Summer 2010

Coex-rank: an approach for microarray combined analysis - applications to PPARγ related datasets

Jinlu Cai
University of Iowa

COEX-RANK: AN APPROACH FOR MICROARRAY COMBINED ANALYSIS – APPLICATIONS TO PPARγ RELATED DATASETS

by

Jinlu Cai

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Genetics (Computational Genetics)
in the Graduate College of
The University of Iowa

July 2010

Thesis Supervisors: Professor Thomas L. Casavant
Professor Curt D. Sigmund

ABSTRACT

Microarrays have been widely used to study differential gene expression at the genomic level. They can also provide genome-wide co-expression information. Robust approaches are needed for integration and validation of independently-collected datasets which may contribute to a common hypothesis. Previously, attempts at meta-analysis have contributed to solutions to this problem. As an alternative, for microarray data from multiple highly similar biological experimental designs, a more direct combined approach is possible. In this thesis, a novel approach is described for *microarray combined analysis*, including gene-level unification into a virtual platform followed by normalization and a method for ranking candidate genes based on co-expression information – called *Coex-Rank*. We applied this approach to our Sppar (a PPARγ mutant) dataset, which illustrated an improvement in statistical power and a complementary advantage of the Coex-Rank method from a biological perspective.

We also performed analysis to other PPARγ-related microarray datasets. From the perspective of gene sets, we observed that up-regulated genes from mice treated with the PPARγ ligand rosiglitazone were significantly down-regulated in mice with a global knock-in dominant-negative mutation of PPARγ. Integrated with publicly available PPRE (PPAR Response Element) datasets, we found that the genes which were most up-regulated by rosiglitazone treatment and which were also down-regulated by the global knock-in mutation of PPARγ were robustly enriched in PPREs near transcription start sites. In addition, we identified several potential PPARγ targets in the aorta and mesenteric artery for further experimental validation, such as Rhobtb1 and Rgs5.

Abstract Approved: ______________________________
Thesis Co-supervisor

Professor, Electrical and Computer Engineering
____
Title and Department

____
Date

____
Thesis Co-supervisor

Professor, Internal Medicine
____
Title and Department

____
Date

COEX-RANK: AN APPROACH FOR MICROARRAY COMBINED ANALYSIS – APPLICATIONS TO PPARγ RELATED DATASETS

by

Jinlu Cai

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Genetics (Computational Genetics)
in the Graduate College of
The University of Iowa

July 2010

Thesis Supervisors: Professor Thomas L. Casavant
Professor Curt D. Sigmund

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

____

PH.D. THESIS

____

This is to certify that the Ph.D. thesis of

Jinlu Cai

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Genetics (Computational Genetics)
at the July 2010 graduation.

Thesis Committee: ___________________________________
Thomas L. Casavant, Thesis Co-Supervisor

____
Curt D. Sigmund, Thesis Co-Supervisor

____
Terry A. Braun

____
Yi Xing

____
Andrew W. Norris

ACKNOWLEDGEMENTS

First of all, I would like to thank my thesis co-advisors – Professors Thomas Casavant and Curt Sigmund. I feel fortunate to join their labs and contribute to this exciting project -- a perfect mix of developing a general computational method and solving a specific biological problem. Both of them have been excellent mentors. I am influenced by their passion for discovery, and exploring the field of Computational Genetics. Only with their guidance and encouragement was I able to accomplish this thesis for my Ph.D. degree.

I would like to express my special thanks to my committee members, Terry Braun, Yi Xing and Andrew Norris. They all have broad knowledge and deep insight in Bioinformatics. With their continuous help, I have integrated better computational approaches into my analysis.

I would like to thank Henry Keen, my supervisor, as well. He has outstanding background in both Biology and Computer Science. He taught me how to use explore a wide range of bioinformatics tools and explained complicated biological models to me in details, introducing me into the amazing world of Bioinformatics.

Moreover, I would like to thank all the current and former members in both the Casavant and Sigmund labs. They shared knowledge from different areas with me, from Computer Science, Engineering and Statistics to Biology. We had many fruitful discussions regarding both research and life. We had fun together.

Finally, I would thank my parents and fiancé, for their self-giving love, for their understanding, for their continuous support. They are always by my side.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

High-throughput microarray technologies have become popular for genome-wide investigation of gene expression profiles. Careful experimental design followed by a variety of proper computational analyses can reveal interaction of genes and related biological pathways [1]. For data analysis, a common goal is to detect differentially expressed genes (DEGs) [1] between controls and cases or in response to specific factors, such as time and dose effects. Different laboratories may carry out microarray experiments with related biological experimental design, but using different types of platforms. Due to the high cost of microarrays, many studies suffer from the problem of small sample size, which may lead to high false discovery rates (FDR) or low sensitivity in determination of DEGs [2]. Combining related but independent microarray datasets increases sample size and theoretically would result in higher reliability of novel gene candidate discovery from a purely statistical view [3]. Such a "combined approach" may be able to detect small but consistent changes. In fact, this is one of the motivating factors for the construction of public microarray databases, such as Gene Expression Omnibus (GEO) [4]. Alternately, successful combined analyses may demonstrate the reproducibility of independent studies [5], which is a fundamental issue in validation of biological experiments.

For combined analysis of microarray studies, however, complications often arise from biological variations and technical differences [6]. Meta-analysis, which has been well-studied in statistics, is a practical way to solve this problem [6]. It involves combining results from independent but related studies. The application of meta-analysis to microarray data has been demonstrated by different groups, yet no consensus has been reached as to the best method. Hong et al. evaluated the performance of different microarray meta-analysis methods and recommended approaches derived from two different philosophies. One is a t-based modeling approach [6], and the other is a rank-product approach. The latter has the advantage of robustness in ranking genes over the t-based method, but only provides relative prioritization of genes [6]. The rank-product approach has also been shown to have similar performance to the rank-aggregation method. [6]

As an alternative to meta-analysis, a more direct combined approach is also possible for datasets resulting from experiments with highly similar biological design. With the continued development of microarray technologies, more comprehensive arrays are becoming available for researchers in biological fields. For example, exon arrays are designed to focus on exon-level analysis, but can also provide accurate assessments for gene-level expression analysis [7]. Thus, there exists a series of microarray datasets with similar biological samples but from significantly different array platforms. Obviously, there are scale and distribution differences among such heterogeneous datasets. To address this issue, gene-level normalization across datasets has typically been performed, but the details of this step have not been widely discussed in the combined analysis of microarray datasets. In this thesis, I describe a novel approach for combined microarray analysis based on gene-level unification followed by normalization and rank-aggregation.

Gene-level normalization is generally the preferred option for microarray analysis in a single study, and this has been revealed by an application of the M-A based loess normalization to a wholly defined control dataset from a “spike-in” experiment [8]. A previous study regarding the comparison of probe level normalization methods suggested that complete data methods including the M-A based loess normalization and the quantile normalization have better performance compared to other methods making use of a baseline array [9]. Therefore, we have adopted both M-A based loess normalization and quantile normalization, and then combined them with scale normalization for a gene-level analysis implementation.

Before normalization, gene-level unification into a virtual platform is demanded, as differences arise from various platform coverages. Therefore, only genes with common annotations either in gene symbol or mRNA accession number are considered to form a virtual platform. For the issue regarding multiple probe-sets matching a unique gene, we employ two strategies as the solutions, including selection based on most significant p-value from a statistical test or highest average of expression levels.

After gene-level normalization, a linear model is devised, which helps to identify lists of differentially expressed genes. Different normalization methods lead to (potentially different) lists of relevant genes, and a rank-aggregation approach is used to merge the power of different normalization methods.

To further complement the rank-aggregation approach, we have incorporated co-expression information to prioritize DEGs. The co-expression pattern of genes at the mRNA level can be recognized from large sets of microarray data. The rich body of data in GEO serves to provide this added dimension to our method. The basic idea is that genes with similar mRNA expression profiles are likely to be regulated via the same mechanism or share common functions [10]. This correlation information is useful for detecting or prioritizing genes with weak differential expression, since these genes are expected to co-express with other highly DEGs [11]. A statistical method of predicting genes with differential expressions based on co-expression patterns has already been proposed [11]. Moreover, rank-aggregation for similar items has been investigated as well [12]. Thus, in our work we have improved upon the rank-aggregation approach adding genome-wide co-expression information, which we term as *Co-expression-Rank-aggregation (Coex-Rank).*

The Coex-Rank approach for combined microarray analysis proceeds from normalization and linear modeling to rank-aggregation. Previous studies have demonstrated that methods using linear modeling have very similar behaviors to the t-based modeling approach [6]. The rank-aggregation method has also been shown to have similar performance to a rank-product approach [6]. Thus, from a theoretical view, our approach addresses the problem from both perspectives.

We applied this Coex-Rank featured approach of microarray combined analysis to our own Sppar (a mutant PPARγ) dataset [13], which illustrated an improvement in statistical power in identification of differentially expressed genes and a complementary advantage of the Coex-Rank approach from a biological perspective. A simulation study was also conducted to demonstrate that the strength of this method is not limited to our specific datasets.

The core novelty of our approach lies in the Coex-Rank process, which assists in the discovery of functional clusters of genes and biological pathways. Coex-Rank is not only flexible regarding different gene-level normalization methods in a combined analysis, but also useful in “merging” the power from different statistical methods within a single dataset analysis. This solution also provides an alternative to a seemingly arbitrary choice among many good computational methods.

We also applied this approach to other PPARγ-related microarray datasets, taking advantage of this merging power from different statistical methods. PPARγ is a transcription factor belonging to the nuclear receptor superfamily [14]. PPARγ heterodimerizes with the retinoid X receptor (RXR) and binds to specific response elements termed PPAR response elements (PPREs) in targeted gene promoters. The activation of target gene transcription depends on the binding of the ligand [14]. The endogenous ligand remains unclear although a number of fatty acids and eicosanoids have been proposed to be endogenous ligands. The synthetic antidiabetic thiazolidinediones (TZDs) represent a group of high affinity ligands for PPARγ [15]. The importance of PPARγ can be gleaned from patients with dominant negative mutations (P467L or V290M) in the ligand binding domain of PPARγ, as these have been reported to cause severe insulin resistance leading to full-blown type II diabetes mellitus and early onset hypertension [16].

From the perspective of gene sets, for example, we observed that up-regulated genes from mice treated with the PPARγ ligand rosiglitazone are significantly down-regulated in mice with a global knock-in dominant-negative mutation of PPARγ. We also integrated DEGs from microarray experiments with publicly available PPRE datasets from ChIP-chip [17] and ChIP-seq [18] and we found that genes which were most up-regulated by rosiglitazone treatment and which were also down-regulated by the global knock-in dominant-negative mutation of PPARγ are robustly enriched in PPREs near transcription start sites.

From this analysis we also identified several potential PPARγ targets in the aorta and mesenteric artery that now require further experimental validation. These genes are therefore candidates for explaining why mice carrying these mutations are more likely to develop hypertension and vascular dysfunction. For instance, Tnnc1, according to the microarray profile, is about 20-fold up-regulated in the aorta from Sppar mice (mice expressing PPARγ mutations in vascular smooth muscle cells). Tnnc1 is known to have mutations that affect the functional properties of Troponin C by increasing the Ca^{2+}-sensitivity of contraction [19]. Another candidate is Rhobtb1 which is part of the Rho Kinase pathway [20]. The Rho Kinase pathway is up-regulated in the aorta of Sppar [13]. We are also investigating Rgs5, a regulator of G protein signaling, which is known to interact with angiotensin AT1 receptors [21]. Rgs5 is down-regulated in the mesenteric artery of the Sppar mice and angiotensin II-mediated contraction is markedly enhanced.

CHAPTER 2: BACKGROUND

2.1 Introduction to PPARγ

2.1.1 All about PPARs

Peroxisome proliferator-activated receptors (PPARs) are a group of nuclear receptor proteins. They serve as transcription factors, modulating gene expression [14]. PPARs play essential roles in a variety of cellular processes, including major metabolic and inflammatory regulations [14].

As a subcellular organelle, the peroxisome plays a crucial role in cellular metabolism. In rodents, peroxisome proliferation can be induced by various chemical compounds [15]. A member of the steroid hormone receptor superfamily in mouse was found to be activated by peroxisome proliferators by Issemann et al and it was named PPAR [22]. There are three major types of PPARs: alpha, gamma, and delta/beta [15]. PPARα (alpha) is expressed in liver, heart, brown adipose tissue and kidney. PPARβ/δ (beta/delta) is expressed in many tissues, but markedly in brain, adipose tissue, skeletal muscle, gut and skin [14]. PPARγ (gamma) has two different forms through alternative promoter usage and differential splicing. PPARγ1 is expressed in virtually all tissues, including heart, skeletal muscle, colon, small and large intestines, kidney, pancreas, and spleen; PPARγ2 is expressed mainly in adipose tissue (30 amino acids longer) [23]. Mouse and human PPARγ are highly homologous with 95% sequence identity at the protein level [24].

All PPARs heterodimerize with the retinoid X receptor (RXR) and bind to response elements termed PPAR response elements (PPREs). The DNA consensus sequence is AGGTCAXAGGTCA, with X being a random nucleotide, known as the direct repeat 1 (DR1) pattern [14] [25]. In general, this sequence occurs in the promoter regions of target genes. With the presence of its ligand, a PPAR binds to PPRE and transcription of its downstream gene is typically activated or increased [14].

2.1.2 PPARγ Overview

2.1.2.1 Protein structure of PPARγ

As a nuclear receptor, PPARγ is modular in structure (See Figure 1) and has two important domains: the DNA-binding domain (DBD) and the ligand binding domain (LBD) [26]. Along with the DBD, the LBD contributes to the dimerization interface of the receptor and binding of coactivator or corepressor proteins [14]. The PPARγ2 protein contains an additional 30 amino acids at the N terminus compared to PPARγ1. The C-terminal region contains the major transcriptional activation domain, termed the activation function 2 (AF2) domain [26].

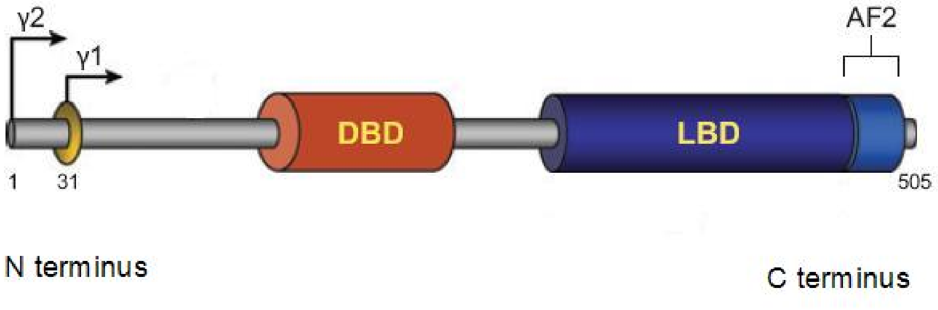


Figure 1. Domain structure of PPARγ. The PPARγ2 protein contains an additional 30 amino acids at the N terminus compared to PPARγ1. DNA binding domain (DBD), ligand-binding domain (LBD) and activation function 2 (AF2) are shown from left to right (N terminus to C terminus). This figure is modified from a review paper by Peter Tontonoz et al [26].

2.1.2.2 Classic working modes of PPARγ

There are two classic working modes of PPARγ: ligand-dependent activation and ligand-independent repression (see Figure 2). Functional significance of interaction between PPARγ and coactivators in transcriptional regulation has been implicated in recent investigation [15]. Steroid receptor coactivator-1 (SRC-1) and cAMP response element binding protein (CREB)-binding protein (CBP)/p300 are known coactivators of PPARγ [27, 28]. In fact, the interaction between PPARγ and CBP/p300 or SRC-1 is initiated by PPARγ coactivator-1 (PGC-1) with the presence of ligand [29]. PPARγ also recruits corepressor, such as the silencing mediator of retinoid and thyroid hormone receptors (SMRT) and the nuclear receptor corepressor (NCoR) [30]. They are capable of down-regulating PPARγ-mediated transcriptional activity in the absence of ligand.

However, by adding PPARγ ligand pioglitazone, these PPARγ-corepressor complexes were shown to be dissociated [30].

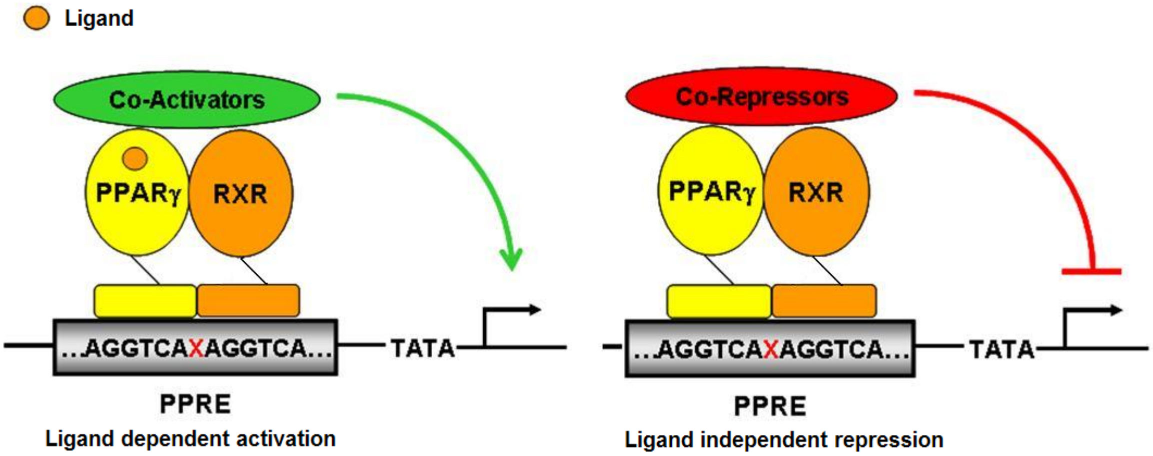


Figure 2. Two classic working modes of PPARγ. On the left panel, with the binding of ligand, PPARγ and RXR heterodimer recruits coactivators and then downstream genes are activated or up-regulated; on the right panel, in the absence of ligand, corepressors are recruited instead and results in the repression or down-regulation of target genes. The figure is modified from Carmen Halabi's [31] thesis.

The first endogenous ligand discovered for PPARγ is the PGJ2 metabolite 15-deoxy-delta 12,14-PGJ2 (PG is short for prostaglandin). It binds directly to PPARγ and promotes efficient adipogenesis [32]. PPARγ also has pharmacological ligands – the antidiabetic thiazolidinediones (TZDs), including troglitazone, pioglitazone, ciglitazone, and rosiglitazone. They bind PPARγ with various affinities, but are thought to be much greater than that of endogenous ligand [15] [33].

