FEBS Letters 580 (2006) 5772–5778

Compositional properties of human cDNA libraries: Practical implications

Stilianos Arhondakis, Oliver Clay, Giorgio Bernardi*

Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, 80121 Naples, Italy

Received 30 June 2006; revised 12 September 2006; accepted 19 September 2006

Available online 27 September 2006

Edited by Takashi Gojobori

Abstract The strikingly wide and bimodal gene distribution exhibited by the human genome has prompted us to study the correlations between EST-counts (expression levels) and base composition of genes, especially since existing data are contradictory. Here we investigate how cDNA library preparation affects the GC distributions of ESTs and/or genes found in the library, and address consequences for expression studies. We observe that strongly anomalous GC distributions often indicate experimental biases or deficits during their preparation. We propose the use of compositional distributions of raw ESTs from a cDNA library, and/or of the genes they represent, as a simple and effective tool for quality control.
© 2006 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

Keywords: EST; GC; Expression level; GC biases; *Sfi I*

1. Introduction

ESTs are single strand reads of transcribed sequences generated from cDNA clones [1,2], and constitute a powerful tool for gene discovery or prediction in genomic studies [3–5]. They also provide an instrument for estimating transcripts' levels and differences in gene expression between different conditions (tissues, pathological states). There are essentially two different types of libraries, non-normalized and normalized. The non-normalized libraries best reflect the population of mRNA sequences in a tissue or sample, giving better estimations of the transcripts' expression profiles and of their differential expression among different conditions [6,7]. The redundancy of highly expressed transcripts and the need to recognize also rarely expressed ones led to the development of experimental procedures, such as normalization, which reduces the frequencies of mRNA species to a narrow range. Similarly, in subtractive hybridization a pool of sequences is removed in order to leave only sequences unique to that library [8,9]. Such procedures provide only an incomplete picture of which genes are expressed at highest levels, i.e., they do not allow detailed quantitative analysis.

Our laboratory has demonstrated (i) that the density of human genes is very low in the isochore families L1, L2 and H1, which represent about 85% of the human genome, and very high in isochore families H2 and H3, which correspond to the remaining 15%; and (ii) that there are compositional correlations between GC_1, GC_2 and GC_3 (the GC levels of positions 1, 2 and 3 of codons) and the GC levels of flanking sequences (see [10] for a review). Therefore, it was proposed that the H3 isochore family presumably has the highest level of transcription because of its very high concentration of genes, especially housekeeping genes [11]. This situation raises the question of the existence of correlations between expression levels and base composition.

In the last years, ESTs, SAGE (serial analysis of gene expression [12]) and microarrrays have been used to quantify the effects of base composition on genes' expression. Estimates of the correlations between genes' expression level and base composition have, however, been often characterized by quantitative and even qualitative discordances. In an early study on mammalian expression and GC content [13], expression levels of genes were estimated from a cDNA array constructed from amygdala of *Rattus norvegicus*, and from a SAGE library of kidney from *Mus musculus*. Despite a technical variability, resulting from the fact that different samples were taken from different species and analyzed using different methods, the authors showed consistently positive correlations between the genes' expression and their base composition. A subsequent publication on gene expression [14] concluded that the human transcriptome map (HTM) contains domains called RIDGES (regions of increased gene expression) that contain several genes with high expression levels, as assessed using the SAGE technique [15]. These authors were apparently not aware of our investigations because they could have found that RIDGES essentially correspond to the GC-richest, gene-richest isochores.

A first result from our laboratory, providing further evidence of a higher transcriptional level in GC-rich mammalian genes, was obtained using human EST data [16]. In this study it was shown that averaged expression level increases steadily for three compositional classes representing GC-poor, medium and rich isochores, with statistically significant differences. This basic result was at variance with some other studies using EST data [17–19], where even a weak negative correlation was reported. The authors of those studies [17] correctly suggested that controversial results obtained using EST data might be related to limitations of ESTs for inferring quantitative expression.

In addition to data from sequencing-based techniques (EST, SAGE), high density oligonucleotide array (Affymetrix) data from a study of the human and mouse transcriptomes [20,21] have been widely used in a series of recent articles to examine relationships between expression levels and base composition

*Corresponding author. Fax: +39 081 7641355.
E-mail address: bernardi@szn.it (G. Bernardi).

0014-5793/$32.00 © 2006 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2006.09.034

[17,22–24], again giving sometimes quantitatively discordant results, even within the same technology.

