Investigating Nrf2 function in disease using enrichment analysis predictions and top-down systems biology

Petri Pölönen
Master´s thesis
Master of Science Program in Biosciences
University of Eastern Finland, Department of Health sciences
University of Eastern Finland
July 2013

UNIVERSITY OF EASTERN FINLAND, Faculty of Health Sciences
Master of Science Program in Biosciences
PETRI PÖLÖNEN: Investigating Nrf2 function in disease using enrichment analysis predictions and top-down systems biology
Master thesis, 90 pages
Instructors: Merja Heinäniemi (PhD), Anna-Liisa Levonen (MD. PhD)
July 2013

Key words: Nrf2, enrichment analysis, top-down systems biology, cancer

Abstract

Nuclear factor (erythroid-derived 2)-like 2 (Nrf2) is a transcription factor which senses oxidative and electrophile stress. When activated, Nrf2 accumulates in the nucleus, where it induces the expression of cytoprotective target genes. Many diseases are associated with oxidative stress and Nrf2 is therefore a potential therapeutic target. However, Nrf2 has been shown to regulate processes other than redox response and could potentially have disease-promoting effects. Accumulating evidence suggests that constitutively active Nrf2 has a pivotal role in cancer as it induces pro-survival genes that promote chemoresistance and cancer cell proliferation. Therefore Nrf2 is a novel oncogenic transcription factor, but the prevalence on Nrf2 dysregulation and functions in cancer have not been fully characterized. We analyzed microarray data of over 900 cancer cell lines in Cancer Cell Line Encyclopedia (CCLE) and created an Nrf2 signature model based on our previous microarray data to identify cancers with overactive Nrf2 status. Four novel cancer types and a total of 77 cancer cell lines were discovered by two individual tools to have overactive Nrf2 status with > 95 % probability or FDR of 0.01. Furthermore, we investigated glioma clinical samples in The Cancer Genome Atlas (TCGA) and found characteristic Nrf2 signature in 60 (ca. 10 %) glioblastoma multiforme samples (FDR 0.05). Metabolic changes, such as the Warburg effect have a crucial role in cancer. Nrf2 has been reported to upregulate pentose phosphate pathway in cancer to produce NADPH and molecular building blocks to support cancer cell proliferation and survival. In addition Nrf2 affects the lipid accumulation and synthesis in atherosclerosis and metabolic syndrome mouse models. However, all Nrf2 regulated metabolic pathways have not been identified and little is known about Nrf2-mediated metabolism and its role in disease. We did pathway analysis to our endothelial cell Nrf2 expression data to identify Nrf2 regulated metabolic pathways using human metabolic pathway reconstruction (Recon1). Many previously reported metabolic pathways were enriched, including the pentose phosphate pathway and fatty acid metabolism pathways. Interestingly, cholesterol homeostasis-related pathways, cholesterol metabolism, steroid metabolism and lysosomal transport pathways were also enriched. Cholesterol homeostasis has a significant role in many diseases, including cancer and atherosclerosis; therefore these pathways were further investigated. Many cholesterol pathway genes were downregulated by Nrf2 during the primary response based on our microarray and confirmed by next generation sequencing methods that assayed both primary transcription and mature transcript levels (RNA- and GRO-seq data). However, during secondary response with Nrf2 overexpression data, many cholesterol pathway genes were upregulated. No direct Nrf2 targeted regulatory elements were found, suggesting that Nrf2 regulates cholesterol homeostasis indirectly. In these computational analyses, publicly available data was integrated with our own data to do unbiased hypothesis tests to understand Nrf2 role in disease. The results motivate several initial conclusions to be investigated further.

Acknowledgements

This study was done in the A.I Virtanen institute, department of Biotechnology and Molecular Medicine, University of Eastern Finland in the autumn 2012 - spring 2013. I would like to acknowledge Cardiovascular signaling group leader Anna-Liisa Levonen (MD. PhD) for giving me the possibility to work in her cutting-edge multidisciplinary group. Many thanks also for her valuable comments and efforts in supervising me.

I want to thank my main supervisor Merja Heinäniemi (PhD) for all the work she has done for me, not only for supervising this master´s thesis but also for orienting me towards future work as a scientist. She has always been supportive and good at motivating me. She was able to teach me bioinformatics and got me interested in the field of computational biology. I want to thank also the rest of the cardiovascular signaling group members for their help and friendliness.

I am grateful for Professor Matti Nykter from Institute of biomedical technology, University of Tampere for his valuable comments on this master´s thesis.

I want to thank also my friend Sean Sapcariu from the Luxembourg center of system biology for providing help with the English grammar.

Last but not the least I want to thank my family and my fiancé Jasmin for their love and support.

Kuopio, July 2013

Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
TF	Transcription Factor
ROS	Reactive Oxygen Species
ARE	Antioxidant Response Element
CCLE	Broad Novartis Cancer Cell Line Encyclopedia
TCGA	the Cancer Genome Atlas
GO	gene ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
GSVA	Gene Set Variation Analysis
GSEA	Gene Set Enrichment Analysis
FDR	False Discovery Rate
OA-NO2	nitro-oleic acid
OA	oleic acid
siRNA	small interfering RNA
adCMV	Adenocytomegalovirus
adNRF2	Adenoviral overexpression of NRF2
siNrf2	Knockdown of Nrf2 using siRNA
siCTRL	siRNA Knockdown Control
MCMC	Markov chain Monte Carlo algorithm
HUVEC	Human umbilical vein endothelial cell
qPCR	quantitative real time polymerase chain reaction
ChIP-seq	chromatin immunoprecipitation sequencing
GRO-seq	global run on sequencing
RNA-seq	RNA sequencing
H3K4me1	Histone-3 Lysine-4 mono-methylation
H3K27ac	Histone-3 Lysine-27 acetylation
IGV	Integrative Genomics Viewer

SVD	Singular value decomposition
PPP	pentose phosphate pathway

genes:

GST	glutathione S-transferase
NQO1	NAD(P)H:quinone oxidoreductase 1
GCLM	glutamate-cysteine ligase
HMOX1	heme oxygenase-1

Table of Contents

1 INTRODUCTION

Advances in high-throughput biological technology have made it possible to produce high amounts of protein, gene expression or metabolite concentration measurements. First it was thought that biological function could be inferred from these measurements directly. However, many biological and technical pitfalls in high-throughput data analysis have been discovered, which can result in false biological interpretations of the measurement data[1].

Gene sets were introduced, because it was difficult to understand biological functions based on differential expression of a single gene[1]. Genes of a gene set share a common function that is derived from the previous knowledge; genes can, for example, belong to the same biological pathway. Originally the approach was developed to address the differences between samples representing two distinct phenotypes. A gene set was associated with the phenotype if the gene set contained more differentially expressed genes than could be expected by chance. However, this approach has limitations, as only two phenotypes can be compared. In some cases it is important to classify data based on expression levels in a variety of phenotypes (for example to identify tumor samples expressing a cancer promoting pathway).

Enrichment analysis capabilities have been extended to cover unclear phenotypes as well, and absolute gene set expression levels can be quantified in a high number of samples[2]. However, the increase in the amount and complexity of data has introduced new challenges in data normalization and integration. Public sources such as the Cancer Genome Atlas (TCGA) have been developed to standardize the biological sample collection, measurement and analysis pipeline to make the data reliable and reproducible.

Single gene level analyses have evolved to gene set (pathway) level analysis. Pathway level analysis has also been extended to a systems biology approach, where the structure of the gene regulatory network (interactions between the genes) is considered. Top-down systems biology uses high amounts of data, previous knowledge and computational models to understand the biological system[3]. Enrichment analysis is useful in top-down studies, because the enriched pathways can be further extended to gene regulatory network analyses.

In this master´s thesis, enrichment analyses were used to study transcription factor Nrf2, which is the master regulator of the antioxidant response pathway. Public data was integrated with our own gene expression data, and computational biology models were used to predict Nrf2 hyperactivity in cancer as well as to discover Nrf2 regulated metabolic pathways. This master´s thesis is divided into two main parts: a literature review and an experimental part. In the literature review, top-down systems biology and public databases are briefly introduced. The main part of the literature review provides an overview of the available enrichment analysis methods that are used in gene expression analysis. In the last part of the literature review, Nrf2 biology and its role in cancer and in metabolism are introduced. The experimental part of the thesis presents the hypotheses, results and discussion of our enrichment analysis. Descriptions of the enrichment tool algorithms used in the analysis are provided in the materials and methods section.

2 REVIEW OF LITERATURE

2.1 Biology of gene expression in eukaryotes

Deoxyribonucleic acid (DNA) is a long double stranded helical molecule that contains the genetic information of the cell. This information is encoded into DNA using four types of nucleotides that differ in their side chain bases: adenine (A), guanine (G), cytosine (C) and thymine (T). These nucleotides are paired between the complementary sense and antisense strands: A is paired to T and G is paired to C. Double stranded DNA is wrapped around the histone proteins in the nucleus to form dense chromatin. Chromatin contains tens of thousands of genes in eukaryotes. A gene is a sequence of DNA that contains enough information to produce a functional unit, typically proteins, but many encode also small nuclear- (snoRNA), transfer- (tRNA), ribosomal- (rRNA) and microRNAs (miRNA). During transcription a gene is converted to messenger ribonucleic acid (mRNA) using RNA polymerases. Transcribed mRNA is complementary to DNA, so genetic information is preserved in the mRNA. Multiple genes can be transcribed simultaneously and these genes form the transcriptome of the cell. The protein-coding mRNA transcripts are processed by post-transcriptional modifications to produce the mature coding mRNA: both a 5´cap and 3´poly-A tail is added to stabilize and direct mRNA and non-coding introns are removed from the transcript during splicing to combine gene coding exons. Exons can be combined in different order by alternative splicing. Processed mRNA is transported from the nucleus to the cytoplasm and eventually to ribosomes, where mRNA is translated to proteins. During translation proteins are synthesized according to the mRNA genetic code. Proteins consist of 20 types of amino acids, each of which is bound to specific transfer RNA (tRNA). These tRNAs recognize a specific three nucleotide long codon in the mRNA and are thereby able to bring amino acids corresponding to the mRNA code to ribosomes which bind them together via peptide-bond forming amino acid chains. These polypeptides are further folded and modified to functional proteins. Proteins represent the main functional molecules of cells and are commonly categorized according to their functions, as for example enzymes, receptors, structural proteins, signaling proteins and transcription factors (TF).

Cells can alter gene expression by numerous mechanisms, including: Changing the DNA and chromatin structure and density, regulation of transcription, post-transcriptional regulation, and regulation of translation. TFs are key components in transcriptional regulation. They are proteins that bind to regulatory regions of the target genes, such as enhancer sites or promoters. TFs regulate gene expression by promoting or preventing accumulation of transcriptional machinery consisting of RNA polymerase and general TFs to the site or by changing chromatin structure via epigenetic changes (Figure 1.). TF binding sites are short regions of DNA with specific sequences that make the protein-DNA molecular interaction possible. Densely packed chromatin can limit TF binding and prevent transcription. Some TFs are able to interact with proteins that regulate chromatin density through epigenetic changes, such as acetylation or methylation of the histone proteins[4].

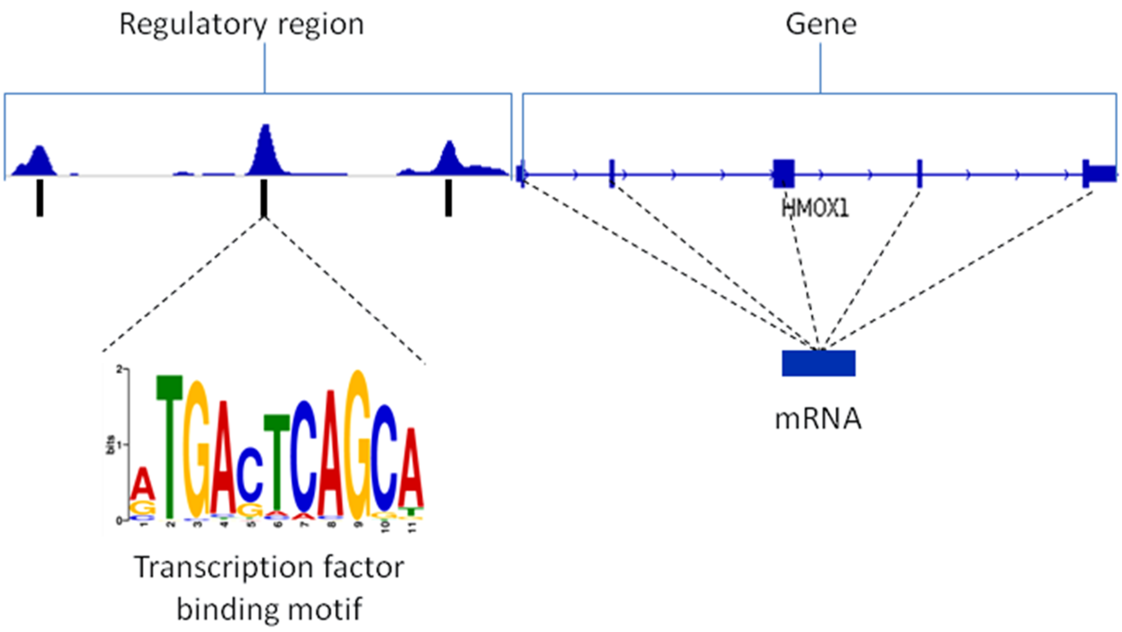


Figure 1. Overview of gene structure and regulation of gene expression. The gene shown contains exons (thicker lines) and introns (thinner lines) that form the primary transcript. The non-coding introns are removed during splicing, to form gene-coding mRNA that can be translated into a protein. Gene expression can be regulated from the gene regulatory regions. These regions commonly have epigenetic markers, such as Histone-3 Lysine-4 mono-methylation (H3K4me1) and Histone-3 Lysine-27 acetylation (H3K27ac) that mark open chromatin for TF binding where specific DNA sequences recognized by different TFs (motifs) can be found. In the figure, the H3K4me1 marker peaks reveal active regulatory regions of the *HMOX1* gene. TFs with a motif specific (complementary) DNA binding domain can attach to this regulatory site and alter *HMOX1* expression levels.

2.2 Top-down systems biology

2.2.1 Introduction to systems biology

Biological systems, such as a living cell, have commonly been studied in steady-state conditions, mainly to understand the role of biological sequences; DNA, proteins and their interactions. Molecular biology methods have made discoveries of fundamental biological systems possible. However, now it is possible to investigate deeper into biological systems, because technology has advanced to produce genome-wide high-throughput data. This data can be used to study the dynamics of the biological processes and complex interactions between the biological molecules.

Genome-wide methods produce vast amounts of information about molecular concentrations, such as mRNA, protein or metabolite levels, which created new challenges for computational biology to analyze and to interpret the data. Mathematical models and computational simulation techniques have been introduced to solve these problems and they have proven to be useful in biological research[3]. In addition, public availability of data expands research even further beyond the limits of molecular biology.

Systems biology aims to understand the functions of a living organism by observing molecular interactions and their mechanistic dynamics[3]. The focus can be on a whole living organism, sub-part or a functional pathway, but the main idea is to explain how the higher-level functions in these systems arise from the lower-level molecular interactions and molecular dynamics. Top-down systems biology starts the analysis from the integration of large genome-wide data sets, providing an unbiased overview of what can be observed from the data[3]. During data integration, molecular concentrations are correlated with molecular pathway activities to formulate hypotheses concerning regulatory relationships between the molecules[3]. This regulatory network can be used to understand how the system works and further developed to predict how the system is regulated under different conditions.

2.2.2 Public data availability

The amount of public data has expanded as techniques in molecular biology have developed. Expression data is available for multiple cell lines, data types and experimental conditions; this data can be used to discover new hypotheses, compare analysis results, integrate data, or predict outcome based on predefined experimental conditions and hypotheses.

Many large-scale projects collect data for public use: the most relevant ones for this thesis are briefly introduced here. The Encyclopedia of DNA Elements (ENCODE) project aims to locate and identify functional elements in the human genome by using genome-wide high-throughput methods[5]. It contains publicly available data on multiple cell lines to identify TF binding sites, regulatory elements and chromatin status markers[5]. The Gene Expression Omnibus (GEO) database contains functional genomics data, such as microarray and sequence-based data[6] published in biological journals by multiple individual research groups. Cancer Cell Line Encyclopedia (CCLE) is specialized in providing public data on different cancer cell line samples including expression data, mutation data, pharmacological profiles and DNA copy number data[7]. The Cancer Genome Atlas (TCGA) collects homogenous specimens on different cancer types from patient tumors. The project focuses on mapping the genetic mutations to different cancer types to understand the genetic basis of cancer. TCGA provides high amounts of cancer specific expression data, mutation data, DNA copy number data and miRNA profile data.

Public data provides more input for computational analysis, enabling more powerful statistical analyses. Public resources, such as TCGA and CCLE, can provide biological test cases to perform predictions based on models that capture previous knowledge or experimental observations. These concepts and public data sources are used in the experimental part of this thesis to perform enrichment analyses, providing a good example of how public data and enrichment analysis benefit a top-down analysis of Nrf2-mediated gene regulation.

2.3 Gene set enrichment methods

While the fundamentals of different enrichment analysis techniques are similar, they tend to focus on the different specialties in biology. In my thesis, I will focus on the use of enrichment analysis on gene expression profiles.

In transcriptomics the gene expression levels can be measured for each gene using microarrays or sequencing-based method (RNA-seq). Comparison of gene expression levels between phenotypes, such as drug treated samples and control samples, can be used to associate changes in the gene expression with the phenotype. However, it has been recognized that making conclusions based on the expression level of a highly differentially expressed gene or the top list of genes (with some cutoff threshold) can result in false biological interpretations due to oversimplification or failure to detect relevant differential expression[8–10]: There are problems with the reproducibility of the analysis due to the requirement to define a cutoff threshold; many genes with moderate but meaningful changes in the expression levels could be lost by setting a cutoff value that is too strict.

Gene expression analysis evolved from single genes to gene sets, whereby modest changes in the set of functionally related genes are considered to explain changes in the phenotype more accurately than the list of differentially expressed genes[10,11]. A gene set is a group of genes that share a function according to a *priori* defined knowledge[8]. Many databases have been created to provide biologically annotated gene sets, including MSigDB V3.0 [11], which provides gene sets with GO (gene ontology) categories and GeneSigDB[12], which provides literature-derived curated gene sets. Some databases also provide additional information about the gene set, such as compartmentalization and molecular interactions. Examples of these databases are KEGG (Kyoto Encyclopedia of Genes and Genomes)[13] pathways and Reactome[14].

A commonly used method for analyzing over-presentation of gene sets in expression data is Gene Set Enrichment Analysis (GSEA)[8], which is widely used because it is easily accessible (R, matlab and java graphical user interphase) and does not require programming skills. It provides graphs that are ready to be published and therefore it is a good compromise for common user. Other popular tools are the Database for Annotation, Visualization and Integrated Discovery (DAVID) and Ingenuity Pathway Analysis (IPA), both of which are intuitive to use but are restricted to specific pathway sources, such as KEGG, GO and IPA's own pathway curations. The enrichment method used in DAVID is a hypergeometric test, while in IPA either a hypergeometric test or a method similar to GSEA is used.

Gene sets can also be called gene expression signatures. In this thesis, the term gene set is used to refer to a list of related genes with a common function. When using the term signature, a group of genes that can distinguish phenotypes based on expression values is meant.

There are two types of questions that are commonly answered using enrichment analysis tools:

Q1: Which gene sets are enriched in a list of differentially expressed genes? (Class discovery)

Q2: Which samples indicate that a gene expression signature is active in them? (Class prediction)

Question 1 is commonly answered by unsupervised enrichment methods (Figure 2.). Unsupervised methods are used to find hidden structures without the evaluation of correct or incorrect solutions. Question 1 is therefore ideal for class discovery, to formulate novel hypotheses in the data. This concept can also be called a pathway analysis as gene sets often contain pathway reconstructions. Samples are typically assigned to pre-defined phenotypes, which are compared to each other to evaluate the enrichment relative to the other phenotype. Amounts of samples and comparable phenotypes are typically limited.

Question 2 can be answered using supervised learning methods and some unsupervised methods (Figure 2). Learning with evaluation of correct or incorrect solutions is used to train a model to predict classes. The general framework is to identify the genes that are important for the classification and then use this information to predict the class in other samples. Multiple samples with different phenotypes can be used, but there must be pre-defined information to create an informative (class-distinguishing) gene expression signature; for example, a signature to indicate samples where a given pathway is active or inactive.

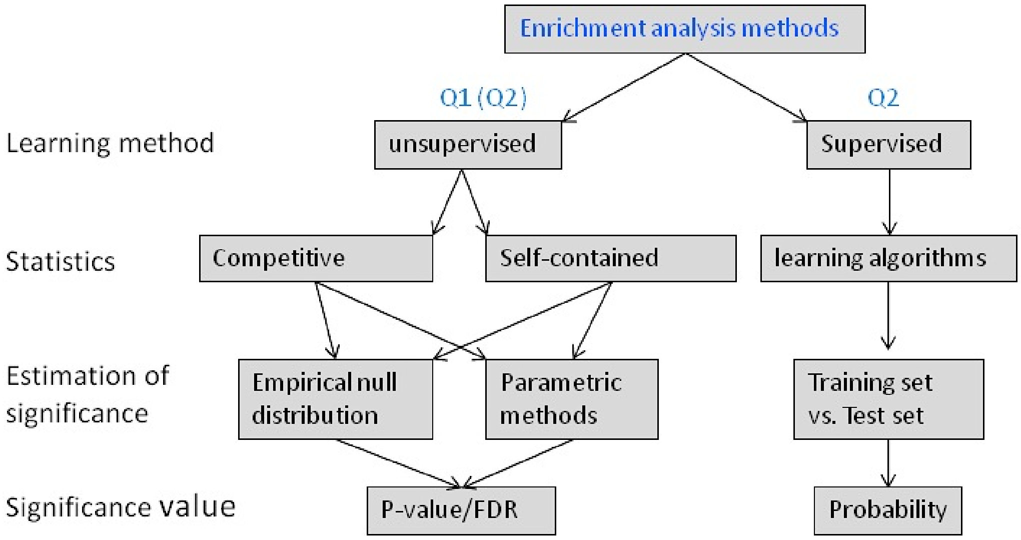


Figure 2. Overview of enrichment method classification. The enrichment method is chosen according to the question to be answered (Q1/Q2). In unsupervised methods, the statistics are chosen based on the null hypothesis and significance is estimated based on parametric or non-parametric methods. Non-parametric methods estimate the significance using an empirical null distribution. Unsupervised methods provide a P-value, or a false discovery rate (FDR) as a measure of significance. Significance is determined by rejecting the null hypothesis at some P-value/FDR cutoff value, typically less than 0.05. In supervised methods many algorithms, such as Bayesian regression, can be used to predict class in the samples. The measure of significance can be the probability of the result, but also other measures of significance are available.

2.4 Unsupervised enrichment methods

Unsupervised gene set enrichment methods are divided between competitive and self-contained methods. Unsupervised methods are typically used in the Q1 type of questions, but a few methods are developed to answer the Q2 type of questions. Distinguishing between competitive and self-contained methods can be done according to the selected null hypothesis for statistics[15]:

Competitive null hypothesis: genes in the gene set can be differentially expressed at most as often as the genes in the background set

Self-contained null hypothesis: genes in the gene set are not differentially expressed

Gene enrichment methods can also use local and global statistics for the enrichment calculation, typically depending on the null hypothesis. Parametric tests estimate the significance by assuming analytic distributions of the test data, such as a normal distribution, whereas non-parametric test estimates significance empirically by permuting gene or sample labels of the test data to create a null distribution (Figure 3.).

