21 Lakshmipathy, U. and Campbell, C. (1999) Double strand break rejoining by mammalian mitochondrial extracts. *Nucleic Acids Res.* 27, 1198–1204
22 Thacker, J. *et al.* (1992) A mechanism for deletion formation in DNA by human cell extracts: the involvement of short sequence repeats. *Nucleic Acids Res.* 20, 6183–6188
23 Larrson, N.G. and Holme, E. (1992) Multiple short direct repeats associated with single mtDNA deletions. *Biochim. Biophys. Acta* 1139, 311–314
24 Chung, S.S. *et al.* (1996) Analysis of age-associated mitochondrial DNA deletion breakpoint regions from mice suggests a novel model of deletion formation. *Age* 19, 117–128
25 Brossas, J.Y. *et al.* (1994) Multiple deletions in mitochondrial DNA are present in senescent mouse brain. *Biochem. Biophys. Res. Commun.* 202, 654–659
26 Wang, E. *et al.* (1997) The rate of mitochondrial mutagenesis is faster in mice than humans. *Mutat. Res.* 377, 157–166
27 Nowak, R.M. (1999) *Walker's Mammals of the World* (Vol. 1–2), 6th edn, Johns Hopkins University Press
28 Parker, S.P. ed. (1990) *Grzimek's Encyclopedia of Mammals* (Vol. 1–5) McGraw-Hill
29 Perrin, W.F., *et al.* eds (2002) *Encyclopedia of Marine Mammals* Academic Press
30 Bi, X. and Liu, L.F. (1996) A replicational model for DNA recombination between direct repeats. *J. Mol. Biol.* 256, 849–858
31 Rocha, E.P.C. (2003) An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res.* 13, 1123–1132
32 Khrapko, K. *et al.* (1999) Cell-by-cell scanning of whole mitochondrial genomes in aged human heart reveals a significant fraction of myocytes with clonally expanded deletions. *Nucleic Acids Res.* 27, 2434–2441
33 Elson, J.L. *et al.* (2001) Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am. J. Hum. Genet.* 68, 802–806
34 Tang, Y. *et al.* (2000) Maintenance of human rearranged mitochondrial DNAs in long-term cultured transmitochondrial cell lines. *Mol. Biol. Cell* 11, 2349–2358
35 Khrapko, K. *et al.* (2003) Clonal expansions of mitochondrial genomes: implications for *in vivo* mutational spectra. *Mutat. Res.* 522, 13–19

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.03.003

Evidence that functional transcription units cover at least half of the human genome

Marie Sémon and Laurent Duret

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Université Claude Bernard Lyon 1, 16 rue Raphaël Dubois, 69622 Villeurbanne Cedex, France

Transcriptome analyses have revealed that a large proportion of the human genome is transcribed. However, many of these transcripts might be functionless. To distinguish functional transcription units (FTUs) from spurious transcripts, we searched for the hallmarks of selective pressure against mutations that impair transcription. We analyzed the distribution of transposable elements, which are counterselected within FTUs. We show that these features are sufficiently informative to predict whether a sequence is transcribed and, if transcribed, in which orientation. Our results indicate that FTUs constitute at least 50% of the genome and that approximately one-third of these transcripts apparently do not encode proteins.

Analyses of the human genome sequence demonstrated that protein-coding regions constitute ~1.5% of human chromosomes [1]. Given the estimated number and the average length of protein-coding genes, protein-coding transcription units should comprise 30%–40% of our genome [1,2]. However, it is much more difficult to estimate the number of transcription units corresponding to non-coding RNA (ncRNA) genes. On chromosome 7, >200 putative ncRNA genes have been identified, comprising ~2% of the chromosome; however, it is possible that many others remain undiscovered [2]. Large-scale cDNA sequencing projects have been established to provide a complete picture of transcriptomes, and recently new methods have been developed to detect rare transcripts and longer cDNAs [3,4]. These studies revealed that ncRNAs are a major component of the mammalian transcriptome [3,4].

