DOMAIN – WIDE LANDSCAPE OF HUMAN GENOME

D.A. Maximov, V.N. Babenko

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: bob@bionet.nsc.ru

Two facts are currently inferred in the course of eukaryotic genome investigations. The first finding is that the highly expressed genes in eukaryotes maintain short introns (Petrov *et al.*, 1998; Castillo-Davis *et al.*, 2002). The second observation is that there is domain-wide regulation of gene expression in human, which comprises regions of ~80–90 genes per domain on average, exhibiting a particular level of integral expression (Gierman *et al.*, 2007; Huvet *et al.*, 2007). In this work we analyzed the features of genes in regard to the domains identified in the papers mentioned above.

We have found that there are 2 distinct groups of genes in low expressed domains. One contains extremely long genes. We observed expression preference in the brain tissue for them. Another group comprises short genes featured as cluster-structured gene loci with various activities, including testis specific and liver-specific genes.

Key words: human, chromatin, replication start, gene structure, gene length, testis-specific genes.

Introduction

It's long appreciated that there is a higher order transcriptome regulation on the level of chromatin state and its exploration is going on (Schübeler, 2007). The range of authors currently incline that the transcriptional domains are regulated to a large extent by histone modification. There is an extensive volume of papers devoted to the subject (Schübeler, 2007), but the locations and roles of histone modifications elsewhere in the genome remain unclear (Heintzman *et al.*, 2007).

Analysis of recent experimental timing data (Woodfine *et al.*, 2005) confirmed that, in a number of cases, domain borders coincide with replication initiation zones active in the early S phase (Yurov, Liapunova, 1977), whereas the center regions replicate in the late S phase (Huvet *et al.*, 2007). Around the putative replication origins, genes are abundant and broadly expressed. These features weaken progressively with the distance from putative replication origins. At the center of domains, genes are rare and expressed in few tissues.

682 successive N-shaped nucleotide compositional skew domains were identified in Huvet *et al.*, 2007. We explored the features of the genes located within the N-shaped regions (N-domains) and gained some insights on the specific structural traits for them. The genes located within domain center maintain significantly longer intergenic length. There is an abundance of very short and extra long genes which usually implies their intron length. We checked GNF atlas expression resource (Su *et al.*, 2002) to define gene expression in various parts of the domains and found that tissue-specific genes, placenta genes and liver-specific genes are presented in significant abundance in the center of the domains. Flanking genes corresponded to housekeeping genes which is consistent with the previous observations (Huvet *et al.*, 2007). It is worth noting that overall the tissue-specific genes in the center area are considered low expressed (Castillo-Davis *et al.*, 2002) compared with their flank counterparts, but they are quite intense in the terms of timing, e.g. they are highly expressed within short period of time (Huang, Niu, 2008). Genes containing long introns were specifically expressed in the nervous tissue.

We extensively explored testis-specific genes since they exhibit vivid positive selection evolution mode (Kouprina *et al.*, 2004, 2007). The intron length features of the center domain genes sought to be under neutral/disruption selection mode what makes them either extremely long or quite short (Hughes *et al.*, 2008).

Materials and Methods

Oligonucleotide microarray data were extracted from the Gene Expression Atlas [http://expression.gnf.org, (Su *et al.*, 2002)] that contains 25 human and 45 mouse non-tumoral tissues. The sample replicates corresponding to the same tissue were averaged. The signals of probes corresponding to the same gene were averaged. In total, 7735 different human mRNAs and 5297 mouse mRNAs are represented in the resulting data set. As recommended by the authors (Su *et al.*, 2002), genes whose expression level exceeded 200 arbitrary units were noted as expressed. Microarray data were available for 1276 genes in 22 normal tissues belonging to N-domains. Testis specific, liver specific and placenta specific genes were defined by 3-times larger than average log normalized expression in the corresponding tissue with at most one alternate tissue expression instance.