Despite quantitative variation among studies, the results relating gene expression and base composition generally support the existence of a higher expression of GC-rich genes compared to GC-poor genes. The remaining discordance among EST studies [16–19] motivated us to examine the expression levels of human genes and their base composition in more detail, using data collected from a variety of tissues and by different laboratories. In particular, we examined the differences among EST/cDNA libraries' compositional properties, the reasons for those differences, and the way in which they might affect conclusions concerning base composition and expression.

We observed that EST-based estimates of genes' expression levels were often affected by strong experimental variability. The general view that transcripts of GC-rich genes tend to be more abundant than those of GC-poor genes was supported after the experimentally unreliable libraries were identified and removed. Our observations also led us to the conclusion that compositional histograms of the GC levels of ESTs, and/or of the GC_3 (third codon position GC) levels of the genes they represent, can be used to assess the quality of cDNA libraries and to recognize experimental deficits during their construction.

2. Materials and methods

In this study we analyzed ESTs as described in a recent publication of our laboratory [16]. The EST data representing different cDNA libraries were retrieved from the TIGR database, human gene index (HGI) release 16.0 (February 22, 2005; [25]). We selected cDNA libraries from the TIGR database (HGI), including only non-normalized libraries [18,19,26,27]. Indeed, it is known that normalized libraries tend to under-represent the clones of highly expressed genes, and such EST data can therefore lead to systematic underestimates of the high expression levels, i.e., only non-normalized libraries allow a more detailed reliable quantification of the detected highly expressed genes. We retained only libraries from normal adults, with one exception, fetal brain. Our final set consisted of 28 non-normalized libraries representing various tissues or samples (for experimental details see: http://cgap.nci.nih.gov/, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene, http://merops.sanger.ac.uk/ and Ref. [28]).

Each library is labelled by a catalogue number or library identifier (CAT#; those used are listed in the Supplementary Table). A cDNA library is generally assumed to be a random sample of the mRNA population for the tissue under consideration, so the number of ESTs from a given gene should ideally reflect the number of transcripts present per cell. To compute how many ESTs can be associated to each coding sequence we used the tentative human consensus (THC) sequences provided by TIGR, each of which represents an assembly of ESTs.

We will use the term 'indicative expression level' (IEL) to denote an indicator of gene expression that may, or may not, have a known and precise quantitative relationship to actual transcript levels. This term should prevent misunderstandings also when one compares expression results from different platforms that use different (and not always equally reliable) ways of gauging expression levels. For example, in an Affymetrix experiment the signal (S), or its logarithm, is an IEL, although it may not be related to the actual transcript counts by any simple, known formula.

In the CATs or libraries studied here, we chose as an indicative expression level (IEL), for each detected gene or THC, the logarithm of the quantity [16]

$$A = \frac{(\text{ESTs in THC})}{(\text{total ESTs of the CAT} - \text{singletons})}.$$

Here, 'ESTs in THC' is the number of ESTs assembled in this particular THC from the CAT and 'total ESTs of the CAT' is the total number of ESTs obtained from the CAT. This formula removes the 'singletons' that the TIGR database reports, since they apparently often represent contamination, or real but very rare transcripts [29,30].

To allow a cross-check with our previous study [16], we also monitored organism-wide expression levels, in addition to tissue- or library-specific expression levels, using 17 of the 28 tissues, each of which represents a given cDNA library and a unique tissue. The organism wide indicative expression level, E, of a gene was estimated by the formula

$$E = \frac{(A1 + A2 + A3 + \cdots + An)}{(\#\text{ of tissues where the gene is expressed})}.$$

Here, A_i is the value of A for the gene of interest in the ith tissue, and $i = 1,2,\ldots,n$ indicate the tissues in the body where the gene is expressed/detected. In the 17 unique tissues we examined, a total of 5742 CDSs were detected.

3. Results

3.1. Correlations and compositional noise

In a first analysis we evaluated organism-wide expression levels of human genes from a set of 17 cDNA libraries, each representing a unique tissue. The overall organism indicative expression level or organism IEL for these libraries (log of E; see Section 2) was estimated for the 5742 genes that were found to be expressed in one or more of the 17 tissues, and plotted against their GC_3. The weak positive correlation observed was significant ($R = 0.03$, $P < 0.01$; data not shown), but did not yet suggest any particularly strong relation for these averaged EST data, which appeared to follow a trend of weak relations reported in the past [16], in which higher GC_3 levels are associated with slightly higher average expression levels.

