Combining DGE and RNA-sequencing data to identify new polyA+ non-coding transcripts in the human genome

Nicolas Philippe, Elias Bou Samra, Anthony Boureux, Alban Mancheron, Florence Rufflé, Qiang Bai, John de Vos, Eric Rivals, Thérèse Commes

2820–2832 Nucleic Acids Research, 2014, Vol. 42, No. 5 Published online 18 December 2013
doi:10.1093/nar/gkt1300

Combining DGE and RNA-sequencing data to identify new polyA+ non-coding transcripts in the human genome

Nicolas Philippe[1,2,3,†], Elias Bou Samra[1,2,†], Anthony Boureux[1,2,3], Alban Mancheron[3,4], Florence Rufflé[1,2,3], Qiang Bai[5], John De Vos[3,5], Eric Rivals[3,4] and Thérèse Commes[1,2,3,*]

[1]Transcriptomics, bioinformatics and myeloid leukaemia, INSERM, U1040, Institute for Research in Biotherapy, Montpellier F-34197, France, [2]Université Montpellier 2, Montpellier, France, [3]Institut de Biologie Computationnelle, Maison de la modélisation, Université Montpellier 2, France, [4]LIRMM, MAB, CNRS UMR 5506, Université Montpellier 2, Montpellier, France and [5]Genomic instability of pluripotent stem cells, INSERM, U1040, Institute for Research in Biotherapy, Montpellier F-34197, France

Received March 1, 2013; Revised November 18, 2013; Accepted November 22, 2013

ABSTRACT

Recent sequencing technologies that allow massive parallel production of short reads are the method of choice for transcriptome analysis. Particularly, digital gene expression (DGE) technologies produce a large dynamic range of expression data by generating short tag signatures for each cell transcript. These tags can be mapped back to a reference genome to identify new transcribed regions that can be further covered by RNA-sequencing (RNA-Seq) reads. Here, we applied an integrated bioinformatics approach that combines DGE tags, RNA-Seq, tiling array expression data and species-comparison to explore new transcriptional regions and their specific biological features, particularly tissue expression or conservation. We analysed tags from a large DGE data set (designated as 'TranscriRef'). We then annotated 750 000 tags that were uniquely mapped to the human genome according to Ensembl. We retained transcripts originating from both DNA strands and categorized tags corresponding to protein-coding genes, antisense, intronic- or intergenic-transcribed regions and computed their overlap with annotated non-coding transcripts. Using this bioinformatics approach, we identified ~34 000 novel transcribed regions located outside the boundaries of known protein-coding genes. As demonstrated using sequencing data from human pluripotent stem cells for biological validation, the method could be easily applied for the selection of tissue-specific candidate transcripts. DigitagCT is available at http://cractools.gforge.inria.fr/softwares/digitagct.

INTRODUCTION

Although the fraction of protein-coding sequences is limited to ~2–3% of the whole human genome, the transcript repertoire is much more diverse and complex than anticipated. Growing evidence suggests that most of the genome is pervasively transcribed (pervasive transcription, known also as 'dark matter') (1–3).

The first genome-wide transcription studies performed using complementary DNA (cDNA) sequencing and tiling microarrays showed that a significant fraction of the genome gives rise to RNAs with reduced protein-coding potential (1,4,5). Thereafter, the rapid development of next-generation sequencing technologies provided new tools to thoroughly profile all aspects of transcription diversity at unprecedented resolution. However, using these new technologies, Van Bakel *et al.* (6) concluded that widespread transcription was mainly associated with known genes. This conclusion was refuted by Clark *et al.* (7) who showed that the existence of pervasive transcription is supported by multiple independent techniques, and by Kapranov *et al.* (8) who provided estimates of the relative mass of the 'dark matter' RNA by sequencing total RNA. More recently, GENCODE v7 provided a

*To whom correspondence should be addressed. Tel: +33 4 67 33 04 74; Fax: +33 4 67 33 04 59; Email: commes@univ-montp2.fr

†These authors contributed equally to the paper as first authors.

catalogue of human long non-coding RNAs (lncRNAs) (9), and several reports described the roles of lncRNAs in gene expression and epigenetic regulation (10–12), arguing in favour of the biological significance of pervasive transcription (13,14).

For a decade, several novel technologies have permitted genome-wide investigations of the transcriptome. Each technology comes with its pros and cons, its limitations and its possible artefacts. For instance, Digital Gene Expression (DGE) delivers short sequence signatures with known strand orientation, the quantification of which gives a reliable and comparable measure of a transcript expression level. On the other hand, RNA-sequencing (RNA-Seq) generates reads that cover almost entirely the sequenced RNAs and requires more complex methods, like RPKM/FPFM, for quantification (15). However, RNA-Seq is the only technique that can differentiate between overlapping transcripts at a specific genomic position and can thus distinguish frequent splice variants.

Each of these technologies (whole-genome tiling arrays, DGE and RNA-Seq) provides a global view of the transcriptome, but may miss interesting novel RNAs. Owing to their specific limitations, these technologies may complement each other for RNA discovery. Therefore, it seems reasonable to combine data from different sources and techniques to improve the prediction and reconstruction of novel RNA transcripts with accuracy. In this work, we examined whether integrating various types of transcriptomic data might improve the identification of novel non-coding RNAs (ncRNAs). In addition, we wanted to determine whether the short sequences (tags) generated by the DGE method could be useful to address the still debated issue of whether pervasive transcription is biologically relevant or originates from sequencing artefacts and/or spurious transcriptional noise (6,7,16–18).

To this aim, we developed a new integrated transcriptome analysis procedure in which DGE data are first analysed using a perfect mapping approach to reduce random annotations. The procedure includes the computation of false-positive tag locations (2% in the human genome) and the analysis of a large number of oriented orphan tags (i.e. without genomic annotation) (19). The transcriptional information given by the annotated DGE tags is then completed by integrating expression data obtained by using other techniques (RNA-Seq and tiling arrays). Currently, one of the major difficulties in characterizing new transcripts is the absence of information on their expression levels, which may help assessing their biological relevance. From a computational point of view, tags are instrumental for measuring and comparing the expression level of transcripts in different tissues. To validate our approach, DGE data from 54 publicly available libraries from normal (including human pluripotent stem cells [hpSCs]) and cancer tissues were used for transcript detection and tissue expression comparison. We characterized ~34 000 new potential non-coding transcribed regions (genomic location, conservation in the mouse and human genomes and tissue-specificity) and identified >1121 transcribed regions that are abundantly expressed in hpSCs.

MATERIALS AND METHODS

DigitagCT pipeline to combine different transcriptome data

We developed a computational bioinformatics pipeline that combines DGE tag expression data from different samples with the available annotation resources to obtain a general view of the transcription landscape in a reference genome (Figure 1). The pipeline uses two mandatory arguments: a sequence alignment/map (SAM) input file of mapped DGE reads and a general feature format (GFF) file that must contain at least the required features on 'exon', 'mRNA' and 'gene'. Others features, such as 'cds' (coding sequence), '3′UTR' or '5′UTR', can be added to give more information about the annotation (see http://www.sequenceontology.org/gff3.shtml for more detail). When RNA-Seq and DGE data are combined, the pipeline uses a non-mandatory argument to integrate the SAM file of mapped RNA-Seq reads. Although the pipeline can accept SAM files from any

