

# Exploratory pathway analysis

## Citation for published version (APA):

Kelder, T. A. J. (2011). Exploratory pathway analysis. Maastricht: Maastricht University.

## Document status and date:

Published: 01/01/2011

## Document Version:

Publisher's PDF, also known as Version of record

## Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

## General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

## Take down policy

If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.

# **Exploratory Pathway Analysis**



# Exploratory Pathway Analysis

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,  
op gezag van de Rector Magnificus, Prof. mr. G.P.M.F. Mols,  
volgens het besluit van het College van Decanen,  
in het openbaar te verdedigen  
op vrijdag 8 juli 2011 om 14.00 uur

door

Thomas Adriaan Johan Kelder

**Promotor:**

Prof. dr. F.J. van Schooten

**Copromotor:**

Dr. ir. C.T.A. Evelo

**Beoordelingscommissie:**

Prof. dr. P. Lambin (voorzitter)

Prof. dr. E.A.L. Biessen

Prof. dr. D. Cavalieri, University of Florence, Italy

Dr. ir. D.G.J. Jennen

Dr. ir. P.D. Moerland, AMC Amsterdam

The study presented in this thesis was performed within NUTRIM School for Nutrition, Toxicology and Metabolism which participates in the Graduate School VLAG (Food Technology, Agrobiotechnology, Nutrition and Health Sciences), accredited by the Royal Netherlands Academy of Arts and Sciences.

This work was supported by the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre.

Cover design and photography by Thomas Kelder. Lego minifigures kindly provided by Jasper and Thom Smits.

© Thomas Kelder, 2011

This work is licensed under a Creative Commons Attribution 3.0 Unported License.

# Contents

<b>CHAPTER 1</b> .....	7
General Introduction	
<b>CHAPTER 2</b> .....	13
Finding The Right Questions: Exploratory Pathway Analysis To Enhance Biological Discovery In Large Datasets	
<b>CHAPTER 3</b> .....	23
WikiPathways: Pathway Editing For The People	
<b>CHAPTER 4</b> .....	31
Mining Biological Pathways Using WikiPathways Web Services	
<b>CHAPTER 5</b> .....	41
Pathway Interactions In Insulin Resistant Mouse Liver	
<b>CHAPTER 6</b> .....	67
General Discussion	
<b>Abbreviations</b> .....	79
<b>Samenvatting</b> .....	81
<b>Acknowledgements</b> .....	85
<b>Curriculum Vitae</b> .....	87
<b>Publications</b> .....	89



# CHAPTER 1

## General Introduction



An important goal in biology is to understand life at the molecular level, to understand how cells and organisms are built from complex mechanisms involving DNA, proteins and other molecules, and how these mechanisms dysfunction to cause disease. The complexity of these mechanisms is enormous. For example, the human genome contains roughly three billion base pairs [1] and encodes for more than twenty-nine thousand proteins [2]. Extra dimensions to these numbers are added by the presence of different protein states and abundance which vary over time and cellular localization, and in turn depend on interactions with DNA, other proteins, molecules and environmental factors. Decades of research in this area have led to a vast base of knowledge about these complex systems. However, this knowledge probably merely scratches the surface of the underlying complexity and we are still unable to completely understand complex, multifactorial diseases, let alone to predict the effects of a novel drug to cure these diseases.

Biological pathways are often used to abstract and modularize our knowledge about molecular biology. A pathway organizes our knowledge with respect to a functional mechanism and describes events at the molecular level that compose this mechanism. For example, the steps by which the molecular machinery in a cell operates to replicate DNA [3], control the cell division [3], or degrade glucose to produce energy [4] may each be represented as a pathway. A pathway typically describes a set of molecular entities, their relations and interactions, and how they vary in spatial dimensions within a given context. For example, a basic version of the intrinsic apoptosis pathway would describe the following events [3]:

- Cytochrome-c is released from the mitochondria and binds to Apaf-1
- Apaf-1 molecules bound to Cytochrome-c aggregate and bind to Procaspase-9
- Procaspase-9 is cleaved into activated Caspase-9
- Caspase-9 in turn cleaves other proteins from the Caspase family, initiating a cascade
- These Caspases in turn cleave several proteins to set in motion the dismantling of the cell

Grouping these relations and events in a pathway summarizes our knowledge about genes, proteins and other molecules at a functional level. Although the boundaries of a pathway are strictly taken arbitrary, a set of canonical pathways has evolved over the years, comprising many generic mechanisms in the cell [5-7].

Pathways are a powerful visual aid when represented as a diagram [8,9]. For example, the apoptosis example from the previous paragraph can be depicted as the diagram shown in Figure 1A, which gives an overview of the order of events, participants and their relations in a single image that is relatively fast and easy to comprehend. In this case, the depicted mechanism is relatively small and simplified, but depending on the goal of the diagram, a pathway diagram can be made more detailed. For example, Gerhard Michal created an extensive metabolic pathway diagram often referred to as the Boehringer-Mannheim biochemical pathways chart (Figure 1B), covering over 700 enzymes and their reactants [10]. Although harder to interpret, such a complex diagram is still a powerful complement to textual representations of the knowledge it represents. Indeed, Michal's metabolic map is printed as wall chart and hangs above many a researcher's desk as reference. The diagram makes it possible to track different paths from one metabolite to another, which would otherwise require looking up bits of information dispersed over textbooks or journal publications. Along the same line, the diagram makes it easier to lookup the involvement of a single enzyme in conversion of different

metabolites. In addition, when looking at the diagram from a distance, general structures become visible (such as the cyclic form of the TCA cycle) and bottlenecks or focal entities are easier to spot. Therefore, pathway diagrams facilitate comprehending the overall structure of a biological mechanism and predicting possible implications of perturbations (e.g. when one of the proteins is mutated in a disease or targeted by a drug), thereby providing a stepping stone for further research based on existing knowledge.

In the last decades, the amount of data that researchers in biology can measure experimentally has greatly increased thanks to the development of high-throughput experimental techniques. These so-called ‘omics’ experiments measure tens of thousands of entities in parallel, allowing to study for example gene expression [11], protein or metabolite abundance [12] in many different biological samples. In addition, functional data can be measured directly at a large scale, such as protein interactions [13] and protein-DNA binding [14]. The large

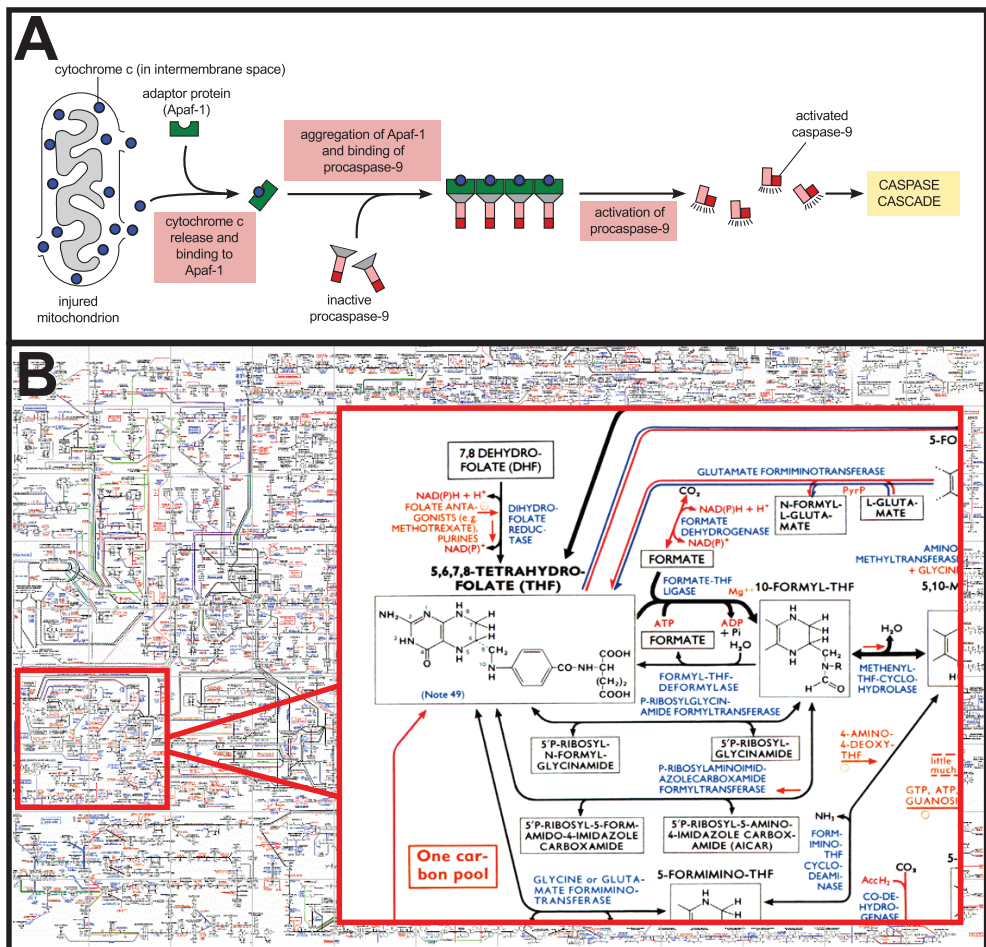


Figure 1: Two examples of a pathway diagram, differing in detail and visual style. A: The intrinsic apoptosis pathway, adapted from [3]. B: The metabolic pathway map created by Gerhard Michal.

coverage and relatively high noise level in the data gathered by these techniques make them especially useful for exploratory analyses, to identify relevant biological mechanisms and generate hypotheses. These hypotheses can then be targeted with smaller scale, but more sensitive experiments. This approach improves our ability to research multifactorial diseases or phenotypes, where the molecular basis does not limit to a single gene, process or tissue. The sheer size of these datasets makes it practically impossible to manually comprehend them and understand the underlying biology, which calls for bioinformatics approaches to assist the human mind.

A critical step in the analysis of a dataset is to integrate it with existing knowledge. Even when the dimensionality of the data has been reduced with data-driven methods such as clustering or pattern recognition, the size and complexity of our knowledge make it hard to manually fit these results into what we already know. For example, MEDLINE, the database for scientific literature in the life sciences field contains over 20 million articles as of 2010 and shows a continuously increasing growth [15]. Typically, all this knowledge is not readily available in a researcher's mind, which often results in focusing on small groups of proteins or a single mechanism that is suspected to play a role, or that is within the field of expertise of the researcher. Unfortunately, that approach does not exploit the key feature of high-throughput techniques (their ability to measure a large diversity of entities) and stands in the way of a more holistic view of molecular biology. Therefore bioinformatics methods may be necessary in this step of the analysis as well, for example to integrate, summarize and present the data in context of biological knowledge. Biological pathways could provide a powerful medium for utilizing existing knowledge in bioinformatics approaches for exploratory data analysis. They are already an abstraction of our knowledge that is more amenable to computing than purely textual information. Furthermore, pathway diagrams provide a way to combine biological knowledge with data visualization to aid human interpretation of the results.

This thesis focuses on applying the concept of biological pathways to bioinformatics approaches in exploratory data analysis. The main goal of this thesis is to improve data analysis by utilizing existing knowledge in the form of biological pathways. Chapter 2 summarizes work that has already been done in this area and discusses several guidelines and prerequisites for utilizing pathways in exploratory data analysis.

One requirement for applying pathways in bioinformatics research is the availability of computable versions of biological pathways that can be kept up-to-date with the latest knowledge. Archiving and curating available literature as pathways is an immense task, requiring expertise from researchers across the world and is therefore especially challenging [16]. Chapter 3 describes a novel platform for collecting, curating and sharing knowledge as biological pathways in a digital form, which applies a community-based approach [17] where the research community is directly engaged in improving pathway information.

A key aspect of exploratory data analysis is to look at the data from different points of view, which requires different bioinformatics software, resources and workflows. To optimally integrate pathways in these tools, the pathway information needs to be accessible programmatically. Furthermore, pathway information evolves as our knowledge grows, requiring these tools to use up-to-date information. Chapter 4 describes a web interface that improves programmatic access to pathways and integration in existing and future bioinformatics tools and analysis workflows.

Multifactorial diseases and phenotypes typically do not result from a single mechanism, but result from interplay between different pathways. Therefore, to improve understanding of a dataset based on a complex disease, we need to look beyond pathway boundaries at interactions between pathways. Chapter 5 applies an exploratory data analysis approach to find interactions between pathways in the context of obesity and insulin resistance. This analysis demonstrates that combining different bioinformatics tools and integrating pathways with additional information, such as protein interactions, provides new perspectives on an experimental dataset and improves our ability to generate new hypotheses.

## References

1. Finishing the euchromatic sequence of the human genome. (2004) *Nature* 431: 931-45.
2. Müller A, MacCallum RM, Sternberg MJE (2002) Structural characterization of the human proteome. *Genome Res* 12: 1625-41.
3. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, et al. (2002) *Molecular Biology of the Cell*, Fourth Edition. Garland Science. 1616 p.
4. Garrett RH, Grisham CM (2008) *Biochemistry*. Brooks Cole. 1184 p.
5. Michal G (1998) *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Wiley-Spektrum. 277 p.
6. Hancock J (2010) *Cell Signalling*. Oxford University Press. 352 p.
7. McMurry J, Begley T (2005) *The Organic Chemistry of Biological Pathways*. Roberts & Company Publishers. 490 p.
8. Michal G (1998) On representation of metabolic pathways. *Biosystems* 47: 1-7.
9. Saraiya P, North C, Duca K (2005) Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Inf Vis* 4: 191-205.
10. Michal G (1993) *Biochemical Pathways* (wall chart). Boehringer Mannheim GmbH.
11. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, et al. (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol* 16: 301-6.
12. Patterson SD, Aebersold RH (2003) Proteomics: the first decade and beyond. *Nat Genet* 33 Suppl: 311-23.
13. Bonetta L (2010) Protein-protein interactions: Interactome under construction. *Nature* 468: 851-854.
14. Gilchrist DA, Fargo DC, Adelman K (2009) Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods* 48: 398-408.
15. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 39: D38-51.
16. Adriaens ME, Jaillard M, Waagmeester A, Coort SLM, Pico AR, et al. (2008) The public road to high-quality curated biological pathways. *Drug Discov Today* 13: 856-62.
17. Doerr A (2008) We the curators. *Nat Methods* 5: 754-755.



## CHAPTER 2

# **Finding The Right Questions: Exploratory Pathway Analysis To Enhance Biological Discovery In Large Datasets**

Thomas Kelder<sup>1</sup>, Bruce R. Conklin<sup>2</sup>, Chris T. Evelo<sup>1</sup>, Alexander R. Pico<sup>2</sup>

1. Department of Bioinformatics – BiGCaT, Maastricht University, The Netherlands
2. Gladstone Institute of Cardiovascular Disease, San Francisco, USA

Published in the August 2010 Issue of PLoS Biology.

At a time when biological data are increasingly digital and thus amenable to computationally driven statistical analysis, it is easy to lose sight of the important role of data exploration. Succinctly defined over 30 years ago by John Tukey [1,2], exploratory data analysis is an approach to data analysis that focuses on finding the right question, rather than the right answer. In contrast to confirmatory analysis, which involves testing preconceived hypotheses, exploratory data analysis involves a broad investigation, a key component of which may be visual display. Though his arguments predate personal computing and thus focus on graph paper and ink, the point still stands: good data visualization leads to simpler (better) descriptions and underlying fundamental concepts. Today, there is tremendous potential for computational biologists, bioinformaticians, and related software developers to shape and direct scientific discovery by designing data visualization tools that facilitate exploratory analysis and fuel the cycle of ideas and experiments that gets refined into well-formed hypotheses, robust analyses, and confident results.

## Pathways for Exploratory Data Analysis

A rich source of visual material relevant to the study of biology is pathway diagrams. Pathways map our understanding about connections and processes underlying biological function. They are powerful models for exploring, interpreting, and analyzing biological datasets and provide a medium to apply Tukey's exploratory data analysis principles to the present-day study of biology (Figure 1). Pathways organize and visualize data and provide a model that both computers and humans can work with, since they are abstract enough to allow for semi-automatic integration and querying in a biological context, and biologists are by and large familiar with pathway diagrams. Ongoing efforts to capture biological knowledge in pathway databases [3] and data exchange formats [4] demonstrate growing interest in applying pathway visualization and analysis to biology research.

Currently, several bioinformatics tools provide pathway visualization to support the exploration of datasets [5,6]. DeRisi et al. projected the changes in mRNA expression on the carbon and energy metabolism pathway to create a visual representation of the properties of metabolic reprogramming during the diauxic shift of yeast [7]. Bensellam et al. applied similar visualization techniques to connect beta cell physiology to specific metabolic and signaling pathways in rat islet cells [8]. A pathway also incorporates a collection or set of biological entities (e.g., genes, proteins, metabolites) that function in the biological process described by the pathway. This information can be used to reduce the dimensionality of



**Figure 1: Pathways for exploratory data analysis.** Biological pathways are powerful visualization tools for data exploration, focused on finding the right question.



large datasets. Identifying pathways that are overrepresented with entities showing interesting behavior gives an overview of global patterns among different biological processes. Many tools and techniques implement this principle [6,9], and it has become an integral part of gene expression data analysis [10]. Recent innovations utilize connectivity and weighting in the calculation of pathway impact [11]. These techniques produce a list of putatively affected pathways that serves as a basis for researchers to develop testable hypotheses of mechanism or direct further exploration. Importantly, when pathway representations are employed in exploratory data analysis, the goal is not a statistical solution, but rather an investigation of the scope of the data and relevant patterns. Pathways serve as the medium for communication, in which the biological story is extracted from the data, prior knowledge is integrated and understanding is constructed [12].

## Challenge

An important goal of “-omics” experiments is to generate directed hypotheses based on relatively noisy but large-scale datasets, which can then be tested in targeted experiments. In this respect, exploratory and confirmatory approaches are complementary, where applying exploratory techniques is a logical first step in the analysis [2]. The relationship is actually more iterative than sequential, where a certain level of statistical analysis or reduction might be required before applying an exploratory technique. But in the overall trajectory from exploratory to confirmatory, exploration is most important in forming a conclusive statistical approach. In the field of pathway analysis, there is active research in developing new techniques and tools from the confirmatory paradigm, using pathways to improve statistical power on specific hypotheses [9,11,13–16]. The value of these techniques for exploratory analysis, however, is limited in the absence of a comprehensive framework for exploration and visualization. The challenge we face now is to fill this gap and to develop flexible tools and pathway content based on the exploratory data analysis paradigm.

Looking at hallmarks of exploratory data analysis may suggest ways that pathways can be more effectively used in data exploration. We will discuss three properties that typify both the exploratory technique and analyst: flexibility, interactivity, and effectiveness. By relating properties of exploratory data analysis to the current state of pathway analysis techniques, we hope to guide researchers in how to best utilize pathway information in exploratory data analysis and help focus future tool development towards better exploratory pathway analysis techniques.

## Flexibility

Exploratory analysis is not a linear start to end process with fixed analysis steps but requires flexibility from both researchers and tools. The decision on what will be the next step in an exploratory analysis is guided by the data and observations rather than by a predefined plan, as is the choice for the technique that is most suitable for highlighting the features under investigation. In exploratory data analysis, we look at the data from many different points of view, few of which actually lead to new or relevant observations. But knowing that a certain description of the data does not lead to a new or relevant observation is itself a step forward in the analysis. The following analogy from Tukey illustrates this:

“As detective stories remind us, many of the circumstances surrounding a crime are



accidental or misleading. Equally, many indications to be discerned in bodies of data are accidental or misleading. To accept all appearances as conclusive would be destructively foolish, either in crime detection or in data analysis. To fail to collect all appearances because some—or even most—are only accidents would, however, be gross misfeasance...” [1].

Thus, open-mindedness is important when using pathways for exploratory data analysis and provides software developers with both a challenge and an opportunity. It is hard to create versatile software that does not restrict researchers to a single workflow. A more generic, flexible framework to support various pathway analysis procedures would be very powerful and would provide a basis for developing new and better pathway analysis techniques. Therefore, instead of aiming for a single, isolated software package, developers should implement flexible solutions that can be integrated in a larger toolbox for pathway analysis, in which each tool provides a different perspective on the dataset. In turn, rather than depending on a single program or algorithm to produce a publishable statistic, biologists should seek tools that help comprehend the data, view it from different angles, and thereby lead to greater understanding of what’s going on.

Consider canonical pathways. These pathways summarize complex biological processes in a comprehensible way, however, these summaries may omit important details by grouping entities, leaving out alternative routes, and imposing artificial boundaries. By limiting analysis to canonical pathways, a researcher is less flexible, fixated on well-described knowledge, and blind to less certain, but potentially more interesting clues. Reality is much more complex than what is depicted in the typical canonical pathway, as has been demonstrated by available protein–protein interaction networks [17] and curated interaction databases, such as Reactome [18]. However, visualizing every possible interaction or entity that might contribute to a process can lead to large incomprehensible “hairball” networks that do not facilitate exploratory analysis. How can we optimally use both types of information in an exploratory analysis?

One option might be to consider canonical pathways as a starting point in the analysis, based on solid foundations from which we might explore less known but potentially interesting areas. For example, a pathway could be dynamically extended with interactions from other pathways, protein–protein interactions, or relations from literature, based on a set of entities that show interesting behavior in the dataset under investigation. In that way, the researcher can explore instances or interactions that might not be integral to the canonical pathway, but might still be relevant to the observations in the pathway. This process could become data-driven, by highlighting and filtering information that is potentially interesting based on the experimental data and context, instead of showing all available information. An analysis environment that exploits both canonical pathways and detailed interaction networks would encourage researchers to take a flexible, exploratory attitude and facilitate construction of an understandable biological story from complex data.

For developers, realizing that exploratory pathway analysis tools might be used not only in isolation but also with other software and different types of data in a flexible analysis setup might guide software design and implementation. For example, providing an application programming interface (API) in addition to the user interface greatly enhances the flexibility to adapt a tool for customized analyses or to reuse components. Reusability of software

components that perform common tasks and define general data models leads to more unity among pathway analysis tools. For example, a data format will be more easily adopted by other developers when an API is available to read, modify and write it. In addition, providing an API opens up the possibility for scripting to automate tasks and combine functionalities of different tools. This introduces a nearly unlimited flexibility and allows a developer to focus on the main functions of a tool and keep the user interface simple and focused, while keeping the option open for advanced users to automate and combine standard features of different tools to perform a novel type of analysis.

## Interactivity

An exploratory analysis is not an automatic process, but relies on decisions by the researcher. Where calculation or visualization tasks may fall to the computer, the researcher controls interpretation and decisions on what data should be viewed, from which angle and in which context. Graphical representations of data are important. As Tukey notes, a good visualization “forces us to notice what we never expected to see,” and “The graph paper (or visualization software) is there, not as a technique, but rather as recognition that the picture-examining eye is the best finder we have of the wholly unanticipated” [2]. Interactive graphics allow the researcher to take control of how the data are visualized and stimulates the researcher to change the visualization perspective based on previous observations.

Pathway analysis techniques that allow the researcher to explore data interactively (rather than delivering a static view) will facilitate exploration and increase the chance of finding interesting observations or patterns. There are several opportunities to improve interactivity of pathway visualizations and highlight features relevant to the question being asked while, just as importantly, filtering out irrelevant features.

Geographical maps illustrate the advantages of interactivity provided by effective visualization software. Paper maps divide the world into multiple views of fixed scope and scale. You can look at a map of the complete world with limited detail or a city map without context. But paper maps are cumbersome and lack critical interactivity (folding a map doesn't count). Digital maps, on the other hand, have several advantages, such as the ability to switch scale through interactive zooming, so you can scroll the viewport to trace a possible route or track your real-time location with GPS information. The integration of information, in general, is yet another advantage, as you can add and remove layers of information on the same map. Such integrated information can be interactively queried to find a particular intersection, a high concentration of public parks, or the best route through traffic. The parallels to biological pathways are obvious and should be exploited at every opportunity in the design of pathway analysis tools. The example of traffic overlays even hints at the dynamics of biological processes, e.g., the flow of biochemistry through metabolic pathways.

