Gene Expression Phenotypes of Atherosclerosis
David Seo, Tao Wang, Holly Dressman, Edward E. Herderick, Edwin S. Iversen, Chunming Dong, Korkut Vata, Carmelo A. Milano, Fabio Rigat, Jennifer Pittman, Joseph R. Nevins, Mike West and Pascal J. Goldschmidt-Clermont
DOI: 10.1161/01.ATV.0000141358.65242.1f
Arteriosclerosis, Thrombosis, and Vascular Biology is published by the American Heart Association.
7272 Greenville Avenue, Dallas, TX 75214
Copyright © 2004 American Heart Association. All rights reserved. Print ISSN: 1079-5642. Online ISSN: 1524-4636

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://atvb.ahajournals.org/cgi/content/full/24/10/1922
Gene Expression Phenotypes of Atherosclerosis

David Seo, Tao Wang, Holly Dressman, Edward E. Herderick, Edwin S. Iversen, Chunming Dong, Korkut Vata, Carmelo A. Milano, Fabio Rigat, Jennifer Pittman, Joseph R. Nevins, Mike West, Pascal J. Goldschmidt-Clermont

Objective—Fulfilling the promise of personalized medicine by developing individualized diagnostic and therapeutic strategies for atherosclerosis will depend on a detailed understanding of the genes and gene variants that contribute to disease susceptibility and progression. To that end, our group has developed a nonbiased approach congruent with the multigenic concept of complex diseases by identifying gene expression patterns highly associated with disease states in human target tissues.

Methods and Results—We have analyzed a collection of human aorta samples with varying degrees of atherosclerosis to identify gene expression patterns that predict a disease state or potential susceptibility. We find gene expression signatures that relate to each of these disease measures and are reliable and robust in predicting the classification for new samples with >93% in each analysis. The genes that provide the predictive power include many previously suspected to play a role in atherosclerosis and additional genes without prior association with atherosclerosis.

Conclusion—Hence, we are reporting a novel method for generating a molecular phenotype of disease and then identifying genes whose discriminatory capability strongly implicates their potential roles in human atherosclerosis. (Arterioscler Thromb Vasc Biol. 2004;24:1922-1927.)

Key Words: clinical human atherosclerosis genomic gene expression profile disease genetic susceptibility molecular signature

Atherosclerosis is the leading cause of morbidity and mortality in the industrialized world. Although we have made substantial progress in the treatment and prevention of atherosclerosis and the related thromboembolic complications, there remains an urgent need to develop individualized prognostic tools and therapeutic plans. Progress in this direction will come with an improved understanding of the genetic foundations of the disease.1–4

See page 1746

As with all complex human diseases, atherosclerosis likely results from the interaction between multiple genetic and environmental factors. Unlike Mendelian disorders like familial hypercholesterolemia and Tangier disease, development of “garden variety” atherosclerosis is not attributable to single genes. More likely, it is the inheritance of an ensemble of gene variants in the form of single nucleotide polymorphisms (SNPs) that define an individual’s atherosclerotic susceptibility by modulating responses to environmental factors, such as diet and tobacco use. Individually, SNPs generally have a mild to moderate effect on the function or quantity of encoded proteins. However, specific combinations of SNPs may have a dominant effect on the development of atherosclerosis. Knowledge of the crucial genes and gene variants will allow us to generate a genomic patient phenotype that will provide the level of detail necessary for dramatically improving diagnosis and prognosis, and for developing personalized treatment regimens.

In this article, we describe a novel method to identify potential genetic factors that determine an individual’s predisposition to atherosclerosis. We generated gene expression data using RNA extracted from a unique collection of freshly harvested human aortas with varying degrees of atherosclerosis. Using a novel analytical approach, we have identified gene expression signatures that differentiate between atherosclerotic disease states in the human aortas studied. In fact, these genomic phenotypes allow us to predict the disease state of an unknown aorta sample with a high degree of accuracy. Furthermore, we have identified the genes whose expression patterns provide this predictive capability, thus strongly implicating their role in atherosclerosis. The genomic information from this study will serve as the foundation for more detailed patient characterization that can be combined with traditional demographic and risk factor information to create individualized patient phenotypes.
Materials and Methods

