Analysis of tissue specific regulatory targets of co-factor Pgc-1α using bioinformatics methods

Krista Kokki
Master's thesis
Master of Science program in Biosciences, major in Bioscience
University of Finland, Faculty of Science and Forestry
University of Eastern Finland
October 2015

UNIVERSITY OF EASTERN FINLAND, Faculty of Health Sciences
Master of Science Program in Biosciences
KRISTA KOKKI: Analysis of tissue specific regulatory targets of Pgc-1α using bioinformatics methods
Master's thesis, 103 pages
Instructors: Merja Heinäniemi (Docent), Petri Pölönen (M.Sc)
October 2015

Key words: Pgc-1α, enrichment analysis, cardiac hypertrophy, circadian rhythm

Abstract

Background. Peroxisome proliferator-activated receptor gamma co-activator-1 alpha (Pgc-1α) is coactivator heavily involved in cellular metabolism and energy homeostasis. It has been linked to hypertrophy in muscle tissues and identified as a putative target for treatment, but it remains unclear if it could be used for this purpose. The effect of overexpression of Pgc-1α across tissues is not known either, nor is its resemblance to physiological and pathological hypertrophy.

Aims. The goal of the study was to investigate if Pgc-1α overexpression would be beneficial in treating cardiac hypertrophy by applying bioinformatics methods on genome-wide RNA expression profiles. The effect of Pgc-1α overexpression between heart and skeletal muscle was investigated in the respective tissues, and resemblance to physiological and pathological hypertrophy was studied.

Methods. RNA-seq Pgc-1α overexpression dataset from mice was studied in comparison to publicly available RNA-seq and microarray experiments. The data was computationally processed, and results were analyzed by using variety of bioinformatics methods, such as gene set enrichment.

Results. Gene set enrichment and pathway analysis revealed metabolic differences between Pgc-1α overexpression in heart and skeletal muscle. As expected, statistical analysis revealed Pgc-1α overexpression to resemble physiological rather than pathological cardiac hypertrophy. Surprisingly, Pgc-1α overexpression was also found to cause downregulation of the circadian clock genes.

Conclusions. The effect Pgc-1α overexpression was found to differ between heart and skeletal muscle, and it was found to resemble physiological rather than pathological cardiac hypertrophy. However, it seems that uncontrolled Pgc-1α overexpression disrupts circadian rhythm and thus affects its possibility as treatment target. Despite this, Pgc-1α may be a possible target in treating cardiac hypertrophy, but the success may lay in greatly controlling the overexpression, thus making clinical window likely to be very narrow. Nonetheless, further studies concerning the effects of Pgc-1α overexpression in circadian rhythm are necessary before approving or discarding its possibility as treatment target.

Acknowledgements

This study was done in the University of Eastern Finland, Institute of Medicine and A.I Virtanen institute, department of Biotechnology and Molecular Medicine. I would like to acknowledge Molecular Physiology group leader Pasi Tavi (PhD) for giving me the possibility to work with their data.

I am ever grateful for my main supervisor Merja Heinäniemi for allowing me to work in her group and providing opportunity to learn in an excellent team. I also want to thank her for all the work she's done for me, not only supervising this thesis but also helping me to grow as a scientist.

I also want to thank my supervisor and co-worker Petri Pölönen for his comments, guidance and help with this thesis and programming in general. Both of my supervisors have supported and motivated me through the projects.

Last but not least, I want to thank the entire Systems Genomics group for excellent work environment, help and friendliness.

Helsinki, October 2015

Abbreviations

DEG	differentially expressed gene
TF	transcription factor
DNA	deoxyribonucleic acid
SCN	suprachiasmatic nucleus
RNA	ribonucleic acid
mRNA	mature RNA
RNA-seq	RNA sequencing
cDNA	complimentary DNA
rRNA	ribosomal RNA
KEGG	Kyoto Encyclopedia for Genes and Genomes
GSEA	gene set enrichment analysis
GSA	gene set analysis
FDR	false discovery rate

Genes:

Pgc-1α	peroxisome proliferator-activated receptor gamma co-activator-1 alpha
Nrf	nuclear respiratory factor
Ppar	peroxisome proliferator-activated receptor
Clock	circadian locomoter output cycles protein kaput
Bmal1/Arntl	aryl hydrocarbon receptor nuclear translocator-like protein 1
Per	period
Cry	cryptochrome

ROR	retinoid orphan receptor
Rev-erb	nuclear receptor subfamily 1

Table of Contents

1 INTRODUCTION

High-throughput experiments, such as next-generation sequencing, have generated large amounts of genome-wide expression data. These are collected in public databases, available to everyone. The challenge no longer lies in generating the data but interpreting the results in biologically meaningful way. At first, it was thought that biological mechanisms could be detected from the genes showing the largest differences. This, however, proved to have technical and biological limitations, resulting in false biological interpretations at its worst[1–3].

Gene sets were introduced to overcome the challenges that focusing to a few genes brought up. Gene sets are group of genes sharing the same function, defined based on prior knowledge, such as biological pathway. The approach was developed to discover differences between two distinct phenotypes, such as wild type versus tumor sample. Roughly, if gene set is associated with phenotype, enriched in other words, it contains more differentially expressed genes (DEGs) that could be expected by chance alone. Enrichment analysis have also been extended to classification based on expression levels (such as identification of tumor samples) and even to gene regulatory network analyses[3].

In this master's thesis, gene set enrichment analysis was used to study the overexpression of coactivator Pgc-1α, master regulator of energy metabolism. Our own gene expression data was compared to publicly available genome-wide datasets by using computational methods to discover and compare the regulated pathways between different tissues and disease states.

This master's thesis is divided into two main parts: literature review and experimental part. In literature review, biology of Pgc-1α and its role in heart, in ground and hypertrophy states along with function in muscle tissue and importance in regulation of the circadian rhythm is introduced. The main part of the literature review focuses on gene set enrichment analyses. Genome-wide expression data is also introduced. The experimental part consists of presentation of the aims, results and discussion of conclusions. Descriptions of the algorithms used in analysis of gene expression data and gene set enrichment used in this thesis are depicted in materials and methods section.

2 REVIEW OF LITERATURE

2.1 Introduction to physiological role of Pgc-1α

2.1.1 Normal state

Pgc-1α, peroxisome proliferator-activated receptor gamma co-activator-1 alpha, is a co-activator involved in variety of regulatory functions in cellular metabolism and energy homeostasis. Due to its nature as co-activator, Pgc-1α regulates gene expression through protein-protein interactions with transcription factors (TFs) that possess deoxyribonucleic acid (DNA) binding domains rather directly binding to DNA. A single co-activator is capable of interacting with variety of TFs and thus regulating the expression of numerous genes and biological processes. It has also been suggested that co-activators can be post-translationally modulated by intracellular pathways and targeted by ubiquitination[4].

Pgc-1α is preferentially expressed in tissues with high oxidative capacity, such as heart, skeletal muscle and brown adipose tissue but high expression has also been detected in brain and kidney. In these tissues, Pgc-1α has critical role in the regulation of mitochondrial biogenesis and energy metabolism[5].

Moreover, Pgc-1α is one of the main regulators of nuclear respiratory factor-1 (Nrf-1) and -2 (Nrf-2) and has been shown to increase and co-activate these TFs and their target genes[5,6]. Nrfs have been widely studied due to their role in mitochondrial biogenesis. By regulating the expression of mitochondrial transcription factor A (Tfam), coordinator between mitochondrial and nuclear activation during mitochondrial biogenesis, they are responsible for transcription and replication of mitochondrial genes[6,7]. The expression of nuclear respiratory chain subunits and other proteins required for mitochondrial functions are also controlled by Nrfs[5].

Nrfs are the key interaction partners of Pgc-1α but the prominent effect of Pgc-1α on biological processes can't be explained by these interactions alone. Pgc-1α has been linked to variety of other biological energy metabolism processes within and outside mitochondria. These include mitochondrial fatty acid oxidation and oxidative phosphorylation inside the mitochondria and gluconeogenesis, cellular respiration and electron transport chain outside the mitochondrion[4,5,8].

Pgc-1α has also been shown to widely co-operate with nuclear receptor family members, such as glucocorticoid receptor and peroxisome proliferator-activated receptors (Ppar) α and -γ. However, it has also been suggested that these interactions are species-specific. According to a study which examined interactions of transcriptionally regulated proteins, there's no interaction between Pgc-1α and Ppar-α and -γ in mouse, whereas in human, these interactions occur[9].

2.1.2 Skeletal muscle - Pgc-1α in exercise

Physical exercise and training have been linked with lower mortality and reduced prevalence of metabolic diseases. Physical inactivity, accompanied with low whole-aerobic capacity, muscle mitochondrial content and oxidative activity have been associated with development of metabolic disorders. Hence the improvement of skeletal muscle function, especially its oxidative metabolism, is considered as a possible intervention point in treatment and prevention of metabolic diseases[10].

It is also known that increased contractile activity, such as endurance exercise training, promotes fibre-type transformation in skeletal muscle. One of the key players in this transformation is Pgc-1α. Exercise training induces the upregulation of muscle Pgc-1α levels which improves not only muscle fibre-type switching (from high speed glycolytic towards high endurance oxidative fibres) through calcium cascade but also mitochondrial biogenesis, fatty acid oxidation and variety of other important pathways. The overexpression of Pgc-1α in skeletal muscle has also been demonstrated to increase glucose uptake, causing prevention of depletion of glycogen and thus improving performance. Moreover, skeletal muscle specific Pgc-1α knockout mice have been shown to have abnormal glucose homeostasis[10–12].

2.1.3 Heart physiology and exercise - hypertrophy and Pgc-1α

Heart is an organ with excessive energy requirements. These needs are met by high-capacity mitochondrial system, accounting for over 90 % of energy production for cardiac muscle[13]. One of the main regulators of this mitochondrial biogenesis is Pgc-1α, cofactor highly expressed in the heart.

Decrease of Pgc-1α has been linked with conversion of fatty acid oxidation to glycolytic metabolism which causes cardiac hypertrophy, thickening of the heart muscle[13]. However, there are two types of hypertrophy: physiological and pathological. Physiological hypertrophy is a natural state of a heart which takes place when there's physiological increase in demand of the heart. This may happen during training or pregnancy, for example. On the other hand, pathological hypertrophy is severe state associated with loss of cardiomyocytes and heart failure[14].

Whereas decrease in Pgc-1α expression has been linked with conversion from fatty acid oxidation to glycolytic metabolism, overexpression of the cofactor has been associated with elevated levels of mitochondrial biogenesis, fatty acid beta-oxidation[8] and electron transport chain[15], for example. Improvement of mitochondrial function in cardiac diseases such as hypertrophy may be reached through Pgc-1α overexpression, marking it as potential treatment target[8].

However, massive overexpression of Pgc-1α has been linked with dilated cardiomyopathy, disease in which heart muscle doesn't contract normally, resulting in inefficient blood pumping[8]. Therefore, therapeutic overexpression of Pgc-1α should be moderate and approached with caution.

2.2 Role in the regulation of the circadian rhythm

Circadian rhythm is a biological process controlled by circadian clock, which responds to self-sustained day/night cycle of the organism's environment, naturally ~24 hour rhythm. In addition to the light, nutrient availability is the most important entrainment cue. This light-dark

cycle, which most living things, including animals, plants and most microbes possess, drastically affects bodily functions, such as behavior, metabolism and body temperature[16–21]. The disruptions of the system have been linked to variety of diseases, including obesity, diabetes, certain cancers and mental problems, such as depression and schizophrenia[16,19]. Pgc-1α, key regulator of cellular metabolism, has been identified as the key link between changes in metabolism and circadian clock.

At the molecular level, the circadian clock represents a complex gene regulatory network composed of positive and negative feedback loops (Figure 1.). The major purpose of the circadian clock is to produce rhythms in behavior and physiology. This can be achieved through rhythmic expression of genes encoding regulators and enzymes of various metabolic pathways which must have different phases across the organism depending of the tissue. Therefore, the so-called "master clock" suprachiasmatic nucleus (SCN), synchronizes the "peripheral clock", system present virtually in all tissues[16,17]. According to the current orchestra model, each peripheral clock plays its own "instrument" but central clock guides the "melody", in this case physiological output rhythms. Thus, each peripheral clock adapts to its own internal and external stimuli, such as feeding cues from liver and kidney, but light-dark cues are sensed by central clock[22]. Competing model is known as master-slave model and suggests that peripheral clocks are synchronized by SCN and not affected by external or internal stimuli. However, recent studies support the former hypothesis rather than the latter[22]. It is also justifiable in biological sense; the major purpose of the circadian clock is, after all, to produce rhythms in behavior and physiology. This can be achieved through rhythmic expression of genes encoding regulators and enzymes of various metabolic pathways. These output pathways must have different phases across the organism depending of the tissue and thus, a single circadian transcription factor with inflexible activity phase would do a poor job[22].

Despite the extensive research of circadian clock, not all of its components are yet known. Currently, two genes lie at the core of the regulation: *Clock* (Circadian locomoter output cycles protein kaput) and *Bmal1* (also known as *Arntl*; aryl hydrocarbon receptor nuclear translocator-like protein 1). They encode TFs that activate the transcription of co-repressors belonging to the *Period* (Per1, Per2, Per3) and *Cryptochrome* (Cry1, Cry2) gene families, which in turn represses Clock/Bmal1 activity, thus causing inhibition of their own expression[18–20]. This forms the first feedback loop. Another feedback loop, formed by family members of ROR (retinoid orphan receptor) and Rev-erb (nuclear receptor subfamily 1) controls the expression of *Bmal1* and *Clock* through feedback loop[23]. Both REV-ERBα (also known as NR1D2; nuclear receptor

subfamily 1, group D, member 2) and RORa (RAR related orphan receptor A) directly affect the circadian clock by regulating Bmal1. This promotion of *Bmal1* is the approach used by Pgc-1α[18–20]. Due to these two feedback loops, the clockwork can regulate the expression of target genes through two widely different phases. In addition, the accumulation of some proteins requires time, delaying the phase of some circadian output regulators, further driving rhythmic transcription[23]. Interestingly, it has been suggested that effects of Pgc-1α, at least on *Per* and *Cry*, are tissue specific[21].

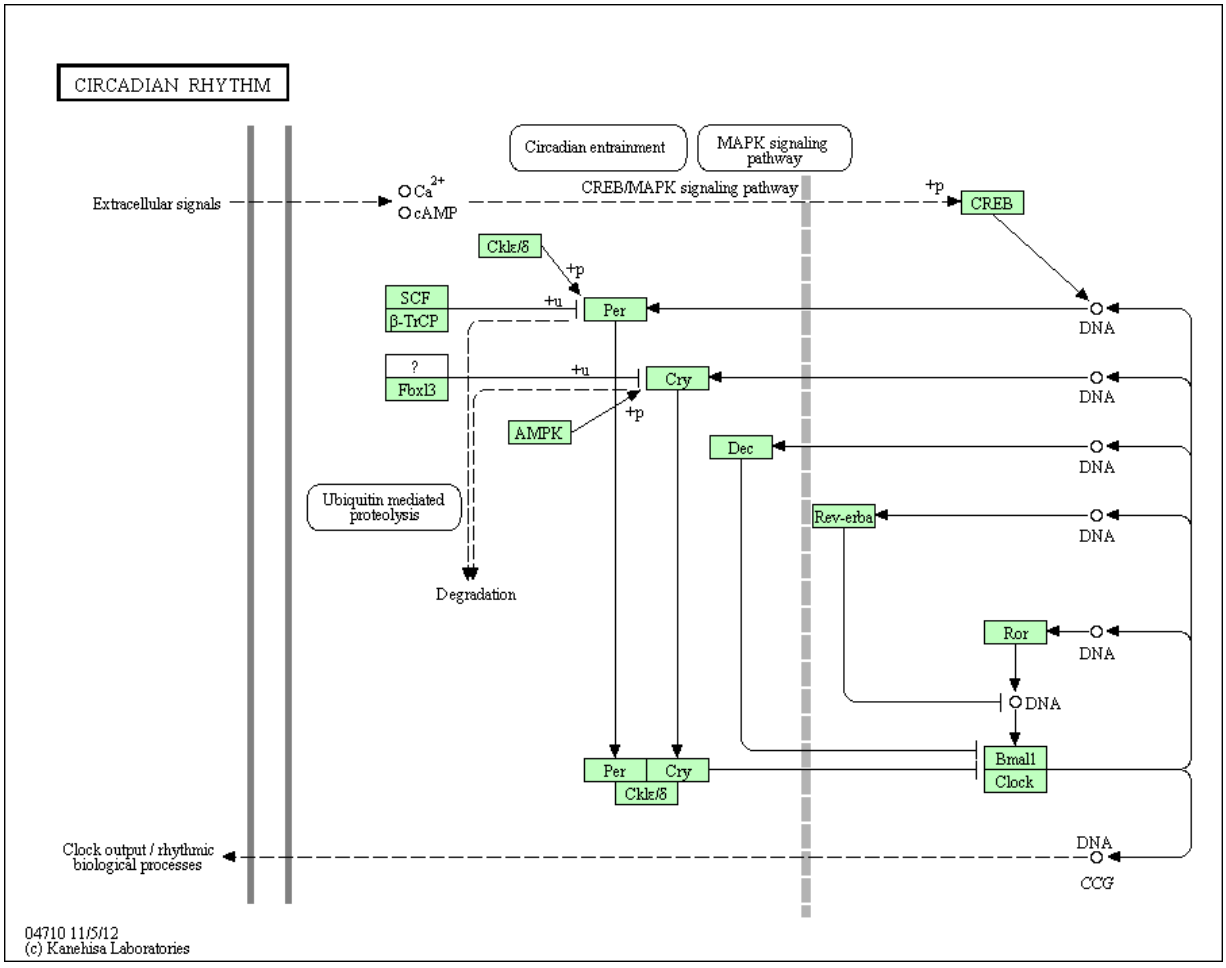

Figure 1. Schematic picture of the mouse circadian pathway from KEGG. According to pathway database Kyoto Encyclopedia of Genes and Genomes (KEGG), the pathway is entirely conserved with human.

2.3 Genome-wide gene expression data

The development of the high-throughput experiments during the last decade has enabled the possibility to inspect the whole genome at once, instead of looking only one or two genes. Variety of genome-wide methods have been implemented, allowing the expression changes to be measured on many levels; for example, microarray and RNA-sequencing (ribonucleic acid) measure mature RNA (mRNA) levels whereas chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) measures protein-DNA interaction.

Due to these genome-wide methods, biology has become a rather data-rich field[1]. This large amount of data is collected in public archives, granting worldwide access to everyone[2]. Therefore the challenge no longer lies in generating the data but in analyzing it. Computational methods have become necessary in handling, processing and analyzing the data, leading to implementation of numerous tools[3] – for example, TopHat is able to align the RNA-sequencing (RNA-seq) reads[24] and Integrative Genomics Viewer[25] can be used to visualize genomic datasets.

High-throughput genome-wide experimentation has also led to the characterization of most components of organisms and therefore the focus has shifted from molecules to networks. In other words, it is of interest to understand how these molecules work with each other as a part of a whole organism instead of studying the components one by one[1]. One way to understand these biological networks better is pathway analysis which often derives from gene set enrichment analysis. These results can be further visualized in, for example, Cytoscape[26].

2.4 Microarrays

Microarrays are one of the most popular high-throughput methods and like other genome-wide methods, they allow the investigation of thousands of genes at a time. Microarrays have been successfully used for detecting gene expression, single nucleotide polymorphisms (SNPs), alternative RNA splicing and so on[27].

There are mainly two types of DNA arrays. The first type, preferred in clinical research, uses small single-stranded oligonucleotides whereas the second uses complementary DNA (cDNA) to measure the level of mRNA[27]. With microarray, it is also possible to measure exon-level expression. These exon arrays differ from traditional microarrays in terms of design of control probes for background correction and in number and placement of the oligonucleotide probes. Due to these more evenly distributed probes and their higher coverage, it has been estimated that these exon arrays are able to provide more accurate measurements of gene expression than traditional microarrays[28,29].

The general microarray experiment process is shown in Figure 2. The first step of the generic microarray experiment is to collect mRNA molecules present in the cell at time point of interest. To determine which genes are expressed in the cell and which are not, mRNA molecules are labeled with reverse transcriptase which generates complementary cDNA to mRNA. During this process, either extracted mRNA or cDNA is dyed with fluorescence. Depending of the experimental design, researcher may use more than one dye in the experiment. In comparison studies, for example, control samples can be dyed with green and treated samples with red.

After dyeing labeled cDNAs are placed onto microarray slide and hybridized by incubating. Next, the array is washed to remove non-specific hybridization. After this, the light generated by fluorescent dye/dyes is detected by scanner and digital image is generated. A very bright fluorescent area corresponds to high amount of mRNA which in turn corresponds to more labeled cDNAs. Genes that are less active produce less mRNA, corresponding to less cDNA which shows as dimmer fluorescent area. In short, the brighter the area, the higher the gene expression. Finally, the digital image is transformed to numerical reading for each spot and processed by integration of intensities and subtraction of the background noise. The final value is then proportional to the concentration of the target sequence of the sample. At last, these values need to be computationally analyzed to gain the results[27].

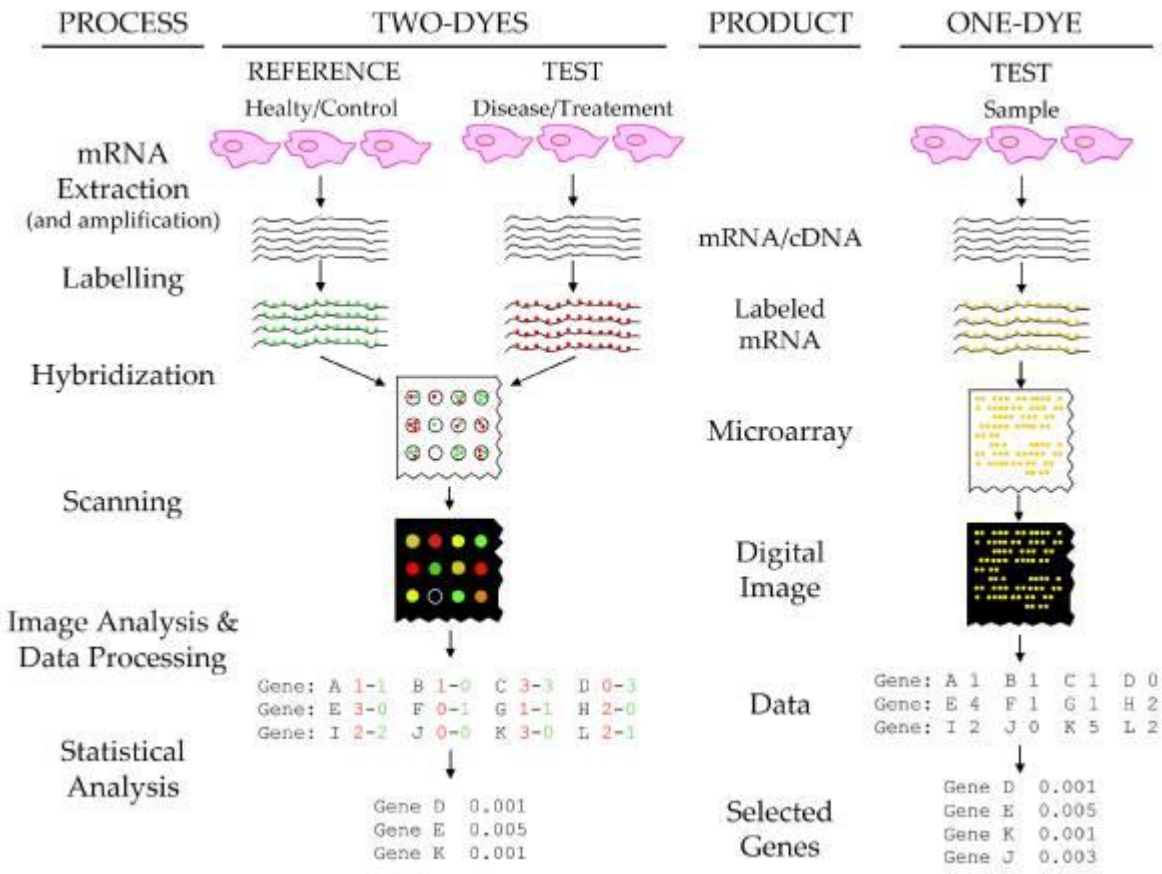

Figure 2. Representation of a general microarray experiments. Arrows represent process and pictures or text represent the product. Left figure represents two-dye experiment and the right figure one-dye experiment. Figure taken from reference 27.

Although microarrays are effective and widely used method, they suffer for quite a few shortcomings, such as low sensitivity for genes expressed in high and low levels[30] and specificity due to non-specific hybridization[31]. Signals of greater intensity (bright spots), in other words genes with high expression, saturate due to large dynamic range of gene expression[32], whereas genes with low expression are often lost in corrections for the background[33]. However, one of the biggest issues of microarrays lies in probe design. In general, each probe represents gene or transcript of interest. These probes differ in their hybridization properties and arrays are limited to interrogating only those genes for which probes are designed[34]. Each probe is part of a probe set, a collection of probes to interrogate target sequence, such as gene or group of highly similar genes. Differently designed probe sets result in different results, naturally[35].

The probe design is where the traditional 3' and exon arrays differ. In exon arrays, up to four probes are selected for the exonic region whereas traditional 3' expression arrays only target the end of mRNA sequence. The biggest difference lies in this; whereas traditional 3' arrays are designed to detect only the gene expression level, the exon array is able to detect the expression of each exon (Figure 3.). In both, the background is determined with separate probes to which none of the gene transcripts binds[29].

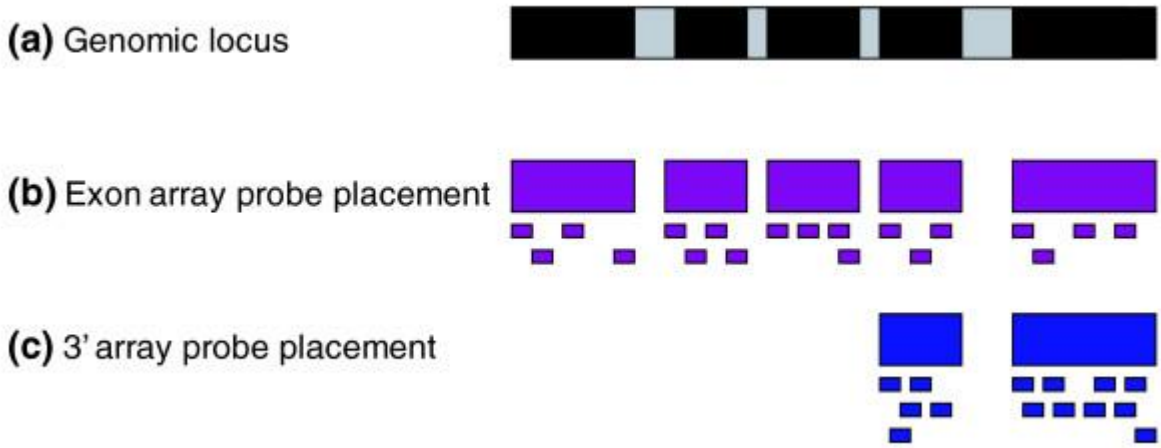

Figure 3. Probe design of exon arrays. a) Exon-intron structure of a gene. Black boxes represent exons. Gray boxes represent introns. Introns are not drawn to scale. b) Probe design of exon arrays. Four probes target each putative exon. c) Probe design of 3' expression arrays. Probes target the 3' end of the mRNA sequence. Figure taken from reference 29.

The expression changes in genes regulated at the post-translational level cannot be detected with arrays[36] either because measuring mRNA simply doesn't reveal the post-transcriptional expression changes. Therefore, it is crucial to understand what can and cannot be measured with method of choice in order to design effective experiment.

2.5 RNA-seq

RNA-sequencing is a genome-wide high-throughput method which approaches to avoid weaknesses of microarrays. In comparison to microarrays, RNA-seq has advantages, such as very low background signal, large range of expression levels over the detection of transcripts and high accuracy of expression levels. Most importantly, RNA-seq doesn't require probe design like microarrays and is therefore devoid of its issues[1].

