Supplementary Online Text

Transcription-induced chimeras

We identified 136 loci (110 if we removed the target loci that are in gene clusters) with RACE extensions reaching exons of upstream loci (i.e. having RACEfrags overlapping upstream exons), thus creating transcripts that possibly encode chimeric versions of already annotated proteins. Often chimeric transcripts are tissue specific (45%, 51 out of 110 loci not part of gene clusters), but when chimeras for a given locus are detected in several tissues (chimeras are detected on average in 2.9 tissues/cell lines out of 15), they are preponderantly similar (an example is illustrated in Figure 3) and link the same two loci (49% of the loci for which we identified a chimera in multiple tissues exhibit only one type of chimera, 29 out of 59 loci). We have identified by RACEfrags a total of 155 different transcription-induced chimeras (1.4 chimeras per locus). Most of them link 2 loci (79%, 123/155), but some are linking 3, 4 or even 5 loci together (17.5%, 2% and 1.5% respectively). Remarkably, some genes tend to be incorporated in different chimeras (two on average); 174 unique genes are linked into the 155 chimeras, which contain 249 genes. 13 chimeras were validated by sequencing (11 of them linking adjacent loci, and 2 skipping other loci). Only one of these incorporates a novel exon, while the others only link known exons of the fused loci together.

Sequence features of the novel RACEfrags

The vast majority of the 225 novel exons are interrogated by the ENCODE tiling array over more than 50% of their length (Figure S1). As expected some do not overlap RACEfrags because they map to repeat-masked regions not present on the tiling chip, or because they are too short and GC poor. It is important to emphasize that the ENCODE Affymetrix tiling array comprises 14,707,189 bp from the non-repeated portion of the 44 ENCODE regions (49% of the entire ENCODE sequences){Consortium, 2007 #739}. As expected a limited subset (5%) of novel exons does not overlap RACEfrags but only RT-PCRfrags (probes hybridized by the RT-PCR reactions to check connectivity, see above and Figure 1).

The 57 novel internal (the first and the last exons of the RT-PCR products were not considered) exons have an average and median length of 149 and 118 bp, respectively (Figure S3A), a size comparable to the average and median size (145 and 122 bp, respectively) of internal exons reported by the human genome consortium {Lander, 2001 #57}, or the richer GENCODE annotation (177 and 124 bp, respectively). Interestingly, we observe that the pool of novel exons is significantly enriched in short exons (<60 bp; p<0.001; see Figure S3A). They have a decreased average (49.7%) and median (49.5%) GC content than those reported for GENCODE annotated exons (average: 53.5% and median: 54.1%; see Figure S3B). 146 novel introns were identified from the RT-PCR sequences. They have an average and median lengths of 32 kb and 12.3 kb, respectively, far above the 4.6 kb and 0.9 kb recorded for GENCODE annotated introns {Harrow, 2006 #726}. This bias towards long introns was expected, because we preferentially targeted distal extensions found by RACEfrags. Noticeably, 11% (16 out of 146) of the novel introns harbor non-canonical splice sites, a proportion much higher than that reported by the whole GENCODE annotation (most of those introns have one annotated canonical splice site and one novel non canonical splice site). Neverthless, 14% (20/146) of these new introns are supported by ESTs, most of them submitted after the GENCODE annotation release. Considering only the entirely novel exons (i.e. not a single nucleotide overlapping an already annotated exon) and both splice sites novel: 90 donors (89 canonical) and 48 acceptors (all canonical) were scored according to the human splice site substitution matrices (see Supplementary Online Materials and Methods section) and compared to the scores of GENCODE splice sites and false splice sites (Figure S4). The novel acceptors have higher scores than false acceptors (random AG; p< 2.2e-16), in the range of GENCODE UTR acceptors (p= 0.6241), and slightly lower than those of GENCODE CDS acceptors (p=0.04716). Similarly, novel donors score higher than false donors (random GT) (p< 2.2e-16), but as well as both GENCODE UTR and GENCODE CDS donors (p= 0.6041 and p=0.1070, respectively). In summary, the novel splice sites score as well as annotated UTR splice sites.

Supplementary Online Materials and Methods

RACE/array analysis of known protein-coding genes.

5'-RACEs were performed on polyA$^+$ RNAs from 12 human tissues (brain, heart, kidney, spleen, liver, colon, small intestine, muscle, lung, stomach, testis, placenta, all BD Clontech) and 3 cell lines (GM06990, HL60 and HeLaS3) using the BD SMART^{TM} RACE cDNA amplification kit (BD Clontech Cat. No.634914). Double-stranded cDNA synthesis, adaptor ligations to the synthesized cDNA and 25 μl final volume RACE reactions were performed according to the manufacturers' instructions. RACE oligonucleotides were designed with primer 3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) with the following parameters: 23 ≤ primer size ≤ 27, optimal size=25, 68°C ≤ primer Tm ≤ 72°C, optimal Tm = 70°C, 50% ≤ primer GC percentage ≤ 70%. 15 ul aliquots of 80 to 100 RACE reactions performed with oligonucleotides specific to non-neighboring genes and on the same tissue/cell line cDNA were assembled in pools, precipitated with ethanol and resuspended in water. 25 μg of RACE amplicons were fragmented with DNAse I to the size of 50-100 bp, denatured by heating to 99^{o}C for 10 minutes and end labeled with biotin using terminal transferase (TdT; Roche) in 35 μl under the following conditions: 1X TdT reaction buffer (Roche), 2.5 mM $CoCl_2$, 1.15 nmoles of Affymetrix DNA Labeling Reagent (DLR, cat. # 900542) per 1 μg of fragmented DNA and 200 units of TdT. The reactions were incubated for 2 hrs at 37^{o}C. 20 μg of labeled RACE DNA was hybridized to ENCODE tiling arrays as described in (Kapranov et al. 2005). RACE maps were generated using Tiling Analysis Software (TAS, http://www.affymetrix.com/support/developer/downloads/TilingArrayTools/index.affx). The maps were generated with no smoothing (bandwidth = 1) and no CEL file normalization. The RACEfrags were generated using probe intensity threshold of 100; maxgap = 30 and minrun = 20. Thus, minimal RACEfrag would contain two consecutive positive probes.

RT-PCR of RACEfrags

538 RACEfrags were selected for independent verification of their connectivity with the original annotated gene: (set 1) 255 RACEfrags corresponding to the longest extension in a tissue; (set 2a) 120 of these distal RACEfrags which are supported in at least two tissues (if not in set 1); (set 2b) 40 RACEfrags which appear most frequently in the highest number of tissues (if not in set 2a); (set 3) 90 RACEfrags which correspond to the second longest tissue-specific extension; and (set 4) 33 intronic RACEfrags. RT-PCR to verify these 538 RACEfrags were done either in Affymetrix Inc., Santa Clara (lab.A 282 RACEfrags) or the Universities of Geneva and Lausanne, Switzerland (lab.B 300 RACEfrags, 40 overlaps). RT-PCRs in lab B were performed on the oligo dT-primed cDNA using BD-advantage II polymerase mix and following the manufacturers' instructions (25 ul final volume). Note that the RNA used was the same as for the RACE reaction in which the RACEfrag was identified. The right primer was the original RACE primer and the left primer was designed with the same characteristics (see above) in the RACEfrag to be verified. ENCODE tiling arrays were used as a readout of the RT-PCR reactions. 15 ul aliquots of RT-PCR reactions were assembled in pools which contained a single reaction per ENCODE region. Pools of RT-PCR reactions were ethanol precipitated, resuspended in water, labeled and hybridized to the microarray as described above to control the connectivity between the RACEfrags and the original exon chosen to design the RACE oligonucleotide.

