25 Oudejans, C.B. *et al*. (2001) Allelic IGF2R repression does not correlate with expression of antisense RNA in human extraembryonic tissues. *Genomics* 73, 331–337
26 Mattick, J.S. and Makunin, I.V. (2005) Small regulatory RNAs in mammals. *Hum. Mol. Genet.* 14, R121–R132
27 Griffiths-Jones, S. (2004) The microRNA Registry. *Nucleic Acids Res.* 32 (Database issue), 109–111
28 Pang, K.C. *et al*. (2005) RNAdb – a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res.* 33 (Database issue), 125–130
29 Ravasi, T. *et al*. Experimental validation of the regulated expression of large numbers of noncoding RNAs from the mouse genome. *Genome Res.* (in press)
30 Imanishi, T. *et al*. (2004) Integrative annotation of 21 037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, e162
31 Ambros, V. *et al*. (2003) A uniform system for microRNA annotation. *RNA* 9, 277–279
32 Kiss, A.M. *et al*. (2004) Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell. Biol.* 24, 5797–5807
33 Borsani, G. *et al*. (1991) Characterization of a murine gene expressed from the inactive X chromosome. *Nature* 351, 325–329
34 Sleutels, F. *et al*. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* 415, 810–813
35 Wolf, S. *et al*. (2001) B-cell neoplasia associated gene with multiple splicing (BCMS): the candidate B-CLL gene on 13q14 comprises more than 560 kb covering all critical regions. *Hum. Mol. Genet.* 10, 1275–1285
36 Ji, P. *et al*. (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041
37 Chan, A.S. *et al*. (2002) Identification of a novel gene NCRMS on chromosome 12q21 with differential expression between rhabdomyosarcoma subtypes. *Oncogene* 21, 3029–3037
38 Millar, J.K. *et al*. (2000) Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* 9, 1415–1423
39 Lewis, B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20
40 Bentwich, I. *et al*. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* 37, 766–770
41 Kiyosawa, H. *et al*. (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* 13, 1324–1334
42 Workman, C. and Krogh, A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* 27, 4816–4822
43 Rivas, E. and Eddy, S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 16, 583–605
44 Washietl, S. *et al*. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2454–2459
45 Mattick, J.S. and Gagen, M.J. (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* 18, 1611–1630
46 Willingham, A.T. *et al*. (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* 309, 1570–1573
47 Mattick, J.S. (2005) The functional genomics of noncoding RNA. *Science* 309, 1527–1528

0168-9525/$ - see front matter
doi:10.1016/j.tig.2005.10.003

Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key

Tanya Vavouri[1,3], Gayle K. McEwen[1,2], Adam Woolfe[1,3], Walter R. Gilks[2] and Greg Elgar[3]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, CB10 1SA
[2]MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, UK, CB2 2SR
[3]School of Biological Sciences, Queen Mary, University of London, Mile End Road, London, UK, E1 4NS

Many conserved non-coding elements (CNEs) in vertebrate genomes have been shown to function as tissue-specific enhancers. However, the target genes of most CNEs are unknown. Here we show that the target genes of duplicated CNEs can be predicted by considering their neighbouring paralogous genes. This enables us to provide the first systematic estimate of the genomic range for distal *cis*-regulatory interactions in the human genome: half of CNEs are >250 kb away from their associated gene.

Corresponding author: Vavouri, T. (tv1@sanger.ac.uk).
Available online 10 November 2005

Introduction

Candidate regulatory elements are often identified using comparative genomics because sequence conservation is considered to indicate negative selection and functional constraint [1–3]. However, the assignment of regulatory elements to genes is a challenging and laborious task. Distal *cis*-regulatory elements in vertebrates are often located far from the genes they interact with and, in some cases, they are found within the introns of neighbouring genes [4–8]. For example, in the human genome, an enhancer of the Sonic Hedgehog gene (*SHH*) is found within an intron of a gene that is located 1 Mb away from *SHH* [4]. Similarly, an enhancer of *PAX6* is 200-kb

downstream of *PAX6* [6]. This makes the association of genes with their potential regulatory elements a significant problem in the human genome.

