FunHoP analysis reveals upregulation of mitochondrial genes in prostate cancer

Kjersti Rise^{1*}, May-Britt Tessem^{2,3}, Finn Drabløs¹, Morten B. Rye^{1, 3, 4, 5*}

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

Abstract

Mitochondrial activity in cancer cells has been central to cancer research since Otto Warburg first published his thesis on the topic in 1956. Although Warburg proposed that oxidative phosphorylation in the tricarboxylic acid (TCA) cycle was perturbed in cancer, later research has shown that oxidative phosphorylation is activated in most cancers, including prostate cancer (PCa). However, more detailed knowledge on mitochondrial metabolism and metabolic pathways in cancers is still lacking. In this study we expand our previously developed method for analyzing functional homologous proteins (FunHoP), which can provide a more detailed view of metabolic pathways. FunHoP uses results from differential expression analysis of RNA-Seq data to improve pathway analysis. By adding information on subcellular localization based on experimental data and computational predictions we can use FunHoP to differentiate between mitochondrial and non-mitochondrial processes in cancerous and normal prostate cell lines. Our results show that mitochondrial pathways are upregulated in PCa and that splitting metabolic pathways into mitochondrial and non-mitochondrial counterparts using FunHoP adds to the interpretation of the metabolic properties of PCa cells.

Introduction

The prostate is an exceptional gland in the male body when it comes to metabolism, both in normal and cancerous cells. Most normal human cells use the tricarboxylic acid (TCA) cycle and oxidative phosphorylation to harvest energy from food. The prostate cells, however, have a different approach. When acetyl-CoA enters the TCA cycle and is added to oxaloacetate to become citrate, the citrate is secreted rather than oxidized [1]. The secreted citrate is an essential part of the prostatic fluid. Zinc ions are crucial in secretion, as these ions inhibit ACO2, the protein that converts citrate to isocitrate [2, 3]. However, when the prostate cells become cancerous, mitochondrial activity increases [1], which is the opposite of the ‘Warburg effect’ exhibited by most cancerous cells [4]. The Warburg effect is found in most tumors, where the tumor cells produce lactate even when oxygen is present and limit the energy metabolism primarily to glycolysis [5].

Studying protein subcellular localization is essential to learn more about how the cell works. Several resources have measured or can predict subcellular protein localization. Experimentally localizations can be found for instance by using isotope-labeled C-atoms [6], antibodies and immunofluorescence [7], or mass spectrometry (MS), which is the method

used by the SubCellBarCode (SubCell) resource [8]. SubCell uses cell fractionation combined with in-depth quantitative MS which is used as input to a bioinformatics pipeline. This gives a database of subcellular localizations. Another resource with subcellular localization data is the Human Protein Atlas (HPA), which integrates various omics technologies to create an open-source protein map [9].

The growing field of computational tools for predicting subcellular localizations also leads to new insights. The Bologna Unified Subcellular Component Annotator (BUSCA) can be used to predict the localization of proteins by using existing knowledge of amino acid patterns such as GPI anchors, signal and transit peptides, as well as transmembrane domains such as alpha-helices and beta-barrels. BUSCA uses multiple tools in its prediction – five for predictions based on the protein sequence and three for those based on the gene sequence. Especially when combined, experimental and predicted data can reveal new knowledge about the localizations of proteins. Our study combines data on localizations from two datasets of experimental data (SubCell and HPA) with predicted data (BUSCA) to find subcellular localizations for proteins from the genes under study. The localization data is used in metabolic pathway analysis, which puts the differential expression analysis from RNA-Seq into biological contexts.

Our group has previously developed a method for investigating functional homologous proteins (FunHoP), a tool for metabolic pathway analysis which improves biological interpretations by utilizing information from both pathway and gene expression data [10]. FunHoP uses biological pathways from the KEGG database [11, 12] in combination with visualizations in Cytoscape [13] using the KEGGScape app [14].

The default KEGG pathways shown both in Cytoscape and on the KEGG website display only one gene in each node in the network, even though many biological reactions can be catalyzed by alternative enzymes or homologs. In a series of steps, FunHoP extracts knowledge on these alternative homologs to expand the nodes into showing all the relevant genes. When all the genes are visible the user can apply different styles based on p-value or read counts and get a better understanding of how the genes are regulated. Alternative or homologous genes within a node can be regulated in opposite directions or found not to be significantly regulated. In this study we add a layer of subcellular localization to FunHoP. By using FunHoP to separate metabolic pathways into mitochondrial and non-mitochondrial counterparts we identify mitochondrial pathways showing that PCa cells tend to activate the

TCA cycle for energy production, as well as mitochondrial-specific sub-pathways which provide the precursors to this activation.