Tissue Collection and Characterization

The aortas of heart donors were collected at organ harvest in University of Wisconsin solution on ice to minimize postmortem changes. Figure 1 shows the aorta sectioning. Strips A and B were frozen in liquid nitrogen for RNA extraction. Strip C was preserved in formalin for scoring atherosclerotic burden. The Pathobiological Determinants of Atherosclerosis in Youth Study (PDAY) investigators showed that atherosclerosis displays increasing severity from proximal to distal locations. Also, they found sagittal symmetry with respect to disease burden across the longitudinal midline that approached 0.99. The mirroring of disease allowed us to ascribe the clinical atherosclerosis phenotype from one side to the gene expression pattern of the other. Early atherosclerotic plaques were assessed with imaging processing software by quantifying the area of Sudan IV staining. Advanced disease was quantified using PDAY methodologies to evaluate raised lesions. The data were expressed as a ratio of affected area over total surface of the studied section. Permission to procure these tissues was obtained from the North Carolina Organ Donor Services executive committee and approved by the institutional review board of the Duke University Medical Center.

RNA Preparation and Microarray Processing

Techniques for microarray assays have been previously reported by our group. Aortic tissue was ground in liquid nitrogen. The RNA was extracted by the Trizol protocol and further purified with the Qiagen RNeasy kit. Quality was assessed with the Agilent Bioanalyzer and Affymetrix Test3 chips. The targets for microarray analysis were hybridized to U95Av2 Affymetrix microarrays and processed with the GeneChip system. The signal intensity values were converted to a log2 scale after quantile normalization. Quantitative real-time polymerase chain reaction reactions were performed on 13 aorta samples and compared with microarray expression measures with a high degree of correlation across these genes. Replicate microarray assays were performed also showing a high degree of correlation in signal intensities (data not shown).

Design of Phenotyping Studies

We prioritized genes by their ability to predict 2 clinical phenotypes: disease extent and aorta location. Disease extent was scored by combining Sudan IV staining and raised lesion data. The “minimally diseased” group showed less than 5% Sudan IV staining and contained no raised lesions. The “severely diseased” group contained both raised lesions and extensive Sudan IV staining. We analyzed sections from 2 identical locations in all the aortas, a proximal and distal section. From this pool, we identified 15 minimally diseased and 16 severely diseased sections for this analysis. Nine of these sections came from single aortas. The size of a particular section used in the analysis was quite small, on average 10 mm by 5 mm, making Sudanaphilia and raised lesion content homogeneous throughout the section.

The second phenotype was the location of the section within the thoracic aorta as a surrogate for disease susceptibility. This assumption is based on the conclusive evidence from the PDAY Study that progression of disease advances from the distal to proximal areas of the aorta, suggesting that distal regions are more susceptible to disease development. As stated above, we analyzed sections from identical locations in all the aortas. There were 31 proximal (1A) sections and 32 distal (4B) sections in our analysis of aorta location. We used the same pool of aorta sections for both analyses.

Statistical Analysis

Statistical analysis was performed using the metagene construction and binary prediction tree analysis used previously in our analysis of gene expression patterns predictive of breast cancer outcomes. The initial step filtered out genes whose maximum expression did not exceed the median value of expression or did not vary more than 1-fold across the samples to remove genes with extremely low levels.
TABLE 1. Analysis of Disease Burden: Demographics and Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Minimally Diseased</th>
<th>Severely Diseased</th>
<th>P Value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, %</td>
<td>50</td>
<td>50</td>
<td>NS</td>
</tr>
<tr>
<td>Age, mean (range)</td>
<td>38.2 (20–62)</td>
<td>55.6 (39–64)</td>
<td>&lt;.00005</td>
</tr>
<tr>
<td>Sudan IV, * mean±sem</td>
<td>1.0%±0.0%</td>
<td>19.3%±5.0%</td>
<td>&lt;.00005</td>
</tr>
<tr>
<td>Raised lesion,* mean±sem</td>
<td>0.0%±0.0%</td>
<td>42.4%±7.5%</td>
<td>&lt;.0000003</td>
</tr>
</tbody>
</table>

*Percent of the total aorta affected. †Student t test.

of expression or little variance. After the filter was applied, 7470 of 12 563 total genes remained in the analysis.