Developers of exploratory pathway analysis tools could borrow concepts from the analogy with geographical maps. For example, enrichment analysis techniques group genes, proteins, and metabolites at the level of pathways ranked by activity. This provides a global “world map” view, showing which pathways may be affected while discarding information about the inner workings of these pathways. This scale may hold information on how each pathway acts as a unit in a specific context and how these units relate to each other. Such relationships could include child–parent relations (glucose metabolism and fatty acid metabolism are both

metabolic pathways), the flow of substances (the output of glycolysis is an input for the TCA cycle) or causal relations (the P53 pathway regulates apoptosis). In contrast to the global scale, techniques based on the constituents of pathways provide a more mechanistic “city map” view by relating data to localized interactions and reactions. Continuing to zoom to the molecular level reveals protein domains, the exon structure of splice variants, and polymorphisms. Interactivity may be improved by allowing seamless transitions between these scales by utilizing semantic zooming [19], where the displayed features and level of detail change automatically along with the zoom level and context. Given that most analysis tools focus on pathway information at a single scale, switching between these scales within an exploratory analysis is far from trivial.

## Effectiveness

The interactive, user-directed character of exploratory data analysis imposes stricter criteria on the effectiveness of exploratory techniques. The techniques described in Tukey’s textbook on exploratory data analysis are surprisingly simple and easy to apply merely with paper and pencil. This allows the researcher to take a quick look at typical questions—“could it be that...?” or “what if it is the case that...?”—without investing days of work on that single question. Effective techniques that are relatively easy to apply and work in a transparent way encourage the researcher to take a true exploratory attitude instead of following well-trod paths while ignoring side roads that may reveal unexpected but interesting aspects of the data.

Of course, if the chance of finding an interesting observation in the data does not outweigh the efforts to perform an analysis technique, researchers may decide not to use the technique. This problem may be less relevant in confirmatory approaches, where investing a large effort in a single technique is often justified because the effort versus results can be weighed during planning. However, in exploratory analysis, a single technique is only a small part of the whole analysis (many clues need to be considered, with different techniques), and the yield is often unpredictable (many clues lead to dead ends). Therefore, the acceptable maximum effort is very low, and to make pathway analysis techniques suitable for true exploratory analysis, this should be taken into account.

Unfortunately, many obstacles and annoyances exist when applying current pathway analysis techniques. While modern computers allow fast data processing and visualization, there remain numerous hurdles beyond the need to install and train on multiple software packages and the need to format and reformat datasets into specific input formats. Reordering data columns might not be a major hurdle—spreadsheet software that performs this task is widely available. But mapping data to different identifier systems or applying calculations on the data is less trivial and more prone to error, often requiring specific bioinformatics skills. Pathway analysis tools should aim to remove the responsibility of data reformatting from the researcher by making tools more flexible to different types of input data or to adhere to widely adopted standards. Generic libraries and services that might assist the developer in this task are already available, such as BridgeDb [20] for identifier mapping (to support multiple identifier systems), Web services to access the latest pathway information [21–24], or paxtools [25] for reading pathways in the BioPAX standard.

The pathways themselves require library-like organization and curation. A handful of projects have undertaken the task of capturing and curating this knowledge as semantic content

that is amenable to computation [18,21,26–28]. Unlike systems biology networks, pathways cannot be directly inferred from high-throughput data, but rather require the synthesis of multiple discoveries, insights, and diverse data types spanning years, or even decades, of work by multiple groups, offering an opportunity for tool developers to facilitate the entry, curation, and distribution of pathway content in effective formats [4,28,29]. BioPAX and SBGN are particular examples of community-driven formats for pathway semantics and graphical notation, respectively. Pathways should be understandable by researchers who may not be fully familiar with the biological process that is described, enabling researchers to look at data in context of knowledge outside the scope of their specialty [5]. The most effective pathways are self-explanatory, contain detailed information about biological context, and reference relevant primary data sources and literature.

Another opportunity to make exploratory pathway analysis techniques more effective is to work on better integration with public data resources. Biologists create a wealth of data, which is often available in a public repository, such as ArrayExpress or GEO for transcriptomics datasets [30,31]. During an exploratory analysis, it can be valuable to extend beyond the researcher's own data to consider relevant orthogonal or correlated datasets. However, this is an inefficient process. The researcher must manually find the right datasets, download the data files from the repository, reformat the data, and import it in the pathway analysis tool. An increasing number of public repositories support Web service queries, assisting developers in building tools that perform these tasks programmatically [32]. Repositories and tools that expose data and methods through Web services can readily be integrated into effective, reusable workflows in pathway analysis tools, leading to high-order standards in data analysis.

Effective data integration is a significant hurdle in working with different datasets and pathways in exploratory analysis. Determining what to integrate and how to present it to the user depends on the context and the question being asked. However, this context is typically defined at the semantic level and, thus, is hard for computers to work with. For example, a computer can easily handle the command “hide everything above a certain  $p$ -value threshold,” but has trouble with “show me all data related to cancer.” In an ideal situation, the data are annotated with this information, but the computer still needs to deal with synonyms or subtypes of the word “cancer.” It becomes even more complex when integrating data at the pathway level, where the researcher could ask something like “show me all studies in which MYC is activated by MAPK.” Such questions require correctly annotated pathway information and must deal with information at the semantic level (which interactions “activate”) and synonym or identifier mapping problems (which entities map to “MAPK”).

Recent developments begin to address these issues. Ontologies help in dealing with information at the semantic level. For example, a disease ontology could tell the computer that melanoma is a subtype of cancer, and an event ontology could tell the computer that activation could include phosphorylation, translation or receptor binding interactions. Standards for ontologies, such as the OBO format, and resources that provide access to different ontologies through unified Web services [33] provide the necessary interfaces for tool developers to improve integration of different types of data in pathway analysis tools. In addition, data repositories are actively working on annotating raw datasets to provide better context [34,35], ready to be queried by pathway analysis tools through Web interfaces. Sometimes referred to as integromics, or multi-omics, the integration of annotations and data is critical to extracting

the full potential from large and high-throughput datasets [9,36,37]. Effective construction, analysis and visualization of multi-omic datasets depend on innovative software. These tools must understand what is going in (i.e., with the help of ontologies and data exchange standards), know how to merge and normalize across orthogonal data types, and be adept at displaying multi-dimensional information in meaningful and intuitive contexts. This is a particularly ripe area for exploratory tool developers.

## Conclusion

Biological pathways are a powerful medium in the exploratory analysis of biological datasets, providing a conceptual framework that is familiar to biologists, visually oriented and increasingly available in digital formats that allow interactive display and analysis. By discussing properties of exploratory data analysis in the light of pathways, we highlighted several opportunities for researchers and developers to use pathway analysis in an exploratory setting. Rather than trying to provide a complete overview of pathway analysis approaches, we discussed several ideas and recent developments that lay out a path towards a powerful set of pathway analysis tools developed from an exploratory analysis paradigm. A critical recurring issue is that current pathway analysis tools are rather isolated and hard to combine within an analysis. This may discourage researchers to follow clues that require the use of a different tool to view the data from another perspective, thereby standing in the way of a true exploratory attitude. The field of exploratory pathway analysis is still in its beginning, but with focused and coordinated development, it may eventually play an important role in providing the right questions for confirmatory approaches.

## Funding

This work was supported by the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre and by National Institutes of Health (GM080223).

## References

1. Tukey JW (1977) Exploratory data analysis. Reading, MA: Addison-Wesley. 688 p.
2. Tukey JW (1980) We need both exploratory and confirmatory. *Am Stat* 34: 4.
3. Pathguide. Available: <http://www.pathguide.org>. Accessed 29 July 2010.
4. BioPAX. Available: <http://www.biopax.org>. Accessed 29 July 2010.
5. Saraiya P, North, C, Duca, K. (2005) Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Inf Vis* 4: 15.
6. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, et al. (2010) Visualization of omics data for systems biology. *Nat Methods* 7: S56-68.
7. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680-686.
8. Bensellam M, Van Lommel L, Overbergh L, Schuit FC, Jonas JC (2009) Cluster analysis of rat pancreatic islet gene mRNA levels after culture in low-, intermediate- and high-glucose concentrations. *Diabetologia* 52: 463-476.
9. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1-13.

10. Slonim DK, Yanai I (2009) Getting started in gene expression microarray analysis. *PLoS Comput Biol* 5: e1000543. doi:10.1371/journal.pcbi.1000543
11. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, et al. (2009) A novel signaling pathway impact analysis. *Bioinformatics* 25: 75-82.
12. Penders B, Horstman K, Vos R. (2008) Walking the Line between Lab and Computation: The “Moist” Zone. *BioScience* 58.
13. Sartor MA, Leikauf GD, Medvedovic M (2009) LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics* 25: 211-217.
14. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ (2009) GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 10: 161.
15. Bussemaker HJ, Ward LD, Boorsma A (2007) Dissecting complex transcriptional responses using pathway-level scores based on prior information. *BMC Bioinformatics* 8 Suppl 6: S6.
16. Gold DL, Miecznikowski JC, Liu S (2009) Error control variability in pathway-based microarray analysis. *Bioinformatics* 25: 2216-2221.
17. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 37: D412-416.
18. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, et al. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 8: R39.
19. Hu Z, Mellor J, Wu J, Kanehisa M, Stuart JM, et al. (2007) Towards zoomable multidimensional maps of the cell. *Nat Biotechnol* 25: 547-554.
20. van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, et al. (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11: 5.
21. Pathway Commons. Available: <http://www.pathwaycommons.org>. Accessed 29 July 2010.
22. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37: D619-622.
23. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, et al. (2009) Mining biological pathways using WikiPathways web services. *PLoS One* 4: e6447. doi:10.1371/journal.pone.0006447
24. Kawashima S, Katayama T, Sato Y, Kanehisa M (2003) KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System. *Genome Informatics* 14: 673-674.
25. Paxtools. Available: <http://www.biopax.org/paxtools>. Accessed 29 July 2010.
26. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27-30.
27. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 32: D438-442.
28. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, et al. (2008) WikiPathways: pathway editing for the people. *PLoS Biol* 6: e184. doi:10.1371/journal.pbio.0060184
29. Le Novere N, Hucka M, Mi H, Moodie S, Schreiber F, et al. (2009) The Systems

- Biology Graphical Notation. *Nat Biotechnol* 27: 735-741.
30. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30: 207-210.
  31. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 31: 68-71.
  32. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, et al. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 34: W729-732.
  33. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, et al. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 37: W170-173.
  34. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, et al. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* 38: D690-698.
  35. Sage Commons. Available: <http://www.sagebase.org/COMMONS/Mission.html>. Accessed 29 July 2010
  36. Werner T (2008) Bioinformatics applications for pathway analysis of microarray data. *Curr Opin Biotechnol* 19: 50-54.
  37. Wheelock CE, Wheelock AM, Kawashima S, Diez D, Kanehisa M, et al. (2009) Systems biology approaches and pathway tools for investigating cardiovascular disease. *Mol Biosyst* 5: 588-602.

# CHAPTER 3

## WikiPathways: Pathway Editing For The People

Alexander R. Pico<sup>1,2</sup>, Thomas Kelder<sup>3</sup>, Martijn P. van Iersel<sup>3</sup>, Kristina Hanspers<sup>1,2</sup>, Bruce R. Conklin<sup>1,2</sup>, Chris Evelo<sup>3</sup>

1. Gladstone Institute of Cardiovascular Disease, San Francisco, USA
2. Departments of Medicine, and Molecular and Cellular Pharmacology, University of California San Francisco, USA
3. Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands

 These authors contributed equally to this work.

Published in the July 2008 Issue of PLoS Biology.



The exponential growth of diverse types of biological data presents the research community with an unprecedented challenge and opportunity. The challenge is to stay afloat in the flood of biological data, keeping it as accessible, up-to-date, and integrated as possible. The opportunity is to cultivate new models of data curation and exchange that take advantage of direct participation by a greater portion of the community.

This combination of challenge and opportunity is especially relevant to the task of collecting biological pathway information. Pathways are critical to understanding the functions of individual genes and proteins in terms of systems and processes that contribute to normal physiology and to disease. Each biological pathway must be hewn from a mass of biological information distributed across multiple publications and databases.

The particular challenge of pathway curation is amplified, because pathways are often presented as static images that are not amenable to computation, integration, or data exchange. Furthermore, pathway experts are distributed throughout the world, and most have limited time to learn about complex databases that need their expertise. This challenge can be met by taking the opportunity to develop a new community-based model for pathway curation.

One way to engage the community is with a wiki model, as exemplified by Wikipedia [1]. We see the potential for a wiki-based pathway curation resource, coupled with an embedded graphical pathway editing tool, to meet the growing challenge presented by the influx of biological data and to provide an innovative example of content curation by the biology community (Figure 1).

## Facing the Challenge

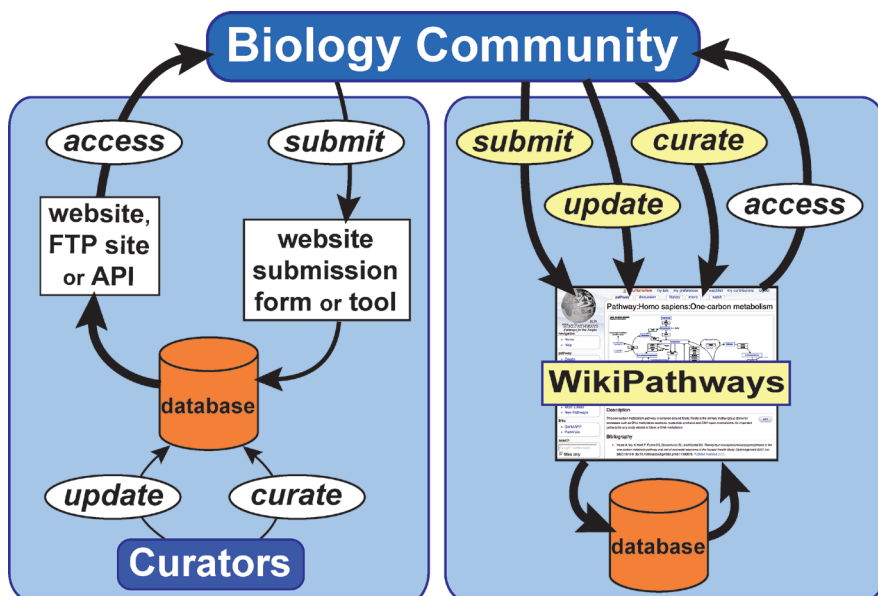
The research community is experiencing massive growth in biological data, from genome and metagenome sequencing to high-throughput assays and microarray studies. This growth has created a need for models of data storage and distribution that support a continuous stream of end-user submissions, frequent updates, integrated search across databases, and access to data formats (preferably community standards) that are amenable to computational analyses. By and large, the need is being met for certain types of biological data: sequences go to GenBank and European Molecular Biology Laboratory (EMBL)-Bank, protein structures go to the Protein Data Bank (PDB), and microarray results go to Gene Expression Omnibus (GEO) and ArrayExpress. But as the influx and complexity of biological data continues to grow, so will the challenge of organizing and maintaining these databases.

Fortunately, the biology community can provide an answer that will scale with the challenge: community curation. There is a growing tendency toward information exchange that supports open access, higher-order organization, community-defined data formats, and collaborative online environments. This trend is most apparent with the growing number of open-access journals (see the Directory of Open Access Journals at <http://www.doaj.org/>), public databases [2], and data exchange formats [3] and ontologies [4]. To promote community curation, database maintainers must be willing to relinquish some control. After removing logistical as well as technical barriers with creative support tools, the data producer will also be the data organizer. Despite initial misgivings, such projects do succeed with the right balance of infrastructure, participation, and administrative principles, as demonstrated by Wikipedia [1], numerous open-source software projects (e.g., Linux, Apache, MySQL, Firefox), and countless

scientific collaborations, including the Internet, itself. The idea of using wiki technology for biological information has been proposed in other areas, for example, genome annotation [5,6]. The EcoliWiki provides a working example of community curation focused on *Escherichia coli* (EcoliWiki, <http://ecoliwiki.net>). And WikiProteins aims to combine automated text mining with community curation to annotate biomedical concepts, including protein functions, interactions, and disease relationships [7].

## Representing Biological Pathways

Biological pathways present a special case in which the information is not directly coupled to data collection. One does not sequence or measure a pathway. Pathways comprise a myriad of interactions, reactions, and regulations, which are often identified piecemeal over extended periods and by a variety of researchers. As a result, pathway information is particularly challenging to compile and curate. Furthermore, biological pathways are often captured only as static images for publications or presentations. Consider the pathway illustrations common in textbooks and review articles that document any given field of biology. A typical signaling pathway, for example, represents receptor-binding events, protein complexes, phosphorylation reactions, translocations, and transcriptional regulation, with only a minimal set of symbols, lines, and arrows. While these simple images are powerful visual and conceptual representations, they cannot be connected to relevant biological annotations or analyzed with respect to experimental data.



**Figure 1: Two Models for Managing Biological Data.** Current biological databases provide the community with data submission forms or tools and access to the compiled data via websites, FTP sites, and, sometimes, programmatic interfaces (API). Internal curation teams organize and update the data. A wiki model for biological databases, such as WikiPathways, provides a single, intuitive interface for submitting, updating, organizing, and accessing data, allowing the community to participate in the curation process and keep up with the influx of new data. The widths of the arrows represent the relative capacity for data management in the two models.

A number of groups have taken on the challenge of curating and archiving biological pathways [8]. Those efforts mainly rely on internally supported teams of biologists or contracts with volunteer experts in particular fields of biology. Their curation tools typically require download, installation, and specialized training and are not designed for broad or collaborative use. Often, the barrier is simply too high for the average biologist to consider contributing their own pathway knowledge. Even when it is contributed, pathway information can remain untouched for years in the current databases, quickly becoming outdated and out of sync with the continuing stream of published discoveries. Some of us (BRC, KH, ARP) have first-hand experience with maintaining the GenMAPP [9] pathway archive in this fashion over the past 8 years. The task of submitting and updating content inevitably falls on a handful of specialists who have invested significant time installing and learning how to use the curation tools. This approach is not sustainable in the face of the growing influx of biological data. Clearly, curating all of biology is a Herculean task for any single group.

## Pathway Editing for the People

To facilitate the contribution and maintenance of pathway information by the biology community, we established WikiPathways (<http://www.wikipathways.org>). WikiPathways is an open, collaborative platform dedicated to the curation of biological pathways. WikiPathways thus presents a new model for pathway databases that enhances and complements ongoing efforts (see Kyoto Encyclopedia of Genes and Genomes (KEGG) at <http://www.genome.jp/kegg>, Pathway Commons at <http://www.pathwaycommons.org/pc/>, and [10]) . Building on the same MediaWiki open-source software that powers Wikipedia, we added a custom graphical pathway editing tool and integrated databases covering major gene, protein, and small-molecule systems. The familiar web-based format of WikiPathways greatly reduces the barrier that prevents participation in pathway curation. More importantly, the open, public approach of WikiPathways allows for broader participation by the entire community, ranging from students to senior experts in each field. This approach also shifts the bulk of peer review, editorial curation, and maintenance to the community.

Each pathway at WikiPathways has a dedicated wiki page, displaying the current diagram, description, references, version history, and component gene, protein, and metabolite lists (Figure 2). Any pathway can be edited from within its wiki page by activating an embedded pathway editor. WikiPathways uses an applet version of PathVisio—a pathway-drawing tool we developed for pathway curation (see PathVisio at <http://www.pathvisio.org> and unpublished data). PathVisio provides a basic palette of objects and annotations needed to represent biological processes. Gene, protein, and metabolite objects directly map to biological annotations from multiple public databases through an extensible identifier synonym database maintained at WikiPathways. The editing tool facilitates annotation with keyword search and auto-completion. Relationships between entities can easily be drawn using “smart connectors” that snap into place. Lines can even connect to other lines to intuitively represent catalysis or other mediated processes. Entities can be grouped to represent complexes and related collections of genes. The editing tool also makes it easy to annotate these entities and relationships with peer-reviewed literature references. The “Help” section of WikiPathways provides guidelines and tutorials for how to use the editor and how to best represent pathway information, as well as how to download and use the pathways in GenMAPP analyses.

After editing, an updated pathway image is displayed on the wiki page along with the version history and list of components. Users can easily monitor and undo changes, compare differences, and search for overlapping pathways. Any registered user can add a pathway to their “watch list” so that they receive email when the pathway is changed. All changes can be reversed, restoring the pathway to an earlier version. Different versions of pathways can be compared side-by-side using an integrated difference-viewing tool, customized for graphical pathway information. Using the search feature, one can locate particular pathways by name, by the genes and proteins they contain, or by the text displayed in their descriptions and comments. One can also browse the collection of pathways with combinations of species names and ontology-based categories. Currently, WikiPathways contains 537 species-specific pathways for human, mouse, rat, zebrafish, fruit fly, worm, and yeast. The mouse pathways, for example, contain 3,741 unique genes (~13% of the mouse genome). The pathway collection was nucleated with GenMAPP pathways, which were collected over the past decade from GenMAPP users. Now at WikiPathways, the collection is growing and improving through new contributions and curation, at an unprecedented rate, which we expect to dramatically increase as community participation grows.

The screenshot shows the WikiPathways interface for the pathway "Homo sapiens:One-carbon metabolism". The page layout includes a navigation sidebar on the left with sections for "navigation", "pathway", "overview", and "community". The main content area features a detailed metabolic pathway diagram with various enzymes and metabolites. A "Pathway Editor" inset is overlaid on the right side of the diagram, showing a zoomed-in view of the pathway with a red box highlighting the editor interface. Below the diagram, there is a "Description" section and a "Bibliography" section with a single reference. The page also includes a search bar and a "Google Custom Search" link.

**Figure 2. Sample WikiPathways Page.** Each pathway has a dedicated page for viewing and editing content. The pathway diagram is edited with an embedded applet version of PathVisio (inset). The Description and Bibliography sections can be edited in-page as well through applets that facilitate entry. Additional information about the pathway components and version history continue on the page (not shown).

The pathway content at WikiPathways is freely available for download in a variety of data and image formats, including GenMAPP Pathway Markup Language (GPML), which is a custom XML format that is compatible with pathway visualization and analysis tools such as Cytoscape [11], GenMAPP [9], and PathVisio (<http://www.pathvisio.org>). GPML allows researchers to draw and identify the molecular participants in a pathway, as well as the relationships among the participants. GPML is a work in progress, and though it does not yet have the full expressiveness of BioPAX (see the BioPAX Wiki at <http://biopaxwiki.org/cgi-bin/moin.cgi>) or Systems Biology Markup Language (SBML) [12], it provides the basic functionality for researchers to create appealing pathway diagrams and to perform basic statistical tests on pathways such as overrepresentation analysis. The goal of GPML is to bridge the simple elegance of a pathway drawn on a napkin by a biologist (including its rich, human interpretability) and the growing databases of gene and protein annotations, interactions, and experimental data. We prioritized the development of GPML based on what is already available and what is most useful to the average biologist: connecting intuitive, human-readable graphics to standardized identifiers from popular databases. This allows users to accurately label entities on pathways and computationally map them to experimental data using pathway analysis software. GPML also supports the representation of relationships between entities to allow network-based visualization and analysis. In a recent “community curation event” at WikiPathways, we formalized network relationships in the human pathway archive. We plan to include a number of BioPAX elements into GPML to support data exchange, but the overriding goal for GPML is to lower the barrier for contributors of pathway information by keeping it simple. This approach resonates with the large portion of the biology community interested in basic statistical pathway analyses and figures for publications and presentations.

