FEBS Letters 582 (2008) 2441–2444

Prediction of replication time zones at single nucleotide resolution in the human genome

Feng Gao, Chun-Ting Zhang*

Department of Physics, Tianjin University, Tianjin 300072, China

Received 6 March 2008; revised 3 June 2008; accepted 4 June 2008

Available online 12 June 2008

Edited by Takashi Gojobori

Abstract The human genome is structured at multiple levels: it is organized into a series of replication time zones, and meanwhile it is composed of isochores. Accumulating evidence suggests a match between these two genome features. Based on newly developed software GC-Profile, we obtained a complete coverage of the human genome by 3198 isochores with boundaries at single nucleotide resolution. Interestingly, the experimentally confirmed replication timing sites in the regions of 1p36.1, 6p21.32, 17q11.2 and 22q12.1 nearly all coincide with the determined isochore boundaries. The precise boundaries of the 3198 isochores are available via the website: http://tubic.tju.edu.cn/isomap/.

Keywords: Human genome; Isochore; Replication timing; Segmentation; G + C content

1. Introduction

Since the pioneer work of Bernardi and co-workers, it has been well established that the mammalian genomes are composed of large sequence segments of fairly homogeneous G + C content that was revealed by the analytical ultracentrifugation of bulk DNA in the mid 1970s [1]. The long DNA segments (on average >300 kb) of fairly homogeneous G + C content lately were given the name 'isochores' [2]. Since then, the issues of isochores in mammalian and other eukaryotic genomes have attracted wide attention [3–17].

One of challenging problems in isochore studies is to determine how many isochores there are in the human genome, and the solution to this problem largely depends on the definition of isochores. Bernardi and co-workers recently obtained a complete coverage of the human genome by about 3200 isochores [10,11], and they also found the rules according to which the 3200 isochores of the human genome are assembled in high resolution chromosome bands (850 bands) [12]. Being different from Bernardi and co-workers [10,12], here we present an alternative solution. A large body of evidence has shown that the mammalian chromosome is organized into a number of large replication time zones [18,19], ranging from 100 kb to 2 Mb in length roughly [20]. The replication time zones (RT-zones) themselves are not the minimum replication units; rather, most of them are composed of several tandem clustered replicons [21]. It was reported that the RT-zones with higher G + C content replicate early, whereas the RT-zones with lower G + C content replicate late [6,7]. Many experiments revealed that the replication timing switch sites are consistent with the transitions of G + C content along the chromosomes investigated [6,7]. Very recent experiment shows that two adjacent replication timing sites constitute an RT-zone, which is exactly an isochore [7]. In the current work, isochores in the human genome were determined such that their boundaries coincide with replication timing sites that have been experimentally confirmed. Therefore, other isochore boundaries, in addition to those coincide with confirmed replication timing sites, are possible to have the same role, thereby boosting the number of potential replication timing sites for further experimental validation.

In our previous papers [16,17], only 56 isochores with length longer than 3000 kb were studied, which cover about 21% of the whole human genome, whereas in the present study, 3198 isochores with length longer than 30 kb were identified, which cover nearly 100% (except gaps) of the whole human genome.

2. Materials and methods

The whole human genome sequences were downloaded from http://hgdownload.cse.ucsc.edu/downloads.html#human released in March 2006. Two independent methods were used to identify boundaries of isochores. The first method is called the cumulative GC-profile (z' curve) method [15]. For a given genome or chromosome, there is a unique cumulative GC-profile or z' curve corresponding to it. The z' curve or the cumulative GC-profile is used interchangeably in this paper. Note that the essence of cumulative GC-profile is to intuitively display the variations of the G + C content along a genome or chromosome. It is not the G + C content itself. Rather, the derivative of z' curve with respect to the base position n is negatively proportional to the G + C content at the given position, i.e., $G+C \propto -dz'/dn$. Therefore, the average slope of the z' curve within a region reflects the average G + C content of the sequence within this region. If the z' curve in a region is an approximately straight line, the G + C content keeps approximately constant within this region. A jump (drop) in the z' curve indicates a decrease (increase) of the G + C content. A turning point in the z' curve indicates a switch site, at which the G + C content undergoes an abrupt change from a relatively GC-poor (GC-rich) region to a relatively GC-rich (GC-poor) region. The point at which the derivate of the z' curve is not continuous is called a turning point or segmentation point. However, a more convenient method to find the coordinates is to use a newly developed segmentation algorithm [22], which is implemented using a computer program, called GC-Profile [23]. The cumulative GC-profile, the distribution of G + C content, the isochores and their boundary coordinates for each of the human

*Corresponding author. Fax: +86 22 2740 2697.
E-mail addresses: ctzhang@tju.edu.cn (C.-T. Zhang), ren_zhang@yahoo.com (C.-T. Zhang).

0014-5793/$34.00
doi:10.1016/j.febslet.2008.06.008

chromosomes as well as the features of isochores are available online by visiting http://tubic.tju.edu.cn/isomap/. Consequently, the total number of isochores is 3198 for the whole human genome (hg18).

