Whole Genome Arrays for Osteolysis: Consequences of Filtering Microarray Data

Arun S. Shanbhag, Kunal Vaidya*, Samreen Jatana*, Neha Girish*, Praful Katta*, G. K. Prabhu*, Mahito Kuwahara and Harry E Rubash

Biomaterials Laboratory, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA,
*Department of Biomedical Engineering, Manipal Institute of Technology, Manipal 576104, Karnataka, India

ABSTRACT

Microarrays allow us to study the expression profile of a large number of genes. However, the voluminous data needs to be efficiently processed if we are to make meaningful inferences about the treatment or disease under investigation. The ultimate goals of a microarray investigation are to identify genes, which are differentially expressed with the pathological process and develop biomarkers for therapeutic intervention.

In studies conducted in our laboratory, we have used Affymetrix Gene Chips to define the gene expression profile of aseptic loosening and osteolysis around total hip replacements. This gene chip contains over 54,000 probe sets representing over 38,500 well characterized genes. Thus judicious filtering of the large data set is necessary to arrive at a more manageable number of genes for further detailed investigations. In this paper we have discussed the consequences of various filtering techniques utilized to remove nearly 95% of the genes from further consideration. These filtering techniques yield us about 2,100 genes which likely define the molecular basis of the pathophysiology of aseptic loosening and osteolysis around total hip replacements. This subset of genes is more amenable to a detailed study of specific genes associated with pathological processes.

1. INTRODUCTION

Total hip replacements (THR) have a long history of successfully reducing pain and improving the quality of life of patient with debilitating arthritis and other joint diseases. In a small percentage, but a large number of patients, failure of their implants is associated with bone loss requiring more complicated revision surgery to remove and replace the components. Thus there is an increasing need to understand the biological aspects of implant failure and use therapeutic modalities to circumvent the process.

Failure of THR is caused by an active bone resorptive process around the implants. Numerous in-vitro and animal models indicate that the biological response to wear debris generated at the articulating surface is the primary driver of the bone resorptive process [1]. Macrophages phagocytize the wear debris and are stimulated to release a variety of inflammatory mediators such as interleukin (IL) IL-1 and IL-6 [2-5]. These mediators are also potent stimulators of osteoclastogenesis, recruiting precursors and facilitating their differentiation to mature osteoclasts. The active resorption process compromises implant stability, initiating a cycle of more wear, further release of inflammatory mediators and bone resorption, ultimately resulting in a painfully loose implant [1].

Based on earlier experimental models, the peri-implant granulomatous tissues surrounding the loose implant itself hold the key to understanding the pathophysiology of osteolysis [6]. Earlier investigations have relied on measuring levels of selected proteins in these tissues as well as the expression levels of associated genes. Such a selective approach has been unsuccessful in previous attempts to identify the key molecular drivers of osteolysis [7]. The overall purpose of this study was to use genome-wide gene expression profiling to comprehensively define the molecular participants in this pathology.

2. MATERIALS AND METHODS:

2.1 Clinical Materials:

Peri-implant granulomatous tissues, also referred to as interfacial tissues, were harvested from patients (n=9) undergoing revision surgery for aseptic loosening of their THR. For control comparisons, capsular tissues were harvested from patients (n=5) with end stage osteoarthritis during their primary THR procedure. In the operating room, tissue samples were flash frozen in liquid nitrogen and stored at (-) 76°C awaiting RNA extraction. Approximately 1 g of tissue was homogenized in the presence of 2 mL Trizol reagent. RNA was extracted using established procedures and additionally purified using RNeasy Mini Spin columns (Qiagen Inc., Valencia, CA).

2.2 Microarray Procedure:

RNA quality was determined spectrometrically as the ratio of absorbance at A260/A280 nm (1.9 – 2.1). Additionally, clearly

Dr Arun Shanbhag is Assistant Professor of Orthopaedic Surgery, Harvard Medical School, Massachusetts General Hospital, Boston, MA
Kunal Vaidya is a Senior Biomedical Engineering Student at the Manipal Institute of Technology, Manipal, India
Samreen Jatana is a Senior Biomedical Engineering Student at the Manipal Institute of Technology, Manipal, India
Neha Girish is a Senior Biomedical Engineering Student at the Manipal Institute of Technology, Manipal, India
Praful Katta is a Senior Biomedical Engineering Student at the Manipal Institute of Technology, Manipal, India
Dr G K Prabhu is Associate Director (R&D) and Professor of Biomedical Engineering at the Manipal Institute of Technology, Manipal, India
Dr Mahito Kuwahara is a Research Fellow in the Department of Orthopaedic Surgery, Massachusetts General Hospital, Boston, MA
Dr Harry Rubash is the Edith Ashley Professor at Harvard Medical School and Chief of the Department of Orthopaedic Surgery, Massachusetts General Hospital, Boston, MA

