

Doctoral thesis

Doctoral theses at NTNU, 2022:212

Kjersti Rise

# Development of an improved pathway analysis - The FunHoP story

**NTNU**  
Norwegian University of Science and  
Technology  
Thesis for the Degree of  
Philosophiae Doctor  
Faculty of Medicine and Health Sciences  
Department of Clinical and Molecular Medicine



Norwegian University of  
Science and Technology



Kjersti Rise

# **Development of an improved pathway analysis - The FunHoP story**

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2022

Norwegian University of Science and Technology  
Faculty of Medicine and Health Sciences  
Department of Clinical and Molecular Medicine



**NTNU**  
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences  
Department of Clinical and Molecular Medicine

© Kjersti Rise

ISBN 978-82-326-5652-3 (printed ver.)  
ISBN 978-82-326-5524-3 (electronic ver.)  
ISSN 1503-8181 (printed ver.)  
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:212

Printed by NTNU Grafisk senter



---

"We're all stories in the end.  
Just make it a good one, eh?"

---

The 11th Doctor, *Doctor Who*, E13S5

---

---

---

---

# Summary

The work in this thesis revolves around the metabolism of prostate cancer, mainly by using and improving biological pathway analysis. A large part of the thesis is about the development of the method FunHoP, and how this method can be used in different ways and provide new biological insight. FunHoP is a Python based method that uses metabolic pathways from KEGG, along with read counts from RNA-sequencing. The basis for the thesis is three scientific studies.

The first study is about metabolism in samples from prostate cancer grouped by their content of reactive stroma. 108 samples were histopathologically evaluated and graded by their content of reactive stroma. Out of these, metabolites were measured in 85 samples and gene expression in 78 samples. Multivariate metabolomics and transcriptomics were used to compare groups with low stroma content ( $\leq 15\%$ ) to groups with high reactive stroma ( $\geq 16\%$ ). We found that groups with high content of reactive stroma had upregulated both genes and metabolites related to functions in the immune system and extracellular matrix. This study was a good introduction to metabolism in prostate cancer, and demonstrated how different types of omics can be used together to give new understanding of how the biology works.

In the second study, development of FunHoP was the main topic. Visualisation is a great tool in analysis of big data, and a well-known method is to use data to colour nodes in a network to show differential expression, using tools such as Cytoscape. A problem with the combination of KEGG, KEGGScape (which is used to load KEGG files into Cytoscape), and Cytoscape is that only the first gene/protein in each node is shown. This makes all reactions look as if there is only one enzyme able to catalyze the reaction. In many cases, this representation is not biologically correct. FunHoP expands the nodes to include all genes, shows the user how the genes are differentially expressed as well as their read counts, before they are all joined together and differential expression can be calculated on node level. This study shows how FunHoP was developed, and also contains two case studies where we show how FunHoP provides results that both fits better into the known biology, and also gives a better visual understanding to the viewer.

In the final study, FunHoP was used in an alternative way to bring out a new level of biological insight. By including cellular localisation data it became possible to differentiate between mitochondrial and non-mitochondrial biological paths, along with those that are a mixture, and see how differentially expressed genes possibly changed between the two location groups. Here we used gene expression from normal and cancerous cell lines, along with a consensus of localisation from both experiments and predictions. This study shows how FunHoP could be used in alternative ways, that mitochondrial pathways are generally upregulated in prostate cancer, and that use of localisation data can give a wider biological insight.

---

---

# Sammendrag

Arbeidet som presenteres i denne avhandlingen omhandler metabolisme i prostatakraft, hovedsakelig i form av bruk og forbedring av analyse av biologiske spor. En stor del av oppgaven handler om utvikling av metoden FunHoP, og hvordan denne kan brukes på forskjellige måter og gi ny biologisk innsikt. FunHoP er en Python-basert metode som bruker metabolske spor fra KEGG, sammen med transkripsjonsuttrykk fra RNA-sekvensering. Basis for avhandlingen er tre vitenskapelige studier.

Den første studien handler om metabolisme i prostatakraftprøver gruppert etter innhold av reaktivt stroma. 108 prøver ble histopatologisk evaluert og gradert etter innhold av reaktivt stroma. Av disse ble det målt metabolitter i 85 prøver mens det ble målt genuttrykk i 78 prøver. Multivariat metabolomikk og transkriptomikk ble brukt for å sammenligne grupper med lav andel av stroma ( $\leq 15\%$ ) mot grupper med høy andel reaktivt stroma ( $\geq 16\%$ ). Det ble vist at i grupper med høy andel reaktivt stroma var både gener og metabolitter med tilknytning til funksjoner i immunforsvaret og ekstracellulær matrise oppregulert. Denne studien gav en god introduksjon til metabolisme i prostatakraft, og demonstrerte også hvordan forskjellige typer omics kan brukes sammen for å gi økt forståelse av hvordan biologien henger sammen.

I den andre studien sto utvikling og demonstrasjon av FunHoP i fokus. Visualisering er et godt hjelpemiddel i analyse av store mengder data, og en mye brukt metode er å bruke data til å f.eks farge noder for å vise differensielt uttrykte gener, ved hjelp av verktøy som Cytoscape. En ulempe med kombinasjonen KEGG, KEGGScape (som laster inn KEGG-filer i Cytoscape), og Cytoscape er at bare det første genet/proteinet i en node vises. Dette gjør at alle reaksjoner ser ut til å bare kunne katalyseres av ett enzym. Dette stemmer i mange tilfeller ikke overens med biologien. FunHoP utvider noder til å inkludere alle gener i en node, viser brukeren hvordan genene er differensielt uttrykt og hvilken read count de har, før alle genene til slutt slås sammen og differensielt uttrykk på node-nivå kan beregnes. Denne studien viser hvordan FunHoP ble utviklet, og har også to eksempler hvor vi viser hvordan FunHoP gir resultater som både stemmer bedre overens med kjent biologi og gir en bedre visuell forståelse av biologien.

I den siste studien ble FunHoP brukt på en alternativ måte for å få fram et nytt nivå av biologisk innsikt. Ved å inkludere lokasjonsdata ble det mulig å differensiere mellom mitokondrielle og ikke-mitokondrielle biologiske spor, samt identifisere de som var en blanding, og se på hvordan differensielt genuttrykk eventuelt endret seg i forskjellige lokasjoner. Her ble genuttryksdata fra normal- og kreftceller brukt, sammen med en konsensus av lokasjonsdata fra både eksperimenter og prediksjon. Denne studien viste hvordan FunHoP kunne brukes på alternative måter, at mitokondrielle spor er generelt oppregulert i prostatakraft, og at bruk av lokasjonsdata kan gi mer biologisk innsikt.

---

---

# Acknowledgements

Finding the words for this acknowledgement was surprisingly difficult. How is one supposed to sum up everyone who helped along a (more than) six year long journey? A journey consisting of learning, thinking, drinking beer, travels, gaining and losing family and friends, a pandemic, and now war?

For starters I would like to thank my supervisors Morten Beck Rye and Finn Drabløs for all their thoughts and inputs over the years. To all my former co-workers and co-authors, thank you for being part of my journey, and letting me be a part of yours.

Thank you to my friends and family for being part of this in so many different ways. From babysitting my son while I worked, listening to me complain when times were rough, sharing experience and knowledge, bragging about me to the neighbours (thanks Grandpa, I miss you!), to my sister and all her editing and proof-reading, thank you.

A great thank you to my amazing little man Mimir. Thank you for showing me the important parts of life, and reminding me of the magic in the world.

And most of all. Einar Johan, the guy who made FunHoP possible. The guy who fixed all my computer issues and all my code issues, put a smile on my face every day, has the best hugs, and the most amazing brain. Thank you for sticking up with all of this, and for being who you are. I could not have done any of this without you.

And finally, thanks to Pengvin for all the adventures and all the fish.

Kjersti Rise  
Trondheim, March 2022

---



# Table of Contents

<b>Summary</b>	<b>i</b>
<b>Sammendrag</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Table of Contents</b>	<b>viii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>Articles</b>	<b>xii</b>
<b>Abbreviations</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Aims of the thesis</b>	<b>3</b>
<b>3 Summary of papers</b>	<b>5</b>
3.1 Paper 1 . . . . .	5
3.2 Paper 2 . . . . .	5
3.3 Paper 3 . . . . .	6
<b>4 Background</b>	<b>7</b>
4.1 Networks . . . . .	7
4.1.1 KEGG . . . . .	9
4.1.2 Cytoscape . . . . .	11
4.2 The cell . . . . .	11
4.3 Metabolism . . . . .	13
4.3.1 Carbohydrates . . . . .	15
4.3.2 Metabolic networks . . . . .	21
4.3.3 Gene expression . . . . .	22
4.4 The prostate . . . . .	24
4.4.1 The normal prostate . . . . .	24
4.4.2 Prostate cancer and altered metabolism . . . . .	25

---

<b>5</b>	<b>Methods and development</b>	<b>27</b>
5.1	Techniques . . . . .	27
5.1.1	Identification of metabolites . . . . .	27
5.1.2	Measuring gene expression . . . . .	27
5.1.3	Using Cytoscape . . . . .	30
5.2	Developing FunHoP . . . . .	30
5.2.1	Changing the XML files . . . . .	31
5.2.2	Using the read counts . . . . .	35
<b>6</b>	<b>Summary of the results</b>	<b>37</b>
<b>7</b>	<b>Discussion</b>	<b>39</b>
7.1	Common challenges for all three papers . . . . .	39
7.2	Paper 1 . . . . .	41
7.3	Paper 2 . . . . .	42
7.3.1	Colours in Cytoscape . . . . .	42
7.3.2	What the pathways are not telling us . . . . .	43
7.4	Paper 3 . . . . .	43
7.4.1	Hard coding the XML files . . . . .	44
7.5	FunHoP . . . . .	46
7.5.1	Removing the hard coding . . . . .	46
7.5.2	Including more databases . . . . .	46
7.5.3	Coordinates . . . . .	47
7.5.4	Example datasets . . . . .	47
7.5.5	Alternative use of FunHoP . . . . .	47
7.5.6	How FunHoP could have been created differently . . . . .	48
<b>8</b>	<b>Future work</b>	<b>49</b>
<b>9</b>	<b>Conclusion</b>	<b>51</b>
	<b>Bibliography</b>	<b>52</b>
<b>A</b>	<b>Appendix: The papers</b>	

# List of Figures

4.1	Matrix representation of networks . . . . .	8
4.2	Network features: Hub and bottleneck . . . . .	9
4.3	KEGG network website representation . . . . .	10
4.4	The human cell . . . . .	12
4.5	Overview of the human metabolism . . . . .	14
4.6	Glycolysis pathway . . . . .	16
4.7	The TCA cycle . . . . .	18
4.8	Oxidative Phosphorylation . . . . .	19
4.9	How nutrients become energy . . . . .	20
4.10	mRNA processing in the nucleus . . . . .	22
4.11	The prostate gland . . . . .	24
5.1	The process of microarray hybridization . . . . .	28
5.2	The process of RNA-sequencing . . . . .	29
5.3	P-value colour scale used in Cytoscape . . . . .	30
5.4	Read count colour scale used in Cytoscape . . . . .	30
5.5	Common (and mistaken) use of KEGG . . . . .	31
5.6	FunHoP: Original KEGG XML node . . . . .	32
5.7	FunHoP: Expanding the graphics name . . . . .	32
5.8	FunHoP: Keeping only the first gene . . . . .	33
5.9	FunHoP: Creating new entries . . . . .	33
5.10	FunHoP: Connecting genes into a new node . . . . .	34
5.11	FunHoP: Collapsing nodes . . . . .	34

---

# List of Tables

7.1 Comparison of differential expression . . . . .	44
---	----

---

# Articles

1. Andersen, M. K., Rise, K., Giskeødegård, G. F., Richardsen, E., Bertilsson, H., Størkersen, Ø., Bathen, T. F., Rye, M., and Tessem, M.-B. (2018). Integrative metabolic and transcriptomic profiling of prostate cancer tissue containing reactive stroma. *Scientific reports*, 8(1):1–11.
2. Rise, K., Tessem, M.-B., Drabløs, F., and Rye, M. B. (2021). FunHoP - enhanced visualization and analysis of functionally homologous proteins in complex metabolic networks. *Genomics, Proteomics & Bioinformatics, in press*
3. Rise, K., Tessem, M.-B., Drabløs, F., Rye, M. B. FunHoP analysis reveals upregulation of mitochondrial genes in prostate cancer. *Submitted*.

---



---

# Abbreviations

Acetyl-CoA	=	Acetyl coenzyme A
ADP	=	Adenosine diphosphate
ATP	=	Adenosine triphosphate
BPH	=	Benign prostatic hyperplasia
BUSCA	=	Bologna Unified Subcellular Component Annotator
cDNA	=	Complementary DNA
COVID-19	=	Coronavirus disease 2019
CZ	=	Central zone of the prostate
DAVID	=	Database for Annotation, Visualisation and Integrated Discovery
DNA	=	Deoxyribonucleic Acid
ECM	=	Extracellular matrix
ER	=	Endoplasmic reticulum
FADH <sub>2</sub>	=	Flavin adenine dinucleotide
FunHoP	=	Functionally Homologous Proteins
GDP	=	Guanosine diphosphate
GPC	=	Glycerophosphocholine
GTP	=	Guanosine triphosphate
GO	=	Gene Ontologies
HR-MAS	=	High-resolution Magic Angle Spinning
KEGG	=	Kyoto Encyclopaedia for Genes and Genomes
KGML	=	KEGG Markup Language
MALDI	=	Matrix-Assisted Laser Desorption-Ionization
mRNA	=	Messenger RNA
miRNA	=	MicroRNA
NADH	=	Nicotinamide adenine dinucleotide
PCa	=	Prostate Cancer
PZ	=	Peripheral zone of the prostate
RNA	=	Ribonucleic Acid
RNA-Seq	=	RNA-Sequencing
RSG	=	Reactive Stroma Grade
siRNA	=	Small interfering RNA
TCA cycle	=	Tricarboxylic Acid Cycle
TME	=	Tumor microenvironment
TZ	=	Transitional zone of the prostate
UTR	=	Untranslated regions
XML	=	Extensible Markup Language

---

# Introduction

The overall aim of this thesis is to utilize and improve biological pathway analysis. The work revolves mainly around the development and use of Functionally Homologous Proteins, FunHoP, a Python-based method that works on metabolic pathways from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) in Cytoscape. The thesis examines gene expression and differentially expressed genes in prostate cancer (PCa), and explores how the addition of new information can make pathway analysis more biologically correct.

This thesis is based on three papers. In the first paper, ‘Integrative Metabolic and Transcriptomic Profiling of PCa Tissue Containing Reactive Stroma’ (Andersen et al., 2018), we examined metabolism and gene expression in PCa. This collaborative work was based on the different gradings of reactive stroma; my contribution was performing the differential expression analysis comparing different groups of reactive stroma and the enrichment analysis of gene ontologies (GO).

The second paper presents the main work of this thesis, ‘FunHoP – Enhanced Visualization and Analysis of Functionally Homologous Proteins in Complex Metabolic Networks’ (Rise et al., 2021). This section examines how FunHoP was made and how it expands the pathways and uses read counts from RNA sequencing (RNA-Seq) in a new way. We present two case studies to show how FunHoP improves biological pathway analysis, using PCa as an example.

The final paper, ‘The Upregulation of Mitochondrially Located Genes in Prostate Cancer: A FunHoP Discovery’, further expanded the usage of FunHoP. This study discusses how FunHoP can be used in combination with localisation data to show how mitochondrial genes, and thus mitochondrial pathways, are upregulated in PCa. The addition of localisation data shows how our interpretations of the pathways can be improved and how FunHoP can provide more biological information.

To introduce this thesis, I give a brief overview of the aims of the studies and a summary of each paper. This is followed by a longer chapter explaining the background of the main topics discussed, to give the reader the explanation of networks, metabolism, RNA-sequencing, and the prostate, needed for reading and understanding the thesis. Following is a section on the methods used for measuring metabolites and

gene expression, as well as the development of FunHoP in more detail. This is followed by a general discussion, notes on each paper, suggestions for possible future research, and a conclusion. Finally, the three papers will follow as appendices.

## Aims of the thesis

This research aims to improve pathway analysis by developing a new tool, FunHoP, and to show how the three papers provide three levels of integration of biological data, which makes analysis more biologically relevant. At the primary level, FunHoP improves pathway analysis by expanding the display of a pathway to show all genes of each node, rather than just a single gene. The second level is added by showing the expression level of each gene, making it possible to identify dominant genes. The expression level is based on read counts from RNA-Seq. The read counts are also used in combination for all genes within a node and to perform differential expression at node level. The final level of pathway analysis improvement involves adding protein localisation and making it possible to divide pathways into subsets, for example mitochondrial vs non-mitochondrial versions. These levels can be used partially, separately, or together to give a deeper understanding of biological pathways.

The main focus of the first study is an analysis of reactive stroma in PCa tissue. We combined differential gene expression analysis and metabolite analysis with histopathology to compare differences between the gradients of reactive stroma and find related pathways. Thus, this paper constitutes an introduction to genes, metabolites, and pathways in PCa.

The second study examined the subject of differentially expressed genes and pathways in greater detail. It was separated into two parts: the first section discussed the development of FunHoP and finding solutions that would create an overall improvement of pathway analysis; the second section used FunHoP in two case studies to show how it works and to give new biological insights. This paper shows how FunHoP and read counts can be used in combination.

The third study investigated pathway analysis by studying how protein location can alter our understanding of the pathways in even greater detail. By adding localisation data to FunHoP, the pathways could be divided into one mitochondrial and one non-mitochondrial version and more could be learnt about protein localisation and gene regulation within the different parts of the cell.



## Summary of papers

### 3.1 Paper 1

The main topic of this paper is metabolism in different groups of reactive stroma. Several analyses compared the differential gene expression between groups of tissue type and identified the metabolic pathways of particular interest in those groups. Reactive stroma were histopathologically evaluated in 108 fresh frozen PCa tissue samples from 43 patients and divided into four groups, ranging from reactive stroma grade (RSG) 0 to 3. Metabolites were measured in 85 samples using HR-MAS MRS. In 78 samples, the transcriptome was analysed using RNA microarray. Multivariate metabolomics and transcriptomics were used to compare low reactive stroma content (<15%) to high reactive stroma content (>16%). Both metabolites and genes linked to immune functions and extracellular matrix remodelling were significantly upregulated in samples with high RSG. This study showed how different omics can complement each other in the search for a more complete biological picture.

### 3.2 Paper 2

Pathway analysis is an essential tool when analysing large amounts of data. A standard tool for visualisation and analysis is Cytoscape, which can be used in combination with pathways from the KEGG database. Each of the many pathways in KEGG can be downloaded as XML files written in KEGG Markup Language (KGML). By using the KGML reader KEGGScape, these pathways can be opened up and viewed in Cytoscape. However, although multiple genes can be responsible for the protein that catalyses the reaction, KEGGScape shows only one. Only showing the first gene can lead to incorrect interpretations of the pathway. By contrast, FunHoP shows all genes, giving a broader picture that can be interpreted more accurately. To determine how the pathway is regulated, FunHoP collapses all genes in a node into one measurement using RNA-Seq read counts. Assuming that activity for an enzymatic reaction depends mainly on the gene with the highest number of reads, as well as weighting reads according to gene length and ratio, a new expression value is calculated for the node as a whole. Differential gene expression is then applied to the whole network. Using PCa as a model, we integrated RNA-Seq data from two patient cohorts and metabolism data from the literature. We could then give plausible explanations as to how the

metabolic paths of histidine metabolism and a minor part of glycerophosphocholine (GPC) were regulated.

### **3.3 Paper 3**

Mitochondrial activity in cancer cells has been central to cancer research since Otto Warburg first published his thesis on the topic in 1956. In this study, we expanded the usage of our method FunHoP. We used RNA-Seq data from cancerous and normal prostate cell lines. By adding localisation data based on experimental data and computational predictions, we could differentiate between mitochondrial and non-mitochondrial processes in PCa. Our results showed that mitochondrial pathways are generally upregulated in PCa and that splitting metabolic pathways into mitochondrial and non-mitochondrial counterparts using FunHoP enables more accurate interpretation of the metabolic make-up of PCa cells.



# Background

Pathway analysis and gene expression are the common focus of all three studies described in this thesis, in addition to the development and usage of FunHoP. In this chapter, I briefly explain five of the related topics: networks, the cell, metabolism, gene expression, and the prostate and prostate cancer.

The first section discusses networks and pathways – how they are created and used, how Cytoscape works, and the possibilities it contains. The second section looks briefly into the cell, before I discuss the basics of the metabolism of the carbohydrates studied in **Paper 3**. The fourth section looks into gene expression, while the fifth and final section describes the prostate, how a normal prostate differs from other glands, and several ways in which metabolism is altered by prostate cancer.

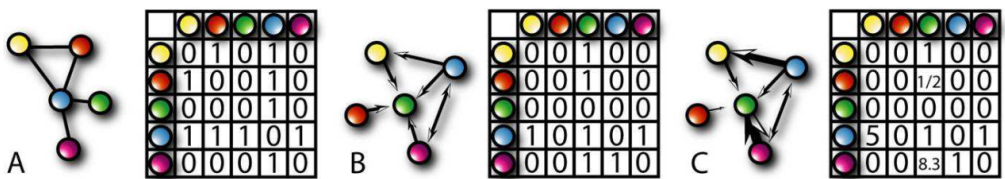
## 4.1 Networks

The main topic of this thesis is improving the analysis of biological pathways. Biological pathways are the defined steps between two given compounds, with various genes and intermediate compounds between them. These pathways can be represented as networks, which are a collection of nodes connected by links. Network representation can be used in various ways in biology, such as showing nodes as muscles with skeleton links or showing how COVID-19 spreads (the ‘links’) between humans (the ‘nodes’) in a population. These representations can contain various levels of information, depending on the available data.

Mathematically, a network can be represented by an adjacency matrix with nodes on each axis and links in the intersections, where a 1 indicates a link and a 0 indicates no link. In the simplest representations, links are simply undirected and unweighed connecting nodes, as shown in **figure 4.1A**. In this representation, it does not matter in which direction the links between the nodes go, as they are all equal, carry no information, and are either there or not. The muscular-skeletal network is an example of this type of network, as the skeleton does not have any particular direction — the bones are either connected to the muscles or not. The adjacency matrix is identical on both sides of the diagonal.

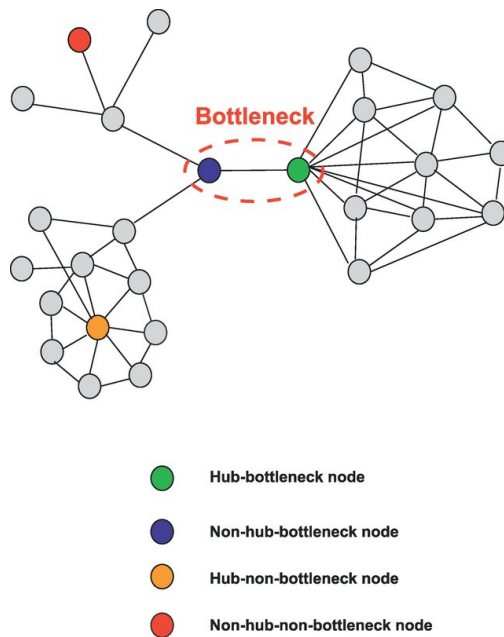
Adding direction to the links can be useful, for example to describe the spread of COVID-19 from human to human. Who carried the virus and how they were infected becomes relevant. The direction is important and visible in the adjacency matrix, which is no longer equal on both sides of the diagonal (**figure 4.1B**). Tracing the spread of COVID-19 would be impossible without knowing the direction of these links.

If more information is available, the links can also be weighted, as in **figure 4.1C**. In a network for predicting where COVID-19 would appear next, the links could be weighted based on the length of the connection between two nodes, for example contact time or physical distance, or to represent various mutations of the virus. Longer exposure to a more infectious mutation would increase the probability of getting COVID-19 from an infected person, and the weighted network would indicate who should be prioritized in getting tested for the disease.



**Figure 4.1:** A network can be represented as a matrix. In A, the nodes are connected with undirected and unweighed links, making the matrix identical across the diagonal. This pattern is lost in B and C, where all the links are directed, and for C also weighted (Almaas, 2013).

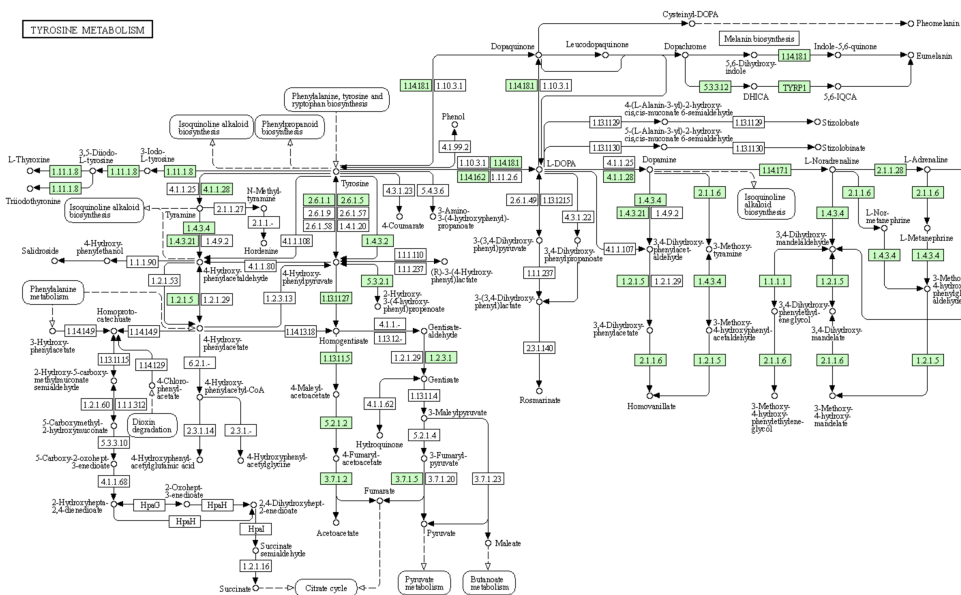
Networks can also include important topological features, such as patterns, node importance, and node interactions. One of these features is the number of links connected to a given node, for instance, if a node becomes a so-called hub (**figure 4.2**). A hub is a central node in the network with multiple links connected to it, and removing the node would break up the network. In the COVID-19 example, a hub could be an infected person who talked to almost everyone at a party. Another feature of networks is bottlenecks, or nodes that determine the rate of flow in the network (**figure 4.2**). Nodes can also exhibit other features, depending on the network in question. Knowledge about the nodes can be represented by using various sizes, colours, or shapes. Similar to an undirected and unweighed link, nodes about which no specific knowledge is available can be indicated by a default shape, colour, or size.



**Figure 4.2:** Patterns and features of a network. The connecting link between the green and the blue node becomes a bottleneck, while the orange with multiple connected nodes is a hub (Yu et al., 2007), CC BY 4.0.

#### 4.1.1 KEGG

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa, 2019; Kanehisa et al., 2021) is one of the main resources used in this thesis. The KEGG database includes data for a wide range of organisms and contains information on pathways, genomes, and compounds (such as metabolites), as well as disease-specific information. **Figure 4.3** shows KEGG's representation of histidine metabolism, studied and discussed further in **Paper 2**.



**Figure 4.3:** KEGG website representation of tyrosine metabolism. The green nodes are the ones found in humans, while the total network covers other species as well (KEGG, 2022b).

KEGG uses rectangles to symbolize genes (or proteins), with circles representing compounds. In **figure 4.3**, the ‘Homo sapiens mode’ has been chosen, and the genes belonging to Homo sapiens are marked in green. This way of representing a biological pathway sadly reduces the chance of patterns such as bottlenecks or hubs, as the different paths are spread out and not intertwined. The same gene can occur on multiple ‘branches’, and hence, any potential patterns are lost. It does however make it easy to investigate the many possible branches.

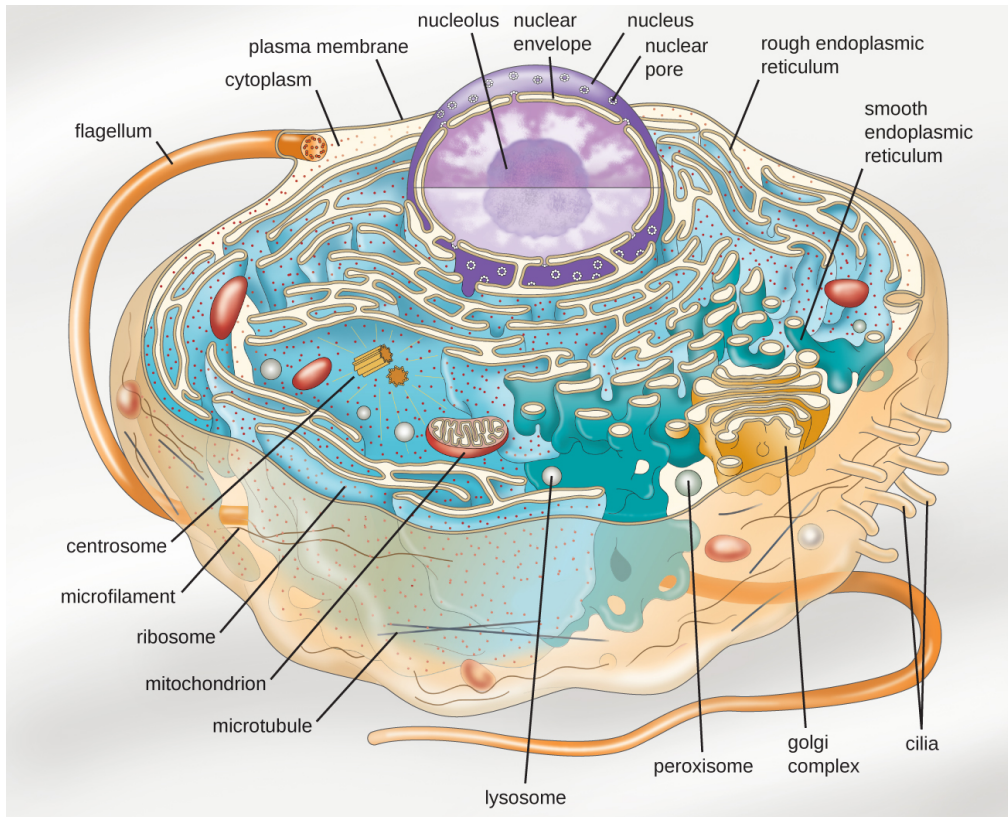
KEGG pathways can be downloaded as XML files, where the network is represented in tree structures. This feature is utilized in **Paper 2 and 3**, where the XML files are altered to show more information (**Paper 2**) or separated into versions representing different subcellular locations (**Paper 3**). The way we use these XMLs are shown and discussed more in **sections Changing the XML files and Hard coding the XML files**.

