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

Sarah Djebali, Philipp Kapranov, Sylvain Foissac, Julien Lagarde, Alexandre Reymond, Catherine Ucla, Carine Wyss, Jorg Drenkow, Erica Dumais, Ryan R Murray, Chenwei Lin, David Szeto, France Denoeud, Miquel Calvo, Adam Frankish, Jennifer Harrow, Periklis Makrythanasis, Marc Vidal, Kourosh Salehi-Ashtiani, Stylianos E Antonarakis, Thomas R Gingeras & Roderic Guigó

Supplementary figures and text:

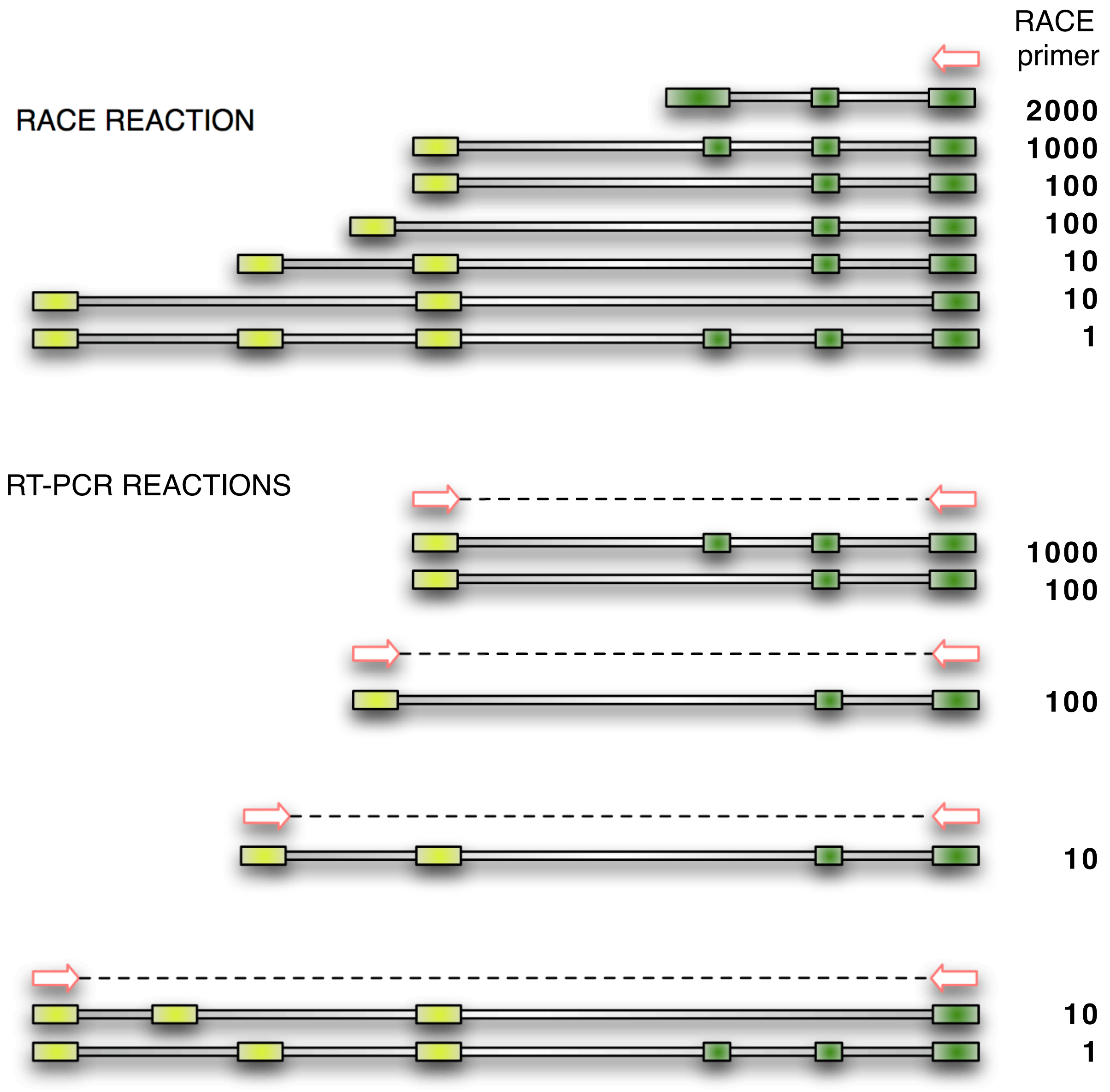

Figure S1: **efficiency of the RACEarray strategy to discover new transcript variants.** Let's assume that a RACE reaction from an annotated exon (in dark green, annotated exons) amplifies the transcripts depicted at the left, with their relative abundances registered in fold increase over the less abundant transcript. If we select 40 clones at random for sequencing, the probability that we pick the least abundant transcript clone is very low (0.01, assuming a multinomial distribution). However, let's assume that we are able to specifically identify the novel exons in the RACE mixture (for instance, through a RACEarray experiment), and that we perform targeted RT-PCR to amplify selectively all transcripts which link each of the novel exons with the originally RACEd exon. Now, if we select 40 clones for sequencing (10 at random from each RT-PCR reaction) the probability of sequencing the less abundant transcript is much larger (0.6). Of course, there is no guarantee that the highest and lowest abundant transcript end up in the same RT-PCR reaction, but in general we expect the dynamic range of each of the RT-PCR mixtures to be much smaller than the dynamic range of the original RACE mixture.

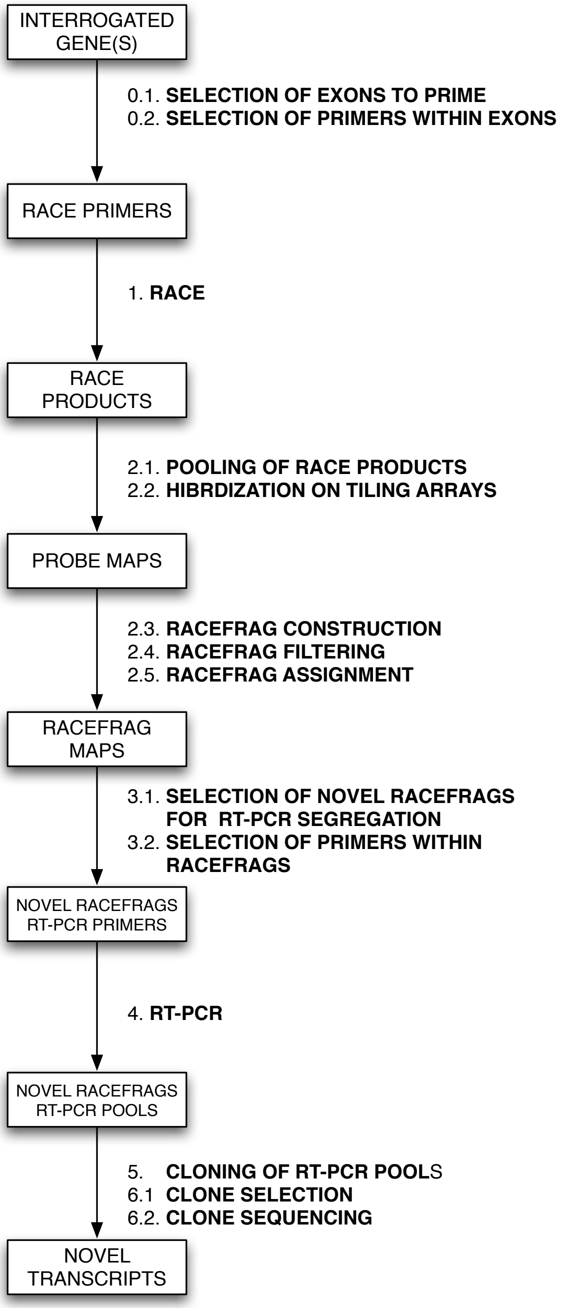
INTERROGATED GENE(S)
0.1. SELECTION OF EXONS TO PRIME
0.2. SELECTION OF PRIMERS WITHIN EXONS
RACE PRIMERS
1. RACE
RACE PRODUCTS
2.1. POOLING OF RACE PRODUCTS
2.2. HIBRDIZATION ON TILING ARRAYS
PROBE MAPS
2.3. RACEFRAG CONSTRUCTION
2.4. RACEFRAG FILTERING
2.5. RACEFRAG ASSIGNMENT
RACEFRAG MAPS
3.1. SELECTION OF NOVEL RACEFRAGS FOR RT-PCR SEGREGATION
3.2. SELECTION OF PRIMERS WITHIN RACEFRAGS
NOVEL RACEFRAGS RT-PCR PRIMERS
4. RT-PCR
NOVEL RACEFRAGS RT-PCR POOLS
5. CLONING OF RT-PCR POOLS
6.1 CLONE SELECTION
6.2. CLONE SEQUENCING
NOVEL TRANSCRIPTS

Figure S2: **the RACEarray strategy.** This figure lists the different steps of the RACEarray strategy. Steps 0.1, 2.4 and 3.1 are described in more details in Supplementary section .

