Published in final edited form as:

Nat Methods. 2008 July ; 5(7): 629–635. doi:10.1038/nmeth.1216.

Efficient targeted transcript discovery via array-based normalization of RACE libraries

Sarah Djebali[1,9], **Philipp Kapranov**[2,9], **Sylvain Foissac**[3,9], **Julien Lagarde**[1,9], **Alexandre Reymond**[4,9], **Catherine Ucla**[5], **Carine Wyss**[5], **Jorg Drenkow**[2], **Erica Dumais**[2], **Ryan R. Murray**[6], **Chenwei Lin**[6], **David Szeto**[6], **France Denoeud**[1], **Miquel Calvo**[7], **Adam Frankish**[8], **Jennifer Harrow**[8], **Periklis Makrythanasis**[5], **Marc Vidal**[6], **Kourosh Salehi-Ashtiani**[6], **Stylianos E. Antonarakis**[5], **Thomas R. Gingeras**[2], and **Roderic Guigó**[1,3]

[1]Grup de Recerca en Informática Biomédica, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra, Dr. Aiguader 88, 08003 Barcelona, Catalonia, Spain [2]Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California 95051, USA [3]Center for Genomic Regulation, Dr. Aiguader 88, 08003 Barcelona, Catalonia, Spain [4]Center for Integrative Genomics, University of Lausanne, Genopole Building, 1015 Lausanne, Switzerland [5]Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel Servet, 1211 Geneva, Switzerland [6]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, 44 Binney Street, Boston, MA 02115-6084, USA [7]Departament d'Estadística, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Catalonia, Spain [8]HAVANA Group, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1HH, United Kingdom.

Abstract

RACE (Rapid Amplification of cDNA Ends) is a widely used approach for transcript identification. Random clone selection from the RACE mixture, however, is an ineffective sampling strategy if the dynamic range of transcript abundances is large. Here, we describe a strategy that uses array hybridization to improve sampling efficiency of human transcripts. The products of the RACE reaction are hybridized onto tiling arrays, and the exons detected are used to delineate a series of RT-PCR reactions, through which the original RACE mixture is segregated into simpler RT-PCR reactions. These are independently cloned, and randomly selected clones are sequenced. This approach is superior to direct cloning and sequencing of RACE products: it specifically targets novel transcripts, and often results in overall normalization of transcript abundances. We show theoretically and experimentally that this strategy leads indeed to efficient sampling of novel transcripts, and we investigate multiplexing it by pooling RACE reactions from multiple interrogated loci prior to hybridization.

Correspondence to: Roderic Guigó.

[9]These authors contributed equally to this work.

INTRODUCTION

Determining the RNA complement of a cell is a pre-requisite for fully understanding its biology, and for translating this understanding to technical applications in medicine, agriculture, and biotechnology. Systematic sequencing of cDNA libraries has been the main approach for transcript characterization. One popular strategy is based on the identification by a "single pass" sequencing of random cDNA clones that results in short partial transcript sequences, known as Expressed Sequence Tag1 (EST). ESTs can then be used to identify clones suitable for full-length sequencing, as in the MGC initiative2. More recently, methods for full-length isolation and sequencing of random clones from cDNA libraries have also been developed3,4. The wide dynamic range of mRNA abundances in cells, however, makes random clone selection inefficient for discovering relatively rare transcripts, because predominantly the abundant cDNAs will be sequenced. To overcome this limitation, procedures which increase the likelihood of sampling rare or tissue specific transcripts such as normalization and subtraction have been developed5,6. But, while these procedures may be effective in normalizing the abundance of representative transcripts from different genes, they are less effective in normalizing the abundances of alternative transcripts within a given gene, which usually share a substantial fraction of sequence. Methods have been developed to enrich cDNA libraries for alternative splice transcripts7,8, but they preferentially target interal exons, whereas a large part of transcript variability resides at the 5’ and 3’ ends of the genes9.

Recent data, obtained through a variety of approaches, strongly suggest the existence of a wealth of transcripts, which had so far escaped detection through systematic sequencing of cDNA libraries10–15. In particular, experiments in which the products of RACE reactions originating from primers anchored in annotated genes are hybridized onto genome tiling arrays have uncovered many previously undetected exons16. Here, building on this approach (see also Kapranov et al.17), we develop an array-based normalization strategy of RACE (Rapid Amplification of cDNA Ends18) reactions, which is very efficient for targeted discovery of novel transcripts. RACE amplifies all transcript sequences in a given RNA sample that include an index exon within which the RACE primer has been designed. Following amplification, RACE products are typically cloned, and random clones are subsequently sequenced. However, the dynamic range in the RACE reaction may still be very large, and random clone sequencing may predominantly yield high-copy number variants—the most likely to have been already identified. To overcome this limitation, we introduce an intermediate step in which the RACE mixture is hybridized onto high-density tiling arrays. This serves to highlight putative novel exons that are used to delineate a set of conventional RT-PCR reactions, from which clones to be sequenced are randomly selected. Here, we show both theoretically and experimentally that this strategy leads to the specific amplification of novel transcripts and to the homogenization of their relative abundances, and that it is indeed very efficient in sampling novel transcript species. We also show that, under the appropriate conditions, it can be efficiently multiplexed by interrogating multiple loci simultaneously.

RESULTS

The “RACEarray” normalization strategy

Figure 1 schematizes the RACEarray strategy (Supplementary Figs 1, 2, Supplementary Methods). Given a locus, we first select the exons in which the RACE primers will be designed (Supplementary Fig 3), and carry out the RACE reactions. Second, we hybridize the RACE products onto tiling arrays, and we build the sites of transcription, the so-called RACEfrags (RACE positive fragments)16, from the probe hybridization intensities. A number of filters can be applied to RACEfrags to account for highly expressed genes in the original RNA source, or to identify RACEfrags produced by the amplification of non-targeted loci (Supplementary Fig 4). If RACE reactions from multiple primers have been pooled together before hybridization, complex assignment procedures may need to be employed to assign RACEfrags to interrogated primers. Third, we use the resulting RACEfrags (Supplementary Fig 5) to delineate RT-PCR reactions. One of the primers for each of the reactions is the original RACE primer, and the second primer is designed within each novel RACEfrag. Strategies based on the pattern of co-occurrence of RACEfrags across different assayed conditions can be designed to select the subset of RACEfrags maximizing transcript discovery (Supplementary Fig. 6). Fourth, we clone the products of the resulting RT-PCR into “mini-pools”: each mini-pool contains the amplified transcripts connecting an index exon with a novel RACEfrag. Finally, we randomly select clones from these pools for sequencing.

