

The Bioconductor Project for Reproducible Analysis of High Throughput Genomic Data

Martin Morgan (mtmorgan@fhcrc.org)
Fred Hutchinson Cancer Research Center

19-21 January, 2011

Analysis and Comprehension of High Throughput Genomic Data

Hallmarks of effective computational software

1. Extensive: data, annotation
2. Statistical: volume, technology, experimental design
3. Reproducible: long-term, multi-participant science
4. Leading edge: novel, technology-driven
5. Accessible: affordable, transparent, usable

1. Extensive Data and Annotation

Data

- ▶ Expression, tiling, methylation, custom arrays.
- ▶ Sequence analysis, e.g., ChIP-, RNA-seq
- ▶ Other high-throughput assays, e.g., flow cytometry, mass spec., imaging
- ▶ Public repositories, e.g., GEO, ArrayExpress

Annotation, e.g.,

- ▶ Well-curated: NCBI, Biomart, UCSC, MsigDB, GO, KEGG
- ▶ Loosely curated: emerging, specialized, & lab-based
- ▶ Consortium: HapMap, 1000 genomes, TCGA

Bioconductor

Goal Help biologists understand their data

Focus

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

Themes

- ▶ Contributions from 'core' members and (primarily academic) user community
- ▶ Based on the *R* programming language – statistics, visualization, interoperability
- ▶ Reproducible – scripts, *vignettes*, packages
- ▶ Open source / open development

Success > 400 packages; publications; 8,000 web visits / week; 75,000 unique IP downloads / year; very active mailing list; annual conferences; courses; ...

Bioconductor: Sample Work Flow

```
> ##
> ## Pre-processing
> library(affy)
> eset <- just.rma()
> ##
> ## Quality assessment
> library(arrayQualityMetrics)
> arrayQualityMetrics(eset)
> ##
> ## Differential expression
> library(limma)
> status <-
+   c("Trt", "Trt", "Trt", "Ctrl", "Ctrl", "Ctrl")
> design <- model.matrix( ~status )
> fit <- eBayes(lmFit(eset, design))
> topTable(fit, coef=2)
```

2. Statistical

Technology

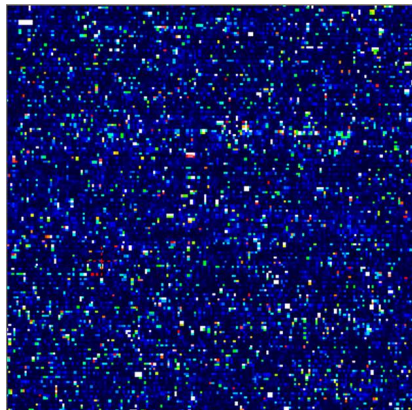
- ▶ Acknowledging artifacts and biases
- ▶ Accomodate using statistical models, e.g., RMA

Volume of data

- ▶ Data reduction essential

Experimental design

- ▶ Exploratory analysis
- ▶ Hypothesis-driven; designed experiments
- ▶ Cost-effective, but not too clever



Expression array. Pseudocolors represent hybridisation intensities of RNA to features. Source: url

Statistical

Technology

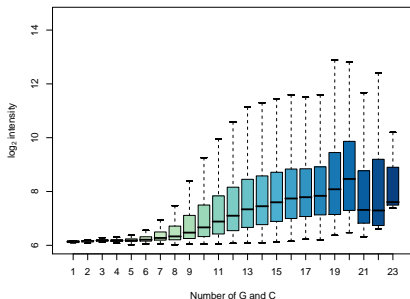
- ▶ Acknowledging artifacts and biases
- ▶ Accomodate using statistical models, e.g., RMA

Volume of data

- ▶ Data reduction essential

Experimental design

- ▶ Exploratory analysis
- ▶ Hypothesis-driven; designed experiments
- ▶ Cost-effective, but not too clever



Measured intensity increases with GC content; Chronic Lymphocytic Leukemia (CLL) dataset.