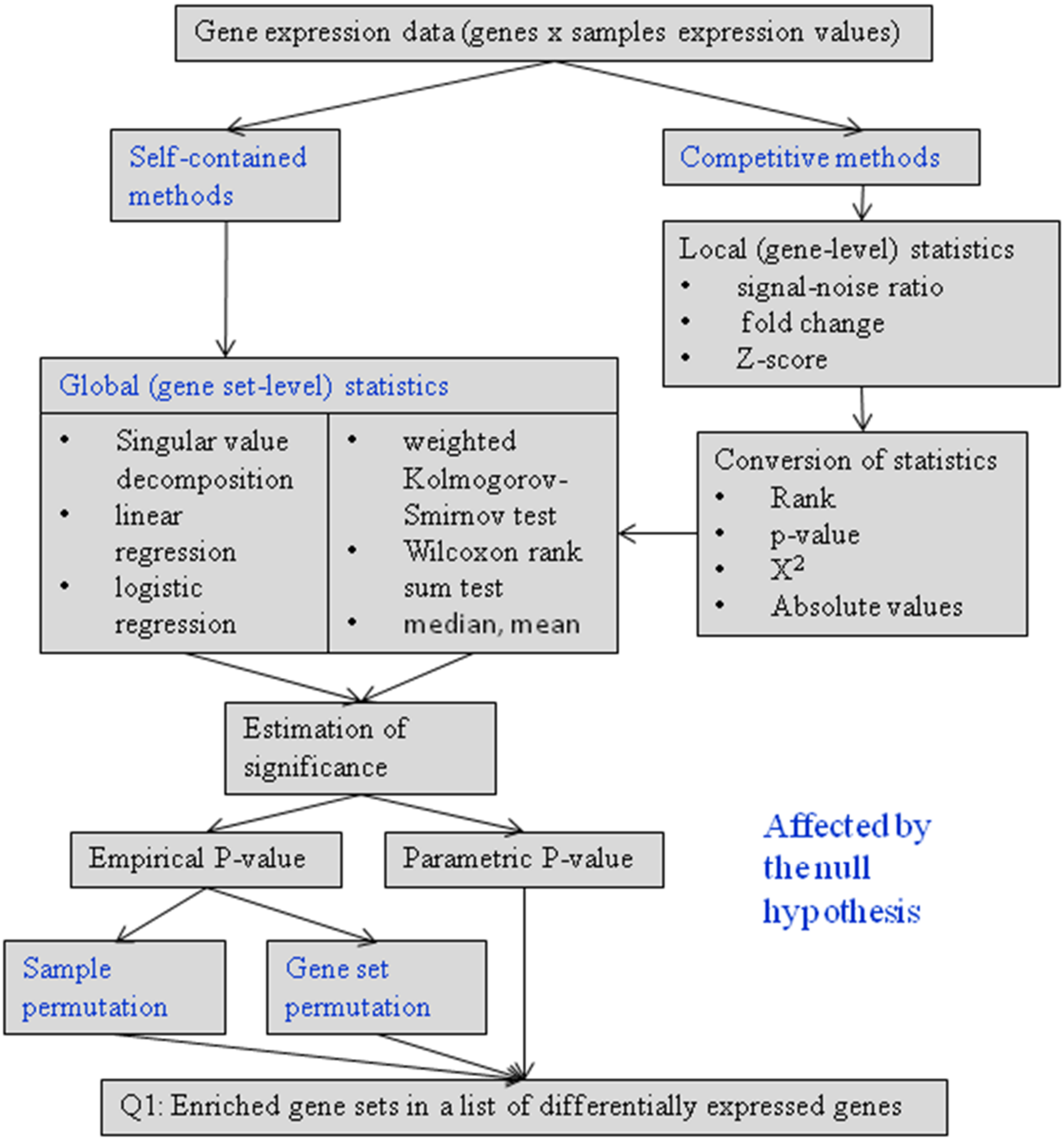

Gene expression data (genes x samples expression values)
Self-contained methods
Competitive methods
Local (gene-level) statistics
• signal-noise ratio
• fold change
• Z-score
Global (gene set-level) statistics
• Singular value decomposition
• linear regression
• logistic regression
• weighted Kolmogorov-Smirnov test
• Wilcoxon rank sum test
• median, mean
Conversion of statistics
• Rank
• p-value
• X2
• Absolute values
Estimation of significance
Empirical P-value
Parametric P-value
Affected by the null hypothesis
Sample permutation
Gene set permutation
Q1: Enriched gene sets in a list of differentially expressed genes

Figure 3. Schematic overview of performing enrichment analysis in Q1 type analysis.

2.4.1 Competitive methods with sample or gene set randomization

Competitive methods test the enrichment of differentially expressed genes in a gene set relative to the background set. A simple example of a competitive method is the hypergeometric test, which is commonly used in testing for over-presentation (Figure 4.) and will be used also in the experimental part of this thesis.

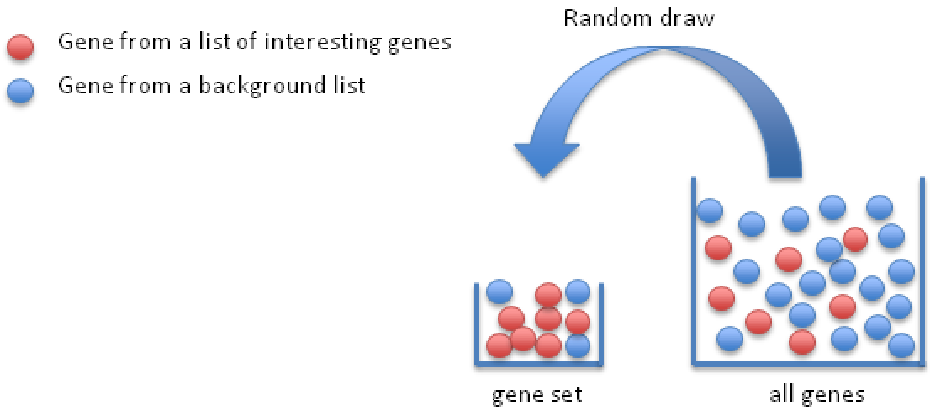

Gene from a list of interesting genes
Gene from a background list
Random draw
gene set
all genes

Figure 4. Schematic description of competitive null hypothesis using hypergeometric test. Genes measured are separated into a list of interesting genes and a background list of genes by some cutoff. Interesting genes can be for example the differentially expressed genes obtained from a t-test between samples and the cut-off used is a certain P-value. If a gene set contains more interesting genes than could be expected by random draw from all genes, genes in that gene set are overpresented. Over representation rejects the null hypothesis, if P-value of over representation is below significance cutoff (for example 0.05).

Another method used in the experimental part is GSEA that uses competitive testing and gene or sample randomization. In the GSEA, expression levels are converted to signal-to-noise ratios and genes are ranked based on the best distinction between the two phenotypes. The phenotypes can be for example treated and untreated samples. Differential expression of a gene set is determined by correlating genes in the ranked list with the phenotype: If multiple genes of the gene set are found from the top or bottom of the ranked gene list, correlation of the gene set with phenotype is high and vice versa. This level of correlation is the GSEA enrichment score, which is calculated by using weighted Kolmogorov-Smirnov statistics (Figure 5). The algorithm goes through a list of ranked and correlated genes and increases a running-sum statistic when gene is included in the gene set and decreases it when it is not. The magnitude of the running-sum is weighted so that genes with a high correlation with the phenotype (up-down in the ranked list) get higher increase in running-sums. Statistical significance of the enrichment score is assessed by simulating a null distribution for the enrichment score. This is achieved by shuffling the phenotype labels or the gene set compositions, calculating the differential expression of the genes and ranking them and then calculating enrichment scores for permutated case. Typically 1000 permutations are done to obtain the null distribution of enrichment scores. The permutation of sample labels preserves the complex correlation structures in the ranked gene list and is thought to provide biologically meaningful significance assessment[10,16]. An empirical P-value is calculated relative to the null distribution: the P-value is defined as the fraction of shuffles in simulations that are needed to produce the actual enrichment score. Also correction for multiple testing can be done in GSEA, by normalizing the enrichment score and then comparing the tails of the observed and the null distribution of the normalized enrichment scores.[8]

The basics behind GSEA are important, as vast majority of the enrichment methods available implement the GSEA algorithm and about half of them simulate the empirical P-value by permuting samples as in the original GSEA[17]. Variants of GSEA include tools such as SAFE[18], GSA[19], ASSESS[20] which have been compared in multiple articles using simulated and experimental data [10,19,21]. GSEA variants differ mainly in their method for calculating the enrichment score. Instead of the weighted Kolmogorov-Smirnov test, many methods use a mean test, a median test, a Wilcoxon rank sum test or a maxmean test. They also have differences in how the significance is estimated and how the multiple hypothesis correction is done.

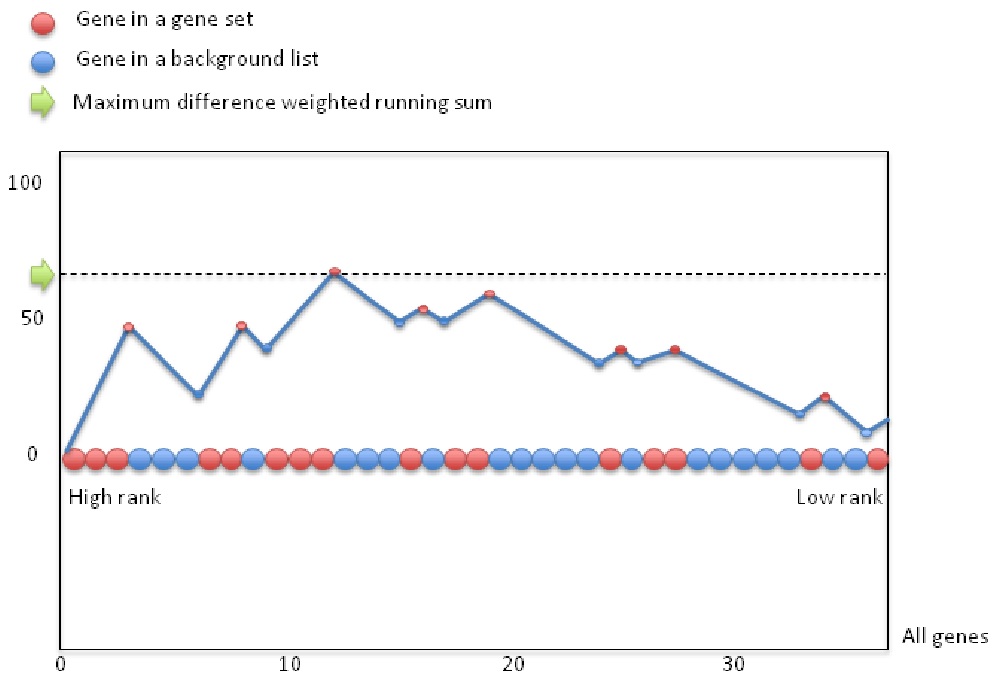

Gene in a gene set
Gene in a background list
Maximum difference weighted running sum
100
50
0
High rank
Low rank
All genes
0
10
20
30

Figure 5. Schematic description of the weighted Kolmogorov-Smirnov random walk statistics. All genes are ranked based on differential expression: high difference gets a high rank and no difference gets a low rank. Random walk is done to each gene and if the gene is in the gene set, running sum is increased and if not, it is decreased. High rank genes in the gene set are weighted to increase the running sum more than the low rank genes. If the maximum difference from zero of the running sum is unusually high (rejects null hypothesis), gene set is enriched. Unusually high running sum means that permutation of sample labels results in a random rank lists, which gets as high or higher running sums than observed at most as often as the significance cutoff states.

Comparisons of gene enrichment analysis tools have revealed that different methods have often poor overlap, especially when results between the competitive and self-contained methods are compared[17]. In addition simpler methods and self-contained methods often outperformed GSEA type methods in simulated data and in experimental data GSEA and the variants seemed to be better. There are multiple suggestions and all the tools seem to outperform others in their own benchmark testing. Possible reason for contradictory enrichment tool suggestions is that gene enrichment tools are often benchmarked using simulated data sets[21]. The benefit of simulated data is that numbers of true positives and true negatives can be controlled, whereas experimental data sets are biologically complex and lacks standards and therefore it is challenging to know which method performs the best. However, simulated data sets might model biological complexity poorly and lack fundamental interactions. Therefore also experimental data should be used when testing the tools[21].

2.4.2 Self-contained methods with sample randomization

In contrast to competitive methods, self-contained methods, such as Globaltest[22], PLAGE[23], ANCOVA[24] do not need comparison of the gene set to the background set. Expression levels of genes in a gene set are associated with the phenotype directly by using global statistical tests instead of calculating the gene-level enrichments[25]. The self-contained null hypothesis is statistically more sensitive than the competitive null hypothesis, as the background set is not considered and the null hypothesis is therefore more restricted[15,17]. If a gene list for example contains many differentially expressed genes, in the competitive null hypothesis enrichment score is reduced, as the background set is likely to contain differentially expressed genes. Controversially, if the gene set contains only a few differentially expressed genes, the self-contained null hypothesis might be over-sensitive in associating the gene set with the phenotype[10,26,27]. An additional reason for the higher sensitivity is that self-contained analysis can be done to a specific gene set, without the need of multiple corrections for many other gene sets[25].

The Globaltest method is based on empirical Bayesian generalized linear model, which uses linear and logistic regression to predict if phenotype (1 or 0) or clinical status is dependent on measured gene expression[22]. Global test is motivated by assumption that a gene set can be used to distinguish phenotypes and that changes in expression patterns should change also the phenotype. Therefore null hypothesis in this method is that none of the genes in the gene set are correlated with the phenotype 1. To reject the null hypothesis, genes in the gene set do not need to have similar expression patterns, but many of the genes needs to be correlated with the phenotype 1. Differential expression between two phenotypes is modeled using random effect logistic regression[22]. Generalized linear model can be used to predict the phenotype or clinical outcome by estimating regression parameters from the training data and then to compute the correlation with the phenotype. The model parameters becomes non-estimable when there are far less samples than genes in the gene set, because there are too few degrees of freedom[22]. In Global test regression is possible by assuming that the regression coefficients for each genes in the gene set follow same distribution with mean of zero and an unknown variance[22]. ANCOVA is very similar to Global test, but it test whether gene sets with similar phenotypes or clinical outcomes have similar expression patterns[28]. Therefore roles of expression levels and phenotypes are exchanged relative to Global test. Both methods use expression values directly for predictions; therefore normalization of the samples needs to be robust[26]. Self-contained methods have been compared and their results were comparable after proper standardization[26].

To address the problem in analyzing enrichment for large gene sets from experiments with small number of samples, singular value decomposition (SVD) was introduced to reduce dimensions when using linear regression in PLAGE[23]. Expression levels are first standardized to Z-scores over the samples, which indicate how many standard deviations the expression level is above or below the sample mean. Then genes in the gene sets are converted to eigenvectors (“metagenes”) using SVD and the first eigenvector (with the highest eigenvalue) is used to define the whole gene set activity level in the sample[23]. SVD is commonly used in enrichment analysis tools and details about the method are provided in the experimental part of the thesis.

Self-contained methods are not limited to address Q1 type hypothesis. Gene Set Variation Analysis (GSVA[29]) is a non-parametric and unsupervised enrichment method that uses similar statistical methods to GSEA, but GSVA is tuned from gene-level statistics to utilize global gene set-level statistics to calculate relative enrichment score of gene set activity across multiple samples. It will be used in the experimental part of this thesis for a Q2 type analysis. GSVA is not suitable for Q1 type class discovery analysis with small sample sizes and experiments where two phenotypes are compared; other unsupervised methods are more suitable for that. However, GSVA enrichment analysis is can be used for quantifying enrichment in large public datasets such as TCGA to predict pathway activity in heterogenic phenotypes.

2.4.3 Parametric methods

The non-parametric methods impose a high computational cost when significance is estimated by permutation. Simple parametric Z-score and X^2 test calculations were introduced as a powerful enrichment analysis methods[30]. The use of analytic distribution, such as the normal distribution reduce the computation time and makes it possible to infer very low P-values, which would require extreme amounts of permutations for the background simulation. However, when simulated and analytic backgrounds were compared, it was clear in many cases that the analytic background is inferior in accuracy[21]. These methods were also criticized of ignoring the gene-gene correlations and the correctness of the analytic assumptions in the calculation of the significance have been questioned[31]. PAGE is one example of a recent parametric method that uses averaging of the local gene-level statistics and comparing it to standard normal distribution to estimate significance[32].

2.5 Supervised learning enrichment methods

Complexity of experimental designs has increased due to the reduced costs in the genome-wide expression analysis. Phenotypes cannot always be classified into a few categories, if RNA has been extracted from clinical samples or from undefined disease data. Complexity of the phenotypic categories and the large sample sizes limits usability of the Q1 type class discovery analysis, but the Q2 type phenotype prediction can be done to classify samples. The Q2 type questions are typically answered using supervised machine learning methods that are not as common as unsupervised methods, but increasingly interesting as sample databases such as TCGA are becoming general. Supervised methods have been applied for example to predict clinical outcome based on expression profiles[33], to detect secondary activation of endogenous signaling pathways[34] and to predict TF activation signatures in specific cancer types[35].

Basic methodology behind the supervised machine learning is to first create a phenotype classifier from a training data set. The phenotype classifier is then used in a test data set to test how probable it is that the test set has same phenotype as the positive training samples[36]. These methods are supervised, because they create a gene set from pre-defined experimental conditions where phenotypes can be strictly controlled, such as cases where a pathway is active in one phenotype and inactive in another or measurements of known pathway response to a defined stimulus. The prior knowledge must be correct to train the model; otherwise the learning method will classify data based on false information[36]. As the expression level changes are known in pre-defined experiments for phenotypes, expression can be directly measured in the other samples[2]. This enables quantitative prediction of the gene set activity in the samples. The most widely used learning algorithms are based on Bayesian models, linear regression, logistic regression, naive Bayes models and support vector machines[37,38].

A typical problem in supervised learning methods is the over fitting of the data, which means that the algorithm predicts well in the training set, but badly in the test set[16]. The over fitting occurs when the model is too complex, for example has too many parameters relative to the number of samples. In a gene expression data this occurs commonly: the number of samples is much smaller than the number of genes. When the algorithm is trained using the training data, it can "memorize" the gene expression levels. When the model is then used in the test set, it cannot find similar expression levels as in the training data, which results in poor predictive power. Over fitting can be prevented by reducing the complexity and dimensionality of the model and by cross-validating the model using a test set[36]. Dimensions can be reduced using SVD and by setting a smaller set of genes to be collected by the training set[2]. Also additional procedures can be done, including regularization, pruning and model training with noise[36]. The supervised machine-learning tool SIGNATURE is used in Q2 type enrichment analysis in the experimental part of this thesis. The method uses Bayesian probit regression model, SVD and Markov Chain Monte Carlo (MCMC) algorithm and more details about these methods are provided in the materials and methods part of the thesis.

2.6 Introduction to Nrf2 biology

2.6.1 Keap1-Nrf2 stress response pathway

Living organisms are exposed to chemicals that are not produced by the organism itself. These xenobiotics can be found in drugs, pollutants and food, some of which can be hazardous for the organism. Organism can also produce harmful substances such as ROS (reactive oxygen species) as a side product of metabolic reactions. Detoxification is considered to consist of phase I, II and III metabolic processes. In the phase I reaction, functional and polar groups are added to the xenobiotics, commonly by the family of cytochrome P-450 enzymes. Some of these compounds are cytotoxic and carcinogenic, causing DNA damage and protein modifications. Therefore these electrophilic xenobiotics and ROS are inducing the phase II response, which further inactivates the electrophilic metabolites. In the phase III the membrane transporters of the multidrug resistance protein family excrete the inactive metabolites outside of the cell.[39,40]

In the mammalian body, liver is the main organ to detoxify drugs and pollutants, but the antioxidant response against oxidative and electrophilic stresses with the phases II and III enzymes is crucial for the cellular health and in the prevention of tissue injury in all cell types. The stress response pathway is activated by inducing the expression of the phase II cytoprotective genes that are involved in the glutathione synthesis, neutralization of reactive oxygen species (ROS) and xenobiotic metabolism[39,41]. The stress response specific TFs bind to the regulatory regions of these cytoprotective target genes, which promote the general transcriptional machinery assembly on the site and starting of the target gene transcription. The binding sites of the oxidative and electrophilic stress response specific TFs are called AREs (antioxidant response element) and they have been found in the regulatory regions of the phase II cytoprotective genes, such as glutathione S-transferase (GST), NAD(P)H:quinone oxidoreductase 1 (NQO1), glutamate-cysteine ligase (GCLM), and heme oxygenase-1 (HMOX1)[42].

One of the master TFs of the phase II genes and redox response is the nuclear factor (erythroid-derived 2)-like 2 (Nrf2) (Figure 6.). It forms a heterodimer with small Maf proteins when it is in the nucleus and bind to ARE elements[43]. Nrf2 belongs to the CNC (cap 'n' collar) family of b-Zip TFs, which contain a highly conserved basic leusine zipper structure. Other family members include p45, NF-E2, Nrf1 and Nrf3.[41] During the basal state Kelch like-ECH-associated protein 1 (Keap1) binds Nrf2 in the cytosol and promotes the ubiquitination of Nrf2, causing proteosomal degradation[44]. As Nrf2 activity is regulated by degradation and new Nrf2 is *de novo* synthesized, Nrf2 response to the oxidative and electrophilic stress is fast and sensitive. Electrophilic agents disrupt Keap1 interaction with Nrf2 by modifying cysteine thiols, which prevents proteosomal degradation of Nrf2. Nrf2 can also be activated by proteins that prevent Keap1 from binding to Nrf2[42].

Multiple diseases have increased oxidative stress levels, including atherosclerosis, metabolic syndrome, type II diabetes and cancer. Nrf2 is therefore a potential therapeutic target for multiple diseases to reduce oxidative stress levels. Many potent Nrf2 inducers have been introduced, such as sulforaphane, nitro-oleic acid (OA-NO2)[45], oltipraz and CDDO-IM[46]. Sulforaphane and OA-NO2 are commonly used compounds to activate Nrf2 in experiments.

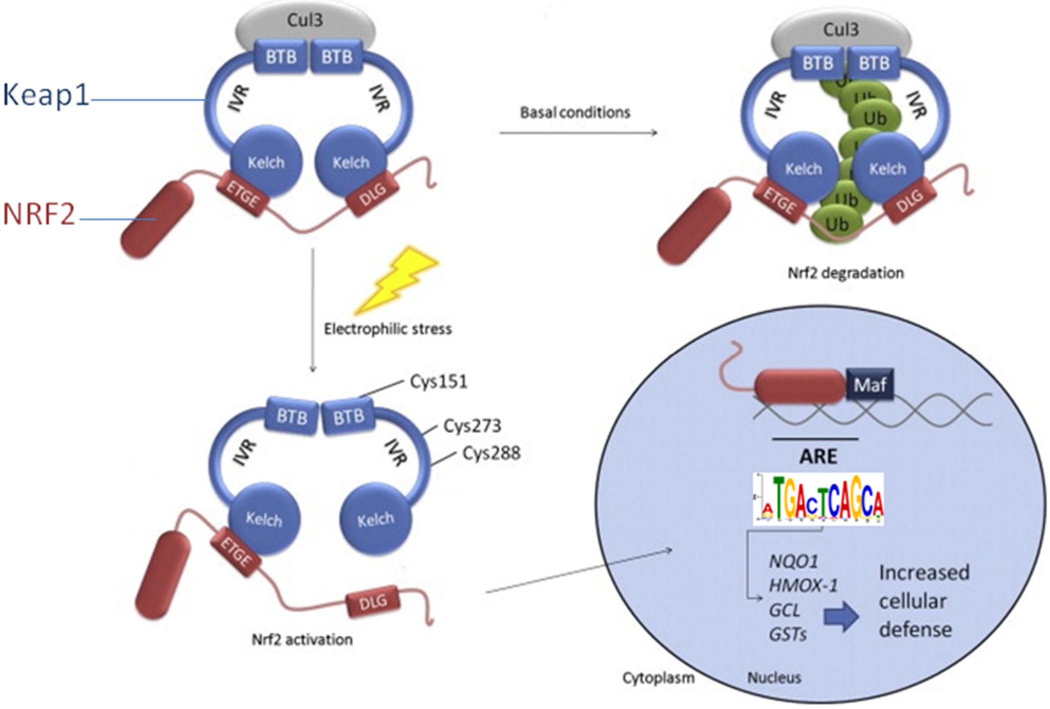


Figure 6. Nrf2 basal and induced conditions. During electrophilic stress Nrf2 enters the nucleus and forms a heterodimer with small Maf proteins. The Maf-Nrf2 heterodimer binds to the antioxidant response element (ARE) to upregulate the phase II cytoprotective target genes, such as *NQO1*, *HMOX1*, *GCL* and *GST*. In basal conditions, Nrf2 forms a complex with Kelch like-ECH-associated protein 1 (Keap1) in the cytosol, which is ubiqitinated and proteasomally degraded. Activating agents and stressors disrupts the Nrf2-Keap1 complex, as Keap1 cysteine thiols are modified. Illustration modified from ref7.

Nrf2 has been shown to regulate a substantial number of genes. Microarray analysis revealed thousands of putative targets from Nrf2 deficient (Nrf2 null) mouse data, based on expression level changes[51]. To detect direct targets (regulatory regions of genes bound by Nrf2) experimentally, the method of chromatin immunoprecipitation (ChIP) was developed. In ChIP proteins are covalently fixed to DNA and chromatin is fragmented. An antibody against the protein of interest (here Nrf2) is used to precipitate the protein-DNA complexes from the lysate and proteins are removed from the DNA to quantify the DNA levels, which can be interpreted as the amount of protein binding in the region. Over 600 targets were identified in mouse embryonic fibroblast from Nrf2 ChIP-sequencing data analysis[47] and over 200 Nrf2 targets with ARE homologous sequences were identified in the human lymphoblast ChIP-seq analysis, suggesting that Nrf2 has vast amount of direct targets[48]. In addition a ChIP-seq analysis has been done for the mouse hepatoma and embryonic fibroblast to study the Nrf2-Maf heterodimer targets[49]. Therefore Nrf2 is likely to have high number of direct targets and even more indirect targets. Nrf2 target composition might depend on the cell type, basal vs. induced conditions, activator concentrations and duration of the activation. However, knowledge about these events is very limited and therefore it might be too optimistic to assume that Nrf2 could be used to prevent and treat diseases. Many individual Nrf2 targets have been identified mainly using expression data and ChIP-seq, but only a few studies have been done to study Nrf2 regulated pathways and Nrf2 interactions[50]. Systems biology might be able to elucidate the role of Nrf2 regulatory network and its dysfunction in different diseases.

