Research

Heart-specific genes revealed by expressed sequence tag (EST) sampling

Karine Mégy, Stéphane Audic and Jean-Michel Claverie

Address: Information Génétique et Structurale - CNRS/UMR 1889, 31 chemin Joseph Aiguier, 13402 Marseille Cedex 20, France.

Correspondence: Jean-Michel Claverie. E-mail: Jean-Michel.Claverie@igs.cnrs-mrs.fr

Published: 25 November 2002

Genome ***Biology*** 2002, **3(12)**:research0074.1–0074.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0074

Received: 31 July 2002
Revised: 18 September 2002
Accepted: 11 October 2002

A previous version of this manuscript was made available before peer review but has been withdrawn by the authors

Abstract

Background: Cardiovascular diseases are the primary cause of death worldwide; the identification of genes specifically expressed in the heart is thus of major biomedical interest. We carried out a comprehensive analysis of gene-expression profiles using expressed sequence tags (ESTs) to identify genes overexpressed in the human adult heart. The initial set of genes expressed in the heart was constructed by clustering and assembling ESTs from heart cDNA libraries. Expression profiles were then generated for each gene by counting their cognate ESTs in all libraries. Differential expression was assessed by applying a previously published statistical procedure to these profiles.

Results: We identified 35 cardiac-specific genes overexpressed in the heart, some of which displayed significant coexpression. Some genes had no previously recognized cardiac function. Of the 35 genes, 32 were mapped back onto the human genome sequence. According to Online Mendelian Inheritance in Man (OMIM), five genes were previously known as heart-disease genes and one gene was located in the locus of a bleeding disorder. Analysis of the promoter regions of this collection of genes provides the first list of putative regulatory elements associated with differential cardiac expression.

Conclusion: This study shows that ESTs are still a powerful tool to identify differentially expressed genes. We present a list of genes specifically expressed in the human heart, one of which is a candidate for a bleeding disorder. In addition, we provide the first set of putative regulatory elements, the combination of which appears correlated with heart-specific gene expression.

Background

The notion of differential expression is central to the identification of target genes of biomedical interest [1]. At the transcriptome level, differentially expressed genes are defined as those exhibiting markedly different amounts of cognate mRNA in one particular tissue, disease or developmental stage, compared with other tissues. Differential expression is involved in various processes, such as development, metabolism, cellular differentiation or pathological states. The identification of differentially expressed genes directs the development of functional assays to understand those biological processes. The analysis of coexpression of differentially expressed genes is also of biomedical interest. Coexpressed genes are defined by their correlated expression patterns in

time or in space. Genes involved in a common cellular function (for example, within a metabolic pathway or encoding interacting proteins) often display significant coexpression. Coexpressed genes might be co-regulated, through common regulatory elements (transcription factor, transcription signal). Coexpression can be used to guide the comparative analyses of promoter regions by enhancing our capacity to identify the weak signature of regulation elements [2].

The most convenient approaches to study differential expression and coexpression use measurements of mRNA abundance. Although mRNA is not the ultimate product of a gene, and that mature protein levels in a cell do not show perfect correlation with the abundance of mRNA [3,4], transcriptional activity and its variation are useful indicators of the involvement of a given gene in a given physiological process. Gene-expression profiling is increasingly carried out using hybridization on various types of microarrays and DNA chips [5]. These techniques are, however, not yet widely available, and have intrinsic limitations of cost and reproducibility. They are not convenient for the comparison of multiple tissues, as they usually require normalization to the same control sample. Moreover, they require access to human tissue samples and can only measure the expression level of the predetermined set of genes spotted on the array.

In 1992, Okubo *et al.* [6] proposed the use of large-scale random 3′-end cDNA library sequencing (3′ expressed sequence tags; ESTs) as a way of estimating the level of gene expression. The abundance of mRNA is simply estimated from the number of cognate ESTs found in each library, under the assumption that it is proportional to the transcript frequencies. This 'digital' approach has become very popular (reviewed in [1]); it has the advantage that the expression data for each gene in every tissue tested can be stored in easily accessible databases, once and for all. The detection of mRNA does not depend on a common control, in contrast to the popular microarray protocol [7], nor does it require access to actual tissue samples. In addition, the EST approach allows the detection of novel genes (or splice variants) expressed in a given sample. As we do not know in advance which genes are to be found expressed in the heart, EST data is well suited to our project.

As cardiovascular diseases are the first cause of death worldwide, the analysis of cardiac genes is of major interest. In this study we have identified genes differentially expressed in the human heart, using their EST frequency in various human adult libraries. We followed a four-step protocol: first, the ESTs detected in the various heart cDNA libraries were assembled in separate contigs, representing genes expressed in the heart; second, we calculated the probability of differential expression in the heart for each contig, using our previously published statistical test [8]; third, we computed the pairwise correlation of the expression profiles of the contigs found to be differentially expressed in the heart; fourth, we clustered the contigs according to their coexpression level, and represented them as a dendrogram.

On the basis of this approach, we identified a set of new genes specifically overexpressed in the heart and with no previously recognized cardiac function. By locating these genes in the human genome sequence, we identified candidate genes for cardiovascular diseases and gathered the first collection of cardiac gene promoter sequences. The analysis of this collection for the first time suggests a pattern of regulatory elements that appear to characterize the promoters of at least a subset of heart-specific genes.

Results

Heart expression of contigs

We first generated contigs representing genes expressed in the heart, as described in Materials and methods. Cardiac ESTs were grouped into 194 clusters, from which 220 contigs were constructed. Such contigs represent heart transcripts.