To assist pathway authors and curators, we are developing “bots” to survey the content and identify potential inconsistencies, redundancies, and incomplete data. The first of these bots identifies all the genes, proteins, and metabolites in any pathway that are not connected to a synonym database identifier. These reports along with additional curator tools will help contributors to submit high-quality content and make corrections where needed. We also plan to use standard biomedical ontologies to structure the content of WikiPathways and to provide organization that can scale with rapidly growing and interrelated information.

Researchers interested in particular interactions or pathways can use WikiPathways as a resource for up-to-date pathway information and as a repository for their own findings that, in turn, are immediately available in multiple data formats for analysis as well as image formats for publication. WikiPathways can be used collaboratively to create, edit, and share pathway information with any colleague who has access to a Web browser. For sensitive content that is proprietary or must first be published as an original finding, pathways can be saved locally in the GPML format, ready to be uploaded and made public at a later time. Expert curators can use WikiPathways to monitor and update pathway information associated with their fields of interest. WikiPathways is also useful to students and professors of biology, providing pathways as educational materials and the editing history of a given pathway as an example of how scientific knowledge iteratively progresses.

To encourage participation by the community we have built templates for “User pages” and “Portals.” User pages help users identify themselves and their work, whereas Portals help entire communities of users to identify themselves collectively and focus on particular pathway

domains, such as diabetes-related pathways or plant pathways. By using the template, users can build a site within WikiPathways dedicated to their lab, organization, or area of interest within minutes. We are also organizing community curation events as a way to introduce new users to the curation tools and, at the same time, improve the quality of the pathway content. Future community curation events will focus on adding annotation, group representations, and literature references.

Even prior to this publication introducing WikiPathways, we have seen strong signs of community participation. Outside of the immediate group of developers, WikiPathways has already attracted ten new mouse pathways, nine new human pathways, six new zebrafish pathways, three new rat pathways, and one Portal for the micro-nutrients community. There are dozens of *E. coli* and plant pathways currently being converted, and 3 new Portals under construction. The site has over 220 registered users (10% contributing users) and has attracted developers through the Google Summer of Code program.

We envision WikiPathways being part of a broader effort to extend curation capacity to larger groups and communities. This effort does not replace current approaches involving centralized teams of curators, but rather it complements and extends them. Eventually, we would like to see wiki solutions like WikiPathways used by current databases and curation sources. Our future directions include supporting “reference” pathways contributed by other pathway databases, and private workspaces for groups to collaboratively work on pathways before making them public. One could also imagine organizations installing local instances of WikiPathways for internal projects at research institutes or biotechnology companies. A journal, for example, could host a version of WikiPathways that only contributing authors can edit. Where the same wiki technology is used, there are opportunities for seamless integration and controlled sharing of content when it is ready to be published or released to the public site. We will continue to work toward supporting broad implementations of WikiPathways to promote contributions from established and diverse sources.

WikiPathways is an experiment. We have considerable work ahead of us in developing the GPML data model, implementing critical features and, most importantly, building a community of users and contributors. The success of WikiPathways will depend on the overall quality of its content, which will be a function of the infrastructure and administrative principles we use in addition to community participation. Features such as database connectivity, automatic consistency checks, curation tools, reversible edits, the visual difference viewer, and support by literature references will assist in tracking and reverting errant contributions, stimulating curation by the community. We anticipate that lowering the entry barrier for participation will allow for a greater capacity of curation, broader consensus on content, and ultimately, higher quality control. We are confident that WikiPathways will be a powerful resource for the research community and a vital forum for pathway curation. And we are hopeful that it will serve as an example for how the continuing flood of biological data can be managed and utilized by the community to irrigate future hypotheses and discoveries.

## Source Code

We are committed to open access and open source. All content is available under a Creative Commons License (<http://creativecommons.org/licenses/by-nc-sa/3.0/>). All source code for WikiPathways and the PathVisio applet is available under the Apache License, Version 2.0



(<http://www.apache.org/licenses/>). You can download the code from:

<http://svn.bigcat.unimaas.nl/wikipathways>

<http://svn.bigcat.unimaas.nl/pathvisio>

## Acknowledgments and funding



We thank Susan Coort, Nathan Salomonis, Andra Waagmeester, Grant Yang, Sam Rudy, Gary Howard, Allan Kuchinsky (Agilent Technologies), and members of the Conklin and Evelo labs for valuable advice and support. This work was supported by grants from the National Institutes of Health (GM080223, HG003053), the Agilent Technologies Foundation, the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre, and the Google Summer of Code program.

## References

1. Giles J (2005) Internet encyclopaedias go head to head. *Nature* 438: 900-901.
2. Galperin MY (2007) The molecular biology database collection: 2008 update. *Nucleic Acids Res* 36: D2-D4.
3. Strömbäck L, Hall D, Lambrix P (2007) A review of standards for data exchange within systems biology. *Proteomics* 7: 857-867.
4. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech* 25: 1251-1255.
5. Salzberg SL (2007) Genome re-annotation: a wiki solution? *Genome Biol* 8: 102.
6. Hu JC, Aramayo R, Bolser D, Conway T, Elsik CG, et al. (2008) The emerging world of wikis. *Science* 320: 1289b-1290.
7. Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, et al. (2008) Calling on a million minds for community annotation in wikiproteins. *Genome Biology* 9: R89.
8. Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* 34: D504-506.
9. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, et al. (2007). Genmapp 2: New features and resources for pathway analysis. *BMC Bioinformatics* 8: 217.
10. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, et al. (2007) Reactome: a knowledgebase of biological pathways and processes. *Genome Biology* 8: R39.
11. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using cytoscape. *Nat Protoc* 2: 2366-2382.
12. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19:524-531.

# CHAPTER 4

## Mining Biological Pathways Using WikiPathways Web Services

Thomas Kelder<sup>3</sup>, Alexander R. Pico<sup>1,2</sup>, Kristina Hanspers<sup>1,2</sup>, Martijn P. van Iersel<sup>3</sup>, Chris Evelo<sup>3</sup>, Bruce R. Conklin<sup>1,2</sup>

1. Gladstone Institute of Cardiovascular Disease, San Francisco, USA
2. Departments of Medicine, and Molecular and Cellular Pharmacology, University of California San Francisco, USA
3. Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands

 These authors contributed equally to this work.

Published in PLoS ONE, July 30, 2009.



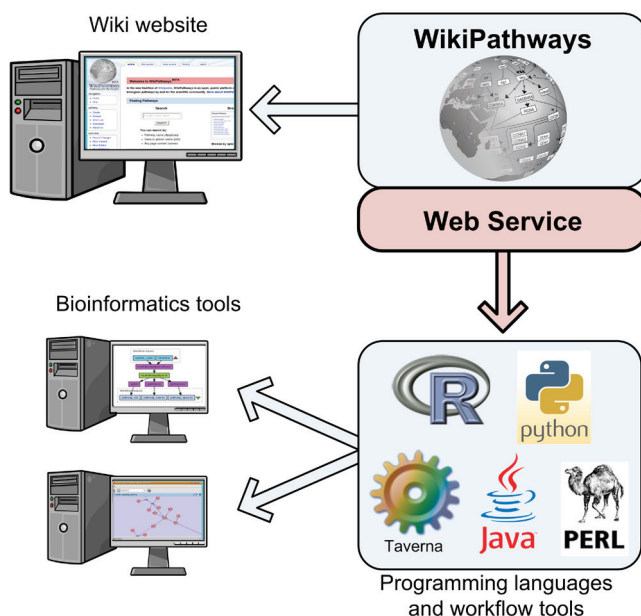
## Abstract

WikiPathways is a platform for creating, updating, and sharing biological pathways [1]. Pathways can be edited and downloaded using the wiki-style website. Here we present a SOAP web service that provides programmatic access to WikiPathways that is complementary to the website. We describe the functionality that this web service offers and discuss several use cases in detail. Exposing WikiPathways through a web service opens up new ways of utilizing pathway information and assisting the community curation process.

## Introduction

WikiPathways is an online resource for biological pathway information and a platform for community-based curation [1]. Currently, WikiPathways contains more than 600 pathways, representing various species from bacteria and fungi to plants and animals. New species are being added on demand, and pathways are being created and edited almost daily. These pathways combine different types of biological knowledge, including protein interactions, metabolic reactions, annotations from gene, protein and metabolite databases and references to scientific literature. Using a Java-based pathway editor, researchers can create, edit and annotate pathways directly on the website [2].

WikiPathways is accessed by the biology community mainly via a wiki-style website. In addition to the website, we recently implemented a web service that provides programmatic access to WikiPathways (figure 1). This makes it easier to integrate biological pathways in existing



**Figure 1:** WikiPathways can be accessed by end-users from the wiki-style website. In addition, the WikiPathways web service provides a programmatic interface that can be used in many programming languages, including R, python, Java and perl and in workflow tools such as Taverna. Using this interface, new pathway analysis tools can be built and existing bioinformatics tools can be extended with pathway-based functionality.

applications and provides a framework for pathway-centered analysis of experimental data. In contrast to other pathway resources with managed curation teams (e.g., KEGG [3] and Reactome [4]) or focused on distributing and querying pathway information (e.g., Pathway Commons [5]), WikiPathways is a primary source for richly annotated pathway content open to community curation. With the addition of web services, users have expanded access and advanced methods that uniquely apply to WikiPathways content. In this article, we discuss the web service functionality in more detail and highlight several use cases.

Supplementary data, including full documentation, example client implementations, and source code, are available at <http://www.wikipathways.org/webservice>. The source code of the web service and all examples are licensed under the Apache 2.0 open-source license.

## Results and Discussion

### Web service functionality

The web service provides an interface to WikiPathways that can be accessed through the Simple Object Access protocol (SOAP), and the data structure and available functions are described in the Web Service Description Language (WSDL). Both SOAP and WSDL are widely supported standards. The web service provides access to all pathway information on WikiPathways in several different forms. Complete pathways can be downloaded in XML format (GPML) or a plain text file listing the biological entities and their identifiers. Image versions of a pathway can be retrieved in several graphics formats, including rasterized formats (e.g., Portable Network Graphics (PNG)) and vector graphics formats (e.g., Scalable Vector Graphics (SVG) and Portable Document Format (PDF)). Additionally, color information can be specified to highlight specific elements of the pathway (e.g., to color protein entities according to their measured expression). Individual interactions between biological entities that are defined in the pathways can be retrieved separately. Furthermore, information related to community-based curation, such as revision history and recently edited pathways, can be queried. A full list of available functions can be found in the supplementary data.

An index of all pathways is maintained using the Apache Lucene library [6]. This feature makes it possible to perform advanced search queries through the web service. All textual information on the pathway is included in the index to allow for keyword searches. In addition, the index uses synonym databases [2] to cross-reference between various biological databases and the entities on the pathways. This allows for queries to find all pathways for any given biological identifier, regardless of the identifier system that is used to annotate the pathway. For example, when the query is an Affymetrix probeset identifier, all pathways containing genes that map to that probeset will be returned.

The web service also allows client software to publish information to WikiPathways. New pathways can be uploaded and pathways can be modified or labeled according to quality standards. This enables scripts to perform quality monitoring and notification to assist the manual community-based curation process. This concept has already been successfully applied to other wiki's, such as Wikipedia [7]. To prevent scripts from systematically overwriting pathways with invalid data, write access to the web service is restricted to a subset of user accounts. Users who require write access for their script can request it from the WikiPathways site administrators.

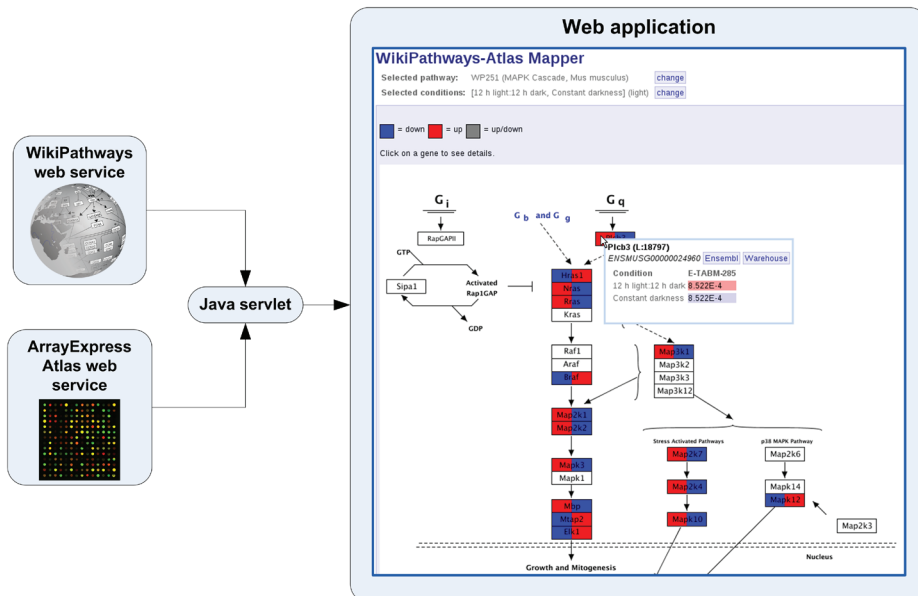
To assist programmers in building applications that use the WikiPathways web service, several toolkits and programming libraries exist. Libraries to handle SOAP requests and responses are available for practically any programming language. Additionally, several bioinformatics tools, such as Taverna and GenePattern, support plugging in SOAP web services by writing only little or no extra code. This makes it easy to integrate WikiPathways in existing pipelines. To facilitate working with the GPML pathway format, we maintain an open-source Java library that provides a high-level API to process GPML. This library contains methods to read and write pathways in several file formats and to modify information in the pathway. Furthermore, it provides an object-oriented interface to the WikiPathways web service, including support for caching downloaded pathway information locally to improve performance. For each described use-case, example code is provided that demonstrates the use of available toolkits and libraries. The supplementary data include a list of useful libraries in several programming languages.

## Web applications

The web service can be used to build web applications that provide end-users access to specific WikiPathways functionality. Research groups can build a website that queries, processes and presents information from WikiPathways in a fully customized way. As an example, we implemented two web applications, each highlighting a unique functionality of the web service. The first application is an improved search application with more advanced functionality, available at <http://search.wikipathways.org>. The default WikiPathways search is based on a Google Custom Search Engine and allows searching through all the text on pathway pages, but doesn't use the biological context that is defined in the GPML format. The improved search application uses the web service functions that query the Lucene index and allows users to perform searches with biological context, such as using biological identifiers and filtering by species. Furthermore, the results are presented as thumbnail images, making it easier to choose the relevant result.

The second example is a web application that demonstrates integration of pathway information with other types of data. This application visualizes gene expression information from ArrayExpress Atlas [8] on a WikiPathways pathway. ArrayExpress Atlas is a curated set of gene expression datasets that are publicly available. In this example, the user can specify a pathway from WikiPathways and a set of experimental conditions defined in ArrayExpress Atlas. First, all gene identifiers on the pathway will be mapped to Ensembl using the synonym database. The resulting Ensembl identifiers are passed to the ArrayExpress Atlas web service, which returns the corresponding experiments and p-values for the differentially expressed genes. Second, the WikiPathways web service will be used to download a colored version of the pathway image that will be displayed to the user (figure 2). This application can be used to get a quick overview of how known pathway interactions relate to observed gene expression in a given experimental condition and is available at <http://atlas.wikipathways.org>. Currently, 716 of the 1000 experiments on ArrayExpress Atlas can be mapped onto at least one of the WikiPathways pathways, representing 9357 unique genes affected in one or more experiments across 7 different species.

Research groups are encouraged to build their own client-side web applications based on the WikiPathways web service and our open source libraries. This could include applications that



**Figure 2:** Web application that integrates pathways with gene expression information. Pathway and gene expression information are retrieved from the WikiPathways and ArrayExpress Atlas web services respectively. A Java servlet integrates this information and publishes it to an interactive web application. In this web application, users can view the information on an interactive pathway diagram.

present WikiPathways content in a customized way or integrate pathways with other data. For example, research groups focused on metabolic pathways could create an application that presents the pathways in combination with detailed enzymatic information, while the genetics community could create a web application that combines polymorphism information with the genes from a pathway.

### Assisting community-based curation

The biology community manually moderates the content on WikiPathways. Being able to quickly respond to mistakes as well as acts of vandalism is an important aspect of community-based curation [9]. Many types of erroneous data can be identified automatically. For example, wrongly annotated gene, protein, or metabolite entities are identified by a simple computer script. The web service enables us to create “bots” that continually watch the pathway content for errors and notify curators when incorrect information is introduced. Currently, various types of bots run on a daily basis and check for incorrect database annotations, missing literature references, or incorrectly defined interactions. Each bot generates an HTML report that provides statistics and an overview of pathways that need to be corrected. By monitoring these reports over time, we can assess the effectiveness of community-based curation.

Community involvement in pathway curation on WikiPathways could be further stimulated by using the web service to build applets that can be included in any webpage or desktop. These applets could display user-specific information, such as recent changes on pathways

that have been edited by the user, recent discussion items, or listing other users that are interested in similar pathways. Users could install this applet on their own webpage or on their desktop, for example using Google Gadget interface [10].

## Querying interactions in Cytoscape

WikiPathways includes interactions between biological entities that can be represented as interaction networks. For example, metabolic reactions or protein activation events that are defined in a pathway can be viewed as binary interactions that form a graph. This enables network-based analysis [11] and integration with various datasets, such as protein-protein interactions [12,13]. Cytoscape [14] is a widely used tool for visualization and analysis of biological networks. The core Cytoscape functionality can be extended by external developers via a plug-in mechanism. We used this mechanism to build a plug-in that allows Cytoscape to work with GPML pathways [2]. We recently extended the plug-in with functionality from the WikiPathways web service. This new functionality makes it possible to search for pathways on WikiPathways directly from within Cytoscape, and these pathways can be loaded as an interaction network and then analyzed and integrated with experimental data [15]. Furthermore, existing networks in Cytoscape can be expanded with interactions defined in WikiPathways. Network analysis of biological pathways can also be a useful tool to augment pathways with knowledge from various resources. For example, cross-talk between biological pathways can be identified by projecting pathways on large protein-protein interaction networks [16]. The Cytoscape plug-in is a typical example of how the web service can be used to enrich external analysis tools with pathway content.

Future versions of WikiPathways will allow the user to define the semantic meaning for each interaction in a pathway. This can be used to improve the web service with functions that query and filter interactions based on this semantic information. For example, queries, such as “show me all proteins that inhibit phosphorylation of protein X” could be performed. Including semantics also opens new possibilities for analysis in Cytoscape, for example the signaling pathway impact factor analysis method [17], which can use the distinction between activation or inhibition interactions.

## Integration in bioinformatics workflows

A primary difficulty in bioinformatics is integrating erratically formatted data from different resources [17,18]. The development of web services for biological databases makes this easier and enables usage of workflow tools, such as Taverna [19]. Taverna provides a framework for creating workflows that integrate and process different types of data. For instance, a workflow could be used to integrate microarray experiment data with quantitative trait loci to filter for relevant genes [20]. Knowledge represented as a pathway is especially suited for integration with other biological data, since it typically covers multiple levels of the biological information hierarchy. Transcriptomics, proteomics, or metabolomics data can be integrated with pathway information to aid in the understanding of the underlying biological mechanisms. We created several Taverna workflows based on the WikiPathways web service, such as a workflow that finds relevant pathways based on a list of over-expressed genes, proteins, or metabolites. These workflows can be downloaded from <http://www.myexperiment.org/packs/30>. Other possible workflows that could be implemented in Taverna include cross-species comparisons

of pathway information, or batch visualization of expression data.

The WikiPathways web service could be used as framework for building data analysis tools that make use of pathway information. Pathways can be used for visualizing experimental data in a biological context [21], finding relevant biological mechanisms and improving statistical power of an experimental analysis [22,23]. The WikiPathways web service provides all of the functionality necessary to perform these methods. Pathways can be downloaded as images where the pathway components can be given user defined colors, allowing for experimental data analysis. Genes, proteins and metabolites are linked to several identifier systems, providing information to perform enrichment analyses.

### **Integration with online databases**

Pathways typically consist of different types of biological entities, such as genes, proteins and metabolites. For each entity type, different biological databases are available, and each presents unique information about the entities in a different way. With the WikiPathways web service, we aim to encourage database developers to integrate pathway information into their online data presentation to provide more biological context. WikiPathways links pathway components to over 50 supported biological databases, including Entrez Gene, Ensembl, Affymetrix, ZFIN, TAIR, and ChEBI. Tools that use information from any of these databases can use the web service to retrieve relevant pathway information per biological entity. For example, the list of pathways that contain a given gene identified by an Ensembl ID could be retrieved and displayed on the Ensembl web page for that gene or any website indexed by Ensembl identifiers. The website could display the pathway name, URL and/or thumbnail image of the pathway. A similar approach could be taken by metabolic databases (e.g., PubChem, ChEBI, or ChemSpider), protein databases (e.g., UniProt), model organism databases (e.g., MGI, ZFIN, WormBase, FlyBase, or TAIR), measurement platforms (e.g., Affymetrix, Illumina, or Agilent), or even literature databases, such as PubMed.

Future versions of WikiPathways will support the export of pathways in BioPAX format. This will make it easier to integrate WikiPathways with other pathway databases and resources, such as PathwayCommons. This functionality will be available in the web service, so that integrated pathway resources can easily keep the pathway information from WikiPathways up-to-date.

### **Conclusions**

The WikiPathways web service provides an interface for programmatic access to community-curated pathway information. It provides a flexible framework for building or extending tools that use pathway information from WikiPathways. The web service can be used by software developers to build or extend tools for analysis and integration of pathways, interaction networks and experimental data. The web services are also useful for assisting and monitoring the community-based curation process. By providing this web service, we hope to help researchers and developers build tools for pathway-based research and data analysis.

## Author contributions

Wrote the paper: TK ARP KH MPvI CE BRC. Designed and implemented the software: TK ARP KH MPvI.

## Funding

This work was supported by the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre and grants from the National Institutes of Health (GM080223, HG003053).

## References

1. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, et al. (2008) WikiPathways: pathway editing for the people. *PLoS Biol* 6: e184.
2. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, et al. (2008) Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* 9: 399.
3. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27-30.
4. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37: D619-622.
5. <http://www.pathwaycommons.org/>.
6. <http://lucene.apache.org/>.
7. Huss JW, 3rd, Orozco C, Goodale J, Wu C, Batalov S, et al. (2008) A gene wiki for community annotation of gene function. *PLoS Biol* 6: e175.
8. <http://www.ebi.ac.uk/microarray-as/atlas/>.
9. Doerr A We the curators. *Nature Methods* 5: 754-755.
10. <http://code.google.com/apis/gadgets/>.
11. Ideker T, Sharan R (2008) Protein networks in disease. *Genome Res* 18: 644-652.
12. Keskin O, Tsai CJ, Wolfson H, Nussinov R (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci* 13: 1043-1055.
13. Tarassov K, Messier V, Landry C, Radinovic S, Molina M, et al. (2008) An in Vivo Map of the Yeast Protein Interactome. *Science* 320: 1465-1470.
14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498-2504.
15. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2: 2366-2382.
16. Li Y, Agarwal P, Rajagopalan D (2008) A Global Pathway Crosstalk Network. *Bioinformatics*: btn200.
17. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, et al. (2008) A Novel Signaling Pathway Impact Analysis (SPIA). *Bioinformatics*.
18. Stein LD (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet* 9: 678-688.
19. Oinn T, Addis M, Ferris J, Marvin D, Senger M, et al. (2004) Taverna: a tool for the

- composition and enactment of bioinformatics workflows. *Bioinformatics* 20: 3045-3054.
20. Fisher P, Hedeler C, Wolstencroft K, Hulme H, Noyes H, et al. (2007) A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Res* 35: 5625-5633.
  21. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, et al. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* 8: 217.
  22. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, et al. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* 4: R7.
  23. Nam D, Kim S-Y (2008) Gene-set approach for expression pattern analysis. *Brief Bioinform*: bbn001.