However, it is not clear whether all of these transcripts are functional. Some spurious transcripts might result from the activity of cryptic promoters [e.g. originating from transposable elements (TEs) or from recent pseudogenes] or from the illegitimate extension of transcription downstream of genes with weak polyadenylation signals. Contrary to functional transcription units (FTUs), these spurious transcripts are unnecessary for the proper functioning of genomes and hence are not subject to selective pressure. Thus, one possible way to distinguish FTUs from spurious transcripts is to find evidence that they are under selective pressure to be transcribed. Interestingly, comparisons of the distribution of TEs within introns and intergenic regions have indicated that there is a selective pressure against insertions of TEs within FTUs [5,6]. This is probably because the regulatory elements of such TEs (e.g. polyadenylation signals and promoters) might interfere with the proper expression of FTUs [5,6]. In this article, we describe how we took advantage of this peculiar distribution of TEs to build a model to predict FTUs, and thus evaluated the

Corresponding author: Laurent Duret (duret@biomserv.univ-lyon1.fr).

fraction of the human genome that contains FTUs (i.e. that is under selective pressure to be transcribed).

Analysis and prediction of FTUs

To analyze the features that discriminate FTUs from other genomic sequences, we first had to prepare a set of sequences corresponding to known FTUs and a set of sequences that do not correspond to any known FTU. Our aim was to identify all FTUs, not only from protein-coding genes but also from ncRNA genes. At present, there are insufficient known ncRNA genes to be used for such analysis. However, it is important to note that FTUs that contain protein-coding genes are essentially composed of non-coding sequences: on average, introns constitute 95% of the length of these FTUs [1]. We therefore decided to consider the intron sequences of known protein-coding genes as a model of FTUs for protein-coding genes and ncRNA genes. We used intergenic sequences (i.e. sequences located between transcription units annotated in databases) as a model of non-FTUs. It should be stressed that database annotations might be incomplete, and therefore these 'intergenic' sequences might contain some transcription units that have not been identified. We extracted 3753 intergenic sequences that were >5 kb from the Hovergen database (http://pbil.univ-lyon1.fr/databases/hovergen.html) [7]. We also extracted and concatenated introns from 2506 protein-coding genes (~17 500 introns). This dataset was randomly selected from experimentally characterized protein-coding genes (to avoid potential errors in predicted genes) for which the total intron length was >5 kb. One-third of these data were isolated to be used as a test set. The rest was used for the study of compositional features and for the training of the predictive model.

The most striking difference between FTUs and intergenic sequences is the distribution of TEs. As noted previously [5,6], long interspersed nuclear elements (LINEs) and long terminal repeat retrotransposon (LTR) elements

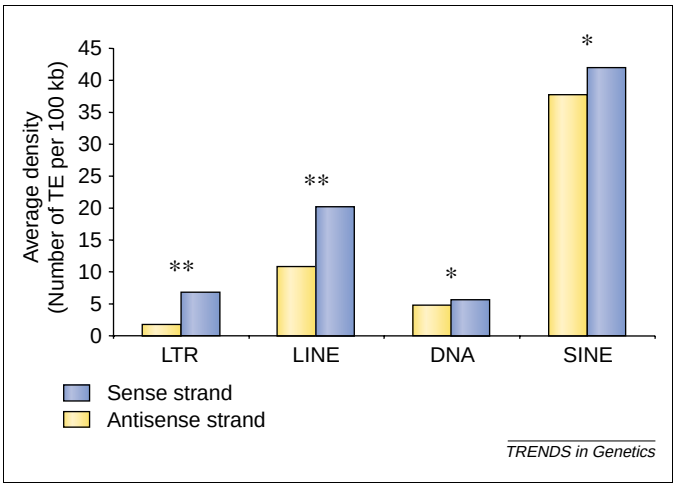

Figure 1. Densities of transposable elements (TEs) on both strands in transcribed regions. There is a strong deficit of LINE and LTR elements on the sense strand compared with the antisense strand of transcripts. This deficit probably results from the fact that these TEs contain polyadenylation signals that can cause premature termination of transcription and, hence, their insertion in the sense strand is conterselected [5]. SINEs and DNA transposons (that have no or weak polyadenylation signals [5]) show a much less pronounced strand bias. The significance of Wilcoxon tests is indicated (*, $P<5\%$; **, $P<1\%$). Abbreviations: DNA, DNA transposons; LINE, long interspersed element; LTR, long terminal repeat retrotransposon; SINE, short interspersed element.

are rare in introns compared with intergenic regions (Table 1). Notably, LTR elements occur twice as frequently in intergenic regions compared with those in FTUs. Moreover, their distribution is asymmetric: LINEs and LTR elements occur 1.5 and 5.5 times more frequently on the sense strand of transcripts than on the antisense strand, respectively (Table 1; Figure 1). This difference between the two strands is probably because these TEs contain polyadenylation signals and insertion of such signals in the sense strand might cause the premature termination of the transcript and therefore be counterselected [5].