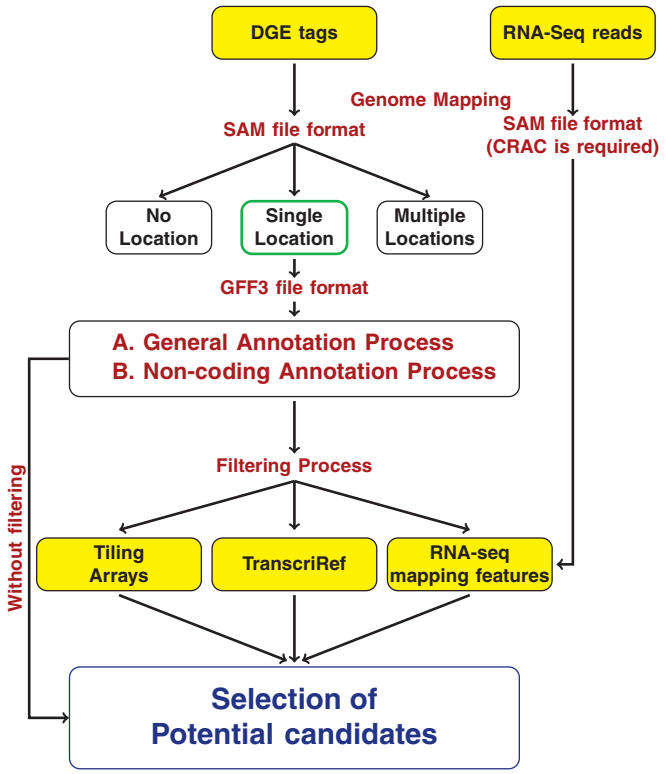


Figure 1. The digitagCT pipeline. Schematic representation of the integrative digitagCT pipeline to analyse high-throughput sequencing data (DGE and RNA-Seq). First, DGE tags are mapped to the reference genome by using the crac software with the –emt argument tool (described in 'Materials and Methods' section). The pipeline uses existent annotation sources (Ensembl annotations for human and mouse) and interrogates the subset of uniquely mapped tags using the Ensembl API. In output, a selection of transcripts can be retained with or without filtering. In the filtering process, retained DGE transcripts are compared with RNA-Seq data, which are integrated in the pipeline by using the CRAC (described in 'Materials and Methods' section), tiling arrays and other DGE data. Users may configure several parameters according to their needs, such as by defining an occurrence threshold for DGE tags or a specific class of transcript (intergenic, non-coding, etc.).

tools, it is advisable to provide a SAM file from CRAC (20). In fact, one of the CRAC specificities is to deliver computational predictions for point mutations, indels, sequence errors, normal and chimeric splice junctions in a single run and to list this information, which can be used to characterize the nature of new transcripts, in extra columns of the SAM file. For example, the combination of a splice junction and an intergenic tag could indicate a new long intergenic non-coding RNA (lincRNA). Our pipeline is called digitagCT and is part of the CracTools suite (not published). DigitagCT is available at http://cractools.gforge.inria.fr/softwares/digitagct/.

Here, we used the digitagCT pipeline to combine the strengths of the DGE and RNA-Seq methods. First, sequences were mapped back to the reference genome using the CRAC software (with the '—emt' argument for an exact matching tool that is particularly suitable for DGE data), available at http://crac.gforge.inria.fr (20).

As the DGE method often implies that each transcript is represented by one tag, DGE data were annotated according to a GFF file from Ensembl Genome Browser by giving priority first to location in exons and then in intronic or intergenic regions. The GFF file was built from Ensembl API (version 66). First the protein-coding gene and pseudogene categories were considered to determine whether the tags corresponded to intragenic (cds, UTR and introns) or intergenic sequences (process A) (Figure 2). For intergenic regions, tags were then classified as proximal (intergene proximal), when the distance between the tag and the 5′ of the next gene was <5 kb, or as distal (intergene distal and intergene EST), when such distance was > 5 kb. The procedure could also distinguish between sense and antisense transcripts originating from both DNA strands because the DGE protocol generates oriented tags (Figure 2). Then, we classified the tags by giving higher priority to gene versus intergenic annotations, and to annotations on the same strand rather than to annotations on the opposite strand. The classification algorithm proceeded as follows. If the gene and the tag were in the same orientation (sense tags), a tag located within a gene could be exonic (tag1, tag2, tag3) when entirely within an exon, inxonic (tag4) if it covered an intron–exon junction or intronic (tag5). The same approach was used with genes on the opposite strand (if any): anti-sense tags could thus be exonic (tag6, tag7, tag8), inxonic (tag9) or intronic (tag10). If a tag was not annotated, we assessed its possible intergenic localization and then classified the tag as proximal (tag11) or distal (tag12) relative to a 3′ gene.

The second step (process B) of our pipeline allowed compiling all the previously Ensembl-annotated, non-coding and unclassified transcripts to specify their

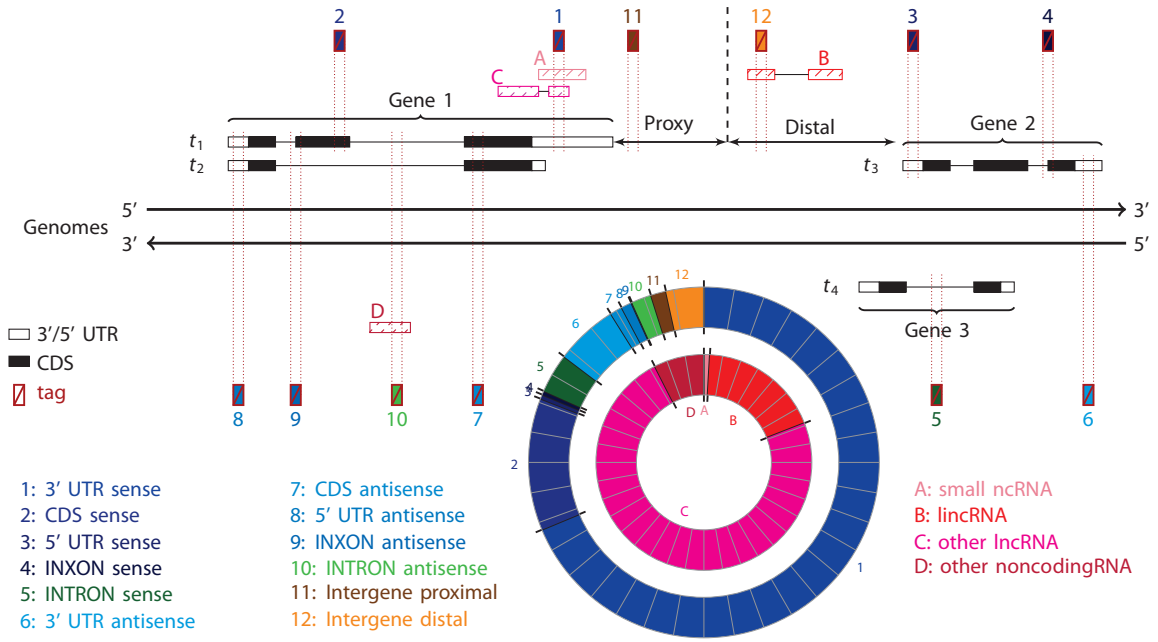


Figure 2. Classification of 'TranscriRef' sequences using process A and process B of the digitagCT pipeline. We adopted a two-step strategy to annotate the subset of 'TranscriRef' tags (from Ensembl annotation) that were uniquely mapped to the human genome using CRAC: a general annotation process that considers protein-coding genes and pseudogenes (process A) and a non-coding annotation process that considers only non-coding genes (process B) (see schematic representation in Figure 1 and 'Materials and Methods' section for details). (Process A) A tag located in a protein-coding gene or a pseudogene (sense orientation) could thus be categorized as exonic (tag1, tag2, tag3), inxonic (tag4) or intronic (tag5). Similarly, a tag located in a gene, but on the opposite strand, could be exonic (tag6, tag7, tag8), inxonic (tag9) or intronic (tag10). A tag outside a gene (intergenic localization) was classified as proximal (tag11) or distal (tag12) to the nearest 3' gene. The external pie chart shows the genomic distribution of DGE sequences assigned to coding-genes based on the tag classification. (process B) Then, tags that overlapped with sequences of non-coding genes were classified in (**A**) small ncRNAs, (**B**) lincRNAs, (**C**) other lncRNAs and (**D**) other ncRNAs. A tag could identify both a non-coding and a protein-coding gene (e.g. tag 1 corresponds to the 3' UTR region of a coding transcript and also to a non-coding transcript). In this case, we consider that the non-coding transcript overlaps with a protein-coding gene. The internal pie chart shows the global genomic distribution of DGE sequences assigned to non-coding transcripts.

genomic proportion and distribution (Figure 2). Most ncRNAs were annotated by aligning their genomic sequence against RFAM using BLASTN. MicroRNAs (miRNA) were predicted by BLASTN of genomic sequence slices against miRBase sequences. LincRNAs were predicted using the lincRNA Ensembl gene annotation, cDNA alignments and chromatin-state maps (21). For any tag, the digitagCT transcriptome pipeline could also analyse RNA-Seq data to complement the DGE information and help constructing the transcript sequence by generating supplementary features, such as splice variants or polymorphisms. Hence, we could intersect any DGE tag position with those of neighbouring RNA-Seq reads. Moreover, a filtering process that integrated other transcriptome expression data (i.e. TranscriRef, tiling arrays) was used to improve the relevance of our analysis (Figure 1). For each tag and according to its location, a value was computed to indicate the detection of a transcribed region with each technology (DGE, RNA-Seq and tiling). Future users may adjust the parameters of the transcriptomic pipeline functions according to their requirements. An optional filtering process for species-comparison analysis may be performed by choosing the genome of the studied species and using a selection of DGE tags as input.

