<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>A bootstrap-based regression method for comprehensive discovery of differential gene expressions: An application to the osteoporosis study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Lu, Y; Liu, YZ; Liu, PY; Dvornyk, V; Deng, HW</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>European Journal Of Medical Genetics, 2011, v. 54 n. 6, p. e560-e564</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2011</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/140927">http://hdl.handle.net/10722/140927</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>NOTICE: this is the author’s version of a work that was accepted for publication in European Journal of Medical Genetics. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in European Journal of Medical Genetics, 2011, v. 54 n. 6, p. e560-e564. DOI: 10.1016/j.ejmg.2011.07.002; This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
A bootstrap-based regression method for comprehensive discovery of
differential gene expressions: an application to the osteoporosis study

Yan Lu¹², Yao-Zhong Liu³, Peng-Yuan Liu²,
Volodymyr Dvornyk⁴, and Hong-Wen Deng¹³

1. College of Life Sciences and Bioengineering, Beijing Jiaotong University, Beijing 100044, P. R.
   China
2. Department of Physiology and the Cancer Center, Medical College of Wisconsin, Milwaukee, WI
   53226, USA
3. Department of Biostatistics and Bioinformatics & Center for Bioinformatics and Genomics, Tulane
   University School of Public Health, New Orleans, LA 70112, USA
4. School of Biological Sciences, University of Hong Kong, Pokfulam Rd., Pokfulam, Hong Kong, P.R.
   China

Running title: Bootstrap-based regression method for microarray analysis

Key words: microarray, bootstrap, regression, osteoporosis

Corresponding author:
Hong-Wen Deng, Ph. D.
Department of Biostatistics and Bioinformatics & Center for Bioinformatics and Genomics,
Tulane University School of Public Health, 1440 Canal Street, Suite 2001, New Orleans, LA
70112
Tel: 504-988-1310 Email: hdeng2@tulane.edu
Abstract

A common purpose of microarray experiments is to study the variation in gene expression across the categories of an experimental factor such as tissue types and drug treatments. However, it is not uncommon that the studied experimental factor is a quantitative variable rather than categorical variable. Loss of information would occur by comparing gene-expression levels between groups that are factitiously defined according to the quantitative threshold values of an experimental factor. Additionally, lack of control for some sensitive clinical factors may bring serious false positive or negative findings.

In the present study, we described a bootstrap-based regression method for analyzing gene expression data from the non-categorical microarray experiments. To illustrate the utility of this method, we applied it to our recent gene-expression study of circulating monocytes in subjects with a wide range of variations in bone mineral density (BMD). This method allows a comprehensive discovery of gene expressions associated with osteoporosis-related traits while controlling other common confounding factors such as height, weight and age. Several genes identified in our study are involved in osteoblast and osteoclast functions and bone remodeling and/or menopause-associated estrogen-dependent pathways, which provide important clues to understand the etiology of osteoporosis.

Availability: SAS code is available from the authors upon request.
Introduction

DNA microarrays are a powerful tool to provide a comprehensive picture of cell function as they can assay expression of tens of thousands of genes simultaneously. A typical microarray experiment may involve a comparison between disease and normal tissues, or a comparison between a strain grown under an experimental treatment and the same strain under a control condition. Various fold-change algorithms or t statistics were used for statistical analyses of this kind of two-sample experiment designs (1, 2). More complex experimental designs, in contrast, may comprise more than two samples as characterized by their genotypes, environments or developmental stages (3-5). A common purpose of these microarray experiments is to study the variation in gene expression across the categories of an experimental factor such as the above mentioned tissue types and drug treatments. However, an experimental factor is often a quantitative variable rather than a categorical variable. Loss of information would occur by comparing gene-expression levels between groups that are factitiously defined according to quantitative threshold values of an experimental factor. Additionally, lack of control for some sensitive clinical factors such as height, weight and age may yield serious false positive or negative findings.