This study was then extended to include 11 more cDNA libraries, and EST data were now also investigated independently for each cDNA library. When we plotted our IEL, i.e., the log of the A value (library-specific expression, as described in Section 2), against the GC_3 of the genes, for each library, we observed different and often stronger positive correlations than when the same data were pooled (Supplementary Table). Such overall discrepancies between organism-wide and sample-specific results are partly expected, because pooling data from different libraries (sources) leads to an increase of sample quantity yet introduces noise into the IELs [24,31,32]. Among the 28 non-normalized libraries that we independently analyzed, we found that 9 libraries were characterized by significant positive correlations between IEL and GC_3, 2 libraries by significant negative correlations, and the remaining 17 libraries by no significant correlations (Supplementary Table). A clear dependence of the significant correlations' signs and strengths on base composition (see below) renders it unlikely that they had arisen just by chance.

This initial analysis alone suggested, but did not yet demonstrate, a persisting tendency, even if many more libraries had significant positive than significant negative correlations. The pronounced variability among the correlations and GC_3 means prompted us to carefully screen, as a next step, the compositional GC_3 distributions of the identified genes and the GC distributions of the raw EST sequences in each cDNA library. We found that the distributions were indeed often very different among libraries, also where the libraries represented the same tissue. Many of them were characterized by GC-poor or

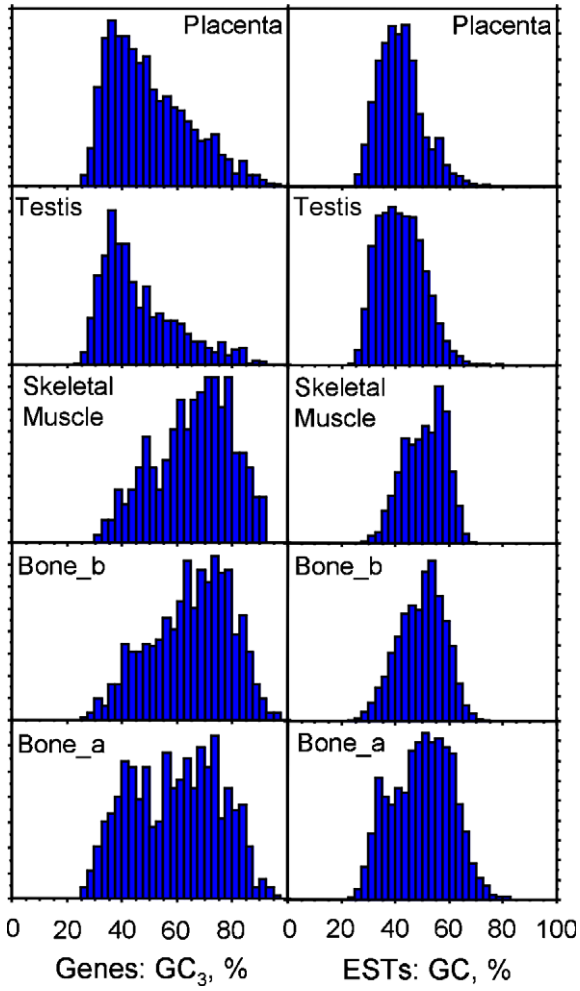

Fig. 1. Compositional distributions of the GC_3 levels of the genes expressed in five different libraries (left panel), and of the GC levels of their corresponding ESTs (raw sequences; right panel). Two of the libraries represent the same tissue (bone_a, LD97; bone_b, #A5A). The vertical axis shows the frequencies in arbitrary units (normalized to approximately same heights).

GC-rich biases that were visible as uniformly skewed histograms. Some examples are shown in Fig. 1.