In practice, the first step is the same in both microarray and RNA-seq; the collection of mRNA. After the collection, RNAs, total or fragmented, are converted into cDNA library. Sequencing adaptors are added to cDNA fragments and short sequences (30-400 bp, depending of the used technology) are obtained via high-throughput sequencing technology. Sequences can be obtained from one end (single-end sequencing) or both ends (pair-end sequencing). The resulting reads are traditionally aligned to reference genome or reference transcripts and classified as three types: exonic reads, junction reads and poly(A) end –reads. With these, expression profile is generated for each transcript[30]. A typical RNA-seq experiment is depicted in Figure 4.

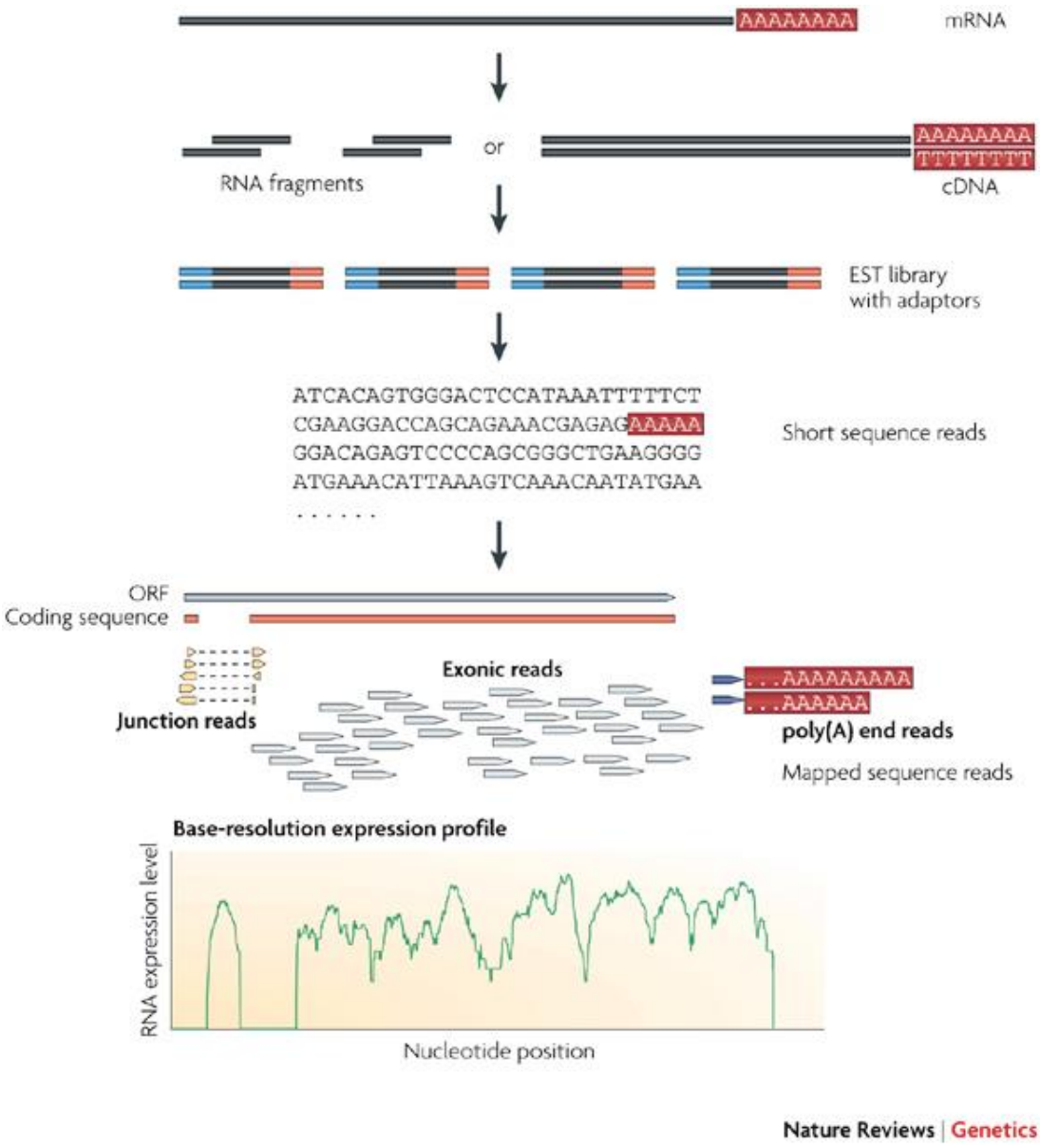

Figure 4. A typical RNA-seq experiment. Briefly, mRNA is converted into cDNA library with adaptors (blue) and adaptors are added to the reads. The short reads are aligned with the reference genome or transcriptome and classified as junction, exonic or poly(A) end reads. These are used to create base-resolution expression profile. Figure taken from reference 9.

While RNA-seq is superior to many methods, it isn't devoid of issues. Even if RNA-seq has lower technical variation, high level of reproducibility for both technical and biological replicates and smaller requirement for the amount of RNA samples, RNA-seq data has GC bias, it can suffer from mapping ambiguity for paralogous sequences and higher statistical power is needed to detect changes at higher counts[30,37]. There are also informatics challenges for complex and large transcriptomes due to numerous sequence read matches in multiple locations

of the genome[30]. Therefore the challenge lies in developing computational methods that can take care of these problems and are still simple enough to use. The biggest problem of RNA-seq is, however, abundance of ribosomal RNA (rRNA). RRNA is the most abundant RNA type, constituting 70-80 % of RNA in most species. Unless researcher is interested in rRNA, these must be removed from total RNA before sequencing to assure sufficient coverage of mRNA. Variety of methods have been implemented to overcome this challenge, such as enrichment of poly-A RNA transcripts[38–41]. Unlike RNA-seq, microarray is devoid of this problem due to pre-designed probes.

Nevertheless, both microarray and RNA-seq are robust, extensively used techniques and while they haven't commonly been integrated, it has been reported that they complement each other in transcriptome profiling and even in finding target genes of a transcription factor[41].

2.6 Gene set enrichment methods

Genome-wide expression analysis, such as microarray or RNA-seq, has become widely employed in research. A successful experiment results in long lists of differentially expressed genes. These DEGs are genes from collection of samples belonging to one or two classes, for example drug treated samples versus control samples. With statistical method of choice, for there are variety of them available, these individual genes have been extracted. Today, the challenge lies in interpreting these results and gaining insight of biological mechanisms beneath.

One common approach is to focus on top and bottom genes of the list but this approach has its issues. By making conclusions solely based on expression levels of the genes showing the largest difference, the obtained results suffer greatly from poorly reproducible results and great information loss of associated genes due to strict cut-off and weak connection with the phenotype[3,42,43].

To overcome these issues shift from single genes to gene sets took place. Usage of gene sets makes it possible to gather also weak expression changes due to large set of genes showing significant pattern[43]. Gene sets, groups of genes sharing same function, are defined based on prior knowledge and constructed without reference to the data[3]. The gene sets needed in

analysis are usually obtained from databases such as Kyoto Encyclopedia for Genes and Genomes (KEGG)[44] and Biocarta[45].

A variety of methods exists to analyze statistical over-representations of genes in gene sets. Naturally, the result varies depending of the chosen method. One of the most common is Gene Set Enrichment Analysis (GSEA)[3] which can be incorporated with programs such as R/Bioconductor and Java but also has a graphical user interface which doesn't require any programming skills. Other easy and commonly used analyzing tools include Graphiteweb[46] and Database for Annotation, Visualization and Integrated Discovery (DAVID)[47]. These two are public web servers for analysis and visualization of pathways. However, this kind of public and easy to use tools have limitations – for example, DAVID limits the maximum numbers of genes in a list and uses its own test (variation of hypergeometric test). In comparison, in Graphiteweb the analysis method can be chosen but there are only two pathway databases and three species of which to choose from.

In general, there are two types of enrichment analysis: class prediction and class discovery. The two answer to different type of questions.

Unsupervised class discovery searches for unknown biologically relevant taxonomy identified by set of co-expressed genes, for example. Question to which this type of analysis can be, for example: *"Which gene sets are enriched in a list of differentially expressed genes?"*

In supervised class prediction, the idea is fundamentally different. Class prediction methods aim to build up a model that can be used for classification and prediction of sample classes. This can be achieved through supervised learning, by usage of training data, for example. Therefore, research question could be: *"In which samples is pathway X active?"*

However, despite the fundamental difference between these two learning methods, the basic idea of the analysis' workflow is the same (Figure 5.).

Figure 5. Overview of gene enrichment method pipeline. The learning method is chosen based on the research question. After choosing the learning method, statistical method is chosen based on null hypothesis and significance is estimated with method of choice. Finally, significance value is calculated.

2.6.1 Pathway/gene set databases

Gene sets are groups of genes sharing the function. One example of a gene set is a pathway, which can be described as a set of biochemical reactions that are linked: one product is reactant or result of a subsequent reaction. In pathway database, this information is stored to describe biochemical reactions. The description is often of metabolic pathways, but may also be something else[48]. Examples of the databases include KEGG[44], Reactome[49] and Wikipathways[50].

There are variety of ways to build up a gene set. For example, in The Molecular Signatures Database (MSig)[3], a collection of annotated gene sets, the gene sets are divided into eight major collections. These collections include curated and motif gene sets, oncogenic and immunologic signatures, for example. Curated gene sets are collections from various online pathway databases and publications whereas motif gene sets contain genes that share conserved cis-regulatory motif across certain species. Oncogenic signatures represent signatures of cellular pathways whose function has been impaired in cancer, generated mainly from microarray data from National Center for Biotechnology Information whereas immunologic signatures represent cell states and disruptions within the immune system, generated by manual curation from the published immunology studies.

When running analysis' using gene sets from databases it is important to understand that even if the databases seemingly have the same pathway, the results may differ. This is mainly due to different annotations which derive from differences in the references to which the pathways are based on. Some pathways may also lack crucial information, such as citations and connections between the genes. Therefore, it is essential to not blindly trust the data and critically think the biological sensibility of the result.

2.6.2 Unsupervised learning methods

Unsupervised class discovery may be separated into two classes: competitive and self-contained methods. The biggest difference between the two is how the null hypothesis, which affects the choice of statistics, is formulated at the beginning of an experiment.

Competitive null hypothesis may be: *"H0. The genes in the gene set of interest are at most as often differentially expressed as the genes in the background gene set."*

Whereas in the self-contained null hypothesis may be*: "H0. No genes in the gene set are differentially expressed."*[51]

Depending of the null hypothesis and the method of choice, local and/or global statistics are used for statistical calculations of the enrichment (Figure 6.).

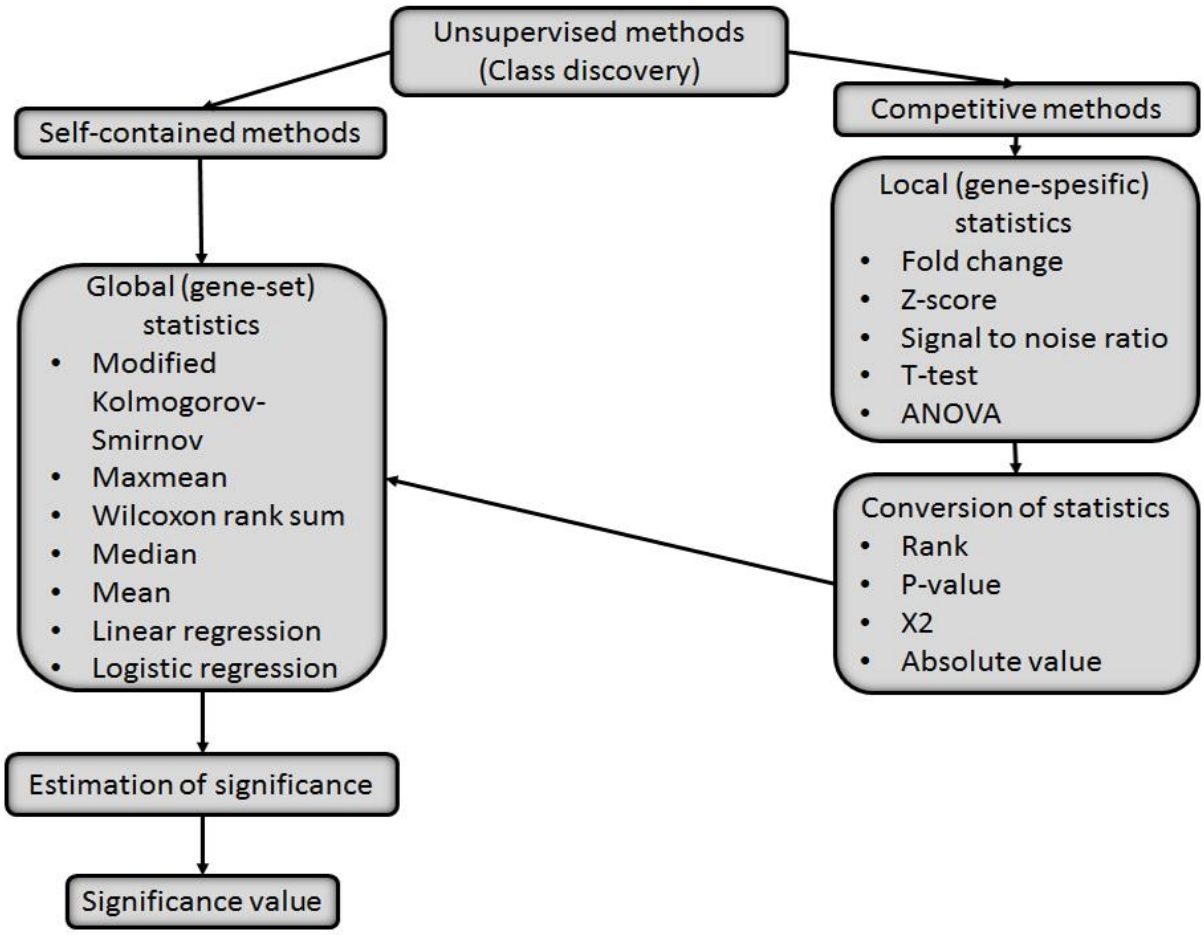

Figure 6. Schematic overview of unsupervised enrichment analysis, focus on statistics. In unsupervised methods, competitive or self-contained method is chosen based on null hypothesis. In competitive methods, local (gene-specific) statistics are first calculated and then converted so that global statistics (gene-set) can be calculated. In self-contained methods, global statistics are calculated without calculating local statistics first. For each statistic, variety of methods are available. After calculating global statistics, significance is estimated and eventually, significance value is calculated for each gene-set.

After statistical calculations, significance of the result is calculated. There are non-parametric and parametric methods for significance calculation. In non-parametric methods, significance value is calculated via permutation or rotation (Figure 7.).

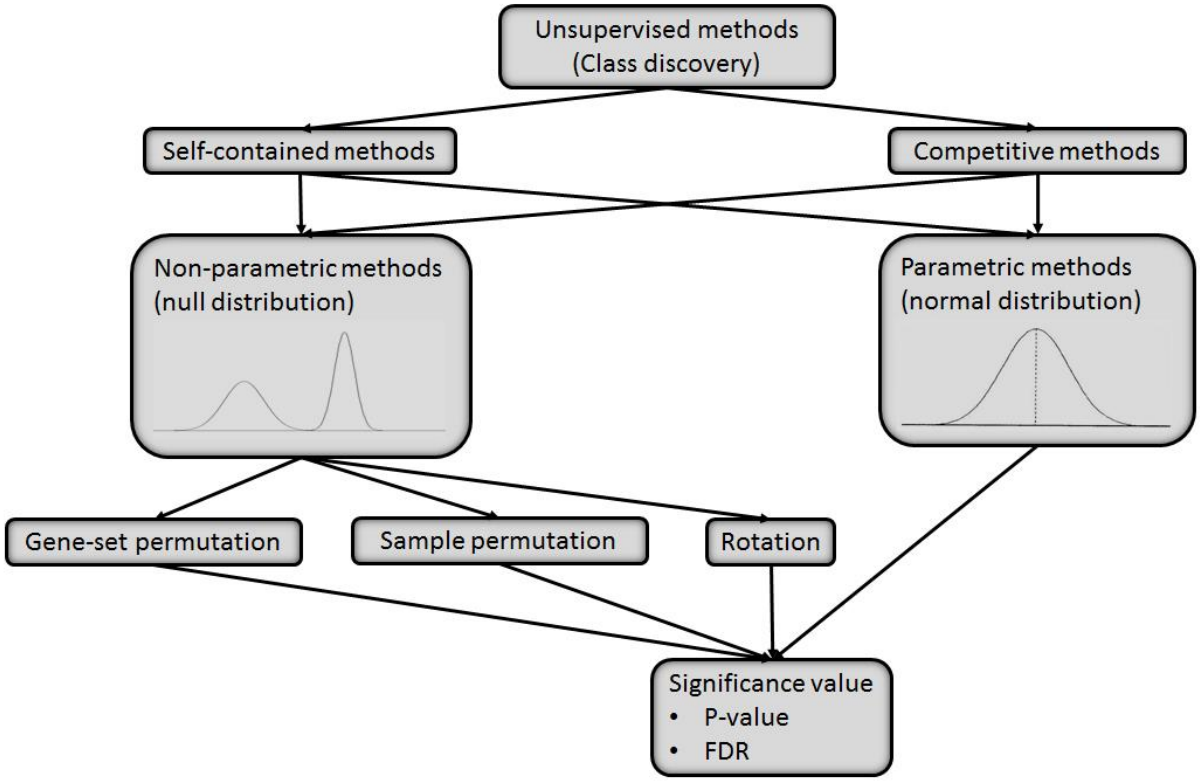

Figure 7. Schematic overview of unsupervised enrichment analysis, focus on significance calculation. In unsupervised methods, significance can be calculated with either non-parametric or parametric methods. Parametric methods assume that gene sets follow predefined distribution whereas non-parametric methods make no prior assumptions of the gene-set distribution. This null distribution is generated from permuting either gene-sets (row-wise randomization) or samples (column-wise randomization) or by rotation. Eventually significance value, traditionally P-value, for each gene-set is generated. Based on the significance value, the null hypothesis can be rejected at cutoff value, typically less than 0.05.

2.6.2.1 Competitive methods

Competitive test compares the differential expression of a gene set to its background set. Competitive tests are more popular than self-contained and there are many available, most common examples being hypergeometric test and GSEA[3].

Hypergeometric test tests statistical significance of successes of random draws (Figure 8.). In gene set enrichment, the question is as competitive null hypothesis.

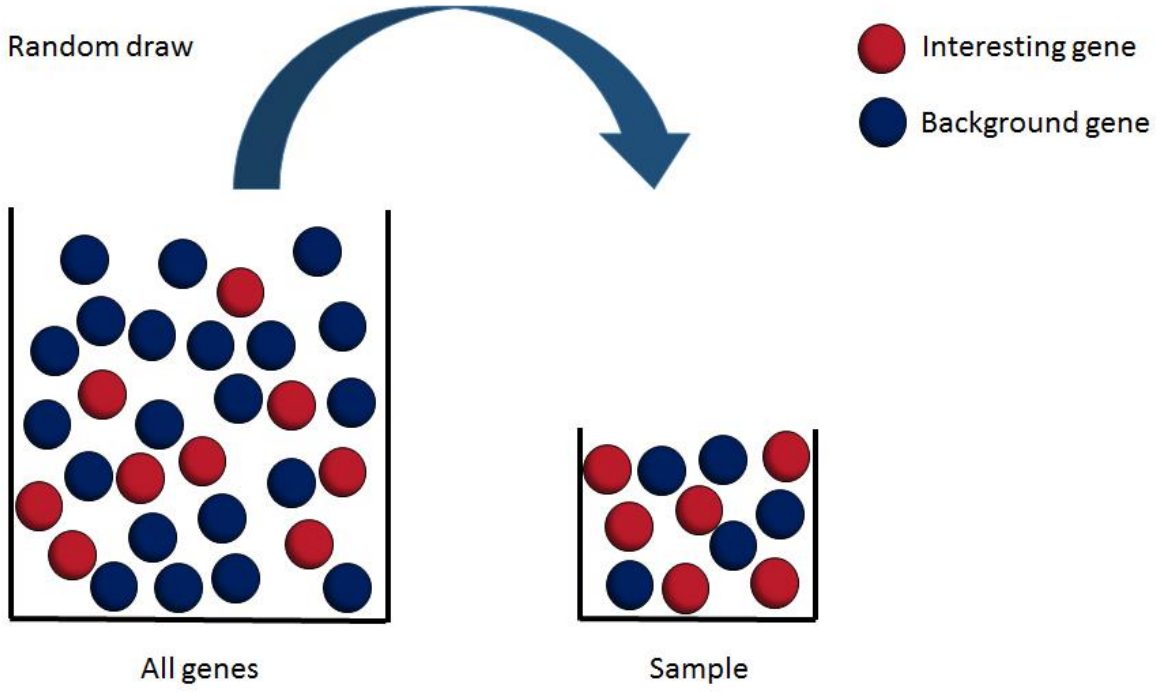

Figure 8. Schematic description of hypergeometric test. Measured genes are separated into two groups, interesting genes and background genes, by chosen cutoff. If the gene set (drawn set) contains more interesting genes than what would be expected from by random draw from all of the genes (background), genes of the gene set are overrepresented and thus gene set is enriched and null hypothesis is rejected.

Other commonly used competitive gene set enrichment method is GSEA which uses its own algorithm. GSEA converts the expression levels into signal-to-noise ratio which is used to rank the genes based on the best distinction between two phenotypes. These phenotypes may be, for example, treated and untreated samples. The ranked list tells how different the two phenotypes are: If the genes of the gene set are found multiple times from the ranked list, top or bottom, the

correlation with the phenotype is high. This is the enrichment score computed by GSEA, calculated by using a weighted Kolmogorov-Smirnov-like (KS) statistic. The algorithm calculates the score by going through the ranked list, increasing running-sum statistic when a gene in gene set is encountered and decreasing statistic when encountering genes that are not in gene set. The statistical significance is estimated by first creating null distribution. The null distribution is generated from permuting the phenotype labels and re-computing the enrichment scores. Typically, 1000 permutations are computed and recorded to obtain the null distribution of enrichment scores. The permutation of class labels is thought to preserve gene-gene correlations and provide biologically reasonable results. The empirical, nominal P-value is calculated relative to the null distribution. It is also possible to correct for multiple hypothesis testing by normalizing the enrichment score for each gene set and comparing the tails of the observed and null distributions of the normalized enrichment score[3].

Traditionally, GSEA is sample randomization, which means samples aka phenotypes are permuted. Other way to randomize is gene sampling. In the former, the phenotype is taken as the sampling unit when calculating P-value whereas in the latter, gene is the sampling unit[51]. Gene sampling methods are more popular than sample randomization ones, mostly due to small minimum sample requirement in practice. Gene sampling methods allow fairly small sample sizes, whereas subject sampling needs a high number of samples to perform properly. With gene sampling, however, one loses the correlations between the genes upon randomization[51]. This is unfortunate because in biological networks, only a handful of genes work alone. Therefore, it is recommended to use subject sampling when possible.

As mentioned before, GSEA is one of the most popular gene set enrichment methods, but there are other similar ones. These include Gene Set Analysis (GSA)[52], Significance Analysis of Function and Enrichment (SAFE)[53] and Gene Set Variation Analysis (GSVA)[54].

For example, instead of KS statistics, in GSA and SAFE, user can choose the settings among the alternatives. In both, different combinations of local (gene-specific) and global (gene-set) are provided. For GSA, the default test statistics in R/Bioconductor package are mean and gene randomization, whereas for SAFE, t-test and Wilcoxon rank sum test are set as default. GSA was used in experimental part of this thesis.

More complex example of gene set enrichment is GSVA, gene set variation analysis. GSVA is a non-parametric and unsupervised analysis. It starts by evaluating whether gene is highly or lowly expressed by using non-parametric kernel estimation. In microarray data, Gaussian kernel

is used, whereas in case of RNA-seq data, discrete Poisson kernel is used. Then, expression level statistics are condensed into gene sets and sample-wise enrichment scores are calculated in order to up-weight the two tails of rank distribution. The enrichment score similar to GSEA's; it is calculated by using KS statistic. Finally, KS statistics are turned to GSVA scores by using either the classical maximum deviation method or normalized enrichment statistic. The choice of the last statistic depends of what is wanted: if the gene sets are explicitly separated into "up" or "down", the normalized GSVA should be used. Sometimes, however, pathways have genes that act strongly in both directions and under these circumstances, usage of maximum deviation is advised[54].

The comparison between different methods is, however, difficult. In competitive methods, it seems that their results have poor overlap, especially when compared with self-contained methods[55].

Furthermore, it is technically impossible to say which method is the best due to the lack of a gold standard. Some have tested variety of tools by with simulated datasets, and that's when mean test outperformed GSEA's KS, for example[56]. However, GSEA seems to outperform other tools when experimental datasets are used. Naturally, experimental dataset is more biologically relevant than simulated, but with the former results are harder to interpret. This is the case because there is no gold standard – it can be almost impossible to tell which gene sets are true positives and which true negatives. On the other hand, simulated datasets cannot substitute for experimental ones due to complexity of biological systems. It can also be argued that rejecting false positives is more important than detecting low true positives. Therefore, both simulated and experimental data should be used when testing the tools[57].

2.6.2.2 Self-contained methods

A self-contained method does not compare the gene set to the background. Unlike competitive methods, self-contained methods don't use any information of the genes outside the gene set. A self-contained test has more power than competitive one, which is due to the hypothesis being more restrictive. In some cases, however, the test may be too powerful: in case of many DEGs, it may call almost all gene sets as significant even if they were not. A self-contained test also doesn't treat gene set differently from a single gene, which is something competitive test takes

into account. Even if the gene set has only one or few genes, self-contained test will call the gene significant if the P-value is just below the defined cut-off value. Moreover, the self-contained test looks all of the genes on the chip. It tests the global hypothesis, meaning that the test could be used, for example, for quality check of the data or as prediction interpretation[51]. Unlike competitive methods, self-contained methods are comparable and similar in performance of size and power when properly standardized[58].

Examples of self-contained methods include Globaltest[59], Analysis of covariance (ANCOVA)[58] and Pathway Level Analysis for Gene Expression (PLAGE)[60]. In addition, variety of other statistical methods, such as KS test, Fisher's test and tail strength can be modified to perform self-contained gene set analysis[61].

Globaltest is based on the idea of having close connection between finding DEGs and predicting clinical outcome. In other words, it tests if clinical status or phenotype (set as 0 and 1) is dependent of gene set. Should the gene expression patterns differ for clinical outcomes, group of genes, gene sets in other words, can be used to predict the outcome. This makes the null hypothesis so that none of the genes are correlated with the phenotype 1. To reject the null hypothesis, the genes in the gene set don't need to have similar expression patterns, only many of the genes need to be correlated with the clinical outcome (phenotype 1).

Mathematically Globaltest is modified generalized linear model which includes linear regression and logistic regression. By estimating regression parameters from the training data, general linear model can be used to predict the phenotype or clinical outcome and eventually compute the correlation with the phenotype. Despite Globaltest being a good self-contained method, it doesn't come without limitations. Possibly the greatest drawbacks occurs if the sample size is small and there are lot of genes in the gene set – the total number of permutations won't be large enough to attain low significance levels[59].

ANCOVA is very similar to Globaltest, but instead of testing if phenotype is dependent of gene expression, it tests if the gene sets with similar phenotypes have similar expression patterns. In other words, the roles of phenotypes and genes are exchanged in regression models[58]. Other examples of self-contained methods are Rotation gene set testing (ROAST)[62] and PLAGE. Both ROAST and PLAGE are able to perform even if the number of samples is relatively low. In ROAST, this ability is based on usage of rotation (multivariate regression) instead of permutation and t-statistics. ROAST's rotation, a Monte Carlo technology for multivariate regression, resembles fractional permutation. Due to this, it doesn't depend on sample size and

thus there is no limit to the number of rotations. Traditionally, 10 000 rotations are used[62]. In permutation on the other hand, a number of randomly picked genes (number based on the number of genes in the gene set and in the background gene set) are used to calculate enrichment score. Traditionally this is repeated at least 1000 times, and the gained values are used as a background for calculating P-values.

In PLAGE, gene expression levels are first standardized to z-scores. Then, gene sets are converted to eigenfactors, "metagenes", by using singular value decomposition. The first eigenvector with the highest eigenvalue, "metagene", in the sample is used to define the activity level in the sample[60].

2.6.2.3 Parametric vs. non-parametric tests

In competitive and self-contained methods, statistical significance can be calculated with parametric or non-parametric methods. Non-parametric methods make no prior assumptions of the distribution of the data whereas in parametric methods, assumption of the distribution is made. The genes in the gene set are expected to follow this, for example normal or bimodal, distribution. In order to have meaningful results, it is crucial that the data follows the presumed distribution.