Through this process, the original RACE population (likely to be dominated by the most abundant transcripts) is segregated into a number of RT-PCR populations, each one designed to include at least one (but probably more) novel transcripts. Sampling from the RT-PCR subpopulations increases the probability of selecting at least one clone of each novel transcript variant, in contrast to directly sampling from the original RACE population. By modeling random clone selection as a multinomial process, we show that the probability of obtaining at least one clone of each novel species after randomly sampling a number of clones increases as the probabilities of the known mRNA species decreases (Supplementary Methods). We can also show that, when sampling from the population of novel transcripts, the probability of obtaining at least one clone from each novel species increases as the probabilities of the novel transcripts approach homogeneity. We have evidence that the RACE array normalization will often, though not necessarily, lead to the overall homogenization of transcript abundances (since transcripts within an RT-PCR subpopulation are likely to share a larger number of exons, and may have in consequence more similar abundances), and we have carried out extensive simulations (see Supplementary Methods) which show that when the transcript abundances within the segregated subpopulations are homogeneous, sampling from them is in general more efficient that sampling from the original population.

RACEarray discovery of transcript isoforms

As a proof of concept, we used RACEarray normalization to interrogate a single gene: MECP2. Mutations in this gene cause the Rett syndrome. MECP2 has two known transcript variants: the longer form has four exons, the shorter form skips the 2nd exon (Figure 2a). We

performed 3’ and 5’ RACE from the exon number 3 in 16 different tissues (see Methods). We additionally performed 5’ RACE from exons 2, 3 and 4 in fetal brain. RACE reactions were separately hybridized on the ENCODE arrays containing the region ENm006, in which the MECP2 gene resides. The raw hybridization data appears on Figure 2a. Seventy one RACEfrags were detected. Eight 5’ RACEfrags were selected for RT-PCR verification. All eight gave at least one RT-PCR product, which was either cloned or sequenced directly. In total, 15 novel isoforms including 14 novel exons were discovered in this way. Most these isoforms are partial, since many of the novel RACEfrags interrogated are likely to correspond to internal exons. The majority of them use canonical splice sites and a few could be coding for proteins. Therefore, through a limited exploration using the RACEarray normalization strategy, we have discovered many novel isoforms for an important disease gene.

To further test the utility of our approach and to demonstrate that it may be effective in exploring protein coding genes, we tested 10 novel RACEfrags connecting to index exons in 9 diferent genes. The RACEfrags were randomly selected from a subpopulation of RACEfrags flanked by good splice sites and within 500Kb from the RACE index exon (see Methods). We performed the RT-PCR reactions using total RNA from two pools of tissues. Positive RT-PCRs were cloned into a mini-pool, and 32 clones were randomly selected in each case for sequencing (Supplementary Table 1). Thirty four novel variants were uncovered for these 9 loci, compared with 59 previously known. Some of these variants correspond to complex transcriptional events including long-range exon sharing (Fig. 2b). Virtually all cases were positive in the two pools, and nearly all sequences aligned to the genome with canonical splice sites. The novel transcripts discovered did not result in novel ORFs for the interrogate genes, being either non-coding variants, or variants extending the UTRs.

Multiplexing of RACEarray normalization

While the strategy designed here is particularly useful to exhaustively characterize the transcript complement of individual loci, a few steps can be efficiently multiplexed allowing for the simultaneous interrogation of multiple loci. This is mostly accomplished by pooling together RACE reactions from different loci prior to the hybridization onto the array. However given the long extensions previously observed in the ENCODE regions16, pooling together RACE reactions originating from index genes in close proximity in the genome sequence might confound assignment of RACEfrags to RACE index exons. It is therefore helpful to estimate the range of genomic distances used by primary transcripts of protein coding genes to design an optimal pooling strategy. In addition, both the number and combination of tissues on which the original RACE reactions are performed, as well as the number and distribution of primers along the interrogated loci influence the ability to survey transcript diversity.

Optimal number and combination of tissues

We performed both 5' and 3' RACE of 12 genes mapping on human chromosomes 21 and 22 on polyA+ RNA of 48 cell types (see Table 1 and Methods). Both 5’ and 3’ RACE reactions for three widely spaced genes per chromosome were pooled and hybridized onto a high-

density tiling array of human chromosomes 21 and 22 with 17-nucleotide interrogation resolution. Detailed results are provided in Supplementary Results, but figure 3 summarizes the main findings. Figure 3a plots the genomic coverage of RACEfrags as a function of the tissue in which the RACE reaction was performed. Not surprisingly, tissues exhibit, in general, higher transcriptional diversity than cell lines, but large variations in the amount of transcribed bases are observed between both tissues and cell lines, consistent with previous results19. Figure 3b plots the cumulative genomic coverage as a function of the combination of tissues. As shown, a combination of about 16 cell types already captures about 90% of all detected transcribed nucleotides.

Optimal distribution of exons to RACE per locus

We carried out, 5' RACE on 10 exons, evenly distributed 3' to 5', of 44 genes, each gene mapping to a different ENCODE region20,21 (Table 1). We used polyA+ RNA from 12 human tissues, and the RACE reactions were pooled before being hybridized to the ENCODE arrays16,17. Each pool contained 44 RACE reactions, each one originating from one exon from a different gene, and thus from a different ENCODE region. Detailed results are presented in Supplementary Results. Figure 4 displays the proportion of all RACEfrags that originate from primers in exons from the 3' to 5'. The cumulative distribution, although inconclusive, suggest and optimal interrogation strategy, in which RACE of the most 5' and 3' exons is likely to give rise to a larger number of novel RACEfrags, compared with RACE of internal exons (see Supplementary Figure 3).

Pooling of RACE reactions – Genomic extent of loci

We conducted 5' RACE on 96 genes in human chromosomes 21 and 22 (Table 1). Reactions were carried out individually on polyA+ RNA from 12 different tissues and subsequently pooled. RACE reactions from different tissues originating from genes each separated by 10 Mb were pooled in groups of 6 on the same chip. Results (Supplementary Results and Figure 5) show that transcripts may span very large genomic space, with about 50% of the RACEfrags more than 3MB away from the index gene. These results need further validation, but they could potentially challenge our current understanding of the structure and organization of transcripts encoded in the human genome, suggesting that distal regions may be connected into individual transcripts more often than previously expected. They also make very challenging the delineation of an effective pooling strategy since only a very sparse pooling appears to guarantee a robust assignment of RACEfrags to primers.

