Prediction of Cell Type-Specific Gene Modules: Identification and Initial Characterization of a Core Set of Smooth Muscle-Specific Genes

Sven Nelander, Petter Mostad and Per Lindahl

Genome Res. 2003 13: 1838-1854
Access the most recent version at doi:10.1101/gr.1197303

Letter

Prediction of Cell Type-Specific Gene Modules: Identification and Initial Characterization of a Core Set of Smooth Muscle-Specific Genes

Sven Nelander,[1] Petter Mostad,[2] and Per Lindahl[1,3]

[1]*Department of Medical Biochemistry, Göteborg University, SE 405 30 Gothenburg, Sweden;* [2]*Department of Mathematical Statistics, Chalmers University of Technology, SE 412 96 Gothenburg, Sweden*

Genes that are expressed in the same subset of cells potentially constitute a module regulated by shared *cis*-regulatory elements and a distinct set of transcription factors. Identifying such units is an important entry point to the molecular study of cell differentiation. We developed a general method to classify cell type-specific genes from expressed sequence tag (EST) data, and we optimized it for identification of smooth muscle cell (SMC)-specific genes. Expression profiles were derived from the quantitative distribution of EST data in mouse, and genes were classified based on their profile similarity to known reference genes, in this case smooth muscle myosin heavy chain. A large majority (>90%) of known SMC-specific genes were identified, together with novel candidates. Extensive experimental validation confirmed SMC-specific expression of candidates, for example, lipoma preferred partner (LPP) and a novel SMC-specific putative monoamine oxidase, SMAO. Our method performed considerably better than other computational methods in an objective cross validation comparison. The total number of SMC-specific genes is estimated to be ~50.

[Supplemental material is available online at www.genome.org. A program package, uni_extract, for extraction of data and data preparation, and a MATLAB program package, QRISP, for data transformation, probability estimation, cross validation, and visualization of data, is available at http://cbz.gu.se/Lindahl/QRISP. A gene expression pattern prediction server will be available at www.qrisp.com.]

A principal aim of modern developmental biology is to understand the differentiation of cell lineages in molecular detail. Regulation of cell type-specific genes is frequently used as a molecular model of cell differentiation. From a systems perspective, it is therefore interesting to identify complete sets of cell type-specific genes, and ask whether the same or different pathways regulate the correlated expression of such sets. The hierarchy of transcription factors that control expression of contractile proteins in skeletal muscle serves as one of the better understood molecular prototypes of mammalian cell lineage determination (Pownall et al. 2002), and comparative analyses of *cis*-regulatory elements indicate a common mechanism for several marker genes (Konig et al. 2002; Wasserman et al. 2000). Less is known about smooth muscle cells (SMCs). There has been considerable progress in mapping *cis*-regulatory elements required for expression of SMC-specific proteins (Mack et al. 2000; Chang et al. 2001; Manabe and Owens 2001a,b; Strobeck et al. 2001), but the presumed SMC-specific transcription factors that bind these elements have not been identified. Shared *cis*-regulatory elements (CArG boxes) in SMC marker promoters, in combination with correlated developmental onset of gene expression of some SMC markers, argue that there may be a modular regulation of the SMC genes. Still, the number of known SMC-specific genes is rather small, which limits the scope of molecular investigation, and SMC marker gene regulation has been studied from a single-gene perspective rather than by systematic side-by-side comparison. The present project was initiated to identify a near complete set of SMC-specific genes to serve as an entry point for a system-oriented analysis of SMC-specific transcriptional regulation.

To identify SMC-specific genes is technically challenging. The broad tissue representation and the tight association between SMCs and other cell types make experimental screening difficult. SMC samples must be compared with a wide range of tissues, preferentially free of vasculature, to claim cell type specificity. In vitro experiments are of limited relevance because the cells are plastic and adopt a fibroblast appearance when cultured (Schwartz et al. 2000). The present work explores the potential of computational identification of SMC-specific genes from expressed sequence tag (EST) data as an alternative to experimental approaches.

Bioinformatic cell type specificity screens have been reported for a range of cells such as endothelial markers, colon cancer cells, and cardiac muscle (Wang et al. 2001; De Young et al. 2002; Huminiecki et al. 2002). Existing protocols for relating EST detection to cell type-specific expression of genes, such as keyword searches (Schuler 1997), and Digital Differential Display (Scheurle et al. 2000), rely on a well-characterized or obvious relationship between the expression pattern of interest and the EST libraries under study: Heart-specific genes are present in heart-derived libraries and absent from others, et cetera. The relationship between gene expression and library representation is less obvious in the SMC case, and more sophisticated analysis tools are therefore required.

An alternative approach for classification of EST data is to perform a multivariate analysis of expression profiles (a set of measurements of a gene's transcript abundance across a range of biological situations). Ewing and co-workers (Ewing et al. 1999; Ewing and Claverie 2000) have demonstrated the potential of unsupervised clustering of quantitative EST data in their analysis

[3]**Corresponding author.**
E-MAIL Per.Lindahl@medkem.gu.se; FAX 46-31-416108.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1197303. Article published online before print in July 2003.

13:1838–1854 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

of the rice genome. We have optimized this approach and introduced a classification procedure that allows cell type-specific genes (e.g., SMC markers) to be discriminated from nonspecific genes. The key step in the optimization is an objective cross validation of alternative feature extraction methods (alternative ways to construct expression profiles) and alternative distance metrics (alternative ways to measure expression profile similarity).

The optimized procedure identified SMC-specific genes from mouse EST data with 99% specificity at 94% sensitivity. Extensive validation by in situ hybridization (ISH) confirmed that predictions of novel SMC marker genes were correct. Furthermore, the classification procedure outlined a module of ~50 SMC-specific genes in the genome, also in agreement with the ISH data.

In conclusion, the present work conceptually links profile-based EST analysis with classification/supervised learning and demonstrates its combined use in detecting modules of genes that are expressed in specific cell types with complex tissue representation. Classification, as opposed to cluster analysis, results in testable hypotheses on cell type specificity, which can be experimentally validated in a nonambiguous manner at success rates that can be estimated in advance.

RESULTS

Quantitative Expression Profiles Generated From Public EST Data Discriminate SMC-Specific Genes

The mouse UniGene data contain 2.7 million ESTs from 639 different libraries (Fig. 1), and thus constitute an enormous source of transcript distribution information. SMCs are represented in practically all tissues, but in substantially different proportions. We therefore asked whether the differential presence of SMCs between organs is sufficient to affect the library distribution of ESTs, and to generate an SMC-specific signature.

Expression profiles were constructed from the distribution of ESTs for each gene across the 639 libraries. The profiles of two known SMC markers, smooth muscle myosin heavy chain (SM-MHC) and leiomodin 1, showed a high degree of correlation, indicating that profiles for SMC-specific genes might have characteristic features (Fig. 2). The correlation was primarily due to abundant transcript detection in urinary bladder and colon libraries. However, the degree of correlation between these two markers was clearly dependent on the type of expression profile that was constructed from the EST data (definitions of different data construction principles are given in Methods). Profiles that were derived from raw data (number of detections/library) and from transformed data (statistic estimate of gene representation/library) correlated very well. Profiles derived from frequency data (number of detections/library, normalized for library size) and binary data (presence/absence of gene/library, data not shown) did not show convincing correlation.

Detection of SMC-Specific Gene Expression Can Be Defined as a Classification/Supervised Learning Problem

The pair-wise correlation between SMC marker profiles implied that SMC-specific genes might be detected by a classification (supervised learning) method. In such methods, user-provided examples ("training data") are used to predict the properties of novel objects. In this case, existing examples of known SMC markers, and examples of genes known *not* to be SMC markers can be used to estimate the probability for an undocumented gene to be SMC specific. We therefore developed a classifier that was optimized with respect to SMC marker identification. The classification has three components: (1) derivation of expression profiles from EST data, (2) computation of profile distances to a reference marker, and (3) estimation of probabilities (based on training data) for genes to have the same expression pattern as the reference marker.

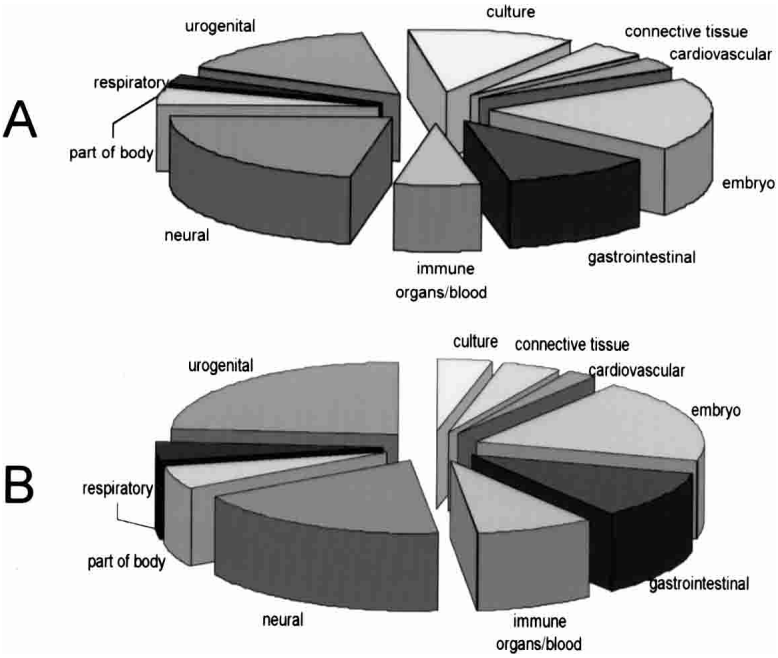

Figure 1 Tissue sources represented in UniGene data. Based on tissue source information for the 639 mouse EST libraries, libraries were grouped into different organ systems (*A*). Embryo, whole embryo or part of embryo regardless of stage; part of body, combinations of adult organs or body parts; culture, cell culture, including stem cell libraries. Based on this grouping, the relative contribution from each tissue to the 2.7 million ESTs in the data was calculated (*B*).

Systematic Evaluation of Alternative Classification Methods Established an Optimal Protocol for SMC Marker Prediction

To maximize sensitivity and specificity of discrimination, alternative technical parameters were systematically compared: The two parameters with the greatest impact on classification performance were (1) the type of profile and (2) the choice of distance metric.

The different profile types (raw, transformed, frequency, and binary) and the different distance metrics (permutation testing, Pearson's correlation coefficient, covariance, Fischer's test) are defined in Methods. In an informal evaluation, we ranked genes according to profile similarity to the SM-MHC gene and registered the ranks of 10 positive controls using different methods. This revealed a striking dependency on technical parameters (Fig. 3A). In a formal evaluation, we again used SM-MHC as the reference marker and estimated the probabilities for genes to be SMC specific using logistic regression fitted by unbiased training data as defined in Methods. This returned the estimated probability for different genes to be SMC specific, on the basis of alternative methods. In an optimal classification, the estimated probability for negative controls should be 0, and the estimated probability for positive controls (SMC mark-