2.1.2.3 PPARγ in adipocyte differentiation and glucose homeostasis

PPARγ is a master regulator of adipocyte differentiation [26]. The direct binding of a TZD drug – rosiglitazone has been demonstrated. These drugs act as agonists for PPARγ and lead to the differentiation of adipose cells [33]. During the natural differentiation of pre-adipocytes into adipocytes, PPARγ is induced, which is highly expressed in both white and brown adipose tissues [34]. With the ectopic expression of PPARγ, expressions of adipose-specific genes were shown to be induced in fibroblasts and morphologic differentiation could be observed [35]. Later, a study reported that adipogenesis of cultured pre-adipocytes would be inhibited by PPARγ with dominant negative mutation [16]. Moreover, for the cells of a PPARγ-null model, formation of adipocytes was also abolished [36].

As mentioned above, PPARγ is a biological receptor for TZD drugs. The TZDs were found to have the capability of lowering glucose levels in rodents at first [26]. Later, in human, they were confirmed to have the function of improving insulin sensitivity [37]. TZDs exert their biological effects on insulin sensitivity through binding to PPARγ [26], with evidences shown as below: 1.) PPARγ with non-TZD agonists (rationally designed) led to improvement of insulin sensitivity [38]; 2.) Mutated PPARγ resulted in insulin resistance in both rodents and humans [16].

2.1.2.4 PPARγ in inflammation and atherosclerosis

While PPARγ plays an important role in the regulation of adipogenesis and glucose homeostasis, a great deal of evidence has emerged, supporting an essential role of PPARγ in inflammation and atherosclerosis [26].

PPARγ is also induced during monocyte differentiation into macrophages. It is revealed to have a high expression level in activated macrophages, such as the foam cells of atherosclerosis lesions [39]. Macrophages are able to detect and clear pathogens and release immune modulators at inflammatory sites. It has been shown that target genes of PPARγ with up-regulated pattern in macrophages are largely overlapping with those target genes arisen from adipose tissues, including Fabp4 and Cd36 [40]. Cd36 serves as a transporter of fatty acids. Consistent with this function, it has been demonstrated that lipid uptake has been to be promoted in PPARγ agonist treatment of macrophages [39]. Glass's group reported that macrophage inflammatory genes, such as TNFα and MMP-9, were inhibited by PPARγ ligands [41]. PPARγ was illustrated to cross-talk with NF-κB on the promoters of these inflammatory genes, which is responsible for the gene repression [42]. However, PPREs have not generally been found in the promoter regions of these repressed genes. Without direct binding to DNA sequences, the inhibitory ability of PPARγ is termed as transrepression [26].

For the development of atherosclerotic lesion, inflammatory signals in the vessel wall are recognized as a critical part [26]. Low density lipoprotein (LDL) is directly involved in atherosclerosis, because LDL-cholesterol accumulates in the blood. For male LDL-receptor deficient mice, administration of rosiglitazone could reduce the development of atherosclerosis [43]. The reduction in numbers and sizes of lesions was coupled with improvement of insulin sensitivity. At the same time, expressions of certain inflammatory marker genes were decreased as well [43].

2.1.2.5 PPARγ in endothelial and vascular smooth muscle cells

PPARγ appears to have effects in endothelial cells and vascular smooth muscle cells (VSMCs) as well, with impact to cardiovascular diseases [15]. It was reported that treatment of several cultured endothelial cells (both bovine and human) with troglitazone and pioglitazone enhanced the secretion of the vasodilator C-type natriuretic peptide (CNP) and decreased expression of endothelin-1 (ET-1), a potent vasoconstrictor [44]. Expression of vascular cell adhesion molecule-1 (VCAM-1) can be inhibited by PPARγ activators, such as ciglitazone and troglitazone, which exert beneficial effects in limiting chronic inflammation mediated by VCAM-1 [45]. Migration of vascular smooth muscle cells is a critical step in the formation of atherosclerosis and TZDs have been shown to function as inhibitors to VSMCs migration pathway [46].

2.1.2.6 PPARγ and hypertension

PPARγ activation with TZDs has been shown to attenuate hypertension in both animal models and human [47]. TZDs attenuated the development of hypertension in angiotensin II-infused rats [48]. Structural abnormalities and endothelial dysfunction of these rats were corrected as well [48]. Fullert et al designed a placebo-controlled double-blinded study to test the efficacy of TZDs. Patients with hypertension but not diabetes were treated with pioglitazone and 6-mm Hg greater reduction in diastolic blood pressure (DBP) over placebo was observed [49]. In another large clinical trial of 5238 patients with type 2 diabetes mellitus, systolic blood pressure was observed to be lowered by pioglitazone by 3 mm Hg [50]. Generally, the lowering blood pressure effect of TZDs is modest, but small decreases in arterial pressure are still beneficial [47][51, 52].

Interestingly, two separate dominant negative mutations of human PPARγ (V290M and P467L) have been described. Patients carrying one of these mutations have severe insulin resistance and develop full-blown type II diabetes mellitus and early onset hypertension later [16]. With mutations in the LBD, they have aberrantly high affinity for corepressor molecules instead of coactivators [16]. In addition, the P467L mutation is dominant negative, because its more reduced promoter turnover rate makes it out-compete the wild type (WT) receptor for promoter binding [53].

From the evidence above, we continue to hypothesize that PPARγ plays an important role in vascular function and hypertension. However, further studies are required to understand the role and mechanisms by which PPARγ exerts its effect in genome-wide transcriptional regulation of its target genes.

2.2 Co-expression indicates co-function

Genes sharing common biological functions have similar expression pattern. This feature is conserved from prokaryotic cells to eukaryotic ones, but with different regulatory mechanisms [10]. From the aspect of energy allocation, similar regulation of genes involved in the same biological process saves unnecessary energy usage for a cell, which is critical to cell’s survival [54].

2.2.1 Example from prokaryotes

In bacteria, operon refers to a cluster of genes with common function under the same control of transcription. Genes are physically located next to each other, sharing a single regulatory signal or promoter. As the result, they are transcribed together and demonstrate the pattern of co-expression [54].

One of the classic examples is the *lac* operon of the bacteria *Escherichia coli* (*E. coli*) [54]. There are three genes encoded by the lac operon: lacZ, lacY, and lacA. which are transcribed in the presence of lactose. lacZ makes the β-galactosidase, an intracellular enzyme that cleaves the lactose. Plus, lacY produces the lactose permease enzyme and it is responsible for transporting lactose into the cell [54]. Therefore, proteins encoded by *lac* operon genes are involved in the processing of lactose, sharing common functions.

2.2.2 An instance in Eukaryotes

For eukaryotes, the co-regulation of gene expressions is generally achieved through common regulatory elements [54]. Take three genes GAL1, GAL7 and GAL10 of yeast for illustration, they share upstream activator sequence for GAL (UAS_G). Their transcription can be initiated in the presence of galactose. They produce proteins involved in galactose processing, respectively galactose transferase, galactose epimerase, and galactokinase [54]. Thus, co-expression patterns of genes indicate co-regulatory mechanisms and shared functions.

2.3 Introduction to microarrays

2.3.1 General information of microarray experiments

High-throughput microarray technologies have become popular for genome-wide investigation of gene expression profiles. Careful experimental design followed by a variety of proper computational analyses can reveal interaction of genes and related biological pathways [1]. For the experimental procedure, the needed raw materials are different biological RNA samples. These isolated RNA samples are labeled and then hybridized to arrays with tens of thousands of probes [1]. Next, the arrays are scanned to generate images which provide information of relative fluorescence intensities for each element (refers to probe). Using software package for image quantitation, the relative amount of original input RNA (indication of gene expression levels) can be robustly measured [1]. Generally, there are two strategies regarding the hybridization process. One-channel microarray is designed to use a single label and independent array for each sample, while two-channel microarrays can afford two samples on one array at the same time, using distinguishable fluorescent dye labels [1].

For the data analysis, a common goal is to detect differentially expressed genes (DEGs) [1] between/among different groups of samples, such as control v.s. cases and distinctive responses to a specific factor. Before the appropriate comparison can be carried out, quantitative transformation of raw data should be performed to balance intensities from different arrays, which is termed normalization [1]. Why data must be normalized? Because differences might arise from unequal starting amount of RNA, varied efficiencies of fluorescent labeling, image detection and so on [1]. Normalization is based on the assumption that the total intensities summed over all elements on an array should be the same for every biological sample in a microarray experiment [1].

2.3.2 Combined analysis of multiple sets of microarray data

Different laboratories may carry out microarray experiments with related biological experimental design, but using different types of platforms. Due to the high cost of microarrays, many studies suffer from the problem of small sample size, which may lead to a high false discovery rate (FDR) in determination of DEGs [2]. Combining related but independent microarray datasets increases sample size and may result in higher reliability of novel gene candidate discovery from a statistical view [3]. For example, a combined approach may be able to detect small but consistent changes. In fact, this is one of the motivating factors for the construction of public microarray databases, such as GEO (Gene Expression Omnibus) [4]. In another way, successful combined analysis demonstrates the reproducibility of these studies [5], which is a fundamental issue in validation of biological experiments.

However, rarely is a direct combined analysis suitable for microarray studies, as complications arise from biological variations and technical differences [6]. Meta-analysis, which has been well-studied in statistics, is a practical way to solve this problem [6]. The application of meta-analysis to microarray data has been demonstrated by different groups, yet no consensus has been reached as to the best method. Hong,F. et al. evaluated the performance of different microarray meta-analysis methods and recommended approaches derived from two different philosophies. One is the t-based modeling approach, which offers a comparison of the magnitudes of a treatment on different genes [6]. Methods using linear models or from a Bayesian perspective have very similar behaviors to the t-based modeling approach [6]. The other is a rank-product approach, which shows indistinguishable performance with the rank-aggregation method [6]. This approach has the advantage of robustness in ranking genes over the t-based method, but only provides relative prioritization of genes [6].

2.3.3 Co-expression information in microarray analysis

Microarray data not only provide us with gene expression profiles but also co-expression information of genes investigated by certain platforms. Especially, for a large set of microarrays, which is available from the rich body data in GEO [4], robust co-expression information of genes can be derived.

As we have introduced in Section 2.2, genes with similar mRNA expression profiles across different tissues are likely to be regulated via the same mechanism or share common functions [10]. This correlated information is useful for detecting or prioritizing genes with relatively weaker differential expression levels, since these genes are expected to co-express with other highly DEGs [11]. A statistical method of predicting genes with differential expressions based on co-expression patterns in a single dataset analysis has already been proposed [11]. A microarray data set collected for a prostate cancer study was demonstrated as an application, in which the proposed method identified many genes with weak differential expressions and several of these genes were known in literature to be associated with the disease [11].

2.4 Genome-wide profiling of DNA-binding proteins

To fully understand the mechanism of transcriptional regulation, it is essential to obtain the genome-wide mapping of protein-DNA interactions [55]. The genome-wide regulatory network of genes can be revealed by a precise map of binding sites for transcription factors, core transcriptional machinery and other DNA-binding proteins [55].

2.4.1 ChIP-chip

Chromatin immunoprecipitation (ChIP) is the major tool to investigate these mechanisms [55]. It is a technique for assaying protein-DNA binding *in vivo*, which gives the interactive picture occurring inside the nucleus of living cells [56]. In a ChIP experiment, DNA-binding proteins in living cells are cross-linked to the DNA with the treatment of formaldehyde at the beginning. Next, the cells are lysed and the DNA is sonicated into small pieces (about 0.2–1 kb long). By using an antibody specific to a putative protein, the target protein-DNA complexes can be pulled out, which are later separated through a process of heat driven cross-link reversal [56].

To identify these DNA fragments isolated from ChIP assay on a genome-wide scale, high-density DNA tiling arrays can be used [57]. For the tiling arrays, oligonucleotide probes are placed across an entire genome or selected chromosomal regions, such as promoter regions [55]. DNA fragments are hybridized to the tiling arrays. Followed by computational analysis, enriched genomic regions can be determined [55].

2.4.2 ChIP-seq

With the rapid development of next-generation sequencing technology, the dream of sequencing tens or hundreds of millions of short DNA fragments in a single run becomes true [55]. This technique has been widely applied to whole genome sequencing, mRNA sequencing for gene expression profiling (RNA-seq) and DNA sequencing from ChIP assay (ChIP-seq) [55].

For ChIP-seq, sequencing of DNA fragments of interests is launched after the ChIP experiment instead hybridization to tiling arrays [58]. Computational analysis to identify the DNA-binding sites is also a critical part of this method. This method has been shown to have higher resolution, fewer artifacts and greater coverage compared to ChIP-chip approach [55].

2.5 Rank aggregation

Combining several ordered lists in a proper and efficient manner is a real challenge in the field of bioinformatics. Rank aggregation provides a solution with a general framework and flexibility [59]. The goal of rank aggregation is to find a "robust" list, which is as close as possible to all individual input lists simultaneously [60]. To cast the problem into optimization area, the objective function is defined as the following formula:

$$\delta = \arg\min \sum_{i=1}^{m} w_i d(\delta, L_i) \ .$$

In this formula, δ is the robust list to be determined. The number of input lists is m and L_i stands for the i_{th} input list. The function *d(,)* calculates the distance between the robust list and one of the input lists and w_i is the weight associated with the corresponding list [61].

In the literature, there are many choices for a distance function. One of the most popular methods is Spearman footrule distance [60], which is defined as below:

$$d(\delta, L_i) = \sum_{t \in L_i \mathrm{U} \delta} \left| \delta\text{-}rank(t) - L_i\text{-}rank(t) \right| .$$

In this formula, *t* stands for a gene in the union of both L_i and δ lists and L_i*-rank(t)* gives the rank of gene *t* on list L_i. If gene *t* is absent on list L_i, its rank will be assigned as the length of L_i plus 1. As an intuitive metric for comparing two ordered lists, Spearman footrule distance sums up the absolute differences between the ranks of all unique elements from both input lists [60].

To determine the robust list, a framework should be constructed to search the solution space. The Cross-Entropy Monte Carlo algorithm has been demonstrated to be effective in discovering the optimal ordering of elements in the output list [60]. The main steps of this algorithm are described with a simple example as follows:

(1) Initialization: At each rank position, every element has the same probability to be selected [60]. For example, consider input lists each with three genes, and there are five genes in the union of all the lists. Therefore, at each rank position, the probability is 0.20 (see Table 1).

Table 1. Probability Matrix at the initialization stage.

	Rank 1	Rank 2	Rank 3
Gene_a	0.20	0.20	0.20
Gene_b	0.20	0.20	0.20
Gene_c	0.20	0.20	0.20
Gene_d	0.20	0.20	0.20
Gene_e	0.20	0.20	0.20

Note: A simple example with each list having 3 genes and 5 genes in total. At each rank, every gene has the same probability to be selected.

(2) Sampling: Based on the current probability matrix, candidate lists are generated for evaluation by the objective function [60]. To select elements for a candidate list, the procedure starts from the first rank to the last one. In the following example, the goal is to choose a gene at “Rank 1” position. Table 2 contains partial information – only “Rank 1” from the probability matrix and then the cumulative probability is calculated for each gene from the top row to the bottom row. The gene for “Rank 1” is selected via conditional random generation. A number is drawn randomly from a uniform distribution between 0 and 1. If the variate is x, we choose the gene with the smallest cumulative probability that is larger than or equal to x and in this case (i.e., x=0.25), the result is Gene_b with cumulative probability 0.40 (see Table 2). After fulfilling the position of “Rank 1”, the probability matrix should be re-scaled using the remaining genes to continue the sampling process.

Table 2. Probability Matrix for the sampling stage.

	Rank 1	Cumulative Prob.
Gene_a	0.20	0.20
Gene_b	0.20	0.40
Gene_c	0.20	0.60
Gene_d	0.20	0.80
Gene_e	0.20	1.00

Note: A simple example with each list having 3 genes and 5 genes in total. Cumulative probabilities at “Rank 1” are shown.

(3) Updating: In practice, tens of thousands of candidate lists (default: n=5×k×u, where k is the size of each input list and u is the number of unique items in the union of all the input lists) are generated by the above sampling method. These lists are evaluated by the objective function and the probability matrix is updated based on lists with the highest scores (default: n×1%) [60]. At the same time, the list with the highest score (the top one) is identified as the optimal list. The following simple example is shown with each list having 3 genes and 5 unique genes in total. We focus on the 6 candidate lists with highest scores generated by the sampling process. The numbers of each gene at each rank are summed up, shown in Table 3, on the left panel. The corresponding probabilities are calculated, shown in Table 3, on the right panel.

(4) Convergence: The searching procedure will be terminated when the optimal list does not change in a fixed number of iterations [60].

Table 3. Probability Matrix at the updating stage.

Top 6 candidate lists					Prob. Matrix			
	Rank 1	Rank 2	Rank 3			Rank 1	Rank 2	Rank 3
Gene_a	5				Gene_a	0.833		
Gene_b	1	1	1		Gene_b	0.167	0.167	0.167
Gene_c		1	1		Gene_c		0.167	0.167
Gene_d		1	2		Gene_d		0.167	0.333
Gene_e		3	2		Gene_e		0.500	0.333

Note: A simple example with each list having 3 genes and 5 genes in total. The numbers of each gene at each rank are shown on the left panel and the corresponding probabilities are shown on the right panel.

CHAPTER 3: COEX-RANK, AN APPROACH FOR MICROARRAY COMBINED ANALYSIS

3.1 Motivating datasets

As mentioned in Chapter 1, microarrays have been widely used to study differential gene expression at the genomic level and biologically related datasets from independent studies are publicly available. This situation requires robust combined approaches for integration and validation. Previously, meta-analysis has been adopted to solve this problem [6]. As an alternative, for microarray data with high similarity in biological experimental design, a more direct combined method based on gene-level unification followed by normalization is feasible.

Before introducing the pipeline of methods, I will first provide two motivating datasets, with similar, but non-identical experimental designs and/or platforms. The Sigmund laboratory has generated transgenic mice with dominant negative PPARγ (P467L) targeted to vascular smooth muscle cells (VSMCs) and these mice (called Sppar mice) have been shown to exhibit severe aortic dysfunction [13]. Two independent microarray experiments were carried out using RNA from the thoracic aortas of these mice, compared to wild-type littermate controls (denoted as Sppar datasets). The first experiment was performed using the Affymetrix mouse genome 430 2.0 array (referred to as expression array), with only 2 control and 3 transgenic samples. The second set of samples from the same mice took advantage of the more recently available Affymetrix mouse exon 1.0 ST array (referred to as exon array); this time with 5 control and 7 transgenic samples. In fact, this experiment was originally designed as 6 control and 6 transgenic. Upon validation of the genotypes from tails of mice collected at sacrifice, 1 control was found to be transgenic and thus it was transferred to the transgenic group. In the expression array experiment, each hybridization contained RNA from 8 pooled aortas, while in the exon array experiment, the RNA was extracted from a single aorta followed by amplification. Clearly, there are multiple challenges presented by this heterogeneity of experimental designs.