DISCUSSION

We have shown here that array-based normalization of RACE reactions is a very efficient strategy in discovering previously unknown transcript variants of protein coding loci. Our experiment yielded about one novel transcript variant per 10 clones sequenced, and while the RACEfrags assayed here for the nine tested genes were selected from a subpopulation of high confident RACEfrags, a large fraction (60%) of the loci interrogated have assigned at least one such RACEfrag. We do not know of any other strategy, which is able to survey the transcriptional diversity of protein coding loci with this level of detail and efficiency. We have compared the results of the hybridization of RACE reactions to tiling arrays with those

from other high-throughput transcript interrogation surveys using distinct technologies: CAGE tags22, GIS PET ditags11 and ESTs (Supplementary Table 2). While there is significant overlap between our RACEfrags and the sites of transcription detected by these other technologies there is still a substantial number of RACEfrags (about 17% of those detected through all the experiments performed here), which are not detected by alternative methods.

The RACE array normalization strategy relies on fairly well-established molecular techniques and can be performed by any basic molecular biology laboratory. It is cost-effective, since the total cost of interrogating a single locus amounts to less than $1,000 (assuming $225-$400 for the array, $50 for RACE reaction and array hybridization/labeling, and the remaining 650 for RT-PCR and sequencing of RT-PCR products). The approach can be scaled to whole genome large-scale transcript discovery. Indeed, our results indicate that, through the interrogation of a relatively small number of properly spaced exons from annotated loci in a relatively small number of tissues and cellular types, it is possible to recover a substantial fraction of the transcript diversity associated to a given locus. Given the large space that protein coding loci seem to span, however, high pooling density of RACE reactions prior to hybridization can only be achieved when the hybridization experiments are performed multiple times in different conditions or cell types, using multiple primers from the same locus. Indeed, the pattern of co-occurrence of primers and RACEfrags across the different conditions provides additional information about their connectivity. Pooling of RACE reactions could reduce the array cost by about two orders of magnitude. “Next generation” sequencing platforms could, in turn, be used to sequence in parallel thousands of clones. If these have been pooled judiciously, short read sequences can be unambiguously assembled into their respective full-length contigs, decreasing the cost of sequencing also several orders of magnitude, and making genome wide RACEarray exploration feasible.

The issue obviously arises as to why so many transcripts of otherwise well-annotated protein coding loci had systematically been missed by the large unbiased cDNA and EST sequencing projects? Many reasons may have contributed. First, cDNA libraries are not exhaustive. If not normalized, random clone selection will only yield high-copy number variants. This is compounded by the fact that many existing cDNA libraries are obtained from tissues characterized by a complex and heterogeneous transcript complement. In this respect, targeted interrogation of a single locus is intrinsically more sensitive in discovering the full complexity of transcripts at that locus than shotgun sequencing of the entire library. Second, normalization based on hybridization may have the undesirable effect of decreasing the likelihood of sampling low or medium copy alternative splice forms or long chimeric transcripts16,23, since these can be selectively eliminated with their higher copy variants with which they may share a large fraction of the their sequence. Third, cDNA and EST libraries are often obtained through oligo-dT primed reverse transcriptase reactions. The single short read sequences originated in this way might not be long enough to reach the 5’ ends of long mRNA sequences, or the junction between exons from different loci, that we are predominantly discovering here.

The many novel transcript variants discovered here are not necessarily in low copy number. In fact, while novel RACEfrags show a restricted expression pattern when compared to annotated exons (6.8 vs. 13.0 tissues on average, respectively, see Supplementary Results, Supplementary Figure 7), a substantial fraction of them (58%) seem to be expressed in more than one tissue (compared with 62% of the annotated exons; see Supplementary Results). Unfortunately, because of the many amplification steps involved, and as a drawback of our approach, RACEarray normalization does not provide a good estimate of the expression levels of the identified transcripts. We have, however, attempted to reconstruct the expression level of RACEfrags by using transcriptional maps obtained within the ENCODE project. Our results (see Supplementary Results, Supplementary Figure 8) indicate that novel RACEfrags are in general in lower expression levels than exonic RACEfrags, but that for about a third of loci interrogated in this study there are no significant differences between the expression levels of exonic and of novel RACEfrags. In any case, whether low copy number can be taken as an indication of lack of functionality or not, we would like to stress that our method is a transcript surveying tool—as, for instance, EST and CAGE sequencing are—and, like these methods, does not attempt to provide evidence of functionality.

In summary, RACEarray normalization can be used to efficiently explore how the transcript complement of loci changes under different cellular conditions, or varies between different cell types or individuals. With the appropriate experimental design, the strategy can be effectively multiplexed by high-density pooling of RACE reactions, and therefore can be used for genome scale transcript discovery.

METHODS

RACE reactions

See Supplementary Methods for the tissues and cell lines used, as well as for the conditions under which the RACE reactions were carried out.

Hybridization of RACE products onto tiling arrays and delineation of RACEarray maps

The products of RACE reactions were pooled and purified by ethanol precipitation. The pooled amplicons were then fragmented, and subsequently labeled for direct array hybridization (see Supplementary Methods).

The maps of probe intensities versus the genomic positions were generated using Tiling array Software (TAS; http://www.affymetrix.com/support/developer/downloads/TilingArrayTools/index.affx, see Supplementary Methods). RACEfrags were built using a 99.7% percentile in the probe intensity value as a threshold. Two probes are included in the same RACEfrag if they are less than 25 nucleotides away (*maxgap*), and the minimum length of a probe (*minrun*) is 25 nucleotides. RACEfrags were filtered using RACEarray *in silico* simulator that aims at reducing the RACEfrag false-positive rate due to RACE mis-priming, as well as array cross-hybridization (see Supplementary Methods). Surviving RACEfrags were assigned to the closest interrogated primer in genomic space.

Selection of novel exons for RT-PCR

Verification of RACEfrag/index exon connectivity was carried out in a nested RT-PCR experiment on 10 cases. These cases corresponded the experiments described in the section "Pooling of RACE reactions-genomic extent of loci". We considered only RACEfrags with very good putative donor sites (with score over 2.4 as computed by the geneid program24) in the vicinity (within −10 to +30 nucleotides) of the 3' end of the RACEfrag, and within 500 Kb from the index RACE exon. The 189 RACEfrags surviving these criteria corresponded to 58 loci, and the 10 test cases were randomly selected from this population.