# CHAPTER 5

## Pathway Interactions In Insulin Resistant Mouse Liver

Thomas Kelder<sup>1</sup>, Lars Eijssen<sup>1</sup>, Robert Kleemann<sup>2</sup>, Marjan van Erk<sup>3</sup>, Teake Kooistra<sup>2</sup>, Chris Evelo<sup>1</sup>

1. Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands
2. TNO, Metabolic Health Research, Leiden, The Netherlands
3. TNO, Research Group Microbiology & Systems Biology, Zeist, The Netherlands

Supplementary figures, tables and data are available online at:  
<http://code.google.com/p/tkelder/wiki/PathwayInteractions>

Under review.

## Abstract

**Background:** Complex phenotypes such as insulin resistance involve different biological pathways that may interact and influence each other. Interpretation of related experimental data would be facilitated by identifying relevant pathway interactions in the context of the dataset. **Results:** We developed an analysis approach to study interactions between pathways by integrating gene and protein interaction networks, biological pathway information and high-throughput data. This approach was applied to a transcriptomics dataset to investigate pathway interactions in insulin resistant mouse liver in response to a glucose challenge. We identified regulated pathway interactions at different time points following the glucose challenge and also studied the underlying protein interactions to find possible mechanisms and key proteins involved in pathway cross-talk. **Conclusions:** Studying pathway interactions provides a new perspective on the data that complements established pathway analysis methods such as enrichment analysis. This study provided new insights in how interactions between pathways may be affected by insulin resistance. In addition, the analysis approach described here can be generally applied to different types of high-throughput data and will therefore be useful for analysis of other complex datasets as well.

## Introduction

Biological pathways provide a powerful medium to explore and reduce the complexity of large datasets. Pathways organize genes, proteins, metabolites and their interactions into functional groups, often visualized as diagrams or networks. A commonly employed analysis technique using pathways is enrichment analysis, where pathways are represented as gene sets and where the aim is to find those sets that are enriched with entities of interest, such as differentially expressed genes [1]. This allows a researcher to get an overview of biological processes that are likely to play a role in the studied phenomenon. The result of enrichment analysis is a sorted list of pathways, which is easier to interpret than a list of thousands of individual significantly expressed genes. However, each pathway in this list is presented as an isolated entity, while in reality these pathways can interact, for example through interacting or shared proteins and metabolites. To aid further exploration and interpretation of gene set enrichment results, it would be useful to get insight in possible relations or interactions between pathways and how these are affected in the context of the studied phenotype.

One way to get insight in possible relationships between pathways is to look at their overlap in gene, protein or metabolite content. Pathways with a high overlap might be related by shared paths. Tools such as ClueGO [2] and EnrichmentMap [3] allow the user to convert the list of enriched pathways into a network by calculating overlap between the sets. We used another approach with bi-partite graphs to create a network based on overlap in significantly regulated genes [4].

Another more functionally based approach is to find possible pathway cross-talk by looking at protein interactions between pathways. Cross-talk allows multiple pathways to exchange signals and influence each other. For example, the P53 pathway can control the Cell Cycle pathway by regulating the expression of p21 and can itself be activated by several pathways, for example the MAPK pathway. Metabolic pathways may share enzymatic reactions and may influence each other by influencing the availability of a substrate. These forms of pathway cross-talk are highly context dependent, for example, interactions between the P53 pathway

and Cell Cycle depend on several external stress factors such as DNA damage or oxidative stress. Previous studies have already used protein interaction networks to study pathway interactions in a given context. Kelley and Ideker [5] integrated protein interactions with data from genetic interaction analyses in yeast and found that the largest part of genetic interactions could be explained by interactions between pathways. In two other studies [6,7] a pathway cross-talk network was built based on direct interactions between the proteins in human pathways. Both studies were based on the assumption that a pair of pathways is likely to interact when a higher number of protein-protein interactions are found between them than would be expected by chance. The work of Li et al. resulted in a scale-free pathway cross-talk network in which pathways in the same broad functional category indeed cluster together. Transcriptomics data was integrated into this network by finding cliques (a subset of pathways in which every two pathways are connected) which contained highly enriched pathways. Huang et al. performed a similar study, but also considered overlapping proteins between the pathways and integrated transcriptomics data at the protein level by counting only interactions between proteins encoded by differentially expressed genes. This resulted in context specific networks, reflecting interactions between pathways within the context of the dataset.

In this study, we investigate interactions between pathways by finding regulated paths between pathways for a given transcriptomics dataset. While the methods of Li et al. and Huang et al. only consider direct protein interactions between pathways, considering paths spanning multiple interactions may detect indirect interactions as well. Indirect interactions consist of paths including one or more proteins that are not annotated to a pathway, but do have known interaction or binding partners. In case the majority of the genes encoding these proteins and their interacting partners in two different pathways would be differentially expressed in a given condition, this indicates a potentially relevant path through which these pathways interact. An algorithm to detect pairs of pathways that contain such a path would allow researchers to directly explore and visualize possible paths along which pathways might interact in a given context. Furthermore, the inclusion of proteins that have not yet been annotated to a pathway in the analysis makes it possible to look beyond well studied processes, increasing the chance of generating novel hypotheses.

We designed a method to detect both direct and indirect interactions between pathways and visualize the resulting paths and applied this method to a transcriptomics dataset from the European Nutrigenomics Organization (NuGO) PPS2 study [8] to investigate the response to a glucose challenge in liver samples from obese and insulin resistant, as well as normal mice. Insulin resistance is a complex disease, not limiting to metabolism, but also associated to for example inflammation [9]. Interactions between different pathways might be especially relevant in the context of such complex phenotypes. Using this dataset, we hope to gain insight in the regulated biological processes during the response to the metabolic stressor glucose and the influence of a pretreatment with a high-fat diet on this response.

## Results

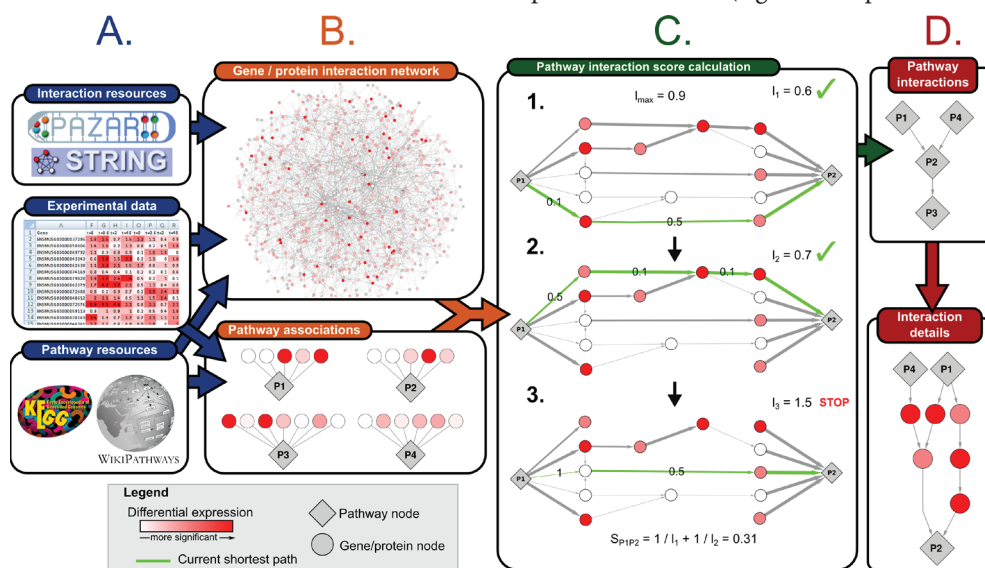
To find potentially regulated paths between different pathways, we designed a generic method based on non-redundant shortest paths in a weighted graph. The required input data are:

- A set of pathways, each pathway consisting of a collection of gene, protein and/or metabolite entities.

- An interaction network, providing interactions or functional associations between genes, proteins and/or metabolites, which can be compiled from different resources.
- A weight value for each edge in the interaction network, indicating how much an edge will contribute to the length of a path. This can be based on experimental data, for example the expression of the genes that the edge connects.

For each pathway pair, a score is calculated by finding the set of non-redundant weighted shortest paths that have a length which is smaller than a given threshold and do not cross any other pathway. The length of a path is defined by the sum of the edge weights in the path. The score for a pathway pair depends on the number of paths found and their length (Figure 1C). The more and shorter the paths found between two pathways, the higher the score. To indicate the significance of an interaction between two pathways, a p-value is calculated by comparing the score to an empirical null distribution based on re-sampling of the pathways (see methods). A pathway network is generated by creating an edge between a pathway pair if their p-value is smaller than a given significance threshold.

This method was applied to pathways from the WikiPathways [10] and KEGG databases [11], which after merging strongly overlapping pathways resulted in a set of 236 pathways covering 6953 genes. A directed interaction network was generated based on reactions in these pathways and extended with reactions, functional associations between proteins, protein-protein interactions and transcription factor targets from several public databases (see methods). The resulting interaction network consisted of 6893 proteins and 138,105 interactions. Because a subset of interactions in this network have a specified direction (e.g. transcription factor



**Figure 1:** Overview of the analysis approach to investigate interactions between pathways. A: Information from different resources and experimental data is integrated into a weighted gene/protein interaction network and a set of pathways and their associated genes and proteins. B: Based on the interaction network, an interaction score and significance is calculated for each pathway pair. C: Example of the process of identifying a set of non-redundant shortest paths for the interaction of pathway P1 to P2. This panel shows step 5-7 of the calculation as described in the Methods section. D: Two representations of the resulting pathway interactions. The top panel shows the pathway interaction network, where each edge represents a significant interaction between two pathways. The bottom panel shows a detailed network showing the identified shortest paths between pathways P1, P4 and P2.

targets), the identified interactions between pathways are directed as well.

Pathway interactions were investigated in the context of a transcriptomics dataset of mouse liver samples. This dataset contains gene expression measurements before ( $t=0$ ) and at 0.6 hour ( $t=0.6$ ), 2 hours ( $t=2$ ) and 48 hours ( $t=48$ ) after a glucose challenge in two groups of mice fed either a low fat (LF) or high fat (HF) diet for 12 weeks before the challenge. After these 12 weeks all mice fed with HF diet had become obese and had developed insulin resistance [12]. To test for differential gene expression between different groups, in total 7 comparisons were made. Firstly, to determine the baseline effect of the different diets on gene expression, the measured genes were tested for differential expression between the HF and LF samples at  $t=0$ . Secondly, to determine the changes in gene expression following the glucose challenge in each individual diet group, the genes were tested for differential expression between  $t=0$  and each other time point ( $t=0.6$ ,  $t=2$  and  $t=48$  hours). Each comparison resulted in a T-statistic for every gene representing the significance and direction of differential expression. The number of differentially expressed genes ( $q < 0.01$ ) is shown in Table 1. For the number of differentially expressed genes in the glucose challenge response, the two diet groups display a different trend. In the LF group, the number of significant genes grows over time and is highest at  $t=48$ , indicating that even 48 hours after the challenge gene regulation has not returned to the initial situation. However, in the HF group, the number of significant genes peaks at  $t=2$  and decreases again at  $t=48$ . For each of the comparisons, pathways were identified that were enriched with differentially expressed genes (Supplementary figure 1 and Table 1). The trend observed with respect to the number of differentially regulated genes in response to the glucose challenge is not apparent in the number of enriched pathways. Both the LF and HF groups show an increasing trend, where the number of enriched pathways at  $t=48$  is higher than preceding time points in both groups.

Edge weights were derived from the T statistics of the comparisons to generate a weighted interaction network for each test. For each edge, the weight value inversely depends on the value of the T statistic of the target gene (see methods for details), receiving a low weight value if the gene is differentially expressed. Thereby, paths that include genes (or the proteins they encode for) that are differentially expressed will have a shorter length. The number of edges in each interaction network that have a weight smaller than the maximum path length and hence may contribute to a path between two pathways is shown in Table 1. These edges cover between 3.5% and almost 19% of the complete interaction network. The algorithm to

Test	Significant genes, $q < 0.01$	Enriched pathways, $p < 0.05$	Edges with weight $\leq l_{max}$
HF vs LF, $t=0$	1971 (17.3%)	54 (22.9%)	24055 (16.8%)
LF $t=0$ vs $t=0.6$	24 (0.2%)	17 (7.2%)	5054 (3.5%)
LF $t=0$ vs $t=2$	573 (5.0%)	43 (18.2%)	13251 (9.2%)
LF $t=0$ vs $t=48$	1607 (14.1%)	57 (24.15%)	20220 (14.1%)
HF $t=0$ vs $t=0.6$	773 (6.7%)	13 (5.5%)	15784 (11.0%)
HF $t=0$ vs $t=2$	2815 (24.7%)	17 (7.2%)	26845 (18.7%)
HF $t=0$ vs $t=48$	736 (6.4%)	40 (17.0%)	16857 (11.7%)

**Table 1:** Gene and pathway statistics for each comparison: Number of significant genes, enriched pathways and edges in the interaction network that have a weight shorter than the maximal path length and hence contribute to the pathway interactions. Percentages are relative to the total number of measured genes, the total number of pathways and the total number of edges in the interaction network respectively.

detect pathway interactions was run for each comparison, resulting in 7 pathway interaction networks. A Cytoscape [13] session file that contains an interactive visualization for each network is available as supplementary data.

The results of the algorithm can be represented as two different types of networks (Figure 1D). First are the pathway interaction networks that provide a global overview of how pathways might relate without showing the underlying protein interactions. Hence, each node in this pathway interaction network represents a pathway and each edge is a significant interaction between them. Second are detailed networks that also show the protein interactions that compose the paths between the pathways. These networks consist of two types of nodes, representing either a pathway or a protein. An edge between a pathway node and protein node represents association of the protein to the pathway and an edge between two protein nodes represents an interaction between these proteins. The next paragraph provides an analysis of the generated pathway interaction networks by identifying pathways with a more central role in the network. This is followed by a global analysis of the protein interactions that form these pathway interactions, to identify proteins that may play an important role in pathway-crosstalk in this dataset. Finally, several potentially interesting pathway interactions will be highlighted and investigated more closely by zooming in to their protein interactions via the detailed network representation.

## Global analysis of the pathway interaction networks

The number of identified pathway interactions ranges from 25 to 207 across the different networks (Table 2). The differential expression between the HF and LF groups at t=0 (before glucose challenge, but after 12 weeks of diet intervention) results in the most interactions, while the early response to the glucose challenge in the LF group results in the fewest interactions. For 66.7% of the pathways enriched with differentially expressed genes between the diet groups at t=0, at least one significant interaction with another pathway could be found. This number is much lower for the early response to the glucose challenge in the LF group (20% and 27.9%) and HF group at t=0.6 (23.1%).

To identify potential focal pathways in the network, we can look at several node centrality measures. Supplementary figure 2 shows the degree (number of interactions with other pathways) of the most connected pathways. Since the pathway interactions are directional, we can make the distinction between in-degree and out-degree. A pathway with a high in-degree

Test	Nodes (pathways >= 1 interaction)	Enriched nodes	Edges ( $p < 0.001$ )
HF vs LF, t=0	123 (52.1%)	36 (66.7%)	207
LF t=0 vs t=0.6	17 (7.2%)	5 (20.0%)	25
LF t=0 vs t=2	57 (24.2%)	12 (27.9%)	52
LF t=0 vs t=48	82 (34.8%)	27 (47.4%)	108
HF t=0 vs t=0.6	43 (18.2%)	3 (23.1%)	54
HF t=0 vs t=2	111 (47.0%)	10 (58.8%)	160
HF t=0 vs t=48	82 (34.8%)	19 (47.5%)	92

**Table 2:** Statistics for each pathway interaction network: Number of pathways and enriched pathways in the network that have at least one significant interaction and the number of significant pathway interactions. Percentages are relative to the total number of pathways and total number of enriched pathways respectively.

is the target of many pathway interactions, which may indicate that the pathway is strongly regulated. A pathway with a high out-degree is the source of many pathway interactions, which may indicate a role in regulation of different targets. The three pathways with the highest in-degree for the network resulting from the comparison between diets at  $t=0$  are three stress response and apoptosis related pathways: Oxidative damage and apoptosis (14 interactions), FAS pathway and stress induction of HSP70 (14 interactions) and Apoptosis and its regulation by HSP70 (10 interactions). The first two pathways share 11 neighbors among their incoming interactions and 7 neighbors are shared among all three pathways. Interestingly, 5 of these 7 shared neighbors are in the top pathways with highest out-degree and 4 of those are also enriched with differentially expressed genes ( $p < 0.05$ ). For most of the pathways with the highest in-degree, the out-degree is zero or very low and the corresponding node is acting like a sink, or end point. The same holds the other way around, where pathways with the highest out-degree tend to act as a source. To find pathways that may have a gatekeeper role in the network, we can look at the betweenness centrality, which measures how often a pathway occurs on shortest paths between the other pathway nodes in the network. The higher the betweenness, the more the pathway can control interactions between other pathways in the network. Supplementary figure 3 shows the betweenness centrality of the pathways with highest betweenness. Notable are the many cytokine related pathways that all have a high betweenness in the network for the response at  $t=2$  in the HF diet group. None of these centrality measures seem to correlate with enrichment of the pathway and a pathway does not have to be enriched with differentially expressed genes to exhibit a central position in the pathway interaction network.

### Protein interactions contributing to regulated paths between pathways

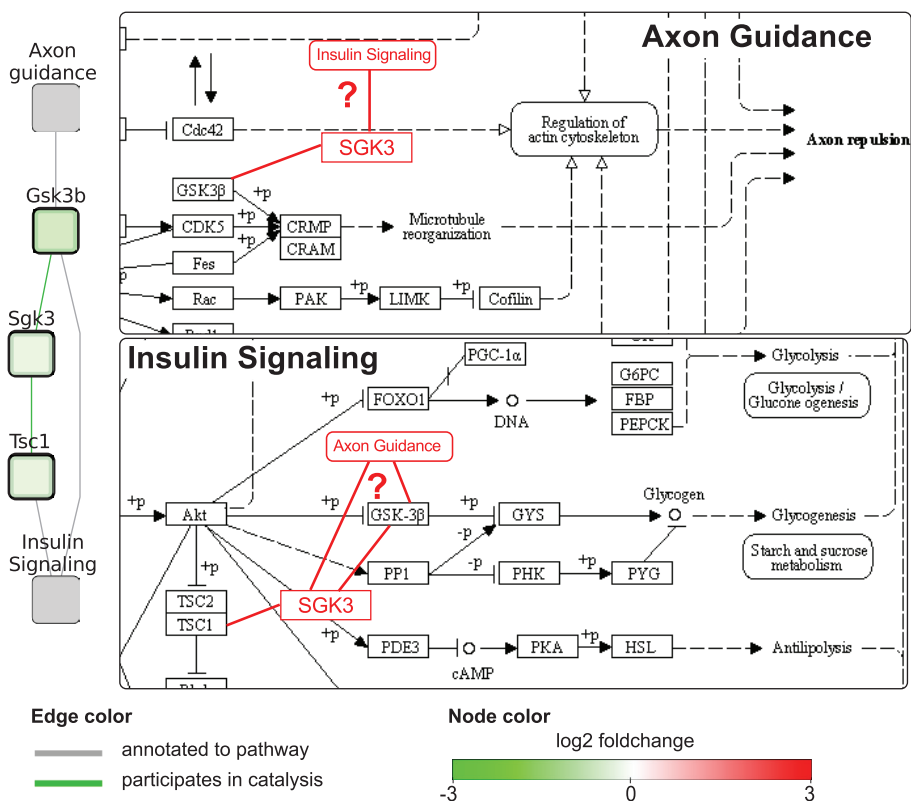
Studying the protein interactions and functional associations that form differentially regulated paths between pathways may provide insights in the mechanisms of interacting pathways and reveal potential key regulators of pathway cross-talk that play a role in the dataset. The number of paths contributing to each pathway interaction network ranges from 52 for the smallest network (LF response at  $t=0.6$ ) to over 1500 for the largest (comparison between diets at  $t=0$ ). Most paths consist of direct interactions between proteins in the interacting pathways (Supplementary figure 4). Only 17% of the paths over all networks are indirect and have one intermediate protein, while only 1 path has two intermediate proteins. The network for the response in the LF group at  $t=48$  consists of almost 50% of indirect paths, a considerably larger part than the other networks.

Test	Unique protein interactions	Number of unique proteins	Unique proteins not in a pathway
HF vs LF, $t=0$	440	198	25
LF $t=0$ vs $t=0.6$	8	6	0
LF $t=0$ vs $t=2$	47	43	1
LF $t=0$ vs $t=48$	275	142	14
HF $t=0$ vs $t=0.6$	53	47	0
HF $t=0$ vs $t=2$	400	200	18
HF $t=0$ vs $t=48$	118	86	2

Table 3: Number of unique proteins and protein interactions that contribute to the paths between pathways for each network.



There is a lot of overlap among the paths between pathways, since the number of unique protein interactions that form these paths is much lower than the number of paths itself (Table 3). For example, for the comparison between the diet groups at  $t=0$ , the 1562 paths between pathways are formed by only 440 unique protein interactions. The high overlap of protein interactions in the paths between pathways indicates the presence of proteins that are involved in multiple paths and might act as key regulators of pathway interactions. Supplementary Table 1 shows the proteins that are involved in most pathway interactions per network. Most of these proteins participate in interactions with highly connected pathways and a large part is involved in well-known signaling processes such as *Kras* and *Mapk1*, transcription factors such as *Jun* and cytokines such as *Il1a* and *Il1b*. There is some overlap of these proteins between the different networks, but there are no proteins that are in the top ten for each time point. One protein that plays a significant role in multiple networks is *Pik3r1*, which is known to play a role in insulin signaling by activation through IRS-1 [14] and participates in most of the paths for the networks for the response in the HF group at  $t = 2$  and  $t = 48$ .



**Figure 2:** An indirect interaction between the Axon Guidance and Insulin Signaling pathways in the network for the comparison between HF and LF diet at  $t=0$ . Left: Network representation of the identified path between the two pathways, consisting of three proteins Gsk3b, Sgk3 and Tsc1. Right: The location of these proteins in the KEGG pathway diagrams. The newly found indirect interactions have been added in red.

Proteins that act as intermediate in indirect interactions between pathways are of special interest. They are not annotated to any pathway yet, while they do seem to play a role in the context of this dataset, so studying their interactions may lead to new findings that would have been missed by looking at the annotations in the pathways alone. Table 3 lists the number of such proteins per network. Interestingly, while the network for LF at t=48 contains relatively many indirect interactions, the number of proteins playing a role in indirect interactions is lower than expected. Upon closer inspection, the high number of indirect paths is mainly due to binding interactions of a single set of proteins (including differentially expressed *Mapre3*, *Cep57*, *Tubg1* and *Nedd1*) that are not annotated to a pathway, but do all bind to a few proteins that play a role in many different pathway interactions, such as *Ywhab*, *Prkaca*, *Tubb5* and *Hsp90ab1*. One of the proteins that is not annotated to any pathway is *Sgk3* which plays a role in several pathway interactions for the response to the glucose challenge in the LF diet group at t=2 and t=48 (Supplementary figure 5) as well as the comparison between the diet groups at t=0 (Figure 2). In the latter, *Sgk3* interacts with *Gsk3b* and *Tsc1* to form one of the paths that form the identified interaction between the Axon guidance and Insulin signaling pathways. The genes encoding for these three proteins are all down-regulated in the HF diet group ( $p < 0.01$ ). Since *Gsk3b* is present in both the Axon guidance and Insulin signaling pathways, in addition to the possible interaction between the two pathways via *Sgk3*, it might also form an alternative route within the Insulin signaling pathway that has been differentially regulated in the HF diet group (Figure 2).