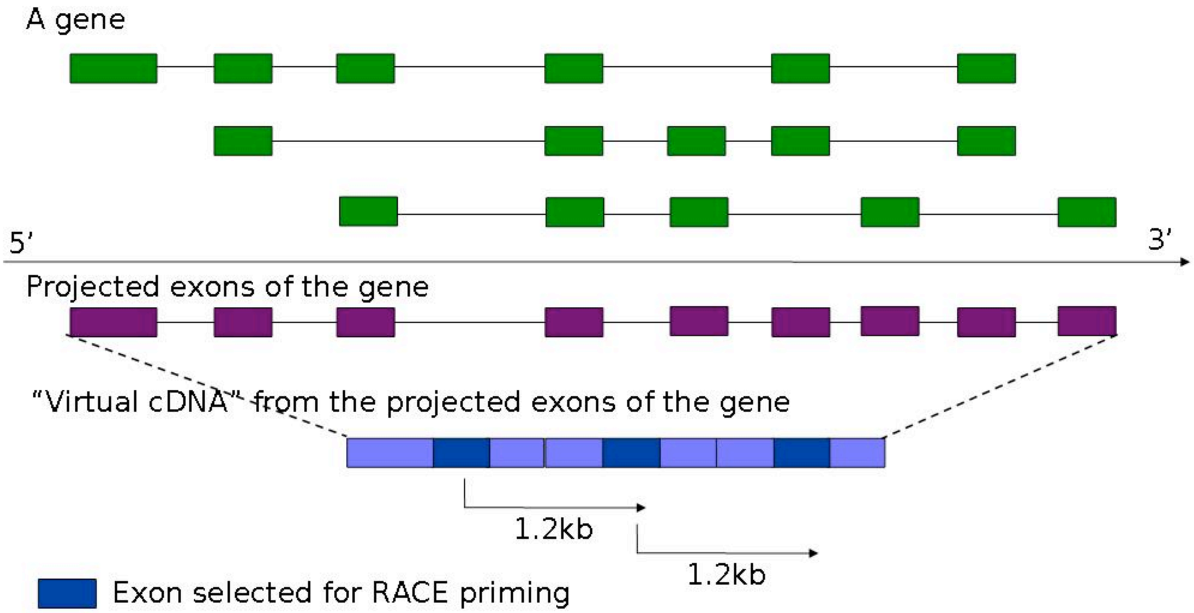
A gene
5'
3'
Projected exons of the gene
"Virtual cDNA" from the projected exons of the gene
1.2kb
1.2kb
Exon selected for RACE priming

Figure S3: **a strategy to select index exons for RACE priming.** In green is a gene made of three different transcripts. The proposed strategy to select exons for RACE priming within this gene is the following: (1) project all the exons of the gene on the genomic sequence, and concatenate them into a ”virtual cDNA”, (2) take the most $5'$ exon (of a transcript) included in the more different transcripts, (3) from the position defined by this exon in the virtual cDNA, go along the virtual cDNA from $5'$ to $3'$ and select one exon every $1.2kb$ (in case of many possible exons take the one included in more transcripts).

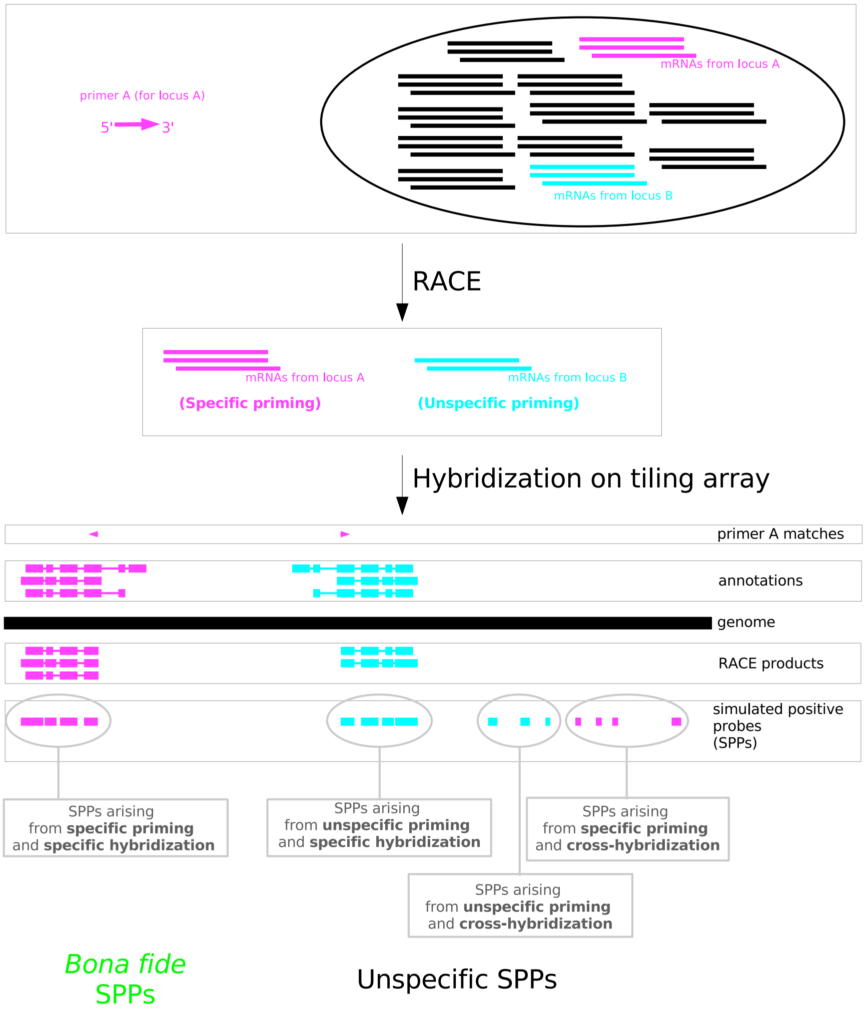
primer A (for locus A)
5'
3'
mRNAs from locus A
mRNAs from locus B
RACE
mRNAs from locus A
mRNAs from locus B
(Specific priming)
(Unspecific priming)
Hybridization on tiling array
primer A matches
annotations
genome
RACE products
simulated positive probes (SPPs)
SPPs arising from **specific priming** and **specific hybridization**
SPPs arising from **unspecific priming** and **specific hybridization**
SPPs arising from **specific priming** and **cross-hybridization**
SPPs arising from **unspecific priming** and **cross-hybridization**
Bona fide SPPs
Unspecific SPPs

Figure S4: **the RACEarray simulator.** This simulation involves two steps: (1) RACE from a set of primers and a set of known transcripts; (2) hybridization of the obtained RACE products on tiling arrays. The RACEarray simulator generates a set of tiling array probes that are highlighted by the RACE products and that we call simulated positive probes (SPPs). These SPPs can be further divided into two categories: (1) bona fide SPPs, i.e. overlapping an exon of the target locus; (2) unspecific SPPs, also called USPPs, i.e. mapping outside of the target locus exons. In our model these USPPs correspond to false positives that originate from RACE mispriming and/or from array cross-hybridization (see text for a more detailed explanation).

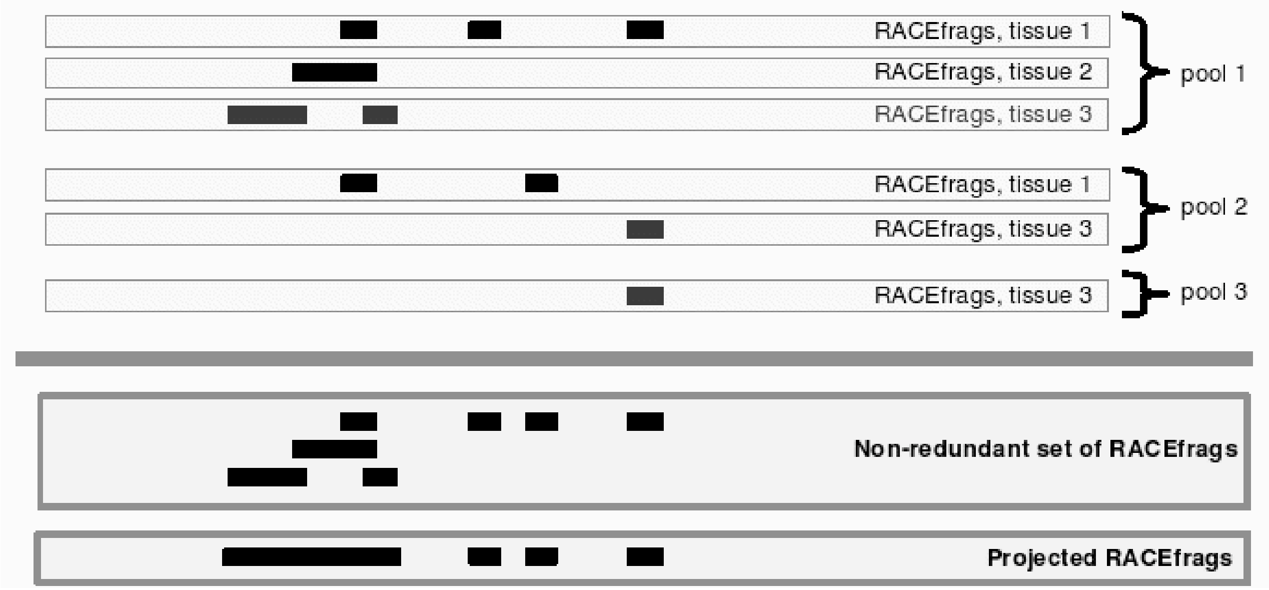

Figure S5: **definitions of RACEfrag, non redundant RACEfrag, projected RACEfrag.** RACEfrags originating from RACE reactions performed in different experiments (pools 1, 2, 3, tissues 1, 2, 3) are represented as black boxes on the top of the figure. As we can see, different experiments can identify the same or almost the same genomic region as a RACEfrag. To remove this redundancy, we define two kinds of objects: the non redundant RACEfrag and the projected RACEfrag. For example, the last RACEfrags of (pool 1, tissue 1), (pool 2, tissue 3) and (pool 3, tissue 3) have the same boundaries: they define the same non redundant RACEfrag. On the other hand, RACEfrags that transitively overlap on the genomic sequence define the same projected RACEfrag, as for example the four first RACEfrags of pool 1 and the first RACEfrag of pool 2, tissue 1.

