Genome Analysis

Regulatory context is a crucial part of gene function

Sabine Fessele, Holger Maier, Christian Zischek, Peter J. Nelson and Thomas Werner

Information about the time and place of gene transcription, which until recently was only possible by extensive experimental analysis, can now be predicted through *in silico* analysis. Using the human RANTES/CCL5 promoter, we show that organizational features of promoters derived from promoter sequences contain information about the spatial and temporal 'functional context' of expression.

Gene function has generally been thought to reside in the protein encoded by the gene. However, because most genes require a network of other genes to exert biological function, gene regulation is also a crucial part of gene function. For example, transcriptional co-regulation can ensure that interacting proteins coordinate to form a functional complex, or to ensure that a substrate is processed along a metabolic pathway.

Promoters, enhancers, locus control regions and matrix attachment regions are key components of transcriptional control [1]. Among these, promoters are unique because they integrate the information encoded by the other control elements to influence gene transcription. Promoter function in terms of temporal and/or spatial tissue-specific transcription is not obvious from the primary DNA sequence. Even co-regulated promoters often show no significant overall sequence similarity. Sets of different transcription factor binding sites are central to promoter function. Growing evidence suggests that the specificity of promoter-controlled gene regulation depends more on the relative organization of these elements within the promoter than on an individual element [2–6]. We refer to an organized group of regulatory elements as a promoter module if their transcriptional function requires the presence of all elements (where synergistic or antagonistic action has been experimentally verified) [3]. This definition is more restricted than the module definition used by Wasserman [7,8], Firulli [9] and Davidson [10], who do not consider organizational features such as distance and order. Genes expressed in the same functional context (e.g. co-regulated genes) often share promoter modules [3,4,11], and recent developments in computer modelling now allow the organization of these modular promoters to be analysed [3,4,12,13].

The crucial transcription elements that make up functional promoter modules can be detected either through *in silico* comparative promoter analysis [14,15] or experimentally. Here, we use the human RANTES/CCL5 promoter as an example of how experimentally defined elements can be used to determine information about functional context.

RANTES/CCL5 is a member of the CC- or β-subfamily of chemotactic cytokines (chemokines) and plays diverse roles in the pathology of inflammatory disease [16,17]. Inflammation (a protective reaction to tissue damage, infection or foreign substances) is a complex biological process that leads to the coordinated regulation of diverse sets of genes. The activity of the human RANTES/CCL5 promoter has been characterized by northern blotting, RNase protection assay, DNase I footprinting, EMSA, selective mutation and transient transfection of reporter constructs in a series of cell types [2,18–22] (C. Zischek and P.J. Nelson, unpublished). To a significant degree, the transcriptional control for human RANTES/CCL5 appears to be mediated through six functionally characterized short regulatory elements (Fig. 1). Not all six elements are functional in the specific cell types analysed, and the individual elements often contain overlapping binding specificities for different classes of transcription factor. In this regard, the human RANTES/CCL5 chemokine gene illustrates the flexibility and selectivity that can underlie tissue- and signal-specific regulation of gene expression.

The chemokine genes are generally found in two gene clusters. This has led to speculation that there might be coordinated regulation at both the individual and regional levels, analogous to that seen for other gene-family clusters [23]. Indeed, rather than acting as single molecules, it now appears that chemokines might be selectively co-regulated in groups that then activate common groups of chemokine receptors [16].

Figure 1 summarizes the individual control elements important for transcription of RANTES/CCL5 in five different tissue types [2,18–22]. Deletion experiments (5′–3′) with RANTES/CCL5 promoter–reporter gene constructs demonstrated that this tissue-specific regulation is encoded in <300 nucleotides [2,18–22]. The binding elements and specific transcription factors that bind to overlapping binding sites within these elements are combined differently in the five different tissues. These differences can be used to derive cell type-specific submodels *in silico* (Fig. 1). Initial models were built with information about individual binding factors (experimentally verified) and their corresponding binding sites, represented by weight matrices, were located in the sequence by computer (for reviews about weight matrix-detection methods see [24,25]). Strand orientation, relative order and distances between binding sites were determined from the human RANTES/CCL5 promoter sequence. Database searches were carried out based on the scoring algorithm of ModelInspector, which combines matrix-similarity measures of individual binding sites into a summary model score [26]. Models were initiated with default matrix-similarity thresholds, which were usually lower than the scores found with the human RANTES/CCL5 promoter sequence. All promoter modelling and subsequent database searches were carried out with the GEMS Launcher 1.0 software package, Genomatix Software GmbH, Munich (http://www.genomatix.de).

