the “normal” value is estimated from a pool of normal profiles run by the instrument manufacturer for calibration purposes. The ratios reported are typcially attenuated from what you might expect, in that if there are 4 copies of a gene in a tumor sample and 2 copies in a normal sample, we might get a value of 0.7 instead of 1. Part of this has to due with “tissue contamination” - almost all tumor “samples” contain both tumor and normal cells.

Methylation measurements estimate the fraction of reads at a given site which are methylated (M) as opposed to unmethylated (U); the reported ratio M/(M+U) is generally called the “beta value”. Values below 0.1 are essentially unmethylated (“hypomethylated”), and those above 0.9 are almost all methylated (“hypermethylated”). Methylation levels at sites which are physically very close to each other tend (unsurprisingly) to be positively correlated.

In terms of estimating survival, if only one of “days to death” or “days to last followup” is present, the patient is to be treated as dead or alive, respectively. If both entries are present and days to last followup exceeds days to death, well, we still have some cleaning to do (sorry!).

The gene names differ by assay.

Welcome to the world of genomic annotation. As we acquire more information, sometimes gene names are changed to reflect what we now know about their structure. If a large dataset is mid-construction, however, the old name may continue to be used to avoid breaking things. This means that it's important to know which annotation version was in place when various quantifications were assembled. When the genome is updated, this can also affect the base pair coordinates specifying where a gene is positioned within a chromosome. Here, just 2 of the 25 genes we're looking at have had their names change over the past few years: MLL2 and MLL3 are now known as KMT2D and KMT2C, respectively. A good place for disambiguation is GeneCards. The table "geneNamesByAssay" lists what names are used for indexing by what assays.

Some of the mutation data requires parsing.

The mutation data readily available from the TumorPortal specifies quite a bit of information about the alterations seen - the base pair coordinates, what the "reference" (healthy) allele is thought to be, and what the "new base" is. For our purposes, the main column of interest involves the "type" of mutation seen - we want to ignore "Silent" mutations (changes in the DNA nucleotide sequence that don't produce a change in the amino acid sequence of the protein produced).

Also, be aware - some of the mutation data is *not* from TCGA (though the vast majority is), and as a result, the names for these samples won't be TCGA barcodes! Right now, the TumorPortal hosts mutation data for about 4500 samples (accurately calling mutations takes more time than running some of the other assays). More recent data is being assembled as part of iCoMut - warning, this is in beta!

Some tumor types can be further subdivided.

When we look for patterns of alteration and try to figure out what's going on, it's typically important to treat different tissues as distinct - the differences in context can mean the effects of a treatment will not transfer that well. Similarly, we try to identify subtypes within a tissue type that might benefit from different drugs.

Breast cancer (BRCA) is a prime example of this. In most breast cancers, the estrogen receptor ESR1 is strongly overexpressed, making the tumor hyperreactive to growth signals estrogen provides. These tumors are said to be "ER-positive" (ER+). Fortunately, we have drugs such as tamoxifen and aromatase inhibitors that can bind to the excess receptors and slow or stop the excess growth. Similarly, other breast cancers undergo amplification at the DNA level of the gene for another receptor, HER2, aka ERBB2, which also leads to overproduction of a receptor and hypersensitivity to growth signals. These tumors are said to be "HER2-positive". Again, a drug (Herceptin / trastuzumab) targeting these receptors exists. That said, tamoxifen won't be of use to women with ER- disease, and Herceptin may not work as well in HER2- disease (it may still have some effects, and this is being explored, but less). A similar case exists for the progesterone receptor (PR), but PR+ is almost always ER+ as well. At present, the toughest type of breast cancer to treat is "triple negative": ER-, PR-, and HER2-, because we don't have as clear an understanding of what drives these tumors and we don't have a drug.

If we want to identify the drivers (e.g. mutations) specific to triple negative disease, simply treating BRCA as one disease may muddy the waters. We've included information about known subtypes for three of the diseases in TCGA: BRCA (BRCAreceptorStatus), Sarcoma (SARC, SARChistologicalType), and Uterine (UCEC, UCEChistologicalType). Sarcoma has lots of subtypes, but only two have been profiled in large numbers: liposarcomas (lipo, or fat), and leiomyosarcomas (smooth muscle, most commonly seen on the uterus). Uterine cancer is often referred to as endometrial cancer, and the most common subtype of UCEC is "endometriod", or EEA. Most of the others are "serous", and have a worse prognosis.

There may be others, but that'll do to start…

Onwards!