Non-parametric tests are, in general, hard to compute and thus, time-consuming. Simple parametric methods, such as $\chi 2$ and z-score, have been implemented in order to overcome this challenge. In this case, significance (P-value) can be computed analytically, which makes the computational calculations robust[63]. However, analytical background has been concluded to be less accurate than simulated one. Moreover, parametric methods ignore the gene-gene correlations, which affects the estimation of the significance of gene set enrichment. It has even been suggested that methods like this such be avoided[64].

Examples of parametric methods include Parametric Analysis of Gene set Enrichment (PAGE)[65] and its variant, Generally Applicable Gene set Enrichment (GAGE)[66]. PAGE uses fold change between sample groups to calculate z-score and statistical significance[65]. GAGE, on the other hand, is a variant of PAGE and uses two-sample t-test instead of z-score with assumption that genes come from different distribution[66].

2.6.3 Supervised vs. unsupervised learning methods

The idea between unsupervised and supervised learning methods is fundamentally different. Whereas unsupervised learning searches for unknown biological relevances, supervised learning aims to predict sample classes.

Unsupervised methods are unbiased and allow identification of complex datasets without any prior assumptions. Supervised learning methods, on the other hand, aim is often to build a classifier or a predictor from training data. In supervised methods, samples are labeled to belong to a class whereas in unsupervised method, the differences are looked into without labeling. Naturally, distinction between different sample groups, such as treated and untreated, is often of interest, but this is achieved without labeling. Supervised method could be used, for example, to predict if Pgc-1α is over-expressed in the gene set or not. The same way, it could be used to predict whether hypertrophy is physiological or pathological.

As mentioned, supervised methods need prior information about which samples or genes are grouped together. In terms of prediction of hypertrophy, this would mean variety of samples of both states with knowledge of corresponding hypertrophy states. These samples are used as a training set to build a classifier and therefore it is important to have "correct" classification for at least some of the samples. Due to this, the accuracy of supervised learning method depends heavily on the quality of the training set. Once the classifier has been built, it must be tested with independent test set, such as datasets with known physiological and pathological hypertrophy samples to estimate classification error and later to predict classes in other samples[67,68]. Overview of the method is shown in Figure 9.

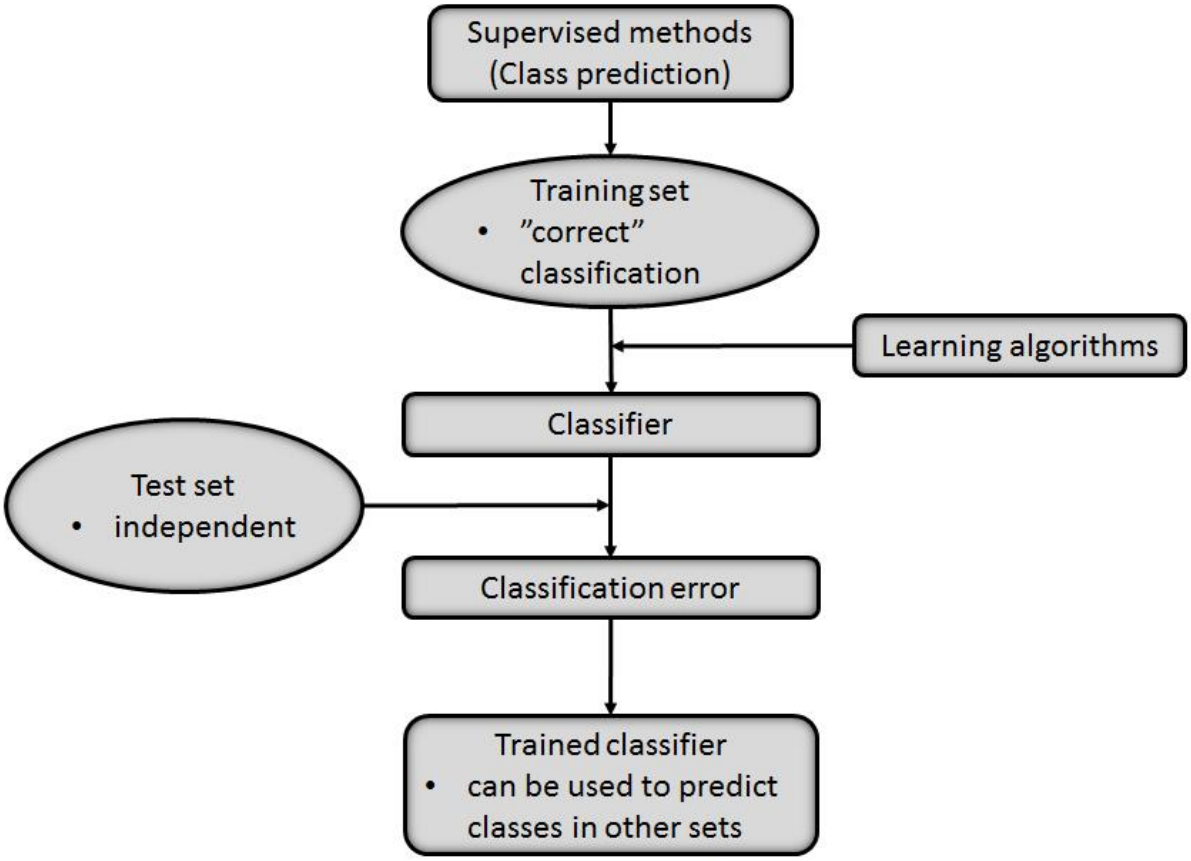

Figure 9. Schematic overview of supervised training method. With learning algorithm, training set with "correct" classification is used to build a classifier. Independent test set is used to test the classifier. Once classifier has been trained, it can be used to predict classes in other sets.

Supervised learning methods have applications in variety of bioinformatics fields. For example in genomics they are used in prediction of splice sites along with identification of motifs and protein coding regions.

Other fields of application include proteomics (prediction of function and secondary structure proteins), systems biology (inference of gene networks and metabolic pathways), microarrays (pre-processing, analysis), evolution studies (phylogenetic trees construction) and primer design[69].

The typical problem in supervised classification is overfitting of the data. This occurs when the model is too complex and has, for example, too many parameters compared to the sample size. In this case the model fits the training data, from which it has been developed, well. It is, however, unable to fit to the test set, resulting in poor predictive power. This problem is

common in gene expression data which traditionally suffers from small sample sizes relative to number of genes. With too many parameters, the model ends up trying to find gene expression levels instead of wanted patterns. This problem can be avoided with dimensionality reduction and cross-validation with test set[67,70].

Whereas unsupervised methods are good starting point of the analysis, supervised methods aim to answer more specific questions (*"Are there enriched pathways in my hypertrophy dataset?"* vs. *"Is the state of hypertrophy in this sample physiological or pathological?"*). Generating the classifier also is more demanding and time-consuming than basic gene set enrichment analysis, but on the other hand, it is capable of answering to specific question of interest. Naturally, generation of a working classifier also requires more data than unsupervised gene set enrichment. In the end, both supervised and unsupervised methods have their pros and cons, and in order to achieve meaningful results, the choice of the method should always be based on the research question and hypothesis.

3 AIMS OF THE STUDY

There are three main questions this thesis aims to answer to:

1) Is the effect of Pgc-1α overexpression on gene expression the same in cardiomyocytes and skeletal muscle?

2) Does Pgc-1α overexpression resemble more physiological than pathological hypertrophy based on gene set enrichment analysis?

3) Could Pgc-1α overexpression be used in treating cardiac hypertrophy?

    3.1) What is the effect on key pathways that are regulated in disease?

    3.2) Are there side effect causing pathways?

The first aim arises from previous studies. It has been indicated that overexpression of Pgc-1α has similar effect in both heart and skeletal muscle[12,71]. Moreover, it has been indicated that the targets of Pgc-1α are the same in both tissues[5,7,11]. According to our hypothesis, this is not the case.

The second and third aim, latter of which is the core of this thesis, are heavily linked together. Should Pgc-1α overexpression resemble pathological rather than physiological hypertrophy and therefore drive for pathological state of cardiomyocyte, it would be dangerous and potentially lethal for the organism. In this case, Pgc-1α overexpression should not be used in treating cardiac hypertrophy. Our hypothesis is that the state caused by Pgc-1α overexpression resembles more physiological than pathological hypertrophy and in that sense, it could be used as a potential treatment.

Upon analyzing the pathways affected by Pgc-1α overexpression, circadian rhythm arose unexpectedly. This significant effect piqued our interest because, as explained in the literature review, circadian rhythm is essential to health and body functions, so heavy disruption of this system could make Pgc-1α overexpression a poor treatment. Thus more datasets were included and further studies were concluded.

4 MATERIALS AND METHODS

4.1 Datasets

4.1.1 Pgc-1α overexpression dataset (Tavi et al.)

Data from four genome-wide experiments were used in this project (Table 1.). The dataset of most interest was the RNA-seq data from Tavi et al. (unpublished) in which the impact of Pgc-1α overexpression on cardiomyocytes was studied. C57HBL/&JolaHsd mice carried MCK-PGC-1α mutation. For RNA-seq, hearts were collected from 16-week-old mice. RNA libraries were prepared according to dUTP protocol, generating paired and single end data.

The dataset from Tavi et al. was computationally compared to three different datasets in order to gain answers to the research questions of this project.

Table 1. Table of datasets used in this thesis. The "name" of the dataset, main contributor of the article and its accession number along with the database the data was extracted from are presented in the table. The used technique, number of replicates and tissue where the samples were extracted from are also displayed. All experiments were performed with mice.

Name	**Accession**	**Contributor**	**Technique**		
			RNA-seq	Microarray	Other
Pgc-1α overexpression	unpublished	Tavi et al.	x		
Skeletal muscle	GSE40439 (GEO)	Pérez-Schindler et al.[72]		x	
Exercise	ERA037989 (DNAnexus)	Song et al.[73]	x		
Circadian 1	GSE43073 (GEO)	Young et al.[74]		x	
Circadian 2		Wu et al.[75]			x

4.1.2 Skeletal muscle dataset (Pérez-Schindler et al.)

One of the datasets used in this thesis analysis' was skeletal muscle dataset from Pérez-Schindler et al[72]. In the original study, comparison between nuclear receptor corepressor 1 (NCoR1) muscle specific knockout mice were compared with Pgc-1α overexpression mice. *NCoR1*$^{IoxP/IoxP}$ specific mice had been generated like in previous study[76], and to create NCoR1 MKO mice, they were crossed with HSA-Cre transgenic animals. Pgc-1α muscle-specific transgenic (mTg) mice were generated like in previous studies[77]. The mice performed exercise by running treadmill for two days with variety of inclines and time. RNA was isolated from multiple organs, and microarray was performed with GeneChip Gene 1.0 ST Array System (Affymetrix) by using the RNA isolated from gastrocnemius.

In this thesis, the Pgc-1α overexpression data from Pérez-Schindler et al. was analyzed and compared to Pasi et al. Pgc-1α overexpression data in order to see whether and how the effect of Pgc-1α overexpression varies between cardiomyocyte and skeletal muscle.

4.1.3 Exercise dataset (Song et al.)

In order to find out whether Pgc-1α resembles more physiological or pathological cardiac hypertrophy, Pgc-1α overexpression dataset was compared with microarray exercise data from Song et al[73]. Song et al. studied physiological and pathological hypertrophy by using RNA-seq. C57BL/6J mice were purchased and used in the study. Cardiac hypertrophy was induced to pathological mice group and its control group with intraperitoneal injection, as described in another study[78]. Physiological mice group swam for 4 weeks as described in their previous study[79]. cDNA libraries were prepared according to instructions from sample preparation kit (Illumina, San Diego, CA).

4.1.4 Circadian rhythm datasets (Young et al. + Wu et al.)

To gain better understanding of the effect of Pgc-1α overexpression in circadian rhythm, the data from Tavi et al. was compared to circadian rhythm dataset from Young et al[74]. This is referred as circadian dataset 1. Young et al. studied the effect on Bmal1 on circadian rhythm. As mentioned in the literature review, Bmal1 is one of the core clock components and known target of Pgc-1α. Bmal1 knockout mice, from C57B1/6J and wild-types from FVB/N background, were enforced into strict 12-hour light/12-hour dark cycle (Zeitgeber time 0). Microarray analysis were performed with Ref-8 BeadChips and the BeadStation System (Illumina, Inc., San Diego, CA) to ventricular tissue collected for every three hours for 24 hour period. The data was computationally analyzed and compared with the data from Tavi et al.

The results from circadian rhythm dataset from Wu et al[75] were also compared to Pgc-1α overexpression in order to identify the timepoint most affected by overexpression. The mutant mice (*PER2*^{s662G}, *PER2*^{s662D}, Pgc1α transgenic and overtime) were from C57BL/6J background, and were enforced to 12-hour light/12-hour dark cycle. Gastrocnemius and heart muscles were collected every 4 hours for 24 hours.

Wu et al. identified seven circadian clock genes affected by Pgc-1α overexpression. The mice strain used in the study was the same as in experiment from Tavi et al. Due to this, experiment results from Wu et al. were of interest, despite the study being real-time qPCR with statistical analysis of ANOVA and Student t-test. This is referred as circadian dataset 2.

4.2. Computational pre-processing of RNA-seq and microarray in differential expression analysis

Computational part of both RNA-seq and microarray analysis starts after gaining millions of shorts reads from sequencing or after gaining the raw probe-level expression data from the array image. Different tools are used in pre-processing of RNA-seq and microarray, but the end results are the same. In order to gain differentially expressed genes or more accurately, transcripts from the statistical analysis, expression levels of the genes need to be calculated. The basic overview of both RNA-seq and microarray pipelines is depicted in Figure 10.

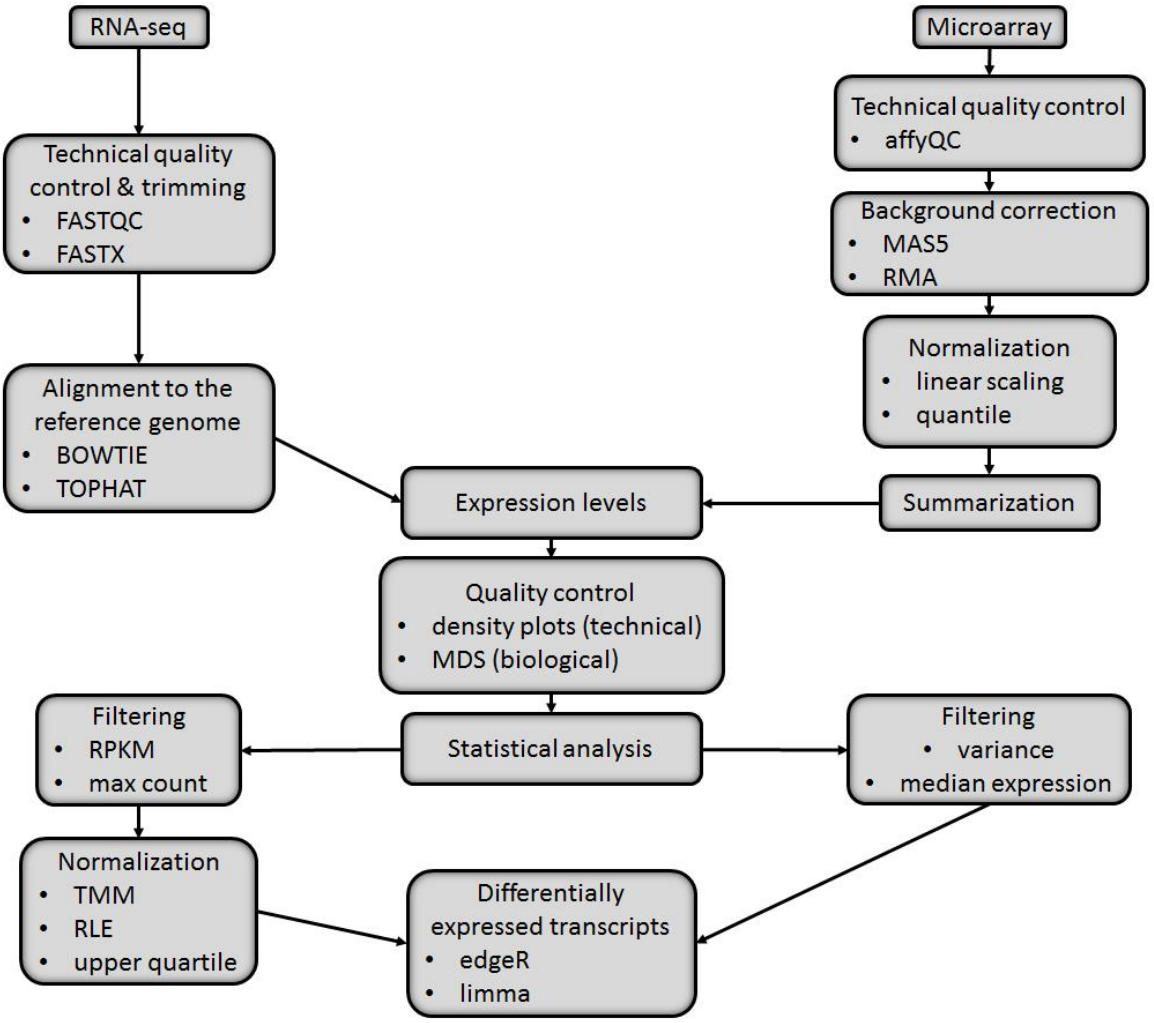

Figure 10. Schematic overview of RNA-seq and Affymetrix microarray analysis pipelines. In both, the raw data is first transformed to expression levels with respective quality control, both technical and biological, and is followed by statistical analysis, leading to identification of differentially expressed transcripts.

With RNA-seq data, the first step is to check the quality of the raw data by using FASTQC[80], a quality control tool for high-throughput data. FASTQC provides visual output of the quality which can be used to determine whether the reads require trimming or not. The low quality base reads should be filtered away by trimming because they may cause otherwise mappable sequence to fail aligning to the reference genome. The optional trimming can be executed with tools such as FASTX[81]. Despite the popularity of RNA-seq and read trimming, there are no specific guidelines for how strict trimming should be performed and thus, it is up to the researcher to determine the requirements.

After optional trimming, RNA-seq reads are aligned to the reference genome. In other words, unique location where a short read is identical to the reference is found. One of the most used programs is TopHat[82], which aligns reads to the genome and discovers transcript splice sites. TopHat uses program called Bowtie[83] for alignment and breaks up the reads Bowtie is unable to align to smaller pieces since often these pieces, when mapped separately, can be aligned to the genome. TopHat also estimates the junction splice sites, allowing the discovery of alternative splicing sites. The aligned reads can tell many things about the sample: mismatches, insertions and deletions can be used to identify polymorphisms whereas reads that align outside annotated genes may be evidence of new protein-coding genes and non-coding RNAs.

After discovering transcript splice sites, Cufflinks[84] can be used to map this against the reference genome to find transcripts. Cufflinks assembles individual transcripts that have been aligned to the genome and quantifies expression levels of each full-length transcript.

Another tool used for quantification of the reads is HOMER[85], which has two alternative programs to quantify the RNA reads in the genome. They count the reads in regions and produce gene expression matrix. There are also variety of options available in the tool. One can, for example, count exons instead of genes. After calculating the expression matrix, quality control of sample levels can be performed for both technical and biological variation. Neither TopHat nor Homer, however, produce differential expression matrix, and thus such statistics must be calculated with other programs.

Two RNA-seq datasets were used in this thesis: the Pgc-1α expression and exercise datasets. Pgc-1α overexpression dataset had been analyzed before the start of this thesis. The quality of the raw reads from both datasets was confirmed using FASQC and NGSQC Toolkit software[86]. Bases with poor quality scores were trimmed with FASTX toolkit; both datasets were required to have minimum of 96 % (exercise dataset) or 97 % (Pgc-1α overexpression) of all bases in one read to have minimum quality score of 10. The reads also had to be at least 25 of length. The Tophat software (version 2.0.9) was used for alignment, allowing up to 3 mismatches, 1 valid alignments and with minimum filtering score of 2.

Similarly to RNA-seq, the arrays can also be quality controlled and outliers may be removed. Pre-processing of microarray experiment starts from background correction, which is performed to reduce the background noise caused by laser reflection on the surface. The background correction isn’t compulsory, and sometimes background detection hasn’t been executed for one reason or the other, but it is highly recommended. These corrected values are

normalized to improve the sensitivity to detect genes. Finally, data is summarized. The summarization combines preprocessed probes and computes expression value for each probe set on the array. Again, quality controls of sample levels may be performed before the statistical analysis.

In this thesis, there were two microarray datasets, circadian dataset 1 and skeletal muscle dataset. The former experiment had been performed with Illumina microarray chip, and was processed with R/Bioconductor and the latter research used Affymetrix chips and the data was thus processed with Affymetrix power tools.

In circadian dataset 1, there were no control probes, so background couldn't be detected.

The skeletal muscle dataset had been processed before the start of this thesis. For the Affymetrix chip, the quality of the probes was tested with R/Bioconductor after the full quantile normalization with Affymetrix power tools. The dabg quantification was performed before statistical analysis with edgeR package on R/Bioconductor. All the R/Bioconductor packages used in this thesis are in table 2.

Table 2. The R/Bioconductor packages used in this thesis with short description.

R. package	**Description**
AnnotationDbi	Annotation of data packages
biomaRt	Retrieval of large amounts of data from databases
edgeR	Differential expression and statistical analysis of RNA-seq
gplots	Programming tools for plotting data
hom.Hs.inp.db	Homology information for human
hom.Mm.inp.db	Homology information for mouse
limma	Data analysis, linear models and differential expression for microarray data
lumi	Illumina microarray data analysis
lumiMouseAll.db	Illumina Mouse expression annotation data
lumiMouseIDMapping	Mapping information between Illumina IDs Mouse chips, nuIDs and RefseqIDs for Illumina Mouse chips
org.Hs.eg.db	Genome-wide annotation for human
org.Mm.eg.db	Genome-wide annotation for mouse
piano	Gene set analysis using various statistical methods
RColorBrewer	Color schemes for graphics
snow	Parallel computations
snowfall	Easier development of parallel R programs (based on snow)
VennDiagram	High-resolution Venn and Euler plots with extensive customization of the plot

4.3 Statistical analysis of RNA-seq and microarray in differential expression analysis

4.3.1 Statistical analysis of RNA-seq

Often the interest in biological questions lies in comparing two or more groups and thus differentially expressed transcripts/genes are what researcher is interested in. Unfortunately, programs such as HOMER don't produce these results, and therefore it is necessary to use other tools.

In RNA-seq, it is possible to filter out the genes that are not expressed in the samples. The filtering can be executed in multiple ways, one of which is rpkm (reads per kilobase per million) value. Genes can also be filtered based on how low is the maximal count in the count matrix. If the rpkm value and/or max count is low, it can be argumented that the gene is not expressed and should be filtered out.

In both RNA-seq and microarray, before the detection of DEGs the data is normalized. The normalization enables comparisons between and within samples, and is essential for differential expression analysis. After normalization, read counts are converted to log-counts per millions. Then, differential expression can be calculated.

RNA-seq data gives discrete measurement for each gene (counts) and thus doesn't follow normal distribution. In the RNA-seq, the Poisson distribution forms the general bases in modeling the count data. Poisson distribution captures the technical variability, but biological variability less effectively because Poisson distribution expects all genes to have the same variance. In reality, some genes may fluctuate more or less than the others. This makes the Poisson-based biological analyses prone to high false positive rates[37]. One solution to this is to use negative binomial distribution, which is extension model to the Poisson. It takes the aforementioned biological variability - greater observed variation than the mean - into account by using gamma distribution. Negative binomial model is also known as gamma-Poisson model. This negative binomial approach is implemented in R/Bioconductor package edgeR, and was used in this thesis.

The last approach, also used in this thesis, is to convert raw RNA-seq counts straight into log-counts per million with associated precision weights. After this, the RNA-seq data can be

analyzed as microarray data[87]. This approach is implemented in R/Bioconductor package limma.

Of RNA-seq datasets, the Pgc-1α overexpression dataset had been processed before the start of this thesis with R/Bioconductor with package edgeR. Transcripts with rpkm > 1 in at least three samples and with at least 50 read within the quantified region in any sample were used in statistical analysis.

In the exercise dataset used in this thesis, statistics were calculated with packages edgeR and limma. Limma was used instead of only edgeR because it is designed for complex experiments and variety of experimental conditions. Thus, it was of interest to test and implement the package.

With exercise dataset, transcripts expressed at the level of rpkm > 1 in at least one of the sample groups with at least 10 of the reads within the quantified region were used for statistical analysis. The normalization was calculated with trimmed mean of M-values (TMM), the voom transformation from limma package was applied to the read counts and eBayes function from limma package was used to calculate statistics. Contrast matrix was used as design matrix for differential expression calculation in order to make pairwise comparisons between the groups. Transcripts with adjusted p-value < 0.05 were defined as differentially expressed.

In order to use GSA as gene enrichment method, statistical analysis for the Pgc-1α overexpression dataset was performed using limma with same parameters as for the exercise dataset. It was crucial to redo the statistical analysis with limma for the gene enrichment analysis because with the chosen settings for GSA, t-statistics were required. T-statistics can only be calculated from data following normal distribution, and as explained above, edgeR's function exactTest expects data to follow negative binomial distribution.

4.3.2 Statistical analysis of microarray

Despite their similarity, RNA-seq and microarray are methodologically different. This shows in differential expression calculations. Whereas RNA-seq data gives discrete measurement for each gene (counts), microarray intensities have a continuous distribution (color intensity). In other words, microarray results follow normal distribution whereas the case with RNA-seq isn't as straightforward.

In microarray analysis, the expression values are transformed to log-scores to make the fold-change values symmetric. Data is then normalized to minimize technical bias of the data. Lastly, linear model (for each gene) is fitted to the data and fold changes and standard errors are calculated.

The two traditional 3' microarray datasets, one Affymetrix and one Illumina microarray, used in this thesis were analyzed with R/Bioconductor. In Illumina microarray, the data was transformed to log-scores and normalized with robust spline normalization (rsn) with R/Bioconductor package lumi. For Affymetrix microarray, the normalization was performed with robust multi-array average (rma), which consists of background detection, quantile normalization and summarization. For both datasets, the differential expression statistics were calculated by fitting the least squares linear model and using Bayes statistics for differential expression calculation with R/Bioconductor limma package.

4.4 Computational gene enrichment analysis

In order to understand the biological and metabolic changes caused by Pgc-1α overexpression, gene enrichment analyses were performed. Most of the pathway databases use canonical pathways for human and therefore mouse gene Refseq IDs were converted into human gene symbols before the analysis. This conversion was achieved with function inpIDMapper from R/Bioconductor package AnnotationDbi. The pathways used in the analysis included all pathways from Biocarta[52], KEGG[54], Pathway Commons[88], Gene Ontology (GO)[89], Wikipathways[50], MSig[3] and Reactome[49] databases, downloaded from their respective databases.

After testing a variety of gene set enrichment methods, it was decided to use R/Bioconductor package piano and its function runGSA. RunGSA was chosen because programmer can control variety of parameters and it's relatively easy to use. GSEA was chosen as test for statistical enrichment for its popularity and good performance. The chosen method for gene significance assessment was gene sampling due to high number of genes and low number of samples. In order to gain directional p-values, t-statistics were used in calculating enrichment score with false discovery rate (FDR) adjusting method and 1000 permutations.

4.4.1 Description of algorithms in GSA

GSA[52] is a competitive gene set enrichment analysis, so it can be used to answer the following question: *"Are there enriched pathways in my dataset?"* This creates the corresponding null hypothesis: *"H0. The genes in the gene set of interest are at most as often differentially expressed as the genes in the background gene set."* If the null hypothesis is rejected, there are enriched gene sets (pathways) in the dataset.

GSA uses P-values, t-values or F-values, depending of the enrichment method of choice. When using algorithm of GSEA, method of choice in this thesis, t-values are the only option in GSA. In most of the cases, the t-statistic has already been calculated beforehand in order to find differentially expressed genes. This is also the case in this thesis: by the time enrichment analysis was performed, the statistics for the datasets, including t-statistics, had already been calculated with R/Bioconductor (Table 3). The list of genes in the gene set, pathway, had also been extracted from the database(s) (Table 4). The following tables don't represent results of this thesis and are extremely simplified. In practice, lists of DEGs and gene sets are far larger than the ones shown here. For the sake of an example, however, they are kept small.

