Gene expression data analysis
SaferNanoDesign 29.05.2018

Dr. Friederike Ehrhart
Department for Bioinformatics – BiGCaT
Maastricht University
friederike.ehrhart@maastrichtuniversity.nl
ORCID: 0000-0002-7770-620X
Content

1. The dogma of molecular biology – DNA/RNA/protein relationship
2. Data and omics data – and their research strategies
3. Bioinformatics and databases
4. How to do gene expression data analysis – ArrayAnalysis.org and PathVisio
5. Limitations and pitfalls
1. The dogma of molecular biology

http://www.youtube.com/watch?feature=playr_detailpage&v=9kOGOY7vthke
DNA and genes

A gene is a locus (or region) of the DNA that encodes a functional RNA.

CDS = Coding DNA Sequence
UTR = UnTranslated region

www.carolguze.com
## Number of genes per species

<table>
<thead>
<tr>
<th>Species and Common Name</th>
<th>Estimated Total Size of Genome (bp)*</th>
<th>Estimated Number of Protein-Encoding Genes*</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Saccharomyces cerevisiae</em> (unicellular budding yeast)</td>
<td>12 million</td>
<td>6,000</td>
</tr>
<tr>
<td><em>Trichomonas vaginalis</em></td>
<td>160 million</td>
<td>60,000</td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em> ( unicellular malaria parasite)</td>
<td>23 million</td>
<td>5,000</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em> (nematode)</td>
<td>95.5 million</td>
<td>18,000</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em> (fruit fly)</td>
<td>170 million</td>
<td>14,000</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em> (mustard; thale cress)</td>
<td>125 million</td>
<td>25,000</td>
</tr>
<tr>
<td><em>Oryza sativa</em> (rice)</td>
<td>470 million</td>
<td>51,000</td>
</tr>
<tr>
<td><em>Gallus gallus</em> (chicken)</td>
<td>1 billion</td>
<td>20,000-23,000</td>
</tr>
<tr>
<td><em>Canis familiaris</em> (domestic dog)</td>
<td>2.4 billion</td>
<td>19,000</td>
</tr>
<tr>
<td><em>Mus musculus</em> (laboratory mouse)</td>
<td>2.5 billion</td>
<td>30,000</td>
</tr>
</tbody>
</table>

Plants and amphibians with huge genomes (not in table) do not have huge amounts of genes.

Organization of the human genome

- Human Genome: 3100 Mb
  - Genes and Gene Related Sequences: 1200 Mb
    - Genes (Exons): 48 Mb
    - Related Sequences: 1152 Mb
  - Intergenic DNA: 1000 Mb
    - Interpersed Repeats: 400 Mb
    - Other Intergenic Regions: 600 Mb
  - Pseudogenes
  - LINES: 640 Mb
  - SINES
  - Various

- DNA of Human Genome:
  - Junk DNA: 97-98%
  - Protein Synthesis DNA: 2-3%
RNA

**mRNA**: messenger RNA – will be translated into protein

**rRNA**: ribosomal RNA – forms ribosomes

**tRNA**: transfer RNA – brings the right amino acids to the ribosomes

**siRNA**: silencer RNA – blocks specific mRNA

**miRNA**: micro RNA – regulative effect on specific translation of proteins

**lncRNA**: long non-coding RNA – regulative effect on specific DNA regions
Protein

21 different amino acids

Protein function:

• Enzymes: catalysis of chemical reactions (alcohol dehydrogenase, lactase)
• Cell signaling: transmit signals within cells (antibodies, insulin)
• Structural proteins: form extracellular matrix (collagen) or generate mechanical force (myosin)
• [https://www.youtube.com/watch?v=yKW4F0Nu-UY](https://www.youtube.com/watch?v=yKW4F0Nu-UY)
2. Data and omics data

Single biomolecules

DNA
RNA
PROTEIN

Sequencing and gene identification
Sequencing and gene expression
Identification and structure determination

High throughput

GENOME
TRANSCRIPTOME
PROTEOME
The size of omics data

- 3,100,000,000 bp DNA per individual
- 22,000 (protein coding) genes
- 120,000 transcripts
- 100,000 proteins
- 40,000 metabolites
COMICS
the best kind of omics!

- genomics
- transcriptomics
- proteomics
- phenomics
- epigenomics
- metabolomics

http://www.stathiskanterakis.com/?p=286
Research strategies

Hypothesis-driven research “reductionistic”

Hypothesis/Theory

Experiment

Falsification

Some hard thinking

New hypothesis

Data driven research “holistic”

Samples

Unbiased omics strategies

Data driven results

Data driven hypothesis/theory

Some hard thinking
Example: Nanomaterial toxicity assessment

Hypothesis – driven: Silver nanoparticles increases oxidative stress in Caco-2 cells

- In vitro assay – exposure scenario
- Oxidative stress assessment
  - Chance in SOD protein expression
  - Level of ROS
- Does the results confirm the hypothesis?

Data – driven: What happens to the transcriptome of Caco-2 cells after exposure to silver nanoparticles

- In vitro assay – exposure scenario
- Collect transcriptome (RNA) and run e.g. RNA-sequencing or microarray analysis
- Data analysis
- Result: list of changed gene expression
- Interpretation
3. What is Bioinformatics?

Bioinformatics uses "informatics" techniques (from applied math, computer science, statistics, etc.) to understand and organize biological information, like genes, proteins and molecules on a large-scale.
Why Bioinformatics?