Statistical

Technology

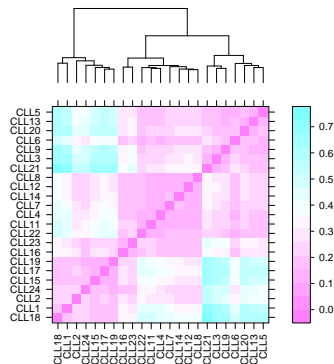
- ▶ Acknowledging artifacts and biases
- ▶ Accomodate using statistical models, e.g., RMA

Volume of data

- ▶ Data reduction essential

Experimental design

- ▶ Exploratory analysis
- ▶ Hypothesis-driven; designed experiments
- ▶ Cost-effective, but not too clever



Heatmap summarizing distance between CLL arrays

Statistical

Technology

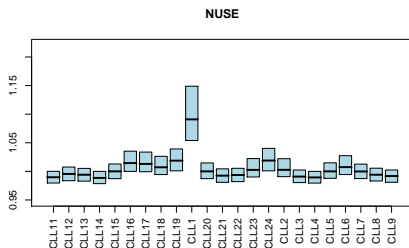
- ▶ Acknowledging artifacts and biases
- ▶ Accomodate using statistical models, e.g., RMA

Volume of data

- ▶ Data reduction essential

Experimental design

- ▶ Exploratory analysis
- ▶ Hypothesis-driven; designed experiments
- ▶ Cost-effective, but not too clever



Normalized unscaled standard error (NUSE) suggests array CLL1 is an outlier.

Statistical

Technology

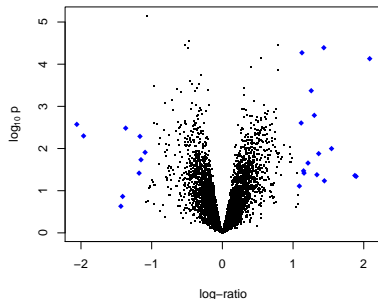
- ▶ Acknowledging artifacts and biases
- ▶ Accomodate using statistical models, e.g., RMA

Volume of data

- ▶ Data reduction essential

Experimental design

- ▶ Exploratory analysis
- ▶ Hypothesis-driven; designed experiments
- ▶ Cost-effective, but not too clever



‘Progressive’ vs. ‘stable’ status.
 $\log P$ vs. $\log\text{-fold change}$, CLL
data set. Probe sets with extreme
differentiation highlighted.

3. Reproducible Research

Long-term

- ▶ Returning to analysis after days, weeks, months of other activity

Multi-participant: communicating with...

- ▶ Other statisticians / bioinformaticians
- ▶ Biologists and others without specialized statistical knowledge

Science: reproducibility...

- ▶ Facilitates third-party verification
- ▶ Allows critical assessment
- ▶ Challenging, even in high-profile journals requiring archived raw data (Ioannidis *et al.*, 2009, Nat Genet 41: 149-155).

Reproducible Research: Case Study

Original research

- ▶ Potti *et al.*, 2006; Hsu *et al.*, 2007
- ▶ NCI60 cell line drug sensitivity signature
- ▶ Clinical trial allocation

Reproducibility

- ▶ Baggerly & Coombes, 2009
- ▶ Off-by-one cisplatin gene signature
- ▶ Four 'interesting' genes not supported by analysis (two not on array)

References

- ▶ Potti *et al.* 2006 Nat Med 12: 1294-1300; (retracted)
- ▶ Hsu *et al.* 2007 J Clin Oncol 25: 4350-4357. (retracted)
- ▶ Baggerly & Coombes 2009 Ann Appl Stat 3: 1309-1334

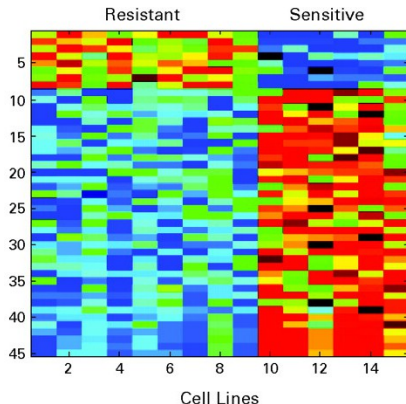
Reproducible Research: Case Study

Original research

- ▶ Potti *et al.*, 2006; Hsu *et al.*, 2007
- ▶ **NCI60 cell line drug sensitivity signature**
- ▶ Clinical trial allocation

Reproducibility

- ▶ Baggerly & Coombes, 2009
- ▶ Off-by-one cisplatin gene signature
- ▶ Four 'interesting' genes not supported by analysis (two not on array)