## Visual exploration of the pathway interaction networks

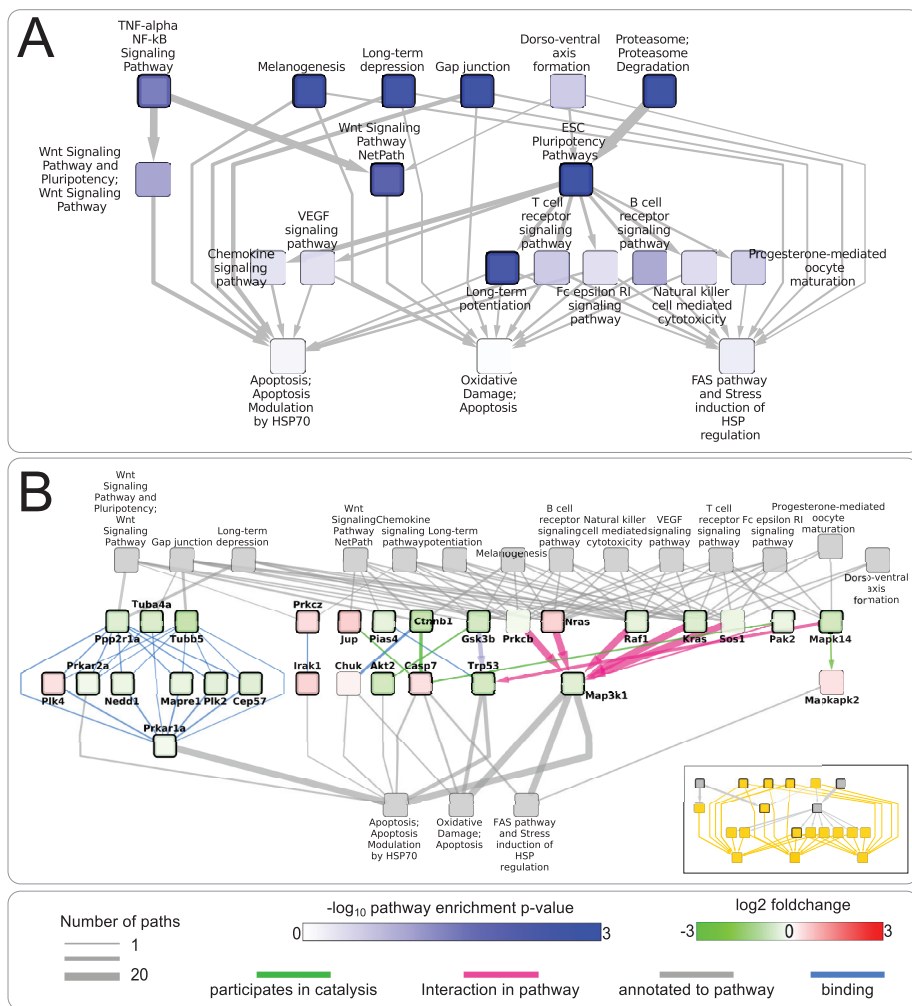
The pathway interaction networks might be powerful tools for interactive exploratory analysis, by browsing and filtering the network visualizations in different ways to find interesting structures. The structural properties discussed above may guide to the most relevant parts of each network, which can then be explored in more detail. In this section we highlight several notable pathway interactions for each network and investigate the protein interactions that form the paths between the pathways.

## Differential expression between the LF and HF group

The network for the comparison between the diet groups at t=0 may provide insight in the changes contributing to and resulting from the development of insulin resistance after feeding of a HF diet. Due to the large size of this network, and to aid visual exploration, we only include pathway interactions for which at least one of the participating pathways is significantly enriched ( $p < 0.05$ ) with differentially expressed genes. Based on this filtered network, we study several potentially interesting interactions and subgraphs. A good starting point for this exploration might be the 3 apoptosis and stress response related pathways that have the highest in-degree in the network. Interestingly, 8 out of 14 pathways interacting with at least one of the three apoptosis related pathways also have an incoming interaction from the ESC pluripotency pathway, which is significantly enriched ( $p < 0.01$ ). This pathway has a strong interaction (consisting of 38 paths) with the Proteasome pathway, which is also significantly enriched ( $p < 0.01$ ). The remaining pathways interacting with the 3 apoptosis related pathways include 4 pathways with a relatively high out-degree of which 3 are significantly enriched ( $p < 0.05$ ), and two versions of the Wnt signaling pathway, which both have a strong connection

(25 and 27 paths) with the significantly enriched TNF-alpha NF-kB Signaling Pathway ( $p < 0.05$ ). The subgraph containing these pathways and their interactions is shown in Figure 3A.

Based on this subgraph, we can “zoom in” to view the individual protein interactions that comprise the paths between the interacting pathways. For example, the paths between the 3 apoptosis related pathways and their neighbors are shown in Figure 3B. Despite the many neighbors of these three pathways, the number of unique proteins in these neighbors contributing to the interaction is relatively low (15 proteins for 14 pathways) and hence many



**Figure 3:** Pathway interactions and detailed network visualization for the interactions with the three apoptosis related pathways for the comparison between HF and LF diet at  $t = 0$ . **A:** Subgraph of the pathway interaction network, based on incoming interactions to three stress response and apoptosis pathways with the highest in-degree. Pathway nodes with a thick border are significantly enriched ( $p < 0.05$ ). **B:** The protein interactions that compose the interactions between the three apoptosis related pathways and their neighbors in the subgraph as shown in box A (see inset, included interactions are colored orange). Protein nodes have a thick border when they are significantly differentially expressed ( $q < 0.05$ ).

of these proteins are annotated to more than one of the source pathways. This indicates the presence of common paths, shared by different pathway interactions. Indeed, there seem to be two main paths shared across different pathway interactions. Firstly, *Raf1*, *Kras*, *Nras* and *Prkcb* are present in more than 10 of the source pathways and all form a path to the three target pathways via *Map3k1*, which participates in 53 paths between the pathways. Secondly, there is a distinct path through *Prkar1a*, which forms an interface to the Gap junction, Long term depression and Wnt signaling pathway via indirect binding interactions with *Ppp2r1a*, *Tuba4a* and *Tubb5*. *Ppp2r1a* may also play a role in the interaction with the TNF-alpha NF-kB Signaling pathway and two versions of the Wnt pathway via indirect binding interactions with *Ywhab* (Supplementary figure 6C). Since this is an indirect interaction, involving proteins that have not been annotated to one of the pathways, this is a target for further study to elucidate the potential role of these proteins in interaction between these two pathways. The link between the Proteasome and ESC pluripotency pathway can be largely explained by interactions between differentially expressed subunits of the proteasome and *Gsk3b*, *Apc* and *Cttnb1* (beta-catenin). The pathway diagram of the ESC pluripotency pathway reveals that these interactions are part of Wnt signaling and can influence the concentration of beta-catenin by targeting it for proteolysis after phosphorylation by *Gsk3b* and *Apc*. Interestingly, *Cttnb1* and *Gsk3b* are also present in other interactions, connecting the TNF-alpha NF-kB signaling pathway to the Wnt signaling pathways, the B-cell and T-cell receptor pathways to the Oxidative damage and apoptosis pathways via *Akt2*, *Casp7* and *Trp53*, and the ESC pluripotency pathway to the B-cell and T-cell signaling and Chemokine signaling pathways via *Crk*, *Nfatc2* and *Chuk*.

In addition to the pathways with a high degree, the network contains a potentially interesting component consisting of 10 metabolic pathways for which interactions with the Primary bile acid biosynthesis and Peroxisome pathways have been found. Out of the 12 pathways in this component, 9 are significantly enriched ( $p < 0.05$ ). When looking at the protein interactions that form the interactions between these pathways, it turns out to be comprised of two binding interactions between *Hadh*, *Hsd17b10* and *Hsd17b4* (Supplementary figure 7). These binding interactions are assigned by the STRING database [15] as subunits of a single enzyme complex, but in reality they are three distinct isozymes of 3-hydroxyacyl-CoA dehydrogenase [16,17]. Therefore, these proteins are probably not binding, but rather catalyzing a similar reaction in each of these pathways. While these proteins are differentially expressed and may play a role in insulin resistance and response to HF diet, they are not likely to contribute to an interaction between these pathways.

### Early response to the glucose challenge

The networks based on differential expression between  $t=0$  and each other time point after the glucose challenge in the LF group provide insight in pathway interactions that play a role during the response to the glucose bolus in mouse liver. The low differential expression at  $t=0.6$  in the LF group (only 24 significant genes,  $q < 0.05$ ) is reflected in the small size of the corresponding network. Studying the protein interactions for that network reveals that all identified pathway interactions include *Jun* and *Fos*, for which the gene expression has both increased more than 5-fold compared to  $t=0$ , and *Il1b* with a 2.5 fold increase (Supplementary Figure 8). Part of the interactions also includes *Il1a*, which is differentially expressed with a more than 2-fold increase. As a dimer, *Jun* and *Fos* form the AP-1 transcription factor,

which responds to several stimuli including cytokines and controls several processes such as apoptosis and proliferation, for all of which related pathway interactions are identified. The network for the HF group at  $t=0.6$  shows a high overlap with the corresponding network for LF, including almost all edges of the LF network. The protein interactions and expression for these overlapping edges are similar to those in LF (Supplementary Figure 8), indicating that these interactions via AP-1 remain unaffected by obesity and insulin resistance. In addition to the largest component of the network for HF at  $t=0.6$ , which is largely spanned by the interactions with *Jun* and *Fos*, there are several small components of isolated interactions between two or three pathways, which are not present in the LF network. For example, one of the components is composed of the Antigen processing and presentation and Spliceosome pathways, both interacting with the Lysosome pathway. These interactions have been identified because two genes encoding for proteins of the HSP70 family, *Hspa1a* and *Hspa1b*, are differentially expressed ( $>4$  fold) and interact with *Ap1s1*, *Ap1b1* and *Cltc*. These proteins are involved in transport of lysosomal enzymes and show moderate, but not significantly differential expression (Supplementary figure 9). Another potentially interesting component that consists of three pathways provides an interaction between the S1P receptor pathway, the Ribosomal proteins pathway and Insulin signaling pathway via *Mapk7*, *Nr4a1* and *Sgk1*, which are all significantly up-regulated compared to  $t=0$  (Supplementary figure 10), and *Rps6kb2*, which is up-regulated albeit not significantly ( $q = 0.066$ ). Based on the pathway diagram, *Mapk7* is a downstream target of the S1P receptor pathway, but because the other proteins in this pathway do not show significantly differential expression, it does not seem likely that the differential expression of *Mapk7* is related to upstream activities in this pathway. However, since *Mapk7* is present in the insulin signaling pathway as well, this might point to an alternative path within the insulin signaling cascade, which may be affected by HF diet in response to the glucose challenge. The networks for  $t=2$  show much more difference between the HF and LF diet. There are only two overlapping pathway interactions and in addition, the network for LF is much less connected. A large part of the connectivity in the HF network situates around the merged version of the Cytokine-cytokine receptor interaction pathway and Cytokines and inflammation pathways (Cytokine pathway), which has the highest betweenness centrality. Several receptors listed in this pathway are differentially expressed, such as *Il1r1* and *Il1r2*, involved in a path to Apoptosis and oxidative stress via *Casp8* and *Casp9*. All other receptors that contribute to the pathway interactions, including up-regulated *Igf1r*, *ErbB3*, *Egfr* and down-regulated *Kdr* and *Fgfr4*, are receptor tyrosine kinases and form a path to several specific interleukin signaling pathways, T-cell and Kit receptor pathways via the *Cbl* protein. *Cbl* is a ubiquitin-protein ligase that can target these receptors for endocytosis [18]. Its expression profile shows a peak at  $t=2$  in the HF group only, where the expression has increased almost 2-fold compared to  $t=0$ . We also found that the Endocytosis pathway, which indeed contains *Cbl* and the receptor tyrosine kinases as well, shares all its interacting pathways with the Cytokine pathway. In addition to *Cbl*, the interactions between the Endocytosis pathway and its neighbors also include the up-regulated *Cltc* and *Cltb*, which encode for the two light chains of clathrin, as well as several other proteins that are involved in clathrin dependent endocytosis (Supplementary figure 11). This endocytosis mechanism is known to be responsible for transporting receptor tyrosine kinases to the endosome after ubiquitination by *Cbl* [18]. Together, these interactions indicate that removal of these receptors from the cell membrane is affected at this point in the glucose challenge in the HF diet group specifically.

## Late response to the glucose challenge

A notable aspect of the late response to glucose challenge is that the LF group contains many more differentially expressed genes than at the early time points, while in the HF group the number peaks at  $t=2$  and is lowest at  $t=48$ . This difference is not reflected in the pathway interaction networks for  $t=48$ , which both contain almost the same number of nodes and edges. However, the overlap between these networks is small; only 4 edges are present in both networks, including the interaction between the TGF-beta receptor signaling Pathway and Proteasome. Interestingly, the proteins in the paths between these two pathways with the TGF-beta and Proteasome pathways behave differently between the HF and LF group (Supplementary figure 12). All genes encoding the participating proteins in the Proteasome pathway are down-regulated in the LF network and up-regulated in the HF network. The proteins *Cdc27* and *Ctnnb1* in the TGF-beta pathway show a similar pattern. The interaction in the LF network contains another protein, *Axin1*, which is significantly up-regulated in the LF group and slightly down-regulated in the HF group is not included in the latter interaction due to its low significance.

The network for the LF group contains one tightly connected cluster, held together by the Antigen processing and presentation and T Cell receptor signaling pathways, including interactions with the enriched Cell cycle and TNF-alpha NF-kB signaling pathways. There are several central proteins in these interactions, including a set of shared binding partners between *Ywhab* in the Cell cycle and TNF-alpha pathways and *Tubb5* in the T Cell Receptor pathway, or *Hsp90ab1* in the Antigen processing pathway (Supplementary figure 13). The *Gab1* protein in the TNF-alpha pathway directly interacts with several proteins in the T Cell pathway, including the down-regulated *Fyn* and *Sos2*, and up-regulated *Pik3r2* and *Pxn*. In addition, *Prkaca* is present in most neighbors of the Antigen processing and presentation and T Cell receptor signaling pathways, making it the protein involved in most paths for this network. Another interesting structure in the network is the second order neighborhood of the Insulin signaling pathway, containing the Wnt and IL-6 signaling pathways and the Starch and sucrose metabolism pathway, which are all significantly enriched (Supplementary figure 14). A large part of the involved proteins are present in multiple of these pathways. For example, half of the proteins of the IL-6 pathway are also present in the Insulin signaling pathway. One of the proteins present in both the Wnt and IL-6 pathways, *Gsk3b*, forms an indirect path to the Insulin signaling pathway via *Sgk3* which interacts with *Trib3* and *Tsc1*, as well as a direct interaction via *Tsc1*. Two enzymes *Pygl* and *Gys2* participate in the interactions with the Starch and sucrose metabolism pathway, interacting with several proteins in the Insulin, IL-6 and Wnt signaling pathways, which might indicate activated mechanisms that regulate these enzymes. The up-regulated *Pygl* gene encodes for glycogen phosphorylase which catalyzes the rate-limiting step in glycogen degradation. The down-regulated *Gys2* gene encodes for glycogen synthase, involved in conversion from glucose to glycogen. Together, this indicates that at  $t=48$  the liver switches from storage of excess glucose to usage of the stored glycogen, possibly influenced by the pathway interactions identified here. The pathway interaction network for the HF response is typed by several key proteins including *Pik3r1*, *Kras*, and *Ctnnb1*. *Kras* and *Ctnnb1* both have a protein interaction with *Pik3r1* which plays a key role in many of the identified pathway interactions in this network. In addition, the up-regulated protein *Mapk14* also interacts with *Kras* and 8 out of 12 pathways that contain *Mapk14* also contain *Pik3r1*.



## Discussion

The goal of this study is two-fold. Firstly, we aimed to design methods for exploring possible pathway interactions in a specific context, which can be defined by an experimental dataset. Secondly, we explored pathway interactions relevant to glucose response in mouse liver and how these interactions are affected by high fat feeding-induced insulin resistance. We will first highlight several of the identified interactions and follow with a discussion about the described approach in general.

The high number of interactions with three apoptosis related pathways in the network for the comparison between diets at  $t=0$  suggest involvement of apoptosis in insulin resistant liver. Even though these pathways are not enriched with differentially expressed genes, they are the end point of many pathway interactions from upstream signaling pathways, which suggests that apoptosis is differentially regulated in the insulin resistant mice. Indeed, a relation between apoptosis and insulin resistance has been described before [19] and a recent study has shown that apoptosis might be increased in insulin resistant liver in humans [20]. The proteins identified could help to find mechanistic explanation for this observation, but it is hard to identify the most crucial step based on transcriptomics data alone, since many of the signaling events take place via post-transcriptional modifications.

The response to the glucose challenge at  $t=0.6$  in both LF and HF is typed by the up-regulation of *Jun* and *Fos*, together forming the transcription factor AP-1, and *Il1a* and *Il1b*. This might indicate that AP-1 initiates several processes in the early response to glucose, possibly activated by *Il1*, and that this mechanism is not influenced by obesity or insulin resistance. A previous study has demonstrated that *Il1* might indeed activate AP-1 in a hepatocyte cell line [21]. Immediate early genes such as *Jun* and *Fos* are genes that are activated transiently and rapidly in response to a wide variety of cellular stimuli, including glucose [22]. They represent a standing response mechanism that is activated at the transcription level in the first round of response to stimuli, before any new proteins are synthesized. Our interaction network contains 7 target genes of AP-1, of which 4 are present in the dataset, but none are differentially expressed during the early response. A possible cause might be that AP-1 is not activated and therefore its expression has no consequences, however we cannot determine this with the current dataset. Because of the general role of AP-1, there are probably more targets than annotated in our interaction network, so it may be of interest to analyze this effect in more detail using more specialized bioinformatics approaches or identify possible targets in a follow-up experiment. In addition to this shared mechanism, we also identified several interactions unique for the HF group. Firstly, the expression of two proteins of the HSP70 family show a distinct peak of over 4-fold increased expression at  $t=0.6$  in the HF group, but not in the LF group. We identified a possible interaction of these proteins with the Lysosome pathway that may indeed play a role in this pathway by uncoating of clathrin-coated vesicles [23]. However, based on our data it is hard to judge the relevance of this finding, since the three proteins in the Lysosome pathway that may interact with the HSP70 proteins show only moderate, not significantly differential expression at  $t=0.6$ , their expression profiles do not correlate, and only a few additional genes in the Lysosome pathway itself show significantly differential expression at  $t=0.6$ . It may be more likely that another function of these HSP70 proteins, for example their role as chaperones, could be related to their expression profile specific to the HF group. In addition, one of the two HSP70 proteins, *Hspa1a* (also known as

HSP72), was found to play a role in preventing insulin resistance and blocking inflammation in human muscle by preventing c-jun amino terminal kinase (JNK) phosphorylation [24]. Since JNK can activate the c-Jun transcription factor, which was also found to be differentially expressed at t=0.6, there may be a relation with the HSP70 proteins. Unfortunately neither the pathway collection nor the interaction network we used here contains any relevant interactions which are altered at the transcriptional level and could provide more insight in this possible relation. Secondly, we identified an interaction between *Mapk7*, *Rps6kb* and *Sgk1* in the Insulin signaling pathway and the transcription factor *Nr4a1* in the Nuclear receptors pathway. Both *Nr4a1* and *Sgk1* show an expression profile with a peak at t=0.6 in both the HF and LF group, however *Mapk7* and *Rps6kb2* are only up-regulated in the HF group. This difference might indicate the presence of an alternative mechanism through which *Nr4a1* can be activated. *Nr4a1* is a known transcriptional regulator of genes involved in glucose metabolism in mouse liver [25]. Because of the absence of any targets for this transcription factor in our interaction network we could not determine whether any downstream interactions to the Insulin signaling pathway are affected. Both the mRNA processing and Spliceosome pathways are significantly enriched for the differentially expressed genes at t=0.6 and together with the presence of several differentially expressed transcription factors and the HSP70 chaperones this indicates that the response to the glucose challenge at t=0.6 is mainly typed by a stress response and possibly transcriptional activation of genes to initiate the further response.

The most notable property of the pathway interaction networks for the glucose response at t=2 is the high centrality of the Cytokine pathway. This pathway summarizes different cytokines and their receptors, and since several of these receptors are differentially expressed it makes sense that this pathway fulfills a central role in the network, by connecting to the different processes these receptors play a role in. We identified a regulated interaction between several cytokine receptor pathways and the Endocytosis pathway, which might point to increased removal of receptor tyrosine kinases from the cell membrane in the HF group after ubiquitination with the up-regulated *Cbl* protein. Several of the proteins participating in the interactions with the cytokine receptor and Endocytosis pathways are known to be related to insulin or even annotated to the Insulin signaling pathway. For example, *Cbl* plays an important role in the CAP/Cbl/TC10 dependent transport of GLUT4 vesicles [26]. This could point to an interaction with the Insulin signaling pathway, but we did not identify a significant interaction in our network. Upon closer inspection, however, an interaction with the Cytokine receptor pathway has been found, but is just above our significance threshold ( $p = 0.0018$ ). This interaction involves *Cbl*, *Egfr* and several other proteins in the Insulin Signaling pathway, including three subunits of PI3K, *Crk* and *Igf1r*. Although this interaction was not considered significant, it might be biologically relevant given the high differential expression of *Cbl* and *Egfr*. One possible scenario is that the up-regulation of *Cbl* is an attempt to compensate for a defect in insulin signaling in the HF group, but causes side effects, such as interactions with the receptors as listed in the Cytokine receptor pathway.

The switch from glycogen storage to glycogen breakdown observed at t=48 can be expected after 48 hour following the glucose challenge. However, the link with the Insulin and Il-6 signaling pathways is interesting, since the Il-6 signaling pathway has already been linked to insulin actions and glycogen metabolism in vitro [27]. Another interesting observation is a difference in gene expression between the HF and LF groups in the interactions with the Proteasome pathway at t=48, which is possibly related to the TGF-beta pathway. Degradation



by the proteasome is an important mechanism in the TGF-beta pathway [28], which may be affected in different ways in the HF and LF groups. An identified interaction between the Wnt and Insulin signaling pathways might also be related to the observed changes in expression in the proteasome pathway, since *Gsk3b* is involved in degradation of beta-catenin by proteolysis, a process which can be inhibited by Wnt stimulation to stabilize beta-catenin, and which subsequently activates transcription of target genes [29]. Inspecting the expression profiles of *Ctnnb1* and its associated proteasome related genes shows a different expression in the LF compared to the HF group at t=48, so this mechanism could be affected by diet in the late response to the glucose challenge. Interestingly, one path between the Wnt and Insulin signaling pathways is the indirect path between *Gsk3b*, *Sgk3* and *Tsc1*, which was also observed in the comparison between the HF and LF group at t=0. At t=0 the expression of these three genes is lower in HF than in LF, but at t=48 the expression is almost equal. This might indicate a common long term stress response in the HF group that is also triggered in the LF group following the glucose challenge. *Sgk3* can phosphorylate *Gsk3b* and may function parallel to AKT in PI3K-dependent signaling [30]. Both *Akt2* and *Pik3c3* show a similar expression profile. *Pik3c3* encodes for the PI3K class III enzyme, involved in vesicular trafficking [31] and likely also in the immune system [32], but the activation mechanisms of PI3K class III enzymes are largely unknown.

As illustrated in the analysis described here, the method we developed is especially useful for exploring large and complex datasets to find interesting aspects and generate hypotheses. It can be used to add an extra dimension to differential gene expression and pathway enrichment results and may complement the search for activated or regulated genes and pathways by taking into account how they may interact. It allowed us to explore the NuGO PPS dataset and identify relevant gene and protein interactions. It also helped to hypothesize about underlying mechanisms and possible downstream effects of certain groups of differentially regulated genes, thereby providing starting points for more focused follow-up studies or experiments.