2.6.2 Nrf2 in cancer

Cancer cells have altered cellular metabolism to produce more ATP, increase biosynthesis and, in order to maintain their proliferative phenotype, to regulate appropriate cellular ROS (reactive oxygen species) levels. The Warburg effect is a well-characterized metabolic phenotype in cancer to produce ATP through an inefficient aerobic glycolysis even in the normal oxygen level environments. The gain of this metabolic shift is that the activation of glycolysis leads to the activation of the pentose phosphate pathway (PPP) and other metabolic pathways that produce molecular building blocks needed in the synthesis of proteins, lipids and nucleotides important for cell proliferation[51] (Figure 7.). Furthermore it is becoming clear that cancer cells not only need increased amounts of metabolites to proliferate, but also a tight maintenance of the ROS balance[52]. High amounts of ROS are produced by rapid protein translation and proliferation rates typical for cancer cells. As ROS damages DNA and proteins, it can induce apoptosis and senescence. On the other hand, cancer cells are thought to tolerate increased ROS levels, as ROS induced mutagenesis and proliferative stimulus can be beneficial for the cancer development[52] (Figure 8). Activated PPP produces NADPH that is an important reducing agent for enzymes that have a crucial role in the anabolic reactions and in ROS neutralization[52]. Therefore ROS sensing pathways in cancer cells may provide a growth advantage by maintaining sufficient expression of antioxidant enzymes to prevent ROS from accumulating at hazardous levels[52].

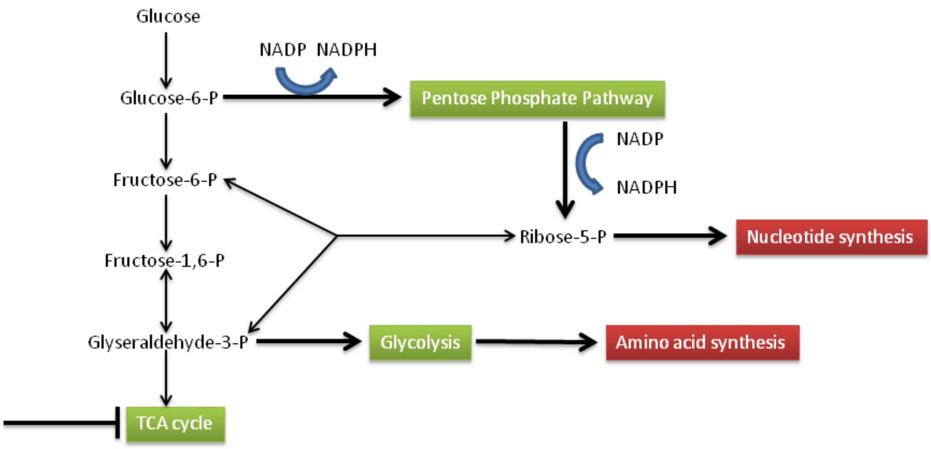


Figure 7. A simplified diagram of aerobic glycolysis (Warburg effect) and its link to the pentose phosphate pathway (PPP). In cancer metabolic fluxes through PPP and glycolysis have increased due to the cell proliferation stimulating signaling and possible lack of metabolic pathway controlling enzymes. Growth factor stimulus and resulting tyrosine kinase signaling prevents glyceraldehydes-3-P from entering the TCA cycle and its flux is directed to the glycolysis[51]. Late stages of glycolysis are directed to produce intermediates that are needed for amino acid synthesis. Accumulation of Glucose-6-P activates the PPP, which produces NADPH and results in intermediates for nucleotide synthesis[53]. NADPH is essential cofactor that provides the reducing power for glutathione and thioredoxin to neutralize ROS and also other enzymatic reactions needed in macromolecular biosynthesis[52].

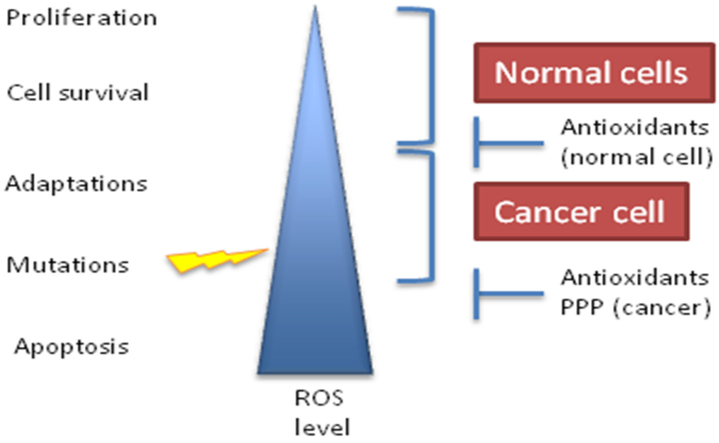


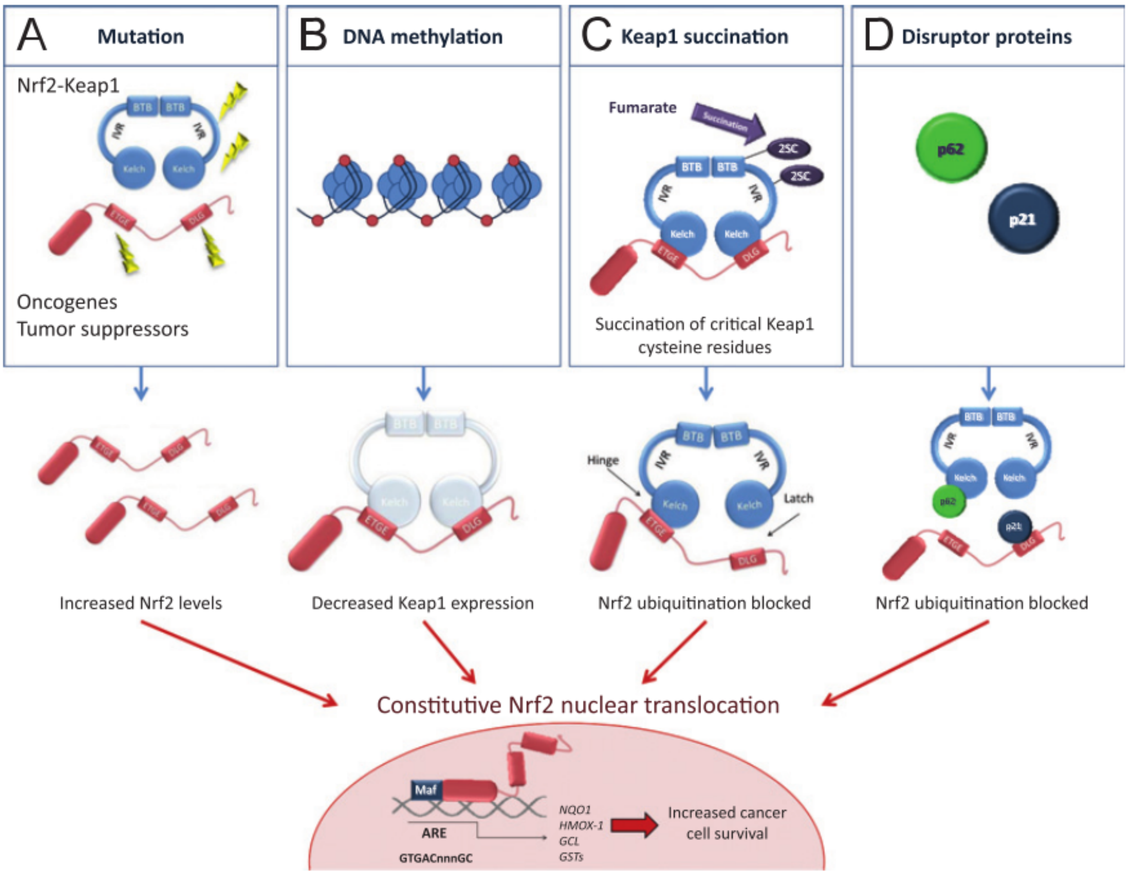
Figure 8. Maintenance of ROS levels in normal and cancer cell. ROS can support cell proliferation and survival pathways by inducing post translational modifications in tyrosine kinases and phosphatases[52,54]. Antioxidants prevent mutations and apoptosis by decreasing the ROS levels. In cancer cells metabolism and proliferation produce ROS, but adaptations and beneficial effects of ROS (proliferative stimulus, survival signals and mutagenesis), makes cancer cells more tolerant to increased ROS levels[52,55]. However, deadly amounts of ROS are neutralized by antioxidants and PPP production of cofactor NADPH[53].

Nrf2 is a key regulator of the antioxidant response and it has been reported that Nrf2 is constitutively activated by various distinct mechanisms, including mutations, epigenetic changes and disruptor proteins in many cancer types (Figure 8.). Also Nrf2 and Keap1 imbalance caused by NF-kB induction of Nrf2 expression has been shown to cause constitutive Nrf2 activity in acute myeloid leukemia[56]. Constitutively activated Nrf2 increases chemoresistance in cancer cells by inducing the detoxifying phase II enzymes[57,58]. Moreover, Nrf2 influences cell proliferation by directing glucose and glutamine to anabolic pathways and promoting the PPP activation[59]. The combined effect of the resistance to drugs, pro-survival signals and influences in the cancer proliferation gives an advantage to cancer types with constitutive Nrf2 status. Unsurprisingly cancers with high Nrf2 levels have a poor prognosis[59]. There is a need for Nrf2 specific inhibitors, but developing such a drug has been challenging due to the similarity of Nrf2 with other bZip family members[60].

This is contradictory to the vast amount of evidence that Nrf2 can suppress carcinogenesis that motivated the development of numerous Nrf2 activator drugs[61,62]. Lack of Nrf2 activity was reported to increase carcinogenesis in Nrf2 knockout mice and increased Nrf2 activity decrease carcinogenesis[61]. This paradox of the dual role of Nrf2 in cancer is not fully understood, but one suggestion is that increased ROS is important in the early development of tumors and therefore activation of Nrf2 prevents tumorigenesis[61]. During further development of cancer, adaptations and mutations creates a new steady state for ROS levels: ROS levels are increased, and constitutive activation of Nrf2 regulated antioxidant systems prevents the ROS from accumulating at high levels[52,61]. Nrf2 might also have an effect on cancer metabolism or in promoting metastasis formation[61]. Increased Nrf2 activity could also be a response to other oncogenic changes in tumors or the role of Nrf2 might be highly cancer type or even subtype specific.

From a systems biology point of view, the Nrf2 regulatory network might have many steady states, which could explain the controversial roles. In normal conditions, Nrf2 regulatory network is activated only a short period at a time and mainly direct Nrf2 targets with ideal ARE binding sites are induced. If Nrf2 is constitutively activated, also weaker ARE elements can be bound by Nrf2, as more Nrf2 is accumulating to the nucleus. In addition, indirect Nrf2 targets will get regulated. Therefore, the Nrf2 regulatory network will expand and could obtain new steady states with cancer promoting effects. In addition, many Nrf2 regulatory network steady states might have a role in supporting normal cell proliferation and differentiation.

Activation or inhibition of Nrf2 has a high potential in the cancer treatment and more information about Nrf2 activity in cancer must be obtained to choose best suitable treatments for patients. In this project Nrf2 overactive cancers are predicted in CCLE and TCGA data sets using supervised and unsupervised machine learning enrichment analyses.



ref[60]

Figure 8. Constitutive activation of Nrf2 in cancer. A. Mutations in Keap1 or Nrf2 are common in cancers and these mutations disrupt the Keap1-Nrf2 complex therefore causing constitutive Nrf2 activation in non-small cell lung cancers[57]. B. Keap1 promoter regions are hypermethylated in lung and prostate cancers, which reduces Keap1 expression levels and therefore increases Nrf2 nuclear translocation[63]. C. Decreased activity of fumarate hydratase causes the accumulation of fumarate, and subsequently succination of Keap1 cysteines. Reduced levels of functional Keap1 increases Nrf2 translocation in the nucleus. This mechanism has been reported in papillary renal carcinoma[64]. D. Increased amounts of disruptor proteins such as p21 and p62 in cancer can also disturb Nrf2-Keap1 interaction by competitive binding to Nrf2 or Keap1[65].

2.6.3 Nrf2 in metabolism

The main role of Nrf2 is the activation of the redox response pathway, but Nrf2 has been shown to regulate other metabolic pathways as well. Nrf2 directs glucose and glutamine to anabolic pathways and promotes the PPP activation, as noted before. However, there may be other pathways relevant in the context of disease, which further analysis of metabolic gene expression could reveal. Nrf2 deficient (Nrf2 null) mice fed with high fat diets have been used to investigate the effect of Nrf2 in metabolism by our group and others. Nrf2 null mice have decreased levels of antioxidant genes, but also altered lipid metabolism. Specifically, Nrf2 was shown to inhibit lipid accumulation and lipid synthesis in mouse liver[66,67]. Proteomics analysis confirmed that Nrf2 regulates the synthesis and metabolism of fatty acids and other lipids in the liver, which affects the cellular lipid disposition[68]. As Nrf2 has been shown to be involved in lipid metabolism, Nrf2 has been investigated in mouse adipose tissue and shown to inhibit its development, supporting an important role in the maintenance of glucose and lipid homeostasis[69]. On the contrary, adipogenesis stimulatory roles have also been reported[70]. Based on Nrf2 ChIP-sequencing in human lymphoblast data, Nrf2 was proposed to regulate adipogenesis via regulating *RXRa* expression[48]. In mouse 3T3L1 cells, Nrf2 activation inhibited adipogenesis by downregulating *RXRa*[48]. Interestingly, RXRa has also been reported to inhibit ARE-driven gene expression by directly binding to Nrf2[71]. Therefore Nrf2-RXRa may form a regulatory loop in lipid metabolic reactions. Also, one key heterodimeric partner TF of RXR in adipocytes, PPARg, has been suggested to interact with Nrf2[50], which could explain the complex role of Nrf2 in the regulation of metabolism in various cell types.

As described above, most of the studies focusing on metabolic effects have used Nrf2 null mice on a high fat diet or cancer cells. As Nrf2 has been shown to be involved in numerous pathways, many levels of metabolic pathways could be affected in the Nrf2 deficient mice. The mice might also have adapted to regulate oxidative stress by compensatory pathways, which could affect metabolic processes. In cancer cells, the Nrf2 regulatory network might be different from nonmalignant cells due to the high constitutive expression of Nrf2, as well as other genomic alterations. More studies must also be conducted in human primary cells to understand Nrf2 role in metabolism during basal, induced and constitutively active states. Metabolic reactions are well defined and therefore it is possible to model the metabolic fluxes using systems biology, which might reveal disease promoting metabolic changes. In this project, Nrf2 regulated pathways are identified from datasets collected from primary human umbilical vein endothelial cells (HUVECs) using enrichment analysis tools.

3 AIMS OF THE STUDY

The general idea of this thesis and related future work is to discover diseases in which Nrf2 can play a central role, and how Nrf2 is involved in the pathogenesis of those diseases. The effect of Nrf2 in disease can depend on cell or tissue type. However, Nrf2-dependent protection is crucial in all cell types and therefore the pathway should be ubiquitous. Our hypothesis is that Nrf2 signature can be identified from Nrf2-activated samples and in cancer cells overexpressing Nrf2. Our second hypothesis is that Nrf2 regulates metabolic pathways. The two aims that are investigated in the experimental part of this thesis to verify these hypotheses are:

Aim1: Prediction of constitutively active Nrf2 in cancers using an Nrf2 signature

Aim2: Identification of Nrf2-regulated metabolic pathways in normal human endothelial cells by pathway analysis

In the first aim, the computational model used aims to capture from the data a subset of expression profiles typical of the constitutive activation of Nrf2 (hypothesized to occur in cancer) using overexpression and activation by OA-NO2 data in HUVECs. This regulatory signature is subsequently tested in many different cancer types: if Nrf2 expression profiles are similar to the Nrf2 activation signature, the particular cancer cell type is likely to constitutively express Nrf2. In the second aim, perturbations of the Nrf2 regulatory network are studied using many different conditions, as samples are collected from activation, overexpression and knockdown experiments in HUVECs. Each of these experiments provides evidence for Nrf2-dependency of gene expression. Combined enrichment analysis is used to reveal Nrf2-dependency at the pathway level. Consequently, Nrf2-regulated pathways are further explored to identify putative direct targets by examining next generation sequencing data sets to reveal candidate regulatory regions and ARE motifs.

4 MATERIALS AND METHODS

4.1 Genome-wide gene expression data

Microarray data from Kansanen et al. [72] and Jyrkkänen et al. (unpublished data) were used in these analyses (Table 1.). These represent studies were Nrf2 was activated transiently by activating ligands (OA-NO2) or constitutively using adenoviral overexpression, or silenced using siRNAs. Microarray samples were chosen according to the experimental design of the enrichment analysis, described in the corresponding section. Data could be used to assess the activating ligand effects: by comparing OA and OA-NO2 samples, ligand dependence could be assessed as OA does not activate Nrf2. Also untreated samples could be compared to OA to study its effect. As siRNA mediated knockdown of Nrf2 is not 100 % efficient, the siNrf2 samples may show Nrf2-dependent gene expression by OA-NO2 but at reduced levels.

Table 1. Definitions of the HUVEC samples used in the enrichment analyses

Platform	type	Experiment	transfection/ transduction	timepoint	timepoint ligand	treatment
Microarray *	Affymetrix hgu133Plus2	Nrf2 Overexpression	adCMV (ctrl)	36 h, 72 h	-	-
			adNrf2	36 h, 72 h	-	-
Microarray **	affymetrix hgu133Plus2	Nrf2 knockdown	siCTRL	24 h	8h	untreated, OA-NO2, OA
			siNrf2	24 h	8h	untreated, OA-NO2, OA

All microarray samples were triplicates. Oleic acid (OA) does not activate Nrf2, nitro-oleic acid (OA-NO2) activates Nrf2, and methanol was control for OA-NO2.In adNrf2, Nrf2 is overexpressed by adenoviral transduction. adCMV does not contain genes, but was used as a control for the transduction effects. In siNrf2, Nrf2 is silenced by a specific siRNA.

* Data from Jyrkkänen et al. unpublished

** Data from Kansanen et al. [72]

4.1.1 Microarray data analysis

Raw data files were normalized using GC-RMA. R 2.14 and Bioconductor were used for data processing and quality control. Limma package was used for statistical analysis. Two-tailed t-test was used to compare specific samples to respective control samples. The Benjamini-Hochberg FDR method was used for adjusting P-values for multiple comparisons.

4.1.2 RNA and GRO-sequencing visualization tracks

HUVECs were treated with OA-NO2 or solvent (metOH). The time point for GRO-seq was 2h and 12h and for RNA-seq 8h - data from Kansanen and Kaikkonen et al. unpublished. RNA-seq measures the mature transcript levels for coding RNAs mainly in the cytosol. Mature transcripts are processed and transported to the cytosol and therefore 8h after the OA-NO2 activation the mature transcripts can be measured. GRO-seq measures the primary transcript levels in the nucleus. Therefore primary transcripts are measured after 2h and 12h to detect changes in the primary transcript expression and seeing the return to basal state. GRO-seq data can also reveal short non-coding enhancer RNA (eRNA) sites that can be used in detecting active enhancer sites[73]. eRNAs are transcribed using DNA as a template on enhancer regions. Their function is not fully characterized.

RNA- and GRO-seq data libraries were created and Illumina Genome-analyser II was used for the sequencing. The data was analyzed according to Wang et. al[73]. Briefly, base calling was done to identify reads and reads were mapped to reference sequence (hg19) using the Bowtie tool. Read counts were normalized to the total number of mapped reads for each sample and counts per genomic position were quantified to produce a signal track across each chromosome. RNA, GRO, ChIP, FAIRE and DNAse sequencing data are used in this Master’s Thesis for visualization of activity of gene regulatory regions. More information about the methods can be found in [73–76], respectively.

4.1.3 CCLE and TCGA data sets

917 human cancer cell lines included in Broad Novartis Cancer Cell Line Encyclopedia (CCLE, GEO ID: GSE36133, http://www.ncbi.nlm.nih.gov/gds) were analyzed to identify cancer cell lines with increased Nrf2 activity. Raw data files were normalized using GC-RMA in R.

The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga) data for 604 glioma patient samples were used to identify glioma samples with overactive Nrf2. Data downloaded was already processed to level 3 data, which means that data has been normalized and that it is ready to use.

4.2 Nrf2 signature prediction in CCLE

4.2.1 SIGNATURE analysis

An accurate indicator of Nrf2 activity had to be developed to predict Nrf2 activity in CCLE samples. Nrf2 activation creates a characteristic signature in the genes that are directly or indirectly regulated by Nrf2. SIGNATURE software available in genepattern (https://genepattern.genome.duke.edu)[2] and its module “createsignature” was used to create the Nrf2 signature model by using microarray data from our previous studies (Table 1). Signature is used to distinguish two biological states from each other by a supervised machine learning algorithm, so two training sets was created from the microarray experiments: samples were Nrf2 pathway is active and where it is inactive. Active training set contained triplicates of adNrf2 overexpression data (36h, 72h) and siCTRL data with OA-NO2 treatment for 8 h, which activates Nrf2. Inactive training set contained triplicates of adCMV samples and siCTRL data with no treatment and OA treatment, which has no effect on Nrf2 activity (Table 1.). Pathway had to be active or inactive in all representative samples; therefore Nrf2 knockdown samples were not used in the training sets, as knockdown is rarely 100 % and off target effects are common.

Probability of active Nrf2 signature was computed in each CCLE cell lines (test set). The training sets and two metagenes were used to create the Nrf2 signature model, which was set to contain a total of 100 genes. These parameters simplify the supervised learning model and are used in SVD dimension reduction to prevent overfitting of the data. Quantile and Shift-Scale normalization was also applied to the microarray data to successfully integrate training data with the CCLE data. 1000 runs were used for burn-in and 5000 samples were collected for the model. Markov Chain Monte Carlo (MCMC) algorithm used in SIGNATURE uses burn-in period to discard initial runs that might not be stationary before collecting specified number of samples, because the Markov Chain is not stabilized in the early runs. Other parameters were set to default.

4.2.2 Definition of the Bayesian probit regression model in SIGNATURE

A Bayesian model specifies a prior probability for some event and updates the model when relevant evidence is given to provide new posterior probability for the event. Bayesian probit regression model is a statistical method that is suitable for high dimensional gene expression data. In supervised machine learning the probit model estimates the likelihood of the sample to be in one of the binary categories, such as Nrf2 active and Nrf2 inactive categories. There are many steps to infer this information. Posterior probability is conditional on the evidence obtained, such as gene expression profile of the sample. The idea is to first learn about the training data by inferring the regression parameters in each training sample. As learning about the data is done in predefined binary conditions, the regression parameters are constrained and easier to infer. Bayesian context and SVD provides additional constraints to help learning about the data. After genes in the training set are fitted to binary conditions, Markov Chain Monte Carlo (MCMC) with Gibbs sampling algorithm is used to simulate the posterior distribution of the test set regression parameters from exact posterior distribution of the regression parameters in the training set. Each sampling step simulates draws that are approximately from a posterior distribution and these draws can be used to compute the likelihood of sample to belong to either of the binary groups and also their 95 % credible intervals. These probabilities can be interpreted as the probability of an active signature pathway in the corresponding sample. Bayesian probit regression model, SVD and standard iterative MCMC has been used similarly as in SIGNATURE, for more information about the algorithms refer to references[34,77–79].

Bayesian probit regression model:

Bayesian probit regression model can be used to predict the posterior probability of a random event using prior probabilities. Posterior probability is conditional on the evidence obtained. Therefore regression parameters **γ** must be estimated from the evidence data prior probabilities to train the model to compute the posterior probabilities.

Training set contains expression values for n samples and p genes in a matrix X. Sample i in X is then a vector of expression values ($X_{1,i...}X_{p,i}$). It is assumed that the samples can be divided into binary classes, such as signature on/off. This is a vector ($Y_{1...}Y_n$), where $Y_i = 1$ if i:th sample belongs to class 1 and if not sample belongs to class 0. The entire set of p genes is included as the predictor variables for the sample class. There is a regression parameter for each gene, γ = ($\gamma_{1...}\gamma_p$).

Cumulative distribution function in probit regression model is a sigmoidal function with values ranging from 0 to 1. It describes the probability that random variable that is normally distributed will be found at a value less than or equal to value $X_i`\gamma$. Therefore higher values for $X`\gamma$ results in a higher posterior probability for $P(Y_i=1)$.