3.2 Methods

3.2.1 Gene-level unification into a virtual platform

To generate gene-level expression values, we used the RMA (Robust Multi-chip Average) algorithm [62, 63]. For expression array data, the implementation was carried out in R using the *affy* package and resulted in 45,101 probe-sets. The Affymetrix Expression Console software (http://www.affymetrix.com/products_services/software/Specific/expression_console_software.affx) was applied to data from the exon arrays and 101,176 gene-level probe-set records were generated. Demanded by comparison across different platforms, we considered genes with common annotions to form a virtual platform.

We attempted to remove redundant and ambiguous probe-sets (See Figure 3). First of all, probe-sets without annotations such as gene symbols or mRNA accession information (according to Affymetrix annotation, release 30) were removed. As the annotation information is not always perfect, there might be duplicated items regarding the same record (see Figure 3), which should be cleared before further processing.

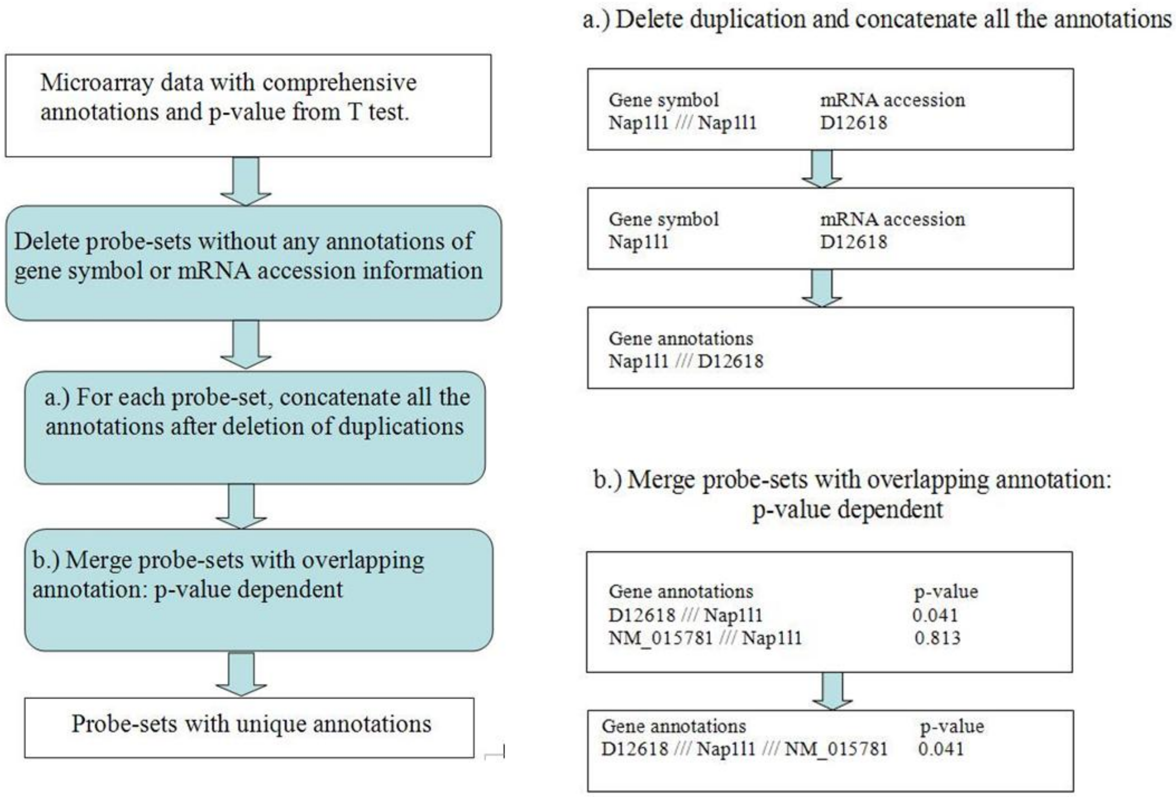

Microarray data with comprehensive annotations and p-value from T test.
Delete probe-sets without any annotations of gene symbol or mRNA accession information
a.) For each probe-set, concatenate all the annotations after deletion of duplications
b.) Merge probe-sets with overlapping annotation: p-value dependent
Probe-sets with unique annotations
a.) Delete duplication and concatenate all the annotations
Gene symbol
Nap1l1 /// Nap1l1
mRNA accession
D12618
Gene symbol
Nap1l1
mRNA accession
D12618
Gene annotations
Nap1l1 /// D12618
b.) Merge probe-sets with overlapping annotation: p-value dependent
Gene annotations
D12618 /// Nap1l1
NM_015781 /// Nap1l1
p-value
0.041
0.813
Gene annotations
D12618 /// Nap1l1 /// NM_015781
p-value
0.041

Figure 3. Pre-processing of probe-sets on a microarray platform. The general steps are listed on the left panel, while a simple example is shown on the right panel using best p-value selection strategy.

In the case of multiple probe-sets matching the same gene, a method was needed to remove this redundancy. We explored using both the most significant p-value and the highest average expression value as determining factors. Student’s T test with equal variance was used to calculate the p-value, comparing control vs. transgenic samples. For example (see Figure 3), there is one record with annotations “NM_015781 /// Nap1l1” and another record annotated as “D12618 /// Nap1l1”. Therefore, they are merged into a new record as they share the same gene symbol “Nap1l1”. In this example, we demonstrate with the p-value dependent approach. As the record with “D12618 /// Nap1l1” has the more significant p-value, the raw data for this record are selected. Moreover, its annotation is extended to incorporate “NM_015781” to provide more comprehensive information. Through the above steps, 26,766 probe-sets on the expression array and 33,312 probe-sets on exon array were retained. Then, we combined probe-sets from two datasets if they had overlapping annotations. Following this rule, we finally generated two combined datasets, each with 18,204 records. One is selected based on the best p-value and the other is chosen according to the highest average expression level.

3.2.2 Normalization

Normalization is naturally driven by the relative scale or differences in the distribution of expression levels among arrays from multiple studies. In the case of Sppar data for example (see Figure 4), the distributions of gene expression intensities are markedly dissimilar between the two platforms, and if analysis proceeded without normalization, the results would be invalid and very misleading. In our implementation, we applied scale normalization first, which is capable of correcting linear variations, followed by either quantile or M-A based loess normalization [9].

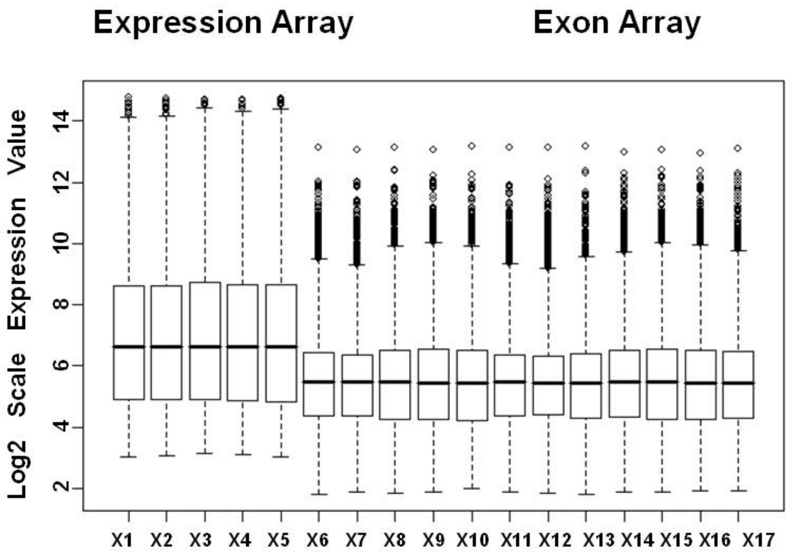


Figure 4. The boxplots of all 17 arrays from Sppar datasets. X1-X5 refer to data from the expression arrays and they show different distributions from X6-X17 plots of the exon arrays

Scale normalization is sometimes referred as global normalization, which enforces an equal median or mean intensity criteria for all arrays [9]. In our implementation, we selected a method based on median, which is less sensitive to extreme data points. The details are explained as follows:

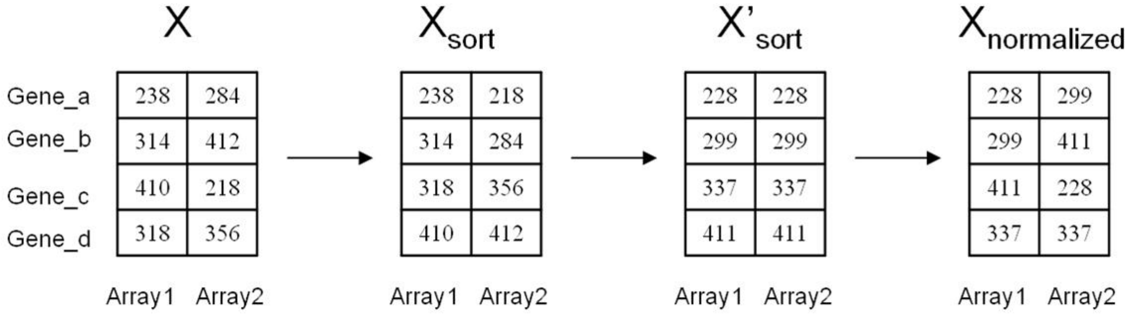
1.) calculate the median intensity of each array;

2.) select the array having the median of the median intensities as the baseline array;

3.) normalize each of the remaining arrays to the baseline array by multiplying by a coefficient β_i :

$$\beta_{\mathrm{i}} = \frac{\text{median(array\_baseline)}}{\text{median}\left(\text{array\_i}\right)}$$, i = 1,2…n (n is the number of all arrays).

Quantile normalization enforces an equal distribution of intensity values across all the arrays [9]. Here is one example of its use (see Figure 5). This is a simple case with only 4 genes and 2 arrays. Table X is the original data. First we sort the intensities from low to high on each array and to get Table Xsort. Next, we calculate the average for each gene and use the mean values instead of the previous intensities to generate Table X'sort. Finally, we move each gene to its original rank position to achieve Table Xnormalized.



X

	Array1	Array2
Gene_a	238	284
Gene_b	314	412
Gene_c	410	218
Gene_d	318	356

X_{sort}

Array1	Array2
238	218
314	284
318	356
410	412

X'_{sort}

Array1	Array2
228	228
299	299
337	337
411	411

$X_{normalized}$

Array1	Array2
228	299
299	411
411	228
337	337

Figure 5. An example of quantile normalization illustrated by 4 genes from 2 arrays.

M-A based Loess normalization is a classical method for cDNA array normalization and can also be applied to two one-channel arrays [64]. First, Y and X denote the $\log_2$-scaled expression values from two arrays. M denotes the difference between Y and X, while A represents the average of Y and X. That is, M=Y–X and A=(Y+X)/2. The M'-A' plot after loess regression should show a cloud of points scattered about the M'=0 axis. Y', X' are then generated [64].

Loess normalization can be realized via two different approaches – either a median-base method or a trim-mean method. For the median-base method, consider the Sppar combined data mentioned above. In each iteration, Y proceeds from array X_1 to array X_{17}, while X is the array storing the median of the median intensities of all arrays (termed as X_{base}), therefore there are 17 rounds of loess regressions. For each loess regression, X is selected dynamically based on the current expression values of all arrays, and both Y' and X' are used to update Y and X. The pseudo code of this algorithm is as follows:

```
for ( i in 1:#interation ) /*the number of iterations*/
{
    for (j in 1:#sample)  /* sample size=17 in our Sppar dataset*/
    {
        $Y=X_j$ ; $X=X_{base}$ ;
        Loess normalization using Y and X ;
    }
}
```

For the trim-mean method, in each iteration, Y proceeds from array X_1 to array X_{17}, while X is the reference array, dynamically generated consisting of the 0.05 trim mean of all 17 arrays. As X is only a series of reference arrays, only Y is updated using Y'.

For loess normalization, the regression can also be performed using only rank-invariant genes. The size of the rank-invariant gene set is data dependent. Genes are defined as rank-invariant as described in a previous study [65].

3.2.3 Linear model

After normalizing using different methods, we generated lists of significantly changed genes for further comparison or validation by a simple linear model. A variety of complex methods have been proposed, but they do not necessarily perform better than a simple one. Further, complex methods may add background noise and even induce bias if all assumptions are not satisfied [66]. For example, consider our Sppar data, a linear model can be constructed for each gene by the following formula:

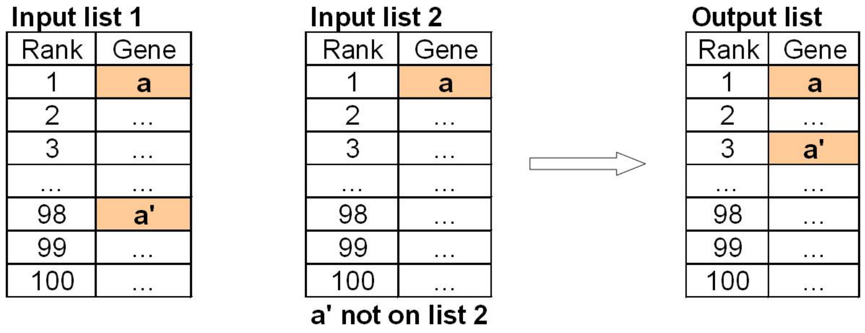
$$Y= b + a_1 \times X_1 + a_2 \times X_2 ,$$

where Y is the observed value of gene expression and b is the baseline level of gene expression. Data from expression array and wild type are considered as the baselines. The exon array effect is indicated by a_1 and X_1 = (0 or 1). The Sppar mutant effect is measured by a_2 and X_2 = (0 or 1) as well. The regression is carried out using R function *lm( )* and then ANOVA is used to test the statistical significance of a_2. The +/- sign of a_2 indicates up or down regulation and the absolute value of a_2 indicates fold-change, which is different from the original scale but can still be used to rank genes or indicate relative changes.

3.2.4 Co-expression-Rank-aggregation (Coex-Rank)

Multiple lists of up/down-regulated genes can be generated from different normalization methods. To take advantage of the power from merging all these lists, we investigated the rank-aggregation method [60], which focuses on finding a robust list with minimum distance among all available ordered lists of genes. We chose to use R *RankAggreg* package from CRAN which is publicly available [60]. One of the most popular distance functions – Spearman footrule distance has been implemented in this package. The realization of rank-aggregation is provided with two different algorithms and the Cross-entropy Monte Carlo algorithm (CE) is better recommended [60].

For the Coex-Rank approach, we modified the R implementation of rank-aggregation by incorporating co-expression information. The goal of Coex-Rank is to prioritize genes highly correlated with genes which are already highly ranked. For instance (see Figure 6), Gene_a and Gene_a' are highly correlated in expression across different tissues of a specific species. Gene_a is a highly-ranked gene on all input lists for Coex-Rank, but Gene_a' is present at the bottom of some of the input lists. Through our Coex-Rank process, Gene_a' will be pulled up near the top of the output list.



Input list 1

Rank	Gene
1	**a**
2	…
3	…
…	…
98	**a'**
99	…
100	…

Input list 2

Rank	Gene
1	**a**
2	…
3	…
…	…
98	…
99	…
100	…

a' not on list 2

Output list

Rank	Gene
1	**a**
2	…
3	**a'**
…	…
98	…
99	…
100	…

Figure 6. Demonstration of the Coex-Rank approach. Gene_a is a highly-ranked gene on all input lists for Coex-Rank, but Gene_a' is present near the bottom of some of the input lists. The Coex-Rank approach enhances the priority of Gene_a' that is highly correlated with an already-highly-ranked Gene_a.

For our implementation, the co-expression information is included in the distance calculation step. This information is obtained from a combination of microarray datasets with samples from different tissues of the same species to avoid bias. To be consistent with our case study, mouse Sppar data, we used two datasets available from GEO: GSE10246 with 182 samples on the Affymetrix mouse genome 430 2.0 array and GSE15998 with 106 samples on the Affymetrix mouse exon 1.0 ST array [67]. The co-expression coefficients calculation was based on the probe-sets matching with the final combined Sppar dataset as described in Section 2.1.2. Then, for any two genes, the Pearson's correlation coefficient was calculated from 288 pairs of records.

Distance calculation with co-expression information is the heart of the Coex-Rank algorithm. The distance D( ) between two ranked gene lists L_1 and L_2, given the co-expression coefficients, is defined as follows:

$$D(L_1,L_2)= \frac{1}{2} \times (F(L_1,L_1\text{-co}) + F(L_2,L_2\text{-co})),$$

where F( ) is the Spearman footrule distance of two lists [12]. List L_1-co contains all the genes from list L_1 but the rank information is obtained from list L_2. For genes also present on list L_2, their ranks remain the same, while for genes only present on list L_1 but not on list L_2, the ranks of their highly correlated genes from list L_2 are used instead. In practice, we need to determine a cut-off value for co-expression coefficients for consideration. For example, Gene_a is only present in list L_1, and it has n highly correlated genes on list L_2. The rank of Gene_a on list L_1-co is defined as follows:

$$L_1\text{-co-rank (Gene\_a)} = \frac{1}{n} \times \sum_{i=1}^{n} [\frac{L_2\text{-Rank }(\text{Gene\_i})}{\text{Co}(\text{Gene\_a},\text{Gene\_i})}],\ i = 1,2, \ldots n.$$

Co(Gene_a,Gene_i) denotes the co-expression coefficient between Gene_a and Gene_i (we used the Pearson correlation coefficient in our implementation) and L_2-Rank (Gene_i) is the rank of Gene_i on list L_2. For genes only present on list L_1 but not on list L_2, if they do not have any highly correlated genes from list L_2, their ranks are assigned as Length(L_2)+1, where Length( ) is the length of the gene list.

The R program is freely available for download from http://genome.uiowa.edu/Coex-Rank with simple data as an example.

3.3 Results

All the results below are based on the combined dataset formed by removing redundant probes based on best p-value as described previously.

3.3.1 Similar effect of different normalization methods

For our mouse Sppar data, 10 different normalization methods were implemented as follows: 1.) quantile normalization; 2.) loess-median-base normalization; 3.) loess-median-base-invariant normalization; 4.) loess-trim-mean normalization; 5.) loess-trim-mean-invariant normalization. In addition, the same 5 methods with first round scale normalization were also used.

For the loess-median-base approach, with or without an invariant gene set, 10 iterations were chosen for normalization. More iterations should result in more similar distributions of intensities from different arrays. As many as 50 iterations were calculated, but no significant improvement in the results was observed, shown by an example using scale-loess-median-base-invariant normalization method (see Figures 7-9).