Conserved non-coding elements (CNEs) in vertebrate genomes have been found to cluster near transcription factors and developmental regulators, indicating that they are involved in vertebrate development [7,9–11]. Indeed, most of the experimentally tested CNEs in transient transfection assays appear to function as tissue-specific enhancers [7,11,13]. Subsets of CNEs have been found to share sequence similarity and to reside next or within a few genes from transcription factors from the same protein families [9–13] (G.K. McEwen *et al.*, unpublished data). For example, in a recent analysis, five sets of CNEs were found to share >75% identity over an alignment length of at least 50 bp [10]. These findings suggest that duplicated CNEs (dCNEs) might be *cis*-regulatory elements that direct tissue-specific expression and, assuming that sequence similarity indicates similarity of function, they are expected to be shared between paralogous genes with common expression patterns [14]. In this article, we propose the comparison of all neighbouring protein-coding genes with each other as an unbiased method to assign dCNEs to specific genes, even when the dCNEs are located hundreds of kilobases away from their predicted targets. Assigning dCNEs to individual genes enables us to calculate the distances that separate these elements from their predicted targets. Assuming that dCNEs function as *cis*-regulatory elements, we present the first computational analysis that aims to define the genomic radius of regulatory activity for *cis*-elements involved in early development in the human genome.

dCNEs are associated with duplicated genes

To test whether duplicated CNEs are the result of retention of regulatory elements after gene duplication, we used a set of CNEs that is conserved between the human and *Fugu* genomes, identified by a more sensitive search than previously described in Ref. [11] (supplementary material online and G.K. McEwen *et al.*, unpublished data). The resulting set of DNA elements consists of 267 dCNEs that can be grouped into 129 families of two-to-four members (a mean of two dCNEs per family). For every family, there is at least one dCNE with a match in the *Fugu* genome; the rest possibly represent more recent duplications. All dCNEs from the same family (except for one pair of dCNEs) are located on different chromosomes in the human genome. The smallest dCNE in the human genome is 42 bp and has an 85% identical match in the *Fugu* genome. The mean dCNE length is 178 bp and the mean percent identity of the dCNEs when compared with their *Fugu* counterparts is 84%. The longest dCNE in the human genome is 737 bp and has an 83% identical match in the *Fugu* genome. As further evidence that the dCNEs are potential regulatory elements, 95% (253/267) of the dCNEs overlap 100-bp intervals of the human genome that have a positive three-way regulatory potential (RP) score [15]. Sequences with positive RP scores are considered useful for identifying putative regulatory elements [16].

Having first established a set of dCNEs, we then set out to define their potential target genes using a method we term 'paralogy mapping'. The most distal regulatory element documented in the human genome is the Sonic Hedgehog (SHH) enhancer, which is 1 Mb from its target gene [4]. Because all known *cis*-regulatory elements are located within 1 Mb from their associated genes, we retrieved all protein-coding genes that are found up to 1 Mb from each dCNE in the human genome (using Ensembl v29 [17]). These genomic regions contain between one and 57 genes (with a mean of 12 genes). We then identified the genes that have paralogues in the regions that neighbour at least two dCNEs from the same family (Figure 1a). Protein paralogy for this analysis was defined according to the TRIBE-MCL clustering of the human protein-coding genes in Ensembl [18]. Based on paralogy mapping, most dCNEs (63%) are associated with just one gene, indicating that dCNEs are regulatory elements retained after duplication with their target gene (Figure 1b). Furthermore, 246 dCNEs (92%) can be associated with one or more genes (Figure 1b). Although our analysis cannot rule out the possibility that dCNEs are *cis*-regulatory elements affecting the gene expression of several neighbouring genes, most dCNEs are found close to one or more paralogues, providing unbiased evidence that CNEs are not genomic features independent of their genic environment. Instead CNEs, when duplicated, are retained with their neighbouring genes.