Next, we clustered the genes into groups based on their expression patterns with the notion that related genes share similar variances in expression using K-means clustering. This algorithm randomly places genes into a predetermined number of groups. The genes are then shuffled among the groups in an iterative fashion to maximize the distinction between each group. The number of designated clusters was also then varied iteratively to further maximize differences between the clusters. The resulting clusters contained anywhere from 20 to 50 genes and represented a unique gene expression pattern.

Single value decomposition was performed on each cluster to generate a single factor, called a metagene. The metagene is the dominant expression pattern of a cluster and represents a group of genes that share a common gene expression signature in the context of a particular experimental condition. The metagenes are then used in binary decision trees to partition the samples into subgroups. In the trees, a metagene is used at a branch point to partition samples to 1 of 2 classifications based on similarity or dissimilarity of a sample’s gene expression pattern to the metagene. Each tree had several of these branches, and hundreds of trees were generated to determine the metagenes that did the best job of partitioning the samples. Within each metagene, we then identified the genes that lend the most weight to the dominant expression pattern.

To guard against over-fitting given the disproportionate number of variables to samples, we performed honest, out-of-sample cross validation analysis to test the stability and predictive capability of our model. Each aorta section was left out of the data set one at a time. The model was refitted (both the metagene factors and the partitions used) using the remaining samples, and the phenotype of the held out case was then predicted and the certainty of the classification was calculated.

Gene Annotation
Candidate gene annotation was performed using the Duke Integrated Genomics database (https://dig.cgt.duke.edu).

Results
Assessment of Atherosclerosis Burden in Aorta Samples
Two analyses were performed: a comparison of minimally versus severely diseased sections and a comparison of proximal and distal sections of the thoracic aorta. The first analysis identified gene signatures associated with atherosclerotic severity. The 2 groups of aorta sections used were significantly different. There was significantly less Sudan IV staining (1.0%±0.0% versus 19.3%±5.0%) and raised lesions (0.0%±0.0% versus 42.4%±7.5%) in the minimally diseased samples (Table 1). The second analysis identified gene signatures associated with the proximal and distal locations within the thoracic aorta as a possible metric of atherosclerotic susceptibility. There were no significant differences in the characteristics of the 2 locations in either Sudan IV staining (12.4%±3.0% versus 12.6%±3.1%) or raised lesions (10.1%±4.7% versus 13.2%±5.3%). There was also no significant difference in the gender or ages of the donor pools (Table 2).

 Metagene Patterns Predictive of Extent of Atherosclerotic Lesions
Figure 2 displays the results from the analysis of disease severity where the predictive model correctly classifies 93.5% (29 of 31 sections) of the sections as minimally or severely diseased based solely on their gene expression profiles. The figure shows results of the hold-one-out cross validation analysis where we constructed the model from 30 samples and used it to predict the phenotype of the 31st sample. The plot represents the probability that the unknown sample is severely diseased. The red numbers represent the severely diseased section with 95% CIs; the blue numbers represent minimally diseased samples.

The gene prioritization process identified a set of 208 genes whose expression patterns provide the power to discriminate and predict disease states in our aorta samples (Table I, available online at http://atvb.ahajournals.org). These genes encode proteins previously suspected to play a role in atherosclerosis including apoE, osteopontin, and the oxidized LDL receptor 1 (olr1). We performed a query
against gene ontology databases to determine the important biological processes represented in the analysis. We found that the genes reflected processes that we would infer from our current understanding of atherosclerosis, such as cell cycle regulation and inflammatory response (Table II, available online at http://atvb.ahajournals.org). Genes in these categories without direct links to atherosclerosis could be novel candidates for study. Such genes include capg, gm2 ganglioside activator protein, matrix metalloproteinase (MMP) 9 (mmp9), and chemokine (C-C motif) receptor-like 2 (ccrl2).