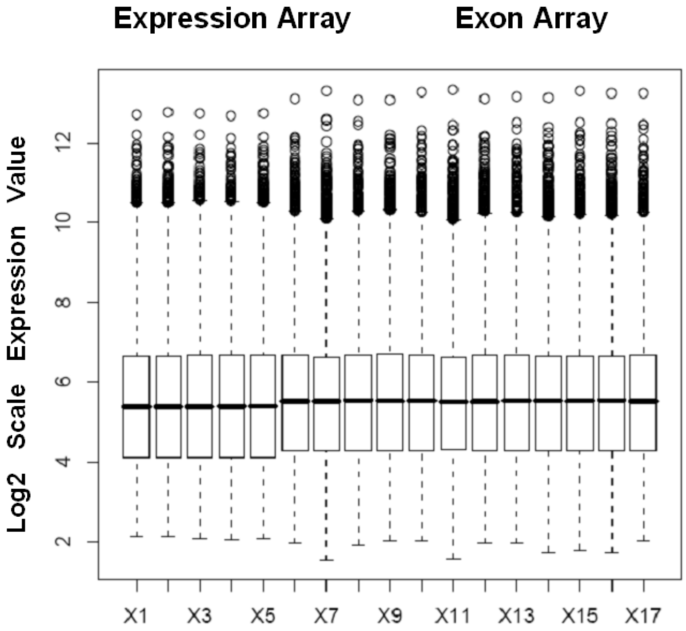


Figure 7. The boxplots of all 17 arrays from Sppar datasets after scale-loess-median-base-invariant normalization with 10 iterations. X1-X5 refer to data from the 5 expression arrays and X6-X17 plots are according to data from the 12 exon arrays.

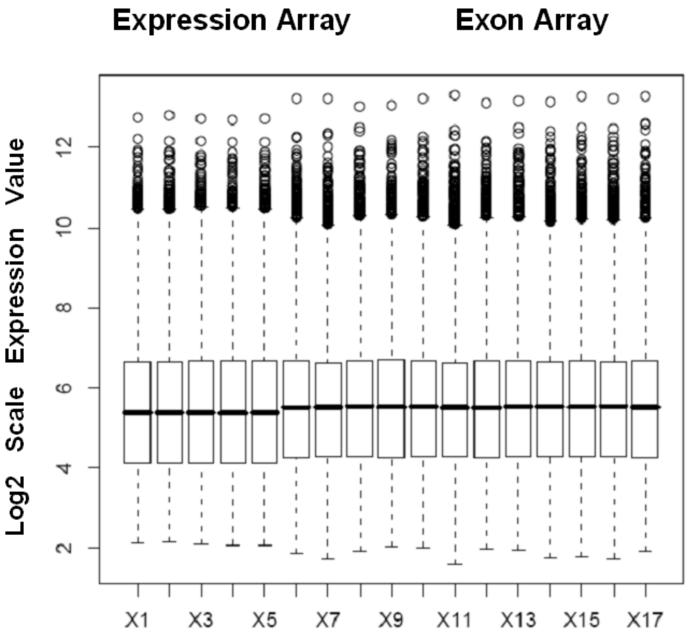


Figure 8. The boxplots of all 17 arrays from Sppar datasets after scale-loess-median-base-invariant normalization with 20 iterations. X1-X5 refer to data from the 5 expression arrays and X6-X17 plots are according to data from the 12 exon arrays.

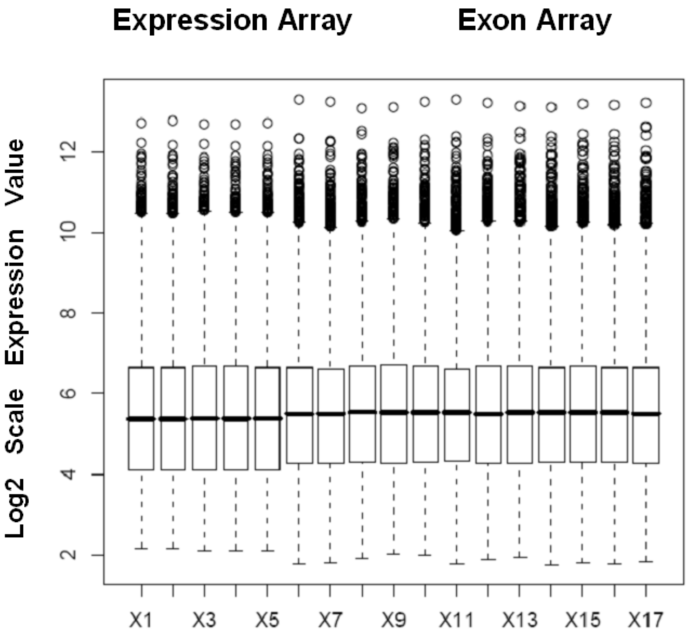


Figure 9. The boxplots of all 17 arrays from Sppar datasets after scale-loess-median-base-invariant normalization with 50 iterations. X1-X5 refer to data from the 5 expression arrays and X6-X17 plots are according to data from the 12 exon arrays.

After applying the scale-loess-median-base-invariant method for normalization, with 10, 20 or 50 iterations, up/down-regulated genes lists were generated from these normalizations, each list with the 100 most highly-ranked genes. Comparisons were carried out separately for up-regulated or down-regulated gene lists (see Table 4). The mimimum size of overlapping geneset was 91 (intersection of genes from 10 iterations and 50 iterations), indicating that increasing the number of iterations does not dramatically improve the normalization. Note: these iterations are computationally demanding.

Table 4. Size of intersection of any two lists from different iterations in normalization.

	Up-regulated genes		Down-regulated genes	
Iterations	20	50	20	50
10	97	95	95	91

The loess-trim-mean approach narrows the distribution of intensities after a sufficiently large number of iterations. In an extreme example of 50 iterations, the boxplots of intensities degenerate into many repeated data points. Thus, we selected 5 iterations for the Sppar data, which produced similar distributions of intensities as other approaches.

For the loess regression based on rank-invariant genes, a separate analysis (data shown in Section 3.3.2) showed that no more than 1,000 genes significantly changed between control and transgenic groups. Thus, we used 17,000 as the size of our rank-invariant gene-set.

After normalization using each method, linear models were created for each gene. An ANOVA test was applied to generate lists of up/down-regulated genes due to the Sppar effect. Next, a comparison of 10 up-regulated lists was performed, each with the 100 most highly-ranked genes based on p-value (see Table 5). In the table showing the size of the union of any two gene lists, the largest set contains 129 genes, which indicates that lists from any two normalization methods have about 70% overlapping genes at least.

Table 5. Size of union of any two lists from different normalizations for up genes.

M	1	2	3	4	5	6	7	8	9	10
1	quantile	100	113	112	110	108	129	123	127	122
2	scale-quantile		113	112	110	108	129	123	127	122
3	loess-trim-mean			102	111	109	128	129	125	126
4	scale-loess-trim-mean				110	108	127	127	125	124
5	loess-trim-mean-invariant					103	127	123	126	119
6	scale-loess-trim-mean-invariant						126	123	126	120
7	loess-median-base							128	115	127
8	scale-loess-median-base								127	122
9	loess-median-base-invariant									127
10	scale-loess-median-base-invariant									

Note: The first row and the first column show the index of normalization Methods (M).The maximum union size is 129, highlighted in yellow.

For down-regulated genes, the results are similar (see Table 6). Though different normalization methods were applied, similar gene lists were generated, which motivated our selection of the rank-aggregation approach to incorporate information from all the normalization methods.

Table 6. Size of union of any two lists from different normalizations for down genes.

M	1	2	3	4	5	6	7	8	9	10
1	quantile	101	116	114	112	111	127	122	123	117
2	scale-quantile		116	114	112	111	127	123	123	118
3	loess-trim-mean			102	105	106	130	127	133	123
4	scale-loess-trim-mean				104	104	128	125	131	121
5	loess-trim-mean-invariant					102	128	125	129	121
6	scale-loess-trim-mean-invariant						127	124	128	120
7	loess-median-base							130	124	124
8	scale-loess-median-base								128	114
9	loess-median-base-invariant									122
10	scale-loess-median-base-invariant									

Note: The first row and the first column show the index of normalization Methods (M). The maximum union size is 133, highlighted in yellow.

3.3.2 Combined analysis increases statistical power

In general, increasing the sample size will result in an increase in statistical power of an analysis. For our Sppar data, the combined analysis has a sample size of 17, while the separate datasets have sample sizes of 5 and 12 respectively. Comparison of the two different analyses demonstrates (in this case) the benefit of the larger sample size.

For the case of separate analyses, student's T test with equal variance was used to compare control vs. transgenic samples. This statistical test is mathematically equivalent to a one-way ANOVA test. When we selected a p-value<0.005 as a cut-off value, we could achieve roughly twice the number of genes via combined analysis, compared to the separate approach (see Table 7). The statistics of the combined analysis were based on the scale-loess-trim-mean-invariant normalization method; other normalization methods resulted in similar numbers.

Table 7. Comparison of combined and separate analyses of Sppar data.

	Expression Array		Exon Array		Combined Analysis	
	p-value	FDR	p-value	FDR	p-value	FDR
#Total	288	5	218	23	583	286
#Up	200	2	115	9	283	140
#Down	88	3	103	14	300	146

Note: Numbers of total DEGs and up/down-regulated genes are shown separately for expression/exon array data and combined analysis. The cut-off value: p-value < 0.005 and FDR < 0.05 are used.

We also corrected for multiple comparisons using the R *qvalue* package [68]. When we set FDR (False Discovery Rate, also called q-value) < 0.05, we could see a dramatic improvement with the combined analysis; from 5 genes from the expression arrays, 23 genes from the exon arrays to 286 genes from the combined analysis (see Table 7). Two different cut-off values were set and more genes were selected as DEGs in the combined analysis, indicating the increased statistical power of this approach.

3.3.3 Complementary advantage of Coex-Rank

Gene lists from 5 normalization methods starting with scale normalization were used as the input for both the rank-aggregation and the Coex-Rank approaches. For example, for the up-regulated genes, considering p-value < 0.005 as the cut-off, 5 gene lists were generated and then the genes were ranked either by p-value or fold change, which resulted in 10 different lists. The 100 most highly-ranked genes were selected from each list and then served as the input for both the rank-aggregation and the Coex-Rank approaches. Although the choice of 100 as the number of genes to consider was somewhat arbitrary, it is noteworthy that these genes were all significantly up-regulated according to the FDR < 0.05 cut-off value. Thus, we will continue to use this convenient list size for the remainder of this example presentation.

The parameter settings for the rank-aggregation step were the default values (Spearman footrule distance and cross-entropy algorithm), except that the maximum-iteration was increased from 1000 to 1500 for our Sppar data. For the Coex-Rank approach, one more parameter for the cut-off value of co-expression coefficients was set to 0.7 for our Sppar data. As there were 198 unique up-regulated genes from 10 different lists (each with 100 genes), there were 197×198/2=19,503 pairs of genes for correlation calculation and 572 pairs resulted in coefficients greater than 0.7. Similarly, for down-regulated genes, there were 199 unique genes, which led to 19,701 co-expression coefficients; 954 of them were greater than 0.7. Therefore, the 0.7 cut-off value roughly selected the highest 3-5% genes based on the values of the correlation coefficients. The output of both the rank-aggregation and the Coex-Rank approaches were lists, each with 100 genes.

We note that the rank-aggregation and the Coex-Rank methods, both generated different lists of genes, but that they shared about 70% genes in common (73 for up-regulated genes and 71 for down-regulated genes). To investigate the biological significance of these genes, we focused on the enrichment of their public annotations. We compared the gene lists from two approaches by clusters generated by DAVID [69] (the default low classification stringency was used) and pathways provided by exon array annotations from Affymetrix. The Coex-Rank approach led to greater enrichments (see Table 8) due to the incorporation of co-expression information.

Table 8. Comparison of annotation enrichment for both aggregation approaches.

	Up-regulated genes		Down-regulated genes	
	#Clusters	#Pathways	#Clusters	#Pathways
Rank-aggregation	5	17	6	11
Coex-Rank	7	21	6	14

However, the Coex-Rank approach prioritizes genes highly correlated with already-highly-ranked genes on the input lists at the cost of sometimes excluding the already-highly-ranked genes. These scenarios arise from the optimization process of rank-aggregation. The Cross-Entropy algorithm generates a series of candidate lists for evaluation (which finds a super list with minimum distance among all input lists). If one candidate list lacks some of the highly-ranked-genes from most input lists, but has genes with counterparts that are already highly ranked, we may still choose this as the super-list. For example, in our analysis gene “Tes” was the 25th ranked gene on the list of up-regulated genes from the rank-aggregation process, but it was absent on the list from the Coex-Rank approach. Its highly correlated counterpart gene “Runx1” was prioritized at rank 25 by the Coex-Rank approach, but it was absent from the list of up-regulated genes from the rank-aggregation process. Therefore, we decided to add non-overlapping genes from the Coex-Rank approach to those 100 genes selected by the rank-aggregation method and in total we promoted 127 up-regulated genes and 129 down-regulated genes to the final reported lists. These up-regulated genes generate 10 clusters according to DAVID (with low classification stringency). For down-regulated genes, 8 clusters were generated. While the above discussion primarily considers sensitivity as the quality metric for evaluating our approach, it should be mentioned that specificity may likewise be controlled by requiring greater concordance among lists, and thus, not including genes in the final list which did not overlap.

3.4 Discussion

3.4.1 Simulation: advantage of combined analysis

To increase confidence that the results described in Section 3.3.2 regarding the advantage of combined analysis over separate analysis was not dataset dependent, we conducted a simulation study consisting of one dataset from exon arrays and one dataset from expression arrays. Each dataset had six samples, three controls v.s. three treatments and each sample covered 18,204 genes. Consider for example a simulated exon array dataset generated as follows:

(1) The sample means μ_i (i = 1, 2, 3…18,204) were from a real dataset. Four arrays using mammary gland were exacted from GSE10246 and the same probe-sets were selected as in our Sppar case study. Sample means were calculated for 18,204 genes separately.

(2) Background variations were added according to the following formulas:

$Y_{ij}=\mu_i+Z_{ij}$ (i=1,2,3…18, 204,j=1,2,3,4,5,6),

$Z_{ij} \sim N(0,\sigma^2)$,

$\sigma = \alpha\times(0.3–0.02\times\mu_i)\times G_i$, G_i~Gamma(5).

Y_{ij} refers to the expression value of the i^{th} gene from the j^{th} sample and α is a parameter controlling the scale of variation [2]. We evaluated α = 0.1, 0.2, and 0.3 to demonstrate different levels of background noise. Here, we also made the assumption that the amount of variation is μ_i dependent. As it is often seen in real data, genes with smaller expression values may exhibit larger proportional variations [2].

(3) The first 200 genes from treated samples were added with differential expression values as follows [2]:

$Y_{ij}=\mu_i+Z_{ij}+\delta_{ij}$ (i = 1,2,3 ……200, j = 4,5,6),

$\delta_{ij} = 0.2\times(2\times B_{ij} - 1) \times G_i$,

B_{ij}~Bernoulli(0.5), $G_i \sim$ Gamma(5).

The simulation data from expression arrays were generated in a similar way. At step 1, the four arrays using mammary gland were extracted from GSE15998 and at step

3, the differential expression value for a specific gene was scaled by the ratio of sample means from two platforms.

We then generated 10 datasets for each platform. We applied both separate and combined analyses including normalization and linear regression followed by an ANOVA test as described in our Methods Section. We used a p-value cutoff of 0.001 to select significantly changed genes. The number of differentially expressed genes was averaged for calculation of sensitivity and specificity and FDR (False Discovery Rate) respectively for expression array data, exon array data and a combined dataset. As shown in Table 9, the combined analysis increases the sensitivity and reduces the FDR compared to the separate analysis, with specificity remaining consistent (around 0.99) at different levels of background noise. The consistency of specificity is due to the nature of microarray data, as the expression levels of most genes are unchanged.

Table 9. Comparison of combined and separate analyses based on simulation.

	Expression Array		Exon Array		Combined Analysis	
	Sensitivity	FDR	Sensitivity	FDR	Sensitivity	FDR
$\alpha = 0.1$	0.89	0.09	0.69	0.11	0.94	0.07
$\alpha = 0.2$	0.67	0.12	0.40	0.17	0.82	0.07
$\alpha = 0.3$	0.47	0.16	0.24	0.28	0.64	0.09

Note: Combined analysis has advantages in increasing of sensitivity and decreasing of FDR. Different background variation has been evaluated via α = 0.1, 0.2 and 0.3.

3.4.2 Similar results from different selection of probe-sets

In Section 3.2.1, we described two approaches to resolve the issue of ambiguous mapping of genes to probe sets. One approach is p-value dependent, in which the probe-set with the most significant p-value is selected; the other is determined by expression value, in which the probe-set with the highest average value of expression is chosen. However, in Section 3.3, only the p-value dependent probe-sets were used in the presentation of our Sppar case study. In fact, the analysis proceeding from the highest average expression value selection approach was repeated and the results were found to be similar.

As mentioned previously, after the p-value dependent dataset was processed through the pipeline of normalization, linear regression and rank-aggregation (including the Coex-Rank method) procedures, 127 up-regulated genes and 129 down-regulated genes were selected. For the original dataset using probe-sets chosen dependent on expression values, 124 genes were significantly up-regulated and 128 genes were down-regulated. The two sets of up-regulated genes held 100 genes (around 80%) in common and the two down-regulated gene sets overlapped with 93 genes (around 72%).

If we only focus on the comparison directly after the probe-set selection step, both approaches selected 18,204 probe-sets from the expression array data respectively and resulted in 13,141 probe-sets (around 72%) in common. In the case of exon array data, the two approaches led to as many as 16,897 overlapping probe-sets (around 92%). At least in this case, it appears that the exon array data are more robust compared to expression array data, however, this may be in part due to the larger sample size.

3.4.3 Similar effects of different statistical tests

In Section 3.2.3, we described a simple linear model followed by an ANOVA test to generate lists of significantly changed genes. We also took advantage of an R package *limma*, which implements a Bayesian linear model [70]. We assumed that 1% of the genes were differentially expressed, as the default setting. The moderated-F test was employed, which is similar to the ordinary F-statistic from ANOVA except that the denominator mean squares are moderated across genes [70]. Following the same normalization schema, two different statistical methods resulted in similar lists of up/down-regulated genes in our Sppar dataset. For two lists of the most highly-ranked 100 genes from each method, they produced at least 70% genes in common (see Table 10, which shows the size of the union set of the two lists from different statistical methods).

Table 10. Size of the union of two lists from different statistical tests.

	NM_1	NM_2	NM_3	NM_4	NM_5
Up-regulated genes	122	126	128	126	126
Down-regulated genes	127	126	128	130	126

Note: The normalization methods (NM) 1-5 are scale-quantile, scale-loess-trim-mean, scale-loess-trim-mean-invariant, scale-loess-median-base, and scale-loess-median-base-invariant sequentially.