### 4.1.2 Cytoscape

Visualisation is an important tool in analysing biological pathways, and Cytoscape (Shannon et al., 2003) is one of many platforms that can help in this regard. Cytoscape was created for biological research, but today it is used as a general platform for complex network analysis and visualisation. The Cytoscape core is open source and provides basic features for building networks and integrating, analysing, and visualising data. Networks can be built from the outset based on the theory shown in **figure 4.1**. It is also possible to use ready-made networks from databases such as KEGG, WikiPathways (Martens et al., 2021), or Reactome (Fabregat et al., 2018). In addition to the core, a wide variety of applications can be downloaded. These applications are available for layouts, network and molecular analyses, additional file format support, scripting, and connections to databases. Being open source means that anyone can access the code and modify it, come with suggestions or create Cytoscape apps for others to use. Cytoscape's Java-based open API is used for app creation, and the app community is always open to new members.

## 4.2 The cell

The biological pathways mentioned so often throughout this thesis are found within human cells, in our studies they are either coming from tissue samples or cell cultures. In general, a human cell is bounded by a semi-permeable plasma membrane, consisting of a bilayer of phospholipids. Small uncharged molecules can pass through the membrane, and there are various proteins and protein complexes that serve as gates for larger and charged molecules. The interior of the cell is known as cytoplasm, and in eukaryotic cells such as human cells, the cytoplasm refers to the area between the nucleus and the cell membrane, as can be seen in **figure 4.4** (Reece et al., 2014).



**Figure 4.4:** The view of a generalized, eukaryotic cell. Here we see some of the basic structure of a human cell, surrounded by a plasma membrane, containing organelles such as the nucleus, ER, golgi, and mitochondria, among others (Parker et al., 2016). CC BY 4.0, access figure for free at <https://openstax.org/books/microbiology/pages/3-4-unique-characteristics-of-eukaryotic-cells>.

Within the cytoplasm are multiple organelles, all with a specialised shape and function. These are membrane-bound structures that all serve a particular purpose for the cell and the organism. For instance, the mRNA we measure in our studies is made in the nucleus, where the DNA is stored. Surrounding the nucleus is rough and smooth endoplasmic reticulum (ER), which is active in membrane synthesis, and the ribosomes that make the ER rough are the ones that translate the mRNA into proteins. An important organelle in **Paper 3** is the mitochondria, which is known as "the power house of the cell", due to being the host for glycolysis and oxidative phosphorylation. As seen in **figure 4.4**, there are also multiple other organelles and structures, such as the cytoskeleton, the golgi apparatus, peroxisomes, and lysosomes (Reece et al., 2014). All of these, including the ones briefly mentioned, could be discussed through many books. However, here I focus on measuring the mRNAs and the products of them.

In order to learn more about how the cell works, it is important to know more about the subcellular localisation of proteins. In **Paper 3** we look into subcellular localisation. Here we combine the localisation data with differential expression from RNA-seq and metabolic pathways from KEGG, and combine it all with our method FunHoP. Finding the subcellular localisation can be done either experimentally or predicted by computers. Experimentally, it is possible to use for instance isotope-labeled C-atoms (Chokkathukalam et al., 2014), antibodies and immunofluorescence (Lundberg and Borner, 2019), or mass spectrometry, which is the method used by the SubCellBarCode (Orre et al., 2019) used in **Paper 3**. Another method used in this study is the Bologna Unified Subcellular Component Annotator (BUSCA), which predicts localisation based on known amino acid patterns such as GPI anchors, and signal and transit peptides. It is also possible to use transmembrane domains like alpha-helices and beta-barrels (Savojarado et al., 2018).

And with that brief introduction to the cell and its components, it is time to move over to metabolism.

### 4.3 Metabolism

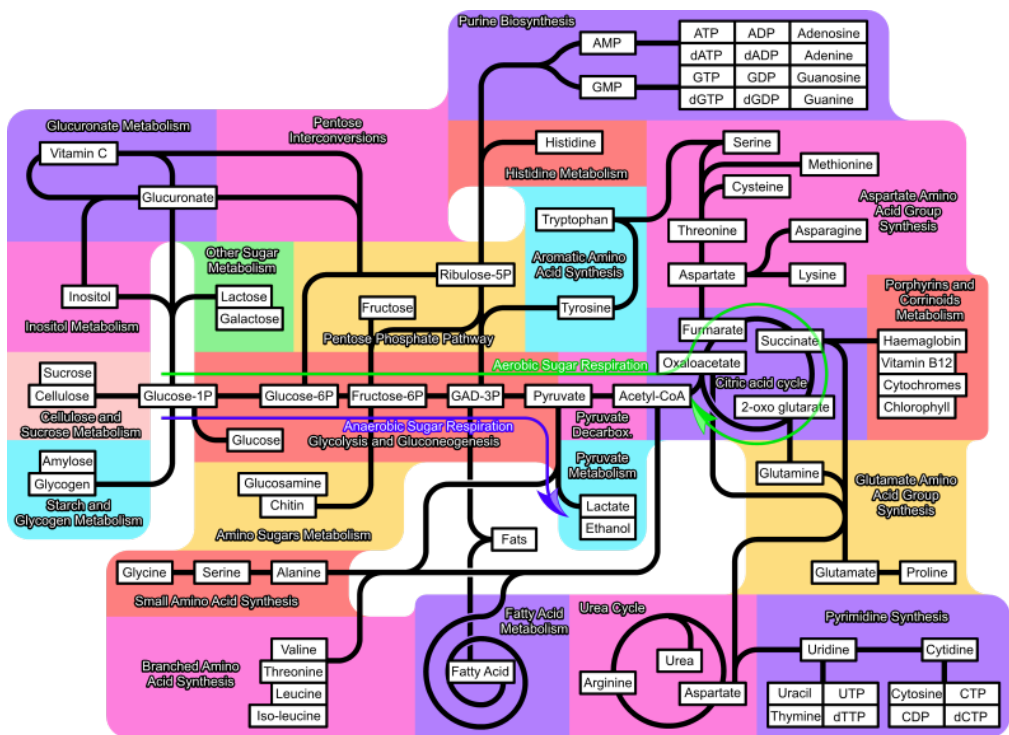
The word ‘metabolism’ comes from the Greek word *metabolē*, which means ‘change’. Metabolic changes are either anabolic (‘building up’) or catabolic (‘breaking down’). The purpose of metabolism is to harvest energy from food, to create building blocks in the form of proteins, lipids, carbohydrates as well as nucleic acids, and to remove metabolic waste (Nelson and Cox, 2008). The reactants, products, and intermediates in these processes are called metabolites, which can be sorted into two groups, namely primary (central) and secondary (specialised) metabolites.

Primary or central metabolites are directly involved in growth and development processes. They include amino acids, nucleotides, sugars, as well as mono-, di-, and tricarboxylic acids. Primary metabolites are produced in large quantities, making them easy to extract, and are not species-specific, meaning they can be found in multiple organisms. Secondary or specialised metabolites are compounds not directly involved in growth or development. They are generally produced in smaller quantities and can be harder to extract. They can also be species-specific, such as certain antibiotics (Reece et al., 2014).

A metabolic pathway links the chemical reactions that occur to create or break down a specific metabolite. Yielding energy is the main goal of catabolic pathways, along with obtaining smaller components for anabolic pathways. Organic compounds contain potential energy in the bonds between their atoms, and a gradual harvest of energy can be achieved through the creation of adenosine triphosphate (ATP) and reduced nicotinamide adenine dinucleotide (NADH). ATP is either created by adding a phosphate group to adenosine diphosphate (ADP) or through chemiosmosis driven by the electron transport chain in oxidative phosphorylation. NADH is created when  $\text{NAD}^+$  is reduced and gains a hydrogen atom and two electrons. The easy cycling between oxidized ( $\text{NAD}^+$ ) and reduced (NADH) states is what makes  $\text{NAD}^+$  a well-suited electron carrier (Reece et al., 2014). ATP and NADH are involved in all pathways, even if they are not always shown: for example, in the KEGG pathways which are central to this thesis.

However, energy yield should be kept in mind, even if it is not visible. This lack of information will be further discussed in the **Discussion**.

The metabolic pathways of a human cell are intertwined in a complex system (**figure 4.5**). Products or intermediates of one reaction can be reactants in another, and vice versa. This must always be kept in mind when studying isolated pathways, and it is important not to be overly focused on a single pathway. However, when working with them, it is easier to separate metabolic pathways into categories based on their reactants or products.



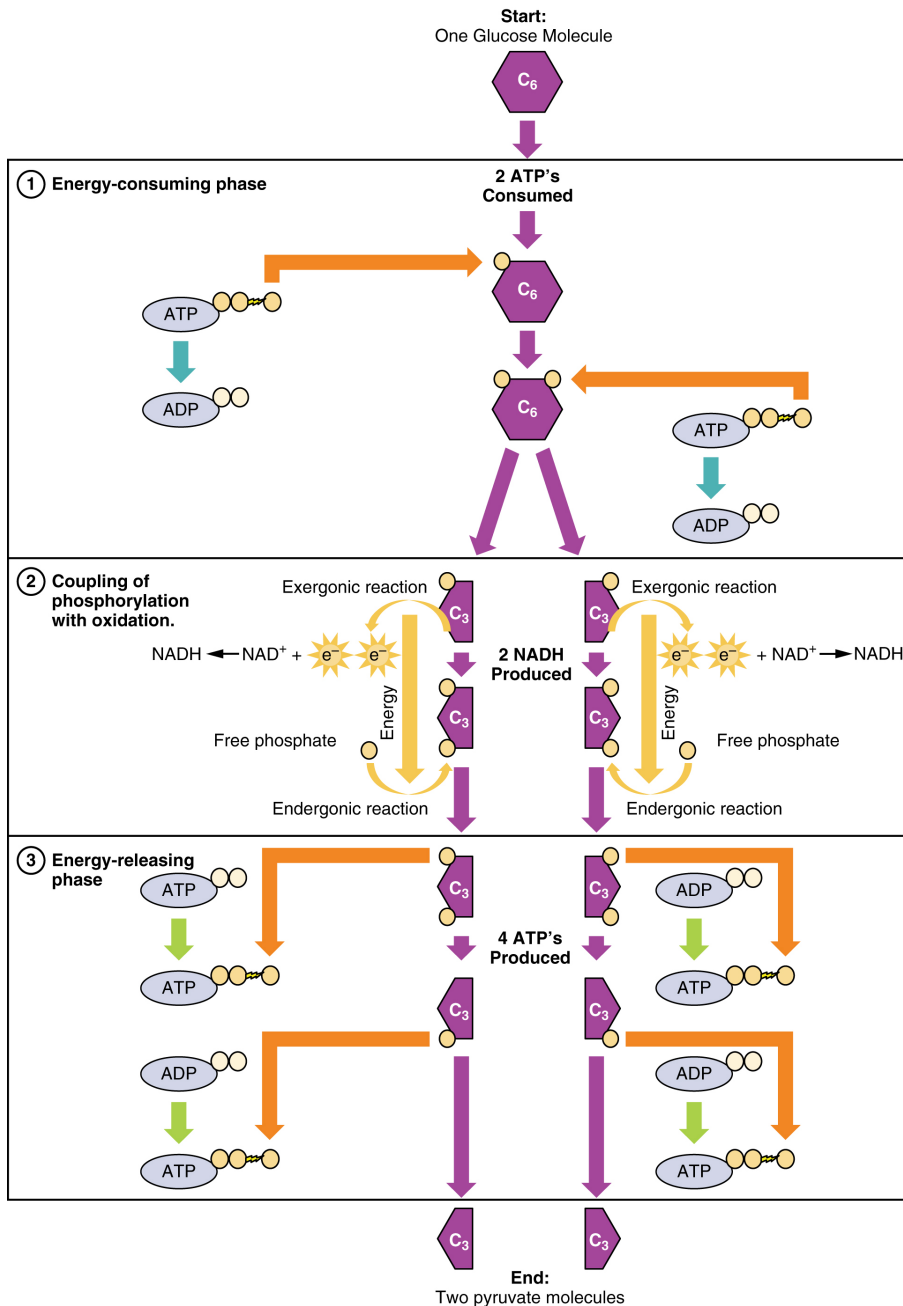
**Figure 4.5:** The massive and complex human metabolism consists of multiple different types of metabolism, such as nucleotide, lipid, energy, and amino acid metabolism, all connected either directly or indirectly (Zephyris, 2022). Figure is licensed under CC BY-SA 3.0.

**Figure 4.5** shows how some of the more central pathways in metabolism are marked, such as lipid metabolism, carbohydrate metabolism, and nucleotide metabolism. As they are relevant for **Paper 2 and 3**, glycolysis and the tricarboxylic acid cycle (TCA) will be discussed in greater detail in the following section.



### 4.3.1 Carbohydrates

Carbohydrates are one of the most easily catalysed sources of energy in cells. Larger polysaccharides such as starch and glycogen are hydrolysed to glucose, a six-carbon carbohydrate. Taking place in the cytosol, glycolysis breaks down glucose into two molecules of pyruvate, a three-carbon sugar, as shown in **figure 4.6** (Reece et al., 2014).

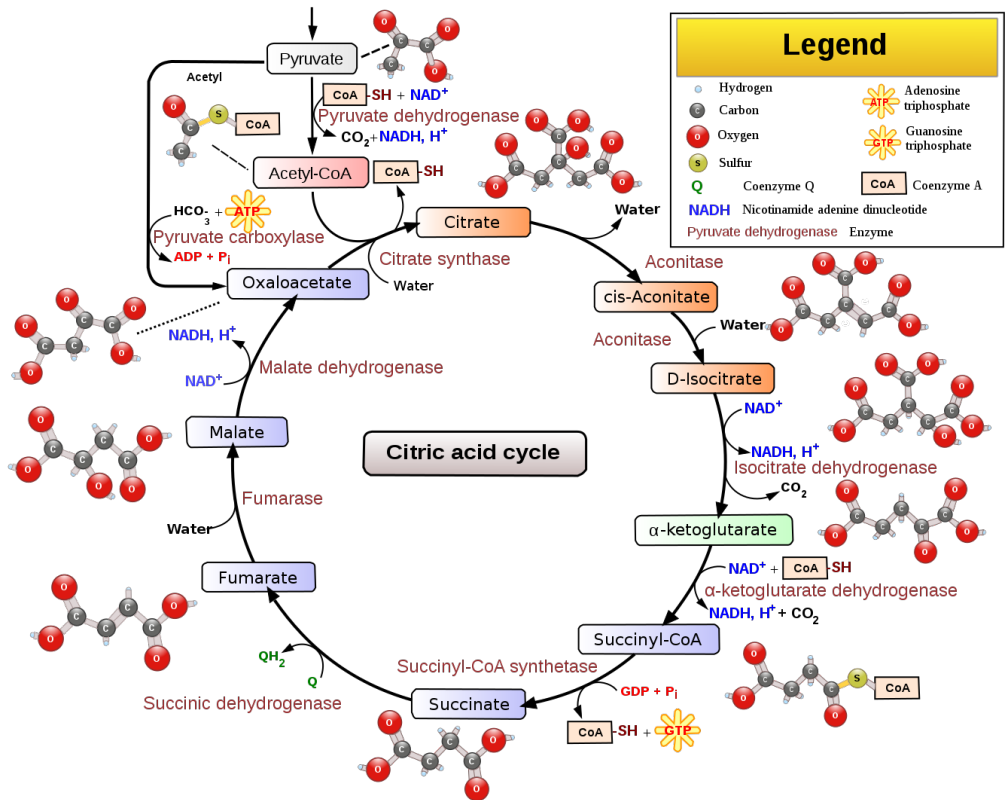


**Figure 4.6:** Glycolysis is the process where glucose is converted into pyruvate. Energy is used to split the circular glucose molecule into two linear ones, and energy is harvested by making ATP while two pyruvate molecules are made (Gordon Betts et al., 2022). CC BY 4.0, access for free at <https://openstax.org/books/anatomy-and-physiology-2e/pages/24-2-carbohydrate-metabolism>.

In the first steps of the glycolysis, ATP is used to transfer a phosphate group to the glucose molecule, making it more chemically reactive and keeping it inside the cell due to the charge on the phosphate. Two ATP molecules are used to add two phosphate groups, after which the sugar molecule is split into two molecules of three-carbon sugars. During the further oxidation of these three-carbon sugars, four ATP molecules are created, in addition to two NADH molecules, which can later be used in harvesting energy during oxidative phosphorylation (or in multiple other reactions in the cell). Provided oxygen is available, the two pyruvate molecules can be transferred into a mitochondrion and oxidized into acetyl coenzyme A (acetyl-CoA), which can be further used in TCA. This is the first step in which  $\text{CO}_2$  is released (Reece et al., 2014).

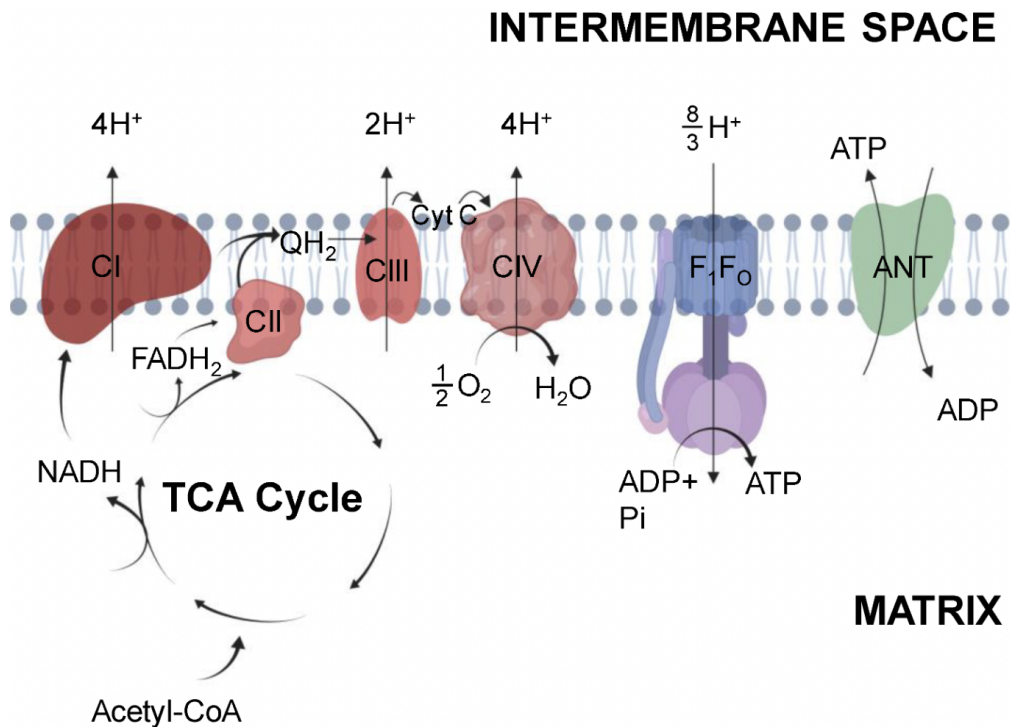
Once acetyl-CoA is available, its two-carbon acetyl group can be added to oxaloacetate, which creates citrate, as shown in **figure 4.7**. Step 2a is a dehydration reaction, where citrate is converted into isocitrate by the enzyme aconitase (ACO2). This has an important function in PCa, which will be discussed in more detail in **Prostate cancer and altered metabolism** (Reece et al., 2014).

**Figure 4.7** also shows how three NADH molecules are created by the reduction of  $\text{NAD}^+$ , in addition to one  $\text{FADH}_2$  and one molecule of ATP via guanosine diphosphate (GDP) and guanosine triphosphate (GTP). It also shows how more  $\text{CO}_2$  is released and how  $\text{H}_2\text{O}$  is both used and released. The remaining molecule at the end of the cycle is oxaloacetate, which can be added to new acetyl-CoA (for instance, from pyruvate), and the whole circle can be repeated. In this manner, energy is extracted from glucose and stored mainly in NADH and  $\text{FAD}_2$  molecules. To obtain more energy in the form of ATP from NADH and  $\text{FAD}_2$ , these two electron transporters must release their energy in the electron transport chain (Reece et al., 2014).



**Figure 4.7:** Acetyl-CoA enters the TCA cycle and is combined with oxaloacetate and becomes citrate. During a full circle, NADH, FADH<sub>2</sub>, and ATP/GTP is created, H<sub>2</sub>O and CO<sub>2</sub> is released, and the final compound is oxaloacetate, which can start the whole thing over again with more Acetyl-CoA (Narayanese, 2008). Figure is licensed under CC BY-SA 3.0.

The electron transport chain consists of a collection of protein complexes in the inner mitochondrial membrane. As the inner mitochondrial membrane is folded, its areal surface is increased, creating space for thousands of copies of the electron transport chain’s protein complexes. These protein complexes pump H<sup>+</sup> across the membrane, while the electrons are passed along the chain to their final acceptor O<sub>2</sub>, creating H<sub>2</sub>O, as shown in **figure 4.8** (Reece et al., 2014).



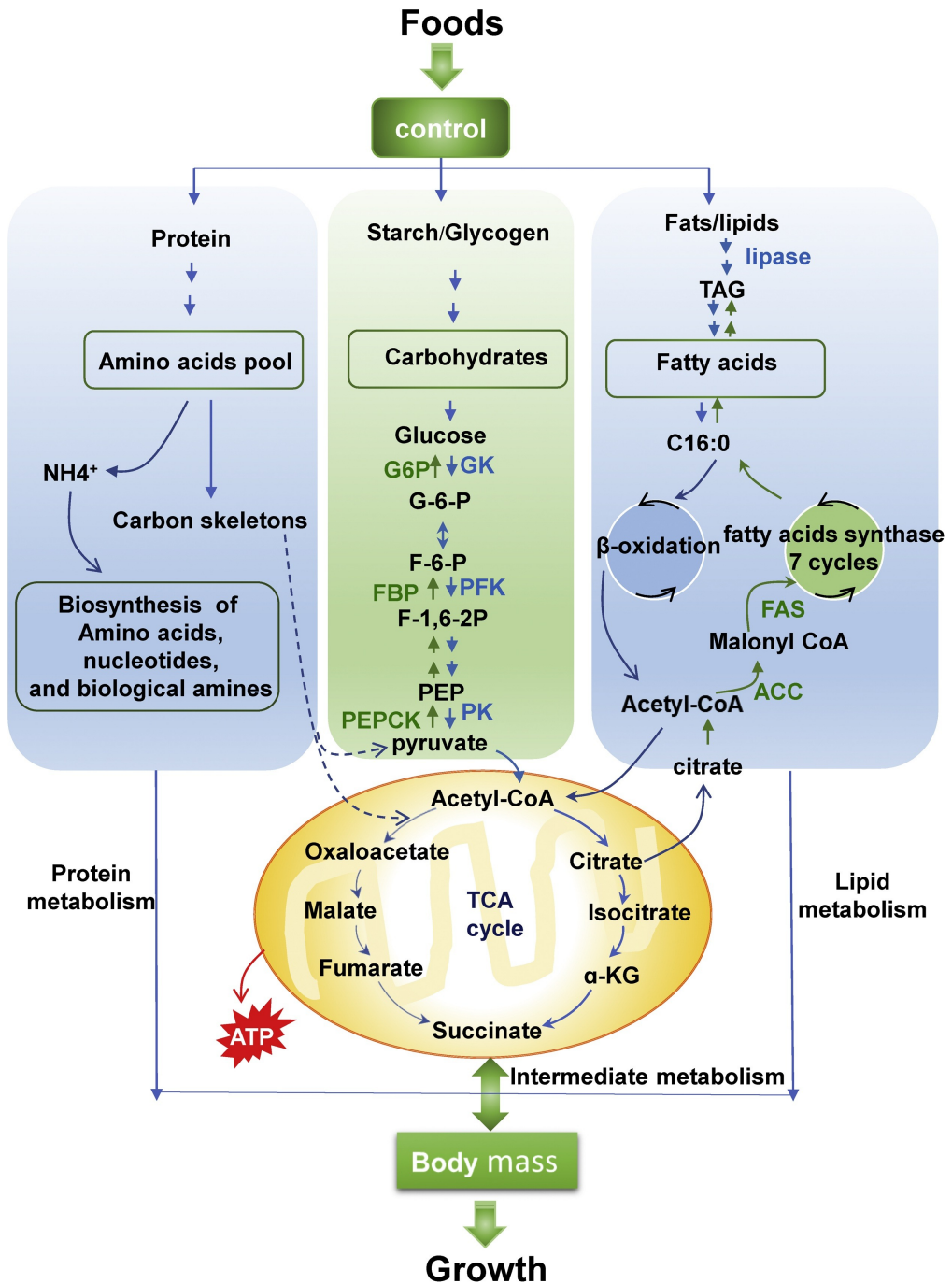
**Figure 4.8:** Oxidative Phosphorylation takes place in the inner mitochondrial membrane. Pumping  $H^+$  across the membrane creates an osmotic potential, which drives the ATPase further down in the chain, creating ATP (Lewis et al., 2019), CC BY 4.0.

The pumping of  $H^+$  across the membrane creates a difference in  $H^+$  on opposite sides of the membrane, and this difference pushes the  $H^+$  through ATP synthase, which in turn creates ATP. To summarize, energy has been harvested from a glucose molecule and 30–32 molecules of ATP have been created (Reece et al., 2014).

As mentioned earlier, this aerobic oxidation only happens if oxygen is present, working as an electronegative pull for the electrons in the transport chain. If oxygen is not available, the whole cycle stops. NADH is not oxidized, and there is no  $NAD^+$  to harvest energy. However, as certain organisms live under anaerobic conditions, a solution must exist. In this case, the solution is fermentation. Glycolysis still takes place during fermentation, but the pyruvate molecules are not transported into the mitochondria and converted to acetyl-CoA. Instead, they stay in the cytosol. Two of the most common types of fermentation are alcohol fermentation, where pyruvate is converted to ethanol, or lactic acid fermentation, where pyruvate is reduced directly to NADH and lactate. The latter happens in human muscles during a hard workout (Reece et al., 2014). Lactic acid fermentation will be discussed further in 4.4.2.

Energy sources other than carbohydrates also exist. **Figure 4.9** shows the four processes already mentioned, namely glycolysis, conversion to acetyl-CoA, TCA, and oxidative phosphorylation. It also illustrates how

carbohydrates, proteins, and fatty acids can enter these processes at various stages.



**Figure 4.9:** Nutrients such as proteins, carbohydrates, and fats can be catabolized and incorporated at different places of the energy metabolism chain (Zhang et al., 2020), CC BY-NC-ND 4.0.

### 4.3.2 Metabolic networks

The networks representing human metabolism are intended to show the metabolic biochemistry that takes place in a human cell, as explained in the beginning of **section 4.3**. Creating metabolic pathways is a way of sorting out specific components from start (substrate) to finish (product) via a path (enzymes and intermediates), from the intricate mixture of metabolites inside a human cell. Yet cellular metabolism is much more plastic and complex than its linear textbook representation (Schuster et al., 2000; Yarmush and Berthiaume, 1997).

A main goal of today's functional genomics is to complete the reconstruction of metabolic pathways, but researchers studied metabolism long before the dawn of the genomic era. In 1945, Horowitz made one of the first attempts at describing a metabolic pathway when he described the retrograde hypothesis. This suggests that if the biosynthesis of compound A requires sequential transformations by B, C, and D via corresponding enzymes, the final product A would have been the first compound (of these) used by the primordial heterotrophs. According to Horowitz, if A was essential for the survival of primordial cells and the primitive soup was depleted of it, selective pressure and the production of cells able to transform B into A would lead to the creation of the very first pathway. This would include enzyme **a** to catalyse the transformation from A to B. Variants with enzyme **b** in addition to **a** would have possibilities to create more B and build up a more complex pathway. With A and B being chemically related, **a** and **b** would also be related, and the theory is that a duplication of the a gene, **a**, leads to the creation of the b gene, **b**. The theory further suggests that these duplication events leading to similar genes means that similar genes are to be found within the same areas (Horowitz, 1945).

