



Tim Roloff Handschin

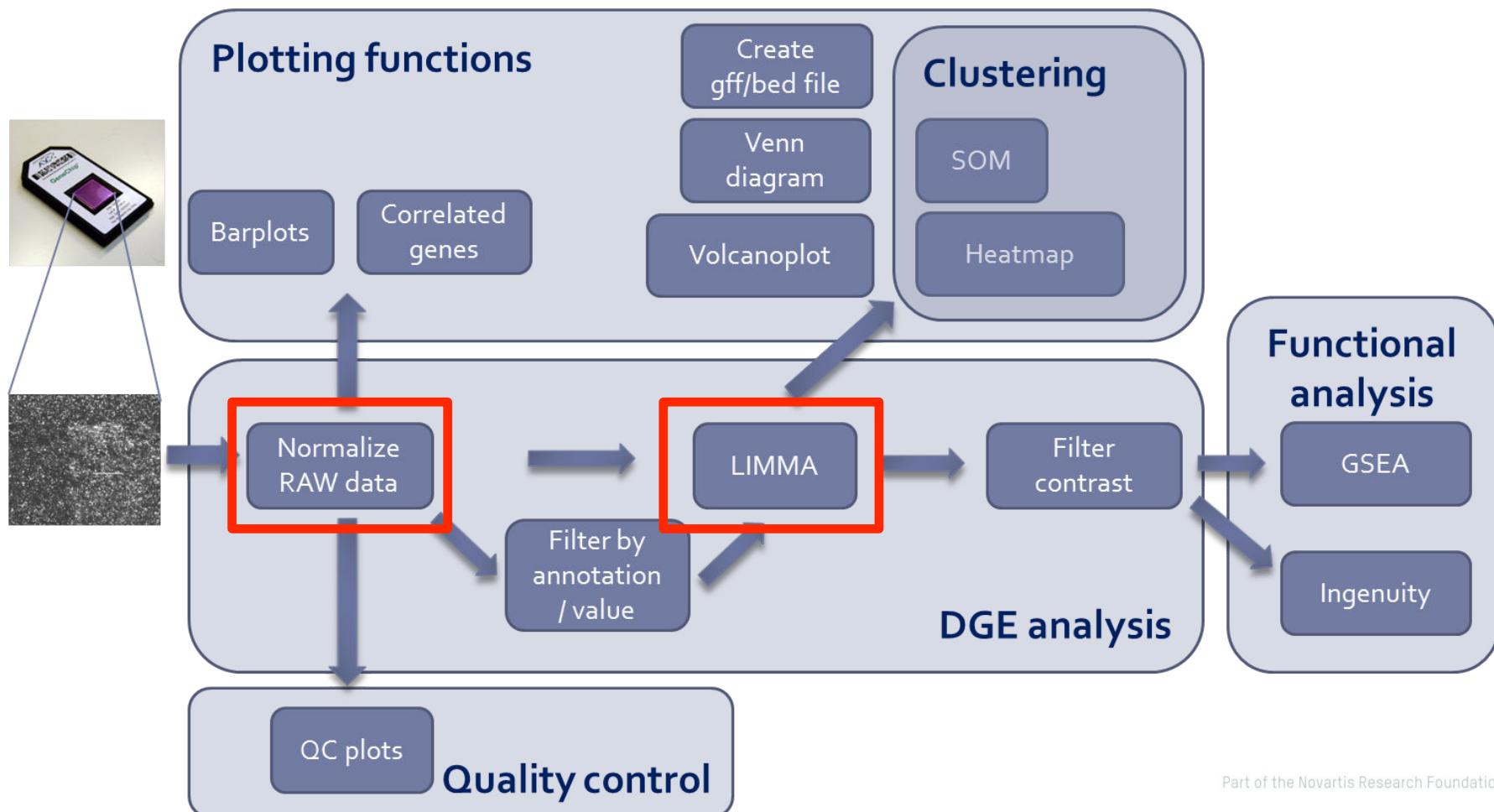
Galaxy workshop Bern
Oct 3, 2012

Reduce standard analyses for bioinformaticians

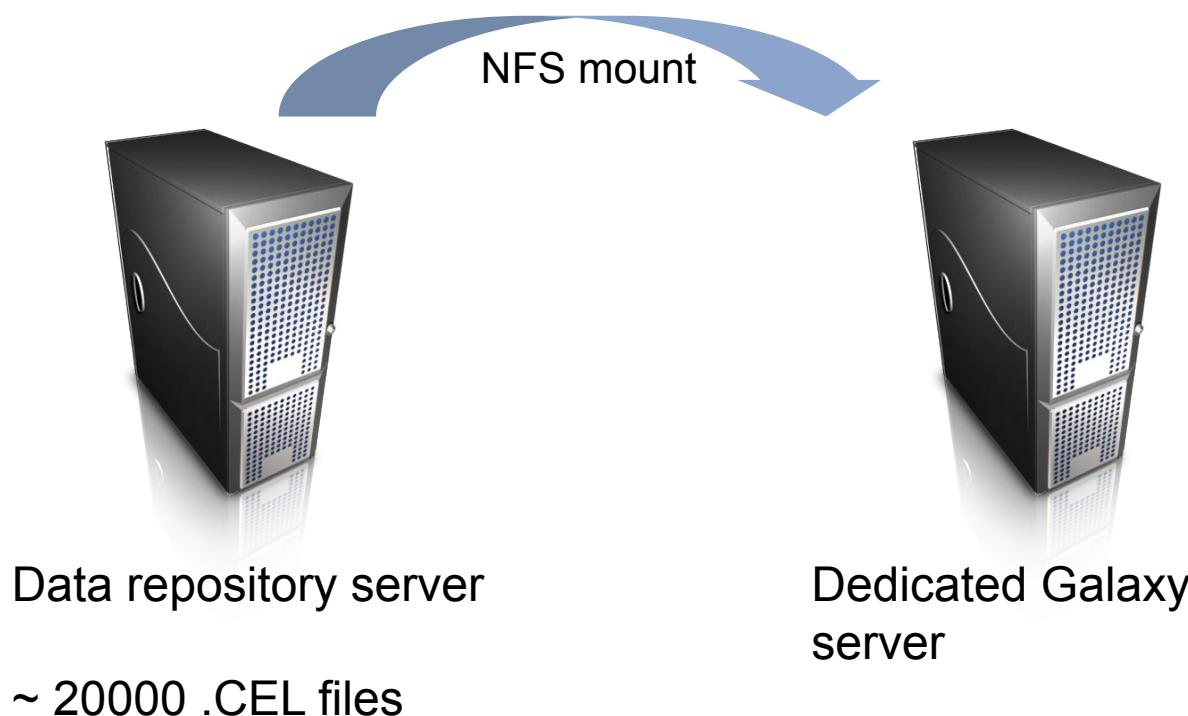


Reduce standard analyses for bioinformaticians

- Wrap R/Bioconductor scripts for Affymetrix microarray analyses in Galaxy



Raw data stored outside galaxy



File paths passed as input parameter to tool

xml file created by cron job for each group in galaxy_dist/tool-data

```
<option name="GSE10895" value="GSE10895">
  <option type="meta_key" name="GSM276175.CEL" value="/work/affy/external/GEO/GSE10895/GSM276175.CEL.gz"/>
  <option type="meta_key" name="GSM276176.CEL" value="/work/affy/external/GEO/GSE10895/GSM276176.CEL.gz"/>
  <option type="meta_key" name="GSM276177.CEL" value="/work/affy/external/GEO/GSE10895/GSM276177.CEL.gz"/>
  <option type="meta_key" name="GSM276178.CEL" value="/work/affy/external/GEO/GSE10895/GSM276178.CEL.gz"/>
</option>
```

Tool xml file

```
<conditional>
  </when>
  <when value="external">
    <repeat name="conditions" title="Condition" help="Select the arrays for this condition">
      <param name="condName" type="text" label="Condition Name" help="Name for this group">
        <validator type="length" min="1" max="12" message="Please enter a short name for each condition!" />
      </param>
      <param name="cels" type="drill_down" display="checkbox" hierarchy="recurse" multiple="true" label="CEL files" from_file="fungen/external.xml">
        <validator type="no_options" message="Select at least 2 .CEL files for each condition!" />
      </param>
    </repeat>
  </when>
</conditional>
```

From CEL files to ExpressionSet

Galaxy / FMI-Xenon1

Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Tools Options

Normalize arrays (version 1.0.0)

Select your group: External

Conditions Select the arrays for this condition

Condition 1

Condition Name: Brain

Name for this group use short names, avoid special characters and numbers at the beginning of the name (The names might be changed by the program to make them conform to processing in R.)