Of the 300 RACEfrags, oligonucleotides could only be selected for 283 by lab A. The 283 reactions in lab A were performed using gene-specific oligonucleotides for cDNA synthesis. cDNA synthesis was conducted on 10 ng of polyA+ RNA from a tissue where a corresponding RACEfrag was detected using the same oligonucleotide as used for 5' RACE analysis. The cDNA synthesis was performed with Thermoscript reverse transcriptase (Invitrogen) using the same conditions as described in (Kapranov et al. 2005) for 5' RACE cDNA synthesis. The cDNA reactions were purified using QIAquick 96 (Qiagen) and ½ of each purified reaction was used as a starting material for RT-PCR. For each RACEfrag, two rounds of nested RT-PCR reactions were performed. The products of first round of RT-PCR were purified using QIAquick 96 system, eluted in 80 μl and 0.01 μl of the first round reaction was used for the second round RT-PCR. Each round of amplification consisted of 30 cycles of PCR (94^{o}C for 20 sec; 60^{o}C for 30 sec; 72^{o}C for 2 min) followed by 10 min at 72^{o}C. Products of the final round of RT-PCRs were purified using QIAquick 96, pooled using the same strategy as in the lab B and hybridized to ENCODE arrays as described above.

In addition, RT-PCR reactions for 96 RACEfrags in lab A were done using oligo-dT cDNA as a substrate. PolyA+ RNA from brain, colon, heart, kidney, liver, lung and muscle were pooled and used for cDNA synthesis following the procedure used for cDNA synthesis for 3’RACE described in (Consortium 2007). The resulting cDNA was used for RT-PCR following the same PCR conditions as above. The RT-PCRfrags were generated using the same parameters as the 5’ RACEfrags for the known genes (see above) for both sets (Labs A & B).

Assignment of RT-PCRfrags

To score an RT-PCR as positive based on the profile of microarray hybridization, we used a two-way approach. First, an RT-PCR reaction was considered as positive if RT-PCRfrags could be found within 1 kb from both forward and reverse RT-PCR primer. 33% of the reactions were positive following this criterion. A separate scoring strategy was used on the reactions that did not pass this filter to account for the cases where an RT-PCR oligonucleotide was picked close to the boundary of the target RACEfrag or the target exon, thus resulting in the absence of RT-PCRfrags immediately proximal to the primer position: if 3 or more of the RT-PCR frags were overlapping the original RACEfrags from the tissue where the RT-PCR was performed, the reaction was recalled positive. Using both scoring strategies combined, about 58% of RT-PCRs were scored positive.

Cloning and Sequences of the RACE/array products

Two different strategies were employed to sequence the amplified transcripts that link tested RACEfrags and known exons. The RT-PCR reactions that appeared as single bands on agarose gel were selected for direct sequencing, while the others were cloned into pDRIVE following manufacturer's instructions (Qiagen) before sequencing of a minimum of eight clones. The reads were assembled after masking of the vector and mapped to the human genome using exonerate (unmasked, max intron length=1.5 Mb) to identify the best hit. The hit has to be more than 100 bp long, and with a %identity greater than 95%. From 2354 assembled sequences, 703 were spliced and mapped in the right target, corresponding to 353 non-redundant sequences (when several sequences were identical or included in each other, only one representative was kept). The following two filtering steps were applied to remove truncated sequences and those not reaching the borders of the target regions (the cloning could lead to a partial loss of the insert). First, at least 90% of the genomic span of a target region has to be covered by the RT-PCR sequence. Secondly, the RT-PCR sequence must not extend further than 100 bp outside the target genomic span. After these filtering steps, 175 unique sequences remained. They are deposited in GenBank under accession numbers DQ655905-DQ656069 and EF070113-070122. They were inspected manually by the annotators who provided the GENCODE annotation (Harrow et al. 2006) and dubious mappings were discarded, leading to a final set of 132 unique sequences. They correspond to 89 RT-PCR reactions and 69 loci. Note that the GENCODE annotation team gold standards to accept transcript sequences as evidence are conservative. For example they reject most of the transcripts with non-canonical splice sites. It is therefore possible that more sequences correspond to *bona fide* RT-PCR products, but that they escaped further analysis because they present some characteristics that are different from the ones harbor by known transcripts. Detailed information about the 132 sequences is available in Table S2, which provides links to the UCSC browser for vizualisation (available on line at http://genome.imim.es/GENCODE/RACEdb/Sequences_Description.html).

Overlaps of RACEfrags with other datasets: RACEfrags from 12 tissues

The RACEfrags were overlapped with 5'end related datasets produced by the ENCODE consortium: TSS 5'end clusters derived from CAGE (5'-specific Cap Analysis Gene Expression) tags and 5'PETs (Paired-End 5' and 3' di-Tags), composite promoters derived from ChIP-on-chip hits and DNAse I Hypersensitive sites (Hss) (Consortium

2007).

Four sets of RACEfrags were used:

1. 1390 projected RACEfrags from 12 tissues (on which the RT-PCR were performed) external to the locus, not yet annotated as 5' ends (i.e not overlapping annotated first exons): they represent a mixture of 5'ends and internal new exons.
2. 60 RACEfrags corresponding to the subset of the 1390 RACEfrags that are in the set of sequenced exons (obtained by RT-PCR followed by cloning and sequencing) from the considered experiment and not yet annotated as 5'ends.
3. 584 RACEfrags corresponding to the RACEfrags that are the most distal for each locus per tissue were extracted: this set (subset of the set of 1390 RACEfrags) does not necessarily contain only 5'ends because the length of the ENCODE regions and the distance between genes in the pools limit the size of the observable extensions, and also because of the conservative filtering of RACEfrags, that could have discarded the most distal ones. However, it is likely to be enriched in 5'ends compared to the previous set.
4. 31 RACEfrags corresponding to the subset of 584 RACEfrags that are in the set of sequenced exons (obtained by RT-PCR followed by cloning and sequencing) from the considered experiment and not yet annotated as 5'ends.

The percentages of RACEfrags having 1bp overlap with the other sets (stranded when the dataset contained a strand information) were calculated for the three RACEfrags sets as well as for random sets (100 random sets mimicking each of the sets) to compare the random overlap to the observed overlap. All overlaps are significant (P-values<0.01) except the overlap of the 60 RACEfrags with TSS for which the P-value is 0.06 (Figure 6).

