Computers in Biology and Medicine 113 (2019) 103389

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Gene set analysis and reduction for a continuous phenotype: Identifying markers of birth weight variation based on embryonic stem cells and immunologic signatures

Shabnam Vatanpour[a], Saumyadipta Pyne[b], Ana Paula Leite[c], Irina Dinu[a,*]

[a] School of Public Health, University of Alberta, AB, Canada
[b] Public Health Dynamics Laboratory, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, PA, USA
[c] University College London Cancer Institute, London, UK

ARTICLE INFO

Keywords:
DNA microarray
Gene expression
Gene set reduction
Continuous phenotype
Pathways

ABSTRACT

Background: Gene set analysis is a popular approach to examine the association between a predefined gene set and a phenotype. Few methods have been developed for a continuous phenotype. However, often not all the genes within a significant gene set contribute to its significance. There is no gene set reduction method developed for continuous phenotype. We developed a computationally efficient analytical tool, called linear combination test for gene set reduction (LCT-GSR) to identify core subsets of gene sets associated with a continuous phenotype. Identifying the core subset enhances our understanding of the biological mechanism and reduces costs of disease risk assessment, diagnosis and treatment.

Results: We evaluated the performance of our analytical tool by applying it to two real microarray studies. In the first application, we analyzed pathway expression measurements in newborns' blood to discover core genes contributing to the variation in birth weight. On average, we were able to reduce the number of genes in the 33 significant gene sets of embryonic stem cell signatures by 84.3% resulting in 229 unique genes. Using immunologic signatures, on average we reduced the number of genes in the 210 significant gene sets by 89% leading to 1603 unique genes. There were 180 unique core genes overlapping across the two databases. In the second application, we analyzed pathway expression measurements in a cohort of lethal prostate cancer patients from Swedish Watchful Waiting cohort to identify main genes associated with tumor volume. On average, we were able to reduce the number of genes in the 17 gene sets by 90% resulting in 47 unique genes.

Conclusions: We conclude that LCT-GSR is a statistically sound analytical tool that can be used to extract core genes associated with a continuous phenotype. It can be applied to a wide range of studies in which dichotomizing the continuous phenotype is neither easy nor meaningful. Reduction to the most predictive genes is crucial in advancing our understanding of issues such as disease prevention, faster and more efficient diagnosis, intervention strategies and personalized medicine.

1. Background

With the advent of DNA microarray technology, scientists are able to study and analyze thousand of genes at the same time leading to early and more accurate disease diagnosis as well as personalized treatment. There are two general approaches to study associations of gene expression measurements with phenotypes in microarray data analysis: Individual Gene Analysis (IGA) and Gene Set Analysis (GSA). A comprehensive review of these methods is given by Goeman and Buhlmann [1]. IGA examines each gene individually to find differentially expressed genes associated with phenotypes or characteristics. Once a list of significant genes is assembled, there is a need to identify biological functions or pathways that are over-represented in a given list. An alternative approach is to identify sets of functionally related genes in advance and to assess whether these gene sets show differential expression. The interest in expression data analysis has changed from single gene to gene set level in recent years because many phenotypes are believed to be associated with modest regulation in a set of related genes rather than a strong increase in a single gene [2]. However, both approaches can be effective and sometimes their combination is more powerful.

Molecular biologists have compiled lists of genes grouped by their

* Corresponding author.
E-mail addresses: spyne@pitt.edu (S. Pyne), apoliveiraleite@gmail.com (A.P. Leite), idinu@ualberta.ca (I. Dinu).

https://doi.org/10.1016/j.compbiomed.2019.103389
Received 2 April 2019; Received in revised form 14 August 2019; Accepted 15 August 2019
Available online 17 August 2019

common biological functions, i.e. biological pathways. There are various pathway databases that are freely available for microarray data analysis, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [3], Gene Expression Omnibus [4], Biocarta [5], and Molecular Signature Data Base [6]. A variety of Gene Set Analysis (GSA) methods have been developed with the aim to identify a-priori defined gene sets associated with phenotypes, in DNA microarray studies. These methods incorporate previous biological knowledge of presumably related genes within a gene set, and hence are more powerful in finding associations with phenotypes.

GSA methods are different in terms of the methodological assumptions related to definition of a sample and formulation of the null hypothesis. Extensive methodological discussions and reviews are given by Goeman and Buhlmann [1], Nam and Kim [7], and Maciejewski [8]. We briefly summarize important aspects of GSA methods. The GSA methods are broadly classified as 'self-contained' or 'competitive'. Competitive methods compare the associations for genes within the gene set with associations for genes in the gene set complement, to determine whether genes in a particular gene set are associated more with a phenotype, as compared to genes outside the gene set. Examples of competitive gene set methods for analysis of gene expression studies are SAFE [9], Random set methods [10], and GSA [11]. In contrast, self-contained methods assess the association between the phenotype and expression of the gene set of interest, ignoring other genes that are not in the gene set. Examples include Global test [12], ANCOVA [13], SAM-GS [14], and LCT [15]. A hybrid between competitive and self-contained methods is Gene Set Enrichment Analysis (GSEA) [2].

The key methodological distinction between the two approaches is inherent to gene-sampling versus subject-sampling concept. The term 'sampling' refers to the permutation test used in GSA methods to estimate the null distribution. Competitive methods use genes as the sampling units, whereas self-contained methods use subjects as sampling units. Under the self-contained null hypothesis of no association between the gene sets and the phenotype, labels are interchangeable and the null distribution is estimated based on permuting the labels of subjects. Under the competitive null hypothesis of no differential expression of genes in the gene set of interest, compared with expression of genes not in the set, we assume that genes are independent and the null distribution is estimated based on permuting the genes [1].

Geoman and Buhlmann [1] strongly discourage using competitive methods because of the untenable statistical independence assumption across genes. Delongchamp et al. [16] commented on how ignoring the correlations within the gene sets can overstate significance and proposed meta-analysis methods for combining p-values with a modification to adjust for correlation. Chen et al. [17] argue their preference for self-contained hypothesis over competitive one, because the p-values computed under former are consistent with the principle of statistical significance testing, while the p-values computed under latter do not take into account correlations among genes. Our focus here is on self-contained methods, which preserve correlations within gene sets.

Most of GSA methods have been developed for binary or categorical phenotypes. The urge of improving methods for continuous phenotype is increasing, on the ground that, quite often, the outcome of interest is measured as a continuous variable, for example, tumor volume, birth weight, metabolites or proteins. In such cases, it is neither easy nor meaningful to dichotomize or categorize continuous phenotypes. Some specific ranges may fail to express the underlying biological function for each subject. Moreover, these ranges are arbitrarily defined by specialists, and different specialists might use different ranges according to the patient's health condition. It would be beneficial to directly analyze continuous phenotypes in DNA microarray studies. Few methods have been proposed for continuous phenotypes such as SAM-GS [14], Global test [18] and Linear Combination Test (LCT) [15].

SAM-GS is an extension of Significance Analysis of Microarrays (SAM) [19], a moderated t-statistic, calculated based on permutations of the group labels. The Global test is based on the generalized linear regression framework, in which the distribution of the phenotype is modelled as a function of the covariates. LCT considers the linear combination with the maximum correlation with the phenotype among all possible linear combinations. Among these methods, LCT efficiently incorporates the gene expression covariance matrix into the test statistic, considering the correlation among gene expressions. This characteristic is desired in GSA methods because it leads to a powerful and computationally efficient approach for evaluating the association of a gene set with a continuous phenotype.