DATA KNOWLEDGE

CONVERT DATA TO KNOWLEDGE
GENERATE NEW HYPOTHESES

Experimental

Public Databases

DATA

22,000 genes
120,000 transcripts
100,000 proteins
40,000 metabolites

KNOWLEDGE

DESIGN NEW EXPERIMENTS
Bioinformatics tools and methods

- pattern recognition
- data mining
- machine learning algorithms
- visualization

Examples:
- sequence alignment (BLAST)
- gene finding
- genome assembly
- drug design, drug discovery
- protein structure alignment
- protein structure prediction
- prediction of gene expression and protein–protein interactions
- genome-wide association studies
- the modeling of evolution and cell division/mitosis.
Biological databases

1. Global nucleotide/protein sequence storage databases:
   - GenBank of NCBI (National Center for Biotechnology Information)
   - The European Molecular Biology Laboratory (EMBL) Ensembl database
   - The DNA Data Bank of Japan (DDBJ)

2. Genome-centered databases
   - NCBI genomes
   - Ensembl Genome Browser
   - UCSC Genome Bioinformatics Site

3. Protein Databases
   - UniProt

4. Metabolite Databases
   - HMDB, ChEBI

5. Interaction Databases
   - Pathways: WikiPathways, KEGG, Reactome

6. Nanomaterial Databases
   - eNanoMapper, Nanowerk, nature.nano
Genome centered databases

UCSC
http://genome.ucsc.edu/

NCBI

Ensembl
http://www.ensembl.org/
Example: DHH

Location  Gene  Transcript

Even more information

Exon  Intron
Where does all this information come from?

- Submissions (e.g. Sequences)
- Literature
- Curators and contributors
- Automated generation by computer tools
- High-throughput lab screenings
- Individual contributions and large scale contributions
Example: DHH

Identifiers

Learn more in the practical!
Unique identifiers – a game of names

- **RefSeq:**
  - Chromosome: NC_
  - mRNA: NM_
  - Protein: NP_

- **Genbank:**
  - Many types of IDs

- **NCBI gene ID:**
  - Number

- **OMIM ID:**
  - Number

- **Pubmed ID:**
  - Number

- No common identifier for nanoparticles yet
Ensembl identifiers

ENSG### Ensembl Gene ID
ENST### Ensembl Transcript ID
ENSP### Ensembl Peptide ID
ENSE### Ensembl Exon ID

For other species than human a suffix is added:
MUS (*Mus musculus*) for mouse: ENSMUSG###
DAR (*Danio rerio*) for zebrafish: ENSDARG###, etc.
### Identifier Mapping

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Probeset</td>
<td>GSM143409</td>
<td>GSM143411</td>
<td>GSM143412</td>
<td>GSM143413</td>
<td>GSM143414</td>
</tr>
<tr>
<td>2</td>
<td>1415670_at</td>
<td>208.9</td>
<td>171.3</td>
<td>185.1</td>
<td>179.7</td>
<td>226.2</td>
</tr>
<tr>
<td>3</td>
<td>1415671_at</td>
<td>330.7</td>
<td>281.9</td>
<td>301.1</td>
<td>355.4</td>
<td>300.9</td>
</tr>
<tr>
<td>4</td>
<td>1415672_at</td>
<td>488.5</td>
<td>453.5</td>
<td>474.8</td>
<td>477.8</td>
<td>477.5</td>
</tr>
<tr>
<td>5</td>
<td>1415673_at</td>
<td>90.1</td>
<td>103.2</td>
<td>64.3</td>
<td>90.8</td>
<td>96.8</td>
</tr>
<tr>
<td>6</td>
<td>1415674_at</td>
<td>167.5</td>
<td>187.9</td>
<td>168.3</td>
<td>174.8</td>
<td>148.5</td>
</tr>
<tr>
<td>7</td>
<td>1415675_at</td>
<td>70.1</td>
<td>95.6</td>
<td>81.1</td>
<td>66.6</td>
<td>70.6</td>
</tr>
<tr>
<td>8</td>
<td>1415676_at</td>
<td>1142.2</td>
<td>123.2</td>
<td>123.2</td>
<td>123.2</td>
<td>123.2</td>
</tr>
</tbody>
</table>

_Ensambl_  
**Home > Human**  
**Location:** X:19,271,972-19,287,886  
**Gene:** PDHA1  
- **Gene summary**  
  - Splice variants (5)  
  - Supporting evidence  
  - Sequence  
  - External references (7)  
**Gene:** PDHA1 (ENSG00000131828)  
Pyruvate dehydrogenase E1 component subunit  
Source: UniProtKB/Swiss-Prot P68559  
**Location:** Chromosome X: 19,271,972-19,287,886  
**Transcripts:** There are 5 transcripts

**Annotation:** ENSG00000131828

**Mapping database:** BridgeDb.org
Data integration by ontology

**Ontology:**
- Controlled language
- Relationship between terms
  - Hierarchy
  - Is_a/has_a
- Machine readable (OWL)
- Repositories
  - AberOwl
  - Bioportal
  - OLS (EBI ontology lookup service)
Anatomy of a GO term

id: GO:0006094
name: gluconeogenesis
namespace: process
def: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol.
[http://cancerweb.ncl.ac.uk/omd/index.html]
exact_synonym: glucose biosynthesis
xref_analog: MetaCyc:GLUCONEO-PWY
is_a: GO:0006006  (glucose metabolic process)
is_a: GO:0006092  (main pathway of carbohydrate metabolism)
The 3 Gene Ontologies

- **Molecular Function** = elemental activity/task
  - the tasks performed by individual gene products; examples are *carbohydrate binding* and *ATPase activity*

- **Biological Process** = biological goal or objective
  - broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

- **Cellular Component** = location or complex
  - subcellular structures, locations, and macromolecular complexes; examples include *nucleus, telomere*, and *RNA polymerase II holoenzyme*
Searching and Browsing GO

- AmiGO
  - http://www.godatabase.org

- Downloads
  - http://www.godatabase.org/dev/database/
  - XML or as a MySQL database dump

  - Annotate gene by sequence similarity.
Back to databases...