During the analysis we mainly followed a fixed approach to explore the pathway interactions consisting of three main steps. First, we studied the pathway interaction network globally to identify highly connected or central pathways and proteins in the network. This, for example, pointed us to the apoptosis pathways in the comparison between the two diet groups at t=0, which would not have been found by looking at pathway enrichment alone. In the second step, we interactively explored the interaction network using Cytoscape and zoomed in to specific pathway interactions or subgraphs by visualizing the protein interactions that form the paths between the pathways. This step is partly guided by the results of the previous step, since neighborhoods of the central pathways or proteins often point to interesting interactions. In addition, we looked for further interactions based on pathways that are known or expected to be involved in insulin resistance. For example, this led us to the interactions with the insulin signaling pathways in the late HF response, which were not central in the network but appeared to be relevant because of enrichment of all involved pathways and the central role of insulin in this analysis. Finally, we often referred back to the original pathway diagrams to understand the context of the interacting proteins within the pathway and look for any upstream or downstream effects (Figure 2). Together, these steps helped us to better understand the generated pathway interaction networks and the biology behind the transcriptomics dataset.

In addition to the new perspective this analysis provides on the dataset, there are several advantages that make this method a useful complement to existing pathway analysis techniques. Firstly, compared to an analysis limiting to pathway annotations, this combinatory analysis integrates additional information by using other sources of protein interactions and allowing indirect interactions including proteins that are not annotated to any pathway. This extends the coverage of the analysis with an additional 1660 genes that are in the interaction network but not annotated to any pathway, of which 929 are available in our transcriptomics dataset. This allowed us to move beyond well annotated knowledge in pathways, while still benefiting from the framework of canonical pathways and their intuitive visualizations. While the interaction network we used here is still relatively small compared to the number of interactions we expect to exist, current developments in measuring protein interactions [33] and identifying transcription factor targets [34] will further increase the coverage in the future. Secondly, when focusing on enrichment of gene sets or pathways alone, results might be missed since pathways may be relevant or activated without being enriched with significantly expressed genes, for example in case of post-translational regulation. For example, the three apoptosis related pathways identified in this analysis showed little differential expression, however given the number of other pathways it connects through via differentially expressed genes might indicate they may play a role in the context of this dataset. Finally, since genes and proteins often participate in different pathways and may have diverse functions, looking at individual pathway diagrams or lists of enriched pathways can be confusing or even misleading. Our method provides insight in the multiple roles of a protein in the context of the studied dataset. In case a gene is differentially expressed and also interacts with differentially expressed genes in multiple pathways, this will typically show up in the pathway interaction network, indicating that it acts on the verge of different pathways.

Besides several advantages, we also identified several possible improvements that may be made to this approach. Firstly, false positive pathway interactions were identified on occasion. One cause can be falsely annotated interactions between proteins in the input interaction network. This is the main reason we already excluded associations based on text-mining only from the STRING database, however remaining data may still contain false associations. For example, the component of metabolic pathways positioned around the Primary bile acid biosynthesis pathway turned out to be a false result after close inspection, due to falsely associated enzymes in the STRING database. Over time, the quality of interaction resources will probably improve, also helped by curation initiatives such as WikiPathways. Another cause of false or misleading results is the heterogeneity of many of the pathways. Pathway databases often include context specific pathways, which provide a summary of processes that are relevant to a specific disease or cell type or down-stream signaling of a specific messenger. However, the processes included in these pathways are often generic and play a role in other contexts as well, so finding an interaction with such a pathway does not automatically mean that its context is also relevant to the studied data. For example, we identified several interactions with the ESC Pluripotency pathway, which summarizes several signaling cascades regulating pluripotency in embryonic stem cells. It is very unlikely that embryonic stem cell functioning plays a role in our dataset, but still several interactions with this pathway were identified. These turned out to involve only specific parts of the signaling cascades, especially part of the Wnt signaling pathway. This is also a common problem in enrichment analysis, and our method does not provide a solution yet. Using a set of smallest generic core pathways with as

little overlap as possible might lead to cleaner results. Possible alternative sources of pathway information that might provide better results than the databases used here could be Gene Ontology [35] or Reactome [36], which provide hierarchical gene sets and pathways, allowing us to move down in the hierarchy to more specific modules if necessary. However, this would require adaptations to the methodology in order to deal with this hierarchical structure. Altogether, these points show that the analysis approach presented here (as well as most currently available pathway analysis method) is meant to be used for exploratory analysis and does not offer statistical proof for any of the findings. However, in the broader investigation of complex data towards more specific questions and testable hypothesis, this method provides a unique way to explore the data in a biological context.

The main shortcoming of using only transcriptomics datasets in pathway-based approaches is that these provide insight in changes at the mRNA level only. In this study, each identified pathway interaction is based on the assumption that if a group of interacting genes is consistently differentially expressed it is likely that changes are reflected at the protein level as well. However, this may not always be the case and it is not possible to investigate beyond this assumption using such datasets. In addition, a large part of signaling and pathway cross-talk probably works at a different level, through protein activation by phosphorylation or other post-translational (or post-transcriptional) regulatory mechanisms. If such interactions play a role, they cannot be identified with this dataset unless the interaction involves gene expression changes as well. The interactions that are most relevant to this dataset are transcriptional regulation, but unfortunately our interaction network contains relatively few transcription factor targets and we were unable to find a resource to increase this number in addition to the PAZAR database [37] used in this study. Despite this shortcoming, we were still able to show that it allows finding several relevant pathway interactions that help focus follow-up experiments.

Although in this particular study we used a transcriptomics dataset only, the method described here can directly be applied to other types of data or combinations of different data types as well. For example, the interaction network could be extended using the STITCH database [38], which defines interactions between chemicals and proteins. Thereby metabolomics data could be incorporated by providing weights for these interactions. In addition, some types of data can even be used to define more specific edge weights, depending on what action the interaction represents. For example, if protein kinase activity measurements are available, this data could be used to obtain more accurate edge weights of specific protein interactions that represent activation. This could then even be combined with proteomics data to incorporate protein abundance in the edge weights as well and thereby obtaining more specific interactions by narrowing down the possible substrates of a protein kinase. Future publicly available datasets might cover these types of data at a larger scale and the method described here may prove useful in analyzing this complex data at a functional level. To this end, all scripts and input data used for this analysis are open-source and freely available (see supplementary data).

## Conclusions

We designed an analysis approach to identify interactions between biological pathways in a specific context. By applying this method to a transcriptomics dataset, we identified relevant pathway interactions and possible key proteins involved in pathway cross-talk in the context of insulin resistant mouse liver, and at specific time points following a glucose challenge response. In addition, the analysis approach presented here can be applied to different types of high-throughput data and will be of more general use to facilitate interpretation of other complex datasets.

## Methods

### Microarray experiment design

The gene expression microarray data from the NuGO Proof of Principle Study 2 [8] was used. In this experiment, C57BL/6J mice were fed two different diets for 12 weeks, containing either 10% (LF) or 45% (HF) fat by energy. Animals fed the HF diet had developed insulin resistance by the end of this feeding trial. After 12 weeks, both diet groups were subjected to a glucose tolerance test and liver tissue was removed 0, 0.6, 2 and 48 hours after the glucose challenge and hybridized to NuGO Affymetrix Mouse GeneChip arrays (NuGO\_Mm1a520177). The resulting dataset consists of 8 biological replicates per diet and time point (totalling 64 samples). The data was normalized using the GC-robust multi-array analysis (GCRMA) algorithm [39] and probesets were redefined and annotated to Entrez Gene identifiers using the Custom CDF version 11.0.2 [40]. Full details of the experiment and preprocessing of the microarrays are described in more detail elsewhere [8,12]. The resulting microarray dataset is available from the ArrayExpress repository (<http://www.ebi.ac.uk/arrayexpress>, accession number: E-MTAB-601).

### Microarray gene-level statistics

The different groups were statistically compared using the *limma* R package [41]. To study the initial difference between the two diets, before the glucose challenge, we tested for differential expression between HF and LF groups at  $t=0$ . To study the response to the glucose challenge, we tested for differential expression between  $t=0$  and each other time point in the HF and LF groups. This resulted in the following statistics:

$T_{\text{HF vs LF, } t=0}$ : The moderated t-statistic, representing the significance and direction of differential expression between the HF and LF group at  $t=0$ .

$T_{\text{LF, } t_x \text{ vs } t_0}$ : The moderated t-statistic, representing the significance and direction of differential expression between each time point after the challenge ( $t=0.6$ ,  $t=2$  or  $t=48$ ) and the time point just before the challenge ( $t=0$ ) in the LF group.

$T_{\text{HF, } t_x \text{ vs } t_0}$ : The moderated t-statistic, representing the significance and direction of differential expression between each time point after the challenge ( $t=0.6$ ,  $t=2$  or  $t=48$ ) and the time point just before the challenge ( $t=0$ ) in the HF group.

Based on the corresponding p-values as provided by the *limma* package, q-values were calculated using the *qvalue* package [42] to correct for multiple testing.

## Pathways

Pathways from WikiPathways (analysis collection, v2010-12-07) [10] and KEGG (v2010-06-14) [11] were used. Pathways categorized under “Human diseases” at the KEGG website were excluded from the analysis.

Pathways with high overlap in protein content were merged into a single pathway using a step-wise procedure:

1. Calculate overlap between each pathway pair.
2. For each pathway, find other pathways that have  $\geq 75\%$  overlap and merge it with the most overlapping pathway.
3. Repeat step 1-2 until no pathway pair with  $\geq 75\%$  overlap is left.

## Pathway enrichment

To find pathways that are enriched with genes that score high in the differential expression test, the *geneSetTest* from the *limma* package was used. This test calculates a test statistic for each pathway by taking the mean of the absolute T-statistic of the genes in the pathway and calculates P-values by comparing the test statistic to an empirical null distribution based on 10,000 permutations with random gene sets. The resulting p-value for each pathway represents the significance of its enrichment with differentially expressed genes, regardless of their direction.

## Interaction network

The interaction network used in this study combines the following databases:

- Protein functional associations from the STRING database (v8.3) [14]. This database aggregates interactions from several protein-protein interaction repositories and pathway databases. Interactions based on text mining alone and those with a confidence score  $< 0.4$  were excluded.
- Transcription factor targets from the PAZAR database (v2010-08-15) [36].
- Manually curated reactions and interactions from the pathways described in the section ‘pathways.’

These interactions were merged into a single directed network. For interactions without a defined direction (e.g. binding), two directed edges in opposite direction were added.

## Gene and protein identifier mapping

Genes and proteins were mapped to a common database using the BridgeDb library [43] and accompanying synonym databases. All probesets in the transcriptomics dataset and genes and proteins in the interaction network were mapped to Ensembl gene identifiers using the *Mm\_Derby\_20090720.bridge* database. These mapped gene and protein identifiers will be referred to as *xref* in the following description of the algorithm.

## Finding interactions between pathways

To find possible interactions between pathways, we performed the following steps:

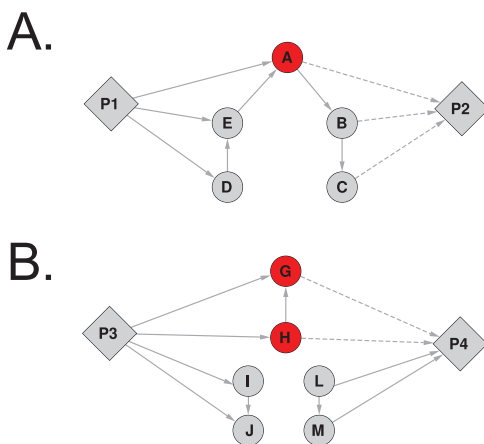
1. A graph  $G_{PX}$  containing both the *xref* interactions and pathway-*xref* associations was created. This graph contains two types of nodes, either representing an *xref* or a pathway. Edges were added between the *xref* nodes based on the interaction network. In addition, each pathway node was connected to each of the *xref* nodes that are associated with that pathway.
2. Edge weights were calculated for  $G_{PX}$ , based on one of the gene-level statistics described in the section 'Microarray gene-level statistics'. First, a transformation using a sigmoid function was applied to the T statistic:

$$f(T) = 1 - \frac{1}{1 + e^{-\alpha(|T| - \mu)}}$$

This function ensures that the transformed values will range from 0 to 1 and applies a soft threshold, where the threshold center is determined by  $\mu$  and the steepness by  $\alpha$ . To transform the T statistic,  $\mu=3$  and  $\alpha=2$  were used. As result, genes with an absolute T statistic  $\geq 3$  (corresponding to unadjusted p-value  $\leq 0.004$ ) were emphasized by receiving a lower weight (Supplementary figure 16).

Secondly, edge weights were assigned based on the transformed T statistic  $f(T)$  of the target *xref* nodes they connect to, or 1 if no data was available for the target *xref*. If the target of an edge was a pathway node, the weight was set to 0.

3. For each pathway node  $P_n$  in  $G_{PX}$  a subgraph  $G_{P_n}$  was created containing only nodes that could be reached within  $nb$  steps form  $P_n$ . In this study, parameter  $nb$  was set to 5, limiting the number of edges between *xref* nodes in a path to maximal three.



**Figure 4:** Two examples of overlap between pathways with respect to the algorithm for finding pathway interactions. Red nodes are proteins that are present in both pathways, dashed edges will be removed before finding shortest paths. A: A pathway pair with one overlapping protein (protein A) which connects to distinct paths within the two pathways. The overlap itself is not counted as interaction between the pathways, but by removing only the dashed edge A still acts as an indirect intermediate to establish a path between P1 and P2 via proteins B and E. B: A pathway pair which shares two proteins that do not connect to distinct paths within each pathway. In this case these proteins do not contribute to the pathway interaction, because the dashed edges are removed.

4. For each pathway node  $Pm$  in  $G_{Pn}$  another subgraph was created containing nodes  $Pn$ ,  $Pm$ , the direct neighbors of  $Pn$  and  $Pm$  (the  $xrefs$  in these pathways) and all  $xref$  nodes that are not direct neighbors of a pathway node ( $xrefs$  that are not in any pathway). Excluding  $xref$  nodes that are present in pathways other than  $Pm$  and  $Pn$  is necessary to prevent detection of false interactions between  $Pn$  and  $Pm$  that cross other pathways. Such an interaction should not be represented as direct interaction between  $Pn$  and  $Pm$ , but rather as two interactions with the other pathway as intermediate between  $Pn$  and  $Pm$ . For  $xref$  nodes that were connected to both  $Pn$  and  $Pm$ , the edge to  $Pm$  was removed to prevent paths to go directly through an overlapping  $xref$  (Figure 4).

5. The shortest path between  $Pn$  and  $Pm$  in  $G_{Pn}$  was found, taking into account the edge weights (the length of a path is the sum of its edge weights). The path length was stored and the edges of the path, except those that connect  $Pn$  or  $Pm$ , were deleted from  $G_{Pn}$  (Figure 1C).

6. Step 5 was repeated until a preset maximum path length  $l_{max}$  was reached (Figure 1C). In this study,  $l_{max}$  was set to 0.9, ensuring that paths between  $xrefs$  with low significance or no measured data (which will have a weight close to 1, see step 2) do not contribute to the interaction score.

7. For each pathway pair  $P_n P_m$  a list of path lengths  $l_i$  is now available. These were converted to a single interaction score  $S_{PnPm}$  for each pathway pair by calculating the sum of the inverted lengths:

$$S_{PnPm} = \sum_i \frac{1}{l_i}$$

8. Significance of the interaction was calculated by comparing the interaction score to an empirical null distribution. This null distribution was generated by repeating step 3-7 to recalculate an interaction score between  $Pn$  and  $Pr$  (a randomized version of  $Pm$ ), resulting in a collection of randomized interaction scores  $S_{PnPr}$ . For each permutation,  $Pr$  was generated by reconnecting the endpoints of its adjacent edges (which are the nodes representing the  $xrefs$  associated to the pathway) to different  $xref$  nodes from the interaction network that have a similar strength (also weighted degree [44]) as the original  $xref$  node. This way, the randomized pathway has a connectivity that is similar to the original pathway. Empirical p-values were calculated for each pathway pair by counting the number of times  $S_{PnPr}$  is higher or equal to  $S_{PnPm}$  and dividing this by the number of permutations. Firstly, p-values for each pathway pair were calculated using 100 permutations. Secondly, the resolution of the p-value of potentially significant pathway pairs was increased by adding another 1000 permutations for pairs that had a p-value  $< 0.1$  after 100 permutations.

This algorithm was implemented in the statistical language *R* [45] and uses the *igraph* package [46] for handling graph structures and finding shortest paths between two nodes. The *R* source code and utility scripts to generate and format the data are available at <http://code.google.com/p/tkelder/wiki/PathwayInteractions>.

Parameter  $nb$  limits the maximum allowed number of edges in a path, which significantly reduces the calculation time. This also filters out paths with many edges that are less likely to be biologically relevant, even though their weighted path length might be low.



To reduce the contribution of overlapping proteins between pathways to the interactions between pathways, the edge from an overlapping protein to one of the pathways in a pair is removed in step 4. While pathways with more than 75% overlap were merged, some pathways might still have much overlap that may outweigh the influence of pathway interactions via protein interactions. However, to completely discard overlapping genes might be too conservative, since an overlapping protein may be a key regulator that might connect two distinct parts of the different pathways. For example, in Figure 4A, two pathways both contain protein A, which connects to a distinct path in each pathway. If protein A would be excluded from the analysis (as in Li et al. [6]), a relevant interaction between the two distinct paths will be missed. However, in Figure 4B, two pathways share a complete path of three proteins which does not connect to any other distinct part within each pathway and hence these proteins are unlikely to contribute to an interaction between the pathways. In this case, our approach of removing one of the edges with the pathway will give equal results as excluding the proteins.

## Network analysis and visualization

All networks were visualized using Cytoscape [13]. Network properties such as average path length, clustering coefficient, node degree and betweenness centrality were calculated using the *igraph* library in R. Betweenness centrality as displayed in Supplementary figure 3 was normalized by dividing the value calculated with the *igraph* function *betweenness* by the maximum centrality, which is the total number of node pairs in the network, excluding the node for which the betweenness centrality is calculated.

## Author contributions

Designed and performed computational analysis: TK. Designed and performed transcriptomics experiment: MvE RK TeK. Normalization and quality control of transcriptomics data: MvE. Extensive reviewing and editing of the manuscript: CE, LE. Drafted the manuscript: TK. All authors read and approved the final manuscript.

## Acknowledgements and funding

We thank Dr. Tony Travis at the University of Aberdeen Rowett Institute of Nutrition and Health (RINH) and Biomathematics and Statistics Scotland (BioSS) for technical assistance on using the RINH/BioSS Beowulf cluster (<http://bioinformatics.rii.sari.ac.uk>). This work was supported by the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre. The authors gratefully acknowledge additional grant support from the European Nutrigenomics Organisation (NuGO, CT-2004-505944; Proof-of-Principle Study PPS2).

## References

1. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1-13.
2. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, et al. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25: 1091-1093.



3. Isserlin R, Merico D, Alikhani-Koupaei R, Gramolini A, Bader GD, et al. (2010) Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. *Proteomics* 10: 1316-27.
4. Kleemann R, Erk M van, Verschuren L, Hoek AM van den, Koek M, et al. (2010) Time-resolved and tissue-specific systems analysis of the pathogenesis of insulin resistance. *PloS One* 5: e8817.
5. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 23: 561-6.
6. Li Y, Agarwal P, Rajagopalan D (2008) A global pathway crosstalk network. *Bioinformatics* 24: 1442-7.
7. Huang Y, Li S (2010) Detection of characteristic sub pathway network for angiogenesis based on the comprehensive pathway network. *BMC Bioinformatics* 11 Suppl 1: S32.
8. Baccini M, Bachmaier E-M, Biggeri A, Boekschoten MV, Bouwman FG, et al. (2008) The NuGO proof of principle study package: a collaborative research effort of the European Nutrigenomics Organisation. *Genes Nutr* 3: 147-151.
9. Hotamisligil GS (2006) Inflammation and metabolic disorders. *Nature* 444: 860-7.
10. Pico AR, Kelder T, Iersel MP van, Hanspers K, Conklin BR, et al. (2008) WikiPathways: pathway editing for the people. *PLoS Biol* 6: e184.
11. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27-30.
12. Rubio-Aliaga I, Roos B de, Sailer M, McLoughlin GA, Boekschoten MV, et al. (2011) Alterations in hepatic one-carbon metabolism and related pathways following a high fat dietary intervention. *Physiol Genomics*.
13. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* 13: 2498-2504.
14. Kadowaki T, Tobe K, Honda-Yamamoto R, Tamemoto H, Kaburagi Y, et al. (1996) Signal transduction mechanism of insulin and insulin-like growth factor-1. *Endocrine J* 43 Suppl: S33-41.
15. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 37: D412-6.
16. Möller G, Leenders F, Grunsven EG van, Dolez V, Qualmann B, et al. (1999) Characterization of the HSD17B4 gene: D-specific multifunctional protein 2/17beta-hydroxysteroid dehydrogenase IV. *J Steroid Biochem Mol Biol* 69: 441-6.
17. Yang S-Y, He X-Y, Schulz H (2005) 3-Hydroxyacyl-CoA dehydrogenase and short chain 3-hydroxyacyl-CoA dehydrogenase in human health and disease. *FEBS J* 272: 4874-83.
18. Schmidt MHH, Dikic I (2005) The Cbl interactome and its functions. *Nat Rev Mol Cell Biol* 6: 907-18.
19. Schattenberg JM, Schuchmann M (2009) Diabetes and apoptosis: liver. *Apoptosis* 14: 1459-71.
20. Civera M, Urios A, Garcia-Torres ML, Ortega J, Martinez-Valls J, et al. (2010) Relationship between insulin resistance, inflammation and liver cell apoptosis in patients with severe obesity. *Diabetes Metab Res Rev* 26: 187-92.

21. Muegge K, Vila M, Gusella GL, Musso T, Herrlich P, et al. (1993) Interleukin 1 induction of the c-jun promoter. *Proc Natl Acad Sci U S A* 90: 7054-8.
22. Zimitat C, Nixon PF (2001) Glucose induced IEG expression in the thiamin-deficient rat brain. *Brain Res* 892: 218-27.
23. Ungewickell E, Ungewickell H, Holstein SE (1997) Functional interaction of the auxilin J domain with the nucleotide- and substrate-binding modules of Hsc70. *J Biol Chem* 272: 19594-600.
24. Chung J, Nguyen A-K, Henstridge DC, Holmes AG, Chan MHS, et al. (2008) HSP72 protects against obesity-induced insulin resistance. *Proc Natl Acad Sci U S A* 105: 1739-44.
25. Chao LC, Wroblewski K, Zhang Z, Pei L, Vergnes L, et al. (2009) Insulin resistance and altered systemic glucose metabolism in mice lacking Nur77. *Diabetes* 58: 2788.
26. Chiang SH, Baumann CA, Kanzaki M, Thurmond DC, Watson RT, et al. (2001) Insulin-stimulated GLUT4 translocation requires the CAP-dependent activation of TC10. *Nature* 410: 944-8.
27. Senn JJ, Klover PJ, Nowak IA, Mooney RA (2002) Interleukin-6 induces cellular insulin resistance in hepatocytes. *Diabetes* 51: 3391-9.
28. Zhang F, Laiho M (2003) On and off: proteasome and TGF- $\beta$  signaling. *Exp Cell Res* 291: 275-281.
29. Jope RS, Johnson GVW (2004) The glamour and gloom of glycogen synthase kinase-3. *Trends Biochem Sci* 29: 95-102.
30. Dai F, Yu L, He H, Chen Y, Yu J, et al. (2002) Human serum and glucocorticoid-inducible kinase-like kinase (SGKL) phosphorylates glycogen synthases kinase 3 beta (GSK-3beta) at serine-9 through direct interaction. *Biochem Biophys Res Commun* 293: 1191-6.
31. Leegers SJ, Vanhaesebroeck B, Waterfield MD (1999) Signalling through phosphoinositide 3-kinases: the lipids take centre stage. *Curr Opin Cell Biol* 11: 219-25.
32. Koyasu S (2003) The role of PI3K in immune cells. *Nat Immunol* 4: 313-9.
33. Bonetta L (2010) Protein-protein interactions: Interactome under construction. *Nature* 468: 851-854.
34. Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10: 605-16.
35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-9.
36. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 39: D691-697.
37. Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, et al. (2009) The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res* 37: D54-60.
38. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, Mering C von, et al. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res* 38: D552-6.
39. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat*

- Assoc 99: 909-917.
40. Dai M, Wang P, Boyd AD, Kostov G, Athey B, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33: e175.
  41. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3.
  42. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100: 9440-5.
  43. Iersel M van, Pico A, Kelder T, Gao J, Ho I, et al. (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11: 5.
  44. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci U S A* 101: 3747-52.
  45. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. Available: <http://www.r-project.org>.
  46. Csárdi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal Complex Systems* 1695: 1695.