### Metagene Patterns Predictive of Susceptibility to Atherosclerosis

In our second analysis, we were able to predict the location of a sample within the thoracic aorta with 93.6% accuracy (59 of 63 sections correctly classified). The location may be a surrogate for disease susceptibility. Figure 3 is a plot of the hold-one-out cross validation analysis that shows the probability that an unknown sample is from the distal aorta with 95% CIs. The red numbers represent samples from the distal location, and the blue numbers are from the proximal aorta. Figure 4 shows expression levels by color display of the genes in the key predictive metagene and illustrates the differential expression patterns between proximal and distal tissues.

Twenty-eight genes were identified that provided the predictive power in the analysis (Table III, available online at http://atvb.ahajournals.org). Some of the genes identified in this analysis, such as superoxide dismutase 3 (sod3) and protein C receptor (procr), have previously been associated with atherosclerosis. Interestingly, many genes that populate the dominant metagene have cellular roles that could be associated with atherosclerotic disease initiation, such as homeobox-containing genes and gata2. An analysis of the biological processes represented by our gene list showed a preponderance of candidates relevant to regulation of transcription and signal transduction within our short list (Table IV, available online at http://atvb.ahajournals.org).

### Discussion

We have described a novel and unbiased approach for developing detailed genomic phenotypes that are exquisitely predictive for clinical disease states in human aorta samples. Importantly, the analysis is driven by the ability to predict phenotypic characteristics, rather than merely identifying genes whose expression differs by an arbitrary amount for a given clinical phenotype. We identified patterns of gene expression involving multiple interacting genes that define the association with the clinical phenotypes. Most importantly, our analytic approach is rigorous in assessing the validity of these patterns through cross-validation analyses. This establishes the robust conclusion that these patterns of gene expression have the capacity to accurately predict the clinical phenotype, and hence underpins the inferred validity of the connection between these genes and the phenotype. The method described could serve as a model for studying other complex cardiovascular diseases where target tissues are genomically phenotyped to identify high-priority candidate genes. These genes are now the focus of SNP studies and are the basis for studying clinical patient populations.

The statistical methodology described in this report has been successfully applied to identify gene expression patterns capable of discriminating breast cancer samples on the basis of estrogen receptor status and predicting both recurrence and extent of lymph node metastasis in primary breast cancer.9,10 The use of metagenes in our statistical approach places the emphasis on the differential coexpression of multiple genes acting in concert, consonant with the biological model of complex diseases. The decision tree models used in this study showed considerable predictive value using the metagene patterns identified. In the cross validation analysis, we were able to correctly classify unknown samples with 93.5%...
accuracy for extent of disease severity and 93.6% accuracy for aortic location. Even with the heterogeneous nature of human aorta tissues and of atherosclerotic lesions therein, the metagenes identified through this approach clearly exhibit reliable discriminatory patterns. One point in the disease burden analysis is that we used sections from 2 different locations within multiple aortas. Because we did not use a single location, the inherent heterogeneity within an individual aorta could have influenced our results. However, the genes that predicted aortic location in our second analysis were quite different from those in our first analysis, suggesting that location was not a major factor in the burden analysis.

Our disease burden analysis identified genes associated with minimal versus severe atherosclerosis as measured, primarily, by the extent of advanced disease plaques. The samples included in the analysis were classified by quantitative measures of atherosclerosis, identifying genes that reflect extent of disease. Although our analysis did identify genes whose expression is limited in the tissue (ie, endothelium-derived), it may be biased toward genes expressed in the more abundant media, especially the inner media, which may be most interactive with the atheromas. Still, we intentionally used the full thickness of the vessel wall in this initial analysis, as the disease process likely involves complex interactions across the tissue layers through autocrine and other mechanisms. The expression patterns identified, therefore, may represent the net gene expression effect contributed from the different tissue layers.

The second analysis identified genes associated with location within the aortas. The regional differences within an aorta could be related to disease susceptibility, given results from the PDAY Study that showed atherosclerotic development progresses from the distal to the proximal aorta. In our analysis of the proximal and distal sections of the human aortas, there was no significant difference in the age of the donor patients and the atherosclerotic burden of the proximal and distal sections used by Sudan IV and raised lesion scores. This would seem to indicate that the differential gene expression patterns we found were indicative of location rather than age or presence of disease. This may reflect inherent differences in the mechanical stress from hemodynamic factors experienced by the different locations or possibly differences in the clonal origin of the cells populating the proximal or distal areas.

