Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 367 (2008) 542–545
BBRC
www.elsevier.com/locate/ybbrc

GC level and expression of human coding sequences

Stilianos Arhondakis, Oliver Clay, Giorgio Bernardi *

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

Received 14 December 2007
Available online 3 January 2008

Abstract

Several groups have addressed the issue of the influence of GC on expression levels in mammalian genes. In general, GC-rich genes appeared to be more expressed than GC-poor ones. Recently, expression levels of GC_3-rich and GC_3-poor versions of genes (GC_3 is the third codon position GC), inserted in vector plasmids, were compared in order to eliminate differences associated with their genomic context. Transfection experiments showed that GC_3-rich genes were expressed more efficiently than their GC_3-poor counterparts, indicating that GC_3 dramatically and intrinsically boosts expression efficiency. Here we show that, while the protocols used eliminated the original genomic context, they replaced it with the plasmid contexts whose compositional properties affected the results.

Keywords: GC level; Expression; Vector plasmids

The genomes of mammals are mosaics of isochores, fairly homogeneous, mega-size sequences covering a broad GC range [1–3]. The assessment of gene density in compositional DNA fractions led earlier to the discovery [4–6] that genes are not uniformly distributed in mammalian genomes. Indeed, in the human genome almost two thirds of the protein-coding genes are concentrated in the GC-richest isochore families H2 and H3 (the ''genome core'', which only represents 15% of the genome), the rest being spread over the vast GC-poor part (the ''genome desert''), which consists of the GC-poor isochore families L1, L2, and H1. These two gene spaces differ not only in gene density, but also in a number of other basic properties. Indeed, the genome core is characterized by shorter introns, high CpG, methylation and recombination levels, abundant CpG islands, early replication, and, of particular importance here, an open chromatin structure.

In the genome core, the higher gene density would already lead to higher transcription levels per-megabase if all genes had the same per-gene transcription level. The hypothesis that, in addition, the GC-richer genes present in such GC-rich isochores would be more highly expressed on a per-gene basis was suggested earlier [7]. Moreover, different groups [8–16] have indeed reported modest correlations between expression and GC levels in human cells/tissues under physiological conditions.

In order to study the effects of GC level on expression, a very different approach was advocated in a recent study [17]. Natural and artificially synthesized sequences having different GC_3 levels of three gene sets (Heat shock proteins, Hsp70; green fluorescent protein, GFP; and interleukin, IL-2) were inserted into plasmids and transiently or stably transfected into mammalian cells. This showed that within each group GC_3-rich sequences had higher expression levels compared to their GC_3-poor counterparts. Since such higher levels exclusively concerned mRNA production, Kudla et al. [17] interpreted them as pure effects of GC_3, with no interference by genomic contexts. The strong influence of GC_3 was seen in each gene comparison within each group, independently of the cell type and of chromosomal/extra-chromosomal locations of constructs. In the present study, we show that the compositional properties of plasmids and/or of the new genomic environment in which a construct is embedded influences gene expression, and conse-

* Corresponding author. Fax: +39 0812455807.
E-mail address: bernardi@szn.it (G. Bernardi).

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.12.155

quently, the conclusions of Kudla et al. [17] are not warranted.

Materials and methods

We retrieved the sequences of the coding regions for the three gene sets (HSP70, GFP, and IL-2) as well as those of the vectors used by Kudla et al. (2006). We then reproduced *in silico*, using the plasmid editor ApE (http://www.biology.utah.edu/jorgensen/wayned/ape/), the gene expression constructs used by the authors (for details see Ref. [17]).

Each *in silico* construct (plasmid plus insert) was then subjected to a compositional sequence analysis using the CpGPlot/CpGReport/Isochore tool (http://www.ebi.ac.uk/emboss/cpgplot/) from EMBOSS (European Molecular Biology Open Software Suite: http://www.ebi.ac.uk/emboss/; [18]) and from http://bioweb.pasteur.fr/seqanal/interfaces/isochore.html).

