# Software tools in bioinformatics

## Introduction

October 2019

UNIVERSITÄT BIELEFELD

Olga Zolotareva
Konstantinos Tzanakis

# Goals of the seminar

- reading, writing and understanding of scientific texts
- composition of a short written report (~8 pages)
  1. briefly explain the problem solved by a selected tool
  2. describe the algorithm(s) implemented in this tool
  3. explain the choice of the tool among its analogs
- oral presentation (~25min)
- peer-reviewing

# Areas of study

Olga

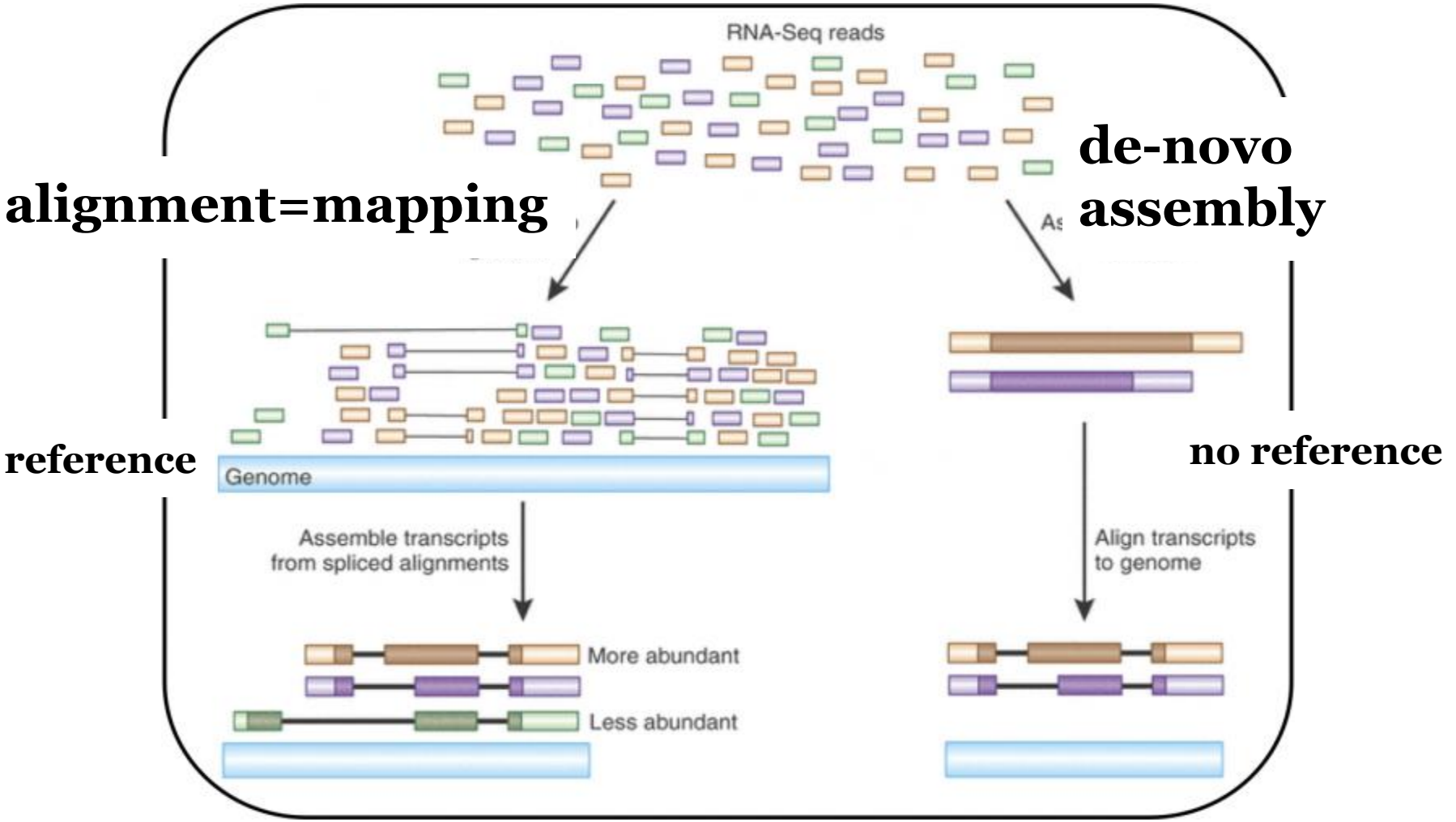
- Biclustering
- Read Alignment
- SNV Calling
- Expression Quantification and Differential Expression Analysis
- Gene Prioritization