For the test cases, the left primers were designed on the RACEfrag sequence, whereas the right primers were selected within the corresponding index exon. The external right primers were chosen to be the same as in the RACEarray experiment. For the positive controls, nested RT-PCR primers were designed in the 60 5'-most and 60 3'-most nucleotides of the target full-length mRNA. In all cases, the primer3 program 25 was used to pick primers.

Additional Methods

RT-PCR, Cloning and Sequencing (Supplementary Methods)

Data availability

The list of genes, exons, and RACE primers used in these experiments is available at http://genome.imim.es/datasets/racearrays2007. Processed and unprocessed micro-array data, as well as RT-PCR primers and the sequences of the resulting RT-PCR products are also available at this address. Primary array data will be deposited in the National Center for Biotechnology Information's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). The sequence data of the RT-PCR products from this study have been submitted to DDBJ/GenBank/EMBL under accession numbers FE530175 - FE530855, with the exception of the MECP2 sequences, which are in the process of being submitted. RACEfrag data and maps are being deposited to the UCSC browser.

In Supplementary Methods we describe in detail the mathematical formalization of the RACEarray sampling problem, and the computer simulations, as well as the detailed protocols for the RACE, array hybridization, RT-PCR, cloning and sequencing experiments

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The project at IMIM, CRG, the Universities of Lausanne and Geneva, and Affymetrix is supported by grants U01HG003150 and U01HG003147 from the National Human Genome Research Institute, NIH. Institut Municipal d'Investigació Mèdica and Center for Genomic Regulation (CRG) have also been funded by grant BIO2006-03380 from the Spanish Ministry of Education and Science and from the European BioSapiens Consortium. The Universities of Lausanne and Geneve have also been funded by the Swiss National Science Foundation, the EU AnEUploidy project, and the NCCR Frontiers in Genetics. Affymetrix has also received funds from the National Cancer Institute, NIH, under contract no. N01-CO-12400 and by Affymetrix, Inc. The portion of this work carried out at Center for Cancer Systems Biology was funded by a grant from the Ellison Foundation (awarded to MV) and as Institute Sponsored Research from the Dana Farber Cancer Institute Strategic Initiative. We gratefully acknowledge Dr. J.M. Oller from the University of Barcelona for his review of the probabilistic results, and Robert

Castelo from the University Pompeu Fabra, Cédric Howald from the University of Lausanne, and David Martin from the CRG, for useful suggestions.

References

1. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nat Genet. 1993; 4:373–380. [PubMed: 8401585]
2. Gerhard DS, et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res. 2004; 14:2121–2127. [PubMed: 15489334]
3. Kawai J, et al. Functional annotation of a full-length mouse cDNA collection. Nature. 2001; 409:685–690. [PubMed: 11217851]
4. Carninci P, et al. The transcriptional landscape of the mammalian genome. Science. 2005; 309:1559–1563. [PubMed: 16141072]
5. Bonaldo MF, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 1996; 6:791–806. [PubMed: 8889548]
6. Soares MB, et al. Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci U S A. 1994; 91:9228–9232. [PubMed: 7937745]
7. Thill G, et al. ASEtrap: a biological method for speeding up the exploration of spliceomes. Genome res. 2006; 16:776–786. [PubMed: 16682744]
8. Watahiki A, et al. Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. Nature methods. 2004; 1:233–239. [PubMed: 15782199]
9. Harrow J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biology. 2006; 7(Suppl 1):1–9. S4.
10. Shiraki T, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003; 100:15776–15781. [PubMed: 14663149]
11. Ng P, et al. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods. 2005; 2:105–111. [PubMed: 15782207]
12. Peters LM, et al. Signatures from tissue-specific MPSS libraries identify transcripts preferentially expressed in the mouse inner ear. Genomics. 2007; 89:197–206. [PubMed: 17049805]
13. Roma G, et al. A novel view of the transcriptome revealed from gene trapping in mouse embryonic stem cells. Genome Res. 2007; 17:1051–1060. [PubMed: 17540781]
14. Kapranov P, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007; 316:1484–1488. [PubMed: 17510325]
15. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
16. Denoeud F, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res. 2007; 17:746–759. [PubMed: 17567994]
17. Kapranov P, et al. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res. 2005; 15:987–997. [PubMed: 15998911]
18. Frohman MA, Dush MK, Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A. 1988; 85:8998–9002. [PubMed: 2461560]
19. Reymond A, et al. Human chromosome 21 gene expression atlas in the mouse. Nature. 2002; 420:582–586. [PubMed: 12466854]
20. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004; 306:636–640. [PubMed: 15499007]
21. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
22. Kodzius R, et al. CAGE: cap analysis of gene expression. Nature methods. 2006; 3:211–222. [PubMed: 16489339]

23. Parra G, et al. Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res. 2006; 16:37–44. [PubMed: 16344564]

24. Parra G, Blanco E, Guigo R. GeneID in Drosophila. Genome Res. 2000; 10:511–515. [PubMed: 10779490]

25. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000; 132:365–386. [PubMed: 10547847]