The similarity between these two different statistical methods demonstrated by our Sppar case study is at an identical level with the resemblance of different normalization approaches applied to the same set of data. Thus, we could extend the Coex-Rank process to accept input lists from different statistical methods as well.

CHAPTER 4: ANALYSES OF PPARγ RELATED MICROARRAY DATASETS

As we have mentioned in Section 2.1, PPARγ plays an important role in the regulation of vascular function and blood pressure. Previous clinical studies reported that patients with dominant negative (DN) mutations (P467L or V290M) in the ligand binding domain of PPARγ had type II diabetes and early onset hypertension [16]. TZD drugs (e.g. rosiglitazone) are pharmacological ligands of PPARγ. The ligand-mediated activation of PPARγ has been shown to have beneficial effects in lowering blood pressure [47]. To better understand the molecular mechanisms of PPARγ exerting its effects on the genome-wide regulation of transcription, we took advantage of microarray technologies to determine gene expression profiles in mouse thoracic aortas in response to ligand activation or interference with different models of dysfunctional PPARγ. We also examined mesenteric arteries in one of these models.

4.1 Microarray datasets information

4.1.1 Rosi Dataset

Adult male mice (aged 5-7 months) from C57BL/6J strain were used in this experiment. PPARγ was activated by rosiglitazone treatment in a time and dose dependent manner. The administration of rosiglitazone was for either 2 or 14 days at a dose of 3 or 10 mg/kg/day via a custom-made diet otherwise identical to standard chow. This resulted in four different treatment groups. Control mice were fed standard chow. Mouse RNA extracted from thoracic aortas were hybridized to Affymetrix GeneChip

Mouse Genome 430 2.0 array. For each treatment/control group, 3 biological replicates were used, except the group with 14 days rosiglitazone treatment at the dose of 10 mg/kg/day having only 2 samples. In each sample, RNA was pooled from 8-9 different mouse thoracic aortas.

4.1.2 Gppar Dataset

To model the effect of the P467L mutation of PPARγ, Tsai et al generated knock-in mice, replacing one normal PPARγ allele with a P465L allele, which is equivalent to P467L in human [71]. In these mice, the interference with the PPARγ signaling pathway is in all the tissues normally expressing PPARγ, so we consider these mice to be model of "global PPARγ interference" and refer them as the "Gppar" mice. Gppar mice were reported to have hypertension at baseline by Tsai et al [71]. Our lab also confirmed that a 10-mm Hg increase of arterial pressure in male mice [72]. The genetic background of Gppar mice are F1 of 129/SvEv and C57BL/6J strains.

In our microarray study, adult male mice with ages from 5 to 7 months were used. RNA extracted from thoracic aortas were hybridized to Affymetrix GeneChip Mouse Genome 430 2.0 array. There were two groups in this experiment, controls vs. mutant mice, each group with 3 replicated samples. In each sample, RNA was pooled from 8 different mouse thoracic aortas.

4.1.3 Sppar Datasets

General information of Sppar Datasets has been provided in details in Section 3.1 – Motivating Datasets. These mice have been backcrossed to C57BL/6J strain. All mice used in the experiment were male and aged from 5 to 7 months.

4.1.4 Msppar Dataset

The mutant mice involved in this experiment were from the same model as those in Sppar experiment. We used RNA isolated from mesenteric arteries for microarray experiment. The hybridization was performed using Affymetrix mouse exon 1.0 ST array. There were two groups for comparison, 4 controls v.s. 3 transgenic samples. In each sample, RNA was extracted from a single mouse.

4.1.5 Ksppar Dataset

Chang et al generated mice with vascular smooth muscle cell-selective deletion of PPARγ using Cre-loxP system [73]. Unlike our transgenic Sppar mice, these mice were reported to have hypotension, with 14-mm Hg decrease in 6-month-old male mice [73]. As this is a knock-out model and specific to the smooth muscle cells, we label them as “Ksppar” mice. These mice have been backcrossed to C57BL/6J strain. Mice with ages from 5-7 months were used in our microarray study. Mouse RNA from thoracic aortas was hybridized to Affymetrix mouse exon 1.0 ST array. For the experimental design, we had three groups, two groups of controls and one group of smooth muscle cells specific PPARγ knock-out mice. For Cre-control group, the sample size was 5 and the sample size was 3 for flox-control group. The knock-out mice group had 6 samples in total. In each sample, RNA was extracted from a single mouse.

4.2 Methods

4.2.1 Affymetrix Present/Absent Calls Detection Method

A detection call helps to answer the question that whether a gene is expressed in a particular biological sample. Present call refers to the expressed status of a gene, while absent call indicates non-distinguishable from background noise. For expression arrays, Affymetrix MAS 5.0 algorithm can be used to achieve the detection calls [74]. Though MAS 5.0 algorithm for generating expression summaries has been criticized for high False Positive, its approach to determine present/absent calls is still widely used [75]. MAS 5.0 algorithm calculates a discrimination score for each probe pair using both Perfect Match (PM) and Mismatch (MM) probes' information. Next, each probe-set is assigned a detection P-value from one-sided Wilcoxon's signed rank test, based on previous discrimination scores. To make the call, two significance levels $\alpha_1 = 0.04$ and $\alpha_2 = 0.06$ are set as the default. If P-value < α_1, it is labeled "Present"; if P-value >= α_2, it is called "Absent"; otherwise, it is assigned "Marginal" [76]. The software is available as an R package called "*affy*", which can be directly installed from Bio-conductor. The specific function is "*mas5calls( )*".

For exon arrays, the corresponding algorithm is termed as Detection Above BackGround (DABG) [77]. A detection metric of each PM probe is generated based on comparison to a distribution of background pool with the same GC content. Then, interrelated probe level p-values are combined into a probe-set level p-value using

Fisher’s method [77]. The software is available from an Affymetrix tool – Expression Console. Along with exon-level normalization, the output file of DABG will be automatically generated, in which a p-value is associated with each exon.

4.2.2 Extension of Coex-Rank featured approach

As we have discussed in Chapter 3, Coex-Rank was applied to generate robust results from different normalization methods. It can be applicable to merge gene lists from different comparison groups or multiple statistical tests as well. Thus, our Coex-Rank solution also provides an alternative to a seemingly arbitrary choice among many good approaches.

4.2.3 Gene set enrichment analysis (GSEA)

Normally, analysis at single-gene level reveals little similarity between/among independent microarray studies, even with related biological experimental design. Gene set enrichment analysis (GSEA) [78] gains its power by taking advantage of gene-sets. Genes within a set share common biological function or locate near each other on the same chromosome or demonstrate similar regulatory pattern. Typically, genes from microarray experiments are ranked according to their differential expression between control and treatment groups [78]. The core of GSEA is to determine if a set of candidate genes tend to be near the top/bottom of the pre-ranked gene list or just randomly distribute through the whole gene list. This approach helps to interpret biological significance of the microarray data from a gene-set perspective [78]. For the mathematical description, an enrichment score is calculated for a gene set by walking down the pre-ranked gene list. When we encounter a gene in the gene set, the running-sum statistic increases; otherwise, it decreases. The final score is the maximum absolute value of the running-sum and this is actually a weighted Kolmogorov-Smirnov-like statistic [78]. Next, the enrichment score is normalized, taking the size of the gene set into account. Finally, False Discovery Rate (FDR) is computed based on permutations [78]. The software is freely available for download (http://www.broadinstitute.org/gsea/msigdb/downloads.jsp).

4.2.4 DAVID, a tool for functional enrichment analysis

Microarray technologies usually produce lists of interesting genes under a certain biological condition. The Database for Annotation, Visualization and Integrated Discovery (DAVID) [69] provides a promising strategy to extract biological meanings out of these gene lists. Compared to other publicly available tools, DAVID has some advanced capabilities, especially in a comprehensive database and a novel module-centric algorithm [69]. The database is known as DAVID knowledgebase. It integrates more than 40 heterogeneous gene annotation resources, particularly across NCBI and UniProt systems. This wide range of information facilitates high throughput gene functional analysis.

For the novel algorithm, it is designed to group functionally related genes into biological modules to identify pertinent biological processes in a study [69]. Before grouping related genes, a method of measuring gene-gene similarity is implemented, based on the assumption that functionally related genes share global annotation profiles with each other. Then a DAVID agglomeration algorithm is applied for gene grouping, which allows a gene to participate in more than one functional group. This fuzziness feature better reflects the nature of genes – playing multiple roles in a biological system. The software is freely available online (http://david.abcc.ncifcrf.gov/).

4.3 Results

4.3.1 Present/absent status of genes in microarrays

The sets of microarray data involve mice with different genetic backgrounds (including C57BL/6J, backcrossed to C57BL/6J and F1 of 129/SvEv and C57BL/6J strains) and various vascular tissues (thoracic aortas v.s. mesenteric arteries). To obtain the present/absent calls of genes on microarrays, we took advantage of MAS5 and DABG algorithms, respectively for expression arrays and exon arrays. We used genetically-matched wild-type mice for these analyses and only focused those genes covered by both platforms (18,204 common genes determined by highest average expression levels as described Section 3.2.1).

We used the function "*mas5calls( )*" from "*affy*" R package to generate output files containing present/absent information of genes for data from expression arrays. There were three different labels in these files, "Present", "Absent" and "Marginal". We considered the "Marginal" label the same as "Absent" to make a relatively strict rule. The numbers of genes with "Present" calls are shown in Figure 10 for Rosi dataset, Gppar dataset and the part of expression arrays from Sppar dataset. The genetic background of Gppar mice is different from the other two groups of mice. The numbers of genes in the intersection of any pair of datasets were also calculated, shown with percentage in Figure

10. On average, 95% genes are in common between any two groups, though there are two different genetic backgrounds involved.

	Rosi	Gppar	Sppar
#genes	9979	9873	10532

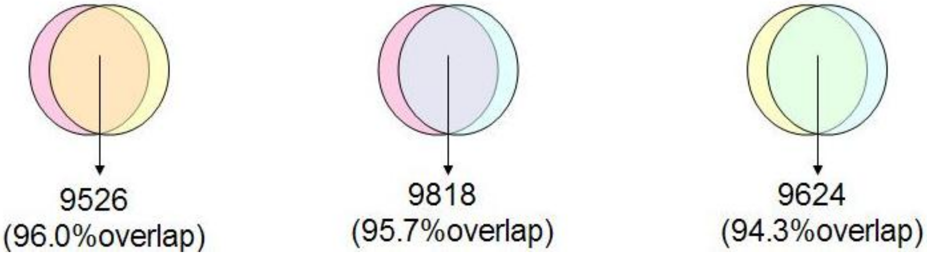
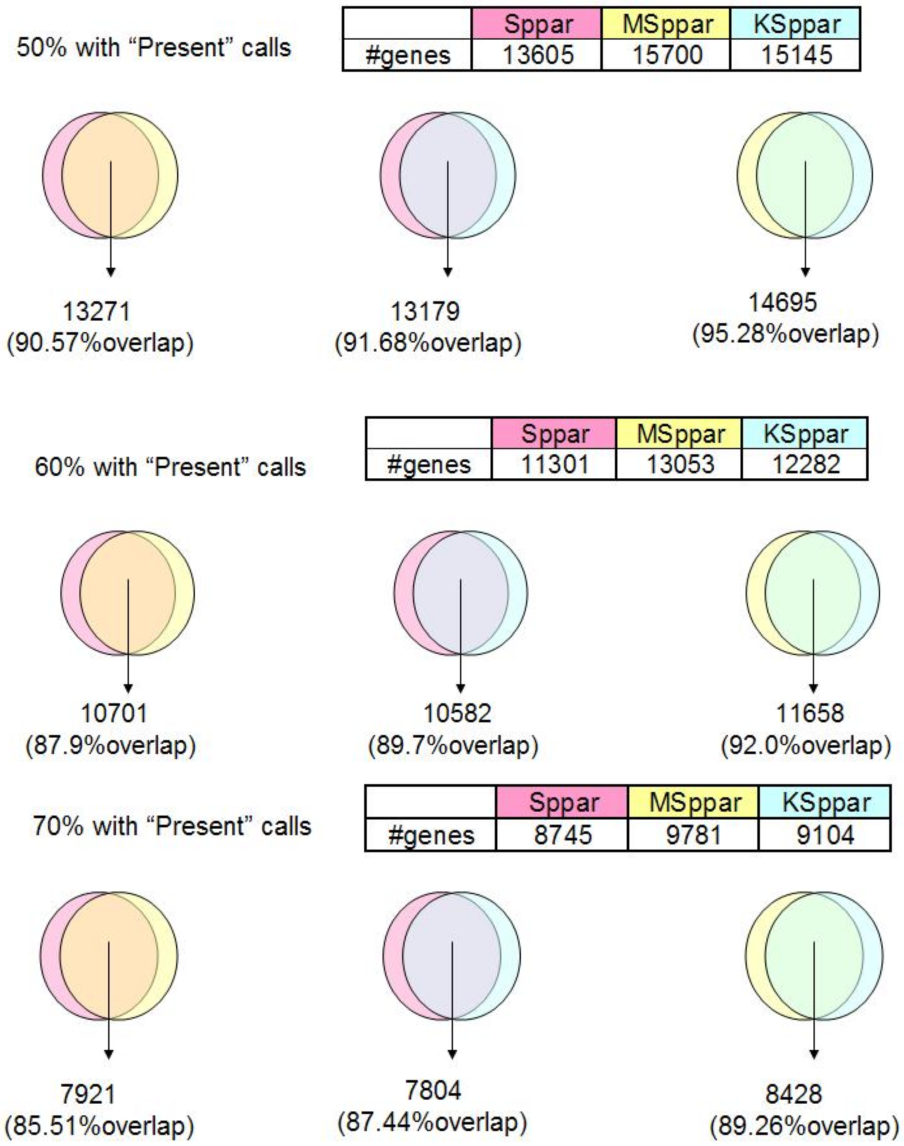


Figure 10. Intersection of expressed genes from expression array experiments. Pink stands for Rosi dataset, yellow stands for Gppar dataset and blue stands for the part of expression arrays from Sppar dataset.

For the present/absent call analysis regarding the exon arrays, DABG method only provides p-values that indicate expression statuses at exon level, not at gene level. As it is recommended by a recent study for exon array analysis, an exon with p-value < 0.05 can be considered as “Present” and a gene having 50% of its exons expressed can be assigned a “Present” call [79]. We followed this method and also explored 60% and 70% cutoff values in addition to 50%. The results are shown in Figure 11.



	Sppar	MSppar	KSppar
#genes	13605	15700	15145

	Sppar	MSppar	KSppar
#genes	11301	13053	12282

	Sppar	MSppar	KSppar
#genes	8745	9781	9104

Figure 11. Intersection of expressed genes from exon array experiments. Different parameters used in determination of genes' expression status. A gene having 50%, 60% or 70% of its exons expressed is assigned a "Present" call.

For the three datasets in our analysis, Msppar experiment used mouse mesenteric arteries and the other two used mouse thoracic aortas. However, the gene expression status in the Msppar experiment was slightly more similar to Ksppar experiment, compared with other pair-wise groups, no matter which cutoff value was used. This scenario can not be explained by difference of vascular tissues involved in these experiments. It might be due to variations from experimental processing, as RNA of Msppar and Ksppar experiments were extracted by the same person and microarray hybridizations were carried out at almost the same time.

4.3.2 Identification of DEGs from each microarray dataset

DEGs from Sppar dataset were generated as described in Chapter 3. For the other four experiments, we extended the Coex-Rank featured approach to identify DEGs from each microarray dataset. As different experimental designs were involved in all the five datasets, different strategies were selected for each analysis. For the pre-processing part of microarray data, the RMA algorithm was used as described in Section 3.2.1, which resulted in 45,101 probe-sets on expression array platform and 101,176 gene-level probe-set records on exon array platform. For exon array data, probe-sets without annotations such as gene symbols or mRNA accession information (according to Affymetrix annotation, release 30) were removed and resulted in smaller datasets, each with 40,434 probe-sets. In these four datasets, we retained multiple probe-sets matching the same genes to avoid introducing any bias in selection of probe-sets, which was required in combining datasets from different microarray platforms.

For the Rosi dataset, there were four groups of different rosiglitazone treatments and only one group of controls. Therefore, we compared each treatment group with the control group and resulted in four separate comparisons. For each comparison, as the sample size was small, we took advantage of an R package *limma*, which implements a Bayesian linear model [70]. This method has been recommended as a significance test for microarray data with small sample size [2]. For example, for up-regulated probe-sets, considering p-value < 0.01 as the cut-off, 4 probe-set lists were generated, each from one comparison group. Then, these probe-sets were ranked either by p-value or fold change, which resulted in 8 different lists. To generate similar number of DEGs as in Sppar dataset, we also selected 100 most highly-ranked probe-sets from each list as the input for both the rank-aggregation and the Coex-Rank approaches.

For Gppar dataset, only one group of comparison was available and the sample size was also small. We applied both Student's T test with equal variance and *limma* method to calculate the p-values. After filtering by p-value < 0.01, probe-sets were ranked either by p-value or fold change, which resulted in 4 different lists for up/down-regulation. To be consistent, we focused on 100 most highly-ranked probe-sets for both the rank-aggregation and the Coex-Rank approaches.

For Sppar dataset, details have been provided in Chapter 3. We also carried out analysis only based on data from exon arrays using the same approach applied to the above Gppar dataset.

For Msppar dataset, the same strategy as the Gppar dataset was adopted as well. The only difference was the selection of threshold p-value < 0.02. Different p-value cut-off values were used for each dataset, because we aimed to select a reasonable number of probe-sets (200-300 probe-sets), which could be ranked by either p-value or fold change later.

For Ksppar dataset, there were two groups of controls and only one group of knock-out mice. First, we compared the two control groups and found 1209 probe-sets to be differentially expressed according to p-value < 0.01 from Student's T test. If we compared one group of controls with the knock-out group, using the same cut-off values, we could select 1037 probe-sets in comparison using flox-controls and 1052 probe-sets with cre-controls. The levels of differences were quite similar. Therefore, we focused on those probe-sets with robust behaviors in both comparisons and chose a less stringent criterion p-value < 0.05. Ranking by either p-value or fold change and 100 most highly-ranked probe-sets selection were used as routines to generate input lists of both the rank-aggregation and the Coex-Rank processes.

The final reported lists of probe-sets were merged from the output of both the rank-aggregation and the Coex-Rank processes. Then we took care of the issue regarding multiple probe-sets matching a unique gene. The numbers of DEGs identified in each dataset were summarized in Table 11.

Table 11. Numbers of unique DEGs identified in each microarray dataset.