The sequence similarity of CNEs between the human and the *Fugu* genome suggests that they are functional in both species. It would be reasonable then to assume that the predicted target genes would also be the same in both species. The draft sequence of the *Fugu* genome is assembled into ~8000 scaffolds, each between 2 kb and 1 Mb in length [19]. Therefore, to test whether the *Fugu* dCNEs are within the same scaffold as the *Fugu* orthologue of their predicted target, we analysed 158 dCNEs that have a single predicted target within 1 Mb in human and have a hit in the *Fugu* genome (using BLAST, database size = 330 Mb, e-value $\leq 10^{-4}$). The *Fugu* orthologues of the human predicted-dCNE targets were retrieved from Ensembl (using unique best reciprocal hit and reciprocal hit based on synteny). In 84/158 examples, the *Fugu* orthologue is found in a scaffold that contains a match with the human dCNE. For most of the remaining dCNEs (70/74), the distance between the dCNE and the end of the *Fugu* scaffold was less than the distance between the dCNE and the human predicted target gene. Therefore, these examples do not enable any conclusions to be made about the association of dCNEs and the predicted targets in *Fugu*. For three out of the four remaining dCNEs, another member of the same family has a 'hit' and a predicted orthologous target gene in the same scaffold in *Fugu*. Hence, the results of this analysis are consistent with the initial hypothesis that the genes associated by paralogy mapping with dCNEs in human are their *bona fide* targets.

dCNEs are retained with duplicated transcription factors

It has previously been shown that genes adjacent to CNEs usually encode transcription factors [9–11]. We confirmed this 'enrichment' for transcription factors (*P*-value