Bayesian fitting of a standard binary probit regression model is done for the samples in the training set:

$$\mathrm{P}(\mathrm{Y}_i = 1|\gamma) = \sigma(X_i`\gamma) \quad (1)$$

Y Binary class predictor value

$\mathbf{P(Y_{\mathit{i}} = 1|\gamma)}$ Is the posterior probability that sample i belongs to class 1, when parameters γ are given as evidence

σ cumulative distribution function of normal distribution

$\mathbf{X_i}$ is vector of gene expression levels for i:th sample n

γ vector of p unknown regression parameters

SVD reduction of dimensionality and estimating posterior probabilities

There are several thousand dimensions as $n << p$, which makes the estimation of the regression parameters $\boldsymbol{\gamma}$ unreliable. Therefore, a special projection is done and genes in the gene sets are converted to eigenvectors (“metagenes”) using linear algebra factorization SVD and first eigenvector with highest eigenvalue is used to define the whole gene set activity level in the sample. Further details about SVD are presented in references [33,34].

$$X=ADF \qquad (2)$$

A is p x k (orthonormal matrix)

D is k x k (diagonal matrix

F is n x k (orthonormal square matrix)

Where k is the number of eigenvector and eigenvalues to include in the model (typically 1~5)

X is included in the regression model $\mathrm{P}(\mathrm{Y}_i = 1|\gamma) = \sigma(X_i`\gamma)$:

$$\mathrm{P}(\mathrm{Y}_i = 1|\beta) = \sigma(\mathrm{F}'\mathrm{DA}'\gamma) = \sigma(\mathrm{F}'\beta) \qquad (3)$$

Y Binary class predictor value

$\mathbf{P}(\mathbf{Y}_i = \mathbf{1}|\boldsymbol{\beta})$ Is the posterior probability that sample i belongs to class 1, when parameters γ are given as evidence

σ cumulative density function of normal distribution

β=DA′ γ are the regression parameters that have been reduced using SVD

Regressions on genes are reduced to regression on “metagenes” and estimation of regression parameters β is less problematic. SVD produces a special projection where sample descriptors are orthogonal to each other, which makes the computationally efficient use of standard MCMC possible. MCMC is used in test set to compute approximation of the posterior probability using draws from the exact posterior probability of β.

4.3 GSVA enrichment analysis in CCLE and TCGA datasets

GSVA[29] implementation available as an R/Bioconductor package was used to confirm the SIGNATURE result in CCLE data set and also to predict overactive Nrf2 signature in glioma patient samples from a TCGA data set. A gene list was created in SIGNATURE as described earlier, using Pearson correlation to obtain genes with best distinction between phenotypes. The same genes were chosen for the GSVA analysis to be able to compare the tools directly. However, only the upregulated genes of the signature model (80 %) were selected for the analysis. This separation needed to be done, because genes with expression levels in both directions would cancel the effect of each other during enrichment calculations following the main GSVA algorithm.

Random gene sets were permutated 1000 times to obtain the empirical null distribution. Genes in the gene set were randomly selected, so that each gene could be added to the gene set only once. Each gene set contained 80 genes because the same amount of genes was originally included in the Nrf2 signature upregulated genes. The gene sets where then analyzed in GSVA to derive the enrichment scores for each TCGA or CCLE samples. A P-value could simply be calculated by counting the number of higher than observed scores in permutations and divided that number by the number of permutations.

R code:

```
# simulating empirical null distribution

eFDR=function(i){length(which(simulated_ES[,i]>observed_ES[1,i]))/length(simulated_ES[,1])}

m=unlist(lapply(1:length(observed_ES[1,]), eFDR))

# normal distribution

normald=function(i){pnorm(observed_ES[1,i],mean=mean(observed_ES),sd=sd(as.numeric(observed_ES)), lower.tail=F)}

m=unlist(lapply(1:length(observed_ES[1,]), normald))
```

4.3.1 Description of the algorithms in GSVA

Both SIGNATURE and GSVA tools are developed to predict classes in Q2 type questions. However, GSVA is very different from SIGNATURE, as it is an unsupervised method and uses a competitive test. GSVA transforms the gene by sample matrix to gene set by sample matrix, without using prior knowledge about the phenotypes. GSVA can be used in large data sets to compute the degree of coordinately up/downregulated gene sets within a sample. Therefore if many Nrf2 signature genes are highly expressed in the sample, Nrf2 is likely to be active.

Description of GSVA enrichment score calculation

Matrix X contains expression values for j samples and i genes. GSVA estimates a cumulative distribution function for each gene expression profile $X_i = (X_{i1...}X_{in})$ by using a Gaussian kernel. This sets expression profiles to a similar scale to be able to determine if gene i expression levels are highly or lowly expressed in sample j compared to population distribution of all samples (4). Therefore each gene gets values ranging from 0 to 1 based on population distribution (row-wise operation). Cumulative distribution function can also be estimated for RNA-seq data, using a discrete Poisson kernel[29].

Gaussian kernel function:

$$\hat{F}_{h_i}(x_{ij}) = \frac{1}{n}\sum_{k=1}^{n}\int_{-\infty}^{\frac{x_{ij}-x_{ik}}{h_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

(4)

ref[29]

$\mathbf{\hat{F}_{h_i}(Xij)}$ is the expression level statistics for gene i in the sample population distribution

i genes, j samples

$\boldsymbol{x_i}$ is gene expression profile for gene i

$\boldsymbol{h_i} = \boldsymbol{s_i}/4$ is the resolution of the kernel estimation for gene i

$\boldsymbol{s_i}$ is sample standard deviation of the i-th gene

Sample-wise operation is done to expression level statistics $\hat{F}_{h_i}(X_{ij})$ (denoted now as Z_{ij}). Z_{ij} is converted to ranks $Z_{(i)j}$ and normalized symmetrically around zero (5). The purpose of this step is to up-weight the two tails of the ranked distribution to weight the high or low rank genes during enrichment score calculation.

(5) $rij = |p/2-Z_{ij}|$

$\boldsymbol{r_{ij}}$ is the ranked and normalized expression level statistics for the sample

$\mathbf{Z_{(i)j}}$ is rank of gene expression level statistics for each sample

$\boldsymbol{p}$ is number of genes in the dataset

Weighted Kolmogorov-Smirnov (KS) random walk statistics are commonly used for enrichment score computation[8,18,20]. During the enrichment score computation, the sample l gene expression level statistics $\boldsymbol{r_{ij}}$ are converted to the Weighted KS random walk statistics (6). These operations describe the fraction (or a distribution) of high or low ranked genes in a gene set in the corresponding sample.

$$\nu_{jk}(\ell) = \frac{\sum_{i=1}^{\ell} |r_{ij}|^{\tau} I(g_{(i)} \in \gamma_k)}{\sum_{i=1}^{p} |r_{ij}|^{\tau} I(g_{(i)} \in \gamma_k)} - \frac{\sum_{i=1}^{\ell} I(g_{(i)} \notin \gamma_k)}{p - |\gamma_k|}$$

(6) ref[29]

$\boldsymbol{V_{jk}(l)}$ is enrichment score

$\boldsymbol{r_{ij}}$ is ranked and normalized expression level statistics

τ is a parameter describing the random walk weight of the tail

γk is the k-th gene set

I(g(i) ∈ γk) is the indicator function (i gene belong to gene set γk)

|γ**k|** is the number of genes in k-th gene set

p is number of genes in the dataset

l is number of samples in the dataset

KS statistics can be turned into GSVA scores (enrichment statistics) by two methods: similar to GSEA by the maximum deviation from zero method or using the GSVA normalized enrichment statistics (figure S1). Purpose of this step is to evaluate, whether the sample is negatively or positively correlated with the gene set. The enrichment statistics also produce distribution of scores that can be used to compare gene set enrichment in the sample relative to all samples and to determine the significance cutoff for the results.

Maximum deviation of the random walk from zero for the j-th sample and k-th gene set:

$$ES_{jk}^{\max} = \nu_{jk}[\arg\max_{\ell=1,\dots,p} \left|\nu_{jk}(\ell)\right|]$$

(7) ref[29]

$\boldsymbol{ES_{jk}^{max}}$ score, maximum deviation from zero method

$\boldsymbol{V_{jk}(l)}$ is enrichment score

Or GSVA score (default):

$$ES_{jk}^{\text{diff}} = \left|ES_{jk}^{+}\right| - \left|ES_{jk}^{-}\right| = \max_{\ell=1,\dots,p}(0,\nu_{jk}(\ell)) - \min_{\ell=1,\dots,p}(0,\nu_{jk}(\ell))$$

(8) ref[29]

$\boldsymbol{ES_{jk}^{+/-}}$ are largest positive and negative random walk deviations from zero for sample *j* and gene set *k*

4.4 Stouffer´s method and Hypergeometric test for metabolic pathway analysis

Nrf2 regulated metabolic pathways were identified by performing hypergeometric tests on the human metabolic reconstruction (Recon1)[80] metabolic pathways. First, to obtain lists of differentially expressed genes, sample-wise T-tests were done to microarray data. Microarray samples that were compared were: siNrf2 -siCTRL untreated, siNrf2-siCTRL OA, siNrf2-siCTRL OA-NO2, siNrf2-siNrf2 OA-NO2 - OA, siCTRL-siCTRL OA-NO2 - OA, adNrf2-adCMV 36h, adNrf2-adCMV 72h (Table 1). Each of these experiments tests the null hypothesis that gene expression is not dependent on Nrf2-perturbation (alternative hypothesis being that Nrf2-dependence is observed). P-values obtained from T-tests were adjusted for multiple testing and subsequently combined using Stouffers´s method. P-values were combined, because it is desirable to neglect the direction of differential expression to identify pathways with both up- and downregulated genes, as Nrf2 perturbations (such as Nrf2 knockdown and Nrf2 overexpression) will have different effects on gene expression levels. Genes that were significantly differentially expressed, but not consistently expressed in different perturbations, were discarded (for example: gene is upregulated during Nrf2 activation but downregulated during overexpression). The Stouffer´s test function was included in the R/Bioconductor MADAM package. P-value of 0.05 was chosen as a cutoff to define the list of interesting genes for a hypergeometric test.

Stouffer’s method

$$Z = \frac{\sum_{i=1}^{k} z_i}{\sqrt{k}} \qquad (9)$$

$\mathbf{Z}$ is Z-score

$\boldsymbol{Z_i}$ is $\Phi - 1(1 - p_i)$, where Φ is the cumulative density function of normal distribution

p_i is P-value for the i:th hypothesis test

Hypergeometric test was also done to all samples individually to test for overrepresentation of pathway terms. In addition, tests were done to separate lists of upregulated and downregulated genes from individual samples to determine if there is a pattern in the enriched pathways based on the experimental conditions, such as knockdown and overexpression and ligand effects on the pathways. Metabolic pathways were obtained from Recon1[80], which contains a total of 9812 genes. Hypergeometric test was done to all Recon1 pathways containing differentially expressed genes. Gene set minimum size was set to 4 (unique gene symbols, contains alternative transcripts).

Hypergeometric test:

$$\mathrm{Pval}(X = A) = \frac{\binom{B}{A}\binom{C-B}{D-A}}{\binom{C}{D}} \qquad (10)$$

A is the number of successes

B is the number of successes in the population

C is the size of the population

D is the number of draws

Example of Hypergeometric test for cholesterol metabolism pathway enrichment:

R code:

```
A=length(grep("Glutathione Metabolism", Total.met.genes$pathway, fixed=T)) # 14

B=length(grep("Glutathione Metabolism", All.met.genes$pathway, fixed=T)) # 29

C=length(All.met.genes$pathway) # 5119

D=length(Total.met.genes$pathway) # 1193

phyper(A-1, B, C-B, D, lower.tail=F) # 0.002768845
```

4.5 GSEA for metabolic pathway analysis

GSEA is commonly used for enrichment analysis. Java based GSEA-P program from http://www.broadinstitute.org/gsea/downloads.jsp was used for the enrichment analysis. Significance was estimated using 1000 permutations of the sample labels.

GSEA requires distinguishable phenotype for the analysis. In addition, it requires more than 7 samples per phenotype if sample permutation is done instead of gene set permutation. The same samples that were used in SIGNATURE analysis were used for the analysis: siCTRL untreated, siCTRL OA, adCMV 36h, adCMV 72h for phenotype 0 and adNrf2 36h, adNrf2 72h, siCTRL OA-NO2 for phenotype 1 (Table 1.). Phenotype 1 was compared to phenotype 0, specified by the custom class file. Custom gene sets were created from Recon1 metabolic pathways. Gene set minimum size was set to 4, as many of the Recon1 gene sets are smaller than the default parameter 15. All the other parameters were set to default.

For more information about running GSEA and input file generation refer to:

http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Main_Page

5 RESULTS

5.1 Nrf2 hyperactive cell lines in CCLE

Cancer cell lines with constitutive Nrf2 were identified using the SIGNATURE software: Characteristic Nrf2 signature was created in predefined training sets where Nrf2 redox pathway is on/off. Nrf2 signature was used to classify cancer cell lines based on Nrf2 activity using CCLE dataset. Furthermore SIGNATURE results were confirmed using GSVA. In GSVA samples are not classified based on predefined categories, and it provides relative enrichment between the samples.

From a total of 917 human cancer cell lines in CCLE, 98 (ca. 11%) had over 95 % probability in SIGNATURE results to have constitutive Nrf2 activity (Table 2). 77 (ca. 80 %) of the cell lines predicted by SIGNATURE were also discovered by GSVA (eFDR 0.01) (Figure 8.).

Lung cancers had frequently overactive Nrf2 status based on both analysis methods (Table 1). This could be expected as it has been previously reported that mutations in Keap1 is common in non-small cell lung cancer cell line A-549, which was among the top cell lines with overactive Nrf2. These mutations disrupt the Keap1-Nrf2 complex ubiquitination process and therefore cause constitutive Nrf2 activation[57].

GSVA was able to detect 139 cell lines with overactive Nrf2 that did not pass the 95 % threshold in the SIGNATURE analysis (Table 2.). These cell lines could mostly be included in the same cancer type categories as with SIGNATURE, excluding 33 cell lines that formed 5 new categories. Most interesting new category in GSVA results was malignant melanoma (25 cell lines), which was predicted to have constitutive Nrf2 activity in 43 % of the CCLE melanoma cell lines. The percentage of overactive Nrf2 cell lines from cell lines representing the same cancer type was highest for oesophagus carcinoma, kidney carcinoma and glioma (Table 2.).

There were five novel cancer types with overactive Nrf2 status: urinary tract carcinoma, glioma, mesothelioma, melanoma, and thyroid carcinoma (Table 1). Most interesting of these cell lines were gliomas (hyperactive in 10 cell lines, 21% of gliomas), as none of the glioma cell lines has previously been reported to have constitutive Nrf2 activity. GSVA predicted even more glioma cell lines with constitutive Nrf2 activity (28 cell lines, 60% of gliomas), including the ones already found by the SIGNATURE.

Table 2. Counts and percentages of cancer types in CCLE with overactive Nrf2
Cell type counts predicted to have overactive Nrf2 status according to SIGNATURE (probability > 95 %) and GSVA (FDR 0.01) in the CCLE data set. Cancer categories in bold include cancer cell lines that have not been previously characterized to have overactive Nrf2. In addition 43 % of melanoma samples had overactive Nrf2 status according to GSVA.

Cancer site	**Cancer Type**	**Total count in 98 & 216 Nrf2 overactive CCLE cell lines**		**Percentage in the same cancer types in CCLE (%)**	
		SIGNATURE	**GSVA**	**SIGNATURE**	**GSVA**
lung	carcinoma	42	47	25.5	28.5
oesophagus	carcinoma	11	11	44	44
central_nervous_system	**glioma**	10	28	21.3	59.6
liver	carcinoma	7	9	26.9	34.6
kidney	carcinoma	5	13	25	65
breast	carcinoma	5	18	8.9	32.1
upper_aerodigestive_tract	carcinoma	4	11	13.8	37.9
ovary	carcinoma	3	9	6.8	20.5
stomach	carcinoma	2	4	5.9	11.8
pleura	**mesothelioma**	2	5	22.2	55.6
large_intestine	carcinoma	2	5	3.6	9.1
urinary_tract	carcinoma	1	4	4.5	18.2
thyroid	carcinoma	1	5	9.1	45.5
pancreas	carcinoma	1	3	2.3	7
bone	giant_cell_tumour	1	10	4	40
biliary_tract	carcinoma	1	1	14.3	14.3

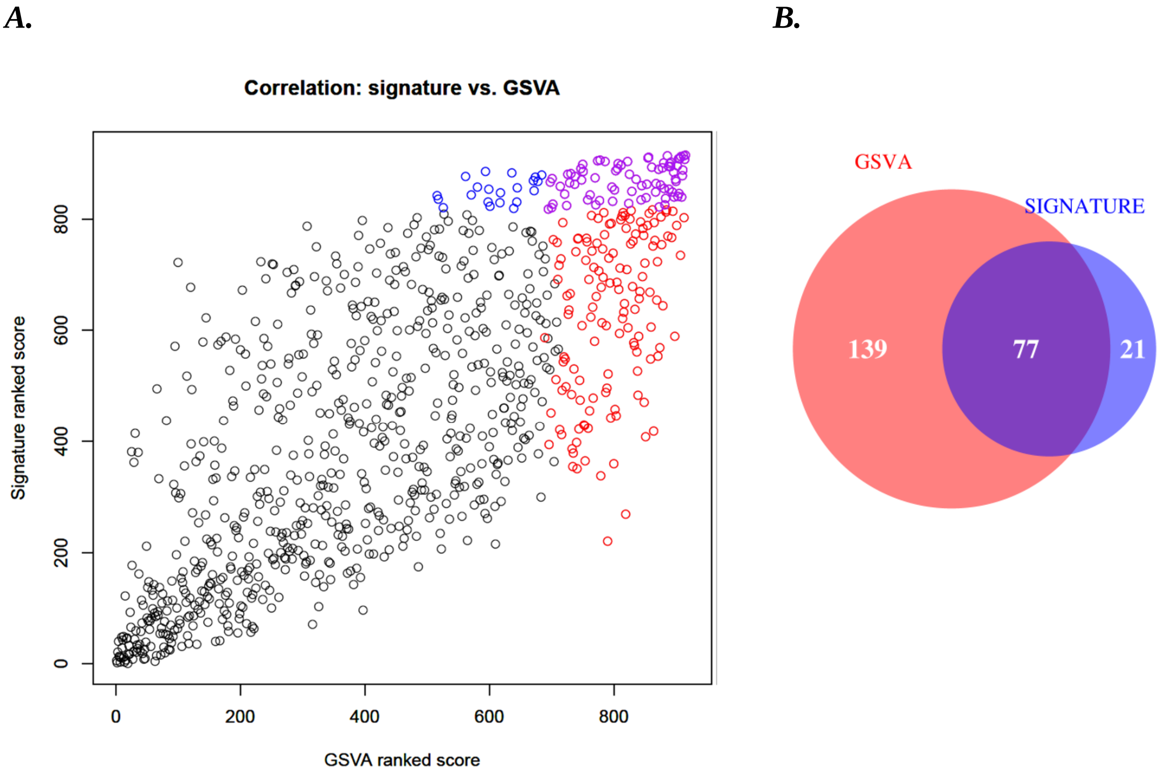


Figure 8. A. Scatter plot of correlation between GSVA and SIGNATURE scores. Ranked SIGNATURE probabilities for Nrf2 overactive cell lines are plotted on the Y-axis and corresponding ranked GSVA enrichment scores are plotted on the X-axis. Blue dots are CCLE cell lines with >95 % probability based on the SIGNATURE analysis and red dots represent samples with GSVA scores P-values <0.01. Purple dots are the overlapping CCLE cell lines high ranked by both of the two tools. B. Venn diagram of SIGNATURE and GSVA ranks, showing numbers of overlapping cell lines between tools on > 95 % and <0.01 P-value cutoffs.

5.2 Hyperactive Nrf2 glioma samples in TCGA

As many of the novel Nrf2 expressing cancer cell lines were gliomas, GSVA was done also for glioma patient samples collected in TCGA. SIGNATURE was not used for analyzing TCGA data, because of technical reasons (SIGNATURE can´t be directly used with Agilent microarray and RNA-seq platforms).

From a total of 604 glioblastoma multiforme samples, 60 (ca. 10 %) were predicted to have an Nrf2 overactive status (FDR 0.05) (Figure 9). Compared to the fraction of positive glioma cell lines, GSVA predicted a higher percentage of cell lines with overactive Nrf2 than observed in patient samples. Nevertheless, overactive Nrf2 signature was confirmed also in a 10 % proportion of clinical patient samples.

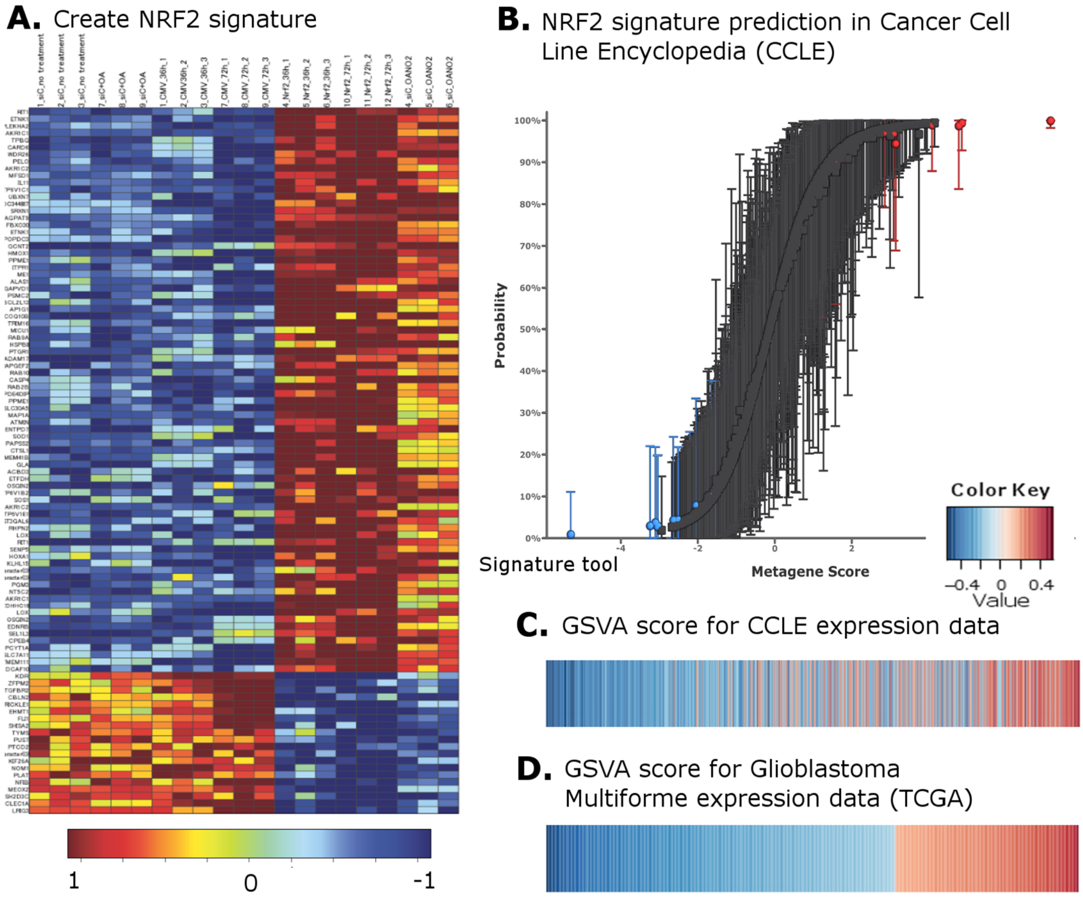


Figure 9. Prediction of Nrf2 overactivity in CCLE and TGCA data using the SIGNATURE and GSVA tools. **A**, A heatmap was created from the Nrf2 signature scores; warm colors (red and yellow) signify high expression and cool colors (shades of blue), low expression. The signature model training sets consist of 12 microarray samples with inactive and 10 samples with active Nrf2 status, which correspond to the columns of the heatmap in respective order. The expression profile of 80 upregulated and 20 downregulated genes relative to inactive samples are shown in rows. **B**, Cancer cell lines available from the CCLE dataset are shown ranked based on the probability of an active Nrf2 target gene signature. The plot shows individual samples on the X-axis and the probabilities and their confidence intervals on the Y-axis. Inactive training sets are marked as blue dots and active training sets as red dots. From the total of 917 samples, 98 (ca. 11%) had active Nrf2 signature with over 95 % probability. **C**, An independent analysis using GSVA was done to verify the SIGNATURE tool result. Cell lines where sorted as in the SIGNATURE plot and the corresponding GSVA scores are shown as a heatmap, where red corresponds to a higher score and blue represents lower score. From the total of 917 samples, 216 (ca. 24%) had active Nrf2 signature with FDR 0.01. **D**, GSVA analysis for the Nrf2 signature was done in glioma clinical samples in TCGA. GSVA scores were sorted from low to high and plotted in heatmap. From the total of 604 samples, 60 (ca. 10%) had Nrf2 active signature with FDR 0.05.

5.3 Metabolic pathway analysis

Nrf2 activity has a role in cancer metabolism as Nrf2 has been reported to regulate PPP in addition to the antioxidant response pathway. In normal cells, Nrf2 has been shown to regulate lipid metabolism in the liver and in adipocytes. Metabolic changes are common in cancer and hyperactive Nrf2 could promote metabolic reprogramming via changes in steady states. Moreover, other Nrf2 relevant diseases could be affected by Nrf2 regulated metabolic pathways. The first step of the pathway analysis was to identify the pathways that are enriched during Nrf2 pathway activation. The second step was to evaluate the reliability of the results and determine which pathways are consistently regulated by Nrf2 in different perturbations of Nrf2 status (activation, knockdown, and overexpression) to detect pathways that are primarily Nrf2 regulated. In the third step we tried to elucidate whether the consistently regulated and top pathways could be directly regulated by Nrf2. Recon1 metabolic pathway reconstruction was used to create the gene sets and to visualize selected metabolic network.