A gene set can be significant only because a subset of genes within the set is actually differentially expressed, and the rest of the genes may not be contributing to its significance. In fact, a large set may be easily identified as significant, only because one gene is associated with the phenotype. It is very important to investigate significant gene sets to identify only those core members that are associated with the phenotype, as a core subset. Dinu et al. [20] developed a gene set reduction method, referred to as SAM-GS reduction (SAM-GSR), for extracting core subset for a binary phenotype. To the best of our knowledge, there are no methods addressing gene set reduction for a continuous phenotype yet. Our proposed algorithm was designed to fill this gap.

2. Methods

We propose here an extension of the LCT method for gene set reduction, called LCT-GSR. Genes within gene sets are expected to be correlated because they share similar biological functions, and same chromosomal locations. Among GSA methods for continuous phenotypes, the LCT method [15] efficiently incorporates the gene expression covariance matrix into the test statistic, resulting in improved performance over other existing GSA methods. It is a powerful and efficient microarray data analysis, for determining significant associations between gene sets and a continuous phenotype. Since the number of genes in the gene sets is much larger than the number of subjects, the covariance matrix is singular. To overcome this problem, a shrinkage covariance matrix estimator is used. Then, eigenvalue decomposition of the shrinkage covariance matrix is performed for the original data, reducing the high computational cost of integrating this estimator. Let $\mathbf{X} = (X_1, ..., X_p)$ denote the gene expression matrix corresponding to p genes in a set, where each vector $X_1, ..., X_p$ is measured on each of the n subjects, $\mathbf{Y}$ denote the phenotype or outcome measured on each of the n subjects, and the covariance matrix decomposition be represented as $\hat{\Omega}^* = UDU^T$; then the orthogonal basis vectors are $(V_1, ..., V_p) = (X_1, ..., X_p)UD^{-1/2}$. Therefore, the LCT statistic is defined by:

$$\rho^2(\boldsymbol{\gamma}^*) = c \sum_{j=1}^{p} Cov(\boldsymbol{Y}, \boldsymbol{V}_j)^2,$$

where $\gamma = D^{1/2}U^T\beta$ and β is the vector of regression coefficients. A permutation test is used to assess the statistical significance against the null hypothesis.

For each significant gene set, we repeat the following steps to extract the core genes. We apply Significance Analysis of Microarrays (SAM) to measure the strength of association between a single gene expression and a phenotype. SAM is a popular analytical tool for DNA microarray data analysis at individual gene level. It assigns a score to each gene, based on the change in gene expression, relative to the standard deviation of repeated measurements. SAM avoids parametric assumptions about the distribution of individual genes, by using a non-parametric statistic d_i:

$$d_i = \frac{r_i}{s_i + s_0}, \quad i = 1,2, ..., p,$$

where r_i is the linear regression coefficient of expression measurements for gene i on the phenotype, s_i is the pooled standard error of r_i, and $s_0$is the exchangeability factor, or a small positive constant, adjusting for

the variability in the microarray measurements.

Given the significant gene set S with s genes, we use the following steps sequentially. These steps are motivated by the fact that d_i is the contribution of the gene i, to the association of the gene set S with the phenotype, in the LCT test statistic.

1. Compute SAM statistic d_i for each gene in the gene set S.
2. For $k = 1,2, \ldots, s-1$, select the first k genes with largest statistic $|d_i|$ to form a reduced set R_k. Let $\bar{R}_k$ be the complement gene set of R_k in S, and c_k be the corresponding LCT p-value of the complement gene set.
3. Select the reduced set when c_k is larger than a pre-specified threshold c, chosen by the analyst.

We used the SAM-R package available in R to compute the SAM statistic values. We can get both FDR values and p-values from the SAM output. However, the FDR values and p-values can be similar for most of the genes, and the ranking process of genes based on their significance would be a problem. We prefer to use SAM statistic values d_i, which are the scores assigned to each gene based on the changes in gene expression, relative to the standard error.

We order the genes within the gene set, according to the absolute value of their SAM values, $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(p)}$, to gradually discover the core genes with the largest $|d_i|$, apply the LCT analysis to the complement gene set $\bar{R}_k$, and calculate its p-value c_k. If $c_k < c$, we still have significant members within the complement gene set, that are associated with the phenotype, therefore making the whole set statistically significant. If $c_k > c$, there are no significant genes remained contributing to the significance of the complement gene set and we stop the procedure. When we reach the threshold, the genes within R_k represent the core subset.

Similarly to Dinu et al. [20], we used a cutoff c= 0.1. We used a threshold slightly more conservative than 0.05, to ensure we included genes that individually may not be strongly associated with the phenotype, but collectively have a biological impact on the phenotype of interest.

An important issue in GSA is adjusting for multiple testing of thousands of gene sets. Each statistical test reports the probability of observing a test score by chance, assuming no association between gene set expressions and the phenotype of interest. Among 10,000 independent tests, even if we set the threshold for p-values as low as 0.01, we will identify 100 of those as “significant” genes, just by chance. A popular approach is to control the false discovery rate (FDR) which measures the proportions of false positives among all gene sets called significant [21].

We applied our analytical tool LCT-GSR to two real microarray studies; a) birth weight and b) lethal prostate cancer, to evaluate its performance. Gene set p-values and multiple comparison adjustments using FDR values were performed.

3. Results

3.1. Birth weight study

3.1.1. Background

Individuals born small for gestational age (SGA) are at greater risk of multiple chronic illnesses, later in their life [22]. The link between low birth weight and adult illness might be explained by uteroplacental insufficiency that alters organ function and hormonal milieu to make the individual more susceptible to disease [23]. In addition, genetic or epigenetic factors may exist that both reduce fetal growth and increase predisposition to disease later in life [24]. The change in DNA methylation is known as the cause of some newborn illnesses and growth disorders. While DNA methylation is important in developmental processes, and its variation in blood lymphocytes has been associated with adult body mass index (BMI) [25], analysis of DNA methylation patterns with respect to birth weight have produced mixed results.

DNA Methylation ultimately affects mRNA production and resultant protein production, both of which are complex processes. Therefore, variation in gene expression levels is one step closer to a direct biological effect than DNA methylation, and might exhibit a stronger association with birth weight variation [26]. Many more significant associations between birth weight and gene expression have been published over the last decade relative to DNA methylation, suggesting the need for further investigation at gene expression level. In a recent study of 201 newborns, ranged in birth weight from 2.1 to 5 kg, Adkins et al. [26] did not identify strong genome-wide association of birth weight with gene expression measurements. The analysis in this study was focused on identifying individual genes that are associated with birth weight among a cohort of clinically normal newborns. We believe that correlation among genes, especially those within biological pathways, might impact the association with birth weight. An analysis at the individual gene level does not take into account correlations across sets. We examined the associations in the same study using the LCT-GSR analytical tool to identify biomarkers that may contribute to variation in birth weight, and thereby predisposition to diseases.

3.1.1.1. Data description. The birth weight data set is part of a larger longitudinal cohort study of human development from pregnancy to age 3, the Conditions Affecting Neurocognitive Development and Learning in Early Childhood (CANDLE) (http://candlestudy.org/). CANDLE was performed in Shelby County, Tennessee. Written informed consents were obtained from all mothers, and this study was approved by the institutional review boards of all the participating hospitals [26]. Data on maternal age, gestational age, race, and baby's gender are also available. We obtained approval from the University of Tennessee Health Science Center for accessing data on continuous phenotype birth weight measured on newborn blood.

The selection criteria for the cohort were: maternal age 18–40 years, singleton pregnancy, complete data on birth weight and maternal prepregnancy weight, and absence of several complications, specifically sexually transmitted diseases, diabetes, oligohydramnios, preeclampsia, placental abruption, tocolytics, and cervical cerclage. We selected gestational ages of 35–42 weeks and mothers whose self-declared race was only Caucasian, or only African-American. After applying these additional criteria, the final sample size was 114. This data set consists of 24,924 gene expression measurements from blood sample for 114 newborns, 67 African-American and 47 Caucasion, with mean birth weight of 3340 (SD: 490) grams. The mother's mean age is 27 years old and the mean gestational age is 39 weeks.