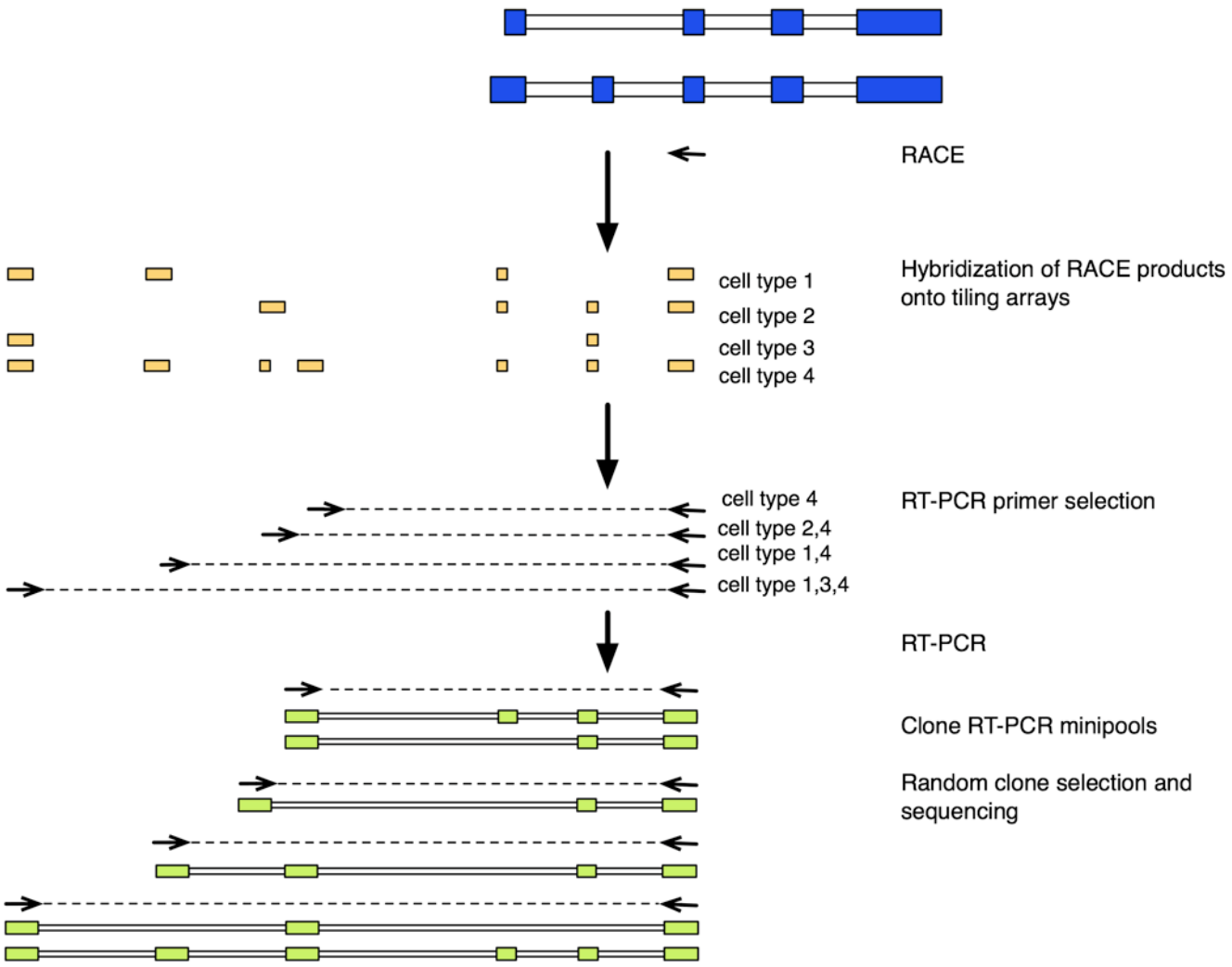


Figure 1. strategy for comprehensive characterization of novel isoforms from annotated genes

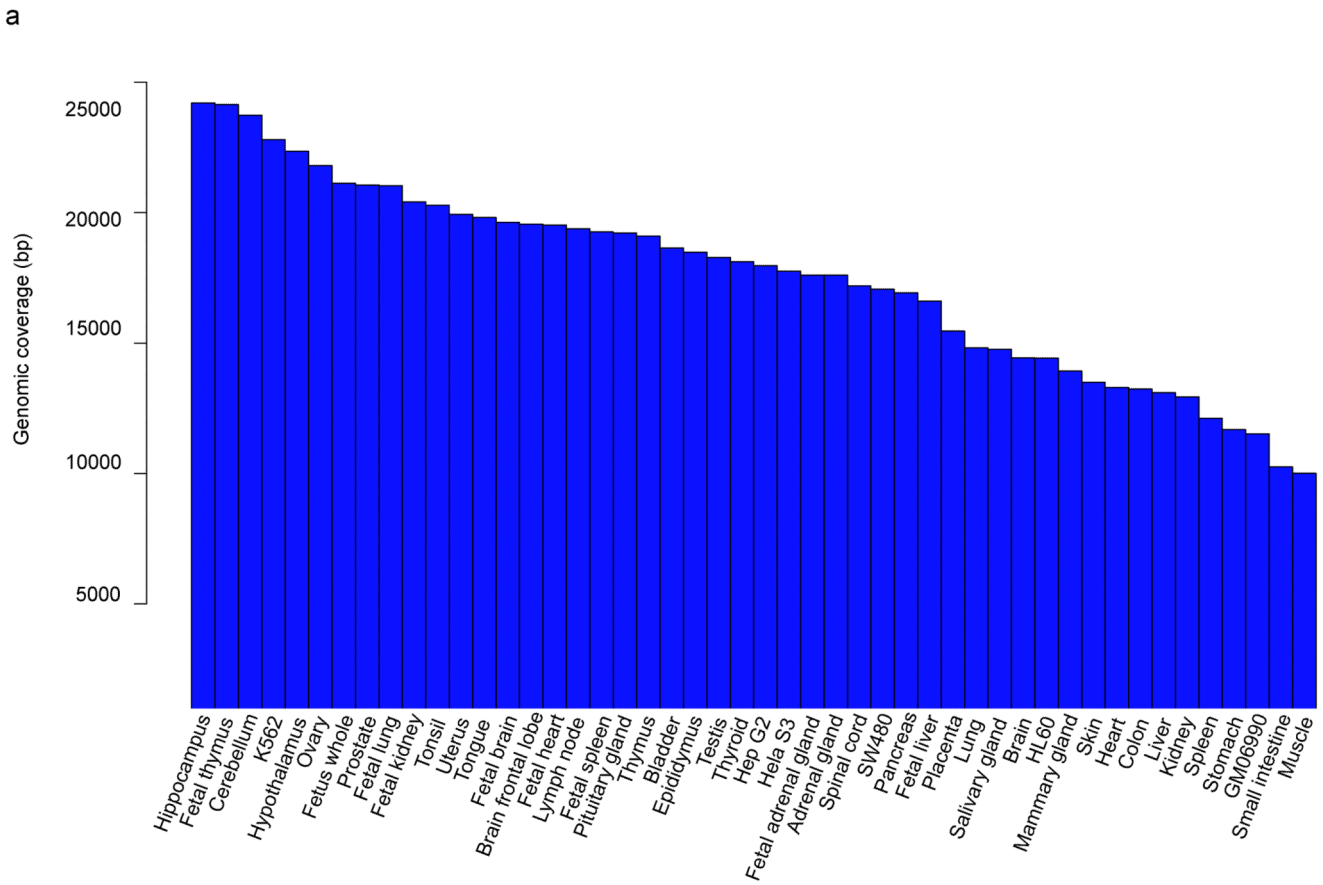
First, RACE (5', 3' or both) is performed with primers (black arrows) from one or more annotated exons of known loci. Second, the RACE products are hybridized onto a tiling array, possibly across different cell types and conditions. Third, the detected sites of transcription (RACEfrags, in yellow in the figure) are used to design RT-PCR primers (black arrows). Primers are designed only on RACEfrags corresponding to previously undetected exons. Fourth, one RT-PCR reaction is performed for each primer in a novel RACEfrag, using the original RACE primer as the second primer. Fith, each RT-PCR reaction is cloned separately into a mini-pool. Finally, clones are randomly selected from the RT-PCR mini-pools and sequenced, leading to the identification of novel transcripts.