5.3.1 Identifying the metabolic pathways with Nrf2 dependent regulation

In the first step GSEA analysis was performed to the same samples used in the signature creation (Nrf2 ligand activation, overexpression and respective controls, Table 1.) to provide two clear phenotypes for the analysis and enough samples for both phenotypes to do sample permutation for significance assessment. ROS detoxification was not included in the significant pathways according to the GSEA analysis (Table 3.). However, glutathione metabolism got high enrichment scores as the pathway contains many known target genes of Nrf2, such as glutamate-cysteine ligases (GCLC and GCLM)[57]. Also heme degradation pathway contains known target of Nrf2, HMOX1. Other enriched pathways were related to transport, amino acid metabolism and lipid metabolism. We wanted to enrich metabolic pathways that were Nrf2 regulated and for GSEA analysis we were able to use only a limited amount of data. Therefore a second pathway analysis was done to extend the analysis to a wider range of Nrf2 perturbations by combining P-values from many different tests.

Multiple metabolic pathways were enriched using the hypergeometric test for genes with combined P-value less than 0.01 (Figure 10). Nrf2 related cytoprotective pathways were enriched, including ROS detoxification, and glutathione metabolism. The ROS detoxification has higher P-value than many other pathways that are not directly related to Nrf2 activation. Many fatty and amino acid metabolism related anabolic pathways were enriched. In addition, the previously reported Nrf2 regulated PPP was enriched. There was a good correlation with the GSEA tool, as 7/11 pathways were also found in Hypergeometric test: lysosomal transport, glutathione metabolism, glutamate metabolism, aminosugar metabolism, tyrosine metabolism, ascorbate and aldarate metabolism, and steroid metabolism. Interestingly also extracellular transport and cholesterol metabolism pathways were enriched.

Table 3. GSEA analysis results for significant metabolic pathways. Enriched pathways from GSEA using the significance cutoff for FDR Q-value 0.25 as suggested by the user guide.

Name	Size	Es	Nom P-Val	Fdr Q-Val	Rank At Max
Transport, Lysosomal	24	0.635	0.002	0.069	2820
Glutathione Metabolism	13	0.593	0.024	0.115	868
Aminosugar Metabolism	23	0.529	0.041	0.25	1365
Glutamate Metabolism	14	0.643	0.031	0.242	3034
Tyrosine Metabolism	40	0.49	0.06	0.244	262
Heme Degradation	4	0.82	0.051	0.206	563
Glycine, Serine, And Threonine Metabolism	23	0.506	0.045	0.236	2131
Fatty Acid Activation	8	0.656	0.103	0.224	3081
Glycerophospholipid Metabolism	60	0.409	0.073	0.201	1907
Ascorbate And Aldarate Metabolism	13	0.467	0.093	0.236	1607
Steroid Metabolism	31	0.477	0.037	0.24	518

In the second step pathways that were enriched in combined P-values and Hypergeometric test and GSEA were further analyzed to select interesting pathways for Recon1 network analysis. Hypergeometric test was performed also for individual comparisons and up- or downregulated genes (Figure 10.). If the pathway that was enriched in combined P-value test is not enriched in any individual tests, result might not be reliable, because extremely low P-values for a gene in a single Nrf2 perturbation could make the combined P-value for the gene significant. In addition there are many perturbations and if multiple genes will get significant P-value based on a single test, the pathway will get enriched. These individual tests could be used together with GSEA and Hypergeometric test results to evaluate consistency of the enriched pathways in different perturbations and to get an idea whether the pathway is up- or downregulated in each perturbation. In addition the individual tests could be used to evaluate whether the pathways are likely to be regulated by Nrf2 or if they are enriched due to expression changes caused by off-target effects, ligand effects, overactivated with viruses or potentially be regulated by another TF (such as only secondary response in overexpression).

ROS detoxification is consistent in the individual comparisons: The pathway is enriched in overexpression tests and also in upregulated genes in Nrf2 activator and overexpression tests. (Figure 10). PPP is enriched in upregulated genes during overexpression and downregulated when Nrf2 is reduced by siRNA. Lysosomal transport is enriched in upregulated genes with ligand activation and overexpression, suggesting a direct regulation. Ascorbate and aldarate metabolism and blood group biosynthesis pathways are enriched in many individual tests and upregulated genes. On the other hand pathways such as cholesterol and steroid metabolism and few other pathways are enriched only in overexpression tests. There are also many pathways, including fatty acid oxidation and methionine metabolism that are not enriched in any individual tests. These pathways could be enriched because of combining P-values: Numerous different genes might be significantly differentially expressed in one test, but not in the other test and therefore combining the P-values provides more significantly differentially expressed genes for the enrichment analysis.

After inspecting different analysis results, Transport Lysosomal, Glutamate metabolism, Ascorbate and Aldarate Metabolism pathways were selected for further analysis as a potentially directly Nrf2 regulated pathways. Cholesterol Metabolism and steroid metabolism pathways were studied as potential indirectly regulated pathways.

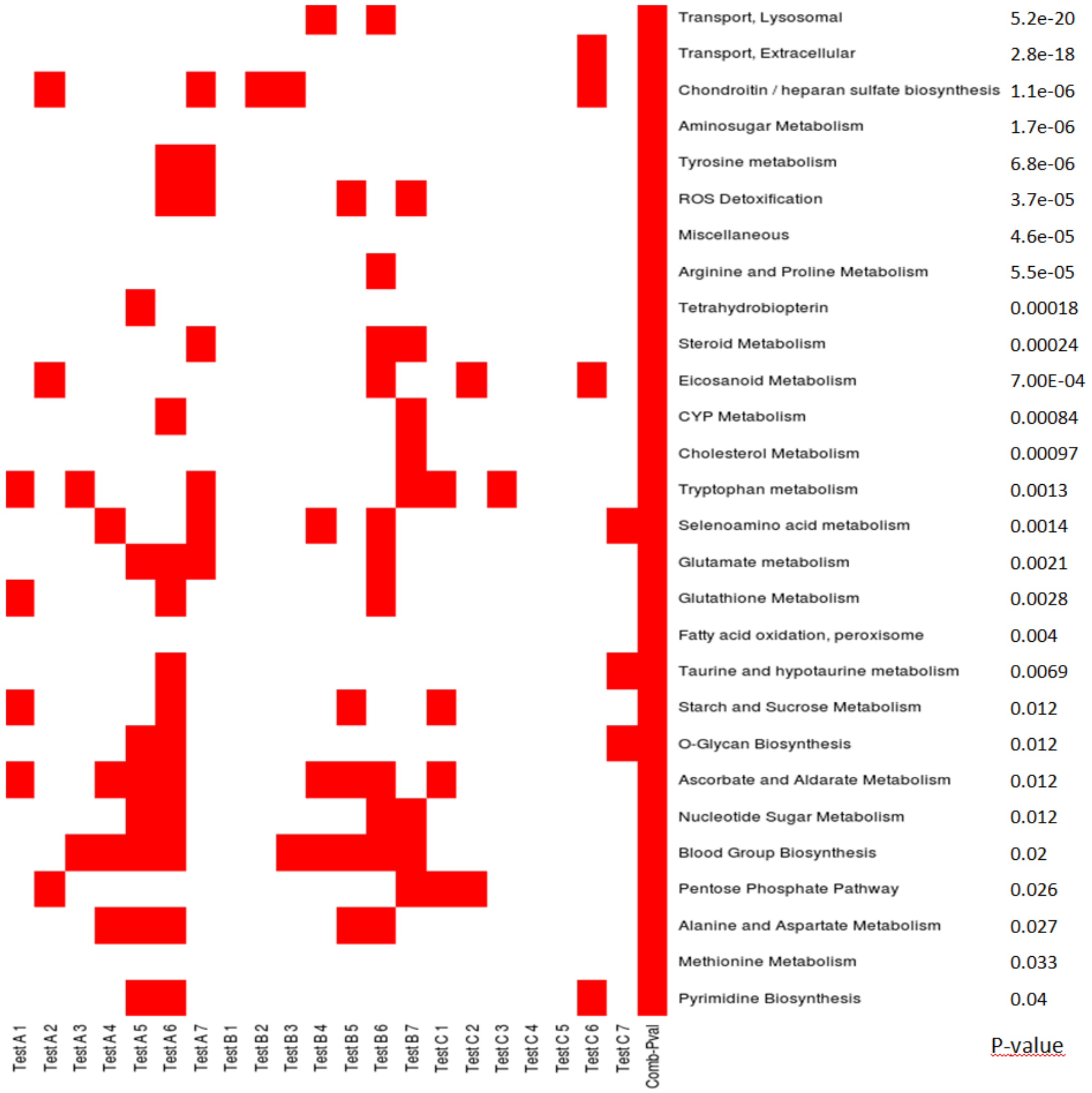


Figure 10. Metabolic pathway analysis results for combined P-value (FDR 0.01) and enrichment in different test cases. Red bars reveal pathways that had Hypergeometric test P-value less than 0.05. Pathways are ranked based on Hypergeometric test P-value scores and the test was done to genes with combined P-value, less than 0.01. P-values were combined from many different tests using the Stouffers method. Tests: siNrf2 -siCTRL untreated (1), siNrf2-siCTRL OA (2), siNrf2-siCTRL OA-NO2 (3), siNrf2-siNrf2 OA-NO2 - OA (4), siCTRL-siCTRL OA-NO2 - OA (5), adNrf2-adCMV 36h (6), adNrf2-adCMV 72h (7) (Table 1). Hypergeometric test was performed also for each test individually: Test A contains enrichment scores for each individual test. For tests B and C genes in individual tests were further divided into upregulated and downregulated genes respectively.

5.3.2 Data integration and network visualization on selected metabolic pathways

Glutathione pathway is enriched in GSEA (Table3.) and in Hypergeometric tests and is predicted to be upregulated by Nrf2 (Figure 10.). Glutathione pathway gene expression levels and Nrf2 dependence was verified using public and our own data. During Nrf2 activation and overexpression *GCLM* is upregulated and Nrf2 knockdown (not complete) decrease the expression (Figure 11.). However, adenoviral overexpression decreases the basal expression in control samples (compared to siRNA samples), likely due to viral transduction. Glutathione pathway gene *GCLM* is also Nrf2 regulated based on our RNA- GRO-seq data (Figure 11). Also *GCLC* was directly regulated by Nrf2 (Nrf2 Chip-seq) and contained ARE motif in the gene regulatory region (Figure 11.). Both of these genes were highly ranked in the list of most significant genes with combined P-values. Therefore *GCLM* shows consistency in various data: These results can be used as a benchmark results for studying other potential Nrf2 regulated pathways and genes.

In the third step, similar approaches as for *GCLM* identification was used in selected pathways to detect directly Nrf2 regulated pathways and to study potential indirectly regulated pathways. Enrichment analyses, barplots on gene expression values and RNA, GRO-seq were used to discover consistently Nrf2 regulated pathways and to study pathway gene regulation. Barplots are available for significant genes (combined P-value 0.01) for Transport Lysosomal, Glutamate metabolism, Ascorbate and Aldarate Metabolism, Cholesterol Metabolism and steroid metabolism pathways (Figure S3, Gene name abbreviations Table S1).

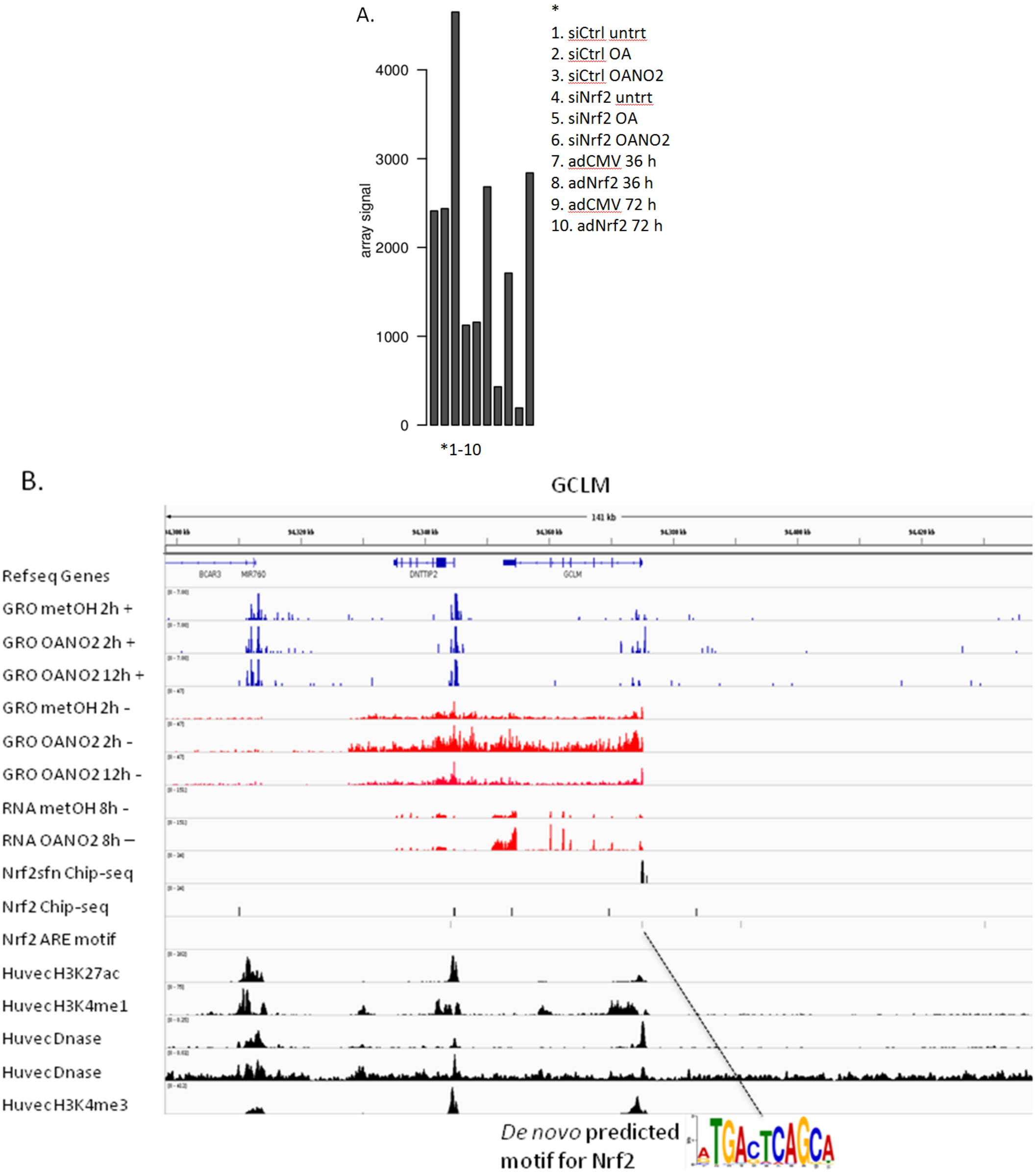


Figure 11. Induction of GCLM by Nrf2. A. Bar plot of GCLM expression values in our microarray data. OA-NO2 upregulates Nrf2 and OA has no effect. siNrf2 decreases the expression of *GCLM* and adNrf2 increases the expression. Basal expression is decreased in adenoviral overexpression B. H3K27ac and H3K4me1 data from ENCODE reveals active regulatory sites for *GCLM* in Human umbilical vein endothelial cells (HUVEC). H3K4me3 marker reveals active transcription start sites. Nrf2 ChIP-seq data available in GEO can reveal sites where Nrf2 is bound. In lymphoblast ChIP-seq data Nrf2 has been activated using sulforaphane (sfn) and is compared to basal control samples (Nrf2 ChIP-seq vs.Ctrl ChIP-seq). In addition, lymphoblast Nrf2 data was used in MEME-ChIP[81] tool to predict Nrf2 binding sites using *De novo* detection of Nrf2 binding motifs. *GCLM* is upregulated in our GRO-seq and RNA-sequencing data, which can be detected by comparing Nrf2 activated nitroOA samples in different time points to control samples (GRO metOH vs. GRO nitroOA and RNA metOH vs. RNA nitroOA). When these different data tracks are visualized in Integrative Genomics Viewer (IGV), Nrf2 binding site homologous motif can be found in the *GCLM* active enhancer site. Furthermore, Nrf2 abundance in the enhancer is increased, as ChIP-seq peaks are larger in activated state than in basal state. In addition Nrf2 activation is shown to upregulate *GCLM*, which can be explained by an increase in Nrf2 abundance in *GCLM* regulatory site. FAIRE and DNAse-seq can provide additional proof for active regulatory sites and TF occupancy on the chromatin.

Lysosomal pathway genes were consistently Nrf2 regulated by activation with OA-NO2 and overexpression, but knockdown did not affect the expression levels, suggesting an indirect effect (Figure S3). Also Nrf2 ARE elements and ChIP-seq peaks were not found in the gene regulatory regions. Glutamate metabolism pathway contained evidence on direct Nrf2 regulation for *GSR* (Figure S2), which is also in top of combined P-value list (Table S5). No other direct targets were identified from visualization tracks for glutamate pathway. Steroid metabolism pathway gene *AKR1C1* was highly induced during Nrf2 overexpression and activation by OA-NO2. *AKR1C1* contained potential ARE element in the gene regulatory region (Figure S2). ChIP-seq peaks were not found, as the gene is not expressed in lymphoblast according to the epigenetic markers. There are eRNA signals showing the increased enhancer activity during activation. Similarly, also ascorbate and aldarate metabolism pathway where highly induced by activation and overexpression, but knockdown effects were low or caused inconsistent regulation compared to control. In addition genes did not contain regulatory elements for Nrf2. *ALDH2* and *ALDH3A2* got high combined P-values and the genes where involved in many different pathways (Table S5). As a summary, gene expression levels where verified to be consistent with enrichment analysis results. Potential direct targets were identified for glutamate and steroid metabolism pathways, but we were not able to exclude possibility of indirect regulation for other selected pathways.

Cholesterol pathway contains in total of 28 genes and their alternative transcripts. Statistically significant genes after in our list of combined P-values ($P<0.01$) included *MSMO1, SOAT1, SQLE, TM7SF2, FDPS, DHCR24, DHCR7,* and *LSS*. To better understand the regulatory effect Nrf2 has on the cholesterol pathway a metabolic network was created using Recon1 reactions and metabolites (Figure 12). Full list of metabolite and gene names and Recon1 reactions can be found in Table S2-S4. Pathway contains the metabolites, enzymes and their expression levels and metabolic reactions that are mainly related to cholesterol synthesis and processing. Interestingly reactions and enzymes in ER (reaction[r]) are downregulated during OA-NO2 activation; including the rate limiting enzymes *DHCR24* and *DHCR7* (Figure 12). Highly similar results were obtained also with RNA/GRO-seq data, when OA-NO2 activated samples were compared to metOH samples (Figure S4). The downregulation could be detected in the primary transcript expression after 2h of Nrf2 activation, suggesting a fast TF (possibly Nrf2) mediated regulation. However, ChIP-seq peaks and promising ARE sites were not found in the gene regulatory regions and therefore we could not exclude indirect regulation. Also Nrf2 overexpression results were not consistent with redox response activation with OA-NO2, as many genes where upregulated by overexpression. Additional proof of indirect regulation was that siNrf2 effect was not clear in most cases, only *FDPS* was upregulated when Nrf2 was silenced, but other genes were downregulated or unchanged by siNrf2 when compared to siCtrl.

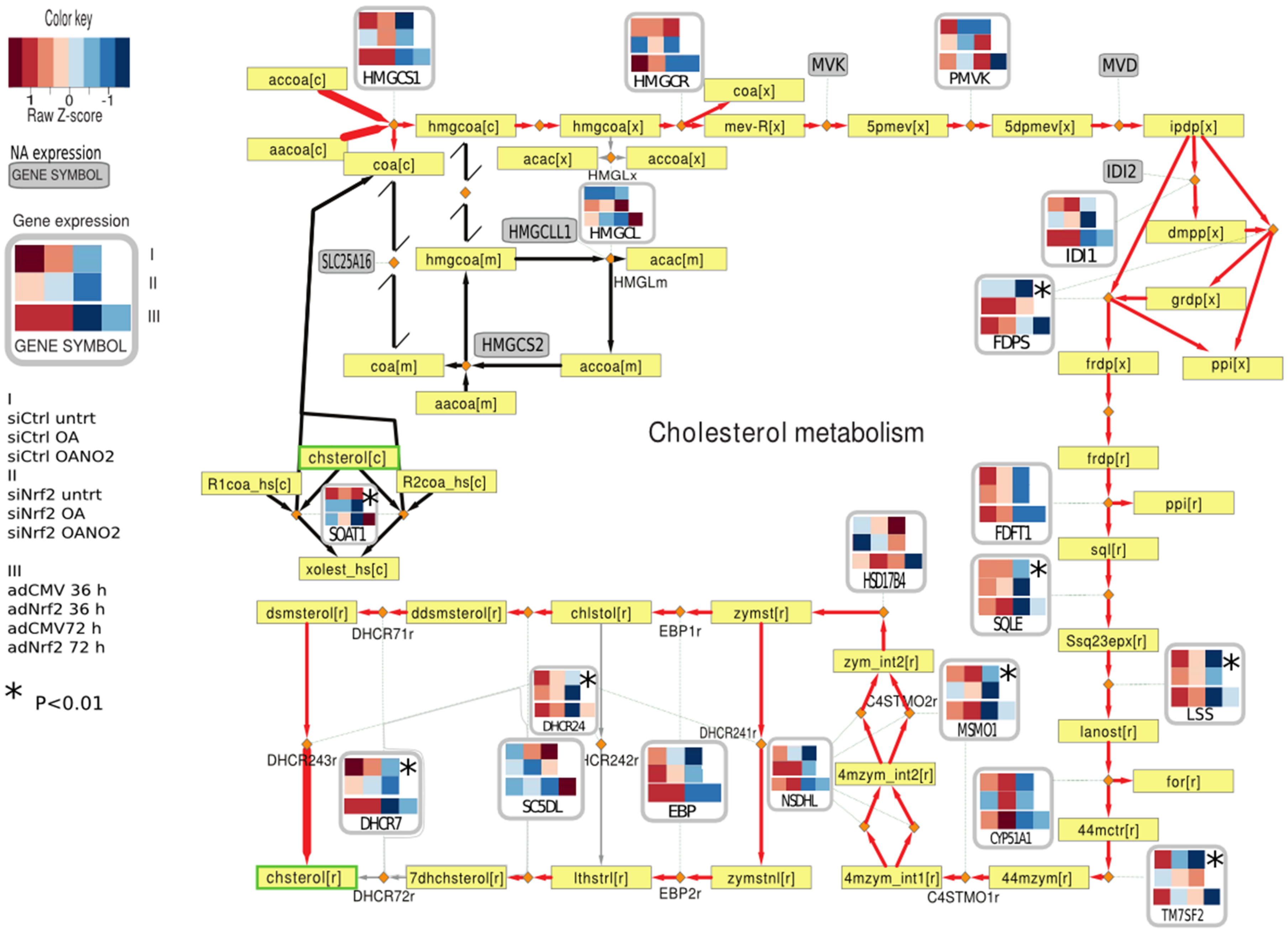

Color key
1
0
-1
Raw Z-score
NA expression
GENE SYMBOL
Gene expression
I
II
III
GENE SYMBOL
I
siCtrl untrt
siCtrl OA
siCtrl OANO2
II
siNrf2 untrt
siNrf2 OA
siNrf2 OANO2
III
adCMV 36 h
adNrf2 36 h
adCMV72 h
adNrf2 72 h
* P<0.01
Cholesterol metabolism
accoa[c]
aacoa[c]
HMGCS1
hmgcoa[c]
coa[c]
hmgcoa[x]
acac[x]
accoa[x]
HMGLx
HMGCR
coa[x]
mev-R[x]
MVK
5pmev[x]
PMVK
5dpmev[x]
MVD
ipdp[x]
IDI2
IDI1
dmpp[x]
grdp[x]
FDPS
frdp[x]
ppi[x]
SLC25A16
hmgcoa[m]
HMGCLL1
HMGCL
acac[m]
HMGLm
HMGCS2
coa[m]
accoa[m]
aacoa[m]
chsterol[c]
R1coa_hs[c]
R2coa_hs[c]
SOAT1
xolest_hs[c]
frdp[r]
FDFT1
ppi[r]
sql[r]
SQLE
Ssq23epx[r]
LSS
lanost[r]
for[r]
CYP51A1
44mctr[r]
TM7SF2
44mzym[r]
C4STMO1r
4mzym_int1[r]
4mzym_int2[r]
C4STMO2r
MSMO1
zym_int2[r]
HSD17B4
NSDHL
zymst[r]
zymstnl[r]
DHCR241r
EBP1r
EBP
EBP2r
chlstol[r]
lthstrl[r]
DHCR242r
DHCR24
SC5DL
ddsmsterol[r]
7dhchsterol[r]
DHCR71r
DHCR7
DHCR72r
dsmsterol[r]
chsterol[r]
DHCR243r

Figure 12. Visualization of the cholesterol pathway and gene expression levels. Cholesterol metabolism network was visualized based on Recon1[80] metabolic network. Yellow rectangular nodes represent metabolites and orange diamond nodes represent reactions. Metabolite[r][m][c][e], r – endoplasmic reticulum, m – mitochondria, c – cytosol, e - extracellular. Enzymes are annotated with the reaction nodes and if the same enzyme catalyzes many reactions, also reaction names are added. Red edges represents cholesterol synthesis pathway and black edges additional cholesterol synthesis related pathways. Thick arrows indicate the start and end of cholesterol synthesis pathway. Adjustments needed to be made for the original network: ACAT2, CYP4F8 and GGPS1related pathways were removed as they were not connected to the main cholesterol synthesis pathway. Also reaction P450SCC1m was removed as no enzymes were annotated with the reaction. Genes with no expression values are marked as grey filled boxes with gene symbols inside. *HMGCS2*, *IDI2*, *CYP4F8* did not have H3K4me3 active TSS markers and are therefore unlikely to be expressed in HUVECs. Expression levels are visualized as heatmaps for enzymes of the pathway in different Nrf2 experiments. Heatmaps I and II are relative to each other and they can be used to evaluate ligand dependence. Untreated and OA treated samples are controls for OA-NO2 activated samples. As Nrf2 knockdown is not complete, also siNrf2 OA-NO2 samples can be compared to OA and untreated samples. Heatmap I and II can be compared to evaluate the effect of Nrf2 knockdown. Heatmap III can be used to evaluate the effect of Nrf2 overexpression (adCMV vs. adNRF2). The full list of metabolite and gene names and Recon1 reactions can be found in Table S2-S4.