Single entity + interaction
= biological pathways!
WikiPathways and How to Change the World

(or at least a small corner of the world)
Nanomaterial database
How to represent nanomaterials in a database?

- **Nanomaterials**
  - Core
  - Coating(s)
  - Linkage
  - Impurities
  - Components, internal structure, etc.

- **Typical assay description**
  - Property – value (range of values) – units (*Excel templates*)

- **More complex description:**
  - Experimental graph (*ISA-TAB / ISA-TAB-nano*)

- **Existing data models**
  - *BioAssay Ontology*
  - *OECD Harmonized Templates*
  - *CoDATA UDS*
  - *ISA-TAB- Nano*

- **Commonalities:**
  - Materials sample
  - Protocols, protocol parameters
  - Experimental conditions
  - Readouts
    - Measurements,
    - Measurement groups,
    - Raw data, derived data
https://data.enanomapper.net/
Integrated view of eNanoMapper database [contributors] and caNanoLab

Current Selection
(x) silver
(x) substanceType:NPO_1892

- Data sources
- Nanomaterial type
  - silver 27
  - P-CHEM (27)
  - TOX (82)
- Cell
- Species
- Results
- References
- Protocols
- Instruments

< 1 2 > displaying 1 to 20 of 27

- **Ag (Harper2011 9)**
P-CHEM. Nanomaterial surface chemistry ATOMIC COMPOSITION = more
- **Ag (Harper2011 8)**
P-CHEM. Nanomaterial surface chemistry ATOMIC COMPOSITION = more
- **Ag (Harper2011 7)**
P-CHEM. Nanomaterial surface chemistry ATOMIC COMPOSITION = more
- **Ag (Harper2011 6)**
P-CHEM. Nanomaterial surface chemistry ATOMIC COMPOSITION = more
- **Ag (Harper2011 5)**
P-CHEM. Nanomaterial surface chemistry ATOMIC COMPOSITION = more
- **Ag (Harper2011 4)**
P-CHEM. Nanomaterial surface chemistry ATOMIC COMPOSITION = more

- **material composition study**
Search by phys-chem parameter or biol. effect
ENM - ontology

Details

**ACRONYM**
ENM

**VISIBILITY**
Public

**BIOPORTAL PURL**
http://purl.bioontology.org/ontology/ENM

**DESCRIPTION**
The eNanoMapper ontology covers the full scope of terminology needed to support research into nanomaterial safety. It builds on multiple pre-existing external ontologies such as the NanoParticle Ontology.

**STATUS**
Alpha

**FORMAT**
OWL

**CONTACT**
Egon Willighagen, egon.willighagen@maastrichtuniversity.nl
Friederike Ehrhart, friedriek.ehrhart@maastrichtuniversity.nl
Garath Owen, gowen@ebi.ac.uk
Linda Rieswijk, linda.rieswijk@maastrichtuniversity.nl
Jie Kong, jkong@ebi.ac.uk
Janna Hastings, hastings@ebi.ac.uk

**HOME PAGE**
https://github.com/enanomapper/ontologies

**PUBLICATIONS PAGE**
http://enanomapper.net/library

**DOCUMENTATION PAGE**
http://enanomapper.net/ontology

**CATEGORIES**
Chemical, Health

**GROUPS**

Metrics

| NUMBER OF CLASSES: | 7937 |
| NUMBER OF INDIVIDUALS: | 196 |
| NUMBER OF PROPERTIES: | 2 |
| MAXIMUM DEPTH: | 10 |
| MAXIMUM NUMBER OF CHILDREN: | 1419 |
| AVERAGE NUMBER OF CHILDREN: | 6 |
| CLASSES WITH A SINGLE CHILD: | 441 |
| CLASSES WITH MORE THAN 25 CHILDREN: | 56 |
| CLASSES WITH NO DEFINITION: | 2756 |

Visits

Download as CSV

No reviews available.

Submissions

<table>
<thead>
<tr>
<th>SUBMISSION</th>
<th>RELEASE DATE</th>
<th>UPLOAD DATE</th>
<th>DOWNLOADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0 (Parsed, Indexed, Metrics, Annotator)</td>
<td>05/22/2016</td>
<td>05/22/2016</td>
<td>OWL</td>
</tr>
<tr>
<td>3.0 (archived)</td>
<td>03/01/2016</td>
<td>03/01/2016</td>
<td>OWL</td>
</tr>
<tr>
<td>2 (archived)</td>
<td>01/29/2016</td>
<td>01/29/2016</td>
<td>OWL</td>
</tr>
<tr>
<td>1 (archived)</td>
<td>01/28/2016</td>
<td>01/28/2016</td>
<td>OWL</td>
</tr>
</tbody>
</table>
Nanoparticle

ENM - eNanoMapper

- Label: nanoparticle
- OID: NPO:707
- SubClassOf: primary particle
- Class: http://purl.bioontology.org/ontology/npo#NPO_707
- http://purl.bioontology.org/ontology/npo#code: npo_707
- http://purl.bioontology.org/ontology/npo#preferred_Name: nanoparticle
- Ontology: ENM
- RDFS:Comment: Definition is partly based on ASTM E 2456-06 (towards nanotechnology).
- Type: class
How to use this interactive machine readable biological data?