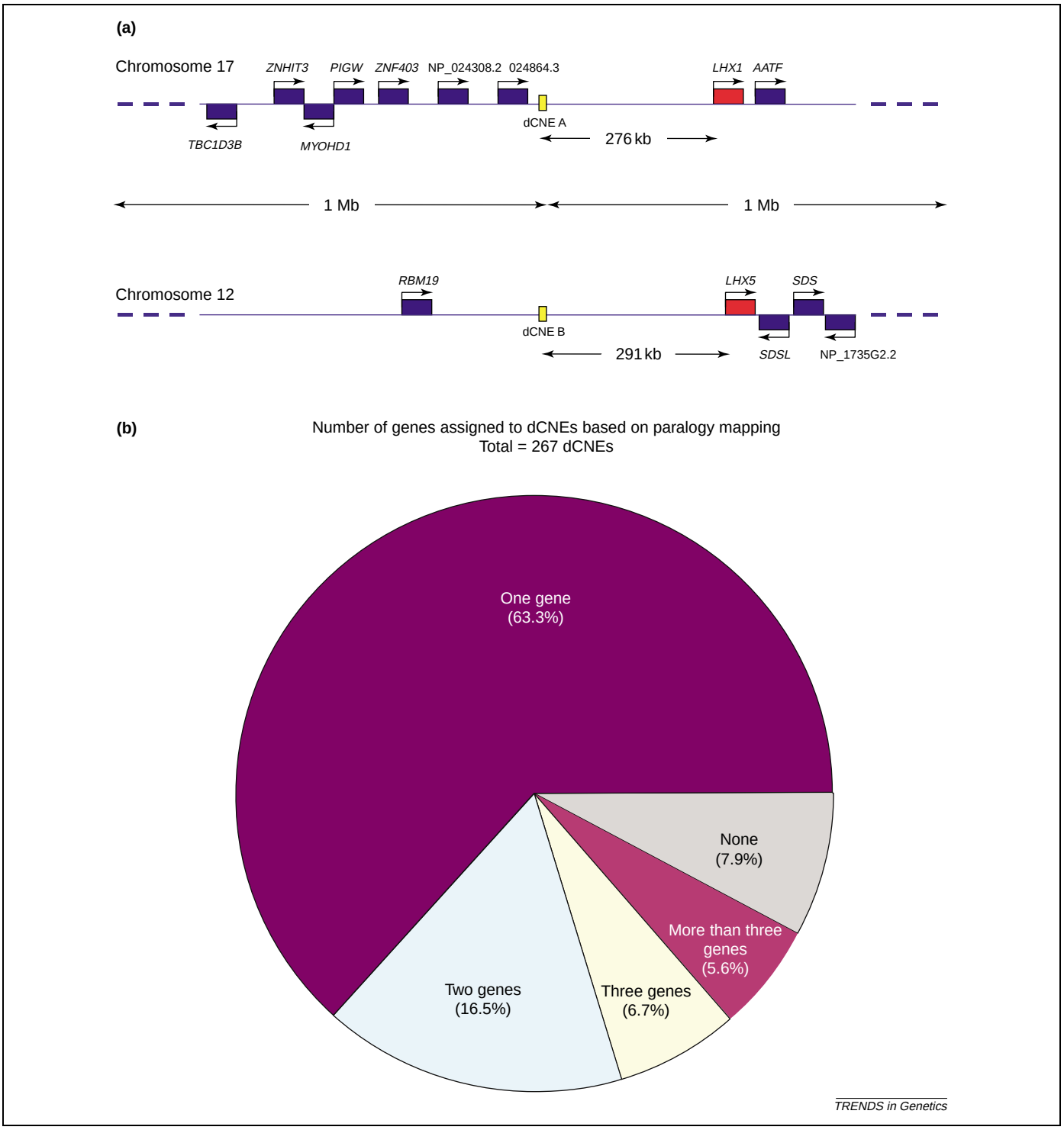

Figure 1. Predicting target genes for dCNEs by paralogy mapping. **(a)** Two dCNEs that share sequence similarity were assigned to a single gene by paralogy mapping. The dCNEs and the protein-coding genes that are located within 400 kb are shown. dCNE A and dCNE B (yellow) are 60% identical over their entire length. Only two genes (*LHX1* and *LHX5*, shown in red) that are found within 1 Mb of these dCNEs belong to the same protein family. Both dCNEs are located ~300 kb from their predicted target. LHX1 and LHX5 belong to the LIM protein family, which contains 16 members in humans. These two genes are expressed in similar patterns in the early development of the vertebrate brain [25]. The diagram is not drawn to scale. **(b)** We assumed that dCNEs are *cis*-regulatory elements, retained with duplicated genes, and we counted the number of paralogous genes that are adjacent to or neighbour similar dCNEs. This pie chart shows the number of genes assigned to dCNEs that are located within 2 Mb (i.e. 1 Mb upstream and 1 Mb downstream). For the majority of dCNEs (shown in purple), there is a single predicted target. At this genomic radius, only 8% of dCNEs cannot be assigned to any probable targets on the basis of paralogy mapping. Abbreviations: AATF, apoptosis antagonizing transcription factor; LHX1, LIM homeobox 1; LHX5, LIM homeobox 5; MYOHD1, myosin head domain containing 1; PIGW, phosphatidylinositol glycan, class W; RBM19, RNA-binding motif protein 19; SDS, serine dehydratase; SDSL, serine dehydratase-like; TBC1D3B, TBC1 domain family member 3B; ZNF403, zinc finger protein 403; ZNHIT3, zinc finger hit type 3.

$<10^{-22}$) by comparing the Gene Ontology (GO) annotation [20] of all *Fugu* genes within 1 Mb of the dCNEs with those in the human genome using GOstat [21] (Table 1 in the supplementary material online). We then assessed the transcription-factor enrichment of genes identified by paralogy mapping compared with that of genes adjacent to dCNEs. This comparison revealed an even stronger enrichment for transcription-factor activity

(P-value $<10^{-73}$, Table 1 in the supplementary material online). This result is not a consequence of transcription factors being over-represented in multigene families in the human genome: 46% of the predicted dCNE target genes are annotated with the GO term 'transcription factor activity' compared with 7% of human genes in multigene families. From an examination of all genes identified by paralogy mapping (Table 2 in the supplementary material online) the enrichment for transcription factors is prevalent even for dCNEs with several predicted targets. These include genes in the HOX and the Iroquois-related homeobox (IRX) clusters. Therefore, it is possible that the examples we have identified of several genes assigned to a single dCNE represent examples of enhancer sharing or global control regions [22,23] rather than false predictions of our method. Thus, our analysis has strengthened the case for the association of highly conserved non-coding elements with the regulation of genes that are important in early vertebrate development.