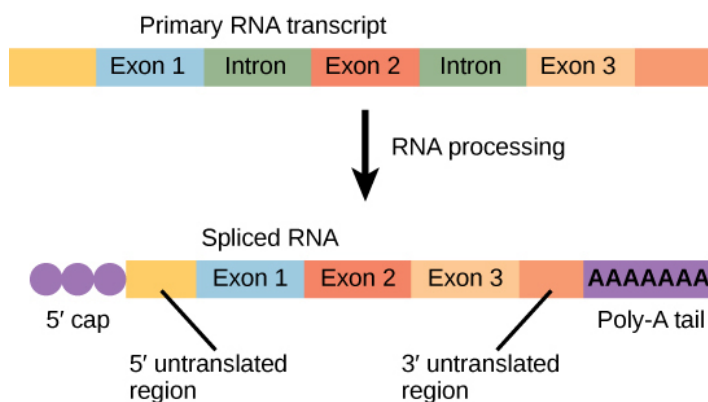
After the early 1960's discovery of operons, a prokaryotic feature in which similar genes are controlled by the same operator, Horowitz argued that genes belonging to the same operon or the same pathway were a result of series or tandem duplications which created a paralogous gene family (Horowitz, 1965). Other hypotheses regarding gene duplication and the formation of pathways have since been published, such as the patchwork hypothesis by Ycas and Jensen (Yčas, 1974; Jensen, 1976). They suggested that an ancestral enzyme, E0, had a very low substrate specificity and could bind to three different substrates, thereby catalysing three different yet similar reactions. Duplications of E0 would lead to more specific enzymes which would have higher substrate specificity and lead to different metabolic routes.

Researchers have been studying metabolic pathways for decades. Examples include early studies on the catalytic action of chymotrypsin (Kraut, 1977; Bender et al., 1973) or glycolysis, which has been researched extensively (Chen and Geiling, 1946; Wu et al., 1964; Villar-Palasi and Lerner, 1970; Ottaway and Mowbray, 1977; Melendez-Hevia and Siverio, 1984; Fernie et al., 2004). With the central goal of functional genomics being to determine the metabolic routes from a specific start to an end product, such research will continue (Schuster et al., 2000). Metabolic flux balance analysis and genome-scale reconstructions are growing fields of research, meaning that our knowledge of the complexity of metabolism is increasing (Schilling and Palsson, 1998; Förster et al., 2003; Orth et al., 2010; Brunk et al., 2018).

### 4.3.3 Gene expression

A key element of this thesis is gene expression. Gene expression comparison between different states and groups is the basis for finding pathways in **Paper 1**, for colouring nodes in **Paper 2 and 3**, and for suggestions on location-specific regulation in **Paper 3**. In **Paper 1**, gene expression has been measured using RNA microarray, the previous gold standard for gene expression, whereas **Paper 2 and 3** use RNA-sequencing. Both these techniques will be discussed below after a general explanation of the process.

The DNA contains the recipe for all the possible proteins a cell can make. Protein levels are regulated on multiple levels. On the catabolic side, ingested proteins are broken down into amino acids, which can be catabolized again into smaller molecules and enter the energy harvest cycle at various stages. Humans require 20 amino acids to make proteins and can synthesize 12 of these. The remaining eight must be obtained by ingesting proteins. Most animal products, such as eggs and meat, contain all 20 amino acids, including the eight essential ones. Having amino acids available is crucial for building new proteins on demand. Most human cells have a nucleus in which DNA is transcribed into primary RNA. The primary RNA is processed to create a finished mRNA, which leaves the cell and is translated into a protein. The finished mRNA has a ‘cap’ at the 5’ end (‘the beginning’), and a poly-adenine tail at the 3’ end (‘the end’), and untranslated regions (UTRs) can be found between the cap and the tail and the actual coding segment. The 3’ UTR contains information regarding the mRNA’s location (Alberts, 2008). The process of going from primary RNA to mRNA is shown in **figure 4.10** and is explained further in the **Discussion**.



**Figure 4.10:** The primary RNA transcript with its mixture of introns and exons needs to be processed before the mRNA can leave the nucleus. A 5’ cap and a 3’ poly-A tail is added, and exons are combined into the finished mRNA while the introns are left behind. Each end of the mRNA contains an UTR, and the 3’ end UTR contains information regarding the mRNA’s location (Fowler et al., 2013). CC BY 4.0, access for free at <https://openstax.org/books/concepts-biology/pages/9-3-transcription>.

In contrast to prokaryotes, eukaryotes have exons and introns in their DNA. As exons are spliced together to form the final mRNA, and due to alternative splicing, the same gene can become different mRNAs, and hence different proteins. Once the mRNA is fully processed, it leaves the nucleus and is translated by

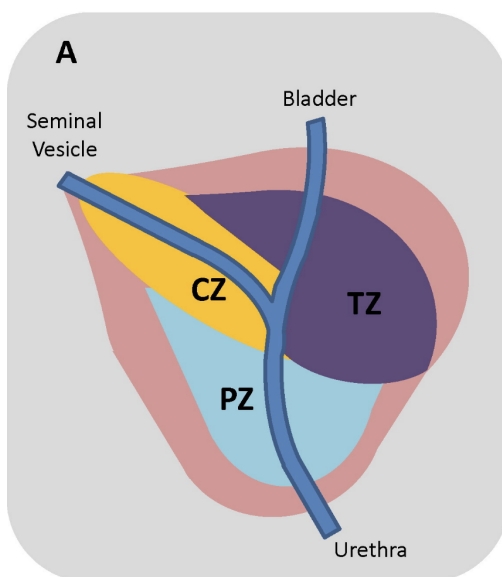


ribosomes. As the mRNA moves through a ribosome, codons on the mRNA are translated into an amino acid chain one at a time. The carboxyl end of one amino acid binds to the amino end of another amino acid through a peptide bond. This linear chain of amino acids is known as the primary structure of the protein, and it dictates the secondary and tertiary structure due to the chemical nature of the polypeptide backbone and the R groups of the amino acids. The secondary structure consists of alpha-helix coils and beta sheets and is a result of hydrogen bonds between the repeated constituents in the backbone. The R groups of the various amino acids also interact and shape the polypeptide, thus creating the tertiary structure. Both hydrophobic interactions and disulfide bridges are among the interactions that contribute to the shaping of the protein. The main challenge of using mRNAs as a measure of proteins is the lack of a known relationship between gene and final protein, level of gene expression and level of active protein. However, in this thesis, we still use mRNA levels as an approximation of protein levels, as it is easier to identify and measure than its end products (Nelson and Cox, 2008).

## 4.4 The prostate

### 4.4.1 The normal prostate

The prostate gland is a walnut-sized gland located below the bladder and surrounding the urethra in mammalian males (Marker et al., 2003). It is the largest accessory gland of the male genital tract (Kosova et al., 2014). As shown in **figure 4.11**, the prostate has three distinct zones, namely the central zone (CZ), the peripheral zone (PZ), and the transitional zone (TZ), and its main purpose is to produce and contain prostatic fluids. A unique trait of the prostate concerns its citrate production and levels – the prostatic fluid contains high levels of citrate, 20–70 times higher than usually found in tissue and 400–1,500 times higher than in blood plasma (Costello and Franklin, 2009; Kavanagh, 1985). This citrate is accumulated in the TCA cycle, where m-aconitase (ACO2, reaction 2a in 4.7) in most cells would convert citrate to isocitrate, which would then be converted further and release energy. However, the PZ glandular epithelial cells in the prostate accumulate high levels of zinc, which inhibits aconitase. Therefore, citrate is not converted. The zinc uptake in these prostate epithelial cells is a result of the expression and activity of ZIP1, a zinc uptake transporter (the *SLC39A1* gene; (Costello and Franklin, 2009).



**Figure 4.11:** The prostate gland contains three zones, namely the central zone (CZ), the transitional zone (TZ), and the peripheral zone (PZ). (Packer and Maitland, 2016), Elsevier open access license.

When citrate is secreted instead of being used in the TCA cycle, the benign prostate cells do not use oxidative phosphorylation as their main resource for energy harvesting. This is in contrast to most human cells, where the TCA cycle and the subsequent oxidative phosphorylation is the most common way of utilizing carbohydrates. The fact that prostate metabolism is different from other cells makes it important to get a better understanding of it, and we can to a lesser extent just use data from other cell types. We have to get as much information as possible out of prostate cells.

The prostate can cause health problems later in life, and most men who reach the age of 80 are affected by prostate disease. The most common type is benign prostatic hyperplasia (BPH;(Berry et al., 1984) (Fitzpatrick, 2006)). BPH is a non-cancerous enlargement of the prostate, and symptoms of BPH are similar to those of PCa (Kim et al., 2016). However, BPH is not cancer, despite the similar symptoms. It is therefore important to learn as much as possible about the features of prostate cancer, and especially to find non-invasive biological markers that can be used to differentiate between cancer and BPH.

#### **4.4.2 Prostate cancer and altered metabolism**

PCa is the most common type of cancer in males, with 5030 diagnosed in Norway in 2019. About 90% of these men were over 60. PCa is generally slow-growing, and 95.5% of patients are still alive five years after diagnosis (Kreftforeningen, 2022). Symptoms will normally not be visible in the early stages of PCa, but as the cancer grows, patients might experience a weak urine flow and frequent urination, trouble with emptying the bladder, and blood in the urine. If the cancer has progressed, patients might also experience back pain (Kreftforeningen, 2022). If cancer is suspected, primary examinations will be performed, including a rectal examination and a blood sample. Screening for elevations in prostate-specific antigens can be deceptive and should not be used as the only source in PCa screening.

During malignant transformation, or the act of growing into cancerous cells, cells gradually evolve from benign to malignant (Brawer, 2005). PCa begins in the peripheral zone epithelium cells, which are programmed to produce and not oxidize citrate (Costello et al., 2004; Costello and Franklin, 2016). During this process, the epithelial cells stop secreting citrate, instead reactivating the TCA cycle and starting to oxidize citrate (Zadra et al., 2013). The high levels of zinc must be decreased to avoid apoptosis, and it has been suggested that this decrease is caused by alterations in ZIP1 (Franz et al., 2013; Feng et al., 2002). Zinc levels must also be kept low to alter the inhibition of aconitase, the enzyme that converts citrate to isocitrate.

In **Paper 1** we study the reactive stroma. The stroma is generally known as the part of the tissue with a structural or connective role, and consists of the non-epithelial components, such as blood vessels, fibroblasts, extracellular matrix (ECM), immune cells, and nerves. The tumor microenvironment (TME) of PCa has chronic inflammation, and is referred to as reactive stroma. The transition from healthy to reactive stroma is part of the cancer process, and reactive stroma appears to play an important part in cancer development (Tuxhorn et al., 2002; Barron and Rowley, 2012). Using the reactive stroma grading

system (Ayala et al., 2003), we were able to divide the samples in **Paper 1** into four groups, and use these as a basis for the differential expression analysis and the correlation analysis.

The aforementioned lactic acid fermentation is the result of the so-called Warburg effect, which can be observed in many types of solid tumours. This effect shows that many tumours have an increased glycolytic rate despite having lower access to O<sub>2</sub>. This means aerobic glycolysis is the dominant ATP-producing pathway, and that cancerous cells take up more glucose to meet their need for ATP. Warburg suggested that cancerous cells sustain irreversible damage at some point during oxidative phosphorylation (Warburg, 1956). ATP synthase consists of many subunits, and many studies show that some of these subunits are downregulated in cancers. For instance, Cuezva et al. have found a reduction in mitochondrial markers, such as the beta-catalytic subunit of H<sup>+</sup>-ATP synthase ( $\beta$ -F1 ATPase) in the human liver, kidney, and colon (Cuezva et al., 2002). Another study, by Isidoro et al., has analysed both mitochondrial and glycolytic protein markers from gastric, prostate, and breast adenocarcinomas and squamous oesophageal and lung carcinomas. These markers include the  $\beta$ -F1 ATPase and HS P60 of mitochondrial marker proteins and GAPDH and PK in the cytosolic proteins. In all cancers except PCa,  $\beta$ -F1 ATPase is found to be downregulated, with no difference (Isidoro et al., 2004). This trait distinguishes PCa from other types of cancers.

Amino acids are the building blocks of proteins, and the utilization of particular amino acids can be observed in the promotion of cancer cell growth. They can also be used in determining the aggressiveness of cancer (Wang et al., 2013). For instance, glutamine is involved in multiple pathways in the cell. Glutamine uptake is found to be upregulated in multiple types of cancer, including PCa, where it can be used in *de novo* lipid biosynthesis (Eidelman et al., 2017). Glutaminolysis is a way for cancer cells to produce ATP, and upregulation of the glutaminase-1 responsible has been found (Moncada et al., 2012; Pan et al., 2015). An *in vitro* study (Wang et al., 2015) has also shown that the inhibition of glutamine uptake limits proliferation and invasiveness.

Another relevant amino acid is arginine, which can be converted to both glutamine and proline and is important in the generation of nitric oxide (NO). Neither arginine nor NO's role in cancer is fully known, although NO is thought to play an important role (Qiu et al., 2015), and *in vitro* studies have shown that starving cells of arginine kills the cells and that high availability of arginine is necessary for cell survival (Feun et al., 2008; Kim et al., 2009).

# Methods and development

## 5.1 Techniques

As already shown, the human metabolome is massive. In addition to the features discussed in previous sections, it also includes acids, amines, vitamins, minerals, drugs, food additives, and other compounds that humans ingest and/or metabolize. A problem with studying the metabolome is that the levels of expressed genes, which are easier to measure, cannot be directly correlated with the levels of metabolites and proteins. Hence, what can be measured does not easily translate to the knowledge that researchers are seeking.

In this chapter I will generally present some of the techniques for measuring metabolites and gene expression, as well as going through the development of our new method FunHoP.

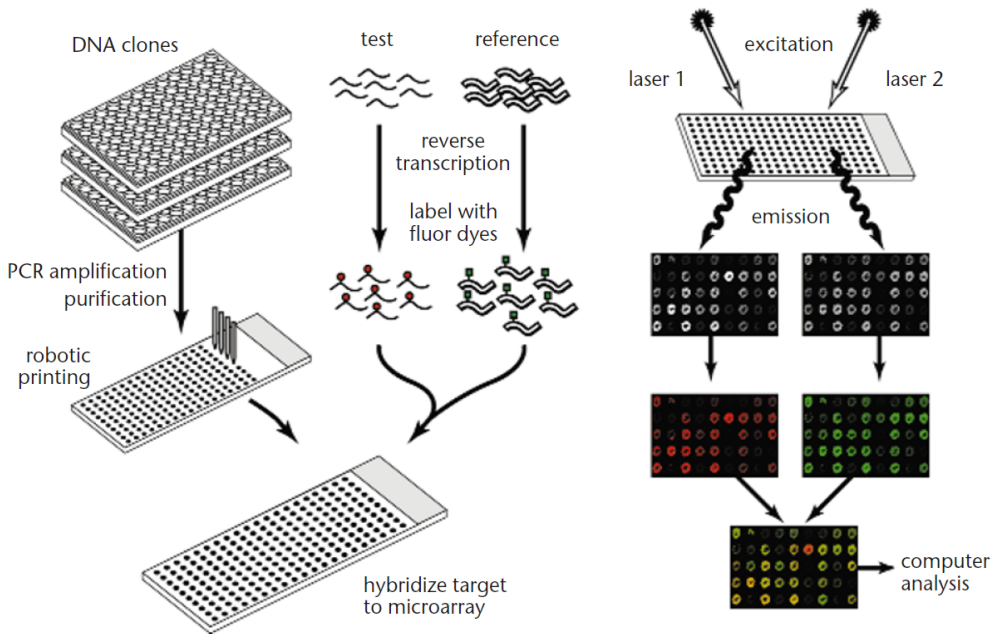
### 5.1.1 Identification of metabolites

The metabolite data in **Paper 1** was obtained by high-resolution magic angle spinning magnetic resonance spectroscopy (HR-MAS MRS), which is a well-established technique for analysing biological tissue (Giskeødegård et al., 2013). HR-MAS is a non-destructive method that provides a snapshot of the metabolic status. As the sample remains intact, it can be used for other types of analysis such as proteomics or genomics data, and data from the same sample will be more comparable than data from different samples (Moestue et al., 2011).

### 5.1.2 Measuring gene expression

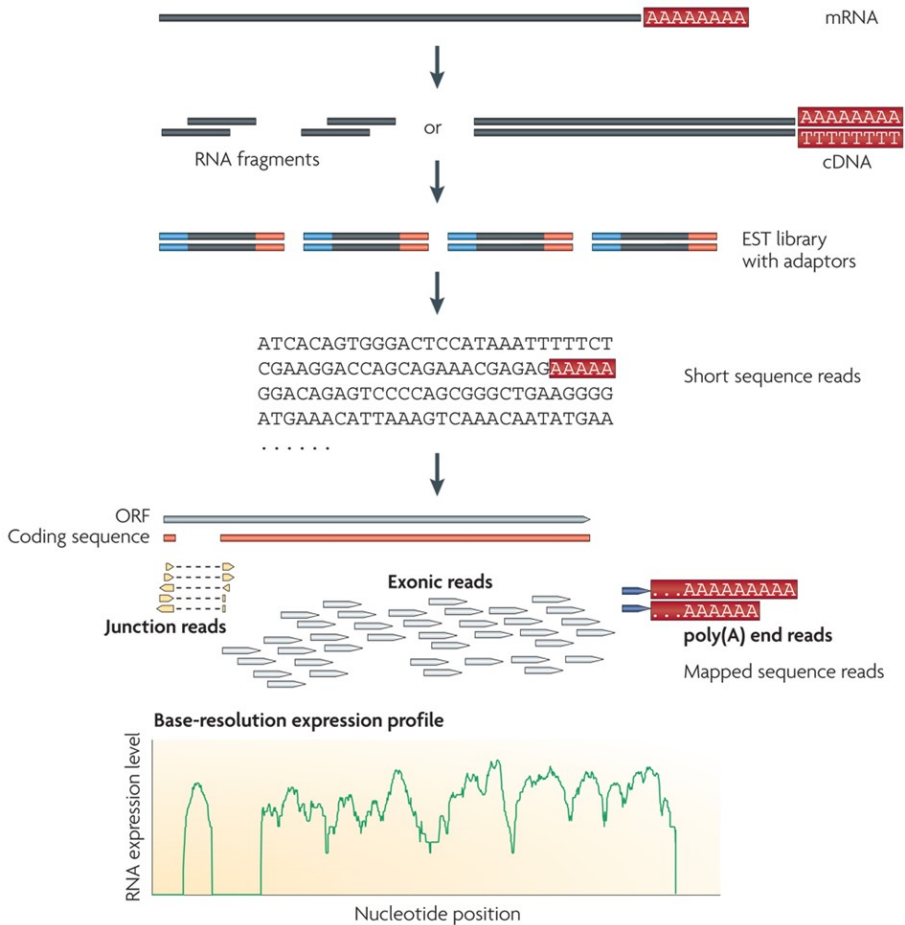
As previously mentioned, the two applied techniques for measuring gene expression in this thesis are microarray and RNA-sequencing. Microarray hybridization was the gold standard for gene expression from the late 1970s until the 2000s (Bumgarner, 2013), even if Fredrick Sanger provided the prototype of Sanger Sequencing in 1977 (Sanger et al., 1977). **Figure 5.1** shows a microscope slide containing multiple DNA fragments in specific positions. The mRNA molecules are converted into complementary DNA (cDNA) and labelled with fluorescent dyes (here red and green), before being hybridized

to the microarray. Lasers are used to measure the expression of each gene, and computer analysis is used to compare the two samples. If the expression is equal between the two samples, a spot will appear yellow, otherwise the sample with the highest expression of the gene will determine the colour of the spot (red or green, respectfully). The relationship between the two samples is known as the fold change.



**Figure 5.1:** cDNA of the samples are labelled with two different colours, before hybridized to the microarray. Lasers are used to measure the expression, and the two samples can be compared. The two colours will appear according to expression. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Genetics, (Duggan et al., 1999).

Sequencing became increasingly popular during the 2000s, as the technology became more available and affordable. **Figure 5.2** provides a brief overview of how RNA-sequencing works.

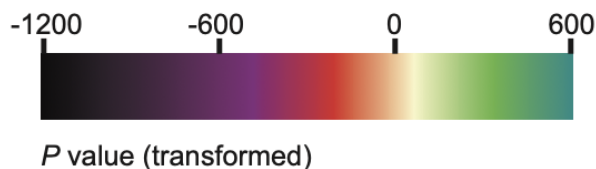


**Figure 5.2:** mRNA is converted into cDNA fragments, and made into a library with adaptors in the ends of each fragment. The molecules are sequenced from one or both ends, and the reads are either aligned to a reference genome or assembled *de novo*. This creates a genome-scale transcription map that contains the level of expression for each gene (the read count). Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Genetics, (Wang et al., 2009)

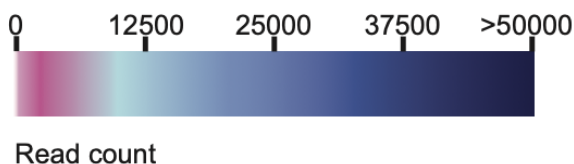
A library is created from the fragmented mRNAs from the sample, and the short sequences are read and mapped. This gives a read count for each of the genes, which can be used either for the sample alone or in comparison with another sample. The latter is used in **Paper 2**. Variations of this procedure can be applied to sequence genomic DNA, for example in search for variations, or perform single cell transcriptomics.

### 5.1.3 Using Cytoscape

For this thesis, the KEGGScape app (Nishida et al., 2014) was used to load KEGG XML files. Cytoscape has a feature that links directly to KEGG, but this would not allow access to FunHoP before loading. The networks in **Paper 2 and 3** were styled based on two types of data: transformed p-values from differential gene expression, and/or read counts from RNA-Seq. To achieve consistent styling, a unique colouring style was devised for each type of data (**figures 5.3 and 5.4**) and applied to all networks of each type.



**Figure 5.3:** Transformed P-values can be found on a scale from  $-1200$  (black) to  $600$  (dark blue-green), with a light yellow at zero and red or purple indicating downregulation and green indicating upregulation.



**Figure 5.4:** The scale for read counts goes from white at zero to a dark blue at above 50,000, with a light to a bright pink representing the numbers from 1 to 4000. Most of the genes were found in this area.

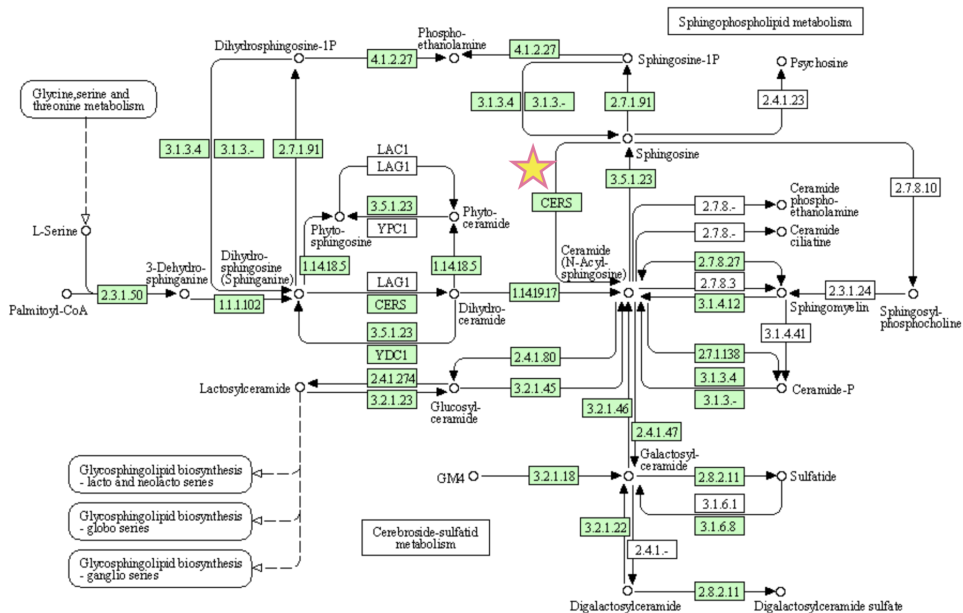
Additionally, the genes within nodes of any size were indicated by a rectangle with grey or purple as the default colour. This colour was retained for genes which were either not found or not significant in differential expression, read counts, or both. Links were shown in black and in the same size, with default KEGG directions where available. No additional information was added to the links. Metabolites were shown as small circles with grey or purple as the default colour.

## 5.2 Developing FunHoP

The underlying idea for FunHoP was conceived at a presentation by a medical doctor and researcher at the Norwegian University of Science and Technology in December 2015. The medical doctor showed a screenshot of a KEGG pathway and had marked one of the nodes with a star (similar to the illustration in **figure 5.5**). This star was supposed to indicate that the node was upregulated, and from this, conclusions as to the overall regulation of the pathway were drawn. This solution seemed far too simple, and I examined the pathway in question later that day. The star-marked node was found to contain multiple genes that



were not included in the earlier presentation. The examination of other nodes in the downstream pathway showed more nodes with many ‘hidden’ genes. Therefore, the conclusion about the overall regulation of the pathway was based on an unsatisfactory level of information.



**Figure 5.5:** Using sphingophospholipid metabolism as an example, we see how a gene (here *CERS*) can be marked by a star, which is supposed to indicate some form of important information. Additionally, this node, which seems to have only one gene, actually contains six (KEGG, 2022a).

It could be misleading for a scientist to see a pathway (such as that in **figure 5.5**) without the further addition of all the proteins that can participate in the reaction. It would be more useful to see all genes within a node and then determine which one of them were most likely doing the job.

### 5.2.1 Changing the XML files

Based on these observations, an idea took hold. When examining the KEGG pathways with KEGGScape in Cytoscape, only one gene was shown, regardless of the number of genes within the node. I wondered what would happen if the pathway files were simply changed to show all the genes, and whether some form of decision could be included regarding which genes within such a multigene node were responsible for the enzymatic activity in the node.

As demonstrated in **Paper 2**, the answer to this last question was yes. Both the KEGG website and a pathway loaded into Cytoscape will show only one gene in each node. However, the KEGG XML files contain all the information needed to see all genes within a node **figure 5.6**, and FunHoP performs a step-by-step modification of the XML files to show all the genes.

XML files are accessed through the ElementTree XML API, which is part of the Python Standard Library. **Figure 5.6** shows how the XML nodes are built, using entry ID 56 as an example. Each of these nodes, known as a ‘child’ in ElementTree, have a number of features for us to look at, with some particular ones that have been coloured. For instance they have their own entry ID (purple), name (green), and type (orange), along with reaction, and link. The ‘underchild’ is a section within the child, which contains the graphics information, such as name (red, this is shown in Cytoscape), a font colour, a background colour, a type, two coordinates (which determines the place within the network), and two variables for determining size (height in light blue). FunHoP leaves most of these untouched, but the first notable aspect is a comparison of the child name and the graphics name. The child name shows four gene IDs within the node, meaning that four genes are considered as homologs in this reaction. The graphics name shows the familiar ALDH3A1, followed by the genes’ names in other organisms. Cytoscape chooses the first of these and uses it as the name for the node.

```
<entry id="56" name="hsa:218 hsa:220 hsa:221 hsa:222" type="gene" reaction="rn:R04996"
  link="http://www.kegg.jp/dbget-bin/www_bget?hsa:218+hsa:220+hsa:221+hsa:222">
  <graphics name="ALDH3A1, ALDH3, ALDHIII..." fgcolor="#000000"
    bgcolor="#BFFFBF" type="rectangle" x="574" y="520" width="46" height="17"/>
</entry>
```

**Figure 5.6:** Original KEGG XML node. Notice how the node ID is marked with purple, the name in green, and the type in orange. There is also an inner “child”, which has the details for the graphics. This includes the name in red, and the height in light blue, which will be changed with FunHoP

The first stage in FunHoP is to change the graphic names. This is done by identifying all children with ‘type=“gene”’ with a name that contains more than one gene name (meaning that it contains an open space, which occurs only between two or more names that have been manipulated by FunHoP, in contrast to the original which has commas). As the child name contains the IDs, the corresponding gene names can be found on the KEGG website. These relations can be downloaded and are used to expand the graphics name string, as shown in **figure 5.7**.

```
<entry id="56" name="hsa:218 hsa:220 hsa:221 hsa:222" type="gene" reaction="rn:R04996"
  link="http://www.kegg.jp/dbget-bin/www_bget?hsa:218+hsa:220+hsa:221+hsa:222" >
  <graphics name="ALDH3A1 ALDH1A3 ALDH3B1 ALDH3B2" fgcolor="#000000"
    bgcolor="#BFFFBF" type="rectangle" x="574" y="520" width="46" height="17" />
</entry>
```

**Figure 5.7:** Expanding the graphics name (red), notice how it has changed into containing one name for each of the genes in the child name (green)

When the graphics name has been expanded (**figure 5.7**), it contains the names of all the genes. At this point in the process, nothing else has been changed. The next step is to expand the number of nodes to include all these genes. FunHoP accomplishes this by using the same strategy that KEGG uses for gene

complexes. This means creating a new child for each gene, and then connecting them in a new, larger group node. The first step in this part of the process is to remove three of the four genes from the node, leaving only the first one, as shown in **figure 5.8**.

```
<entry id="56" name="hsa:218" type="gene" reaction="rn:R04996"
link="http://www.kegg.jp/dbget-bin/www\_bget?hsa:218+hsa:220+hsa:221+hsa:222">
  <graphics name="ALDH3A1" fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle"
    x="574" y="639" width="46" height="17" />
</entry>
```

**Figure 5.8:** In this node, only the first gene is kept. Notice how the both name (green) and graphical name (red) now only contains data for the first gene from the original node. The remaining three will get their own nodes, before they are all linked in a group

The three remaining genes will then need their own sibling entries. These are shown in **figure 5.9**, all with type 'gene'. They all have only one gene in both name (green) and graphical name (red).

```
<entry id="132" name="hsa:220" type="gene" reaction="rn:R04996"
link="http://www.kegg.jp/dbget-bin/www\_bget?hsa:218+hsa:220+hsa:221+hsa:222" >
  <graphics name="ALDH1A3" fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle"
    x="574" y="656" width="46" height="17" />
</entry>
<entry id="133" name="hsa:221" type="gene" reaction="rn:R04996"
link="http://www.kegg.jp/dbget-bin/www\_bget?hsa:218+hsa:220+hsa:221+hsa:222" >
  <graphics name="ALDH3B1" fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle"
    x="574" y="673" width="46" height="17" />
</entry>
<entry id="134" name="hsa:222" type="gene" reaction="rn:R04996"
link="http://www.kegg.jp/dbget-bin/www\_bget?hsa:218+hsa:220+hsa:221+hsa:222" >
  <graphics name="ALDH3B2" fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle"
    x="574" y="690" width="46" height="17" />
</entry>
```

**Figure 5.9:** Three new entries have been created, for the three remaining genes from the original node. Notice how they have entries with a higher number than the original 56, these are added to the bottom of the file, and all new entries get their own number

It is noticeable that the new entries have much higher ID numbers than the original node from which they came. FunHoP allocates new IDs based on the first available ID and continues from there. The entry name and graphics name are also changed, while the other details are retained from the original. With the new entries made, the four entries in **figure 5.8** and **figure 5.9** are joined into a multigene node, as shown in **figure 5.10**.

```
<entry id="135" name="undefined" type="group">  
  <graphics fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle"  
    x="574" y="657.75" width="55" height="68"/> <component id="56" /> <component id="132" />  
  <component id="133"/><component id="134" />  
</entry>
```

**Figure 5.10:** Connecting the genes in a multigene node using KEGG’s gene complex strategy to link genes together in a group. Notice how the name (green) is set to "undefined", and the type (orange) is set to "group". The height (light blue) represents all four nodes ( $4 \times 17 = 68$ ), and four components with IDs belonging to the original node as well as the three newly created ones

The entry that combines the genes in **figure 5.10** is different from the gene type entries. It has less information, and links to the IDs it consists of. The type is changed into ‘group’, and the height has been changed to represent all the genes. The original height of a gene node is set to 17 (presumably pixels), and with four genes in the group, the new height is set to 68. The multigene nodes are separated from the protein complexes by a broader edge around them, with additional white space.

