WikiPathways Summit 2018

Workshop 1: Introduction to
Pathway Enrichment Analysis and
Data Overlays for Non-Programmers

Martina Summer-Kutmon
Maastricht University

Overview

1. Introduction
2. Pathway analysis
3. Pathways as tools for network analysis
4. Workshop setup and dataset

WP Summit 2018 – Workshop 1

2

Overview

1. **Introduction**
2. Pathway analysis
3. Pathways as tools for network analysis
4. Workshop setup and dataset
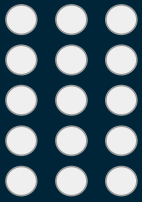
WP Summit 2018 – Workshop 1

3

Introduction

Quantitative measurements
Isolated data points

WP Summit 2018 – Workshop 1   4

Introduction

Comparative statistics
Isolated lists
Clustering
Isolated groups
Gene sets
Isolated lists
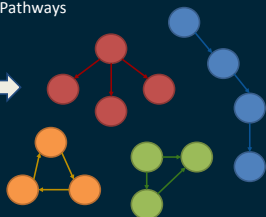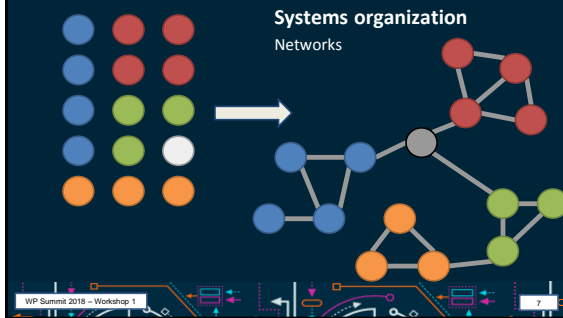
WP Summit 2018 – Workshop 1   5

Introduction

Functional organization
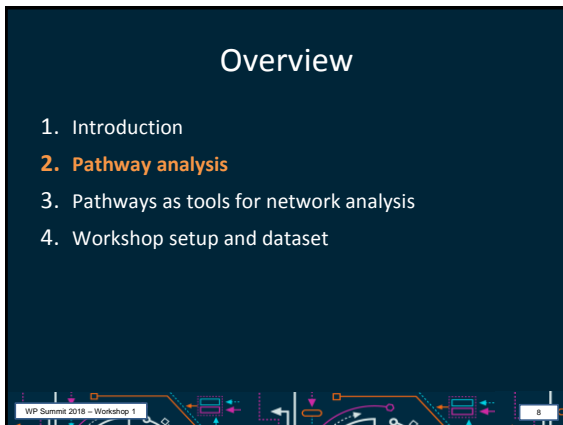Pathways

WP Summit 2018 – Workshop 1   6

Introduction

Systems organization
Networks

WP Summit 2018 – Workshop 1
7

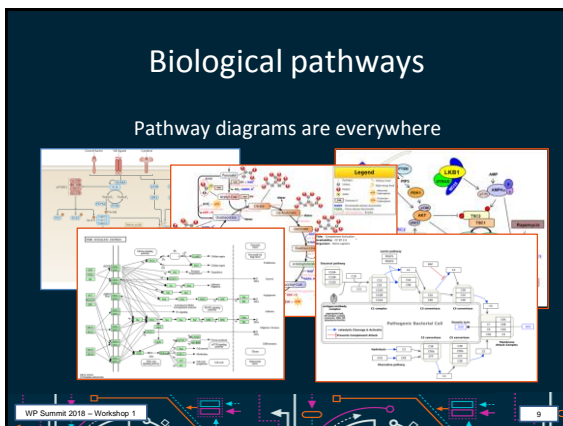## Overview

1. Introduction
2. **Pathway analysis**
3. Pathways as tools for network analysis
4. Workshop setup and dataset

WP Summit 2018 – Workshop 1
8

Biological pathways

Pathway diagrams are everywhere

WP Summit 2018 – Workshop 1
9

Biological pathways

Pathways are found everywhere

⬇

Utility to biologists as conceptual models is obvious

⬇

<u>If modeled properly</u> - immensely useful for computational analysis and interpretation of large-scale experimental data
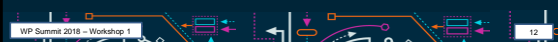
WP Summit 2018 – Workshop 1

10

Why pathway analysis?