Half of dCNEs are associated with genes that are found >250 kb away

So far, we have considered for each dCNE all the genes that are in the genomic radius defined by the most distal enhancer documented in the human genome. To assess the number of predicted target genes for each dCNE across a range of distances, we repeated our analysis every 250 kb up to 2 Mb away (Figure 2). Only half of the dCNEs have predicted target genes within the first 250 kb, whereas 95% of the dCNEs have predicted targets within 1.25 Mb. After that distance, the number of unassigned dCNEs remains approximately the same up to 2 Mb. Looking at the genomic space up to 1 Mb away from the dCNE maximises the number of dCNEs associated with a single gene, but minimises the number of dCNEs without any genes associated with them. Searching too far from each dCNE for potential targets carries the risk of identifying genes that belong to the same family by chance. Nonetheless, we found that the number of dCNEs with at least one target identified by paralogy mapping that is up to 2 Mb from its dCNE was significantly larger than expected by chance ($P<0.001$ based on 1000 randomisations; Figure 1 in the supplementary material online).

We then examined the distribution of distances between dCNEs and their assigned genes. For reasons of simplicity, we only considered the examples where a dCNE can be assigned to a single gene within 1 Mb. Our results show that less than a third of dCNEs (50/169) are within 100 kb from their assigned gene (Figure 3). Although these distances appear to be large, they might be slightly underestimated because our previous analysis showed that searching further than 1 Mb from each dCNE yields more potential target genes. We can thus conclude that analyses of short stretches of upstream sequence are inadequate for the identification of *cis*-regulatory elements, especially when the genes under consideration encode transcription factors that are involved in early development. A more appropriate genomic range to analyse would be up to ~0.5-1 Mb upstream and downstream of the gene, although there might still be a small fraction of potential regulatory elements that are further away. The reported range of distances should also be considered when analysing non-coding mutations causing diseases and breakpoint-associated disorders, especially when the candidate gene is a transcription factor.

There are reports of several diseases that are associated with genomic disruptions hundreds of kilobases away from affected genes [3,5]. For example, Townes-Brocks syndrome can be caused by a balanced translocation that is at least 180-kb telomeric to the sal-like 1 gene (*SALL1*) [24]. Interestingly, our data set of dCNEs assigned to single genes includes two dCNEs that were assigned to