Address Correspondence to:

Arun Shanbhag, PhD, MBA
GRJ 1115, 55 Fruit St
Boston, MA 02114
(617) 724-1923
shanbhag@helix.mgh.harvard.edu

defined 18S and 28S ribosomal peaks were analyzed using the Agilent 2100 Bioanalyzer. Double stranded cDNA was synthesized sequentially by first strand and second strand using protocols established by Affymetrix (Santa Clara, CA). Labeled and fragmented cDNA were mixed with control oligonucleotides and internal controls to create a hybridization cocktail. A test chip was run to verify sample quality and the hybridization procedure was repeated with Affymetrix HG-U133A Plus 2.0 Chipset (Affymetrix, Santa Clara, CA). This gene chip provides a comprehensive coverage of the transcribed human genome on a single chip with 54,120 probe sets; including 38,500 well-characterized human genes.

2.3 Data Analyses:

The microarray intensity data was analyzed using the Affymetrix MAS 5.0 software and after preliminary data normalization, Call Definitions were determined for each gene set.

On an Affymetrix gene chip, eleven pairs of oligonucleotides probes are used to measure the levels of transcription of each gene sequence represented on the gene chip. To determine the specificity of hybridization, every set of perfect match probes has a corresponding set of mismatch probes. A mismatch probe is constructed from the same nucleotide sequence as its perfect match probe partner, except that the middle (usually the 13 th) base pair has been switched to result in an alphabet mismatch. Signals associated with the mismatched probe are attributed to a stray signal. Using a statistical-decision tree, signal intensities of the perfect and mismatched probes are compared and a decision (or a call) is made to identify the signal for each probe as either ‘Present,’ ‘Absent’ or ‘Marginal.’ While a Present call indicates statistical surety of the gene expression intensity, an Absent call does not indicate an absence of gene expression, but rather an unreliable level of gene expression measurement relative to the mismatched probe [8].

A major drawback of performing a whole genome array is the large number of available genes for analyses. We developed systematic procedures to evaluate the various filtering techniques to judiciously reduce the data set. During the first step, we used the Present call as a primary filtering technique. We selected the osteolysis cases (n=9) as our target population and identified those genes which were ‘Present’ in all n=9 osteolysis cases with no regard to their presence in the control samples. The Present call provided the most dramatic reduction in genes under consideration and resulted in 11,080 genes which were carried forward for further analyses. Figure 1 illustrates this drastic reduction in genes as the number of present calls is increased. According to the literature, a 25% Present call in a treatment group is sufficient to dramatically reduce the false selection rate of genes in a microarray experiment [8]. Our selection of 100% Present call in osteolysis cases essentially reduces the false discovery rate to 0.00008 at a probability level of $p < 0.01$.

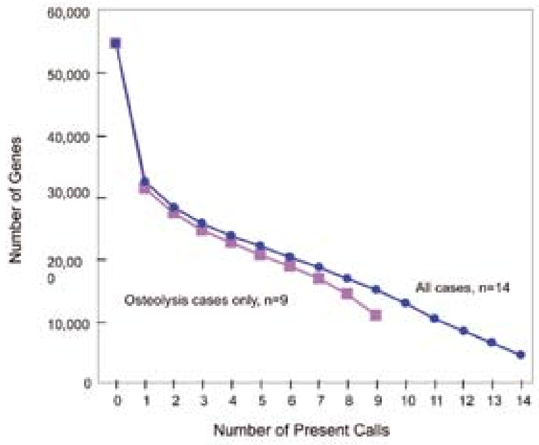

Figure 1. Variation in the number of genes with respect to changes in number of Present Calls.

3. FILTERING OPERATIONS

On this data set we applied various filtering parameters such as p-value based on the t-statistic, Osteolysis/Control (O/C) ratio and the case mean. The goal was to study the effect of these robust estimators of significance and variance on the number of genes that are screened.

3.1 p Value Based on t-Statistic:

For each of the 11,080 genes, the means for the osteolysis and the control groups were compared using the t-statistic. The t-statistic accounts for the variability of gene expression for each gene. An important drawback is that since a large number of comparisons are required, equal to the number of genes being compared, there is a higher likelihood of finding a statistical significance by chance alone. Further since the t-statistic relies on the estimation of central tendency (mean) and dispersion (standard deviation), it is strongly affected by outliers [9]. In our microarray data outliers are the highest levels of up- or down- regulated genes resulting in an increase in the estimation of standard deviation, which reduces the power of the test. Figure 2 is the graphical representation of filtering results obtained by varying the p-value for the t-statistic in our data set. The graphs indicate that the region of the curve between the p values of 0.001 and 0.005 are most sensitive, causing a significant reduction of genes.