Hsu *et al.*, cisplatin, fig. 1a

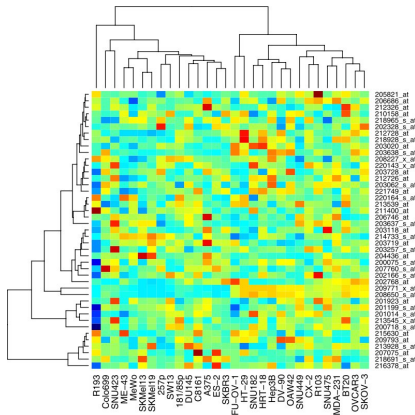
Reproducible Research: Case Study

Original research

- ▶ Potti *et al.*, 2006; Hsu *et al.*, 2007
- ▶ NCI60 cell line drug sensitivity signature
- ▶ Clinical trial allocation

Reproducibility

- ▶ Baggerly & Coombes, 2009
- ▶ **Off-by-one cisplatin gene signature**
- ▶ Four 'interesting' genes not supported by analysis (two not on array)



Baggerly & Coombes, fig. 2a

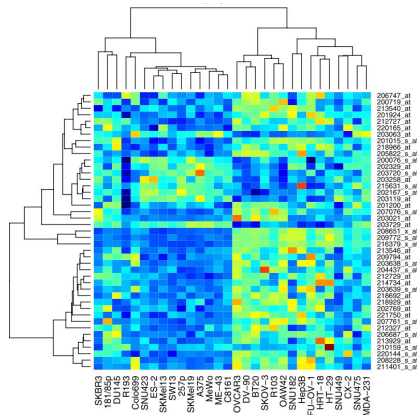
Reproducible Research: Case Study

Original research

- ▶ Potti *et al.*, 2006; Hsu *et al.*, 2007
- ▶ NCI60 cell line drug sensitivity signature
- ▶ Clinical trial allocation

Reproducibility

- ▶ Baggerly & Coombes, 2009
- ▶ **Off-by-one cisplatin gene signature**
- ▶ Four 'interesting' genes not supported by analysis (two not on array)



Baggerly & Coombes, fig. 2b

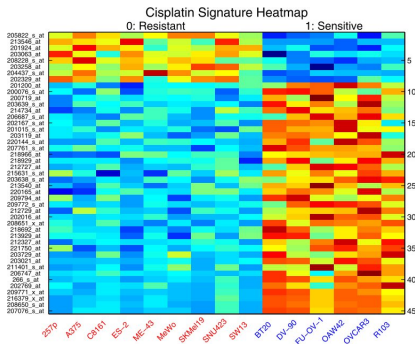
Reproducible Research: Case Study

Original research

- ▶ Potti *et al.*, 2006; Hsu *et al.*, 2007
- ▶ NCI60 cell line drug sensitivity signature
- ▶ Clinical trial allocation

Reproducibility

- ▶ Baggerly & Coombes, 2009
- ▶ **Off-by-one cisplatin gene signature**
- ▶ Four 'interesting' genes not supported by analysis (two not on array)



Baggerly & Coombes, fig. 2d

Reproducible Research: Case Study

Original research

- ▶ Potti *et al.*, 2006; Hsu *et al.*, 2007
- ▶ NCI60 cell line drug sensitivity signature
- ▶ Clinical trial allocation

Reproducibility

- ▶ Baggerly & Coombes, 2009
- ▶ Off-by-one cisplatin gene signature
- ▶ Four 'interesting' genes not supported by analysis (two not on array)

... results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common – Baggerly & Coombes, 2009

Reproducible Research: *Bioconductor*

- Script-based Data transformations *necessarily* documented
- 'Literate programming' Text documents embed scripts, scripts *evaluated* when text document processed
- Versioned software and repositories Record which package versions used, and retrieve from *Bioconductor* archives
- Integrated data containers Sample descriptions and expression data in a single object. Subsetting expression data automatically subsets sample descriptions