Rates of low birth weight vary among women of different origins. It has been long observed that the rate of low birth weight among African-American mothers is twice that of Caucasian women [27]. On the other hand, birth weight has consistently been shown to be higher in males than in females [28]. Mean birth weight of 3190.1 g for African-American mothers is lower than mean birth weight of 3553.7 for Caucasian mothers and the difference is statistically significant ($t = -4.2$, p-value = 0.0001). There is no significant difference between birth weight of male and female newborns ($t = 0.09$, p-value = 0.927). We examined whether the effect of race on birth weight is modified by gender, and the interaction was not significant ($t = 1.01$, p-value = 0.314). Since gender and race are important characteristics influencing the birth weight, we adjusted for both variables in our analysis.

3.1.1.2. Pathway databases. Since low birth weight was previously associated with the risk of developing immune diseases, we chose immunologic signatures catalog, C7. We downloaded the most recent list of gene sets in the Molecular Signature Database C7 catalog (accessed on May 2015) from Broad Institute (http://www.broadinstitute.org/gsea/msigdb). The C7 catalog represents immunologic signatures collected from immunologic studies.

Table 1
Gene sets in stem cell signatures associated with birth weight phenotype based on the LCT analysis at a p-value cutoff of 0.01 (FDR < 0.003).

Gene set name	Gene set size	LCT p-value
IPA_affects differentiation of embryonic stem cells	41	0
StemCell_Kasper06_30genes_16880536-Table 1	30	0.001
DMAP_MEGA_UP	46	0.001
DMAP_MONO1_DN	47	0.001
DMAP_PRE_BCELL2_UP	44	0.001
DMAP_PRE_BCELL3_DN	44	0.001
StemCell_Lim08_50genes_18510698-Table 1	47	0.002
Ben-Porath_MYC_TARGETS_WITH_EBOX	226	0.002
DB_ESR1-15608294	88	0.002
StemCell_Kocer08_87genes_18667080-Table S6	71	0.003
StemCell_Shim04_25genes_15246160-table6	22	0.003
StemCell_Fruehauf06_110genes_16863911-Table 1	97	0.003
DMAP_ERY_UP	45	0.003
DMAP_GM_EARLY_DN	42	0.003
DMAP_PRE_BCELL_UP	39	0.003
DMAP_BCELL_DN	44	0.003
DMAP_TCELLA6_DN	45	0.003
StemCell_Tondreau08_52genes_18405367-Table2b	41	0.004
DMAP_BCELLA2_UP	49	0.005
DMAP_TCELLA6_UP	44	0.005
IPA_affects differentiation of stem cells	72	0.006
DMAP_ERY4_DN	47	0.007
IPA_decreases differentiation of stem cells	18	0.007
StemCell_Colombo09_111genes_19123479-Table S1	92	0.008
StemCell_Lim08_25genes_18510698-Table 2	25	0.008
DMAP_ERY_DN	46	0.008
DMAP_GM_EARLY_UP	40	0.008
DMAP_HSC1_DN	48	0.008
DMAP_HSC3_UP	48	0.008
DB_PPARG-19300518	194	0.008
StemCell_Bhattacharya05_2843genes_16207381-Table1Sa	312	0.01
DMAP_MONO2_DN	40	0.01
DMAP_TCELLA2_DN	47	0.01

We also used the list of stem cell signatures consisting of 457 gene sets, collected from manuscripts (Leite & Pyne, manuscript in preparation) and others from the Differentiation Map portal [29], Ingenuity Pathway Analysis tool (http://www.ingenuity.com/), and ChIP-X database [30].

We restricted the size of gene sets in both lists to be between 15 and 500. Within this range, there are 1910 gene sets for the C7 catalog and 251 gene sets for the stem cell signatures.

3.1.1.3. Gene set reduction results. We applied LCT-GSR to the gene expression data set from CANDLE Study, adjusting for race and gender. We performed a logarithmic transformation on the gene expression values to increase the normality of the distribution across individuals. The LCT analysis revealed 33 gene sets in the stem cell signatures (FDR < 0.003) and 210 gene sets in the C7 catalog (FDR < 0.004), which are associated with birth weight at a cut-off p-value of 0.01, presented in Table 1 and Supplementary Material Table S1, respectively.

The next step is to use the list of significant gene sets and perform gene set reduction. We demonstrate the gene set reduction method for the significant gene sets *StemCell_Kasper06_16880536* pathway, composed of 30 genes, as defined in the stem cell signatures. We ranked the absolute value of SAM statistic for these 30 genes in a decreasing order. First, we selected the gene with the largest absolute value, ULK1 with $|d_{(1)}| = 2.78$ to form the core subset, and the rest of the genes within the gene set to form the complement set. We applied the LCT method to the complement set, and compared the LCT p-value with a pre-specified cut-off value of 0.1. Since the p-value is smaller than 0.1, we selected the gene with the second largest absolute value of SAM statistic, i.e., EGR3 with $|d_{(2)}| = 2.51$. We sequentially added the gene to the core subset, and test the complement set until we reached the cut-off threshold. The p-value of the complement set is greater than 0.1, after taking out the third gene IDS with $|d_{(3)}| = 2.44$. Genes within the complement set, are collectively not associated with the phenotype, and represent the redundant set. Therefore, the core subset contains three genes ULK1, EGR3 and IDS. Fig. 1 shows each step of the linear combination gene set reduction procedure.

Table 2 shows the summary of the LCT-GSR for stem cell signatures, including the list of gene sets along with the gene set size, core set size, percent reduction, and the core pathway members. Core set size indicates the number of core genes obtained from each significant gene set, according to LCT-GSR algorithm. Percent reduction is computed as the number of genes eliminated (in the complement set) divided by the total number of genes in a set, multiplied by 100. The core pathways show the core genes collectively contributing to the association with birth weight, excluding the redundant genes from the significant gene sets.

There are 33 significant gene sets within stem cell signatures (p-value < 0.01) associated with variation in birth weight, after adjusting for race and gender. There are 228 genes identified to be significantly associated with variation in birth weight from these gene sets, after adjusting for the race and gender variables. On average, we were able to reduce the number of genes in the 33 significant gene sets of stem cell signatures by 84.3%, using the cut-off value of 0.1.

Supplementary Material Table S2 shows the summary of the LCT-GSR for C7 catalog. There are 210 significant gene sets within C7 catalog (p-value < 0.01) associated with variation in birth weight, after adjusting for race and gender. There are 1603 genes identified to be significantly associated with variation in birth weight from these gene sets, after adjusting for race and gender. On average, we were able to reduce the number of genes in the 210 significant gene sets of C7 catalog by 89%, using the cut-off value of 0.1.

There are a total of 229 unique genes identified in the reduced subsets. In the stem cell signature, the most frequent core genes are *Kruppel-Like Factor 6* (KLF6), *Diazepam Binding Inhibitor* (DBI), *Early Growth Response 3* (EGR3), and *Jun Proto-Oncogene* (JUN). They show up in the core subsets of four significant gene sets.

In the C7 catalog, there are a total of 1603 unique genes identified in the reduced subsets. The core genes *Lectin, Galactoside-Binding, Soluble, 3* (LGALS3) and *G0/G1 Switch 2* (G0S2) are the most frequent genes extracted from 17 significant gene sets. The core gene *Endothelial PAS Domain Protein 1* (EPAS1) appeared in 16 significant gene sets, *Iduronate 2-Sulfatase* (IDS) and *Chemokine (C-X-C Motif) Ligand 8* (CXCL8) appeared in 15 significant gene sets.