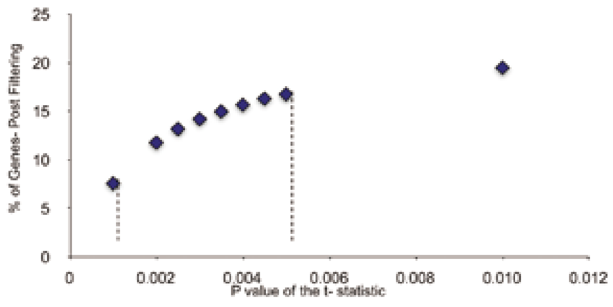

Figure 2. Changes in the percentage of genes of the data set with changing p-value.

3.2 Case Mean and Standard Deviation:

We used the arithmetic mean (or Case Mean) of the gene expression intensities of the nine osteolysis samples as filter parameter. It essentially tells us how highly a particular gene is expressed in the entire group relative to the control samples. Thus filtering using the case mean was an important parameter

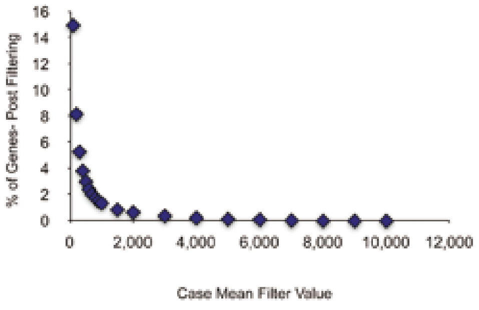

Figure 3. Percentage of genes of the entire data set with respect to case mean filter value.

directly identifying highly expressed genes. As presented in Figure 3, we observed that high values of case mean reduced the number of genes under consideration; yet it was not advisable to filter at a high case mean threshold, as it would eliminate important genes responsible for the disease. Based on trends in the literature we settled for a case mean value of 100 [10]. This threshold eliminates only those genes with a lower level of expression and ensures that significant genes are still a part of the data set to be used for post filtering analyses.

3.3 Osteolysis/Control Ratio:

The osteolysis/control (O/C) ratio is obtained by simply dividing the mean intensity of the osteolysis cases by the mean intensity for the control cases, for each gene. This ratio accounts for consistently up- or down- regulated genes in both cases, and provides a quick assessment of how intensely a particular gene is upregulated in the osteolysis cases as compared to the control case. Genes with O/C ratio greater than one are considered upregulated and those with O/C less than one are down regulated. An O/C ratio of 2 or higher is generally used. While it would be tempting to use a higher O/C ratio, it could also screen out important genes which are tightly regulated, as would be expected in a chronic pathological case. Another practical drawback of this parameter is that it has to be applied in conjunction with an acceptable t-statistic p-value.

A cautionary note and a potential drawback is that since our controls are not healthy individuals but represent tissues affected by severe degenerative joint disease, many inflammatory genes would be expected to be upregulated in the controls as well. Thus it is likely that the O/C ratio could be artifactually lower in many related genes and not make the threshold. This is another reason for a reasonably lower O/C ratio. Figure 4 demonstrates the effect of varying the Osteolysis/Control ratio on the total number of screened genes. It is evident that the O/C parameter is more sensitive than the case mean at filtering out genes.

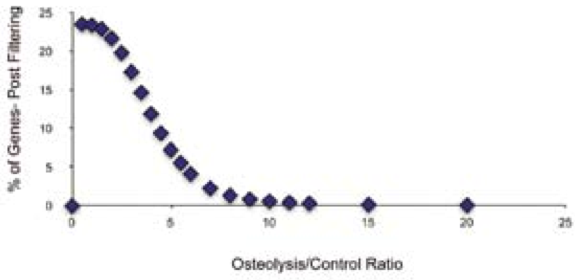

Figure 4. Filtering genes based on the Osteolysis/Control ratio.

3.4 Trade Offs in Filtering:

Optimal filtering of genes is a trade off between the various filtering parameters discussed here. Depending on particulars of different pathologies to be investigated, individual investigators may include other more pertinent parameters. Figure 5 shows the combined effect of the parameters we used: t-statistic p value, O/C ratio and the case mean. These were applied in addition to the nine Present Call. When used in combination, the filtering yielded a data set with a significantly reduced set of genes, which could be easily managed in the next stage of analyses, namely, clustering of genes and identification of the biochemical pathways. Our results demonstrate that while the osteolysis/control ratio was useful in filtering, the t-statistic had the most profound effect in controlling the number of genes to screen.