Kostas

- Noise filtering and Baseline Correction
- Peak detection
- Peak alignment
- Metabolite identification
- Metabolomics data pre-treatment (normalization, scaling, transformation)

| Topic | Review(s) | Tools |
|---|---|---|
| Noise filtering and Baseline Correction | Piasecka et al., 2019<br>Spicer et al., 2017<br>Lazar et al., 2015 | OpenMs, XCMS, Workflow4Metabolomics, MeltDB, MetAlign |
| Peak detection | | |
| Metabolite identification | | |
| Metabolomics data pre-treatment | | |
| Peak alignment | Lange et al., 2008 | XCMS, OpenMs, MZmine |
| De novo transcriptome assembly | Hölzer & Marz, 2019 | Trinity, SPAdes |
| Read Mapping | Canzar & Salzberg, 2015<br>Reinert et al., 2015 | bwa, bowtie2, HISAT, STAR, GMAP |
| SNV Calling | Xu 2018 | VarScan2, Strelka, MuTect, DeepVariant |
| Expression Quantification and DE analysis | Lowe et al., 2017<br>Finotello & Di Camillo, 2016 | kallisto, limma-voom, DESeq2, EdgeR |
| Biclustering | Pontes et al., 2015 | QUBIC, DeBi, COALESCE, FABIA |
| Gene prioritization | Zolotareva & Kleine 2019 | Endeavour, PhenoRank, pBrit |

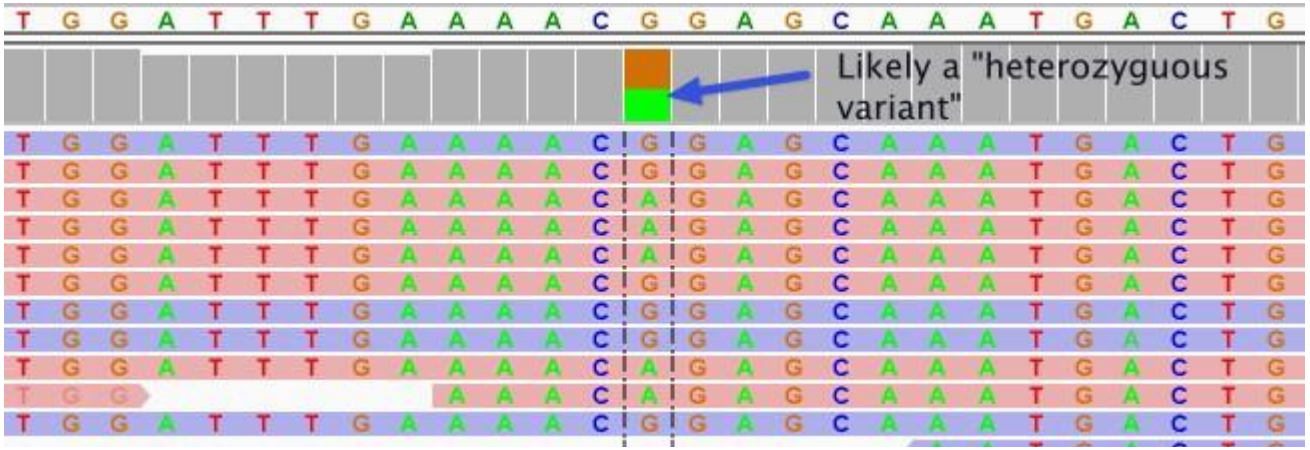
# NGS data analysis




*from:* https://www.ebi.ac.uk/training/online/course/

# Downstream analysis

## Variant calling

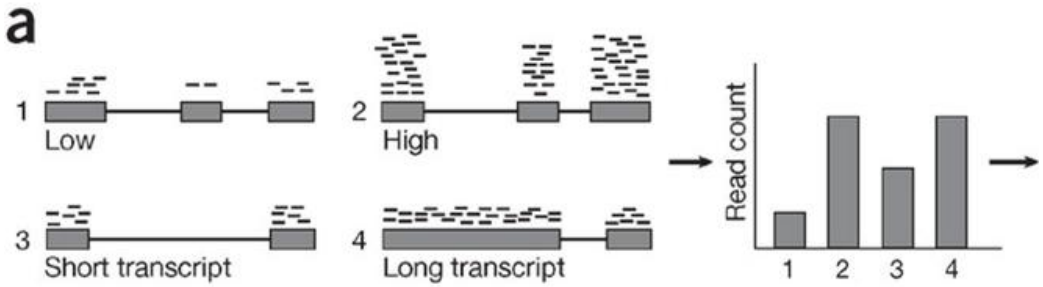



***from:*** https://towardsdatascience.com/simple-convolution-neural-network-for-genomic-variant-calling-with-tensorflow-c085dbc2026f