Methods

An essential part of this study has been to add another layer of information on top of the improvements in pathway analysis that FunHoP could provide. Pathway data in XML format were downloaded from KEGG and run through FunHoP, which expanded all nodes with more than one gene, so that all genes were included for each node. This expansion made it possible to visualize all functional homologs within each pathway simultaneously. The expanded XML files were loaded into Cytoscape (version 3.4.0) via KEGGScape (version 0.7.0) and colored using transformed p -values from differential expression, using a red color range for downregulated genes and a green color range for upregulated genes. In this study the differential expression was calculated using RNA-Seq raw reads from two normal prostate cell lines (RWPE and PrEC) and two PCa cell lines (LNCaP and VCaP). Raw RNA-Seq SRA files were downloaded from Gene Expression Omnibus with accession GSE25183[15]. Raw RNA-Seq reads were mapped to the hg19 transcriptome using TopHat2 [16], and featureCounts [17] was used to assign the reads to each gene. We used Voom [18] to perform differential expression analysis between prostate cancer and regular cell lines. Differentially expressed genes with a p -value below 0.05 were extracted, and p -values were log2 transformed by:

$$\text{value} = \log_2 p\text{-value} \times (-10) \times (\text{regulation})$$

where regulation was defined as 1 for upregulated genes (positive fold-change) and -1 for downregulated genes (negative fold-change).

The proteins created from these gene homologs have potentially different subcellular localizations, which obviously is relevant to our view of the pathways. We used three different data sources for identifying the subcellular localization of proteins from each gene.

SubCell - We downloaded from SubCell ([8], <https://www.subcellbarcode.org/>) the subcellular localization data for the five cell lines A431, MCF7, H322, U251, and HCC827. We then made a SubCell consensus set from these data. For each gene we excluded any cell line where the localization was Unassign. If the localization was the same for all the remaining cell lines, or if one of the localizations was more frequent than the others (*i.e.*, the most frequent localization of this gene was seen more often than the second most frequent

one), then this was used as the preferred localization for the gene product. Otherwise, the localization was classified as Uncertain, and this was the case for approximately 4% of the genes.

HPA - From Human Protein Atlas (HPA) ([9], <https://www.proteinatlas.org/>) we downloaded subcellular localization data based on 69 cell lines representing different tissue types. This dataset showed only minor overlap with SubCell concerning cell lines, as only MCF7 was used in both datasets. In HPA more than one localization can be assigned to each protein, and 35 different subcellular localizations are defined. Therefore, we first re-coded the annotated localizations to a simplified list of four localizations (mitochondria, cytoplasm, nucleus, and secretory), using organelle proteomes and their grouping as provided by HPA (<https://www.proteinatlas.org/humanproteome/cell>), but with mitochondria as a separate group. Next, we used this simplified list for each gene with more than one localization to identify a preferred localization. If the list was consistent (*i.e.*, all the simplified localizations were the same), or if one of the localizations was more frequent than the others (*i.e.*, the most frequent simplified localization of this gene was seen more often than the second most frequent one), this was used as the preferred localization for the protein. Otherwise, the localization was classified as 'uncertain', which occurred for approximately 15% of the genes.

BUSCA - We used BUSCA to generate predicted subcellular localization of proteins for a total of 3,341 genes with gene products belonging to the set of 85 metabolic KEGG pathways. We used Ensembl BioMart [19] to download protein sequences for this set of KEGG-relevant genes, using HGNC [20] gene names as identifiers. Initially, we downloaded proteins for MANE Select transcripts ([21], <https://www.ncbi.nlm.nih.gov/refseq/MANE/>), providing one preferred transcript per gene. This gave protein sequences for 2,929 genes, or 87.7% of the relevant genes. For the remaining genes we identified the protein-coding transcript in each case that provided the longest protein sequence and used that for prediction. Finally, the complete set of sequences was used as input to the BUSCA server.

A consensus of the three localization data sets was made, using the most frequent localization in each case. For cases where each method showed a different localization (or localization was missing), the prediction from BUSCA was used, as the prediction data would be available for all genes.