Data sets

SAGE/DGE data were collected from publicly available repositories: the CAGP project (Sage genie: ftp://ftp1.nci.nih.gov/pub/SAGE/) for human and mouse data and the kidney DGE library described in Philippe *et al*. (19). One hundred base pairs of RNA-Seq paired-end reads of the HD291 human embryonic stem cell (hESC) line were obtained using an Illumina HiSeq2000 device at DNAVision (Gosselies, Belgium). The DGE libraries for the induced pluripotent stem (iPS) cell line M4C2 and human foreskin fibroblasts were generated at the MGX platform (Montpellier, France) using the NlaIII restriction enzyme (CATG sites) (unpublished data, Bai *et al.)*. RNA-Seq data from acute myeloid leukaemia (AML) primary cells were produced in the laboratory. The compilation of all the used DGE libraries (54 libraries; list in Supplementary Table S2) was designated as 'TranscriRef'. *Homo sapiens* chromosome sequences (GRCh37 version) were retrieved from the Genome Reference Consortium (http://www.ncbi.nlm.*nih.gov*/*projects*/*genome*/*assembly*/*grc*/*human*/*) and *Mus musculus* chromosome sequences (NCBI37/mm9 version) from the UCSC genome browser Web site (http://genome.ucsc.edu/). Annotations and orthology information were retrieved from Ensembl (E66 version) and tiling arrays data from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu/). We used transcriptional active region (TAR) data from the Affymetrix Transcriptome Project Phase2, Affymetrix PolyA+ RNA transfrags, Yale RNA TARs and Yale Maskless Array synthesizer experiments [see Rivals *et al*. for detailed information (22)].

Annotations and ncRNA categories

For clarity, 'tag' defines a representative sequence and 'occurrence' indicates all sequences that are identical to that tag. Thus, the number of occurrences (occnb) indicates the number of times a tag is found in the library collection and it is considered a measure of its biological validity. Specifically, a tag observed only once may be the result of a sequencing error, whereas a tag observed at least 10 times is more likely to be a valid biological observation.

Non-coding transcripts were classified in four groups according to the Ensembl annotation and the annotation proposed by ENCODE: (i) small ncRNAs (mainly miRNAs, snoRNAs and snRNAs); (ii) lincRNAs defined by Ensembl; (iii) other lncRNAs: this is an ENCODE category that includes processed, antisense and sense intronic transcripts and is described in (9); and (iv) other ncRNAs: this category is defined by Ensembl. The full definition of each category can be found in the Ensembl website. Potential new transcripts that could not be annotated according to the Ensembl human genome annotation (process B) were classified in the intergenic distal category.

Concerning intergenic tags, we always reported their distance from the 3′- and 5′-ends of known genes, irrespective of the strand orientation (Supplementary Figure S1), because 3' extension of known genes is one of the potential sources of annotation artefacts that generate false intergenic transcription (6). The tag distribution allowed the easy distinction of the two classes of intergenic tags defined as proximal (intergene proximal) and distal (intergene distal). For intergene distal tags, we could compute the overlap with EST (intergene EST). Distal tags were equally distributed in the genome, whereas proximal tags were heterogeneously distributed with high density at the 3′-ends of known genes. Thus, we decided to consider only intergenic distal tags for subsequent analyses.

Manual validation and curation

To select specific candidate tissue-specific transcripts, the corresponding DGE and RNA-Seq libraries were analysed with the pipeline using the filtering process (Figure 1). We retained all tags seen at least 10 times (i.e. occnb $\geq$10) in at least one DGE sample and covered by at least 3 RNA-Seq reads. Each tag was annotated to give its chromosome position, orientation, expression level in 'TranscriRef', tiling array information, proximity to coding and non-coding genes (at both 3′- and 5′-ends) and coverage by RNA-Seq reads. These features were then used for biological selection and validation by real-time quantitative PCR (qPCR). The selection criteria can be adjusted according to the user's requirements. Selected candidates were verified *in silico* by integrating our personal DGE and RNA-Seq data in the Ensembl Genome Browser using a DAS server. For tags corresponding to protein-coding genes that are conserved in human and mouse, enrichment analysis of the predicted GO-MF was carried out using the DAVID database with default parameters and functional annotation chart report (23).

Clustering

Hierarchical clustering was performed with the Cluster and Treeview software packages (24). To obtain a

homogeneous proportion of data from different tissues, we removed the over-represented libraries. Thus, only 33 of the 54 available libraries (highlighted in green, Supplementary Table S1) were used, and each individual tissue was represented at most by three different libraries. Tissue-specific tags were selected with the percentile method. As each tissue was represented in ~10% of all libraries (3/33), a 90% percentile was used for clustering. A compressed archive (clustering-digitagCT_transcriRef_34000intergenic_tags.tar.gz) is available at http://www.get.univ-montp2.fr/en/ncRNA with the 'Clustering-digitagCT_transcriRef_34000intergenic_tags.txt' file containing the 34 000 tags used for clustering and all the files generated by the clustering program. To cluster the 34 000 tags, the following command line was used with the clustering program: cluster –f Clustering-digitagCT_transcriRef_34000intergenic_tags.txt -l -cg a -ca a -g 1 -e 1 -m a.

Cell lines and culture

The embryonic kidney HEK 293, the human chronic myelogenous leukaemia K562 and the human AML U937 cells are routinely grown in our laboratory. Briefly, cells were cultured in RPMI 1640 (Invitrogen, Carlsbad, CA, USA) supplemented with 10% foetal calf serum at 37°C in a humidified atmosphere containing 5% CO_2 and 95% air. U937 cells were treated (U937 DIFF) or not (U937 NT) with 10 μM 1,25-$(OH)_2$ Vitamin D3 (Sigma), 10 μM TTNBP {4-[E-2-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-1-propenyl]} benzoic acid and 1 μM targretin (LGD1069), a selective retinoid X receptor agonist (both from Roche Pharmaceutics), for 48 h.

The HD291 hESC line was derived in the Institute for Research in Biotherapy (Montpellier, France) according to our previous report (25), and the M4C2 human iPS cell line was obtained by reprogramming new-born human foreskin fibroblasts using lentiviral vectors to express human OCT4/POU5F1, SOX2, NANOG and LIN28 (26). HD291 and M4C2 cells were cultured in 80% KO-DMEM, 20% KOSR, 2 mM L-glutamine, 1% non-essential amino acids, 0.5 mM β-mercaptoethanol (all from Gibco Invitrogen, Cergy-Pontoise, France), complemented with 10 ng/ml bFGF (Abcys, Paris, France) and maintained on irradiated (40 Gy) human foreskin fibroblast feeders. Both lines were mechanically passaged weekly. We also used SH-SY5Y (human neuroblastoma cancer), MCF7 (human breast cancer) and MDAPCa 1 (human prostate cancer) cells that were provided by S. Marchal and D. Noel (Montpellier, France).

RNA extraction, reverse transcription and real-time quantitative PCR

Total RNA was extracted with the RNeasy kit and included a DNase treatment (Qiagen). RNA quality and quantity were analysed using a 2100-Bioanalyzer (Agilent Technologies, Waldronn, Germany). Reverse transcription was performed with random primers (High-capacity cDNA Archive kit; Applied Biosystems, Courtaboeuf, France) using 1 μg of total RNA (or water, negative control) according to the manufacturer's instructions. qPCR was performed in 384-well plates (Sorenson BioScience, Inc.) on a Lightcycler® 480 Real-Time PCR System (Roche Diagnostics). One microlitre of each cDNA sample (1/10 dilution) was added to a 5 μl of reaction mix containing 3 μl of Master Mix (Roche Diagnostics) and 0.33 μM forward and reverse primers. Primer sequences were designed using the Primer3Plus software and are listed in Supplementary Table S3. Amplifications were carried out according to the following conditions: 95°C for 5 min, then 55 cycles as follows: 95°C for 10 s, 60°C for 10 s and 72°C for 10 s. At the end, a melting curve from 95°C to 65°C was performed to control primer specificity. The relative quantity (RQ) of gene expression was analysed using the $2^{-\Delta\Delta Ct}$ method (27). Transcriptional modulation ($\log_{10}$RQ) was calculated using the expression data of *RPS19* as endogenous control.