Figure 2. Examples of novel RACEfrags verified by RT-PCR, cloning and sequencing

(**a**) RACEarray interrogation of the MECP2 locus. Probe intensity values of RACE products originating from annotated exons of the MECP2 locus hybridized into the ENCODE tiling-array including the region in which the MECP2 locus resides. Two isoforms are known for this gene. Fifteen new transcript sequences have been discovered through RACEarray normalization (cloned RT-PCR products).

(**b**) RACEarray interrogation of the CHAF1B locus (UCSC Genome Browser screenshot). RACEfrags are depicted in orange, RT-PCR primer pairs in cyan, sequenced RT-PCR products in purple, and index genes ("Q3 target genes" track) in blue. A 5' RACE reaction originating from an exon of gene CHAF1B (a.k.a. NM_005441) produced the RACEfrags showed in orange. The most distal RACEfrag—overlapping an exon from upstream gene ZCWCC3 (a.k.a. NM_015358, on the same strand as CHAF1B)—was chosen for RT-PCR

verification. RT-PCR products were cloned. Sixteen clones were selected at random and sequenced. Eight different sequences were obtained (under “RT-PCR products”; two end sequences from one clone could not be assembled and are not shown here). All the novel sequences connect the two loci using a variety of novel exon combinations. No previous evidence existed supporting these transcripts. For reference, various UCSC annotation tracks are also represented at the bottom of each screenshot. Some tracks (“Q3 RxFRAGS”, “Human mRNAs + ESTs”) were collapsed (“Dense” mode) for clarity purposes (a more detailed figure as well as other examples are available at http://genome.imim.es/datasets/racearrays2007/).



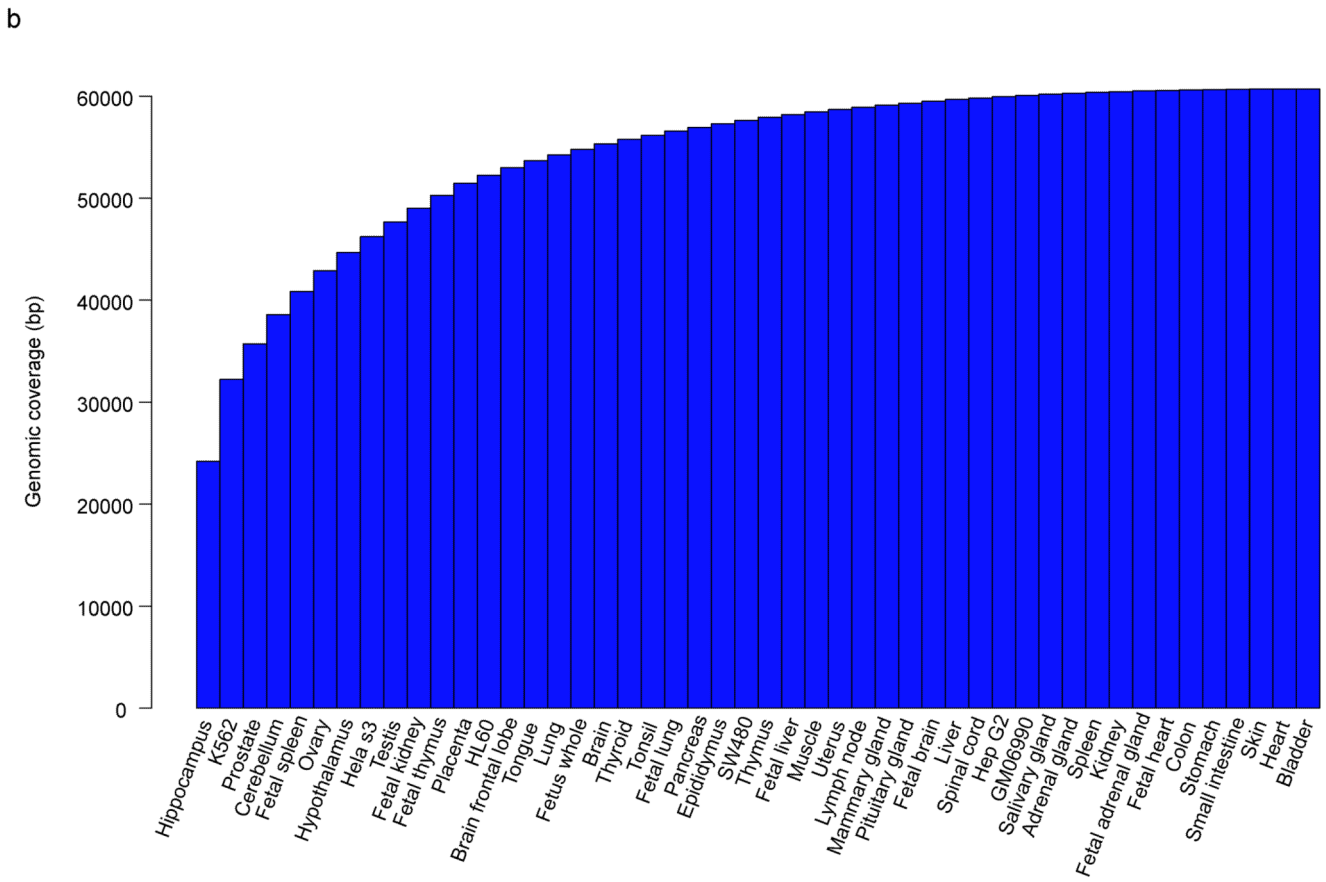
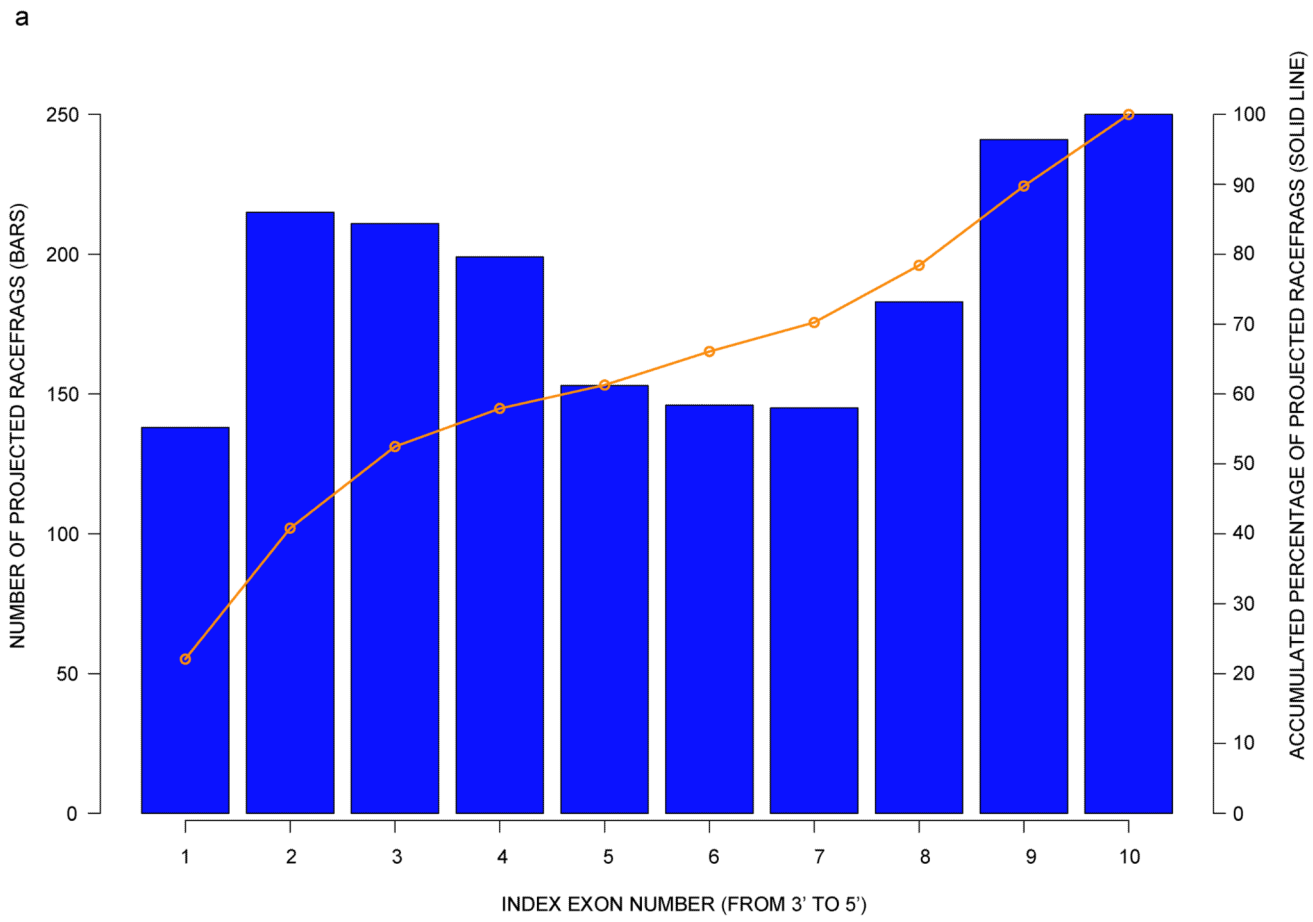