Another process that takes place in FunHoP at this point is cleaning up the XML files by removing orthologs and metabolites that are included in the file but not connected to any of the nodes (this step is not shown in the figures in this thesis, but is described in more detail in **Paper 2**).

In the final stage of FunHoP, multigene nodes are collapsed once more into single-gene nodes. As the read counts for each gene are combined and differential expression is calculated from the new value, the nodes in the XML files must change their names to match the differential expression results. A collapsed node is shown in **figure 5.11**.

```
<entry id="56" name="hsa:218 hsa:220 hsa:221 hsa:222" type="gene" reaction="rn:R04996"  
link="http://www.kegg.jp/dbget-bin/www_bget?hsa:218+hsa:220+hsa:221+hsa:222" >  
  <graphics fgcolor="#000000" bgcolor="#BFFFBF" name="ALDH3A1-B4" type="rectangle"  
    x="574" y="639" width="46" height="17" />  
</entry>
```

**Figure 5.11:** In order to make the pathway less comprehensive, we collapse the nodes back into the original size. The collapsed node has the same ID as the original, and almost every other detail is the same. However, the graphics name has been changed to reflect the name of the first gene in the node, along with the total number of genes

The collapsed entry is a copy of the original, but the child name contains all the gene IDs. By simply counting the genes ( $X$ ), the graphics name can be changed to include ‘-BX’ at the end. There are four genes in **figure 5.11**, making the name of the collapsed node ‘ALDH3A1-B4’. As the calculations of read counts (**section 5.2.2**) create a dataset with differential expression on the node level with compatible names, the nodes can be styled in Cytoscape, as previously.

### 5.2.2 Using the read counts

Although it is possible to use parts of FunHoP without them, the read counts bring an interesting layer to the analysis. The gold standard of microarrays and fold change could show the comparison of the same gene in two different samples in the form of transformed p-values from differential expression, but read counts add a new dimension.

One of the advantages of using RNA-seq as measurement method for mRNA is the advantage you get from the read counts. Read counts provide a number of how many reads were mapped to a gene, and can be used as a measurement of how many mRNAs the sample contained for each gene. A gene with a higher number of reads indicates more mRNA was made for this gene than a gene with a lower read count. Although enzyme kinetics play an important role here, it is still possible to assume that in the case of homologs within a node, a higher number of read counts for a particular gene means the gene is actually upregulated. This is shown and discussed further in **Paper 2** and the **Discussion**. As we saw in **Paper 2**, adding a read count value to each gene provides some interesting views — in many cases, it shows a different picture than the fold changes.

The read counts were used in two different ways in FunHoP. The first was to show the read count for each individual gene in the pathways. Using the Cytoscape style shown in **figure 5.4**, each gene in the node could be coloured based on the read counts. This means it becomes possible to see the regulation for each gene. The second was the combination of the values into a single value for all genes in the node, which was then used to calculate differential expression at node level. The final step of changes to the XML files, as shown in **figure 5.11**, was particularly important here, as the new name, 'gene-BX', became the name of the node. This had to be taken into consideration when calculating the results from the differential expression. The problems related to this particular step are further explored in **section 7.5.1**.



## Summary of the results

All the results from the three studies are presented in greater detail in their respective articles. However, a short presentation of the main results from each paper will be presented in this section.

Our metabolic and transcriptomic profiling of prostate cancer tissue in **Paper 1** provided us with more knowledge about the differences between high and low stromal content in the tissue samples. We found that high stromal content has upregulated genes and metabolites linked to ECM remodelling and the immune system.

The main result in **Paper 2** was FunHoP, which takes metabolic pathway XMLs from KEGG and read counts from RNA-sequencing, and extend and combine these two into a new way of utilizing biological pathways analysis. The user can take a closer look at all genes within a node, both from differential expression and read counts point of views, and learn more about both the pathway and the gene expression. A new value based on all the read counts from all the genes within a node is calculated, and can be used for differential expression analysis on node-level. Our two case studies show plausible explanations to how histidine and GPC can be elevated in PCa.

In **Paper 3** we use parts of FunHoP combined with data on subcellular localisation. We used this data to divide the pathways into mitochondrial and non-mitochondrial editions, which in turn showed how mitochondrial paths are upregulated in PCa.





## Discussion

In this chapter, I will discuss some of the challenges we met in these studies. I will look at the common challenges that all three studies encountered, and then each paper will be discussed in turn.

The PCa transcriptome was used as a basis for the research in this thesis. Two different types of transcriptome data were used, namely RNA microarray and RNA-Seq. Different types of additional data were applied to the studies wherever useful and available, such as histopathology and recurrence-free survival in **Paper 1**, and metabolomics data in all three papers, both from our own samples and from the literature. The prostate was a source of transcriptome, both in the form of tissue data in **Paper 1** and **2** and from cell lines in (**Paper 3**). Metabolism and metabolic changes were the focus of all three papers.

### 7.1 Common challenges for all three papers

All three papers have transcriptome data as either a partial or major source of data. Four challenges arose regarding the data from RNA microarrays and RNA-Seq used in this thesis: The first is that we studied the templates (mRNA) while trying to understand the functions of the products (proteins and metabolites), but the relationship between RNA, proteins, and metabolites is not linear. The second challenge is that the transcriptome is a snapshot of the condition in the cells or tissue, and it does not show variations. It will not show you the situation in the cell the minute before or after the sample was taken, it can only show that exact moment, and that might not be representative. The third is that the data is not location-specific within the cell or tissue. We tried to find a partial solution for this problem in **Paper 3**. The final challenge is that, in many cases, the results will be an average over several cells and cell types.

For the first challenge, we look at the relationship between genes and phenotypes. Ideally, scientists would have a good understanding of the relationship between mRNAs and final proteins, and all the steps in this process, but this is not the case. Even if the goal is to understand the functions of the proteins and how they affect metabolism, knowledge of transcriptomics is also important. Learning how the cell changes gene expression when becoming cancerous is of interest, and this type of data is currently available. In our studies, we used mRNA levels, which can be measured, to come to a conclusion about protein levels and work towards a greater understanding of metabolite levels, which is the phenotype we want to comprehend.

Although these levels are all very much connected, multiple factors affect the relationship between them. This complexity becomes clear from a brief look at RNA regulation. Four levels of control exist between an mRNA molecule and the finished protein, namely RNA transport and localisation control, mRNA degradation control, translation control, and protein activity control (Alberts, 2008). The localisation of mRNA is regulated by many distinct mechanisms that require specific signals found in the 3' UTR of the mRNA (Lipshitz and Smibert, 2000; Alberts, 2008; Chin and Lécuyer, 2017). Localisation is an important factor, as proteins are needed in specific locations, as shown in **Paper 3**. The degradation of mRNA determines how long the mRNA is available for translation. Small RNA molecules like microRNAs (miRNA) or small interfering RNAs (siRNAs) affect both mRNA degradation and translation, as well as chromatin structure. Therefore, they have an impact on both transcription and translation rates (Valencia-Sanchez et al., 2006). Baudrimont et al. have found that the median half-life of an mRNA is around two minutes in yeast (Baudrimont et al., 2017), although this is a debated topic. In addition to being affected by miRNAs, translation can also be regulated by other translational repressors which bind to the 5' end and inhibit translation (Alberts, 2008).

Additionally, multiple forms of regulation exist on the protein itself. Activity by the finished proteins is also regulated by activation or cofactors, among others. Adding or removing phosphate is one way of activating a protein, and this has a major impact on the overall activity within a pathway. When all of this is considered, and knowing that this is just a brief overview of all the mechanisms occurring between mRNAs and protein functions, it becomes clear that in most projects one will not have enough data or knowledge to fully understand how levels of active protein may be reflected in the measured levels of mRNA. In **Paper 2**, we used known accumulated metabolites (histidine and glycerophosphocholine) as 'proof' of some form of upregulated activity. As the paths leading towards these metabolites contain upregulated genes, a plausible explanation is that the upregulated genes lead to more proteins that are, in turn, responsible for the upregulation of metabolites. It is possible that proteomics would indicate something different. This is a possible area for future study. Ideally, it would be possible to add data to all the omics layers and combine them to determine if our hypotheses are correct.

The second challenge is related to using 'snapshots' to describe a living, dynamic system, which is a common challenge in systems biology. When using transcriptomics data from microarrays or RNA sequencing, as in this thesis, the data is not time-specific. It does not say anything about changes or flux in the system, only what the status is at a specific moment, or more precisely, the average over many different points in time. In the case of **Paper 1 and 2**, which are based on prostatic tissue, this is still a challenge without an easy (or ethically approved) solution. It is possible that single-cell analysis and spatial data can provide more information for experiments like these. For the cell lines in **Paper 1**, it would be possible to take samples and measurements from the same culture over time. For instance, one could measure the metabolites in the growth medium to determine how they change over time (Halldorsson et al., 2017). However, this was not done and, to our knowledge, these types of data on PCa are not available for common use.

The third challenge is related to how the transcriptomics analysis is performed. Traditionally this type of analysis involves homogenizing the tissue and extracting mRNA from this mixture (Bertilsson et al., 2011). New technology on spatial transcriptomics on tissue biopsies (Ståhl et al., 2016; Berglund et al., 2018; Friedrich and Sonnhammer, 2020) will make it possible to learn more about where in the tissue the different pathways are found to be active. Single-cell sequencing (Stegle et al., 2015) also makes it possible to look at transcriptomics within a specific subtype of tissue. These techniques will add more depth to the analysis and understanding of metabolism. Given the emergence of technologies such as spatial proteomics with MALDI imaging (Cornett et al., 2007), combining these and other techniques will lead to even more knowledge. However, this was not an option when the work for this thesis was done. For future research, studying these types of data could be an option.

The fourth challenge is closely related to the third, and shares a similar solution. In traditional methods, the sample contains multiple different cells, even if they are from similar areas of the tumour. Using a mixture of cells means the final numbers used to define regulation of a gene will be based on an average from all these potentially different cells. If single cell transcriptomics were an option, the results could vary more between the cell types and perhaps reveal more interesting features. Along with spatial transcriptomics and proteomics, this holds potential for future research.

From this point on, I will look into each of the papers and their individual challenges.

## 7.2 Paper 1

This study was one of many prostate-related studies from the MR Cancer Group, using techniques and data they had collected over many years. Having these different types of data made it possible to combine them within the same study, which is often lacking in other projects. My role involved analysing differential gene expression and finding relevant GO terms based on different groups of reacting stroma. The study used a different technique for gene expression than the others mentioned in this thesis, as it was based on microarrays and not on RNA-Seq. Using a microarray means that only transformed p-values were available and not read counts, which makes it impossible to use all the steps of FunHoP. When finding relevant GO terms, some terms with equivalent pathways in KEGG appeared, such as the 04660 T cell receptor signalling pathway. The GO terms were found in three categories, namely the immune system, cell signalling, and the extracellular matrix. KEGG includes sections on both the immune system and cell signalling. It would be possible to use the first part of FunHoP, which involves expanding pathways and examining all genes within the pathway, for the relevant paths found based on GO terms. However, FunHoP was not fully developed at the time this study was performed and could not be used.

One of the main arguments in **Paper 2** is that RNA-Seq is a better technique than microarrays, as it creates read counts which not only look at the differential expression but also at the actual read count number. This argument still stands. However, having different types of data, including expressions using microarrays from the same tissue, was such an advantage that other limitations of the microarray technique

were acceptable in this study.

**Paper 1** is a more biologically focused paper than the other two, which deal more with the technical aspects of development and analytical methods. Instead, this research tends to use results from established methods for analysis, providing a valuable introduction to the complexity of studying PCa. It provides a good starting point for the continued creation of new analysis tools for different pathways, metabolites, and genes that play an important role in PCa.

### 7.3 Paper 2

In **Paper 2**, we made a few necessary assumptions that went against certain widely accepted techniques and theories. The first assumption was that we used read counts from RNA-Seq to colour the nodes directly. The reason was that using only the p-value could be misleading. If a gene is considered as having five read counts in condition X and 10 in condition Y, this means it has twice as many reads in condition Y. However, there are still only a few read counts, even if the relative difference between them is large. If another gene has 250 read counts in condition X and 350 in condition Y, the relative difference between them is smaller, and therefore, the p-value between them is larger. This means that the p-value indicates a larger difference between the genes in condition X. If these two genes are considered functional homologs and can be expected to have a similar kinetic rate, the gene with >250 read counts will be expressed more in both samples than the first gene, despite the p-values suggesting the first gene is more important. Of course, in reality, more than one sample exists of each condition, and therefore, p-values are affected by the similarity of read counts between samples of the same condition. However, the principle remains the same.

This leads to the second assumption, namely a similar rate for proteins within a node. If the genes from the above example were found within the same node, we assumed they were so closely related that they could catalyse the same reaction and would do so at a similar rate. This is a simplification, as there is not enough conversion rate data to confirm or disprove this. However, since we had to assume that the proteins were similar enough to perform the same task, we decided it would be reasonable to assume they would have the same rate. For the two genes in question, this is of importance. Even if a doubling occurs between the conditions in gene 1, the total number of read counts in gene 2 is much higher. If these two genes have the same rate, the gene with 250 or 350 copies will be able to catalyse a great deal more than the gene with 5 or 10 copies. Therefore, it can be said that a dominant gene can be found within a node.

#### 7.3.1 Colours in Cytoscape

An important comment from one of the reviewers concerned the choice of colour scale for the transformed p-values. It was pointed out that the colour scale we used would be problematic for the colour blind, and that it was more common to use this range in the opposite direction to ours. Our colour scales (**figure 5.3**) range from a faint yellow at zero to a dark green (value 600) via bright green in the positive direction, and to a black (-1200) via red and purple (-600) in the negative direction. For most of the genes, this would mean that an upregulated gene would be bright green and a downregulated gene would be bright

red, or fainter variations of these colours. As the reviewer pointed out, red-green is the most common type of colour blindness. Colour blindness itself is relatively common, affecting approximately 8% of males and 0.4% of females (Birch, 2012). Our choice of colours could therefore become a problem for those who are colour blind, as it would be harder to distinguish between the ranges. We have not found any evidence to prove the reviewer's comment about the colour range being the opposite of the standard in the field, which supposedly uses red to indicate upregulation and green for downregulation. However, we do acknowledge that a majority of studies seem to have used the opposite colour range, and to ensure accessibility for all readers we should have changed our colour scale.

### 7.3.2 What the pathways are not telling us

A major limitation in defining a path from A to E via B, C, and D is that the path is rarely linear and isolated. B might be the substrate for creating compound K, creating a side branch, while C can also be created via the path of L, which creates another branch. These side branches need to be investigated to make sure no important information is sidelined. This can be seen in the studies outlined in **Paper 2**, where the measured levels of histidine and glycerophosphocholine were used as precursors for trying to explain the regulations and flows in the pathways. Although the two main pathways were properly shown in the case studies, we also considered the potential other paths in which these two metabolites participated to make sure our suggestions were plausible. As none of the other pathways had any particular regulated genes that could be responsible for the elevated metabolites, our interpretation of the results from the case studies were more likely to be plausible.

Another limitation in the pathways from KEGG is the lack of additional information about for instance energy transfer. When looking at pathways in a textbook (or at **figures 4.6, 4.7, or 4.8**), it is clear how much more detailed these pathways are. They include the usage and creation of ATP/NADH, H<sub>2</sub>O, and CO<sub>2</sub>, among others, and it would be helpful to see whether an area with many up- or downregulated genes would be affected or would affect the production or usage of these additional compounds.

## 7.4 Paper 3

In this study, an alternative usage of FunHoP was found. Another level of detail was added to the ongoing discussion on multigene nodes, and finding a solution to how we could use these details was fascinating. **Paper 2** showed that the concept of multigene nodes was somewhat controversial, especially when we compared the enzyme kinetics of the genes within the node. However, when we started splitting the multigene nodes into mitochondrial and non-mitochondrial parts, the reactions from colleagues and other scientists were far more positive. Linking our findings to the biological literature inspired some interesting thoughts. Despite **Paper 3** not being peer-reviewed and published yet, and hence there being no comments from reviewers to discuss, there are some points that can be examined.

We know that, in the normal prostate, the Krebs cycle is stopped at early stages due to zinc inhibition of aconitase (ACO2). This can also be seen in the fact that citrate is the major anion in the sperm fluids. In

our analysis, we saw that the Krebs cycle is, for the most part, upregulated. This would indicate that the uptake transporter for zinc, ZIP1, was downregulated or inhibited in some way. *SLC39A1*, the gene that translates into ZIP1, is upregulated in the prostate cancer cell lines, as shown in **table 7.1**.

**Table 7.1:** Comparing the differential expression of two genes; *ACO2* and *SLC39A1*. *ACO2* has a higher fold change and a lower p-value than *SLC39A1*

<b>Gene</b>	<b>Fold change</b>	<b>p-value</b>
<i>ACO2</i>	0.989	1.736e-09
<i>SLC39A1</i>	0.907	1.508e-05

This shows that *ACO2* has a slightly higher fold change than *SLC39A1*, meaning there is slightly more of it. The p-value is also lower for *ACO2*. If these were to have a one-to-one relationship in regard to inhibition (one zinc uptake transporter takes up one zinc atom, which inhibits one aconitase), this could be a possible explanation. It would be plausible to say that in cancer the levels of zinc uptake transporters decrease, so there is a slightly lower amount of zinc to inhibit aconitase. Without full inhibition, the Krebs cycle can continue, as seen in PCa. However, this theory relies on multiple assumptions, including directly comparing the kinetics of two very different enzymes.

#### **7.4.1 Hard coding the XML files**

In order to create two sets from the same pathway, one mitochondrial and one non-mitochondrial, the XML file was physically hard coded to represent both. This was done by going through all the genes within the pathway, determining location, and then removing either mitochondrial or non-mitochondrial genes, depending on which file. This included removing some compounds and links, and the process was cumbersome in the case of pathways that lacked a certain location from the beginning. The chance of human error is large in detailed work such as dividing the XMLs, and it would have been preferable to automate this process.

Dividing the XML files could also have created problems in cases where the three location sources disagreed. If one said 'mitochondrial', the other 'non-mitochondrial', and the third 'unknown', it would be hard to determine the actual location. Our decision to use the predictions from the Bologna Unified Subcellular Component Annotator (BUSCA) as a casting vote meant making some decisions that would have given a different result if we had chosen just one of the sources. For instance, if the BUSCA casting vote was 'mitochondrial' for a single gene with no homologs, and the removal of the non-mitochondrial

gene made the whole pathway fall apart, this would be considered a much larger consequence than if the gene was part of a multigene node. Additionally, all links had to be removed for single genes.

## 7.5 FunHoP

FunHoP changes the KEGG pathways and adds more information. It is simple to use, the results are easy to understand (at least to the degree that pathways can be easily interpreted), and it brings added value to the field. As seen in **Developing FunHoP**, the process where FunHoP expands nodes is fairly simple and gives the user the opportunity to edit the pathway XML files to their own liking. However, FunHoP has not yet reached its full potential, and it could be improved by a variety of additions. The first step is to turn FunHoP into a functional Python package. This would make it even easier to use FunHoP and make it more easily available to users. Making FunHoP into a Python package would also include removing the steps that are currently hard-coded, and it would be possible to include more databases. Adjusting coordinates, adding example data, and looking at alternative uses for FunHoP are also of interest. This section will end with a brief look at possible fundamental changes to FunHoP.

### 7.5.1 Removing the hard coding

These hard-coded sections are found in the section where multigene nodes are created, as shown in **figure 5.11**. Each multigene node is named ‘gene-BX’, where X is the number of genes within the multigene node. This means that for two nodes with identical first genes and the same number of genes within, the name would be the same. This was checked manually and the second node was renamed ‘Gene-BX-2’. No cases were found of three nodes with a similar name, and only eight cases of duplicate names were discovered, so this was relatively easy to fix manually. It is crucial that the node that becomes the second in the read counts list has the correct corresponding nodes in the XML files. In case of a mismatch, the colouring might be wrong. Ideally, this should be solved automatically. This is one of the improvements that are necessary before publishing FunHoP as a Python package.

### 7.5.2 Including more databases

FunHoP currently only works for XML files from KEGG, which limits its usage. Ideally, users should be able to choose which database they want to use, and FunHoP would automatically adjust to their choice. As the databases use different methods for creating their pathways, this presents a difficult, although not impossible, challenge. Making specific versions of FunHoP for each of the major databases (such as Reactome, WikiPathways, and Panther) and letting the user choose based on the type of input would be possible. Another possibility is to use one specific database format as standard and provide access to other databases through conversion to this format. If it were possible to convert other formats into one that FunHoP can recognize and work with, this would be a better and more stable solution than creating a different version of FunHoP for each individual database.



### 7.5.3 Coordinates

A major difficulty in the current version of FunHoP is the coordinates of the nodes. Each node in the original XML files has both an x- and a y-coordinate, placing the node within the graphical display of the network. When nodes are expanded, they become larger. The expanded node retains the coordinates of the original node, meaning that the top of the node is in the same location as before, but the bottom of the node might be expanded downwards, sometimes creating conflict with the nodes, metabolites, or links that were placed below the original node. For nodes that are only expanded to include two genes, this is usually not an issue, but with any more than two this tends to become a problem, which only grows with an increase in the number of genes. In **Paper 2**, some of the largest multigene nodes in the glycerophosphocholine metabolism case study have 21 genes. These come into conflict with all other nodes, making the network hard to work with.

The interim solution to the coordinate problem for an expanded node was to keep the x-coordinates and change the y-coordinates for all nodes below, based on the number of genes in the multigene nodes. Changing the y-coordinates stretches the entire network vertically and means that the user must subsequently make a slight adjustment, but this stretching makes adjustment easier, as it separates the nodes and links from each other.

This solution is not perfect. It stretches the network into a shape that can be easily adjusted, but manual adjustment by the user is still required. Ideally, all components of the network would have the ‘perfect’ coordinates, similar to the original files. We still do not know how KEGG creates its XML files, or whether any apps exist that can redraw a network from an XML file into something more useful than the coordinates in the XML, but to my knowledge, this is complicated. Finding the current solution took a great deal of trial and error, and it does work, albeit imperfectly. If FunHoP were to become a Cytoscape app, this would be one of the main areas for improvement.

### 7.5.4 Example datasets

One of the biggest challenges for potential FunHoP users at present is the lack of example data on which to test it. Ideally, a simple dataset could ensure that FunHoP is working as intended on local machines, and the user could see how it was intended to work. A small set of gene expression data, the gene ID list from KEGG, and a number of KEGG XML files could be used, in addition to output files and finished figures, to show how FunHoP functions. This is one of the main priorities for future improvements.

### 7.5.5 Alternative use of FunHoP

FunHoP can be used on many levels, and it is not necessary to go through all the steps in order to gain more information. For instance, it would be possible to use only the read counts to find interesting pathways in DAVID (Dennis et al., 2003) or Enrichr (Chen et al., 2013). This might give a different view than doing the same with transformed p-values. Another example is shown in **Paper 3**, where only the first part of

FunHoP is used to expand the networks, and these are then used without read counts or collapsed nodes and p-values on the node level. This only shows the expanded pathways and neither of the other options (original or collapsed). Using FunHoP without the read count section means that all types of data that connect a value to a gene or protein can be visualized in Cytoscape. Therefore, proteomics measurements could easily be used as a basis for colouring KEGG pathways.

If more data became available, applying them to the networks would be possible. Cytoscape has features for changing both nodes and links, and if data on enzymatic rates were available, it could be added to the networks by changing the thickness of links or the borders of the nodes.

### **7.5.6 How FunHoP could have been created differently**

With regard to writing FunHoP, certain things could have been done differently. FunHoP is written in Python, whereas Cytoscape apps are usually written in Java. It would have been an advantage if FunHoP was written in Java, as it could be converted into a Cytoscape app much more easily. However, due to a lack of both Java and app design experience, this was not an option. In its current form, FunHoP can still be converted into a Python package.

In order to create the best version of FunHoP, a more active testing regime should have been in place from the start. A number of tests were conducted during the creation process, but without truly satisfactory structure. A complete testing system would also be important in allowing others to use FunHoP to its fullest extent.

If the exact nature of FunHoP had been determined before the work started, many things could have been done differently. However, as the final product was quite different from the initial concept, and it was made by non-developers, several of these things were simply unknown unknowns. The important thing is to focus on the lessons that have been learnt and the fun I had during the process.

## Future work

In terms of future work, several paths can be followed. FunHoP in its current state is far from fully developed and although it is ready to use, it has not reached its full potential. Adjusting coordinates, removing hardcoding, adding example data, and including other databases are among the many possibilities for improving the method. The top priority could be turning FunHoP into a Python package, which would make it easier to use. FunHoP could be greatly improved by collaborating with a group with more experience in software development, specifically a group that has worked on Cytoscape and app development. This would be extremely helpful in improving some of the more challenging aspects of FunHoP, such as the coordinates.

This project was initially intended to be about the spatial distribution of genes and metabolites in PCa tissue, and the first few months we were focusing on this topic. However, the financial planning did not go through, and the plans had to change. This research is now being conducted by our partners in May-Brit Tessem's group (Tessem, 2022), and their work with MALDI imaging will hopefully yield more location data in the future. Combined with spatial transcriptomics, spatial metabolomics is a promising topic for further study. This could take the work described in **Paper 3** further and also combine different types of data from the same tissue samples. FunHoP would be useful in this endeavour.

More studies on enzyme kinetics could also be of interest. Our assumptions from **Paper 2** would need a great deal of work to be considered valid. Yet even if they were found to be invalid, the new information on kinetics could be incorporated into the networks to provide even more information.

With time and resources, improving the manual hard-coding from **Paper 3** would be an interesting project. Getting FunHoP to identify the location of the gene products and modifying the XML files accordingly would require intensive research and development, but at the same time be a very fun project to take on. This project could also be extended to look at more specific subcellular localisations, perhaps doing more than two categories for instance, and see if we could separate the pathways even further. All in all, there are many exciting topics waiting for someone to pick them up!



## Conclusion

In the three projects in this thesis we have studied the metabolism of prostate cancer, from different angles and with different approaches. All three studies have brought us new insight, and the second and the third have also given us new tools for further studies.

In the first study we looked at differences in the stromal environment in PCa. Our 108 samples were evaluated using histopathology, and out of these, metabolites were measured in 85 and gene expression was measured in 78. Multivariate metabolomics and transcriptomics were used to compare the samples with high levels of reactive stroma versus samples with low levels of reactive stroma. We found that the samples with high levels of reactive stroma had upregulated genes and metabolites connected to functions in the immune system and extracellular matrix. This study was an exciting introduction to prostate cancer, biological and statistical methods, and metabolism.

In the second study we looked into pathway analysis. Visualisation is an important tool when working with big data, and a well-known method is to use biological pathways from KEGG in Cytoscape. Unfortunately, both the KEGG webpage and KEGG in Cytoscape (via the KEGGscape app) shows only one gene in each node, regardless of the number of homologs capable to participate enzymatically in the reaction. Our new method FunHoP fixes this by expanding each node to show all genes, making it possible for the user to look at and style the nodes based on their own data. If the user has data from RNA-seq, FunHoP can use the read counts to show the number for each genes, to see if a node has a dominating gene, and finally it collapses all the genes in each node into a new number, which can be used in differential expression. Our two case studies showed how FunHoP could bring new biological insight, as well as improving the visual understanding for the viewer.

In the final study we took FunHoP a step further, bringing yet another layer of insight to the biological picture. By adding localisation data from both experimental and prediction data, we could differentiate between mitochondrial and non-mitochondrial processes, and those that are a mixture. By altering the KEGG pathways based on our findings, we could compare differential gene expression from cell lines, based on localisation. We saw that mitochondrial processes are generally upregulated in PCa, and that localisation data brought a new and interesting aspect to the pathway analysis.