Results

Fig. 1 illustrates the compositional landscapes for one gene pair studied by Kudla et al. [17] in their transient and stable transfection experiments (Panels A and A′), and in the site-directed integration into human chromosomes (panel B; see also Supplementary Figure S1). A 400-bp window was used to scan the sequence because of the stability of the GC profile as seen through windows comprised between 300 and 500-bp. We remark that the inserted GC-rich gene matches the high GC level of kanamycin/neomycin resistance gene, as well as the overall GC-richness of the plasmid, whereas this is not the case for the GC-poor gene. Likewise, the GC-rich insert in the site-directed integration (panel B) matches the GC level of the other expressed genes (hygromycin resistance gene, LacZ–Zeocin fusion gene and, to a lesser extent, ampicillin resistance gene), whereas this is not the case for the GC-poor genes. To sum up, the compositional landscapes are dramatically different for the two kinds of inserts.

We note, in addition, that the integration of the vectors carrying the genes of interest into the chromosomes, assumed by the authors to be "random" preferentially occurs into open chromatin regions, which correspond to GC-rich isochores. This is indicated by consistent observations that retroviral sequences preferentially integrate in GC-rich regions of the host genomes characterized by an open chromatin structure. Moreover, integration into compositionally matching chromosomal environments allows expression whereas that in non-matching regions does not [1,3,19–23]. The preferential expression of GC-rich integrants, i.e., the plasmids carrying the genes of interest, reported by Kudla et al. [17] is therefore easily understood as a result of the matching chromosomal environment. In the case of site-directed integration (Panel B in Fig. 1, and lower panel in Supplementary Figure S1), the Flp recombination target (FRT) is thought to be located in a transcriptionally active region (http://www.invitrogen.com/content/sfs/manuals/flpintrexcells_man.pdf, "Growth and Maintenance of the Flp-In T-Rex-293 Cell Line"), i.e., in a GC-rich region.

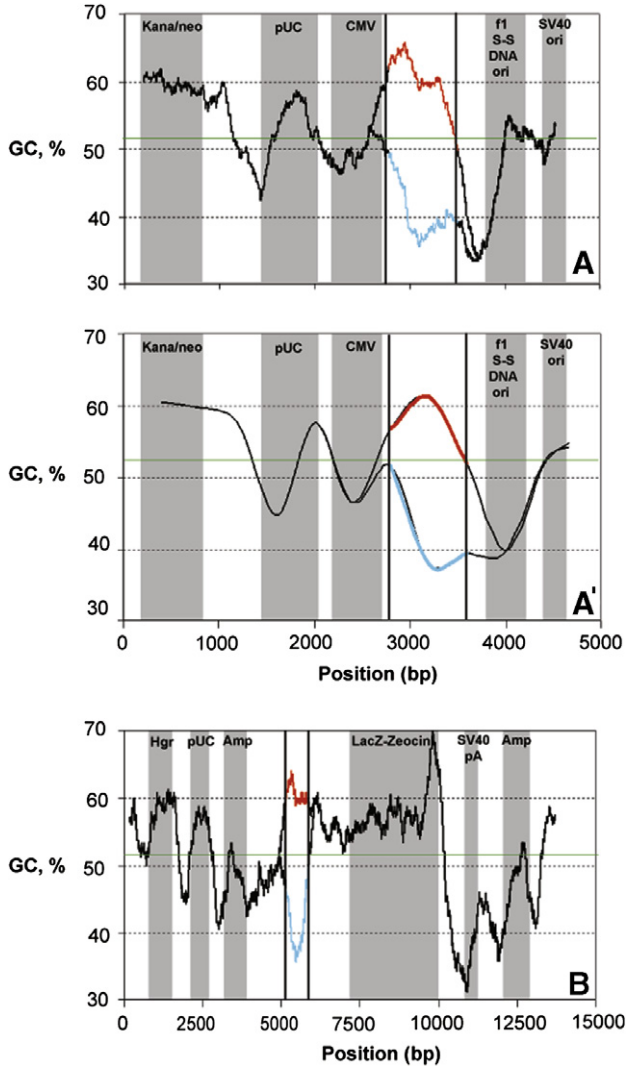

Fig. 1. Compositional patterns of the plasmids carrying the two green fluorescence protein genes, indicated by the vertical black lines, the GC-poor GFP gene, pGFP-N2 (blue profile), and its GC-rich counterpart EGFP, pEGFP-N2 (red profile). Panels A, and A' represents, the same plasmid, used in both transient and "random" stable transfection experiments, as obtained using an overlapping window of 400-bp (1-nucleotide step; panel A) or a non-overlapping 400-bp window (panel A′). Slight differences between panels A and A′ are due to the different type of windows. Panel B shows the pattern in the site-directed experiments using a 400-bp overlapping window (with 1-nucleotide step; panel B). In all panels the plasmid sequences are shown in black, their genes and elements are shadowed, while horizontal green lines indicate the average GC level of the plasmid (as estimated without the insert).