For the consensus set the fraction of genes for proteins with mitochondrial localization was plotted against the number of genes (both genes in general and genes for proteins with mitochondrial localization) (Supplementary material, Supplementary Figures S1, S2 and S3). This was used to suggest specific pathways for further study. Four pathways were chosen as examples based on the fraction of mitochondrial proteins and previous knowledge on subcellular localization of each pathway. For each of the four chosen example pathways, the consensus was used to determine which proteins could belong to a mitochondrial version of the pathways and which proteins could not. The pathway XML files were then modified manually to create the mitochondrial and non-mitochondrial versions, which were run through the expansion part of FunHoP to get a more complete overview of the pathways. Finally, the XML files were loaded into Cytoscape for display and colored based on differential expression.

Results

Identification of proteins with mitochondrial localization

All the 3,341 genes from the 85 metabolic KEGG pathways were extracted to create the KEGG gene list used for the analysis. To identify proteins that localize to mitochondria, we used localization data from three sources BUSCA, HPA and Subcell. First, experimental localization data were obtained from SubCell and HPA. Since the experimental sources included localization data for only 53% (HPA) and 61% (SubCell) of the 3341 relevant KEGG pathway proteins, we also predicted subcellular localization for all the KEGG pathway genes with BUSCA, using the MANE selected transcripts if available (2,929 genes, or 87.7%), and the longest known transcript for the rest. The fraction of proteins located in mitochondria was similar in the experimental and predicted data (Table 1). The protein localization was in general quite consistent between the different methods (Supplementary material, Supplementary Table S1), but with some variation. The three datasets were therefore merged into a consensus set, as described under Methods.

Since a large fraction of the localizations will be based on predictions rather than experimental data the quality of these predictions will be important. To check the reliability of the predictions we compared the predicted data (BUSCA) to the experimental data from SubCell and HPA (Supplementary material, Supplementary Table S2), using data on whether localization was mitochondrial or not. The comparison showed an average specificity of 0.93 and sensitivity of 0.74 for the BUSCA predictions compared with SubCell

and HPA, confirming that BUSCA does a prediction of mitochondrial localization that is consistent with the experimental data. We also compared the experimental data (HPA to SubCell, and vice versa) in the same way. This showed an average specificity of 0.97 and sensitivity of 0.81. Although there is somewhat better specificity and sensitivity in the experimental data, we concluded that the similarity in performance for experimental and predicted data was sufficiently robust to use the consensus localization data in the further pathway analysis.

Table 1 – Summary of localization data

	Genes ^{a,i}	Mito ^b	% ^c	KEGG ^{d,i}	% ^e	KEGG & Mito ^f	% ^g
SubCellBarCode	11613	737	6.3%	2039	61.0%	233	11.4%
Human Protein Atlas	10837	819	7.6%	1782	53.3%	224	12.6%
Predictions (BUSCA)				3284	98.3%	367	11.4%
KEGG ^h				3341	100.0%		

^aNumber of genes in dataset. ^bNumber of genes in dataset where the gene product has mitochondrial localization. ^cPercentage with mitochondrial localization. ^dNumber of genes in dataset belonging to the set of KEGG genes. ^ePercentage of KEGG genes. ^fNumber of genes in dataset being KEGG genes and with mitochondrial localization. ^gPercentage with mitochondrial localization. ^hThe set of genes found in the KEGG pathways selected for analysis. ⁱExcluding genes without localization (unassigned or uncertain, see Methods).

Mitochondrial pathways are upregulated in prostate cancer cell lines

We performed differential gene expression analysis between PCa cell lines and prostate normal cell lines and mapped the results on the metabolic genes from KEGG. We observed a clear enrichment of upregulated genes in the mitochondria compared with other compartments. This trend was evident regardless of the resource used to determine localization (Figure 1). There also seems to be more genes outside of mitochondria that are downregulated, although this effect is less pronounced. We believe that the increase in upregulated secretory genes in the prediction data (BUSCA) compared to the experimental data may be an artefact, as most of the gene products predicted as ‘extracellular’ or ‘membrane’ by BUSCA were not in the experimental data.

Figure 1 — Up- and downregulated genes for various subcellular localizations in the cell. All three datasets show that the number of upregulated genes (blue bars) is higher than the number of downregulated genes (orange bars) for proteins localized to mitochondria, whereas proteins with non-mitochondrial localization are more equally distributed.

Using the fraction of mitochondrial proteins in each pathway we selected four pathways as particularly interesting (Table 2). First, we selected the TCA pathway as it is known to occur inside the mitochondria. The TCA cycle has 23 out of the 29 proteins usually located inside the mitochondria and has also been shown in the literature to occur there. Second, in contrast to the first pathway we selected Glycolysis which is known to be located outside of the mitochondria. Glycolysis is generally found in the cytoplasm, and with 52 out of its 68 proteins having a non-mitochondrial localization, it was a good example of a non-mitochondrial pathway. Pyruvate metabolism and Alanine, aspartate, and glutamate metabolism were chosen as examples of pathways with a mixed localization.