Table 3. Output from statistics. The imaginary table of output from statistical analysis for gene set enrichment with gene symbols and their respective t-statistics.

Gene symbol	**t-statistic**
HIF1A	-6.03
FBXL3	-3.85
CRY2	-5.59
CRY1	-5.09
HCCS	0.69
DBP	-4.80
PER2	-6.20
CLOCK	-8.80
PPARA	2.30

Table 4. Table of the genes in the pathway. The imaginary list of the genes in the pathway of interest. Pathways and their respective genes can be extracted from pathway databases such as KEGG and Biocarta.

Custom pathway gene IDs	
PPARA	HIF1A
CSNK1E	PER2
FBXL3	CLOCK
CRY1	NAMP
CRY2	

First, enrichment score is calculated for the dataset. In GSA, there are variety of enrichment statistic choices available, but as mentioned, GSEA[3,90] is the one used in this thesis. Therefore, it's the one presented here.

GSEA starts by sorting the genes based on their values in descending order (Table 5.).

Table 5. Sorted table of statistical output. The genes along with their statistical values are sorted from highest to the lowest.

Gene symbol	**t-statistic**
PPARA	2.30
HCCS	0.69
FBXL3	-3.85
DBP	-4.80
CRY1	-5.09
CRY2	-5.59
HIF1A	-6.03
PER2	-6.20
CLOCK	-8.80

Next, the enrichment score is calculated with weighted KS statistic, which can be formalized as:

$$P_{hit}(S,i) = \sum_{\substack{g_i \in S \\ j \le i}} \frac{|r_j|^P}{N_R}, where\ N_R = \sum_{g_j \in S} |r_j|^P$$

$$P_{miss}(S,i) = \sum_{\substack{g_i \in S \\ j \le i}} \frac{1}{(N - N_H)}$$

with P_{hit} denoting score of the gene that is on the ranked list and in the gene set.
S denotes gene set and i denotes index of the gene in the ranked list.
r_j denotes value of the gene of the ranked list and N_R is the sum of values of the genes that are in the ranked list and in the gene set.
P_{miss} denotes score of the gene that is on the ranked list but not in the gene set. N denotes the number of genes in ranked list, and N_H denotes number of genes from the ranked list that are in the gene set as well.

Originally, GSEA uses P-values instead of t-statistics to calculate directional gene set enrichment scores. However, in order to calculate directional gene set enrichment scores with combination of GSA and GSEA, t-statistics must be used. Therefore absolute values are used in calculation of the enrichment statistic.

7 genes of ranked list of the example belong in the gene set, so sum of absolute values of the genes (N_R) is calculated as follows:

N_R = 2.3 + |-3.85| + |-5.09| + |-5.59| + |-6.03| + |-6.20| + |-8.80| = 37.86

The enrichment score is calculated by walking down the ranked list, and when gene that belongs to the gene set (pathway) is encountered, the score is increased.

The first gene of the ranked list is *PPARA*, and that is also in the gene set (pathway). The absolute value of the gene is divided by sum of absolute values of the genes. The first step of the random walk, running sum, calculated.

$P_{hit}(S,i)$ = 2.3 / 37.86 = 0.0608

On the contrary, if gene that does not belong to the gene set is encountered, the score is decreased.

The second gene of the ranked list is *HCCS*, which is not in the gene set. Thus, the score for missing genes, P_{miss}, is calculated by dividing 1 with the number of genes of the ranked list that are not in the gene set (*HCCS*, *DPB*). The calculated score is then decreased from the running sum.

$P_{miss}(S,i) = 1 / (9-7) = 0.5$

0.06075013 - 0.5 = -0.4392

The next gene of the ranked list, *FBXL3*, is in the gene set, thus its P_{hit} value is calculated and added to the enrichment score. The fourth gene of the ranked list, *DBP*, isn't in the gene set and thus P_{miss} value is subtracted from the current score. This is continued until the end of the ranked list (Table 6).

Table 6. Weighted running sums. Running sums after each step of the calculation, corresponding to the gene at hand.

Gene symbol	**Running sum**
PPARA	0.0608
HCCS	-0.4392
FBXL3	-0.3376
DBP	-0.8376
CRY1	-0.7031
CRY2	-0.5555
HIF1A	-0.3962
PER2	-0.2324
CLOCK	5.5511E-17

The enrichment score is the maximum deviation from the random walk statistic, corresponding to weighted Kolmogorov-Smirnov statistic. In the case of example data, as seen from table 6, the value is: ES = -0.8376. The schematic overview of the KS statistic is represented in Figure 11.

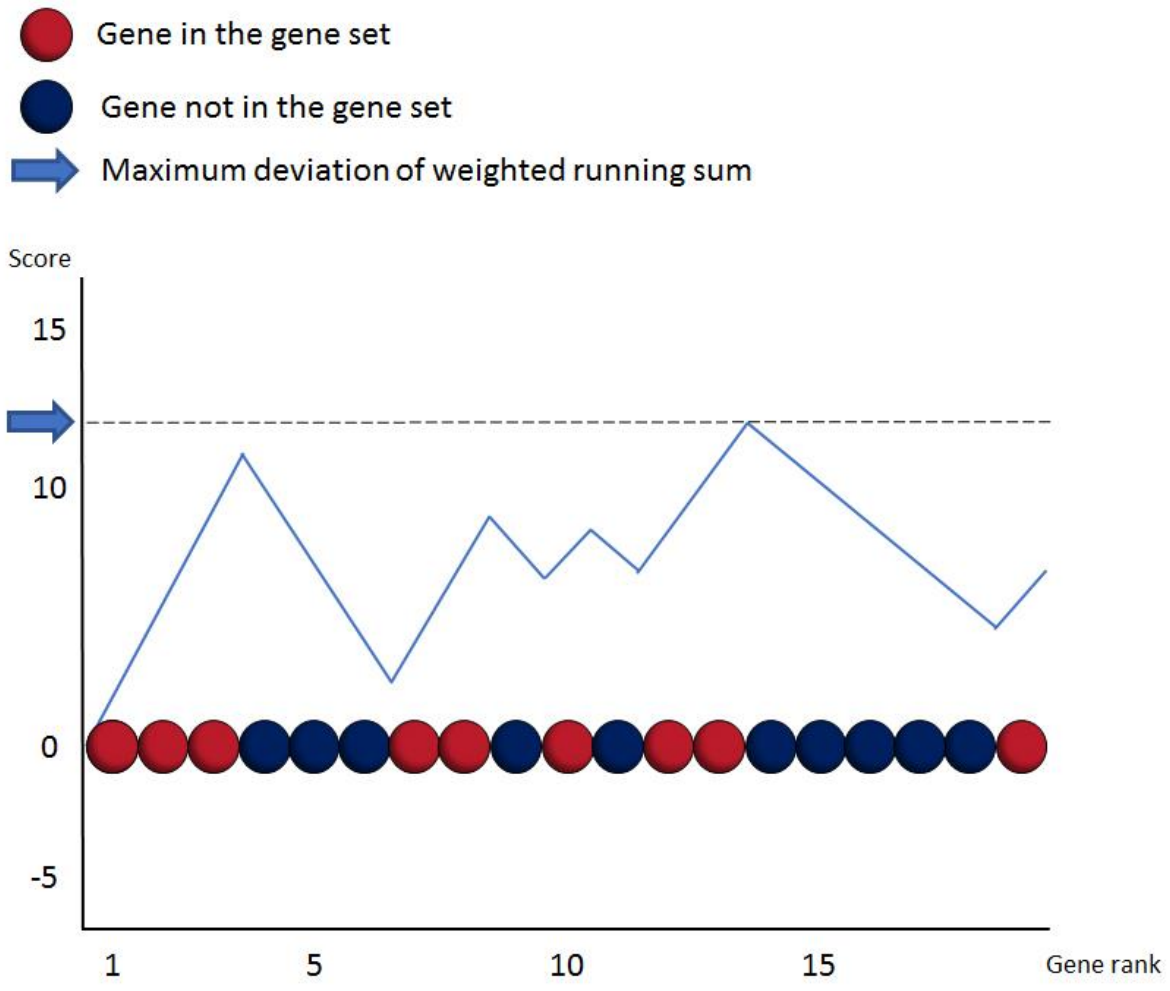

Figure 11. The schematic overview of Kolmogorov-Smirnov statistic. The algorithm calculates score by walking through the ranked list, increasing running-sum statistic if the encountered gene belongs to the gene set and decreasing if not. The genes with higher rank are weighted to increase the running sum more than low rank genes. The maximum deviation is used as enrichment score. If the maximum deviation is unusually high, gene set is often enriched.

After calculating the enrichment score, GSEA's algorithm starts permuting values. In terms of this example and thesis, gene permutation is used. If there were enough samples, sample permutation would also be an option.

GSEA randomly takes the number of the genes that are both in the ranked list and in the gene set from the ranked list for permutation.

In this example case, there are total of 9 genes in our ranked list. 7 of these are in the gene set. So, seven genes from the list of ranked genes (total of 9) are taken to the permutation. For example, in the first round, the genes can be *HCCS*, *CRY2*, *FBXL3*, *DBP*, *PPARA*, *HIF1A*. And on the next round, 7/9 genes are again randomly picked. For example, the genes can, in this time, be: *DBP*, *HCCS*, *PER2*, *CRY1*, *PPARA*, *FBXL3*, *CLOCK*.

Next, enrichment score is calculated for the randomly picked genes like before but without ranking them first. In this light, the ES for the first is 1.0, for second 0.5273 and so on. For the sake of example, the values for 5 permutations are shown in table 7. It is important to note that 5 is extremely low permutation number. Traditionally 1000 permutations are used and GSA function in R requires minimum of 100 permutations.

Table 7. The enrichment scores of permuted values.

Permutation values
1.0000
0.5273
-0.5000
0.7172
0.5000

The permutation values are used as a background to calculate statistical significance (P-value). They are compared to the enrichment score. This differs slightly depending of the direction of the ES. If ES is negative, negative values are used whereas if it's positive, positive values are used. The number of permutations equal to or below/above ES from the dataset are divided by the number of all negative/positive permutations.

As calculated before (Table 6.) the ES score of the dataset is -0.8376. There is only one negative permutation value, -0.5. This value isn't lower than ES score. Thus: 0 / 1= 0.00.

For the sake of an example, let's pretend that the ES from the dataset was 0.6. If the ES is positive, the number of permutation values above or equal to it are divided by the number of all

positive permutations. In our example, we have 4 positive permutations values, two of which are higher than 0.6.

In this case, the calculation would be as follows: 2 / 4 = 0.50.

Finally, the P-values can be adjusted with the method of choice. In the case of this thesis, the chosen adjusting method was false discovery rate, FDR[91].

$$P_i(\frac{N}{i}) \leq q$$

with P_i denoting adjusted P-value.
N denotes the number of P-values whereas i is assigned rank (ordinal number) of P-value.
q denotes the indicated threshold of FDR.

To calculate FDR, one must have more than one gene set. For the sake of example, let's pretend that we had three gene sets instead of one, P-values being the following: "Custom Pathway": 0.00 ; "Pathway 2": 0.50 ; "Pathway 3": 0.30.

FDR is calculated by ranking the P-values in decreasing order. The smallest gets rank 1, second 2 and largest rank N. Then, each P-value is multiplied by N and divided by its assigned rank to gain adjusted P-values.

In our example, the decreasing order is: Custom Pathway, Pathway 3 and Pathway 2. Their adjusted P-values are:
Custom Pathway: 0.00*3 / 1= 0.00 ; Pathway 3: 0.3*3 / 2 = 0.45 ; Pathway 2: 0.5*3 / 3 = 0.5.

Finally, the cut-off is set to define which gene sets are enriched and which not. Usually, this cut-off is FDR < 0.05. In this example, Pathway 3 and Pathway 2 are above the threshold (0.45 > 0.05 ; 0.5 > 0.05) whereas Custom Pathway is below the threshold (0.00 < 0.05). If the adjusted P-value is below the threshold, gene set is defined as enriched and null hypothesis (*"H0. The genes in the gene set of interest are at most as often differentially expressed as the genes in the background gene set."*) is rejected. In this case, Custom Pathway is significantly enriched in the dataset whereas two others are not. The ES of Custom Pathway was -0.8376, sign denoting the direction of regulation; in this case, the Custom Pathway is negatively regulated.

5 RESULTS

5.1 Pgc-1α overexpression in heart

The main results of this thesis focus on the RNA-seq of Pgc-1α overexpression in heart (Tavi et al., unpublished) and its comparison to physiological/pathological hypertrophy states. Different comparisons were performed to determine the possibility of Pgc-1α overexpression as a treatment target for cardiac hypertrophy. Moreover, the comparison to skeletal muscle dataset was also performed in order to elucidate possible tissue-specific effects in the function of Pgc-1α.

Towards this end, the Pgc-1α overexpression dataset was processed and analyzed as outlined next. First the RNA-seq reads were trimmed, aligned and processed as explained in the Materials and Methods -section, resulting in millions of reads used in the analysis. Next, quality control was performed. One of the three replicates of the dataset clustered poorly in comparison with the other two, implying possibility of an outlier (Figure S1A.). However, with minimum number of replicates (N = 3), it is impossible to be certain. Removing the possible outlier would've also reduced the number of replicates to two, resulting in drastic diminution of the statistical analyses. Therefore all three replicates were retained and the statistical analysis resulted in 618 differentially expressed, unique gene symbols (adj. P-val. < 0.05).

In order to discover significantly enriched pathways and identify their directionality under Pgc-1α overexpression in heart, GSA's gene set enrichment analysis with GSEA's algorithm was performed (see Methods). In the analysis, pathways with adjusted P-value < 0.05 were defined as significant.

A total of 75 pathways were significantly enriched, 62 of which were up- and 14 downregulated. (Table S1., Table 8.) These upregulated pathways include glycolysis, beta oxidation and electron transport chain whereas further significant, interesting downregulated pathways include muscle contraction, PPAR, phosphatidylinositol 3-kinase (PI3K) and calcium signaling (Table 8.). Unexpectedly, two significantly downregulated pathways were involved in circadian rhythm.

Table 8. Downregulated, enriched pathways of Pgc-1α overexpression in heart. Enriched, downregulated pathways of Pgc-1α overexpression in heart with their respective adjusted P-values. The pathways of most interest are colored.

Pathway	**Adj. P-value**
Akt pathway (BIOCARTA)	0.0118
EGFR interacts with phospholipase C-gamma (PWC)	0.0176
Diurnally regulated genes with circadian orthologs (WIKIPW)	0.0176
Signaling by EGFR (PWC)	0.0221
Ppara pathway (BIOCARTA)	0.0248
Adipogenesis (WIKIPW)	0.0285
PI3K events in ErbB4 signaling (REACTOME)	0.0304
PI3K AKT activation (REACTOME)	0.0308
PI3K events in Erbb2 signaling (REACTOME)	0.0325
Striated muscle contraction (WIKIPW)	0.0364
Bmal1 Clock Npas2 activates circadian expression (REACTOME)	0.0385
Ca-dependent events (PWC)	0.0419
Raccycd pathway (BIOCARTA)	0.0428
CaM pathway (PWC)	0.0471

In order to further study the effect of Pgc-1α overexpression in heart, ten pathways of interest were chosen based on the known biology (Table S2.). Different pathway databases have different genes associated per pathway, so in order to optimally reflect the biology, pathways were custom curated across pathway databases (Table S3.). The percentage of significantly regulated genes among detected genes was summarized for each pathway (Figure 12.).

Many of the pre-defined curated pathways contain genes that were both up- and down-regulated, although most of the genes didn't appear to be regulated. Of these predetermined pathways, the circadian rhythm had the highest percentage of regulated genes (22.6 %) and down-regulated genes (21.0 %).

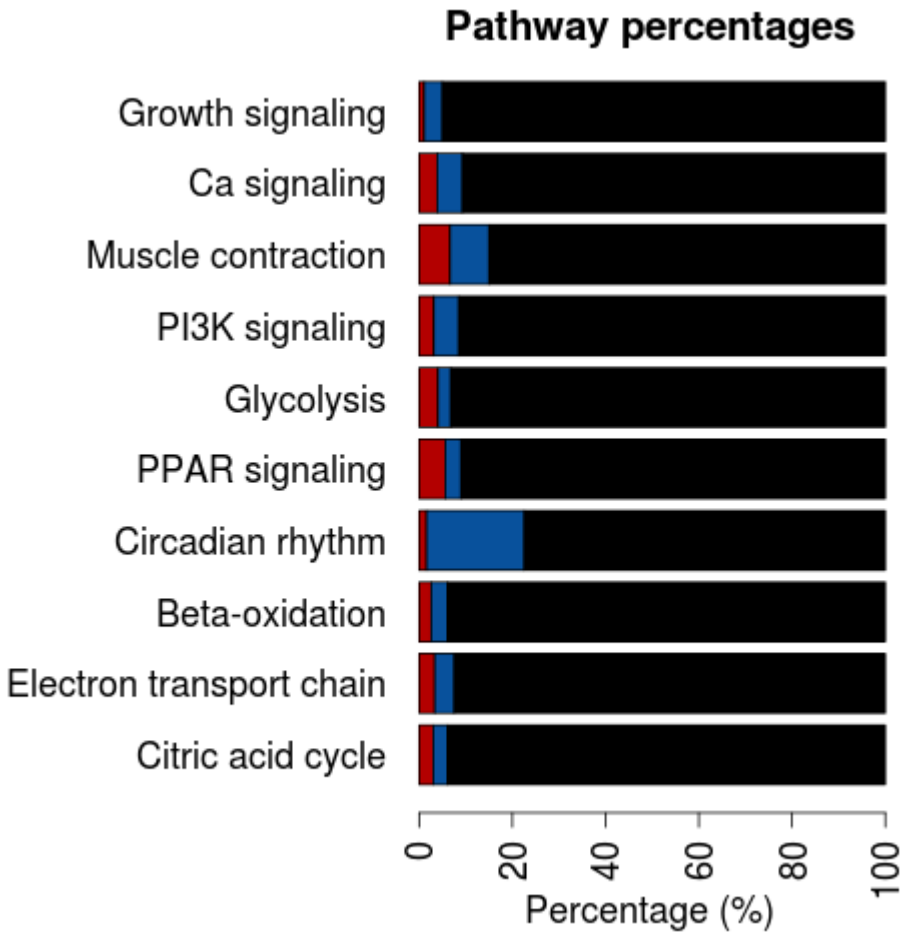

Figure 12. The extent of differential expression in pathways affected by Pgc-1α overexpression in heart. Selected pathways curated from KEGG, Wikipathways, GO, Biocarta, Pathway Commons, MSig and Reactome databases (Table S3.) are shown as barplots. In each plot, the proportion of differentially expressed up- (in red) and downregulated (in blue) genes is compared to non-regulated genes (in black).

Heatmaps of differentially expressed genes were produced for each curated pathway. In the curated circadian rhythm pathway, there were 10 DEGs, 9 of which were downregulated, further implying downregulation of the pathway (Fig. 13). This, along with gene set enrichment results, piqued our interest and led to further investigation of the pathway. For other pathways, however, it was difficult to make reliable assumption in terms of regulation due to uniformly distributed genes in both up- and downregulation.

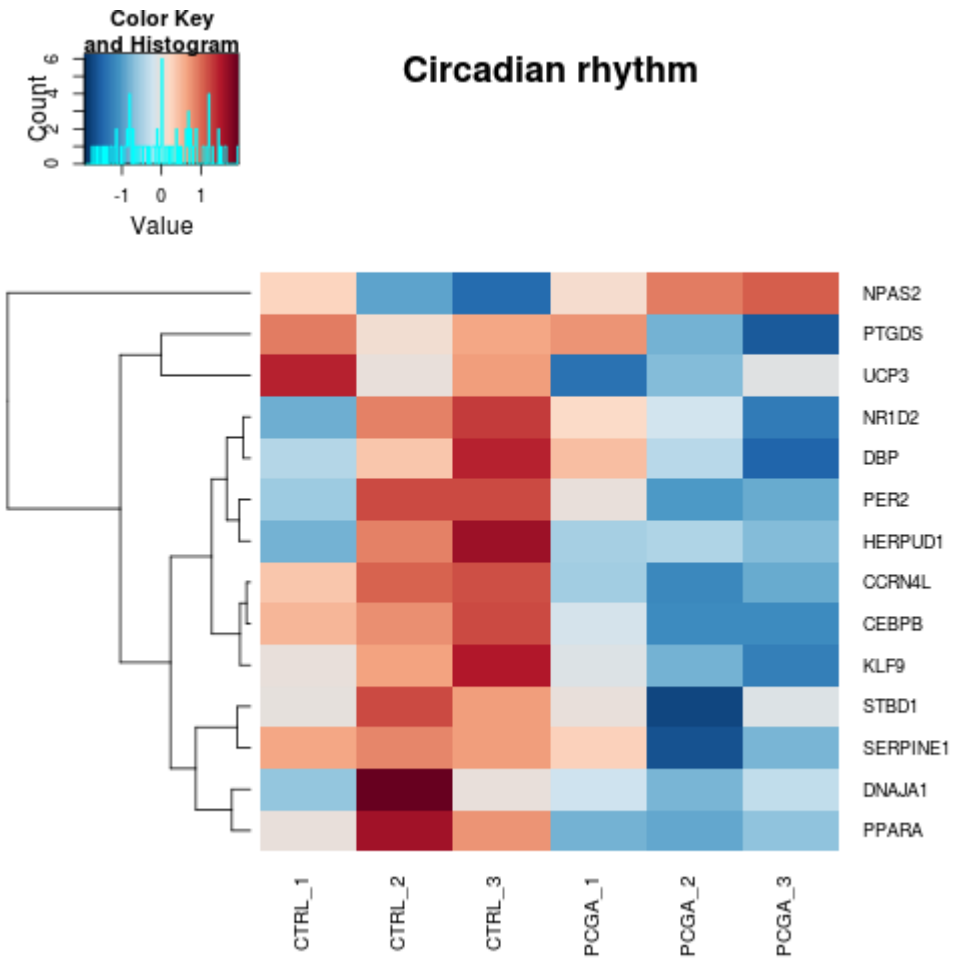

Figure 13. Expression levels of differentially expressed genes in Pgc-1α overexpression in curated circadian rhythm pathway. Hierarchical clustering of scaled expression levels of the differentially expressed genes (adj. P-val. < 0.05) in Pgc-1α overexpression in the heart in the curated circadian rhythm pathway shown as heatmap. Rows correspond to genes and columns to samples whereas red and blue colors indicate the up- and down-regulation during Pgc-1α overexpression in heart.

Due to the extensive downregulation of the circadian rhythm and the significant enrichment results, the circadian rhythm pathway with the smallest adjusted P-value ("Diurnally regulated genes with circadian orthologs" (WIKIPW)) was studied further. The identified pathway was from human but comparison to orthologous pathway in mouse revealed them to be highly conserved. Some genes are always lost in conversions, especially so in between organisms and therefore the orthologous pathway from mouse was selected for further analysis.

In order to study the pathway in Pathvisio[92], open-source biological pathway analysis software, Refseq IDs were converted to ensembl gene IDs with R/Bioconductor's package biomaRt. The expression and significance of the genes in the Pgc-1α overexpression dataset in the pathway

of interest were studied. As seen from the Figure 14., many genes of the pathway are affected by overexpression of Pgc-1α. Majority of the significant DEGs are downregulated. Interestingly, these include core clock components, such as *Per1*, *Per2* and *Arntl*.

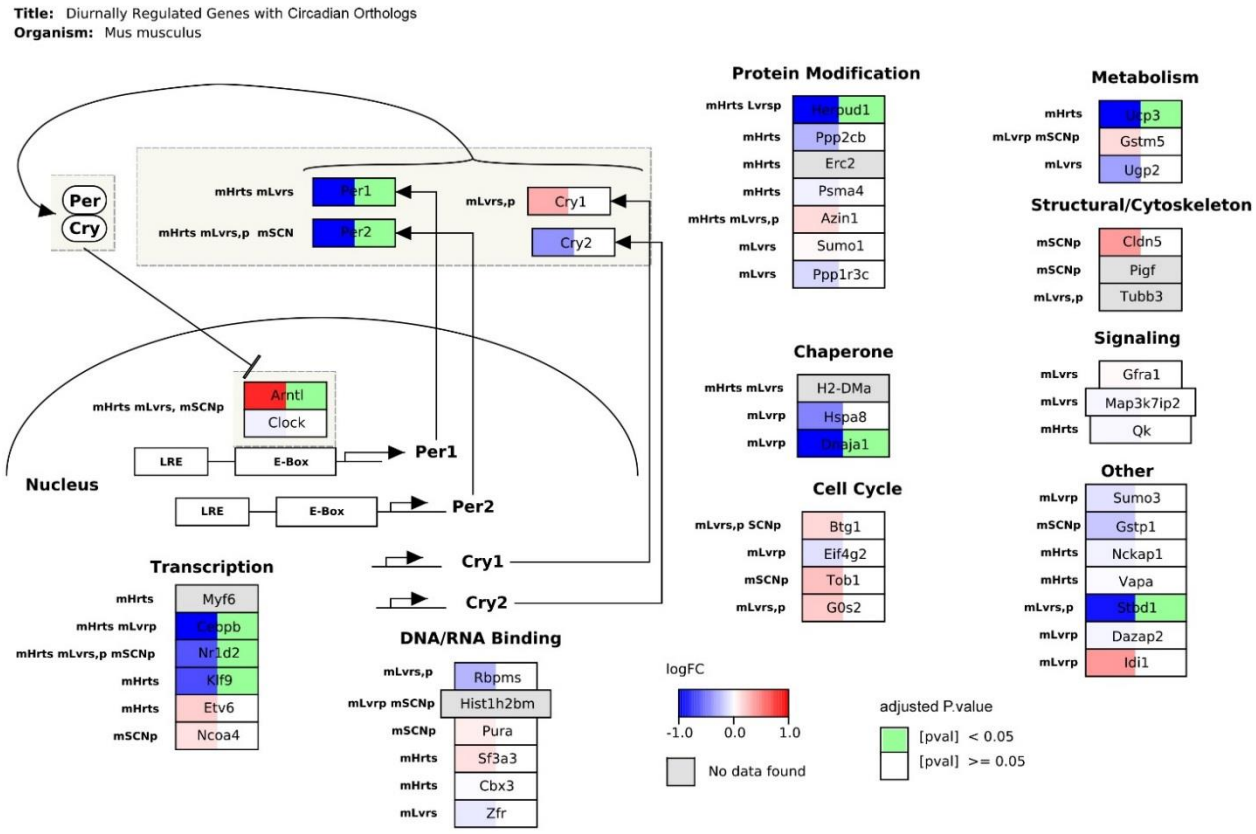

Figure 14. Circadian rhythm pathway affected by Pgc-1α overexpression. Genes expressed in the Pgc-1α overexpression dataset are colored. Blue and red correspond to low and high expression levels whereas green and white indicate the significance of the gene in the dataset. Grey indicates that the gene was not expressed in the dataset.

5.2 Pgc-1α overexpression in cardiomyocyte vs. in skeletal muscle

In order to investigate the tissue-specific effects of Pgc-1α overexpression between cardiomyocyte and skeletal muscle, the publicly available microarray dataset from Pérez-Schidler et al. was downloaded, processed and analyzed with aforementioned settings. Before

the statistical analysis, quality of the data was confirmed to be satisfactory (Figure S1B.). Finally, the statistical analysis yielded 6318 differentially expressed genes (adj. P-val. <0.05).

Next, gene set enrichment analysis was performed for skeletal muscle dataset. Again, threshold for significance was set at adj. P-val. <0.05. The result revealed 69 significantly upregulated pathways (Table S4.).

Computational comparison of the results from tissues showed that 51 of the significantly upregulated pathways were same in both tissues whereas 18 were unique to skeletal muscle and 11 to heart (Figure 14.). The shared pathways included those of glycolysis, beta-oxidation, electron transport chain and citric acid cycle. Surprisingly, many of the upregulated, significant pathways unique to skeletal muscle were involved with beta-oxidation. Other interesting, significant and unique pathways to skeletal muscle included electron transport chain, citric acid cycle and Ppar signaling (Table 9.). Closer inspection revealed the pathway in the heart ("Respiratory electron transport (Reactome)") as a subset of the pathway enriched in the skeletal muscle ("Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins (Reactome)").

