Softwareproject Bio-Medical Informatics 2 (SWPBMI2, WS2009/10)

Analyzing Microarray Data using Bioconductor (two ovarian studies)

Markus Neuner, Roman Landsteiner, Markus Hofer, Roland Kienast

University for Health Sciences, Medical Informatics and Technology (UMIT)

January 2009

Project description I

Goal

- get familiar with R, Bioconductor in general
- successfully use, normalize, analyze, and process microarray expression data to identify interesting genes and to classify samples based on their gene expression.

Documentation

- with LaTeX / Sweave under Debian-Linux and versioned with subversion
- mainly based on the book "Bioconductor Case Studies"

Project description II

Organization

- Team sessions in short intervals (usually one week)
- Each session a goal was defined (e.g. read chapter X until next session, collect information on a topic) and during the session the R-code was created and documented.
- After the session the generated R-code and documentation was compiled into a Sweave file and typeset by Markus Neuner (responible for documentation and R-code hacking).

Main question

Do certain genes show very high (or low) expression for cancer samples compared to normal samples (without high risk)?

Workflow I

The following workflow was performed for each study:

1. Import of **raw data** (with *ArrayExpress*) and **preparation** (subsetting)
2. **Quality Assessment** before normalization
   - `arrayQualityMetrics` convenience functions with automatic outlier detection
3. **Preprocessing**: background correction, normalization, summarization
   - RMA only (Robust Multi-Array Average expression measure, `rma` from package `affy`)
4. **Quality Assessment** after preprocessing
5. **Quality Control**: Remove possible outlier(s) (repeat preprocessing and quality assessment)
6. Nonspecific **filtering**: remove control probes, high variability, genes that do not match to identifiers
   - `nsFilter` from the package `genefilter`

Workflow II

7. **Testing and ranking**: differential expressed genes
   - Moderated t-test from package limma (lmFit, eBayes, topTable)
   - t-statistic from package genefilter and multtest (rowttests, mt.rawp2adjp) only for comparison
8. **Annotation**: number of the microarray is replaced by something more informative (e.g. gene name, Entrez gene ID, chromosome, location, description)
   - with package `annotate, annaffy`
9. Build **classificator** for prediction with *leave-one-out cross-validation* (LOOCV)
   - `MLearn` from package `MLInterfaces`
   - Classificators: k-nearest neighbor, naive Bayes, random forest
10. Biomedical **validation**:
    - Gene Set Enrichment Analysis with GO (`hyperGTest` from the package `GOstats, plotGOTermGrah` from the `Rgraphviz` package)
    - Downloading and filtering information from PubMed (`getPMID, pm-getabsts` from package `annotate`)

Studies I

ID	Tumor samples	Normal samples
Study 1: E-GEOD-15578	4 (4 subjects)	6 (4 subjects) 7 high-risk
Study 2: E-GEOD-6008	99 (99 subjects) (37 endometrioid, 41 serous, 13 mucinous and 8 clear cell carcinomas)	4 (4 subjects)

Study 2 stages

```
1  1a  1A  1c  1C   2  2a  2A  2B  2c  2C   3  3B  3c  3C  3D
5   3  11   4   8   1   1   3   1   2   3   8   1  13  21   1
4 I a I c N/A
9   1   3   4
```

Problems I

General

- No red line in the Book (what, when, why, where !?, solution: decition for one method)
- Vignettes often not usable (solution: google, try and error/fail)
- Too much methods available...