The models were used to search the human, rodent, 'other mammalian' and 'other vertebrate' sections of the EMBL Nucleotide Sequence Database of the European Bioinformatics Institute [27] (http://www.ebi.ac.uk/embl) and the

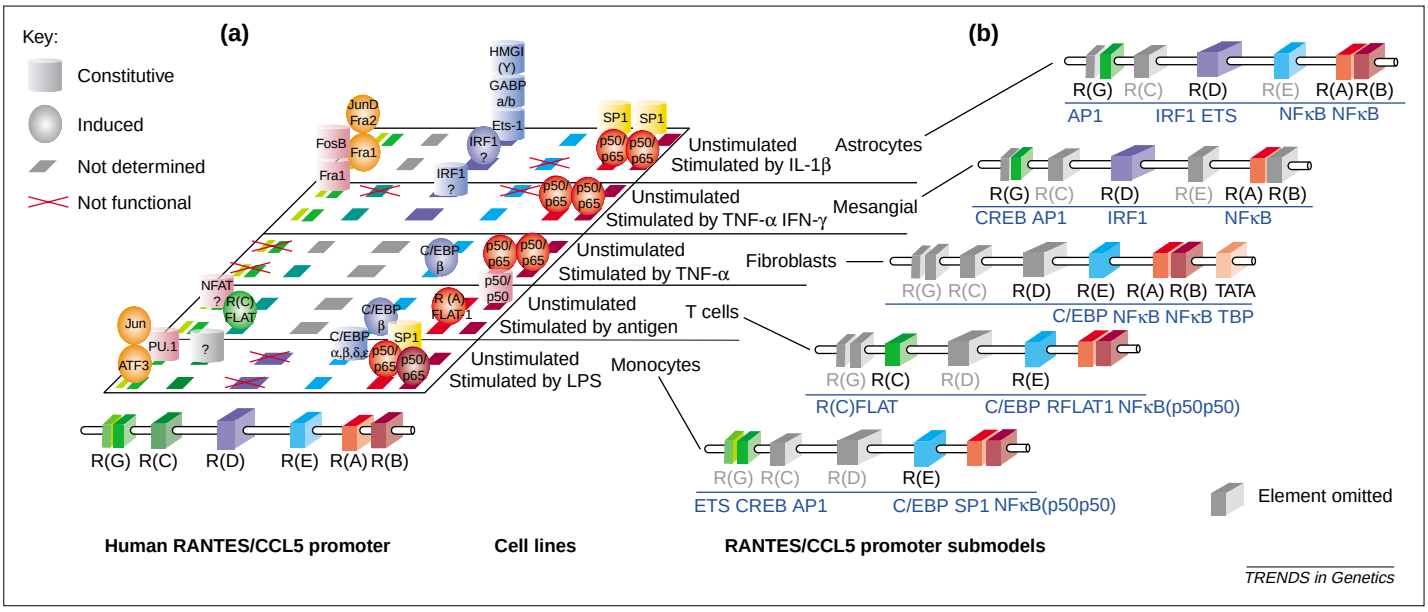

Fig. 1. Model of the functional organization of the human RANTES/CCL5 promoter. (a) Transcription factor binding and usage has been assessed experimentally in five human cell types (astrocytes, mesangial cells, fibroblasts, T cells and monocytes) under both unstimulated and stimulated conditions, and is summarized in the area above the model [2,18–22] (C. Zischek and P.J. Nelson, unpublished). The flat squares represent the binding regions from the summary model. Grey binding sites have not been assessed experimentally for protein binding under the conditions specified, binding sites crossed out were found to have no influence under the respective conditions (point mutations). Known transcription factors are indicated by their names on their respective symbols. Two symbols are used to differentiate binding under unstimulated (cylinders) and stimulated (oval spheres) conditions. (b) Partial cell type-specific organizational promoter models. Five tissue-specific submodels of the human RANTES/CCL5 promoter were derived based on experimental data [2,18–22]. The coloured elements represent the specific submodels. Grey nonfunctional elements are shown to facilitate visualization of the promoter region. Below the binding regions individual factor binding sites used in the models are shown in dark blue. Note that even models containing the same binding regions might use different binding sites. Abbreviations: ATF, activating transcription factor; C/EBP, CAAT/enhancer binding protein; CREB, cyclic AMP-response element-binding protein; GABP, GA binding protein; HMG, high-mobility group protein; IFN-γ, interferon-γ; IL-1β, interleukin-1β; IRF, interferon regulatory factor; Jun/Fos/Fra, AP1 family members; LPS, lipopolysaccharide; NFAT, nuclear factor of activated T cells; NFκB, nuclear factor kappa B (Rel family dimer); p50/p50, NFκB p50 subunit homodimer; p50/p65, NFκB Rel family members; PU.1/Ets-1, Ets family members; R(A)FLAT-1, R(A) factor of late activated T cells-1; RANTES/CCL5, regulated upon activation normal T-cell expressed and secreted/CC chemokine ligand 5; R(C)FLAT, R(C) factor of late activated T cells; R(X), RANTES region X; Sp1, stimulating protein 1; TBP, TATA-box binding protein; TNF-α, tumour necrosis factor-α.