• Databases
  – Information (API – manually)

• Tools
  – Re-analysis
  – Modelling
  – Reference materials
4. How to **DO** data analysis?

Just with one button, you do all analysis! Right?

Look! Do you see this red button? I especially made it for you. Everything is just done with this button, so easy!
Microarray experiments workflow

E.g. Caco-2 cells exposed to silver nanoparticles
- Caco-2 small intestine cell line (human)
- Exposed to
  - 2.5 µg/ml Ag nanoparticles
  - 25 µg/ml Ag nanoparticles
  - 0.5 µg/ml AgNO₃ (soluble)
- Control without exposure
- Data from GEO: GSE62253
- Publication:

 Nanotoxicology

Molecular mechanism of silver nanoparticles in human intestinal cells
Linda Böhmert, Birgit Niemann, Dajana Lichtenstein, Sabine Juling, and Alfonso Lampen

Department Food Safety, Federal Institute for Risk Assessment, Berlin, Germany
Microarray data analysis

- Image analysis
- Quality Control
- Pre-processing
  - Background correction
  - Normalisation
  - Filtering
  - Annotation
Microarray data analysis

- Statistical evaluation
  - T-test
  - ANOVA / modelling
- Further analysis
  - Significantly changed genes
  - LogFC (log2)
Microarray data analysis

- Biological evaluation
  - Pathway analysis
  - Gene Ontology analysis
  - Network analysis
  - Etc...

- Validation
  - Technology
  - Biology / Literature
Microarray data analysis

- With publication of the paper, also the data has to be published: obligatory!
  - ArrayExpress at EBI
  - Gene Expression Omnibus (GEO) at NCBI
- Standard for publication: MIAME
Dual channel (‘two colour’) gene expression microarrays – ‘spotted arrays’

For Affymetrix chips each gene is measured by dozens of probes that are randomly distributed across the chip; these probes together form a probeset.
Affymetrix Chips

Image of Hybridized Probe Array

1.28cm
Image analysis

- Start with a scanned microarray image
- Use software packages to recognise spots and compute (raw) intensities
Example raw data file of a dual channel array

...are these values of good quality?
Quality control

• Check for technical failures or biological outliers
• Check abnormalities on the array or dissimilarities between the arrays
• Decide to repeat / reject arrays if needed
• Think careful before repeating: introducing possible bias
example: QC for Affymetrix arrays using http://www.arrayanalysis.org
Tables and images of QC criteria

**Affymetrix criteria**

- Sample prep controls Lys < Phe < Thr < Dap
- Lys present
- Beta Actin 3’/5’ ≤ 3
- GAPDH 3’/5’ ≤ 1.25
- Hybridisation controls BioB < BioC < BioD < Crex
- BioB present
- Percentage present within 10%
- Background within 20 units
- Scaling factors within 3-fold from the average
RNA degradation

- RNA degradation starts 5’ -> 3’
- Less fragments of 5’ end than 3’ indicates degradation
RNA degradation analysis

3′/5′ and 3′/M′ ratios for beta-actin

beta-actin QC: OK
all 3′/5′ ratios < 3
Average intensity boxplot

- Maximum
- 3rd quartile
- Median
- 1st quartile
- Minimum
This is a simplified example: in reality PCA is used to reduce the dimensions of a multi-dimensional data set to 2 or 3 dimensions.
Now we reduced the two dimensional data set to one dimension, thereby explaining (keeping) 70% of the original variation.
Making all your data comparable:

PRE-PROCESSING
QC and pre-processing

• Ensure signal comparability within each array
  – Stains on the array
  – Gradient over the array

• Ensure comparable signals between all arrays
  – Degraded / low quality sample
  – Failed hybridisation
  – Too low or high overall intensity

• Some effects can be corrected for, others require removal of data from the set
Background correction

• Background signal needs to be corrected for
  – for example signal of remaining non-hybridized mRNA

• Three types of background
  – Overall slide background
    • Can be corrected for by subtracting mean background, or by subtracting mean of empty spots
  – Local slide background
    • Same as previous, but per slide region
  – Specific background
    • For example cross-hybridization, can be corrected for by mismatch probes (in case of Affymetrix arrays)
The importance of background
Uneven background
Background correction

- Measure the intensity of the background around the spot as well as the intensity of the spot itself

- Reported intensity = spot intensity – background intensity
- More advanced methods are generally used
  - prevent negative values
Normalisation

• Adjusting values

• **Between-slide** normalisation: correct for experimental differences between slides
  – e.g. one may have an overall higher signal due to differences in hybridisation

• **Within-slide** normalisation: correct for within slide variations
  – by applying normalisation per region, per spot group etc.

• For dual channel arrays: **between-channel** normalisation
Boxplot before and after normalization

**before**

**after**
Finding differentially expressed genes:

STATISTICAL ANALYSIS
Which genes have changed? (I)

- “Every gene that has changed two-fold is relevant”
- Doesn’t take variation into account
Often people use both the difference and statistical significance between two groups to determine the list of **differentially expressed** genes.