	Rosi data	Gppar data	Sppar data	Msppar data	Ksppar data
# up genes	129	139	127	119	135
# down genes	141	138	129	121	131

DAVID analysis was performed on each list of DEGs and we obtained some interesting biological explanations about these genes. The default lowest clustering criterion of DAVID was used, as we would like to explore as much biological information as possible, which might help us to make hypothesis for future experimental investigation. Briefly, for genes up-regulated in Rosi experiment, they are enriched with annotations especially related to PPAR signaling pathway and positive regulation of transcription, which fits with the molecular mechanism of ligand dependent activation of PPARγ. For down-regulated genes in Gppar experiment, they are also enriched with PPAR signaling pathway annotation. Moreover, they have another gene cluster annotated with negative regulation of transcription, which may explain the down-regulation pattern. The most attractive functional characteristic of Sppar and Msppar datasets is that up-regulated genes are enriched with calcium ion binding. The most up-regulated gene in Sppar experiment is Tnnc1 (troponin C type 1) with ~ 20 fold change robustly in both expression array and exon array platforms. It has also been validated by real-time PCR to have more than 100-fold change (see Figure 12).

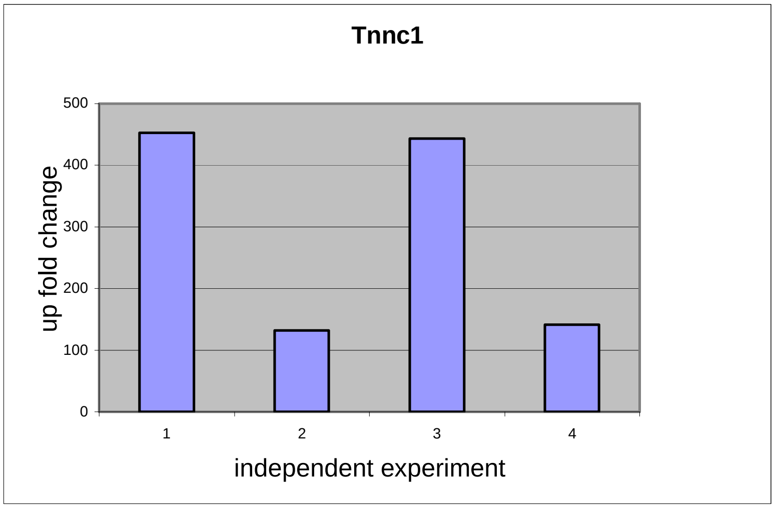


Figure 12. Real-time PCR validation of up-regulated Tnnc1 in Sppar mice. Four independent experiments show robustly more than 100-fold up-regulation of Tnnc1.

According to the Gene database of NCBI (National Center for Biotechnology Information), the protein encoded by Tnnc1 (Troponin C) is a subunit of Troponin, which is a central regulatory protein of striated muscle contraction. The binding of calcium to Troponin C abolishes the function of Troponin I (an inhibitor of actomyosin ATPase), thus allowing the interaction of actin with myosin, the hydrolysis of ATP, and leading to the generation of muscle contraction. Tnnc1 has been classified as one of the hypertrophic cardiomyopathy susceptibility genes, mutations of which (A8V and D145E) affect the functional properties of Troponin C by increasing the Ca^{2+}-sensitivity of contraction [19].

4.3.3 Comparisons of all PPARγ related datasets

We performed pair-wise comparisons of these 10 lists of DEGs from 5 microarray datasets, but limited numbers of genes (0~5 genes) were found to be in the intersections. This scenario is quite common in microarray studies. Previous research has shown that microarray analysis at single-gene level reveals little similarity even with related biological experimental design. In fact, this is the motivation for the development of Gene Set Enrichment Analysis (GSEA) [78], which gains power by taking advantage of gene sets. Therefore, we utilized this tool to observe general biological patterns in comparisons of all the 5 PPARγ related microarray datasets. We focused only on the genes covered by both expression and exon arrays (18,204 common genes determined by highest average expression levels as Section 3.2.1), because the comparisons were across all the datasets.

As the input required by GSEA, there should be a pre-ranked list of all genes involved in a microarray experiment and several gene-sets. GSEA helps to answer the question that if genes in a set have differentially expressed patterns in a microarray experiment. Generally, there are two approaches to rank all the genes in an experiment, either by p-value from a statistical test or fold change. We explored both of them to draw robust conclusions.

When we used p-value to rank all genes in a microarray experiment, the query gene sets were also selected by most significant p-value strategy in order to be consistent. The p-values based on which we ranked the genes were generated from Students' T test with equal variance. For some datasets (e.g. Rosi experiment including 4 comparisons of controls v.s. treatments), we simply averaged p-values from each comparison. The most significantly up-regulated genes were ranked near the top of the list, while the most significantly down-regulated genes were near the bottom. In total, we had 5 pre-ranked gene lists and 10 gene sets. For gene sets, each was with 100 most up/down-regulated genes from an experiment. All the implementations were carried out in a similar way for fold change ranking approach.

For the parameter setting of GSEA, the default 1,000 permutations were used to calculate False Discovery Rate (FDR). The results were organized into Table 12 and Table 13, respectively for both p-value and fold change approaches. In Table 12, the first row refers to the gene sets. For instance, rosi_up gene set contains 100 most up-regulated genes according p-value in Rosi dataset. The first column refers to the pre-ranked gene lists. For example, Rosi_expression list has all the genes ranked by p-value, with up-regulated genes near the top and down-regulated genes near the bottom. All the numbers in the table are FDRs from statistical tests. A value of 0.00000 can be observed, which indicates the real number is < 0.00001. The plus (+) or minus (-) sign before the number indicates up or down regulation, respectively. In fact, it was not necessary to run rosi_up and rosi_down gene sets with pre-ranked lists from Rosi experiment, they were known to be near the top/bottom of the list. We still included them in our analysis, as we considered them as internal controls. In the result tables, we highlight all these types of internal controls using yellow color. We used FDR < 0.01 as cut-off values. We highlight all the robust (consistent result from both p-value and fold change approaches) and significant FDRs with either pink or blue colors, where pink indicates up-regulation and blue refers to down-regulation.

Table 12. Summary table of FDRs from GSEA using p-value ranking method.

	rosi_up	rosi_down	gppar_up	gppar_down	sppar_up
Rosi_exp	+0.00000	-0.00000	-0.00140	+0.00000	-0.00000
Gppar_exp	-0.00000	-0.00196	+0.00000	-0.00000	+0.28406
Sppar_com	-0.00000	+0.00154	+0.37313	-0.00531	+0.00000
Msppar_exon	-0.00000	-0.06255	+0.70554	-0.00281	+0.00000
Ksppar_exon	-0.00000	-0.21224	+0.79238	-0.00000	-0.00000
	sppar_down	msppar_up	msppar_down	ksppar_up	ksppar_down
Rosi_exp	-0.00104	+0.12024	+0.00050	+0.00674	+0.00038
Gppar_exp	+0.00432	+0.18745	+0.37467	+0.00000	-0.00065
Sppar_com	-0.00000	+0.43570	-0.00000	-0.00000	+0.28070
Msppar_exon	-0.00000	+0.00000	-0.00000	-0.00000	-0.00000
Ksppar_exon	+0.00000	+0.00949	+0.00000	+0.00000	-0.00000

Note: We used FDR < 0.01 as cut-off values and a value of 0.00000 indicates the real number is < 0.00001. The plus (+) or minus (-) sign before the number indicates up or down regulation, respectively. Yellow indicates internal controls, pink stands for up-regulation and blue refers to down-regulation.

Table 13. Summary table of FDRs from GSEA using fold change ranking method.

	rosi_up	rosi_down	gppar_up	gppar_down	sppar_up
Rosi_exp	+0.00000	-0.00000	-0.01412	+0.00950	-0.00000
Gppar_exp	-0.00000	+0.00000	+0.00000	-0.00000	+0.04008
Sppar_com	-0.00000	+0.00832	-0.43329	-0.65847	+0.00000
Msppar_exon	-0.00075	+0.29549	+0.78901	-0.48735	+0.11544
Ksppar_exon	-0.00027	-0.02466	+0.17062	-0.88936	-0.00000
	sppar_down	msppar_up	msppar_down	ksppar_up	ksppar_down
Rosi_exp	-0.00000	+0.54647	-0.10437	-0.91511	-0.00108
Gppar_exp	+0.00063	-0.88050	-0.92694	+0.00000	+0.45364
Sppar_com	-0.00000	+0.00000	-0.00000	-0.00000	+0.00014
Msppar_exon	-0.00000	+0.00000	-0.00000	-0.26204	-0.28184
Ksppar_exon	+0.00000	+0.23956	+0.03070	+0.0000	-0.00000

Note: We used FDR < 0.01 as cut-off values and a value of 0.00000 indicates the real number is < 0.00001. The plus (+) or minus (-) sign before the number indicates up or down regulation, respectively. Yellow indicates internal controls, pink stands for up-regulation and blue refers to down-regulation.

Based on these summary tables of FDRs, we could identify several pair-wise patterns with biological meanings, shown as the following:

(1) First of all, rosi_up gene-set was significantly down-regulated in Gppar experiment and gppar_down gene-set was significantly up-regulated in Rosi experiment. These genes that were up-regulated by rosiglitazone treatment but down-regulated in Gppar mutant mice might be direct targets of PPARγ.

(2) Secondly, rosi_down gene-set was significantly up-regulated in Sppar experiment and sppar_up gene-set was significantly down-regulated in Rosi experiment. These genes were down-regulated by rosiglitazone treatment and up-regulated in Sppar mutant mice. They might be affected indirectly by PPARγ transcriptional regulation.

(3) The third pattern was that sppar_down gene-set was significantly down-regulated in Msppar experiment and msppar_down gene-set was also significantly down-regulated in Sppar experiment. This similarity might be due to the same transgenic mice model used in two experiments, suggesting that dominant negative mutation of PPARγ is responsible for down-regulation of genes in different blood vessels.

(4) The last one involved significantly down-regulated genes in Sppar experiment and significantly up-regulated genes in Ksppar experiment. And these genes might help to explain the different phenotypes of two PPARγ dysfunctional mice models, specifically, hypertension v.s. hypotension.

To further extract biological meanings out of these gene lists with interesting pair-wise regulation patterns, we utilized a tool called DAVID to generate gene clusters based on functional annotations. The default lowest clustering criterion was used as well. For example, for genes with up-regulated pattern in Rosi experiment and down-regulated feature in Gppar experiment, we had two separate gene sets, one from p-value ranking and the other from fold change ranking. We used both of them for DAVID analysis and reported the consistent functional enrichment from different input lists of genes.

According DAVID analysis, genes with expression pattern “rosi_up and gppar_down” generated two clusters (shown in Table 14 and Table 15) that are functional enriched by annotation “PPAR signaling pathway”. Genes in Table 14 are also enriched with annotations, such as fatty-acid binding, lipid binding and cytoplasm. Rbp7 (retinol binding protein 7) has been shown to be highly correlated with PPARγ at the levels of expression across different tissues [80] and the other two genes are classic targets of

PPARγ. Genes in Table 15 are functional enriched with glycoprotein and transmembrane annotations.

Table 14. First cluster of genes enriched with “PPAR signaling pathway”.

Probe-set ID	Gene Name
1449461_at	retinol binding protein 7, cellular
1416023_at	fatty acid binding protein 3, muscle and heart
1417023_a_at	fatty acid binding protein 4, adipocyte

Note: Genes are also enriched with annotations, such as fatty-acid binding, lipid binding and cytoplasm.

Table 15. Second cluster of genes enriched with “PPAR signaling pathway”.

Probe-set ID	Gene Name
1450883_a_at	CD36 antigen
1418848_at	aquaporin 7
1418197_at	uncoupling protein 1 (mitochondrial, proton carrier)
1417130_s_at	angiopoietin-like 4

Note: Genes are functional enriched with glycoprotein and transmembrane annotations.

Genes with expression pattern “sppar_down and msppar_down” led to a cluster with “GTPase” annotation (see Table 16). Previous study in our laboratory showed that the response of thoracic aortas from Sppar mice to a vasoconstrictor – the peptide endothelin-1 (ET-1) was remarkably augmented, compared with normal mice. Treatment with the Rho kinase-specific inhibitor Y27632 resulted in a significant inhibition of the contractile response, which suggested the dependence upon Rho kinase activity [13]. Furthermore, the increasing activity of Rho kinase in Sppar mice has been confirmed recently by our laboratory. Rho kinase is known as an effector of RhoA. Rho subfamily of small GTPases can be divided into five groups: the Rho-like (RhoA-C), Cdc42-like, Rac-like, Rnd-like and RhoBTB (RhoBTB1-3) proteins [20]. The functions of RhoBTB proteins have not been defined yet. One member, RhoBTB2 has been identified as a component of cullin3-dependent ubiquitin ligase complexes [20]. This property is highly probably shared with other RhoBTB members [20]. Cullin3 is the core subunit of CRL3 (Cullin-RING ubiquitin ligase 3), which was discovered by Chen et al to target the RhoA for degradation [81]. Therefore, we hypothesize that the Rho kinase activity may lie downstream of RhoBTB1 in Sppar mice. The down-regulation RhoBTB1 might cause the increasing of Rho kinase activity via RhoA/Rho kinase signaling pathway. The down-regulation of RhoBTB1 has already been confirmed by real-time PCR in Sppar mice (see Figure 13).

Table 16. A cluster of genes enriched with “GTPase” annotation.

Probe-set ID	Gene Name
1422562_at	Ras-related associated with diabetes
1428067_at	RAS-like, family 12
1437100_x_at	proviral integration site 3
1429206_at	Rho-related BTB domain containing 1
1432415_at	RAB3C, member RAS oncogene family

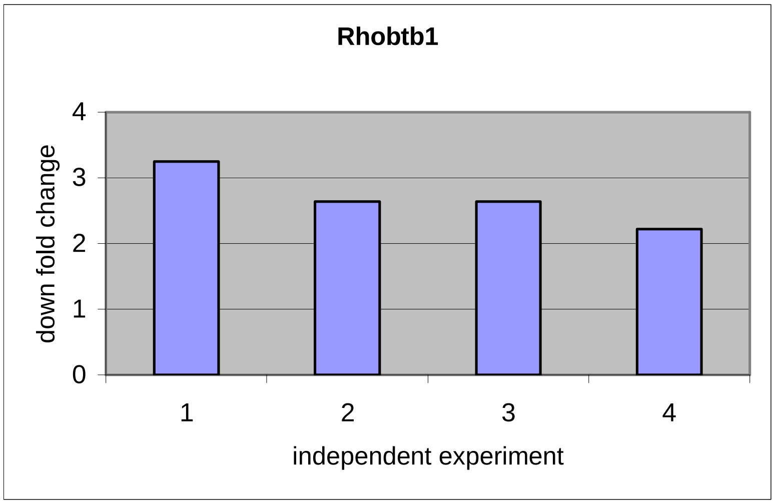


Figure 13. Real-time PCR validation of down-regulated Rhobtb1 in Sppar mice. Four independent experiments show robustly more than two fold down-regulation of Tnnc1.

Genes with expression pattern “sppar_down and ksppar_up” resulted in a cluster with “kinase” annotation (Table 17). Genes produce proteins with functions of phosphorylation might be targets for future investigation. As we have discussed above, Rho kinase activity is increased in Sppar mice, which is related with vasoconstriction. Take Camk4 (calcium/calmodulin-dependent protein kinase IV) for example, the product of this gene belongs to the Ca^{2+}/calmodulin-dependent protein kinase subfamily [82]. Though it has not been demonstrated to be involved in blood pressure regulation, its sibling Camk2 (calcium/calmodulin-dependent protein kinase II) is implicated to inhibit vascular smooth muscle contraction [82]. As shown in Figure 14, phosphorylation of MLC (light chain of myosin) by MLCK (myosin light chain kinase) allows interaction of myosin and actin, which results in vascular smooth muscle contraction. MLCK is activated by the Ca^{2+}-CaM (Ca^{2+}-calmodulin complex) and it can be inhibited by Camk2 [82].

Table 17. A cluster of genes enriched with “kinase” annotation.

Probe-set ID	Gene Name
1435319_at	inositol hexaphosphate kinase 2
1416069_at	phosphofructokinase, platelet
1451140_s_at	protein kinase, AMP-activated, gamma 2 non-catalytic subunit
1437100_x_at	proviral integration site 3
1434513_at	ATPase type 13A3
1422084_at	BMX non-receptor tyrosine kinase
1453817_at	ATP-binding cassette, sub-family A (ABC1), member 6
1452572_at	calcium/calmodulin-dependent protein kinase IV
1431167_at	diacylglycerol kinase, gamma

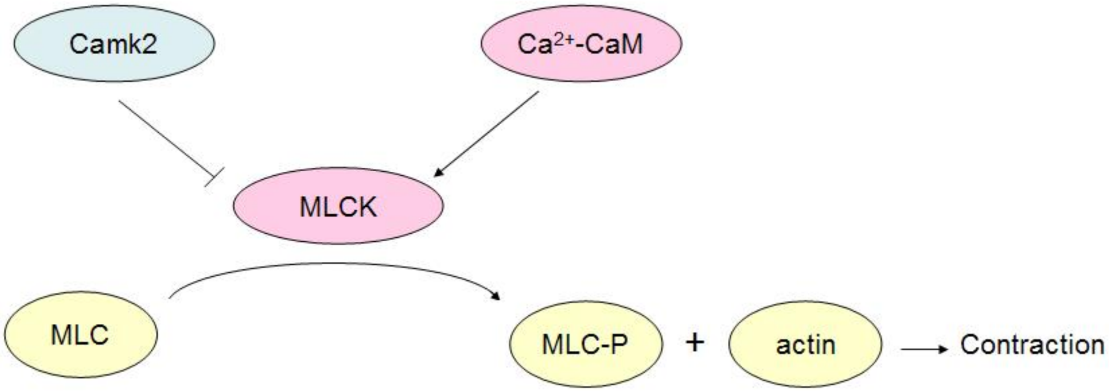


Figure 14. Camk2 involved in biological pathway of vascular smooth muscle contraction.

In addition to these pair-wise patterns, there are two other interesting patterns involved more experiments at the same time. The first is that rosi_up gene-set are robustly and significantly down-regulated in all the other four experiments. This is biological reasonable, as all the other four experiments using mice with dysfunctional PPARγ or even without PPARγ. When we focused on individual genes that are satisfied with this criterion, only a few genes could be selected, shown in Table 18.

Cd36 (CD36 antigen) and Fabp4 (fatty acid binding protein 4) are classic target genes of PPARγ; Ppargc1b (peroxisome proliferative activated receptor, gamma, coactivator 1 beta) is the co-activator of PPARγ. Therefore, their down-regulation fits with the PPARγ dysfunctional models reasonably. Inactivation of RhoA by PKA (Protein Kinase A) / PKG (Protein Kinase G) has been reported, but the underlying mechanism is still not fully understood. Prkar2b (protein kinase, cAMP dependent regulatory, type II beta) belongs to PKA and its down-regulation re-indicates the activation of RhoA/Rho kinase signaling pathway in PPARγ dysfunctional models [83].