Eukaryotic Promoter Database of the ISREC [28] (http://www.epd.isb-sib.ch). Resulting lists were manually inspected for: (1) genes known to be co-regulated with RANTES/CCL5, (2) genes correlated with inflammation in general, and (3) genes sharing up- or downstream pathways with RANTES/CCL5 (e.g. transcription factors). Such matches were used to generate small training sets of sequences (~10). These sets were used solely to increase the selectivity of the models by fine tuning the distance ranges and matrix thresholds, which does not change search results qualitatively. To facilitate evaluation, matches to the models were restricted to those occurring within annotated promoters. This search resulted in a combined total of only 53 matches. Interestingly, >70% of these promoters can be linked to either chemokine genes or genes associated with the immediate functional context of chemokines, including inflammatory mediators, signal transduction proteins and transcription factors.

Six chemokine genes (human RANTES/CCL5, human PARC, murine and rat GRO/KC, rat and murine MIP-2, murine IP-10 and murine eotaxin) were found to show a similar organization of elements. Of these, RANTES/CCL5, PARC, IP-10 and eotaxin are known to share functional context (i.e. they are expressed by similar tissues and recruit overlapping subpopulations of leukocytes) [16,17]. The biology of the remaining chemokines GRO/KC, MIP-2 and RANTES/CCL5 tells a striking story about contextual gene function. As the mouse chemokines MIP-2 and KC are human GRO homologues [16,17], comparison is restricted to RANTES/CCL5 and GRO/KC genes in Fig. 2.

Human and mouse RANTES/CCL5 are orthologous genes that show significant sequence similarity within their coding sequences (Fig. 2). However, analysis for matches to tissue-specific submodels of the human RANTES/CCL5 promoter indicates that rat and mouse GRO/KC are apparently closer in functional regulation to the human RANTES/CCL5 promoter than is the mouse orthologue of RANTES/CCL5, in spite of their lack of overall promoter sequence similarity (Fig. 2). All chemokine promoter sequences appear approximately equally distant from each other (except for the rodent GRO/KCs). Experimental results support the *in silico* analysis. In mouse, RANTES/CCL5 does not bind to the same receptors as in humans [29,30]. The activity in humans is mediated by a different set of chemokines, including members the GRO family. The correlation between mouse RANTES/CCL5 and the human GRO gene has been verified experimentally [16,29–31]. This example demonstrates that protein sequence similarity does not always guarantee functional similarity. Here, functional similarity is better described by the modular organization of the promoters than by protein-sequence similarity. A similar example of prevalence of gene regulation over protein sequence has been demonstrated with the engrailed genes, *En1* and *En2*: *En2* functionally rescues *En1*-knockout mutants if placed under the control of the *En1* promoter, even though *En1* and *En2* have distinct functions in brain development in wild-type mice [32,33].