Heart-specific expression of contigs

We then identified contigs specifically expressed in the heart. Contigs were classified as differentially expressed in the heart on the basis of our previously published statistical test [8]. This test calculates the probability for a gene to be equally expressed in two different conditions given the observed distribution of tag counts. Small probabilities (p-values) are thus associated with asymmetrical distributions characterizing differentially expressed genes. The probability of differential expression was here computed between cardiac versus non-cardiac libraries. Contigs associated with p-values < 0.03 of being expressed at the same level in the two pools were classified as differentially expressed in the heart (see Materials and methods). Of the 220 contigs, 68 were defined as significantly overexpressed in the heart (Table 1), no underexpressed contigs were found.

The expression profile was then derived for each of the 68 contigs. For each of the 438 EST libraries, the contig expression level was computed as the fraction of ESTs matching the contig, relative to the total number of ESTs in the library.

Table 1

Statistics on EST clusters and contigs

Human heart ESTs	4,303
Human heart ESTs cleaned	4,115
Clusters	194
Contigs	223
Contigs cleaned	220
Contigs extended	220
Contigs differentially expressed	68
Genes differentially expressed = unique contigs	39

The final gene-expression data was thus transformed in a 68 contigs x 438 libraries matrix. This matrix was the basis of all subsequent computations.

Coexpression of contigs and dendrogram representation

Coexpression of contigs

To identify coexpressed contigs, that is contigs exhibiting a similar expression pattern, we looked for significant similarities between their expression profiles, that is, between the rows of the 68 x 438 matrix described above. The pairwise Pearson linear correlation coefficient was computed between all pairs of expression profiles resulting into a 68 x 68 symmetrical matrix. The coefficient ranges from -1 (opposite pattern) to +1 (fully correlated pattern). Near-zero values indicate no association. The coexpression between the 68 genes was displayed in a dendrogram (Figure 1) built from the 68 x 68 correlation matrix (see Materials and methods). The branch length between two contigs reflects their distance in terms of expression pattern; thus highly coexpressed contigs are close to one another in the dendrogram.

Function of contigs

In the dendrogram, the putative functions of the contigs were annotated according to their best match in GenBank release 128.0 (see Materials and methods). When a high similarity was found between a contig and a GenBank entry, we assigned the describe function to the contig. 'X' was used to denote the absence of known function. NHF (no hits found) was used to denote the absence, or low quality, of matches. Then, cross-comparison of contigs was used to eliminate those originating from the same genes. As expected, these contigs were found very close to each other on the dendrogram, thus serving as internal controls for our study. The elimination of these duplicates resulted in a final total of 39 unique contigs, qualified as 'genes' (Figure 1, Table 2).

Analysis of function

Among these genes, 15 corresponded to functions expected to be highly expressed in the heart, such as NADH dehydrogenase ubiquinone (NM_002489) for energy production, myosin (XM_027060, XM_033374, XM_032189 and XM_004995), tropomyosin (NM_000366) and actin (BC009978), involved in muscle contraction, and to proteins such as colligin (NM_001235) that interact with the abundant heart collagen. Thirteen of the remaining genes exhibited no obvious relationship to the physiology of the heart. Finally, 11 genes had no functional attribute (noted as X, NHF, KIAA or HSP).

Analysis of the dendrogram

Interestingly, most of the genes with related functions clustered together in the dendrogram. Four clusters were obvious: one of contractile proteins (isoforms of tropomyosin - TG72_14, TG114, TG131 - and myosin TG132_7), a second of troponin isoforms (TG13 and TG154_9), a third of genes without a match in GenBank (NHF), and a fourth due to contamination by *Escherichia coli* sequences. These vector sequences were not masked in the previous step because they were not included in the RepeatMasker and RepBase databases. These four *E. coli* sequences were removed from further consideration.

Muscle contribution of cardiac-specific genes

The goal of this study is to identify genes specifically related to the heart physiology. However, this organ is a muscle, and is likely to share similar gene expression with other muscles. The protocol we used to identify heart-specific genes involved the comparison of EST frequencies computed in the merged heart libraries versus the frequencies computed in all other tissue types. The later pool included muscle libraries, but their contribution was diluted with other tissue types. It was thus possible that genes generally overexpressed in muscles were not eliminated in the initial identification of the heart genes. To address this problem, muscle-specific genes were identified by comparison of EST frequencies between all non-muscular tissues versus non-cardiac muscle libraries. Using the same *p*-value threshold of 0.03, 1,156 contigs were found overexpressed in muscle. Five of the previously identified heart-specific genes were found within this list (Figure 1). Unsurprisingly, they all corresponded to contractile proteins and four of them clustered together on the dendrogram.

Chromosomal localization

To determine the genomic location of the 35 remaining candidate heart-specific genes, they were used as query sequences to search the human genome sequence. A cognate match was found for 32/35 (89%) of the genes. As expected, most of the genes had several partial matches in close proximity along the human genome, separated by some hundreds of nucleotides, corresponding to introns. Most of these genes have been previously linked to a specific chromosome by different techniques. To avoid mapping our candidate genes to potential pseudogenes, we only retained matches on the specific chromosome. The absence of matches can be attributed to the incomplete status of the human genome sequence and/or to the difficulty of identifying short segments of similarity spread over long genomic regions.