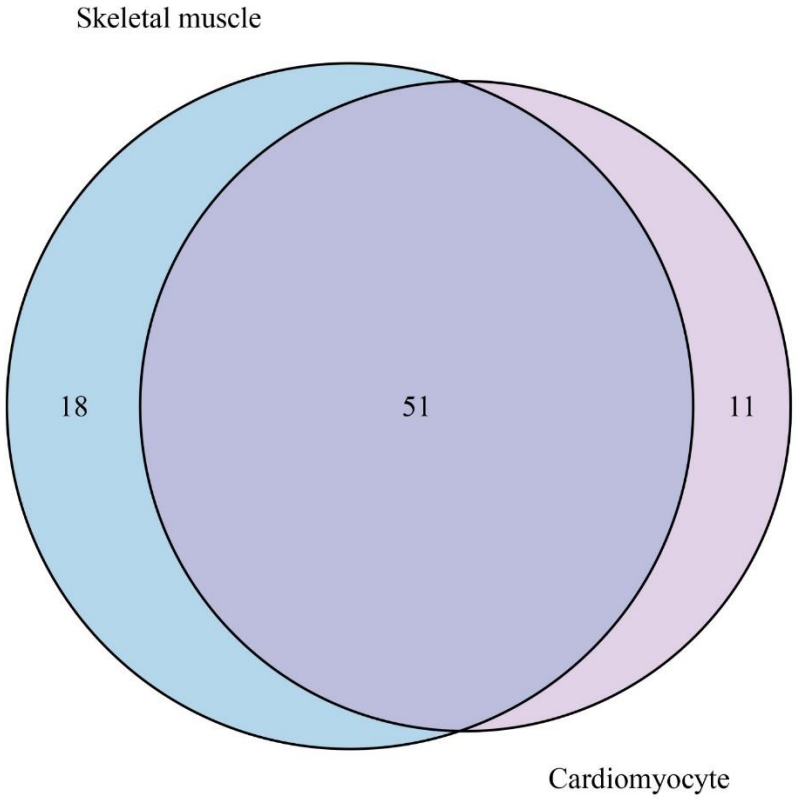

Figure 14. Venn diagram of gene enrichment results. The significantly enriched, upregulated pathways of Pgc-1α overexpression in heart and skeletal muscle represented as proportional Venn diagram.

Table 9. Upregulated, significantly enriched pathways of Pgc-1α overexpression in skeletal muscle. Table of upregulated, significantly (adj. P- val. < 0.05) enriched pathways unique to skeletal muscle dataset in comparison to heart with their respective adjusted P-values. The pathways of most interest are colored.

Pathway	**Adj. P-value**
Metabolism of lipids and lipoproteins (PWC)	0.0000
Mitochondrial fatty acid beta-oxidation (PWC)	0.0000
Mitochondrial fatty acid beta-oxidation of unsaturated fatty acids (PWC)	0.0000
Glucose metabolism (PWC)	0.0000
Activated AMPK stimulates fatty-acid oxidation in muscle (PWC)	0.0000
Integration of energy metabolism (PWC)	0.0000
Diabetes pathways (PWC)	0.0000
Huntingtons disease (KEGG)	0.0001
Mitochondrial fatty acid beta-oxidation of saturated fatty acids (PWC)	0.0001
Beta oxidation of palmitoyl-CoA to myristoyl-CoA (PWC)	0.0001
Import of palmitoyl-CoA into the mitochondrial matrix (PWC)	0.0001
Alzheimers disease (KEGG)	0.0001
Respiratory electron transport atp synthesis by chemiosmotic coupling and heat production by uncoupling proteins (REACTOME)	0.0028
Valine leucine and isoleucine degradation (KEGG)	0.0101
Metabolism of amino acids and derivatives (REACTOME)	0.0256
TCA Cycle (WIKIPW)	0.0411
Metabolism of lipids and lipoproteins (REACTOME)	0.0461
Ppara activates gene expression (REACTOME)	0.0487

The distribution of DEGs across curated pathways (Table S3.) was also investigated similar as for cardiomyocyte data. Interestingly, there were considerably more regulated genes in skeletal muscle than in the heart. Unlike in the heart, glycolysis, beta-oxidation, electron transport chain and citric acid cycle pathways were remarkably positively regulated (> 50 %). Growth and PI3K signaling seem to be downregulated in skeletal muscle whereas the direction of genes in calcium and PPAR signaling seem uniformly distributed (Figure 15.).

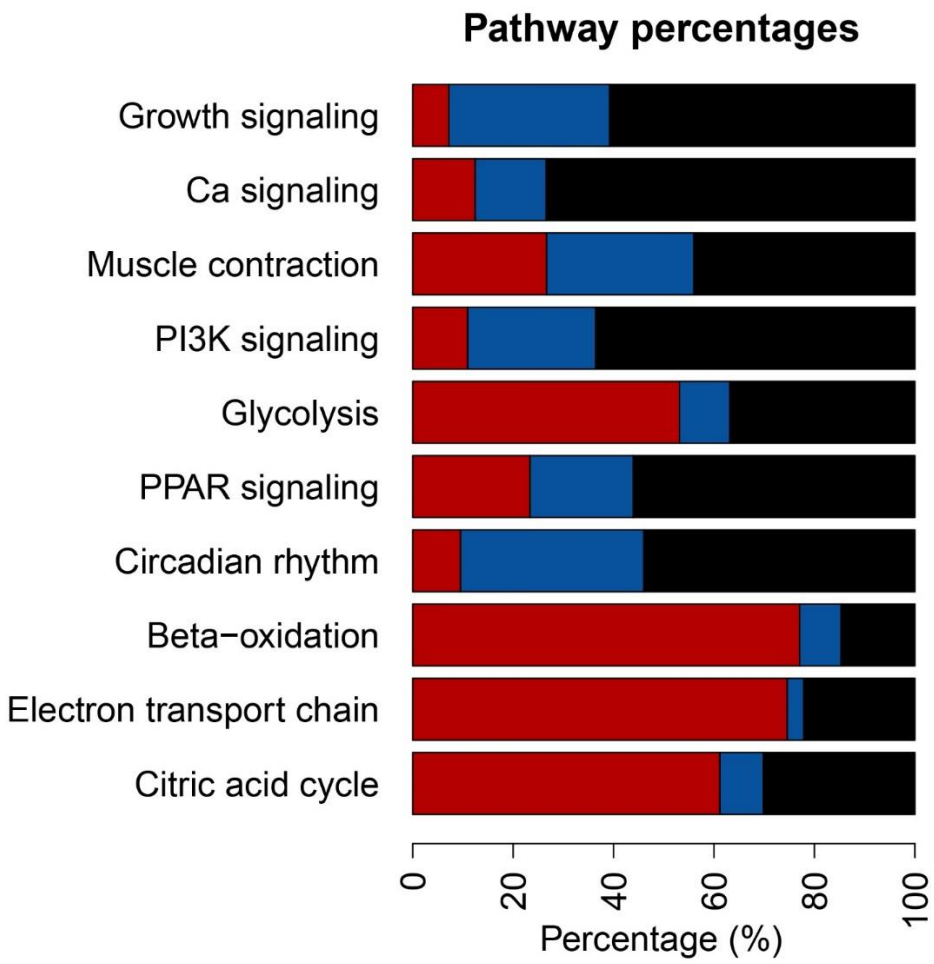

Figure 15. The extent of differential expression in pathways affected by Pgc-1α overexpression in skeletal muscle. Selected pathways curated from KEGG, Wikipathways, GO, Biocarta, Pathway Commons and Reactome databases (Table S3.) are shown as barplots. In each plot, the proportion of differentially expressed up- (in red) and downregulated (in blue) genes is compared to non-regulated genes (in black).

To summarize, in comparison to cardiomyocyte, there are shared and uniquely upregulated pathways concerning fatty acid beta-oxidation, citric acid cycle and gluconeogenesis in skeletal muscle. In addition, distribution of DEGs in the curated pathways in skeletal muscle is higher than in the heart. These findings imply that those pathways are upregulated to greater extent in skeletal muscle in comparison to cardiomyocyte.

Interestingly, apart from upregulation of Ppar signaling in skeletal muscle, none of the downregulated pathways in heart (muscle contraction, circadian rhythm, PI3K and calcium signaling) are significantly enriched in skeletal muscle. Based on the distribution of DEGs in curated pathways, however, circadian rhythm and PI3K signaling would be downregulated in skeletal muscle.

Nevertheless, these findings support the hypothesis that Pgc-1α overexpression has similar but also tissue-specific effects between cardiomyocyte and skeletal muscle.

5.3 Pgc-1α overexpression in heart vs. physiological and pathological hypertrophy

Overexpression of Pgc-1α has been linked with cardiac hypertrophy and even hypothesized as potential treatment target for the disease[8].

However, there is a drastic difference between physiological and pathological cardiac hypertrophy. Whereas physiological cardiac hypertrophy is a natural state that occurs during exercise and pregnancy, the pathological state is associated with heart failure. Thus it is critical to investigate the resemblance of Pgc-1α overexpression to both states in order to ascertain its possibility as a treatment target. To achieve this, publicly available microarray hypertrophy data of both states (Song et al.[73]), was downloaded, processed and analyzed. Before statistical analysis, the quality of the data was confirmed to be acceptable. (Figure S1C.)

First, gene set enrichment analysis was performed for the hypertrophy dataset. Like before, adj. P-val. < 0.05 was defined as significant. Results revealed pathological and physiological hypertrophy to have 408 and 105 significantly enriched pathways, respectively. Majority of these were upregulated (265 and 92, respectively). These significantly upregulated pathways are presented on Supplementary tables S5. and S6.

Comparison to upregulated, significantly enriched pathways in Pgc-1α overexpression in heart showed that only two pathways were shared with pathological hypertrophy whereas 38 were the same as in physiological state (Figure 16.). Interestingly, these shared upregulated pathways with physiological cardiomyopathy included those of electron transport chain, citric acid cycle, beta-oxidation and glycolysis. One downregulated pathway was the same in Pgc-1α overexpression and physiological hypertrophy (“Diurnally regulated genes with circadian orthologs (WIKIPW)”) whereas none were shared with pathological state.

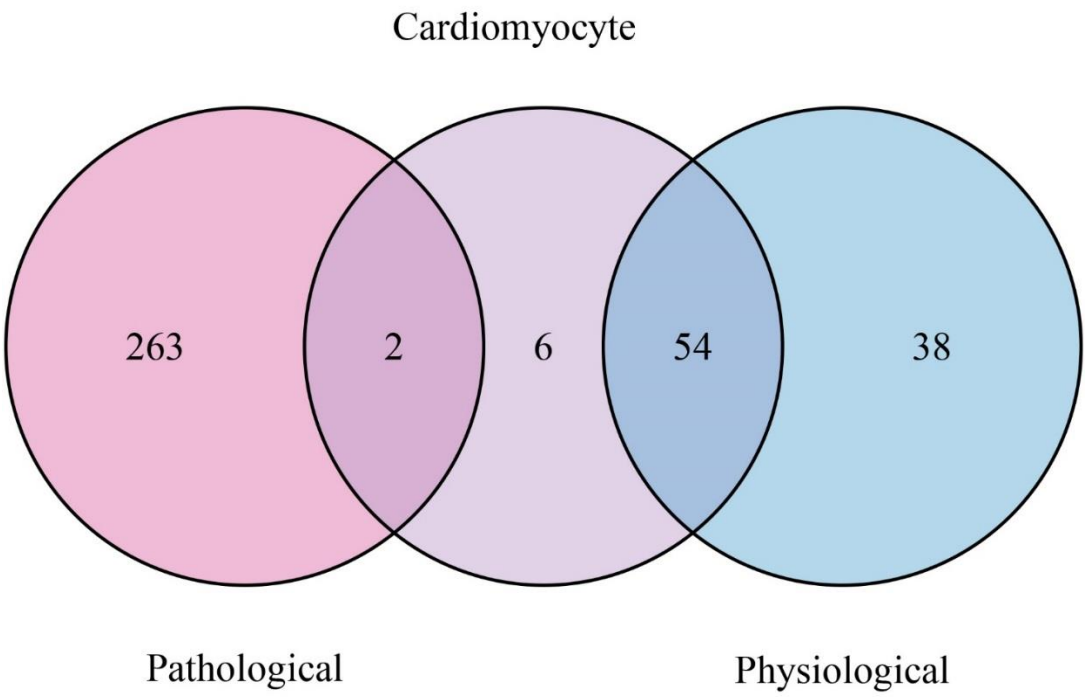

Figure 16. Venn diagram of gene enrichment results. The upregulated, significantly enriched pathways (adj. P-val. < 0.05) of Pgc-1α overexpression in cardiomyocyte, pathological and physiological cardiomyopathy are represented as Venn diagram.

To further compare Pgc-1α overexpression with hypertrophy states in the heart tissue, the differentially expressed genes (adj. P-val. <0.05) were compared. State of pathological hypertrophy, however, resulted in very high number of DEGs (4381) whereas physiological state only to 148.

There is vast difference between the number of differentially expressed genes in physiological and pathological hypertrophy states (4381 vs. 148). This extremely high number of DEGs in pathological cardiomyopathy is reasonable though, because pathological cardiomyopathy is a severe state that results in death. It causes drastic gene expression alteration which leads to stress and apoptosis, for example.

One way to examine Pgc-1α overexpression and hypertrophy states would simply be to compare the number of DEGs shared between the Pgc-1α overexpression and the two hypertrophy states. Yet in this case, it would not be purposeful. Due to extreme differences in the number of regulated genes (4381 vs. 148), Pgc-1α would have more genes shared with pathological state by chance alone, hence producing result that may not be biologically meaningful.

To take this great variety in the number of DEGs into account, hypergeometric test was performed. Differentially expressed genes of Pgc-1α overexpression were compared to the ones of physiological or pathological cardiomyopathy while using all of the expressed genes shared between the experiments (10395) as background. 40 and 262 of the genes were shared between Pgc-1α overexpression and physiological/pathological hypertrophy, respectively (Figure 17.). The same was tested by taking the directionality of the genes into account, but all four possibilities were significant (Figure S2.).

Hypergeometric test revealed the differentially expressed genes of Pgc-1α overexpression to be significantly enriched in the state of pathological hypertrophy (P-val. = 4.0389^{-07}) but even more so in physiological hypertrophy (P-val. = 0.00). This, along with gene set enrichment results, concludes that Pgc-1α overexpression resembles physiological rather than pathological hypertrophy.

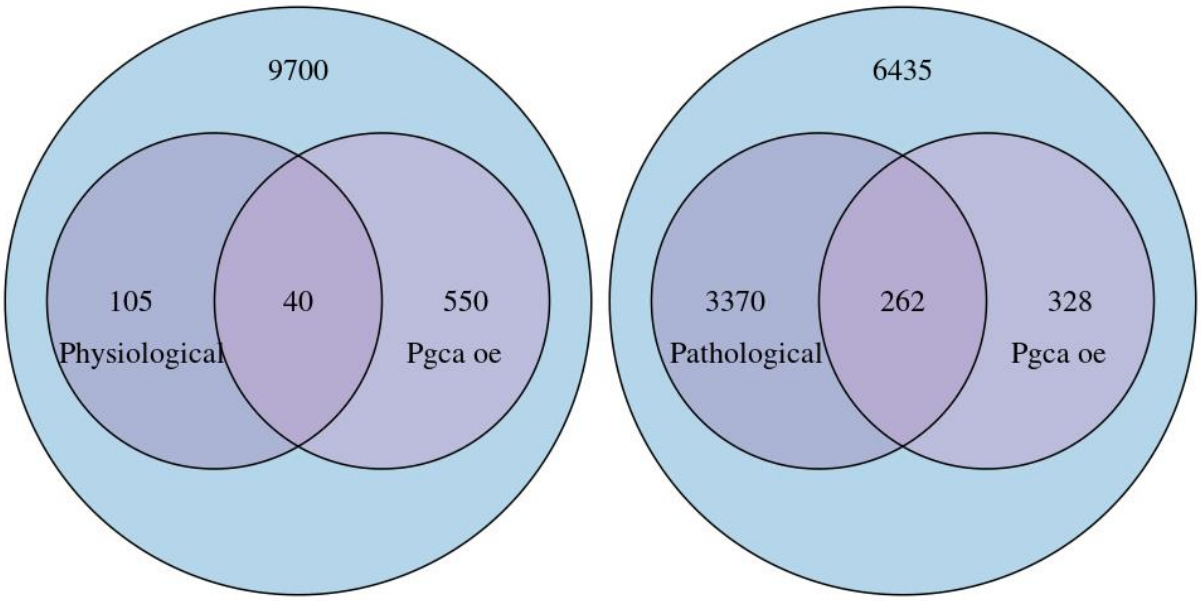

Figure 17. Hypergeometric distribution of Pgc-1α overexpression and physiological/pathological cardiomyopathy. The differentially expressed genes (adj. P-val. <0.05) of Pgc-1α overexpression were compared with DEGs of physiological and pathological cardiomyopathy in comparison to expressed genes across both experiments.

5.4 Pgc-1α overexpression vs. circadian rhythm

In the very beginning of the analysis of Pgc-1α overexpression results, the circadian rhythm unexpectedly arose as one of the most influenced pathways. As mentioned before, disruptions in circadian rhythm cause severe metabolic and physiological effects on organism, thus likely affecting the possibility of Pgc-1α as treatment target for hypertrophy. Due to this it was decided to compare Pgc-1α overexpression with two publicly available circadian rhythm datasets, one performed with microarray (Young et al.[74]) and one with qPCR (Wu et al.[75]).

The microarray circadian dataset 1 was downloaded, processed and analyzed as explained before. Like before, the quality control was performed and confirmed to be satisfactory (Figure S1D.).

Circadian microarray and Pgc-1α overexpression datasets were compared with gene set enrichment and hypergeometric test whereas the reported expression of circadian clock genes affected by Pgc-1α from study by Wu et al.[75] was compared to corresponding Pgc-1α overexpression results.

One of the ideas was to try to identify the timepoint most affected by Pgc-1α overexpression. However, since the aim to compare these wasn't originally in the aims but added after intriguing results, it posed challenges. The biggest of these was the design: Pgc-1α overexpression dataset wasn't time-series data and the timepoint in which the samples were collected was unknown.

Nevertheless, first approach was to perform gene set enrichment analysis for microarray circadian dataset. Unfortunately, only two of eight timepoints had enriched pathways.

Second approach was to apply hypergeometric test to determine the enrichment of genes affected by Pgc-1α overexpression in the circadian rhythm dataset per timepoint. Interestingly, the result revealed it to be significant in every single time point (P-val < 0.1E-8). According to this, genes regulated by Pgc-1α are detected through the day. This effect further suggests the importance of Pgc-1α in regulation of the circadian rhythm.

Third method was to calculate the percentage of DEGs (adj. P-val. < 0.05) shared between the datasets per timepoint. The highest percentage, 17.25 %, is at time point Zt 6. However, five of the timepoints have percentage of 15-18 (Table 10.).

Table 10. The percentage of differentially expressed genes shared between Pgc-1α overexpression and circadian rhythm per time point.

Timepoint	Percentage
Zt 0	10.0349
Zt 3	17.1913
Zt 6	17.2507
Zt 9	10.1947
Zt 12	14.1243
Zt 15	16.3347
Zt 18	17.1004
Zt 21	17.2414

Wu et al. studied the effect of Pgc-1α overexpression in mice in cardiomyocyte by following the circadian rhythm of selected genes with qPCR, successfully identifying seven circadian clock genes affected by this condition. The conditions of the experiment, including the background and tissue were the same as in the study of Tavi et al. and thus it was of interest to compare the results (Figure 18.).

Based on the directionality between the seven circadian clock genes (Figure 18.), it may be that timepoint most affected by Pgc-1α overexpression is somewhere between Zt 3 and Zt 8. However, due to small number of replicates and design of the study, it is impossible have certainty in the conclusion. Nevertheless, the results further show the importance of Pgc-1α on the regulation of the circadian clock.

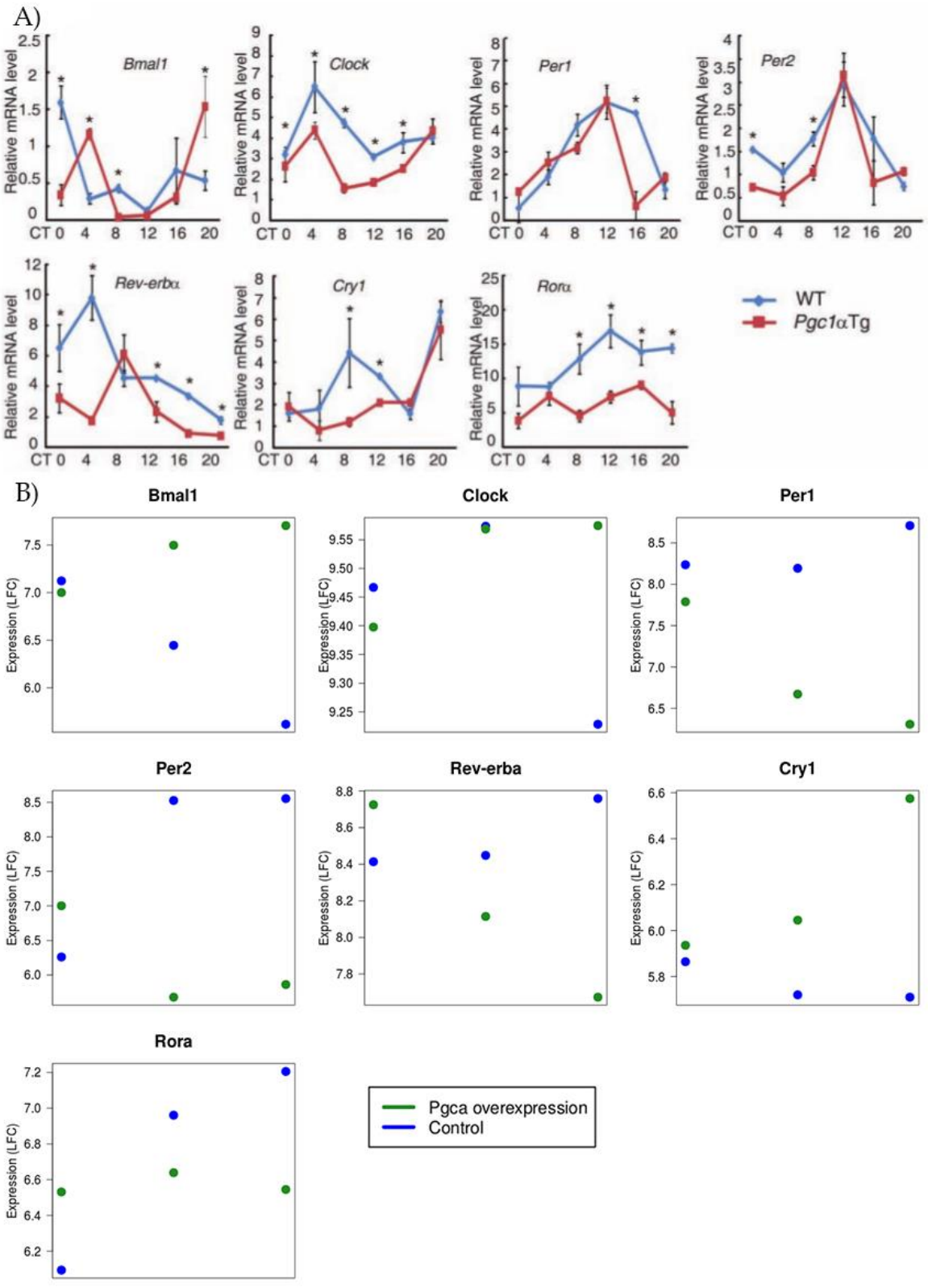

Figure 18. Expression of circadian clock genes under Pgc-1α overexpression. A) Wu et al. identified the disrupted circadian clock genes under Pgc-1α overexpression. The relative mRNA profiles are shown as mean +/- SEM for time points. Figure taken from reference 75. B) Log fold changes of disrupted circadian clock genes of replicates of Pgc-1α overexpression dataset.

6 DISCUSSION

Pgc-1α overexpression dataset was successfully compared to three publicly available genome-wide datasets, allowing the confirmation or rejection of the hypotheses. Enriched pathways were identified for Pgc-1α overexpression from cardiomyocyte and skeletal muscle datasets, along with physiological and pathological hypertrophy dataset. The tissue-specific effects of Pgc-1α overexpression between heart and skeletal muscle were discovered and similarity to physiological/pathological hypertrophy was confirmed. The unexpected, severe effect of Pgc-1α overexpression to circadian rhythm was identified. Unfortunately, the Pgc-1α dataset was not time-series data and the comparison to circadian rhythm datasets was not taken into account while planning the experiment. Due to this, the identification of the most affected timepoint in circadian rhythm by Pgc-1α overexpression was not confirmed. The results, however, further support the importance of Pgc-1α in regulation of the circadian rhythm.

The characteristics of the genome-wide datasets were studied before the analyses with several methods. Among these, one of the most important sample distance visualizations was MDS. MDS first calculates (dis)similarity matrix among observations, in this case samples, and plots samples in two-dimensional space. The resulting graph allows visual observation of the similarity between the samples. If the samples with same treatment cluster together, the quality of the data is good. Apart from the dataset of Pgc-1α overexpression in cardiomyocyte, the quality of the datasets was satisfactory (Figure S1). Sample number 1 of Pgc-1α overexpression in the dataset of Tavi et al. didn't cluster well with the others, implying possibility of an outlier. However, due to small number of samples (N = 3), certainty couldn't be reached and all three were kept in the analysis. Naturally, this lowered the quality and certainty of the performed analysis' but they still serve as a good template for future studies.

In this thesis, two kinds of genome-wide datasets were used: RNA-seq and microarray. As discussed in the literature review, both have their pros and cons, but they complement each other[41]. Therefore, after proper data processing, comparison between the results should pose trustworthy results.

All datasets used in this thesis were performed in mice. The strains across experiments differ, but according to the literature the difference between them is relatively low, especially in muscle tissue[93]. However, most of these differences occur in BL6 strain, breed used in datasets

from Tavi et al. and Wu et al. Fortunately, the biggest variability occurs in genes coding for structural proteins, determined by low expression[93].

Gene set enrichment methods can be used to explore the biological differences between phenotypes and datasets. The identification of these pre-constructed gene sets, or pathways, enable the detection of weak but consistent expression changes across a set of genes, resulting in better reproducibility and lower information loss in comparison to conclusions solely based on expression levels of individually analyzed genes[3,42,43]. There are variety of methods available, the choice depending of the research question and null hypothesis. The method of choice in this thesis was unsupervised, competitive, non-parametric GSA with GSEA as the statistical enrichment method. The method of choice can always be argued since the best performing method differs from study to study. GSEA isn't the perfect approach and it has been outperformed with simulated datasets but it has surpassed others with the use of experimental datasets[57]. It is also a widely accessed method and therefore is our enrichment method of choice. The algorithm of the combination of GSA and GSEA is discussed in materials and methods section. The biggest drawback of gene set enrichment settings of this thesis is, however, the use of gene randomization. As explained on the literature review, in gene randomization, gene is the sampling unit for permutation. In this type of randomization, gene-gene correlations are lost. Therefore, it is preferred to use sample randomization[51]. However, sometimes it is impossible to use sample randomization instead of gene sampling. This is due to overly complex phenotypes (such as analysis across all cell lines) or, more often, lack of replicates[51] ($N > 7$ for GSEA[94]). On the other hand, gene sampling methods have been suggested to be too powerful, declaring set as false positive based on only a few differentially expressed genes[55]. The threshold of significance is also subject of argumentation. Too high threshold may result in false negatives whereas too lenient one winds up with false positives. Again, there is no common guideline but generally used threshold is adj. P-val. < 0.05, and that is also the one used in this thesis.

Gene set enrichment analysis was performed for all high-throughput datasets. In overexpression of Pgc-1α in heart and muscle, 75 significantly enriched pathways were identified. The differences and similarities between these hasn't been widely studied but increased Pgc-1α expression has been linked with increase in mitochondrial biogenesis in both mouse heart and skeletal muscle[7,11,71]. The pathways linked to mitochondrial biogenesis include respiratory chain and fatty acid oxidation, for example[95,96]. The gene set enrichment and the distribution of differentially expressed genes across curated pathways in heart revealed circadian rhythm to be

downregulated under Pgc-1α overexpression. This finding stirred our interest and lead to further studies. Interestingly, while the same effect was implicated in skeletal muscle by distribution of DEGs, it was not among the enriched gene sets. This implies that the effect of Pgc-1α on circadian rhythm may be tissue specific.

