Available online at www.sciencedirect.com
ScienceDirect

Gene 419 (2008) 66–69

GENE

www.elsevier.com/locate/gene

Extrapolating ENCODE data to the whole human genome

Maria Costantini, Miriam Di Filippo, Giorgio Bernardi *

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

Received 20 December 2007; received in revised form 4 February 2008; accepted 5 February 2008
Available online 21 February 2008
Received by Takashi Gojobori

Abstract

The ENCODE (ENCyclopedia Of DNA Elements) project was launched three years ago with the purpose of identifying all of the functional elements in the human genome. ENCODE was started with 44 target sequences, which comprise 1% of the human genome. A crucial question about ENCODE is how representative it is of the human genome. Indeed, this is not a negligible problem if one considers that only 1% of the genome was selected for the project, and, more importantly, that the choice of the large DNA segments was based on two major criteria, namely the presence of extensively characterized genes and/or other functional elements, and the availability of a substantial amount of comparative sequence data. We found that the ENCODE data lead to an unbalanced representation of the compositional pattern of the human genome, especially for the GC-poorest and GC-richest regions. This unbalanced representativity of ENCODE can, however, be corrected by multiplying ENCODE data by a G/E factor (the ratio of whole genome data over ENCODE data), so amplifying the potential interest of ENCODE.

Keywords: ENCODE; Isochores; Human genome; Compositional patterns

1. Introduction

The ENCODE (ENCyclopedia Of DNA Elements) project was launched three years ago with the purpose of identifying all of the functional elements in the human genome (The ENCODE Project Consortium 2004). ENCODE was started with 44 target sequences, which comprise 1% of the human genome (about 30 Mb of 0.5–2 Mb regions). This approach already led to a number of interesting results concerning replication timing, transcription, histone methylation and acetylation, DNase I hypersensitivity and regulatory factor binding (Sabo et al., 2006; Crawford et al., 2006; Koch et al., 2007; The ENCODE Project Consortium, 2007; see also Henikoff, 2007, for comments).

A crucial question about ENCODE is how representative it is of the human genome. Indeed, this is not a negligible problem if one considers that only 1% of the genome was selected for the project, and, more importantly, that the choice of the large DNA segments was based on two major criteria, namely the presence of extensively characterized genes and/or other functional elements, and the availability of a substantial amount of comparative sequence data (The ENCODE Project Consortium 2004).

It is evident that an answer to this question can only come from comparisons of ENCODE data with analogous data at the whole genome level. The best possibility which is available is to compare the compositional distributions of the 44 ENCODE targets (downloaded from http://genome.ucsc.edu/ENCODE/) with that of isochores, which we have recently mapped on human chromosomes (Costantini et al., 2006; 2007). Isochores have been defined in terms of number, ~ 3200, and average size, ~ 1 Mb, and have been confirmed to belong in five families characterized by different GC levels and different relative amounts (Costantini et al., 2006; 2007). The rationale for a comparative approach using compositional patterns as the criterion is not only the robustness of the patterns but also the fact that a number of structural and functional properties of the

Abbreviations: ENCODE; ENCyclopedia, Of DNA Elements.

* Corresponding author. Tel.: +39 081 5833402; fax: +39 081 2455807.
E-mail addresses: mcosta@szn.it (M. Costantini), miriam@szn.it (M. Di Filippo), bernardi@szn.it (G. Bernardi).