- given **mapped reads**, identify sequence **variants**

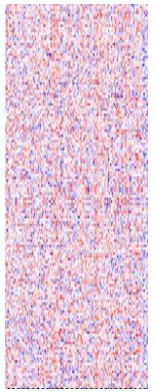
# Downstream analysis

- **Expression quantification**




***from:*** Garber et al. 2011

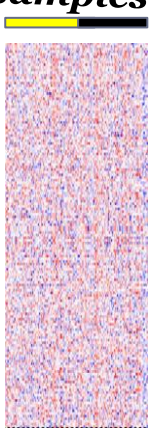
***N samples***



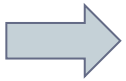
***~10⁴ genes***

- **Identification of Differentially Expressed genes**

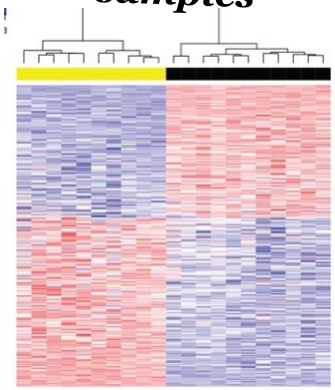
***samples***



***~10⁴ genes***



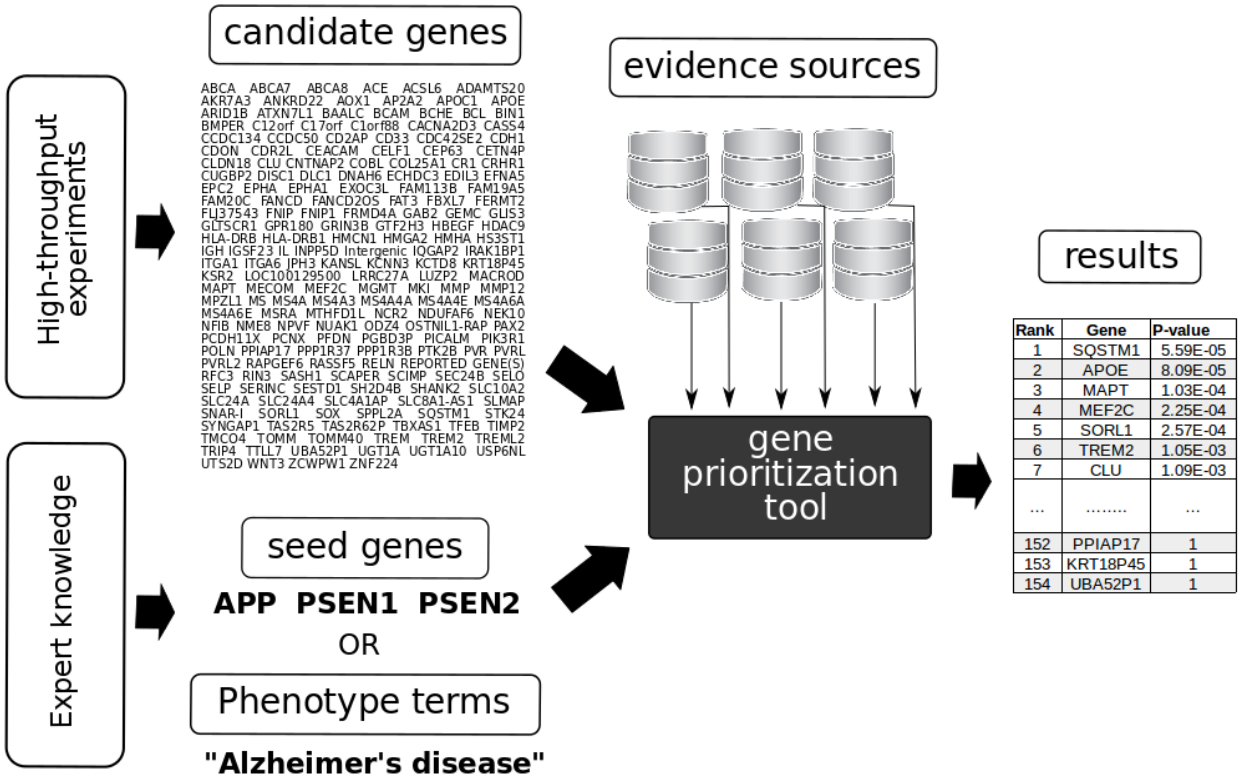
***samples***



***genes***

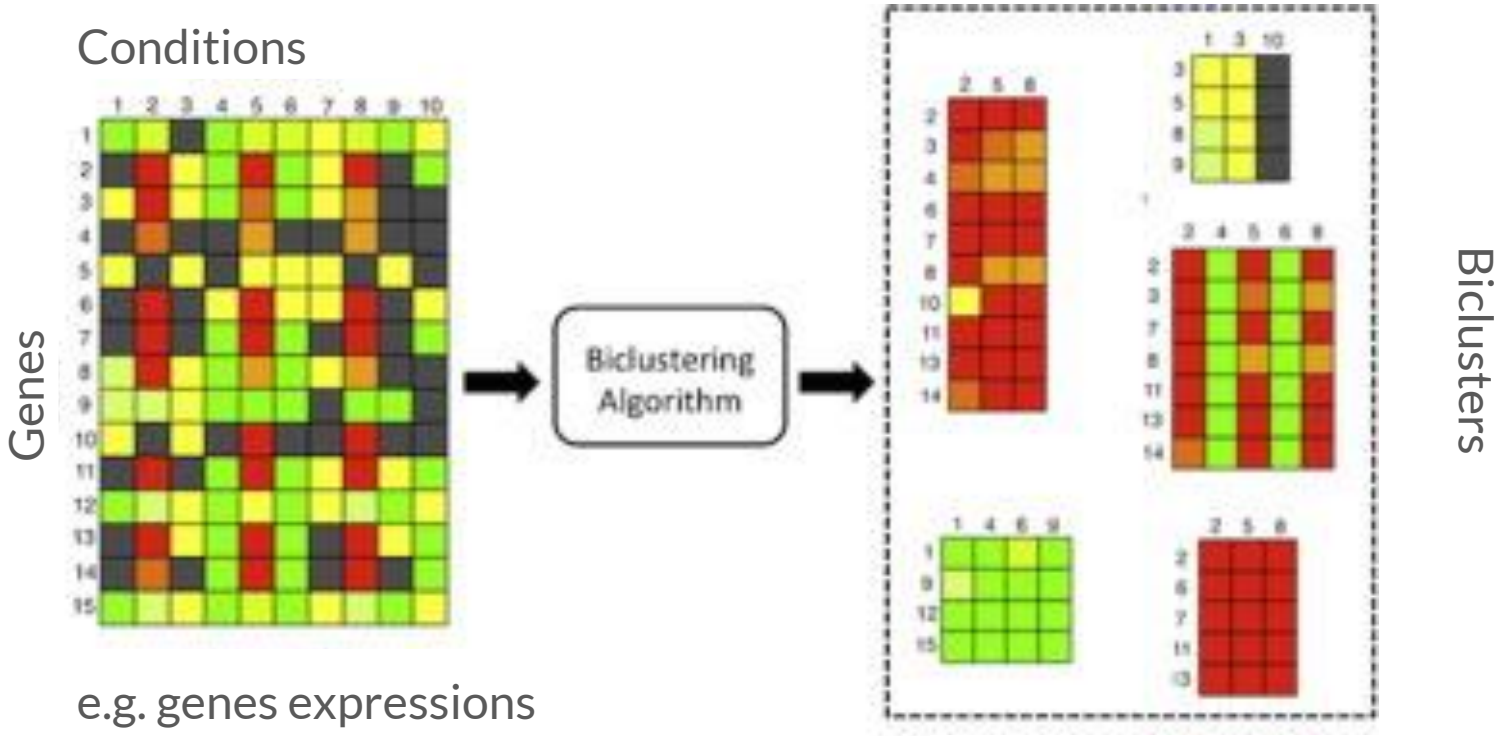
***from:*** *www.rna-seqblog.com*

# Gene Prioritization




- **arrange** candidate **genes** in order of their **relevance to the phenotype** decreasing, given **prior knowledge** about genes and phenotypes