Figure 3. genomic coverage of RACEfrags originating from different tissues or combinations of tissues

(**a**)total number of nucleotides in RACEfrags as a function of the tissue in which the RACE reaction has been performed.

(**b**)cumulative number of transcribed nucleotides detected by RACEfrags per tissue. The cumulative coverage is obtained iteratively. At each step, a new tissue is included in the carrying combination of tissues. The tissue included at each step is the one for which RACEfrags include the maximum number of nucleotides in the genome, not previously included in the carrying combination of tissues. Correspondingly, tissues are ordered on the X-axis from left to right, so that the tissue at a given position is the one producing more novel RACEfrags with respect to the RACEfrags produced by tissues to its left on the axis. While this is a heuristic approach that does not guarantee optimality, we believe that for this particular problem, it will certainly produce a nearly optimal ranking.



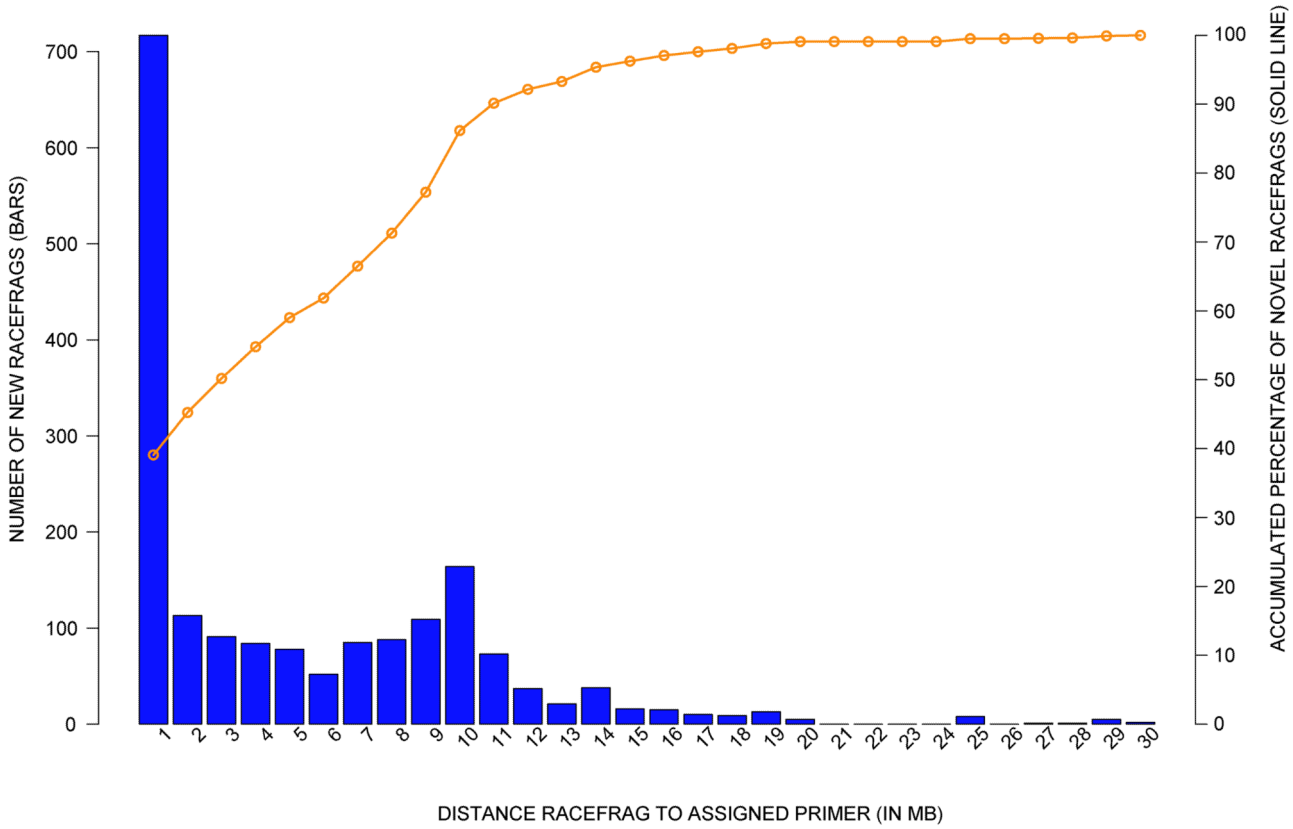


Figure 4. absolute number and cumulative proportion of all (a) and novel (b) projected RACEfrags originating from index exons

A projected RACEfrag is a maximal set of RACEfrags that transitively overlap (see Supplementary Methods and Supplementary Figure 3). In the X-axis, index exons are ordered from 3' to 5' of the gene. For instance, the most 3’ exon generates 22% of all projected RACEfrags. The two most 3’ exons generate 40 % of them, and so on. Adding the 5’ most index exon generates 10% of all projected RACEfrags.