Table 2 — Distribution of mitochondrial and non-mitochondrial proteins within the chosen pathways

Pathway	Mitochondrial ^a	Non-mitochondrial ^a
TCA cycle (hsa00020)	79% (23)	21% (6)
Glycolysis (hsa00010)	24% (16)	76% (52)
Pyruvate metabolism (hsa00620)	48% (19)	52% (21)
Alanine, aspartate, and glutamate metabolism (hsa00250)	30% (11)	70% (26)

^aShows percentage and number of genes.

All four pathway XML files were modified manually to determine how the pathways were affected by sorting the nodes according to protein localization, whether mitochondrial or non-mitochondrial. For this dataset most of the non-mitochondrial localizations were to the cytosol (59%). For TCA and Glycolysis, this modification meant removing non-mitochondrial and mitochondrial proteins, respectively. For Pyruvate metabolism and Alanine, aspartate, and glutamate metabolism, one mitochondrial and one non-mitochondrial version of each pathway was created. The following sections will look deeper into these four cases.

The TCA cycle

The TCA cycle (Figure 2) is a typical example of a mitochondrial pathway that is upregulated in the PCa cells. As expected, this pathway remains intact also when only mitochondrial genes are included, and nearly all the genes are upregulated in the pathway, including ACO2, which converts citrate to isocitrate. Thus, the combination of intact pathways and upregulated ACO2 confirms the current notion that the forward TCA cycle is indeed activated in PCa.

Figure 2 — The TCA cycle, mitochondrial version.

1

2 In the non-expanded TCA cycle version displayed by KEGG the mitochondrial version of the
3 pathway is broken, since six of the proteins displayed are classified as cytoplasmic.
4 However, five of the six proteins are really from multigene nodes, where the mitochondrial
5 counterparts are left out. Using FunHoP to show all relevant proteins and localization layers
6 creates awareness of all the genes included in the pathway. By adding these layers we show
7 that the mitochondrial TCA-cycle is not broken, since the mitochondrial proteins are included
8 in the pathway model.

9

10 *Glycolysis/gluconeogenesis*

11 By contrast, Glycolysis/gluconeogenesis is an example of a pathway that is mostly non-
12 mitochondrial (Figure 3).

Figure 3 — Glycolysis/gluconeogenesis, non-mitochondrial version.

The non-mitochondrial part of the pathway shown in Figure 3 remains connected and intact, with one exception. The original figure shows how pyruvate is converted into acetyl-CoA. As this takes place in the mitochondria, this conversion is not shown in the non-mitochondrial version, leaving pyruvate only with the conversion towards lactate. No significant upregulation of lactate dehydrogenases (*LDHA-C* and *LDHAL6A*) was found in cancer cell lines compared with normal cell lines, supporting the notion that activation of the TCA cycle is the preferred mode of energy production in the PCa cells rather than anaerobic glycolysis. The switch from *HK2* to *HK3* mediated glucose conversion in cancer cells is also interesting

in this respect, since *HK2* is important in mediating the Warburg effect in prostate cancer [22].

The TCA cycle and Glycolysis/gluconeogenesis pathways remain intact in their respective compartments and serve as proof of principle that our experimental and predicted data on enzyme localization could be used to perform compartmentalized pathway analysis with the expected biological behavior. When separating pathways into mitochondrial and non-mitochondrial counterparts, an important goal was to learn more about pathways where there is a mixture of mitochondrial and non-mitochondrial genes. We describe two pathways to exemplify this, namely Pyruvate metabolism (Figure 4) and Alanine, glutamine and glutamate metabolism (Figure 5).

Pyruvate metabolism

1 b)

2 c)

3 **Figure 4** — The FunHoP-expanded original pathway of Pyruvate metabolism is shown in
4 Figure 4a, showing both up- and downregulated genes, independent of localization. The
5 original picture changes when sorting the genes on localization into mitochondrial versus

non-mitochondrial categories, and then splitting the pathway into a mitochondrial version (Figure 4b) and a non-mitochondrial version (Figure 4c).