- "A picture is worth a thousand words."
  - Intuitive and simple
  - Puts data into a biological context
  - More efficient than looking up single gene information

  - Reduces complexity → several hundred pathways instead of thousands of genes
  - Higher explanatory power than a simple gene list
  - Visual representation
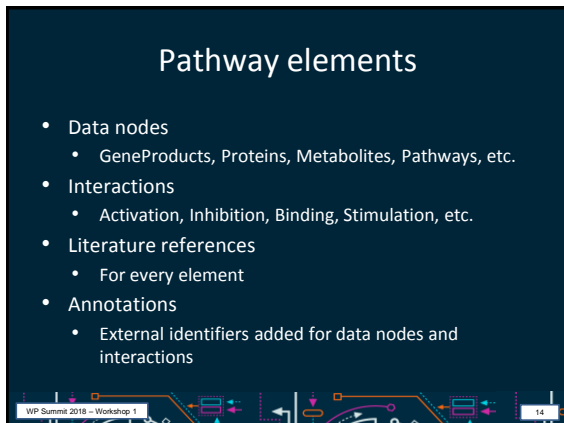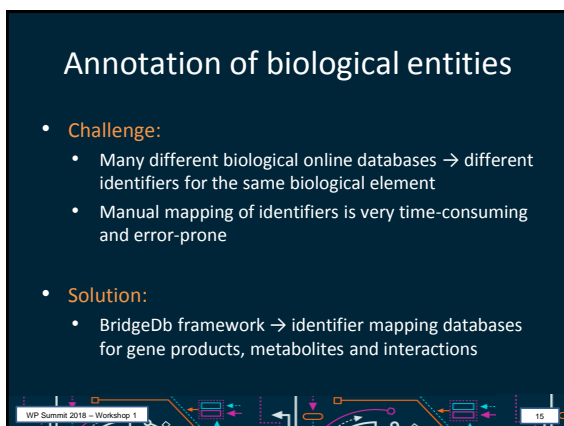
WP Summit 2018 – Workshop 1

11
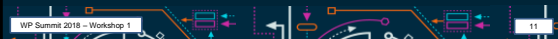
PathVisio

- Pathway editor, analysis and visualization toolbox
- Java standalone desktop application
- Functionality
  - Design biological pathway models ⟵ **Workshop 3**
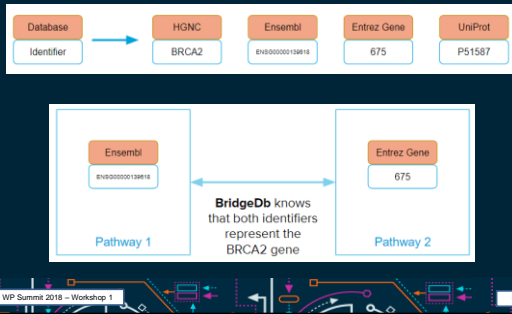 - Visualize your experimental data
  - Perform pathway statistics **Workshop 1**

WP Summit 2018 – Workshop 1

12

PathVisio

Toolbar: drawing objects, layout, view, visualization

Pathway display area

Sidepanel: Drawing, Editing, Backpage, Search, Legend

Status bar: Loaded databases, data sets

WP Summit 2018 – Workshop 1   13

Pathway elements

- Data nodes
  - GeneProducts, Proteins, Metabolites, Pathways, etc.
- Interactions
  - Activation, Inhibition, Binding, Stimulation, etc.
- Literature references
  - For every element
- Annotations
  - External identifiers added for data nodes and interactions

WP Summit 2018 – Workshop 1   14

Annotation of biological entities

- Challenge:
  - Many different biological online databases → different identifiers for the same biological element
  - Manual mapping of identifiers is very time-consuming and error-prone

- Solution:
  - BridgeDb framework → identifier mapping databases for gene products, metabolites and interactions

WP Summit 2018 – Workshop 1   15

Annotation of biological entities
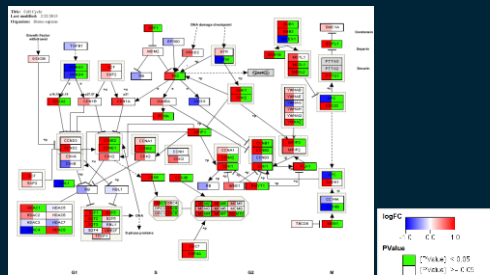

WP Summit 2018 – Workshop 1   16

Data visualization

- Data can be visualized on data nodes, interactions
- Color gradient and color-rules

- Basic and advanced options
  - Multi-omics data visualization
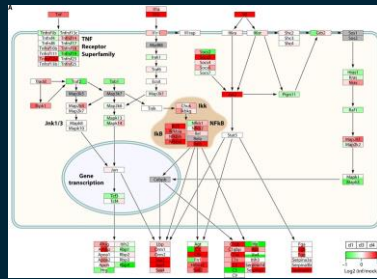 - Time-series data visualization
  - …