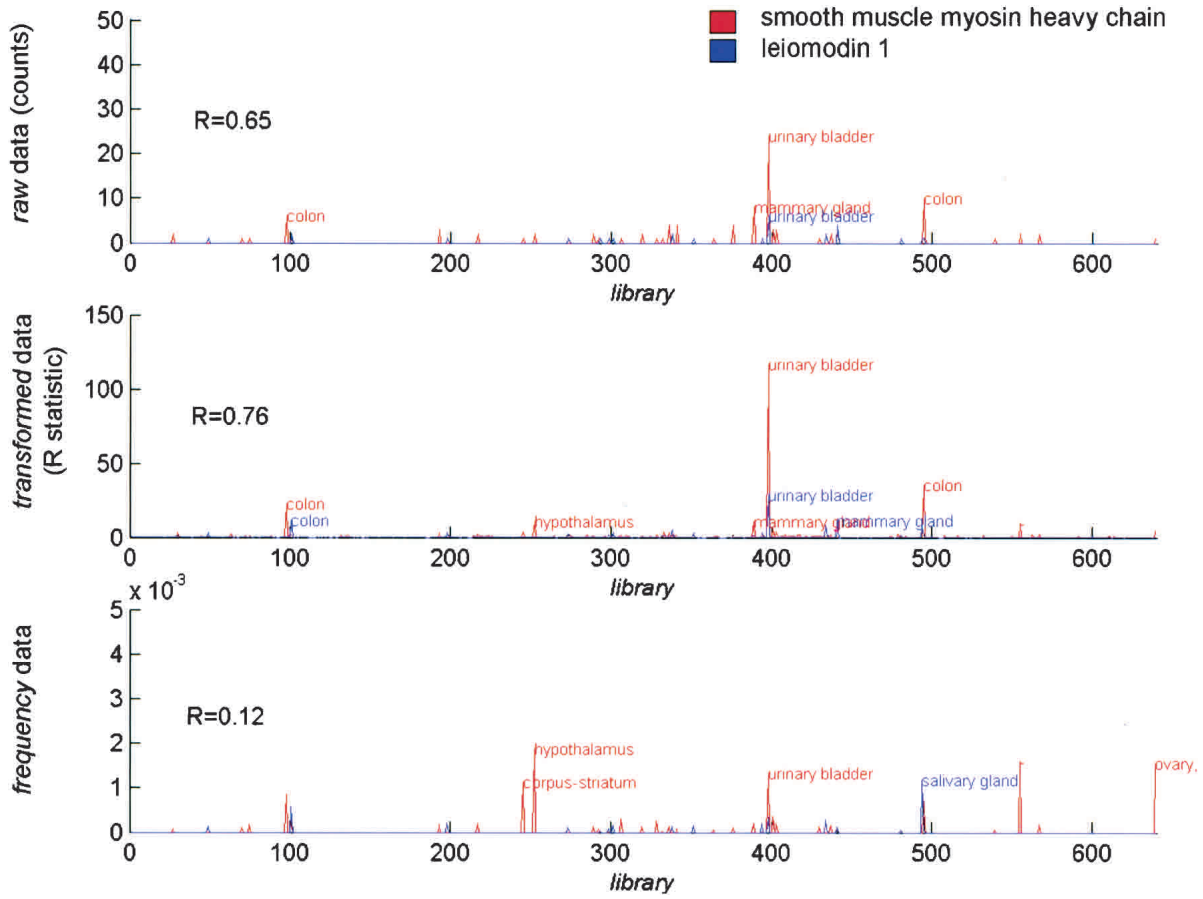

Figure 2 Examples of EST-derived expression profiles for two SMC-specific genes: *SM-MHC* (103 detections in UniGene, red curve) and *leiomodin 1* (28 detections in UniGene, blue curve). *X*-axis, The 639 mouse EST libraries in arbitrary order; *Y*-axis, EST-derived expression profiles according to definitions in Methods. For the *raw* and *transformed* profiles, the two SMC-specific genes have related profiles (correlation 0.65 and 0.76, respectively). Transcripts are primarily detected in certain urinary bladder and colon libraries. The two genes do not correlate in the frequency data, which also contain signals from unexpected organs such as the hypothalamus.

ers that were not included in fitting of the regression curve) should be 1.

Raw data and Pearson's correlation performed best, estimating the average SMC marker to 37% probability and the average nonmarker to a 1% probability (Fig. 3B,C). The use of binary profiles and the Fischer metric correspond to the hitherto best performing established method, GBA (Walker et al. 1999). However, this method only had an 8% probability for the average SMC marker and a 1% probability for nonmarkers, demonstrating that GBA should not be applied to SMC marker prediction. The informal evaluation (Fig. 3A) showed that, when using GBA, the top 3100 genes would need to be screened in order to detect ~90% of the known SMC-specific genes, compared to the top 280 genes when using raw data and Pearson's correlation. The formal sensitivity/specificity relationship for raw data and Pearson's correlation on a test set was 99% specificity at 94% sensitivity, and this method was used in all subsequent experiments. Predicted genes and probability estimates are shown in Table 1.

In Situ Hybridization Confirmed SMC-Specific Expression of Candidate Genes and Revealed Systematic Patterns of Ectopic Expression

Forty-six genes with estimated probabilities between 1% and 77% were selected for further validation using ISH. The selection was fairly randomized with a slight bias toward abundant genes, genes with "interesting features," and previously undocumented genes. The established SMC markers SM-MHC, α-smooth muscle actin (ASMA), vinculin, and smoothelin were included to provide reference data. ISH was performed on mouse late embryonic (E17.5) sagittal sections, which allows for a wide range of organs to be screened on one slide. Embryonic expression is also relevant to validate a potential role in SMC development. ISH results were informative (presence of signal and low background) for 29 genes including the four controls. The result was used to classify the genes into three groups:

1. SMC marker group (four controls and five candidates). This group contained genes with an expression pattern highly concordant with expression of the SMC marker controls (Figs. 4, 5; Table 2). The SMC marker group displayed reoccurring patterns of ectopic expression in specific locations (Fig. 6).
2. SMC marker-related group (five candidates). This group contained genes with an SMC marker-concordant expression pattern but with additional atypical ectopic expression in a limited set of other locations (Fig. 7).
3. Genes with other expression patterns.

None of the known markers were strictly confined to SMCs but were to variable extents also expressed in cardiac and skeletal muscle (Figs. 4, 6). The markers were further expressed in some nonmuscle cells (SM-MHC, ASMA, and smoothelin stained sub-

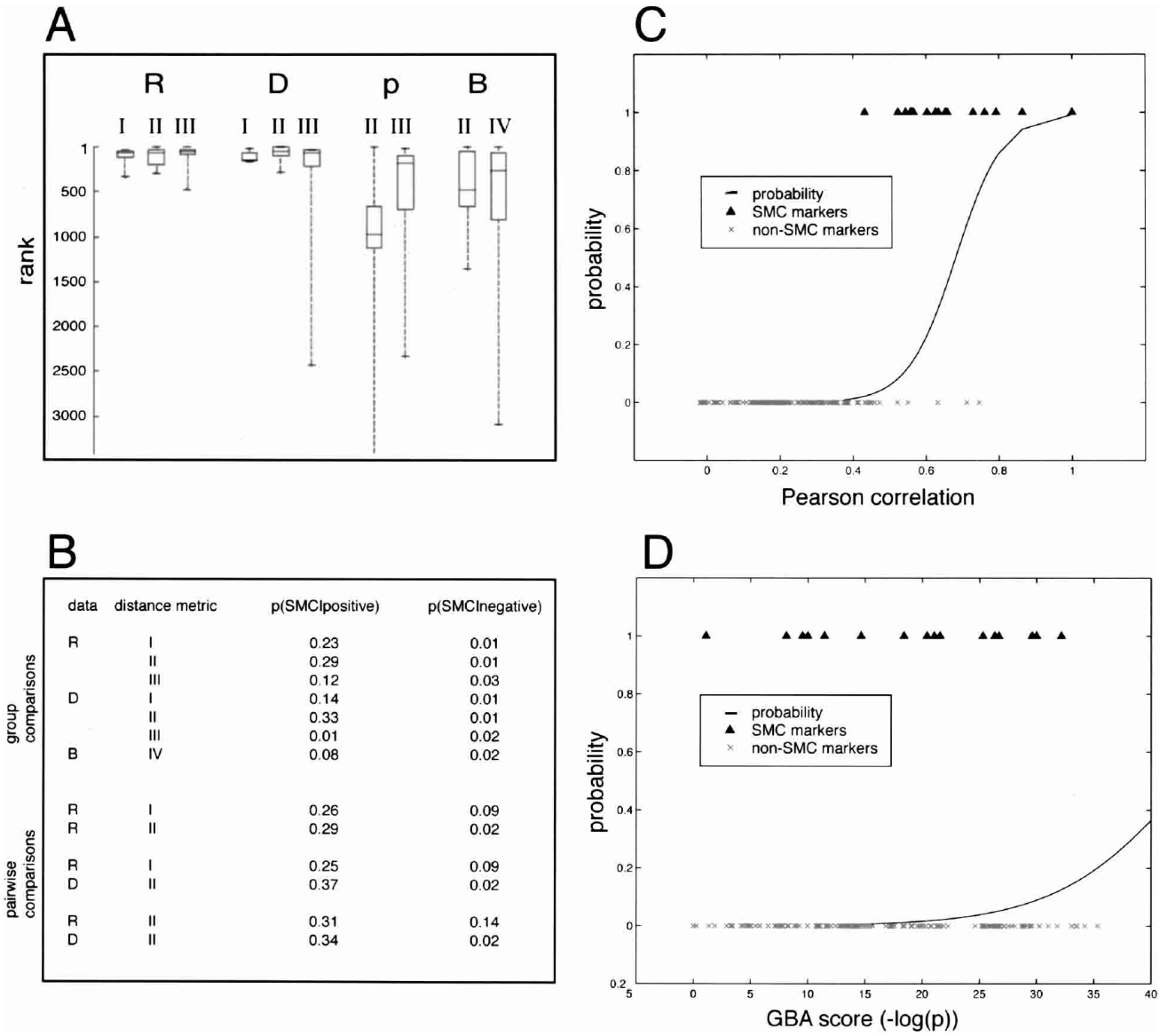

Figure 3 Systematic evaluation of alternative data types and distance metrics with respect to prediction of SMC-specific genes. R, transformed data; D, raw data; p, frequency data; B, binary data. Roman numerals indicate different distance metrics. I, Permutation; II, Pearson's correlation; III, covariance; IV, GBA (Walker et al. 1999). Euclidean distance was also evaluated, giving nonspecific results (not shown). For full definition of terms, see Methods. (*A*) Nearest neighbor searches were performed on all genes with at least five ESTs in UniGene (n = 29,812). Genes were ranked according to their profile similarity to SM-MHC. Ranks for 10 SMC markers are displayed on the *Y*-axis as box blots. Boxes, Central 20th to 80th percentiles; whiskers, the full range of observations. (*B*) Alternative methods (combinations of data type and distance metric) were evaluated with a logistic regression model. This model was used to compute probabilities for genes to be SMC markers based on their profile distance to SM-MHC. P(SMC|positive) denotes the model average probabilities for positive controls to be SMC markers, and (SMC|negative) denotes the corresponding probability for negative controls. Models were first compared group-wise using the full test set. This identified three preferred models (R-I, R-II, and D-II). These three models were then compared in a pair-wise fashion. To avoid bias, subsets of the reference gene set were used (see Methods). The preferred method was data = D, distance = II (Pearson's correlation), with expectations 0.37 for the markers and 0.02 for the nonmarkers. (*C*, *D*) Logistic regression curves for raw data/Pearson's correlation (*C*) and GBA (*D*). Triangles, Known SMC markers (probability = 1); crosses, genes with other expression pattern (probability = 0); curve, estimated relationship between a gene's correlation against SM-MHC and its probability to be an SMC marker.

mandibular gland epithelium; Fig. 5), ASMA and smoothelin were expressed in hair follicle dermal sheets, and ASMA finally was expressed in portions of the central nervous system (CNS). The CNS expression was not confined to vascular SMC.

The candidate genes were similarly expressed in all muscle lineages and in some additional cell types (Figs. 4, 5, 6; Table 2). Genes that recapitulated the SMC marker typical pattern of ectopic expression in submandibular glands, hair follicle dermal sheaths, and regions in the CNS were classified as novel SMC markers (Fig. 5; Table 2). Genes with additional expression in atypical ectopic sites, for example, brown fat, bladder epithelium, and pancreas epithelium, were classified as SMC marker related (Fig. 7; Table 2).