Choose CEL files:

- [+] Affymetrix
- [+] GEO
- [+] GSE10797
- [+] GSE10895
- [+] GSE12306
- GSM309430.CEL
- GSM309434.CEL
- GSM309436.CEL
- GSM309438.CEL
- GSM309441.CEL
- GSM309450.CEL
- GSM309451.CEL
- GSM309452.CEL
- [+] GSE14548
- [+] GSE15907
- [+] GSE16923
- [+] GSE21577
- [+] GSE26730
- [+] GSE27896
- [+] GSE37975
- [+] GSE38650
- [+] GSE5764
- [+] GSE5847
- [+] GSE6532
- [+] GSE9012

Select the raw data CEL files for the arrays to be loaded.

Condition 2

Condition Name: Muscle

Name for this group use short names, avoid special characters and numbers at the beginning of the name (The names might be changed by the program to make them conform to processing in R.)

Choose CEL files:

- [+] Affymetrix
- [+] GEO

Select the raw data CEL files for the arrays to be loaded.

History Options

1: Contrast data

11: R list with contrasts

10: Probesets with similar expression pattern

9: Expression plots for similar genes

8: Filtered expressionset

7: Filtered expressionset

6: expression barplots on ExpressionSet of RMA normalized data

5: microarray PCA on ExpressionSet of RMA normalized data

4: Pairwise correlation plot for arrays 1,2,3,4,5,6,7,8,9

3: RMA normalized data

34,761 lines
format: tabular, database: ?
Info: WARNING: ignoring environment value of R_HOME
Background correcting
Normalizing
Calculating Expression

1	2	3	4
probeSet_id	MouseTP_Brain_01_n2GENE.CEL	MouseTP_Brain_02_n2GENE.CEL	MouseTP_Brain_03_n2GENE.CEL
mouseTP_Brain_01_n2GENE.CEL	12.1592323216505	12.1592323216505	11.5539544503765
10338001	10.1827423313285	9.5721702361821	9.50086763341435
10338004	9.09842940454923	9.11524886401235	9.94047223733102
10338017	12.3853425428555	13.225798746805	13.193571744357

2: Annotation of RMA normalized data

9 lines
format: exphe, database: ?
Info: WARNING: ignoring environment value of R_HOME
Background correcting
Normalizing
Calculating Expression

1_CelFileName	2_ArrayId	3_Condition
MouseTP_Brain_01_n2GENE.CEL	1	Brain
MouseTP_Brain_02_n2GENE.CEL	2	Brain
MouseTP_Brain_03_n2GENE.CEL	3	Brain
MouseTP_SkeletalMuscle_01_n2GENE.CEL	4	Muscle
MouseTP_SkeletalMuscle_02_n2GENE.CEL	5	Muscle
MouseTP_SkeletalMuscle_03_n2GENE.CEL	6	Muscle

1: ExpressionSet of RMA normalized data

2.4 Mb
format: exset, database: ?
Info: WARNING: ignoring environment value of R_HOME
Background correcting
Normalizing
Calculating Expression

Binary ExpressionSet data file

Custom datatypes for R objects

Normalization

History item	Comment
ExpressionSet of normalized data	Binary dataset, used as input for following tools
Annotation of normalized data	Tabular file listing the normalized CEL files and conditions
Normalized data	Large tabular file with normalized expression values for each probeset on each array

Linear modeling (LIMMA)

History item	Comment
Contrast data (R list)	Binary dataset with contrast data; to be used as input for subsequent tools
Contrast data	Tabular dataset with normalized expression values for all arrays, logFCs, p-values, adj p-values for each contrast and annotation
Contrast Information	Tabular dataset with contrast names

Additional concepts used

- Access rights controlled via \$userEmail and lookup table
- In R:
 - R objects stored as binary files using save(): faster reading and writing
 - Data
 - Annotation
 - suppressMessages() and suppressWarnings() to avoid output to standard error
- External data (GEO, ArrayExpress, Affymetrix) downloaded to data repository

Status quo:

- 21 tools implemented
- > 500 analyses run

We are currently working on:

- QC for raw data
- Plot probesets for genomic regions
- Meta analyses
- Time-course analyses
- Data preparation for external tools / webservices
 - David
 - Broad GSEA
 - Mara
- Data preparation for GEO submission

Thanks to

- **Hans-Rudolf Hotz**
- **FMI IT support**
- **FMI Galaxy users (155)**