We then searched for potential links with cardiovascular diseases. We compared the location of our heart-specific genes to the genetic loci of monogenic diseases associated with cardiovascular defects using Online Mendelian Inheritance in Man (OMIM) [9]. Among the 32 genes, five had been previously reported as responsible for such a disease. We also identified a gene within a locus of a bleeding disorder (Modifier of von Willebrand factor, OMIM: 601628). On the dendrogram, this gene is close to two of the genes responsible for cardiovascular diseases (Table 3). About 300 diseases with cardiovascular symptoms are known. If we assume that each of these diseases is due to just one gene, then one gene in a hundred is involved

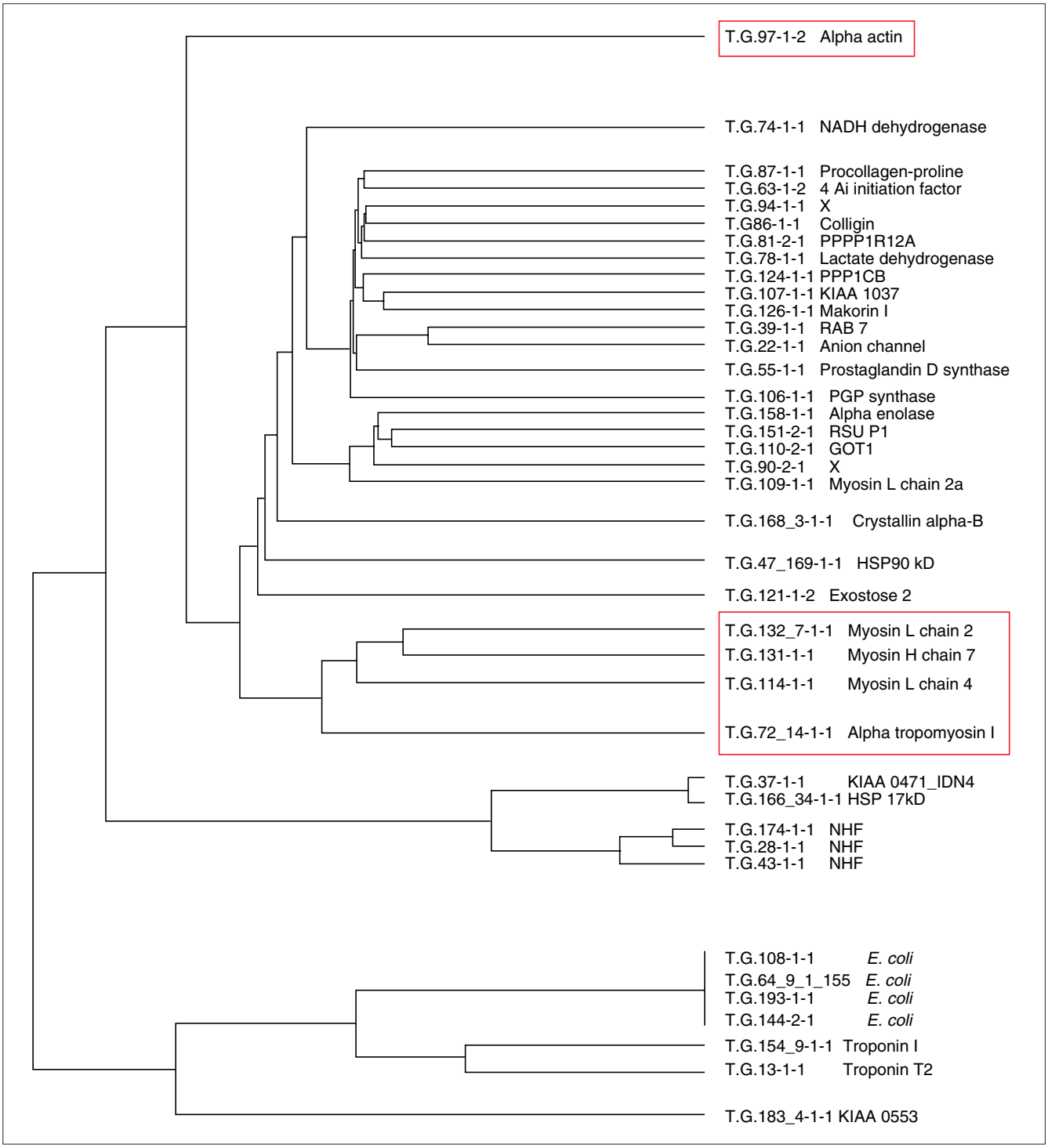

Figure 1
Dendrogram of correlation in the expression of 'heart-specific' genes. Each gene is identified by a number (see Table 3) and its function. X denotes an unknown function, NHF (no hits found) denotes genes with no similar sequence in GenBank, *E. coli* denotes probable *E. coli* sequence contamination. Genes exhibiting significant overexpression in both heart and generic muscle libraries are boxed in red.

in a cardiovascular disease (300 cardiovascular diseases over the 30,000 predicted genes of the human genome). Out of 32 genes, we found six genes, one being a candidate, associated with cardiovascular disorders. There is thus a low probability that the candidate gene is located by chance near a locus associated with a bleeding disorder.

Table 2

Accession number in GenBank and function of the heart-specific genes from Figure 1

