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.

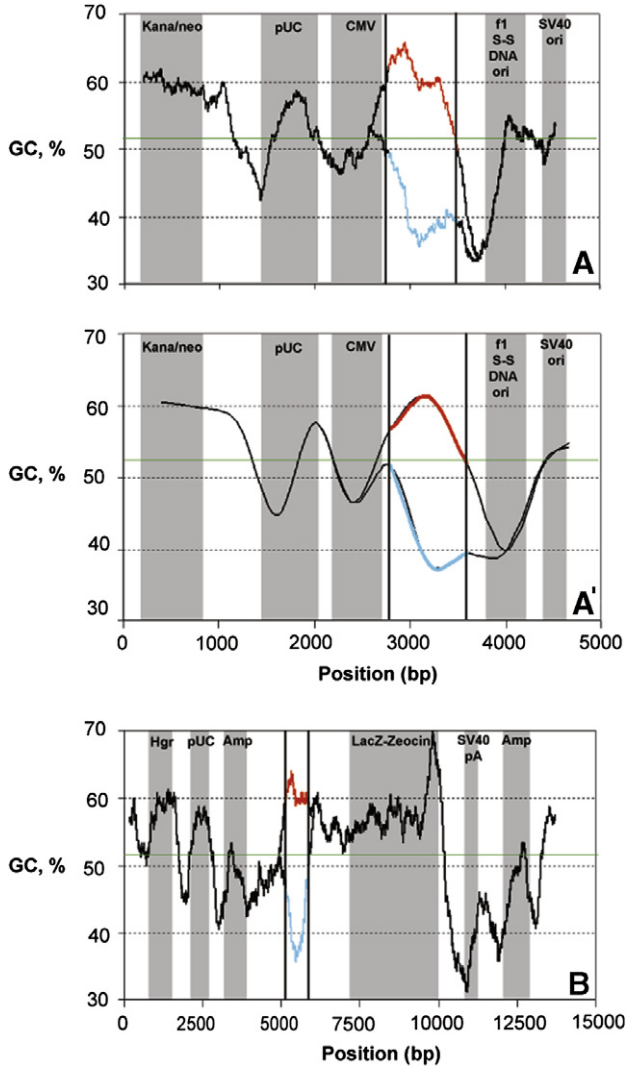

Fig. 1. Compositional patterns of the plasmids carrying the two green fluorescence protein genes, indicated by the vertical black lines, the GC-poor GFP gene, pGFP-N2 (blue profile), and its GC-rich counterpart EGFP, pEGFP-N2 (red profile). Panels A, and A' represents, the same plasmid, used in both 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