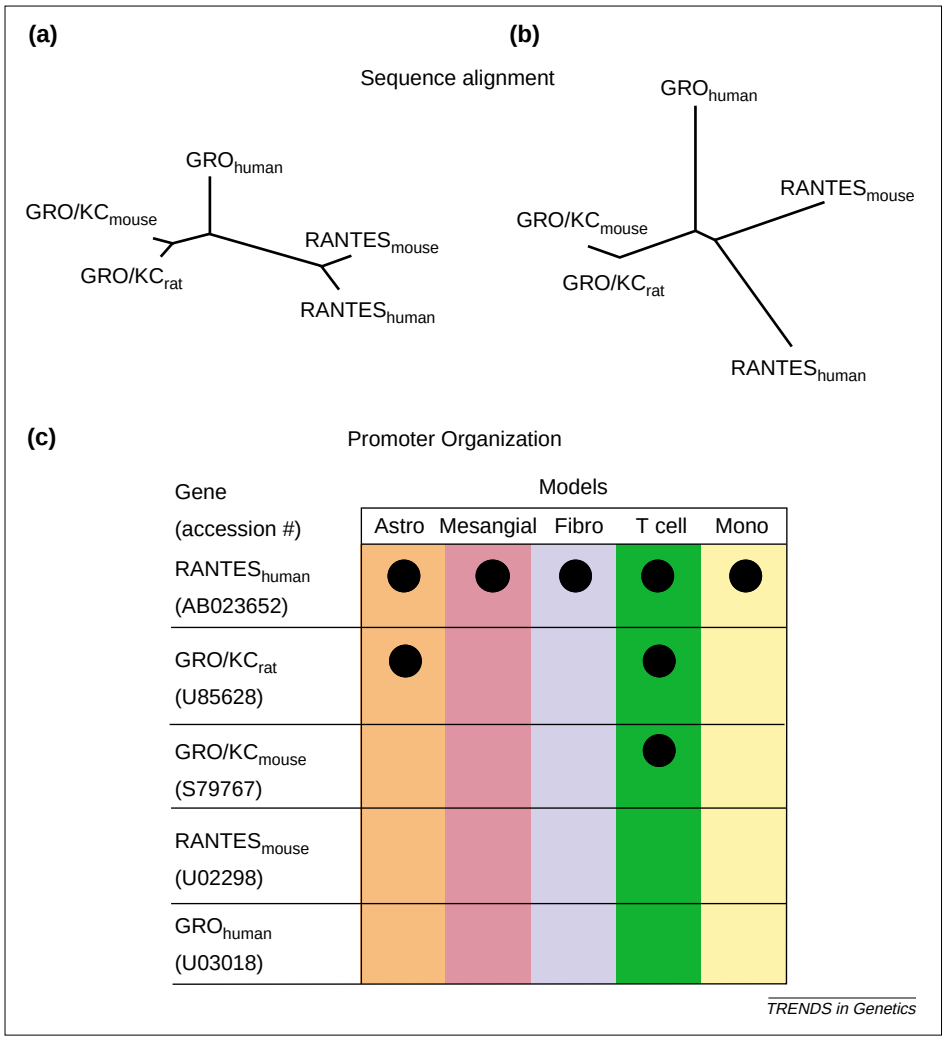

Fig. 2. Comparison of RANTES/CCL5 and GRO/KC from human, mouse and rat. (a) Unrooted tree derived from an alignment of the protein amino acid sequences of the five proteins is shown. Note that homologous proteins clearly cluster together. (b) Alignment of the corresponding promoter nucleotide sequences of the five genes. Note that all promoter sequences appear approximately equally distant (except the rodent GRO/KC genes). (c) Similarity in promoter organization derived *in silico* from tissue type-specific promoter models. The black dots indicate which of the five tissue-specific submodels of the human RANTES/CCL5 promoter recognizes the other chemokine promoters. A higher number of matches indicates organizational similarity of a larger part of the promoter sequences.

Analysis of promoter sequences for organizational structures can help to elucidate the functional context of genes by indicating those that are potentially co-regulated. It is now also possible to derive organizational models solely by *in silico* analysis from groups of genes known to be co-regulated [14,15] (e.g. models of mammalian actin gene promoters and Lentiviral long terminal repeats obtained by automatic modelling [34,35]). Groups of co-regulated genes can be derived, for example, from expression arrays and do not require elaborate experimental analysis of individual promoters [14,15]. Therefore, promoter modules, in addition to more complex organizational models, provide a new general approach towards the elucidation of regulatory networks.