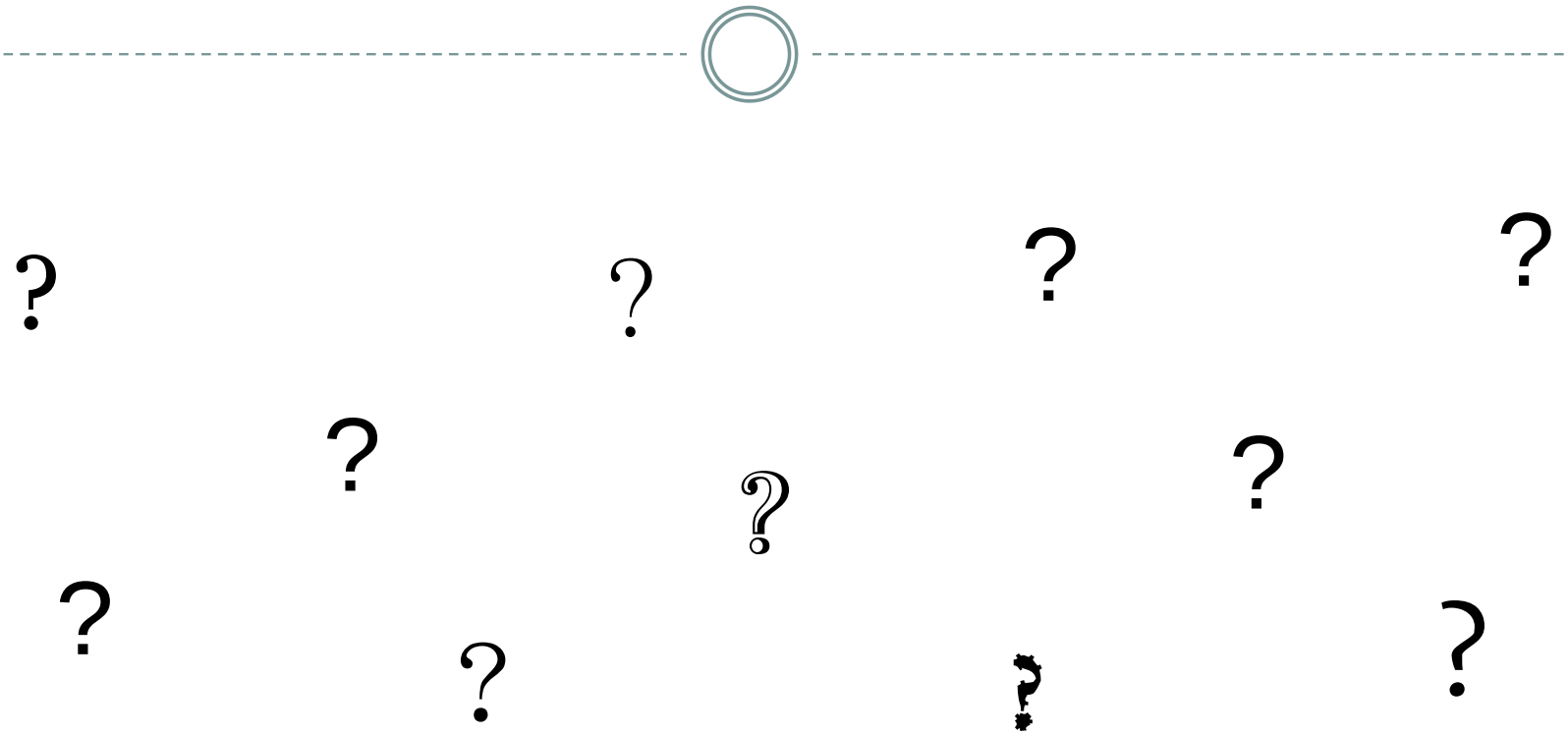
# Biclustering



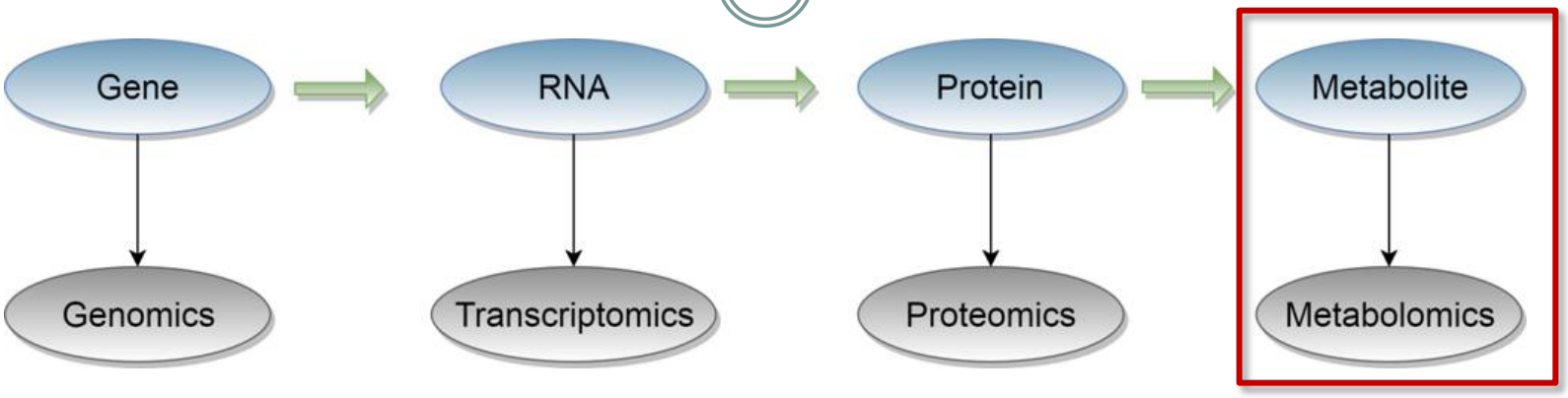

*from: Pontes et al. 2015*

- searching for **subgroups of rows** (genes) demonstrating a **pattern** in a **subgroup of columns** (samples)

