

# The *Bioconductor* Project

Martin Morgan  
Fred Hutchinson Cancer Research Center

19-21 January, 2011

# 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;  
...

# The *Bioconductor* Web Site

- ▶ Finding and installing packages
- ▶ Work flows
- ▶ Finding help – in and outside *R*
- ▶ The *Bioconductor* release schedule
- ▶ Developer support
- ▶ Courses and conferences

# 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; systems biology
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
```

# 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*, ...

# 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. 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.