Functional analysis generally focuses on context-dependent functions rather than on intrinsic functions that derive from the amino acid sequence. Although elucidation of metabolic pathways and circuits is extending the functional context at the level of the protein, inclusion of regulatory networks and signalling cascades is required to complete the functional context of genes. Analysis of promoters for organizational features provides a crucial link between the static nucleotide sequence of the genome and the dynamic aspects of gene regulation and expression. In a recent study, Pilpel *et al.* [36] demonstrated that yeast expression data can be successfully analysed by clustering according to common transcription-factor binding-sites, with subsequent addition of expression levels as a second criterion for clustering. They also showed transcription factors to be part of a network connecting genes via these protein factors. Thus, a combination of *in silico* analyses on amino acid sequences and functional comparisons of regulatory sequences (e.g. promoters) will be required to understand fully the functional range of many genes in the genome.

Acknowledgements

This work was supported by the DFG (WE2370/1-2), BMFT (Verbundprojekt FANGREB BEO/31 0311641) and SFB 469 and 571 (P.J.N.).

References

1 Boulikas, T. (1996) Common structural features of replication origins in all life forms. *J. Cell Biochem.* 60, 297–316
2 Fessele, S. *et al.* (2001) Molecular and *in silico* characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. *FASEB J.* 15, 577–579
3 Klingenhoff, A. *et al.* (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* 15, 180–186
4 Kel, A.E. *et al.* (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J. Mol. Biol.* 309, 99–120
5 Fickett, J.W. and Wasserman, W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* 11, 19–24
6 Wasserman, W.W. *et al.* (2000) Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* 26, 225–228
7 Krivan, W. and Wasserman, W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* 11, 1559–1566
8 Wasserman, W.W. and Fickett, J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278, 167–181
9 Firulli, A.B. and Olson, E.N. (1997) Modular regulation of muscle gene transcription: a mechanism for muscle cell diversity. *Trends Genet.* 13, 364–369
10 Yuh, C.H. *et al.* (1998) Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279, 1896–1902
11 Kel-Margoulis, O.V. *et al.* (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.* 28, 311–315
12 Frech, K. *et al.* (1997) Software for the analysis of DNA sequence elements of transcription. *Comput. Appl. Biosci.* 13, 89–97
13 Lavorgna, G. *et al.* (1999) TargetFinder: searching annotated sequence databases for target genes of transcription factors. *Bioinformatics* 15, 172–173

14 Werner, T. (2001) Target gene identification from expression array data by promoter analysis. *Biomol. Eng.* 17, 87–94
15 Werner, T. (2001) Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* 2, 25–36
16 Murphy, P.M. *et al.* (2000) International Union of Pharmacology. XXII. Nomenclature for chemokine receptors. *Pharmacol. Rev.* 52, 145–176
17 Gerard, C. and Rollins, B.J. (2001) Chemokines and disease. *Nat. Immunol.* 2, 108–115
18 Boehlk, S. *et al.* (2000) ATF and Jun transcription factors, acting through an Ets/CRE promoter module, mediate lipopolysaccharide inducibility of the chemokine RANTES in monocytic Mono Mac 6 cells. *Eur. J. Immunol.* 30, 1102–1112
19 Miyamoto, N.G. *et al.* (2000) Interleukin-1β induction of the chemokine RANTES promoter in the human astrocytoma line CH235 requires both constitutive and inducible transcription factors. *J. Neuroimmunol.* 105, 78–90
20 Ortiz, B.D. *et al.* (1996) Kinetics of transcription factors regulating the RANTES chemokine gene reveal a developmental switch in nuclear events during T-lymphocyte maturation. *Mol. Cell. Biol.* 16, 202–210
21 Ortiz, B.D. *et al.* (1997) Switching gears during T-cell maturation: RANTES and late transcription. *Immunol. Today* 18, 468–471
22 Song, A. *et al.* (1999) RFLAT-1: a new zinc finger transcription factor that activates RANTES gene expression in T lymphocytes. *Immunity* 10, 93–103
23 Niehrs, C. and Pollet, N. (1999) Synexpression groups in eukaryotes. *Nature* 402, 483–487
24 Frech, K. *et al.* (1997) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.* 22, 103–104
25 Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23
26 Frech, K. *et al.* (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.* 270, 674–687
27 Stoesser, G. *et al.* (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.* 29, 17–21
28 Perier, R.C. *et al.* (2000) The eukaryotic promoter database (EPD). *Nucleic Acids Res.* 28, 302–303
29 Gao, J.L. *et al.* (1997) Impaired host defense, hematopoiesis, granulomatous inflammation and type 1–type 2 cytokine balance in mice lacking CC chemokine receptor 1. *J. Exp. Med.* 185, 1959–1968
30 Gao, J.L. and Murphy, P.M. (1995) Cloning and differential tissue-specific expression of three mouse β chemokine receptor-like genes, including the gene for a functional macrophage inflammatory protein-1α receptor. *J. Biol. Chem.* 270, 17494–17501
31 Zhang, S. *et al.* (1999) Differential effects of leukotactin-1 and macrophage inflammatory protein-1α on neutrophils mediated by CCR1. *J. Immunol.* 162, 4938–4942
32 Hanks, M. *et al.* (1995) Rescue of the En-1 mutant phenotype by replacement of En-1 with En-2. *Science* 269, 679–682
33 Hanks, M.C. *et al.* (1998) *Drosophila* engrailed can substitute for mouse Engrailed1 function in mid-hindbrain, but not limb development. *Development* 125, 4521–4530
34 Frech, K. *et al.* (1996) Common modular structure of lentivirus LTRs. *Virology* 224, 256–267
35 Frech, K. *et al.* (1998) Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol.* 1, 29–38
36 Pilpel, Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159