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
```

4. Leading Edge

Technological innovations

- ▶ E.g., SNP, miRNA arrays
- ▶ E.g., lab sequencing platforms; novel protocols

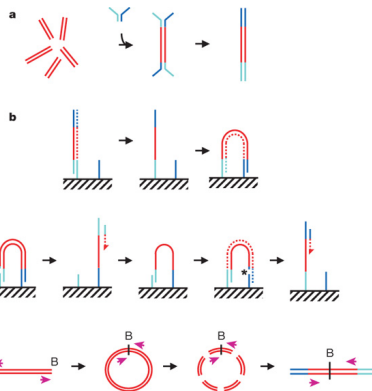
Fast-changing

- ▶ Commercial software products not yet developed, or already out-of-date
- ▶ Research questions require novel solutions

Leading Edge: Illustration

Sequencing technologies

- ▶ Historically (e.g., 2 years ago): short reads, low 'tail' quality, tail base call bias, data volume
- ▶ Current: count models, read bias, designed experiments, variant representations, annotation

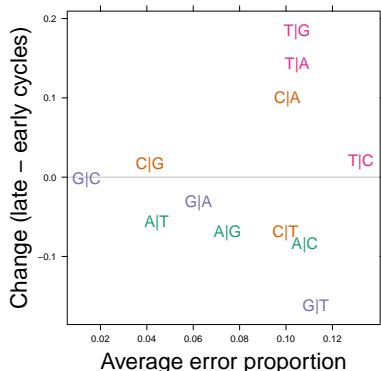


Bentley et al., 2008, Nature 456:
53-9

Leading Edge: Illustration

Sequencing technologies

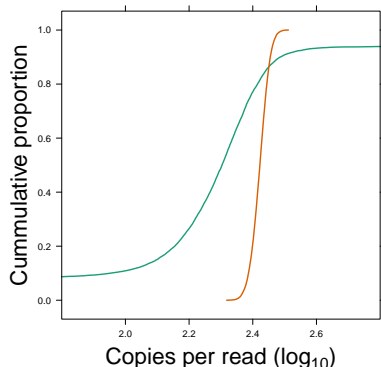
- ▶ **Historically** (e.g., 2 years ago): short reads, low 'tail' quality, tail base call bias, data volume
- ▶ Current: count models, read bias, designed experiments, variant representations, annotation



Leading Edge: Illustration

Sequencing technologies

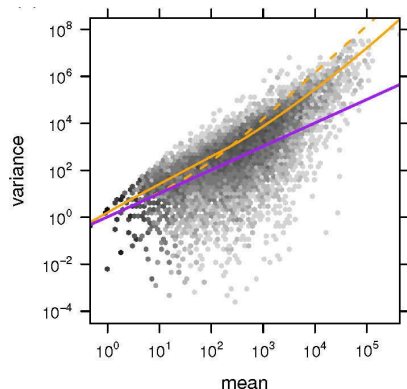
- ▶ **Historically** (e.g., 2 years ago): short reads, low 'tail' quality, tail base call bias, data volume
- ▶ **Current**: count models, read bias, designed experiments, variant representations, annotation



Leading Edge: Illustration

Sequencing technologies

- ▶ Historically (e.g., 2 years ago): short reads, low 'tail' quality, tail base call bias, data volume
- ▶ **Current:** count models, read bias, designed experiments, variant representations, annotation



Poisson (purple) and negative binomial (orange) fit to RNA-seq data. Anders & Huber, 2010, Genome Biol, 11:R106

5. Accessible

Affordable

- ▶ Purchase / licensing; time

Transparent

- ▶ Algorithms, e.g., RMA
- ▶ Code reuse

Challenges and solutions

- ▶ Research questions requiring 'one-off' solutions
- ▶ Software bugs

Usable

- ▶ Documentation
- ▶ Training, such as today!

Accessible

Affordable

- ▶ Purchase / licensing; time

Transparent

- ▶ Algorithms, e.g., RMA
- ▶ Code reuse

Challenges and solutions

- ▶ Research questions requiring 'one-off' solutions
- ▶ Software bugs

Usable

- ▶ Documentation
- ▶ Training, such as today!

Documentation

- ▶ Help pages
- ▶ Vignettes
- ▶ Archived course and conference material
- ▶ Mailing list

BioC2011

- ▶ Annual conference – user and scientific presentations, workshops, poster session
- ▶ Seattle July 27-29

Analysis and Comprehension of High Throughput Genomic Data

Hallmarks of effective computational software

1. Extensive: data, annotation
2. Statistical: volume, technology, experimental design
3. Reproducible: long-term, multi-participant science
4. Leading edge: novel, technology-driven
5. Accessible: affordable, transparent, usable

Acknowledgments

- ▶ Vince Carey (Brigham & Womens, Harvard), Wolfgang Huber (EBI), Rafael Irizzary (JHU), Robert Gentleman (Genentec)
- ▶ Hervé Pagès, Marc Carlson, Nishant Gopalakrishnan, Chao-Jen Wong, Dan Tenenbaum, Valerie Obenchain
- ▶ Patrick Aboyoun, Seth Falcon, Michael Lawrence, Deepayan Sarkar, Florian Hahne
- ▶ Sean Davis, James MacDonald