Candidate genes that were classified to group (3) displayed the full spectrum of patterns from borderline cases of SMC-specific expression to ubiquitous expression. Approximately half

Table 1. Prediction of Smooth Muscle-Specific Genes

Expression	Correlation	Probability	UniGene ID	Annotation
105	1.00	0.99	Mm.3153	Myosin heavy chain 11, smooth muscle
174	0.86	0.94	Mm.16562	Actin gamma 2, smooth muscle, enteric
196	0.79	0.85	Mm.2006	Transgelin/SM-22 alpha
49	0.79	0.84	Mm.43921	RIKEN cDNA 4631426E05 gene
68	0.78	0.83	Mm.99349	*Svs7*
33	0.77	0.80	Mm.95705	Gamma-2 syntrophin
58	0.77	0.78	Mm.3089	ESTs
91	0.76	0.77	Mm.31259	SMAO
31	0.76	0.77	Mm.99224	ESTs
24	0.75	0.75	Mm.46214	*Uroplakin 3*
57	0.74	0.70	Mm.45481	RIKEN cDNA 1110032A04 gene
292	0.73	0.68	Mm.6712	Desmin
26	0.72	0.66	Mm.4351	Arginine vasopressin receptor 1A
26	0.70	0.58	Mm.100108	RIKEN cDNA 5730414M22 gene
31	0.68	0.50	Mm.152763	*Svs3*
32	0.66	0.43	Mm.3787	*Svs6*
344	0.66	0.42	Mm.16537	Actin, alpha 2, smooth muscle, aorta
31	0.65	0.40	Mm.57221	Homeobox B13
27	0.65	0.40	Mm.180124	Leiomodin 1 (smooth muscle)
41	0.64	0.34	Mm.86380	ESTs
29	0.63	0.33	Mm.34359	ESTs
26	0.63	0.32	Mm.89958	MRV integration site 1
54	0.63	0.30	Mm.25722	Purinergic receptor P2X
76	0.62	0.27	Mm.17807	ESTs
22	0.61	0.25	Mm.30933	ESTs
99	0.60	0.24	Mm.4356	Calponin 1
30	0.60	0.23	Mm.88548	ESTs
71	0.59	0.21	Mm.24684	ESTs
219	0.59	0.19	Mm.10661	Solute carrier family 2 member 4
222	0.58	0.19	Mm.34060	RIKEN cDNA 5830475I06 gene
55	0.58	0.17	Mm.45019	Cancer related gene-liver 1
286	0.58	0.17	Mm.28278	Caveolin, caveolae protein, 22 kD
50	0.57	0.17	Mm.27165	ESTs
134	0.57	0.16	Mm.28649	ESTs
189	0.57	0.16	Mm.38877	ESTs
60	0.57	0.15	Mm.23983	ESTs
81	0.57	0.15	Mm.34171	ESTs
43	0.57	0.15	Mm.83243	ESTs
49	0.57	0.15	Mm.12966	Plasma membrane associated protein, S3–12
284	0.57	0.15	Mm.205997	LPP
34	0.57	0.15	Mm.34461	ESTs
139	0.56	0.15	Mm.22941	RIKEN cDNA 4930435F02 gene
307	0.56	0.14	Mm.36850	Smoothelin
36	0.56	0.14	Mm.85304	Gob-5
147	0.56	0.14	Mm.7342	PDZ and LIM domain 3
51	0.56	0.13	Mm.314	ESTs
31	0.56	0.13	Mm.74658	ESTs
76	0.56	0.13	Mm.119265	ESTs
466	0.55	0.12	Mm.1987	*Decorin*
26	0.55	0.12	Mm.55337	ESTs
44	0.55	0.12	Mm.87506	ESTs
458	0.54	0.11	Mm.196484	Cysteine rich protein 1
94	0.54	0.11	Mm.1237	Peripheral myelin protein, 22 kD
51	0.54	0.11	Mm.35290	Expressed sequence AI552420
111	0.54	0.11	Mm.136207	Expressed sequence AA986860
66	0.54	0.10	Mm.183030	RIKEN cDNA 5430429D21 gene
52	0.54	0.10	Mm.44235	HNF-3/forkhead homolog 1 like
65	0.53	0.10	Mm.29953	RIKEN cDNA 5430437E11 gene
30	0.53	0.09	Mm.38127	Expressed sequence AI642000
118	0.53	0.09	Mm.27390	ESTs
22	0.52	0.08	Mm.112236	ESTs
253	0.52	0.08	Mm.36769	SLAP
58	0.52	0.08	Mm.31552	ESTs
49	0.52	0.08	Mm.23873	ESTs
…	…	…	… [88 genes]	
…	…	…	…	
39	0.43	0.02	Mm.87553	Protocadherin beta 17

(continued)

of these genes showed some degree of SMC selectivity (data not shown). Three genes, finally, displayed restricted, but muscle unrelated, patterns.

The number of genes that experimentally qualified as being SMC markers was in agreement with the prediction

The Degree of Ectopic Expression of Candidate Genes Is Recapitulated in the Profile's Distance to SM-MHC

Is there a correlation between the degree of SMC-specific expression and the distance to SM-MHC? ISH results were visualized by plotting SM-MHC profile similarity on the *Y*-axis versus gene abundance on the *X*-axis (Fig. 8). Gene abundance was defined as the logarithm of the total number of EST detections for a gene in UniGene. The ISH results revealed a logic relationship between the EST profile similarity to SM-MHC and the degree of SMC specificity. The expression in nonmuscle cells gradually increased with decreasing mathematical correlation to SM-MHC. The SMC marker group (1) had correlation ranging from 0.43–0.86, average 0.61. The SMC marker-related group (2) ranged between 0.52 and 0.57, average 0.54. Finally, the heterogeneous nonselective group (3) spread between 0.37 and 0.65, average 0.49. For genes with >50 detections, the separation between SMC markers and nonmarkers occurred in a correlation interval between 0.4 and 0.6. Only two low-abundance genes were experimentally validated; both were classified in the nonselective group (3). Data from the literature-based reference set were distributed in good accordance with the ISH data. However, the distribution also revealed a number of "false positives", that is, non-SMC markers with high correlation. These were found, with rare exceptions, in the low-intensity range.

The Mouse Genome Is Estimated to Encode ~50–60 SMC Markers

Probability estimation for genes to be SMC markers allows for a global estimate of the total number of SMC-specific genes (see Methods). We estimated probabilities by logistic regression that was fitted with the combined ISH and literature data (regression curve shown in Fig. 3C). Integrating probabilities gave a model estimate of 56 genes in the mouse genome that encode SMC markers (standard error 4.97). Of these, 34 are predicted to be among the top 161 genes with a profile correlation above 0.5, of which 12 are known markers. This indi-

Table 1. *Continued*

Expression	Correlation	Probability	UniGene ID	Annotation
25	0.43	0.02	Mm.67539	RIKEN cDNA 4933406D09 gene
25	0.43	0.02	Mm.116986	Secreted and transmembrane 1
216	0.43	0.02	Mm.12842	Vinculin

Top-ranking genes based on profile similarity to SM-MHC. Smooth muscle cell markers underlined. Likely false positives (the Svs and decorin) in italics. Expression, Number of detections in UniGene; correlation, Pearson's correlation in the raw data set (see Methods); probability, logistic regression probability estimate for a gene to be an SMC marker. For brevity, genes with fewer than 20 detections have been excluded. The logistic regression estimate of the total number of SMC markers in the list shown is 20.4; standard error 3.06.

cates that a majority of the unknown markers can be identified in a systematic screen of high-correlating genes.

Quantitative EST Profiling Identified Structural and Regulatory Genes in Heart and Skeletal Muscle, Retina, and Lens, Demonstrating General Applicability

The method is stated in a general way and could theoretically be applied to any tissue or cell type. The only a priori information required is a marker gene that can be used as discriminator. Obviously, the lack of cell type-specific libraries reduces the resolution for some tissues: It is not possible to discriminate between closely associated cell types unless at least one library is enriched for one of the cells.

We profiled all genes with at least five UniGene detections (n = 29,812) against a panel of genes with tissue-specific expression patterns (see Methods and Supplementary Material, available online at www.genome.org). Examples of predictions based on skeletal muscle, heart muscle, photoreceptor, and lens markers are shown in Table 3A–D. Previously known markers for the respective cell type are consequently clustered at the top of the lists, thus confirming the accuracy. Fourteen of the top 19 genes with highest correlation to the cone-rod containing homeobox gene have a previous record of retina-specific expression. Compared with the SMC case, the Pearson's correlation coefficients in this case are higher, possibly indicating less complex expression patterns for these genes. Several well-documented key regulators of skeletal and cardiac muscle differentiation were identified among the structural genes.

DISCUSSION

Concordance Between Mathematical Predictions and Experimental Results

We have developed a new method for mathematical prediction of SMC-specific genes. Five SMC markers and five SMC marker-related genes with selective but not specific SMC expression were validated by ISH. The proportion of SMC markers versus non-markers in the experimentally evaluated gene set agrees well with the a priori estimated probabilities.

ISH offers a resolution and an organ overview that alternative methods such as Northern blot and RT-PCR cannot match. Still, classification of genes as SMC specific or not based on ISH is bound to be wrong sometimes. The screening was performed solely on embryonic sections and on the subset of organs that were simultaneously represented on single slides. Technical failures that result in artefact staining are sometimes mistaken for ubiquitous expression. We have consciously chosen to be conservative in judgments and to classify genes that have an SMC/muscle selective staining but high background staining as non-muscle, which might lead to underestimation of specificity and to underestimation of the total number of SMC markers. We have also been careful not to claim technical failure unless it is obvious. The most common type of technical failure was complete lack of signal. Genes that are SMC specific only in certain splice isoforms (such as vinculin) may give ISH results that depend on the choice of probe. Our vinculin probe consisted of a 3′ fragment of the vinculin mRNA, and should in theory not be able to differ between SMC-specific (metavinculin) and broadly expressed vinculin isoforms. Still, our ISH result for vinculin was highly SMC specific, possibly explained by a dominant contribution of metavinculin at this time point.

Data Quantity and Library Diversity Are the Main Limiting Factors of Method Performance

The method is formulated in a general way to classify genes according to coexpression with a discriminator gene. This classification was used to identify cell type-specific markers, but theoretically it can be used to predict similarity to any discriminator gene expression pattern. The potential of the method in any particular case will depend on three factors: (1) the tissue sources underlying the data in relation to the expression pattern of interest, (2) the amount and quality of the data, and (3) the analysis method itself.

Predictions can only be made for cell types present in the data bank. Similarly, morphological resolution of the method depends on the presence of tissue- or cell type-specific libraries. Clearly, if two cell types are present in amounts that have the same ratio for all libraries, then the method cannot separate between genes specific for one or the other of these cell types. The efficiency of finding genes specific for cell type C depends on in how many libraries, and to what extent, the amount of C varies in relation to other cell types. It is, however, not necessary to have libraries containing only C. The sizes, that is, the total counts, of the libraries where C has a higher proportion cannot be too small. If they are, then the stochastic noise from the random selection of the sequenced tags will overwhelm any signal of differential expression. For the same reason, the C-specific genes that are to be detected cannot have too small expression rates in C compared with the sizes of the libraries where C has a higher proportion. Thus, the method is more likely to detect highly expressed C-specific genes than lowly expressed ones.