3.2. Detecting effects of specific restriction enzymes on GC distributions

As mentioned above, only two libraries, from testis and skeletal muscle (#6JA; $R = -0.07$, $P < 0.02$; #9FS; $R = -0.12$, $P < 0.0001$), showed significant negative correlations between IEL and GC_3 of the genes. Both were characterized by a GC poor bias, as is shown in Fig. 1 for the testis library, and we also noticed that both of these libraries had been constructed using only the specific restriction enzyme *Sfi I*. This enzyme has a very GC-rich recognition sequence (*ggccnnnn|nggcc*), and therefore acts preferentially in GC-rich regions. The same enzyme, *Sfi I*, was also used in kidney, lung, liver and placenta libraries (#6LH; #6LI; #6QD; #6LJ; see Supplementary Table), all of which were characterized by not significant correlations that tend to be negative and by GC-poor biases. The mean GC_3 levels in those cDNA libraries where *Sfi I* was used were much lower than in the other libraries, even where they were constructed from the same tissues (see Supplementary Table). Other libraries were also characterized by different degrees of GC bias, GC-poor or GC-rich, although this remaining variability could not be traced to any particular restriction enzyme other than *Sfi I*. Fig. 1 shows two libraries from a single tissue (bone) and one from skeletal muscle, with their different correlations between the IEL and GC_3 (bone_a, LD97; bone_b, #A5A; skeletal muscle, LA1; see Supplementary Table). The two libraries with significant positive correlations are strongly biased toward GC rich genes (bone_b, skeletal muscle). The strong compositional difference observed between the two bone libraries cannot be justified by any biological variability, nor by clustering or matching errors, since it persists also for the raw EST sequences extracted from the TIGR database. In this case the difference may not be as easily explained as for the *Sfi I* libraries, but presumably it can also be traced to experimental reasons, since GC-poor regions from the cDNA were apparently eliminated during the preparation of the severely biased bone library (bone_b).

3.3. Specific restriction enzyme effects within a single tissue (prostate)

In order to track and understand the compositional effects of the restriction enzyme *Sfi I*, we enlarged our data set for a single tissue. For this particular analysis we included also libraries from pathological states, as well as normalized libraries. We selected six libraries from the same tissue, prostate, of which two libraries had been constructed using the *Sfi I* enzyme (#6LF, #6JB). The selection criterion was a high and similar number of EST sequences (>7000).

First, for each of these prostate libraries we estimated the mean GC, using raw EST sequences (i.e., not the full gene sequences that they represent). We found that the two *Sfi I* libraries had lower means (#6LF, 45.5; #6JB, 45.9) than three of the four libraries constructed using other enzymes (#8C9, 49.5; #DPH, 50.3; #8K2, 56.3; the one exception was LE55, 45.8). GC means of the tentative human consensus (THC) sequences of each library were then evaluated. The lower GC means and strong asymmetries in the GC distributions corresponding to the *Sfi I* libraries were again very noticeable. Fig. 2 shows the results obtained for a *Sfi I* library (#6JB) having 3826 THCs and a GC mean of 45.3%, and a library obtained without *Sfi I* (#DPH), having 3553 THCs and a mean of 50.0% GC.

In order to exclude the possibility of database-related artefacts related to EST "cleaning" and/or assembly methods (THCs) that might have been used by a particular database, we also looked for the same libraries in databases other than TIGR, but found no indication of database-specific biases. In particular, we randomly selected libraries from TIGR and cross-checked the raw EST sequences by re-estimating GC means for corresponding EST sets provided from a different and well-known database, UniGene-dbEST (http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi). For example, the two prostate libraries shown in Fig. 2 maintained, when retrieved from dbEST instead of TIGR, a similar mean GC whereas the *Sfi I* library gave a slightly higher mean than in TIGR (library #DPH gave 50.2% for dbEST id 14129, as in TIGR; library #6JB gave 46.6% for dbEST id 6763, versus 45.9% in TIGR). The results did not reveal any notable differences even if the number of ESTs reported for the same library often varied between these two databases.

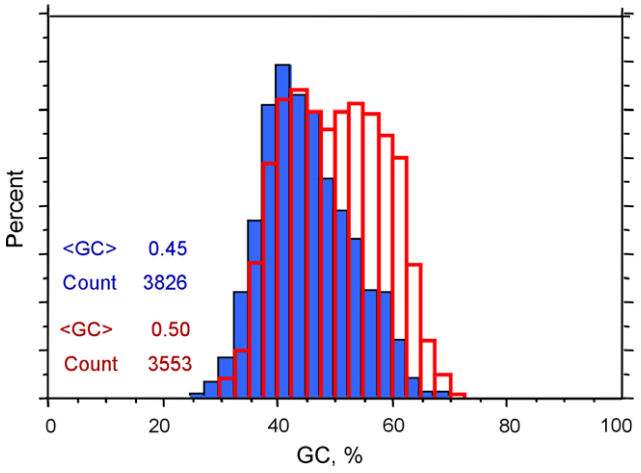

Fig. 2. Compositional distribution of the GC levels of the tentative human consensus sequences (THCs) of an *Sfi I* prostate library (#6JB, blue histogram) and a prostate library constructed without using *Sfi I* (#DPH, red transparent histogram). The vertical axis shows the frequencies in arbitrary units (normalized to approximately same heights). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Summarizing, the GC means of raw ESTs, from the 6 prostate libraries and the 28 other libraries, exhibited high intra-tissue variability, although this was apparently created largely by *Sfi I* use. More precisely, the *Sfi I* libraries were constantly found to have a GC mean below 46%, whereas the full range extended up to about 56% GC (Fig. 3), underlining the contribution of experimental GC effects. Moreover, GC effects related to another restriction enzyme, *Not*I (with the GC-rich recognition site gc|ggccgc), have been reported in a recent study of full length cDNA sequences in cow [33].