The pyruvate metabolic pathway contains many genes classified as either mitochondrial or non-mitochondrial, creating substantially different pathways in these two compartments. The mitochondrial pathway remains continuous, with multiple branches surrounding pyruvate. In this pathway, conversion of D-lactate to pyruvate and further to oxaloacetate is increased in cancer cells by upregulation of *HAGH*, *LDHD*, and *PC*. Oxaloacetate is a precursor to and carrier of substrates in the TCA cycle, supporting the other results showing increased TCA cycle activity. The mitochondrial version also shows increased conversion of pyruvate to acetyl-CoA in cancer cells through upregulation of *PDHB* and *DLAT*.

By contrast, the non-mitochondrial version of the pyruvate pathway is broken into three separate pathways. First, there is increased conversion in cancer cells to D-lactate in the cytosol by upregulation of *HAGHL*, but no further conversions to oxaloacetate or acetyl-CoA. Instead, acetyl-CoA in the cytosol is prioritized for fatty acid synthesis by upregulation of *ACACA*. Note that the highly prioritized acetyl-CoA conversion by *ACACA*, leading to the activation of fatty acid synthesis is only evident when looking at the non-mitochondrial version of the pathway, since all alternative conversion branches from acetyl-CoA are downregulated in the cancer cells. By contrast, the unmodified pathway includes a complex mixture of both up- and downregulated genes in several branches, making it more difficult to conclude on a prioritized path for acetyl-CoA. An upregulation of a smaller pathway leading towards D-lactate in the top right corner is also observed.

In the pyruvate pathway many of the two-gene nodes from the original unmodified pathway typically contain one mitochondrial and one non-mitochondrial enzyme, for example *MDH1/2*, *ACAT1/2*, and *ACSS1/2*. When comparing prostate cancer cell lines to normal cell lines it can often be observed that the mitochondrial genes are upregulated in cancer, while the non-mitochondrial genes are downregulated.

Alanine, glutamine, and glutamate metabolism

Finally, the alanine/glutamine/glutamate (AGG) metabolic pathway (Figure 5) is examined. AGG metabolism is a complex pathway with several branches and sub-pathways surrounding the TCA cycle and providing it with metabolites affecting its activity. Again, we will focus on oxaloacetate in addition to glutamate/2-oxoglutarate.

1

Figure 5 — Alanine, glutamine and glutamate metabolism, a) original; b) mitochondrial; c) non-mitochondrial.

The precursor for oxaloacetate in the AGG pathway is L-aspartate (L-Asp). Three genes can convert L-Asp to oxaloacetate, namely *GOT1*, *IL4I1* (cytosolic), and *GOT2* (mitochondrial). Of these, only *GOT1* is upregulated in PCa cells, while *GOT2* and *IL4I1* are unchanged. This indicates that the increased conversion of oxaloacetate from L-Asp in cancer occurs mainly in the cytosol and not in the mitochondria, where *GOT2* is the only enzyme acting on L-Asp. Instead, oxaloacetate can be produced in the mitochondria through the pyruvate pathway, as described previously. Alternative conversions of L-Asp are exclusive to the cytosol. Another alternative source for TCA cycle intermediates is glutamate through its conversion to 2-oxoglutarate [23-25]. The precursor for glutamate is L-glutamine, and the path from L-glutamine via glutamate to 2-oxoglutarate is exclusive to the mitochondrial pathways through the genes *GLUD1/GLUD2* and *GLS2/GLS*. Though there is no net upregulation of these genes, a switch occurs where *GLS2* is preferred to *GLS* in the cancer cells. L-glutamine can be converted to glutamate in the cytosolic pathway (by the enzyme *GLUL*), but it is not converted further to 2-oxoglutarate. The latter can be produced by a separate cytosolic path, including *NIT2*, but it would then be transported to the mitochondria to be utilized in the TCA cycle.

Discussion

Defining the localization of gene products

When collecting the subcellular localization from the experimental data, it became clear that many genes still lack a defined localization. The two experimental datasets used in this study have a considerable overlap in ‘unknowns’, meaning that some data will be missing even in a combined dataset. Missing data can be complemented with tools such as BUSCA, which can be used to predict localizations based on gene or protein sequence.

A major issue when working with localization data turned out to be how to find a consensus among categories from different datasets. We decided to simplify this task in our analysis of pathways, as proteins were classified as either mitochondrial or not, and the latter were just used as a non-mitochondrial category. However, the different resources may operate with several different localizations. The histograms in Figure 1 show how all three datasets have been reduced to a few main categories: cytoplasm, mitochondria, nuclear, and secretory,

along with an ‘unknown’ category in the experimental data sets. For the experimental data, this reduction means that the ‘secreted’ group also will include extracellular – but not necessarily secreted – proteins. By contrast, the ‘mitochondrial’ group contains proteins found both within the mitochondria and in the mitochondrial membrane, with the equivalent solution for the nucleus. The predicted data from BUSCA does not contain an ‘unknown’ category but has a separate category for membranes. However, datasets from HPA and SubCellBarCode (SubCell) have both ‘uncertain’ (not an exact localization) and ‘unknown’. A comparison of these datasets regarding ‘uncertain’ or ‘unknown’ show a considerable overlap of genes in this category. This supports our decision to use predicted localization data in addition to the experimental data.