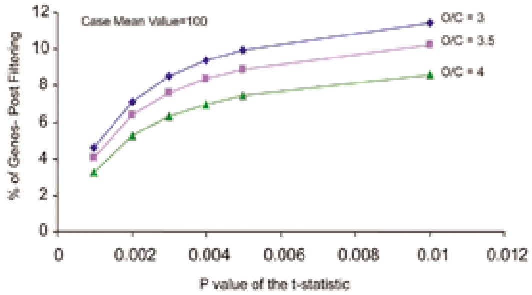

Figure 5. The combined effect of t-statistic and Osteolysis/Control ratio.

4. POST FILTERING OPERATIONS ON MICROARRAY DATA

Reducing the data set by over 95%, from 54,000 probes to less than 2,500 genes represents a major step in analyzing microarray data. The next step is to pinpoint genes responsible for osteolysis. Clustering methods attempt to identify genes that behave similarly across a range of conditions or samples. The motivation to find such genes is driven by the assumption that genes that demonstrate similar patterns of expression share common characteristics, such as common regulatory elements, common functions, or in the case of tissue studies as ours, common cellular origins. Several clustering methods can be applied to gene expression data, hierarchical clustering being the most popular. But others methods include *k*-means, deterministic annealing, self organizing maps, combinatorial methods and graph theoretical approaches [10]. Clustering results are directly dependent on the input data in the clustering algorithms, which reiterates the importance of filtering as an important preliminary step in data analyses - eliminating genes not responsible for the disease under study. As every subsequent step in our analyses of genes that cause osteolysis is inter-dependent, hence the number of genes obtained after our filtering steps bear a direct consequence on the results obtained by the study of biochemical pathways.

5. CONCLUSION

The purpose of our gene expression profiling investigation was to identify genes responsible for driving peri-implant osteolysis. The data set obtained after multiple filtering operations contains a set of genes most likely associated with osteolysis. This limited data set needs to be processed further in order to identify a more relevant and smaller set of genes. Further analyses include Clustering methods (such as hierarchical clustering, k-means clustering) and followed by the identification of biochemical pathways. Our final picks for effective filtering are processing a nine Present call data set, a case mean of 100, an osteolysis/controls C ratio of 3 and p-value less than 0.001. These values are not absolutely rigid and can be modified during subsequent clustering operations. Thus the results of clustering can be assessed and the filtering parameters can be varied in an iterative process.

References

1. **Shanbhag AS, Sethi MK, Rubash HE.** Biological Response to Wear Debris: Cellular Interactions Causing Osteolysis, In the Adult Hip. 2nd Edition; J Callaghan, HE Rubash & AJ Rosenberg (Eds.) Lippincott Williams & Wilkins, Philadelphia, PA; vol 1, Chapter 22, 286-303. 2006
2. **Dorr LD, Bloebaum R, Emmanual J, Meldrum R.** Histologic, biochemical and ion analysis of tissue and fluids retrieved during total hip arthroplasty. *Clin Orthop Relat Res*. 1990; 261:82-95.
3. **Goodman SB, Chin RC, Chiou SS, Schurman DJ, Woolson ST, Masada MP.** A clinical-pathologic-biochemical study of the membrane surrounding loosened and nonloosened total hip arthroplasties.*Clin Orthop Relat Res*.1989; 244:182-187.
4. **Kim KJ, Rubash HE, Wilson SC, D'Antonio JA, McClain EJ.** A histologic and biochemical comparison of the interface tissues in cementless and cemented hip prosthesis. *Clin Orthop*.1993; 287:142-152.
5. **Shanbhag AS, Jacobs JJ, Black J, Galante JO, Glant TT.** Cellular mediators secreted by interfacial membranes obtained at revision total hip arthroplasty. *J Arthroplasty*. 1995; 10(4):498-506.
6. **Goldring SR, Schiller AL, Roelke M, Rourke CM, O'Neil DA, Harris WH.** The synovial-like membrane at the bone-cement interface in loose total hip replacements and its proposed role in bone lysis. *J Bone Joint Surg [Am]*. 1983; 65:575-584.
7. **Schwarz EM, Campbell D, Totterman S, Boyd A, O'Keefe RJ, Looney RJ.** Use of volumetric computerized tomography as a primary outcome measure to evaluate drug efficacy in the prevention of peri-prosthetic osteolysis: a 1-year clinical pilot of etanercept vs. placebo. *J Orthop Res* 2003; 21(6):1049-1055.
8. **McClintick JN, Edenberg HJ.** Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics*. 2006; 7:49.
9. **John Quackenbush.** Microarray data normalization and transformation. *Nature Genetics*. 2002; 32 Suppl: 496-501.
10. **Kaminski N, Friedman N.** Practical approaches to analyzing results of microarray experiments. *Am. J. Respir. Cell Mol Biol*. 2002; 27(2):125-132.