---

Overall, this thesis adds a few more pieces to the prostate cancer metabolism puzzle. With many pieces left, it is my hope that my new tool FunHoP and the approach of using localisation data can be further adapted and developed into finding the remaining ones.

# Bibliography

- Alberts, B. (2008). *Molecular Biology of the Cell*, chapter 6, 7. Garland Science, 5 edition.
- Almaas, E. (2013). *Compendium for TBT4165 Systems biology and biological networks*. NTNU.
- Andersen, M. K., Rise, K., Giskeødegård, G. F., Richardsen, E., Bertilsson, H., Størkersen, Ø., Bathen, T. F., Rye, M., and Tessem, M.-B. (2018). Integrative metabolic and transcriptomic profiling of prostate cancer tissue containing reactive stroma. *Scientific reports*, 8(1):1–11.
- Ayala, G., Tuxhorn, J. A., Wheeler, T. M., Frolov, A., Scardino, P. T., Ohori, M., Wheeler, M., Spitler, J., and Rowley, D. R. (2003). Reactive stroma as a predictor of biochemical-free recurrence in prostate cancer. *Clinical Cancer Research*, 9(13):4792–4801.
- Barron, D. A. and Rowley, D. R. (2012). The reactive stroma microenvironment and prostate cancer progression. *Endocrine-related cancer*, 19(6):R187–R204.
- Baudrimont, A., Voegeli, S., Vilorio, E. C., Stritt, F., Lenon, M., Wada, T., Jaquet, V., and Becskei, A. (2017). Multiplexed gene control reveals rapid mRNA turnover. *Science advances*, 3(7):e1700006.
- Bender, M., Killheffer, J., and Cohen, S. (1973). Chymotrypsin. *CRC critical reviews in biochemistry*, 1(2):149–199.
- Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidis, A., Vickovic, S., Larsson, L., et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature communications*, 9(1):1–13.
- Berry, S. J., Coffey, D. S., Walsh, P. C., and Ewing, L. L. (1984). The development of human benign prostatic hyperplasia with age. *The Journal of urology*, 132(3):474–479.
- Bertilsson, H., Angelsen, A., Viset, T., Skogseth, H., Tessem, M.-B., and Halgunset, J. (2011). A new method to provide a fresh frozen prostate slice suitable for gene expression study and mr spectroscopy. *The Prostate*, 71(5):461–469.
- Birch, J. (2012). Worldwide prevalence of red-green color deficiency. *JOSA A*, 29(3):313–320.
- Brawer, M. K. (2005). Prostatic intraepithelial neoplasia: an overview. *Reviews in urology*, 7(Suppl 3):S11.

- 
- Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Gonzalez, G. A. P., Aurich, M. K., et al. (2018). Recon3d enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36(3):272–281.
- Bumgarner, R. (2013). Overview of dna microarrays: types, applications, and their future. *Current protocols in molecular biology*, 101(1):22–1.
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):1–14.
- Chen, G. and Geiling, E. (1946). Glycolysis in trypanosoma equiperdum. *Proceedings of the Society for Experimental Biology and Medicine*, 63(2):486–487.
- Chin, A. and Lécuyer, E. (2017). Rna localization: Making its way to the center stage. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861(11):2956–2970.
- Chokkathukalam, A., Kim, D., Barrett, M., Breitling, R., and Creek, D. (2014). Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. *Bioanalysis*, 6(4):511–524.
- Cornett, D. S., Reyzer, M. L., Chaurand, P., and Caprioli, R. M. (2007). Maldi imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature methods*, 4(10):828–833.
- Costello, L., Feng, P., Milon, B., Tan, M., and Franklin, R. (2004). Role of zinc in the pathogenesis and treatment of prostate cancer: critical issues to resolve. *Prostate cancer and prostatic diseases*, 7(2):111–117.
- Costello, L. and Franklin, R. (2009). Prostatic fluid electrolyte composition for the screening of prostate cancer: a potential solution to a major problem. *Prostate cancer and prostatic diseases*, 12(1):17–24.
- Costello, L. C. and Franklin, R. B. (2016). A comprehensive review of the role of zinc in normal prostate function and metabolism; and its implications in prostate cancer. *Archives of biochemistry and biophysics*, 611:100–112.
- Cuezva, J. M., Krajewska, M., de Heredia, M. L., Krajewski, S., Santamaría, G., Kim, H., Zapata, J. M., Marusawa, H., Chamorro, M., and Reed, J. C. (2002). The bioenergetic signature of cancer: a marker of tumor progression. *Cancer research*, 62(22):6674–6681.
- Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). David: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(9):1–11.
- Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cdna microarrays. *Nature genetics*, 21(1):10–14.
- Eidelman, E., Twum-Ampofo, J., Ansari, J., and Siddiqui, M. M. (2017). The metabolic phenotype of prostate cancer. *Frontiers in oncology*, 7:131.



- 
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655.
- Feng, P., Li, T.-L., Guan, Z.-X., Franklin, R. B., and Costello, L. C. (2002). Direct effect of zinc on mitochondrial apoptosis in prostate cells. *The Prostate*, 52(4):311–318.
- Fernie, A. R., Carrari, F., and Sweetlove, L. J. (2004). Respiratory metabolism: glycolysis, the tca cycle and mitochondrial electron transport. *Curr Opin Plant Biol*, 7(3):254–261.
- Feun, L., You, M., Wu, C., Kuo, M., Wangpaichitr, M., Spector, S., and Savaraj, N. (2008). Arginine deprivation as a targeted therapy for cancer. *Current pharmaceutical design*, 14(11):1049–1057.
- Fitzpatrick, J. M. (2006). The natural history of benign prostatic hyperplasia. *BJU international*, 97:3–6.
- Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. (2003). Genome-scale reconstruction of the *saccharomyces cerevisiae* metabolic network. *Genome research*, 13(2):244–253.
- Fowler, S., Roush, R., and Wise, J. (2013). Transcription. <https://openstax.org/books/concepts-biology/pages/9-3-transcription>.
- Franz, M.-C., Anderle, P., Bürzle, M., Suzuki, Y., Freeman, M., Hediger, M., and Kovacs, G. (2013). Zinc transporters in prostate cancer. *Molecular aspects of medicine*, 34(2-3):735–741.
- Friedrich, S. and Sonnhhammer, E. L. (2020). Fusion transcript detection using spatial transcriptomics. *BMC medical genomics*, 13(1):1–11.
- Giskeødegård, G. F., Bertilsson, H., Selnæs, K. M., Wright, A. J., Bathen, T. F., Viset, T., Halgunset, J., Angelsen, A., Gribbestad, I. S., and Tessem, M.-B. (2013). Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness. *PLoS one*, 8(4):e62375.
- Gordon Betts, J., Young, K., Wise, J. A., Johnson, E., Poe, B., Kruse, D., Korol, O., Johnson, J., Womble, M., and DeSaix, P. (2022). Glycolysis. <https://openstax.org/books/anatomy-and-physiology-2e/pages/24-2-carbohydrate-metabolism>.
- Halldorsson, S., Rohatgi, N., Magnusdottir, M., Choudhary, K. S., Gudjonsson, T., Knutsen, E., Barkovskaya, A., Hilmarsdottir, B., Perander, M., Mælandsmo, G. M., et al. (2017). Metabolic re-wiring of isogenic breast epithelial cell lines following epithelial to mesenchymal transition. *Cancer letters*, 396:117–129.
- Horowitz, N. (1965). The evolution of biochemical syntheses—retrospect and prospect. In *Evolving genes and proteins*, pages 15–23. Elsevier.
- Horowitz, N. H. (1945). On the evolution of biochemical syntheses. *Proceedings of the National Academy of Sciences of the United States of America*, 31(6):153.
- Isidoro, A., Martínez, M., Fernández, P. L., Ortega, A. D., Santamaría, G., Chamorro, M., Reed, J. C., and Cuezva, J. M. (2004). Alteration of the bioenergetic phenotype of mitochondria is a hallmark of breast, gastric, lung and oesophageal cancer. *Biochemical Journal*, 378(1):17–20.
-

- 
- Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual review of microbiology*, 30(1):409–425.
- Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951.
- Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021). Kegg: integrating viruses and cellular organisms. *Nucleic acids research*, 49(D1):D545–D551.
- Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.
- Kavanagh, J. (1985). Sodium, potassium, calcium, magnesium, zinc, citrate and chloride content of human prostatic and seminal fluid. *Reproduction*, 75(1):35–41.
- KEGG (2022a). Sphingophospholipid metabolism. <https://www.genome.jp/kegg-bin/show%5Fpathway?hsa00600>.
- KEGG (2022b). Tyrosine metabolism. <https://www.genome.jp/kegg/pathway/map/hsa00350.html>.
- Kim, E. H., Larson, J. A., and Andriole, G. L. (2016). Management of benign prostatic hyperplasia. *Annual review of medicine*, 67:137–151.
- Kim, R. H., Coates, J. M., Bowles, T. L., McNerney, G. P., Sutcliffe, J., Jung, J. U., Gandour-Edwards, R., Chuang, F. Y., Bold, R. J., and Kung, H.-J. (2009). Arginine deiminase as a novel therapy for prostate cancer induces autophagy and caspase-independent apoptosis. *Cancer research*, 69(2):700–708.
- Kosova, F., Temeltaş, G., Ari, Z., and Lekili, M. (2014). Possible relations between oxidative damage and apoptosis in benign prostate hyperplasia and prostate cancer patients. *Tumor Biology*, 35(5):4295–4299.
- Kraut, J. (1977). Serine proteases: structure and mechanism of catalysis. *Annual review of biochemistry*, 46(1):331–358.
- Kreftforeningen (2022). <https://kreftforeningen.no/om-kreft/kreftformer/prostatakreft/>.
- Lewis, M. T., Kasper, J. D., Bazil, J. N., Frisbee, J. C., and Wiseman, R. W. (2019). Quantification of mitochondrial oxidative phosphorylation in metabolic disease: Application to type 2 diabetes. *International journal of molecular sciences*, 20(21):5271.
- Lipshitz, H. D. and Smibert, C. A. (2000). Mechanisms of rna localization and translational regulation. *Current opinion in genetics & development*, 10(5):476–488.
- Lundberg, E. and Borner, G. H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology*, 20(5):285–302.
- Marker, P. C., Donjacour, A. A., Dahiya, R., and Cunha, G. R. (2003). Hormonal, cellular, and molecular control of prostatic development. *Developmental biology*, 253(2):165–174.

- 
- Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., A. Miller, R., Digles, D., Lopes, E. N., Ehrhart, F., et al. (2021). Wikipathways: connecting communities. *Nucleic Acids Research*, 49(D1):D613–D621.
- Melendez-Hevia, E. and Siverio, J. M. (1984). Studies on glycolysis in vitro: role of glucose phosphorylation and phosphofructokinase activity on total velocity. *International Journal of Biochemistry*, 16(5):469–476.
- Moestue, S., Sitter, B., Frost Bathen, T., Tessem, M.-B., and Susann Gribbestad, I. (2011). Hr mas mr spectroscopy in metabolic characterization of cancer. *Current topics in medicinal chemistry*, 11(1):2–26.
- Moncada, S., Higgs, E. A., and Colombo, S. L. (2012). Fulfilling the metabolic requirements for cell proliferation. *Biochemical Journal*, 446(1):1–7.
- Narayanese (2008). The tca cycle. [https://en.wikipedia.org/wiki/File:Citric\\_acid\\_cycle\\_with\\_aconitate\\_2.svg](https://en.wikipedia.org/wiki/File:Citric_acid_cycle_with_aconitate_2.svg).
- Nelson, D. and Cox, M. (2008). *Lehninger principles of biochemistry*, chapter 3, 4, 7. W.H Freeman and company, 5 edition.
- Nishida, K., Ono, K., Kanaya, S., and Takahashi, K. (2014). Keggscape: a cytoscape app for pathway data integration. *F1000Research*, 3(144).
- Orre, L. M., Vesterlund, M., Pan, Y., Arslan, T., Zhu, Y., Woodbridge, A. F., Frings, O., Fredlund, E., and Lehtiö, J. (2019). Subcellbarcode: proteome-wide mapping of protein localization and relocation. *Molecular cell*, 73(1):166–182.
- Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3):245–248.
- Ottaway, J. and Mowbray, J. (1977). The role of compartmentation in the control of glycolysis. *Current topics in cellular regulation*, 12:107–208.
- Packer, J. R. and Maitland, N. J. (2016). The molecular and cellular origin of human prostate cancer. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1863(6):1238–1260.
- Pan, T., Gao, L., Wu, G., Shen, G., Xie, S., Wen, H., Yang, J., Zhou, Y., Tu, Z., and Qian, W. (2015). Elevated expression of glutaminase confers glucose utilization via glutaminolysis in prostate cancer. *Biochemical and biophysical research communications*, 456(1):452–458.
- Parker, N., Schneegurt, M., Tu, A.-H. T., Lister, P., and Forster, B. M. (2016). Microbiology. <https://openstax.org/books/microbiology/pages/3-4-unique-characteristics-of-eukaryotic-cells>.
- Qiu, F., Huang, J., and Sui, M. (2015). Targeting arginine metabolism pathway to treat arginine-dependent cancers. *Cancer letters*, 364(1):1–7.
- Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B. (2014). *Campbell biology*, chapter 7, 8, 17, 18. Pearson.

- 
- Rise, K., Tessem, M.-B., Drabløs, F., and Rye, M. B. (2021). Funhop: Enhanced visualization and analysis of functionally homologous proteins in complex metabolic networks. *Genomics, Proteomics & Bioinformatics*.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.
- Savojardo, C., Martelli, P. L., Fariselli, P., Profti, G., and Casadio, R. (2018). Busca: an integrative web server to predict subcellular localization of proteins. *Nucleic acids research*, 46(W1):W459–W466.
- Schilling, C. H. and Palsson, B. O. (1998). The underlying pathway structure of biochemical reaction networks. *Proceedings of the National Academy of Sciences*, 95(8):4193–4198.
- Schuster, S., Fell, D. A., and Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(3):326–332.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.
- Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82.
- Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145.
- Tessem, M.-B. (2022). Prostromics group. <https://www.ntnu.edu/isb/prostromics#/view/about>.
- Tuxhorn, J. A., Ayala, G. E., Smith, M. J., Smith, V. C., Dang, T. D., and Rowley, D. R. (2002). Reactive stroma in human prostate cancer: induction of myofibroblast phenotype and extracellular matrix remodeling. *Clinical Cancer Research*, 8(9):2912–2923.
- Valencia-Sanchez, M. A., Liu, J., Hannon, G. J., and Parker, R. (2006). Control of translation and mrna degradation by mirnas and sirnas. *Genes & development*, 20(5):515–524.
- Villar-Palasi, C. and Lerner, J. (1970). Glycogen metabolism and glycolytic enzymes. *Annual review of biochemistry*, 39(1):639–672.
- Wang, Q., Hardie, R.-A., Hoy, A. J., Van Geldermalsen, M., Gao, D., Fazli, L., Sadowski, M. C., Balaban, S., Schreuder, M., Nagarajah, R., et al. (2015). Targeting asct2-mediated glutamine uptake blocks prostate cancer growth and tumour development. *The Journal of pathology*, 236(3):278–289.
- Wang, Q., Tiffen, J., Bailey, C. G., Lehman, M. L., Ritchie, W., Fazli, L., Metierre, C., Feng, Y., Li, E., Gleave, M., et al. (2013). Targeting amino acid transport in metastatic castration-resistant prostate

- 
- cancer: effects on cell cycle, cell growth, and tumor development. *Journal of the National Cancer Institute*, 105(19):1463–1473.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.
- Warburg, O. (1956). On the origin of cancer cells. *Science*, 123(3191):309–314.
- Wu, R., Sessa, G., and Hamerman, D. (1964). Pi transport and glycolysis in leucocytes and platelets. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 93(3):614–624.
- Yarmush, M. L. and Berthiaume, F. (1997). Metabolic engineering and human disease. *Nature biotechnology*, 15(6):525–528.
- Yčas, M. (1974). On earlier states of the biochemical system. *Journal of Theoretical Biology*, 44(1):145–160.
- Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59.
- Zadra, G., Photopoulos, C., and Loda, M. (2013). The fat side of prostate cancer. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1831(10):1518–1532.
- Zephyris (2022). Metabolism. [https://commons.wikimedia.org/wiki/File:Metabolism\\_790px.png](https://commons.wikimedia.org/wiki/File:Metabolism_790px.png).
- Zhang, Y., Lu, R., Qin, C., and Nie, G. (2020). Precision nutritional regulation and aquaculture. *Aquaculture Reports*, 18:100496.

---

Appendix **A**

## Appendix: The papers

SCIENTIFIC REPORTS **OPEN** Integrative metabolic and transcriptomic profiling of prostate cancer tissue containing reactive stroma

Received: 5 June 2018  
Accepted: 10 September 2018  
Published online: 24 September 2018

Maria K. Andersen<sup>1</sup>, Kjersti Rise<sup>2</sup>, Guro F. Giskeødegård<sup>1</sup>, Elin Richardsen<sup>3,4</sup>, Helena Bertilsson<sup>2,5</sup>, Øystein Størkersen<sup>6</sup>, Tone F. Bathen<sup>1</sup>, Morten Rye<sup>2,7</sup> & May-Britt Tessem<sup>1</sup>

Reactive stroma is a tissue feature commonly observed in the tumor microenvironment of prostate cancer and has previously been associated with more aggressive tumors. The aim of this study was to detect differentially expressed genes and metabolites according to reactive stroma content measured on the exact same prostate cancer tissue sample. Reactive stroma was evaluated using histopathology from 108 fresh frozen prostate cancer samples gathered from 43 patients after prostatectomy (Biobank1). A subset of the samples was analyzed both for metabolic (n = 85) and transcriptomic alterations (n = 78) using high resolution magic angle spinning magnetic resonance spectroscopy (HR-MAS MRS) and RNA microarray, respectively. Recurrence-free survival was assessed in patients with clinical follow-up of minimum five years (n = 38) using biochemical recurrence (BCR) as endpoint. Multivariate metabolomics and gene expression analysis compared low ( $\leq 15\%$ ) against high reactive stroma content ( $\geq 16\%$ ). High reactive stroma content was associated with BCR in prostate cancer patients even when accounting for the influence of Grade Group (Cox hazard proportional analysis,  $p = 0.013$ ). In samples with high reactive stroma content, metabolites and genes linked to immune functions and extracellular matrix (ECM) remodeling were significantly upregulated. Future validation of these findings is important to reveal novel biomarkers and drug targets connected to immune mechanisms and ECM in prostate cancer. The fact that high reactive stroma grading is connected to BCR adds further support for the clinical integration of this histopathological evaluation.

The tumor microenvironment (TME) has in recent years gained attention for its role in cancer cell and tumor development. TME, considered to consist of non-malignant cells and their products, is more genetically stable than cancer cells and supports and allows cancer cells to develop<sup>1,2</sup>. In prostate tumors, TME include activated fibroblasts called cancer associated fibroblasts (CAFs), immune cells and vasculature cells. It is often the site of chronic inflammation and extracellular matrix (ECM) remodeling, similar to what occurs during wound-healing with an increase of activated fibroblasts<sup>2,3</sup>. Such inflammatory TME is usually referred to as 'reactive stroma'. In prostate cancer, a transition from healthy stroma to reactive stroma has been characterized by a replacement of smooth muscle cells by CAFs and immune cells<sup>3</sup>.

For prostate cancer, the current gold standard for predicting clinical outcome is histopathological evaluation through the Grade Group system<sup>4</sup>. This system sets a grade based on the morphological appearance of prostate glands and cancerous epithelial cells. However, the tumor area can contain clinically relevant histopathologic information that is not captured by the current grading system. Ayala *et al.* were the first to develop a grading system for reactive stroma in prostate cancer and to show that a higher level of reactive stromal response is

<sup>1</sup>Department of Circulation and Medical Imaging, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. <sup>2</sup>Department of Clinical and Molecular Medicine, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. <sup>3</sup>Department of Medical Biology, UiT The Arctic University of Norway, Tromsø, Norway. <sup>4</sup>Department of Clinical Pathology, University Hospital of North Norway, UHN Tromsø, Norway. <sup>5</sup>Department of Urology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. <sup>6</sup>Department of Pathology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. <sup>7</sup>Clinic of Surgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. Correspondence and requests for materials should be addressed to M.K.A. (email: [maria.k.andersen@ntnu.no](mailto:maria.k.andersen@ntnu.no)) or M.-B.T. (email: [may-britt.tessem@ntnu.no](mailto:may-britt.tessem@ntnu.no))



connected to biochemical recurrence (BCR)<sup>5</sup>. Since then, several studies have linked high reactive stroma content to a worse clinical outcome, including BCR<sup>6–9</sup>, development of castration-resistant prostate cancer<sup>10</sup> and prostate cancer-specific mortality<sup>11</sup>. In particular, evaluating the tumor stroma was shown to be of extra value in cases where the Grade Group system failed to accurately predict outcome<sup>6</sup>. Although validation and standardization is needed, incorporating reactive stroma into the clinical histopathology evaluation, along with Grade Group, shows potential to optimize prognostic stratification of prostate cancer patients.

As reactive stroma appears to play a significant role in cancer development<sup>12</sup>, it is of interest to understand its underlying molecular mechanisms. These insights may provide new prognostic markers and therapeutic targets. Some molecular features of reactive stroma have already been identified. Smooth muscle differentiation markers such as calponin and desmin are commonly reduced in reactive stroma<sup>3,5,9,10,13</sup>. In contrast, vimentin, pro-collagen and tenascin-C, markers for activated fibroblasts and ECM remodeling, are elevated in reactive stroma in prostate cancer tissue<sup>3,10,13,14</sup>. Reactive stromal cells have also been suggested to promote angiogenesis in the tumor area<sup>15</sup>. Dakova *et al.*<sup>16</sup> performed global gene expression on laser dissected prostate tissue samples, identifying several differentially expressed genes between reactive and normal stroma. These included genes related to functions such as neurogenesis and DNA repair<sup>16</sup>. Thus, research on proteins and gene expression has revealed changes associated with ECM remodeling, angiogenesis and DNA repair. In contrast, metabolic patterns related to reactive stroma content in prostate cancer tissue are currently unknown. Metabolic reprogramming is a hallmark of cancer and several metabolic alterations has been identified in prostate cancer tissue compared to normal tissue through metabolic profiling, including increase of choline<sup>17</sup> and sarcosine<sup>18</sup>, and decrease of polyamine and citrate levels<sup>19</sup>.

The aim of our study was to combine histopathology determined reactive stromal grading (RSG) with integrative analysis of metabolomics and transcriptomics data from the same prostate cancer tissue sample, thereby investigating the molecular characteristics of reactive stroma in prostate cancer. Further we investigated how the expression of significant genes and metabolites of reactive stroma are correlated, and investigated biochemical recurrence of patients with high reactive stroma content.

## Methods

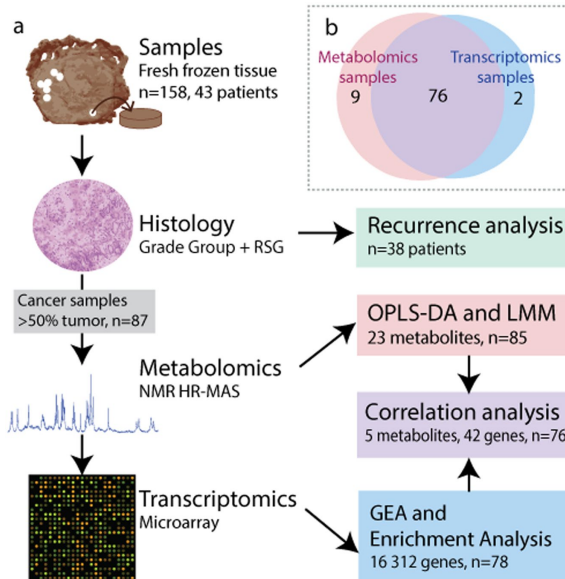
**Patients and tissue collection.** This study was approved by the Regional committee for Medical and Health Research Ethics (REC) central Norway (identifier 4.2007.1890). All experiments were carried out in accordance to the ethical regulations of REC. All tissue donors signed a written informed consent.

Tissue used for this study was donated and collected in 2007 and 2008 ensuing radical prostatectomy. None of the patients received neoadjuvant therapy prior to surgery. A two mm thick tissue slice was cut from the middle of the prostate gland perpendicular to the urethra. The slice was snap frozen in liquid nitrogen on average 15 minutes after surgical removal and stored at  $-80^{\circ}\text{C}$  as previously described by Bertilsson *et al.*<sup>20</sup>. Between four and eleven core tissue samples (three mm diameter) were later collected from each prostate slice (Fig. 1a). In total 158 samples were collected from 43 patients. We obtained at least five years of clinical follow-up from the hospital patient records (Braadland *et al.*<sup>21</sup>) including T-stage, clinical Gleason score (postoperative), tumor volume, preoperative serum prostate specific antigen (PSA) measurements and biochemical recurrence (defined as PSA  $\geq 0.2$  ng/ml). The clinical Gleason scores were translated into the new Grade Group system as described by Gordetsky and Epstein<sup>4</sup>.

**Histopathological evaluation.** From one side of each fresh frozen tissue sample, a four  $\mu\text{m}$  thick cryosection was stained with hematoxylin and eosin (HE). All HE-stained slides ( $n = 158$ ) were evaluated independently by two experienced uropathologists (E.R. and Ø.S.). Percentage of cancer, normal epithelium and healthy stroma were determined along with Grade Group<sup>4</sup>. Reactive stroma content was defined as the percentage of stroma that was reactive within the tumor area, according to the reactive stroma grade (RSG) system developed by Ayala *et al.*<sup>5</sup>. Each sample was given a grade ranging from 0 to 3: RSG 0 containing 0–5% reactive stroma; RSG 1, 6–15% reactive stroma; RSG 2, 16–50% reactive stroma and RSG 3, 51–100% reactive stroma. Normal prostatic stroma with a high number of smooth muscle cells were characterized by a strong red eosinophilic staining, and the cells by having a large cytoplasm, rounded nuclei and organization into bundles (Fig. 2a). When the stroma gets reactive there will be a replacement of smooth muscle cells by CAFs and immune cells, and the stroma will appear with a paler eosinophilic coloring (Fig. 2b–d). Kappa-statistics was used to calculate a quality score between the two pathologists for both Grade Group and RSG<sup>22</sup>. Later, consensus was reached between the pathologists when there was disagreement on RSG. With disagreement on Grade Group, an independent previous histopathological evaluation by a third pathologist was used to find consensus<sup>20</sup>.

**Metabolomics.** Metabolite data was obtained by high-resolution magic angle spinning magnetic resonance spectroscopy (HR-MAS MRS) on fresh frozen tissue samples. HR-MAS MRS spectra were acquired on a Bruker Advance DRX600 (14.1 T) spectrometer (Bruker BioSpin, Germany) with a  $^1\text{H}/^{13}\text{C}$  MAS probe. LCModel was applied to quantify 23 metabolites from the spectra<sup>23,24</sup>. Further details of the HR-MAS MRS procedure, spectral pre-processing and metabolite quantification are described by Giskeødegård *et al.*<sup>19</sup>. Furthermore, samples containing  $>50\%$  tumor ( $n = 85$ ) were selected for molecular and statistical analysis to ensure that the metabolomics profiles mainly represented tumor areas.

**RNA microarray.** After HR-MAS MRS, the tissue samples were homogenized and mRNA was extracted. Isolated mRNA was amplified with Illumina TotalPrep RNA amplification Kit (Ambion Inc.) and relative gene expression was subsequently measured with Illumina Human HT-12v4 Expression Bead Chip (Illumina). A comprehensive overview of the protocol and data preprocessing is reported by Bertilsson *et al.*<sup>25</sup>. Here we also selected

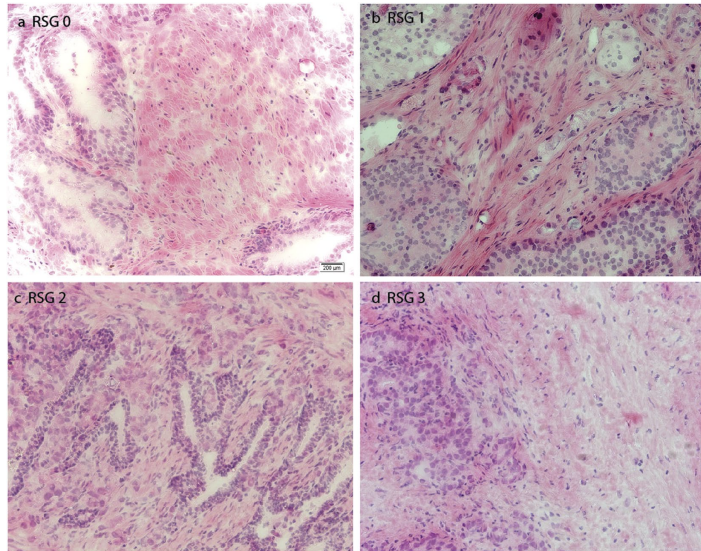


**Figure 1.** Methodology flowchart. (a) Samples were collected from fresh frozen human prostate tissue and cryosections were stained with hematoxylin and eosine. Two pathologists evaluated Grade Group and reactive stroma grade (RSG). Samples with >50% tumor content were selected for further metabolomics and transcriptomics analysis. Data analysis included survival analysis (Kaplan-Meier and Cox hazard proportional analysis) with biochemical recurrence as endpoint, multivariate orthogonal partial least squares discriminant analysis (OPLS-DA), linear mixed models (LMM), gene expression analysis (GEA) and Pearson correlation between selected genes and metabolites. GEA results were used for enrichment analysis. (b) Venn diagram of samples used for metabolomics, transcriptomics and both.

samples with >50% tumor content ( $n = 78$ ) for further gene expression analysis (GEA). There was an overlap of 76 samples which were subjected to both metabolite and gene expression analysis (Fig. 1b).