The results from both gene set databases show 180 common core genes extracted from the significant gene sets. There are genes among core subsets that are not associated with the birth weight, at individual gene level analysis; for example, *N(Alpha)-Acetyltransferase 35* (NAA35) and *GABA(A) Receptor-Associated Protein-Like 2* (GABARAPL2) with the SAM p-value 1.0 and FDR 59.6%, *Heparan Sulfate (Glucosamine) 3-O-Sulfotransferase 3A1* (HS3ST3A1) and *Par-3 Family Cell Polarity Regulator* (PARD3) with the SAM p-value 1.0 and FDR 54.6%, POU2AF1 with the SAM p-value 0.07 and FDR 20.5% in the C7 catalog. However, they contributed to the significant association with birth weight, jointly with other genes within the core subset. For example, gene GABARAPL2 shows up in 4 different core subsets. This underlines the importance of gene set analysis over individual gene analysis. Our results show an advantage over analysis at the individual level, particularly for a microarray dataset with subtle signal, where small but coordinated differences cannot be captured at the individual level.

Leptin (LEP) is identified to be associated with birth weight in both gene set databases (p-value = 0.003, FDR = 0.02). *Leptin* encodes a protein, which acts through the leptin receptor that is secreted by white adipocytes, and which plays a major role in the regulation of body weight. This protein is involved in the regulation of immune and inflammatory responses, angiogenesis and wound healing. Mutations in this gene and/or its regulatory regions cause severe obesity, and morbid

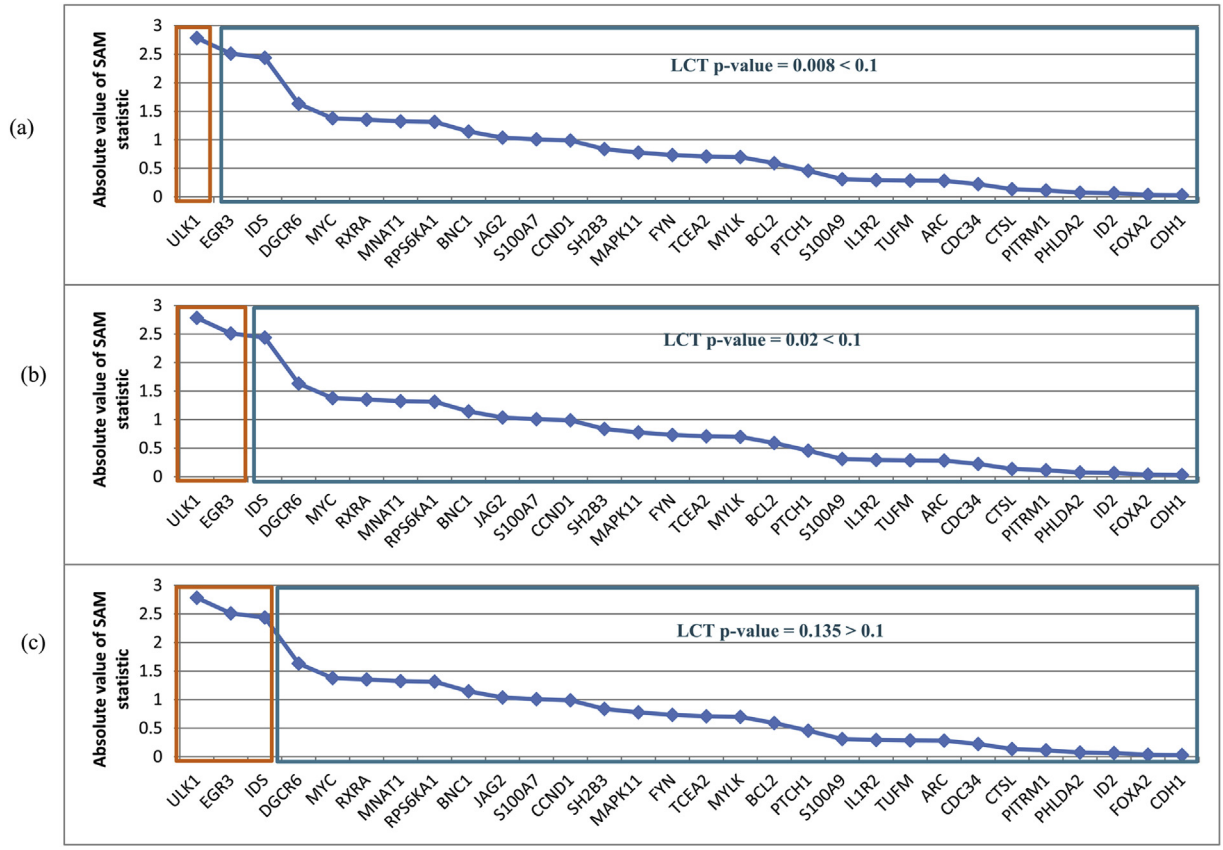

Fig. 1. An example of linear combination test gene set reduction (*StemCell_Kasper06_16880536*).

obesity with hypogonadism. This gene has also been linked to type 2 diabetes mellitus development (Genecards). *Early growth response 3* (EGR3) is another core gene (p-value = 0.001, FDR = 0.01) that plays a role in a wide variety of processes including muscle development, lymphocyte development, endothelial cell growth and migration, and neuronal development (Genecards).

3.2. Lethal prostate cancer

3.2.1. Background

Prostate cancer is the most common cancer in men. One in eight men will be diagnosed with the disease in their lifetime. It is estimated that in 2017, 21300 Canadian men will be diagnosed with prostate cancer and 4100 will die from the disease according to Prostate Cancer Canada. A major dilemma in prostate cancer management is how to treat patients with clinically localized disease. The death rate can be significantly reduced by improved testing and better treatment options.

3.2.2. Data description

The prostate cancer data set is part of the Swedish Watchful cohort study, nested in a cohort of men with localized prostate cancer (1977–1999) with up to 30 years of clinical follow up [31]. The study design was approved by the Ethical Review Boards in Örebro and Linköping. The cohort consists of 255 patients' expression measurements on 6014 genes and histopathologic features, such as Gleason score and tumor volume. The patients were categorized into lethal and indolent prostate cancer. We selected 145 patients with lethal cancer to create a homogenous cohort, based on the phenotype. We downloaded the expression data file, as well as histopathologic features from Gene Expression Omnibus, with accession ID GSE16560 [32].

3.2.3. Pathway databases

We downloaded the C2 catalog, an extensive collection of metabolic and signaling pathways, and gene sets from the Molecular Signature Database of Broad Institute of MIT and Harvard (http://www.broadinstitute.org/gsea/msigdb). The C2 catalog consists of 1892 gene sets, collected from online pathway databases, including The National Cancer Institute Pathway Interaction Database, Biocarta, KEGG, biomedical literature, and contributions from domain experts [5].

3.2.4. Gene set reduction results

We screened the C2 catalog for associations with tumor volume which has been found to be associated with development of prostate cancer. We restricted the size of the gene sets in the C2 catalog between 15 and 500, following Subramanian et al. [2]. There were 1263 gene sets within this range. There were 145 patients with lethal prostate cancer. LCT analysis revealed 15 gene sets among 1263 in the C2 catalog that are significantly associated with tumor volume at a cut-off p-value of 0.01 (FDR < 0.35).

The next step is to use the list of significant gene sets and perform gene set reduction. Given a significant gene set, we used the SAM statistic as a measure of association between each gene within the gene set and the tumor size. For reducing the significant gene set, we ranked the absolute values of the SAM statistic in a decreasing order for the genes within the gene set to help us gradually discover the core genes associated with the tumor size. We used the SAM-R package available in R to compute the SAM statistic values. We can get both FDR values and p-values from the SAM output. However, the FDR values and p-values can be similar for most of the genes, and the ranking process of genes based on their significance would be a problem. We prefer to use the SAM statistic values d, which are the scores assigned to each gene, based on the changes in gene expression relative to the standard error.