# Metabolomics



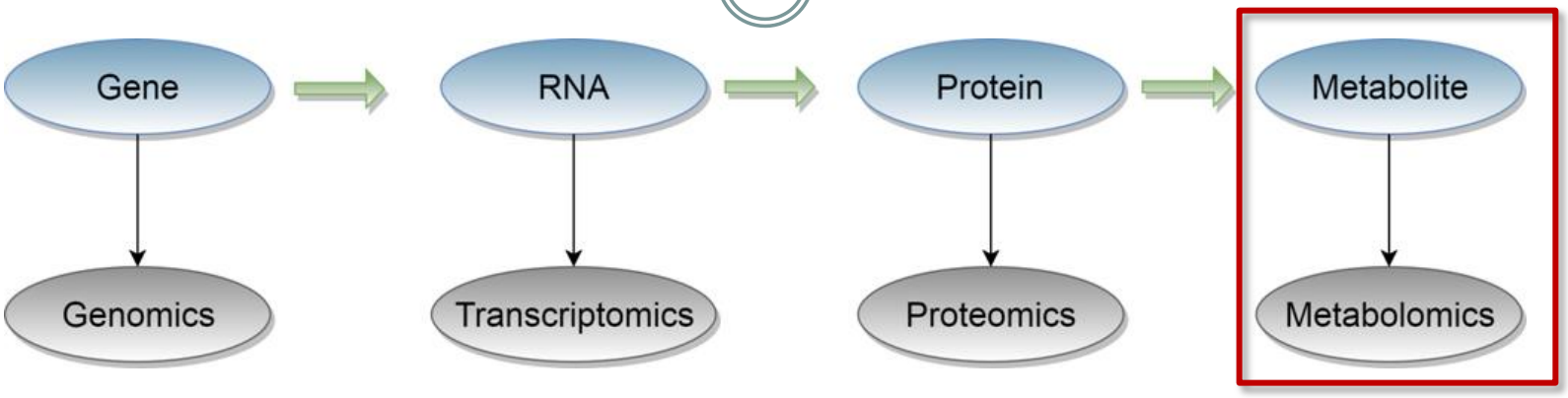
# - Omics




**What is metabolomics?**

Study and analysis of the quantitative and qualitative collection of all metabolites in the cell.