Discussion

As already anticipated in the Introduction, the recent study by Kudla et al. [17] revisited the influence of GC_3 on expression levels of genes. The authors first eliminated any genomic compensation, primarily cis-regulatory influences, that may act in the natural context of a gene, by inserting coding sequences differing in GC_3 into the same plasmids, and then compared expression levels. They concluded that GC_3 dramatically and intrinsically boosts gene

expression efficiency, and that genomic compensations masked this correlation in previous studies.

A sequence analysis of the constructs used shows, that the GC-rich plasmid and/or the GC-rich genomic environment in which the coding sequences were inserted could exert a strong influence on expression. Indeed, although the protocols used eliminated genomic compensations that may act in the natural context of the genes, they replaced them with artificial ones, in which the GC-rich sequences of the vectors provided a compositional context that favors the expression of GC-rich genes over that of GC-poor genes (see Fig. 1 and Supplementary Figure S1). As already mentioned, this effect is well known from studies on the expression of retroviral sequences integrated in compositionally matching or non-matching isochores of mammalian genomes [1,3,19–23]. Indeed retroviral sequences behave like host genes, in that the latter are expressed in a matching genomic environment [24].

Furthermore, several studies reported effects on transcription which are the consequence of structural events, such as mini-chromatinization, nucleosome assembly, that occur in transiently transfected plasmids in the nucleus [25–28]; (see Ref. [29], for a review). Similar effects will influence the expression of stably integrated transgenes [29]. Such influence of GC on the chromatin configurations that the constructs or transgenes adopt once they have reached their target positions in the interphase nucleus or on chromosomes, is relevant for both episomal and stably integrated chromosomal contexts. Constructs with a GC-rich gene inserted, but not those with a GC-poor one, will create a long, GC-rich region. This long-range difference is likely to lead to a more open chromatin structure and, the plasmid environment could thus effectively boost the transcription of the GC-richer gene and/or dampen that of the GC-poor one.

The above observations show that the conclusion of Kudla et al. [17] is not warranted. In addition, a crucial control experiment, in which GC-poor coding sequences would be inserted in a GC-poor plasmid and assessed for expression, was not done.

In conclusion, the findings of Kudla et al. [17] do not disprove the fact that the compositional context is very relevant as far as gene expression is concerned. Indeed, the simplest explanation for the high expression of the GC-rich genes is that their compositional context corresponds to open chromatin regions, favoring the accessibility of transcription factors. Finally, our approach provides an explanation, which the authors did not provide (see Ref. [30]) on why GC_3-rich genes were transcribed more than their GC_3-poor counterpart in their experiments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.12.155.