LINEs and LTR elements are shorter in both strands of FTUs than in intergenic sequences (Table 1). Notably, the

Table 1. Distribution and length of TEs and base-composition skews in transcribed (Tr) and non-transcribed (Ntr) regions on both strands[a,b]

Transposable elements		Sense strand			Antisense strand		
		Tr[c]	NTr[c]	Ratio (Ntr over Tr)[d,f]	Tr[c]	Ntr[c]	Ratio (Ntr over Tr)[d,f]
Average density (number of TE per 100 kb)	LTR	1.4 (0)	7.7 (6.3)	5.60**	5.6 (4.5)	8.2 (6.8)	1.55**
	LINE	7.8 (7.2)	12.1 (10.8)	1.55**	13.3 (12.1)	12.9 (11.6)	0.97 NS
	DNA	4.3 (3.6)	4.1 (3.5)	1.02 NS	4.8 (4.0)	4.3 (3.5)	0.90*
	SINE	36.0 (29.3)	39 (31)	1.03 NS	39.6 (33.2)	37.5 (29.6)	0.94* NS
Average length (bp)	LTR	463.7 (369.0)	658.4 (380.0)	1.42**	505.6 (374.0)	635.6 (379.0)	1.26 *
	LINE	467.8 (230.0)	690.1 (310.5)	1.48**	599.2 (285.0)	676.0 (307.0)	1.12**
	DNA	207.2 (163.0)	218.7 (161.0)	1.06 NS	221.1 (167)	213.1 (161.0)	0.97 NS
	SINE	231.2 (283.0)	235.4 (287.0)	1.02*	237.6 (288.0)	235.3 (287.0)	0.99*
Frequency of 5′ complete LINEs		3.3%	6.9%	2.09**[e]	5.2%	6.7%	1.29**[e]
$ATskew = \frac{A-T}{A+T}$		−0.050 (−0.052)	−0.0075 (−0.0084)	0.15**	0.050 (0.050)	0.00077 (0.00041)	0.15**
$GCskew = \frac{C-G}{C+G}$		−0.021 (−0.021)	−0.0031 (−0.0032)	0.14**	0.022 (0.021)	0.0039 (0.0036)	0.18**

[a]Abbreviations: DNA, DNA transposons; LINE, long interspersed element; LTR, long terminal repeat retrotransposon; SINE, short interspersed element; TEs, transposable elements.
[b]The orientation of Ntr regions is indicated relative to that of the closest flanking gene.
[c]Median values are indicated in parenthesis.
[d]Comparison of transcribed versus untranscribed regions using the Wilcoxon test (except for [e]).
[e]Chi-square test.
[f]Significance: *, $P<5\%$; ** $P<1\%$; NS, not significant.

proportion of LINE insertions resulting from complete retrotranscription events (i.e. including their 5′ end) is about twice as low in FTUs than in intergenic regions (Table 1). The shortening of TEs in FTUs compared with those of intergenic regions might be explained in part by a selective pressure to decrease the cost of transcription [8]. However, such a selective pressure is expected to act independently of the orientation of TEs. Contrary to this prediction, we noticed that in FTUs, LINEs and LTR elements are shorter on the sense strand than on the antisense strand, and the deficit in 5′-complete LINEs is higher on the sense strand than on the antisense strand (Wilcoxon test, P-value $= 10^{-6}$). This peculiar distribution of TE length on both strands therefore suggests that there is a selection against insertions of regulatory elements (notably promoter elements located in 5′ part of LINEs) within FTUs.