<table>
<thead>
<tr>
<th></th>
<th>Large difference</th>
<th>Small difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Significant</strong></td>
<td>![Stars] (Large)</td>
<td>![Stars] (Small)</td>
</tr>
<tr>
<td><strong>Non-Significant</strong></td>
<td>![Stars] (Small)</td>
<td>![Stars] (Small)</td>
</tr>
</tbody>
</table>
Comparing experimental groups

- The ratio between the (average) expression in two experimental groups is generally called the **fold change**

- Generally we do not work directly with the fold, but with the logged fold change, which is called the **log ratio or log fold change**

- Why?
  - Intuitive understanding log ratio -> + is up, - down
  - Visualization
Considerations for the t-test

• Requirements
  – Adequate sample size (n)
  – Knowledge on the experimental grouping
  – Normally distributed data
    • “Gaussian curve”
  – P value
### Example results table

<table>
<thead>
<tr>
<th></th>
<th>logFC</th>
<th>Fold Chan</th>
<th>AveExpr</th>
<th>P.Value</th>
<th>adj.P.Val</th>
<th>B</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-7.23489</td>
<td>-150.633</td>
<td>7.863147</td>
<td>2.37E-11</td>
<td>1.02E-07</td>
<td>15.14212</td>
<td>170496</td>
</tr>
<tr>
<td>3</td>
<td>-2.67637</td>
<td>-6.39244</td>
<td>11.12548</td>
<td>-20.2298</td>
<td>1.53E-09</td>
<td>3.30E-06</td>
<td>12.15308</td>
</tr>
<tr>
<td>4</td>
<td>-4.92565</td>
<td>-30.3926</td>
<td>7.786763</td>
<td>-19.26</td>
<td>2.49E-09</td>
<td>3.58E-06</td>
<td>11.75189</td>
</tr>
<tr>
<td>5</td>
<td>-2.7171</td>
<td>-6.7555</td>
<td>7.613864</td>
<td>-17.415</td>
<td>6.73E-09</td>
<td>7.27E-06</td>
<td>10.90447</td>
</tr>
<tr>
<td>6</td>
<td>-4.97497</td>
<td>-31.4458</td>
<td>9.115975</td>
<td>-16.431</td>
<td>1.19E-08</td>
<td>1.03E-05</td>
<td>10.40304</td>
</tr>
<tr>
<td>7</td>
<td>-2.14602</td>
<td>-4.42604</td>
<td>9.631401</td>
<td>-15.1037</td>
<td>2.72E-08</td>
<td>1.96E-05</td>
<td>9.657943</td>
</tr>
<tr>
<td>9</td>
<td>2.696893</td>
<td>6.484038</td>
<td>6.947894</td>
<td>13.93027</td>
<td>5.98E-08</td>
<td>3.23E-05</td>
<td>8.930261</td>
</tr>
<tr>
<td>10</td>
<td>2.0373</td>
<td>4.104765</td>
<td>7.227198</td>
<td>13.19942</td>
<td>1.01E-07</td>
<td>4.84E-05</td>
<td>8.439403</td>
</tr>
<tr>
<td>11</td>
<td>1.848977</td>
<td>3.602446</td>
<td>7.935563</td>
<td>12.39145</td>
<td>1.85E-07</td>
<td>7.99E-05</td>
<td>7.859381</td>
</tr>
<tr>
<td>12</td>
<td>-4.19194</td>
<td>-18.2768</td>
<td>7.498633</td>
<td>-12.1536</td>
<td>2.23E-07</td>
<td>8.60E-05</td>
<td>7.680627</td>
</tr>
<tr>
<td>13</td>
<td>-2.54811</td>
<td>-5.84867</td>
<td>6.233037</td>
<td>-12.0639</td>
<td>2.39E-07</td>
<td>8.60E-05</td>
<td>7.612213</td>
</tr>
<tr>
<td>14</td>
<td>-1.1417</td>
<td>-2.20649</td>
<td>5.738167</td>
<td>-11.8145</td>
<td>2.92E-07</td>
<td>9.69E-05</td>
<td>7.419026</td>
</tr>
<tr>
<td>15</td>
<td>2.512379</td>
<td>5.621646</td>
<td>8.801788</td>
<td>-1.18545</td>
<td>1.03E-07</td>
<td>1.48E-05</td>
<td>1.251565</td>
</tr>
<tr>
<td>16</td>
<td>-2.05217</td>
<td>-15.4767</td>
<td>7.790343</td>
<td>-7.8274</td>
<td>4.85E-08</td>
<td>2.20E-05</td>
<td>1.251565</td>
</tr>
<tr>
<td>17</td>
<td>1.684944</td>
<td>4.353204</td>
<td>9.176813</td>
<td>14.2989</td>
<td>1.02E-07</td>
<td>7.97E-05</td>
<td>7.859381</td>
</tr>
<tr>
<td>18</td>
<td>-1.41899</td>
<td>-11.3647</td>
<td>10.45552</td>
<td>-12.1536</td>
<td>2.23E-07</td>
<td>8.60E-05</td>
<td>7.680627</td>
</tr>
<tr>
<td>19</td>
<td>-2.07639</td>
<td>-5.84867</td>
<td>6.233037</td>
<td>-12.0639</td>
<td>2.39E-07</td>
<td>8.60E-05</td>
<td>7.612213</td>
</tr>
<tr>
<td>20</td>
<td>1.421941</td>
<td>5.621646</td>
<td>8.801788</td>
<td>14.2989</td>
<td>1.02E-07</td>
<td>7.97E-05</td>
<td>7.859381</td>
</tr>
<tr>
<td>21</td>
<td>3.428312</td>
<td>16.0097</td>
<td>11.8656</td>
<td>14.2989</td>
<td>1.02E-07</td>
<td>7.97E-05</td>
<td>7.859381</td>
</tr>
<tr>
<td>22</td>
<td>-1.73615</td>
<td>-11.3647</td>
<td>10.45552</td>
<td>-12.1536</td>
<td>2.23E-07</td>
<td>8.60E-05</td>
<td>7.680627</td>
</tr>
<tr>
<td>23</td>
<td>2.274691</td>
<td>5.621646</td>
<td>8.801788</td>
<td>14.2989</td>
<td>1.02E-07</td>
<td>7.97E-05</td>
<td>7.859381</td>
</tr>
<tr>
<td>24</td>
<td>1.58322</td>
<td>5.621646</td>
<td>8.801788</td>
<td>14.2989</td>
<td>1.02E-07</td>
<td>7.97E-05</td>
<td>7.859381</td>
</tr>
</tbody>
</table>