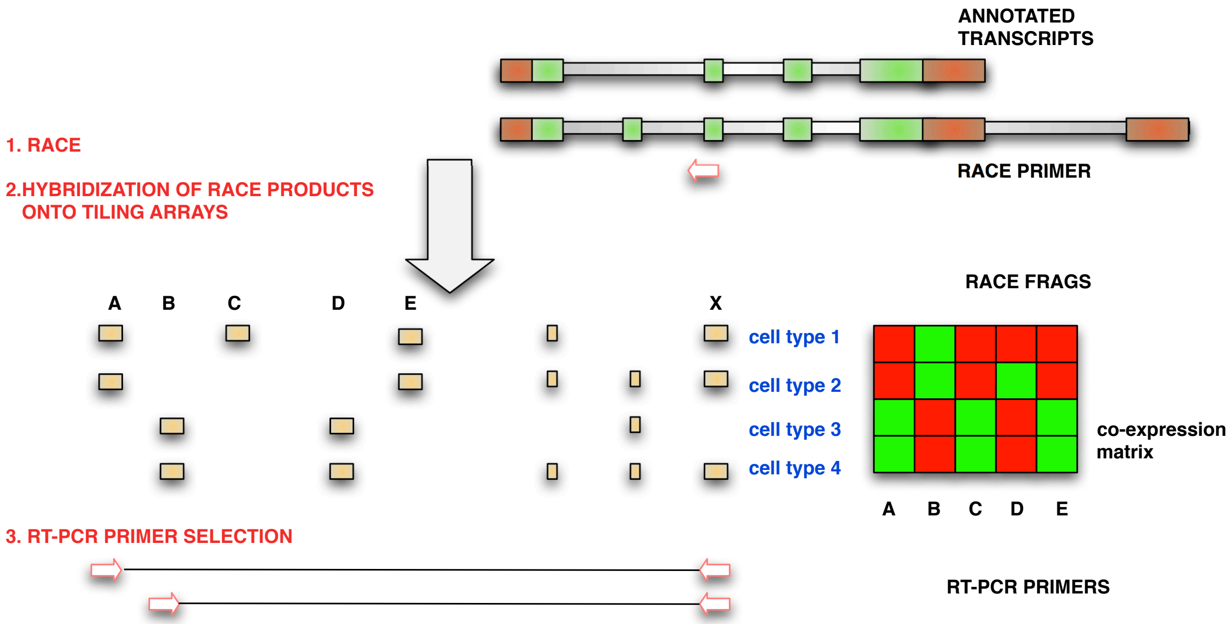

Figure S6: **co-occurrence of RACEfrags across cell lines can be used to minimize the number of RACEfrags to RT-PCR to uncover novel transcripts.** In the figure, RACEfrags C and E are only present when RACEfrag A is present, whereas RACEfrag D is only present when RACEfrag B is present. Therefore RT-PCR reactions from A to X and from B to X may suffice to recover all transcript diversity. RACEfrag co-occurrence can be quantified, if needed, from the ”co-expression” matrix on the right (see also main document).

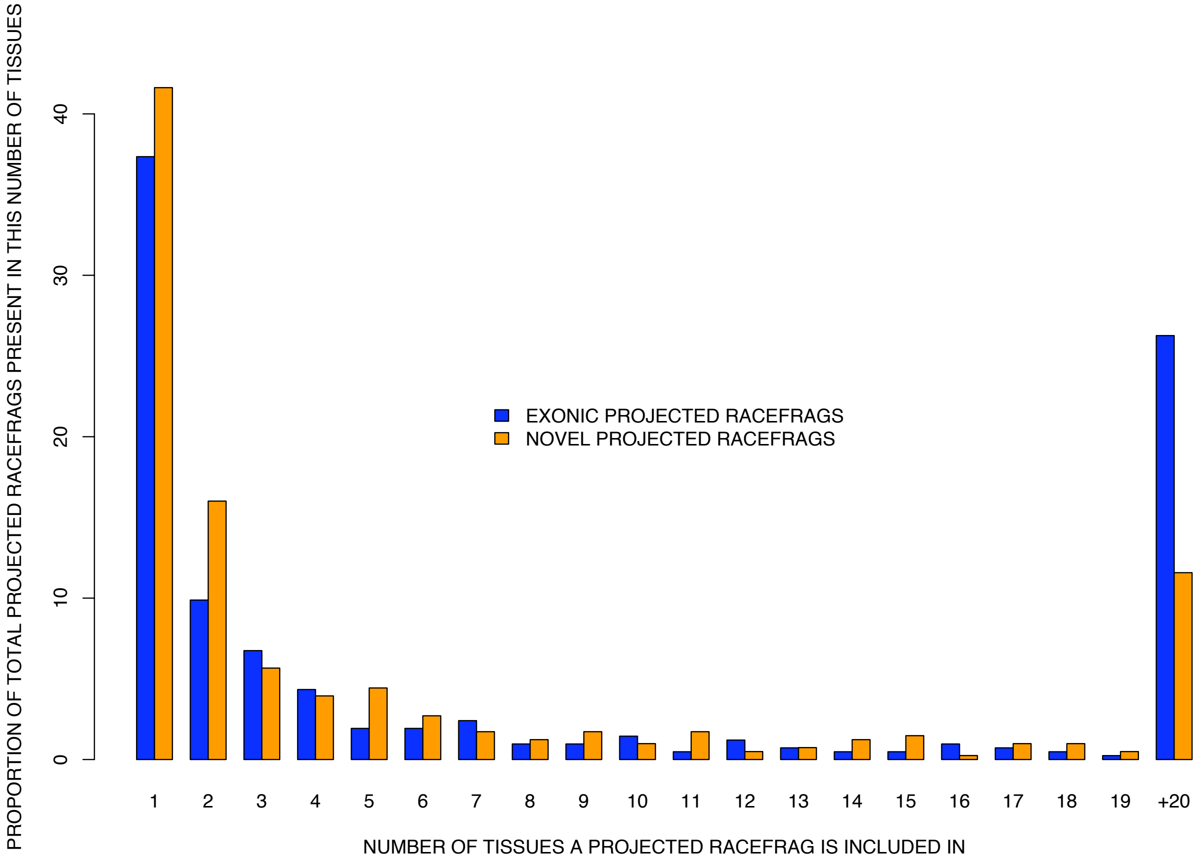

Figure S7: **expression pattern for exonic and novel RACEfrags.** For each number of tissues the proportion of total projected RACEfrag present in this number of tissues is plotted, separately for exonic and novel projected RACEfrags. While novel projected RACEfrags show a restricted expression pattern when compared to annotated projected RACEfrags (6.8 vs. 13.0 tissues on average, respectively), a substantial fraction of them (58%) seem to be expressed in more than one tissue (compared with 62% of the annotated projected RACEfrags).

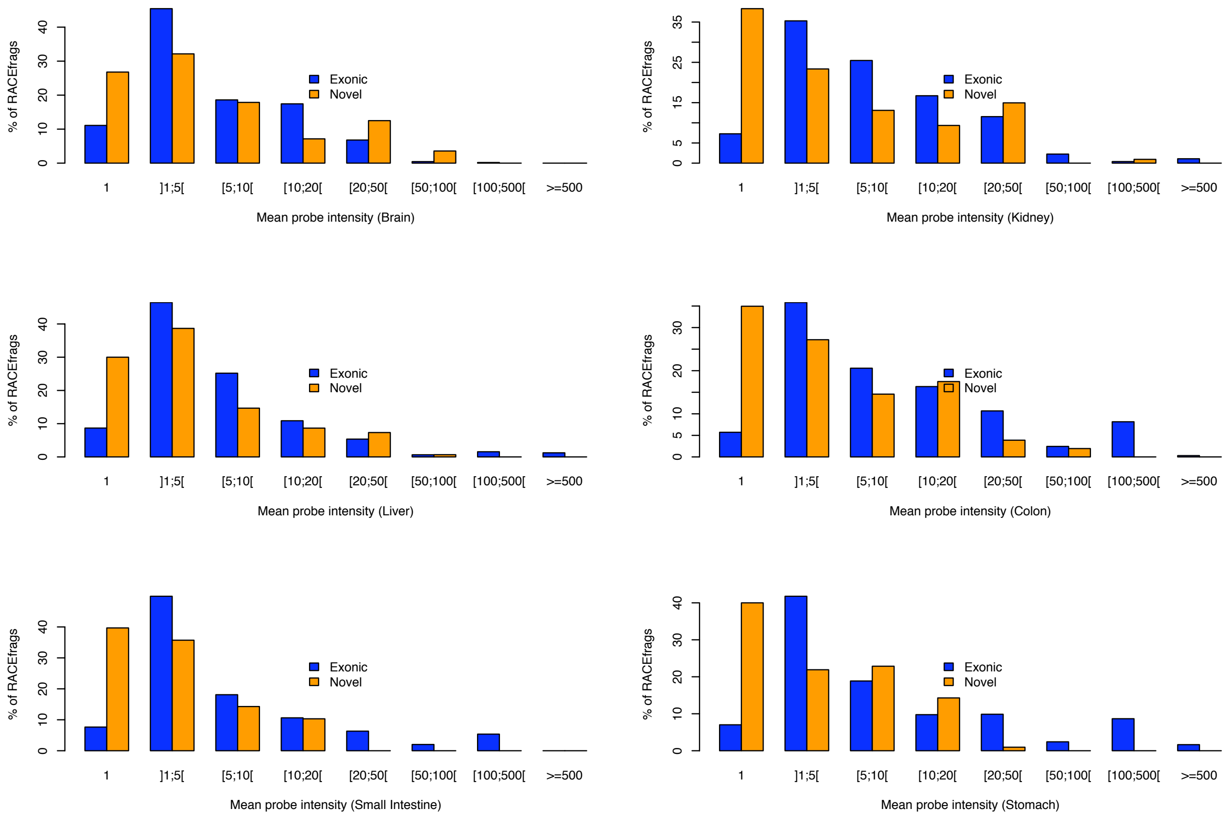

Figure S8: **RACEfrag expression levels from RNA transcriptional maps.** The ENCODE array RACEfrags have been compared to the transcriptome maps on the ENCODE regions generated from polyA+ RNA in six different tissues (brain, kidney, liver, colon, small intestine, stomach, see [4]). More precisely, each RACEfrag has been attributed an expression value which is the geometric mean of the intensities of the transcriptome map probes overlapped by this RACEfrag in the corresponding tissue. As we can see, exonic RACEfrags are in general more highly expressed than novel ones.