# - Omics




**What is metabolomics?**

Study and analysis of the quantitative and qualitative collection of all metabolites in the cell.

**What is metabolites?**

Intermediates and products of metabolism that refer to small molecules (<1.5kDa).

**Importance**

- Disease/Drug Biomarkers
- Plant Biotechnology and Crop Breeding

**Biological Sample**

## Biological Sample

## Data Acquisition

- Separation: LC or GC
- Detection: MS

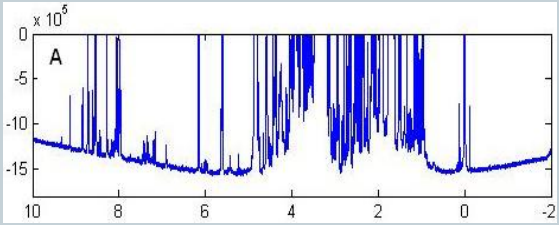
*Xi, Yuanxin & M Rocke, David. (2008). Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. BMC bioinformatics. 9. 324. 10.1186/1471-2105-9-324.*

## Biological Sample

## Data Acquisition

- Separation: LC or GC
- Detection: MS

### Mass Spectrum



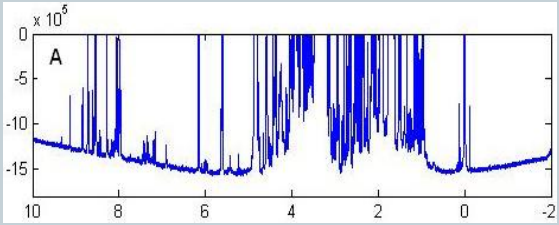

*Xi, Yuanxin & M Rocke, David. (2008). Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. BMC bioinformatics. 9. 324. 10.1186/1471-2105-9-324.*

## Biological Sample

## Data Acquisition

- Separation: LC or GC
- Detection: MS

### Mass Spectrum




## Processing

### Spectral Processing *(Quantification)*

- Noise Filtering and Baseline Correction
- Peak Detection
- Peak Alignment

Features

Samples

Peaks

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1165.434 | 280.2641 | 1165.56 | 1165.617 | 280.2645 | 1323.76 | 1165.544 | 280.2646 | 1288.39 |
| 68.92787 | 204.0566 | 79.4584 | 68.8607 | 204.0573 | 117.571 | 68.89248 | 204.0564 | 101.048 |
| 1146.27 | 165.1288 | 89.4447 | 1147.402 | 165.1278 | 100.562 | 1145.886 | 165.1284 | 99.7233 |
| 1065.567 | 522.3657 | 1926.19 | 1064.956 | 522.3672 | 8189.13 | 1067.06 | 522.3668 | 8523.84 |
| 662.0922 | 585.2739 | 78.4788 | 662.181 | 585.2733 | 83.6998 | 662.1596 | 585.2749 | 37.3213 |
| 1086.891 | 228.2309 | 9362.18 | 1087.255 | 228.2316 | 9373.18 | 1086.956 | 228.232 | 7959.34 |
| 1228.891 | 256.1717 | 186.113 | 1230.075 | 256.1681 | 112.88 | 1229.034 | 256.1734 | 141.097 |
| 948.4953 | 361.2741 | 501.383 | 948.2905 | 361.2748 | 437.427 | 948.4788 | 361.2744 | 270.058 |
| 782.7708 | 286.1423 | 5178.75 | 783.1694 | 286.1431 | 4975.11 | 782.8954 | 286.1428 | 5055.95 |
| 64.16562 | 258.8998 | 279.911 | 64.34053 | 258.8996 | 289.793 | 64.61098 | 258.9001 | 303.655 |
| 65.08186 | 96.92269 | 229.869 | 65.4986 | 96.92273 | 157.423 | 65.53273 | 96.92288 | 262.262 |
| 983.4194 | 454.2942 | 111.342 | 983.4808 | 454.2936 | 161.106 | 983.6665 | 454.2957 | 120.46 |
| 984.3786 | 464.3131 | 160.855 | 983.4808 | 464.3152 | 129.868 | 982.943 | 464.3183 | 120.737 |

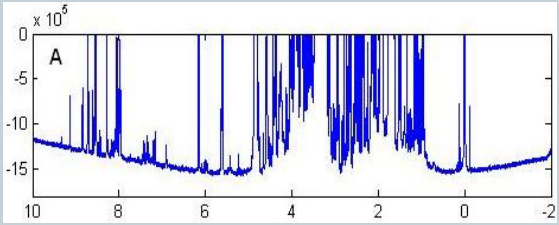
*Xi, Yuanxin & M Rocke, David. (2008). Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. BMC bioinformatics. 9. 324. 10.1186/1471-2105-9-324.*