In the present study, we described a bootstrap-based regression method for analyzing DNA expression data from the non-categorical microarray experiments. Bootstrap is an interesting method to select covariables in multivariable models. It allows increasing internal validity of models (6) and has been used widely in biology (7, 8). To illustrate the utility of this method, we applied it to our recent gene-expression study of circulating monocytes in subjects with a wide range of variations in bone mineral density (BMD). BMD, a quantitative clinical phenotype, is a major risk factor for osteoporosis in the elderly especially in the postmenopausal women (9). BMD has a large genetic determinant while is significantly affected by height, weight, age, and life-style factors such as smoking, exercise and alcohol consumption (10-14). Menopause is a major physiological event associated with accelerated bone loss in females (15). Using the proposed
method, we attempted to identify differentially expressed genes associated with BMD variation and menopausal events while controlling other confounding factors such as height, weight and age.

**Methods**

*Subjects and measurement*

The study subjects came from an expanding database being created for genetic studies of osteoporosis, which are underway in the Osteoporosis Research Center of Creighton University. Since our major goal is to find genes related to osteoporosis, the exclusion criteria to exclude diseases or medications known to affect bone metabolism were used. Given that monocytes were used in this study, the additional exclusion criteria were also implemented to exclude those diseases/conditions, which may lead to gene expression changes of blood monocytes. All the exclusion criteria were detailed earlier by Liu et al. (16).

We analyzed gene expressions of blood monocytes using Affymetrix HG-U133A GeneChip®, containing probes for 14,500 genes. The raw fluorescence intensity data within CEL files were processed with Robust Multichip Average (RMA) algorithm (17), as implemented with R packages from Bioconductor (www.bioconductor.org). Blood monocytes were obtained from 19 otherwise healthy women, each woman had total hip BMD, lumbar spine BMD (L1-L4), weight and height measurements, and completed a questionnaire with age, menopause status and years since menopause for postmenopausal women. Hip and spine BMD were measured with Hologic 4500 dual energy X ray absorptiometry (DXA) scanners (Hologic Corporation, Waltham, Massachusetts, USA). The machine was calibrated daily. The coefficient of variation (CV) values of the DXA measurements for BMD is 0.9%. Weight was measured using a calibrated balance beam scale; height was measured using a calibrated stadiometer. Table 1 presents the basic characteristics of the study subjects.

*Statistical methods*
Let $y_n$ be the expression level for gene $n$ and $x$ be a vector of experimental factors of interest (such as BMD, menopause status and years since menopause) and potential clinical covariates (such as height, weight and age). Then, the model for a multiple linear regression takes the form,

$$y_n = \alpha + \beta x + \epsilon$$

where $\alpha$ is regression intercept, $\beta$ are regression coefficients and $\epsilon$ is residual. Note that the years since menopause for premenopausal women were set to zero. The stepwise method was used for choosing independent variables in this multiple regression model. The selection $p$ value for both inclusion and exclusion of a variable in the regression model was set as 0.05. In our preliminary data, there were 6363 genes for each of which at least one independent variable was selected in the model. To obtain a robust list of important genes in relation to the independent variables, 1000 bootstrap resampling was used to estimate bootstrap frequencies of each variable selected in the regression model for each gene. We observed that the random sampling errors are very low when using 1000 bootstrap samples. If a variable has strong bootstrap support (frequency greater than 80%), this variable is regarded to be related with the tested gene. Although 80% bootstrap frequency was an empirical threshold, we believe this threshold is very stringent for choosing BMD-related genes. All statistical analyses were implemented in the SAS 8.0e (SAS Institute, Cary, N.C., USA).

**Results and Discussion**

Using the standard of bootstrap frequency greater than 80%, we found 75 hip BMD-related genes, 173 spine BMD-related genes, 221 menopause-related genes, and 139 years since menopause-related genes (Supplemental Table 1).