Table 2
Extracting core subsets of stem cell signatures associated with birth weight.

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
IPA_affects differentiation of embryonic stem cells	41	5	87.8	RNF2, ANGPT1, TLN1, NANOG, SOX2
StemCell_Kasper06_30genes_16880536-Table 1	30	3	90.0	ULK1, EGR3, IDS
DMAP_MEGA_UP	46	13	71.7	NUDT6, LOC55338, AGT, CALD1, SIX3, POLH, SSX1, TNP2, TFAP2A, PCP4, LAMB4, TBCE, LOC57399
DMAP_MONO1_DN	47	21	55.3	BRD8, GIMAP5, BTG1, ZCCHC6, MAP7D1, MICB, PREP, IQSEC1, ZFP36L2, ACOX1, IRF2, RNASEL, SARS, GEMIN6, HLA-A, DUSP10, KCNJ2, APOL2, TM2D3, SELPLG, TLR1
DMAP_PRE_BCELL2_UP	44	8	81.8	ZNF124, SERPINA5, MFSD6, 654056, PHF20L1, GNG11, ARHGEF17, CSPP1
DMAP_PRE_BCELL3_DN	44	8	81.8	ZNF124, SERPINA5, MFSD6, 654056, PHF20L1, GNG11, ARHGEF17, CSPP1
StemCell_Lim08_50genes_18510698-Table 1	47	9	80.9	GP5, PLEK, GABRE, LRP12, SLC44A1, CALD1, SCD, PDE5A, CXCL3
Ben-Porath_MYC_TARGETS_WITH_EBOX	226	24	89.4	APP, BAX, GSTP1, MNX1, EGR3, JUN, MST1, DBI, RHOG, CD79B, SNHG5, CD2, HDAC3, PRTN3, MUC1, HSPA8, HMBS, MPO, HIST1H4E, SERPINE1, TXN, NBN, PPID, BCL3
DB_ESR1-15608294	88	14	84.1	CRCP, SIRT3, SERPINB9, BRCA1, TRIP10, BRIP1, SERPINE1, ZNF600, ENSA, CASP8AP2, AGT, LTF, DCC, PGR
StemCell_Kocer08_87genes_18667080-Table S6	71	6	91.5	HSPA1B, CTSB, MCC, ACTR2, BTG1, KIAA0020
StemCell_Shim04_25genes_15246160-table6	22	3	86.4	KLF6, JUN, IDS
StemCell_Fruehauf06_110genes_16863911-Table 1	97	9	90.7	CLK4, HSPA1B, MS4A3, RNASE3, EGR3, HIST1H2BK, RNASE2, MPO, ELL2
DMAP_ERY_UP	45	9	80.0	XK, TRAK2, ARHGEF12, RHCE, TMCC2, GYPE, ACSL6, ANK1, HBBP1
DMAP_GM_EARLY_DN	42	10	76.2	DMP1, KCNH6, NAG18, ASCC2, EPB41L4A, LOC55338, SIX3, POLH, SEMA3C, SSX1
DMAP_PRE_BCELL_UP	39	7	82.1	LOC55338, CLDN14, POLR3G, POLH, DPYS, TFAP2A, LAMB4
DMAP_BCELL_DN	44	5	88.6	ACTN1, DMP1, NUCB2, BLZF1, ASCC2
DMAP_TCELLA6_DN	45	7	84.4	KLF6, CD58, DBI, FAS, SYT11, YWHAQ, AUTS2
StemCell_Tondreau08_52genes_18405367-Table2b	41	6	85.4	IGFBP7, MFAP5, COL8A2, HAS3, CALD1, PAWR
DMAP_BCELLA2_UP	49	6	87.8	CTSB, EGR3, CD1C, GIMAP5, DSE, IDH3A
DMAP_TCELLA6_UP	44	5	88.6	FKTN, GP5, CEPT1, IGF1R, NET1
IPA_affects differentiation of stem cells	72	5	93.1	RNF2, ANGPT1, GATA2, TLN1, NANOG
DMAP_ERY4_DN	47	5	89.4	HLA-DPB1, LILRA6, ACSM5, HLA-DMA, C4BPA
IPA_decreases differentiation of stem cells	18	5	72.2	JUN, DKK1, LIF, IL6ST, NEUROG1
StemCell_Colombo09_111genes_19123479-Table S1	92	8	91.3	OTUD1, HBP1, MGAT1, MTMR3, CHIC2, MIS12, TRIB1, FIP1L1
StemCell_Lim08_25genes_18510698-Table 2	25	3	88.0	MS4A3, JUN, ALOX5
DMAP_ERY_DN	46	13	71.7	EIF4B, ACSL5, GMFG, PDCD4, DBI, TES, RPL39, RPS3A, ZFP36L2, TRIM44, SMAD3, DICER1, RPL13A
DMAP_GM_EARLY_UP	40	10	75.0	GRHPR, TMEM156, EHD4, CR2, DYRK4, MRPS18B, GTF2H5, QTRTD1, BET1, SHMT2
DMAP_HSC1_DN	48	6	87.5	PLEK, ARAP3, TIMP3, PRKAR2B, DNAJA1, DNAJC6
DMAP_HSC3_UP	48	6	87.5	PLEK, ARAP3, TIMP3, PRKAR2B, DNAJA1, DNAJC6
DB_PPARG-19300518	194	17	91.2	MCM2, ATP1A2, NDUFV1, SMARCA4, DBI, CHIC2, G0S2, SDHC, LEP, COX15, RCL1, PDZRN3, FGF10, S100A8, UBE2I, ALDH3A1, ACADVL
StemCell_Bhattacharya05_2843genes_16207381-Table1Sa	312	24	92.3	IDI1, MCM2, WDR18, SOAT1, KLF6, YIPF1, FAR2, KIAA0020, ZCCHC6, GGCT, CD79B, TCEB3, GYG2, MAP4K4, MSMO1, CHMP2B, MTHFD2, HEATR5B, SNRPA, PICALM, THRAP3 STON1-GTF2A1L, JARID2, PREP
DMAP_MONO2_DN	40	7	82.5	ATP5J2, PRIM2, ZNF43, CUL7, TCIRG1, MYO15B, NDUFS6
DMAP_TCELLA2_DN	47	2	95.7	KLF6, CD58