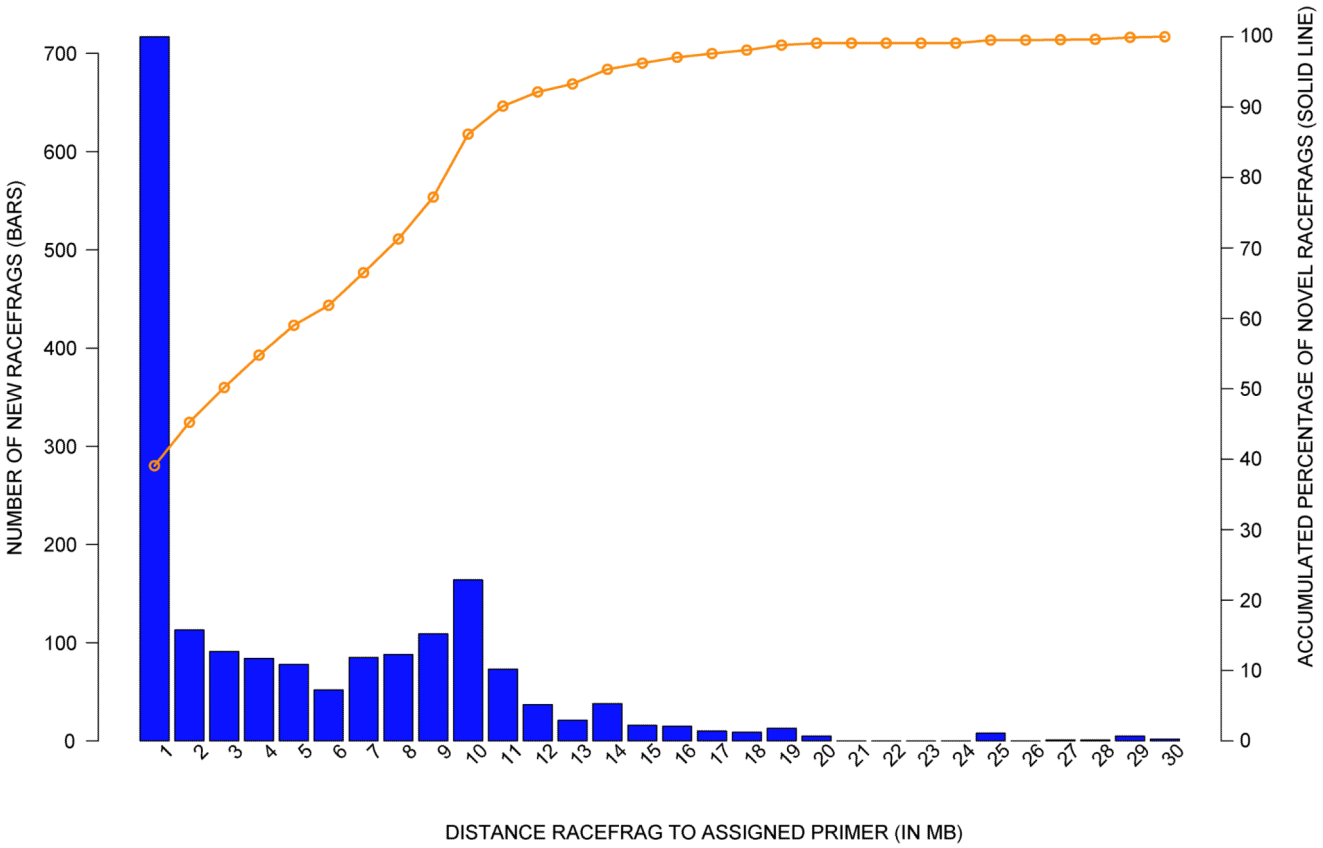


Figure 5. distribution of distances of RACEfrags to assigned index exons

The blue bars reflect the frequency of extension lengths among different length classes. The solid yellow line shows the cumulative frequency of extensions of that length or greater. Most of the RACEfrags map within 1MB of the index primer. The pick at 10MB could reflect RACEfrags 3' to the 5' index primer (originated by circularization of the cDNA); note that the average distance between pooled primers is about 10MB.

Table 1

Summary of experiments for large-scale transcript characterization.

	genes	exons	primers		tissues	RACE		pooling	arrays	# arrays (RACE/pooling)
			5’	3’		5’	3’			
Optimal combination of tissues	**12**	**12**	**12**	**12**	**48**	**576** (12 × 48)	**576** (12 × 48)	6 genes per pool. 5’ and 3’ RACE together	**C21-C22**	**96** (576 / 6)
Optimal distribution of interrogated exons	**44**	**440** (10 × 44)	**440**	—	**12**	**5,280** (440 ×12)	—	44 exons per pool, each exon from one gene.	**ENCODE**	**120** (5,280 / 44)
Genomic extent of loci	**96**	**96**	**96**	—	**12**	**1,152** (96 ×12)	—	6 genes per pool. Tissues pooled together.	**C21-C22**	**16** ((1,152 / 6)/12)