0378-1119/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.02.013

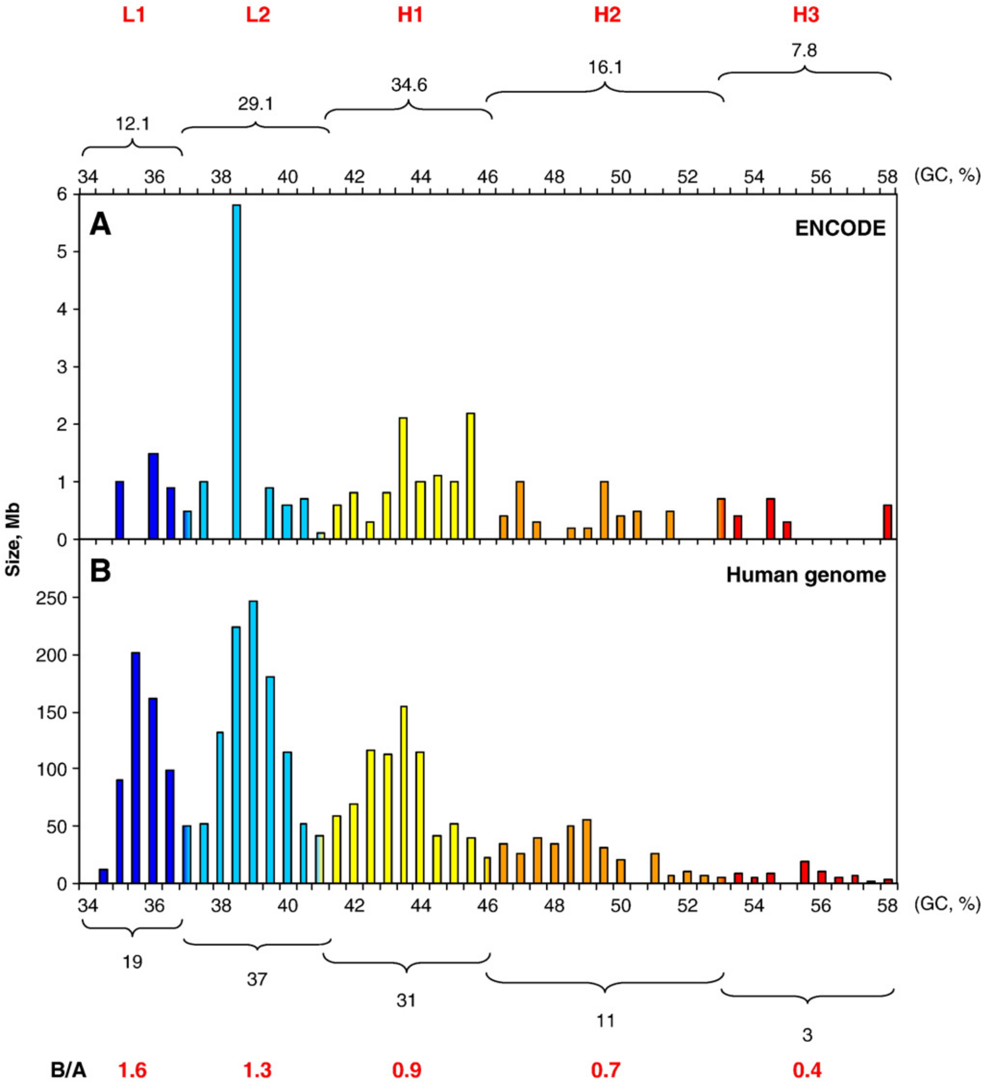

Fig. 1. (A) The compositional distribution of (A) the 44 ENCODE targets (as calculated from our isochore map; see Supplementary Table T1) is compared with (B) that of the human isochores (Costantini et al., 2006). The bar heights show the amount of DNA in each compositional interval. When the ENCODE targets covered isochore borders, they were split and assigned separately to the two corresponding compositional intervals (see Supplementary Table T1). The relative amounts of DNA per isochore family of A (top values in brackets) are compared with those of B (bottom values in brackets), the ratios B/A being given on the bottom line.

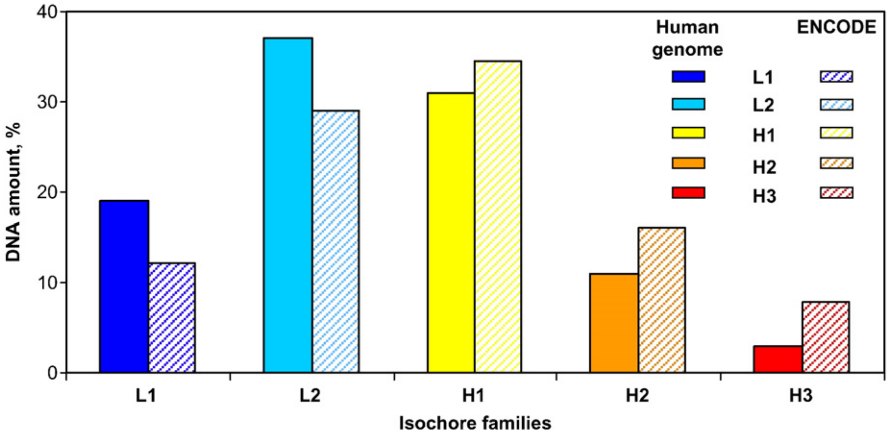

Fig. 2. The histogram shows the relative DNA amount in each isochore family as calculated from whole genome data (Costantini et al., 2006; left-hand set of bars) and from ENCODE data (right-hand set of bars).