RESULTS

Global transcription distribution in the human genome (process A)

We developed a computational bioinformatics tool called digitagCT pipeline (Figure 1) that combines DGE data with the available Ensembl annotations to obtain an overall view of transcription distribution across the human reference genome. For this purpose, DGE data from 54 human tissue libraries (25 normal and 29 cancer tissues) were compiled in the 'TranscriRef' data set. These DGE libraries represent ~268 million 21-bp-long sequences and correspond to ~5 million distinct tags (Supplementary Table S1). Tags were then mapped to the human genome using the crac software in perfect match mode. Of all sequences, ~147 million (750 110 distinct tags) matched to unique genomic locations, ~69 million (142 786 distinct tags) showed multiple matches and ~50 million (4 247 867 distinct tags) had no match in the genome. Although only 15% of distinct tags had a single genomic location, they represented >50% of all sequences (147 million over 268 million occurrences). As mentioned in previous reports, the major cause of unmapped tags is sequence errors, and erroneous tag locations have been estimated to concern only 2% of the located sequences (19,28).

For each unique mapped sequence, a genomic annotation (according to protein-coding information) was extracted (process A of Figure 1 and Table 1) (21). Briefly, priority was given to tags located in gene exons in the sense orientation, then to intronic or antisense positions and finally to intergenic regions (Figure 2). The distribution of genomic annotations of the 'TranscriRef' data set represented the abundance of the different classes in this transcriptional repertoire (Figure 2, pie chart). The vast majority of mapped sequences (~89%) originated from exons of protein-coding genes and their precise location was also established (i.e. 3′UTR, CDS and 5′ UTR) (Figure 2, pie chart). As the DGE technology generates stranded tags preferentially in the 3′ part of polyA+ mRNAs, most of them matched the 3′ UTR (~69%) and less frequently the CDS (12.19%) or the 5′ UTR

Table 1. Occurrence, proportion (%) and number of tags of all 'TranscriRef' sequence according to the specific annotation processes

A (Process A)

Type	Nb Occ.	(in %)	Nb Tags
1: 3' UTR sense	101 651 692	68.73	78 227
2: CDS sense	18 032 580	12.19	44 334
3: 5' UTR sense	594 999	0.40	4690
4: INXON sense	526 835	0.36	3529
5: INTRON sense	5 375 383	3.64	202 446
6: 3' UTR antisense	8 543 983	5.78	47 610
7: CDS antisense	1 668 396	1.13	33 765
8: 5' UTR antisense	1 429 397	0.97	16 072
9: INXON antisense	125 319	0.08	2712
10: INTRON antisense	2 716 927	1.84	123 399
11: Intergene proximal	2 087 546	1.41	25428
12: Intergene distal	5 138 478	3.47	167 879
Total	147 891 535	100.00	750 091

B (Process B)

Type	Nb Occ.	(in %)	Nb Tags
A: Small ncRNA	31 579	0.89	450
B: lincRNA	648 260	18.18	10 260
C: Other lncRNA	2 617 656	73.42	29 770
D: Other non-codingRNA	267 929	7.51	772
Total	3 565 424	100.00	41 252

C (Process A x B)

Class	A: Small ncRNA	B: lincRNA	C: Other lncRNA	D: Other ncRNA	Total
1: 3' UTR sense	3668	750	77 072	2368	83 858
2: CDS sense	0	17	177 899	554	178 470
3: 5' UTR sense	126	13	1439	72	1650
4: INXON sense	23	1610	9046	8	10 687
5: INTRON sense	7700	15 566	146 950	1070	171 286
6: 3' UTR antisense	1457	29 014	473 029	2443	505 943
7: CDS antisense	117	3798	63 238	329	67 482
8: 5' UTR antisense	225	1361	166 397	126 087	294 070
9: INXON antisense	0	101	59 609	278	59 988
10: INTRON antisense	1400	8528	215 160	6876	231 964
11: Intergene proximal	2323	32 943	129 600	71 994	236 860
12: Intergene distal	14 540	554 559	1 098 217	55 850	1 723 166
Total	31 579	648 260	2 617 656	267 929	3 565 424

Type	Nb Occ.	(in %)	Nb Tags
1: 3' UTR sense	83 858	2.35	1089
2: CDS sense	178 470	5.01	155
3: 5' UTR sense	1650	0.04	57
4: INXON sense	10 687	0.30	65
5: INTRON sense	171 286	4.81	3146
6: 3' UTR antisense	505 943	14.19	3080
7: CDS antisense	67 482	1.89	1504
8: 5' UTR antisense	294 070	8.26	959
9: INXON antisense	59 988	1.68	160
10: INTRON antisense	231 964	6.50	6773
11: Intergene proximal	236 860	6.64	1366
12: Intergene distal	1 723 166	48.33	22898
Total	3 565 424	100.00	41 252

Detailed distribution, proportion and occurrence of 'TranscriRef' DGE tags with a unique match on the human genome. (A) Genomic distribution and occurrences of DGE tags assigned to coding transcripts (process A). (B) Genomic distribution and occurrences of DGE tags assigned to non-coding transcripts (process B). (C) Global distribution and occurrences of DGE tags assigned to non-coding transcripts (process A × B).

Table 2. Comparison of the 'TranscriRef' and Ensembl annotations

Categories	TranscriRef tag	TranscriRef gene	Ensembl gene	Percentage (TranscriRef/Ensembl)
Process A				
Protein coding	548 998	18 282	20 075	91.07
Pseudogene/polymorphic	7 672	2 786	12 550	22.06
IG/TR gene/pseudogene	114	64	562	11.38
Total	556 784	21 132	33 196	63.66
Process B				
Small ncRNAs				
miRNA	31	31	1 756	1.77
rRNA	10	8	530	1.51
snoRNA	213	90	1 521	5.92
snRNA	135	10	1 944	0,51
Small nc-pseudogene	61	61	1 835	3,32
lincRNAs				
lincRNA	10 260	2 450	4 883	50.17
Other lncRNAs				
Sense intronic	481	203	456	44.52
Processed transcript	16 332	1 645	2 076	79.24
Antisense	12 957	2 499	3 892	64.21
Other ncRNAs				
Sense overlapping	358	109	136	80.15
Non-coding	345	53	101	52.48
misc RNA and pseudogene	28	5	1 190	0,42
ncRNA host	41	9	19	47.37
Total	41 252	7 194	20 365	35.33

The output of the 'TranscriRef' library (after process A or B) was compared with the Ensembl database version 66. The first two columns describe the distribution of 'TranscriRef' tags and genes, respectively. The third column describes the distribution of Ensembl genes and the fourth column represents the 'TranscriRef' genes/Ensembl genes ratio.

(0.4%) of a gene sequence (29). The protein-coding gene fraction included 556 784 tags corresponding to 91% of the protein-coding HUGO terms identified in Ensembl (18 282 of the 20 075 HUGO terms), thus indicating that 'TranscriRef' represents a large transcriptome data set (Table 2). Transcribed regions were also identified in the intronic (5.48%) and intergenic (4.88%) categories as well as in the exon antisense class (7.96% of all TranscriRef occurrences with a unique match in the human genome). In parallel, we also analysed separately two DGE libraries included in the 'TranscriRef' data set to study transcriptome variability in specific tissues. The first library was generated using normal hESCs (hESC-hs0238: 3 636 083 total sequences and 293 179 unique tags) and the second one using a peripheral blood sample from a patient with AML (AML-hs0430: 6 399 705 total sequences and 204 169 unique tags). The distribution of tags in exonic, intronic and intergenic regions (90%, 4–5% and 5–6%, respectively) in these two libraries was comparable with the overall profile obtained for the whole 'TranscriRef' data set (Supplementary Figures S2 and S3).

Distribution of non-coding annotated transcripts in the human genome (process B)

We then evaluated the proportion of Ensembl-annotated non-coding transcripts by considering the 'TranscriRef' tags that matched non-coding sequences (process B of Figures 1 and 2). Non-coding transcripts amounted to 41 252 distinct tags, corresponding to 7194 non-coding genes among the 20 365 (35.33%) listed in Ensembl (Table 2). Non-coding transcripts were classified in four groups (as defined in 'Materials and Methods' section): (i) small ncRNAs (mainly miRNAs, snoRNAs and snRNAs), (ii) lincRNAs, (iii) other lncRNAs (processed, antisense and sense intronic transcripts) and (iv) other ncRNAs (Figure 2 and Table 2). As expected for DGE data, the proportion of small ncRNAs was negligible (0.96%). Within the 'TranscriRef' data set, lincRNAs represented ~34% of all non-coding genes (2450 of 7194) and were thus over-represented compared with Ensembl data set (24%: 4883 of 20 365 non-coding genes) (Table 2).