### Voila!
**List of differentially expressed genes!**

**Possible filtering on:**
- **P value**
  - Only significant changes
- **logFC or FC**
  - Is a significant change with a FC of 1.001 relevant?
- **Average expression**
  - Recall that lowly expressed genes are less reliably measured
Gene ontology enrichment analysis (overrepresentation analysis (ORA))

- NNT  NAD(P) transhydrogenase
- DHRS2  Dehydrogenase/reductase SDR family member 2
- ME3  NADP-dependent malic enzyme
- SDHC  Succinate dehydrogenase cytochrome b560 subunit
- BCO2  Beta,beta-carotene 9',10'-oxygenase
- SURF1  Surfeit locus protein 1

GO: mitochondria oxidation-reduction process!
GO-Elite: a flexible solution for pathway and ontology over-representation

Alexander C. Zambon\(^1\), Stan Gaj\(^2\), Isaac Ho\(^3\), Kristina Hanspers\(^3\), Karen Vranizan\(^3\), Chris T. Evelo\(^2\), Bruce R. Conklin\(^3,4\), Alexander R. Pico\(^3\) and Nathan Salomonis\(^3,\ast\)

Gene Ontology enRIchment anaLysis and visuaLizAtion tool
1. **Step 1: Choose organism**
   - Homo sapiens

2. **Step 2: Choose running mode**
   - Single ranked list of genes
   - Two unranked lists of genes (target and background lists)

3. **Step 3: Paste a ranked list of gene/protein names**
   - **Target set**: list of changed genes
   - **Background set**: list of all investigated genes

4. **Step 4: Choose an ontology**
   - Process
   - Function
   - Component
   - All

   [Search Enriched GO terms]
<table>
<thead>
<tr>
<th>GO term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0070647</td>
<td>protein modification by small protein conjugation or removal</td>
</tr>
<tr>
<td>GO:0006996</td>
<td>organelle organization</td>
</tr>
<tr>
<td>GO:0044260</td>
<td>cellular macromolecule metabolic process</td>
</tr>
<tr>
<td>GO:0044238</td>
<td>primary metabolic process</td>
</tr>
<tr>
<td>GO:0008104</td>
<td>protein localization</td>
</tr>
<tr>
<td>GO:0044267</td>
<td>cellular protein metabolic process</td>
</tr>
<tr>
<td>GO:0033036</td>
<td>macromolecule localization</td>
</tr>
<tr>
<td>GO:0006464</td>
<td>cellular protein modification process</td>
</tr>
<tr>
<td>GO:0036211</td>
<td>protein modification process</td>
</tr>
<tr>
<td>GO:0044237</td>
<td>cellular metabolic process</td>
</tr>
<tr>
<td>GO:0043170</td>
<td>macromolecule metabolic process</td>
</tr>
<tr>
<td>GO:0043412</td>
<td>macromolecule modification</td>
</tr>
<tr>
<td>GO:0048193</td>
<td>Golgi vesicle transport</td>
</tr>
<tr>
<td>GO:0032446</td>
<td>protein modification by small protein conjugation</td>
</tr>
<tr>
<td>GO:0071704</td>
<td>organic substance metabolic process</td>
</tr>
</tbody>
</table>
Pathway analysis
Why Pathway Analysis?

• Intuitive to biologists
  – Puts data in biological context
  – More intuitive way of looking at your data
  – More efficient than looking up gene-by-gene

• Computational analysis
  – Overrepresentation analysis
  – Network analysis
PathVisio

- http://www.pathvisio.org/downloads/
- PathVisio is a free open-source biological pathway analysis software that allows you to **draw, edit and analyze biological pathways**.
- Direct down- and upload to WikiPathways.org via WikiPathways App
Biological Context

• Statistical results:
  – 1,300 genes are significantly regulated after treatment with X

• Biological Meaning:
  – Is a certain biological pathway activated or deactivated?
  – Which genes in these pathway are significantly changed?
How to use PathVisio

1. Prepare your data
2. Import your data
3. Find enriched pathways
4. Create a visualization
5. Export pathway
1. Prepare your data!