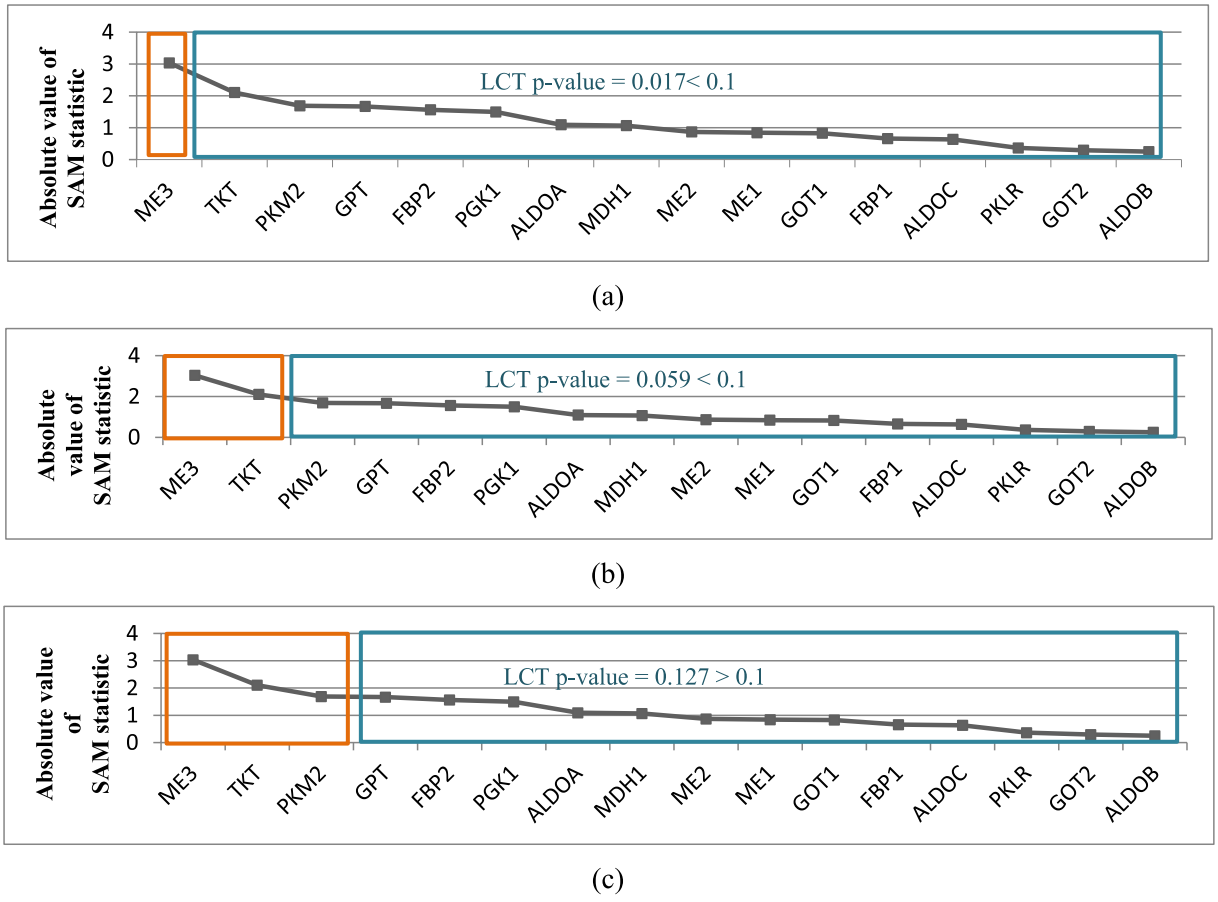

Fig. 2. An example of linear combination test gene set reduction. We used CARBON FIXATION gene set, identified to be significant by LCT. Each plot shows the absolute value of SAM statistic for genes within this gene set in a decreasing order. In this example we required three consecutive iterations of the gene set reduction method.

We demonstrate the gene set reduction method for the significant gene sets Carbon Fixation pathway, composed of 16 genes as defined in the C2 catalog. We rank the absolute value of SAM statistic for these 16 genes. First, we select the gene with the largest absolute value, ME3 with $|d_{(1)}| = 3.04$ to form the core subset, and the rest of the genes within the gene set to form the complement set. We apply the LCT method to the complement set and compare the LCT p-value with a pre-specified cut-off value of 0.1. Since the p-value is smaller than 0.1, we select the gene with the second largest absolute value of SAM statistic, i.e., TKT with $|d_{(2)}| = 2.10$. We sequentially add the gene to the core subset and test the complement set, until we reach the cut-off threshold. The p-value of the complement set is greater than 0.1, after taking out the third gene PKM2 with $|d_{(3)}| = 1.69$. Genes within the complement set, collectively are not associated with the phenotype and represent the redundant set. Therefore, the core subset contains three genes ME3, TKT and PKM2. Fig. 2 shows each step of the LCT-GSR algorithm.

Table 3 shows the summary of the LCT-GSR, including the list of gene sets along with the gene set size, LCT p-value, core set size, percent reduction, and the core pathway members. Core set size indicates the number of core genes obtained from each significant gene set, after applying the LCT-GSR algorithm. Percent reduction is computed by number of genes eliminated (in the complement set) divided by the total number of genes in a set, multiplied by 100. Core pathway shows the core genes collectively contributing to the association with tumor volume, excluding the redundant genes from the significant gene sets.

On average, we were able to reduce the number of genes in the 15 gene sets by 90%, using the threshold value of 0.1. We observed a situation where a whole gene set is reduced to a single gene. That suggests the genes within the complement subset are not associated with the phenotype. If the significance of a set is due to only one gene, the set should be investigated with caution. Biological functional role of the significant gene within the gene set should be considered.

There are 47 core genes obtained from the LCT-GSR method. The core gene *Malic Enzyme 3* (ME3) is the most frequent gene appearing in the reduced subset of three significant gene sets. The genes *Axis Inhibition Protein* (AXIN1), *Insulin-Like Growth Factor Binding Protein 6* (IGFBP6), *Arachidonate 15-Lipoxygenase, Type B* (ALOX15B), *Upstream Binding Transcription Factor* (UBTF), *High Mobility Group Nucleosomal Binding Domain 4* (HMGN4), *Pyruvate Kinase* Muscle (P KM2), *Cell Division Cycle 16* (CDC16), and *Transketolase* (TKT) appeared two times. The rest of thirty eight core genes appeared once in the significant gene set.

Our method identified pathways and genes that were previously discovered to be associated with the tumor volume, as well as new markers that need to be further validated. *Malic Enzyme 3*, a gene known to have an important role in cancer cell proliferation [33], appears most frequently in the three core subsets. Some well-characterized regulators of tumor volume, showing up in the core subsets, include: *Insulin-Like Growth Factor Binding Protein 6* [34], *Cell Division Cycle 16, Axis Inhibition Protein, Transketolase* and *Pyruvate Kinase Muscle* [35].

4. Discussion

We developed the LCT-GSR based on two computationally efficient and powerful methods, SAM and LCT. By using self-contained methods, we acknowledge that genes are not independent, and consider the coordination and network among genes, especially those that share

Table 3
Extracting core subsets for tumor volume for gene sets significant at a cut-off p-value of 0.01 (FDR < 0.35).

Gene set name	Gene set size	LCT p-value	Core pathway size	Percent reduction	Core pathway member
ELECTRON_TRANSPORTER_ACTIVITY	89	0.008	3	96.6	TSTA3, ME3, ALOX15B
CASPASEPATHWAY	19	0.005	1	94.7	BIRC2
GNATENKO_PLATELET	30	0.002	2	93.3	RGS10, SPARC
CARBON_FIXATION	16	0.001	3	81.3	ME3, TKT, PKM2
ZHAN_MMPC_EARLYVS	45	0.008	3	93.3	SPIB, SNRPC, SLC7A6
FALT_BCLL_DN	36	0.007	6	83.3	HEBP2,IFI6,HMGN4,SERP1,NPC2, PUM1
TPA_RESIST_EARLY_DN	65	0.003	3	95.4	ME3,POMZP3, DPP6
METASTASIS_ADENOCARC_DN	32	0.005	2	93.8	DLG3, RNASE1
AGED_RHESUS_DN	101	0.007	8	92.1	AXIN1, UBE2D2, DPP4, HMGN4, CDC16, RARRES2, JARID1C, SPARC
UVC_HIGH_D5_DN	23	0.006	3	87.0	SFRS3, DYRK1A, UBTF
XPB_TTD-CS_UP	19	0.003	2	89.5	PRKCZ, PTN
INNEREAR_UP	19	0.002	3	84.2	IGFBP6,RPS5, VAMP5
BCNU_GLIOMA_MGMT_48HRS_DN	123	0.005	5	95.9	ALOX15B,CRABP1,EPHX1,KIF5A, GP1BB
GH_EXOGENOUS_ALL_UP	22	0.007	2	90.9	NOS1, POU2F2
HSA00710_CARBON_FIXATION	16	0.001	3	81.3	ME3,TKT, PKM2

biological pathways. An important limitation of the self-contained approaches is that only a few genes, even one gene, can drive the association between the gene set and the phenotype. In such cases, post-hoc analysis can be useful to extract significant subsets associated with the phenotype. LCT-GSR is a simple analytical tool to reduce gene sets that have been found associated with the phenotype to smaller core sets, by gradually exploring the association of remaining genes, as a set, with the phenotype. The analyst can choose multiple cut-offs as stopping rule, moving from more conservative to more liberal values allowing for a flexible reduction process. Scientists can focus on biological interpretation of the reduced sets, instead of the whole sets.