WP Summit 2018 – Workshop 1   17

Data visualization

Transcriptomics cancer dataset visualized on the human Cell Cycle Pathway:
https://www.wikipathways.org/instance/WP179

WP Summit 2018 – Workshop 1   18

Slide 19

Data visualization

Time-series data visualization (virus infection over multiple days)
Into the eye of the cytokine storm. https://doi.org/10.1128/MMBR.05015-11

WP Summit 2018 – Workshop 1

19

Slide 20

Pathway statistics

Gene expression

Pathway database

Differentially expressed genes
⇩
Z-score per pathway

PathVisio

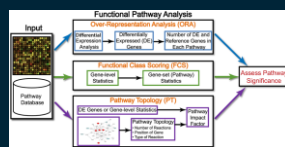Ranking of pathways based on Z-score
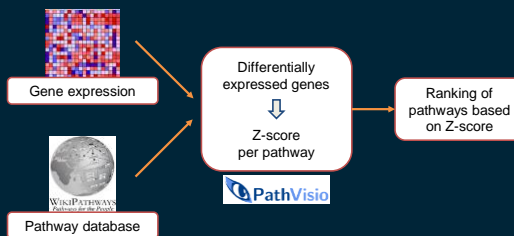
WP Summit 2018 – Workshop 1

20

Slide 21

Pathway statistics

- Overrepresentation analysis (ORA)
- Functional Class Scoring (FCS)
- Pathway Topology Based (PT)

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Computational Biology, 8(2), e1002375.

WP Summit 2018 – Workshop 1

21

Pathway statistics

- **Overrepresentation analysis:**
  - Input list → e.g. significantly changed genes
  - Background list → e.g. all measured genes
  - Statistical test → e.g. Fisher's exact test (hypergeometric test)

$$Z\text{-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

  - Z-Score is calculated for each pathway
    - Results in ranked list of pathways
  - Four variables in the formula: N, R, n, r
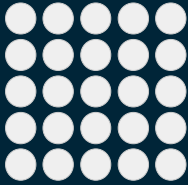
WP Summit 2018 – Workshop 1   22

Z-Score ORA

$$Z\text{-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

N = 25
background list (total number of measured genes in experiment)

WP Summit 2018 – Workshop 1   23

Z-Score ORA

$$Z\text{-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

N = 25
background list (total number of measured genes in experiment)

R = 9
input list (number of changed genes in experiment)

WP Summit 2018 – Workshop 1   24

Slide 25

Z-Score ORA

$$Z\text{-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

N = 25
background list (total number of measured genes in experiment)

R = 9
input list (number of changed genes in experiment)

n = 9
total number of genes in pathway

Pathway X

WP Summit 2018 – Workshop 1   25

Slide 26

Z-Score ORA

$$Z\text{-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

N = 25
background list (total number of measured genes in experiment)

R = 9
input list (number of changed genes in experiment)

n = 9
total number of genes in pathway

r = 6
number of changed genes in pathway

Pathway X

WP Summit 2018 – Workshop 1   26

Slide 27

Z-Score ORA

$$Z\text{-score} = \frac{(r - n\frac{R}{N})}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

N = 25
background list (total number of measured genes in experiment)

R = 9
input list (number of changed genes in experiment)

n = 9
total number of genes in pathway

r = 6
number of changed genes in pathway

Pathway X

Z-Score for pathway X = 2.347

WP Summit 2018 – Workshop 1   27

Z-Score ORA

Be aware!!

- **What does the Z-Score tell you?**
  - Z-Score > 1.96
    - Significantly more genes than expected are changed in the pathway → altered pathways in the experiment (different between the groups)
  - **Z-Score = 0**
    - Distribution of changed genes in the pathway is the same as in the complete dataset
  - **Z-Score = < -1.96**
    - Significantly less genes than expected are changed in the pathway → very stable pathway (not affected in experiment)

WP Summit 2018 – Workshop 1   28

Z-Score ORA

- **Be aware!**
  - ORA and FCS **do not** take pathway topology into account!
  - You don't know yet where the changes occur in the pathway.
  - Always look at the pathway diagrams and study the changes to make the right conclusions!