Supplementary tables

TARGET LOCUS	ANNOTATED ISOFORMS OF TARGET LOCUS	DISTANCE RACEFRAG TO INDEX PRIMER	cDNA LENGTH	NUMBER OF SEQUENCES
NM_020528	6 UCSC	$2.5kb$	260	30
APP	5 UCSC	$7kb$	$694 + 652$	$29 + 29$
IL10RB	2 UCSC / 5 Gencode	$15kb$	$696 + 652$ 171	$2 + 2$ 10
NM_005441	1 UCSC	$15kb$	867 1028 673 934 875 750 513 652	5 7 6 3 1 3 9 6
APP	5 UCSC	$17kb$	361	32
NM_013396	6 UCSC	$45kb$	775 615 584 757	8 7 1 16
RP5-858B16.1	5 UCSC / 19 Gencode	$80kb$	338 501	16 15
NM_001011545	2 UCSC	$127kb$	325 315 485 349	21 4 1 5
NM_170736	1 UCSC	$145kb$	599 626 350 304 740 555 413 313 563 627	2 1 1 1 1 10 2 6 1 2

Table S1: **RT-PCR cloning and sequencing from novel RACEfrags.** Each row corresponds to a novel transcript sequence; the number of clones supporting it is given in the last column. The number of previously annotated transcripts per each annotated locus and the genomic distance between the RACEfrag and the index exon are reported for each locus. For each novel transcript we report the length of the corresponding cDNA, and the number of sequences supporting it. In two instances (targets APP and IL10RB) none of the forward reads could be assembled with their respective reverse counterparts. In such cases the ”cDNA length” and ”Number of sequences” features are described separately for forward and reverse reads (within the same cell).

	Number of features	Overlap with CAGE data (%)	Overlap with PET data (%)
Most 5’ RACEfrags of each gene and each tissue	858	37.9	22.5
Non most 5’ RACEfrags of each gene and each tissue	16,438	11.5	4.8
Random set of RACEfrags generated from the most 5’ RACEfrags of each gene and each tissue	858	1.0	0.2

Table S2: **Comparison of RACEfrags with sites of transcription detected by other unbiased interrogation techniques.** The most $5'$ ENCODE array RACEfrags of each gene and each tissue (first class), as well as the non most $5'$ RACEfrags of each gene and each tissue (second class), have been compared to CAGE tags and PET ditag sequences. The proportion of RACEfrags overlapping this data is much higher for the first than for the second class, and is also much higher for the first class than expected by chance.

Supplementary methods

A Probabilistic model for array based normalization

In this section, we first lay out a formal framework to investigate sampling from a population of transcript species. Following previous work (see, for instance [15]), we model sampling from such a population as a multinomial process. Assuming that the population is made of known and novel transcript species, we first derive an expression for the probability of sampling at least one exemplar of each novel species after random sampling a fixed number of clones. Using this expression, we investigate the conditions in the structure of the sampled population that optimize such probability. These conditions are the expected outcome of the RACEarray strategy, through the segregation of the initial population of transcripts into smaller subpopulations–each one containing transcript from a subset of species. Therefore, we next derive an expression for the probability of sampling at least one exemplar of each novel species after random sampling from the segregated subpopulations. Finally, we have carried out extensive simulations in different scenarios in which we have compared this probability with the initial probability before segregations, and we have assessed in this way the sampling efficiency achieved by segregation. In our case, the original transcript population is the population of transcripts in a RACE reaction, while the segregated populations are the transcript populations in the RT-PCR reactions delineated after the hybridization of the RACE reaction into a tiling array.

Formalization of the problem

Let's assume that the result of a given RACE reaction generates RNA species $a_1 \cdots a_s$, with probabilities $p_1 \cdots p_s$. We assume that all other non-amplified transcript species have probability zero. Let's assume that the RACE transcripts are subsequently cloned, and that the clone population maintains the probabilities. Random selection from such a population can be modeled as a multinomial process [15]. Indeed let's assume that we select n clones at random from the population, and let X_i be the number of times a clone containing the species a_i is selected among these n clones. Assuming independence between each clone, then $(X_1, \cdots, X_s)$ is multinomially distributed. Therefore, the probability to obtain $x_1, \cdots, x_s$ clones of the RNA species $a_1, \cdots, a_s$ is:

$$P(X_1 = x_1, \cdots, X_s = x_s) = \frac{n!}{x_1! \cdots x_s!} p_1^{x_1} \cdots p_s^{x_s} = \frac{n!}{\prod_{j=1}^{s} x_j!} \prod_{j=1}^{s} p_j^{x_j} \tag{1}$$

Along this document, we will denote with $\vec{p} = (p_1, p_2, \ldots, p_s)$ the vector of probabilities of the different mRNA species.

Let's assume that the species $a_1 \cdots a_r$ are novel, and the species $a_{r+1}, \cdots, a_s$ have already been previously identified. Let us denote with $prob(A(s, r, n|\vec{p}))$ the probability of obtaining at least one clone of each novel species after random sampling n clones. This probability is:

$$prob(A(s,r,n|\vec{p})) = prob\left(\bigcap_{i=1}^{r}(X_i \geq 1) \Bigg/ \sum_{j=1}^{s} X_j = n\right) \tag{2}$$

Some analytical results

In this section we briefly present some key analytical results, without detailed demonstrations. After obtaining a compact expression for $prob(A(s,r,n|\vec{p})))$, we show that in a population of transcripts, the more homogeneous the frequencies of the different transcript species, the higher the probability of obtaining at least one exemplar of each species after random sampling the population. In addition, we show that this probability also increases as the frequencies of the known species decrease.

Compact expressions for $prob(A(s,r,n|\vec{p}))$ Combining expressions (1) and (2), and applying the *inclusion-exclusion principle*, we obtain after some simplifications:

$$prob(A(s,r,n|\vec{p})) = 1 + \sum_{j=1}^{r}(-1)^j \left(\sum_{(i_1,\ldots,i_j)\in C(j,r)} (1 - \sum_{k=1}^{j} p_{i_k})^n\right) \tag{3}$$

where the set of index $\mathbf{C}(a,b)$ is defined for $a, b \in \mathbb{N}$ and $a \leq b$ as:

$$\mathbf{C}(a,b) = \{(i_1, i_2, \ldots, i_a) \in \mathbb{N}^a : 1 \leq i_j \leq i_{j+1} \leq b \quad \text{for} \quad 1 \leq j \leq a\}$$

A more detailed expression equivalent to (3) is:

$$\begin{array}{rcl} prob(A(s,r,n|\vec{p})) & = & 1 - (1-p_1)^n - (1-p_2)^n - \ldots - (1-p_r)^n \\ & + & (1-p_1-p_2)^n + (1-p_1-p_3)^n + \ldots + (1-p_1-p_r)^n \\ & + & (1-p_2-p_3)^n + (1-p_2-p_4)^n + \ldots + (1-p_2-p_r)^n \\ & + & \ldots + (1-p_{r-1}-p_r)^n \\ & - & (1-p_1-p_2-p_3)^n - \ldots - (1-p_{r-2}-p_{r-1}-p_r)^n \\ & + & \ldots + (-1)^r(1-p_1-p_2-\ldots-p_r)^n \end{array}$$

Note that when $s = r$ the last term vanishes. When the first r species are equiprobable, that is, $\vec{p} = (q^{-1}, q^{-1}, \ldots, q^{-1}, p_{r+1}, \ldots, p_s)$ the expression (3) simplifies to

$$prob(A(s,r,n|\vec{p})) = 1 + \sum_{j=1}^{r}(-1)^j \binom{r}{j} \left(1 - \frac{j}{q}\right)^n$$

and, in particular, if $r = s$, then $\vec{p}$ becomes $(r^{-1}, r^{-1}, \ldots, r^{-1})$, and therefore

$$prob(A(s,r,n|\vec{p})) = 1 + \frac{1}{r^n}\sum_{j=1}^{r-1}(-1)^j \binom{r}{j}(r-j)^n$$

On the increment of $prob(A(s,r,n|\vec{p}))$ when the relative abundances of RNA species $1,\ldots,r$ tend to $1/r$ The basic result here is the demonstration that given two RNA species a_1 and a_2 with probabilities $p_1 < p_2$, and assuming that the probabilities of all other species in the population remain unchanged, the increase in the probability of the lower probability species a_1, and the decrease in the same amount of the probability of the higher probability species a_2, leads to an increase of the probability of sampling at least one exemplar of each species in the population, $prob(A(s,r,n|\vec{p}))$. More specifically, assuming $r \geq 2$, $p_1 < p_2$ and $\Delta > 0$ with $p_1+\Delta \leq p_2-\Delta$. We have demonstrated that

$$prob(A(s,r,n|p_1,p_2,\ldots,p_r,\ldots,p_s)) < prob(A(s,r,n|p_1+\Delta,p_2-\Delta,\ldots,p_r,\ldots,p_s)) \qquad (4)$$

The result is valid along the range of probabilities between p_1 and the *relative equiprobability* of the two RNA species, that is, $(p_1+p_2)/2$.