6 DISCUSSION

Four novel cancer types, urinary tract carcinoma, glioma, mesothelioma, and thyroid carcinoma and a total of 77 cancer cell lines were discovered by two individual tools SIGNATURE and GSVA to have overactive Nrf2 status with > 95 % probability or empirical FDR of 0.01. Furthermore, characteristic Nrf2 overexpression signature was found in 60 (ca. 10 %) glioblastoma multiforme samples (FDR 0.05) in The Cancer Genome Atlas (TCGA) clinical samples. Cancers with overactive Nrf2 have poor prognosis and therefore it is important to identify cell lines with Nrf2 hyperactivity to develop cancer type selective treatments targeting Nrf2 pathway. Nrf2 inhibitors could be used to treat cancers with Nrf2 hyperactivity and to increase the effectiveness of conventional treatments.

Nrf2 is a novel oncogenic TF and based on our prediction it is frequently overactive in multiple cancers. Environmental factors have a major role in cancer onset. Environmental burden, such as cigarette smoke, UV-light or xenobiotics typically increase Nrf2 expression to protect the cell from damage. Hence Nrf2 should have preventive effects in cancers during early events of cancer development. Interestingly, it appears that the Nrf2 response becomes chronically activated during cancer adaptations and mutations during later phases of cancer development. Suggested benefits of constitutive Nrf2 include proliferative stimulus and control of ROS levels from accumulating at apoptosis inducing levels[61]. As Nrf2 has effects in many pathways, also other oncogenic functions of Nrf2 are likely to be discovered. Nrf2 can potentially alter cancer metabolism and be involved in the formation of metastases. It is likely that constitutive activation of Nrf2 is achieved by various cancer type specific mechanisms that have not been found.

Higher-grade gliomas are extremely malignant tumors and they have poor prognosis. In addition gliomas respond poorly to conventional treatments such as chemotherapy and radiation treatments. Nrf2 could have a substantial role in glioma chemoresistance and malignancy and should therefore be investigated further to verify the Nrf2 activity and the mechanism of how Nrf2 is constantly active to provide better treatments. Nrf2 status in predicted constitutive cell lines can be validated using lentiviral ARE-luciferase reporter constructs[82]. Also negative control glioma cell lines could be selected from cell lines predicted not to have overactive Nrf2 and also A549 is a good positive control for Nrf2 activity, as it has high expression of Nrf2. After Nrf2 activity has been validated, multiple experiments can be done to understand the mechanism of Nrf2 activity. Also drug sensitivity assays can be done to assess the effect of Nrf2 on chemoresistance capabilities. In addition lentiviral ARE-thymidine kinase cancer suicide gene therapy can be utilized, if Nrf2 is confirmed to be constitutively active in gliomas[82]: Increased Nrf2 activity produces high amounts of thymidine kinase in the cancer cell. When a pro-drug ganciclovir will be introduced to the cells, thymidine kinase activity phosphorylates the drug, which induces apoptosis of the cancer cells via signaling cascade and also the nearby cells due to bystander effect[82]. This method could be applied *in vitro* to cell cultures and also *in vivo* to glioma mouse models.

In our results Nrf2 was overactive in many gliomas in CCLE, but not as abundant in TCGA data. One technical reason could be that Nrf2 activity is increased in most glioma samples, and as GSVA provides information about relative expression within samples (see equation 4) a large portion of TCGA samples will get lower score, even if their activity would be increased compared to all cancer samples. A biological reason could be that in CCLE cell lines gliomas are not histologically graded as specifically as in TCGA. Hence CCLE is likely to contain glioma cell lines from various grades. By predicting Nrf2 activity also in lower grade gliomas, it would be possible to study if there is a pattern in Nrf2 activity in different glioma grades. If Nrf2 overactivity is more abundant in higher-grade gliomas, Nrf2 might have a role in the extreme malignancy of higher-grade gliomas. On the other hand if Nrf2 overactivity is more prevalent in lower-grade gliomas, Nrf2 might have a role in glioma transformation to higher-grade. Both preventing and promoting effects are possible, as reduced ROS levels can decrease mutagenesis and also to prevent apoptosis. The SIGNATURE and GSVA results agreed to a large extent but differed in the CCLE melanoma predictions, as it did not pass the SIGNATURE significance threshold but was abundant in the GSVA results. To explore the relevance of these predictions in patient data, TCGA data for melanoma could be used similarly as in gliomas to test if melanomas with overactive Nrf2 could be found in a large portion of patient samples.

TCGA samples with predicted Nrf2 overactivity could be collected for specific cancers such as gliomas and also to many different cancer types. TCGA data could be used in computational analysis to elucidate the mechanism and effect of Nrf2 hyperactivity. Explanations for Nrf2 hyperactivity could be related to Keap1-Nrf2 balance: altered gene-copy numbers, mutations, epigenetic changes in the gene regulatory regions, or differential expression of miRNAs. Such experimental data is available in TCGA and therefore it is a powerful resource for additional analysis.

Supervised learning methods used in Q2 type data classification require informative training sets that can be used to classify the test set data. Our training sets were created based on genes that can best distinct two phenotypes, instead of selecting genes that have highest expression differences between the phenotypes or predefined gene annotations to redox pathway. Therefore Nrf2 signature should be an effective Nrf2 status classifier containing gene-gene correlation structures. In addition the training sets were created using expression data from primary HUVEC cells, which has advantages compared to the use of immortalized cell cultures and cancer cell lines. Cancer is descended from normal cells and therefore results obtained from normal primary cells can be applied to a wide spectrum of cancers, instead of using cancer specific model. HUVEC cell-line pathways are also close to physiological states and therefore resemble Nrf2 activity accurately. The downside of using HUVECs is the tissue specific expression.

GSVA does not offer statistical evaluation of the enrichment score due to the differences between different scoring methods that can be selected. For any of the methods, significance can be estimated by computing an empirical null distribution (such as results obtained using equation 7, Figure S1). On the other hand asymptotic assumption of distribution can be made if the scores are approximately normally distributed (equation 8, Figure S1). Results were compared when normalized enrichment score (following the main GSVA algorithm, equation 8.) significance was estimated using Gaussian distribution assumption and using empirical null distribution in cases where enrichment was calculated using equation 7 and 8 (Figure S1). Results obtained from equation 7 and empirical null distribution found in total of 220 cell lines significant (overlap with the SIGNATURE 80/98) in CCLE and using equation 8 and empirical null distribution found 216 cell lines significant (overlap with the SIGNATURE 77/98) at eFDR 0.01. Asymptotic normal distribution classified 61 cell lines significant (overlap with the SIGNATURE 36/98). We chose equation 8 results with significance estimated using empirical null distribution, as it had high overlap with the significant SIGNATURE results.

Correlation between SIGNATURE and GSVA was good when the cell lines with overactive Nrf2 were statistically significant. However, there was variation in the number of significant results (216 and 98) and the correlation was poor between the tools in the cell lines that were not significant, which is likely resulting from the methodological differences between the tools. One clear difference is that SIGNATURE used genes that were upregulated and downregulated in the scoring and GSVA was performed using only the upregulated genes (80%) of the same Nrf2 signature gene list. This choice was made, as up- or downregulated genes would block the effect of each other during enrichment score calculation. Hence GSVA estimation of enrichment lacks some predictive information that SIGNATURE has. One additional reason could be that GSVA is more sensitive or prone to false positive results, because it tests for competitive null hypothesis as FDR was estimated in GSVA by permuting genes randomly in the gene set. As described in the introduction, genes in the test gene set are more highly correlated compared to random gene sets used to generate the empirical null distribution, P-values will be lower for test gene set, resulting in an increased rate of false positive results. FDR was set to a strict cutoff of 0.01 in GSVA to reduce false positive results. A better estimate of the null distribution in GSVA would have been obtained by random permutation of the samples (used in SIGNATURE) and therefore generating random signatures. However, as SIGNATURE results were used as a benchmark result, it was not essential to obtain the most accurate null distribution and cutoff value.

Nrf2 was identified to be overactivated in many cancers. Therefore additional enrichment analyses were done to understand, which effect Nrf2 could have in the cells at a pathway level. The focus was in metabolic pathways, as metabolic reprogramming is often observed in cancers to support proliferation. Constitutively active Nrf2 could cause metabolic reprogramming, as it has been shown to activate PPP pathway and fatty acid metabolism pathway. This analysis is a Q1 type question (class discovery) to identify Nrf2 regulated pathways and two methods were compared, Hypergeometric test and GSEA.

In order to do metabolic pathway analysis (and create Nrf2 signature) and identify direct Nrf2 targets, experiments needed to be selected specifically for each analysis, as there are downsides in each experiments. In microarray and RNA-seq data, ligand activation of Nrf2 with OA-NO2 is not specific and therefore activation of other TFs or signaling pathways is possible. In addition many RNAs are degraded after 8h time point and many primary effects can be lost in RNA expression data. GRO-sequencing can provide information about primary TF mediated gene regulation, as it can be done in early time points when RNA is not degraded. However, also other TFs can be activated with OA-NO2. Overexpression of Nrf2 does not activate other TFs, but secondary effects are common due to long time points. siRNA knockdown of Nrf2 should alter basal and ligand activation and therefore provide proof of direct Nrf2 regulation. However, siRNAs can have off-target effects. ChIP-seq can be used to associate altered gene expression with increased abundance of TF in the regulatory region. However, antibodies not only enrich specific TFs, but also precipitate impurities and add noise to data. There are many downsides in each experiment and therefore it is important to integrate different data to reason which effect is consistently observed and which are caused by off-target, secondary effects, other TFs, or impurities. Therefore data can be used more efficiently to identify the direct, indirect and off-target effects.

Our results support the previous findings as the PPP and many lipid metabolism pathways were enriched in Hypergeometric test and lipid metabolism related pathway was enriched in GSEA. The PPP was upregulated during overexpression of Nrf2 and downregulated during Nrf2 knockdown (Table 10). Therefore constitutive activation of Nrf2 is likely to cause the PPP activation, which could support the cancer cell proliferation and protection against oxidative stress. Surprisingly, ROS detoxification pathway was not among top enriched pathways in Hypergeometric tests and was missing in GSEA results. Pathway describing gene sets are often constructed by annotating genes to pathways based on literature proof. Many of the key redox pathway genes were not included in the ROS detoxification pathway, but instead assigned to other metabolic processes in the Recon1 annotations. The ROS pathway contained superoxide dismutases that are important enzymes to neutralize ROS, but inadequate alone to describe ROS detoxification pathway. Glutathione metabolism (containing many known Nrf2 target genes, such as GCLM, GCLC) was enriched in Hypergeometric test and in GSEA and therefore seems to be a better indicator of Nrf2 activity in Recon1.

Our metabolic pathway analysis suggests that many metabolic pathways could be Nrf2 regulated. Many pathways can be directly Nrf2 regulated, because the pathways contain enzymes directly involved in detoxification, such as glutamate and glutathione metabolism pathway. However, Nrf2 could regulate also other redox response independent pathways. In our enrichment analysis many amino- and fatty acid and steroid metabolism related pathways were enriched and Nrf2 might affect the transport pathways of the cell. Interestingly cholesterol pathway, lysosomal transport and steroid metabolism pathways are all linked in maintaining cholesterol homeostasis in the cell. In normal conditions cells take in cholesterol packed in low-density lipoprotein (LDL) by endocytosis or synthesize it and use it as a structural component of the cell lipid membranes, and in some tissues cholesterol is a precursor to synthesize steroid hormones. Cholesterol homeostasis is maintained by de novo synthesis of cholesterol, LDL uptake, cholesterol esterification and reverse cholesterol transport[83].

An imbalance between cholesterol influx and efflux is a well-characterized property in atherosclerotic lesion: Elevated cholesterol levels cause accumulation of cholesterol and other lipids to large arteries, which progressively forms fibrous lesions that contain a complex mixture of cells such as macrophage foam cells and smooth muscles cells, oxidized lipids and collagen. Macrophages ingest cholesterol containing oxidized LDL and become foam cells, which is a hallmark of atherosclerotic lesions[84]. Rupture of these lesions and formation of a thrombus causes the clinical complications, such as myocardial infarction and stroke[84]. Foam cells are formed because scavenger receptors collect more cholesterol but cholesterol efflux is low causing systemic cholesterol imbalance[84]. The absence of macrophage Nrf2 has been shown to promote early atherogenesis and therefore Nrf2 activity could have a substantial role in macrophage cholesterol homeostasis[85]. Also endothelial cell dysfunction is one of the initial stages of atherosclerosis and therefore our analysis result fits in this context well. However, cholesterol does not accumulate in the endothelial cells, as they can shut down cholesterol synthesis and increase cholesterol efflux from the cell[83]. Endothelial cholesterol is transported to lysosomes, however cholesterol is not hydrolyzed, but sent to other lipid membranes, mainly to the plasma membrane, but also to Golgi, mitochondria and endoplasmic reticulum and therefore membranes can be storage particles for the cholesterol[86]. In the membranes, cholesterol has many crucial functions, such as regulation of membrane fluidity and formation of lipid rafts that have an important role in signal transduction[87]. For example activation of pro-survival mTOR pathway is dependent on cholesterol trafficking in endothelial cells[87]. In addition sterol accumulation in endothelial cells has been shown to reduce eNOS production, which synthesize nitric oxide that can increase vasodilatation via relaxation of smooth muscle cells[83]. The lack of NO is a prominent feature of endothelial dysfunction.

Cholesterol homeostasis and trafficking can also have a high significance in cancer. Interestingly Nrf2 hyperactivity in cancer and the activation of mTOR pathway was linked, suggested mechanism included increased expression of RagD, which is an activator of mTOR, but was not a direct target of Nrf2[88]. One possible reason is that Nrf2 overactivity upregulates cholesterol pathway causing increased cholesterol trafficking, lipid raft formation and increased signaling, which could activate the mTOR pathway. Similarly Nrf2 could have an effect also on activating other signaling pathways that can promote survival and proliferation via upregulation of cholesterol pathway.

Direct targets of Nrf2 could not be found in cholesterol pathway, suggesting an indirect regulation. It is much more challenging to characterize how Nrf2 regulates cholesterol pathway, as direct Nrf2 targets could not be detected. Interestingly cholesterol pathway was consistently downregulated during early time points (OA-NO2 activation), but upregulated in late time points with overexpressed Nrf2. Nrf2 activator ligands are not specific and therefore activation of other TFs or signaling pathways is probable reason for a quick response to OA-NO2 treatment. It is also possible that Nrf2 mediates cholesterol gene regulation via protein-protein interactions, which are not well characterized for Nrf2. There are many different factors that could have a role in this system, including cholesterol pathway and LDL-receptor mediator SREBP-2[83] and SREBP-2 inhibitors INSIG1 and INSIG2. As noted in the introduction, also Nrf2-RXRa may form a regulatory loop in lipid metabolic reactions and imbalance in this system might have an effect especially in the secondary response during Nrf2 overactivation. LXR is a partner of RXRa and is thought to mediate cholesterol efflux[83]. Downregulation of RXRa via constitutively active Nrf2 could prevent cholesterol efflux and therefore increase cholesterol trafficking, lipid raft formation and signaling. These changes could have high impact on cellular functions and promote diseases, such as atherosclerosis and cancer.

Cholesterol enzymes in the ER were downregulated during OA-NO2 activation (Figure 12.). Before cholesterol synthesis pathway is entering the ER, FDPS converts Dimethylallyl diphosphate (dmpp[x]) or Geranyl diphosphate (grdp[x]) to Farnesyl-diphosphate (frdp[r], Figure 12). Interestingly *FDPS* was consistently downregulated by Nrf2 in all experiments and also the Nrf2 knockdown upregulated FDPS. However, direct binding sites were not identified. FDPS forms a branch from FPP to isoprenoids (Farnesyl, Dolichol, Ubiquinone) pathways[89]. FPP also serves as a substrate for protein farnesylation and geranylgeranylation[89]. These metabolites can participate in multiple cellular processes including cell growth, differentiation, and vesicle trafficking and have central role in many cancers (such as prostate cancer[90]) and are likely to be involved in atherosclerosis[89]. Prenylation of mutated RAS is needed in many cancers to transform tumor malignant and many inhibitors have been developed with good *in vitro* results but lack of clinical success[91]. KRAS has been shown to elevate the basal Nrf2 levels in oncogene primary murine cells[92]. Therefore Nrf2 mediated cholesterol pathway regulation might have a link to prenylation of key disease promoting proteins.

The metabolic pathway analysis revealed several common caveats encountered in selecting a proper enrichment method. For the hypergeometric test a cutoff for P-values has to be selected. There is no general way of choosing the cutoff and some pathways with low but meaningful expression differences might be lost, as multiple testing corrections might be too conservative. To address this, we used the approach to combine evidence from multiple experiments by combining P-values from comparisons where the same null hypothesis was tested. In total of 472 genes with combined P-values were significant and many metabolic pathways were enriched. Many known Nrf2 target genes were highly ranked and got low P-value, suggesting that the combined P-value method was successfully combining proof from multiple Nrf2 perturbations (Table S4). According to Hypergeometric test, many metabolic pathways were enriched. Although straightforward to use, the hypergeometric test assumes gene independence, which is not true in biology. Metabolic pathways are likely to have a high number of correlations between the genes and the pathways. In addition hypergeometric test does not weight highly ranked genes and therefore many pathways with direct Nrf2 targets with high P-values can have low enrichment scores.

GSEA takes the gene-gene correlation structures into account when sample permutation is used for significance assessment and also weights the high ranked genes. GSEA or similar methods have been recommended for the Q1 type enrichment analysis by many method comparing articles and frameworks for using GSEA type methods has been published[10,17,93]. In our analysis GSEA seemed to lack sensitivity and few metabolic pathways were estimated significant. Similarly GSEA has been criticized to be overly complicated, heavy and insensitive[30]. GSEA user guide recommend high FDR threshold of 0.25 as the most suitable, which suggests a lack of sensitivity. GSEA is also outperformed in many articles and other methods have been suggested instead of GSEA[93]. GSEA can have low power because GSEA and some of its variants are really hybrids of the competitive and self-contained methods and they do not really test the competitive null hypothesis, because a gene vs. background gene model is used in the weighted Kolmogorov-Smirnov test, but shuffling of samples is used for the significance estimation[15]. In addition Small sample sizes and lack of replicates limit the use of the GSEA significance estimation and it was challenging to construct two phenotypes from our data with enough samples (N=3 in most experimental conditions). Gene randomization parameter was added to the tool to overcome the small sample size problem, but gene randomization loses the gene-gene correlation structures and therefore increases the detection of false positive gene sets[10,16] and in this respect does not constitute an improvement over the hypergeometric test.

When interpreting the biological significance of the enrichment score it should consider that GSEA can have two types of results that are not distinguishable: gene sets with few highly differentially expressed genes or many genes with small changes in expression[10]. The weighted running sum increases the enrichment score for high ranked genes, which could underline why GSEA detected the HMOX1 containing heme degradation pathway enriched (total number of unique genes 4), as the pathway contained only three other genes that were not differentially expressed and therefore not associated with either phenotypes. HMOX1 is a known target of Nrf2 and it is highly induced during Nrf2 activation, which increases the pathway enrichment score. Another major problem is that GSEA is not powerful to detect gene sets that have genes positively and negatively associated with the phenotype[94], so the pathways with both up- and downregulated genes will get decreased scores. GSEA failed to detect cholesterol pathway, likely due to this effect, as Nrf2 activation by OA-NO2 and overexpression affected the gene expression levels inconsistently. This behavior and lack of sensitivity of GSEA might also be favorable for some users, as it can reduce the amount of false positive results and indirectly regulated pathways therefore making the result interpretation more straightforward.

An alternative approach on metabolic pathway analysis would be to include network structure of Recon1 metabolic pathways in the model and to set up constraints based on this information to gain better understanding of the metabolic (and other) consequences of Nrf2 overactivity. As a follow-up computational work, systems biology models, such as CBM[95,96], could be used to infer metabolic activities from expression levels to understand the effects of altered expression levels in context of the other active metabolic reactions. This might be useful in understanding the effect of Nrf2 for the pathway dynamics of and could lead to the identification of new disease promoting steady states of the metabolic network. Many microarray experiments on Nrf2 perturbations are publicly available and these experiments could be collected for additional information about Nrf2 regulated pathways to provide additional proof for CBM. One clear problem will then be the integration of the data. The data should be comparable so data normalization procedures are important. Even after successful normalization the data might not be comparable, especially when the data is produced by different platforms or technologies or (depending on the model assumptions) when data is from different cells or experimental designs[97]. Therefore data only from HUVECs and the Affymetrix platform should be used as additional data. Data reliability and reproducibility presents a problem, because data is obtained from other laboratories[97]. These problems in data integration has been recognized and therefore databases, such as ENCODE and TCGA have started to standardize sample collection and data analysis and validate available data to make it more reliable and reproducible. Other challenges are that mathematical modeling is relatively new tool in biology; so all pitfalls have not been mapped. The amount of data is huge, it is possible to find biologically irrelevant correlations using mathematical models and therefore make false conclusions. Therefore the model must be designed well to provide biologically relevant data for reliable results.

For building a systems biology model of Nrf2-dependent gene regulation, methods from the top down systems biology field could be used to create accurate lists of the Nrf2 regulatory network components, using data integration and combining P-values of many genome-wide experiments, in particular by utilizing the power of several next generation sequencing methods to capture the different steps of gene regulation. This list would ideally be divided into direct and indirect targets, that could initially be based on time points of activation and motif analysis tools, as direct targets get typically regulated faster and contain ARE homologous elements. This distinction would be relevant for the model because their activation can be associated directly with Nrf2 activation without considering the presence of additional factors. In constitutive Nrf2 activation regulatory network steady state has changed, which changes the structure of the regulatory network. Measurements of mRNA levels (using RNA-seq and microarray data) over a time series could be used to as means to distinguish indirect targets that are expected to appear only in later time points and in combination with GRO-seq data that provides information about ongoing transcription, the early events of transcription could be confirmed without the need to correct for different mRNA stabilities. GRO-seq can also be used in discovering and quantifying enhancer sites that are activated in an Nrf2-dependent manner based on eRNA expression. Initially, DNAse or FAIRE-seq data could be used in discovering the enhancer locations. DNAse-seq on Nrf2 activation samples would reveal additional information about Nrf2 binding sites, because increased Nrf2 binding would provide higher signal compared to control. Deep-sequencing could provide enough information about the specific binding site and motif predictions in these sites could be used to detect putative Nrf2 binding sites and also other TFs that are regulated by activation of the Electrophile Response. Combination of GRO-seq and DNAse-seq therefore represents an efficient means to identify the enhancers responsible for observed expression changes, first by screening which enhancers are activated from GRO-seq and then checking if DNAse-seq signal is increased in the region and if the region contains ARE elements or binding sites for other TFs.

In summary, the need of sensitivity, choice of null hypothesis, amount of samples and the questions to be answered guide the enrichment tool selection. Also user friendly tools that are easily accessible are important features for selecting the enrichment tool. Especially supervised learning methods have not been available for common users, because there has not been software available. Nowadays also supervised learning methods, such as the SIGNATURE tool are available for common users and they can be used to answer increasingly popular Q2 type questions, where GSEA type methods are not appropriate. When choosing an enrichment method, those using gene randomization should be avoided in general, as they ignore the gene-gene correlation structure[15]. However, as encountered here, sample randomization is not always possible due to complex phenotypes (for analysis across all cancer cell lines) or the lack of replicates (N>7 is not typical for most common experimental setups). To address this issue, methods that allow gene-gene correlations to be estimated have been proposed[98,99]. Competitive methods with gene randomization produce easily interpretable and sensitive P-values, so solutions for gene-gene correlation problem has been developed in CAMERA[99]. Alternatively, analytical background distributions can be used when data is confirmed to correspond to the assumed distribution (as in GSVA). Also it could be informative to include multiple independent enrichment tools to prevent method dependent biases and to compare overlap between the predictions[17]. Next generation sequencing is increasingly popular and therefore many methods need to be updated to make the enrichment analysis possible. Typical procedure is to convert count data to log2-count-per-million values (limma, voom() function) or RPKM values (edgeR) and use tools such as GSVA, ROAST and CAMERA. In addition to tool development, the pathway databases should be improved to provide more accurate pathway constructions with cell line specific pathway information for different conditions. Pathway genes are not always annotated correctly and in disease pathway annotations about role of the pathway might be highly context dependent. Benchmark data sets derived from biological data could be constructed to support testing of the enrichment methods.