## Biological Sample

## Data Acquisition

- Separation: LC or GC
- Detection: MS

### Mass Spectrum




## Processing

### Spectral Processing *(Quantification)*

- Noise Filtering and Baseline Correction
- Peak Detection
- Peak Alignment

Features

Samples

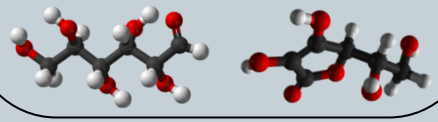
Peaks

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1165.434 | 280.2641 | 1165.56 | 1165.617 | 280.2645 | 1323.76 | 1165.544 | 280.2646 | 1288.39 |
| 68.92787 | 204.0566 | 79.4584 | 68.8607 | 204.0573 | 117.571 | 68.89248 | 204.0564 | 101.048 |
| 1146.27 | 165.1288 | 89.4447 | 1147.402 | 165.1278 | 100.562 | 1145.886 | 165.1284 | 99.7233 |
| 1065.567 | 522.3657 | 1926.19 | 1064.956 | 522.3672 | 8189.13 | 1067.06 | 522.3668 | 8523.84 |
| 662.0922 | 585.2739 | 78.4788 | 662.181 | 585.2733 | 83.6998 | 662.1596 | 585.2749 | 37.3213 |
| 1086.891 | 228.2309 | 9362.18 | 1087.255 | 228.2316 | 9373.18 | 1086.956 | 228.232 | 7959.34 |
| 1228.891 | 256.1717 | 186.113 | 1230.075 | 256.1681 | 112.88 | 1229.034 | 256.1734 | 141.097 |
| 948.4953 | 361.2741 | 501.383 | 948.2905 | 361.2748 | 437.427 | 948.4788 | 361.2744 | 270.058 |
| 782.7708 | 286.1423 | 5178.75 | 783.1694 | 286.1431 | 4975.11 | 782.8954 | 286.1428 | 5055.95 |
| 64.16562 | 258.8998 | 279.911 | 64.34053 | 258.8996 | 289.793 | 64.61098 | 258.9001 | 303.655 |
| 65.08186 | 96.92269 | 229.869 | 65.4986 | 96.92273 | 157.423 | 65.53273 | 96.92288 | 262.262 |
| 983.4194 | 454.2942 | 111.342 | 983.4808 | 454.2936 | 161.106 | 983.6665 | 454.2957 | 120.46 |
| 984.3786 | 464.3131 | 160.855 | 983.4808 | 464.3152 | 129.868 | 982.943 | 464.3183 | 120.737 |

### Identification

| Spectral Databases | Dataset Features |
|---|---|
| hmdb The Human Metabolome Database<br>METLIN | • Retention Time<br>• Relative Intensity<br>• m/z ratio |

Identified Metabolites



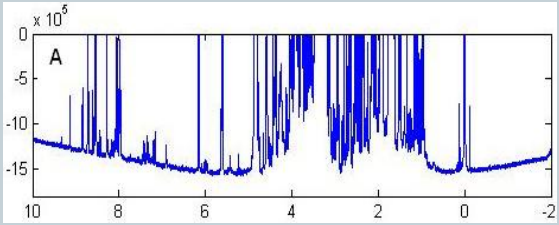
*Xi, Yuanxin & M Rocke, David. (2008). Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. BMC bioinformatics. 9. 324. 10.1186/1471-2105-9-324.*

## Biological Sample

## Data Acquisition

- Separation: LC or GC
- Detection: MS

### Mass Spectrum




## Processing

### Spectral Processing *(Quantification)*

- Noise Filtering and Baseline Correction
- Peak Detection
- Peak Alignment

Features

Samples

Peaks

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1165.434 | 280.2641 | 1165.56 | 1165.617 | 280.2645 | 1323.76 | 1165.544 | 280.2646 | 1288.39 |
| 68.92787 | 204.0566 | 79.4584 | 68.8607 | 204.0573 | 117.571 | 68.89248 | 204.0564 | 101.048 |
| 1146.27 | 165.1288 | 89.4447 | 1147.402 | 165.1278 | 100.562 | 1145.886 | 165.1284 | 99.7233 |
| 1065.567 | 522.3657 | 1926.19 | 1064.956 | 522.3672 | 8189.13 | 1067.06 | 522.3668 | 8523.84 |
| 662.0922 | 585.2739 | 78.4788 | 662.181 | 585.2733 | 83.6998 | 662.1596 | 585.2749 | 37.3213 |
| 1086.891 | 228.2309 | 9362.18 | 1087.255 | 228.2316 | 9373.18 | 1086.956 | 228.232 | 7959.34 |
| 1228.891 | 256.1717 | 186.113 | 1230.075 | 256.1681 | 112.88 | 1229.034 | 256.1734 | 141.097 |
| 948.4953 | 361.2741 | 501.383 | 948.2905 | 361.2748 | 437.427 | 948.4788 | 361.2744 | 270.058 |
| 782.7708 | 286.1423 | 5178.75 | 783.1694 | 286.1431 | 4975.11 | 782.8954 | 286.1428 | 5055.95 |
| 64.16562 | 258.8998 | 279.911 | 64.34053 | 258.8996 | 289.793 | 64.61098 | 258.9001 | 303.655 |
| 65.08186 | 96.92269 | 229.869 | 65.4986 | 96.92273 | 157.423 | 65.53273 | 96.92288 | 262.262 |
| 983.4194 | 454.2942 | 111.342 | 983.4808 | 454.2936 | 161.106 | 983.6665 | 454.2957 | 120.46 |
| 984.3786 | 464.3131 | 160.855 | 983.4808 | 464.3152 | 129.868 | 982.943 | 464.3183 | 120.737 |