Overlaps of RACEfrags with other datasets: HL60 RACEfrags

All HL60 RACEfrags not yet annotated as first exons (791) were overlapped with ChIP-on-chip hits (Consortium 2007) obtained from HL60 cell line, using the same Affymetrix chips as were used for RACEfrags. The ChiP-on-chip hits coordinates were downloaded at the UCSC genome browser (http://genome.ucsc.edu/encode/), they correspond to the following tracks:

1. Brg1 retinoic acid-treated HL-60, 0hrs (Brahma-related Gene 1)
2. CEBPe retinoic acid-treated HL-60, 0hrs (CCAAT-enhancer binding protein-epsilon
3. CTCF retinoic acid-treated HL-60, 0hrs (CCTC binding factor)
1. H3K27me3 retinoic acid-treated HL-60, 0hrs (Histone H3 tri-methylated lysine 27)
2. H4Kac4 retinoic acid-treated HL-60, 0hrs (Histone H4 tetra-acetylated lysine)
3. P300 retinoic acid-treated HL-60, 0hrs (E1A-binding protein, 300-KD)
4. PU1 retinoic acid-treated HL-60, 0hrs (Spleen focus forming virus proviral integration oncogene)
5. Pol2 8WG16 antibody, retinoic acid-treated HL-60, 0hrs (RNA Polymerase II, 8WG16 ab against pre-initiation complex form)
6. RARA retinoic acid-treated HL-60, 0hrs (Retinoic Acid Receptor-Alpha)
7. SIRT1 retinoic acid-treated HL-60, 0hrs (Sirtuin-1)
8. H3K9K14ac2, retinoic acid-treated HL-60, 0hrs Strict Sites (Histone H3 K9 K14 Di-Acetylated)
9. H4Kac4, retinoic acid-treated HL-60, 0hrs Strict Sites (Histone H4 tetra-acetylated lysine)
10. Pol2, retinoic acid-treated HL-60, 0hrs Strict Sites (RNA Polymerase II, 8WG16 ab against pre-initiation complex form)
11. actinomycin-D treated p63 HL-60 Strict Sites (p63 with actinomycin D treatment)
12. p63, HL-60 Strict Sites (p63 without actinomycin D treatment)

The proportion of RACEfrags overlapping the hits on at least one base pair was compared to the overlap obtained from a set of randomly distributed RACEfrags mimicking HL60 RACEfrags in order to calculate the significance of the overlaps (Figure 7).

Supplementary Online References:

The ENCODE Consortium. 2007. The ENCODE pilot project: Identification and analysis of functional elements in 1% of the human genome. *Nature* submitted.

Harrow, J., F. Denoeud, A. Frankish, A. Reymond, C.K. Chen, J. Chrast, J. Lagarde, J.G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, and R. Guigo. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7 Suppl 1: S4 1-9.

Kapranov, P., J. Drenkow, J. Cheng, J. Long, G. Helt, S. Dike, and T.R. Gingeras. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* 15: 987-997.

Table S1 : Description of the transcripts obtained by sequencing of RT-PCR products on RACEfrags

	Among 132 RT-PCR sequences obtained				Among 69 loci with RT-PCR sequence(s) obtained *			
Transcript structure	Number of (re) annotated transcripts	Number of transcripts (re)annotated as Coding (a CDS was assigned)	Number of transcripts with Novel CDS assigned (% of all)	Number of transcripts for which a potential new CDS is detected but was not annotated (% of all)	Number of loci with (re) annotated transcripts	Number of loci with transcripts (re)annotated as Coding	Number of loci with Novel CDS assigned (% of all)	Number of loci for which a potential new CDS is detected but was not annotated (% of all)
Only new internal exons	15	5	1 (6.7%)	6 (40%)	15	5	1 (6.7%)	5 (33.3%)
Extension of the first exon	24	17	1 (4.2%)	3 (12.5%)	18	14	1 (5.6%)	3 (16.7%)
New 5' exons (not chimeric)	65	23	8 (12.3%)	35 (53.8%)	34	18	7 (20.6%)	16 (47.1%)
Chimeric transcripts	28	15	14 (50%)	6 (21.4%)	13	7	6 (46.1%)	3 (23.1%)
Total*	132	60	24 (18.2%)	50 (37.9%)	69	40	16 (23.2%)	25 (36.2%)

Table S2 : description of the 132 sequences obtained from RT-PCR followed by cloning and sequencing.
(This table is also available at http://genome.imim.es/GENCODE/RACEdb/Sequences_Description.html)

RT-PCR ID	Locus ID	Gencode locus ID	Internal sequence ID	Genbank AC	(re)annotated Gencode transcript ID	RT-PCR set	Extension length (from RACEfrags)	Structure of the (re) annotated transcript	Type of the (re)annotated transcript	Novel CDS annotated	Potential new CDS (identified by automatic pipeline)	Novel exons ?
5RACE208_chr19_59847174_59847202	LILRB4	AC011515.1	UGL10a09	DQ655905	AC011515.1-010	2ab	18217	New exons upstream (not chimeric)	Coding	No	New M (ATG) upstream	yes, not coding
5RACE179_chr21_32906405_32906494	TCP10L	AP000274.7	UGLsupplB	DQ656039	AP000274.7-004	1	26780	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE179_chr21_32906301_32906360	TCP10L	AP000274.7	Affy08248C06	DQ655906	AP000274.7-004	2a	26646	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE179_chr21_32906301_32906360	TCP10L	AP000274.7	UGL9b12	DQ655908	AP000274.7-005	2a	26646	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE179_chr21_32901846_32901892	TCP10L	AP000274.7	Affy08252B09	DQ655909	AP000274.7-006	2b	22178	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE179_chr21_32901846_32901892	TCP10L	AP000274.7	UGL9a11	DQ655910	AP000274.7-004	2b	22178	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE318_chr5_142057771_142057784	FGF1	AC005370.1	Affy08248A11	DQ655914	AC005370.1-006	2ab	14	extension of the first exon	Coding	No		no
5RACE318_chr5_142057771_142057784	FGF1	AC005370.1	Affy08248D11	DQ655915	AC005370.1-010	2ab	14	extension of the first exon	Coding	yes (exon skipped)		no
5RACE318_chr5_142057771_142057784	FGF1	AC005370.1	UGL16a06	DQ655916	AC005370.1-011	2ab	14	extension of the first exon	Not coding	No		no
5RACE318_chr5_142057771_142057784	FGF1	AC005370.1	UGL16c06	DQ655917	AC005370.1-012	2ab	14	extension of the first exon	Not coding	No		no
5RACE318_chr5_142057771_142057784	FGF1	AC005370.1	UGL16d06	DQ655918	AC005370.1-003	2ab	14	extension of the first exon	Not coding	No		no

5RACE369_chr7_115906409_115906409	MET	AC002543.3	UGL369-F-G5	DQ656040	AC002543.3-005	2ab	1	intronic	Coding	Yes (new internal exon)		yes, partly coding (ATG)
5RACE027_chrX_122819563_122819627	STAG2	RP11-517O1.1	UGL1a10	DQ655919	RP11-517O1.1-020	1	382	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE027_chrX_122819904_122819944	STAG2	RP11-517O1.1	UGL23a02	DQ655921	RP11-517O1.1-002	2ab	41	extension of the first exon	Coding	No		no
5RACE027_chrX_122819904_122819944	STAG2	RP11-517O1.1	UGL23e02	DQ655922	RP11-517O1.1-006	2ab	41	extension of the first exon	Coding	No		no
5RACE145_chr5_131919535_131919635	RAD50	AC004041.1	UGL7g06	DQ655924	AC004041.1-006	2a	994	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE018_chr1_147983375_147983547	PIP5K1A	RP11-68I18.9	UGL018-A-G2	DQ656041	RP11-68I18.9-011	2ab	725	New exons upstream (not chimeric)	Coding	Yes (new ATG upstream)	New M (ATG) upstream	yes, partly coding (ATG)
5RACE145_chr5_131919940_131919961	RAD50	AC004041.1	UGL7g07	DQ655928	AC004041.1-006	2b	589	New exons upstream (not chimeric)	Coding	Yes (ATG downstream: shorter CDS)		no
5RACE290_chr7_126825724_126825744	FSCN3	AC073934.3	UGL15a05	DQ655929	AC073934.3-005	3	1916	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE183_chr21_33107870_33107935	C21orf62	AP000280.67	UGLsupplC	DQ656042	AP000280.67-001	2ab	66	extension of the first exon	Coding	No		no
5RACE261_chr7_26987975_26988041	HOXA9	AC004080.4	UGL14a03	DQ655931	AC004080.4-006	3	4684	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE185_chr21_33775307_33775371	C21orf4	AP000300.7	UGL10a02	DQ655932	AP000300.7-010	1	1216	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE290_chr7_126825405_126825453	FSCN3	AC073934.3	UGL22a07	DQ655934	AC073934.3-005	1	2235	chimeric	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE023_chr1_148067170_148067191	RP11-126K1.3	RP11-126K1.3	UGL1d08	DQ655935	RP11-126K1.3-004	2b	646	New exons upstream (not chimeric)	Coding	No	New M (ATG) upstream	yes, not coding
5RACE185_chr21_33774156_33774188	C21orf4	AP000300.7	UGLsupplD	DQ656043	AP000300.7-001	2ab	33	extension of the first exon	Coding	No		no
5RACE073_chr22_30216509_30216653	EIF4ENIF1	RP11-247I13.2	UGL073-A-G6	DQ656044	RP11-247I13.2-008	1	6289	New exons upstream (not chimeric)	Coding	No		yes, not coding