Our results and observations make it clear that EST data should, and can, be carefully pre-filtered to check for quality, for example by viewing the libraries' GC_3 distributions before applying them to further analyses involving IELs, expression breadth (number of tissues in which a gene is expressed) and/or base composition. The examples presented above show that careful preliminary screening using compositional distributions of the raw ESTs or of the expressed genes can be a well-suited tool for detecting experimental biases or deficits. By such screening one can significantly increase the reliability and compositional representativity of EST data, despite their persisting and well-known limitations.

3.4. Inter-technological variability of correlations and transcriptome composition

Despite the counterexamples reported above, there is a clearly visible trend of EST data to produce positive correlations and "avoid" negative ones, and the gene sets with mean GC_3 levels below 55% were almost all from *Sfi I* libraries (Fig. 4 and Supplementary Table). Furthermore, the libraries with GC_3 means above 55% yielded nine significant positive correlations (IEL vs GC_3) and no significantly negative ones.

After excluding the *Sfi I* libraries, the lower threshold for the libraries' mean GC_3 coincides remarkably well with a lower bound for transcriptomes that we inferred from Affymetrix arrays (from both array generations, U95Av2 and U133A; S. Arhondakis, thesis in preparation). More precisely, we analyzed data from 201 arrays (replicates) spanning a wide range of tissues [20,21,34,35]. Of these 201 replicates, 187 (93%) showed positive correlations between genes' expression level and GC_3, while only 14 showed negative correlations (for significant correlations the counts were 184 and 9, respectively). These parallel findings reinforce a lower compositional limit (mean $GC_3 \approx 55\%$) of the transcriptome, which seems to be independent of the technology used, but also a clear, platform-independent tendency of positive correlations.

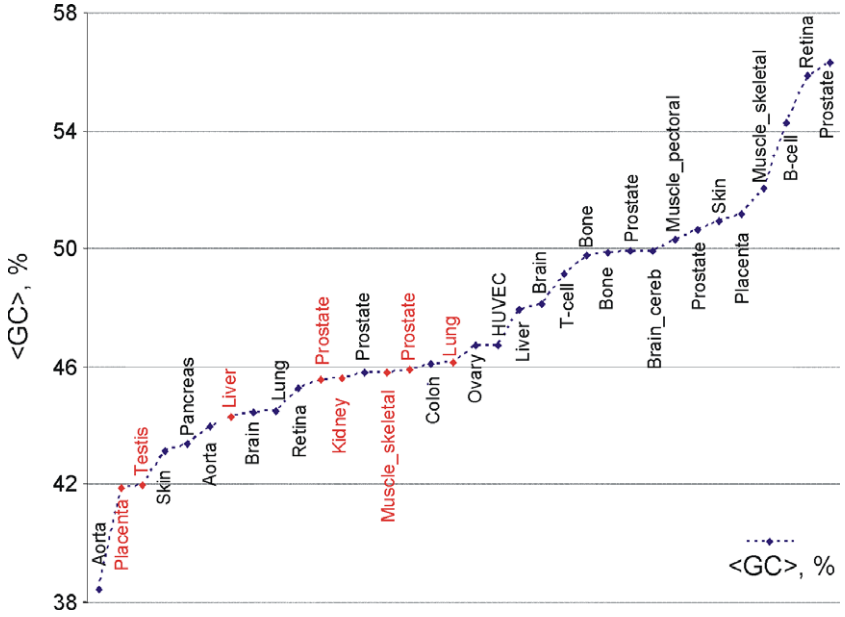

Fig. 3. Rank plot of GC means (<GC>, %), as estimated using raw ESTs, of all 34 cDNA libraries. Red names of tissues indicate the libraries constructed using the specific restriction enzyme *Sfi I*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