In the hESC-hs0238 and AML-hs0430 DGE libraries, 4% of all tags corresponded to annotated ncRNAs, independently of the cell type (normal or cancer), and were similarly distributed as those obtained for 'TranscriRef' (Supplementary Figures S2 and S3).

We then repeated the analysis by allowing a tag to overlap both coding and non-coding annotations. This procedure enabled to extract not only overlapping annotations but also transcripts overlapping with non-annotated genomic regions (Figure 3A and Table 1). A negligible fraction of the non-coding sequences (5%) was in cds regions, irrespective of the tag orientation. Conversely, non-coding sequences were abundantly represented in the antisense categories. For example, 20% of the 5′ UTR antisense occurrences (i.e. 294 070/1 429 397) overlapped with non-coding transcripts.

As expected, most non-coding transcripts (33.5%) overlapped with intergenic distal regions (1 723 166 of the 5 138 478 occurrences): ~50% were lincRNAs and the rest belonged mostly to the 'other lncRNAs' category. Interestingly, in this category, 3 415 312 sequences remained without annotation and could thus

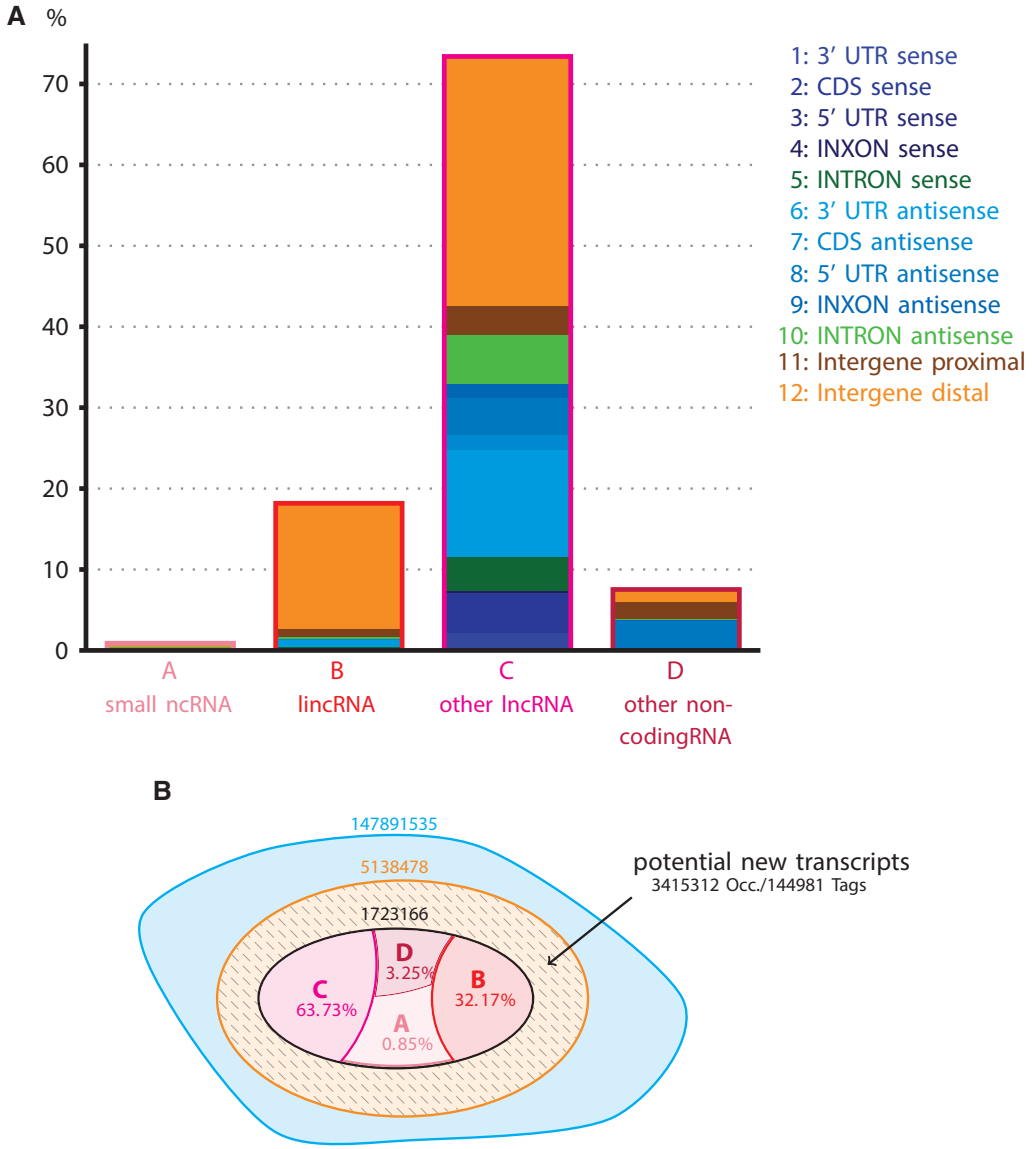


Figure 3. Distribution of overlapping non-coding transcripts (process A x B) and non-annotated sequences in 'TranscriRef'. (**A**) Bar chart describing the proportion of DGE sequences assigned to different non-coding transcript categories in each genomic region. LincRNA and other lncRNA sequences are more abundant in the intergenic distal regions. (**B**) Pie chart showing the distribution of all 'TranscriRef' sequences in the human genome. The first inset pie chart represents all intergenic 'TranscriRef' sequences. The second inset corresponds to the four identified categories of non-coding transcripts in intergenic regions (small ncRNAs, lincRNAs, other lncRNAs and ncRNAs) with their relative proportions (%). The difference between the two pie charts represents orphan sequences without annotation that could correspond to potential new transcripts or methodological artefacts. New non-annotated sequences are more abundant than the identified non-coding sequences.

correspond to either potential new transcripts or methodological artefacts (Figure 3B).These transcripts corresponded to 144 981 ('TranscriRef' dataset), 8131 (AML-hs0430 library) and 6273 (hpESC-hs0238 library) intergenic tags (Supplementary Figures S2 and S3).

The filtering step of the digitagCT pipeline: selection with high accuracy of new tissue-specific intergenic transcripts

To address the question of the biological significance of intergenic transcripts and to avoid the problem of false locations that could generate random annotations (methodological artefacts), we optimized our data mining procedure by integrating other available transcriptomic data, as described in 'Materials and Methods' section (filtering and species-comparison processes).

We first tested the species-comparison process (an optional filtering procedure) to validate the hypothesis that sequence conservation could provide interesting biological information. For this purpose, we investigated tags conserved between the human and the mouse genomes. We intersected all the human 'TranscriRef' tags with a unique location with a collection of murine DGE/SAGE libraries to select tags that are expressed in both species. In all, 8705 distinct tags corresponding to 6 363 596 input sequences were expressed and located in

both genomes. The distribution of annotation categories for this set of tags was comparable with that of the whole 'TranscriRef' data set (Supplementary Figure S5). We then considered the subset of common exonic tags (in the sense orientation for orthology analysis). Among the 3268 protein-coding gene tags (4 005 620 sequences), 96.1% corresponded to orthologous genes. We then thoroughly analysed their functional relevance using the DAVID database. The same analysis was performed with common intronic or antisense tags, and the enrichment in GO molecular functions is reported in Supplementary Table S2. Not surprisingly, this analysis revealed enrichment in GO annotations that correspond to highly conserved molecular functions, thus validating our selection and annotation processes. The three categories of conserved tags (exon, intron and antisense) showed common GO function enrichment profiles, such as nucleotide binding, transcription factor activity or RNA polymerase activity. Interestingly, the antisense group presented specific annotations, such as pyrophosphatase activity, hydrolase activity or nuclear hormone receptor binding. Finally, species-comparison selection also revealed 293 conserved non-coding transcripts, including 40 lincRNAs and 222 other lncRNAs. Moreover, we identified 486 potential new intergenic transcripts that were conserved and expressed in both species.