**Multivariate and statistical analysis.** Biochemical recurrence (BCR)-free survival analysis included Kaplan Meier and Cox proportional hazards analysis and were performed with the *survival* package in the R environment. BCR was defined as serum PSA  $>0.2$  ng/mL, confirmed by two independent measurements. Time-to-event was set as the number of days between radical prostatectomy and confirmed BCR. Three patients were lost to follow-up and two patients received adjuvant treatment before BCR. As the adjuvant treatment could be influencing the time to BCR, these patients were removed from survival analysis, resulting in a total of 38 patients. As multiple samples were collected from each patient, the sample with the highest RSG was selected as representative for a patient in survival analysis (patient RSG). Patients were divided into a *low RSG* (RSG 0 and 1) group and *high RSG* (RSG 2 and 3) group due to the low numbers of RSG 0 and RSG 3 patients. Covariates included in Cox proportional hazard was *low vs high RSG* and clinical Grade Group. For Kaplan-Meier, a log-rank test was used to calculate significance. In addition, to correct for the possible confounding effect of clinical Grade Group and T-stage, a second Kaplan-Meier analysis was performed after removing patients with clinical Grade Group  $\geq 4$ , as this produced the same median Grade Group and T-stage in both the *low* and *high RSG* group. Pearson correlations between RSG and clinical Grade Group, and RSG and preoperative PSA of the patients were also performed.

Multivariate analysis of the metabolite dataset (23 metabolites,  $n = 85$ ) was performed in PLSToolbox in the MatLab 8.6.0 (The Mathworks, Inc, USA) environment. The dataset was preprocessed by autoscaling. Supervised orthogonal partial least squares discriminant analysis (OPLS-DA) was used to examine metabolic differences between *high* and *low RSG* using leave-10%-of-patients-out cross-validation and permutation testing for analyzing model reliability (1000 permutations).



**Figure 2.** Photomicrographs (x20) of representative hematoxylin and eosinophil stained slides of histopathology of prostate tissue cryosections with reactive stroma grade (RSG) 0–3. (a) Normal prostatic tissue with reactive stromal grade (RSG) 0 (<5% reactive stroma). Stroma is mostly consisting of smooth muscle cells making up bundles. (b) RSG 1 (6–15% reactive stroma) and Grade Group 4. The majority of stroma still has a strong eosinophilic stain, with a few cells with paler staining appearing, in addition to the presence of more fibroblasts. (c) RSG 2 (16–50% reactive stroma) and Grade Group 3. The reactive stroma is more prominent by a weaker eosinophilic stain. (d) Sample with RSG 3 (>50% reactive stroma) and Grade Group 3. Here, nearly all normal stroma is replaced by reactive stroma with pale eosinophil staining.

Univariate analysis of the 23 quantified log-transformed metabolites was performed with linear mixed models (LMM) in R with the *nlme* package<sup>26</sup>. The relationship between each metabolite concentration and RSG was modeled while correcting for multiple samples per patient. Correct model assumptions were confirmed by qq-plots of model residuals.

Univariate GEA was carried out with the *lumi* and *limma* packages in R for the 23 444 probes, representing 16 312 genes. Samples with *low* RSG were compared to samples with *high* RSG. The result of the GEA was further used to remove duplicated probes so that the dataset only contained one probe per gene. The probe with the lowest adjusted p-value from the GEA was selected for further analysis. The significantly upregulated and downregulated genes were separated into two gene lists, and used for enrichment analysis with Enrichr<sup>27,28</sup>. Results from the background library Gene Ontology (GO) Biological Process 2018 were exported.

Pearson correlation between significant metabolites ( $n = 5$ ) and the most significantly expressed genes involved in relevant biological processes ( $n = 42$ ) was calculated in R. Due to lack of normal distribution, the metabolite data was log<sub>2</sub>-transformed prior to correlation analysis. Five metabolites were selected based on significance in LMM analysis and/or a loading score of  $\geq \pm 3.0$  (first latent variable, OPLS-DA). The genes were selected based on an adjusted p-value < 0.001 from GEA ( $n = 98$ , Supplementary Table S1). These genes were manually annotated through geneCards.org, and genes with a clear relation to biologically relevant processes were selected for correlation analysis ( $n = 42$ ).

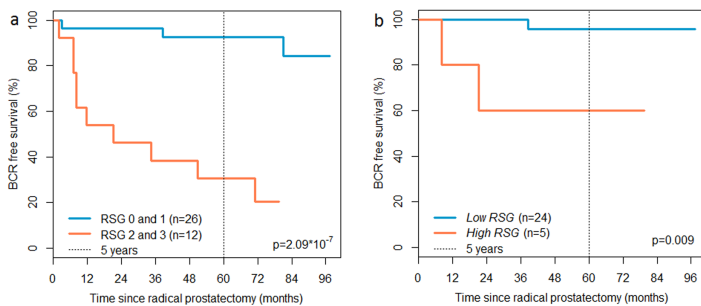
Unadjusted p-values of  $\leq 0.05$  were considered significant for univariate tests and LMM on the metabolic dataset due to a low number of variables ( $n = 23$ ). Benjamini-Hochberg adjusted p-values  $\leq 0.05$  were considered significant for GEA, enrichment analysis and gene-metabolite correlations. All confidence intervals (CI) were 95%.

## Results

**Histopathology.** A total of 158 samples from 43 patients were histologically evaluated for Grade Group, tumor content and RSG (Fig. 1). Before consensus between pathologists was reached on tumor containing samples ( $n = 108$ ), the original evaluations gave a kappa score of 0.64 and 0.30 for Grade Group and RSG, respectively. An overview of histopathology and clinical data are listed in Table 1. The majority of samples ( $n = 48$ , 55.2%) and patients ( $n = 24$ ,

	RSG 0	RSG 1	RSG 2	RSG 3	Total
<b>Samples with &gt;50% tumor used for metabolomics (n = 85)</b>					
Samples (percent)	11 (12.9%)	47 (55.3%)	23 (27.1%)	4 (4.7%)	85
Median Grade Group (range)	3 (1–5)	1 (1–5)	3 (1–5)	4.5 (3–5)	2 (1–5)
Mean tumor percent (range)	89.5 (70–100)	82.3 (60–92.5)	83.2 (62.5–100)	88.1 (72.5–97.5)	83.6 (60–100)
<b>Samples with &gt;50% tumor used for transcriptomics (n = 78)</b>					
Samples (percent)	10 (12.8%)	41 (52.3%)	23 (29.5%)	4 (5.2%)	78
Median Grade Group (range)	2.5 (1–5)	1 (1–5)	3 (1–4)	4.5 (3–5)	2 (6–10)
Mean tumor percent (range)	87.5 (70–100)	83.2 (57.5–95)	83.2 (62.5–100)	88.1 (72.5–97.5)	84.0 (57.5–100)
<b>Clinical variables of Patients (n = 38)</b>					
Patients (percent)	2 (5.3%)	24 (63.2%)	10 (26.3%)	2 (5.3%)	38
Recurrence, 5 year follow-up (percent)	0	1 (4.2%)	7 (70.0%)	1 (50.0%)	11 (28.9%)
Mean age at operation (range)	58.5 (56–61)	61.4 (48–69)	62.1 (48–68)	68.5 (68–69)	61.8 (48–69)
Median Grade Group (range)	2 (2)	2 (1–5)	3.5 (1–5)	5 (5)	3 (1–5)
Median pathological stage (range)	T2c (T2c)	T2c (T2a–T3b)	T3a (T2c–T3b)	T3b (T3a–T3b)	T2c (T2a–T3b)
Mean preoperative serum PSA (range)	8.0 (5.2–10.7)	10.8 (3.7–45.8)	10.6 (5.2–17.0)	9.75 (5.6–13.9)	10.3 (3.7–48.8)

**Table 1.** Histology of samples and clinical data of patients. RSG = reactive stroma grade, PSA = prostate specific antigen.

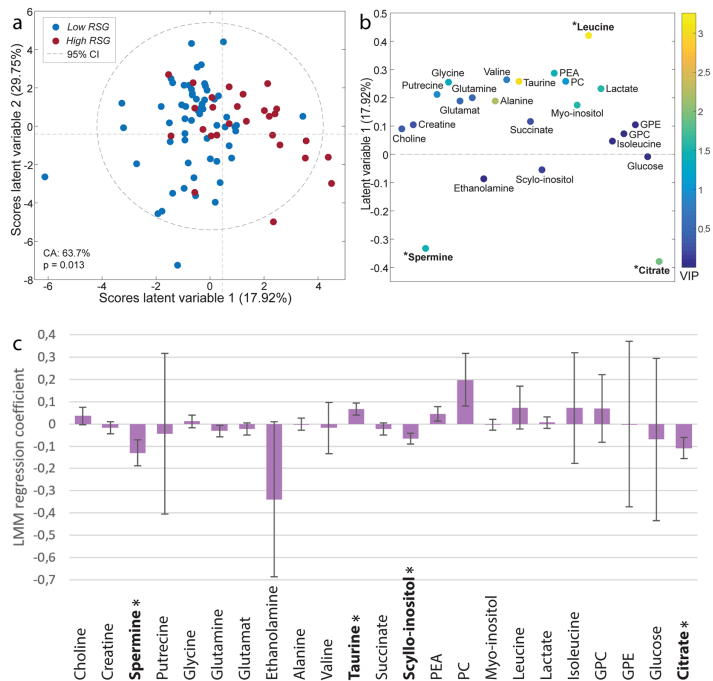


**Figure 3.** Kaplan-Meier plots of biochemical recurrence (BCR). Kaplan-Meier analysis was performed on (a) all patients (n = 38) and (b) patients with low-to-medium Grade Group ( $\leq 3$ ) (n = 29).

63.2%) were scored as RSG 1, while the least prevalent score was RSG 3 with four samples (4.6%) and two patients (5.3%). There was a clear correlation between clinical Grade Group and RSG of patients ( $R = 0.56$ ,  $p = 0.23 \times 10^{-3}$ ). There was also a weak, but significant, correlation between RSG and Grade Group in the samples ( $R = 0.25$ ,  $p = 0.018$ ). There was no correlation between preoperative PSA levels and patient RSG ( $R = 0.015$ ,  $p = 0.93$ ).

**High RSG predict shorter BCR-free survival independent of Grade Group.** A total of 38 patients had sufficient clinical follow-up data and were included in survival analysis where *low RSG* (n = 26) was compared to *high RSG* (n = 12). Kaplan-Meier analysis showed significantly better BCR-free survival in patients with *low RSG*, having 92.3% recurrence-free survival, and *high RSG* patients having 25.0% recurrence-free survival after 5 years of follow-up ( $p = 2.09 \times 10^{-7}$ ) (Fig. 3a). However, the *low* and *high RSG* patient groups had a different median clinical Grade Group of 2 and 4, respectively (two-sided t-test,  $p = 0.013$ ). In addition, these two groups also had a significant different median T-stage of T2c for *low RSG* and T3a for *high RSG* (two-sided t-test,  $p = 0.16 \times 10^{-3}$ ). A second Kaplan-Meier analysis was therefore performed for patients with Grade Group  $\leq 3$ , resulting in a total of 29 patients. This second selection of *low* (n = 24) and *high RSG* (n = 5) patients had the same median Grade Group of 2 and median T-stage of T2c and still displayed a significant recurrence-free 5-year survival difference (BCR-free survival 95.8% for *low RSG* and 60% for *high RSG*,  $p = 0.009$ ) (Fig. 3b). Multivariate Cox proportional hazard model of all 38 patients provided hazard ratios of 16.44 ( $p = 0.013$ , CI = 1.81–149.20) for RSG and 1.95 ( $p = 0.018$ , CI = 1.12–3.40) for Grade Group.

**Reactive stroma shows metabolic alteration.** Multivariate OPLS-DA analysis using quantified values for 23 metabolites showed a significant difference between *high* and *low RSG* ( $p = 0.014$ , accuracy 64.9%, sensitivity 75.0% and specificity 54.9%, Fig. 4a,b). The loadings depicted in Fig. 4b show that there are lower levels of citrate and spermine and higher levels of leucine in samples with *high RSG*.



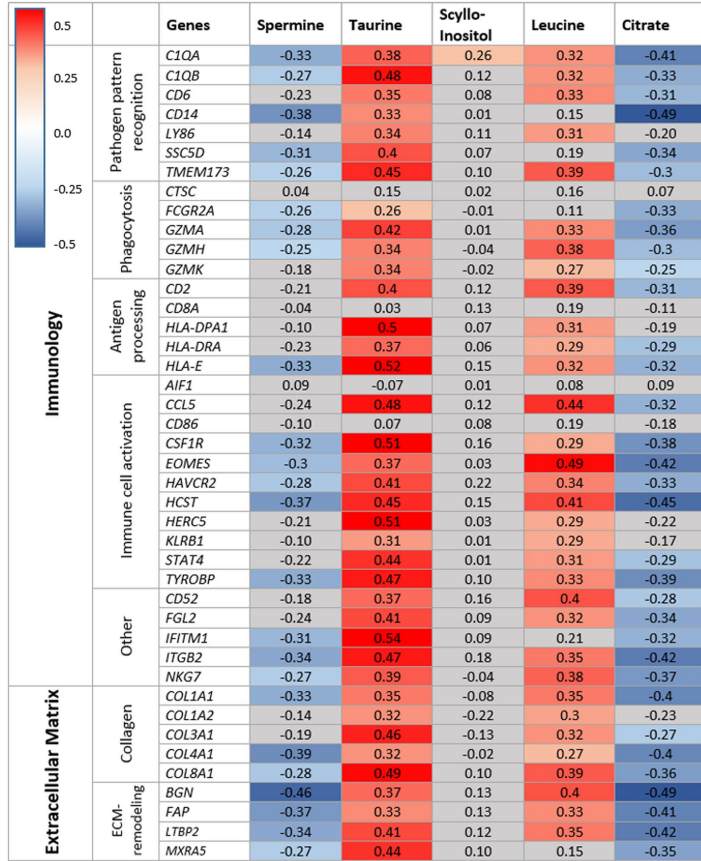
**Figure 4.** Metabolite analysis in samples with *high* and *low* reactive stroma grading (RSG). (a) Scores plot and (b) loadings plot from OPLS-DA model where *low* RSG (RSG 0 and 1,  $n = 58$ ) were compared to *high* RSG (RSG 2 and 3,  $n = 27$ ). Variables in the loadings plot are color-coded by variable importance in the projection (VIP), which is an estimate of each variables contribution to the model. Metabolites with a loadings score  $\geq 3.0$  or  $\leq -3.0$  are indicated by \* (c) Univariate linear mixed model (LMM) regression coefficients with increasing RSG. Error bars represent standard error and significant metabolites are indicated by \*. Abbreviations: CA = classification accuracy, GPC = Glycerophosphocholine, GPE = Glycerophosphoethanol, PC = phosphocholine, PEA = phosphoethanolamine and GPE = Glycerophosphoethanol.

Univariate LMM testing of each quantified metabolite modeled against RSG values 0–3 resulted in four significant metabolites (Fig. 4c). Taurine ( $p = 0.018$ ) was found at elevated levels, while citrate ( $p = 0.027$ ), spermine ( $p = 0.031$ ) and scyllo-inositol ( $p = 0.009$ ) were found at lower levels with increasing RSG.

**Genes involved in immune responses and ECM remodeling are upregulated in reactive stroma.** Gene expression analysis (GEA) was performed comparing *high* RSG to *low* RSG. A total of 609 and 471 genes were up- and downregulated, respectively. Enrichment analysis was performed with Enrichr using gene lists of significantly up- and downregulated genes, which produced 339 significantly upregulated and seven significantly downregulated enriched biological process terms in *high* compared to *low* RSG (Supplementary Table S2). All biological terms with a combined score (calculated by Enrichr) over 30 are presented in Fig. 5. Of these terms ( $n = 22$ ), all were upregulated and 18 were related to the immune system, three to cell signaling and one was related to extracellular matrix.

**Correlation between selected genes and metabolites.** A total of 42 upregulated genes and five metabolites (spermine, taurine, scyllo-inositol, leucine and citrate) were selected for correlation analysis. Immunology and ECM were considered relevant biological processes to reactive stroma based on our enrichment results and the literature<sup>29,30</sup>, and were along with level of significance, used as selection criteria for the genes. Nine genes were related to ECM and 33 were related to immunology, which could be further categorized into various different functions of the immune system and ECM (Fig. 6). Of the selected genes, four immunology-related genes,





**Figure 6.** Correlation analysis between selected metabolites and genes. Values are Pearson correlation coefficients. Values marked with red (positive correlation) or blue color (negative correlation) were significant after Benjamini-Hochberg adjustment, while values marked with light grey were non-significant. Color intensity corresponds to the correlation coefficient value.

samples (Fig. 4b). Leucine is a key amino acid of proteoglycans such as decorin and biglycan. These molecules function as building blocks during ECM remodeling and are found with elevated expression in tumor stroma<sup>36</sup>. Biglycan expression was upregulated among *high* RSG samples in this study ( $p = 1.23 \times 10^{-4}$ ) and is previously reported to attract pro-inflammatory macrophages in both cell culture and mice<sup>37</sup>. In sum, our metabolic profile appears to be linked to inflammation and ECM remodeling.

The results from gene enrichment analysis indicated that genes involved in immunity, cell signaling and extra cellular matrix were particularly important when comparing *low* and *high* RSG (Fig. 5). Cellular signaling pathways are known to be reprogrammed in cancer cells<sup>38</sup> and our results from the enrichment analysis may therefore represent both cancer cell and the cross-talk between cancer cells and reactive stroma. One known example is transforming growth factor- $\beta$  (TGF- $\beta$ ), significantly upregulated in our GEA ( $p = 0.003$ , Supplementary Data S1), which is secreted by cancer cells, activates fibroblasts and promotes ECM remodeling<sup>39</sup>. Remodeling of the ECM is, together with inflammation, a feature of the reactive stroma<sup>40</sup>, and is parallel to chronic wound repair.



Among the genes which were selected based on level of significance between *low* and *high* RSG and their involvement in immunity and ECM remodeling, we found 12 genes that were specifically involved in pathogen responses, such as phagocytosis, pathogen pattern recognition and antigen processing (Fig. 6). Additionally, biological processes related to interferon signaling were particularly enriched (Fig. 5). Interferons are a group of signaling proteins which are secreted from cells as a response to pathogen infections<sup>41</sup>. Infectious pathogens like bacteria and viruses may be involved in chronic inflammation and further progression of cancer<sup>42</sup>. In previous studies different pathogens were correlated with prostate cancer initiation, including high risk human papilloma virus (HR-HPV)<sup>43</sup>, *Enterobacteriaceae* species<sup>44</sup> and *Porphyromonas acnes*<sup>44–47</sup>. In our study, both the genes *CD6* and *CD14*, which are directly involved in recognition of surface bound bacterial lipopolysaccharide (LPS)<sup>48,49</sup>, were expressed higher in *high* RSG samples. The fact that genes specifically involved in both recognition and destruction of pathogens are among the most highly expressed genes in *high* RSG, suggest the presence of infectious agents contributing to the reactive stromal response. Future studies using sensitive methods suitable for detecting suspected pathogens are needed.

Several genes involved in immune cell activation were differentially expressed between *high* and *low* RSG (Fig. 6). Many of these genes are involved in regulation of inflammatory responses, by modifying the functions of T-cells, macrophages and natural killer cells, and can either be pro-inflammatory or inhibit immune responses. These genes include *CCL5* and *CSF1R*. *CCL5* is a pro-inflammatory chemokine that attracts immune cells such as macrophages, T-leukocytes, eosinophils and basophiles<sup>50</sup>. *CCL5* has previously been linked to cancer progression in prostate<sup>51</sup>. *CSF1R* is a pro-inflammatory receptor mainly found on macrophages and monocytes. It is thought to trigger recruitment, growth and proliferation of these cells in cancer, and blocking this receptor was found to suppress tumor growth in combination with irradiation therapy in prostate cancer patients<sup>52</sup>. These data indicate that the tissue is inflamed by the actions of an array of immune cells.

Several genes related to remodeling of the ECM were upregulated in reactive stroma in this study (Fig. 6). One of the key contributors to reactive stroma is a group of activated fibroblasts, CAFs. The function of these cells is to remodel the ECM<sup>53</sup>. CAFs have an elevated production of  $\alpha$  smooth muscle actin ( $\alpha$ -SMA) and fibroblast activation protein (FAP) compared to other cells in the tissue<sup>53,54</sup>. *FAP* is selectively expressed by activated fibroblasts during either wound-healing responses or by CAFs in epithelial cancers<sup>55,56</sup>, and was found to have increased expression in *high* RSG in our cohort ( $p = 0.001$ ). Expression of  $\alpha$ -SMA is also a key characteristic of CAFs, but it was not found to be differentially expressed in reactive stroma of this study ( $p = 0.34$ ). A possible explanation for this observation is that  $\alpha$ -SMA is also produced by smooth muscle cells<sup>40</sup>, so any increase in fibroblast-derived  $\alpha$ -SMA may be hidden by a reduction of smooth muscle-derived  $\alpha$ -SMA.

Collagen is the most abundant type of protein making up the ECM, and various collagen genes had increased expression in *high* RSG samples in our study. In cancer, breakdown and re-deposition of collagen is common and causes cancer progression through destabilization of cell polarity and cell-to-cell adhesion<sup>57</sup>. Collagen building is thought to be partly organized by the proteoglycan biglycan<sup>58</sup>. Biglycan is encoded by the gene *BGN*, which was higher expressed in *high* RSG ( $p = 0.12 * 10^{-5}$ ). Up-regulation of *BGN* has previously been linked to poor prognosis in prostate cancer<sup>59</sup>. Another proteoglycan encoding gene which were higher expressed in *high* RSG, *MXRA5*, has a similar function to *BGN* and is associated with several forms of cancers<sup>60</sup>. These findings reflect the remodeling of ECM which occurs in reactive stroma, and suggest that a higher number of CAFs are likely present due to the high expression of *FAB*, a selective marker for activated fibroblasts.

Even though stromal grading shows clinical potential, RSG evaluation will still need standardization before it can be implemented in the clinic, clearly indicated by the kappa score for RSG ( $\kappa = 0.30$ ) which was considerably lower than for Grade Group ( $\kappa = 0.64$ ). To our knowledge, no kappa score was included in any of the previous published studies, and it is therefore not possible to compare the robustness of our evaluation to others. Progress are being made to optimize characterization of reactive stroma<sup>61</sup> and there is a need to quality check and quantify the variation between individual pathologists. In addition, evaluating RSG on cryosections caused further limitation in this study due to common lower staining quality compared to sections from formalin fixed paraffin embedded tissue. There is higher requirement for section quality and staining when assessing RSG compared to assessing Grade Group.

In this study we have demonstrated that reactive stroma grading of prostate cancer offer additional prognostic value as a supplement to the clinical Grade Group assessment. However, for applying RSG in the routine clinical assessment, more standardized scoring criteria is needed. Metabolic and translational differences between samples with *high* and *low* reactive stroma content were also identified. In particular, genes related to immunology and ECM remodeling were upregulated in samples with high reactive stroma content. Molecular understanding of the reactive stroma may lead to new diagnostic and therapeutic tools. Identifying therapeutic targets residing in reactive stroma, could be of particular benefit due to the higher degree of genetic stability compared to cancer cells. Hence, such therapeutic targets might be less prone to treatment resistance.

## References

- Palumbo, A., de Oliveira Meireles Da Costa, N., Bonamino, M. H., Ribeiro Pinto, L. F. & Nasciutti, L. E. Genetic instability in the tumor microenvironment: a new look at an old neighbor. *Mol. Cancer* **14**, 145, <https://doi.org/10.1186/s12943-015-0409-y> (2015).
- Bianchi-Frias, D. *et al.* Cells comprising the prostate cancer microenvironment lack recurrent clonal somatic genomic aberrations. *Mol. Cancer Res.* **14**, 374–384 (2016).
- Tuxhorn, J. A. *et al.* Reactive stroma in human prostate cancer induction of myofibroblast phenotype and extracellular matrix remodeling. *Clin. Cancer Res.* **8**, 2912–2923 (2002).
- Gordetsky, J. & Epstein, J. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagn. Pathol.* **11**, 25, <https://doi.org/10.1186/s13000-016-0478-2> (2016).
- Ayala, G. *et al.* Reactive stroma as a predictor of biochemical-free recurrence in prostate cancer. *Clin. Cancer Res.* **9**, 4792–4801 (2003).



6. McKenney, J. K. *et al.* Histologic grading of prostatic adenocarcinoma can be further optimized: analysis of the relative prognostic strength of individual architectural patterns in 1275 patients from the Canary retrospective cohort. *Am. J. Surg. Pathol.* **40**, 1439–1456 (2016).
7. Yanagisawa, N. *et al.* Stromogenic prostatic carcinoma pattern (carcinomas with reactive stromal grade 3) in needle biopsies predicts biochemical recurrence-free survival in patients after radical prostatectomy. *Hum. Pathol.* **38**, 1611–1620 (2007).
8. Billis, A. *et al.* Adenocarcinoma on needle prostatic biopsies: does reactive stroma predicts biochemical recurrence in patients following radical prostatectomy? *Int. Braz. J. Urol.* **39**, 320–327 (2013).
9. Tomas, D. *et al.* Intensity of stromal changes predicts biochemical recurrence-free survival in prostatic carcinoma. *Scand. J. Urol. Nephrol.* **44**, 284–290 (2010).
10. Wu, J. P. *et al.* Intensity of stromal changes is associated with tumor relapse in clinically advanced prostate cancer after castration therapy. *Asian J. Androl.* **16**, 710–714 (2014).
11. Sæter, T. *et al.* The prognostic value of reactive stroma on prostate needle biopsy: A population-based study. *Prostate* **75**, 662–671 (2015).
12. Shiao, S. L., Chu, G. C.-Y. & Chung, L. W. K. Regulation of prostate cancer progression by the tumor microenvironment. *Cancer Lett.* **380**, 340–348 (2016).
13. Tomas, D. & Kruslin, B. The potential value of (Myo)fibroblastic stromal reaction in the diagnosis of prostatic adenocarcinoma. *Prostate* **61**, 324–331 (2004).
14. Silva, M. M. Jr. *et al.* Characterization of reactive stroma in prostate cancer: involvement of growth factors, metalloproteinase matrix, sexual hormones receptors and prostatic stem cells. *Int. Braz. J. Urol.* **41**, 849–858 (2015).
15. Yang, F. *et al.* Stromal expression of connective tissue growth factor promotes angiogenesis and prostate cancer tumorigenesis. *Cancer Res.* **65**, 8887–8895 (2005).
16. Dakhova, O. *et al.* Global gene expression analysis of reactive stroma in prostate cancer. *Clin. Cancer Res.* **15**, 3979–3989 (2009).
17. Awwad, H. M., Geisel, J. & Obeid, R. The role of choline in prostate cancer. *Clin. Biochem.* **45**, 1548–1553 (2012).
18. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457**, 910–914 (2009).
19. Giskeødegård, G. F. *et al.* Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness. *PLoS One* **8**, e62375, <https://doi.org/10.1371/journal.pone.0062375> (2013).
20. Bertilsson, P. R. *et al.* A new method to provide a fresh frozen prostate slice suitable for gene expression study and MR spectroscopy. *Prostate* **71**, 461–469 (2011).
21. Braadland, P. R. *et al.* *Ex vivo* metabolic fingerprinting identifies biomarkers predictive of prostate cancer recurrence following radical prostatectomy. *Br. J. Cancer* **117**, 1656, <https://doi.org/10.1038/bjc.2017.346> (2017).
22. Cross, S. S. Kappa statistics as indicators of quality assurance in histopathology and cytopathology. *J. Clin. Pathol.* **49**, 597–599 (1996).
23. Provencher, S. W. Estimation of metabolite concentrations from localized *in vivo* proton NMR spectra. *Magn. Reson. Med.* **30**, 672–679 (1993).
24. Opstad, K. S., Wright, A. J., Bell, B. A., Griffiths, J. R. & Howe, F. A. Correlations between *in vivo* 1H MRS and *ex vivo* 1H HRMAS metabolite measurements in adult human gliomas. *J. Magn. Reson. Imaging* **31**, 289–297 (2010).
25. Bertilsson, H. *et al.* Changes in gene transcription underlying the aberrant citrate and choline metabolism in human prostate cancer samples. *Clin. Cancer Res.* **18**, 3261–3269 (2012).
26. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1–117, <http://CRAN.R-project.org/package=nlme>.
27. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128, <https://doi.org/10.1186/1471-2105-14-128> (2013).
28. Kulshow, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–97, <https://doi.org/10.1093/nar/gkw377> (2016).
29. Kruslin, B., Ulaevec, M. & Tomas, D. Prostate cancer stroma: an important factor in cancer growth and progression. *Bosnian J. Basic Med.* **15**, 1–8 (2015).
30. Bussard, K. M., Mutkus, L., Stumpf, K., Gomez-Manzano, C. & Marini, F. C. Tumor-associated stromal cells as key contributors to the tumor microenvironment. *Breast Cancer Res.* **18**, 84, <https://doi.org/10.1186/s13058-016-0740-2> (2016).
31. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).
32. Lynch, M. J. & Nicholson, J. K. Proton MRS of human prostatic fluid: Correlations between citrate, spermine, and myo-inositol levels and changes with disease. *Prostate* **30**, 248–255 (1997).
33. Marcinkiewicz, J. & Kontny, E. Taurine and inflammatory diseases. *Amino Acids* **46**, 7–20 (2014).
34. Swanson, M. G. *et al.* Proton HR-MAS spectroscopy and quantitative pathologic analysis of MRI/3D-MRSI-targeted postsurgical prostate tissues. *Magn. Reson. Med.* **50**, 944–954 (2003).
35. Hahn, P. *et al.* The classification of benign and malignant human prostate tissue by multivariate analysis of 1H magnetic resonance spectra. *Cancer Res.* **57**, 3398–3401 (1997).
36. Bi, X. L. & Yang, W. Biological functions of decorin in cancer. *Chin. J. Cancer* **32**, 266–269 (2013).
37. Schaefer, L. *et al.* The matrix component biglycan is proinflammatory and signals through Toll-like receptors 4 and 2 in macrophages. *J. Clin. Invest.* **115**, 2223–2233 (2005).
38. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
39. Rowley, D. R. Transforming Growth Factor- $\beta$  in Cancer Therapy. Volume II. 30, 475–505 (Springer, 2008).
40. Barron, D. A. & Rowley, D. R. The reactive stroma microenvironment and prostate cancer progression. *Endocr. Relat. Cancer* **19**, R187–R204 (2012).
41. Pestka, S., Krause, C. D. & Walter, M. R. Interferons, interferon-like cytokines, and their receptors. *Immunol. Rev.* **202**, 8–32 (2004).
42. Vandeven, N. & Nghiem, P. Pathogen-driven cancers and emerging immune therapeutic strategies. *Cancer Immun. Res.* **2**, 9–14 (2014).
43. Singh, N. *et al.* Implication of high risk human papillomavirus HR-HPV infection in prostate cancer in Indian population: a pioneering case-control analysis. *Sci. Rep.* **5**, 7822, <https://doi.org/10.1038/srep07822> (2015).
44. Yow, M. A. *et al.* Characterisation of microbial communities within aggressive prostate cancer tissues. *Infect. Agent. Cancer* **12**, 4, <https://doi.org/10.1186/s13027-016-0112-7> (2017).
45. Bae, Y. *et al.* Intracellular Propionibacterium acnes infection in glandular epithelium and stromal macrophages of the prostate with or without cancer. *PLoS One* **9**, e90324, <https://doi.org/10.1371/journal.pone.0090324> (2014).
46. Kakegawa, T. *et al.* Frequency of Propionibacterium acnes Infection in Prostate Glands with Negative Biopsy Results Is an Independent Risk Factor for Prostate Cancer in Patients with Increased Serum PSA Titers. *PLoS One* **12**, e0169984, <https://doi.org/10.1371/journal.pone.0169984> (2017).
47. Cavarretta, I. *et al.* The Microbiome of the Prostate Tumor Microenvironment. *Eur. Urol.* **72**, 625–631 (2017).
48. Sarrias, M. R. *et al.* CD6 binds to pathogen-associated molecular patterns and protects from LPS-induced septic shock. *Proc. Natl. Acad. Sci. USA* **104**, 11724–11729 (2007).
49. Triantafyllou, M. & Triantafyllou, K. Lipopolysaccharide recognition: CD14, TLRs and the LPS-activation cluster. *Trends Immunol.* **23**, 301–304 (2002).