5RACE292_chr7_126665891_126665935	AC000123.1	AC000123.1	UGL15b06	DQ655936	AC000123.1-008	3	39212	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE004_chr1_148296037_148296096	CGN	RP11-74C1.3	UGL004-A-B1	DQ656045	RP11-74C1.3-005	2a	909	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE073_chr22_30210448_30210475	EIF4ENIF1	RP11-247I13.2	UGL2a12	DQ655937	RP11-247I13.2-009	2ab	111	New exons upstream (not chimeric)	Coding	No	New M (ATG) upstream (entirely open)	yes, not coding
5RACE004_chr1_148296464_148296527	CGN	RP11-74C1.3	UGL004-A-C1	DQ656046	RP11-74C1.3-006	2b	482	New exons upstream (not chimeric)	Coding	No	New M (ATG) upstream (entirely open)	yes, not coding
5RACE028_chr1_147980842_147980907	TCFL1	RP11-68I18.8	UGLsupplA	DQ656047	RP11-68I18.8-004	1	5168	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)(entirely open)	yes, not coding
5RACE074_chr22_30382940_30382976	PISD	RP5-858B16.2	UGL3a01	DQ655940	RP5-858B16.2-001	2a	196	extension of the first exon	Coding	No		no
5RACE028_chr1_147975740_147975772	TCFL1	RP11-68I18.8	UGL1a11	DQ655941	RP11-68I18.8-001	2ab	33	extension of the first exon	Coding	No		no
5RACE074_chr22_30382781_30382823	PISD	RP5-858B16.2	UGL074-A-H6	DQ656048	RP5-858B16.2-001	2b	43	extension of the first exon	Coding	No		no
5RACE112_chrX_152883245_152883267	MECP2	AF030876.1	UGL112-B-D5	DQ656049	AF030876.1-006	4	0	intronic	Not coding	No		yes, not coding
5RACE055_chr20_33725862_33725884	CPNE1	RP1-309K20.2	UGL055-A-E5	DQ656050	RP1-309K20.2-029	1	9592	chimeric	Not coding	No		no
5RACE112_chrX_152923396_152923422	MECP2	AF030876.1	UGL112-B-E5	DQ656051	AF030876.1-007	1	39363	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG) (entirely open)	yes, not coding
5RACE227_chr19_59573942_59573977	LAIR1	AC008746.1	UGL12d04	DQ655942	AC008746.1-005	2b	36	extension of the first exon	Coding	No		no
5RACE055_chr20_33725711_33725737	CPNE1	RP1-309K20.2	UGL2a06	DQ655943	RP1-309K20.2-031	3	9445	chimeric	Not coding	No	Skipping of the exon containing the M (ATG)	no
5RACE055_chr20_33725711_33725737	CPNE1	RP1-309K20.2	UGL2d06	DQ655945	RP1-309K20.2-029	3	9445	chimeric	Not coding	No		no

5RACE055_chr20_33725711_33725737	CPNE1	RP1-309K20.2	UGL2e06	DQ655946	RP1-309K20.2-030	3	9445	chimeric	Not coding	No	Skipping of the exon containing the M (ATG)	no
5RACE083_chr22_31778284_31778399	SYN3	LL22NC03-28H9.1	UGL3c10	DQ655947	LL22NC03-28H9.1-009	4	0	intronic	Coding	No		yes, not coding
5RACE241_chr16_386727_386755	NME4	Z97634.4	UGL12a10	DQ655948	Z97634.4-002	2ab	29	extension of the first exon	Not coding	No		no
5RACE201_chr19_59618039_59618247	TTYH1	AC008746.2	UGL10d04	DQ655949	AC008746.2-010	2ab	378	New exons upstream (not chimeric)	Coding	Yes (new ATG upstream)	Skipping of the exon containing the M (ATG)	yes, partly coding (ATG)
5RACE062_chr20_33792299_33792321	RNPC2	RP11-353C18.2	UGL062-A-H5	DQ656052	RP11-353C18.2-038	4	0	intronic	Not coding	No	New M (ATG) upstream (entirely open)	no (extends an existing exon)
5RACE069_chr22_30763573_30763634	SLC5A1	RP1-127L4.1	UGL069-A-E6	DQ656053	RP1-127L4.1-001	2ab	62	extension of the first exon	Coding	No		no
5RACE372_chr7_116699774_116699815	CFTR	AC000061.1	UGLsupplH	DQ656054	AC000061.1-003	2ab	14194	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE005_chr9_128951111_128951115	CRAT	RP11-247A12.5	UGL1c02	DQ655951	RP11-247A12.5-009	4	0	intronic	Not coding	No		no
5RACE302_chr18_59778397_59778482	SERPINB8	AC009802.3	UGL302-E-B6	DQ656055	AC009802.2-002	1	9742	chimeric (already annotated from the locus upstream)	Not coding	No		no
5RACE012_chr9_128783287_128783347	NUP188	RP11-167N5.2	UGL012-A-B2	DQ656056	RP11-167N5.2-009	1	6245	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE334_chr15_41747163_41747188	CATSPER2	AC011330.3	UGL17a09	DQ655952	AC011330.3-012	4	0	intronic	Not coding	No		yes, not coding
5RACE012_chr9_128789806_128790202	NUP188	RP11-167N5.2	UGL012-A-C2	DQ656057	RP11-167N5.2-010	4	0	intronic	Not coding	No	New exon inside the CDS part	yes, not coding
5RACE297_chr18_59465796_59465822	SERPINB11	AC069356.3	UGLsupplF	DQ656058	AC069356.3-003	1	62612	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE207_chr19_59833688_59833727	LILRB1	AC009892.6	UGL10e08	DQ655953	AC009892.6-001	4	0	intronic	Coding	No	New M (ATG) upstream	no (extends an existing exon)