From this basic result, we have been able to derive some other analytical properties of $prob(A(s,r,n|\vec{p}))$. We summarize here only the most relevant ones in the context of the RACEarray application:

1. Let's assume that $\vec{p}$ and $\vec{q}$ are two probability vectors such that the last $s-r$ components are identical

   $$\begin{aligned} \vec{p} &= (p_1,p_2,\ldots,p_r,p_{r+1},\ldots,p_s) \\ \vec{q} &= (q_1,q_2,\ldots,q_r,p_{r+1},\ldots,p_s) \end{aligned}$$

   if $d(\cdot,\cdot)$ is the Euclidean distance in $\mathbb{R}^s$, $\vec{r} = (r^{-1},r^{-1},\ldots,r^{-1})$ and if for any $1 \leq j \leq s$

   $$\begin{aligned} sign(p_j - r^{-1}) &= sign(q_j - r^{-1}) \quad \text{and} \\ d(\vec{p},\vec{r}) &> d(\vec{q},\vec{r}) \end{aligned}$$

   then

   $$prob(A(s,r,n|\vec{p})) \leq prob(A(s,r,n|\vec{q}))$$

   That is, the closer to the equiprobability the frequencies of the transcript species, the more likely we are to select at least on exemplar of each species after selecting a fixed number of clones. Homogenization of transcript abundances, through segregation into smaller subpopulations, is actually one of the outcomes often expected from the array based normalization strategy described here (see next section).

2. Let's assume $r < s$, and let be $\Delta > 0$ with $p_1 + \Delta \leq 1$ and $p_s - \Delta \geq 0$, therefore

   $$prob(A(s,r,n|p_1,p_2,\ldots,p_s)) < prob(A(s,r,n|p_1+\Delta,p_2,\ldots,p_s-\Delta)) \qquad (5)$$

   This simply formalizes the obvious observation that, as the probability of the already known species decrease in the population, the probability increases of sampling at least one exemplar of each novel species. Note that, in the RACEarray strategy, the probabilities of the already know species are, by construction of the RT-PCR reactions, null.

3. The global maximum of $prob(A(s,r,n|\vec{p}))$ is attained at

$$\vec{p} = r^{-1}(\delta_{1,r}, \delta_{2,r}, \ldots, \delta_{s,r}) \tag{6}$$

   where $\delta_{i,r} = 1$ if $i \leq r$, $\delta_{i,r} = 0$ otherwise. Therefore, the probability of sampling at least one exemplar of each novel species is maximal when $p_1 = \cdots = p_r = 1/r$, that is, when all novel species are equiprobable and all known species have probability zero.

Sampling probabilities in segregated populations Let us assume k different subpopulations $A_1, \ldots, A_k$ each one with r_i species:

$$\begin{array}{lll}
A_1\text{:} & r_1 & \text{species with probability} \quad p_1^{(1)}, \ldots, p_{r_1}^{(1)} \\
A_2\text{:} & r_2 & \text{species with probability} \quad p_1^{(2)}, \ldots, p_{r_2}^{(2)} \\
\ldots & & \\
A_k\text{:} & r_k & \text{species with probability} \quad p_1^{(k)}, \ldots, p_{r_k}^{(k)}
\end{array}$$

with $\sum_{i=1}^{k} r_i = r$ and $\sum_{i=1}^{k} \sum_{j=1}^{r_i} p_j^{(i)} = 1$. $A_1, \cdots, A_k$ are the transcript subpopulations generated by the RT-PCR reactions originating in each novel RACEfrag.

The new probabilities $p_j^{(i)*}$ within each of the segregated subpopulations are:

$$A_i : p_j^{(i)*} = \frac{p_j^{(i)}}{\sum_{j=1}^{r_i} p_j^{(i)}} \tag{7}$$

Let us assume that the original n replicates are sampled equally in each of the subsets, that is, the effective sampling on each subpopulation is $\frac{n}{k}$. Let us denote with $prob(B(r,n,k|\vec{p}))$ the probability of the event *after segregation, all the $r_1, \ldots, r_k$ species of all the k subpopulations A_i are observed at least once*. This probability can be expressed as:

$$prob(B(r,n,k|\vec{p})) = \prod_{m=1}^{k} \left(1 + \sum_{j=1}^{r_m} (-1)^j \left(\sum_{(i_1,\ldots,i_j) \in C(j,r_i)} \left(1 - \sum_{b=1}^{j} \frac{p_{i_b}^{(m)}}{\sum_{c=1}^{r_m} p_c^{(m)}} \right)^{\frac{n}{k}} \right) \right) \tag{8}$$

since the samples are obtained independently in each subpopulation.

In the next section, we have used this expression to explore, through exhaustive simulations, the sampling efficiency of the segregation of the initial RACE population of transcripts into the RT-PCR subpopulations.

Some numerical results about $prob(A(s,r,n|\vec{p}))$ and $prob(B(r,n,k|\vec{p}))$

We had explored numerically the parameter space of the multinomial distribution quantifying the differences in $prob(A(s,r,n|\vec{p}))$ and $prob(B(r,n,k|\vec{p}))$, that is the differences in the probability of sampling at least one exemplar of each species in the population before and after segregation into subpopulations. Simulations are described at `http://genome.imim.es/datasets/racearrays2007/`. In summary, the simulation results suggest that the segregation into subpopulations is, in general, an effective strategy to exhaustively sample across all transcript species. Segregation appears to be particularly effective, first, when the number of species across subpopulations is similar, and, second, when the probabilities of the species within subpopulations are homogeneous. Two outcomes that we believe are sensible to expect from the array induced segregation of the RACE population into RT-PCR subpopulations (see Supplementary Fig. S1 for a cartoon example, and the main text for justification).

RACEarray normalization

Supplementary figure S2 schematizes the RACEarray normalization strategy targeted novel transcript discovery (see main document for a description; the numbering of the steps matches the numbering in Fig. 1). Many of the components of the strategy (RACE, RT-PCR, hybridization on tiling arrays, cloning, and sequencing) are performed following quite standard procedures and it is mostly their combination through the pipeline which converts them into a powerful approach for novel transcript discovery. We have developed specific algorithms and bioinformatics applications to support some of the experimental steps in the pipeline. Specifically, we have delineated a protocol to distribute RACE primers in exons along the interrogated locus to maximize transcript discovery, we have implemented an *in silico* RACEarray simulator to identify RACEfrags that can be produced by the amplification of non-targeted loci through mis-priming, and we have conceived an optimal protocol to minimize redundancy when delineating the RT-PCR segregation reactions. We describe these algorithms and programs next. All of them are accessible through the web site http://genome.imim.es/datasets/racearrays2007/

Selection of exons within genes for RACE priming (0.1)

The main results of the experiments performed to investigate the optimal distribution of primers along exons for RACEarray transcript discovery can be summarized as follows (see "optimal distribution of exons to RACE per locus" in main document results section, and Supplementary section):

- an exon generates a larger number of RACEfrags (i.e., is more informative) as it is included in a greater number of alternative transcripts,
- an exon is able to produce RACEfrags at a distance within a cDNA sequence of up to $1.2kb$,
- within a gene, the exons that generate a larger number of RACEfrags (i.e. the most informative) are the terminal exons (Fig. 4).

Based on these results, we propose the following strategy to select index exons within genes for $5'$ RACE priming (see Supplementary Fig. S3):

1. project all the exons of the gene, and concatenate the obtained projected exons into a ”virtual cDNA”,

2. select the most $5'$ exon (of a transcript) included in a larger number different alternative transcript isoforms,

3. from the position defined by this exon in the virtual cDNA, go along the virtual cDNA from $5'$ to $3'$, and select one exon every $1.2kb$ (if there is a choice between several different exons, then take the one included in a larger number of different transcripts).

The converse protocol can be applied to select exons for $3'$ RACE priming.

in silico **RACEarray simulator (2.4)**

We have conceived a program that simulates *in silico* the RACEarray experiments. Starting from a known set of transcripts, primers and array probes, it generates simulated RACEfrag maps from which we can confidently discriminate between *bona fide* RACEfrags (i.e., originating from specific priming and specific array hybridization) and artefactual ones (i.e., arising from RACE mis-priming and/or array cross-hybridization).

This simulation involves two steps (see Supplementary Fig. S4):

1. *in silico* RACE, which consists in searching each RACE primer against the transcriptome (we use RefSeq RNAs [13] + Unigene [1] + Gencode [5] transcripts). A primer is considered matching an mRNA if it aligns to it with more than 95% identity over more than 60% of its sequence, and if the alignment includes one of the 4 $3'$-most nucleotides of the primer. When a match is found, the primer is elongated until the end of the transcript, in either the $5'$ or the $3'$ direction. At this point, and for each primer, we end up with a population of RACE products, arising from both specific and unspecific priming.

2. *in silico* array hybridization, which consists in scanning the two strands of all these RACE products with all probes of the tiling array. We can allow for 0 or more mismatches for a probe to be considered positive.

Since our starting material consists in known transcripts that have been manually mapped to the genome, we do not expect any of the RACE products obtained at step 1 to highlight probes outside of the annotated exons of the target locus; if they do, these highlighted probes will be considered unspecific.

The result of this step is a set of Simulated Positive Probes (SPPs), that we can split into the following categories: Bona-fide SPPs: these are SPPs that overlap the annotated target locus exons; Unspecific SPPs (USPPs): these are SPPs that map outside of the target locus exons. In our model, they clearly correspond to false positives, originating from RACE mis-priming and/or array cross-hybridization.

We have used the *in silico* RACEarray simulator in all the RACEarray experiments described in this paper (see “Multiplexing of RACEarray normalization” in main document results section, and Supplementary section). Results indicate that the RACEarray simulator filters out between about 2 and 10% of all the RACEfrags (see tables W2, W3 and W4 at http://genome.imim.es/datasets/racearrays2007/).