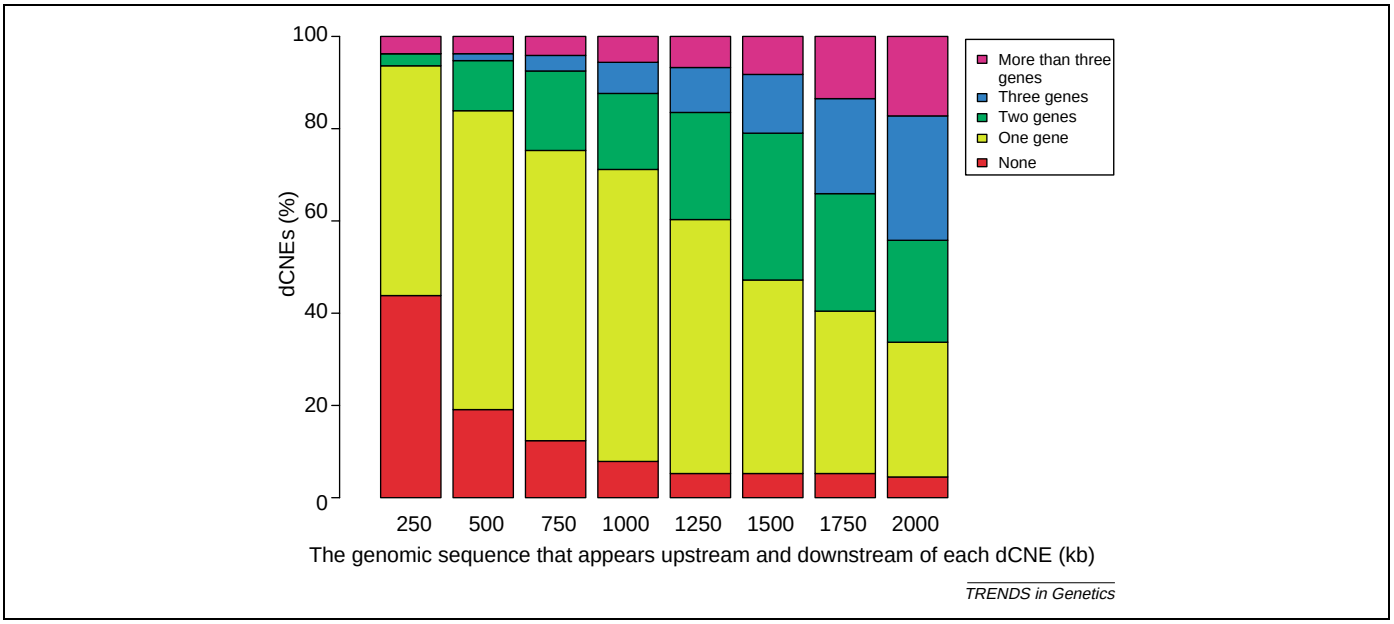

Figure 2. Unbiased mapping of dCNEs to paralogous genes at different distances. Almost half of dCNEs have no associated genes when the first 250 kb of upstream and downstream sequence are analysed. However, when analysing distances that are >1 Mb from the dCNE <10% of dCNEs have no predicted targets. The genes assigned to dCNEs within 1 Mb are highly enriched for transcription factors. At this distance, the role of the predicted targets in transcription regulation is prevalent, even in the examples where more than one gene has been assigned to a dCNE (e.g. genes in HOX clusters).