The three datasets did not agree in all cases. This was solved by making a consensus dataset. If two out of three lists agreed, consensus went to these two, even in cases where the predictions from BUSCA differed. In cases where all three datasets differed, BUSCA was used as a casting vote. This decision was based on the number of unknown localizations from the two experimental datasets SubCell and HPA. Since BUSCA uses prediction, such predictions could be generated for all proteins with missing data in SubCell or HPA. Therefore, BUSCA was in many cases the only dataset with a suggested localization. Hence, it provided the casting vote if no two similar localizations could be found. However, for some proteins, such as ACACA, both experimental datasets found the localization to be cytoplasmic, while BUSCA predicted it to be mitochondrial. This shows that, although BUSCA’s performance seems to be quite good (see Results), its predictions were not always consistent with other data. In cases like ACACA, with consistent experimental data, the experimental localizations would be used.

Interpretations of split networks using FunHoP

In this paper we illustrate how FunHoP in combination with data on subcellular localization can be used to generate more biological insights. When all multigene nodes are expanded with FunHoP, even its intermediate steps provide more information. In this study, we chose not to do a full FunHoP analysis, excluding the steps with expanded nodes and read counts, as well as the final collapsed nodes with a differential expression calculated at node level. However, using only the first and second steps of FunHoP also shows how moving from the original to the expanded view can provide the user with more biological information. For cases such as those shown in Figure 2 and Figure 3, with paths that are expected to occur mainly in a specific localization, FunHoP can provide comprehensive information by showing

all genes within each node. However, in Figure 4 and Figure 5, many of the nodes with two genes had one gene found within the mitochondria and one outside (mainly cytosol), but default visualization was inconsistent in terms of which gene was displayed. If the user does not see all the genes, making a list based on only one gene in each node means dividing the paths into a mitochondrial and a non-mitochondrial pathway, which may lead to broken paths. For instance, in the TCA cycle pathway, which is usually a mitochondrial pathway, four out of five two-gene nodes with one non-mitochondrial and one mitochondrial gene showed the non-mitochondrial gene first. Without FunHoP to expand the path to show the hidden mitochondrial genes, a similar localization analysis would create an incomplete pathway with four “holes” in the path.

A notable aspect of Figure 3 is that even though most of the path is intact, starting with glucose, it stops at pyruvate, although the link to the TCA cycle is still there. This link is missing from the non-mitochondrial version, while the original provides the conversion from pyruvate to acetyl-CoA. The removal of this conversion is a good illustration of potential improvements in pathway analysis that can be achieved by adding localizations to the analysis. Even such minor improvements can make the visualizations of pathways more biologically correct.

In terms of the split networks, a notable issue is that some connections seem impossible unless a transport mechanism across the cellular compartments (for example, from the cytosol to mitochondria) exists. Thus, a possible expansion to pathway analysis includes the integration of transport steps (with associated genes) into the pathway networks where this is necessary for the reaction to occur. This information is currently not included in most pathway analyses, though its biological importance can be considerable.

Cell lines vs. in vivo tissue data

Finally, the analysis in this study is performed on cell lines rather than on tissue samples. Cell lines used for prostate cancer research are often from metastatic cancer [26], where metabolism may be different from *in vivo*. Although cell lines in general often retain similarity to their primary tissues [27], it has also been shown that many cell lines exhibit gene expression and regulatory changes that clearly distinguish them from their origin [28], sometimes leading to only minimal similarity in biological processes [29]. Thus, the observed trends must be validated in tissue samples, for example by using publicly available datasets. However, the interpretation of metabolic pathways in tissue data can also be challenging

due to heterogeneous mixtures of tissue types and other confounding factors [30]. Nevertheless, cell cultures are homogenous and work well as a model system to demonstrate proof of principle, which was the primary goal of this study.

Conclusion

In this study, we have shown how using a combination of experimental and computational data can create a reliable consensus on the localization of gene products related to cellular metabolism. We have used this data in combination with differential expression from cancerous and normal cell lines and found that mitochondrial genes are generally upregulated in PCa cell lines. We have also shown that our program FunHoP can be used to investigate expanded networks where all relevant genes within a pathway node are shown, and that dividing such pathways into sub-pathways based on subcellular localization can provide novel biological insights.

