

Introduction to Bioconductor

Martin Morgan
Fred Hutchinson Cancer Research Center

9-10 December, 2010

Bioconductor

- Project Motivation and Success

- Representative Packages (Microarrays)

Microarrays: Work Flows

- Pre-processing

- Quality Assessment

- Filtering

Resources

- Lab Activity

- References

Bioconductor: Analysis and Comprehension of High Throughput Genetic Data

Goal Help biologists understand their data

Focus

- ▶ Expression and other microarray; flow cytometry
- ▶ High-throughput sequencing

Themes

- ▶ Open source / open development
- ▶ Code reuse – statistics, visualization, domain-specific applications, e.g., *limma*
- ▶ Interoperability
- ▶ Reproducible – scripts, *vignettes*, packages

Success > 400 packages; very active mailing list; annual conferences (BioC2011, Seattle, July 27-29); courses;
...

Representative Packages (Microarrays)

Pre-processing *affy*, *oligo*, *lumi*, *beadarray*, *limma*, *genefilter*, ...

Machine learning *MLInterfaces*, *CMA*

Differential expression *limma*, ...

Gene set enrichment *topGO*, *GOstats*, *GSEABase*, ...

Annotation *AnnotationDbi*, 'chip', 'org' and *BSgenome* packages

'Domain-specific' *DNAcopy*, *snpMatrix*, ...

Work Flow: Expression Microarrays

Prior to analysis

- ▶ Biological experimental design – treatments, replication, etc.
- ▶ Microarray preparation – especially two-channel

Analysis

1. Pre-processing (normalization); quality assessment; exploratory analysis
2. Differential expression; machine learning (clustering and classification)
3. Annotation
4. Gene set enrichment analysis
5. ...

<http://bioconductor.org/workflows> for common analyses.

Example Data

Chiaretti et al., 2005 [1]

- ▶ 128 adult patients, newly diagnosed for ALL
- ▶ B- and T-lineage; various molecular and cytological characteristics.
- ▶ HG-U95Av2
- ▶ Pre-processed (background correction, normalization, summarization into probe sets).

The ALL dataset

```
> library(ALL); data(ALL); ALL
```

```
ExpressionSet (storageMode: lockedEnvironment)
```

```
assayData: 12625 features, 128 samples
```

```
  element names: exprs
```

```
protocolData: none
```

```
phenoData
```

```
  sampleNames: 01005 01010 ... LAL4
```

```
    (128 total)
```

```
  varLabels: cod diagnosis ... date
```

```
    last seen (21 total)
```

```
  varMetadata: labelDescription
```

```
featureData: none
```

```
experimentData: use 'experimentData(object)'
```

```
  pubMedIds: 14684422 16243790
```

```
Annotation: hgu95av2
```

ExpressionSet

Example of an 'S4' class

- ▶ Coordinate different types of data (assay, phenotype, feature, experiment) into a single container
- ▶ Reduces clerical errors, enhances interoperability and reproducibility

Manipulate with *accessors* and *subsetting*

```
> m <- exprs(ALL)      # matrix of expression values  
> adf <- phenoData(ALL) # data frame-like sample descriptions  
> some <- ALL[,1:10]    # first 10 samples
```

Metadata, e.g., adf

- ▶ pData(adf): 1 row per sample, columns representing measured attributes, e.g., sex, age, 'fusion protein'.
- ▶ Data about the columns, e.g., varMetadata(adf)

Warning: different S4 classes have different conventions

Pre-processing

Background correction

- ▶ One-channel: PM / MM probes
- ▶ Two-channel: background vs. foreground intensities

Normalization

- ▶ Key assumption: most probe sets not differentially expressed; distribution of intensities approximately equal across arrays

Summarization

- ▶ One-channel: from probes to probesets (approximately, genes)

One channel Affymetrix 3' expression arrays

- ▶ In practice:

```
> ## assume phenoData is an AnnotatedDataFrame  
> ## "/celfile/directory" contains CEL files  
> setwd("/your/celfile/directory")  
> library(affy)  
> eset <- just.rma(phenoData=phenoData)
```

- ▶ Other normalizations, e.g., `just.gcrma`, `vsn2`; *affyPLM* for detailed probe models; *oligo* for recent arrays.
- ▶ *limma* for two-channel arrays.

Example: RMA (robust multi-chip average)

Background correction

- ▶ Observation: using MM probes is problematic when $MM > PM$.
- ▶ Model PM probes as exponentially distributed signal, plus normal noise, $\exp(\alpha) + N(\mu, \sigma^2)$.

Normalization

- ▶ Quantile normalization – force the *distribution* of background-corrected expression values of each array to have exactly the same.

Summarization

- ▶ Estimate probeset effect by fitting a linear model to all probes in each probe set, across array.

Quality assessment

- ▶ In practice:

```
> library(arrayQualityMetrics)
> rpt <- arrayQualityMetrics(eset)
> ## or, as appropriate,
> ##   rpt <- arrayQualityMetrics(abatch)
> ##   rpt <- arrayQualityMetrics(rg)
> browseURL(rpt)
```

- ▶ QC summary statistics: acceptable ranges for 'control' probes
- ▶ Between-array distances: no unintended association with experimental conditions, e.g., run date.
- ▶ NUSE (normalized unscaled standard error) and RLE (relative log expression) plots: consistent expression and variability across arrays.

Filtering probe sets

Use `nsFilter` from the *genefilter* package to filter out probes that:

- ▶ Lack variability
- ▶ Are without an Entrez Gene ID annotation
- ▶ Map to the same Entrez Gene ID
- ▶ Are not annotated to GO (or other) terms

```
> library(genefilter)
> filteredESet <- nsFilter(eset, require.entrez=TRUE,
+   remove.dupEntrez=TRUE, feature.exclude="^AFFX")
```

Lab activity

Goal: learn to work with S4 classes, especially *ExpressionSet*

1. Load and explore ALL object, including finding help on S4 objects.
2. Extract `mol.biol` `phenoData`, subset samples to include only BCR/ABL or NEG.
3. Perform quality assessment.
4. Filter (remove) probes without gene-level annotation

References



S. Chiaretti, X. Li, R. Gentleman, A. Vitale, K. S. Wang, F. Mandelli, R. Foa, and J. Ritz.

Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation.

Clin. Cancer Res., 11:7209–7219, Oct 2005.

sessionInfo

- ▶ R version 2.12.0 Patched (2010-11-28 r53696),
i386-pc-mingw32
- ▶ Locale: LC_COLLATE=C,
LC_CTYPE=English_United States.1252,
LC_MONETARY=English_United States.1252,
LC_NUMERIC=C, LC_TIME=English_United States.1252
- ▶ Base packages: base, datasets, grDevices, graphics, methods,
stats, utils
- ▶ Other packages: ALL 1.4.7, AnnotationDbi 1.12.0,
Biobase 2.10.0, DBI 0.2-5, GO.db 2.4.5, RSQLite 0.9-4,
SeattleIntro2010 0.0.41, biomaRt 2.6.0, genefilter 1.32.0,
hgu95av2.db 2.4.5, org.Hs.eg.db 2.4.6
- ▶ Loaded via a namespace (and not attached): RCurl 1.5-0.1,
XML 3.2-0.2, annotate 1.28.0, splines 2.12.0, survival 2.36-2,
tools 2.12.0, xtable 1.5-6