File Format

- PathVisio accepts **Tab delimited text files**
- Prepare and export from Excel

![Excel screenshot with file name GSE12817_anova.txt and save as type Text (Tab delimited) (*.txt).]
File Format

- Export from R

```r
write.table(myTable, file = txtFile,
            col.names = NA, sep = "\t", quote = FALSE, na = "NaN")
```
Identifier Systems

PathVisio accepts many identifier systems:

- Probes
  - Affymetrix, Illumina, Agilent,...

- Genes and Proteins
  - Entrez Gene, Ensembl, UniProt, HUGO,...

- Metabolites
  - ChEBI, HMDB, PubChem,...
2. Import Expression Data

Choose file locations

Input file: /omas/PathVisio-Data/expression datasets/GSE12817/GSE12817_anova.txt
Output file: /has/PathVisio-Data/expression datasets/GSE12817/GSE12817_anova.png
Gene database: /home/thomas/PathVisio-Data/gene databases/Rn_Derby_20090508.pgdb
Identifier mapping database

Your data

Entrez Gene

5326
153
4357
65543
2094
90218
...

A pathway

Load BridgeDB files
ID mapping database

• Download from
  http://www.pathvisio.org/downloads/download-bridgedbs/

• OR
  http://www.bridgedb.org/mapping-databases/

32 species supported
Identifier and System Code

---

**Choose column types**

- **Select primary identifier column:** ENTREZG_ID
- **Select a column to specify system code:** LogFC
- **Use the same system code for all rows**

---

### Table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ENTREZG_ID</td>
<td>0.5919216625</td>
<td>6.63933242...</td>
</tr>
<tr>
<td>2</td>
<td>5791</td>
<td>0.5806104979</td>
<td>0.0000004681</td>
</tr>
<tr>
<td>3</td>
<td>1318</td>
<td>3.0834714719</td>
<td>4.78608750...</td>
</tr>
<tr>
<td>4</td>
<td>3290</td>
<td>0.7155023711</td>
<td>4.87264987...</td>
</tr>
<tr>
<td>5</td>
<td>6717</td>
<td>0.7777755536</td>
<td>7.07682042...</td>
</tr>
<tr>
<td>6</td>
<td>29940</td>
<td>0.7781936177</td>
<td>8.03384940...</td>
</tr>
<tr>
<td>7</td>
<td>51762</td>
<td>-1.3223803...</td>
<td>1.01799150...</td>
</tr>
<tr>
<td>8</td>
<td>6653</td>
<td>-1.60943715...</td>
<td>1.00000000...</td>
</tr>
</tbody>
</table>

---

**Fill in correct database!!**
Exception File

Perform import

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>X:1367454 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>12</td>
<td>X:1367462 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>21</td>
<td>X:1367471 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>24</td>
<td>X:1367474 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>25</td>
<td>X:1367475 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>30</td>
<td>X:1367480 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>34</td>
<td>X:1367484 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>35</td>
<td>X:1367485 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>39</td>
<td>X:1367489 at Could not look up this identifier in the synonym database</td>
</tr>
<tr>
<td>40</td>
<td>X:1367490 at Could not look up this identifier in the synonym database</td>
</tr>
</tbody>
</table>
3. Find „enriched“ pathways by applying pathway statistics

**Question:**
- Does the small circle have a higher percentage of changed genes than the large circle?
- Is this difference significant?
Calculate Z-scores

- The Z-score can be used as a measure for how much a subset of genes is different from the rest.

\[
z\text{score} = \frac{(r - n \frac{R}{N})}{\sqrt{n \frac{R}{N}(1 - \frac{R}{N})(1 - \frac{n-1}{N-1})}}
\]

- \(r\) = changed genes in Pathway
- \(n\) = total genes in Pathway
- \(R\) = changed genes
- \(N\) = total genes

Other enrichment calculation methods
Ackermann M et al., A general modular framework for gene set enrichment analysis, BMC bioinformatics, 2009
Z-score

• The Z-score is a ranking method.
  
  – High Z-score → selection is very different from the rest of the dataset
  
  – Z-score = 0 → selection is not different at all
Define criterion and select pathway collection

\[ ([\text{LogFC}] < -1 \text{ OR } [\text{LogFC}] > 1) \text{ AND } [\text{P.Value}] < 0.05 \]
### Z-score Calculation

\[ r = \text{changed genes in Pathway} \]
\[ n = \text{total genes in Pathway} \]