Gene name in Figure 1	Accession number in GenBank	Function	Heart physiology relation
TG131	XM_033374	Myosin heavy chain 7 (MYH 7)	Expected
TG132_7	XM_027060	Myosin light chain 2 (MYL 2)	Expected
TG109	XM_004995	Myosin light chain 2a (MYL 2a)	Expected
TG114	XM_032189	Myosin light chain polypeptide 4 (MYL 4)	Expected
TG72	NM_000366	Alpha-tropomyosin 1	Expected
TG97	BC009978	Alpha-actin (ACTC)	Expected
TG13_8	X74819	Troponin T2	Expected
TG154	XM_012849	Troponin I	Expected
TG86	NM_001235	Colligin	Expected
TG166_34	U15590	HSP 17 kD	Expected
TG47_169	D87666	HSP 90 kD	Expected
TG78	NM_005566	Lactate dehydrogenase	Expected
TG158	NM_001428	Alpha-enolase 1 (ENO1)	Expected
TG22	L06132	Voltage-dependent anion channel 1 (VDAC-1)	Expected
TG74	NM_002489	NADH dehydrogenase ubiquinone 1, alpha subcomplex, 4	Expected
TG81	NM_002709	Protein phosphatase 1, regulatory (inhibitor) subunit (PPP1 R)	Not obvious
TG124	X80910	Protein phosphatase 1, catalytic subunit - beta isoform (PPP1 CB)	Not obvious
TG87	BC010859	Procollagen-proline, 2-oxoglutarate 4-dioxygenase	Not obvious
TG106	NM_024419	Phosphatidyl glycerophosphate synthase	Not obvious
TG110	XM_005783	Glutamic oxaloacetic transaminase (GOT1)	**Not obvious**
TG55	XM_038059	Prostaglandine D synthase	**Not obvious**
TG63	XM_029592	Elongation factor eiF 4A	**Not obvious**
TG121	XM_031239	Exostose 2 (EXT2)	Not obvious
TG3_168	BC007008	Crystallin alpha B	**Not obvious**
TG90	AF151904	'CGI 146 protein'	Not obvious
TG126	XM_035119	Makorin 1	**Not obvious**
TG151	L12535	RSU-1	**Not obvious**
TG39	BC008721	RAB-7	**Not obvious**
TG107	NM_015023	KIAA1037	**New**
TG183_4	XM_045981	KIAA0553	**New**
TG37	NM_014857	KIAA0471_IDN4	New
TG94	AL356299	X	New

Relationship between gene names in Figure 1, GenBank references and functions. The relationship with heart physiology is denoted 'expected' when the gene is known to be widely expressed in the heart, 'not obvious' when the gene is not known to be widely expressed in the heart, and 'new' when the gene has no known function. Genes in bold were experimentally validated later and were found to be heart-specific.

Collection and analysis of promoter sequences

As the 32 genes all appear (statistically) to be specifically expressed in the heart, one might suspect that they share some regulatory elements. Our final step was to analyze their regulatory regions. Core promoter regions were operationally defined as the 1,000 bases upstream of the transcription start site (TSS), if known, up to the end of the 5′-UTR (that is, to the site of translation initiation). The core promoter regions of 17 cardiac genes with known TSSs were extracted. This collection of promoters is the first database of heart promoter sequences and is available at [10]. Three types of regulatory element were searched for: polymerase II promoter elements, known transcription factor sites and new motifs common to most of these promoter sequences.

Core promoter regions were analyzed by first locating the polymerase II promoter elements using Tfbind [11] with the TRANSFAC matrix V$CAAT, V$CAP, V$TATA and V$GC. Known transcription factor sites were searched for using MatInspector [12] and only those common to the 17 promoter sequences were considered. New motifs were searched for using MEME and two complex combinations of motifs common to five promoter sequences were detected. The combination pattern and location of the detected

Table 3

Genes linked to monogenic diseases with associated cardiovascular defects

Gene name	OMIM reference for the disease	Identification of the responsible gene
TG132	160781 (heart formation defect)	+
TG154	191044 (cardiomyopathy)	+
TG131	160760 (cardiomyopathy)	+
TG13	191045 (cardiomyopathy)	+
TG78	150000 (enzyme deficiency)	+
TG4_183	601628 (bleeding disorder)	- Modifier of von Willebrand factor

Gene names are from Figure 1. The cognate diseases and their references in OMIM are noted. +, Previously known relationship; -, unknown relationship (candidate gene).

elements are shown in Figure 2. Over-represented oligonucleotides ('words') were also identified using RSA-tools [13] and are listed in Table 4. None of the over-represented word or MEME combination corresponds to previously known patterns and may define new regulatory elements. Interestingly, we noticed that the combination of at least three elements out of five (GKLF, Tcf11_Rora, Tcf11_AP1C and both MEME combinations) were present in seven promoter sequences, five of them being associated with genes known to be highly expressed in the heart.

Discussion

The present analysis of cardiac ESTs identified 35 genes as being differentially expressed in the heart. After clustering these genes on the basis of the correlation of their expression profiles, genes with known related function appeared as close neighbors on the resulting tree. They might thus share regulatory regions. The initial analysis was done with a dbEST release of September 1999. To ensure the specificity of the originally identified 32 heart-specific genes, we revalidated them in the light of the most recent dbEST release (February 2002).

Studies based on ESTs and on variation in expression require rigorous statistical validation. As the identification of changes in expression level on the basis of quotients of very small relative abundance is not very meaningful, many methods have been developed to evaluate variation in expression level (see [1] for a review). We used a previously published test [8] estimating the probability of differential expression for a gene between two pools of ESTs (cardiac versus non-cardiac) and able to detect a weak differential expression provided the absolute number of tag counts is large enough. Assuming a probabilistic model, this test calculates the probability of observing y ESTs in library B given that we observed x ESTs in library A, a low probability indicating a high differential expression of the cognate gene over the two libraries. This test was independently validated by others and, in a comparative analysis of statistical tests evaluating differential gene expression, it was found to be the most appropriate for pairwise comparisons of EST libraries [14].

Our approach suffers from several obvious limitations, shared by all EST-based analyses. First, mRNA level does not always correlate with protein abundance in the cell; thus the EST analyses are not representative of the proteome of a cell. However, these problems also apply to the more expensive and sophisticated microarray techniques. Second, the abundance of transcripts detected depends on the initial EST number: starting with 4,303 cardiac ESTs, only highly expressed genes are expected to show up in our final list of genes. Nevertheless, highly expressed genes are expected to have a significant impact on the physiology and pathology of the heart. Unlike other studies [15,16], we removed from our study genes represented by a single EST (singletons), thus decreasing the overall number of ESTs in consideration, but considerably reducing the danger of taking in the artifacts induced by this unreliable data.