7 CONCLUSION

Genome-wide HUVEC datasets from perturbations that affect Nrf2 activity were used together with public data (CCLE, TCGA, Recon1) in different contexts: first to address whether Nrf2 is active in cancers and subsequently to investigate how Nrf2 may affect metabolic pathways. These questions were answered by using enrichment tools to obtain an unbiased overview of what should be investigated further. The results suggest that a constitutive Nrf2 signature can be found in multiple cancers. There were also four novel cell lines discovered by two independent enrichment tools. Several conclusions from this study should be further investigated. The novel cancer cell lines, especially the ones that were detected by two computational analyses, are good candidates to validate further using experimental methods.

The metabolic pathway analyses predicted that Nrf2 regulates many fatty and amino acid metabolism pathways, as well as the pentose phosphate pathway. Interestingly, cholesterol, lysosomal transport and steroid metabolism pathways were all enriched; as these pathways are highly connected in cholesterol homeostasis, they could play a significant role in many diseases, including atherosclerosis and cancer. The link of Nrf2 to the cholesterol pathway has not been previously reported. We were not able to exclude indirect regulation and therefore further studies must be conducted to understand the role of Nrf2 in cholesterol homeostasis.

Nrf2 is known to have a central role in maintaining the cell homeostasis via the stress response pathway. However, Nrf2 might also have a crucial role in pathogenesis of cancer and other diseases as it has numerous target genes and it is involved in many disease-promoting pathways. This study aids in characterizing Nrf2 activity in cancers and provides insight in how Nrf2 activity can regulate the cellular functions. This information is needed to develop cancer selective inhibitors targeting Nrf2 pathway and to understand Nrf2 function in disease.

8 REFERENCES

1. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. & Ebert, B. L. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. (2005).

2. Chang, J. T. *et al.* SIGNATURE: a workbench for gene expression signature analysis. *BMC bioinformatics* **12,** 443 (2011).

3. Bruggeman, F. J. & Westerhoff, H. V. The nature of systems biology. *Trends in microbiology* **15,** 45–50 (2007).

4. Clayton, A. L., Hazzalin, C. a & Mahadevan, L. C. Enhanced histone acetylation and transcription: a dynamic perspective. *Molecular cell* **23,** 289–96 (2006).

5. Material, S. O., Web, S., Press, H., York, N. & Nw, A. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)* **306,** 636–40 (2004).

6. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic acids research* **39,** D1005–10 (2011).

7. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–7 (2012).

8. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. & Ebert, B. L. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. (2005).

9. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics (Oxford, England)* **21,** 171–8 (2005).

10. Nam, D. & Kim, S.-Y. Gene-set approach for expression pattern analysis. *Briefings in bioinformatics* **9,** 189–97 (2008).

11. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)* **27,** 1739–40 (2011).

12. Culhane, A. C. *et al.* GeneSigDB--a curated database of gene expression signatures. *Nucleic acids research* **38,** D716–25 (2010).

13. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30 (2000).

14. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39,** D691–7 (2011).

15. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)* **23,** 980–7 (2007).

16. Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews. Genetics* **7,** 55–65 (2006).

17. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics* (2013). doi:10.1093/bib/bbt002

18. Barry, W. T., Nobel, A. B. & Wright, F. a. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics (Oxford, England)* **21,** 1943–9 (2005).

19. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *The Annals of Applied Statistics* **1,** 107–129 (2007).

20. Edelman, E. *et al.* Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics (Oxford, England)* **22,** e108–16 (2006).

21. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z. & DeLisi, C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics* **13,** 281–91 (2012).

22. Goeman, J. J., van de Geer, S. a., de Kort, F. & van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20,** 93–99 (2003).

23. Tomfohr, J., Lu, J. & Kepler, T. B. Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics* **6,** 225 (2005).

24. Mansmann, U. & Meister, R. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods of information in medicine* **44,** 449–53 (2005).

25. Fridley, B. L., Jenkins, G. D. & Biernacka, J. M. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PloS one* **5,** (2010).

26. Liu, Q., Dinu, I., Adewale, A. J., Potter, J. D. & Yasui, Y. Comparative evaluation of gene-set analysis methods. *BMC bioinformatics* **8,** 431 (2007).

27. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)* **23,** 980–7 (2007).

28. Mansmann, U. & Meister, R. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods of information in medicine* **44,** 449–53 (2005).

29. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC bioinformatics* **14,** 7 (2013).

30. Irizarry, R. a, Wang, C., Zhou, Y. & Speed, T. P. Gene set enrichment analysis made simple. *Statistical methods in medical research* **18,** 565–75 (2009).

31. Tamayo, P., Steinhardt, G., Liberzon, A. & Mesirov, J. P. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical methods in medical research* (2012). doi:10.1177/0962280212460441

32. Kim, S.-Y. & Volsky, D. J. PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics* **6,** 144 (2005).

33. West, M. *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 11462–7 (2001).

34. Liu, Z. *et al.* Singular value decomposition-based regression identifies activation of endogenous signaling pathways in vivo. *Genome biology* **9,** R180 (2008).

35. Furge, K. a *et al.* Detection of DNA copy number changes and oncogenic signaling abnormalities from gene expression data reveals MYC activation in high-grade papillary renal cell carcinoma. *Cancer research* **67,** 3171–6 (2007).

36. Ringnér, M., Peterson, C. & Khan, J. Analyzing array data using supervised methods. *Pharmacogenomics* **3,** 403–15 (2002).

37. Larranaga, P. Machine learning in bioinformatics. *Briefings in Bioinformatics* **7,** 86–112 (2006).

38. Wilkinson, D. J. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics* **8,** 109–16 (2007).

39. Schopfer, F. J., Cipollina, C. & Freeman, B. a. Formation and signaling actions of electrophilic lipids. *Chemical reviews* **111,** 5997–6021 (2011).

40. Okawa, H. *et al.* Hepatocyte-specific deletion of the keap1 gene activates Nrf2 and confers potent resistance against acute drug toxicity. *Biochemical and biophysical research communications* **339,** 79–88 (2006).

41. Motohashi, H. & Yamamoto, M. Nrf2-Keap1 defines a physiologically important stress response mechanism. *Trends in molecular medicine* **10,** 549–57 (2004).

42. Kansanen, E. *et al.* Electrophilic nitro-fatty acids activate NRF2 by a KEAP1 cysteine 151-independent mechanism. *The Journal of biological chemistry* **286,** 14019–27 (2011).

43. Motohashi, H., O'Connor, T. & Katsuoka, F. Integration and diversity of the regulatory network composed of Maf and CNC families of transcription factors. *Gene* **294,** 1–12 (2002).

44. Wakabayashi, N. *et al.* Keap1-null mutation leads to postnatal lethality due to constitutive Nrf2 activation. *Nature genetics* **35,** 238–45 (2003).

45. Kansanen, E., Jyrkkänen, H.-K. & Levonen, A.-L. Activation of stress signaling pathways by electrophilic oxidized and nitrated lipids. *Free radical biology & medicine* **52,** 973–82 (2012).

46. Vomhof-Dekrey, E. E. & Picklo, M. J. The Nrf2-antioxidant response element pathway: a target for regulating energy metabolism. *The Journal of nutritional biochemistry* **23,** 1201–6 (2012).

47. Malhotra, D. *et al.* Global mapping of binding sites for Nrf2 identifies novel targets in cell survival response through ChIP-Seq profiling and network analysis. *Nucleic acids research* **38,** 5718–34 (2010).

48. Chorley, B. N. *et al.* Identification of novel NRF2-regulated genes by ChIP-Seq: influence on retinoid X receptor alpha. *Nucleic acids research* **40,** 7416–29 (2012).

49. Hirotsu, Y. *et al.* Nrf2-MafG heterodimers contribute globally to antioxidant and metabolic networks. *Nucleic acids research* 1–12 (2012). doi:10.1093/nar/gks827

50. Papp, D. *et al.* The NRF2-related interactome and regulome contain multifunctional proteins and fine-tuned autoregulatory loops. *FEBS letters* **586,** 1795–802 (2012).

51. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)* **324,** 1029–33 (2009).

52. Cairns, R. a, Harris, I. S. & Mak, T. W. Regulation of cancer cell metabolism. *Nature reviews. Cancer* **11,** 85–95 (2011).

53. Grant, C. M. Metabolic reconfiguration is a regulated response to oxidative stress. *Journal of biology* **7,** 1 (2008).

54. Giannoni, E., Buricchi, F., Raugei, G., Ramponi, G. & Chiarugi, P. Intracellular Reactive Oxygen Species Activate Src Tyrosine Kinase during Cell Adhesion and Anchorage-Dependent Cell Growth Intracellular Reactive Oxygen Species Activate Src Tyrosine Kinase during Cell Adhesion and Anchorage-Dependent Cell Growth †. *Society* (2005). doi:10.1128/MCB.25.15.6391

55. Giannoni, E., Buricchi, F., Raugei, G., Ramponi, G. & Chiarugi, P. Intracellular Reactive Oxygen Species Activate Src Tyrosine Kinase during Cell Adhesion and Anchorage-Dependent Cell Growth Intracellular Reactive Oxygen Species Activate Src Tyrosine Kinase during Cell Adhesion and Anchorage-Dependent Cell Growth †. *Society* (2005). doi:10.1128/MCB.25.15.6391

56. Rushworth, S. a *et al.* The high Nrf2 expression in human acute myeloid leukemia is driven by NF-κB and underlies its chemo-resistance. *Blood* **120,** 5188–98 (2012).

57. Taguchi, K., Motohashi, H. & Yamamoto, M. Molecular mechanisms of the Keap1–Nrf2 pathway in stress response and cancer evolution. *Genes to cells : devoted to molecular & cellular mechanisms* **16,** 123–40 (2011).

58. Wang, X.-J. *et al.* Nrf2 enhances resistance of cancer cells to chemotherapeutic drugs, the dark side of Nrf2. *Carcinogenesis* **29,** 1235–43 (2008).

59. Mitsuishi, Y. *et al.* Nrf2 redirects glucose and glutamine into anabolic pathways in metabolic reprogramming. *Cancer cell* **22,** 66–79 (2012).

60. Kansanen, E., Kuosmanen, S. M., Leinonen, H. & Levonen, A.-L. The Keap1-Nrf2 pathway: Mechanisms of activation and dysregulation incancer. *Redox Biology* **1,** 45–49 (2013).

61. Sporn, M. B. & Liby, K. T. NRF2 and cancer: the good, the bad and the importance of context. *Nature reviews. Cancer* **12,** 564–71 (2012).

62. Hayes, J. D., Mcmahon, M., Chowdhry, S. & Dinkova-kostova, A. T. Through the Keap1 – Nrf2 Pathway. **13,** (2010).

63. Hanada, N. *et al.* Methylation of the KEAP1 gene promoter region in human colorectal cancer. *BMC cancer* **12,** 66 (2012).

64. Adam, J. *et al.* Renal cyst formation in Fh1-deficient mice is independent of the Hif/Phd pathway: roles for fumarate in KEAP1 succination and Nrf2 signaling. *Cancer cell* **20,** 524–37 (2011).

65. Ma, Q. & He, X. Molecular basis of electrophilic and oxidative defense: promises and perils of Nrf2. *Pharmacological reviews* **64,** 1055–81 (2012).

66. Tanaka, Y. *et al.* NF-E2-Related Factor 2 Inhibits Lipid Accumulation and Oxidative Stress in Mice Fed a High-Fat Diet □. *Pharmacology* (2008). doi:10.1124/jpet.107.135822.

67. Okada, K. *et al.* Deletion of Nrf2 leads to rapid progression of steatohepatitis in mice fed atherogenic plus high-fat diet. *Journal of gastroenterology* (2012). doi:10.1007/s00535-012-0659-z

68. Kitteringham, N. R. *et al.* Proteomic analysis of Nrf2 deficient transgenic mice reveals cellular defence and lipid metabolism as primary Nrf2-dependent pathways in the liver. *Journal of proteomics* **73,** 1612–31 (2010).

69. Xue, P. *et al.* Adipose Deficiency of Nrf2 in ob/ob Mice Results in Severe Metabolic Syndrome. *Diabetes* **62,** 845–54 (2013).

70. Shin, S. *et al.* NRF2 modulates aryl hydrocarbon receptor signaling: influence on adipogenesis. *Molecular and cellular biology* **27,** 7188–97 (2007).

71. Wang, H. *et al.* RXRα Inhibits the NRF2-ARE Signalling Pathway Through A Direct Interaction With the Neh7 domain of NRF2. *Cancer research* (2013). doi:10.1158/0008-5472.CAN-12-3386

72. Kansanen, E. *et al.* Nrf2-dependent and -independent responses to nitro-fatty acids in human endothelial cells: identification of heat shock response as the major pathway

activated by nitro-oleic acid. *The Journal of biological chemistry* **284,** 33233–41 (2009).

73. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474,** 390–4 (2011).

74. Nagalakshmi, U. *et al.* The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320 ,** 1344–1349 (2008).

75. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)* **316,** 1497–502 (2007).

76. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. 1757–1767 (2011). doi:10.1101/gr.121541.111.Freely

77. Spang, R. *et al.* Prediction and uncertainty in the analysis of gene expression profiles. *In silico biology* **2,** 369–81 (2002).

78. Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature genetics* **34,** 226–30 (2003).

79. Albert, J. & Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical …* **88,** 669–679 (1993).

80. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 1777–82 (2007).

81. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics (Oxford, England)* **27,** 1696–7 (2011).

82. Leinonen, H. M. *et al.* Oxidative stress-regulated lentiviral TK/GCV gene therapy for lung cancer treatment. *Cancer research* **72,** 6227–35 (2012).

83. Hassan, H. H., Denis, M., Krimbou, L., Marcil, M. & Genest, J. Cellular cholesterol homeostasis in vascular endothelial cells. *The Canadian journal of cardiology* **22 Suppl B,** 35B–40B (2006).

84. Glass, C. K. & Witztum, J. L. Atherosclerosis : The Road Ahead Review. **104,** 503–516 (2001).

85. Ruotsalainen, A.-K. *et al.* The absence of macrophage Nrf2 promotes early atherogenesis. *Cardiovascular Research* (2013). doi:10.1093/cvr/cvt008

86. Lange, Y., Ye, J. & Steck, T. L. Circulation of cholesterol between lysosomes and the plasma membrane. *The Journal of biological chemistry* **273,** 18915–22 (1998).

87. Xu, J., Dang, Y., Ren, Y. R. & Liu, J. O. Cholesterol trafficking is required for mTOR activation in endothelial cells. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 4764–9 (2010).

88. Shibata, T. *et al.* Global downstream pathway analysis reveals a dependence of oncogenic NF-E2-related factor 2 mutation on the mTOR growth signaling pathway. *Cancer research* **70,** 9095–105 (2010).

89. McTaggart, S. J. Isoprenylated proteins. *Cellular and molecular life sciences : CMLS* **63,** 255–67 (2006).

90. Jiang, F. *et al.* Farnesyl diphosphate synthase is abundantly expressed and regulated by androgen in rat prostatic epithelial cells. *The Journal of steroid biochemistry and molecular biology* **78,** 123–30 (2001).

91. Berndt, N., Hamilton, A. D. & Sebti, S. M. Targeting protein prenylation for cancer therapy. *Nat Rev Cancer* **11,** 775–791 (2011).

92. DeNicola, G. M. *et al.* Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature* **475,** 106–9 (2011).

93. Ackermann, M. & Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC bioinformatics* **10,** 47 (2009).

94. Dinu, I. *et al.* Improving gene set analysis of microarray data by SAM-GS. *BMC bioinformatics* **8,** 242 (2007).

95. Rossell, S., Huynen, M. a. & Notebaart, R. a. Inferring Metabolic States in Uncharacterized Environments Using Gene-Expression Measurements. *PLoS Computational Biology* **9,** e1002988 (2013).

96. Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø. & Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology* **26,** 1003–10 (2008).

97. Joyce, A. R. & Palsson, B. Ø. The model organism as a system: integrating “omics” data sets. *Nature reviews. Molecular cell biology* **7,** 198–210 (2006).

98. Wu, D. *et al.* ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics (Oxford, England)* **26,** 2176–82 (2010).

99. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research* **40,** e133 (2012).

9 SUPPLEMENTARY TABLES AND FIGURES

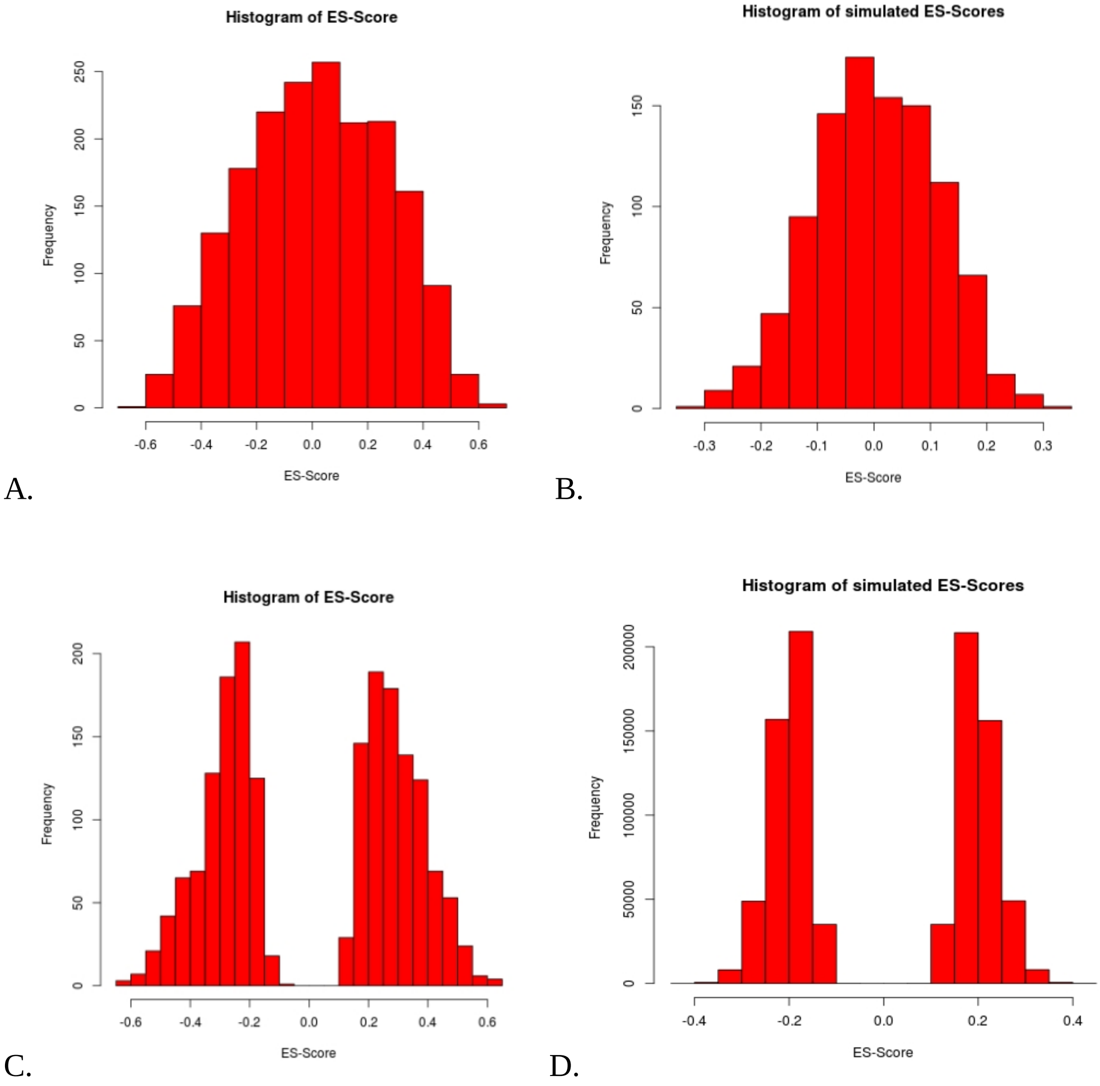


Figure S1. Distribution of Enrichment Scores (ES). A. Histogram of ES when ES is computed using difference between largest positive and negative deviations. B. Histogram of ES when ES is simulated by permuting gene labels (difference between largest positive and negative deviations). C. Histogram of ES when ES is computed using maximum deviation from zero method. B. Histogram of ES when ES is simulated by permuting gene labels (maximum deviation from zero).

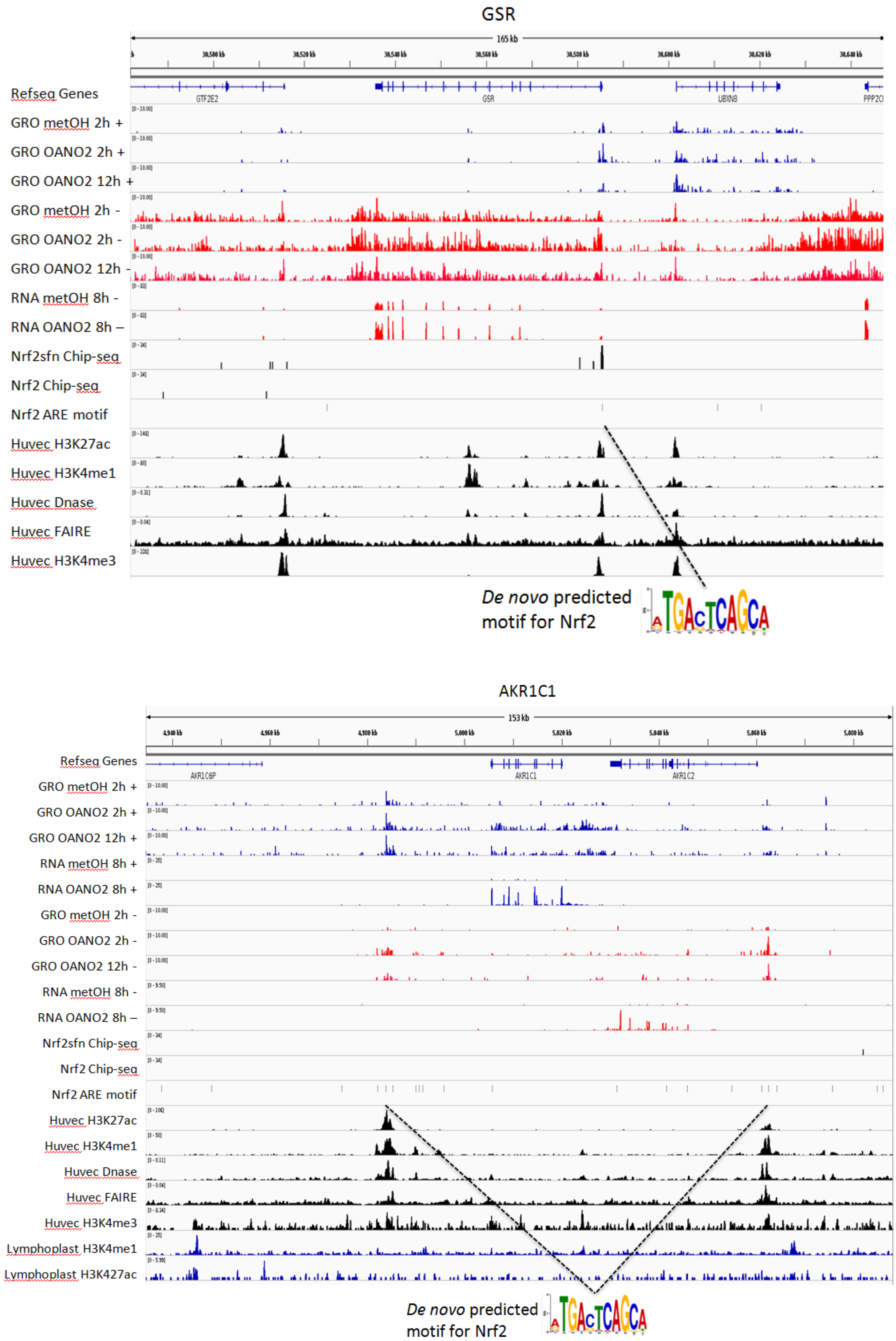


Figure S2. Potential direct targets of Nrf2

Table S1. Gene abbreviations for Glutamate metabolism, Steroid Metabolism, Transport, Lysosomal, Ascorbate and Aldarate Metabolism genes