While interpreting the results of gene set enrichment results, it should be also taken into consideration that the original construction of the pathways in the databases may be unreliable. This was also the case in three seemingly enriched pathways. Originally, gene set enrichment revealed heart tissue to have significant downregulation of circadian rhythm, muscle contraction, calcium and PPARα signaling pathways. Further inspection of expressed genes in muscle contraction and PPAR signaling pathways showed them to be false positives. Detailed discussion is presented below.

Traditionally, overexpression of Pgc-1α has been linked with induced PPAR signaling. Pgc-1α binds to and co-activates PPARα, thus inducing fatty acid oxidation[97–99]. However, Pgc-1α is able to induce beta-oxidation through other ways as well. While upregulation of PPAR pathway in skeletal muscle partially explains the heightened expression of fatty acid oxidation, downregulation of Pparα pathway (BIOCARTA) makes no sense in heart while upregulation of fatty acid oxidation is clearly enriched. The expressed genes of the pathway were further studied by heatmap (Figure S2). According to the heatmap, Pgc-1α may have an effect on the regulation of the pathway, but with only three samples it is rather feeble and thus the interpretation is challenging. The effect of many DEGs in the activation of Pparα pathway are also unknown and therefore it is difficult to have certainty whether they affect the activation/inactivation of the pathway. Still, this could be further studied by manually constructing list of Pparα targets and testing the enrichment of these genes in the Pgc-1α overexpression dataset. By knowing the state of the pathway (active/inactive), the effect of other genes would also be more thoroughly understood.

Interestingly, muscle contraction and calcium signaling, both significantly downregulated in heart under Pgc-1α overexpression, are heavily linked. In exercise, muscle fiber type changes towards more oxidative type which has greater endurance capacity instead of glycolytic. This corresponds to the oxidative effects of Pgc-1α[100,101]. Gene set enrichment revealed muscle contraction pathway ("Striated muscle contraction (WIKIPW)") to be significant in the heart. Sometimes, skeletal muscles are referred as striated muscles but in reality, this is not the case. Both cardiac and skeletal muscles have striations and can be referred as striated muscles,

although they differ in histology and physiology, making distinction crucial[102,103]. Due to this, the pathway was inspected more closely. Closer inspection revealed that the pathway indeed takes both skeletal and cardiac striated muscles into account. However, the pathway itself lacked connected lines and citations, implying unreliability. The DEGs of the Pgc-1α overexpression dataset in the pathway were also studied, revealing there to be only a few. Based on these discoveries, it is unlikely that the pathway is truly enriched in cardiac muscle. Muscle contraction pathway wasn't significantly enriched in the skeletal muscle either although literature implies otherwise[104].

Calcium signaling has been linked with muscle contraction and Pgc-1α. In endurance exercise, basal level of Pgc-1α is increased and only small amounts of calcium are released. In strength exercise on the other hand, calcium levels are elevated[100,101]. Upregulation of these CaMK (calcium/calmodulin-dependent protein kinase)-signaling pathways are known to stimulate MEF2 (myocyte enhancer factor 2) activity, which in turn induces Pgc-1α[105–108], driving towards more oxidative fibre-types and greater endurance capacity[100,101]. The role of calcium signaling in hypertrophy isn't, however, clear. While other studies have shown decrease in calcium activity, others have shown activation or no change at all[7]. Interestingly, according to our results of gene set enrichment, under overexpression of Pgc-1α calcium signaling is downregulated in heart but unchanged in skeletal muscle.

The effect of Pgc-1α in growth signaling was also studied. There were no significant pathways of growth signaling in neither heart nor in skeletal muscle. However, the downregulation was implicated in skeletal muscle by the distribution of DEGs. Literature also supports this implication. Fatty acid oxidation, which is clearly upregulated in both heart and skeletal muscle, is known to promote SIRT1 (NAD-dependent protein deacetylase sirtuin-1) activity which, at least in skeletal muscle, decreases growth[109,110]. Upregulation of SIRT1 has also been linked with inhibition of PI3K[111] which, according to the results, is significantly downregulated in heart. Its downregulation is also implicated in skeletal muscle by the distribution of DEGs.

In cardiac muscle, upregulation of Pgc-1α has also been linked with downregulation of PI3K and Akt signaling[112], which is also the case according to our gene set enrichment results. Interestingly, in hearts, the downregulation of these two is also associated with insulin resistance[113]. In skeletal muscle, reduction of PI3K signaling, also implicated in our results, has been suggested to play a role in skeletal muscle[114]. These findings imply that overexpression of Pgc-1α influences insulin resistance, at least in cardiac muscle.

Second enrichment method used in this study was unsupervised, competitive hypergeometric test. Straightforward hypergeometric test assumes gene independence, which in general is not true in biological systems. It also suffers for not weighting highly ranked genes, and therefore may produce too pessimistic outcomes, resulting in false negatives[64]. Here, hypergeometric test was used to test the significance of enrichment of DEGs upon Pgc-1α overexpression in physiological and pathological hypertrophy, revealing significant enrichment in both states, even if more so in the physiological (Figure 17.). As mentioned, hypergeometric test doesn't take the directionality of the genes into account. This was taken into consideration by testing separately for up- and downregulated genes. The result mimicked the former one: Pgc-1α overexpression was significant in up- and downregulated genes in both physiological and pathological hypertrophy (Figure S2.). This bias may be due to the fact that hypergeometric test completely ignores the relations of the genes.

According to the gene set enrichment however, state of Pgc-1α overexpression is drastically different of pathological hypertrophy but greatly resembles physiological hypertrophy. This also makes sense biologically: according to the literature, downregulation of Pgc-1α has been linked with pathological cardiac hypertrophy. Literature also implies that downregulated muscle contraction has been linked with pathological hypertrophy whereas in physiological state, muscle contraction is either upregulated or unchanged[104,115]. The results of this thesis support this: in Pgc-1α overexpression and physiological hypertrophy, there are no significant changes in muscle contraction whereas in pathological state it is significantly downregulated ("Cardiac muscle contraction (KEGG)").

The results of Ppar signaling also support this assumption: in pathological hypertrophy, Ppar signaling is reported to be downregulated[98]. As mentioned before, according to our results, under Pgc-1α overexpression Ppar signaling remains mainly unchanged in the heart.

Moreover, pathological cardiac hypertrophy is accompanied with downregulation of fatty acid oxidation whereas in physiological state, fatty acid oxidation is reported to be upregulated[98,116]. This is also the case in our results: only pathological state has reduced fatty acid oxidation. These results confirm that Pgc-1α overexpression resembles physiological rather than pathological cardiomyopathy and in that sense, it may be used as a treatment target. Naturally, disruptions of circadian rhythm and insulin resistance reduce this compatibility.

However, according to the result of gene set enrichment and regulation of the differentially expressed genes among curated pathways, overexpression of Pgc-1α causes downregulation of

circadian rhythm pathway. As mentioned in the literature review, disruptions of circadian rhythm cause changes in bodily functions and have been linked to variety of diseases, including obesity and mental illnesses. Two circadian rhythm related pathways were also significantly enriched in the gene set enrichment results. One of these pathways ("Diurnally regulated pathways with circadian orthologs (WIKIPW)" was further studied. According to the results, *Bmal1* is significantly upregulated under the overexpression of Pgc-1α whereas other core components of the circadian clock, apart from *Clock,* are downregulated. This event is also supported by the literature: Pgc-1α upregulates *Bmal1* by activating RORs and while transcription of *Bmal1* is highest, Pgc-1α protein peaks[117]. According to the visualization of significantly enriched circadian rhythm pathway ("Diurnally regulated genes with circadian orthologs (WIKIPW)"), the overexpression of Pgc-1α significantly affects *Bmal1, Per1* and *Per2*. It also seems to affect *Cry1* and *Cry2,* all of which belong to the core clock. This highlights the importance of Pgc-1α and suggests the possibility that Pgc-1α may also belong to the core clock components. Therefore, it would be of an interest to study this intriguing effect further.

In order to study this unexpected disruption of circadian rhythm even further, the dataset of Pgc-1α overexpression in heart was compared with circadian rhythm datasets. While our results support the conclusion that Pgc-1α overexpression causes disruptions in circadian rhythm by causing downregulation of the pathway, identification of time point was not reliable; neither gene set enrichment nor hypergeometric test or manual comparison revealed reliable results. The low number of replicates with possible outlier affected the comparison and ignorance of timepoint in which the samples were collected provided extra challenges. The biggest problem, however, was the design of the experiment. In order to properly study the effect the dynamics of circadian regulation the experiment should be re-designed as a time-series with higher number of samples and replicates.

In addition to previously mentioned methods, the effect of Pgc-1α overexpression to circadian rhythm could also be studied computationally. This would be especially effective because circadian rhythm is one of the most complicated pathways due to its hefty size and heavy regulation of autonomous transcription-translation feedback loops. Based on pathway databases and literature, a model of circadian rhythm pathway could be built. Then this model could be disturbed and the effect of disruption, such as Pgc-1α overexpression, could be studied first by modeling simulations and then confirming it experimentally. The computational model studies aim to uncover general principles of circadian clock and provide more abstract interpretations

in the systems view. However, the models are often simplified, making them mathematically tractable and require no extraneous details. The system generates predicted outcomes provided by training data and due to this, thus in theoretical point of view, whole system can treated as mechanistic "black box" as long as it generates the predictions. This synthetic approach has been used to mimic circadian clock and investigate rhythmic outcomes of generated by topological schemes, for example[118].

Luckily, the circadian clock has been studied as modeled for centuries and thus there already are models to use. Therefore, the first step would be looking into the existing models and choosing one for adaptation in order to investigate more specific questions, such as the effect of Pgc-1α overexpression. Example of a potential model could be from Regorio et al[119]. Their model includes the core clock genes and the two main feedback loops, and has been tested by comparing results from mutation data from knockout mice and verified with human osteocarcinoma cells. Naturally, this is something to take into account while testing and building up the model: circadian rhythm is known to have tissue-dependent effects. This knowledge motivates the collection and research of variety of tissues in order to study tissue specific effects of circadian rhythm. We have shown that overexpression of Pgc-1α are, at least partially, tissue specific. After successfully conducting the modeling of the effects of Pgc-1α in cardiomyocytes, it would be interesting to test the tissue specificity with skeletal muscle dataset, as well.

The design of the model of Regorio et al. is based on ordinary differential equations (ODEs) which is one of the two widely used methods in computational modeling for system biology. Descriptions with ODE system may take detailed knowledge, such as concentrations of the substrates, individual protein-protein functions and gene regulatory mechanisms into account. Based on these, dynamics of the mRNA concentrations of the system can be presented. Challenge in this is the lack of information – not much is known about kinetic constants and often functions of many proteins and their interactions are uncertain. Time-resolved concentration data is also challenging to measure[120], and it is what would be required to study the effect of Pgc-1α overexpression.

It can also be argued that ODEs assumptions of continuous and deterministic concentration are not valid when it comes to gene expression. On single cell level, the abundance of molecules is often low. Moreover, the abundance of mRNA molecules is also often below detection limit, causing uncertainty whether process is actually present or not. Furthermore, in transcription, it

takes time from initiation until termination. Therefore it is debatable whether the process can be regarded as continuous like it is described in ODE models. One way to overcome these challenges is stochastic stimulation which computes concentration for each molecule along time[120].

Another way to build up a network model would be to construct a Boolean network in which all the states are binary (Figure 19.). Boolean models are simpler than ODEs and therefore require less time and effort, but they suffer from one major limitation: in Boolean model, genes are either active or inactive. Biologically, this is rarely the case. Therefore, for modeling circadian rhythm and the effect Pgc-1α on it, ODE model would be better. Although, as mentioned before, it requires more time and effort, both with the modeling and gathering the data in the laboratory.

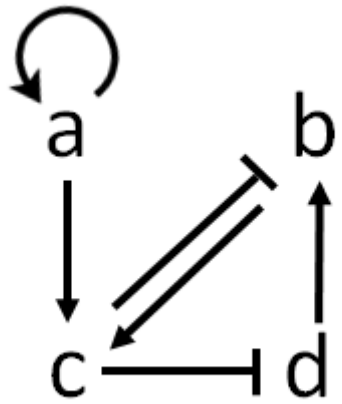

Figure 19. Boolean gene regulatory network.
Rules:
a(t+1) = a(t) *b(t+1) = not c(t) and d(t)*
c(t+1) = a(t) and b(t) *d(t+1) = not c(t)*
In Boolean regulatory network, all states are binary. Statements are defined by using operators "and", "or" or "not" and their combinations. In modeling of gene regulation networks, genes are the elements of the network. They have two states, expressed (assigned value 1) or not expressed (assigned value 0). Gene can either activate, dissuade or not affect another gene based on the rules of the network. The network drives the system towards a steady state which is constant and no changes of expression are happening. If, however, a gene is activated by external stimulation for example, it may drive towards an exit from this attractor and eventually to a new steady state by affecting expression of other genes which affect further genes and so on until steady state (either the original or a new one) is reached[120]. In this particular network, or *b* to be active, there should be *d* available whereas for *c* to be active, there should be both *b* and *d* available. This system, like any other, evolves towards steady

states through attractors. Based on the rules, if initial state of *a* is 1 (switched on), at first *a* is active and other genes are switched off (1000). Then, *d* is activated because it is not repressed (1001). Next, *d* activates *b* (1101). After this *a* and *b* and a activate *c*, resulting in state where all genes are active (1111). Then *c* inhibits *b* and *d* (1010). Because *b* is not active, no *c* is produced and gene becomes inactive (1000). Thus: 1000 → 1001 → 1101 → 1111 → 1010 → 1000. On the other hand, if initial state of *a* is 0, *d* is active and activates *b* (0101). *c* isn't activated because there is no *a*.

After choosing the model, conduction of experiments can begin. Whereas the generalized mathematical models can provide directions and highlight the importances of networks, it is essential to experimentally validate these models and understand the precise molecular basis of regulation[118,121]. It is also crucial to choose the model to be adopted before conducting experiments because then the parameters needed for the modeling, such as degradation and inhibition rates, and its experimental work needed for adjustments are known. In the case of adapting the model for studying of the effect of Pgc-1α, the experimental analysis to identify and ensure its link to circadian rhythm through time-series experiment would be necessary. A potential link between the Pgc-1α and the circadian clock could be its effect through RORs, which leads to heightened *Bmal1* expression. This is not only suggested in the literature[117] but also in our results.

After this, the model could be used to predict the effects of Pgc-1α overexpression, for example. Lastly, the model should be verified by conducting time-series experiment of Pgc-1α overexpression on circadian rhythm. If the model is able to predict the results from the experiment, it is validated and could be further used to predict other perturbations as well, such as reduced expression or knockout of Pgc-1α.

An intriguing, more experimental way to test the effect and importance of Pgc-1α on circadian rhythm are so-called resonance experiments. Pgc-1α has been suggested to be a pheripheral oscillator[122], player of the melody that SCN guides. These circadian oscillators have evolved to anticipate organism's biological needs during the light-dark cycle, causing the period length (tau) to be in resonance with period of the light-dark cycle (T cycle). By creating several mutant organisms with different tau, let's say 20, 24 and 30 h, their performance under the corresponding T cycle could be tested. For example, organism with 24 tau should perform better under T cycle of 24 h than 20 or 30 h. Likewise, organisms with 20 tau and 30 tau should outperform others under T cycle 20 and 30 h, respectively. If the resonance of tau with T cycles increases and decreases organism's fitness and even survival, then cause is the interaction of

the clock with environmental rhythms rather than the mutation of the gene or TF itself[23]. Studies like this have been conducted in cyanobacteria, but not yet in mammals[23]. According to our results, overexpression of Pgc-1α significantly affects many of the core clock genes and thus is likely to affect period lengths. By first identifying the period length change caused by overexpression Pgc-1α, its importance as an oscillator could be further studied by these resonance experiments.

By understanding how Pgc-1α regulates expression of circadian rhythm and its components, the potential of Pgc-1α as treatment target could be concluded. Naturally, experiments like this would also benefit the circadian rhythm research (possibly even introduce a new core clock gene) and may even uncover treatment targets for circadian rhythm disorders.

7 CONCLUSION

Genome-wide RNA-seq and microarray datasets from different tissues were used to answer the three hypothesis in this thesis. The hypotheses were confirmed by using unsupervised gene set enrichment tools, hypergeometric test and curated pathway analysis.

The pathways affected by Pgc-1α overexpression were identified and confirmed, with surprising result of downregulation of circadian rhythm. Interestingly, Pgc-1α overexpression seemed to cause insulin resistance in mice. However, further studies should be conducted to verify and further investigate this result.

The tissue-dependent effect of Pgc-1α overexpression was confirmed with GSA's gene set enrichment with GSEA's algorithm. The biggest difference was the upregulation of PPAR signaling in skeletal muscle and higher upregulation of fatty acid oxidation in skeletal muscle.

Furthermore, Pgc-1α overexpression was shown to resemble physiological rather than pathological hypertrophy, suggesting its safety as treatment target. The downregulation of circadian rhythm and possibility of abnormal insulin regulation, however, threatens this safety. While Pgc-1α overexpression clearly disrupts circadian rhythm, the severity of it remains unconfirmed. The association with insulin resistance also requires closer inspection. The success of targeting Pgc-1α, however, may lay in controlling the overexpression, and potentially

the clinical window will therefore be very narrow. Nevertheless, further studies, both computational modeling and laboratory work are necessary to confirm this.

In summary, the results obtained in this thesis allowed identification of the effects of Pgc-1α overexpression on gene expression in heart tissue, identified tissue-dependent effects in comparison with skeletal muscle and provided insight of its possibility as a treatment target for cardiac hypertrophy. This information is needed to further investigate the effects of Pgc-1α levels on tissue physiology. By providing clues on the key pathways for deeper investigation of Pgc-1α, this study works as a beneficial template for future studies.

8 REFERENCES

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

2. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14,** 89–99 (2013).

3. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–50 (2005).

4. Puigserver, P. & Spiegelman, B. M. Peroxisome proliferator-activated receptor-gamma coactivator 1 alpha (PGC-1 alpha): transcriptional coactivator and metabolic regulator. *Endocr. Rev.* **24,** 78–90 (2003).

5. Finck, B. N. & Kelly, D. P. PGC-1 coactivators: inducible regulators of energy metabolism in health and disease. *J. Clin. Invest.* **116,** 615–22 (2006).

6. Austin, S. & St-Pierre, J. PGC1α and mitochondrial metabolism--emerging concepts and relevance in ageing and neurodegenerative disorders. *J. Cell Sci.* **125,** 4963–71 (2012).

7. Ventura-Clapier, R., Garnier, A. & Veksler, V. Transcriptional control of mitochondrial biogenesis: the central role of PGC-1alpha. *Cardiovasc. Res.* **79,** 208–17 (2008).

8. Ichida, M., Nemoto, S. & Finkel, T. Identification of a specific molecular repressor of the peroxisome proliferator-activated receptor gamma Coactivator-1 alpha (PGC-1alpha). *J. Biol. Chem.* **277,** 50991–5 (2002).

9. Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140,** 744–52 (2010).

10. Schuler, M. *et al.* PGC1alpha expression is controlled in skeletal muscles by PPARbeta, whose ablation results in fiber-type switching, obesity, and type 2 diabetes. *Cell Metab.* **4,** 407–14 (2006).

11. Dillon, L. M., Rebelo, A. P. & Moraes, C. T. The role of PGC-1 coactivators in aging skeletal muscle and heart. *IUBMB Life* **64,** 231–41 (2012).

12. Lira, V. A., Benton, C. R., Yan, Z. & Bonen, A. PGC-1alpha regulation by exercise training and its influences on muscle function and insulin sensitivity. *Am. J. Physiol. Endocrinol. Metab.* **299,** E145–61 (2010).

13. Ventura-Clapier, R., Garnier, A. & Veksler, V. Energy metabolism in heart failure. *J. Physiol.* **555,** 1–13 (2004).

14. Trivedi, C. M. & Epstein, J. A. Heart-healthy hypertrophy. *Cell Metab.* **13,** 3–4 (2011).

15. Marín-García, J. *Mitochondria and Their Role in Cardiovascular Disease*. **19,** (Springer Science & Business Media, 2012).

16. Jagannath, A., Peirson, S. N. & Foster, R. G. Sleep and circadian rhythm disruption in neuropsychiatric illness. *Curr. Opin. Neurobiol.* **23,** 888–94 (2013).

17. Mohawk, J. A., Green, C. B. & Takahashi, J. S. Central and peripheral circadian clocks in mammals. *Annu. Rev. Neurosci.* **35,** 445–62 (2012).

18. Zhao, X. *et al.* Nuclear receptors rock around the clock. *EMBO Rep.* **15,** 518–28 (2014).

19. Bass, J. & Takahashi, J. S. Circadian integration of metabolism and energetics. *Science* **330,** 1349–54 (2010).

20. Andrews, J. L. *et al.* CLOCK and BMAL1 regulate MyoD and are necessary for maintenance of skeletal muscle phenotype and function. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 19090–5 (2010).

21. Liu, C., Li, S., Liu, T., Borjigin, J. & Lin, J. D. Transcriptional coactivator PGC-1alpha integrates the mammalian clock and energy metabolism. *Nature* **447,** 477–81 (2007).

22. Richards, J. & Gumz, M. L. Advances in understanding the peripheral circadian clocks. *FASEB J.* **26,** 3602–13 (2012).

23. Asher, G. & Schibler, U. Crosstalk between components of circadian and metabolic cycles in mammals. *Cell Metab.* **13,** 125–37 (2011).

24. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

25. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–6 (2011).

26. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–504 (2003).

27. Trevino, V., Falciani, F. & Barrera-Saldaña, H. A. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol. Med.* **13,** 527–41

28. Karin Zimmermann, U. L. Analysis of Affymetrix Exon Arrays. at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.175.7485>

29. Kapur, K., Xing, Y., Ouyang, Z. & Wong, W. H. Exon arrays provide accurate assessments of gene expression. *Genome Biol.* **8,** R82 (2007).

30. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

31. Draghici, S., Khatri, P., Eklund, A. C. & Szallasi, Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **22,** 101–9 (2006).

32. García de la Nava, J., van Hijum, S. & Trelles, O. Saturation and quantization reduction in microarray experiments using two scans at different sensitivities. *Stat. Appl. Genet. Mol. Biol.* **3,** Article11 (2004).

33. Tran, P. H. *et al.* Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res.* **30,** e54 (2002).

34. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9,** e78644 (2014).

35. Xing, Y., Kapur, K. & Wong, W. H. Probe selection and expression index computation of Affymetrix Exon Arrays. *PLoS One* **1,** e88 (2006).

36. Kel, A., Voss, N., Jauregui, R., Kel-Margoulis, O. & Wingender, E. Beyond microarrays: find key transcription factors controlling signal transduction pathways. *BMC Bioinformatics* **7 Suppl 2,** S13 (2006).

37. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biol.* **11,** 220 (2010).

38. Peano, C. *et al.* An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb. Inform. Exp.* **3,** 1 (2013).

39. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15,** 419 (2014).

40. Miller, D. F. B. *et al.* A new method for stranded whole transcriptome RNA-seq. *Methods* **63,** 126–34 (2013).

41. Kogenaru, S., Qing, Y., Guo, Y. & Wang, N. RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genomics* **13,** 629 (2012).

42. Abatangelo, L. *et al.* Comparative study of gene set enrichment methods. *BMC Bioinformatics* **10,** 275 (2009).

43. Nam, D. & Kim, S.-Y. Gene-set approach for expression pattern analysis. *Brief. Bioinform.* **9,** 189–97 (2008).

44. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199–205 (2014).

45. Nishimura, D. BioCarta. *Biotech Softw. Internet Rep.* **2,** 117–120 (2001).

46. Sales, G., Calura, E., Martini, P. & Romualdi, C. Graphite Web: Web tool for gene set analysis exploiting pathway topology. *Nucleic Acids Res.* **41,** W89–97 (2013).

47. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4,** P3 (2003).

48. Karp, P. D. Pathway databases: a case study in computational symbolic theories. *Science* **293,** 2040–4 (2001).

49. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472–7 (2014).

50. Kelder, T. *et al.* WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* **40,** D1301–7 (2012).

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

52. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1,** 107–129 (2007).

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

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

55. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.* **15,** 504–18 (2014).

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

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

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

59. 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).

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

61. 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).

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

63. Irizarry, R. A., Wang, C., Zhou, Y. & Speed, T. P. Gene set enrichment analysis made simple. *Stat. Methods Med. Res.* **18,** 565–75 (2009).

64. Tamayo, P., Steinhardt, G., Liberzon, A. & Mesirov, J. P. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat. Methods Med. Res.* (2012). doi:10.1177/0962280212460441

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

66. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10,** 161 (2009).

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

68. *Genomic and Personalized Medicine, Volumes 1-2*. **11,** (Academic Press, 2008).

69. Larranaga, P. Machine learning in bioinformatics. *Brief. Bioinform.* **7,** 86–112 (2006).

70. Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **7,** 55–65 (2006).

71. Nikolić, N. *et al.* Overexpression of PGC-1α increases fatty acid oxidative capacity of human skeletal muscle cells. *Biochem. Res. Int.* **2012,** 714074 (2012).

72. Pérez-Schindler, J. *et al.* The corepressor NCoR1 antagonizes PGC-1α and estrogen-related receptor α in the regulation of skeletal muscle function and oxidative metabolism. *Mol. Cell. Biol.* **32,** 4913–24 (2012).

73. Song, H. K., Hong, S.-E., Kim, T. & Kim, D. H. Deep RNA sequencing reveals novel cardiac transcriptomic signatures for physiological and pathological hypertrophy. *PLoS One* **7,** e35552 (2012).

74. Young, M. E. *et al.* Cardiomyocyte-specific BMAL1 plays critical roles in metabolism, signaling, and maintenance of contractile function of the heart. *J. Biol. Rhythms* **29,** 257–76 (2014).

75. Wu, X. *et al.* The circadian clock influences heart performance. *J. Biol. Rhythms* **26,** 402–11 (2011).

76. Yamamoto, H. *et al.* NCoR1 is a conserved physiological modulator of muscle mass and oxidative function. *Cell* **147,** 827–39 (2011).

77. Lin, J. *et al.* Transcriptional co-activator PGC-1 alpha drives the formation of slow-twitch muscle fibres. *Nature* **418,** 797–801 (2002).

78. Cha, H. *et al.* PICOT is a critical regulator of cardiac hypertrophy and cardiomyocyte contractility. *J. Mol. Cell. Cardiol.* **45,** 796–803 (2008).

79. McMullen, J. R. *et al.* Phosphoinositide 3-kinase(p110alpha) plays a critical role for the induction of physiological, but not pathological, cardiac hypertrophy. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 12355–60 (2003).

80. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. at <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>

81. FASTX-Toolkit. at <http://hannonlab.cshl.edu/fastx_toolkit/>

82. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

83. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

84. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–5 (2010).

85. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–89 (2010).

86. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7,** e30619 (2012).

87. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).

88. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39,** D685–90 (2011).

89. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43,** D1049–1056 (2014).

90. Yang, D. A Critical Review of Gene Set Enrichment Analysis: Development and Improvement. (2012). at <http://biochem218.stanford.edu/Projects 2012/Yang.pdf>

91. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57,** 289 – 300 (1995).

92. Kutmon, M. *et al.* PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput. Biol.* **11,** e1004085 (2015).

93. Turk, R. *et al.* Gene expression variation between mouse inbred strains. *BMC Genomics* **5,** 57 (2004).

94. *Gene Regulatory Network Reconstruction and Pathway Inference from High Throughput Gene Expression Data*. (ProQuest, 2008). at <https://books.google.com/books?id=uw_7kMtjn6sC&pgis=1>

95. Scarpulla, R. C. Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiol. Rev.* **88,** 611–38 (2008).

96. Kang, C. & Ji, L. L. Role of PGC-1α in muscle function and aging. *J. Sport Heal. Sci.* **2,** 81–86 (2013).

97. Finck, B. N. & Kelly, D. P. Peroxisome proliferator-activated receptor gamma coactivator-1 (PGC-1) regulatory cascade in cardiac physiology and disease. *Circulation* **115,** 2540–8 (2007).

98. *Molecular Defects in Cardiovascular Disease*. (Springer Science & Business Media, 2011). at <https://books.google.com/books?id=EUOgGY6gz4kC&pgis=1>

99. Rowe, G. C., Jiang, A. & Arany, Z. PGC-1 coactivators in cardiac development and disease. *Circ. Res.* **107,** 825–38 (2010).