References

[1] G. Bernardi, Structural and Evolutionary Genomics, Natural Selection in Genome Evolution, Elsevier, Amsterdam, 2005.
[2] M. Costantini, O. Clay, F. Auletta, G. Bernardi, An isochore map of human chromosomes, Genome Res. 16 (2006) 536–541.
[3] G. Bernardi, The neoselectionist theory of genome evolution, Proc. Natl. Acad. Sci. USA 104 (2007) 8385–8390.
[4] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, F. Rodier, The mosaic genome of warm-blooded vertebrates, Science 228 (1985) 953–958.
[5] D. Mouchiroud, G. D'Onofrio, B. Aïssani, G. Macaya, C. Gautier, G. Bernardi, The distribution of genes in the human genome, Gene 100 (1991) 181–187.
[6] S. Zoubak, O. Clay, G. Bernardi, The gene distribution of the human genome, Gene 174 (1996) 95–102.
[7] G. Bernardi, The vertebrate genome: isochores and evolution, Mol. Biol. Evol. 10 (1993) 186–204.
[8] O. Konu, M.D. Li, Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents, J. Mol. Evol. 54 (2002) 35–41.
[9] L. Duret, Evolution of synonymous codon usage in metazoans, Curr. Opin. Genet. Dev. 12 (2002) 640–649.
[10] R. Versteeg, B.D. vanSchaik, M.F. vanBatenburg, M. Roos, R. Monajemi, H. Caron, H.J. Bussemaker, A.H. van Kampen, 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 (2003) 1998–2004.
[11] A.E. Vinogradov, Isochores and tissue-specificity, Nucleic Acids Res. 31 (2003) 5212–5220.
[12] A.E. Vinogradov, Dualism and GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth, Trends Genet. 21 (2005) 639–643.
[13] J.M. Comeron, Selective and mutational patterns associated with gene expression in humans: influences of synonymous composition and introns, Genetics 167 (2004) 1293–1304.
[14] M. Semon, D. Mouchiroud, L. Duret, Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance, Hum. Mol. Genet. 14 (2005) 421–427.
[15] S. Arhondakis, F. Auletta, G. Torelli, G. D'Onofrio, Base composition and expression level of human genes, Gene 325 (2004) 165–169.
[16] S. Arhondakis, O. Clay, G. Bernardi, Compositional properties of human cDNA libraries: practical implications, FEBS Lett. 580 (2006) 5772–5778.
[17] G. Kudla, L. Lipinski, F. Caffin, A. Helwak, M. Zylicz, High guanine and cytosine content increases mRNA levels in mammalian cells, PLoS Biol. 4 (2006) e180.
[18] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European molecular biology open software suite, Trends Genet. 16 (2000) 276–277.
[19] S. Zoubak, J.H. Richardson, A. Rynditch, P. Höllsberg, D.A. Hafler, E. Boeri, A.M. Lever, G. Bernardi, Regional specificity of HTLV-I proviral integration in the human genome, Gene 143 (1994) 155–163.
[20] A.V. Rynditch, S. Zoubak, L. Tsyba, N. Tryapitsina-Guley, G. Bernardi, The regional integration of retroviral sequences into the mosaic genomes of mammals, Gene 222 (1998) 1–16.
[21] D. Zink, A. Bolzer, C. Mayr, W. Hofmann, N. Sadoni, K. Überla, Mammalian genome organization and its implications for the development of gene therapy vectors, Gene Ther. Mol. Biol. 6 (2001) 1–24.
[22] A.R. Schröder, P. Shinn, H. Chen, C. Berry, J.R. Ecker, F. Bushman, HIV-1 integration in the human genome favors active genes and local hotspots, Cell 110 (2002) 521–529.
[23] D. Elleder, A. Pavlicek, J. Paces, J. Hejnar, Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence, FEBS Lett. 517 (2002) 285–286.

[24] M. Costantini, G. Bernardi, Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes. Gene (in press).
[25] S. Cereghini, M. Yaniv, Assembly of transfected DNA into chromatin: structural changes in the origin-promoter–enhancer region upon replication, EMBO J. 3 (1984) 1243–1253.
[26] R. Reeves, C.M. Gorman, B. Howard, Minichromosome assembly of non-integrated plasmid DNA transfected into mammalian cells, Nucleic Acids Res. 13 (1985) 3599–3615.
[27] S. Jeong, A. Stein, Micrococcal nuclease digestion of nuclei reveals extended nucleosome ladders having anomalous DNA lengths for chromatin assembled on non-replicating plasmids in transfected cells, Nucleic Acids Res. 22 (1994) 370–375.
[28] F. Recillas-Targa, A.C. Bell, G. Felsenfeld, Positional enhancer-blocking activity of the chicken beta-globin insulator in transiently transfected cells, Proc. Natl. Acad. Sci USA 96 (1999) 14354–14359.
[29] F. Recillas-Targa, Multiple strategies for gene transfer, expression, knockdown, and chromatin influence in mammalian cell lines and transgenic animals, Mol. Biotechnol. 34 (2006) 337–354.
[30] R. Robinson, More GC Means More RNA, PLoS Biol. 4 (2006) e206, doi:10.1371/journal.pbio.0040206.