# CHAPTER 6

## General Discussion

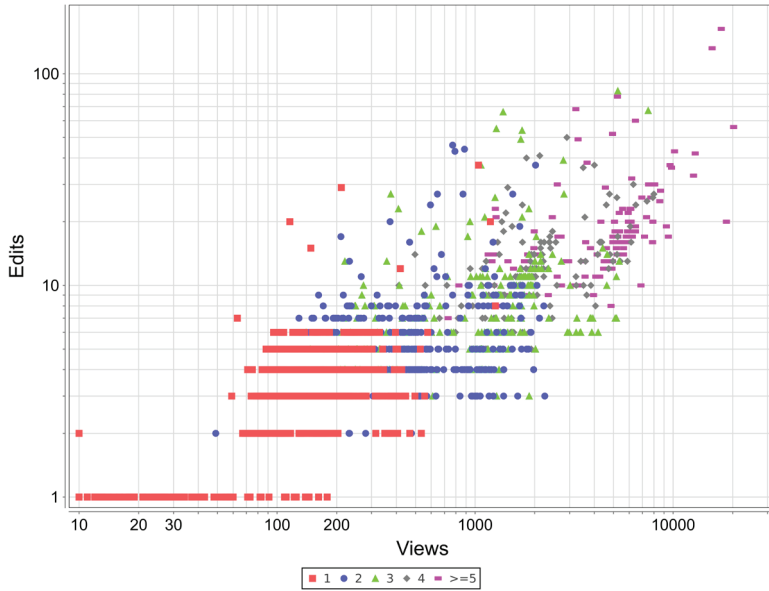
Integration of existing knowledge in exploratory data analysis is an important step to make optimal use of the current possibilities to generate large datasets in biology research and to understand the underlying biology. This thesis aims to improve this step by utilizing the concept of biological pathways. In chapter 2 several hallmarks of the exploratory data analysis paradigm were applied to biological pathways and chapters 3 and 4 presented two bioinformatics tools that were designed to enhance the ability to include pathways in exploratory data analysis. Chapter 5 describes an application of an exploratory analysis using the developed resources integrated with additional public data. This chapter will evaluate the developed tools in context of the key aspects of exploratory data analysis.

## **Maintaining a collection of pathways**

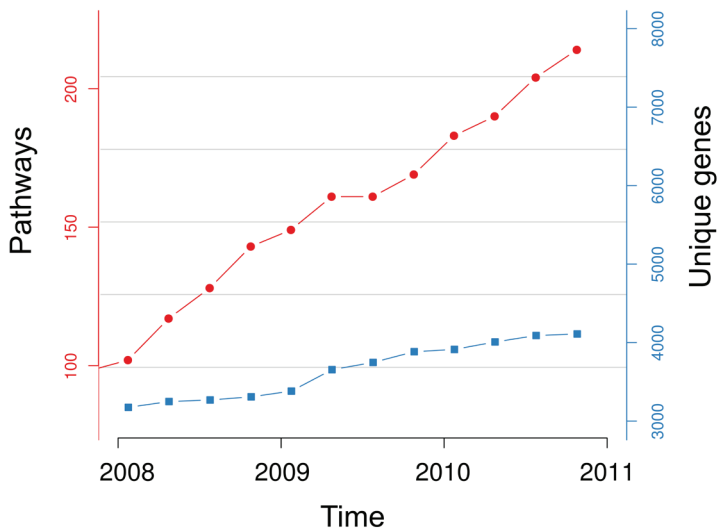
One of the prerequisites for effectively utilizing biological pathways in a data analysis is the availability of an open repository of digitalized pathways that can be applied to bioinformatics approaches. Curating and maintaining a public collection of pathways is a large and never-ending task, as a continuous stream of new knowledge is being generated. Given the current exponential growth of knowledge, centralized curation as employed in existing pathway resources does not scale [1]. The WikiPathways database described in this thesis provides an original approach to maintaining a pathway collection. It offers a free and open pathway repository in the form of a wiki that makes it possible for any researcher to take on the task of curating the content. WikiPathways and the wiki approach applied to curation of biological information in general, received positive attention in commentaries in several widely read scientific journals [2-6], indicating that the research community is open to novel solutions towards more scalable curation.

Wikis provide an effective platform for community-based curation. Users can directly contribute, update and correct the content, which in the long term improves comprehensiveness and quality. The mechanism behind this can be described as a positive feedback loop [7]. The wiki starts with some initial content which attracts users. A small fraction of these users will also correct or extend the content, which on average leads to an improvement of the quality and usefulness of the content. This will attract more users, thereby also increasing the number of potential editors, which improves the content even more. An example of how this mechanism has worked very well is Wikipedia, which has led to a comprehensive encyclopedia that compares in quality to more established, centrally curated alternatives [8]. Based on statistics gathered during the short history of WikiPathways, it might be possible to obtain more insight in how well the wiki approach has worked so far for biological pathways.

If a relation would exist between usage and contribution in WikiPathways, the number of edits to a pathway is expected to increase with the number of views. Indeed, pathways that have many views are also among the ones with the most edits and have been edited by a larger group of authors (Figure 1). Following the positive feedback mechanism, usage would also result in a growth of content. Since the number of site visits has increased from over 1100 per month over the three months prior to publication to almost 5400 per month over the last three months of 2010, a growth and improvement of content would also be expected. Indeed, compared to the initial content of WikiPathways, the number of human pathways has grown with 128% and the number of annotated human genes in these pathways has increased with 30% (Figure 2). To put this number in perspective, the pathway collection for GenMAPP, on



**Figure 1:** Relation between views and edits for the pathways at WikiPathways. Each dot represents a pathway, the x-axis represents the number of times the pathway has been viewed and the y-axis represents the number of times the pathway has been edited. The color and shape of each dot indicates the number of unique users that edited the pathway.



**Figure 2:** Two indicators of WikiPathways growth over the period 2008-2010 for human specific content. The red line and left axis represents the number of human pathways. The blue line and right axis represents the number of unique human genes across all pathways.

which the initial WikiPathways content was based, grew with only 1% in number of pathways and 5% in number of genes in the three years preceding WikiPathways [9]. In addition, since the number of pathways in KEGG, a comparable, but centrally curated resource, has grown with 14% in the same period, the community-based curation mechanism seems quite effective.

In addition to the usage and content growth of WikiPathways, the size and activity of its community have also grown. In 2008, on average 10 users per month made one or more edits to a pathway and 87 edits were made per month. These numbers have grown to on average 16 editing users and 261 edits per month over 2010, a growth of 56% and 200% respectively. Although the active community shows growth, it remains relatively small, probably too small to effectively keep up with the growth of biological knowledge that can be captured in pathways. The size and activity of the community can be improved in two main ways that reinforce each other. Firstly, the part of users that become contributor can be improved to increase the activity of the community. Secondly, improving usability of the pathways will increase the size of the community, since the number of users is proportional to utility of the content.

The barrier to contribute seems fairly high, since on average only 0.36% of the website visitors actually edited a pathway one or more times. However, compared to Wikipedia, a wiki with a very active community, the barrier is not that high. For the English Wikipedia, only 0.02-0.03% of the visitors is active contributor (defined as at least 5 edits in a given month [10]) and for WikiPathways this translates to almost 0.19% averaged over 2010. However, in contrast to Wikipedia, the content at WikiPathways is focused on a smaller domain and a large part of the target audience is expert in this domain. Therefore, it will probably be worth trying to lower this barrier even further. One improvement could be on speed, stability and user-friendliness of the editor. To optimally re-use code and programming efforts, the current editor is running on Java applet technology and directly based on the stand-alone application PathVisio [11]. This causes long load times, lacks integration with the browser, and requires Java to be installed and properly configured for the web browser. A better solution would be to implement an editor as a pure web application which is fully integrated in the webpage that displays the pathway. This way, the editor as separate entity will disappear and each part of the page will be directly editable, improving the consistency of the user interface and decreasing load times. The new HTML5 standard [12] that is currently being adopted by most browsers may provide the right platform.

Another way to lower the barrier is to assist and challenge users to contribute by pointing out which parts of the content are in need for improvement. For example, some potential errors, such as wrong or missing gene annotations, can be detected automatically and presented to the user. Several automated scripts called 'bots' have already been implemented to find such errors and put a banner on the pathway page providing instructions on how to fix them. Such corrections are generally straightforward, take relatively little time and may therefore lower the barrier for new contributors. Another approach, taken by the TFe wiki project [13], could be to add a clear indicator of completeness of the content. This assigns a completion percentage to each page and lists steps that can be taken to increase this score. For pathways, such a score could be based on basic parts such as a textual description, categorization or reference to literature. Again, this approach could encourage users to contribute by splitting up the curation into small, manageable tasks. Finally, an approach taken by Gene Wiki project [7]

to improve curation of gene articles at Wikipedia is the automatic creation of initial pages, called 'stubs'. These pages are typically based on content that is automatically extracted from authoritative sources and provide an initial starting content. Since the barrier for starting a completely new article is usually higher than extending an existing article, this greatly increased the number of contributors [14]. This approach will be harder to apply to pathways, since in contrast to genes, there is no complete list of possible pathways. One option might be to find proteins that are not annotated in any pathway, but do occur with a high frequency scientific publications, indicating that they represent relevant knowledge yet to be translated into a pathway. Small pathway stubs could then be generated for these proteins by for example including potential interacting proteins from sources such as STRING [15] or GeneMania [16]. Although these will probably not reflect a proper pathway and contain false positive or non-relevant interactions, this might challenge and engage users to correct or extend these stubs.

Since the release of WikiPathways, several use cases have emerged that illustrate the utility of pathway content at WikiPathways. Defining these use cases and providing tools to support them might help to attract new users and grow the active WikiPathways community. WikiPathways has mainly been used as resource, publishing tool, collaborative editor and educational tool. The following sections will highlight examples of these use-cases and pointers to improve them.

## **WikiPathways as resource**

WikiPathways can be used as resource for researchers in two ways. Firstly, the website provides a reference for biological knowledge. Each pathway page provides a useful starting point to lookup knowledge on a specific biological mechanism, including the pathway diagram, hyperlinks to detailed information about genes, proteins and metabolites, and relevant literature references. Website statistics indicate that WikiPathways is indeed being used this way, since over 40% of the visits originated from Google searches on keywords related to biology, and the currently most visited pathway (One Carbon Metabolism [17]) has been visited over 11,000 times. Secondly, WikiPathways can be used as repository for pathways for bioinformatics research and exploratory data analysis. Pathway archives can be downloaded in computable formats and are also distributed with tools such as GenMAPP [18] and Path-Visio [11]. With these tools, researchers can perform overrepresentation analysis to prioritize biological pathways based on the data and visualize the data on pathway diagrams. This approach has been taken in several studies [19-21], including two studies where I participated as bioinformatics researcher in the exploratory data analysis phase [22,23]. In addition to tools that bundle WikiPathways content, the web service as described in chapter 4 provides a resource of pathway information via a programmatic interface, allowing integration in tools and scripts. This web service processed over 45,000 requests per month over 2010, excluding requests from maintenance scripts, indicating that WikiPathways is also being used as resource for bioinformatics research.

A common concern about wikis is the difficulty to ensure quality and accuracy of the content, because of the lack of an appointed group of expert curators. Therefore providing tools to get background information on the quality of the content may make it more attractive and convince skeptical users to use WikiPathways as a resource. A mechanism that has been very effective so far is organization of the content into several quality levels via revision specific



tagging of pathways. Currently, the two main quality levels are featured pathways and analysis collection pathways, where the latter may be less well curated and visually appealing, but do have properly annotated genes and proteins thereby being useful for analysis. Download archives provide snapshots of a subset of pathways and specific revisions that correspond to the quality level. While these quality levels apply to whole pathways, a more sophisticated method would be to provide levels of accuracy on specific parts of the pathway. Algorithms to derive this information have already been implemented to process textual content for Wikipedia by WikiTrust [24] and for WikiGenes [25]. Both approaches are based on tracking text origin and author information, making it is possible to trace back each part of the text to the original author and revision and thereby providing a tool to judge quality of the content. WikiTrust also uses this mechanism to provide a trust score for each part of the text, based on the assumption that content which is stable over different revisions is more trustworthy than content that undergoes many changes. It also includes author reputation based on the content of the wiki, which increases when the author's contributions are preserved and decreases when they are reverted. It might be possible to adapt this algorithm to work with pathways. Being able to find the origin, authors and revision history of a specific interaction or protein will be a powerful tool to judge trustworthiness of pathway information in analyses. For example, in the analysis described in chapter 5 of this thesis, a trust score for interactions could be integrated in the edge weights to reduce false positive results.

## **WikiPathways as publishing tool**

Pathway diagrams are widely used in scientific publications as figures accompanying the text. These figures are being translated into annotated pathways in digital form, however this is a rather tedious and often error prone task. Using WikiPathways, authors could directly create a properly annotated pathway and either save it as image to include in a publication or provide it as supplementary material. Several authors have already taken this approach [26,27], which offers additional advantages from the author's point of view. Firstly, the pathway page at WikiPathways improves the reader experience, by providing links to additional detailed information and unambiguously annotated genes, proteins and metabolites. In contrast, when using a static image, manual searching is required to find more information about a protein on the diagram and ambiguous protein names may lead to incorrect interpretations. Secondly, the online pathway improves visibility and usability of the author's results, by making it searchable via internet search engines such as Google and redistributing it with several bioinformatics tools for use in data analyses of other researchers. When such analyses lead to new publications, the original source of the pathway can be tracked and cited, thereby increasing its measurable impact. Finally, once at WikiPathways the pathway can be updated by both the original author and community based on new research results that may become available over time to provide a better representation of the biological mechanism.

The usability of WikiPathways as publishing tool has recently been improved by implementing the possibility to temporarily hide a pathway from public view. Thereby authors can retrieve a stable identifier that can be used to reference the pathway from the manuscript and allow reviewers to inspect the pathway, without giving away their results before publication. Another way to support publication better is to allow more advanced graphical styles in the pathway diagrams. Currently, the style used in WikiPathways is focused on data analysis and very minimal to support data visualization, while the typical style in publications is more

colorful and contains more complex illustrations to clarify mechanisms of action. Although the GPML format is internally very flexible and allows creation of complex graphics, drawing these is not straightforward in the editor. WikiPathways might be able to reach a wider audience if the editor would provide more control over curved paths, allow inclusion of custom graphics or provide a set of templates that match styles commonly used in publication better.

## **WikiPathways as collaborative editor**

WikiPathways can be used by a group of researchers to build and maintain a set of focused pathways supporting ongoing projects or research collaborations. For example, the Micronutrient Genomics project [28] aims to provide a resource for knowledge on the biological context of micronutrients and uses WikiPathways to collaboratively edit a subset of pathways related to these topics. A core team of experts builds pathways and streamlines contributions from the community. Along the same lines, the California Institute for Regenerative Medicine (CIRM) has adopted WikiPathways to highlight a subset of pathways contributed by the stem cell research community. For initiatives like these, WikiPathways provides the option to create a portal page, which provides an access point to the subset of pathways and can be customized with the project logo and announcements. Such portals make the content more attractive for a specific group of users because it provides a more convenient entry point that is focused on their research subject.

Another way WikiPathways can be used as a collaborative editor is by providing a framework for collecting community contributions for centrally curated databases. In this setup, the content of the database is mirrored on WikiPathways which provides a medium for its users to contribute corrections or additions. This way, WikiPathways complements centrally curated databases by providing a staging ground for new content that can then be reviewed by appointed curators for inclusion in the database. The Reactome database is currently in the process of setting up this workflow using WikiPathways to improve their ability to collect community contributions [29]. A similar approach is taken by the maintainers of the PluriNetwork [30], an electronic resource for curated protein interactions relevant to pluripotency, for which a version is maintained at WikiPathways that allows users to contribute new interactions. Partnering with other pathway or interaction resources directly increases the community and contribution rate of WikiPathways. In addition, it increases database inter-compatibility and makes each contribution from the community more valuable, because it will eventually be distributed over different pathway resources.

## **WikiPathways in education**

Building a pathway diagram from existing knowledge is a good educational exercise and works well in combination with already existing literature research projects for undergraduate students. By creating a pathway, students learn about the biology and become familiar with bioinformatics tools and resources, which are important for analyzing experimental data in their future research career. WikiPathways has been used for educational projects and courses at Maastricht University at a small scale. Several pathways at WikiPathways are direct results of undergraduate projects [31,32] where a student performed an extensive literature review about a specialized topic. Other pathways resulted from shorter teaching sessions where multiple students worked on a single pathway [33,34]. Recently, the EcoliWiki

project has adapted the education model at a larger scale, introducing a competition-based element that comes closer to the community curation model [35]. In this project, students contribute content to the wiki, but are also stimulated to correct the contributions of their fellow students. This directly benefits motivation of the students and improves quality of the contributions. In addition, exposing students to the principles of a wiki helps cultivating a future research community that is better prepared for an open-data culture and new ways of sharing and curating knowledge.

Altogether, the website statistics and the four use cases discussed here, indicate that WikiPathways is being adapted by the research community as a versatile tool. Eventually, both usage and contribution need to reach a critical mass to establish a stable active community that can keep up with the continuously growing amount of knowledge and publications. As each of the different use cases develops further and new ways of using the content emerge, it is likely that the community will continue to grow and contributions will increase. The open nature of the WikiPathways project (both in content, code and collaborations) allows any user group to adapt and implement features to support a specific use case, and stimulates new communities to adapt WikiPathways for their research. In the end, this will lead to a better representation of our knowledge as biological pathways, and contribute to improving exploratory pathway analysis.

## **Integration of pathways in bioinformatics tools and workflows**

To effectively utilize pathway content in exploratory data analysis it should be possible to integrate it into bioinformatics tools, scripts, workflows and to combine it with other types of information and data. The web service described in chapter 4 provides an infrastructure for programmatic access to WikiPathways content. In addition to its utility in assisting curation tasks (the Reactome workflow and the bots as discussed before are both supported by the web service), the web service was also used for integration of pathway information in external applications. Several online resources and databases integrate pathway information from WikiPathways, including NCBI BioSystems [36], BioPortal [37] and CitedIn [38]. NCBI BioSystems integrates different pathway resources with the Entrez database. This allows researchers to directly view pathway annotations at different NCBI resources, for example a gene page at Entrez Gene or a compound page at PubChem. The CitedIn website provides a novel measure of research impact based on internet citations and uses WikiPathways as a citation resource, by querying literature references from pathways via the web service. BioPortal is a resource that integrates and annotates various ontologies which organize concepts in topics such as taxonomy, tissues or diseases. It uses WikiPathways to derive functional annotations for ontology terms. These examples demonstrate how the WikiPathways web service supports online resources in extracting different types of information from pathways to enrich their content.

Another use case that directly benefits exploratory pathway analysis is the integration into bioinformatics tools and workflows. An example of a web application that uses the WikiPathways web service for pathway analysis is WebGestalt [39], which allows researchers to find out which pathways are overrepresented with genes from a user-specified input list, such as a list of differentially expressed genes. This web application uses WikiPathways as one of the pathway resources and uses the built-in visualization method of the web service to directly

visualize the results on a pathway diagram in the web browser. An example of an offline tool that integrates WikiPathways content via the web service is DomainGraph [40]. DomainGraph is implemented as a plugin for the widely used network analysis tool Cytoscape [41] and can be used to visualize alternative-splicing data. To help finding possible functional effects of alternative splicing it links the data to biological pathways and directly visualizes it on pathway diagrams from WikiPathways. These examples show that the web service improves integration of pathways in different specialized analysis tools.

The WikiPathways web service is only part of a bigger trend of making data available through web services [42]. This will significantly enhance exploratory data analysis by improving unified data access from different tools and facilitating integration of different resources in an analysis workflow. For example, Cytoscape already provides a set of web service clients to import and integrate data from different resources, including the WikiPathways. This allows tool developers to focus on the core features of a new analysis method and minimize time spent on implementing data import features. For example, the ExprEssence plugin for Cytoscape [43] only implements new analysis features and uses the functionality of the already existing web service clients to support different pathway resources including WikiPathways. Design choices that utilize web services also improve homogeneity of data import procedures from a user perspective, increasing the user friendliness of a tool and minimizing the need to manually download data and convert between different data formats.

## **Exploratory analysis to find interactions between pathways**

A hallmark of exploratory data analysis is to reformat and present the data in different ways to provide new perspectives that allow different observations which together help developing vague questions and ideas into more focused and testable hypotheses. The analysis described in chapter 5 presents a new pathway-centered perspective on the data that complements existing methods such as enrichment analysis. By finding regulated paths between pathways, it shows the dataset from a higher perspective, showing how functional modules might interact, cooperate and regulate each other to generate a complex phenotype. Zooming in to the protein interactions composing these paths also provides possible mechanisms for pathway interactions at a more detailed level. This has led to several interesting observations about activated mechanisms of pathway cross-talk. It also made it possible to identify relevant pathways that were not enriched with differentially expressed genes, but did contain many regulated paths between other pathways. By limiting to established analysis techniques such as enrichment analysis, these aspects of the data would most likely not have been observed. Although this analysis approach alone did not lead to a full understanding of the data, when combined with methods that provide different perspectives it might be possible to gather enough pieces of the puzzle to better explain the complex biology underlying the dataset.

From a developer point of view, the analysis described in chapter 5 demonstrates the importance of flexibility and reusability of available tools and methods. Firstly, this approach combines several different types of information, including pathways and interactions from different resources. Importing and mapping identifiers was achieved with minimal scripting, because the bulk of the work had already been implemented by the BridgeDb [44] and PathVisio [11] libraries. Furthermore, graph algorithms and visualization tools were readily available from packages such as igraph [45] and Cytoscape. Therefore, the overhead of

implementing functionality secondary to the core analysis method could be kept to a minimum. Secondly, the analysis method itself has grown out of a simpler visualization method that was implemented ad-hoc during preceding data analyses [22,46], which was only feasible because the main parts were already available as open-source libraries and tools. Based on these preliminary results, a decision could be made whether to develop it further into a more sophisticated analysis method. Since the source code of the analysis method itself is publicly available, it can be developed further into a more generic tool or library once its usefulness has been demonstrated. This shows an interesting cycle of developing new analysis methods into generic tools that is being fueled by making analysis scripts, source-code and programming libraries publicly available.

## Conclusion

This thesis covered merely a fraction of the current developments towards better exploratory data analysis methods in the context of biological pathways. Pathway based approaches for data analysis and visualization continue to mature and become more established and standardized. This work contributed to this process by providing a platform for maintaining pathway information, an infrastructure to use it computationally and by demonstrating how these flexible tools and resources can be used to push the boundaries of pathway analysis. Our ability to measure biological phenomena in more detail and higher quantities keeps improving and the computational power to process and integrate this data continues to grow. However, interpretation largely remains a human task. By providing an interface between computer and human, the role of biological pathways in the process of exploring and understanding this data becomes increasingly important.