Gene and intron length profiles were built along the human chromosomes with overlapping window of 1 Mb in size and shift of 10 kb. The number of gene (intron) starts per window was calculated for the gene (intron) profile plot. Each gene was represented by the longest transcript unless mentioned otherwise.

Results and Discussion

We binned the genes of the 682 N-domain regions into «Late replication (LR)» gene sample as those located within 50 % of domain length in the center of the regions, and the genes comprised in 50 % of the domain territory flanking the center region («Early replication (ER)» gene sample).

Overall gene abundance (gene density) in LR set was significantly lower than in the ER gene set (2184 vs 3274 genes, $p < 1e\text{-}31$ with binomial test).

Next we compared the distribution of genes length within and between the bins (Fig. 1). It was revealed that the genes less than 5 kb in length were abundant in LR sample in comparison with ER sample (Fig. 1, $p < 0.0001$ with χ^2 2×2 table test). While overall genes' deficiency in LR region over the lengths of 7–24 kb had been observed, the extra long genes were also abundant in LR set (Fig. 1). Nearly half of the short genes were single exon genes, the abundance was also very high (Table 1). The number of merged genes was preferred in the ER genes.

Testis-specific genes are fast evolving genes possessing several tissue-specific features, namely: a) the vast majority of them are short (Su *et al.*, 2008), e.g. less than 10kb in length; b) they are organized in small clusters containing 1–5 genes which are dispersed along chromosomes (Su *et al.*, 2008). They evolved rapidly (Kouprina *et al.*, 2004, 2007; Thurman *et al.*, 2008), and usually possess rather strong tissue-specific expression level (Su *et al.*, 2008).

It has been shown that chromatin remodeling is a specific epigenetic feature of spermatogenesis (Pradeepa, Rao, 2007; Tachiwana *et al.*, 2008) as well as other tissue specific expression (Thurman *et al.*, 2008). They are present in eukaryotic genomes

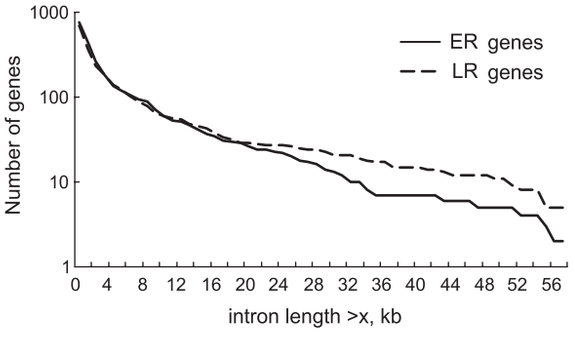

Fig. 1. Comparison of Late replication and Early replication gene sets. The maximal transcript was taken for each gene. Each point represents number of transcripts possessing the length greater than the abscissa value.

Table 1

Comparison of specific gene numbers for a range of features

Replication timing	Number of merged genes comprised within gene loci	Number of single-exon genes	Number of genes total
ER genes	199	128	3274
LR genes	82	211	2184

Given the total number of ER genes and LR genes we assess the χ^2 2×2 tables test value equal to 12,8, $p < 0{,}0002$ to reject equal distribution null hypothesis for merged genes against total gene number. The single exon genes are represented skewed towards abundance in the LR regions as well with $\chi^2 = 65{,}0$, $p < 1e\text{–}7$.

from mammals down to insects (Spellman, Rubin, 2002). While relatively constant gene number, the gene content evolves rapidly in Drosophila group (Spellman, Rubin *et al.*, 2002).

We calculated the number of testis-specific genes in the LR regions and found it significantly abundant compared to their number in the flanking regions with the p-value 1e-8 (785 vs 324 testis-specific genes). Thus we can say that testis-specific genes are attributable to the LR region.

Conclusion

We sought that the late replication origins are the key regulation units that are responsible for tissue-specific expression. We found the key structural end expression gene features are similar within ER and LR domains. Therefore we may say the domain wide chromatin structural features we observe are quite widespread phenomenon in humans and probably, in all eukaryotic species that modulate stage and tissue expression. We believe that the future studies will gain much more insight on the stage-specific mechanics of such structures.