Other parameters could be used and collected for each intergenic tag, based on its chromosome position and strand orientation. Complementary data could be associated, including expression level (i.e. occurrence in SAGE Genie and 'TranscriRef' data sets), tiling array information, proximity to coding and non-coding genes on both 3′- and 5′-ends and coverage by RNA-Seq reads. To validate these parameters in the filtering process for identifying potential tissue-specific transcripts, first the hESC-hs0238 DGE and hpSC RNA-Seq libraries were analysed and 36 potentially relevant intergenic tags were randomly chosen (see 'Manual validation and curation' in the Materials and Methods section for the selected parameters). Then, analysis of their distribution in the 54 available libraries allowed to classify them as ubiquitous or ESC-specific (Figure 5A). The 36 selected intergenic candidates were not annotated by Ensembl version 62 at the time of the screening. Following the release of the current version 66, 16 of these tags have been annotated (6 are associated with antisense transcripts, 3 with intronic transcripts and 7 with newly annotated genes or pseudogenes), whereas the other 20 tags are still non-annotated. We measured their expression level by qPCR in normal and cancer cell lines and validated 78% of all candidates (90% when considering only the 20 non-annotated intergenic tags) (Supplementary Table S3). We further categorized the selected candidates by manual curation using the Ensembl display window with the DGE and RNA-Seq data. Among the annotated tags, some corresponded to new non-coding variants (Figure 4A), and among the non-annotated ones, a large fraction of candidates were lncRNAs (longer than 500 bp) (Figure 4B and Supplementary Figure S4) and a few could correspond to short exons or small RNAs (Figure 4B and Supplementary Figure S4).

As the coverage by RNA-Seq data could be used for tissue-specific selection, we tested specifically the hpSC data. Clustering analysis was performed on a selected data set of intergenic distal tags (see 'Materials and Methods' section for details). This set included ~34 000 potential new transcribed regions (Figure 5B). Clustering allowed 'checking' tissue specificity. As highlighted by the focus in Figure 5B, all embryonic cell libraries clustered together and showed a tissue-specific pattern with 1121 highly expressed tags. When applying the digitagCT pipeline filtering process (by comparing tags with the HD291 human embryonic stem cell RNA-Seq reads) to extract the more reliable transcripts, we identified 524 intergenic tags that were covered by both DGE and RNA-Seq (~50% of the 1121 highly expressed tags). Of note, the majority of the 1121 intergenic tags represented novel ncRNAs (931 candidates including 423 tags detected by RNA-Seq), whereas only 190 tags corresponded to already annotated non-coding transcripts (Supplementary Figure S6). All these data are available at http://www.get.univ-montp2.fr/en/ncRNA.

Looking at the nearest 5′ and 3′ genes, we observed that intergenic tags were often neighbours of lincRNAs, indicating transcriptional clusters. The search of neighbouring genes could be an interesting mean to provide potential functional information, as illustrated by the case of the *PIWIL4* neighbours that are specifically expressed in hESCs (Figure 4A).

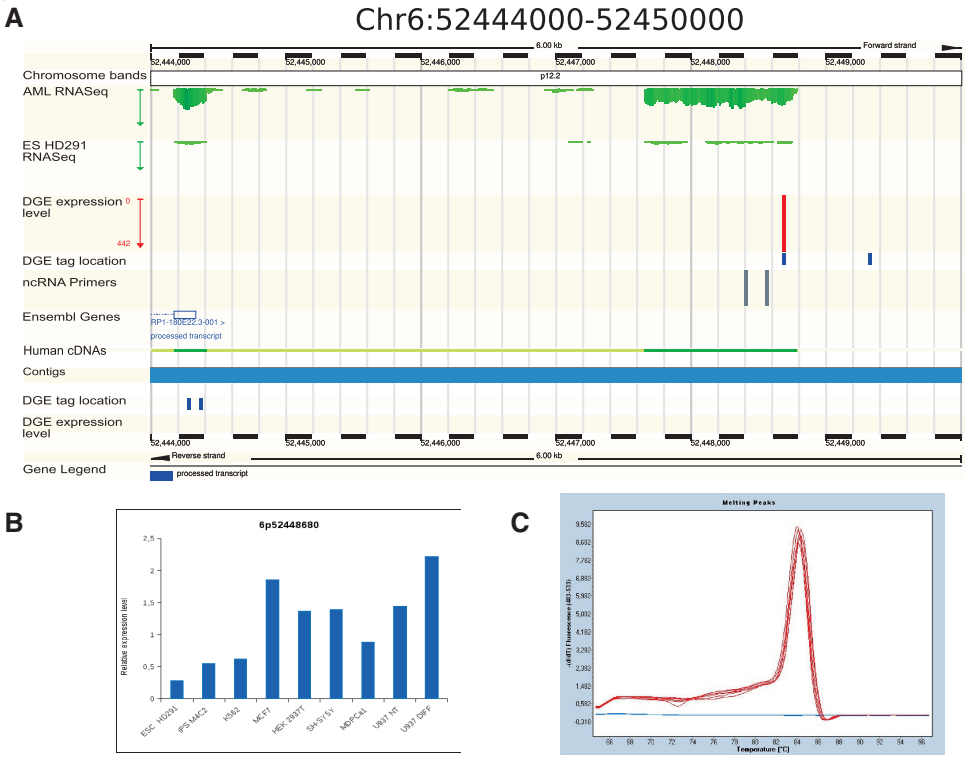
DISCUSSION

By combining gene expression data and next-generation sequencing-based assays, our digitagCT pipeline provides a high-quality catalogue of human transcriptome data and allows the easy selection of tissue-specific candidates. The high-throughput assays generate huge data volumes, but also many artefacts that blur the biological signal. Comparison and integration of complementary transcription data eliminate technology-specific errors and also reinforce the biological significance of newly identified RNAs. Moreover, our filtering process allowed determining the expression profile (in a variety of conditions/tissues, or only in specific tissues) of these newly identified RNAs, suggesting that they are transcriptionally regulated. This is possible only by simultaneously interrogating distinct gene expression data sources.

The proposed strategy avoids false locations and annotations

Although the analysis of DGE tags for transcript identification and characterization is now considered as a 'resolved' issue (30), most procedures include an approximate genomic localization that generates false positives. In the present report, the procedure for annotating DGE transcripts is based on previous results showing that the probability of false location in the human genome for a tag of 21 nt in length is minimized by using a perfect match approach (19). As a consequence, most of the erroneous tags were not mapped to the genome and were discarded. We also considered the impact of tags