GSR	Homo sapiens glutathione reductase (GSR), transcript variant 1, mRNA
AKR1C1	Homo sapiens aldo-keto reductase family 1, member C1 (dihydrodiol dehydrogenase 1; 20-alpha (3-alpha)-hydroxysteroid dehydrogenase) (AKR1C1), mRNA
ATP6V1B2	Homo sapiens ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2 (ATP6V1B2), mRNA
GLRX	Homo sapiens glutaredoxin (thioltransferase) (GLRX), transcript variant 3, mRNA
ATP6V1D	Homo sapiens ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D (ATP6V1D), mRNA
SULT1E1	Homo sapiens sulfotransferase family 1E, estrogen-preferring, member 1 (SULT1E1), mRNA
ATP6V1H	Homo sapiens ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H (ATP6V1H), transcript variant 1, mRNA
ATP6V0B	Homo sapiens ATPase, H+ transporting, lysosomal 21kDa, V0 subunit b (ATP6V0B), transcript variant 2, mRNA
ATP6V1C1	Homo sapiens ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1 (ATP6V1C1), mRNA
GLS	Homo sapiens glutaminase (GLS), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA
ATP6V1E1	Homo sapiens ATPase, H+ transporting, lysosomal 31kDa, V1 subunit E1 (ATP6V1E1), transcript variant 2, mRNA
ATP6V0D1	Homo sapiens ATPase, H+ transporting, lysosomal 38kDa, V0 subunit d1 (ATP6V0D1), mRNA
GPT2	Homo sapiens glutamic pyruvate transaminase (alanine aminotransferase) 2 (GPT2), transcript variant 2, mRNA
SLC29A3	Homo sapiens solute carrier family 29 (nucleoside transporters), member 3 (SLC29A3), transcript variant 2, mRNA

Table S2. Recon1 cholesterol pathway metabolite abbreviations

metabolite ID	**metabolite name**
44mzym[r]	4,4-dimethylzymosterol
4mzym_int1[r]	4-Methylzymosterol intermediate 1
4mzym_int2[r]	4-Methylzymosterol intermediate 2
5dpmev[x]	(R)-5-Diphosphomevalonate
5pmev[x]	(R)-5-Phosphomevalonate
7dhchsterol[r]	7-Dehydrocholesterol
R1coa_hs[c]	R group 1 Coenzyme A homo sapiens
R2coa_hs[c]	R group 2 Coenzyme A homo sapiens
Ssq23epx[r]	(S)-Squalene-2,3-epoxide
aacoa[c]	Acetoacetyl-CoA
aacoa[m]	Acetoacetyl-CoA
acac[m]	Acetoacetate
acac[x]	Acetoacetate
accoa[c]	Acetyl-CoA
accoa[m]	Acetyl-CoA
accoa[x]	Acetyl-CoA
chlstol[r]	Cholesta-7,24-dien-3beta-ol
chsterol[c]	Cholesterol
chsterol[r]	Cholesterol
coa[c]	Coenzyme A
coa[m]	Coenzyme A
coa[x]	Coenzyme A
ddsmsterol[r]	7-Dehydrodesmosterol
dmpp[x]	Dimethylallyl diphosphate
dsmsterol[r]	Desmosterol
for[r]	Formate
frdp[r]	Farnesyl diphosphate
frdp[x]	Farnesyl diphosphate
grdp[x]	Geranyl diphosphate
hmgcoa[c]	Hydroxymethylglutaryl-CoA
hmgcoa[m]	Hydroxymethylglutaryl-CoA
hmgcoa[x]	Hydroxymethylglutaryl-CoA

ipdp[x]	Isopentenyl diphosphate
lanost[r]	Lanosterol
lthstrl[r]	5alpha-Cholest-7-en-3beta-ol
mev-R[x]	(R)-Mevalonate
ppi[r]	Diphosphate
ppi[x]	Diphosphate
sql[r]	Squalene
xolest_hs[c]	cholesterol ester
zym_int2[r]	zymosterol intermediate 2
zymst[r]	zymosterol
zymstnl[r]	Zymostenol

Table S3. Recon1 cholesterol pathway reactions

Reaction ID	Reaction	Reaction ID	Reaction	Pathway
116	ACACT1	1452	GRTTx	Cholesterol Metabolism
403	C14STRr	1524	HMGCOARx	Cholesterol Metabolism
426	C3STDH1Pr	1525	HMGCOASi	Cholesterol Metabolism
427	C3STDH1r	1526	HMGCOASim	Cholesterol Metabolism
428	C3STKR2r	1529	HMGLm	Cholesterol Metabolism
429	C4STMO1r	1530	HMGLx	Cholesterol Metabolism
430	C4STMO2Pr	1587	IPDDIx	Cholesterol Metabolism
431	C4STMO2r	1642	LNS14DMr	Cholesterol Metabolism
635	DHCR241r	1643	LNSTLSr	Cholesterol Metabolism
636	DHCR242r	1656	LSTO1r	Cholesterol Metabolism
637	DHCR243r	1657	LSTO2r	Cholesterol Metabolism
638	DHCR71r	1710	MEVK1x	Cholesterol Metabolism
639	DHCR72r	2038	PMEVKx	Cholesterol Metabolism
661	DMATTx	2237	SOAT11	Cholesterol Metabolism
774	DPMVDx	2238	SOAT12	Cholesterol Metabolism
800	EBP1r	2259	SQLEr	Cholesterol Metabolism
801	EBP2r	2260	SQLSr	Cholesterol Metabolism

Table S4. Cholesterol pathway gene abreviations

TM7SF2	Transmembrane 7 superfamily member 2
NSDHL	NAD(P) dependent steroid dehydrogenase-like
HSD17B4	Hydroxysteroid (17-beta) dehydrogenase 4
MSMO1	Methylsterol monooxygenase 1
DHCR24	24-dehydrocholesterol reductase
DHCR7	7-dehydrocholesterol reductase
GGPS1	Geranylgeranyl diphosphate synthase 1
FDPS	Farnesyl diphosphate synthase (farnesyl pyrophosphate synthetase, dimethylallyltranstransferase, geranyltranstransferase)
MVD	Mevalonate (diphospho) decarboxylase
EBP	Emopamil binding protein
HMGCR	3-hydroxy-3-methylglutaryl-Coenzyme A reductase
HMGCS1	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble)
HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)
HMGCL	3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase (hydroxymethylglutaricaciduria)
HMGCLL1	3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase-like 1
IDI1	Isopentenyl-diphosphate delta isomerase 1
IDI2	Isopentenyl-diphosphate delta isomerase 2
CYP4F8	Cytochrome P450, family 4, subfamily F, polypeptide 8
CYP51A1	Cytochrome P450, family 51, subfamily A, polypeptide 1
LSS	Lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)
SC5DL	Sterol-C5-desaturase (ERG3 delta-5-desaturase homolog, S. cerevisiae)-like
MVK	Mevalonate kinase (mevalonic aciduria)
PMVK	Phosphomevalonate kinase
SOAT1	Sterol O-acyltransferase (acyl-Coenzyme A: cholesterol acyltransferase) 1
SQLE	Squalene epoxidase
FDFT1	Farnesyl-diphosphate farnesyltransferase 1

Table S5. 100 most significant genes (combined P-value) in Recon1. Table contains combined P-values (Stouffer´s method). Many top ranked genes in the list are verified direct targets of Nrf2. Some genes are involved in many pathways (marked with *number in Table S5).

symbols	name	comb.pval	pathway
GLA	Homo sapiens galactosidase, alpha (GLA), mRNA	0	Sphingolipid Metabolism
GSR	Homo sapiens glutathione reductase (GSR), transcript variant 1, mRNA	0	Glutamate metabolism
PHYH	Homo sapiens phytanoyl-CoA 2-hydroxylase (PHYH), transcript variant 2, mRNA	0	Fatty acid oxidation, peroxisome
TXNRD1	Homo sapiens thioredoxin reductase 1 (TXNRD1), transcript variant 5, mRNA	0	*1
AKR1C1	Homo sapiens aldo-keto reductase family 1, member C1 (dihydrodiol dehydrogenase 1; 20-alpha (3-alpha)-hydroxysteroid dehydrogenase) (AKR1C1), mRNA	0	Steroid Metabolism
GCNT2	Homo sapiens glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group) (GCNT2), transcript variant 2, mRNA	0	Blood Group Biosynthesis
CBR1	Homo sapiens carbonyl reductase 1 (CBR1), mRNA	0	Eicosanoid Metabolism
GCLM	Homo sapiens glutamate-cysteine ligase, modifier subunit (GCLM), mRNA	0	Glutathione Metabolism
HMOX1	Homo sapiens heme oxygenase (decycling) 1 (HMOX1), mRNA	0	Heme Degradation
ME1	Homo sapiens malic enzyme 1, NADP(+)-dependent, cytosolic (ME1), mRNA	0	Pyruvate Metabolism
SLC7A11	Homo sapiens solute carrier family 7 (anionic amino acid transporter light chain, xc- system), member 11 (SLC7A11), mRNA	0	Transport, Extracellular
GCLC	Homo sapiens glutamate-cysteine ligase, catalytic subunit (GCLC), transcript variant 2, mRNA	5.6e-16	Glutathione Metabolism
SLCO2A1	Homo sapiens solute carrier organic anion transporter family, member 2A1 (SLCO2A1), mRNA	4.1e-15	Transport, Extracellular
PLOD2	Homo sapiens procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2 (PLOD2), transcript variant 2, mRNA	1.1e-14	Lysine Metabolism
ELOVL4	Homo sapiens ELOVL fatty acid elongase 4 (ELOVL4), mRNA	1.3e-14	Fatty acid elongation
TALDO1	Homo sapiens transaldolase 1 (TALDO1), mRNA	1.9e-14	Pentose Phosphate Pathway
UGDH	Homo sapiens UDP-glucose 6-dehydrogenase (UGDH), transcript variant 2, mRNA	2,00E-14	Starch and Sucrose Metabolism
SLC3A2	Homo sapiens solute carrier family 3 (activators of dibasic and neutral amino acid transport), member 2 (SLC3A2), transcript variant 2, mRNA	4.2e-14	*2
SLC27A3	Homo sapiens solute carrier family 27 (fatty acid transporter), member 3 (SLC27A3), mRNA	1.6e-13	Transport, Extracellular
RDH11	Homo sapiens retinol dehydrogenase 11 (all-trans/9-cis/11-cis) (RDH11), transcript variant 2, mRNA	2,00E-13	Vitamin A Metabolism
GGCT	Homo sapiens gamma-glutamylcyclotransferase (GGCT), transcript variant 2, mRNA	2.5e-13	Glutathione Metabolism
GNE	Homo sapiens glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase (GNE), transcript variant 1, mRNA	3.5e-13	Aminosugar Metabolism
SLC19A2	Homo sapiens solute carrier family 19 (thiamine	7.3e-13	Transport, Extracellular

	transporter), member 2 (SLC19A2), mRNA		
CHSY3	Homo sapiens chondroitin sulfate synthase 3 (CHSY3), mRNA	3.3e-12	Chondroitin / heparan sulfate biosynthesis
INPP4B	Homo sapiens inositol polyphosphate-4-phosphatase, type II, 105kDa (INPP4B), transcript variant 2, mRNA	3.9e-12	Inositol Phosphate Metabolism
ST3GAL6	Homo sapiens ST3 beta-galactoside alpha-2,3-sialyltransferase 6 (ST3GAL6), transcript variant 1, mRNA	4,00E-12	Blood Group Biosynthesis
AADAC	Homo sapiens arylacetamide deacetylase (AADAC), mRNA	4.5e-12	Alkaloid biosynthesis II
PIK3C2B	Homo sapiens phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 beta (PIK3C2B), mRNA	3.3e-11	Inositol Phosphate Metabolism
CSGALNACT1	Homo sapiens chondroitin sulfate N-acetylgalactosaminyltransferase 1 (CSGALNACT1), transcript variant 2, mRNA	4.2e-11	Chondroitin / heparan sulfate biosynthesis
IDS	Homo sapiens iduronate 2-sulfatase (IDS), transcript variant 1, mRNA	4.4e-11	*3
HS3ST1	Homo sapiens heparan sulfate (glucosamine) 3-O-sulfotransferase 1 (HS3ST1), mRNA	1.1e-10	Chondroitin / heparan sulfate biosynthesis
LIPG	Homo sapiens lipase, endothelial (LIPG), mRNA	1.7e-10	Triacylglycerol Synthesis
CTH	Homo sapiens cystathionase (cystathionine gamma-lyase) (CTH), transcript variant 3, mRNA	1.8e-10	*4
TUSC3	Homo sapiens tumor suppressor candidate 3 (TUSC3), transcript variant 1, mRNA	2.8e-10	Oxidative Phosphorylation
SLC33A1	Homo sapiens solute carrier family 33 (acetyl-CoA transporter), member 1 (SLC33A1), transcript variant 1, mRNA	3.5e-10	*5
UXS1	Homo sapiens UDP-glucuronate decarboxylase 1 (UXS1), transcript variant 1, mRNA	5.9e-10	Nucleotide Sugar Metabolism
ALDH3A2	Homo sapiens aldehyde dehydrogenase 3 family, member A2 (ALDH3A2), transcript variant 2, mRNA	9.3e-10	*6
AK4	Homo sapiens adenylate kinase 4 (AK4), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA	1,00E-09	Nucleotides
RRM2	Homo sapiens ribonucleotide reductase M2 (RRM2), transcript variant 2, mRNA	1.2e-09	Nucleotides
GFPT1	Homo sapiens glutamine--fructose-6-phosphate transaminase 1 (GFPT1), transcript variant 1, mRNA	1.2e-09	Aminosugar Metabolism
GCNT1	Homo sapiens glucosaminyl (N-acetyl) transferase 1, core 2 (GCNT1), transcript variant 3, mRNA	1.3e-09	O-Glycan Biosynthesis
BHMT2	Homo sapiens betaine--homocysteine S-methyltransferase 2 (BHMT2), transcript variant 2, mRNA	1.3e-09	Glycine, Serine, and Threonine Metabolism
SAT1	Homo sapiens spermidine/spermine N1-acetyltransferase 1 (SAT1), transcript variant 1, mRNA	1.9e-09	Arginine and Proline Metabolism
PLA2G12A	Homo sapiens phospholipase A2, group XIIA (PLA2G12A), mRNA	3.1e-09	Glycerophospholipid Metabolism
RDH10	Homo sapiens retinol dehydrogenase 10 (all-trans) (RDH10), mRNA	4.8e-09	Vitamin A Metabolism
ALAS1	Homo sapiens aminolevulinate, delta-, synthase 1 (ALAS1), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA	5.3e-09	Glycine, Serine, and Threonine Metabolism
SULT1A1	Homo sapiens sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1 (SULT1A1), transcript	8.9e-09	*7

	variant 1, mRNA		
GK	Homo sapiens glycerol kinase (GK), transcript variant 2, mRNA	1,00E-08	Glycerophospholipid Metabolism
FADS1	Homo sapiens fatty acid desaturase 1 (FADS1), mRNA	1,00E-08	Fatty acid elongation
PAPSS2	Homo sapiens 3-phosphoadenosine 5-phosphosulfate synthase 2 (PAPSS2), transcript variant 2, mRNA	1,00E-08	*8
MTHFD2L	Homo sapiens methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2-like (MTHFD2L), mRNA	1.4e-08	Folate Metabolism
CHST14	Homo sapiens carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 14 (CHST14), mRNA	1.5e-08	Chondroitin / heparan sulfate biosynthesis
ADSS	Homo sapiens adenylosuccinate synthase (ADSS), mRNA	1.9e-08	Nucleotides
ALDH6A1	Homo sapiens aldehyde dehydrogenase 6 family, member A1 (ALDH6A1), nuclear gene encoding mitochondrial protein, mRNA	2.9e-08	*9
P4HA2	Homo sapiens prolyl 4-hydroxylase, alpha polypeptide II (P4HA2), transcript variant 2, mRNA	3,00E-08	Arginine and Proline Metabolism
ATP6V1B2	Homo sapiens ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2 (ATP6V1B2), mRNA	3.6e-08	Transport, Lysosomal
ATP1B1	Homo sapiens ATPase, Na+/K+ transporting, beta 1 polypeptide (ATP1B1), mRNA	4,00E-08	Transport, Extracellular
CYB5D1	Homo sapiens cytochrome b5 domain containing 1 (CYB5D1), mRNA	4.2e-08	Pyruvate Metabolism
ASNS	Homo sapiens asparagine synthetase (glutamine-hydrolyzing) (ASNS), transcript variant 4, mRNA	4.8e-08	Alanine and Aspartate Metabolism
G6PD	Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), transcript variant 1, mRNA	5.2e-08	Pentose Phosphate Pathway
SLC22A4	Homo sapiens solute carrier family 22 (organic cation/ergothioneine transporter), member 4 (SLC22A4), mRNA	6.2e-08	Transport, Extracellular
ALDH2	Homo sapiens aldehyde dehydrogenase 2 family (mitochondrial) (ALDH2), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA	7,00E-08	*10
GLRX	Homo sapiens glutaredoxin (thioltransferase) (GLRX), transcript variant 3, mRNA	7.7e-08	Ascorbate and Aldarate Metabolism
DGKH	Homo sapiens diacylglycerol kinase, eta (DGKH), transcript variant 3, mRNA	1.4e-07	Glycerophospholipid Metabolism
ETNK1	Homo sapiens ethanolamine kinase 1 (ETNK1), transcript variant 1, mRNA	1.8e-07	Glycerophospholipid Metabolism
HK2	Homo sapiens hexokinase 2 (HK2), mRNA	2.5e-07	*11
RFK	Homo sapiens riboflavin kinase (RFK), mRNA	2.5e-07	Riboflavin Metabolism
KYNU	Homo sapiens kynureninase (KYNU), transcript variant 3, mRNA	2.7e-07	Tryptophan metabolism
PLD1	Homo sapiens phospholipase D1, phosphatidylcholine-specific (PLD1), transcript variant 2, mRNA	4.3e-07	Glycerophospholipid Metabolism
GALK2	Homo sapiens galactokinase 2 (GALK2), transcript variant 2, mRNA	5.2e-07	Galactose metabolism
TYMS	Homo sapiens thymidylate synthetase (TYMS), mRNA	5.3e-07	Nucleotides
PGD	Homo sapiens phosphogluconate dehydrogenase (PGD), mRNA	6.5e-07	Pentose Phosphate Pathway
ADK	Homo sapiens adenosine kinase (ADK), transcript variant 4, mRNA	1,00E-06	Nucleotides
SLC7A5	Homo sapiens solute carrier family 7 (amino acid	1.1e-06	Transport, Extracellular

	transporter light chain, L system), member 5 (SLC7A5), mRNA		
ENO2	Homo sapiens enolase 2 (gamma, neuronal) (ENO2), mRNA	1.2e-06	Glycolysis/Gluconeogenesis
GBE1	Homo sapiens glucan (1,4-alpha-), branching enzyme 1 (GBE1), mRNA	1.2e-06	Starch and Sucrose Metabolism
IP6K2	Homo sapiens inositol hexakisphosphate kinase 2 (IP6K2), transcript variant 2, mRNA	1.2e-06	Inositol Phosphate Metabolism
MAN1C1	Homo sapiens mannosidase, alpha, class 1C, member 1 (MAN1C1), mRNA	1.2e-06	N-Glycan Biosynthesis
GPX3	Homo sapiens glutathione peroxidase 3 (plasma) (GPX3), mRNA	1.2e-06	Glutathione Metabolism
ATP6V1D	Homo sapiens ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D (ATP6V1D), mRNA	1.3e-06	Transport, Lysosomal
UCK2	Homo sapiens uridine-cytidine kinase 2 (UCK2), mRNA	1.4e-06	*12
PIGW	Homo sapiens phosphatidylinositol glycan anchor biosynthesis, class W (PIGW), mRNA	1.4e-06	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
PPAP2B	Homo sapiens phosphatidic acid phosphatase type 2B (PPAP2B), mRNA	1.5e-06	Triacylglycerol Synthesis
SOD1	Homo sapiens superoxide dismutase 1, soluble (SOD1), mRNA	1.9e-06	ROS Detoxification
MAOA	Homo sapiens monoamine oxidase A (MAOA), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA	2.1e-06	*13
ETFDH	Homo sapiens electron-transferring-flavoprotein dehydrogenase (ETFDH), nuclear gene encoding mitochondrial protein, mRNA	2.1e-06	Fatty acid oxidation
ST3GAL5	Homo sapiens ST3 beta-galactoside alpha-2,3-sialyltransferase 5 (ST3GAL5), transcript variant 2, mRNA	2.8e-06	Sphingolipid Metabolism
CHSY1	Homo sapiens chondroitin sulfate synthase 1 (CHSY1), mRNA	3.2e-06	Chondroitin / heparan sulfate biosynthesis
SUOX	Homo sapiens sulfite oxidase (SUOX), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA	4.1e-06	Cysteine Metabolism
MGAT4A	Homo sapiens mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme A (MGAT4A), transcript variant 1, mRNA	4.3e-06	N-Glycan Biosynthesis
SLC26A11	Homo sapiens solute carrier family 26, member 11 (SLC26A11), transcript variant 1, mRNA	4.6e-06	Transport, Extracellular
PYGL	Homo sapiens phosphorylase, glycogen, liver (PYGL), transcript variant 2, mRNA	5.1e-06	Starch and Sucrose Metabolism
PPAT	Homo sapiens phosphoribosyl pyrophosphate amidotransferase (PPAT), mRNA	5.2e-06	IMP Biosynthesis
MSMO1	Homo sapiens methylsterol monooxygenase 1 (MSMO1), transcript variant 2, mRNA	5.6e-06	Cholesterol Metabolism
AKR1B1	Homo sapiens aldo-keto reductase family 1, member B1 (aldose reductase) (AKR1B1), mRNA	5.6e-06	*14
SLC26A2	Homo sapiens solute carrier family 26 (sulfate transporter), member 2 (SLC26A2), mRNA	5.8e-06	Transport, Extracellular
AGPAT5	Homo sapiens 1-acylglycerol-3-phosphate O-acyltransferase 5 (lysophosphatidic acid acyltransferase, epsilon) (AGPAT5), mRNA	7,00E-06	Triacylglycerol Synthesis
CMAS	Homo sapiens cytidine monophosphate N-acetylneuraminic acid synthetase (CMAS), mRNA	9,00E-06	Aminosugar Metabolism
SLC12A6	Homo sapiens solute carrier family 12	9.2e-06	Transport, Extracellular

	(potassium/chloride transporters), member 6 (SLC12A6), transcript variant 3, mRNA		

*Table S5, *.*

*1	*2	*3
TXNRD1	SLC3A2	IDS
Nucleotides	Transport, Extracellular	Heparan sulfate degradation
Miscellaneous	Starch and Sucrose Metabolism	Chondroitin sulfate degradation
*4	*5	*6
CTH	SLC33A1	ALDH3A2
Cysteine Metabolism	Transport, Extracellular	Tryptophan metabolism
Selenoamino acid metabolism	Transport, Endoplasmic Reticular	Fatty Acid Metabolism
	Sphingolipid Metabolism	Glycolysis/Gluconeogenesis
		beta-Alanine metabolism
		Glyoxylate and Dicarboxylate Metabolism
		Ascorbate and Aldarate Metabolism
		Histidine Metabolism
		Pyruvate Metabolism
		Arginine and Proline Metabolism
		Limonene and pinene degradation
*7	*8	*9
SULT1A1	PAPSS2	ALDH6A1
CYP Metabolism	Selenoamino acid metabolism	Glyoxylate and Dicarboxylate Metabolism
Tyrosine metabolism	Nucleotides	Pyruvate Metabolism
Steroid Metabolism		Valine, Leucine, and Isoleucine Metabolism
		Propanoate Metabolism
*10	*11	*12
ALDH2	HK2	UCK2
Tryptophan metabolism	Glycolysis/Gluconeogenesis	Nucleotides
Glycolysis/Gluconeogenesis	Aminosugar Metabolism	Pyrimidine Biosynthesis
beta-Alanine metabolism	Fructose and Mannose Metabolism	
Glyoxylate and Dicarboxylate Metabolism		
Ascorbate and Aldarate Metabolism		
Histidine Metabolism		
Pyruvate Metabolism		
Arginine and Proline Metabolism		
Limonene and pinene degradation		
*13	*14	
MAOA	AKR1B1	

Tryptophan metabolism	Glycine, Serine, and Threonine Metabolism*
Tyrosine metabolism	Pyruvate Metabolism
Arginine and Proline Metabolism	Pentose and Glucuronate Interconversions
Phenylalanine metabolism	Galactose metabolism
	Fructose and Mannose Metabolism

Figure S4 Barplots on significant genes (combined P-value <0.01) in Transport Lysosomal, Glutamate metabolism, Ascorbate and Aldarate Metabolism, Cholesterol Metabolism and steroid metabolism pathways

Available in:

https://www.wuala.com/ppolonen/Supplementary%20material%20(Petri%20Pölönen´s%20Master´s%20Thesis%202013)/?key=4D2bmwkTewaR

Or contact petri.polonen@uef.fi

Figure S5 visualization tracks for cholesterol metabolism pathway RNA- and GRO-seq data and ENCODE chromatin markers.

Available in:

https://www.wuala.com/ppolonen/Supplementary%20material%20(Petri%20Pölönen´s%20Master´s%20Thesis%202013)/?key=4D2bmwkTewaR

Or contact petri.polonen@uef.fi