Acknowledges

We are grateful to the BGRS 2008 chair committee for providing the opportunity to report and publish this study. We thank the referees for quite useful comments. We also thank Institute of Computational Mathematics and Mathematical Geophysics SB RAS (www.sscc.ru) for providing computing facilities. This work was partly supported by Integration Grant from Russian Academy of Science «Biodiversity and Genome Dynamics».

References

Castillo-Davis C.I., Mekhedov S.L., Hartl D.L. *et al*. Selection for short introns in highly expressed genes // Nat. Genet. 2002. V. 31. № 4. P. 415–418.

Heintzman N.D., Stuart R.K., Hon G. *et al*. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome // Nat. Genet. 2007. V. 3. P. 311–318.

Huang Y.F., Niu D.K. Evidence against the energetic cost hypothesis for the short introns in highly expressed genes // BMC Evol. Biol. 2008. V. 8. P. 154.

Hughes S.S., Buckley C.O., Neafsey D.E. Complex selection on intron size in *Cryptococcus neoformans* // Mol. Biol. Evol. 2008. 25(2). P. 247–253.

Huvet M., Nicolay S., Touchon M. *et al*. Human gene organization driven by the coordination of replication and transcription // Genome Res. 2007. V. 17. № 9. P. 1278–1285.

Gierman H.J., Indemans M.H., Koster J. *et al*. Domain-wide regulation of gene expression in the human genome // Genome Res. 2007. V. 17. № 9. P. 1286–1289.

Kouprina N., Mullokandov M., Rogozin I.B. *et al*. The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids // Proc. Natl Acad. Sci. USA. 2004. V. 101. № 9. P. 3077–3082.

Kouprina N., Noskov V.N., Pavlicek A. *et al*. Evolutionary diversification of SPANX-N sperm protein gene structure and expression // PLoS ONE. 2007. V. 2. № 4. P. e359.

Petrov D.A., Lozovskaya E.R., Hartl D.L. High intrinsic rate of DNA loss in *Drosophila* // Nature. 1998. V. 384. P. 346–349.

Pradeepa M.M., Rao M.R. Chromatin remodeling during mammalian spermatogenesis: role of testis specific histone variants and transition proteins // Soc. Reprod. Fertil. Suppl. 2007. V. 63. P. 1-10.

Schübeler D. Enhancing genome annotation with chromatin // Nat Genet. 2007. V. 39. № 3. P. 284–2855.

Spellman P.T., Rubin G.M. Evidence for large domains of similarly expressed genes in the Drosophila genome // J. Biol. 2002. V. 1(1). P. 5.

Su A.I., Cooke M.P., Ching K.A. *et al.* Large-scale analysis of the human and mouse transcriptomes // Proc. Natl Acad. Sci. USA. 2002. № 99. P. 4465–4470.

Su W.L., Modrek B., GuhaThakurta D. *et al*. Exon and junction microarrays detect widespread mouse strain- and sex-bias expression differences // BMC Genomics. 2008. V. 9. P. 273.

Tachiwana H., Osakabe A., Kimura H., Kurumizaka H. Nucleosome formation with the testis-specific histone H3 variant, H3t, by human nucleosome assembly proteins in vitro // Nucl. Acids Res. 2008. V. 36. № 7. P. 2208–2218.

Thurman R.E., Noble W.S., Stamatoyannopoulos J.M. Multi-scale correlations in continuous genomic data // Pac. Symp. Biocomput. 2008. P. 201–215.

Woodfine K., Beare D.M., Ichimura K. *et al*. Replication timing of human chromosome 6 // Cell Cycle. 2005. V. 4. P. 172–176.

Yurov Y.B., Liapunova N.A. The units of DNA replication in the mammalian chromosomes: evidence for a large size of replication units // Chromosoma. 1977. V. 60. P. 253–267.