The current representation of tissues in mouse UniGene libraries (see Fig. 1) indicates that our prediction may be biased toward efficient detection of genes with expression in visceral and urogenital SMC and less efficient detection of genes that are preferentially expressed in vascular and pulmonary SMC. This indicates that the detection of SMC subtype-specific genes may be problematic with the current data set. However, the intention in this experiment has been to characterize a core set of general SMC markers that is present in all subtypes, and not to identify genes that are specifically expressed in certain subtypes. The UniGene data are a secondary data set derived from dbEST, a constantly growing public EST database, which indicates that resolution and tissue representation will improve over time. Other data sets that can be analyzed with the method in its current design are UniGene data for other species, or SAGE data. Potentially, data from several sources could be combined to enhance the power of the method. Improvements that could be attempted

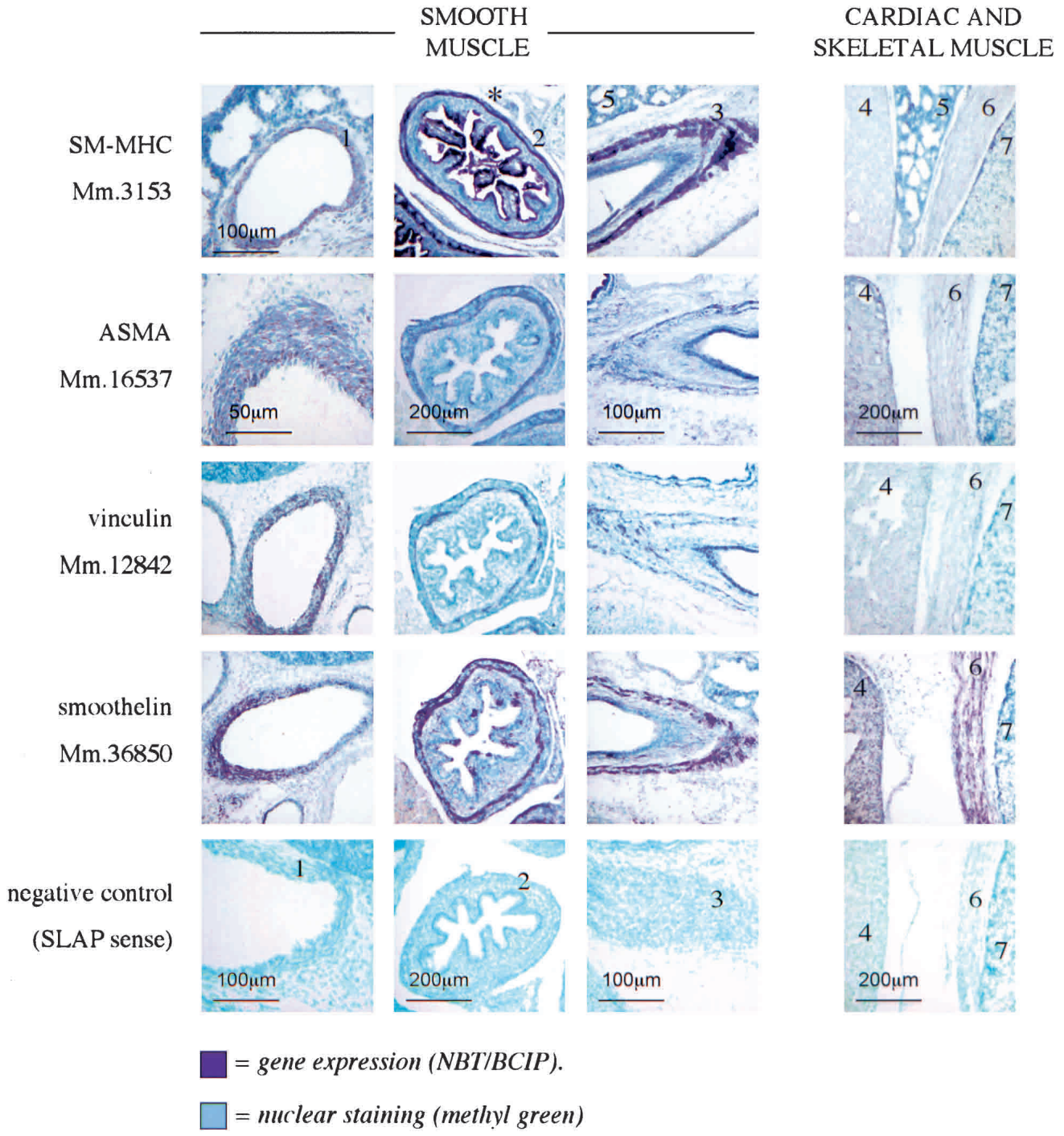

Figure 4 In situ hybridization (ISH) results for selected genes. Expression of control genes. 1, Vascular SMC; 2, intestinal SMC; 3, esophageal SMC; 4, cardiac muscle; 5, lung parenchyma; 6, diaphragm (skeletal muscle); 7, liver. A summary of ISH results is given in Table 2. They are also graphically depicted in Fig. 8. Intestinal lumen artefact staining is due to endogenous alkaline phosphatase activity (asterisk).

include mapping of SAGE data onto UniGene, addition of private-domain EST data to the public set, and cross comparing predictions versus microarray results. Further, it would be worth considering combinations of the method presented here with prediction methods based on other features, such as promoter sequences.

The libraries included are heterogeneous in terms of sampling depth and construction (sequencing direction, normalization, et cetera), which is a potential source of systematic error. The annotation is another source of error: The one-to-one correspondence between genes and the EST clusters in UniGene can be questioned.

The predictive power using different data pretreatments and distance metrics was systematically evaluated. GBA, an established method for identification of expression modules, was clearly not performing well on genes with a more complex tissue

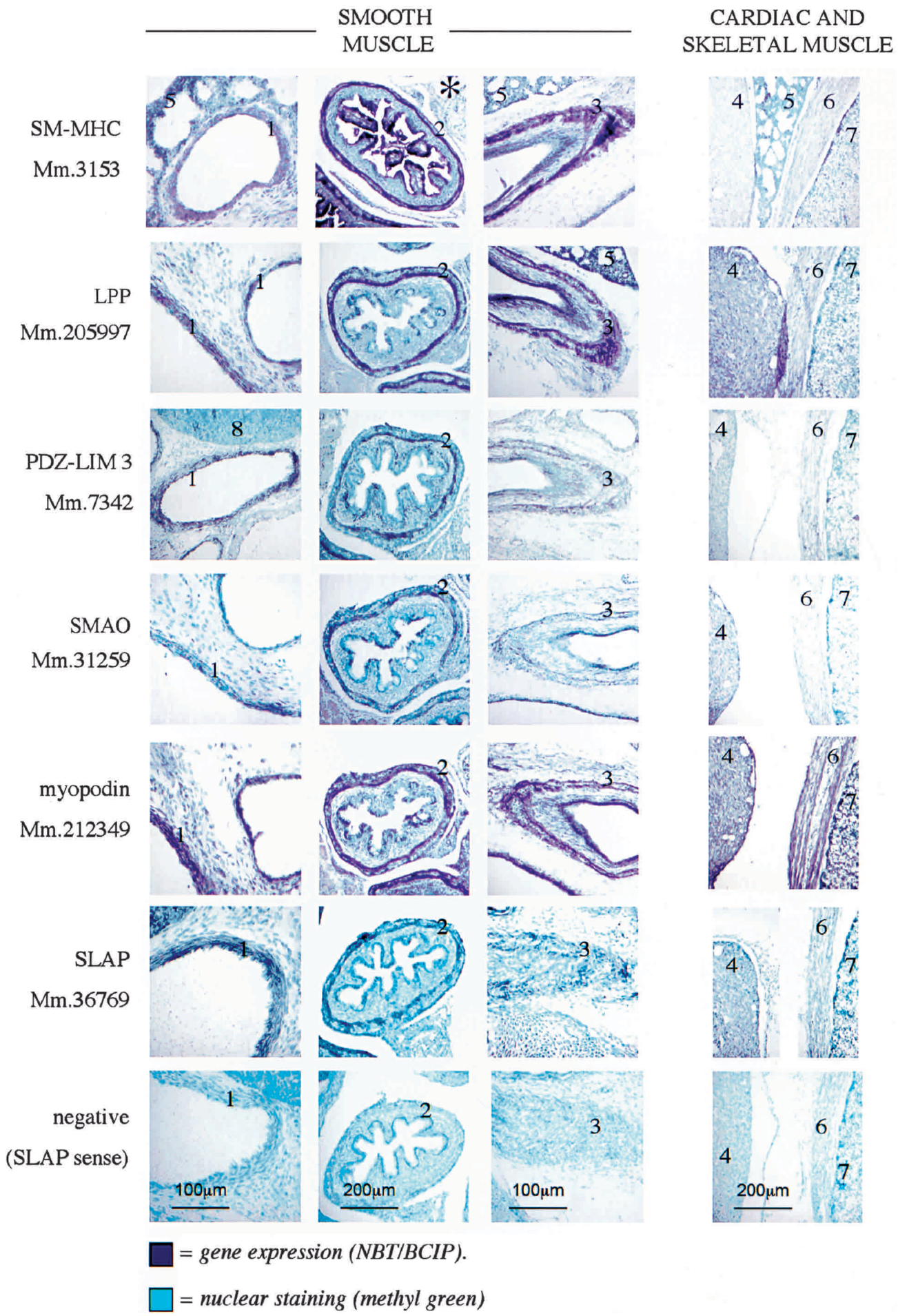

Figure 5 (Legend on next page)

Table 2. Summary of In Situ Hybridization Results

Expression	Correlation	Probability	UniGene ID	Annotation	IMAGE clone	In situ hybridization result
105	1.00	0.99	Mm.3153	Smooth muscle myosin heavy chain	1547017	SMC marker
91	0.76	0.77	Mm.31259	SM-MAO	1166836	SMC marker
344	0.66	0.42	Mm.16537	Alpha smooth muscle actin	1138966	SMC marker
284	0.57	0.15	Mm.205997	LPP	2698658	SMC marker
307	0.56	0.14	Mm.36850	Smoothelin	1162749	SMC marker
147	0.56	0.14	Mm.7342	PDZ and LIM domain 3	332852	SMC marker
29	0.56	0.13	Mm.212349	Myopodin	1617930	SMC marker
253	0.52	0.08	Mm.36769	SLAP	1025661	SMC marker
216	0.43	0.02	Mm.12842	Vinculin	1096898	SMC marker
134	0.57	0.16	Mm.28649	Putative serine protease	5057524	SMC marker-related
76	0.56	0.13	Mm.119265	No annotation	347893	SMC marker-related
94	0.54	0.11	Mm.1237	Pmp22	1121305	SMC marker-related
118	0.53	0.09	Mm.27390	Similar to human myoneurin	1195532	SMC marker-related
58	0.52	0.08	Mm.31552	No annotation	1365251	SMC marker-related
9	0.75	0.75	Mm.156846	Putative myosin ATPase	1137251	Nonselective
10	0.65	0.40	Mm.160140	No annotation	3412898	Nonselective
76	0.62	0.27	Mm.17807	No annotation	1532956	Nonselective
6	0.58	0.18	Mm.31071	No annotation	1052804	Nonselective
34	0.57	0.15	Mm.34461	Putative ATPase	643914	Nonselective
51	0.56	0.13	Mm.314	No annotation	1529589	Nonselective
66	0.54	0.10	Mm.183030	Weakly similar to ankyrin	1494583	Nonselective
77	0.50	0.06	Mm.41387	No annotation	337790	Nonselective
229	0.48	0.04	Mm.28406	unc93 homolog B 86% similarity	1245118	Nonselective
352	0.48	0.04	Mm.184314	Putative dioxygenase	479719	Nonselective
69	0.48	0.04	Mm.45173	Msr2	390123	Nonselective
101	0.47	0.04	Mm.22588	Similarity to human supervillin	949239	Nonselective
214	0.45	0.03	Mm.10211	Entpd5	579841	Nonselective
117	0.41	0.02	Mm.32011	Kelch-related	458907	Nonselective
349	0.38	0.01	Mm.34315	Similar to synembryn	5055713	Nonselective
410	0.37	0.01	Mm.28793	Snap25bp	465348	Nonselective
31	0.65	0.40	Mm.57221	Hoxb13, homeobox B13	1026377	Technical failure
222	0.58	0.19	Mm.34060	No annotation	679316	Technical failure
55	0.58	0.17	Mm.45019	No annotation	2938912	Technical failure
7	0.58	0.17	Mm.139005	No annotation	2780347	Technical failure
189	0.57	0.16	Mm.38877	No annotation	4000106	Technical failure
60	0.57	0.15	Mm.23983	Niban protein	1344864	Technical failure
139	0.56	0.15	Mm.22941	Putative methyltransferase	1449933	Technical failure
7	0.51	0.07	Mm.173134	No annotation	577206	Technical failure
100	0.49	0.05	Mm.196527	Ankyring repeat domain protein	1514780	Technical failure
26	0.48	0.04	Mm.33974	Unc gene homolog	1617070	Technical failure
121	0.44	0.02	Mm.30693	Basic Krüppel-like factor 4	3464873	Technical failure
181	0.43	0.02	Mm.25707	No annotation	1066724	Technical failure
245	0.42	0.02	Mm.29865	Zinc finger domain-containing	425825	Technical failure
80	0.41	0.02	Mm.139238	No annotation	402913	Technical failure
211	0.40	0.01	Mm.28119	Sgpl1	1531669	Technical failure
245	0.37	0.01	Mm.31353	Vps16 gene product	4317642	Technical failure

In situ hybridization validation of predicted SMC-specific genes. Forty-six EST clusters were chosen for validation by in situ hybridization in mouse E17.5 embryos. Genes were classified as SMC markers, SMC marker-related, nonselective, and technical failure, according to criteria described in Results. Expression, Number of detections in UniGene; correlation, Pearson's correlation with SM-MHC in the raw data set (see Methods); probability, logistic regression probability for a gene to be an SMC marker; IMAGE clone, probe template clone ID (see Methods). These results are visualized in Fig. 8. The logistic regression estimate of the total number of SMC markers in the list shown is 7.4; standard error 2.1.

distribution. Profiles generated from binary data and frequency data generally performed poorly. Similarly, established distance metrics such as Euclidean distance and covariance were less suitable. With the current data set, we saw no clear benefit in using statistical transformation or permutation testing (Fig. 3) in comparison with simpler approaches such as Pearson's correlation. The fact that the size of an EST library (10^2–10^5 tags) is small in relation to the often minute frequencies at which individual genes are represented in the transcriptome (10^{-6} to 10^{-1}) indicates that expression levels can only be crudely estimated from EST data.