Our first analysis for extent of disease identified genes reproducibly associated with atherosclerosis in the literature, such as apoE, osteopontin, and olr1. As one would expect from a biological context, these genes were upregulated in the diseased sections of the aorta. Importantly for our study, we also identified genes such as capg, gm2, mmp9, and ccr12 that encode proteins whose function is consistent with a role in atherogenesis but have not been previously linked to atherosclerosis. CapG is a key regulatory protein for actin and membrane phospholipids within migrating phagocytes. The gene gm2 has been well studied in the oncology field for its role in cell proliferation, adhesion, and chemotaxis. In addition, it has been shown to be present in aorta of apoE knock-out mice. The proteins MMP1, MMP2, MMP9, and MMP13 have all been associated with atheroma progression, presumably through vascular remodeling. The gene product of ccr12 serves as a receptor for monocyte chemotactic protein 1 (MCP1) and other chemokines and may be important for processes, such as vascular infiltration by monocytes and intimal hyperplasia. An ontologic analysis of the gene list confirmed much of what has already been reported regarding the processes involved in atherosclerosis. However, the number of genes related to cell growth/motility, transcriptional regulation, and signal transduction seems to indicate that even in the advanced stages, the disease is active and proliferating. This would suggest that molecular targeted interventions could still be of some benefit to this population of patients.

The second analysis looked for genes related to disease susceptibility. As before, there were genes whose translated products are directly associated with atherosclerosis, such as sod3 and procr. In addition, the analysis identified genes whose function has been described in other human disorders, but are involved in inflammatory, growth signaling, and cell–cell communication pathways that would be important for early atherosclerosis initiation and progression. Key examples are a number of homeobox-containing genes that have been found in vascular smooth muscle and endothelial cells and have been linked to cellular proliferation, migration, and differentiation as well as vascular remodeling. Another gene, gata2, is a transcription factor that regulates endothelin-1 production in endothelial cells. Endothelin-1 is a vasoactive peptide that has been highly associated with atherosclerosis. An intriguing observation is the great percentage of genes that are related to regulation of transcription and signal transduction without a great number of genes related to inflammation. One explanation could be that the genes identified by our analysis could be upstream effectors that modulate the inflammatory pathways. For example, the identification of the homeobox genes and gata2 is intriguing given the potential role in modulation of vascular cell adhesion molecule-1 and intercellular adhesion molecule-1, two proteins linked to atherosclerotic prelesion and early lesion formation. Homeobox genes, particularly the C class, have been associated with the increased and decreased expression of intercellular adhesion molecule-1. The gene gata2 has been shown to mediate vascular cell adhesion molecule induction in response to thrombin, estrogen, and glucocorticoids.

The identification of genes that predict atherosclerotic phenotypes, whether they are genes already known to function in the disease process or novel genes not previously linked to the disease, represents an important initial step toward an improved understanding of the disease. Our method as described is not meant initially to serve as a diagnostic test, but rather as a means to prioritize genes and allow us to focus our research efforts for identification of SNPs for large-scale analyses of gene variants. As well, identification of important genes advances our understanding of the biological pathways relevant to atherosclerosis. Importantly, our analytical approach may identify not only the initial steps in biological pathways but the secondary and tertiary events as well. As such, the analysis provides a much richer data set than merely identifying the immediate effec-
tors of a process. Many of the genes we have identified are likely to be causative and may be relevant to future therapeutic interventions. Furthermore, by finding the polymorphisms within these high priority genes, we may begin to identify combinations of SNPs that, when taken in concert with clinical cardiovascular risk factors, may lead to the development of new diagnostic and prognostic tools for cardiovascular events. Clearly, a next step must be to develop a deeper understanding of the biological pathways implicated by these analyses and to begin the process of investigating the role of these genes in the development of vascular disease.

**Acknowledgments**

This study was supported in part by National Institutes of Health program grant HL073042-02 through the National Heart, Lung, and Blood Institute.

**References**