100. Handschin, C. & Spiegelman, B. M. The role of exercise and PGC1alpha in inflammation and chronic disease. *Nature* **454,** 463–9 (2008).

101. *Histone Deacetylases: the Biology and Clinical Implication*. (Springer Science & Business Media, 2011). at <https://books.google.com/books?id=ojYtE3IxYTEC&pgis=1>

102. Striated Muscle - Anatomy, Histology, Function | Kenhub. at <https://www.kenhub.com/en/library/anatomy/striated-musculature>

103. Lommel, A. T. L. Van. *From Cells to Organs: A Histology Textbook and Atlas*. (Springer Science & Business Media, 2012). at <https://books.google.com/books?id=nHbgBwAAQBAJ&pgis=1>

104. Hou, J. & Kang, Y. J. Regression of pathological cardiac hypertrophy: signaling pathways and therapeutic targets. *Pharmacol. Ther.* **135,** 337–54 (2012).

105. Wu, H. *et al.* Regulation of mitochondrial biogenesis in skeletal muscle by CaMK. *Science* **296,** 349–52 (2002).

106. Czubryt, M. P., McAnally, J., Fishman, G. I. & Olson, E. N. Regulation of peroxisome proliferator-activated receptor gamma coactivator 1 alpha (PGC-1 alpha) and mitochondrial function by MEF2 and HDAC5. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 1711–6 (2003).

107. Roberts-Wilson, T. K. *et al.* Calcineurin signaling and PGC-1alpha expression are suppressed during muscle atrophy due to diabetes. *Biochim. Biophys. Acta* **1803,** 960–7 (2010).

108. Antico Arciuch, V. G., Elguero, M. E., Poderoso, J. J. & Carreras, M. C. Mitochondrial regulation of cell cycle and proliferation. *Antioxid. Redox Signal.* **16,** 1150–80 (2012).

109. Rollo, C. D. Aging and the Mammalian regulatory triumvirate. *Aging Dis.* **1,** 105–38 (2010).

110. Corbi, G. *et al.* Adrenergic signaling and oxidative stress: a role for sirtuins? *Front. Physiol.* **4,** 324 (2013).

111. Kwon, H.-S. & Ott, M. The ups and downs of SIRT1. *Trends Biochem. Sci.* **33,** 517–25 (2008).

112. Borniquel, S. *et al.* Inactivation of Foxo3a and subsequent downregulation of PGC-1 alpha mediate nitric oxide-induced endothelial cell migration. *Mol. Cell. Biol.* **30,** 4035–44 (2010).

113. Guo, S. Insulin signaling, resistance, and the metabolic syndrome: insights from mouse models into disease mechanisms. *J. Endocrinol.* **220,** T1–T23 (2014).

114. Kim, Y. B., Nikoulina, S. E., Ciaraldi, T. P., Henry, R. R. & Kahn, B. B. Normal insulin-dependent activation of Akt/protein kinase B, with diminished activation of phosphoinositide 3-kinase, in muscle in type 2 diabetes. *J. Clin. Invest.* **104,** 733–41 (1999).

115. Abel, E. D. & Doenst, T. Mitochondrial adaptations to physiological vs. pathological cardiac hypertrophy. *Cardiovasc. Res.* **90,** 234–42 (2011).

116. Marks, A. R., Investigation, A. S. for C. & Neill, U. S. *Science In Medicine: The JCI Textbook Of Molecular Medicine*. **1,** (Jones & Bartlett Learning, 2007).

117. Pavlath, G. K. *Myogenesis*. (Academic Press, 2011). at <https://books.google.com/books?id=kuU4fxqkiJcC&pgis=1>

118. Zhang, E. E. & Kay, S. A. Clocks not winding down: unravelling circadian networks. *Nat. Rev. Mol. Cell Biol.* **11,** 764–76 (2010).

119. Relógio, A. *et al.* Tuning the Mammalian Circadian Clock: Robust Synergy of Two Loops. *PLoS Comput. Biol.* **7,** e1002309 (2011).

120. Klipp, E. *et al. Systems Biology: A Textbook*. *Time* (2009). doi:10.3797/scipharm

121. Middleton, A. M., Farcot, E., Owen, M. R. & Vernoux, T. Modeling Regulatory Networks to Understand Plant Development: Small Is Beautiful. *Plant Cell* **24,** 3876–3891 (2012).

122. *Issues in Eating Disorders, Nutrition, and Digestive Medicine: 2013 Edition*. (ScholarlyEditions, 2013). at <https://books.google.com/books?id=_LXpmRaFmrYC&pgis=1>

9 SUPPLEMENTARY MATERIAL

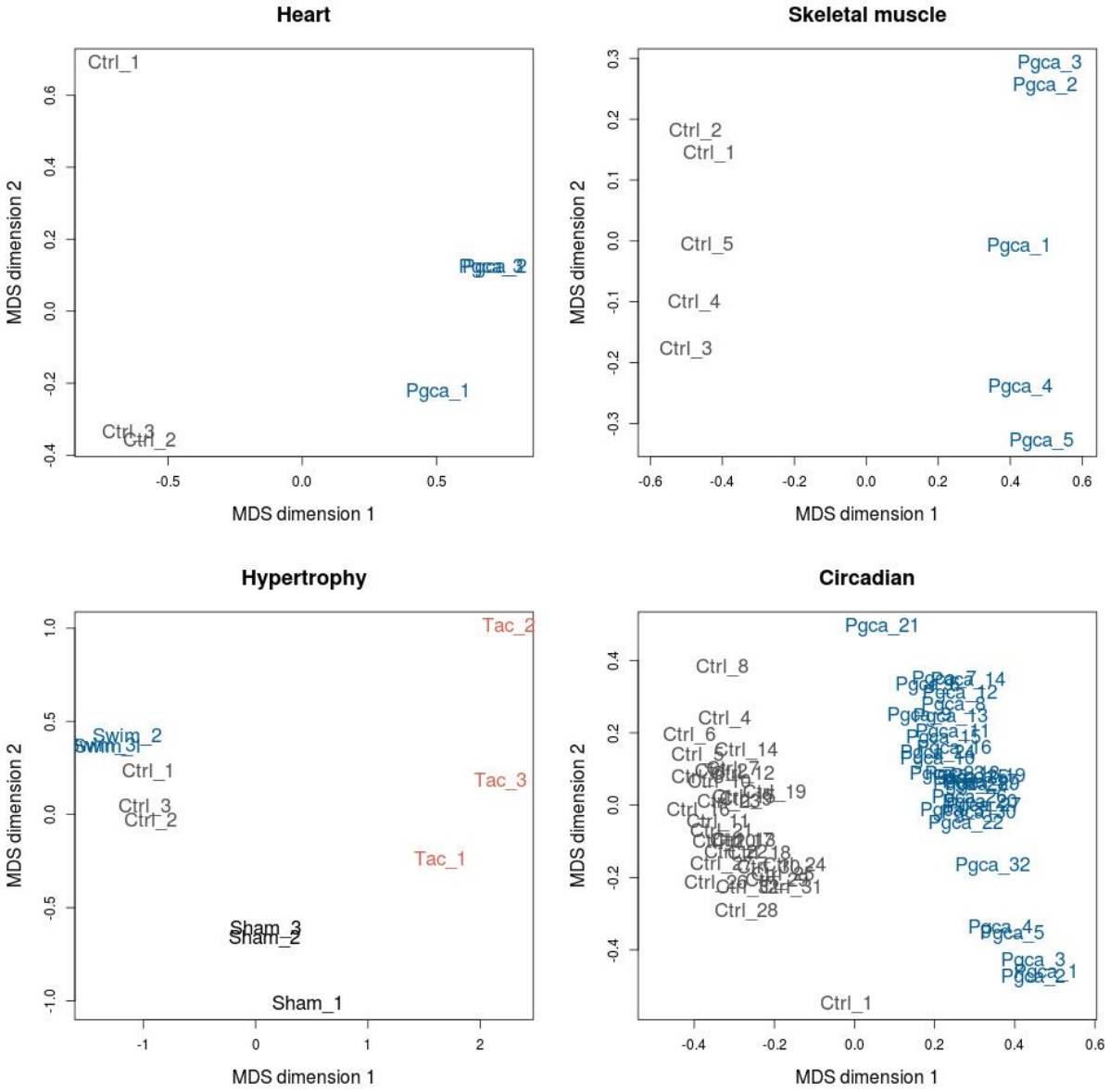

Figure S1. Multidimensional scaling plots of genome-wide datasets. The samples of the datasets represented as MDS plot. In A, B and D dark grey corresponds to control samples whereas blue corresponds to overexpression/knockout in heart, skeletal muscle and circadian datasets. In C, blue and dark grey correspond to physiological hypertrophy and its control whereas red and black correspond to pathological hypertrophy and its control in hypertrophy dataset, respectively.

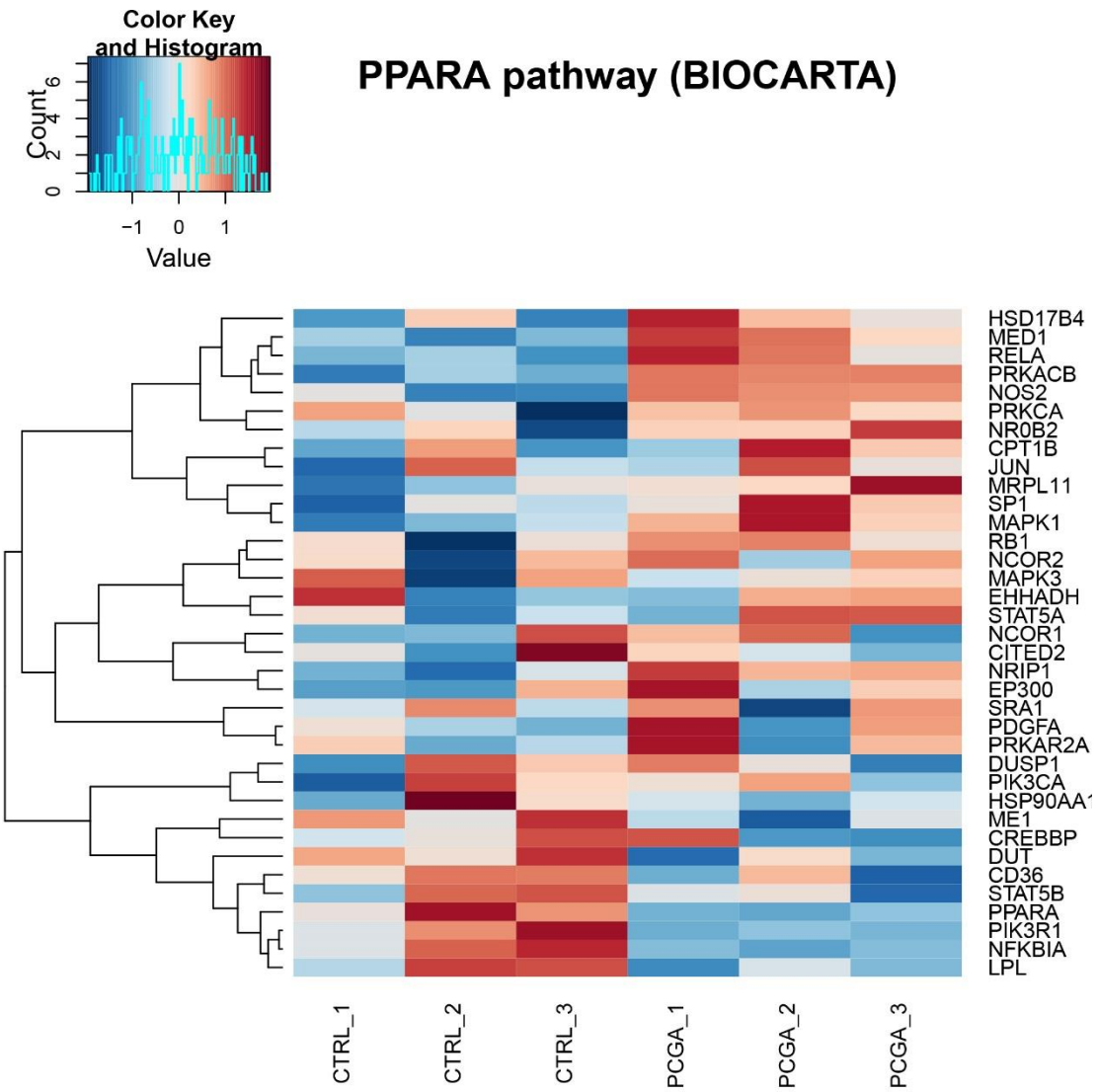

Figure S2. Expression levels of expressed genes in Pgc-1α overexpression in significantly enriched PPARA pathway (BIOCARTA). Hierarchical clustering of scaled expression levels of expressed genes in Pgc-1α overexpression in the heart in significantly enriched (adj. P-val. <0.05), downregulated pathway. The enrichment analysis was performed as explained in the materials and methods section. Rows correspond to genes and columns to samples whereas red and blue colors indicate the up- and down-regulation during Pgc-1α overexpression in heart.

Table S1. Upregulated, enriched pathways of Pgc-1α overexpression in heart. Enriched, upregulated pathways of Pgc-1a overexpression in heart with their respective adjusted P-values.

Pathway	**Adj.P.value**
Class a1 rhodopsin like receptors (REACTOME)	0.0026
Oxidative decarboxylation of alpha-ketoglutarate to succinyl CoA by alpha-ketoglutarate dehydrogenase (PWC)	0.0041
Citric acid cycle (TCA cycle) (PWC)	0.0041
Pentose phosphate pathway (hexose monophosphate shunt) (PWC)	0.0041
Insulin effects increased synthesis of Xylulose-5-Phosphate (PWC)	0.0041
PP2A-mediated dephosphorylation of key metabolic factors (PWC)	0.0041
Glucose + ATP => glucose-6-phosphate + ADP (PWC)	0.0046
Gluconeogenesis (PWC)	0.0045
Oxidative decarboxylation of alpha-ketoadipate to glutaryl CoA by alpha-ketoglutarate dehydrogenase (PWC)	0.0046
Fructose catabolism (PWC)	0.0047
Electron Transport Chain (WIKIPW)	0.0048
Glycolysis (PWC)	0.0048
ChREBP activates metabolic gene expression (PWC)	0.0050
Dihydroxyacetone phosphate is isomerized to form glyceraldehyde-3-phosphate (PWC)	0.0051
Glucose 6-phosphate is isomerized to form fructose-6-phosphate (PWC)	0.0051
Respiratory electron transport (REACTOME)	0.0051
Beta oxidation of butanoyl-CoA to acetyl-CoA (PWC)	0.0053
Glucose Regulation of Insulin Secretion (PWC)	0.0055
Glucose is carried across the plasma membrane by a glucose transport protein (GLUT) (PWC)	0.0057
Pyruvate metabolism (PWC)	0.0057
Pyruvate metabolism and TCA cycle (PWC)	0.0057
Oxidative phosphorylation (KEGG)	0.0061
Oxidative decarboxylation of alpha-keto-beta-methylvalerate to alpha-methylbutyryl-CoA by branched-chain alpha-ketoacid dehydrogenase (PWC)	0.0065
Alanine and aspartate metabolism (WIKIPW)	0.0065
Propionyl-CoA catabolism (PWC)	0.0071
Isoleucine catabolism (PWC)	0.0071
Regulation of Insulin Secretion (PWC)	0.0072
Transcriptional activation of glucose metabolism genes by ChREBP:MLX (PWC)	0.0077
Phosphoenolpyruvate and ADP react to form pyruvate and ATP (PWC)	0.0077
1,3-bisphosphoglycerate and ADP react to form 3-phosphoglycerate and ATP (PWC)	0.0077
Alanine metabolism (PWC)	0.0077
Oxidative decarboxylation of pyruvate to acetyl CoA by pyruvate dehydrogenase (PWC)	0.0085
Beta oxidation of hexanoyl-CoA to butanoyl-CoA (PWC)	0.0088

Glyceraldehyde 3-phosphate, NAD+, and orthophosphate react to form 1,3-bisphosphoglycerate, NADH, and H+ (PWC)	0.0106
Regulation of pyruvate dehydrogenase complex (PDC) (PWC)	0.0121
Electron transport chain (PWC)	0.0132
Oxidative decarboxylation of alpha-ketoisovalerate to isobutyryl-CoA by branched-chain alpha-ketoacid dehydrogenase (PWC)	0.0141
Lysine catabolism (PWC)	0.0152
Regulation of beta-cell development (PWC)	0.0152
GPCRs, class A rhodopsin-like (WIKIPW)	0.0156
Oxidative phosphorylation (WIKIPW)	0.0156
Metabolism of amino acids (PWC)	0.0157
Parkinsons disease (KEGG)	0.0158
Aspartate, asparagine, glutamate, and glutamine metabolism (PWC)	0.0161
Beta oxidation of octanoyl-CoA to hexanoyl-CoA (PWC)	0.0182
Beta oxidation of decanoyl-CoA to octanoyl-CoA-CoA (PWC)	0.0182
Branched-chain amino acid catabolism (PWC)	0.0186
Gpcr ligand binding (REACTOME)	0.0191
Valine catabolism (PWC)	0.0212
Peptide ligand binding receptors (REACTOME)	0.0229
Glucose uptake (PWC)	0.0246
Interferon alpha beta signaling (REACTOME)	0.0250
Regulation of gene expression in beta cells (PWC)	0.0280
Beta oxidation of myristoyl-CoA to lauroyl-CoA (PWC)	0.0281
Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA (PWC)	0.0281
Tca cycle and respiratory electron transport (REACTOME)	0.0334
Butanoate metabolism (KEGG)	0.0334
Metabolism of carbohydrates (PWC)	0.0350
Sensory perception (GO)	0.0357
Type II interferon signaling (IFNG) (WIKIPW)	0.0390
Mitochondrial protein import (REACTOME)	0.0403
Citrate cycle tca cycle (KEGG)	0.0454

Table S2. The pathways chosen for curated pathway analysis. The literature based on which the curated pathways were chosen, with the respective pathway, tissue in which the study was conducted, name of the article and author.

Curated pathway	Article name	Article author	Tissue
Growth signaling	Mitochondrial biogenesis in cardiac pathophysiology	Rimbaud et al.	heart
	Mechanisms regulating skeletal muscle growth and atrophy	Schiaffino et al.	skeletal muscle
Calcium signaling	Mitochondrial biogenesis and turnover	Diaz F. et al.	skeletal muscle
	Role of PGC-1a in signaling skeletal muscle health and disease	Kang et al.	skeletal muscle
	Mitochondrial biogenesis in cardiac pathophysiology	Rimbaud et al.	heart
Muscle contraction	Mitochondrial biogenesis and turnover	Diaz F. et al.	skeletal muscle
	PGC-1a: a key regulator of energy metabolism	Liang et al.	skeletal muscle
	Mitochondrial biogenesis in cardiac pathophysiology	Rimbaud et al.	heart
PI3K signaling	Molecular Defects in Cardiovascular Disease	Dhalla et al.	heart
	Signaling pathways controlling skeletal muscle mass	Egerman et al.	skeletal muscle
Glycolysis	PGC-1 coactivators: inducible regulators of energy metabolism in health and disease	Fink et al.	heart
	Skeletal muscle PGC-1α controls whole-body lactate homeostasis through estrogen-related receptor α-dependent activation of LDH B and repression of LDH A	Summermatter et al.	skeletal muscle
PPAR signaling	Role of PGC-1a in signaling skeletal muscle health and disease	Kang et al.	skeletal muscle
	The Coactivator PGC-1 Cooperates with Peroxisome Proliferator-Activated Receptor α in Transcriptional Control of Nuclear Genes Encoding Mitochondrial Fatty Acid Oxidation Enzymes	Vega et al.	heart
	PGC-1 coactivators: inducible regulators of energy metabolism in health and disease	Fink et al.	heart
Circadian rhythm	Transcriptional coactivator PGC-1a integrates the mammalian clock and energy metabolism	Liu et al.	skeletal muscle
	Circadian rhythms, Wnt/beta-catenin pathway and PPAR alpha/gamma	Lecarpentier	heart

	profiles in diseases with primary or secondary cardiac dysfunction		
Fatty acid beta-oxidation	The Coactivator PGC-1 Cooperates with Peroxisome Proliferator-Activated Receptor α in Transcriptional Control of Nuclear Genes Encoding Mitochondrial Fatty Acid Oxidation Enzymes	Vega et al.	heart
	PGC-1 coactivators: inducible regulators of energy metabolism in health and disease	Fink et al.	heart
	Regulation of skeletal muscle mitochondrial fatty acid metabolism in lean and obese individuals	Holloway et al.	skeletal muscle
Electron transport chain	PGC-1 coactivators: inducible regulators of energy metabolism in health and disease	Fink et al.	heart
	PGC-1 coactivators in cardiac development and disease	Rowe et al.	heart
	Regulation of skeletal muscle mitochondrial fatty acid metabolism in lean and obese individuals	Holloway et al.	skeletal muscle
Citric acid cycle	Metabolomic Analysis of the Skeletal Muscle of Mice Overexpressing PGC-1α	Hatazawa et al.	skeletal muscle
	PGC-1 coactivators in cardiac development and disease	Rowe et al.	heart

Table S3. Curated pathways of the pathway analysis with their respective databases.

Growth signaling	**PI3K signaling**	**Circadian rhythm**
Regulation of transforming growth factor beta receptor signaling pathway (GO)	IL2 signaling events mediated by PI3K (PWC)	Circadian rhythm mammal (KEGG)
Regulation of growth (GO)	Class I PI3K signaling events mediated by Akt (PWC)	Circadian rhythm pathway (PWC)
Negative regulation of growth (GO)	Class I PI3K signaling events (PWC)	Circadian rhythm (GO)
Epidermal growth factor receptor signaling pathway (GO)	Trk receptor signaling mediated by PI3K and PLC-gamma (PWC)	Diurnally regulated genes with circadian orthologs (WIKIPW)
Transforming growth factor beta receptor signaling pathway (GO)	PIK3 events in ERBB4 signaling (REACTOME)	Bmal1 Clock NPAS2 activates circadian expression (REACTOME)
Developmental growth (GO)	PIK3 events in ERBB2 signaling (REACTOME)	RORA activates circadian expression (REACTOME)
Regulation of cell growth (GO)	Negative regulation of the PI3K AKT network (REACTOME)	Circadian expression of expression by REV-ERBA (REACTOME)
Cleavage of growing transcript in the termination region (PWC)	PI3K AKT activation (REACTOME)	Circadian clock (REACTOME)
Signaling events activated by hepatocyte growth factor receptor (c-Met) (PWC)	G beta gamma signaling through PI3Kgamma (REACTOME)	
Growth (GO)	CD28 dependent PI3K AKT signaling (REACTOME)	
Signaling of hepatocyte growth factor receptor (WIKIPW)	PI3K cascade (REACTOME)	
Growth hormone receptor signaling (REACTOME)		
Regulation of insulin like growth factor IGF activity by insulin like growth factor binding proteins IGFBPS (REACTOME)		
Cleavage of growing transcript in the termination region (REACTOME)		
Beta-oxidation	**Muscle contradiction**	**Citric acid cycle**
Fatty acid beta oxidation (WIKIPW)	Cardiac muscle contraction (KEGG)	Citrate cycle tca cycle (KEGG)
Mitochondrial LC-fatty acid beta-oxidation (WIKIPW)	Muscle contraction (PWC)	Citric acid cycle (TCA cycle) (PWC)
Fatty acid beta oxidation (GO)	Regulation of muscle contraction (GO)	Pyruvate metabolism and TCA cycle (PWC)
Beta-oxidation of very long chain fatty acids (PWC)	Regulation of heart contraction (GO)	TCA cycle (WIKIPW)
Mitochondrial fatty acid beta-oxidation of unsaturated fatty acids (PWC)	Striated muscle contraction go 0006941 (GO)	Pyruate metabolism and citric acid TCA cycle (REACTOME)
Mitochondrial fatty acid beta-oxidation (PWC)	Striated muscle contraction (REACTOME)	TCA cycle and respiratory electron transport (REACTOME)
Activated AMPK stimulates fatty acid oxidation in muscle (REACTOME)	Muscle contraction (REACTOME)	Citric acid cycle TCA cycle (REACTOME)
Mitochondrial fatty acid beta oxidation (REACTOME)		

Ca signaling	**Electron transport chain**	**Glycolysis**
Calcium signaling pathway (KEGG)	Electron transport chain (WIKIPW)	Glycolysis gluconeogenesis (KEGG)
Calcium signaling in the CD4+TCR pathway (PWC)	Electron transport go 0006118 (GO)	Glycolysis pathway (BIOCARTA)
Calcium mediated signaling (GO)	Electron transport chain (PWC)	Glycolysis (PWC)
Calcium ion transport (GO)	TCA cycle and respiratory electron transport (REACTOME)	Glycolysis and Gluconeogenesis (WIKIPW)
Calcium independent cell cell adhesion (GO)	Respiratory electron transport ATP synthesis by chemiosmotic coupling and hear production by uncoupling proteins (REACTOME)	Glycolysis (REACTOME)
Calcium regulation in the cardiac cell (WIKIPW)		
PPAR signaling		
Ppar signaling pathway (KEGG)		
PPARA activates gene expression (REACTOME)		

Table S4. Upregulated, enriched pathways of Pgc-1α overexpression in skeletal muscle. Enriched, upregulated pathways of Pgc-1α overexpression in skeletal muscle with their respective adjusted P-values.

Pathway	**Adj. P-value**
Oxidative decarboxylation of alpha-ketoadipate to glutaryl CoA by alpha-ketoglutarate dehydrogenase (PWC)	0.0000
Oxidative decarboxylation of alpha-ketoglutarate to succinyl CoA by alpha-ketoglutarate dehydrogenase (PWC)	0.0000
Valine catabolism (PWC)	0.0000
Oxidative decarboxylation of alpha-ketoisovalerate to isobutyryl-CoA by branched-chain alpha-ketoacid dehydrogenase (PWC)	0.0000
Transcriptional activation of glucose metabolism genes by ChREBP:MLX (PWC)	0.0000
Regulation of gene expression in beta cells (PWC)	0.0000
Glycolysis (PWC)	0.0000
Phosphoenolpyruvate and ADP react to form pyruvate and ATP (PWC)	0.0000
Glyceraldehyde 3-phosphate, NAD+, and orthophosphate react to form 1,3-bisphosphoglycerate, NADH, and H+ (PWC)	0.0000
1,3-bisphosphoglycerate and ADP react to form 3-phosphoglycerate and ATP (PWC)	0.0000
Dihydroxyacetone phosphate is isomerized to form glyceraldehyde-3-phosphate (PWC)	0.0000
Glucose 6-phosphate is isomerized to form fructose-6-phosphate (PWC)	0.0000
Citric acid cycle (TCA cycle) (PWC)	0.0000

Propionyl-CoA catabolism (PWC)	0.0000
Regulation of insulin secretion (PWC)	0.0000
Glucose regulation of insulin secretion (PWC)	0.0000
Oxidative decarboxylation of pyruvate to acetyl CoA by pyruvate dehydrogenase (PWC)	0.0000
Electron transport chain (PWC)	0.0000
Pentose phosphate pathway (hexose monophosphate shunt) (PWC)	0.0000
Metabolism of lipids and lipoproteins (PWC)	0.0000
Mitochondrial fatty acid beta-Oxidation (PWC)	0.0000
Mitochondrial fatty acid beta-oxidation of unsaturated fatty acids (PWC)	0.0000
Oxidative decarboxylation of alpha-keto-beta-methylvalerate to alpha-methylbutyryl-CoA by branched-chain alpha-ketoacid dehydrogenase (PWC)	0.0000
Gluconeogenesis (PWC)	0.0000
Lysine catabolism (PWC)	0.0000
Glucose metabolism (PWC)	0.0000
Pyruvate metabolism (PWC)	0.0000
Glucose uptake (PWC)	0.0000
Glucose is carried across the plasma membrane by a glucose transport protein (GLUT) (PWC)	0.0000
Glucose + ATP => glucose-6-phosphate + ADP (PWC)	0.0000
Isoleucine catabolism (PWC)	0.0000
ChREBP activates metabolic gene expression (PWC)	0.0000
Regulation of beta-cell development (PWC)	0.0000
Regulation of pyruvate dehydrogenase complex (PDC) (PWC)	0.0000
Pyruvate metabolism and TCA cycle (PWC)	0.0000
Insulin effects increased synthesis of Xylulose-5-Phosphate (PWC)	0.0000
Activated AMPK stimulates fatty-acid oxidation in muscle (PWC)	0.0000
Fructose catabolism (PWC)	0.0000
Branched-chain amino acid catabolism (PWC)	0.0000
Alanine metabolism (PWC)	0.0000
Integration of energy metabolism (PWC)	0.0000
PP2A-mediated dephosphorylation of key metabolic factors (PWC)	0.0000
Diabetes pathways (PWC)	0.0000
Metabolism of amino acids (PWC)	0.0000
Tca cycle and respiratory electron transport (REACTOME)	0.0000
Oxidative phosphorylation (KEGG)	9.9168e-05
Huntingtons disease (KEGG)	0.0001
Metabolism of carbohydrates (PWC)	0.0001
mitochondrial fatty acid beta-oxidation of saturated fatty acids (PWC)	0.0001
Beta oxidation of palmitoyl-CoA to myristoyl-CoA (PWC)	0.0001
Import of palmitoyl-CoA into the mitochondrial matrix (PWC)	0.0001

Alzheimers disease (KEGG)	0.0001
Aspartate, asparagine, glutamate, and glutamine metabolism (PWC)	0.0001
Beta oxidation of octanoyl-CoA to hexanoyl-CoA (PWC)	0.0002
Beta oxidation of hexanoyl-CoA to butanoyl-CoA (PWC)	0.0002
Beta oxidation of butanoyl-CoA to acetyl-CoA (PWC)	0.0002
Beta oxidation of myristoyl-CoA to lauroyl-CoA (PWC)	0.0002
Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA (PWC)	0.0002
Beta oxidation of decanoyl-CoA to octanoyl-CoA-CoA (PWC)	0.0002
Parkinsons disease (KEGG)	0.0002
Electron transport chain (WIKIPW)	0.0002
Oxidative phosphorylation (WIKIPW)	0.0009
Respiratory electron transport atp synthesis by chemiosmotic coupling and heat production by uncoupling proteins (REACTOME)	0.0028
Valine leucine and isoleucine degradation (KEGG)	0.0101
Metabolism of amino acids and derivatives (REACTOME)	0.0256
Citrate cycle tca cycle (KEGG)	0.0411
TCA Cycle (WIKIPW)	0.0411
Metabolism of lipids and lipoproteins (REACTOME)	0.0461
Ppara activates gene expression (REACTOME)	0.0487

Table S5. Upregulated, enriched pathways of pathological hypertrophy. Enriched, upregulated pathways of pathological hypertrophy in skeletal muscle with their respective adjusted P-values.