Using an EST approach, Hwang *et al.* [16] characterized gene transcription and identified genes overexpressed in cardiac hypertrophy. They generated about 77,000 ESTs, half of them corresponding to 5,000 unique known genes and expressed in the heart. A large fraction of those genes may be represented by very-low-copy ESTs (singletons) that may arise from tissue contamination (for example by blood). Moreover, this count cannot be used to estimate the expression level of cardiac genes because the EST frequency is not given. A larger number of ESTs would thus be required to increase the sensitivity of our study. Although more cardiac ESTs were recently generated, none of them was used in our study because they were all generated from normalized libraries.

Protocols using EST numbers to estimate gene-expression levels were successfully used in previous studies [17-19]. For instance, Bortoluzzi *et al.* [19] reconstructed the human adult skeletal muscle transcriptional profile where they found a good agreement between their results and a SAGE (serial analysis of gene expression) experiment.

We compared our results to previous EST analyses of the heart transcriptome [15,16,20]. Each cardiac gene (except TG106, TG110, TG90 and TG94) given in our list was found in at least one of these studies. We also noticed that these lists are not exactly similar and that some genes were detected in one study only.

We also compared our results to experimental studies of the human transcriptome. We considered three analyses, two based on EST counting [18,19] and one based on microarray data [21]. For the 15 genes found to be specifically expressed

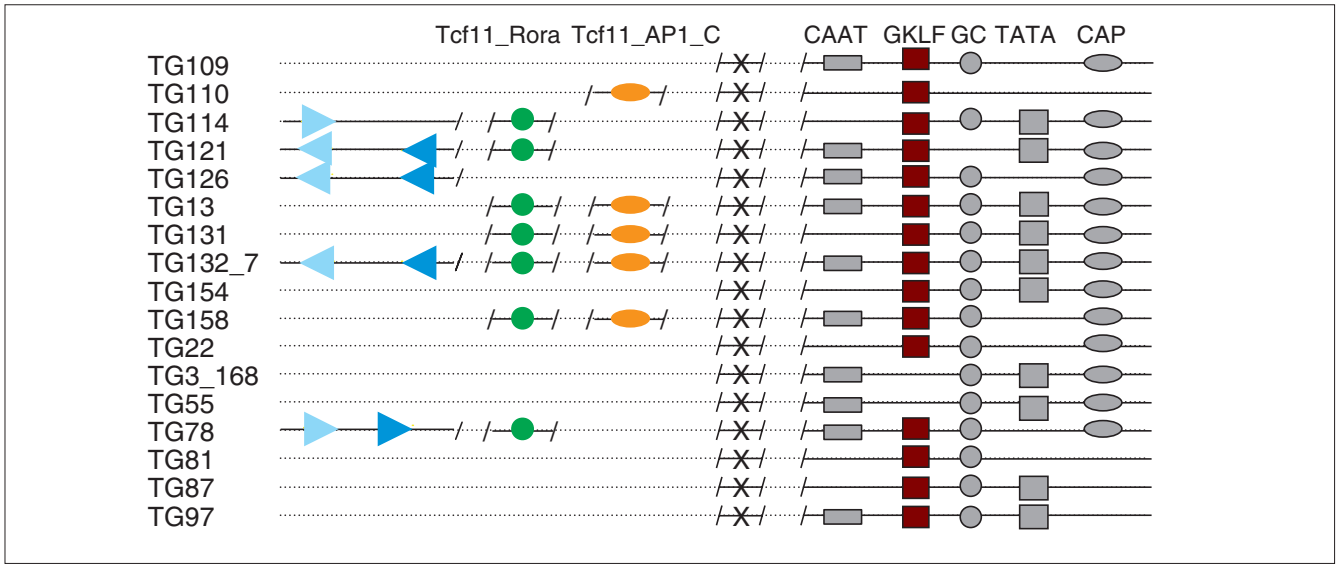

Figure 2
Topology of the putative regulatory elements identified in 17 cardiac promoters. Gray boxes are polymerase II promoter elements at correct locations. TRANSFAC transcription factor sites and MEME motifs (arrows) are colored. Eight transcription factor sites are common to all sequences and their combination is represented by X. They are AP-2 and AP-4 (activator protein), SP1 (stimulating protein), NFAT (nuclear factor of activated T cells), the GC box, NF1_Q6 (nuclear factor of activated T cells), TCF11 (zinc-finger protein) and MZF1 (myeloid zinc finger). Brown squares are a transcription factor site (for GKLF (gut-enriched Krüppel-like factor)) located at the same position (-70) in at least 15 promoters. Green and orange symbols are couples of motifs that are separated by a constant width and are present in the maximum number of promoter sequences (seven and eight respectively). Both couples involve the TCF11 zinc finger with RORA (RAR-related orphan receptor alpha1) or AP1_C (activator protein 1). Blue arrows are MEME combinations of motifs. Light-blue arrows are the combinations of motifs M2-M1-M5 (M2 and M1 on the strand S+ and M5 on the strand S-, or vice versa). Dark-blue arrows are the combination of motifs M1-M3-M2-M6 (M1 and M2 on S+ and M3 and M6 on S-, or vice versa). Consensus sequences of the motifs are:
M1: TC[CTA][TG]G[GA]G[TG]CTGAGG[GC][AG][GA]GAGG[AG][GT][CT][TG][GC][GT][GT]G;
M2: [AC][CG][TA]CCAGCCTGGG[CT][GA]ACA[GT]AG[CT][GA]A[GA]A[CT]CC[TC][GT]T[CT]T[CT][AT]A[AC][AT]AAAA[AC][AT]AA[AT]AA;
M3: GTGG[CT]G[CT]GATC[AT][CT]GGCTCACTGCA[GA]CCTC[GC][AG]CCT[CT]C[CT]G;
M5: GCTGGGA[TC]TACAGGC[GA]TG[AC]GCCAC[CG][GA][TC]GC;
M6: TTT[TAG][AT][TA]TT[AT][AT]TT[TC]TT[TC][CA][TG][TC]T[TA][TAG][CT][TA][CT][TA][TG][AC]C.
The figure is not to scale.

in the heart and with no previously known cardiac specificity, nine were found upregulated in the heart by at least one of these studies. Three were not found to be specific to the heart, but as they have a function related to contractile proteins (myosin regulation or similarity with a protein binding actin), their expression in the heart is not surprising. Three others were not found to be heart-specific in any of these studies (Table 5).