5RACE299_chr18_59689889_59689930	SERPINB2	AC072051.2	UGL15a11	DQ655954	AC072051.2-005	1	16025	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE370_chr7_116045050_116045099	CAPZA2	AC002543.1	UGL19a07	DQ655955	AC002543.1-009	1	51467	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)(entirely open)	yes, not coding
5RACE370_chr7_116045050_116045099	CAPZA2	AC002543.1	UGL19f07	DQ655956	AC002543.1-010	1	51467	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE095_chrX_153234089_153234161	XX-FW81657B9.1	XX-FW81657B9.1	UGL4b09	DQ655957	XX-FW81657B9.1-009	2ab	73	extension of the first exon	Not coding	No		no
5RACE299_chr18_59708066_59708087	SERPINB2	AC072051.2	UGL299-E-A6	DQ656059	AC072051.2-005	4	0	intronic	Coding	No		yes, not coding
5RACE173_chr21_33542890_33542928	IL10RB	AP000295.8	UGL9b05	DQ655960	AP000295.8-006	2a	17643	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE173_chr21_33542890_33542928	IL10RB	AP000295.8	UGL9d05	DQ655961	AP000295.8-007	2a	17643	chimeric	Not coding	No	Skipping of the exon containing the M (ATG)	no
5RACE287_chr7_89837715_89837806	PFTK1	AC084381.2	UGL15a02	DQ655962	AC084381.2-008	2b	32747	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE287_chr7_89837715_89837806	PFTK1	AC084381.2	UGL15b02	DQ655963	AC084381.2-007	2b	32747	chimeric	Not coding	No	Skipping of the exon containing the M (ATG)	no
5RACE287_chr7_89837715_89837806	PFTK1	AC084381.2	UGL15d02	DQ655964	AC084381.2-009	2b	32747	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE287_chr7_89837715_89837806	PFTK1	AC084381.2	UGL15e02	DQ655965	AC084381.2-010	2b	32747	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE287_chr7_89837715_89837806	PFTK1	AC084381.2	UGL15f02	DQ655966	AC084381.2-011	2b	32747	New exons upstream (not chimeric)	Not coding	No	New exon inside the CDS part,Skipping of the exon containing the M (ATG)	yes, not coding

5RACE422_chr2_234466320_234466396	AC006985.5	AC006985.5	UGL20a03	DQ655967	AC006985.5-007	1	2580	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG) (entirely open)	yes, not coding
5RACE422_chr2_234466320_234466396	AC006985.5	AC006985.5	UGL20c03	DQ655968	AC006985.5-006	1	2580	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE422_chr2_234466320_234466396	AC006985.5	AC006985.5	UGL20g03	DQ655969	AC006985.5-008	1	2580	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE287_chr7_89740378_89740436	PFTK1	AC084381.2	UGL287-E-D4	DQ656060	AC084381.2-007	2a	130084	chimeric	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding (intervening exon between the 2 loci)
5RACE174_chr21_33618563_33618802	IFNAR1	AP000298.3	UGL9a08	DQ655972	AP000298.3-002	2ab	516	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE155_chr5_131907481_131907503	IL5	AC116366.2	UGL8e03	DQ655974	AC116366.2-003	2ab	390	extension of the first exon	Not coding	No		no
5RACE283_chr7_89489182_89489284	STEAP2	AC002064.1	UGLsupplE	DQ656061	AC002064.1-008	4	0	intronic	Not coding	No		no (extends an existing exon)
5RACE283_chr7_89441551_89441572	STEAP2	AC002064.1	UGL283-E-A4	DQ656062	AC002064.1-009	1	44100	New exons upstream (not chimeric)	Coding	Yes (new ATG upstream)	New M (ATG) upstream	yes, partly coding (ATG)
5RACE126_chrX_153543534_153543572	GAB3	CTD-2173L12.1	UGL126-B-H6	DQ656063	CTD-2173L12.1-003	2ab	52	extension of the first exon	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE294_chr2_118288445_118288510	DDX18	AC009404.1	Affy08248B12	DQ655975	AC009404.1-001	1	66	extension of the first exon	Coding	No		no
5RACE294_chr2_118288865_118288887	DDX18	AC009404.1	Affy08254D01	DQ655976	AC009404.1-006	4	0	intronic	Not coding	No	New exon inside the CDS part	yes, not coding
5RACE175_chr21_33679163_33679252	IFNGR2	AP000300.6	Affy08254E03	DQ655977	AP000300.6-006	3	17909	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE245_chr16_42113_42151	POLR3K	Z69719.1	Affy08248B08	DQ655978	Z69719.1-002	4	0	intronic	Not coding	No		no (extends an existing exon)
5RACE152_chr5_131375690_131375772	ACSL6	AC034228.1	Affy08246A12	DQ655980	AC034228.1-016	2b	94	New exons upstream (not chimeric)	Coding	No		yes, not coding

5RACE152_chr5_131375690_131375772	ACSL6	AC034228.1	Affy08246F12	DQ655981	AC034228.1-017	2b	94	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE409_chr11_5213134_5213197	HBD	AC104389.18	Affy08244A08	DQ655982	AC104389.18-004	3	743	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE408_chr11_5207176_5207199	HBB	AC104389.17	Affy08250G03	DQ655985	AC104389.17-004	3	2196	New exons upstream (not chimeric)	Coding	No		yes, not coding
5RACE259_chr7_26965357_26965495	HOXA6	AC004080.2	Affy08242H09	DQ655986	AC004080.2-004	2ab	1659	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE006_chr9_128922932_128922935	DOLPP1	RP11-167N5.4	Affy08254D10	DQ655991	RP11-167N5.4-001	2b	4	extension of the first exon	Coding	No		no
5RACE286_chr7_89657675_89657716	CLDN12	AC006153.2	Affy08254F11	DQ655993	AC006153.2-014	1	19724	chimeric	Not coding	No		no
5RACE286_chr7_89657675_89657716	CLDN12	AC006153.2	Affy08254G11	DQ655994	AC006153.2-015	1	19724	chimeric	Not coding	No		no
5RACE286_chr7_89677341_89677398	CLDN12	AC006153.2	Affy08242E11	DQ655995	AC006153.2-001	2ab	58	extension of the first exon	Coding	No		no
5RACE300_chr18_59715383_59715425	SERPINB10	AC009802.1	Affy08246B09	DQ655996	AC009802.1-002	1	5338	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE300_chr18_59715383_59715425	SERPINB10	AC009802.1	Affy08246G09	DQ655997	AC009802.1-003	1	5338	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE300_chr18_59720027_59720048	SERPINB10	AC009802.1	Affy08246E10	DQ655998	AC009802.1-003	3	694	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE300_chr18_59720027_59720048	SERPINB10	AC009802.1	Affy08246F10	DQ655999	AC009802.1-002	3	694	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE239_chr16_258721_258773	ARHGDIG	LA16c-314G4.1	Affy08250B05	DQ656000	LA16c-314G4.1-005	2a	11730	chimeric	Coding	Yes (links CDS of the two adjacent loci)	continuous ORF (entirely open)	no
5RACE239_chr16_258721_258773	ARHGDIG	LA16c-314G4.1	Affy08250F05	DQ656001	LA16c-314G4.1-006	2a	11730	chimeric	Coding	Yes (links CDS of the two adjacent	continuous ORF (entirely open)	no

										loci)		
5RACE203_chr19_59700906_59701072	LAIR2	AC008746.8	Affy08256A02	DQ656002	AC008746.8-003	1	4919	New exons upstream (not chimeric)	Coding	Yes (new ATG upstream)	Skipping of the exon containing the M (ATG)	yes, partly coding (ATG)
5RACE203_chr19_59700906_59701072	LAIR2	AC008746.8	Affy08256C02	DQ656003	AC008746.8-004	1	4919	New exons upstream (not chimeric)	Coding	Yes (new ATG upstream)	Skipping of the exon containing the M (ATG)	yes, partly coding (ATG)
5RACE323_chr5_56303234_56303255	AC008937.5	AC008937.5	Affy08256B03	DQ656004	AC008937.5-011	1	19480	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE323_chr5_56303234_56303255	AC008937.5	AC008937.5	Affy08256D03	DQ656005	AC008937.5-009	1	19480	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE323_chr5_56303234_56303255	AC008937.5	AC008937.5	Affy08256F03	DQ656006	AC008937.5-010	1	19480	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG) (entirely open)	yes, not coding
5RACE404_chr11_5102288_5102328	AC113331.10	AC113331.10	Affy08256A04	DQ656008	AC113331.10-002	1	2944	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE404_chr11_5102288_5102328	AC113331.10	AC113331.10	Affy08256B04	DQ656009	AC113331.10-003	1	2944	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	Affy08256G06	DQ656013	AC006159.1-007	1	211174	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5A05	EF070113	AC006159.1-008	1	211174	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5A06	EF070114	AC006159.1-009	1	211174	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5A08	EF070115	AC006159.1-010	1	211174	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5B05	EF070116	AC006159.1-011	1	211174	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5C05	EF070117	AC006159.1-012	1	211174	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding

5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5C07	EF070118	AC006159.1-013	1	211174	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5D05	EF070119	AC006159.1-014	1	211174	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5D07	EF070120	AC006159.1-015	1	211174	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5E05	EF070121	AC006159.1-016	1	211174	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE367_chr7_115522073_115522115	CAV2	AC006159.1	5E06	EF070122	AC006159.1-017	1	211174	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE367_chr7_115523874_115523910	CAV2	AC006159.1	Affy08256B07	DQ656015	AC006159.1-007	3	209373	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE367_chr7_115523874_115523910	CAV2	AC006159.1	Affy08256F07	DQ656017	AC006159.1-006	3	209373	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE367_chr7_115523874_115523910	CAV2	AC006159.1	Affy3F10-1	DQ656068	AC006159.1-004	3	209373	New exons upstream (not chimeric)	Not coding	No	New exon inside the CDS part	yes, not coding
5RACE367_chr7_115523874_115523910	CAV2	AC006159.1	Affy3F10-7	DQ656069	AC006159.1-005	3	209373	New exons upstream (not chimeric)	Not coding	No		yes, not coding
5RACE411_chr11_5623571_5623597	HBG2	AC104389.21	Affy2H6-2	DQ656066	AC104389.21-001	3	140186	chimeric	Coding	No		yes, not coding
5RACE007_chr6_41645877_41645938	FOXP4	RP11-328M4.1	Affy08256A09	DQ656018	RP11-328M4.1-004	4	0	intronic	Not coding	No	New exon inside the CDS part	yes, not coding
5RACE374_chr7_116662093_116662131	ASZ1	AC002465.3	Affy08256D10	DQ656019	AC002465.3-005	1	619	New exons upstream (not chimeric)	Coding	Yes (ATG downstream: shorter CDS)	Skipping of the exon containing the M (ATG)	no
5RACE374_chr7_116661513_116661518	ASZ1	AC002465.3	Affy08256A11	DQ656020	AC002465.3-001	3	6	extension of the first exon	Coding	No		no
5RACE190_chr21_34209825_34209901	ATP5O	AP000313.5	Affy08250G07	DQ656023	AP000313.5-007	4	0	intronic	Not coding	No	New exon inside the CDS part	no (extends exon into intron)
5RACE014_chr6_41829745_41829856	PGC	RP11-298J23.1	Affy08256B12	DQ656024	RP11-298J23.1-004	2ab	6743	New exons upstream (not chimeric)	Coding	Yes (new ATG upstream)	Skipping of the exon containing the M (ATG)	yes, partly coding (ATG)

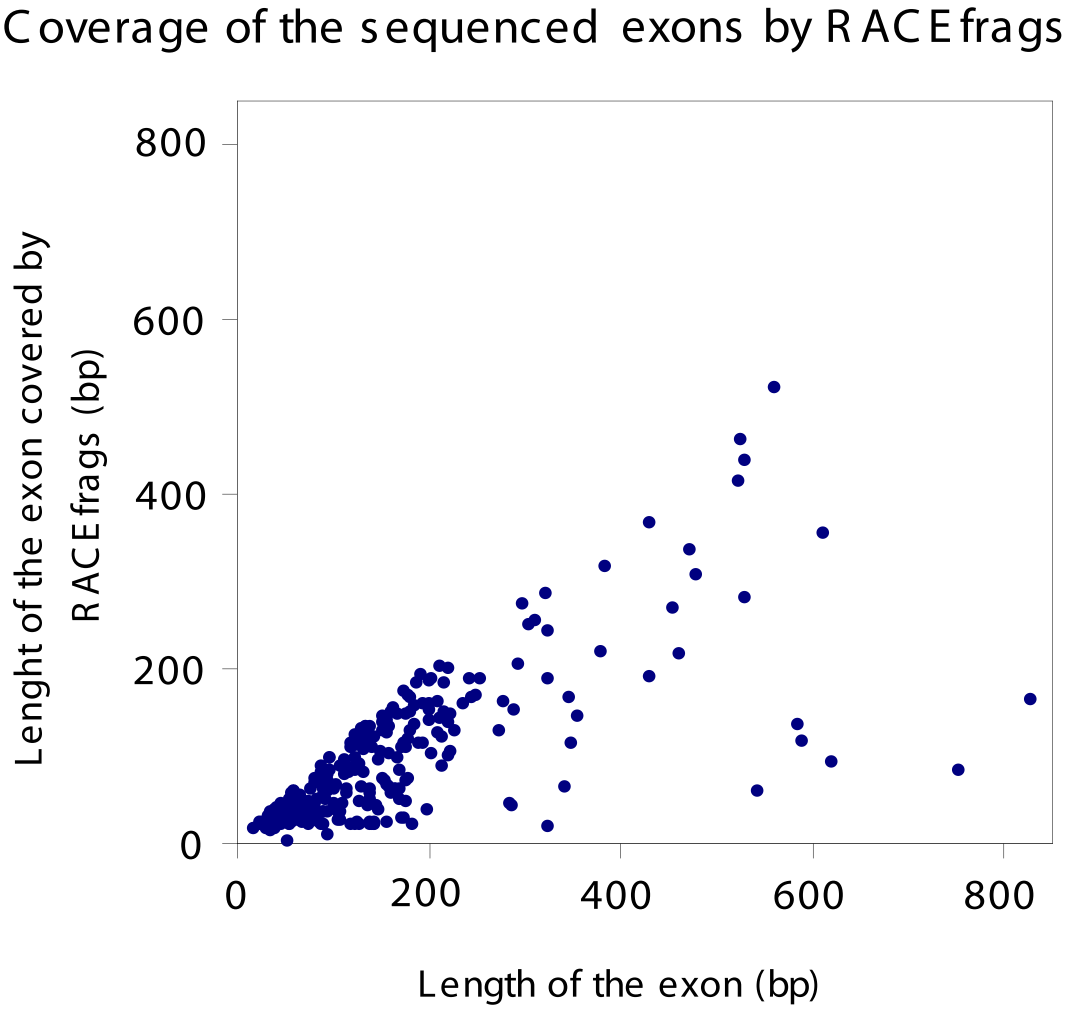
5RACE188_chr21_34206481_34206506	DONSON	AP000304.9	Affy08248B10	DQ656027	AP000304.11-012	1	316865	chimeric	Coding	Yes (links CDS of the two adjacent loci)	Skipping of the M (entirely open),Chimeric transcript with continuous ORF (entirely open)	no
5RACE285_chr7_89614334_89614360	AC006153.3	AC006153.3	Affy08258B04	DQ656028	AC006153.3-007	3	6266	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE285_chr7_89614334_89614360	AC006153.3	AC006153.3	Affy08258C04	DQ656029	AC006153.3-008	3	6266	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE285_chr7_89614334_89614360	AC006153.3	AC006153.3	Affy08258D04	DQ656030	AC006153.3-010	3	6266	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE285_chr7_89614334_89614360	AC006153.3	AC006153.3	Affy08258H04	DQ656031	AC006153.3-009	3	6266	New exons upstream (not chimeric)	Not coding	No	Skipping of the exon containing the M (ATG)	yes, not coding
5RACE393_chr11_5685993_5686053	OR56B1	AC131574.4	Affy08250A12	DQ656032	AC131574.4-003	2a	28330	chimeric	Not coding	No		yes, not coding
5RACE393_chr11_5685993_5686053	OR56B1	AC131574.4	Affy08250B12	DQ656033	AC131574.4-002	2a	28330	chimeric	Not coding	No		yes, not coding
5RACE011_chr6_41411364_41411474	NCR2	RP1-149M18.2	Affy08252A03	DQ656036	RP1-149M18.2-001	1	141	extension of the first exon	Coding	No	New M (ATG) upstream	no
5RACE133_chr6_74228516_74228563	MTO1	RP11-505P4.1	Affy08242A03	DQ656037	RP11-505P4.1-011	4	0	intronic	Coding	No		no (extends an existing exon)
5RACE202_chr19_59652582_59652671	LENG8	AC008746.4	Affy08258B01	DQ656038	AC008746.4-003	4	0	extension of the first exon	Coding	No	New M (ATG) upstream (entirely open)	no