Genes differ markedly in their absolute expression level. Rare transcripts have a low representation in EST libraries, are less well detected in assays such as ISH and Northern blotting, and give weak signals with a large experimental error on fluorescence-based chip assays such as cDNA microarrays. As with any other method, the EST screening procedure is likely to be more error prone with rare transcripts. Our definition of expression level is based on the sum of UniGene detections. In the analysis presented earlier, we anticipated that any cluster with fewer than five (arbitrary limit) members was unlikely to give reliable results. Filtering the data gave 29,812 clusters with five or more mem-

Figure 5 ISH data for genes that were classified as SMC markers. Vascular SMC; 2, intestinal SMC; 3, esophageal SMC; 4, cardiac muscle; 5, lung parenchyma; 6, diaphragm (skeletal muscle); 7, liver. Intestinal lumen artefact staining is due to endogenous alkaline phosphatase activity (asterisk).

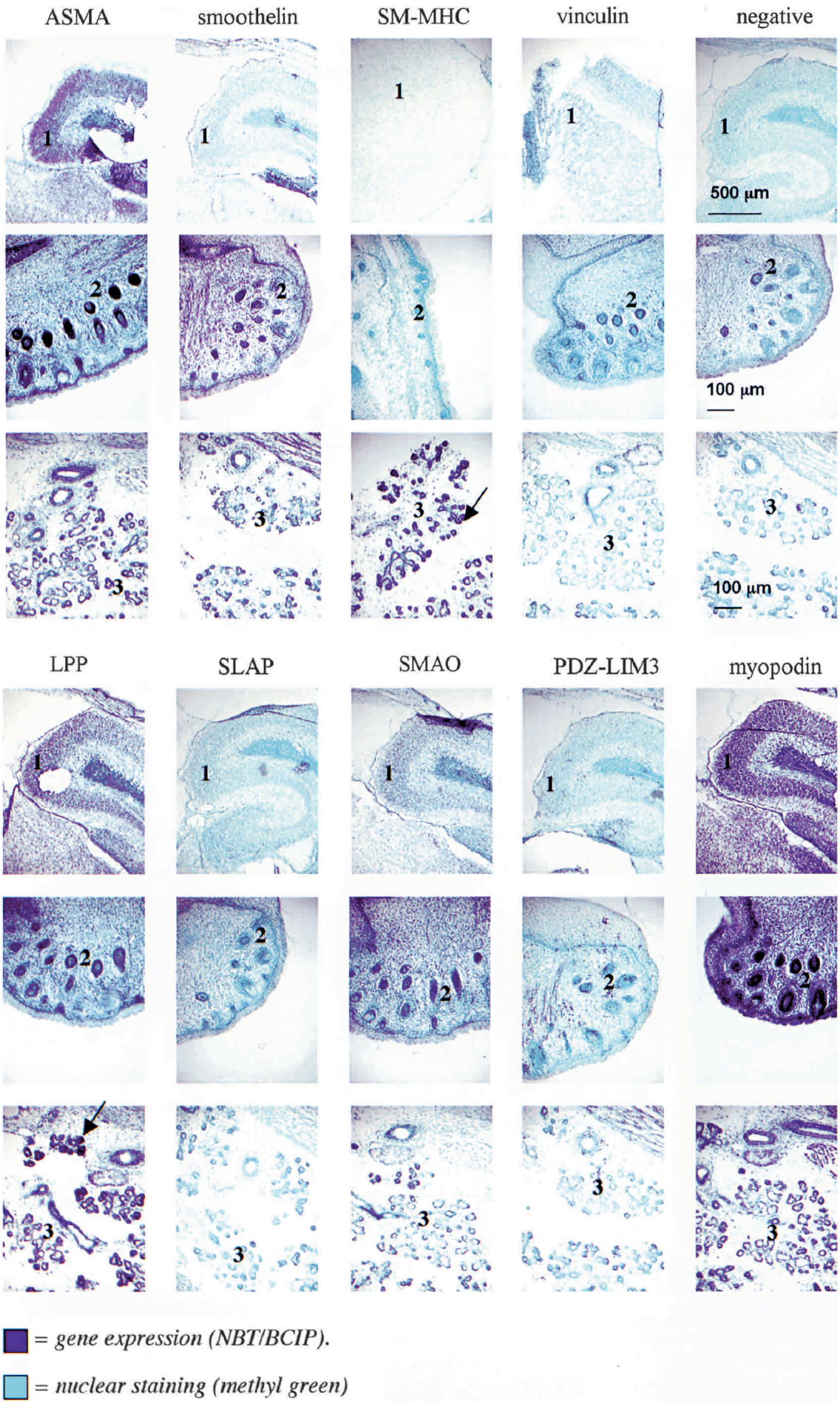

Figure 6 (Legend on next page)

bers, which is in parity to the estimated 30,000–40,000 genes in mammals.

Lipoma Preferred Partner and Smooth Muscle Monoamine Oxidase: Two SMC-Specific Genes With Potential Regulatory Functions

On the basis of protein sequence and previously published data, two of the newly detected SMC markers stand out as interesting genes with potential regulative functions in SMC biology: lipoma preferred partner (LPP) and smooth muscle monoamine oxidase (SMAO).

LPP

LPP belongs to the zyxin protein family (Petit et al. 1996). The family members zyxin, Trip6, ajuba, LIMD1, and LPP have similar structures with a proline-rich domain at the N terminus and three successive LIM domains at the C terminus. LPP and other members of the family are shuttled between focal adhesions and the nucleus in vitro (Petit et al. 2000; Petit et al. 2003). The significance of the nuclear localization is not clear, but LPP strongly enhances transcription in a Gal4-reporter system. Previously published adult human Northern blot data (Petit et al. 1996) are compatible with smooth muscle-specific expression, although the connection was not made. The name LPP was motivated by frequent fusion between LPP and the HMGIC gene in lipomas, indicating a possible role in tumor development (Petit et al. 1996).

Recently, Chang et al. (2003) showed that CRP-1 and 2 act as adapters between serum response factor (SRF) and GATA4/6 to form a transcription factor complex that induces expression of SMC-specific genes such as SM-MHC in vitro. CRP-1 and 2 also shuttles between the nucleus and the cytoplasm and has been shown to bind zyxin (Sadler et al. 1992). Hypothetically, LPP is the endogenous binding partner of the SRF-CRP-GATA complex in SMC and a coregulator of SMC-specific transcription. Targeted disruption of the zyxin gene in mice did not generate any obvious phenotypic changes, possibly due to function overlap between zyxin family members (Hoffman et al. 2003).

SMAO

The SMAO gene is previously undocumented, and belongs to a single-gene family. Domain annotation identified a C-terminal flavin-containing monoamine oxidase (MAO) domain. A Psi-BLAST search against the National Center for Biotechnology Information (NCBI) protein database showed that the predicted MAO domain matches MAO domains of numerous other proteins at 20%–35% sequence identity, including mammalian MAO-A and MAO-B proteins. Mouse SMAO has unique rat and human orthologs, with ~90% global protein sequence identity. On the basis of domain information, it is possible that SMAO functions as a smooth muscle-specific or smooth muscle-selective MAO enzyme. As such, it may contribute to the MAO activity registered in, for example, aortic SMCs (Jaakkola et al. 1999). MAO activity in vascular SMC has been implicated as a potential source of advanced glycosylation end products (Mathys et al. 2002). SMC specificity, a potential for pharmacological manipulation, and coupling to medically relevant processes motivate further studies of this protein.

Potential Roles in SMC Biology for Previously Described Muscle Proteins and for Pmp22

An unexpected finding was that peripheral myelin protein 22 (Pmp22) showed strong selective muscular expression, with expression in specific SMC subsets at E17.5 (Fig. 7). This gene has previously been reported to be expressed at nonneural locations during embryogenesis, indicating a function not restricted to neural processes (Baechner et al. 1995). Another member of the Pmp22 family, epithelial membrane protein 3 (Emp3) is predicted as one of the extreme top correlators of fast skeletal muscle myosin light chain (Table 3B). An intriguing hypothesis is that different members of the Pmp22 family of proteins participate in muscle processes, in a subtype-dependent manner.

The ISH data demonstrated SMC/muscle selectivity for a number of genes with existing documentation in cardiac and skeletal muscle: sarcolemmal-associated protein, PDZ-LIM3 or alpha actinin-associated LIM protein, and myopodin (Fig. 5). These may represent a subset of muscular proteins that participate in SMC processes. Furthermore, the gene matching the cluster Mm. 119265 (Fig. 7) is homologous to the actin-bundling genes palladin and myopalladin, indicating a role in actin bundling.

Can Upstream Developmental Regulators Be Extrapolated From a Core Set of Cell Type-Specific Genes?

This work was initiated to identify a core set of SMC-specific genes. The expression of these genes is controlled, directly or indirectly, by transcription factors involved in the developmental regulation of the cell phenotype. A critical question is whether upstream regulators coexpress with downstream (e.g., structural) genes to such a degree that they may be detected by EST screening.

A majority of documented SMC-specific genes depend on SRF for their expression (Mack et al. 2000; Chang et al. 2001; Manabe and Owens 2001a,b; Strobeck et al. 2001). SRF is not SMC specific but is expressed more widely in other muscle types and in other tissues (Chai and Tarnawski 2002). In the SM-MHC-based search, SRF appears on rank 333, slightly below the top 1 percentile, which is compatible with an expression pattern that is related to SM-MHC. The idea that upstream developmental regulators may be found among top-ranking genes is supported by the presence of several well-documented regulators in the top 1 percentile of skeletal- and heart muscle-specific genes: The transcription factors myogenin, MyoD1, and MRF4 are key regulators of skeletal muscle induction in vivo (Pownall et al. 2002). A search based on skeletal muscle myosin light chain (UniGene ID Mm.1000) detected these regulators at ranks 248, 43, and 19, respectively. A fourth factor, Myf5 (Pownall et al. 2002), was not represented in the data analyzed. A search based on cardiac myosin heavy chain (UniGene ID Mm.3776) identified the following developmental regulators at top positions: the mouse *tinman* homolog Nkx2.5 (Cripps and Olson 2002) at rank 27, the Xin (Wang et al. 1999) at rank 50, and the transcription factor myocardin (Wang et al. 2001) at rank 80. Other factors indicated in the same process, such as MEF2C, dHand, eHand, Gata4, and Gata6 (Cripps and Olson 2002) do not appear in top-ranking positions in this search, possibly indicating a more general expression pattern or a temporally restricted heart-specific expression. For SMC biology, the obvious future prospect of formalized EST screening is to pinpoint SMC-specific candidate genes with potential developmental and regulatory functions. However, SMC-specific expression does not necessarily require SMC-specific regulators, but may result from combinatorial interac-

Figure 6 Ectopic expression of SMC markers in the CNS, hair follicles, and the submandibular gland. Expression in the submandibular gland epithelium was confined to the distal tubular system (arrows). 1: CNS, 2: hair follicle dermal sheaths, 3: submandibular glands.