We reason that traditional statistical methods cannot be applied here, because the number of genes in the gene sets is much larger than number of subjects, an issue also dubbed as the “high dimensionality problem”. Traditional statistical methods fail to converge when the number of features is larger than the number of subjects. As explained in the methods section, LCT incorporates the gene expression covariance matrix into the test statistic, and uses eigenvalue decomposition and shrinkage to maximize the correlation between phenotype and linear combinations of gene expressions. Permutation based approach is used to calculate p-values and FDR values. Therefore, LCT is a GSA method that incorporates the gene expressions correlations into the test statistic. This usually comes with a high computational cost, especially when permutation-based tests are used. However, we noticed the eigenvalue decomposition combined with an orthogonal transformation of the gene expression matrix, give LCT an important computational advantage: we do not have to decompose the covariance matrix for all the permuted versions of the dataset, but only once, for the original version of the dataset.

We selected the LCT approach among the other GSA methods to identify significant gene sets, as it outperformed other existing self-contained methods for a continuous phenotype in a simulation study [15]. The LCT method efficiently incorporates correlations among the genes in a set into the test statistic, while the other methods do not have this feature. Incorporating the covariance matrix into the test statistic and using permutation test results in better power [15]. The covariance matrix is singular when genes in a set are larger than the sample size, and this is a common situation in GSA. Shrinkage covariance matrix estimator can deal with this problem, but the computational cost of this approach is usually high. Orthogonal transformation of the gene expression is used to make this approach computationally efficient. As a result, the eigenvalue decomposition of the shrinkage covariance matrix is performed only once, for the real gene expression data, and there is no need to estimate it for each permuted version of the data.

We identified many important genes from the significant gene sets. Understanding biological function of these genes provides useful information on the underlying mechanism of birth weight, and their links to other diseases.

The main strength of our gene set reduction approach is integration of the biological information in the construction of the pathways. Identifying the core subsets of significant gene sets for a continuous phenotype has an important advantage of extracting biological information efficiently from extremely noisy microarray data, by interpreting only differentially expressed core sets. Our application to markers of variation in birth weight represents a situation in which genes show no or weak signals at an individual gene analysis level, but coordinating with other genes within a pathway, they show moderate to strong signals. The method is powerful in detecting biomarkers of complex diseases because it considers biological networks between genes. However, for prostate cancer data, the signal at the individual gene level seems to be strong, and in such situations, GSA will fail to show a significant advantage over an individual gene analysis methods, such as SAM. In such cases, the core sets will consist of only a few genes, or even a single gene.

Reducing significant gene sets to smaller sets can reduce costs of disease diagnosis and treatment, by focusing on smaller number of genes in screening massive databases for association with a continuous phenotype. Examination of redundant genes’ expression levels increases unnecessary costs, without a significant improvement in clinical decisions. Reduction to the most predictive genes is crucial in advancing our understanding of issues such as disease prevention, faster and more efficient diagnosis, intervention strategies and tailored treatment. It can lead to a change of platform from high-dimensional microarray technology to alternative methods, such as real time polymerase chain reaction (PCR) assays that are cheaper and faster. This alternative method is easily applicable to routine clinical setting for diagnosis purposes [36–38].

Please note LCT-GSR is a post-hoc analysis to LCT method. While simulations were possible for LCT, this is not the case for LCT-GSR. Similar to model selection where there is no unique solution for a “best” model, there may be more than one core subset that can be extracted from a significant set. This can be particularly challenging when the phenotype – gene expression signal is subtle, and in the presence of correlations across genes in a set, which are two fundamental reasons motivating GSA methodology development.

As mentioned in the introduction, there is no other GSA method available for gene set reduction, when the outcome is continuous. Our choice of LCT among GSA methods, as a basis for the core selection approach is motivated from previous simulation studies showing superior performance of LCT compared to other GSA self-contained methods. Published comparisons across GSA methods focus on either self-contained or competitive approaches. The literature on machine

learning techniques for high dimensional data abounds, however, the main goal of such methods is different than the main goal of GSA. Existing machine learning techniques, such as support vector regression, principal component analysis, Lasso, Elastic net [39], and several modifications of mixed effects models to accommodate clustered high dimensional data [40–43], do not assess significance of a pre-defined set of genes, but rather focus on extracting a set of features out of a list. In addition, many of these methods do not model correlations across a set of genes, but rather select only one out of a group of strongly correlated genes. Some of the methods can only select a number of genes smaller than the sample size. A recent article [44] that appeared in the preprint server for biology, BioRvix, not yet peer-reviewed, focuses on using regularized regression methods, such as Elastic net, to extract a minimum number of non-overlapping *sets* of genes from a catalog. It is beyond the scope of our paper to extend such methods to assess gene-set significance and/or extract core subsets and compare them with our proposed approach.

The methodological approach to gene set reduction for continuous phenotypes can be applied to a wide range of common situations in which dichotomizing the continuous phenotype is neither easy nor meaningful. The variable may not be informative about the disease mechanism after categorization based on arbitrary or less meaningful cut-off values. We hypothesized roles of gene expression variability and gene expression correlations with each other in the development of outcomes or diseases.

Significant gene sets and core subsets may have potential roles in occurrence and development of diseases, and further function enrichment analysis based on KEGG and GO, such as David platform, are necessary. However, the goal of our paper was to propose an analytical tool for post-hoc GSA to extract core subsets; comprehensive biological interpretations of significant gene sets for our applications on birth weight and prostate cancer are beyond the scope of our paper.

We were able to reduce the significant gene sets by 80%–90% in the CANDLE study. These genes need to be further investigated by experts to understand biomarkers contributing to low birth weight. Our method is based on a linear model, LCT, which is powerful but has its limitations. The LCT tests only linear associations between sets and a continuous phenotype. To check the linearity assumption, exploratory data analysis needs to be done. On the other hand, a small number of samples can be a limitation to check for non-linearity. LCT method can be extended to non-linear model if we can collect a large number of samples, however, this is not practical in real situations. We used the logarithmic transformation of the gene expression data and phenotypes to provide more support to the linearity assumption.

5. Conclusions

In conclusion, LCT-GSR is a powerful analytical tool based on a computationally efficient LCT method that can be used to extract core genes associated with a continuous phenotype. It can be applied to a wide range of studies in which dichotomizing the continuous phenotype is neither easy nor meaningful. Reduction to the most predictive genes is crucial in advancing our understanding of issues such as disease prevention, faster and more efficient diagnosis, intervention strategies and personalized medicine.

Availability of data and material

The birth weight data supporting the findings of this study is available from the CANDLE Study, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors, upon reasonable request and with permission of the CANDLE Study Team.

We downloaded the prostate cancer expression data file, as well as histopathologic features from Gene Expression Omnibus with accession ID GSE16560.

Authors' contributions

SV and ID developed the LCT-GSR methodology and designed/conducted the methodological study. SV was responsible for programming and conducting the data analysis, interpretation and presentation of the results. SP provided the study data and research question. AL provided the list of embryonic stem cell signatures. The manuscript was written primarily by SV and critically reviewed and revised by all authors. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the CANDLE study team for providing us with the birth weight data set. We thank the reviewers for their comments, which improved the manuscript significantly.