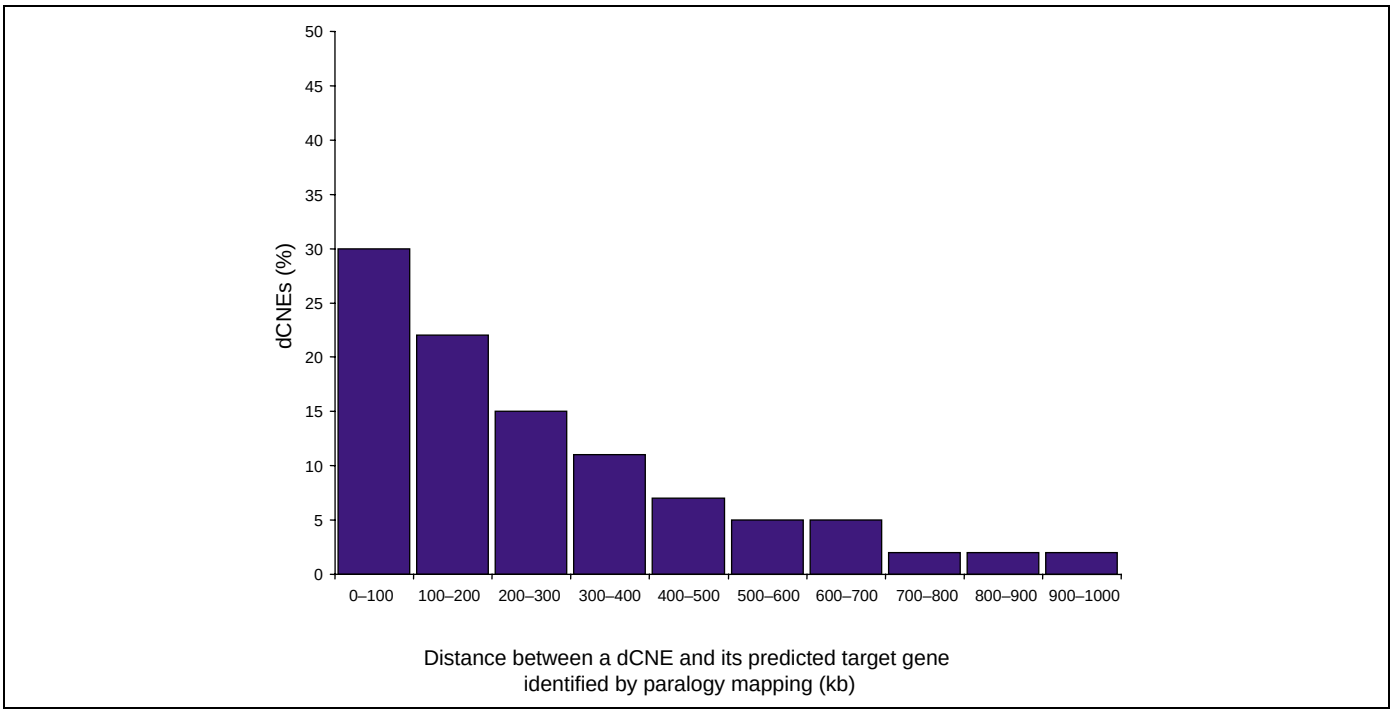

Figure 3. Genomic radius for long-range enhancer action estimated from dCNEs associated with target genes using paralogy mapping. The distances plotted correspond to 169 dCNEs that were assigned to single genes within 1 Mb of upstream and downstream genomic sequence. The results show that the predicted target genes tend to be located close to dCNEs, however only 30% of them are within 100 kb. In a few examples (2%), dCNEs are separated by 800 kb –1 Mb from their predicted targets.

SALL1 and two assigned to sal-like 3 (*SALL3*) that are in the region reported to be affected by the translocation. These SALL-associated dCNEs are between 278-kb and 908-kb upstream of *SALL1*. Our analysis has therefore identified these sequences as candidates for SALL1 and SALL3 *cis*-regulatory elements that are translocated in certain cases of Townes-Brocks syndrome.

Concluding remarks

Non-coding elements that are conserved between human and *Fugu* are considered to have great regulatory potential. We assigned conserved elements that exist at low copy numbers in the human genome to their probable targets, based on paralogy mapping of all neighbouring protein-coding genes. Our results have shown that these elements are strongly associated with duplicated transcription factors. Most of our candidate regulatory elements could be assigned to individual genes, even when analysing a total of 2 Mb of upstream and downstream genomic sequence. Approximately half of the regulatory sequences are likely to be found up to 250-kb upstream of the start of a gene encoding a transcription factor, and the rest of the regulatory sequences are likely to be found up to 1 Mb away from the gene. A few experimentally verified enhancers are known to be located at such long distances from their targets (e.g. the enhancer of *SHH*). The genomic radius of activity for *cis*-regulatory elements is an important aspect of their mechanism and is vital for understanding eukaryotic transcription and human genetic disorders.

Acknowledgements

We thank Ben Lehner for stimulating discussions and for critically reading the article. Part of the analysis described in this article was carried out at the MRC Rosalind Franklin Centre for Genomics Research. This work was supported by the U.K. MRC. T.V. is a Predoctoral Fellow funded by the MRC.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.10.005

References

1 Pennacchio, L.A. and Rubin, E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* 2, 100–109
2 Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 16, 369–372
3 Ahituv, N. *et al*. (2004) Exploiting human–fish genome comparisons for deciphering gene regulation. *Hum Mol Genet* 13(Suppl. 2), R261–R266
4 Lettice, L.A. *et al*. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 12, 1725–1735
5 Kleinjan, D.A. and van Heyningen, V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* 76, 8–32
6 Griffin, C. *et al*. (2002) New 3′ elements control Pax6 expression in the developing pretectum, neural retina and olfactory region. *Mech. Dev.* 112, 89–100
7 Nobrega, M.A. *et al*. (2003) Scanning human gene deserts for long-range enhancers. *Science* 302, 413
8 Thomas, J.W. *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788–793
9 Bejerano, G. *et al*. (2004) Ultraconserved elements in the human genome. *Science* 304, 1321–1325
10 Sandelin, A. *et al*. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* 5, 99
11 Woolfe, A. *et al*. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3, e7