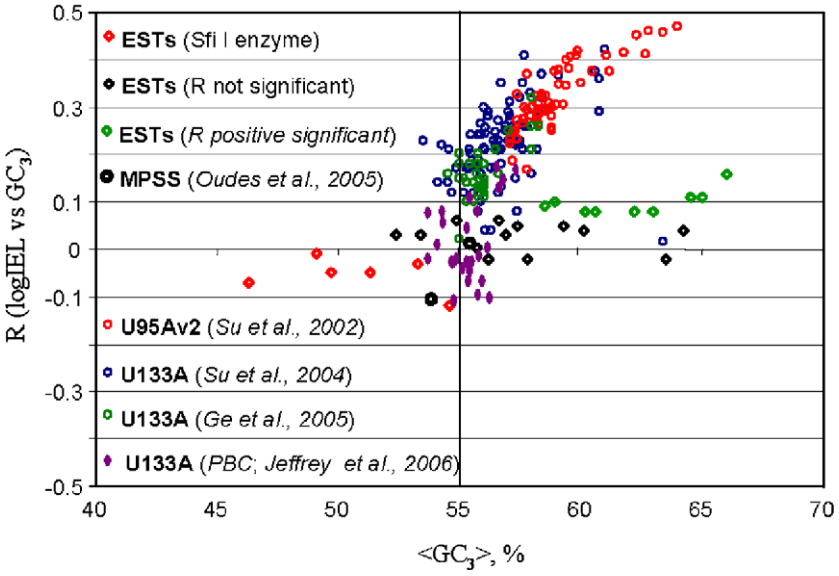

Fig. 4. Overview of compositional properties of human transcriptomes as reported by different technologies. The correlation coefficients R between IEL and GC_3 of the genes identified as present are shown plotted against their mean GC_3, for ESTs/cDNA libraries (from TIGR database, Human Gene Index/HGI, see Section 2), samples/replicates analyzed by Affymetrix arrays (U133A, two labs [21,34]; U95Av2, one lab [20]), immune system/peripheral blood cells analyzed by Affymetrix arrays (U133A, one lab [35]), and finally data from MPSS (one lab, [37]). Except for the biases introduced by the restriction enzyme *Sfi I* (red open lozenges) there is a clear, platform-independent tendency of the correlations to be positive and the GC_3 levels to be above 55%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4 reports a comparison among results from ESTs, short-oligo (Affymetrix) microarrays (both described above), and massive parallel signature sequencing (MPSS; [36]) technique. The mean GC_3 and the correlations between expression level (Affymetrix: log of Signal; MPSS: log of tpm; ESTs: log of A) and GC_3 are shown for output data sets from these three technologies. Massively parallel signature sequencing (MPSS) is represented here by two samples from cancer cell lines (LNCaP and C4-2; [37]). LNCaP has a GC_3 mean just below 55% (≈54%), while C4-2 is just above this threshold (≈56%), and correspondingly they show significant negative ($R = -0.11$) and not significant but marginally positive ($R = 0.016$) correlations, respectively. Their low GC_3 means and weak correlation coefficients may in part be related to a GC-poor bias that is visible in these data sets, especially for the LNCaP cell line (see also Figure S1 of Supplementary Material 1), or to the fact that cancer cell lines were used. It is well known that cell lines may undergo de novo methylation processes and alteration of chromatin structure [38], and genes' usual expression patterns often change when cells are kept ex vivo [39,40]. More data will be needed to faithfully represent the MPSS technique on the diagram, especially because only cultured cell lines were analyzed for this technique, in contrast to most other data represented in the scatterplot (although the Affymetrix data contain a few arrays for immune/peripheral blood cells, which may also give anomalous results cf. Ref. [35]). Details on the analysis of the two MPSS data sets and discussions are given in Supplementary Material 1.

4. Discussion

The detailed analyses presented here show that indicative expression levels of genes, as assessed by ESTs, are often affected by strong variability of cDNA libraries related to experimental protocols. This observation should in part explain the differences among the correlations that were estimated using EST data in some recent studies [16–19] and even among those obtained using different techniques within a same laboratory [17]. Technology-specific or experimental limitations involving or affecting GC may also explain the generally low concordance among expression levels reported by different technologies ([41,36]; see also Fig. 4). Indeed, GC level can be an important key for identifying methodological limitations and discrepancies [42].

GC biases related to experimental procedures, well known for the SAGE technique ([43], cf. also [44]), have been detected also for other technologies that monitor the transcriptome ([45,46] and the present work). Such biases can affect detection sensitivity, expression levels, and consequently also correlations with base composition.