WP Summit 2018 – Workshop 1   29

Pathway analysis methods

- **Overrepresentation analysis:**
  - Ranked list of pathways

Up-regulated pathways (log2FC > 2, p-value < 0.05)	Z-score	Perm. p-value
Cell Cycle	6.12	0.001
G1 to S cell cycle control	4.26	0.002
Synaptic Vesicle Pathway	3.89	0.001
DNA Damage Response	3.88	0.002
ATM Signaling Pathway	3.80	0.001

Down-regulated pathways (log2FC < -2, p-value < 0.05)	Z-score	Perm. p-value
Complement and Coagulation Cascades	5.87	0.001
Complement Activation	5.84	0.001
Adipogenesis	5.49	0.001
Differentiation of white and brown adipocyte	5.44	0.001
Triacylglyceride Synthesis	4.53	0.001

Same method is also used for Gene Ontology analysis

WP Summit 2018 – Workshop 1   30

Overview

1. Introduction
2. Pathway analysis
3. **Pathways as tools for network analysis**
4. Workshop setup and dataset

WP Summit 2018 – Workshop 1

31

WikiPathways app

- WikiPathways app
  - Load pathways as networks in Cytoscape
  - Use online database query in Cytoscape or open local GPML file
 - apps.cytoscape.org/apps/wikipathways

WP Summit 2018 – Workshop 1

32

WikiPathways app

- App supports two views:
  - Pathway view
    - complete visual appearance
    - ideal for data visualization
  - Network view
    - simplified networks without any of the graphical elements of the original pathway diagram
    - ideal for topological analyses, network merging and automatic layout

WP Summit 2018 – Workshop 1

33

WikiPathways app

Cytoscape 3.6.0+
WikiPathways app 3.3.0+
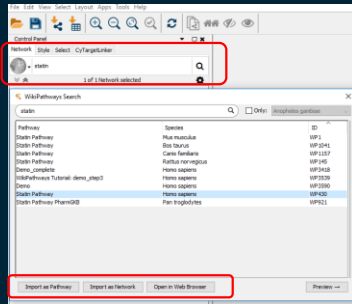
WP Summit 2018 – Workshop 1

34

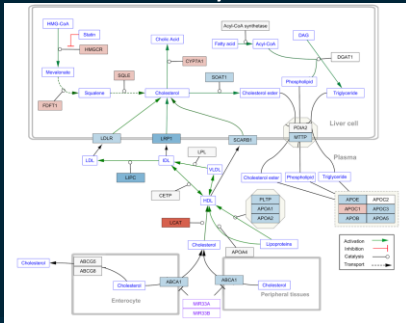

Pathway view
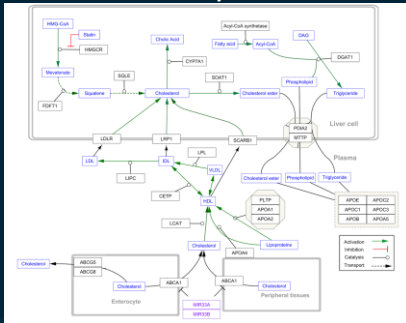
WP Summit 2018 – Workshop 1

35

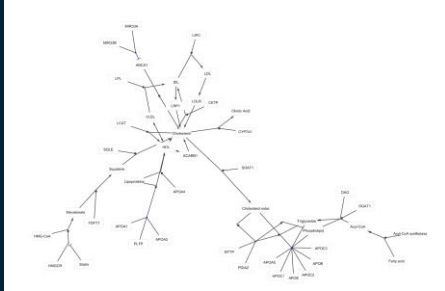Pathway view
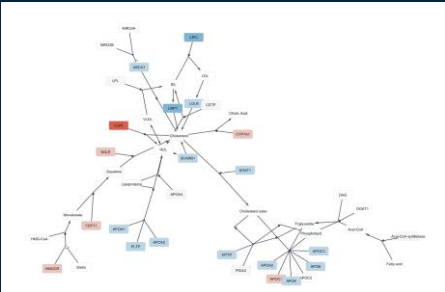
WP Summit 2018 – Workshop 1