Sabine Fessele
Christian Zischek
Peter J. Nelson
Medizinische Poliklinik der Ludwig-Maximilians-Universität München, Schillerstr. 42, D-80336, Munich, Germany.

Holger Maier
GSF-National Research Center for Environment and Health, Institute of Experimental Genetics, Ingolstädter Landstr. 1, D-85764, Neuherberg, Germany.

Thomas Werner*
Genomatix Software GmbH, Landsberger Str. 6, D-80339 Munich, Germany.
*e-mail: werner@gsf.de

Antisense transcripts in the human genome

Ben Lehner, Gary Williams, R. Duncan Campbell and Christopher M. Sanderson

By a systematic search of vertebrate mRNA sequences, we have identified a surprisingly large number of human antisense transcripts. These data suggest that regulation of gene expression by antisense and double-stranded RNAs could be a common phenomenon in mammalian cells.

Although there is abundant evidence to show that natural antisense transcripts (NATs) regulate gene expression in prokaryotic cells [1], there are few reported examples of eukaryotic NATs [2]. Nevertheless, there is good reason to believe that functionally diverse antisense transcripts can operate in higher-eukaryotic cells:

(1) Exogenous antisense RNAs can be used to regulate the expression of an endogenous sense RNA in mammalian cells.
(2) Fifteen percent of imprinted genes have associated antisense transcripts [3]. For example, mutational loss of the *igf2r* antisense transcript prevents imprinting of the *igf2r* gene and upregulates expression of the *igf2r* sense transcript [4].
(3) Introduction of double-stranded RNAs into many organisms including mouse embryos and cultured mammalian cells leads to the degradation of complementary mRNAs by a process called RNA interference (RNAi) [5].
(4) NATs that are complementary to the 3′ untranslated regions (UTRs) of sense mRNAs inhibit expression of the encoded proteins in *Caenorhabditis elegans*. Significantly, an orthologue of one such *C. elegans* antisense transcript is conserved in the human genome [6], and many further potential examples have been discovered recently [14].
(5) Mammalian mRNAs that form sense–antisense pairs (referred to here as NAT pairs) frequently exhibit reciprocal expression patterns [2].

Intrigued by this fragmentary evidence, we instigated a search to identify novel vertebrate NATs and assess the prevalence of putative regulatory RNAs within the human transcriptome.

We used the BLAST algorithm (with an Expect cutoff value of 10^{-9}) to identify regions of complementarity between vertebrate mRNAs from two mRNA databases: RefSeq and our own subset of complete mRNAs. RefSeq [7] (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html) is a curated, nonredundant database containing the most complete example of each sequenced mammalian mRNA. Our set of complete or 'full length' mRNAs consists of all the vertebrate mRNAs annotated as 'complete cds' from the EMBL database. We did not use expressed sequence tag (EST) databases because of the uncertainties concerning the correct orientation of ESTs. We used mRNAs because most of the known vertebrate NATs contain an open reading frame [2]. After excluding vector sequences, we identified a partially redundant set of

  PII: S0168-9525(01)02598-7