50. Aldinucci, D. & Colombatti, A. The inflammatory chemokine CCL5 and cancer progression. *Mediators Inflamm.* **2014**, 292376, <https://doi.org/10.1155/2014/292376> (2014).
51. Vaday, G. G., Peehl, D. M., Kadam, P. A. & Lawrence, D. M. Expression of CCL5 (RANTES) and CCR5 in prostate cancer. *Prostate* **66**, 124–134 (2006).
52. Xu, J. *et al.* CSF1R signaling blockade stanches tumor-infiltrating myeloid cells and improves the efficacy of radiotherapy in prostate cancer. *Cancer Res.* **73**, 2782–2794 (2013).
53. Öhlund, D., Elyada, E. & Tuveson, D. Fibroblast heterogeneity in the cancer wound. *J. Exp. Med.* **211**, 1503–1523 (2014).
54. Levesque, C. & Nelson, P. S. Cellular constituents of the prostate stroma: Key contributors to prostate cancer progression and therapy resistance. *Cold Spring Harb. Perspect. Med.*, <https://doi.org/10.1101/cshperspect.a030510> (2017).
55. Park, J. E. *et al.* Fibroblast activation protein, a dual specificity serine protease expressed in reactive human tumor stromal fibroblasts. *J. Biol. Chem.* **274**, 36505–36512 (1999).
56. Brennen, W. N., Isaacs, J. T. & Denmeade, S. R. Rationale behind targeting fibroblast activation protein-expressing carcinoma-associated fibroblasts as a novel chemotherapeutic strategy. *Mol. Cancer Ther.* **11**, 257–266 (2012).
57. Fang, M., Yuan, J., Peng, C. & Li, Y. Collagen as a double-edged sword in tumor progression. *Tumour Biol.* **35**, 2871–2882 (2014).
58. Ameye, L. *et al.* Abnormal collagen fibrils in tendons of biglycan/fibromodulin-deficient mice lead to gait impairment, ectopic ossification, and osteoarthritis. *FASEB J.* **16**, 673–680 (2002).
59. Jacobsen, F. *et al.* Up-regulation of biglycan is associated with poor prognosis and PTEN deletion in patients with prostate cancer. *Neoplasia* **19**, 707–715 (2017).
60. He, Y. *et al.* Matrix-remodeling associated 5 as a novel tissue biomarker predicts poor prognosis in non-small cell lung cancers. *Cancer Biomark.* **15**, 645–651 (2015).
61. De Vivan, A. D. *et al.* Histologic features of stromogenic carcinoma of the prostate (carcinomas with reactive stroma grade 3). *Hum. Pathol.* **63**, 202–211 (2017).

### Acknowledgements

This research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 758306), Norwegian University of Science and Technology (NTNU), the Liaison Committee between the Central Norway Regional Health Authority (RHA) and NTNU, Norwegian Cancer Society and The Northern Health Administration, UiT - The Arctic University of Norway. All tissue samples were collected and stored by Biobank1, St. Olav's Hospital. HR-MAS MRS were performed at the MR Core Facility, NTNU. RNA microarray measurements the Genomics Core facility, NTNU, and Norwegian Microarray Consortium (NMC), a national platform supported by the functional genomics program (FUGE) of the research council of Norway. We would like to thank Alan Wright for quantification of metabolites with LCMoDel and Trond Viset for histopathological evaluation.

### Author Contributions


M.K.A., G.F.G., T.E.B., M.B.R. and M.-B.T. contributed to the design of the study. H.B. and M.-B.T. developed and performed wet lab experiments. Histopathology evaluations were performed by E.R. and Ø.S. Data analysis was performed by M.K.A. and K.R. with the guidance of G.F.G., M.B.R. and M.-B.T. The manuscript was written by M.K.A., and all authors edited and approved the final manuscript.

### Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-32549-1>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018



## METHOD

# FunHoP: Enhanced Visualization and Analysis of Functionally Homologous Proteins in Complex Metabolic Networks

Kjersti Rise <sup>1,\*</sup>, May-Britt Tessem <sup>2</sup>, Finn Drablos <sup>1</sup>, Morten B. Rye <sup>1,3,\*</sup><sup>1</sup> Department of Clinical and Molecular Medicine, NTNU – Norwegian University of Science and Technology, Trondheim NO-7491, Norway<sup>2</sup> Department of Circulation and Medical Imaging, NTNU – Norwegian University of Science and Technology, Trondheim NO-7491, Norway<sup>3</sup> Clinic of Surgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim NO-7491, Norway

Received 3 August 2018; revised 8 May 2019; accepted 18 August 2019

Available online xxx

Handled by Henning Hermjakob

## KEYWORDS

Homologous proteins;  
Metabolic network;  
Pathway visualization and analysis;  
RNA-seq;  
KEGG;  
Cytoscape

**Abstract** Cytoscape is often used for visualization and analysis of metabolic pathways. For example, based on KEGG data, a reader for KEGG Markup Language (KGML) is used to load files into Cytoscape. However, although multiple genes can be responsible for the same reaction, the KGML-reader KEGGScape only presents the first listed gene in a network node for a given reaction. This can lead to incorrect interpretations of the pathways. Our new method, FunHoP, shows all possible genes in each node, making the pathways more complete. FunHoP collapses all genes in a node into one measurement using read counts from RNA-seq. Assuming that activity for an enzymatic reaction mainly depends upon the gene with the highest number of reads, and weighting the reads on gene length and ratio, a new expression value is calculated for the node as a whole. Differential expression at node level is then applied to the networks. Using prostate cancer as model, we integrate RNA-seq data from two patient cohorts with metabolism data from literature. Here we show that FunHoP gives more consistent pathways that are easier to interpret biologically. Code and documentation for running FunHoP can be found at <https://github.com/kjerstirise/FunHoP>.

## Introduction

Metabolic pathway analysis is a common framework for interpreting large-scale omics data and revealing functional trends and patterns in known biological multi-gene pathways. Important curated resources of metabolic pathways are the Kyoto Encyclopedia of Genes and Genomes (KEGG) [1,2],

\* Corresponding authors.

E-mail: [kjersti.rise@ntnu.no](mailto:kjersti.rise@ntnu.no) (Rise K), [morten.rye@ntnu.no](mailto:morten.rye@ntnu.no) (Rye MB).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

<https://doi.org/10.1016/j.gpb.2021.03.003>

1672-0229 © 2021 The Author. Production and hosting by Elsevier B.V. on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).Please cite this article as: K. Rise, M. B. Tessem, F. Drablos et al., FunHoP: Enhanced Visualization and Analysis of Functionally Homologous Proteins in Complex Metabolic Networks, Genomics Proteomics Bioinformatics, <https://doi.org/10.1016/j.gpb.2021.03.003>

Reactome [3], Panther [4], and similar knowledge bases [5]. Such resources are increasingly integrated with other knowledge bases, as can be seen for example for KEGG [6]. Several approaches can be used for analyzing metabolic pathways in the context of general network representations [7], and recent tools like eXamine [8] and Orthoscape [9] are relevant examples. For transcriptomics, an often-used approach is to map differentially expressed genes (DEGs) to known biological pathways, for example from KEGG. Such pathway representations can then be analyzed and visualized with commercial tools like Pathway Studio ([www.pathwaystudio.com/](http://www.pathwaystudio.com/)) or iPathwayGuide ([www.advaitabio.com/ipathwayguide.html](http://www.advaitabio.com/ipathwayguide.html)), or free tools like CellDesigner [10] or Cytoscape [11].

In these tools, metabolic pathways are generally displayed as a network of metabolic transitions, where each transition is associated with a node representing the enzyme responsible for the transition. Each node typically represents a separate child from a structured pathway file, such as XML format. However, a challenge occurs when a transition from one metabolite to another can be catalyzed by more than one possible enzyme, *i.e.*, by functionally homologous protein families, or functional homologs [12]. This is best illustrated by a typical example from KEGG. In the histidine metabolic pathway (KEGG: hsa00340), the four paralogs of NAD(P)<sup>+</sup> dependent aldehyde dehydrogenase (*ALDH3A1*, *ALDH1A3*, *ALDH3B1*, and *ALDH3B2*, KEGG node index 1.2.1.5, Figure 1A) can all catalyze the transition from methylimidazole acetaldehyde to methylimidazole acetate. However, KEGG displays only the first gene, *ALDH3A1*, both in the website and in the XML file. In the website, the user can hover the mouse pointer over the gene in question to see any functional homologs, and the XML file does contain the KEGG IDs to all of them, although the corresponding gene names are not available in the file. In most conditions and cell types, one of these paralogs might be the preferred for the enzymatic transition, but in certain conditions one or several of the other three paralogs may become important, which should be taken into account. Though the selected example contains only four paralogs, the number of alternative enzymes can exceed 30 for some transitions, which complicates both visualization and interpretation of such nodes in the current framework. An example of a large node is the *PLA2G4B* node with 21 genes shown Figure 1B. In particular, the conclusion as to whether a node is overall up- or down-regulated will depend on the degree of differential expression of each gene (fold change and/or *P* value), the relative expression level of each gene in the node, and the enzymatic efficacy of the protein. The challenges regarding nodes with multiple genes are thus twofold. First, there is a need for data that can help us identify the most important enzyme(s) in conditions where multiple genes are able to perform the same reaction. Second, there is a need for improved visualization strategies to convey the relative importance of different enzymes with overlapping function when viewing biological networks from databases such as KEGG.

Cytoscape is a common tool for pathway visualization and analysis, often with data from KEGG. Pathways of choice can be downloaded from KEGG as KGML XML files (KEGG Markup Language, in XML format) and imported into Cytoscape using one of the many apps, such as KEGGScape [13]. In Cytoscape, the user can define styles, highlight nodes and/or edges, or change properties (*e.g.*, color, thickness, or shape of both nodes and edges based on uploaded data, such as gene

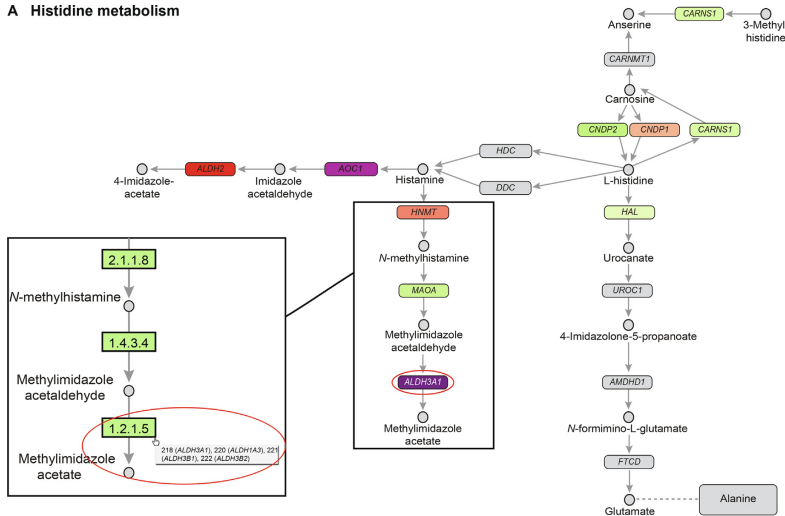
expression or protein data). Layouts, statistical analyses, or specific apps with certain abilities can be applied to analyze the network in question. Importing the pathway is a crucial part of the analysis. The limitation in the KEGG XML files and/or KEGGScape of only showing the first of potentially multiple genes in each node has consequences for both analysis and interpretation (Figure 1), since the missing expression data of the remaining genes in the node makes it impossible to conclude on the overall gene expression associated with each node. It would be a huge advantage if one could expand the analysis to include differential expression of all genes in a node, and visualize the expression levels and associated differences for nodes consisting of multiple genes. This can be used to conclude on the overall up- or down-regulation at the node level, and suggest which gene(s) in the node that may have the largest influence on the overall activity.

Other options for importing KEGG XML files are CyKEGGParser [14] and CytoKEGG (<http://apps.cytoscape.org/apps/cytokegg>). CyKEGGParser discusses the topic of paralogs being grouped into single nodes, and their solution is to create new separate nodes for each of the genes within a multi-gene node. CytoKEGG is used to search and import KEGG pathways into Cytoscape. Dealing with multiple genes in the same node has also been discussed by others in a non-Cytoscape related context. The Bioconductor package Graphite [15] converts pathway topology to gene networks, and uses a combination of data from three curated databases (KEGG, Reactome, and BioCarta/NCI/NPID [16]) to create more complete networks. For the pathways from KEGG, Sales et al. [15] discuss how nodes with multiple genes may represent two different types of groups: protein complexes ("AND groups", all genes should be considered together) or alternative proteins for the same function (functional homologs; "OR groups", considering one gene at the time). This second group (OR) can be expanded into pathways without any connections between the alternative genes/proteins. In another publication, Wang et al. [17] acknowledge nodes with multiple genes by coloring the same node with multiple colors representing the different gene expression values. In addition, the number of genes in each node is displayed next to it. Although this approach can work for nodes with a limited number of genes, it will become harder to interpret when the number of genes increases. Additionally, neither of these approaches show the expression level for each gene, which can help to identify the genes that are most likely to be responsible for the reaction in a given node.

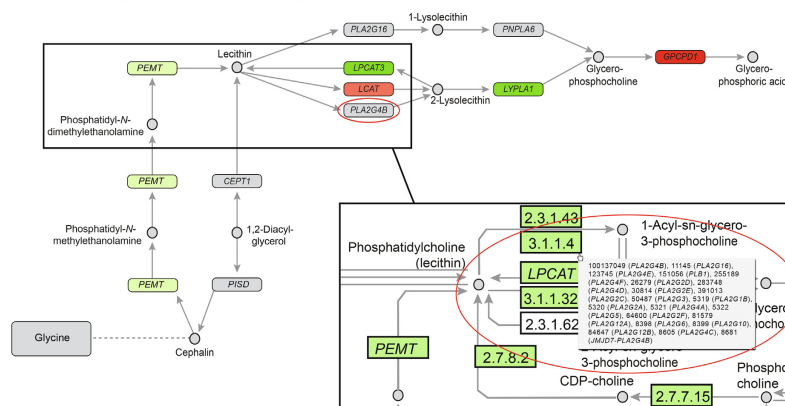
In nodes with multiple functional homologs, the relative expression levels of the genes in a node can be an accessible and useful measure to assess the relative importance of the individual enzymes for a given condition. For microarrays, the previous golden standard for gene expression analysis, differences in probe-affinities made it difficult to assess the relative expression levels between genes in an experiment [18]. However, the replacement of microarrays by RNA sequencing (RNA-seq) has now made comparison of expression levels feasible [18–21]. Data from RNA-seq could therefore be utilized to improve the analysis of the overall node activity, as well as the individual contribution of each gene in the node for a given metabolic pathway.

Here we present Functionally Homologous Proteins (FunHoP): a method to improve gene expression pathway analysis and visualization. FunHoP improves the network visualization

## A Histidine metabolism



## B Glycerophosphocholine metabolism (part)



**Figure 1 Comparison of pathway XML files in Cytoscape to the same pathways in KEGG**

**A.** A schematic of histidine metabolism pathway. All nodes in the original Cytoscape display show one single gene, including the *ALDH3A1* node. The *ALDH3A1* node from KEGG actually contains four genes: *ALDH3A1*, *ALDH1A3*, *ALDH3B1*, and *ALDH3B2*. **B.** A schematic of glycerophosphocholine metabolism pathway (part). The *PLA2G4B* node contains 21 genes, despite only showing one in KEGG.

and analysis with respect to differential expression of nodes with multiple genes, and the relative contribution of each gene in a node. In particular, FunHoP aggregates gene information for each KEGG node consisting of multiple genes by using

RNA-seq gene expression data for each gene, assuming that genes in the same node represent overlapping enzymatic potential (*i.e.*, functional homologs). We show that prioritizing genes based on read counts from RNA-seq will improve the

interpretation of differential expression results when analyzed with KEGG metabolic pathways. By gaining information from multiple genes for each node as input for differential expression analysis, we receive more biologically relevant and reliable pathways. Using prostate cancer (PCa) as a model system, we present two case studies showing how gene expression data are able to explain previously observed metabolic changes when FunHoP is applied.

## Method

RNA-seq data for PCa (read counts and gene identifiers) were downloaded from The Cancer Genome Atlas (TCGA) [22] at <https://portal.gdc.cancer.gov/repository>. For the Prensner cohort [23], RNA-seq raw reads in *fastq*-format were downloaded with approval from The database of Genotypes and Phenotypes (dbGap: phs000443.v1.p1, project #5870) at [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000443.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000443.v1.p1).

Raw RNA-seq reads were mapped to the hg19 transcriptome using TopHat2 [24], and featureCounts [25] was used to assign the reads to each gene. Voom [26] was further used for differential expression analysis. DEGs with a  $P$  value below 0.05 were extracted, and  $P$  values were  $\log_2$  transformed by:

$$\text{Value} = \log_2 P \text{ value} \times (-10) \times (\text{regulation}) \quad (1)$$

where regulation was defined as 1 for upregulated genes (positive fold-change) and  $-1$  for downregulated genes (negative fold-change). Average RNA-seq read count for each gene was calculated using the mean of the two average values calculated over cancer and normal samples, respectively. All read counts were adjusted for gene lengths by a factor estimated by taking the gene length of the respective gene (sum of exons) divided by the average gene length over all genes.

In this study, 85 pathways of relevance to human metabolism, from subcategories 1.1 up to and including 1.11, were downloaded with human genes from the KEGG pathway database [27]. 71 of these did not contain any “line” nodes, and were used further (see Tables S1 and S2). The initial pathway analysis was performed by loading original KEGG XML files into Cytoscape (v. 3.4.) via the KEGGScape app and using a color gradient based on differential expression. All displays of differential expression used the same gradient: values were found on a scale from  $-1200$  (black) to  $600$  (dark green), via  $-600$  (purple),  $-300$  (bright red),  $0$  (light yellow), and  $300$  (bright green). All values below zero showed downregulated gene expression, and all values above zero showed upregulated gene expression.

To expand the XML files to show all the genes in all the nodes, the list of human IDs and corresponding gene names was downloaded [28]. Using the ElementTree XML API, name strings in nodes with more than one gene were extended to include only the human names for all these genes (File S1). KEGG’s solution to protein complexes was used as a base, and nodes with more than one gene were expanded. The expanded nodes were made by creating a new child for each of the genes that were not included in the initial child, and combining the new children along with the old child in a common node. The gene nodes use the same coordinates as the original gene, making it appear in the same place. To distinguish gene nodes from protein complexes, the gene nodes were

made bigger than the default size, giving them a white field on each side. Differential gene expression was first used in combination with the expanded networks, showing how all the genes in the pathway were expressed.

To make more interpretable networks yet containing all the information, all genes within a node were aggregated into one. Name strings were extracted from all the network files, and the lists of unique names were defined as unique nodes. These included both single-gene and multiple-gene nodes. All gene nodes were named on the form “gene1-Bx”, where “gene1” is the name of the first gene in the gene-name string for a given node, and “x” is the number of genes within the node. For single-gene nodes  $x$  is 1. The total read count for a node was found by adding the read counts for all genes in the node, and this value was used for differential expression analysis at the node level. The contribution of each gene to expression level within a node was calculated as the fraction of the read count for that gene to the total read count of the node. The read count for each gene was used to style for expanded networks according to the relative expression levels of the genes. Nodes were colored on a scale from 0 to  $\geq 50,000$  read counts, changing from white to dark blue via shades of pink and blue.

The aggregated network files were adapted to work with the aggregated network gene node names. Changing the name strings to reflect the first gene name and the number of genes made the string similar to the gene node name format, and the network files could again be used together with the output from differential expression. The previously used style for differential expression was again used for the aggregated networks.

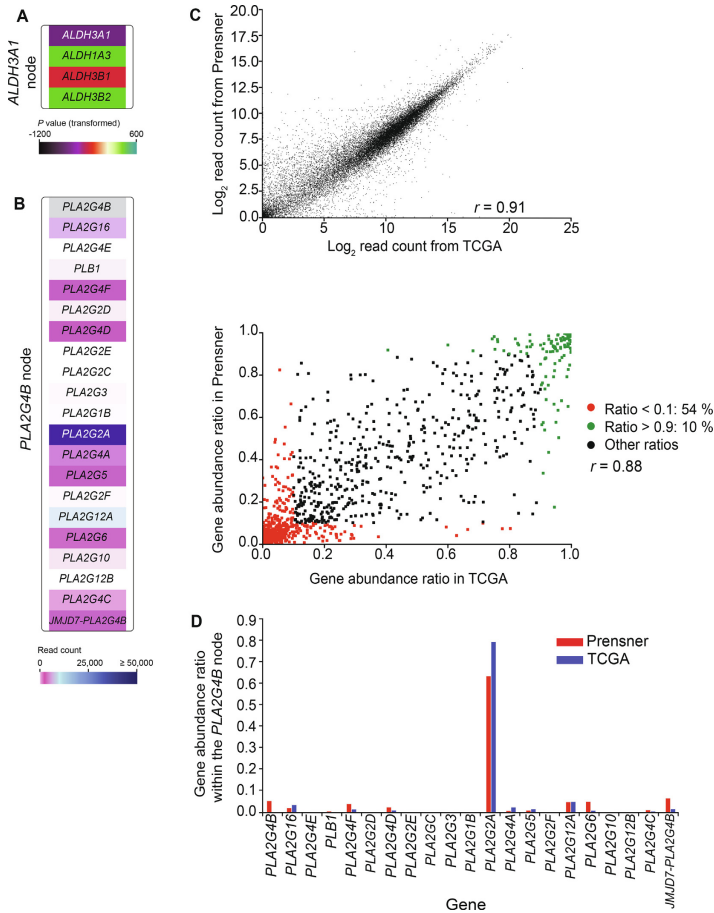
To show that the method works, two case studies were performed: the histidine metabolism pathway and a minor part of the metabolic pathway for glycerophosphocholine (GPC). The original files from KEGG were run through FunHoP’s steps of creating expanded and aggregated networks, as explained above, analyzed with differential expression, and visualized as expanded networks at the gene level and aggregated networks at the node level.

## Results

### Expanding nodes and using RNA-seq counts to improve pathway analysis

To visualize KEGG pathways using information from all the individual genes involved, each node containing multiple functional homologs was expanded to show all genes in the node. Nodes were expanded by adding a new child for each gene belonging to the node, in addition to the existing child representing the default gene displayed in the pathway from KEGG in Cytoscape by KEGGScape. Old and new children of a node were then connected in a type “group” child, using the same strategy as for protein complexes (AND groups). To visually distinguish nodes with functional homologs (OR) from protein complexes (AND), the nodes were made bigger than the default size, giving them a white border on each side (Figure 2A).

We then used the average RNA-seq read count for each gene (normalized against gene length), generated from patient samples in two available PCa cohorts [22,23]. Aggregated average read counts for all genes in each node were used to define the



**Figure 2** Validation of the FunHoP approach

**A.** The *ALDH3A1* node expanded to show individual genes styled by differential gene expression. **B.** The expanded *PLA2G4B* node with its 21 genes. The *PLA2G4B* gene itself is not found in the dataset, leading to the whole node to be seen as not significant when only *PLA2G4B* is shown, although the other hidden genes are significantly differentially expressed. **C.** Plots of log<sub>2</sub> read count from TCGA and Prensner (top) and gene abundance ratio in TCGA vs in Prensner (bottom). **D.** Gene abundance ratios within the *PLA2G4B* node from Figure 2B are comparable between the TCGA and Prensner cohorts.

total expression level of each node in the network. Moreover, the relative read count for each gene in a node divided by the total read count for the node was used to define the relative expression contribution from each gene in a node.

To show the effect of FunHoP, original pathways were color-coded according to log-transformed *P* values from differential gene expression analysis, here comparing PCa tissue

with normal prostate tissue. For showing individual genes within an expanded node, each gene was color-coded by both *P* values and the average read count for the gene to indicate expression level, giving two expanded networks that were comparable. The final representation shows the network with aggregated nodes, color-coded by differential expression based on overall read counts within a node.



### Assumptions regarding gene families and expression levels

We introduce two important assumptions for the biological interpretations in FunHoP. First, we observe that genes assigned to the same node usually belong to the same functional gene family or are closely related, as in the case of the nodes for the aldehyde dehydrogenases (Figure 2A) and phospholipases (Figure 2B). Thus, we make the assumption that the gene products also have similar function, in particular that they are able to catalyze the same main reaction, and describe them as enzymes with homologous function, or functional homologs. Therefore, we assume that each homolog can catalyze the reaction at a comparable rate. This is obviously an oversimplification, but also a necessary simplification given the general lack of rate data for most cellular processes.

Second, we assume that read counts from RNA-seq are indicative of the relative expression level of genes within a sample cohort. To check this assumption, we used RNA-seq read counts from two independent datasets. We see that the gene expression levels based on RNA-seq read counts are highly correlated (Figure 2C). We also find that expression ratios for individual genes in a node are correlated (Figure 2D). In particular, there is a very good correspondence for genes having particularly high ( $>0.9$ ) or low ( $<0.1$ ) ratios, which shows that RNA-seq data can robustly identify genes with a very high or very low relative abundance. This pattern is also evident when looking at individual genes within a node with high complexity, as the ratio for each gene within the node follows the same trend independent of which dataset we used (Figure 2D). The highly expressed *PLA2G2A* is clearly dominant in both datasets, the genes with very low number of read counts are the same, and the genes identified with few and intermediate number of read counts are also the same, though the relative ratios vary somewhat among the intermediate genes in the two datasets.

Under these assumptions, a gene's contribution to the overall node activity is proportional to its expression level. This information becomes particularly useful in situations where one specific gene is dominating within a node. An example of this is the *PLA2G4B* node in the glycerophospholipid metabolism pathway (KEGG: hsa00564). The current Cytoscape/KEGGscape/KEGG framework only shows *PLA2G4B*, which is not found in the TCGA dataset, and hence the node seems to be not significant in the pathway. When the node is expanded, we see all 21 genes or functional homologs. By comparing the read counts for each gene, we see how *PLA2G2A* is expressed at a level that is ten times higher than the second one on the list (Figure 2B). Here, the darkest blue corresponds to  $\geq 50,000$  read counts, whereas the white/pink/light blue corresponds to  $< 5000$  read counts. The genes indicated in light pink have  $< 10$  read counts, and the ones in white are not expressed. These genes will most likely not contribute significantly to the pathway in this case. The KEGG default gene *PLA2G4B* is not found in the TCGA dataset, and has a low expression in the Prensner dataset. In this case, it is reasonable to assume that *PLA2G2A* is the main driving force for the transition represented by the node.