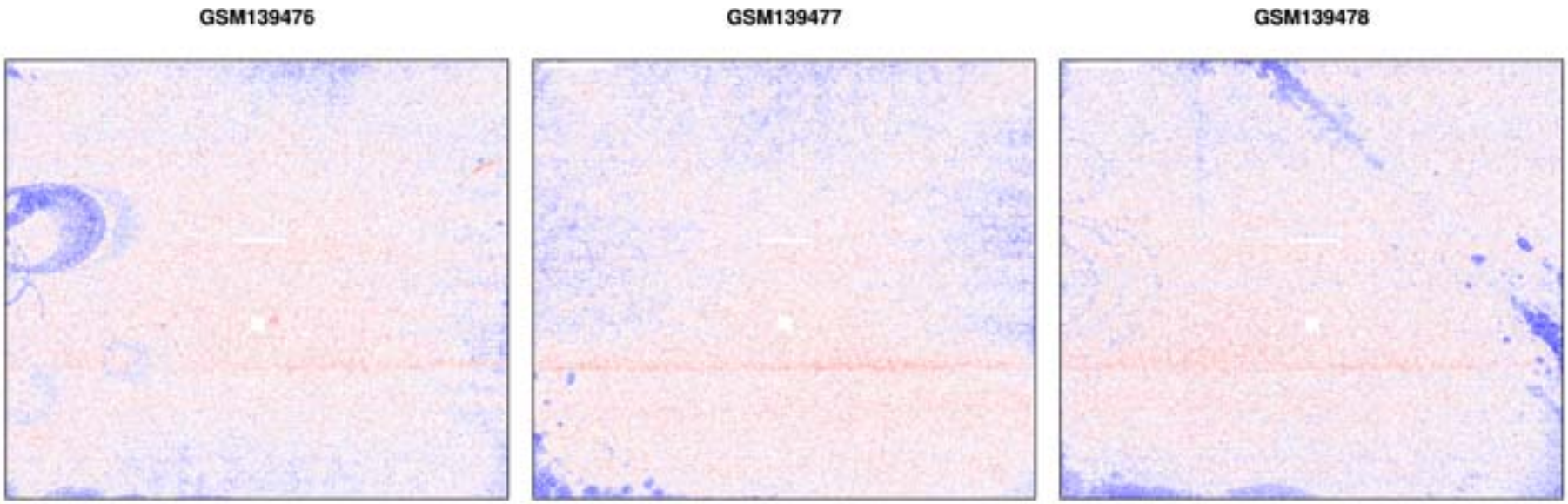
Study 1: E-GEOD-15578

- almost no problems (one normal sample removed at quality control)
- small sample size (4 cancer, 5 normal)

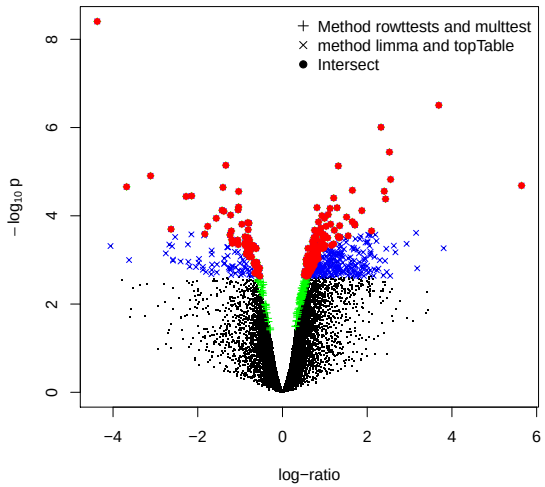
Problems II

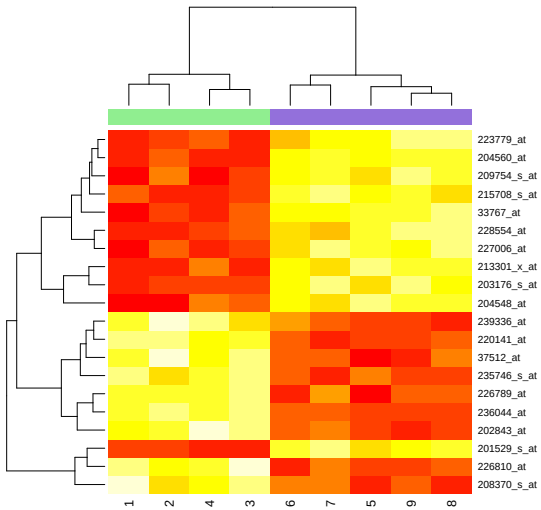
Study 2: E-GEOD-6008

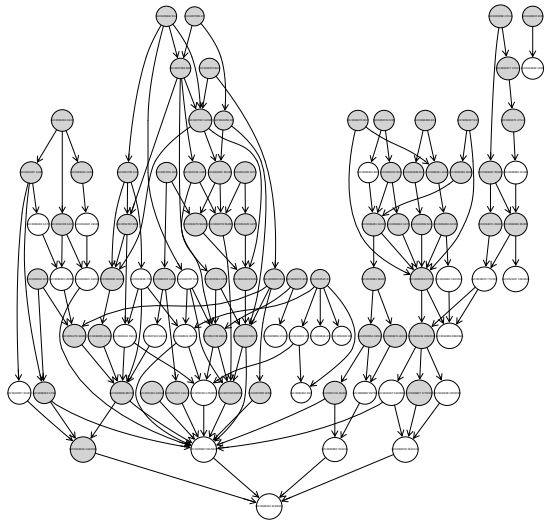
- Comparison cancer stage 1c vs. normal was not possible (3 of 4 normal samples were detected as outliers)
- Alternatively cancer stage 1c was compared against cancer stage 4 (one sample from cancer 1c and 3 samples from cancer 4 were removed at quality control, result 14 samples cancer1c and 6 samples cancer4)



ENDE







Vielen Dank für Ihre Aufmerksamkeit!