Acknowledgements

We would like to thank the BUSCA team, especially Castrense Savojardo and Pier Luigi Martelli, for their help with the prediction of subcellular localization. This work was supported by the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU) to [MBR]; PhD position from Enabling Technologies, Norwegian University of Science and Technology (NTNU) to [KR], the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 758306) [MBT] and The Norwegian Cancer society [MBT].

Competing Financial Interests

The authors declare no competing financial interests

Author contributions

Kjersti Rise (KR): Conceptualization, software, validation, formal analysis, data curation, writing – original draft, writing – review & editing, visualization

May-Britt Tessem (MBT): Writing – review & editing, funding acquisition

Finn Drabløs (FD): Conceptualization, validation, formal analysis, writing – review & editing, supervision, funding acquisition

Morten Beck Rye (MBR): Conceptualization, writing – review & editing, supervision, funding acquisition

References

1. Costello LC, Franklin RB. A comprehensive review of the role of zinc in normal prostate function and metabolism; and its implications in prostate cancer. *Arch Biochem Biophys.* 2016;611:100-12. Epub 2016/05/02. doi: 10.1016/j.abb.2016.04.014. PubMed PMID: 27132038; PubMed Central PMCID: PMC5083243.
2. Eidelman E, Twum-Ampofo J, Ansari J, Siddiqui MM. The Metabolic Phenotype of Prostate Cancer. *Front Oncol.* 2017;7:131. Epub 2017/07/05. doi: 10.3389/fonc.2017.00131. PubMed PMID: 28674679; PubMed Central PMCID: PMC5474672.
3. Kavanagh JP. Sodium, potassium, calcium, magnesium, zinc, citrate and chloride content of human prostatic and seminal fluid. *J Reprod Fertil.* 1985;75(1):35-41. Epub 1985/09/01. doi: 10.1530/jrf.0.0750035. PubMed PMID: 4032375.
4. Warburg O. On the origin of cancer cells. *Science.* 1956;123(3191):309-14. Epub 1956/02/24. doi: 10.1126/science.123.3191.309. PubMed PMID: 13298683.
5. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* 2011;144(5):646-74. Epub 2011/03/08. doi: 10.1016/j.cell.2011.02.013. PubMed PMID: 21376230.
6. Chokkathukalam A, Kim DH, Barrett MP, Breitling R, Creek DJ. Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. *Bioanalysis.* 2014;6(4):511-24. Epub 2014/02/27. doi: 10.4155/bio.13.348. PubMed PMID: 24568354; PubMed Central PMCID: PMC4048731.
7. Lundberg E, Borner GHH. Spatial proteomics: a powerful discovery tool for cell biology. *Nat Rev Mol Cell Biol.* 2019;20(5):285-302. Epub 2019/01/20. doi: 10.1038/s41580-018-0094-y. PubMed PMID: 30659282.
8. Orre LM, Vesterlund M, Pan Y, Arslan T, Zhu Y, Fernandez Woodbridge A, et al. SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol Cell.* 2019;73(1):166-82 e7. Epub 2019/01/05. doi: 10.1016/j.molcel.2018.11.035. PubMed PMID: 30609389.
9. Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. *Science.* 2017;356(6340). Epub 2017/05/13. doi: 10.1126/science.aal3321. PubMed PMID: 28495876.
10. Rise K, Tessem M-B, Drabløs F, Rye MB. FunHoP: Enhanced Visualization and Analysis of Functionally Homologous Proteins in Complex Metabolic Networks. *Genomics, Proteomics & Bioinformatics.* 2021;In press. doi: 10.1016/j.gpb.2021.03.003.
11. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45(D1):D353-D61. Epub 2016/12/03. doi: 10.1093/nar/gkw1092. PubMed PMID: 27899662; PubMed Central PMCID: PMC5210567.
12. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1999;27(1):29-34. Epub 1998/12/10. doi: 10.1093/nar/27.1.29. PubMed PMID: 9847135; PubMed Central PMCID: PMC148090.
13. 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(11):2498-504. Epub 2003/11/05. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PubMed Central PMCID: PMC403769.
14. Nishida K, Ono K, Kanaya S, Takahashi K. KEGGscape: a Cytoscape app for pathway data integration. *F1000Res.* 2014;3:144. Epub 2014/09/02. doi: 10.12688/f1000research.4524.1. PubMed PMID: 25177485; PubMed Central PMCID: PMC4141640.
15. 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(8):742-9. Epub 2011/08/02. doi: 10.1038/nbt.1914. PubMed PMID: 21804560; PubMed Central PMCID: PMC3152676.
16. 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(4):R36. Epub 2013/04/27. doi: 10.1186/gb-2013-14-4-r36. PubMed PMID: 23618408.

17. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-30. Epub 2013/11/15. doi: 10.1093/bioinformatics/btt656. PubMed PMID: 24227677.
18. 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(2):R29. Epub 2014/02/04. doi: 10.1186/gb-2014-15-2-r29. PubMed PMID: 24485249; PubMed Central PMCID: PMC4053721.
19. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res*. 2015;43(W1):W589-98. Epub 2015/04/22. doi: 10.1093/nar/gkv350. PubMed PMID: 25897122; PubMed Central PMCID: PMCPMC4489294.
20. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res*. 2021;49(D1):D939-D46. Epub 2020/11/06. doi: 10.1093/nar/gkaa980. PubMed PMID: 33152070; PubMed Central PMCID: PMCPMC7779007.
21. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res*. 2021;49(D1):D1046-D57. Epub 2020/11/23. doi: 10.1093/nar/gkaa1070. PubMed PMID: 33221922; PubMed Central PMCID: PMCPMC7779060.
22. Wang L, Xiong H, Wu F, Zhang Y, Wang J, Zhao L, et al. Hexokinase 2-mediated Warburg effect is required for PTEN- and p53-deficiency-driven prostate cancer growth. *Cell Rep*. 2014;8(5):1461-74. Epub 2014/09/02. doi: 10.1016/j.celrep.2014.07.053. PubMed PMID: 25176644; PubMed Central PMCID: PMCPMC4360961.
23. Gao P, Tchernyshyov I, Chang TC, Lee YS, Kita K, Ochi T, et al. c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature*. 2009;458(7239):762-5. Epub 2009/02/17. doi: 10.1038/nature07823. PubMed PMID: 19219026; PubMed Central PMCID: PMCPMC2729443.
24. 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(7231):910-4. Epub 2009/02/13. doi: 10.1038/nature07762. PubMed PMID: 19212411; PubMed Central PMCID: PMCPMC2724746.
25. Taylor BS, Pal M, Yu J, Laxman B, Kalyana-Sundaram S, Zhao R, et al. Humoral response profiling reveals pathways to prostate cancer progression. *Mol Cell Proteomics*. 2008;7(3):600-11. Epub 2007/12/14. doi: 10.1074/mcp.M700263-MCP200. PubMed PMID: 18077443.
26. Cunningham D, You Z. In vitro and in vivo model systems used in prostate cancer research. *J Biol Methods*. 2015;2(1). Epub 2015/07/07. doi: 10.14440/jbm.2015.63. PubMed PMID: 26146646; PubMed Central PMCID: PMCPMC4487886.
27. Gillet JP, Varma S, Gottesman MM. The clinical relevance of cancer cell lines. *J Natl Cancer Inst*. 2013;105(7):452-8. Epub 2013/02/26. doi: 10.1093/jnci/djt007. PubMed PMID: 23434901; PubMed Central PMCID: PMCPMC3691946.
28. Lopes-Ramos CM, Paulson JN, Chen CY, Kuijjer ML, Fagny M, Platig J, et al. Regulatory network changes between cell lines and their tissues of origin. *BMC Genomics*. 2017;18(1):723. Epub 2017/09/14. doi: 10.1186/s12864-017-4111-x. PubMed PMID: 28899340; PubMed Central PMCID: PMCPMC5596945.
29. Tran V, Kim R, Maertens M, Hartung T, Maertens A. Similarities and Differences in Gene Expression Networks Between the Breast Cancer Cell Line Michigan Cancer Foundation-7 and Invasive Human Breast Cancer Tissues. *Front Artif Intell*. 2021;4:674370. Epub 2021/06/01. doi: 10.3389/frai.2021.674370. PubMed PMID: 34056582; PubMed Central PMCID: PMCPMC8155268.
30. Tessem MB, Bertilsson H, Angelsen A, Bathen TF, Drablos F, Rye MB. A Balanced Tissue Composition Reveals New Metabolic and Gene Expression Markers in Prostate Cancer. *PLoS One*. 2016;11(4):e0153727. Epub 2016/04/23. doi: 10.1371/journal.pone.0153727. PubMed PMID: 27100877; PubMed Central PMCID: PMC4839647.