Selection of novel exons for RT-PCR (3.1)

Ideally, one RT-PCR reaction should be designed for each novel RACEfrag. However, since RT-PCR reactions originating from different novel RACEfrags may produce largely overlapping transcript populations, more efficient strategies can be designed, based on the pattern of co-occurrence of RACEfrags across different assayed conditions, to minimize redundant RT-PCR reactions. Indeed, different RACEfrags may correspond to the same underlying transcripts; for instance, in those cases in which we have events of double exon inclusion/skipping. We can predict cases of correlated exon inclusion by observing the pattern of co-occurrence of RACEfrags across cell conditions. Supplementary figure S6 depicts an hypothetical example. A RACEarray experiment in four cell lines originating from a primer in annotated exon X has generated 5 novel RACEfrags (A,B,C,D,E). In principle, to recover all associated novel transcript sequences, RT-PCR reactions need to be carried out for each of these novel RACEfrags. However, by investigating the degree of co-occurrence between novel RACEfrags across the cell lines, we infer that RT-PCRs from RACEfrags A and B will likely suffice to capture most of the associated transcript diversity. This can be formalized in the following way. Let R be the set of all RACEfrags generated in a hybridization experiment from a $5'$ primer in a gene oriented in the forward strand in the reference genes, and carried out in tissues $t_1, \cdots, t_n$. Let’s represent a RACEfrag r as a tuple $r = (r_1, \cdots, r_{n+2})$, with r_1 and r_2, the genomic coordinates of the RACEfrag, and $r_{i+2} = 1$ $(i \in \{1, \cdots, n\})$ if and only if RACEfrag r is expressed in tissue t_i. Let use R_i to denote the set of RACEfrags expressed in tissue t_i, that is $R_i = \{r \in R \mid r_{i+2} = 1\}$. We will say that RACEfrag x is included in RACEfrag y , $x \subset y$, if and only if $y_2 < x_1$ and if $y \in R_i$, then $x \in R_i$ (for all $i \in \{1, \cdots, n\}$). Then, the set of optimal RACEfrags for RT-PCR can be defined as $R_0 = \{x \in R \mid x \not\subset y, \forall y \in R\}$.

RACEarray experimental methods

We describe here the experimental protocols under which the novel isoform discovery and of the multiplexing of RACEarray normalization experiments, were actually performed.

RACE reactions

To uncover the ideal number and choice of tissues to get the best coverage of transcription (both annotated and newly-identified exons), we performed $5'$ and $3'$ RACEs of 12 genes on polyA+ RNAs from 48 different cell/tissue types, i.e. 33 human adult tissues (Adrenal Gland, Bladder, Brain, Brain Frontal Lobe, Brain Hippocampus, Brain Hypothalamus, Cerebellum, Colon, Epididymus, Heart, Kidney, Liver, Lu, Lymph Node, Mammary Gland, Muscle, Ovary, Pancreas, Pituitary Gland, Placenta, Prostate, Salivary Gland, Skin, Small Intestine, Spinal Cord, Spleen, Stomach, Testis, Thymus, Thyroid, Tongue, Tonsil, Uterus; all BD Clontech), 9 fetal tissues (Fetal Adrenal Gland, Fetal Brain, Fetal Heart, Fetal Kidney, Fetal Liver, Fetal Lung, Fetal Spleen, Fetal Thymus, Whole Fetus; all BD Clontech) and 6 cell lines (GM06990, HeLaS3, HepG2, HL60, K562, SW480) using the BD SMARTTM RACE cDNA amplification kit (BD Clontech Cat. No.634914). To define how many exons per gene should be interrogated, we performed $5'$ RACE experiments in ten exons of a set of 44 genes, each mapping to a different ENCODE region [3, 2] on polyA+ RNAs from 12 tissues (brain, heart, kidney, spleen, liver, colon, small intestine, muscle, lung, stomach, testis, placenta, all BD Clontech). The same set of 12 RNAs and $5'$ RACE of 96 genes mapping to HSA21 and HSA22 was used to estimate the genomic space occupied by a gene. 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 `primer3` (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) with the following parameters: $23 \leq primer\ size \leq 27, optimal\ size = 25, 68^oC \leq primer\ Tm \leq 72^oC, optimal\ Tm = 70^oC, 50\% = primer\ GC\ percentage = 70\%$.

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

The products of RACE PCR reactions were pooled and purified by ethanol precipitation. The pooled amplicons were then fragmented to $\sim 50-100$ nucleotides using DNAse I (Epicenter) at concentration of $0.00625U/\mu g$ of DNA for 8 minutes at 37^oC followed by boiling for 10 min and cooling on ice. The fragmented amplicons were labeled in the following conditions: $1X$ terminal transferase (TdT) reaction buffer (Roche); $2.5mM$ CoCl2, DNA labeling reagent (DLR; Affymetrix; 1.125 nmol per $1\mu g$ of fragmented DNA) and 200 units of recombinant TdT (Roche) for 3 hours at 37^oC. The labeled products were used directly for array hybridization in TMAC-based buffer as described previously [9, 8].

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). The graphs were generated without smoothing and median scaling. Each value on the graph represents intensity of the perfect match (PM) probe minus the intensity of the cognate mismatch (MM) probe. The coordinates of the probe always represent the ”left” genomic coordinate of a 25-mer. All data is in the coordinates of the `hs.NCBIv35` version of the human genome. To minimize the number of falsely positive probes arising from cross-hybridization on the tiling array, we discard all positive probes having more than one perfect hit without gap in the whole genome.

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 section for details). 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 to loci on chromosomes 21 and 22, and were obtained through the experiments described in the section ”Pooling of RACE reactions-genomic extent of loci”. In these experiments, 96 protein coding loci were interrogated, and they generated $3,292$ novel RACEfrags (that is, not corresponding to annotated exons from the index loci). We further considered only RACEfrags with very good putative donor sites (with score over 2.4 as computed by the geneid program [7, 12]) in the vicinity (within -10 to $+30$ nucleotides) of the $3'$ end of the RACEfrag, and within $500kb$ 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 (up to the predicted donor site, as predicted by the geneid program), 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 was used to pick primers.

RT-PCR

Total RNA pooled from five tissues (brain, testis, liver, placenta, heart) or 20 tissues (brain, testis, liver, placenta, adipose, bladder, cervix, colon, esophagus, heart, kidney, lung, ovary, prostate, skeletal muscle, small intestine, spleen, thymus, thyroid, trachea) were reverse transcribed by priming the reaction by dT_{16} ($1\ \mu M$) or random hexamers ($5\ \mu M$) at final RNA concentration of $0.3\ \mu g/\mu l$. For each RT-PCR amplification, nested primers were designed such that the internal primers contain Gateway tails to facilitate cloning of the products. ”Touchdown” PCR was done using ”hot start” KOD polymerase (Novagen Corp.) with the first 10 cycles done at 10^oC above the primer Tm and then $25-35$ cycles at the Tm. For the second PCR, a small portion of the first PCR ($\sim 0.5\mu l$) was transferred to new PCR plates, and the nested gene-specific primers, which contain the Gateway tails, were added. PCR products were verified by gel electrophoresis. For the MECP2 experiments, the cell types used were Testis, TerBJ, Prostate, Placenta, Ovary, K562, HepG2, HeLaS3, GM06990, Fetal Thymus, Fetal Spleen, Fetal Kidney, Hypothalamus, Hyppocampus, Frontal lobe and Cerebellum.

Cloning and sequencing

Briefly, PCR products (carrying Gateway recombination sites at their extremities) were used directly in Gateway BP reactions with `pDONR223` vector as described in Rual et al. [14, 6]. The products from the BP reactions were used to transform chemically competent $DH5\alpha$ E. coli, and plated on spectinomycin LB plates for growth and selection. Eight to 16 colonies from each transformation reaction were picked, grown in LB-spectinomycin media, then, a small portion of the liquid media containing the bacteria were used as ”template” to PCR amplify the cloned insert using universal vector primers ($M13$ forward and reverse). The amplified inserts were sequenced at Agencourt Bioscience Corp (Beverly, MA, USA).

Supplementary results

The three following subsections describe the results of three sets of experiments that were carried out in order to investigate the conditions for efficient multiplexing of the RACEarray strategy. We specifically investigated the optimal combination of tissues, and the optimal distribution of primers to maximize transcript discovery, and investigate genomic distance between adjacent genes in order to delineate an optimal pooling strategy. The sets of experiments are summarized in “Multiplexing of RACEarray normalization” in main document results section and table 1; here we describe them in some more details. Each set of experiments generated a set of RACEfrags. Because the same genomic space can be highlighted by RACEfrags from different pools and cell types, we introduce the following terminology (see Supplementary Fig. S5):

Given a specific hybridization experiment,

- a RACEfrag is a genomic region (a pair of integer coordinates: start, end) identified as transcribed/expressed according to the results of the hybridization and the particular criteria to process them (see Supplementary section).

Different experiments within the same set of experiments (for instance, hybridization with RNA from different tissues) may identify the same genomic region as a RACEfrag. Therefore, given a set of experiments, and the set of all RACEfrags produced by them we introduce the following concepts:

- a non redundant RACEfrag (nrRF) is a pair of (start,end) genomic positions from the set of all (start,end) genomic positions of all RACEfrags, (that is, RACEfrags from different experiments delimited by the same genomic coordinates will represent the same non redundant RACEfrag);
- a projected RACEfrag (projRF) is a maximal set of RACEfrags that transitively overlap on the genomic sequence.