36

Network view

Network view

Connecting pathways

Legend:
- Elements linking two or more pathways
- DNA replication
- Cell cycle
- Histone modifications
- G1 to S cell cycle control
- DNA damage response
- ATM signaling pathway
- Fluoropyrimidine Activity
- AhR signaling pathway
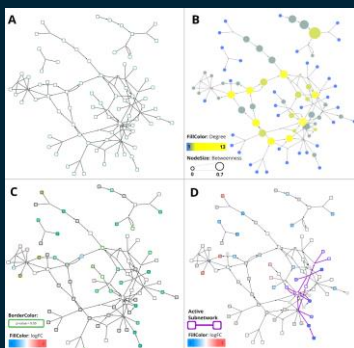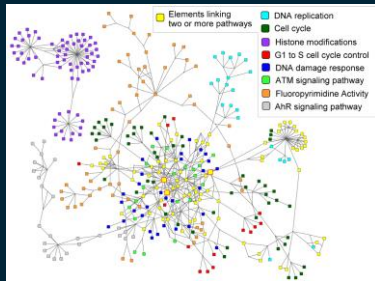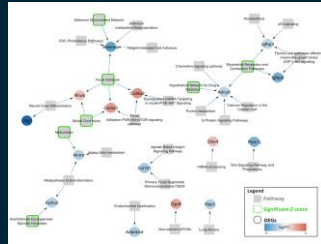
Kutmon, M. et al.
Integrative network-based analysis of mRNA and microRNA
expression in 1,25-dihydroxyvitamin D3-treated cancer cells
https://doi.org/10.1007/s12263-015-0484-0

WP Summit 2018 – Workshop 1

40

Pathway associations

- Extend molecular networks with pathway associations using CyTargetLinker app

CyTargetLinker: a flexible solution for
network extension
https://doi.org/10.12688/f1000research.14613.1

WP Summit 2018 – Workshop 1

41

Pathway associations

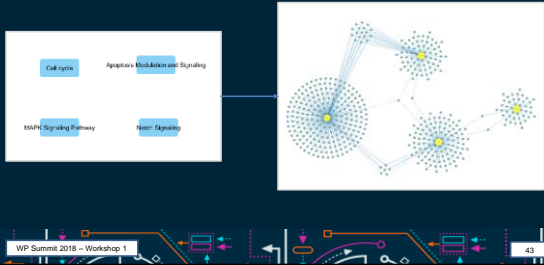- Visualize gene overlap between significant pathways with CyTargetLinker
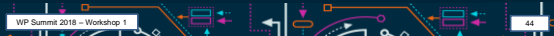

WP Summit 2018 – Workshop 1

42

14

Pathway associations

- Visualize gene overlap between significant pathways with CyTargetLinker

43

Cytoscape automation

- Available since the last major release (3.6.0+)
- Execute commands from within R or Python
- WikiPathways and CyTargetLinker app both have automation interface

44

Overview

1. Introduction
2. Pathway analysis
3. Pathways as tools for network analysis
4. **Workshop setup and dataset**

45
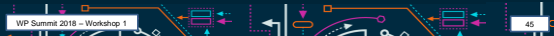
Workshop setup

- Step-by-step instructions
- Required data provided in a zip file
- https://github.com/PathVisio/tutorials/2018-WP-Summit

- Feel free to ask questions at any point
- Three instructors will help you out

Tina   Kristina   Denise

WP Summit 2018 – Workshop 1   46

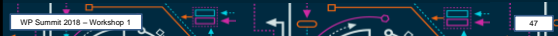## Workshop setup

- Data visualization on pathways
- Identifying affected pathways
- Using pathways as networks
- Extension of pathways with regulatory information
- Pathway-gene associations

WP Summit 2018 – Workshop 1   47

Dataset

Data retrieved from lung cancer subjects in TCGA

Primary tumor tissue sample

Solid tissue normal sample

RNA-sequencing

RNA-sequencing

Gene expression
Raw counts

Gene expression
Raw counts

Data pre-processing and statistical analysis with edgeR in R

Differential gene expression
Comparison between groups

WP Summit 2018 – Workshop 1   48

Dataset

GeneID	GeneName	log2FC	P.Value	adj.P.Value
ENSG00000230657	PRB4	13.27397592	0.003920605	0.097861917

- **GeneID** → identifier in online database
- **GeneName** → official gene symbol
- **log2FC** → log2 of fold change (ratio of the differences between cancer and healthy samples)
- **P.Value** → significance level of comparison
- **adj.P.Value** → corrected P.Value for multiple testing

WP Summit 2018 – Workshop 1   49

Dataset

- log2FC
  - Is the gene more or less expressed in the cancer tissue sample compared to the normal tissue sample?

Negative log2FC downregulation in cancer sample	**Positive log2FC** upregulation in cancer sample

Example
$log2FC = 1$ → $fold\ change = 2^1 = 2$
gene is twice as expression cancer
compared to healthy tissue sample

Why do we use the log2?
easier for interpretation – 0 = no chance,
1 is upregulation, -1 is downregulation

WP Summit 2018 – Workshop 1   50

Questions

https://github.com/PathVisio/tutorials/2018-WP-Summit

Martina Summer-Kutmon
email: martina.kutmon@maastrichtuniversity.nl

WP Summit 2018 – Workshop 1   51