Table 18. Genes up-regulated in Rosi experiment, but down in others.

Probe-set ID	Gene Name	
1450883_a_at	CD36 antigen	
1417765_a_at	amylase 1, salivary	
1417561_at	apolipoprotein C-I	
1456611_at	family with sequence similarity 13, member A	
1417023_a_at	fatty acid binding protein 4, adipocyte	
1449945_at	peroxisome proliferative activated receptor, gamma, coactivator 1 beta	
1423972_at	electron transferring flavoprotein, alpha polypeptide	
1438664_at	protein kinase, cAMP dependent regulatory, type II beta	
Probe-set ID	Functional Annotations	Pathway
1450883_a_at	lipoprotein, transport	PPAR signaling pathway
1417765_a_at	Calcium, carbohydrate metabolism	Starch and sucrose metabolism
1417561_at	lipid transport	
1456611_at	phosphoprotein	
1417023_a_at	lipid binding, transport	PPAR signaling pathway
1449945_at	activator, transcription	
1423972_at	electron transport, flavoprotein	
1438664_at	acetylation, cAMP binding	Insulin signaling pathway

Another interesting pattern is that “rosi_up and msppar_down and ksppar_up” gene-sets were all significantly down-regulated in Sppar experiment. The scenario perfectly matches with a predictive model of PPARγ’s molecular mechanism [47]. As shown in Figure 15, for Sppar/Msppar mice with DN mutation PPARγ, without binding of ligands, corepressors and histone deacetylase are recruited, which results in repression of transcription of PPARγ target genes. For wild-type mice, treatment using rosiglitazone provides ligands for PPARγ. Therefore, coactivators, such as PGC1α and SRC, are recruited instead of corepressors, and PPARγ target genes are activated or up-regulated. For Ksppar mice, PPARγ is knocked out, there might be other transcription factors binding to PPRE, like heterodimer of RXRs, which leads to intermediate level of transcription of PPARγ target genes. This model was proposed to explain the hypotension phenotype of Ksppar mice [47].

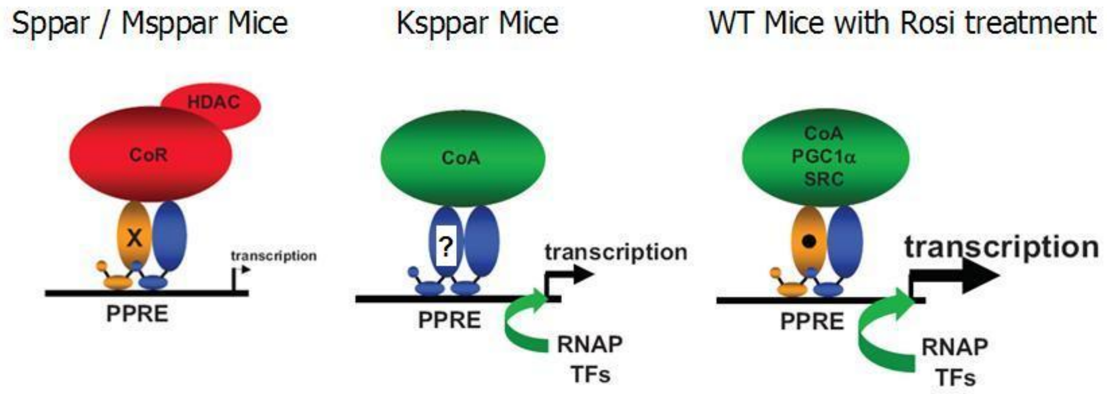


Figure 15. Postulated model of PPARγ's molecular mechanism. The length and width of the arrows reflect levels of transcription. PPARγ is orange and RXR is blue. CoR indicates corepressor; HDAC, histone deacetylase; CoA, coactivator; PGC1α, PPARγ coactivator-1 alpha; SRC, steroid receptor coactivator; RNAP, RNA polymerase; TF, transcription factors; PPRE, PPARγ response element. The figure is modified from a review paper [47].

Genes falling in this criterion are shown in Table 19. Serpine1 (serine/cysteine peptidase inhibitor, clade E, member 1) has the function of regulation of angiogenesis, which is a physiological process involving the growth of new blood vessels from pre-existing vessels, according to Gene database of NCBI. It is also annotated with positive regulation of coagulation (a complex process of blood forming clots), in the same database.

Table 19. Genes consistent with a predictive working model of PPARγ.

Probe-set ID	Gene Name
1418719_at	4HAUS augmin-like complex, subunit 8
1450554_at	defensin beta 2
1416021_a_at	fatty acid binding protein 5, epidermal
1452388_at	heat shock protein 1A
1437100_x_at	proviral integration site 3
1417466_at	regulator of G-protein signaling 5
1419149_at	serine (or cysteine) peptidase inhibitor, clade E, member 1

Probe-set ID	Functional Annotations	Pathway
1418719_at	alternative splicing, mitosis	
1450554_at	antibiotic, defensin	
1416021_a_at	lipid-binding, transport	PPAR signaling pathway,
1452388_at	atp-binding, chaperone	MAPK signaling pathway, endocytosis
1437100_x_at	atp-binding, protein kinase	
1417466_at	signal transduction inhibitor	
1419149_at	plasminogen activation	p53 signaling pathway, coagulation cascades

Rgs5 (regulator of G-protein signaling 5) is involved in the regulation of heterotrimeric G proteins by acting as GTPase activators. Rgs5 has been identified as an antihypertensive target (down-regulated in hypertensive models) by vascular microarray profiling in two rat models of hypertension [84]. Changes in RGS (regulator of G-protein signaling) expression and function correlate with vascular remodeling. Xi Wang et al showed that RGS5 inhibits angiotensin II (angII)-induced signaling in smooth muscle cells by *in vitro* overexpression studies [21]. It is known that angII leads to vasoconstriction [82].

CHAPTER 5: INVESTIGATION OF PPARγ RELATED DEGS WITH PPARES

As introduced in Chapter 2, PPARγ heterodimerizes with RXR and binds to DNA elements called PPAR Response Elements (PPREs). The known PPREs have a consensus sequence of AGGTCAXAGGTCA, with X being a random nucleotide. This initial conclusion was based on analysis of approximately 30 well-characterized PPARγ target genes [25]. More recent studies, as described below, have characterized PPARγ binding sites on a genome-wide level.

5.1 Genome-wide binding sites of PPARγ from literature

5.1.1 PPREs identified by ChIP-chip technique

Leferova et al used ChIP (chromatin immunoprecipitation) followed by hybridization to whole genome tiling arrays (ChIP-chip) to determine the genome-wide binding sites of PPARγ [17]. PPARγ is expressed at highest levels during differentiation of mouse fibroblasts into adipocytes. The ChIP assay was performed using 3T3-L1 adipocytes harvested at day 10 post-hormonal induction of adipogenesis. Immuno-precipitated chromatin fragments were hybridized to the whole genome Mouse Tiling 2.0R Array Set (Affymetrix). The model-based analysis of tiling arrays (MAT) was employed to determine genome-wide binding sites of PPARγ. Filtering by the cutoffs of FDR (False Discovery Rate) <=1% and enrichment of PPARγ signal over control IgG equal to or greater than two fold, 5299 PPREs were identified (mapped to reference genome mm8), with each region about 1000 base pair (bp) long [17].

The results from ChIP-chip experiment were validated by ChIP-quantitative PCR (QPCR) and with a different PPARγ antibody. These additional studies suggested that the actual FDR was around 5%. The previously known motif for PPREs was found at most of the identified binding sites. Moreover, ChIP for RXRα showed a co-localization pattern with PPARγ at nearly all the tested binding sites [17].

5.1.2 PPREs identified by ChIP-seq technique

Nielsen et al took advantage of deep sequencing technology, combined with ChIP assay to map genome-wide PPARγ binding sites [18]. The samples for ChIP were 3T3-L1 cells at day 6 post differentiation. For data analysis, mapping short sequences to mouse genome was performed by the Illumina Analysis Pipeline and detection of binding sites was carried out using FindPeaks at an FDR level < 0.001. They identified 6952 binding sites for PPARγ (mapped to mm9 reference genome), most of which were co-localized with RXRα binding sites. Motif search was also carried out and showed positive re-confirmation, matching with the known PPRE [18].

5.2 Methods

5.2.1 Bowtie – a tool for alignment

With the development of high-throughput sequencing technology, it is necessary to optimize the alignment of large sets of short DNA sequences (reads) to large genomes. Bowtie is an ultrafast, memory-efficient short read aligner, which facilitates further sequence analysis [85]. For example, Bowtie aligns 35-base-pair reads to the human genome at a rate of 25 million reads per hour on a typical workstation. It has been shown to consume less time and computational resources than other similar tools, such as Maq and SOAP [85]. The efficiency of Bowtie lies in a novel indexing strategy – Burrows-Wheeler index to keep its memory footprint small. However, Bowtie is not a general-purpose alignment tool like BLAST. For the current version, it does not yet report gapped alignments [85]. The software is freely available from http://bowtie-bio.sourceforge.net.

5.2.2 MACS – a tool for peak detection in ChIP-seq

Chromatin immunoprecipitation (ChIP), integrated with advanced next-generation sequencing technology (ChIP-seq), [86] provides data that only represent the ends (about 30 bp) of the ChIP fragments. These short sequences (called tags) are aligned to a reference genome and then their accumulations at different genomic loci are detected as the binding sites of transcription factors. There are various algorithms for the peak detection part [86].

Generally, ChIP-seq experiments can be performed with or without control samples. For data without controls, peaks are detected against background information. To model the background noise, Poisson distribution is preferred to uniform distribution, which has been used previously [86]. According to a recent study comparing different methods for peak detection, two methods – MACS and SISSRs are available for the public ChIP-seq data analysis. The better performance of MACS over SISSRs is supported by reproducibility of the detections and external validation using qPCR, at least in a third party comparison study [86].

MACS stands for Model-based Analysis of ChIP-Seq. It improves the spatial resolution of predicted binding sites by modeling the shift size of tags empirically [87]. Furthermore, the dynamic Poisson distribution employed by MACS can effectively capture local biases in the genome. The software is available from http://liulab.dfci.harvard.edu/MACS/.

5.3 Results

5.3.1 Analysis of ChIP-seq data for identification of PPREs

The raw data are available for ChIP-seq experiment regarding PPREs identification from http://www.ncbi.nlm.nih.gov/traces/sra/ with study accession number SRP000630. Thus, we decided to carry out the data analysis using the tools for both sequence alignment and peak detection described earlier.

In our application of bowtie, according to the online manual (http://bowtie-bio.sourceforge.net/manual.shtml), we used the following options in parameter setting:

--best, which guarantees the best reported singleton alignment. “Best” is defined in terms of the number of mismatches or the quality values at the mismatched positions. This mode also removes all strand bias.

-n, which stands for the maximum number of mismatches for alignment. To be consistent with original paper, we allowed one mismatch in our analysis.

-m, which keeps Bowtie from reporting any alignments for reads having more than a certain number of reportable alignments. We set m=1 for this parameter, which was able to report a loose “unique” alignments.

As the result, 35.70% of original short reads (5,420,479) were reported to have unique alignments mapping to mm9 reference genome.

In application of MACS, we kept all the default parameter settings. By using the recommended cutoffs – p-value < 1×10^{-5} and fold enrichment > 32, we identified 4656 PPARγ binding sites (labeled as ChIP-seq-1). According to a recent study comparing different methods for peak detection [86], the 1000~2000 most significant binding sites of a transcription factor are robustly detected by a variety of computational approaches. So we also used a strict criterion that p-value < 1×10^{-8} and fold enrichment > 38 to determine 1647 PPREs (labeled as ChIP-seq-2).

5.3.2 Result of PPRE enrichment analysis

All the PPARγ related microarrays were carried out using RNAs extracted from blood vessels, but the genome-wide sets of PPREs were identified from ChIP-seq/ChIP-chip using DNAs from 3T3-L1 adipocytes. Therefore, we limited our PPREs enrichment analysis to those genes also expressed in 3T3-L1 adipocytes. The expression status (present or absent) of a gene in 3T3-L1 adipocytes was determined using datasets GSE14004 and GSE8682, which are publicly available from GEO. We used ten lists of DEGs we had already identified (see details in Chapter 4). For example, list Rosi_up stands for the list of genes up-regulated in Rosi experiment. As the sets of PPREs were based on either mm8 or mm9 reference genome, we obtained both versions of the genomic locations of genes' TSS (transcription start site) from UCSC Genome Browser. For each list of DEGs, we calculated the percentage of genes having at least one PPRE within a certain distance. We used three scales of distances, 10k, 50k and 100k from TSS, including both up-stream and down-stream directions. To assess the statistical significance of the PPRE enrichment, we calculated the empirical p-values from random permutations.

The results of PPRE enrichment analysis using ChIP-seq-1 dataset is shown in Table 20, with the percentage of genes having at least one PPRE within a certain distance. The corresponding empirical p-values were obtained from 10,000 random permutations (see Table 21). Up-regulated genes from Rosi experiment are statistically enriched for PPREs within the distance of 50k.

Table 20. Percentage of genes with PPRE in analysis using ChIP-seq-1 dataset.

Distance	Rosi_up	Rosi_down	Gppar_up	Gppar_down	Sppar_up
10k	6.98%	3.66%	6.45%	4.55%	7.55%
50k	43.02%	21.95%	35.48%	25.00%	30.19%
100k	54.65%	40.24%	50.00%	50.00%	41.51%
Distance	Sppar_down	Msppar_up	Msppar_down	Ksppar_up	Ksppar_down
10k	6.67%	6.06%	2.44%	3.57%	8.33%
50k	37.78%	30.3%	36.59%	35.71%	20.83%
100k	51.11%	45.45%	46.34%	35.71%	41.67%

Table 21. P-values of PPRE enrichment analysis using ChIP-seq-1 dataset.

Distance	Rosi_up	Rosi_down	Gppar_up	Gppar_down	Sppar_up
10k	0.4818	0.8961	0.4818	0.7878	0.3305
50k	0.0005	0.9401	0.0612	0.7360	0.3289
100k	0.0330	0.8248	0.1466	0.1466	0.7713
Distance	Sppar_down	Msppar_up	Msppar_down	Ksppar_up	Ksppar_down
10k	0.4818	0.4818	0.9653	0.8961	0.2076
50k	0.0211	0.3289	0.0367	0.0612	0.9649
100k	0.1048	0.4713	0.3947	0.9757	0.7713

The results of PPRE enrichment analysis using ChIP-seq-2 dataset is shown is Table 22 and Table 23. These p-values were obtained from 10,000 permutations as well. Up-regulated genes from Rosi experiment are also statistically enriched with PPREs within the distance of 50k.

Table 22. Percentage of genes with PPRE in analysis using ChIP-seq-2 dataset

Distance	Rosi_up	Rosi_down	Gppar_up	Gppar_down	Sppar_up
10k	2.33%	1.22%	3.23%	2.27%	0.00%
50k	22.09%	7.32%	12.90%	11.36%	9.43%
100k	29.07%	13.41%	20.97%	18.18%	16.98%
Distance	Sppar_down	Msppar_up	Msppar_down	Ksppar_up	Ksppar_down
10k	0.00%	3.03%	0.00%	0.00%	6.25%
50k	13.33%	12.12%	12.20%	17.86%	12.50%
100k	17.78%	15.15%	14.63%	25.00%	20.83%

Table 23. P-values of PPRE enrichment analysis using ChIP-seq-2 dataset.

Distance	Rosi_up	Rosi_down	Gppar_up	Gppar_down	Sppar_up
10k	0.3378	0.6075	0.1508	0.3378	0.8775
50k	0.0001	0.8319	0.2267	0.3355	0.5954
100k	0.0044	0.9193	0.3157	0.5114	0.7124
Distance	Sppar_down	Msppar_up	Msppar_down	Ksppar_up	Ksppar_down
10k	0.8775	0.1508	0.8775	0.8775	0.0043
50k	0.1480	0.2267	0.2267	0.0136	0.2267
100k	0.6183	0.7977	0.8655	0.0464	0.3157

The results of PPRE enrichment analysis using ChIP-chip dataset is shown is Table 24 and Table 25. These p-values were obtained from 100,000 permutations. Up-regulated genes from Rosi experiment are statistically enriched with PPREs within the all three different distances. The number 0.00000 refers to p-value < 0.00001.

Table 24. Percentage of genes with PPRE in analysis using ChIP-chip dataset.

Distance	Rosi_up	Rosi_down	Gppar_up	Gppar_down	Sppar_up
10k	45.24%	13.41%	8.93%	20.45%	17.65%
50k	70.24%	43.90%	39.29%	45.45%	49.02%
100k	83.33%	57.32%	50.00%	63.64%	62.75%
Distance	Sppar_down	Msppar_up	Msppar_down	Ksppar_up	Ksppar_down
10k	21.43%	12.12%	21.62%	7.14%	24.44%
50k	52.38%	48.48%	51.35%	39.29%	51.11%
100k	59.52%	63.64%	62.17%	42.86%	66.67%

Table 25. P-values of PPRE enrichment analysis using ChIP-seq-2 dataset.

Distance	Rosi_up	Rosi_down	Gppar_up	Gppar_down	Sppar_up
10k	0.00000	0.37909	0.89601	0.01216	0.07319
50k	0.00000	0.01429	0.08335	0.00459	0.00032
100k	0.00000	0.02158	0.26940	0.00054	0.00126
Distance	Sppar_down	Msppar_up	Msppar_down	Ksppar_up	Ksppar_down
10k	0.00589	0.49615	0.00589	0.94567	0.00047
50k	0.00004	0.00064	0.00008	0.08335	0.00008
100k	0.00768	0.00054	0.00126	0.83528	0.00005

In fact, we previously performed 10,000 permutations as we did to two ChIP-seq datasets. However, not only Rosi_up genes but also some other gene-sets, including Sppar_down, Msppar_down and Ksppar_down genes, turned out to have p-values < 0.0001 as well. According to the original data (refers to percentage of genes with PPRE), we found that 70.24% of Rosi_up genes are associated with PPREs within 50k bp distance, compared to around 50% of Sppar_down, Msppar_down and Ksppar_down genes having PPREs. Obviously, Rosi_up genes are much more enriched with PPREs than the other three sets of genes. Therefore, we carried out 100,000 permutations, which helped us to obtain distinguishable p-values. Therefore, up-regulated genes from Rosi experiment are robustly and significantly enriched with PPREs from both ChIP-seq and ChIP-chip datasets.