First example



Second example

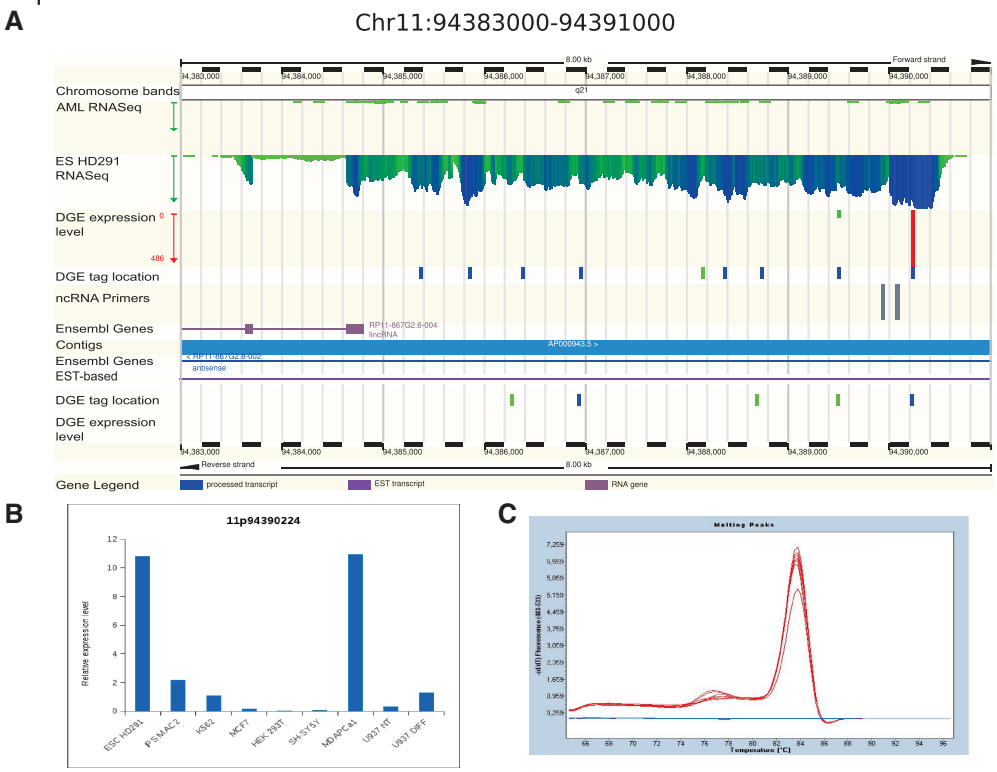


Figure 4. Examples of new non-annotated transcripts. (**A**) Display of the Ensembl Genome Browser web page for new non-annotated transcripts. The blue horizontal bars represent chromosomes. Gene structures ('Ensembl Genes', 'Human cDNAs', 'EST-based' tracks) are annotated by Ensembl. Public and private DGE data ('DGE tag location' track: blue rectangle for occurrence ≥2, green for occurrence = 1) are displayed on both strands of the chromosome with their relative occurrences (histogram of 'DGE expression level' track) using a private DAS server. The histogram of the RNA-Seq coverage (private data: RNA-Seq for hpSC and AML) in the chromosomal region is displayed on the top ('ES HD291 RNASeq' and 'AML RNASeq' tracks). (**B**) Relative expression of new transcripts in different cell lines validated by qPCR. (**C**) The corresponding melting curve analysis.

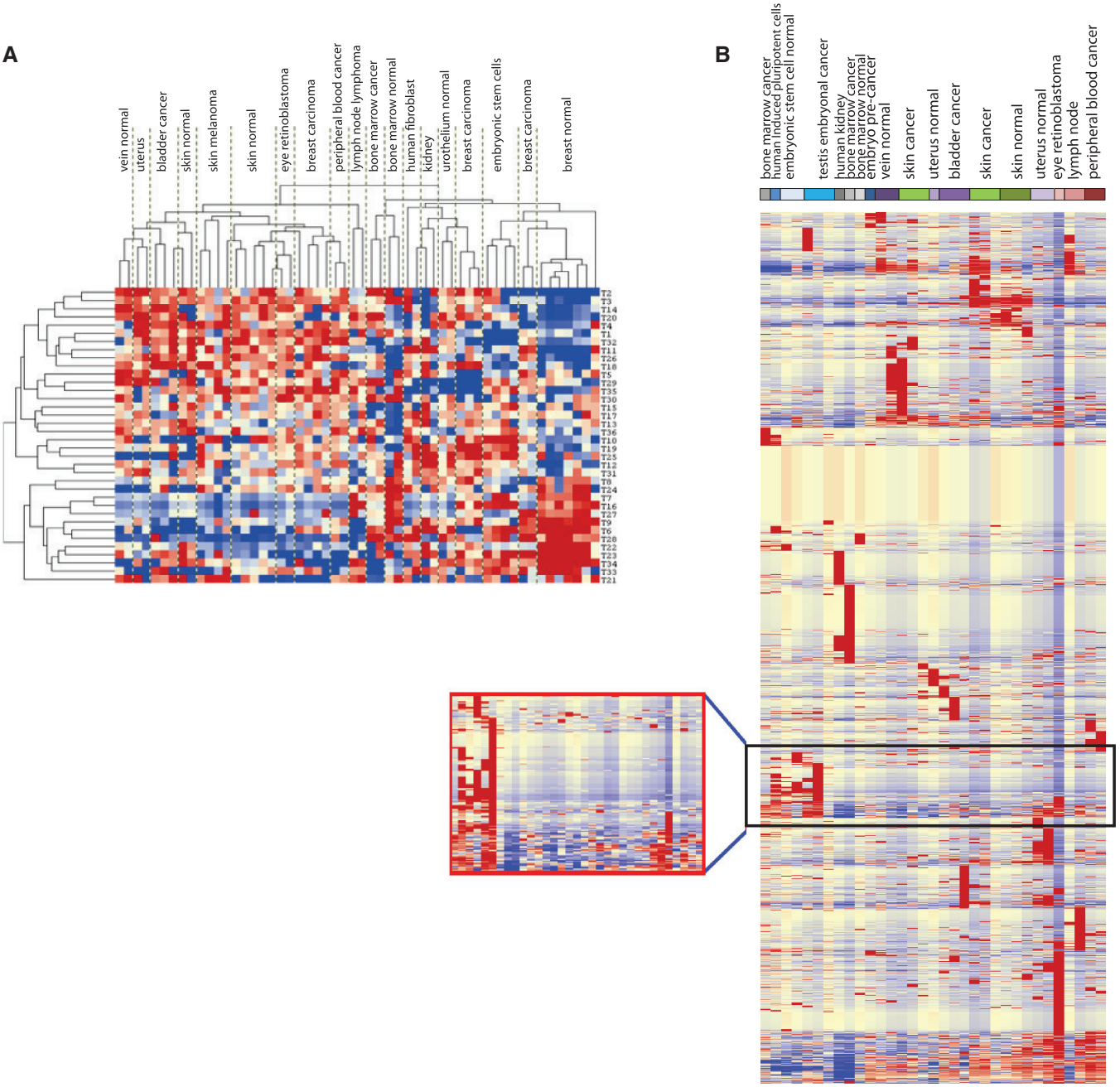


Figure 5. Clustering of 'TranscriRef' intergenic tags across different tissues. (**A**) Heat map of the expression level of the 36 tags randomly selected from the hESC-hs0238 DGE and hpSC RNA-Seq libraries in the 54 DGE libraries. (**B**) Heat map showing the expression level of the 34 000 intergenic tags used for the clustering analysis in 33 DGE libraries.

generated from inner enzymatic sites when the anchoring enzyme fails to cleave the expected canonical site. We have previously estimated this experimental artefact and showed that it could correspond to a frequency of 0.1% (31). This effect can be minimized by taking into account the tag occurrence. For instance, for a highly expressed transcript (10 000 occurrences of the canonical tag), the non-canonical enzymatic cutting generated only 10 occurrences, thus strongly reducing the weight of second rank tags in the global sequence distribution. Moreover, the total number of tags from 'TranscriRef' with a unique location (~750 000) included in the analysis represents an interesting biological data set of human transcription with strand orientation information.

Why combining DGE and RNA-Seq data?

By using RNA-Seq and DGE technologies, it is now possible to obtain accurate and dynamic ranges of transcript expression levels at the genome-wide scale with a level of sensitivity that is unachievable with any other technology. DGE reads arise from polyA+ RNAs within

the total RNA starting material and can be used to detect stranded transcripts and to provide an interesting approach for overall gene expression data mining. A DGE tag is a transcript signature. Hence, the tag allows determining the tissue expression profile of a transcript by examining its number of occurrence in each available DGE library and comparing it in all available libraries. DGE facilitates the comparison of large transcriptomic data sets and allows estimating the tissue-specific distributions and abundances. These features are particularly important, especially when looking for new non-coding transcripts, because they can suggest regulated and differential expression. On the other hand, RNA-Seq provides a digital measure of RNA abundance, which is represented by the sequence read counts in a region of interest and is independent from any pre-existing knowledge about the studied transcriptome. Considering how difficult it is to determine transcript boundaries especially in non-coding regions when transcription coverage is low (30) and how difficult it is to discriminate various isoforms at a specific gene location with the DGE method, the usage of RNA-Seq is a valuable complementary approach to discover and characterize new transcripts in specific tissues.

Potential novel ncRNAs constitute a significant fraction of polyA+ transcription

Several studies have shown that polyA+ intergenic transcripts are less varied and abundant than exonic ones, which codes for proteins (6), and that intergenic transcripts located near genes (which are qualified as proximal) are extensions of protein-coding genes. Our analysis confirmed this view and provided new insights on the distribution of overlapping coding and non-coding transcripts, thereby bringing converging evidence that the transcriptome is more complex than thought, as recently discussed in the review by Dinger *et al.* and largely demonstrated by the ENCODE project (32). Moreover, we report here ~34 000 intergenic transcripts that are located away from annotated genes, thus ruling out an origin as extension of protein-coding transcripts. This number exceeds the value published previously by van Bakel *et al.* and could be partly explained by the size and complexity of the compendium of DGE data used here. The extent and function of pervasive transcription is still a matter of debate, but our data provide new estimates on new polyA+ non-coding transcription that are in agreement with other studies (16,17). This is in line with the plethora of novel lncRNAs reported by Mercer *et al.* (13) showing that the non-coding part of a genome is a reservoir of lincRNAs. This finding was recently confirmed by the GENCODE v7 project (9). Moreover, our results on the mouse–human genome comparison are in favour of a conserved function of a non-negligible subset of lncRNA.