We compared the base composition (GC-content and dinucleotide content) in FTUs and intergenic regions. One noteworthy observation is that in FTUs, the frequency of A and G in the sense strand is higher than the frequency of T and C. These AT and GC skews between the two strands are not observed in intergenic sequences (Table 1). Such skews probably result from the asymmetry of the transcription process, which might affect the pattern of mutation or the efficiency of DNA repair on the sense strand compared with that of the antisense strand [9,10]. The AT and GC skews observed in FTUs, although significant, are not sufficient to be used to predict the location of transcribed regions. However, they are reliable predictors of the orientation of transcribed regions (discussed in the following sections).

Because the insertion of TEs is expected to be counter-selected in FTUs but not in spurious transcripts or untranscribed regions, we used these features to differentiate FTUs from other sequences. For this purpose, we developed a prediction model (generalized linear model [11]) based on the analysis of TE distribution. We took into account the ten most discriminating parameters: the densities of LINEs, short interspersed nuclear elements (SINEs), DNA and LTR elements on each DNA strand, and the AT and GC skews. We introduced the AT and GC skews to predict the orientation of transcripts. Two predictive models, one for each orientation of transcription, were trained on the learning set of FTUs and intergenic sequences. A sliding window (20 kb) was moved along the sequence and, for each window, two scores were computed (one for each transcription orientation).

The efficiency of the model to detect FTUs was first evaluated on the test set; the sensitivity (proportion of the known transcripts that are correctly predicted as transcribed) and specificity (proportion of the FTU predictions that correspond to known transcripts) of the method are ~65%. Note that the specificity is probably underestimated because some 'false positive' predictions in intergenic regions might in fact correspond to true, but unannotated, FTUs. The model performs well in determining the orientation of transcription: 90% of the predictions correspond to the annotations, and 5% of the sequences are predicted to be transcribed on both strands.

We also evaluated the efficiency of the method on the 20 human ncRNA genes, from the Noncoding RNAs Database (http://biobases.ibch.poznan.pl/ncRNA/), that could be located on the human genome [12]. The sensitivity of the method appears to be similar for protein-coding genes and ncRNA genes: 60% of ncRNA genes were predicted to be transcribed and 92% of them in the correct orientation.

Quantification of FTUs in the human genome

We then used our model on the whole human genome [1]. We used annotations to locate previously known transcribed sequences (http://www.ensembl.org/). 37% of the windows contain at least one annotated transcribed region that covers >2 kb within the 20-kb window. Approximately half (53%) of these sequences were recognized as transcribed sequences by the model. Moreover, 24% of the other windows are also predicted to be transcribed. Given the sensitivity and specificity of the method (65%), this approach is not sufficiently accurate to be of use in the automated annotation of the genome. However, this model enabled us to evaluate the fraction of the human genome that is covered by FTUs. Taking into account the sensitivity and specificity measured previously, our results suggest that, overall, 45% of the windows contain FTUs – approximately half of the human genome is under selective pressure to be transcribed. This is significantly more than the proportion of the genome corresponding to the known protein-coding FTUs but less than previous estimations, suggesting that as much as 90% of the human genome might be transcribed [13]. This does not mean that 45% of the genome is under strong selective constraints. Indeed, several recent studies have shown that constrained, evolutionary conserved sequences constitute only 3–5% of mammalian genomes [14–16]. The difference between this figure and the total coverage of FTUs in the genome is because constrained sequences often only constitute a small proportion of a FTU: typically, in a protein-coding FTU, only the exons and some intronic regulatory elements are conserved by evolution, whereas the other parts of the transcription unit evolve rapidly. In other words, in these unconstrained regions of FTUs, many mutations can freely accumulate except those mutations that impair transcription (such as TE insertions). Thus, these regions can be distinguished from spurious transcription units by their requirement for transcription to ensure the expression of mature mRNAs.

We compared the location of the FTUs that we predicted on chromosome 22, with that of transcripts identified by Rinn and colleagues using DNA microarrays [17]. Approximately 87% of FTUs identified by our study overlap with those identified by Rinn and coworkers, which confirms that most of the predicted FTUs are transcribed. However, a large fraction (66%) of transcripts identified by Rinn *et al*. are not located within predicted FTUs. Given the sensitivity of our method, we would have expected only 35% of FTUs to be false negatives. It is likely that some of the transcripts identified in Rinn's study are *bona fide* FTUs and are too short to be detected by our method. However, a large proportion of the transcripts might be spurious, functionless RNAs. Rinn and colleagues argued that their transcripts are probably functional because a significant fraction (44%) show sequence conservation

with orthologous loci in the mouse genome. However, given the evolutionary distance between rodents and primates, even functionless sequences (such as pseudogenes or ancient repeats) might have remained conserved. Indeed, it has been shown that ~40% of the human genome forms alignments with the mouse genome sequence but only 3–5% is under selective pressure [14–16].