Among BMD-related genes, 37.4% of them were associated with cellular protein metabolisms based on the inference of gene ontology (GO). Particularly, *BMPRIA*, *TNFRSF10C*, *TNF*, *FIP2*, *TGFBR1*, *CCL11*, *PTHB1*, *CXCL3* and *MTHFR* are potentially involved in osteoblast and osteoclast functions and bone
remodeling according to the currently available literatures. BMPR1A is bone morphogenetic protein receptor (BMP), type 1A. A recent mouse mutation study demonstrated essential and age-dependent roles for BMP signaling mediated by BMPR1A in osteoblasts for bone remodeling (18). BMPR1A is located on the human chromosome 10q22.3. Klein et al. reported two mouse QTLs for whole body BMD in the genomic regions homologous to human 10q21-24 and 10q23-26 (19, 20). TNFRSF10C, TNF and FIP2 are members of tumor necrosis factor (TNF) superfamily. TNF is a proinflammatory cytokine that promotes osteoclastic bone resorption and inhibits osteoblast differentiation (21-23). CCL11 is another proinflammatory cytokine important to osteoarthritis. The CCL11 expresses eotaxin-1 induced by treatment with interleukin-1β and TNF-α and plays an important role in cartilage degradation in osteoarthritis. TGFBR1 is a transforming growth factor (TGF) beta receptor whose expression is important in mediating 1α, 25(OH) 2D3-associated changes in the growth rate of osteoblasts (24). PTHB1 (parathyroid hormone-responsive B1) is downregulated by parathyroid hormone in osteoblastic cells, and therefore, is thought to be involved in parathyroid hormone action in bones. CXCL3 is a chemokine (C-X-C motif) ligand. CXCL3 is potentially involved in chemokine activity based on the NCBI GO inference, which is known to participate in bone monocyte recruitment (25, 26). MTHFR (5,10-methylenetetrahydrofolate reductase) can affect the methylation of homocysteine to methionine and high serum homocysteine concentrations have adverse effects on bone (27, 28). A polymorphism of the MTHFR gene, C677T, which causes an alanine to valve substitution and gives rise to a thermolabile variant of the MTHFR protein with reduced activity (29), was associated with elevated levels of circulating homocysteine (30) and lumbar spine BMD (31). During the past two years, a number of association studies appeared with regard to its relevance to osteoporosis (32-35). It is worth noting that several osteoblast-related genes such as BMPR1A and PTHB1 were found in circulating monocytes which are early precursors of osteoclasts. Bone loss in osteoporosis is due to the persistent excess of bone resorption over bone formation. Osteoblasts play a central role in bone formation by synthesizing multiple bone matrix proteins, while they also regulate osteoclast maturation by soluble factors and cognate interactions, resulting in bone resorption. For example, while FGF-2 induces RANKL expression by
osteoblasts, it also inhibits osteoclast differentiation directly by interfering with the action of M-CSF (36). Thus, osteoclast differentiation, formation, and, to a lesser degree, activation depend upon the proximity and products of the osteoblast. Therefore, it is not unexpected to identify osteoblast-related genes differentially expressed in monocytes in the present study. These genes may regulate the balance between osteoblast and osteoclast activity that is important for bone remodeling and health.

Among genes that are related to menopausal status or years since menopause, several are upstream and downstream targets of estrogen receptors, including MAPK1, U29725, laminin α2, laminin β1, PLA2, PRKD3, ALPP, ADAM2, ADAM21, ADAM22, ADAMTS2, ADAMTS6, TNFSF13, TNFRSF13B, TNFRSF25, TRADD, TNFAIP6 and IL22R. The expression changes of these genes may be due to the fact that ovaries reduce and stop producing estrogen in postmenopausal women. In our data, the status of menopause had effects on MAPK1 gene expression (bootstrap frequency=83.5%) and years since menopause affected U29725 (MAPK7) gene expression (bootstrap frequency=93.5%). Estrogen can activate the MAPK family member extracellular regulated kinase-1 (ERK-1). Increases in ERK activation coincided with increased ER-α phosphorylation. Reduced availability of this pathway when estrogen levels are reduced could explain diminished effectiveness of mechanically related control of bone architecture after the menopause (37). Two laminin genes, laminin α2 and laminin β1, were detected to be associated with menopausal status and years since menopause, which were potentially involved in maintenance of estrogen receptor alpha expression (38). PLA2 (urokinase-type plasminogen activator) was shown to be downregulated by ovariectomy but restored with estrogen during fracture healing (39). PRKD3 is a member of protein kinase C (PKC) family. Recent data revealed a direct PKCα-c-Src-ERα interaction, which may be crucial in the modulation of estrogen responsiveness and the differentiation process in osteoblasts (40). ALPP (alkaline phosphatase) is a bone formation marker. Bone alkaline phosphatase was significantly increased in postmenopausal women (41). Interestingly, we found three genes (ADAM2, ADAM21 and ADAM22) that encode a member of the ADAM (a disintegrin and metalloprotease domain) family and two genes (ADAMTS2 and ADAMTS6) that encode a
member of the ADAM with thrombospondin motifs (ADAMTS) protein family. Both ADAM and ADAMTS
gene families have been shown to play a role in bone osteoblast function in several recent studies (42-45). It
has been suggested that one of the mechanisms by which estrogen protects against postmenopausal
osteoporosis is by modulating the production of cytokines, such as tumor necrosis factors (TNF), interleukin-
1 (IL-1) and interleukin-6 (IL-6), in the bone microenvironment (46). In our data, five TNF superfamily
genes (TNFSF13, TNFRSF13B, TNFRSF25, TRADD and TNFAIP6) and one interleukin cytokine (IL22R)
were associated with menopausal female events.