Table S3 : Probe intensities from RNA hybridization in 4 sets of RACEfrags

tissue	Type of RACEfrag	Average intensity of all probes	Average intensity of positive probes	Median intensity of all probes	Median intensity of positive probes	% of positive probes
Brain	exonic	9.88	16.08	2	6.3	58.9%
Brain	Novel intronic	2.93	7.49	1	3	29.7%
Brain	Novel external	7.83	16.83	1	6.5	43.4%
Brain	chimeric	18.05	21.90	8	11	81.6%
Kidney	exonic	19.38	30.10	3	8	63.2%
Kidney	Novel intronic	2.60	5.89	1	3	32.7%
Kidney	Novel external	9.37	20.96	1	5	41.9%
Kidney	chimeric	34.40	41.48	11	16.65	82.5%
Small intestine	exonic	12.07	19.47	2.3	6	59.9%
Small intestine	Novel intronic	2.48	5.79	1	3.3	30.9%
Small intestine	Novel external	12.21	28.36	1	9.3	41.0%
Small intestine	chimeric	29.90	35.58	9.4	12.73	83.6%
Colon	exonic	16.14	25.17	3	7.3	62.6%
Colon	Novel intronic	3.03	8.25	1	3	27.9%
Colon	Novel external	11.28	25.84	1	4	41.4%
Colon	chimeric	98.85	110.37	24	28.85	89.5%
Liver	exonic	20.90	32.33	3	7	63.5%
Liver	Novel intronic	4.06	9.32	1	3	36.8%
Liver	Novel external	23.29	37.15	3	8.075	61.7%
Liver	chimeric	70.84	79.72	27.4	37	88.7%
Stomach	exonic	11.82	20.24	2	5.3	56.2%
Stomach	Novel intronic	2.47	5.74	1	3	31.0%
Stomach	Novel external	15.32	31.73	1	9	46.6%
Stomach	chimeric	58.17	78.91	5.5	10.85	73.4%

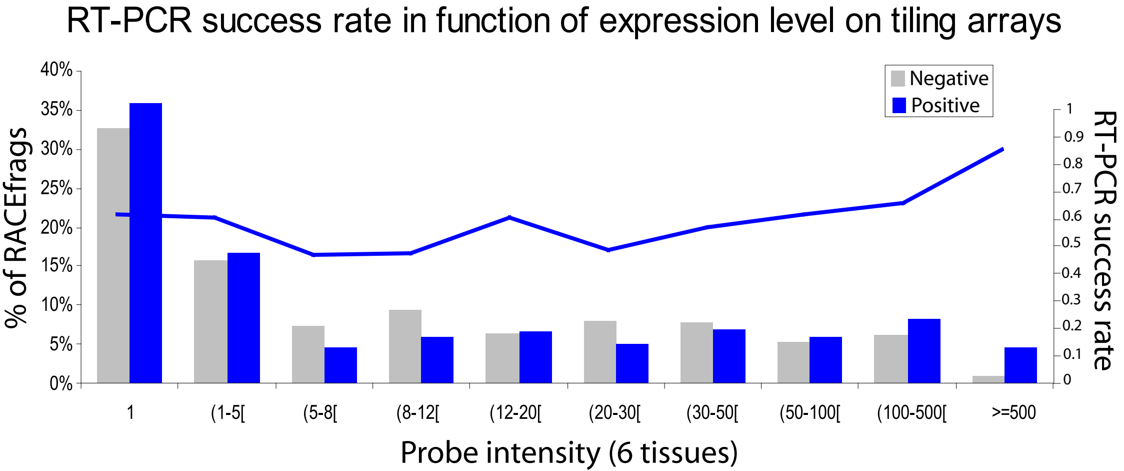
Supplementary Figures

Supplementary Figure S1: *Portion of novel exons covered by RACEfrags*
Comparison of the lengths of the sequenced novel exons (horizontal axis) and the length of the portion of these exons covered by RACEfrags (vertical axis). Each point represents one sequenced exon.


Coverage of the sequenced exons by RACEfrags
Lenght of the exon covered by RACEfrags (bp)
800
600
400
200
0
0
200
400
600
800
Length of the exon (bp)

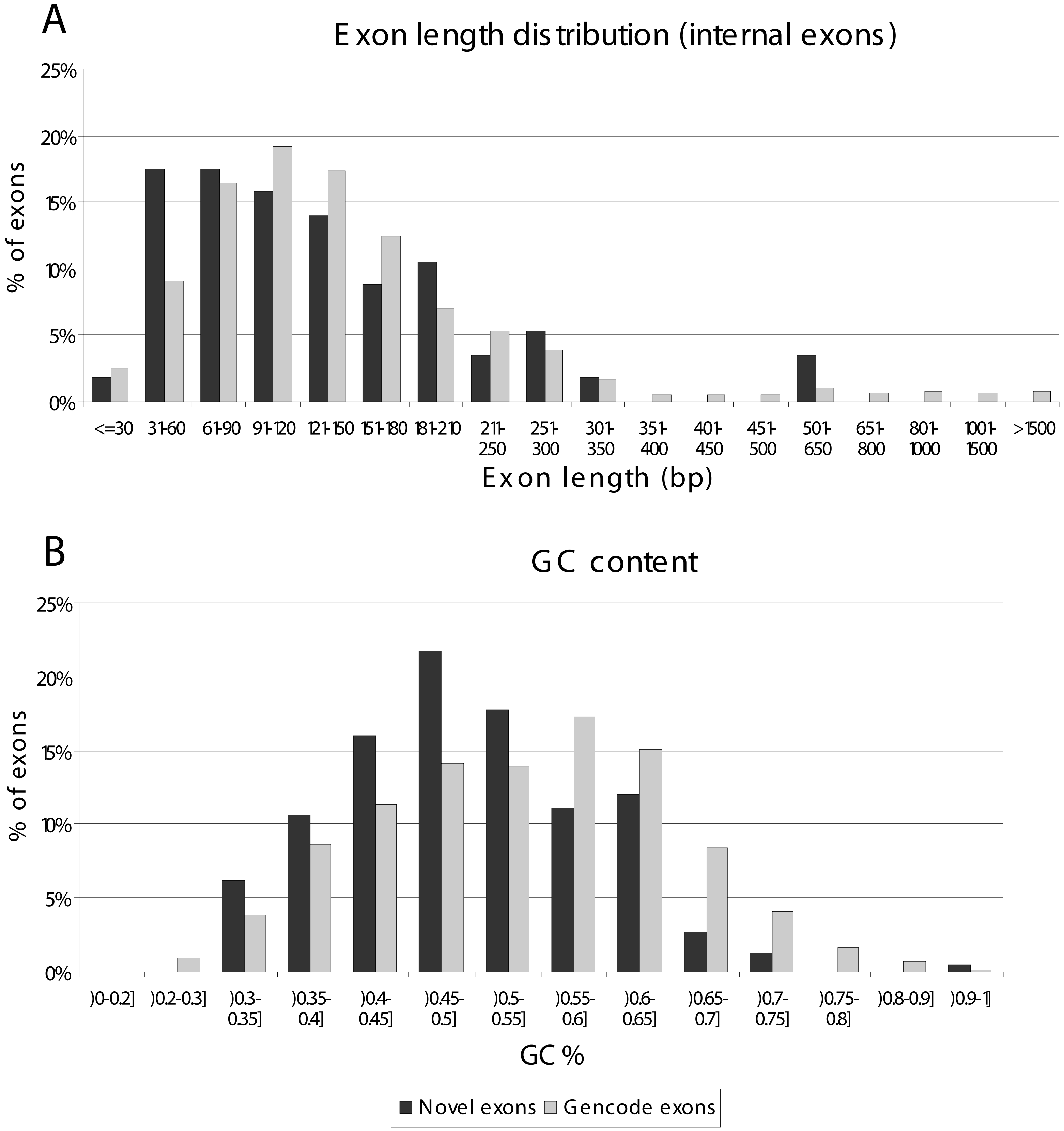
Supplementary Figure S2:

Distribution of RACEfrags tested by RT-PCR (positive reactions in blue, negative reactions in grey) according to the intensity signals measured on probes overlapping the regions where they map, in six tissues. Intensity values are represented on the X-axis. Values of 1 mean no signal (ratio of 1 compared to control) : positive probes are probes with intensity > 1. The-axis shows the % of RACEfrags in each intensity bin on the left, and the success rate of the RT-PCR reactions on the right.


RT-PCR success rate in function of expression level on tiling arrays
% of RACEfrags
40%
35%
30%
25%
20%
15%
10%
5%
0%
Negative
Positive
RT-PCR success rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
(1-5[
(5-8[
(8-12[
(12-20[
(20-30[
(30-50[
(50-100[
(100-500[
>=500
Probe intensity (6 tissues)

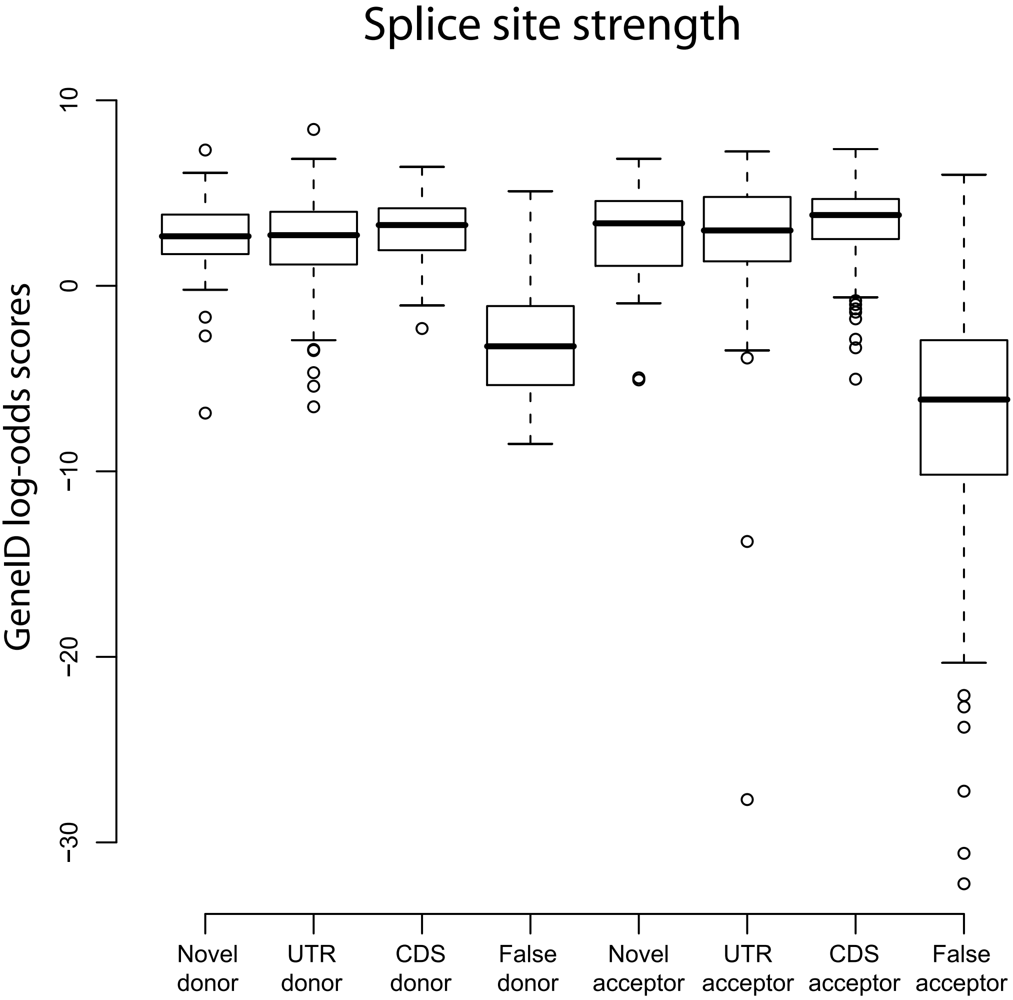
Supplementary Figure S3: *Characteristics of the novel exons*

Distributions of exon lengths (A) and GC contents (B) of novel exons identified by RACE/array (dark grey columns) and annotated by GENCODE (light grey columns) (Harrow et al. 2006).


A
Exon length distribution (internal exons)
25%
20%
15%
10%
5%
0%
% of exons
<=30
31-60
61-90
91-120
121-150
151-180
181-210
211-250
251-300
301-350
351-400
401-450
451-500
501-650
651-800
801-1000
1001-1500
>1500
Exon length (bp)
B
GC content
25%
20%
15%
10%
5%
0%
% of exons
)0-0.2]
)0.2-0.3]
)0.3-0.35]
)0.35-0.4]
)0.4-0.45]
)0.45-0.5]
)0.5-0.55]
)0.55-0.6]
)0.6-0.65]
)0.65-0.7]
)0.7-0.75]
)0.75-0.8]
)0.8-0.9]
)0.9-1]
GC %
■ Novel exons ■ Gencode exons

Supplementary Figure S4: *Splice site strength of novel exons*

Boxplots representing the distribution of log-odds scores for donor and acceptor sites as reported by GeneID are shown for each dataset. False splice sites were picked at random from the set of all GT or AG dinucleotides in ENCODE regions which do not overlap GENCODE-annotated exons or repeats. The heavy black line marks the median score, the box contains the 2nd and 3rd quartiles and whiskers mark the 5th and 95th percentiles.


Splice site strength
GeneID log-odds scores
10
0
−10
−20
−30
Novel donor
UTR donor
CDS donor
False donor
Novel acceptor
UTR acceptor
CDS acceptor
False acceptor