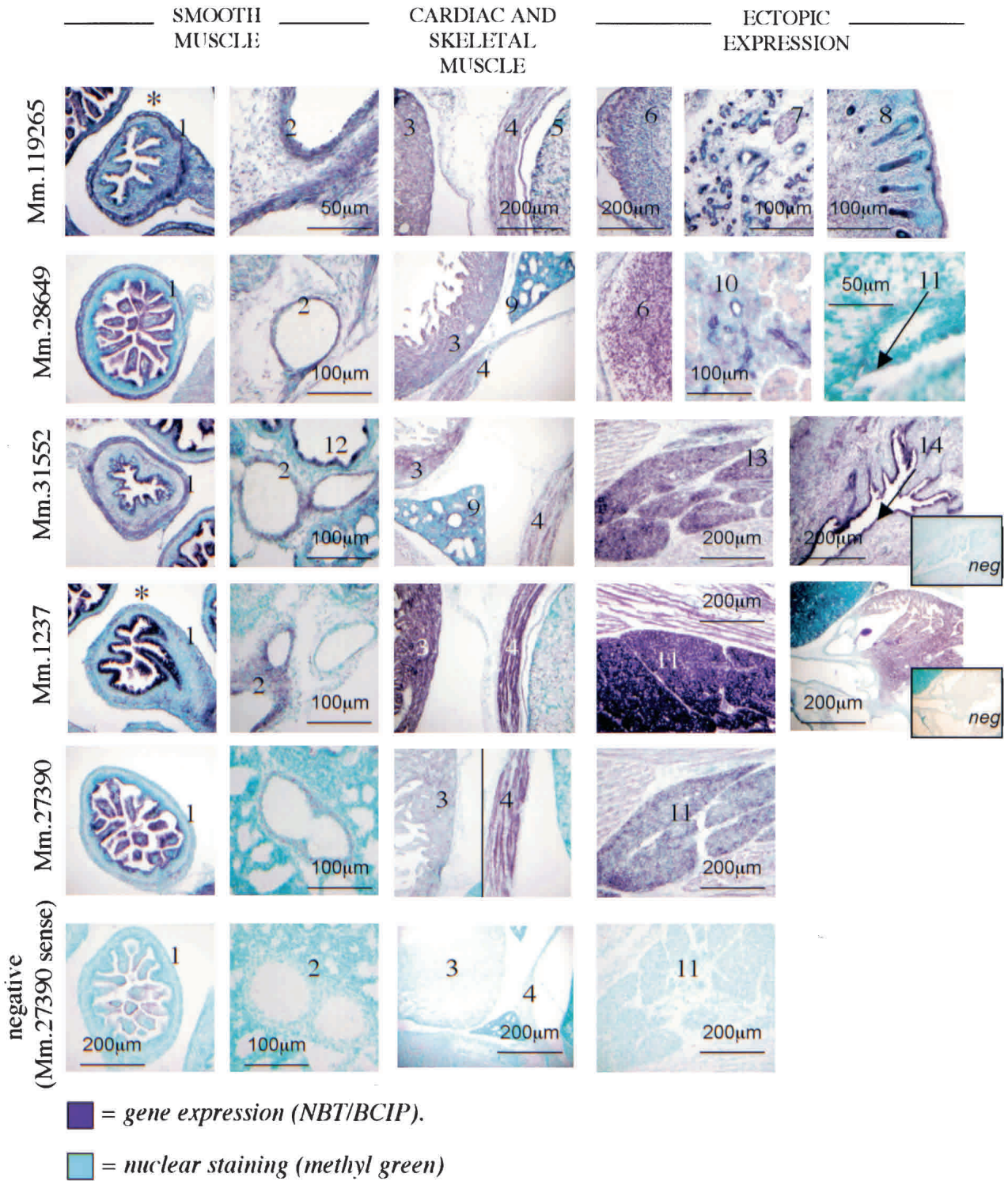

Figure 7 ISH data for SMC marker-related genes. Expression was prominent in muscle cells but was also seen in a range of ectopic sites. 1, Intestinal SMC; 2, vascular SMC; 3, cardiac muscle; 4, diaphragm (skeletal muscle); 5, liver parenchyma; 6, CNS; 7, submandibular gland; 8, hair follicle (general mesenchymal expression seen); 9, lung parenchyma; 10, pancreas; 11, liver; 12, airway; 13, brown adipose tissue; 14, urinary bladder. Intestinal lumen artefact staining is due to endogenous alkaline phosphatase activity (asterisk).

tions between factors with less restricted expression patterns. Nkx3–1 and SRF form a complex that has been shown to regulate the smooth muscle γ-actin gene in vitro (Carson et al. 2000). Similarly, the previously mentioned SRF-CRP1/2-GATA4/6 complex induced the expression of several SMC-specific genes including SM-MHC in vitro (Chang et al. 2003). Our EST screening is optimized to predict genes with correlating expression boundaries and may not identify less restricted regulators.

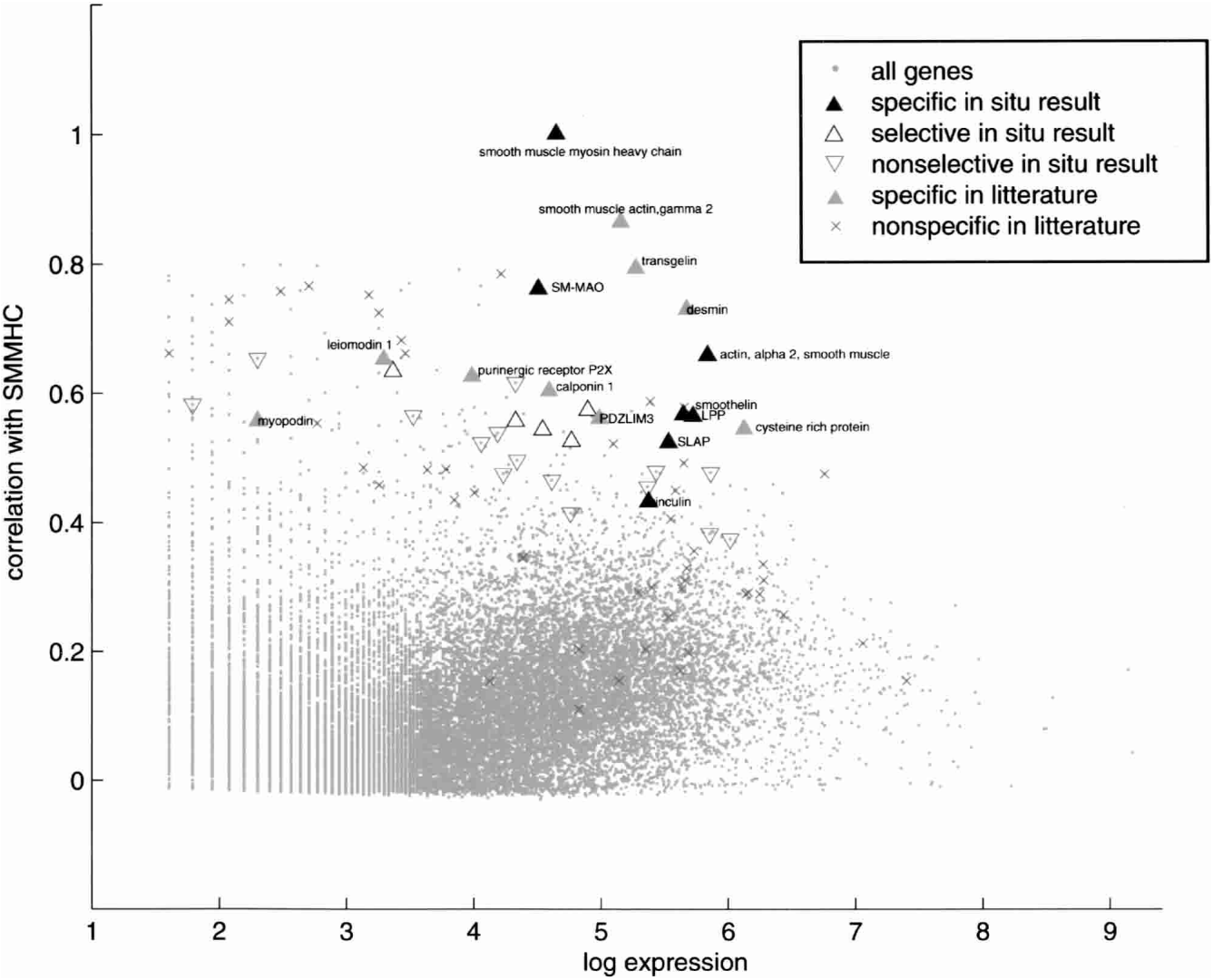

Figure 8 ISH results in relation to profile similarity to SM-MHC (*Y*-axis) and to expression level (*X*-axis). Expression level is defined as the logarithm of the number of UniGene detections for a gene, and profile similarity is defined as Pearson's correlation in the *raw* data set. Gray dots, All 29,812 genes with at least five UniGene detections; dark upward-pointing triangles, ISH-detected SMC markers; open upward-pointing triangles, ISH-detected SMC marker-related genes; downward-pointing triangles, genes with a nonselective expression pattern in the ISH experiment; grey upward-pointing triangles, SMC markers from the literature; crosses, nonselective genes from the literature.

Support for Extended SMC Properties of Myoepithelial Cells

Starfish-shaped cells in the outer layers of exocrine gland epithelium share features of both epithelium and SMCs. In the mammary gland, these myoepithelial cells have been shown to express several SMC markers including ASMA, SM-MHC, calponin, and h-caldesmon (Lazard et al. 1993). Myoepithelial cells in other exocrine glands have been shown to express ASMA and calponin (Ogawa et al. 1999). The myoepithelial cells have ultrastructural features that are typical of SMC and they have been shown to contract (Redman 1994). Our data show that, in addition to the previously known genes ASMA and SM-MHC, several SMC-specific genes are expressed in the submandibular gland epithelium, including LPP, SMAO, and myopodin. The resolution of our data is not sufficient to exactly define the spatial location of the SMC marker expressing cells within the epithelium. It seems likely that they are synonymous with the previously described myoepithelial cells, or progenitors of these cells. The growing number of SMC-specific genes that are expressed in myoepithelium indicates that this cell type should be more consciously considered in studies of SMC differentiation and in studies of regulation of SMC-specific genes.

Ectopic Expression in Nonmuscle Cells Indicates Coregulation of SMC-Specific Genes?

Our ISH experiments showed frequent expression of the known markers in nonmuscle tissues such as the submandibular gland epithelium. Interestingly, the new markers were detected in the same ectopic locations. Coordinated expression of a group of genes is either a result of shared upstream regulators, or an effect of convergent evolution where unique combinations of *cis*-regulatory elements in the individual genes result in overlapping expression. Large numbers of coexpressed genes and a complex expression pattern decreases the probability of convergent evolution being the explanatory factor. SMC-specific genes are expressed in selective subsets of cells originating from all three germ layers: neural crest-derived vascular SMC, mesoderm-