Figure 1 shows hierarchical clustering for 19 samples using differential expression associated with
BMD variation, menopausal status, and years since menopause. According to the results of the cluster
analysis, women with hip BMD value greater than 0.9 were classified into one group and those smaller than
0.9 were classified into another group. Pre- and postmenopausal women also had quite distinct patterns of
gene expression and therefore were clearly clustered into two well-defined groups. Based on the expression
data associated with years since menopause, those women who have not experienced or recently experienced
menopause were classified into one group, while those women who have experienced menopause for a long
time (at least more than 4 years) were classified into another group.

Conclusions

In summary, we described a bootstrap-based regression method for handling expression data from the
non-categorical microarray experiments. This method allows a comprehensive discovery of gene expressions
associated with osteoporosis-related traits while controlling other common confounding factors such as
height, weight and age. We prioritized a small list of candidate genes for future confirmation studies in terms
of their functional relevance to osteoblast and osteoclast functions and bone remodeling and/or menopause-
associated estrogen-dependent pathways. These gene expression data provide important clues to understand
the etiology of osteoporosis.
Acknowledgments

Investigators of this work were partially supported by grants from NIH, State of Nebraska and CNSF. The study was also benefited by grants from Huo Ying Dong Education Foundation, the Ministry of Education of China, and Xi’an Jiaotong University.
Reference List

(1) Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 2002 Sep, **18**(9), 1207-1215.


(32) Abrahamsen B, Madsen JS, Tofteng CL, et al. A common methylenetetrahydrofolate reductase (C677T) polymorphism is associated with low bone mineral density and increased fracture incidence


**Figure legend**

**Figure 1** Hierarchical clustering for 19 samples using differential expression associated with (A) BMD variation, (B) menopausal status, and (C) years since menopause. Red indicates upregulated genes, and green represents down-regulated genes. At the top of panel A, the numbers represent hip BMD value of each woman. At the top of panel B, the letter “pre” represents premenopausal women and “post” represents postmenopausal women. At the top of panel C, the numbers represents years since menopause for each woman. Note that the years since menopause for premenopausal women were set to zero.
Table 1 Basic characteristics of the study subjects

<table>
<thead>
<tr>
<th>Trait</th>
<th>Mean ± SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hip BMD (g/cm(^2))</td>
<td>0.932 ± 0.169</td>
<td>0.721-1.339</td>
</tr>
<tr>
<td>Spine BMD (g/cm(^2))</td>
<td>1.023 ± 0.234</td>
<td>0.748-1.428</td>
</tr>
<tr>
<td>Age (years)</td>
<td>51 ± 2.54</td>
<td>47-55</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>167.0 ± 7.5</td>
<td>154.8-177.4</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>77.0 ± 21.7</td>
<td>48.7-126.1</td>
</tr>
<tr>
<td>Years since menopause (years)</td>
<td>6.4 ± 4.7</td>
<td>1-15</td>
</tr>
</tbody>
</table>