What are the biological features of novel ncRNA candidates?

The integration of other mammalian genomes could help investigating non-coding transcript conservation in several species and selecting new specific potential transcripts. In the present work, we investigated transcript conservation in the human and mouse genomes because their evolutionary distance is considered sufficiently high to allow separating conserved functional sequences that are under purifying selection from background neutral DNA (33). Using the digitagCT pipeline to investigate the conservation of protein-coding genes in the human and mouse genomes, we found significant enrichment of GO molecular functions. Moreover, by comparing conserved exonic, intronic and antisense transcripts, we show the presence of specific and common molecular functions, suggesting that conserved non-coding transcripts have essential functions. Specifically, conservation analysis of intergenic non-coding transcripts revealed 303 expressed and conserved novel ncRNAs, the function and tissue-specific expression of which could be further investigated. To our knowledge, this is the first demonstration that a 21-bp DGE tag can be efficiently used to target transcripts that are expressed and conserved between species. Moreover, the proportion of conserved intergenic tags was smaller than that of conserved protein-coding tags, in agreement with the lower conservation of lncRNAs compared with protein-coding genes (9,10).

As expected, annotated lincRNAs were in intergenic regions, and by mining 'TranscriRef', we could detect 50% of the annotated Ensembl data set. However, this DGE data set includes only polyA RNA+, while many non-coding transcripts have also been found using total RNA (8). It could be interesting to determine whether the lincRNAs described by the ENCODE project are present in our data set as well (9). A common feature of non-coding transcripts, and particularly of lincRNAs, is their low expression level. This led to the conclusion that their transcription level is not the best parameter for evaluating the biological involvement of non-coding transcripts (9,10). Another feature is their strong tissue-specific expression, illustrated in the present study by the 1121 regions that are specifically transcribed only in hpSCs.

Finally, our method is based on the use of a flexible and rigorous computing process for the analysis of a biological data set. This analysis is in itself a source of information that could be used to help optimizing the selection of new non-coding transcripts. Thus, our method provides new information on the diversity of the transcriptional repertoire in the human genome and an easy-to-use tool for selecting new tissue-specific non-coding transcripts by combining DGE and RNA-Seq transcription data. DGE data sets from other species (like parasites or plants) could readily be mined by using the digitagCT pipeline. Moreover, this approach could be extended to integrate other types of biological information, such as transcription binding sites and chromatin marks. Overall, we believe that this report brings convincing arguments in favour of data integration as a key for a more exhaustive exploitation of the data delivered by high-throughput technologies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Jérôme Audoux for help in packaging the digitagCT pipeline software. They acknowledge the Languedoc-Roussillon facility Montpellier GenomiX (MGX) and the Q-PCR facility.

FUNDING

Ligue Regionale contre le Cancer Languedoc-Roussillon, GEFLUC-Montpellier, Canceropole Grand Sud Ouest (GSO), Groupe Ouest Est d'Etudes des Leucémies et Autres Maladies du Sang (GOELAMS), CS Université Montpellier 2 (2011), CNRS INS2I [PEPS BFC: 66293]; Institute of Computational Biology, Investissement d'Avenir. Association de Recherche contre le Cancer [PDF20101202345 to N.P.]. Funding for open access charge: CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Johnson,J.M., Edwards,S., Shoemaker,D. and Schadt,E.E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, **21**, 93–102.
2. Willingham,A.T. and Gingeras,T.R. (2006) TUF love for 'Junk' DNA. *Cell*, **125**, 1215–1220.
3. Gerstein,M.B., Bruce,C., Rozowsky,J.S., Zheng,D., Du,J., Korbel,J.O., Emanuelsson,O., Zhang,Z.D., Weissman,S. and Snyder,M. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, **17**, 669–681.
4. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermüller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
5. Moran,V.A., Perera,R.J. and Khalil,A.M. (2012) Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res.*, **40**, 6391–6400.
6. Van Bakel,H., Nislow,C., Blencowe,B.J. and Hughes,T.R. (2010) Most 'Dark Matter' transcripts are associated with known genes. *PLoS Biol.*, **8**, e1000371.
7. Clark,M.B., Amaral,P.P., Schlesinger,F.J., Dinger,M.E., Taft,R.J., Rinn,J.L., Ponting,C.P., Stadler,P.F., Morris,K.V., Morillon,A. *et al.* (2011) The Reality of Pervasive Transcription. *PLoS Biol.*, **9**, e1000625.
8. Kapranov,P., St Laurent,G., Raz,T., Ozsolak,F., Reynolds,C.P., Sorensen,P.H., Reaman,G., Milos,P., Arceci,R.J., Thompson,J.F. *et al.* (2010) The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.*, **8**, 149.
9. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.
10. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
11. Rinn,J.L., Kertesz,M., Wang,J.K., Squazzo,S.L., Xu,X., Brugmann,S.A., Goodnough,L.H., Helms,J.A., Farnham,P.J., Segal,E. *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, **129**, 1311–1323.
12. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
13. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
14. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
15. Van Verk,M.C., Hickman,R., Pieterse,C.M.J. and Van Wees,S.C.M. (2013) RNA-Seq: revelation of the messengers. *Trends Plant Sci.*, **18**, 175–179.
16. Kapranov,P. and Laurent,G.S. (2012) Genomic 'dark matter': implications for understanding human disease mechanisms, diagnostics, and cures. *Front. Genet.*, **3**, 95.
17. Kowalczyk,M.S., Higgs,D.R. and Gingeras,T.R. (2012) Molecular biology: RNA discrimination. *Nature*, **482**, 310–311.
18. St Laurent,G., Savva,Y.A. and Kapranov,P. (2012) Dark matter RNA: an intelligent scaffold for the dynamic regulation of the nuclear information landscape. *Front. Genet.*, **3**, 57.
19. Philippe,N., Boureux,A., Bréhélin,L., Tarhio,J., Commes,T. and Rivals,É. (2009) Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Res.*, **37**, e104–e104.
20. Philippe,N., Salson,M., Commes,T. and Rivals,E. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.*, **14**, R30.
21. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
22. Rivals,E., Boureux,A., Lejeune,M., Ottones,F., Pérez,O.P., Tarhio,J., Pierrat,F., Ruffle,F., Commes,T. and Marti,J. (2007) Transcriptome annotation using tandem SAGE tags. *Nucleic Acids Res.*, **35**, e108–e108.
23. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
24. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
25. Assou,S., Cerecedo,D., Tondeur,S., Pantesco,V., Hovatta,O., Klein,B., Hamamah,S. and De Vos,J. (2009) A gene expression signature shared by human mature oocytes and embryonic stem cells. *BMC Genomics*, **10**, 10.
26. Bai,Q., Assou,S., Haouzi,D., Ramirez,J.-M., Monzo,C., Becker,F., Gerbal-Chaloin,S., Hamamah,S. and De Vos,J. (2012) Dissecting the first transcriptional divergence during human embryonic development. *Stem. Cell Rev.*, **8**, 150–162.
27. Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*, **25**, 402–408.
28. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
29. Morrissy,A.S., Morin,R.D., Delaney,A., Zeng,T., McDonald,H., Jones,S., Zhao,Y., Hirst,M. and Marra,M.A. (2009) Next-generation tag sequencing for cancer gene expression profiling. *Genome Res.*, **19**, 1825–1835.
30. Raz,T., Kapranov,P., Lipson,D., Letovsky,S., Milos,P.M. and Thompson,J.F. (2011) Protocol dependence of sequencing-based gene expression measurements. *PLoS One*, **6**, e19287.
31. Piquemal,D., Commes,T., Manchon,L., Lejeune,M., Ferraz,C., Pugnère,D., Demaille,J., Elalouf,J.-M. and Marti,J. (2002) Transcriptome analysis of monocytic leukemia cell differentiation. *Genomics*, **80**, 361–371.
32. Dinger,M.E., Gascoigne,D.K. and Mattick,J.S. (2011) The evolution of RNAs with multiple functions. *Biochimie*, **93**, 2013–2018.
33. Zhang,Z., Pang,A. and Gerstein,M. (2007) Comparative analysis of genome tiling array data reveals many novel primate-specific functional RNAs in human. *BMC Evol. Biol.*, **7**, S14.