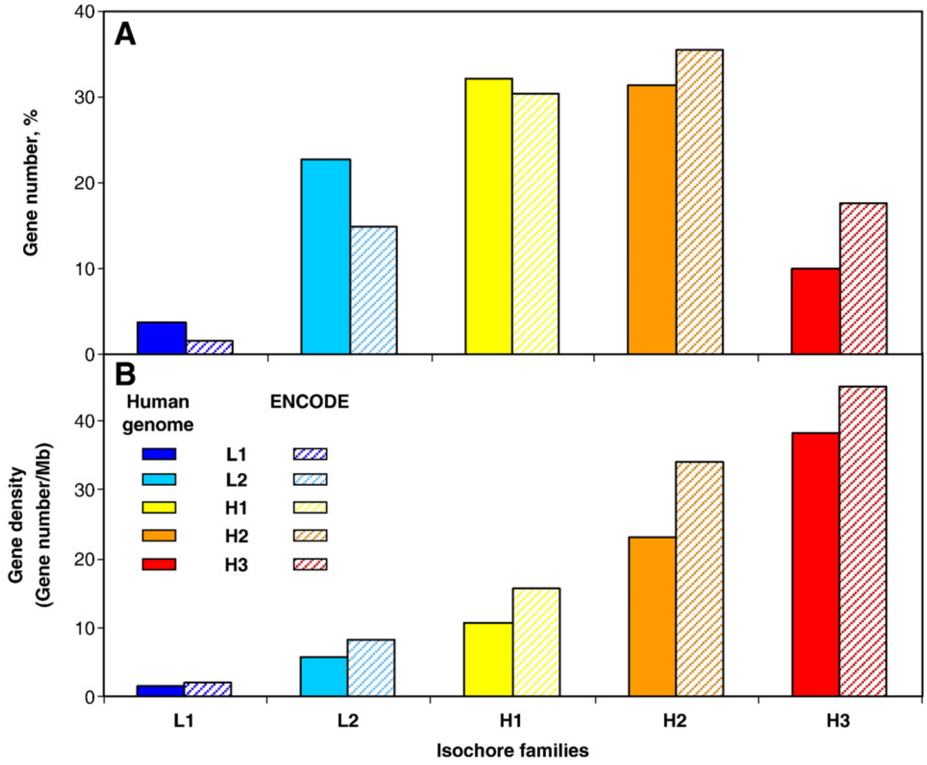

Fig. 3. (A) The histogram shows the gene number (%) in each isochore family. (B) Gene density is reported as gene number/megabases from both whole genome and ENCODE. In both histograms the left-hand set of bars refers to whole genome data (see Costantini et al., 2006), the right-hand set to ENCODE data.

genome and in particular, gene density, which was one of the two criteria used in setting up ENCODE (see above), are correlated with isochores (Bernardi, 2004; 2007).

2. Results and discussion

With this comparison in mind, we have mapped the 44 targets of ENCODE on our isochore map (see Supplementary Table T1) and put their GC levels in bins of 1% GC. We then compared the histogram so obtained with that of the isochores of the human genome. This comparison (Fig. 1) shows some similarity but also a number of differences, among which the most remarkable ones are that the GC-poorest L1 family is underrepresented (12.1% vs. 19.0%), whereas the GC-richest H3 family is overrepresented (7.8% vs. 3%) in ENCODE relative to the whole genome (see also Fig. 2).

This unbalanced representation has two implications. The first one is that, while the results presented in Fig. 1 do not question the assessment of the functional properties as directly derived from the targets, they warn against quantitatively extrapolating those properties to the whole human genome. For instance, the number of genes in H3 isochores as estimated by extrapolating from ENCODE to the whole genome would be almost double the real number, whereas it would be less than half in L1 isochores (see Fig. 3). Needless to say, these problems will tend to disappear as more sequences will be added to ENCODE and if the isochore structure of the genome will be taken as a major criterion for the choice of targets. Incidentally, the approach just outlined could also be done by comparing the gene numbers from the whole genome and as extrapolated from ENCODE data. This approach has, however, the problem of our uncertainty about the total number of genes. Finally, it worth noting that discrepancies between ENCODE and whole genome data are less striking for some isochore families in the case of gene densities (see Fig. 3).

The second implication, and the important one for practical applications, is that the data of Fig. 1 can be used to correct for the biased representativity of ENCODE. This is easily done if we use as correction factors the ratios G/E (B/A of Fig. 1) of isochore families in the whole genome over the corresponding ENCODE data. Indeed, by multiplying the ENCODE data by the G/E ratio one can correct ENCODE, so expanding the interest of ENCODE. An example of such a correction is given in Di Filippo et al. (see following paper) for DNase I hypersensitive sites (Crawford et al., 2006).

Acknowledgements

We thank Oliver Clay for comments and helpful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2008.02.013.

References

Bernardi, G., 2004. Structural and evolutionary genomics. Natural Selection in Genome Evolution. Elsevier, Amsterdam. The Netherlands, reprinted in 2005.

Bernardi, G., 2007. The neo-selectionist theory of genome evolution. Proc. Natl. Acad. Sci. USA 104 (20), 8385–8390.

Costantini, M., Clay, O., Auletta, F., Bernardi, G., 2006. An isochore map of human chromosomes. Genome Res. 16, 536–541.

Costantini, M., Clay, O., Federico, C., Saccone, S., Auletta, F., Bernardi, G., 2007. Human chromosomal bands: nested structure, high-definition map and molecular basis. Chromosoma 116 (1), 29–40.

Crawford, G.E., et al., 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat. Methods 3, 503–509.

The ENCODE Project Consortium, 2004. The ENCODE (encyclopedia of DNA elements) project. Science 306, 636–640.

The ENCODE Project Consortium, 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 779–816.

Henikoff, S., 2007. ENCODE and our very busy genome. Nat. Genet. 39, 817–818.

Koch, C.M., et al., 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res. 17, 691–707.

Sabo, P.J., et al., 2006. Genome-scale mapping of DNaseI sensitivity in vivo using tiling DNA microarrays. Nat. Methods 3, 511–518.