### Case studies

To investigate the impact of FunHoP on real biological interpretation of networks, we used PCa as a model system for two

case studies. Metabolic studies have identified significant changes in metabolites in both histidine and glycerophospholipid metabolism pathways, but gene expression changes in the original network models were unable to explain the observed metabolic differences. Our aim was to investigate if FunHoP could identify the possible changes in expression levels leading to the observed changes in metabolites. The dataset from TCGA was further used in the following case studies due to its high number of samples and thereby statistical power.

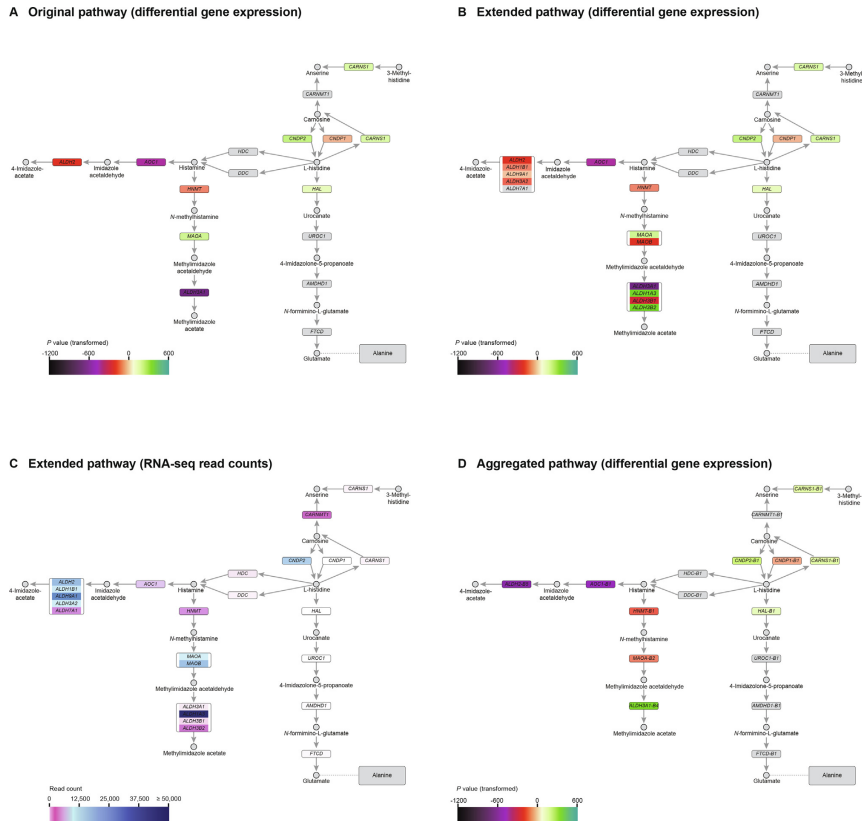
#### Case study 1: histidine metabolism

The first case study looks at the histidine metabolism pathway. It has been shown that histidine is elevated in PCa compared to normal prostate tissue [29]. This elevation cannot be explained by differential changes in gene expression using the original pathway (Figure 3A). In the original pathway, histidine is produced from carnosine in two paths, by *CNDP2* or *CNDP1*. Histidine can then be converted back to carnosine through a loop by *CARNS1* (Figure 3A). Looking at the *P* values for the respective genes shows that *CNDP1* is downregulated, and *CNDP2* and *CARNS1* are upregulated with *P* values within the same order of magnitude. Moreover, of the genes in the other paths leading away from histidine, *HAL* is upregulated while *HDC* and *DDC* are unchanged (Figure 3A). Overall, this pathway is not compatible with the observed increase in histidine levels in PCa.

However, when using the FunHoP-expanded pathway to visualize all genes and nodes, we can see how the original pathway is an oversimplification of a more complex pathway including three nodes with multiple genes (Figure 3B). Many of these nodes have genes that are up- or down-regulated. As there are no functional homologs in the nodes most directly linked to histidine, expanding nodes alone does not lead to any improved interpretation in this case. However, when RNA-seq read counts are shown together with differential expression in the extended pathway, we are able to provide a possible explanation as to how histidine level may be elevated (Figure 3C).

For the paths leading to histidine synthesis, the most dominant gene in read counts is *CNDP2*, which is upregulated and has about 11,000 reads. Upregulation of *CNDP2* pushes carnosine conversion to histidine. The downregulated *CNDP1* has close to zero read counts and can be ignored. *CARNS1*, responsible for the loop back towards carnosine, has less than 100 reads, and is probably less influential than *CNDP2*. We can therefore assume that upregulation of the highly expressed *CNDP2* most likely leads to increased production of histidine. For the paths leading away from histidine, all genes in the path leading towards glutamate (including the upregulated *HAL*) have close to zero read counts, and can be ignored. With *HDC* and *DDC* remaining unchanged, there is no net change in histidine consumption. Increased histidine production through the highly expressed *CNDP2* combined with ignorable changes in histidine consumption, leads to a possible explanation for how histidine accumulates. Moreover, the genes further downstream of histamine (i.e., *HNMT* and *AOCT*) are downregulated with higher read counts (2519 and 3636 read counts, respectively), creating a bottleneck in the influx/efflux balance, which can lead to further increase in histidine levels. The overall read counts in the pathway seems to push towards





**Figure 3 Pathway of histidine metabolism**

A. Original pathway colored by differential gene expression on a log-scale. B. Expanded pathway colored by differential gene expression. C. Expanded pathway colored by RNA-seq read counts. D. Aggregated pathway colored by differential gene expression at the node level.

accumulation of histidine, which is not used further downstream in any direction, allowing a build-up of histidine to happen. The histidine pathway also shows examples of nodes with high difference in read counts between genes in the node. One example is the *ALDH3A1* node, where *ALDH3A1* dominates with 46,595 read counts, while the three remaining genes have less than 1000 reads each. This further strengthens the idea that the differential expression of the dominant gene will determine the overall expression of the node.

The conclusions from the expanded network are also evident in the aggregated network at the node level (Figure 3D), where *CNDP2* is clearly highlighted, especially when looking at the pathway styled with read counts. The aggregated network shows how nodes that appear to be upregulated in the

original network are shown to be downregulated, and vice versa. Overall, FunHoP provides a more complete pathway analysis, and is able to give a more precise explanation on how histidine can be elevated in PCA.

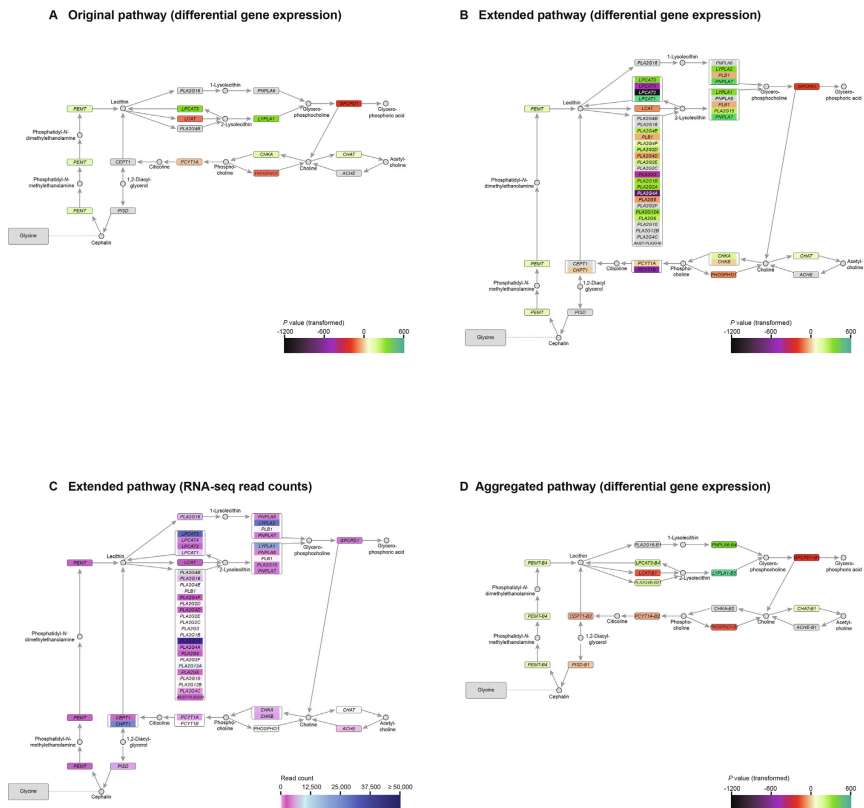
#### Case study 2: glycerophospholipid metabolism

The second case study looks at part of the glycerophospholipid metabolism. The complete pathway is extensive and contains several complex nodes with up to 21 genes, which makes visualization and analysis challenging. Previous studies have shown elevated levels of GPC in PCA [30–32], and the original Cytoscape network from KEGG colored by differential expression is shown in Figure 4A.

The initial pathway does not provide an explanation to how GPC can be elevated based on differential gene expression values. In the reaction paths leading from lecithin towards GPC, three displayed genes are not significantly differentially expressed (*PLA2G16* and *PNPLA6* via 1-lysolecithin, and *PLA2G4B* towards 2-lysolecithin), along with one upregulated gene functioning in the opposite direction (upregulated *LPCAT3* from 2-lysolecithin back to lecithin). This indicates that even if the conversion from 2-lysolecithin to GPC is upregulated by *LYPLA1*, the reaction is just as much pushed away from 2-lysolecithin and back towards lecithin by *LPCAT3*, instead of towards GPC. Overall, this does not explain how GPC can be accumulated. However, when expanding the network, a more complex picture emerges, with

more genes involved in several nodes and huge differences in RNA-seq read counts among genes and nodes (Figure 4B and C). The expanded networks provide a different interpretation of several nodes in the pathway.

A particularly complex node of 21 genes appears in the original *PLA2G4B* node. This node contains both non-significant genes and up-/down-regulated genes, with average RNA-seq read counts varying over many orders of magnitude. The clearly dominant gene is *PLA2G2A* with 71,482 read counts (Figure 4C), which is also upregulated (Figure 4B). Since none of the other genes have read counts of comparable magnitude (the second highest is *PLA2G12A* with 4502 read counts), we see how this node changes from non-significant in the original network to upregulated in the aggregated network (Figure 4D). The other nodes in the paths leading



**Figure 4** Pathway of glycerophospholipid metabolism (part)

A. Original pathway colored by differential gene expression on a log-scale. B. Expanded pathway colored by differential gene expression. C. Expanded pathway colored by RNA-seq read counts D. Aggregated pathway colored by differential gene expression at the node level.

to production of GPC also show both up- and down-regulated genes. The most dominant gene in the *PNPLA6* node is *LYPLA2* (Figure 4C), which is upregulated (Figure 4B). For the *LYPLA1* node, the dominant gene is *LYPLA1* (Figure 4C), which is upregulated as in the original network (Figure 4A and B). Both dominant genes lead to their respective nodes being upregulated in the aggregated network (Figure 4D), enabling two possible paths towards production of GPC. Both paths via 1-lysolecithin and 2-lysolecithin, respectively, are now upregulated. Even though *PLA2G16* in the 1-lysolecithin path is not significant, *PLA2G16* still has 3246 reads, which indicates a flux through this path. For the path via 2-lysolecithin, the upregulation of both the *PLA2G4B* and *LYPLA1* nodes reveals an unambiguous upregulated path from lecithin to GPC. Though the *LPCAT3* node looping backwards upstream of GPC is also upregulated, the pathway as a whole shows a net unambiguous flux towards GPC through several possible paths, explaining why GPC would accumulate in PCa. The expanded network also shows that *LCAT* (the sole gene of the *LCAT* node) has fewer reads than the *PLA2G4B* and *LPCAT3* nodes, making its downregulation less important (Figure 4C).

Alongside providing more biological information, the GPC example also illustrates the possible complexity of nodes. With the full pathway having four highly complex multi-gene nodes of 21 genes, as well as several nodes with 4–6 genes, we see how difficult pathways can be to interpret. Using FunHoP, we show how we can gain important additional information by expanding the networks to show all genes, and by looking at differential expression and gene expression level simultaneously for network interpretation.

In order to validate our conclusions on the two case studies using data from TCGA, we performed the same analysis with the data from the Prensner cohort. Due to the generally smaller number of samples in the Prensner data, in addition to lower sequencing depth, many of the significant changes observed in TCGA were not statistically significant in Prensner. However, the overall patterns are also evident in both case studies, both in terms of the dominant genes within the pathway and the differential expression (Figures S1 and S2), which supports the conclusions on which genes can contribute to the elevated levels of histidine and GPC in PCa.

## Discussion

Metabolic pathway analysis is an important approach for analyzing gene expression. With the constantly growing amount of available data, we can improve our understanding of the complexity in biological systems, and continuously develop models to capture and utilize new data and information. However, the most commonly used pathway representations from databases and associated tools often give a simplified picture of metabolic pathways, focusing on only one gene in each network node, despite the fact that more genes may be able to perform the same enzymatic reaction. One example, which we have focused on in this study, is the current integration of KEGG and Cytoscape using KEGGScape.

We have therefore implemented a strategy for including all functional homologs of a gene in the analysis, based on the following assumptions:

First, we have to assume that the relevant genes in an expanded node indeed are functional homologs, *i.e.*, with similar function. KEGG networks are manually curated, and documentation can be found within KEGG for genes, compounds, and reactions. When KEGGScape places a gene within a certain node, we assume that this gene is able to produce an enzyme that can catalyze the transition represented by the node. In FunHoP, we have implicitly made an assumption that the different genes within a node representing an enzymatic reaction also catalyze the reaction at a similar rate. This is a simplification, and to model the enzyme activity one should ideally also include enzyme efficiency and kinetics for the given situation. However, data on enzyme kinetics are usually not available, or very hard to obtain. We believe that our assumption on the enzyme activity correlating with expression level is at least reasonable for differences spanning several orders of magnitude, and represents a model improvement compared to networks where expression levels are not considered at all. Supporting this assumption is the observation that genes in a node usually belong to the same gene family. For example, for the node in the histidine metabolism pathway with *ALDH3A1* on top, all the other genes are aldehyde dehydrogenase paralogs that are able to catalyze the same reaction (Figure 2A).

Secondly, we have to assume that we actually can estimate relative expression levels of relevant genes. With microarrays being the previous gold standard to measure changes in gene expression, differential expression analysis and subsequent network mapping were limited to fold changes and *P* values. Variations in probe affinities made it difficult to assume anything about the real expression level differences between genes. However, with RNA-seq, one should be able to provide relative expression level measurements with much improved correlation to the real relative mRNA levels compared to microarrays.

Using the two assumptions on relative expression levels and similarity in enzyme efficiency described above, we can predict which of the genes is/are most likely to be responsible for a given reaction in a node. Especially for cases where the read-count difference for two genes in the same node spans several orders of magnitude, we find it likely that difference in expression level will take precedence over reaction efficiency. We have shown that read counts are highly reproducible for two independent patient cohorts for PCa. We observe that many pathway nodes typically consist of one or a few dominant genes in terms of expression level, supporting our claim that this is a highly relevant measure to include when evaluating the contribution from different enzymes in a node. For the single-gene nodes, the approach of looking at absolute gene expression can also reveal patterns in the pathways that are not evident from comparing *P* values alone. By using read counts, we are also capable of determining whether some paths are turned completely off, as in the case for the path leading from histidine to glutamine (Figure 3C).

A possible limitation of our approach is to which degree tissue-specific isoforms affect enzymatic activity and estimated expression levels of the genes represented in the nodes. Not all isoforms of a gene are necessarily enzymatically active. However, KEGG does not currently provide curated information on enzyme activity of isoforms. We have thus limited analysis to the gene level. However, an expansion to isoforms is concep-

tually possible within the FunHoP framework if such data become available. Another isoform related limitation is that genes with particularly short or long dominant isoforms compared to the canonical isoform model may lead to aberrant expression level estimation for the genes affected. In addition, tissue-specific isoform switches can potentially affect results from differential expression analysis. In this study, we have assumed that genes are presented by their canonical isoform.

The starting point for network analysis is usually an expression table with samples and genes, which for RNA-seq is presented as a table of read counts. It is thus preferable that the network analysis is reproducible with respect to RNA-seq RNA selection protocols, sequence length, library size, choice of alignment, and mapping tools. It was not possible to systematically investigate many settings (mostly due to lack of available data on prostate), but we demonstrate that the results are reproducible in two independent PCa cohorts with different properties. Both cohorts use poly-A selection of transcripts, but differ in sequence length, library size, and alignment/mapping tools.

We also assume that changes in transcript level are informative about changes in protein level. It is well known that a direct association between mRNA expression level, protein level, and subsequent protein activity is inaccurate, for example because of the effects of post-transcriptional and post-translational regulation of proteins on enzyme kinetics; however, the reasons in most cases are unknown. We cannot say with absolute certainty that an upregulated pathway with multiple read counts will result in a similar increased number of metabolites. A study by Schwanhäusser et al. [33] shows a correlation between mRNA and protein copy numbers in NIH3T3 mouse fibroblasts, which was found to be 0.41. When considering translation rate constants, the correlation went up to 0.95.

Other studies in different organisms have also shown correlations, although this is organism dependent [32,34]. FunHoP does not pretend to describe the complete picture, but still represents a significant improvement compared to analyses where all genes are assumed to have the same expression level, or where multiple genes in the same node are not taken into account at all.

In the KEGG database, histidine is also involved in two other pathways that can affect the overall levels of this metabolite. In aminoacyl-tRNA biosynthesis (KEGG: hsa00970), histidine is converted to L-histidyl-tRNA(his), catalyzed by *HARS* and *HARS2*. Neither of these genes show significant changes, which indicates that this does not affect the level of histidine between the samples. In beta-alanine metabolism (KEGG: hsa00910), histidine is involved in the same step as the one in our case study, although we here see a more complete picture of carnosine being converted into histidine and beta-alanine. This is performed by the same enzymes as in the case study (*CNDP1/CNDP2*). As we know, *CNDP1* is downregulated and has close to zero read counts, and *CNDP2* is upregulated with 11,217 read counts. This should indicate that beta-alanine is also elevated in PCa, which was confirmed by the same study [29]. Overall, we see how the case study provides a possible explanation on how histidine can be elevated in PCa, and our solution also fits with other available measurements of related metabolites [29].

GPC is also involved in another pathway: ether lipid metabolism (KEGG: hsa00565), where GPC can also be produced

by conversion of 1-(1-alkenyl)-sn-glycero-3-phosphocholine by *TMEM86B*. However, this gene does not show a significant change between PCa tissue and normal tissue, and hence we can explain the elevated levels of GPC by the extracted part of glycerophospholipid metabolism shown in the case study. Another possibility for GPC to be elevated using the original network would be if the level of 2-lysoecithin was high, the upregulated *LYPLA1* converted 2-lysoecithin to GPC. To our knowledge, 2-lysoecithin has not been documented as high in PCa.

Choline metabolism in PCa is a well-studied topic, especially in regard to relevant metabolites and identification of potential biomarkers [30–32,35,36]. The pathways involved in the metabolism are still not fully covered, and our findings from case study #2 are therefore of special interest. These results will be focused on in later studies.

The current version of FunHoP supports the human metabolic pathways found in KEGG, with exception of the glycan-related pathways (mostly found in “Glycan biosynthesis and metabolism”, category 1.7), which uses a different type of visualization (“lines” instead of the traditional “rectangles”). These lines cannot be colored and expanded similarly to gene nodes, and are hence not suitable for pathway analysis in Cytoscape. Another challenge with the glycan-related pathways is that many of the children lack reactions in the downloaded XML files, even if the genes are presented as rectangles, and hence parts of the networks seem to consist of random genes with no connection to the path. It is possible to extend and style these gene nodes like in other pathways, but the missing reactions will still be lost. These problems are due to the way KEGG builds the XML file and how the file is read by KEGGScape.

As seen in the Tables S1 and S2, a total of 64 out of the 71 pathways contain at least one multi-gene node. The 71 pathways contain a total of 1974 nodes, of which 768 has multiple genes. Even though this only accounts for 39% of all nodes, it still means that for 90% of the pathways there is a possibility that not all relevant data will be included in the analysis. As we have shown, a single multi-gene node can change the entire interpretation of the pathway when all genes are included. Having at least one multi-gene node for 90% of the metabolism-related pathways used in this study demonstrates the importance of developing tools like FunHoP.

The first part of FunHoP, which deals with expanding the nodes with multiple genes, could possibly be solved also with other KGML-readers. CyKEGGParser has similar functions where all the “hidden” genes get a new node, with its own edges. This displays all the genes within a node as separate nodes, and these nodes can be colored and analyzed similarly as the original ones. However, as this study has shown, there are some nodes that contain a very large number of genes, which makes the analysis and interpretation challenging without further filtering by read counts from RNA-seq. CyKEGGParser is a KGML-reader/tweaker and does not have any of the features of FunHoP with regard to using reads from RNA-seq to determine an overall expression value for all genes in a multi-gene node. KEGGScape is a pure KGML-reader, which allows for running the pathway XML files through FunHoP locally and using KEGGScape to import the improved files. KEGGScape does not bend the edges the way CyKEGGParser does, and does not separate the functional homologs when reading the KEGG XML files. However, CyKEGGPar-

ser has many useful features such as corrections of inconsistencies in pathways and tissue-specific tuning, and these features could be interesting to consider in future studies.

For the cases where a KEGG protein complex contains nodes with multiple genes, it is dealt with by adding an expanded node on top of the protein complex. All genes can hence be seen and colored by expression, although the user may have to do a bit of manual editing of the network. This is a challenge in visualization of the networks, as an expanded node will be placed in the same position as the original node, but in most cases, it takes up more space than what was originally allocated.

To improve FunHoP and make it easier for others to use it, solutions for the problems above are under development. Converting FunHoP into a Cytoscape app is also in development, which will make it easier for all users to apply this method to their own analyses.

## Conclusion

In this study, we have shown how FunHoP can be used to expand nodes from KEGG in Cytoscape to include all alternative genes present in a node. We have shown how *P* values from differential expression are not sufficient to determine regulation in a pathway, and how using the read counts from RNA-seq can facilitate metabolic network interpretation. Finally, we have shown that information in the extended networks can be aggregated to create more simplified networks at the node level, taking data from all genes into account.

By comparison of measured values of histidine and GPC in PCa and healthy prostate tissue from literature, we have shown how our analysis can explain why these metabolites are elevated, whereas the original pathway representations could not. We have also managed to show how differential expression based on *P* values does not differentiate between highly expressed genes and lowly expressed genes. By incorporating RNA-seq read counts into the analysis, we have highlighted genes that are highly expressed and more likely to dominate within a pathway. Overall, we show that FunHoP, by incorporating more biological information on network nodes and genes from KEGG, is able to provide improved pathway analysis.

## Code availability

Code and documentation for running FunHoP can be found at <https://github.com/kjerstirise/FunHoP>.

## CRedit author statement

**Kjersti Rise:** Conceptualization, Software, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization. **May-Britt Tessem:** Writing - review & editing, Funding acquisition. **Finn Drablos:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition. **Morten B. Rye:** Conceptualization, Software, Formal analysis, Writing - review & editing, Visualization, Supervision, Funding acquisition. All authors have read and approved the final manuscript.

## Declaration of Competing Interest

The authors have declared no competing interest.

## Acknowledgments

This work was supported by a PhD position from Enabling Technologies, Norwegian University of Science and Technology (NTNU), and by the Department of Clinical and Molecular Medicine (IKOM), NTNU to KR. Funding was given from the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the NTNU to MBR. European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 758306) gave funding to MBT. Funding support for MPC\_Transcriptome sequencing to identify non-coding RNAs in prostate cancer was provided through the NIH Prostate SPOR (Grant Nos. P50CA69568 and R01R01CA132874), the Early Detection Research Network (Grant No. U01 CA111275), the Department of Defense Grant (Grant No. W81XWH-11-1-0331), and the National Center for Functional Genomics (Grant No. W81XWH-11-1-0520). The results shown here are in part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>. The authors would also like to thank Cato Lauvli for his contributions to the figures, and Einar Johan Sømåen, Torje Digernes, and Endre Stovner for their valuable thoughts and inputs.

## Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gpb.2021.03.003>.

## ORCID

ORCID 0000-0003-1822-2675 (Rise, K)  
ORCID 0000-0001-5734-2157 (Tessem, M-B)  
ORCID 0000-0001-5794-828X (Drablos, F)  
ORCID 0000-0002-4405-1008 (Rye, MB)

## References

- [1] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999;27:29–34.
- [2] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353–61.
- [3] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;46:D649–55.
- [4] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 2019;47:D419–26.
- [5] Chowdhury S, Sarkar RR. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database* 2015, bau126.

- [6] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res* 2019;47:D590–5.
- [7] Villaveces JM, Koti P, Habermann BH. Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Adv Appl Bioinform Chem* 2015;8:11–22.
- [8] Spohr P, Dinkla K, Klau GW, El-Kebir M. eXamine: Visualizing annotated networks in Cytoscape. *F1000Res* 2018;7:519.
- [9] Mustafin ZS, Lashin SA, Matushkin YG, Gunbin KV, Afonnikov DA. Orthoscape: a cytoscape application for grouping and visualization KEGG based gene networks by taxonomy and homology principles. *BMC Bioinformatics* 2017;18:1427.
- [10] Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. Cell Designer 3.5: A versatile modeling tool for biochemical networks. *Proc IEEE* 2008;96:1254–65.
- [11] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–504.
- [12] Bryant DH, Moll M, Chen BY, Fofanov VY, Kavradi LE. Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. *BMC Bioinformatics* 2010;11:242.
- [13] Nishida K, Ono K, Kanaya S, Takahashi K. KEGGscape: a Cytoscape app for pathway data integration. *F1000Res* 2014;3:144.
- [14] Nersisyan L, Samsonyan R, Arakelyan A. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows. *F1000Res* 2014;3:145.
- [15] Sales G, Calura E, Cavalieri D, Romualdi C. graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* 2012;13:20.
- [16] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. *Nucleic Acids Res* 2009;37:D674–9.
- [17] Wang J, Suzuki T, Dohra H, Takigami S, Kako H, Soga A, et al. Analysis of ethanol fermentation mechanism of ethanol producing white-rot fungus *Phlebia* sp. MG-60 by RNA-seq. *BMC Genomics* 2016;17:616.
- [18] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006;7:55–65.
- [19] Shendure J. The beginning of the end for microarrays? *Nat Methods* 2008;5:585–7.
- [20] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57–63.
- [21] Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 2014;9:e78644.
- [22] Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* 2015;163:1011–25.
- [23] Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* 2011;29:742–9.
- [24] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;14:R36.
- [25] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;30:923–30.
- [26] Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29.
- [27] Kanehisa Laboratories. KEGG Pathway Database. <http://www.genome.jp/kegg/pathway.html> (2016-01-28 2016, date last accessed).
- [28] Kanehisa Laboratories. Hsa ID to gene name. <http://rest.kegg.jp/list/hsa> (2016-01-28 last accessed).
- [29] Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 2009;457:910–4.
- [30] Bertilsson H, Tessem MB, Flatberg A, Viset T, Gribbestad I, Angelsen A, et al. Changes in gene transcription underlying the aberrant citrate and choline metabolism in human prostate cancer samples. *Clin Cancer Res* 2012;18:3261–9.
- [31] Giskeodegård GF, Bertilsson H, Selnaes KM, Wright AJ, Bathen TF, Viset T, et al. Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness. *PLoS One* 2013;8:e62375.
- [32] Swanson MG, Keshari KR, Tabatabai ZL, Simko JP, Shinohara K, Carroll PR, et al. Quantification of choline- and ethanolamine-containing metabolites in human prostate tissues using 1H HR-MAS total correlation spectroscopy. *Magn Reson Med* 2008;60:33–40.
- [33] Schwanhauser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. *Nature* 2011;473:337–42.
- [34] Lee MV, Topper SE, Hubler SL, Hose J, Wenger CD, Coon JJ, et al. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol Syst Biol* 2011;7:514.
- [35] Cheng M, Bhujwala ZM, Glunde K. Targeting phospholipid metabolism in cancer. *Front Oncol* 2016;6:266.
- [36] Glunde K, Penet MF, Jiang L, Jacobs MA, Bhujwala ZM. Choline metabolism-based molecular diagnosis of cancer: an update. *Expert Rev Mol Diagn* 2015;15:735–47.

## Paper 3

### **FunHoP analysis reveals upregulation of mitochondrial genes in prostate cancer**

*Kjersti Rise<sup>1\*</sup>, May-Britt Tessem<sup>2,3</sup>, Finn Drabløs<sup>1</sup>, Morten B. Rye<sup>1, 3, 4, 5\*</sup>*

1. Department of Clinical and Molecular Medicine, NTNU – Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

2. Department of Circulation and Medical Imaging, NTNU – Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.

3. Clinic of Surgery, St. Olavs Hospital, Trondheim University Hospital, NO-7030 Trondheim, Norway

4. Clinic of Laboratory Medicine, St. Olavs Hospital, Trondheim University Hospital, NO-7030 Trondheim, Norway

5. BioCore - Bioinformatics Core Facility, NTNU – Norwegian University of Science and Technology, P.O. Box 8905, NO-7491, Trondheim, Norway

KR - kjersti.rise@ntnu.no

MBT - may-britt.tessem@ntnu.no

FD - finn.drablos@ntnu.no

MBR - morten.rye@ntnu.no

\*Corresponding author:

Morten Beck Rye

Department of Clinical and Molecular Medicine

NTNU – Norwegian University of Science and Technology

P.O. Box 8905

NO-7491 Trondheim

Norway

---

---



ISBN 978-82-326-5652-3 (printed ver.)  
ISBN 978-82-326-5524-3 (electronic ver.)  
ISSN 1503-8181 (printed ver.)  
ISSN 2703-8084 (online ver.)



**NTNU**

Norwegian University of  
Science and Technology