Abbreviations

CANDLE	Conditions affecting neurocognitive development and learning in early childhood
FDR	False Discovery Rate
GSA	Gene set analysis
GSEA	Gene Set Enrichment Analysis
IGA	Individual Gene Analysis
LCT	Linear combination test
SAM	Significance analysis of microarrays
SAM-GS	Significance analysis of microarrays for gene sets
SAM-GSR	Significance analysis of microarrays for gene set reduction

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.103389.

References

[1] J.J. Goeman, P. Buhlmann, Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics 23 (2007) 980–987.
[2] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 15545–15550.
[3] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000 Jan 1) 27–30.
[4] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (1) (2002 Jan 1) 207–210.
[5] D. Nishimura, BioCarta, Biotech. Software Internet Rep. 2 (3) (June 2001) 117–120.
[6] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, et al., Molecular signature database (MSigDB) 3.0, Bioinformatics 27 (12) (2011) 1739–1740.
[7] D. Nam, S.Y. Kim, Gene-set approach for expression pattern analysis, Briefings Bioinf. 9 (2008) 189–197.
[8] H. Maciejewski, Gene set analysis methods: statistical models and methodological differences, Briefings Bioinf. 15 (4) (2014 Jul) 504–518.
[9] W.T. Barry, A.B. Nobel, F.A. Wright, Significance analysis of functional categories in gene expression studies: a structured permutation approach, Bioinformatics 21 (9) (2005) 1943–1949.
[10] M.A. Newton, F.A. Quintana, J.A. den Boon, Random set methods identify distinct aspects of the enrichment signal in gene-set analysis, Ann. Appl. Stat. 1 (1) (2007) 85–106.
[11] B. Efron, R. Tibshirani, On testing the significance of sets of genes, Ann. Appl. Stat. 1 (1) (2007) 107–129.
[12] J.J. Goeman, S.A. van de Geer, F. de Kort, et al., A global test for groups of genes: testing association with clinical outcome, Bioinformatics 20 (1) (2004) 93–99.
[13] U. Mansmann, R. Meister, Testing differential gene expression in functional groups: Goeman's global test versus an ANCOVA approach, Methods Inf. Med. 44 (3) (2005) 449–453.
[14] I. Dinu, J.D. Potter, T. Mueller, et al., Improving gene set analysis of microarray data by SAM-GS, BMC Bioinf. 8 (2007) 242.
[15] I. Dinu, X. Wang, L.E. Kelemen, S. Vatanpour, S. Pyne, Linear combination test for gene set analysis of a continuous phenotype, BMC Bioinf. 14 (2013 Jul 1) 212.
[16] R. Delongchamp, T. Lee, C. Velasco, A method for computing the overall statistical significance of a treatment effect among a group of genes, BMC Bioinf. 7 (Suppl. 2)

(2006) S11.
[17] J.J. Chen, T. Lee, et al., Significance analysis of groups of genes in expression profiling studies, Bioinformatics 23 (2007) 2104–2112.
[18] J.J. Goeman, S.A. Van de Geer, F. de Kort, et al., A global test for groups of genes: testing association with clinical outcome, Bioinformatics 20 (1) (2004) 93–99.
[19] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 5116–5121.
[20] I. Dinu, J.D. Potter, T. Mueller, Q. Liu, A.J. Adewale, G.S. Jhangri, et al., Gene set analysis and reduction, Briefings Bioinf. 10 (1) (2008) 24–34.
[21] D.J. Storey, A direct approach to false discovery rates, J. R. Stat. Soc. B. 64 (3) (2002) 479–498.
[22] M.W. Gillman, D. Barker, D. Bier, F. Cagampang, et al., Meeting report on the 3rd international congress on developmental origins of health and disease (DOHaD), Pediatr. Res. 61 (5 Pt 1) (2007 May) 625–629.
[23] D.J. Barker, The developmental origins of adult disease, J. Am. Coll. Nutr. 23 (2004) 588S95S.
[24] O. Basso, A.J. Wilcox, C.R. Weinberg, Birth weight and mortality: causality or confounding? Am. J. Epidemiol. 164 (4) (2006 Aug 15) 303–311.
[25] A.P. Feinberg, R.A. Irizarry, D. Fradin, M.J. Aryee, et al., Personalized epigenomic signatures that are stable over time and covary with body mass index, Sci. Transl. Med. 2 (49) (2010 Sep 15) 49ra67.
[26] R.M. Adkins, F.A. Tylavsky, J. Krushkal, Newborn umbilical cord blood DNA methylation and gene expression levels exhibit limited association with birth weight, Chem. Biodivers. 9 (5) (2012) 888–899.
[27] J.W. Collins, R.J. David, A. Handler, S. Wall, S. Andes, Very low birthweight in African American infants: the role of maternal exposure to interpersonal racial discrimination, Am. J. Public Health 94 (12) (2004 December) 2132–2138.
[28] G. Van Vliet, S. Liu, M.S. Kramer, Decreasing sex difference in birth weight, Epidemiology 20 (4) (2009) 622.
[29] N. Novershtern, A. Subramanian, L.N. Lawton, R.H. Mak, et al., Densely interconnected transcriptional circuits control cell states in human hematopoiesis, Cell 144 (2) (2011 Jan 21) 296–309.
[30] A. Lachmann, H. Xu, J. Krishnan, S.I. Berger, et al., ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments, Bioinformatics 26 (19) (2010 Oct 1) 2438–2444.
[31] A. Sboner, F. Demichelis, S. Calza, Y. Pawitan, et al., Molecular sampling of prostate cancer: a dilemma for predicting disease progression, BMC Med. Genomics 3 (2010) 8.
[32] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (1) (2002 Jan 1) 207–210.
[33] F.J. Zheng, H.B. Ye, M.S. Wu, Y.F. Lian, et al., Repressing malic enzyme 1 redirects glucose metabolism, unbalances the redox state, and attenuates migratory and invasive abilities in nasopharyngeal carcinoma cell lines, Chin. J. Canc. 31 (11) (2012) 519–531.
[34] H. Koike, K. Ito, Y. Takezawa, T. Oyama, et al., Insulin-like growth factor binding protein-6 inhibits prostate cancer cell proliferation: implication for anticancer effect of diethylstilbestrol in hormone refractory prostate cancer, Br. J. Canc. 92 (8) (2005 Apr 25) 1538–1544.
[35] The human protein atlas, Available online: http://www.proteinatlas.org.
[36] M. West, G.S. Ginsburg, A.T. Huang, et al., Embracing the complexity of genomic data for personalized medicine, Genome Res. 16 (5) (2006) 559–566.
[37] J. Pittman, E. Huang, H. Dressman, et al., Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes, Proc. Natl. Acad. Sci. U.S.A. 101 (22) (2004) 8431–8436.
[38] L. Ein-Dor, O. Yuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome of cancer, Proc. Natl. Acad. Sci. U.S.A. 103 (15) (2006) 5923–5928.
[39] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Ser. Soc. B Stat. Methodol. (2005) 301–320.
[40] P.R. Loh, G. Tucker, et al., Efficient Bayesian mixed-model analysis increases association power in large cohorts, Nat. Genet. 47 (2015) 284–290.
[41] Z. Tan, K. Roche, et al., Scalable algorithms for learning high-dimensional linear mixed models, http://auai.org/uai2018/proceedings/papers/99.pdf.
[42] Z. Zhang, E. Ersoz, et al., Mixed linear model approach adapted for genome-wide association studies, Nat. Genet. 42 (4) (2010) 355–360.
[43] J. Schelldorfer, P. Buhlmann, et al., Estimation for high-dimensional linear mixed-effects models using $\ell 1$-penalization, Scand. J. Stat. 38 (2) (2011) 197–214.
[44] T. Fang, I. Davydov, D. Marbach, J.D. Zhang, Gene-set enrichment with regularized regression. BioRxiv, The Preprint Sever for Biology https://doi.org/10.1101/659920.