3. Results and discussion

3.1. Comparisons with experimental evidence

In what follows, we will show evidence that the boundaries of some isochores obtained here are in accordance with the known replication timing sites confirmed by experiments. Please refer to the related materials (such as Supplementary Figures 1–4) from http://tubic.tju.edu.cn/isomap/supplementary while reading the following text. The first evidence we show here concerns the replication timing switch site in the human major histocompatibility complex (MHC) sequence. One of the replication timing sites was found experimentally within this sequence [6]. Readers are suggested to refer to the UCSC Genome Browser in the region of chr6: 31,647,700–33,232,435, or Supplementary Figure 1. The replication timing switch region is between the gene *NOTCH4* (chr6: 32,270,599–32,299,822) in MHC class III (isochore H3) and the gene *HLA-DRA* (chr6: 32,515,625–32,520,799) in MHC class II (isochore L2). To be more accurate, the switch region should be located within the interval 32,280–32,320 kb (please refer to Fig. 6 or Fig. 8 in Ref. [6]), while the segmentation point obtained by our software GC-Profile, 32,300,166 bp within the 6p21.32 band, is situated exactly within the above interval. Therefore, the boundary between the predicted isochore 6-30 and isochore 6-31 is precisely consistent with the replication timing site confirmed experimentally [6].

The second evidence is related to the transitions in the replication timing in a 340 kb region of the human chromosome R-band 1p36.1 [24]. Please refer to the UCSC Genome Browser in the region of chr1:17,212,621–19,418,116 or Supplementary Figure 2. According to the result by GC-Profile, the precise coordinate of the segmentation point within this region is 19,288,551. Previous experimental evidence showed that there is a switch region (also called replication fork barrier by these authors) between the gene *ALDH4A1* (chr1:19,070,513–19,101,659) and the gene *RBAF600* (chr1: 19,417,172–19,450,633) [24]. Obviously, the predicted replication timing site, i.e., 19,288,551, is exactly situated within the above interval (chr1: 19,101,659–19,417,172).

The third evidence we present here is regarding the replication timing region on the human chromosome 17q11.2, in which the *NF1* gene resides. Schmegner and co-workers found that (i) a transition from a GC-poor isochore to a GC-rich one in the *NF1* region occurs within 5 kb; (ii) at the isochore transition the replication fork is stalled in the mid-S phase of cell cycle, which can be visualized by fiber-FISH techniques as a Y-shaped structure [25]. In other words, the boundary of the two isochores is also the replication timing site. It was found that the boundary between the two isochores is sharp and is located exactly at the 14 kb intergenic region between the gene *NF1* and gene *RAB11FIP4*. Please refer to the figure of UCSC Genome Browser (hg18) in the region of chr17: 26,457,053–26,954,701, or Supplementary Figure 3. The intergenic region is between the gene *NF1* (chr17: 26,446,121–26,728,820) and the gene *RAB11FIP4* (chr17: 26,742,768–26,889,352), with 14 kb in length, while the segmentation point obtained by the software GC-Profile, i.e., 26,735,738, is exactly within this region, which is just the transition site at which the G + C content varies from 37% to 51% between the isochore 17-74 (26,457,053–26,735,738) and the isochore 17-75 (26,735,739–26,954,701). The two isochores identified here, i.e., the isochore 17-74 and isochore 17-75, are exactly the two isochores studied by Schmegner et al. [25].

The fourth evidence we show here concerns the replication timing of the *MN1/PITPNB* gene region on chromosome 22 [7]. The analysis of the G + C content for this fragment of DNA sequence showed that there are four distinct sub-regions with different G + C content. The proximal sub-region has a G + C content of 50.1%, and the G + C content of the second, third and fourth sub-regions are 39.5%, 53.0% and 39.5%, respectively. The four sub-regions are termed A-, B- C- and D-isochores, respectively. Perhaps, the most surprising finding of the work is that the four isochores match the replication timing zones with such a degree that the authors called the match "perfect" [7]. The authors found that the A- and C-isochores replicate early, whereas the B- and D-isochores replicate late during the S phase of the cell cycle. There are three sharp boundaries (transitions) of the four isochores, i.e., the boundaries between isochores A/B, B/C and C/D, respectively. The analysis of the G + C content (using the window-based method) showed that the transition occurs within a region of few kb. Interestingly, the four replication zones or isochores the authors studied [7] are basically the four isochores obtained in this work. Please refer to the figure of UCSC Genome Browser (hg18) in the region of chr22: 26,129,874–26,866,656, or Supplementary Figure 4. This region consists of four distinct isochores, i.e., the isochores 22-32, 22-33, 22-34 and 22-35, respectively, separated by three sharp isochore transitions. The four isochores (RT-zones) are arranged in the following order: the GC-rich proximal isochore, the GC-poor isochore, the short GC-rich isochore and the distal isochore, respectively. Obviously, the isochore 22-32 corresponds to the isochore A; the isochore 22-33 corresponds to the isochore B; the isochore 22-34 corresponds to the isochore C and the isochore 22-35 corresponds to the isochore D. The first switch region is between the gene *MN1* (chr22: 26,474,266–26,527,486) and the gene *PITPNB* (chr22: 26,577,658–26,645,255), while the segmentation point obtained by the software GC-Profile, i.e., the predicted replication timing site, 26,577,053, is within this intergenic region. The second switch region is near the proximal part of *KIAA1043* (RP3-477H23.1-001) gene (chr22: 26,707,254–26,889,455), while the segmentation point obtained by GC-Profile, 26,703,751, is precisely located here. The third switch region is within the transition between the isochores C and D. The corresponding segmentation point is found to be 26,757,747 by GC-Profile, which is exactly the replication timing site between the isochores C and D confirmed by experiment with isochore C replicating early and isochore D late [7].

Replication timing of the human X-inactivation center (XIC) region on chromosome X was studied 7 years ago before the completion of the Human Genome Project [26]. The authors found two regions where the replication timing changes from the early to late period during the S phase of the cell cycle. The first region they found is located near a large inverted duplication segment proximal to the XIC, and the second is near the XIST locus. However, the predicted positions of replication timing have ~500 kb displacements with those reported by these authors. One of possible explanations for