On the other hand, with respect to the annotation, RACEfrags (as well as nrRFs, and projRFs) can be classified as exonic, intronic, and intergenic. A RACEfrag is classified as exonic if it overlaps (whatever the criteria for overlap are) an annotated exon. A RACEfrag is classified as intronic if it is within the boundaries of an annoted gene, and it is not exonic. All other RACEfrags are classified as intergenic. Exonic and intronic RACEfrags are genic RACEfrags as opposed to intergenic, while intronic and intergenic RACEfrags are classfied as novel as opposed to exonic RACEfrags, which are annotated. Given an assignment of RACEfrags to index exons and genes, RACEfrags can be internal, if they map within the boundaries of the index gene (whatever the criteria for overlap are), or external if they map outside the boundaries of the index gene. Note that internal RACEfrags are always genic, while external RACEfrags can be genic or intergenic (an external RACEfrag can either map within the boundaries of a gene, or outside of the boundaries of any gene).

Optimal number and combination of tissues for RACEarray based transcript discovery

To estimate the optimal number and combination of tissues for RACEarray based transcript discovery, $5'$ and $3'$ RACEs were performed on 12 different genes from chromosomes 21 and 22 (6 from each chromosome), in 48 tissues. Both direction RACEs of 3 genes for each chromosome were pooled and hybridized to the chromosome 21 and 22 arrays, separately for the 48 tissues (96 arrays in total). These experiments gave rise to a total of $11,434$ RACEfrags which cover 0.2% of interrogated nucleotides. Eighty one percent of the RACEfrags overlap with either EST, CAGE [10] or PET [11] sequences. The remaining RACEfrags correspond to 414 non redundant RACEfrags of which 405 are novel. Detailed results can be found at `http://genome.imim.es/datasets/racearrays2007/`. Figure 3 plots the individual and cumulative genomic RACEfrag coverage per tissue. As it is possible to see, large variations in the amount of transcribed bases are observed between tissues, with about 2-fold difference between the number of transcribed bases in the human brain's hippocampus compared to muscle. The RACEfrags originating from brain's hippocampus cover 40% of the maximum cumulative genomic coverage produced by all RACEfrags considered together. As it is possible to see in Figure 3b, the cumulative genomic coverage reaches a plateau at the addition of the 16_{th} best tissue. 95% of the exons of the target genes and 64% of their nucleotides are covered by RACEfrags from at least one tissue. The tissue providing the maximal coverage of target gene exons is fetal kidney, with RACEfrags originating from this tissue covering 71% of the maximum cumulative coverage obtained with all RACEfrags considered together. Ten tissues are common to the 16 best tissues/cell types regarding total genomic coverage and the 16 best tissues/cell types regarding genomic coverage of target gene exons: Brain_Cerebelum, Brain_Frontal_lob, Fetal_Kidney, K562, Lung, Ovary, Placenta, Prostate, Testis, Tongue. Target gene exon length is covered by RACEfrags on 64%, and 95% of target gene exons are hit by RACEfrags. Differences are also observed between the relative fraction of novel transcription detected across different tissues: brain cerebellum is the tissue in which the largest fraction of detected transcription is novel (36% of the of the nucleotides covered by RACEfrags from this tissue correspond to novel transcripts (i.e. that have not been previously annotated)), while muscle is the tissue in which the smallest fraction of detected transcription is novel (10.0% of the nucleotides covered by RACEfrags from this tissue correspond to novel transcripts). While the main goal of this set of experiments was to identify the optimal number and combination of tissues that provide maximum transcript coverage, we have also used the results to compare and investigate the differences in expression pattern between novel and annotated RACEfrags. The results, which can not be extrapolated due to the small number of genes interrogated, are displayed in Supplementary figure S7. As it is possible to see, annotated RACEfrags have a wider pattern of expression than novel RACEfrags: on average annotated RACEfrags are expressed in 13.0 tissues, while novel RACEfrags are expressed in 6.8 tissues, and 62% of the annotated RACEfrags are expressed in more than one tissue, while 58% of the novel RACEfrags do.

Optimal distribution of primers along exons for RACEarray based transcript discovery

To estimate the optimal distribution of primers along exons for RACEarray based transcript discovery, $5'$ RACE reactions were performed on 10 index exons from 44 different genes, each mapping to one of the 44 ENCODE regions [4]. These reactions were performed in 12 different tissues and hybridized separately to 120 ENCODE arrays (44 reactions per pool). These experiments (Q2) lead to a total of $17,740$ RACEfrags. Detailed results can be found at `http://genome.imim.es/datasets/racearrays2007/`. The USPP filter eliminates 1.5% of the RACEfrags, which means that very few RACEfrags arise from unspecific priming and/or array cross-hybridization. Four fifth of the projected RACEfrags are internal to the gene they have been assigned to (internal projRF). This may be biased by the fact that ENCODE regions are quite small ($500kb$ for most of them). Within the last fifth of projected RACEfrags, called external, the vast majority are genic (85.7%), and within this call most of them are exonic (88.1%). A total of 437 projected RACEfrags are novel (not overlapping any annotated exon) and may represent new exons. Note that nearly 90% of them are intronic.

Interestingly, our experiments suggests that the larger the number of alternative transcripts which a given exon belongs to, the larger the number of projected RACEfrags originated by the exon). We also found a lower boundary for the maximal cDNA length an exon is able to produce, which can in turn be used to estimate, for any given gene, how many exons should be used for RACE to achieve the optimal transcript discovery. Indeed, if we consider exons included into only one transcript isoform (in order not to introduce any bias due to alternative splicing), and if we compute the total length of projRF generated by any such exon, then we find a maximal value of $1.2kb$. Finally we have found that the more internal to a gene an exon is, the less RACEfrags it generates. Indeed, if we consider genes where 10 exons have been used for RACE, order these exons according to their position within the gene (from $3'$ to $5'$), and compute the cumulative number of projRF generated by each kind of exon (orange curve in Fig. 4a), then we observe a plateau for the exons number 4 to 7, a tendency which is even clearer for novel projRF (Fig. 4b). This shows that the external exons of a gene are more informative when generating novel exons through RACE than the internal ones. These three observations can be used to design a strategy to select exons within genes for RACE priming (see “Optimal distribution of exons to RACE per locus” in main document results section, and Supplementary section and Fig. S3).

The specific design of this set of experiments allowed us to investigate the efficiency of the USSP filter (see Supplementary section), as well as the rate of false positive calls in the hybridization of RACE reactions into tiling arrays. First, we realized that a number of RACEfrags mapped $3'$ of the interrogated exon within the ENCODE region which the exon belonged to. Since only $5'$ RACE was carried on these experiments, one should in principle assume - and ignoring from the time being other molecular possibilities, such as the formation of circular cDNAs- that such RACEfrags could not be generated by the exon interrogated in this region. In total we found, before any filtering, $1,004$ such $3'$ prime incompatible RACEfrags out of a total of $17,740$ RACEfrags (5.57%). Interestingly, after filtering using the *in silico* RACEarray simulator, 185 of the $3'$ incompatible RACEfrags were eliminated (18.45%) versus only 72 of the compatible $5'$ RACEfrags (0.45%). That is, the USSP filter eliminated almost 40-fold more incompatible $3'$ RACEfrags than $5'$ compatible RACEfrags. This suggests both that a substantial fraction of the $3'$ incompatible RACEfrags are likely to originate from the amplification of unintended targets, and that the USSP filter is able to successfully identify them. After the filtering, 819 $3'$ incompatible RACEfrags remained (that is, 4.7% of the filtered RACEfrags). Since the genomic space which lies $3'$ of the interrogated exons is about $1/3$ of the total interrogated, and assuming the rate of false positives to be constant across the genome, we would initially estimate a false positive rate of about 14%.

Second, these experiments were designed under the assumption that 10 exons per gene were interrogated. Since one and only one gene was chosen from each ENCODE region, 440 exons should have been initially selected to initiate the RACE reactions. However, in a number of cases it was not possible to find a gene with at least 10 exons for some ENCODE regions (3 ENCODE regions corresponding to 30 exons), and in some other cases, even though the gene was found, it was not possible to design an appropriate primer within some of the exons selected (28 ENCODE regions corresponding to 79 exons). In total, only 331 exons were effectively used to initiate $5'$ RACE reactions. Therefore, in some hybridization experiments, less than 44 RACE reactions were pooled together before hybridization, and the corresponding ENCODE regions were not effectively interrogated. And still, we observe the presence of RACEfrags mapping to these regions that were nor interrogated. We will call such RACEfrags, ”phantom” RACEfrags, since, in principle, they should not exist. We identified 181 phantom RACEfrags (1.0% of the total). We can assume these to correspond to false positives, and assuming the false positive ratio to be constant across ENCODE regions, we expect therefore, 731 false positive RACEfrags ($181\times440/109$) for the entire set of experiments, corresponding to a RACEfrag false positive ratio of 4.2% ($731/17477\times100$).

However, the actual false positive rate is likely to be smaller, since given the long genomic space that RACEfrags appear to reach, as suggested by our experiments (see Supplementary section), it can not be ruled out that $3'$ incompatible and phantom RACEfrags are indeed bona fide $5'$ RACEfrags originating from exons of genes located in other ENCODE regions, and which RACE products have been pooled together before hybridization. Indeed, we have observed that both the number of $3'$ incompatible RACEfrags and of phantom RACEfrags is larger in the cases in which the interrogated exon maps to ENCODE regions in a chromosome containing many other ENCODE regions. If we only use the ENCODE regions mapping to chromosomes harboring only one ENCODE region (10 regions), the proportion of $3'$ incompatible RACEfrags is only 3.0%, and the corresponding estimate false positive rate would thus be of 9.0%, and the proportion of phantom RACEfrags is 0.5%, and the corresponding estimate false positive rate would be 2.0%. Some of these false positive RACEfrags are likely to be caused by amplification of the unintended primers. Indeed, while we test for uniqueness on the transcriptome, we did not test for primer uniqueness on the genome, and 25 out of the 331 primers have multiple perfect matches on the genome. Given the pervasive transcription observed across the entire human genome [3], it can not be ruled out that some of these primers amplified novel transcripts. Such amplification of unintended targets can not be detected by our RACEarray *in silico* simulator. On the other hand, the discrepancies between the false positive estimates from $3'$ incompatible and phantom RACEfrags may be revealing some events of circularization of the cDNAs amplified through RACE (which would result in $3'$ incompatible RACEfrags).