As we discussed in Chapter 4, up-regulated genes in Rosi experiment, as a gene-set, were found to be down-regulated in Gppar experiment. However, the Gppar_down genes did not show up with the pattern of PPRE enrichment. Thus, we divided these genes into two sets, one with the trend of down-regulation in Gppar experiment and the other with up-regulated direction. Then, PPRE enrichment analysis was carried out for all three sets. As shown in Table 26, only up-regulated genes in Rosi experiment that also have a trend of down-regulation in the Gppar dataset are robustly and significantly enriched with PPREs, consistent with a previous analysis [80]. This scenario fits with the conventional model of PPARγ action, in which activation of its target genes demands ligand binding to PPREs, as we have described previously in Chapter 2.

Table 26. P-values of PPRE enrichment analysis for Rosi_up genes.

	ChIP-seq-1		ChIP-seq-2		ChIP-chip	
Distance	R_up_G_down	R_up_G_up	R_up_G_down	R_up_G_up	R_up_G_down	R_up_G_up
10k	0.0612	0.6450	0.0536	0.8775	0.00000	0.11960
50k	0.0001	0.3289	0.0000	0.4631	0.00000	0.00811
100k	0.0043	0.4713	0.0003	0.7977	0.00000	0.00230

Note: All three sets of PPREs were used, including ChIP-seq-1, ChIP-seq-2 and ChIP-chip. For p-value=0.0000 means p-value < 0.0001, which was calculated from 10,000 permutations and for p-value=0.00000 means p-value < 0.00001, which was calculated from 100,000 permutations.

REFERENCES

1. Quackenbush J: **Microarray data normalization and transformation.** Nat Genet 2002, **32 Suppl**:496-501.

2. Kooperberg C, Aragaki A, Strand AD, Olson JM: **Significance testing for small microarray experiments.** Stat Med 2005, **24**(15):2281-2298.

3. Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM: **Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer.** Funct Integr Genomics 2003, **3**(4):180-188.

4. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** Nucleic Acids Res 2002, **30**(1):207-210.

5. Zhang M, Zhang L, Zou J, Yao C, Xiao H, Liu Q, Wang J, Wang D, Wang C, Guo Z: **Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes.** Bioinformatics 2009, **25**(13):1662-1668.

6. Hong F, Breitling R: **A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.** Bioinformatics 2008, **24**(3):374-382.

7. Kapur K, Xing Y, Ouyang Z, Wong WH: **Exon arrays provide accurate assessments of gene expression.** Genome Biol 2007, **8**(5):R82.

8. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** Genome Biol 2005, **6**(2):R16.

9. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** Bioinformatics 2003, **19**(2):185-193.

10. Allocco DJ, Kohane IS, Butte AJ: **Quantifying the relationship between co-expression, co-regulation and gene function.** BMC Bioinformatics 2004, **5**:18.

11. Lai Y: **Genome-wide co-expression based prediction of differential expressions.** Bioinformatics 2008, **24**(5):666-673.

12. Sculley D: **Rank Aggregation for Similar Items**. In Proceedings of 2007 SIAM International Conference on Data Mining, 2007 April.

13. Halabi CM, Beyer AM, de Lange WJ, Keen HL, Baumbach GL, Faraci FM, Sigmund CD: **Interference with PPAR gamma function in smooth muscle causes vascular dysfunction and hypertension.** Cell Metab 2008, **7**(3):215-226.

14. Feige JN, Gelman L, Michalik L, Desvergne B, Wahli W: **From molecular action to physiological outputs: peroxisome proliferator-activated receptors are nuclear receptors at the crossroads of key cellular functions.** Prog Lipid Res 2006, **45**(2):120-159.

15. Takano H, Komuro I: **Roles of peroxisome proliferator-activated receptor gamma in cardiovascular disease.** J Diabetes Complications 2002, **16**(1):108-114.

16. Barroso I, Gurnell M, Crowley VE, Agostini M, Schwabe JW, Soos MA, Maslen GL, Williams TD, Lewis H, Schafer AJ, Chatterjee VK, O'Rahilly S: **Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension.** Nature 1999, **402**(6764):880-883.

17. Lefterova MI, Zhang Y, Steger DJ, Schupp M, Schug J, Cristancho A, Feng D, Zhuo D, Stoeckert CJ,Jr, Liu XS, Lazar MA: **PPARgamma and C/EBP factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale.** Genes Dev 2008, **22**(21):2941-2952.

18. Nielsen R, Pedersen TA, Hagenbeek D, Moulos P, Siersbaek R, Megens E, Denissov S, Borgesen M, Francoijs KJ, Mandrup S, Stunnenberg HG: **Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis.** Genes Dev 2008, **22**(21):2953-2967.

19. Landstrom AP, Parvatiyar MS, Pinto JR, Marquardt ML, Bos JM, Tester DJ, Ommen SR, Potter JD, Ackerman MJ: **Molecular and functional characterization of novel hypertrophic cardiomyopathy susceptibility mutations in TNNC1-encoded troponin C.** J Mol Cell Cardiol 2008, **45**(2):281-288.

20. Berthold J, Schenkova K, Rivero F: **Rho GTPases of the RhoBTB subfamily and tumorigenesis.** Acta Pharmacol Sin 2008, **29**(3):285-295.

21. Wang X, Adams LD, Pabon LM, Mahoney WM,Jr, Beaudry D, Gunaje J, Geary RL, Deblois D, Schwartz SM: **RGS5, RGS4, and RGS2 expression and aortic contractibility are dynamically co-regulated during aortic banding-induced hypertrophy.** J Mol Cell Cardiol 2008, **44**(3):539-550.

22. Issemann I, Green S: **Activation of a member of the steroid hormone receptor superfamily by peroxisome proliferators.** Nature 1990, **347**(6294):645-650.

23. Berger J, Moller DE: **The mechanisms of action of PPARs.** Annu Rev Med 2002, **53**:409-435.

24. Fajas L, Auboeuf D, Raspe E, Schoonjans K, Lefebvre AM, Saladin R, Najib J, Laville M, Fruchart JC, Deeb S, Vidal-Puig A, Flier J, Briggs MR, Staels B, Vidal H, Auwerx J: **The organization, promoter analysis, and expression of the human PPARgamma gene.** J Biol Chem 1997, **272**(30):18779-18789.

25. Palmer CN, Hsu MH, Griffin HJ, Johnson EF: **Novel sequence determinants in peroxisome proliferator signaling.** J Biol Chem 1995, **270**(27):16114-16121.

26. Tontonoz P, Spiegelman BM: **Fat and beyond: the diverse biology of PPARgamma.** Annu Rev Biochem 2008, **77**:289-312.

27. Mizukami J, Taniguchi T: **The antidiabetic agent thiazolidinedione stimulates the interaction between PPAR gamma and CBP.** Biochem Biophys Res Commun 1997, **240**(1):61-64.

28. Zhu Y, Qi C, Calandra C, Rao MS, Reddy JK: **Cloning and identification of mouse steroid receptor coactivator-1 (mSRC-1), as a coactivator of peroxisome proliferator-activated receptor gamma.** Gene Expr 1996, **6**(3):185-195.

29. Puigserver P, Adelmant G, Wu Z, Fan M, Xu J, O'Malley B, Spiegelman BM: **Activation of PPARgamma coactivator-1 through transcription factor docking.** Science 1999, **286**(5443):1368-1371.

30. Yu C, Markan K, Temple KA, Deplewski D, Brady MJ, Cohen RN: **The nuclear receptor corepressors NCoR and SMRT decrease peroxisome proliferator-activated receptor gamma transcriptional activity and repress 3T3-L1 adipogenesis.** J Biol Chem 2005, **280**(14):13600-13605.

31. Halabi CM: **Interference with Peroxisome Proliferator Activated Receptor Gamma Function in Smooth Muscle Causes Vascular Dysfunction and Hypertension.** 2008

32. Kliewer SA, Lenhard JM, Willson TM, Patel I, Morris DC, Lehmann JM: **A prostaglandin J2 metabolite binds peroxisome proliferator-activated receptor gamma and promotes adipocyte differentiation.** Cell 1995, **83**(5):813-819.

33. Lehmann JM, Moore LB, Smith-Oliver TA, Wilkison WO, Willson TM, Kliewer SA: **An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPAR gamma).** J Biol Chem 1995, **270**(22):12953-12956.

34. Tontonoz P, Hu E, Graves RA, Budavari AI, Spiegelman BM: **mPPAR gamma 2: tissue-specific regulator of an adipocyte enhancer.** Genes Dev 1994, **8**(10):1224-1234.

35. Tontonoz P, Hu E, Spiegelman BM: **Stimulation of adipogenesis in fibroblasts by PPAR gamma 2, a lipid-activated transcription factor.** Cell 1994, **79**(7):1147-1156.

36. Rosen ED, Sarraf P, Troy AE, Bradwin G, Moore K, Milstone DS, Spiegelman BM, Mortensen RM: **PPAR gamma is required for the differentiation of adipose tissue in vivo and in vitro.** Mol Cell 1999, **4**(4):611-617.

37. Nolan JJ, Ludvik B, Beerdsen P, Joyce M, Olefsky J: **Improvement in glucose tolerance and insulin resistance in obese subjects treated with troglitazone.** N Engl J Med 1994, **331**(18):1188-1193.

38. Henke BR, Blanchard SG, Brackeen MF, Brown KK, Cobb JE, Collins JL, Harrington WW,Jr, Hashim MA, Hull-Ryde EA, Kaldor I, Kliewer SA, Lake DH, Leesnitzer LM, Lehmann JM, Lenhard JM, Orband-Miller LA, Miller JF, Mook RA,Jr, Noble SA, Oliver W,Jr, Parks DJ, Plunket KD, Szewczyk JR, Willson TM: **N-(2-Benzoylphenyl)-L-tyrosine PPARgamma agonists. 1. Discovery of a novel series of potent antihyperglycemic and antihyperlipidemic agents.** J Med Chem 1998, **41**(25):5020-5036.

39. Tontonoz P, Nagy L, Alvarez JG, Thomazy VA, Evans RM: **PPARgamma promotes monocyte/macrophage differentiation and uptake of oxidized LDL.** Cell 1998, **93**(2):241-252.

40. Castrillo A, Tontonoz P: **Nuclear receptors in macrophage biology: at the crossroads of lipid metabolism and inflammation.** Annu Rev Cell Dev Biol 2004, **20**:455-480.

41. Ricote M, Li AC, Willson TM, Kelly CJ, Glass CK: **The peroxisome proliferator-activated receptor-gamma is a negative regulator of macrophage activation.** Nature 1998, **391**(6662):79-82.

42. Ogawa S, Lozach J, Benner C, Pascual G, Tangirala RK, Westin S, Hoffmann A, Subramaniam S, David M, Rosenfeld MG, Glass CK: **Molecular determinants of crosstalk between nuclear receptors and toll-like receptors.** Cell 2005, **122**(5):707-721.

43. Li AC, Brown KK, Silvestre MJ, Willson TM, Palinski W, Glass CK: **Peroxisome proliferator-activated receptor gamma ligands inhibit development of atherosclerosis in LDL receptor-deficient mice.** J Clin Invest 2000, **106**(4):523-531.

44. Fukunaga Y, Itoh H, Doi K, Tanaka T, Yamashita J, Chun TH, Inoue M, Masatsugu K, Sawada N, Saito T, Hosoda K, Kook H, Ueda M, Nakao K: **Thiazolidinediones, peroxisome proliferator-activated receptor gamma agonists, regulate endothelial cell growth and secretion of vasoactive peptides.** Atherosclerosis 2001, **158**(1):113-119.

45. Jackson SM, Parhami F, Xi XP, Berliner JA, Hsueh WA, Law RE, Demer LL: **Peroxisome proliferator-activated receptor activators target human endothelial cells to inhibit leukocyte-endothelial cell interaction.** Arterioscler Thromb Vasc Biol 1999, **19**(9):2094-2104.

46. Goetze S, Xi XP, Kawano H, Gotlibowski T, Fleck E, Hsueh WA, Law RE: **PPAR gamma-ligands inhibit migration mediated by multiple chemoattractants in vascular smooth muscle cells.** J Cardiovasc Pharmacol 1999, **33**(5):798-806.

47. Sigmund CD: **Endothelial and vascular muscle PPARgamma in arterial pressure regulation: lessons from genetic interference and deficiency.** Hypertension 2010, **55**(2):437-444.

48. Diep QN, El Mabrouk M, Cohn JS, Endemann D, Amiri F, Virdis A, Neves MF, Schiffrin EL: **Structure, endothelial function, cell growth, and inflammation in blood vessels of angiotensin II-infused rats: role of peroxisome proliferator-activated receptor-gamma.** Circulation 2002, **105**(19):2296-2302.

49. Fullert S, Schneider F, Haak E, Rau H, Badenhoop K, Lubben G, Usadel KH, Konrad T: **Effects of pioglitazone in nondiabetic patients with arterial hypertension: a double-blind, placebo-controlled study.** J Clin Endocrinol Metab 2002, **87**(12):5503-5506.

50. Dormandy JA, Charbonnel B, Eckland DJ, Erdmann E, Massi-Benedetti M, Moules IK, Skene AM, Tan MH, Lefebvre PJ, Murray GD, Standl E, Wilcox RG, Wilhelmsen L, Betteridge J, Birkeland K, Golay A, Heine RJ, Koranyi L, Laakso M, Mokan M, Norkus A, Pirags V, Podar T, Scheen A, Scherbaum W, Schernthaner G, Schmitz O, Skrha J, Smith U, Taton J, PROactive investigators: **Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial.** Lancet 2005, **366**(9493):1279-1289.

51. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R, Prospective Studies Collaboration: **Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies.** Lancet 2002, **360**(9349):1903-1913.

52. Collins R, Peto R, MacMahon S, Hebert P, Fiebach NH, Eberlein KA, Godwin J, Qizilbash N, Taylor JO, Hennekens CH: **Blood pressure, stroke, and coronary heart disease. Part 2, Short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context.** Lancet 1990, **335**(8693):827-838.

53. Li G, Leff T: **Altered promoter recycling rates contribute to dominant-negative activity of human peroxisome proliferator-activated receptor-gamma mutations associated with diabetes.** Mol Endocrinol 2007, **21**(4):857-864.

54. P.J. Russell: *iGenetics, A Mendelian Approach:* Pearson Education Inc.; 2006.

55. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.** Nat Rev Genet 2009, **10**(10):669-680.

56. Solomon MJ, Larsen PL, Varshavsky A: **Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene.** Cell 1988, **53**(6):937-947.

57. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins.** Science 2000, **290**(5500):2306-2309.

58. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** Science 2007, **316**(5830):1497-1502.

59. DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R: **Combining results of microarray experiments: a rank aggregation approach.** Stat Appl Genet Mol Biol 2006, **5**:Article15.

60. Pihur V, Datta S, Datta S: **RankAggreg, an R package for weighted rank aggregation.** BMC Bioinformatics 2009, **10**:62.

61. Pihur V, Datta S, Datta S: **Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach.** Bioinformatics 2007, **23**(13):1607-1615.

62. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** Nucleic Acids Res 2003, **31**(4):e15.

63. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** Biostatistics 2003, **4**(2):249-264.

64. Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, Terence P. Speed: **Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments.** 2002, **12**(1):111-139.

65. Pelz CR, Kulesz-Martin M, Bagby G, Sears RC: **Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data.** BMC Bioinformatics 2008, **9**:520.

66. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** Nat Rev Genet 2006, **7**(1):55-65.

67. Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, Sweet MJ: **Expression analysis of G Protein-Coupled Receptors in mouse macrophages.** Immunome Res 2008, **4**(1):5.

68. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** Proc Natl Acad Sci U S A 2003, **100**(16):9440-9445.

69. Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** Genome Biol 2007, **8**(9):R183.

70. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** Stat Appl Genet Mol Biol 2004, **3**:Article3.

71. Tsai YS, Kim HJ, Takahashi N, Kim HS, Hagaman JR, Kim JK, Maeda N: **Hypertension and abnormal fat distribution but not insulin resistance in mice with P465L PPARgamma.** J Clin Invest 2004, **114**(2):240-249.

72. Beyer AM, Baumbach GL, Halabi CM, Modrick ML, Lynch CM, Gerhold TD, Ghoneim SM, de Lange WJ, Keen HL, Tsai YS, Maeda N, Sigmund CD, Faraci FM: **Interference with PPARgamma signaling causes cerebral vascular dysfunction, hypertrophy, and remodeling.** Hypertension 2008, **51**(4):867-871.

73. Chang L, Villacorta L, Zhang J, Garcia-Barrio MT, Yang K, Hamblin M, Whitesall SE, D'Alecy LG, Chen YE: **Vascular smooth muscle cell-selective peroxisome proliferator-activated receptor-gamma deletion leads to hypotension.** Circulation 2009, **119**(16):2161-2169.

74. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP: **Analysis of high density expression microarrays with signed-rank call algorithms.** Bioinformatics 2002, **18**(12):1593-1599.

75. Pepper SD, Saunders EK, Edwards LE, Wilson CL, Miller CJ: **The utility of MAS5 expression summary and detection call algorithms.** BMC Bioinformatics 2007, **8**:273.

76. Affymetrix Inc.: **Statistical Algorithms Description Document.** 2002.

77. Affymetrix Inc.: **Exon Array Background Correction.** 2005.

78. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** Proc Natl Acad Sci U S A 2005, **102**(43):15545-15550.

79. Gellert P, Uchida S, Braun T: **Exon Array Analyzer: a web interface for Affymetrix exon array analysis.** Bioinformatics 2009, **25**(24):3323-3324.

80. Keen HL, Halabi CM, Beyer AM, de Lange WJ, Liu X, Maeda N, Faraci FM, Casavant TL, Sigmund CD: **Bioinformatic analysis of gene sets regulated by ligand-activated and dominant-negative peroxisome proliferator-activated receptor gamma in mouse aorta.** Arterioscler Thromb Vasc Biol 2010, **30**(3):518-525.

81. Chen Y, Yang Z, Meng M, Zhao Y, Dong N, Yan H, Liu L, Ding M, Peng HB, Shao F: **Cullin mediates degradation of RhoA through evolutionarily conserved BTB adaptors to control actin cytoskeleton structure and cell movement.** Mol Cell 2009, **35**(6):841-855.

82. Joseph L. Izzo Jr., Domenic A. Sica, Henry R. Black: *Hypertension Primer:* 4th ed. Lippincott Williams and Wilkins (LLW); 2008.

83. Puetz S, Lubomirov LT, Pfitzer G: **Regulation of smooth muscle contraction by small GTPases.** Physiology (Bethesda) 2009, **24**:342-356.

84. Grayson TH, Ohms SJ, Brackenbury TD, Meaney KR, Peng K, Pittelkow YE, Wilson SR, Sandow SL, Hill CE: **Vascular microarray profiling in two models of hypertension identifies caveolin-1, Rgs2 and Rgs5 as antihypertensive targets.** BMC Genomics 2007, **8**:404.

85. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** Genome Biol 2009, **10**(3):R25.

86. Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo LL: **A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments.** BMC Genomics 2009, **10**:618.

87. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** Genome Biol 2008, **9**(9):R137.