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

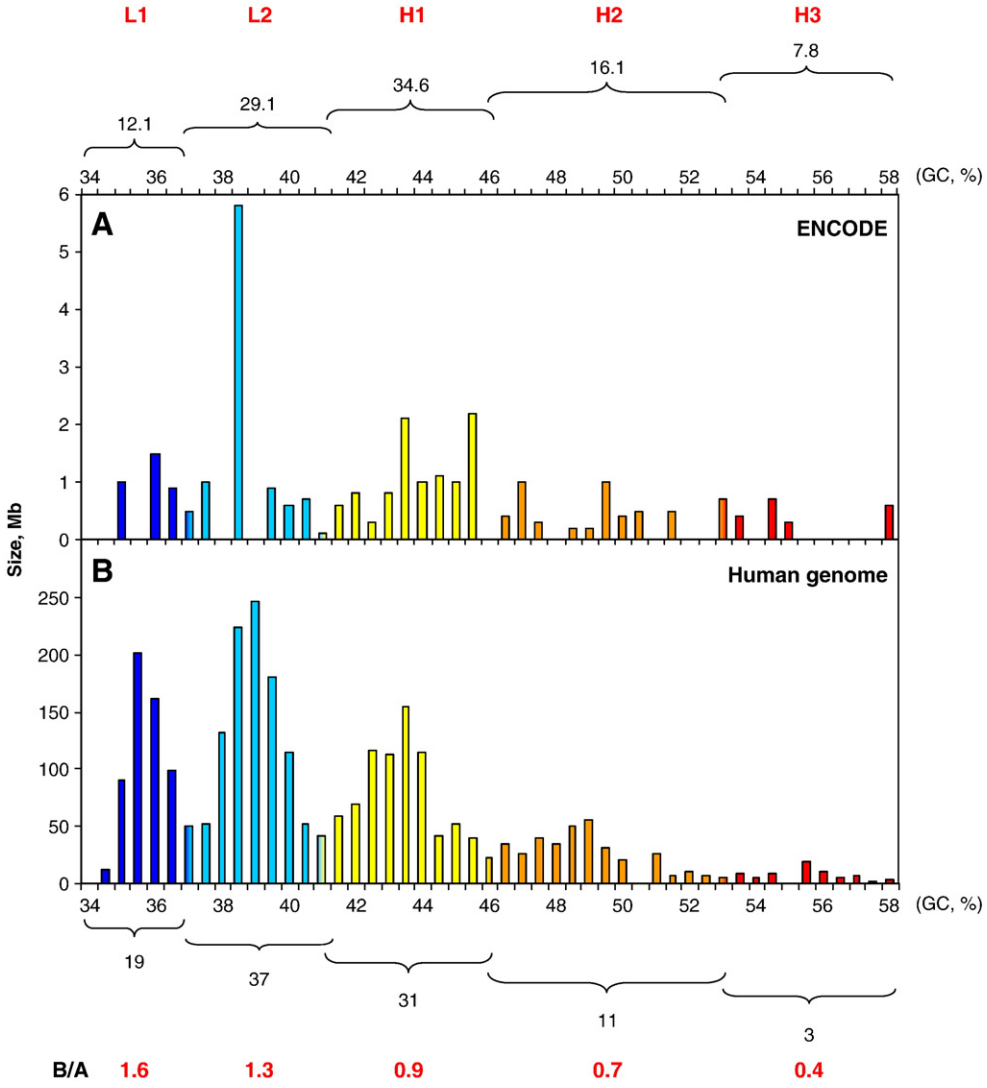

Fig. 1. (A) The compositional distribution of (A) the 44 ENCODE targets (as calculated from our isochore map; see Supplementary Table T1) is compared with (B) that of the human isochores (Costantini et al., 2006). The bar heights show the amount of DNA in each compositional interval. When the ENCODE 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.

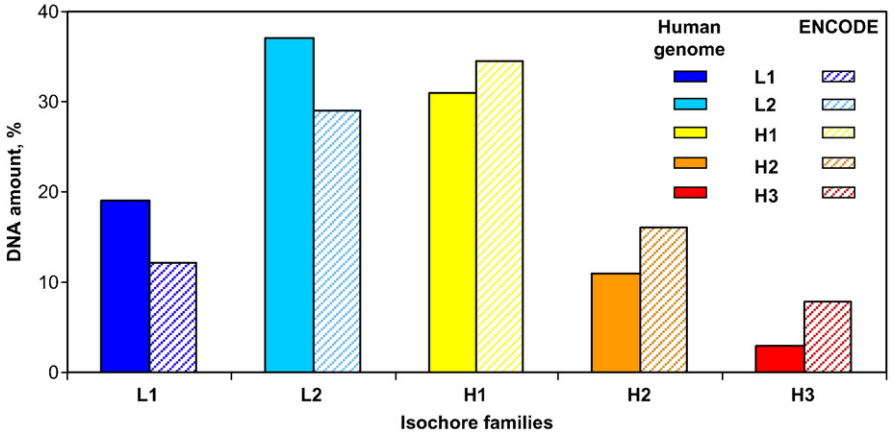

Fig. 2. The histogram shows the relative DNA amount in each isochore family as calculated from whole genome data (Costantini et al., 2006; left-hand set of bars) and from ENCODE data (right-hand set of bars).