Concordance of RACEfrags with sites of transcription detected by other unbiased high-throughput interrogation techniques

We have used the experiments in the ENCODE arrays to investigate in more detail the concordance between the results of the RACEarray interrogation and of other high-throughput transcription interrogation technologies. We have specifically compared the RACEfrags with CAGE tags [10] and PET di-tags [11], as well as with all available EST data. $66,283$ CAGE tags obtained from 29 distinct RNA libraries corresponding to 15 tissues [3], and $1,339,679$ PET ditags from HCT116, MCF7, Hes3 and MCF7Estr, map to the ENCODE regions. Since CAGE and PET tags map specifically to the $5'$ end of the genes (PET tags also map to the $3'$ end, but only $5'$ RACE extensions have been attempted on the ENCODE arrays), we have considered separately all RACEfrags, and the $5'$ most RACEfrags per gene and tissue. We have identified 858 most $5'$ RACEfrags corresponding to 235 unique most $5'$ RACEfrags. Of these, 26 correspond to novel $5'$ ends of the interrogated genes. Supplementary table S2 shows the results of the comparison. Sequences are taken as overlapping if they share at least 1 bp. There is good concordance between the RACEfrags and the CAGE tags and the PET ditags. Thirty eight percent of the $5'$ most RACEfrags overlap CAGE tags, which is a proportion much larger than the one obtained when comparing the CAGE tags with a set of randomly generated RACEfrags. To generate random RACEfrags we use a program that generates a set of features of the same sizes as the the input features and located in a set of allowed regions (here the ENCODE regions). As expected, there is less overlap with the non $5'$ most RACEfrags (less enriched for real $5'$ ends of transcripts): only 11% of those overlap CAGE tags. A similar behaviour is observed when considering PET ditags: the overlap of RACEfrags (and in particular of $5'$ most RACEfrags) with them is much higher than expected by chance. On the other hand, the RACEfrag data reveals novel $5'$ ends of transcripts, which do not appear to be captured by CAGE or PET tags. Of the 26 novel non redundant $5'$ most RACEfrags, 19 do not overlap neither CAGE nor PET tags.

We have also compared the RACEfrags with all available EST data ($1,062,267$ spliced ESTs and $256,601$ unspliced ESTs map to the ENCODE regions). We have found that a large fraction of RACEfrags overlap ESTs (91.6%), but there is still a substantial number of RACEfrags that have not been detected by the large scale EST sequencing projects. Overall 414 of the novel non redundant RACEfrags do not overlap neither annotated exons, not CAGE, PET or EST sequences, and they, therefore, correspond to previously undetected exons. This clearly indicates that even in genomic regions as heavily investigated as the ENCODE regions, the RACEarray normalization is still a source of transcriptional novelty.

Reconstruction of RACEfrag expression levels from RNA transcriptional maps

We have also used the experiments in the ENCODE arrays in an attempt to reconstruct the expression values of the novel RACEfrags and compare them with those of previously annotated exons. Because of the many amplification steps involved, a serious drawback of the RACEarray normalization approach, and in general of all normalization approaches, is that quantitative information is lost on the expression levels of transcripts. We have attempted to reconstruct the expression levels of novel and annotated exons by using available transcriptome maps on the ENCODE regions generated from polyA+ RNA from a number of tissues and cell lines [4]. In particular, we have used transcript maps obtained from brain, kidney, small intestine, colon, liver, and stomach. Within each tissue, each RACEfrag has been assigned an expression value, which is the geometric mean of the intensity values of all probes overlapping that RACEfrag. Supplementary figure S8 compares the resulting expression values on annotated vs. unannotated (novel) RACEfrags whithin each of the considered tissues. RACEfrags overlapping annotated exons are in general more highly expressed than novel RACEfrags. Indeed, while only 8% of the exonic RACEfrags have no detected expresion in the transcript maps for any of the tissues investigated, this proportion is 35% for the novel RACEfrags. To get a better estimate of the expression level of the novel RACEfrags assigned to a given locus compared with the expression levels of the exonic RACEfrags in that locus, we have compared within each locus, using the Wilcoson-Mann-Whitney test, the distribution of intensity values in probes overlapping exonic RACEfrags, and in probes overlapping novel RACEfrags. At a significance level of 0.05, for 13 out of the 31 genes with novel RACEfrags assigned, we can conclude that there is no difference in the probe intensities between exonic and novel RACEfrags.

Pooling of RACE reactions. Genomic extent of loci

To estimate the range of genomic extent of loci, $5'$ RACE reactions were performed on 96 genes of chromosomes 21 and 22 in 12 different tissues. More precisely these 96 genes were divided into 16 pools so that two consecutive genes of a given pool are separated by a distance of at least $10Mb$. The 96 genes were used for RACE in 12 tissues and the reactions obtained for each pool were pooled together and hybridized in 16 different chromosome 21 and 22 arrays. From these experiments (Q3), a total of $4,012$ RACEfrags were obtained, which characteristics are described at `http://genome.imim.es/datasets/racearrays2007/`. Seventy nine percent of the RACEfrags overlap with either EST, CAGE or PET sequences. The remaining RACEfrags correspond to 693 non redundant RACEfrags, of which 682 are novel. As tissues were pooled together, and as only one index exon per gene was used, the RACEfrag dataset is not very redundant (1.3 RACEfrags per projected RACEfrag). Less than one fifth of Q3 projected RACEfrags are internal to the gene they have been assigned to, which is the opposite of the situation obtained in Q2 (Supplementary section). The other four fifth of the projected RACEfrags, which are external, are mostly genic, and then most of them are exonic. A total of $1,805$ projected RACEfrags are novel (i.e. not overlapping any annotated exon), and thus represent potentially new exons. Also, these are almost equally divided into intronic and intergenic projected RACEfrags.Strikingly, we observe RACEfrags assigned to index loci over very large genomic distances (see Fig. 5): about 50% of the novel RACEfrags are more than $3Mb$ away from the index gene. This certainly compounds the delineation of an efficient pooling strategy, since only a very sparse pooling appears to guarantee a robust assignment of RACEfrags to primers (in particular for RACEfrags mapping distal to the interrogated exons). Indeed, as Figure 5 indicates, pooling together RACE reactions originating from primers spaced at about $10Mb$ on the human genome sequence would lead only to about 85% of the novel RACEfrags correctly assigned to the closest compatible primer, which constitutes already, for such a very sparse pooling, a large degree of uncertainty. Still, we believe that higher pooling density can be achieved by an experimental design in which multiple primers per locus are independently interrogated (both in $5'$ and $3'$ orientations) across multiple conditions. Under such a scenario, the confidence of a particular connection between a RACEfrag and a primer can be scored. Connections between RACEfrags and primers observed across multiple conditions would be given more weight that connections observed in only one or a few conditions - even though very condition specific RACEfrags may actually exist. To maximize assignment resolution, primers from the same locus should be obviously hybridized in different pools, and if primers from two different loci are hybridized together in a pool, other primers from these two loci should be hybridized, whenever possible, in different pools. We are working in the delineation of a pooling strategy that takes into account the usage of multiple primers per locus across different conditions, and the associated assignment algorithm. We believe that, in this way, we can achieve high pooling density, while maintaining at the same time high assignment confidence.

References

[1] *The NCBI Handbook*, chapter UniGene: a unified view of the transcriptome. Bethesda (MD), 2003.

[2] ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science*, 2004.

[3] T.E.P. Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 2007.

[4] F. Denoeud et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in encode regions. *Genome Research*, 2007.

[5] J. Harrow et al. Gencode: producing a reference annotation for encode. *Genome Biology*, 2006.

[6] J.F. Rual et al. Human orfeome version 1.1: a platform for reverse proteomics. *Genome research*, 2004.

[7] R. Guigó, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *J Mol Biol.*, 1992.

[8] P. Kapranov, S.E. Cawley, J. Drenkow, S. Bekiranov, R.L. Strausberg, S.P.A. Fodor, and T.R. Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 2002.

[9] P. Kapranov, J. Drenkow, J. Cheng, J. Long, G. Helt, S. Dike, and T.R. Gingeras. Examples of the complex architecture of the human transcriptome revealed by race and high-density tiling arrays. *Genome research*, 2005.

[10] R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki, and P. Carninci. Cage: cap analysis of gene expression. *Nature Methods*, 2006.

[11] P. Ng, C.L. Wei, W.K. Sung, K.P. Chiu, L. Lipovich, C.C. Ang, S. Gupta, A. Shahab, A. Ridwan, C.H. Wong, E.T. Liu, and Y. Ruan. Gene identification signature (gis) analysis for transcriptome characterization and genome annotation. *Nature Methods*, 2005.

[12] G. Parra, E. Blanco, and R. Guigó. GeneID in Drosophila. *Genome Research*, 10(4):511–515, Apr 2000.

[13] K.D. Pruitt, T. Tatusova, and D.R. Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 2007.

[14] J.-F. Rual, D.E. Hill, and M. Vidal. Orfeome projects: gateway between genomics and omics. *Current Opinion in Chemical Biology*, 2004.

[15] E. Susko and A.J. Roger. Estimating and comparing the rates of gene discovery and expressed sequence tag (est) frequencies in est surveys. *Bioinformatics*, 2004.