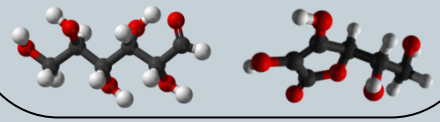
### Data pre-treatment

- Normalization
- Scaling • Transformation

### Identification

| Spectral Databases | Dataset Features |
|---|---|
| hmdb The Human Metabolome Database; METLIN | • Retention Time<br>• Relative Intensity<br>• m/z ratio |

Identified Metabolites



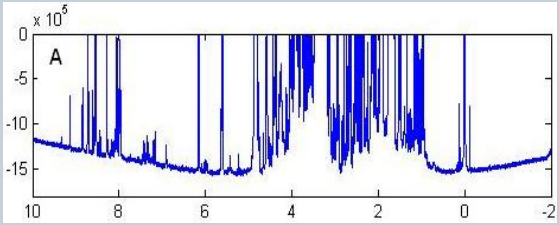
*Xi, Yuanxin & M Rocke, David. (2008). Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. BMC bioinformatics. 9. 324. 10.1186/1471-2105-9-324.*

## Biological Sample

## Data Acquisition

- Separation: LC or GC
- Detection: MS

### Mass Spectrum




## Statistical Analysis

**Univariate Analysis** | **Multivariate Analysis**

## Processing

### Spectral Processing *(Quantification)*

- Noise Filtering and Baseline Correction
- Peak Detection
- Peak Alignment

Features

Samples

Peaks

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1165.434 | 280.2641 | 1165.56 | 1165.617 | 280.2645 | 1323.76 | 1165.544 | 280.2646 | 1288.39 |
| 68.92787 | 204.0566 | 79.4584 | 68.8607 | 204.0573 | 117.571 | 68.89248 | 204.0564 | 101.048 |
| 1146.27 | 165.1288 | 89.4447 | 1147.402 | 165.1278 | 100.562 | 1145.886 | 165.1284 | 99.7233 |
| 1065.567 | 522.3657 | 1926.19 | 1064.956 | 522.3672 | 8189.13 | 1067.06 | 522.3668 | 8523.84 |
| 662.0922 | 585.2739 | 78.4788 | 662.181 | 585.2733 | 83.6998 | 662.1596 | 585.2749 | 37.3213 |
| 1086.891 | 228.2309 | 9362.18 | 1087.255 | 228.2316 | 9373.18 | 1086.956 | 228.232 | 7959.34 |
| 1228.891 | 256.1717 | 186.113 | 1230.075 | 256.1681 | 112.88 | 1229.034 | 256.1734 | 141.097 |
| 948.4953 | 361.2741 | 501.383 | 948.2905 | 361.2748 | 437.427 | 948.4788 | 361.2744 | 270.058 |
| 782.7708 | 286.1423 | 5178.75 | 783.1694 | 286.1431 | 4975.11 | 782.8954 | 286.1428 | 5055.95 |
| 64.16562 | 258.8998 | 279.911 | 64.34053 | 258.8996 | 289.793 | 64.61098 | 258.9001 | 303.655 |
| 65.08186 | 96.92269 | 229.869 | 65.4986 | 96.92273 | 157.423 | 65.53273 | 96.92288 | 262.262 |
| 983.4194 | 454.2942 | 111.342 | 983.4808 | 454.2936 | 161.106 | 983.6665 | 454.2957 | 120.46 |
| 984.3786 | 464.3131 | 160.855 | 983.4808 | 464.3152 | 129.868 | 982.943 | 464.3183 | 120.737 |

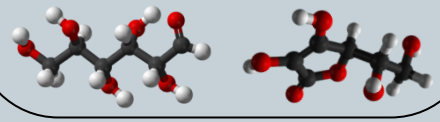
### Data pre-treatment

- Normalization
- Scaling • Transformation

### Identification

| Spectral Databases | Dataset Features |
|---|---|
| hmdb The Human Metabolome Database; METLIN | • Retention Time<br>• Relative Intensity<br>• m/z ratio |

Identified Metabolites



*Xi, Yuanxin & M Rocke, David. (2008). Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. BMC bioinformatics. 9. 324. 10.1186/1471-2105-9-324.*

# Information

Name: Olga Zolotareva
Office: D5-117
e-mail: olya.zolotareva@gmail.com

Name: Konstantinos Tzanakis
Office: U10-144
e-mail: ktzan@cebitec.uni-bielefeld.de

**Website**:

https://www.didy.uni-bielefeld.de/teaching/2019winter/softwaretoolsinbioinformatics

***Thank you ! ! !***