The analysis and results reported here, and independent results that we obtained via high density oligonucleotide arrays (S. Arhondakis, thesis in preparation), generally point towards a persistence of positive correlations. In the case of Affymetrix, probes' GC-related artefacts may or may not have strengthened the correlations we observed, while in the case of ESTs the use of some restriction enzymes can weaken them, as we have seen here. Different inter- and intra-technological limitations of current methods for monitoring expression levels do not yet, however, allow faithful quantitative estimates of correlations between expression levels and GC, so reported values must be interpreted cautiously.

The specific restriction enzyme GC effects that we report for cDNA libraries were obtained by using, as a reference, the well-known and established bimodal, wide GC_3 distribution of human genes [42,47,48]. Such a wide compositional distribution may create technical problems, since experimental condi-

tions that are optimal for GC-poor genes may not be optimal for the GC-richest genes and vice versa. The findings reported here allow us to propose the use of compositional distributions of ESTs, or of the genes that they represent, as a simple yet effective tool for quickly visualizing and monitoring cDNA libraries, and for detecting experimental biases or flaws during their experimental preparation.

Acknowledgments: We thank Fernando Alvarez (Facultad de Ciencias, Uruguay) for constructive criticism and discussions, and our colleagues Fabio Auletta and Giuseppe Torelli for excellent bioinformatic support and for automating analyses. We also thank Dr. Margherita Branno (Laboratory of Molecular Biology, SZN) for helpful information on experimental protocols for cDNA libraries.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2006.09.034.

References

[1] Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, J.C. (1992) Sequence identification of 2,375 human brain genes. Nature 355, 632–634.
[2] Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B. and Moreno, R.F., et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656.
[3] Boguski, M.S., Tolstoshev, C.M. and Bassett Jr., D.E. (1994) Gene discovery in dbEST. Science 265, 1993–1994.
[4] Gibson, G. and Muse, S.V. (2004) A primer of genome science, 2nd ed, Sinauer Associates, Sunderland, MA.
[5] Bailey Jr., L.C., Searls, D.B. and Overton, G.C. (1998) Analysis of EST-driven gene annotation in human genomic sequence. Genome Res. 8, 362–376.
[6] Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. Genome Res. 7, 986–995.
[7] Schmitt, A.O., Specht, T., Beckmann, G., Dahl, E., Pilarsky, C.P., Hinzmann, B. and Rosenthal, A. (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nucleic Acids Res. 27, 4251–4260.
[8] Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L. and Efstratiadis, A. (1994) Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. USA 91, 9228–9232.
[9] Bonaldo, M.F., Lennon, G. and Soares, M.B. (1996) Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6, 791–806.
[10] Bernardi, G. (2004). Structural and Evolutionary Genomics Natural Selection in Genome Evolution, Elsevier, Amsterdam, The Netherlands.
[11] Bernardi, G. (1993) The vertebrate genome: isochores and evolution. Mol. Biol. Evol. 10, 186–204.
[12] Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. Science 270, 484–487.
[13] Konu, O. and Li, M.D. (2002) Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. J. Mol. Evol. 54, 35–41.
[14] Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J. and van Kampen, A.H. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. 13, 1998–2004.
[15] Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heisterkamp, S., van Kampen, A. and Versteeg, R. (2001) The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291, 1289–1292.
[16] Arhondakis, S., Auletta, F., Torelli, G. and D'Onofrio, G. (2004) Base composition and expression level of human genes. Gene 325, 165–169.
[17] Semon, M., Mouchiroud, D. and Duret, L. (2005) Relationship between gene expres-sion and GC-content in mammals: statistical significance and biological relevance. Hum. Mol. Genet. 14, 421–427.
[18] Duret, L. and Mouchiroud, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17, 68–74.
[19] Duret, L. (2002) Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12, 640–649.
[20] Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., Patapoutian, A., Hampton, G.M., Schultz, P.G. and Hogenesch, J.B. (2002) Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA 99, 4465–4470.
[21] Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., So-den, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R. and Hogenesch, J.B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA 101, 6062–6067.
[22] Vinogradov, A.E. (2003) Isochores and tissue-specificity. Nucleic Acids Res. 31, 5212–5220.
[23] Vinogradov, A.E. (2005) Dualism and GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. Trends Genet. 21, 639–643.
[24] Comeron, J.M. (2004) Selective and mutational patterns associated with gene expression in Humans: Influences of synonymous composition and introns. Genetics 167, 1293–1304.
[25] Quackenbush, J., Liang, F., Holt, I., Pertea, G. and Upton, J. (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 28, 141–145.
[26] Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. and Kondrashov, F.A. (2002) Selection for short introns in highly expressed genes. Nat. Genet. 31, 415–418.
[27] Megy, K., Audic, S. and Claverie, J-M. (2002) Heart-specific genes revealed by expressed sequence tags (EST) sampling. Genome Biol. 3, research0074.1–research0074.11.
[28] Rawlings, N.D., Tolle, D.P. and Barrett, A.J. (2004) MEROPS: the peptidase data-base. Nucleic Acids Res. 32, D160–D164.
[29] Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L. and Quackenbush, J. (2000) An optimized protocol for analysis of EST sequences. Nucleic Acids Res. 28, 3657–3665.
[30] Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Petres, G., Sultana, R. and White, J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. 29, 159–164.
[31] Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. Genome Res. 13, 2260–2264.
[32] Liu, D. and Graber, J.H. (2006) Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. BMC Bioinformatics 7, 77.
[33] Harhay, G.P., Sonstegard, T.S., Keele, J.W., Heaton, M.P., Clawson, M.L., Snelling, W.M., Wiedmann, R.T., Ven Tassell, C.P. and Smith, T.P. (2005) Characterization of 945 full-CDS cDNA sequences. BMC Genomics 6, 16.
[34] Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S.M. and Aburatani, H. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 86, 127–141.
[35] Jeffrey, K.L., Brummer, T., Rolph, M.S., Liu, S.M., Callejas, N.A., Grumont, R.J., Gillieron, C., Mackay, F., Grey, S., Camps, M., Rommel, C., Gerondakis, S.D. and Mackay, C.R. (2006) Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat. Immunol. 7, 274–283.