## References

1. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, et al. (2008) Big data: The future of biocuration. *Nature* 455: 47-50.
2. Doerr A (2008) We the curators. *Nature Methods* 5: 754-755.
3. Callaway E (2010) No rest for the bio-wikis. *Nature* 468: 359-60.
4. Ledford H (2008) Molecular biology gets wikified. *Nature News*, 23 July.
5. Waldrop M (2008) Wikiomics. *Nature* 455: 22-25.
6. Zekowitz R (2008) WikiPathways Debuts. *Science* 321: 623c-623c.
7. Huss JW, Orozco C, Goodale J, Wu C, Batalov S, et al. (2008) A gene wiki for community annotation of gene function. *PLoS Biol* 6: e175.
8. Giles J (2005) Internet encyclopaedias go head to head. *Nature* 438: 900-1.
9. Pico A (2010) WikiPathways: Community Curation of Biological Pathways. *Nature Precedings*, Available: <http://dx.doi.org/10.1038/npre.2010.5361.1>
10. MediaWiki strategic planning (n.d.). Available: [http://strategy.wikimedia.org/wiki/Wikimedia\\_users#Contributors\\_as\\_a\\_percentage\\_of\\_all\\_visitors](http://strategy.wikimedia.org/wiki/Wikimedia_users#Contributors_as_a_percentage_of_all_visitors). Accessed 26 Jan 2011.
11. Iersel M van, Kelder T, Pico A, Hanspers K, Coort S, et al. (2008) Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* 9.
12. HTML5 (n.d.). Available: <http://www.w3.org/TR/html5/>. Accessed 26 Jan 2011.
13. Transcription Factor Encyclopedia (n.d.). Available: <http://www.cisreg.ca/tfe>. Accessed 26 Jan 2011.

14. Huss JW, Lindenbaum P, Martone M, Roberts D, Pizarro A, et al. (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res* 38: D633-9.
15. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8 - a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 37: D412-6.
16. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38: W214-W220.
17. Adriaens M, Evelo C, Frank F, Kelder T, Txr24, et al. (n.d.) One Carbon Metabolism (Homo sapiens). Available: <http://www.wikipathways.org/index.php/Pathway:WP241>. Accessed 26 Jan 2011.
18. Salomonis N, Hanspers K, Zambon A, Vranizan K, Lawlor S, et al. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* 8.
19. Coort SLM, Iersel MP van, Erk M van, Kooistra T, Kleemann R, et al. (2008) Bioinformatics for the NuGO proof of principle study: analysis of gene expression in muscle of ApoE3\*Leiden mice on a high-fat diet using PathVisio. *Genes Nutr* 3: 185-91.
20. Kursawe R, Eszlinger M, Narayan D, Liu T, Bazuine M, et al. (2010) Cellularity and Adipogenic Profile of the Abdominal Subcutaneous Adipose Tissue From Obese Adolescents: Association With Insulin Resistance and Hepatic Steatosis. *Diabetes* 59: 2288-2296.
21. Rubio-Aliaga I, Roos B de, Sailer M, McLoughlin GA, Boekschoten MV, et al. (2011) Alterations in hepatic one-carbon metabolism and related pathways following a high fat dietary intervention. *Physiol Genomics*.
22. Kleemann R, Erk M van, Verschuren L, Hoek AM van den, Koek M, et al. (2010) Time-resolved and tissue-specific systems analysis of the pathogenesis of insulin resistance. *PLoS One* 5: e8817.
23. Caesar R, Manieri M, Kelder T, Boekschoten M, Evelo C, et al. (2010) A Combined Transcriptomics and Lipidomics Analysis of Subcutaneous, Epididymal and Mesenteric Adipose Tissue Reveals Marked Functional Differences. *PLoS One* 5: e11525.
24. Kramer M, Gregorowicz A, Iyer B. (2008) Wiki trust metrics based on phrasal analysis. *Proceedings of the 4th International Symposium on Wikis, 2008:1*.
25. Hoffmann R (2008) A wiki for the life sciences where authorship matters. *Nat Genet* 40: 1047-1051.
26. Rückert F, Dawelbait G, Winter C, Hartmann A, Denz A, et al. (2010) Examination of Apoptosis Signaling in Pancreatic Cancer by Computational Signal Transduction Analysis. *PLoS One* 5: e12243.
27. Jennen DGJ, Gaj S, Giesbertz PJ, Delft JHM van, Evelo CT, et al. (2010) Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes. *Drug Discov Today* 15: 851-8.
28. Ommen B, El-Sohemy A, Hesketh J, Kaput J, Fenech M, et al. (2010) The Micronutrient Genomics Project: a community-driven knowledge base for micronutrient research. *Genes Nutr* 5: 285-296-296.
29. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2010) Reactome: a database of



- reactions, pathways and biological processes. *Nucleic Acids Res* 39: D691-697.
30. Som A, Harder C, Greber B, Siatkowski M, Paudel Y, et al. (2010) The PluriNetWork: An Electronic Representation of the Network Underlying Pluripotency in Mouse, and Its Applications. *PLoS One* 5: e15165.
  31. Grinsven M van, Kelder T, Rishi R, Evelo C (n.d.) Renin - Angiotensin System (*Rattus norvegicus*). Available: <http://www.wikipathways.org/index.php/Pathway:WP376>. Accessed 26 Jan 2011.
  32. Krumeich J, Gaj S, Kutmon M (n.d.) miRNA regulation of DNA Damage Response (*Homo sapiens*). Available: <http://www.wikipathways.org/index.php/Pathway:WP1530>. Accessed 26 Jan 2011.
  33. Hermans K, Hanspers K, Iersel M van (n.d.) p53 pathway (*Rattus norvegicus*). Available: <http://www.wikipathways.org/index.php/Pathway:WP655>. Accessed 26 Jan 2011.
  34. Coort S, with Kelder T, Evelo C, Iersel M van, Arik (n.d.) Cytochrome P450 (*Homo sapiens*). Available: <http://www.wikipathways.org/index.php/Pathway:WP43>. Accessed 26 Jan 2011.
  35. Renfro D, Siegele DA, Hu JC (2010) Extending Mediawiki for community annotation. *Network Tools and Applications in Biology NETTAB-BBCC 2010 Biological Wikis*, pp. 21-24.
  36. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, et al. (2010) The NCBI BioSystems database. *Nucleic Acids Res* 38: D492-6.
  37. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, et al. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 37: W170-3.
  38. CitedIn (n.d.). Available: <http://www.citedin.org>. Accessed 15 Feb 2011.
  39. Duncan D, Prodduturi N, Zhang B (2010) WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit. *BMC Bioinformatics* 11: P10.
  40. Emig D, Salomonis N, Baumbach J, Lengauer T, Conklin BR, et al. (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res* 38 Suppl: W755-62.
  41. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* 13: 2498-2504.
  42. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orłowski J, et al. (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res* 38 Suppl: W689-94.
  43. Warsow G, Greber B, Falk SSI, Harder C, Siatkowski M, et al. (2010) ExprEssence - Revealing the essence of differential experimental data in the context of an interaction/regulation network. *BMC Syst Biol* 4: 164.
  44. Iersel M van, Pico A, Kelder T, Gao J, Ho I, et al. (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11: 5.
  45. Csárdi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal Complex Systems* 1695: 1695.
  46. Bachmair E-M, Bots ML, Mennen LI, Kelder T, Evelo C, et al. (n.d.) Effect of cis9, trans11 conjugated linoleic acid supplementation on the human platelet proteome. In preparation.

# Abbreviations



AP-1	Activator Protein 1
API	Application Programming Interface
DNA	Deoxyribonucleic acid
FTP	File Transfer Protocol
GenMAPP	Gene Map Annotator and Pathway Profiler
GEO	Gene Expression Omnibus
GLUT4	Glucose Transporter Type 4
GPML	GenMAPP Pathway Markup Language
GPS	Global Positioning System
HF	High Fat (experimental group)
HSP70	70 Kilodalton Heat Shock Proteins
HTML	Hypertext Markup Language
JNK	c-Jun Amino Terminal Kinase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LF	Low Fat (experimental group)
MAPK	Mitogen-activated protein kinase
MYC	Myc proto-oncogene protein
NCBI	National Center for Biotechnology Information
NuGO	Nutrigenomics Organization
OBO	Open Biological and Biomedical Ontologies
PI3K	Phosphoinositide 3-kinase
PPS	NuGO Proof of Principle Study
SBGN	Systems Biology Graphical Notation
SOAP	Simple Object Access Protocol
STITCH	Search Tool for Interactions of Chemicals
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TCA Cycle	Tricarboxylic Acid Cycle
URL	Uniform Resource Locator
WSDL	Web Service Definition Language
XML	Extensible Markup Language

# Samenvatting

Een belangrijk doel van biologisch onderzoek is om beter te begrijpen hoe interacties tussen verschillende moleculen het functioneren en disfunctioneren van organismen bepalen. Deze mechanismen zijn vaak zeer complex. Ter illustratie: Het humane DNA bestaat uit ongeveer drie miljard base paren die coderen voor meer dan negenentwintig duizend eiwitten. Deze eiwitten komen voor in verschillende hoeveelheden en varianten, die afhankelijk zijn van omgevingsfactoren, cellulaire locatie en onderlinge interacties.

Er is al veel bekend over deze interacties en mechanismen in de context van verschillende organismen en ziekten. Een veelgebruikte manier om deze kennis weer te geven is het concept biologisch pathway. Een typisch pathway beschrijft een set moleculaire entiteiten (e.g. genen, eiwitten of metabolieten), hun onderlinge relaties of interacties, en variaties in tijd of locatie. De verschillende stappen in een pathway worden vaak op een grafische manier weergegeven in de vorm van een diagram. In een pathway wordt onze kennis over de moleculaire biologie op een functioneel niveau beschreven en georganiseerd. Over de jaren is er een canon van pathways ontstaan die de werking van veel voorkomende biologische mechanismen samenvatten.

Dankzij de ontwikkeling van nieuwe technieken is het de laatste decennia mogelijk geworden om relatief gemakkelijk grote hoeveelheden experimentele data te genereren. De zogenoemde 'high-throughput' of 'omics' technieken stellen biologen in staat om metingen te doen op duizenden entiteiten binnen een enkel experiment. Hierdoor kan bijvoorbeeld de expressie van alle genen of de aanwezigheid van vele eiwitten tegelijk gemeten worden. Door deze metingen uit te voeren op verschillende tijdstippen, of door het vergelijken van verschillende cel typen, kan bepaald worden welke veranderingen er optreden op moleculair niveau. Door het grote bereik maar relatief hoge ruis in dit type experimenten zijn deze uitermate geschikt voor exploratieve analyse. Dit soort analyse heeft als doel om gerichte vragen en hypothesen te genereren die vervolgens getest kunnen worden in meer gerichte experimenten.

De hoeveelheden experimentele data zijn zo groot dat ze niet meer efficiënt met de hand geanalyseerd kunnen worden. Daarom worden er binnen de bioinformatica algoritmen en computer software ontworpen en toegepast om het interpreteren van deze data mogelijk te maken. Een belangrijke stap in een analyse is het combineren van de data met de kennis die we al hebben, om zo tot nieuwe biologische inzichten te komen. Dit proefschrift focust op dat deel van de analyse, waarbij pathways worden gebruikt om bestaande kennis te integreren met bioinformatica toepassingen.

Het doel van dit proefschrift is om exploratieve data analyse te verbeteren door het gebruiken van biologische pathways. Hoofdstuk 2 bevat een korte samenvatting van bestaande pathway analyse technieken. Twee veelgebruikte toepassingen zijn het visualiseren van data op pathway diagrammen en het uitvoeren van over-representatie analyse, waarbij wordt gezocht naar pathways die waarschijnlijk betrokken zijn in het bestudeerde biologische mechanisme. Verder zijn een aantal kenmerken en vereisten van exploratieve data analyse beschreven, die verder in het proefschrift worden gebruikt om analyse methoden te verbeteren en toe te passen.

Een van de vereisten om pathways te kunnen gebruiken in data analyse is dat onze huidige kennis over moleculaire biologie wordt gerepresenteerd als pathways. Het verzamelen van biologische pathways is een immense taak, omdat deze niet automatisch kunnen worden gegenereerd,

maar handmatig moeten worden gecompileerd uit kennis verspreid over wetenschappelijke literatuur en verschillende domein experts. De pathway database WikiPathways die is beschreven in hoofdstuk 3 is een unieke benadering om deze grote hoeveelheden biologische kennis te verzamelen in een vorm die gebruikt kan worden in bioinformatica toepassingen. WikiPathways is een wiki waarbij, zoals op Wikipedia, alle informatie direct te bewerken is door gebruikers. Dit geeft de wetenschappelijke gemeenschap zelf controle om de informatie te verbeteren in plaats van dit aan een kleine groep curators over te laten. Hiermee biedt WikiPathways op termijn een beter schaalbare oplossing om de exponentiële groei van nieuwe kennis aan te kunnen. Een vereiste hiervoor is wel dat het aantal actieve gebruikers groot genoeg is om een goede kwaliteit en actuele inhoud te verkrijgen. Twee voorbeelden van punten die dit moeten bevorderen zijn een directe bruikbaarheid van de pathway informatie in onderzoek en een laagdrempelige interface om pathway informatie in te voeren in de vorm van een intuïtief werkend tekenprogramma. Gezien het kort bestaan van WikiPathways is het moeilijk te zeggen of een gebruikersgroep van voldoende grootte bereikt kan worden, echter zowel de inhoud als het aantal actieve gebruikers laten een positieve groei zien.

In hoofdstuk 4 wordt een webservice beschreven die het mogelijk maakt effectiever gebruik te maken van pathway informatie in exploratieve data analyse. Deze webservice geeft computerprogramma's direct toegang tot de informatie op WikiPathways. Dit heeft diverse toepassingen. Ten eerste wordt het hierdoor mogelijk om pathways te integreren in verschillende bioinformatica software. Zo is het bijvoorbeeld mogelijk om vanuit het netwerk visualisatie programma Cytoscape direct de nieuwste versie van een pathway op WikiPathways te in laden. Ook geeft het de mogelijkheid om bijvoorbeeld op de NCBI website informatie over een specifiek gen uit te breiden met de verschillende pathways waarin het een rol speelt. Ten tweede maakt de webservice het ook makkelijker de pathways direct te integreren in een analyse script of workflow en eventueel te combineren met andere typen informatie of publieke datasets die ook via webservices beschikbaar zijn. Als laatste maakt de webservice het ook mogelijk om curatie taken op WikiPathways te assisteren. Zo draaien er bijvoorbeeld voortdurend computerprogramma's op de achtergrond die bijdragen van gebruikers controleren op veel voorkomende fouten en deze terugrapporteren zodat ze verbeterd kunnen worden.

Hoofdstuk 5 geeft een voorbeeld van een exploratieve data analyse die is opgebouwd rondom pathway informatie. Het doel van deze analyse is om inzicht te krijgen in mogelijke interacties tussen pathways in de context van insuline resistentie. Zoals veel complexe ziekten spelen bij insuline resistentie waarschijnlijk verschillende pathways een rol en is het dus nuttig om te kijken naar het samenspel tussen pathways. Veel bestaande pathway gebaseerde methoden gebruiken pathways als geïsoleerde entiteiten, dus biedt deze analyse een nieuw perspectief op de data. De analyse gaat uit van een gen expressie dataset gemeten in lever weefsel van gezonde en insuline resistente muizen, voorafgaand aan en op verschillende tijdstippen na het uitvoeren van een glucose tolerantie test. Op basis van een netwerk met interacties tussen genen en hun eiwit producten wordt tussen elk paar pathways een set kortste paden gezocht waarbij de lengte van een pad de som is van de gewichten van de verbindende eiwit interacties. Interacties tussen eiwitten waarvan de coderende genen differentieel tot expressie komen krijgen een lager gewicht toegewezen, waardoor paden met veel veranderde genen een kortere lengte krijgen. Op basis van de aanname dat hoe meer en kortere paden tussen twee pathways gevonden worden, hoe waarschijnlijker het is dat deze pathways een interactie

aangaan, kan dan een netwerk van pathway interacties worden afgeleid. Op deze manier zijn diverse mogelijke pathway interacties gevonden voor verschillende vergelijkingen van de gen expressie tussen gezonde en insuline resistente groepen en tussen de verschillende tijdstippen. Bovendien is er ingezoomd op de onderliggende eiwit interacties om mogelijke mechanismen en eiwitten die betrokken zijn bij communicatie tussen pathways te bestuderen. Dit heeft tot nieuwe inzichten en hypothesen geleid over het mogelijk samenspel tussen verschillende pathways in insuline resistentie en hoe deze verandert als gevolg van een glucose stimulans.

Tezamen heeft het onderzoek zoals beschreven in dit proefschrift op een aantal punten bijgedragen aan de verbetering van exploratieve data analyse met behulp van biologische pathways. Er is een platform gecreëerd voor het verzamelen en verbeteren van pathway informatie die geschikt is voor gebruik in bioinformatica toepassingen. Er is een infrastructuur gebouwd om deze informatie programmatisch toegankelijk te maken en te integreren in analyses. Tot slot is gedemonstreerd hoe de ontwikkelde hulpmiddelen rondom pathway informatie gebruikt kunnen worden om originele analyse methoden te ontwikkelen die nieuwe perspectieven bieden op een bestaande dataset.

# Acknowledgements

Ongeveer zes jaar geleden kwam ik tijdens mijn studie voor het eerst in aanraking met de BiGCaT Bioinformatica groep. Wat me meteen aantrok was het praktisch gerichte onderzoek, dichtbij de toepassingen in de biologie (iets wat in Eindhoven soms nogal ver te zoeken was). Na mijn afstuderen was ik dan ook blij dat ik meteen als promovendus bij BiGCaT aan de slag kon. Chris, bedankt voor je begeleiding en de vrijheid die je gaf om mijn eigen weg te zoeken en vinden binnen het bioinformatica onderzoek. Martijn, bedankt voor de goede samenwerking, jouw visie op software ontwikkeling heeft me veel geleerd. Andra, Lars, Susan en Tina, bedankt voor de nuttige samenwerkingen waar onze projecten overlaptten. Jos, bedankt voor je hulp met het nodige papierwerk. Jahn, thanks for introducing me to the sweet sounds of the ukulele. Michiel, bedankt voor je uitstekende muziek en film tips. Everyone at BiGCaT, thanks for your collegiality, chats over coffee and fun activities after work, which altogether made my years at BiGCaT a great time to remember!

Alex, Kristina and Bruce, thanks for our long distance but nevertheless very close collaboration, working with you was a great pleasure and I hope it will not end here. I would like to thank everyone in the Conklin lab for welcoming me in your group and showing me a great time in San Francisco. Rudy, thanks for your friendship and hospitality during my stay in San Francisco, I'll never forget our road trip to Yosemite and the critical mass bike ride.

Robert, Marjan, Teake en iedereen van het NuGO PPS team, bedankt voor het beschikbaar maken van jullie dataset en het waardevolle commentaar op de biologische aspecten van de analyse.

Verder wil ik mijn familie en schoonfamilie bedanken voor het tonen van interesse in mijn werk, ik kan me voorstellen dat het voor jullie vaak een ver van je bed show was en niet de meest interessante gespreksstof.

Ise, bedankt voor je schaterlach, je ondeugende blik, je gebrabbel, knuffels en kusjes, allemaal heel effectieve en welkome middelen om even niet aan dit proefschrift te hebben hoeven denken het afgelopen jaar. Karin, bedankt voor je geweldige steun, hulp en liefde en het inzicht dat het leven meer biedt dan werk alleen. Zonder jou was dit proefschrift er nooit gekomen.

# Curriculum Vitae



Thomas Adriaan Johan Kelder was born in 's-Hertogenbosch, on February 4, 1983. He spent his early years playing with Lego, collecting rocks and dinosaur figurines, and trying to stay unnoticed in order to avoid having to participate in outdoor games with the other kids. Thomas finished the Atheneum at the Willem van Oranje college in Waalwijk, after which he went to Eindhoven to study Biomedical Engineering. There he developed interest in bioinformatics and molecular biology. During the first internship of the MSc. program, Thomas gained invaluable experience in wetlab work at MAASTRO Lab in Maastricht, but also learned that this was not his calling. In 2005, Thomas visited Italy for three months to do bioinformatics research in the Department of Pharmacology at the University of Florence. There he learned about different aspects of analyzing microarray data, from preprocessing and quality control to biological interpretation. For the final project in the MSc program, Thomas went to the BiGCaT Bioinformatics group at Maastricht University to work on development of PathVisio, a tool for visualization of experimental data on biological pathways. In 2007 he started on his PhD project in the same group and in close collaboration with the Conklin group at Gladstone Institutes in San Francisco, resulting in this thesis. At the time of writing this, Thomas is still looking for opportunities to continue his career in bioinformatics and is experiencing how strange it feels to write about himself in the third person.

# Publications

1. **Kelder T**, Eijssen L, Kleemann R, van Erk M, Kooistra T, Evelo C. (2011) Pathway interactions in insulin resistant mouse liver. *Under review*.
2. **Kelder T**, Conklin BR, Evelo CT, Pico AR. (2010) Finding the Right Questions: Exploratory Pathway Analysis to Enhance Biological Discovery in Large Datasets. *PLoS Biology* 8(8):e1000472.
3. Caesar R, Manieri M, **Kelder T**, et al. (2010) A Combined Transcriptomics and Lipidomics Analysis of Subcutaneous, Epididymal and Mesenteric Adipose Tissue Reveals Marked Functional Differences. *PLoS One* 5(7):e11525.
4. Kleemann R, van Erk M, Verschuren L, van den Hoek AM, Koek M, Wielinga PY, Jie A, Pellis L, Bobeldijk-Pastorova I, **Kelder T**, et al. (2010) Time-resolved and tissue-specific systems analysis of the pathogenesis of insulin resistance. *PLoS One*. 5(1):e8817.
5. Iersel M van, Pico A, **Kelder T**, et al. (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11(1):5.
6. **Kelder T**<sup>Ⓢ</sup>, Pico AR<sup>Ⓢ</sup>, Hanspers K, et al. (2009) Mining biological pathways using WikiPathways web services. *PLoS One* 4(7):e6447.
7. Baccini M, Bachmaier EM, Biggeri A, Boekschoten MV, Bouwman FG, Brennan L, Caesar R, Cinti S, Coort SL, Crosley K, Daniel H, Drevon CA, Duthie S, Eijssen L, Elliott RM, van Erk M, Evelo C, Gibney M, Heim C, Horgan GW, Johnson IT, **Kelder T**, et al. (2008) The NuGO proof of principle study package: a collaborative research effort of the European Nutrigenomics Organisation. *Genes & Nutrition* 3(3-4):147-151.
8. Waagmeester A, **Kelder T**, Evelo C. (2008) The role of bioinformatics in pathway curation. *Genes & Nutrition* 3(3-4):139-142.
9. Hu JC, Aramayo R, Bolser D, Conway T, Elsik CG, Gribskov M, **Kelder T**, et al. (2008) The emerging world of wikis. *Science* 320(5881):1289-90.
10. Iersel M van, **Kelder T**, Pico A, et al. (2008) Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* 9(1).
11. Pico AR<sup>Ⓢ</sup>, **Kelder T**<sup>Ⓢ</sup>, Iersel MP van, et al. (2008) WikiPathways: pathway editing for the people. *PLoS Biology* 6(7):e184.
12. Cavalieri D, Castagnini C, Toti S, Maciag K, **Kelder T**, et al. (2007) EuGene Analyzer a tool for integrating gene expression data with pathway databases. *Bioinformatics* 23(19):2631-2632.

Ⓢ These authors contributed equally to this work.