<table>
<thead>
<tr>
<th>Pathway</th>
<th>pos</th>
<th>meas</th>
<th>total</th>
<th>%</th>
<th>Z Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Factors</td>
<td>29</td>
<td>50</td>
<td>50</td>
<td>58.00%</td>
<td>4.55</td>
</tr>
<tr>
<td>mRNA processing</td>
<td>58</td>
<td>125</td>
<td>131</td>
<td>46.40%</td>
<td>4.36</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>43</td>
<td>89</td>
<td>94</td>
<td>48.31%</td>
<td>4.07</td>
</tr>
<tr>
<td>DNA Replication</td>
<td>23</td>
<td>42</td>
<td>49</td>
<td>54.76%</td>
<td>3.70</td>
</tr>
<tr>
<td>G1 to S cell cycle control</td>
<td>33</td>
<td>67</td>
<td>69</td>
<td>49.25%</td>
<td>3.69</td>
</tr>
<tr>
<td>DNA damage response</td>
<td>32</td>
<td>67</td>
<td>70</td>
<td>47.76%</td>
<td>3.42</td>
</tr>
<tr>
<td>TNF-alpha/NF-κB Signaling Pathway</td>
<td>71</td>
<td>186</td>
<td>188</td>
<td>38.17%</td>
<td>2.83</td>
</tr>
<tr>
<td>Tamoxifen metabolism</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>100.00%</td>
<td>2.71</td>
</tr>
<tr>
<td>FAS pathway and Stress induction of HSP...</td>
<td>18</td>
<td>38</td>
<td>43</td>
<td>47.37%</td>
<td>2.51</td>
</tr>
<tr>
<td>TGF-beta Receptor Signaling Pathway</td>
<td>56</td>
<td>151</td>
<td>152</td>
<td>37.09%</td>
<td>2.24</td>
</tr>
<tr>
<td>Wnt Signaling Pathway NetPath</td>
<td>42</td>
<td>109</td>
<td>110</td>
<td>38.53%</td>
<td>2.23</td>
</tr>
<tr>
<td>Androgen Receptor Signaling Pathway</td>
<td>41</td>
<td>112</td>
<td>114</td>
<td>36.61%</td>
<td>1.80</td>
</tr>
<tr>
<td>Delta–Notch Signaling Pathway</td>
<td>32</td>
<td>85</td>
<td>85</td>
<td>37.65%</td>
<td>1.78</td>
</tr>
<tr>
<td>IL-7 Signaling Pathway</td>
<td>18</td>
<td>44</td>
<td>45</td>
<td>40.91%</td>
<td>1.75</td>
</tr>
<tr>
<td>Keap1–Nrf2</td>
<td>7</td>
<td>14</td>
<td>16</td>
<td>50.00%</td>
<td>1.73</td>
</tr>
<tr>
<td>Glucuronidation</td>
<td>10</td>
<td>22</td>
<td>41</td>
<td>45.45%</td>
<td>1.71</td>
</tr>
<tr>
<td>Estrogen metabolism</td>
<td>8</td>
<td>17</td>
<td>33</td>
<td>47.06%</td>
<td>1.64</td>
</tr>
<tr>
<td>DNA damage response (SubATM dataset)</td>
<td>13</td>
<td>32</td>
<td>45</td>
<td>28.89%</td>
<td>1.33</td>
</tr>
</tbody>
</table>
Z-score Calculation

Rows in data (N): 3607
Rows meeting criterion (R): 1046

<table>
<thead>
<tr>
<th>Pathway</th>
<th>position</th>
<th>measured</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Factors</td>
<td>29</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>mRNA processing</td>
<td>58</td>
<td>125</td>
<td>131</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>43</td>
<td>89</td>
<td>94</td>
</tr>
<tr>
<td>DNA Replication</td>
<td>23</td>
<td>42</td>
<td>49</td>
</tr>
<tr>
<td>G1 to S cell cycle control</td>
<td>33</td>
<td>67</td>
<td>69</td>
</tr>
<tr>
<td>DNA damage response</td>
<td>32</td>
<td>67</td>
<td>70</td>
</tr>
<tr>
<td>TNF-alpha/NF-kB Signaling Pathway</td>
<td>71</td>
<td>186</td>
<td>188</td>
</tr>
<tr>
<td>Tamoxifen metabolism</td>
<td>3</td>
<td>3</td>
<td>31</td>
</tr>
<tr>
<td>FAS pathway and Stress induction of HSP</td>
<td>18</td>
<td>38</td>
<td>43</td>
</tr>
<tr>
<td>TGF-beta Receptor Signaling Pathway</td>
<td>56</td>
<td>151</td>
<td>157</td>
</tr>
</tbody>
</table>
4. Create a Visualization

Activate visualization options

Add/Remove Visualizations

Expression as numerical value: Display a numerical value next to a DataNode

Expression as color: Color DataNodes by their expression value

Text label: Draws a label
Visualizations

- Gradient based
  - Fold-change

- Rule based
  - Significant genes
Color by Data Values
Color Set based on Criterion
Color Set based on Gradient
Gradient based
Rule based
5. Export Pathway

- Export to image formats

![Diagram showing file export options: SVG, PNG, PDF]
Which pathways do you expect to change when cells/tissues are exposed to nanoparticles?

- Oxidative stress
- DNA reparation
- Apoptosis
- Necrosis
- Fibrosis
- Inflammation
- Cytoskeleton
- Metal homeostasis
- Cholesterol metabolism
6. Limitations and pitfalls

• Tissue – cells specific gene expression
  – PathVisio tissue analyzer

• Snap shot
  – RNA half-life
  – Transcription onset
  – Sample preparation

• Dependent on database (+) or (-)
We want you to know:

• Know the basic biological molecules DNA-RNA-Protein and how they interact
• Have an idea about basic protein reactions (conversion of metabolites, signaling)
• Know about the most commonly investigated effects of nanoparticles *in vitro* and *in vivo*
• Recognize these effects in biological pathways
  – Know that typical affected pathways are oxidative stress, apoptosis, metal ion response
• Know about the variety of omics data and how to use it
• Have heard about a variety of tools and methods to assess the effects on pathway/system level: omics data, especially microarray and RNA-seq
• Know the limitations and pitfalls of omics data/systems biology analysis
• Know why databases are useful for biologic research
  – Remembers some of the databases for single entities and pathways
• Know about the basics of semantic web and data integration using ontologies
• Have heard about data repositories like ArrayExpress, GEO, eNanoMapper
Acknowledgements

Thanks for slides and support:

Chris T. Evelo
Egon Willighagen
Susan Coort
Lars Eijssen
Martina Summer-Kutmon
Andra Waagmeester