[36] Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. 18, 630–634.

[37] Oudes, A.J., Roach, J.C., Walashek, L.S., Eichner, L.J., True, L.D., Vessella, R.L. and Liu, A.Y. (2005) Application of Affymetrix array and Massively Parallel Signature Sequencing for identification of genes involved in prostate cancer progression. BMC Cancer 5, 86.

[38] Antequera, F., Boyes, J. and Bird, A. (1990) High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. Cell 62, 503–514.

[39] Baechler, E.C., Batliwalla, F.M., Karypis, G., Gaffney, P.M., Moser, K., Ortmann, W.A., Espe, K.J., Balasubramanian, S., Hughes, K.M., Chan, J.P., Begovich, A., Chang, S.Y., Gregersen, P.K. and Behrens, T.W. (2004) Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. Genes Immun. 5, 347–353.

[40] Moschella, F., Catanzaro, R.P., Bisikirska, B., Sawczuk, I.S., Papadapoulos, K.P., Ferrante Jr., A.W., McKiernan, J.M., Hesdorffer, C.S., Harris, P.E. and Maffei, A. (2003) Shifting gene expression profiles during ex vivo culture of renal tumor cells: implications for cancer immunotherapy. Oncol. Res. 14, 133–145.

[41] Haverty, P.M., Hsiao, L.L., Gullans, S.R., Hansen, U. and Weng, Z. (2004) Limited agreement among three global gene expression methods highlights the requirement for non-global validation. Bioinformatics 20, 3431–3441.

[42] Cruveiller, S., Jabbari, K., Clay, O. and Bernardi, G. (2003) Compositional features of eukaryotic genomes for checking predicted genes. Brief Bioinform. 4, 43–52.

[43] Margulies, E.H., Kardia, S.L. and Innis, J.W. (2001) Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acids Res. 29, E60-0.

[44] Lercher, M.J., Urrutia, A.O., Pavlicek, A. and Hurst, L.D. (2001) A unification of mosaic structures in the human genome. Hum. Mol. Genet. 12, 2411–2415.

[45] van Haaften, R.I., Schroen, B., Janssen, B.J., van Erk, A., Debets, J.J., Smeets, H.J., Smits, J.F., van den Wijngaard, A., Pinto, Y.M. and Evelo, C.T. (2006) Biologically relevant effects of mRNA amplification on gene expression profiles. BMC Bioinformatics 7, 200.

[46] Siddiqui, A.S., Delaney, A.D., Schnerch, A., Griffith, O.L., Jones, S.J. and Marra, M.A. (2006) Sequence biases in large scale gene expression profiling data. Nucleic Acids Res. 34, e83.

[47] Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C. and Bernardi, G. (1991) The distribution of genes in the human genome. Gene 100, 181–187.

[48] Zoubak, S., Clay, O. and Bernardi, G. (1996) The gene distribution of the human genome. Gene 174, 95–102.