We generated genes starting from ESTs found in cardiac libraries, thus focusing on genes expressed in the heart. In general, these genes can be expressed in other tissues as well. For example, genes involved in energy metabolism and in contraction are highly expressed in the heart, as well as in muscle and sperm [22]. Those cell types require energy and involve muscle contraction or motility. As the heart is a vascular and a muscular organ, identifying 'truly' heart-specific genes (that is, those involved in vascular function) requires elimination of genes specifically overexpressed in generic muscular tissues. Some genes encode proteins with multiple splice isoforms (such as the contractile proteins); some of these are specific to skeletal muscle and others to heart muscle. Such genes were analyzed more precisely to find out exactly which isoform was involved. The isoforms found in this study are known to be expressed in the heart or in all muscle cells: no isoforms specific to skeletal muscle were found.

We identified 35 genes likely to be involved in heart-specific functions, either vascular or neuro-muscular. Five were previously known as disease genes with cardiovascular symptoms, and one gene, clustering close to other disease-linked genes in the dendrogram, lies near a locus associated with a bleeding disorder (OMIM: 60162). This direct validation of our approach allows us to propose that the remaining heart-specific genes identified in this study might be of biomedical interest. Such genes are candidates for further linkage analysis.

Through a computational analysis of the promoter regions of these cardiac genes, we identified a combination of five main motifs that could participate in their specific expression. These motifs should not be considered as individual

Table 4

Over-represented 'words' in the 17 cardiac promoters

4 bp	5 bp	6 bp	7 bp	8 bp
tttt	ttttt	tttttt	ttttttt	tttttttt
aaaa	aaaaa	aaaaaa	aaaaaaa	aaaaaaaa
tggg	agctg	gaagct	gtaccag	tgttaggc
gctg	tgcct	ttcttg	cccaagt	tctgaggg
ctgg	ggtgg	tgtttt	ttgtttt	aggtggcc
gtgg	ttttg	atatat		cccaagtt
ttgg	tggct	tttttg		gaccctga
ctct	ctgag	tcttga		gcctctgc
tgtg	tcttg	agctgg		agctggag
ggct	gctgg	tgtcat		ggggtggg
cttg	gaatg	ttcttg		
ggag	tgggg	agctgg		
tgag	tgtcc	atatat		
agct	tgggt	tcttga		
gctc	ggctc	cagctg		
gcct	gattc	ttgttt		
ctgt				
tgtt				
cccc				
tctg				
gggg				
tggc				
cctc				
cctg				
tttg				
tgct				
gcag				
tcct				
tctc				
tggt				
tgtc				
tgga				
gtgt				

Table 5

Expression in the heart of the 15 genes without any previous known cardiac function

Gene name and function	Gen_Card (EST)	UniGene (EST)	Microarrays
KIAA 1037	+++	+++	0
KIAA 0553	+++	+++	0
Crystallin alpha B	+++	+	ND
Makorin I	++	++	ND
RAB 7	++	++	0
RSU - RSP I	++	+	ND
Initiation factor	+	+++	ND
Prostaglandin D synthase	+	+++	+++
GOT I	+	+	+++
PPPP1-CB (Regulation of myosin)	+	+	ND
PPPP1 -R12A (Regulation of myosin)	+	+	ND
KIAA 0471_IDN4 (Similar to the actin-binding protein of the fly)	0	-	-
Exostose II	+	+	ND
Procollagen proline-2 oxoglutarate-4-dioxygenase	+	+	0
Phosphoglycerophosphate synthase	-	-	ND

+++, Upregulation in the heart. 0, Regular expression in the heart (neither downregulated nor upregulated). -, Weak expression in the heart. ND, no data. Heart-specific genes according to Gen-Card, UniGene and/or microarray resources are shown in bold.

elements but as a module of organized elements (in position, order and/or distance) controlling gene expression [23]. As all the genes specifically expressed in a tissue are not expected to be regulated by the same elements, the combination revealed in this study may be involved in the regulation of a subset of cardiac genes. Of 17 genes, five have this combination in their promoters and may constitute this subset, or a part of this subset. The combination was searched for in the human sequences in the Eukaryotic Promoter Database (EPD). Eighteen promoters from EPD have this combination, half of them being promoters of cardiac genes and thus co-regulated with the genes identified in this study. Others may be promoters of genes whose cardiac activity has not previously been recognized.

Homologous genes in related organisms often share the same regulatory elements. As most of the mouse genome is now available, the motif combination was searched for in the promoters of mouse orthologs. No identical combination of elements was found in the orthologous mouse promoter regions, probably because of evolutionary divergence. The absence of identifiable conservation of these motifs suggests that they may be involved in the fine regulation of a subset of cardiac genes rather than in 'constitutive' cardiac expression.