Table 3. Prediction of Specific Expression in Other Tissues

Expression	Correlation	ID	Annotation
A. Prediction of cardiac-specific genes			
270	1.00	Mm.3776	Myosin heavy chain, cardiac muscle, adult
127	0.82	Mm.155714	Myosin, heavy polypeptide 7, cardiac muscle, beta
59	0.82	Mm.10728	Myosin binding protein C, cardiac
389	0.82	Mm.686	Actin, alpha, cardiac
137	0.72	Mm.1529	Myosin light chain, phosphorylatable, cardiac ventricles
58	0.71	Mm.20271	Fibrillin 2
47	0.69	Mm.19961	ESTs, highly similar to atrial natriuretic factor
201	0.67	Mm.632	Troponin T2, cardiac
51	0.66	Mm.46514	Myosin light chain, regulatory A
11	0.66	Mm.89854	A disintegrin and metalloprotease domain 11
6	0.66	Mm.171448	ESTs, weakly similar to mouse forkhead protein O1A
10	0.64	Mm.207070	*Mus musculus* heart alpha-kinase mRNA, partial cds
43	0.64	Mm.89727	CD8beta opposite strand
5	0.60	Mm.33561	ESTs
5	0.60	Mm.125583	ESTs
5	0.60	Mm.208235	ESTs
6	0.59	Mm.158756	RIKEN cDNA 2610019F11 gene
8	0.59	Mm.27037	ESTs
52	0.59	Mm.17235	Cysteine-rich protein 3
109	0.58	Mm.43	Myosin light chain, alkali, cardiac atria
136	0.57	Mm.6375	Glycogenin 1
7	0.57	Mm.89976	Zinc finger, imprinted 1
11	0.57	Mm.210460	ESTs
5	0.56	Mm.209027	ESTs
12	0.56	Mm.218070	ESTs
36	0.56	Mm.4211	Solute carrier family 8 (sodium/calcium exchanger), member 1
7	0.56	Mm.41974	Nkx2.5/tinman homolog
8	0.55	Mm.116789	Repressor of GATA
5	0.55	Mm.160735	ESTs
5	0.55	Mm.189024	ESTs
B. Prediction of skeletal muscle-specific genes			
341	1.00	Mm.1000	Myosin light chain, alkali, fast skeletal muscle
31	0.80	Mm.57093	Calcium channel, voltage-dependent, gamma subunit 1
18	0.77	Mm.18125	Expressed sequence AI118095
28	0.74	Mm.33171	Ankyrin repeat and SOCs box-containing protein 5
51	0.74	Mm.140604	RIKEN cDNA 1110002H13 gene
428	0.71	Mm.214950	Actin, alpha 1, skeletal muscle
124	0.71	Mm.29358	RIKEN cDNA 2700055K07 gene
156	0.71	Mm.29994	Enolase 3, beta muscle
260	0.70	Mm.14526	Myosin light chain, phosphorylatable, fast skeletal muscle
121	0.69	Mm.14546	Troponin T3, skeletal, fast
107	0.69	Mm.20829	Epithelial membrane protein 3
254	0.68	Mm.39469	Troponin I, skeletal, fast 2
43	0.68	Mm.46232	Integrin beta 1 binding protein 2
38	0.67	Mm.4583	Cholinergic receptor, nicotinic, alpha polypeptide 1 (muscle)
41	0.67	Mm.144259	RIKEN cDNA 2310050C09 gene
52	0.65	Mm.31099	ADP-ribosyltransferase 1
152	0.65	Mm.36900	Troponin I, skeletal, slow 1
135	0.65	Mm.1583	Lymphocyte antigen 6 complex, locus C
12	0.65	Mm.11	MRF4
14	0.64	Mm.58214	RIKEN cDNA 2310002L13 gene
78	0.64	Mm.712	Troponin C, cardiac/slow skeletal
22	0.64	Mm.140151	Small proline-rich protein 1B
51	0.63	Mm.125614	Myosin binding protein H
21	0.63	Mm.2810	Cholinergic receptor, nicotinic, gamma polypeptide
27	0.63	Mm.45734	RIKEN cDNA 2310024D23 gene
110	0.63	Mm.878	Lymphocyte antigen 6 complex, locus D
24	0.63	Mm.45137	RIKEN cDNA 2310032K21 gene
34	0.62	Mm.10194	Myomesin 2
272	0.62	Mm.43831	Lectin, galactose binding, soluble 1
19	0.61	Mm.36668	RIKEN cDNA 1110008K04 gene
C. Prediction of retina-specific genes			
67	1.0	Mm.8008	Cone-rod homeobox containing gene
92	1.0	Mm.1205	Rod photoreceptor 1
94	1.0	Mm.1372	Phosphodiesterase 6B, cGMP, rod receptor, beta
37	1.0	Mm.78749	Retinitis pigmentosa 1 homolog (human)
418	1.0	Mm.69061	Alpha transducin
23	1.0	Mm.206228	ESTs

(continued)

Table 3. *Continued*

Expression	Correlation	ID	Annotation
62	1.0	Mm.59151	ESTs
830	1.0	Mm.2965	Similar to rhodopsin
85	1.0	Mm.151562	Similar to retinol-binding protein
6	1.0	Mm.172488	ESTs
38	1.0	Mm.194050	ESTs
31	1.0	Mm.1370	Phosphodiesterase 6A, cGMP-specific, rod, alpha
15	1.0	Mm.23793	Cyclic nucleotide gated channel, cGMP gated
38	1.0	Mm.42102	Tubby like protein 1
38	1.0	Mm.41982	Retinoschisis 1 homolog (human)
23	1.0	Mm.39200	Phosphodiesterase 6G, cGMP-specific, rod, gamma
114	1.0	Mm.1276	Retinal S-antigen
29	1.0	Mm.95707	RIKEN cDNA A930007I01 gene
82	1.0	Mm.679	Rod outer segment membrane protein 1
59	1.0	Mm.20422	Neural retina leucine zipper gene
D. Prediction of lens-specific genes			
135	1.0	Mm.1228	Crystallin, alpha A
25	0.9	Mm.30374	Crystalline, gamma C
7	0.9	Mm.130559	ESTs
16	0.9	Mm.127184	ESTs, weakly similar to A39757 beta-crystallin
64	0.9	Mm.22830	Crystalline, beta A1
9	0.9	Mm.22861	ESTs
5	0.9	Mm.215166	ESTs
7	0.9	Mm.209940	ESTs
45	0.9	Mm.89477	Crystalline, gamma F
48	0.9	Mm.31625	Major intrinsic protein of eye lens fiber
13	0.9	Mm.95578	ESTs
17	0.9	Mm.86656	Crystalline, beta A2
22	0.9	Mm.168942	BR3B
7	0.9	Mm.180528	ESTs
25	0.9	Mm.138345	ESTs
93	0.9	Mm.127171	Beaded filament structural protein in lens-CP94
5	0.9	Mm.209953	ESTs
26	0.9	Mm.215250	Solute carrier family 7
9	0.9	Mm.137178	RIKEN cDNA E130202I16 gene
11	0.9	Mm.138333	Ras p21 GTPase

Nearest neighbor searches were performed with four different tissue-specific genes: cardiac-specific myosin heavy chain (A), skeletal muscle myosin light chain (B), the retina-specific cone-rod homeobox (C), and the lens-specific genes crystallin alpha A (D). Top 30 genes are shown in A and B. Top 20 genes are shown in C and D. On the basis of literature validation, genes specifically expressed in each tissue were identified (underlined in each list). Expression, Number of detections in UniGene.

derived visceral and vascular SMC, and endoderm-derived submandibular epithelium. This is a highly complex expression pattern, which gives indirect support for the idea of a coregulated SMC expression module.

Perspective

This work establishes mathematical prediction as a mode for expanding the molecular repertoire of SMC biology. We have identified a number of novel markers and estimated the total number of SMC-specific genes to be ~50. We anticipate that the remainder of these genes can be identified by this methodology. Further, comparative studies of *cis*-regulatory elements, for example, by phylogenetic footprinting approaches, in the now expanded core set of SMC markers may provide more information of the gene networks regulating the differentiation of SMCs.

METHODS

1. Bioinformatics and Statistics

Data Preparation

The aim of the data preparation step is to produce a tag sampling data matrix on the form D_{ij} = number of detections of gene i in library j. Mouse UniGene data were downloaded from the NCBI server (ftp://ftp.ncbi.nih.gov). Other potential sources of tag sampling data include SAGEdb (Lash et al. 2000) and BodyMAP (Hishiki et al. 2000). EST counts were extracted from the UniGene flat files using software that can be downloaded from http://cbz.gu.se/Lindahl/EST. In the UniGene-derived matrix D_{ij}, rows correspond to UniGene clusters and columns to cDNA libraries. Subsequent analysis was performed with programs written in the MATLAB (MathWorks, Inc.) language. Data were filtered to exclude UniGene clusters with fewer than five (arbitrary limit) members, leaving 29,812 clusters.

EST–Derived Expression Data

The EST data preparation resulted in a matrix on the form D_{ij} = number of tag detections of gene i in library j. In this matrix, the sum of $D_{.j} = D_{1j} + D_{2j} + \ldots + D_{Nj}$ corresponds to the size of library j, whereas the sum of $D_{i.} = D_{i1} + D_{i2} + \ldots + D_{iM}$ is the number of detections of gene i in the data (N = number of genes, and M = number of libraries). $D_{i.}$ can be seen as a global measure of the abundance of gene i. The analysis is based on interpreting tag counts as expression levels of genes. The naïve use of the tag counts in D_{ij} as the expression signal is termed "raw data" (Fig. 2). A variety of the raw data matrix, B_{ij}, termed "binary data" was calculated using $B_{ij} = 0$ if $D_{ij} = 0$, $B_{ij} = 1$ if $D_{ij} > 0$.

Viewing the number of detections in a certain library as fixed, the distribution of detection events across different genes can be modeled statistically. A test statistic R_{ij}, termed *transformed data* (Fig. 2), was formulated: for each cluster C and each library L, the R statistic tests whether the proportion of the sequences in

L that are in C could be the same as the corresponding proportion for all other libraries combined. More precisely, a generalized likelihood ratio test (Rice 1995) is preformed. Let, P_L, respectively, P_0, be the probability for a sequence in L, respectively in all other libraries, to be in C. Let D_{LC} and D_{L0}, respectively D_{0C} and D_{00}, be the observed counts of sequences in C and out of C for L, respectively, for all other libraries. Assuming binomial distributions for the data, the maximum likelihood estimates become

$$\hat{p}_L = \hat{p}_0 = \frac{D_{LC} + D_{0C}}{D_{LC} + D_{L0} + D_{0C} + D_{00}}$$

under the null hypothesis

$$P_L = P_0,$$

and

$$\tilde{p}_L = \frac{D_{LC}}{D_{LC} + D_{L0}}, \tilde{p}_0 = \frac{D_{0C}}{D_{0C} + D_{00}}$$

otherwise. The generalized likelihood ratio test statistic becomes

$$R = -2 \log \frac{\hat{p}_L^{D_{LC}} (1 - \hat{p}_L)^{D_{L0}} \hat{p}_0^{D_{0C}} (1 - \hat{p}_0)^{D_{00}}}{\tilde{p}_L^{D_{LC}} (1 - \tilde{p}_L)^{D_{L0}} \tilde{p}_0^{D_{0C}} (1 - \tilde{p}_0)^{D_{00}}},$$

with an approximate χ^2 distribution with one degree of freedom under the null hypothesis.

We also calculated the proportion of transcripts for each gene in a library, $p_{ij} = D_{ij}/D_{.j}$, termed "frequency data."

Distance Metrics

The distance metrics evaluated were Pearson's correlation coefficient and covariance. For the binary data, standard correlation and the use of Fischer's exact test (Rice 1995) were applied, the latter corresponding to a previously published method, GBA (Walker et al. 1999). We also constructed an alternative metric based on pair-wise permutation testing, and interpretation of the *P* value as coexpression. More precisely, for each pair of profiles, the correlation between them with the libraries permuted in one of them is computed for a number of permutations. The rank of the actual correlation in this set of numbers can be interpreted as a *P* value for the hypothesis that the profiles are enriched in unrelated libraries. We interpreted this *P* value as a distance between the profiles.

Nearest Neighbor Search and Annotation

Let *v(i)* be the expression profile for gene *i* (i.e., row *i* in one of the data matrices D_{ij}, B_{ij}, R_{ij}, and p_{ij}, discussed earlier). Given one single training example, *b*, and a distance metric, the distance between *v(b)* and *v(1),v(2),…(N)* was calculated (*N* = the number of genes). Genes were then sorted according to distance and ranked. From the top-ranking genes downward, genes were annotated against the Celera Discovery System and public databases. Genes with sufficient published expression information were classified as SMC specific or not.

SMC–Specific Reference Gene Set

The following set of smooth muscle-specific genes was used as a reference set in the evaluation of data treatments and distance metrics below: alpha smooth muscle actin (ASMA, Acta2), smooth muscle myosin heavy chain (SM-MHC, Myh11), gamma smooth muscle actin (Actg2), desmin (Des), leiomodin 1 (Lmod1), SM-22-alpha (transgelin,Tagln), smoothelin (Smtn), calponin 1 (Cnn1), cysteine-rich protein 1 (Crp1), and vinculin (Vcl).

Logistic Regression

Logistic regression was used to model the probability of two genes sharing the same expression pattern, based on profile similarity. The model assumes a linear relationship between profile similarity *x* and probability *p* on the form $ln[p/(1-p)] = a + bx$, where *a* and *b* are the parameters estimated from "training data" (i.e., examples of SMC markers and nonmarkers; see following). Our implementation is based on the MATLAB statistics toolbox. The model returns p_i = the probability for gene i to be an SMC marker based on the profile distance (x_i) to SM-MHC. Parameters of the model were fitted by literature data, or literature data in combination with in situ results.