Pathway	**Adj. P-value**
Dna replication (KEGG)	0.0000
Cell cycle (KEGG)	0.0000
Focal adhesion (KEGG)	0.0000
Ecm receptor interaction (KEGG)	0.0000
Leukocyte transendothelial migration (KEGG)	0.0000
Regulation of actin cytoskeleton (KEGG)	0.0000
M Phase (PWC)	0.0000
Mitotic Prometaphase (PWC)	0.0000
M/G1 Transition (PWC)	0.0000
Activation of the pre-replicative complex (PWC)	0.0000
DNA strand elongation (PWC)	0.0000
Unwinding of DNA (PWC)	0.0000
G1/S Transition (PWC)	0.0000
S Phase (PWC)	0.0000
Synthesis of DNA (PWC)	0.0000
Cell cycle, mitotic (PWC)	0.0000

DNA replication pre-initiation (PWC)	0.0000
Integrin cell surface interactions (PWC)	0.0000
Hemostasis (PWC)	0.0000
Formation of Platelet plug (PWC)	0.0000
DNA Replication (PWC)	0.0000
Aurora B signaling (PWC)	0.0000
Signaling by Aurora kinases (PWC)	0.0000
FOXM1 transcription factor network (PWC)	0.0000
Cytokinesis (GO)	0.0000
M phase (GO)	0.0000
Regulation of mitosis (GO)	0.0000
Cell cycle process (GO)	0.0000
Mitotic cell cycle (GO)	0.0000
Cell cycle phase (GO)	0.0000
Cell division (GO)	0.0000
Mitosis (GO)	0.0000
M phase of mitotic cell cycle (GO)	0.0000
Cell cycle (WIKIPW)	0.0000
Focal Adhesion (WIKIPW)	0.0000
Inflammatory response pathway (WIKIPW)	0.0000
DNA replication (WIKIPW)	0.0000
Activation of the pre replicative complex (REACTOME)	0.0000
Cell cycle (REACTOME)	0.0000
Extracellular matrix organization (REACTOME)	0.0000
Collagen formation (REACTOME)	0.0000
Chondroitin sulfate dermatan sulfate metabolism (REACTOME)	0.0000
Glycosaminoglycan metabolism (REACTOME)	0.0000
Mhc class ii antigen presentation (REACTOME)	0.0000
Integrin cell surface interactions (REACTOME)	0.0000
Cell cycle mitotic (REACTOME)	0.0000
Axon guidance (REACTOME)	0.0000
Synthesis of dna (REACTOME)	0.0000
Mitotic g1 g1 s phases (REACTOME)	0.0000
Mitotic m m g1 phases (REACTOME)	0.0000
Kinesins (REACTOME)	0.0000
Dna replication (REACTOME)	0.0000
Activation of atr in response to replication stress (REACTOME)	0.0000
Mitotic prometaphase (REACTOME)	0.0000
G2 m checkpoints (REACTOME)	0.0000
S phase (REACTOME)	0.0000
Dna strand elongation (REACTOME)	0.0000
Orc1 removal from chromatin (REACTOME)	0.0005
Platelet Activation (PWC)	0.0006
Unwinding of dna (REACTOME)	0.0006
Prostaglandin Synthesis and Regulation (WIKIPW)	0.0006
G1 to S cell cycle control (WIKIPW)	0.0006

G1 s transition (REACTOME)	0.0006
Mitotic cell cycle checkpoint (GO)	0.0006
Mitotic sister chromatid segregation (GO)	0.0006
Cell adhesion molecules cams (KEGG)	0.0006
G2 pathway (BIOCARTA)	0.0006
Immune system process (GO)	0.0007
Sister chromatid segregation (GO)	0.0007
Cell surface interactions at the vascular wall (PWC)	0.0007
G1 Phase (PWC)	0.0007
Cyclin D associated events in G1 (PWC)	0.0007
G1 pathway (BIOCARTA)	0.0007
Cell Cycle Checkpoints (PWC)	0.0009
E2F transcriptional targets at G1/S (PWC)	0.0009
E2F mediated regulation of DNA replication (PWC)	0.0009
Cell cycle go 0007049 (GO)	0.0009
Defense response (GO)	0.0010
Response to elevated platelet cytosolic ca2 (REACTOME)	0.0013
Integrin-mediated cell adhesion (WIKIPW)	0.0013
Chromosome segregation (GO)	0.0013
Signal transduction by l1 (REACTOME)	0.0013
Cell proliferation go 0008283 (GO)	0.0013
L1cam interactions (REACTOME)	0.0014
Platelet activation signaling and aggregation (REACTOME)	0.0014
Cell cycle checkpoints (REACTOME)	0.0015
A tetrasaccharide linker sequence is required for gag synthesis (REACTOME)	0.0015
Semaphorin interactions (REACTOME)	0.0016
Assembly of the pre replicative complex (REACTOME)	0.0018
Generation of second messenger molecules (REACTOME)	0.0018
M g1 transition (REACTOME)	0.0019
Activation of ATR in response to replication stress (PWC)	0.0020
Assembly of the pre-replicative complex (PWC)	0.0021
Signaling in Immune system (PWC)	0.0021
G2/M Checkpoints (PWC)	0.0023
Switching of origins to a post-replicative state (PWC)	0.0023
Orc1 removal from chromatin (PWC)	0.0023
Regulation of DNA replication (PWC)	0.0023
Mcm pathway (BIOCARTA)	0.0024
G1 s specific transcription (REACTOME)	0.0027
Cytoskeleton organization and biogenesis (GO)	0.0028
Immunoregulatory interactions between a lymphoid and a non lymphoid cell (REACTOME)	0.0031
Developmental biology (REACTOME)	0.0032
Hemostasis (REACTOME)	0.0033
Ncam1 interactions (REACTOME)	0.0034
Integrin alphaiib beta3 signaling (REACTOME)	0.0036
Skeletal development (GO)	0.0037

Telomere maintenance (REACTOME)	0.0037
Removal of licensing factors from origins (PWC)	0.0038
Granulocytes pathway (BIOCARTA)	0.0038
E2f mediated regulation of dna replication (REACTOME)	0.0042
amb2 Integrin signaling (PWC)	0.0043
Chondroitin sulfate biosynthesis (REACTOME)	0.0043
Lipoprotein metabolism (REACTOME)	0.0044
Pyrimidine metabolism (PWC)	0.0049
Cell migration (GO)	0.0049
Eukaryotic Translation Elongation (PWC)	0.0050
Cell matrix adhesion (GO)	0.0050
Positive regulation of cell proliferation (GO)	0.0052
Cell cell adhesion (GO)	0.0053
T Cell Receptor Signaling Pathway (WIKIPW)	0.0054
Lysosome (KEGG)	0.0054
Leukocyte migration (GO)	0.0059
Immune response (GO)	0.0059
Alpha6Beta4Integrin (PWC)	0.0060
Fc gamma r mediated phagocytosis (KEGG)	0.0060
Ncam signaling for neurite out growth (REACTOME)	0.0060
Glycosaminoglycan biosynthesis chondroitin sulfate (KEGG)	0.0061
Cell cell communication (REACTOME)	0.0062
P53 pathway (BIOCARTA)	0.0063
Interferon alpha beta signaling (REACTOME)	0.0064
Cell substrate adhesion (GO)	0.0064
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell (PWC)	0.0065
Ucalpain pathway (BIOCARTA)	0.0066
Chromosome condensation (GO)	0.0066
Organ development (GO)	0.0066
Heparan sulfate heparin hs gag metabolism (REACTOME)	0.0071
Platelet activation triggers (PWC)	0.0073
Cyclin a b1 associated events during g2 m transition (REACTOME)	0.0076
IL4-mediated signaling events (PWC)	0.0077
Other glycan degradation (KEGG)	0.0099
Chromosome maintenance (REACTOME)	0.0101
Extension of telomeres (REACTOME)	0.0105
Signaling by robo receptor (REACTOME)	0.0107
G0 and early g1 (REACTOME)	0.0108
Organ morphogenesis (GO)	0.0109
The role of nef in hiv1 replication and disease pathogenesis (REACTOME)	0.0110
Cellcycle pathway (BIOCARTA)	0.0110
Srcrptp pathway (BIOCARTA)	0.0113
Pathogenic escherichia coli infection (KEGG)	0.0114
IL-5 Signaling Pathway (WIKIPW)	0.0120
Peptide chain elongation (REACTOME)	0.0125

Peptide chain elongation (PWC)	0.0133
Immune system (REACTOME)	0.0134
Lagging strand synthesis (REACTOME)	0.0152
Lym pathway (BIOCARTA)	0.0156
IL-3 Signaling Pathway (WIKIPW)	0.0157
Signal regulatory protein sirp family interactions (REACTOME)	0.0158
Natural killer cell mediated cytotoxicity (KEGG)	0.0158
L13a-mediated translational silencing of Ceruloplasmin expression (PWC)	0.0159
Translation initiation complex formation (PWC)	0.0159
3' -UTR-mediated translational regulation (PWC)	0.0159
Translation (PWC)	0.0160
Eukaryotic translation initiation (PWC)	0.0160
Cap-dependent translation initiation (PWC)	0.0160
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S (PWC)	0.0160
Formation of a pool of free 40S subunits (PWC)	0.0160
Pyrimidine salvage reactions (PWC)	0.0161
Epidermis development (GO)	0.0162
Cytokine cytokine receptor interaction (KEGG)	0.0166
Peptidyl tyrosine modification (GO)	0.0170
Alpha6-Beta4 integrin signaling pathway (WIKIPW)	0.0171
Recycling pathway of l1 (REACTOME)	0.0171
B cell receptor signaling pathway (WIKIPW)	0.0172
Cell cycle checkpoint go 0000075 (GO)	0.0180
Innate immune system (REACTOME)	0.0180
Tissue development (GO)	0.0181
Mitotic spindle organization and biogenesis (GO)	0.0183
Spindle organization and biogenesis (GO)	0.0183
Gpvi mediated activation cascade (REACTOME)	0.0183
Cell surface interactions at the vascular wall (REACTOME)	0.0184
P53 signaling pathway (KEGG)	0.0187
Nkcells pathway (BIOCARTA)	0.0187
Plateletapp pathway (BIOCARTA)	0.0187
Leishmania infection (KEGG)	0.0188
Tnfr1 pathway (BIOCARTA)	0.0192
Leading strand synthesis (PWC)	0.0194
Lagging strand synthesis (PWC)	0.0194
VEGFR3 signaling in lymphatic endothelium (PWC)	0.0195
Hematopoietic cell lineage (KEGG)	0.0196
Leukocyte activation (GO)	0.0197
G1 s transition of mitotic cell cycle (GO)	0.0199
3 utr mediated translational regulation (REACTOME)	0.0206
Inflammasomes (REACTOME) Mpr pathway (BIOCARTA)	0.0206 0.0206
Formation of tubulin folding intermediates by cct tric (REACTOME)	0.0209
Regulation of phosphorylation (GO)	0.0211

Positive regulation of cell cycle (GO)	0.0215
Microtubule cytoskeleton organization and biogenesis (GO)	0.0218
Ranms pathway (BIOCARTA)	0.0225
Cytoplasmic ribosomal proteins (WIKIPW)	0.0227
Antigen processing and presentation (KEGG)	0.0229
Ace2 pathway (BIOCARTA)	0.0232
Type II interferon signaling (IFNG) (WIKIPW)	0.0234
Monocyte pathway (BIOCARTA)	0.0236
Tel pathway (BIOCARTA)	0.0237
Pyrimidine biosynthesis (interconversion) (PWC)	0.0245
Interphase of mitotic cell cycle (GO)	0.0255
Interphase (GO)	0.0263
Regulation of lymphocyte activation (GO)	0.0268
Signaling by pdgf (REACTOME)	0.0274
Multicellular organismal development (GO)	0.0275
Reversible phosphorolysis of pyrimidine nucleosides (PWC)	0.0286
Systemic lupus erythematosus (KEGG)	0.0289
Regulation of actin cytoskeleton (WIKIPW)	0.0289
Intrinsic pathway (BIOCARTA)	0.0319
Cell activation (GO)	0.0322
Degradation of the extracellular matrix (REACTOME)	0.0324
Dc pathway (BIOCARTA)	0.0343
Interferon signaling (REACTOME)	0.0346
Peptidyl tyrosine phosphorylation (GO)	0.0348
Reversible phosphorolysis of pyrimidine nucleosides by uridine phosphorylase 1 (PWC)	0.0349
Regulation of i kappab kinase nf kappab cascade (GO)	0.0349
Lair pathway (BIOCARTA)	0.0351
DNA replication initiation (PWC)	0.0351
Hs gag biosynthesis (REACTOME)	0.0353
Establishment of organelle localization (GO)	0.0354
Hdl mediated lipid transport (REACTOME)	0.0354
Platelet aggregation plug formation (REACTOME)	0.0366
Adherens junction (KEGG)	0.0383
Ectoderm development (GO)	0.0386
Grb2 sos provides linkage to mapk signaling for intergrins (REACTOME)	0.0393
Myeloid leukocyte differentiation (GO)	0.0409
Dna packaging (GO)	0.0410
Positive regulation of lymphocyte activation (GO)	0.0410
Nucleosome assembly (GO)	0.0411
Actin filament based process (GO)	0.0412
Regulation of cell proliferation (GO)	0.0414
Cellular defense response (GO)	0.0415
Senescence and autophagy (WIKIPW)	0.0416
Telomere extension by telomerase (PWC)	0.0416
Interferon gamma signaling (REACTOME)	0.0420

Regulation of t cell activation (GO)	0.0423
Sema3a plexin repulsion signaling by inhibiting integrin adhesion (REACTOME)	0.0428
Regulation of ifna signaling (REACTOME)	0.0430
Repair synthesis for gap filling by dna pol in tc ner (REACTOME)	0.0433
Tgf beta signaling pathway (KEGG)	0.0433
Response to biotic stimulus (GO)	0.0434
Eukaryotic translation termination (PWC)	0.0436
System development (GO)	0.04376
The role of Nef in HIV-1 replication and disease pathogenesis (PWC)	0.04641
BARD1 signaling events (PWC)	0.04761
Vasculature development (GO)	0.04782
Endochondral ossification (WIKIPW)	0.04787
Cell death signalling via nrage nrif and nade (REACTOME)	0.04808
APC-Cdc20 mediated degradation of Nek2A (PWC)	0.04814
Golgi associated vesicle biogenesis (REACTOME)	0.04871
Regulation of peptidyl tyrosine phosphorylation (GO)	0.0488
Deposition of new cenpa containing nucleosomes at the centromere (REACTOME)	0.0491
TGFBR (PWC)	0.0491
Cs ds degradation (REACTOME)	0.0492
The nlrp3 inflammasome (REACTOME)	0.0492
Activation of the mrna upon binding of the cap binding complex and eifs and subsequent binding to 43s (REACTOME)	0.0494
Cell junction organization (REACTOME)	0.0495
Adaptive immune system (REACTOME)	0.0500

Table S6. Upregulated, enriched pathways of physiological hypertrophy. Enriched, upregulated pathways of physiological with their respective adjusted P-values.

Pathway	**Adj. P-value**
Oxidative decarboxylation of alpha-ketoadipate to glutaryl CoA by alpha-ketoglutarate dehydrogenase (PWC)	0.0000
Oxidative decarboxylation of alpha-ketoglutarate to succinyl CoA by alpha-ketoglutarate dehydrogenase (PWC)	0.0000
Beta oxidation of octanoyl-CoA to hexanoyl-CoA (PWC)	0.0000
Beta oxidation of hexanoyl-CoA to butanoyl-CoA (PWC)	0.0000
Beta oxidation of butanoyl-CoA to acetyl-CoA (PWC)	0.0000
Beta oxidation of myristoyl-CoA to lauroyl-CoA (PWC)	0.0000
Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA (PWC)	0.0000
Beta oxidation of decanoyl-CoA to octanoyl-CoA-CoA (PWC)	0.0000
Valine catabolism (PWC)	0.0000
Oxidative decarboxylation of alpha-ketoisovalerate to isobutyryl-CoA by branched-chain alpha-ketoacid dehydrogenase (PWC)	0.0000
Transcriptional activation of glucose metabolism genes by ChREBP:MLX (PWC)	0.0000
Glycolysis (PWC)	0.0000
Phosphoenolpyruvate and ADP react to form pyruvate and ATP (PWC)	0.0000
Glyceraldehyde 3-phosphate, NAD+, and orthophosphate react to form 1,3-bisphosphoglycerate, NADH, and H+ (PWC)	0.0000
1,3-bisphosphoglycerate and ADP react to form 3-phosphoglycerate and ATP (PWC)	0.0000
Dihydroxyacetone phosphate is isomerized to form glyceraldehyde-3-phosphate (PWC)	0.0000
Glucose 6-phosphate is isomerized to form fructose-6-phosphate (PWC)	0.0000
tRNA aminoacylation (PWC)	0.0000
Citric acid cycle (TCA cycle) (PWC)	0.0000
Propionyl-CoA catabolism (PWC)	0.0000
Glucose regulation of insulin secretion (PWC)	0.0000
Oxidative decarboxylation of pyruvate to acetyl CoA by pyruvate dehydrogenase (PWC)	0.0000
Electron transport chain (PWC)	0.0000
Pentose phosphate pathway (hexose monophosphate shunt) (PWC)	0.0000
Mitochondrial fatty acid beta-oxidation (PWC)	0.0000
Mitochondrial fatty acid beta-oxidation of unsaturated fatty acids (PWC)	0.0000
Oxidative decarboxylation of alpha-keto-beta-methylvalerate to alpha-methylbutyryl-CoA by branched-chain alpha-ketoacid dehydrogenase (PWC)	0.0000
Glucose metabolism (PWC)	0.0000
Pyruvate metabolism (PWC)	0.0000
Glucose uptake (PWC)	0.0000

Glucose is carried across the plasma membrane by a glucose transport protein (GLUT) (PWC)	0.0000
Glucose + ATP => glucose-6-phosphate + ADP (PWC)	0.0000
Isoleucine catabolism (PWC)	0.0000
ChREBP activates metabolic gene expression (PWC)	0.0000
Regulation of pyruvate dehydrogenase complex (PDC) (PWC)	0.0000
Pyruvate metabolism and TCA cycle (PWC)	0.0000
Insulin effects increased synthesis of Xylulose-5-Phosphate (PWC)	0.0000
Fructose catabolism (PWC)	0.0000
Branched-chain amino acid catabolism (PWC)	0.0000
Alanine metabolism (PWC)	0.0000
PP2A-mediated dephosphorylation of key metabolic factors (PWC)	0.0000
Metabolism of carbohydrates (PWC)	0.0000
Electron transport chain (WIKIPW)	0.0000
Tca cycle and respiratory electron transport (REACTOME)	0.0000
Gluconeogenesis (REACTOME)	0.0000
Trna aminoacylation (REACTOME)	0.0000
Respiratory electron transport (REACTOME)	0.0000
Respiratory electron transport atp synthesis by chemiosmotic coupling and heat production by uncoupling proteins (REACTOME)	0.0000
Glucose metabolism (REACTOME)	0.0000
Aminoacyl trna biosynthesis (KEGG)	0.0003
Aspartate, asparagine, glutamate, and glutamine metabolism (PWC)	0.0003
Glycolysis and gluconeogenesis (WIKIPW)	0.00036
Regulation of insulin secretion (PWC)	0.00036
Oxidative phosphorylation (KEGG)	0.00037
Gluconeogenesis (PWC)	0.00037
Mitochondrial fatty acid beta-oxidation of saturated fatty acids (PWC)	0.00038
Beta oxidation of palmitoyl-CoA to myristoyl-CoA (PWC)	0.00038
Mitochondrial trna aminoacylation (REACTOME)	0.00169
Citrate cycle tca cycle (KEGG)	0.00298
Alanine and aspartate metabolism (WIKIPW)	0.00304
Citric acid cycle tca cycle (REACTOME)	0.00305
Cellular respiration (GO)	0.00404
Lysine catabolism (PWC)	0.00485
TCA Cycle (WIKIPW)	0.00493
Alanine aspartate and glutamate metabolism (KEGG)	0.00502
Parkinsons disease (KEGG)	0.00513
Integration of energy metabolism (PWC)	0.00522
Oxidative phosphorylation (WIKIPW)	0.00526
Aerobic respiration (GO)	0.00534
Activated AMPK stimulates fatty-acid oxidation in muscle (PWC)	0.00543
Regulation of gene expression in beta cells (PWC)	0.00585
Regulation of beta-cell development (PWC)	0.00594
Mitochondrial tRNA aminoacylation (PWC)	0.00603

Huntingtons disease (KEGG)	0.00613
Mitochondrial protein import (REACTOME)	0.00674
Import of palmitoyl-CoA into the mitochondrial matrix (PWC)	0.01004
Pyruvate metabolism and citric acid tca cycle (REACTOME)	0.01311
Regulation of heart contraction (GO)	0.01985
Valine leucine and isoleucine biosynthesis (KEGG)	0.02247
Energy derivation by oxidation of organic compounds (GO)	0.03252
Propanoate metabolism (KEGG)	0.03284
Amino acid catabolic process (GO)	0.03765
Formation of ATP by chemiosmotic coupling (PWC)	0.03774
Starch and sucrose metabolism (KEGG)	0.03784
Diabetes pathways (PWC)	0.03802
Neurotransmitter uptake and metabolism in glial cells (PWC)	0.03813
Metabolism of amino acids (PWC)	0.03824
Insulin-mediated glucose transport (PWC)	0.04124
Cofactor catabolic process (GO)	0.04248
Igf1mtor pathway (BIOCARTA)	0.04539
Glycolysis gluconeogenesis (KEGG)	0.04813
Amino acid synthesis and interconversion transamination (REACTOME)	0.04978

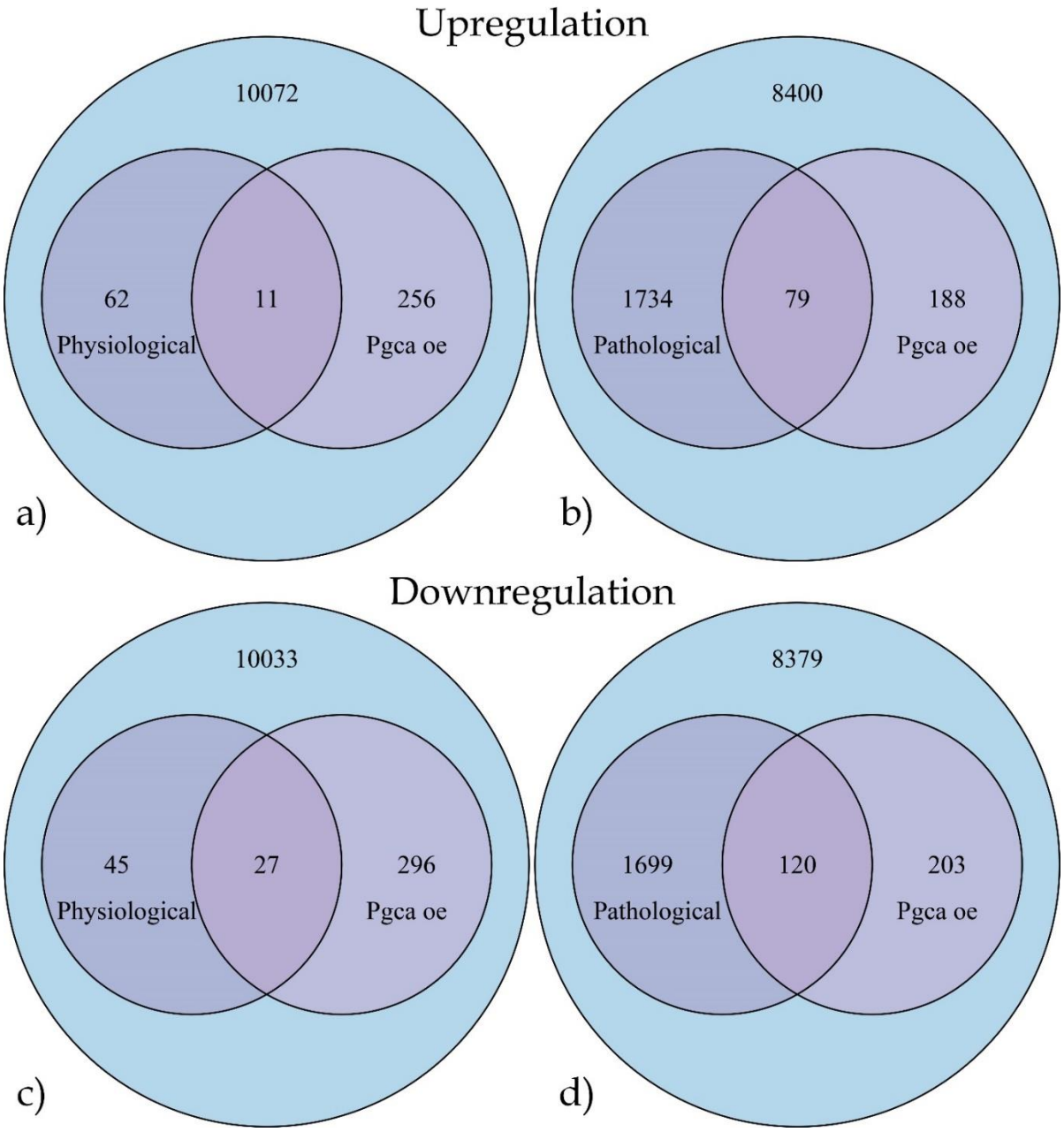

Figure S2. Hypergeometric distribution of Pgc-1α overexpression and physiological/ pathological hypertrophy. The differentially expressed genes (adj. P-val. <0.05) of Pgc-1α overexpression were compared with differentially expressed genes of physiological and pathological cardiomyopathy in comparison to expressed genes across both experiments, separated by up- and downregulation. The respective p-values were the following: a) 2.92^-07; b) 2.45^-07; c) 0.00; d) 0.00.