Materials and methods

EST databases and contigs

Human ESTs were extracted from dbEST (September 1999) [24]. To avoid variation due to disease or development stages we eliminated libraries from diseased, fetal or infant tissues, and libraries pooled, subtracted or normalized. Libraries were eliminated by parsing their description field for the terms: 'sclerosis', 'cancer', 'carcinom', 'tumor', 'tumour', sarcom', 'leukemi', 'melanom', 'lymphom', 'fetal', 'foetal', 'fetus', 'foetus', 'infant', 'neonat', 'post natal', 'placenta', 'umbilical', 'embryo', 'subtracted', 'normaliz' and 'normalis'. ESTs were then cleaned (vector removal, elimination of low-complexity sequences) with RepeatMasker (A.F.A. Smith and P. Green,

unpublished results) and RepBase [25]. After these steps 309,904 ESTs remained. A set of 4,303 'cardiac' ESTs was extracted from this total set, representing 4% of the total human heart ESTs (Table 6). ESTs sharing stretches of high sequence identity (score > 40 and e-value < 10^{-5}) were grouped into 194 clusters using BLAST 2.0.8 [26]. From each cluster, contigs of overlapping ESTs were built using CAP1 [27] and extended using other libraries in an iterative manner; 220 contigs were obtained. ESTs without a match (singletons) after the whole procedure were discarded. Given that the assembly of an EST cluster could result in several consensus sequences, we obtained more contigs than clusters. As one gene corresponds to one cluster, some genes could initially be represented by more than one contig. These 'duplicate' cases served as internal controls (related contigs appearing as coexpressed) and were eliminated at the final stage of our analysis.

Differential gene expression

To assess its differential expression, every contig was compared to the total EST set for high stringency (*P*-value < 10^{-20}). For each contig, the hit list of cognate matches was then separated into two groups: ESTs from cardiac libraries versus any other libraries. Given the count of cognate ESTs and the total number of ESTs in both groups, we evaluated the differential expression of the contig in heart.

The statistical significance of the difference in frequencies (x/N_1, y/N_2) between these two groups was computed according to [8]:

$$Pvalue(x,y) = \sum_{Y=y}^{\infty} p(x/Y)$$

with:

$$p(x/y) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}}$$

where x and y are the numbers of ESTs, respectively, from cardiac libraries and from others matching the contig, and N_1 and N_2 are the total numbers of ESTs, respectively, from cardiac libraries and from others. A *p*-value < 3% threshold was used to classify contigs as selectively expressed in the heart. This limit was chosen as a compromise between the predicted rate of false positives (0.03 x 220 = 6.6) and retaining enough potential candidates for the remainder of the study. The *p*-value here is not taken as a way to assess the actual statistical significance of the result, but as the rational way to prioritize our study [12].

Expression profile and correlation

The expression profile of each contig was computed from the number of cognate ESTs in each library constituting the total EST set relative to the total number of ESTs in the library. All expression profiles were stored in a matrix with rows corresponding to contigs and columns corresponding to libraries. Element M_{ij} of the matrix corresponds to the relative frequency of cognate ESTs for contig *i* in library *j*.

The similarity of expression profile between contigs was estimated by computing the value of Pearson's *r* coefficient in a pairwise manner between each row. This coefficient takes values within the [-1, +1] range. Values close to 0 indicate no correlation, positive values denote a positive correlation (contigs going up and down together), and negative values denote opposite patterns of contig expression. A matrix of pairwise gene distances was then derived from the correlation matrix.

Distance matrix and tree representation

The hierarchical classification of objects requires the calculation of a matrix of their pairwise distances. The contig correlation matrix constructed previously was turned into

Table 6

Human heart EST libraries represented in dbEST

dbEST library identifier	Number of ESTs	Library description
25	371	Human heart cDNA library
88	1,739	Human heart cDNA library
89	653	Atrium cDNA library human heart
218	227	Atrium cDNA library human heart
246	1,041	Adult heart
276	1,935	Fetal heart
285	16,946	Fetal heart
424	38,556	Fetal heart normalized
469	2,716	Human heart cDNA normalized
589	43,460	Pooled human melanocyte, fetal heart, and pregnant uterus
653	20	Pooled human melanocyte, fetal heart, and pregnant uterus
674	47	Fetal heart
680	1,878	Fetal heart
717	85	Adult heart, male 25 years
784	44	Adult heart, subtracted
799	5	Fetal heart, subtracted
847	111	Fetal heart
848	163	Adult heart
1063	10	Fetal heart + brain + liver
1570	4	Adult heart muscle
1763	3	Fetal heart

For each library, the dbEST identifier is shown with a short description from dbEST, if available, and the number of ESTs. Libraries in bold are normalized, subtracted, pooled or from infant, fetal or disease tissues and were eliminated from the study. Data as of October 1999.

such a distance matrix by computing the Euclidean distance *d* between genes X and Y from the columns of the correlation matrix:

$$d(X,Y) = \sqrt{\sum_{i=0}^{i=N} (x_i - y_i)^2}$$

The resulting distance matrix was then displayed as a dendrogram, using the UPGMA (unweighted pair group method with arithmetic means) algorithm [28], as implemented in the Neighbor program [29].

Identification of homologous sequences

Contigs were functionally annotated by querying GenBank (release 128.0) with the program BLAST, version 2.0.8. Cognate matches were initially identified using a threshold of 98% sequence identity, followed by an extensive bibliographical analysis of the matching entry.

To locate the genes on the human genome, we compared their sequences to the human draft, daily updated, and available online at the National Center for Biotechnology Information (NCBI) [30]. If 70% of the query length matched the genomic sequence with a score > 200, the query sequence was considered as being successfully located in the human genome. As genes were built from ESTs, they represent gene transcripts. As expected, many of the genomic matches were found to be separated by intron sequences.

Collection and analysis of promoter sequences

We further analyzed the putative promoter/regulatory sequences of the candidate genes for which a transcription start site (TSS) was previously identified. Seventeen core promoter regions were thus extracted from the human genome sequence assembly as the 1,000 bases upstream of the TSS and the 5′-UTR. These sequences constitute the first collection of putative heart specific promoters and are available at [10]. We analyzed these regulatory regions by locating polymerase II promoter elements and by searching for known and new regulatory elements common to all the sequences. As we were searching for new elements, we did not consider the known cardiac sites (such as GATA and Sp1).