Concluding remarks

One-third of the predicted FTUs does not correspond to known protein-coding genes, and thus might encode ncRNA genes. Our analysis therefore provides independent evidence that a significant fraction of the mammalian transcriptome corresponds to functional ncRNAs [3,4]. We found that 5% of the transcribed sequences are expressed on both strands. This observation is consistent with other reports in the literature [18] and could be biologically relevant: many non-coding RNAs are developmental regulators on the antisense strand of a coding gene [19].

Because the human-gene catalogue is not complete, and the discovery of non-coding RNAs is still in its infancy [3,4], our estimate of the fraction of the genome covered by FTUs is not outstanding. Wong and colleagues [13] estimated that protein-coding FTUs might constitute up to 90% of our genome. They argue that there are some long protein-coding genes that remain to be discovered and that these transcription units might cover a large fraction of our genome. Given the uncertainty surrounding the exact number of long genes, their estimate (90%) should be taken as an upper limit. Conversely, our estimate should be considered conservative. Indeed, because of the size of the window (20 kb), our method can only detect relatively long FTUs. Moreover, as mentioned previously, the specificity of our method is probably underestimated. Taken together, this suggests that FTUs constitute >50% of our genome.

References

1 International Human Genome Sequencing Consortium, (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Scherer, S.W. *et al.* (2003) Human chromosome 7: DNA sequence and biology. *Science* 300, 767–772
3 Okasali, Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature* 420, 563–573
4 Kapranov, P. *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919
5 Smit, A.F. (1999) Interspersed repeats and other mementos of transposable elements in Mamm. Genomes. *Curr. Opin. Genet. Dev.* 9, 657–663
6 Medstrand, P. *et al.* (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res.* 12, 1483–1495
7 Duret, L. *et al.* (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22, 2360–2365
8 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
9 Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 12, 640–649
10 Green, P. *et al.* (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* 33, 514–517
11 McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, Chapman & Hall
12 Szymanski, M. *et al.* (2003) Noncoding regulatory RNAs database. *Nucleic Acids Res.* 31, 429–431
13 Wong, G.K. *et al.* (2001) Most of the human genome is transcribed. *Genome Res.* 11, 1975–1977
14 Thomas, J.W. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788–793
15 Mouse Genome Sequencing Consortium, (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
16 Dermitzakis, E.T. *et al.* (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* 302, 1033–1035
17 Rinn, J.L. *et al.* (2003) The transcriptional activity of human chromosome 22. *Genes Dev.* 17, 529–540
18 Shendure, J. and Church, G.M. (2002) Computational discovery of sense–antisense transcription in the human and mouse genomes. *Genome Biol.* 3, 1–14
19 Eddy, S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* 2, 919–929

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.03.001

The impact of very short alternative splicing on protein structures and functions in the human genome☆

Fang Wen[1,*], Fei Li[1,*], Huiyu Xia[1], Xin Lu[1,2], Xuegong Zhang[1] and Yanda Li[1]

[1]MOE Key Laboratory of Bioinformatics and Department of Automation, Tsinghua University, Beijing 100084, China
[2]Current address: Department of Statistics, Harvard University, Cambridge, MA 02138, USA

The systematic analysis of very short alternative splicing (VSAS) in the human genome showed that VSAS might contribute more to protein-function diversity than expected. More than 65% of VSAS fragments have different secondary structures from flanking regions. They tend to have a non-loop structure and have an important influence on protein functions, as shown by the predicted 3D structure of human IL-4δ2. The observed VSAS events can be classified into two groups depending on whether they insert new structure domains in the proteins, and they might be of different evolutionary status.

It is a surprise that only 30 000–40 000 genes exist in the

☆ Supplementary data associated with this article can be found at doi: 10.1016/j.tig.2004.03.005

* These authors have contributed equally.
Corresponding author: Xuegong Zhang (zhangxg@tsinghua.edu.cn).