12 Bejerano, G. *et al*. (2004) Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics* 20(Suppl. 1), i40–i48
13 de la Calle-Mustienes, E. *et al*. (2005) A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* 15, 1061–1072
14 Prince, V.E. and Pickett, F.B. (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.* 3, 827–837
15 Kolbe, D. *et al*. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.* 14, 700–707
16 King, D.C. *et al*. (2005) Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res.* 15, 1051–1060
17 Hubbard, T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res.* 33 (Database issue), D447–453
18 Enright, A.J. *et al*. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584
19 Aparicio, S. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301–1310
20 Harris, M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32 (Database issue), D258–261
21 Beissbarth, T. and Speed, T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20, 1464–1465
22 Ellies, D.L. *et al*. (1997) Relationship between the genomic organization and the overlapping embryonic expression patterns of the zebrafish *dlx* genes. *Genomics* 45, 580–590
23 Kmita, M. and Duboule, D. (2003) Organizing axes in time and space; 25 years of colinear tinkering. *Science* 301, 331–333
24 Marlin, S. *et al*. (1999) Townes-Brocks syndrome: detection of a *SALL1* mutation hot spot and evidence for a position effect in one patient. *Hum. Mutat.* 14, 377–386
25 Bachy, I. *et al*. (2001) The LIM-homeodomain gene family in the developing *Xenopus* brain: conservation and divergences with the mouse related to the evolution of the forebrain. *J. Neurosci.* 21, 7620–7629

0168-9525/$ - see front matter
doi:10.1016/j.tig.2005.10.005

An evolutionary scenario for one of the largest yeast gene families

Laurence Despons, Bénédicte Wirth, Véronique Leh Louis, Serge Potier and Jean-Luc Souciet

UMR 7156 University Louis Pasteur-CNRS, Department of Microorganisms, Genomes and the Environment, 28 rue Goethe, 67083 Strasbourg Cedex, France

The *DUP* gene family of *Saccharomyces cerevisiae* comprises 23 members that can be divided into two subfamilies – *DUP240* and *DUP380*. The location of the *DUP* loci suggests that at least three mechanisms were responsible for their genomic dispersion: nonreciprocal translocation at chromosomal ends, tandem duplication and Ty-associated duplication. The data we present here suggest that these nonessential genes encode proteins that facilitate membrane trafficking processes. Dup240 proteins have three conserved domains (C1, C2 and C3) and two predicted transmembrane segments (H1 and H2). A direct repetition of the C1–H1–H2–C2 module is observed in Dup380p sequences. In this article, we propose an evolutionary model to account for the emergence of the two gene subfamilies.

Introduction

Gene duplication, one of the main forces driving evolutionary change, generates sets of paralogous genes that can acquire functional specificities by sequence divergence. All the genomes sequenced so far feature redundant genes that can be classified into gene families. The *PAU* (seripauperin) and *DUP* (duplicated) families, with 24 and 23 members, respectively, are the two largest multigene families in the sequenced *Saccharomyces cerevisiae* strain S288C. The *DUP* gene family comprises two subfamilies named *DUP240* and *DUP380* (Figure 1), which encode proteins consisting of ~240 and 380 amino acids, respectively. These genes, whose function remains unknown, have orthologs in various species that belong to the hemiascomycete phylum, a phylum that is as diverse at the molecular level as the entire chordate animal phylum [1].

There are ten *DUP240* paralogs, which are scattered across four chromosomes (Figure 2, Table 1); these are arranged in single open reading frames (ORFs) and tandemly repeated loci [2]. A recent study that detected traces of ancient duplicated *DUP240* copies showed that two putative ORFs are actually pseudogenes, and that three additional gene relics were detected in intergenic regions [3] (Table 1). Sequences related to yeast transposons (Ty) and/or tRNA genes were observed surrounding each tandem repeat and adjacent to all of the isolated *DUP240* copies (Figure 2). The *DUP380* subfamily comprises eleven genes named *COS1–COS11* (after conserved sequence) and two pseudogenes; we refer to the *COS* genes as *DUP380* genes because they encode proteins containing ~308 amino acids (Table 1 indicates the systematic ORF name of each *COS* gene). All the

Corresponding author: Souciet, J.-L. (souciet@gem.u-strasbg.fr).
Available online 2 November 2005