Polymerase II promoter elements were determined with Tfbind [31], considering the TRANSFAC matrix V$CAAT, V$GC, V$TATA, and V$CAP only. We only retained the matches found at the expected locations: in the -105 to -70 region for the CAAT box, in the -74 to -45 region for the GC box, in the -20 to -30 region for the TATA box, and in the -5 to +5 region for the CAP box; +1 being the transcription start site (TSS). Known transcription factor sites (limited to the vertebrate matrix group) were searched for using MatInspector and the TRANSFAC database [32]. Default parameters were used.

Over-represented words were identified using RSA-tools [33]. Oligonucleotides (4 to 8 residues in size) were counted on both strands, to detect orientation-insensitive elements. The expected frequency was calculated from the human promoters of the EPD release 70 [34,35]. Only slight differences were observed when changing pseudo-weights from 0.10 to 0.20. Other parameters were kept to their default values.

MEME version 3.0 [36,37] was used to reveal new motifs that might be common to all the promoter sequences. Motifs with a number of sites between 2 and 300 and a width from 6 to 50 nucleotides were searched for.

The EPD database release 71 [33] was used to get the background sequences.

Acknowledgements

K.M. was supported by a grant from AVENTIS Pharma and the region Provence-Alpes-Côte d'Azur. We thank F. Gosse, C. Notredame, H. Ogata and K. Suhre for critically reading the manuscript.

References

1. Claverie JM: **Computational methods for the identification of differential and coordinated gene expression.** *Hum Mol Genet* 1999, **8:**1821-1832.
2. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale.** *Genome Res* 1998, **8:**1202-1215.
3. Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in human liver.** *Electrophoresis* 1997, **18:**533-537.
4. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19:**1720-1730.
5. Kurian KM, Watson CJ, Wyllie AH: **DNA chip technology.** *J Pathol* 1999, **187:**267-271.
6. Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K: **Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression.** *Nat Genet* 1992, **2:**173-179.
7. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
8. Audic S, Claverie JM: **The significance of digital gene expression profiles**. *Genome Res* 1997, **7:**986-995.
9. **OMIM** [http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM]
10. **Cardiac gene database** [http://igs-server.cnrs-mrs.fr/Card_Gene/]
11. Tsunoda T, Takagi T: **Estimating transcription factor bindability on DNA.** *Bioinformatics* 1999, **15:**622-630.
12. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation**. *Nucleic Acids Res* 2000, **28:**316-319.
13. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281:**827-842.
14. Romualdi C, Bortoluzzi S, Danieli GA: **Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests.** *Hum Mol Genet* 2001, **10:**2133-2141.
15. Bortoluzzi S, d'Alessi F, Danieli GA: **A computational reconstruction of the adult human heart transcriptional profile.** *J Mol Cell Cardiol* 2000, **32:**1931-1938.
16. Hwang DM, Dempsey AA, Lee CY, Liew CC: **Identification of differentially expressed genes in cardiac hypertrophy by analysis of expressed sequence tags**. *Genomics* 2000, **66:**1-14.

17. Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, Hinzmann B, Rosenthal A: **Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues**. *Nucleic Acids Res* 1999, **27:**4251-4260.
18. Stekel DJ, Git Y, Falciani F: **The comparison of gene expression from multiple cDNA libraries.** *Genome Res* 2000, **10:**2055-2061.
19. Bortoluzzi S, d'Alessi F, Romualdi C, Danieli GA: **The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach.** *Genome Res* 2000, **10:**344-349.
20. Liew CC, Hwang DM, Fung YW, Laurenssen C, Cukerman E, Tsui S, Lee CY: **A catalogue of genes in the cardiovascular system as identified by expressed sequence tags.** *Proc Natl Acad Sci USA* 1994, **91:**10645-10649.
21. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, *et al.*: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99:**4465-4470.
22. Aydin S, Yilmaz Y, Odabas O, Sekeroglu R, Tarakcioglu M, Atilla MK: **A further study of seminal plasma: lactate dehydrogenase and lactate dehydrogenase-X activities and diluted semen absorbance.** *Eur J Clin Chem Clin Biochem* 1997, **35:**261-264.
23. Fessele S, Maier H, Zischek C, Nelson PJ, Werner T: **Regulatory context is a crucial part of gene function.** *Trends Genet* 2002, **18:**60-63.
24. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST - database for "expressed sequence tags"**. *Nat Genet* 1993, **4:**332-333.
25. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16:**418-420.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
27. Huang X: **A contig assembly program based on sensitive detection of fragment overlaps.** *Genomics* 1992, **14:**18-25.
28. Sokal R, Michener C: **A statistical method for evaluating systematic relationship.** *Univ Kansas Sci Bull* 1958, **28:**1409-1438.
29. Kuhner MK, Felsenstein J: **A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates [erratum: Mol Biol Evol 1995 May;12(3):525]**. *Mol Biol Evol* 1994, **11:**459-468.
30. **Entrez Genome** [http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search]
31. **Tfbind** [http://tfbind.ims.u-tokyo.ac.jp/]
32. **TRANSFAC - the transcription factor database** [http://transfac.mirror.edu.cn/TRANSFAC/]
33. **Regulatory sequence analysis tools** [http://rsat.ulb.ac.be/rsat/]
34. Praz V, Perier R, Bonnard C, Bucher P: **The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data.** *Nucleic Acids Res* 2002, **30:**322-324.
35. **Eukaryotic promoter database** [http://www.epd.isb-sib.ch/]
36. Bailey TL, Gribskov M: **Methods and statistics for combining motif match scores.** *J Comput Biol* 1998, **5:**211-221.
37. **MEME** [http://meme.sdsc.edu/meme/website]