The expectation of the number of genes with the same expression pattern in a specific set, for example, the top-ranking 200 genes, is given by calculating the sum of p_i's in that set. The standard error for the expectation is estimated as the square root of the sum over all $p_i(1-p_i)$ in the set.

Logistic Regression Training Data

In order to compare alternative methods with respect to classifying genes as SMC markers or non-SMC markers in such a way as to not benefit one specific method, and because it was unfeasible to annotate the full set of genes, the following reference set of genes was constructed. The 1% top-ranking ($n = 289$) genes were identified using seven alternative search methods: 1–6, nearest neighbor searching against SM-MHC in the raw (*Dij*) and transformed (*Rij*) data matrices using Pearson's correlation coefficient, permutation testing, and covariance as distance metrics. 7, nearest neighbor searching against SM-MHC in the binary data (*Bij*) matrix using GBA. Merging lists and removing redundancies produced a list of 975 genes. Annotation of this list resulted in a test set with 14 positive examples (SMC markers) and 134 negative examples (non-SMC markers, that is, genes documented as having another expression pattern). In order to include a reasonable number of negative examples for the remaining $29{,}812 - 975$ genes, a random subset of 200 genes from this set was annotated. Nineteen negative examples and 0 positive examples were found. Based on this, $(29{,}812 - 975 - 200) \times 19/200 = 2720$ genes were randomly selected and used as additional negative examples.

Comparison of Classification Performance

Using distance data from the seven different methods described earlier, the literature reference data were analyzed by logistic regression. In a simulation that was repeated 50 times, the 14 positive and 134 negative reference genes were split into a training set (75% of the genes in each group) and a test set (remaining 25% in each group). The logistic model was then fitted using the training data plus additional random negative examples (in order to avoid overestimation of probabilities due to insufficient annotation, see earlier). Model predictions of the probabilities for genes in the test set to be SMC markers were recorded, and expectations (mean values) calculated. To further compare methods in a pair-wise fashion, subsets of reference genes present in the top 1% list of either method were used.

2. Experimental Validation

DIG–RNA Probe Preparation

Clones in the vector pT3T7-PAC representing candidate UniGene clusters were purchased from Research Genetics, Inc. Plasmid midi-preparations were done according to Sambrook et al. (1989) and plasmids were linearized by restriction digestion with *Not*I (sense probes), and *Eco*RI (antisense probes). One microgram of linearized plasmid was purified with a DNA cleanup kit (Zymo Research, Inc). Digoxigenin-labeled cRNA probes were made with in vitro transcription using T3 (antisense probes) and T7 (sense negative control probes) RNA polymerase (Roche).

In Situ Hybridization

E17.5 mouse embryos were fixed overnight in 4% PFA in PBS, and further incubated in 30% sucrose in PBS. Embryos were mounted in Optimal Cutting Temperature compound (Sakura) and stored at −80°C. Fourteen-micrometer cryosections on SuperFrost plus slides were heated to 55°C for 60 sec and stored at −80°C. incubated at 37°C for 30 min, and washed 5 min each in

PBS, PBS 0.1% tween-20, and PBS. Sections were permeabilized with proteinase K 5µg/mL in TE at 30°C for 30 min, and fixed for 5 min in 4%PFA in PBS. Further steps were performed as described at www.roche-applied-science.com/prod_inf/manuals/InSitu/pdf/ISH_181-188.pdf.

ACKNOWLEDGMENTS

We thank Stefan Scheidl, Holger Gerhart, Mattias Kallen, and Christer Betsholtz for reading the manuscript and suggesting improvements. We thank Olle Nerman for helpful statistical suggestions. This work was funded by the Association for International Cancer Research (AICR), the Swedish Cancer Society, the Swedish Society for Medical Research (SSMF), and SWEGENE.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Baechner, D., Liehr, T., Hameister, H., Altenberger, H., Grehl, H., Suter, U., and Rautenstrauss, B. 1995. Widespread expression of the peripheral myelin protein-22 gene (PMP22) in neural and non-neural tissues during murine development. *J. Neurosci. Res.* **42:** 733–741.

Carson, J.A., Fillmore, R.A., Schwartz, R.J., and Zimmer, W.E. 2000. The smooth muscle γ-actin gene promoter is a molecular target for the mouse bagpipe homologue, mNkx3–1, and serum response factor. *J. Biol. Chem.* **275:** 39061–39072.

Chai, J. and Tarnawski, A.S. 2002. Serum response factor: Discovery, biochemistry, biological roles and implications for tissue injury healing. *J. Physiol. Pharmacol.* **53:** 147–157.

Chang, D.F., Belaguli, N.S., Iyer, D., Roberts, W.B., Wu, S.P., Dong, X.R., Marx, J.G., Moore, M.S., Beckerle, M.C., Majesky, M.W., et al. 2003. Cysteine-rich LIM-only proteins CRP1 and CRP2 are potent smooth muscle differentiation cofactors. *Dev. Cell* **4:** 107–118.

Chang, P.S., Li, L., McAnally, J., and Olson, E.N. 2001. Muscle specificity encoded by specific serum response factor-binding sites. *J. Biol. Chem.* **276:** 17206–17212.

Cripps, R.M. and Olson, E.N. 2002. Control of cardiac development by an evolutionarily conserved transcriptional network. *Dev. Biol.* **246:** 14–28.

De Young, M.P., Damania, H., Scheurle, D., Zylberberg, C., and Narayanan, R. 2002. Bioinformatics-based discovery of a novel factor with apparent specificity to colon cancer. *In Vivo* **16:** 239–248.

Ewing, R.M. and Claverie, J.M. 2000. EST databases as multi-conditional gene expression datasets. *Pac. Symp. Biocomput.* **5:** 430–442.

Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S., and Claverie, J.M. 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* **9:** 950–959.

Hishiki, T., Kawamoto, S., Morishita, S., and Okubo, K. 2000. BodyMap: A human and mouse gene expression database. *Nucleic Acids Res.* **28:** 136–138.

Hoffman, L.M., Nix, D.A., Benson, B., Boot-Hanford, R., Gustafsson, E., Jamora, C., Menzies, A.S., Goh, K.L., Jensen, C.C., Gertler, F.B., et al. 2003. Targeted disruption of the murine zyxin gene. *Mol. Cell. Biol.* **23:** 70–79.

Huminiecki, L., Gorn, M., Suchting, S., Poulsom, R., and Bicknell, R. 2002. Magic roundabout is a new member of the roundabout receptor family that is endothelial specific and expressed at sites of active angiogenesis. *Genomics* **79:** 547–552.

Jaakkola, K., Kaunismaki, K., Tohka, S., Yegutkin, G., Vanttinen, E., Havia, T., Pelliniemi, L.J., Virolainen, M., Jalkanen, S., and Salmi, M. 1999. Human vascular adhesion protein-1 in smooth muscle cells. *Am. J. Pathol.* **155:** 1953–1965.

Konig, S., Burkman, J., Fitzgerald, J., Mitchell, M., Su, L., and Stedman, H. 2002. Modular organization of phylogenetically conserved domains controlling developmental regulation of the human skeletal myosin heavy chain gene family. *J. Biol. Chem.* **277:** 27593–27605.

Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., and Altschul, S.F. 2000. SAGEmap: A public gene expression resource. *Genome Res.* **10:** 1051–1060.

Lazard, D., Sastre, X., Frid, M.G., Glukhova, M.A., Thiery, J.P., and Koteliansky, V.E. 1993. Expression of smooth muscle-specific proteins in myoepithelium and stromal myofibroblasts of normal and malignant human breast tissue. *Proc. Natl. Acad. Sci.* **90:** 999–1003.

Mack, C.P., Thompson, M.M., Lawrenz-Smith, S., and Owens, G.K. 2000. Smooth muscle α-actin CArG elements coordinate formation of a smooth muscle cell-selective, serum response factor-containing activation complex. *Circ. Res.* **86:** 221–232.

Manabe, I. and Owens, G.K. 2001a. CArG elements control smooth muscle subtype-specific expression of smooth muscle myosin in vivo. *J. Clin. Invest.* **107:** 823–834.

———. 2001b. The smooth muscle myosin heavy chain gene exhibits smooth muscle subtype selective modular regulation in vivo. *J. Biol. Chem.* **6:** 6.

Mathys, K.C., Ponnampalam, S.N., Padival, S., and Nagaraj, R.H. 2002. Semicarbazide-sensitive amine oxidase in aortic smooth muscle cells mediates synthesis of a methylglyoxal-AGE: Implications for vascular complications in diabetes. *Biochem. Biophys. Res. Commun.* **297:** 863–869.

Ogawa, Y., Yamauchi, S., Ohnishi, A., Ito, R., and Ijuhin, N. 1999. Immunohistochemistry of myoepithelial cells during development of the rat salivary glands. *Anat. Embryol. (Berl.)* **200:** 215–228.

Petit, M.M., Mols, R., Schoenmakers, E.F., Mandahl, N., and Van de Ven, W.J. 1996. LPP, the preferred fusion partner gene of HMGIC in lipomas, is a novel member of the LIM protein gene family. *Genomics* **36:** 118–129.

Petit, M.M., Fradelizi, J., Golsteyn, R.M., Ayoubi, T.A., Menichi, B., Louvard, D., Van de Ven, W.J., and Friederich, E. 2000. LPP, an actin cytoskeleton protein related to zyxin, harbors a nuclear export signal and transcriptional activation capacity. *Mol. Biol. Cell* **11:** 117–129.

Petit, M.M., Meulemans, S.M., and Van de Ven, W.J. 2003. The focal adhesion and nuclear targeting capacity of the LIM-containing lipoma-preferred partner (LLP) protein. *J. Biol. Chem.* **278:** 2157–2168.

Pownall, M.E., Gustafsson, M.K., and Emerson Jr., C.P. 2002. Myogenic regulatory factors and the specification of muscle progenitors in vertebrate embryos. *Annu. Rev. Cell Dev. Biol.* **18:** 747–783.

Redman, R.S. 1994. Myoepithelium of salivary glands. *Microsc. Res. Tech.* **27:** 25–45.

Rice, J.A. 1995. *Mathematical statistics and data analysis*, pp. 483–485. Duxbury Press, Belmont, CA.

Sadler, I., Crawford, A.W., Michelsen, J.W., and Beckerle, M.C. 1992. Zyxin and cCRP: Two interactive LIM domain proteins associated with the cytoskeleton. *J. Cell Biol.* **119:** 1573–1587.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. *Molecular cloning: A laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Scheurle, D., DeYoung, M.P., Binninger, D.M., Page, H., Jahanzeb, M., and Narayanan, R. 2000. Cancer gene discovery using digital differential display. *Cancer Res.* **60:** 4037–4043.

Schuler, G.D. 1997. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75:** 694–698.

Schwartz, S.M., Virmani, R., and Rosenfeld, M.E. 2000. The good smooth muscle cells in atherosclerosis. *Curr. Atheroscler. Rep.* **2:** 422–429.

Strobeck, M., Kim, S., Zhang, J.C., Clendenin, C., Du, K.L., and Parmacek, M.S. 2001. Binding of serum response factor to CArG box sequences is necessary but not sufficient to restrict gene expression to arterial smooth muscle cells. *J. Biol. Chem.* **276:** 16418–16424.

Walker, M.G., Volkmuth, W., Sprinzak, E., Hodgson, D., and Klingler, T. 1999. Prediction of gene function by genome-scale expression analysis: Prostate cancer-associated genes. *Genome Res.* **9:** 1198–1203.

Wang, D.Z., Reiter, R.S., Lin, J.L., Wang, Q., Williams, H.S., Krob, S.L., Schultheiss, T.M., Evans, S., and Lin, J.J. 1999. Requirement of a novel gene, Xin, in cardiac morphogenesis. *Development* **126:** 1281–1294.

Wang, D., Chang, P.S., Wang, Z., Sutherland, L., Richardson, J.A., Small, E., Krieg, P.A., and Olson, E.N. 2001. Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor. *Cell* **105:** 851–862.

Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225–228.

Received January 25, 2003; accepted in revised form May 27, 2003.